Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

Yingtian Hu	Date

STATISTICAL METHODS FOR ANALYZING COMPOSITIONAL HUMAN MICROBIOME DATA

By

Yingtian Hu

Doctor of Philosophy

Biostatistics

Yijuan Hu, Ph.D.
Advisor

Glen A. Satten, Ph.D.
Co-advisor

Zhaohui (Steve) Qin, Ph.D.
Committee Member

Hao Wu, Ph.D.
Committee Member

Accepted:

Kimberly J. Arriola, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

STATISTICAL METHODS FOR ANALYZING COMPOSITIONAL HUMAN MICROBIOME DATA

By

Yingtian Hu
M.S., Emory University, 2020
B.S., Nankai University, 2016

Advisors: Yijuan Hu, Ph.D. and Glen A. Satten, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2022

Abstract

STATISTICAL METHODS FOR ANALYZING COMPOSITIONAL HUMAN MICROBIOME DATA

By

Yingtian Hu

With recent development in high-throughput sequencing technologies, human microbiome data are becoming more readily available, which drives people's interest in studying relationships between human microbiome and host diseases. Despite the fact that research in this field is booming, due to complex features of microbiome data, namely, compositionality, high dimensionality, sparsity, overdispersion, and experimental bias, researchers face many statistical challenges when analyzing the data. In particular, compositionality of microbiome data refers to the fact that the sequencing depth (library size) of each sample is noninformative, and converting the read counts into relative abundances yields compositional data. In this dissertation, we propose novel statistical methods for analyzing compositional human microbiome data. The dissertation is composed of three topics.

In the first topic, we address the problem of detecting differentially abundant bacterial taxa, i.e., taxa whose abundances are associated with the trait (condition) of interest. Our goal is to detect the taxa that initially respond to the condition change, not the taxa that show changes in relative abundance because of the compositional constraint. In this case, the null hypothesis that is tested at a taxon is that the ratio of the relative abundances at the taxon against some null taxon is unchanged. Existing methods tend to produce excessive false positive findings because they may improperly handle the sparsity of data, incorrectly identify the reference taxon, and fail to account for the experimental bias. To address these issues, we develop a novel method for compositional analysis of differential abundance, based on a robust version of logistic regression that we call LOCOM (**LO**gistic **COM**positional analysis). Our method circumvents the use of pseudocount, does not require the reference taxon to be null, and does not require normalization of the data. Further, it is applicable to a variety of microbiome studies with binary or continuous traits of interest and can account for potentially confounding covariates. We present simulation results to explicitly demonstrate the advantages of our proposed methods in terms of higher sensitivity and well-controlled false discovery rate (FDR) compared with other methods. We apply our method to two real microbiome datasets and compare with existing methods. LOCOM identifies more biologically meaningful differential abundant taxa.

In the second topic, we evaluate the impact of interactive bias on compositional analysis methods in testing differential abundance of taxa. Microbiome data are subject to experimental bias. However, this important feature has often been ignored in the development of statistical methods for analyzing microbiome data. McLaren, Willis and Callahan (2019) proposed a model (which we call the MWC model) for how such bias affects the measured taxonomic profiles, which assumes no taxon-taxon interactions. Our newly developed method, LOCOM, is robust to the experimental bias that follows the MWC model. However, there

is evidence for taxon-taxon interactions, so it is of interest to re-evaluate LOCOM and other compositional analysis methods in the presence of the interactive bias. We propose a model to describe the experimental bias in the measurement of a taxon that allows the contributions from the other taxa. Using this model, we conduct simulation studies to evaluate the impact of such experimental bias on the performance of LOCOM, as well as other compositional analysis methods. Our simulation results indicate that LOCOM is robust to any main bias and a reasonable range of interactive bias. The other methods tend to have inflated FDR even when there is only main bias. LOCOM maintains the highest sensitivity among all methods even when the other methods cannot control the FDR.

In the third topic, we study the association between microbiome composition and survival outcomes. Existing methods for survival outcomes are restricted to testing associations at the community level and do not provide results at the individual taxon level. An ad hoc approach testing taxon-level association using the Cox proportional hazard model may not perform well in the microbiome setting with sparse count data and small sample sizes. Here we develop a unified approach, an extension of the linear decomposition model (LDM) that allows testing both community-level and taxon-level association, to test survival outcomes. We propose to use the Martingale residuals or the deviance residuals obtained from the Cox model as continuous covariates in the LDM. We further construct tests that combine the results of analyzing each set of residuals separately. We also extend PERMANOVA, the most commonly used distance-based method for testing community-level hypotheses, to handle survival outcomes in a similar manner. Simulation results demonstrate that the LDM-based tests preserve the FDR for testing individual taxa and have good sensitivity. The LDM-based community-level tests and PERMANOVA-based tests have comparable or better power than competing methods. An analysis of data on the association of the gut microbiome and the time to acute graft-versus-host disease reveals several dozen associated taxa and improved community-level tests.

Acknowledgement

First and foremost, I would like to express my deepest gratitude to my advisors, Drs. Yijuan Hu and Glen A. Satten for their guidance, patience, inspiration, encouragement and tremendous support over the past years. It has been a great honor to work with them. They taught me how to think logically, communicate efficiently, be more organized, and pay attention to detail. They were always very patient and supportive when I had difficulties during my research. I will never forget their advice in my future endeavors.

I would like to thank my dissertation committee members, Drs. Zhaohui Qin and Hao Wu, for their precious time, thoughtful comments, and creative ideas. Their comments and suggestions have led to significantly improvement in this dissertation.

I would like to thank Dr. Jeanie Park for providing me the great opportunity to work as a biostatistician in her lab. The opportunity provides me with invaluable exposures to practical research and experience in analyzing real clinical data.

I would like to extend my appreciation to all faculty and staff members in the Department of Biostatistics and Bioinformatics at Emory, and to all my friends for great memories we have shared during the past six years.

Finally, I would like to thank my family for their unconditional love and support throughout my whole life.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview of human microbiome data

Human microbiome refers to the genetic material of all the microbota, including but not limited to bacteria, fungi, protozoa and virus that live on and inside the human body. Research shows that there are at least 10-100 trillion human microbota, an enormous number, in one person. This number is 10-fold more than the number of human cells (Ley et al., 2006; Turnbaugh et al., 2007).

Microbiome have large variation across different human populations. The sets of microbiome presenting in a given habitat in all or the vast majority of humans are referred as core human microbiome. Habitat can be the entire body or a specific body site. For example, both the gut and smaller region in gut are habitat. The sets of microbiome presenting in a given habitat in a small group of humans are referred to as variable microbiome. Several causes may contribute to this variation, namely host genotype, host pathobiology (disease status), host lifestyle and so on. Meanwhile, large variation also exists for microbiome in different body sites, like on the skin and in the mouth, stomach, colon and vagina. In order to better understand the microbiome composition across different human population and body sites, the human microbiome project (HMP) was initiated and great progress have been made (Turnbaugh et al., 2007).

A commonly used approach to obtain human microbiome data is 16S ribosomal RNA(rRNA) sequencing technology. The technology utilizes PCR to target and amplify portions of the variable regions (V1-V9) of the 16S rRNA gene (Laudadio et al., 2019), which is found in almost all microorganisms with enough sequence conservation for accurate alignment (Turnbaugh et al., 2007). Molecular barcodes are then given to amplicons, which would be further pooled together and sequenced. After sequencing from separate samples, data are further processed using various publicly available tools, and then clustered into Operational Taxonomic Units (OTUs) and summarized in a OTU count table. The OTU count table, consisted of OTUs of samples, summarizes the sequenced abundance of each microbes in each sample, and would be typically further processed for downstream statistical analysis

of microbiome data. In addition, lineages information of microbes can be organized hierarchically by taxonomic tree and phylogenetic tree according to organism similarities and evolutionary relationships. There are seven different levels of lineages in a taxonomic tree: species, genus, family, order, class, phylum, and kingdom.

Microbiome data has a number of important features. The first feature is compositionality. Because not only the number of sequenced reads varies across different samples, it is also limited by the capacity of the instrument. The observed read count can not actually reflect the absolute abundance, but can only reflect relative abundance of microbiome taxa in the original environment. The second notable features of microbiome data is its high dimensionality. The number of taxa may range from several hundreds to tens of thousands, or even larger while the sample size is usually no more than a few hundreds. In some cases, only dozens of samples are available for analysis. The third feature is high overdispersion. For the same taxa, even the read depth are very close among samples, observed read count number can still possibly range from 0 to 1000 across different samples. The fourth feature is sparsity. Researchers have shown that the percentage of zeros in the data ranges between 50% and 90% for many different studies. Zero counts occur for two reasons: (1) samples have low read depths so that rare taxa can not be captured; (2) due to the variability of the distribution of microbiome, some taxa are not shared by the entire population and some samples have 0 count for those taxa. The former is referred as "technical zeros" and the latter is called "structural zeros". The fifth feature is the complex correlation structure within data. Taxa can be correlated with multiple other taxa, and the correlation between different taxa can be positive or negative. Last but not the least, the measurements of microbiome data are biased that relative abundances measurement of the taxa in the sample systematically distort from their true values (Brooks, 2016; Sinha et al., 2017). The bias is resulted from preferentially measurement of some taxa during each step in an experimental workflow or protocol (Brooks, 2016; Hugerth and Andersson, 2017; Pollock et al., 2018).

1.2 Association analysis of compositional human microbiome data

Many researchers have shown that human microbiome are associated with human health status, including risk of disease development. For example, Vandeputte et al. (2017) discovered that changes in the microbial ecology of fecal samples were associated with changes in composition of the gut microbiome of patients diagnosed with Crohn's disease. Teo et al. (2015) conducted the study investigating changes in the microbial ecology of the lower respiratory tract of children. They found that the composition of microbiome had effect on infection severity and pathogen spread to lower airways. They also found that the development of asthma at later years may be caused by infections at the lower respiratory tract at a young age. Fettweis et al. (2019) found that women with a lower vaginal levels of Lactobacillus crispatus were more likely to deliver preterm. The acidic environment created by these bacteria in the vagina may protect the vagina and the womb against harmful microbes.

Therefore, building a better understanding of how human microbiome are associated with clinical outcomes can help monitor human health status and develop better diagnosis and treatment of human diseases. Currently, there are two main types of analysis discovering the association between microbiome composition and clinical outcomes: community-level association and taxon-level association. Community-level association measures the overall association of microbial community profiles with clinical outcomes. Taxon-level association measures the association between individual taxa and clinical outcomes. Taxa with significant association are usually named as differentially abundant taxa.

1.2.1 Two biological models

There are (at least) two biological models that can evaluate associations between microbiome compositions and clinical outcomes, i.e, how microbial community/taxa may change when comparing populations with different phenotypes or along a phenotypic gradient. In

one model, the null hypothesis at community level is that relative abundance of all taxa remain the same. The "no differential abundance" tested at a taxon is that the taxon relative abundance remains the same, i.e., any changes in taxon relative abundance across conditions are of interest. We refer to this hypothesis as "relative abundance hypothesis" in the following sections of Introduction. The other model only assumes that a few key taxa vary across different populations, while the rest show changes in relative abundance because of the compositional constraint. Thus, the second null hypothesis is that the ratios of relative abundances of taxa and reference non-differentially abundant taxa are unchanged, or the absolute abundances of taxa are unchanged. Because this hypothesis accounts for the compositional constraint that a change in relative abundance for one taxon leads to a counterbalancing change in other taxa, a statistical phenomenon known as compositional effects, it is generally referred to as compositional analysis. We refer to this hypothesis as "compositional hypothesis" in the following sections of Introduction.

To test the "relative abundance hypothesis", total sum scaling (TSS) normalization, calculated from dividing the counts by the library size, is usually adopted to obtain the relative abundance microbiome data. To test the "compositional hypothesis", a commonly-used strategy is to conduct log-ratio transformation of the mibrobiome data. This strategy is inspired by Aitchison's methodology for compositional data that only ratios are well-defined in compositional data (Aitchison, 1986). Here we give a brief review of two most popular types of log-ratio transformation. Suppose we have a compositional vector

$$\boldsymbol{x} = [x_1, x_2, \ldots, x_J] | x_i > 0, i = 1, 2, \ldots, D; \sum_{i=1}^{D} x_i = k.$$

Additive log-ratio (alr) transformation is defined by

$$alr(\boldsymbol{x}) = [log(\frac{x_1}{x_J}), log(\frac{x_2}{x_J}), \ldots, log(\frac{x_{J-1}}{x_J})].$$

The denominator can be an arbitrary component in the vector. The center log ratio (clr)

transformation is both an isomorphism and an isometry where $clr : S^D \rightarrow R^D$, which is defined by

$$clr(\boldsymbol{x}) = [log(\frac{x_1}{g(\boldsymbol{x})}), log(\frac{x_2}{g(\boldsymbol{x})}), \ldots, log(\frac{x_{J-1}}{g(\boldsymbol{x})})],$$

where $g(\boldsymbol{x})$ is the geometric mean of all components in $\boldsymbol{x}$.

1.2.2 Community-level tests

Community-level association test, also named as global test for global composition, identifies the overall association of microbial community profiles with clinical outcomes. We first review methods for testing the "relative abundance hypothesis". Two popular methods are Permutational Multivariate Analysis of Variance (PERMANOVA) (Anderson, 2005) and Microbiome Regression-based Kernel Association Test (MiRKAT) (Zhao et al., 2015). Both methods are distance-based methods that use ecologically meaningful dissimilarity metrics to measure the phylogenetic or taxonomic dissimilarity between different samples (beta diversity). The distances between each pair of samples will then be compared to the distribution of clinical outcomes of interest via multivariate testing approach. Despite their popularity, the power of distanced-based testing methods highly depends on the choice of the dissimilarity metric. Popular choices are Bray-Curtis distance, unweighted Unifrac distance, and weighted Unifrac distance. Generalized Unifrac is another commonly used metric that proposes to balance between weighted and unweighted Unifrac distance. How to choose the best dissimilarity metric is still a question in practice because it requires prior knowledge of how the microbiome influences the clinical outcome. Researchers have been working on handling this issue. Zhao et al. (2015) proposed optimal MiRKAT to consider multiple possible dissimilarity metric simultaneously. Tang, Chen and Alekseyenko (2016) proposed a method called PERMANOVA-S that ensembles multiple distances metrices to improve the power of PERMANOVA. Besides PERMANOVA and MiRKAT, Hu and Satten (2020) proposed a linear decomposition model (LDM) to test global association by building linear model on relative abundance data or transformation of relative abundance data. They illustrated

the relationship between LDM with other two distance-based methods, PERMANOVA and MiRKAT. Meanwhile, they also proposed PERMANOVAFL, a new and more powerful version of PERMANOVA with Freedman Lane permutation scheme (Freedman and Lane, 1983).

Besides these distance-based methods, there are other parametric models to test community-level association. La Rosa et. suggested (La Rosa et al., 2012) using Dirichlet-multinomial (DM) model to model the distribution of microbiome count data. They assumed that each count vector should follow a multinomial distribution with underlying proportion parameters sampled from a Dirichlet distribution. The method can test not only the association between the frequency of microbiome profile and the outcome of interest, but also the association between dispersion and the outcome of interest. However, since Dirichlet-multinomial assumes negative correlation between different taxa, but the correlation between taxa could be both positive and negative, it may not be a proper probability distribution to model the microbiome data in practice. In addition, current implementation of DM test can not adjust for confounding effects. Tang and Chen (2019) proposed another probability distribution, zero-inflated generalized Dirichlet multinomial distribution (ZIGDM) to model the multivariate taxon counts. Compared with DM, ZIGDM allows incorporating additional parameters to accommodate the complex correlation structure, over-dispersion and zero-inflation of the microbiome data. As a multivariate test, ZIGDM cannot handle high-dimensional microbial taxa.

All aforementioned methods focus on identifying the community-level association between binary, categorical or continuous clinical outcome. Since finding microbiome associations with possibly censored survival times is also an important and interesting problem, researchers have developed MiRKAT-S (Plantinga et al., 2017) and OMiSA (Koh et al., 2018) to test associations between microbiome data and survival outcomes at the community level. Both approaches first fit a Cox model to account for the relationship between any fixed covariates (excluding microbiome variables) and survival times. Then, under the random-effects framework, they compare variance-covariance matrix of the (Martingale) residuals from the

Cox model with the between-sample distance matrix calculated using the microbiome data: the similarity between the two matrices represents the extent of association between the microbiome and the survival outcome. In particular, MiRKAT-S uses an arbitrary distance matrix. OMiSA is an extension of MiRKAT-S that allows a family of power transformations of the relative abundance data to weigh abundant taxa and rare taxa differently.

To test the "compositional hypothesis", (Gloor et al., 2017) suggested applying PERMANOVA to clr transformed count data. Since log ratio transformation is not applicable to zero count, pseudo count is required before transformation.

1.2.3 Taxon-level association tests

Taxon-level association test, also named as individual test, identifies the association between individual taxa and clinical outcomes. The idea behind is very similar to differential abundance (DA) analysis in the analysis of other high-throughput sequencing data, such as microarray and RNA-seq. To test the "relative abundance hypothesis", a simple approach for DA when there is only a single binary covariate is to apply Wilcoxon rank-sum test to the relative abundance data directly. Hu and Satten (2020) proposed to use LDM model, a more sophisticated approach to test this hypothesis when covariates exist. This approach models the relative abundance through a linear decomposition model.

For the "compositional hypothesis", due to the compositional effects, analyses directly based on relative abundance are not applicable for testing this null hypothesis and are likely to produce a large number of false positive findings. Therefore, researchers have developed different strategies to handle this compositionality issue and test this "compositional hypothesis".

One of most commonly used strategies to accommodate compositional effect is robust normalization. The idea is to calculate a normalization factor (scale factor), dividing by which helps reduce or remove the compositional effect so that the abundance of non-differential taxa are still comparable and the differences of differential taxa are retained. After robust

normalization, people can apply standard statistical tools for testing the "compositional hypothesis". People have proposed a variety of approaches to normalize microbiome data, such as cumulative-sum scaling (CSS) (Paulson et al., 2013), GMPR (Chen et al., 2018), WRENCH (Kumar et al., 2018) and DACOMP (Brill et al., 2019). These approaches usually require the assumption that only a small fraction of taxa are differentially abundant.

Another commonly used strategy is to conduct log-ratio transformation of the mibrobiome data as we mentioned earlier. However, there are two important issues using log-ratio transformation. First, log-ratio transformation can not be applied to zero count. Pseudo count is usually required before the transformation. Secondly, although ratio keeps the absolute abundance information, it couples the information about two or more taxa so that we are unable to make inference about single taxon to decide whether it is differentially abundant in terms of absolute abundance. For example, if we use alr transformation and find the log ratio of taxa i and j are differential across conditions, we still have no information whether taxa i or taxa j is differentially abundant taxa respectively or both of them are. However, we can make inference about single taxon if we have a good reference to compare. In this example, if we know taxa j is non-differentially abundant, we then know that taxa i is differentially abundant. Therefore, in order to make valid inference for single taxa, a good reference or a robust normalization factor is still needed for methods based on log-ratio transformation.

Researchers have introduced log-ratio based methods with different strategies to solve the above two issues. Fernandes et al. (2014) proposed ALDEx2, a Bayesian compositional data analysis tool. ALDEx2 firstly converts count values to probabilities via Monte Carlo sampling from the Dirichlet distribution with the addition of a uniform prior. The addition of the uniform prior solves the zero issue. The tool then applies clr transformation to the sampled data and treats the geometric mean as a good reference or a robust normalization factor. Further inferences are conducted on clr-tranformed data. Mandal et al. (2015) proposed ANCOM, which now has became one of the most popular methods for differential abundance

analysis in the microbiome data. ANCOM uses alr transformation to obtain log-ratios with a small positive pseudo count added to zero to handle zero issue in the transformation. When it comes to identifying differential abundant taxa, ANCOM does not just choose one single taxon as the reference, but combines results from all taxa. Therefore, the method requires fitting $K(K-1)/2$ models to conduct pairwise comparison for K taxa. Recently, (Lin and Peddada, 2020) proposed a bias-corrected version of ANCOM, ANCOM-BC that models log-transformed taxa count with pseudocount under a linear regression framework, and estimates the unknown sampling fraction (bias term) through EM algorithm. The sampling fraction can be considered as normalization factor to help recover the absolute abundances at the log scale. Inspired by ANCOM and ANCOM-BC, Zhou, Wang, Zhao and Wang (2022) proposed fastANCOM. Similar to ANCOM-BC, fastANCOM also adopts the linear regression framework to model the log-transformed taxa count with pseudocount. But unlike ANCOM-BC, fastANCOM uses a simpler way to estimate the bias term by first identifying a subset of potentially non-differential taxa, and then estimating the mean difference in the unknown sampling fraction term between populations. Compared with ANCOM, fastANCOM only need fitting K models for K taxa and can provide p-value for each taxa. Zhou, He, Chen and Zhang (2022) proposed another method based on bias-correction, LinDA that fits linear regression models on the clr transformed data with pseudocount and then conducts bias-correction. They demonstrated that LinDA enjoys asymptotic false discovery control (FDR) control and can be extended to analyze correlated microbiome data. In general, the most common practice to resolve zero issue is to add a pseudocount, most frequently 1 or 0.5 or even smaller values, to the zeros or all entries of the taxa count table. However, there is no consensus on how to choose the pseudocount, and it has been shown that the choice of pseudocount can affect the conclusions of a compositional analysis (Costea et al., 2014; Paulson et al., 2014).

1.2.4 Unified approach to testing associations at both the community and taxon levels

Most methods for microbiome association test can only be conducted at either community level or taxon level. Since the findings of a community-level test could sometimes be inconsistent with findings of another taxon-level test, no unified testing approaches bring challenges to resolve these inconsistency. In order to fill this gap, Hu and Satten (Hu and Satten, 2020) proposed a linear decomposition model (LDM) that can not only perform testing at the community level but also at the individual taxon level, with additional control for FDR. LDM is based on a linear model that regresses the microbial data at each taxon on the (confounding) covariates that we wish to adjust for and the outcome variable(s) that we wish to test with. Inference is based on permutation to circumvent making parametric assumptions about the distribution of the taxon-level data. The LDM can be used for testing associations of microbiome with continuous or categorical (including binary) outcomes. In addition, LDM is highly versatile: it can analyze the taxon-level data at the relative abundance scale, the arcsin-root-transformed relative abundance scale (which is variance-stabilizing for Multinomial and Dirichlet-Multinomial count data) or any other transformation, as well as the presence-absence scale (Hu, Lane and Satten, 2021), and can also accommodate clustered samples (Zhu et al., 2021). The LDM is designed for testing the "relative abundance hypothesis". No existing approach has unified the two-level associations between microbiome compositions and survival outcomes to test the "relative abundance hypothesis". In addition, there are no unified approaches to test the "compositional hypothesis".

1.3 Experimental bias in microbiome data

Research have already shown that microbiome studies are biased, and biases are possibly introduced in almost every step in the experimental pipeline, resulting from preferentially measurement of some taxa during each step in an experimental workflow or protocol. For

example, different bacterial species have different tendencies to lyse so that the DNA they yield are different during DNA extraction. The number of 16S rRNA gene copies and PCR products can also be different. Upon that, different bacterial sequences may bind to primers in different ways so that some taxa are preferentially amplified compared to others. Sequencing platforms also have different abilities to read DNA with high GC content. Sources of biases in the bioinformatic processing pipeline include but no limited to read filtering, trimming, deduplication, read mapping. Due to these protocol-dependent and taxon-dependent biases, microbiome data generated from different protocols are incomparable. Therefore, without taking bias into account, analyse of microbiome data can lead to false conclusions. However, modeling every possible sources of bias factors is a complicated task.

McLaren et al. (2019) recently proposed a simple model, which we refer to as the MWC model here, to model the bias generation process in the microbiome studies. They demonstrated that MWC can estimate the experimental bias by analyzing mock community data where true relative abundances are known. In their approach, the observed relative abundance of each taxon is the product of the true relative abundance and a taxon-specific bias factor normalized by all taxa in the same sample. They assumed 1) biases at all steps in the experimental pipeline are multiplicative; 2) there are no taxon-taxon interactions, i.e., one taxon has no effects on the bias factor of the another taxon; 3) there are no covariate affecting bias factors (like plate effects or variations in extraction protocol. Under these assumptions, bias factors can be presented as taxon-specific. Later, Zhao and Satten (2021a) proposed a statistical model that generalizes the MWC model so that more complex questions about bias factors can be addressed. Their model generalized the MWC model to include covariates which may affect bias factors, and proposed permutation-based inference procedure to test complex hypothesis of bias factors across taxa and protocols. One interesting hypothesis they proposed to test is whether there is taxon-taxon interaction on bias factors.

Although it is well-known that microbiome data are subject to experimental bias, this crucial feature has often been ignored in the development of many statistical methods, es-

pecially the methods for differential abundance testing. Therefore, there is an increasing demand to develop statistical methods that are robust to different types of experimental bias.

1.4 Outline

In Chapter 2, we develop a novel method for compositional analysis of differential abundance, at both the taxon level and the community level, based on a robust version of logistic regression that called LOCOM (**LO**gistic **COM**positional analysis). In the Methodology section, we illustrate the motivations for adopting logistic regression to minimize the effect of experimental bias in analyzing microbiome data, and describe the details of our framework. In the Simulations and Data analysis section, we present simulation studies that compare the performance of LOCOM to other compositional methods. We also compare results from LOCOM and other methods in two real data application. We conclude with a Discussion section.

In Chapter 3, we evaluate the impact of interactive bias on compositional analysis methods in testing differential abundance of taxa. In the Methodology section, we start with the MWC model that includes only the main bias and then generalize the model to incorporate interactive bias. In the Simulations section, we present simulation studies that evaluate the performance of LOCOM, as well as other compositional analysis methods, in the presence of interactive bias. We conclude with a Discussion section.

In Chapter 4, we propose a unified approach, an extension of LDM, to test the two-level associations between microbiome data and survival outcomes. In the Methodology section, we describe our tests based on the Martingale residuals (LDM-m), showing their connection to MiRKAT-S, OMiSA, and the taxon-by-taxon Cox regression. Then we extend the tests to use the deviance residuals (LDM-d) and then construct combination tests (LDM-c) that combine the results from tests using the two types of residuals. In the Simulations and Data

analysis section, we first present simulation studies and then an application of all methods to data on a real microbiome data. We conclude with a brief Discussion section.

Chapter 2

LOCOM: A logistic regression model for testing differential abundance in compositional microbiome data with false discovery rate control

2.1 Introduction

Microbiome association studies are useful for the development of microbial biomarkers for prognosis and diagnosis of a disease or for the development of microbial targets (e.g., pathogenic or probiotic bacteria) for drug discovery, by detecting the taxa that are most strongly associated with the trait of interest (e.g., a clinical outcome or environmental factor). Read count data from 16S amplicon or metagenomic sequencing are typically summarized in a taxa count (or feature) table. Because the total sample read count (library size) is an experimental artifact, only the relative abundances of taxa, not absolute abundances, can be measured. Thus, microbial data are compositional (constrained to sum to 1). Analysis of microbial associations is further encumbered by data sparsity (having 50–90% zero counts in the taxa count table), high-dimensionality (having hundreds to thousands of taxa), and overdispersion. In addition, most microbiome association studies have relatively small sample sizes; further complications arise as the traits of interest may be either binary or continuous, and the detected associations may need to be adjusted for confounding covariates. Finally, any method for detecting taxon-trait associations should control the false discovery rate (FDR) (Hawinkel et al., 2017). The capability to handle all these features is essential for any statistical method to be practically useful.

There are (at least) two biological models for how microbial communities may change when comparing groups with different phenotypes or along a phenotypic gradient. In one model, a substantial proportion of the taxa in the community change; the concept "community state types" exemplifies this approach (see e.g., (Arumugam et al., 2011; Koren et al., 2013)). The null hypothesis of "no differential abundance" that is tested at a taxon is that the taxon relative abundance remains the same, i.e., any change in taxon relative abundance across conditions is of interest. Methods for testing this hypothesis include the linear decomposition model (LDM) (Hu and Satten, 2020) and direct application of non-parametric tests (e.g., the Wilcoxon rank-sum test) to relative abundance data or rarefied count data. In the other model, only a few key taxa are considered to change, while the other taxa show

changes in relative abundance because of the compositional constraint (Kumar et al., 2018; Brill et al., 2019). Thus, the null hypothesis that is tested at a taxon is that the *ratio* of the relative abundances at the taxon against some null taxon is unchanged. Methods for testing this hypothesis include ANCOM (Mandal et al., 2015), ANCOM-BC (Lin and Peddada, 2020), ALDEx2 (Fernandes et al., 2014), WRENCH (Kumar et al., 2018), and DACOMP (Brill et al., 2019). Because the hypothesis in the second model accounts for the compositional constraint that a change in relative abundance for one taxon necessarily implies a counterbalancing change in other taxa, it is generally referred to as *compositional analysis* (Gloor et al., 2017).

Methods for compositional analysis are typically based on some form of log-ratio transformation of the read count data. The ratio can be formed against a reference taxon or the geometric mean of relative abundances of all taxa, referred to as additive log-ratio (alr) or centered log-ratio (clr) transformation, respectively (Aitchison, 1986). Thus, zero count data, which cannot be log-transformed, is the major challenge in using compositional methods on microbiome data. A common practice is to add a *pseudocount*, most frequently 1 or 0.5 or even smaller values, to the zeros or all entries of the taxa count table (Aitchison, 1986; Paulson et al., 2013; Mandal et al., 2015; Zhao et al., 2018; Sohn and Li, 2019; Lin and Peddada, 2020). However, there is no consensus on how to choose the pseudocount, and it has been shown that the choice of pseudocount can affect the conclusions of a compositional analysis (Costea et al., 2014; Paulson et al., 2014).

The most popular pseudocount-based method for compositional analysis is perhaps ANCOM (Mandal et al., 2015), which has now evolved into ANCOM-BC (Lin and Peddada, 2020). After adding 0.001 to all count data, ANCOM performs the alr transformation and treats the transformed data as the response of the linear regression model that includes the traits of interest and confounding variables as covariates. For each taxon, ANCOM uses all other taxa, one at a time, as the reference in forming the alr transformation, and then it employs a heuristic strategy to declare taxa that are significantly differentially abundant

(outputting rankings of taxa instead of p-values). ANCOM-BC first estimates sampling fractions that are different across samples, and then models the log of read count data, in which zeros are replaced by pseudocount 1, through a linear regression model including the estimated sampling fraction as an offset term. This is essentially a normalization approach that first attempts to recover the absolute abundances of taxa and then test hypotheses about the absolute abundances. Unlike ANCOM, ANCOM-BC provides p-values for individual taxa. Both ANCOM and ANCOM-BC are restricted to group comparisons and can not handle continuous traits of interest, although adjustment for confounding covariates is supported.

Several methods have been developed that circumvent the use of pseudocount. ALDEx2 (Fernandes et al., 2014) first draws Monte-Carlo samples of non-zero relative abundances from Dirichlet distributions (with parameters constructed from read count data plus a uniform prior 0.5). Then, the sampled relative abundances are clr transformed and tested against the traits of interest via linear regression to yield p-values and adjusted p-values by the Benjamini-Hochberg (BH) procedure (Benjamini and Hochberg, 1995a), both of which are averaged over sampling replicates to give the final p-values and adjusted p-values. However, the sampling process adds noise to the data which may cause loss of power. In addition, by using the clr transformation, ALDEx2 is designed to identify differential abundant taxa relative to the *mean* of all taxa, which may be sensitive to outliers. DACOMP (Brill et al., 2019) is a normalization approach that first selects a set of null reference taxa by a data-adaptive procedure and then normalizes read count data by rarefaction so that each taxon within the reference has similar counts across samples. However, the selected reference set may mistakenly contain causal taxa, which may compromise the performance of the normalization. In addition, adjustment for confounding covariates is not supported, although continuous traits of interest are allowed. WRENCH (Kumar et al., 2018) is also a normalization approach that estimates group-specific compositional factors to bring the read counts of null taxa across groups to a similar level and employs DESeq2 to detect differentially abundant taxa. It is limited to group comparisons without confounding covariates.

It is also of interest to test differential abundance at the community (i.e., global) level, rather than taxon by taxon, using the compositional analysis approach. The most commonly used method for testing community-level hypotheses about the microbiome is PERMANOVA (McArdle and Anderson, 2001), which is a distance-based version of ANOVA. For compositional analysis, use of the Aitchison distance is recommended (Gloor et al., 2017), which is simply the Euclidean distance applied to the clr transformed data (Aitchison et al., 2000). Again, the clr transformation necessitates the use of pseudocount, so the choice of pseudocount may affect the outcome of the test.

Finally, it is of vital interest to develop a method that can provide valid inference even in the presence of experimental bias. Experimental bias is ubiquitous because each step in the sequencing experimental workflow (i.e., DNA extraction, PCR amplification, amplicon or metagenomic sequencing, and bioinformatics processing) preferentially measures (i.e., extracts, amplifies, sequences, and bioinformatically identifies) some taxa over others (McLaren et al., 2019; Brooks, 2016; Hugerth and Andersson, 2017; Pollock et al., 2018). For example, bacterial species differ in how easily they are lysed and therefore how much DNA they yield during DNA extraction (Costea et al., 2017). As a result, the bias distorts the *measured* taxon relative abundances from their *actual* values.

We are particularly interested in the case of differential bias, where the bias of taxa that are associated with a trait is systematically different from the bias of null taxa. A concrete example of this is the differential bias between bacteria in the phyla *Bacteroidetes* and *Firmicutes*. *Bacteroidetes* are gram-negative, while *Firmicutes* are gram-positive. It is known that gram-positive bacteria have strong cell walls and are hence harder to lyse than gram-negative bacteria; thus gram-positive bacteria may be underrepresented due to bias in the extraction step. The *Bacteroidetes-Firmicutes* ratio has been implicated in a number of studies of the gut microbiome (e.g., (Mariat et al., 2009; Magne et al., 2020)). Thus, studies that compare *Bacteroidetes* to *Firmicutes* may be affected by differential extraction bias. In some of our simulations, we consider the effect this kind of differential bias can have on the

FDR.

In this article, we develop a novel method for compositional analysis of differential abundance, at both the taxon level and the global level, based on a robust version of logistic regression that we call LOCOM (LOgistic COMpositional analysis). Our method circumvents the use of pseudocount, does not require the reference taxon to be null, and does not require normalization of the data. Further, it is applicable to a variety of microbiome studies with binary or continuous traits of interest and can account for potentially confounding covariates. In the methods section, we give the motivation for using logistic regression as a way to minimize the effect of experimental bias in analyzing microbiome data, and describe the details of our approach. In the results section, we present simulation studies that compare the performance of LOCOM to other compositional methods. We also compare results from LOCOM and other methods in the analysis of two microbiome datasets. We conclude with a discussion section.

2.2 Methodology

Let Y_{ij} be the read count of the jth taxon ($j = 1, \ldots, J$) in the ith sample ($i = 1, \ldots, n$) and N_i the library size of the ith sample. Because N_i can vary widely between samples, we focus on the relative abundance data as a form of normalized data. We denote by P_{ij} the observed relative abundance, given by Y_{ij}/N_i. We let X_i be a vector of q covariates including the (possibly multiple) traits of interest and other (confounding) covariates that we wish to adjust for, but excluding the intercept.

2.2.1 Motivation

Our starting point is the model of McClaren, Willis and Callahan (McLaren et al., 2019), as expanded by Zhao and Satten (Zhao and Satten, 2021b), which relates the expected value of the observed relative abundance, denoted by p_{ij}, to the true relative abundance we would

measure in an experiment with no experimental bias, denoted by π_{ij}. In particular, this model assumes that

$$\log(p_{ij}) = \log(\pi_{ij}) + \gamma_j + \alpha_i, \tag{2.1}$$

where γ_j is the taxon-specific *bias factor* that describes how the relative abundance is distorted by the bias, and α_i is the sample-specific *normalization factor* that ensures the composition constraint $\sum_{j=1}^{J} p_{ij} = 1$. Following (Zhao and Satten, 2021b), we further assume that the true relative abundance π_{ij} can be described by a baseline relative abundance π_j^0 that would characterize the true relative abundance of taxon j for a sample having $X_i = 0$, and a term that describes how the baseline relative abundance is changed in the presence of covariates $X_i \neq 0$. Then, we can replace (2.1) by

$$\log(p_{ij}) = \log(\pi_j^0) + X_i^{\mathrm{T}}\beta_j + \gamma_j + \alpha_i, \tag{2.2}$$

where β_j describes the way the true relative abundance changes with covariates X_i and is our parameter of interest. The presence of bias factors in (2.1) and (2.2) imply that inference based on the observed relative abundances P_{ij} may not give valid inference on β_j. It is clear that, without knowing the bias factor γ_j, we cannot estimate $\log(\pi_j^0)$ as $\log(\pi_j^0)$ and γ_j always appear together as a sum.

We can examine equation (2.2) to see if there are any combinations of parameters that could potentially be estimated without knowing the bias factors. Analyzing log probability ratios such as $\log(p_{ij}/p_{ij'})$ removes the effect of α_i (which depends on bias factors through normalization) but does not remove the effect of γ_j. However, if we use (2.2) to write log odds ratios of observed relative abundances for two different taxa and two different samples, we find

$$\log\left(\frac{p_{ij}p_{i'j'}}{p_{ij'}p_{i'j}}\right) = (X_i - X_{i'})^{\mathrm{T}}(\beta_j - \beta_{j'}), \tag{2.3}$$

which *is* independent of bias factors. This motivates the choice of logistic regression to analyze microbiome count data.

Note that testing $\beta_j - \beta_{j'} = 0$ in (2.3) corresponds to testing $p_{ij}/p_{ij'} = p_{i'j}/p_{i'j'}$, which is exactly the null hypothesis in a compositional analysis, e.g., in popular compositional models of the microbiome such as ANCOM and ALDEx2. As a result, logistic regression based on (2.3) is of interest even without the bias-removal motivation provided here.

2.2.2 Multivariate logistic regression model

Equation (2.3) implies a polychotomous logistic regression of the full $n \times J$ taxa count table. This is numerically difficult as the analysis of each taxon potentially requires all β_j parameters. Instead, we follow Begg and Grey (Begg and Gray, 1984) and analyze data using separate or "individualized" logistic regressions, each using data from just two taxa at a time. Rather than considering all possible pairs of taxa, we choose one taxon (without loss of generality, the Jth taxon) to be a reference taxon, and compare all other taxa to the reference taxon. Then, if we define $\mu_{ij} = p_{ij}/(p_{ij} + p_{iJ})$, equation (2.2) implies

$$\log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) = \theta_j + X_i^{\mathrm{T}}(\beta_j - \beta_J), \qquad 1 \le j \le J-1 \tag{2.4}$$

where the intercepts $\theta_j = \left[\log(\pi_j^0) - \log(\pi_J^0)\right] + (\gamma_j - \gamma_J)$ are treated as nuisance parameters since estimation of the γ_js is not possible when the π_j^0s are not known. As written, the model is over-parameterized because only the $J-1$ log odds ratios $\beta_j - \beta_J$ are identifiable. To make the full set of β_js identifiable requires a constraint; we temporarily use $\beta_J = 0$ with the understanding that β_j then refers to an odds ratio that compares taxon j to the reference taxon J. According to (Begg and Gray, 1984), the efficiency of individualized logistic regression highly depends on the prevalence (relative abundance) of the reference category, so we recommend that the reference taxon be a common taxon that is present in a large number of samples.

To avoid distributional assumptions in a standard logistic regression, we consider the score functions as estimating functions. When a taxon is rare and/or the sample size is

small, it may occur that all (or nearly all) counts for that taxon are zero in one group (e.g., the case or control group), which is referred to as separation in the literature on logistic regression. It is known that the Firth bias correction (Firth, 1993), when applied to logistic regression (Georg and Michael, 2002), solves the problem of separation. Hence, we estimate (θ_j, β_j) by solving the Firth-corrected score equations

$$U_j(\theta_j, \beta_j) = \sum_{i=1}^{n} \Big[Y_{ij} - M_{ij}\mu_{ij} + h_i\big(0.5 - \mu_{ij}\big) \Big] \begin{pmatrix} 1 \\ X_i \end{pmatrix} = 0,$$

where $M_{ij} = Y_{ij} + Y_{iJ}$ and h_i is the ith diagonal element of the weighted hat matrix $W_j^{\frac{1}{2}} X (X^{\mathrm{T}} W_j X)^{-1} X^{\mathrm{T}} W_j^{\frac{1}{2}}$ with the design matrix X (including a column of 1's corresponding to the intercept) and the diagonal weight matrix $W_j = \mathrm{diag}\,\{M_{1j}\mu_{1j}(1-\mu_{1j}), \ldots, M_{nj}\mu_{nj}(1-\mu_{nj})\}$. We let $\widehat{\beta}_j$ denote the estimator of β_j obtained by solving the above equations.

2.2.3 Testing hypotheses at individual taxa

Now we describe the formula for the null hypotheses we test to decide which taxa are "null", i.e. have no effect. Write $\beta_j = (\beta_{j,1}, \beta_{j,-1})$, where $\beta_{j,1}$ is the coefficient for the trait of interest and $\beta_{j,-1}$ for the other covariates. We assume the trait of interest has only one component, but the approach can be generalized to test multiple traits simultaneously (see Discussion). The naive formula $\beta_{j,1} = 0$ only implies that the effect of the trait on the jth taxon is the same as the effect of the trait on the reference taxon; thus testing $\beta_{j,1} = 0$ only identifies null taxa when the reference taxon used in (2.4) is itself null.

As we have no *a priori* knowledge about whether the reference taxon is null or causal, we seek an approach that does not require such knowledge; in addition, we need a test for the reference taxon itself. To this end, we make the assumption that more than half of the taxa are null taxa, which has been frequently adopted in compositional methods (Brill et al., 2019; Kumar et al., 2018). With this assumption, we can expect $\mathrm{median}_{j'=1,\ldots,J}\{\beta_{j',1}\}$ to correspond to the value of $\beta_{j^*,1}$ for some taxon j^* that is null. If we then consider

parameters $\tilde{\beta}_{j,1} = \beta_{j,1} - \text{median}_{j'=1,\ldots,J}\{\beta_{j',1}\} = \beta_{j,1} - \beta_{j^*,1}$ in place of parameters $\beta_j - \beta_J$ in (2.4), then using $\widetilde{\beta}_{j,1} = 0$ as a null hypothesis *does* correspond to testing whether taxon j is null. Thus, we wish to test the null hypotheses

$$H_{j0} : \beta_{j,1} - \text{median}_{j'=1,\ldots,J}\{\beta_{j',1}\} = 0.$$

Note that centering by the median can also be thought of as replacing the constraint $\beta_{J,1} = 0$ to identify $\beta_{j,1}$s for all j. To test these null hypotheses, we use the statistic

$$\mathbb{Z}_j = \widehat{\beta}_{j,1} - \text{median}_{j'=1,\ldots,J}\{\widehat{\beta}_{j',1}\}.$$

Note that we use $\widehat{\beta}_{J,1} = 0$ both when calculating the median and obtaining $\mathbb{Z}_J$ for the reference taxon. Also note that the $\mathbb{Z}_j$s are reminiscent of centered log-ratios, in which the log of the abundance is centered by the *mean* of the log abundances. Use of the median in place of the mean for our centering is advantageous as the mean is sensitive to large or outlying observations that do not affect the median. Since the odds ratios we estimate each use data from only two taxa, our method is subcompositionally coherent in the sense of Atchinson (2005).

In the simplest case testing a binary trait that takes values 0 and 1, with no other covariates, $\mathbb{Z}_j$ is invariant to different choices of the reference taxon. This is because in this simple case, all estimated (pairwise) log odds ratios are of the form $\left(\widehat{\beta}_{j,1} - \widehat{\beta}_{j',1}\right) = \log\{n_{1j}n_{0j'}/(n_{0j}n_{1j'})\}$, where $n_{xj} = \sum_{i:X_i=x} Y_{ij}$ and so are completely free of the reference taxon. This holds even if the Firth-corrected estimator is used because, in this simple case, the Firth-corrected estimator corresponds to adding 1/2 to each n_{xj} (Firth, 1993; Georg and Michael, 2002); note that n_{xj} is an aggregated read count in a group of samples and thus this result is fundamentally different from the aforementioned pseudocount approach that adds a pseudocount to each read count of a sample. For the general case, we evaluate the dependence of $\mathbb{Z}_j$ on the reference taxon via simulations.

