©Copyright 2022
David Clausen

Measurement Error in Microbiome Sequencing Experiments: Statistical and Scientific Considerations

David Clausen

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

Doctor of Philosophy

University of Washington

2022

Reading Committee:
Amy D. Willis, Chair
Scott S. Emerson
James Hughes

Program Authorized to Offer Degree:
Department of Biostatistics

University of Washington

Abstract

Measurement Error in
Microbiome Sequencing Experiments:
Statistical and Scientific Considerations

David Clausen

Chair of the Supervisory Committee:
Assistant Professor Amy D. Willis
Department of Biostatistics

Next-generation sequencing (NGS) methods have become an essential tool in the study of complex microbial communities known as microbiomes. Because their near ubiquity, such communities have been the focus of substantial research aiming to elucidate their structure and yield new insights into public health, medicine, and agriculture, among other fields. However, the relationship between the true composition of biological samples on which sequencing is performed and sequencing output is complex and only partially understood. As a result, it is often unclear to what extent experimental results in microbiome science reflect underlying biology rather than technical artifacts of a complex measurement process.

To address this uncertainty, we analyze a large NGS dataset generated by a multi-laboratory study of measurement error in microbiome sequencing data. We find, in replicate measurements on identical biological specimens, that distinctions between specimens apparent in measurements taken by one laboratory are not reliably resolved in measurements taken by others, with the degree of discordance varying with the taxonomic level and scale at which distinctions are made. Hence, our finding suggests that comparisons across groups in microbiome studies may not dependably reflect biology.

We next present a statistical model appropriate for NGS data subject both to detection effects – multiplicative over- and under-detection of microbial taxa relative to their true

abundances – and to potential contamination by taxa not present in specimens of interest. Our model uses experimental covariates and measurements on communities of known composition (also called positive controls) to estimate community composition in specimens of interest as well as detection effects and the form and intensity of contamination. We show via analysis of real datasets as well as through simulation that this model substantially outperforms standard estimators of microbial relative abundance in data subject to detection effects and contamination. In particular, we demonstrate that our model can exploit the structure of dilution series experiments to accurately identify contamination, even in the absence of positive control measurements. However, the same is not true for detection effects, which in general can only be estimated among microbial taxa present in communities of known composition.

To address this limitation, we develop a log-linear model to estimate means of outcomes observed up to unknown sample-specific scalings and subject to detection effects, taking as our motivating example estimation on the basis of NGS data of differences in log mean microbial cell concentrations across covariates of interest. The presence of unknown scalings renders our estimand only partially identiable. We address this by imposing simple constraints, which may be modified to suit differing scientific contexts. We validate this model via simulations and illustrate its use with a whole-genome-sequencing dataset collated from multiple studies associations between colorectal cancer and the human gut microbiome.

Taken together, this work identifies measurement error as a key consideration in the design, analysis, and interpretation of microbiome sequencing experiments, and in addition provides novel statistical methods to characterize and account for this error.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I wish to thank my doctoral supervisory committee for their valuable guidance, feedback, and time. In particular, I am grateful to the chair, my advisor Dr. Amy Willis, without whose support, encouragement, and patience none of the work in this document would have been possible.

I am grateful to my colleagues in the StatDivLab, Maria Valdez Cabrera, Jess Kunke, Sarah Teichman, and Pauline Trinh for camaraderie, feedback on numerous projects and presentations, and a seemingly endless supply of pet pictures. In particular, I would like to thank Pauline Trinh, with whom it has been an honor to collaborate and whose willingness to field my naive questions about metagenomics I greatly appreciate.

I am grateful to the many faculty as well. Drs. Lianne Sheppard and Adam Szpiro – thank you for your support. I have very much enjoyed learning about air pollution epidemiology and ethical biostatistical research from you. I owe Drs. Anthony Lioi (The Juilliard School) and Rick Chappell (UW-Madison) a debt of gratitude for their support early in my academic career.

I am indebted as well to many friends – Sheridan Grant for enduring years of academic rants, Hayden Windrow for perspective when I needed it and numerous dinners I would have been too tired to make myself, Andrew Herstein for intellectual company and commiseration over many a brunch, and many others. Thanks also to Diego Gomez for serving as a reminder that there is life beyond graduate school.

To my boyfriend, Josh Milber – thank you for your patience, understanding, and support.

I could not have pursued this degree without the love and encouragement of my family – my Aunt Jane, whose support allowed me to pursue statistical education in the first place,

my parents, Ray and Liz, who supported me through thick and thin, and my brother Elliot, whose wry humor brightened many difficult days.

This work was supported by the National Institute of General Medical Sciences (NIGMS) of the NIH under award number R35 GM133420, by the National Institute of Environmental Health Sciences (NIEHS) under award number T32ES015459, and by the Seattle Chapter of the ARCS Foundation.

Chapter 1

INTRODUCTION

In the past two decades, next-generation sequencing (NGS) methods have improved the characterization of complex microbial communities commonly known as microbiomes. As such communities are present in the human body (including in the digestive, urogenital and pulmonary tracts) as well as in natural environments such as soils, salt- and freshwater systems, and aquifers, better understanding their composition and functional capabilities may yield advances in medicine, public health, and agriculture, among other fields.

However, the NGS methods most commonly employed in microbiome studies – marker gene and whole genome sequencing – do not provide entirely unambiguous characterizations of microbial communities.[1] Performing marker gene sequencing, in which a highly conserved gene[2] is sequenced in an attempt to quantify microbial composition of a sample, or whole genome sequencing, in which a random subset of all genetic material in a sample is sequenced, entails a complex multi-step process. Samples to be measured must be carefully collected and stored before DNA is extracted from them and prepared for sequencing. Sequencing itself must be performed, after which the resulting raw read data is subjected to bioinformatics postprocessing to produce observations in the form of an abundance table. (For more comprehensive description of these measurement techniques, see Nearing et al. [2021] or Knight et al. [2018].) The structure and complexity of this measurement process limit the information available in its output; marker gene and whole-genome sequencing are, for example, generally unable to resolve microbe concentrations and instead provide information only re-

[1]NGS, as this sentence suggests, refers to a broad category of sequencing techniques that includes marker gene and whole genome sequencing. As marker gene and whole genome sequencing are the focus of this dissertation, for simplicity we use the term NGS to refer to these techniques.

[2]Frequently the 16S gene, which is present in Bacteria and Archaea, is used.

garding the proportional microbial composition of each sequenced sample. Moreover, NGS measurements are subject to substantial distortion compared to the communities they represent, with the form of this distortion depending on choice of protocol at each measurement step, [Nearing et al., 2021, Pollock et al., 2018, Sinha et al., 2017], and deconvolving biological signals from measurement error in NGS data is an area of ongoing investigation. The quality of these measurements and, in particular, how they relate to underlying biological quantities of interest is the topic of this dissertation.

Our first project attempts to characterize the impact of measurement error on experimental findings in microbiome science. To accomplish this, we quantify to what degree structure in NGS data is robust to measurement protocol and laboratory in a dataset consisting of measurements taken by multiple sequencing laboratories on samples from a common set of microbiome specimens. Using a classification-based approach, we assess how well distinctions between specimens estimated on the basis of measurements taken by a single sequencing laboratory replicate in additional sets of measurements either taken by the same laboratory or by another laboratory. We find higher replicability within than between sequencing laboratory, with the degree of replicability depending both on what data transformation is used and on the level of taxonomic aggregation at which data is considered. This finding suggests that, as between-specimen structure is not reliably conserved under differing sequencing protocols, observed between-group effects in microbiome studies may be driven in part by choice of sequencing protocol and laboratory.

We next present a method to estimate sample microbial relative abundances in the presence of two forms of measurement error commonly found in microbiome data: detection effects (whereby microbial taxa are multiplicatively over- and under-represented in measurements relative to their true abundance); and contamination of measurements by genetic material not present in samples of interest. That is, our model attempts to reconstruct composition of specimens from which samples have been taken for sequencing, an approach which we anticipate may find use either in settings in which accurately characterizing the compositions of individual specimens is of interest or in which improving downstream esti-

mation of some population quantity is the primary goal. Notably, this method is able to exploit dilution series data to identify contamination introduced during sequencing, a substantial concern in studies of low biomass microbial ecosystems. A major limitation of this method, however, is that detection effects are typically only identifiable among microbial taxa present in an artificial communities of known composition (i.e., "positive controls") sequenced alongside samples from natural communities under study. Artificial communities currently can only be constructed from microbes we are able to isolate and culture, and hence estimation of and adjustment for detection effects is not currently scalable to non-culturable organisms, limiting the applications in which sample relative abundances can be deconvolved from detection effects.

We address this limitation in our last chapter, which introduces a partially identified log-linear modeling framework for inference on population means of quantities (e.g., concentration of microbial cells in a large number of different species) that are observed only up to an unknown sample-specific scaling term and subject to detection effects. This, we argue, is a reasonable model in many microbiome experiments where measurements are taken via NGS; these techniques more easily detect some taxa than others, and they yield measurements that are uninformative with respect to total microbial concentration – i.e., that in each sample are scaled relative to total concentration by an unknown term. In contrast to our approach in the previous chapter, this method does not require estimation of sample compositions and instead links observations directly to a population mean of interest. In a similar vein, detection effects are not estimated directly in this approach but instead are absorbed into an intercept term that we regard as a nuisance parameter. We apply this method to perform a large meta-analysis of fecal samples taken from patients with colorectal cancer as well as from healthy controls. We also examine empirical performance through simulation.

We conclude with a brief discussion of outstanding statistical and scientific challenges posed by measurement error in microbiome science, together with recommendations for improving research practices to support rigorous interrogation of microbial communities.

Chapter 2

ASSESSING REPLICABILITY IN MICROBIOME DATA

2.1 Introduction

In this chapter, we examine data produced by the Microbiome Quality Control (MBQC) Project, a large collaborative study of cross-laboratory comparability of human microbiome 16S sequencing measurements. 16S sequencing is a widely employed approach that attempts to quantify microbial abundances by sequencing a hypervariable region of the 16S rRNA gene. The data that results from a 16S sequencing experiment is the number of times each 16S sequence variant j was observed in each sample i, which we call W_{ij}. Many steps are involved in generating a taxon abundance table $\{W_{ij}\}$ from specimens, including sample storage, DNA extraction, sequencing, and raw data processing (bioinformatics), and all of these steps are known to impact the resulting profiles [Pollock et al., 2018, Hugerth and Andersson, 2017, Sinha et al., 2017, Gibbons et al., 2018a].

We analyze this data both because at present no other dataset provides comparable insight into measurement error in human gut microbiome studies and also because its analysis published by Sinha et al. [2017] does not support various conclusions that Sinha et al. [2017] draw from it.

In broad terms, both our analysis and that of Sinha et al. [2017] aim to investigate to what extent observed structure in samples is preserved under differing measurement protocols as performed by various sequencing and bioinformatics laboratories. Accordingly, the MBQC dataset contains observations on 22 extremely dissimilar specimens: 2 are low-complexity artificial communities, 2 are drawn from a bioreactor, and 18 are from human subjects. Of the 18 human specimens, 9 are male, 5 are female, with donor sex unknown for the remainder. The age range of human specimens spans 2 years to 70 years, and approximately

1/3 of samples are from ICU cases, 1/3 are pre- or post-surgery, and 1/3 are healthy.

The disparate sources of samples in this dataset should lead to observable differences in 16S sequencing data. Indeed, we are able to identify distinctions between samples that hold in individual sequencing laboratories (low within-laboratory technical variation). However, we also find very different distinctions between samples across laboratories (high cross-laboratory variation). Hence, while within-laboratory technical variability is low enough to allow differentiation of samples, observed distinctions between samples are not robust to between-laboratory technical variation, suggesting that measurement error may mask or bias between-subject and between-group comparisons in human 16S studies.

2.2 Data and Model

2.2.1 Dataset

The MBQC Project was established by the MBQC Consortium to "comprehensively evaluate methods for measuring the human microbiome" [Sinha et al., 2015, pp. 1-2]. A major objective was to quantitatively compare the results of 16S sequencing as implemented by multiple research groups on identical samples. To this end, Sinha et al. [2017] distributed identical microbiome sample sets comprising samples from 22 unique specimens to 15 participating sequencing laboratories that were blinded to the samples' labels. Each laboratory prepared and analyzed sample sets according to a sequencing protocol of their choice. Raw data from each sequencing laboratory was then sent to 9 bioinformatics laboratories, which were blind to sequencing laboratory identity as well as specimen origin of samples. Bioinformatics laboratories processed the data according to an analysis protocol of their choosing. The taxon abundance tables from each sequencing-bioinformatics combination were submitted in standardized format.

We denote the taxon abundance data collected by Sinha et al. [2017] $\{W_{ijklm}\}$, where $i \in \{1, \ldots, I\}$ indexes the specimen from which a sample was taken, and $j \in \{1, \ldots, J\}$ indexes the operational taxonomic unit (OTU) to which the count is attributed. We fur-

ther index taxon abundance data according to the sequencing laboratory $k \in \{1, \ldots, K\}$ and bioinformatics laboratory $l \in \{1, \ldots, L\}$ that generated it. $m \in \{1, \ldots, M_{ikl}\}$ indexes replicate measurements on specimen i within sequencing laboratory $k \times$ bioinformatics laboratory l. We note in particular the distinction between "specimen" and "sample" here, which we will maintain throughout this chapter: a "specimen" is a unique source of genetic material, portions of which may be extracted for sequencing; a "sample" is such an extracted portion – i.e., it is the unit of sequencing.

2.2.2 Prior Analyses

Our analysis of the MBQC dataset was in part motivated by weaknesses in the original analyses of Sinha et al. [2017]. We briefly review these analyses here.

Sinha et al. [2017] present two major groups of analyses. The first of these concerns cross-laboratory similarity of various diversity indices – that is, of common univariate summaries of specimen complexity (as well as of pairwise similarity across specimens). This chapter does not attempt to recapitulate these diversity analyses, though we note that a conclusion Sinha et al. [2017] draw from them, that in general "relative diversity levels remain consistent"[1] across sequencing laboratory (i.e., regardless of which laboratory performs sequencing) holds only insofar as that in pairwise comparisons across sequencing laboratory, Spearman correlation of diversity indices is generally (though not always) positive. (The median pairwise Spearman correlation is 0.44 with IQR $0.29 - 0.66$.)

In addition to these analyses, Sinha et al. [2017] also fit linear mixed models separately to arcsine-square-root transformed observed relative abundance data for 4 phyla[2] (Firmicutes, Bacteroidetes, Proteobacteria, and Actinobacteria). Individually on each of these phyla, they fit two linear mixed effects models: the first including terms for sequencing and bioinformatics

[1] i.e., that comparisons in diversity across samples are preserved: e.g., if measurements from laboratory A indicate that specimen 1 has greater diversity than specimen 2, the same will generally hold in measurements taken by laboratory B.

[2] i.e., in each sequenced sample, the result of arcsine-square-root transforming the proportion of observed reads assigned to a bacterium in Firmicutes, Bacteroidetes, Proteobacteria, or Actinobacteria

laboratories as fixed (main) effects with a random intercept for specimen; and the second including terms for sequencing and bioinformatics *protocols* also with a random intercept for specimen. (Collinearity of laboratory and protocol precluded fitting models containing terms in both.) On the basis of these model fits, Sinha et al. [2017] conclude that "the greatest variability in microbiome profiling was assigned to biological differences between specimen sources and handling laboratories" and later assert "relative, not absolute, measures are comparable between protocols" [Sinha et al., 2017, p. 1084], ("relative" here refers to sample-to-sample comparisons).

For a number of reasons, it is difficult to relate the results of the analysis of Sinha et al. [2017] to human microbiome experiments as they are practically carried out. Firstly, the analysis is conducted on a fairly infrequently used transformation of 16S data, in this case explicitly chosen for variance stabilization (rather than on the basis of a relationship with some underlying biological quantity of interest). Additionally, only variability in transformed phylum data is considered; phyla are very broad taxonomic groupings and behavior of measurements at this level of aggregation are unlikely to be informative with respect to measurements analyzed on finer scales (such as species or genus, both commonly a focus of human 16S microbiome studies). Lastly, while it is not entirely clear how broadly Sinha et al. [2017] intend readers to construe their comments about relative structure between specimens being conserved across sequencing laboratory and protocol, the mixed models they fit are categorically unable to support such claims; they do not estimate any quantity from which we might draw such a conclusion. Sinha et al. [2017] summarize impact of measurement protocol (and laboratory) by appealing to variance explained by main effects in measurement protocol (or laboratory), but estimates of such effects provide no insight into whether measurement protocol or laboratory *differentially* impact different specimens (i.e., whether relative structure is preserved). The effect relevant to this question is instead an interaction between protocol and specimen, but no interactions are modeled by Sinha et al. [2017].

In view of these difficulties, in this chapter we attempt to more fully investigate to what degree apparent structure in measurements taken on different samples replicates across

sequencing laboratory.

2.2.3 Model and Model Evaluation

We wish to evaluate the widely-held belief that "each protocol will have a set of biases that affect all samples equally" (Sinha et al. [2017, p. 1081]), and in particular that on this basis measurement error may safely be disregarded in microbiome analyses. To this end, we propose a statistical model for the MBQC data to formalize and explore the implications of this claim. Let ρ_{ij} be the true, unknown relative abundance of taxon j in specimen i (so $\sum_{j=1}^{J} \rho_{ij} = 1$), and $\vec{W}_{i\cdot klm} = (W_{i1klm} \ldots, W_{iJklm})$ be the observed counts from all taxa in specimen i by sequencing laboratory k, and bioinformatics laboratory l. For simplicity, we omit further reference bioinformatics laboratories l, which our analysis treats as dependent replicates (see Section 2.3.3), and we let M_{ik} represent the number of replicate observations on specimen i reported for laboratory k. (Although all sequencing laboratories were sent identical sample sets, some laboratories were sent multiple sets; number of bioinformatics replicates per sample is relatively consistent in the subset of data we analyze, as discussed in Section A.2.) We represent transformed sequencing count data $\vec{W}_{i\cdot km}$ as a sum of transformed true relative abundances $\vec{\rho}_{i\cdot}$ and an error term $\vec{\epsilon}_{\Psi i\cdot km}$:

$$\vec{\Psi}(\vec{W}_{i\cdot km}) = \vec{\Psi}(\vec{\rho}_{i\cdot}) + \vec{\epsilon}_{\Psi i\cdot km}, \tag{2.1}$$

where Ψ is some scale-invariant transformation of the count data (e.g., relative abundance or centered log-ratio [3]; we consider only scale-invariant transformations because $\sum_{j=1}^{J} W_{ij}$ is an artifact of the sequencing experiment). We formalize the idea of protocol (or laboratory) biases affecting specimens equally via assumptions on the expectation of the error term $\vec{\epsilon}_{\Psi i\cdot km}$:

$$\mathbb{E}[\vec{\epsilon}_{\Psi i\cdot km}] = \vec{c}_{\Psi\cdot km} \tag{2.2}$$

That is, $\vec{\Psi}(\vec{W}_{i\cdot km})$ is a biased estimate of $\vec{\Psi}(\vec{\rho}_{i\cdot})$, with bias $\vec{c}_{\Psi\cdot k}$ constant across samples i but potentially differing across taxa j. (Note that bias of $\vec{\Psi}(\vec{W}_{i\cdot km})$ for $\vec{\Psi}(\vec{\rho}_{i\cdot})$ includes bias

[3] For $\mathbf{x} \in \mathbb{R}_{+}^{k}$, the centered log-ratio transformation of $\mathbf{x}$ is given by $g(\mathbf{x}) = \log\mathbf{x} - \frac{1}{k}\sum_{d=1}^{k} \log x_d$.

resulting from nonlinearity of $\vec{\Psi}$.) If we lift the first moment assumption on $\vec{\epsilon}_{\Psi i \cdot km}$, the model above is completely general; in particular, we note that, with or without first-moment conditions, this model makes no assumptions on similarity of error distributions across taxa.

We do not know the true abundances $\vec{\rho}_{i\cdot}$ for the MBQC data, nor in most 16S experiments. However, because we are interested in evaluating replicability, we do not attempt to estimate the $\vec{\rho}_{i\cdot}$'s. Instead, we propose a statistical machine learning approach to evaluate a key implication of model (2.1) - (2.2), with errors modeled as independent across samples.

We first note that under the above model, the expected difference between transformed observed abundances and transformed true abundances depends on sequencing laboratory k and the chosen transformation Ψ, but not on the sample composition i. Consequently, differences in transformed measurements between specimens are invariant in expectation across laboratory: $\mathbb{E}[\Psi(\vec{W}_{i\cdot k}) - \Psi(\vec{W}_{i'\cdot k})] = \mathbb{E}[\Psi(\vec{W}_{i\cdot k'}) - \Psi(\vec{W}_{i'\cdot k'})]$ (for simplicity, we suppress replicate index m here).

Therefore, we can evaluate the model in the absence of knowing $\vec{\rho}_{i\cdot}$ by assessing if observed between-specimen structure is conserved across laboratories.

To do this, we split data as follows. As each sequencing laboratory received multiple physical samples of each of the 22 unique specimens included in this study, we divide data produced by sample. We assign each sample sequenced by a given laboratory either to the training or the test set for that laboratory, with each sample assigned to training or test sets with equal probability and each training and test set containing at least one sample from each unique specimen. All bioinformatics results reported for a given sample share the set assignment of the sample.

This procedure produces, for each sequencing laboratory k, a training set W_k^{train} and a test set W_k^{test}. Separately on each of these sets we calculate "sample-centered" transformed measurements (Section 2.3.2), which we call $\Psi'(W)$. On each training set W_k^{train}, we attempt to find a function $\phi_k : range(\Psi') \to \{1, \ldots, I\}$ such that expected misclassification error

$$\frac{1}{\sum_{i=1}^{I} M_{ik}} \sum_{i=1}^{I} \sum_{m=1}^{M_{ik}} \mathbb{E}\left[\mathbf{1}_{\left[\phi_k(\Psi'(W_{i\cdot km})) \neq i\right]}\right] \tag{2.3}$$

is minimized.[4] In other words, ϕ_k is a classifier that identifies specimen label i, and was trained on data from sequencing laboratory k (Section 2.3.3). We then use ϕ_k to predict specimen labels on *every* sample-centered test set W_k^{test}: both for the laboratory that the classifier was trained on, and all other laboratories. [5]

Finally, we analyze the magnitude and patterns in the classifiers' misclassification rates, including which transformations better preserve distinctions between samples; at which taxonomic resolution we observe highest replicability; whether some laboratories detect more replicable distinctions between samples than others; and the difference in classifier performance within- versus across-laboratories. Figure 2.1 illustrates how our classifier-based approach distinguishes between structure-preserving and structure-distorting measurement error across sequencing laboratories.

In short, our approach allows us to evaluate how similar measurement error is across sequencing laboratories in terms of our ability to distinguish specimens after we have removed a form of bias that is irrelevant to estimating differences between specimens. When the distribution of measurement error in laboratories k and k' is identical up to a constant (i.e., up to $\vec{c}_{\Psi \cdot k} - \vec{c}_{\Psi \cdot k'}$), we expect classifiers to perform equally well in either laboratory. When measurement error distributions are not equal up to a mean shift, our approach provides an empirical measure of the degree of difference between them. We express this result in terms of "centered" error terms $\epsilon_{\Psi i \cdot km} - \mathbb{E}[\epsilon_{\Psi i \cdot km}]$, where expectation is taken over the joint distribution of class (specimen) labels i and class-and-laboratory-specific error terms $\epsilon_{\Psi i \cdot km}$ (note that by design the distribution of class labels i is identical across laboratories). In particular, if we let Q_k denote the joint distribution of centered error terms $\epsilon_{\Psi i \cdot km} - \mathbb{E}[\epsilon_{\Psi i \cdot km}]$ and the class labels i in laboratory k, and analogously for laboratory k' with $Q_{k'}$, we have

[4] We note that ϕ_k, as defined here, is not unique without some limitations imposed on the function class from which it is chosen. To explore sensitivity to choice of function class (and hence of non-uniqueness of ϕ_k), we compare results using classifiers selecting ϕ_k from disparate function classes.

[5] All bioinformatics laboratories included in our analysis produced OTU tables using a common reference database [Sinha et al., 2017, p. 1078], so taxonomic assignments are generally comparable across data reported from these laboratories.

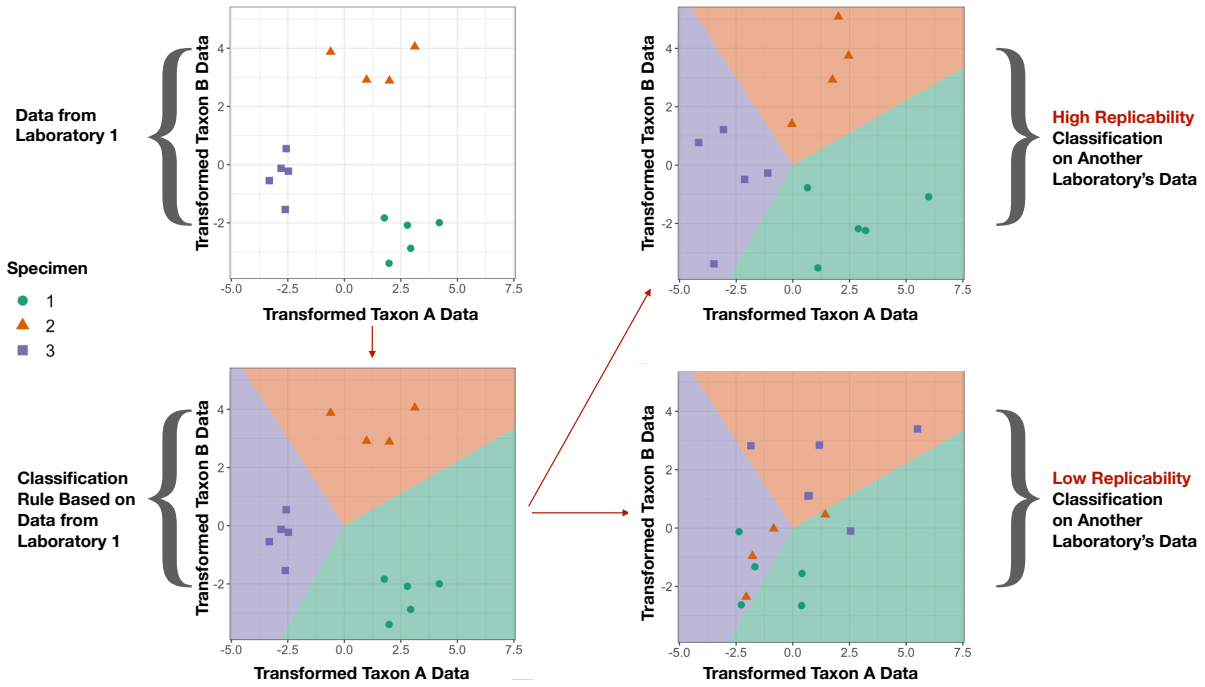

Figure 2.1: Performance of a classifier trained on centered transformed taxon abundances in Laboratory A indicates the degree to which similar between-specimen structure is able to be resolved in data from another sequencing laboratory.

for any classification rule $\phi(X)$ that $|\mathbb{E}_{Q_k}[\mathbf{1}_{[\phi(X)\neq I]}] - \mathbb{E}_{Q_{k'}}[\mathbf{1}_{[\phi(X)\neq I]}]| \leq d_{\mathrm{TV}}(Q_k, Q_{k'})$. That is, the absolute value of the difference in expected misclassification rates lower bounds the total variation distance between Q_k and $Q_{k'}$. In practice, we center by sample averages since $\mathbb{E}[\epsilon_{\Psi i \cdot km}]$ is not known, and so the inequality holds up to $\mathcal{O}(1/\sqrt{n_k \vee n_{k'}})$ term, if n_k is the number of samples sequenced by laboratory k, and similarly for $n_{k'}$ and laboratory k'. (See Appendix A.1 for a formal statement and proof).

2.3 Methods

2.3.1 Data Sources and Treatment of Missing Data

We considered OTU count data published by Sinha et al. [2017] in Nature Biotechnology (Supplementary Data set 6 available at `https://www.nature.com/articles/nbt.3981`). As explained in Sinha, et al., data from bioinformatics laboratories BL-3 and BL-5 was excluded from Supplementary Data set 6 as these laboratories did not report counts in standardized format. We stored available data as a table with columns representing variables and rows observations using the data.table package (version 1.12.6) [Dowle and Srinivasan, 2019] in R (version 3.6.1) [R Core Team, 2019]. We then filtered out all rows for which sequencing lab, bioinformatics lab, or specimen was listed as missing or unknown, as well as negative controls and pre-extracted DNA samples. As reported in Sinha et al. [2017, pg. 1079], some bioinformatics groups did not report results for samples with read counts below a given threshold. This induced missingness in some combinations of sequencing lab, dry lab, and specimen. To avoid confounding bioinformatics and sequencing laboratory effects, as well as to ensure that sufficient data was available to train and validate classifiers in each sequencing laboratory we considered, we limited our analysis to a subset of sequencing and bioinformatics laboratories with sufficient completeness in each combination of sequencing and bioinformatics laboratory. Details of the procedure used to select this subset of laboratories are available in Appendix A.2.

2.3.2 Data Transformations

Each laboratory provided taxon abundance tables with 16S sequence variants j attributed to operational taxonomic units (OTUs), a unit based on observed 16S gene sequence similarity. OTUs can be organized according to a taxonomy comprising (from finest to broadest) species, genus, family, order, class, and phylum. To assess degree of replicability at finer or broader aggregations, we trained classifiers for every level of taxonomic aggregation.

For every taxonomic aggregation level, we consider two transformations Ψ that are commonly used in microbiome analyses. We consider the proportion transformation

$$\Psi_1 : \mathbb{R}^J \to \mathbb{S}^{J-1}; \Psi_1(\vec{W}_{i\cdot k}) = \left(\frac{W_{i1k}}{\sum_{j=1}^J W_{ijk}}, \ldots, \frac{W_{iJk}}{\sum_{j=1}^J W_{ijk}} \right),$$

and the centered log-ratio transformation

$$\Psi_2 : \ \mathbb{R}^J \to \mathbb{R}^J; \ \Psi_2(\vec{W}_{i\cdot k}) = \left(\log \frac{W_{i1k}}{\left(\prod_{j=1}^J W_{ijk}\right)^{1/J}}, \ldots, \log \frac{W_{iJk}}{\left(\prod_{j=1}^J W_{ijk}\right)^{1/J}} \right).$$

Note that the centered log-ratio transform is only defined when $W_{ijk} > 0$ for all j, so in practice, zero counts in taxon tables are frequently replaced by a small positive "pseudocount" prior to transformation [Quinn et al., 2019]. In accordance with this practice, we replaced all zero counts with a pseudocount of 1 before transformation. We investigated the sensitivity of our results to the choice of pseudocount and found negligible differences in classifier performance. We note that adding a pseudocount breaks the scale invariance of Ψ_2.

To ensure that classifiers learn features of the data that reflect between-specimen structure and do not depend on $\vec{c}_{\Psi\cdot k}$, for transformations Ψ_1 and Ψ_2 we center measurements on samples from specimens 1 through I from every test or training set by subtracting $\frac{1}{I}\sum_{i=1}^I \frac{1}{M_{ik}^{\text{set}}} \sum_{m'=1}^{M_{ik}^{\text{set}}} \Psi(\vec{W}_{i\cdot km'})$ from each observation, where M_{ik}^{set} is the number of replicate measurements for specimen i in the sample set (either a training or test set). We performed this centering to ensure that the resulting centered quantities have expectation that does not

depend on $\vec{c}_{\Psi\cdot k}$:

$$\Psi(\vec{W}_{i\cdot km}) - \frac{1}{I}\sum_{i=1}^{I}\frac{1}{M_{ik}^{\text{set}}}\sum_{m'=1}^{M_{ik}^{\text{set}}}\Psi(\vec{W}_{i\cdot km'}) \tag{2.4}$$

$$= \left[\Psi(\vec{\rho}_{i\cdot}) + \epsilon_{\Psi i\cdot km}\right] - \frac{1}{I}\sum_{i'=1}^{I}\frac{1}{M_{ik}^{\text{set}}}\sum_{m'=1}^{M_{ik}^{\text{set}}}\left[\Psi(\vec{\rho}_{i'\cdot}) + \epsilon_{\Psi i'\cdot km'}\right] \tag{2.5}$$

$$= \left[\Psi(\vec{\rho}_{i\cdot}) - \frac{1}{I}\sum_{i'=1}^{I}\Psi(\vec{\rho}_{i'\cdot})\right] + \left[\epsilon_{\Psi i\cdot km} - \frac{1}{I}\sum_{i'=1}^{I}\frac{1}{M_{ik}^{\text{set}}}\sum_{m'=1}^{M_{ik}^{\text{set}}}\epsilon_{\Psi,i'\cdot km'}\right] := \Psi'(\vec{\rho}_{i\cdot}) + \epsilon'_{\Psi i\cdot km} \tag{2.6}$$

where $\Psi'(\vec{\rho}_{i\cdot})$ and $\epsilon'_{\Psi i\cdot km}$ are defined to be the first and second bracketed terms of the LHS of (2.5), respectively. We additionally define $\Psi'(\vec{W}_{i\cdot km})$ calculated on either a test or training from sequencing laboratory k: $\Psi'(\vec{W}_{i\cdot km}) := \Psi(\vec{W}_{i\cdot km}) - \frac{1}{I}\sum_{i=1}^{I}\frac{1}{M_{ik}^{\text{set}}}\sum_{m'=1}^{M_{ik}^{\text{set}}}\Psi(\vec{W}_{i\cdot km'})$.

If the model given in Section 2.2.3 holds, we have

$$\mathbb{E}\epsilon'_{\Psi i\cdot km} := \mathbb{E}\Big[\Big[\epsilon_{\Psi i\cdot km} - \frac{1}{I}\sum_{i'=1}^{I}\frac{1}{M_{ik}^{\text{set}}}\sum_{m'=1}^{M_{ik}^{\text{set}}}\epsilon_{\Psi,i'\cdot km'}\Big]\Big] \tag{2.7}$$

$$= c_{\psi\cdot k} - \frac{1}{I}\sum_{i'=1}^{I}\frac{1}{M_{ik}^{\text{set}}}\sum_{m'=1}^{M_{ik}^{\text{set}}}c_{\psi\cdot k} = \vec{\mathbf{0}} \tag{2.8}$$

Hence, the centered transformed measurements are equal to centered transformed true relative abundances plus a mean-zero error under model (2.1) - (2.2).

2.3.3 Training and Validation of Boosted Tree and Elastic Net Classifiers

To train and validate classifier performance, we first assign samples sequenced by each sequencing laboratory to either the test or the training set for that laboratory. For each unique specimen, samples taken from that specimen are assigned to training or test sets with equal probability, and each training and test set contains at least one sample from each unique specimen. All bioinformatics results reported for a given sample share the set assignment of the sample.

On each training set, we trained boosted regression tree classifiers with R package xgboost (version 0.90.0.2) [Chen et al., 2019] and elastic net classifiers with R package glmnet (version

2.0.18) [Friedman et al., 2010] to classify samples according to which specimen they were taken from. We then predicted on the test set to obtain estimates of within-laboratory misclassification rates for each classifier. We also predicted specimen label on test sets for sequencing laboratories not used to train the classifier to obtain estimates of cross-laboratory misclassification rates.

For both boosted tree and elastic net classifiers, we selected parameters via 10-fold cross-validation on training sets. Details of parameter selection are provided in Appendix A.2. For each combination of k, Ψ and taxonomic level, a distinct classifier was trained.

We briefly note here that overfitting (i.e., selection of high-variance classifiers) is a common concern in high-dimensional settings. For this reason, we chose classifiers (boosted tree and elastic net) and a training procedure (10-fold cross-validation) with some robustness to this problem [Bühlmann et al., 2007, Friedman et al., 2004, Zou and Hastie, 2005]. Moreover, we constructed training and test sets so that classifiers are trained and evaluated on measurements taken on *completely disjoint* sets of samples, so misclassification error on test sets is unbiased for population misclassification error if errors $\epsilon_{\psi i \cdot km}$ are independent across samples m. Conversely, if dependence of errors $\epsilon_{\psi i \cdot km}$ within sequencing laboratory drives lower within- than across-laboratory misclassification error, this indicates that observed distinctions between samples depend on sequencing laboratory – the very phenomenon we aim to investigate in this chapter.

To investigate conservation of observed distinctions among more disparate samples, we conducted an additional analysis focusing on specimen type rather than specimen. For this analysis, we categorized the 22 unique specimens analyzed by the MBQC into four broad types: human (18 specimens); chemostat (2 specimens); artificial fecal community (1 specimen); artificial oral community (1 specimen). We performed sample centering using these types, trained classifiers with type labels, and evaluated classifier performance predicting specimen type on held-out test sets from each laboratory. In all other respects, we observed the same protocol as in our primary analysis.

All misclassification results shown in Section 2.4 are based on the testing sets (no mis-

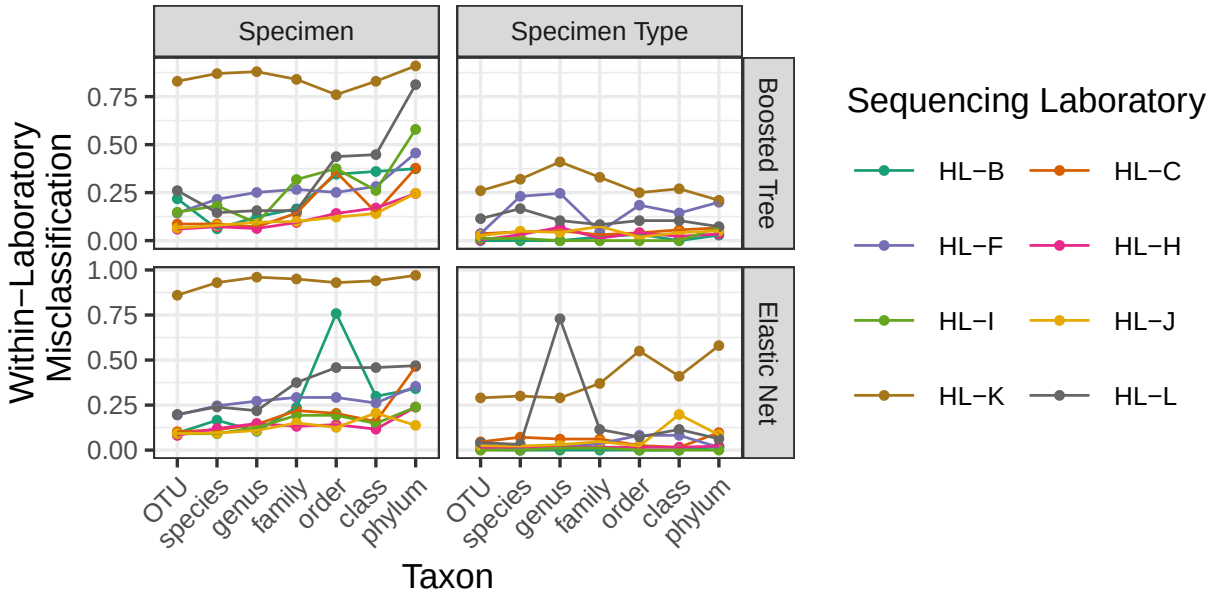

Figure 2.2: Within-laboratory misclassification for boosted tree and elastic net classifiers predicting specimen or specimen type on proportion data plotted against level of taxonomic aggregation. Color indicates the laboratory that the classifier was trained and evaluated on.

classification rates for training data are shown).

Code to reproduce the analysis is available at `github.com/statdivlab/mbqc_supplementary`.

2.4 Results

2.4.1 Proportion-Scale Data

We first examine performance of each classifier on held-out test data from the laboratory on which the classifier was trained. Within-laboratory, out-of-sample predictions provide a baseline against which to compare cross-laboratory performance. Within-laboratory misclassification rates on proportion-scale data (Ψ_1) are shown in Figure 2.2.

Within sequencing laboratory, signals distinguishing specimen types exhibit generally low misclassification error: median 5% (IQR 2% - 11%) for boosted tree signals and 3% (IQR

1% - 8%) for elastic net. This low misclassification is likely due both to the relatively large biological differences between specimen types (human, oral mock, fecal mock, or chemostat) and to the composition of sample sets sent to sequencing laboratories, which were composed of $\sim 75\%$ human fecal samples on average.

Specimen misclassification is generally higher than specimen type misclassification on within-laboratory replication, with median specimen misclassification 22% (IQR 12% - 38%) for boosted tree classifiers. This higher misclassification reflects both the increased biological similarity between specimens (versus specimen types) and the relatively even distribution of specimens across samples: no specimen accounts for more than 6% of samples sent to a sequencing laboratory.

Additionally, within-laboratory replicability of between-specimen signals appears to decrease as taxonomy coarsens. Median within-laboratory misclassification of boosted tree classifiers is 15% (IQR 8% - 23%) on OTU-level data, rising to 42% (IQR 34% - 64%) on phylum data. The corresponding figures for elastic net classifiers are 10% (IQR 9% - 20%) and 35% (IQR 24% - 47%), respectively.

Spikes in misclassification of elastic net classifiers on HL-B and HL-L (for specimen and specimen type classification, respectively) are explored in greater detail in Appendix A.5. In short, they likely result from three sources: relatively high within-laboratory technical variation in HL-L; probable splitting of (unlabeled) batches [6] in HL-B across training and test sets; and the sensitivity of the elastic net to the distribution of measurement error under the sum-to-one constraint imposed at the proportion scale. The consistently high within-laboratory misclassification in HL-K may be due to mislabeling of samples by this laboratory.

To investigate the degree to which between-specimen structure was conserved across sequencing laboratory, we used each of the classifiers trained on proportion data to predict

[6] That is, HL-B processed multiple sample sets, likely in different sequencing runs (i.e., batches), but we do not have access to which samples were sequenced in which batch and so cannot take batch structure into account in our analysis.

specimen using proportion data from every other laboratory. The performance of these classifiers in terms of misclassification on OTU, genus, order, and phylum data is summarized in Figure 2.3. See Appendix A.4 for results for specimen and specimen type at all taxonomic levels.

Between-specimen signals learned from proportion data replicate far less strongly across- than within-sequencing laboratory. Median within-laboratory misclassification for boosted tree classifiers is 22% (IQR 12% - 38%), compared to 60% (IQR 42% - 75%) across-laboratory.

Within- and cross-laboratory misclassification both generally increase with coarsening taxonomy. For the boosted tree classifiers, median misclassification on OTU-level data is 15% within-laboratory (IQR 8% - 23%) versus 55% cross-laboratory (IQR 30% - 75%). On phylum-level data, these figures are 35% (IQR 24% - 47%) and 82% (IQR 66% - 88%), respectively. On phylum-level data, elastic net classifiers perform similarly, attaining 30% (IQR 24% - 42%) median within-laboratory and 71% (IQR 55% - %) cross-laboratory misclassification. On OTU-level data, elastic net classifiers marginally outperform boosted tree classifiers, with median within-laboratory misclassification 10% (IQR 9% - 20%) and median cross-laboratory misclassification 34% (IQR 24% - 61%).

2.4.2 Log-Ratio Transformed Data

16S data is also frequently analyzed after a log-ratio transformation, an approach from the compositional data literature (see, e.g., Aitchison [1982]). In this section, we examine replicability of between-specimen signals on the centered log-ratio scale.

As with proportion-scale data, between-specimen signals learned from centered log-ratio data replicate more strongly within than across sequencing laboratory (Figure 2.3; see Figure A.8 for more detailed within-laboratory results). For boosted tree classifiers, median within-laboratory misclassification is 24% (IQR 17% - 31%), in contrast to 60% (IQR 42% - 75%) cross-laboratory misclassification.

Within- and cross-laboratory misclassification both generally increase with coarsening taxonomy (Figure 2.3). For the boosted tree classifiers, median within-laboratory misclas-

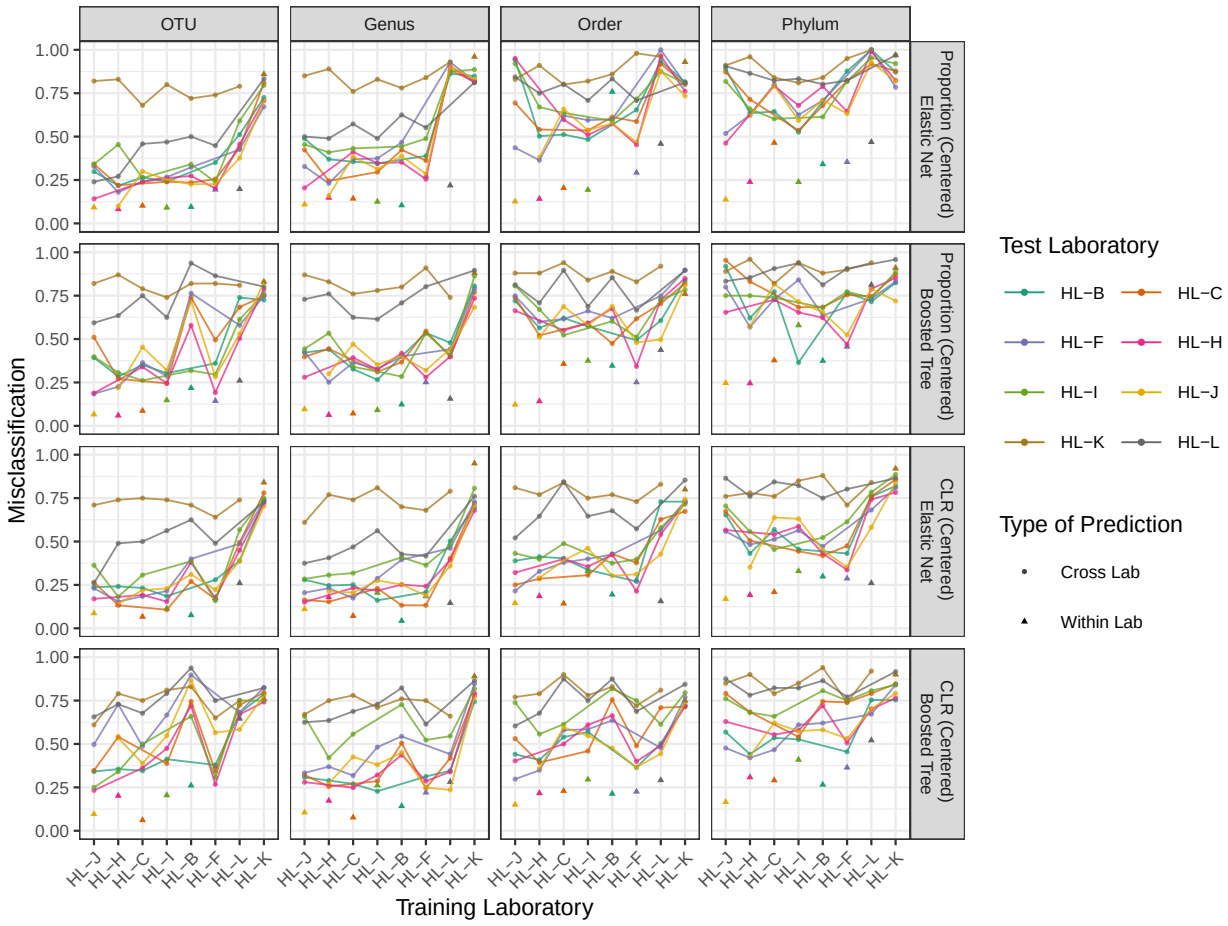

Figure 2.3: Within-laboratory (solid triangles) and between-laboratory (lines) misclassification for boosted tree and elastic net classifiers predicting specimen on sample-centered proportion (first two rows) and centered log-ratio (third and fourth rows) data plotted against level of taxonomic aggregation. Color indicates the laboratory whose data the classifier was trained on.

