Distribution Agreement

Signature:

____________________________ _______________

Ye Yue Date

New Statistical Methods for Analyzing Microbiome Data

By

Ye Yue

Doctor of Philosophy

Biostatistics

Yijuan Hu, Ph.D.
Advisor

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

Glen A. Satten, Ph.D.
Committee Member

Jingjing Yang, Ph.D.
Committee Member

Accepted:

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

Date

New Statistical Methods for Analyzing Microbiome Data

By

Ye Yue
M.A., Columbia University, 2019

Advisor: Yijuan Hu, 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
2023

Abstract

New Statistical Methods for Analyzing Microbiome Data

By

Ye Yue

Microbiome research has proliferated due to booming interests in the scientific community, increasing power of high-throughput sequencing, and rapid advancement of data analytics. The analysis for microbiome data from sequencing studies is challenging because of high-dimensionality, overdispersion, sparsity, compositionality, and experimental bias. In addition, microbiome studies typically have small sample, complex traits of interest and confounding covariates. New methods that can fully account for the complexities of data are needed.

In the first topic, we develop a new statistical method for testing mediation effects of microbiome at both the community and individual taxon levels. We have seen a rapidly growing volume of evidence linking the microbiome and human diseases or clinical outcomes, as well as evidence linking the microbiome and environmental exposures. Understanding whether and which microbes played a mediating role between an exposure and a disease outcome are essential for researchers to develop clinical interventions by modulating the microbes. Our new method allows an arbitrary number of taxa to be tested simultaneously, supports different types of exposures and outcomes, and so on.

In the second topic, we extend the most commonly used distance-based method PERMANOVA to testing microbiome mediation effects at the community level. Use of distance matrices is a popular approach to analyzing complex microbiome data. Our extension allows adjustment of confounders, accommodates various types of exposures and outcomes, and provides an omnibus test that combines the results from analyzing multiple distance matrices.

In the third topic, we develop a novel method for integrative analysis of datasets generated by both 16 marker-gene sequencing and shotgun metagenomics sequencing. Many microbiome studies have performed both experiments on the same cohort of samples. The two datasets often yield consistent patterns; however, each is subject to distinct experimental biases in an experiment-specific manner. These experimental biases, together with partially overlapping samples and differential library sizes between the two datasets, pose tremendous challenges when combining the datasets. Our new method combines data from both experiments for differential abundance tests, while accounting for differential experimental biases, assigning adaptive weights to each observation, and accommodating samples and taxa unique to an experiment.

New Statistical Methods for Analyzing Microbiome Data

By

Ye Yue
MA, Columbia University, 2019

Advisor: Yijuan Hu, Ph.D.

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
2023

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction to high-throughput microbiome data

Thanks to technological advances in high-throughput sequencing, microbiome research has proliferated in the past decade and revealed differences in human microbiome are associated with many diseases and conditions such as inflammatory bowl diseases (Gevers et al., 2014), obesity and type II diabetes (Hartstra et al., 2015), and even cancers (Marchesi et al., 2011) as well as with environmental exposures such as diet (McDonald et al., 2018). Nowadays, the microbiome becomes a particularly attractive target for establishing new biomarkers for disease diagnosis and prognosis, and for developing low-cost, low-risk interventions.

Data in microbiome research is collected by two commonly used sequencing approaches which are marker-gene sequencing and metagenome sequencing (Weinstock, 2012). Marker-gene sequencing targets and amplifies portions of the hypervariable regions of a specific gene, such as the 16S ribosomal RNA (rRNA) gene, while metagenome sequencing sequences all of the microbial genes from a sample. Microbiome data from sequencing studies are processed into a taxa count table. The taxa count data are high-dimensional with typically many more taxa than samples. The data are also sparse (having 50-90% zero counts), compositional (measuring relative abundances that sum to one), and highly overdispersed. In addition, microbiome studies may have complex exposures or outcomes that can be either continuous or discrete, as well as multivariate (comprising multiple components such as categorical variables with more than two levels); the outcome can even be censored survival times (Spencer et al., 2021; Jenq et al., 2015). These studies often have potential confounding covariates, small sample sizes (e.g., 30–100) and complex designs (e.g., clustered data (Hu and Satten, 2020), matched sets (Zhu et al., 2021), longitudinal sampling (Hu, Li, Satten and Hu, 2022)). The capability to handle all these features is essential for any statistical method to be practically useful.

Microbiome data analysis methods appear to be categorized into two groups. One group tests the association between the variables of interest and the overall microbial compositions, such as PERMANOVA (McArdle and Anderson, 2001), MiRKAT (Zhao et al., 2015), aMiSPU (Wu et al., 2016), pairNM (Shi and Li, 2017), Linear Decomposition Model (LDM) (Hu and Satten, 2020) and logistic compositional analysis (LOCOM) (Hu, Satten and Hu, 2022), which is called community-level testing. The other one is taxon-level testing, which targets on the effect or contributions of the individual taxon. In microbiome research, there exist at least two biological models that explain how microbial communities can change when comparing groups with different phenotypes or along a phenotypic gradient. In one model, a significant portion of the taxa within the community undergoes changes. The null hypothesis, tested at the taxon level, assumes no differential abundance and states that the relative abundance of a taxon remains constant. Therefore, any observed changes in the relative abundance of a taxon across conditions are of particular interest in this model. Various methods can be utilized to test this hypothesis, including LDM (Hu and Satten, 2020) and methods developed for RNA-Seq data , such as DESeq2 (Love et al., 2014), edgeR (Robinson et al., 2010) and metagenomeSeq (Paulson et al., 2013), as well as the direct application of nonparametric tests (such as the Wilcoxon rank-sum test) to relative abundance data or rarefied count data. In the alternative model, it is posited that only a small subset of key taxa undergo changes, while the remaining taxa exhibit variations in relative abundance due to compositional constraints. In this case, the null hypothesis tested at the taxon level posits that the ratio of relative abundances at a specific taxon, compared to a chosen null taxon, remains unchanged. Assessing whether the ratio between the relative abundances of a taxon and the null taxon alters provides crucial insights within this model. There are multiple methods available for testing this hypothesis, including the analysis of composition of microbiomes ANCOM (Mandal et al. (2015), Kaul et al. (2017)), ANOVA-Like Differential Expression tool (ALDEx2) (Fernandes et al., 2014), WRENCH (Kumar et al., 2018), Testing for Differential Abundance in Compositional Counts Data (DACOMP) (Brill et al., 2020) and LOCOM (Hu, Satten and Hu, 2022). Notably, the second model accounts for the compositional constraint where a change in the relative abundance of one taxon implies a compensating change in other taxa. Consequently, this model is commonly referred to as compositional analysis (Gloor et al., 2017).

1.2 Mediation analysis of microbiome data

While most microbiome studies conducted so far have focused on bivariate associations between the microbiome and the covariates of interest (e.g., environmental factors, clinical outcomes) (Bai et al., 2019; Dunlop et al., 2021), increasing studies have emerged recently to elucidate the biological mechanisms underlying the complex interplay between environmental exposures, the microbiome, and clinical outcomes. In many cases, it is of interest to understand whether the microbiome play a mediating role between an exposure and an outcome (Pope et al., 2017; Dolan and Chang, 2017; Wang et al., 2020), as depicted in Figure 1.1(a). For example, does diet have any effect on inflammatory bowel diseases that is mediated through the perturbation of the gut microbiome (Dolan and Chang, 2017)? How does the change in the gut microbiome due to antibiotic exposure cause the change in mouse body weight (Wang et al., 2020)?

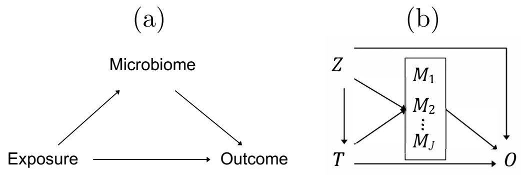


Figure 1.1: (a) Multiple microbes mediate the effect of the exposure on the outcome. (b) T denotes the exposure, $(M_1, \dots, M_J)$ the microbes, O the outcome, and Z the confounders.

Compared to the test of bivariate associations, one challenge in the test of mediation is the composite null hypothesis. Let T denote the exposure (or treatment), $M = (M_1, \dots, M_J)$ the J mediators, O the outcome, and Z the confounding covariates; using this notation, the mediation relationships are shown in Figure 1.1(b). To claim a mediation effect of a microbe, both the exposure-microbe and microbe-outcome associations (given the exposure) are required to be significant. Thus, the null hypothesis of no mediation at microbe j is a composite null that consists of no microbe-outcome association, no exposure-microbe association, or neither:

$$T \to M_j \not\to O,\ T \not\to M_j \to O,\ \text{or}\ T \not\to M_j \not\to O,$$

which are referred to as the type-I, type-II, and type-III null hypotheses, respectively. It is highly likely that different microbes are under different types of null. For example, antibiotic use may perturb a large number of microbes but most of them do no modify mouse body weight, whereas some microbes remain intact from antibiotic use but do interact with the body weight; of course, there are microbes that are not associated with either factor. In this example, we have all three types of null, and a valid analysis should acknowledge that.

1.2.1 Mediation analysis at both the taxon and community levels

In addition to the community-level mediation effect of the microbiome, it is of particular importance to identify the specific microbes that are responsible for the overall mediation effect, which is essential for researchers to develop clinical interventions to modify the outcome by modulating the mediating microbes, e.g., through antibiotics or probiotics that directly modify the number of the microbes, or prebiotics that modify microbial products such as metabolites (Berg et al., 2020; Quigley and Gajula, 2020). The two existing methods, MedTest and MODIMA , are restricted to testing the overall mediation effect at the community level. Although other methods, namely CMM (Sohn and Li, 2019) (and CMMB, the extension for binary outcomes (Sohn et al., 2022)), SparseMCMM (Wang et al., 2020), and Zhang's method (Zhang et al., 2021) attempt to identify individual mediating taxa, they have no control of any error rate (e.g., the FDR).

LDM was originally developed to test associations between the microbiome and the covariates of interest, providing a unified framework that allows for both community-level and taxon-level testing. It is based on a linear model that regresses the microbiome data at each taxon on the sequentially orthogonalized covariates that include first the confounding covariates that we wish to adjust for and then the covariates that we wish to test. Specifically,

$$Y = X\beta + \epsilon,$$

where X is the design matrix of all covariates and the columns of X is grouped into K submodels, i.e., $X = (X_1, X_2, \ldots, X_K)$. Each submodel includes components that will be tested jointly, such as a single covariate, multiple covariates, or multiple indicators for a categorical covariate. The submodels are first processed into sequentially orthogonal, unit vectors by the Gram-Schmidt process, so that the partition of the distance matrix is unambiguous. This requires that the covariates in X follow a scientifically meaningful order; for example, the confounders should enter first. β is an $r \times J$ matrix that should be estimated, and ϵ is an $r \times J$ matrix of error terms with $E(\epsilon|X) = 0$. For J models considered for the columns of Y, the j^{th} column of β represents the regression coefficients specific to the j^{th} regression model. It allows an arbitrary number of taxa (including arbitrarily rare taxa) to be tested simultaneously. The covariates can be continuous, discrete, or multivariate variables, or even censored survival times; note that the survival times and censoring statuses are first fit by a Cox model to be converted to the Martingale or deviance residuals, which are then used as a generic continuous covariate in the LDM (Hu, Li, Satten and Hu, 2022). The taxon data can be at the relative abundance scale, arcsin-root-transformed relative abundance scale, or the presence-absence scale (Hu and Satten, 2021), and their results can be combined to provide omnibus tests (Zhu, Satten and Hu, 2022). The inference of associations in the LDM is based on permutation (i.e., permuting the orthogonalized covariates) to circumvent making parametric assumptions about the distribution of the microbiome data. Thus, the inference is robust to sparse and overdispersed count data, as well as small sample sizes, and the LDM always has good control of the FDR. Also, the permutation can be conducted to preserve the sample structure (e.g., clustered data (Hu and Satten, 2020), matched sets (Zhu et al., 2021)), so the LDM can accommodate certain complex designs. The covariate types, taxon data scales, and sample structures that the LDM supports were summarized in Figure 1 of (Zhu, Satten and Hu, 2022).

1.2.2 Mediation analysis based on distance matrices

To circumvent the complexities of microbiome count data, a popular approach is to first summarize the taxon-level data into a $n \times n$ distance (dissimilarity) matrix D calculated among n samples that measures the pairwise dissimilarity in the microbiome profiles, and then base the analysis of microbiome data on the distance matrix (Legendre and Anderson, 1999; McArdle and Anderson, 2001; Zhao et al., 2015; Alekseyenko, 2016; Zhang et al., 2017). This approach provides results at the community-level, which is usually the first step in an analytical pipeline. Numerous distance measures, with different properties, have been proposed to detect diverse patterns in microbiome data; the most commonly used ones include Jaccard (Jaccard, 1912), Bray-Curtis (Bray and Curtis, 1957), and weighted or unweighted UniFrac (Lozupone and Knight, 2005; Chen et al., 2012). It is well acknowledged that the optimal choice of a distance measure depends on the underlying variation pattern in a particular dataset, which is unknown a priori. Therefore, it is a common practice to construct an omnibus test that combines the results from analyzing different distance matrices.

Two existing methods, MedTest (Zhang et al., 2018) and MODIMA (Hamidi et al., 2019), adopted such a distance-based approach to mediation analysis of microbiome data. Specifically, MedTest uses the principal components (PCs) of a given distance matrix as multiple mediators and tests their joint mediation effects. MedTest considers microbiome "features" to be the eigenvectors of the Gower-centered distance matrix Δ, denoted by $u_1, u_2, \ldots, u_L$, that are associated with the L positive eigenvalues, denoted by $\lambda_1, \lambda_2, \ldots, \lambda_L$ and it assumes that these microbiome features are the units through which the microbiome exert the mediation effect. Thus, MedTest adopts a test statistic that is a sum of feature-specific mediation effects, each weighted by λ_l (the percentage of variance explained by that feature):

$$\mathbb{U}_{\text{MedTest}} = \sum_{l=1}^{L} \lambda_l |u_l^{\mathrm{T}} T_r||u_l^{\mathrm{T}} O_r|,$$

where $|.|$ is the absolute value function, T_r is the residual of T after orthogonalizing against Z and O_r is the residual of O after orthogonalizing against (Z, T). Note that $u_l^{\mathrm{T}} T_r$ and $u_l^{\mathrm{T}} O_r$ are the sample Pearson correlation coefficients that measure the associations between the lth feature and the exposure and the outcome, respectively; the sample Pearson correlation coefficient does not easily accommodate multivariate exposure or outcome variables. Because the null hypothesis of no mediation is a composite null that consists of three types of null hypothesis. MedTest calculates the maximum of the statistics corresponding to the three types of null hypotheses for the bth permutation replicate:

$$\mathbb{U}_{\text{MedTest}}^{(b)} = \max\Big\{\sum_{l=1}^{L}\lambda_l|u_l^{\mathrm{T}}T_r^{(b)}||u_l^{\mathrm{T}}O_r|,\ \ \sum_{l=1}^{L}\lambda_l|u_l^{\mathrm{T}}T_r||u_l^{\mathrm{T}}O_r^{(b)}|,\ \ \sum_{l=1}^{L}\lambda_l|u_l^{\mathrm{T}}T_r^{(b)}||u_l^{\mathrm{T}}O_r^{(b)}|\Big\},$$

where $T_r^{(b)}$ and $O_r^{(b)}$ are permuted vectors of T_r and O_r, respectively. Finally, the p-value is obtained as the proportion of $\mathbb{U}_{\text{MedTest}}^{(b)}$ that are equal to or larger than the observed statistic $\mathbb{U}_{\text{MedTest}}$. The power of MedTest may critically depend on whether the exposure-microbiome association and the microbiome-outcome association coincide at the same set of PCs. Further, when the true mediators in the community are rare taxa, the PCs may not effectively capture the variation at these mediators. However, the assumption that the exposure-microbiome association and the microbiome-outcome association coincide at the same set of PCs may be overly optimistic. Also, the PCs may not capture mediation effects at rare taxa. Moreover, MedTest does not accommodate multivariate exposures and outcomes in its current form.

In addition to the distance matrix D from the microbiome profiles, MODIMA also requires the $n \times n$ distance matrices (usually the Euclidean distance) being calculated from the exposure data and the outcome data, separately, which we denote by D_T and D_O. These distance matrices naturally accommodate multivariate variables. Then, MODIMA uses the distance correlation (Székely and Rizzo, 2009), $\text{dCor}(D_T, D)$, for measuring the exposure-microbiome association, which parallels the Pearson correlation with the major difference being that the centered product moment transformation is applied to the distance matrices rather than data vectors. MODIMA uses the partial distance correlation (Székely and Rizzo, 2014), $\text{pdCor}(D_O, D|D_T)$, for measuring the microbiome-outcome association conditional on the exposure, which parallels the Pearson partial correlation. MODIMA adopts the test statistic

$$\mathbb{U}_{\text{MODIMA}} = \text{dCor}(D_T, D) \times \text{pdCor}(D_O, D|D_T),$$

and the statistic for the bth permutation replicate,

$$\mathbb{U}_{\text{MODIMA}}^{(b)} = \begin{cases} \text{dCor}(D_T^{(b)}, D) \times \text{pdCor}(D_O, D|D_T), & \text{if } \text{dCor}(D_T, D) \le \text{pdCor}(D_O, D|D_T) \\ \text{dCor}(D_T, D) \times \text{pdCor}(D_O^{(b)}, D|D_T), & \text{if } \text{dCor}(D_T, D) > \text{pdCor}(D_O, D|D_T), \end{cases}$$

where $D_T^{(b)}$ and $D_O^{(b)}$ are obtained by permuting both rows and columns of the D_T and D_O matrices, respectively. Finally, the p-value is calculated as the proportion of $\mathbb{U}_{\text{MODIMA}}^{(b)}$ that are equal to or larger than the observed statistic $\mathbb{U}_{\text{MODIMA}}$. Note that, in this process, the confounding covariate Z cannot be adjusted and MODIMA does not provide an omnibus test. Further, the MODIMA paper pointed out a lack of correspondence between conditional independence and zero partial distance correlation, e.g., a non-zero partial correlation in scenarios with conditionally independent variables. It implies that MODIMA may generate false positive findings under the null hypothesis of no mediation. Finally, neither MedTest nor MODIMA can handle censored survival times.

PERMANOVA (McArdle and Anderson, 2001) is currently the most commonly used distance-based method in analysis of microbiome data. PERMANOVA is based on a linear model of covariates that partition a given distance matrix along each covariate. In particular, when the Euclidean distance measure is used on the relative abundance data, it is the total variance of relative abundance data across all taxa that is partitioned into variance explained by each covariate. Using the same notation as in the LDM, we denote the design matrix of all covariates as $X = (X_1, X_2, \ldots, X_K)$, where the columns of X are grouped into K submodels. The $n \times n$ distance matrix D is often Gower-centered (Gower, 1966) to become $\Delta = -0.5 \left(I - n^{-1}11'\right) D^2 \left(I - n^{-1}11'\right)$, where D^2 is the element-wise squared D, I is the identity matrix, and 1 is a vector of n ones. The "residual" distance matrix after projecting off all submodels except the kth one takes the form $\widetilde{\Delta}_k = \Big(I - \sum_{k'=1,\ldots,K,k'\neq k} X_{k'}X_{k'}^{\mathrm{T}}\Big)\Delta\Big(I - \sum_{k'=1,\ldots,K,k'\neq k} X_{k'}X_{k'}^{\mathrm{T}}\Big)$ by noting that $X_kX_k^{\mathrm{T}}$ is the hat matrix for the kth submodel. Then, PERMANOVA tests the effect of the kth submodel by using the F-statistic

$$F_k \propto \frac{\operatorname{Tr}\left[X_kX_k^{\mathrm{T}}\widetilde{\Delta}_kX_kX_k^{\mathrm{T}}\right]}{\operatorname{Tr}\left[\left(I - \sum_{k'=1}^{K} X_{k'}X_{k'}^{\mathrm{T}}\right)\widetilde{\Delta}_k\left(I - \sum_{k'=1}^{K} X_{k'}X_{k'}^{\mathrm{T}}\right)\right]}, \tag{1.1}$$

where $\operatorname{Tr}(\cdot)$ is the trace operation. PERMANOVA assesses the significance of the F-statistic via permutation, particularly the Freedman-Lane permutation scheme (Freedman and Lane, 1983) as implemented in "permanovaFL". The Supplementary Materials of (Hu and Satten, 2020) showed that the Freedman-Lane scheme is equivalent to forming the following statistic for the bth permutation replicate:

$$F_k^{(b)} \propto \frac{\operatorname{Tr}\left[X_k^{(b)}X_k^{(b)^{\mathrm{T}}}\widetilde{\Delta}_kX_k^{(b)}X_k^{(b)^{\mathrm{T}}}\right]}{\operatorname{Tr}\left[\left(I - \sum_{k'=1}^{K} X_{k'}^{(b)}X_{k'}^{(b)^{\mathrm{T}}}\right)\widetilde{\Delta}_k\left(I - \sum_{k'=1}^{K} X_{k'}^{(b)}X_{k'}^{(b)^{\mathrm{T}}}\right)\right]}, \tag{1.2}$$

where $X_k^{(b)}$ is a row-permuted version of X_k and thus the columns of $X_k^{(b)}$ remain orthogonal. Note that the residual distance matrices $\widetilde{\Delta}_k$s do not need to be recalculated for each replicate. In contrast, the permutation scheme implemented in adonis2 replaces all $\widetilde{\Delta}_k$s in F_k and $F_k^{(b)}$ by the raw distance matrix Δ.

