Scalable statistical methods for microbial metagenomics

Sarah Teichman

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2024

Reading Committee:
Amy Willis, Chair
Alex Luedtke
Tyler McCormick

Program Authorized to Offer Degree:
Department of Statistics

University of Washington

Abstract

Scalable statistical methods for microbial metagenomics

Sarah Teichman

Chair of the Supervisory Committee:
Amy Willis
Department of Biostatistics

Scientific interest in microbiomes (communities of microscopic organisms in a given environment) has recently expanded due to the growing understanding of the role of the microbiome in human and environmental health, and in conjunction with the decreasing costs of metagenomic sequencing. However, there are several complications of the data that we observe from sequencing microbial samples that preclude the use of off-the-shelf statistical methods. Therefore, there is a high demand for statistical methods that are tailored to address scientific questions about microbiomes while accounting for relevant features of how the data are collected and processed. These methods must also be feasible and computationally efficient for the large scale of data that metagenomic sequencing produces.

In my first project, I present a visualization method to compare estimated gene-level evolutionary histories to estimated genome-level evolutionary histories. Evolutionary histories are best represented by phylogentic trees, which are complex graph objects made up of nodes that represent biological categories, referred to as taxa, and edges that represent the evolutionary relationships between taxa. I use a local linear approximation of phylogenetic tree space to visualize estimated gene trees as points in a low-dimensional Euclidean space. I demonstrate the utility of my proposed visualization approach through two microbial data analyses. This visualization approach is scalable for large sets of gene trees that encode a large number of taxa.

Next, I present another computationally scalable method for the analysis of metagenomic sequencing data. I extend the method of Clausen and Willis [2024] for taxonomic differential abundance analysis in order to make it computationally efficient for datasets with thousands of taxa. Through simulation, I demonstrate that my scalable method achieves similar Type I error rate control and power to the original method, and through data analyses I demonstrate that the two methods lead to very similar differential abundance conclusions. The differential abundance estimand in my method is defined with respect to a small set of reference taxa, and I suggest several approaches to choosing such a set and investigate how these approaches affect estimates and inference results through simulation and in a small data analysis.

In my third project, I consider differential abundance analyses of molecular functions. I propose a novel functional abundance model, and show that in this model, the identifiable differential abundance parameter is a function of both biological parameters and unknown sequencing effects. I develop a framework to simulate data under my functional abundance model, and use this framework to study how different magnitudes of sequencing effects affect estimation and inference of these differential abundance parameters, relative to the true biological fold-differences in abundance that are scientifically relevant. In these simulations, I find that inference on the identifiable differential abundance parameter cannot reliably be used to draw conclusions about biological fold-differences in abundance, especially in the presence of sequencing effects with large magnitudes. To address this, I suggest careful interpretation of results from the differential abundance analysis of functional data in terms of a parameter that combines biological signal with sequencing artifacts.

As a whole this dissertation presents three methods that address complex scientific questions with applications to microbiome science, each of which accounts for the effects of sequencing on microbiome data and is computationally efficient for the large scale of a typical metagenomic dataset.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank my doctoral committee: Sean Gibbons, Alex Luedtke, Tyler McCormick, Lauren Rajakovich, and Amy Willis. I am particularly appreciative of my advisor, Amy Willis, for the rigorous statistical training she has provided, as well as her infectious enthusiasm for developing scientifically relevant statistical methods.

I am very grateful for my fellow members of the Statistical Diversity Lab, and the ways in which they've shaped my growth as a statistician and a scientist. I especially appreciate Pauline Trinh, for sharing with me her wealth of knowledge about microbiome science, and Maria Valdez Cabrera, for cheering me on through the most difficult parts of my dissertation.

I would like to thank my parents and grandparents for supporting my education throughout my life. They have taught me that hard work pays off, while also demonstrating the importance of taking time to enjoy pursuits outside of work. I would like to thank my friends for bringing me joy throughout my time in graduate school. I especially appreciate my friend Adam for his willingness to engage with the many challenges that I've brought to him over the years, both statistical and personal. I would like to thank my partner, Pete, for his endless patience and kindness. Finally, I would like to thank my sister Emily for inspiring me with her determination, resilience, and passion for her work in pursuit of her own PhD.

Chapter 1

INTRODUCTION

The study of microbiomes, or communities of microbes within an environment, has become a major focus of scientific inquiry due to the increasing awareness of links between microbiomes and human and environmental health. This field was transformed by the development of next-generation sequencing (NGS) technology in the early 2000's, which drastically increased the speed and lowered the cost of large scale parallel sequencing of genetic material in microbial samples [Hu et al., 2021]. For many years, most microbiome datasets were generated using amplicon sequencing, which is targeted sequencing of the 16S ribosomal rRNA gene that is present in all bacteria [Ranjan et al., 2016]. This type of sequencing data can be used to learn about microbial community composition, as well as the abundance of each taxon in that community. An alternate sequencing approach is whole genome sequencing, in which all genes are sequenced. Whole genome sequencing data can provide taxonomic classification at a finer resolution, as well as insights into the functional capabilities of a microbiome. Over the past few years, technology and bioinformatic tools for whole genome sequencing have improved, and many microbiologists have shifted their focus from asking questions about which taxa exist in a microbiome to asking questions about what those taxa are capable of doing.

Despite major improvements in sequencing technology, metagenomic sequencing data (from both amplicon and whole genome sequencing) are subject to unknown sequencing effects. Some of these effects come from the limited capacities of sequencing instruments, which make the total abundances reported from sequencing data uninformative for total abundances in a microbial sample as defined by cell counts per unit volume [Gloor et al., 2017]. Other effects come from the differential detection of taxa across nearly all steps of the

sequencing process [McLaren et al., 2019]. Therefore, any statistical method developed with applications to the study of microbiomes with metagenomic sequencing data must address the ways in which this data reflects both true biology and artifacts of the sequencing process.

Additionally, any analysis of microbiome data requires distilling a large amount of information into interpretable summaries that can be used to draw conclusions. The process of identifying the best statistical estimand based on the scientific question and limitations of metagenomic sequencing data is rarely straightforward, and requires knowledge of the scientific context and the sequencing process. In this dissertation, I present work that aims to characterize microbiomes in terms of both taxonomic composition and function, while accounting for unknown sequencing effects. I develop three methods that each target different complex microbial estimands: gene-level phylogenies, fold-differences in taxon abundances, and fold-differences in molecular function abundances.

In the next chapter, I propose a visualization method to compare estimated gene-level evolutionary histories to estimated genome-level evolutionary histories. This is motivated by the fact that different genes in a single genome can be subject to different evolutionary pressures, which can result in distinct gene-level evolutionary histories. To address this challenge, I propose to treat estimated gene-level phylogenies as data objects, and present an interactive method for the analysis of a collection of gene phylogenies. I use a local linear approximation of phylogenetic tree space to visualize estimated gene trees as points in low-dimensional Euclidean space, and address important practical limitations of existing related approaches, allowing an intuitive visualization of complex data objects. I demonstrate the utility of my proposed approach through microbial data analyses, including by identifying outlying gene histories in strains of *Prevotella*, and by contrasting *Streptococcus* phylogenies estimated using different gene sets. In this chapter, I show that the analysis and comparison of a set of estimated gene phylogenies is a useful complement to the analysis of a single genome-level phylogeny, especially for hypothesis generation.

In the following chapter, I develop a computationally efficient method to perform inference on differential abundance parameters for analyses of taxonomic abundance from

metagenomic sequencing data. I extend the method of Clausen and Willis [2024] to make it computationally feasible for the analyses of thousands of taxa. Through simulation, I demonstrate that my inference method achieves similar Type I error rate control and power to Clausen and Willis' method. The differential abundance estimand in my method is defined with respect to a small set of reference taxa. I propose several ways to choose this reference set, and compare results through simulation and in a small data analysis. I apply my method to a differential abundance analysis of approximately $8,000$ taxa, and compare the results to those from other popular differential abundance methods. In this data analysis, my method leads to very similar estimation and inferential results as Clausen and Willis' method, while requiring a small fraction of the computational time and resources.

In my final chapter, I apply my scalable differential abundance method to differential abundance analyses of molecular function. I propose a model for functional abundances from metagenomic sequencing data that makes similar assumptions to the taxonomic abundance model studied in my third chapter. I show that unlike in the taxonomic abundance model, the identifiable differential abundance parameter in my functional abundance model is affected by taxon detection efficiencies. Through simulation, I explore the effects of varying taxon efficiency magnitudes on the differences between the identifiable parameters in my model and the true biological fold-differences in abundance that are of scientific interest. In these simulations, I find that estimation and inference based on my identifiable differential abundance parameter cannot reliably be used to draw conclusions about biological fold-differences in abundance, especially in the presence of sequencing effects with large magnitudes. I apply this approach to a functional differential abundance analysis of a large metagenomic sequencing dataset. In this analysis, I am careful to interpret results in terms of the fold-difference parameter that is affected by both biological signal and sequencing artifacts.

I conclude by drawing connections between the chapters of my dissertation and discussing ways in which scientific progress will create new and exciting open questions for statisticians in the field of microbiome science.

Chapter 2

ANALYZING MICROBIAL EVOLUTION THROUGH GENE AND GENOME PHYLOGENIES

2.1 Introduction

Microbiomes are often studied through the lens of microbial evolution, which is essential for appropriately assigning taxonomic labels to organisms, comparing communities, and reconstructing a microbial tree of life [Parks et al., 2018, Hug et al., 2016]. However, most evolutionary methods were originally developed to study either multicellular organisms, and are not always applicable to bacterial and archaeal populations [Matsen, 2015]. Thus, with the growing interest in studying microbiomes, there is an increasing need for methods designed specifically to study bacterial and archaeal evolution.

Phylogenetic trees are a key estimand in studies of evolution. A phylogenetic tree (or *phylogeny*) organizes entities (e.g., individuals, strains, or species) into a tree that reflects shared ancestry between these entities through both its branching structure (*topology*) and branch lengths. We will consider phylogenetic trees in which these entities are either genes or some representation of genomes. We refer to phylogenies that reflect relationships at the gene level as gene trees and to phylogenies that reflect relationships at the genome-level as phylogenomic trees. When pursuing phylogenomics, most microbial investigations estimate a single phylogeny that summarizes the evolution of a set of organisms (see, e.g., Brown et al. [2015], Parks et al. [2018], Imachi et al. [2020]).

While phylogenomic trees summarize evolution at the level of the organisms' genomes, individual genes in a microbial genome frequently have different evolutionary histories due to incomplete lineage sorting and/or biological processes including gene duplication and horizontal gene transfer (HGT). This can lead to discrepancies between any individual gene

tree and a phylogenomic tree, which has led to controversy over the meaning and utility of a genome-level evolutionary summary for microbes (e.g., Bapteste et al. [2009], Boto [2010], Puigbo et al. [2010]). Controversy aside, it has been noted that "mainstream applications of [genome-level] phylogenetics to [...] microbes have typically been with the idea of finding 'the' tree of such a collection rather than explicitly exploring divergence between various gene trees" [Matsen, 2015]. Exploring the divergence between gene trees is the primary motivation for this work.

In this chapter, we take the approach of viewing estimated gene trees as complex data objects that necessitate statistical tools for their analysis, and introduce a method and software for investigating gene-level evolutionary divergence in microbial genomes. We propose first estimating both a phylogenomic tree and individual gene-level phylogenetic trees. We then map this set of trees into Euclidean space and use dimension reduction methods to provide a visual comparison between genes' evolutionary histories. In contrast to traditional approaches to estimating phylogenomic relationships, which combine prespecified sets of genes to estimate a single tree, our approach uses a more expansive collection of gene sequence data to also estimate individual phylogenies. Our approach complements single-gene analyses and phylogenomic analyses by providing insight into varied mechanisms of evolution in bacterial and archaeal organisms. Specifically, it can be used to identify genes with unusual evolutionary histories. This information can then be used to remove genes deemed to be unusual or problematic from a phylogenomic approach. Additionally, this tool can be used to identify genes with common evolutionary histories and potential anomalies in the gene identification, annotation, alignment, or tree estimation processes.

2.2 Proposal

2.2.1 Overview

We propose an interactive, visualization-based approach to exploring gene-level and phylogenomic trees. While it is extremely challenging to compare large collections of phylogenetic

trees directly (e.g., through plots or summary statistics), our approach lets the user visualize a set of gene trees and a phylogenomic tree as points in two dimensions. The interactive elements of our tool allow users to (i) identify genes of interest based on the two-dimensional visualization, (ii) examine the phylogeny of the identified genes, and (iii) potentially exclude any outlying genes from an estimate of the phylogenomic tree. We believe that genes with outlying trees (e.g. those depicted as relatively distant points in the low-dimensional representation) are more likely to have been subject to notably different evolutionary pressures and therefore reflect different data generating processes, or encountered potential problems in their gene identification or alignment processes or tree estimation. Thus, outlying (or *incongruous*) trees reflect genes that the researcher may not want to include when attempting to estimate the genome-level evolutionary history of the organisms under consideration.

We give a brief overview of the approach before describing each component in greater detail. Define $\mathcal{T}_{m+3}$ as the set of edge-weighted phylogenetic trees with $m+3$ tips, $m \geq 2$. For n genes and $m+3$ genomes, we start by estimating $T_i \in \mathcal{T}_{m+3}$, the true (unknown) phylogeny of gene i, for genes $i = 1, \ldots, n$. We denote the estimate of T_i by $\hat{T}_i = \hat{T}_i(S_i)$ to reflect that this estimator is a function of the aligned gene i sequence data, which we denote by S_i. We also calculate the tree-valued summary measure $\bar{T}_p = \bar{T}_p(S_1, \ldots, S_n)$ as the phylogenomic tree estimated using gene-level sequence data $S_1, \ldots, S_n$. Once we have our set of n gene trees $\hat{T}_1, \ldots, \hat{T}_n$ along with $\bar{T}_p$, we propose two dimension reduction methods to visualize this set of trees as a scatterplot in $\mathbb{R}^q$ for small q (e.g., $q = 2$ or $q = 3$). Users then can interact with the low dimensional visualization, such as by hovering their cursor over each point on the scatterplot to view the corresponding gene tree, as shown in Figure 2.1. Users can also investigate the individual gene trees. Given this overview, we now describe each component of the method in greater detail.

2.2.2 Methodology

Our method begins by estimating the phylogenies of distinct genes in a given collection of microbial genomes and estimating a phylogenomic tree using information from all incorporated

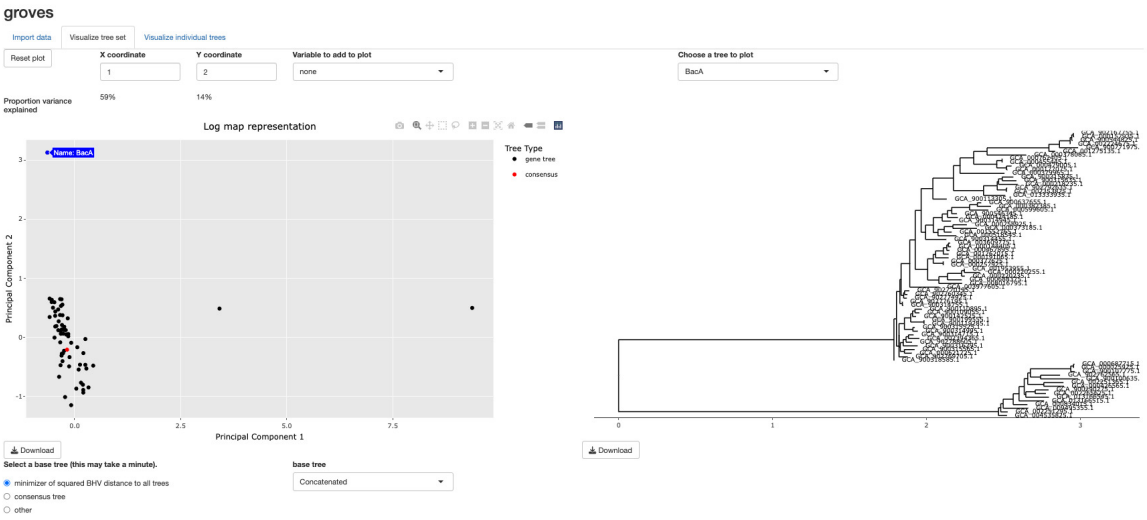


Figure 2.1: A screenshot of the interactive tool. A scatterplot showing the relationships between a collection of trees (left) can be shown alongside a selected individual gene tree (or collection of individual gene trees) (right). Additional gene-level variables such as functional annotation can also be visualized.

genes. These can be estimated via the user's choice of tree estimator, and the phylogenomic tree does not need to be estimated using the same methodology as the gene-level phylogenies. As is common in studies of microbial evolution [Wu and Eisen, 2008, Segata et al., 2013, Asnicar et al., 2020], in Section 2.3 we estimate the consensus tree via concatenation. We use the phylogenomics workflow GToTree [Lee, 2019], which by default uses prodigal [Hyatt et al., 2010] to predict genes on input genomes, HMMER3 [Eddy, 2011] to identify target genes, muscle [Edgar, 2021] to align genes, trimal [Capella-Gutiérrez et al., 2009] to trim alignments, and FastTree2 [Price et al., 2010] for approximate maximum-likelihood tree estimation. We use IQTREE2 [Minh et al., 2020] to estimate gene trees via maximum likelihood.

After estimating gene trees, we propose to consider them as objects in the Billera-Holmes-Vogtmann (BHV) space; compute their log maps with respect to a central tree; and then use principal components analysis to visualize them in a low-dimensional Euclidean space. BHV space [Billera et al., 2001], also referred to as "tree space", is a metric space for the internal branches of phylogenetic trees with the same m tips. An *internal branch* is a branch that does not lead to a tip, while a *external branch* does lead to a tip. The BHV distance $\gamma(T_i, T_{i'})$ between two trees accounts for differences in both their topologies (branching structure) and internal branch lengths. Tree space is constructed by representing each possible tree topology by a single non-negative Euclidean orthant, where each coordinate represents the length of an internal branch. The orthants are "glued together" along nearest neighbor interchange topologies. The distance between two trees is defined to be the L_2-length of the shortest possible path between them (see Figure 2.2). The shortest path between two trees is called the *geodesic*, and is unique [Billera et al., 2001, Owen and Provan, 2011]. For two trees with the same topology, the BHV distance between them is simply the L_2-distance between their internal branch lengths. In contrast, for two trees in different topologies, the geodesic path will traverse multiple orthants, and the BHV distance is the sum of the lengths of the linear path segments in each orthant that the geodesic passes through. In this way, the BHV distance encodes information about both branch length differences and topological

differences between trees. However, external branch lengths are not encoded in tree space, which we discuss later in this section.

Because of its complex combinatorial structure, it is challenging to visualize trees in their native BHV space. Instead, we find a local Euclidean approximation to BHV space around a central tree, then rely on Euclidean analysis tools such as principal coordinates analysis and scatterplots. To do this, we consider the log map [Barden et al., 2018], which is a mapping from $\mathcal{T}_{m+3}$ to $\mathbb{R}^m$ that captures both geodesic distance and local direction around a base tree. Barden et al. [2018] define the log map of a tree T from a base tree T^* as $\log_{T^*}(T) = \gamma(T^*, T)\mathbf{v}_{T^*}(T)$, in which $\gamma(T^*, T)$ is the geodesic distance between the two trees, and $\mathbf{v}_{T^*}(T)$ is a unit vector that represents the direction of the first linear segment of the geodesic path from T^* to T. The modified log map (see Figure 2.2) can be defined similarly, but as $\log^m_{T^*}(T) = \log_{T^*}(T) + \mathbf{t}^*$, where $\mathbf{t}^*$ is the coordinate in $\mathbb{R}^m$ that encodes the internal branch lengths of the base tree [Willis, 2019].

While the log map and modified log map contain information about the topology of a tree and its internal branches, they do not contain information about the external branches. However, external branch lengths provide information about evolutionary distance between entities, and we therefore believe that external branch information should be reflected in our visualizations. We therefore also define the augmented log map, $\log^a_{T^*}(T) : \left(\mathcal{T}_{m+3} \times \mathbb{R}^{m+3}_{\geq 0}\right) \to \mathbb{R}^{2m+3}$, in which the first m coordinates are given by the modified log map and the next $m+3$ coordinates are the lengths of the external branches in a predefined order (e.g., based on alphabetical ordering of the leaf labels). Thus, the augmented log map allows us to embed a collection of phylogenetic trees on $m+3$ tips into Euclidean space in a way that preserves geodesic distances and local directions from a chosen base tree, as well as external branch lengths.

Having defined the necessary mathematical infrastructure to describe our mapping from tree space to Euclidean space, we are now able to describe our method. We propose to take the augmented log map of $\hat{T}_1, \ldots, \hat{T}_n, \bar{T}_p$ to obtain $n+1$ vectors in $\mathbb{R}^{2m+3}$ representing each estimated gene tree and the phylogenomic tree. We propose two options for choosing the

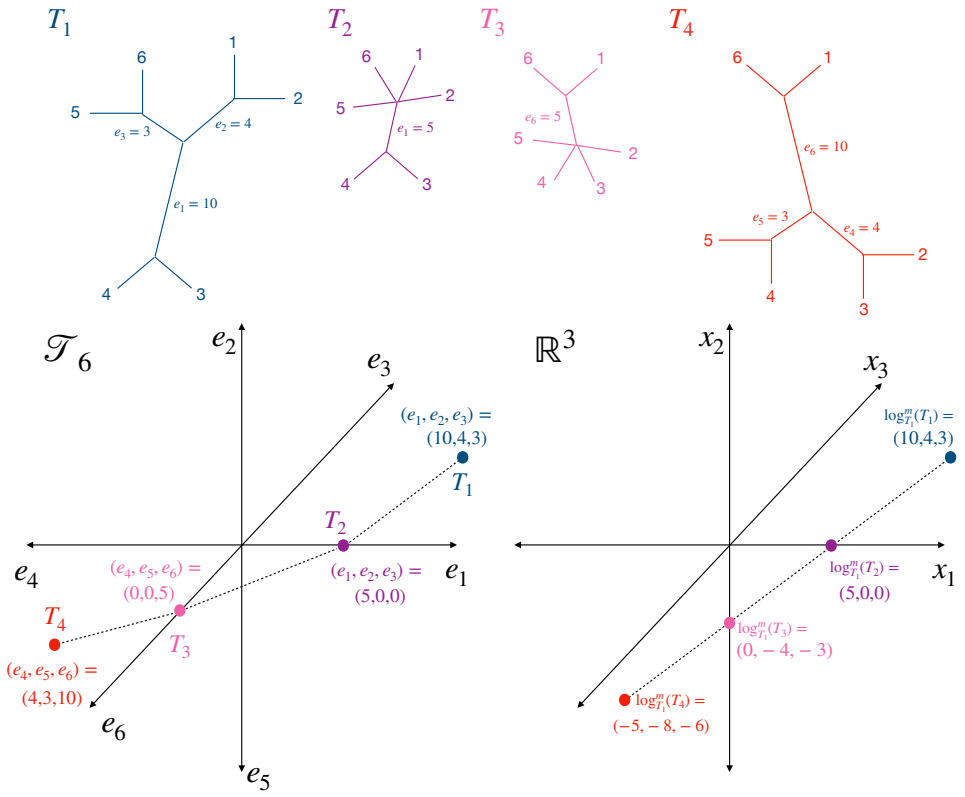


Figure 2.2: (Top panel) The geodesic path between T_1 and T_4 passes through T_2 and T_3. (Bottom left panel) A representation of the path between T_1 and T_4 in $\mathcal{T}_6$, and (bottom right panel) a mapping of $\mathcal{T}_6$ around T_1 in $\mathbb{R}^3$ via the modified log map. Each binary tree topology corresponds to a single non-negative orthant in $\mathbb{R}^3$, and the orthants are joined along axes corresponding to common branches.

base tree of the augmented log map. The first is to choose the base tree as the tree $t \in \mathbf{T}$ that minimizes $\sum_{s \in \mathbf{T}} \gamma(s,t)^2$ where $\mathbf{T}$ is the subset of trees $\{\hat{T}_1, \ldots, \hat{T}_n, \bar{T}_p\}$ that are binary; that is, have no internal branches of length zero. This will select the most central binary tree in our dataset as the base tree.

An alternative approach to selecting the base tree t is to set $t = \bar{T}_p$; that is, to use the phylogenomic tree as the base tree when computing log maps. While this tree might not be the most central (with respect to squared BHV distance), choosing $t = \bar{T}_p$ can highlight the differences between the phylogenomic tree and estimated gene trees. This can be especially useful when investigating whether the phylogenomic tree differs substantially from most gene trees, which may occur in situations when the phylogenomic tree estimation process is variable. This often occurs when the tree estimation procedure inexhaustively or stochastically explores tree space, and may be particularly an issue when using maximum likelihood for tree estimation when m is large. We provide an example and discussion of this behavior in Appendix Section A.3, along with a discussion of the advantages and disadvantages of our two proposed base tree options. Note, however, that our implementation makes it easy to toggle between different selections of the base tree. We suggest that users use this feature to investigate the sensitivity of the visualization to the base tree. We give an example of this type of sensitivity analysis in Appendix Sections A.1 and A.2.

After finding $\log_t^a(T_1), \ldots, \log_t^a(T_n), \log_t^a(\bar{T}_p) \in \mathbb{R}^{2m+3}$ for a given choice of t, we perform dimension reduction to create a low-dimensional representation of the set of trees that reflects both the trees' topologies and branch lengths. Our tool implements principal components analysis (PCA) of augmented log map vectors for dimension reduction. In our data examples shown in Section 2.3, we use the first two principal coordinates of our $n+1$ $2m+3$-dimensional vectors to visualize our tree-valued data objects.

An advantage of our BHV space-based approach is that it enables the exploration of variation in phylogenies with respect to differences in tree topologies and branch lengths. Even for users who are primarily interested in comparing trees topologically, we believe that branch lengths provide information about both the extent of evolution as well as the

uncertainty in tree estimation. For example, long branches represent alignments with more bases or amino acids that support a given split in the tree, which corresponds to relatively low uncertainty in the presence of that split in the tree. Correspondingly, shorter branches represent splits with fewer divergent sites. Thus, even for researchers who are primarily interested in exploring differences in tree topologies, our proposed approach still provides important information about uncertainty in topology via the magnitudes of branch lengths. This interplay between topology and branch lengths is a key advantage of BHV space-based analysis: a branch length shrinking to zero is equivalent to moving towards a topological change.

While the above describes our proposed method, we also permit a number of modifications that provide more flexibility for users. For example, we allow the option to construct a multidimensional scaling (MDS) visualization using either BHV distances or Robinson-Foulds (RF) distances between trees. MDS with the RF distance provides a visualization method for researchers who are interested in comparing trees based on topology alone (without regard for branch lengths). In addition, we provide various options for MDS, including both metric MDS and nonmetric MDS. Our software implementation is easily extensible, allowing for further expansion of the possible variations of our general methodology. In practice we find that our proposed approach is highly interpretable, producing scatterplot visualizations where most trees are distributed in all directions around the base tree. This distribution of points in the scatterplot is intuitive and allows for straightforward visual diagnosis of outliers.

2.2.3 Relationship to other work

While a number of studies and tools also make use of low-dimensional representations of a set of trees for analysis, we believe that the novelty of our proposal is three-fold. Firstly, while other methods compute distances between trees and use MDS for dimension reduction [Amenta and Klingner, 2002, Holmes, 2006, Chakerian and Holmes, 2012, Kendall and Colijn, 2016, Gori et al., 2016, Huang et al., 2016, Jombart et al., 2017], we use the log map to find

an approximation in Euclidean space for each tree and use PCA for dimension reduction. We believe that PCA has many advantages not shared by MDS, which scales poorly when the number of trees is large because all pairwise distances between trees must be computed. In addition, PCA has advantages for reproducibility when adding new trees to a visualization or comparing across different studies [Willis and Bell, 2018].

Secondly, our approach addresses a key practical limitation of the related work of Willis and Bell [2018], who also proposed PCA on log map-transformed gene trees. Willis and Bell [2018] proposed to use the Fréchet mean of the gene trees as the base tree for computing log maps. Unfortunately, Fréchet means of microbial gene trees are almost always non-binary, including in both examples considered in Section 2.3. Non-binary trees fall in low-dimensional strata of BHV tree space, resulting in additional distortion and compression of tree space in comparison to log map projections from binary base trees. Therefore, to address the limitation of non-binary Fréchet mean trees, we proposed two alternative options as base trees: the phylogenomic tree, or the binary tree that minimizes the sum of squared BHV distances. The minimization approach is necessarily binary by construction, and we have never seen an example of a non-binary phylogenomic tree. The latter can be explained by the prevalence of concatenation for microbial phylogenomics, which results in many genetic sites for tree estimation and thus the identification of distinct lineage events. Thus, our proposal addresses a prohibitive limitation of the Willis and Bell [2018] methodology, making it applicable to a wider variety of datasets while retaining the advantages of a BHV-based analysis.

Finally, while a number of tree-based visualization methods have analyzed eukaryotic evolution [Hillis et al., 2005, Nye, 2011, Kendall and Colijn, 2016, Gori et al., 2016, Willis and Bell, 2018], our focus on bacterial and archaeal gene trees supports the continuing expansion of microbiome research. Bacterial and archaeal gene trees typically display a greater degree of incongruity than eukaryotic gene trees due to the comparatively high frequency of gene duplication, loss, and horizontal transfer in these domains. Our emphasis on detecting gene-level phylogeny outliers reflects the greater variability in the distribution of gene

trees of bacteria and archaea. Unlike other modern tools for viewing a collection of trees [Huang et al., 2016, Jombart et al., 2017], our tool lets the user choose a tree in the low-dimensional visualization and see its estimated phylogeny, providing rapid insight into why certain trees might appear as outliers. By visually identifying the phylogenomic tree in the two-dimensional visualization and supporting gene selection for phylogenomic tree estimation, our tool is well-suited to improving modern microbial phylogenetic estimation as it is currently practiced.

2.3 Data Analyses

We envisage three primary use cases for our exploratory method. We briefly describe these settings before illustrating the approach on bacterial genome datasets.

- Identifying gene-level phylogeny outliers: Any individual gene tree can differ from the remaining trees in the collection through its topology, branch lengths, or both. Our method can be used to identify outlying gene trees via the low-dimensional visualization, then explore more deeply by plotting the trees itself. This approach can be used to generate hypotheses about drift-based or selective evolution on specific microbial traits, or help identify genes that may have undergone HGT. In Section 2.3.1 we analyze Prevotella genomes and identify and interrogate three genes that are outlying in their phylogenies.

- Contrasting multiple sets of genes for phylogenomics: In microbiome data analysis, phylogenetic trees are commonly estimated by concatenating alignments of multiple genes. However, it is not common practice to investigate the robustness of the estimated tree to the choice of genes that are input to the alignment. Our tool streamlines investigating robustness by easily allowing the removal of genes from a concatenated alignment then re-estimating the phylogenomic tree. We illustrate this approach by comparing the phylogenomic tree estimated using only ribosomal genes for Streptococcus genomes to a tree estimated using a fuller set of single-copy orthologous genes in Section 2.3.2.

- Highlighting anomalies in preprocessing: Finally, our tool can flag potential issues with gene identification, sequence alignments, and tree estimation. For example, an outlying tree may be caused by a tree estimation algorithm that failed to converge, or an issue performing the multiple sequence alignment. We provide a brief example of this use case in Section 2.4.

For full details on software and database versions used in this section, please see Supplementary File `software-version.csv`.

2.3.1 Prevotella

Prevotella is a genus of gram-negative anaerobic bacteria that are present in the human gut, oral, and vaginal microbiomes. The abundance of Prevotella in the human gut microbiome is associated with geography, lifestyle, and diet [Tett et al., 2021]. Understanding the relatedness of different strains within this genus is essential in studying concepts such as patterns of ecological niches, genomic elasticity, diversification, and host-microbe-phage interactions.

We wish to estimate a phylogenomic tree for the Prevotella genus, to explore gene-level histories of Prevotella genes, and to understand the concordance between the gene-level and phylogenomic trees. In order to have broad representation across the Prevotella genus and work with high quality genomes, we consider the 383 species representatives in the Genome Taxonomy Database (GTDB) [Parks et al., 2018, 2020]. To develop a gene set, we first use HMMER [Eddy, 2011] to identify all protein families from the Pfam database [Mistry et al., 2021] in our Prevotella genomes that are present in a single copy in a subset of the genomes. We order the Pfams (genes) by their prevalence in the Prevotella genomes, omitting those in the lowest 10% of prevalence, resulting in 63 genes for analysis. We then consider only genomes for which all of these genes are present, leaving us with 78 genomes and 63 genes. These choices result in a set of genomes that each contain all of our genes of interest, covering much of the breadth of the Prevotella genus as well as a set of genes that are broadly prevalent across the genus. Furthermore, manual comparison of 63 trees with 78 tips would

be intractable, motivating our streamlined method for their analysis.

After constructing our gene and genome set, we build gene level and concatenated alignments using GToTree [Lee, 2019], which uses prodigal [Hyatt et al., 2010] to predict genes on input genomes, HMMER3 [Eddy, 2011] to identify target genes, muscle [Edgar, 2021] to align genes, and trimal [Capella-Gutiérrez et al., 2009] to trim alignments. We use IQ-Tree [Minh et al., 2020] to estimate $\hat{T}_1, \ldots, \hat{T}_{63}$ and $\bar{T}_p^{\text{full}}$, where $\bar{T}_p^{\text{full}} = \bar{T}_p(S_1, \ldots, S_{63})$. We then construct the visualization described in Section 2.2.2, choosing the minimal BHV-distance binary tree as the base tree for our log maps. Interestingly, for this analysis, this tree is $\bar{T}_p^{\text{full}}$, and thus, our two options for choosing the base tree (the phylogenomic tree or the binary tree that minimizes the sum of squared BHV distances) coincide in this instance. A static rendering of our visualization is shown in Figure 2.3. The first principal component explains 59% of the variance in the log map vectors for the set of trees, and the second principal component explains 14%. Note that these are the proportions of variance in the log map vectors, and not proportions of the variance between trees in BHV tree space. Because the log map transformation can map multiple trees in BHV space to the same vector, these proportions likely overestimate the variance between the trees explained by each principal component.

In Figure 2.3, we clearly observe three genes that have outlying phylogenies relative to the full gene set. Two of these outliers are with respect to the first principal component and correspond to genes DMRL_synthase and GTP_cyclohydroI. The third outlier is with respect to the second principal component and corresponds to the gene BacA. The rest of the gene trees are clustered together in the bottom left of the plot, with $\bar{T}_p^{\text{full}}$ shown in red in the middle of the cluster of gene trees. The placement of $\bar{T}_p^{\text{full}}$ in the middle of the cluster of non-outlying gene trees suggests that there are not major differences in branch lengths nor topology between the concatenated tree and the non-outlying gene trees. To investigate the sensitivity of our estimated phylogenomic tree to the outlying genes, we re-estimated it after removing the three declared outlying genes from the concatenated alignment. We call this tree $\bar{T}_p^{\text{reduced}}$, and plot its log map in Figure 2.3 in green. These two concatenated trees

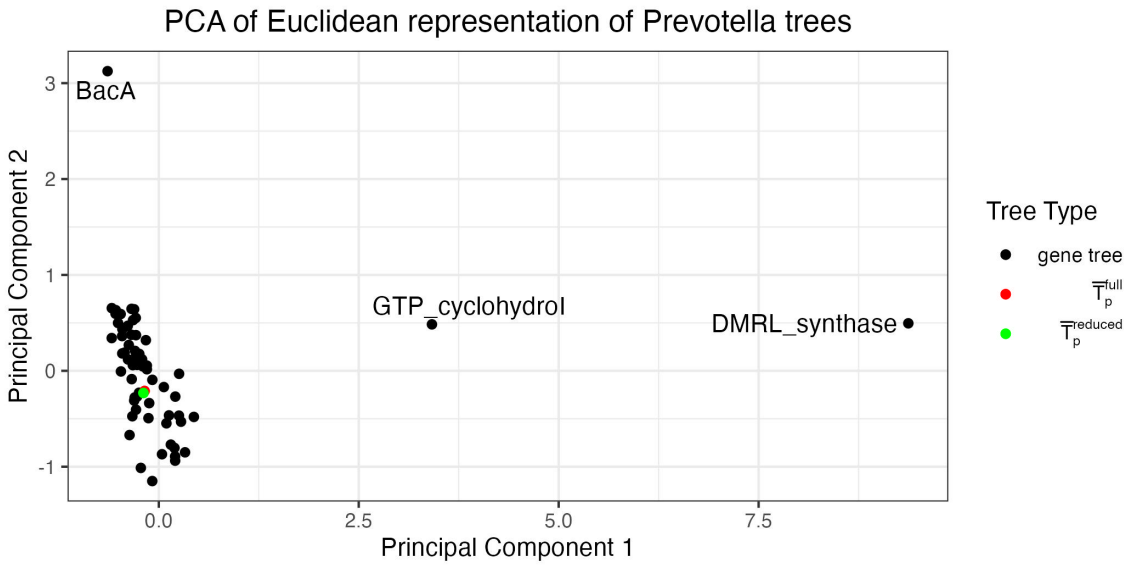
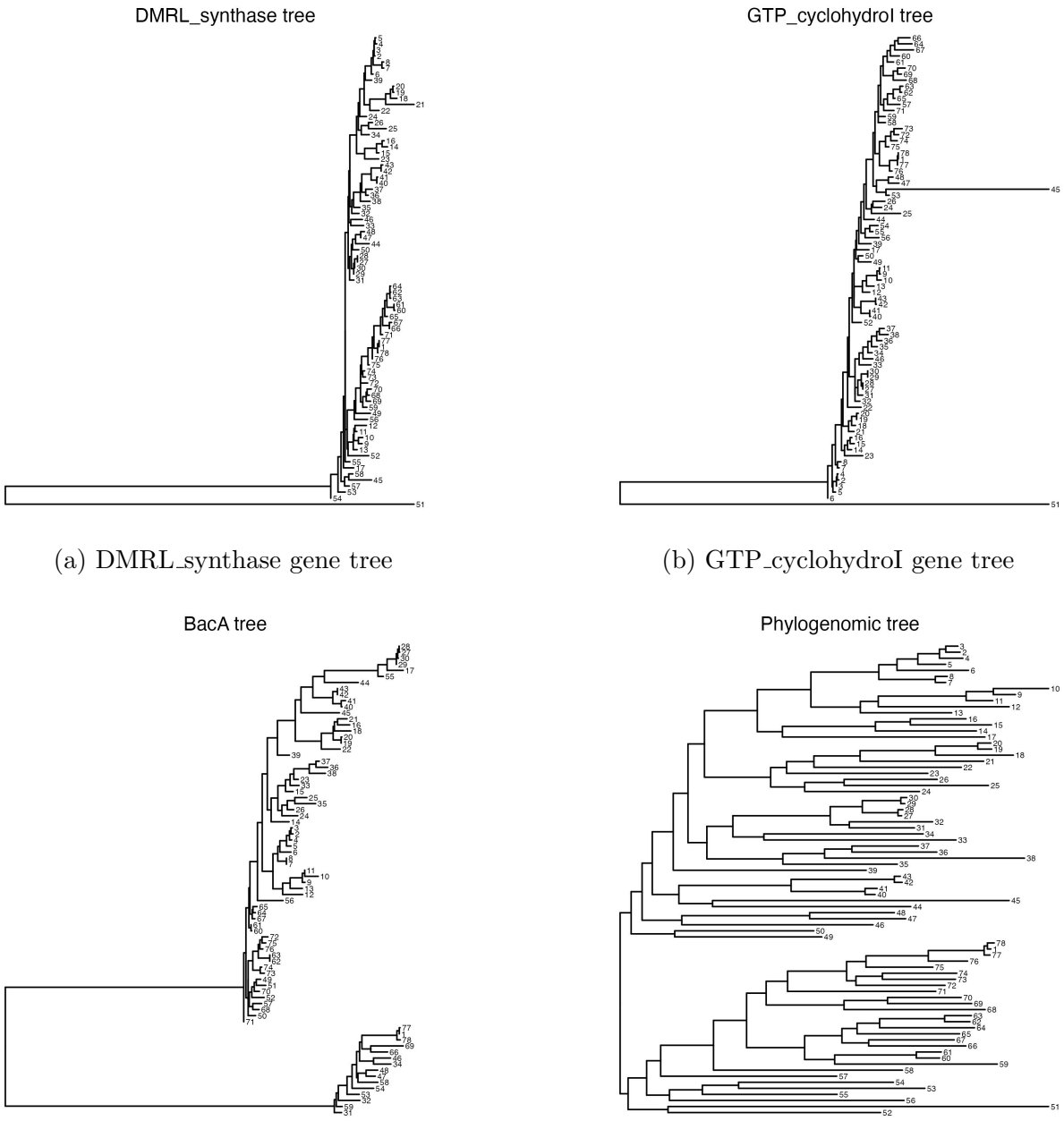


Figure 2.3: The proposed visualization of 63 gene trees constructed from 78 genomes from the Prevotella genus, depicted by a two-dimensional scatterplot. Three visibly outlying gene trees are labeled. A phylogenomic tree constructed from the full gene set is shown in red and a phylogenomic tree constructed from a reduced gene set after removing the three outlying genes is shown in green.

are in almost the same location in the log map visualization, with a RF distance of 14 and a BHV distance of 0.15 (which are the smallest RF and BHV distances between any pairs of trees in this analysis), suggesting that the initial phylogenomic tree is robust (on the scale of topological variation present in the dataset) to the inclusion of the outlying DMRL_synthase, GTP_cyclohydroI and BacA genes.

We now investigate the outlying gene trees further. The three gene tree outliers along with $\bar{T}_p^{\text{full}}$ are shown in Figures 2.4a–2.4d. All trees are visualized using ggtree [Yu et al., 2017]. For ease of viewing, we replace tip labels with numeric labels (see Supplementary File `Prevotella-data.csv` for a key to match labels with accession numbers). We see from Figures 2.4a and 2.4b that both of these estimated gene trees have a single long external branch leading to tip 51, which corresponds to a genome classified as Prevotellaceae bacterium UBA4332 (GCA_900316295.1). This branch length has a first principal component loading of 1.00, relative to a total sum of loading magnitudes of 1.51 across all log map features, giving it a relative weight of 0.66. Our visualization tool provides us an easy way to identify that this long branch is common to these two gene trees and substantially large in comparison with the branches in other gene trees and the phylogenomic trees. This genome (GCA_900316295.1) could be investigated further to understand why DMRL_synthase and GTO_cyclohydroI seem to have distinct evolutionary histories as compared to the rest of the incorporated genes. Visualizations of the alignments for these two genes can be seen in the Supplementary Material (alignment visualizations are generated with NCBI's Multiple Sequence Alignment Viewer [Sayers et al., 2019]).

Similarly, the second principal component in our plot differentiates the gene BacA from the rest of the estimated gene and phylogenomic trees. In Figure 2.4c we observe one long internal branch that separates this gene in 15 genomes from the gene in the remaining 63, while in the phylogenomic tree these 15 are spread across the two major clades (Figure 2.4d). Furthermore, gene-level differences in each of these genomes can be clearly observed in the amino acid alignments, which can also be seen in the Supplementary Material. The supplementary file `Prevotella-data.csv` holds additional information for the incorporated



(a) DMRL_synthase gene tree

(b) GTP_cyclohydroI gene tree

(c) BacA gene tree

(d) Phylogenomic tree

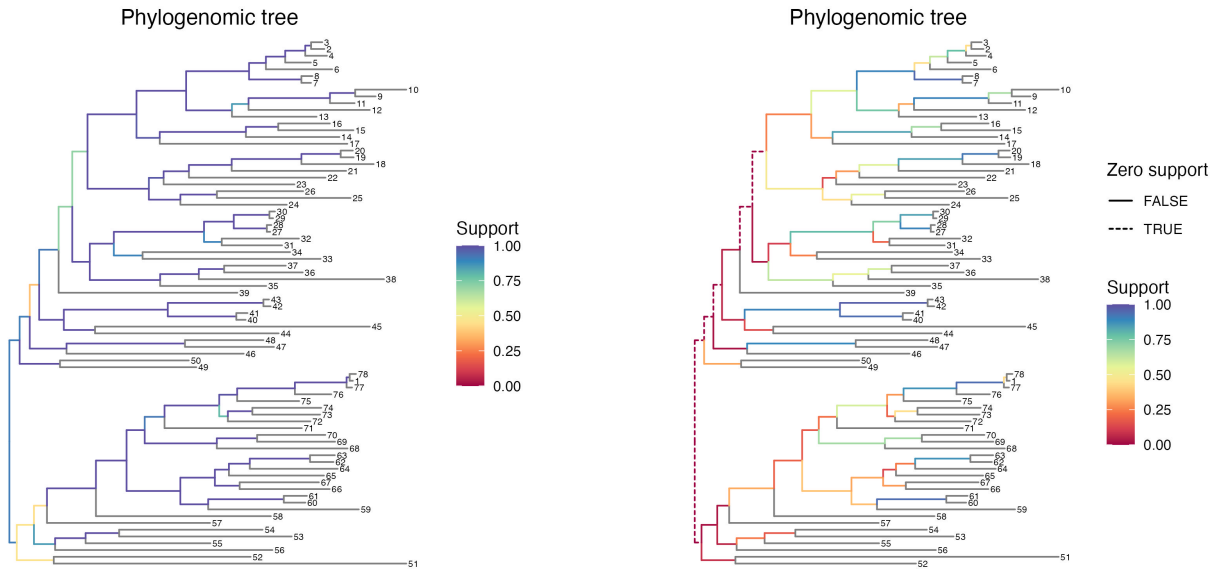
Figure 2.4: The estimated gene trees for the three outlying *Prevotella* genes identified in Figure 2.3, as well as the estimated phylogenomic tree (bottom right). All trees are rooted at their mid-point. The DMRL_synthase (top left) and GTP_cyclohydroI (top right) gene trees both include one long branch leading to tip 51. The BacA (bottom left) gene tree includes one long branch separating two clades of tips. No strikingly long branches are present in the phylogenomic tree.

genomes (e.g., source of the genome), but no clear trends emerge that might contribute to this gene-level phylogeny. It is possible this particular gene may be commonly transferred horizontally. Ultimately, the easy identification of these genes by our approach motivates further exploration into their sequence divergence for the two sets of genomes that are depicted by the large clades in Figure 2.4c.

Finally, we investigate the concordance between $\bar{T}_p^{\text{full}}$ and the gene trees. While our low-dimensional visualizations of the collection of gene trees may suggest that there is good clustering of gene trees around the phylogenomic tree, there is actually very low agreement between many of the splits in $\bar{T}_p^{\text{full}}$ and the splits present in the gene trees. We contrast the bootstrap support for $\bar{T}_p^{\text{full}}$ (computed using Hoang et al. [2018]) against the percentage of gene trees containing each split in Figures 2.5a and 2.5b. We observe the "tree of tips" phenomenon described by Thiergart et al. [2014], in which splits near the tips on the tree have higher gene tree support and splits near the center of the tree have lower support. However, as noted by Avni and Snir [2020], due to high rates of HGT, a difference between individual gene trees and phylogenomic trees is to be expected. This phenomenon indicates that while the estimated gene trees agree with the phylogenomic tree for the edges that split a small number of genomes, there is much more discordance among gene trees for splits that separate larger clusters of genomes from each other. The low gene tree support for many branches on the tree contrasts starkly with the very high bootstrap support for many of these splits. As a result, we advise caution when interpreting high bootstrap support values on phylogenomic trees, as they do not necessarily imply that the individual gene phylogenies support these splits.

2.3.2 Streptococcus

The above analysis provided an example of how to use our proposed method for generating hypotheses about differences in evolutionary histories between genes. We now illustrate how to assess the robustness of the estimated phylogenomic tree to different gene sets. In this analysis, we investigate the differences in phylogenomic trees built from a large set of genes



(a) Bootstrap support

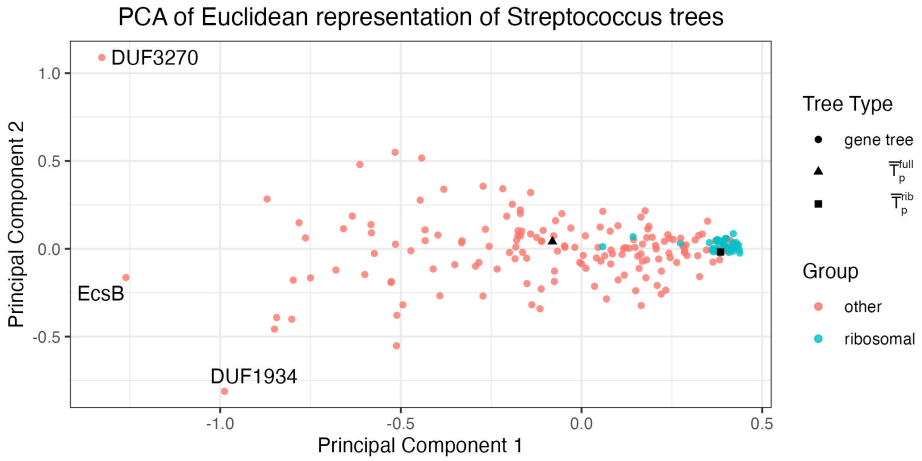
(b) Gene tree support

Figure 2.5: The phylogenomic tree with splits colored by bootstrap support (left) and gene tree support (right). Bootstrap support is high for all splits towards the tips and for most splits farther from the tips. Gene tree support is high for splits near the tips and lower throughout the rest of the tree. We observe zero gene tree support for several splits farthest from the tips.

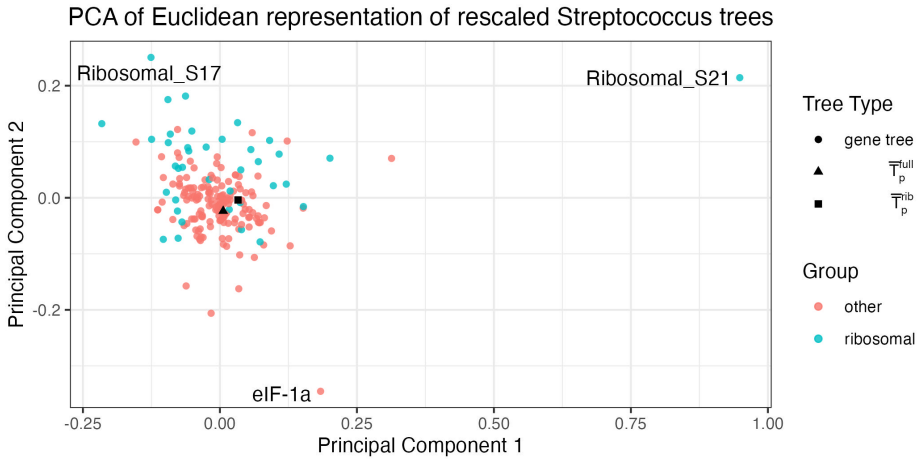
compared to a smaller set of functionally specific (ribosomal) genes. We consider the genus Streptococcus, which is of particular relevance in public health and medicine because of its potential pathogenicity.

Similar to Section 2.3.1, we consider all 301 GTDB representative species genomes for Streptococcus; identify 196 protein families (genes) that appear in a single copy in more than 90% of the representative genomes; and analyze the 106 genomes that contain all 196 genes of interest. This large set of genes and genomes provides a good test of our tool because trees with 106 tips are difficult to compare visually, especially a set of 196 trees of this size. Out of the 196 genes that we identified, 38 are ribosomal genes. In general, ribosomal genes are present in bacterial genomes in single copy and are considered essential "core" genes. As a result, they are commonly chosen as an appropriate gene set for constructing a concatenated alignment. We are interested in comparing the phylogenomic tree estimated using the 38 ribosomal genes with the phylogenomic tree estimated with the full set of 196 genes, which represent a wider range of genomic functions. We use GToTree and IQ-Tree to estimate $\hat{T}_1, \ldots, \hat{T}_{196}$, $\bar{T}_p^{\text{full}}$, and $\bar{T}_p^{\text{rib}}$, where $\bar{T}_p^{\text{full}}$ is built from the full set of 196 genes and $\bar{T}_p^{\text{rib}}$ from the set of 38 ribosomal genes.