To avoid distributional assumptions in sparse microbiome data, we assess the significance of $\mathbb{Z}_j$ using the permutation scheme for logistic regression proposed by Potter (Potter, 2005), which is described as follows. The covariate vector X_i is partitioned into (T_i, C_i) where T_i denotes the trait of interest and C_i the other covariates. A linear regression of T_i on C_i and an intercept is fit to obtain the residual T_{ir}, which is then permuted to obtain $T_{ir}^{(b)}$ and to construct the new covariate vector $X_i^{(b)} = \left(T_{ir}^{(b)}, C_i\right)$. We follow the same procedure as for the observed dataset to obtain the estimate of $\beta_{j,1}$ from the bth permutation replicate, denoted by $\widehat{\beta}_{j,1}^{(b)}$, and the corresponding statistic $\mathbb{Z}_j^{(b)} = \widehat{\beta}_{j,1}^{(b)} - \text{median}_{j'}\{\widehat{\beta}_{j',1}^{(b)}\}$. We adopt Sandve's sequential stopping rule (Sandve et al., 2011) with a minor modification to stop the permutation procedure, which is described below. For each taxon j, after the Bth permutation we store the (cumulative) number of times that $\mathbb{Z}_j^{(b)}$ falls on the left (i.e., is less than) and right side (i.e. is greater than) of $\mathbb{Z}_j$ which we denote by L_j and R_j, respectively. We count the number of *rejections* to be $2\min(L_j, R_j)$. The p-value based on B permutations is given by $p_j = [2\min(L_j, R_j)+1]/(B+1)$ and the q-value is calculated according to (Sandve et al., 2011). The permutation procedure is continued until every taxon either has a q-value below the nominal FDR level or has accumulated a number of rejections exceeding a pre-specified value (e.g., 100). This stopping rule is slightly different from Sandve's in that we obtain $\widehat{\beta}_{j,1}^{(b)}$ for every taxon at every permutation, rather than stopping permutation early for some taxa, because the median calculation requires $\widehat{\beta}_{j,1}^{(b)}$ from all taxa.

2.2.4 Testing the global hypothesis

The global null hypothesis is that there are no differentially abundant taxa, i.e., H_{j0} holds for every taxon. Given the p-values at individual taxa, it is straightforward to construct a global test statistic by combining the individual p-values. Here we adopt the harmonic-mean approach to combining p-values proposed by Wilson et al. (Wilson, 2019a), which is more robust to the dependence structure among taxa than Fisher's method, and has more focus on the smallest p-value(s) (i.e., more power for scenarios with sparse, strong signals) than

Fisher's method. The harmonic mean of the p_js is $J/\big(\sum_{j=1}^{J} p_j^{-1}\big)$, for which smaller values correspond to stronger evidence against the null hypothesis. To have a test statistic with the "usual" directionality, we choose $\mathbb{Z}_{\text{global}} = \sum_{j=1}^{J} p_j^{-1}$. We use all permutation replicates generated for taxon-level tests, say B replicates, to assess the significance of $\mathbb{Z}_{\text{global}}$. At the bth replicate, the test statistic is $\mathbb{Z}_{\text{global}}^{(b)} = \sum_{j=1}^{J} \big\{p_j^{(b)}\big\}^{-1}$, where $p_j^{(b)}$ is the p-value of taxon j for this null replicate. Following (Westfall and Young, 1993), we calculate the null p-value $p_j^{(b)}$ using the rank statistic to be $p_j^{(b)} = 2B^{-1} \min\Big\{\Big[\text{rank}\big(\mathbb{Z}_j^{(b)}\big) - 0.5\Big], \Big[B - \text{rank}\big(\mathbb{Z}_j^{(b)}\big) + 0.5\Big]\Big\}$, where $\text{rank}\big(\mathbb{Z}_j^{(b)}\big)$ is the rank of $\mathbb{Z}_j^{(b)}$ among B such statistics. Let R_{global} be the number of times that $\mathbb{Z}_{\text{global}}^{(b)}$ falls on the right hand side of $\mathbb{Z}_{\text{global}}$. Then, the global p-value is given by $\big(R_{\text{global}} + 1\big)/(B + 1)$.

2.3 Simulations

2.3.1 Simulation studies

We used simulation studies to evaluate the performance of LOCOM and compare its performance to other compositional analysis packages. We based our simulations on data on 856 taxa of the upper-respiratory-tract (URT) microbiome; these taxa correspond to the "OTUs" in the original report on these data by Charlson et al. (Charlson et al., 2010). We considered both binary and continuous traits of interest and both binary and continuous confounders, as well as the case of no confounder. We mainly focused on two causal mechanisms. For the first mechanism (referred to as M1), we randomly sampled 20 taxa (after excluding the most abundant taxon) whose mean relative abundances were greater than 0.005 as observed in the URT data (i.e., ranking among the top 40 most abundant taxa) to be *causal* (i.e., associated with the trait of interest). For the second mechanism (referred to as M2), we selected the top five most abundant taxa (having mean relative abundance 0.105, 0.062, 0.054, 0.050, and 0.049) to be *causal*. In some cases, we also considered two variations of M1, one randomly sampling 500 taxa (again excluding the top one) to be causal to create

a scenario that violated our assumption that more than half of the taxa are null, and one randomly sampling 20 rare taxa (whose mean relative abundances were between 0.001 and 0.002) to be causal, which are referred to as M1-500 and M1-rare, respectively. For simulations with a confounding covariate, we assumed the confounder was associated with 20 taxa under M1 (10 sampled at random from the 20 causal taxa and 10 from the null taxa) and 5 taxa under M2 (2 from the 5 causal taxa and 3 from the null taxa). We simulated most data without adding experimental bias, but did conduct one set of simulations having differential experimental bias. We focused on datasets having 100 observations but also simulated some datasets with 50 or 200 observations.

To be specific, we let T_i denote the trait and C_i the confounder for the ith sample. To generate a binary trait, we selected an equal number of samples with $T_i = 1$ and $T_i = 0$. When a binary confounder was present, we drew C_i from the Bernoulli distribution with probability 0.2 in samples with $T_i = 0$ and from the Bernoulli distribution with probability 0.8 in samples with $T_i = 1$. When a continuous confounder was present, we drew C_i from the uniform distribution $U[-1, 1]$ in samples with $T_i = 0$ and $U[0, 2]$ in samples with $T_i = 1$. To generate a continuous trait, we sampled it from $U[-1, 1]$ when there was no confounder. When there was a binary confounder, we used the aforementioned data generated for a binary trait and a continuous confounder but exchanged the roles of trait and confounder. When there was a continuous confounder, we generated T_i from $U[-1, 1]$ and a third variable Z_i from $U[-1, 1]$ independently of T_i, and then constructed the confounder $C_i = \rho T_i + \sqrt{1 - \rho^2} Z_i$, where ρ was fixed at 0.5.

To simulate read count data for the 856 taxa, we first sampled the *baseline* (when $T_i = 0$ and $C_i = 0$) relative abundances $\pi_i^{(0)} = \left(\pi_{i1}^{(0)}, \pi_{i2}^{(0)}, \ldots, \pi_{iJ}^{(0)}\right)$ of all taxa for each sample from the Dirichlet distribution $Dirichlet(\overline{\pi}, \theta)$, where the mean parameter $\overline{\pi}$ and overdispersion parameter θ took the estimated mean and overdispersion (0.02) from fitting the Dirichlet-Multinomial (DM) model to the URT data. We formed the relative abundances p_{ij} for all taxa by spiking the j'th causal taxon with an $\exp(\beta_{j',1})$ fold change and the j''th

confounder-associated taxon with an $\exp(\beta_{j'',2})$ fold change, and then re-normalizing the relative abundances, so that

$$p_{ij} = \frac{\exp\left(\gamma_j + \beta_{j,1}T_i + \beta_{j,2}C_i\right)\pi_{ij}^{(0)}}{\sum_{j'=1}^{J}\exp\left(\gamma_{j'} + \beta_{j',1}T_i + \beta_{j',2}C_i\right)\pi_{ij'}^{(0)}} \quad , \tag{2.5}$$

where γ_j was the bias factor for the jth taxon. Note that $\beta_{j,1} = 0$ for null taxa, $\beta_{j,2} = 0$ for confounder-independent taxa, and $\gamma_j = 0$ for all taxa for data without experimental bias. In most cases, for simplicity, we set $\beta_{j,1} = \beta$ for all causal taxa, and thus β is a single parameter that we refer to as the effect size; we refer to $\exp(\beta)$ as the fold change. In some cases, we also considered the more general scenario when different values were sampled for different $\beta_{j,1}$. We fixed $\beta_{j,2} = \log(2)$ for all confounder-associated taxa. When there was no confounder, we simply dropped the term $\beta_{j,2}C_i$ (or equivalently, set $\beta_{j,2} = 0$ for all js) in calculating p_{ij}. In cases with differential experimental bias, we drew γ_j from $N(0, 0.8^2)$ for non-causal taxa and from $N(1, 0.8^2)$ for causal taxa; thus the bias-related fold changes varied roughly between 0.2 and 5 for most (95%) non-causal taxa and between 0.55 and 13.5 for most causal taxa, which are within a reasonable range according to (McLaren et al., 2019). Finally, we generated the taxon count data for each sample using the Multinomial model with mean $p_i = (p_{i1}, p_{i2}, \ldots, p_{iJ})$ and library size sampled from $N(10000, (10000/3)^2)$ and left-truncated at 2000.

In order to evaluate the robustness of our simulation results, we changed our simulation procedure in the following ways. First, we replaced the compositional model (2.5) by a model that generates microbiome-trait associations by assigning differential relative abundances only at causal taxa (which corresponds to the first biological model introduced in Background and constitutes a mis-specified model for LOCOM and other compositional analyses). Second, we replaced the DM model by a Poisson log-normal mixture (PLNM) model (which can impose any pre-specified correlation structure across taxa) for generating read count data. Both replacement models are described in Supplementary Text S2 of our

LDM paper (Hu and Satten, 2020). We also followed the LDM paper by basing our simulations on its association scenarios, which were denoted by S1 and S2. Scenario S1 assumed a large number of causal taxa (428 taxa in the LDM paper, which we modified here to 500 to create violation of our assumption that fewer than half the taxa are causal). Scenario S2 chose the top 10 most abundant taxa to be causal; here we will refer to the two scenarios as S1-500 and S2. Note that the data simulated using the PLNM model appeared to be less overdispersed and less sparse compared to data simulated using the DM model.

We applied two versions of LOCOM: one used the most abundant null taxon as the reference, which is referred to as LOCOM-null, and one used the most abundant causal taxon as the reference, referred to as LOCOM-causal. Both versions use the median of $\widehat{\beta}_{j',1}$s to compute the test statistic. Of course, in a real application, we would not know whether or not the reference taxon we had chosen was null or causal; we differentiate these two "versions" of LOCOM here to show that LOCOM is robust to whether the reference taxon is null or causal. In practice, when the most abundant taxon is chosen as the reference, the results from LOCOM would correspond to LOCOM-null in M1 and to LOCOM-causal in M2.

For testing the global hypothesis, we compared LOCOM to PERMANOVA (the `adonis2` function in the `vegan` R package) based on the Aitchison distance, which is referred to as PERMANOVA-half and PERMANOVA-one corresponding to adding pseudocount 0.5 and 1, respectively, to all cells. The type I error and power of the global tests were assessed at the nominal level 0.05 based on 5000 and 1000 replicates of data, respectively.

For testing individual taxa, we compared LOCOM to ANCOM, ANCOM-BC, ALDEx2, DACOMP, and WRENCH. However, ANCOM, ANCOM-BC, and WRENCH cannot handle continuous traits; DACOMP and WRENCH cannot adjust for other covariates. Prior to analysis, we removed taxa having fewer than 20% presence (i.e., present in fewer than 20% of samples) in each simulated dataset. For ANCOM and ANCOM-BC, we also considered their own filtering criterion with 10% presence as the cutoff and refer to these meth-

ods as ANCOM^o and ANCOM-BC^o. In the case with a binary trait only, we considered two additional pseudocount-based methods, Wilcox-alr-half and Wilcox-alr-one, which add pseudocount 0.5 and 1, respectively, to all cells, form the alr using the most abundant null taxon as the reference, perform the Wilcoxon rank-sum test at individual log ratios, and correct multiple comparisons using the BH procedure. Because the reference was selected to be a taxon known to be null, these methods are not applicable to real studies but were included in the simulations here to assess the properties of the pseudocount approach to testing individual taxa. In the case with a binary trait only, we also applied the Wilcoxon test directly to relative abundance data, i.e., data with total-sum scaling (TSS); although not a compositional method, this is commonly used in microbiome studies. The sensitivity (proportion of truly causal taxa that were detected) and empirical FDR were assessed at nominal level 20% based on 1000 replicates of data. We chose a relatively high nominal FDR level because the numbers of causal taxa in both M1 and M2 were small. In some cases, we also considered a lower nominal FDR level of 10%.

2.3.2 Simulation results

The type I error of the global tests for all simulation scenarios are summarized in Table A.1. In all scenarios, LOCOM-null and LOCOM-causal yielded type I error rates that were close to the nominal level and generally closer for sample size 200 than 100. Note that, in cases when there was a confounder, there was substantial inflation of type I error when the confounder was not accounted for (Table A.2), demonstrating that LOCOM is effective in adjusting for confounders. The PERMANOVA tests also controlled type I error. In cases without any confounder, the zero data were similarly distributed across trait values under the (global) null, so the effect of adding pseudocount is non-differential. In cases with a confounder, the taxa associated with the confounder caused the zeros to be differentially distributed across trait values, so that adding pseudocount had a differential effect for different trait values; however, this difference was adjusted by including the confounder as a

covariate in the model. Note that, although the pseudocount approach did not lead to invalid global tests, it did lead to invalid tests at individual taxa (in the presence of causal taxa), as indicated in the empirical FDR of Wilcox-alr-one and Wilcox-alr-half (e.g., Figures 2.1).

Figures 2.1–2.4 present power of the global tests and sensitivity and empirical FDR of the individual taxon tests, for a binary or continuous trait without and with a binary confounder, in scenarios M1 and M2 without experimental bias. The results for cases with a continuous confounder are deferred to Figures A.1–A.2, which show similar patterns of results to their counterparts with a binary confounder (Figures 2.2 and 2.4). The results in Figures 1-4 all have sample size 100 and FDR level 20%. To explore the effects of changing sample size and FDR level, we restricted to the two most important scenarios, one with a binary trait with no confounder in which all methods are applicable, and one with a binary trait and a binary confounder which is very common in real data. We changed the sample size to 50 (Figures A.3–A.4) or 200 (Figures A.5–A.6), then changed the nominal FDR level to 10% (Figures A.7–A.8). In general, those results show similar patterns to their counterparts with sample size 100 and nominal FDR level 20%.

In the simplest scenario with a binary trait and no confounder (Figures 2.1, A.3, and A.5), LOCOM-null and LOCOM-causal yielded identical type I error and power; in fact, the two methods gave identical p-values for every dataset in this case, which corroborated our claim that the test is invariant to different reference taxa. In other scenarios, LOCOM-null and LOCOM-causal produced similar results although the one using the more abundant taxon as the reference (LOCOM-null in M1 and LOCOM-causal in M2) tended to be more powerful and more sensitive. The aforementioned figures (Figures 2.1–2.4, A.1–A.8) show that the LOCOM tests yielded (almost) the highest power for testing the global hypothesis; LOCOM always controlled the FDR for testing individual taxa (even with the sample size 50) and had the highest sensitivity among methods that also controlled the FDR.

The competing methods generally have limited application to the scenarios we considered and significantly inferior performance to LOCOM. PERMANOVA had similar power to the

LOCOM global test in M1 but lost substantial power to LOCOM in M2 (e.g., Figures 2.1–2.4), likely because the Aitchison distance used by PERMANOVA may not be efficient in capturing sparse signals (only 5 causal taxa in M2) whereas the harmonic mean p-value combination method that LOCOM uses focuses on the strongest signal(s). For testing individual taxa, ALDEx2 is the only method that is applicable to all scenarios we considered; however, it tended to lose control of FDR when the effect size β was large (e.g., Figures 2.1–2.2) and it had much lower sensitivity than LOCOM in all cases. ANCOM and ANCOM-BC are only applicable for testing binary traits, with or without confounders. ANCOM easily lost control of FDR when the effect size was small, especially with their own, less stringent filtering criterion (e.g., Figures 2.1–2.2). ANCOM-BC tended to lose control of FDR when the effect size was large, especially when there was a confounder (e.g., Figure 2.2). Both ANCOM and ANCOM-BC had substantially lower sensitivity than LOCOM when they controlled the FDR. DACOMP is applicable for testing both binary and continuous traits but does not allow adjustment of any confounder. In scenarios without a confounder, DACOMP had good control of FDR, and while the sensitivity of DACOMP tended to be the largest among all competing methods, it was noticeably lower than that of LOCOM (e.g., Figures 2.1 and 2.3). WRENCH is only applicable to one scenario (with a binary trait and no confounder) in which case it had inflated FDR and nevertheless low sensitivity (e.g., Figure 2.1). The pseudocount methods, Wilcox-alr-half and Wilcox-alr-one, almost always produced inflated FDR, especially when the effect size was large so that zeros at null taxa were more differentially distributed across trait values (e.g., Figure 2.1). As expected, the Wilcox-TSS method had inflated FDR in simulations based on the compositional model (2.5) (e.g., Figure 2.1) but controlled the FDR in simulations based on differential relative abundances (Figure 2.6).

Results for simulated data with differential experimental bias (and a binary trait and no confounder) are shown in Figure 2.5. These simulations showed that, while LOCOM, ANCOM, and DACOMP were unaffected by differential bias, all other methods were sensitive to differential bias and yielded significantly inflated FDR in the presence of such bias.

Results for simulations based on the differential relative abundance model and the PLNM model are shown in Figure 2.6. In this setting, Wilcox-TSS is the most appropriate method. Indeed, it always controlled the FDR and yielded the highest sensitivity (except for the pseudocount methods which had inflated FDR). Interestingly, LOCOM controlled the FDR in both S1-500 and S2, even when S1-500 assumed 500 taxa to be causal; the reason might be that most causal taxa in this setting had very weak signals and act almost like null taxa. Note that LOCOM generated similar sensitivity to the "gold standard" (Wilcox-TSS). The PERMANOVA tests had higher power than the LOCOM global tests in S1-500 likely because the signals were very dense there.

Results for simulated data generated under M1-500 and M1-rare are shown in Figure A.9. When our assumption that more than half of the taxa are null was violated (M1-500), LOCOM lost control of the FDR as expected. However, the FDR inflation of LOCOM appears to be smaller than most competing methods and LOCOM maintained good sensitivity. When the causal taxa were all rare (M1-rare), LOCOM still yielded the highest sensitivity while controlling the FDR, although the absolute sensitivity values were low.

Results for simulated data with heterogeneous $\beta_{j,1}$ values are displayed in Figure A.10. The patterns we observed with heterogeneous $\beta_{j,1}$ values were similar to those seen in the analogous simulations with homogeneous $\beta_{j,1}$ values (Figure 2.2).

2.4 Data analysis

2.4.1 URT microbiome Data

The data for our first example were generated as part of a study to examine the effect of cigarette smoking on the oropharyngeal and nasopharyngeal microbiome (Charlson et al., 2010). We focused on the left oropharyngeal microbiome in this analysis. The 16S sequence data were summarized into a taxa count table consisting of data from 60 samples and 856 taxa. The trait of interest was a binary variable for smoking status, which divided the par-

ticipants into 28 smokers and 32 nonsmokers. Other covariates include gender and antibiotic use within the last 3 months. There was an imbalance in the proportion of males by smoking status (75% in smokers, 56% in non-smokers), indicating a potential confounding effect of gender. Since there were only three samples who used antibiotics within the last 3 months, we excluded these samples from our analysis and adjusted for gender only. We adopted the same filter (20% presence) as in the simulation studies, which resulted in 111 taxa for downstream analysis. We applied LOCOM with the most abundant taxon (having mean relative abundance 10.5% before filtering and 11.4% after filtering) as the reference. Given the need to adjust for gender, we only applied ANCOM, ANCOM-BC, and ALDEx2 as a comparison. The nominal FDR was set at 10%.

As shown in the upper panel of Table 2.1, the global p-value of LOCOM is 0.0045, which indicates a significant difference in the overall microbiome profile between smokers and non-smokers after adjusting for gender. At the taxon level, LOCOM, ALDEx2, ANCOM, and ANCOM-BC detected 6, 0, 2, and 2 taxa, respectively; Figure 2.7 displays a Venn diagram of these sets of taxa; Table A.3 lists information on the 6 taxa detected by LOCOM. Figure 2.8 shows the distributions of relative abundance across four covariate groups cross-classified by smoking status and gender, for taxa detected by LOCOM, ANCOM, and ANCOM-BC, as well as for two null taxa. One null taxon is the taxon with the median $\widehat{\beta}_{j,1}$ value. The other is the average of a group of null taxa for improved stability. The two null taxa both had lower relative abundance in smokers than in non-smokers, among either females or males. The six taxa detected by LOCOM all had the opposite trend (i.e., higher relative abundance in smokers than in non-smokers), indicating that these taxa are likely to be real signals (i.e., overgrew in smokers). The taxon detected by ANCOM only also had the opposite trend to the null taxa, but it was not detected by LOCOM because the adjusted p-value (0.137) by LOCOM did not meet the nominal FDR. The taxon detected by ANCOM-BC only had a similar trend as the null taxa, suggesting that this taxon may actually be a null taxon; indeed, the adjusted p-value by LOCOM is 0.674. Note that the difference in relative

abundance distributions between smokers and non-smokers at null taxa may be considered as the counterbalancing change that the null taxa underwent in response to the changes at the causal taxa.

The original analysis of this dataset (Charlson et al., 2010) reported that *Megasphaera* and *Veillonella spp.* were most enriched in the left oropharynx of smokers compared to non-smokers. Later, a large study of oral microbiome (from oral wash samples) in 1204 American adults (Wu, Peters, Dominianni, Zhang, Pei, Yang, Ma, Purdue, Jacobs, Gapstur et al., 2016) reported enrichment of *Atopobium*, *Streptococcus*, and *Veillonella* in smokers compared to non-smokers. More recently, a shotgun metagenomic sequencing study of salivary microbiome in Hungary population (Wirth et al., 2020) reported enrichment of *Prevotella* and *Megasphaera* in smokers compared to non-smokers. Thus, all six taxa detected by LOCOM have been implicated in the literature, even if we only consider the latter two independent studies. These taxa were largely missed by ANCOM and ANCOM-BC.

2.4.2 PPI microbiome data

The data for our second example were generated in a study of the association between the mucosal microbiome in the prepouch-ileum (PPI) and host gene expression among patients with IBD (Morgan et al., 2015). The PPI microbiome data from 196 IBD patients were summarized in a taxa count table with 7,000 taxa classified at the genus level. The gene expression data at 33,297 host transcripts, as well as clinical metadata such as antibiotic use (yes/no), inflammation score (0–9), and disease type (familial adenomatous polyposis/FAP and non-FAP) were also available. The data also included nine gene principal components (gPCs) that together explained 50% of the total variance in host gene expression. Here, we included all nine gPCs as multiple traits of interest into one model while adjusting for the three potentially confounding covariates. We filtered out taxa based on our previous filtering criterion, which resulted in 507 taxa to be included in the analysis. We applied LOCOM with the most abundant (8.2%) taxon as the reference. Given the continuous traits of interest

and the three covariates, we only considered ALDEx2 for comparison. The nominal FDR was set at 10%.

The results of PPI data analysis are presented in the lower panel of Table 2.1. LOCOM discovered that gPC2, gPC3, and gPC5 had significant associations with the overall microbial profiles at the $\alpha = 0.05$ level. LOCOM detected 2, 2, and 32 taxa as associated with gPC2, gPC3, and gPC5, respectively, at the 10% FDR level, and did not detect any taxa for the gPCs that were not found to be associated with the microbiome by the global test. Among the 32 taxa associated with gPC5, 15 belong to the genus *Escherichia* (Table A.3), which appeared frequently in the literature of IBD according to a highly-cited review article (Ni et al., 2017). ALDEx2 failed to detect any taxa.

2.5 Discussion

We have presented LOCOM, a novel compositional approach for testing differential abundance in the microbiome data, at both the taxon level and the global level. The global statistic is an aggregate of p-values from tests of individual taxa, so results from the taxon-level and global tests are coherent. LOCOM allows both binary and continuous traits of interest, can test multiple traits simultaneously, and can adjust for confounding covariates. In our simulations, the taxa detected by LOCOM always preserved FDR while those identified by the competing methods did not, even though LOCOM had clearly superior sensitivity. In addition, LOCOM also provided a global test that always controlled the type I error and had good power compared to PERMANOVA. In analysis of the URT microbiome data, we demonstrated that the taxa detected by LOCOM were likely to be real signals while those detected by ANCOM and/or ANCOM-BC but not LOCOM may be false positives. In analysis of the PPI microbiome data, since global and taxon-specific tests were coherent, LOCOM identified significant taxa only for gene principal components that were globally significant.

Like many compositional methods (e.g., DACOMP and WRENCH), LOCOM adopts the

assumption that more than half of the taxa in the community are null. This assumption may not be valid in some cases, for example, in testing higher taxonomic levels such as the class or phylum level. In theory, when this assumption does not hold, LOCOM, which always compares each taxon with the "median" taxon (with the median effect size estimate $\widehat{\beta}_{j,1}$), would find differences at truly null taxa. In our simulations, however, we found that, when most causal taxa had very weak signals, LOCOM still controlled the FDR (Figures 2.6 and A.9).

We showed both theoretically and with simulation studies that LOCOM is unaffected by experimental bias, even when bias factors are differentially distributed between causal and non-causal taxa. While some competing compositional methods (ANCOM and DACOMP) share this robustness, others (ANCOM-BC, ALDEx2, and WRENCH) do not. The problem in ALDEx2 may be related to the choice of centering; in general, the centered log ratio will not be robust when there are cells with zero counts, since this centering will depend on the set of taxa seen in each sample even if a pseudocount is used. Thus, the centering may not cancel out when comparing log ratios from different samples, leaving these comparisons affected by the particular bias factors that characterize the data being analyzed. Note that any compositional method should perform well when the bias is non-differential, since the centering will be the same on average in each sample.

It is possible to generalize LOCOM to test a trait with more than one component, such as a categorical trait with more than two levels. While ordered categories could be handled in the framework presented here by assigning an appropriate score to each category and then treating this score as a continuous variable, a categorical trait with K unordered categories would presumably require testing $K-1$ components to fully describe the variable. Within the framework presented here, we could then compare some summary (e.g., max or mean) of these test statistics to their equivalent value in the null permutations. Although this "better" analysis would require some software development and simulation testing, a simpler proposal could provide results within the existing framework, by calculating separate (marginal) p-

values for each of the $K-1$ components and then combining these p-values into a single test statistic, e.g., by using the harmonic mean statistic we used to form our global test. Choosing these $K-1$ components to be orthogonal may be helpful here. We hope to modify LOCOM to incorporate multi-component traits such as multi-category variables in future work.

Our filtering criterion to exclude taxa with fewer than 20% presence in the sample worked well for the extensive simulation studies we conducted. In fact, a compositional analysis performs best when non-null taxa are relatively common throughout all samples. Analyses that look for the effect of rare taxa should probably be focussed on a presence-absence analysis (Hu, Lane and Satten, 2021; Hu and Satten, 2021), or on a method based directly on relative abundances.

The compositional null hypothesis considered here is also appropriate in other experimental settings, such as studies of gene expression. This hypothesis corresponds to the scenario that a small number of microbes have "bloomed" while the absolute counts of the others have not changed; this is the reason we made the assumption that more than half of the taxa are null taxa, which is commonly made in other compositional methods. In the gene expression experiment, we often see only a few genes that are differentially expressed; the majority of genes have the same expression in cases and controls. However, it is not completely clear that the compositional hypothesis is applicable to microbiome data because, unlike genes, microbes interact with each other: not only do they compete for resources, but they also change their environment in ways that favor some microbes and suppress others. For example, *Lactobacilli* generally make lactic acid, which changes the pH of the environment. This suppresses microbes that do not thrive in an acidic environment while encouraging growth of microbes that do. Because the microbiota are a community, it is not unreasonable to expect that potentially every taxon changes between cases and controls. The "community change" null hypothesis may also be reasonable because, when comparing the alpha diversity with causal taxa spiked in to a case group, the control group would have a lower alpha diversity

(i.e., lower evenness); if this change in alpha diversity is meaningful, then the "community change" null hypothesis is appropriate. Note that, unlike the compositional null, the "community change" null hypothesis will consider *all* taxon relative abundances to be potentially changed if extra counts of a small number of taxa are "spiked in". When the "community change" null hypothesis seems more reasonable than the compositional null hypothesis, then a method that applies directly to relative abundance data such as the LDM is more appropriate. However, the LDM when applied to relative abundance data is not invariant to experimental bias the way LOCOM is; in fact, hypotheses based on differences in relative abundances typically require tests based on unbiased data to be valid.

Like LOCOM, ANCOM is based on comparing pairs of taxa. However, ANCOM yielded lower sensitivity than LOCOM in our simulations (e.g., Figures 2.1 and 2.2). There are several possible reasons. First, LOCOM analyzes the count data using logistic regression which downweighs zero counts, while ANCOM analyzes alr-transformed count data using linear regression which makes data with zero or very small counts more influential; the former is based on transformation of parameters (i.e., true relative abundances), while the latter is based on transformation of data. Second, ANCOM's approach of adding pseudocounts further introduces noise and possibly bias to the data. Third, LOCOM uses the most abundant taxon as the reference while ANCOM looks at all possible pairs of taxa, which can lead to unstable log ratios when both taxa are rare. Finally, ANCOM's strategy to declare differentially abundant taxa uses an arbitrary cutoff which may not be well-calibrated.

We have implemented our method in the R package LOCOM, which is available on GitHub at `https://github.com/yijuanhu/LOCOM` in formats appropriate for Macintosh or Windows. LOCOM is computationally efficient for data with small sample sizes but can take longer for larger sample sizes. For example, using parallel computing (by parallelizing permutation replicates) with 4 cores of a MacBook Pro laptop (1.4 GHz Quad-Core Intel Core i5, 8GB memory), it took 11s to analyze a simulated dataset with 100 samples, 11s to analyze the URT data, and 40 mins to analyze the PPI data. In considering this last timing,

it should be noted that the analysis considered 9 traits simultaneously in the presence of 3 confounding covariates, and as such is more complex than the typical microbiome analysis. In addition, LOCOM could be further parallelized by splitting the data into subsets with sets of taxa that only share the reference taxon and then combining the values of $\beta_{j,1}$ from each dataset (care should be taken to use the same seed for each analysis so that the same set of permutations is used).

Table 2.1: Results in analysis of the two real datasets

	Global p-value	Number of detected taxa			
Trait	LOCOM	LOCOM	ALDEx2	ANCOM	ANCOM-BC
URT microbiome data					
Smoking	0.0045	6	0	2	2
PPI microbiome data					
gPC1	0.70	0	0	NA	NA
gPC2	0.020	2	0	NA	NA
gPC3	0.018	2	0	NA	NA
gPC4	0.16	0	0	NA	NA
gPC5	0.0070	32	0	NA	NA
gPC6	0.59	0	0	NA	NA
gPC7	0.11	0	0	NA	NA
gPC8	0.21	0	0	NA	NA
gPC9	0.11	0	0	NA	NA

Note: ANCOM and ANCOM-BC are not applicable for testing continuous traits.

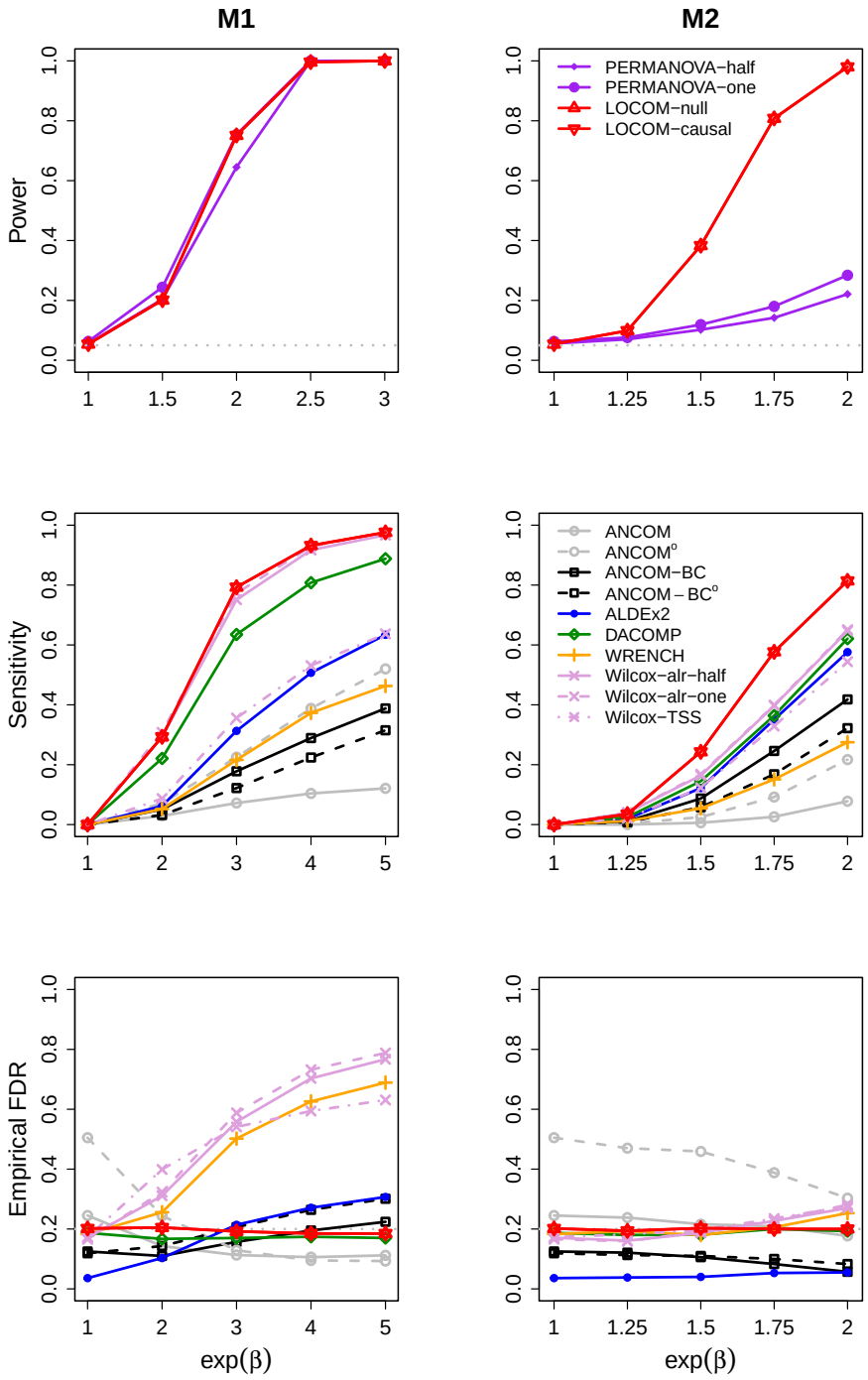


Figure 2.1: Simulation results for data ($n = 100$) with a binary trait (and no confounder). The power at $\exp(\beta) = 1$ corresponds to the type I error. The gray dotted line indicates the nominal type I error 0.05 in the first row and the nominal FDR 20% in the last row.

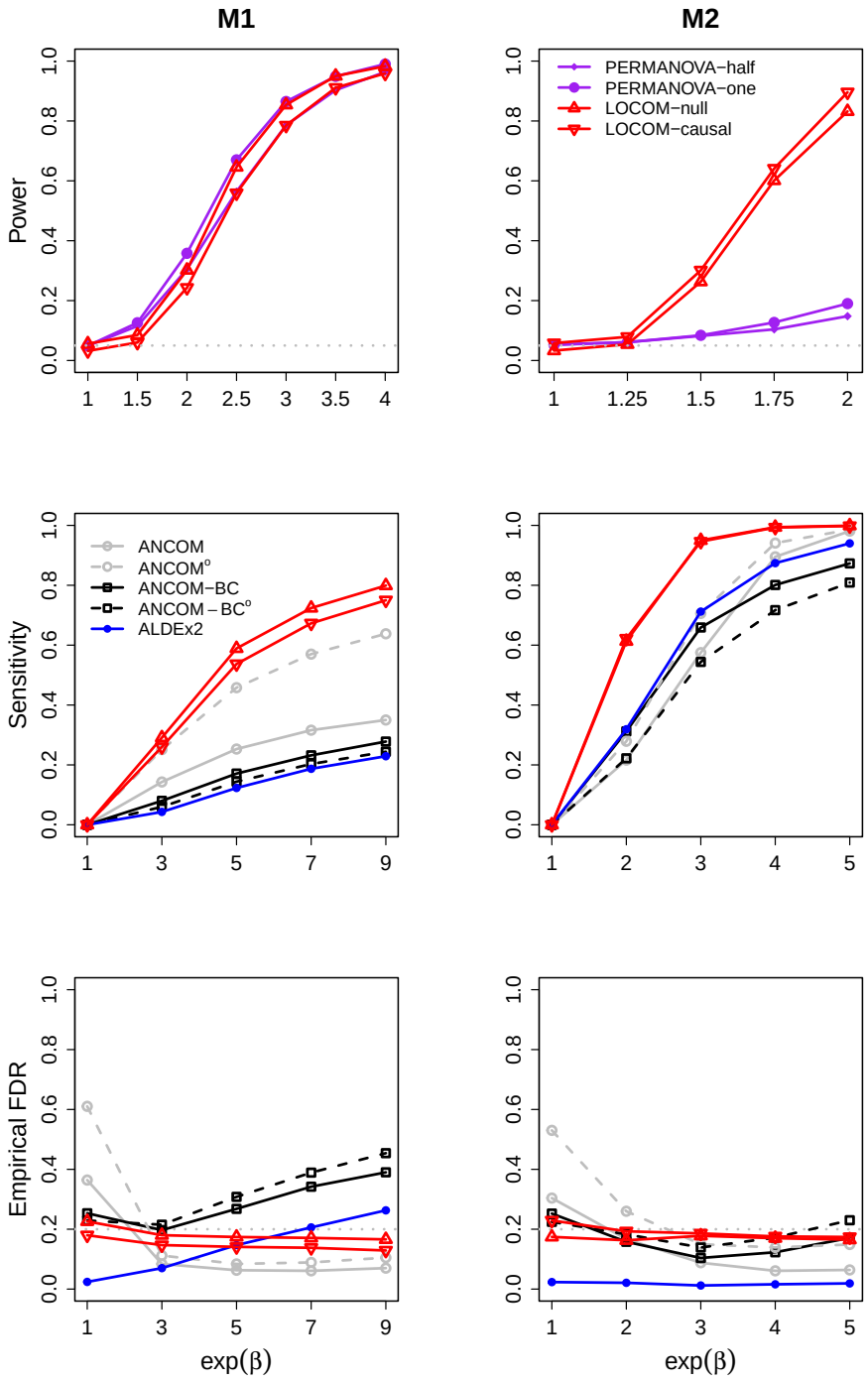


Figure 2.2: Simulation results for data ($n = 100$) with a binary trait and a binary confounder.

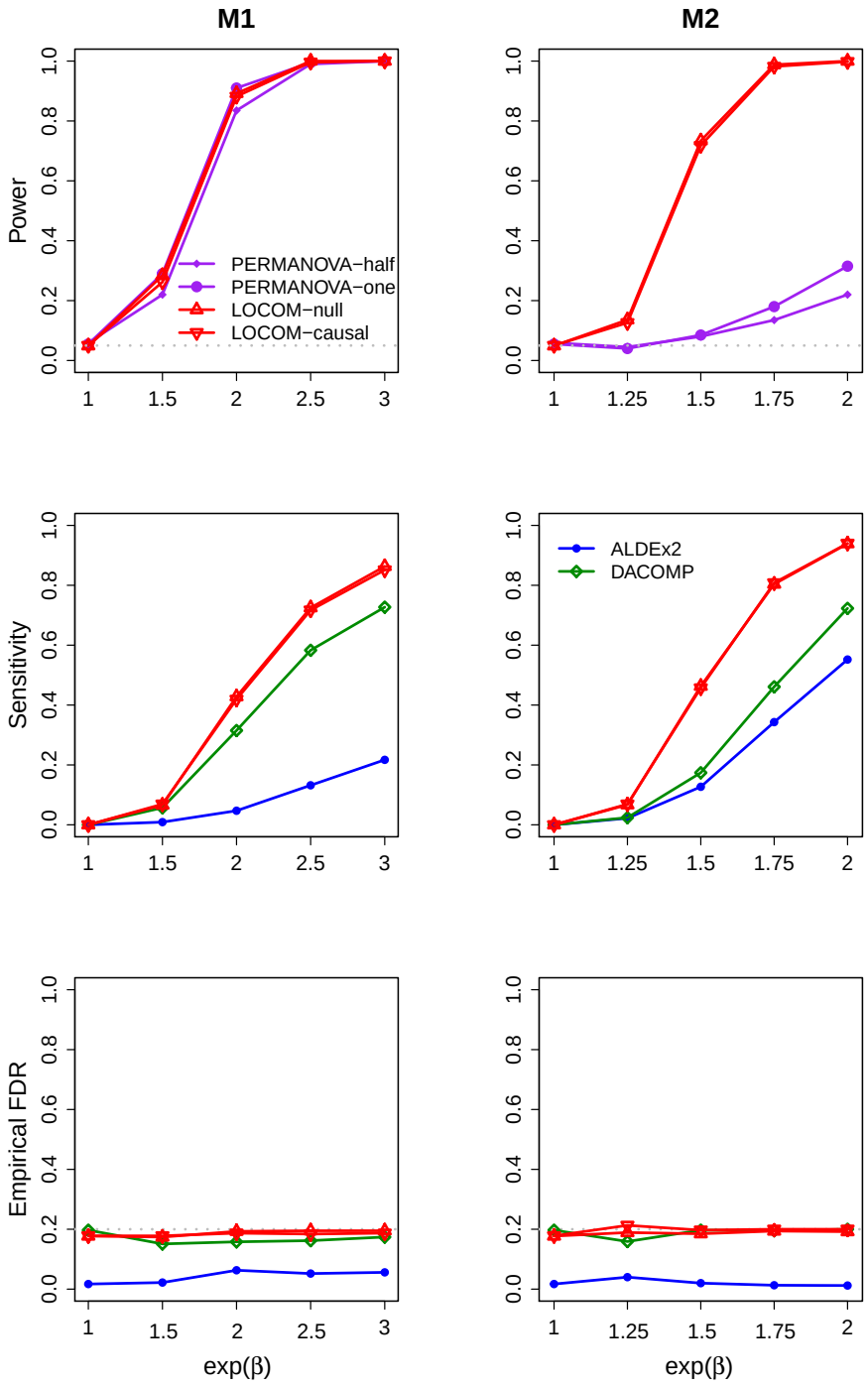


Figure 2.3: Simulation results for data ($n = 100$) with a continuous trait (and no confounder).