PERMANOVA is very versatile. It can handle censored survival times. As proposed in (Hu, Li, Satten and Hu, 2022), the survival times and censoring statuses are first fit by a Cox model (including non-microbiome risk predictors as covariates) to be converted into the Martingale or deviance residuals, which are then used as a generic continuous covariate in PERMANOVA. Because PERMANOVA bases its inference on permutation, it is robust to small sample sizes. The permutation replicates can also be readily used to construct an omnibus test of multiple distance matrices, which uses the minimum of the p-values obtained from analyzing each distance matrix as the final test statistic and uses the corresponding minima from the permutation replicates to simulate the null distribution. In addition, the permutation can be conducted in ways that preserve the correlation found in the original data, so PERMANOVA can accommodate certain structures of samples such as clustered samples (Hu and Satten, 2020) and matched sets (Zhu et al., 2021).

Although it was originally developed for testing microbiome associations, we find that we can extend PERMANOVA to testing microbiome mediation effects by using the idea of inverse regression and including both the exposure and the outcome as covariates whose F-statistics capture the exposure-microbiome association and the microbiome-outcome association conditional on the exposure, respectively. This extension of PERMANOVA would naturally inherit all the features of PERMANOVA, some of which have been a focus of recent development, including adjustment of confounders (Hu and Satten, 2020), test of multivariate covariates, test of censored survival times (Hu, Li, Satten and Hu, 2022), and an omnibus test of multiple distance matrices (Tang et al., 2016). Thus, the extension of PERMANOVA would be very appealing to researchers who routinely use PERMANOVA.

In Topic 1, we focus on testing, rather than estimation, of mediation effects at individual taxa with a goal of controlling the FDR. This strategy is very common in the initial *scan* of high-dimensional features in omic studies (Asher et al., 2009; Hu and Lin, 2010; Hu et al., 2015); “fine mapping” of mechanistic mediators and formal estimation of their mediation effects can be performed more easily after the dimension is greatly reduced. We find that, the testing objective can be facilitated by using inverse regression that regresses the microbiome data at each taxon on the exposure and the exposure-adjusted outcome. We implement the inverse regression model using the LDM framework (Hu and Satten, 2020; Zhu et al., 2021; Hu and Satten, 2021) that we developed originally for testing microbiome associations. As the LDM was designed to specifically handle the microbiome data complexities (e.g., high-dimensionality, sparsity, and overdispersion), our LDM-based mediation analysis naturally inherits these features. As the LDM models each taxon separately, our approach allows different taxa to be under different types of null. Finally, like MedTest and MODIMA, we also develop a global test of community-level mediation; our global test statistic is a coherent combination of our taxon-specific statistics. The main advantage of our approach is that results for individual taxa are available; neither MedTest and MODIMA provide taxon-specific results. In the Method section, we first give the motivation for using inverse regression. Then, we consider four ways of testing individual taxa for mediation and then a method that aggregates the taxon-level information to test the overall mediation in a community. In the Numerical Studies section, we first present simulation results in which we numerically compared the four ways of testing individual taxa and selected the one with the best performance, and we compared our global test to existing tests. Then, we present the application to a real study on murine microbiome. We conclude with a Remarks section.

In Topic 2, we present PERMANOVA-med, the extension of PERMANOVA to testing the community-level mediation effect of the microbiome. We base PERMANOVA-med on our implementation of PERMANOVA through the function “permanovaFL” in our R package LDM (Hu and Satten, 2020), which differs from the “adonis2” implementation in the R package vegan in the permutation scheme and outperformed adonis2 in many situations (Hu and Satten, 2020; Zhu et al., 2021; Hu and Satten, 2021). All the features that PERMANOVA supports were summarized in Figure 1.2. In the Method section, we first motivate the use of inverse regression and then show how to extend PERMANOVA to PERMANOVA-med. In the Numerical Studies section, we present extensive simulation studies in which we numerically compared PERMANOVA-med to MedTest and MODIMA. We also demonstrate the wide applicability of PERMANOVA-med through 16 different mediation analyses of the real data on melanoma immunotherapy response. We conclude with a Remarks section.

1.3 Integrative analysis of 16S marker-gene and shotgun metagenomic sequencing data

The most widely used technologies for profiling microbial communities are 16S marker-gene sequencing (16S) and shotgun metagenomic sequencing (SMS) (Knight et al., 2018). The 16S method employs primers that target a highly variable region of the 16S ribosomal RNA gene, which is then

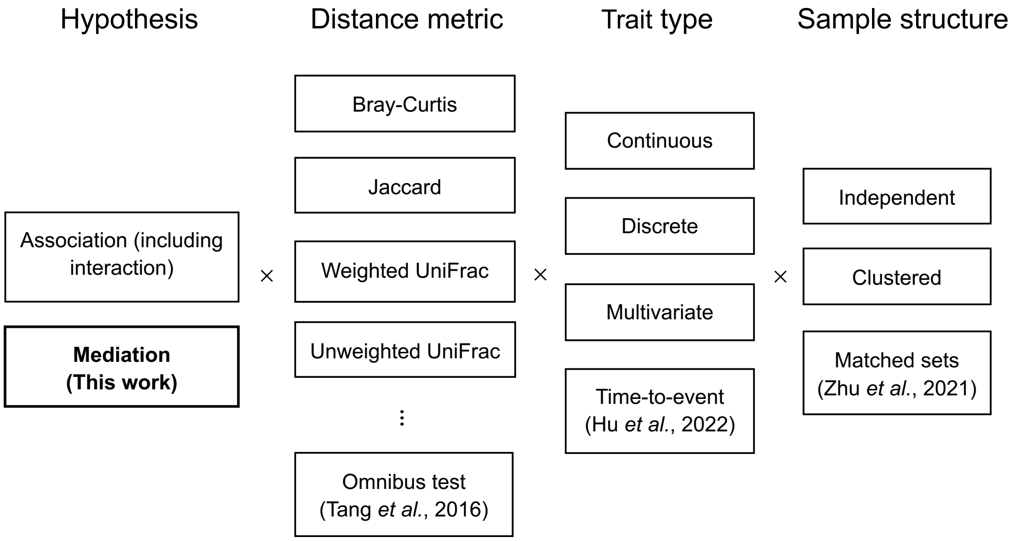


Figure 1.2: Analyses supported by permanovaFL. Analysis types without a citation were introduced in the original LDM paper Hu and Satten (2020). “Clustered” refers to analyses of clustered data where traits of interest vary by cluster or vary both by and within clusters (some analyses may require special structure or additional assumptions). “Matched sets” is a special type of clustered data in which all traits of interest vary within sets.

PCR amplified, sequenced, and classified. This approach is well-tested, fast and cost effective, and provides a low-resolution view of a microbial community, typically at the genus level. On the other hand, SMS extracts all microbial genomes within a sample, which are then fragmented, sequenced, and assembled. This technique offers detailed genomic information, including higher taxonomic resolution and additional functional capability, but it is 10 to 30 times more expensive to prepare and sequence samples and much more challenging to conduct bioinformatics analysis.

Both 16S and SMS introduce *experimental bias* at every step of the experiment (i.e., DNA extraction, PCR amplification, amplicon or metagenomic sequencing, and bioinformatics processing), as each step preferentially measures (i.e., extracts, amplifies, sequences, and bioinformatically identifies) certain taxa over others (McLaren et al., 2019). This bias systematically distorts the measurements (e.g., taxon relative abundances) from their actual values. Moreover, the bias differs significantly between 16S and SMS, as they comprise different steps (e.g., PCR amplification vs. no PCR) and even for the same step they adopt different protocols (e.g., bioinformatics pipelines) (Nearing et al., 2021). As a result, each experiment leads to some taxa being underrepresented or even entirely missed (Peterson et al., 2021). The complementary nature of these taxa suggests that combing both 16S and SMS methods could enhance the profiling of complex microbial communities.

As it turns out, many microbiome studies have performed both 16S and SMS on the same cohort of samples. In Qiita, currently the largest open-source microbial study management platform (https://qiita.ucsd.edu) (Gonzalez et al., 2018), 26 out of a total of 660 studies (as of June 2022 [update]) have both datasets. In fact, there should be more studies with both datasets, but they may have only deposited the one used in the final publication. There are at least two scenarios in which this occurs. In one scenario, a study initially performs 16S and later expands its aims to investigate higher-resolution taxa or biological functions by performing SMS on all or a subset of the samples. In another scenario, a study initially plans to perform SMS but adds 16S due to its low cost.

The 16S data are routinely summarized into a taxa count table by the popular analysis platform QIIME2 (Bokulich et al., 2018). Although there is less consensus, the SMS data can also be classified into taxonomies using tools such as MetaPhlAn (Segata et al., 2012; Truong et al., 2015; Beghini et al., 2021), Kraken (Wood and Salzberg, 2014; Wood et al., 2019), or more recently Woltka (Zhu, Huang, Gonzalez, McGrath, McDonald, Haiminen, Armstrong, Vázquez-Baeza, Yu, Kuczynski et al., 2022). Aside from the experiment-specific differences mentioned earlier, a large number of studies (Clooney et al., 2016; Hillmann et al., 2018; Mas-Lloret et al., 2020; Peterson et al., 2021; Durazzi et al., 2021; Biegert et al., 2021; Zuo et al., 2022; de Vries et al., 2023) have found that taxonomic profiles generated from 16S and SMS yield consistent patterns of microbiome signatures. Thus, it is expected that an integrative analysis of both taxonomic profiles could enhance the power in testing such patterns of microbiome signatures, particularly when assessing differential abundance at a given taxon level (e.g., genus) and the community level against sample-level covariates (e.g., environmental factors or clinical outcomes). To the best of our knowledge, there is currently no statistical method available for performing such an integrative analysis.

LOCOM, a recent developed model by Hu et al. (Hu, Satten and Hu, 2022), is specifically designed for microbiome data analysis from a single experiment. It enables compositional analysis of differential abundance at both the taxon level and the global level, effectively addressing the influence of experimental bias. Notably, LOCOM eliminates the need for pseudocounts which can affect the conclusions of a compositional analysis (Costea et al. (2014), Paulson et al. (2014)) and does not rely on the assumption of a null reference taxon. The commonly used compositional analysis methods, such as ANCOM and ANCOM-BC (Mandal et al. (2015), Kaul et al. (2017)), add pseudocounts to address zero values in count data and are primarily designed for group comparisons. However, these methods lack the capability to accommodate continuous traits of interest in the analysis. The ALDEx2 method (Fernandes et al., 2014) utilizes a sampling process to sample nonzero relative abundances that introduces noise to the data, potentially leading to a loss of statistical power. Furthermore, ALDEx2 relies on the clr transformation to identify differentially abundant taxa compared to the overall mean of all taxa, making it susceptible to the influence of outliers. Both WRENCH (Kumar et al., 2018) and DACOMP (Brill et al., 2020) introduce the requirement of choosing a set of null reference taxa for read count data normalization. Nevertheless, the selection of this reference set poses a risk of including causal taxa, potentially impacting the performance. Additionally, WRENCH is constrained to group comparisons and lacks the ability to account for confounding covariates.

LOCOM employs logistic regression for testing differential abundance of taxa and the whole community. It starts with model

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

where p_{ij} is the expected value of the observed relative abundance for taxon j $(j = 1, \ldots, J)$ in sample i $(i = 1, \ldots, n)$. The term π_j^0 represents the baseline relative abundance that characterizes the true relative abundance of taxon j when the covariates $X_i = 0$. The coefficient β_j captures how the true relative abundance of taxon j changes with X_i. The parameter γ_j is the taxon-specific bias factor, reflecting how the relative abundance is influenced by distortions. Additionally, the term α_i represents the sample-specific normalization factor, ensuring the composition constraint $\sum_j^J p_{ij} = 1$. Then, LOCOM introduces the variable μ_{ij}, $\mu_{ij} = p_{ij}/(p_{ij} + p_{iJ})$ and utilizes the logistic model

$$log(\frac{\mu_{ij}}{1-\mu_{ij}}) = \theta_j + X_i^T(\beta_j - \beta_J), \quad 1 \leq j \leq J-1,$$

where $\theta_j = [log(\pi_j^0) - log(\pi_J^0)] + (\gamma_j - \gamma_J)$ are treated as nuisance parameters. LOCOM uses the Firth-corrected score equation

$$\mathbb{U}_j(\theta_j, \beta_j) = \sum_{i=1}^{n} (Y_{ij} - M_{ij}\mu_{ij} + h_i(0.5 - \mu_{ij})) \begin{bmatrix} 1 \\ X_i \end{bmatrix} = 0 \tag{1.3}$$

to estimate the parameters $(\theta_j,\ \beta_j)$ robustly, without relying on specific distributional assumptions in a standard logistic regression and addressing separation issues, where separation refers to the situation where all (or nearly all) counts for a taxon are zero in one group (e.g., the case or control group). In the estimating equation (1.3), $M_{ij} = Y_{ij} + Y_{iJ}$, where Y_{ij} is the read count of the jth taxon $(j = 1, ..., J)$ in the ith sample. h_i corresponds to the ith diagonal element of the weighted hat matrix $W_j^{\frac{1}{2}}X(X^TW_jX)^{-1}X^TW_j^{\frac{1}{2}}$ with the design matrix X and the diagonal weight matrix $W_j = \text{diag}\{M_{1j}\mu_{1j}(1-\mu_{1j}), \ldots, M_{nj}\mu_{nj}(1-\mu_{nj})\}$. To avoid relying on a specific taxon being considered as null, LOCOM tests hypotheses at individual taxa with a null hypothesis centering by the median

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

The inference of associations in the LOCOM is performed through permutation to address overdispersion and small sample sizes. LOCOM partitions the covariate vector X_i into the trait of interest T_i and the other covariates C_i. Then, it permutes the residual of the trait of interest obtained by regressing the trait on the other covariates C_i and an intercept, denoted by $T_i^{(b)}$. LOCOM constructs a new covariate vector $X_i^{(b)}$ by combining $T_i^{(b)}$ with C_i. Finally, it computes the permutation coefficients and test statistics based on this new covariate vector. LOCOM also provides a community-level test by combining the P values from tests of individual taxa using the test statistic $\sum_{j=1}^{J} p_j^{(-1)}$. LOCOM is flexible in handling both binary and continuous traits of interest, allowing for the simultaneous testing of multiple traits. It is also robust against experimental bias, even when bias factors vary between causal and non-causal taxa. Additionally, LOCOM offers the capability to adjust for potential confounding covariates.

In Topic 3, we propose a method, named LOCOM-I, to fill the significant gap of integrative analysis of 16S marker-gene and shotgun metagenomic sequencing data. We extend LOCOM to combining data from both 16S and SMS experiments, while allowing for differential experimental bias and assigning adaptive weights to each observation. Our method utilizes all available samples and taxa, whether they overlap or not between the two experiments, and adopt a permutation procedure that preserves any specific sample structure. To benchmark the performance of the new method, we introduce two additional ad hoc approaches: pooling read counts and combining p-values. In the Numerical Studies section, we present extensive simulation results and findings from analyzing the ORIGINS and dietary studies found in Qiita. We conclude with a Remarks section.

.

Chapter 2

Topic 1: A new approach to testing mediation of the microbiome at both the community and individual taxon levels

2.1 Method

2.1.1 Motivation

Our starting point is the following classical model for multiple mediators (VanderWeele and Vansteelandt, 2014). For a continuous outcome and J continuous (potential) mediators with no exposure-mediator and mediator-mediator interactions, the model specifies a linear model for each mediator and a linear model for the outcome that includes the effects of all mediators:

$$\mathrm{E}(M_j|Z,T) = \alpha_{0,j} + \alpha_{\mathrm{Z},j}^{\mathrm{T}} Z + \alpha_{1,j} T, \tag{2.1}$$

$$\mathrm{E}(O|Z,T,M_1,\ldots,M_J) = \theta_0 + \theta_{\mathrm{Z}}^{\mathrm{T}} Z + \theta_1 T + \textstyle\sum_{j=1}^{J} \theta_{2,j} M_j, \tag{2.2}$$

where the notation was introduced in Figure 1.1(b). It can be derived that the overall (total) mediation effect through $(M_1,\ldots,M_J)$ takes the form $\sum_{j=1}^{J}\alpha_{1,j}\theta_{2,j}$ (VanderWeele and Vansteelandt, 2014); note that $\alpha_{1,j}$ characterizes the effect of T on M_j given Z, and $\theta_{2,j}$ characterizes the effect of M_j on O given Z and T and all other M_js. When the mediators are independent of one another conditional on Z and T, each product term $\alpha_{1,j}\theta_{2,j}$ can be interpreted as the mediation effect through a single mediator M_j. Even if the mediators are not conditionally independent, a non-zero value of $\alpha_{1,j}\theta_{2,j}$ indicates a non-zero contribution of M_j to the overall mediation effect. Thus, our objective can be achieved by testing whether $\alpha_{1,j}\theta_{2,j} = 0$ at each potential mediator. However, the *forward* outcome model (2.2), although describing the mediation process in a natural order and enabling intuitive forms for the mediation effects, are not easily generalizable to an outcome that is a discrete, multivariate, or censored-survival-time variable. In addition, model (2.2) does not permit a large number of mediators, e.g., more mediators than samples, unless some regularization is imposed.

2.1.2 Inverse regression model

The limitations of the *forward* outcome model motivated us to adopt the *inverse* regression model that exchanges the positions of the outcome and mediators. Inverse regression is a commonly used approach, which, for example, has been widely used in genetics studies of multiple phenotypes (O'Reilly et al., 2012; Wu and Pankow, 2015; Majumdar et al., 2015). It has a key advantage of accommodating different types of outcomes. Also, it allows a large number of microbial taxa to be analyzed simultaneously by treating each taxon as the response variable in the regression, one at a time.

Here we assume that a mediating taxon acts through its relative abundance, so we let M_j denote the relative abundance of taxon j, although our methodology can easily accommodate presence-absence data (M_j taking value 1 or 0 indicating non-zero read count of taxon j in a sample). We find that, by properly orthogonalizing the exposure variable T and outcome variable O, we can obtain an inverse regression model that "merges" both the mediator model (2.1) and the forward regression model (2.2) into one regression. To this end, we define T_r to be the residual of T after orthogonalizing against Z, and O_r to be the residual of O after orthogonalizing against (Z, T). We consider the inverse regression model for taxon j

$$\mathrm{E}(M_j|Z,T,O) = \beta_{0,j} + \beta_{\mathrm{Z},j}^{\mathrm{T}} Z + \beta_{1,j} T_r + \beta_{2,j} O_r. \tag{2.3}$$

We show in Appendix Text A1 that $\beta_{1,j} = \alpha_{1,j}$ and that $\beta_{2,j} = 0$ and $\theta_{2,j} = 0$ coincide. As a result, testing

$$H_{0j} : \beta_{1,j}\beta_{2,j} = 0 \tag{2.4}$$

is equivalent to testing $\alpha_{1,j}\theta_{2,j} = 0$, i.e., whether there exists a mediation effect through taxon j. We can test (2.4) by obtaining the least-squares estimates from (2.3), denoted by $\widehat{\beta}_{1,j}$ and $\widehat{\beta}_{2,j}$, forming the test statistic $|\widehat{\beta}_{1,j}\widehat{\beta}_{2,j}|$, and using permutation to provide the null distribution of the test statistic. All of these can be achieved by using the LDM framework with minor modifications.