A static output of our tool is shown in Figure 2.6a. The first principal component explains 44% of the variance in the log map vectors for the set of trees and the second principal component explains 10%. We observe a tight cluster of ribosomal gene trees and a more distributed cloud of non-ribosomal trees, suggesting that the estimated ribosomal gene trees are generally more similar to each other than to the other estimated gene trees. Due to relatively high purifying selective pressures, ribosomal proteins tend to evolve more slowly than the majority of genes. Therefore we generally expect their phylogenies to have shorter branch lengths than genes that are less functionally constrained. However, this pattern is not the case for every ribosomal gene, and we observe two ribosomal genes (Ribosomal_S30AE and Ribosomal_L9_C) that are further from the ribosomal gene tree cluster with respect to the first principal component. While these appear to be outlying amongst the ribosomal genes, they are not unusual when compared to the entire collection of gene trees. We also



(a) Visualization of estimated trees



(b) Visualization of estimated and rescaled trees

Figure 2.6: The proposed visualization of 196 gene trees estimated from 106 genomes in the Streptococcus genus shown as a two-dimensional scatterplot (upper). The same visualization is shown after rescaling the trees (lower), where the rescaling is performed by dividing all branch lengths on each tree by the sum of the branch lengths for that tree. A phylogenomic tree constructed from the full gene set is shown as a black triangle and a phylogenomic tree constructed from a subset of ribosomal genes is shown as a black square. Gene trees are colored by whether or not they are ribosomal genes.

observe two non-ribosomal gene trees, DUF1934 and DUF3270, which appear as potential outliers amongst all genes.

We can also see a noticeable difference between $\bar{T}_p^{\text{full}}$ and $\bar{T}_p^{\text{rib}}$ in Figure 2.6a. In order to compare $\bar{T}_p^{\text{full}}$ and $\bar{T}_p^{\text{rib}}$ we can compare their distances to distances among other pairs of trees. The BHV distance between the two phylogenomic trees is 0.43, which is smaller than 75% of the BHV distances between gene trees and phylogenomic trees in this analysis. However, the RF distance (based solely on topology) between the two phylogenomic trees is 140, which is smaller than 99.9% of distances between pairs of trees in the analysis. Finally, the difference in the sum of squared branch lengths between the two phylogenomic trees is 2.93, which is smaller than 43% of the distances between trees in the analysis. In general, the estimated ribosomal tree had shorter branches compared to the full gene set tree (ribosomal vs. full tree branch length mean: 0.003 vs. 0.017; median: 0.0008 vs. 0.0090). Taken together, this suggests that these two phylogenomic trees are very similar in their topologies, but differ in their branch lengths. We conclude that the separation implied by the log map visualization is most likely driven by differences in branch lengths, not topology, between these two trees.

Motivated by our observation that branch lengths likely drive the difference between the ribosomal and full phylogenomic trees, we investigate the impact of rescaling *all* trees' branch lengths on the visualization. We rescale each tree by dividing each individual branch length by the sum of the branch lengths on the given tree. This modification means that our visualization highlights differences between trees with respect to topology and relative branch lengths, but not absolute branch lengths. This is also an option within the interactive visualization. We show the resulting visualization in Figure 2.6b. We see that the ribosomal gene trees are much more evenly spread among the other gene trees than in Figure 2.6a, and that the two phylogenomic trees are visibly closer. This provides further evidence that much of the differences between the ribosomal and full phylogenomic trees that we observed in Figure 2.6a can be attributed to differences in branch lengths. This is unsurprising because as mentioned above, ribosomal genes are typically under greater purifying selective pressures. In the rescaled visualization, there remain a few noticeable outliers that could be

investigated further: Ribosomal_S21, Ribosomal_S17, and eIF-1a. These trees all have one or two branches that are especially large relative to other branches in the tree. Notably, we did not detect these differences in the original visualization because none of these long branches were long in comparison to branches on other gene trees prior to rescaling. All trees and rescaled trees are available as Supplementary Data.

2.4 Discussion

In this chapter, we introduced an easy-to-use, interactive method for visualizing a set of trees. Our method was motivated by expanding interest in interrogating microbial evolution, specifically in the context of comparing individual gene evolutionary histories, and the relatively few tools available to facilitate such analyses. Our approach is to represent each phylogeny as a vector in $\mathbb{R}^{2m+3}$ by taking a local linear approximation to tree space around a central tree, then performing dimension reduction (such as PCA) to view each tree in low-dimensional Euclidean space. The approach of performing PCA on tree approximations is notably different from the approach of Nye [2011], in which a version of PCA is performed in tree space to identify sources of variation in a set of trees. This is also distinct from most other tools for visualizing a set of phylogenies, which compute dissimilarities between each tree and perform MDS on the dissimilarity matrix [Daubin et al., 2002, Amenta and Klingner, 2002, Hillis et al., 2005, Holmes, 2006, Chakerian and Holmes, 2012, Kendall and Colijn, 2016, Jombart et al., 2017, Zhu et al., 2019]. Our approach also contrasts with the related method of Willis and Bell [2018] by addressing its main limitation of attempting to locally project tree space around the Fréchet mean tree, which commonly lies at the boundary of two orthants and results in a poor local approximation of the metric space. While our method could be applied to the analysis of any collection of genes, it is especially well-suited to microbial datasets because it integrates with common methods for estimating phylogenomic trees via a concatenated alignment, provides tools for investigating the robustness of the phylogenomic tree estimates to the incorporated genes, and is computationally feasible for analyzing all genes that are common to a set of microbial genomes.

Despite substantial advantages in ease-of-use and intuitive interpretations of the visualization, our approach inherits the limitations of BHV tree space. In particular, BHV geodesics and log maps are only defined for trees that share identical leaf sets. As a result, our tool is restricted to analyzing sets of genes and genomes for which all genes are present in all genomes. In practice, we address this by limiting our analysis to genes present in 90% of genomes, then restricting to genomes that contain all of these genes. However, users could use their own approach if they have specific genes or genomes that they would like to include in their analysis. One implication of this is that more genes can be studied when the incorporated genomes are more closely related evolutionarily. In contrast, datasets with taxa that span phyla or domains will have fewer genes common to many genomes, and may result in a small number of genes for analysis. This limitation is shared by many cross-genome comparative (often called *pangenomic*) analyses. While the absence of a gene in a genome may reflect true biological variation, genes may also be absent due to limitations of genome sequencing and reconstruction. Thus, there are many reasons why a microbial gene may be excluded from a given analysis, and the limitations of BHV space in comparing genomes with different gene sets is one possible reason. Ongoing work to reconcile BHV space across trees with different leaf sets [Grindstaff and Owen, 2020, Ren et al., 2017] may ultimately enable local linear approximations of BHV spaces of different leaf sets, and thus more comprehensive comparisons of gene sets, in the future.

The goal of this work is to demonstrate that investigating gene trees as well as summary phylogenomic trees can uncover interesting evolutionary signal and to offer a method to investigate and analyse sets of trees. For an unsupervised exploration and visualization of trees we prefer to use PCA on log map vectors instead of MDS on distances between trees for reasons outlined in Section 2.2.3. However, we think that there are situations in which MDS, or other methods such as UMAP or tSNE may work better for a set of gene trees and uncover interesting features of the data that may not be seen in PCA of the log map vectors. Additionally, if the goal of an analysis is not to construct a low-dimensional visualization and instead to classify gene trees or perform another supervised task, other multivariate analysis

methods tailored to that task could be used with the log map vectors as input. Our software includes a function that outputs the log map vectors for trees used in an analysis, which could then be used as input to software that performs other types of analyses outside of our proposed methodology and MDS of BHV distances between trees.

We provide an easy-to-use software implementation of our proposed methodology. This tool enables a visual exploration of gene-level evolutionary differences across sets of trees, which facilitates the identification of genes with unique (or anomalous) evolutionary histories. Our software also allows for the straightforward refinement of genes being used in estimating a phylogenomic tree, including sensitivity analyses for adding or removing genes. As described in Section 2.2.2, it is also easily extensible to alternative visualization approaches (e.g., MDS instead of PCA) and tree metrics (e.g., RF distance rather than BHV distance). Our tool is available as both a workflow or shiny app in our open-source R package at `github.com/statdivlab/groves`. Code to reproduce the data analyses in Section 2.3 is available at `github.com/statdivlab/groves_supplementary`.

Chapter 3

SCALABLE DIFFERENTIAL ABUNDANCE ANALYSIS FOR MICROBIAL METAGENOMICS

3.1 Introduction

The complex ways in which a microbiome can affect and be affected by changes in its host organism or the biome in which it exists are in large part driven by its community composition. For example, a healthy human gut is typically considered to be one that has high taxonomic diversity, which means that it includes many microbial species, and is dominated by three major microbial phyla [Hou et al., 2022]. Changes to this gut microbiome composition have been linked to many gastrointestinal diseases [Bull and Plummer, 2014]. In marine environments, microorganisms perform important functions in the ecosystem, including fixing carbon and nitrogen. However, increasing warming and acidification of oceans impact marine microbiome composition, which has effects on these microorganisms' ability to perform these functions [Cavicchioli et al., 2019]. One way to study differences in microbial community composition across changes in an environment is through the lens of differential abundance. Understanding which microbial taxa (groups used to classify microbes) have increased or decreased abundance across covariate levels can provide important scientific insights.

Although differential abundance analyses are one of the most common types of analyses performed on microbiome data, there is little consensus about how to define and evaluate differential abundance [Nearing et al., 2022]. This is because of the challenges that metagenomic sequencing data present. An ideal differential abundance estimand would link abundances of taxa in microbiome samples, in terms of number of cells of that taxon per unit volume, to sample covariates. However, we cannot directly observe these abundances on the absolute

scale of cells per unit volume from data generated from metagenomic sequencing. Instead, the counts that we observe in this sequencing data are distortions of the true abundances, impacted by sample-specific and taxon-specific unknown sequencing effects. This makes their magnitudes uninformative for the absolute abundances. However, this data still contains valuable information about relative changes in abundance. Therefore, the major aim of a differential abundance method should be to define a useful differential abundance estimand that can be estimated and tested using metagenomic sequencing data.

Many differential abundance methods exist to account for these sample-specific and taxon-specific sequencing effects, as well as the natural sparsity of microbiome data, including ALDEx2 [Fernandes et al., 2013], ANCOM [Mandal et al., 2015] and its sucessors ANCOM-BC Lin and Peddada [2020] and ANCOM-BC2 [Lin and Peddada, 2024], and DESeq2 [Love et al., 2014]. We believe that an ideal differential abundance method should target differential abundance estimands that are unaffected by unknown sequencing effects and should provide a clear interpretation of these estimands and their limitations. As a statistical method, we would also like it to handle sparisty without pseudocounts (small values added to observed zero counts), avoid strong assumptions about the distribution of the data, and control the Type I error rate, even in small samples. While many differential abundance methods achieve some of these goals, one method that achieves all of them is presented by Clausen and Willis [2024].

Clausen and Willis introduce a method to estimate and test an estimand that represents the expected log fold-difference in abundance of a given microbial taxon across covariate levels, subject to an identifiability constraint defined over expected log-fold differences for all taxa in an analysis. They let the user of their method specify this constraint function, and therefore the target estimand, but they recommend using a smoothed median constraint in order to compare each expected log fold-difference in abundance to the "typical" expected log fold-difference in abundance across taxa. They show that, under identifiability assumptions motivated by the limitations of metagenomic sequencing data, while fold-differences in abundance are *not* identifiable, they can estimate *differences* in abundance across covariate

levels, relative to the smoothed median fold-difference across taxa.

Once their target estimand is defined, Clausen and Willis employ a flexible and assumption-light approach to estimation and testing. They propose a joint mean model for all samples and taxa, sharing information about sample-specific sequencing effects across taxa, and develop an algorithm to estimate parameters in this model. Only a mean model is assumed, with testing that is robust to differences between the parametric model used for estimation and the true unknown distribution of the data. Clausen and Willis show through simulations that the robust score test procedure that they propose controls the Type I error rate, even for small sample sizes.

Unfortunately, this method is extremely computationally expensive for analyses of a large number of taxa. This limits the situations in which this method can be applied, because many metagenomic sequencing datasets involve thousands of taxa. The primary computational burden of this method comes from the robust score tests. A score test requires the estimation of parameters in the model under the null hypothesis. Therefore, performing J different score tests, for J taxa in a metagenomic sequencing dataset, involves estimating parameters under the null hypothesis J separate times. Because Clausen and Willis consider a joint model over $p \times J$ parameters, where p is the number of columns in the design matrix, an increase in the number of taxa included in an analysis will increase the number of parameters that need to be estimated, leading to an increased computational burden for each score test. In this chapter, we develop a method to perform robust score tests that tests the same null hypotheses as Clausen and Willis' robust score tests, while requring a fraction of the computational time and resources.

The key to the improved computational efficiency of our robust score tests is that the process of estimation under the null hypothesis that is required for each test only involves estimating $p \times J'$ parameters for $J' << J$. We prove that for each null hypothesis, we can define a reduced model with only $p \times J'$ parameters, with which we can test the null hypothesis defined using the full model over all samples and taxa. We demonstrate through simulation that for appropriately chosen reduced mdoels, we can perform highly efficient

robust score tests that retain the Type I error rate control and power of Clausen and Willis' robust score tests. In comparisons of the results of these two robust score test procedures in data analyses, we show that they lead to similar differential abundance conclusions. The major challenge in our method is that these computational advantages are only possible for estimands that are defined using an identifiability constraint on log fold-differences for a small set of reference taxa. We propose several ways in which to choose this set of reference taxa, and investigate how the choice of reference set and consequent target estimand affects the estimation and inference results of our method.

This chapter is organized as follows. We define the reduced models in Section 3.2, and prove that they can be used to test the same hypotheses as those tested with Clausen and Willis' full model. In Section 3.3, we present simulation results to compare the Type I error rate control and power of our efficient robust score tests to those of Clausen and Willis' robust score tests. We suggest ways in which to choose a reference set and compare the results of this choice via simulation and in a small data analysis in Section 3.4. In Section 3.5, we apply our method to a large differential abundance analysis of a metagenomic sequencing dataset, and compare our results to those from other differential abundance methods. We conclude with a discussion of challenges and future work in Section 3.6.

3.2 Methods

3.2.1 Model

Consider absolute concentations $W_i \in \mathbb{R}^J_{\geq 0}$ and covariate vectors $X_i \in \mathbb{R}^p$ for samples $i \in \{1, \ldots, n\}$. We assume the following mean regression model,

$$\log \mathbb{E}[W_i | X_i, \beta^*] = X_i \beta^*, \tag{3.1}$$

where $\beta^* \in \mathbb{R}^{p \times J}$. The parameter ${\beta^*}^j_k$ for $k > 0$ represents the expected log fold-difference in cell count per unit volume of taxon j between a sample with covariates $X = x$ and a sample with covariates $X_{-k} = x_{-k}, X_k = x_k + 1$.

In practice, we do not observe W from WGS data. The counts that we observe from metagenomic sequencing data, $Y_i \in \mathbb{R}^J_{\geq 0}$, are disorted measurements of W_i, affected by sample-specific and taxon-specific unknown sequencing effects. We assume that these sequencing effects have multiplicative effects on expected counts Y_{ij}, such that,

$$\log \mathbb{E}[Y_i | X_i, \beta^*, z_i, \delta] = z_i \mathbf{1}_J^T + X_i \beta^* + \delta^T, \tag{3.2}$$

in which e^{z_i} represents the sample effect and e^{δ_j} represents each taxon effect for $j \in \{1, \ldots, J\}$.

Clausen and Willis [2024] provide identifiability results for this mean model. Assuming that the design matrix includes a column of ones, they note that the taxon effect δ_j and intercept β_0^{*j} only appear in the model through the sum $\delta_j + \beta_0^{*j}$ and therefore neither parameter can be identified. The intercepts are typically not of scientific interest, so we define $\beta \in \mathbb{R}^{p \times J}$ such that $\beta_0 = \beta_0^* + \delta^T$ and $\beta = [\beta_0^T, \beta_1^{*T}, \ldots, \beta_{p-1}^{*T}]^T$. This substitution results in the updated mean model,

$$\log \mathbb{E}[Y_i | X_i, \beta, z_i] = z_i \mathbf{1}_J^T + X_i \beta \cdot \tag{3.3}$$

More consequentially, Clausen and Willis show that this mean model is only partially identifiable up to equivalence classes of β parameters. In order to make this mean model identifiable, they propose including an identifiability constraint of the form $g(\beta_k) = 0$ on each row of β, such that each equivalence class of β only includes one element that satisfies the constraint. They require this function $g(\cdot)$ to be smooth and satisfy $g(\beta_k + a) = g(\beta_k) + a$ for $a \in \mathbb{R}$, in order to avoid additional inferential and computational complications.

This constraint function $g(\cdot)$ affects parameter interpretation, as the identifiable parameter in this model has the form $\beta_k^j - g(\beta_k)$. In order to compare log fold-differences for each taxon to typical log fold-differences in the dataset, Clausen and Willis propose using the pseduo-Huber loss $g_p(\cdot)$ over all categories as the constraint function, which is a smooth function that approximates the median. This function is defined in Appendix B.1. Using the constraint function $g_p(\cdot)$, $\beta_k^j - g_p(\beta_k)$ can be interpreted as the expected log fold-difference in cell count per unit volume of taxon j between a sample with covariates $X = x$ and a sample

with covariates $X_{-k} = x_{-k}, X_k = x_k + 1$, relative to the typical expected log fold-difference in abundance across all taxa.

3.2.2 Estimation and inference

Clausen and Willis [2024] develop procedures to estimate parameters $\beta_k - g(\beta_k)$ and perform statistical inference for null hypotheses of the form $H_0^j : \beta_k^j - g(\beta_k) = 0$. They propose using the Poisson log likelihood with the mean model given by (3.3) subject to the identifiability constraint $g(\beta_k) = 0 \; \forall k$ to define estimating equations. They show that after profiling the nuisance parameters z out of the Poisson likelihood, the resulting likelihood is equivalent to a multinomial log likelihood with a logistic link. In order to guarantee finite parameter estimates, they add a Firth penalty [Firth, 1993] to the multinomial log likelihood-based estimating equations. They apply results from Kosmidis and Firth [2011] to develop an efficient algorithm to obtain penalized maximum likelihood estimates of $\beta_k - g(\beta_k)$.

Clausen and Willis provide two methods to test the hypothesis $H_0^j : \beta_k^j - g(\beta_k) = 0$: a robust score test and a robust Wald test. For $h(\beta) = \beta_{kj} - g(\beta_k)$, the robust score test uses the test statistic given in [White, 1982], using the adjustment suggested by Guo et al. [2005]:

$$T_{RS} = \frac{n}{n-1} \cdot S_{H_0}^T I_{H_0}^{-1} F_{H_0}^T (F_{H_0} I_{H_0}^{-1} D_{H_0} I_{H_0}^{-1} F_{H_0}^T)^{-1} F_{H_0} I_{H_0}^{-1} S_{H_0}, \tag{3.4}$$

in which S_{H_0} is the score evaluated at the MLEs under the null, I_{H_0} is a consistent estimate of the information matrix under the null, D_{H_0} is a consistent estimate of the covariance matrix of the score equations under the null, and F_{H_0} is $\frac{\partial h}{\partial \beta^T}$ evaluated at the MLEs estimated under the null. The robust Wald test statistic has the form

$$T_{RW} = h(\beta)^T (F_{H_A}^T I_{H_A}^{-1} D_{H_A} I_{H_A}^{-1} F_{H_A})^{-1} h(\beta) \tag{3.5}$$

where F_{H_A}, I_{H_A}, and D_{H_A} are defined as in the score test, but evaluated using MLEs estimated under the alternative hypothesis.

These tests are considered robust because unlike the model-based score test and Wald tests, they do not assume that the variance of the score vector is approximated by the information matrix defined by the Poisson likelihood used to motivate the estimating equations.

Instead, these test statistics involves estimates of the approximate asymptotic covariance matrix of the score vector (under the alternative hypothesis for the Wald test and under the null hypothesis for the score test) [Boos, 1992]. In this way, Clausen and Willis use the Poisson likelihood to define an estimator, but do not need to assume that the Y_{ij} values are Poisson-distributed random variables for valid inference. The only assumptions required for valid inference are that the mean model (3.3) holds and that the Y_i vectors are independent of each other. For known correlation in the Y_i's, the test statistics can be generalized to retain error rate control.

Regardless of the choice of test, estimation under the alternative hypothesis needs to be performed in order to generate parameter estimates to be interpreted in terms of covariates in the model. This means that the robust score tests require an additional step compared to the robust Wald tests, because they also require parameter estimates under the null hypothesis. Although the process of estimating parameters under the null hypothesis is simpler than the process of estimating parameters under the alternate hypothesis in many model settings, that is not the case for this model. Estimation under the null is complicated in this setting by the interplay between the parameter constraints imposed by the identifiability functions $g(\beta_k) = 0 \; \forall k \in \{0, \ldots, p-1\}$ and the parameter constraints imposed by the null hypothesis $\beta_k^j - g(\beta_k) = 0$. Clausen and Willis develop an augmented Lagrangian algorithm to perform estimation under the null hypothesis. This algorithm is more complex than the algorithm for estimation under the alternative hypothesis, and typically requires more computation time to converge. Therefore, in this setting a single robust score test has a higher computational burden than a single robust Wald test. Additionally, this algorithm for estimation under the null hypothesis needs to be run J different times for tests of all J taxa.

Despite the greater computational burden of the robust score tests compared to the robust Wald tests, Clausen and Willis recommend using the robust score tests in practice, especially for small or moderate sample sizes. This is because in simulation they show that the robust score test is conservative under the null hypothesis, especially with smaller sample sizes and non-Poisson distributed data. They find that in those settings the robust Wald

test is anticonservative and fails to control the Type I error rate. We confirm these findings with our own simulations in Section 3.3.

3.2.3 Scalable Inference

Although Clausen and Willis recommend using the robust score test instead of the robust Wald test in most settings, it is computationally intensive to run these tests on all taxa in a dataset that includes many thousands of taxa. This is because parameter estimation under the null hypothesis must be repeated for each of the J tests. Additionally, the time and memory required to estimate parameters for a single score test typically increase with J, because it requires more resources to estimate a model with a larger number of parameters. Because estimation for a model with a large number of parameters requires a large amount of memory, for a large enough J it is not possible to run many robust score tests in parallel without access to high performance computing resources.

In order to utilize the advantages of the robust score tests without needing the same scale of computational resources, we introduce a procedure to run robust score tests that involve the estimation of fewer parameters under the null hypothesis for each test. We achieve this by defined a reduced model for each test that has pJ' parameters for $J' << J$. We prove that we can define the reduced model for the test of the hypothesis $\beta_k^j - g(\beta_k) = 0$ in such a way that a parameter equivalent to $\beta_k^j - g(\beta_k)$ can be identified in and therefore tested with this reduced model.

3.2.4 Construction of reduced joint models

The idea of decreasing computational burden by defining reduced models with parameters that are equivalent to parameters of interest from larger models is used in the work of Begg and Gray [1984]. Begg and Gray show that the regression parameters from a multinomial logistic regression model are equivalent to the regression parameters from individual logistic regression models that compare each category to a baseline category. They argue that maximum likelihood estimates (MLEs) of the parameters from the individual logistic regression

models will be asymptotically unbiased for the parameters in the multinomial logistic regression model, and show through simulation that their logistic regression MLEs loses little efficiency compared to the MLEs from the multinomial logistic regression model. Hu et al. [2022] use Begg and Gray's logistic regression approach in LOCOM, another microbiome differential abundance method, in order to avoid fitting a model on $p \times J$ parameters. We apply a similar approach in this section.

Throughout this section, we will state and prove results using the identifiability constraint $g(\beta_k) = \beta_k^1 = 0 \ \forall k$. The following section generalizes these results to a more general class of constraint functions.

Consider the mean model given in (3.3), written slightly differently,

$$\log \mathbb{E}[Y_{ij}|X_i, \beta, z_i] = z_i + X_i(\beta^j - \beta^1), \ i \in \{1, \ldots, n\}, \ j \in \{1, \ldots, J\}. \tag{3.6}$$

We will refer to this model as M_F, for full model.

We will assume that when we consider only a subset of taxa $S_R \subset \{1, \ldots, J\}$ with $1 \in S_R$ and $|S_R| = J_R$, all of the assumptions that we made in order to construct mean model (3.3) are still true. Therefore, we define a reduced model M_R as the following,

$$\log \mathbb{E}[Y_{ij}|X_i, \alpha, v_i] = v_i + X_i(\alpha^j - \alpha^1), \ i \in \{1, \ldots, n\}, \ j \in S_R. \tag{3.7}$$

Proposition 1. *$\alpha_k^j - \alpha_k^1 = \beta_k^j - \beta_k^1$ for all $k \in \{0, \ldots, p-1\}$ and for all $j \in S_R$.*

Proof: We will first consider the parameters z_i and v_i for all $i \in \{1, \ldots, n\}$. Because we are using the first category constraints $\beta_k^1 = 0 \ \forall k$ for M_F and $\alpha_k^1 = 0 \ \forall k$ for M_R, this means that in their respective models, z_i and v_i both represent intercept terms for taxon 1 and sample i. This implies that $z_i = \log \mathbb{E}[Y_{i1}]$ and $v_i = \log \mathbb{E}[Y_{i1}]$ for all i, and therefore $z_i = v_i$ for all i. Next, we will consider the interpretation of the $\beta_k^j - \beta_k^1$ and $\alpha_k^j - \alpha_k^1$ parameters in their respective models.

$$\beta_k^j - \beta_k^1 = \log \frac{\mathbb{E}[Y_{ij}|z_i = z, X_{im} = x_{im} \forall m \neq k, X_{ik} = x+1]}{\mathbb{E}[Y_{ij}|z_i = z, X_{im} = x_{im} \forall m \neq k, X_{ik} = x]} \; \forall k, \forall j \in \{1, \ldots, J\}$$
$$\alpha_k^j - \alpha_k^1 = \log \frac{\mathbb{E}[Y_{ij}|v_i = v, X_{im} = x_{im} \forall m \neq k, X_{ik} = x+1]}{\mathbb{E}[Y_{ij}|v_i = v, X_{im} = x_{im} \forall m \neq k, X_{ik} = x]} \; \forall k, \forall j \in S_R$$

Finally, we will note that because $z_i = v_i$ for all samples i, then we can replace the second line above with,

$$\alpha_k^j - \alpha_k^1 = \log \frac{\mathbb{E}[Y_{ij}|v_i = z, X_{im} = x_{im} \forall m \neq k, X_{ik} = x+1]}{\mathbb{E}[Y_{ij}|v_i = z, X_{im} = x_{im} \forall m \neq k, X_{ik} = x]} \; \forall k, \forall j \in S_R.$$

This implies that $\beta_k^j - \beta_k^1 = \alpha_k^j - \alpha_k^1$ for all $k \in \{0, \ldots, p-1\}$ and for all $j \in S_R$. □

Therefore, the null hypothesis $H_0^{j(F)} : \beta_k^j - \beta_k^1 = 0$ is equivalent the the null hypothesis $H_0^{j(R)} : \alpha_k^j - \alpha_k^1 = 0$ for $j \in S_R$. As these hypotheses are equivalent, we propose testing $H_0^{j(R)}$ instead of $H_0^{j(F)}$.

Let $\hat{\alpha}_{H_0^{j(R)}}$ be the maximum likelihood estimate of parameter α using model M_R under null hypothesis $H_0^{j(R)}$ and identifiability constraint $\hat{\alpha}^1_{H_0^{j(R)}} = 0$. Define the robust score statistic $T_{RS}^{H_0^{j(R)}}$ of the form:

$$T_{RS}^{H_0^{j(R)}} = S^T_{H_0^{j(R)}} I^{-1}_{H_0^{j(R)}} F^T_{H_0^{j(R)}} (F_{H_0^{j(R)}} I^{-1}_{H_0^{j(R)}} D_{H_0^{j(R)}} I^{-1}_{H_0^{j(R)}} F^T_{H_0^{j(R)}})^{-1} F_{H_0^{j(R)}} I^{-1}_{H_0^{j(R)}} S_{H_0^{j(R)}}, \qquad (3.8)$$

where $S_{H_0^{j(R)}}$ is the score evaluated at $\hat{\alpha}_{H_0^{j(R)}}$, $I^{-1}_{H_0^{j(R)}}$ is a consistent estimate of the information matrix under $H_0^{j(R)}$, $F_{H_0^{j(R)}}$ is $\frac{\partial h}{\partial \alpha^T}$ for $h(\alpha) = \alpha_k^j - \alpha_k^1$ evaluated at $\hat{\alpha}_{H_0^{j(R)}}$, and $D_{H_0^{j(R)}}$ is the sum of the outer products of the score for each sample evaluated at $\hat{\alpha}_{H_0^{j(R)}}$.

Proposition 2. *Under the null hypothesis* $H_0^{j(F)}$, $T_{RS}^{H_0^{j(R)}} \xrightarrow{d} \chi^2_1$ *for* $j \in S_R$.

Proof: We have shown that $H_0^{j(R)}$ and $H_0^{j(F)}$ are equivalent. Therefore, $H_0^{j(F)}$ is true if and only if $H_0^{j(R)}$. Assume that $H_0^{j(F)}$ is true. This means that $H_0^{j(R)}$ is also true. $T_{RS}^{H_0^{j(R)}}$ is the robust score statistic given by White [1982] evaluated at the MLE under the restriction of the null hypothesis $H_0^{j(R)}$ and we are assuming that $H_0^{j(R)}$ is true, so we can apply Theorem 3.5 [White, 1982], which states that $T_{RS}^{H_0^{j(R)}} \xrightarrow{d} \chi^2_1$. □

Critically, if $J_R = |S_R|$ is small relative to J, then using the reduced model M_R instead of the full model M_F will reduce the computation time needed to run a robust score test.

Derivation of reduced model using the multinomial distribution

In their estimation algorithms, Clausen and Willis use the fact that after profiling z parameters out of the Poisson likelihood function with mean model given by (3.3), the resulting function is a multinomial logistic regression likelihood. This fact does not require the parametric assumption that $Y_{ij} \big| \sum_{j'=1}^{J} Y_{ij'}$ follows a multinomial distribution. However, if we do assume that $Y_{ij} \sim$ Poisson and $Y_{ij} \big| \sum_{j'=1}^{J} Y_{ij'} \sim$ Multinomial, then we can show that the reduced model M_R is the mean model for $\mathbb{E}[Y_{ij} | X_i, \beta, \sum_{j^*=1}^{J} Y_{ij^*} = y, Y_{ij'} = y_{ij'} \ \forall j' \in S_C]$, where we define $S_C = \{1, \ldots, J\} \setminus S_R$. This connection is further described in Appendix B.2.

3.2.5 Identifiability under the reduced model

In the previous section, we demonstrated that we can define a reduced model M_R on a subset of taxa $S_R \subset \{1, \ldots, J\}$, such that for parameters α in the reduced model, $\alpha_k^j - \alpha_k^1 = \beta_k^j - \beta_k^1$ for all $j \in S_R$. We can see that this statement would not be true if we use the pseudo-Huber loss over β_k^j parameters for all $j \in \{1, \ldots, J\}$ as the constraint function, as recommended by Clausen and Willis. In that case we would be comparing different parameters in the two models, because even if $\beta_k^j - \beta_k^1 = \alpha_k^j - \alpha_k^1$ for all $j \in S_R$, the pseudo-Huber loss over parameters $\{\beta_k^j - \beta_k^1 : j \in \{1, \ldots, J\}\}$ and the pseudo-Huber loss over parameters $\{\alpha_k^j - \alpha_k^1 : j \in S_R\}$ will be different for most realizations of $\beta_k^j - \beta_k$ parameters.

To address this, we consider the class of pairs of constraint functions $g_F : \mathbb{R}^J \to \mathbb{R}$ and $g_R : \mathbb{R}^{J_R} \to \mathbb{R}$ that guarantee that $\beta_k^j - g_F(\beta_k) = \alpha_k^j - g_R(\alpha_k)$ for all $j \in S_R$. Let $v[S] \in \mathbb{R}^{|S|}$ denote the subset of elements of $v \in \mathbb{R}^J$ that correspond to the $s \in S$-th entries. Then, define the class $\mathcal{G}_{S_R} = \{\{g_F, g_R\} : g_F(v) = g_R(v[S_R]) \ \forall v \in \mathbb{R}^J\}$.

Proposition 3. *For $S_R \subset \{1, \ldots, J\}$ with $1 \in S_R$ and any pair of constraint functions $\{g_F, g_R\} \in \mathcal{G}_{S_R}$, define the full model M_F and the reduced model M_R as in (3.6) and (3.7). Then, $\alpha_k^j - g_R(\alpha_k) = \beta_k^j - g_F(\beta_k)$ for all k and all $j \in S_R$.*

Proof: In the previous section, we proved that under these models $\alpha_k^j - \alpha_k^1 = \beta_k^j - \beta_k^1$ for all $j \in S_R$. We will apply the definition of $\mathcal{G}_{S_R}$ to note that $g_F(v) = g_R(v[S_R])$ for all

$v \in \mathbb{R}^J$. We will also use the fact we've defined our set of constraint functions such that $g(v + a) = g(v) + a$ for $a \in \mathbb{R}$.

$$\begin{aligned}
\beta_k^j - g_F(\beta_k) &= \beta_k^j - \beta_k^1 - g_F(\beta_k - \beta_k^1) \\
&= \beta_k^j - \beta_k^1 - g_R((\beta_k - \beta_k^1)[S_R]) \\
&= \alpha_k^j - \alpha_k^1 - g_R(\alpha_k - \alpha_k^1) \\
&= \alpha_k^j - g_F(\alpha_k).\square
\end{aligned}$$

3.2.6 Approach

Our goal in defining reduced models is to develop a method of testing the null hypothesis $H_0^j : \beta_k^j - g_F(\beta_k) = 0$ for all $j \in \{1, \ldots, J\}$ with robust score tests in a way that requires fewer computational resources than Clausen and Willis' robust score tests. We will start by choosing a set of taxa S_{ref} such that $|S_{ref}| << J$, which we will refer to as a reference set. In Section 3.4, we investigate ways to choose a reference set and consequent parameter that is useful and interpretable in practice. We will then choose a constraint function $g_{\text{ref}} : \mathbb{R}^{|S_{ref}|} \to \mathbb{R}$ on this set of reference taxa. We suggest choosing g_{ref} to be the pseudo-Huber loss over β_k^j parameters for $j \in S_{ref}$. Then, we define the constraint function for the full model $g_F(\cdot)$ to be a function such that $g_F(v) = g_{\text{ref}}(v[S_{ref}])$. Going forward, we will use the notation $g_F^p(v^j : j \in S)$ to refer to the function $g_F : \mathbb{R}^J \to \mathbb{R}$ that returns the pseudo-Huber loss over the vector $v[S] \in \mathbb{R}^{|S|}$. Our approach can then be implemented as follows:

1. Estimate parameters $\beta_k - g_F^p(\beta_k^j : j \in S_{ref})$ for all $k \in \{0, \ldots, p-1\}$ using Clausen and Willis' algorithm to estimate penalized MLEs under the alternative hypothesis.

2. For each taxon $j \in \{1, \ldots, J\}$, do the following:

   (a) Define the set $S_R = S_{ref} \cup \{j\}$ (or as $S_R = S_{ref}$ if $j \in S_{ref}$). Define the model M_R on the set of taxa S_R with identifiability constraint g_R such that $g_R(v[S_R]) = g_{\text{ref}}(v[S_{ref}])$ for all $v \in \mathbb{R}^{|S_R|}$.

(b) Estimate parameters $\alpha_k - g_R^p(\alpha_k^j : j \in S_{ref})$ for all k under the null hypothesis $H_0^{j(R)} : \alpha_k^j - g_R^p(\alpha_k^j : j \in S_{ref}) = 0$ using Clausen and Willis' algorithm to estimate penalized MLEs under the null hypothesis.

(c) Use penalized MLEs estimated under the null hypothesis to construct the robust score statistic $T_{RS}^{H_0^{j(R)}}$, as defined in (3.8). Calculate the probability of observing a random variable from a χ_1^2 distribution that is larger than the robust score test statistic to obtain the robust score test p-value.

In practice, we often must make a small modification to this approach. We can only use Clausen and Willis' estimation algorithms with a data matrix Y for which all samples i have at least one non-zero taxon count. This is because when a sample Y_i consists only of zero values, the information matrix defined using the multinomial logistic regression log likelihood is not positive definite and cannot be inverted, and this inversion step is required by both estimation algorithms. Often the reduced data matrix $Y_{S_R^j}$ will fail to satisfy this condition, especially for small sets S_R. When this happens, we modify the set S_R^j to include additional taxa. We sequentially add taxa $j \in \{1, \ldots, J\} \setminus S_R^j$ to S_R^j, starting with the taxon with the highest prevalence (the remaining taxon j with the maximum $\sum_{i=1}^n \mathbb{I}\{Y_{ij} > 0\}$) in the dataset, and stopping when the updated data matrix $Y_{S_R^j}$ has a non-zero count for each sample i. In practice, we find that few taxa need to be added to the reduced models in order to achieve this condition. For example, the dataset analyzed in Section 3.5 has an abundance table with 73% of the entries equal to zero and includes $> 8,000$ taxa. In this analysis, only two taxa would be needed to be added to the selected reference set of 50 taxa in order for all sample abundance vectors Y_i to include at least one non-zero value.

3.3 Simulations

In this section, we investigate the Type I error rate control and power of our proposed robust score tests using reduced models and compare their performance to that of robust score tests using the full model, Clausen and Willis' robust Wald tests, as well as testing procedures im-

plemented in two popular differential abundance methods, ALDEx2 [Fernandes et al., 2013] and ANCOM-BC2 [Lin and Peddada, 2024]. The log fold-difference estimands in ALDEx2 and ANCOM-BC2 are not directly comparable to each other or to $\beta_k^j - g_F^p(\beta_k^j : j \in S_{ref})$ because they are a product of different modeling assumptions. However, in the context of their respective models, the ALDEx2 and ANCOM-BC2 estimands for taxon j both represent the expected log fold-difference of taxon j across covariate levels, relative to a measure of center across expected log fold-differences for all taxa. Therefore, we expect to observe an association between estimates and p-values from our method and estimates and p-values from these methods. We run ALDEx2 with a centered log ratio (CLR) transformation and 128 Monte Carlo iterations. To ensure comparability with other methods, we run ANCOM-BC2 without data manipulation in advance of testing (i.e., no "prevalence filtering", "structural zero detection", or sensitivity analysis screening with pseudocounts).

3.3.1 Data generation

We simulate data under several different settings. We consider sample sizes $n \in \{10, 50, 250\}$, total number of taxa $J \in \{50, 250\}$, and data that is drawn from Poisson and zero-inflated negative binomial (ZINB) distributions. We include the ZINB settings to evaluate the performance of our method on data with high levels of sparsity, which is a common characteristic of microbial abundances. We use a balanced design matrix that includes an intercept and one binary covariate.

We describe our simulation framework in detail in Appendix B.3. We construct β_1 such that the values range within the set $[-5, 5]$, in which few elements have values with large magnitudes and most elements have values with small magnitudes. We set β_1^{125} to 0 under the null hypothesis and to b under alternate hypotheses. We use a set of 24 categories as the reference set S_{ref} with log fold-differences that range across much of the range of β_1 values, but have smoothed median $g_F^p(\beta_1^j : j \in S_{ref}) = 0$.

We generate data by drawing each Y_{ij} value independently from either a Poisson distribution or a zero-inflated negative binomial (ZINB) distribution with mean given by the

taxonomic abundance model described in Section 3.2.1. Approximately 60% of counts in the simulated ZINB data are zero.

3.3.2 Type 1 error simulations

In the Type I error rate simulations, we test the null hypothesis $H_0^{125} : \beta_1^{125} - g_F^p(\beta_1^j : j \in S_{ref}) = 0$, for data generated according to this null hypothesis. We run $10,000$ trials for this simulation.

We first compare p-values from the robust score tests using the full model and robust score tests using the reduced models, applied to the same sets of simulated data. Aggregated over all simulation settings, these two sets of p-values have a Pearson correlation of 0.993 and a root mean squared difference of 0.034. Correlations and root mean squared differences separated by simulation setting are reported in Appendix **??**.

We now confirm that our proposal controls Type 1 error rate. The Type I error rate quantile-quantile plots can be seen in Figure 3.1. These plots compare the quantiles of p-values calculated from each of our methods on data simulated under the null hypothesis to quantiles of a Uniform$(0, 1)$ distribution. The robust score tests control or very nearly control the Type I error rate at a 0.05 level for all simulation settings. The robust Wald test fails to control Type I error at a 0.05 level for both settings with sample sizes of 25 and for the setting with $n = 50$ and ZINB data. This aligns with known behavior of robust Wald tests to typically be anticonservative for small sample sizes [Guo et al., 2005]. These simulation results let us draw similar conclusions to Clausen and Willis; that both versions of the robust score test control Type I error across these simulation settings, and the robust Wald test fails to control Type I error across most of these settings.

ALDEx2 is quite conservative in all settings except for Poisson data with $n = 250$. This corresponds with a downwards bias in this setting, in which the mean ALDEx2 estimate for β_1^{125} (subject to the identifiability constraint used by ALDEx2) is -0.39 across 500 trials ALDEx2 shows a similar downwards bias in all other Poisson settings (and a more subtle downwards bias in the ZINB settings), although the conservative testing procedure and lower

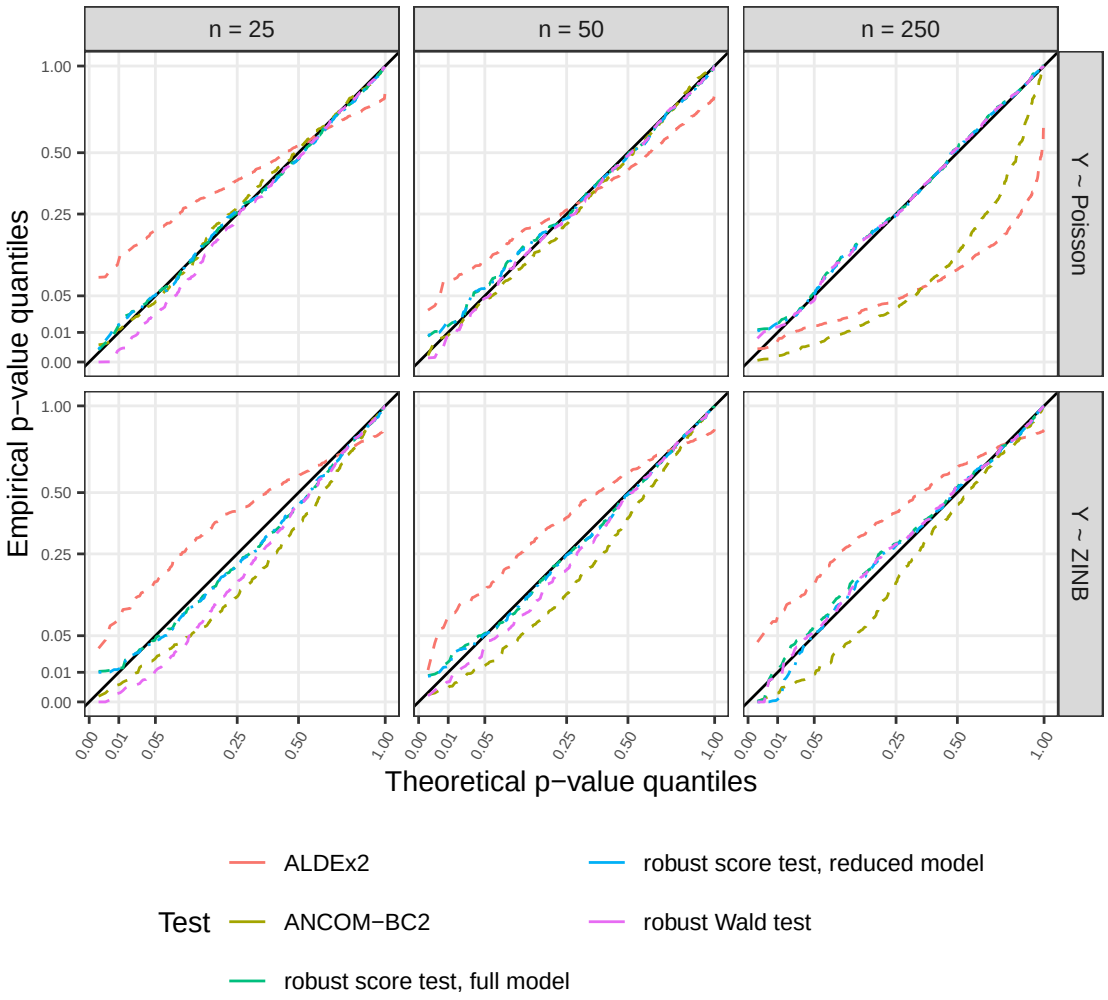


Figure 3.1: Quantiles of p-values obtained from the Type I error rate simulation compared to quantiles of a Uniform$(0, 1)$ distribution. Data is simulated under the null hypothesis using the Clausen and Willis mean model and draws from a Poisson or zero-inflated Negative Binomial (ZINB) distribution. P-values are generated from robust score tests from full and reduced models, as well as robust Wald tests, ALDEx2, and ANCOM-BC2. The $x = y$ line is shown in black, and represents quantiles of p-values from a test that controls Type I error rate at a nominal rate across the full range of p-values. Tests corresponding to lines above the $x = y$ line are conservative and control the Type I error rate and tests corresponding to lines below the $x = y$ line are anticonservative and fail to control the Type I error rate. Results come from a simulation with 500 trials.

power in these settings leads to conservative tests. Additionally, ALDEx2 does not produce p-values with a uniform$(0, 1)$ distribution under the null hypothesis. Across all simulation settings, the maximum p-value from ALDEx2 is 0.85.

ANCOM-BC2 fails to control the Type I error rate in all ZINB data settings and for Poisson data with $n = 250$. ANCOM-BC2 does not appear to be biased, and this anticonservative behavior likely comes from underestimated standard errors, especially for sparse data in the ZINB settings.

3.3.3 Power simulations

We test the same hypothesis, $H_0^{125} : \beta_1^{125} - g_F^p(\beta_1^j : j \in S_{ref}) = 0$ in our power simulations, but generate data according to specific alternate hypotheses $H_{A_b}^{125} : \beta_1^{125} = b$, for b values that take on twenty evenly spaced values between 0.25 and 5.00. We run 500 trials for this simulation.

As in the Type I error rate simulations, we first compare p-values from the robust score tests using the full model and robust score tests using reduced models, applied to the same sets of simulated data. The p-values from the two approaches have a Pearson correlation of 0.995 and a root mean squared difference of 0.018, aggregated over all simulation settings.

Power curves can be seen in Figure 3.2. For Poisson data, both robust score tests have high power in most settings, with the exception of settings with $n = 25$ and β_1^{125} magnitudes less than 1. For ZINB data with $n \in \{25, 50\}$, the power of the test ranges between approximately 5% and 95% across β_1^{26} magnitudes. The test has a power of 90% and higher for ZINB data with $n = 250$ and β_1^{125} magnitudes of 1 and greater.

The power for the robust Wald test, ALDEx2, and ANCOM-BC2, are only shown for simulation settings in which they control the Type I error rate at a 0.05 level. The robust score tests have power that approximately the same as ANCOM-BC2 or the robust Wald test across β_1^{125} magnitudes when these tests are included. ALDEx2 has lower power than all other methods for Poisson data with β_1^{125} magnitudes less than 1 and for all settings with ZINB data.

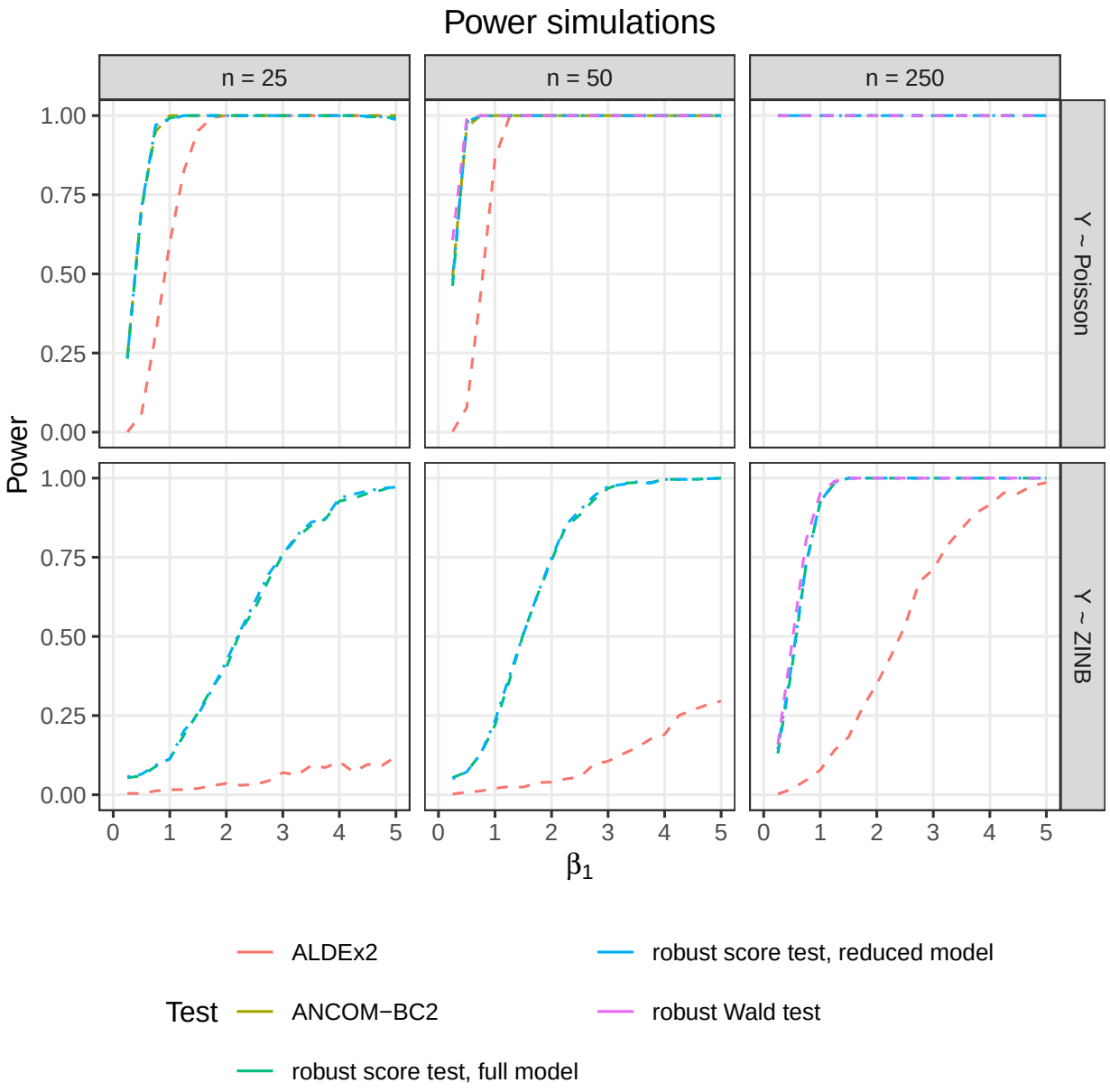


Figure 3.2: Power curve plots using p-values from power simulations. Data is simulated under the 20 different specific alternatives based on different magnitudes of β_1^{125} using the Clausen and Willis mean model and draws from a Poisson or zero-inflated Negative Binomial (ZINB) distribution. P-values are generated from robust score tests from the full and reduced model, as well as the robust Wald test on the full model, ALDEx2, and ANCOM-BC2. Tests are only shown for simulation settings in which they control the Type I error rate at a 0.05 level. Results come from a simulation with 500 trials.