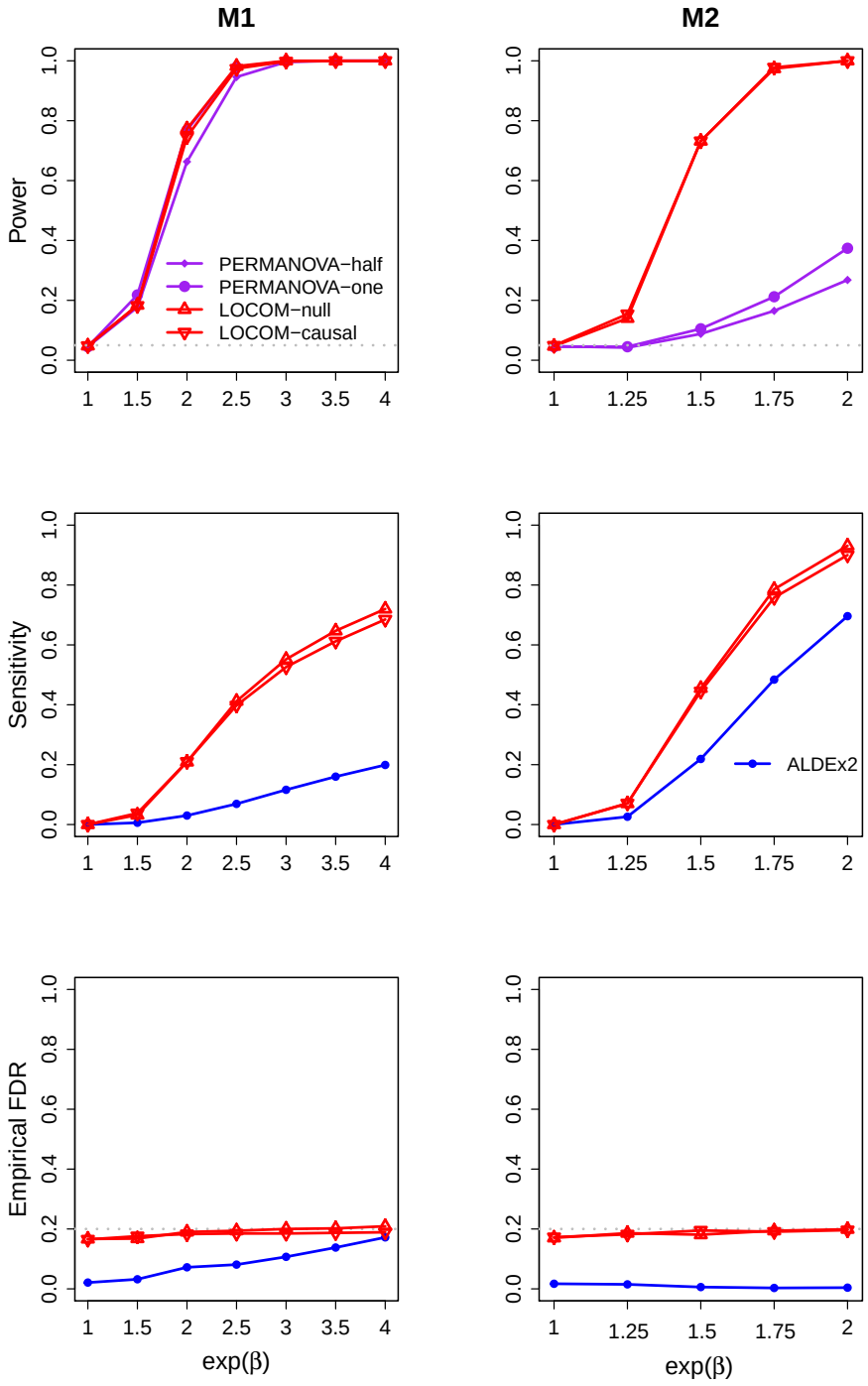


Figure 2.4: Simulation results for data ($n = 100$) with a continuous trait and a binary confounder.

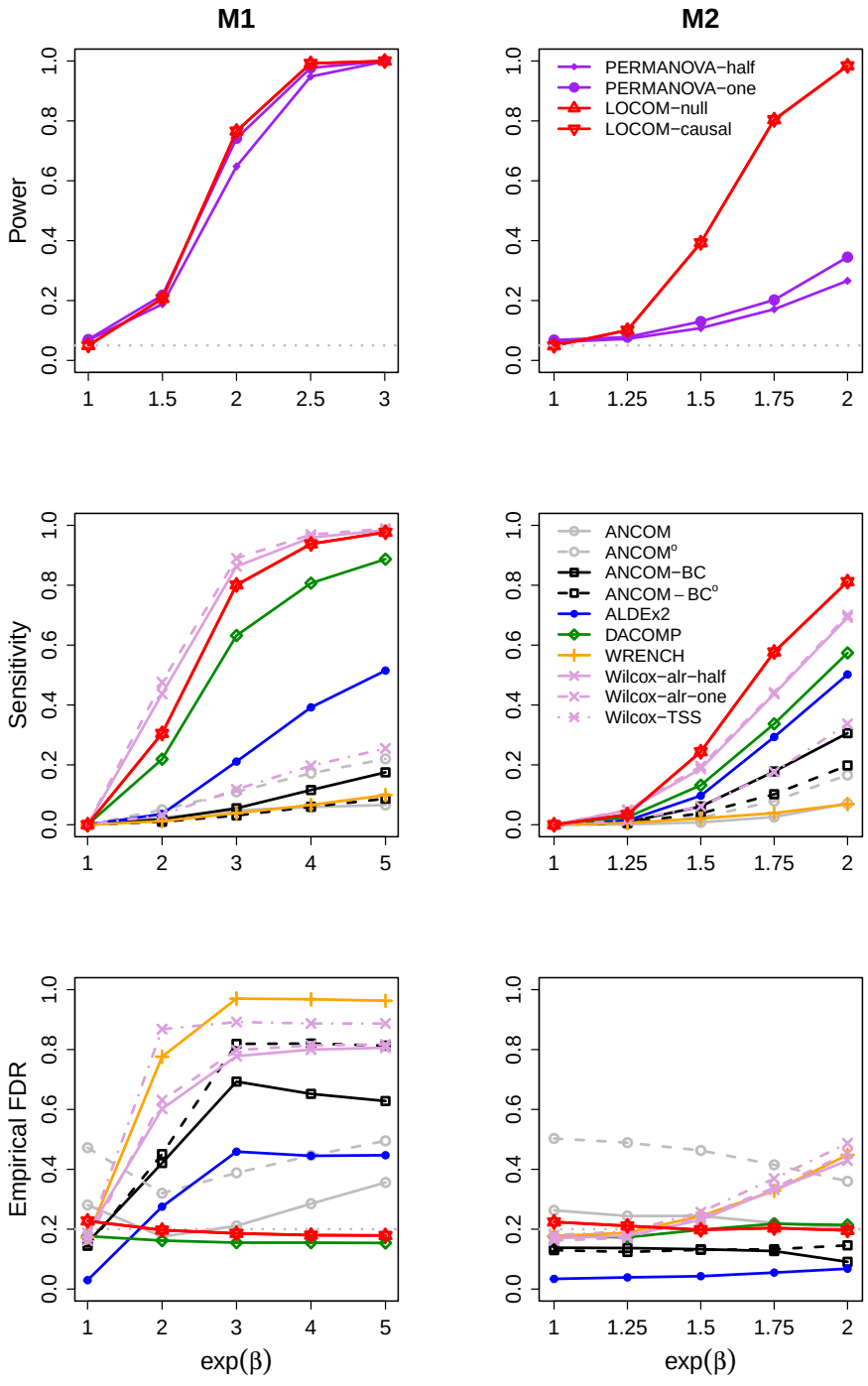


Figure 2.5: Simulation results for data ($n = 100$) with differential experimental bias in the binary-trait setting (no confounder).

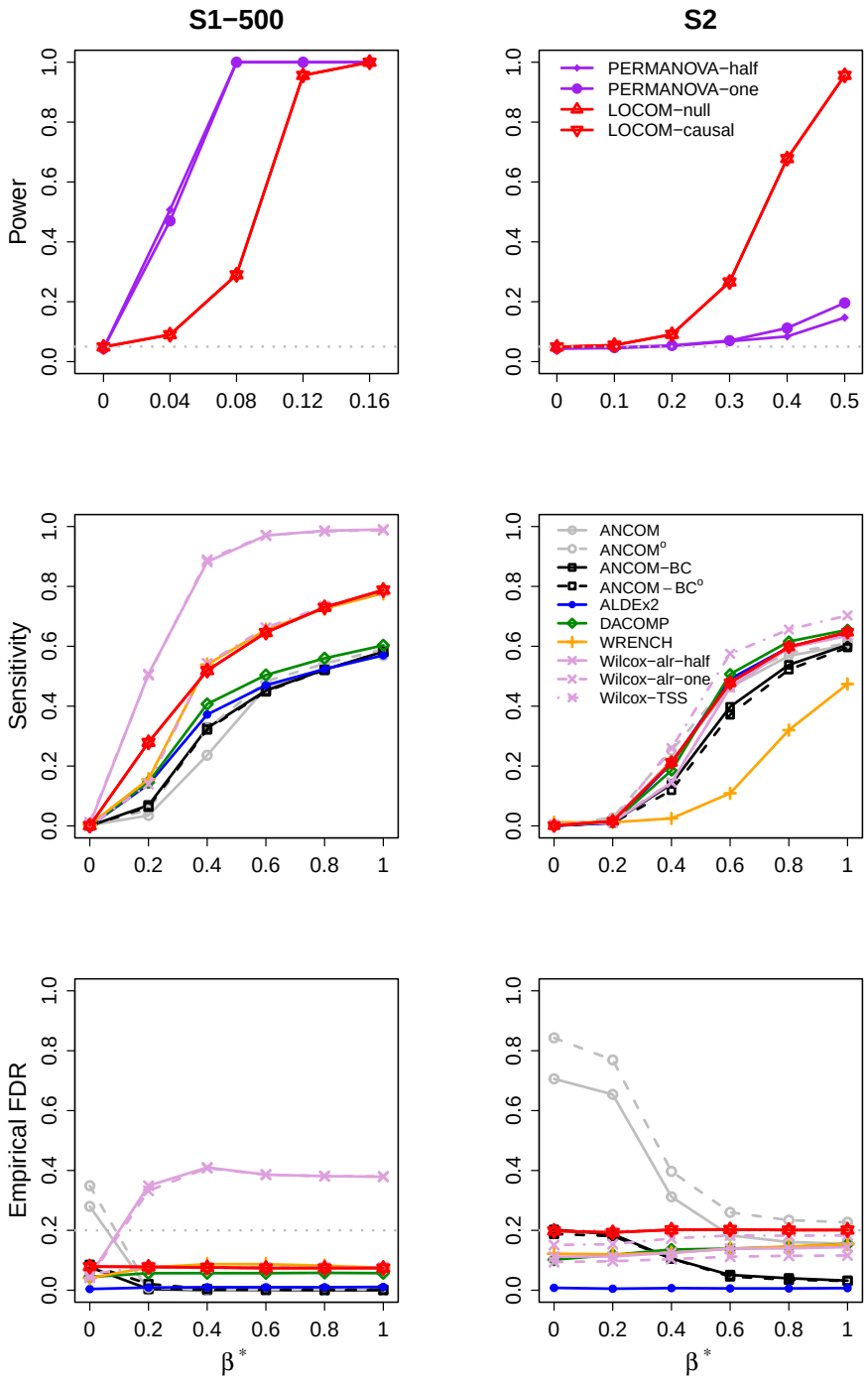


Figure 2.6: Simulation results for data ($n = 100$) generated from the differential relative abundance model and the PLNM model in the binary-trait setting (no confounder). Here β^* corresponds to the effect size β used in the LDM paper (Hu and Satten, 2020); S1-500 and S2 correspond to scenarios S1 and S2 in the LDM paper, except that in S1-500 there are 500 causal taxa.

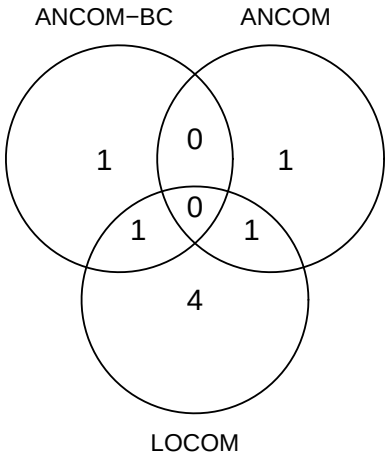


Figure 2.7: Taxa detected to be differentially abundant in the URT data.

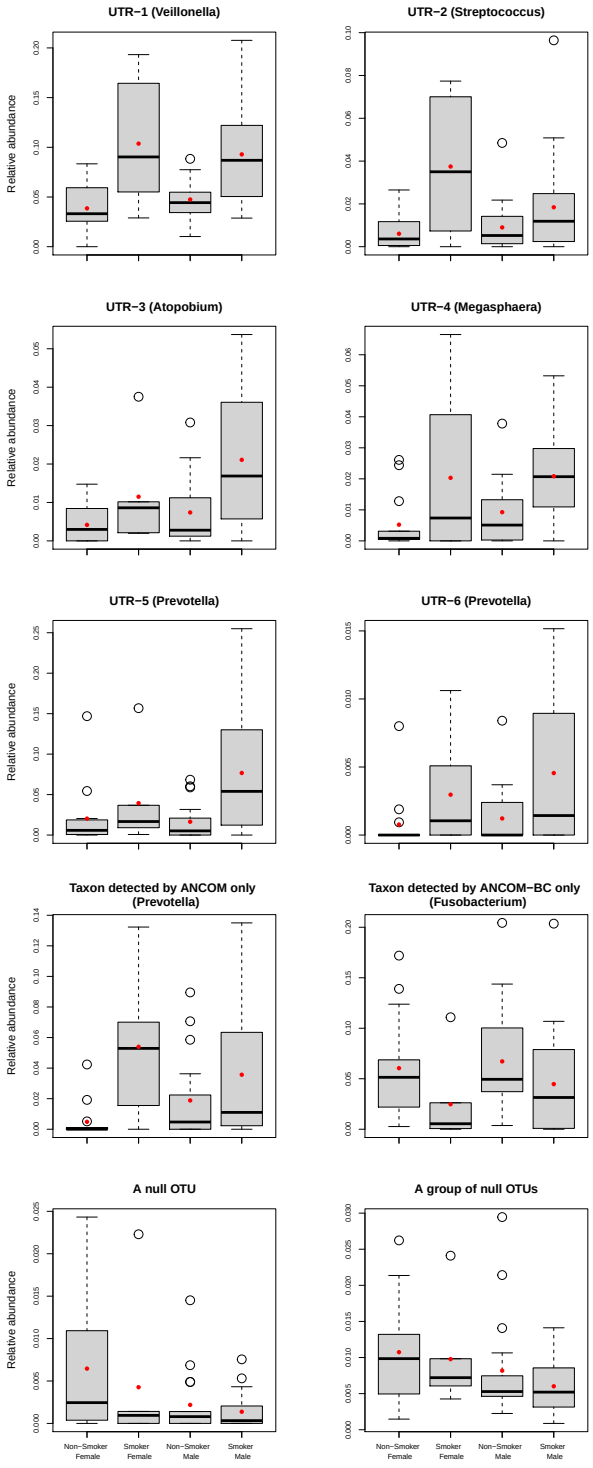


Figure 2.8: Distributions of relative abundances for taxa in the URT data. The red dots represent the means. The six taxa in rows 1-3 were detected by LOCOM; among these, URT-1 was also detected by ANCOM-BC and URT-5 was also detected by ANCOM. In the last row, "A null taxon" corresponds to the taxon (*Shigella*) with the median $\widehat{\beta}_{j,1}$ value. "A group of null taxa" include the taxon with the median $\widehat{\beta}_{j,1}$ value and 20 taxa with $\widehat{\beta}_{j,1}$ values closest to (10 less than and 10 greater than) the median; their relative abundances were averaged.

Chapter 3

Impact of experimental bias on compositional analysis of microbiome data

3.1 Introduction

High-throughput 16S rRNA gene sequencing and shotgun metagenomics sequencing provides an unprecedented opportunity to discover microbial taxa associated with traits such as clinical outcomes or environmental factors. Read count data for microbial taxa from either sequencing platform are typically summarized in a taxa count (or feature) table. It is well acknowledged that the microbiome data are compositional, because the total read count per sample is an experimental artifact and only the relative abundances of taxa, not absolute abundances, can be measured. In addition, the microbiome data are sparse (having 50–90% zero counts in a taxa count table), high-dimensional (having hundreds to thousands of taxa), and highly overdispersed (having much more variability than Poisson count data). Commonly used statistical methods for analyzing microbiome data, especially methods for testing differential abundance of individual taxa (Paulson et al., 2013; Fernandes et al., 2014; Mandal et al., 2015; Hu and Satten, 2020), have generally taken these data features into account.

An important feature in the microbiome data that has often been ignored in the development of statistical methods is experimental bias. The data generated from either 16S rRNA gene sequencing or shotgun metagenomics sequencing are subject to experimental bias, which is introduced in every step of the experimental workflow, including DNA extraction, PCR amplification, DNA sequencing, and bioinformatics processing. Each step preferentially measures (i.e., extracts, amplifies, sequences, and bioinformatically identifies) some taxa over others (McLaren et al., 2019; Brooks, 2016; Hugerth and Andersson, 2017; Pollock et al., 2018). For example, DNA extraction generates more DNA yields for bacterial species (e.g., Gram-negative) that are more easily lysed Costea et al. (2017).

McLaren, Willis and Callahan (McLaren et al., 2019) proposed a model, which we refer to as the MWC model here, for how the experimental bias affects the measured taxonomic profiles. In their model, the measured relative abundance of each taxon is a product of the taxon's true relative abundance and a taxon-specific bias factor, normalized over all

taxa observed in the sample. Each taxon-specific bias factor represents the accumulation of multiplicative biases over all the steps in the experimental pipeline, so that multiple sources of bias are described by a single factor. The MWC model has two important properties. First, the bias factor at one taxon is independent of the existence or abundance of the other taxa, i.e., there are no taxon-taxon interactions. Second, after normalization, the magnitude and even direction of the difference of the measured and true relative abundances does depend on the other taxa and are sample-specific; thus, analyses based on changes in taxon relative abundances are invalid, whereas analyses based on changes in ratios of pairs of taxa are insensitive to experimental bias. McLaren, Willis and Callahan (McLaren et al., 2019) validated this model using two mock community datasets, one from a 16S rRNA gene sequencing study (Brooks, 2016) and one from a shotgun metagenomics sequencing study (Costea et al., 2017), although their validation was limited to graphical demonstration.

Motivated by the MWC model, we previously developed LOCOM (logistic regression for compositional analysis) (Hu, Satten and Hu, 2021a), a logistic regression model for testing differential abundance that is based on changes in pairwise taxon ratios. We demonstrated, both analytically and numerically, that LOCOM is robust to the experimental bias that follows the MWC model. Since LOCOM tests the change in taxon ratios against the change in the trait of interest, it falls into the scope of *compositional analysis* (Gloor et al., 2017). Compositional analysis addresses the compositional effects that a change in the relative abundance of one taxon necessarily results in counterbalancing changes in all other taxa, and thus bases the analysis on taxon ratios or normalized count data. Several other methods, including ANCOM (Mandal et al., 2015), ANCOM-BC (Lin and Peddada, 2020), fastANCOM (Zhou, Wang, Zhao and Wang, 2022), ALDEx2 (Fernandes et al., 2014), WRENCH (Kumar et al., 2018), DACOMP (Brill et al., 2019), and LinDA(Zhou, He, Chen and Zhang, 2022), have been developed for the compositional analysis of differential abundance. ANCOM forms pairwise taxon ratios of raw count data after adding a pseudocount to all count data; fastANCOM is a fast implementation of ANCOM and also improved ANCOM by providing

p-values for the tests of differential abundance. ALDEx2 first draws Monte-Carlo samples of non-zero relative abundances from Dirichlet distributions and employs a different type of taxon ratio, centered log-ratio (clr), that uses the geometric mean of all sampled relative abundances as the reference. LinDA fits linear regression models to the clr-transformed raw count data after filling zero count data by the pseudocount or imputation approach adaptively. ANCOM-BC, DACOMP, and WRENCH all adopt some normalization techniques to account for the compositional effects. A natural question arised as to whether these methods are robust to the experimental bias that follows the MWC model, as LOCOM is. We found, using the same simulated data as for evaluating LOCOM, that DACOMP was robust to this type of bias, while ANCOM, ANCOM-BC, ALDEx2, and WRENCH were not (Hu, Satten and Hu, 2021a). However, fastANCOM and LinDA were not included in that evaluation and their performance remain unknown.

The MWC model assumes no taxon-taxon interactions, i.e., the presence or abundance of one taxon does not affect the bias factors of any other taxa in the sample. To be specific, we refer to the bias factor in the MWC model as the "main" bias and the interactions as the "interactive" bias. Zhao and Satten (Zhao and Satten, 2021a) expanded the MWC model to allow estimation and testing of the two types of bias. They had several important findings (see Table 4 of (Zhao and Satten, 2021a)) from analyzing the Brooks mock-community data (Brooks et al., 2015) using their model. First, there is some evidence for the existence of interactive bias, but the magnitude of the interactive bias is usually much smaller than the main bias. This is expected because the main bias is dominated by the bias in DNA extraction, which could vary by 10-fold across different taxa (Costea et al., 2017). Second, the presence of a taxon tends to uniformly suppress or promote the measurements of all other taxa, i.e., the interactions with all other taxa have the same direction (sign). For example, the taxon may compete with the other taxa for resources that are required for DNA extraction or PCR amplification, which would suppress extraction or amplification of the other taxa.

In this article, we evaluate the impact of experimental bias on compositional analysis methods in testing differential abundance of taxa, and we focus on the interactive bias. In the Methods section, we start with the MWC model that includes the main bias only and then generalize the MWC model to incorporate the interactive bias. In the Results section, we present simulation studies in which we simulated a wide range of experimental bias to evaluate the performance of LOCOM as well as other compositional analysis methods. We conclude with a discussion section.

3.2 Methodology

3.2.1 MWC model for experimental bias

Let p_{ij} denote the *measured* relative abundance (i.e., expected value of the observed relative abundance) of taxon j in sample i, π_{ij} denote the *true* relative abundance in the biological specimen, and γ_j denote the main bias factor that is taxon-specific. The MWC model relates the measured and true relative abundances through the formula

$$\log(p_{ij}) = \log(\pi_{ij}) + \gamma_j + \alpha_i \ , \tag{3.1}$$

where α_i is the sample-specific *normalization factor* that ensures the composition constraint $\sum_{j=1}^{J} p_{ij} = 1$.

Let X_i be a vector of covariates including the (possibly multiple) traits of interest that we wish to test and other (confounding) covariates that we wish to adjust for, but excluding the intercept. Following (Zhao and Satten, 2021a), we further assume that the true relative abundance π_{ij} can be described by a *baseline* relative abundance $\pi_j^{(0)}$ that would characterize the true relative abundance of taxon j for a sample having $X_i = 0$, and a term that describes how the baseline relative abundance is changed in the presence of covariates $X_i \neq 0$. Then,

we can replace (3.1) by

$$\log(p_{ij}) = \log(\pi_j^{(0)}) + \beta_j^{\mathrm{T}} X_i + \gamma_j + \alpha_i \ , \tag{3.2}$$

where β_j describes the way the true relative abundance changes with covariates X_i and is the parameter of interest for testing differential abundance in a compositional analysis. Here we abuse the notation a little and use α_i again to denote the normalization factor, although it is different from the one in (3.1).

LOCOM is based on model (3.2) and the taxon ratio p_{ij}/p_{iJ}, where J (without loss of generality) is the working reference taxon that is not required to be a null taxon but preferably the most abundant taxon. This gives a generalized logistic regression model

$$\log(p_{ij}/p_{iJ}) = \left\{\log(\pi_j^{(0)}) - \log(\pi_J^{(0)}) + \gamma_j - \gamma_J\right\} + (\beta_j - \beta_J)^{\mathrm{T}} X_i \ .$$

In this model, the intercept $\left\{\log(\pi_j^{(0)}) - \log(\pi_J^{(0)}) + \gamma_j - \gamma_J\right\}$ is treated as a nuisance parameter. Thus, there is no need to estimate the baseline relative abundance $\pi_j^{(0)}$ and the bias factor γ_j separately. The inference of LOCOM is based on the odds ratio $\log(p_{ij}/p_{iJ}) - \log(p_{i'j}/p_{i'J}) = (\beta_j - \beta_J)^{\mathrm{T}}(X_i - X_{i'})$, which is independent of the bias factors. In summary, LOCOM is robust to the experimental bias that follows the MWC model.

3.2.2 A general model for experimental bias

Let γ_{ij} denote the (total) interactive bias in the measurement of taxon j in sample i, i.e., the bias that is attributable to the interactions with the other taxa in the sample. Note that γ_{ij} is also indexed by i because the interactive bias depends on the composition of the other taxa in sample i. With the total bias $\gamma_j + \gamma_{ij}$, we rewrite p_{ij} in (3.2) as follows:

$$\log(p_{ij}) = \log(\pi_j^{(0)}) + \beta_j^{\mathrm{T}} X_i + \gamma_j + \gamma_{ij} + \alpha_i \ . \tag{3.3}$$

We further formulate the interactive bias γ_{ij} as a linear function of the true relative abundances π_{ij^*}s for all other taxa j^*s:

$$\gamma_{ij} = \sum_{j^* \in \mathcal{A}} \gamma_{jj^*} \pi_{ij^*} \ , \tag{3.4}$$

where $\mathcal{A}$ denotes all taxa in the sample and γ_{jj^*} is the interaction between taxa j^* and j; in fact, γ_{jj^*} is the uni-directional effect of taxon j^* on the bias of taxon j. We set $\gamma_{jj^*} = 0$ for $j^* = j$. It is possible to allow no interaction between taxa j^* and j by setting the corresponding $\gamma_{jj^*} = 0$. Note that, it is sensible for the bias γ_{ij} to depend on the true relative abundances of the other taxa rather than the measured ones, the latter of which would result in a circular dependency. In addition, model (3.4) implies that the interactive bias caused by taxon j^* is small if taxon j^* is rare.

In a simple scenario, we obtain an analytical result on the impact of the interactive bias on LOCOM. We consider the case when X_i consists of a binary trait only, denoted as T_i, and hence β_j consists of only one component $\beta_{j,1}$. We also consider a modification of model (3.4), assuming that the interactive bias γ_{ij} depends on the true *absolute* abundance of all other taxa in the specimen to obtain $\gamma_{ij} = \sum_{j^* \in \mathcal{A}} \gamma_{jj^*} \exp\left(\beta_{j^*,1} T_i\right) \pi_{j^*}^{(0)}$. Since T_i is a 0-1 variable, we have $\exp\left(\beta_{j^*,1} T_i\right) = 1 + (\exp \beta_{j^*,1} - 1) T_i$ and re-write (3.3) to be

$$\log(p_{ij}) = \left\{ \sum_{j^* \in \mathcal{A}} \gamma_{jj^*} \pi_{j^*}^{(0)} + \log(\pi_j^{(0)}) \right\} + \left\{ \sum_{j^* \in \mathcal{A}} \gamma_{jj^*} (\exp \beta_{j^*,1} - 1) \pi_{j^*}^{(0)} + \beta_{j,1} \right\} T_i + \gamma_j + \alpha_i \ . \tag{3.5}$$

Comparing model (3.5) to model (3.2), we find that each baseline relative abundance $\pi_j^{(0)}$ is replaced by $\exp\left\{ \sum_{j^* \in \mathcal{A}} \gamma_{jj^*} \pi_{j^*}^{(0)} \right\} \pi_j^{(0)}$, which does not affect the test of differential abundance since the baseline relative abundances are treated as nuisance parameters. More importantly, the coefficient of T_i has an additional term

$$\sum_{j^* \in \mathcal{A}} \gamma_{jj^*} (\exp \beta_{j^*,1} - 1) \pi_{j^*}^{(0)} \ , \tag{3.6}$$

which has a direct impact on the test of differential abundance. Recall that LOCOM assumes that more than half of the taxa are null taxa. If the term (3.6) is constant for at least all null taxa, it can be completely removed for all null taxa by subtracting the median of all estimated coefficients of T_i, a standard procedure in LOCOM. In such a case, LOCOM can control the FDR, regardless of the magnitude of γ_{jj^*}s and $\beta_{j^*,1}$s. For the term (3.6) to be constant for null taxa, γ_{jj*}s for each j^* are required to take the same value across all null taxa js, which corresponds to the special case that taxon j^* has the same effect on the bias of all null taxa.

In general scenarios, we expect that the FDR of LOCOM may be inflated. We use simulation studies to evaluate to which extent the FDR of LOCOM is still robust to the interactive bias and identify what are the magnitude of γ_{jj^*} and $\beta_{j^*,1}$ that lead to significant FDR inflation. The analytical result above serves as a motivation for designing our simulation studies.

3.3 Simulations

3.3.1 Simulation studies

Our simulations were based on data on 856 taxa of the upper-respiratory-tract (URT) microbiome; these taxa correspond to the "OTUs" in the original report on these data by Charlson et al. (Charlson et al., 2010). We considered both binary and continuous traits of interest without any confounder, and we also considered a binary confounder when the trait was binary. We used the same two casual mechanisms that were used in the LOCOM paper (Hu, Satten and Hu, 2021b). For the first mechanism (referred to as M1), we randomly sampled 20 taxa (after excluding the most abundant taxon) whose mean relative abundances were greater than 0.005 as observed in the URT data (i.e., ranking among the top 40 most abundant taxa) to be *causal* (i.e., associated with the trait of interest). For the second mechanism (referred to as M2), we selected the top five most abundant taxa (having mean

relative abundances 0.105, 0.062, 0.054, 0.050, and 0.049) to be *causal*. For simulations with a binary confounder, we assumed that the confounder was associated with 20 taxa under M1 (10 sampled at random from the 20 causal taxa and 10 from the null taxa) and 5 taxa under M2 (2 from the 5 causal taxa and 3 from the null taxa). We introduced the experimental bias, including both the main and interactive bias, into the simulated data. We generated datasets each having 100 samples.

To be specific, we let T_i denote the trait and C_i the confounder for the ith sample. To generate a binary trait, we selected an equal number of samples with $T_i = 1$ and $T_i = 0$. When a binary confounder was present, we drew C_i from the Bernoulli distribution with probability 0.2 in samples with $T_i = 0$ and from the Bernoulli distribution with probability 0.8 in samples with $T_i = 1$. To generate a continuous trait, we sampled T_i from $U[-1, 1]$. To simulate read count data for the 856 taxa, we first sampled the baseline (when $T_i = 0$ and $C_i = 0$) relative abundances $\pi_i^{(0)} = \left(\pi_{i1}^{(0)}, \pi_{i2}^{(0)}, \ldots, \pi_{iJ}^{(0)}\right)$ of all taxa for each sample from the Dirichlet distribution $Dirichlet(\overline{\pi}, \theta)$, in which the mean parameter $\overline{\pi}$ and overdispersion parameter θ took the estimated mean and overdispersion (0.02) from ftting the Dirichlet-Multinomial model to the URT data. We formed the measured relative abundances p_{ij} for all taxa by spiking the j'th causal taxon with an $\exp(\beta_{j',1})$ fold change and the j''th confounder-associated taxon with an $\exp(\beta_{j'',2})$ fold change, then adding the experimental bias $\gamma_j + \gamma_{ij}$, and finally re-normalizing the relative abundances, so that

$$p_{ij} = \frac{\exp\left(\gamma_j + \gamma_{ij} + \beta_{j,1}T_i + \beta_{j,2}C_i\right)\pi_{ij}^{(0)}}{\sum_{j'=1}^{J}\exp\left(\gamma_{j'} + \gamma_{ij'} + \beta_{j',1}T_i + \beta_{j',2}C_i\right)\pi_{ij'}^{(0)}} \ .$$

Note that $\beta_{j,1} = 0$ for null taxa, and $\beta_{j,2} = 0$ for confounder-independent taxa. For simplicity, we set $\beta_{j,1} = \beta$ for all causal taxa, and thus β is a single parameter that we refer to as the effect size; $\exp(\beta)$ is referred to as the fold change. We generated the main bias γ_j from $N(0, 0.8^2)$, which corresponds to a range between 0.2 and 5 for most (95%) fold changes $(\exp\gamma_j)$ by the main bias. We considered several scenarios for generating the interactive

bias γ_{ij}, which will be described below. Finally, we generated the taxon count data for each sample using the Multinomial model with mean $p_i = (p_{i1}, p_{i2}, \ldots, p_{iJ})$ and library size sampled from $N(10000, (10000/3)^2)$ and left-truncated at 2000.

We adopted the following model for generating the interaction between taxa j (whose bias is being altered) and j^* (which causes the bias in taxon j):

$$\gamma_{jj^*} = \phi z_{j^*} \epsilon_{jj^*} |\gamma_j|,$$

in which $|.|$ is the absolute value function, ϵ_{jj^*} is a non-negative perturbation term that has mean one, z_{j^*} takes values 1 or -1 that is independent of ϵ_{jj^*} and represents the direction of the interaction, and ϕ is a non-negative constant reflecting the magnitude of the interaction relative to the main bias, as $|\mathrm{E}(\gamma_{jj^*}|\phi, \gamma_j)| = \phi|\gamma_j|\mathrm{E}(\epsilon_{jj^*}) = \phi|\gamma_j|$. From the findings by Zhao and Satten (Zhao and Satten, 2021a), z_{j^*} was determined from the simulated main bias γ_{j^*} by $z_{j^*} = -\mathrm{sign}(\gamma_{j^*})$.

To motivate our choice for the values of ϕ, we considered the main and interactive bias for the seven taxa in the Brooks data that were estimated by Zhao and Satten (Zhao and Satten, 2021a) and presented in their Table 4. Note that their interactive bias depends on the presence-absence statuses (0/1) of the other taxa, rather than the true relative abundance as we assumed in (3.4), so the "interaction effect" in their Table 4 corresponds to $\gamma_{jj^*}\pi_{ij^*}$ in our model. Because the Brooks data are dominated by community samples that have three taxa with equal proportions, i.e., $\pi_{ij^*} = 1/3$, and the mean of the "interaction effect"-"main effect" ratios in Table 4 of (Zhao and Satten, 2021a) is 0.204, we obtained the mean of γ_{jj^*}/γ_j using our model to be 0.612, which can be viewed as an estimate of ϕ based on the Brooks data. In our simulations, we varied the value of ϕ from 0 to 4, which well covered the value 0.612 but also extended to a much wider range to explore scenarios not covered by the Brooks data.

We considered four scenarios for generating the perturbation term ϵ_{jj^*}. In the first

scenario, referred to as S-nondiff, we sampled $\epsilon_{jj^*}/2$ from $N(0.5, 0.1^2)$ for all taxa js and j^*s. In the second scenario, referred to as S-diff-causal, we modified S-nondiff to sample $\epsilon_{jj^*}/2$ from $Beta(0.5, 0.5)$ for causal taxa js. Both distributions have mean 0.5; however, $N(0.5, 0.1^2)$ has one mode at 0.5 and $Beta(0.5, 0.5)$ has two modes 0 and 1, resulting in differential variability of ϵ_{jj^*} between null and causal taxa. In the third scenario, referred to as S-diff-half, we used one of these distributions for half of randomly selected taxa js and the other distribution for the remaining half of taxa. Finally, we considered a "null" scenario, referred to as S-null, in which we set $\epsilon_{jj^*} = 0$ for null taxa js and generated ϵ_{jj^*} in the same way as in S-nondiff for causal taxa js. In Figures B.1–B.2, we displayed the distribution of the total interactive bias γ_{ij} across samples for each taxon j (that passed our filter of taxa) and contrasted the distribution to the main bias γ_j, using one replicate of simulated data in each scenario. As expected, in all scenarios and for all taxa js, the mean of γ_{ij} is approximately zero due to the average of contributions with different directionalities, and the variability of γ_{ij} increases as γ_j increases. S-nondiff and S-diff-causal have the same γ_{ij} values at the null taxa, and only differ at the causal taxa. In S-diff-half, the γ_{ij} distributions for the second half of taxa show larger variability than the first half of taxa. In S-null, the null taxa have no interactive bias.

Prior to analysis, taxa that were present in fewer than 20% of samples were filtered out in each simulated dataset. Then, we applied LOCOM using the most abundant taxon as the reference taxon. We also applied ANCOM, ANCOM-BC, fastANCOM, ALDEx2, LinDA, DACOMP (v1.26), WRENCH, Wilcox-alr-half, and Wilcox-alr-one. The latter two are pseudocount-based methods that first add pseudocount 0.5 and 1, respectively, to all count data, then perform the Wilcoxon rank-sum test at individual additive log ratios with the most abundant null taxon (known in simulated data) as the reference, and apply the Benjamini-Hochberg (Benjamini and Hochberg, 1995a) procedure to correct for multiple testing. Note that fastANCOM is applicable to data with continuous traits while ANCOM and ANCOM-BC are not; neither are WRENCH, Wilcox-alr-half, and Wilcox-alr-one. In

addition, DACOMP, WRENCH, Wilcox-alr-half, and Wilcox-alr-one cannot adjust for confounders. For ANCOM, ANCOM-BC, fastANCOM, and LinDA, we also considered their own filtering criterion with 10% presence as the cutoff and denoted them as ANCOM^o, ANCOM-BC^o, fastANCOM^o, and LinDA^o, respectively. The empirical FDR and sensitivity (proportion of truly causal taxa that were detected) were evaluated at nominal level 20% based on 1000 replicates of data. A relatively high nominal FDR level was chosen due to the small numbers of causal taxa in both M1 and M2.

3.3.2 Simulations results

Figures 3.1–3.4 display the results of empirical FDR and sensitivity under scenarios M1 and M2, for the case of a binary trait without any confounder. Each figure shows results along increasing magnitude of the interactive bias ϕ and at three effect sizes β. Note that $\phi = 0$ corresponds to the scenario with no interactive bias (i.e., the MWC model), and $\exp(\beta) = 1$ corresponds to the global null (i.e., having no differential abundant taxa). We first describe the results of LOCOM. Under the global null (the first row of each figure), the FDR of LOCOM was controlled regardless of the magnitude and distribution of the interactive bias. In S-null (the last column of each figure), the FDR of LOCOM was never inflated regardless of the values of ϕ and β. In all other scenarios, the FDR of LOCOM stayed at the nominal level when ϕ was less than 1, which covers the empirical value (0.612) observed in the Brooks data. It was only when both ϕ and β became (unrealistically) large that we observed significant inflation in the FDR of LOCOM. In both M1 and M2, the inflation was similar in S-nondiff and S-diff-causal, which had the same interactive bias at the null taxa; the inflation was the largest in S-diff-half, which had the largest variability in the interactive bias at the null taxa. The sensitivity of LOCOM decreased as ϕ increased, even in S-null. The seemingly less drop of sensitivity in S-diff-half compared to S-diff-causal and S-null is likely attributable to the inflated FDR in S-diff-half.

The FDR and sensitivity of the other methods showed a similar trend as LOCOM, as ϕ

increased. In fact, these methods had very different performance even when there was no interactive bias ($\phi = 0$). In both M1 and M2, ANCOM^o had inflated FDR when β was small, and LinDA^o, WRENCH, Wilcox-alr-half, and Wilcox-alr-one had inflated FDR when β was large; in M1 but not in M2, ALDEx2, ANCOM-BC^o, and fastANCOM^o had inflated FDR when β was large. These findings are consistent with those in the LOCOM paper. The only exception is DACOMP. We used the latest version of DACOMP (v1.26) here and an older version (v1.1) in the LOCOM paper. Unlike its satisfactory performance in the LOCOM paper, DACOMP had high FDR and low sensitivity in M2 here; this is because DACOMP incorrectly selected some causal taxa as their reference taxa, which are required to be null taxa. In general, ANCOM, ANCOM-BC, fastANCOM and LinDA had better FDR control when a more stringent filtering criterion was applied.

The results for simulated data with a binary trait and a binary confounder were summarized in Figures B.3–B.6; the results for simulated data with a continuous trait only were summarized in Figures B.7–B.10. These results showed similar patterns as those for data with a binary trait only (Figures 3.1–3.4).

3.4 Discussion

In this article, we considered the experimental bias that is not only due to taxon-specific attributes (main bias) but also caused by interactions with other taxa (interactive bias). We used extensive simulation studies to evaluate the impact of the experimental bias, focusing on the interactive bias, on LOCOM as well as on other compositional analysis methods. We found that LOCOM was robust to any main bias and a reasonable range of interactive bias. Other than LOCOM, the other methods tended to have inflated FDR even when there was only main bias. LOCOM maintained the highest sensitivity among all methods even when the other methods cannot control the FDR. Therefore, LOCOM seems to be the best choice for compositional analysis.

The robust performance of all methods to the interactive bias are likely due to two reasons. First, our model (3.4) implies that each interactive bias $\gamma_{jj^*}\pi_{ij^*}$ (that taxon j^* exerts on taxon j) is small because π_{ij^*} is generally small. Second, the contributions from all taxa to the total interactive bias γ_{ij} of taxon j have different directions and tend to cancel out to a large extend.

We used the Brooks mock community data to motivate our simulation studies, which may have two limitations. First, the Brooks data contained at most seven taxa in their samples. More mock community datasets, especially datasets that contain a large number of taxa and that mimick the real microbiome data, should be used to study the interactive bias. Second, it has been found that bacteria in a real microbial community may have different interactions from a mock community to affect the experimental bias. In summary, study of the experimental bias and its consequence in downstream analysis continue to be an important topic in the foreseeable future.

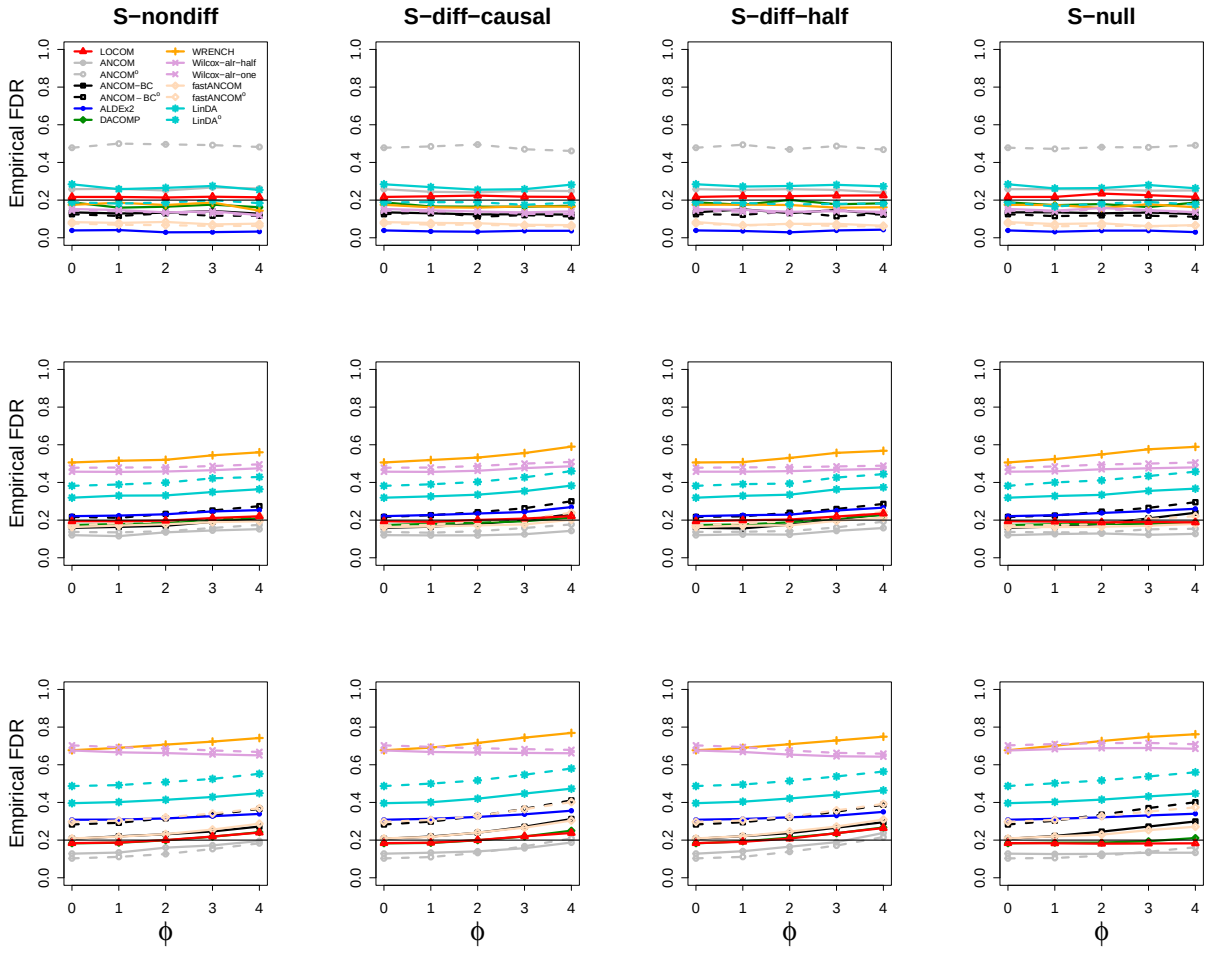


Figure 3.1: Empirical FDR results for data generated under M1. The rows correspond to fold change $\exp(\beta) = 1$, $\exp(\beta) = 3$ and $\exp(\beta) = 5$, respectively.