2.1.3 Testing mediation effects at individual taxa

As mentioned after equation (2.4), it is most natural to consider the following statistic for testing the mediation effect at taxon j:

$$\mathbb{U}_j = |\widehat{\beta}_{1,j}\widehat{\beta}_{2,j}|.$$

To provide a reference distribution for this statistic under the composite null of no mediation, we calculate the following statistic under the bth ($b = 1, \dots, B$) permutation:

$$\mathbb{U}_j^{(b)} = \max\left\{|\widehat{\beta}_{1,j}\widehat{\beta}_{2,j}^{(b)}|,\ |\widehat{\beta}_{1,j}^{(b)}\widehat{\beta}_{2,j}|,\ |\widehat{\beta}_{1,j}^{(b)}\widehat{\beta}_{2,j}^{(b)}|\right\},$$

where $\widehat{\beta}_{1,j}^{(b)}$ and $\widehat{\beta}_{2,j}^{(b)}$ are obtained by permuting T_r and O_r, separately, to break the T-M_j association given Z and the M_j-O association given (Z, T), respectively, and they are directly available from the LDM. The three product terms in $\mathbb{U}_j^{(b)}$ correspond to the test statistics under the type-I, type-II, and type-III null hypotheses. Because $\mathbb{U}_j^{(b)}$ is the maximum of three statistics whereas $\mathbb{U}_j$ is not,

$\mathbb{U}_j^{(b)}$ is inherently conservative in the sense that its distribution is more spread out than the *true* distribution of $\mathbb{U}_j$ under a specific type of null (unknown). Finally, the permutation p-value for taxon j is calculated to be $p_j = B^{-1}\sum_{b=1}^{B}\mathbb{I}\{\mathbb{U}_j^{(b)} \geq \mathbb{U}_j\}$, which is then corrected for multiple testing by Sandve's sequential stopping rule (Sandve et al., 2011) as implemented in the LDM. We refer to this approach to testing individual taxa as LDM-med-product. However, it is unclear how to handle multivariate exposures or outcomes, in which case there are more than one element in $\beta_{1,j}$ or $\beta_{2,j}$.

A second way is to base the test statistic on the p-values $p_{1,j}$ and $p_{2,j}$ for testing $\beta_{1,j} = 0$ and $\beta_{2,j} = 0$, respectively, which naturally accommodate multivariate exposures or outcomes and are directly available from the LDM. Now we consider the test statistic

$$\mathbb{Z}_j = \max(p_{1,j}, p_{2,j}),$$

and assess the significance of $\mathbb{Z}_j$ by using the same permutation procedure as above and calculating the statistic

$$\mathbb{Z}_j^{(b)} = \min\left\{\max(p_{1,j}, p_{2,j}^{(b)}), \max(p_{1,j}^{(b)}, p_{2,j}), \max(p_{1,j}^{(b)}, p_{2,j}^{(b)})\right\},$$

where the null p-values $p_{1,j}^{(b)}$ and $p_{2,j}^{(b)}$ are based on the rank statistics of $\widehat{\beta}_{1,j}^{(b)}$ and $\widehat{\beta}_{2,j}^{(b)}$, respectively, among all permutation replicates (Westfall and Young, 1993). Note that $\max(p_{1,j}, p_{2,j})$ can also be directly used as an analytical p-value for testing a single mediator (Boca et al., 2014), but here we choose permutation for inference because permutation is more robust and the permutation replicates are readily available from the LDM. Similarly to $\mathbb{U}_j^{(b)}$, the statistic $\mathbb{Z}_j^{(b)}$ is inherently conservative. Finally, the permutation p-value is calculated to be $p_j = B^{-1}\sum_{b=1}^{B}\mathbb{I}\{\mathbb{Z}_j^{(b)} \leq \mathbb{Z}_j\}$ and corrected for multiple testing by Sandve's sequential stopping rule (Sandve et al., 2011) as implemented in the LDM. We refer to this approach as LDM-med-maxP. In fact, this approach was found to be equivalent to LDM-med-product in simple settings, for example, when all variables are normally distributed (Boca et al., 2014). However, besides the conservative $\mathbb{U}_j^{(b)}$ and $\mathbb{Z}_j^{(b)}$, the stringent correction of all J tests in both LDM-med-product and LDM-med-maxP tends to make them even more inefficient.

A third approach is to directly apply the MultiMed procedure (Sampson et al., 2018) to the LDM p-values $p_{1,j}$ and $p_{2,j}$, which was developed to improve the efficiency of testing a large number of mediators. The idea is to restrict the mediation testing to a subset of taxa that have relatively small $p_{1,j}$ and $p_{2,j}$. Here, we briefly describe this procedure; the theoretical properties that guarantee the FDR control can be found in the original papers (Sampson et al., 2018; Bogomolov and Heller, 2018). First, for a nominal FDR level α, find the subset of taxa with relatively small $p_{1,j}$ to be

$\omega_{S1} = \{j : p_{1,j} < \alpha/2\}$, and denote the cardinality of the subset by $S_1 = \mathcal{C}(\omega_{S1})$. Similarly, find the subset with relatively small $p_{2,j}$ to be $\omega_{S2} = \{j : p_{2,j} < \alpha/2\}$ and denote $S_2 = \mathcal{C}(\omega_{S2})$. Then, the downstream testing of mediation is restricted to taxa at the intersection of the two subsets, which can greatly alleviate multiple testing correction. For taxon $j \in \omega_{S1} \cap \omega_{S2}$, define the subset-adjusted p-value

$$p_{S,j} = 2\max(S_2 p_{1,j}, S_1 p_{2,j}).$$

Taxon j is declared to be a mediator if the FDR-adjusted p-value

$$p_{D,j} = \min_{j' : p_{S,j'} \geq p_{S,j}} p_{S,j'}/\mathrm{rank}(p_{S,j'}) \leq \alpha.$$

We call this approach LDM-med-subset. Although the subset-based approach has shown to be more efficient than the approach based on the product of coefficients (similar to our first approach) in the context of controlling the family-wise error rate (Sampson et al., 2018), it is of interest to re-evaluate these approaches in the context of controlling the FDR.

A fourth approach is to directly apply the HDMT procedure (Dai et al., 2022) to the LDM p-values $p_{1,j}$ and $p_{2,j}$, which was developed to overcome the conservativeness of a mediation test due to the composite null. The core of the HDMT procedure is based on estimating the proportions of the three types of null and then the underlying mixture null distribution of the statistic $\max(p_{1,j}, p_{2,j})$. We call this approach LDM-med-HDMT.

2.1.4 Testing the overall mediation effect in a community

If every taxon in a community is under some type of null (not necessarily the same type), we declare a null community with no mediation effect. Recall that $\mathbb{Z}_j = \max(p_{1,j}, p_{2,j})$ has frequently been used as a p-value for testing a single mediator (Boca et al., 2014). Given these "p-values" at individual taxa, it is straightforward to construct a global test statistic by combining these "p-values". Here we adopt the Harmonic mean method (Wilson, 2019) to aggregate $\mathbb{Z}_j$s, which is more robust to the dependence structure among taxa than Fisher's method. The harmonic mean of $\mathbb{Z}_j$s is $J/\big(\sum_{j=1}^{J} \mathbb{Z}_j^{-1}\big)$, where a smaller value corresponds to a stronger evidence against the null hypothesis. To have a usual test statistic with a reverse directionality, we choose the statistic for the global test to be

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

We assess the significance of $\mathbb{Z}_{\text{global}}$ via permutation, since permutation is more robust and the permutation replicates are readily available. The statistic based on bth permutation replicate is $\mathbb{Z}_{\text{global}}^{(b)} = \sum_{j=1}^{J} \left\{ \mathbb{Z}_{j}^{(b)} \right\}^{-1}$, where $\mathbb{Z}_{j}^{(b)}$ has been introduced earlier. Finally, the permutation p-value for the global test is given by $p_{\text{global}} = B^{-1} \sum_{b=1}^{B} \mathbb{I} \left\{ \mathbb{Z}_{\text{global}}^{(b)} \geq \mathbb{Z}_{\text{global}} \right\}$. We call this test LDM-med-global, which is a natural extension of LDM-med-maxP but is also compatible with LDM-med-subset and LDM-med-HDMT in the sense that all are based on the p-values $p_{1,j}$ and $p_{2,j}$.

2.2 Numerical Studies

2.2.1 Simulation studies

Our simulations were based on data on 856 taxa of the upper-respiratory-tract (URT) microbiome (Charlson et al., 2010), and the mediator model (2.1) and the forward outcome model (2.2) as generative models. We focused on the sample size 100 but also considered 30 in some cases, because our murine microbiome dataset has 36 samples. Suppose that the exposure variable T_i is binary and that an equal number of samples were exposed ($T_i = 1$) and unexposed ($T_i = 0$). We considered continuous outcomes as well as binary outcomes. In what follows, we number the taxa by decreasing relative abundance so that taxon 1 is the most abundant. We considered three mediation mechanisms, in which we assumed the mediating taxa were the top five most abundant taxa (taxa 1–5), five less abundant taxa (taxa 51–55), and a mixture of the two sets (taxa 4–5 and 51–52); we refer to them as M-common, M-rare, and M-mixed, respectively. In all scenarios, we selected taxa 6–10 to be associated with the exposure but not with the outcome, and taxa 11–15 to be associated with the outcome but not with the exposure, corresponding to the type-I and type-II null taxa, respectively.

To generate the read count data, we first set the *baseline* (when $T_i = 0$) relative abundances of all taxa for all samples, denoted by $\overline{\pi}_i = (\overline{\pi}_{i1}, \overline{\pi}_{i2}, \ldots, \overline{\pi}_{iJ})$, to the population means that were estimated from the URT data. To induce the effects of the exposure T_i on a set of associated taxa (e.g., the mediating taxa or type-I null taxa), for those unexposed we kept $\overline{\pi}_i$ unchanged; for those exposed we decreased $\overline{\pi}_{ij}$ for some of the associated taxa by a percentage, which equals β_{TM} for the mediating taxa and α_{TM} (0 or 0.6) for the type-I null taxa, and we redistributed the decreased amount evenly over the remaining of the associated taxa. This way ensures that the relative abundances of non-associated taxa remain intact and the modified $\overline{\pi}_i$ satisfies the compositional constraint (unit sum). Note that β_{TM} captures the effects of the exposure on the mediating taxa and α_{TM} captures the effects of the exposure on the type-I null taxa. Specifically, in M-common, the increasing set of the mediating taxa was taxa 1–2 and the decreasing set was taxa 3–5; in M-rare, the two sets were taxa 51–52 and 53–55; in M-mixed, the two sets were taxa 4 and 52 and taxa 5 and 51. Among the type-I null taxa, the two sets were taxa 6–7 and 8–10. The modified $\overline{\pi}_i$ represents the *mean* relative abundances conditional on the exposure value. Then, we introduced sample heterogeneity by drawing the sample-specific composition $\pi_i = (\pi_{i1}, \pi_{i2}, \ldots, \pi_{iJ})$ from the Dirichlet distribution $Dir(\overline{\pi}_i, \theta)$ with mean $\overline{\pi}_i$ (after modification) and overdispersion θ that was set to 0.02 (as estimated from the URT data). Finally, we generated the read count data $M_i = (M_{i1}, M_{i2}, \ldots, M_{iJ})$ using the Multinomial distribution with mean π_i and the library sizes (sequencing depth) sampled from $N(10000, (10000/3)^2)$ and left truncated at 500.

To generate the outcome that is influenced by the mediating taxa, denoted by $\mathcal{M}$, and the type-II null taxa, denoted by $\mathcal{N}$, we partitioned each set of taxa into two subsets ($\mathcal{M}_1$ and $\mathcal{M}_2$, $\mathcal{N}_1$ and $\mathcal{N}_2$) with approximately equal total relative abundance. In particular, we set $\mathcal{M}_1$ and $\mathcal{M}_2$ to be the increasing and decreasing sets, respectively, that were determined earlier relative to the exposure and have similar total relative abundance; we set $\mathcal{N}_1$ and $\mathcal{N}_2$ to be taxa 11–12 and 13–15, respectively. To simulate a continuous outcome, we used the model

$$O_i = \beta_{\text{TO}} T_i + \beta_{\text{MO}} \text{scale}\left(\sum_{j \in \mathcal{M}_1} \pi_{ij} - \sum_{j \in \mathcal{M}_2} \pi_{ij} \right) + \alpha_{\text{MO}} \text{scale}\left(\sum_{j \in \mathcal{N}_1} \pi_{ij} - \sum_{j \in \mathcal{N}_2} \pi_{ij} \right) + \epsilon_i, \quad (2.5)$$

where scale(.) is a scaling function that standardizes a variable to have mean 0 and standard deviance 1, β_{TO} characterizes the direct effect of the exposure on the outcome and was fixed at 0.2 here, β_{MO} characterizes the effects of the mediating taxa on the outcome, α_{MO} characterizes the effects of the type-II taxa and was fixed at 0 or 0.4, and the error term ϵ_i was drawn from $N(0, 0.5^2)$. It can be verified that the taxa that are neither mediators nor type-II null taxa were uncorrelated with the outcome after controlling for T_i, owing to the counterbalancing effects of taxa in $\mathcal{M}_1$ and $\mathcal{M}_2$ (or $\mathcal{N}_1$ and $\mathcal{N}_2$) on the outcome. To simulate a binary outcome, we calculated the probability $\Pr(O_i = 1 | T_i, \pi_i) = \exp(\mu_i)/\{1+\exp(\mu_i)\}$ with μ_i being the same linear predictor as in (2.5), without the error term ϵ_i.

To simulate a confounder, we note that a confounder has effects on the exposure, the microbiome, and the outcome (Figure 1.1(b)). Thus, we first simulated the binary confounder Z_i with 70% "success" rate among the exposed and 30% among the unexposed. Then, we used the same decreasing and increasing sets of the mediating taxa as determined earlier, now with the deduction percentage $\gamma_{\text{ZM}} = 0.3$, and the same operation as for the exposure to further modify $\overline{\pi}_i$ for those with $Z_i = 1$. Finally, we modified the linear predictor in the outcome model (2.5) to include the term $\gamma_{\text{ZO}} Z_i$ with

Table 2.1: Type I error (at level 0.05) of the global tests in M-mixed with a continuous outcome and no confounder, in 12 scenarios under the global null

					Method		
β_{TM}	β_{MO}	α_{TM}	α_{MO}	Type(s) of null	LDM-med-global	MedTest	MODIMA
0.0	0.4	0.0	0.0	II	0.010	0.024	0.044
			0.4	II	0.007	0.024	0.048
		0.6	0.0	I, II	0.010	0.504	0.936
			0.4	I, II	0.010	0.547	0.985
0.6	0.0	0.0	0.0	I	0.004	0.031	0.042
			0.4	I, II	0.007	0.270	0.720
		0.6	0.0	I	0.008	0.038	0.051
			0.4	I, II	0.010	0.282	0.811
0.0	0.0	0.0	0.0	III	0.000	0.004	0.003
			0.4	II	0.005	0.018	0.039
		0.6	0.0	I	0.006	0.030	0.053
			0.4	I, II	0.009	0.317	0.813

Note: MedTest is the omnibus test that combines results from analyzing the Bray-Curtis and Jaccard distances. MODIMA is based on the Bray-Curtis distance. The parameters β_{TM} and β_{MO} determine the type of null that the pre-selected mediating taxa reduce to; α_{TM} controls the existence of the pre-selected type-I null taxa and α_{MO} controls the existence of the pre-selected type-II null taxa.

γ_{ZO} fixed at 0.7.

Prior to analysis, we filtered out taxa that were found in fewer than 5 subjects in the dataset, which resulted in ∼460 taxa remaining in analysis. For testing mediation effects at individual taxa, we compared our four approaches: LDM-med-maxP, LDM-med-product, LDM-med-subset, and LDM-med-HDMT (using the asymptotic version as recommended because the proportions of the type-I and type-II null taxa are small in all scenarios here). The sensitivity (proportion of the truly mediating taxa that were detected) and empirical FDR were assessed at the nominal level of 10% based on 1000 replicates of data. Note that none of CMM, SparseMCMM, and Zhang's method worked for our simulated data, as they either gave errors (due to the large number of taxa or extensive zero count data) or ran more than 10 hours. For testing the overall mediation effect, we applied LDM-med-global and compared it to MedTest and MODIMA whenever the latter were applicable. For MedTest, we adopted the omnibus test based on both the Bray-Curtis and Jaccard distance matrices, which would work well when mediating taxa are abundant and less abundant, respectively, and thus form a complementary pair. For MODIMA, we chose Bray-Curtis, as MODIMA allows one distance measure only and Bray-Curtis is the most commonly used distance in the literature and was also frequently used in the MODIMA paper. The type I error and power were assessed at the nominal level 0.05 based on 10000 and 1000 replicates of data, respectively.

2.2.2 Simulation results

For testing mediation at individual taxa, the subset approach (LDM-med-subset) had substantially improved sensitivity over the product (LDM-med-product) and maxP (LDM-med-maxP) approaches in all scenarios, while the latter two always had similar performance (Figures 2.1–2.5, A1–A2). As

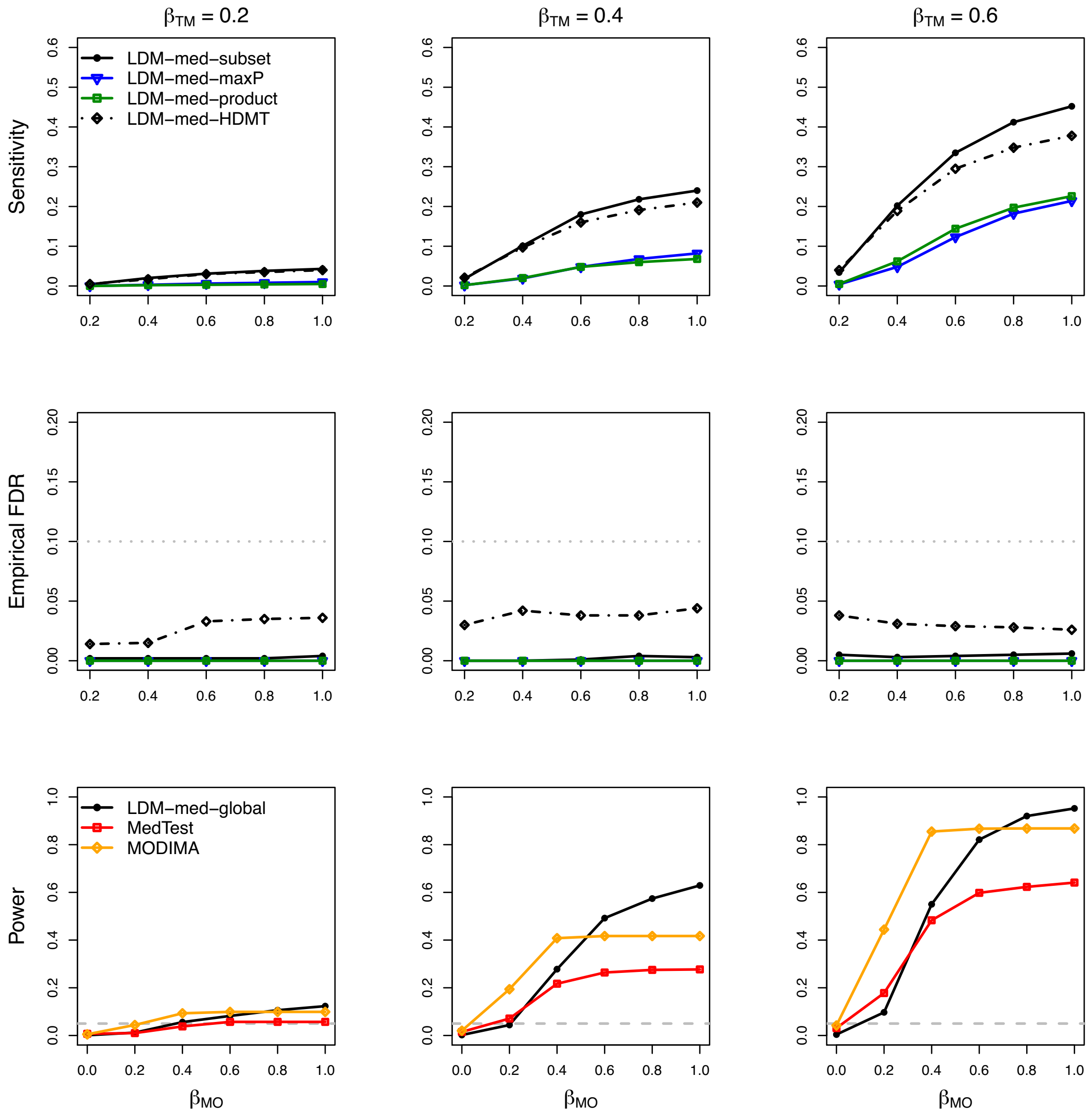


Figure 2.1: Simulation results in M-mixed with a continuous outcome and no confounder, in the absence of type-I and type-II null taxa ($\alpha_{\text{TM}} = 0$ and $\alpha_{\text{MO}} = 0$). The upper and middle panels pertain to sensitivity and empirical FDR, respectively, of the four approaches to testing individual taxa: LDM-med-product, LDM-med-maxP, LDM-med-subset, and LDM-med-HDMT, which are based on the product of coefficients as the test statistic, the maximum of coefficient p-values as the test statistic, a subset of promising taxa, and the HDMT procedure, respectively. The gray dotted line in the middle panel represents the nominal level of 10% for the FDR. The lower panel pertains to power of the proposed global test, LDM-med-global, and the existing global tests, MedTest and MODIMA. The gray dashed line there represents the nominal level 0.05 for the type I error.