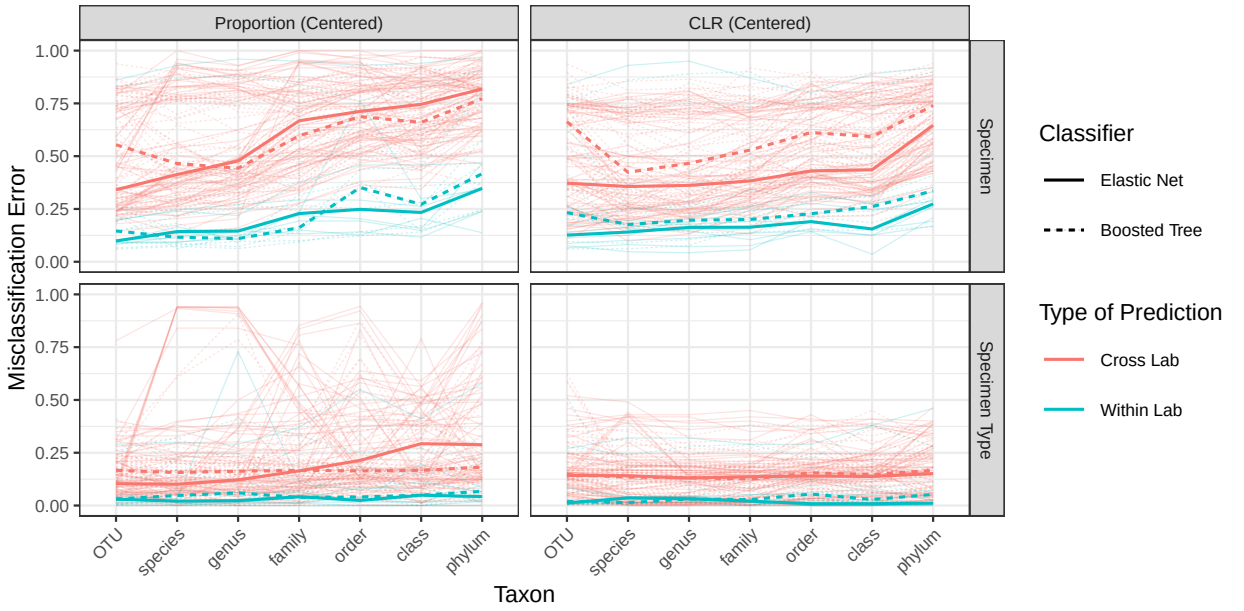

Figure 2.4: The within-laboratory (aqua) and cross-laboratory (red) misclassification for centered proportion and centered log ratio data for classifiers predicting both specimen and specimen type. The misclassification rate is shown for boosted tree (solid lines) and elastic net (dotted lines) classifiers for every combination of laboratories (thin lines) and is also summarized as a median across laboratory combinations (thick lines).

sification on OTU-level data is 23% (IQR 17% - 44%) versus 66% (IQR 41% - 75%) for cross-laboratory misclassification. On phylum data, these figures are 33% (IQR 28% - 44%) and 74% (IQR 58% - 81%), respectively.

2.4.3 Summary of findings

Figure 2.4 summarizes our findings, showing the misclassification error for each pair of training and test laboratories rendered as a thin line segment. For each combination of transformation and classification task (specimen vs. specimen type), median within- and cross-laboratory misclassification is plotted against taxon as a bold line.

Regardless of classifier, for each transformation and at every level of taxonomic aggregation we considered, median *within-laboratory* specimen misclassification is substantially lower than median *cross-laboratory* specimen misclassification. That is, sequencing laboratories find distinctions between samples that cannot be replicated in other laboratories. Whether this occurs because some sequencing laboratories report spurious reads (e.g., contamination) or because laboratories are differentially able to measure various features of microbial communities is not clear from this analysis. Both sources of error likely contribute to the patterns observed in this analysis, and the scale and level of taxonomic aggregation at which data are considered may determine the relative importance of each source.

For each transformation and classifier we considered, within- and cross-laboratory specimen misclassification generally increase with increasing level of taxonomic aggregation, although the strength of this trend varies by transformation. That is, replicability decreases with increasing taxonomic aggregation. This result most likely reflects information lost as a result of aggregation. Additionally, the pattern of cross-laboratory misclassification suggests that measurement error is not mitigated by taxonomic aggregation. That is, our findings suggest that between-phylum signals are generally less replicable than signals at finer taxonomic levels, possibly because the effects of measurement error are more similar in closely related taxa than in distantly related taxa.[7] We therefore recommend analysis on data at the genus level or finer.

Within-laboratory misclassification was generally lower than cross-laboratory misclassification under every data transformation we considered. With respect to both within- and cross-laboratory misclassification, however, proportion data were particularly unreliable, with specimen type classifiers frequently failing to outperform simply classifying all samples as human. Measurement error due to differential detection efficiency and contamination are nonlinear on this scale, which may explain for this behavior. Hence, we expect statistical inference on means of 16S proportion data to be particularly sensitive to measurement error.

[7]That is, at high levels of taxonomic aggregation, measurement error is summed across contributions across constituent (finer) taxa which may be quite differently impacted by error under a given protocol.

2.4.4 Supplemental Analyses

We perform several analyses to investigate the sensitivity of our analysis to various modeling choices. Appendix A.4 explores whether we can find a subgroup of sequencing laboratories across which between-specimen signals replicate well. In Appendix A.6, we assess the influence of pooling results from multiple bioinformatics laboratories via a sensitivity analysis in which we fit and predict from classifiers using only data reported by bioinformatics laboratories individually. In Appendix A.7, we fit classifiers to centered and rescaled sequencing data to investigate whether laboratory- and taxon-specific scalings could explain our results, and we additionally provide descriptive plots of measured abundance across four major phyla in fresh human specimens by sequencing laboratory and specimen. Appendix A.7 reports results for within- and cross-laboratory performance of classifiers trained on presence-absence data, another commonly used transformation of sequencing read data.

2.5 Discussion

Our objective in this chapter was to assess the replicability in 16S data, and to this end, we focused on the inter-laboratory similarity of measured distinctions between specimens. Focusing on between-specimen distinctions allowed us to assess evidence for the claim that "each protocol will have a set of biases that affect all samples equally" [Sinha et al., 2017, p. 1081] and that on this basis between-group comparisons can be resolved in the presence of measurement error; we find little evidence to support this claim. Our analysis also estimates a lower bound on the total variation distance between the residual measurement error distributions across sequencing laboratories after a laboratory-specific bias term has been canceled.

We found that replication of between-specimen signals is stronger within- than across-sequencing laboratory, even when laboratories analyze identical specimens. This is consistent with the recent work of Wirbel et al. [2021], who found substantially lower cross-study than within-study performance of classifiers trained to identify disease states on the basis of data

from a single study. Notably, cross-study predictive performance improved when training data was augmented with external study data. Both these results and our own suggest that laboratory-specific measurement error may mask or distort between-group comparisons of 16S data.

While we did not directly test the applicability of any given model connecting sample composition $\rho_{i\cdot}$ to taxon abundance tables $W_{i\cdot}$, our findings were broadly consistent with the multiplicative detection effect model of McLaren et al. [2019]. This model predicts that between-laboratory replicability should be greater on centered log-ratio data, where, given sufficiently deep sequencing, multiplicative distortions should cancel in our analysis. In centered log-ratio data, this canceling should occur primarily at fine levels of taxonomy, as the multiplicative detection effects described in McLaren et al. [2019] properly apply to microbes at the strain level; at higher taxonomic levels, strains varying in degree of detectability will be grouped together, and the model of McLaren et al. [2019] will fit less well. We do in fact observe these patterns (in particular, compare classifier performance on centered log ratio versus proportion data with and without sample centering, as reported in Appendix A.9), suggesting that differential detection of certain taxa by protocols may drive some of the measurement error we observe.

2.5.1 Limitations & interpretation of results

In this analysis, we set out to assess conservation of between-specimen signals within and across sequencing laboratories. As the form of these signals was not known a priori, we used flexible classifiers to learn between-specimen distinctions and then assessed replicability in terms of misclassification error of classifiers trained on training sets from each sequencing laboratory. In order to train classifiers, we pooled data across bioinformatics laboratories for use as replicates. While these decisions allowed us to illustrate the impact of measurement error on between-specimen signals, they each introduce limitations into our analysis.

While our analytical approach allowed us to flexibly learn distinctions between specimens, this flexibility prevented our analysis from highlighting any specific set of taxa as

contributing to replication failure. This reflects our goal of evaluating the replicability of measured between-specimen distinctions, though it renders the applicability of our findings to any particular experimental result more difficult to determine.

In addition, while we chose to examine 16S data under two transformations commonly used in microbiome analyses, microbiome studies employ a wide range of analytic techniques. These include approaches borrowed from the RNA-seq literature such as limma, an empirical Bayes method that can incorporate observation reweighting via an estimated mean-variance relationship [Law et al., 2014], as well as DESeq2 and edgeR, which attempt to account for technical variation via normalizations either applied directly to data as transformations or included as terms in a model [Robinson et al., 2010, Love et al., 2014], Though we did not explore the replicability of observed differences between specimens after application of normalizations commonly used in RNA-seq analyses, we note that RNA-seq (as well as microbiome) methods have previously been found not to control type-1 error in microbiome data [Hawinkel et al., 2019]. In a similar vein, we were unable to address the totality of statistical methods developed specifically for 16S data in our analysis, and it is unclear how our results will generalize across methods.

We chose to use taxon abundance data reported by participating bioinformatics laboratories rather than reprocess raw read data. Accordingly, our results, particularly for fine taxonomic levels, do not reflect recent developments in bioinformatics protocol, such as the ability to identify exact 16S sequence variants [Callahan et al., 2016, 2017]. Furthermore, as we estimate and validate signals over bioinformatics replicates, we estimate misclassification error averaged over bioinformatics laboratories. We investigate the sensitivity of our analysis to pooling of bioinformatics in Appendix A.6.

The conclusions we can draw from this analysis are also limited by the MBQC dataset. As the MBQC Project was not a designed experiment with respect to laboratory protocol (laboratories chose their own protocols), laboratory effects may confound any observed protocol effects. Additionally, the distribution of some protocol variables is highly unbalanced, rendering inference imprecise, even in the absence of confounding. For example, among the

laboratories we included in our analysis, six out of eight used the same 16S primer, with the other two each using a distinct primer. For these reasons, we chose not to estimate effects due to protocol variables directly.

Additionally, the population of laboratories included in the MBQC may not represent the global population of laboratories that generate microbiome data. For instance, laboratories participating in the MBQC study may have differed from the typical laboratory conducting 16S sequencing in terms of funding and academic profile. Furthermore, participating laboratories knew that they were participating, which may have changed their behavior (referred to as the Hawthorne effect). Relatedly, bioinformatics laboratories discarded samples as quality control, and it is unclear if bioinformatics teams working in collaboration with sequencing teams would discard samples as readily. For these reasons, we chose to present descriptive summaries of classifier performance rather than perform inference on misclassification rates.

The generalizability of our analysis is also limited by the range of specimens included in the MBQC. As all human specimens included in this study were fecal samples, our results are most relevant to studies of the human gut microbiome. In a similar vein, as the true composition of these specimens is unknown, we were only able to assess consistency of measurements across laboratories. Therefore we are not able to recommend any particular sequencing protocol for estimating true sample composition.

2.6 Conclusion

In the past two decades, failures of replication in quantitative disciplines ranging from social science to biomedical research have attracted considerable scientific and public attention [Loken and Gelman, 2017, Ioannidis et al., 2001, Simmons et al., 2011]. Concern over replication failures is well-founded: if independent groups of researchers cannot replicate scientific findings, this calls into question to what extent the published literature reflects objective reality.

In this chapter we evaluated the replicability of high-dimensional microbiome data obtained from 16S sequencing. By analyzing a dataset wherein identical samples were dis-

tributed to different sequencing laboratories, we demonstrated that measurement error in 16S studies degrades the replicability of measured distinctions between specimens. The degree of non-replicability depends on the data transformation and level of taxonomic aggregation used in analysis. We derived this result by training flexible classifiers to identify specimens using data from individual sequencing laboratories and comparing their performance on held-out test sets taken from the laboratories on which they were trained versus on test sets from other laboratories. On species-level data, the classifiers *correctly* classified a median of 64% and 56% of specimens predicting on, respectively, centered log ratio and proportion data based on test data from a *different* laboratory than the laboratory that generated the training data. These figures were substantially lower than the corresponding figures for classification on out-of-sample test data from the *same* laboratory used to train classifiers, 84% and 87%.

In general we observed larger misclassification errors at coarser levels of taxonomy. For example, on phylum-level centered log-ratio data, classifiers correctly classified a median of 71% of specimens within laboratory, but only 30% across laboratory. In addition, we found that even when observed within-laboratory technical variation was low, replication of between-specimen structure suffered in cross-laboratory comparisons.

These results suggest that measurement error in 16S studies may mask or distort distinctions between specimens. Hence, in our view 16S profiles are best understood as providing a noisy, likely distorted picture of microbial communities, and caution should be exercised when interpreting the results of a 16S analysis. Accordingly, we advocate for the independent validation of conclusions drawn from 16S sequencing [Minot and Willis, 2020].

Our findings highlight the need for further research in the characterization of measurement error in sequencing of microbial communities. For example, McLaren et al. [2019] recently demonstrated that observed profiles of simple communities differ from the true profiles by taxon-specific multiplicative factors that can be attributed to components of the sequencing workflow (e.g. extraction and amplification). The model of McLaren et al. [2019] may partially explain some of our findings. If the model of McLaren et al. [2019] applied

perfectly to the MBQC data, we would expect similar levels of cross-laboratory and within-laboratory misclassification in CLR-transformed species-level abundance data. However, we observe a median correct classification of 64% and 84% across- and within-laboratory misclassification, respectively, a discrepancy highlighting the need for further characterization of measurement error in human 16S data. We consider this challenge further in chapter 3.

More positively, our findings suggest that certain analyses may be more robust to measurement error than others. In particular, our results point to analyses at fine taxonomies on a log-ratio scale as more likely to replicate, although this pattern did not hold in every cross-laboratory comparison in our analysis. This has a number of practical implications for many different types of microbiome analysis, including inference (log-ratio based models may be more robust to measurement error than relative-abundance models) and visualization (ordination using Aitchison distance may be more appropriate than using other measures of dissimilarity).

Experimental techniques to study microbial communities continue to be developed, and whole-genome "shotgun" sequencing, long-read sequencing, microbial single-cell sequencing and microbial transcriptomics are becoming increasingly prevalent approaches to surveying microbiomes. To our knowledge, a study where identical samples were distributed to data collection centers that use these alternative microbial community profiling techniques has not been performed. As our analysis indicates that within-laboratory consistency does not in general guarantee cross-laboratory consistency, we encourage microbial ecologists to consider the potential for cross-laboratory inconsistency in emerging experimental techniques until cross-laboratory consistency has been demonstrated.

Chapter 3

MODELING COMPLEX MEASUREMENT ERROR IN MICROBIOME SEQUENCING DATA

3.1 Introduction

In this chapter, we present a method for deconvolution of certain forms of measurement error from biological signals in high-throughput sequencing data generated by microbiome experiments. While our method is novel, the problem of measurement error has attracted substantial attention in the microbiome literature, and numerous approaches have been proposed to address measurement error in high-throughput sequencing studies. A particularly common concern are "batch effects" – systemic distortions in observed abundance data due either to true biological variation (e.g., cage/tank effects in model organism studies) or measurement error (e.g., lot-to-lot variation in reagents). The majority of batch effect methods used in microbiome studies were originally developed for RNA-seq and microarray analysis, such as surrogate variable analysis [Leek and Storey, 2007], ComBat [Johnson et al., 2007], remove unwanted variation [Gagnon-Bartsch and Speed, 2012] and batch mean centering [Sims et al., 2008]. Batch effects methods developed specifically for microbiome studies include percentile-normalization for case-control meta-analyses [Gibbons et al., 2018b], and a Bayesian multinomial-Dirichlet model [Dai et al., 2019].

Another line of research aims to directly estimate microbial abundances from sequencing data. In an early effort in this area, Brooks et al. [2015], fit linear models to model the empirical proportions of bacteria in communities of known composition. However, the complexity of these models, which incorporated second-order species interactions, limited their generalizability and interpretability. Using data published by Brooks et al. [2015], McLaren et al. [2019] proposed a parsimonious model for microbial abundance data in which observed

ratios of bacterial strains are multiplicatively distorted due to the over-detection of some strains relative to others. This approach yielded improved predictive performance, particularly in the form of cross-sample generalizability, as it was able to disentangle measurement protocol effects from sample composition. Generalizing this approach remains an active area of research, and multiple methods now exist to estimate multiplicative detection effects [Silverman et al., 2021, Zhao and Satten, 2021].

However, unequal detection of bacterial strains is not the only impediment to estimating microbial abundances from NGS data. Contamination introduced during sample handling, preparation and sequencing can be a significant source of measurement error, particularly in low biomass settings [Willner et al., 2012, Salter et al., 2014, Weiss et al., 2014]. Methods to identify and remove contamination include SourceTracker [Knights et al., 2011] and FEAST [Shenhav et al., 2019], which model contamination as a mixture of contributions from an arbitrary number of known sources in addition to a single unknown source. Another method, decontam [Davis et al., 2018], uses measured DNA concentrations to identify contaminants under the assumption that total contaminant DNA concentration is approximately constant across samples.

In this chapter, we present a method to account both for systematic over- and under-detection of taxa as well as contamination in NGS data. The framework we present allows for principled inclusion of covariates governing both detection effects and contamination, such as batch covariates. We develop a stable algorithm for constrained estimation of relative abundance and measurement error parameters, and present inferential procedures that remain valid when relative abundance parameters lie on the boundary of the parameter space. We demonstrate applications of our method to relative abundance estimation under complex measurement error, and to model evaluation and hypothesis testing in cross-protocol measurement experiments. We show consistency and weak convergence of our estimators under mild conditions and empirically evaluate their performance on a dataset of measurements on specimens of known composition and via simulation. We conclude with a discussion of advantages of our approach and areas for future research.

3.2 A measurement error model for microbiome data

We propose a measurement error model that accounts for both contaminant reads and taxon-dependent distortion of sample community structure. We use the term "read" as a catch-all for the measurement output of microbiome sequencing experiments, but do not require that observations be integer-valued. Similarly, "sample" refers to the unit of sequencing; "specimen" refers to a unique source of genetic material that may be repeatedly sampled for sequencing; and "taxon" refers to a grouping of organisms (e.g., species, strains or cell types).

Let $W_i = (W_{i1}, \ldots, W_{iJ})$ denote observed reads from taxa $1, \ldots, J$ in sample i. A common modeling assumption for high-throughput sequencing data is

$$\mathbb{E}[W_i|\mathbf{p}_i, \gamma_i] = \exp(\gamma_i)\mathbf{p}_i \tag{3.1}$$

where $\mathbf{p}_i \in \mathbb{S}^{J-1}$ is the unknown relative abundances of taxa $1, \ldots, J$ and $\exp(\gamma_i) \in \mathbb{R}^+$ is a sampling intensity parameter (throughout, we let $\mathbb{S}^{J-1}$ denote the closed $J-1$-dimensional simplex). While the simplicity of this model is appealing, it poorly describes actual microbiome sequencing data because taxa are not all detected equally well [McLaren et al., 2019]. To account for this, we begin by considering the model

$$\mathbb{E}[W_i|\mathbf{p}_i, \beta, \gamma_i] = \exp(\gamma_i)(\exp(\beta) \circ \mathbf{p}_i) \tag{3.2}$$

where $\beta = \beta_1, \ldots, \beta_J$ represents the detection effects for taxa $1, \ldots, J$ in a given experiment ($\circ$ indicates element-wise multiplication). As an identifiability constraint, we set $\beta_J = 0$, and so interpret $\exp(\beta_j), j = 1, \ldots, J$, as the degree of multiplicative over- or under-detection of taxon j relative to taxon J.

We now generalize model (3.2) to multiple samples and one or more experimental protocols. For a study involving n samples of K unique sources of samples ($n \geq K$), we define the *sample design matrix* $\mathbf{Z} \in \mathbb{R}^{n \times K}$ to link samples to specimens. In most experiments, $Z_{ik} = \mathbf{1}_{\{\text{sample } i \text{ taken from specimen } k\}}$, but we note that more complex designs are possible. For

example, if sample i is a $1:1$ mixture of microbial communities k and k', we may specify $Z_{ik} = Z_{ik'} = \frac{1}{2}$. Letting $\mathbf{p}$ be a $K \times J$ matrix such that the k-th row of $\mathbf{p}$ gives the true relative abundances of taxa in specimen k, we have that the relative abundance vector for sample i is $(\mathbf{Zp})_{i\cdot}$. To reflect that a sample's relative abundance must be a convex combination of K sources' relative abundances, we require $\mathbf{Z}_{i\cdot} \in \mathbb{S}^{K-1}$. We allow differing detections across samples to be specified via the *detection design matrix* $\mathbf{X} \in \mathbb{R}^{n \times p}$. For example, if samples are processed using one of p different protocols, we might specify $X_{iq} = \mathbf{1}_{\{\text{sample } i \text{ processed with protocol } q\}}$. Accordingly, we now consider a detection effect matrix $\boldsymbol{\beta} \in \mathbb{R}^{p \times J}$. As above, we impose identifiability constraint $\boldsymbol{\beta}_{\cdot J} = \mathbf{0}_p$. Therefore, one generalization of model (3.2) is

$$\mathbb{E}\left[\text{reads due to contributing samples} | \boldsymbol{\beta}, \mathbf{p}, \boldsymbol{\gamma}, \mathbf{Z}, \mathbf{X}\right] = \mathbf{D}_{\boldsymbol{\gamma}}\left(\mathbf{Zp}\right) \circ \exp\left(\mathbf{X}\boldsymbol{\beta}\right), \tag{3.3}$$

where $\mathbf{D}_{\boldsymbol{\gamma}} = \text{diag}(\exp(\boldsymbol{\gamma}))$, and where exponentiation is element-wise.

We now extend this model to reflect contributions of contaminant sources to the expected number of observed reads. We consider $\tilde{K}$ sources of contamination with relative abundance profiles given in the rows of $\tilde{\mathbf{p}} \in \mathbb{R}^{\tilde{K} \times J}$.[1] To link sources of contamination to samples, we let $\tilde{\mathbf{Z}} \in \mathbb{R}^{n \times \tilde{K}}$ be a *spurious read design* matrix. Most commonly we expect $\tilde{Z}_{i\tilde{k}} = \mathbf{1}_{\{\text{source } \tilde{k} \text{ may contribute reads to sample } i\}}$, but we give an example of an analysis with more complex $\tilde{\mathbf{Z}}$ in Section 3.5.2. Then, along with contaminant read intensities $\tilde{\boldsymbol{\gamma}} = [\tilde{\gamma}_1, \ldots, \tilde{\gamma}_{\tilde{k}}]^{\mathrm{T}}$, we propose to model

$$\mathbb{E}\left[\text{reads due to spurious sources} | \boldsymbol{\gamma}, \tilde{\boldsymbol{\gamma}}, \tilde{\mathbf{p}}, \tilde{\mathbf{Z}}\right] = \mathbf{D}_{\boldsymbol{\gamma}}\tilde{\mathbf{Z}}\left[\tilde{\mathbf{p}} \circ \exp\left(\tilde{\boldsymbol{\gamma}}\mathbf{1}_J^{\mathrm{T}}\right)\right]. \tag{3.4}$$

While we could incorporate a detection design matrix for contaminant reads (replacing (3.4) with $\mathbf{D}_{\boldsymbol{\gamma}}\tilde{\mathbf{Z}}\left[\tilde{\mathbf{p}} \circ \exp\left(\tilde{\boldsymbol{\gamma}}\mathbf{1}_J^T + \tilde{\mathbf{X}}\boldsymbol{\beta}\right)\right]$ for $\tilde{\mathbf{X}} \in \mathbb{R}^{\tilde{K} \times p}$), for most practical applications it is sufficient to identify $\tilde{\mathbf{p}}$ up to detection distortion. Therefore, combining models (3.3) and (3.4),

[1] We require $\tilde{K}$ to be treated as known. In practice, it is typically unknown. In some cases, this may matter fairly little, as we can treat multiple sources of contamination introduced at similar points in a measurement process as a single cumulative contaminant source. Accurately characterizing component sources of contamination may be more important when contamination occurs at differing measurement steps and hence potentially has systematically different impact depending on the step at which it occurs.

we propose the following mean model for next-generation sequencing data $\mathbf{W} \in \mathbb{R}^{n \times J}$:

$$\boldsymbol{\mu} := \mathbb{E}\left[\mathbf{W} | \boldsymbol{\beta}, \mathbf{p}, \boldsymbol{\gamma}, \tilde{\mathbf{p}}, \tilde{\boldsymbol{\gamma}}\right] = \mathbf{D}_{\boldsymbol{\gamma}}\left(\mathbf{Z}\mathbf{p}\right) \circ \exp\left(\mathbf{X}\boldsymbol{\beta}\right) + \mathbf{D}_{\boldsymbol{\gamma}}\tilde{\mathbf{Z}}\left[\tilde{\mathbf{p}} \circ \exp\left(\tilde{\boldsymbol{\gamma}}\mathbf{1}_J^T\right)\right]. \tag{3.5}$$

3.3 Estimation and optimization

We propose to estimate parameters $\boldsymbol{\theta}^\star := (\boldsymbol{\theta}, \boldsymbol{\gamma}) := (\boldsymbol{\beta}, \mathbf{p}, \tilde{\mathbf{p}}, \tilde{\boldsymbol{\gamma}}, \boldsymbol{\gamma})$ either by maximum likelihood or by maximum weighted likelihood. We use likelihoods to define M-estimators and do not require or assume that the distribution of $\mathbf{W}$ lies in any particular parametric class. We show consistency and weak convergence of our estimators of $\boldsymbol{\theta} := (\boldsymbol{\beta}, \mathbf{p}, \tilde{\mathbf{p}}, \tilde{\boldsymbol{\gamma}})$ under mild conditions in Supporting Information (SI) Section 2. Note that elements of $\boldsymbol{\gamma} \in \mathbb{R}^n$ are sample-specific nuisance parameters. We use $\boldsymbol{\theta}_0$ to denote the true value of $\boldsymbol{\theta}$. Our unweighted objective is given by a Poisson log-likelihood:

$$M_n^\star(\boldsymbol{\theta}^\star) := \frac{1}{n} l_n(\boldsymbol{\theta}^\star) = \frac{1}{n}\mathbf{1}^T\left[\mathrm{vec}\left(\mathbf{W}\right) \circ \log\left(\mathrm{vec}\left(\boldsymbol{\mu}\left(\boldsymbol{\theta}^\star\right)\right)\right) - \mathrm{vec}\left(\boldsymbol{\mu}\left(\boldsymbol{\theta}^\star\right)\right)\right]. \tag{3.6}$$

We use $M_n(\boldsymbol{\theta})$ to indicate the profile log-likelihood $\sup_{\boldsymbol{\gamma} \in \mathbb{R}^n} M_n^\star(\boldsymbol{\theta}, \boldsymbol{\gamma})$. Consistency of the maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}_0$ does not require W_{ij} to follow a Poisson distribution. However, the Poisson estimator will in general be inefficient if the relationship between $\mathbb{E}[W_{ij} | Z_i, X_i, \tilde{Z}_i, \gamma_i, \boldsymbol{\theta}_0]$ and $\mathrm{Var}[W_{ij} | Z_i, X_i, \tilde{Z}_i, \gamma_i, \boldsymbol{\theta}_0]$ is not linear [McCullagh, 1983]. Therefore, to obtain a more efficient estimator, we also consider maximizing a reweighted Poisson log-likelihood, with weights chosen on the basis of a flexibly estimated mean-variance relationship. We motivate our specific choice of weights via the Poisson score equations (see SI Section 1). For weighting vector $\hat{\mathbf{v}} \in \mathbb{R}_+^{nJ}$ such that $\mathbf{1}^T\hat{\mathbf{v}} = nJ$, we define the weighted Poisson log-likelihood as

$$M_n^{\star\hat{\mathbf{v}}}(\boldsymbol{\theta}^\star) := \frac{1}{n}\hat{\mathbf{v}}^T\left[\mathrm{vec}\left(\mathbf{W}\right) \circ \log\left(\mathrm{vec}\left(\boldsymbol{\mu}\left(\boldsymbol{\theta}^\star\right)\right)\right) - \mathrm{vec}\left(\boldsymbol{\mu}\left(\boldsymbol{\theta}^\star\right)\right)\right]. \tag{3.7}$$

We define $M_n^{\hat{\mathbf{v}}_n}(\boldsymbol{\theta})$ by analogy with $M_n(\boldsymbol{\theta})$ above. We select $\hat{\mathbf{v}}$ via a centered isotonic regression [Oron and Flournoy, 2017] of squared residuals $\mathrm{vec}\left[(\mathbf{W} - \hat{\boldsymbol{\mu}})^2\right]$ on fitted means $\mathrm{vec}\left[\hat{\boldsymbol{\mu}}\right]$ obtained from the unweighted objective. Full details are given in SI Section 1, but briefly, we set $\hat{v}_{ij} \propto \frac{\hat{\mu}_{ij}+1}{\hat{\sigma}_{ij}^2+1}$, where $\hat{\sigma}_{ij}^2$ is the monotone regression fitted value for μ_{ij}.

When the mean model is correctly specified, the estimators defined by optima of the weighted or unweighted Poisson likelihoods given above are consistent for the true value of θ and converge weakly to well-defined limiting distributions at $\sqrt{n}$ rate (the form of this distribution in general depends on the true value of θ). We leverage an approach from Van der Vaart [2000] to prove consistency, and we combine a bracketing argument with a directional delta method theorem of Dümbgen [1993] to show weak convergence. Details of conditions and proofs are given in Supporting Information Section 2.

Computing maximum (weighted) likelihood estimates of $\boldsymbol{\theta}^\star$ is a constrained optimization problem, as the relative abundance parameters in our model are simplex-valued, and the estimate may lie on the boundary of the simplex. Therefore, we minimize $f_n(\boldsymbol{\theta}^\star) = -M_n^\star(\boldsymbol{\theta}^\star)$ or $f_n(\boldsymbol{\theta}^\star) = -M_n^{\star\hat{\mathbf{v}}_n}(\boldsymbol{\theta}^\star)$ in two steps. In the first step, we employ the barrier method, converting our constrained optimization problem into a sequence of unconstrained optimizations, permitting solutions progressively closer to the boundary. That is, for barrier penalty parameter t, we update $\boldsymbol{\theta}$ as

$$\boldsymbol{\theta}^{\star(r+1)} = \arg\min_{\boldsymbol{\theta}} \left(f_n(\boldsymbol{\theta}^\star) + \frac{1}{t^{(r)}} \left[\sum_{k=1}^{K} \sum_{j=1}^{J} -\log p_{kj} + \sum_{\tilde{k}=1}^{\tilde{K}} \sum_{j=1}^{J} -\log \tilde{p}_{\tilde{k}j} \right] \right) \tag{3.8}$$

$$\text{subject to } \sum_{j=1}^{J} p_{kj} = 1 \text{ and } \sum_{j=1}^{J} \tilde{p}_{\tilde{k}j} = 1 \text{ for all } k, \tilde{k} \text{ s.t. } \mathbf{p}_k, \tilde{\mathbf{p}}_{\tilde{k}} \text{ unknown} \tag{3.9}$$

and set $t^{(r+1)} = at^{(r)}$ where $a > 1$ is a prespecified incrementing factor, iterating until $t^{(r)} > t_{\text{cutoff}}$ for large t_{cutoff}. In practice we find that $t^{(0)} = 1$, $a = 10$, and $t_{\text{cutoff}} = 10^{12}$ yield good performance. We enforce the sum-to-one constraints (3.9) by reparametrizing $\mathbf{p}$ and $\tilde{\mathbf{p}}$ as $\boldsymbol{\rho}$ and $\tilde{\boldsymbol{\rho}}$, with $\rho_{kj} := \log \frac{p_{kj}}{p_{kJ}}$ and $\tilde{\rho}_{\tilde{k}j} := \log \frac{\tilde{p}_{\tilde{k}j}}{\tilde{p}_{\tilde{k}J}}$ for $j = 1, \ldots, J-1$, which are well-defined because of the logarithmic penalty terms in (3.8).

In the second step of our optimization procedure, we apply a constrained Newton algorithm within an augmented Lagrangian algorithm to allow elements of $\hat{\mathbf{p}}$ and $\hat{\tilde{\mathbf{p}}}$ to equal zero.

Iteratively in each row $\mathbf{p}_k$ of $\mathbf{p}$, we approximately solve

$$\arg\min_{\mathbf{p}_k} f_n(\mathbf{p}_k) \quad \text{subject to } \sum_{j=1}^{J} p_{kj} = 1,\ p_{kj} \geq 0 \text{ for } j = 1, \dots J \tag{3.10}$$

where $f_n(\mathbf{p}_k)$ is the objective function f_n considered as a function of only $\mathbf{p}_k$, with all other parameters fixed at values obtained in previous optimization steps. To do this, we choose update directions for $\mathbf{p}_k$ via an augmented Lagrangian algorithm of Bazaraa [2006] applied to

$$\mathcal{L}_k := Q_k^{(t)} + \nu \left[\sum_{j=1}^{J} p_{kj} - 1 \right] + \mu \left[\sum_{j=1}^{J} p_{kj} - 1 \right]^2 \tag{3.11}$$

where $Q_k^{(t)}$ is a quadratic approximation to $f_n(\mathbf{p}_k)$ at $\mathbf{p}_k^{(t)}$ and ν and μ are chosen using the algorithm of Bazaraa [2006]. The augmented Lagrangian algorithm iteratively updates ν and μ until solutions to $\mathcal{L}_k$ satisfy $|\sum_{j=1}^{J} p_{kj} - 1| < \epsilon$ for a small prespecified value of ϵ (we use 10^{-10} by default). Within each iteration of the augmented Lagrangian algorithm, we minimize $\mathcal{L}_k$ via fast non-negative least squares to preserve nonnegativity of $\mathbf{p}_k$. Through the augmented Lagrangian algorithm, we obtain a value $\mathbf{p}_{\mathcal{L}_k}^{(t)}$ of $\mathbf{p}_k$ that minimizes $\mathcal{L}_k^{(t)}$ (at final values of ν and μ) subject to nonnegativity constraints. Our update direction for $\mathbf{p}_k$ is then given by $\mathbf{s}_k^{(t)} = \mathbf{p}_{\mathcal{L}_k}^{(t)} - \mathbf{p}_k^{(t)}$. We conduct a backtracking line search in direction $\mathbf{s}_k^{(t)}$ to find an update $\mathbf{p}_k^{(t+1)}$ that decreases $f_n(\mathbf{p}_k)$.[2]

3.4 Inference for $\mathbf{p}$ *and* $\boldsymbol{\beta}$

We now address construction of confidence intervals and hypothesis tests. We focus on parameters $\boldsymbol{\beta}$ and $\mathbf{p}$, which we believe to be the most common targets for inference. To derive both marginal confidence intervals and more complex hypothesis tests, we consider a general setting in which we observe some estimate $\hat{\phi} = \phi(\mathbb{P}_n)$ of population quantity $\phi = \phi(P)$, where $\mathbb{P}_n$ is the empirical distribution corresponding to a sample $\left\{ \left(\mathbf{W}_i, \mathbf{Z}_i, \mathbf{X}_i, \tilde{\mathbf{Z}}_i \right) \right\}_{i=1}^{n}$,

[2]We typically do not repeat this procedure for $\tilde{\mathbf{p}}$, as $\tilde{\mathbf{p}}$ is a nuisance parameter and it suffices to approximate the MLE for $\tilde{p}$ with a value arbitrarily close to (but not lying on) the boundary of the simplex. However, this second optimization step can be applied to $\tilde{\mathbf{p}}$ as well if desired.

P is its population analogue, and ϕ is a Hadamard directionally differentiable map into the parameter space Θ or into $\mathbb{R}$. To derive marginal confidence intervals for $\mathbf{p}$ and $\boldsymbol{\beta}$, we let $\phi(\mathbb{P}_n) = \hat{\boldsymbol{\theta}}_n$ with $\phi(P) = \boldsymbol{\theta}_0$. For hypothesis tests involving multiple parameters, we specify $\phi(\mathbb{P}_n)$ as $2\left[\sup_{\boldsymbol{\theta}\in\Theta} M_n(\boldsymbol{\theta}) - \sup_{\boldsymbol{\theta}\in\Theta_0} M_n(\boldsymbol{\theta})\right]$ with population analogue $\phi(P) = 0$ under $H_0 : \boldsymbol{\theta} \in \Theta_0$. In each case, we estimate the asymptotic distribution of $a(n)\left(\phi(\mathbb{P}_n) - \phi(P)\right)$ for an appropriately chosen $a(n) \to \infty$.

As our model includes parameters that may lie on the boundary of the parameter space, the limiting distributions of our estimators and test statistics in general do not have a simple distributional form [Geyer, 1994], and the multinomial bootstrap will fail to produce asymptotically valid inference [Andrews, 2000]. To address this, we employ a Bayesian subsampled bootstrap [Ishwaran et al., 2009], which consistently estimates the asymptotic distribution of our estimators when the true parameter is on the boundary. Let $\mathbb{P}_n^{\boldsymbol{\xi}}$ be a weighted empirical distribution $\sum_{i=1}^n \xi_{i,n}\mathbf{1}_{\mathbf{W}_i}$ with weights $\boldsymbol{\xi} \sim G$ for $G \sim \text{Dirichlet}\left(\frac{m}{n}\mathbf{1}_n\right)$. Then the bootstrap estimator $a(m)\left(\phi\left(\mathbb{P}_n^{\boldsymbol{\xi}}\right) - \phi\left(\mathbb{P}_n\right)\right)$ converges weakly to the limiting distribution of $a(n)\left(\phi\left(\mathbb{P}_n\right) - \phi\left(P\right)\right)$ if we choose $m = m(n)$ such that $\lim_{n\to\infty} m = \infty$ and $\lim_{n\to\infty} \frac{m}{n} = 0$ [Ishwaran et al., 2009]. We explore finite-sample behavior of the proposed bootstrap estimators with $m = \sqrt{n}$ in Section 3.6, finding good Type 1 error control. We note as well that if a reweighted likelihood is used, we condition on weights used to construct this likelihood in our bootstrap and do not recalculate these weights inside each iteration.

To derive marginal confidence intervals for elements of $\boldsymbol{\theta}$, we let $a(n) = \sqrt{n}$ and $\phi(P) = \boldsymbol{\theta}(P)$. Then $\sqrt{m}(\hat{\boldsymbol{\theta}}_n^{\xi} - \hat{\boldsymbol{\theta}}_n)$ has the same limiting distribution as $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$. Therefore, for $\hat{L}_c^q$ the c-th bootstrap quantile of the q-th element of $\sqrt{m}(\hat{\boldsymbol{\theta}}_n^{\xi} - \hat{\boldsymbol{\theta}}_n)$ and $\hat{\theta}_q$ the q-th element of the maximum (weighted) likelihood estimate $\hat{\boldsymbol{\theta}}$, an asymptotically $100(1-\alpha)\%$ marginal confidence interval for θ_q is given by $\left(\hat{\theta}_q - \frac{1}{\sqrt{n}}\hat{L}^q_{1-\alpha/2},\ \hat{\theta}_q - \frac{1}{\sqrt{n}}\hat{L}^q_{\alpha/2}\right)$.

As it may be of interest to test hypotheses about multiple parameters while leaving other parameters unrestricted (e.g., $\boldsymbol{\beta} = \mathbf{0}$ with unrestricted elements of $\mathbf{p}$), we also develop a procedure to test hypotheses of the form $H_0 : \{\theta_k = c_k : k \in \mathcal{K}_0\}$ for a set of parameter indices $\mathcal{K}_0$ against alternatives with $\boldsymbol{\theta}$ unrestricted. Letting Θ_0 indicate the parameter space

under H_0 and Θ indicate the full parameter space, we conduct tests using test statistic $T_n := nT(\mathbb{P}_n) := 2n[\sup_{\boldsymbol{\theta}\in\Theta} M_n(\boldsymbol{\theta}) - \sup_{\boldsymbol{\theta}\in\Theta_0} M_n(\boldsymbol{\theta})]$. As noted above, T_n is in general not asymptotically χ^2 if (unknown) elements of $\mathbf{p}$ or $\tilde{\mathbf{p}}$ lie at the boundary, and so we instead approximate the null distribution of T_n by bootstrap resampling from an empirical distribution projected onto an approximate null; this is closely related to the approach suggested by Hinkley [1988]. Let $\mathring{\mu}_{ij}$ denote $\exp(-\gamma_i)$ times the expectation of W_{ij} under the full model; $\mathring{\mu}^0_{ij}$ denote the analogous quantity under H_0; and define $W^0_{ij} = W_{ij}\frac{\mathring{\mu}^0_{ij}}{\mathring{\mu}_{ij}}$ if $\mathring{\mu}_{ij} > 0$ and otherwise set $W^0_{ij} = W_{ij} = 0$. In practice, we do not know $\mathring{\mu}_{ij}$ or $\mathring{\mu}^0_{ij}$, so we replace W^0_{ij} with $\hat{W}^0_{ij} = \frac{\hat{\mathring{\mu}}^0_{ij}}{\hat{\mathring{\mu}}_{ij}}$, where $\hat{\mathring{\mu}}_{ij}$ and $\hat{\mathring{\mu}}^0_{ij}$ are, up to proportionality constant $\exp(\hat{\gamma}_i)$, fitted means for W_{ij} under the full and null models. After constructing $\hat{\mathbf{W}}^0$, we rescale its rows so row sums of $\hat{\mathbf{W}}^0$ and $\mathbf{W}$ are equal. We then approximate the null distribution of T via bootstrap draws from $mT(\mathbb{P}^{\boldsymbol{\xi}}_{0n})$ where $\mathbb{P}^{\boldsymbol{\xi}}_{0n}$ is the Bayesian subsampled bootstrap distribution on $\hat{\mathbf{W}}^0_n$. We reject H_0 at level α if the observed likelihood ratio test statistic is larger than the $1-\alpha$ quantile of the bootstrap estimate of its null distribution, or equivalently, if $T_n \geq \tilde{L}^m_{1-\alpha}$ for $\tilde{L}^m_{1-\alpha}$ the $1-\alpha$ quantile of $mT(\mathbb{P}^{\boldsymbol{\xi}}_{0n})$.

3.5 Data Examples

3.5.1 Comparing detection effects across experiments

We now demonstrate the utility of our model in comparing different experimental protocols for high-throughput sequencing of microbial communities. We consider data generated in the Phase 2 experiment of Costea et al. [2017] (see also McLaren et al. [2019]), wherein ten human fecal specimens (labeled $1, 2, \ldots, 8$, A and B) were mixed with a synthetic community of 10 taxa and prepared for shotgun metagenomic sequencing according to three different sample preparations (labeled H, Q, and W; samples A and B were only analyzed with preparation Q). The synthetic community was also sequenced alone. Raw sequencing data was processed into taxon abundance data using MetaPhlAn2 [Truong et al., 2015] by McLaren et al. [2019]. In addition to sequencing data, taxon abundances in the synthetic

community were also measured using flow cytometry. We treat both sequencing and flow cytometry measurements as outcomes $\mathbf{W}$. We are interested in comparing the detection of taxa in the synthetic community across protocols H, Q and W relative to flow cytometry. We are specifically interested in testing the null hypothesis that all sequencing protocols share the same detection effects. To accomplish this, we estimate the 3×10 matrix $\boldsymbol{\beta}$ (we set $\boldsymbol{\beta}_{\cdot 10} = \mathbf{0}_3$ to ensure identifiability). Each row of $\boldsymbol{\beta}$ corresponds to a sequencing protocol, and each column corresponds to a taxon. For details regarding the specification of $\mathbf{X}$ and other model parameters, see SI Section 4.1. Under our model, $\exp(\beta_{1j})$, $\exp(\beta_{1j} + \beta_{2j})$, and $\exp(\beta_{1j} + \beta_{3j})$ give the degree of over- or under-detection of taxon j relative to taxon 10 under protocols H, Q, and W, respectively. We compare this model to a submodel in which $\beta_{kj} = 0$ for $k = 2, 3$ and all j. Under this null hypothesis, taxon detections relative to flow cytometry do not differ across protocols.