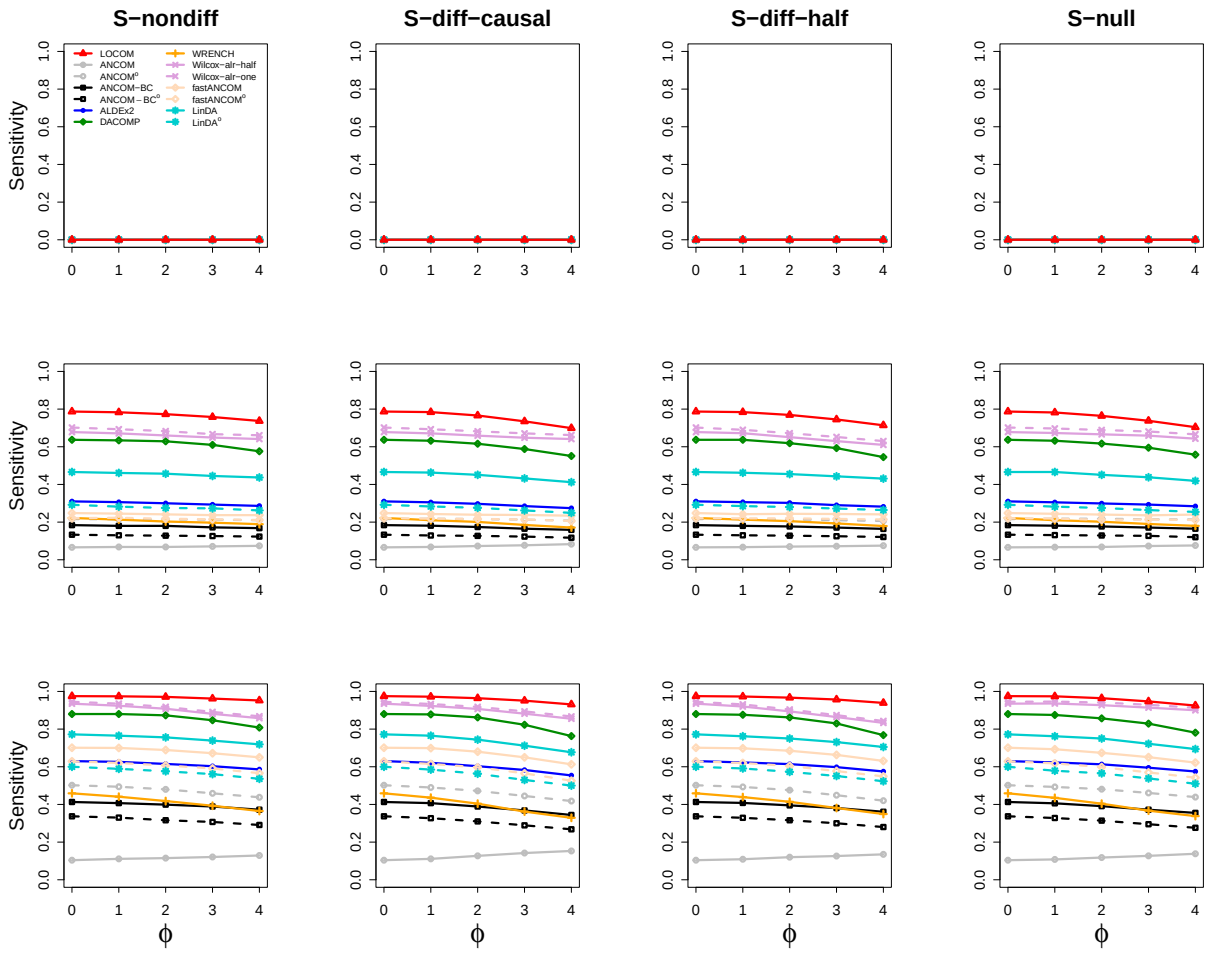


Figure 3.2: Sensitivity results for data generated under M1. The rows correspond to fold change $\exp(\beta) = 1$, $\exp(\beta) = 3$ and $\exp(\beta) = 5$, respectively.

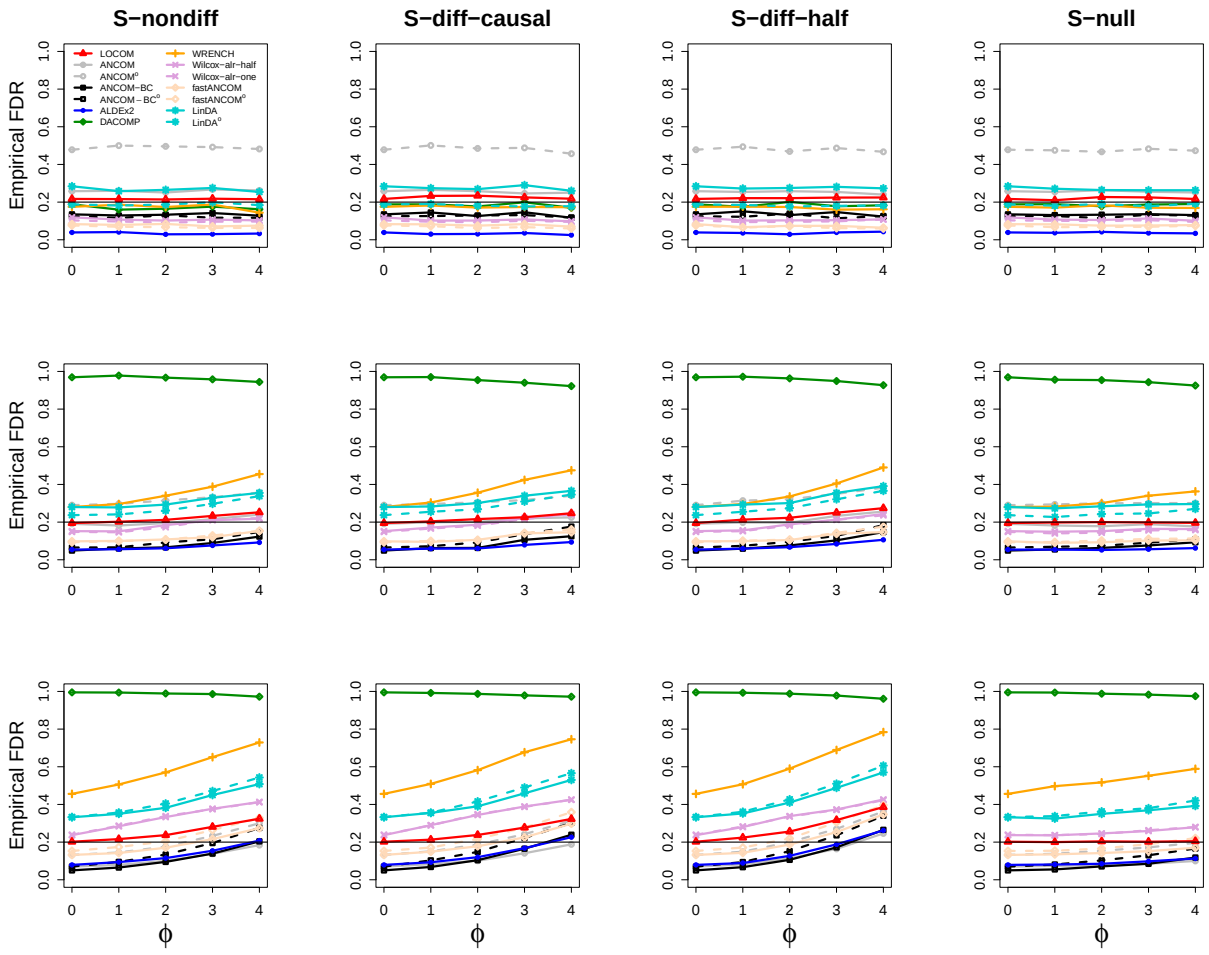


Figure 3.3: Empirical FDR results for data generated under M2. The rows correspond to fold change $\exp(\beta) = 1$, $\exp(\beta) = 2$ and $\exp(\beta) = 3$, respectively.

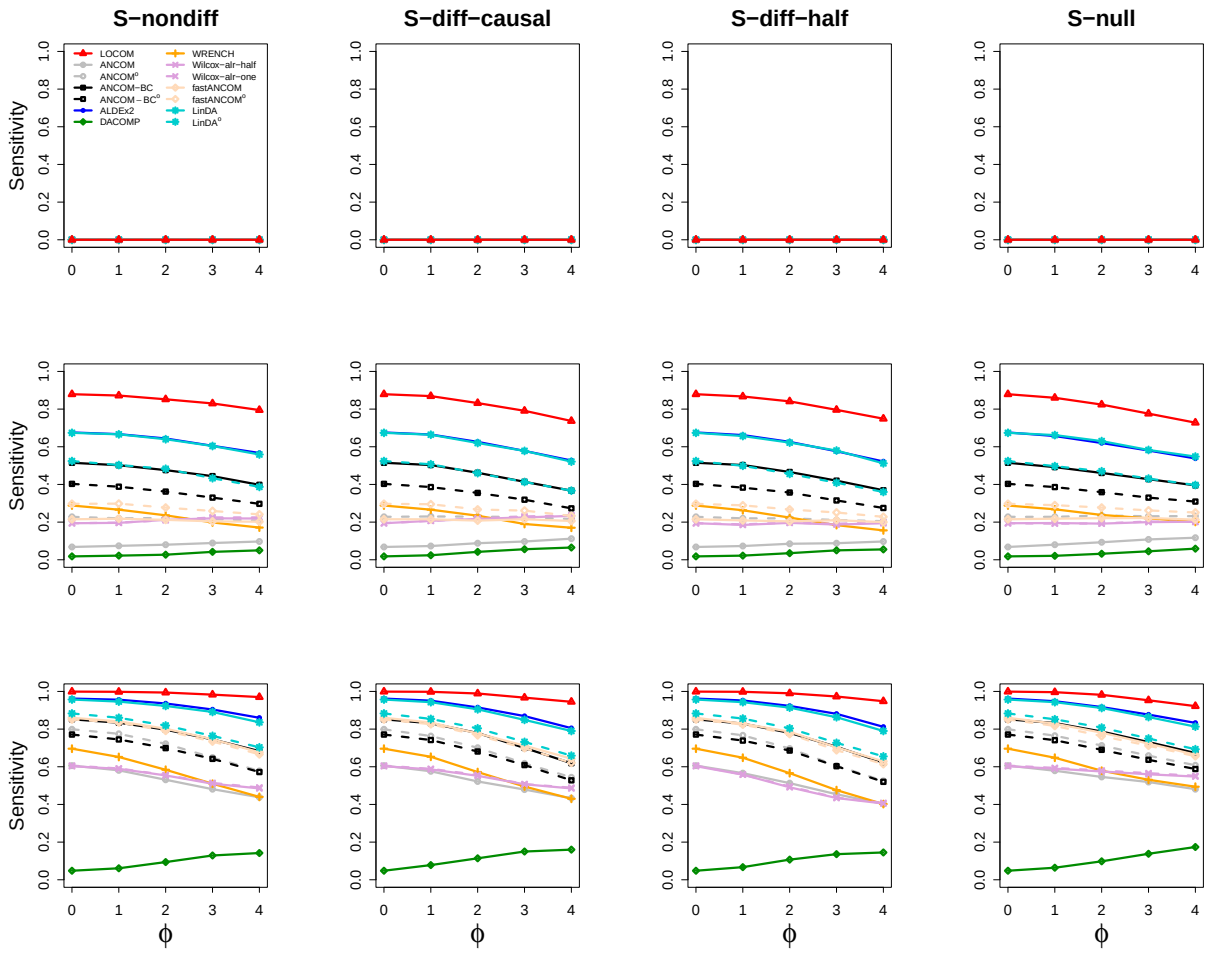


Figure 3.4: Sensitivity results for data generated under M2. The rows correspond to fold change $\exp(\beta) = 1$, $\exp(\beta) = 2$ and $\exp(\beta) = 3$, respectively.

Chapter 4

Testing microbiome associations with survival times at both the community and individual taxon levels

4.1 Introduction

Advances in sequecing technologies for profiling human microbiomes have led to the discovery of numerous microbiome associations with clinical responses (Gopalakrishnan et al., 2018; Routy et al., 2018; Matson et al., 2018). These successes suggest that microbial taxa may be useful as biomakers for disease prognosis, or targets for therapeutic interventions (Veziant et al., 2021) . For example, the miCARE study is attempting to find whether the gut microbiome can be used to predict colorectal cancer recurrence (Principle Inversitagor: Dr. Veronika Fedirko, personal communication). Like the miCARE study, studies conducted to establish these links would collect the subjects' times to an event of interest (i.e., survival times) as the outcomes some of which may have censored values. For the success of this research, finding microbiome associations with the survival outcomes only at the community level may be less importatnt than finding associations with individual taxa (we use "taxon" generically to refer to any feature such as amplicon sequence variants or any other taxonomic or functional grouping of bacterial sequences.)

However, data form micorbiome association studies can be difficult to analyze, because the taxa count may have hundreds to thousands of taxa and 50-90% zero counts, and are typically highly overdispersed. In addition, there generally exist confounders, such as previous treatment history or current medications, that correlate with both the microbiome composition and the survival outcome and so must be properly adjusted for. Finally, the sample size in a microbiome association study is typically not large and the event rate may be low, especially for rare diseases such as cancers. Analysis methods that cannot account for these data complexities will typically not yield robust and clinically meaningful results.

Two methods have been proposed specifically for testing association between the microbiome and survival outcomes: MiRKAT-S (Plantinga et al., 2017) and OMiSA(Koh et al., 2018). Unfortunately, both methods are restricted to community-level (global) association tests. While OMiSA does allow testing pre-determined sets of taxa such as taxonomic classes, it requires each set to be comprised of multiple taxa. As a result, neither MiRKAT-S nor

OMiSA can be used to find individual taxa that can act as biomarkers. A third, ad hoc, approach is to apply the Cox proportional hazard model (Cox, 1972) in a taxon-by-taxon manner (Salosensaari et al., 2021; Han et al., 2014). However, the performance of this approach has not been formally evaluated in the microbiome context, although it is known that small sample sizes and sparse count data may lead to inflated type I error when using the Cox model [10, 11]. Unfortunately, permutation-based inference, which might improve the performance of the ad hoc approach, is difficult for survival outcomes.

We previously proposed the linear decomposition model (LDM) (Hu and Satten, 2020) for testing microbiome associations with continuous or categorical (including binary) outcomes, which not only performs the test at the community level but also at the individual taxon level with false discovery rate (FDR) control. Here, we extend the LDM to survival outcomes, in order to allow a unified test framework to test both community-level and taxon-specific associations for survival outcomes. The LDM is based on a linear model that regresses that microbial data at each taxon on the (confounding) covariates that we wish to adjust for and the outcome variable(s) that we wish to test. Inference is based on permutation on circumvent making parametric assumptions about the distribution of the taxon-level data. In addition, the LDM is highly versatile: it can analyze the taxon-level data at the relative abundance scale, the arcsin-root-transformed relative abundance scale (which is variance-stabilizing for Multinomial and Dirichlet-Multinomial count data) or any other transformation, as well as the presence-absence scale (Hu, Lane and Satten, 2021), and can also accommodate clustered samples (Hu and Satten, 2020; Zhu et al., 2021) .

Our extension of the LDM was motivated by ideas developed in MiRKAT-S and OMiSA. Both of these tests first fit a Cox model to account for the relationship between any fixed covariates (excluding microbiome variables) and survival times. Then, using a random-effects framework, the variance-covariance matrix of the (Martingale) residuals from the Cox model are compared to a between-sample distance matrix calculated using the microbiome data; the similarity between these two matrices indicates the extent of association between the

microbiome and the survival outcome. MiRKAT-S allows an arbitrary distance matrix, most commonly, the Bray-Curtis or Jaccard distance matrix. OMiSA extends MiRKAT-S by using a family of power transformations of the relative abundance data to weigh abundant and rare taxa differently but calculating the MiRKAT-S test statistic based on the Euclidean distance only. Our generalization of the LDM to survival outcomes is also based on obtaining residuals from the Cox model; however, we use these residuals as covariates in the LDM to directly assess the association between the microbiome and the survival outcome. In this way, we are able to use the LDM to test both community-level and taxon-level associations with a survival outcome. In a similar manner, we also extend PERMANOVA (McArdle and Anderson, 2001), the most commonly used method for testing microbiome associations, to handle survival outcomes, although the test is at the community level and distance-based like MiRKAT-S.

The rest of this paper is organized as follows. In the Methods section, we first describe our tests based on the Martingale residuals, showing their connection to MiRKAT-S, OMiSA, and the taxon-by-taxon Cox regression. Then we extend the tests to use the deviance residuals, which are transformations of the Martingale residuals that are more symmetric above zero, and then construct combination tests that combine the results from tests using the two types of residuals. In the Results section, we first present simulation studies and then an application of all methods to data on acute graft-versus-host disease (aGVHD) after allogeneic blood or marrow transplantation (Jenq et al., 2015). We conclude with a brief discussion section.

4.2 Methodology

Suppose that, for n unrelated subjects, we have data on the time to an event of interest (e.g., disease onset or relapse) that may be subject to random censoring. For $i = 1, 2, \ldots, n$, let T_i be the (underlying) time to event for the ith subject and C_i be the corresponding

censoring time. Instead of observing T_i and C_i, we only observed the time $U_i = min(T_i, C_i)$ and the indicator $\delta_i = I(T_i \leq C_i)$ that indicates whether U_i corresponds to the event or to censoring. Further, let X_i be a set of possibly confounding covariates, which does not include the intercept. For $j = 1, 2, \ldots, J$, let Z_{ij} denote the microbiome data on taxon j from subject i, which can be the relative abundance, arcsin-root-transformed relative abundance, presence-absence status, or any (e.g, additive or centered) log-ratio transformed data. Following the conventions used in the LDM, we assume that both X_i and Z_{ij} are centered to have mean zero, i.e., $\sum_{i=1}^{n} X_i = 0$ and $\sum_{i=1}^{n} Z_{ij} = 0$ for any j.

Because survival times are censored, it is difficult to include them in the linear model framework used by the LDM. Following MiRKAT-S (Plantinga et al., 2017), we resolve this issue by first fitting a Cox model to the survival outcomes (U_i, δ_i) and covariate data X_i; we then use the residuals from this model as a covariate in the LDM (Hu and Satten, 2020). Because no microbiome data is used in the Cox model, the residuals should be associated with the microbiome data if the microbiome affects the survival outcome. If we use the Martingale residuals, denoted by M_i for subject i, we propose to test the association of taxon j with the Martingale residuals while adjusting for covariates X_i by using the LDM to fit the following linear model:

$$Z_{ij} = \beta_{X,j}^{T} X_i + \beta_j M_i + \epsilon_i, \tag{4.1}$$

where ϵ_i is the error term with mean zero and a constant variance (the only distributional assumption we make). Note that the Martingale residuals have the properties that $\sum_{i=1}^{n} M_i = 0$ and $\sum_{i=1}^{n} M_i X_i = 0$ (Bi et al., 2020).

To test $H_0 : \beta_j = 0$, the LDM uses an F-statistic, the numerator of which is proportional to the square of $\mathbb{U}_j$ given by

$$\mathbb{U}_j = \sum_{i=1}^{n} M_i (Z_{ij} - \widehat{\beta}_{X,j}^{\mathrm{T}} X_i) = \sum_{i=1}^{n} M_i Z_{ij},$$

where $\hat{\beta}_{X,j}$ is the lease squares estimator of $\beta_{X,j}$ under the null model of (3.1). Further, the numerator of the global test statistics for testing the global association between the Martingale residuals and the microbiome is

$$\mathbb{U}^2_{\text{global}} = \sum_{j=1}^{J} \mathbb{U}^2_j.$$

These test statistics can be used to show a connection between our approach and existing methods. First, the global statistics $\mathbb{U}^2_{\text{global}}$ agrees with the variance-component score statistics in MiRKAT-S when the Euclidean distance (the linear kernel) is used, as well as the variance-component score statistics in OMiSA (the OMiSALN part) for untransformed data. Second, letting $\lambda(.)$ denote the hazard function for a survival analysis, the taxon-specific $\mathbb{U}_j$ coincides with the score statistics for testing $\alpha_j = 0$ in the Cox model $\lambda(t; X_i, Z_{ij}) = \lambda_0(t)\exp(\alpha_X^{\mathrm{T}} X_i + \alpha_j Z_{ij})$(Bi et al., 2020), which includes both the covariates and the microbiome data from the jth taxon as explanatory variables in the hazard function. Theses connections justify the use of the Martingale residual as a covariate in the LDM.

The main advantage of our approach is that results for individual taxa are available, and that the global test statistics is a coherent combination of these taxon-specific statistics; neither MiRKAT-S nor OMiSA provide taxon-specific results. However, the LDM is based on the Euclidean distance for combining taxon-specific statistics, while MiRKAT-S can use arbitrary distances. For this reason, we also provide an extension of PERMANOVA for testing survival outcomes that can be used with arbitrary distances, at the end of this section.

An important feature of our approach is that, although the effect of X_i has been removed form M_i (i.e. M_i and X_i are uncorrelated), we still include X_i in (3.1). In the Appendix, we show how including this term allows our permutation tests to achieve higher power than the permutation tests currently available in MiRKAT-S. We further show how to obtain global tests with power similar to what we achieve using the original MiRKAT method (Zhao et al.,

2015) with the Martingale residual as a continuous outcome.

Compared with the ad hoc approach of fitting a Cox model for each taxon, our permutationbased inference is robust to small sample size, low event rate, and sparse count data, while the Cox model is known to have inflated type I error in these situations (Chen et al., 2014; Bi et al., 2020). Compared with the ad hoc approach, both MiRKAT-S and our approach share the huge computational advantage that the Cox model only needs to be fit once. In addition, both methods only depend on the presence of an association between the Martingale residuals and the microbiome measures, and so do not depend on the correct specification of the Cox model for validity (i.e., type I error control), although power may be lost if the Cox model provides a poor fit to the data.

One deficiency of the Martingale residual is its skewness, because it has a maximum value 1 but a minimum value $-\infty$. Because a residual measure with a more normal-like distribution may perform better in downstream analyses, Therneau et al. (1990) introduced the deviance residual for a Cox model:

$$D_i = \text{sign}(M_i)\sqrt{-2\{M_i + \Delta_i \log(\Delta_i - M_i)\}},$$

which is a non-linear transformation of the Martingale residual M_i . Therneau et al. (1990) found that with less than 25% censoring, the deviance residual is approximately normally distributed; with more than 40% censoring, too many points will lie near 0 making the distribution non-normal, although the deviance residuals remain approximately symmetric about 0. Therefore, we also consider a variation of our method that replaces M_i by D_i in the linear model (3.1). Although D_i is not orthogonal to X_i , we can still use the LDM to fit (3.1) as long as X_i enters the model before D_i because, in this case, the LDM will make D_i orthogonal to X_i before testing for association with Z_{ij}. In our simulations, use of the Martingale residual sometimes gave better power and sensitivity; in other situations the deviance residual performed better. Since we cannot characterize those scenarios a priori, we

also combine the results from analyzing each residual separately into a single combination test. To account for differences in residual scale, we take the minimum of the p-values obtained from analyzing each residual separately, and use the corresponding minima of null p-values for each test from the permutation replicates to simulate the null distribution; the null p-value is calculated based on the rank of the test statistic among all permutation replicates (Westfall and Young, 1993).

We extend PERMANOVA to analyzing survival outcomes in a similar way. Like MiRKAT-S, PERMANOVA is distance-based and offers a global test of the association at the community level. To explain the variability in a given distance matrix, we use a similar linear model as in (3.1) that includes the covariates X_i and the Martingale residual M_i as explanatory variables. We obtain the p-value for testing M_i , repeat the procedure with the deviance residual D_i , and then construct a combination test that take the minimum of the two p-values as the final test statistic. A common use of PERMANOVA is through the function "adonis2" in the R package vegan. We have also presented another implementation of PERMANOVA through the function "permanovaFL" in our R package LDM (Hu and Satten, 2020), which differs from adonis2 in terms of the permutation scheme and outperformed adonis2 in many occasions (Hu and Satten, 2020; Zhu et al., 2021; Hu and Satten, 2021) .

4.3 Simulations

4.3.1 Simulation designs

We conducted simulation studies to evaluate the properties of our approach and compare our results to those obtained using competing methods. Our simulations were based on data on 856 taxa of the upper-respiratory-tract (URT) microbiome (Charlson et al., 2010) that were also used in the MiRKAT-S paper. We considered a binary confounder X_i and assumed equal numbers of subjects with $X_i = 1$ and $X_i = 0$. We randomly sampled 100 taxa to be associated with X_i and generated their associations as follows. We first set

two vectors, π_1 and π_2, to the taxon frequencies (i.e., relative abundances) estimated from the URT microbiome data, and then permuted the frequencies in π_2 that belong to the set of 100 taxa selected to be associated with X_i , which ensured the same frequencies in π_1 and π_2 for taxa not selected. Next, we defined a subject-specific frequency vector to be $\tilde{\pi}(X_i) = (1 - \beta_{XZ}X_i)\pi_1 + \beta_{XZ}X_i\pi_2$, in which β_{XZ} can be interpreted as the effect of X_i on the selected taxa. When $\beta_{XZ} = 0$, there was no association between X_i and the microbiome, in which case X_i reduced to a simple covariate for the survival outcome instead of a confounder. Finally, we generated the taxon count data for each subject using the Dirichlet-Multinomial model with mean $\tilde{\pi}(Xi)$, overdispersion 0.02, and library size sampled from $N(10000, (10000/3)2)$ and left-truncated at 1000.

We considered two models, M1 and M2, for simulating the survival outcome. In what follows, we number the taxa by decreasing relative abundance so that taxon 1 is the most abundant. In model M1, we assumed that the relative abundances of taxa 1–10 determined the association with the survival outcome; in model M2, we assumed that the presence or absence of 10 randomly selected taxa, selected from taxa 11–100, determined this association. Specifically, we defined $S_i = \sum_{j\in\mathcal{A}} \delta_j Z_{ij}/\overline{Z}_j$ under M1 and $S_i = \sum_{j\in\mathcal{A}} \delta_j \mathbb{I}(Z_{ij} > 0)$ under M2, where δ_js were directions taking values 1 and -1 with equal probabilities (and fixed across replicates of data), $\mathcal{A}$ was the set of taxa selected to be associated with the outcome, Z_{ij} was the observed frequency (taxon count divided by the library size), and $\overline{Z}_j$ was the average frequency for the jth taxon across subjects. Then, we simulated the time to event from the Cox model with the baseline hazard following the Weibull distribution $\mathcal{W}(2, 0.01)$, namely, $T_i = \left[\frac{-\log(V_i)}{0.01\exp\{\beta_{XS}\text{scale}(X_i)+\beta\text{scale}(S_i)\}}\right]^{1/2}$, where V_i was sampled from the uniform distribution $U[0, 1]$ and $B_i = exp\beta_{XS}scale(X_i) + \beta scaleS_i$ with β characterizing the effects of the "causal" taxa on the event time, β_{XS} being fixed at 0.5, and scale(.) standardizing the input vector to have mean 0 and standard deviation 1. The censoring time was simulated independently from the Exponential distribution $Exp(\mu)$, where μ was set to 0.03, 0.08, and 0.2 to achieve approximately 25%, 50% and 75% censoring. Using this procedure, we generated $n = 100$ or

50 subjects for each replicate of data. To evaluate robusteness of our methods to violation of the proportional hazard (PH) assumption, we also simulated the event time from the accelerated hazard (AH) model (Chen and Wang, 2000) with the baseline hazard following the lognormal distribution, namely, $T_i = B_i^{-1} exp\{\phi^{-1}[1 - exp(B_i log V_i)]\}$, where ϕ^{-1} is the inverse cumulative distribution of the standard normal distribution. The censoring time was simulated as before using $\mu = 0.5$ to achieve approximately 50% censoring. The AH model generated data that strongly violated the PH assumption (specifically, 28.8% rejection rate for testing the PH assumption (Grambsch and Therneau, 1994) using our simulated data, which was much higher than the nominal value 5% of the test) and even had crossing survival curves.

Prior to analysis, we filtered out taxa that were found in fewer than 5 subjects in the dataset. We used the R package Survival to obtain the Martingale and deviance residuals, M_i and D_i , from fitting the Cox model for the survival outcomes with X_i as the explanatory covariate.

For testing individual taxa, we applied the LDM with either X_i and M_i as covariates or X_i and D_i as covariates in the linear regression model (1), and refer to them as LDM-m and LDM-d, respectively. Specifically, for data generated under model M1, we applied the LDM to the relative abundance data and arcsin-root-transformed relative abundance data separately and used the omnibus test that combined their results; for data generated under model M2, we applied the LDM to the presence-absence data. We also obtained the combination test that combines the results from LDM-m and LMD-d, and refer to it as LDM-c. To evaluate the ad hoc approach, we fit the Cox model and the Firth-corrected Cox model (using the "coxphf" function in the R package coxphf) taxon by taxon, using X_i and the taxon relative abundance under model M1 or taxon presence-absence status under model M2 as covariates; the p-values for these taxon-specific tests were then adjusted for multiple testing using the Benjamini Hochberg procedure Benjamini and Hochberg (1995b). We evaluated the sensitivity and empirical FDR at nominal level 20% for all taxon-specific

tests, using 1000 replicates of data. We chose a relatively high nominal level for FDR because the numbers of "causal" taxa in both M1 and M2 were small (i.e., 10).

For testing global association, we obtained these results from LDM-m, LDM-d, and LDM-c, and we also applied permanovaFL in a similar way to obtain permanovaFL-m, permanovaFLd, and permanovaFL-c. For permanovaFL-based tests and all other distance-based tests described below, we used the Bray-Curtis distance under model M1 and the Jaccard distance under model M2. For comparison, we applied MiRKAT-S using the permutation p-value, which was based on the Martingale residual only. We also applied OMiSA, specifically OMiSALN, the part of OMiSA that combines the results from analyzing differently power-transformed relative abundance data (with the default set of power values), which always analyzes data at the relative abundance scale even under model M2. In addition, we considered a number of secondary tests to gain more insights. To verify the equivalence of MiRKAT-S to an implementation of MiRKAT, we applied MiRKAT with a linear regression model that used the Martingale residual as the continuous outcome and the microbiome profile as the covariates without adjusting for X_i , and refer to this test as MiRKAT-m1. We also applied a variation of MiRKAT-m1 that additionally adjusted X_i in the linear regression, referred to as MiRKAT-m, and a variation of MiRKAT-m that replaced the Martingale residual by the deviance residual, referred to as MiRKAT-d. Finally, we applied PERMANOVA implemented in adonis2, with either X_i and M_i as covariates or X_i and D_i as covariates to obtain adonis2-m and adonis2-d. All global tests were evaluated on their type I error and power at the nominal level 0.05, based on 10000 and 1000 replicates of data, respectively.

4.3.2 Simulation results

We focus on the results from simulated data with 50% censoring and sample size 100; the results when the censoring rate was varied to 75% or 25% or the sample size was reduced to 50 showed the same patterns and are thus deferred to Supplementary Materials (Table S1, Figures C.3, C.4, and C.5). Figure 1 shows the sensitivity and empirical FDR results

for the taxon-specific tests. In both scenarios M1 and M2, the deviance residual (LDM-d) corresponds to higher sensitivity than the Martingale residual (LDM-m), although the difference was small. We explored two more scenarios, one assuming taxon 11 to be associated with the event time (referred to as M3) and one assuming taxon 21 to be associated (referred to as M4), in which data were analyzed at the relative abundance scale and the presence-absence scale, respectively. The results were displayed in Figure C.1. We found that the Martingale residual led to higher sensitivity than the deviance residual under M3 and the two residuals performed very differently under M4. Fortunately, the combination test LDM-c tracked the results of the best-performing residual in all cases. As expected, all LDM tests controlled the FDR (except for some minor inflation when the sensitivity was extremely low). The ad hoc Cox regression had very inflated FDR in all cases. The Firth-corrected Cox model had close sensitivity, smaller while still inflated FDR compared with Cox regression.

The type I error results of the global tests are summarized in Table 1, which shows that the LDM- and permanovaFL-related tests all yielded type I error close to the nominal level 0.05. MiRKAT-S and OMiSA produced conservative type I errors when X_i was a confounder; for example, their type I error rates were 0.007 and 0.034 under model M2. Note that all these tests yielded highly inflated type I error (> 0.4) when the confounder was not adjusted for in the entire analysis, confirming that we have generated substantial confounding effects. The type I error rate of all these tests were robust to violation of the PH assumption when the event times were instead simulated using the AH model.

Figure 2 displays the power for the global tests. Using either the LDM or permanovaFL, the Martingale and deviance residuals, as well as the combination test, all led to similar power. The similar power between the LDM and permanovaFL was a coincidence here and is not guaranteed in general, since permanovaFL results will vary depending on the distance measure used. MiRKAT-S had similar power to permanovaFL-m when X_i was a simple covariate (i.e., not correlated with the microbiome data) but had lower power than permanovaFL-m when X_i was a confounder (especially under Model M2), which is consistent

with its conservative type I error results in this situation. OMiSA had very low power in both scenarios M1 and M2. We explored an additional scenario in which rare taxa (taxa 91–100) were associated with the event time; OMiSA yielded good power among all tests when the data were simulated and analyzed based on the relative abundance scale (Figure S2).

Figure 3 displays the power for the secondary global tests and included MiRKAT-S again as a calibration. Indeed, MiRKAT-S had equivalent power to MiRKAT-m1 in all cases. MIRKAT-m and MiRKAT-d always had very similar power to permanovaFL-m and permanovaFLd, respectively, which was expected given the equivalent performance of MiRKAT and permanovaFL we have consistently observed in the context of testing continuous or binary outcomes. These results confirmed that the improvement in the power of permanovaFLm over MiRKAT-S was truly due to its inclusion of X_i in the linear regression model (3.1). Lastly, adonis2-m and adonis2-d occasionally had lower power than permanovaFL-m and permanovaFL-d, as seen before (Hu and Satten, 2020; Zhu et al., 2021; Hu and Satten, 2021).

4.4 Data analysis

4.4.1 Analysis of the aGVHD data

We analyzed the same data on aGVHD that were also analyzed in the MiRKAT-S paper. We first followed the same procedure as in the MiRKAT-S paper to obtain 2436 operational taxanomic unites (OTUs) in 94 subjects, and then removed subjects with library sizes less than 1000 and excluded OTUs that were found in fewer than 5 subjects to obtaina final set of 88 subjects and 441 OTUs for our analysis. We tested the association of the gut microbiome with two survival outcomes separately, the overall survival and the time to stage-III aGVHD, both adjusting for age and gender. The censoring rates for the overall survival the and time to stage-III aGCHD were 52.3% and 42.0%, respectively. The Martingale and deviance

residuals obtained from the Cox model with age and gender as covariates were displayed in Figure S6, which shows that neither residuals were normally or symmetrically distributed in this dataset.

We applied the LDM only for testing individual OTUs and the LDM, permanovaFL, MiRKAT-S and OMiSA (the OMiSALN part only) for testing the global association. We applied these methods to both relative-abundance and presence-absence data scales, in the same way as in the simulation studies. For the presence-absence analyses, we considered both rarefied and unrarefied data for all methods. The unrarefied data may be subject to confounding by the library size, which varied considerably between 1,274 and 265,352 in this dataset. In the rarefaction-based analysis with rarefaction depth 1,274, the LDM was based on all rarefied OTU tables (the LDM-A method in Hu, Lane and Satten (2021)), and permanovaFL and MiRKAT-S were based on the expected Jaccard distance matrix over all rarefied OTU tables (Hu and Satten, 2021). Note that OMiSA was not applicable for presence-absence analysis.

All test results were summarized in Table 4.2. The LDM or permanovaFL combination tests (LDM-c, permanovaFL-c) always tracked the better results obtained using the Martingale residual and the deviance residual, so we focus on their combination tests hereafter. Among the different analyses we performed, presence-absence analyses based on all rarefied OTU tables consistently led to the most significant results for all tests. Specifically, LDM-c detected 17 OTUs associated with the overall survival and 29 OTUs associated with the time to stage-III aGVHD; the survival functions stratified by the presence and absence status of each detected OTU (based on a singly rarefied OTU table) were plotted in Figures S7 and S8, which showed a clear separation in each case. LDM-c, permanovaFL-c, and MiRKAT-S yielded p-values 0.0002, 0.0006, and 0, respectively, for testing the global association of the gut microbiome with the overall survival, and 0.0006, 0.0016, and 0.003 for the global association with the time to stage-III aGVHD. The substantial difference in results between the rarefied and unrarefied analyses implied that differences in the library size played an impor-

tant, although undesired, role in the unrarefied analysis. Based on the relative abundance data and a nominal significance level 0.05, LDM-c and permanovaFL-c declared a significant global association of the gut microbiome with the time to stage-III aGVHD but failed for the overall survival; MiRKAT-S failed for both outcomes; OMiSA was significant for both outcomes.

4.5 Discussion

We have presented an approach that can be used in the LDM and PERMANOVA frameworks to testing microbiome associations with survival outcomes. This approach is based on a linear model treating both the Martingale and deviance residuals from the Cox proportional hazards model as continuous covariates. Unlike existing methods which only give communitylevel (global) tests, our extension of the LDM gives both community-level and taxon-level association tests. Further, we find that the LDM global test and permanovaFL outperform the existing permutation-based global tests, MiRKAT-S and OMiSA, when there are strong confounders. Although the analysis of a single type of residuals can make use of existing code of the LDM or permanovaFL, the test that combines the two, which is recommended over each single test, does entail additional programming and has been added to the LDM package. Note that the only additional computational burden for testing survival outcomes in the LDM framework is the single calculation of the Cox model residuals and the calculation of the combination test, which is a negligible addition in computation.

In our simulation studies and real data application, we have analyzed the microbiome data at the relative abundance scale and the presence-absence scale separately. In theory, those p-values should be adjusted for multiple comparisons. An alternative would be to use a test based on the minimum of p-values at each scale. We have recently developed such a test for the LDM that combines the results of analyses at three different scales, namely, the relative abundance, arcsin-root transformed relative abundance, and presence-absence scales;

we call this test LDM-omni3 (Zhu et al., 2022). A version of LDM-omni3 that applies our survival-based LDM-c to analysis results on the three scales is available in the most recent release ($\geq v5.0$) of the LDM package. Similarly, an omnibus test based on permanovaFL that combines the results of permanovaFL-c applied to different distance matrices is also available in the LDM package.

Table 4.1: Type I error of the global tests for simulated data with 50% censoring and $n = 100$

Hazards			LDM			permanovaFL				
model	Scenario	β_{XZ}	-c	-m	-d	-c	-m	-d	MiRKATS	OMiSALN
Cox	M1	0	0.051	0.049	0.047	0.051	0.051	0.048	0.052	0.050
		0.8	0.050	0.047	0.048	0.052	0.051	0.050	0.032	0.034
		0.8*	0.626	0.634	0.563	0.450	0.453	0.418	0.471	0.518
	M2	0	0.044	0.042	0.044	0.046	0.046	0.044	0.048	0.050
		0.8.	0.048	0.050	0.047	0.049	0.050	0.046	0.007	0.034
		0.8*	0.805	0.808	0.74	0.814	0.817	0.74	0.818	0.518
AH	M1	0	0.050	0.049	0.047	0.051	0.051	0.048	0.052	0.050
		0.8	0.050	0.047	0.048	0.052	0.051	0.050	0.032	0.034
	M2	0	0.050	0.042	0.044	0.046	0.046	0.044	0.048	0.050
		0.8	0.052	0.050	0.047	0.049	0.050	0.046	0.007	0.034

Note: AH is the accelerated hazards model (Chen and Wang, 2000). When $\beta_{XZ} = 0$, X was a simple covariate (i.e., not correlated with the microbiome data); when $\beta_{XZ} = 0.8$, X_i was a confounder; when $\beta_{XZ} = 0.8^*$, X_i was a confounder but omitted in the entire analysis.

Table 4.2: Results in analysis of the aGVHD data

			Relative abundance	Presence-absence (unrarefied)	Presence-absence (rarefied)
Overall survival	Global	LDM-c	0.0640	0.0456	0.000200
	p-value	LDM-m	0.0565	0.0385	0.000200
		LDM-d	0.0965	0.0737	0.000400
		permanovaFL-c	0.0785	0.0376	0.000600
		permanovaFL-m	0.0665	0.0316	0.000800
		permanovaFL-d	0.132	0.0411	0.000400
		MiRKATS	0.0581	0.0290	0
		OMiSALN	0.002	-	-
	Number of	LDM-c	2	2	17
	detected	LDM-m	5	10	28
	OTUs	LDM-d	0	1	3
Stage-III aGVHD	Global	LDM-c	0.0376	0.0365	0.000600
	p-value	LDM-m	0.0315	0.0323	0.000600
		LDM-d	0.0591	0.0668	0.00180
		permanovaFL-c	0.0366	0.0411	0.00160
		permanovaFL-m	0.0310	0.0355	0.00140
		permanovaFL-d	0.0604	0.0624	0.00260
		MiRKATS	0.0711	0.0300	0.00300
		OMiSALN	0.004	-	-
	Number of	LDM-c	12	12	29
	detected	LDM-m	50	15	64
	OTUs	LDM-d	0	8	8

Note: The OTUs were detected at the nominal FDR = 10%. MiRKATS and permanovaFL results were based on the Bray-Curtis distance in analysis of relative abundance data and the Jaccard distance in analysis of presence-absence data. In rarefaction-based presence-absence analysis, the results of MiRKATS were based on one rarefied OTU table and those of LDM and permanovaFL were based on all rarefied OTU tables.