The tests using the reduced model are faster than the tests using the full model in 75% of trials across simulation settings, and on average the reduced model tests run 9 times faster than the full model tests. These time savings with the reduced model are modest for 250 categories but far more impactful for thousands of categories, as seen in Section 3.5.

3.4 Parameter choice

In Section 3.2.6, the first step of our approach is to identify a reference set of taxa S_{ref}, such that $|S_{ref}| << J$. We will use the notation $g_F^p(v^{j'} : j' \in S)$ to refer to the function $g_F^p : \mathbb{R}^J \to \mathbb{R}$ that returns the pseudo-Huber loss over the vector $v[S] \in \mathbb{R}^{|S|}$, and simplify this to $g_F^p(\beta_k)$ for $S = \{1, \ldots, J\}$. We suggest defining our target parameter as $\beta_k^j - g_F^p(\beta_k^{j'} : j' \in S_{ref})$. In this section, we propose several methods to choose this reference set and investigate the behavior of these methods in simulation and in a metagenomic sequencing dataset.

This decision is trivial if we have additional data from a study's experimental design or scientific knowledge about the differential abundance of taxa in our dataset. If a synthetic compound was added to all samples of the same known quantity or if absolute abundances were measured for some taxa with quantitative PCR or droplet digital PCR, we could use this information to choose our reference set. We could also apply knowledge from previous studies of the same biome and covariate. If a previous study concluded that all taxa from a certain genus had little differential abundance across the same covariate, we could use taxa from that genus as a reference set. In the case of a single reference taxon, the constraint would be set to the β_k^j value for that taxon. If we know our reference taxon or taxa to be differentially abundant at a certain non-zero level c, we could adjust $g_F(\cdot)$ to be $g_F^p(\beta_k^{j'} - c : j' \in S_{ref})$. However, this type of experimental design and previous scientific knowledge are rare.

A naive approach is to randomly choose a subset of taxa to use as the reference set. If we would like to define a parameter that is similar to $\beta_k^j - g_F^p(\beta_k)$ using a small reference set, we could instead target the parameter $\beta_k^j - g_F^p(\beta_k^{j'} : j' \in S_{rand})$, where S_{rand} is a randomly chosen subset of taxa of a prespecified size. As $|S_{rand}| \to J$, the parameters $\beta_k^j - g_F^p(\beta_k)$ and $\beta_k^j - g_F^p(\beta_k^{j'} : j' \in S_{rand})$ would get closer and closer, although the computational advantages

of using reduced models for score testing would decrease. One drawback of this approach is its randomness, as the reference set and consequently the target parameter will change each time that a random subset of a given size is selected. This random parameter is less interpretable and meaningful than the one that includes all taxa in the dataset. It is also possible to randomly choose a "bad" reference set such that $g_F^p(\beta_k^{j'} : j' \in S_{ref})$ differs substantially from $g_F^p(\beta_k)$.

One advantage of the parameter $\beta_k^j - g_F^p(\beta_k)$ is the interpretation. We interpret the pseudo-Huber loss over the vector β_k to be the "typical log fold-difference" of taxa in the set, as it is a smoothed measure of center that is resistant to outliers. Therefore, our parameter $\beta_k^j - g_F^p(\beta_k)$ can be interpreted as the log fold-difference for taxon j, relative to the typical log fold-difference across all taxa. If we follow similar logic for a small reference set, we may want to choose a reference set made up of taxa that can all be considered "typical taxa". Consider a setting in which we know which taxa have the smallest log fold-differences relative to the pseudo-Huber loss over the vector β_k. We will refer to this known reference set as S_{kr}. The parameter $\beta_k^j - g_F^p(\beta_k^{j'} : j' \in S_{kr})$ will then have a very similar interpretation to $\beta_k^j - g_F^p(\beta_k)$, because $g_F^p(\beta_k^{j'} : j' \in S_{kr})$ and $g_F^p(\beta_k)$ both can be interpreted as "typical" log fold-differences across all taxa.

One plausible assumption is that in most microbiomes, the majority of taxa have no or low differential abundance across covariate levels. A relaxed version of this assumption is that the majority of taxa have similar differential abundances, which would account for scenarios in which most taxa increase or decrease in abundance at the same rate across covariate levels. Under either of these assumptions, and with a large set of taxa, we expect there to be many taxa j such that $|\beta_k^j - g_F^p(\beta_k)|$ is small. We can expect that in these cases $g_F^p(\beta_k)$ and $g_F^p(\beta_k^j : j \in S_{kr})$ will take on very similar values, and the shift in parameter values associated with using the reference set S_{kr} instead of the full set of taxa when defining the parameters will be quite small. Therefore, this approach would target a similar parameter to the parameter suggested by Clausen and Willis.

In practice, we do not know the set S_{kr} of taxa with the smallest log fold-differences

relative to the pseudo-Huber loss over β_k. Instead, we must estimate this set. We will refer to this approach as using a data-driven reference set. In this approach, we can start by using Clausen and Willis' algorithm to estimate $\hat{\beta}_k^j - g_F^p(\hat{\beta}_k)$. We can use these parameter estimates to identify a small subset of taxa with the smallest $|\hat{\beta}_k^j - g_F^p(\hat{\beta}_k^j)|$ values. We will call this subset S_{dd} and use it as the reference set. We can then shift all estimates to align with this new reference set and consequent identifiability constraint. This is computationally trivial, because in order to compute estimated parameters for a new constraint, we can simply set $\hat{\beta}_k^{j,\,\text{new}} = \hat{\beta}_k^{j,\,\text{old}} - g^{\text{new}}(\hat{\beta}_k^{\text{old}})$. Under the same assumption that we describe above, in which we expect the majority of taxa in an analysis to have no or low differential abundance across covariate levels, then we would expect that our parameter estimates would shift little when comparing $\hat{\beta}_k^j - g_F^p(\hat{\beta}_k)$ and $\hat{\beta}_k^j - g_F^p(\hat{\beta}_k^{j'} : j' \in S_{dd})$.

We believe there are two major benefits to using S_{dd} as a reference set. The first is that we believe that in most cases, both the true parameters and the estimated parameters will shift by a small amount when using the full set of taxa for parameter definition compared to the reference set S_{dd}. Therefore, we expect to typically observe similar estimation and inference results when using Clausen and Willis' approach with parameter $\beta_k^j - g_F^p(\beta_k)$ and our approach with parameter $\beta_k^j - g_F^p(\beta_k^{j'} : j' \in S_{kr})$ and estimate $\hat{\beta}_k^j - g_F^p(\hat{\beta}_k^{j'} : j' \in S_{dd})$. The second is that these parameters have similar interpretations. Both compare the log fold-difference of each taxon to the typical log fold-difference across all taxa, using a measure of center that is robust to outliers. We can use analyses with either parameter in order to identify the taxa that are the most enriched or depleted across covariate levels.

This data-driven approach that we describe is a clear case of "double dipping", because the same data is used for two separate tasks in an analysis. Double dipping causes invalid inference in many settings, and has inspired the field of selective inference, which involves the study of how to rigorously "measure the strength of the resulting selections" after "searching through the data for the strongest associations" [Taylor and Tibshirani, 2015]. However, the major difference between the problems tackled in selective inference and our suggestion is that for most selective inference tasks, the first step is using the data to find interesting

associations and the second step is testing those associations. In our setting, we are not looking for interesting associations. We are instead searching for taxa that are the *least* interesting to inform a measure of center. We then use this baseline as a way to define a meaningful estimand.

A major concern about double dipping is losing Type I error rate control for a test that would otherwise be valid. However, in order to consider Type I error rates, we must be able to define the null hypothesis that we are testing with the robust score test. An additional concern with using the reference set S_{dd} for parameter definition and then running robust score tests is that the parameter depends on the data, and therefore there is not a clear null hypothesis in terms of parameters that we are testing with the robust score test. We address both of these concerns, about precise hypothesis definition and loss of Type I error rate control, by showing that the distribution of robust score test statistics, constructed using parameter estimates $\hat{\beta}_k - g_F^p(\hat{\beta}_k^{j'} : j' \in S_{dd})$ under the restriction that $\hat{\beta}_k^j - g_F^p(\hat{\beta}_k^{j'} : j' \in S_{dd}) = 0$, approximates the distribution of robust score test statistics constructed using parameter estimates $\hat{\beta}_k - g_F^p(\hat{\beta}_k^{j'} : j' \in S_{kr})$ under the restriction that $\hat{\beta}_k^j - g_F^p(\hat{\beta}_k^{j'} : j' \in S_{kr}) = 0$. We show this in finite samples through simulations in Section 3.4.1. Asymptotically, we will prove that we have valid tests of the null hypothesis, $H_0 : \hat{\beta}_k^j - g_F^p(\hat{\beta}_k^{j'} : j' \in S_{kr}) = 0$, using test statistics constructed using the reference set S_{dd}.

First, we will assume that no taxa have exactly the same magnitudes $|\beta_k^j - g_F^p(\beta_k)|$, and therefore there is a unique known reference set S_{kr} of size J_r for $S_{kr} \equiv \{j : rank(|\beta_k^j - g_F^p(\beta_k)|) \leq J_r\}$. This assumption will simplify the following propositions and proofs, although similar results could be derived for a relaxed version of this assumption, in which taxa can have the same parameter magnitudes. We will also assume that for all data realizations, our estimates $|\hat{\beta}_k^j - g_F^p(\hat{\beta}_k)| \neq |\hat{\beta}_k^{j''} - g_F^p(\hat{\beta}_k)|$ for all pairs of taxa $j, j'' \in \{1, \ldots, J\}$.

Proposition 4. *When we choose a data-driven reference set S_{dd} to be the J_r taxa with the smallest values of $|\hat{\beta}_k^j - g_F^p(\hat{\beta}_k)|$, then $Pr(S_{dd} = S_{kr}) \to 1$ as $n \to \infty$.*

The proof of this proposition is in Appendix B.4.

Proposition 5. *Define the random variables T_{dd}^{full} and T_{dd}^{red} as the robust score test statistics using the full model and the reduced model respectively that are constructed using parameter estimates $\hat{\beta}_k - g_F^p(\hat{\beta}_k^j : j \in S_{dd})$ under the restriction that $\hat{\beta}_k^j - g_F^p(\hat{\beta}_k^j : j \in S_{dd}) = 0$. Similarly define T_{kr}^{full} and T_{kr}^{red} using the reference set S_{kr}. Under the null hypothesis $H_0 : \beta_k^j - g_F^p(\beta_k^j : j \in S_{kr}) = 0$, $T_{dd}^{full} \xrightarrow{d} \chi_1^2$ and $T_{dd}^{red} \xrightarrow{d} \chi_1^2$.*

Proof: Under the null hypothesis $H_0 : \beta_k^j - g_F^p(\beta_k^j : j \in S_{kr}) = 0$, we know that $T_{kr}^{full} \xrightarrow{d} \chi_1^2$ based on theory about the asymptotic behavior of the robust score test under the null hypothesis for a predetermined identifiability constraint. We know that $T_{kr}^{red} \xrightarrow{d} \chi_1^2$ from Proposition 2. Additionally, we can show that $|T_{dd}^{full} - T_{kr}^{full}| \xrightarrow{p} 0$ and $|T_{dd}^{red} - T_{kr}^{red}| \xrightarrow{p} 0$. We will show the former, and the latter follows from the same logic. For any $\epsilon > 0$,

$$Pr(|T_{dd}^{full} - T_{kr}^{full}| \geq \epsilon) = Pr(|T_{dd}^{full} - T_{kr}^{full}| \geq \epsilon | S_{dd} = S_{kr}) Pr(S_{dd} = S_{kr}) \tag{3.9}$$

$$+ Pr(|T_{dd}^{full} - T_{kr}^{full}| \geq \epsilon | S_{dd} \neq S_{kr}) Pr(S_{dd} \neq S_{kr}) \tag{3.10}$$

$$= Pr(|T_{kr}^{full} - T_{kr}^{full}| \geq \epsilon | S_{dd} = S_{kr}) Pr(S_{dd} = S_{kr}) \tag{3.11}$$

$$+ Pr(|T_{dd}^{full} - T_{kr}^{full}| \geq \epsilon | S_{dd} \neq S_{kr}) Pr(S_{dd} \neq S_{kr}) \tag{3.12}$$

$$= 0 + Pr(|T_{dd}^{full} - T_{kr}^{full}| \geq \epsilon | S_{dd} \neq S_{kr}) Pr(S_{dd} \neq S_{kr}) \tag{3.13}$$

$$\leq 0 + Pr(S_{dd} \neq S_{kr}) \tag{3.14}$$

$$\lim_{n \to \infty} Pr(|T_{dd}^{full} - T_{kr}^{full}| \geq \epsilon) \leq \lim_{n \to \infty} Pr(S_{dd} \neq S_{kr}) = 0 \tag{3.15}$$

Therefore, $|T_{dd}^{full} - T_{kr}^{full}| \xrightarrow{p} 0$. Finally, we will apply Slutsky's theorem to this convergence in probability result and the convergence in distribution result about T_{kr}^{full}.

$$T_{dd}^{full} = T_{dd}^{full} - T_{kr}^{full} + T_{kr}^{full} \tag{3.16}$$

$$\xrightarrow{d} \chi_1^2 \tag{3.17}$$

Using the same argument, we can also prove that $T_{dd}^{red} \xrightarrow{d} \chi_1^2$. □

Because T_{dd}^{full} and T_{dd}^{red} both asymptotically have χ_1^2 distributions under the null hypothesis $H_0 : \beta_k^j - g_F^p(\beta_k^j : j \in S_{kr}) = 0$, this approach of using a data-driven reference set S_{dd} is a valid way to test this hypothesis when S_{kr} is unknown, and the asymptotic Type I error

rate control from the robust score test is not affected by using the data to determine the reference set. Despite this result, we expect that some researchers will still be wary of this approach. To address this, we also consider data-driven approaches to choosing a reference set that split the data in order to use part of the data to choose the reference set and the rest of the data for estimation and inference. We consider both sample-splitting and Poisson thinning [Neufeld et al., 2024a] as methods to split our data into two sets.

In sample splitting, we randomly choose a subset of samples of size $n_r = \lceil \epsilon \times n \rceil$ for $\epsilon \in (0, 1)$ to use for reference set selection, and then use the remaining $n_a = n - n_r$ samples for estimation and inference. The ϵ value can be tuned to determine how much data to use for reference set selection versus analysis. This results in reference selection covariate data $X^r \in \mathbb{R}^{n_r \times p}$ and count data $Y^r \in \mathbb{R}^{n_r \times J}$, and analysis covariate data $X^a \in \mathbb{R}^{n_a \times p}$ and count data $Y^a \in \mathbb{R}^{n_a \times J}$. The benefit of sample splitting is that it doesn't require any distributional assumptions for our data.

In Poisson thinning, we separate each Y_{ij} into two counts Y^r_{ij} and Y^a_{ij} such that $Y^r_{ij} + Y^a_{ij} = Y_{ij}$. We use $Y^r \in \mathbb{R}^{n \times p}$ for reference set selection and $Y^a \in \mathbb{R}^{n \times J}$, and use X as covariate data for both tasks. If the values Y_{ij} are truly distributed as Poisson random variables, then the resulting Y^r_{ij} and Y^a_{ij} will also be distributed as Poisson, will be independent, and will have $\mathbb{E}[Y^a_{ij}] = \epsilon \times \mathbb{E}[Y_{ij}]$ and $\mathbb{E}[Y^a_{ij}] = (1 - \epsilon) \times \mathbb{E}[Y_{ij}]$ for $\epsilon \in (0, 1)$, where ϵ is a tuning parameter to determine what approximate proportion of each count should be included in Y^r. However, if the counts are not Poisson distributed, then the two datasets will not be fully independent and Y^r_{ij} and Y^a_{ij} will not be Poisson distributed. For example, Neufeld et al. [2024b] show that when applying Poisson thinning to data that is generated according to a negative binomial distribution with overdispersion paramer b, there will be a positive correlation between the Y^r_{ij} and Y^a_{ij} for all i and j, and the correlation will be a function of ϵ, μ_{ij}, and b. This correlation increases as the overdispersion in the data increases. We could also consider thinning with other distributions, but each will require its own parameteric assumptions.

When using sample splitting or Poisson thinning to produce two separate (and hopefully

independent) datasets, we will use the first set Y^r to choose our reference set. We will do this in the same way as the data-driven approach described above. We will use Clausen and Willis' estimate algorithm to estimate $\beta_k^j - g_F^p(\beta_k)$, identify a small subset of taxa with the smallest $|\beta_k^j - g_F^p(\beta_k)|$ values, and define the reference set as this subset. We will refer to the reference set from sample splitting as S_{ss} and the reference set from Poisson thinning as S_{th}. Once we have a reference set, we will estimate parameters and perform robust score tests using Y^a.

Similarly to the random reference set approach, both sample splitting and thinning are random procedures and repeated analyses could lead to different reference sets and analysis results. Sample splitting is also associated with a loss in power, because we will have fewer observations to use in our analysis dataset. Poisson thinning cannot be directly applied to a non-integer dataset because only intergers can be thinned with the Poisson distribution. We often observe Y_{ij} values in the form of coverages from metagenomic sequencing data, which represent observed abundances and take on non-integer values. In order to apply Poisson thinning to non-integer Y_{ij} values, we must round all Y_{ij} values to the nearest integer to apply Poisson thinning. Despite these drawbacks, the benefit of these approaches are that we can use differential abundance signal in the data to choose a reference set such that $|g_F^p(\beta_k^j : j \in S_{ref}) - g_F^p(\beta_k)|$ is small, without reusing that data for estimation and inference.

Going forward, we will refer to the reference set selection approach that uses the whole dataset as the data-driven approach, and will refer the other two approaches that utilize the data based on their dataset splitting methods, as the sample-splitting approach and the thinning approach.

3.4.1 Reference set selection simulations

In this section we run additional Type I error and power simulations in order to compare the performance of our data-driven reference set approaches to each other and to the Clausen and Willis approach with constraint $g_F^p(\beta_k)$ over all taxa. We will refer to the Clausen and Willis approach as having a reference set of all taxa. We also consider the known reference

set S_{kr}, in which we use information about the true parameter values to define the reference set S_{kr} as the set of taxa with the smallest values of $\beta_k^j - g_F^p(\beta_k)$. When we use a known reference set of size 26, $g_F^p(\beta_k) = g_F^p(\beta_k^{j'} : j' \in S_{kr}) = 0$.

For all constraints based on data-driven reference sets, we consider reference sets of size 25. For both sample splitting and thinning we set the tuning parameter $\epsilon = 0.25$. When using the sample splitting approach, we assign $\lceil 0.25 \times n \rceil = 13$ samples to the reference set selection dataset ($X^r \in \mathbb{R}^{13\times 2}, Y^r \in \mathbb{R}^{13\times 250}$) and the remaining $\lfloor 0.75 \times n \rfloor = 37$ samples to the analysis dataset ($X^a \in \mathbb{R}^{37\times 2}, Y^a \in \mathbb{R}^{37\times 250}$). When using the Poisson thinning approach, we generate $Y^r \in \mathbb{R}^{50\times 250}$ and $Y^a \in \mathbb{R}^{50\times 250}$ such that $Y_{ij}^r + Y_{ij}^a = Y_{ij}$ and $\mathbb{E}[Y_{ij}^a] = 0.25 \times \mathbb{E}[Y_{ij}]$ and $\mathbb{E}[Y_{ij}^a] = 0.75 \times \mathbb{E}[Y_{ij}]$. For simulation settings with Poisson distributed data, Y_{ij}^r and Y_{ij}^a are also Poisson distributed and they are independent. For simulations settings with ZINB data, the expectations $0.25 \times \mathbb{E}[Y_{ij}]$ and $\mathbb{E}[Y_{ij}^a] = 0.75 \times \mathbb{E}[Y_{ij}]$ still hold, but we will not have results about the distribution of Y_{ij}^r and Y_{ij}^a, and there two datasets will be positively correlated [Neufeld et al., 2024b].

In the Type I error rate simulations, we generate data under the null hypothesis that $\beta_1^{J/2} - g_F^p(\beta_1) = \beta_1^{J/2} - g_F^p(\beta_1^{j'} : j' \in S_{kr}) = 0$. Therefore, in these simulations, null hypotheses based on parameters $\beta_1^{J/2} - g_F^p(\beta_1)$ and $\beta_1^{J/2} - g_F^p(\beta_1^{j'} : j' \in S_{kr})$ are both true. In this framework, we will evalulate the ability of each approach to reject this null hypothesis at a nominal level, regardless of whether that approach estimates parameters $\beta_1^{J/2} - g_F^p(\beta_1)$ or $\beta_1^{J/2} - g_F^p(\beta_1^{j'} : j' \in S_{kr})$ that appears in the null hypothesis or the similar parameter $\beta_1^{J/2} - g_F^p(\beta_1^{j'} : j' \in S_{ref})$ for the data-driven reference sets S_{dd}, S_{ss}, and S_{th}.

Quantile-quantile plots of p-values from this simulation can be seen in Figure 3.3. The empirical p-value distributions are nearly the same across all five reference set approaches. For each of these simulation settings, the distribution of p-values under the null hypothesis $\beta_1^{J/2} - g_F^p(\beta_1) = 0$ are approximately Uniform(0, 1). Therefore, in this simulation we do not see a loss of Type I error rate control when using data-driven referense set S_{dd} directly from the data or sets S_{ss} and S_{th}, when testing hypotheses of the form $\beta_k^j - g(\beta_k^{j'} : j' \in S_{kr}) = 0$.

We also run power simulations, in which we generate data such that $\beta_1^{J/2} - g_F^p(\beta_1) = b$ for

$b \in [0.25, 5]$). The power results can be seen in Figure 3.4. In the Poisson settings, the power is nearly the same for all approaches, although it is slightly lower for the sample splitting and thinning approaches for signals with $\beta_1 < 1$. The same pattern holds for ZINB data with $n = 250$. However, for the ZINB data with $n \in \{25, 50\}$, the power for tests with reference sets S_{dd} and S_{kr} is higher than the power for the test that uses all taxa, especially for medium to large signals. These results are explored further in Appendix B.5. In this appendix, we compare test statistic distributions between approaches under the null and specific alternate hypotheses. We find the empirical test statistic distributions for robust score tests using both full and reduced models with reference sets S_{dd} and S_{kr} to be very similar, across hypotheses and simulation settings. We found that in all simulation settings, for specific alternate hypothesis values of 2 and larger, the test statistic distributions shifted towards higher test statistic values for reference sets S_{dd} and S_{kr} in comparison to the approach using all taxa. This explains the difference in power between these approaches in some simulation settings. We hypothesize that this could be driven by a higher "inside" portion of the robust score test statistic for the approach using all taxa, representing a higher estimated standard error of the score vector for this approach, and show simulation results that support this idea.

The mean squared errors (MSEs) for estimating the parameter $\beta_1^{J/2} - g_F^p(\beta_k)$ (or equivalently $\beta_1^{J/2} - g_F^p(\beta_k^{j'} : j' \in S_{kr})$) in these Type I error simulations are contained in Tables 3.1. Across all simulation settings, the estimation MSE is very similar between the approachs that uses all taxa and the reference sets S_{dd} and S_{kr}, and between the sample splitting and thinning approaches. Estimation is more accurate for the all taxa, S_{dd}, and S_{kr} approaches. This implies that the estimates $\hat{\beta}_1^j - g_F^p(\hat{\beta}_1)$, $\hat{\beta}_1^j - g_F^p(\hat{\beta}_1^j : j \in S_{dd})$, and $\hat{\beta}_1^j - g_F^p(\hat{\beta}_1^j : j \in S_{kr})$ have similar accuracy for estimating the parameter $\beta_1^j - g_F^p(\beta_1) = \beta_1^j - g_F^p(\beta_1^j : j \in S_{kr})$ for data generated under these simulation settings.

This simulations demonstrate that for data simulated under our generation process, Type I error is not inflated by using a data-driven method to select a reference set, whether using all the data to select S_{dd} or avoiding double dipping by using sample splitting or thinning to

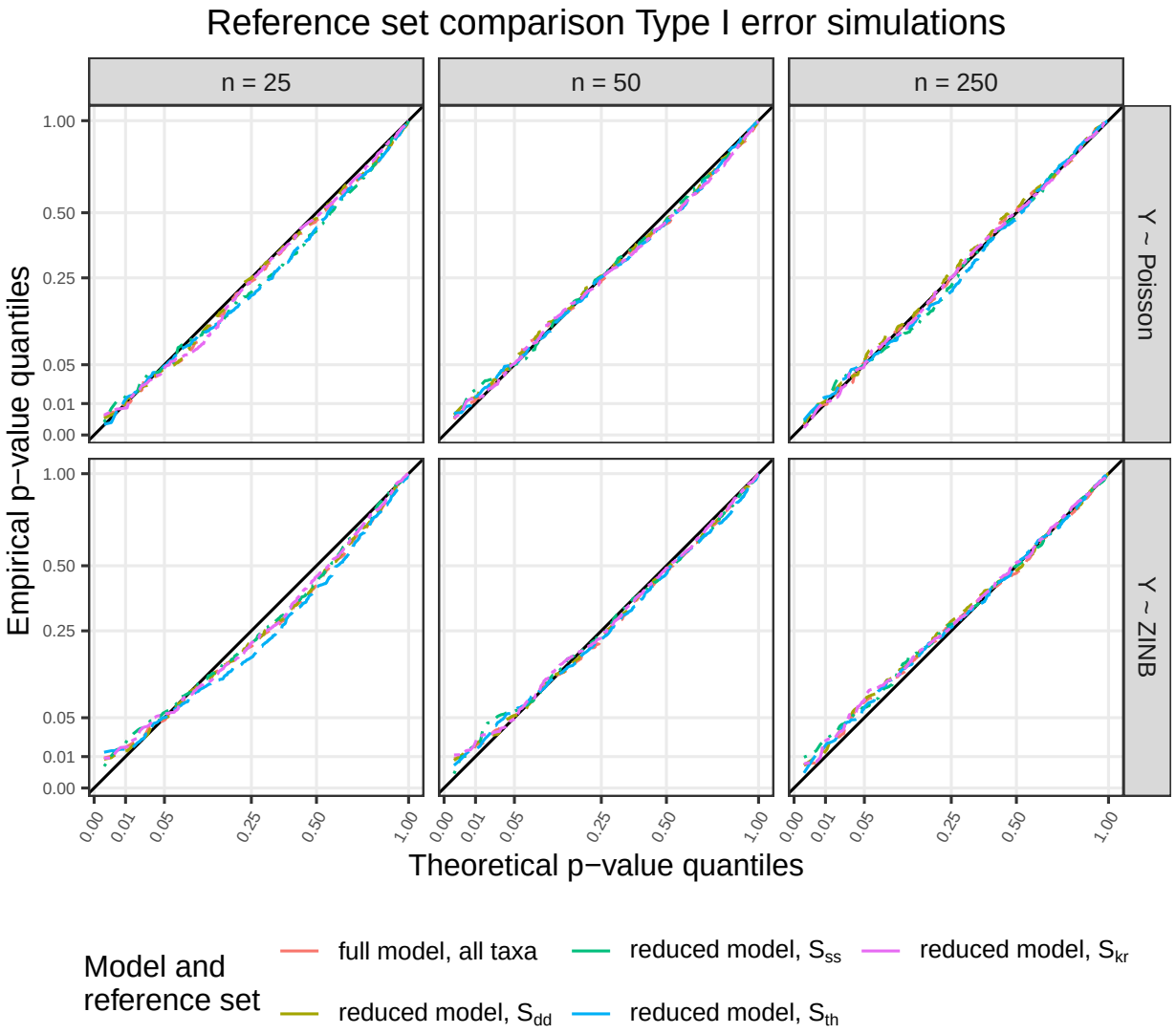


Figure 3.3: Quantiles of p-values obtained from the reference set Type I error rate simulation compared to quantiles of a Uniform$(0, 1)$ distribution. Data is simulated under the null hypothesis $\beta_1^{J/2} - g_F^p(\beta_k) = 0$. P-values are generated from robust score tests on the full model with the constraint $g_F^p(\beta_k)$ and on reduced models with data-driven reference sets S_{dd}, S_{ss}, and S_{th}, and a known reference set S_{kr}. The $x = y$ line is shown in black, and represents quantiles of p-values from a test that controls Type I error rate at a nominal rate across the full range of p-values. Results come from a simulation with 500 trials.

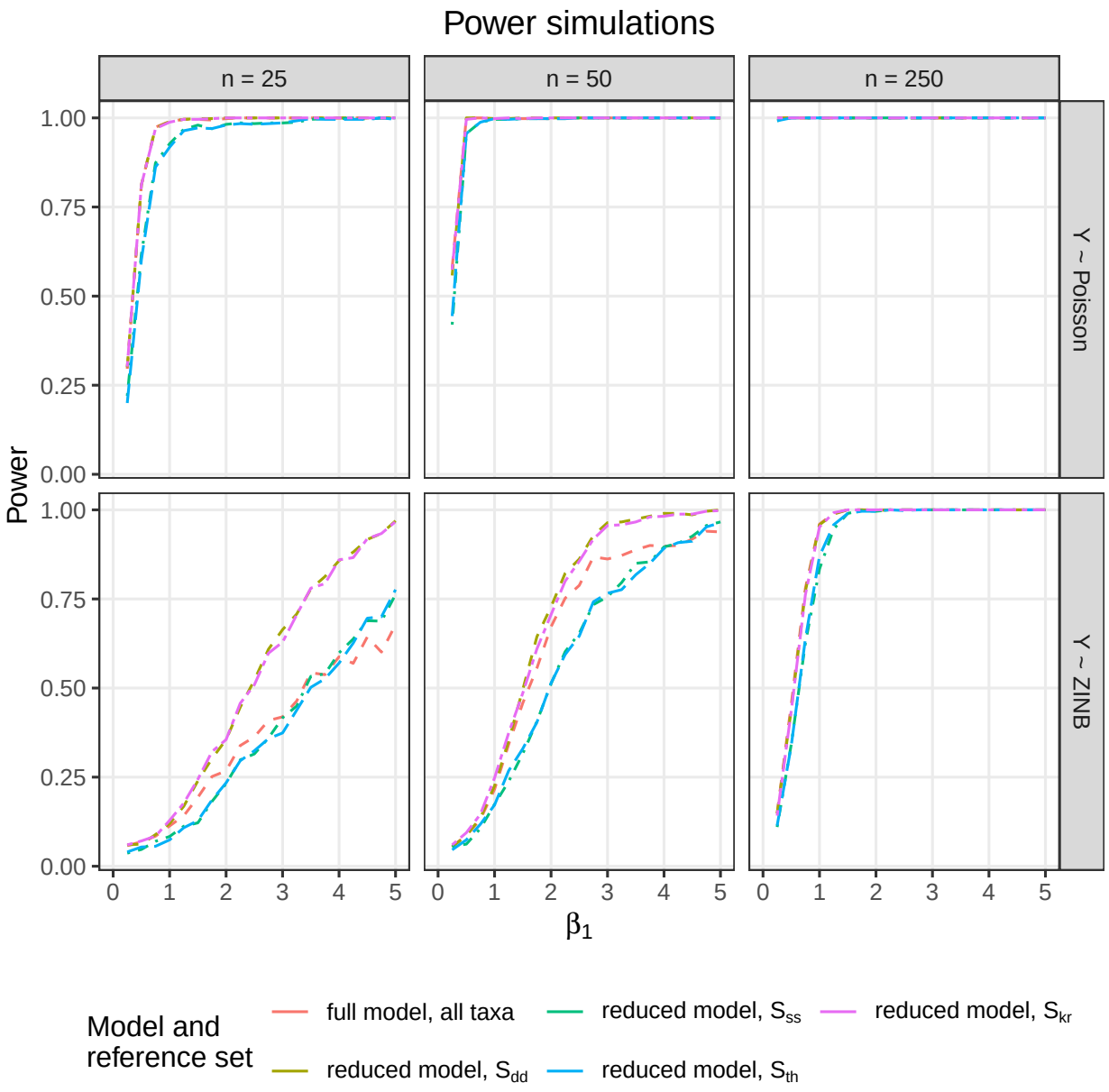


Figure 3.4: Power curve plots using p-values from reference set power simulations. Data is simulated under the 20 different specific alternatives based on different magnitudes of $\beta_1^{J/2} - g_F^p(\beta_k)$. P-values are generated from robust score tests on the full model with the constraint $g_F^p(\beta_k)$ and on reduced models with data-driven reference sets S_{dd}, S_{ss}, and S_{th}, and a known reference set S_{kr}. Results come from a simulation with 500 trials.

n	data	all taxa	S_{dd}	S_{kr}	S_{ss}	S_{th}
25	Poisson	0.025	0.025	0.024	0.040	0.041
50	Poisson	0.011	0.011	0.011	0.015	0.016
250	Poisson	0.002	0.002	0.002	0.003	0.003
25	ZINB	1.061	1.061	1.008	1.954	1.843
50	ZINB	0.363	0.365	0.348	0.555	0.526
250	ZINB	0.069	0.069	0.067	0.091	0.090

Table 3.1: MSE from estimating $\beta_1^{J/2} - g_F^p(\beta_1)$ in Type 1 error simulations from approaches with different reference sets. These results are aggregated across 500 trials.

select S_{ss} or S_{th}. However, the sample splitting and thinning approaches are associated with lower power, which is more noticeable for the ZINB data with smaller sample sizes. The lower power makes sense for sample splitting, as analysis is done using a smaller sample size than the other constraints. However, the cause of reduced power for the thinning approaches is not obvious, because differential abundance analysis should not be affected by approximately scaling every count by a constant value. We also find that for sparse ZINB data and smaller sample sizes, the approach that uses reference set S_{dd} leads to a test with higher power than the approach using the full set of taxa. Therefore, from a pragmatic viewpoint, we recommend using the data-driven reference set from the whole dataset. However, if this double dipping is of concern to the user of our method, we suggest using sample splitting or thinning to choose a reference set.

3.4.2 Reference set selection comparison in a data analysis

We also want to investigate how the choice of reference set and consequent choice of parameter affect estimation and inference using a real dataset. We consider the Wirbel et al. [2019] dataset that Clausen and Willis analyze in their paper. This dataset includes taxon counts

and covariate data from a meta-analysis of several cohorts of patients with and without colorectal cancer. We consider a simpler model than the one fit by Clausen and Willis, with case versus control as the only covariate. We only use samples from one study with a Chinese cohort. The full dataset that we use has 126 samples and 758 taxa. In this section, we are not interested in re-analyzing this dataset in order to identify differentially abundant taxa. Instead, we will use this to understand how results change across different reference sets used for parameter definition.

We consider many different reference sets. First, we consider the set of all taxa, which we will reference to as S_{all}. For the reference sets that are subsets of taxa, we consider subsets of sizes in $\{10, 30, 50, 100\}$. We randomly generate ten reference sets of each size to use as random reference set constraints. We include data-driven reference sets S_{dd} of each size, determined from the full dataset. Finally, we use sample splitting and thinning to each generate an additional ten reference sets S_{ss} and S_{th} of each size. For sample splitting and thinning, we use $\lfloor 0.25 \times n \rfloor = 31$ samples for reference set selection datasets (X^r, Y^r) and $\lceil 0.75 \times n \rceil = 95$ samples for analysis datasets (X^a, Y^a). For thinning, we generate Y^r and Y^a such that $\mathbb{E}[Y^r_{ij}] = 0.25 \times \mathbb{E}[Y_{ij}]$ and $\mathbb{E}[Y^a_{ij}] = 0.75 \times \mathbb{E}[Y_{ij}]$ for all samples i and taxa j. We estimate parameters associated with all reference sets using the Clausen and Willis estimation algorithm. For inference with the reference set S_{all}, we use Clausen and Willis' robust score tests on the full model. For inference with the reference sets that are subsets of taxa, we use our robust score tests on reduced models.

The results of this analysis can be seen in Figures 3.5 and 3.6. In Figure 3.5, we compare all 128 constraint settings. We calculate the root mean squared difference (RMSD) between the estimates and p-values from each reference set and the estimates and p-values using S_{all}. The data-driven reference sets S_{dd} have the lowest estimate and p-value RMSDs. The sample splitting reference sets S_{ss} have the highest estimate p-value RMSDs, while some of the random reference sets also have high p-value RMSDs. All analyses that we use reduced models for take less than 14 hours to run all tests serially using a computing cluster, with most taking less than 4 hours. The reference sets with the longest run times include some

of the random constraints. In comparison, it takes 57 hours to run tests serially using the full model for S_{all}. This decrease in computational time for the reduced models does not account for additional time reductions that can be achieved by running the robust score tests in parallel, which is feasible because they each involve estimating a small number of parameters under the null hypothesis.

In Figure 3.6, we look closer at three reference sets of size 50. We consider S_{dd}, as well as the S_{ss} and S_{th} sets with the lowest estimate RMSDs with respect to S_{all} estimates. The shift between estimates using the reference set of all taxa and S_{dd} is equal to $|g_F^p(\hat{\beta}_1) - g_F^p(\hat{\beta}_1^j : j \in S_{dd})| = 0.003$. The estimates using reference sets S_{ss} and S_{th} are not a constant shift from the estimates using S_{all} because only part of the data is available for estimation when using sample splitting or thinning. The reference set S_{ss} has p-values with the weakest association with S_{all} p-values, and the reference set S_{dd} has p-values with the strongest association. The latter association is stronger and more linear for larger p-values, and weaker for small p-values. This aligns with our results from the Type I error and power simulations in Section 3.4.1. In simulation settings with sparse data, the approaches that use S_{all} and S_{dd} have similar p-value distributions under the null hypothesis that $\beta_k^j - g_F^p(\beta_k) = \beta_k^j - g_F^p(\beta_k^j : j \in S_{kr}) = 0$, and the approach that uses S_{dd} has higher power for medium to strong specific alternative hypotheses.

In this comparative analysis, we show that the approach that uses reference sets S_{dd} identified from the full dataset results in estimates and p-values that are the most similar to estimates and p-values from the approach that uses reference set S_{all}. Because we can run robust score tests on reduced models when using these small reference sets, they are associated with a major reduction in computational burden in comparison to S_{all}, which requires us to run robust score tests with the full model. The random reference sets and sample splitting reference sets S_{ss} have the worst performance in this analysis in terms of the association of their estimates and p-values with the estimates and p-values from the approach with reference set S_{all}.

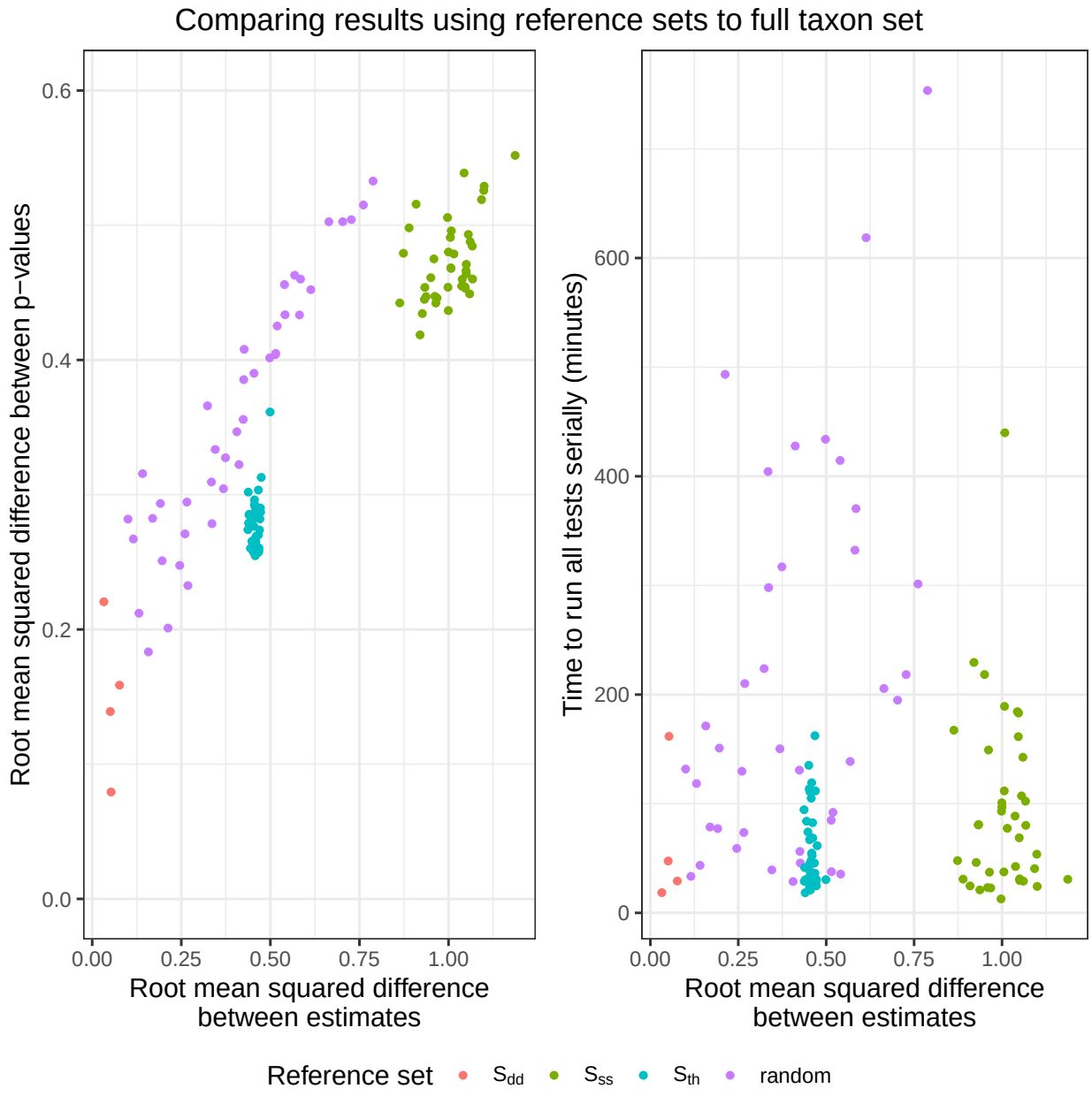


Figure 3.5: Results from comparing reference sets used in a differential abundance analysis using data from [Wirbel et al., 2019]. We consider 128 different reference sets, generated in five different ways. The root mean square distances (RMSDs) are calculated when comparing estimates and p-values from each reference set to the estimates and p-values from the analysis that uses S_{all}.

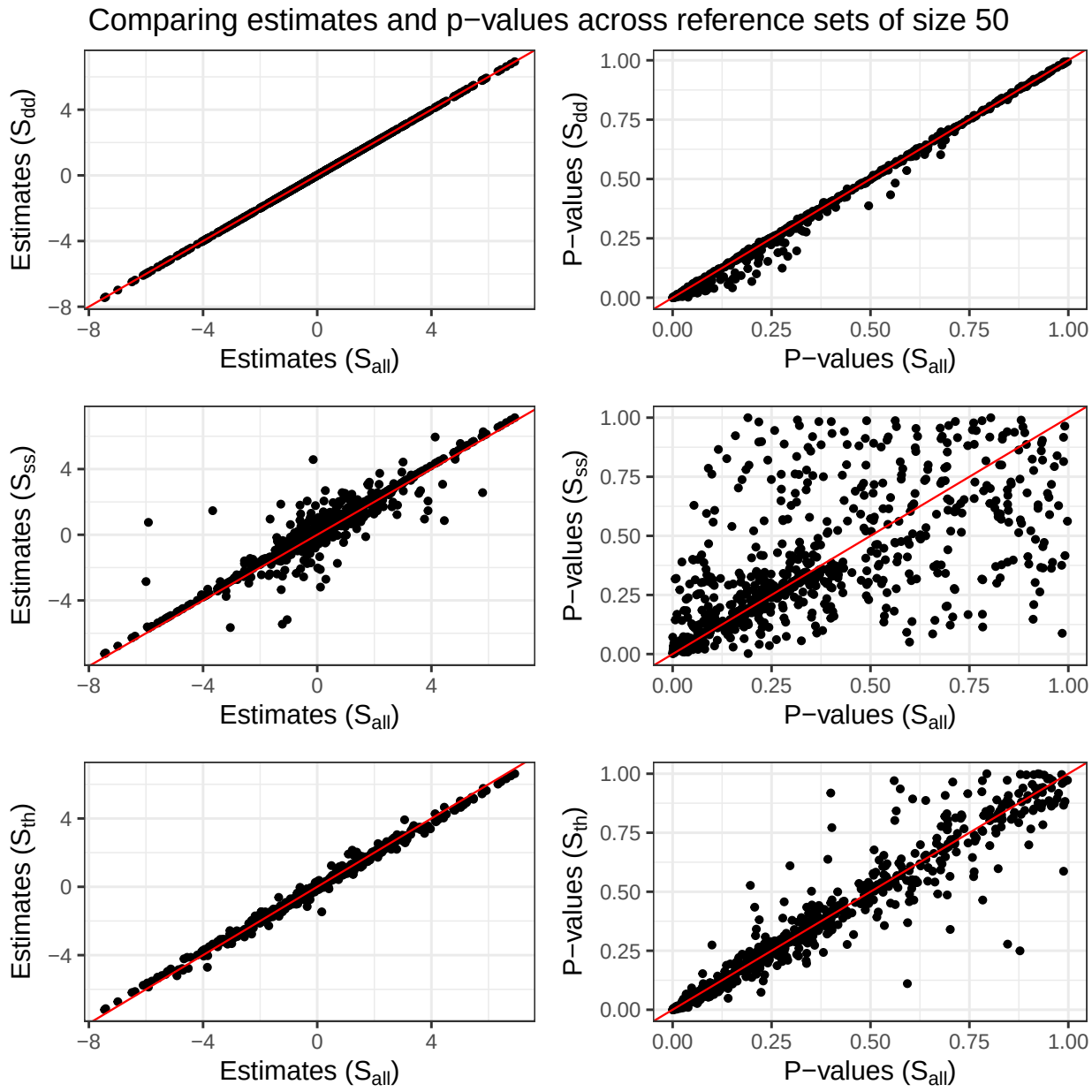


Figure 3.6: A comparison of estimates and p-values across four different reference sets in a differential abundance using data from [Wirbel et al., 2019]. We compare the estimates of $\beta_k^j - g_F^p(\beta_k)$ parameters using the reference set S_{all} and associated p-values to estimates of $\beta_k^j - g_F^p(\beta_k^j : j \in S_{ref})$ for reference sets S_{dd}, S_{ss}, and S_{th} which all have size 50 and associated p-values. We select S_{ss} and S_{th} as the reference sets from sample splitting and thinning of size 50 with the lowest estimation RMSE when comparing to estimates of $\beta_k^j - g_F^p(\beta_k)$.

3.5 Data analysis

In this section, we perform a differential abundance analysis on a large metagenomic sequencing dataset. This dataset is part of the Tara Oceans project [Karsenti et al., 2011], which involved a years long data collection process from oceans worldwide, in order to characterize as much microbial marine biodiversity as possible. Our dataset includes 89 samples, taken across all of the world's oceans, processed to provide coverage data for $8,360$ genome bins. A genome bin represents a taxonomic classifications at a finer level than species, and is defined in terms of procedures used to process whole genome sequencing data.

We fit a model in which we study the differential abundance of genome bins with respect to temperature. Temperature is measured in degrees Celsius and ranges from -2 degrees to 31 degrees. In order to make our differential abundance parameter more interpretable, we shift all temperature measurements by the sample mean of 17 degrees, so that our temperature covariate is mean-centered, and then divide all temperatures by 5. Differential abundance parameters with respect to this adjusted temperature covariate represents the expected log fold change in abundance of a given genome bin with respect to a five degree increase in temperature, relative to the typical log fold difference across genome bins.

3.5.1 Comparison between methods

We consider differential abundance parameters from the model given in Section 3.2.1 using four different identifiability constraints. We consider $g_F^p(\beta_k)$, the pseudo-Huber loss over β_k^j parameters from all taxa, and $g_F^p(\beta_k^j : j \in S_{ref})$ for reference sets S_{dd}, S_{ss}, and S_{th}. These reference sets are selected with data-driven approaches that use the full dataset, sample splitting, and Poisson thinning, respectively. We estimate each set of parameters using Clausen and Willis' estimation algorithm. We perform inference using robust score tests with the full model for the parameter defined with $g_F^p(\beta_k)$, and perform inference using robust score tests with reduced models for the parameters defined with $g_F^p(\beta_k^j : j \in S_{ref})$ for smaller reference sets. We only run robust score tests with the full model for a subset

of 418 (5%) of genome bins, because they require too many computational resources to run for all $8,360$ genome bins. We also compare to p-values from Clausen and Willis' robust Wald tests, and to differential abundance parameter estimates and p-values as calculated by ALDEx2 [Fernandes et al., 2013] and ANCOM-BC2 [Lin and Peddada, 2024].

The Pearson correlations between estimated parameters with different methods can be seen in Table 3.2. Each method estimates a different log fold-difference parameter. There is a correlation of 1.00 between estimates of parameter $\beta_k^j - g_F^p(\beta_k)$ and $\beta_k^j - g_F^p(\beta_k^j : j \in S_{dd})$ because they are related by a constant shift for all genome bins j, of magnitude 0.0004. These estimates have correlations of 0.96 and 0.98 with estimates of parameters $\beta_k^j - g_F^p(\beta_k^j : j \in S_{ss})$ and $\beta_k^j - g_F^p(\beta_k^j : j \in S_{th})$. ALDEx2 estimates have a correlation of 0.74 with estimates of $\beta_k^j - g_F^p(\beta_k)$. ANCOM-BC2 has much lower estimate correlations with all other methods, with a correlation 0.11 with the most similar method. This is because ANCOM-BC2 has a small set of genome bins with estimates that are very large with respect to the majority of genome bins, the largest of which has a magnitude of 530.

The Pearson correlations between p-values from each method can be seen in Table 3.3. Robust score tests using the full model were only run on a subset of 500 ($\approx 6\%$) of genome bins, so those correlations only consider that subset. In that subset, the robust score tests using the full model and the robust score tests reduced models with reference set S_{dd} have a correlation of 0.998. This demonstrates how well we are able to approximate the results of robust score tests with the full model of the parameters $\beta_k^j - g_F^p(\beta_k)$ with robust score tests with reduced models of parameters $\beta_k^j - g_F^p(\beta_k^j : j \in S_{dd})$.

Running robust score tests on reduced models results in a large increase in speed. When comparing median robust score test runtimes for the subset of tests run using the full model, the median time for tests using full models is 85 times larger than the median time for tests using reduced models. This improvement in computational efficiency is even more notable when comparing mean runtimes. The mean time for robust score tests using full models is $1,022$ times larger than the mean time for tests using reduced models. Additionally, the reduced score tests can be run on a computing cluster with 2 GB of memory allocated for

	Method 1	Method 2	Method 3	Method 4	Method 5	Method 6
Method 1	1.00	1.00	0.96	0.98	0.74	0.11
Method 2	1.00	1.00	0.96	0.98	0.74	0.11
Method 3	0.96	0.96	1.00	0.93	0.72	0.10
Method 4	0.98	0.98	0.93	1.00	0.72	0.10
Method 5	0.74	0.74	0.72	0.72	1.00	0.06
Method 6	0.11	0.11	0.10	0.10	0.06	1.00

Table 3.2: Pearson correlations between estimated differential abundance parameters from the analysis of TARA data, across different methods. Methods 1 through 4 use Clausen and Willis' estimation algorithm, to estimate parameters of the form $\beta_k^j - g_F^p(\beta_k^j : j \in S_{ref})$ for reference sets S_{all}, S_{dd}, S_{ss}, and S_{th} respectively. Method 5 and 6 correspond to parameter estimates from ALDEx2 [Fernandes et al., 2013] and ANCOM-BC2 [Lin and Peddada, 2024] respectively.

each test, while the full score tests are each run with 56 GB of memory. This difference means that many more reduced score tests can be run in parallel than full score tests, given a finite amount of memory.