To compare predictive performance of these models, we perform 10-fold cross-validation on each model (see SI Section 4 for details). We use a bootstrapped likelihood ratio test to formally test our full model against the null submodel. We use the Bayesian subsampled bootstrap with $m = \sqrt{n}$ to illustrate its applicability, however, a multinomial bootstrap would also be appropriate as all parameters are in the interior of the parameter space in this case. In addition, we report point estimates and bootstrapped marginal 95% confidence intervals for detection effects estimated for each protocol under the full model. We also compare our results to MetaPhlAn2's "plug-in" estimate of each sample's composition.

Figure 3.1 summarizes 10-fold cross-validated estimated relative abundances from the full model, null model and plug-in estimates. We observe substantially better model fit for the full model (top row) than for the null model (middle row). At each flow cytometric relative abundance, cross-validated estimates from the full model are generally centered around the line $y = x$ (dashed line), whereas estimates from the null model exhibit substantial bias for some taxa. A bootstrapped likelihood ratio test of the null model (i.e., $H_0 : \boldsymbol{\beta}_{2\cdot} = \boldsymbol{\beta}_{3\cdot} = \mathbf{0}$) against the full model reflects this, and we reject the null with $p < 0.001$. Both the full and null models outperform the plug-in estimates of sample composition (bottom row), which

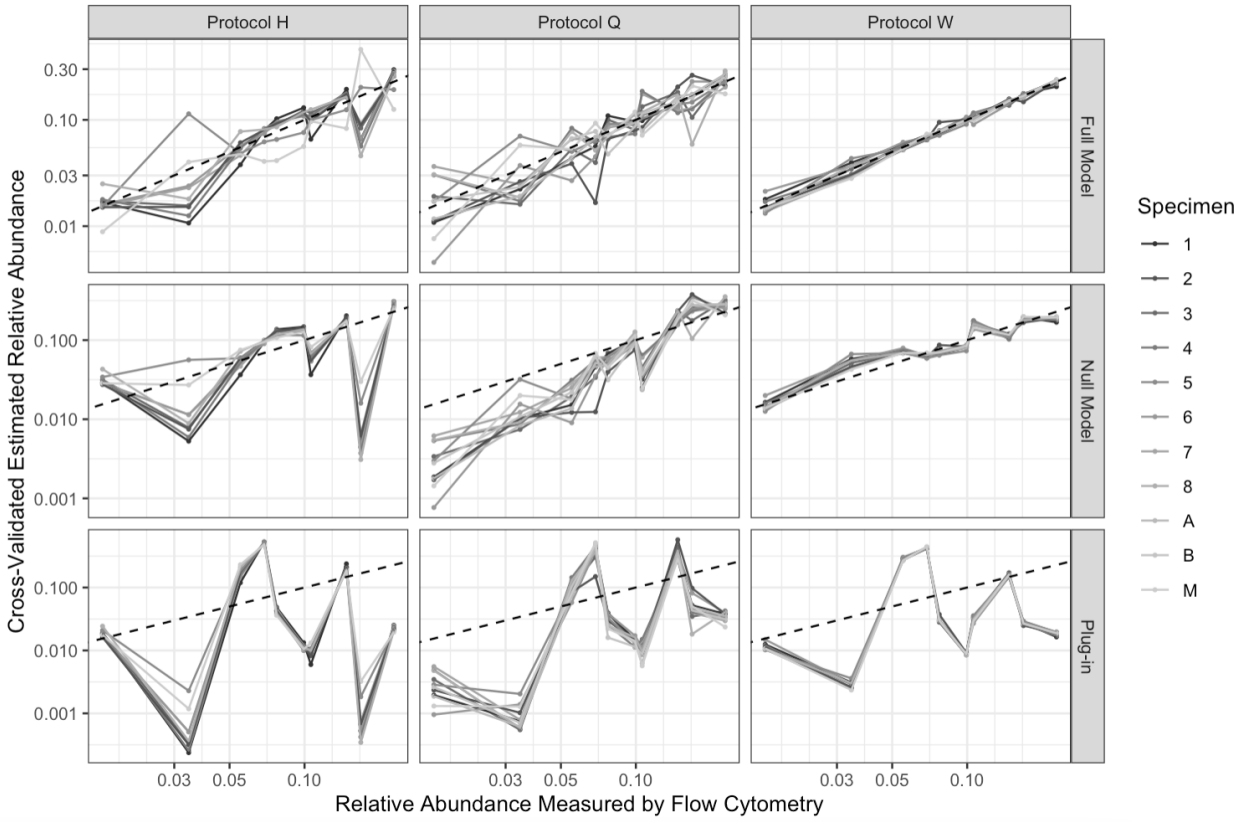

Figure 3.1: 10-fold cross-validated estimates of relative abundance (y-axis) in Costea et al. [2017] samples that were measured by whole-genome sequencing as well as plug-in estimates of relative abundance (bottom row). On the x-axis are relative abundance estimates obtained via flow cytometry (mean concentration in each taxon divided by sum across taxa). The top row contains estimates produced by a model containing separate detection effects $\boldsymbol{\beta}$ for each protocol; the middle row contains output from a model assuming a single common detection effect across protocols; and the bottom row contains "plug-in" estimates from MetaPhlAn2 output. Results for each protocols H, Q, and W are given in the leftmost, center, and rightmost columns, respectively. Within each pane, estimated relative abundances for the same sample are connected by line segments, and the line $y = x$ is indicated with a dotted line.

produces substantially biased estimates of relative abundance relative to a flow cytometry standard. We report point estimates and marginal 95% confidence intervals for $\boldsymbol{\beta}$ in SI Section 4.3.

Using the full model, we estimate relative abundances with substantially greater precision under protocol W (top right) than under either other protocol (top left and center). This appears to be primarily due to lower variability in measurements taken via protocol W (bottom row). Our finding of greater precision of protocol W contrasts with Costea et al. [2017], who recommend protocol Q as a "potential benchmark for new methods" on the basis of median absolute error of centered-log-ratio-transformed plug-in estimates of relative abundance against flow cytometry measurements (as well as on the basis of cross-laboratory measurement reproducibility, which we do not examine here). The recommendations of Costea et al. [2017] are driven by performance of plug-in estimators subject to considerable bias, whereas we are able to model and remove a large degree of bias and can hence focus on residual variation after bias correction. We also note that Costea et al. [2017] did not use MetaPhlAn2 to construct abundance estimates, which may partly account our different conclusions.

3.5.2 Estimating contamination via dilution series

We next illustrate how to use our model to estimate and remove contamination in samples. We consider data from Karstens et al. [2019], who generated 9 samples via three-fold dilutions of a synthetic community containing 8 distinct strains of bacteria which each account for 12.5% of the DNA in the community. Despite only 8 strains being present in the synthetic community, 248 total strains were identified based on sequencing (see SI Section 5.1 for data processing details). We refer to the 8 strains in the synthetic community as "target" taxa and other strains as "off-target." Note that Karstens et al. [2019] identified one strain as a likely mutant of synthetic community member *S. enterica*, and we refer to this strain as *S. [unclassified]*.

To evaluate the performance of our model, we perform three-fold cross-validation and

estimate relative abundance in the hold-out fold. We consider $\mathbf{p} \in \mathbb{R}^{2\times 248}$, where the first row contains the known composition of the training fold $\left(\mathbf{0}_{240}^T \ \frac{1}{8}\mathbf{1}_8^T\right)$ and the second row is unknown and the target of inference (see SI Section 5.2 for full model specification). We evenly balance dilutions across folds, and using our proposed reweighted estimator to fit a model that accounts for the serial dilutions. We set $\tilde{K} = 1$ and let $\tilde{\mathbf{Z}}_i = \exp(\gamma_i) \times 3^{d_i}$, where d_i is the number of three-fold dilutions sample i has undergone. This model reflects the assumption that the ratio of expected contaminant reads to expected non-contaminant reads is proportional to 3^{d_i}. To avoid improperly sharing information about contamination amounts across folds, we further parametrize $\tilde{\mathbf{Z}}$ in terms of a fixed, unknown parameter $\tilde{\alpha} \in \mathbb{R}$ by multiplying each row of $\tilde{\mathbf{Z}}$ corresponding to a held-out sample by $\exp(\tilde{\alpha})$. This preserves information about relative dilution within the held-out fold without treating samples in the training and held-out folds as part of the same dilution series. We model a single differential detection effect β_j for each of the 8 taxa in the synthetic community, setting $\beta_J = 0$ for the reference taxon *L. fermentum* for identifiability. Because β_j is not identifiable for off-target taxa, we also fix $\beta_j = 0$ for $j = 1, \ldots, 240$.

Figure 3.2 shows data from Karstens et al. [2019] along with summaries of our analysis. Our estimate for the relative abundance of taxa in samples in the held-out fold $\mathbf{p}_{2\cdot}$ improves on the performance of plug-in estimators (Figure 3.2, left panels) by taking into account two forms of structure in the Karstens et al. [2019] data. First, in each successive three-fold dilution, we observe approximately three times more contamination relative to the number of non-contaminant reads (Figure 3.2, top right). In addition, our model accounts for the degree of under- (or over-) detection of target taxa relative to *L. fermentum*. We observe that taxon detection is reasonably constant across dilutions (Figure 3.2, bottom right). However, we do observe greater variability in taxon detections at higher dilutions, most likely because we observe comparatively few reads ($\sum_{j=1}^{248} W_{\{i:d_i=1\}j} \approx 227,000$ while $\sum_{j=1}^{248} W_{\{i:d_i=8\}j} \approx 8,000$).

In terms of root mean squared error (RMSE) $\sqrt{1/J \sum_{j=1}^{J} (\hat{p}_j - p_j)^2}$, our cross-validated estimates (0.0037, 0.0073, and 0.0033) substantially outperform the "plug-in" estimates given by sample read proportions in any of these dilutions (median 0.017; range 0.013 – 0.022).

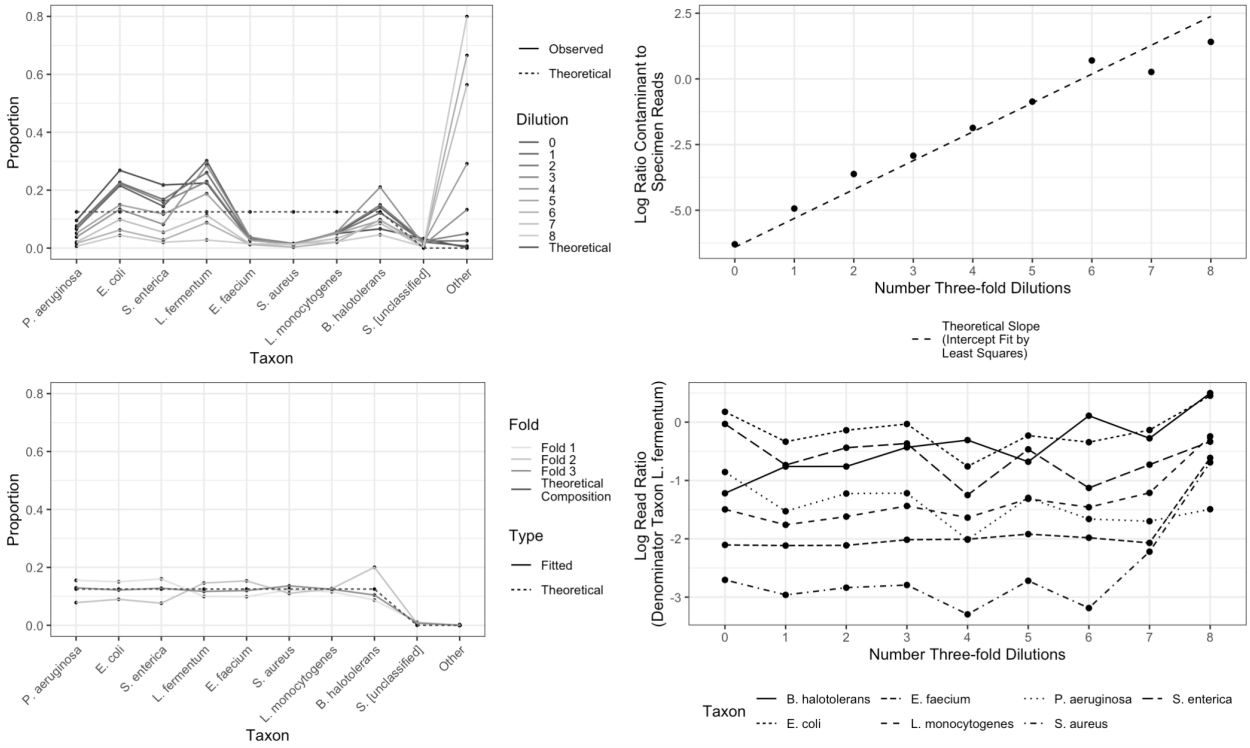

Figure 3.2: Data from Karstens et al. [2019], and estimates and summaries from the fitted model. (Top left) Observed read proportions by taxon and dilution, with theoretical synthetic composition indicated by dotted line. (Top right) The log-ratio of total contaminant reads to total non-contaminant reads (excluding *S. [unclassified]*). The dotted line has slope log(3), with intercept fit via least squares. (Bottom left) Fitted read proportions obtained from each cross-validation fold, with theoretical synthetic composition indicated by dotted line. Every fold produces abundance estimates that improve over observed read proportions. (Bottom right) The log-ratio of reads for target taxa to reference taxon *L. fermentum* is relatively constant across increasing numbers of dilutions.

This is not an artifact of incorporating information from 3 samples in each cross-validation fold, as pooling reads across all samples yields an estimator with RMSE 0.014.

Fitting a model to this relatively small dataset and evaluating its performance using cross validation prohibits the reasonable construction of confidence intervals. Therefore, to evaluate the performance of our proposed approach to generating confidence intervals, we also fit a model which treats all samples as originating from a single specimen of unknown composition. $\boldsymbol{\beta}$ is not identifiable in this setting, so we set it equal to zero and do not estimate it. We set $\tilde{K} = 1$ and $\tilde{\mathbf{Z}}_i = \exp(\gamma_i) \times 3^{d_i}$ as before (the need for $\tilde{\alpha}$ is alleviated). Strikingly, under this model, marginal 95% confidence intervals for elements of $\mathbf{p}$ included zero for 238 out of 240 off-target taxa (empirical coverage of 99.2% when true $p_{kj} = 0$). No interval estimates for target taxa included zero. This suggests that applying our proposed approach to data from a dilution series experiment can aid in evaluating whether taxa detected by next-generation sequencing are actually present in a given specimen.

3.6 Simulations

3.6.1 Sample size and predictive performance

To evaluate the predictive performance of our model on microbiome datasets of varying size, we use data from 65 unique specimens consisting of synthetic communities of 1, 2, 3, 4, or 7 species combined in equal abundances. Brooks et al. [2015] performed 16S amplicon sequencing on 80 samples of these 65 specimens across two plates of 40 samples each. Very few reads in this dataset were ascribed to taxa outside the 7 present by design, so we limit analysis to these taxa. To explore how prediction error varies with number of samples of known composition, we fit models treating randomly selected subsets of $n_{\text{known}} \in \{3, 5, 10, 20\}$ samples per plate as known. In each model, we included one source of unknown contamination for each plate and estimate a detection vector $\boldsymbol{\beta} \in \mathbb{R}^J$. For each n_{known}, we drew 100 independent sets of samples to be treated as known, requiring that each set satisfy an identifiability condition in $\boldsymbol{\beta}$ (see SI Section 6.1). On each set, we fit both the unweighted and reweighted models,

treating n_{known} samples as having known composition and $80 - n_{\text{known}}$ samples as arising from unique specimens of unknown composition.

We observe similar RMSE for the reweighted and unweighted estimators. For n_{known} equal to 3, 5, 10, and 20, we observe RMSE 0.041, 0.037, 0.035, and 0.032 (to 2 decimal places) for both estimators. By comparison, the RMSE for the plugin estimator $\frac{W_{ij}}{\sum_{j=1}^{J} W_{ij}}$ is 0.173. Notably, RMSE decreases but does not approach zero as larger number of samples are treated as known, which reflects that we estimated each relative abundance profile on the basis of a single sample.

With respect to correctly estimating p_{kj} when $p_{kj} = 0$, we again see very similar performance of the reweighted and unweighted Poisson estimators. Out of 100 sets, unweighted estimation yields $\hat{p}_{kj} = 0$ for $53\%, 55\%, 59\%$, and 64% of $\{k, j\}$ pairs for which $p_{kj} = 0$ (n_{known} equal to 3, 5, 10, and 20). The corresponding figures for the reweighted estimator are $53\%, 55\%, 58\%$, and 63%. The plug-in estimator sets 51% of these relative abundances equal to zero. While we observe the proportion of theoretical zero relative abundances estimated to be zero increases in number of samples treated as known regardless of estimator, in general we do not expect this proportion to approach 1 as number of known samples increases. We also note that our model is not designed to produce prediction intervals, and confidence intervals for a parameter estimated from a single observation are unlikely to have reasonable coverage. Finally, we acknowledge that despite the excellent performance of our model on the data of Brooks et al. [2015], our model does not fully capture sample cross-contamination known as index-hopping [Hornung et al., 2019], which likely affects this data.

3.6.2 Type 1 error rate and power

To investigate the Type 1 error rate and power of tests based on reweighted and unweighted estimators, we simulate data arising from a set of hypothetical dilution series. In each simulated dataset, we observe reads from dilution series of four specimens: two specimens of known composition and two specimens of unknown composition (specimens A and B). Each

dilution series consists of four samples: an undiluted sample from a specimen as well as a 9-, 81-, and 729-fold dilution of the specimen. We vary the number of taxa $J \in \{5, 20\}$, as well as the magnitude of elements of $\boldsymbol{\beta}$, the number of samples, and the distribution of $W_{ij}|\mu_{ij}$.

We consider three different values of $\boldsymbol{\beta} \in \mathbb{R}^J$: $\boldsymbol{\beta} = 0\boldsymbol{\beta}^\star$, $\boldsymbol{\beta} = \frac{1}{10}\boldsymbol{\beta}^\star$, and $\boldsymbol{\beta} = \boldsymbol{\beta}^\star$ where $\boldsymbol{\beta}^\star = \begin{pmatrix} 3 & -1 & 1 & -3 & 0 \end{pmatrix}$ when $J = 5$ and $\boldsymbol{\beta}^\star = \begin{pmatrix} 3 & -1 & 1 & -3 & 3 & -1 & \dots & -3 & 0 \end{pmatrix}$ when $J = 20$. We base the magnitude of entries of $\boldsymbol{\beta}^\star$ using the observed magnitude of entries of $\hat{\boldsymbol{\beta}}$ in our analysis of Costea et al. [2017] data. We vary the number of samples between either a single dilution series from each specimen or three dilution series from each specimen (for a total of nine samples per specimen). We draw $W_{ij}|\mu_{ij}$ from either a Poisson(μ_{ij}) distribution or a Negative Binomial distribution with mean parameter μ_{ij} and size parameter $s = 13$. $s = 13$ was chosen to approximate the Karstens et al. [2019] data via a linear regression of fitted mean-centered squared residuals. In all settings we simulate $\{\gamma_i\}_{i=1}^n$ from a log-normal distribution with parameters $\mu = \min(13.5 - 1.5d_i, 12)$ and $\sigma^2 = 0.05$. These values were chosen based on observed trends in reads from target taxa in the data of Karstens et al. [2019] data. In all settings, the first specimen has true relative abundance proportional to $(1, 2^{\frac{4}{J-1}}, 2^{2\cdot\frac{4}{J-1}}, \dots, 2^4)$, that is, taxon J is 16 times more abundant than taxon 1. The second specimen has true relative abundance proportional to $(2^4, \dots, 2^{\frac{4}{J-1}}, 1)$. When $J = 5$, the first two taxa are absent from specimen A, and when $J = 20$, first eight taxa are absent from specimen A. Relative abundances in the remaining taxa form an increasing power series such that the first taxon present in nonzero abundance has relative abundance that is 1/100-th of the relative abundance of taxon J. The relative abundance profile of specimen B is given by the relative abundance vector for specimen A in reverse order. We also simulate the degree of contamination as scaling with dilution. When comparing samples with the same read depth, on average a 9-fold diluted sample will contain 9 times more contaminant reads than an undiluted sample (see Section 5.2 and Figure 3.2, top right). We simulate contamination from a source containing equal relative abundance of all taxa. We set $\tilde{\mathbf{Z}} \in \mathbb{R}^n$ such that $\tilde{Z}_i = 9^{d_i}$ where d_i is the degree of dilutions of sample i, and $\tilde{\gamma} = -3.7$, as we observe in Karstens et al. [2019] data.

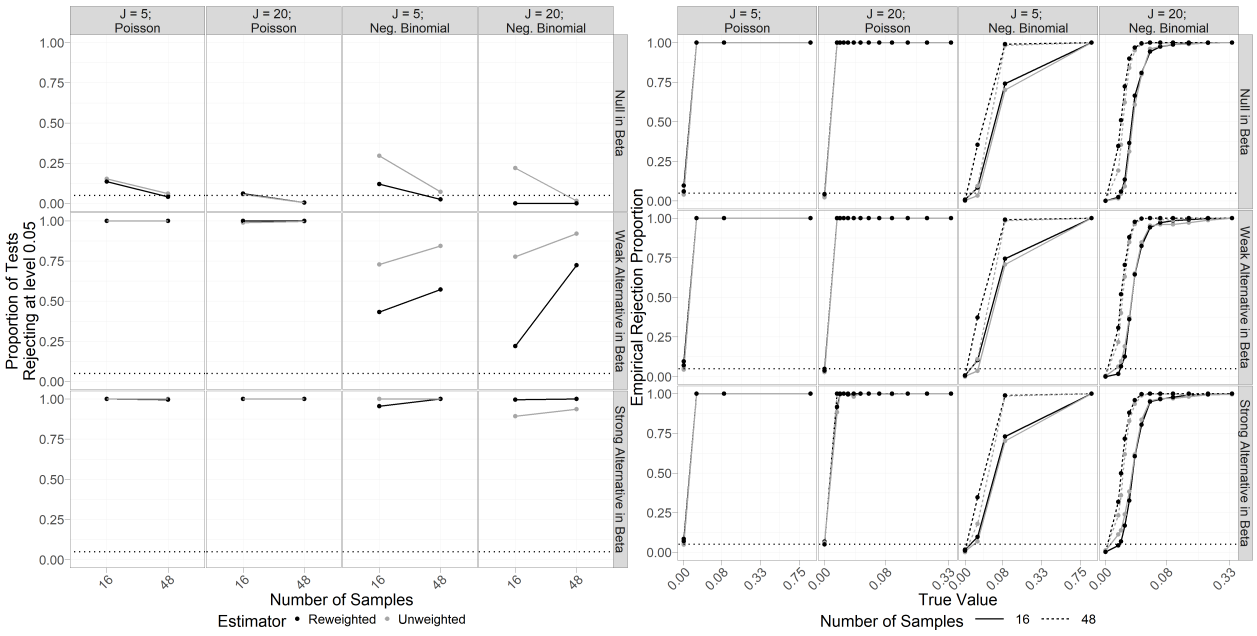

Figure 3.3: At left, the Type 1 error (top row) and power of our proposed likelihood ratio tests for both the unweighted and reweighted estimators. Performance of tests of $H_0 : \boldsymbol{\beta} = 0$ against a general alternative are summarized in terms of empirical rejection rates at level $\alpha = 0.05$ (y-axis), with sample size plotted on the x-axis. Columns give the conditional distribution of data (Poisson or Negative Binomial) and number of taxa J. Rows specify whether data was simulated under the null $\boldsymbol{\beta} = 0$, under a "weak" alternative with $\boldsymbol{\beta} = \frac{1}{10}\boldsymbol{\beta}^\star$ (i.e., $\boldsymbol{\beta} \neq 0$ of small magnitude), or under a "strong" alternative $\boldsymbol{\beta} = \boldsymbol{\beta}^\star$ (i.e., $\boldsymbol{\beta} \neq 0$ of larger magnitude). At right, empirical rejection rate for marginal bootstrap tests of $H_0 : p_{kj} = 0$ at the 0.05 level versus true value of p_{kj} (x-axis). Columns and rows are as specified above.

Figure 3.3 summarizes our results from 250 simulations under each condition. Empirical performance of bootstrapped likelihood ratio tests of $H_0 : \boldsymbol{\beta} = 0$ against $H_A : \boldsymbol{\beta} \neq 0$ at level 0.05 (lefthand pane of Figure 3.3) reveals good performance, even for small sample sizes. Unsurprisingly, we observe improved Type 1 error control and power at larger sample sizes. Tests based on the reweighted estimator generally improved Type 1 error compared to the unweighted estimator, with the greatest improvements observed for data simulated from a negative binomial distribution (mean Type 1 error across simulations was 0.15 for the unweighted estimator and 0.04 for the reweighted estimator). Surprisingly, Type 1 error control appears to improve when the number of taxa J is larger (mean Type 1 error was 0.11 for $J = 5$ and 0.05 for $J = 10$). This may in part be a result of simulating W_{ij} as conditionally independent given γ_i, covariates, and other parameters.[3] Power to reject the null hypothesis is very high when the data generating process is Poisson, as well as for strong alternatives ($\boldsymbol{\beta} = \boldsymbol{\beta}^\star$) when the data follows a negative binomial distribution.

Empirical Type 1 error control for bootstrapped marginal tests of $H_0 : p_{kj} = 0$ against $H_A : p_{kj} > 0$ at the 0.05 level (lefthand pane of Figure 3.3) is generally no larger than nominal, with median empirical Type 1 error 0.02 across all conditions. We observe above-nominal Type 1 error in limited cases, primarily when sample size is small and data is Poisson-distributed (the largest observed Type 1 error rate, of 0.10, occurs for tests based on reweighted estimators at sample size $n = 16$ and number of taxa $J = 5$ when data is Poisson-distributed). Power of marginal bootstrap tests to reject the null is close to 1 for all non-zero values of p_{kj} we consider when data is conditionally Poisson-distributed (median empirical power across conditions 1; minimum 0.88). When data is simulated as negative binomial, unsurprisingly, power appears to increase in magnitude of p_{kj} for tests based on unweighted and reweighted estimators, with somewhat higher power in tests using reweighted estimators. Magnitude of $\boldsymbol{\beta}$ (rows of Figure 3.3) does not appear to affect Type 1 error or power of our tests of $H_0 : p_{kj} = 0$. We also note that empirical coverage of

[3]Under these settings, increasing J provides us more information to, for instance, estimate a mean-variance relationship that might a stabler reweighted estimator.

marginal bootstrapped confidence intervals for p_{kj} is generally lower than nominal in our simulations (see Figure B.2). This aligns with general performance of bootstrap percentile confidence intervals at small sample sizes, and we expect that generating confidence intervals from inverted bootstrapped likelihood ratio tests would yield better coverage in this case (unfortunately this is not feasible for computational reasons).

3.7 Discussion

In this chapter we introduce a statistical method to model measurement error due to contamination and differential detection of taxa in microbiome experiments. Our method builds on previous work in several ways. By directly modeling the output of microbiome experiments, we do not rely on data transformations that discard information regarding measurement precision, such as ratio- or proportion-based transformations. This affords our method the key advantage of estimating relative abundances lying on the boundary of the simplex, which is typically precluded by transformation-based approaches. Accordingly, we implement inference tools appropriate to the non-standard parameter space that we consider. The advantage of estimating relative abundances on the boundary of the simplex is not purely theoretical, and we show that our interval estimates do indeed include boundary values, and demonstrate above-nominal empirical coverage in an analysis of data from Karstens et al. [2019]. Furthermore, our reweighting estimator allows for flexible mean-variance relationships without the need to specify a parametric model. Our approach to parameter estimation does not assume that observations are counts, and therefore our method can be applied to a wide array of microbiome data types, including proportions, coverages and cell concentrations as well as counts. Finally, our method can accommodate complex experimental designs, including analysis of mixtures of samples, technical replicates, dilution series, detection effects that vary by experimental protocol or specimen type, and contamination impacting multiple samples. Contamination is commonly addressed via "pre-processing" data, thereby conditioning on the decontamination step. In contrast, by simultaneously estimating contamination along with all other model parameters, holistic uncertainty in estimation is captured by our ap-

proach.

Another advantage of our methodology is that we do not require the true composition of any specimens to be known. For example, our test of equal detection effects across protocols in Costea et al. [2017] can be performed without knowledge of specimen composition. Accordingly, our approach provides a framework for comparing experimental protocols in the absence of synthetic communities, which can be challenging to construct.

In addition, we expect that our method may have substantial utility applied to dilution series experiments, as illustrated in our analysis of data from Karstens et al. [2019]. Dilution series are relatively low-cost and scalable (especially in comparison to synthetic communities) and may be especially advantageous when the impact of sample contamination on relative abundance estimates is of particular concern. With this said, we strongly recommend against testing the composite null $H_0 : p_{kj} > 0$ against the point alternative $H_A : p_{kj} = 0$, as these hypotheses are statistically indistinguishable on the basis of any finite sample of reads. However, we are able to determine the degree to which our observations are consistent with the *absence* of any given taxon, and therefore we can meaningfully test $H_0 : p_{kj} = 0$ against a general alternative.

We provide a number of suggestions for future research. Firstly, we have not formally identified conditions that ensure the identifiability of parameters of our model. In some situations it is trivial to diagnose non-identifiability through examination of design matrices $\mathbf{Z}$, $\mathbf{X}$ and $\tilde{\mathbf{Z}}$ alongside known entries of $\mathbf{p}$ and $\boldsymbol{\beta}$. In complex situations, this is not straightforward. We defer investigation of these conditions to future work. For now, we note that the parameter estimation algorithm we propose is stable for identifiable means and challenges in model-fitting may betray non-identifiability.

The focus of our chapter was on $\mathbf{p}$ and $\boldsymbol{\beta}$ as targets of inference. Future research could investigate extensions to our model that connect relative abundances $\mathbf{p}$ to covariates of interest, allowing the comparison of average relative abundances across groups defined by covariates, for example. Our proposed bootstrap procedures may also aid the propagation of uncertainty to group-level comparisons of relative abundances in downstream analyses.

In addition, while we focus applications of our model on microbiome data, our model could be applied to a broad variety of data structures obtained from high-throughput sequencing, such as single-cell RNAseq. We leave these applications to future work.

Software implementing the methodology described in this chapter are implemented in a `R` package available at `https://github.com/statdivlab/tinyvamp`. Code to reproduce all simulations and data analyses are available at `https://github.com/statdivlab/tinyvamp_supplementary`.

Chapter 4

LOG-LINEAR MODELS FOR PARTIALLY OBSERVED OUTCOMES

4.1 Introduction

In the previous chapter, we developed a method to address two important forms of measurement error in high-throughput microbiome data: systematic over- and under-detection of microbial taxa, and contamination. Our results suggest that the latter concern can be greatly ameliorated through the use of serial dilution designs in concert with our method. However, in order to estimate and account for detection effects, we require that artificial communities of known composition be sequenced along samples of interest. Such communities are technically demanding to construct, exhibit much lower diversity than samples from many environments of interest, and typically contain only culturable organisms [Cichocki et al., 2020]. One appealing approach to dealing with this limitation is to predict detection effects in microbial taxa not present in artificial communities of known composition using estimated detection effects for closely related microbes present in these communities. After detection effects have been predicted in such a way, analysis may proceed using relative abundances inferred on the basis of predicted effects. The effectiveness of this approach will depend largely on the stability of detection effects over phylogeny; if closely related taxa, under some measurement protocol, are subject to similar detection effects, predictions will likely perform fairly well, and conversely if not. Further research into the relationship between detection effects and phylogeny would help to elucidate the feasibility of this modeling strategy.

This chapter, however, presents a different approach to handling detection effects. We avoid estimation of detection effects altogether and instead target an estimand defined in

terms of means of true sample quantities.

Most generally, this chapter introduces a method for inference on ratios of means of a nonnegative multivariate outcome observed only up to unknown sample-specific proportionality terms. While to our knowledge a similar method does not currently exist, a similar niche in microbiome science is occupied by a range of regression models that rely on log-ratio tansformations to target a similar estimand. These include ANCOM-II, a method which attempts to distinguish sampling, structural, and "outlier" zero values in NGS data prior to transformation and modeling [Kaul et al., 2017]; and ALDEx2, which employs a Bayesian approach to impute nonzero values (under an assumption that all observed zero counts are due to sampling variability) before transformation [Fernandes et al., 2014]. These fall into the larger category of "differential abundance" methods[1], which include LEfSe, DEseq2, edgeR, metagenomeSeq, and corncob [Segata et al., 2011, Love et al., 2014, Robinson et al., 2010, Paulson et al., 2013, Martin et al., 2020]. As these methods target estimands less similar to ours, we will not discuss them further in this chapter.

4.2 Model

We motivate our model in terms of microbiome experiments in which the outcome of primary interest is microbial cell concentration (or a similar quantity) $\overset{\circ}{Y}_{ij}$ in samples $i = 1, \ldots, n$ and taxa $j = 1, \ldots, J$. When NGS measurement techniques are used, we are typically unable to observe $\overset{\circ}{Y} \in \mathbb{R}^{n \times J}$ and instead have access only to sequencing output $Y \in \mathbb{R}^{n \times J}$. Observing Y rather than $\overset{\circ}{Y}$ represents a loss of information in two major ways we will focus on in this chapter:

- the composition of sequencing output Y_i is systematically distorted relative to microbial cell concentrations $\overset{\circ}{Y}_i$ due to (typically unknown) detection effects

- the magnitude of sequencing output Y_i is uninformative with respect to the magnitude

[1] A category subject to some controversy in part on account of the fact that no unambiguous definitions of "differential" or "abundance" are widely agreed upon.

of cell concentrations $\overset{\circ}{Y}_i$

Because our target of inference in this chapter is fold-differences in mean cell concentrations $\mathbb{E}\overset{\circ}{Y}_j$ across levels of some covariate of interest X (e.g., $\mathbb{E}\overset{\circ}{Y}_{j_{\text{cases}}} / \mathbb{E}\overset{\circ}{Y}_{j_{\text{controls}}}$, detection effects are of relatively little concern (as they appear in both the numerator and denominator of fold-differences in the mean). Loss of information about the magnitude of $\overset{\circ}{Y}$ is a more substantial (but not insurmountable) problem, as we will demonstrate below.

Before we more formally lay out our model, we first develop some intuition regarding what kind of quantities related to fold-differences in means of cell concentrations $\mathbb{E}\overset{\circ}{Y}$ we can estimate from sequencing output Y. Consider a simple setting in which we wish to characterize fold differences in mean cell concentrations of three microbial taxa (taxa A, B, and C) across two groups, cases and controls. For simplicity, we ignore sampling variability and detection effects and imagine that we observe mean cell concentration in each group up to an unknown proportionality constant. We will refer to this quantity as a relative abundance. In the simplest case, we observe identical relative abundance among cases and controls. This implies that fold differences in mean cell concentration comparing cases to controls must be equal across taxa A, B, and C, as shown in the top righthand pane of figure 4.1. When relative abundance is not equal in cases and controls, as we see in the bottom left pane of figure 4.1, fold differences in mean cell concentration cannot be equal, but we nonetheless are able to identify patterns of (log) fold differences (cases to controls, for each taxon), shown in the bottom right pane, consistent with observed relative abundances.

In both cases shown in figure 4.1, the groups of log fold differences that are able to explain observed relative abundances differ from each other only by a location shift. This turns out to hold in general. We are able to estimate how much larger a difference in log mean concentrations in one taxon is than in another, but we cannot estimate the differences in log means in isolation. In the context of many microbiome sequencing experiments, this represents a fairly small loss of information if we seek to identify microbial taxa that markedly depart from the norm in terms of how their abundance varies across groups of interest.

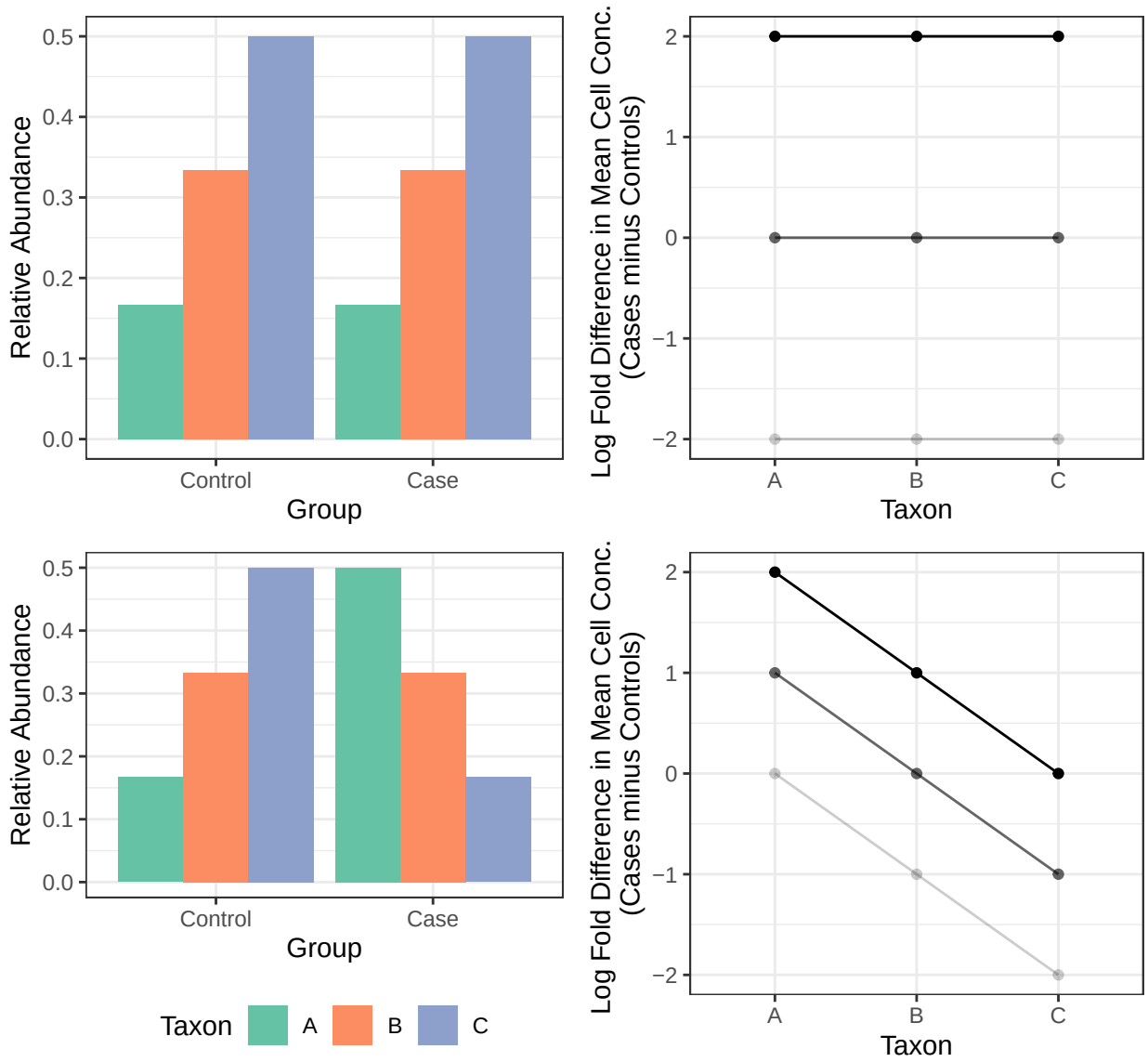

Figure 4.1: Identical relative abundances among a case group and a control group (upper left) are consistent with log mean fold differences in mean cell concentration, though unknown, being equal across taxa (upper right – three possible combinations of log fold differences shown). Differing relative abundances among cases versus controls (lower left) imply differing patterns of log mean fold differences across taxa, explainable by, for example, any of the three groups of fold differences shown in the lower righthand plot.

We now return to the general case and consider an ideal setting in which we observe pairs $(X_i, \overset{\circ}{Y}_i)$ with $X_i \in \mathbb{R}^p$ and $\overset{\circ}{Y}_i \in \mathbb{R}^J_{\geq 0}$ such that

$$\log\mathbb{E}[\overset{\circ}{Y}_i|X_i] = X_i\beta \tag{4.1}$$

Our target of inference here and throughout this chapter is $\beta \in \mathbb{R}^{p\times J}$. We note that this β does *not* correspond to detection effects (also denoted β in the previous chapter); we exclusively use δ to refer to detection effects in this chapter and reserve β to denote the quantity defined above. In applications we consider in this chapter, in place of J-dimensional outcome $\overset{\circ}{Y}_i$, we instead observe a perturbed outcome $Y_i \in \mathbb{R}^J_{\geq 0}$. Specifically, Y_i is perturbed in the following way:

$$\log\mathbb{E}[Y_i|X_i, \beta, z_i, \delta] = z_i\mathbf{1}_J^T + X_i\beta + \delta^T \tag{4.2}$$

where sample-specific scalings $z_i \in \mathbb{R}$ and detection effects $\delta \in \mathbb{R}^J$ are unknown. Hence, roughly speaking, observing Y_i is equivalent to observing $\overset{\circ}{Y}_i$ subject to scaling by factor $\exp(z_i)$ and multiplicative distortion by vector $\exp(\delta^T)$.

When detection effects δ are present, they are absorbed into the intercept row of β, provided it exists. More precisely, for a sample of size n, $\{Y_i, X_i\}_{i=1}^n$, if $\mathbf{X}_n = [X_1^T, \ldots, X_n^T]^T$ contains an intercept column, letting $[\mathbf{1}_n\ \mathbf{W}_n] = \mathbf{X}_n$, $\mathbf{Y}_n = [Y_1^T, \ldots, Y_n^T]^T$, and $\mathbf{z}_n = [z_1, \ldots, z_n]^T$ we can rewrite model (4.2) as follows:

$$\begin{aligned}
\log\mathbb{E}[\mathbf{Y}_n|\mathbf{X}_n] &= \mathbf{z}_n\mathbf{1}_J^T + [\mathbf{1}_n\ \mathbf{W}_n]\beta + \mathbf{1}_n\delta^T && (4.3)\\
&= \mathbf{z}_n\mathbf{1}_J^T + \mathbf{1}_n\big(\beta_0 + \delta^T\big) + \mathbf{W}_n\beta^{(-0)} && (4.4)\\
&= \mathbf{z}_n\mathbf{1}_J^T + \mathbf{1}_n\tilde{\beta}_0 + \mathbf{W}_n\beta^{(-0)} && (4.5)\\
&= \mathbf{z}_n\mathbf{1}_J^T + \mathbf{X}_n\tilde{\beta} && (4.6)
\end{aligned}$$

In the above, $\tilde{\beta} = [\tilde{\beta}_0^T, \beta_1^T, \ldots, \beta_{p-1}^T]^T$ and $\tilde{\beta}_0^T = \beta_0 + \delta^T$. $\beta^{(-0)}$ indicates a matrix consisting of the 2nd through p-th rows of β. Hence we can absorb the detection effect δ into intercept row β_0. For this reason, going forward we assume β includes an intercept row and we omit

reference to δ, writing β as a shorthand for $\tilde{\beta}$ with the understanding that interpretation of the intercept row depends on detection effects δ.

We now turn our attention to z_i's. The presence of scaling terms $z_i \mathbf{1}_J^T$ in mean model (4.2) results in a loss of identifiability. Specifically, for any $\alpha \in \mathbb{R}^p$, we have

$$z_i \mathbf{1}_J^T + X_i(\beta + \alpha \mathbf{1}_J^T) = (z_i + X_i\alpha)\mathbf{1}_J^T + X_i\beta \tag{4.7}$$

$$= z_i^\star \mathbf{1}_J^T + X_i\beta \text{ for } z_i^\star := (z_i + X_i\alpha) \tag{4.8}$$

Hence if $\beta^\star = \beta + \alpha \mathbf{1}_J^T$ for some $\alpha \in \mathbb{R}^p$, we can always find a $z_i^\star \in \mathbb{R}$ such that

$$\log\mathbb{E}[Y_i|X_i, \beta, z_i] = \log\mathbb{E}[Y_i|X_i, \beta^\star, z_i^\star], \tag{4.9}$$

and therefore mean function (4.2) is not identifiable in β and z_i. However, mean function (4.2) is *partially identifiable* in the sense that we can define equivalence classes of βs such that for β and $\beta^\star$ in the same equivalence class, there always exist z_i and $z_i^\star$ such that (4.9) holds, and for β and $\beta^\star$ in distinct classes, there never exist such z_i and $z_i^\star$. We now make the form of these classes explicit and show that they have the properties discussed above.

Theorem 1. *Let* $\mathbf{X}_n := \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} \in \mathbb{R}^{n\times p}$ *and* $\mathbf{Y}_n := \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} \in \mathbb{R}^{n\times J}$ *represent n pairs* (X_i, Y_i). *For arbitrary* $\beta^\dagger \in \mathbb{R}^{p\times J}$, *define* $G_{\beta^\dagger}$ *to be the set*

$$\{\beta \in \mathbb{R}^{p\times J} : \beta = \beta^\dagger + \alpha \mathbf{1}_j^T \text{ for some } \alpha \in \mathbb{R}^p\}.$$

If $\mathbf{X}_n$ *has full column rank, for any* $\beta, \beta^\star \in \mathbb{R}^{p\times J}$ *there exist* $\mathbf{z} \in \mathbb{R}^n$ *and* $\mathbf{z}^\star \in \mathbb{R}^n$ *such that under mean model (4.2)* $log\mathbb{E}[\mathbf{Y}_n|\mathbf{X}_n, \beta, \mathbf{z}] = log\mathbb{E}[\mathbf{Y}_n|\mathbf{X}_n, \beta^\star, \mathbf{z}^\star]$ *if and only if* $\beta^\star \in G_\beta$.