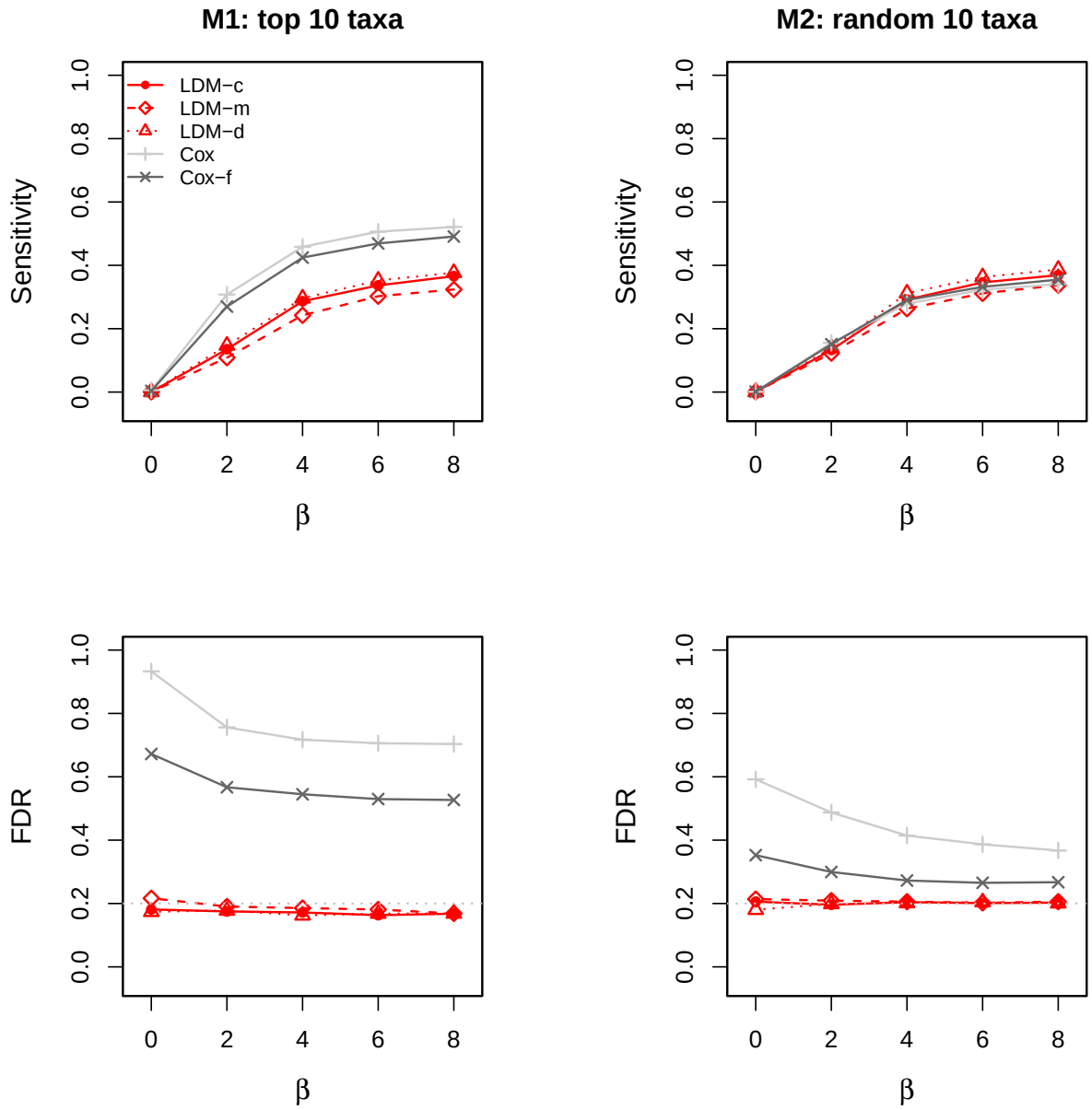


Figure 4.1: Sensitivity and empirical FDR of the taxon-specific tests in analysis of simulated data with a confounder X_i ($\beta_{XZ} = 0.8$), 50% censoring, and n = 100. "Cox-f" is the Firth-corrected Cox model. The gray dotted line represents the nominal FDR level 20%

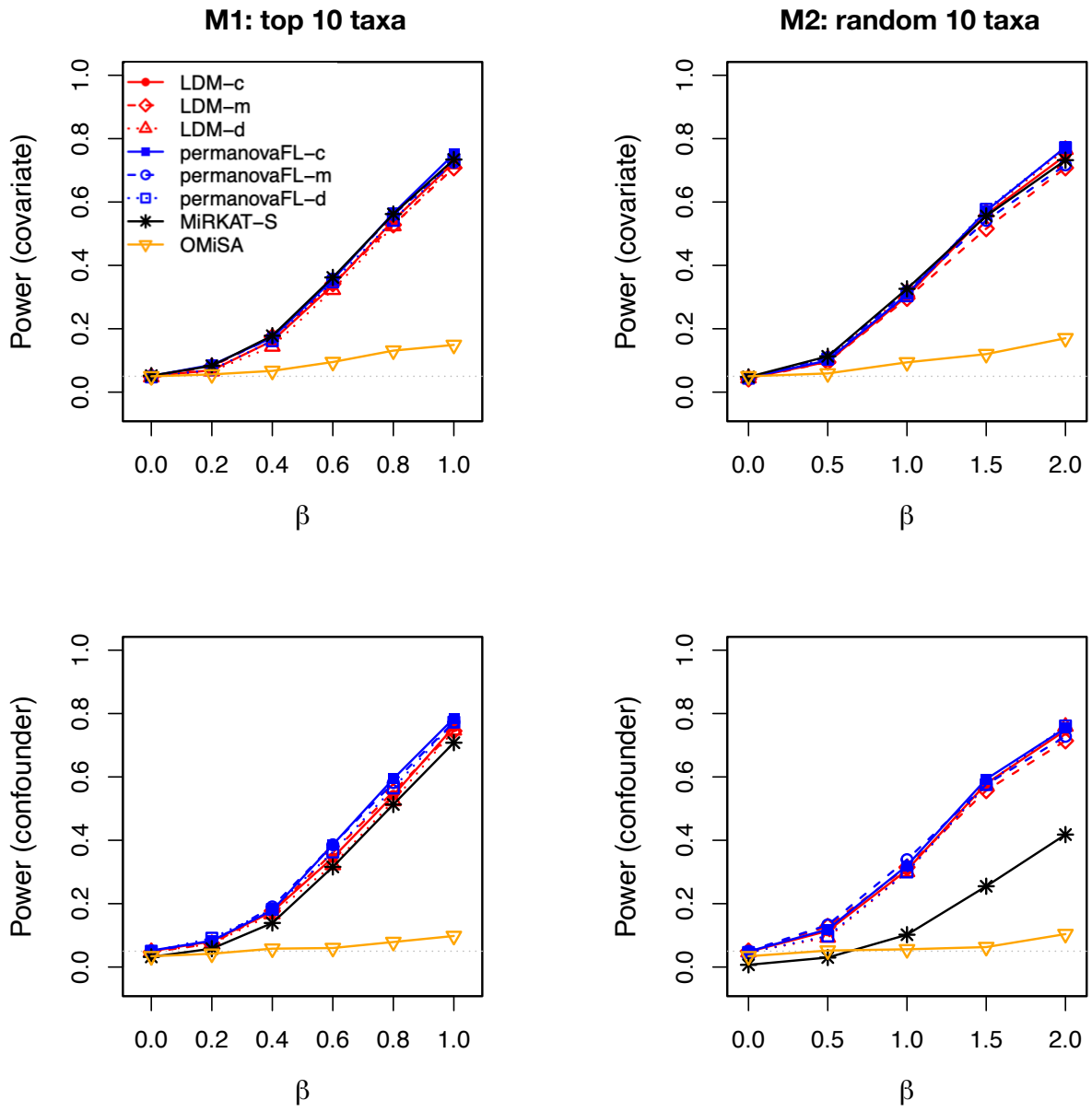


Figure 4.2: Power of the global tests in the presence of a covariate $\beta_{XZ} = 0$ and a confounder $\beta_{XZ} = 0.8$. The data were simulated with 50% censoring and n = 100. The gray dotted line represents the nominal type I error level 0.05.

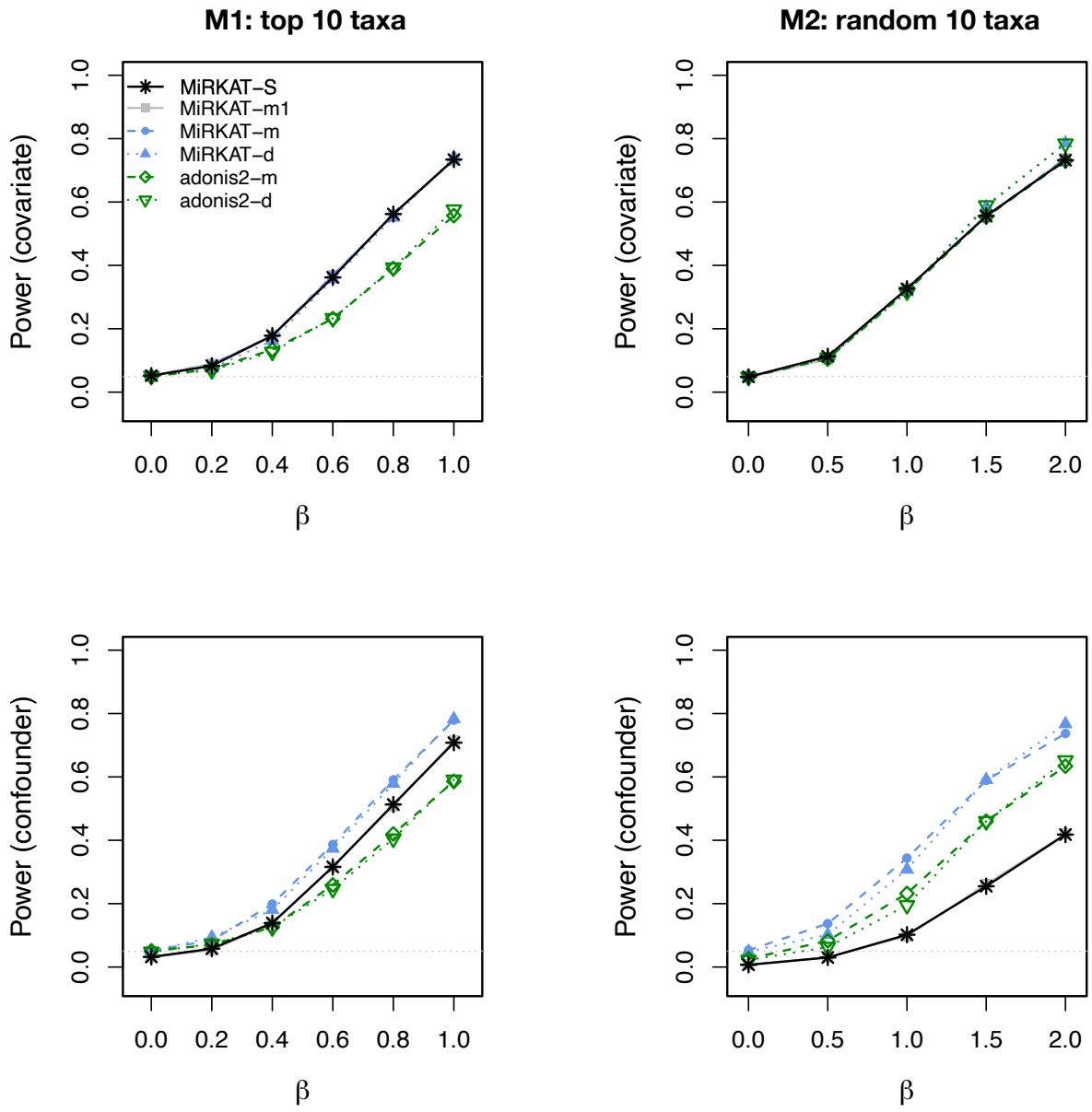


Figure 4.3: See caption to Figure 4.2. The MiRKAT-S results are the same as those in Figure 4.2.

Appendix A

Appendix for Chapter 2

Table A.1: Type I error for testing the global hypothesis at nominal level 0.05

Trait	Confounder	Causal mechanism	n	LOCOM -null	LOCOM -causal	PERMANOVA -half	PERMANOVA -one
Binary	NA	M1, M2	50	0.047	0.047	0.054	0.052
			100	0.052	0.052	0.056	0.063
			200	0.055	0.055	0.048	0.047
*Binary	NA	M1, M2	100	0.040	0.040	0.068	0.070
†Binary	NA	M1, M2	100	0.049	0.049	0.044	0.044
Binary	Binary	M1	50	0.055	0.036	0.045	0.046
			100	0.060	0.037	0.049	0.049
			200	0.050	0.045	0.050	0.050
		M2	50	0.036	0.054	0.045	0.046
			100	0.040	0.059	0.053	0.052
			200	0.043	0.052	0.048	0.049
Binary	Continuous	M1	100	0.049	0.042	0.054	0.053
		M2	100	0.046	0.053	0.049	0.053
Continuous	NA	M1, M2	100	0.049	0.049	0.059	0.055
Continuous	Binary	M1	100	0.048	0.048	0.045	0.045
		M2	100	0.047	0.049	0.047	0.045
Continuous	Continuous	M1	100	0.055	0.049	0.048	0.048
		M2	100	0.043	0.047	0.048	0.048

Note: n is the number of samples. * with differential experimental bias. † data were generated from the differential relative abundance model and the PLNM model. Note that when there is no confounder, the data generated under the global null for scenarios M1 and M2 (as well as M1-500 and M1-rare) are the same.

Table A.2: Type I error for testing the global hypothesis at nominal level 0.05, without adjustment of the confounder

Trait	Confounder	Causal mechanism	LOCOM -null	LOCOM -causal
Binary	Binary	M1	0.132	0.132
		M2	0.154	0.154
Binary	Continuous	M1	0.481	0.481
		M2	0.525	0.525
Continuous	Binary	M1	0.161	0.145
		M2	0.168	0.177
Continuous	Continuous	M1	0.112	0.104
		M2	0.123	0.139

Note: Sample size $n = 100$.

Table A.3: Taxa (in ascending order of the raw p-value) detected by LOCOM in analysis of the two real datasets

Taxon ID	Taxon Assignment	Mean relative abundance	Raw p-value	Adjusted p-value
Taxa associated with smoking in URT microbiome data				
URT-1	*g_Veillonella*	0.067	0.00018	0.018
URT-2	*g_Streptococcus*	0.015	0.00036	0.018
URT-3	*g_Atopobium*	0.012	0.00091	0.031
URT-4	*g_Megasphaera*	0.013	0.00127	0.032
URT-5	*g_Prevotella*	0.041	0.00182	0.037
URT-6	*g_Prevotella*	0.0024	0.00455	0.076
Taxa associated with gPC5 in PPI microbiome data				
PPI-1	*f_Flavobacteriaceae*	0.00098	0.0020	0.025
PPI-2	*g_Escherichia*	0.015	0.0020	0.025
PPI-3	*g_Escherichia*	0.00032	0.0020	0.025
PPI-4	*f_Comamonadaceae*	0.0013	0.0040	0.030
PPI-5	*g_Coprococcus*	0.00051	0.0040	0.030
PPI-6	*g_Escherichia*	0.093	0.0060	0.032
PPI-7	*g_Escherichia*	0.0021	0.0060	0.032
PPI-8	*f_Enterobacteriaceae*	0.00037	0.0014	0.066
PPI-9	*g_Blautia*	0.00015	0.0018	0.070
PPI-10	*g_Acinetobacter*	0.0014	0.0022	0.070
PPI-11	*g_Escherichia*	0.0011	0.0022	0.070
PPI-12	*g_Streptococcus*	0.00023	0.0024	0.070
PPI-13	*g_Escherichia*	0.0032	0.0026	0.070
PPI-14	*g_Escherichia*	0.000058	0.0026	0.070
PPI-15	*g_Escherichia*	0.00028	0.0028	0.070
PPI-16	*g_Escherichia*	0.0053	0.0032	0.075
PPI-17	*g_Escherichia*	0.000071	0.0034	0.075
PPI-18	*Unclassified*	0.00091	0.0044	0.085
PPI-19	*g_Streptococcus*	0.00010	0.0046	0.085
PPI-20	*g_Bifidobacterium*	0.011	0.0052	0.085
PPI-21	*f_Clostridiaceae*	0.0015	0.0054	0.085
PPI-22	*g_Escherichia*	0.00016	0.0056	0.085
PPI-23	*g_Escherichia*	0.000080	0.0058	0.085
PPI-24	*g_Escherichia*	0.00011	0.0062	0.085
PPI-25	*g_Bifidobacterium*	0.0015	0.0064	0.085
PPI-26	*f_Lachnospiraceae*	0.000089	0.0064	0.085
PPI-27	*g_Escherichia*	0.000049	0.0064	0.085
PPI-28	*g_Clostridium*	0.00031	0.0066	0.085
PPI-29	*g_Blautia*	0.0040	0.0066	0.085
PPI-30	*g_Escherichia*	0.00080	0.0068	0.085
PPI-31	*f_Enterobacteriaceae*	0.000040	0.0072	0.088
PPI-32	*f_Enterococcaceae*	0.00076	0.0080	0.094
Taxa associated with gPC2 in PPT microbiome data				
PPI-19	*g_Streptococcus*	0.00010	0.00020	0.047
PPI-33	*g_Streptococcus*	0.00023	0.00020	0.047
Taxa associated with gPC3 in PPT microbiome data				
PPI-19	*g_Streptococcus*	0.00010	0.00020	0.041
PPI-33	*g_Streptococcus*	0.00023	0.00020	0.041

Note: The mean relative abundance was calculated based on the data after filtering out taxa having fewer than 20% presence in the sample. "*g_*" or "*f_*" indicates that the taxon is assigned a genus or a family. The nominal FDR is 10%.

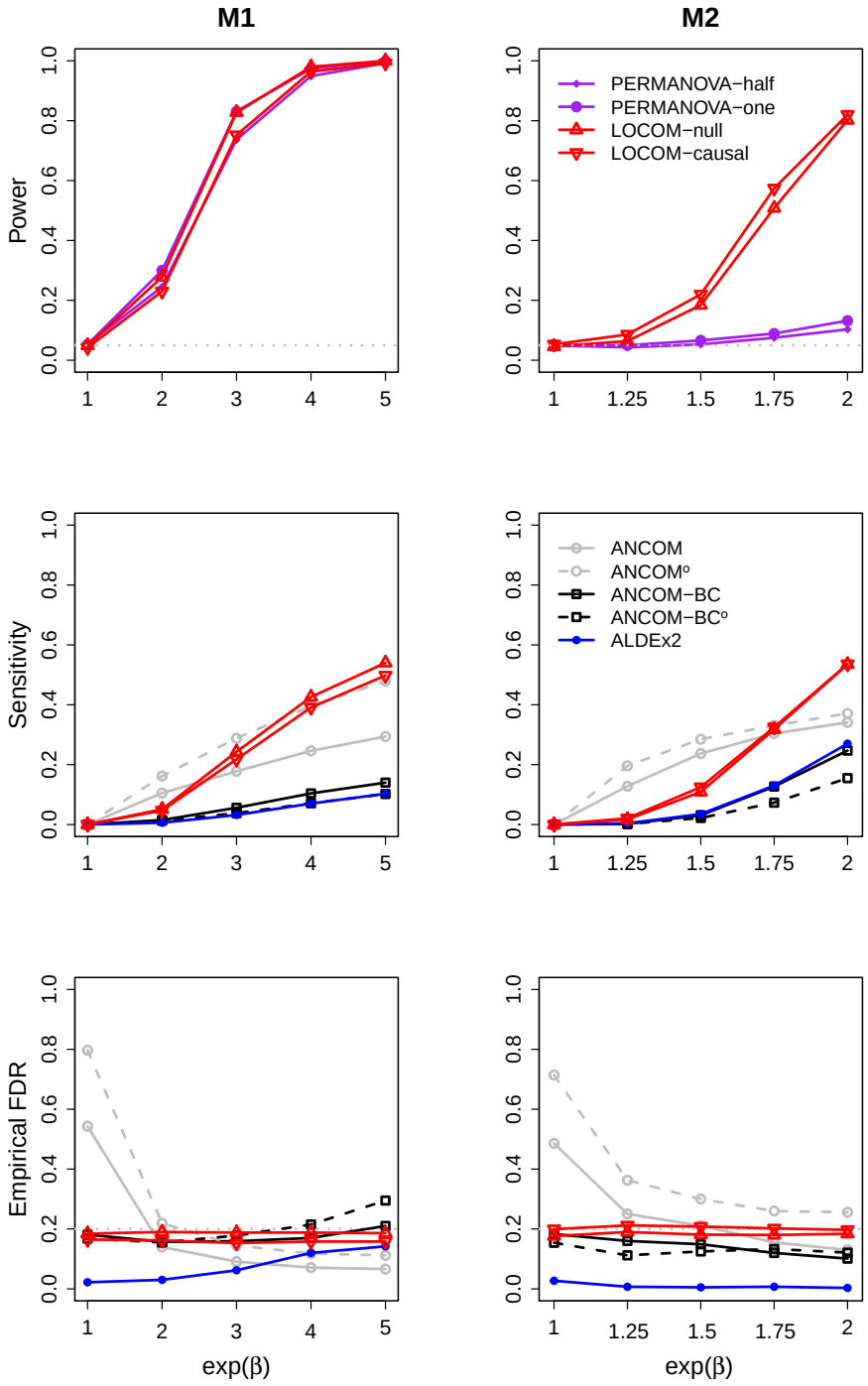


Figure A.1: Simulation results for data ($n = 100$) with a binary trait and a continuous confounder.

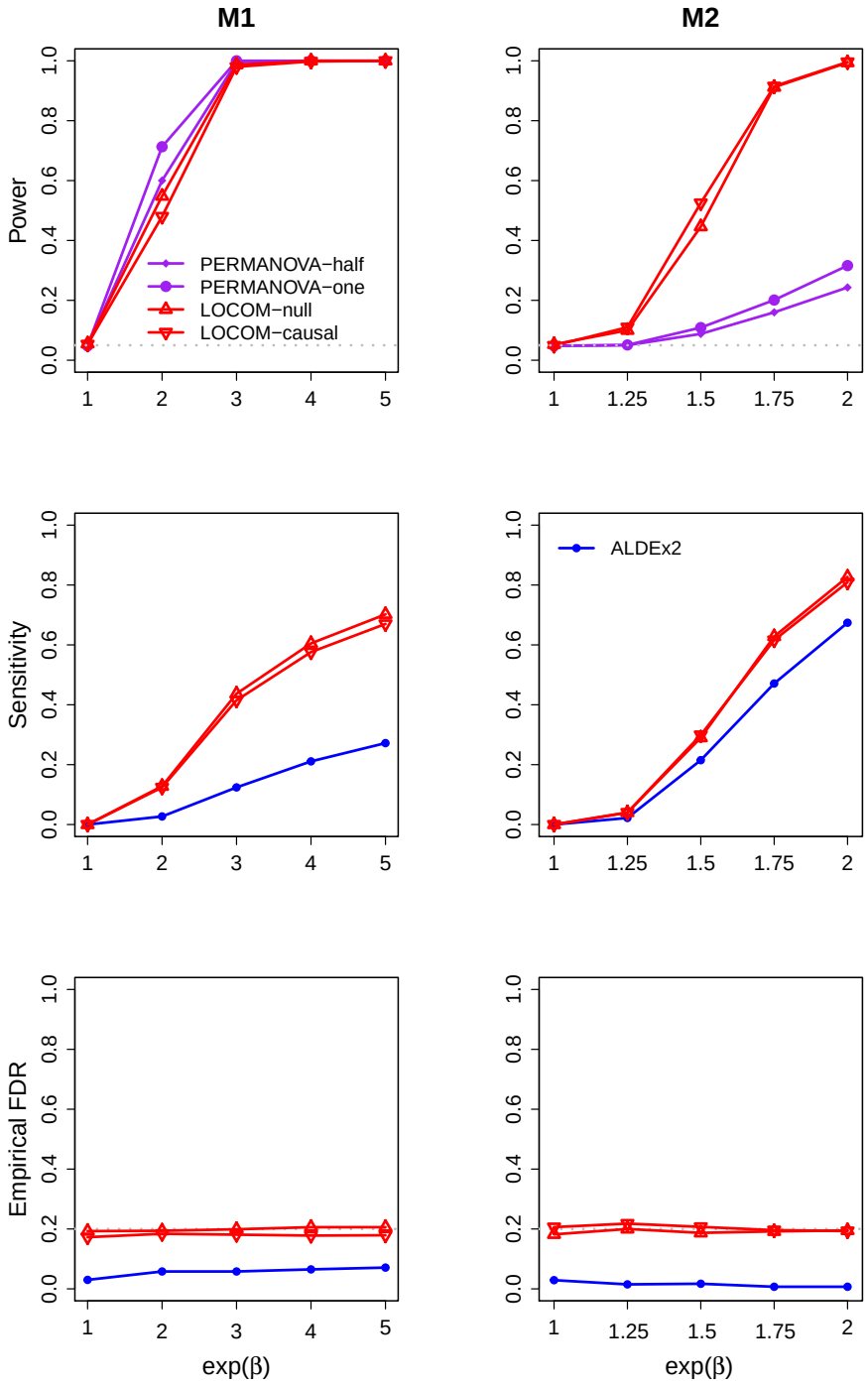


Figure A.2: Simulation results for data ($n = 100$) with a continuous trait and a continuous confounder.

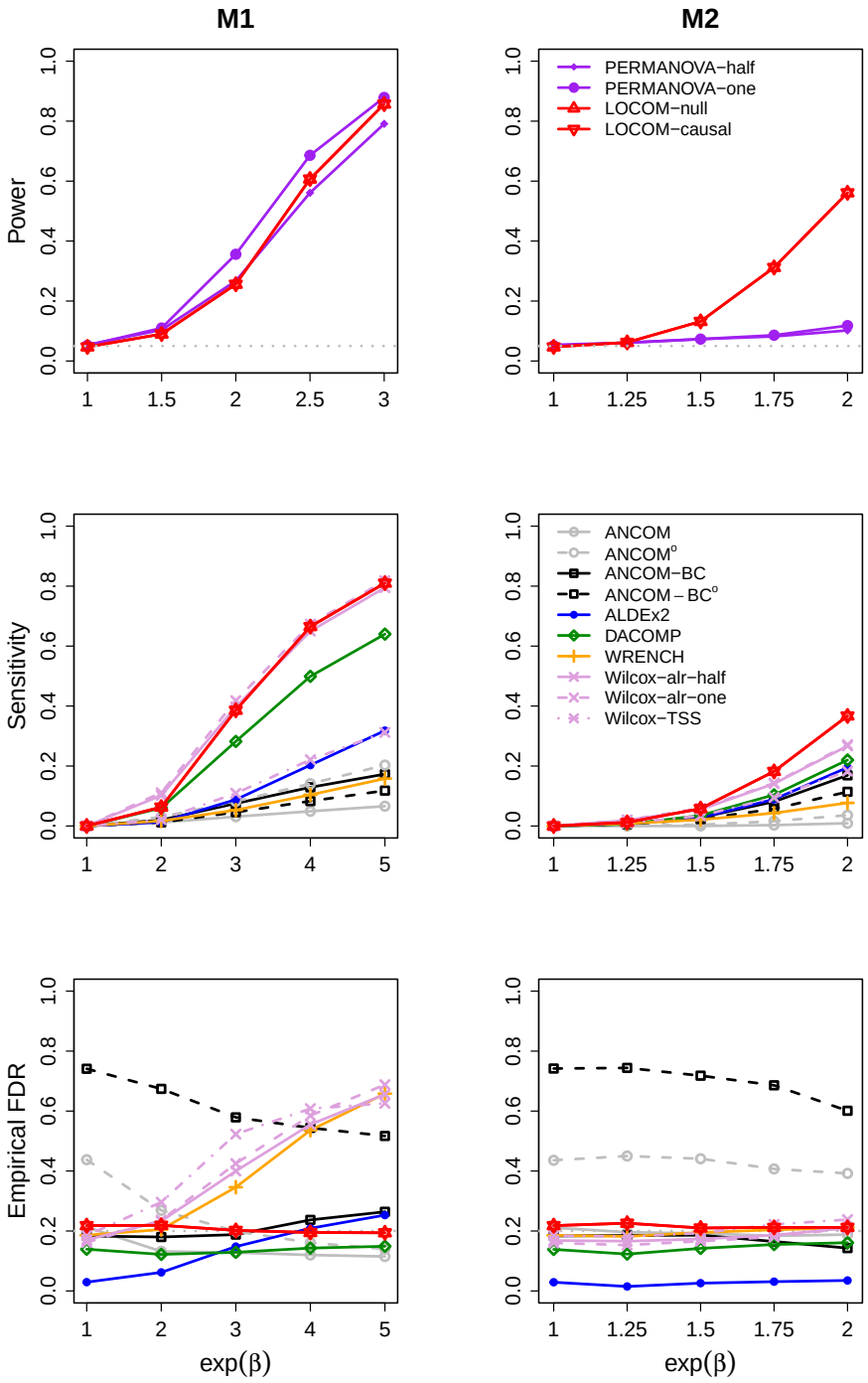


Figure A.3: Simulation results for data ($n = 50$) with a binary trait (and no confounder).

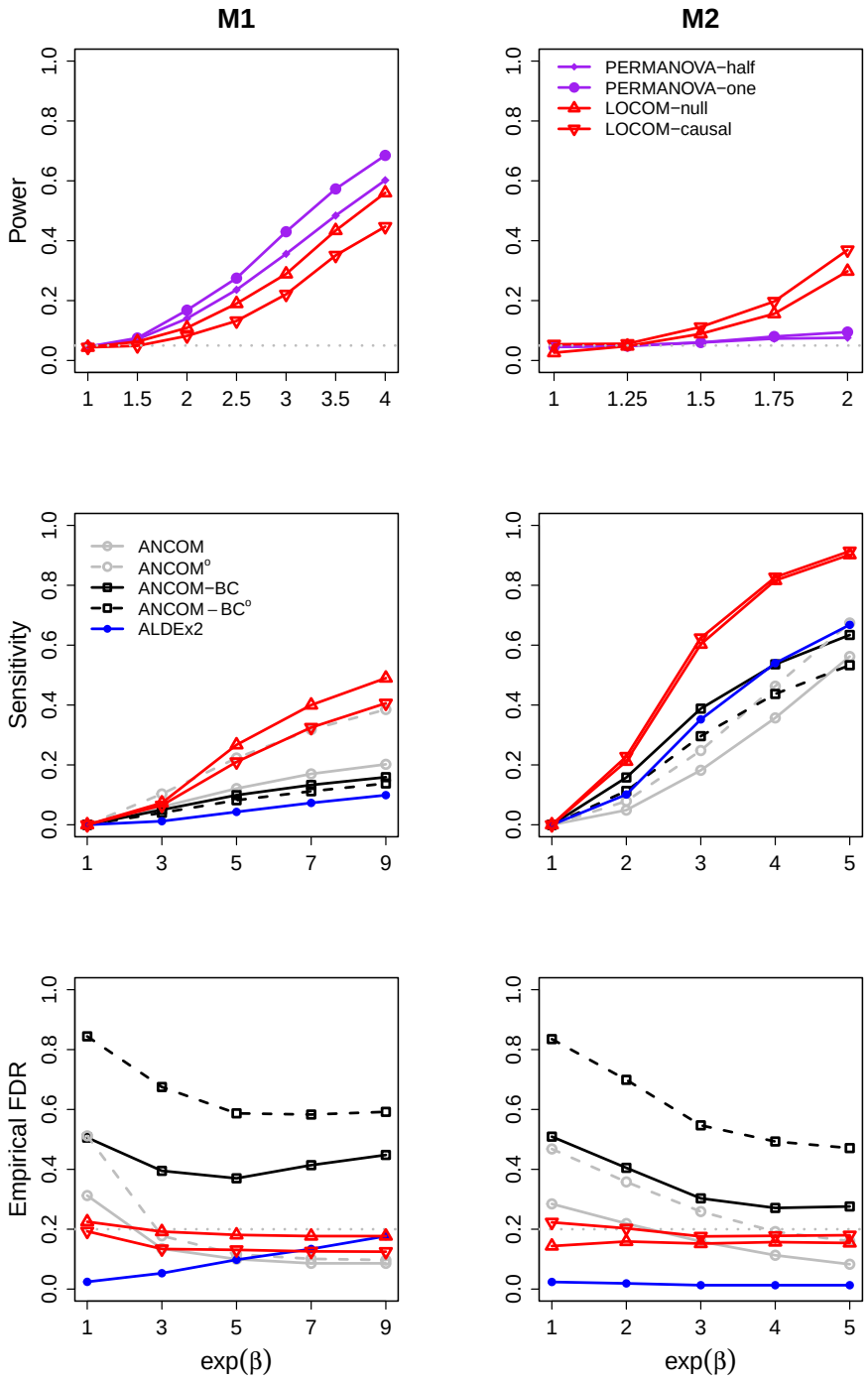


Figure A.4: Simulation results for data ($n = 50$) with a binary trait and a binary confounder.

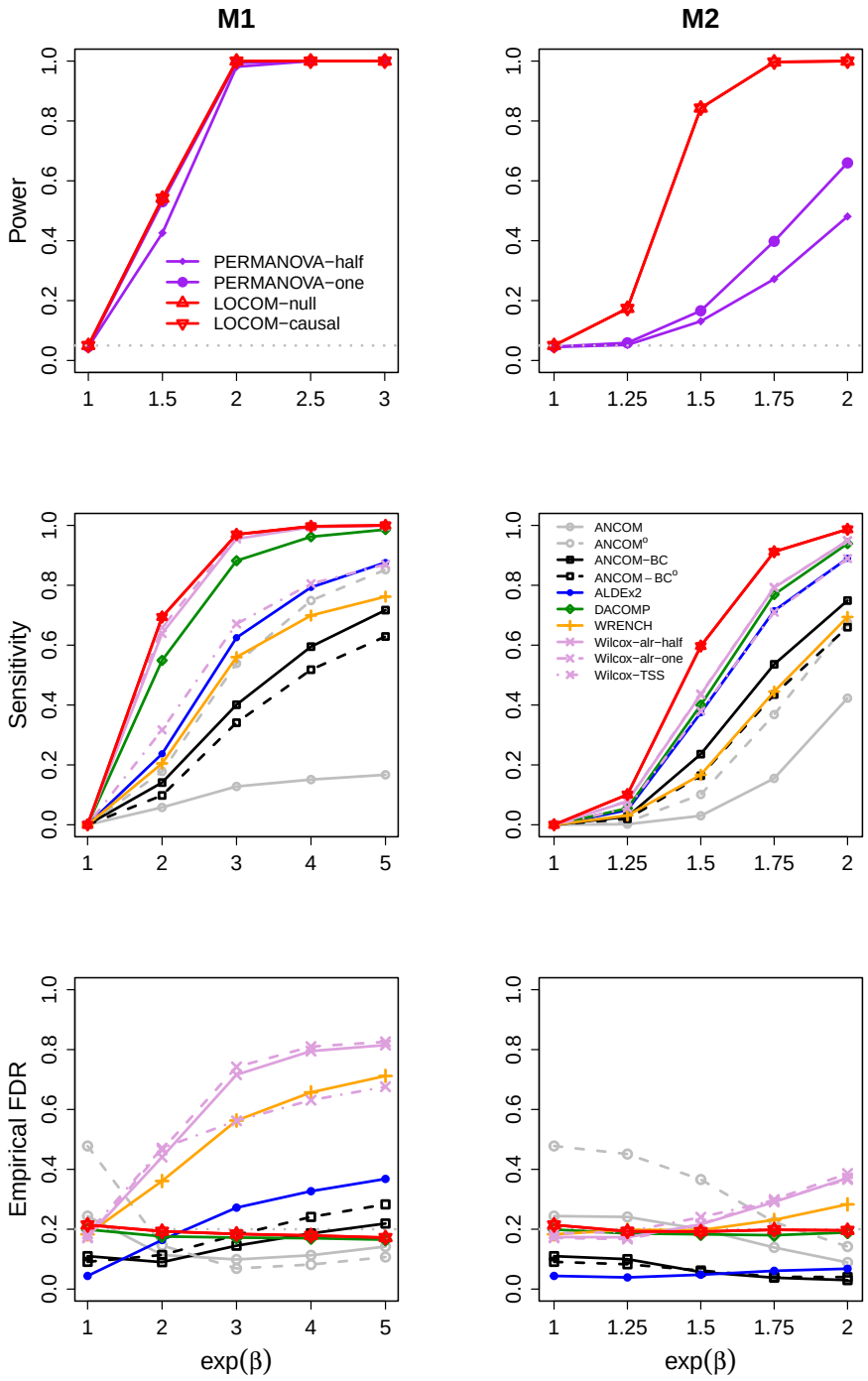


Figure A.5: Simulation results for data ($n = 200$) with a binary trait (and no confounder).

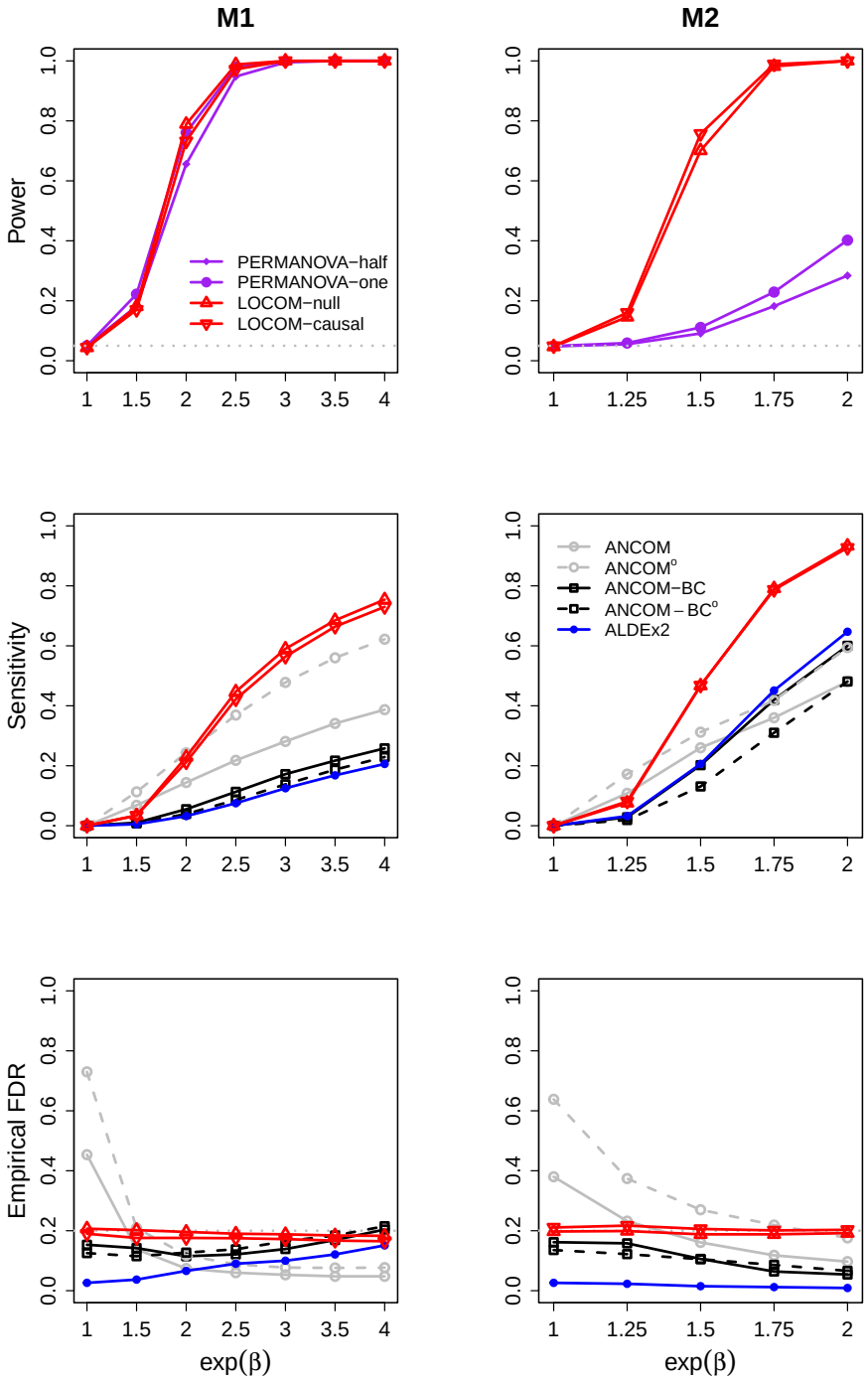


Figure A.6: Simulation results for data ($n = 200$) with a binary trait and a binary confounder.

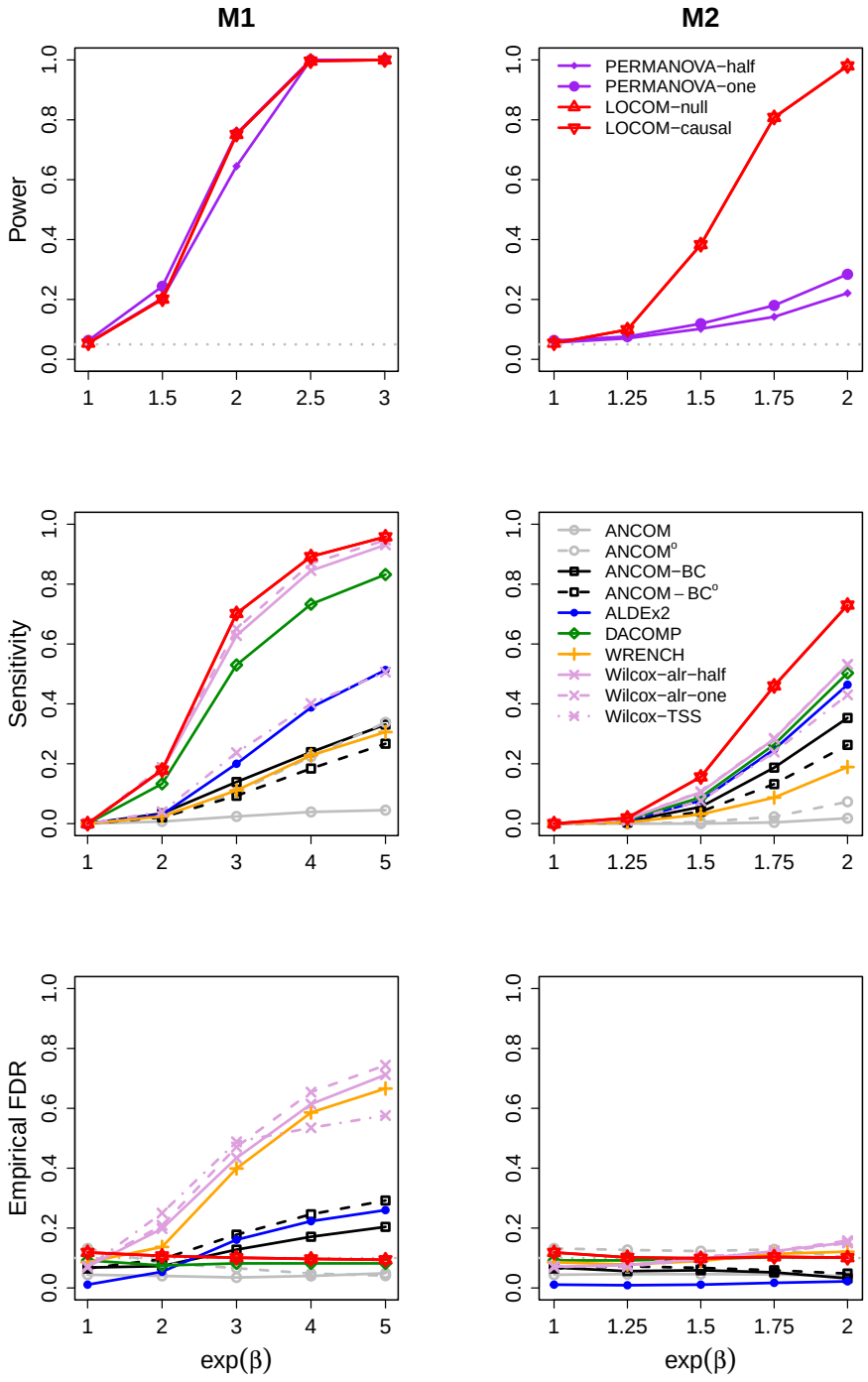


Figure A.7: Simulation results for data ($n = 100$) with a binary trait (and no confounder), and for the nominal FDR level of 10%.

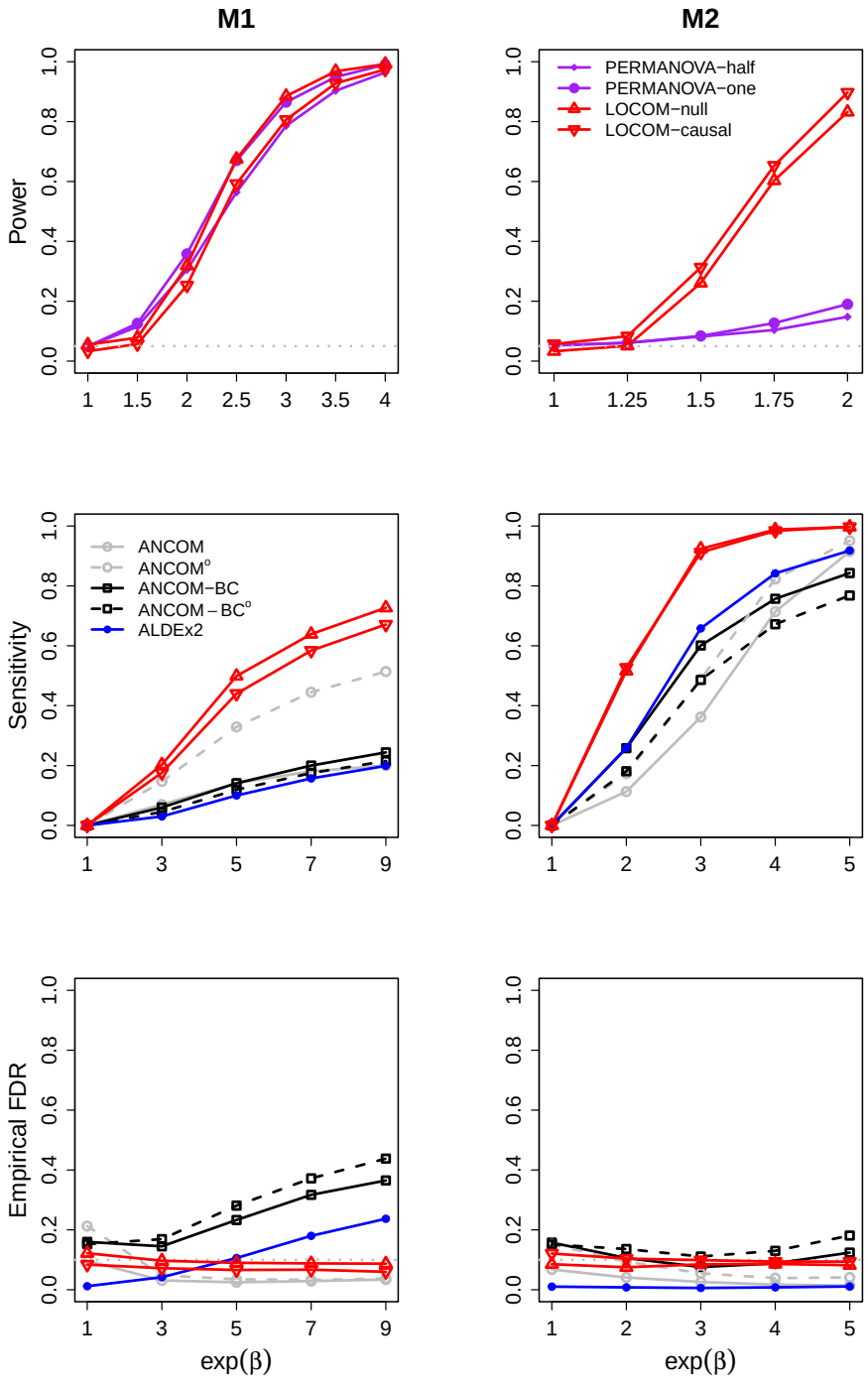


Figure A.8: Simulation results for data ($n = 100$) with a binary trait and a binary confounder, and for the nominal FDR level of 10%.