Finally, we account for running $J = 8,360$ tests for each method by calculating q-values [Storey, 2002]. These q-values are functions of the p-values for each method, and their purpose is to control the false discovery rate across all tests run in an analysis. We use q-values instead the Benjamini-Hochberg (BH) method [Benjamini and Hochberg, 1995] due to their potential gain in power over the BH method, especially when there are a substantial proportion of non-null signals in the dataset.

We find that when using a q-value threshold of 0.01, the robust Wald tests and ANCOM-BC2 result in more than 80% of genome bins with q-values below the 0.01 threshold. This coincides with our findings in Section 3.3 that these methods have high power but do not

	Method 2	Method 3	Method 4	Method 5	Method 6	Method 7
Method 1	1.00	0.52	0.87	0.98	0.25	0.09
Method 2	1.00	0.62	0.79	0.95	0.29	0.15
Method 3	0.62	1.00	0.54	0.57	0.34	0.15
Method 4	0.79	0.54	1.00	0.76	0.29	0.14
Method 5	0.95	0.57	0.76	1.00	0.21	0.09
Method 6	0.29	0.34	0.29	0.21	1.00	0.08
Method 7	0.15	0.15	0.14	0.09	0.08	1.00

Table 3.3: Pearson correlations between p-values from hypothesis tests of differential abundance parameters from the analysis of TARA data, across different methods. Method 1 is robust score tests using the full model. These were only run for $500(\approx 6\%)$ of genome bins, so these correlations are only across this subset of genome bins. Methods 2 through 4 are robust score tests using reduced models, for parameters defined using reference sets S_{dd}, S_{ss}, and S_{th} respectively. Method 5 is Clausen and Willis' robust Wald test. Methods 6 and 7 are inferential procedures implemented in ALDEx2 [Fernandes et al., 2013] and ANCOM-BC2 [Lin and Peddada, 2024] respectively.

always control Type I error rate. The robust score tests that use reference sets S_{dd}, S_{ss}, and S_{th} result in 57%, 19%, and 42% of genome bins with q-values less than 0.01 respectively. This coincides with our findings in Section 3.4 that reference sets chosen with sample splitting and Poisson thinning can lead to lower power than reference sets chosen with the full dataset. ALDEx2 results in only 5% of genome bins with q-values less than 0.01. These q-values are more conservative in terms of false discovery rate control than the other q-values reported for this analysis because the ALDEx2 p-values range from 0 to 0.79 instead of from 0 to 1. Due to this truncated distribution of p-values, the proportion of true null categories cannot be estimated in the q-value proecedure and is instead fixed to 1. This results in a procedure

that is equivalent to the more conservative BH procedure [Storey, 2002].

Method	Proportion q-values ≤ 0.01
robust score test with parameters $\beta_k^j - g_F^p(\beta_k^j : j \in S_{dd})$	$4,2786$ (57%)
robust score test with parameters $\beta_k^j - g_F^p(\beta_k^j : j \in S_{ss})$	$1,558$ (19%)
robust score test with parameters $\beta_k^j - g_F^p(\beta_k^j : j \in S_{th})$	$3,4466$ (42%)
robust Wald test with parameters $\beta_k^j - g_F^p(\beta_k)$	$6,831$ (82%)
ALDEx2	395 (5%)
ANCOM-BC2	$7,336$ (88%)

Table 3.4: Proportion of genome bins from analysis of TARA data with q-values less than 0.01 across inferential methods.

3.5.2 Differential abundance

We proceed with the differential abundance analysis using parameters defined as $\beta_k^j - g_F^p(\beta_k^j : j \in S_{dd})$, and performing inference with robust score tests using reduced models. In Figure 3.7, we plot the distribution of estimated log fold-difference parameters. The estimates range from -25.4 to 8.3. The distribution is bimodal, with two modes of approximately -0.8 and 0.8. A log fold-difference of 0.8 corresponds with an expected $\approx$ 2-fold increase in abundance associated with a five degree increase in ocean temperature. The distribution has a long tail of negative estimates, with another small mode of approximately -6, which corresponds with an expected $\approx$ 400-fold decrease in abundance associated with a five degree increase in temperature.

These parameters are defined as log fold-differences with respect to the approximate smoothed median across all genome bins, and we do not know whether this approximate smoothed median is equal to a true log fold-difference of 0. However, we will consider the biological interpretation if the approximate smoothed median is very close to a true log

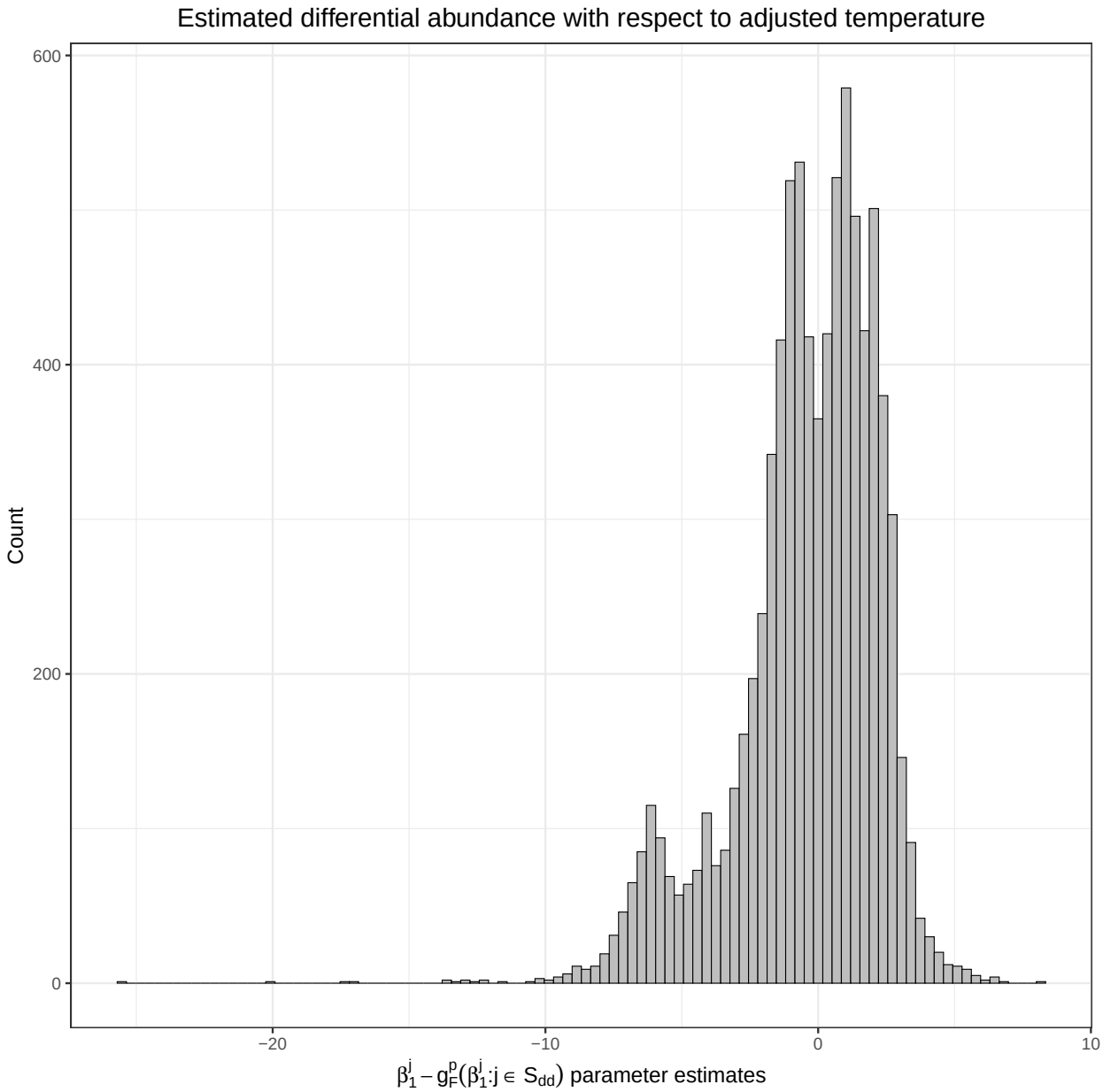


Figure 3.7: Histogram of estimates $\beta_k^j - g(\beta_k^j : S_{dd})$ from the analysis of the Tara oceans data. These represent expected log fold-differences in abundance of genome bins associated with a five degree increase in ocean temperature, relative to the typical log fold difference across genome bins.

fold-difference of 0. In this setting, the two modes in this distribution represent groups of taxa that have moderate expected increases or decreases in abundance associated with a five degree increase in temperature. In the distribution of expected log fold-difference magnitudes, the median magnitude is 1.5 and 75% of genome bins have magnitudes greater than 0.8, which corresponds with an expected $\approx$ 2-fold increase or decrease in abundance. This implies that only a small subset of genome bins have abundances that are relatively stable across temperature increases or decreases of five degrees. This would make sense based on our knowledge that ocean temperature has a major effect on the composition of marine microbial communities.

Using p-values from robust score tests on reduced models with reference set S_{dd}, we identify $4,786(57\%)$ genome bins with q-values less than 0.01. In order to filter this to a smaller set for further investigate, we use a smaller q-value threshold of $1e-5$ to identify the genome bins with the most evidence of differential abundance. There are 141 genome bins that pass this threshold. In Figure 3.8, we plot parameter estimates and confidence intervals computed with robust standard errors for these genome bins by the prevalence of these bins in the 89 Tara ocean samples. We color the genome bins in this plot by whether or not the bin belongs to the *Pelagibacter* genus. We choose this genus because 34% of the genome bins in this analysis belong to this genus, making it by far the most common genus in our analysis. This is unsurprising, as genome bins of the *Pelagibacter* genus account for most of the genome bins from the *Pelagibacterales* order. Most bacterial members of the SAR11 clade belong to the *Pelagibacterales* order. SAR11 bacteria are among the most abundant marine microbes, are found in every ocean, and are important to the ocean carbon cycle [Giovannoni, 2017]. Therefore, this is a particularly interesting genus to study in terms of differential abundance across temperature changes.

In Figure 3.8, we can see that the 141 genome bins with the most evidence of differential abundance include bins with large negative estimates and moderate positive estimates. These genome bins are prevalent in between 20 to 74 of the 89 samples. This is notable because despite not performing prevalence filtering for genome bins, rare genome bins are not included

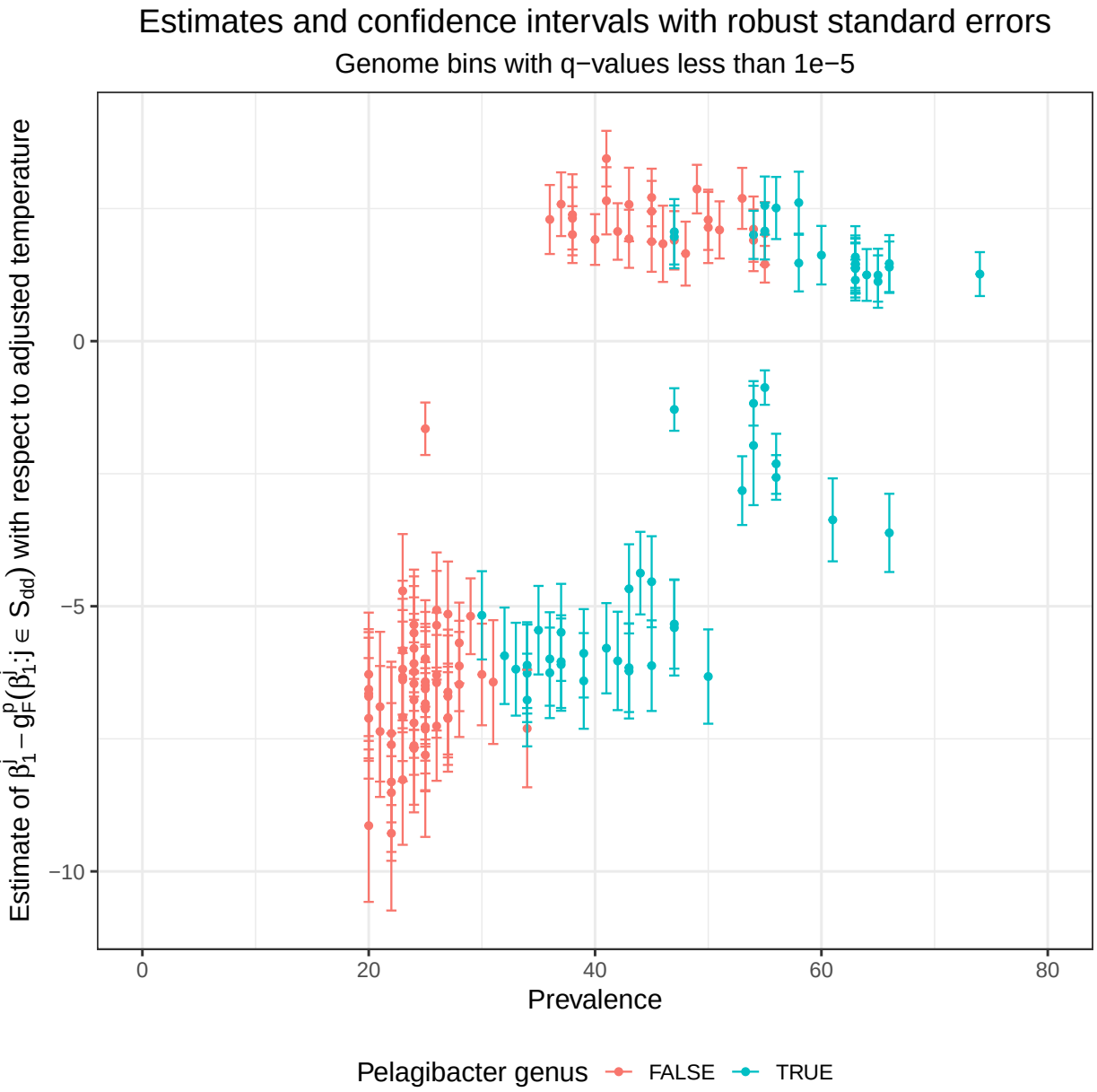


Figure 3.8: Estimates $\beta_k^j - g(\beta_k^j : S_{dd})$, with respect to the adjusted temperature covariate, and confidence intervals generated with robust standard errors. The estimates are plotted by prevalence of that genome bin in the 89 Tara oceans samples that we consider, and colored by whether or not the genome bin belongs to the *Pelagibacter* genus.

in this set of most differentially abundant bins. The bins with positive log fold-differences (these represent bins that are enriched in warmer water) have higher average prevalence compared to the bins with negative log fold-differences. Approximately half of these genome bins with the most evidence of differential abundance belong to the *Pelagibacter* genus. The *Pelagibacter* bins in this subset have both positive and negative differential abundance, and are more prevalent on average than the bins from other genera in this subset.

P-value rank	Estimate	Prevalence	Genus	Species
1	-6.16	43	Pelagibacter	None
2	-6.26	36	Pelagibacter	None
3	-4.67	43	Pelagibacter	Pelagibacter ubique
4	-6.12	45	Pelagibacter	Pelagibacter sp905612495
5	-7.26	26	Arcticimaribacter	None
6	-5.36	26	MS024-2A	MS024-2A sp905480425
7	-6.33	23	SGZJ01	None
8	-6.62	27	SCGC-AAA160-P02	None
9	-7.09	23	HTCC2207	HTCC2207 sp905182275
10	-7.31	28	UWMA-0277	UWMA-0277 sp905181635

Table 3.5: Characteristics of the ten genome bins from the Tara oceans differential abundance analysis with the most evidence of differential abundance based on p-values from robust score tests on reduced models with reference set S_{dd}. Estimates are of parameters $\beta_1^j - g_F^p(\beta_1^j : j \in S_{dd})$ with respect to increases of temperature of 5 degrees. Prevalence is the number of the 89 samples in which the genome bin is found. Genus and species represent the taxonomic classification of the genome bin. All genome bins have p-values and q-values less than $10e-10$.

Finally, we look closer at the genome bins with the most evidence of differential abun-

dance. In Table 3.5, we can see that the ten bins with the most evidence of differential abundance all have negative log fold-differences between -7.5 and -4.5. The four bins with the lowest p-values have higher prevalences and belong to the *Pelagibacter* genus, and the next six bins have lower prevalences and belong to other genera. In Table 3.6, we list the ten genome bins with the most evidence of positive differential abunbdance. These bins are ranked in the sixties to eighties out of all genome bins in terms of p-values. Compared to the bins listed in Table 3.5, they have more moderate estimates and higher prevalences. Three of these genome bins also belong to the *Pelagibacter* genus.

P-value rank	Estimate	Prevalence	Genus	Species
62	2.07	42	TMED80	TMED80 sp002170405
63	2.71	45	CACNYO01	None
67	2.29	50	MEDG-81	MEDG-81 sp902595715
70	2.58	37	GCA-002708715	GCA-002708715 sp902602795
75	2.65	41	MED-G52	MED-G52 sp001627375
77	1.47	58	Pelagibacter	None
78	2.29	36	UBA8592	UBA8592 sp002380145
79	1.37	63	Pelagibacter	None
80	1.97	47	Pelagibacter_A	Pelagibacter_A sp002457475
81	2.87	49	UBA8309	None

Table 3.6: Characteristics of the ten genome bins from the Tara oceans differential abundance analysis with the most evidence of positive differential abundance based on p-values from robust score tests on reduced models with reference set S_{dd}. Estimates are of parameters $\beta_1^j - g_F^p(\beta_1^j : j \in S_{dd})$ with respect to increases of temperature of 5 degrees. Prevalence is the number of the 89 samples in which the genome bin is found. Genus and species represent the taxonomic classification of the genome bin. All genome bins have p-values and q-values less than $10e-5$.

In this analysis, we find that while log fold-difference estimates for the $8,360$ genome bins in this analysis range from -25 to 8, the bins with the most evidence of differential abundance have expected $\approx$ 400-fold decreases in abundance associated with a five degree increase in ocean temperature. These bins represent microbial taxa that have higher expected abundances in cooler water compared to warmer water. Therefore, they might be at risk of being out-competed by other microbial taxa as ocean temperatures rise due to the effects of climate change. Additionally, we find that genome bins that belong to the *Pelagibacter* genus are included in the set of bins with the most evidence for both positive and negative differential abundance with respect to increasing ocean temperature. This is not surprising, as in an analysis of marine metagenomes, Brown et al. [2012] found SAR11 subgroups with strong correlations between abundance and temperature. They found that phylogenetically defined SAR11 subgroups had distinct temperature preferences and that the polar subgroups included genes not shared by the tropical subgroups, suggesting functional differences in these groups. Therefore, it is possible that the genome bins that we have identified as negatively differentially abundant correspond to polar SAR11 subgroups and genome bins that we have identified as positively differentially abundant correspond to tropical SAR11 subgroups.

3.6 Discussion

In this paper, we present a method to perform inference on parameters from the model given in Section 3.2.1 using a robust score test that requires fewer computational resources than the robust score test proposed by Clausen and Willis. We do this by defining a reduced model for each hypothesis test of taxon j that includes a parameter that is equivalent to $\beta_k^j - g_F(\beta_k)$ under requirements on the constraint function g_F, and conclude that we can test the null hypothesis $H_0 : \beta_k^j - g_F(\beta_k) = 0$ with a test constructed with the reduced model. Because the reduced model has pJ' parameters for $J' << J$, these tests can be performed more quickly and use fewer computational resources than tests that use the full model. We show that using robust score tests on reduced models instead of full models does not sacrifice Type I error or power in our simulation settings, while computation is efficient even for a

large number of taxa in our data analysis.

The major disadvantage of our robust score test method compared to the Clausen and Willis method is that we can only estimate and test a parameter for which the constraint function is defined in terms of a small reference set of taxa. We discuss several methods to choose this reference set, and investigate how these different reference set methods lead to similar or different estimation and inference results in simulation and in a small data analysis. We recommend choosing the reference set in a data-driven way in order to define a parameter that closely approximates the log fold-difference of abundance for each taxon across covariate levels, relative to the typical log fold-difference across all taxa in a dataset. However, we acknowledge that this data-driven approach is an instance of double dipping (although it does not invalidate inference in simulation). To address this, we also provide data-driven ways to choose a reference set that use sample splitting and thinning, letting us define separate datasets for reference set selection and for analysis. We compare our methods of performing robust score tests on reduced models and defining a parameter in terms of the reference set S_{dd} in a differential abundance analysis of $8,360$ taxa, and compare estimation and inference results to those from ALDEx2 [Fernandes et al., 2013] and ANCOM-BC2 [Lin and Peddada, 2024].

Although we demonstrate our method on a differential abundance analysis of taxa from metagenomic sequencing data, it could also be used in a wider range of applications. We could use this method to perform differential abundance analyses of other abundance quantities from 'omics data, as long as the modeling assumptions that we make in Section 3.2.1 are reasonable. Additionally, the idea of running robust score tests on reduced models could be applied to multinomial logistic regression on datasets with a large number of categories. This decrease in computational power needed for robust score tests would make it easier to justify using them instead of other types of tests that may be faster but have worse Type I error rate control.

This project has several possible extensions. While we have presented a method in which the time and computational resources required to run a single score remain small as J

increases, we typically want to run J robust score tests in an analysis. As J increases to the hundreds of thousands, regardless of the computational efficiency of a single test, this will at some point become infeasible. We could extend our use of sample splitting or thinning for choosing the reference set to also screen for interesting taxa to test. We could use one subset of the data to run a faster method to screen for potentially differentially abundant taxa, and the other subset of the data to run robust score tests for the categories selected from the first data subset. Having fewer tests to consider when correcting for multiple testing could let us identify more differentially abundant categories, while controlling the false discovery rate. However, this would require investigate to determine whether it would combat the loss of power for a single test that is associated with sample splitting and thinning.

Finally, this chapter demonstrates that choosing a target estimand is an important part of any differential abundance procedure. As described in Section 3.2.1, under our generative model for Y_{ij} from metagenomic sequencing data, log fold-differences in abundance across covariate levels are only identifiable after imposing an identifiability constraint. Choosing a meaningful constraint and correctly interpreting estimation and inference results in terms on this constraint are necessary components to any differential abundance analysis of metagenomic sequencing data. However, other differential abundance methods rarely explicitly state the identifiability constraints they use or assumptions they make, making it difficult for the scientist applying these methods to correctly interpret results that come from these methods. We think that this a major disadvantage of these methods, regardless of how accurately they may be able to estimate and test differential abundance parameters.

The method presented here is available in our open-source R package at `https://github.com/statdivlab/fastEmu`. Code to reproduce all analyses run for this paper is available at `https://github.com/statdivlab/fastEmu_supplementary`.

Chapter 4

A MODEL OF FUNCTIONAL ABUNDANCE FROM METAGENOMIC DATA AND ITS IMPLICATIONS

4.1 Introduction

In the previous chapter, we discussed how differential abundance analysis is a tool for understanding how microbial community composition is associated with host and environmental health. We referenced two microbiome and covariate settings (the human gut microbiome and gastrointestinal disease, and the marine microbiome and climate change) in which differential abundance analyses could provide useful insights into the link between environments and their microbiomes. In both of these settings, we could further explore the link between the microbiome and the environment by considering the functions that can be performed by the microbiome. In the human gut microbiome, the disturbance of the gut microbiome and its associated metabolic functions are correlated with the development of gastrointestinal diseases. In the marine microbiome, the increased warming and acidification of the ocean are associated with changes in the ability of the microbiota to perform important functions, such as carbon fixation [Cavicchioli et al., 2019]. In this chapter, we will consider how to use metagenomic sequencing data to identify molecular functions that differ in abundance (in terms of copies of genes that perform that function) across covariate levels. We will refer to this as functional differential abundance, as distinct from the taxonomic differential abundance that we considered in Chapter 3.

In Section 3.2.1, we considered a taxonomic abundance model, and reviewed the identifiability results of Clausen and Willis [2024] for this model. Clausen and Willis show that under this model, they can identify, estimate, and test a parameter that represents the expected log fold-difference in abundance of a given microbial taxon across covariate levels, relative to

the "typical" expected log fold-difference across taxa, where typical is defined based on an identifiability constraint function. Importantly, this differential abundance estimand is not a function of sample-specific or taxon-specific unknown sequencing effects.

An ideal functional differential abundance estimand would link the abundances of molecular functions in microbial samples (in terms of copies of genes that perform this function in a unit volume) to sample covariates. However, as in the taxonomic abundance setting, we cannot directly observe functional absolute abundances (gene copies per unit volume) from metagenomic sequencing data. Unfortunately, it is not obvious that the results about identifiable parameters in the taxonomic differential abundance model should hold for functional abundance parameters. In the taxonomic model, the taxon-specific sequencing effects, which we will refer to as taxon efficiencies, have a constant effect on observed abundances Y_{ij} across all samples. However, these taxon efficiencies will not have a constant effect on observed abundances for function m, Y_{im}, across all samples. This is because the observed functional abundances Y_{im} are aggregated over Y_{ijm}, the observed abundances of function m within taxa j that can perform that function. Because these taxa have different abundances in each sample, the effect of taxon efficiencies on observed Y_{im} abundances varies across samples.

In order to answer the scientific question "is the molecular function m differentially abundant with respect to covariates?" we would like to consider parameters defined at the function m level and data at the Y_{im} level. We are not interested in drawing conclusions about whether function m within taxon j is differentially abundant with respect to covariates. This is because we care about community-level functional potential, not taxon-level functional potential. However, because we believe that detection efficiencies function on the taxon level, we must define additional parameters at the m and j level and consider data at the Y_{ijm} level. Because of this complication, the functional differential abundance question requires further study.

To our knowledge, the relationship between taxon efficiencies and identifiable functional differential abundance parameters has not been previously studied. Most differential abun-

dance methods have been developed using assumptions about taxonomic abundances, and tested on taxonomic datasets. Two exceptions to this are the methods LEfSe [Segata et al., 2011] and MaAsLin2 [Mallick et al., 2021]. Both of these methods are described as being applicable to differential abundance analyses of both taxa and functions, and the associated papers present analyses of both taxonomic and functional datasets. However, neither of these papers consider how taxon efficiencies might affect the identifiable estimands for functional abundances.

In Section 4.2, we develop a mean model for functional abundance that accounts for the differential effects of taxon efficiencies across samples. We show that a reparameterization of this model for design matrices with a single categorical covariate is equivalent to the taxonomic abundance model, although the identifiable parameters have different interpretations. Under our functional abundance model, this identifiable parameter is a function of both fold-differences in absolute abundances and unknown taxon efficiencies. We argue that we can estimate and test this parameter with the same methods that we use in Chapter 3. In Section 4.3, we develop and apply a simulation framework based on our functional abundance model, which lets us vary the magnitude and structure of taxon efficiencies to investigate their effects on our identifiable differential abundance parameter. We find that estimating and testing this parameter can lead to false discoveries with respect to biological fold-differences, although this can be mitigated by using a conservative effect size threshold. We apply this strategy to a differential abundance analysis of functions from a metagenomic sequencing dataset in Section 4.4, and conclude with a discussion of the implications of this work in Section 4.5.

4.2 Models for taxonomic and functional abundance

4.2.1 A model for taxonomic abundance

In the previous chapter, we considered a model for expected abundances of taxa from metagenomic sequencing data. We review this model here. Let $W_{ij} \geq 0$ represent the total number

of cells of taxon j per unit volume in sample i, for $j \in \{1, \ldots, J\}$ and $i \in \{1, \ldots, n\}$. For each sample i and taxon j, we consider a log linear mean model for W_{ij} of the form,

$$\mathbb{E}[W_{ij}|X_i, \beta_0^j, \beta^j] = e^{\beta_0^j + X_i\beta^j}. \tag{4.1}$$

Here we use slightly different notation from the previous chapter. We write X_i as the length $p-1$ vector of covariate data $X_i = [X_{i1}, \ldots, X_{ip-1}]$, excluding the constant term associated with an intercept. We use β_0^j to denote the intercept for taxon j, and use β^j to denote the vector of remaining $p-1$ parameters for taxon j, $\beta^j = [\beta_1^j, \ldots, \beta_{p-1}^j]$. In metagenomic sequencing data, we observe $Y_{ij} \geq 0$, the count of taxon j in sample i. We assume that Y_{ij} are distorted measurements of W_{ij}, affected by sequencing effects at the sample and taxon level. We will refer to these taxon effects as taxon efficiencies. As proposed by McLaren et al. [2019], we assume that these sequencing effects have multiplicative effects on $\mathbb{E}[Y_{ij}]$ values. These assumptions imply,

$$\mathbb{E}[Y_{ij}|W_{ij}, z_i, \delta_j] = W_{ij}e^{z_i+\delta_j}. \tag{4.2}$$

The mean models given by (4.1) and (4.2) can be combined to create a regression model for observed MGS abundances Y_{ij},

$$\begin{aligned}\mathbb{E}[Y_{ij}|X_i, \beta_0^j, \beta^j, z_i, \delta_j] &= \mathbb{E}[\mathbb{E}[Y_{ij}|W_{ij}, z_i, \delta_j]|X_i, \beta_0^j, \beta^j] &(4.3)\\ &= e^{z_i+\delta_j+\beta_0^j+X_i\beta^j}. &(4.4)\end{aligned}$$

In this model we consider β_0^j and β^j to be biological parameters, because they come from the regression model for $\mathbb{E}[W_{ij}]$, and we consider z_i and δ_j to be sequencing parameters, because they come from the model for $\mathbb{E}[Y_{ij}|W_{ij}]$.

Clausen and Willis [2024] show that the model mean given in (4.4) is only partially identifiable. The parameters δ_j and β_0^j always appear in the model as the sum $\delta_j + \beta_0^j$, and must be considered together as an intercept that includes both biological signal and taxon efficiencies. As suggested by Clausen and Willis, in order to make this model identifiable we impose the identifiability constraint $g(\delta + \beta_0) = 0$ and $g(\beta_k) = 0$ for $k \in \{1, \ldots, p-1\}$

for a function $g(\cdot)$ that is smooth and satisfies $g(\beta_k + \alpha) = g(\beta_k) + \alpha$. The function $g(\cdot)$ determines the interpretation of the parameter β_k^j.

4.2.2 A model for functional abundance

Notation	Meaning
i	sample, with $i \in \{1, \ldots, n\}$
k	column in design matrix X, with $k \in \{1, \ldots, p-1\}$
X_{ik}	covariate data for sample i and covariate k
m	molecular function, with $m \in \{1, \ldots, M\}$
j	taxon, with $j \in \{1, \ldots, J\}$
J_m	all taxa j that include function m
W_{im}	true abundance of function m in sample i
W_{ijm}	true abundance of function m in taxon j in sample i
Y_{im}	observed abundance of function m in sample i
Y_{ijm}	observed abundance of function m in taxon j in sample i
θ_0^m	$\log \mathbb{E}[W_{im}]$ for i such that $X_{ik} = 0$ forall $k \in \{1, \ldots, p-1\}$
θ_k^m	$\log \mathbb{E}[W_{im} \vert X_{ik-} = x_{ik-}, X_{ik} = x+1]) - \log \mathbb{E}[W_{im} \vert X_{ik-} = x_{ik-}, X_{ik} = x])$
θ_0^{jm}	$\log \mathbb{E}[W_{ijm}]$ for i such that $X_{ik} = 0$ forall $k \in \{1, \ldots, p-1\}$
θ_k^{jm}	$\log \mathbb{E}[W_{ijm} \vert X_{ik-} = x_{ik-}, X_{ik} = x+1]) - \log \mathbb{E}[W_{ijm} \vert X_{ik-} = x_{ik-}, X_{ik} = x])$
e^{z_i}	sample specific sequencing effect for sample i
e^{δ_j}	taxon efficiency for taxon j

Table 4.1: Notation used in functional abundance model

In this section, we propose a mean model for expected abundances of molecular functions from metagenomic sequencing data, applying similar assumptions to those used in the taxonomic abundance model. Let $W_{im} \geq 0$ represent the total number of copies of genes that

perform function m per unit volume in sample i, for $m \in \{1, \ldots, M\}$ and $i \in \{1, \ldots, n\}$. As in our taxonomic abundance model, for all samples i and functions m we will consider a log linear regression model for W_{im} of the form,

$$\mathbb{E}[W_{im}] = e^{\theta_0^m + X_i \theta^m}. \tag{4.5}$$

The parameters θ_k^m are the differential abundance parameter of interest in this model. θ_k^m represents the expected log fold-difference in the abundance of function m associated with covariate level k.

We observe counts $Y_{im} \geq 0$ from metagenomic sequencing data, for all samples i and functions m. We assume that Y_{im} is a distorted measurement of W_{im}, disorted by sequencing effects that act on the sample i and a combination of sequencing efficiencies for the taxa in which function m is found in sample i. We can write this as

$$\mathbb{E}[Y_{im} | W_{im}, z_i, \delta_{im}] = W_{im} e^{z_i + \delta_{im}}. \tag{4.6}$$

If we assumed that $\delta_{im} = \delta_m$ for all samples i, then this would be equivalent to (4.2), and we could derive similar equations to (4.3-4.4). This would let us define a model in which the parameter $\theta_k^m - g(\theta_k)$ is identifiable, and we would target this parameter for estimation and inference. We will refer to $\theta_k^m - g(\theta_k)$ as our target biological parameter. However, $\delta_{im} = \delta_m$ for all i is not a reasonable assumption for metagenomic sequencing data.

Therefore, the model given in (4.6) is more complex than the associated taxonomic model given in (4.2) because we would like consider parameters at the functional level, but must account for sequencing efficiencies that work on the taxon level. This is similar to the model for efficiencies for higher-level taxa given by McLaren et al. [2022]. They define the efficiency of a higher-level taxon, such as a genus or phylum, as the abundance-weighted mean of the efficiencies of the species that make up that taxon. We can show that the efficiency for function m in sample i can be defined in a similar way. Define the set $J_m \subseteq \{1, \ldots, J\}$ for each function m as the set of all taxa in which function m could appear in the microbiome of interest. For function m and each $j \in J_m$, define W_{ijm} as the number of copies per unit

volume of genes that perform function m that are found in microbes of taxon j in sample i, such that $W_{im} = \sum_{j \in J_m} W_{ijm}$. We can re-express (4.6) using true abundances W_{ijm}, such that,

$$\mathbb{E}[Y_{im}|W_{im}, z_i, \delta] = e^{z_i} \sum_{j \in J_m} W_{ijm} e^{\delta^j} \tag{4.7}$$

$$= e^{z_i} W_{im} \cdot \frac{\sum_{j \in J_m} e^{\delta_j} W_{ijm}}{\sum_{j \in J_m} W_{ijm}} \tag{4.8}$$

$$\delta_{im} \equiv \frac{\sum_{j \in J_m} e^{\delta_j} W_{ijm}}{\sum_{j \in J_m} W_{ijm}}. \tag{4.9}$$

In a result similar to the efficiency result from McLaren et al. for the higher-level taxa, the efficiency δ_{im} is the abundance-weighted mean of the efficiences of taxa that contain function m.

Equations (4.7-4.9) show that we must consider efficiencies that work on the taxon level. Therefore, we will define a mean model for W_{ijm} abundances. We will assume,

$$\mathbb{E}[W_{ijm}|X_i, \theta^{jm}] = e^{\theta_0^{jm} + X_i \theta^{jm}}. \tag{4.10}$$

For function m we can aggregate over all taxa $j \in J_m$ to get,

$$\mathbb{E}[W_{im}] = \sum_{j \in J_m} \mathbb{E}[W_{ijm}] \tag{4.11}$$

$$= \sum_{j \in J_m} e^{\theta_0^{jm} + X_i \theta^{jm}}. \tag{4.12}$$

Equations (4.5) and (4.12) provide two models for $E[W_{im}]$ at different scales of abundance measurement. We set them equal, which implies,

$$e^{\theta_0^m + X_i \theta^m} = \sum_{j \in J_m} e^{\theta_0^{jm} + X_i \theta^{jm}}. \tag{4.13}$$

This then implies,

$$\theta_0^m = \log \sum_{j \in J_m} e^{\theta_0^{jm}} \tag{4.14}$$

$$X_i \theta^m = \log \frac{\sum_{j \in J_m} e^{\theta_0^{jm} + X_i \theta^{jm}}}{\sum_{j \in J_m} e^{\theta_0^{jm}}}. \tag{4.15}$$

When our design matrix consists of a single categorical covariate, such that $\sum_{k=1}^{p-1} X_{ik} \in \{0, 1\}$ for all $i \in \{1, \ldots, n\}$, we can rewrite our target estimand θ_k^m as,

$$\theta_k^m = \log \frac{\sum_{j \in J_m} e^{\theta_0^{jm} + \theta_k^{jm}}}{\sum_{j \in J_m} e^{\theta_0^{jm}}}, \tag{4.16}$$

for all $k \in \{1, \ldots, p-1\}$.

Now that we have a model for W_{ijm}, we can return to our model for Y_{im}, given in (4.7).

$$\mathbb{E}[Y_{im}|X_i, \theta, z_i, \delta, J_m] = \mathbb{E}[\mathbb{E}[Y_{im}|W_{im}, z_i, \delta]|X_i, \theta] \tag{4.17}$$

$$= e^{z_i} \sum_{j \in J_m} e^{\delta_j + \theta_0^{jm} + X_i \theta^{jm}} \tag{4.18}$$

We will then rearrange terms.

$$\log \mathbb{E}[Y_{im}|X_i, \theta, z_i, \delta, J_m] = z_i + \log \sum_{j \in J_m} e^{\delta_j + \theta_0^{jm}} \tag{4.19}$$

$$= z_i + \log \sum_{j \in J_m} e^{\delta_j + \theta_0^{jm} + X_i \theta^{jm}} + \log \sum_{j \in J_m} e^{\delta_j + \theta_0^{jm}} - \log \sum_{j \in J_m} e^{\delta_j + \theta_0^{jm}} \tag{4.20}$$

$$= z_i + \log \sum_{j \in J_m} e^{\delta_j + \theta_0^{jm}} + \log \frac{\sum_{j \in J_m} e^{\delta_j + \theta_0^{jm} + X_i \theta^{jm}}}{\sum_{j \in J_m} e^{\delta_j + \theta_0^{jm}}} \tag{4.21}$$

$$= z_i + \alpha_m + f_{\delta, \theta, J_m}(X_i), \tag{4.22}$$

for $\alpha_m = \log \sum_{j \in J_m} e^{\delta_j + \theta_0^{jm}}$ and $f_{\delta,\theta,J_m}(X_i) = \log(\sum_{j \in J_m} e^{\delta_j + \theta_0^{jm} + X_i \theta^{jm}} / \sum_{j \in J_m} e^{\delta_j + \theta_0^{jm}})$. Therefore, we have a log-linear model for $\mathbb{E}[Y_{im}]$ in terms of components z_i, α_m, and f_{δ,θ,J_m}, where f_{δ,θ,J_m} is a non-linear function of parameters δ and θ and the data vector $X_i \in \mathbb{R}^{p-1}$. There are several challenges of using this model for $\mathbb{E}[Y_{im}]$. Each covariate vector element X_{ik} for $k \in \{1, \ldots, p-1\}$ appears $|J_m|$ times in the non-linear function f_{δ,θ,J_m}, associated with parameters θ_k^{jm}. Therefore, this model has no natural differential abundance parameter at the covariate k and function m level. If we instead wanted to target parameters θ_k^{jm} at the covariate k, function m, and taxon j scale, we would need more complex identifiability constraints than those needed in the previous chapter. It would also be more challenging to estimate and perform inference on parameters in the model that includes this non-linear

function f_{δ,θ,J_m}, in comparison to estimation and inference for log-linear models that are linear functions of parameters and covariates. In order to avoid these difficulties, we will reduce the class of design matrices X that we consider to only include design matrices with one categorical covariate, such that $\sum_{k=1}^{p-1} X_{ik} \in \{0,1\}$ for all samples i.

Under this restriction,

$$\log \mathbb{E}[Y_{im}|X_i, z_i, \alpha_m, \gamma^m] = z_i + \alpha_m + \sum_{k=1}^{p-1} X_{ik}\gamma_k^m \tag{4.23}$$

$$\gamma_k^m \equiv \log \frac{\sum_{j\in J_m} e^{\delta_j+\theta_0^{jm}+\theta_k^{jm}}}{\sum_{j\in J_m} e^{\delta_j+\theta_0^{jm}}}. \tag{4.24}$$

In (4.23), we have a log linear model for $\mathbb{E}[Y_{im}]$, which is a linear combination of covariates X_{ik} and parameters z_i, α_m, and γ_m. The benefit of this restriction is that this model now has a log fold-difference parameter γ_k^m at the covariate k and function m level that is multiplied by covariate X_{ik}.

Proposition 6. *Parameters of the form $\gamma_k^m - g(\gamma_k)$ are identifiable in model (4.23), for identifiability constraint functions $g(\cdot)$ that are smooth, such that $g(v+a) = g(v) + a$ for $v \in \mathbb{R}^M$ and $a \in \mathbb{R}$.*

Proof: The model given in (4.23), parameterized in terms of z, α, and γ, is equivalent to the taxonomic model given in (4.4), parameterized in terms of z, β_0, and β. Therefore, the same identifiability results as those derived by Clausen and Willis [2024] will apply. This implies that parameters z, α, γ in the mean model in (4.23) are identifiable up to equivalence classes of α and γ, and under appropriate identifiability constraint functions $g(\cdot)$, the parameters $\alpha - g(\alpha)$ and $\gamma_k - g(\gamma)$ for all $k \in \{1,\ldots,p-1\}$ are identifiable. □

A result of this proposition is that we can use methods described in the previous chapter to estimate and perform inference on parameters in this model. This is described further in Section 4.2.3.

Define the set of parameters $\phi = (\delta, \theta_0, \theta_1, \ldots, \theta_{p-1}) \in \mathbb{R}^{J+p\sum_{m=1}^M |J_m|}$, where $\theta_k = (\theta_k^{\cdot 1}, \ldots, \theta_k^{\cdot M})$ with $\theta_k^{\cdot m} = (\theta_k^{jm} : j \in J_m)$ for $k \in \{1,\ldots,p-1\}$. The parameters within the vector ϕ appear

in the mean model given in (4.21), and fully determine the parameters α and γ that appear in the re-parameterized version of the model given in (4.23). Given a constraint function $g(\cdot)$, define the functions $f^{\gamma_k^m}$ and $f^{\theta_k^m}$ as the following:

$$f^{\gamma_k^m}(\phi) = \log \frac{\sum_{j \in J_m} e^{\delta_j + \theta_0^{jm} + \theta_k^{jm}}}{\sum_{j \in J_m} e^{\delta_j + \theta_0^{jm}}} - g\left(\log \frac{\sum_{j \in J.} e^{\delta_j + \theta_0^{j\cdot} + \theta_k^{j\cdot}}}{\sum_{j \in J.} e^{\delta_j + \theta_0^{j\cdot}}} \right) \tag{4.25}$$

$$f^{\theta_k^m}(\phi) = \log \frac{\sum_{j \in J_m} e^{\theta_0^{jm} + \theta_k^{jm}}}{\sum_{j \in J_m} e^{\theta_0^{jm}}} - g\left(\log \frac{\sum_{j \in J.} e^{\theta_0^{j\cdot} + \theta_k^{j\cdot}}}{\sum_{j \in J.} e^{\theta_0^{j\cdot}}} \right). \tag{4.26}$$

Consider the class of parameter vectors $\phi \in \Phi^c \subset \Phi$, where $\Phi \equiv \mathbb{R}^{J + p\sum_{m=1}^{M} |J_m|}$ such that the following assumption holds.

Assumption 1: For all $\phi \in \Phi^c$, for all functions m, one of the following is true:

(a) $\delta_j = c_m \in \mathbb{R}$ for all $j \in J_m$

(b) $\theta_k^{jm} = \theta_k^m$ for all $j \in J_m$ and all $k \in \{1, \ldots, p-1\}$

Claim: For all $\phi \in \Phi^c$, $f^{\gamma_k^m}(\phi) = f^{\theta_k^m}(\phi)$. Therefore, under that restriction that the parameters ϕ in the model in equation (4.21) belong to the set Φ^c, the resulting parameters $\gamma_k^m - g(\gamma_k)$ and $\theta_k^m - g(\theta_k)$ are equivalent.

Proof: Take an arbitrary parameter vector $\phi \in \Phi^c$. First, consider a function m' such that Assumption 1(a) holds.

$$\gamma_k^{m'} = \log \frac{\sum_{j \in J_{m'}} e^{\delta_j + \theta_0^{jm'} + \theta_k^{jm'}}}{\sum_{j \in J_{m'}} e^{\delta_j + \theta_0^{jm'}}} = \log \frac{e^{c_{m'}} \sum_{j \in J_{m'}} e^{\theta_0^{jm'} + \theta_k^{jm'}}}{e^{c_{m'}} \sum_{j \in J_{m'}} e^{\theta_0^{jm'}}} = \theta_k^{m'}$$

Next, consider a function m'' such that Assumption 1(b) holds.

$$\gamma_k^{m''} = \log \frac{\sum_{j \in J_{m''}} e^{\delta_j + \theta_0^{jm''} + \theta_k^{jm''}}}{\sum_{j \in J_{m''}} e^{\delta_j + \theta_0^{jm''}}} = \log \frac{e^{\theta_k^{m''}} \sum_{j \in J_{m''}} e^{\delta_j + \theta_0^{jm''}}}{\sum_{j \in J_{m''}} e^{\delta_j + \theta_0^{jm''}}} = \theta_k^{m''}$$

Therefore, for a set of functions $m \in \{1, \ldots, M\}$ such that each m follows Assumption 1(a) or 1(b), $\gamma_k^m = \theta_k^m$. Therefore, for any constraint function, $g(\gamma_k) = g(\theta_k)$. This means

that $\gamma_k^m - g(\gamma_k) = \theta_k^m - g(\theta_k)$. Because this is true for all $\phi \in \Phi^c$, this means that the functions $f^{\gamma_k^m}(\phi)$ and $f^{\theta_k^m}(\phi)$ are equivalent for all $\phi \in \Phi^c$. $\square$

This means that if we could assume that $\phi \in \Phi^c$ in all biologically plausible scenarios, then the parameters $\gamma_k^m - g(\gamma_k)$ and $\theta_k^m - g(\theta_k)$ would be equivalent, and we could identify the biological fold-difference parameter $\theta_k^m - g(\theta_k)$ in the mean model given in equation (4.23). However, Assumption 1 is not reasonable in most settings. Assumption 1(a) requires all functions to only occur in a set of taxa with the same efficiencies. This would be trivial if $\delta_1 = \ldots = \delta_J$. This assumption is biologically feasible for functions that only occur within a small set of closely related microbes, which are similarly difficult to extract DNA from. Assumption 1(b) requires all functions to only occur in a set of taxa that have the same log fold-differences for that function across covariate levels. This assumption could also hold for functions that appear only in a small subset of taxa, all with similar copy numbers of genes that perform function m and similar taxonomic differential abundances with respect to the covariate. Unfortunately, neither Assumption 1(a) or 1(b) are plausible for all functions $m \in \{1, \ldots, M\}$

Proposition 7. *Given an arbitrary $\phi \in \Phi$, $f^{\gamma_k^m}(\phi)$ and $f^{\theta_k^m}(\phi)$ are not necessarily equal. Therefore, the parameters $\gamma_k^m - g(\gamma_k)$ and $\theta_k^m - g(\theta_k)$ are not equivalent in the general setting that $\phi \in \Phi$.*

Proof: Consider a small example in which $m = 3$ and $J_1 = \{1, 2, 3\}$. We will choose $\phi^t \in \Phi$ such that $\{\delta_1, \delta_2, \delta_3\} = \{\log 2, \log 1, \log 1\}$, $\theta_0^{j1} = \log 1$ for all $j \in J_1$, and $\{\theta_1^{11}, \theta_1^{21}, \theta_1^{31}\} = \{\log 1, \log 4, \log 10\}$.

$$\gamma_1^1 = \frac{2 \cdot 1 \cdot 1 + 1 \cdot 1 \cdot 4 + 1 \cdot 1 \cdot 10}{2 \cdot 1 + 1 \cdot 1 + 1 \cdot 1} = 4$$
$$\theta_1^1 = \frac{1 \cdot 1 + 1 \cdot 4 + 1 \cdot 10}{1 + 1 + 1} = 5$$

In this toy example, $\gamma_1^1 \neq \theta_1^1$. We will assume that in this toy example, we can also choose parameters θ_0^{jm} and θ_1^{jm} for $m \in \{2, 3\}$ so that $\{\gamma_1^1, \gamma_1^2, \gamma_1^3\} = \{4, 3, 3\}$ and $\{\theta_1^1, \theta_1^2, \theta_1^3\} = \{5, 3, 1\}$. Consider several possible identifiability constraint functions $g(\cdot)$.

- $g(\gamma_1) = \gamma_1^2$: $\gamma_k^1 - \gamma_k^2 = 1 \neq 2 = \theta_1^1 - \theta_1^2$
- $g(\gamma_1) = \gamma_1^3$: $\gamma_k^1 - \gamma_k^3 = 1 \neq 4 = \theta_1^1 - \theta_1^3$
- $g(\gamma_1) = \bar{\gamma}_1$: $\gamma_k^1 - \bar{\gamma}_1 = \frac{2}{3} \neq 2 = \theta_1^1 - \bar{\theta}_1$
- $g(\gamma_1) = \text{median}(\gamma)_1$: $\gamma_k^1 - \text{median}(\gamma)_1 = 1 \neq 2 = \theta_1^1 - \text{median}(\theta)_1$

In this toy example, across four different potential identifiability constraints $g(\cdot)$, $f^{\gamma_k^m}(\phi^t) \neq f^{\theta_k^m}(\phi^t)$. Therefore, $\gamma_k^m - g(\gamma_k)$ and $\theta_k^m - g(\theta_k)$ are not equivalent in the general setting of $\phi \in \Phi$. $\square$

Corollary: Because the parameters $\gamma_k^m - g(\gamma_k)$ and $\theta_k^m - g(\theta_k)$ are not equivalent for the general setting that $\phi \in \Phi$, the parameters $\theta_k^m - g(\theta_k)$ do not appear in the model given in equation (4.23), and therefore are not identifiable in this model.

4.2.3 Estimation and inference for the functional abundance model

In the previous section, we state that for the restricted class of design matrices,

$$\log \mathbb{E}[Y_{im}|X_i, z_i, \alpha_m, \gamma_m] = z_i + \alpha_m + \sum_{k=1}^{p-1} X_{ik}\gamma_k^m,$$

for all i and m, where,

$$\alpha_m = \log \sum_{j \in J_m} e^{\delta_j + \theta_0^{jm}}$$
$$\gamma_k^m = \log \frac{\sum_{j \in J_m} e^{\delta_j + \theta_0^{jm} + \theta_k^{jm}}}{\sum_{j \in J_m} e^{\delta_j + \theta_0^{jm}}}.$$

The parameters $\alpha_m - g(\alpha)$ and $\gamma_k^m - g(\gamma_k)$ are identifiable for all $k \in \{1, \ldots, p-1\}$. The form of this model is equivalent to $\log \mathbb{E}[Y_{ij}|X_i, z_i, \delta_j, \beta_0^j, \beta^j] = z_i + (\delta_j + \beta_0^j) + \sum_{k=1}^{p-1} X_{ik}\beta_k^j$ for samples i and taxa j. This is the taxonomic abundance model, with identifiable parameters $\beta_k^j - g(\beta_k)$ for all $k \in \{1, \ldots, p-1\}$. Clausen and Willis [2024] develop an algorithm to estimate penalized MLEs of the parameters in the taxonomic abundance mean model. When

this mean model is true, these penalized MLEs will be consistent for the parameters in this model. They propose an inferential framework which uses a robust score test to test the null hypothesis $H_0 : \beta_j^k - g(\beta_k) = 0$ for each taxon j. They demonstrate that this procedure controls the Type I error rate under a range of data generation settings in simulation. In Chapter 3, we show that our proposed robust score tests using reduced models also control Type I error rate under similar simulation settings. Therefore, we can propose Clausen and Willis' estimation algorithm to estimate parameters $\gamma_1^m - g(\gamma_1)$ from our functional abundance model, and test hypotheses of the form $\gamma_1^m - g(\gamma_1) = 0$ using robust score tests on reduced versions of our functional abundance model. Even though functional abundance datasets tend to be very large, typically including ten thousand functions or more, we have demonstrated that robust score tests with reduced models are computationally efficient for datasets with a large number of categories.