Proof. Suppose $\beta^\star \in G_\beta$. Then there exists an $\alpha \in \mathbb{R}^p$ such that $\beta^\star = \beta + \alpha \mathbf{1}_J^T$ and by (4.7), for any $\mathbf{z}$ and $\mathbf{z}^\star := \mathbf{z} + \mathbf{X}_n\alpha$, $\log\mathbb{E}[\mathbf{Y}_n|\mathbf{X}_n, \beta, \mathbf{z}] = \log\mathbb{E}[\mathbf{Y}_n|\mathbf{X}_n, \beta^\star, \mathbf{z}^\star]$. Now suppose

$\beta^\star \notin G_\beta$. If there exist $\mathbf{z}, \mathbf{z}^\star$ such that $\log\mathbb{E}[\mathbf{Y}_n|\mathbf{X}_n, \beta, \mathbf{z}] = \log\mathbb{E}[\mathbf{Y}_n|\mathbf{X}_n, \beta^\star, \mathbf{z}^\star]$, we have

$$\mathbf{z}\mathbf{1}_J^T + \mathbf{X}_n\beta = \mathbf{z}^\star\mathbf{1}_J^T + \mathbf{X}_n\beta^\star \tag{4.10}$$

$$\Rightarrow \ \mathbf{X}_n(\beta - \beta^\star) = (\mathbf{z}^\star - \mathbf{z})\mathbf{1}_J^T \tag{4.11}$$

$$\Rightarrow \ \mathbf{X}_n^T\mathbf{X}_n(\beta - \beta^\star) = \mathbf{X}_n^T(\mathbf{z}^\star - \mathbf{z})\mathbf{1}_J^T \tag{4.12}$$

$$\Rightarrow \ (\beta - \beta^\star) = (\mathbf{X}_n^T\mathbf{X}_n)^{-1}\mathbf{X}_n^T(\mathbf{z}^\star - \mathbf{z})\mathbf{1}_J^T \tag{4.13}$$

$$\Rightarrow \ \beta^\star = \beta + \tilde{\alpha}\mathbf{1}_J^T \text{ for } \tilde{\alpha} := (\mathbf{X}_n^T\mathbf{X}_n)^{-1}\mathbf{X}_n^T(\mathbf{z}^\star - \mathbf{z}) \in \mathbb{R}^p \tag{4.14}$$

$$\Rightarrow \ \beta^\star \in G_\beta \tag{4.15}$$

which is a contradiction. Hence $\beta^\star \notin G_\beta$ guarantees that for no $\mathbf{z}, \mathbf{z}^\star$ do we have $\log\mathbb{E}[\mathbf{Y}_n|\mathbf{X}_n, \beta, \mathbf{z}] = \log\mathbb{E}[\mathbf{Y}_n|\mathbf{X}_n, \beta^\star, \mathbf{z}^\star]$. □

Theorem 1 demonstrates that while we cannot uniquely estimate β from data generated according to (4.2), we can estimate it up to the addition of constant terms to each of its rows. We will argue that in many settings, this represents a fairly small loss of information, and we are still able to answer meaningful scientific questions by estimating what equivalence class β lies in.

We address this estimation task by imposing identifiability constraints on rows of β such that each G_β contains a single unique element satisfying these constraints. While there are many choices of constraint that satisfy the foregoing condition, we impose constraint $\text{median}_j(\beta_{kj}) = 0$ for $k = 1, \ldots, p$ for this purpose. Under this constraint, we interpret β_{kj} as the difference between the following quantities (first minus second):

- the difference in the log mean of $\overset{\circ}{Y}_j$ comparing two groups differing in covariate X_k by one unit (log mean at smaller value of X_k minus log mean at smaller value; other covariates held constant)

- the median such difference over $j = 1, \ldots, J$

4.3 Estimation

We estimate β (our target of inference) and z (a nuisance parameter) via maximum likelihood or maximum penalized likelihood. In both cases, we consider both unweighted estimation as well as estimation under a (possibly penalized) likelihood weighted according to an estimated mean-variance relationship.

We first address estimation via maximum likelihood (ML). Under the parametrization of the mean model given above, we use the following Poisson likelihood, weighted according to weighting function v, to define ML estimators. (We take $v = 1$ unless otherwise stated; other weightings are discussed below.) We regard these as solutions to score equations induced by the Poisson likelihood and do not assume that Y is Poisson-distributed conditional on our mean model.

$$l_n^v(\beta, z) = \sum_{j=1}^{J} \Big[\sum_{i=1}^{n} v_{ij} \Big(Y_{ij}(X_i\beta^j + z_i) - \exp(X_i\beta^j + z_i) \Big) \Big] \tag{4.16}$$

$$:= \sum_{j=1}^{J} l_{nj}^v(\beta^j, z) \tag{4.17}$$

Because this likelihood decomposes into a sum of terms $l_{nj}^v(\beta^j, z)$, where β^j is the j-th column of β, we can estimate β and z via coordinate descent.

For a fixed β, updates for z exist in closed form:

$$l_i := \sum_{j=1}^{J} \big[v_{ij} \big(Y_{ij}(X_i\beta^j + z_i) - \exp(X_i\beta^j + z_i) \big) \big] \tag{4.18}$$

$$\frac{\partial}{\partial z_i} l_i = \sum_{j=1}^{J} \big[v_{ij} \big(Y_{ij} - \exp(X_i\beta^j + z_i) \big) \big] = 0 \tag{4.19}$$

$$\Rightarrow \hat{z}_i = \log \sum_{j=1}^{J} v_{ij} Y_{ij} - \log \sum_{j=1}^{J} v_{ij} \exp(X_i\beta^j) \tag{4.20}$$

(This maximum is unique by convexity of $-l_i$ in z_i.)

These two update steps together form the backbone of the coordinate descent algorithm (1).

Algorithm 1 Coordinate Descent (Maximum Likelihood)

1. Initiate with data (X, Y), starting value $\beta^{(0)}$ for β, identifiability constraint $g(\beta_k) = 0$ for $k = 1, \ldots, p$, weights v_{ij} for $i = 1, \ldots, n$ and $j = 1, \ldots, J$, convergence tolerance $\epsilon > 0$, and iteration limit $t_{\max}$.

2. Letting $j^\star$ indicate the column index of a column of Y with $j^\star$ chosen to maximize $\sum_{i=1}^{n} \mathbf{1}_{[Y_{ij^\star}>0]}$ over j, set $j^\star$-th column of $\beta^{(0)}$ equal to $\mathbf{0}_p$ as a (temporary) identifiability constraint.

3. Compute $z^{(0)}$ using equation (4.20) applied to data $X, Y^{j^\star}$ with weights $v_{ij^\star}$, $i = 1, \ldots, n$ and $\beta^{j^\star}$ equal to $\mathbf{0}_p$. If $Y^{j^\star}$ contains any zero elements, compute $z^{(0)}$ using data $Y_+^{j^\star}$ consisting of $Y^{j^\star}$ with zero elements replaced with 1. Set $f^{(0)} = l_n^v(\beta^{(0)}, z^{(0)})$.

4. For $t = 1, \ldots, t_{\max}$ or until convergence

   (a) Set $\beta^{(t)} \leftarrow \beta^{(t-1)}$ and $z^{(t)} \leftarrow z^{(t-1)}$.

   (b) For $j \in \{1, \ldots, J\}/\{j^\star\}$

      i. Update jth column of $\beta^{(t)}$ via a Poisson regression of outcome Y^j, the j-th column of Y, on X with weights v^j and offset $z^{(t)}$.

      ii. If $t > 1$, update $z^{(t)}$ using equation (4.20) applied to data X, Y with weights v and $\beta = \beta^{(t)}$. Otherwise update $z^{(t)}$using only the first j columns of Y and $\beta^{(t)}$ (include the $j^\star$-th column as well if $j^\star > j$).

   (c) Set $f^{(t)} = l_n^v(\beta^{(t)}, z^{(t)})$. If $f^{(t)} - f^{(t-1)} < \epsilon$ or $t = t_{max}$, exit iteration and proceed to step 5. Otherwise return to step 4a and increment t.

5. Enforce identifiability constraints: $\beta_k^{(t)} \leftarrow \beta_k^{(t)} - g(\beta_k^{(t)})$ for $k = 1, \ldots, p$ and update $z^{(t)}$ via equation (4.20) with $\beta = \beta^{(t)}$. Return $\beta^{(t)}$ and $z^{(t)}$.

When Y is sparse in some or many columns, the maximum likelihood estimate $\hat{\beta}$ will in general include infinite elements. While $\hat{\beta}$ will exist in these situations so long as only a minority of outcome categories are subject to separation (so row-wise medians of $\hat{\beta}$ are finite), we may nonetheless wish to employ an estimation procedure that ensures elements of $\hat{\beta}$ are finite, particularly if we wish to assume finiteness of elements of β. To accomplish this, we fit the above mean model using a Firth-penalized likelihood.

To describe this penalized likelihood, we rearrange the elements of Y to form column matrix $\tilde{Y} = (Y_1, \ldots, Y_n)^T \in \mathbb{R}^{nj \times 1}$ (where $Y_i = [Y_{i1} \ldots Y_{iJ}]$) and similarly we rearrange β to form column matrix $\tilde{\beta} = (\beta_1, \ldots, \beta_p)^T \in \mathbb{R}^{pJ \times 1}$ (similarly with $\beta_k = [\beta_{k1} \ldots \beta_{kJ}]$). We also define parameter $\theta = (\tilde{\beta}^T z^T)^T \in \mathbb{R}^{(pJ+n)\times 1}$ and construct expanded design matrix $\tilde{X}$ with k-th row is equal to $[(e_j \otimes X_i)\ e_i]$ if $k = (i-1)J + j$. (e_i and e_j are, respectively, $1 \times n$ and $1 \times J$ vectors with ith and jth elements equal to 1 and all others equal to 0; $\otimes$ is the Kronecker product.) That is, we construct $\tilde{X}$ such that $\log \mathbb{E}\tilde{Y} = \tilde{X}\theta$ under our model. The Firth-penalized log likelihood is then given by

$$pl_n^v(\theta) = l_n^v(\theta) + \frac{1}{2} \log |I^v(\theta)|_+ \tag{4.21}$$

where we indicate with $|I(\theta)|_+$ the product of nonzero eigenvalues of information matrix $I^v(\theta) = -\mathbb{E}\frac{\partial^2}{\partial\theta\partial\theta^T} l_n^v(\theta)$ (i.e., the pseudodeterminant of this matrix) [Martin et al., 2013], and we define $l_n^v(\theta) := l_n^v(\beta, z)$ if $\theta = (\tilde{\beta}^T z^T)^T$.

For some choices of identifiability constraint (e.g., $\beta_{kJ} = 0$ for $k = 1, \ldots, p$), we can fairly easily reparametrize our model so that the information matrix is full-rank, in which case (4.21) is a standard Firth penalty. The form given in (4.21) generalizes this penalty to the partially-identified case we consider here, in which the information matrix is not full-rank. We note that that for θ, θ' such that $\tilde{X}\theta = \tilde{X}\theta'$, $|I^v(\theta)|_+ = |I^v(\theta')|_+$ since $I^v(\theta)$ depends on θ only through the linear predictor $\tilde{X}\theta$. Hence the penalty term $\frac{1}{2}\log|I^v(\theta)|_+$ is invariant over the class G_β in the sense that for any $\beta, \beta^\dagger \in G_\beta$ and any $z \in \mathbb{R}^n$, we can find a $z^\dagger \in \mathbb{R}^n$ such that the fitted means in (β, z) and $(\beta^\dagger, z^\dagger)$ are equal, and so the penalty term for each of these pairs is also equal. In particular, this implies that the profile penalized log likelihood

for any $\beta, \beta^\dagger \in G_\beta$ is equal.

To fit our model under a Firth-penalized likelihood, we exploit a representation of the maximum penalized likelihood estimate as a solution to Poisson score equations in data $\tilde{Y}^\dagger$, where

$$\tilde{Y}^\dagger = \tilde{Y} + \text{diag}\Big[W^{1/2}\tilde{X}(\tilde{X}^T W \tilde{X})^- \tilde{X}^T W^{1/2}\Big] \tag{4.22}$$

with $W := \text{diag}(v \circ \exp(\tilde{X}\theta))$ for $v = (v_1, \ldots, v_n)$ an nJ-vector of weights for observations 1 through n in categories 1 through J. (We use $\circ$ to indicate element-wise multiplication.) Accordingly, we find maximum penalized likelihood estimates for θ by iteratively approximately solving score equations in $\tilde{Y}^\dagger$ and updating $\tilde{Y}^\dagger$.

Algorithm 2 Coordinate Descent (Maximum Penalized Likelihood)

1. Initiate with data (X, Y), starting value $\beta^{(0)}$ for β, identifiability constraint $g(\beta_k) = 0$ for $k = 1, \ldots, p$, weights v_{ij} for $i = 1, \ldots, n$ and $j = 1, \ldots, J$, convergence tolerance $\epsilon > 0$, and iteration limit $t_{\max}$.

2. Initiate $z^{(0)}$ at value given given by equation (4.20) with $\beta = \beta^{(0)}$ and set $f^{(0)} = l_n^v(\beta^{(0)}, z^{(0)})$

3. For $t = 1, \ldots, t_{\max}$ or until convergence

   (a) Compute $\tilde{Y}^\dagger$ via (4.22) using $\beta^{(t-1)}$ and $z^{(t-1)}$

   (b) Run algorithm (1) with data $\tilde{Y}^\dagger$ and weights v to obtain $\beta^{(t)}$ and $z^{(t)}$

   (c) Set $f^{(t)} = pl_n^v(\beta^{(t)}, z^{(t)})$. If $f^{(t)} - f^{(t-1)} < \epsilon$, exit with convergence.

To improve efficiency of our estimators, we optionally employ a weighting scheme in which we use fitted values and residuals from an initial model fit to derive weights which we then use as input to a second model fit. This scheme also reduces the influence of large entries

of Y, which can yield stabler estimators if Y exhibits substantial excess-Poisson dispersion. Details of this procedure are given in algorithm 3.

Briefly, this weighting scheme can be motivated in terms of estimating a mean-variance relationship under an assumption that the variance of Y_{ij} is non-decreasing in the mean. The parameter $c > 0$ serves to stabilize the isotonic regression when $\hat{\mu}_{ij}$ is small, and the parameter $u > 0$ limits degree to which any observation can be upweighted (after weights are normalized to sum to 1). We find that $c = 1$ and $u = 10^4$ work well in practice. We note that the implicit mean-variance relationship used to generate weights $\hat{v}$ need not be correctly specified in order for inference to be valid so long as the mean model is correctly specified, but that correct specification will in general lead to more efficient estimation. Accordingly, we condition on these weights in bootstrapping steps (and do not recalculate weights for each bootstrap iteration). When the mean model is misspecified, choice of weights may change what quantity is being estimated, so careful consideration is warranted in this case.

4.4 Inference

It is frequently of interest to evaluate evidence for some difference in $\log \mathbb{E}\overset{\circ}{Y}_j$ across levels of a covariate X_k holding other covariates fixed – that is, to test hypothesis $\overset{\circ}{H}_0 : \log \mathbb{E}[\overset{\circ}{Y}_j | X_k = x_k, X_{-k} = x_{-k}] = \log \mathbb{E}[\overset{\circ}{Y}_j | X_k = x_k + 1, X_{-k} = x_{-k}]$, or expressed in terms of equation (4.1), $\overset{\circ}{H}_0 : \beta_{kj} = 0$.

As discussed above, on the basis of data Y we are typically not able to test $\overset{\circ}{H}_0$ directly, as this requires us to distinguish between between elements of a set G_β over which the mean of Y does not vary. However, if we impose a suitable identifiability constraint on β, which we will generically denote by $g(\beta_k) = 0$ for $k = 1, \ldots, p$, we are able to test a related hypothesis: $H_0^{(g)} : \beta_{kj} = g(\beta_k)$, or equivalently

$$H_0 : \log \mathbb{E}[\overset{\circ}{Y}_j | X_k = x_k, X_{-k} = x_{-k}] - \log \mathbb{E}[\overset{\circ}{Y}_j | X_k = x_k + 1, X_{-k} = x_{-k}] \tag{4.23}$$

$$= g(\log \mathbb{E}[\overset{\circ}{Y} | X_k = x_k, X_{-k} = x_{-k}] - \log \mathbb{E}[\overset{\circ}{Y} | X_k = x_k + 1, X_{-k} = x_{-k}]) \tag{4.24}$$

The interpretation of this test depends on our choice of identifiability constraint. For

Algorithm 3 Weighted Estimation (Maximum Likelihood or Maximum Penalized Likelihood)

1. Initiate with

   - Data (X, Y)
   - Starting value $\beta^{(0)}$ for β
   - Identifiability constraint $g(\beta_k) = 0$ for $k = 1, \ldots, p$
   - Weights $v_{ij} = 1$ for $i = 1, \ldots, n$ and $j = 1, \ldots, J$
   - Convergence tolerance $\epsilon > 0$
   - Iteration limit $t_{\max}$
   - Weight stabilization parameter $c > 0$
   - Weight constraint parameter $u > 0$

2. Compute initial $\hat{\theta}$ via algorithm 1 or 2 with inputs specified in step 1

3. Letting $\hat{\mu} = \exp(\tilde{D}\hat{\theta})$ and $\hat{r} := \tilde{Y} - \hat{\mu}$, estimate variance function $V(\mu)$ via isotonic regression of $\hat{r}^2$ on $\hat{\mu}$.

4. Using $\hat{R}$ estimated in previous to compute weights $\hat{v}_{ij} = \hat{\mu}_{ij}/(\hat{V}(\hat{\mu}_{ij}) + c)$.

5. To prevent weights from growing too large and yielding unstable estimators, constrain $\max(\hat{v}_{ij})/\min(\hat{v}_{ij}) \leq u$ for weight constraint parameter u. If $\max(\hat{v}_{ij})/\min(\hat{v}_{ij}) > u$, impose this constraint by rescaling weights as follows: $v_{ij}^{\text{rescaled}} = \min(v_{ij})(v_{ij}/\min(v_{ij}))^a$ with a chosen so that $(\max(\hat{v}_{ij})/\min(\hat{v}_{ij}))^a = u$.

6. Compute $\hat{\theta}^{\hat{v}}$ via algorithm 1 or 2 using inputs specified in step 1 setting weights equal to $\hat{v}$. Return $\hat{\theta}^{\hat{v}}$

instance, if we use constraint $\text{median}_j(\beta_{kj}) = 0$ for a row of β corresponding to some covariate of interest X_k, then we test a null that population average difference in $\log \mathbb{E}\overset{\circ}{Y}_j$ per unit difference in X_k (holding all other covariates constant) is equal to the median such difference across j. (Choice of constraints on other rows of β does not impact this interpretation.) On the other hand, if we constrain $\beta_{kj'} = 0$ for some $j' \in \{1, \ldots, J\}$, then we test a null that population average difference in $\log \mathbb{E}\overset{\circ}{Y}_j$ per unit difference in X_k is equal to the population average difference in $\log \mathbb{E}\overset{\circ}{Y}_{j'}$ per unit difference in X_k.

We perform inference via bootstrap. Depending on choice of identifiability constraint g, different bootstrap schemes are asymptotically valid. In particular, if g is a smooth function of β, a standard nonparametric bootstrap is sufficient for inference. However, we allow non-smooth g, $g(\beta_k) = \text{median}(\beta_k)$ being one such choice, in which case we employ a studentized Bayesian subsampled bootstrap. The Bayesian subsampled bootstrap is a smoothed m-out-of-n bootstrap valid in settings where standard delta-method arguments fail [Ishwaran et al., 2009]. To improve small-sample performance of this bootstrap, we studentize according to the following procedure. In an outer bootstrap iteration, we refit our model with weights ξ drawn such that $\frac{1}{m}\xi \sim \text{Dirichlet}(\frac{m}{n}\mathbf{1}_n)$ distribution, where m is chosen as a function of n (the sample size) so that $\lim_{n\to\infty} m = \infty$ and $\lim_{n\to\infty} \frac{m}{n} = 0$. (We by default set $m = n^{1/\sqrt{2}}$.) At outer iteration b, this yields bootstrapped $\tilde{\beta}^{(b)}$ fit under bootstrap weights $\xi^{(b)}$. To compute a standard error for each element of $\tilde{\beta}^{(b)}$ (in order to studentize), we perform an inner bootstrap iteration, with weights drawn from a $\text{Dirichlet}(\frac{m}{n}\xi^{(b)})$ distribution. Letting $t_c = (\hat{\theta}_c - \theta_c)/\text{se}(\hat{\theta}_c)$ (where standard error $\text{se}(\hat{\theta}_c$ is computed using bootstrap samples in the outer interation), and $t_c^\xi = (\tilde{\theta}_c^\xi - \hat{\theta}_c)/\text{se}(\tilde{\theta}_c^\xi)$ the bootstrapped version of this quantity (with standard error computed via inner bootstrap iterations) we have $\sqrt{m}t_c$ and $\sqrt{n}t_c^\xi$ converging to the same limiting distribution, and on this basis we construct confidence intervals and compute p-values for marginal tests.

We also allow a block bootstrap in cases where conditional dependence between observations is expected (e.g., replicate measurements on the same specimen, or longitudinal data).

4.5 Simulations

We conduct a small simulation study to examine the type-I error and power of our proposed estimators.

Since applications in microbiome science are a motivation for our method, we attempt to incorporate characteristics of the biological systems studied data generated in microbiome experiments into our simulation design.

In particular, we consider a setting in which we wish to use observations Y – which we can here think of as a table of read counts across some number of microbial taxa – to test hypotheses concerning means of latent J-dimensional outcomes $\overset{\circ}{Y}$. In terms of a microbiome experiment, we relate $\overset{\circ}{Y}$ to (true) cell concentrations in microbial taxa 1 through J for samples 1 through n. In particular, we will test hypotheses concerning how means in $\overset{\circ}{Y}$ (i.e., mean cell concentrations) differ across two groups of interest, which we will refer to as group 1 and group 2. More formally, we generate data according to

$$\log \mathbb{E}[\overset{\circ}{Y}|X, \beta] = X\beta \tag{4.25}$$

$$\log \mathbb{E}[Y|X, \beta, \delta, \mathbf{z}] = X\beta + \mathbf{z}\mathbf{1}_J^T + \mathbf{1}_n\delta^T \tag{4.26}$$

where $X \in \mathbb{R}^{n\times 2}$ consists of an intercept column and a column whose i-th element is equal to zero if observation i belongs to the second group of interest and 0 otherwise. The second row of $\beta \in \mathbb{R}^{2\times J}$ is our target of inference. Nuisance parameter $\mathbf{z} \in \mathbb{R}^n$ is a vector of unknown sample-specific proportionality constants, and $\delta \in \mathbb{R}^J$ in this example corresponds to detection effects [McLaren et al., 2019] in microbiome measurements whereby some microbial taxa are multiplicatively over- or under-represented in read count data relative to their true abundance. As noted above, we do not separately estimate δ and β but instead absorb δ into the intercept row of β.

We conduct simulations letting number of observations $n \in \{50, 200\}$ and dimension of Y_i $J = 60$, and letting $\beta_1 = a(e_1 - e_{J/4} - e_{3J/4} + e_J)$ for $a = 0, 0.5, 2$, with e_k indicating a vector with k-th element 1 and all others 0. In other words, we simulate under a null in

which $\beta_1 = 0$ as well as under sparse alternatives in which some elements of β_1 have small ($a = 0.5$) or moderate ($a = 2$) magnitude. In all simulations β_0 is constructed to reflect a microbial community composed of taxa present a wide range of abundances: we set the first and last $J/10$ elements of β_0 equal to, respectively, -9 and 2; the remaining $8J/10$ terms consist of an increasing arithmetic sequence from -5 to -2.

Since our method is motivated by applications in NGS data generated by microbiome experiments, we attempt to recreate some of the features of that data in our simulated observations Y. Accordingly, we chose a data generation mechanism that produced count data Y with a high degree of sparsity (about 75% of elements of Y equal to zero on average) and substantial excess-Poisson dispersion. In a similar vein, we attempted to imbue in simulated $\overset{\circ}{Y}$ (which we used to generate Y draws) some common features of the biological systems on which microbiome measurements are taken. Specifically, means of $\overset{\circ}{Y}_j$ (which we can imagine to be mean cell concentrations in a particular microbial taxon, for instance) span several orders of magnitude, and we allow positive and negative correlation between columns of $\overset{\circ}{Y}$, corresponding to to correlated concentrations of microbes in different taxa.

To simulate Y, we first simulate $\overset{\circ}{Y}$ as follows. Letting $X_i = [1\ 1_{[\text{i in group 2}]}]$, for a given choice of $\beta \in \mathbb{R}^{2\times J}$ we draw $\overset{\circ}{Y}_i$ as the product $S_i \circ T_i \circ \exp(X_i\beta)$ where $\circ$ indicates elementwise multiplication. S_i is a J-vector of independent (mean-1) draws from a lognormal distribution with parameters $\mu = -1/20$ and $\sigma^2 = \sqrt{1/10}$. The S_i are also drawn independently across i. T_i is chosen so as to induce correlation between some elements of $\overset{\circ}{Y}$ while preserving mean structure. Specifically, we let

$$T_i(\gamma_{i1}, \ldots, \gamma_{iL}) = \Big[\prod_{l=1}^{L} A_l(\gamma_{il})\Big] \tag{4.27}$$

Here, each γ_{il} is an independent draw from a beta$(1/2, 1/2)$ distribution and $A_l(\gamma_{il})$ is a J-vector each of whose elements is equal to either 1, $2\gamma_{il}$, or $2(1-\gamma_{il})$. We select which elements are not set equal to 1 by drawing m elements of $1, \ldots, J$ uniformly without replacement, where m is itself uniformly drawn from $2, \ldots, J/6$. For all simulations, we set $L = 10$, and for each choice of J, we generated a single set $(A_1, \ldots, A_L)$ that we hold fixed for all simulations

under that J. This data generating process yields $\overset{\circ}{Y}$ with low correlation between most elements (approximately 80% of pairwise correlations between -0.05 and 0.05) but with high correlation between some elements (approximately 5% of pairwise correlations greater than 0.5 in absolute value).

Having drawn $\overset{\circ}{Y}$ we simulate Y by letting Y_{ij} be a draw from a negative binomial distribution with size 1/8 and mean $\overset{\circ}{Y}_{ij} \exp(z_i + \delta_j)$, where z is an n-vector of independent draws from a normal distribution with mean 4 and unit variance, and δ is a J-vector whose elements are draws from a normal distribution with mean 0 and variance 4. For each choice of J we draw δ once and fix it throughout simulations.

We marginally test, for $j = 1, \ldots, J$, $H_0 : \beta_{2j} = 0$ under the constraint that $\text{median}_{j' \in \mathcal{J}} \beta_{2j'} = 0$ for $\mathcal{J}$ the $J/3$ taxa with highest baseline abundance β_1. In other words, we test whether the difference $\log \mathbb{E}[\overset{\circ}{Y}_j | \text{group 2}] - \log \mathbb{E}[\overset{\circ}{Y}_j | \text{group 1}]$ deviates from its median value over the top 1/3 most abundant taxa.[2] In all simulations, we use outer bootstrap iterations with 5 inner iterations, and we conduct 250 simulations for each combination of n and magnitude a of deviation from null hypothesis ($a \in \{0, 0.5, 2\}$).

Figure 4.2 shows coverage under the null ($a = 0$) of weighted and unweighted maximum likelihood and maximum Firth-penalized likelihood estimators. We observe in general lower coverage among less abundant taxa ($\mathbb{E}\overset{\circ}{Y}_j$ increases from left to right on x-axis), at smaller sample sizes, and for maximum likelihood estimation. For Firth estimators, reweighting appears to result in substantially lower coverage in some taxa. This phenomenon does not appear to affect estimation in more abundant taxa, which suggests that it may be due to the impact of reweighting on Firth penalization (weights are used in constructing the information matrix form which the penalty is derived).

We note that coverage is influenced by choice of number of inner bootstrap iterations; Supplemental Figure C.1 shows generally lower coverage for simulations conducted with 10

[2]We take a median only over the most abundant taxa to ensure that in each simulation we are formally testing the same hypothesis; not all taxa are observed in every simulation, so taking a median over all (detected) taxa in each simulation would result in slightly different hypotheses being tested in different simulations.

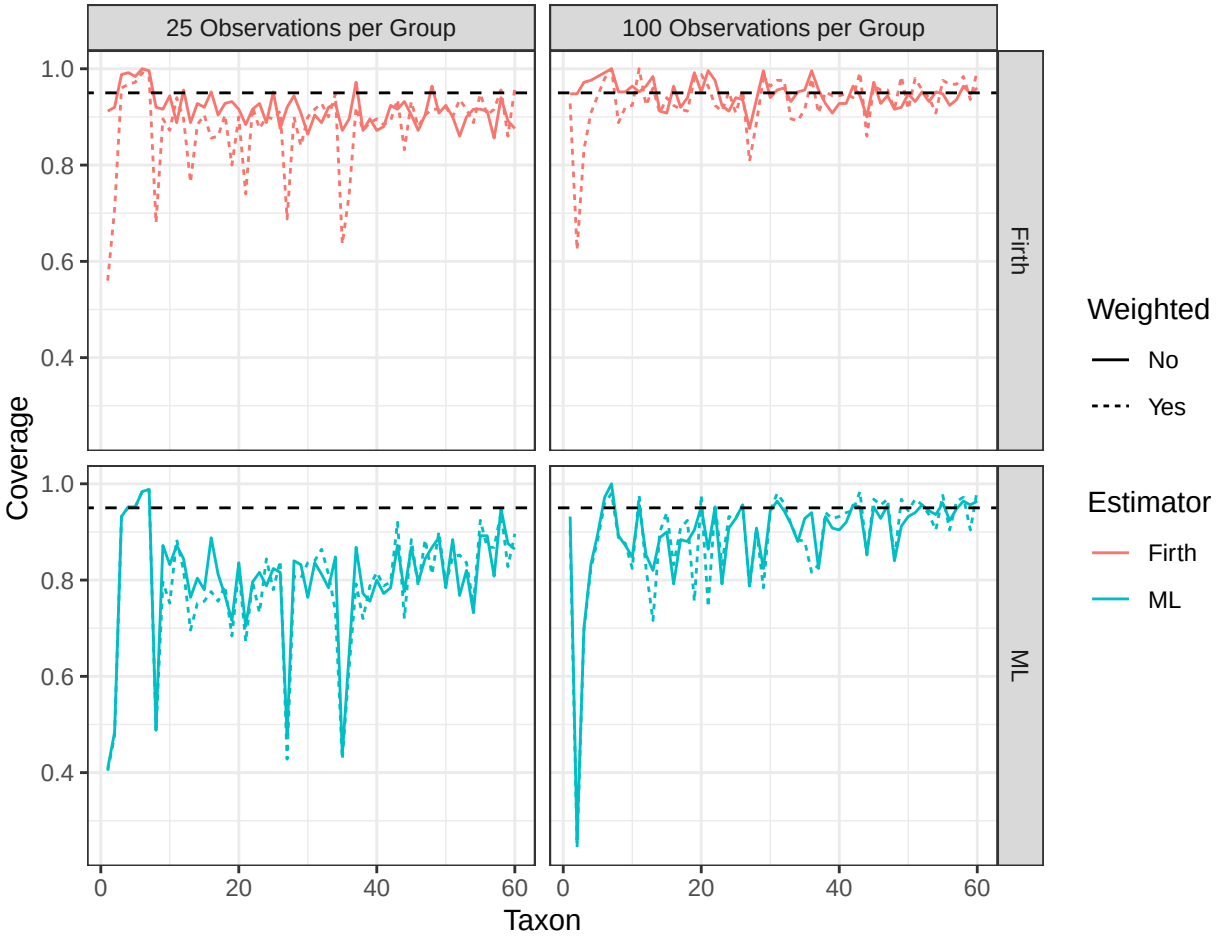

Figure 4.2: Empirical coverage of marginal 95% confidence intervals computed from 250 simulation iterations for elements of second row of β under the null (all elements of 2nd row of β truly equal to zero) by estimator (maximum likelihood in cyan, maximum Firth-penalized likelihood in red), weighting (solid lines for unweighted estimation, dashed lines for weighted estimation), number of observations per group (given in columns), and taxon (x-axis). In all simulations, 500 outer and 5 inner bootstrap iterations are used.

inner iterations than we observe here for simulations with 5 inner iterations, though coverage improves as sample size increases. This is likely due to relatively less stable estimation of standard errors when 5 inner iterations are used, which results in a fatter-tailed bootstrap t distribution.

Figure 4.3 summarizes power under the null and two alternatives by estimator, number of observations, weighting, and taxon. We observe generally higher power in taxa with higher abundance (taxa 45 and 60), as well as generally increasing power as magnitude of effect (x-axis) increases. Somewhat surprisingly, and in a departure from our results in the previous chapter, weighting does not appear to improve either power or type-1 error control in general, perhaps as a result of instability of estimated weights. We note as well that power is higher when 10 rather than 5 inner bootstrap iterations are used (see Supplemental Figure C.2).

4.6 Data analysis

In this section we consider a metagenomic dataset preprocessed and published by Wirbel et al. [2019]. The raw sequencing data from which this dataset was produced was collated from four previous studies targeting possible associations between composition of the human gut microbiome and colorectal cancer (CRC) risk, as well as from a cohort from whom Wirbel et al. [2019] themselves gathered and sequenced samples. In each study, fecal samples were taken from participants with colorectal cancer as well as from healthy controls undergoing colonoscopy for analysis via whole-genome sequencing.

In this analysis, we are interested in identifying microbial strains unusually enriched or depleted in study participants with colorectal cancer as compared to control participants.

4.6.1 Data

Metagenomic data published by Wirbel et al. [2019] that we consider here consists of a mOTU (metagenomic operational taxonomic unit) table containing estimated relative abundances within 849 mOTUs detected across study populations for each study participant, with 575 participants in total. Each row of this table consists of estimated relative abundances of

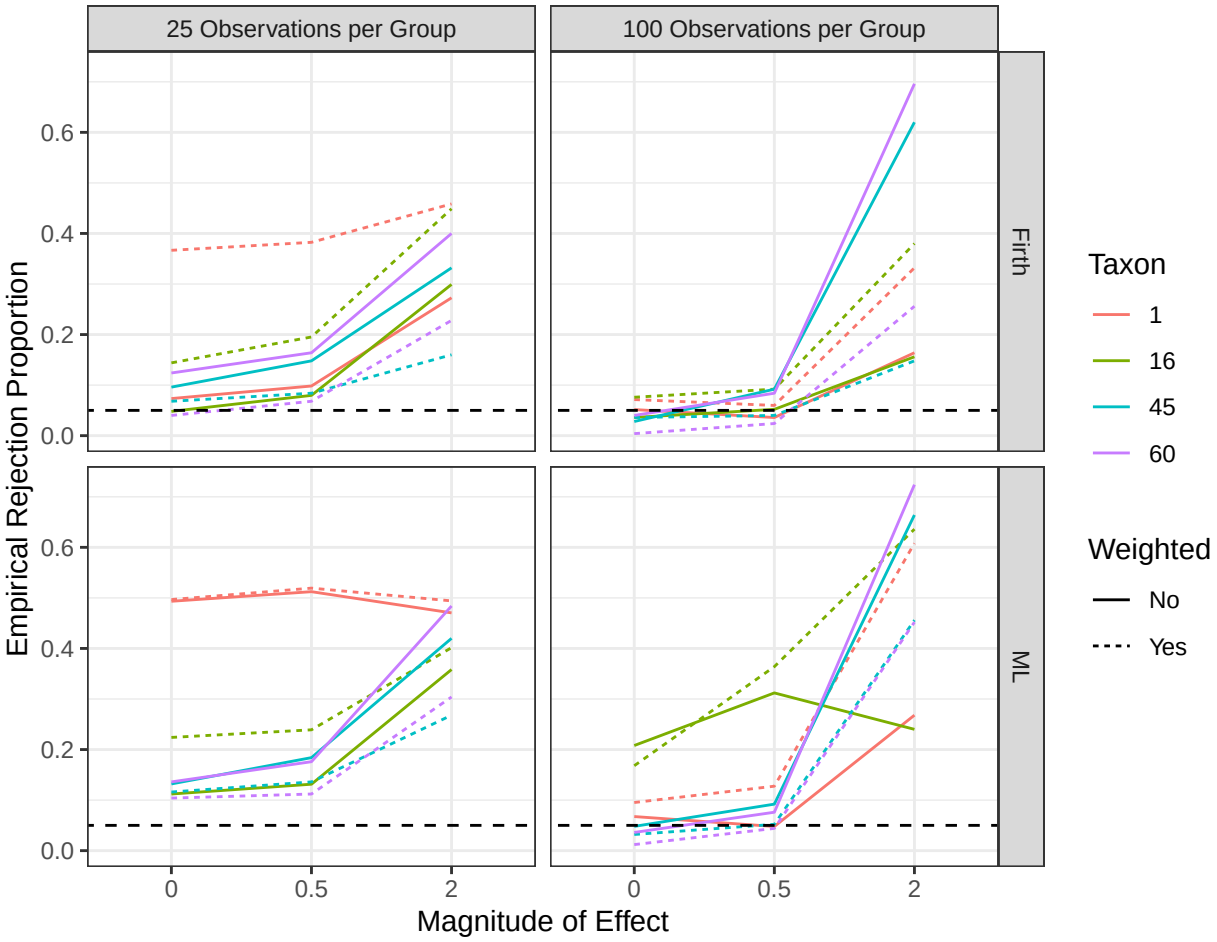

Figure 4.3: Empirical power computed from 250 simulation iterations for potentially nonzero elements of second row of β by magnitude of effect (given on x-axis) by estimator (rows), weighted (solid lines for unweighted estimation, dashed lines for weighted estimation), number of observations per group (given in columns), and taxon (color). In all simulations, 500 outer and 5 inner bootstrap iterations are used.

each of 849 microbial taxa for a particular study participant's gut microbiome. In addition, Wirbel et al. [2019] published clinical and demographic metadata including colorectal cancer status, age in years, binary gender, BMI, and timing of sampling relative to colonoscopy (sample taken prior to or following colonoscopy). In addition, for each participant, library size (i.e., number of reads generated from that participant's sample) was made available.

4.6.2 Modeling

Before fitting a model, we transform estimated relative abundances reported for each participant by multiplying by library size. We note that this does not fully recover the mOTU counts from which relative abundances were obtained, as Wirbel et al. [2019] set all estimated relative abundances lower than 10^{-6} equal to zero. However, rescaling by library size restores at least some information regarding precision that is lost during transformation of mOTU counts to relative abundances. (We note that rescaling is not strictly necessary for model fitting.)

On this transformed data, we fit a Firth-penalized log-linear model (without reweighting) under median-zero constraints on β. We include, in addition to an intercept and a term for CRC status (1 for CRC and 0 otherwise), terms in our model for age (linear spline with a single knot at median age 64), study, and whether fecal samples were taken prior to colonoscopy (0 if prior, 1 otherwise).

We note that study terms here account both for differences in the populations from which each of these studies recruited as well as for differences in measurement protocol across studies. Previous work [McLaren et al., 2019] has indicated that at least in some contexts, the systematic component of measurement error in metagenomic datasets is well-described by log-mean shift δ (see equation 4.2).

4.6.3 Results

Figure 4.4 shows elements of $\hat{\beta}_{\text{group}}$ in four bacterial genera, all of which contain members that previous studies have suggested may be associated with CRC risk [Sun and Kato, 2016].

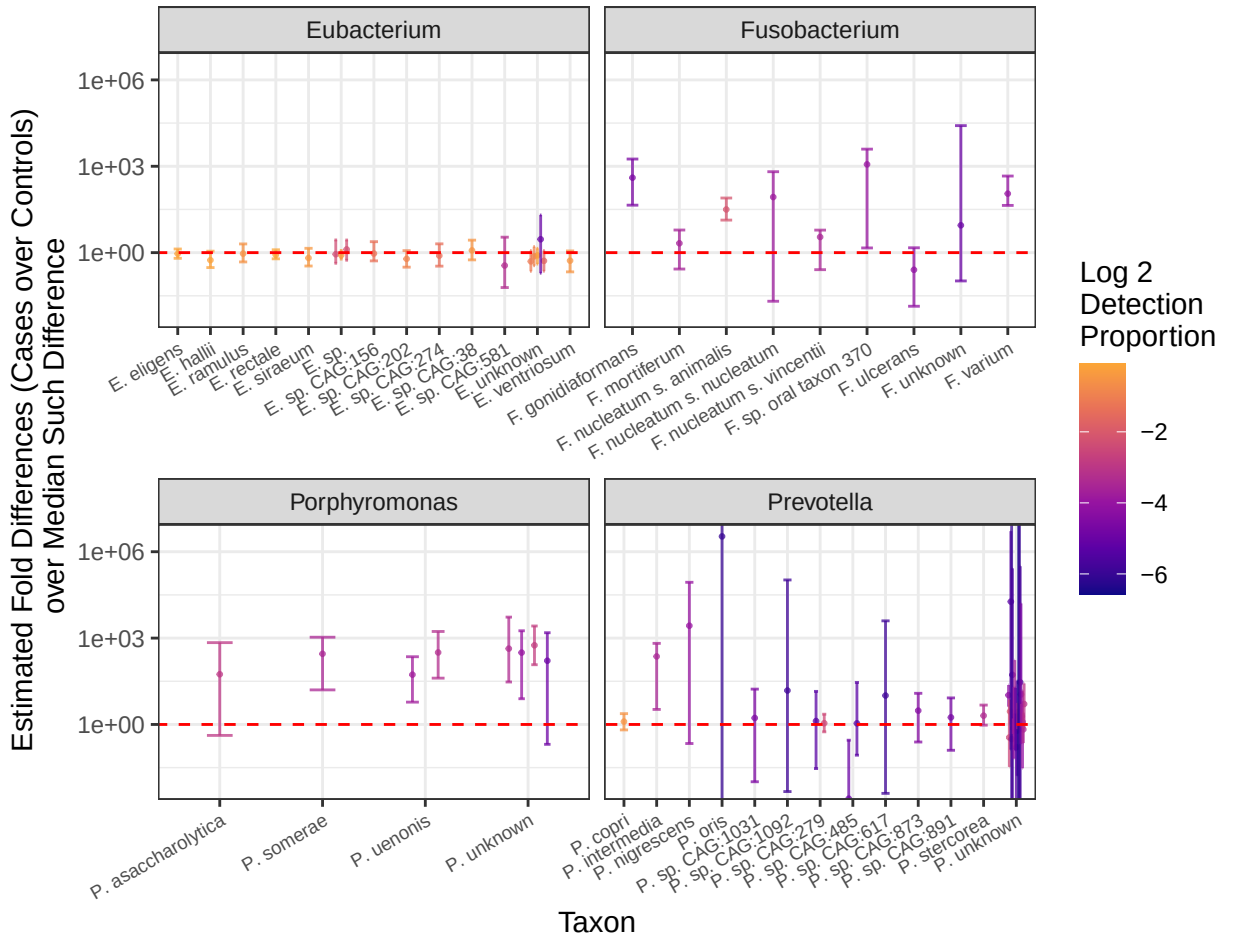

Figure 4.4: Estimates and 95% confidence intervals for fold differences in mean cell concentration among cases and controls (first minus second; model adjusts for age, BMI, study, and timing of sample collection) among mOTUs in 4 bacterial genera. Each pane contains estimates for a single genera, with species assignments given on the x-axis and effect plotted on the y-axis. Color indicates the proportion of study participants in which each mOTU is detected.

Among mOTUs within *Eubacteria*, we observe, on the ratio scale, relatively small, generally negative effects, with 95% confidence intervals including 1, suggesting ratio of mean cell concentration in cases to controls (with both group alike with respect to all other covariates) is approximately the same as (or slightly less than) is typical among the mOTUs detected in the five study cohorts [3] considered here. We find large positive effects among mOTUs in three genera, *Fusobacteria*, *Porphyromonas*, and *Prevotella*, suggesting that the ratio of mean cell concentration in cases to controls among these mOTUs is much larger than is typical. Patterns of association within *Prevotella* depart somewhat from those in *Fusobacteria* and *Porphyromonas* in that we observe mOTUs with large positive and moderately large negative associations in this genus. This appears to be in part due to the fact that many mOTUs in *Prevotella* were detected in a small proportion of study participants, and effect estimates for mOTUs with low detection are generally more variable than for those with high detection. Indeed, for the most commonly detected mOTU in this genus, in species *P. copri*, we observe a modest positive effect with a fairly narrow confidence interval that includes the null.

We note that in bootstrapping on individual observations within each of the five studies from which Wirbel et al. [2019] collected data, we are appealing to asymptotics in which we sample from larger and larger groups recruited from the same populations these studies recruited from. If we wish to generalize beyond these populations, then it is more reasonable to consider the study rather than the participant as the unit of sampling, and a block bootstrap would be appropriate. In other words, we treat study as a fixed effect, but it could also, were data from more studies available, reasonably be treated as a random effect.

4.7 Discussion

In this chapter we describe a regression method for inference on means of a nonnegative mutivariate outcome $\overset{\circ}{Y}$ observed after perturbation by unknown sample-specific scaling terms $\mathbf{z}$ and "detection effects" δ. We draw motivation from microbiome sequencing experiments,

[3] Four cohorts from previous studies and one recruited by Wirbel et al. [2019].

in which outcomes of interest, such as microbial cell concentration across potentially many species, are typically observed indirectly via sequencing output subject to both sample-specific scaling and systematic taxon-specific distortion (i.e., detection effects). However, we anticipate that our method may find application in other settings; in particular, we expect that it may be appropriate for analysis of various types of 'omics measurements, where similar scaling and detection effects are likely at work.