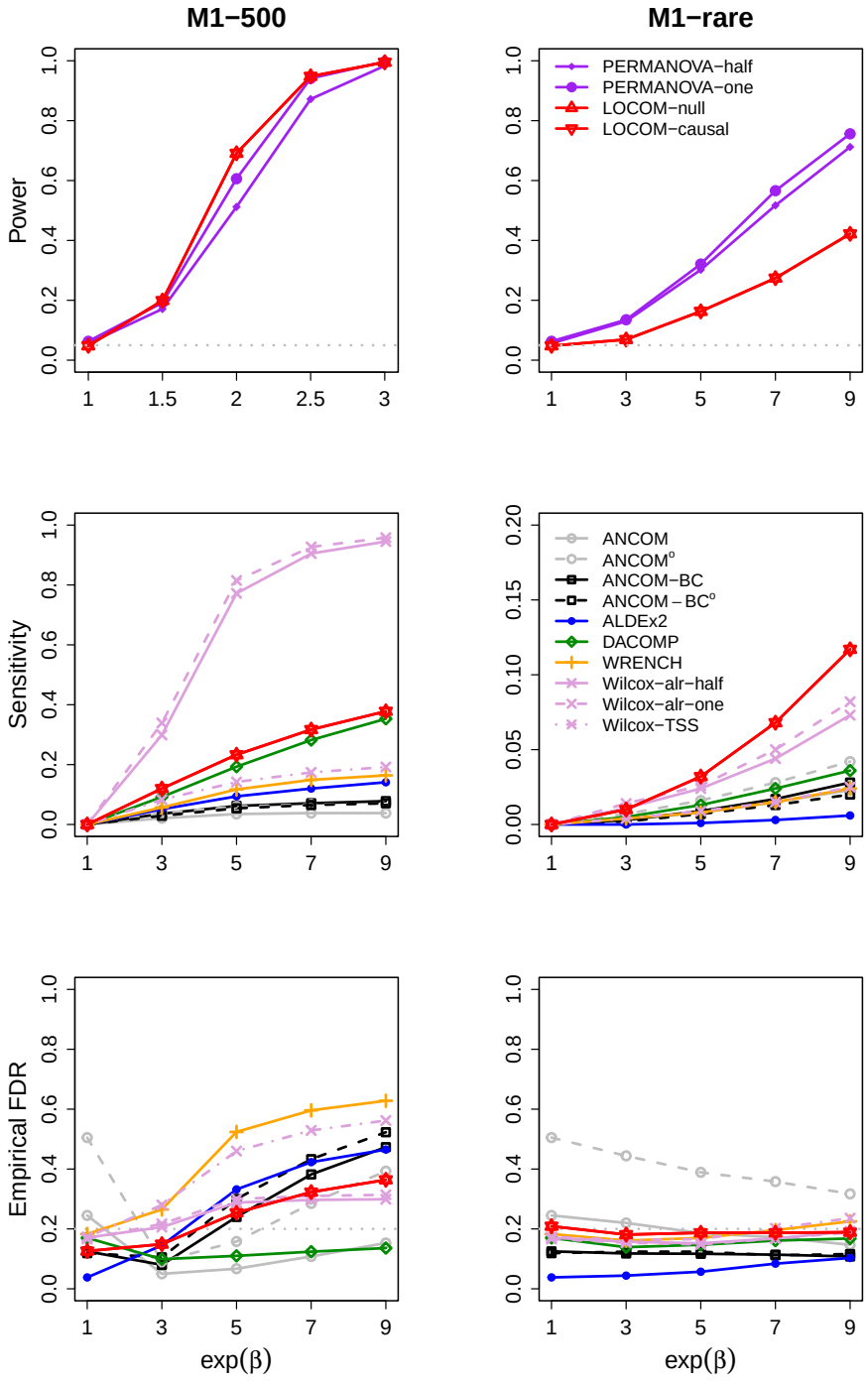


Figure A.9: Simulation results for data ($n = 100$) with a binary trait (and no confounder), generated by modifying M1 to have 500 causal taxa (M1-500) or 20 rare causal taxa (M1-rare).

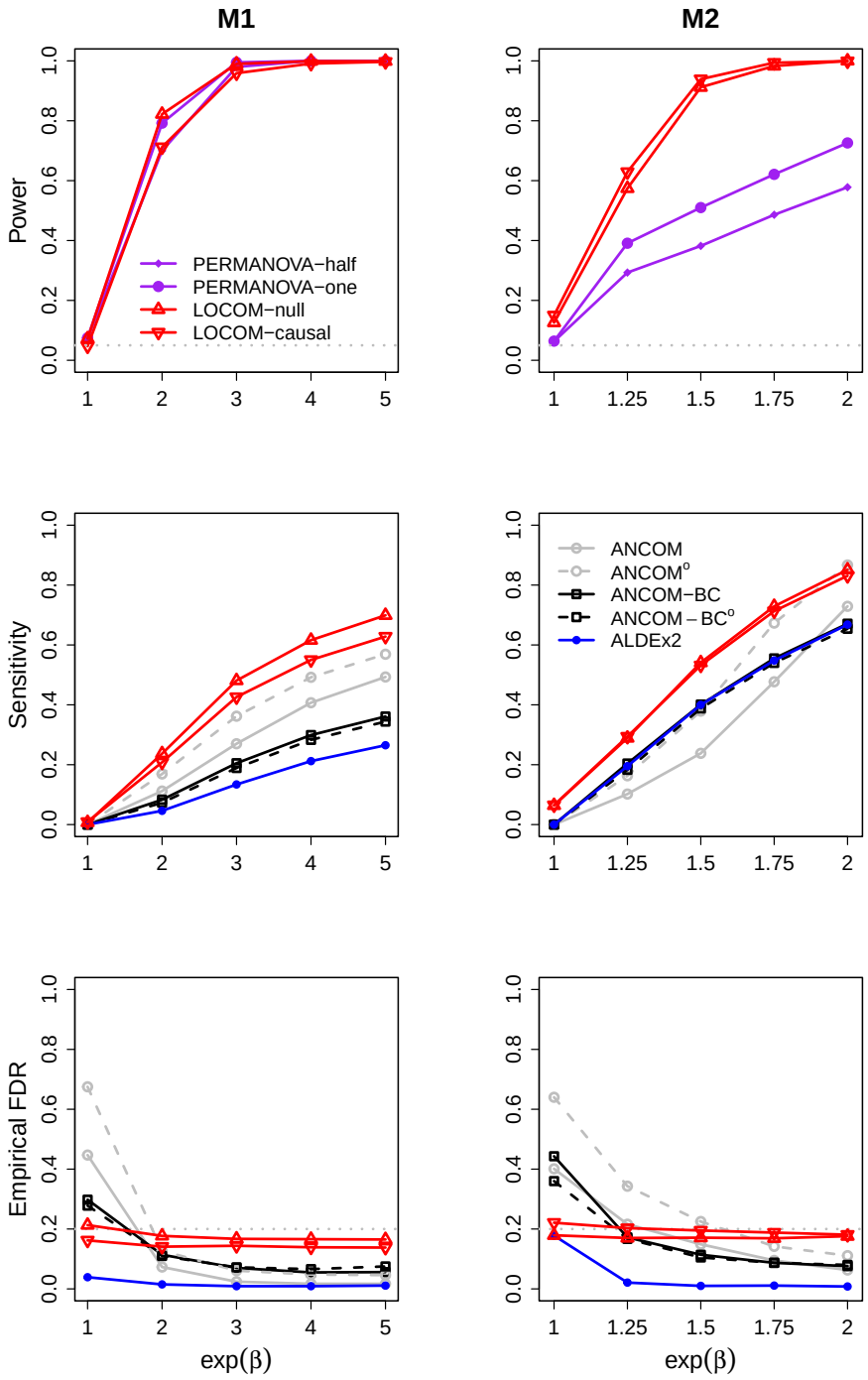


Figure A.10: Simulation results for data ($n = 100$) with a binary trait and a binary confounder, when different values were used for different $\beta_{j,1}$. Specifically, the fold change $\exp(\beta_{j,1})$ was obtained by coupling an initial fold change with the "scale factor" $\exp(\beta)$. The initial fold change was sampled from $U[0.5, 1.5]$, which implies different directions. If the initial fold change was positive (greater than 1), it was up-scaled (multiplied) by the scale factor to give the final fold change; if the initial fold change was negative (less than 1), it was down-scaled (divided) by the scale factor to give the final fold change. Note that the scale factor $\exp(\beta) = 1$ does not correspond to the global null.

Appendix B

Appendix for Chapter 4

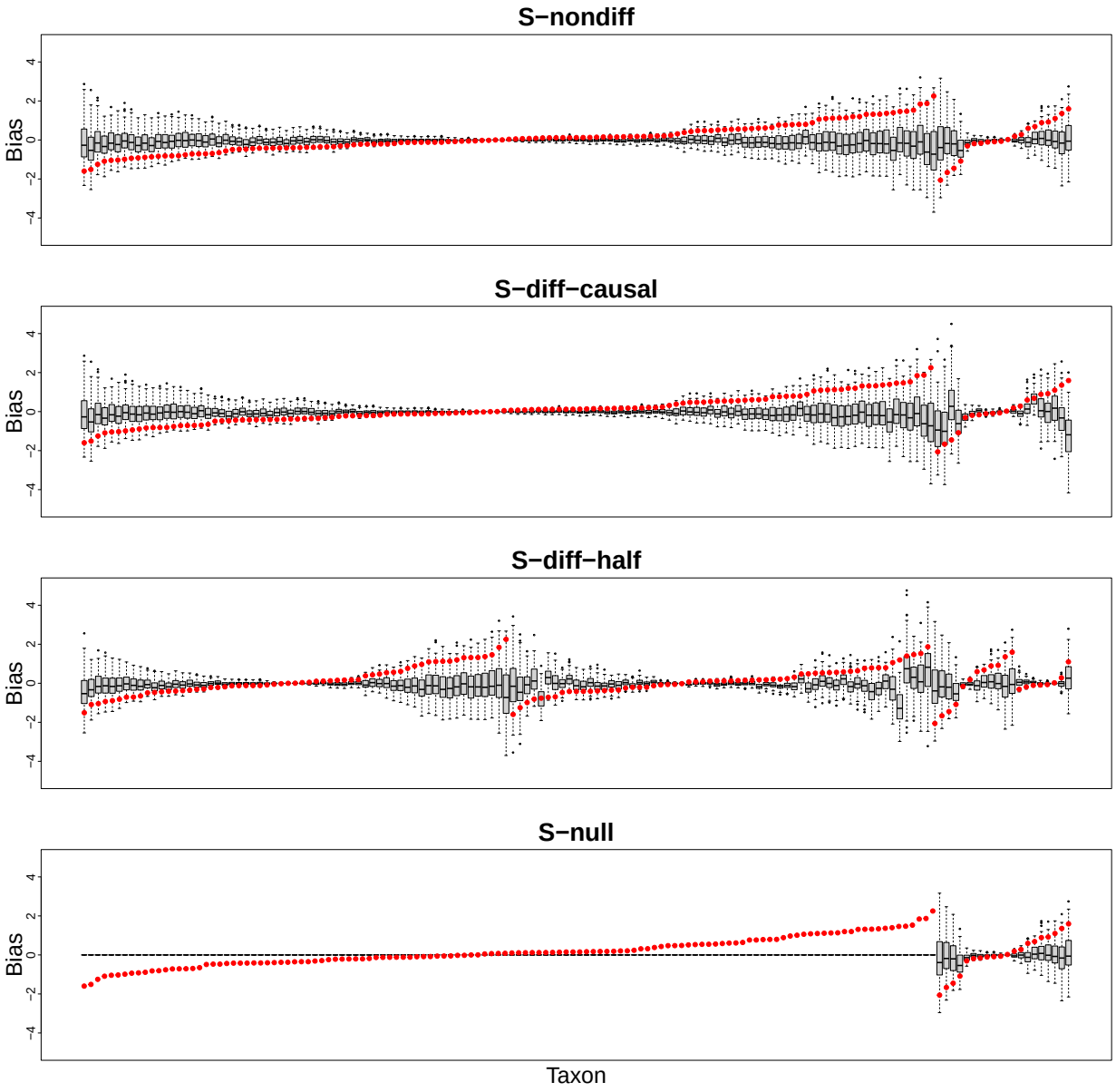


Figure B.1: Distributions of the interactive bias γ_{ij} across samples (box plot) and the taxon-specific main bias γ_j (red dot), for all taxa (that passed our filter) using one replicate of data simulated under M1. For each panel, we sorted the taxa so that the null taxa appeared first and the causal taxa next, and then, within each group, we sorted the taxa in ascending order of the main bias. In S3, we additionally sorted the taxa within each group so that the taxa belonging to the first half of taxa that were randomly selected appeared first and the taxa belonging to the remaining half of taxa appeared next.

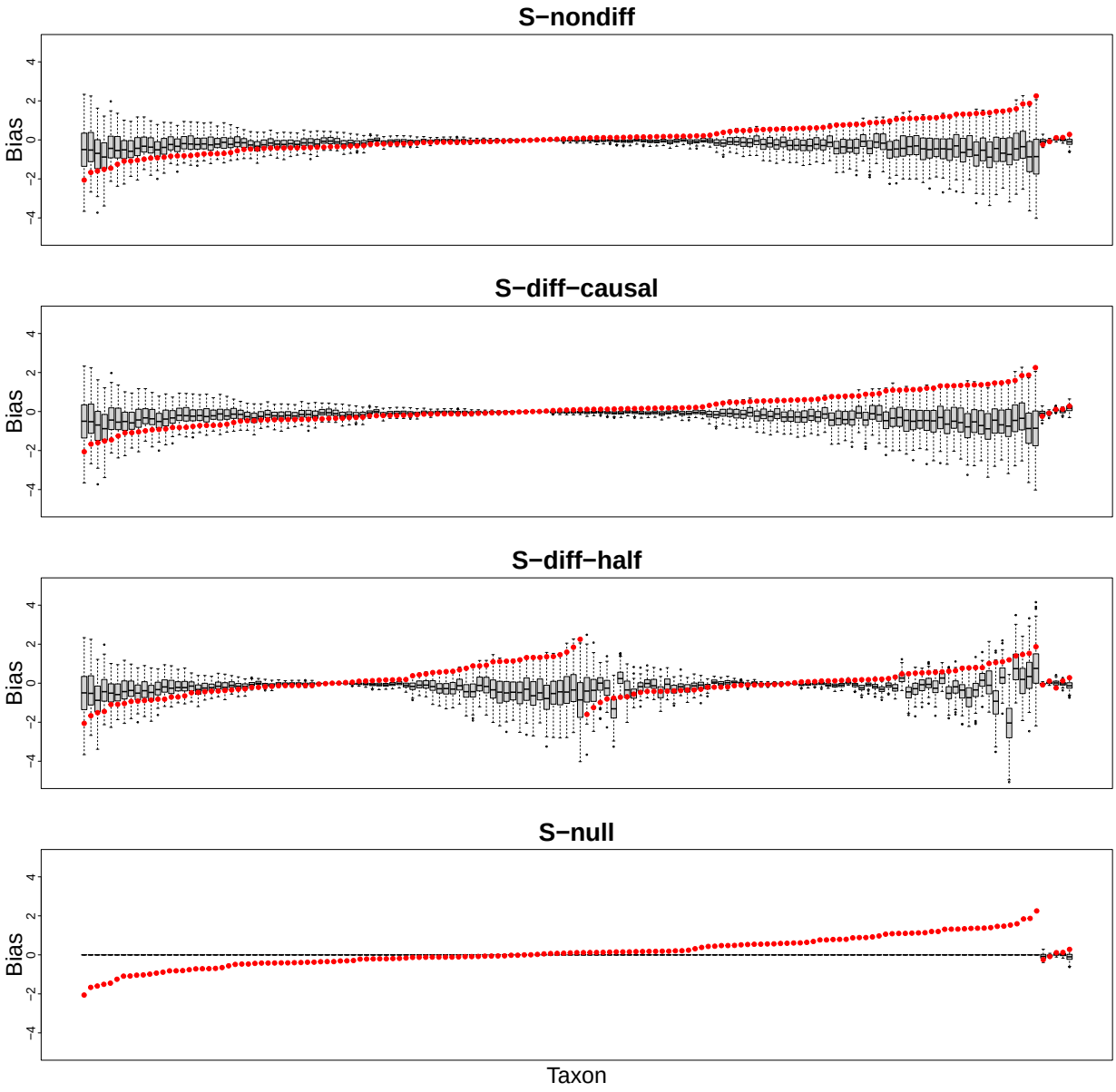


Figure B.2: Distributions of the interactive bias γ_{ij} across samples, for all taxa (that passed our filter) using one replicate of data simulated under M2.

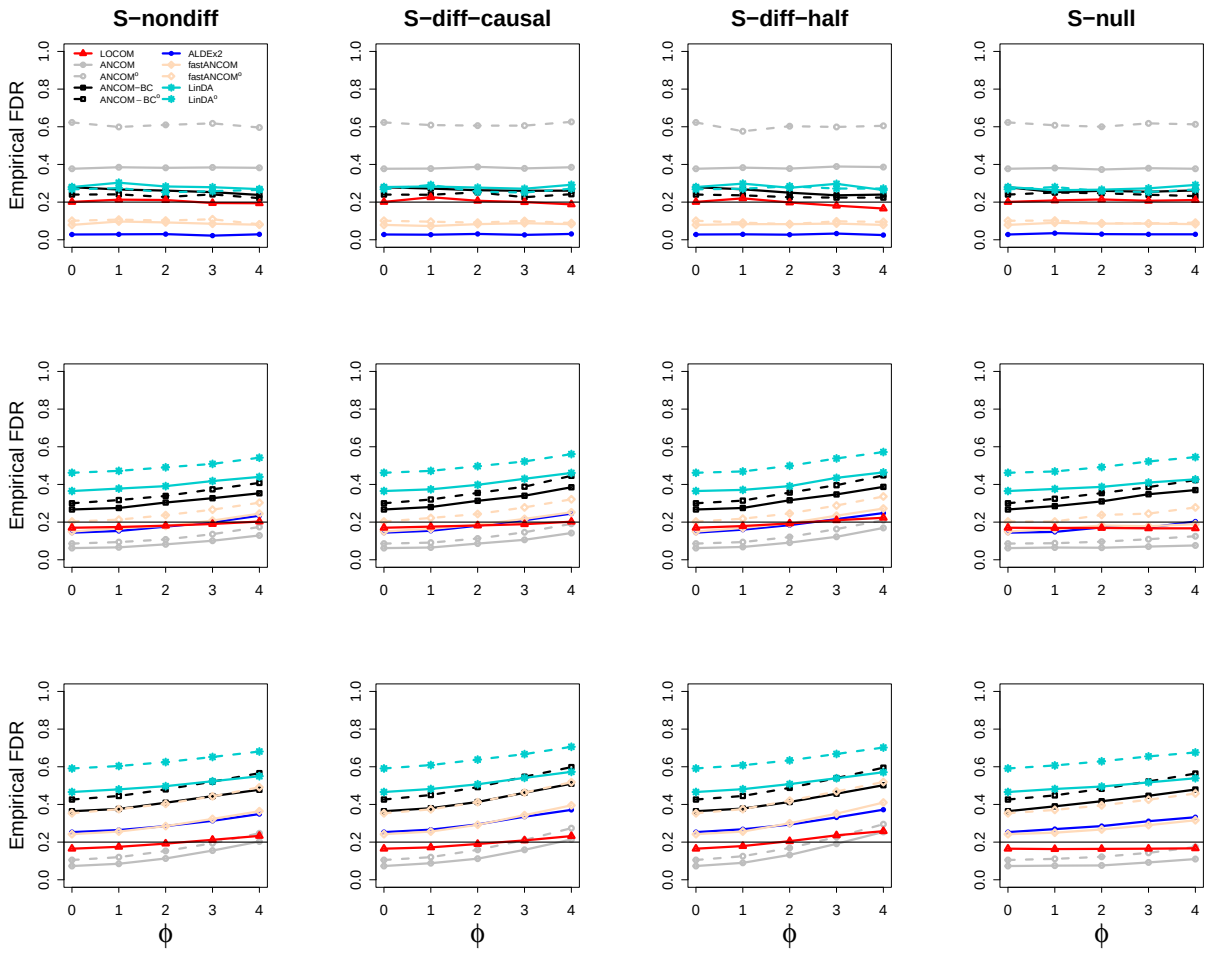


Figure B.3: Empirical FDR results for data generated under M1, with a confounder. The rows correspond to fold change $\exp(\beta) = 1$, $\exp(\beta) = 5$ and $\exp(\beta) = 9$, respectively.

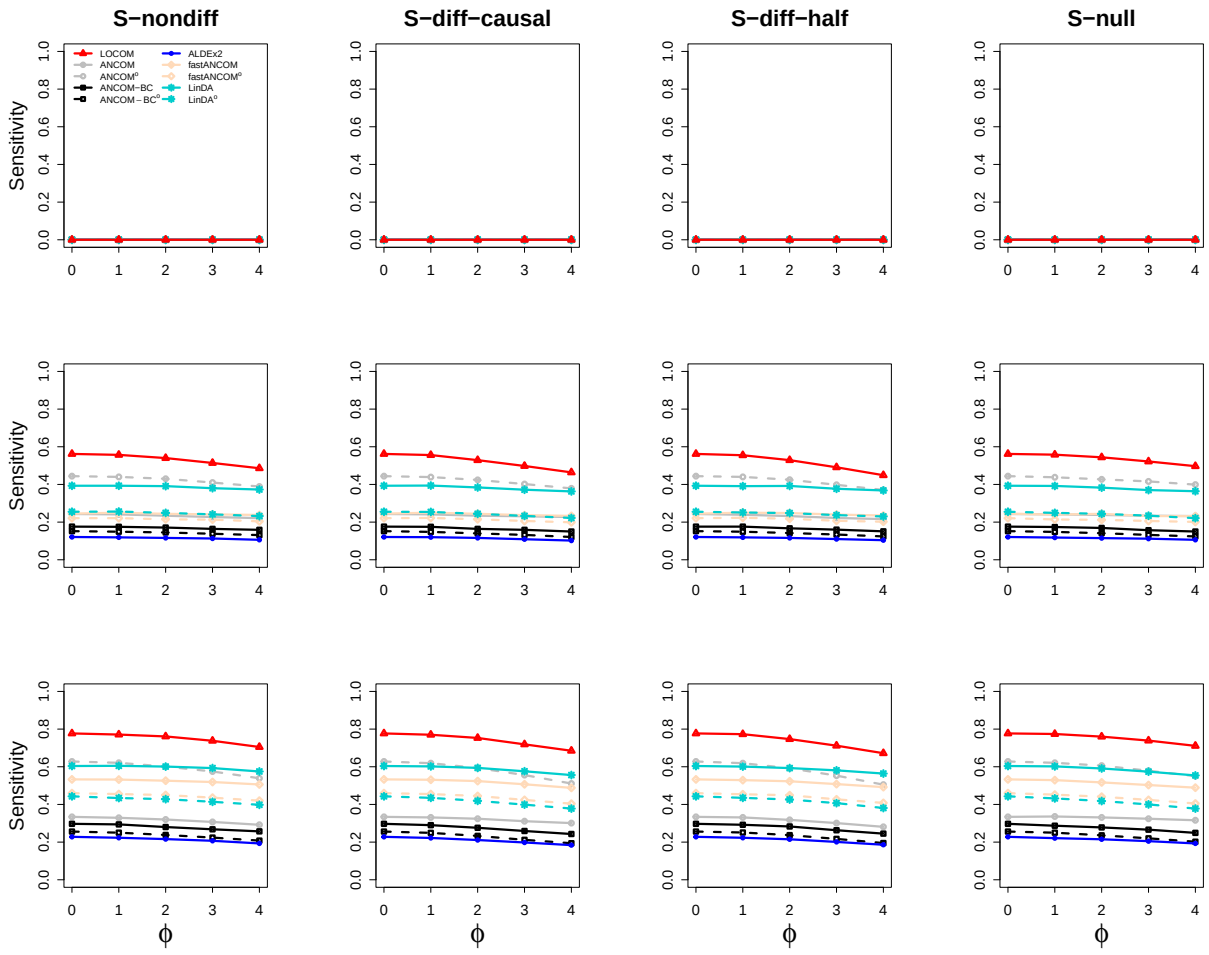


Figure B.4: Sensitivity results for data generated under M1, with a confounder. The rows correspond to fold change $\exp(\beta) = 1$, $\exp(\beta) = 5$ and $\exp(\beta) = 9$, respectively.

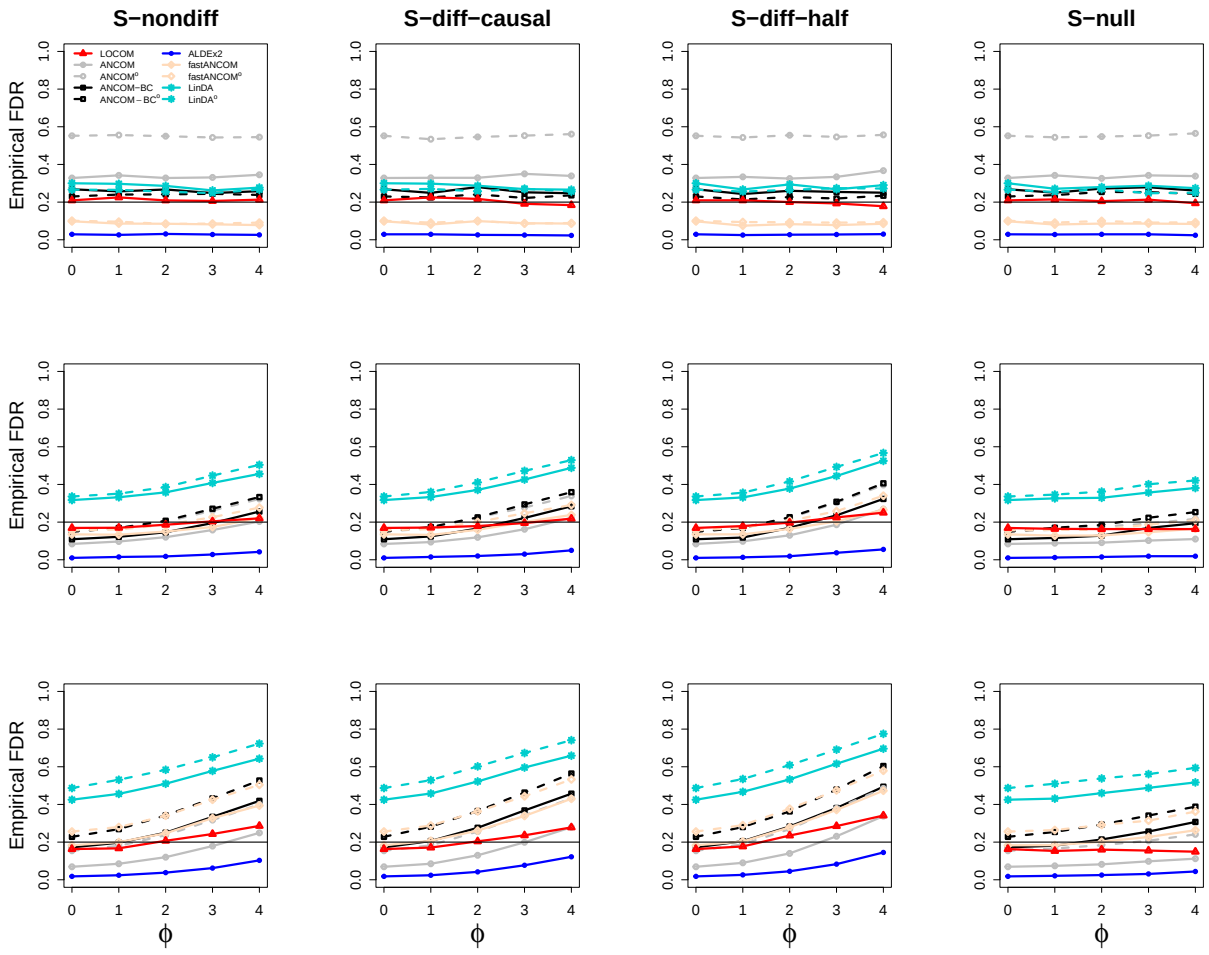


Figure B.5: Empirical FDR results for data generated under M2, with a confounder. The rows correspond to fold change $\exp(\beta) = 1$, $\exp(\beta) = 3$ and $\exp(\beta) = 5$, respectively.

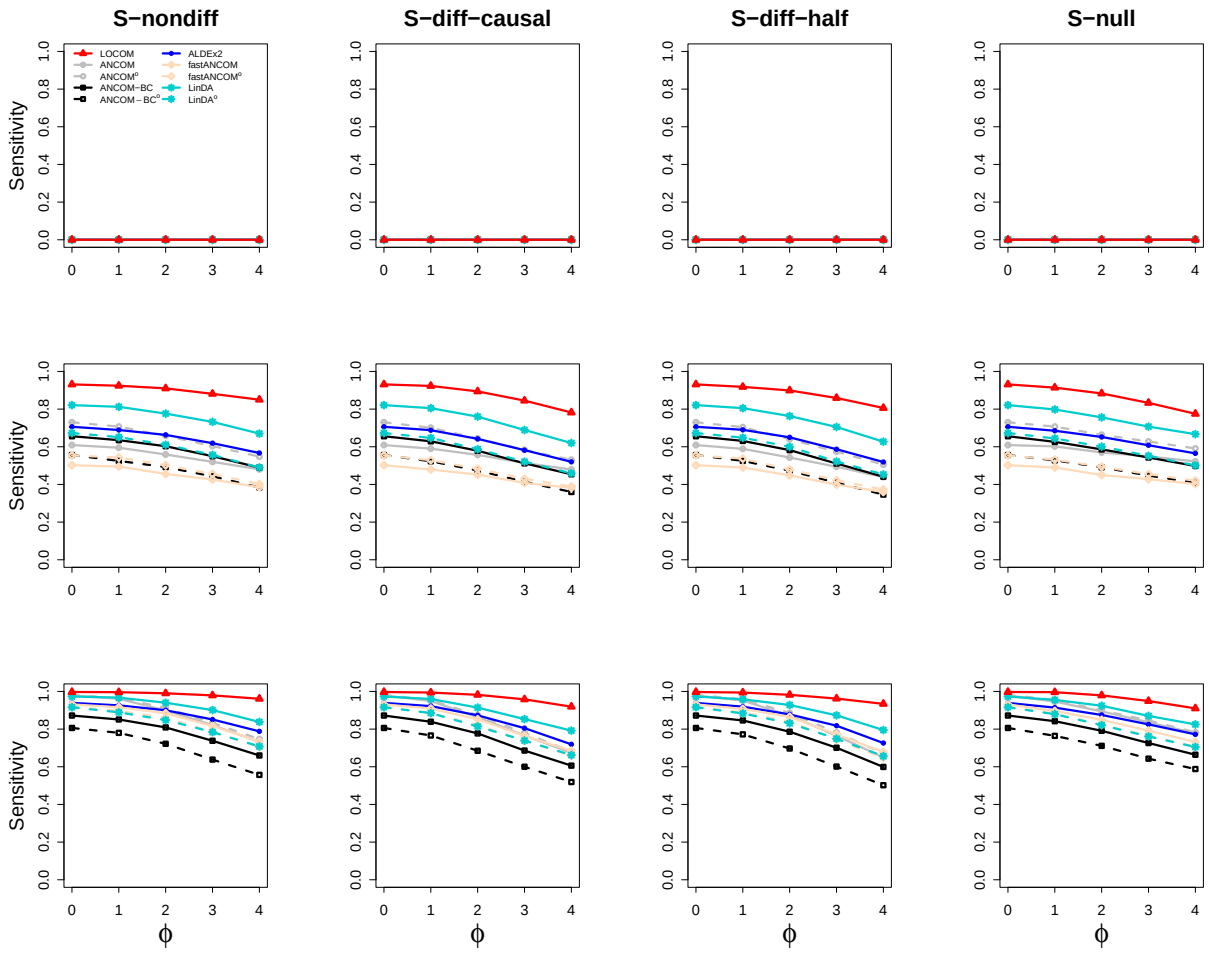


Figure B.6: Sensitivity results for data generated under M2, with a confounder. The rows correspond to fold change $\exp(\beta) = 1$, $\exp(\beta) = 3$ and $\exp(\beta) = 5$, respectively.

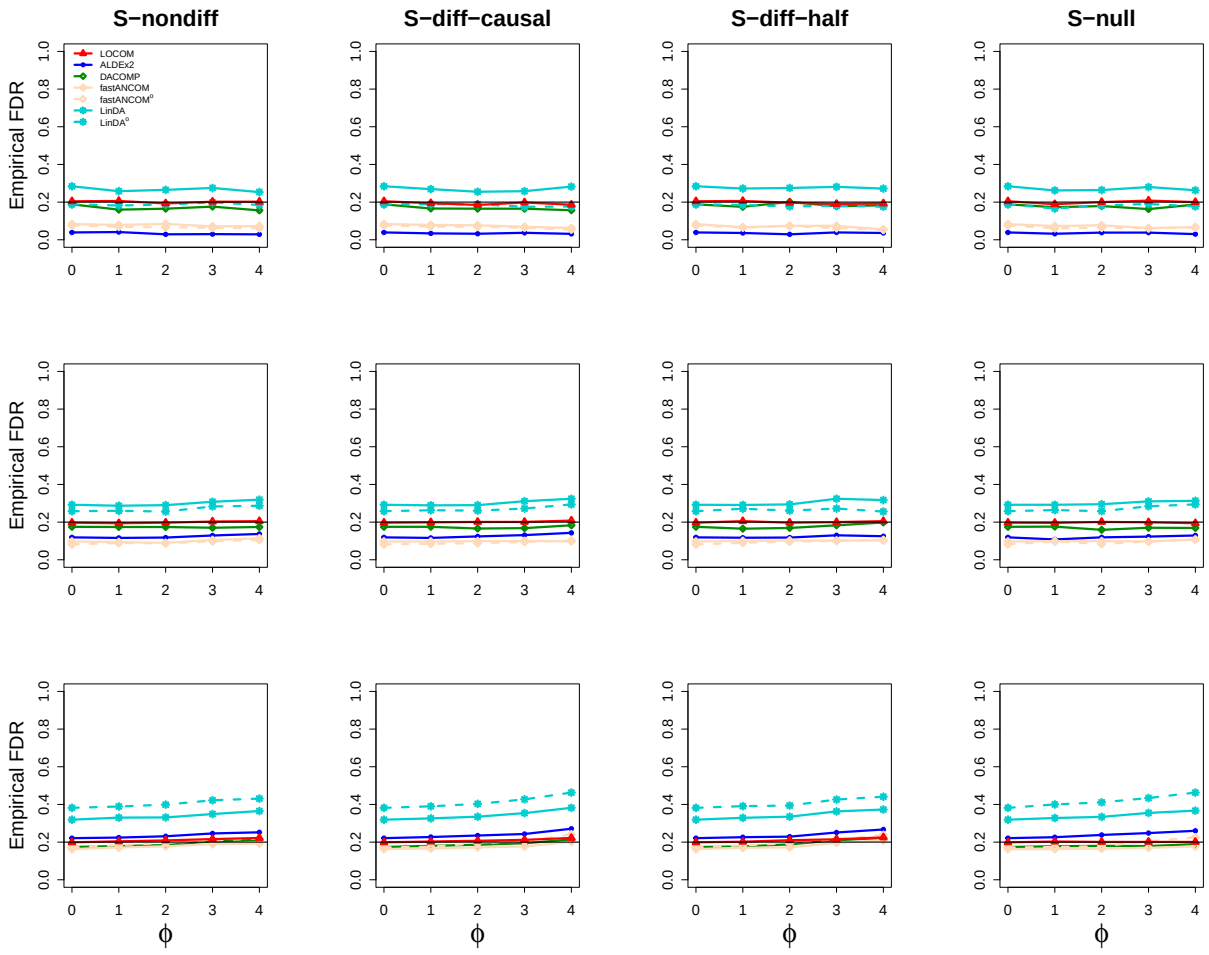


Figure B.7: Sensitivity results for data generated under M1, with a continuous trait. The rows correspond to fold change $\exp(\beta) = 1$, $\exp(\beta) = 2$ and $\exp(\beta) = 3$, respectively.

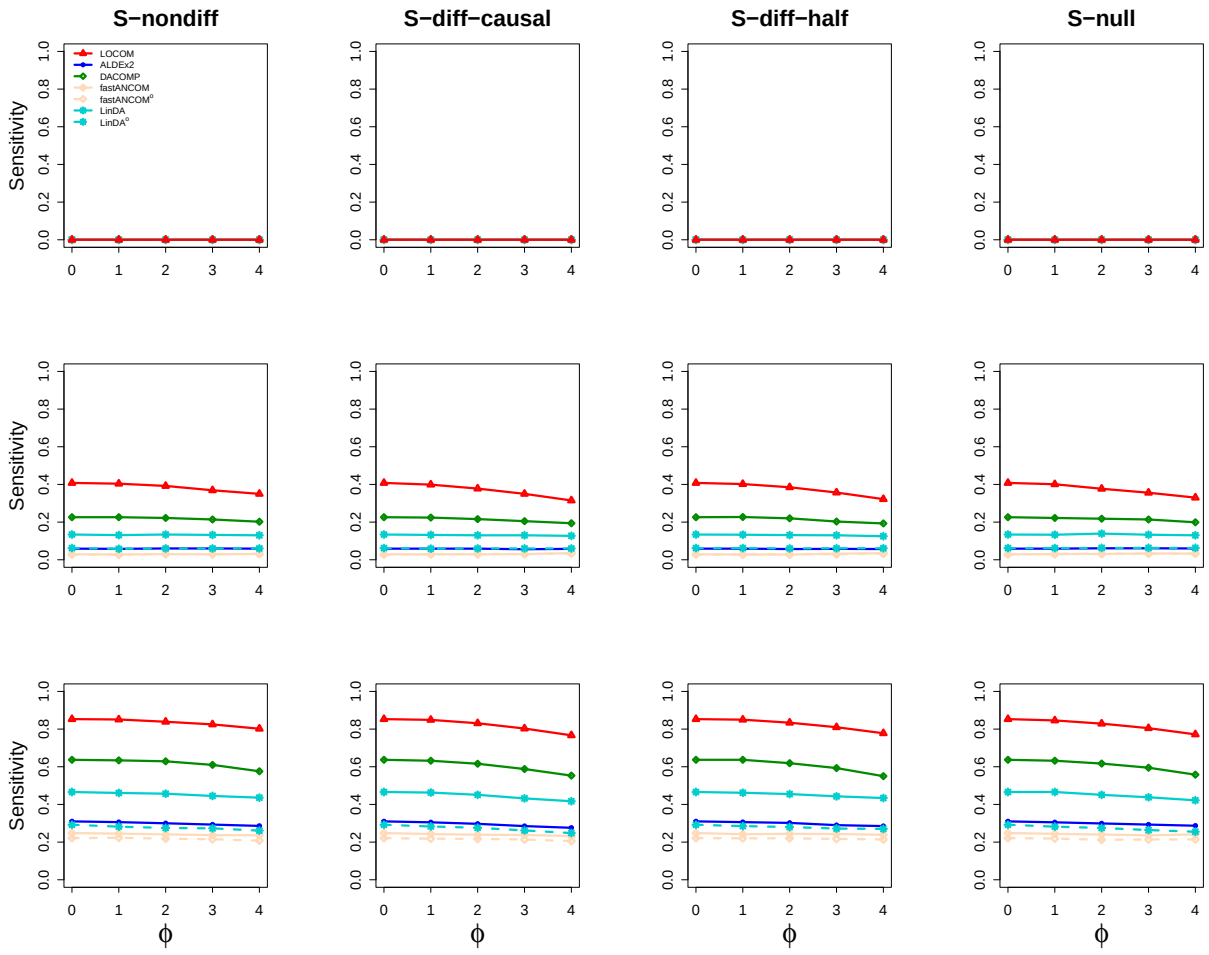


Figure B.8: Sensitivity results for data generated under M1, with a continuous trait. The rows correspond to fold change $\exp(\beta) = 1$, $\exp(\beta) = 2$ and $\exp(\beta) = 3$, respectively.

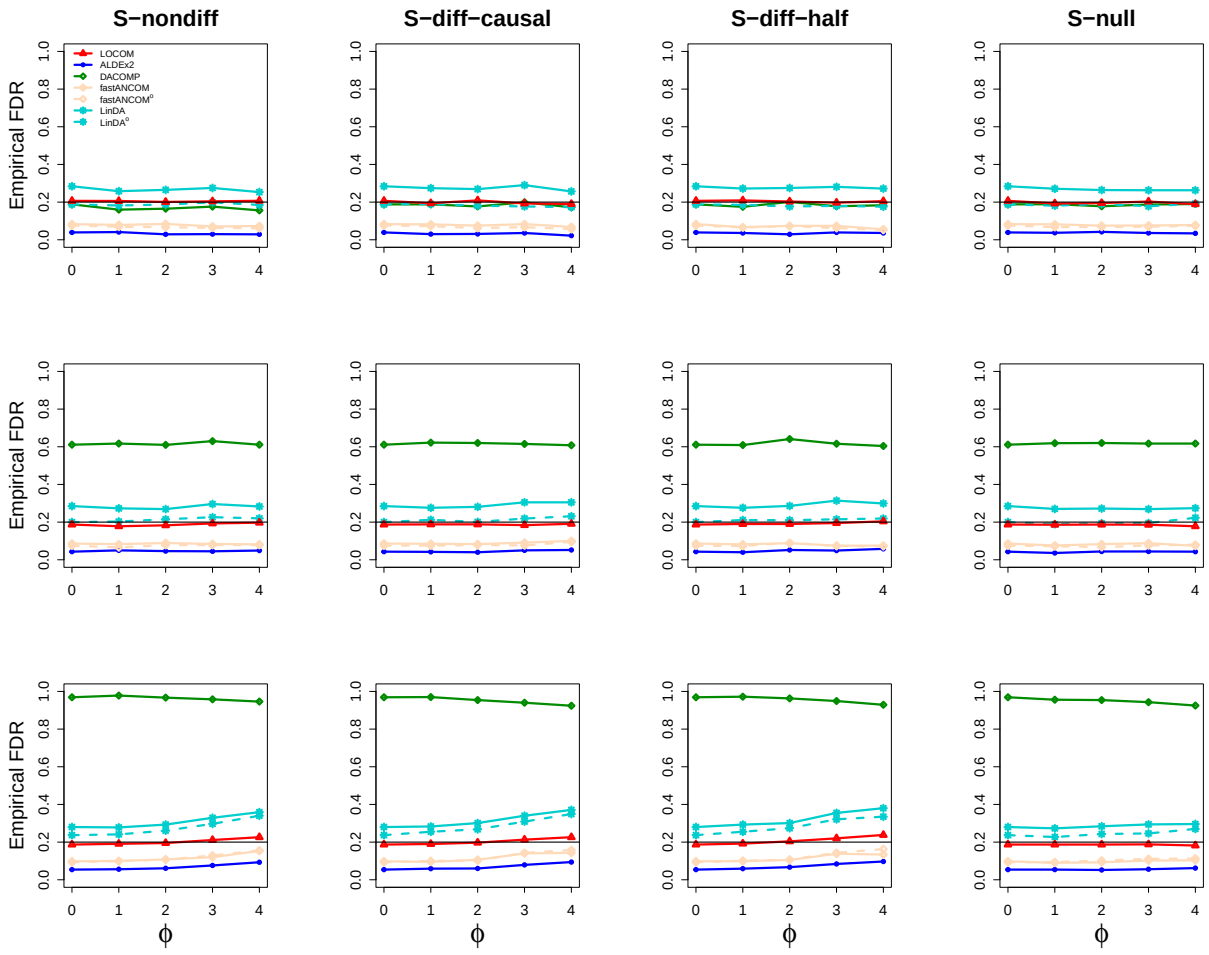


Figure B.9: Empirical FDR results for data generated under M2, with a continuous trait. The rows correspond to fold change $\exp(\beta) = 1$, $\exp(\beta) = 1.5$ and $\exp(\beta) = 2$, respectively.