Proposition 8. *Under design matrix restriction that* $\sum_{k=1}^{p-1} X_{ik} \in \{0, 1\}$ *for all samples* $i \in \{1, \ldots, n\}$*, if we use Clausen and Willis' algorithm to estimate parameters* $\hat{\beta}_k^j - g(\hat{\beta}_k)$ *under the alternative, applied to functional abundance data* $Y \in \mathbb{R}^{n \times M}$*, then estimates* $\hat{\beta}_k^m - g(\hat{\beta}) \xrightarrow{p} \gamma_k^m - g(\gamma_k)$.

Proof: Under these design matrix restrictions, we assume a model in which

$$\log \mathbb{E}[Y_{im} | X_i, z_i, \alpha_m, \gamma^m] = z_i + \alpha_m + \sum_{k=1}^{p-1} X_{ik} \gamma_k^m,$$

subject to identifiability restriction $g(\gamma_k) = 0$. This model is equivalent to the mean model that is assumed by Clausen and Willis. Because the mean model holds, the estimated MLEs $\hat{\beta}_k^m - g(\hat{\beta}_k)$ will be consistent for the log-fold difference parameters in the mean model that are associated with covariates X_{ik}: $\gamma_k^m - g(\gamma_k)$. □

In Chapter 3, we needed to address the challenge of choosing a small subset of taxa to serve as a reference set with which to define our identifiability constraint and subsequently our differential abundance parameter. This is difficult for taxa because we rarely have scientific knowledge to guide this reference set selection. However, we can apply scientific knowledge

about core genes in the functional abundance setting. Core genes are defined in terms of a high-level taxon, such as a phylum or a domain, and appear in every constituent species of that taxon. These genes typically encode essential cellular functions and appear in a single copy within each organism that contains them. Therefore, we expect that the functions that correspond to these core genes will have relatively stable abundances across covariate levels for all environmental covariates (the exception being covariates that are associated with an increase or decrease in all taxa in a microbiome). This makes these core functions an optimal set of functions to serve as a reference set. We will propose setting the constraint function $g(\cdot)$ to be the pseudo-Huber loss over the γ_m parameters for functions m in a reference set of core functions, which we will refer to as S_{core}.

In this section we have proposed a model for $\log \mathbb{E}[Y_{im}]$ for functional abundances, and shown that under a restriction of the class of design matrices that we can consider, this model is linear in the parameters z_i, α_m, and γ_m. We've shown that the parameters $\gamma_k^m - g(\gamma_k)$ are identifiable in this model, and we can use the methodology from Clausen and Willis [2024] and Chapter 3 to estimate this parameter and test hypotheses of the form $\gamma_k^m - g(\gamma_k) = 0$. We have also demonstrated that in plausible biological settings, there exist k and m such that $|\gamma_k^m - g(\gamma_k) - \theta_k^m + g(\theta_k)| > 0$.

The parameters $\gamma_k^m - g(\gamma_k)$ and $\theta_k^m - g(\theta_k)$ have similar forms and interpretations. $\theta_k^m - g(\theta_k)$ is the expected log fold-difference in the true abundance (in terms of gene copies per unit volume) of function m between samples in the baseline level and the kth covariate level, relative to the typical log fold-difference in true abundances (where typical is defined based on the constraint function $g(\cdot)$). $\gamma_k^m - g(\gamma_k)$ is the expected log fold-difference in the observed abundance of function m between samples in the baseline level and in the kth covariate levels with the same sample-specific sequencing effects, relative to the typical observed log fold-difference. However, the difference between these parameters and interpretations is very important, because $\theta_k^m - g(\theta_k)$ is only a function of true biological fold-differences, and $\gamma_k^m - g(\gamma)$ is a function of both true biological fold-differences and taxon efficiencies.

If a researcher were to naively assume that the taxonomic abundance mean model could

be applied to functional abundances Y_{im} (with X in our class of restricted design matrices), without concern for taxonomic efficiencies δ_j, they would estimate and test the parameter $\gamma_k^m - g(\gamma_k)$ but interpret it as if it were $\theta_k^m - g(\theta_k)$. In the next section, we will use simulation to study how different these two sets of parameters will be across different taxon efficiency settings, and investigate the consequences of using $\hat{\gamma}_k^m - g(\hat{\gamma}_k)$ estimates and tests of the hypotheses $H_0^m : \gamma_k^m - g(\gamma_k) = 0$ to draw conclusions about the $\theta_k^m - g(\theta_k)$ parameters.

Estimation for the functional abundance model without design matrix restrictions

Previously in this section, we've considered the case in which the design matrix restriction $\sum_{k=1}^{p-1} X_{ik} \in \{0, 1\}$ holds. Now, we will consider what estimates from Clausen and Willis' estimation under the alternative algorithm will converge to when applied to functional abundance data $Y \in \mathbb{R}^{n \times M}$ without this design matrix restriction. As we describe in Section 4.2.2, without this design matrix restriction there is no obvious parameter in terms of the gene model given in equation 4.22 that we would like to estimate. Therefore, we do not expect the mean model used by Clausen and Willis to apply to this data. We will therefore consider this problem through the lens of model misspecification.

In order to present results about the limit of parameter estimates from Clausen and Willis' model, under model misspecification, we will assume that Y are Poisson distributed. We will then assume the true mean model:

$$Y_{ijm} \overset{ind}{\sim} \text{Poisson}(e^{\tilde{z}_i + \delta_j + X_i^T \theta^{jm}}) \tag{4.27}$$

$$g(\theta_k) = 0 \; for \; \theta_k \in \mathbb{R}^{\sum_m |J_m|} \tag{4.28}$$

$$Y_{im} \overset{ind}{\sim} \text{Poisson}(e^{\tilde{z}_i} \sum_{j \in J_m} e^{\delta_j + X_i^T \theta^{jm}}) \tag{4.29}$$

In this model, we specify an identifiability constraint over vectors $\theta_k \in \mathbb{R}^{\sum_m |J_m|}$. This different identifiability constraint, taken over a vector of log fold-differences for function m and taxon j, makes this a fundamentally different model than the one that we assume when we use the design matrix restrictions. However, we can show that the parameters $\theta_k^{jm} - g(\theta_k)$ are identifiable under this restriction, if we have data Y_{ijm}, stratified by taxon.

When we apply Clausen and Willis' mean model, we will use the misspecified model:

$$Y_{im} \overset{ind}{\sim} \text{Poisson}(e^{z_i + X_i^T \beta^j}) \tag{4.30}$$

$$g(\beta_k) = 0 \quad \text{for} \quad \beta_k \in \mathbb{R}^J \tag{4.31}$$

To simplify notation, we will assume single category constraints for both the true and misspecified models. We will use the constraint $g(\theta_k) = \theta_k^{j_1 1}$, where j_1 is the first element of the set J_1, and the constraint $g(\beta_k) = \beta_k^1$.

In order to derive the asymptotic limits of estimates from the misspecified model, we will use results from [White, 1982]. Under regularity conditions, White states that the MLE $\hat{\theta}$ calculated from a misspecified model f, converges in probability to $\theta*$ as $n \to \infty$, where $\theta*$ is the minimum of the Kullback-Leibler (KL) divergence $D_{KL}(g||f)$, where g is the true model. Therefore, the MLEs $(\hat{z}, \hat{\beta})$ estimated using the misspecified model are consistent for parameters (z, β) expressed in terms of parameters from the true model, $(\tilde{z}, \delta, \theta)$ that solve the system of equations of partial derivatives of the KL divergence between the true and misspecified models. This system of equations is the following:

$$\sum_{m=1}^{M} e^{z_i + X_i(\beta^m - \beta_1)} = \sum_{m=1}^{M} e^{\tilde{z}_i + \log \sum_{j \in J_m} e^{\delta_j + X_i(\theta^{jm} - \theta^{j_1 1})}}, \tag{4.32}$$

$$\forall i \in \{1, \ldots, n\} \tag{4.33}$$

$$\sum_{i=1}^{n} \sum_{m=1}^{M} X_{ik} e^{z_i + X_i(\beta^m - \beta^1)} = \sum_{i=1}^{n} \sum_{m=1}^{M} X_{ik} e^{\tilde{z}_i + \log \sum_{j \in J_m} e^{\delta_j + X_i(\theta^{jm} - \theta^{j_1 1})}}, \tag{4.34}$$

$$\forall k \in \{0, \ldots, p-1\} \tag{4.35}$$

$$\sum_{i=1}^{n} e^{z_i + X_i(\beta^m - \beta_1)} = \sum_{i=1}^{n} e^{\tilde{z}_i + \log \sum_{j \in J_m} e^{\delta_j + X_i(\theta^{jm} - \theta^{j_1 1})}}, \tag{4.36}$$

$$\forall k \in \{0, \ldots, p-1\}, \ \forall m \in \{2, \ldots, M\} \tag{4.37}$$

While there is no closed-form solution to this system of equations for arbitrary design matrix X, this shows that the parameters (z, β) that estimates from the misspecified model $(\hat{z}, \hat{\beta})$ converge to, will involve taxon efficiencies δ, and log fold-differences θ_k^{jm} at the function m and taxon j level.

4.3 Simulations

In equation (4.22) in Section 4.2.2, we showed that under our model for functional abundance with design matrices with a single categorical covariate, the $\gamma_k^m - g(\gamma_k)$ parameter that we are able to identify, estimate, and test is a function of both biological and sequencing parameters. In this section, we will investigate how the magnitudes and structures of taxon efficiencies δ affect this $\gamma_k^m - g(\gamma_k)$ parameter and its relationship with our ideal biological estimand $\theta_k^m - g(\theta_k)$.

4.3.1 Parameter generation

In our simulations, we consider sample sizes $n \in \{25, 50, 250\}$ and use a design matrix with a single binary covariate. We generate data for $M = 1000$ functions that occur in $J = 100$ taxa. We consider the first 25 functions M to be core functions, and assume that they appear in all taxa $\{1, \ldots, J\}$. These core functions will be used to define the identifiability constraint $g_p(\gamma_1^m | m \in \{1, \ldots, 25\})$, for the pseudo-Huber loss function $g_p(\cdot)$. We assume that the remaining 975 functions each appear in 25% of the J taxa. We assign these functions to taxa in increasing order, such that function 26 appears in taxa $\{1, \ldots, 25\}$, function 27 appears in taxa $\{2, \ldots, 26\}$, and so on, such that function 250 appears in taxa $\{75, \ldots, 99\}$. This pattern then restarts again with function 251 appearing in taxa $\{1, \ldots, 25\}$, and continues for all remaining functions. This method of assigning functions to taxa ensures that there are sets of taxa that can perform similar sets of functions. This reflects the biological assumption that there exist groups of molecular functions that are performed in conjunction with each other by sets of similar microbes. Additionally, this gives us the ability to assign similar efficiencies to taxa that perform similar functions.

We consider nine taxon efficiency settings. In the first setting, all efficiencies are set to zero. Under this setting, the identifiable $\gamma_1^m - g(\gamma_1)$ parameter is equal to the ideal biological estimand $\theta_1^m - g(\theta_1)$. While this efficiency setting is implausible with current metagenomic sequencing technology, it serves as a baseline with which to compare other efficiency settings.

The remaining settings have efficiencies generated from the set of J evenly spaced values between $\exp(-f)$ and $\exp(f)$, where $f \in \{0.5, 1.15, 1.5, 2.0\}$. These values of f correspond with efficiencies for which the fold-difference from the taxon with the highest efficiency to the taxon with the lowest efficiency are approximately 3-, 10-, 20-, and 50-fold respectively. For each value of f, we construct one setting in which the f values are applied sequencially to taxa $\{1, \ldots, J\}$ such that taxon 1 has efficiency $\exp(-f)$ and taxon J has efficiency $\exp(f)$. We refer to this as the ordered setting. In conjunction with our method for assigning functions to taxa, this implies that sets of taxa that perform similar functions also have similar efficiencies. In the other setting for each f value, we randomly assign efficiency values in the set defined by f to taxa, which produces no relationship between the efficiencies of two taxa and the functions that they perform. We refer to this as the unordered setting. These efficiency settings are summarized in Table 4.2.

Setting	Meaning
`S1`	0-fold difference
`S2`	3-fold difference, ordered
`S3`	3-fold difference, unordered
`S4`	10-fold difference, ordered
`S5`	10-fold difference, unordered
`S6`	20-fold difference, ordered
`S7`	20-fold difference, unordered
`S8`	50-fold difference, ordered
`S9`	50-fold difference, unordered

Table 4.2: Descriptions of nine efficiency settings used in simulations.

While it is difficult to quantify how much taxon efficiencies range in metagenomic sequencing data, studies that use mock communities can provide some insight. Mock communities

are artificially constructed samples with known composition, that are then sequenced and processed in the same way as all other samples in a metagenomic sequencing dataset. In a reanalysis of a vaginal microbiome dataset with mock community measurements for seven taxa, McLaren et al. [2019] found a 30-fold-difference between the taxa with the highest and lowest efficiencies. In a study of fungal microbes, Leopold and Busby [2020] utilized a mock community with eight fungal commensal species and one fungal pathogen species, and estimated the maximum fold-difference between pairs of commensal species to be 13 and the maximum fold-difference of commensal species with the pathogen species to be 40. Therefore, taxon efficiencies that range between 3- and 50-fold seem biologically plausible.

Our parameter generation process for the α_m and γ_1^m parameters is described in detail in Appendix C.1. These parameters are generated such that $g(\alpha) = 0$, $g(\gamma_1) = 0$, and $g(\theta_1) = 0$. The $\theta_1^m - g(\theta_1)$ parameters take on values in the range $[-5, 5]$. We set the reference set to be functions $m \in \{1, \ldots, 25\}$, and define $g(\cdot)$ to be the pseudo-Huber loss over γ_m parameters in that set. We draw the z parameters independently from a Normal$(3, 1)$ distribution. While the assignment of the $\theta_1^m - g(\theta_1)$ parameters is deterministic, the generation of the α_m and γ_1^m parameters are random, based on randomly generated θ_0^{jm} and θ_1^{jm} parameters, as well as randomly assigned δ_j parameters in the unordered settings. For all simulations we generate random parameters for each trial, in order to investigate the behavior of parameters $\gamma_1^m - g(\gamma_1)$ and estimates $\hat{\gamma}_1^m - g(\hat{\gamma}_1)$ over a wide range of the scenarios that are possible within our parameter generation framework. We do this to avoid drawing conclusions that are true for a specific realization of parameters but not true across parameter realizations.

Before we simulate data, can study how the different taxon efficiency settings lead to differences between the ideal biological estimand $\theta_1^m - g(\theta_1)$ and the identifiable parameter $\gamma_1^m - g(\gamma_1)$ for a function m. We generate parameters 100 times, with 100 different random seeds. For each set of parameters, we calculate the difference $|\theta_1^m - g(\theta) - \gamma_1^m + g(\gamma_1)|$ for each $m \in \{1, \ldots, 1000\}$. Figure 4.1 presents boxplots of these differences for each efficiency setting, aggregated over the M parameters and 100 seeds. The parameter differences are larger for efficiency settings with larger fold-differences and for unordered settings. In all

settings, the lowest 75% of parameter differences $|\theta_1^m - g(\theta) - \gamma_1^m + g(\gamma_1)|$ are less than 0.75, which corresponds to a fold-difference of 2.12. However, outlying parameter differences exist for all settings besides `S1`, and the most extreme have magnitudes of 1.9, 2.5, and 3.1, which correspond to fold-differences of approximately 7, 12, and 22 for `S5`, `S7`, and `S9`. This simulation shows that for most parameters within most efficiency settings, the difference $|\theta_1^m - g(\theta) - \gamma_1^m + g(\gamma_1)|$ is small relative to the range of $\theta_1^m - g(\theta_1)$ parameters, but there are a number of outlying parameters with large differences. These large differences, induced by taxon efficiencies, could make $\gamma_1^m - g(\gamma_1)$ appear large when $\theta_1^m - g(\theta)$ is equal to zero, make functions with large values of $\theta_1^m - g(\theta_1)$ have small values of $\gamma_1^m - g(\gamma_1)$, or make $\gamma_1^m - g(\gamma_1)$ have the opposite sign from $\theta_1^m - g(\theta_1)$. We will investigate these concerns through simulation in the remainder of this section.

4.3.2 Data generation

We generate data in the same way as in Chapter 3. Using our design matrix X and generated parameters z, α, and γ, we can compute means of the form $\mu_{im} = e^{z_i + \alpha_m + X_i(\gamma_m - g(\gamma_m))}$ for each $i \in \{1, \ldots, n\}$ and $m \in \{1, \ldots, M\}$. We draw each Y_{im} value independently from either a Poisson distribution of a zero-inflated negative binomial (ZINB) distribution. The simulated ZINB data has approximately 40% zeroes in each dataset, which is similarly sparse to the functional abundance datasets that we analyze in Section 4.4.

4.3.3 Type I error simulations

In Section 4.2.3, we state that we can use robust score tests to test the null hypothesis $H_0^{m(\gamma)} : \gamma_1^m - g(\gamma_1)$, where $g(\cdot)$ is the pseudo-Huber loss over γ_m parameters for a small set of core functions. However, this null hypothesis is less biologically relevant than $H_0^{m(\theta)} : \theta_1^m - g(\theta_1) = 0$, because γ is a function of both biological and sequencing parameters but θ_1 is a function of only biological parameters. Therefore, when we consider inference for Type I error rate simulations, we will generate data according to null hypotheses that are functions of the biological parameters, $\theta_1^m - g(\theta_1)$.

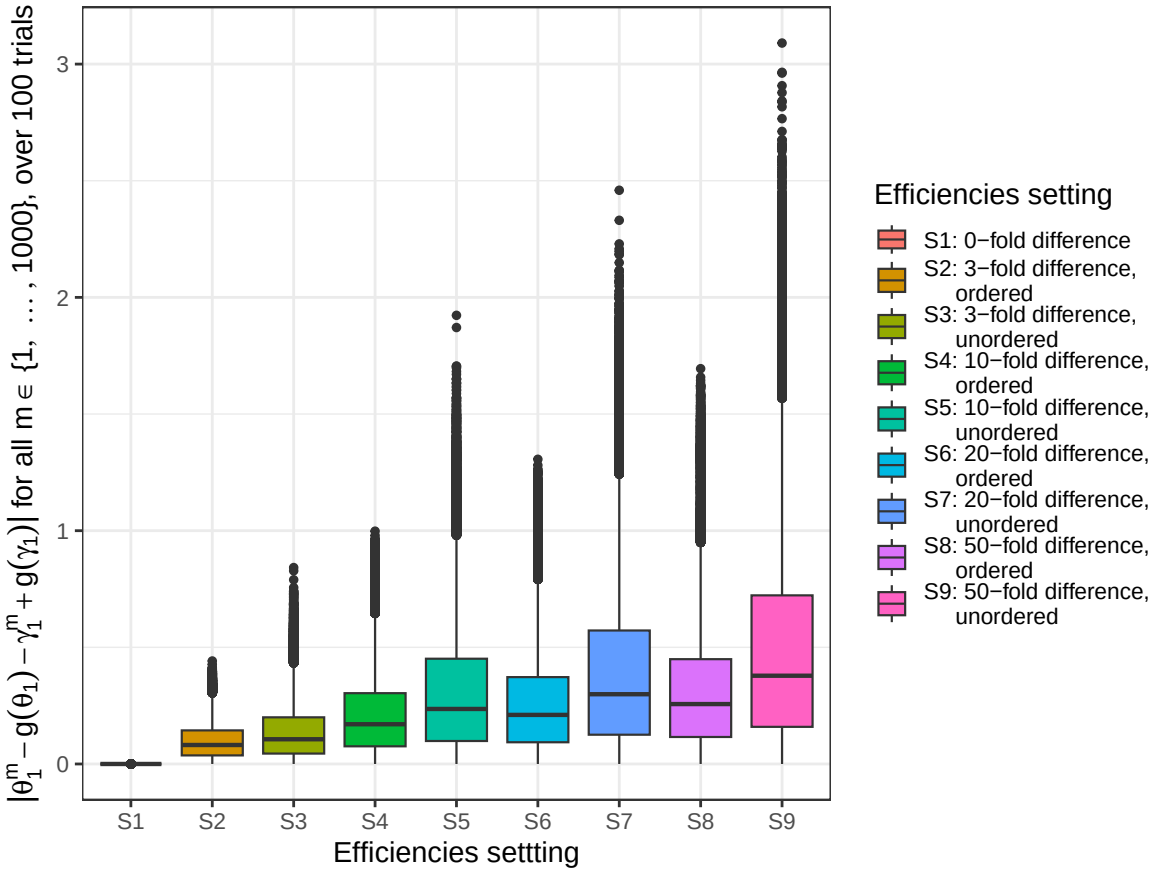


Figure 4.1: Boxplots of simulated differences $|\theta_1^m - g(\theta_1) - \gamma_1^m + g(\gamma_1)|$ between the ideal biological parameter and the identifiable parameter, with parameters generated based on our proposed functional abundance model. Parameters are generated with 100 different random seeds for each taxon efficiency setting. Each boxplot represents one efficiency setting and aggregated parameter differences over M parameters and 100 random seeds.

We consider this to be an investigation of model misspecification in this setting. In our Type I error simulations, we will generate data under the null hypothesis $H_0^{m(\theta)}$. For our robust score tests, we will estimate $\hat{\gamma}_{H_0^{m(\gamma)}}$ such that $g(\hat{\gamma}_{H_0^{m(\gamma)}}) = 0$ under the null hypothesis $H_0^{m(\gamma)}$, and use the $\hat{\gamma}_{H_0^{m(\gamma)}}$ estimates to construct the robust score test statistic $T_{RS}^{m(\gamma)}$. When $\gamma_1^m - g(\gamma_1) = \theta_1^m - g(\theta_1)$, then $T_{RS}^{m(\gamma)} \xrightarrow{\mathcal{D}} \chi_1^2$ as $n \to \infty$. However, when $\gamma_1^m - g(\gamma_1) \neq \theta_1^m - g(\theta_1)$, and we construct score statistic $T_{RS}^{m(\gamma)}$ using data generated under the null hypothesis $H_0^{m(\theta)}$, then we have no asymptotic guarantees for the distribution of $T_{RS}^{m(\gamma)}$ statistics. In this simulation, we will investigate the empirical distribution of the $T_{RS}^{m(\gamma)}$ for finite sample sizes $n \in \{25, 50, 250\}$, for data in which $H_0^{m(\theta)} : \theta_1^m - g(\theta_1) = 0$ is true.

Figure 4.2 presents boxplots of $|\hat{\gamma}_{26} - g(\hat{\gamma})|$ estimates from data simulated across 500 trials. Across all sample size and distribution combinations, the patterns seen in Figure 4.1 are repeated. In all settings, the lowest 75% of estimate magnitudes $|\hat{\gamma}_1^m + g(\hat{\gamma}_1)|$ are less than 0.98, which corresponds to a 2.66-fold difference. The estimate magnitudes are more variable for settings in which efficiencies have larger fold-differences and are unordered. Additionally, the estimate magnitudes are more variable for ZINB data and smaller sample sizes than for Poisson data and larger sample sizes. The estimate with the largest magnitude of 3.30 is from ZINB data with $n = 25$ under efficiency setting eight, and corresponds with a fold-difference of approximately 27.

Figure 4.3 presents quantiles of p-values from the robust score test of the null hypothesis $H_0^{26(\gamma)} : \gamma_{26} - g(\gamma)$ on data simulated under the null hypothesis $H_0^{26(\theta)} = \theta_1^{26} - g(\theta_1) = 0$, compared to quantiles of a Uniform$(0, 1)$ distribution. The tests on data generated under `S1` control the Type I error rate, as well as tests on ZINB data with $n = 25$ and efficiencies with 3-fold-differences. This makes sense, because `S1` is the setting in which $\gamma_1^m - g(\gamma_1) = \theta_1^m - g(\theta_1)$ for all m, and tests of ZINB data with $n = 25$ typically don't have enough power to reject the null hypothesis for small effect sizes. Tests on data generated under all other settings fail to control the Type I error rate. This is especially drastic for Poisson data with larger sample sizes and and for efficiency settings with larger fold-differences. In the most extreme setting, p-values from tests on data generated from Poisson data with $n = 250$ and `S9` are

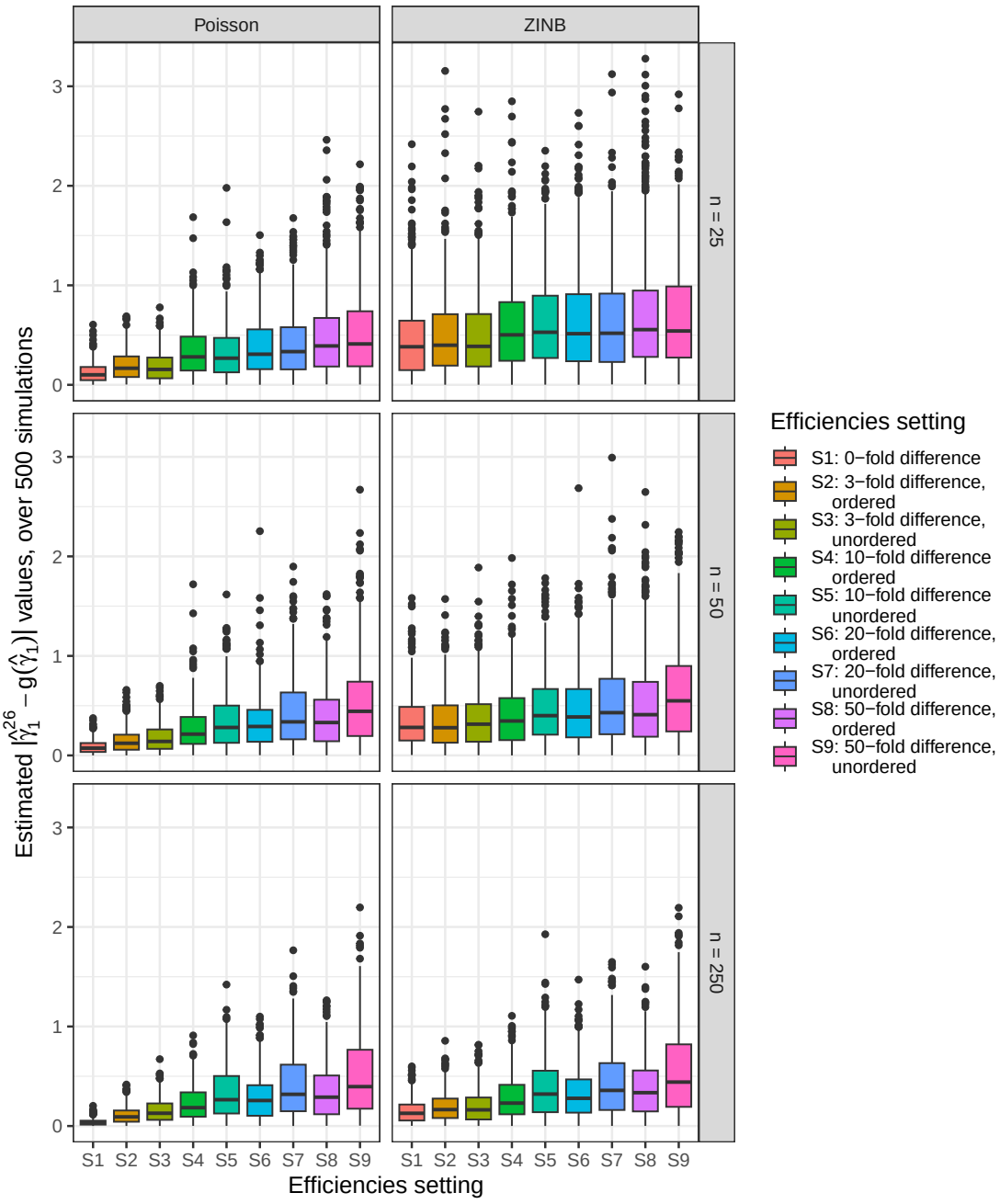


Figure 4.2: Boxplots of $\hat{\gamma}_{26} - g(\hat{\gamma})$ estimates from data simulated from our proposed functional abundance model under the null hypothesis that $\theta_1^{26} - g(\theta_1) = 0$, separated by simulation setting. Results are from 500 randomly generated sets of parameters and data. There are 54 simulation settings that correspond with nine taxon efficiency settings, $n \in \{25, 50, 250\}$, and data generated from Poisson and ZINB models.

less than 0.05 in 86% of trials.

These figures show the magnitude of the $\hat{\gamma}_{26} - g(\hat{\gamma})$ estimates are larger for efficiencies with larger magnitudes that are unordered with respect to taxa, ZINB data, and smaller sample sizes. The robust score test on the reduced model fails to control Type I error rate for the null hypothesis $H_0^{26(\gamma)} : \gamma_1^{26} - g(\gamma_1) = 0$ when performed on data generated under the null hypothesis $H_0^{26(\theta)} : \theta_1^{26} - g(\theta_1) = 0$ in nearly every efficiency setting in which the efficiencies are non-zero. This loss of Type I error rate control is most extreme for efficiencies with larger magnitudes that are unordered with respect to taxa, Poisson data, and larger magnitudes. This suggests that for large sample sizes, functions m for which $\theta_1^m - g(\theta_1) = 0$ will frequently have robust score test p-values less than 0.05, but will nearly always have estimates with magnitudes that are less than 2, which corresponds with a 7-fold-difference. Therefore, in data analyses it would be prudent to identify differentially abundant functions as those that have both small p-values and large estimated effect sizes, in order to avoid false positives.

4.3.4 False discovery rate simulations

In the Type I error rate simulations, we only run tests of the parameter $\gamma_1^{26} - g(\gamma)$ for each trial. However, when we run a differential abundance analysis, we typically estimate and test parameters for every category in the dataset. When we consider a differential abundance analysis of molecular function, this will often result in running approximately ten thousand tests. Multiple testing procedures are typically used to control the false discovery rate over the entire set of tests run within the analysis. Common procedures are the Benjamini-Hochberg (BH) method [Benjamini and Hochberg, 1995] and the q-value method [Storey, 2002]. In our simulations and data analyses we use q-values, due to their potential gain in power over the BH method, especially when there are a substantial proportion of non-null signals in the analysis.

In this simulation, we aim to investigate the effects of different efficiency settings on the conclusions from a functional differential abundance analysis. We do this by testing null

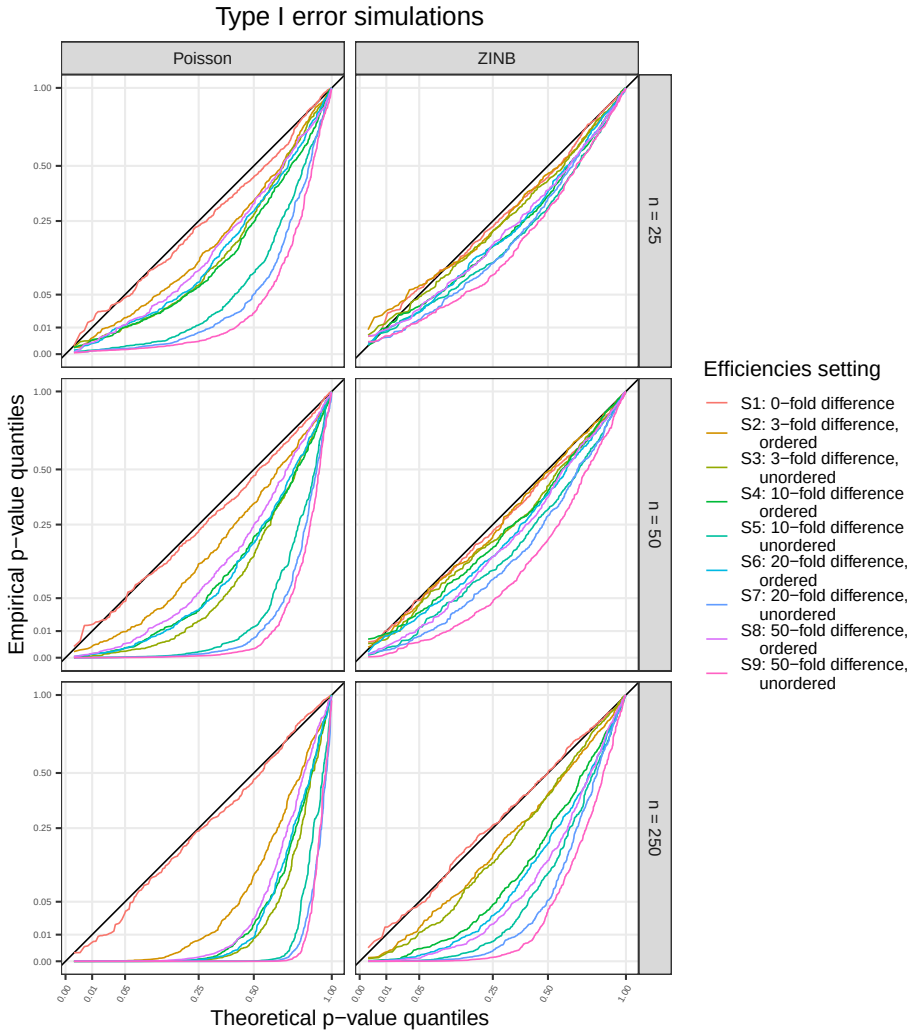


Figure 4.3: Quantiles of p-values obtained from the Type I error rate simulation compared to quantiles of a Uniform(0, 1) distribution. Data are simulated under the null hypothesis that $\theta_1^{26} - g(\theta_1) = 0$ using the functional abundance mean model and draws from a Poisson or zero-inflated Negative Binomial (ZINB) distribution. P-values are generated from robust score tests. The $x = y$ line is shown in black, and represents quantiles of p-values from a test that controls Type I error rate at a nominal rate across the full range of p-values. Tests corresponding to lines above the $x = y$ line are conservative and control the Type I error rate and tests corresponding to lines below the $x = y$ line are anticonservative and fail to control the Type I error rate. Results come from a simulation with 500 trials.

hypotheses $H_0^{m(\gamma)} : \gamma_1^m - g(\gamma) = 0$ for all $m \in \{1, \ldots, 1000\}$, using the set of M robust score test p-values to generate a set of q-values, and drawing conclusions based on a q-value threshold. We generate parameters and data in nearly the same way as described in Sections 4.3.1 and 4.3.2. We simplify parameter generation by setting $\theta_1^m = 0$ for $m \in \{1, \ldots, 25\}$, the categories that represent core functions. We set all other θ_1^m parameters to values in $\{-5, -1, 0, 1, 5\}$, with $\{50, 150, 575, 150, 50\}$ of the remaining parameters randomly assigned to each signal magnitude respectively. This lets us easily classify functions as those that are null, with $\theta_1^m = 0$, those that have low biological signal, with $|\theta_1^m| = 1$, and those that have high biological signal, with $|\theta_1^m = 5|$. We run 10 trials for each simulation setting.

The first thing we investigate in this simulation is how many estimates $\hat{\gamma}_1^m - g(\hat{\gamma}_1)$ have different signs than the associated parameter $\theta_1^m - g(\theta_1)$ for values of $|\theta_1^m - g(\theta_1)| \in \{1, 5\}$. We find that when $|\theta_1^m - g(\theta_1)| = 5$, the estimates of γ parameters always have the same sign as the underlying θ parameters. However, when $|\theta_1^m - g(\theta_1)| = 1$, sign switching does occur. This happens more often for ZINB data than for Poisson data and for efficiency settings with larger magnitudes. The setting with the highest proportion of sign switching for functions m such that $|\theta_1^m - g(\theta_1)| = 1$ is ZINB data with $n = 25$ and taxon efficiencies that are ordered with a 50-fold range. Sign switching happens for 13% of estimates for this setting, aggregated over functions m and the 10 trials.

Figure 4.4 shows proportions of false discoveries and proportions of signals detected, separated by simulation settings and aggregated over the 10 trials for each setting. In this figure, discoveries are defined as functions with q-values less than 0.05 and false discoveries are defined as discoveries with $|\theta_1^m - g(\theta_1)| = 0$. In efficiency setting one, in which all efficiencies are equal, the false discovery rate is controlled at nearly a nominal rate across all distribution and sample size settings. For all other efficiency settings, the false discovery rate is not controlled in any setting, with the exception of data generated with $n = 25$ from the ZINB distribution. The highest proportion of false discoveries comes from the simulation setting with Poisson data with $n = 250$ and efficiencies that are unordered and have a 50-fold-difference. In this simulation setting, aggregated over the 10 trials, 59% of discoveries

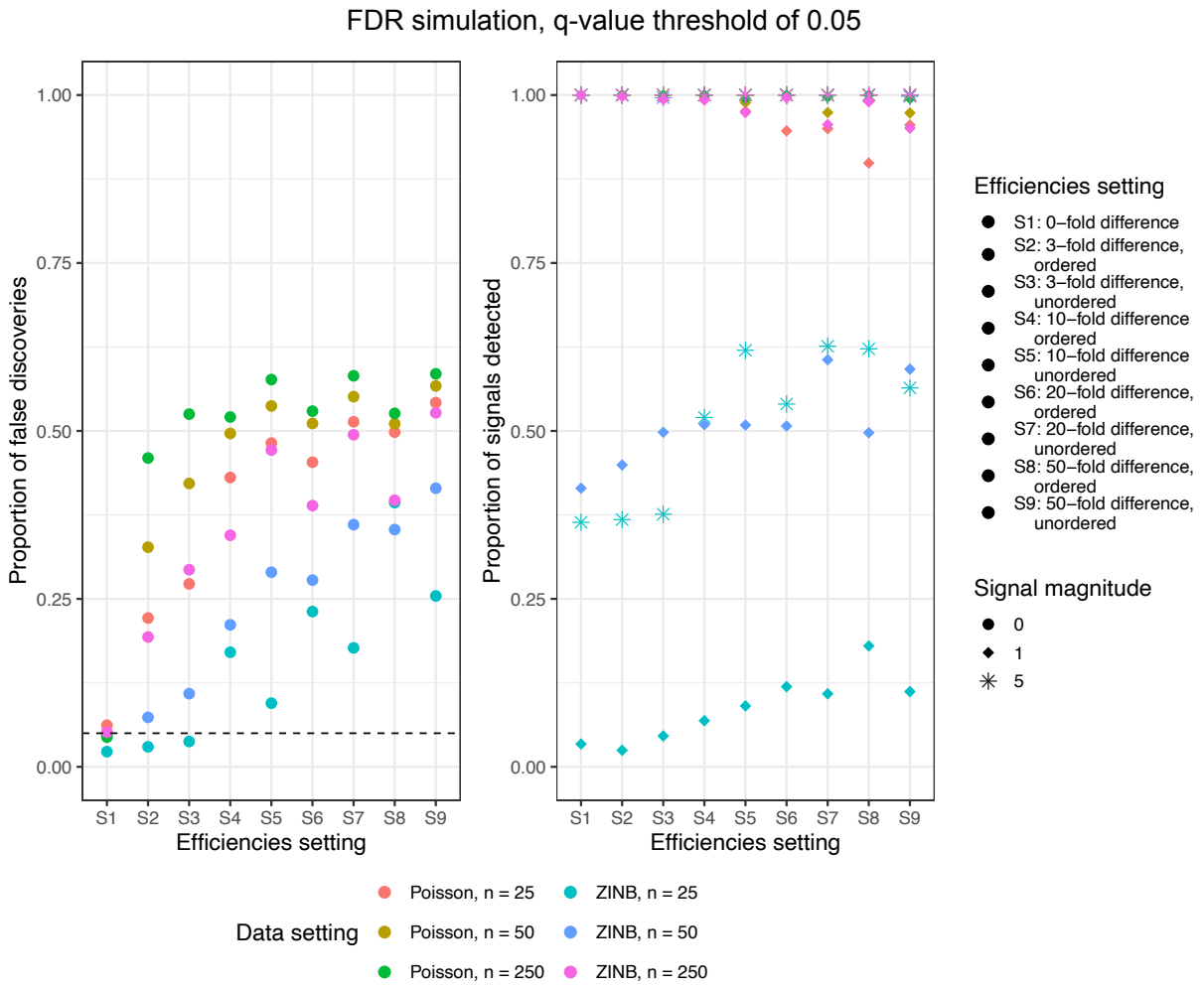


Figure 4.4: Proportions of false discoveries from false discovery rate simulations, separated by simulation setting. Discoveries are defined as functions with q-values less than 0.05, and false discoveries are defined as discoveries with $|\theta_1^m - g(\theta_1)| = 0$. The proportion of false discoveries for each setting, aggregated over the 10 trials, are shown in this figure.

are false discoveries.

There is a high proportion of true signals (functions with $|theta_1^m - g(\theta_1) > 0)$ detected across settings, with the expection of ZINB data with $n \in \{25, 50\}$. In these settings, there is lower power for signals with lower magnitudes of 1. While the proportion of true signals detected is approximately 1 for the other data settings for efficiencies that range up to 10-fold, this proportion drops for efficiencies with larger ranges for Poisson data with $n \in \{25, 50\}$. This simulation suggests that under our parameter and data generation procedures, differences in efficiencies can lead to data analyses in which some or most discoveries are false, even when using methods such as q-values that asymptotically control the false discovery rate for valid tests. However, there are a high proportion of true signals detected in most data settings, even for large efficiency ranges.

One way to try to reduce false discoveries could be to use a more stringent false discovery rate threshold, in hopes that the discoveries with the smallest q-values will be less likely to be false. In Figure 4.5, we use a q-value threshold of $1e-4$ to define discoveries. In this plot, we can see that we avoid false discoveries for ZINB data with $n \in \{25, 50\}$, although in these settings we also fail to detect some to most true signals. In the other data settings, there is still a large proportion of false discoveries. For Poisson data with $n = 250$, the false discovery proportion is over 50% for unordered efficiencies that range 10-fold or more. The proportions of true signals detected are lower than when using a q-value threshold of 0.05, although they are still quite high data-settings with $n = 250$ and signals with large magnitude of 5.

In Section 4.3.3, we found that despite anticonservative tests for most efficiency settings, functions such that $\gamma_1^m - g(\gamma_1) = 0$ rarely had estimated effect sizes with magnitudes greater than 3. Most estimated effect sizes were less than 1. Therefore, we consider the effects of defining a discovery using both a q-value threshold of 0.05 and an estimated effect size threshold of $|\hat{\gamma}_1^m - g(\hat{\gamma}_1)| > 3$. In Figure 4.6, we can see that the false discovery proportions are less than 0.05 in all settings. This discovery definition is quite conservative, as for most data and efficiency settings, the proportion of false discoveries is far closer to 0 than to

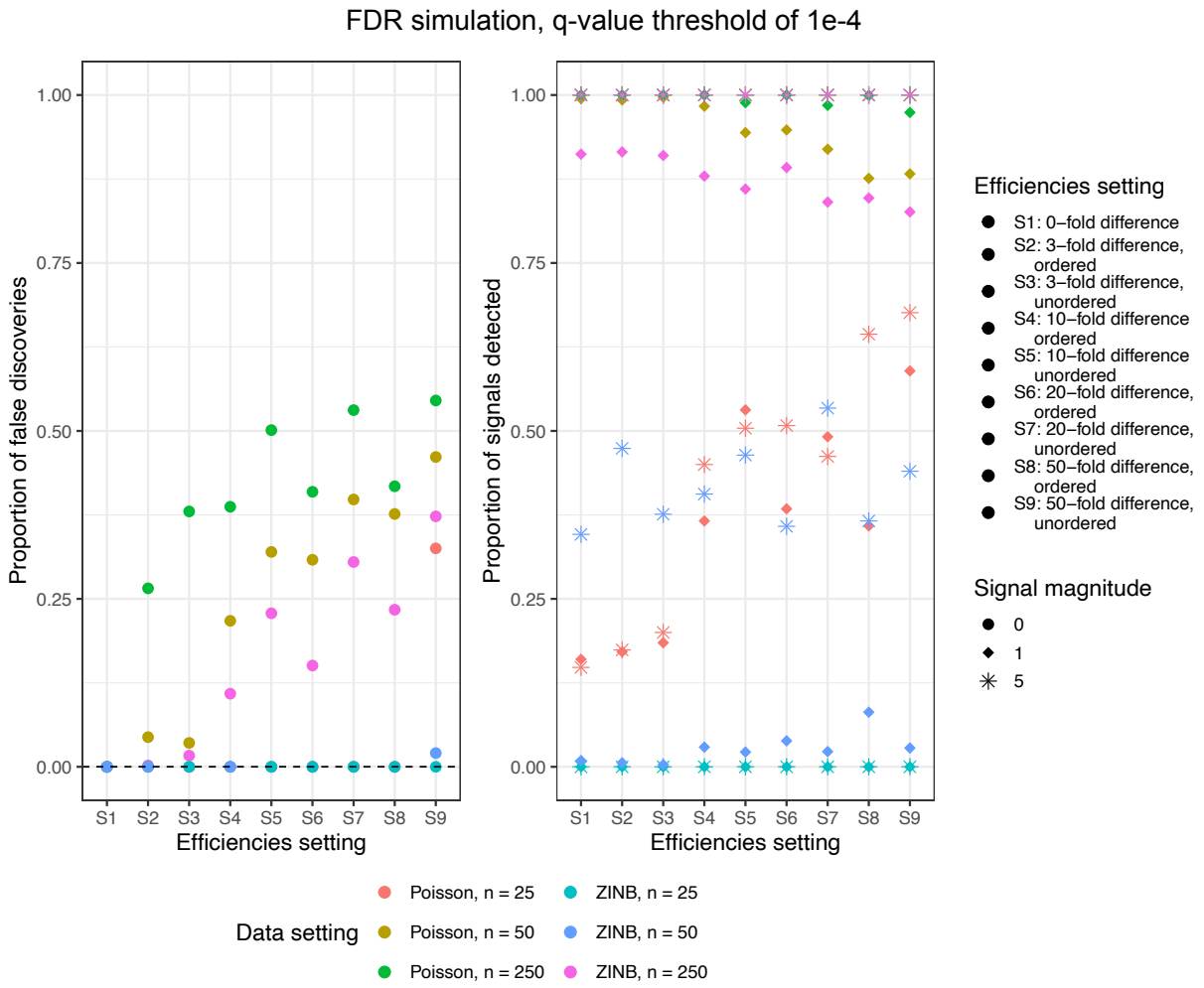


Figure 4.5: Proportions of false discoveries from false discovery rate simulations, separated by simulation setting. Discoveries are defined as functions with q-values less than $1e-4$, and false discoveries are defined as discoveries with $|\theta_1^m - g(\theta_1)| = 0$. The proportion of false discoveries for each setting, aggregated over the 10 trials, are shown in this figure.

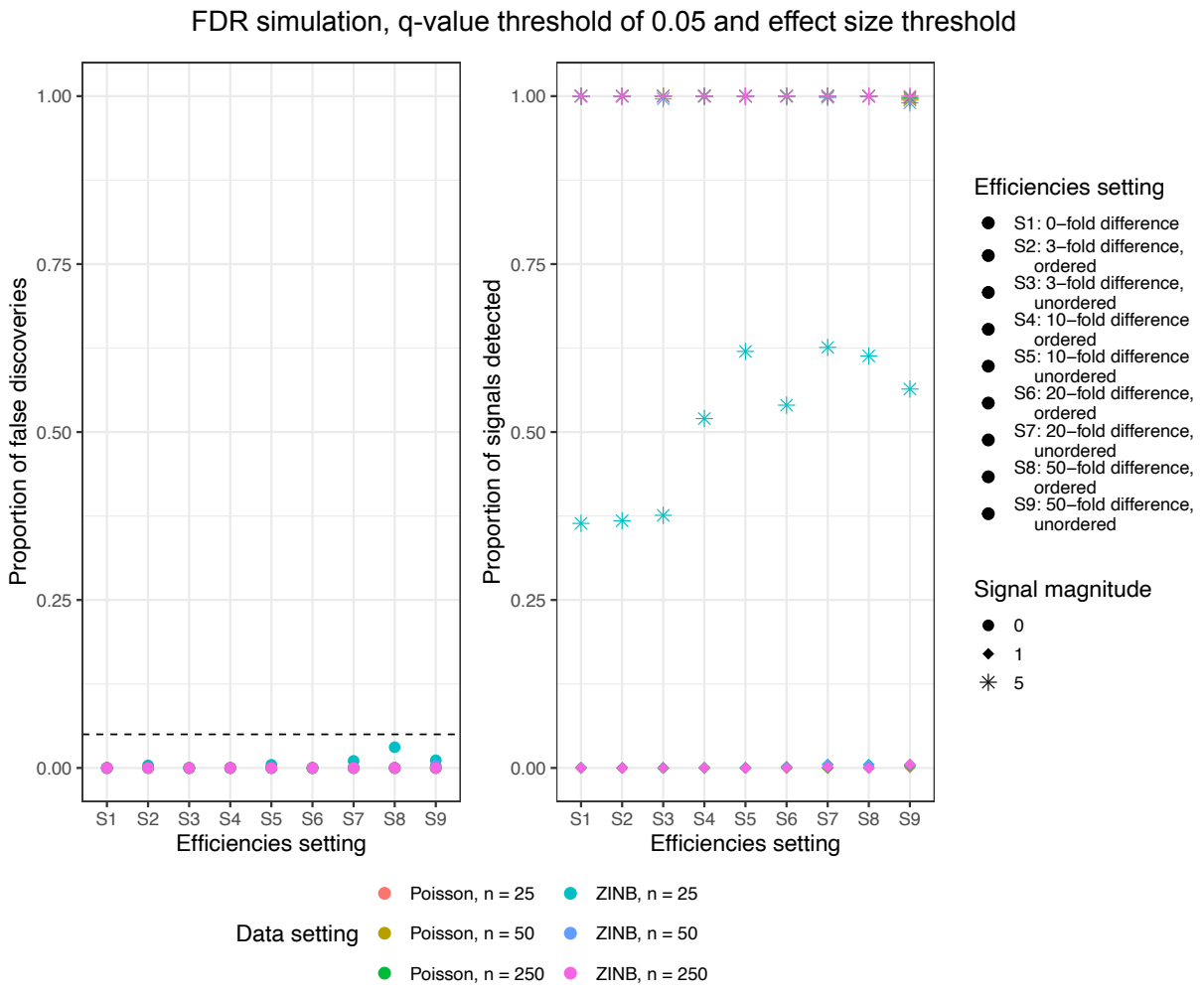


Figure 4.6: Proportions of false discoveries from false discovery rate simulations, separated by simulation setting. Discoveries are defined as functions with q-values less than 0.05 and estimated effect sizes $|\hat{\gamma}_1^m - g(\hat{\gamma}_1)| > 3$, and false discoveries are defined as discoveries with $|\theta_1^m - g(\theta_1)| = 0$. The proportion of false discoveries for each setting, aggregated over the 10 trials, are shown in this figure.