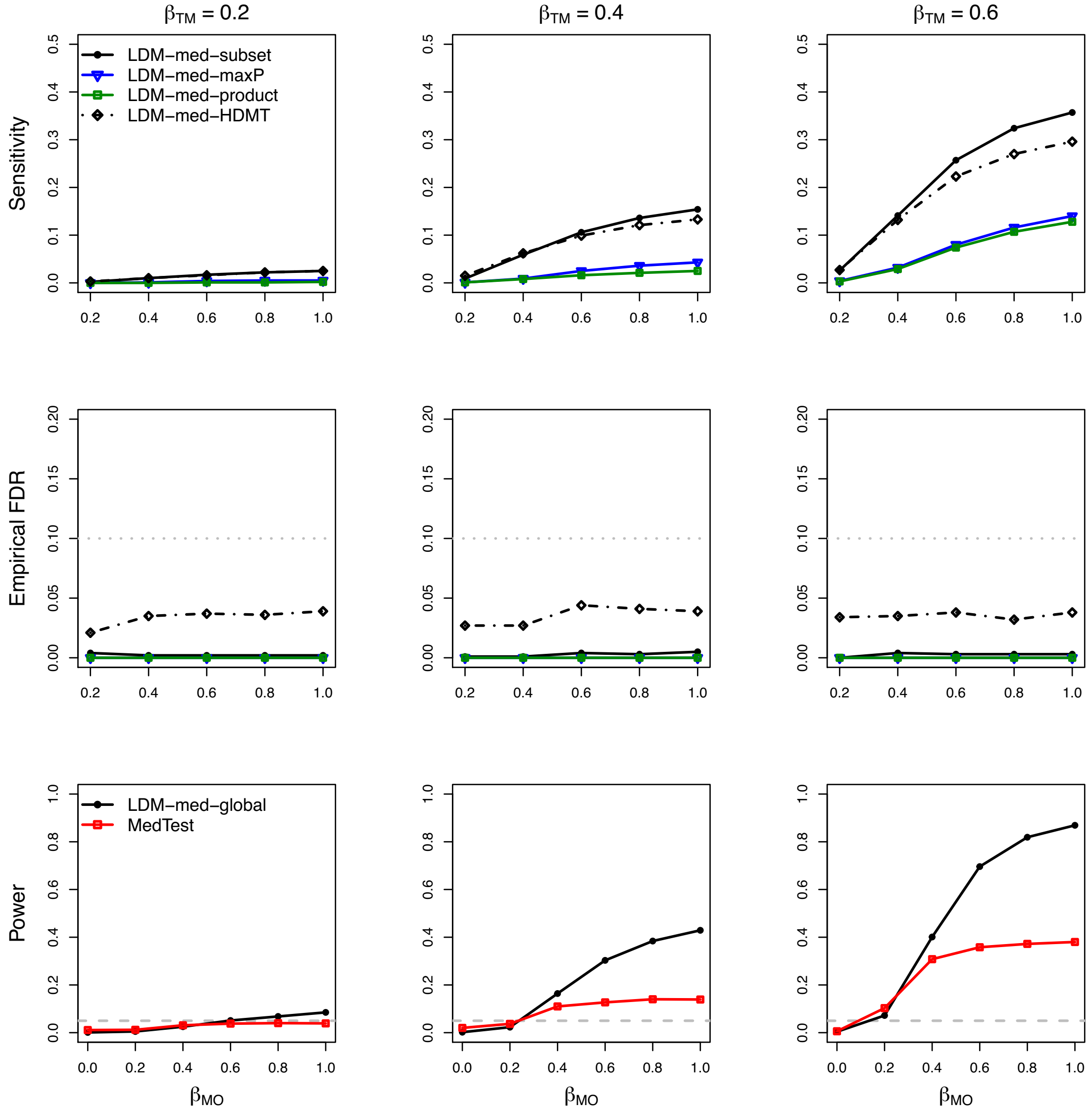


Figure 2.2: Simulation results in M-mixed with a continuous outcome and a confounder, in the absence of type-I and type-II null taxa. MODIMA was excluded because it does not allow adjustment of confounders.

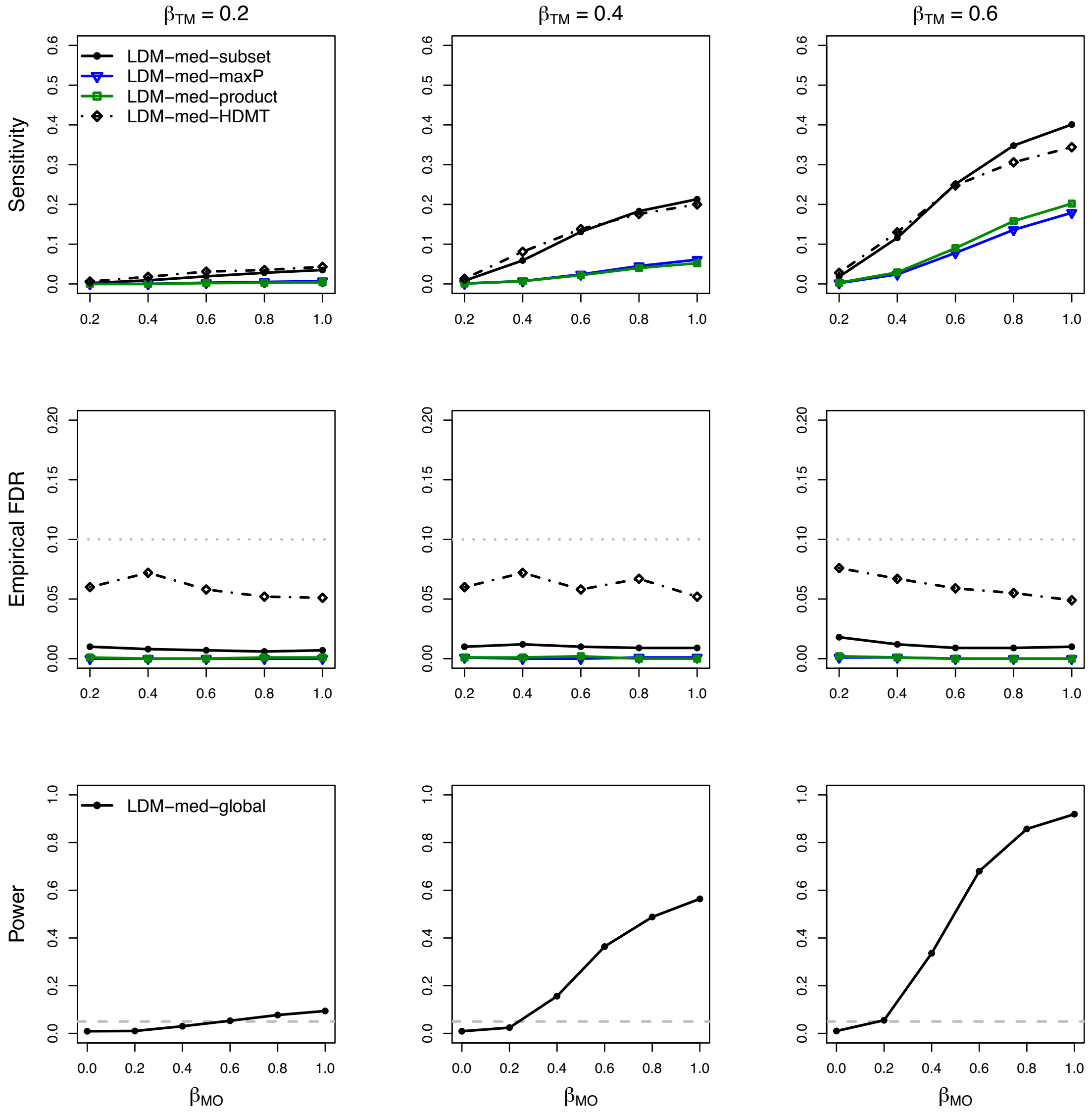


Figure 2.3: Simulation results in M-mixed with a continuous outcome and no confounder, in the presence of type-I and type-II null taxa ($\alpha_{\mathrm{TM}} = 0.6$ and $\alpha_{\mathrm{MO}} = 0.4$). MedTest and MODIMA were both excluded because they did not control the type I error (Table 2.1).

expected, all three approaches yielded conservative empirical FDR in all scenarios. Although the empirical FDR of the HDMT approach (LDM-med-HDMT) are less conservative (i.e., closer to the nominal level), its sensitivity results are generally comparable to those from LDM-med-subset in all scenarios. For these reasons, we always select LDM-med-subset as the recommended method for testing individual taxa.

The type I error results of the global tests in M-common, M-rare, and M-mixed are summarized in Tables 2.1 and A1. We considered 12 scenarios under the global null hypothesis, each corresponding to a specific combination of the three types of null taxa in a simulated community. For example, when $(\beta_{\text{TM}}, \beta_{\text{MO}}, \alpha_{\text{TM}}, \alpha_{\text{MO}}) = (0, 0.4, 0, 0)$, the pre-selected mediating taxa reduced to the type-II null taxa ($\beta_{\text{TM}} = 0$ and $\beta_{\text{MO}} = 0.4$), and both the pre-selected type-I and type-II null taxa reduced to the type-III null taxa ($\alpha_{\text{TM}} = 0$ and $\alpha_{\text{MO}} = 0$); here the type-III null taxa were viewed as a special case of either the type-I or type-II null taxa whichever existed in the community, so this community was determined to have type-II null taxa only. Clearly, MedTest and MODIMA easily lost control of the type I error whenever the type-I and type-II null taxa coexisted in the community. In all scenarios, LDM-med-global controlled the type I error; in fact, it was conservative as expected. In scenarios that consist of a single type of null taxa, MedTest and MODIMA controlled the type I error; then LDM-med-global appeared to have more conservative type I error than MedTest and MODIMA because LDM-med-global still allowed different taxa to be under different types of null.

In the presence of a confounder (Table A2), LDM-med-global controlled the type I error even when the confounder was not adjusted for, due to its conservativeness. As this provided no clue to the extent of the confounding effect and thus the capability of LDM-med-global in adjusting for the confounding effect, we considered a variant of LDM-med-global, called LDM-med-global*, that used the information on the type of null for each taxon (only available in simulation). Specifically, we modified $\mathbb{Z}_j^{(b)}$ to be $\max(p_{1,j}, p_{2,j}^{(b)})$, $\max(p_{1,j}^{(b)}, p_{2,j})$, or $\max(p_{1,j}^{(b)}, p_{2,j}^{(b)})$ depending on the actual type of null at taxon j. LDM-med-global* yielded inflated type I error when the confounder was not adjusted in the regression and type I error close to 0.05 when it was adjusted, demonstrating that LDM-med-global* (and hence LDM-med-global) was effective in adjusting for confounders.

For evaluating power of the global tests, we started with the scenarios when there were neither type-I nor type-II null taxa ($\alpha_{\text{TM}} = \alpha_{\text{MO}} = 0$), under which MedTest and MODIMA were valid. We also started with the simple case of a continuous outcome and no confounder. In M-common (Figure A1), MedTest and MODIMA were more powerful than LDM-med-global, whereas in M-rare (Figure A2), they were much less powerful than LDM-med-global, demonstrating that MedTest and MODIMA were effective in capturing mediation effects in abundant taxa but in not rare ones. In

M-mixed (Figure 2.1), the power of LDM-med-global crossed with that of MedTest and MODIMA; LDM-med-global performed best when β_{MO} was relatively large. The relative power of LDM-med-global and MedTest remained largely unchanged when a confounder was introduced to M-mixed (Figure 2.2); MODIMA was not included for comparison in this case because it cannot adjust for the confounder. When the type-I and type-II null taxa were both introduced (Figure 2.3), they invalidated both MedTest and MODIMA but minimally affected the performance of LDM-med-global. When we switched to a binary outcome (Figure 2.4), LDM-med-global lost power to MedTest and MODIMA. We wanted to know whether the power loss was the price that LDM-med-global paid in order to always allow different types of null at different taxa. To investigate this, we considered another variant of LDM-med-global, called LDM-med-global**, that assumed the same type of null (unknown) across all taxa as assumed in MedTest and MODIMA, and modified $\mathbb{Z}^{(b)}_{\mathrm{global}}$ to be $\max\left\{ \sum_{j=1}^{J}[\max(p_{1,j}, p^{(b)}_{2,j})]^{-1},\ \sum_{j=1}^{J}[\max(p^{(b)}_{1,j}, p_{2,j})]^{-1},\ \sum_{j=1}^{J}[\max(p^{(b)}_{1,j}, p^{(b)}_{2,j})]^{-1} \right\}$. Indeed, LDM-med-global** gained substantially power over LDM-med-global and had comparable or even better power than MedTest. Finally, when the sample size was reduced to merely 30 in the same scenario as in Figure 2.1, we observed similar patterns of results in Figure 2.5 compared to Figure 2.1.

2.2.3 Murine microbiome study

We analyzed the data generated from a murine microbiome study (Schulfer et al., 2019), which was conducted to explore whether the sub-therapeutic antibiotic treatment (STAT) would alter the gut microbiome composition and whether the altered gut microbiome would affect the body weight gain later in life. We focused on male mice for this analysis. The male mice were first randomized into the STAT and control groups, which was used as a binary exposure variable in our analysis. Then, their fecal samples were collected longitudinally at days 21 and 28. Bacterial DNAs were extracted from the fecal samples, sequenced for the 16S rRNA gene, and summarized into a taxa count table that initially contained 149 genera. Samples with less than 1800 reads, and genera with less than 10% presence or 0.01% mean relative abundance were filtered out, so the final taxa count table for our analysis included 41 genera and 36 mice (23 exposed to STAT and 13 unexposed); each mouse had two microbiome measurements at both time points. The mice body weight (in grams) prior to sacrifice was measured and used as a continuous outcome variable in our analysis. There were no additional covariates to be adjusted, as all potential confounders had been well-controlled in the randomized experiment.

It can be seen from Figure A3 that mice exposed to STAT were heavier than the control mice,

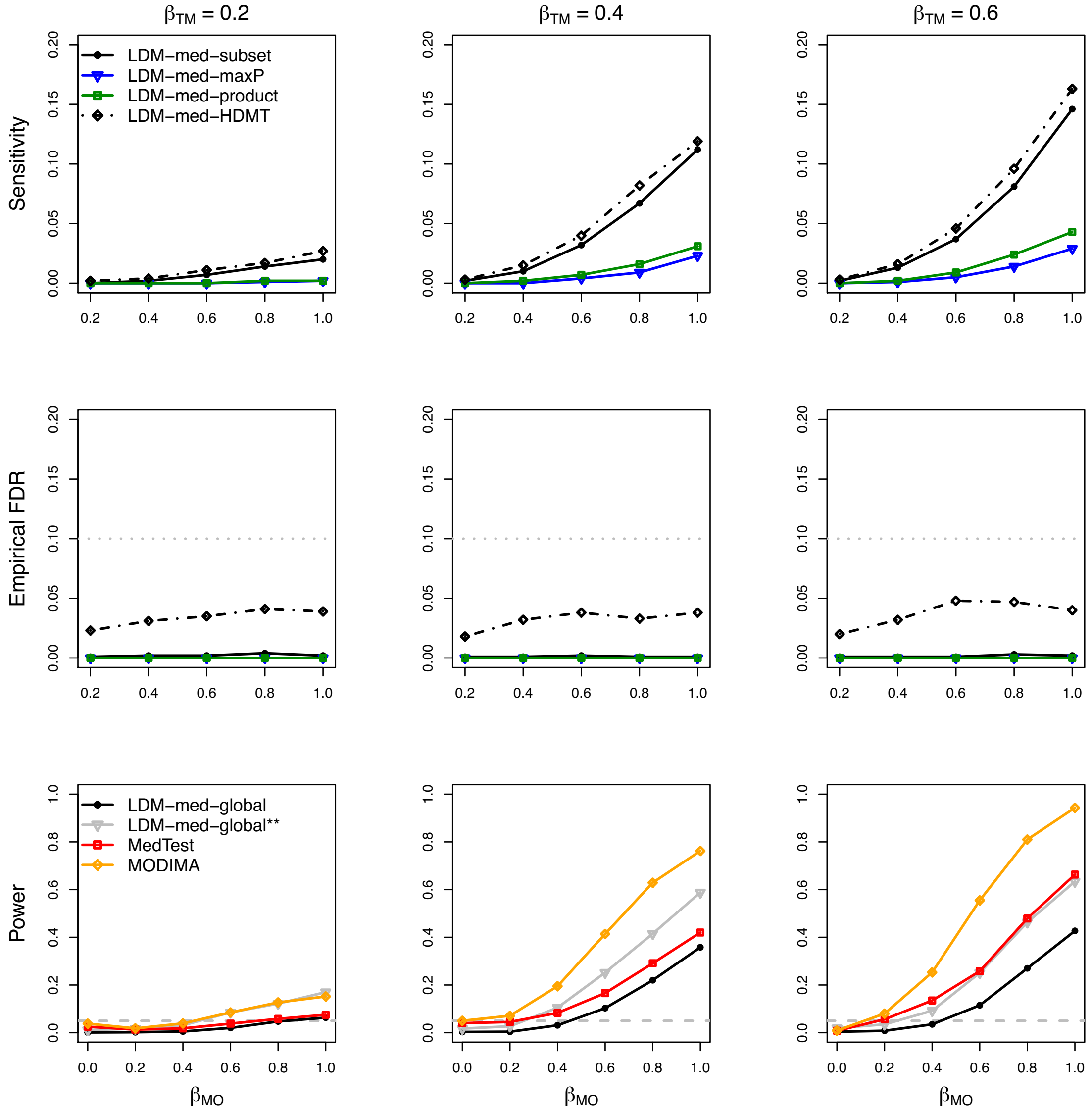


Figure 2.4: Simulation results in M-mixed with a binary outcome and no confounder, in the absence of type-I and type-II null taxa. LDM-med-global** is a variant of LDM-med-global that assumed the same type of null (unknown) for all taxa as was assumed in MedTest. The sample size was increased to 200 to obtain adequate power.