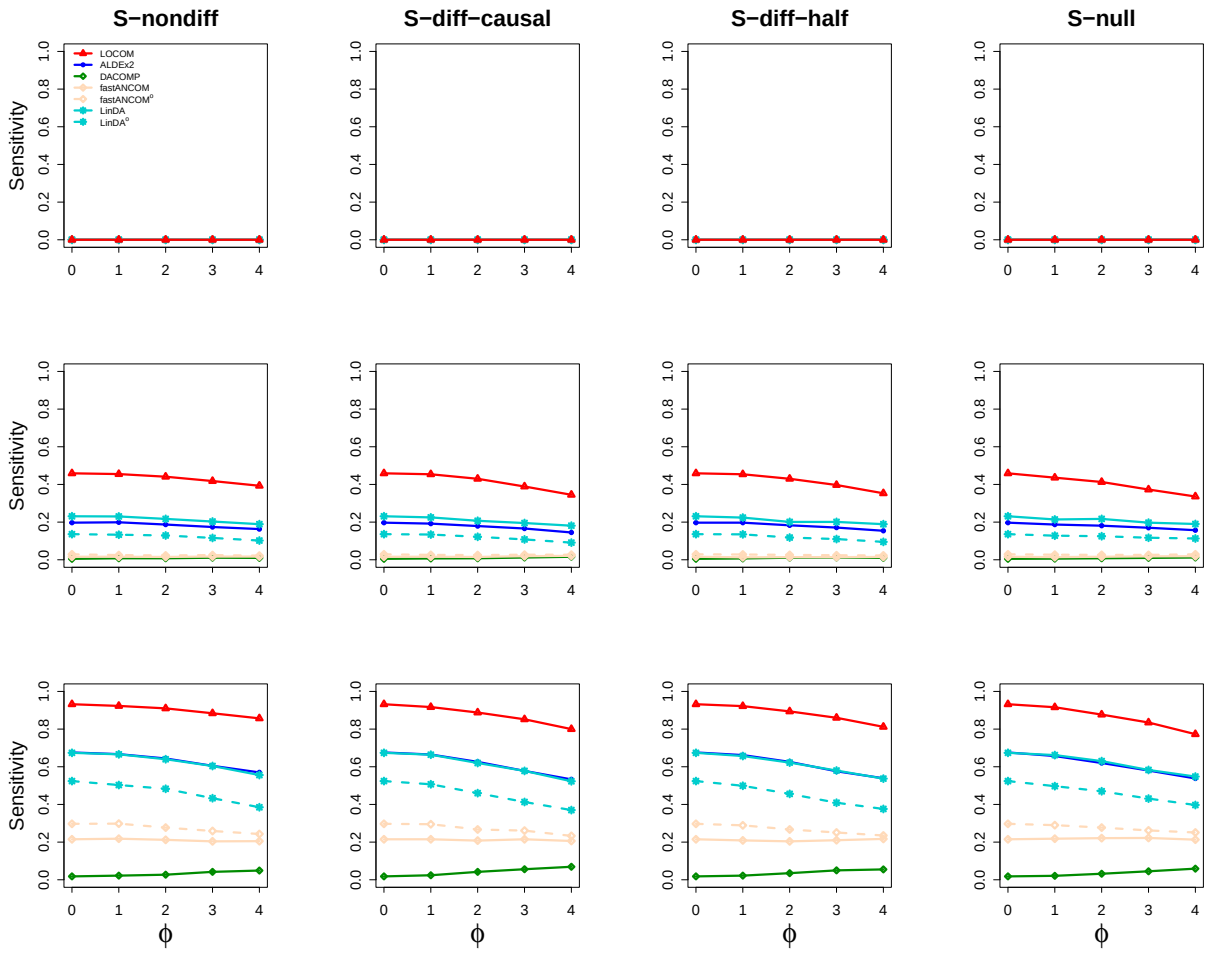


Figure B.10: Sensitivity results for data generated under M2, with a continuous trait. The rows correspond to fold change $\exp(\beta) = 1$, $\exp(\beta) = 1.5$ and $\exp(\beta) = 2$, respectively.

Appendix C

Appendix for Chapter 3

A large number of permutation schemes have been proposed for inference in linear models. We and others (Anderson and Legendre, 1999; Winkler et al., 2014) have found that the Freedman-Lane scheme both preserves type I error and optimizes test power. For example, we showed in Hu and Satten (2020) that permanovaFL, our implementation of PERMANOVA using the Freedman-Lane permutation, had a noticeable increase in power compared to adonis2, the implementation of PERMANOVA in the R package vegan, while still controlling type I error. We found a similar issue with the permutation scheme used in MiRKAT-S.

The test statistics used in MiRKAT-S when permutation-based inference is requested is M^TKM, where M is the vector of Martingale residuals and K is the $n \times n$ distance (kernel) matrix (see Plantinga et al. (2017) for more information on K). The observed value of the test statistics is compared to permutation distribution of $(\mathbb{P}_r M)^T K(\mathbb{P}_r M)$ for $r = 1, \ldots, R$, where $\mathbb{P}_r$ is the rth permutation matrix and R is the total number of permutation replicates. OMiSA adopted the same form of test statistic (except that K is the Euclidean distance matrix of power transformed relative abundance data) and the same permutation scheme, and thus has the same issue that MiRKAT-S has, as demonstrated below.

One way to see the role of permutation scheme on power is to use the original MiRKAT program to conduct a survival analysis by first obtaining the Martingale residuals from the

Cox Model, same as in MiRKAT-S, and then using the Martingale residual as a continuous outcome variable. Since the Martingale residual M_i has already accounted for the effect of covariates X_i, we could use MiRKAT to fit the model

$$M_i = f(Z_i) + \epsilon_i, \tag{C.1}$$

where Z_i denotes the microbiome data of all taxa from subject i and the $f(.)$ function is determined by the distance measure. Alternatively, we could fit the model

$$M_i = \beta_X X_i + f(Z_i) + \epsilon_i, \tag{C.2}$$

in which inclusion of X_i seems redundant as M_i is already a residual after accounting for X_i and, in fact, the two vectors $M = (M_1, \ldots, M_n)^T$ and $X = (X_1, \ldots, X_n)^T$ are orthogonal. However, if permutation is conducted by permuting M_i, i.e, replacing M by $\mathbb{P}_r M$, then these two models are different: model C.1 corresponds to what Winkler et al. (2014) call the "StillWhite" method in their Table 2, while model C.2 generates the Freedman-Lane method. The difference is that, after permuting M_i, $\mathbb{P}_r M$ and X are no longer exactly orthogonal, but including X_i in the model ensures that $\mathbb{P}_r M$ is orthogonal to X when $f(Z_i)$ is fit for each permutation. In our experience, this orthogonality is the source of the power advantage enjoyed by the Freedman-Lane approach. The power advantage is especially large when there is strong confounding effect and when the orthogonalization has a large effect.

Comparing the test statistics of MiRKAT-S with that of MiRKAT, we find that the permutation-based MiRKAT-S corresponds to fitting model C.1 in MiRKAT. If model C.2 is fit in MiRKAT, we find an improvement in the power of MiRKAT to essentially equal the power of our approach (i.e., our adaption of permanovaFL based on the Martingale residuals).

Table C.1: Type I error of the global tests for simulated data in other cases

				LDM-			permanovaFL-				
Censoring	n	Scenario	β_{XZ}	c	m	d	c	m	d	MiRKAT-S	OMiSALN
75%	100	M1	0	0.050	0.049	0.053	0.051	0.051	0.052	0.051	0.053
			0.8	0.052	0.050	0.048	0.050	0.050	0.050	0.032	0.038
			0.8*	0.332	0.329	0.314	0.239	0.228	0.228	0.233	0.261
		M2	0	0.048	0.047	0.049	0.049	0.048	0.050	0.051	0.052
			0.8	0.051	0.053	0.049	0.050	0.050	0.047	0.009	0.038
			0.8*	0.469	0.480	0.442	0.475	0.473	0.437	0.474	0.261
25%	100	M1	0	0.052	0.051	0.052	0.052	0.051	0.051	0.053	0.052
			0.8	0.047	0.044	0.048	0.049	0.046	0.049	0.030	0.033
			0.8*	0.806	0.812	0.725	0.624	0.628	0.593	0.643	0.697
		M2	0	0.043	0.045	0.044	0.045	0.044	0.045	0.047	0.052
			0.8	0.049	0.047	0.046	0.049	0.050	0.047	0.009	0.033
			0.8*	0.912	0.919	0.860	0.925	0.931	0.864	0.931	0.694
50%	100	M1	0	0.045	0.046	0.044	0.045	0.046	0.044	0.051	0.047
			0.8	0.045	0.047	0.047	0.049	0.050	0.048	0.032	0.027
			0.8*	0.335	0.327	0.294	0.218	0.217	0.207	0.225	0.304
		M2	0	0.042	0.044	0.043	0.047	0.046	0.046	0.050	0.047
			0.8	0.044	0.045	0.040	0.046	0.047	0.044	0.006	0.027
			0.8*	0.462	0.471	0.435	0.474	0.477	0.445	0.489	0.304

Note: See the note to Table 4.5. All event times here were simulated from the Cox model.

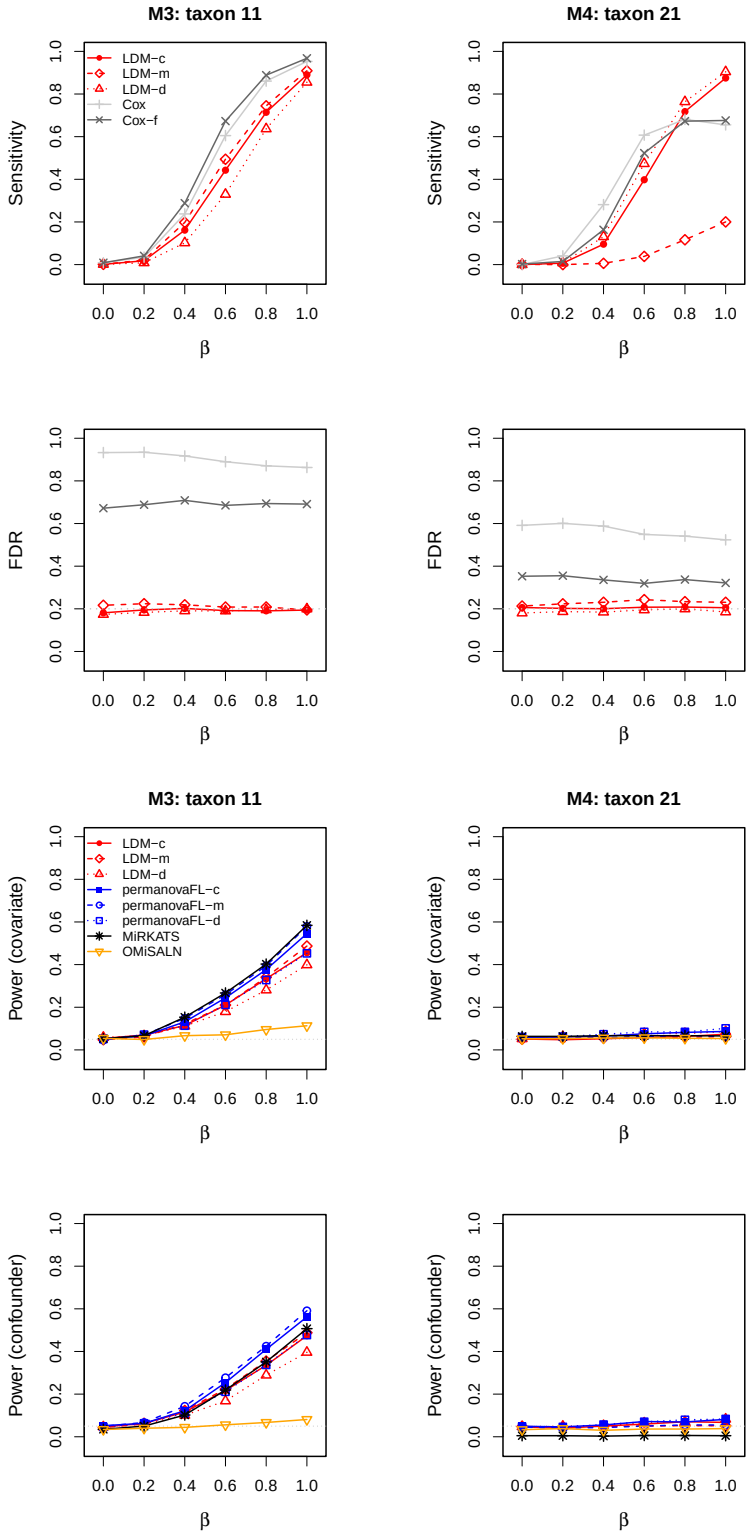


Figure C.1: Sensitivity and empirical FDR of the taxon-level tests in two more scenarios when the 11th and 21th taxa, respectively, were associated with the survival outcome.

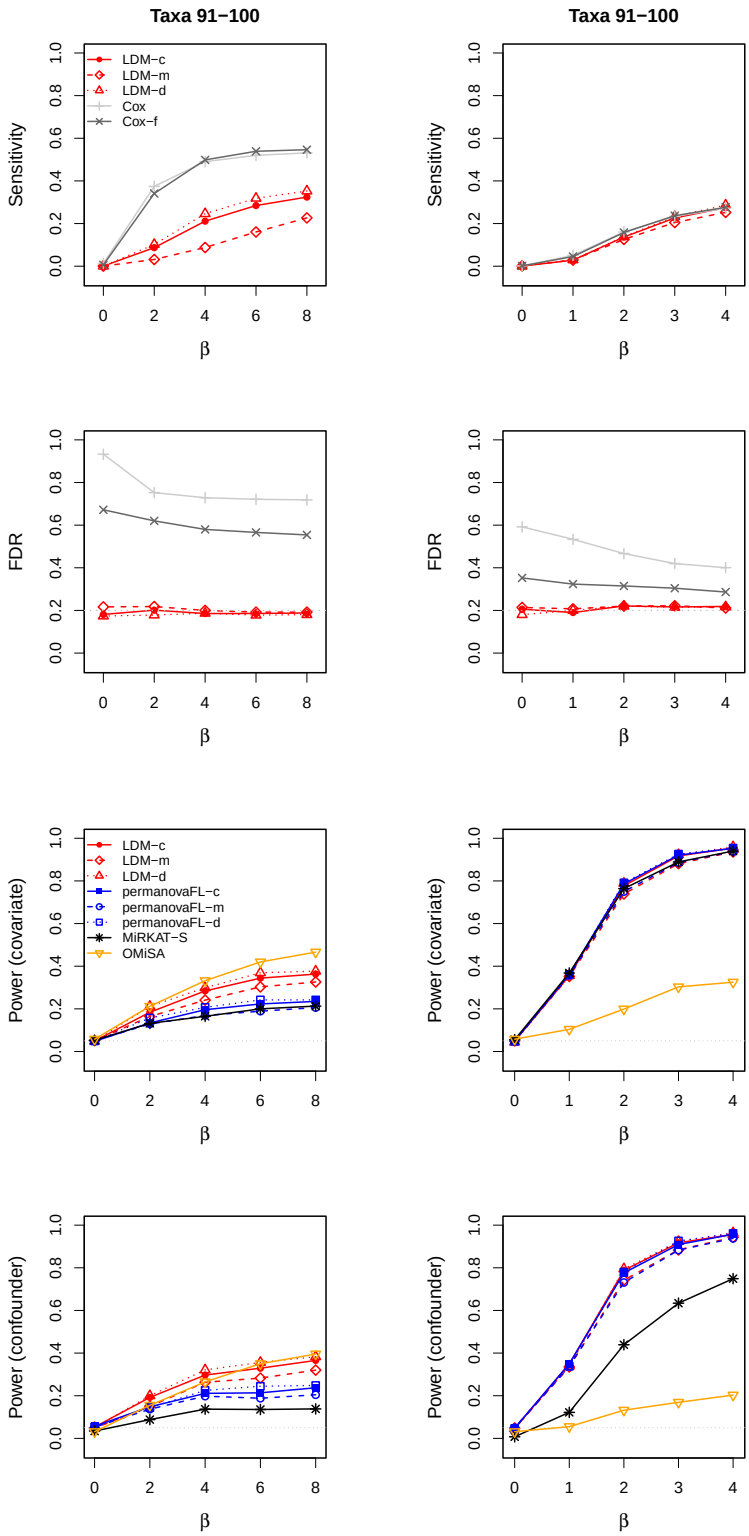


Figure C.2: Results in the scenario when rare taxa (taxa 91-100) were associated with the event time. Left column: data were simulated and analyzed based on the relative abundance scale, same as in model M1. Right column: data were simulated and analyzed based on the presence-absence scale (except for OMiSA), same as in model M2.The censoring rate was 50% and n = 100. Results of sensitivity and empirical FDR were obtained when X_i was a confounder ($\beta_{XZ} = 0.8$).

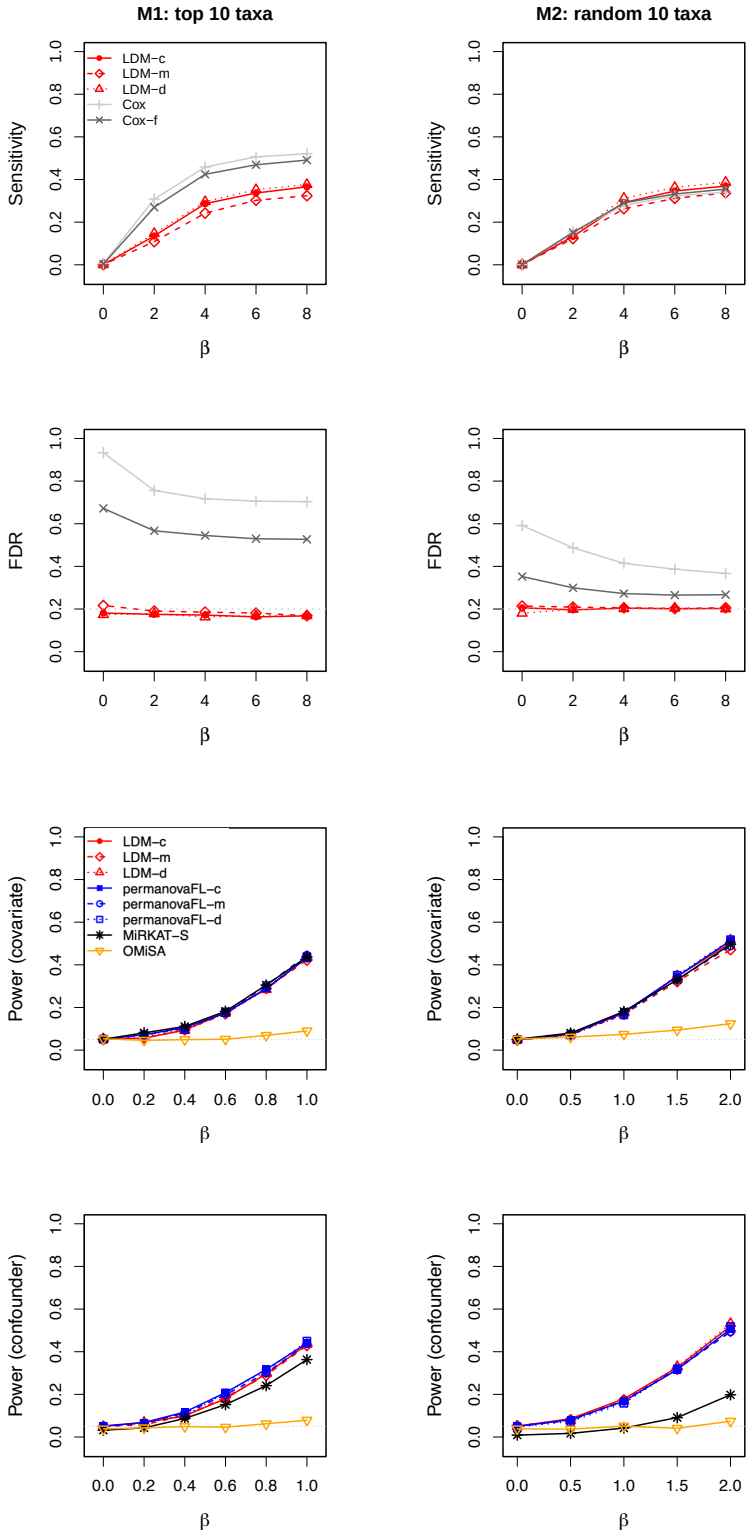


Figure C.3: Results for simulated data with 75% censoring and n = 100. Results of sensitivity and empirical FDR were obtained when X_i was a confounder ($\beta_{XZ} = 0.8$).

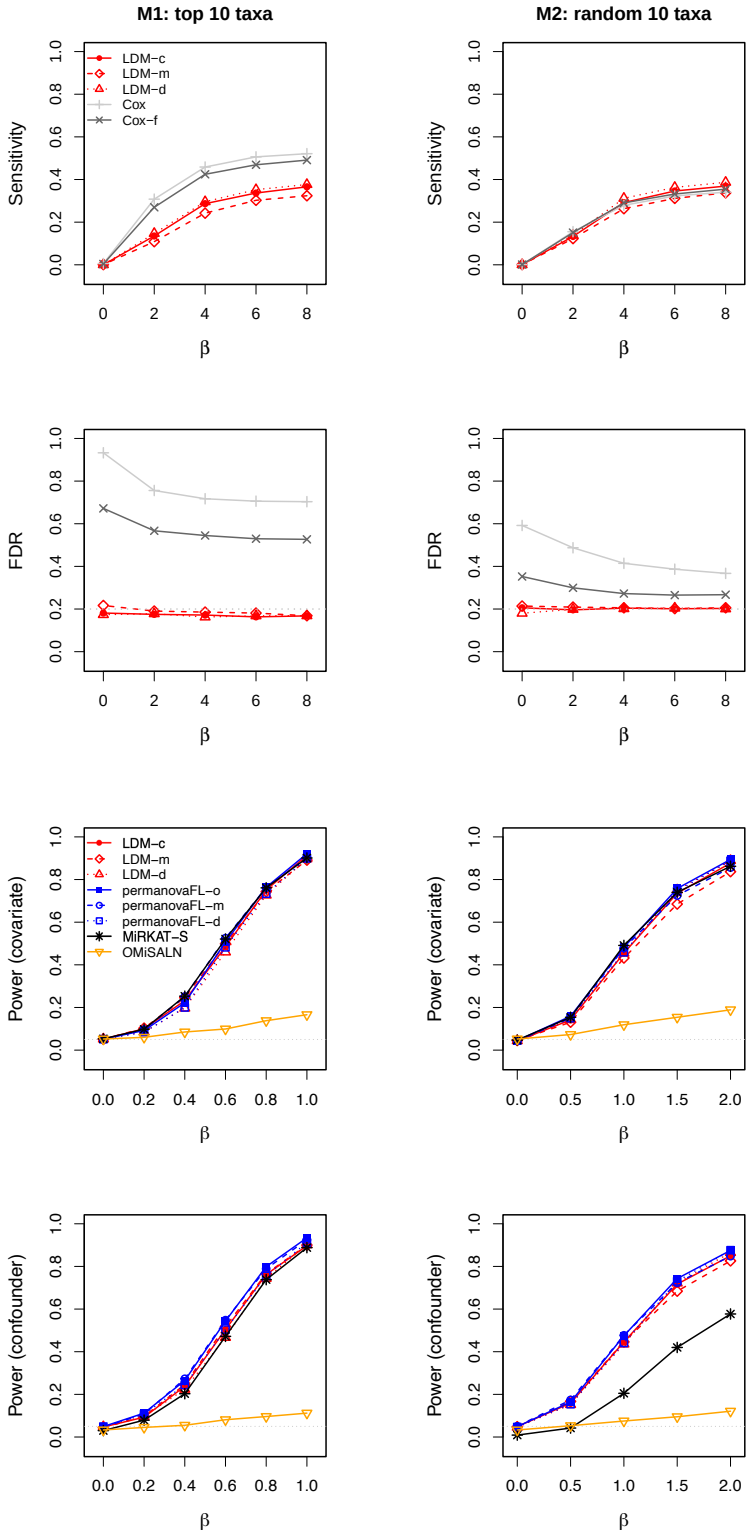


Figure C.4: Results for simulated data with 25% censoring and n = 100. Results of sensitivity and empirical FDR were obtained when X_i was a confounder ($\beta_{XZ} = 0.8$).

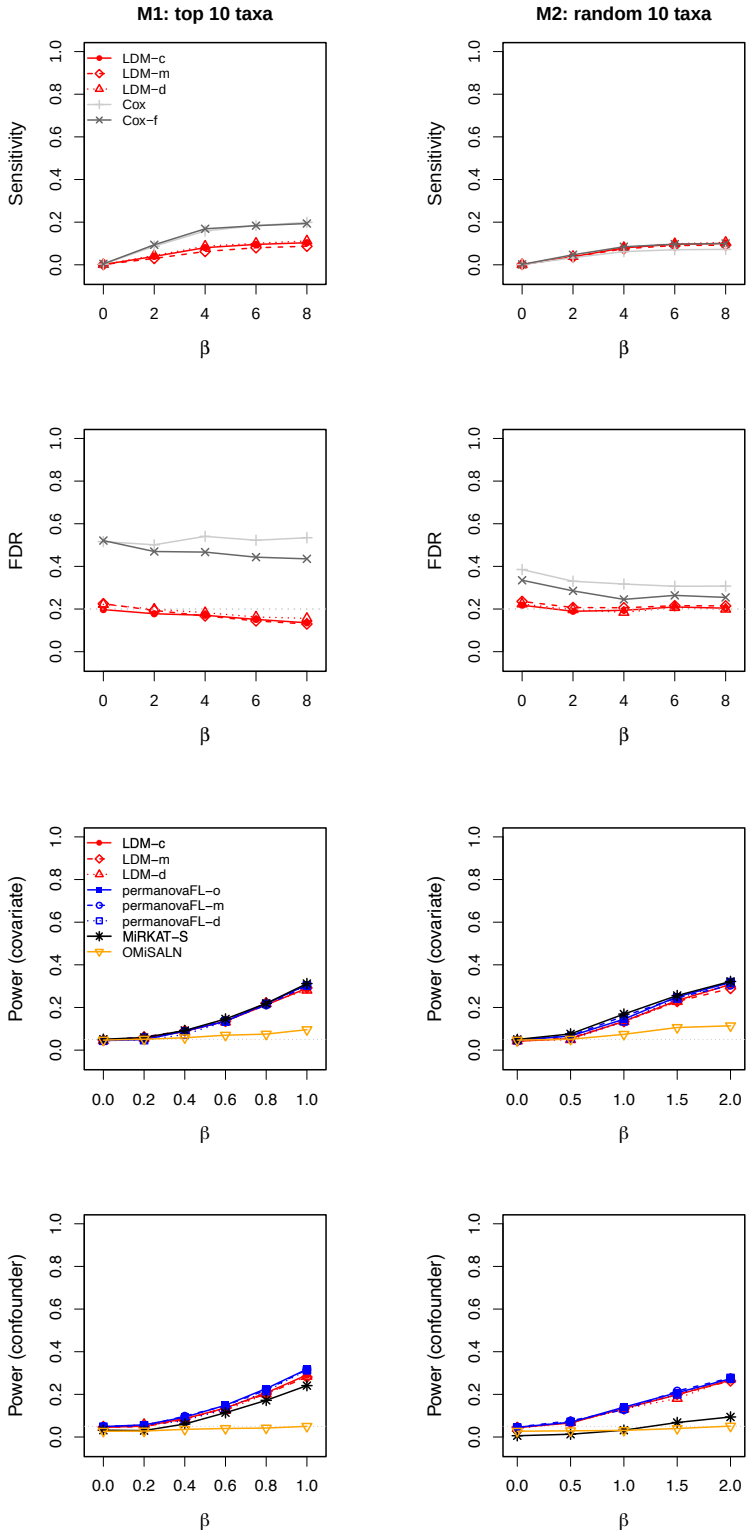
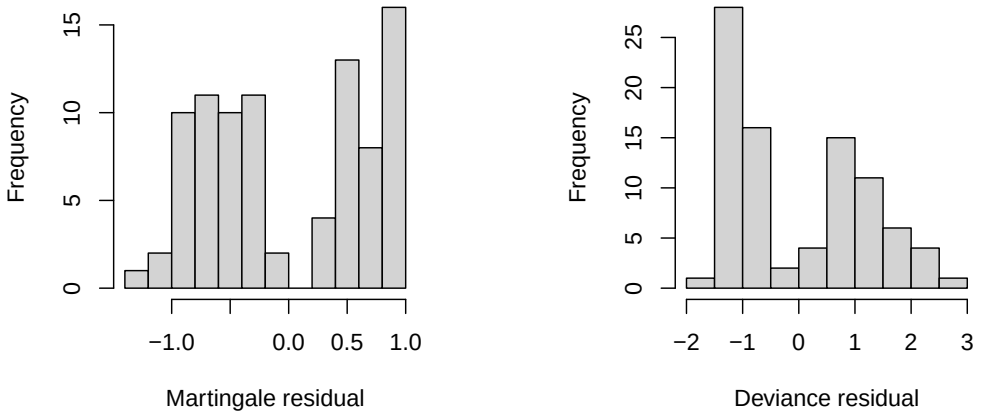


Figure C.5: Results for simulated data with 50% censoring and n = 50. Results of sensitivity and empirical FDR were obtained when X_i was a confounder ($\beta_{XZ} = 0.8$).

(a) Overall survival outcome



(b) Stage-III aGVHD outcome

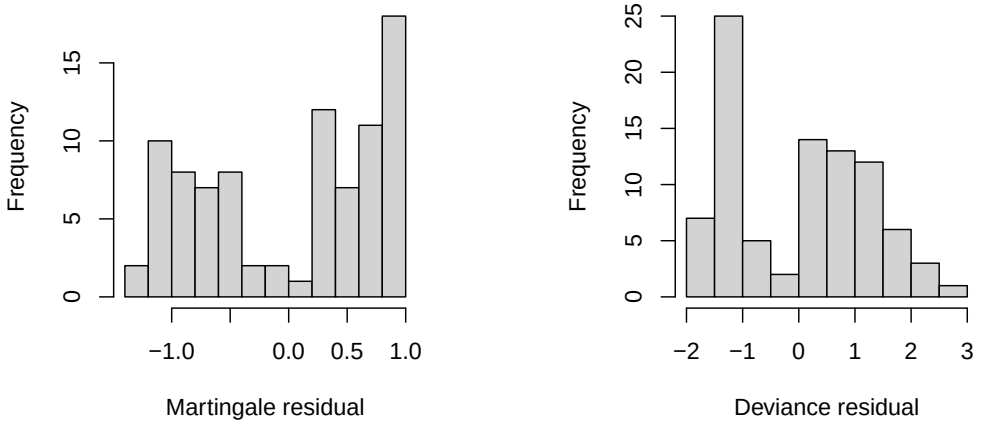


Figure C.6: Martingale and deviance residuals, generated from the Cox model that fit age and gender as covariates in analysis of the aGVHD data.

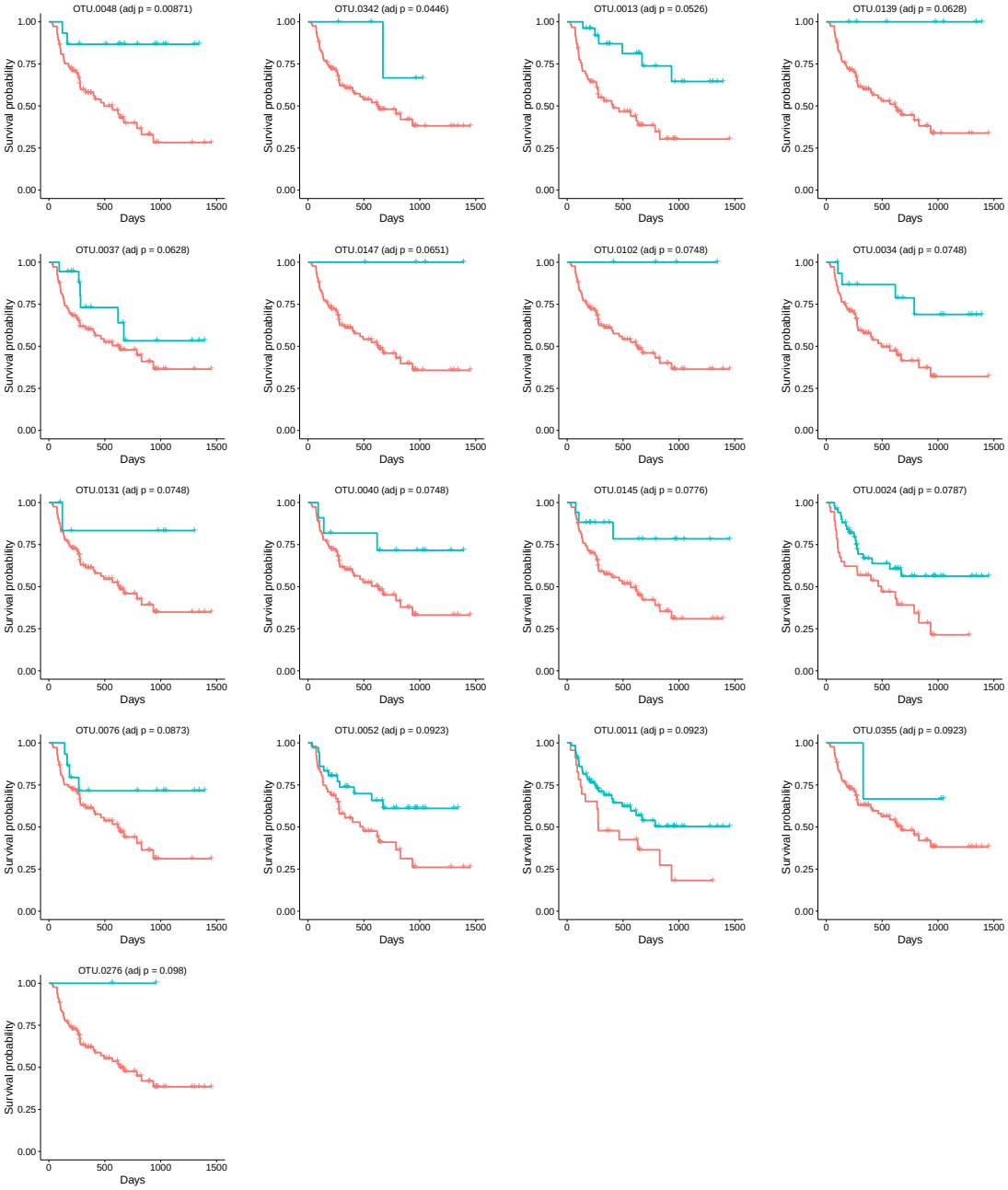


Figure C.7: Survival functions for the overall survival outcome by the presence (blue) and absence (red) status (based on a singly rarefied OTU table) of the OTUs detected by LDM-c. The plots were ordered by the adjusted p-values from LDM-c.

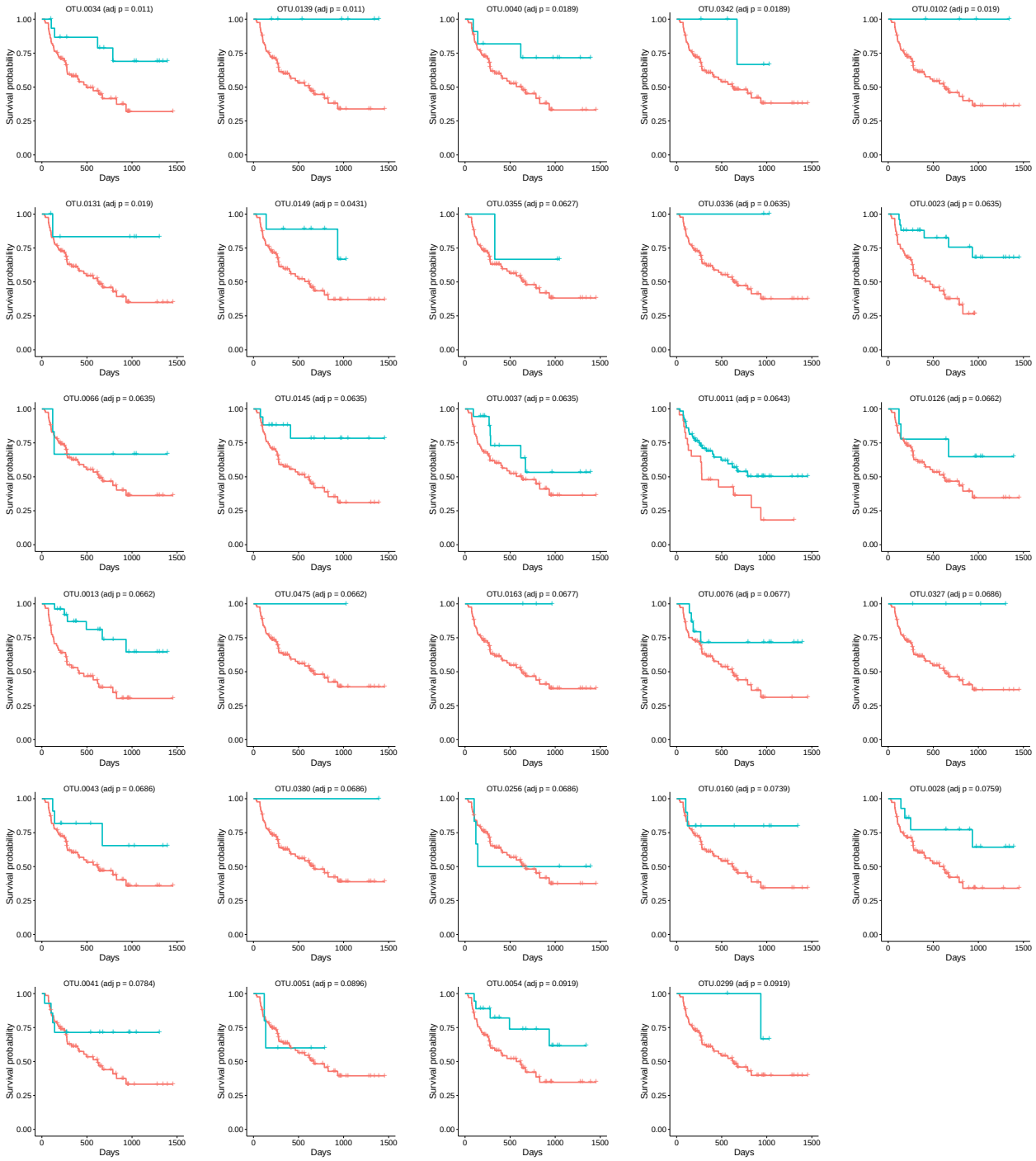


Figure C.8: See the caption to Figure C.7. The outcome is the time to stage-III aGVHD here.

Bibliography

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

Aitchison, J. (1986), The statistical analysis of compositional data, Chapman and Hall, London-New York.

Aitchison, J. and Bacon-Shone, J. (1984), 'Log contrast models for experiments with mixtures', Biometrika **71**(2), 323–330.

Aitchison, J., Barceló-Vidal, C., Martín-Fernández, J. A. and Pawlowsky-Glahn, V. (2000), 'Logratio analysis and compositional distance', Mathematical Geology **32**(3), 271–275.

Aitchison, J. and Ho, C. (1989), 'The multivariate poisson-log normal distribution', Biometrika **76**(4), 643–653.

Anderson, M. J. (2001), 'A new method for non-parametric multivariate analysis of variance', Austral ecology **26**(1), 32–46.

Anderson, M. J. (2005), 'Permutational multivariate analysis of variance', Department of Statistics, University of Auckland, Auckland **26**, 32–46.

Anderson, M. J. and Legendre, P. (1999), 'An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model', Journal of statistical computation and simulation **62**(3), 271–303.

Arumugam, M., Raes, J., Pelletier, E., Paslier, D. L., Yamada, T., Mende, D. R., Fernandes, G. R., Tap, J., Bruls, T., Batto, J.-M., Bertalan, M., Borruel, N., Casellas, F., Fernandez, L., Gautier, L., Hansen, T., Hattori, M., Hayashi, T., Kleerebezem, M., Kurokawa, K., Leclerc, M., Levenez, F., Manichanh, C., Nielsen, H. B., Nielsen, T., Pons, N., Poulain, J., Qin, J., Sicheritz-Ponten, T., Tims, S., Torrents, D., Ugarte, E., Zoetendal, E. G., Wang, J., Guarner, F., Pedersen, O., de Vos, W. M., Brunak, S., Doré, J., (additional members), M. C., Weissenbach, J., Ehrlich, S. D. and Bork, P. (2011), 'Enterotypes of the human gut microbiome', Nature **473**, 174–180.

Asher, J. E., Lamb, J. A., Brocklebank, D., Cazier, J.-B., Maestrini, E., Addis, L., Sen, M., Baron-Cohen, S. and Monaco, A. P. (2009), 'A whole-genome scan and fine-mapping linkage study of auditory-visual synesthesia reveals evidence of linkage to chromosomes 2q24, 5q33, 6p12, and 12p12', The American Journal of Human Genetics **84**(2), 279–285.

Atchinson, J. (2005), Concise guide to compositional data analysis, in 'In2do Compositional Data Analysis Workshop CoDaWork Oct', Vol. 5, pp. 17–21.

Bai, J., Hu, Y. and Bruner, D. (2019), 'Composition of gut microbiota and its association with body mass index and lifestyle factors in a cohort of 7–18 years old children from the american gut project', Pediatric obesity **14**(4), e12480.

BECG, C. B. and Gray, R. (1984), 'Calculation of polychotomous logistic regression parameters using individualized regressions', Biometrika **71**(1), 11–18.

Begg, C. B. and Gray, R. (1984), 'Calculation of polychotomous logistic regression parameters using individualized regressions', Biometrika **71**(1), 11–18.

Benjamini, Y. and Hochberg, Y. (1995a), 'Controlling the false discovery rate: a practical and powerful approach to multiple testing', Journal of the royal statistical society. Series B (Methodological) pp. 289–300.

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

Berg, G., Rybakova, D., Fischer, D., Cernava, T., Vergès, M.-C. C., Charles, T., Chen, X., Cocolin, L., Eversole, K., Corral, G. H. et al. (2020), 'Microbiome definition re-visited: old concepts and new challenges', Microbiome **8**(1), 1–22.

Berkson, J. (1944), 'Application of the logistic function to bio-assay', Journal of the American Statistical Association **39**(227), 357–365.

Berkson, J. (1953), 'A statistically precise and relatively simple method of estimating the bio-assay with quantal response, based on the logistic function', Journal of the American Statistical Association **48**(263), 565–599.

Besag, J. and Clifford, P. (1991), 'Sequential Monte Carlo p-values', Biometrika **78**(2), 301–304.

Bezdek, J. C. and Hathaway, R. J. (2003), 'Convergence of alternating optimization', Neural, Parallel & Scientific Computations **11**(4), 351–368.

Bi, W., Fritsche, L. G., Mukherjee, B., Kim, S. and Lee, S. (2020), 'A fast and accurate method for genome-wide time-to-event data analysis and its application to uk biobank', The American Journal of Human Genetics **107**(2), 222–233.

Boca, S., Heller, R. and Sampson, J. (2018), 'Multimed: Testing multiple biological mediators simultaneously', R package version **2**(0).