0.05. Using this effect size threshold, we can never detect a signal with a small magnitude of 1. However, we detect nearly 100% of signals with large magnitudes of 5 in all data and efficiency settings except for ZINB data with $n = 25$, for which there is not enough power to detect a high proportion of true signals. These results show that using q-value and effect size thresholds in tandem produces conservative inference, in which we have low rates of false discoveries even in the presence of large efficiencies, but can only detect large signals.

In practice, we do not recommend using a strict effect size threshold for inference. This is because in our simulation, we have the prior knowledge to choose a reasonable effect size threshold, based on our knowledge of the differences between estimates $\hat{\gamma}_1^m - g(\hat{\gamma}_1)$ and parameter $\theta_1^m - g(\theta_1)$ from our Type I error rate simulation. However, our simulation cannot fully capture the complicated dynamics of how functions appear in taxa and we do not know how much taxon efficiencies range in specific microbial sequencing datasets. Therefore, we have no strategies to choose a reasonable effect size threshold. However, Figure 4.6 does show that we can be more confident that discoveries that we make reflect true biological signals in terms of parameters $\theta_1^m - g(\theta_1)$ for functions that have small q-values and large estimated effect sizes.

4.4 Data analyses

In this section, we perform a functional differential abundance analysis on a metagenomic sequencing dataset using Clausen and Willis' estimation procedure and the robust score tests on reduced models that we propose in Chapter 3. We believe that the Y_{im} values in this dataset follow the functional abundance mean model, and we can therefore only estimate and test parameters of the form $\gamma_1^m - g(\gamma_1)$. We will therefore make sure to interpret results based on the $\gamma_1^m - g(\gamma_1)$ parameters, that are effected by taxon efficiencies from sequencing, and not the log fold-differences $\theta_1^m - g(\theta_1)$ that represent biological differential abundance. Therefore, we will not know whether any differential abundance signals that we identify are the results of functions with non-zero $\theta_1^m - g(\theta_1)$ parameters, or the results of differential taxon efficiencies.

We also analyze this dataset using ALDEx2 [Fernandes et al., 2013] and ANCOM-BC2 [Lin and Peddada, 2024]. Although we do not derive the estimands of ALDEx2 and ANCOM-BC2 when applied to functional abundance data, their models are also unable to account for taxon efficiencies that affect the data at a different scale than the functional abundances parameters that we aim to estimate and test. Additionally, under taxonomic abundance models, their target estimands are similar to the $\beta_k^j - g(\beta_k)$ estimand from our taxonomic abundance model. Therefore, when applied to functional abundance data, we expect that both the ALDEx2 and ANCOM-BC2 estimands will be functions of both biological and sequencing parameters.

In this data analysis, we observe Y_{ij} values that represent coverages of Kegg Orthology (KO) categories [Kanehisa et al., 2016]. A coverage represents the observed abundance of the category in whole genome sequencing data. A KO is a category that represents a set of functional orthologs, which are genes that perform the same functions across a set of species and have all evolved from the same gene in the most recent common ancestor of that set of species.

4.4.1 Analysis of functional differential abundance between the tongue and gingiva

The functional dataset that we analyze is a subset of data from the Human Microbiome Project (HMP) [Consortium, 2012]. The HMP was a large-scale initiative aimed at characterizing human microbiomes. We consider samples that come from either the tongue or the gingiva of subjects on their first study visit, and investigate functional differential abundance between these two body sites. This dataset includes coverages for $8,152$ KOs from 322 samples, 163 from the tongue and 159 from the gingiva. The tongue is considered the baseline category in this analysis.

We define the differential abundance parameters in our functional abundance model as $\gamma_1^m - g(\gamma_1)$ for $m \in \{1, \ldots, M\}$ and for $g_p(\gamma_1^m : m \in S_{rib})$ the pseudo-Huber loss over γ_m parameter values for a reference set S_{rib} of 31 ribosomal protein KOs (a list of these proteins and KOs is given in Table C.1 in Appendix C.2.1). The $\hat{\gamma}_1^m - g(\hat{\gamma}_1)$ estimates range from

-0.13 to 0.12 for functions m in S_{rib} and from -9.61 to 12.06 across all $m \in \{1, \ldots, M\}$. The functions in S_{rib} have magnitudes that are small and close to each other, making it seem like a good choice for a reference set for this data analysis. One notable feature of these dataset is a large number of KO categories that appear only in tongue samples or only in gingiva samples. There are $2,037$ ($\approx 25\%$) KOs that appear in samples from only one body site.

Figure 4.7 shows estimates $\hat{\gamma}_1 - g(\hat{\gamma}_1)$, distinguished by which body sites they are present in. Each estimate $\hat{\gamma}_1^m - g(\hat{\gamma}_1)$ represents the expected log fold-difference in observed abundance of function m between a sample from the gingiva and a sample from the tongue, where these samples have the same sample-specific sequencing effects, relative to the typical log fold-difference in observed abundance between these body sites for ribosomal proteins in the set S_{rib}. The left panel includes all functions, and the distribution has three modes. There is a large mode at 0 (this corresponds to the smoothed median log fold-difference within the set S_{rib}). This is another sign that these ribosomal proteins have $\gamma_1^m - g(\gamma_1)$ parameters that are a good representation of the typical $\gamma_1^m - g(\gamma_1)$ parameter value in the dataset, as the largest mode of estimates is approximately the same $g_p(\hat{\gamma}_1^m : m \in S_{rib})$. All functions with estimates near this mode appear in both gingiva and tongue samples. There are two more small modes, at approximately 3 and -4.5. Most of the functions near these modes appear only in gingiva samples and only in tongue samples, respectively.

The right panel of Figure 4.7 includes log fold-difference estimates only for functions that appear in more than one sample. Out of the $2,037$ ($\approx 25\%$) KOs that appear in samples from only one body site, only 419 appear in more than one sample. This distribution has a single mode at 0. This demonstrates that the two small modes in the left panel are made up primarily of functions that appear in a single sample.

When performing inference on this dataset, we account for multiple tests by calculating q-values [Storey, 2002] using the $8,152$ robust score test p-values. We use a false discovery rate threshold of 0.01. Using the robust score tests on reduced models, $4,298$ KOs (53%) have q-values less than 0.01. We perform a sensitivity analysis for our chosen reference set

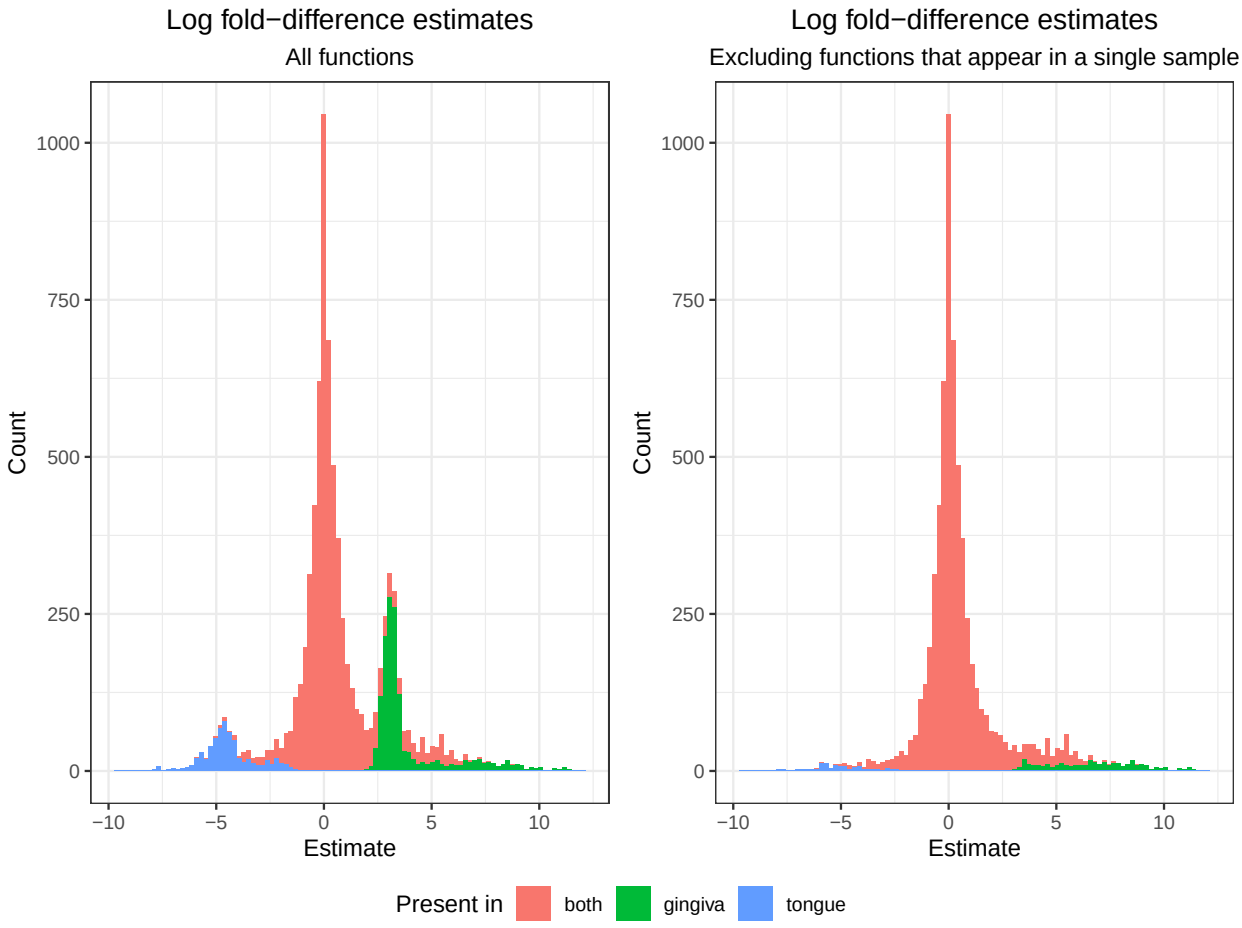


Figure 4.7: The distribution of $\hat{\gamma}_1 - g(\hat{\gamma}_1)$ estimates for $8,152$ parameters in the HMP data analysis. Estimation is done using Clausen and Willis' estimation algorithm. The bars are colored by the proportion of estimates in each bin that correspond to functions that are found in samples from both body sites, functions that are found only in gingiva samples, and functions that are found only in tongue samples.

S_{rib} by defining an alternate reference set, S_{alt} using the set of 16 KOs that correspond to ribosomal proteins used to build a new multidomain phylogeny in [Hug et al., 2016]. We then re-run the robust score tests on reduced models using constraint $g_p(\gamma_1^m : m \in S_{alt})$. There is a Pearson correlation of 0.952 between p-values from these two procedures with the two difference reference sets. Out of the $4,298$ parameters with q-values less than 0.01 from the robust score tests with the original reference set, $4,212$ of them also have q-values less than 0.01 when using the smaller reference set. We also compare the results of the robust score tests with the reduced model to robust score tests with the full model. We find a Pearson correlation of 1.00 between the p-values from the two sets of tests and robust score tests with reduced models run $1,089$ times faster than tests with the full model, on average.

We also analyze this data with Clausen and Willis' robust Wald tests, ALDEx2 [Fernandes et al., 2013], and ANCOM-BC2 [Lin and Peddada, 2024]. There is strong concordance between our $\hat{\gamma}_1^m - g(\hat{\gamma}_1)$ estimates and the parameter estimates from ALDEx2 and ANCOM-BC2, with Pearson correlations of 0.86 and 0.72 respectively. The correlation with ANCOM-BC2 estimates is only over KOs without separation, because ANCOM-BC2 does not produce estimates for separated KOs. The ALDEx2 estimated effect sizes have a similar range to our $\hat{\gamma} - g(\hat{\gamma})$ estimates (from ≈ -8.8 to ≈ 11.2), and the ANCOM-BC2 estimated effect sizes have a smaller range (from ≈ -4.5 to ≈ 4.0).

The robust Wald tests result in $6,304$ KOs (77%) that have q-values less than 0.01. The largest differences between the p-values from the robust Wald tests and robust score tests are for parameters from separated KOs. This is explored further in Appendix C.3.1. There are $4,345$ KOs (54%) with ALDEx2 q-values less than 0.01. These q-values are more conservative than the other q-values reported for this analysis because the ALDEx2 p-values range from 0 to 0.94 instead of from 0 to 1. Due to this truncated distribution of p-values, the proportion of true null KOs cannot be estimated in the q-value procedure and is instead fixed to 1. This results in a procedure that is equivalent to the more conservative BH procedure [Storey, 2002]. $6,034$ KOs (74%) have robust score test q-values and ALDEx2 q-values that are either both greater than 0.01 or both less than 0.01.

There are $4,413$ KOs (54%) with ANCOM-BC2 q-values less than 0.01, and $3,717$ of these KOs (46%) pass ANCOM-BC2's sensitivity analysis to assess the impact of adding pseudocounts to zero counts. There are two ways to deal with separation with ANCOM-BC2. When running ANCOM-BC2 and including the functionality to detect structural zeros, these separated KOs will be declared differentially abundant, and test statistics and p-values will not be calculated for parameters from these KOs. When running ANCOM-BC2 without detecting structural zeros, estimates from these KOs will be set to NA, and the p-values will be set to 1. We run ANCOM-BC2 without detecting structural zeros, and therefore have discordance between the robust score test p-values and ANCOM-BC2 p-values for all parameters from separated KOs. Considering only the KOs without separation and ignoring the results of the pseudocount sensitivity analysis, $4,838$ KOs (79%) have robust score test q-values and ANCOM-BC2 q-values that are either both greater than 0.01 or both less than 0.01.

In Figure 4.8, we look closer at KOs with small q-values based on p-values from robust score tests with reduced models. Both panels includes estimates and confidence intervals for functions with q-values less than 0.01, plotted by prevalence in the dataset and the body sites that the function is found in. The left panel includes functions with q-values less than 0.01, which corresponds with $\approx 53\%$ of KOs in the analysis. The function with the lowest prevalence in this subset appears in 5 samples. In this plot, we can see that functions with q-values less than 0.01 either have low to medium prevalence and large estimated $\hat{\gamma}_1^m - g(\hat{\gamma}_1)$ values, or high prevalence and smaller estimated $\hat{\gamma}_1^m - g(\hat{\gamma}_1)$ values. Functions that only appear in gingiva samples or only appear in tongue samples make up $\approx 4\%$ of the functions in this subset, and have larger effect sizes and smaller prevalences. The right panel uses a much more stringent q-value threshold of $1e-16$, and includes $\approx 10\%$ of KOs in the analysis that have the most evidence for differential abundance. These KOs all appear in 100 or more samples, and most have effect sizes greater than 1. This plot shows that even though KOs that appear only in gingiva samples or only in tongue samples have the largest estimated effect sizes, most of the KOs with the most evidence of differential abundance are present in

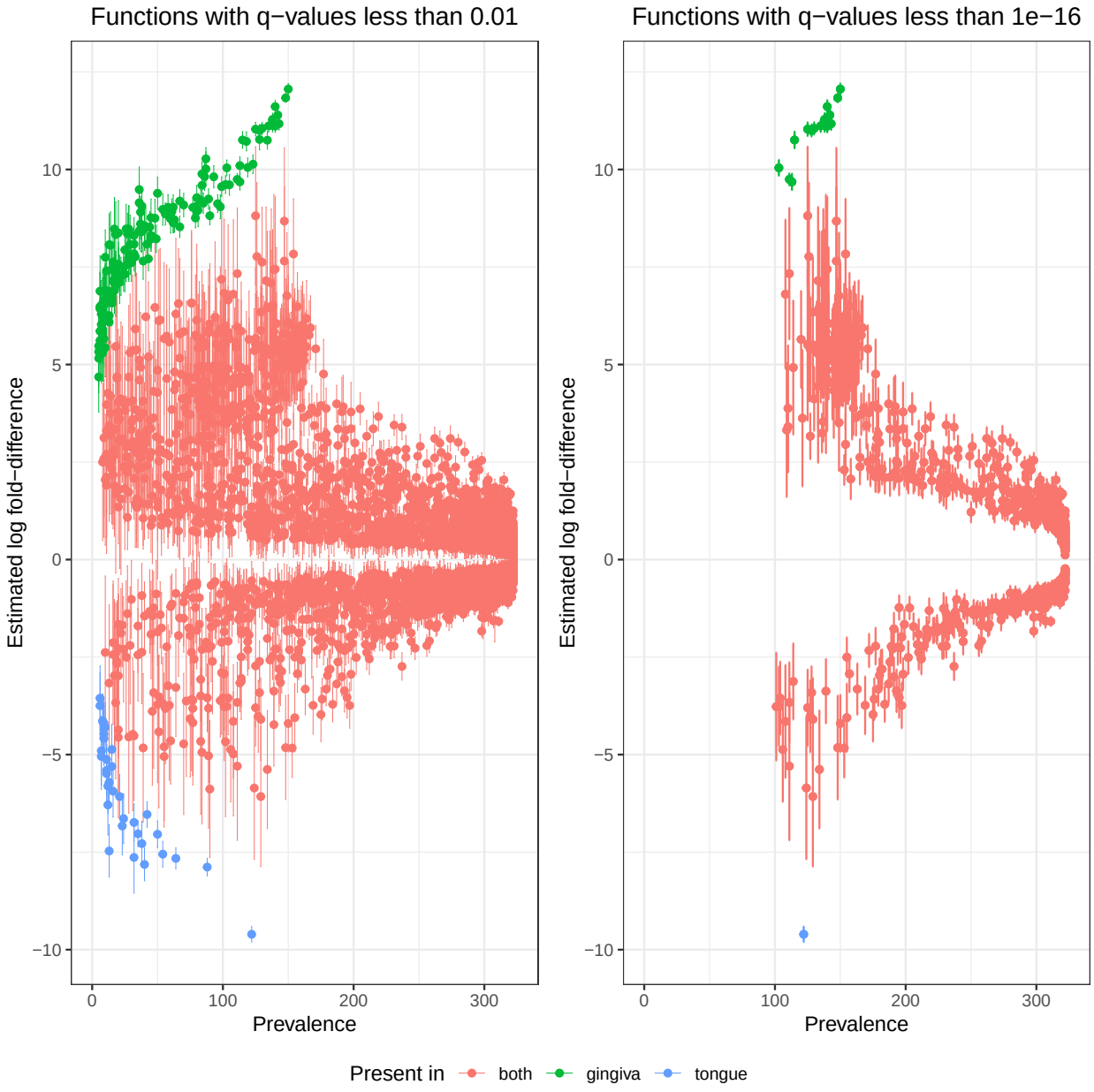


Figure 4.8: Estimated log fold-differences $\hat{\gamma}_1^m - g(\hat{\gamma}_1)$ with confidence intervals using robust standard errors by function prevalence from the HMP dataset. The left panel includes all functions with q-values less than 0.01 (this includes $\approx 53\%$ of functions). The right panel includes all functions with q-values less than $1e-16$ (this includes $\approx 10\%$ of functions). The estimates and error bars are colored by whether a function appears in samples from both body sites, only in gingiva samples, or only in tongue samples.

both body sites and have high prevalence.

KO name	KO function	Estimate	Prevalence
K01990	ABC-2 type transport system ATP-binding protein	1.25	322
K01992	ABC-2 type transport system permease protein	0.81	322
K05602	histidinol-phosphatase [EC:3.1.3.15]	6.18	164
K05988	dextranase [EC:3.2.1.11]	-1.83	298
K06188	aquaporin Z	-2.20	256
K13562	NAD(P)H:FMN oxidoreductase [EC:1.5.1.-]	1.13	322
K16038	N-methyltransferase [EC:2.1.1.-]	3.10	274
K20460	lantibiotic transport system permease protein	4.76	177
K21481	heme oxygenase (mycobilin-producing) [EC:1.14.99.57]	3.99	187
K24180	malate permease and related proteins	-1.58	311

Table 4.3: KOs with significant q-values from the HMP data analysis described in Section 4.4.1. Estimates and p-values are computed using robust score tests on reduced models. Q-values for all of these KOs are less than $1e29$. All KOs appear in samples from both body sites.

Finally, we present the ten KOs with the smallest p-values (and therefore the most evidence of differential abundance) in Table 4.3. These KOs all have q-values less than $1e29$, and appear in samples from both body sites. Seven of the KOs are more abundant in the gingiva, and three of the KOs are more abundant in the tongue. The estimated effect size magnitudes range from 0.81 to 6.18. However, it is important to note that these KOs have strong evidence of differential abundance in terms of log fold-differences in observed abundances from sequencing data. Therefore, we don't know if this reflects differential abundance in terms of the biological log fold-difference parameter that we care about, $\theta_1^m - g(\theta_1)$. Our Type I error and false discovery rate simulations suggested that large effect size

magnitudes often correspond with true discoveries. Even though K05602 is ranked third among these KOs in terms of p-values, we believe that it is more likely that this very small p-value reflects biological differential abundance and is not simply a product of sequencing effects because it has a large effect size compared to K01990 and K01992, which have small effect sizes.

HMP analysis permutation test

We additionally run a permutation test with this HMP dataset to investigate how often KOs are concluded to be significant by each method for random permutations of the site covariate. We permute the site covariate labels ten times, and for each permutation calculate p-values using robust score tests on reduced models, Clausen and Willis' robust Wald tests, ALDEx2, and ANCOM-BC2, and compute q-values using each set of p-values.

Across the ten permutations, the robust score test and ALDEx2 never detect a KO with a robust score q-value less than 0.01. There are between 625 and 681 KOs with robust Wald q-values less than 0.01 across permutations, of which 570 to 645 of these KOs have separation in the permuted data. There are between 465 and 584 KOs with ANCOM-BC2 q-values less than 0.01 across permutations, and between 6 and 65 of these KOs with significant ANCOM-BC2 q-values pass the pseudocount sensitivity analysis.

This permutation analysis demonstrates that the robust score test and ALDEx2 are sufficiently conservative to avoid detecting false signals for this setting in which permutation generates data under the null hypothesis of no differential abundance. ANCOM-BC2 is susceptible to false discoveries, although most of them do not pass the pseudocount sensitivity analysis. The robust Wald test is also susceptible to false signals, especially when there is separation in the data. Even though the Wald test retains Type I error rate control in simulations in Chapter 3 for simulated Poisson data with $n = 250$, this example provides a case in which the robust Wald test is prone to false discoveries in a dataset with a large sample size. This provides another reason to use the robust score test instead of the robust Wald test, even for datasets with large sample sizes.

In Appendix C.4, we present a second functional differential abundance analysis. We do not describe the results here, because the estimation results are similar to those in the HMP data analysis, and the small sample size of $n = 24$ is too small to detect differentially abundant functions with our method for a reasonable q-value threshold.

4.5 Discussion

In this chapter, we consider differential abundance analyses of molecular functions. We define our target differential abundance parameter, $\theta_k^m - g(\theta_k)$, as the expected log fold-difference in the true abundance of function m associated with the kth covariate level relative to the typical log fold-difference in a reference set of functions. We present a mean model for functional abundance, which relies on assumptions about sequencing effects that are considered to be true for taxonomic abundances. We argue that while we like to develop a model at the scale of W_{im}, we must instead model at the scale of W_{ijm} to account for differential taxon efficiencies. Due to the effect of taxon efficiencies, our target differential abundance parameter $\theta_k^m - g(\theta_k)$ does not appear in this model. Instead, the identifiable parameter that corresponds with covariate k in our model is $\gamma_k^m - g(\gamma_k)$, which is a function of both biological parameters and unknown taxon efficiencies. We show that we can estimate and efficiently test the parameter $\gamma - g(\gamma)$ by applying our taxonomic differential abundance method from Chapter 3to data generated from the functional abundance mean model.

Through simulation, we investigate the difference $|\gamma_1^m - g(\gamma_1) - \theta_1^m + g(\theta_1)|$ across a variety of efficiency settings. We find that testing the null hypothesis $H_0^{m(\gamma)} : \gamma_1^m - g(\gamma_1) = 0$ using data that is generated under $H_0^{m(\theta)} : \theta_1^m - g(\theta_1) = 0$ can lead to many false positives and false discoveries, especially for efficiencies with large fold-differences and datasets with larger sample sizes. However, we also find in simulation that when identifying differentially abundant functions based on both a false discovery rate threshold and a conservative estimated effect size threshold, we nearly always identify functions m such that $|\theta_1^m - g(\theta_1)| > 0$. We apply our method to a functional differential abundance analysis, and compare our results to ALDEx2 and ANCOM-BC2. In our analyses, we show that unlike ALDEx2, our method

can be applied to non-integer values and produce valid p-value distributions, and unlike ANCOM-BC2, our method can perform inference in the presence of separation and avoid detecting signals in permuted data.

There are several directions for future statistical work based on the findings that we present in this chapter. While we propose an identifiable model for functional abundance using a design matrix with a single categorical covariate, it would be preferable to have a model that is identifiable for an arbitrary design matrix. This could be used to understand what functional differential abundance parameters can be identified based on reasonable biological assumptions across much more complex experimental designs. Another extension to this project could be to develop more realistic simulation settings, that better reflect the complexity of how functions occur within taxa in real microbiomes.

One technique that has been used to account for the effect of taxon efficiencies in taxonomic abundance models is calibration with mock communities. Methods have been proposed by Brooks et al. [2015], Krehenwinkel et al. [2017], Bell et al. [2019], and McLaren et al. [2019] to use mock communities to estimate taxon efficiencies and apply these efficiency estimates to better estimate taxonomic abundances and microbial community composition. Similar approaches could be developed for functional abundance models. Data would need to be processed a step further than it currently is for these types of analyses, to provide coverages Y_{ijm} of function m in taxon j in sample i, instead of coverages Y_{im} of function m in sample i. Estimated taxon efficiencies could then potentially be used to calibrate a model for functional abundance similar to the one that we propose. However, mock communities are limited in their ability to represent complex microbial communities [Brooks et al., 2015]. Therefore, while mock communities can help us learn about the magnitudes of taxon efficiencies and could potentially be leveraged to calibrate functional abundance models, they are not a feasible solution for most data analyses.

While the parameter vector $\gamma_k - g(\gamma_k)$ in our functional abundance model is affected by taxon efficiencies, it is preferable to perform analyses in which we estimate and test this parameter while being careful to interpret it correctly than to completely avoid functional

differential abundance analyses due to this limitation. However, it is also important to interpret our estimated parameters correctly, as expected log fold-differences in observed abundances, that are affected by taxon efficiencies, and not expected log fold-differences in true abundances. We believe that an analysis performed in this way can lead to useful scientific conclusions about functions that are differentially abundant across a covariate, which can then be investigated further in future studies.

Code to reproduce all analyses run for this paper is available at `https://github.com/statdivlab/functionalEmu_supplementary`.

Chapter 5

CONCLUSION

In this dissertation, I have presented three statistical methods to visualizate, estimate, or test estimands that are relevant to the study of microbiomes. While each of these methods focuses on a different scientific question, they all consider the context in which metagenomic sequencing data are generated and are computationally efficient for the large scale of datasets that are typical in microbiome science.

In my second chapter, I argue that the parallel study of phylogenies at both a gene and genome level can provide complementary information about evolutionary histories. My third and fourth chapters both address differential abundance analyses, for abundances of taxa and molecular functions respectively. I believe that as in my second chapter, these two different perspectives with which to study abundance data could be combined in order to gain additional scientific insight into a microbial community of interest. Analyzing both taxonomic and functional differential abundance from the same dataset could help us better understand the relationship between changes in community composition and differences in the functional potential of a microbiome, as they relate to relevant environmental covariates.

In my third and fourth chapters, I develop methods for the differential abundance analysis of metagenomic sequencing data that do not utilize control data. However, when researchers build control data into their experimental design, we can use them to counteract the effects of unknown sequencing effects on our conclusions. In my third chapter, I study a differential abundance estimand in which I compare the expected log fold-difference in abundance for a given taxon across covariate levels to the typical log fold-difference in the analysis, with the typical fold-difference defined as the median fold-difference over all taxa or a small reference set. However, consider a dataset in which a researcher has added a synthetic compound to

all samples in known quantities. This is referred to as a "spike-in", and has been used to validate differential abundance methods by comparing the estimated differential abundance of this compounds to the known true differential abundance [Hardwick et al., 2018]. In my method, I could define the target estimand to be the log fold-difference of each taxon across covariate levels, relative to the log fold-difference of the synthetic compound. This would provide a baseline that corresponds to known absolute abundances.

In my fourth chapter, I show that the functional differential abundance estimand that I can identify, estimate, and test with my model is affected by unknown taxon efficiencies. In most metagenomic sequencing datasets, there is no way to quantify how large these efficiencies are, or how similar they are between taxa within different levels of classification. However, another type of control data can be useful here. If I had mock community data for a subset of taxa in a dataset, I could estimate the detection efficiencies for these taxa. Unlike in my third chapter, I could not directly apply this information to my choice of estimand. However, I could use this to design a sensitivity analysis, in which I perturb my data many times based on reasonable models of taxon efficiencies using information from the mock community, and investigate which functions have estimated effect sizes that are robust to these perturbations. This could help me understand the effects of differential taxon detection efficiencies on functional differential abundance analyses.

Microbiome science is an exciting area for statistical methods development because of the large scale of data and the importance of accounting for ways in which that data are sequenced and processed. It is the responsibility of statisticians who work in this field to distill this data into useful and interpretable estimands, while recognizing and communicating the limitations that their methods inherit from the limitations of metagenomic sequencing data. As improvements to high-throughput sequencing technology and to the experimental design of microbiome studies lead to more accurate metagenomic sequencing data, there will be opportunities for statisticians to develop new methods that utilize these improvements to expand the class of estimands that we can visualize, estimate, and perform inference on.

BIBLIOGRAPHY

Nina Amenta and Jeff Klingner. Case study: Visualizing sets of evolutionary trees. In IEEE Symposium on Information Visualization, 2002. INFOVIS 2002., pages 71–74. IEEE, 2002.

Francesco Asnicar, Andrew Maltez Thomas, Francesco Beghini, Claudia Mengoni, Serena Manara, Paolo Manghi, Qiyun Zhu, Mattia Bolzan, Fabio Cumbo, Uyen May, et al. Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using PhyloPhlAn 3.0. Nature Communications, 11(2500), 2020.

Eliran Avni and Sagi Snir. A new phylogenomic approach for quantifying horizontal gene transfer trends in prokaryotes. Scientific Reports, 10(12425), 2020.

Eric Bapteste, Maureen A O'Malley, Robert G Beiko, Marc Ereshefsky, J Peter Gogarten, Laura Franklin-Hall, François-Joseph Lapointe, John Dupré, Tal Dagan, Yan Boucher, et al. Prokaryotic evolution and the tree of life are two different things. Biology Direct, 4 (34), 2009.

Dennis Barden, Huiling Le, and Megan Owen. Limiting behaviour of fréchet means in the space of phylogenetic trees. Annals of the Institute of Statistical Mathematics, 70(1): 99–129, 2018.

Colin B Begg and Robert Gray. Calculation of polychotomous logistic regression parameters using individualized regressions. Biometrika, 71(1):11–18, 1984.

Karen L Bell, Kevin S Burgess, Jamieson C Botsch, Emily K Dobbs, Timothy D Read, and Berry J Brosi. Quantitative and qualitative assessment of pollen dna metabarcoding using constructed species mixtures. Molecular Ecology, 28(2):431–455, 2019.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological), 57(1):289–300, 1995.

Louis J Billera, Susan P Holmes, and Karen Vogtmann. Geometry of the space of phylogenetic trees. Advances in Applied Mathematics, 27(4):733–767, 2001.

Dennis D Boos. On generalized score tests. The American Statistician, 46(4):327–333, 1992.

Luis Boto. Horizontal gene transfer in evolution: facts and challenges. Proceedings of the Royal Society B: Biological Sciences, 277(1683):819–827, 2010.

J Paul Brooks, David J Edwards, Michael D Harwich, Maria C Rivera, Jennifer M Fettweis, Myrna G Serrano, Robert A Reris, Nihar U Sheth, Bernice Huang, Philippe Girerd, et al. The truth about metagenomics: quantifying and counteracting bias in 16s rrna studies. BMC microbiology, 15:1–14, 2015.

Christopher T Brown, Laura A Hug, Brian C Thomas, Itai Sharon, Cindy J Castelle, Andrea Singh, Michael J Wilkins, Kelly C Wrighton, Kenneth H Williams, and Jillian F Banfield. Unusual biology across a group comprising more than 15% of domain bacteria. Nature, 523(7559):208–211, 2015.

Mark V Brown, Federico M Lauro, Matthew Z DeMaere, Les Muir, David Wilkins, Torsten Thomas, Martin J Riddle, Jed A Fuhrman, Cynthia Andrews-Pfannkoch, Jeffrey M Hoffman, et al. Global biogeography of sar11 marine bacteria. Molecular systems biology, 8 (1):595, 2012.

Matthew J Bull and Nigel T Plummer. Part 1: the human gut microbimoe in health and disease. Integrative Medicine: A Clinician's Journal, 13(6):17–22, 2014.

Salvador Capella-Gutiérrez, José M Silla-Martínez, and Toni Gabaldón. trimal: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics, 25 (15):1972–1973, 2009.

Ricardo Cavicchioli, William J Ripple, Kenneth N Timmis, Farooq Azam, et al. Scientists' warning to humanity: microorganisms and climate change. Nature reviews microbiology, 17:569–586, 2019.

John Chakerian and Susan Holmes. Computational tools for evaluating phylogenetic and hierarchical clustering trees. Journal of Computational and Graphical Statistics, 21(3): 581–599, 2012.

David S Clausen and Amy D Willis. Estimating fold changes from partially observed outcomes with applications in microbial metagenomics, 2024.

The Human Microbiome Project Consortium. A framework for human microbiome research. Nature, 486(7402):215–221, 2012.

Vincent Daubin, Manolo Gouy, and Guy Perriere. A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. Genome Research, 12 (7):1080–1090, 2002.

Sean R Eddy. Accelerated profile HMM searches. PLoS Computational Biology, 7(10): e1002195, 2011.

Robert C Edgar. Muscle v5 enables improved estimates of phylogenetic tree confidence by ensemble bootstrapping. bioRxiv, 2021.

Andrew D Fernandes, Jean M Macklaim, Thomas G Linn, Gregor Reid, and Gregory B Gloor. Anova-like differential expression (aldex) analysis for mixed population rna-seq. PloS one, 8(7):e67019, 2013.

David Firth. Bias reduction of maximum likelihood estimates. Biometrika, 80(1):27–38, 1993.

Stephen J Giovannoni. Sar11 bacteria: the most abundant plankton in the oceans. Annual review of marine science, 9(1):231–255, 2017.

Gregory B Gloor, Jean M Macklaim, Vera Pawlowsky-Glahn, and Juan J Egozcue. Microbiome datasets are compositional: and this is not optional. Frontiers in microbiology, 8: 2224, 2017.

Kevin Gori, Tomasz Suchan, Nadir Alvarez, Nick Goldman, and Christophe Dessimoz. Clustering genes of common evolutionary history. Molecular Biology and Evolution, 33(6): 1590–1605, 2016.

Gillian Grindstaff and Megan Owen. Geometric comparison of phylogenetic trees with different leaf sets. SIAM Journal on Applied Algebra and Geometry, 3:691–720, 2020.

Xu Guo, Wei Pan, John E Connett, Peter J Hannan, and Simone A French. Small-sample performance of the robust score test and its modifications in generalized estimating equations. Statistics in medicine, 24(22):3479–3495, 2005.

Simon A Hardwick, Wendy Y Chen, Ted Wong, Bindu S Kanakamedala, Ira W Deveson, Sarah E Ongley, Nadia S Santini, Esteban Marcellin, Martin A Smith, Lars K Nielsen, et al. Synthetic microbe communities provide internal reference standards for metagenome sequencing and analysis. Nature communications, 9(1):3096, 2018.

David M Hillis, Tracy A Heath, and Katherine St John. Analysis and visualization of tree space. Systematic Biology, 54(3):471–482, 2005.

Diep Thi Hoang, Olga Chernomor, Arndt Von Haeseler, Bui Quang Minh, and Le Sy Vinh. Ufboot2: improving the ultrafast bootstrap approximation. Molecular Biology and Evolution, 35(2):518–522, 2018.

Susan Holmes. Visualising data. In Statistical Problems In Particle Physics, Astrophysics And Cosmology, pages 197–207. World Scientific, 2006.

Kaijian Hou, Zhuo-Xun Wu, Xuan-Yu Chen, Jing-Quan Wang, Dongya Zhang, Chuanxing Xiao, Dan Zhu, Jagadish B Koya, Liuya Wei, Jilin Li, et al. Microbiota in health and diseases. Signal transduction and targeted therapy, 7(1):1–28, 2022.

Taishan Hu, Nilesh Chitnis, Dimitri Monos, and Anh Dinh. Next-generation sequencing technologies: An overview. Human Immunology, 82(11):801–811, 2021.

Yingtian Hu, Glen A Satten, and Yi-Juan Hu. Locom: A logistic regression model for testing differential abundance in compositional microbiome data with false discovery rate control. Proceedings of the National Academy of Sciences, 119(30):e2122788119, 2022.

Wen Huang, Guifang Zhou, Melissa Marchand, Jeremy R Ash, David Morris, Paul Van Dooren, Jeremy M Brown, Kyle A Gallivan, and Jim C Wilgenbusch. Treescaper: visualizing and extracting phylogenetic signal from sets of trees. Molecular Biology and Evolution, 33(12):3314–3316, 2016.

Laura A Hug, Brett J Baker, Karthik Anantharaman, Christopher T Brown, Alexander J Probst, Cindy J Castelle, Cristina N Butterfield, Alex W Hernsdorf, Yuki Amano, Kotaro Ise, et al. A new view of the tree of life. Nature microbiology, 1(5):1–6, 2016.

Doug Hyatt, Gwo-Liang Chen, Philip F LoCascio, Miriam L Land, Frank W Larimer, and Loren J Hauser. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics, 11(119), 2010.

Hiroyuki Imachi, Masaru K Nobu, Nozomi Nakahara, Yuki Morono, Miyuki Ogawara, Yoshihiro Takaki, Yoshinori Takano, Katsuyuki Uematsu, Tetsuro Ikuta, Motoo Ito, et al. Isolation of an archaeon at the prokaryote–eukaryote interface. Nature, 577(7791):519–525, 2020.

Thibaut Jombart, Michelle Kendall, Jacob Almagro-Garcia, and Caroline Colijn. treespace: Statistical exploration of landscapes of phylogenetic trees. Molecular Ecology Resources, 17(6):1385–1392, 2017.

Minoru Kanehisa, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. Kegg as a reference resource for gene and protein annotation. Nucleic acids research, 44 (D1):D457–D462, 2016.

Eric Karsenti, Silvia G Acinas, Peer Bork, Chris Bowler, Colomban De Vargas, Jeroen Raes, Matthew Sullivan, Detlev Arendt, Francesca Benzoni, Jean-Michel Claverie, et al. A holistic approach to marine eco-systems biology. PLoS biology, 9(10):e1001177, 2011.

Michelle Kendall and Caroline Colijn. Mapping phylogenetic trees to reveal distinct patterns of evolution. Molecular Biology and Evolution, 33(10):2735–2743, 2016.

Tomasz Konopka. umap: Uniform Manifold Approximation and Projection, 2023. URL `https://CRAN.R-project.org/package=umap`. R package version 0.2.10.0.

Ioannis Kosmidis and David Firth. Multinomial logit bias reduction via the poisson log-linear model. Biometrika, 98(3):755–759, 2011.

Henrik Krehenwinkel, Madeline Wolf, Jun Ying Lim, Andrew J Rominger, Warren B Simison, and Rosemary G Gillespie. Estimating and mitigating amplification bias in qualitative and quantitative arthropod metabarcoding. Scientific reports, 7(1):17668, 2017.

Jesse H. Krijthe. Rtsne: T-Distributed Stochastic Neighbor Embedding using Barnes-Hut Implementation, 2015. URL `https://github.com/jkrijthe/Rtsne`. R package version 0.16.

Michael D Lee. Gtotree: a user-friendly workflow for phylogenomics. Bioinformatics, 35(20): 4162–4164, 2019.

Devin R Leopold and Posy E Busby. Host genotype and colonist arrival order jointly govern plant microbiome composition and function. Current Biology, 30(16):3260–3266, 2020.

Andy Liaw and Matthew Wiener. Classification and regression by randomforest. R News, 2 (3):18–22, 2002. URL `https://CRAN.R-project.org/doc/Rnews/`.

Huang Lin and Shyamal Das Peddada. Analysis of compositions of microbiomes with bias correction. Nature communications, 11(1):3514, 2020.

Huang Lin and Shyamal Das Peddada. Multigroup analysis of compositions of microbiomes with covariate adjustments and repeated measures. Nature Methods, 21(1):83–91, 2024.

Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. Genome biology, 15:1–21, 2014.

Himel Mallick, Ali Rahnavard, Lauren J McIver, Siyuan Ma, Yancong Zhang, Long H Nguyen, Timothy L Tickle, George Weingart, Boyu Ren, Emma H Schwager, et al. Multivariable association discovery in population-scale meta-omics studies. PLoS computational biology, 17(11):e1009442, 2021.

Siddhartha Mandal, Will Van Treuren, Richard A White, Merete Eggesbø, Rob Knight, and Shyamal D Peddada. Analysis of composition of microbiomes: a novel method for studying microbial composition. Microbial ecology in health and disease, 26(1):27663, 2015.

Frederick A Matsen. Phylogenetics and the human microbiome. Systematic Biology, 64(1): e26–e41, 2015.

P. McCullagh and J.A Nelder. Generalized Linear Models. Chapman and Hall, London, 1989.

Michael R McLaren, Amy D Willis, and Benjamin J Callahan. Consistent and correctable bias in metagenomic sequencing experiments. Elife, 8:e46923, 2019.

Michael R McLaren, Jacob T Nearing, Amy D Willis, Karen G Lloyd, and Benjamin J Callahan. Implications of taxonomic bias for microbial differential-abundance analysis. bioRxiv, pages 2022–08, 2022.

Bui Quang Minh, Heiko A Schmidt, Olga Chernomor, Dominik Schrempf, Michael D Woodhams, Arndt Von Haeseler, and Robert Lanfear. Iq-tree 2: new models and efficient methods for phylogenetic inference in the genomic era. Molecular Biology and Evolution, 37(5):1530–1534, 2020.

Jaina Mistry, Sara Chuguransky, Lowri Williams, Matloob Qureshi, Gustavo A Salazar, Erik LL Sonnhammer, Silvio CE Tosatto, Lisanna Paladin, Shriya Raj, Lorna J Richardson, et al. Pfam: The protein families database in 2021. Nucleic Acids Research, 49(D1): D412–D419, 2021.

Jacob T Nearing, Gavin M Douglas, Molly G Hayes, Jocelyn MacDonald, Dhwani K Desai, Nicole Allward, Casey MA Jones, Robyn J Wright, Akhilesh S Dhanani, André M Comeau, et al. Microbiome differential abundance methods produce different results across 38 datasets. Nature communications, 13(1):342, 2022.

Anna Neufeld, Ameer Dharamshi, Lucy L Gao, and Daniela Witten. Data thinning for convolution-closed distributions. Journal of Machine Learning Research, 25(57):1–35, 2024a.

Anna Neufeld, Lucy L Gao, Joshua Popp, Alexis Battle, and Daniela Witten. Inference after latent variable estimation for single-cell rna sequencing data. Biostatistics, 25(1):270–287, 2024b.

Tom MW Nye. Principal components analysis in the space of phylogenetic trees. The Annals of Statistics, 39(5):2716–2739, 2011.

Megan Owen and J Scott Provan. A fast algorithm for computing geodesic distances in tree space. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 8(1): 2–13, 2011.

Emmanuel Paradis and Klaus Schliep. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics, 35:526–528, 2019. doi: 10.1093/bioinformatics/bty633.

Donovan H Parks, Maria Chuvochina, David W Waite, Christian Rinke, Adam Skarshewski, Pierre-Alain Chaumeil, and Philip Hugenholtz. A standardized bacterial taxonomy based

on genome phylogeny substantially revises the tree of life. Nature Biotechnology, 36(10): 996–1004, 2018.

Donovan H Parks, Maria Chuvochina, Pierre-Alain Chaumeil, Christian Rinke, Aaron J Mussig, and Philip Hugenholtz. A complete domain-to-species taxonomy for bacteria and archaea. Nature Biotechnology, 38(9):1079–1086, 2020.

Morgan N Price, Paramvir S Dehal, and Adam P Arkin. Fasttree 2–approximately maximum-likelihood trees for large alignments. PLOS ONE, 5(3):e9490, 2010.

Pere Puigbo, Yuri I Wolf, and Eugene V Koonin. The tree and net components of prokaryote evolution. Genome Biology and Evolution, 2:745–756, 2010.

Ravi Ranjan, Asha Rani, Ahmed Metwally, Halvor S McGee, and David L Perkins. Analysis of the microbiome: Advantages of whole genome shotgun versus 16s amplicon sequencing. Biochemical and biophysical research communications, 469(4):967–977, 2016.

Yingying Ren, Sihan Zha, Jingwen Bi, José A Sanchez, Cara Monical, Michelle Delcourt, Rosemary K Guzman, and Ruth Davidson. A combinatorial method for connecting bhv spaces representing different numbers of taxa. arXiv, (1708.02626), 2017.

Eric W Sayers, Richa Agarwala, Evan E Bolton, J Rodney Brister, Kathi Canese, Karen Clark, Ryan Connor, Nicolas Fiorini, Kathryn Funk, Timothy Hefferon, et al. Database resources of the national center for biotechnology information. Nucleic Acids Research, 47 (Database issue):D23, 2019.

Nicola Segata, Jacques Izard, Levi Waldron, Dirk Gevers, Larisa Miropolsky, Wendy S Garrett, and Curtis Huttenhower. Metagenomic biomarker discovery and explanation. Genome biology, 12:1–18, 2011.

Nicola Segata, Daniela Börnigen, Xochitl C Morgan, and Curtis Huttenhower. Phylophlan is a new method for improved phylogenetic and taxonomic placement of microbes. Nature Communications, 4(1):1–11, 2013.

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

Jonathan Taylor and Robert J Tibshirani. Statistical learning and selective inference. Proceedings of the National Academy of Sciences, 112(25):7629–7634, 2015.

Adrian Tett, Edoardo Pasolli, Giulia Masetti, Danilo Ercolini, and Nicola Segata. Prevotella diversity, niches and interactions with the human host. Nature Reviews Microbiology, 19: 585–599, 2021.

Thorsten Thiergart, Giddy Landan, and William F Martin. Concatenated alignments and the case of the disappearing tree. BMC Evolutionary Biology, 14(266), 2014.

W. N. Venables and B. D. Ripley. Modern Applied Statistics with S. Springer, New York, fourth edition, 2002. URL `https://www.stats.ox.ac.uk/pub/MASS4/`. ISBN 0-387-95457-0.

Halbert White. Maximum likelihood estimation of misspecified models. Econometrica: Journal of the econometric society, pages 1–25, 1982.

Amy Willis. Confidence sets for phylogenetic trees. Journal of the American Statistical Association, 114(525):235–244, 2019.

Amy Willis and Rayna Bell. Uncertainty in phylogenetic tree estimates. Journal of Computational and Graphical Statistics, 27(3):542–552, 2018.

Jakob Wirbel, Paul Theodor Pyl, Ece Kartal, Konrad Zych, Alireza Kashani, Alessio Milanese, Jonas S Fleck, Anita Y Voigt, Albert Palleja, Ruby Ponnudurai, et al. Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. Nature medicine, 25(4):679–689, 2019.

Martin Wu and Jonathan A Eisen. A simple, fast, and accurate method of phylogenomic inference. Genome Biology, 9(10):R151, 2008.

Guangchuang Yu, David K Smith, Huachen Zhu, Yi Guan, and Tommy Tsan-Yuk Lam. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods in Ecology and Evolution, 8(1):28–36, 2017.

Qiyun Zhu, Uyen Mai, Wayne Pfeiffer, Stefan Janssen, Francesco Asnicar, Jon G Sanders, Pedro Belda-Ferre, Gabriel A Al-Ghalith, Evguenia Kopylova, Daniel McDonald, et al. Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains bacteria and archaea. Nature Communications, 10(5477), 2019.

Appendix A

ANALYZING MICROBIAL EVOLUTION THROUGH GENE AND GENOME PHYLOGENIES

A.1 Additional analyses of Prevotella data from Section 2.3.1

A.1.1 Sensitivity to the choice of log map base tree

The purpose of this additional analysis of the Prevotella gene trees is to investigate the sensitivity of the proposed visualization to changes in the base tree used in the log map transformation. Figure A.1 shows the proposed visualization constructed using the phylogenomic tree as the base tree (which, in this case, is also the tree with the lowest mean squared BHV distance from the other trees in the set) as well as the visualization constructed with five different base trees. These five trees have the next smallest mean squared BHV distance from other trees in the set after the phylogenomic tree. Each of the visualizations in Figure A.1 show the same three outlying genes (unlabeled for visual simplicity): BacA, DMRL_synthase, and GTP_cyclohydroI. There are subtle differences between the visualizations including the spread of the non-outlying trees in the second principal component and the orientation of the BacA gene (the outlier in the second principal component) compared to the rest of the trees. However, overall, the same conclusions can be drawn by each of the visualizations in Figure A.1.

In Figure A.2, we now show the proposed log map visualization with base trees chosen to be the three outliers visible in Figure A.1. The visualizations that use BacA and GTP_cyclohydrI as base trees are very similar to the one that uses the phylogenomic tree as the base tree. Interestingly, when DMRL_synthase is the base tree, BacA is no longer identifiable as a clear outlier in the second principal component.

Our recommendation for choosing base trees is to choose a tree that is central to the tree

set (in terms of BHV distance). Such a choice will map as much information as possible from BHV tree space to the log map representation. However, we strongly suggest running a sensitivity analysis (similar to those described above) to investigate if the results of the visualization are relatively robust to the base tree choice.

Finally, we investigate the result of choosing a base tree with the topology drawn uniformly-at-random, i.e., a base tree with a topology uninformed by the gene tree data. While gene trees rarely all have the same topology as each other, they tend to have similar topologies relative to the number of possible tree topologies. Thus, they exist in a small number of BHV orthants. We hypothesize that the proposed visualization was relatively robust to choice of base tree in Figures A.1 and A.2 because all of the trees in the analysis exist in this small number of nearby orthants in BHV space. By this rationale, choosing a base tree from an orthant that is farther away from the set of trees could distort the visualization, because the direction of the first segment of the geodesic from the base tree to each tree in the analysis will be similar. Figure A.3 shows the proposed visualization using a randomly generated tree on 78 tips as the base tree. The topology is drawn at random using the function `rtree` from the package ape [Paradis and Schliep, 2019], and the branch lengths are drawn at random from the empirical distribution of branch lengths from all trees in this Prevotella dataset. Results look very similar when different random base trees are generated. This plot shown in Figure A.3 has the same patterns as the majority of the plots in Figures A.1 and A.2, with the exception of the randomly generated base tree, which is an outlier in the second principal component. This plot shows that while the random base tree is clearly an outlier from the set of trees in the analysis, the patterns in the plot are even robust to this choice of topology-uninformed base tree.