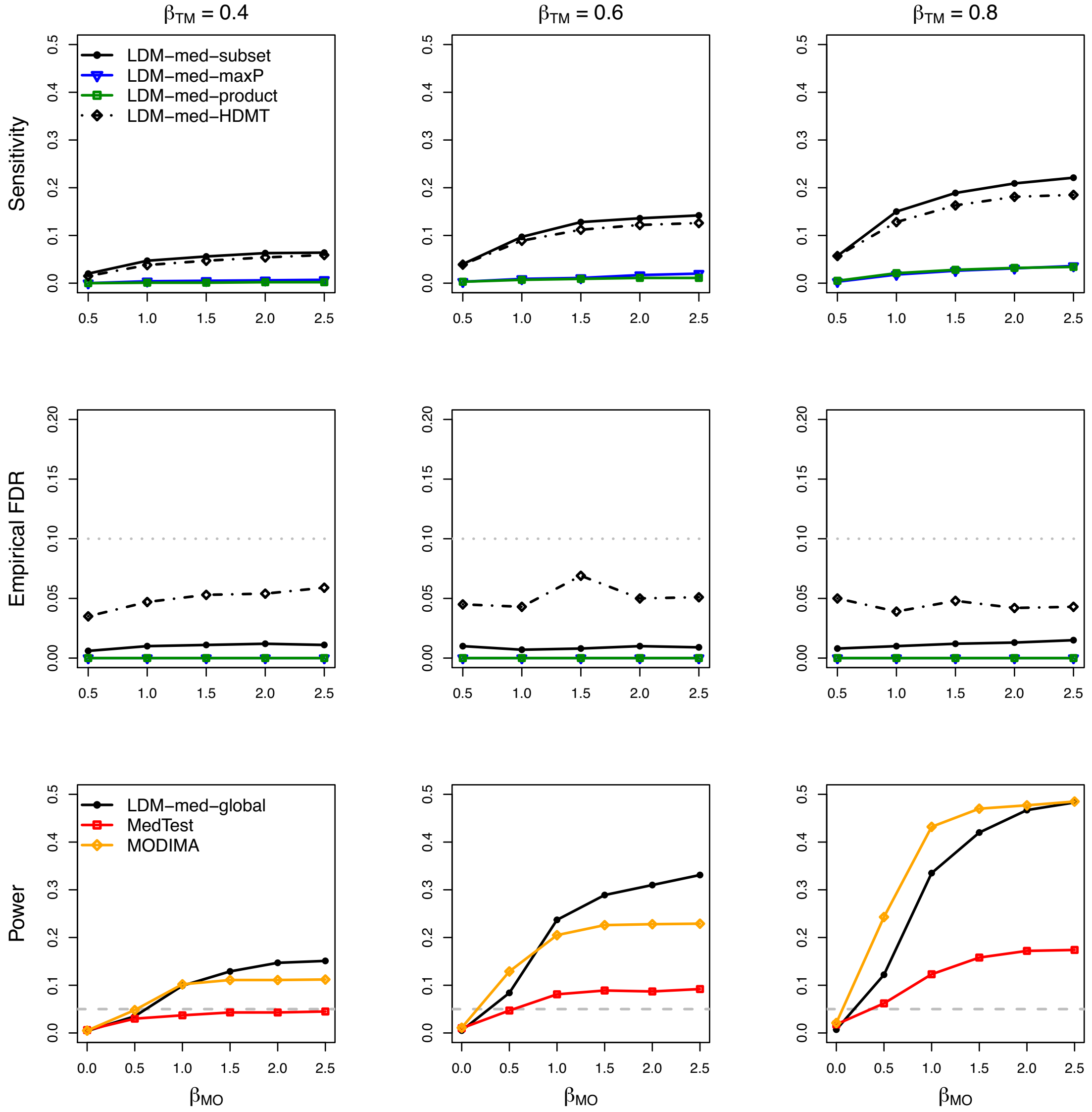


Figure 2.5: Simulation results in the same scenario as in Figure 2.1 but with sample size 30.

with a small Wilcoxon p-value 0.011. This motivated us to test whether this effect of STAT on body weight was mediated through the gut microbiome. For detecting individual mediating taxa (at the nominal FDR level of 20%, which was relatively high because the total number of genera was small), we applied LDM-med-subset and LDM-med-HDMT. For testing the overall mediation effect of the gut microbiome, we applied LDM-med-global, as well as MedTest, MODIMA, and SparseMCMM whenever they were applicable. Note that, although the outcome distribution somewhat deviated from the normal distribution (Figure A3), all methods should be robust to the deviation because LDM-related tests treat the outcome as a covariate, and MedTest, MODIMA, and SparseMCMM all base their inference on permutation.

All main results were summarized in Table 2.2. We first restricted our mediation analysis to the cross-sectional microbiome data at day 28 only. LDM-med-subset detected seven significant mediators, *[Ruminococus]* (a species that is misclassified to the genus *Ruminococcus* and is now awaiting to be formally renamed through the appropriate Code of Nomenclature), *Candidatus Arthromitus*, *Clostridiales*, *Clostridium*, *Ruminococcus*, *Dehalobacterium*, and *Oscillospira*, among which the first three genera were detected by LDM-med-HDMT. If the nominal FDR level of 10% were used, LDM-med-subset would detect one mediator *[Ruminococcus]* while LDM-med-HDMT would detect none. These results provided additional support for selecting LDM-med-subset over LDM-med-HDMT. Although SparseMCMM identified six mediators (shown in their Table A9), two of which (*[Ruminococus]* and *Clostridium*) overlapped with our detection list, SparseMCMM had no control for the FDR. To gain more insights into these results, we performed analysis of the bivariate association between the exposure and the relative abundance of each taxon using the Wilcoxon rank-sum test, and the bivariate association between each taxon and the outcome conditional on the exposure using the standard linear regression (treating the outcome as the response variable, and the exposure and taxon as covariates). We corrected multiple testing in each association analysis by the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) at the nominal FDR level of 20%. As shown in Table A3, 25 taxa were detected to be associated with the exposure, including all seven mediators detected by LDM-med-subset; five of the seven mediators were confirmed to be associated with the outcome, and the other two mediators ranked next but failed to pass the threshold of significance here. Thus, the mediators identified by LDM-med-subset seem plausible. For testing the community-level mediation, LDM-med-global produced a global p-value 0.0351. SparseMCMM yielded a more significant global p-value 0.004. Both MedTest (the omnibus test of Bray-Curtis and Jaccard distances) and MODIMA (based on the Bray-Curtis distance) produced non-significant global p-values 0.379 and 0.133, respectively.

Table 2.2: Mediation analysis of the murine microbiome dataset

		Three analyses		
	Method	Day 28 (continuous outcome)	Days 21 & 28 (clustered samples†)	Day 28 (multivariate outcome‡)
Detected taxa (FDR = 20%)	LDM-med-subset	*[Ruminococcus]** *Candidatus Arthromitus* *Clostridiales* *Clostridium* *Ruminococcus* *Dehalobacterium* *Oscillospira*	*[Ruminococcus]* *Candidatus Arthromitus** *Clostridiales* *Clostridium**	None
	LDM-med-HDMT	*[Ruminococcus]* *Candidatus Arthromitus* *Clostridiales*	*[Ruminococcus]* *Candidatus Arthromitus* *Clostridiales* *Clostridium*	None
Global p-value	LDM-med-global	0.0351	0.0387	0.633
	MedTest	0.379	-	-
	MODIMA	0.133	-	0.177
	SparseMCMM	0.004	-	-

Note: *[Ruminococcus]* is a species that is misclassified to the genus *Ruminococcus* and is now awaiting to be formally renamed through the appropriate Code of Nomenclature. *: taxa that would have been detected at the nominal FDR level of 10%. †: the microbiome data from days 21 and 28 tend to cluster within subjects, i.e., more correlated within subjects. ‡The weight gain outcome values were categorized into three categories by the 33rd and 66th percentiles. The detected taxa are listed such that the common taxa generated from different analyses appear in the same rows.

We also performed mediation analysis of the longitudinal (clustered) microbiome data at both days 21 and 28. Note that the outcome was observed only once per subject. While no other methods exist to analyze mediation of the microbiome data with correlations, LDM-related tests inherited such a capability from the LDM (by setting perm.within.type=”none” and perm.between.type=”free”). Here, a time variable (1/0) indicating day 28 was included as a covariate Z, as the microbiome composition was found to be significantly different between the two times (p-value 0.040 by the LDM for analyzing the matched-pair data). The results of mediation analysis by LDM-related tests were largely consistent with the previous results based on the data at day 28 only. We again performed analysis of bivariate associations between the exposure and each taxon by applying the LDM to the clustered data (adjusted for the time effect); we performed analysis of bivariate associations between each taxon and the outcome conditional on the exposure using the standard linear regression (regressing the outcome variable on the exposure, the relative abundances of the taxon at days 21 and 28, and testing the joint effect of the two relative abundance variables using the F-test). The results were again largely consistent with the previous results on bivariate associations using the data at day 28 only (Table A3).

Finally, to illustrate the capability of LDM-related tests to handle categorical outcome variables, we converted the continuous outcome variable into a three-level categorical variable by the 33rd and 66th percentiles. For this type of outcome variables, only LDM-related tests and MODIMA were applicable, none of which, however, identified any significant mediation effect.

2.3 Remarks

We presented a new approach to mediation analysis of the microbiome that is based on inverse regression and the LDM framework. We call the mediation framework based on the LDM LDM-med, which consists of LDM-med-subset for testing the taxon-level mediation and LDM-med-global for testing the community-level mediation. LDM-med offers maximum robustness to the complex features in the taxa count data (e.g., high-dimensionality, sparsity, and overdispersion), and provides extensive flexibility to accommodate various exposures and outcomes and study designs. Specifically, using the simulated and real data, we demonstrated the capabilities of LDM-med to deal with null taxa under different types of null hypothesis of no mediation, continuous, binary, and multivariate outcomes, clustered data with the exposure and outcome variables varying *between* the clusters, and adjustment of confounding covariates. In addition, LDM-med could also handle clustered data with the exposure and/or outcome variables varying *within* the clusters (Zhu et al., 2021), and perform analysis at the presence-absence scale using a rarefaction-without-resampling approach (Hu and Satten, 2021). In summary, LDM-med can be highly useful in practice.

We have added LDM-med to our existing R package LDM. The computation of LDM-med is as efficient as the LDM. For example, using a single-thread MacBook Pro laptop (2.9 GHz Quad-Core Intel Core i7, 16GB memory), it took 46s to analyze one simulated dataset having 100 samples and ∼460 taxa (after filtering); it took 126s to analyze one simulated dataset having 200 samples and ∼700 taxa (after filtering). The murine dataset was at a smaller scale, consisting of 36 mice and 41 genera, so it took only 5s and 12s to analyze the data at day 28 only and the data at both day 21 and day 28, respectively.

LDM-med tests the marginal mediation effect for each taxon, and thus the identified mediators may not all be true biological mediators, which are called ”probable mediators” but not ”true mediators” (Sampson et al., 2018). This compromise was made in order to obtain controlled FDR for the detected mediators, which we deem as critical in the initial ”scan” of high-dimensional features to generate ”targets” to follow up in the downstream mechanistic study. This strategy has been very common in the analysis of high-dimensional omic data (Asher et al., 2009; Hu et al., 2015; Sampson et al., 2018).

Chapter 3

Topic 2: Extension of PERMANOVA to testing the mediation effect of the microbiome

3.1 Method

3.1.1 Motivation

Assuming a continuous outcome and a continuous mediator and further assuming no exposure-mediator interaction and no unmeasured confounding, the classical mediation model (Baron and Kenny, 1986) specifies a linear model for the mediator and a linear model for the outcome:

$$\mathrm{E}(M|Z,T) = \alpha_0 + \alpha_{\mathrm{Z}}^{\mathrm{T}} Z + \alpha_T T, \tag{3.1}$$

$$\mathrm{E}(O|Z,T,M) = \theta_0 + \theta_{\mathrm{Z}}^{\mathrm{T}} Z + \theta_T T + \theta_M M. \tag{3.2}$$

Note that α_T characterizes the effect of T on M given Z, and θ_M characterizes the effect of M on O given Z and T. Then it can be shown that the mediation effect is given by $\alpha_T \theta_M$ (VanderWeele and Vansteelandt, 2009). However, it is unclear how to use the microbiome composition data, which are represented by a distance matrix here, as a mediator. Also, the *forward* outcome model (3.2) is not easily generalizable to an outcome variable that is discrete, multivariate, or censored survival time.

These limitations motivated us to adopt the *inverse* regression model that exchanges the positions of the outcome and the mediator in model (3.2). Inverse regression is a commonly used approach to testing associations (O'Reilly et al., 2012; Wu and Pankow, 2015; Majumdar et al., 2015). It has a key advantage of accommodating different types of outcome variables including multivariate variables. In what follows, we show that, by proper orthogonalization of the non-microbiome variables, the inverse regression model we consider "merges" both models (3.1) and (3.2) into one regression model, which fits nicely into the framework of PERMANOVA that takes the distance matrix as the response variable. To be specific, we first sequentially orthogonalize variables Z, T, and O, and denote the residual of T after orthogonalizing against Z by T_r and denote the residual of O after orthogonalizing against (Z,T) by O_r. Then, we consider the inverse regression model

$$\mathrm{E}(M|Z,T,O) = \beta_0 + \beta_{\mathrm{Z}}^{\mathrm{T}} Z + \beta_T T_r + \beta_O O_r. \tag{3.3}$$

For now, we view M as a univariate continuous variable, just as in (3.1) and (3.2). Model (3.3) implies that $\mathrm{E}(M|Z,T) = \beta_0 + \beta_{\mathrm{Z}}^{\mathrm{T}} Z + \beta_T T_r$, which is exactly model (3.1) after replacing T by T_r. Thus, we easily obtain that $\beta_T = \alpha_T$. Although it is well known that $\beta_O \neq \theta_M$, we see that $\beta_O = 0$ and $\theta_M = 0$ coincide as they both capture the microbiome-outcome association given (Z,T). As a result, testing $\beta_T \beta_O = 0$ is equivalent to testing $\alpha_T \theta_M = 0$, i.e., whether there exists a mediation effect through M. We find that model (3.3) fits nicely into the PERMANOVA framework, in which we view M as a distance matrix and the linear regression as a partition of M into additive components corresponding to the orthogonal factors (Z, T_r, O_r).

3.1.2 PERMANOVA-med: Extension of PERMANOVA to mediation analysis

Following the same notation in (1.1) and (1.2), under model (3.3), we set submodels $X_1 = Z$, $X_2 = T_r$, and $X_3 = O_r$ and denote the PERMANOVA F-statistics for testing microbiome associations with T_r and O_r by F_T and F_O, respectively. Then, we propose to test the existence of a mediation effect by the microbiome, i.e., $H_0 : \beta_T \beta_O = 0$, using the test statistic

$$\mathbb{U}_{\text{PERMANOVA-med}} = F_T F_O.$$

To claim a mediation effect by the microbiome, both the exposure-microbiome and microbiome-outcome associations (given the exposure) are required to be significant. Thus, the null hypothesis of no mediation is a composite null that consists of no exposure-microbiome association, no microbiome-outcome association, or neither. Accordingly, we construct the statistic for the bth permutation replicate,

$$\mathbb{U}^{(b)}_{\text{PERMANOVA-med}} = \max\left\{F_T^{(b)} F_O, \quad F_T F_O^{(b)}, \quad F_T^{(b)} F_O^{(b)}\right\},$$

where the three product terms correspond to the statistics under the three types of null hypotheses. Then, the p-value is obtained as the proportion of $\mathbb{U}^{(b)}_{\text{PERMANOVA-med}}$ that are equal to or larger than the observed statistic $\mathbb{U}_{\text{PERMANOVA-med}}$. Note that all the F-statistics needed for calculating the p-value are directly available from PERMANOVA. As a result, our mediation analysis implemented in the PERMANOVA framework naturally inherits all the features in PERMANOVA.

3.2 Numerical Studies

3.2.1 Simulation studies

Our simulations were based on data on 856 taxa of the upper-respiratory-tract (URT) microbiome (Charlson et al., 2010), and the mediator model (3.1) and the forward outcome model (3.2) as generative models. We considered both binary and continuous exposure variables, continuous outcome variables, and 100 or 200 sample size (n); note that both MedTest and MODIMA papers considered continuous exposures only. In what follows, we number the taxa by decreasing relative abundance so that taxon 1 is the most abundant. We considered three mediation mechanisms, in which we assumed the mediating taxa were the top five most abundant taxa (taxa 1–5), 100 relatively rare taxa (taxa 51–150), and a mixture of abundant and relatively rare taxa (taxa 4, 5, 51, and 52), which are referred to as M-common, M-rare, and M-mixed, respectively. We further assumed that the mediating taxa played the role through their relative abundances in M-common and M-mixed and through their presence-absence (0/1) statuses in M-rare.

Specifically, for a binary exposure T_i, we assigned half of the samples $T_i = 1$ and the other half $T_i = 0$. For a continuous exposure T_i, we sampled T_i from the Beta(2, 2) distribution. We initially set the baseline relative abundances of all taxa for all samples to the population means that were estimated from the real data, which we denote by $\overline{\pi}_i = (\overline{\pi}_{i1}, \overline{\pi}_{i2}, \ldots, \overline{\pi}_{iJ})$. To induce the effects of the exposure on the mediating taxa, we decreased $\overline{\pi}_{ij}$ by the percentage $\beta_{\mathrm{TM}} T_i$ $(\in [0, 1])$ for taxa 3–5 in M-common and taxa 5 and 51 in M-mixed, and then redistributed the decreased amount evenly over the remaining mediating taxa, i.e., taxa 1–2 in M-common and taxa 4 and 52 in M-mixed. In M-rare, we set $\overline{\pi}_{ij}$ for the mediating taxa to 0 with the probability $\beta_{\mathrm{TM}} T_i$ independently, and increased $\overline{\pi}_{ij}$ of the most abundant taxon by the total mass that had been set to 0 (which did not affect the presence-absence statuses of the most abundant taxon as it was always present). This way of modifying $\overline{\pi}_i$ did not change the relative abundances of non-associated taxa (except for the most abundant taxon in M-rare) and the modified $\overline{\pi}_i$ still satisfied the compositional constraint (unit sum). Note that β_{TM} characterizes the exposure-microbiome (T-M) association and $\beta_{\mathrm{TM}} = 0$ corresponds to no T-M association. Next, we drew the sample-specific composition $\pi_i = (\pi_{i1}, \pi_{i2}, \ldots, \pi_{iJ})$ from the Dirichlet distribution $Dir(\overline{\pi}_i, \theta)$, where the overdispersion parameter θ was set to 0.02 (as estimated from the real data). Then, we generated the read count data using the Multinomial distribution with mean π_i and library size (sequencing depth) sampled from $N(10000; (10000/3)^2)$ and truncated at 2000. Finally, we scaled each read count by the library size to obtain the observed relative abundance, denoted by M_{ij} for taxon j in sample i.

In M-common and M-mixed, we generated the continuous outcome O_i from the following model that allows different directions for the effects of different taxa on the outcome:

$$O_i = \beta_{\mathrm{TO}} T_i + \beta_{\mathrm{MO}} \mathrm{scale} \left(\sum_{j \in \mathcal{A}_1} M_{ij} - \sum_{j \in \mathcal{A}_2} M_{ij} \right) + \epsilon_i, \tag{3.4}$$

where $\mathcal{A}_1$ and $\mathcal{A}_2$ are the “increasing” and “decreasing” subsets of mediating taxa as determined above and $\epsilon_i \sim N(0, 0.5^2)$. In M-rare, we let $\mathcal{A}_1$ and $\mathcal{A}_2$ to include taxa 51–100 and taxa 101–150,

Table 3.1: Type I error (at the level 0.05) in analysis of simulated data without a confounder

Scenario	Exposure	β_{TM}	β_{TO}	n	PERMANOVA-med	MedTest	MODIMA
M-common	Binary	0.2	0.1	100	0.012	0.021	0.017
		0.4	0.1	100	0.044	0.049	0.046
		0.4	0.8	100	0.044	0.049	0.086
		0.4	0.8	200	0.046	0.052	0.126
	Continuous	0.4	0.1	100	0.009	0.016	0.013
		0.6	0.1	100	0.026	0.032	0.025
		0.6	0.8	100	0.026	0.032	0.040
		0.6	0.8	200	0.048	0.045	0.072
M-mixed	Binary	0.4	0.1	100	0.014	0.019	0.017
		0.6	0.1	100	0.039	0.043	0.040
		0.6	0.8	100	0.039	0.043	0.047
		0.6	0.8	200	0.048	0.049	0.068
	Continuous	0.6	0.1	100	0.004	0.010	0.007
		0.8	0.1	100	0.011	0.016	0.013
		0.8	0.8	100	0.011	0.016	0.016
		0.8	0.8	200	0.027	0.033	0.038
M-rare	Binary	0.2	0.1	100	0.039	0.041	0.042
		0.4	0.1	100	0.050	0.028	0.041
		0.4	0.8	100	0.050	0.028	0.088
		0.4	0.8	200	0.052	0.023	0.125
	Continuous	0.6	0.1	100	0.045	0.046	0.042
		0.8	0.1	100	0.044	0.034	0.039
		0.8	0.8	100	0.044	0.034	0.082
		0.8	0.8	200	0.049	0.026	0.125

Note: The type I error results were generated at $\beta_{\text{MO}} = 0$ (i.e., no M-O association), and thus the same for datasets using different sets of taxa for generating the M-O association.

respectively, and replaced M_{ij} in (3.4) by $I(M_{ij} \neq 0)$. We also considered a modification of the microbiome-outcome (M-O) association by restricting $\mathcal{A}_1$ and $\mathcal{A}_2$ to a subset of originally selected taxa, i.e., taxa 4 and 5 in M-common, taxa 51 and 52 in M-mixed, and taxa 101–150 in M-rare.

We simulated a binary confounder Z_i in settings with a binary exposure. Note that a confounder is associated with the exposure, the microbiome, and the outcome simultaneously. First, we generated $Z_i = 1$ with probability 0.7 among samples with $T_i = 1$ and with probability 0.3 among those with $T_i = 0$. Then, we used the same operation as used for simulating the T-M association, except that we replaced $\beta_{\text{TM}}T_i$ by $\gamma_{\text{ZM}}Z_i$ with $\gamma_{\text{ZM}} = 0.6$, to further modify $\overline{\pi}_{ij}$ based on Z_i for the mediating taxa that had been modified based on T_i. Finally, we added the term $\gamma_{\text{ZO}}Z_i$ with $\gamma_{\text{ZO}} = 0.7$ to model (3.4).

We applied PERMANOVA-med and compared it to MedTest and MODIMA, for testing the mediation effect of the microbiome in the simulated data. In M-common and M-mixed, all tests were based on the Bray-Curtis distance. In M-rare, all tests were based on the Jaccard distance. The type I error and power of all tests were assessed at the nominal level 0.05 based on 10000 and 1000 replicates of data, respectively.