Boca, S. M., Sinha, R., Cross, A. J., Moore, S. C. and Sampson, J. N. (2014), 'Testing multiple biological mediators simultaneously', Bioinformatics **30**(2), 214–220.

Bogomolov, M. and Heller, R. (2018), 'Assessing replicability of findings across two studies of multiple features', Biometrika **105**(3), 505–516.

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C., Al-Ghalith, G. A., Alexander, H., Alm, E. J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J. E., Bittinger, K., Brejnrod, A., Brislawn, C. J., Brown, C. T., Callahan, B. J., Caraballo-Rodríguez, A. M., Chase, J., Cope, E., Da Silva, R., Dorrestein, P. C., Douglas, G. M., Durall, D. M., Duvallet, C., Edwardson, C. F., Ernst, M., Estaki, M., Fouquier, J., Gauglitz, J. M., Gibson, D. L., Gonzalez, A., Gorlick, K., Guo, J., Hillmann, B., Holmes, S., Holste, H., Huttenhower, C., Huttley, G., Janssen, S., Jarmusch, A. K., Jiang, L., Kaehler, B., Kang, K. B., Keefe, C. R., Keim, P., Kelley, S. T., Knights, D., Koester, I., Kosciolek, T., Kreps, J., Langille, M. G., Lee, J., Ley, R., Liu, Y., Loftfield, E., Lozupone, C., Maher, M., Marotz, C., Martin, B. D., McDonald, D., McIver, L. J., Melnik, A. V., Metcalf, J. L., Morgan, S. C., Morton, J., Naimey, A. T., Navas-Molina, J. A., Nothias, L. F., Orchanian, S. B., Pearson, T., Peoples, S. L., Petras, D., Preuss, M. L., Pruesse, E., Rasmussen, L. B., Rivers, A., Robeson II, M. S., Rosenthal, P., Segata, N., Shaffer, M., Shiffer, A., Sinha, R., Song, S. J., Spear, J. R., Swafford, A. D., Thompson, L. R., Torres, P. J., Trinh, P., Tripathi, A., Turnbaugh, P. J., Ul-Hasan, S., van der Hooft, J. J., Vargas, F., Vázquez-Baeza, Y., Vogtmann, E., von Hippel, M., Walters, W., Wan, Y., Wang, M., Warren, J., Weber, K. C., Williamson, C. H., Willis, A. D., Xu, Z. Z., Zaneveld, J. R., Zhang, Y., Zhu, Q., Knight, R. and Caporaso, J. G. (2018), QIIME 2: Reproducible, interactive, scalable, and extensible microbiome data science, Technical report, PeerJ Preprints.

Breslow, N. E., Day, N. E., Davis, W. et al. (1980), Statistical methods in cancer research: volume 1-the analysis of case-control studies, Vol. 32, IARC.

Brill, B., Amir, A. and Heller, R. (2019), 'Testing for differential abundance in compositional counts data, with application to microbiome studies', arXiv **XX**(XX).

Brooks, J. P. (2016), 'Challenges for case-control studies with microbiome data', Annals of epidemiology **26**(5), 336–341.

Brooks, J. P., Edwards, D. J., Harwich, M. D., Rivera, M. C., Fettweis, J. M., Serrano, M. G.,

Reris, R. A., Sheth, N. U., Huang, B., Girerd, P. et al. (2015), 'The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies', BMC microbiology **15**(1), 1–14.

Burns, M. B., Lynch, J., Starr, T. K., Knights, D. and Blekhman, R. (2015), 'Virulence genes are a signature of the microbiome in the colorectal tumor microenvironment', Genome Medicine **7**(1), 55.

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A. and Holmes, S. P. (2016), 'Dada2: high-resolution sample inference from illumina amplicon data', Nature methods **13**(7), 581.

Chao, A. and Chiu, C.-H. (2016), 'Bridging the variance and diversity decomposition approaches to beta diversity vis similarity and differentiation measures', Methods in Ecology and Evolution **7**(8), 919–928.

Charlson, E. S., Chen, J., Custers-Allen, R., Bittinger, K., Li, H., Sinha, R., Hwang, J., Bushman, F. D. and Collman, R. G. (2010), 'Disordered microbial communities in the upper respiratory tract of cigarette smokers', PloS one **5**(12), e15216. PMCID: PMC3004851.

Chen, E. Z. and Li, H. (2016), 'A two-part mixed-effects model for analyzing longitudinal microbiome compositional data', Bioinformatics **32**(17), 2611–2617. PMCID: PMC5860434.

Chen, H., Lumley, T., Brody, J., Heard-Costa, N. L., Fox, C. S., Cupples, L. A. and Dupuis, J. (2014), 'Sequence kernel association test for survival traits', Genetic epidemiology **38**(3), 191–197.

Chen, J. and Chen, L. (2017), 'Gmpr: A novel normalization method for microbiome sequencing data', bioRxiv p. 112565.

Chen, J. and Li, H. (2013), 'Variable selection for sparse dirichlet-multinomial regression with an application to microbiome data analysis', The annals of applied statistics **7**(1).

Chen, L., Reeve, J., Zhang, L., Huang, S., Wang, X. and Chen, J. (2018), 'Gmpr: A robust normalization method for zero-inflated count data with application to microbiome sequencing data', PeerJ **6**, e4600.

Chen, Y. Q. and Wang, M.-C. (2000), 'Analysis of accelerated hazards models', Journal of the American Statistical Association **95**(450), 608–618.

Costea, P. I., Zeller, G., Sunagawa, S. and Bork, P. (2014), 'A fair comparison', Nature Methods **11**, 359.

Costea, P. I., Zeller, G., Sunagawa, S., Pelletier, E., Alberti, A., Levenez, F., Tramontano, M., Driessen, M., Hercog, R., Jung, F.-E. et al. (2017), 'Towards standards for human fecal sample processing in metagenomic studies', Nature biotechnology **35**(11), 1069–1076.

Cox, D. R. (1972), 'Regression models and life-tables', Journal of the Royal Statistical Society: Series B (Methodological) **34**(2), 187–202.

Cox, M. J., Cookson, W. O. and Moffatt, M. F. (2013), 'Sequencing the human microbiome in health and disease', Human molecular genetics **22**(R1), R88–R94.

Dolan, K. T. and Chang, E. B. (2017), 'Diet, gut microbes, and the pathogenesis of inflammatory bowel diseases', Molecular nutrition & food research **61**(1), 1600129.

Dunlop, A. L., Satten, G. A., Hu, Y.-J., Knight, A. K., Hill, C. C., Wright, M. L., Smith, A. K., Read, T. D., Pearce, B. D. and Corwin, E. J. (2021), 'Vaginal microbiome composition in early pregnancy and risk of spontaneous preterm and early term birth among african american women', Frontiers in Cellular and Infection Microbiology **11**.

Everson, R. (1998), 'Orthogonal, but not orthonormal, procrustes problems', Advances in computational Mathematics **3**(4).

Fernandes, A. D., Reid, J. N., Macklaim, J. M., McMurrough, T. A., Edgell, D. R. and Gloor, G. B. (2014), '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.

Fettweis, J. M., Serrano, M. G., Brooks, J. P., Edwards, D. J., Girerd, P. H., Parikh, H. I., Huang, B., Arodz, T. J., Edupuganti, L., Glascock, A. L. et al. (2019), 'The vaginal microbiome and preterm birth', Nature medicine **25**(6), 1012–1021.

Firth, D. (1993), 'Bias reduction of maximum likelihood estimates', Biometrika **80**(1), 27–38.

Freedman, D. and Lane, D. (1983), 'A nonstochastic interpretation of reported significance levels', Journal of Business & Economic Statistics **1**(4), 292–298.

Friston, K. J., Penny, W. D. and Glaser, D. E. (2005), 'Conjunction revisited', Neuroimage **25**(3), 661–667.

Gandy, A. and Hahn, G. (2014), 'MMCTest-a safe algorithm for implementing multiple Monte Carlo tests', Scandinavian Journal of Statistics **41**(4), 1083–1101.

Gandy, A. and Hahn, G. (2016), 'A framework for Monte Carlo based multiple testing', Scandinavian Journal of Statistics **43**(4), 1046–1063.

Gandy, A. and Hahn, G. (2017), 'QuickMMCTest: quick multiple Monte Carlo testing', Statistics and Computing **27**(3), 823–832.

Georg, H. and Michael, S. (2002), 'A solution to the problem of separation in logistic regression', Statistics in Medicine **21**, 2409–2419.

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V. and Egozcue, J. J. (2017), 'Microbiome datasets are compositional: and this is not optional', Frontiers in microbiology **8**, 2224.

Gopalakrishnan, V., Spencer, C. N., Nezi, L., Reuben, A., Andrews, M., Karpinets, T., Prieto, P., Vicente, D., Hoffman, K., Wei, S. C. et al. (2018), 'Gut microbiome modulates response to anti–pd-1 immunotherapy in melanoma patients', Science **359**(6371), 97–103.

Gower, J. C. (1966), 'Some distance properties of latent root and vector methods used in multivariate analysis', Biometrika **53**(3-4), 325–338.

Grambsch, P. M. and Therneau, T. M. (1994), 'Proportional hazards tests and diagnostics based on weighted residuals', Biometrika **81**(3), 515–526.

Guo, W. and Peddada, S. (2008), 'Adaptive choice of the number of bootstrap samples in large scale multiple testing', Statistical applications in genetics and molecular biology **7**(1).

Guo, X., Pan, W., Connett, J. E., Hannan, P. J. and French, S. A. (2005), 'Small-sample performance of the robust score test and its modifications in generalized estimating equations', Statistics in medicine **24**(22), 3479–3495.

Haaland, R. E., Fountain, J., Hu, Y., Holder, A., Dinh, C., Hall, L., Pescatore, N. A., Heeke, S., Hart, C. E., Xu, J., Hu, Y.-J. and Kelley, C. (2018), 'Repeated rectal application of a hyperosmolar lubricant is associated with microbiota shifts but does not affect PrEP drug concentrations: results from a randomized trial in men who have sex with men', Journal of the International AIDS Society **21**(10), e25199. PMCID: PMC6207839.

Haldane, J. (1956), 'The estimation and significance of the logarithm of a ratio of frequencies', Annals of human genetics **20**(4), 309–311.

Hamidi, B., Wallace, K. and Alekseyenko, A. V. (2019), 'MODIMA, a method for multivariate omnibus distance mediation analysis, allows for integration of multivariate exposure-mediator-response relationships', Genes **10**(7), 524.

Han, M. K., Zhou, Y., Murray, S., Tayob, N., Noth, I., Lama, V. N., Moore, B. B., White, E. S., Flaherty, K. R., Huffnagle, G. B. et al. (2014), 'Lung microbiome and disease progression in idiopathic pulmonary fibrosis: an analysis of the comet study', The Lancet Respiratory Medicine **2**(7), 548–556.

Hawinkel, S., Mattiello, F., Bijnens, L. and Thas, O. (2017), 'A broken promise: microbiome differential abundance methods do not control the false discovery rate', Briefings in bioinformatics **20**(1), 210–221.

Heard, N. A. and Rubin-Delanchy, P. (2018), 'Choosing between methods of combining-values', Biometrika **105**(1), 239–246.

Heinze, G. and Dunkler, D. (2008), 'Avoiding infinite estimates of time-dependent effects in small-sample survival studies', Statistics in medicine **27**(30), 6455–6469.

Heinze, G. and Schemper, M. (2001), 'A solution to the problem of monotone likelihood in cox regression', Biometrics **57**(1), 114–119.

Holm, S. (1979), 'A simple sequentially rejective multiple test procedure', Scandinavian journal of statistics pp. 65–70.

Hu, J., Koh, H., He, L., Liu, M., Blaser, M. J. and Li, H. (2018), 'A two-stage microbial association mapping framework with advanced FDR control', Microbiome **6**(1), 131. PMCID: PMC6060480.

Hu, Y. J., Berndt, S. I., Gustafsson, S., Ganna, A., Hirschhorn, J., North, K. E., Ingelsson, E. and Lin, D.-Y. (2013), 'Meta-analysis of gene-level associations for rare variants based on single-variant statistics', American Journal of Human Genetics **93**(2), 236–248. PMCID: PMC3738834.

Hu, Y.-J., Lane, A. and Satten, G. A. (2021), 'A rarefaction-based extension of the ldm for testing presence-absence associations in the microbiome', Bioinformatics p. https://doi.org/10.1093/bioinformatics/btab012.

Hu, Y. J., Li, Y., Auer, P. L. and Lin, D. Y. (2015), 'Integrative analysis of sequencing and array genotype data for discovering disease associations with rare mutations', Proceedings of the National Academy of Sciences **112**(4), 1019–1024. PMCID: PMC4313847.

Hu, Y.-J., Liao, P., Johnston, H. R., Allen, A. S. and Satten, G. A. (2016), 'Testing rare-variant association without calling genotypes allows for systematic differences in sequencing between cases and controls', PLoS genetics **12**(5), e1006040. PMCID: PMC4859496.

Hu, Y. J. and Lin, D. Y. (2010), 'Analysis of untyped SNPs: maximum likelihood and imputation methods', Genetic Epidemiology **34**(8), 803–815. PMCID: PMC3030127.

Hu, Y. J., Lin, D. Y., Sun, W. and Zeng, D. (2014), 'A likelihood-based framework for association analysis of allele-specific copy numbers', Journal of the American Statistical Association **109**(508), 1533–1545. PMCID: PMC4315366.

Hu, Y. J., Lin, D. Y. and Zeng, D. (2010), 'A general framework for studying genetic effects and gene–environment interactions with missing data', Biostatistics **11**(4), 583–598. PMCID: PMC3294269.

Hu, Y.-J. and Satten, G. A. (2020), 'Testing hypotheses about the microbiome using the linear decomposition model (ldm)', Bioinformatics **36**(14), 4106–4115.

Hu, Y.-J. and Satten, G. A. (2021), 'A rarefaction-without-resampling extension of permanova for testing presence-absence associations in the microbiome', bioRxiv p. https://doi.org/10.1101/2021.04.06.438671.

Hu, Y.-J., Sun, W., Tzeng, J.-Y. and Perou, C. M. (2015), 'Proper use of allele-specific expression improves statistical power for cis-eQTL mapping with RNA-seq data', Journal of the American Statistical Association **110**(511), 962–974. PMCID: PMC4642818.

Hu, Y., Satten, G. A. and Hu, Y.-J. (2021a), 'Locom: A logistic regression model for testing differential abundance in compositional microbiome data with false discovery rate control', bioRxiv p. doi.org/10.1101/2021.10.03.462964.

Hu, Y., Satten, G. A. and Hu, Y.-J. (2021b), 'Locom: A logistic regression model for testing

differential abundance in compositional microbiome data with false discovery rate control', bioRxiv p. https://doi.org/10.1101/2021.10.03.462964.

Hugerth, L. W. and Andersson, A. F. (2017), 'Analysing microbial community composition through amplicon sequencing: from sampling to hypothesis testing', Frontiers in microbiology **8**, 1561.

Hughes, J. B. and Hellmann, J. J. (2005), 'The application of rarefaction techniques to molecular inventories of microbial diversity', Methods in enzymology **397**, 292–308.

Jenq, R. R., Taur, Y., Devlin, S. M., Ponce, D. M., Goldberg, J. D., Ahr, K. F., Littmann, E. R., Ling, L., Gobourne, A. C., Miller, L. C. et al. (2015), 'Intestinal blautia is associated with reduced death from graft-versus-host disease', Biology of Blood and Marrow Transplantation **21**(8), 1373–1383.

Jiang, H. and Salzman, J. (2012), 'Statistical properties of an early stopping rule for resampling-based multiple testing', Biometrika **99**(4), 973–980.

Kaul, A., Mandal, S., Davidov, O. and Peddada, S. D. (2017), 'Analysis of microbiome data in the presence of excess zeros', Frontiers in microbiology **8**, 2114. PMCID: PMC5682008.

Kim, Y., Lim, J., Lee, J. S. and Jeong, J. (2018), 'Controlling two-dimensional false discovery rates by combining two univariate multiple testing results with an application to mass spectral data', Chemometrics and Intelligent Laboratory Systems **182**, 149–157.

Kleinbaum, D. G., Kupper, L. L., Nizam, A. and Muller, K. G. (2007), Applied Regression Analysis and Other Multivariable Methods, Duxbury Press.

Koeth, R. A., Wang, Z., Levison, B. S., Buffa, J. A., Org, E., Sheehy, B. T., Britt, E. B., Fu, X., Wu, Y., Li, L., Smith, J. D., DiDonato, J. A., Chen, J., Li, H., Wu, G. D., Lewis, J. D., Warrier, M., Brown, J. M., Krauss, R. M., Tang, W. H., Bushman, F. D., Lusis, A. J.

and Hazen, S. L. (2013), 'Intestinal microbiota metabolism of l-carnitine, a nutrient in red meat, promotes atherosclerosis', Nature medicine **19**(5), 576. PMCID: PMC3650111.

Koh, H., Blaser, M. J. and Li, H. (2017), 'A powerful microbiome-based association test and a microbial taxa discovery framework for comprehensive association mapping', Microbiome **5**(1), 45.

Koh, H., Li, Y., Zhan, X., Chen, J. and Zhao, N. (2019), 'A distance-based kernel association test based on the generalized linear mixed model for correlated microbiome studies', Frontiers in genetics **10**, 458. PMCID: PMC6532659.

Koh, H., Livanos, A. E., Blaser, M. J. and Li, H. (2018), 'A highly adaptive microbiome-based association test for survival traits', BMC genomics **19**(1), 1–13.

Koren, O., Knights, D., Gonzalez, A., Waldron, L., Segata, N., Knight, R., Huttenhower, C. and Ley1, R. E. (2013), 'Analysis of microbial community structures in human microbiome datasets', PLOW Computational Biology **9**, e1002863.

Kumar, M. S., Slud, E. V., Okrah, K., Hicks, S. C., Hannenhalli, S. and Bravo, H. C. (2018), 'Analysis and correction of compositional bias in sparse sequencing count data', BMC genomics **19**(1), 799.

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

La Rosa, P. S., Warner, B. B., Zhou, Y., Weinstock, G. M., Sodergren, E., Hall-Moore, C. M., Stevens, H. J., Bennett, W. E., Shaikh, N., Linneman, L. A., Hoffmann, J. A., Hamvas, A., Deych, E., Shands, B. A., Shannon, W. D. and Tarr, P. (2014), 'Patterned progression of bacterial populations in the premature infant gut', Proceedings of the National Academy of Sciences **111**(34), 12522–12527. PMCID: PMC4151715.

Laudadio, I., Fulci, V., Stronati, L. and Carissimi, C. (2019), 'Next-generation metagenomics: Methodological challenges and opportunities', Omics: a journal of integrative biology **23**(7), 327–333.

Legendre, P. and Anderson, M. J. (1999), 'Distance-based redundancy analysis: testing multispecies responses in multifactorial ecological experiments', Ecological monographs **69**(1), 1–24.

Legendre, P. and Gallagher, E. D. (2001), 'Ecologically meaningful transformations for ordination of species data', Oecologia **129**(2), 271–280.

Legendre, P. and Legendre, L. F. (2012), Numerical ecology, Elsevier.

Ley, R. E., Peterson, D. A. and Gordon, J. I. (2006), 'Ecological and evolutionary forces shaping microbial diversity in the human intestine', Cell **124**(4), 837–848.

Li, J., Witten, D. M., Johnstone, I. M. and Tibshirani, R. (2012), 'Normalization, testing, and false discovery rate estimation for RNA-sequencing data', Biostatistics **13**(3), 523–538.

Li, Y., Hu, Y.-J. and Satten, G. A. (2019), 'A bottom-up approach to testing hypotheses that have a branching tree dependence structure, with false discovery rate control', arXiv:1903.06850 .

Liao, P., Satten, G. A. and Hu, Y.-J. (2017a), 'PhredEM: a phred-score-informed genotype-calling approach for next-generation sequencing studies', Genetic epidemiology **41**(5), 375–387. PMCID: PMC5564424.

Liao, P., Satten, G. A. and Hu, Y.-J. (2017b), 'Robust inference of population structure from next-generation sequencing data with systematic differences in sequencing', Bioinformatics **34**(7), 1157–1163. PMCID: PMC6031038.

Lin, D. Y., Hu, Y. J. and Huang, B. E. (2008), 'Simple and efficient analysis of disease

association with missing genotype data', American Journal of Human Genetics **82**(2), 444–452. PMCID: PMC2427170.

Lin, H. and Peddada, S. D. (2020), 'Analysis of compositions of microbiomes with bias correction', Nature communications **11**(1), 1–11.

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

Liu, Y., Chen, S., Li, Z., Morrison, A. C., Boerwinkle, E. and Lin, X. (2019), 'Acat: a fast and powerful p value combination method for rare-variant analysis in sequencing studies', The American Journal of Human Genetics **104**(3), 410–421.

Liu, Y. and Xie, J. (2020), 'Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures', Journal of the American Statistical Association **115**(529), 393–402.

Loughin, T. M. (2004), 'A systematic comparison of methods for combining p-values from independent tests', Computational statistics & data analysis **47**(3), 467–485.

Love, M. I., Huber, W. and Anders, S. (2014), 'Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2', Genome biology **15**(12), 550. PMCID: PMC4302049.

Lozupone, C. A., Hamady, M., Kelley, S. T. and Knight, R. (2007), 'Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities', Applied and environmental microbiology **73**(5), 1576–1585. PMCID: PMC1828774.

Lozupone, C. and Knight, R. (2005), 'UniFrac: a new phylogenetic method for comparing microbial communities', Applied and environmental microbiology **71**(12), 8228–8235. PMCID: PMC1317376.

Lozupone, C., Lladser, M. E., Knights, D., Stombaugh, J. and Knight, R. (2011), 'Unifrac: an effective distance metric for microbial community comparison', The ISME journal **5**(2), 169.

Lu, J., Shi, P. and Li, H. (2019), 'Generalized linear models with linear constraints for microbiome compositional data', Biometrics **75**(1), 235–244.

Magne, F., Gotteland, M., Gauthier, L., Zazueta, A., Pesoa, S., Navarrete, P. and Balamurugan, R. (2020), 'The firmicutes/bacteroidetes ratio: A relevant marker of gut dysbiosis in obese patients?', Nutrients **12**, 1474.

Majumdar, A., Witte, J. S. and Ghosh, S. (2015), 'Semiparametric allelic tests for mapping multiple phenotypes: Binomial regression and Mahalanobis distance', Genetic Epidemiology **39**(8), 635–650.

Mallick, H., Tickle, T., McIver, L., Rahnavard, G., Nguyen, L., Weingart, G., Ma, S., Ren, B., Schwager, E., Subramanian, A., Paulson, J., Franzosa, E., Corrada, B. H. and Huttenhower, C. (2019), 'Multivariable association in population-scale meta'omic surveys', In Submission .

Mancl, L. A. and DeRouen, T. A. (2001), 'A covariance estimator for gee with improved small-sample properties', Biometrics **57**(1), 126–134.

Mandal, S., Van Treuren, W., White, R. A., Eggesbø, M., Knight, R. and Peddada, S. D. (2015), 'Analysis of composition of microbiomes: a novel method for studying microbial composition', Microbial ecology in health and disease **26**(1), 27663.

Mariat, D., Firmesse, O., Levenez, F., Guimarăes, V., Sokol, H., Doré, J., Corthier, G. and Furet, J.-P. (2009), 'The firmicutes/bacteroidetes ratio of the human microbiota changes with age', BMC Microbiology **9**, 123.

Matson, V., Fessler, J., Bao, R., Chongsuwat, T., Zha, Y., Alegre, M.-L., Luke, J. J. and Gajewski, T. F. (2018), 'The commensal microbiome is associated with anti–pd-1 efficacy in metastatic melanoma patients', Science **359**(6371), 104–108.

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

McCullagh, P. and Nelder, J. A. (2019), Generalized linear models, Routledge.

McLaren, M. R., Willis, A. D. and Callahan, B. J. (2019), 'Consistent and correctable bias in metagenomic sequencing experiments', Elife **8**.

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

Mitchell, C. M., Srinivasan, S., Zhan, X., Wu, M. C., Reed, S. D., Guthrie, K. A., LaCroix, A. Z., Fiedler, T., Munch, M., Liu, C. et al. (2017), 'Vaginal microbiota and genitourinary menopausal symptoms: a cross-sectional analysis', Menopause **24**(10), 1160–1166.

Morgan, J. L., Darling, A. E. and Eisen, J. A. (2010), 'Metagenomic sequencing of an in vitro-simulated microbial community', PloS one **5**(4), e10209.

Morgan, X. C., Kabakchiev, B., Waldron, L., Tyler, A. D., Tickle, T. L., Milgrom, R., Stempak, J. M., Gevers, D., Xavier, R. J., Silverberg, M. S. et al. (2015), 'Associations between host gene expression, the mucosal microbiome, and clinical outcome in the pelvic pouch of patients with inflammatory bowel disease', Genome biology **16**(1), 67.

Muller, K. E. and Fetterman, B. A. (2012), Regression and ANOVA: An Integrated Approach using SAS Software, SAS Institute.

Navas-Molina, J. A., Peralta-Sánchez, J. M., González, A., McMurdie, P. J., Vázquez-Baeza, Y., Xu, Z., Ursell, L. K., Lauber, C., Zhou, H., Song, S. J., Huntley, J., Ackermann, G. L., Berg-Lyons, D., Holmes, S., Caporaso, J. G. and Knight, R. (2013), Advancing our

understanding of the human microbiome using QIIME, in 'Methods in enzymology', Vol. 531, Elsevier, pp. 371–444.

Ni, J., Wu, G. D., Albenberg, L. and Tomov, V. T. (2017), 'Gut microbiota and ibd: causation or correlation?', Nature reviews Gastroenterology & hepatology **14**(10), 573–584.

Nichols, T., Brett, M., Andersson, J., Wager, T. and Poline, J.-B. (2005), 'Valid conjunction inference with the minimum statistic', Neuroimage **25**(3), 653–660.

Nyante, S. J., Gammon, M. D., Kaufman, J. S., Bensen, J. T., Lin, D. Y., Barnholtz-Sloan, J. S., Hu, Y., He, Q., Luo, J. and Millikan, R. C. (2011), 'Common genetic variation in adiponectin, leptin, and leptin receptor and association with breast cancer subtypes', Breast Cancer Research and Treatment **129**(2), 593–606.

Olesen, S. W., Vora, S., Techtmann, S. M., Fortney, J. L., Bastidas-Oyanedel, J. R., Rodríguez, J., Hazen, T. C. and Alm, E. J. (2016), 'A novel analysis method for paired-sample microbial ecology experiments', PloS one **11**(5), e0154804.

O'Reilly, P. F., Hoggart, C. J., Pomyen, Y., Calboli, F. C., Elliott, P., Jarvelin, M.-R. and Coin, L. J. (2012), 'MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS', PloS One **7**(5).

Palarea-Albaladejo, J. and Martin-Fernandez, J. A. (2015), 'zCompositions–R package for multivariate imputation of left-censored data under a compositional approach', Chemometrics and Intelligent Laboratory Systems **143**, 85–96.

Paulson, J. N., Bravo, H. C. and Mihai, P. (2014), 'Reply to: "a fair comparison"', Nature Methods **11**, 359–360.

Paulson, J. N., Stine, O. C., Bravo, H. C. and Pop, M. (2013), 'Differential abundance analysis for microbial marker-gene surveys', Nature Methods **10**(12), 1200–1202. PMCID: PMC4010126.

Pawlowsky-Glahn, V., Egozcue, J. J. and Tolosana-Delgado, R. (2015), Modeling and analysis of compositional data, John Wiley & Sons.

Phipson, B. and Smyth, G. K. (2010), 'Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn', Statistical applications in genetics and molecular biology **9**(1).

Plantinga, A., Zhan, X., Zhao, N., Chen, J., Jenq, R. R. and Wu, M. C. (2017), 'Mirkats: a community-level test of association between the microbiota and survival times', Microbiome **5**(1), 1–13.

Pollock, J., Glendinning, L., Wisedchanwet, T. and Watson, M. (2018), 'The madness of microbiome: attempting to find consensus "best practice" for 16s microbiome studies', Applied and environmental microbiology **84**(7), e02627–17.

Pope, J. L., Tomkovich, S., Yang, Y. and Jobin, C. (2017), 'Microbiota as a mediator of cancer progression and therapy', Translational Research **179**, 139–154.

Potter, D. M. (2005), 'A permutation test for inference in logistic regression with small-and moderate-sized data sets', Statistics in medicine **24**(5), 693–708.

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T. et al. (2010), 'A human gut microbial gene catalogue established by metagenomic sequencing', nature **464**(7285), 59–65.

Randolph, T. W., Zhao, S., Copeland, W., Hullar, M. and Shojaie, A. (2018), 'Kernel-penalized regression for analysis of microbiome data', The annals of applied statistics **12**(1), 540.

Relman, D. A. (2012), 'The human microbiome: ecosystem resilience and health', Nutrition reviews **70**(suppl_1), S2–S9.

Robinson, M. D., McCarthy, D. J. and Smyth, G. K. (2010), 'edgeR: a Bioconductor package for differential expression analysis of digital gene expression data', Bioinformatics **26**(1), 139–140. PMCID: PMC2796818.

Routy, B., Le Chatelier, E., Derosa, L., Duong, C. P., Alou, M. T., Daillère, R., Fluckiger, A., Messaoudene, M., Rauber, C., Roberti, M. P. et al. (2018), 'Gut microbiome influences efficacy of pd-1–based immunotherapy against epithelial tumors', Science **359**(6371), 91–97.

Rune Halvorsen, Ø. (2003), 'Partitioning the variation in a plot-by-species data matrix that is related to n sets of explanatory variables', Journal of Vegetation Science **14**(5), 693–700.

Salosensaari, A., Laitinen, V., Havulinna, A. S., Meric, G., Cheng, S., Perola, M., Valsta, L., Alfthan, G., Inouye, M., Watrous, J. D. et al. (2021), 'Taxonomic signatures of cause-specific mortality risk in human gut microbiome', Nature communications **12**(1), 1–8.

Sampson, J. N., Boca, S. M., Moore, S. C. and Heller, R. (2018), 'Fwer and fdr control when testing multiple mediators', Bioinformatics **34**(14), 2418–2424.

Sandve, G. K., Ferkingstad, E. and Nygård, S. (2011), 'Sequential monte carlo multiple testing', Bioinformatics **27**(23), 3235–3241.

Satten, G. A., Kong, M. and Datta, S. (2018), 'Multisample adjusted u-statistics that account for confounding covariates', Statistics in Medicine **37**(2), 3357–3372.

Satten, G. A., Tyx, R. E., Rivera, A. J. and Stanfill, S. (2017), 'Restoring the duality between principal components of a distance matrix and linear combinations of predictors, with application to studies of the microbiome', PloS one **12**(1), e0168131. PMCID: PMC5234780.

Schulfer, A. F., Schluter, J., Zhang, Y., Brown, Q., Pathmasiri, W., McRitchie, S., Sumner, S., Li, H., Xavier, J. B. and Blaser, M. J. (2019), 'The impact of early-life sub-therapeutic

antibiotic treatment (stat) on excessive weight is robust despite transfer of intestinal microbes', The ISME journal **13**(5), 1280–1292.

Shade, A., Peter, H., Allison, S. D., Baho, D., Berga, M., Bürgmann, H., Huber, D. H., Langenheder, S., Lennon, J. T., Martiny, J. B., Matulich, K. L., Schmidt, T. M. and Handelsman, J. (2012), 'Fundamentals of microbial community resistance and resilience', Frontiers in microbiology **3**, 417. PMCID: PMC3525951.

Shi, P. and Li, H. (2017), 'A model for paired-multinomial data and its application to analysis of data on a taxonomic tree', Biometrics **73**(4), 1266–1278. PMCID: PMC5623182.

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

Sinha, R., Abu-Ali, G., Vogtmann, E., Fodor, A. A., Ren, B., Amir, A., Schwager, E., Crabtree, J., Ma, S., Abnet, C. C. et al. (2017), 'Assessment of variation in microbial community amplicon sequencing by the microbiome quality control (mbqc) project consortium', Nature biotechnology **35**(11), 1077–1086.

Sohn, M. B. and Li, H. (2019), 'Compositional mediation analysis for microbiome studies', The Annals of Applied Statistics **13**(1), 661–681.

Sohn, M. B., Lu, J. and Li, H. (2021), 'A compositional mediation model for a binary outcome: Application to microbiome studies', Bioinformatics .

Sonnenburg, J. L. and Bäckhed, F. (2016), 'Diet–microbiota interactions as moderators of human metabolism', Nature **535**(7610), 56. PMCID: PMC5991619.

Stewart, C. J., Ajami, N. J., O?Brien, J. L., Hutchinson, D. S., Smith, D. P., Wong, M. C., Ross, M. C., Lloyd, R. E., Doddapaneni, H., Metcalf, G. A. et al. (2018), 'Temporal development of the gut microbiome in early childhood from the teddy study', Nature **562**(7728), 583.

Sullivan, P. F., de Geus, E. J., Willemsen, G., James, M. R., Smit, J. H., Zandbelt, T., Arolt, V., Baune, B. T., Blackwood, D., Cichon, S. et al. (2009), 'Genome-wide association for major depressive disorder: a possible role for the presynaptic protein piccolo', Molecular psychiatry **14**(4), 359–375.

Tang, Z.-Z. and Chen, G. (2019), 'Zero-inflated generalized dirichlet multinomial regression model for microbiome compositional data analysis', Biostatistics **20**(4), 698–713.

Tang, Z.-Z., Chen, G. and Alekseyenko, A. V. (2016), 'Permanova-s: association test for microbial community composition that accommodates confounders and multiple distances', Bioinformatics **32**(17), 2618–2625.

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

Tang, Z.-Z., Chen, G., Alekseyenko, A. V. and Li, H. (2017), 'A general framework for association analysis of microbial communities on a taxonomic tree', Bioinformatics **33**(9), 1278–1285.

Tang, Z.-Z., Chen, G., Hong, Q., Huang, S., Smith, H. M., Shah, R. D., Scholz, M. B. and Ferguson, J. F. (2019), 'Multi-omic analysis of the microbiome and metabolome in healthy subjects reveals microbiome-dependent relationships between diet and metabolites', Frontiers in genetics **10**, 454.

Teo, S. M., Mok, D., Pham, K., Kusel, M., Serralha, M., Troy, N., Holt, B. J., Hales, B. J., Walker, M. L., Hollams, E. et al. (2015), 'The infant nasopharyngeal microbiome impacts severity of lower respiratory infection and risk of asthma development', Cell host & microbe **17**(5), 704–715.

Therneau, T. M., Grambsch, P. M. and Fleming, T. R. (1990), 'Martingale-based residuals for survival models', Biometrika **77**(1), 147–160.

Thorsen, J., Brejnrod, A., Mortensen, M., Rasmussen, M. A., Stokholm, J., Al-Soud, W. A., Sørensen, S., Bisgaard, H. and Waage, J. (2016), 'Large-scale benchmarking reveals false discoveries and count transformation sensitivity in 16S rRNA gene amplicon data analysis methods used in microbiome studies', Microbiome **4**(1), 62. PMCID: PMC5123278.

Tran, T. T., Corsini, S., Kellingray, L., Hegarty, C., Le Gall, G., Narbad, A., Müller, M., Tejera, N., O?toole, P. W., Minihane, A.-M. et al. (2019), 'Apoe genotype influences the gut microbiome structure and function in humans and mice: relevance for alzheimer?s disease pathophysiology', The FASEB Journal pp. fj–201900071R.

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R. and Gordon, J. I. (2007), 'The human microbiome project', Nature **449**(7164), 804–810.

Tyx, R. E., Stanfill, S. B., Keong, L. M., Rivera, A. J., Satten, G. A. and Watson, C. H. (2016), 'Characterization of bacterial communities in selected smokeless tobacco products using 16s rdna analysis', PLoS One **11**(1), e0146939.

van Winkel, R., Rutten, B. P., Peerbooms, O., Peuskens, J., van Os, J. and De Hert, M. (2010), 'MTHFR and risk of metabolic syndrome in patients with schizophrenia', Schizophrenia Research **121**(1), 193–198.

Vandeputte, D., Kathagen, G., D'hoe, K., Vieira-Silva, S., Valles-Colomer, M., Sabino, J., Wang, J., Tito, R. Y., De Commer, L., Darzi, Y. et al. (2017), 'Quantitative microbiome profiling links gut community variation to microbial load', Nature **551**(7681), 507–511.

VanderWeele, T. J. and Shpitser, I. (2011), 'A new criterion for confounder selection', Biometrics **67**(4), 1406–1413.

VanderWeele, T. J. and Vansteelandt, S. (2009), 'Conceptual issues concerning mediation, interventions and composition', Statistics and its Interface **2**(4), 457–468.

VanderWeele, T. and Vansteelandt, S. (2014), 'Mediation analysis with multiple mediators', Epidemiologic methods **2**(1), 95–115. PMCID: PMC4287269.

Vatanen, T., Franzosa, E. A., Schwager, R., Tripathi, S., Arthur, T. D., Vehik, K., Lernmark, Å., Hagopian, W. A., Rewers, M. J., She, J.-X. et al. (2018), 'The human gut microbiome in early-onset type 1 diabetes from the teddy study', Nature **562**(7728), 589.

Verbeek, E. C., Bakker, I. M., Bevova, M. R., Bochdanovits, Z., Rizzu, P., Sondervan, D., Willemsen, G., De Geus, E. J., Smit, J. H., Penninx, B. W., Boomsma, D., Hoogendijk, W. and Heutink, P. (2012), 'A fine-mapping study of 7 top scoring genes from a GWAS for major depressive disorder', PloS One **7**(5), e37384.

Veziant, J., Villéger, R., Barnich, N. and Bonnet, M. (2021), 'Gut microbiota as potential biomarker and/or therapeutic target to improve the management of cancer: focus on colibactin-producing escherichia coli in colorectal cancer', Cancers **13**(9), 2215.

Vogt, N. M., Kerby, R. L., Dill-McFarland, K. A., Harding, S. J., Merluzzi, A. P., Johnson, S. C., Carlsson, C. M., Asthana, S., Zetterberg, H., Blennow, K. et al. (2017), 'Gut microbiome alterations in alzheimer's disease', Scientific reports **7**(1), 13537.

Wang, C., Hu, J., Blaser, M. J. and Li, H. (2019), 'Estimating and testing the microbial causal mediation effect with high-dimensional and compositional microbiome data', Bioinformatics p. doi: 10.1093/bioinformatics/btz565.

Wang, K. (2014), 'Testing genetic association by regressing genotype over multiple phenotypes', PloS One **9**(9).

Weiss, S., Xu, Z. Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., Lozupone, C., Zaneveld, J. R., Vázquez-Baeza, Y., Birmingham, A., Hyde, E. R. and Knight, R. (2017), 'Normalization and microbial differential abundance strategies depend upon data characteristics', Microbiome **5**(1), 27. PMCID: PMC5335496.

Werft, W. and Benner, A. (2010), 'glmperm: A permutation of regressor residuals test for inference in generalized linear models', The R Journal **2**(1), 39–43.

Westfall, P. H. and Young, S. S. (1993), Resampling-based multiple testing: Examples and methods for p-value adjustment, John Wiley & Sons.

Wijesinha, A., Begg, C. B., Funkenstein, H. H. and McNeil, B. J. (1983), 'Methodology for the differential diagnosis of a complex data set: a case study using data from routine ct scan examinations', Medical Decision Making **3**(2), 133–154.

Wilson, D. J. (2019a), 'The harmonic mean p-value for combining dependent tests', Proceedings of the National Academy of Sciences **116**(4), 1195–1200.

Wilson, D. J. (2019b), 'The harmonic mean p-value for combining dependent tests', Proceedings of the National Academy of Sciences **116**(4), 1195–1200.

Winkler, A. M., Ridgway, G. R., Webster, M. A., Smith, S. M. and Nichols, T. E. (2014), 'Permutation inference for the general linear model', Neuroimage **92**, 381–397.

Wirth, R., Maróti, G., Mihók, R., Simon-Fiala, D., Antal, M., Pap, B., Demcsák, A., Minarovits, J. and Kovács, K. L. (2020), 'A case study of salivary microbiome in smokers and non-smokers in hungary: analysis by shotgun metagenome sequencing', Journal of Oral Microbiology **12**(1), 1773067.

Witkin, S. and Linhares, I. (2017), 'Why do lactobacilli dominate the human vaginal microbiota?', BJOG: An International Journal of Obstetrics & Gynaecology **124**(4), 606–611.

Wu, B. and Pankow, J. S. (2015), 'Statistical methods for association tests of multiple continuous traits in genome-wide association studies', Annals of Human Genetics **79**(4), 282–293.

Wu, C., Chen, J., Kim, J. and Pan, W. (2016), 'An adaptive association test for microbiome data', Genome medicine **8**(1), 56. PMCID: PMC4872356.

Wu, J., Peters, B. A., Dominianni, C., Zhang, Y., Pei, Z., Yang, L., Ma, Y., Purdue, M. P., Jacobs, E. J., Gapstur, S. M. et al. (2016), 'Cigarette smoking and the oral microbiome in a large study of american adults', The ISME journal **10**(10), 2435–2446.

Yue, Y. and Hu, Y. (2021), 'A new approach to testing mediation of the microbiome using the ldm', bioRxiv p. https://doi.org/10.1101/2021.11.12.468449.

Zhai, J., Knox, K., Twigg III, H. L., Zhou, H. and Zhou, J. J. (2019), 'Exact variance component tests for longitudinal microbiome studies', Genetic epidemiology **43**(3), 250–262. PMCID: PMC6416054.

Zhan, X., Tong, X., Zhao, N., Maity, A., Wu, M. C. and Chen, J. (2017), 'A small-sample multivariate kernel machine test for microbiome association studies', Genetic epidemiology **41**(3), 210–220.

Zhang, H., Chen, J., Li, Z. and Liu, L. (2019), 'Testing for mediation effect with application to human microbiome data', Statistics in Biosciences pp. 1–16.

Zhang, J., Wei, Z. and Chen, J. (2018), 'A distance-based approach for testing the mediation effect of the human microbiome', Bioinformatics **34**(11), 1875–1883.

Zhang, X., Pei, Y.-F., Zhang, L., Guo, B., Pendegraft, A., Zhuang, W. and Yi, N. (2018), 'Negative binomial mixed models for analyzing longitudinal microbiome data', Frontiers in microbiology **9**, 1683. PMCID: PMC6070621.

Zhang, Y., Han, S. W., Cox, L. M. and Li, H. (2017), 'A multivariate distance-based analytic framework for microbial interdependence association test in longitudinal study', Genetic epidemiology **41**(8), 769–778.

Zhao, N., Chen, J., Carroll, I. M., Ringel-Kulka, T., Epstein, M. P., Zhou, H., Zhou, J. J., Ringel, Y., Li, H. and Wu, M. C. (2015), 'Testing in microbiome-profiling studies with

mirkat, the microbiome regression-based kernel association test', The American Journal of Human Genetics **96**(5), 797–807.

Zhao, N. and Satten, G. A. (2021a), A log-linear model for inference on bias in microbiome studies, in 'Statistical Analysis of Microbiome Data', Springer, pp. 221–246.

Zhao, N. and Satten, G. A. (2021b), A log-linear model for inference on bias in microbiome studies, in S. Datta and S. Guha, eds, 'Statistical Analysis of Microbiome Data', Springer-Verlag, New York, chapter 9, pp. 221 – 247.

Zhao, N., Zhan, X., Guthrie, K. A., Mitchell, C. M. and Larson, J. (2018), 'Generalized Hotelling's test for paired compositional data with application to human microbiome studies', Genetic epidemiology **42**(5), 459–469.

Zhou, C., Wang, H., Zhao, H. and Wang, T. (2022), 'fastancom: a fast method for analysis of compositions of microbiomes', Bioinformatics **38**(7), 2039–2041.

Zhou, H., He, K., Chen, J. and Zhang, X. (2022), 'Linda: linear models for differential abundance analysis of microbiome compositional data', Genome biology **23**(1), 1–23.

Zhu, Z., Satten, G. A. and Hu, Y.-J. (2022), 'Integrative analysis of relative abundance data and presence–absence data of the microbiome using the ldm', Bioinformatics **38**(10), 2915–2917.

Zhu, Z., Satten, G. A., Mitchell, C. and Hu, Y.-J. (2021), 'Constraining permanova and ldm to within-set comparisons by projection improves the efficiency of analyses of matched sets of microbiome data', Microbiome **9**(1), 1–19.