Our method allows inference on differences in $\log \mathbb{E}\overset{\circ}{Y}_j$ across groups defined by covariates X up to a location shift. That is, letting X_k denote a covariate of interest and X_{-k} denote any other covariates modeled, if β_{kj} is the (population weighted average) difference in $\log \mathbb{E}\overset{\circ}{Y}_j$ at $X_k = x_k, X_{-k} = x_{-k}$ versus at $X_k = x_k - 1, X_{-k} = x_{-k}$ (first quantity minus second), then we are able to identify $\beta_k = (\beta_{k1}, \ldots, \beta_{kJ})$ up to a constant shift in all of its elements: our observations Y provide us a basis on which to distinguish β_k from $\beta_k^\star$ if and only if $\beta_k \neq \beta_k^\star + a\mathbf{1}_J^T$ for any $a \in \mathbb{R}$. In other words, we are able to identify differences between elements of β_k but not the location of any particular element β_{kj}. We address this lack of full identifiability via constraint $g(\beta_k) = 0$ which may be chosen based on scientific context. We interpret $\hat{\beta}_{kj}$ estimated under such a constraint as the estimated difference in log means of $\overset{\circ}{Y}_j$ per unit difference in covariate X_k minus the value of the constraint function g. If we specify $g(\beta_k) = \beta_{k1}$, then $\hat{\beta}_{kj}$ estimates how much larger or smaller the difference in log means of $\overset{\circ}{Y}_j$ per unit difference in covariate X_k is than the corresponding difference in log means of $\overset{\circ}{Y}_1$. We suggest $g(\beta_k) = \text{median}_j \beta_{kj}$ as a reasonable default constraint in many settings, in which case $\hat{\beta}_{kj}$ estimates how much larger or smaller the difference in log means of $\overset{\circ}{Y}_j$ per unit difference in covariate X_k is than the median such difference in log means over $j = 1, \ldots, J$.

Our method departs from other methods commonly applied for similar purposes in this context in that we do not require data transformation. We consider this advantageous for several reasons. First, this allows us to use information regarding the precision of observations that is typically lost under, for instance, a log-ratio transformation. Additionally, while our method requires a choice regarding identifiabiity constraint, the estimates it produces are

invariant in the sense that under constraints g and g', our method will produce estimates $\hat{\beta}$ and $\hat{\beta}'$ that are identical up to a location shift in each row. This simplifies comparison of results across studies even if researchers have used differing constraints. When a log-ratio method is employed, such invariance does not apply to choice of denominator for this transformation, and it is not always straightforward to compare results obtained under differing transformations. Most importantly, our method estimates quantities that exist under realistic conditions. We require (population) means in $\overset{\circ}{Y}_j$ to be positive across groups of interest. Log-ratio methods, by contrast, generally require an assumption that $\overset{\circ}{Y}_{ij} > 0$ for all samples i and outcomes j, and add a small "pseudocount" to elements of Y equal to zero prior to transformation.[4] In the motivating context for our method, microbiome sequencing experiments, assuming $\overset{\circ}{Y}_{ij} > 0$ is often quite restrictive: it amounts to assuming every microbe detected in any sample is present in all samples. When this assumption is not satisfied, it is difficult to relate estimates obtained from transformed data unambiguously to any particular functional of true sample abundances $\overset{\circ}{Y}$.

We propose $\text{median}_j \beta_{kj} = 0$ as a reasonable default choice for identifiability constraint. However, as noted in our examination of simulations empirically validating our model, some care is required in interpretation of estimates and tests performed under this constraint, as the median will typically be taken over outcome categories observed in a given experiment, which may not coincide with a median taken over all relevant outcome categories. Accordingly, in some contexts it may make sense either to constrain the median taken over a subset of $j = 1, \ldots, J$ corresponding to microbial taxa (or other outcome) frequently observed in a particular scientific context, or to employ a different constraint altogether. With this said, we do not anticipate variation in which outcome categories are detected contributing to

[4] In fact, considerable effort has been expended in pursuit of "better" pseudocounts under the assumption that $\overset{\circ}{Y}$ is strictly positive. Such approaches yield results that are difficult to interpret when $\overset{\circ}{Y}_{ij} > 0$ is not satisfied. Other methods [Kaul et al., 2017] have attempted to relax this assumption by classifying zeroes according to whether they are thought to be structural, sampling, or "outlier" zeroes. Leaving aside whether structural and sampling zeroes can in general be reliably distinguished, this strikes us as a Gordian knot that can be cut by abandoning $\mathbb{E} \log \overset{\circ}{Y}$ as an estimand in favor of $\log \mathbb{E} \overset{\circ}{Y}$.

incomparability of results across studies, as estimates of β are invariant up to a row-wise location shift regardless of what identifiability constraint is used. As a result, even if two studies report estimates under differing identifiability constraints, these estimates can be made comparable via recentering.

An additional consideration in practical applications for this method is measurement error. Our method is motivated by a context in which systematic measurement error, which we represent in the foregoing sections by δ, is typically present. Estimation of β is minimally impacted by measurement error of this form. However, no such guarantees exist for other kinds of measurement error – for instance, we do not anticipate that in the context of microbiome analysis that estimation of β will be robust to the presence of microbial contamination in sequencing output. Accordingly, when it is possible to detect and remove other forms of measurement error prior to use of this method, we strongly encourage this (e.g., via methods described in Davis et al. [2018] or chapter 3) and further recommend that the impact on precision of such preprocessing be incorporated into analyses (e.g., via resampling techniques).

Our method admits several possible extensions. In contexts where it is scientifically plausible that parameter of interest β_k is sparse, enforcing sparsity via L1 penalization is an attractive option. We note in particular that under the assumption that most the elements of β_k are equal zero, β_k is identified, since such a sparsity assumption implies $\text{median}_j \beta_{kj} = 0$. In addition, while we focused on tests of marginal hypotheses, as we in general expect these to be of greatest scientific interest, it may also be of interest to test hypotheses of the form $H_0 : \beta_k = 0$ – i.e., that $\beta_{k1} = \cdots = \beta_{kJ} = 0$ simultaneously. We anticipate that a robust score or bootstrapped likelihood ratio test could perform well for this purpose. Lastly, while we allow flexible weighting to improve efficiency in the presence of excess-Poisson dispersion, we do not currently allow differing weighting schemes to be applied across outcome categories j. In some contexts, allowing differing schemes may improve efficiency. However, we anticipate that sharing at least some information across outcome categories will remain necessary to avoid unstable weights.

Chapter 5

CONCLUSION

In this dissertation, we have attempted to explicitly relate measured, sample, and population quantities relevant to microbiome experiments. In the absence of reliable models linking measurements to true sample quantities, it is difficult to justify claims regarding population quantities. Accordingly, we advocate for validation of measurement technologies with robust experimentation and modeling, as well as the adoption of publication standards requiring assumptions on measurement error to be reported as part of any microbiome analysis.

Unfortunately, common practice in microbiome research poses challenges to these goals. Terminology that elides distinctions between measured and true sample quantities, for instance, is quite common: it is, for example, typical to refer to the proportion of reads attributed to a particular taxon in a particular sample as "the" relative abundance of that taxon in that sample rather than the measured relative abundance or the read relative abundance. While we expect that most microbiome researchers understand "the relative abundance" and similar language as shorthand not to be interpreted literally, vague terminology nonetheless makes speaking and reasoning about the relationship between measurements and underlying quantities of interest more difficult. This is evidenced by a variety of common practices in microbiome science that involve statistical manipulation of measurements without clear grounding in any underlying biological quantity. To take one example, many published analyses apply Wilcoxon, Kruskal-Wallis, or similar tests to read proportion data to determine which microbial taxa are "differentially abundant" (see, e.g., Wirbel et al. [2019]). However, even leaving aside whether these procedures test scientifically meaningful hypotheses, this practice runs into conceptual trouble: measurement error (in particular, detection effects) alone may induce orderings of read proportions that substantially deviate from the ordering

of true microbial relative abundances, so it is difficult to relate the results of such a test to any particular population quantity.

An additional challenge lies in the fact that discussions of measurement error in the microbiome literature are often framed primarily in terms of technical variation and batch effects. Sinha et al. [2017], whose data we analyze in the first chapter of this dissertation, conduct an "assessment of variation" – not an assessment of error – in microbial sequencing data, and their analysis reflects this focus, in particular in its dismissal of the idea that measurement error may systematically distort comparisons of scientific interest. Similarly, many microbiome papers and methods have been targeted at batch effects [Dai et al., 2019, Wang and LêCao, 2020, Gibbons et al., 2018b]. This is a reasonable phenomenon to investigate, as systematic differences across sequencing batches are readily observable in microbiome data. However, in our view, such observable differences are a symptom of the larger problem that that measurements systematically deviate from the quantities they are meant to represent. In discussing this problem as a matter of batch effects, we risk creating the impression that but for batch-to-batch variability, our measurements would be largely accurate, or that we can make valid between-group comparisons by removing or adjusting for apparent batch effects without regard to the functional form of such effects or their relationship to sample quantities we wish to measure.

With this said, the problems posed by measurement error are challenging but not insurmountable, and researchers can take steps to limit impact on their work. Firstly, motivating analyses in terms of (biological) sample and population quantities of interest – rather than in terms of some manipulation of sequencing data – on its own can substantially improve interpretability of results. No amount of experimentation or computation can make for a lack of conceptual clarity in a microbiome analysis. Additionally, as we discuss in chapter 4, some regression techniques are fairly robust to detection effects. We are, unsurprisingly, partial to our own method presented in chapter 4, but various other approaches, often described as compositional data analysis methods, also share this robustness quality, though we note that other assumptions some of these methods make (e.g., that all taxa detected in

any sample are present in all samples) may not be appropriate for microbiome analyses. As we also discuss in chapter 4, regression methods that are robust to the presence of detection effects may not be robust to contamination. At a minimum, we suggest addressing problems of contamination by including positive (i.e., mock community) and negative (blank) controls in all sequencing runs and clearly reporting the results on these control samples to increase the odds that contamination, if present, will be detected. Beyond this, we anticipate that dilution series designs – in which diluted samples from some or all specimens interest are sequenced alongside undiluted samples – may prove fruitful in detection and removal of contamination via the model we describe in chapter 3. Using this approach in concert with the regression method we describe in chapter 4 may yield analyses with some robustness to both contamination and detection effects, though further research is needed to establish the performance of this combination of methods. Lastly, we advocate that, to the extent possible, results of microbiome experiments be validated with orthogonal measurement techniques (i.e., with technologies not subject to the same forms of error as those used to take initial measurements) and in independent experiments.

To support rigorous interrogation of microbial communities, this dissertation has developed new methods addressing various challenges in microbiome data analysis, from relative abundance estimation in the presence of contamination and detection effects to inference on scientifically meaningful population quantities when only sequencing data is available. The forms of measurement error we have addressed, though important, are by no means exhaustive, and as research practice and measurement technology evolve, sustained effort will be required to ensure that experimental findings are grounded in reality.

BIBLIOGRAPHY

John Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):139–160, 1982.

Donald WK Andrews. Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica*, pages 399–405, 2000.

Mokhtar S Bazaraa. Nonlinear programming: theory and algorithms, 2006.

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):1–14, 2015.

Peter Bühlmann, Torsten Hothorn, et al. Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science*, 22(4):477–505, 2007.

Benjamin J Callahan, Paul J McMurdie, Michael J Rosen, Andrew W Han, Amy Jo A Johnson, and Susan P Holmes. DADA2: high-resolution sample inference from illumina amplicon data. *Nature Methods*, 13(7):581, 2016.

Benjamin J Callahan, Paul J McMurdie, and Susan P Holmes. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME Journal*, 11(12):2639, 2017.

Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, Rory Mitchell, Ignacio Cano, Tianyi Zhou, Mu Li, Junyuan Xie, Min Lin, Yifeng Geng, and Yutian Li. *xgboost: Extreme Gradient Boosting*, 2019. URL `https://CRAN.R-project.org/package=xgboost`. R package version 0.90.0.2.

Nicolas Cichocki, Thomas Hübschmann, Florian Schattenberg, Frederiek-Maarten Kerckhof, Jörg Overmann, and Susann Müller. Bacterial mock communities as standards for reproducible cytometric microbiome analysis. *Nature Protocols*, 15(9):2788–2812, 2020.

Marcio C Costa, Luis G Arroyo, Emma Allen-Vercoe, Henry R Stämpfli, Peter T Kim, Amy Sturgeon, and J Scott Weese. Comparison of the fecal microbiota of healthy horses and horses with colitis by high throughput sequencing of the V3-V5 region of the 16s rRNA gene. *PLOS ONE*, 7(7):e41484, 2012.

Paul I Costea, Georg Zeller, Shinichi Sunagawa, Eric Pelletier, Adriana Alberti, Florence Levenez, Melanie Tramontano, Marja Driessen, Rajna Hercog, Ferris-Elias Jung, and others. Towards standards for human fecal sample processing in metagenomic studies. *Nature Biotechnology*, 35(11):1069, 2017.

Zhenwei Dai, Sunny H Wong, Jun Yu, and Yingying Wei. Batch effects correction for microbiome data with dirichlet-multinomial regression. *Bioinformatics*, 35(5):807–814, 2019.

Nicole M Davis, Diana M Proctor, Susan P Holmes, David A Relman, and Benjamin J Callahan. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome*, 6(1):1–14, 2018.

Matt Dowle and Arun Srinivasan. *data.table: Extension of 'data.frame'*, 2019. URL `https://CRAN.R-project.org/package=data.table`. R package version 1.12.6.

Lutz Dümbgen. On nondifferentiable functions and the bootstrap. *Probability Theory and Related Fields*, 95(1):125–140, 1993.

Jane Elith, John R Leathwick, and Trevor Hastie. A working guide to boosted regression trees. *Journal of Animal Ecology*, 77(4):802–813, 2008.

Andrew D Fernandes, Jennifer Ns Reid, Jean M Macklaim, Thomas A McMurrough, David R Edgell, and Gregory B Gloor. Unifying the analysis of high-throughput sequencing

datasets: characterizing rna-seq, 16s rrna gene sequencing and selective growth experiments by compositional data analysis. *Microbiome*, 2(1):1–13, 2014.

Jerome Friedman, Trevor Hastie, Saharon Rosset, Robert Tibshirani, and Ji Zhu. Discussion of boosting papers. *Ann. Statist*, 32:102–107, 2004.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1, 2010.

Johann A Gagnon-Bartsch and Terence P Speed. Using control genes to correct for unwanted variation in microarray data. *Biostatistics*, 13(3):539–552, 2012.

Charles J Geyer. On the asymptotics of constrained m-estimation. *The Annals of Statistics*, pages 1993–2010, 1994.

Sean M Gibbons, Claire Duvallet, and Eric J Alm. Correcting for batch effects in case-control microbiome studies. *PLOS Computational Biology*, 14(4):e1006102, 2018a.

Sean M Gibbons, Claire Duvallet, and Eric J Alm. Correcting for batch effects in case-control microbiome studies. *PLoS Computational Biology*, 14(4):e1006102, 2018b.

Stijn Hawinkel, Federico Mattiello, Luc Bijnens, and Olivier Thas. A broken promise: microbiome differential abundance methods do not control the false discovery rate. *Briefings in bioinformatics*, 20(1):210–221, 2019.

David V Hinkley. Bootstrap methods. *Journal of the Royal Statistical Society: Series B (Methodological)*, 50(3):321–337, 1988.

Bastian VH Hornung, Romy D Zwittink, and Ed J Kuijper. Issues and current standards of controls in microbiome research. *FEMS Microbiology Ecology*, 95(5):fiz045, 2019.

Luisa W Hugerth and Anders F Andersson. Analysing microbial community composition through amplicon sequencing: from sampling to hypothesis testing. *Frontiers in Microbiology*, 8:1561, 2017.

John PA Ioannidis, Evangelia E Ntzani, Thomas A Trikalinos, and Despina G Contopoulos-Ioannidis. Replication validity of genetic association studies. *Nature Genetics*, 29(3):306, 2001.

Hemant Ishwaran, Lancelot F James, and Mahmoud Zarepour. An alternative to the m out of n bootstrap. *Journal of Statistical Planning and Inference*, 139(3):788–801, 2009.

W Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127, 2007.

Lisa Karstens, Mark Asquith, Sean Davin, Damien Fair, W Thomas Gregory, Alan J Wolfe, Jonathan Braun, and Shannon McWeeney. Controlling for contaminants in low-biomass 16s rrna gene sequencing experiments. *MSystems*, 4(4):e00290–19, 2019.

Abhishek Kaul, Siddhartha Mandal, Ori Davidov, and Shyamal D Peddada. Analysis of microbiome data in the presence of excess zeros. *Frontiers in microbiology*, 8:2114, 2017.

Nicholas A Kennedy, Alan W Walker, Susan H Berry, Sylvia H Duncan, Freda M Farquarson, Petra Louis, and John M Thomson. The impact of different dna extraction kits and laboratories upon the assessment of human gut microbiota composition by 16s rRNA gene sequencing. *PLOS ONE*, 9(2):e88982, 2014.

Rob Knight, Alison Vrbanac, Bryn C Taylor, Alexander Aksenov, Chris Callewaert, Justine Debelius, Antonio Gonzalez, Tomasz Kosciolek, Laura-Isobel McCall, Daniel McDonald, et al. Best practices for analysing microbiomes. *Nature Reviews Microbiology*, 16(7): 410–422, 2018.

Dan Knights, Justin Kuczynski, Emily S Charlson, Jesse Zaneveld, Michael C Mozer, Ronald G Collman, Frederic D Bushman, Rob Knight, and Scott T Kelley. Bayesian community-wide culture-independent microbial source tracking. *Nature Methods*, 8(9): 761–763, 2011.

Charity W Law, Yunshun Chen, Wei Shi, and Gordon K Smyth. voom: Precision weights unlock linear model analysis tools for rna-seq read counts. *Genome biology*, 15(2):1–17, 2014.

Jeffrey T Leek and John D Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS genetics*, 3(9):e161, 2007.

Eric Loken and Andrew Gelman. Measurement error and the replication crisis. *Science*, 355 (6325):584–585, 2017.

Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):1–21, 2014.

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

Jeremy L Martin, Molly Maxwell, Victor Reiner, and Scott O Wilson. Pseudodeterminants and perfect square spanning tree counts. *arXiv preprint arXiv:1311.6686*, 2013.

Peter McCullagh. Quasi-likelihood functions. *The Annals of Statistics*, 11(1):59–67, 1983.

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

Raphaël Méheust, David Burstein, Cindy J Castelle, and Jillian F Banfield. The distinction of CPR bacteria from other bacteria based on protein family content. *Nature Communications*, 10(1):1–12, 2019.

Samuel S Minot and Amy D Willis. Strategies to facilitate translational advances from microbiome surveys. *Trends in Microbiology*, 2020.

Jacob T Nearing, André M Comeau, and Morgan GI Langille. Identifying biases and their potential solutions in human microbiome studies. *Microbiome*, 9(1):1–22, 2021.

Assaf P Oron and Nancy Flournoy. Centered isotonic regression: point and interval estimation for dose–response studies. *Statistics in Biopharmaceutical Research*, 9(3):258–267, 2017.

Joseph Nathaniel Paulson, Mihai Pop, and Hector Corrada Bravo. metagenomeseq: Statistical analysis for sparse high-throughput sequencing. *Bioconductor package*, 1(0):191, 2013.

Jolinda Pollock, Laura Glendinning, Trong Wisedchanwet, and Mick Watson. The madness of microbiome: attempting to find consensus "best practice" for 16S microbiome studies. *Appl. Environ. Microbiol.*, 84(7):e02627–17, 2018.

Christian Quast, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jörg Peplies, and Frank Oliver Glöckner. The silva ribosomal rna gene database project: improved data processing and web-based tools. *Nucleic Acids Research*, 41(D1):D590–D596, 2012.

Thomas P Quinn, Ionas Erb, Greg Gloor, Cedric Notredame, Mark F Richardson, and Tamsyn M Crowley. A Field Guide for the Compositional Analysis of Any-omics Data. *GigaScience*, 8(9), 2019. ISSN 2047-217X. doi: 10.1093/gigascience/giz107. URL `https://doi.org/10.1093/gigascience/giz107`. giz107.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019. URL `https://www.R-project.org/`.

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

Susannah J Salter, Michael J Cox, Elena M Turek, Szymon T Calus, William O Cookson, Miriam F Moffatt, Paul Turner, Julian Parkhill, Nicholas J Loman, and Alan W Walker.

Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biology*, 12(1):1–12, 2014.

Nicola Segata, Jacques Izard, Levi Waldron, Dirk Gevers, Larisa Miropolsky, Wendy S Garrett, and Curtis Huttenhower. Metagenomic biomarker discovery and explanation. *Genome biology*, 12(6):1–18, 2011.

Liat Shenhav, Mike Thompson, Tyler A Joseph, Leah Briscoe, Ori Furman, David Bogumil, Itzhak Mizrahi, Itsik Pe'er, and Eran Halperin. Feast: fast expectation-maximization for microbial source tracking. *Nature Methods*, 16(7):627–632, 2019.

Justin D Silverman, Rachael J Bloom, Sharon Jiang, Heather K Durand, Eric Dallow, Sayan Mukherjee, and Lawrence A David. Measuring and mitigating pcr bias in microbiota datasets. *PLoS Computational Biology*, 17(7):e1009113, 2021.

Joseph P Simmons, Leif D Nelson, and Uri Simonsohn. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11):1359–1366, 2011.

Andrew H Sims, Graeme J Smethurst, Yvonne Hey, Michal J Okoniewski, Stuart D Pepper, Anthony Howell, Crispin J Miller, and Robert B Clarke. The removal of multiplicative, systematic bias allows integration of breast cancer gene expression datasets–improving meta-analysis and prediction of prognosis. *BMC Medical Genomics*, 1(1):1–14, 2008.

Rashmi Sinha, Christian C Abnet, Owen White, Rob Knight, and Curtis Huttenhower. The Microbiome Quality Control Project: baseline study design and future directions. *Genome Biology*, 16(1):276, 2015.

Rashmi Sinha, Galeb Abu-Ali, Emily Vogtmann, Anthony A Fodor, Boyu Ren, Amnon Amir, Emma Schwager, Jonathan Crabtree, Siyuan Ma, Christian C Abnet, Rob Knight, Owen White, and Curtis Huttenhower. Assessment of variation in microbial community

amplicon sequencing by the Microbiome Quality Control (MBQC) project consortium. *Nature Biotechnology*, 486(11):207, 2017.

Jun Sun and Ikuko Kato. Gut microbiota, inflammation and colorectal cancer. *Genes & diseases*, 3(2):130–143, 2016.

Duy Tin Truong, Eric A Franzosa, Timothy L Tickle, Matthias Scholz, George Weingart, Edoardo Pasolli, Adrian Tett, Curtis Huttenhower, and Nicola Segata. Metaphlan2 for enhanced metagenomic taxonomic profiling. *Nature Methods*, 12(10):902–903, 2015.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Heidi C Vebø, Magdalena Kauczynska Karlsson, Ekaterina Avershina, Lene Finnby, and Knut Rudi. Bead-beating artefacts in the bacteroidetes to firmicutes ratio of the human stool metagenome. *Journal of Microbiological Methods*, 129:78–80, 2016.

Yiwen Wang and Kim-Anh LêCao. Managing batch effects in microbiome data. *Briefings in bioinformatics*, 21(6):1954–1970, 2020.

Sophie Weiss, Amnon Amir, Embriette R Hyde, Jessica L Metcalf, Se Jin Song, and Rob Knight. Tracking down the sources of experimental contamination in microbiome studies. *Genome Biology*, 15(12):1–3, 2014.

Dana Willner, Joshua Daly, David Whiley, Keith Grimwood, Claire E Wainwright, and Philip Hugenholtz. Comparison of dna extraction methods for microbial community profiling with an application to pediatric bronchoalveolar lavage samples. *PloS One*, 7(4):e34605, 2012.

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.

Jakob Wirbel, Konrad Zych, Morgan Essex, Nicolai Karcher, Ece Kartal, Guillem Salazar, Peer Bork, Shinichi Sunagawa, and Georg Zeller. Microbiome meta-analysis and cross-disease comparison enabled by the SIAMCAT machine learning toolbox. *Genome biology*, 22(1):1–27, 2021.

Ni Zhao and Glen A Satten. A log-linear model for inference on bias in microbiome studies. In *Statistical Analysis of Microbiome Data*, pages 221–246. Springer, 2021.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.

Appendix A

ASSESSING REPLICABILITY IN MICROBIOME DATA

A.1 Proofs

Lower bound for total variation distance

Consider an arbitrary measurable function $\phi : \mathbb{R}^J \to \mathcal{Z}$ where $\mathcal{Z} = \{1, \ldots, I\}$, and note that any classification rule on classes $1, \ldots, I$ based on data $X \subset \mathbb{R}^J$ can be expressed as such a function. Letting Q_k be the joint distribution of class labels $Z \in \mathcal{Z}$ and taxonomic covariates X (with sample space $\mathcal{X}$) in a sequencing laboratory k, we have, for the absolute value of the difference of expected misclassification error in laboratories k and k' of an arbitrary classification rule based on test function ϕ:

$$
\begin{aligned}
&|Q_k \mathbf{1}_{[\phi(X)\neq Z]} - Q_{k'} \mathbf{1}_{[\phi(X)\neq Z]}| \\
&\qquad = \; |Q_k(1 - \mathbf{1}_{[\phi(X)=Z]}) - Q_{k'}(1 - \mathbf{1}_{[\phi(X)=Z]})| \\
&\qquad = \; |Q_k \mathbf{1}_{[\phi(X)=Z]} - Q_{k'} \mathbf{1}_{[\phi(X)=Z]}| \\
&\qquad = \; |\int_{\{(x,z)\in\mathcal{X}\times\mathcal{Z}:\ \phi(x)=z\}} dQ_k(x,z) - \int_{\{(x,z)\in\mathcal{X}\times\mathcal{Z}:\ \phi(x)=z\}} dQ_{k'}(x,z)| \\
&\qquad \leq \; \sup_{A\subset\mathcal{X}\times\mathcal{Z}} |\int_A dQ_k(x,z) - \int_A dQ_{k'}(x,z)| \\
&\qquad = \; d_{\mathrm{TV}}(Q_k, Q_{k'})
\end{aligned}
$$

Therefore, letting the joint distribution of covariates (i.e., sample-centered transformed counts) and specimens in a test set in laboratory k follow Q_k and the corresponding data in laboratory k' follow $Q_{k'}$, the difference in misclassification error of an arbitrary classification rule in laboratories k and k' is bounded in expectation by the total variation distance between Q_k and $Q_{k'}$.

A.2 Supplementary Methods

Data Exclusions

To evaluate the consistency of the taxonomic profile of identical biological samples observed by each lab, we trained classifiers to predict the specimen number of each sample (details below). We considered samples from freeze-dried human stool, fresh human stool samples, chemostat and mock communities that were extracted at the laboratory at which they were sequenced.

As few combinations of wet lab and bioinformatics lab had complete data among non-pre-extracted samples (no aliquots excluded by bioinformatics), we adopted a relaxed standard of sufficient completeness. Under this standard, we considered any combination of a wet and a dry lab to have sufficiently complete data for inclusion in our analysis only if more than 75% of aliquots were available, and only if data for every specimen was reported.

The imbalance of bioinformatics laboratory data available across aliquots from sequencing laboratories may create confounding between bioinformatics effects and sequencing laboratory effects. To minimize the impact of this confounding on our results, which concern sequencing laboratory effects, we constructed a procedure to find a subset of the bioinformatics labs and sequencing labs described in Sinha et al. [2017] containing no excluded combination. We excluded all sequencing laboratories for which no bioinformatics laboratory provided sufficiently complete data as defined above. We then exhaustively searched all combinations of sequencing laboratory and bioinformatics laboratory to find sets of sequencing laboratories that had sufficiently complete data across a common set of at least four bioinformatics laboratories. We set this minimum number of bioinformatics laboratories to ensure sufficient data within wet lab to train and evaluate classifiers on. We found two combinations of 8 sequencing laboratories that met these criteria: HL-E and HL-I. We chose to use the combination containing HL-I as HL-I had more complete data than HL-E.

Treatment of Unclassified Reads

Some bioinformatics laboratories classified some reads from sequencing lab data as unclassified at various taxonomic levels. We chose to include such reads classified to an unknown taxon in our analysis. We regard both misclassification of a known organism as an unknown organism and heterogeneity across sequencing labs of organisms classified as unknown as forms of measurement error that impact the replicability of taxonomic profiling via 16S.

Classifier Training and Validation

We selected boosted regression tree classifier parameters for each combination of wet lab, classification task, and level of taxonomic aggregation. There are four parameters to select:

1. Proportion of observations sampled per boosting step (0.5 or 1)
2. Proportion of covariates used at each boosting step (0.25, 0.5, 0.75, or 1)
3. Learning rate
4. Number of boosting steps

For each combination of the first two parameters, we trained a classifier minimizing multinomial logistic loss starting at learning rate 0.1 for 10,000 boosting steps, or until cross-validated misclassification error had not improved for 500 boosting steps, whichever occurred first. We then retrained for up to 5 iterations, with learning rate decreased by $\sqrt{10}$ at each successive iteration. If at any iteration, optimal 10-fold cross-validated misclassification error occurred after at least 1000 boosting steps, following the rule of thumb proposed by Elith et al. [2008], we ended iteration over training rate. We also ended iteration early if a model reached cross-validated misclassification rate 0. We then compared all fitted boosted tree models and selected parameters corresponding to the model with the lowest 10-fold cross-validated misclassification error. We resolved ties in favor of whichever model took the greatest number of boosting steps to reach the lowest cross-validated misclassification error, and any

further ties in favor of, in order, lowest learning rate, variable subsampling, and observation subsampling. We fit our boosted trees using the `R` package xgboost [Chen et al., 2019].

We also selected elastic net parameters for each combination of wet lab, classification task, and level of taxonomic aggregation. There are two elastic net parameters to select:

1. λ: penalty magnitude

2. $\alpha \in \{0, 0.1, 0.2, \ldots, 1\}$: mixing proportion between L_1 and L_2 penalties

For each value of α, we evaluated many values of λ via automatic penalty selection algorithm provided in the `cv.glmnet` function in `glmnet` [Friedman et al., 2010]. We then selected α and λ according to minimum 10-fold cross-validated misclassification error, resolving ties in favor of values of α closer to 0.5, and resolving any further ties in favor of the lower value of α.

A.3 Data Completeness

	HL−A	HL−B	HL−C	HL−D	HL−E	HL−F_1	HL−F_2	HL−H	HL−I	HL−J	HL−K	HL−L	HL−M	HL−N_1	HL−N_2
BL−1	35	106	97	0	106	100	51	159	0	265	53	53	15	26	27
BL−2	35	105	97	0	71	99	51	159	53	265	53	53	15	26	27
BL−3	0	105	97	0	105	95	48	159	0	259	53	0	15	26	27
BL−4	35	105	97	44	106	99	51	159	53	265	53	53	15	26	27
BL−6	35	105	97	44	71	99	51	159	53	265	53	53	15	26	27
BL−8	6	104	97	40	103	91	46	156	53	264	52	53	15	26	27
BL−9A	35	103	97	0	106	99	51	159	0	264	53	53	15	26	27
BL−9B	0	106	97	44	71	101	53	159	53	265	53	0	15	26	27

Figure A.1: The number of raw (i.e., not centrally extracted) aliquots reported for each combination of sequencing (columns) and bioinformatics (rows) laboratories. Raw aliquots were distributed in sets of 53, with some sequencing laboratories analyzing multiple sets.

	HL−A	HL−B	HL−C	HL−D	HL−E	HL−F_1	HL−F_2	HL−H	HL−I	HL−J	HL−K	HL−L	HL−M	HL−N_1	HL−N_2
BL−1	22	22	22	0	22	22	22	22	0	22	22	22	10	18	17
BL−2	22	22	22	0	22	22	22	22	22	22	22	22	10	18	17
BL−3	0	22	22	0	22	22	22	22	0	22	22	0	10	18	17
BL−4	22	22	22	21	22	22	22	22	22	22	22	22	10	18	17
BL−6	22	22	22	21	22	22	22	22	22	22	22	22	10	18	17
BL−8	5	22	22	20	22	22	21	22	22	22	22	22	10	18	17
BL−9A	22	22	22	0	22	22	22	22	0	22	22	22	10	18	17
BL−9B	0	22	22	21	22	22	22	22	22	22	22	0	10	18	17

Figure A.2: The number of unique specimens for which any data was reported in each combination of sequencing (columns) and bioinformatics (rows) laboratories. 22 unique specimens (excluding negative controls) were sent to each sequencing laboratory.

A.4 Additional Results

Replicability within subgroups of laboratories

Our primary analysis admits the possibility that between-specimen signals replicate well within some subgroup of sequencing laboratories. In figure A.3 we present results for four sequencing laboratories with generally low within-laboratory misclassification. Figure A.3 indicates that the gap between within-laboratory and cross-laboratory misclassification is generally largest for signals learned on proportion data. By contrast, between-specimen signals learned on centered log ratio data best replicate within and across laboratories in this comparison. At fine levels of taxonomy, signals learned on presence-absence data exhibit relatively strong within- and cross-laboratory replicability as well.

Alas, under no transformation nor level of taxonomic aggregation does within-laboratory replication well predict cross-laboratory replication. For instance, HL-B exhibits generally low within-laboratory misclassification, but replicates relatively poorly on laboratories F, H, and J. Similarly, replication of a signal detected by one laboratory in another does not guarantee mutual replicability: centered log-ratio signals at fine levels of taxonomy in HL-F

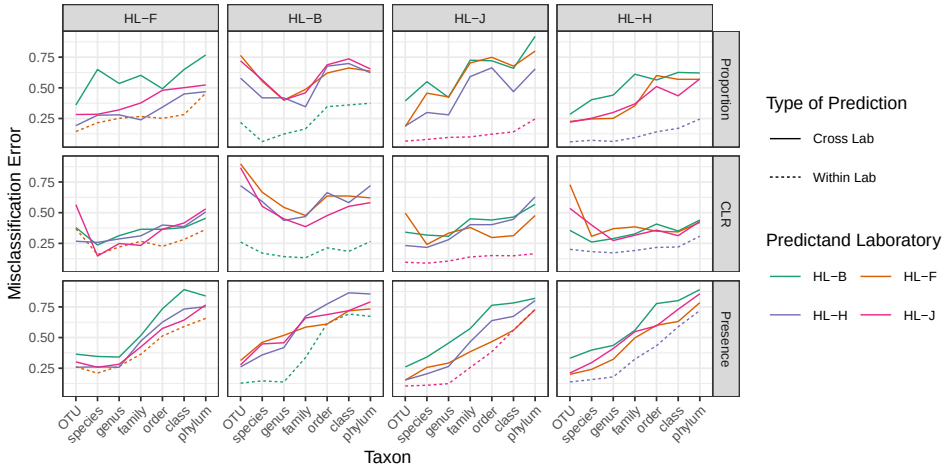

Figure A.3: Misclassification error of boosted tree specimen classifiers trained on centered proportion (top row), centered log ratio (middle row), and presence-absence (bottom row) data from four sequencing laboratories (columns). Across taxa (x-axis), within-laboratory misclassification (dotted lines) is typically lower than cross-laboratory misclassification (solid lines), but the size of this discrepancy depends on predictor laboratory, transformation, and taxonomic aggregation level.

replicate fairly well in HL-B (24% versus 16% misclassification on species-level data from HL-B and HL-F, respectively), but signals learned on HL-B replicate relatively poorly in HL-F (66% versus 17% misclassification on species-level data from HL-F and HL-B, respectively). Taken together, these results suggest that low observed technical variation in a group of sequencing laboratories does not guarantee replicability of between-group signals detected by these laboratories. Moreover, replication of the results of one laboratory by another does not guarantee that the converse replication will hold.

Since our focus is assessing replicability, we chose not to model labs' protocol variables. However, the results presented in Figure A.3 are consistent with known results about which protocol variables influence sample measurements. For instance, HL-F, HL-H, and HL-J generally replicate better on each other than on HL-B. HL-F, HL-H, and HL-J used extraction kits from the same manufacturer, while HL-B used a extraction kit from a different manufacturer. Although we cannot conclusively attribute this pattern to extraction protocol, it is in line with the large extraction effects reported in the microbiome literature [McLaren et al., 2019, Vebø et al., 2016].

A.5 Anomalous misclassification results in centered proportion data

In Figure 2 (main text), we observed a spike in the misclassification of elastic net classifiers trained on HL-B for specimen at the order level. We also observed a spike in the misclassification of elastic net classifiers trained on HL-L for specimen type at the genus level. In this section we investigate the source of these spikes.

In Figure A.4 we see that the within-laboratory elastic net classifier trained on on HL-B centered proportion order-level data erroneously categorizes many aliquots as originating from sample 63. Figure A.5, which shows average values of elastic net linear predictors (by specimen) in the HL-B centered proportion order-level training and test sets, suggests that poor performance on the HL-B test set is due to the influence of differing measurements in order Sphingomonadales between the training and test sets.

In Figure A.6 we see that the within-laboratory elastic net classifier trained on on HL-L centered proportion genus-level data erroneously categorizes many aliquots as originating from samples 61 and 101. Figure A.7, which shows average values of elastic net linear predictors (by specimen) in the HL-L centered proportion genus-level training and test sets, suggests that poor performance on the HL-L test set is due to measurements across various genera being discordant between the training and test sets.

Together these results show that discrepancies in the misclassification level of the elastic classifiers can be due to unusual test/train splits on a transformation of the data (i.e., proportion) at which error in any taxon propagates to all other taxa. However, we observed spikes in the misclassification error in only two of our elastic net classifiers. This gives us confidence that, in general, our classifiers are picking up distinctions between samples based on the data from the training laboratory, and not random noise.

	sample10	sample101	sample102	sample103	sample104	sample11	sample12	sample13	sample14	sample2	sample3	sample4	sample55	sample56	sample57	sample59	sample6	sample61	sample63	sample64	sample8	sample9
sample10	43	0	0	3	0	0	0	0	3	0	0	0	1	8	0	0	0	0	8	0	0	0
sample101	0	57	0	0	0	0	0	0	0	0	0	0	0	11	0	0	1	0	23	0	0	0
sample102	0	0	9	0	0	0	0	0	4	0	0	0	0	7	0	0	0	0	78	0	0	2
sample103	4	0	1	69	0	0	0	0	0	0	0	0	0	12	0	0	0	0	10	0	0	3
sample104	0	0	0	0	84	0	0	0	0	0	0	0	0	8	0	0	0	0	4	0	0	0
sample11	3	0	0	0	0	15	0	0	3	0	0	0	0	12	24	0	1	0	33	4	0	0
sample12	4	0	2	2	0	0	4	0	12	3	0	0	0	8	0	0	4	0	15	4	0	0
sample13	0	0	0	0	4	0	0	41	0	0	2	6	3	8	1	0	0	0	3	0	0	0
sample14	13	1	0	0	0	0	0	0	29	1	0	0	0	8	0	0	1	4	11	0	0	0
sample2	0	3	0	4	0	0	0	0	4	56	0	0	16	8	0	0	0	0	4	4	0	0
sample3	0	5	0	0	0	4	0	0	4	0	29	0	0	9	0	0	1	0	13	3	0	0
sample4	0	2	0	0	0	0	0	0	0	4	0	0	3	11	0	0	12	0	36	0	0	0
sample55	6	0	2	0	0	0	0	4	1	0	0	0	6	6	0	0	1	0	78	0	0	0
sample56	0	0	1	0	0	0	0	0	0	0	0	1	0	8	0	0	0	0	58	0	0	0
sample57	8	0	0	0	0	0	0	0	3	0	0	0	0	8	4	0	0	0	41	0	0	0
sample59	0	9	0	0	0	0	0	0	0	0	8	0	0	4	0	11	16	0	4	48	0	0
sample6	0	0	0	1	0	0	0	0	0	0	31	0	0	9	0	0	9	0	14	0	0	0
sample61	0	0	0	4	0	0	0	0	0	0	0	1	0	9	0	0	0	0	54	0	0	0
sample63	0	4	0	0	0	0	0	0	2	0	0	0	0	8	0	0	1	0	53	0	0	0
sample64	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	60	0	0
sample8	1	2	0	0	0	0	1	0	8	1	0	0	1	12	0	0	2	0	18	4	44	1
sample9	0	0	2	2	0	0	4	0	15	1	0	0	0	4	0	0	4	0	15	8	8	0

Figure A.4: A confusion matrix for within-laboratory elastic net classification on HL-B centered proportion order-level data. True labels are given as row names, and predicted labels are given in column names. This classifier erroneously categorizes many aliquots as originating from sample 63.

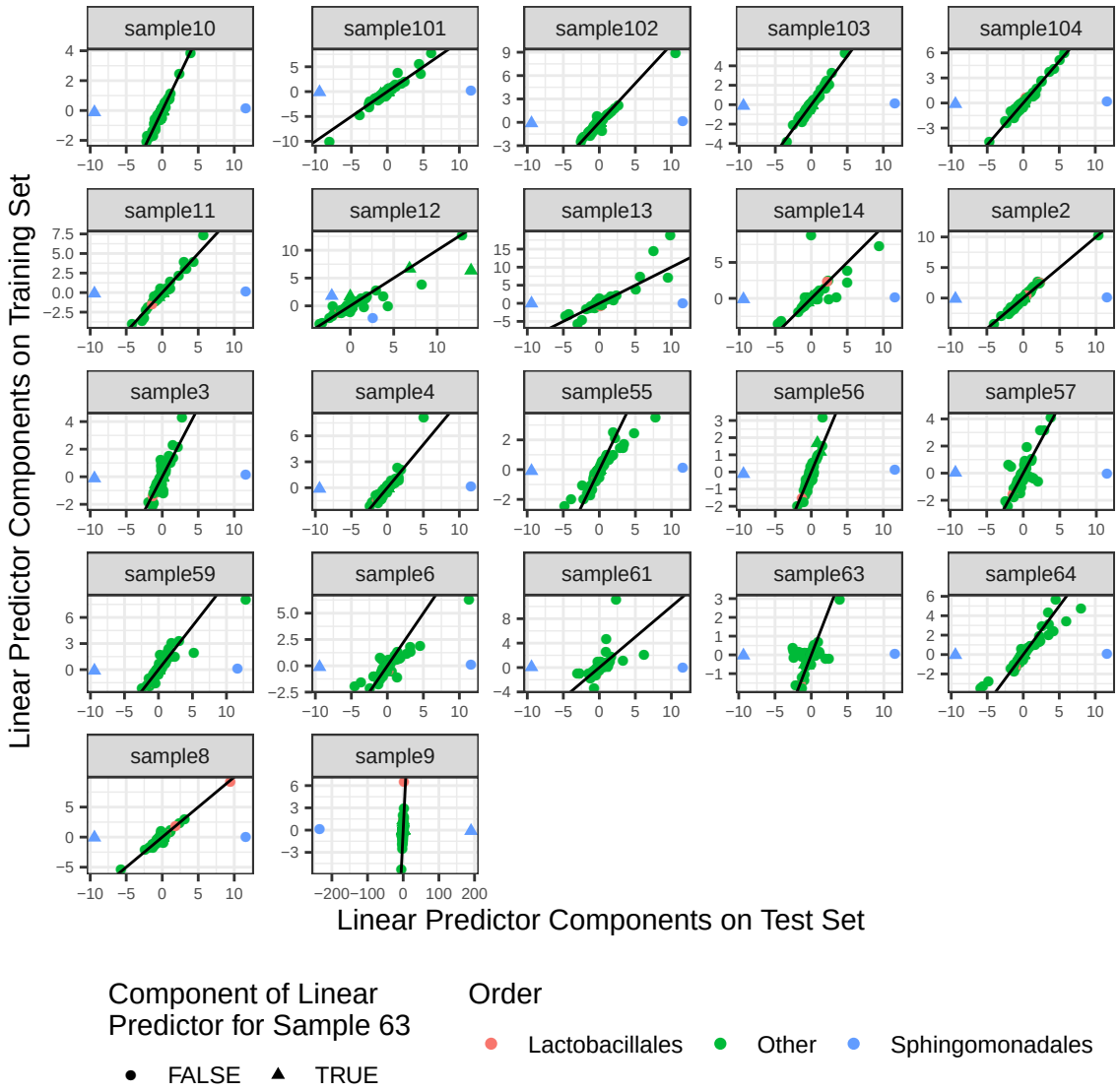

Figure A.5: Mean values of linear predictors (centered proportion multiplied by elastic net coefficients) by specimen for elastic net classifier trained on centered proportion order-level specimen data. The line x = y is shown in black. The linear predictor for order Sphingomonadales is generally large in aliquots in the training set, but not in the test set.

HL-L Genus Within-Lab Performance

	sample10	sample101	sample102	sample103	sample104	sample11	sample12	sample13	sample14	sample2	sample3	sample4	sample55	sample56	sample57	sample59	sample6	sample61	sample63	sample64	sample8	sample9
sample10	4	27	0	0	0	0	0	0	0	0	4	0	2	0	0	0	0	27	2	0	0	0
sample101	0	4	0	0	0	2	0	0	0	0	7	78	1	0	0	0	0	0	0	0	0	0
sample102	0	10	4	0	0	0	0	0	0	0	86	0	0	0	0	0	0	0	0	0	0	0
sample103	0	32	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	63	0	0	0	0
sample104	0	0	0	0	3	0	0	0	0	0	0	93	0	0	0	0	0	0	0	0	0	0
sample11	0	12	0	0	0	56	0	0	0	0	23	0	1	3	0	0	0	0	0	0	0	0
sample12	0	21	0	0	0	1	4	0	0	0	6	3	0	0	3	0	0	19	0	0	0	1
sample13	0	21	0	0	0	0	0	2	0	0	0	4	3	0	0	0	15	23	0	0	0	0
sample14	0	26	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0	35	0	0	0	0
sample2	0	38	1	0	0	2	0	0	0	4	5	0	0	0	0	0	0	47	0	0	2	0
sample3	0	4	0	0	0	1	0	0	0	0	55	4	0	0	0	0	0	4	0	0	0	0
sample4	0	4	0	0	0	2	0	0	0	0	56	0	0	0	0	0	6	0	0	0	0	0
sample55	1	44	0	0	0	14	0	0	0	0	0	0	6	0	0	0	0	39	0	0	0	0
sample56	0	18	0	0	0	0	0	0	0	0	24	0	0	4	0	0	0	20	2	0	0	0
sample57	0	16	0	0	0	17	0	0	0	0	0	0	0	0	4	0	0	27	0	0	0	0
sample59	0	37	0	0	0	0	0	0	0	0	0	0	0	0	0	21	4	32	4	0	0	2
sample6	0	5	0	0	0	0	0	0	0	0	51	0	0	0	0	0	8	0	0	0	0	0
sample61	0	19	0	0	0	12	0	0	0	0	0	0	0	0	0	0	0	37	0	0	0	0
sample63	0	20	0	0	0	12	0	0	2	0	0	0	0	4	0	0	0	20	7	0	0	3
sample64	0	31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	27	4	2	0	0
sample8	0	75	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	15	0	0	4	0
sample9	0	15	0	0	0	1	0	4	0	0	3	0	0	0	0	0	0	40	0	0	0	0

Figure A.6: A confusion matrix for within-laboratory elastic net classification on HL-L centered proportion genus-level data. True labels are given as row names, and predicted labels are given in column names. This classifier erroneously categorizes many aliquots as originating from samples 61 and 101.