Table 3.2: Type I error (at the level 0.05) in analysis of simulated data with a binary exposure and a binary confounder

Scenario	β_{TM}	PERMANOVA-med	MedTest	MODIMA
M-common	0.2	0.008	0.014	0.242
	0.4	0.035	0.040	0.385
M-mixed	0.4	0.006	0.015	0.056
	0.6	0.020	0.029	0.103
M-rare	0.2	0.026	0.036	0.279
	0.4	0.046	0.035	0.238

Note: $\beta_{\mathrm{TO}} = 0.1$, $\beta_{\mathrm{MO}} = 0$, and $n = 100$. MODIMA does not allow adjustment of confounders.

3.2.2 Simulation results

We first present results for the simulated data without a confounder. The power of the PERMANOVA-med, MedTest, and MODIMA with varying values of β_{MO}, β_{TM}, β_{TO}, and sample size n are displayed in Figures 3.1, 3.2, and 3.3 for M-common, M-mixed, and M-rare, respectively. The numerical values of the type I error rates (when $\beta_{\mathrm{MO}} = 0$) shown in these figures are also listed in Table 3.1.

In M-common with a binary exposure, when the same abundant taxa (taxa 1–5) were used to generate both the T-M and M-O associations (Figure 3.1(a)), MedTest was slightly more powerful than PERMANOVA-med, possibly because the top PCs used by MedTest effectively captured both the T-M and M-O associations. When a subset of taxa (taxa 4 and 5) were used for generating the M-O association (Figure 3.1(b)), the power of MedTest declined much more quickly than the power of PERMANOVA-med, as the PCs that captured the T-M association (e.g., PC1) may not coincide with the PCs that captured the M-O association (e.g., PC2). MODIMA seemed to be very powerful in some cases (e.g., Figure 3.1(a)), but its performance was sensitive to the value of β_{TO}. In particular, MODIMA generated inflated type I error when β_{TO} was enlarged to 0.8 and especially when n was also increased from 100 to 200.

In M-common with a continuous exposure, which tended to result in more complex variation patterns in the data than a binary exposure, MedTest (and MODIMA) lost the advantage in power to PERMANOVA-med, even when taxa 1–5 were used for both the T-M and M-O associations (Figure 3.1(c)). Again, MedTest lost further, considerable power to PERMANOVA-med when taxa 4 and 5 were used for the M-O association (Figure 3.1(d)) and MODIMA yielded inflated type I error when β_{TO} and n were both large.

As expected, PERMANOVA-med always had significantly higher power than MedTest in M-mixed (Figure 3.2), and the power difference was more pronounced in M-rare (Figure 3.3), since PCs became less efficient in capturing variations in less abundant taxa. In M-rare, MODIMA was uniformly less powerful than PERMANOVA-med, even its type I error was clearly inflated.

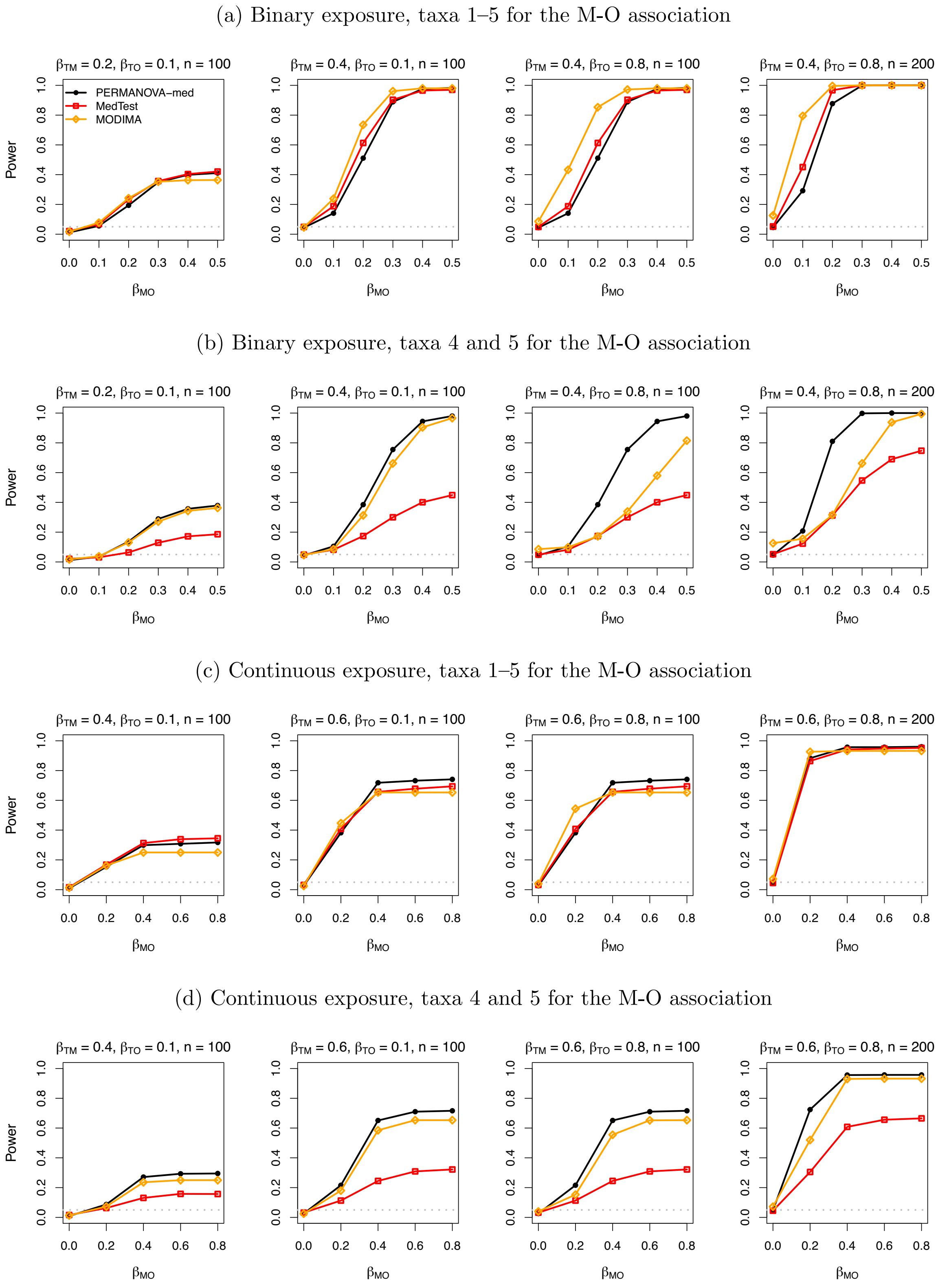
Finally, when a confounder was added to the simulated data, MODIMA, without the capability to adjust for the confounding effect, produced very inflated type I error (Table 3.2). Note that, PERMANOVA-med and MedTest always controlled the type I error below the nominal level, with (Table 3.2) or without (Table 3.1) the confounder.

3.2.3 Real data on melanoma immunotherapy response

The real data (Spencer et al., 2021) we used were generated from a cohort of 167 melanoma patients, who received immune checkpoint blockade (ICB) treatment and were classified as 106 responders and 61 non-responders. Their progression-free survival times (in days) were observed for 61 patients, censored for 49 patients, and missing for 57 patients. Their gut microbiome were profiled via shotgun metagenomic sequencing to generate a taxa count table including 225 taxa (lowest taxon known for a feature, up to species). These patients were further asked to complete a lifestyle survey, which included assessment of dietary fiber intake and use of probiotic supplements within the past month; 110 provided data for probiotic use, 94 provided data for dietary fiber intake, and 89 provided data for both.

Spencer et al. (Spencer et al., 2021) found in this dataset that higher dietary fiber intake was associated with significantly improved progression-free survival, with the most pronounced benefit observed in patients with sufficient dietary fiber intake and no probiotic use. They also found marginal significance for the association of dietary fiber intake and response to ICB. In addition, the influence of the gut microbiome on immunotherapy response has been demonstrated in numerous human cohorts as well as in preclinical models (Routy et al., 2018; Matson et al., 2018), and the human gut microbiome is itself shaped by diet (McDonald et al., 2018) and medication use (Maier et al., 2018). Given this interplay between diet and medication use, gut microbiome, and immunotherapy response, a natural question that arose was then whether some effect of dietary fiber intake and probiotic use on immunotherapy response in this dataset was mediated through the gut microbiome.

We performed a variety of mediation analyses using this dataset. For the outcome, we considered both the progression-free survival and the response to ICB, the former of which is a possibly censored survival time variable and the latter is a binary variable. For the exposure, we considered the dietary fiber intake (sufficient or insufficient), the probiotic use (no/yes), and the four-level categorical variable defined by both dietary fiber intake and probiotics use. Following (Spencer et al., 2021), we additionally compared patients with sufficient dietary fiber intake and no probiotic use to all other three groups. We selected body mass index (BMI), prior treatment, lactate dehydrogenase



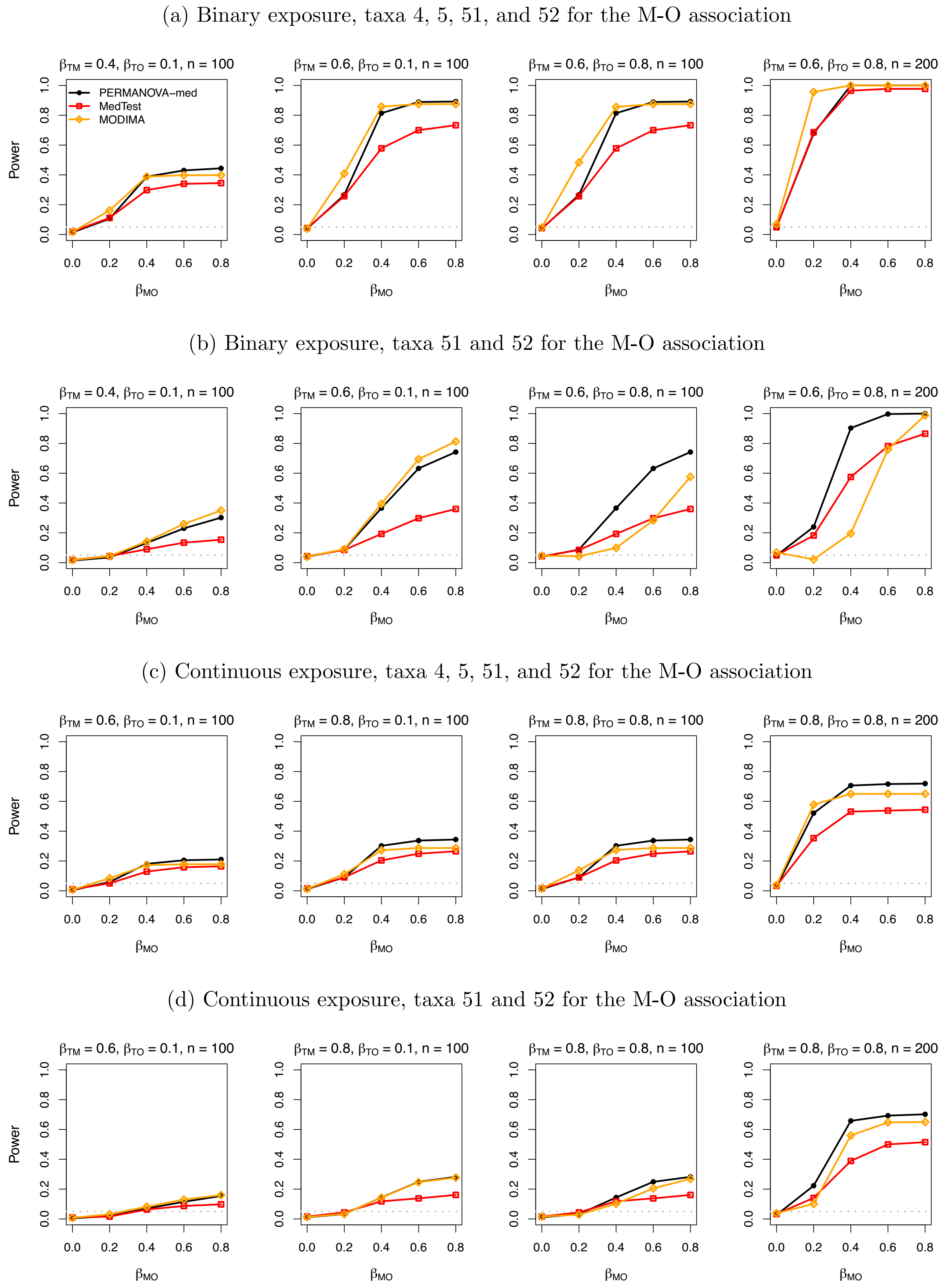
(a) Binary exposure, taxa 1–5 for the M-O association

(b) Binary exposure, taxa 4 and 5 for the M-O association

(c) Continuous exposure, taxa 1–5 for the M-O association

(d) Continuous exposure, taxa 4 and 5 for the M-O association

Figure 3.1: Simulation results in analysis of simulated data under M-common.



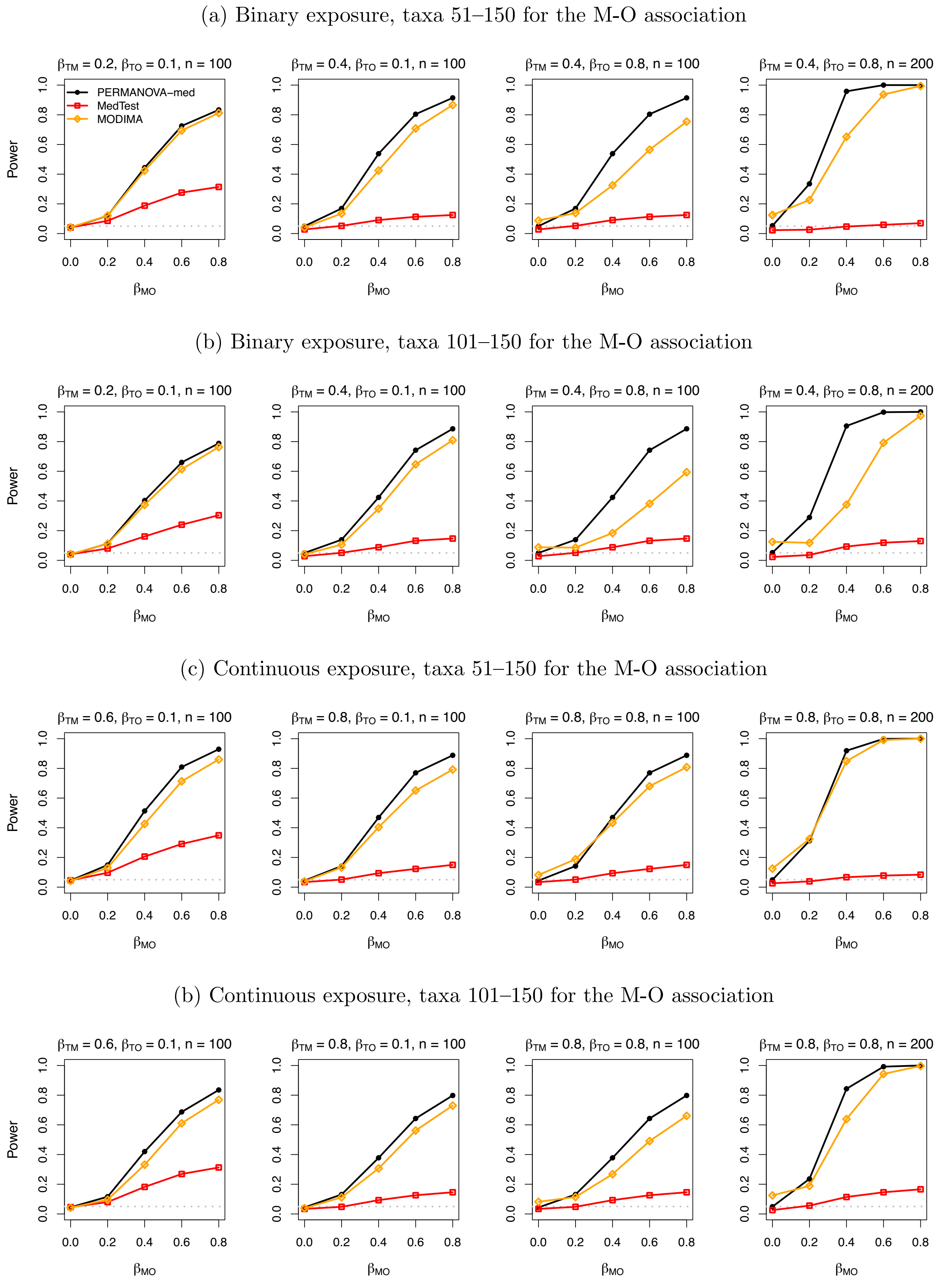
(a) Binary exposure, taxa 4, 5, 51, and 52 for the M-O association

(b) Binary exposure, taxa 51 and 52 for the M-O association

(c) Continuous exposure, taxa 4, 5, 51, and 52 for the M-O association

(d) Continuous exposure, taxa 51 and 52 for the M-O association

Figure 3.2: Simulation results in analysis of simulated data under M-mixed.


(a) Binary exposure, taxa 51–150 for the M-O association

(b) Binary exposure, taxa 101–150 for the M-O association

(c) Continuous exposure, taxa 51–150 for the M-O association

(b) Continuous exposure, taxa 101–150 for the M-O association

Figure 3.3: Simulation results in analysis of simulated data under M-rare.

Table 3.3: P-values from 16 mediation analyses of the data on melanoma immunotherapy response

			PERMANOVA-med			MedTest			MODIMA		
Outcome	Exposure	n	BC	J	Omni	BC	J	Omni	BC	J	Omni
No adjustment of covariates											
Progression-free survival	Fiber intake	89	0.808	0.965	0.958	-	-	-	-	-	-
	Probiotics	110	0.913	0.716	0.899	-	-	-	-	-	-
	Fiber + probiotics (4 levels)	89	0.777	0.975	0.953	-	-	-	-	-	-
	Sufficient fiber + no probiotics	89	0.717	0.965	0.910	-	-	-	-	-	-
Response to ICB	Fiber intake	94	0.727	0.955	0.903	0.624	0.636	0.837	0.384	0.935	-
	Probiotics	110	0.888	0.589	0.794	0.978	0.698	0.898	0.915	0.381	-
	Fiber + probiotics (4 levels)	89	0.620	0.980	0.827	-	-	-	0.430	0.947	-
	Sufficient fiber + no probiotics	89	0.490	0.955	0.697	0.276	0.626	0.455	0.441	0.947	-
Adjusting for BMI, prior treatment, LDH, stage											
Progression-free survival	Fiber intake	89	0.786	0.990	0.936	-	-	-	-	-	-
	Probiotics	110	0.983	0.788	0.947	-	-	-	-	-	-
	Fiber + probiotics (4 levels)	89	0.770	0.995	0.935	-	-	-	-	-	-
	Sufficient fiber + no probiotics	89	0.725	0.980	0.903	-	-	-	-	-	-
Response to ICB	Fiber intake	94	0.870	0.920	0.975	0.832	0.935	0.966	-	-	-
	Probiotics	110	0.973	0.433	0.630	0.911	0.539	0.773	-	-	-
	Fiber + probiotics (4 levels)	89	0.760	0.975	0.928	-	-	-	-	-	-
	Sufficient fiber + no probiotics	89	0.644	0.925	0.850	0.453	0.973	0.682	-	-	-

Note: BC: Bray-Curtis; J: Jaccard; Omni: the omnibus test that combines the results from analyzing the Bray-Curtis and Jaccard distances; n: sample size; $-$: not applicable.

level (LDH), and stage as potential confounders based on our analysis of bivariate associations, and we performed each mediation analysis with and without adjustment of these confounders. In all 16 mediation analyses, we applied PERMANOVA-med, MedTest, and MODIMA whenever they were applicable. For each method, we constructed tests based on the Bray-Curtis and Jaccard distance measures separately, as well as the omnibus test of both distance measures (except for MODIMA).

All results of p-values were summarized in Table 3.3. None of the p-values were significant at the 0.05 level, possibly due to the small sample sizes. Nevertheless, Table 3.3 demonstrated the wide applicability of PERMANOVA-med and the limited capabilities of MedTest and MODIMA. Specifically, neither MedTest nor MODIMA can handle censored survival times (the progression-free survival); MODIMA cannot adjust confounders (BMI et al.) nor provide an omnibus test (that combines Bray-Curtis and Jaccard); MedTest cannot handle multivariate exposures (the four-level categorical variable).