A.1.2 MDS visualization

Figure A.4 provides a visualization of the Prevotella gene trees and phylogenomic tree using metric MDS of BHV distances. This plot shows the same signals as the proposed visualization, with the phylogenomic tree towards the center of a cluster of gene trees, with

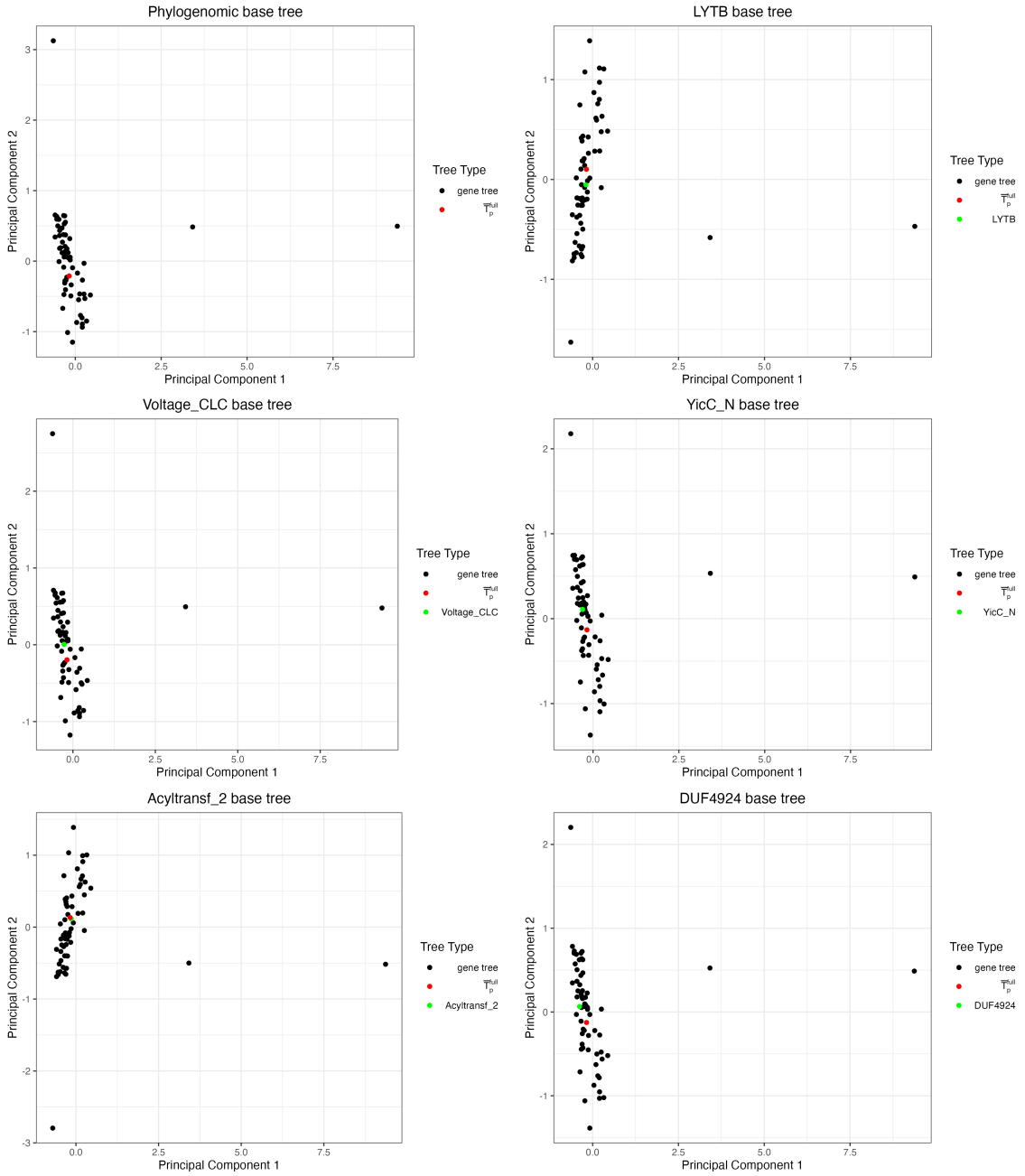


Figure A.1: The proposed visualization of 63 gene trees constructed from 78 genomes from the *Prevotella* genus, depicted by a two-dimensional scatterplot. Each subplot is constructed using a different tree as the base tree in the log map (see subplot titles). The five chosen gene trees (LYTB, Voltage_CLC, YicC_N, Acyltransf_2, and DUF4924) have the smallest mean squared BHV distances from other trees in the dataset.

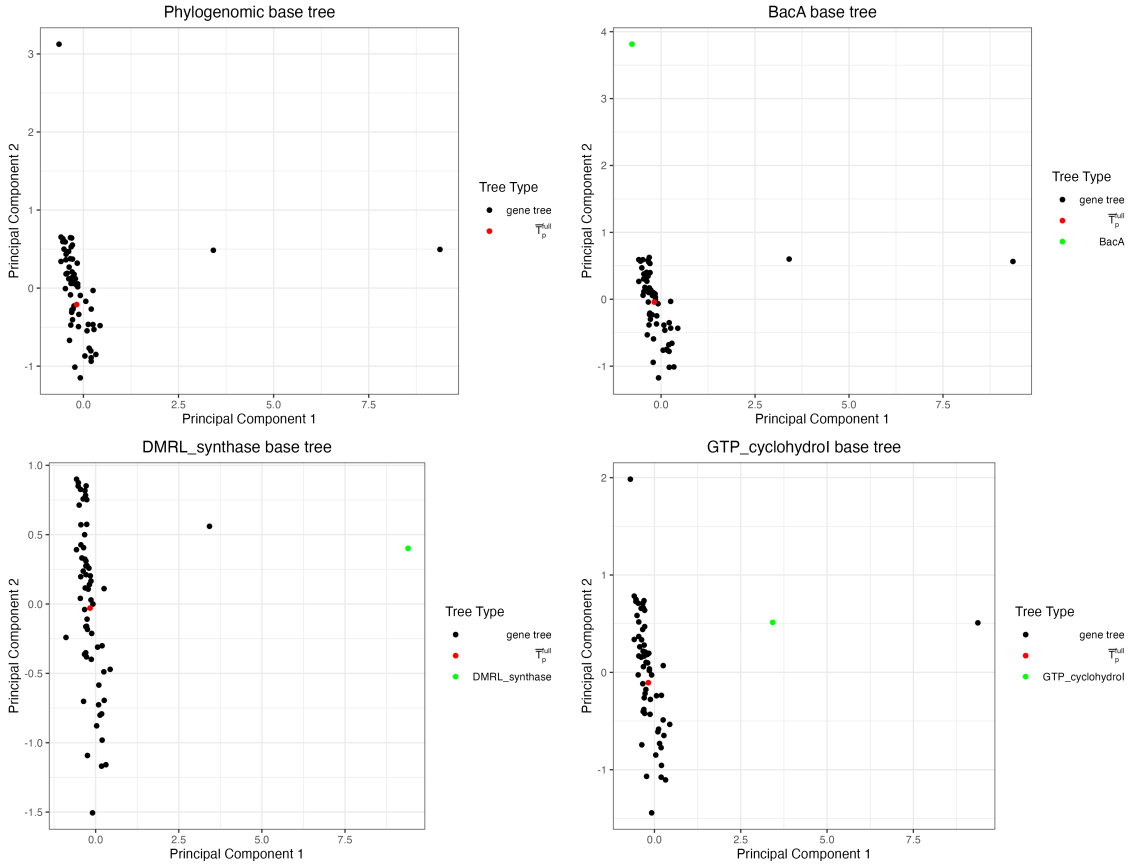


Figure A.2: The proposed visualization of 63 gene trees constructed from 78 genomes from the *Prevotella* genus, depicted by a two-dimensional scatterplot. Each subplot is constructed using a different tree as the base tree in the log map. BacA, DMRL_synthase, and GTP_cyclohydroI are the outlying genes identified in the plot created with the phylogenomic tree as the base tree.

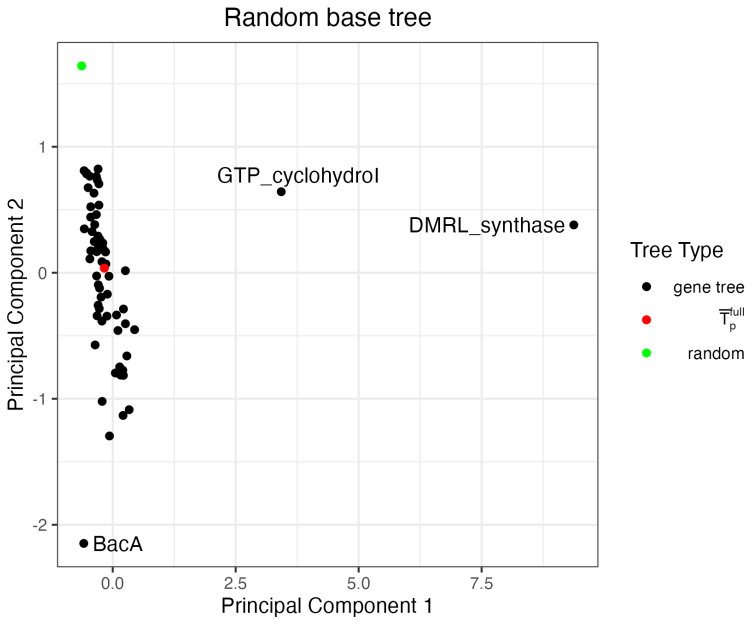


Figure A.3: The proposed visualization of 63 gene trees constructed from 78 genomes from the *Prevotella* genus, depicted by a two-dimensional scatterplot. The base tree is a randomly generated tree with 78 tips.

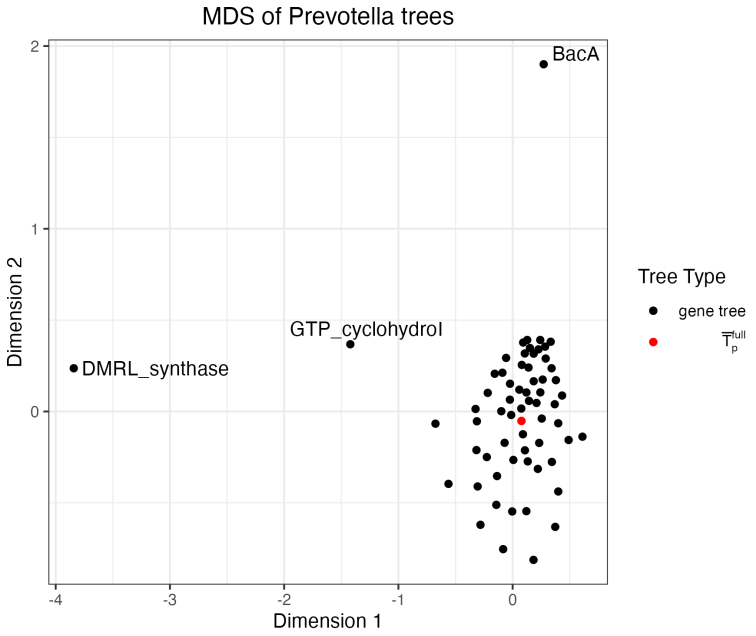


Figure A.4: A visualization of 63 gene trees constructed from 78 genomes from the *Prevotella* genus using MDS of BHV distances between trees, depicted by a two-dimensional scatterplot.

DMRL_synthase and GTP_cyclohydroI as outliers in the first dimension and BacA as an outlier in the second dimension. However, the cloud of non-outlying trees is less dense in the MDS plot in Figure A.4 compared to the plots in Figures A.1 and A.2. The cloud of points is spread out along both MDS dimensions 1 and 2 in Figure A.4, versus only along principal component 2 in the PCA plots.

A.1.3 tSNE and UMAP visualizations

In addition to the more traditional dimension reduction methods of PCA and MDS, we can also consider performing dimension reduction using t-distributed stochastic neighbor embedding (t-SNE) and uniform manifold approximation and projection (UMAP) on the log map vectors. We implement t-SNE using the function `Rtsne` from the R package of the same name [Krijthe, 2015], and UMAP using the function `umap` from the R package of the

same name [Konopka, 2023]. As both of these dimension reduction algorithms are stochastic, we consider visualizations created with four different initial seeds. The t-SNE visualizations can be seen in Figure A.5 and the UMAP visualizations in Figure A.6.

Each of the t-SNE plots show the phylogenomic tree as a central point among the set of points and BacA, DMRL_synthase, and GTP_cyclohydroI are extreme points although not separated from the other points as apparent outliers as seen in the PCA and MDS plots. The relative positions of the other points appear to be slightly different between each initial seed, although in each plot the points are spread out in a sparse pattern throughout the range of the two axes. The UMAP plots have more variation across initial seeds. The point representing the phylogenomic tree is towards the center of the points in one of the plots but is farther from the center in the other three. While the points representing BacA, DMRL_synthase, and GTP_cyclohydroI are closer to the edges of the set of points, they are not the most extreme points in the plots and would not be identified as apparent outliers. Additionally the overall pattern of the other points varies between plots. It is interesting that the plots for this analysis change so much across initial seeds and that they do not share the same outliers as the PCA and MDS plots. This suggests that different dimension reduction tools will provide different insights and may be more or less useful depending on the data being analyzed.

A.2 Additional analyses of Streptococcus data from Section 2.3.2

A.2.1 Sensitivity to base tree in log map transformation

Here we investigate the sensitivity of the proposed Streptococcus visualization to changes in the base tree used in the log map transformation. Figure A.7 shows the proposed visualization constructed using the phylogenomic tree as the base tree (which in this analysis is also the tree with the lowest mean squared BHV distance from the other trees in the set) as well as the visualization constructed with five different base trees. Three of these trees (MnmE_helical, NFACT_N, and NAPRTase) have the smallest mean squared BHV distance

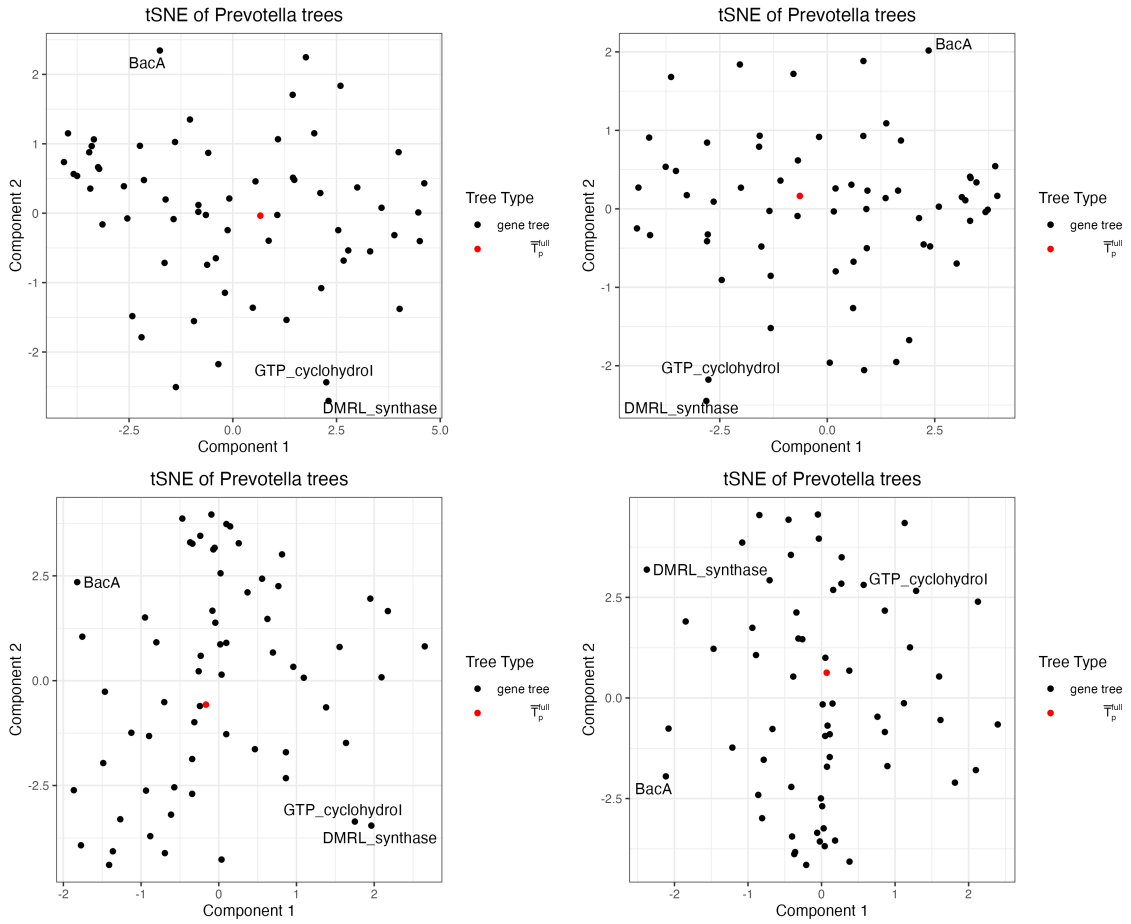


Figure A.5: A visualization of 63 gene trees constructed from 78 genomes from the *Prevotella* genus, depicted by a two-dimensional scatterplot of the first two dimensions from t-SNE. Each subplot is constructed using a different initial seed. BacA, DMRL_synthase, and GTP_cyclohydroI are the outlying genes identified in the plot created with the phylogenomic tree as the base tree.

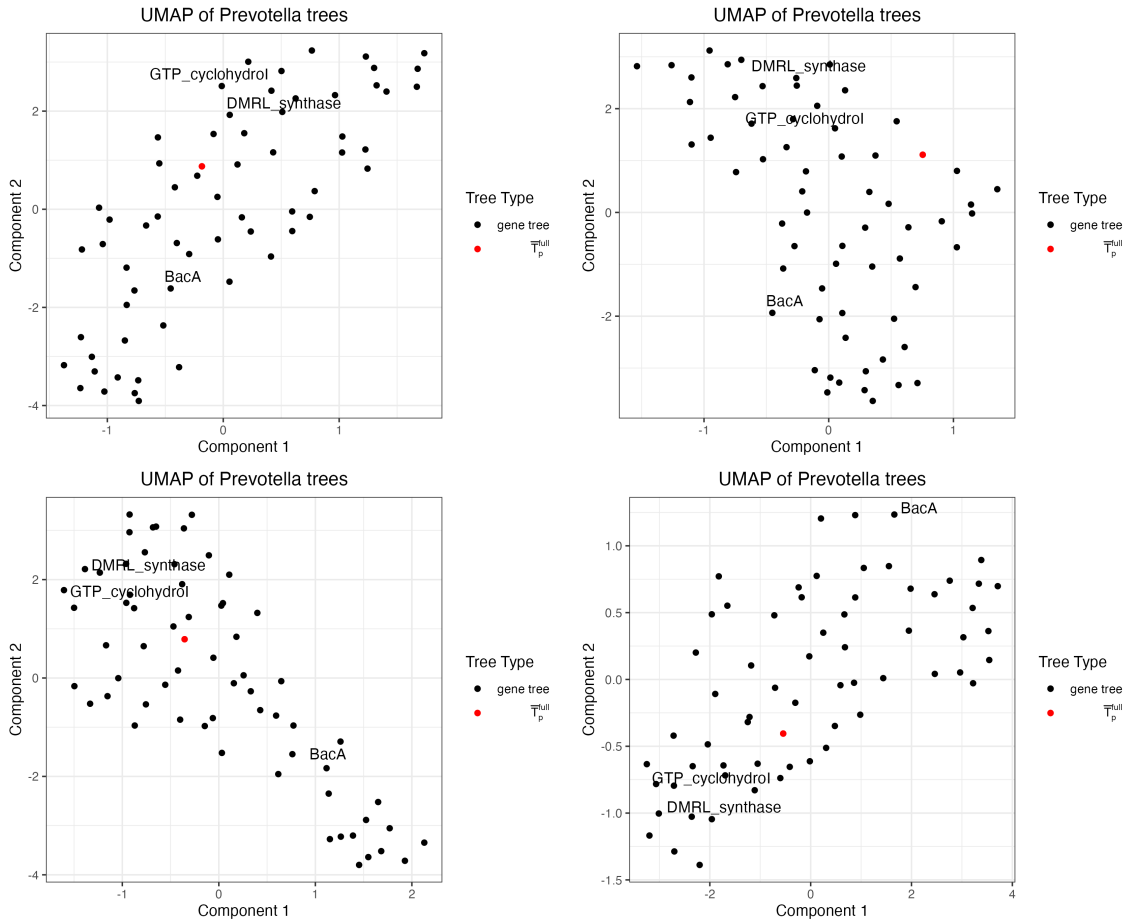


Figure A.6: A visualization of 63 gene trees constructed from 78 genomes from the *Prevotella* genus, depicted by a two-dimensional scatterplot of the first two dimensions from UMAP. Each subplot is constructed using a different initial seed. BacA, DMRL_synthase, and GTP_cyclohydroI are the outlying genes identified in the plot created with the phylogenomic tree as the base tree.

from other trees in the set (Ribosomal_L9_C, Ribosomal_S30AE, and DUF3270 also have two of the smallest mean squared BHV distances from other trees but are non-binary and therefore cannot be base trees in the log map transformation). The other two trees (EcsB and DUF1934) are apparent outliers in the visualization with the phylogenomic tree as base. While there are subtle differences between the plots, including the orientation of the first principal component, they overall show the same relationships between trees, including the dense clustering of ribosomal gene trees. This analysis appears to be robust to the choice of base tree, as trees central to the set in terms of BHV distance and trees on the edges of the visualization across all choices of base tree give similar results.

We also investigate the result of choosing a base tree with a random topology. Figure A.8 shows the proposed visualization using a base tree with a randomly generated topology on 106 tips and branches drawn from the empirical distribution of non-zero branch lengths from trees in this analysis. The branch lengths must be non-zero in order for the tree to be binary and a viable base tree. As in the Prevotella exploration, the random base tree is an outlier (here in the first principal component) but the other patterns seen in Figure A.7 can still be seen in the plot. This includes the clustering of points representing ribosomal vs non-ribosomal trees, the cone-shaped pattern of points, and the more extreme points. This provides another case in which the proposed visualization is robust even to base trees with topologies that are uninformed by the trees in the analysis.

A.2.2 MDS visualization

Figure A.9 provides a visualization of the same set of Streptococcus gene trees and phylogenomic tree using MDS of BHV distances. While there is a similar high density of trees in part of the plot (representing ribosomal gene trees) and the other trees are more spread out, there is a different overall shape than in the log map and PCA plots. The points are less dense than in the log map and PCA plots and collectively have a more circular pattern. While the identified outliers are at the edge of the cloud of trees, they are not easily discernible as outliers.

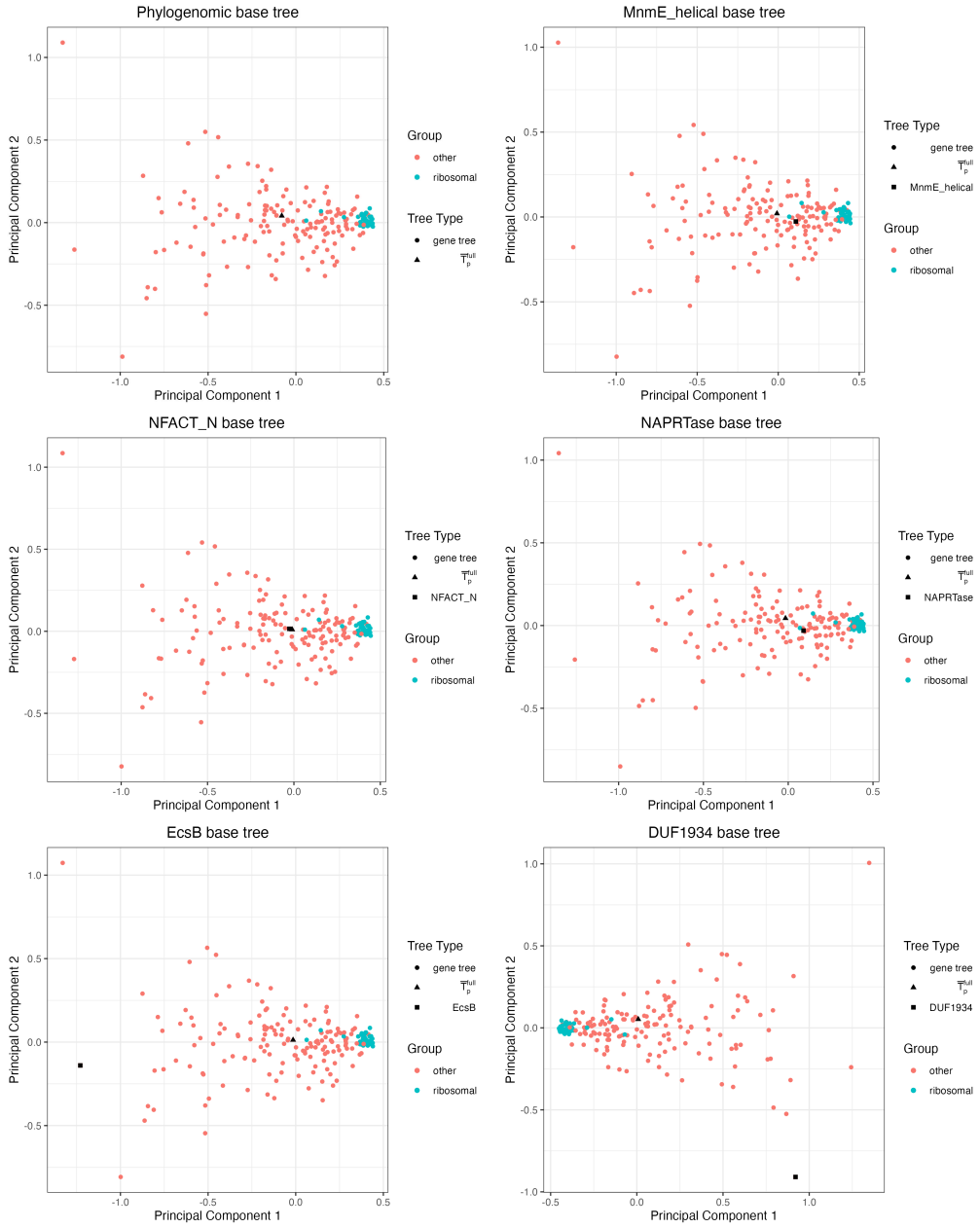


Figure A.7: The proposed visualization of 196 gene trees estimated from 106 genomes in the Streptococcus genus shown as a two-dimensional scatterplot. The points are colored by whether or not they represent a ribosomal gene. Each subplot is constructed using a different tree as the base tree in the log map. Gene trees MnmE_helical, NFACT_N, and NAPRTase have smallest mean squared BHV distances from other trees in the set respectively out of binary trees (non-binary trees are ignored because they cannot be used as the base tree in the log map). EcsB and DUF1934 are outliers identified in the visualization with the phylogenomic tree as the base tree.

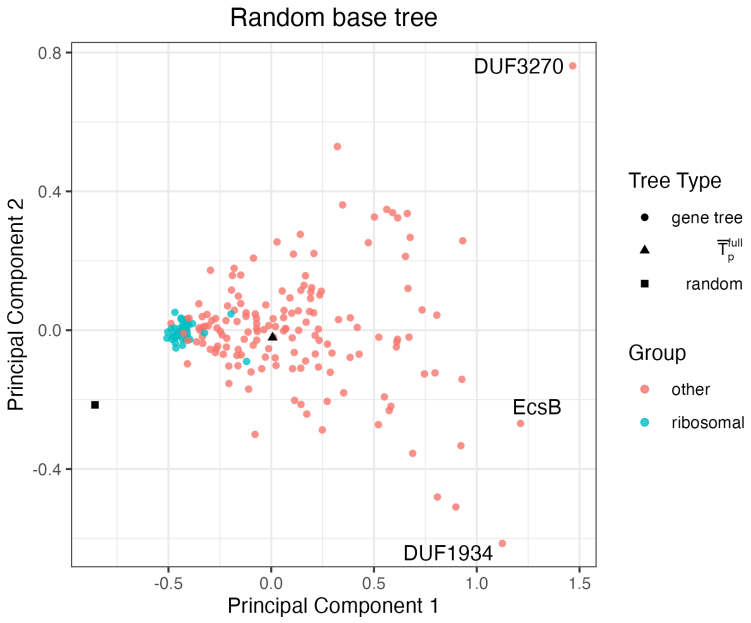


Figure A.8: The proposed visualization of 196 gene trees estimated from 106 genomes in the Streptococcus genus shown as a two-dimensional scatterplot. The base tree is a randomly generated tree with 106 tips. The points are colored by whether or not they represent a ribosomal gene.

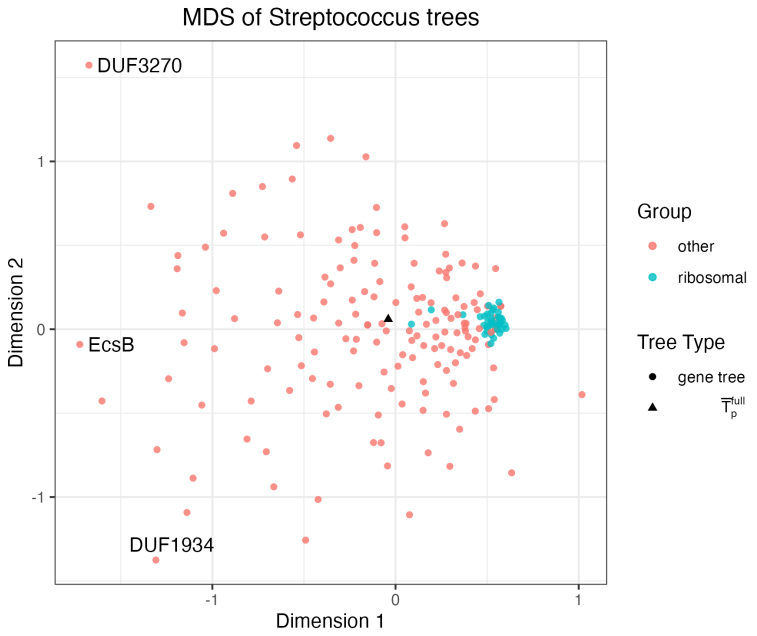


Figure A.9: A visualization of 196 gene trees constructed from 106 genomes from the *Streptococcus* genus using MDS of BHV distances between trees, depicted by a two-dimensional scatterplot. The points are colored by whether or not they represent a ribosomal gene.

A.2.3 tSNE and UMAP visualizations

As in the Prevotella exploration, we also create visualizations of the Streptococcus genomes by performing t-SNE and UMAP on the log map vectors. We use four initial seeds for these visualizations. The t-SNE visualizations can be seen in Figure A.10 and the UMAP visualizations in Figure A.11. Both sets of trees show the ribosomal trees clustered together in one area on the outer area of the plot, the phylogenomic tree farther from the ribosomal trees and towards the center of the cluster of trees, and the points representing genes EcsB, DUF1934, and DUF3270 on the edges of the plot. However, the overall pattern of points in the plots are different. Three of the t-SNE plots have points spread out throughout the two axes while one has more of a diagonal band of points. Conversely, three of the UMAP plots have clearly defined diagonal bands of points while one has the points more spread out throughout the two axes. This exploration provides an example in which the four different dimension reduction techniques (metric MDS of BHV distances and PCA, t-SNE, and UMAP on log map vectors) provide similar insights, although the points themselves are positioned differently.

A.2.4 Classification using log map vectors

Once the log map transformation is used to represent a set of trees with m tips as vectors in $\mathbb{R}^{2m+3}$, PCA is not the only method that can be used to explore the set of data. Consider the Streptococcus dataset, which includes 38 ribosomal genes and 158 non-ribosomal genes. If instead only a fraction of the genes were labeled as ribosomal or non-ribosomal, and a goal of the analysis was to predict the class of the non-labeled genes, this could be addressed with a supervised method. To illustrate this idea, we use Fisher's discriminant analysis and Random Forests to construct classifiers for the log map vectors of the gene trees, using the log map vectors from 148 trees with class labels as the training set and the log map vectors from 49 trees with hidden class labels as the test set.

We run discriminant analysis with the R function `lda` from the package MASS [Venables

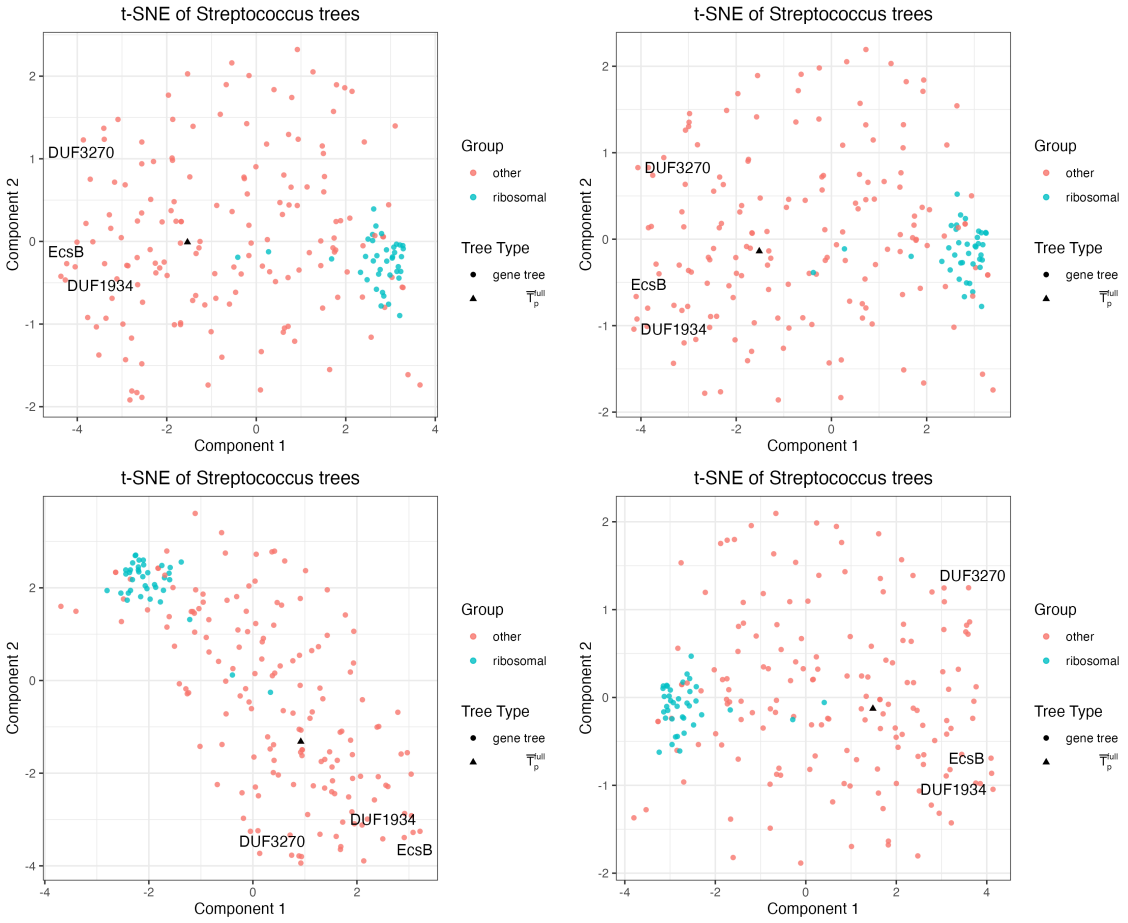


Figure A.10: A visualization of 196 gene trees constructed from 106 genomes from the *Streptococcus* genus, depicted by a two-dimensional scatterplot of the first two dimensions from t-SNE. Each subplot is constructed using a different initial seed. The points are colored by whether or not they represent a ribosomal gene.

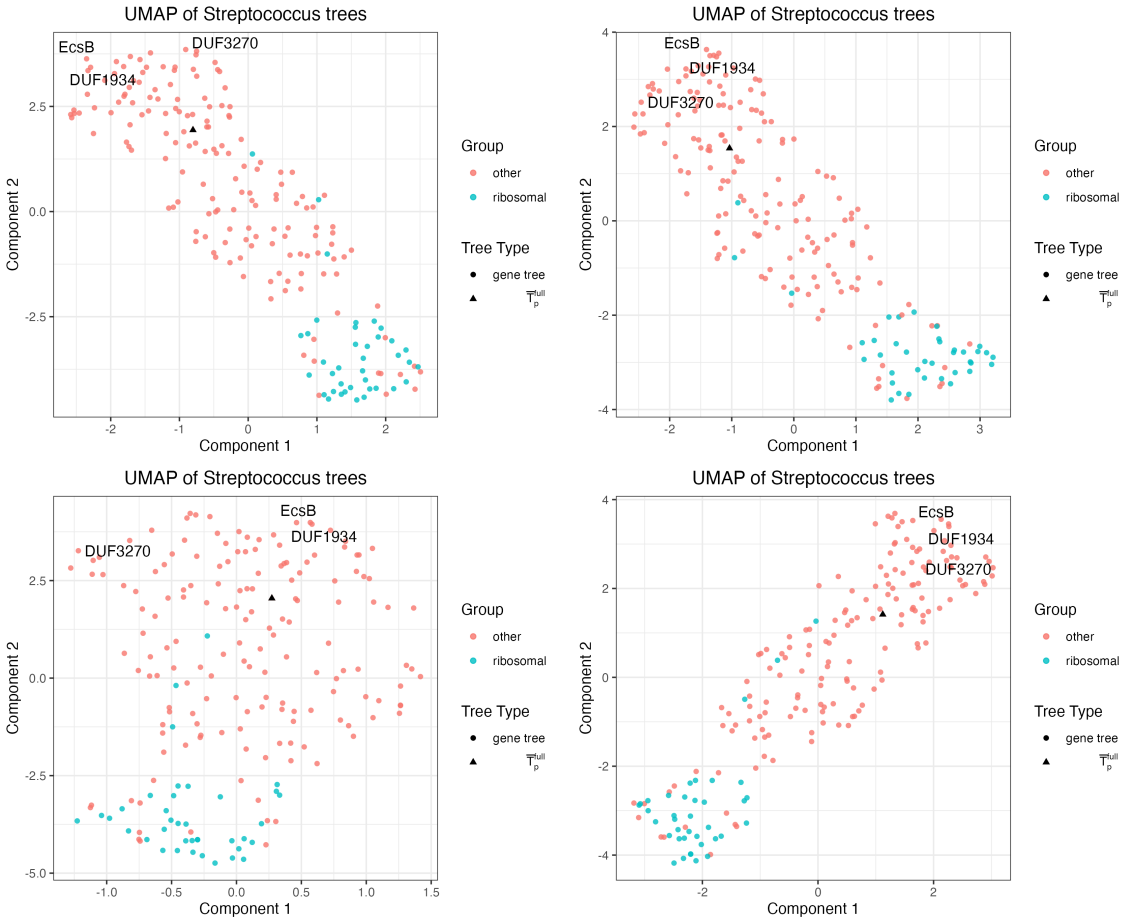


Figure A.11: A visualization of 196 gene trees constructed from 106 genomes from the *Streptococcus* genus, depicted by a two-dimensional scatterplot of the first two dimensions from UMAP. Each subplot is constructed using a different initial seed. The points are colored by whether or not they represent a ribosomal gene.

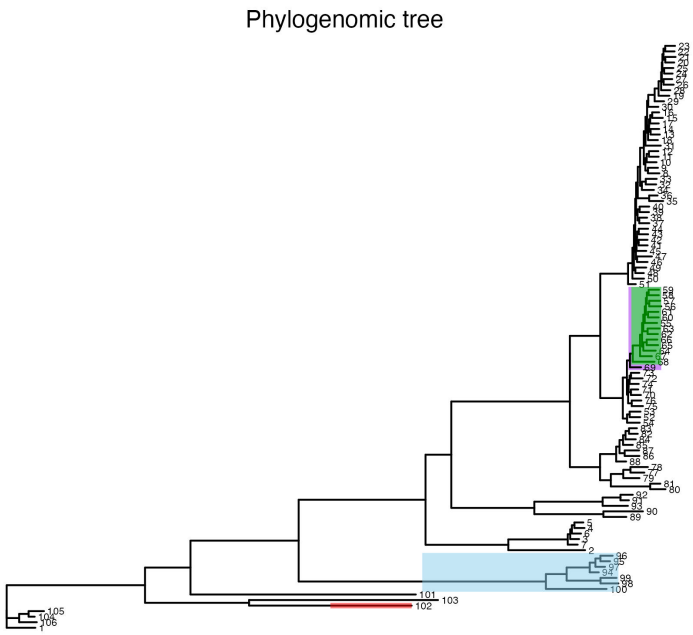


Figure A.12: A phylogenomic tree for 106 genomes from the *Streptococcus* genus. The edges separating colored clades from the rest of the tree represent the edges with the highest variable importances in the Random Forest used to predict whether a gene tree is ribosomal or not.

and Ripley, 2002]. This classifier has a prediction success rate of 63% on the test set. We also do classification with the R function `randomForest` from the package of the same name [Liaw and Wiener, 2002], which has a prediction success rate of 94% on the test set. Additionally, we are able to look at variable importance measures from our Random Forest to see which log map vectors have the largest impact on the classification algorithm. The most influential vectors are those that correspond with edges 100, 88, 205, and 57 in the phylogenomic tree used as base tree. These edges are shown in green, purple, red, and blue respectively in Figure A.12.

A.3 Anomaly detection in a different Prevotella analysis

In Section 2.3 we illustrated the proposed approach by investigating the robustness of the phylogenomic tree estimate to genes with apparently outlying phylogenies, and to the functional constraints on genes used to estimate the phylogenomic tree. In addition, our tool can also be used to detect anomalies in upstream processing of microbial sequence data. For example, when applying this tool to a set of 63 Prevotella genomes (a distinct set of genomes from those analyzed in Section 2.3.1), we found that the phylogenomic tree itself appeared as an outlier, and had a median gene tree support value of 0.00 (Figure A.13a). We could not explain this highly unusual finding until we re-ran our entire analysis and this time estimated a dramatically different phylogenomic tree and log map visualization (Figure A.13b). The second analysis showed a much more typical placement of the phylogenomic tree amongst the gene trees, and the new phylogenomic tree had a median gene tree support value of 0.40. IQTREE2's search for the maximum likelihood tree is stochastic, and in this case failed to produce a good estimate on the first attempt with the settings that we used.

We believe that the choice of base tree will have a relatively small impact on our proposed visualization in most settings, because of the similarity in topology between all trees in an analysis. This is not the case in this example because the base tree used in the first estimation run has a notably different topology from the gene trees in the analysis (this can be seen from the fact that the mean RF distance from trees in the analysis for the base tree in the first estimation run is 108, versus 66 in the second estimation run). This example demonstrates a situation in which the visualization is sensitive to base tree choice and there is a greater than usual difference in topology between the base tree and set of gene trees, unlike the situations shown in Appendix Sections A.1 and A.2.

This exploration highlights how our tool can be additionally useful for diagnosing anomalies in upstream bioinformatic and statistical analyses. We can also see the effect of choosing a base tree that has a large difference in topology from the set of gene trees. It is notable that the BacA gene is an apparent outlier in the set of trees from two distinct analyses of

Prevotella genomes.

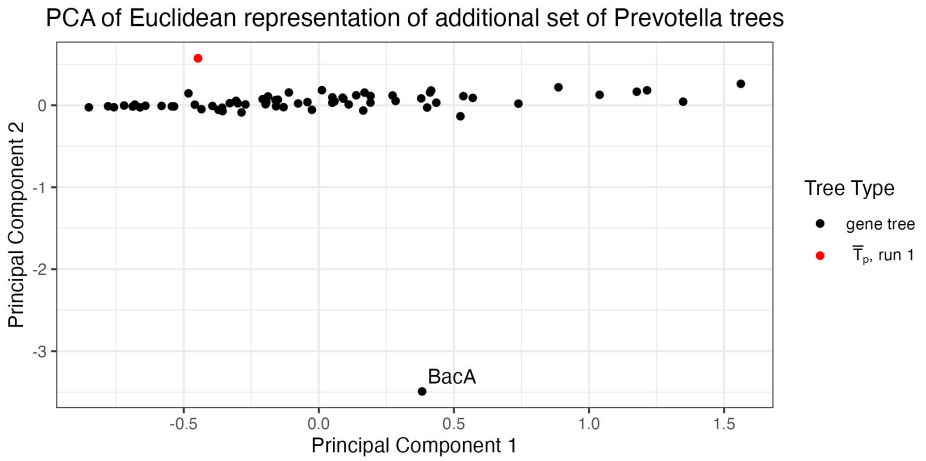
A.4 Taxonomy for genomes used in Section 2.3.1

A table matching tip numbers used in section 2.3.1 with NCBI accession numbers, taxonomic information, sequencing technology, assembly method, and information about sample sites can be found at `https://github.com/statdivlab/groves_supplementary/blob/main/data/prevotella/prevotella-tip-numbers-to-taxonomy.csv`.

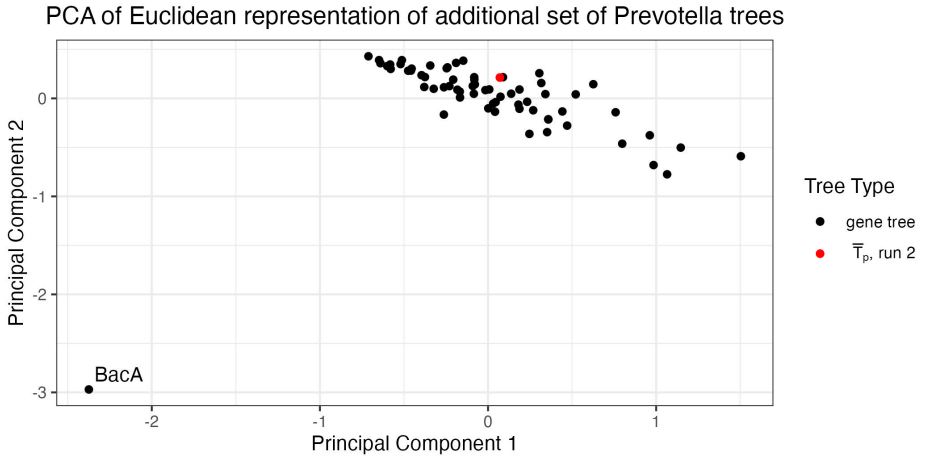
A.5 Software used to generate and analyze data in Section 2.3

Table A.1: Software versions use to generate and analyze the data used in Section 2.3. R packages are denoted with (R).

Sofware	Version
GToTree	1.5.51
HMMER3	3.3.2
Muscle	3.8.1551
TrimAl	1.4.rev15
Prodigal	2.6.3
FastTree 2	2.1.10
GTDB	R202
groves (R)	0.0.0.9000
ggtree (R)	3.4.2
tidyverse (R)	1.3.2
ape (R)	5.6-2
phangorn (R)	2.10.0
latex2exp (R)	0.9.5



(a) First estimation run



(b) Second estimation run

Figure A.13: The proposed visualization of 65 gene trees constructed from 63 genomes from the Prevotella genus and a phylogenomic tree constructed using all genes. The two phylogenomic trees shown in red in both the top and bottom panels were estimated from two different runs of IQTREE2. The placement of the estimated phylogenomic tree on the visualization differs substantially between the two runs.

Appendix B

SCALABLE DIFFERENTIAL ABUNDANCE ANALYSIS FOR MICROBIAL METAGENOMICS

B.1 Pseudo-Huber loss function

Clausen and Willis use the pseudo-Huber loss $g_p(\cdot)$ as the suggested constraint in their method.

$$g_p(\beta_k) = \text{argmin}_c \sum_{j=1}^{J} h_\delta(\beta_k^j - c) \tag{B.1}$$

$$h_\delta(x) = \delta^2 \left(\sqrt{1 + (x/\delta)^2} - 1 \right) \tag{B.2}$$

for $\delta > 0$.

This function is smooth and approximates the median as $\delta \downarrow 0$ and approximates the mean as $\delta \to \infty$. Empirically, Clausen and Willis find that a value of $\delta = 0.1$ works well in practice.

B.2 Results from a parametric multinomial logistic regression setting

Following the approach of Clausen and Willis [2024], we do not assume a specific parameter model for our data. However, we do use a Poisson log likelihood with mean model given by (3.3) to motivate estimating equations. We then note that after profiling nuisance parameters z out of the Poisson log likelihood, it takes on the form of a multinomial logistic regression log likelihood. In this appendix, we do make the assumption that each count $Y_{ij} \sim$ Poisson, and therefore $Y_{ij} \big| \sum_{j'=1}^{J} Y_{ij'} \sim$ multinomial, and we show that we can use results about the multinomial distribution to derive the reduced models M_R.

Consider a random vector $Y_i \in \mathbb{R}^J$ that is distributed according to a multinomial distribution. Assume a multinomial logistic regression model in which the probabilities are connected to data and parameters such that,

$$(Y_{i1}, \ldots, Y_{iJ}) \sim \text{Multinomial}(S_i, \pi_i^1, \ldots, \pi_i^J) \tag{B.3}$$

$$\beta^1 = 0 \tag{B.4}$$

$$\pi_i^j = \frac{e^{X_i \beta^j}}{1 + \sum_{j'=2}^{J} e^{X_i \beta^{j'}}}, \; j \in \{1, \ldots, J\} \tag{B.5}$$

Then, there is a known form for the joint distribution of Y_i^{-J} when conditioning on $Y_{iJ} = y_{iJ}$ [McCullagh and Nelder, 1989]. We will apply this to get,

$$(Y_{i1}, \ldots, Y_{iJ-1} | Y_{iJ} = y_{iJ}) \sim \text{Multinomial}(S_i - y_{iJ}, \frac{\pi_i^1}{1 - \pi_i^J}, \ldots, \frac{\pi_i^{J-1}}{1 - \pi_i^J}) \tag{B.6}$$

$$\alpha^1 = 0 \tag{B.7}$$

$$\frac{\pi_i^j}{1 - \pi_i^J} = \frac{e^{X_i \alpha^j}}{1 + \sum_{j'=2}^{J-1} e^{X_i \alpha^{j'}}}, \; j \in \{1, \ldots, J-1\}. \tag{B.8}$$

Then, for all $j \in \{1, \ldots, J-1\}$, we can calculate the expression $\pi_i^j / (1 - \mu_i^J)$ using equation (B.5) in terms of β parameters, and set it equal to the right hand side of (B.8). This leads to the equality $\alpha_k^j - \alpha_k^1 = \beta_k^j - \beta_k^1$ for all k and all $j \in \{1, \ldots, J-1\}$. We can then generalize this result for any set $S_R \subset \{1, \ldots, J\}$.

B.3 Data generation in Type I error and power simulations

We consider sample sizes $n \in \{10, 50, 250\}$, total number of taxa $J \in \{50, 250\}$, and data that is drawn from Poisson and zero-inflated negative binomial (ZINB) distributions. We use a balanced design matrix that includes an intercept and a binary covariate.

For each J, we construct a $\beta \in \mathbb{R}^{2 \times J}$ matrix. As in Clausen and Willis [2024], we set the first row of β (corresponding to the intercept column of the design matrix) such that the odd elements are an evenly spaced sequence of length $J/2$ from -3 to 3 and the even elements are an evenly spaced sequence of length $J/2$ from 3 to -3. We set the second row

of β (corresponding to the covariate column of the design matrix) to be $5 \times \sinh(x)/\sinh(10)$, in which x is defined as a linearly increasing sequence from -10 to 10 with length J. In the Type I error and power simulations in Sections 3.3.2 and **??**, we test parameter values for $\beta_1^{125} - g_F^p(\beta_1^j : j \in S_{ref})$. We define S_{ref} to be a set of 24 categories that do not include category 125, contain log fold-differences that range across much of the range of β_1 values, and have smoothed median $g_F^p(\beta_1^j : j \in S_{ref}) = 0$ In the Type I error simulations, $\beta_1^{125} = 0$ and $g_F^p(\beta_1^j : j \in S_{ref}) = 0$, so $\beta_1^{125} - g_F^p(\beta_1^j : j \in S_{ref}) = 0$. In the power simulations, $\beta_1^{125} = b$, so $\beta_1^{125} - g_F^p(\beta_1^j : j \in S_{ref}) = b$.