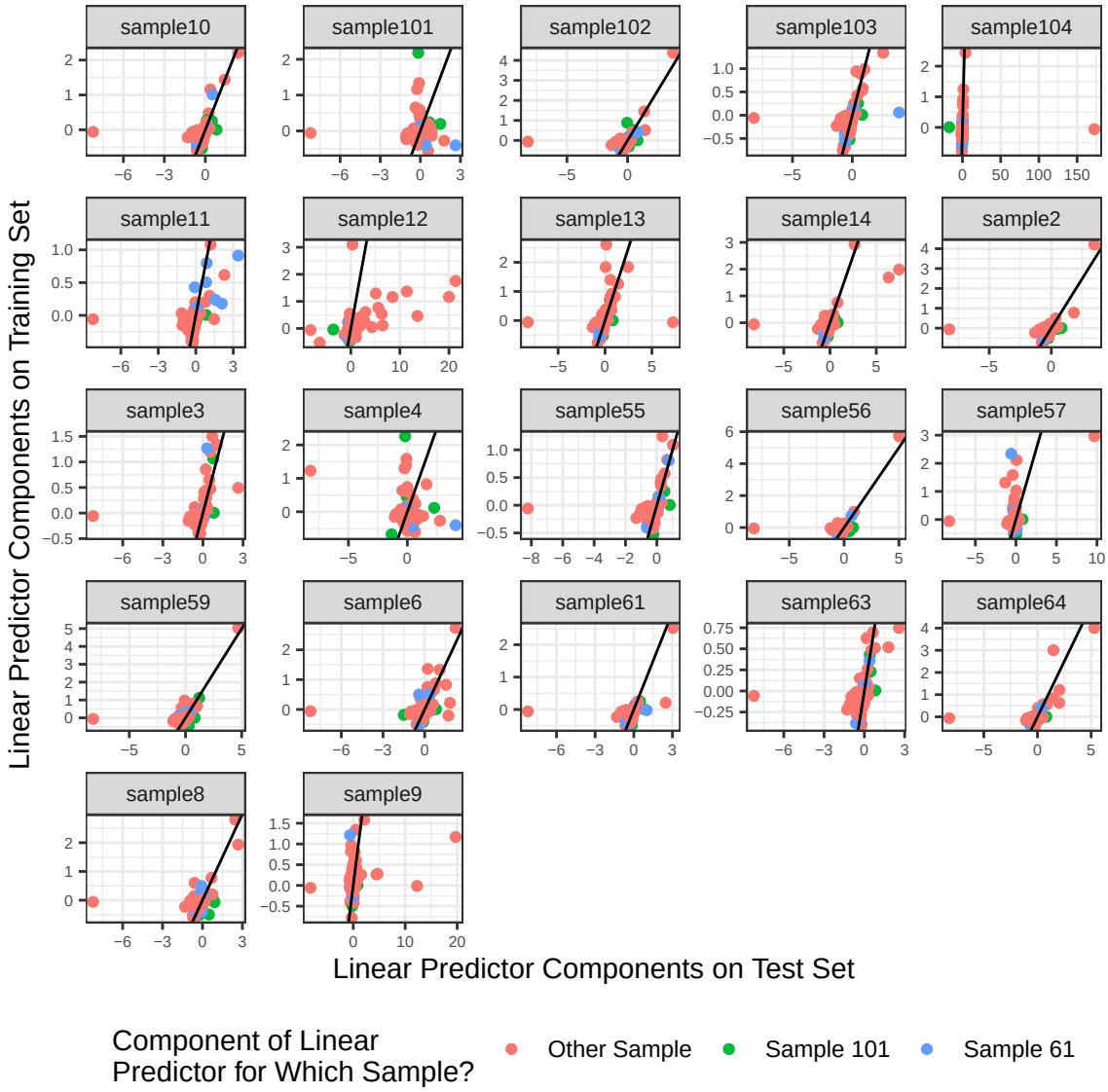

Figure A.7: Mean values of linear predictors (centered proportion multiplied by elastic net coefficients) by specimen for elastic net classifier trained on centered proportion genus-level specimen data. The line x = y is shown in black.

A.6 Bioinformatics Sensitivity Analysis

To assess the influence of our choice to treat results reported from different bioinformatics laboratories on individual samples sequenced by a single sequencing laboratory as replicate measurements, for each bioinformatics laboratory we included in our analysis, we fit and predicted from elastic net classifiers (classifying specimen) using only data reported by that laboratory. We limited this analysis to sequencing laboratories HL-B, HL-C, HL-F, HL-H, and HL-J, as other laboratories did not sequence enough samples to train and test elastic net classifiers on the basis of only one set of bioinformatics results.

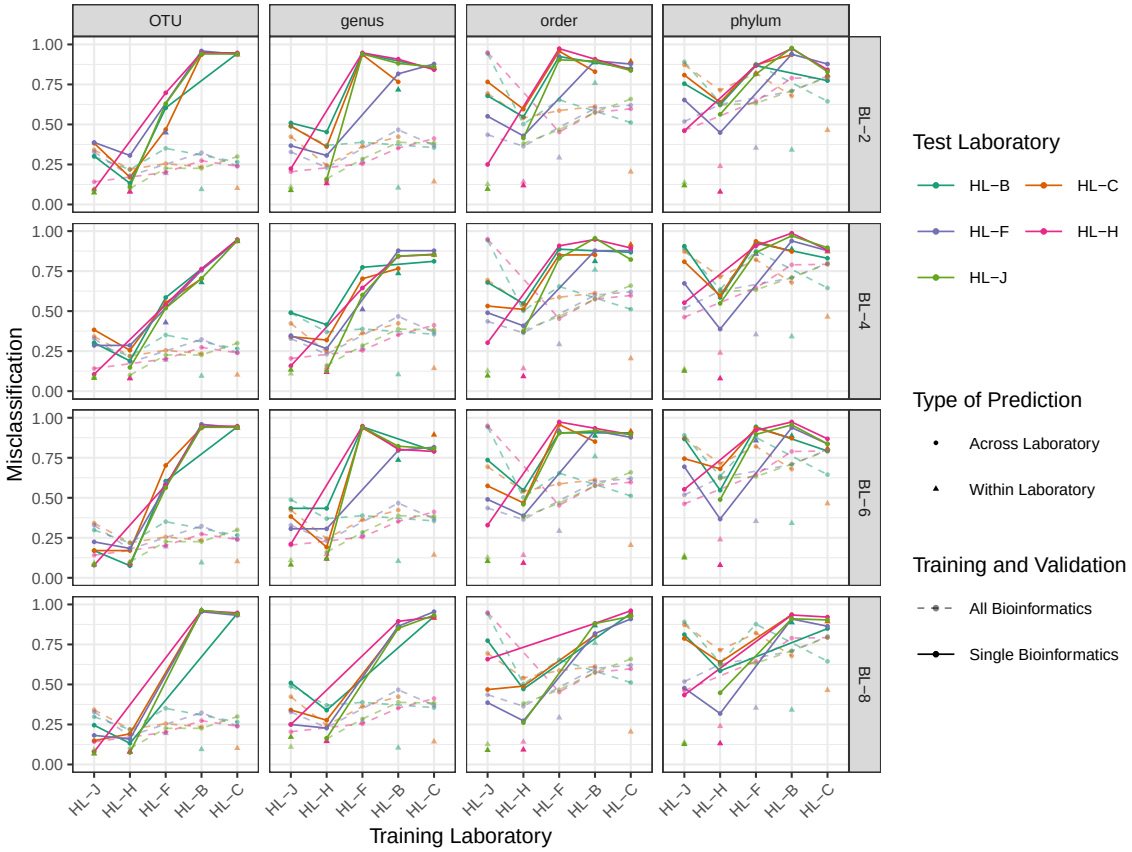

Figure A.8: Test set misclassification error for elastic net classifiers fit and validated on centered proportion data provided by individual bioinformatics laboratories (rows) is shown in bold against misclassification error for elastic net classifiers fit and validated on all four included bioinformatics laboratories, which is indicated by transparent points and lines. Performance of classifiers predicting within-sequencing-laboratory is indicated by triangles; cross-laboratory performance is shown with connected points.

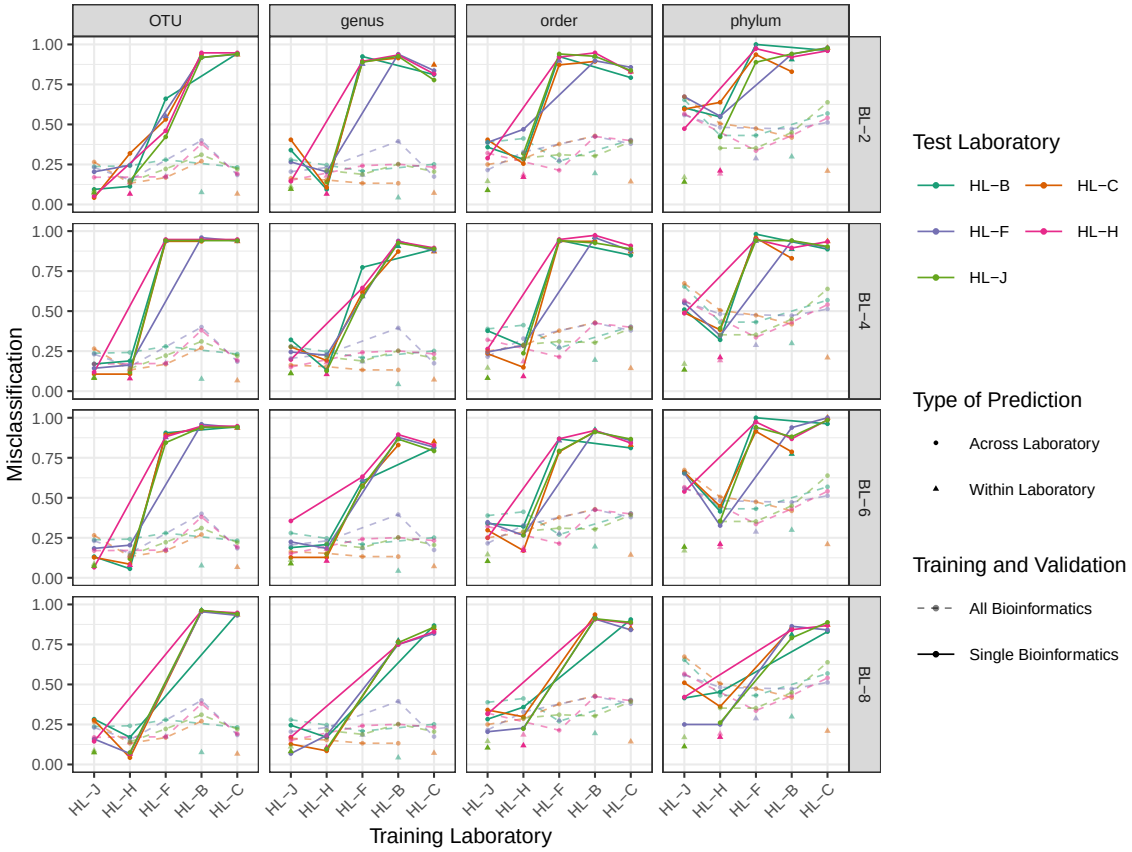

Figure A.9: Test set misclassification for error elastic net classifiers fit and validated on centered CLR data provided by individual bioinformatics laboratories (rows) is shown in bold against misclassification error for elastic net classifiers fit and validated on all four included bioinformatics laboratories, which is indicated by transparent points and lines. Performance of classifiers predicting within-sequencing-laboratory is indicated by triangles; cross-laboratory performance is shown with connected points.

For the two sequencing laboratories that processed the largest number of samples from sample types included in our analysis, HL-H (159 samples) and HL-J (265 samples), we are able to fit elastic net classifiers using data from single bioinformatics laboratories that distinguish (within sequencing laboratory) between specimens about as well as classifiers trained

on all bioinformatics laboratories. We also observe largely similar patterns of within- versus across-sequencing laboratory misclassification for classifiers trained on individual bioinformatics data from HL-H or HL-J as we do for classifiers trained on data from all bioinformatics laboratories for HL-J or HL-H.

For sequencing laboratories that processed fewer samples – HL-B (106 samples), HL-C (97 samples), HL-F (101 samples) – we are in general unable, on the basis of data from individual bioinformatics laboratories, to fit classifiers that distinguish between specimens within-laboratory. This renders comparison of cross-laboratory performance of these classifiers with cross-laboratory performance of classifiers fit on data from all bioinformatics laboratories difficult to interpret.

A.7 Classifier Performance under Centering and Scaling; Descriptive Analysis of Fresh Samples at Phylum Level

To explore whether taxon- and (sequencing) laboratory-specific scaling could explain our main results, we fit and predicted from elastic net classifiers using proportion- and centered log-ratio-transformed data to which we applied a sample centering (as described in section 3 of the manuscript) as well as a sample scaling. Under each transformation (proportion or CLR), in each test or training set, and for each taxon, we calculated scaling factors as the standard deviation of the mean observed value in each specimen (i.e., for each taxon we computed specimen means and calculated the standard deviation of these means). We then, after centering data, rescaled data by dividing by these taxon-specific scalings. We did not rescale taxa for which the computed scaling factor was 0. After applying this rescaling to each test and training set, we fit elastic net classifiers to identify specimen as in our primary analysis. The results of this analysis are shown in the figure below.

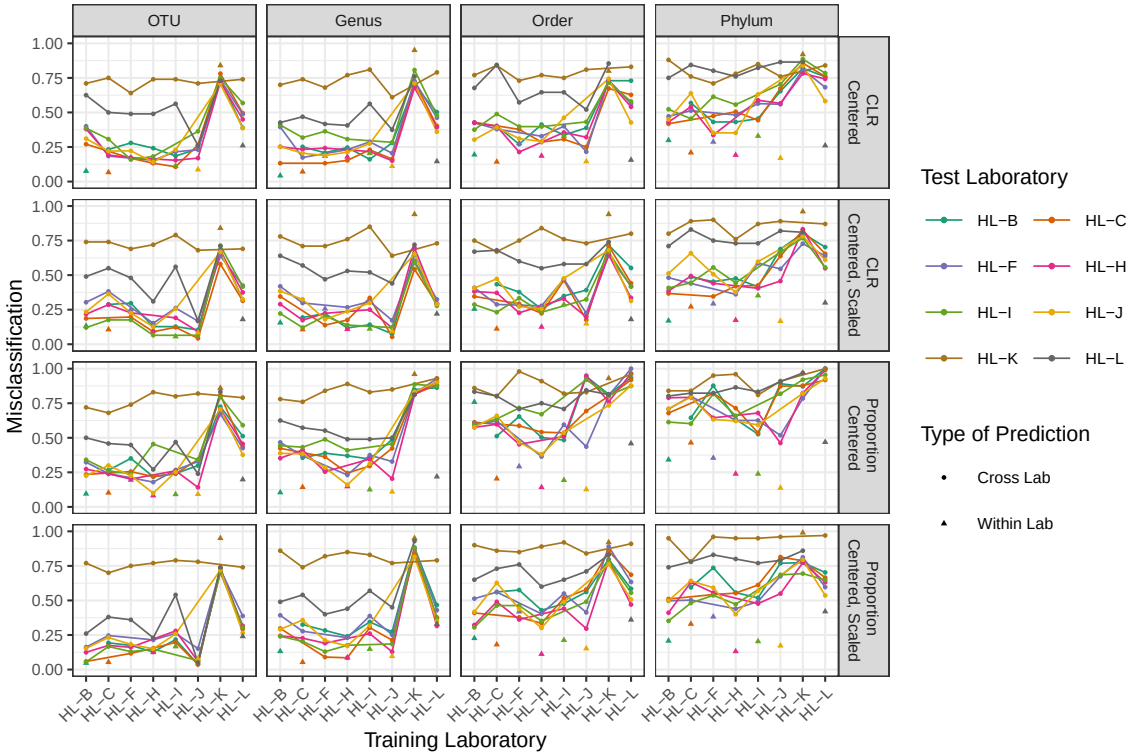

Figure A.10: Test set misclassification error for elastic net classifiers fit and validated on CLR (first two rows) and proportion data (third and fourth rows). For each transformation (CLR or proportion), we consider performance under sample centering (first and third rows) and under sample centering and scaling (second and fourth rows). Within-laboratory misclassification is indicated by triangles, and cross-laboratory misclassification is indicated by connected dots.

The comparison between within-laboratory and cross-laboratory misclassification is quite similar for centered and centered-and-scaled CLR data, with perhaps marginally better performance for classifiers fit and validated on scaled data than on unscaled genus data.

For proportion data, we observe greater improvement in cross-laboratory as compared to within-laboratory misclassification for classifiers fit on scaled data relative to those fit on unscaled data. This may reflect the fact that on the proportion scale, over- (or under-)

detection of an abundant taxon results in a compression (or stretching) of measured proportions in other taxa (on account of the sum-to-one constraint). However, while the gap between within- and cross-laboratory performance narrows for centered and scaled proportion data as compared to centered proportion data, cross-laboratory misclassification remains substantially higher than within-laboratory misclassification.

Descriptive Analyses

In order to provide intuition for the kinds of between-sequencing-laboratory differences we observe in between-specimen structure, we provide a descriptive analysis of abundances in fresh human samples for each of the four phyla (Firmicutes, Bacteroidetes, Actinobacteria, and Proteobacteria) directly analyzed by Sinha et al. [2017]. The following plots show measured abundances on the proportion and centered log-ratio scales in each of these phyla by sequencing laboratory, bioinformatics laboratory, health status, and specimen.

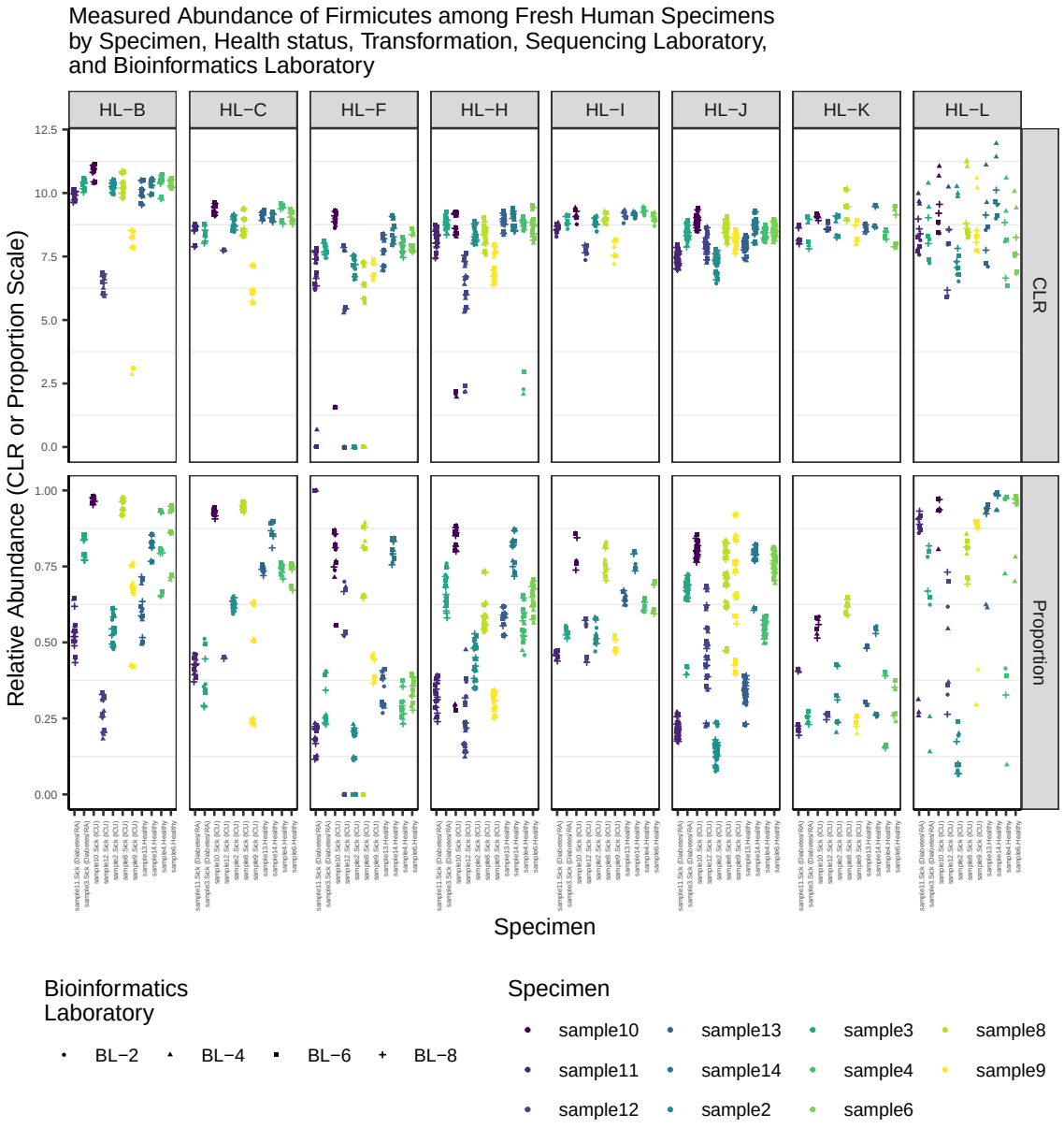

Figure A.11: Measured centered log-ratio (first row) and proportion (second row) Firmicutes abundance by sequencing laboratory (columns), bioinformatics laboratory (point shape), health status (x-axis), and specimen (color).

We observe, on both proportion and centered log-ratio scales, some shared between-specimen structure across sequencing laboratory. However, we also observe substantial qualitative differences across sequencing laboratory. For instance, is the centered log-ratio abundance of Firmicutes higher in specimen 13 (healthy; 4th from right) than in specimen 9 (ICU; 5th from right)? Measurements from HL-B and HL-H suggest so; measurements from HL-J do not. On the proportion scale, we, for instance, may come to substantially different conclusions about how similar the abundance of Firmicutes is in the two diabetes/RA samples depending on sequencing laboratory – HL-C suggests essentially no difference, HL-F suggests perhaps a small difference, and HL-H and HL-J suggest substantial differences.

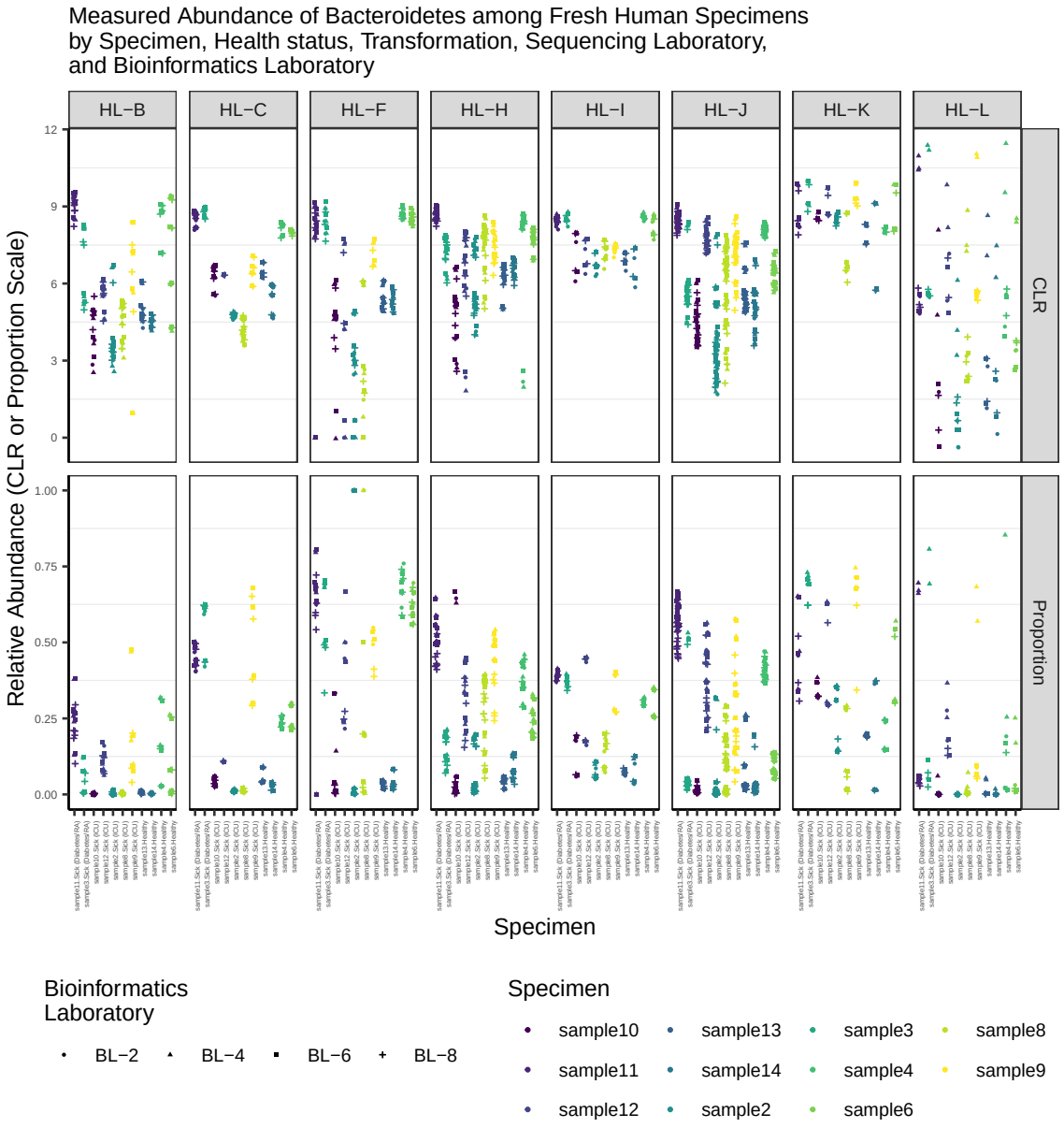

Figure A.12: Measured centered log-ratio (first row) and proportion (second row) Bacteroidetes abundance by sequencing laboratory (columns), bioinformatics laboratory (point shape), health status (x-axis), and specimen (color).

In terms of measured abundance of Bacteroidetes, we again observed partial concordance in terms of between-specimen structure across sequencing laboratories as well as examples of discordance. This is particularly evident at the proportion scale, where, for example, we might conclude that specimen 11 (first on left) has greater Bacteroidetes relative abundance than specimen 3 (second from left) if we consult HL-B or HL-H, whereas in HL-I we observe essentially no difference between these specimens. In a similar vein, determination of which specimen (among all specimens plotted) has highest relative abundance of Bacteroidetes differs across sequencing laboratory, with, notably, no consensus across laboratories regarding whether diabetes/RA (rheumatoid arthritis) specimens (at left) have generally higher or lower Bacteroidetes relative abundance than healthy specimens 4 and 6 (at right).

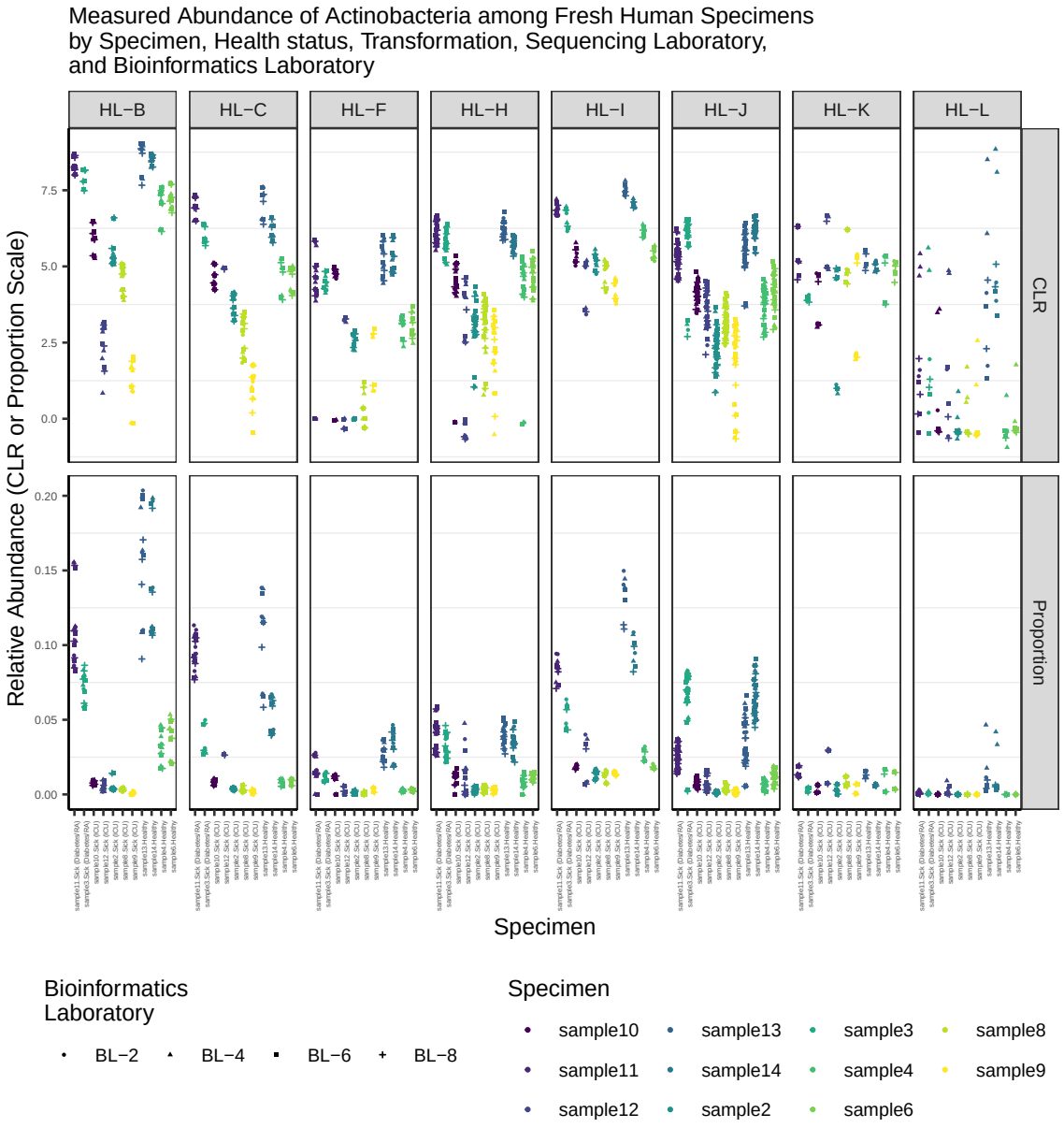

Figure A.13: Measured centered log-ratio (first row) and proportion (second row) Actinobacteria abundance by sequencing laboratory (columns), bioinformatics laboratory (point shape), health status (x-axis), and specimen (color).

We observe generally similar patterns in measured Actinobacteria abundance as we do for Firmicutes and Bacteroidetes. In particular, some comparisons on the proportion scale replicate less well across sequencing laboratory on the proportion than on the centered log-ratio scale. For example, how do specimens 13 and 14 (healthy; 3rd and 4th from right) compare to specimens 11 and 3 (diabetes/RA; 1st and 2nd from left)? Is relative abundance (i.e., proportion) Actinobacteria higher in sample 11 or sample 3?

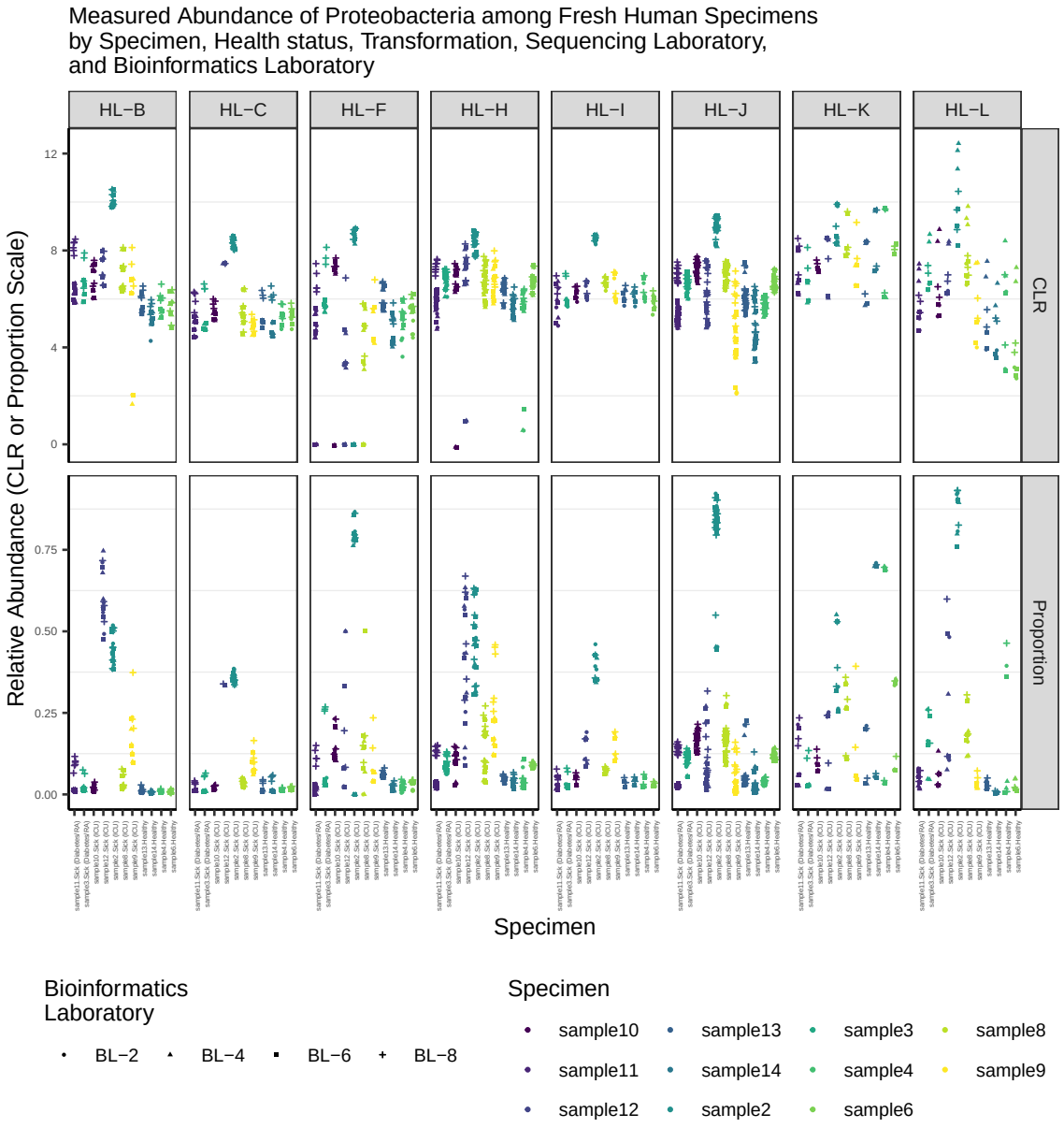

Figure A.14: Measured centered log-ratio (first row) and proportion (second row) Proteobacteria abundance by sequencing laboratory (columns), bioinformatics laboratory (point shape), health status (x-axis), and specimen (color).

On the centered log-ratio scale, all sequencing laboratories agree that specimen 2 (ICU; 5th from left) has highest abundance Proteobacteria, though the degree to which this abundance exceeds abundances in other specimens differs somewhat by sequencing laboratory. On the proportion scale, the picture is not so clear – we might judge specimen 12 to have higher relative abundance Proteobacteria than specimen 2 on the basis of measurements from HL-B, and sequencing laboratories, furthermore, do not agree regarding whether specimen 12 has higher relative abundance Proteobacteria than specimens 8 and 9. On the centered log-ratio scale, HL-B suggests somewhat elevated Proteobacteria in diabetes/RA specimens (1st and 2nd from left) as compared to healthy specimens (rightmost four), but this distinction essentially disappears in HL-C and HL-I.

A.8 Additional Centered Log-Ratio Figures

A.9 Classification on Presence Data

To assess the replicability of between-specimen signals observed under a commonly-used transformation of sequencing count data, the presence-absence transformation, we fit and predicted from classifiers using data at this scale. Concretely, we define the presence-absence transformation as follows:

$$\Psi_{:}\ \mathbb{R}^J \to \{0,1\}^J; \Psi(\vec{W}_{i\cdot k}) = \left(\mathbf{1}_{[W_{i1k}>0]}, \ldots, \mathbf{1}_{[W_{iJk}>0]}\right).$$

Consideration of 16S data on the presence-absence scale was motivated by the practice of treating 16S amplicon data as non-quantitative in the sense measured proportions of 16S variants do not reflect the true proportions of the taxa to which they are assigned [Méheust et al., 2019, Kennedy et al., 2014, Costa et al., 2012]. Hence, this line of reasoning suggests, it may be more appropriate to treat 16S amplicon data as indicating only presence or absence of taxa. That is, in the presence of measurement error, between-specimen comparisons based on presence-absence data should be either invariant or at least less variable than proportion-scale comparisons.

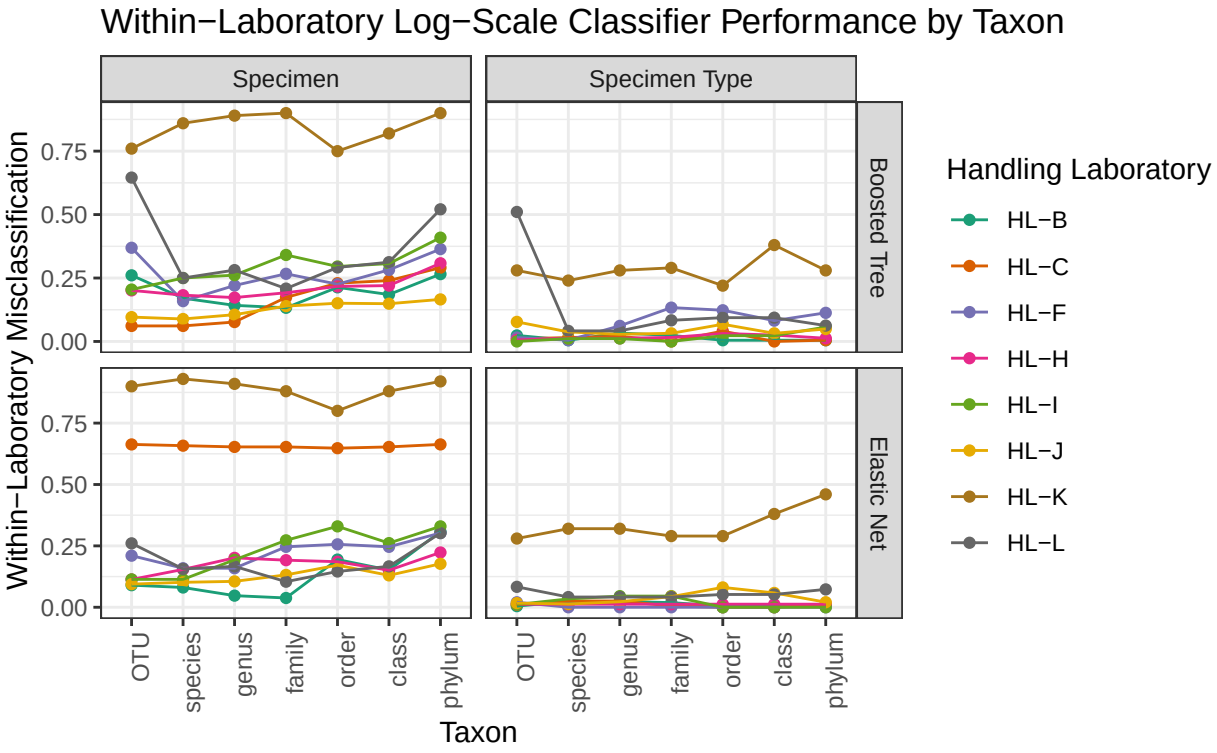

Figure A.15: Within-laboratory misclassification for boosted tree and elastic net classifiers predicting specimen or specimen type on centered log-ratio data plotted against level of taxonomic aggregation. Color indicates the laboratory that the classifier was trained and evaluated on.

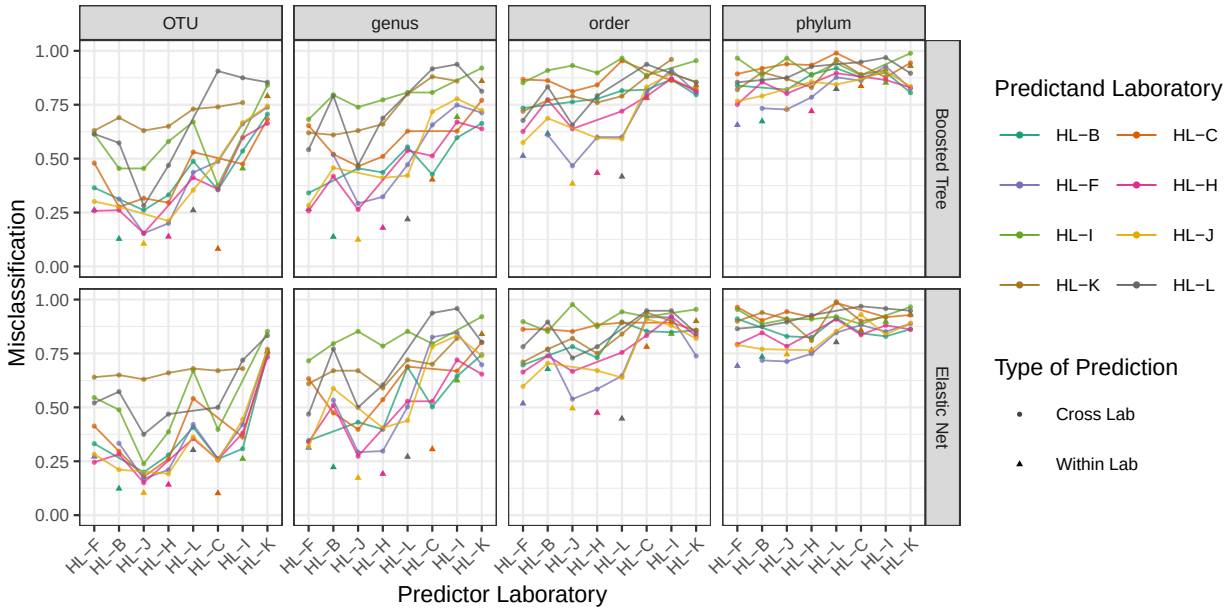

Figure A.16: Within-laboratory (solid triangles) and between-laboratory (lines) misclassification for boosted tree and elastic net classifiers predicting specimen on presence data plotted against level of taxonomic aggregation.

We chose not to center presence-absence data, as this would produce continuous data, and we were interested in the consistency of binary measurements across laboratories. Hence, the error model presented in Section 2.2.3 is not directly relevant here. The results in this section therefore pertain to whether observed patterns of differential presence of taxa across specimens are likely to replicate across different sequencing laboratories.

At every level of taxonomic aggregation, and for both elastic net and boosted tree classification, median cross-laboratory misclassification of specimen classifiers is larger than median within-laboratory misclassification (Figure A.16). Both within- and cross-laboratory misclassification increase with increasing level of taxonomic aggregation. On OTU-level data, median within-laboratory misclassification of boosted tree classifiers is 20% (IQR 12% - 31%), versus 49% (IQR 33% - 66%) median cross-laboratory misclassification. On phylum data,

these figures are 78% (IQR 71% - 84%) and 88% (IQR 84% - 92%), respectively. Similar patterns hold for elastic net classifiers, with 27% (IQR 14% - 39%) and 46% (IQR 28% - 68%) within- and cross-laboratory misclassification on OTU-level data, rising to 78% (IQR 74% - 87%) and 89% (IQR 84% - 92%), respectively, on phylum data.

A.10 Comparison Across Transformations

This section contains within- and cross-laboratory misclassification results for all levels of taxonomic aggregation, as well as results on data that has not been sample-centered.

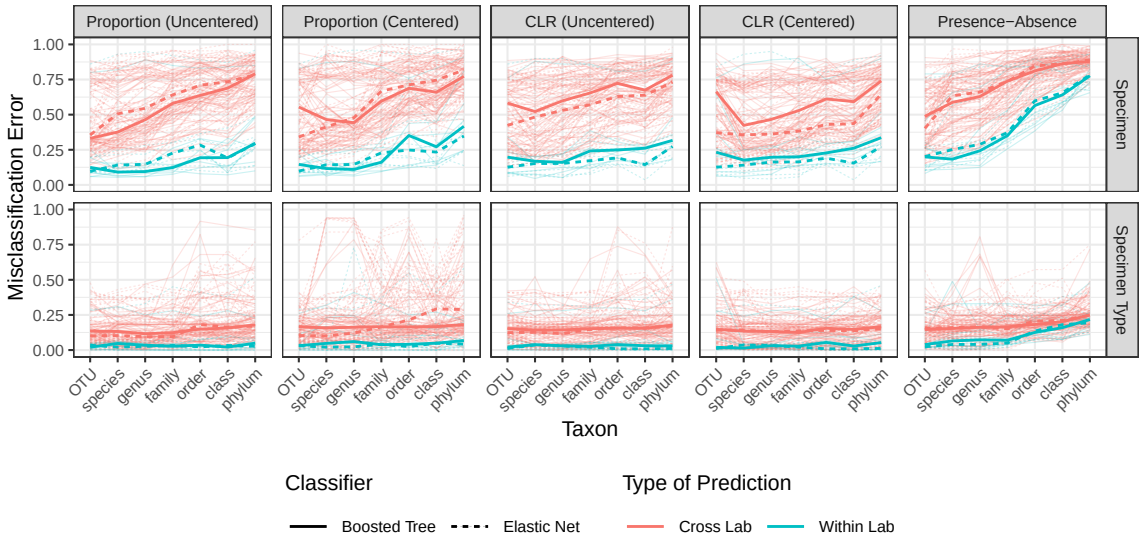

Figure A.17: The within (aqua) and across (red) laboratory misclassification for uncentered proportions, centered proportions, uncentered log ratio, centered log ratio and presence absence data for both classifying both specimen and specimen type. The misclassification rate is shown for boosted tree (solid lines) and elastic net (dotted lines) classifiers for every combination of laboratories (thin lines) and is also summarized as a median across laboratory combinations (thick lines). We see that centering the centered log ratio transformation improves the misclassification rate, but centering the proportions does not improve the misclassification rate.