3.3 Remarks

We presented PERMANOVA-med, an extension of PERMANOVA to mediation analysis of microbiome data. Through extensive simulation studies, we observed that PERMANOVA-med did not uniformly outperform MedTest. However, the scenarios in which PERMANOVA-med did outperform seemed more realistic and more general, e.g., scenarios with a mixture of abundant and less abundant mediating taxa, relatively rare mediating taxa, or different sets of taxa associated with the exposure and the outcome. Even in the single scenario that PERMANOVA-med lost power to MedTest (Figure 3.1(a)), the power loss was relatively small. The power comparison between PERMANOVA-med and MODIMA was more difficult, as MODIMA often lost control of the type I error. Nevertheless, there were many more scenarios in which PERMANOVA-med had higher power than MODIMA than scenarios when it was the opposite.

The main advantage of PERMANOVA-med over MedTest and MODIMA is its wide applicability to a variety of mediation analyses of microbiome data, which was achieved by using our existing function permanovaFL. Through analysis of the simulated data and the real data, we have illustrated most features in Figure 1.2 that are supported by permanovaFL, such as multivariate exposures, survival outcomes, and omnibus tests of multiple distance measures. Although we did not cover clustered or matched-set data in this article, these types of data are emerging rapidly in recent years and may also call for mediation analysis. PERMANOVA-med is well positioned to accommodate such data in its current form. Further, PERMANOVA-med is not constrained to analysis of microbiome data but applicable to any high-dimensional data (e.g., genomic, epigenomic, metabolomic, proteomic, and cytokine data) that can be summarized into distance matrices. PERMANOVA-med has been added to the existing function “permanovaFL” in our R package LDM, which is available on GitHub at https://github.com/yijuanhu/LDM (accessed on 1 May 2022).

Caution is required in interpreting results from PERMANOVA-med (as well as MedTest and MODIMA). Strictly speaking, a significant p-value from PERMANOVA-med only means that the microbiome are associated with both the exposure and the exposure-adjusted outcome. External information on causal direction is needed to declare that the microbiome truly mediate the effect of the exposure on the outcome. Although the causal directions in the exposure–outcome and exposure–microbiome relationships may be evident in many cases, the causal direction between the microbiome and the outcome is often less clear because the change of microbiome may well be a consequence of the change of outcome rather than a cause.

PERMANOVA-med is limited to testing the mediation effect by the microbiome at the community level. Using the idea of inverse regression, we have also extended the LDM, called LDM-med, to testing microbiome mediation at both the community and individual taxon levels; some of those results mirror the results we obtained here. Aside from the capability of LDM-med to detect individual mediating taxa, a major difference between the two works is how we define the mediation effect by the microbiome. In the current work, we declare a mediation effect whenever the exposure perturbs some part of the microbial community and some part of the community influence the outcome; the two parts do not necessarily overlap (e.g., involving different taxa). This definition is reasonable here because, in distance-based analysis, a microbial community is viewed as a whole interconnected entity. The definition of microbiome mediation in LDM-med was more stringent. Because the main focus there was to detect individual taxa that act as mediators, only taxa that are first affected by the exposure and then influence the outcome were declared to be mediating taxa, and only a community that has mediating taxa in it was declared to have a global mediation effect. In practice, how to choose between PERMANOVA-med and LDM-med depends on what type of mediation is of most interest.

PERMANOVA-med is limited to testing the mediation effect by the microbiome at the community level. We have previously developed LDM-med to testing microbiome mediation at both the community and individual taxon levels; some of those results mirror the results we obtained here. Aside from the capability of LDM-med to detect individual mediating taxa, a major difference between the two works is how we define the mediation effect by the microbiome. In the current work, we declare a mediation effect whenever the exposure perturbs some part of the microbial community and some part of the community influence the outcome; the two parts do not necessarily overlap (e.g., involving different taxa). This definition is reasonable here because, in distance-based analysis, a microbial community is viewed as a whole interconnected entity. The definition of microbiome mediation in LDM-med was more stringent. Because the main focus there was to detect individual taxa that act as mediators, only taxa that are first affected by the exposure and then influence the outcome were declared to be mediating taxa, and only a community that has mediating taxa in it was declared to have a global mediation effect. In practice, how to choose between PERMANOVA-med and LDM-med depends on what type of mediation is of most interest.

Chapter 4

Topic 3: Integrative analysis of 16S marker-gene and shotgun metagenomic sequencing data improves efficiency of testing microbiome hypotheses

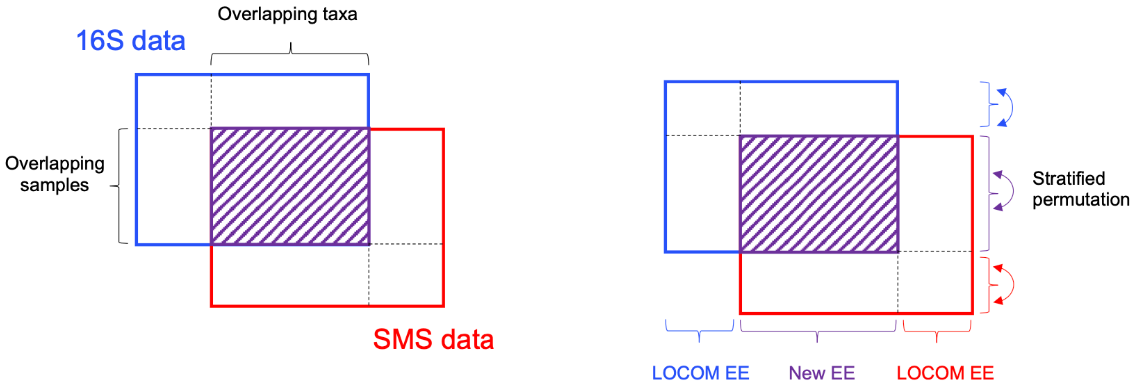


Figure 4.1: Illustration of the data structure (left) and strategies to analyze the data (right). The shaded area indicates the data for overlapping sample and taxa between the 16S and SMS taxa count tables. New EE is Equation (4.2) or (B3) in this manuscript. LOCOM EE is the Firth-corrected score equation in the LOCOM paper (Hu, Satten and Hu, 2022).

4.1 Method

4.1.1 Motivation

To motivate our method, we characterize the taxa count data generated from 16S and SMS for the same cohort of samples. First, we investigate the overlap of samples and genera between the two taxa count tables for the 26 studies found in Qiita. Table B1 shows that, in most cases, there is partial overlap of samples and genera between the two tables, although the sample overlap is typically substantial; this data structure is schematically depicted in Figure 4.1 (left). Note that the library sizes in the two tables can differ by orders of magnitude, ranging from 1.4 to 1500 fold. Then, we compare the observed relative abundances between 16S and SMS for overlapping samples at overlapping genera, using the ORIGINS data (more information about this dataset is provided in the Results section). Figures 4.2 and B1 demonstrate that there are systematic differences in observed relative abundances at many genera even for the second most abundant genus *Haemophilus*, and there are many zeros at some genera from either experiment, such as *Gemella* from 16S; both features can be explained by differential experimental bias in the two experiments. Besides the systematic differences at the taxon level, there are also random differences at the sample level due to variation in sample handling (Nearing et al., 2021), although the agreement tends to be higher for more abundant genera. In what follows, we first present methodologies for the data from overlapping samples and taxa, and then show how to accommodate the data from samples and taxa that are unique to only one experiment.

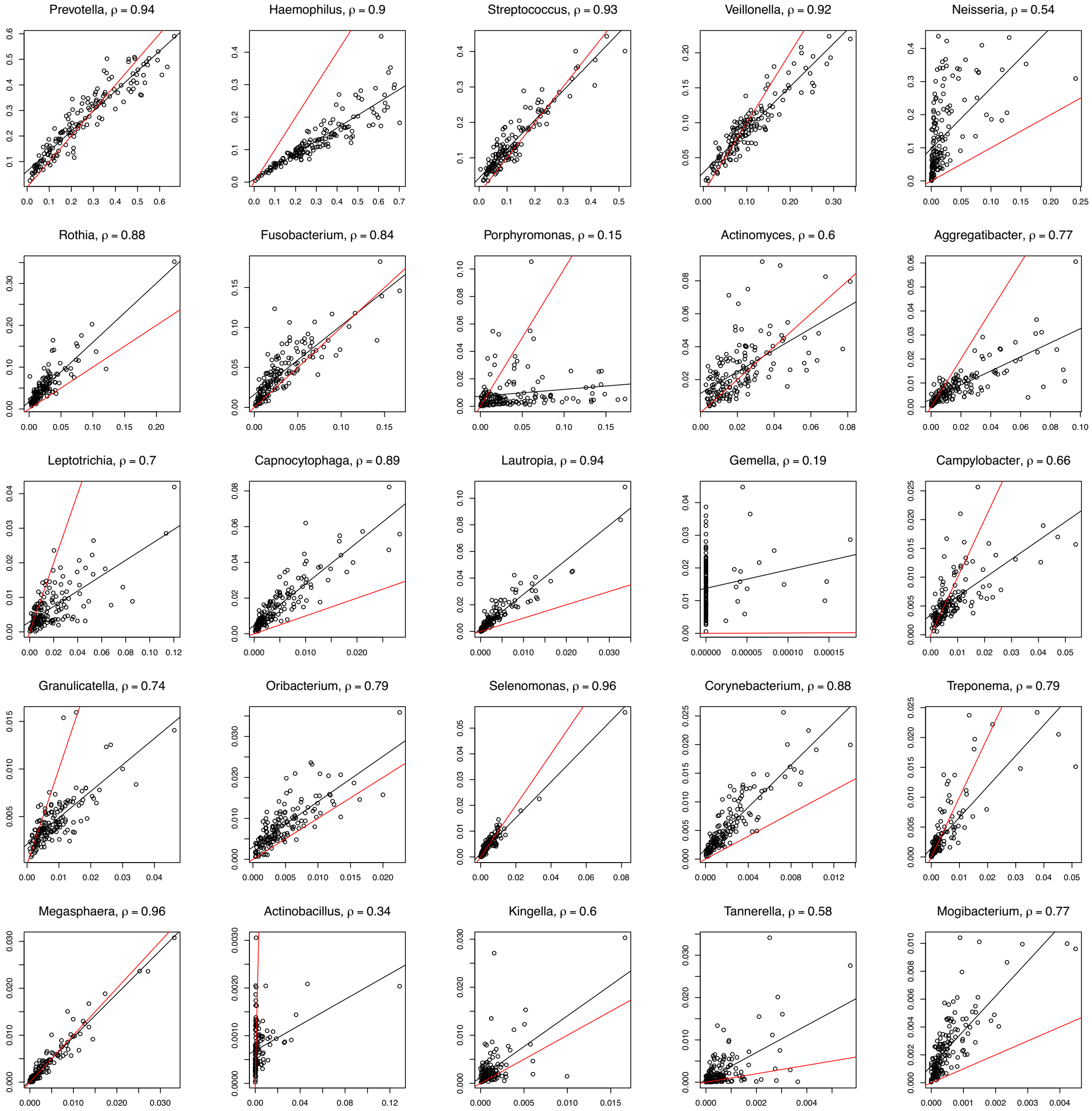


Figure 4.2: Scatter plot of observed relative abundances from 16S (x-axis) and SMS (y-axis) for the top 1–25 most abundant genera (ordered by decreasing abundance) in the ORIGINS data. The ρ value is the Pearson correlation coefficient. The red line is the 45° reference line. The black line depicts a fitted linear regression. The observed relative abundances are subcompositions among the 125 overlapping genera in the 152 overlapping samples.

4.1.2 Integrative analysis of 16S marker-gene and shotgun metagenomic se- quencing data

Let $Y_{ik,j}$ be the read count of the jth taxon ($j = 1, \ldots, J$) in the ith sample ($i = 1, \ldots, n$), generated from the kth experiment ($k = 1, 2$). Let Z_i be a vector of q covariates of the ith sample, which include 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. We consider the following log-linear model that relates the expected value of the observed relative abundances and true relative abundances, allowing experiment-specific bias factors:

$$\log(p_{ik,j}) = \mathbb{I}(k = 1)\gamma_{1,j} + \mathbb{I}(k = 2)\gamma_{2,j} + \log(\pi_{i,j}) + \alpha_{ik}, \tag{4.1}$$

where $\pi_{i,j}$ is the true relative abundance of taxon j in sample i irrespective of any experiment, $p_{ik,j}$ is the expected value of the observed relative abundance from the kth experiment, $\mathbb{I}(.)$ is the indicator function, $\gamma_{1,j}$ and $\gamma_{2,j}$ are taxon- and experiment-specific bias factors that describe how the observed relative abundance is distorted by experimental bias, and α_{ik} is the sample-specific normalization factor that ensures the composition constraint $\sum_{j=1}^{J} p_{ik,j} = 1$ for any i and k. We followed (Zhao and Satten, 2021) to introduce the effects of covariates Z_i on taxon j by replacing $\log(\pi_{i,j})$ with $\log(\pi_j^0) + Z_i^{\mathrm{T}}\beta_j$, where β_j contains the effect sizes and π_j^0 is the true relative abundance of taxon j when $Z_i = 0$. Like LOCOM, instead of fitting the multivariate logistic regression (4.1) to all taxa simultaneously, we fit individualized logistic regressions to each pair of taxa at a time. We choose one taxon (without loss of generality, the Jth taxon) that has the largest mean relative abundance (across both tables) to be the reference taxon, and compare all other taxa to the reference taxon using individualized logistic regressions. Because the most abundant taxa can always be effectively captured by any experiment, our selection criterion would result in a reference taxon that is among the top abundant taxa in each table. Specifically, we consider the subcomposition within taxa j and J: $\mu_{ik,j} = p_{ik,j}/(p_{ik,j} + p_{ik,J})$, and use (4.1) and $\log(\pi_{i,j}) = \log(\pi_j^0) + Z_i^{\mathrm{T}}\beta_j$ to obtain a standard logistic regression

$$\log\{\mu_{ik,j}/(1 - \mu_{ik,j})\} = \eta_{k,j} + Z_i^{\mathrm{T}}\beta_j$$

for each j ($j = 1, 2, \ldots, J - 1$) separately, where the intercept $\eta_{k,j} = \left[\log(\pi_j^0) - \log(\pi_J^0)\right] + \mathbb{I}(k = 1)(\gamma_{1,j} - \gamma_{1,J}) + \mathbb{I}(k = 2)(\gamma_{2,j} - \gamma_{2,J})$ but is considered as a free parameter without the need to distinguish the baseline relative abundances from the bias factors.

We propose to estimate $(\eta_{1,j}, \eta_{2,j}, \beta_j)$ by solving the estimating equation (EE)

$$\mathbb{U}_j(\eta_{1,j}, \eta_{2,j}, \beta_j) = \sum_{i=1,\ldots n, k=1,2} \omega_{ik,j} \left(\frac{Y_{ik,j}}{Y_{ik,j} + Y_{ik,J}} - \mu_{ik,j} \right) \begin{bmatrix} \mathbb{I}(k=1) \\ \mathbb{I}(k=2) \\ Z_i \end{bmatrix} = 0, \tag{4.2}$$

where $\omega_{ik,j}$ is the weight for the observation of sample i from experiment k. The EE of LOCOM (1.3) can be viewed as a special case of Equation (4.2) when $\omega_{ik,j}$ is set to $Y_{ik,j} + Y_{ik,J}$ and all observations are from one experiment. When $\omega_{ik,j} = Y_{ik,j} + Y_{ik,J}$, the summation in Equation (4.2) is at the "count" level, and observations with small $Y_{ik,j}$ and $Y_{ik,J}$ tend to contribute less. In fact, Equation (4.2) with these weights corresponds to score equations for read count data that follow the Binominal distribution (i.e., read assignments are independent of each other). Thus, these weights are optimal when there is minimal overdispersion in the count data (i.e., limited correlation between the read assignments). Additionally, these weights are reasonable for a taxon when the observations from one experiment have sparse count data (e.g., due to extraction or amplification inefficiency) and should contribute less than the observations from the other experiment, which yielded abundant count data.

On the other hand, taxa count data are notoriously overdispersed, and the amount of information quickly reaches a plateau as long as there is moderate coverage of reads. In this case, it is sensible to use $\omega_{ik,j} = 1$, which puts the summation in Equation (4.2) at the "relative abundance" level and treats the two observations of the same sample equally. These weights are particularly important when the library size of SMS is orders of magnitude (e.g., 10 times) higher than that of 16S (Table B1); Equation (4.2) would have been dominated by information from SMS if the count weights were used. Even when the overall library sizes are comparable between the two experiments, there could be substantial variation in read coverage at different taxa due to experimental bias. Since we have no a priori knowledge as for which weighting scheme works better for which taxa, we adopt the approach in which we first obtain results from tests with each weighting scheme separately for each taxon and then construct an omnibus test that combine these results.

The development of tests with a given weighting scheme generally follows the methodologies of LOCOM. We first estimate β_j by solving Equation (4.2) or the bias-corrected estimating equation (derived in Appendix B) when taxon j has sparse count data. Let $\beta_{j,1}$ be the effect size in β_j that corresponds to the trait of interest. Because we have no a priori knowledge about whether the reference taxon is null or causal (i.e., associated with the trait), we cannot test whether $\beta_{j,1} = 0$

directly for the hypothesis that taxon j is null. Instead, we make the assumption that more than half of the taxa are null taxa, which implies that $\text{median}_{j'=1,\ldots,J}\beta_{j,1}$ corresponds to $\beta_{j^*,1}$ for a certain null taxon j^*, and test whether $\beta_{j,1} - \text{median}_{j'=1,\ldots,J}\beta_{j,1} = 0$. To this end, we adopt the median-adjusted estimate of $\beta_{j,1}$ as the test statistic, and calculate the p-value using Potter's permutation scheme for logistic regression (Potter, 2005), which is based on shuffling the trait residuals after regressing out the other covariates. To preserve the correlations between the 16S and SMS observations from the same samples in the permutation replicates, we shuffle as a whole the trait residuals for each pair of 16S and SMS observations, which have identical values as the pair of observations share the same covariate values. The taxa with adjusted p-values, calculated using the Benjamini-Hockberg procedure (Benjamini and Hochberg, 1995) for correcting multiple comparisons, below the nominal FDR level are declared as differentially abundant. For testing the global hypothesis that there are no differentially abundant taxa in the community, we adopt the Harmonic Mean (HM) of the taxon-specific p-values as the test statistic and assess the significance using the existing permutation replicates. The two test methods (each includes the taxon-level tests and a global test) with the count weights and equal weights are referred to as New-count and New-equal, respectively. Finally, we construct an omnibus test for each taxon that uses the smaller p-value of the two tests as the test statistic, thus allowing the optimal weighting scheme to be selected at each taxon, and we assess its significance using the existing permutation replicates; the omnibus global test is based on the HM of the taxon-specific omnibus p-values. We refer to the test method as New-omni.

Next, we extend our methodologies to accommodate the data for samples and taxa that are unique to one experiment. For an overlapping taxon, when there are samples sequenced by only one experiment, we make a small modification to Equation (4.2) to allow this sample to contribute to the summations accordingly. For a non-overlapping taxon, we use the LOCOM EE to estimate β_j. In fact, when a taxon is identified by both experiments but has increasing zero counts in one experiment, Equation (4.2) applied to this taxon converges to the LOCOM EE (with weights $\omega_{ik,j}$). This implies that the non-overlapping taxon can be considered as a special case of an overlapping taxon with all zero data from one experiment. Then, the estimates of β_j from all (overlapping and non-overlapping) taxa are pooled to calculate the median and the test statistics. Care should be taken to preserve the sample structure when generating permutation replicates. The permutation procedure should be stratified to restrict the shuffling of trait residuals in one stratum of samples that are sequenced in both experiments, one stratum of samples sequenced in one experiment only, and one stratum sequenced in the other experiment only. This permutation scheme is illustrated in Figure 4.1 (right).

Finally, we develop a filter to exclude very rare taxa that can jeopardize the validity of our method. Note that the filter in LOCOM removes rare taxa present in fewer than 20% of samples. We continue to apply this filter to non-overlapping taxa that utilize the LOCOM EE. For overlapping taxa, a sample with a non-zero count from either experiment (i.e., based on the pooled read count) contributes to the 3rd equation of (4.2), making it reasonable to retain the taxon if more than 20% of such samples exist. Moreover, we keep the taxon as long as it passes the LOCOM filter applied to either the 16S part of the data or the SMS part of the data, since Equation (4.2) reduces to the LOCOM EE in the worse-case scenario when the data from one experiment consist entirely of zeros.