Each time that our data is generated, we draw n values independently from a Normal distribution to use as the z vector. Once we have X, β, and z, we can compute means of the form $\mu_{ij} = exp(X_i\beta^j + z_i)$ for $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, J\}$. We draw each Y_{ij} value independently from either a Poisson distribution or a zero-inflated negative binomial (ZINB) distribution. The Poisson distribution is fully parameterized by the mean model. In order to draw from the ZINB distribution, we first draw from a Negative Binomial distribution with mean $\mu_{ij}/0.4$ and dispersion parameter $\phi = 5/(0.4^2)$ (this has a variance of $\mu_{ij}/0.4 + \mu_{ij}^2/5$). We then multiply this with an independent draw from a Bernoulli distribution with probabilty of success 0.4. This corresponds to a set of data with roughly 60% zeroes.

B.4 Proof of Proposition 4

We will assume that no taxa have exactly the same parameter magnitudes $|\beta_k^j - g_F^p(\beta_k)$, and therefore there is a unique known reference set S_{kr} of size J_r for $S_{kr} \equiv \{j : rank(|\beta_k^j - g_F^p(\beta_k)|) \leq J_r)$. We will also assume that for all data realizations, our estimates $|\hat{\beta}_k^j - g_F^p(\hat{\beta}_k^{j'} : j' \in S_{dd})| \neq |\hat{\beta}_k^{j''} - g_F^p(\hat{\beta}_k^{j'} : j' \in S_{dd})|$ for all pairs of taxa $1 \leq j < j'' \leq J$.

Proposition 4: *When we choose a data-driven reference set S_{dd} to be the J_r taxa with the smallest values of $|\hat{\beta}_k^j - g_F^p(\hat{\beta}_k)|$, then $Pr(S_{dd} = S_{kr}) \to 1$ as $n \to \infty$.*

Proof: A sufficient condition for choosing S_{dd} to be equal to S_{kr} is accurately estimating the complete ordering of taxa j based on values of $|\beta_k^j - g_F^p(\beta_k)$ using estimates $\hat{\beta}_k^j - g_F^p(\hat{\beta}_k)$. A sufficient condition for this accurate complete ordering is accurate pairwise orderings for

every pair of taxa $j, j' \in \{1, \ldots, J\}$. Therefore, a proof that $\lim_{n\to\infty} Pr(sign(|\hat{\beta}_k^j - g(\hat{\beta}_k)| - |\hat{\beta}_k^{j'} - g(\hat{\beta}_k)|) = sign(|\beta_k^j - g(\beta_k)| - |\beta_k^{j'} - g(\beta_k)|) = 1$ will prove the proposition.

For a given k, define $\delta_{jj'k} = |\beta_k^j - g(\beta_k)| - |\beta_k^{j'} - g(\beta_k)|$, for pairs of taxa $j \neq j'$. Then, define δ_m as the minimum $\delta_{jj'}$ value across all pairs j, j'. We know that $\delta_m > 0$ due to our assumption that no taxa have exactly the same parameter magnitudes.

We know that our MLEs under the alternative hypothesis are consistent for the true parameter values, so that $\hat{\beta}_k^j - g_F^p(\hat{\beta}_k) \xrightarrow{p} \beta_k^j - g_F^p(\beta_k)$ for all taxa j. We will apply the continuous mapping theorem to the absolute value function $g(x) = |x|$ to argue that $|\hat{\beta}_k^j - g_F^p(\hat{\beta}_k)| \xrightarrow{p} |\beta_k^j - g_F^p(\beta_k)|$ for all taxa j. Then, we can directly apply the definition of convergence in probability to argue that $|\hat{\beta}_k^j - g_F^p(\hat{\beta}_k)| - |\hat{\beta}_k^{j'} - g_F^p(\hat{\beta}_k)| \xrightarrow{p} |\beta_k^j - g_F^p(\beta_k)| - |\beta_k^{j'} - g_F^p(\beta_k)|$ for all pairs of taxa $j \neq j'$.

Now, we will return to $\lim_{n\to\infty} Pr(sign(|\hat{\beta}_k^j - g_F^p(\hat{\beta}_k)| - |\hat{\beta}_k^{j'} - g_F^p(\hat{\beta}_k)|) = sign(|\beta_k^j - g_F^p(\beta_k)| - |\beta_k^{j'} - g_F^p(\beta_k)|)$. In order to estimate $sign(|\hat{\beta}_k^j - g_F^p(\hat{\beta}_k)| - |\hat{\beta}_k^{j'} - g_F^p(\hat{\beta}_k)|)$ that does not match $sign(|\beta_k^j - g_F^p(\beta_k)| - |\beta_k^{j'} - g_F^p(\beta_k)|)$, we will need the difference in the estimates, $||\hat{\beta}_k^j - g_F^p(\hat{\beta}_k)| - |\hat{\beta}_k^{j'} - g_F^p(\hat{\beta}_k)||$, to be at least as large as δ_m. This is because in order to estimate the wrong sign, we will need the estimation error to be larger than the true parameter difference, which is bounded below by δ_m.

$$\lim_{n\to\infty} Pr(sign(|\hat{\beta}_k^j - g_F^p(\hat{\beta}_k)| - |\hat{\beta}_k^{j'} - g_F^p(\hat{\beta}_k)|) \neq sign(|\beta_k^j - g_F^p(\beta_k)| - |\beta_k^{j'} - g_F^p(\beta_k)|) \quad \text{(B.9)}$$

$$\leq \lim_{n\to\infty} Pr(||\hat{\beta}_k^j - g_F^p(\hat{\beta}_k)| - |\hat{\beta}_k^{j'} - g_F^p(\hat{\beta}_k)|) - \Big(|\beta_k^j - g_F^p(\beta_k)| - |\beta_k^{j'} - g_F^p(\beta_k)|\Big)| \geq \delta_m) \quad \text{(B.10)}$$

$$= 0 \quad \text{(B.11)}$$

The final line is true due to the convergence in probability result above and the definition of convergence in probability.

B.5 Investigation of reference set comparison Type I error and power simulations

In Section 3.4.1, we compare the performance of robust score tests with Clausen and Willis' recommended constraint $g_F^p(\beta_k)$ over all taxa to that of the robust score tests with difference reference sets, across simulation settings and null and specific alternative hypotheses. We find that under a null hypothesis such that both $\beta_k^j - g_F^p(\beta_k) = 0$ and $\beta_k^j - g_F^p(\beta_k^j : j \in S_{kr}) = 0$, all robust score testing methods have p-value distributions that approximation uniform distributions, and control the Type I error rate. In power simulations, we found that in all settings and across most specific alternate hypothesis parameter values b, the approaches using all taxa and reference sets S_{kr} and S_{dd} have higher power than the approaches using reference sets S_{ss} and S_{th} from sample splitting and thinning. More surprisingly, we found that for ZINB data with $n \in \{25, 50\}$, the approaches with reference sets S_{kr} and S_{dd} have higher power than the approach with all taxa for moderate to large values of b. We investigate this behavior further in this section by showing additional simulation results and comparing robust score test statistic distributions across simulation settings.

We start with by returning to the Type I error rate simulations. In Figure B.1, we can see that the three approaches, robust score tests with the full model and a parameter involving all taxa and robust score tests with reduced models and parameters involving reference sets S_{kr} and S_{dd} all have very similar distributions under the null hypothesis. This aligns with our conclusions from Figure 3.3 in Section 3.4.1.

Next, we consider the power simulations. In these figures, we add two additional approaches for comparison: robust score tests with the full model using reference sets S_{kr} and S_{dd}. In Figure B.2, we can see that the robust score test statistic distributions are very similar between all approaches using reference sets S_{kr} and S_{dd}, and are different from the approach using all taxa. In this figure, results are aggregated over specific alternate hypothesis parameter values b from the set $[0.25, 5]$. We investigate these test statistic distributions further by considering the cases with $n = 250$ and stratified by alternate hypoteshis parameter value b.

Figures B.3 and B.4 show these results for Poisson data and ZINB data respectively. In each of these figures, all test statistic distributions are very similar for $b \leq 2$. For larger values of b, the distributions start to diverge. In these settings, the test statistic distribution densities for approaches using reference sets S_{kr} and S_{dd} are shifted to the right of the density for the approach using the full model and all taxa. This aligns with our conclusions from Figure 3.4 that in some simulation settings, approaches with reference sets S_{kr} and S_{dd} have higher power than the full model with all taxa for moderate to large alternate hypothesis values b. The major difference between the conclusions from Poisson data and ZINB data are that for Poisson data, the distributions become close together again for $b \in \{4.75, 5\}$, and for ZINB data, the distributions remain more separated for these b values.

Finally, we investigate a potential mechanism for these different robust score test statistic distributions and the resulting differences in power in some settings. In a single parameter setting, a score test measures the gradient of the likelihood function, scaled by the curvature of the the likelihood function, evaluated at the maximum likelihood estimate under the null hypothesis. A more curved likelihood function indicates a higher variance of the score value evaluated at the parameter of interest, and less confidence that a large gradient represents a null hypothesized parameter value that is truly different from the actual parameter value. We hypothesize that for moderate to large values of b, the transformation of the score vector used in the robust score test statistic has a higher variance for the parameter that involves all taxa, compared to the parameter that involves a reference set of a small number of taxa.

The robust score test statistic has the form,

$$T_{RS} = \frac{n}{n-1} \cdot S_{H_0}^T I_{H_0}^{-1} F_{H_0}^T (F_{H_0} I_{H_0}^{-1} D_{H_0} I_{H_0}^{-1} F_{H_0}^T)^{-1} F_{H_0} I_{H_0}^{-1} S_{H_0}, \tag{B.12}$$

in which S_{H_0} is the score evaluated at the MLEs under the null, I_{H_0} is a consistent estimate of the information matrix under the null, D_{H_0} is a consistent estimate of the covariance matrix of the score equations under the null, and F_{H_0} is $\frac{\partial h}{\partial \beta^T}$ evaluated at the MLEs estimated under the null. We define $F_{H_0} I_{H_0}^{-1} D_{H_0} I_{H_0}^{-1} F_{H_0}^T$ as the "inside" portion of the robust score test statistic. This portion represents the estimated empirical covariance matrix of the score

vector, transformed by multiplying it on both sides by F_{H_0} and $F_{H_0}^T$. In Figures B.5 and B.6, we present boxplots of this "inside" portion of the robust score test statistic for simulated data with $n = 250$ from Poisson and ZINB distributions respectively. In these plots, we notice that for moderate to large values of b, the distribution of inside test statistic portions for the full model with all taxa are shifted towards larger values compared to the distributions for the approaches with reference sets. This larger inside test statistic portions correspond to larger transformations of the estimated empirical covariance of the score vector, and smaller robust score test statistics.

In this expanded investigation of Type I error rate and power simulation results, we confirm that under both the null and specific alternate hypotheses, the distribution of test statistics from appraoches with the known reference set S_{kr} and estimated reference set S_{dd} are nearly identical. This supports our argument that we can use tests involving parameter estimates $\hat{\beta}_k - g_F^P(\beta_k^{j'} : j' \in S_{dd})$ to test hypotheses of the form $H_0 : \beta_k^j - g_F^p(\beta_k^{j'} : j' \in S_{kr}) = 0$. Additionally, we find that across all simulation settings, for moderate to large values of b, the robust score test statistics are typically larger for approaches with reference sets S_{dd} and S_{kr}, compared to the approach with all taxa. This results in higher power for these approaches in some simulation settings. We hypothesize that these different test statistic distributions could be caused by different estimated empirical score vector variances, and show simulation results that support this idea.

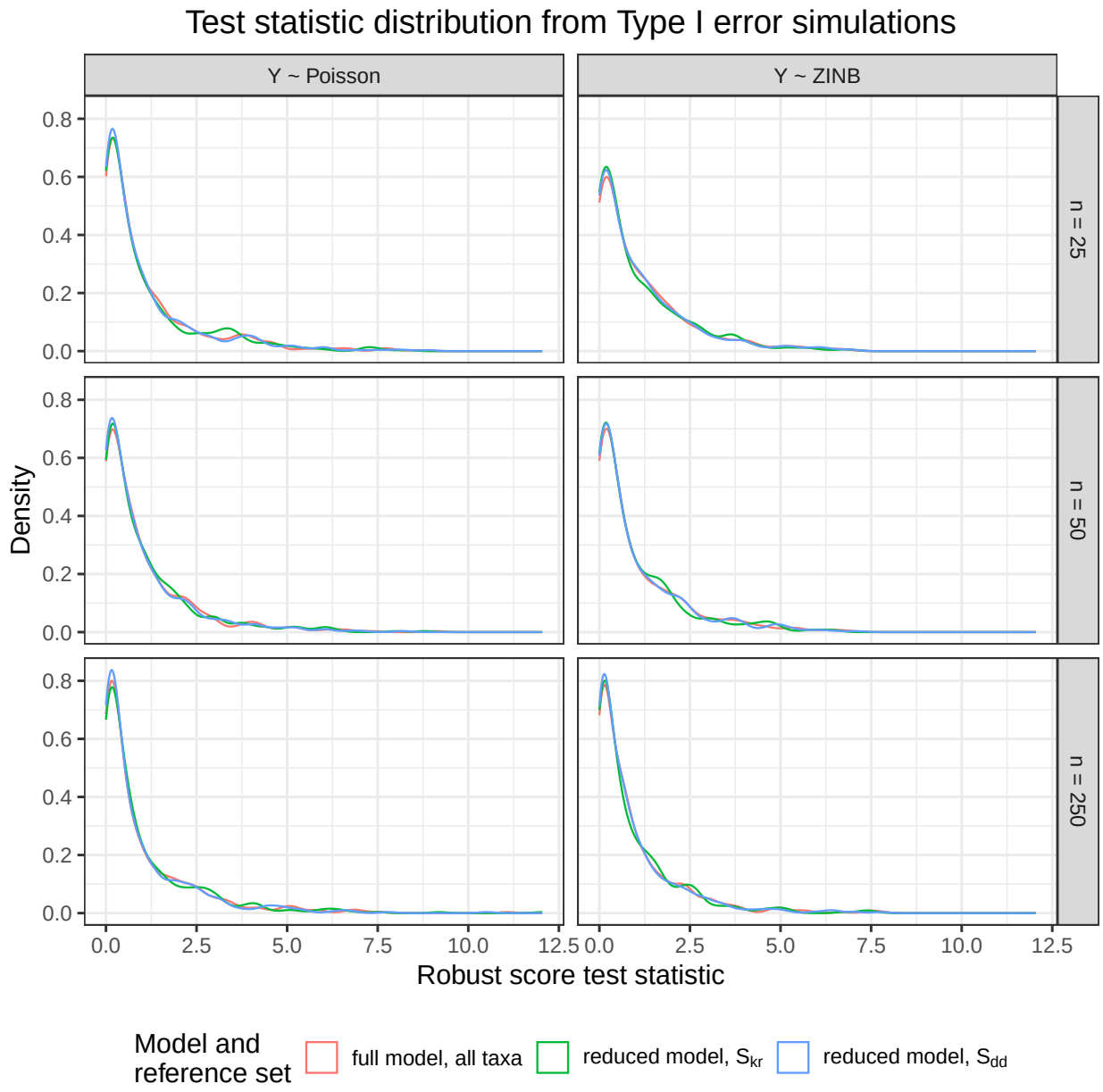


Figure B.1: Estimated densities of distributions of test statistics from Type I error simulations to compare robust score tests using parameters defined with the full set of taxa versus different reference sets. Results come from a simulation with 500 trials.

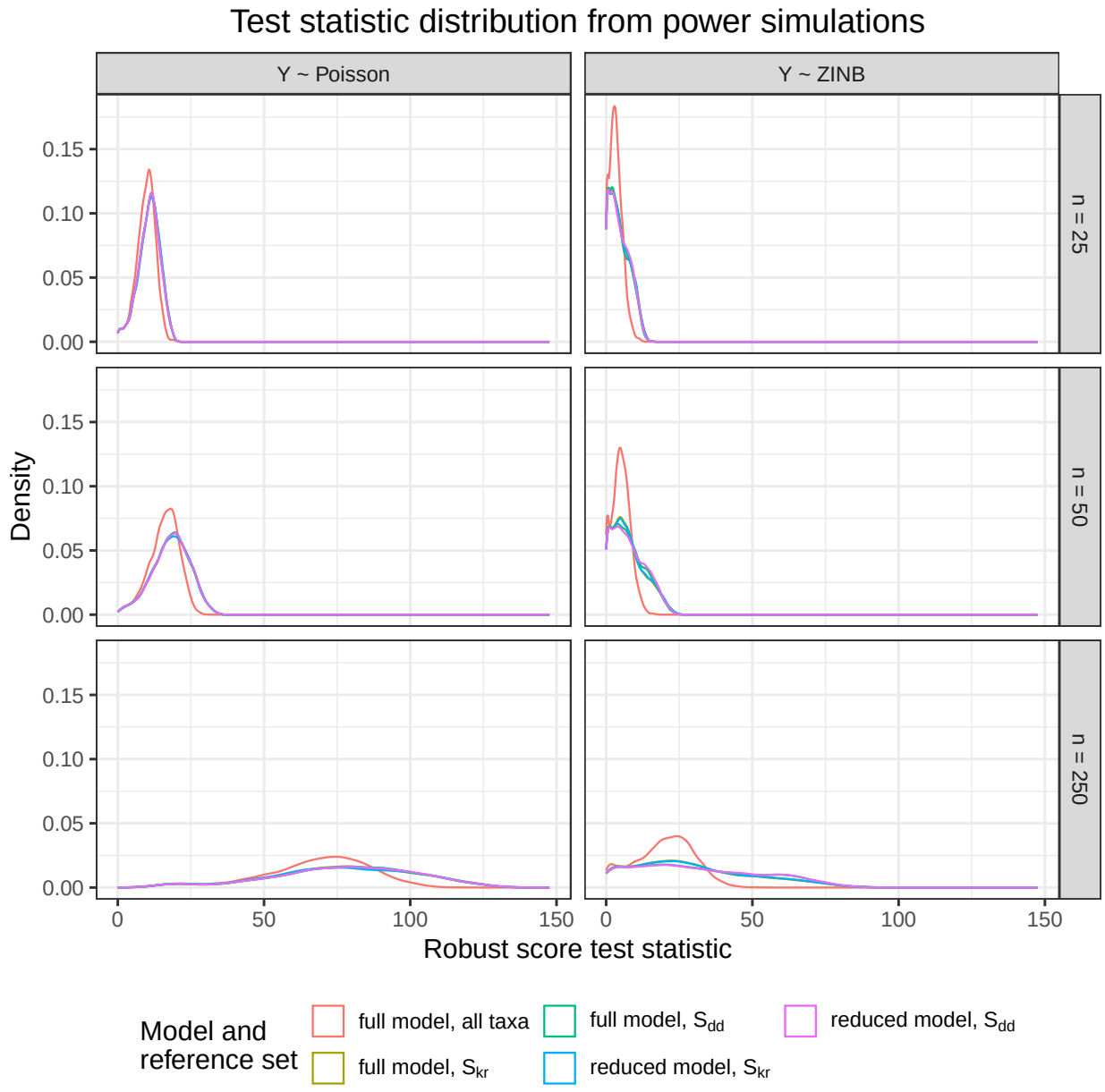


Figure B.2: Estimated densities of distributions of test statistics from power simulations to compare robust score tests using parameters defined with the full set of taxa versus different reference sets. In each setting, results are aggregated over specific alternative hypothesis parameter values in the set $[0.25, 5]$. Results come from a simulation with 500 trials.

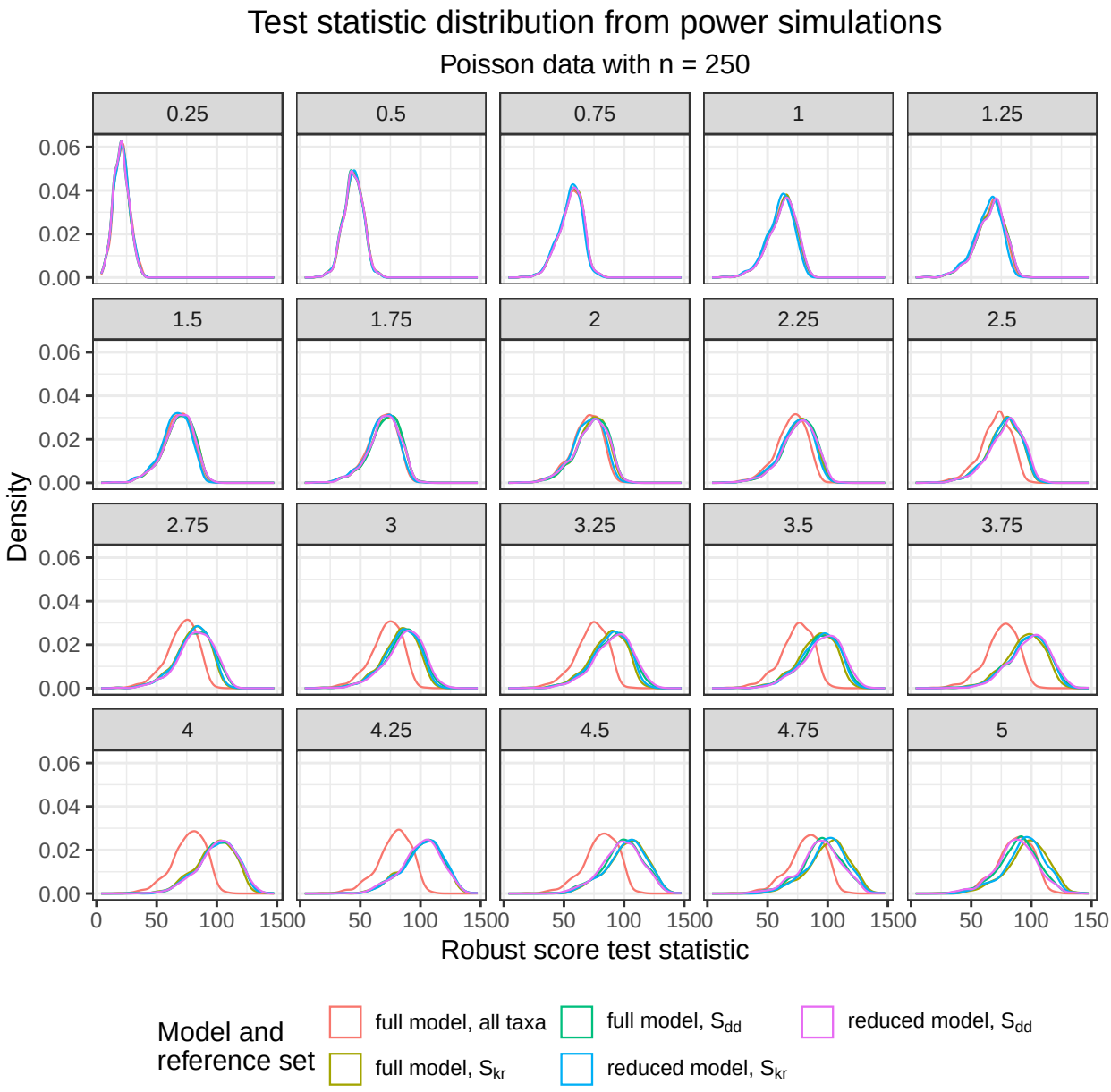


Figure B.3: Estimated densities of distributions of test statistics from power simulations to compare robust score tests using parameters defined with the full set of taxa versus different reference sets. Data are simulated from a Poisson distribution with $n = 250$ and results are stratified by specific alternate hypothesis parameter value. Results come from a simulation with 500 trials.

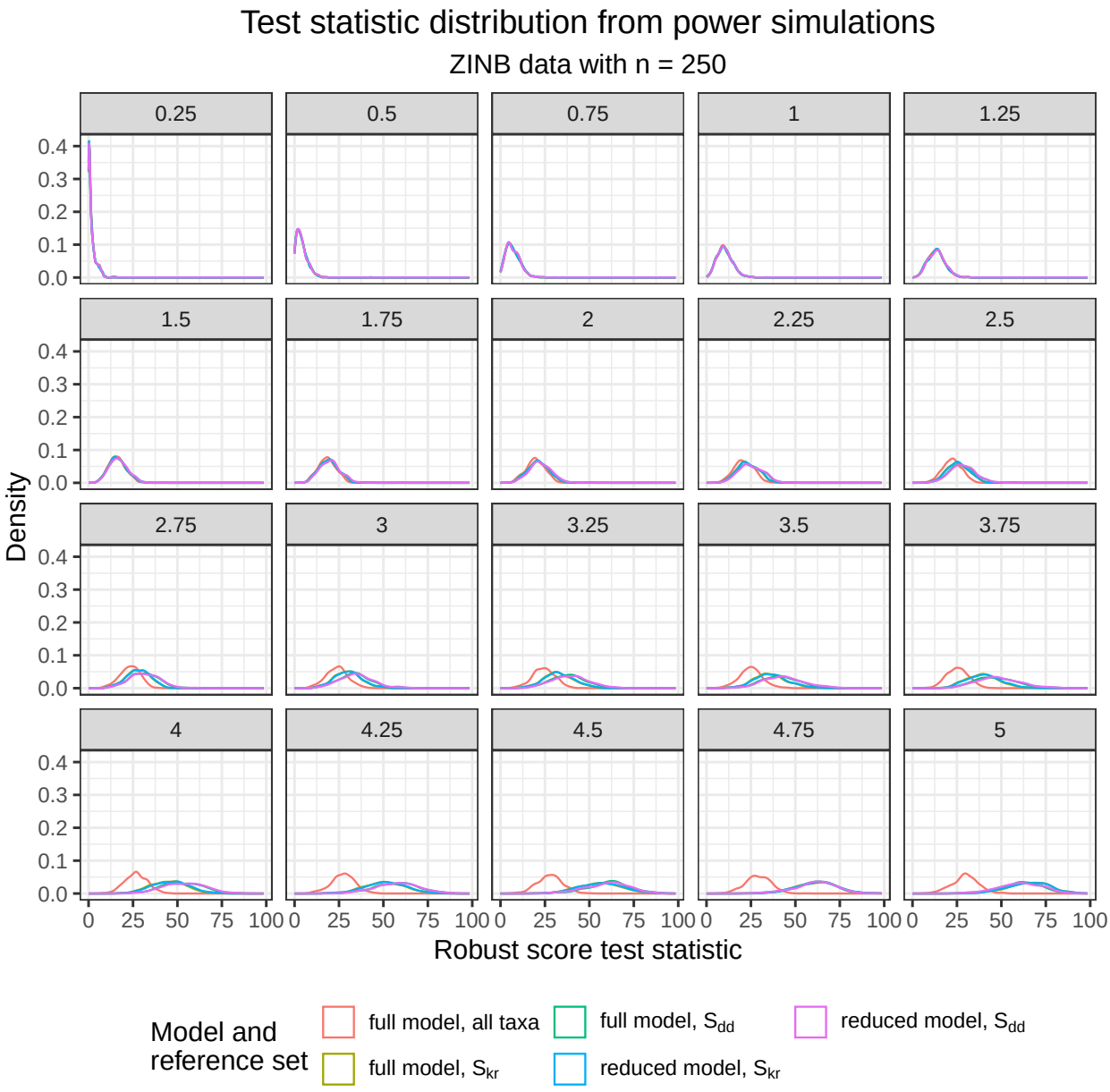


Figure B.4: Estimated densities of distributions of test statistics from power simulations to compare robust score tests using parameters defined with the full set of taxa versus different reference sets. Data are simulated from a zero-inflated negative binomial (ZINB) distribution with $n = 250$ and results are stratified by specific alternate hypothesis parameter value. Results come from a simulation with 500 trials.

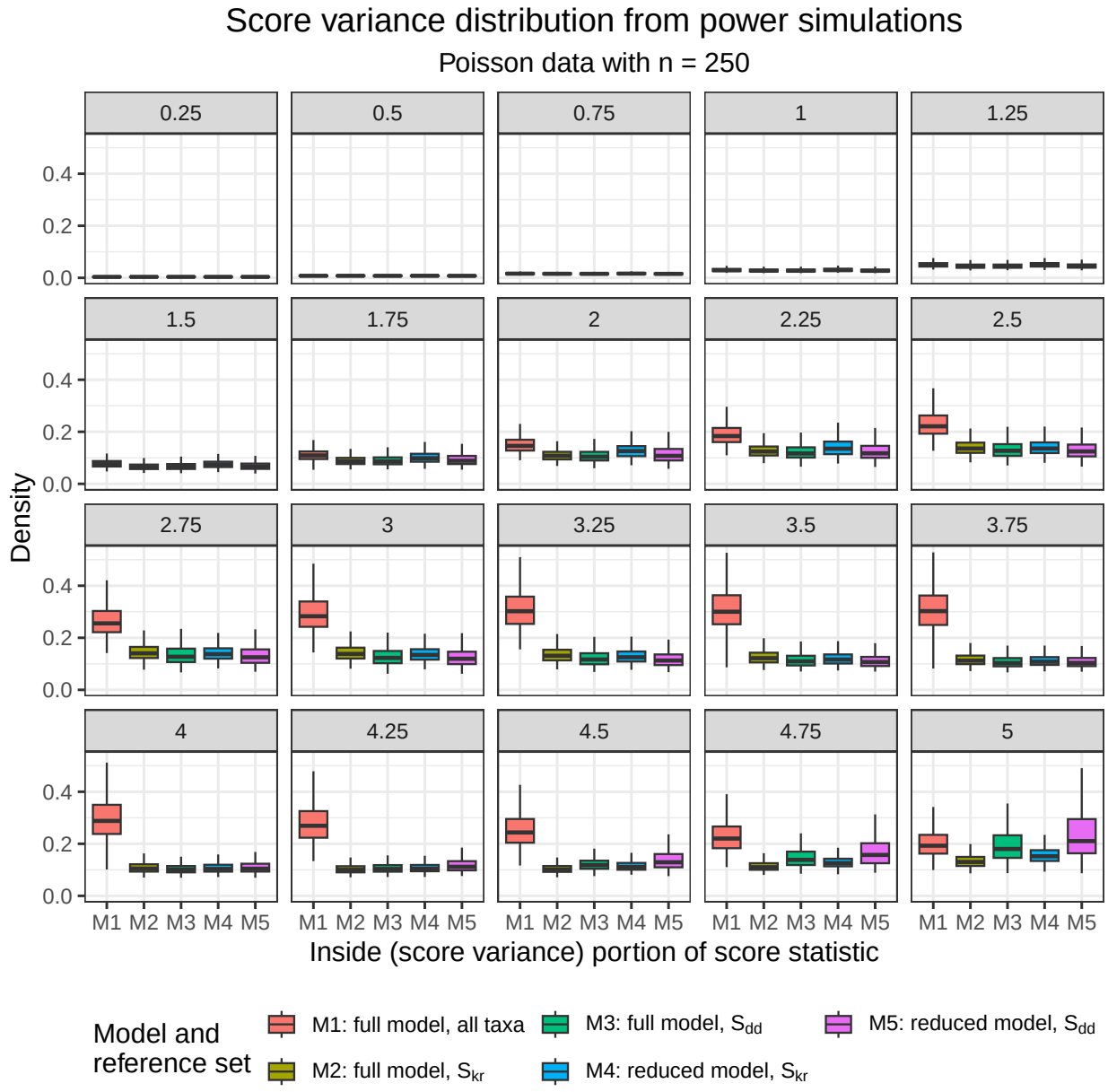


Figure B.5: Estimated densities of distributions of portions of test statistics from power simulations to compare robust score tests using parameters defined with the full set of taxa versus different reference sets. Specifically, "inside" values of the robust score test statistic are plotted, which correspond to the portion of the test statistic that estimates a transformation of the score vector covariance. Data are simulated from a Poisson distribution with $n = 250$ and results are stratified by specific alternate hypothesis parameter value. Results come from a simulation with 500 trials.

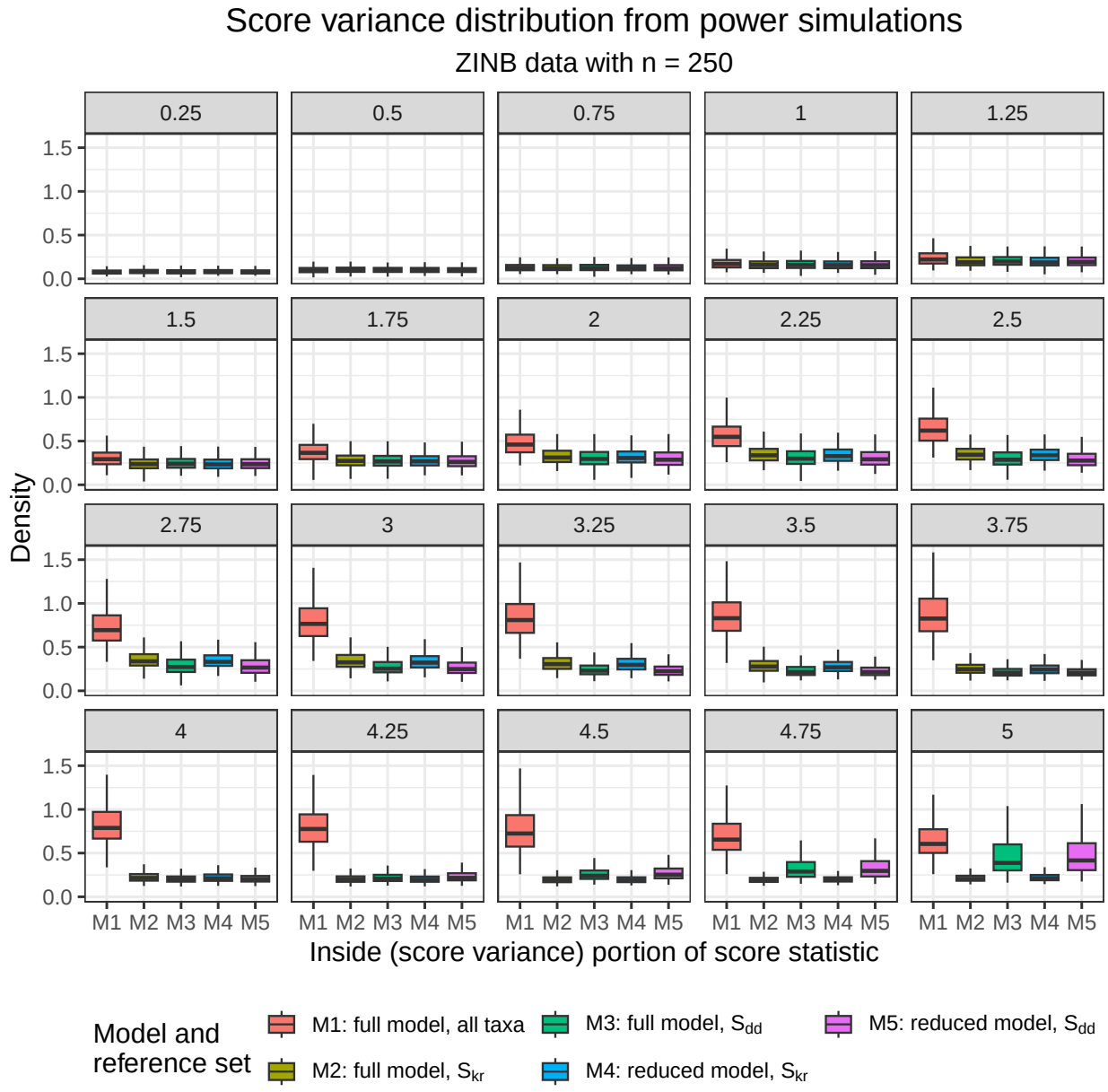


Figure B.6: Estimated densities of distributions of portions of test statistics from power simulations to compare robust score tests using parameters defined with the full set of taxa versus different reference sets. Specifically, "inside" values of the robust score test statistic are plotted, which correspond to the portion of the test statistic that estimates a transformation of the score vector covariance. Data are simulated from a ZINB distribution with $n = 250$ and results are stratified by specific alternate hypothesis parameter value. Results come from a simulation with 500 trials.

Appendix C

A MODEL OF FUNCTIONAL ABUNDANCE FROM METAGENOMIC DATA AND ITS IMPLICATIONS

C.1 Generation of α and γ parameters in simulation

Recall that the θ_0 and θ_1^m parameters are defined as,

$$\theta_0^m = \log \sum_{j \in J_m} e^{\theta_0^{jm}} \tag{C.1}$$

$$\theta_1^m = \log \frac{\sum_{j \in J_m} e^{\theta_0^{jm} + \theta_1^{jm}}}{\sum_{j \in J_m} e^{\theta_0^{jm}}}. \tag{C.2}$$

For the first 25 functions that represent necessary functions, we assume that $\theta_0^{jm} = \theta_0^j$. We randomly assign elements of the vector of evenly spaced values from -3 to 3 of length J to the vector θ_0^j. This represents the functions that each appear in a single copy in all organisms represented in the dataset. We randomly draw values from the vector of evenly spaced values from -3 to 3 of length $1,000$ for the twenty-five θ_0^{jm} values for each remaining function m. For an identifiable gene model, we need $g(\alpha) = 0$, where $\alpha_m = \log \sum_{j \in J_m} e^{\delta_j + \theta_0^{jm}}$. For $g(\cdot)$, we use the pseudo-Huber median function over the first twenty-five functions, as defined in Section B.1. In order to satisfy $g(\alpha) = 0$, we subtract $g(\alpha)$ from each θ_0^{jm} value.

For ease of Type I error rate simulations, we would like to have θ_1^m values such that $\theta_1^m - g(\theta_1) = \theta_1^m$. Therefore, we assign the θ_1 values as follows. We set the first twenty-five elements to $1 \times \sinh(x)/\sinh(10)$, in which x is a linearly increasing sequence from -10 to 10 of length 25. These represent the necessary functions, which we expect to have small log fold changes with respect to any covariate. We set the twenty-sixth element of θ_1 to b based on null hypothesis or specific alternative hypothesis, as we will test hypotheses based on β_1^{26} in our simulations. We then set the remainder of the θ_1 vector to be $5 \times \sinh(x)/\sinh(10)$,

in which x is defined as a linearly increasing sequence from -10 to 10 with length $J - 26$.

Once we have generated θ_0^{jm} values and θ_1^m values, we generate θ_1^{jm} values such that equation (22) holds. For each function m, we start by setting each θ_1^{jm} value to a randomly chosen value from the vector $5 \times \sinh(x)/\sinh(10)$, in which x is defined as a linearly increasing sequence from -10 to 10 with length $|J_m|$ Then, for j_L the last element of J_m, if $\sum_{j \in J_m} e^{\theta_0^{jm}} > \sum_{j \in J_m \setminus j_L} e^{\theta_0^{jm} + \theta_1^{jm}}$, we set $\theta_1^{j_L m} = \log(\sum_{j \in J_m} e^{\theta_0^{jm}} - \sum_{j \in J_m \setminus j_L} e^{\theta_0^{jm} + \theta_1^{jm}}) - \theta_0^{j_L m}$. If the inequality does not hold, then we subtract 0.1 from each θ_1^{jm} until it does, and then set $\theta_1^{j_L m}$ as described above. This causes the inital γ_m^{init} value for each function m to be

$$\gamma_m^{\text{init}} = \log \frac{\sum_{j \in J_m} e^{\delta^j + \theta_0^{jm} + \theta_k^{jm}}}{\sum_{j \in J_m} e^{\delta_j + \theta_0^{jm}}}. \tag{C.3}$$

In order to satisfy the constraint $g(\gamma) = 0$, we set each $\gamma_m = \gamma_m^{\text{init}} - g(\gamma_m^{\text{init}})$.

C.2 Additional data analysis information

C.2.1 List of ribosomal proteins used in the constraint

Protein	KO	Protein	KO
L1	K02863	S2	K02967
L2	K02886	S3	K02982
L3	K02906	S4	K02986
L5	K02931	S5	K02988
L6	K02933	S7	K02992
L10	K02864	S8	K02994
L11	K02867	S9	K02996
L13	K02933	S10	K02946
L14	K02874	S11	K02948
L15	K02876	S12	K02950
L18	K02881	S13	K02952
L22	K02890	S14	K02954
L23	K02892	S15	K02956
L24	K02895	S17	K02961
L29	K02904	S19	K02965
L30	K02907		

Table C.1: Ribosomal proteins and corresponding KO labels used for the constraint in data analyses.

C.2.2 List of ribosomal proteins used in the constraint in sensitivity analysis

Protein	KO
L2	K02886
L3	K02906
L4	K02926
L5	K02931
L6	K02933
L14	K02874
L15	K02876
L16	K02878
L18	K02881
L22	K02890
L24	K02895
S3	K02982
S8	K02994
S10	K02946
S17	K02961
S19	K02965

Table C.2: Ribosomal proteins and corresponding KO labels used for the constraint in sensitivity data analyses, chosen based on findings from Hug et al. [2016].

C.3 Supplemetary information for HMP data analysis

C.3.1 Comparison of robust score test and robust Wald test results for categories with separation in the HMP data analysis

In the analysis of differential abundance of KOs in the HMP dataset, the reduced model robust score tests result in $4,298$ categories (53%) with significant q-values. The robust Wald tests from radEmu result in $6,304$ categories (77%) that have significant q-values. The majority of the disagreement between the robust Wald tests and robust score tests comes from parameters for separated categories. A category is separated if it compares two levels of a covariate in which one level only includes samples with zero counts. This analysis includes $2,037$ parameters from separated categories, for which all Wald q-values are significant and only 394 score q-values (19%) are significant. Out of the 6115 non-separated parameters, 5924 of them (97%) have Wald q-values and score q-values that lead to the same significance conclusions.

In Figure C.1, we compare the ranks of q-values from the robust Wald tests and the robust score tests on all 8152 parameters. The points that are not shown in black represent parameters for separated categories. These points are colored by prevalence. The parameters from separated categories with small score p-value ranks tend to have higher prevalence for separated categories, suggesting true differential abundance and not a result of sparsity or incomplete sampling. In this data analysis, the Wald test detects any type of separation as a signal, while the score test only detects separation as a signal when there is more evidence for differential abundance in the form of a greater prevalence.

C.4 Supplemental analysis of functional differential abundance across ocean depths

The first dataset that we analyze is a metagenomic sequencing dataset collected from the Arctic Ocean with 24 samples and $10,196$ observed KOs. This dataset is part of the Tara Oceans project [Karsenti et al., 2011]. The covariate of interest is ocean depth, which we

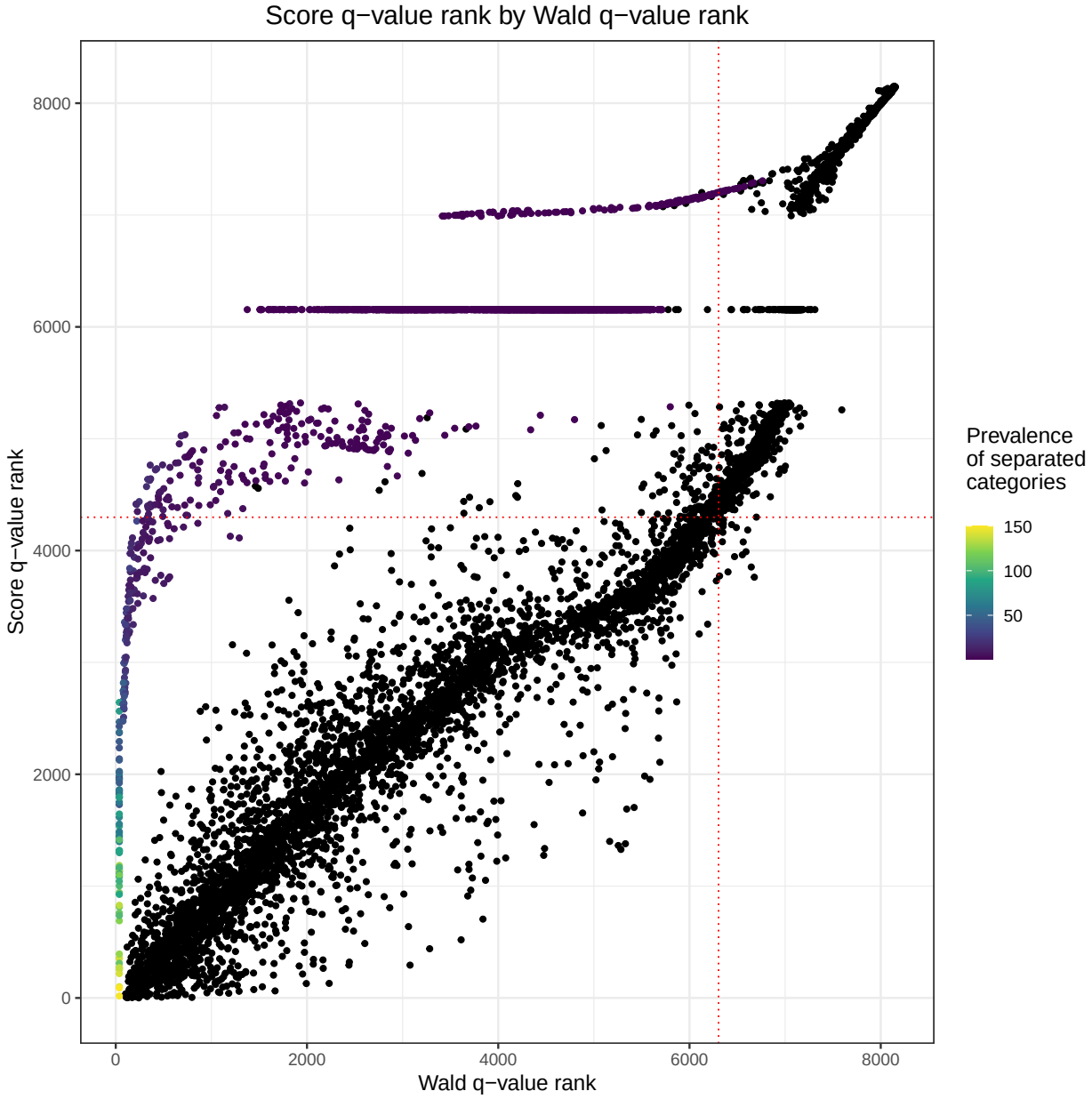


Figure C.1: A comparison of the ranks of q-values from robust Wald tests and robust score tests on 8,152 parameters in the HMP analysis. The points that are not black represent parameters for separated categories. A category is separated if it compares two levels of a covariate in which one only includes samples with zero counts. These points are colored by prevalence. Points to the left of the vertical red dotted line have Wald q-values less than 0.05 and points below the horizontal red dotted line have score q-values less than 0.05.

consider as a categorical variable with two levels. The baseline level is a deep ocean layer (8 samples), which we compare to the surface water layer (16 samples).

We define the differential abundance parameters in our functional abundance model as $\gamma_m - g(\gamma)$ for $m \in \{1, \ldots, M\}$ and for $g(\cdot)$ the pseudo-Huber loss over γ_m parameter values for a reference set of 31 ribosomal protein KOs (a list of these proteins and KOs is given in Table C.1 in Appendix C.2.1. The $\hat{\gamma}_m - g(\hat{\gamma})$ estimates range from -0.15 to 0.13 for the reference set and from -8.47 to 6.53 across all $m \in \{1, \ldots, M\}$. The estimates from the reference set all have small magnitudes compared to the variation in the dataset and similar magnitudes to each other, which makes this a good choice for a reference set in this analysis.

In order to provide another comparison between Clausen and Willis' robust score tests on full models and our robust score tests on reduced models introduced in Chapter 3, we run robust score tests using the full on model on all categories for a randomly chosen subset of 102 KOs (approximately 1% of the dataset). Within this subset, the robust score test p-values from the full model and the reduced model have a Pearson correlation of 1.000. The robust score tests can be run $5,875$ times faster with the reduced model than with the full model on average.

We correct for multiple testing by computing q-values for all parameters [Storey, 2002], using the p-values from the robust score tests on reduced models. The smallest q-value in this set is 0.89, and no parameters are considered significant. As a sensitivity analysis for our reference set choice, we also run our reduced model robust score tests using a different reference set of 16 KOs that correspond to ribosomal proteins used to build a new multidomain phylogeny in [Hug et al., 2016]. This smaller constraint set is nearly a subset of the 31 KOs that we use for our original reference set, and the full list of KOs and associated proteins can be found in Appendix C.2.2. There is a Pearson correlation of 0.996 between p-values from the robust score tests that use each reference set.

We do not run Clausen and Willis' robust Wald test on this dataset because it fails to control the Type I error rate in small sample sizes in our Type I error simulations in Chapter 3. We do run differential abundance analyses of this dataset with ALDEx2 [Fernandes et al.,

2013] and ANCOM-BC2 [Lin and Peddada, 2024]. We run ALDEx2 with a CLR transformation and 128 Monte Carlo iterations. Because ALDEx2 only accepts non-negative counts, we round all coverages to the nearest integer. We run ANCOM-BC2 without prevalence filtering or structural zero detection and we use their sensitivity analysis to assess the impact of pseudocounts on the results for each taxon.

There is a Pearson correlation of 0.63 between parameter estimates from our model and from ALDEx2. Parameters are not estimated by ANCOM-BC2 for categories with separation. A category has separation if it compares two levels of a covariate in which one level only includes samples with zero counts. Ignoring the 702(7%) parameters associated with separated categories, there is a Pearson correlation of 0.78 between estimates from our model and from ANCOM-BC2. No parameters are significant from the ALDEx2 analysis using a q-value threshold of 0.05. There are 62 ($< 1\%$) parameters with ANCOM-BC2 q-values less than 0.05, and 12 ($< 1\%$) of them are still significant in the pseudocount sensitivity analysis. The 12 KOs detected as differentially abundant by ANCOM-BC2 with the pseudocount sensitivity analysis can be seen in Table C.3, along with the functions they perform, their log fold-difference estimates from ANCOM-BC2, and q-values calculated from p-values from ANCOM-BC2. The ANCOM-BC2 effect size estimates for these 12 KOs are all in the 95th quantile of ANCOM-BC2 estimated effect size magnitudes within the dataset. Because we have not explored how taxon efficiencies affect ANCOM-BC2 effect sizes in simulation, we can investigate the $\hat{\gamma}_m - g(\hat{\gamma})$ estimates from our model for these categories. These twelve categories all have $|\hat{\gamma}_m - g(\hat{\gamma})| > 1.9$, which correspond to the 90th quantile of estimated effect size magnitudes in the dataset. If the effect of taxon efficiencies on this dataset is similar to the effect of taxon efficiencies in our simulations, most categories with effect sizes of this magnitude have non-zero values of $\theta_1^m - g(\theta)$. However, it is impossible to determine whether these categories with small ANCOM-BC2 p-values and relatively large estimated effect sizes are the result of biological fold-differences in abundance with respect to ocean depth, differential taxon efficiencies, or both.

It is unsurprising that our robust score tests and ALDEx2's inferential procedure do not

result in any statistically significant tests, as they are both conservative for small sample sizes in the Type I error rate simulations shown in Chapter 3. The small sample size of 24 in this analysis is not large enough to detect signals with small or medium magnitudes.

KO name	KO function	ANCOM-BC2 Estimate
K05781	putative phosphonate transport system ATP-binding protein	1.90
K18657	cell division protein ZapC	-2.77
K18968	diguanylate cyclase [EC:2.7.7.65]	-3.13
K18765	RNase E specificity factor CsrD	-3.12
K07084	putative amino acid transporter	-2.23
K24843	O2-independent ubiquinone biosynthesis accessory factor UbiT	-2.83
K07665	two-component system, OmpR family, copper resistance phosphate regulon response regulator CusR	-2.38
K12069	conjugal transfer pilus assembly protein TraA	-3.07
K12072	conjugative transfer pilus assembly protein TraH	-2.18
K19595	membrane fusion protein, gold/copper resistance efflux system	-2.87
K13652	AraC family transcriptional regulator	-1.27
K09909	uncharacterized protein	-2.75

Table C.3: KOs with significant q-values from the analysis of TARA data described in Section C.4. Estimates and p-values are computed using ANCOM-BC2 [Lin and Peddada, 2024], without prevalence filtering or structural zero detection and with a sensitivity analysis to screen for results affected by pseudocounts.