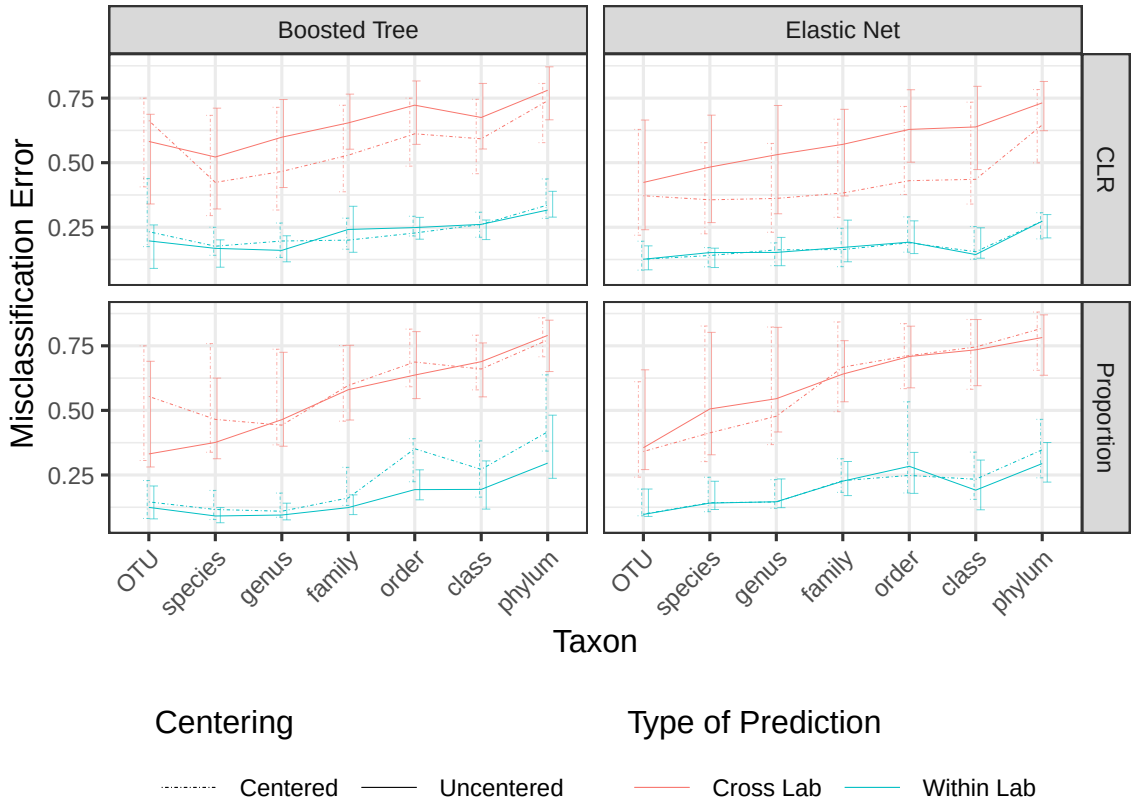

Figure A.18: Median within-laboratory (aqua) and cross-laboratory (red) specimen misclassification for proportion and centered-log-ratio data, with and without sample centering. The interquartile range is shown in brackets at each taxon.

Results for Uncentered Proportion Data

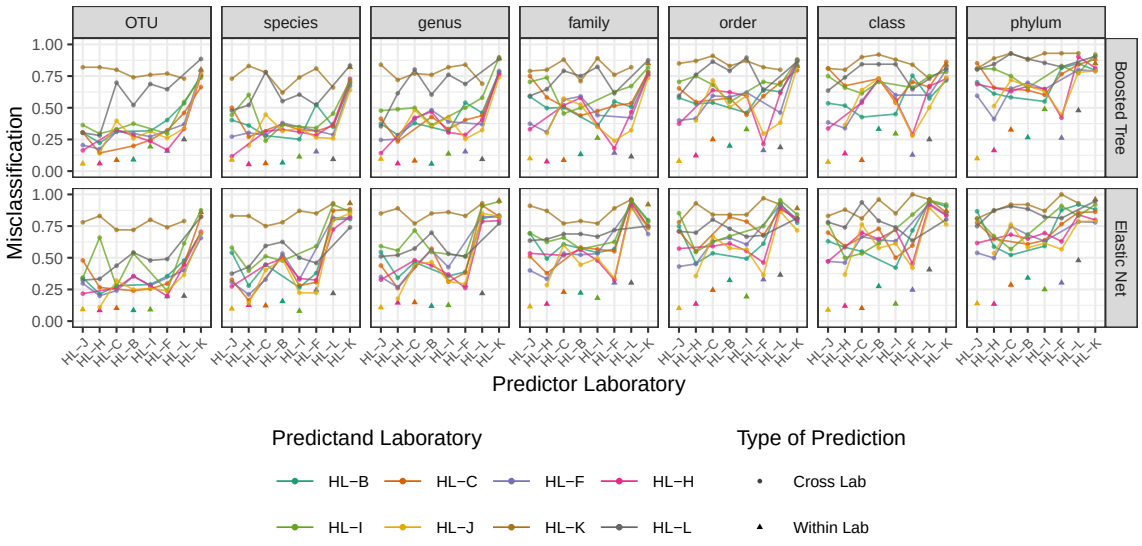

Figure A.19: Within-laboratory (solid triangles) and between-laboratory (lines) specimen misclassification results on uncentered proportion data.

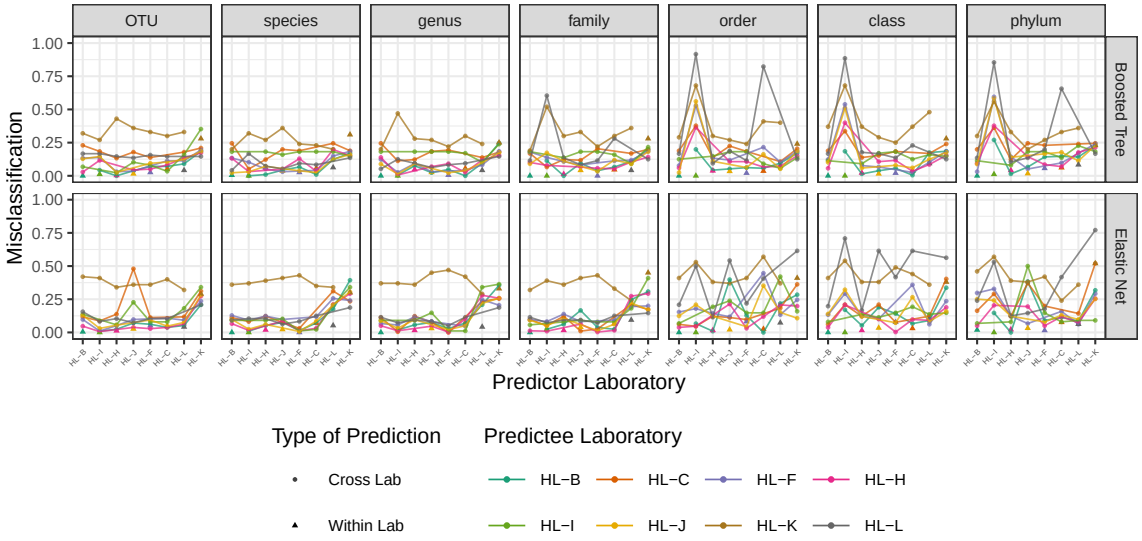

Figure A.20: Within-laboratory (solid triangles) and between-laboratory (lines) specimen type misclassification results on uncentered proportion data.

Results for Centered Proportion-Scale Data

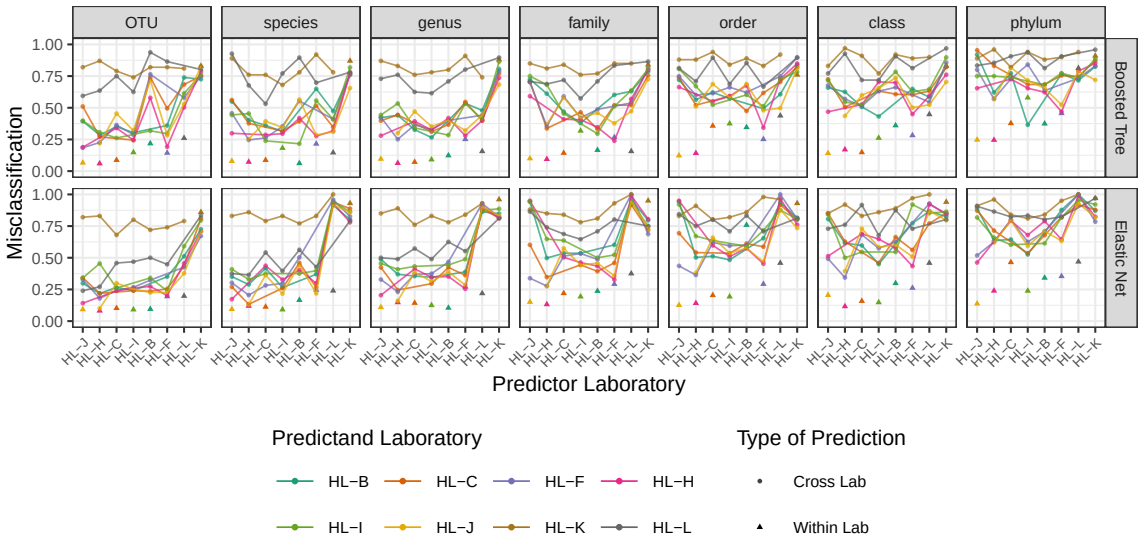

Figure A.21: Within-laboratory (solid triangles) and between-laboratory (lines) specimen misclassification results on centered proportion data.

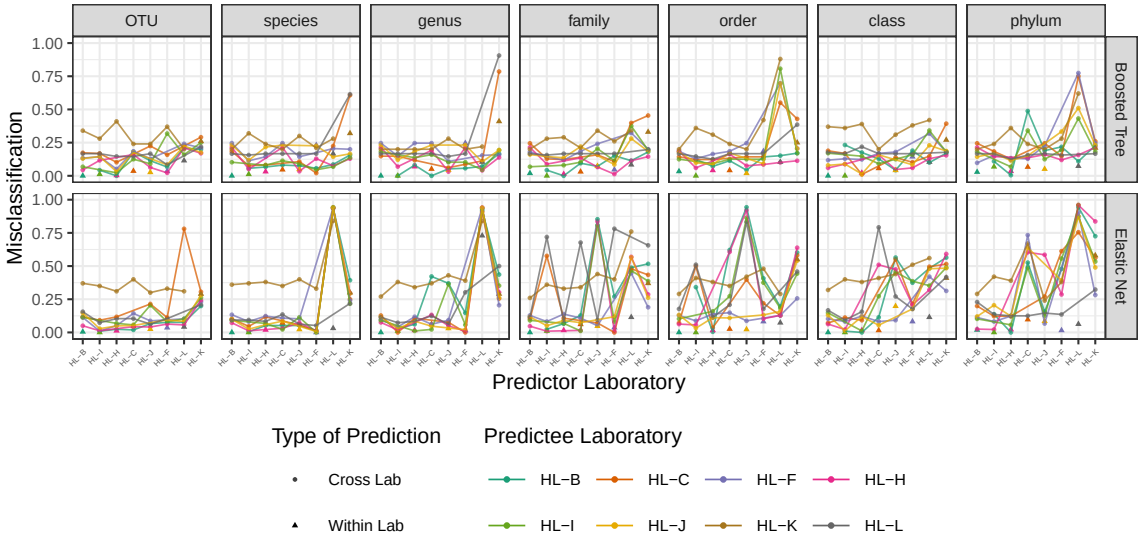

Figure A.22: Within-laboratory (solid triangles) and between-laboratory (lines) specimen type misclassification results on centered proportion data.

Results for Uncentered Log-Ratio Data

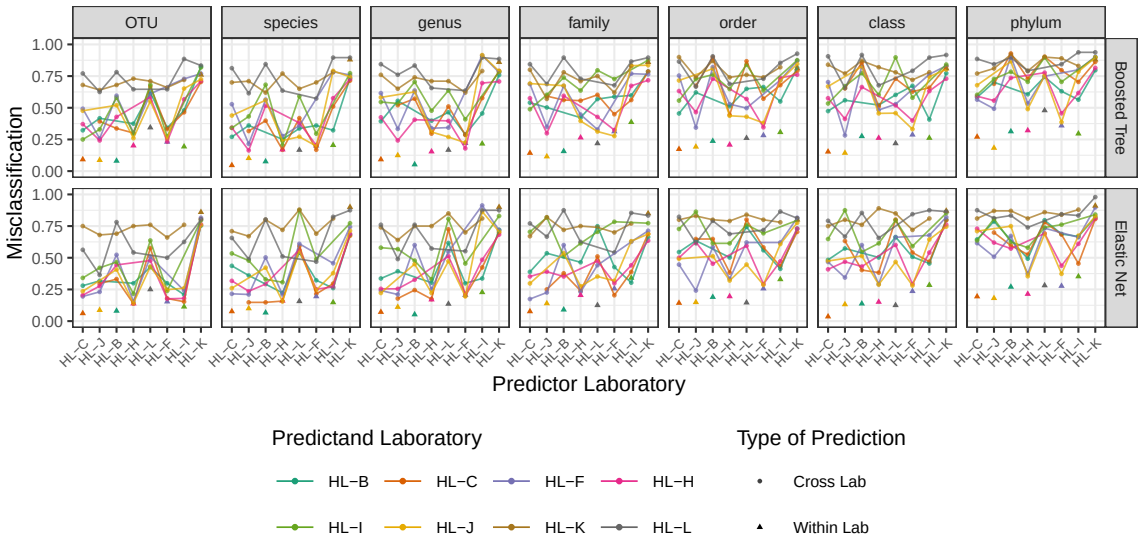

Figure A.23: Within-laboratory (solid triangles) and between-laboratory (lines) specimen misclassification results on log-ratio data (without sample centering).

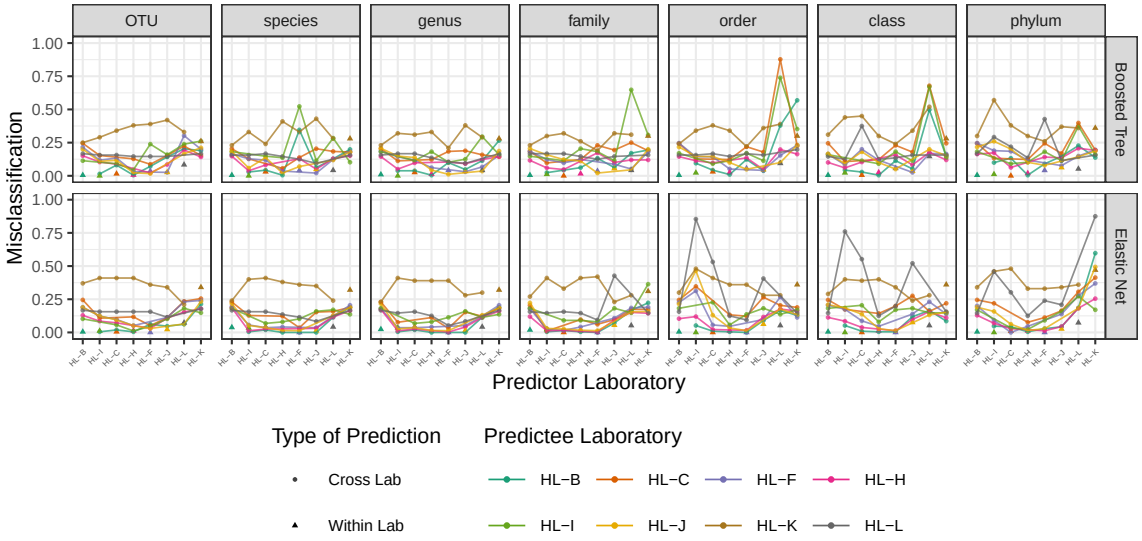

Figure A.24: Within-laboratory (solid triangles) and between-laboratory (lines) specimen type misclassification results on log-ratio data (without sample centering).

Results for Centered Log-Ratio-Scale Data

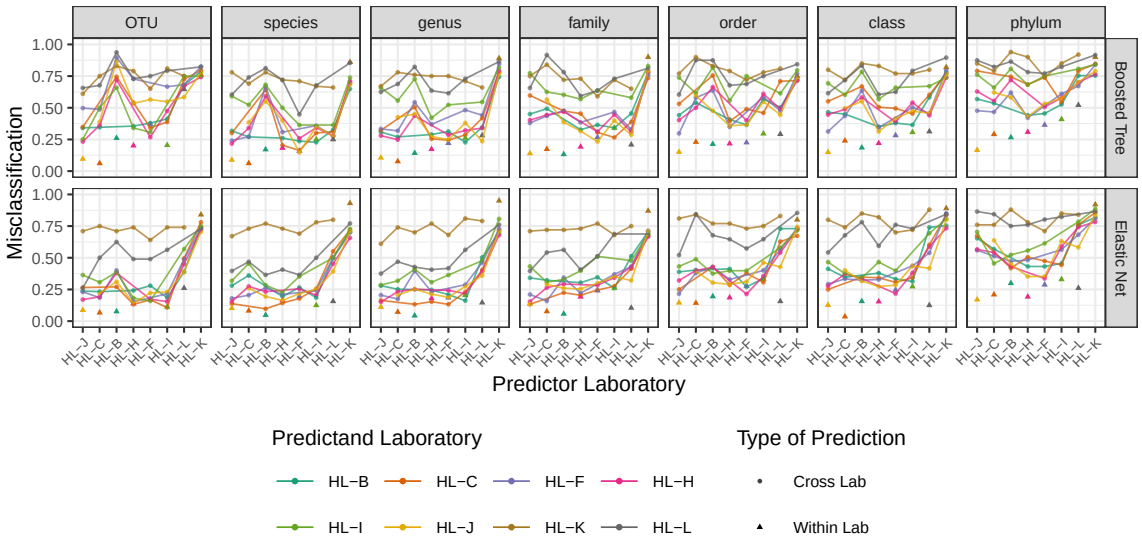

Figure A.25: Within-laboratory (solid triangles) and between-laboratory (lines) specimen misclassification results on log-ratio data (with sample centering).

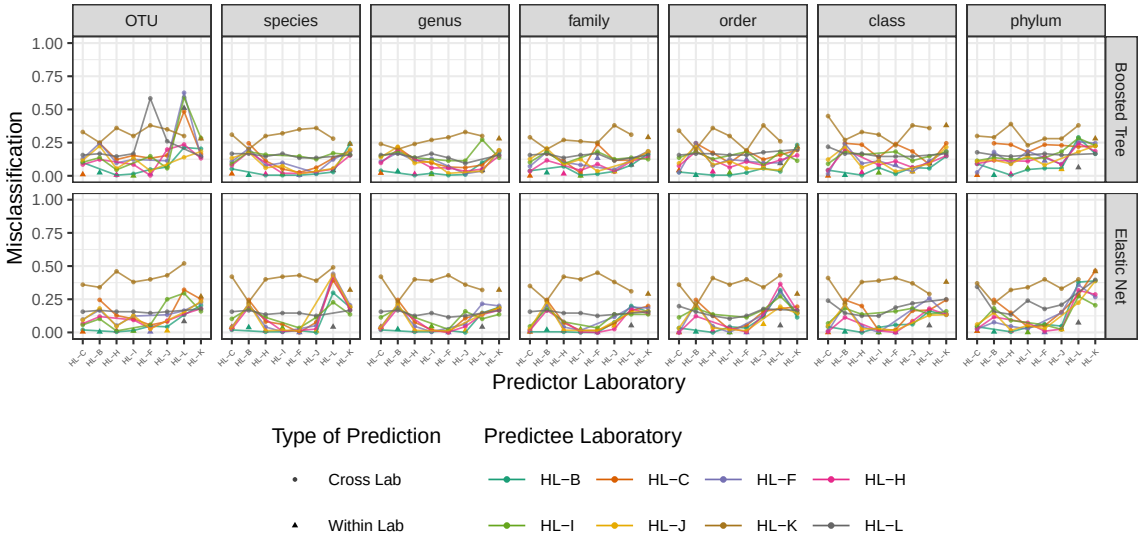

Figure A.26: Within-laboratory (solid triangles) and between-laboratory (lines) specimen type misclassification results on log-ratio data (with sample centering).

Results for Presence-Absence Data

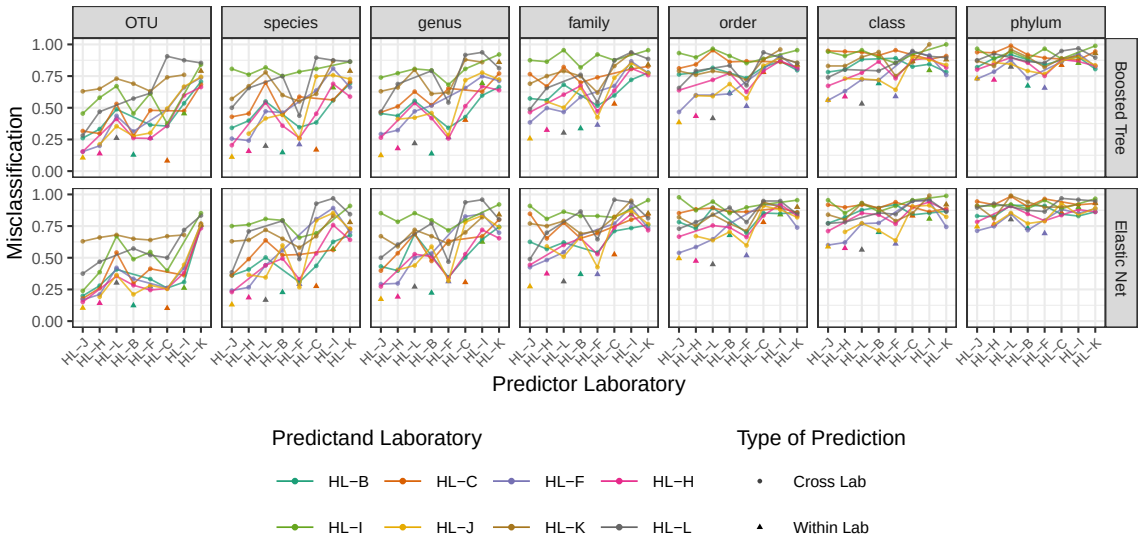

Figure A.27: Within-laboratory (solid triangles) and between-laboratory (lines) specimen misclassification results on presence-absence data (without sample centering).

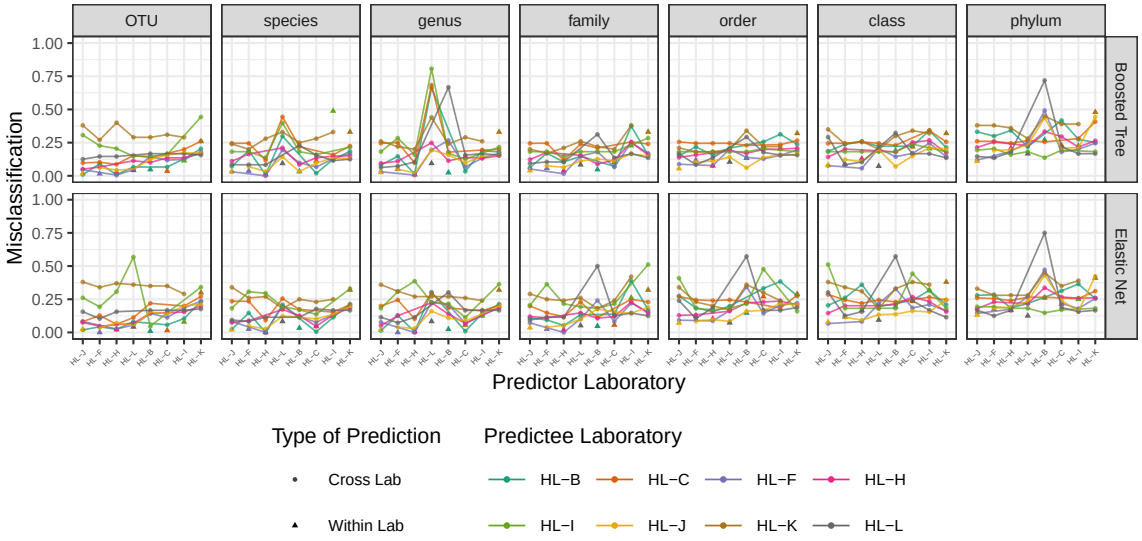

Figure A.28: Within-laboratory (solid triangles) and between-laboratory (lines) specimen type misclassification results on presence-absence data (without sample centering).

Appendix B

MODELING COMPLEX MEASUREMENT ERROR IN MICROBIOME EXPERIMENTS

B.1 Additional details for reweighted estimator

In Section 3, we introduced a weighted Poisson log-likelihood with weight for the likelihood contribution of W_{ij} given by

$$\hat{w}_{ij} = \frac{\hat{\mu}_{ij} + 1}{\hat{\sigma}^2_{ij} + 1}$$

where $\hat{\mu}_{ij}$ is the fitted mean for W_{ij} given parameters $\hat{\theta}$ estimated under a Poisson likelihood and read depth $W_{i\cdot}$. arising from a model fit to $\mathbf{W}$ via a Poisson likelihood (without reweighting) and $\hat{\sigma}^2_{ij}$ is a fitted value from a monotone regression of squared residuals $(W_{ij} - \hat{\mu}_{ij})^2$ on fitted means $\hat{\mu}_{ij}$ (with $i = 1, \ldots, n$ and $j = 1, \ldots, J$). In other words, $\hat{\sigma}^2_{ij}$ is an estimate of $\text{var}(W_{ij}|W_{i\cdot}, \theta)$.

To motivate why this reweighting is reasonable, we consider the case in which θ is in the interior of the parameter space Θ. In this setting we can express the Poisson MLE as a solution to the following score equations:

$$\begin{cases} \sum_{i,j} \frac{1}{\mu_{ij}} [\frac{\partial}{\partial \theta_1} \mu_{ij}](W_{ij} - \mu_{ij}) & = 0 \\ \vdots & \\ \sum_{i,j} \frac{1}{\mu_{ij}} [\frac{\partial}{\partial \theta_L} \mu_{ij}](W_{ij} - \mu_{ij}) & = 0 \end{cases}$$

Equivalently, we can write

$$\sum_{i,j} \frac{1}{\mu_{ij}} \mathbf{g}_{ij}(\theta) = 0$$

letting $\mathbf{g}_{ij} = [\frac{\partial}{\partial \theta^T}\mu_{ij}](W_{ij} - \mu_{ij})$. Hence, we can view this system of equations as a weighted sum of zero expectation terms $\mathbf{g}_{ij}$ with weights given by $\frac{1}{\mu_{ij}}$ — that is, one over a model-based estimate of $\text{Var}(W_{ij} - \mu_{ij})$. In this setting, if the Poisson mean-variance relationship holds and the score equations have a unique solution, we expect the estimator given by this solution to be asymptotically efficient [McCullagh, 1983], whereas when a different mean-variance relationship holds, in general we expect to lose efficiency. In contrast, when the Poisson mean-variance relationship does not hold, we expect to be able to improve efficiency by reweighting the score equations with a more flexible estimator of $\text{Var}(W_{ij} - \mu_{ij})$. To accomplish this, we use a consistent estimator of θ, the Poisson MLE $\hat{\theta}$, to estimate $\boldsymbol{\mu}$ and $\text{Var}(W_{ij}|\mu_{ij})$. Specifically, we estimate $\sigma^2(\hat{\mu}_{ij}) := \text{Var}(W_{ij}|W_{i\cdot}, X_i, Z_i, \tilde{Z}_i, \boldsymbol{\beta}, \mathbf{p}, \tilde{\mathbf{p}}, \tilde{\boldsymbol{\gamma}})$ under the assumption that $\sigma^2(\cdot)$ is an increasing function via a centered isotonic regression of $(W_{ij} - \hat{\mu}_{ij})^2$ on $\hat{\mu}_{ij}$. Weighting the log-likelihood contribution of W_{ij}, $l_{ij} := W_{ij}\log(\mu_{ij}) - \mu_{ij}$, by a factor of $\frac{\hat{\mu}_{ij}}{\hat{\sigma}^2_{ij}}$ then yields reweighted score equations

$$\sum_{i,j} \frac{\hat{\mu}_{ij}}{\hat{\sigma}^2_{ij}} \frac{1}{\mu_{ij}} \mathbf{g}_{ij}(\theta) = \sum_{i,j} \frac{\hat{\mu}_{ij}}{\mu_{ij}} \frac{1}{\hat{\sigma}^2_{ij}} \mathbf{g}_{ij}(\theta)$$

in which each $\mathbf{g}_{ij}$ is, up to a factor of $\frac{\hat{\mu}_{ij}}{\mu_{ij}} \xrightarrow{p} 1$, weighted by the inverse of a flexible estimate of $\text{Var}(W_{ij}|\mu_{ij})$. In practice, however, the weighting above may be unstable when $\hat{\mu}_{ij}$ and $\hat{\sigma}^2_{ij}$ are small. Hence, we weight instead by $\frac{\hat{\mu}_{ij}+1}{\hat{\sigma}^2_{ij}+1}$ to preserve behavior of weights when the estimated mean and variance are both large (where reweighting is typically most important) and stabilizes them when these quantities are small.

B.2 Supporting theory for proposed model and estimators

Throughout this section, we will use the following notation:

- $\mathbf{W}_i = (W_{i1}, \ldots, W_{iJ})$: a measured outcome of interest in sample i across taxa $j = 1, \ldots, J$. We also use $\mathbf{W}$ without subscript i where this does not lead to ambiguity

- $\mathbf{X}_i$ here denotes covariates $(Z_i, X_i, \tilde{Z}_i)$ described in the main text

- $\mathcal{W}$: the support of $\mathbf{W} = (W_1, \ldots, W_J)$
- $\mathcal{X}$: the support of $\mathbf{X} = (X_1, \ldots, X_p)$
- $\mathcal{W}_\Sigma$: the support of $W_\Sigma := \sum_{j=1}^J W_j$
- v: a weighting function from $\mathcal{W}_\Sigma \times \mathcal{X}$ into $\mathbf{R}^J_{>0}$. For simplicity of notation, we frequently suppress dependence on $W_{i\Sigma}$ and $\mathbf{X}_i$ and write v_{ij} to indicate $v_j(W_{i\Sigma}, \mathbf{X}_i)$
- $\hat{v}_n$: an empirical weighting function estimated from a sample of size n
- θ: unknown parameters $(\mathbf{p}, \boldsymbol{\beta}, \tilde{\mathbf{p}}, \tilde{\boldsymbol{\gamma}})$; we denote the true value with θ_0
- $\boldsymbol{\mu}_\theta = (\mu_{\theta 1} \ldots \mu_{\theta J})$: a parametrization of the mean model given in equation (5) in main text; $\mathbb{E}[\mathbf{W}|\mathbf{X}, \gamma, \theta] = \exp(\gamma)\boldsymbol{\mu}_\theta(\mathbf{X})$; when unambiguous, we suppress dependence on $\mathbf{X}$ and write $\boldsymbol{\mu}_\theta$; we also use $\boldsymbol{\mu}_{\theta\Sigma}$ to denote $\sum_{j=1}^J \mu_{\theta j}$
- $M_n^v(\theta)$: profile log likelihood under weighting function v, evaluated at θ on a sample of size n
- $M^v(\theta)$: expected profile log likelihood under weighting function v, evaluated at θ
- m_θ^v: the profile log likelihood under weighting function v as a function from $\mathcal{W} \times \mathcal{X}$ into $\mathbb{R}$; $M_\theta^v = \mathbb{E}_{\mathbf{W},\mathbf{X}} m_\theta^v(\mathbf{W}, \mathbf{X}) := P m_\theta^v$, and similarly we can express $M_n^v(\theta)$ in terms of m_θ^v and the empirical measure $\mathbb{P}_n$: $M_n^v(\theta) = \mathbb{P}_n m_\theta^v$
- $L^\infty(\mathcal{F})$: the set of all uniformly bounded real functions on $\mathcal{F}$

B.2.1 Assumptions

(A) We draw pairs $(\mathbf{W}, \mathbf{X}) \overset{\text{iid}}{\sim} P_{\theta_0}$ where $\mathbf{W}$ has closed, bounded support $\mathcal{W} \subset \mathbb{R}^J_{\geq 0}$ and $\mathbf{X}$ has closed, bounded support $\mathcal{X} \subset \mathbf{R}^p$.

(B) Letting θ denote $(\mathbf{p}, \boldsymbol{\beta}, \tilde{\mathbf{p}}, \tilde{\boldsymbol{\gamma}})$, for a set of known functions from $\mathcal{X}$ to $\mathbb{R}^J_{\geq 0}$ $\{\boldsymbol{\mu}_\theta : \theta \in \Theta\}$ we have that $\mathbb{E}[\mathbf{W}|\mathbf{X}, W_\Sigma] = W_\Sigma \frac{\boldsymbol{\mu}_{\theta_0}(\mathbf{X})}{\boldsymbol{\mu}_{\theta_0 \Sigma}(\mathbf{X})}$ where $\theta_0 \in \Theta \subset \mathbb{R}^d$, with $\mu_\theta(x)$ differentiable in θ for all $x \in \mathcal{X}$ and for each fixed $\theta \in \Theta$, $\mu_\theta(x)$ a bounded function on $\mathcal{X}$.

(C) For almost all $\mathbf{x} \in \mathcal{X}$, $\Pr([\sum_{j=1}^J W_j] > b|\mathbf{X} = x) = 1$ for some $b > 0$.

Note: while the form of the mean model given above differs somewhat from the presentation in the main text, it in fact implies the form in the main text if we introduce random variable Γ and let $\mathbb{E}[W_\Sigma|\mathbf{X} = \mathbf{x}, \Gamma = \gamma] = \exp(\gamma)\boldsymbol{\mu}_{\theta_0 \Sigma}$. However, since this construction in terms of Γ is not necessary for the results that follow, we omit it.

B.2.2 Form of profile log likelihood

We first derive the form of a log likelihood in which nuisance parameters $\{\gamma_i\}_{i=1}^n$, have been profiled out. We characterize population analogue of this log likelihood. The form of this profile log likelihood is as follows:

$$M_n^v(\theta) := \frac{1}{n} \sum_{i=1}^n \sup_{\gamma_i \in \mathbb{R}} \Big[\sum_{j=1}^J v_{ij} \Big(W_{ij} \log[\exp(\gamma_i)\mu_{\theta j}(X_i)] - \exp(\gamma_i)\mu_{\theta j}(X_i) \Big) \Big] \tag{B.1}$$

$$= \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^J \Big[v_{ij} \Big(W_{ij} \log[\frac{\mathbf{v}_i \cdot \mathbf{W}_i}{\mathbf{v}_i \cdot \boldsymbol{\mu}_\theta} \mu_{\theta j}] - \frac{\mathbf{v}_i \cdot \mathbf{W}_i}{\mathbf{v}_i \cdot \boldsymbol{\mu}_\theta} \mu_{\theta j} \Big) \Big] \tag{B.2}$$

where we suppress dependence on $\mathbf{X}_i$ for simplicity in the second row. We derive the profile likelihood in the second row via differentiation with respect to γ_i; the optimum is unique by convexity of $ay - b\exp(y)$ in y when $a, b > 0$. We use $\mathbf{v}_i \cdot \mathbf{W}_i$ to denote $\sum_{j=1}^J v_{ij} W_{ij}$ and similarly for $\mathbf{v}_i \cdot \boldsymbol{\mu}_\theta$.

We now allow weights $\mathbf{v}_i = (v_{i1}, \ldots, v_{iJ})$ to be given as a (bounded positive) function of $\mathbf{X}_i$ and $\mathbf{W}_{i\Sigma} := \sum_{j=1}^J W_{ij}$ and examine the population analogue $M^v(\theta)$ of of the weighted

profile log likelihood $M_n^v(\theta)$.

$$M^v(\theta) = \mathbb{E}_{\mathbf{W},\mathbf{X}}\Big[\sum_{j=1}^{J} v_{ij}\Big(W_{ij}\log[\frac{\mathbf{v}_i\cdot\mathbf{W}_i}{\mathbf{v}_i\cdot\boldsymbol{\mu}_\theta}\mu_{\theta j}] - \frac{\mathbf{v}_i\cdot\mathbf{W}_i}{\mathbf{v}_i\cdot\boldsymbol{\mu}_\theta}\mu_{\theta j}\Big)\Big] \tag{B.3}$$

$$\mathbb{E}_{\mathbf{W}_\Sigma,\mathbf{X}}\mathbb{E}_{\mathbf{W}|\mathbf{W}_\Sigma,\mathbf{X}}\Big[\sum_{j=1}^{J} v_{ij}\Big(W_{ij}\log[\frac{\mathbf{v}_i\cdot\mathbf{W}_i}{\mathbf{v}_i\cdot\boldsymbol{\mu}_\theta}\mu_{\theta j}] - \frac{\mathbf{v}_i\cdot\mathbf{W}_i}{\mathbf{v}_i\cdot\boldsymbol{\mu}_\theta}\mu_{\theta j}\Big)\Big] \tag{B.4}$$

$$= \mathbb{E}_{\mathbf{W},\mathbf{X}}\Big[\sum_{j=1}^{J} v_{ij}W_{ij}\log\mathbf{v}_i\cdot\mathbf{W}_i\Big] \tag{B.5}$$

$$+ \mathbb{E}_{\mathbf{W}_\Sigma,\mathbf{X}}\Big[\sum_{j=1}^{J} v_{ij}\Big(W_{i\Sigma}\frac{\mu_{\theta_0 j}}{\mu_{\theta_0\Sigma}}\log\frac{\mu_{\theta j}}{\mathbf{v}_i\cdot\boldsymbol{\mu}_\theta} - W_{i\Sigma}\frac{\mathbf{v}_i\cdot\boldsymbol{\mu}_{\theta_0}}{\mu_{\theta_0\Sigma}}\frac{\mu_{\theta j}}{\mathbf{v}_i\cdot\boldsymbol{\mu}_\theta}\Big)\Big] \tag{B.6}$$

$$= C + \mathbb{E}_{\mathbf{W}_\Sigma,\mathbf{X}}\Big[W_{i\Sigma}\frac{\mathbf{v}_i\cdot\boldsymbol{\mu}_{\theta_0}}{\mu_{\theta_0\Sigma}}\sum_{j=1}^{J} v_{ij}\Big(\frac{\mu_{\theta_0 j}}{\mathbf{v}_i\cdot\boldsymbol{\mu}_{\theta_0}}\log\frac{\mu_{\theta j}}{\mathbf{v}_i\cdot\boldsymbol{\mu}_\theta} - \frac{\mu_{\theta j}}{\mathbf{v}_i\cdot\boldsymbol{\mu}_\theta}\Big)\Big] \tag{B.7}$$

We note that the term in line 5 above depends on θ_0 but not θ; accordingly, we represent it with constant C on line 7.

B.2.3 Optimizer of profile likelihood

We now show that, under a suitable identifiability condition, a weak condition on $\mathbf{W}$, and a condition on weighting function v, that if the mean model given in assumption B holds at $\theta = \theta_0$, then the unique optimizer of population criterion $M^v(\theta)$ is θ_0.

The additional conditions we need are as follows:

(D) For all $\theta, \theta' \in \Theta$, we have that, for any $a \in \mathbb{R}^+$, $\theta \neq \theta' \Rightarrow \boldsymbol{\mu}_\theta(\mathbf{x}) \neq a\boldsymbol{\mu}_{\theta'}(\mathbf{x})$ holds for all $\mathbf{x} \in A \subset \mathcal{X}$ with $P_{\mathbf{X}}(A) > 0$.

(E) Weighting function $v : \mathcal{X} \times \mathcal{W}_\Sigma \to \mathbb{R}^J_{>0}$, where $\mathcal{W}_\Sigma$ is the support of W_Σ, is continuous and bounded.

We will use the following simple lemma:

Lemma 1. *For every $a \geq 0$, the function defined by $f_a(b) := a\log(b) - b$ is uniquely maximized at $b = a$ (defining $0\log 0 := 0$ and letting $a\log 0 = -\infty$ for every $a > 0$).*

Proof. First consider the case $a > 0$. Since $f_a(0) = -\infty$ in this case and f_a is finite for all $b > 0$, the optimum cannot occur at $b = 0$. Over $b \in \mathbb{R}^+$, $\frac{\partial^2}{\partial b^2} f = -\frac{a}{b^2} < 0$, so f_a is strictly convex over $\mathbb{R}^+$ and hence takes a unique optimum. Setting $\frac{\partial}{\partial b} f = \frac{a}{b} - 1 = 0$ gives us that the optimum occurs at $b = a$.

When $a = 0$, $f_a(b) = \begin{cases} 0 & \text{if } b = 0 \\ -b & \text{if } b > 0 \end{cases}$, so f_a is optimized at 0 since $-b < 0$ when $b > 0$. □

Theorem 2. *Suppose that conditions (A) - (D) are met. Then for any weighting function satisfying (E), the criterion $M^v(\cdot)$ defined above is uniquely optimized at θ_0.*

Proof. From above we have the form of the population criterion $M^v(\theta)$:

$$M^v(\theta) = C + \mathbb{E}_{\mathbf{W}_\Sigma, \mathbf{X}}\Big[W_\Sigma \frac{\mathbf{v} \cdot \boldsymbol{\mu}_{\theta_0}}{\mu_{\theta_0 \Sigma}} \sum_{j=1}^{J} v_{ij}\Big(\frac{\mu_{\theta_0 j}}{\mathbf{v}_i \cdot \boldsymbol{\mu}_{\theta_0}} \log \frac{\mu_{\theta j}}{\mathbf{v}_i \cdot \boldsymbol{\mu}_{\theta}} - \frac{\mu_{\theta j}}{\mathbf{v}_i \cdot \boldsymbol{\mu}_{\theta}}\Big)\Big] \tag{B.8}$$

For each fixed pair $(\mathbf{x}, w_\Sigma) \in \text{supp}(\mathbf{X}, W_\Sigma)$, denote by $\frac{\mu_{\theta_0 j}}{\mathbf{v} \cdot \boldsymbol{\mu}_{\theta_0}}(\mathbf{x}, w_\Sigma)$ the function $\frac{\mu_{\theta_0 j}(\mathbf{x})}{\mathbf{v}(\mathbf{x}, w_\Sigma) \cdot \boldsymbol{\mu}_{\theta_0}(\mathbf{x})}$. Then

$$h_j^v(\mathbf{x}, w_\Sigma, \theta; \theta_0) := \frac{\mu_{\theta_0 j}}{\mathbf{v} \cdot \boldsymbol{\mu}_{\theta_0}}(\mathbf{x}, w_\Sigma) \log \frac{\mu_{\theta j}}{\mathbf{v} \cdot \boldsymbol{\mu}_{\theta}}(\mathbf{x}, w_\Sigma) - \frac{\mu_{\theta j}}{\mathbf{v} \cdot \boldsymbol{\mu}_{\theta}}(\mathbf{x}, w_\Sigma) \tag{B.9}$$

is maximized when $\frac{\mu_{\theta j}}{\mathbf{v}_i \cdot \boldsymbol{\mu}_{\theta}}(\mathbf{x}, w_\Sigma) = \frac{\mu_{\theta_0 j}}{\mathbf{v}_i \cdot \boldsymbol{\mu}_{\theta_0}}(\mathbf{x}, w_\Sigma)$ by lemma 1. □

Before proceeding, we show that $-\infty < M^v(\theta_0) < \infty$. By definition, we have

$$M^v(\theta_0) = \mathbb{E}_{\mathbf{W}, \mathbf{X}} \text{sup}_{\gamma \in \mathbb{R}} \sum_{j=1}^{J} v_j \big(W_j \log\big[\exp(\gamma) \frac{\mathbf{v} \cdot \mathbf{W}}{\mathbf{v} \cdot \boldsymbol{\mu}_{\theta_0}} \mu_{\theta_0 j}\big] - \exp(\gamma) \frac{\mathbf{v} \cdot \mathbf{W}}{\mathbf{v} \cdot \boldsymbol{\mu}_{\theta_0}} \mu_{\theta_0 j}\big) \tag{B.10}$$

$$\geq \mathbb{E}_{\mathbf{W}, \mathbf{X}} \sum_{j=1}^{J} v_j \big(W_j \log\big[\frac{\mathbf{v} \cdot \mathbf{W}}{\mathbf{v} \cdot \boldsymbol{\mu}_{\theta_0}} \mu_{\theta_0 j}\big] - \frac{\mathbf{v} \cdot \mathbf{W}}{\mathbf{v} \cdot \boldsymbol{\mu}_{\theta_0}} \mu_{\theta_0 j}\big) \text{ (setting } \gamma = 0) \tag{B.11}$$

$$= \mathbb{E}_{\mathbf{W}, \mathbf{X}} \sum_{j=1}^{J} v_j W_j \log \mathbf{v} \cdot \mathbf{W} \tag{B.12}$$

$$+ \mathbb{E}_{W_\Sigma, \mathbf{X}} \frac{\mathbf{v} \cdot \boldsymbol{\mu}_{\theta_0}}{\mu_{\theta_0 \Sigma}} W_\Sigma \sum_{j=1}^{J} \big[\frac{\mu_{\theta_0 j}}{\mathbf{v} \cdot \boldsymbol{\mu}_{\theta_0}} \log \frac{\mu_{\theta_0 j}}{\mathbf{v} \cdot \boldsymbol{\mu}_{\theta_0}} - \frac{\mu_{\theta_0 j}}{\mathbf{v} \cdot \boldsymbol{\mu}_{\theta_0}}\big] \tag{B.13}$$

The term in line (B.12) is equal to $\mathbb{E}_{\mathbf{W},\mathbf{X}}\mathbf{v}\cdot\mathbf{W}\log\mathbf{v}\cdot\mathbf{W}$, which as the integral of a bounded function over a bounded domain is finite. By assumption (C), we must have $\sum_j \mu_{\theta_0 j}(\mathbf{X}) > 0$ almost surely, so the term $\frac{\mathbf{v}\cdot\boldsymbol{\mu}_{\theta_0}}{\mu_{\theta_0\Sigma}}W_\Sigma$ in line (B.13) is almost surely bounded by boundedness of μ_θ, v, and $\mathbf{W}$. Inside the sum in this line, we have terms of the form $a\log a - a$, which is a bounded function on any bounded set in $\mathbb{R}_{\geq 0}$. Hence line (B.13) is an integral of a bounded function over a bounded domain and so is also finite, so $M^v(\theta_0) > -\infty$.

Similarly,

$$M^v(\theta_0) = \mathbb{E}_{\mathbf{W},\mathbf{X}}\sup_{\gamma\in\mathbb{R}}\sum_{j=1}^{J} v_j\big(W_j\log\big[\exp(\gamma)\frac{\mathbf{v}\cdot\mathbf{W}}{\mathbf{v}\cdot\boldsymbol{\mu}_{\theta_0}}\mu_{\theta_0 j}\big] - \exp(\gamma)\frac{\mathbf{v}\cdot\mathbf{W}}{\mathbf{v}\cdot\boldsymbol{\mu}_{\theta_0}}\mu_{\theta_0 j}\big) \tag{B.14}$$

$$\leq \mathbb{E}_{\mathbf{W},\mathbf{X}}\sum_{j=1}^{J} v_j\big(W_j\log W_j - W_j\big) \text{ by lemma 1} \tag{B.15}$$

$$< \infty \text{ since } \sum_{j=1}^{J} v_j\big(W_j\log W_j - W_j\big) \text{ is bounded on } \mathcal{W}\times\mathcal{X}\times\mathcal{G} \tag{B.16}$$

Hence $-\infty < M^v(\theta_0) < \infty$, which guarantees that the difference in the following argument is not of the form $\infty - \infty$.

Now, for any $\theta \in \Theta$ with $\theta \neq \theta_0$, we have

$$\begin{aligned}
&M^v(\theta) - M^v(\theta_0)\\
&= \int_A \int w\frac{\mathbf{v}\cdot\boldsymbol{\mu}_{\theta_0}}{\mu_{\theta_0\Sigma}}\sum_{j=1}^{J}\Big[h_j^v(\mathbf{x},w_\Sigma,\theta;\theta_0) - h_j^v(\mathbf{x},w_\Sigma,\theta_0;\theta_0)\Big]dP_{W_\Sigma|\mathbf{X}}(w_\Sigma)dP_{\mathbf{X}}(\mathbf{x})\\
&+ \int_{A^C} \int w\frac{\mathbf{v}\cdot\boldsymbol{\mu}_{\theta_0}}{\mu_{\theta_0\Sigma}}\sum_{j=1}^{J}\Big[h_j^v(\mathbf{x},w_\Sigma,\theta;\theta_0) - h_j^v(\mathbf{x},w_\Sigma,\theta_0;\theta_0)\Big]dP_{W_\Sigma|\mathbf{X}}(w_\Sigma)dP_{\mathbf{X}}(\mathbf{x})\\
&\leq \int_A \int w\frac{\mathbf{v}\cdot\boldsymbol{\mu}_{\theta_0}}{\mu_{\theta_0\Sigma}}\sum_{j=1}^{J}\Big[h_j^v(\mathbf{x},w_\Sigma,\theta;\theta_0) - h_j^v(\mathbf{x},w_\Sigma,\theta_0;\theta_0)\Big]dP_{W_\Sigma|\mathbf{X}}(w_\Sigma)dP_{\mathbf{X}}(\mathbf{x}) \quad (\star)\\
&< 0
\end{aligned}$$

The first inequality is a result of θ_0 maximizing (not necessarily uniquely) h_j^v; i.e., $h_j^v(\mathbf{x},\theta;\theta_0) - h_j^v(\mathbf{x},\theta_0;\theta_0) \leq 0$. Strict inequality holds in the last line because the integrand in $(\star)$ is strictly negative, since $h_j^v(\mathbf{x},w_\Sigma,\theta;\theta_0) - h_j^v(\mathbf{x},w_\Sigma,\theta_0;\theta_0) < 0$ for $\theta \neq \theta_0$ on A, and the term $w\frac{\mathbf{v}\cdot\boldsymbol{\mu}_{\theta_0}}{\mu_{\theta_0\Sigma}}$ is

a.s. strictly positive by assumption (C) and positivity of $\mathbf{v}$. Hence $M^v(\theta_0) > M^v(\theta)$ for all $\theta \neq \theta_0$ in Θ, so θ_0 is the unique maximizer of the population criterion $M^v(\cdot)$.

B.2.4 Consistency of M-estimators

We apply theorem 5.14 of van der Vaart (1998) to show consistency of maximizers $\hat{\theta}_n^v$ of M_n^v for θ_0. We also show consistency of estimators $\hat{\theta}_n^{\hat{v}_n}$, where $\{\hat{v}_n\}$ is a sequence of random continuous positive bounded weighting functions converging uniformly in probability to v.

We first require the following assumption on v,$\hat{v}_n$, and $\boldsymbol{\mu}$:

(F) $\sup_{t \in \mathcal{X} \times \mathcal{W}_\Sigma} |v(t) - \hat{v}_n(t)| \xrightarrow{p} 0$ and every $\hat{v}_n$ is continuous, positive, and bounded.

(G) There exist $\delta > 0$ and $\epsilon > 0$ such that $\min_j \inf_{\mathbf{x} \in \mathcal{X}: \mu_{\theta_0 j}(x) > 0} \frac{\mu_{\theta j}(x)}{\mu_{\theta \Sigma}(x)} \geq \epsilon$ holds for all $\theta \in H_\delta := K_\delta \cap \Theta$ where $K_\delta := \{\theta : d(\theta, \theta_0) \leq \delta\}$ is a closed δ-neighborhood of θ_0 in $\mathbb{R}^d$. Moreover, $\mu_\theta(w, x)$ is Lipschitz continuous in θ on H_δ uniformly over $\text{supp}(W, X)$.

Note: while assumption (G) can likely be loosened, we note that in practice it is not particularly restrictive. In particular, we emphasize that it in no way precludes θ_0 from lying at the boundary of the parameter space; rather, it guarantees that nonzero means are bounded away from zero, which allows us to select neighborhoods of θ_0 in which m_θ^v is bounded.

Theorem 3. *Suppose conditions (A) through (F) are satisfied. Then for* $d(\theta, \theta') = \sum_{k=1}^p |arctan(\theta_k) - arctan(\theta'_k)|$, $Pr(d(\hat{\theta}^v, \theta_0) > \epsilon) \to 0$ *for all* $\epsilon > 0$ *and* $Pr(d(\hat{\theta}^{\hat{v}_n}, \theta_0) > \epsilon) \to 0$ *for all* $\epsilon > 0$.

Proof

We first compactify our parameter space Θ to obtain $\bar{\Theta}$ by allowing elements of unconstrained Euclidean parameters to take values in the extended reals.

A necessary condition for theorem 5.14 is that we have $\mathbb{E}\sup_{\theta \in U} m_\theta^v < \infty$ for $m_\theta^v(\mathbf{w}, \mathbf{x}) := \sum_{j=1}^J v_j \big(w_j \log \big[\frac{\mathbf{v} \cdot \mathbf{w}}{\mathbf{v} \cdot \mu_\theta} \mu_{\theta_j} \big] - \frac{\mathbf{v} \cdot \mathbf{w}}{\mathbf{v} \cdot \mu_\theta} \mu_{\theta_j} \big)$ and a sufficiently small ball $U \in \Theta$. By lemma 1, $\sup_{\theta \in U} m^v(\mathbf{w}, \mathbf{x}) \leq \sum_{j=1}^J v_j \big(w_j \log w_j - w_j \big)$, which is bounded above since v is bounded. Hence by assumption

(A), $\mathbb{E}\sup_{\theta\in U} m_\theta^v \leq P\sum_{j=1}^J v_j\big(w_j \log w_j - w_j\big) < \infty$. We also require $M_n^v(\hat{\theta}^v) \geq M_n^v(\boldsymbol{\theta_0}) + o_P(1)$ which is trivially satisfied since $\hat{\theta}^v$ maximizes M_n^v.

Then letting compact set $K = \bar{\Theta}$, we can directly apply theorem 5.14 to obtain $Pr(d(\hat{\theta}^v, \theta_0) \geq \epsilon) \to 0$ for any $\epsilon > 0$.

To apply theorem 5.14 to $\hat{\theta}^{\hat{v}_n}$, we only need in addition to the above that $M_n^v(\hat{\theta}^{\hat{v}_n}) \geq M_n^v(\theta_0) + o_P(1)$.

For any fixed $\hat{v}$, we have

$$M_n^{\hat{v}}(\hat{\theta}_n^{\hat{v}}) \geq M_n^{\hat{v}}(\theta_0) + o_P(1) \tag{B.17}$$

$$= M_n^v(\theta_0) + o_P(1) + (M_n^v(\theta_0) - M_n^{\hat{v}}(\theta_0)) \tag{B.18}$$

However, the term $(M_n^v(\theta_0) - M_n^{\hat{v}}(\theta_0)) = o_P(1)$ if we let $\hat{v} = \hat{v}_n$ since, letting $l_{ij}(\theta) := W_j \log\frac{\mu_{\theta j}}{\mu_{\theta\Sigma}} - \frac{\mu_{\theta j}}{\mu_{\theta\Sigma}}$,

$$|M_n^{\hat{v}_n}(\theta_0) - M_n^v(\theta_0)| = \Big|\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^J(\hat{v}_{n,ij} - v_i j) l_{ij}(\theta_0)\Big| \tag{B.19}$$

$$\leq \sup_{t\in\mathcal{X}\times\mathcal{W}_\Sigma}|\hat{v}_n(t) - v(t)|\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^J |l_{ij}(\theta_0)| \xrightarrow{p} 0 \tag{B.20}$$

since $\sup_{t\in\mathcal{X}\times\mathcal{W}_\Sigma}|\hat{v}_n(t) - v(t)| \xrightarrow{p} 0$. Hence $M_n^v(\hat{\theta}^{\hat{v}_n}) \geq M_n^v(\theta_0) + o_P(1)$, so $Pr(d(\hat{\theta}^{\hat{v}_n}, \theta_0) \geq \epsilon) \to 0$ for any $\epsilon > 0$.

B.2.5 Convergence in distribution of M-estimators

Theorem 4. *If assumptions A - G are met, $\sqrt{n}(\hat{\theta}^v - \theta_0)$ converges in distribution to a tight limiting distribution in $\mathbb{R}^d$.*

Proof. Without loss of generality, we consider $\mathcal{M}_\delta^v := \{m_\theta^v : \theta \in H_\delta\}$ for H_δ given in assumption G (since by theorem 2, with probability approaching 1, $\hat{\theta}_n^v \in H_\delta$ as n increases without bound).

We begin by considering criterion m_θ without weighting.

Fix $(w, x) \in \text{supp}(W, X)$. We then have, for any $\theta \in H_\delta$,

$$m_\theta(w,x) = \sum_{j=1}^{J} w_j \log \frac{\mu_{\theta j}(x)}{\mu_{\theta\Sigma}(x)} - \frac{\mu_{\theta j}(x)}{\mu_{\theta\Sigma}(x)} \tag{B.21}$$

By Lipschitz continuity of $\mu_\theta(x)$ in θ over H_δ, $m_\theta(w,x)$ is a Lipschitz continuous function at all θ such that all elements of $\mu_\theta(x)$ are nonzero since $g(z) := a\log(z) - z$ is Lipschitz continuous for $z \geq \epsilon > 0$ and bounded $a \geq 0$. When one or more elements of $\mu_\theta(w,x)$ are zero, they must be 0 identically over all $\theta \in H_\delta$; otherwise by continuity of $\mu_\theta(w,x)$ in θ, we are able to select $\theta^\star \in H_\delta$ such that $\mu_{\theta^\star j}(w,x)$ is positive but arbitrarily close to zero, contradicting assumption G.

Since if $\mathbb{E}W_j = 0$, we must have $W_j = 0$ with probability 1 (on account of W_j being a nonnegative random variable), except possibly on a set with measure 0, we have

$$m_\theta(w,x) = \sum_{j=1}^{J} \mathbf{1}_{\mu_{\theta_0 j}(x)>0}\big[w_j \log \frac{\mu_{\theta j}(x)}{\mu_{\theta\Sigma}(x)}\big] - \frac{\mu_{\theta j}(x)}{\mu_{\theta\Sigma}(x)}\big] \tag{B.22}$$

Accordingly, $m_\theta(w,x)$ is Lipschitz continuous in θ for all $\theta \in H_\delta$. Hence $m^v_\theta(w,x)$, as a weighted sum of Lipschitz continuous functions (with bounded weights), must also be Lipschitz continuous θ for all $\theta \in H_\delta$ as well.

We now use a bracketing argument to show that $\mathcal{M}_\delta$ is a Donsker class, which we will use together with a result due to Dümbgen [1993] to show that $\sqrt{n}(\hat{\theta}^v_n - \theta_0)$ converges to a well-defined limiting distribution in $\mathbb{R}^d$.

By uniform Lipschitz continuity of $m^v_\theta(w,x)$ in θ on H_δ, we have

$$|m^v_\theta(w,x) - m^v_{\theta'}(w,x)| \leq C||\theta - \theta'|| \tag{B.23}$$

for some $C < \infty$.

Since by assumption $\text{supp}(W,X)$ is bounded, this implies that $|C|^2$ is integrable, which is sufficient for the bracketing entropy of M_δ to be at most of order $\log(1/\epsilon)$ (see Van der Vaart [2000] example 19.7). Hence by theorem 19.5 of Van der Vaart [2000], $\mathcal{M}_\delta$ is Donsker.

Accordingly, we have $\{\sqrt{n}(\mathbb{P}_n m^v_\theta - P m^v_\theta) : m_\theta \in M_\delta\}$ weakly converging to a tight Gaussian process in $l^\infty(\mathcal{M}_\delta)$ as $n \to \infty$. By proposition 1 of Dümbgen [1993], this, taken

with Hadamard directional differentiability of the map defined by $g(F) := \arg\max_{\theta \in H_\delta} F(\theta)$, gives us

$$\sqrt{n}(\hat{\theta}_n^v - \theta_0) := \sqrt{n}(g(\mathbb{P}_n m_\theta^v) - g(P m_\theta^v)) \rightsquigarrow \mathcal{J} \tag{B.24}$$

for well-defined limiting distribution J on $\mathbb{R}^d$.

□

Theorem 5. *$\sqrt{n}(\hat{\theta}^{\hat{v}_n} - \theta_0)$ converges in distribution to the same limit as $\sqrt{n}(\hat{\theta}^v - \theta_0)$ if assumptions A - G are met.*

Proof.

$$\sqrt{n}(M_n^{\hat{v}_n}(\theta) - M^{\hat{v}_n}(\theta)) \tag{B.25}$$

$$= \sqrt{n}(M_n^v(\theta) - M^v(\theta) + \left[[M_n^{\hat{v}_n}(\theta) - M_n^v(\theta)] - [M^{\hat{v}_n}(\theta) - M^v(\theta)]\right]) \tag{B.26}$$

$$= \sqrt{n}(M_n^v(\theta) - M^v(\theta)) + \sqrt{n}(\mathbb{P}_n - P) m_\theta^{\hat{v}_n - v} \tag{B.27}$$

$$= \sqrt{n}(M_n^v(\theta) - M^v(\theta)) + o_P(1) \tag{B.28}$$

since

$$\sqrt{n}(\mathbb{P}_n - P) m_\theta^{\hat{v}_n - v} = \sqrt{n} \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{J} (\hat{v}_{n;ij} - v_{ij}) l_{ij}(\theta) \tag{B.29}$$

$$\leq \sqrt{n} \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{J} |l_{ij}(\theta)| \sup_{t \in \text{supp}(\mathbf{X}, W_\Sigma)} |\hat{v}_n(t) - v(t)| \tag{B.30}$$

$$\xrightarrow{p} 0 \tag{B.31}$$

Note that $\arg\max_{\theta \in \Theta} M^{\hat{v}_n}(\theta) = \arg\max_{\theta \in \Theta} M^v(\theta) = \theta_0$ by theorem 1. □

B.3 Optimization details

B.3.1 Reparametrization of barrier subproblem

Letting $\theta^\star$ indicate the unknown parameters in our model after reparametrizing $\mathbf{p}$ and $\tilde{\mathbf{p}}$ as $\boldsymbol{\rho}$ and $\tilde{\boldsymbol{\rho}}$, we now have the following unconstrained minimization problem:

$$\arg\min_{\theta^\star} f_n(\theta^\star) + \frac{1}{t^{(r)}}\Big[\sum_k \left(\sum_{j=1}^{J-1} -\rho_{kj} + J\log\left[1+\sum_{j=1}^{J-1}\exp(\rho_{kj})\right]\right) \tag{B.32}$$

$$\sum_{\tilde{k}} \left(\sum_{j=1}^{J-1} -\tilde{\rho}_{\tilde{k}j} + J\log\left[1+\sum_{j=1}^{J-1}\exp(\tilde{\rho}_{\tilde{k}j})\right]\right)\Big] \tag{B.33}$$

B.3.2 Barrier algorithm

Barrier Algorithm

1. Initiate with value of penalty parameter t set to starting value $t^{(0)}$ and values of parameters $\theta^\star$ equal to $\theta^{\star(0)}$. Set iteration $r = 0$.

2. Using current value $t^{(r)}$ of t and starting at parameter estimate $\theta^{\star(r)}$, solve barrier subproblem r given in main text via Fisher scoring. Denote the solution of this subproblem $\theta^\star_{(r+1)}$ and set $t_{(r+1)} = at_{(r)}$ for a prespecified $a > 1$.

3. If $t_{(r+1)} > t_{\max}$ for prespecified $t_{\max}$, return $\theta^\star_{(r+1)}$. Otherwise set iteration $r = r + 1$ and return to step 2.

B.3.3 Constrained Newton within Augmented Lagrangian Algorithm

We calculate update steps from $\mathcal{L}_k$ given in Section 3 of the main text as follows:

Constrained Newton within Augmented Lagrangian Algorithm

1. Initiate with initial values $\nu^{(0)}$ and $\mu^{(0)} > 0$ of penalty coefficients ν and μ

2. Calculate proposed update $\mathbf{p}_k^{\text{update}}$ via nonnegative least squares on $\mathcal{L}_k$ using current values of ν and μ

3. If $|\sum_j^J p_{kj}^{\text{update}} - 1| < \delta$ for some prespecified tolerance $\delta > 0$, set update direction $\mathbf{s}_k := \mathbf{p}_k - \mathbf{p}_k^{\text{update}}$ and proceed to step (3). Otherwise update ν and μ via algorithm given in Bazaraa [2006] (p. 496) and return to step (1).

4. Perform a line search in direction $\mathbf{s}_k$ to determine updated parameter value $\mathbf{p}_k^{\text{updated}} := \mathbf{p}_k + \epsilon \mathbf{s}_k$ that decreases objective $f_n(\mathbf{p}_k)$ for some $0 < \epsilon \leq 1$.

B.3.4 Quadratic approximation to $f_n(\mathbf{p}_k)$.

In Section 3 of the main text, we specify $\mathcal{L}_k$ in terms of a quadratic approximation $Q_k^{(t)}$ to objective $f_n(\mathbf{p}_k)$. In practice we construct $Q_k^{(t)}$ as a slightly modified Taylor expansion of $f_n()$ around the current value of $\mathbf{p}_k^{(t)}$. We use the gradient of f_n with respect to $\mathbf{p}_k$ in the first order term, and in the second order term, and in place of the Hessian, we use (-1 times) the Fisher information matrix in $\mathbf{p}_k$ regularized (for numerical stability) by addition of magnitude of the gradient times an identity matrix.

B.4 Analysis of Costea et al. [2017] data

B.4.1 Details of model specification

Costea et al. [2017] published two flow cytometric readings for every species in the synthetic community with the exception of *V. cholerae*, for which only one reading was published. In all taxa save *V. cholerae*, we take the mean reading as our observation, and we include the resulting vector of readings augmented by the single reading for *V. cholerae* as a row in $\mathbf{W}$.

We anticipate that our use of mean readings represents a fairly small loss of information, as flow cytometric readings did not vary substantially within taxon. However, in a similar setting where multiple sets of flow cytometric readings across all taxa were available, we could include each set as a row of $\mathbf{W}$ to capture variability in these measurements.

To estimate detection effects relative to flow cytometry measurements, we specify $X_1 = \mathbf{0}$. For $i \geq 2$, $\mathbf{X}_{i\cdot} = [1\ \mathbf{1}_Q\ \mathbf{1}_W]$ where $\mathbf{1}_Q$ is an indicator for sample i being processed according to protocol Q, and similarly for $\mathbf{1}_W$.

B.4.2 Cross-validation design

We construct folds for our 10-fold cross-validation on Costea et al. [2017] data so that, with the exception of samples A and B, which we grouped together in a single fold, each fold included all observations for a given specimen. For each fold, we fit a model in which all observations in all other folds, along with flow cytometry readings, were treated as arising from a common specimen (as in fact they do, save for flow cytometry readings, which were taken on specimens mixed to create the mock spike-in). We model each sample in the held-out fold as arising from a distinct specimen of unknown composition to allow our model to estimate a different relative abundance profile for distinct samples processed according to different protocols.

B.4.3 Model summaries

Table B.1: Point estimates and 95% bootstrap confidence intervals for protocol-specific detection effects $\boldsymbol{\beta}$ (with reference taxon Y. pseudotuberculosis) estimated from Costea et al. [2017] data

Taxon	Protocol H	Protocol Q	Protocol W
B. hansenii	-1.61 (-2.00 – -1.16)	-1.55 (-1.75 – -1.31)	-0.08 (-0.16 – 0.00)
C. difficile	-0.18 (-0.30 – 0.01)	-0.57 (-0.79 – -0.41)	1.23 (1.18 – 1.28)
C. perfringens	3.38 (3.27 – 3.57)	2.48 (2.31 – 2.62)	4.05 (4.03 – 4.07)
C. saccharolyticum	-0.19 (-0.23 – -0.16)	-0.01 (-0.12 – 0.1)	-0.10 (-0.13 – -0.06)
F. nucleatum	2.37 (2.28 – 2.44)	0.14 (-0.16 – 0.42)	2.11 (2.05 – 2.16)
L. plantarum	-2.62 (-2.96 – -2.12)	0.72 (0.60 – 0.93)	0.60 (0.56 – 0.63)
P. melaninogenica	4.17 (4.12 – 4.2)	3.88 (3.82 – 4.04)	4.25 (4.23 – 4.27)
S. enterica	2.49 (2.45 – 2.51)	2.74 (2.64 – 2.79)	2.48 (2.46 – 2.51)
V. cholerae	1.54 (1.50 – 1.56)	0.90 (0.78 – 0.99)	1.48 (1.44 – 1.50)

Table B.1 provides point estimates and marginal 95% confidence intervals for the detection effects for each of protocols H, Q, and W estimated via the full model described above. This model was fit with reference taxon *Y. pseudotuberculosis* (i.e., under the constraint that the column of $\boldsymbol{\beta}$ corresponding to this taxon consists of 0 entries). Hence we interpret estimates in this table in terms of degree of over- or under-detection relative to *Y. pseudotuberculosis* – for example, we estimate that, repeated measurement under protocol H of samples consisting of 1:1 mixtures of *B. hansenii* and *Y. pseudotuberculosis*, the mean MetaPhlAn2 estimate of the relative abundance of B. hansenii will be $\exp(-1.61) \approx 0.20$ as large as the mean estimate of the relative abundance of *Y. pseudotuberculosis*.

B.5 Analysis of Karstens et al. [2019] data

B.5.1 Preprocessing

We process raw read data reported by Karstens et al. [2019] using the DADA2 R package (version 1.20.0) [Callahan et al., 2016]. We infer amplicon sequence variants using the *dada* function with option 'pooled = TRUE' and assign taxonomy with the *assignSpecies* function using a SILVA v138 training dataset downloaded from https://benjjneb.github.io/dada2/training.html [Quast et al., 2012].

B.5.2 Model Specification

We conduct a three-fold cross-validation of a model containing both contamination and detection effects. For each held-out fold r, if we let $\mathbf{e}_r := (\mathbf{1}_{[\text{sample 1 in fold } r]}, \ldots, \mathbf{1}_{[\text{sample n in fold } r]})^T$, $\mathbf{d} = (3^0, \ldots, 3^8)^T$, and $\circ$ indicate element-wise multiplication then we specify the model for this fold with

$$\begin{aligned}
\mathbf{Z} &= \begin{bmatrix} \mathbf{1} - \mathbf{e}_r & \mathbf{e}_r \end{bmatrix} \\
\tilde{\mathbf{Z}} &= \begin{bmatrix} \mathbf{d} \circ (\mathbf{1} - \mathbf{e}_r) \end{bmatrix} + \exp(\tilde{\alpha}) \begin{bmatrix} \mathbf{d} \circ \mathbf{e}_r \end{bmatrix} \\
\mathbf{X} &= \vec{\mathbf{1}} \\
\tilde{\mathbf{X}} &= 0
\end{aligned}$$

The relative abundance matrix $\mathbf{p}$ consists of two rows, the first of which is treated as fixed and known and contains the theoretical composition of the mock community used by Karstens et al. [2019]. The second row is to be estimated from observations on samples in the held-out fold. $\tilde{\mathbf{p}}$ consists of a single row, the first 247 elements of which we treat as unknown. We fix $\tilde{p}_{248} = 0$ as an identifiability constraint – identifiability problems arise here because all samples sequenced arise from the same specimen, we lack identifiability over, for any choice of fixed $\tilde{\mathbf{p}}^\star$ and $\mathbf{p}$, the set $\{\tilde{\mathbf{p}} = a * \tilde{\mathbf{p}}^\star + (1-a)\mathbf{p} : 0 \leq a \leq 1\}$. Briefly, we do not consider the assumption $\tilde{p}_{248} = 0$ unrealistic; while in general distinguishing between contaminant and

non-contaminant taxa is challenging, it is fairly frequently the case that choosing a single taxon *unlikely* to be a contaminant is not difficult. Moreover, we anticipate that in most applied settings, more than one specimen will be sequenced and this identifiability problem will hence not arise.

$\boldsymbol{\beta}$ consists of a single row, the first $J - 8 = 240$ elements of which (corresponding to contaminant taxa) are treated as fixed and known parameters equal to 0, as we cannot estimate detection efficiencies in contaminant taxa. The following 7 elements of β are treated as fixed and unknown (to be estimated from data), and $\boldsymbol{\beta}_{248}$ is set equal to 0 as an identifiability constraint. $\tilde{\boldsymbol{\gamma}}$ is specified as a single unknown parameter in $\mathbb{R}$

The full model fit without detection efficiencies $\boldsymbol{\beta}$ is specified by treating $\boldsymbol{\beta}$ as fixed and known with all elements equal to 0. We treat all samples as arising from the same specimen, so $\mathbf{Z} = \mathbf{1}$, $\tilde{\mathbf{Z}} = \mathbf{d}$, and $\mathbf{p}$ consists of a single row treated as an unknown relative abundance. Specifications of $\mathbf{X}$, $\tilde{\mathbf{X}}$, $\tilde{\mathbf{p}}$, and $\tilde{\gamma}$ are specified as above.

B.5.3 Additional summaries

Table B.2: Entries of $\hat{\boldsymbol{\beta}}$ (reference taxon L. fermentum) estimated from Karstens et al. [2019] data

Taxon	Estimate
P. aeruginosa	-1.29
E. coli	-0.20
S. enterica	-0.48
E. faecium	-2.09
S. aureus	-3.05
L. monocytogenes	-1.60
B. halotolerans	-0.74

For each taxon for which β_j is identifiable (i.e., taxa in the mock community), our model produces a point estimate $\hat{\beta}_j$, as shown in table B.2. (The reference taxon, *L. fermentum*, for which we enforce identifiability constraint $\beta_j = 0$, is excluded.) On the basis of this model, we estimate that in an equal mixture of *E. coli* and our reference taxon, *L. fermentum* sequenced by the method used by Karstens et al. [2019], we expect on average to observe $\exp(-0.20) \approx 0.82$ *E. coli* reads for each *L. fermentum* read. In an equal mixture of *S. aureus* and *L. fermentum* similarly sequenced, we expect on average to observe $\exp(-3.05) \approx 0.047$ reads for each *L. fermentum* read.

B.6 Simulation results based on Brooks et al. [2015] data

B.6.1 Identifiability

In our simulation using Brooks et al. [2015] data, we repeatedly selected sets of 3, 5, 10, or 20 samples from each of two plates of 40 samples sequenced by Brooks et al. [2015] to treat as known. For each randomly selected subset of samples to be treated as known, we required that $\boldsymbol{\beta}$ be identifiable on the basis of the taxa present in all known samples on each plate (i.e., identifiable from either group of samples). Briefly, this amounts to requiring that the graph whose nodes are the 7 taxa under consideration with edges between two nodes if the taxa those nodes correspond to are present in the same known sample. When a randomly selected set of samples failed this requirement, we redrew sets of samples until we found one that satisfied it.

B.6.2 Figures

Figure B.1 summarizes performance of cross-validated models fit to Brooks et al. [2015] data. Briefly, we observe very similar performance comparing unweighted and weighted estimators both in terms of root mean square error (RMSE) and proportion of elements of $\mathbf{p}$ estimated to be 0. RMSE is generally smaller for models fit on larger training sets, but it does not approach zero as training set size increases. We also observe a strong relationship between

RMSE and theoretical true relative abundance, which likely reflects a strong mean-variance relationship in the data.

We also observe generally a greater proportion of elements of **p** estimated to be zero with increasing training set size, although the degree to which this occurs depends on taxon. Weighting does not appear to have a large impact on this measure of predictive performance.

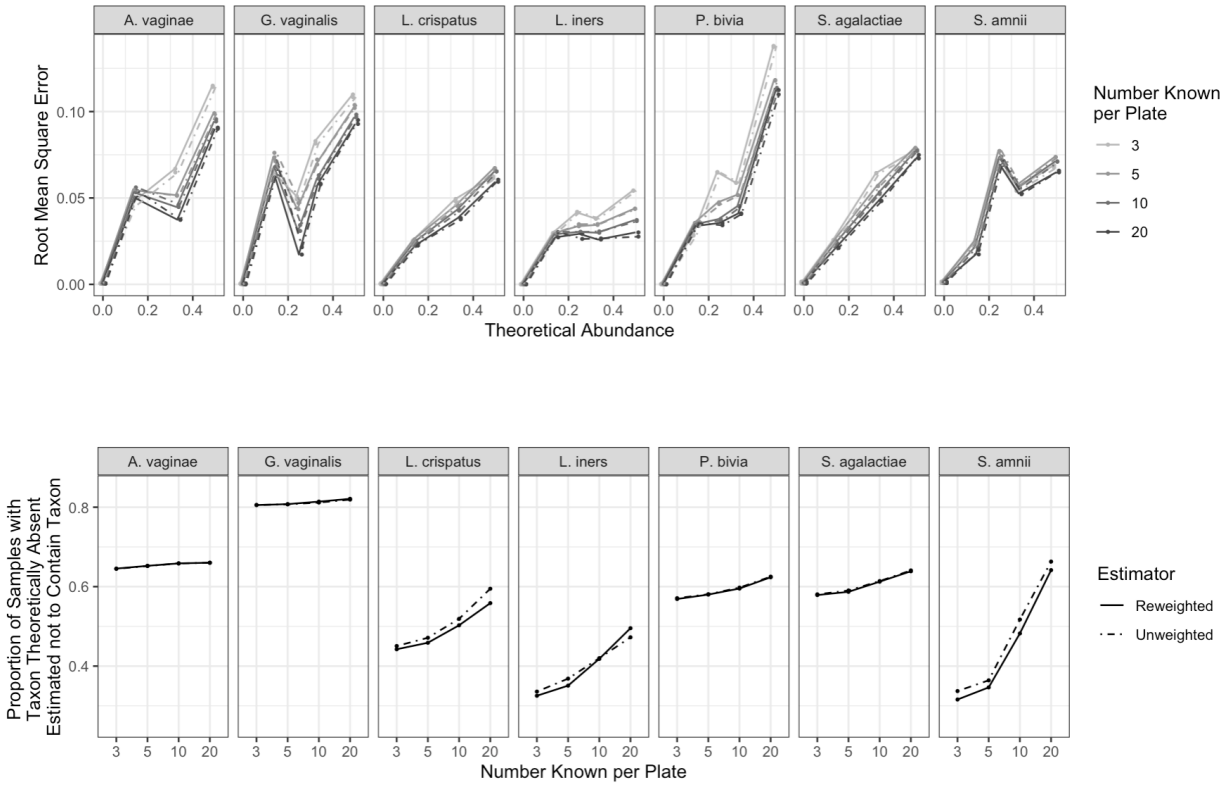

Figure B.1: Predictive performance of models fit on Brooks et al. [2015] data. The upper row includes root mean square error of relative abundance estimates by estimator, taxon, number of samples treated as known per plate, and true relative abundance. True relative abundance is given on the x-axis and root mean square error is plotted on the y-axis; for concision, true relative abundances equal to 1 are plotted at 0. Each column pane contains estimates for a different taxon, estimator is indicated with line type (solid for Poisson and dashed for weighted Poisson), and number of samples known per plate is indicated by color. In the lower row, proportion of elements of $\mathbf{P}$ truly equal to zero estimated to be equal to zero is plotted on the y-axis of each pane, and the x-axis gives number of samples per plate treated as known. Taxon and estimator are represented as in the upper row, and the proportion of nonzero elements of $\mathbf{W}$ corresponding to zero elements of $\mathbf{P}$ for each taxon is plotted as a dotted horizontal line.

B.7 Simulations with Artificial Data

Figure B.2 summarizes empirical coverage of marginal bootstrap 95% confidence intervals for elements p_{kj} of $\mathbf{p}$ obtained from simulations described in the main text. As discussed in the main text, coverage is high for $p_{kj} = 0$ but falls when $p_{kj} > 0$. Unsurprisingly, we observe higher coverage at larger sample sizes. Coverage of intervals based on reweighted estimators appears to be slightly lower than for unweighted estimators when data is Poisson-distributed (lefthand columns), but intervals using reweighted estimators substantially outperform unweighted intervals when data is negative binomial distributed, particularly when number of taxa J is larger.

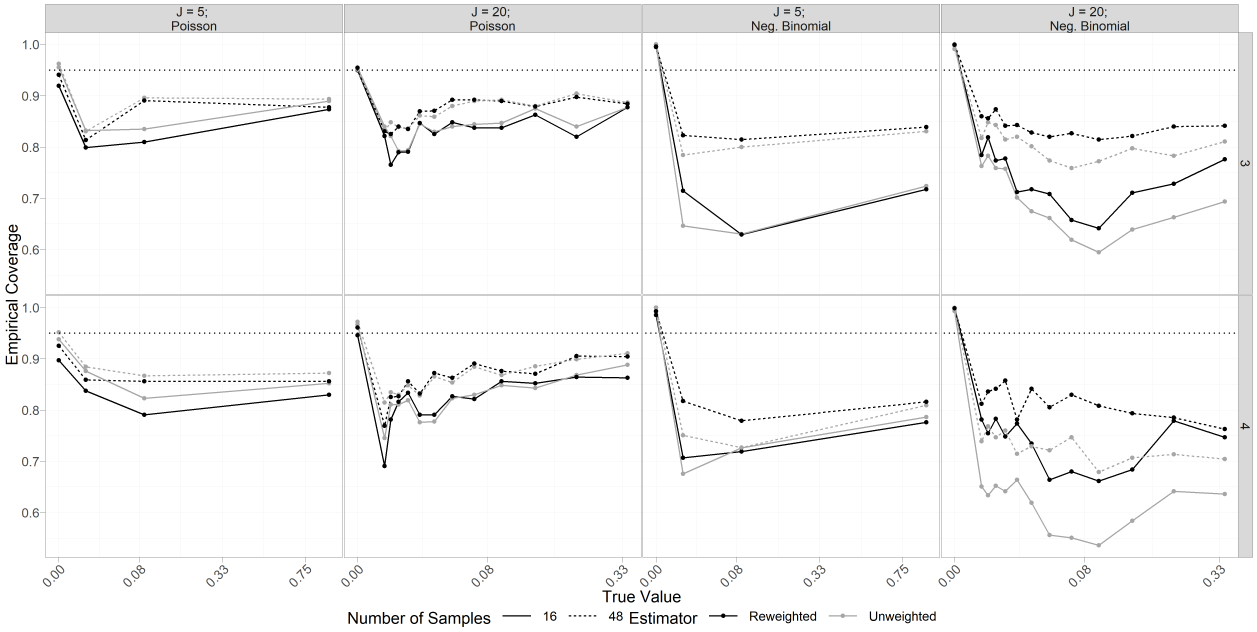

Figure B.2: Empirical coverage of marginal bootstrap 95%confidence intervals for p_{kj} ($k = 3, 4$) versus true value of p_{kj}. Coverage for tests based on unweighted and reweighted estimators are shown in grey and black, respectively. Sample size is indicated by line type (solid for $n = 16$ and dotted for $n = 48$). Columns give the conditional distribution of data (Poisson or Negative Binomial) and number of taxa J. Rows specify which row of $\mathbf{p}$ coverages are computed for.

Appendix C

ADDITIONAL SIMULATION RESULTS FOR PARTIALLY IDENTIFIED LOGLINEAR MODEL

C.1 Coverage under Null for Estimation using 10 Inner Bootstrap iterations

C.2 Power using 10 Inner Bootstrap iterations

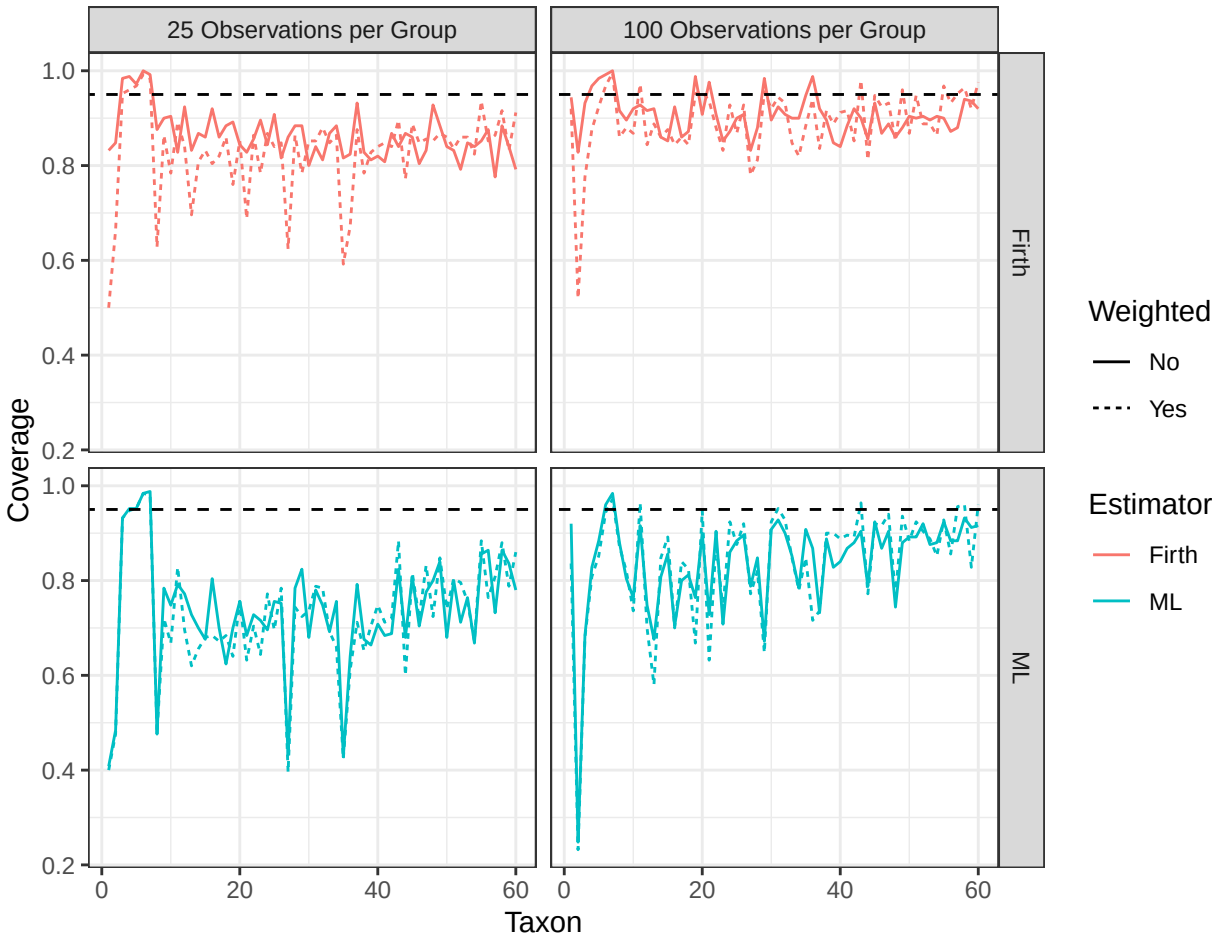

Figure C.1: Empirical coverage of marginal 95% confidence intervals computed from 250 simulation iterations for elements of second row of β under the null (all elements of 2nd row of β truly equal to zero) by estimator (maximum likelihood in cyan, maximum Firth-penalized likelihood in red), weighting (solid lines for unweighted estimation, dashed lines for weighted estimation), number of observations per group (given in columns), and taxon (x-axis). In all simulations, 500 outer and 10 inner bootstrap iterations are used.

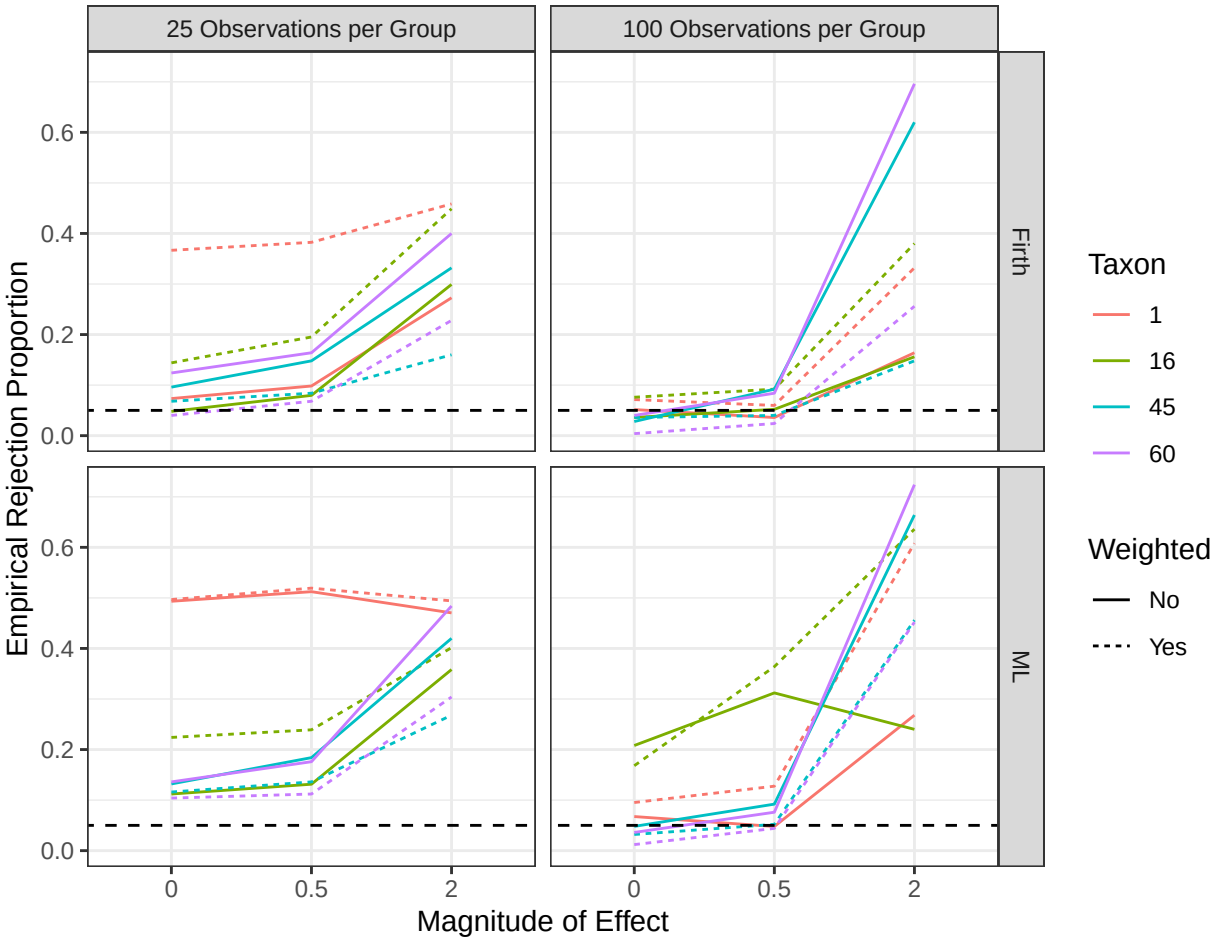

Figure C.2: Empirical power computed from 250 simulation iterations for potentially nonzero elements of second row of β by magnitude of effect (given on x-axis) by estimator (rows), weighted (solid lines for unweighted estimation, dashed lines for weighted estimation), number of observations per group (given in columns), and taxon (color). In all simulations, 500 outer and 10 inner bootstrap iterations are used.