As benchmarks, we consider two alternative approaches to integrative analysis. One naive approach is to first create a "pooled" taxa count table consisting of the union of samples and the union of taxa, pool the reads for overlapping samples and taxa, and fill in zeros for taxa unidentified by an experiment and for samples only sequenced by that experiment. Then, LOCOM can be applied to the pooled table. We refer to this approach as Com-count. The zero-filling strategy may create spurious associations for non-overlapping taxa when samples filled with zeros have a different distribution of trait values from other samples. Moreover, Com-count shares the same drawbacks as New-count in cases when the count weights are suboptimal, because Com-count can be equivalently obtained from New-count by forcing the use of a common intercept and restricting analysis to the data for overlapping samples and taxa. Lastly, the LOCOM filter applied to the pooled table is more stringent than the new filter above, i.e., retaining fewer taxa for analysis.

Another approach is to apply LOCOM separately to the two taxa count tables and then combine the two p-values for overlapping taxa into a single p-value using a p-value combination method. The resulting p-values, along with directly output p-values from LOCOM for non-overlapping taxa, are used to detect differentially abundant taxa and also combined into one p-value for testing the global hypothesis. At each taxon, the two p-values from the same sample are expected to exhibit a strong (positive) correlation; at the global level, the p-values across interacting taxa may also be correlated. Therefore, we opt for the Cauchy (Liu and Xie, 2020) or HM (Wilson, 2019) combination method, which accounts for such correlations, and we use the same method for both levels of p-value combination. This integrative analysis approach is referred to as Com-p-C or Com-p-HM. Note that the HM and Cauchy methods generate p-values based on asymptotic theories, while our new method assesses the significance of the HM statistics via permutation. Com-p-C and Com-p-HM are inherently unable to produce an overall p-value that is more significant than the most significant member p-value, whereas the new methods can. Additionally, Com-p-C and Com-p-HM combine p-values without considering the direction of association in each dataset, whereas the new methods tend to strengthen the signal if it is consistent across both datasets and disregard the signal if it is contradictory. Furthermore, taxa that fail to pass the LOCOM filter based on either taxa count table are completely missed by this approach, but they still have a good chance of passing the new filter based on the pooled data, especially when the sample overlap is substantial.

4.2 Numerical Studies

4.2.1 Simulation studies

Our simulations are based on data on 856 taxa of the upper-respiratory-tract (URT) microbiome by Charlson et al. (Charlson et al., 2010). We fixed the sample size to 100 unless otherwise specified and considered a binary trait T_i throughout the simulations. In some cases, we also simulated a continuous confounder C_i by drawing values from $U[-1,1]$ for samples with $T_i = 0$ and from $U[0,2]$ for those with $T_i = 1$. We used the two sets of causal taxa (i.e., taxa that are associated with the trait) employed in (Hu, Satten and Hu, 2022), namely, a random sample of 20 taxa with mean relative abundances greater than 0.005, as observed in the URT data, and the five most abundant taxa; we refer to these sets as M1 and M2, respectively. While M1 and M2 comprise moderately abundant and very abundant causal taxa, respectively, we also considered a set of rare causal taxa by randomly sampling 50 taxa with mean relative abundances between 0.0005 and 0.001; we refer to this set as M3. When a confounder was present, we randomly sampled 5 taxa with mean relative abundances greater than 0.005 to be associated with the confounder.

We assumed that the 856 taxa form the complete set of underlying taxa in the community and generated bias factors $\gamma_{1,j}$ and $\gamma_{2,j}$ for 16S and SMS, respectively. We set $\gamma_{1,j}$ and $\gamma_{2,j}$ to a very small value of -5 to create missingness for specific taxa in each experiment. Specifically, in M1 and M3, we selected two sets of five non-overlapping causal taxa to be missing in 16S and SMS. In M2, the two sets included two taxa each. Additionally, we sampled 20% of non-causal taxa to be missing in each experiment. We set $\gamma_{1,j}$ and $\gamma_{2,j}$ for the most abundant taxon j to 1, reflecting its efficient capture by both experiments. For all other taxa, we independently drew $\gamma_{1,j}$ and $\gamma_{2,j}$ from $N(0, 0.5^2)$.

We then simulated read count data for the 856 taxa across two experiments, taking into account the effects of the trait and confounder as well as the influences of bias factors. First, we drew the baseline relative abundances $(\pi_{i,1}^{(0)}, \pi_{i,2}^{(0)}, \ldots, \pi_{i,J}^{(0)})$ of all taxa for each sample from the Dirichlet distribution $Dirichlet(\bar{\pi}, \theta)$, where $\bar{\pi}$ contains the mean relative abundances and θ is the overdispersion parameter estimated from fitting the Dirichlet-Multinomial (DM) model to the URT data. Notably, the parameter θ controls sample heterogeneity in baseline relative abundances, excluding the overdispersion in the process of generating count data; thus, we set θ to 0.01, which is half of the total overdispersion (0.02) estimated from the URT data. Then, we formed the expected value of the observed relative abundances obtained from the kth experiment, $p_{ik,j}$, by spiking in the causal taxa and confounder-associated taxa, then imposing bias factors on all taxa, and finally normalizing the relative abundances to have a sum of 1, resulting in the following equation:

$$p_{ik,j} = \frac{\exp\left[\mathbb{I}(k=1)\gamma_{1,j} + \mathbb{I}(k=2)\gamma_{2,j} + \beta_{1,j}T_i + \beta_{2,j}C_i\right]\pi_{ij}^{(0)}}{\sum_{j'=1}^{J}\exp\left[\mathbb{I}(k=1)\gamma_{1,j'} + \mathbb{I}(k=2)\gamma_{2,j'} + \beta_{1,j'}T_i + \beta_{2,j'}C_i\right]\pi_{ij'}^{(0)}}. \tag{4.3}$$

Here $\beta_{1,j} = 0$ for null taxa, and $\beta_{2,j} = 0$ for confounder-independent taxa. For simplicity, we set $\beta_{1,j} = \beta$ for all causal taxa, which is referred to as the *effect size*, and fixed $\beta_{2,j} = \log(1.5)$ for all confounder-associated taxa. Subsequently, we generated the read count data for sample i obtained from the kth experiment using the DM model with mean $(p_{ik,1}, p_{ik,2}, \ldots, p_{ik,J})$, overdispersion parameter τ_k, and library size drawn from $N(\nu_1, (\nu_1/3)^2)$ and $N(\nu_2, (\nu_2/3)^2)$ for 16S and SMS, respectively, with left truncation at 2,000. We fixed $\nu_1 = 10,000$ and varied ν_2 to achieve the depth ratio ν_1:ν_2 = 1:1 or 1:10. The parameter τ_k controls the extend to which the observed relative abundances from the kth experiment deviate from their expected values $p_{ik,j}$. We set $\tau_1 = \tau_2 = \tau$ without loss of generality and varied τ between 0.01 and 0.001 corresponding to large and small deviation.

We began with the complete-overlap case, where data from both experiments were collected for all 100 samples. We then moved to a partial-overlap case by collecting data from both experiments for 40 samples (15 cases and 25 controls), from 16S only for 40 samples (30 cases and 10 controls), and from SMS only for 20 samples (5 cases and 15 controls), which resulted in a dataset with a total of 100 samples, 50 cases and 50 controls, and varying case-control ratios across the three strata of samples. We applied New-omni, New-equal, New-count, Com-count, Com-p-C, and Com-p-HM for the integrative analysis of 16S and SMS data, using LOCOM to analyze each dataset separately as a reference. In the partial-overlap case, we also compared the proposed permutation scheme based on three strata of samples for New-omni, New-equal, and New-count to two alternative schemes: one that combined samples sequenced by only one experiment into a single stratum (two strata), and another that pooled all samples together (one stratum). We evaluated the sensitivity and empirical FDR of each method for testing individual taxa at the nominal FDR level of 20%, as well as type I error and power for testing the global association at the nominal level of 0.05. The type I error results were based on 10,000 replicates of simulated data, while all other results were derived from 1,000 replicates.

4.2.2 Simulation results

To confirm that the simulated data captured the important features of the ORIGINS data shown in Figures 4.2 and B1, we presented similar scatter plots for the simulated data based on the sample size ($n = 152$) of the ORIGINS data and $\tau = 0.001$ in Figures B2 and B3. From the simulated data, we observed that the range of observed relative abundances at each taxon as governed by the overdispersion parameter θ, the deviation of the fitted line from the 45° reference line as determined by the bias factors $\gamma_{1,j}$ and $\gamma_{2,j}$, and the agreement of individual data points along the fitted line as impacted by the overdispersion parameter τ, all resemble those of the real data. Similarly, we presented scatter plots for the simulated data based on the sample size ($n = 76$) of the dietary data and $\tau = 0.01$ in Figures B4 and B5, and found patterns comparable to those in the dietary data shown in Figures B6 and B7.

We concentrate on the results for the complete-overlap case, which provide the most information on method comparison. The type I error results (at $\beta = 0$) from the global tests are shown in Figure 4.3 and B8, while the global power, sensitivity, and FDR for testing individual taxa across various values of β are displayed in Figures 4.4, B9–B11. As expected, New-count and Com-count produced very similar results throughout, so we will focus on the results of New-count in the following comparison. All methods, except Com-p-HM, controlled the type I error at the nominal level, whereas Com-p-HM yielded inflated type I error in every scenario. This is not surprising, as Com-p-HM as well as Com-p-C rely on asymptotic theories, while all other methods are based on permutation. Due to its inflated type I error, Com-p-HM was excluded in the figures for power comparison. In addition, Com-p-HM demonstrated similar sensitivity and FDR as Com-p-C when testing individual taxa, so we will focus on the results of Com-p-C only.

All integrative analyses showed a significant improvement in efficiency for testing hypotheses about the microbiome compared to analyses of a single dataset, at both the taxon and global levels. The increase in sensitivity of detecting causal taxa is substantial, as each experiment failed to capture certain causal taxa but the combination of both experiments provided better coverage. The boost in global power is less pronounced, because both the HM and Cauchy statistics are dominated by a few of the smallest p-values and less influenced by the total number of causal taxa. In all cases, New-omni exhibited the highest or nearly highest power and sensitivity.

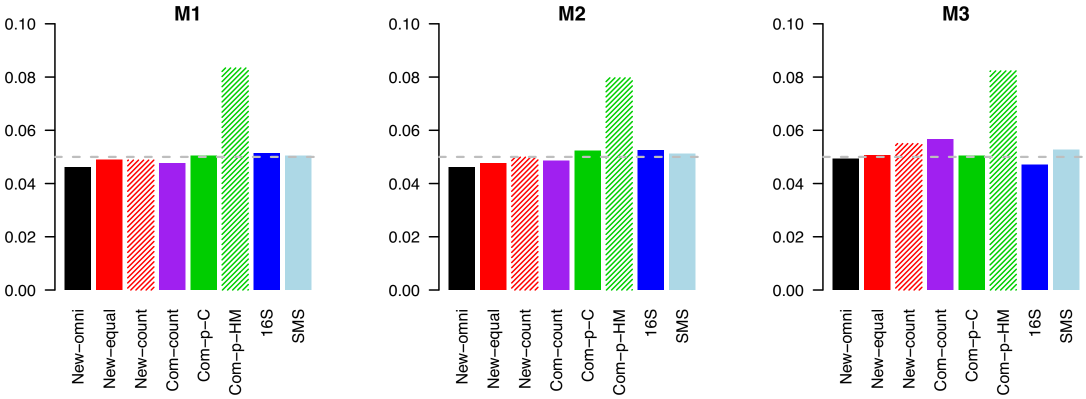


Figure 4.3: Type I error rate (y-axis) of global tests at the nominal level of 0.05 (gray dashed line), based on data simulated with completely overlapping samples, an overdispersion of $\tau = 0.01$, and a depth ratio of 1:10.

In our detailed comparison of various methods for integrative analysis, we will focus on sensitivity, as power differences are generally small. We begin by confirming that all methods controlled the FDR at the nominal level across all scenarios. Among all integrative analyses, New-omni consistently achieved the highest or nearly highest sensitivity in all cases, which was driven at times by New-equal and at other times by New-count. Specifically, New-equal had higher sensitivity than New-count in M1 and M2 with an overdispersion of $\tau = 0.01$ and a depth ratio of 1:10, which was expected as the large overdispersion encouraged observations to be weighted equally, regardless of a 10-fold difference in coverage. New-count had higher sensitivity than New-equal in M3 because, in cases of sparse coverage of reads, every read mattered. Com-p-C lost sensitivity to New-omni in M3 with $\tau = 0.01$ because, in the presence of the large overdispersion, some causal taxa failed the LOCOM filter based on any single dataset but passed the new filter based on pooled data. The sensitivity loss disappeared when τ was decreased to 0.001. Indeed, there were approximately 5 causal taxa missed by Com-p-C in a typical replicate of data with $\tau = 0.01$ and only around one when $\tau = 0.001$. Com-p-C also lost some sensitivity to New-omni in M2 when $\tau = 0.001$.

The same patterns of results persisted when a confounder was simulated (Figure B12). We confirmed that the confounding effect was substantial, leading to highly inflated type I error if it was not controlled for (Figure B13, upper panel). All methods (except for Com-p-HM) yielded proper type I error after adjusting for the confounder (Figure B13, lower panel). Results for the partial-overlap case are displayed in Figures 4.5 and 4.6. In this general case, Com-count failed to control the type I error. The permutation scheme based on three strata of samples is the only one among all schemes that led to the correct type I error.

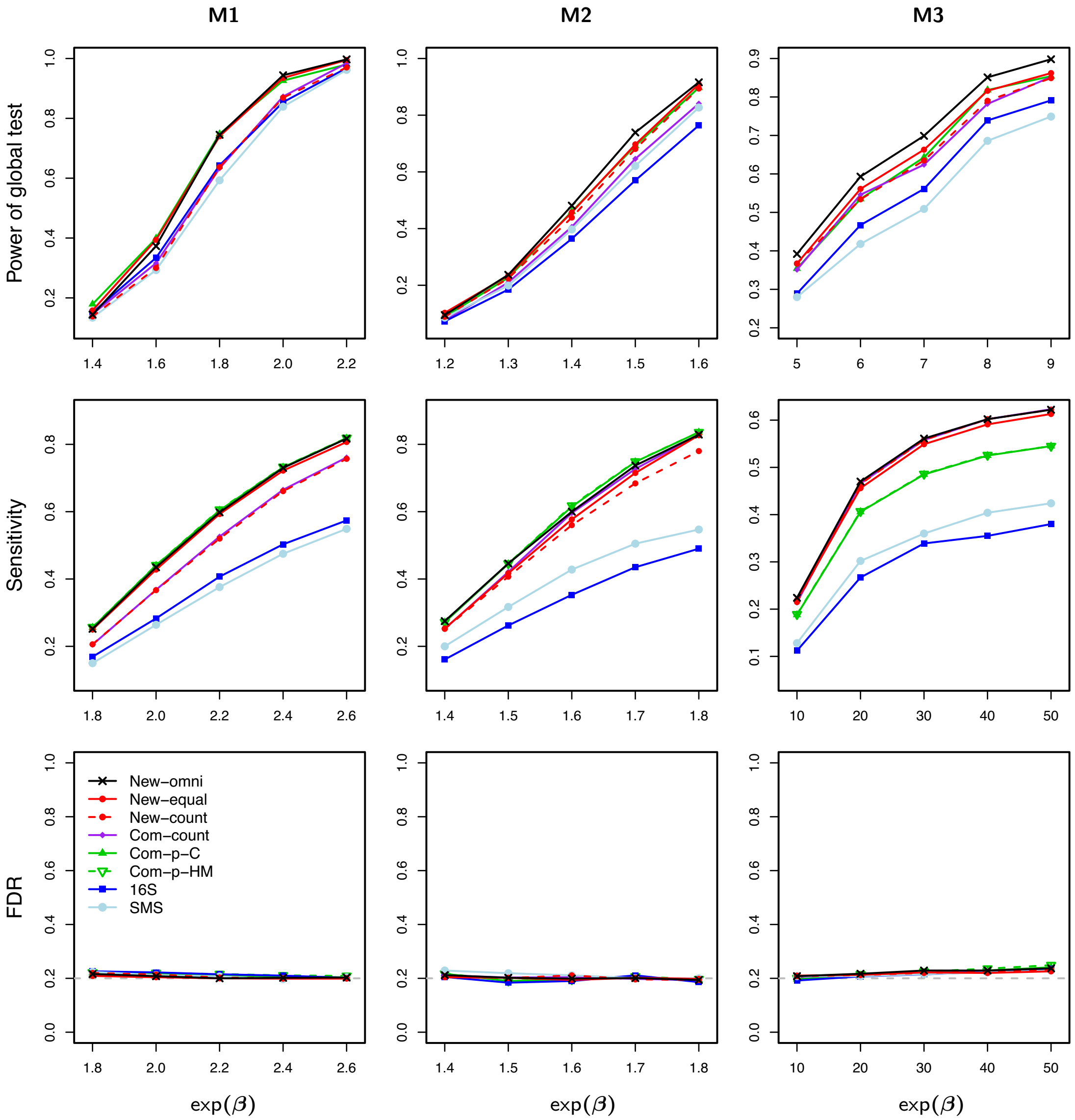


Figure 4.4: Power (upper panel) for testing the global association and sensitivity (middle panel) and empirical FDR (lower panel) for testing individual taxa, based on data simulated with completely overlapping samples, an overdispersion of $\tau = 0.01$, and a depth ratio of 1:10.

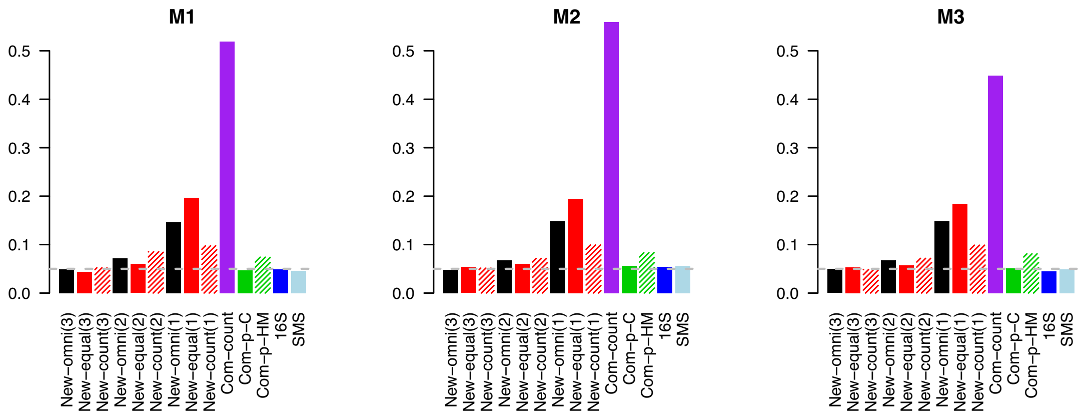


Figure 4.5: Type I error rate (y-axis) of global tests at the nominal level of 0.05 (gray dashed line), based on data simulated with partially overlapping samples, an overdispersion of $\tau = 0.01$, and a depth ratio of 1:10. The methods annotated with “(3)”, “(2)” and “(1)” used permutation schemes that are based on 3 strata, 2 strata, and 1 stratum, respectively.

4.2.3 ORIGINS data

We analyzed data generated from the Oral Infections, Glucose Intolerance, and Insulin Resistance Study (ORIGINS) (Demmer et al., 2015) to investigate the association between periodontal bacteria and prediabetes status (yes or no) among diabetes-free adults, without adjusting for any other risk factors. One subgingival plaque sample was collected from each participant and sequenced by either 16S, SMS, or both. We downloaded both 16S and SMS taxa count tables with study ID 11808 from Qiita. The 16S table contains 271 samples linked with meta data and having adequate (i.e., greater than 5,000) library sizes, as well as 234 genera after quality control (QC)(i.e., excluding genera found in less than 5 samples). The SMS table includes 183 samples linked with meta data and having adequate library size, as well as 756 genera after QC. In total, there are 302 distinct samples (56 cases and 246 controls) and 864 distinct genera, among which 152 samples (47 cases and 105 controls) and 125 genera have both 16S and SMS data. The mean library sizes generated from 16S and SMS are 26,950 and 176,321, respectively, resulting in a depth ratio of 1:6.5.

We began by analyzing the data for overlapping samples and genera to compare different methods. We applied the same methods (except for Com-p-HM) as in the simulation studies and summarized their global p-values and detected genera at the nominal FDR level of 10% in Table 4.1 (upper panel). At the nominal level of 0.05, the analysis of a single 16S or SMS dataset, as well as Com-p-C, produced non-significant global p-values, whereas New-omni, New-equal, New-count, and Com-count yielded significant ones. The genera detected by New-omni encompassed all those identified by either New-equal, New-count, 16S or Com-count, with the exception of *Butyrivibrio*. Both the

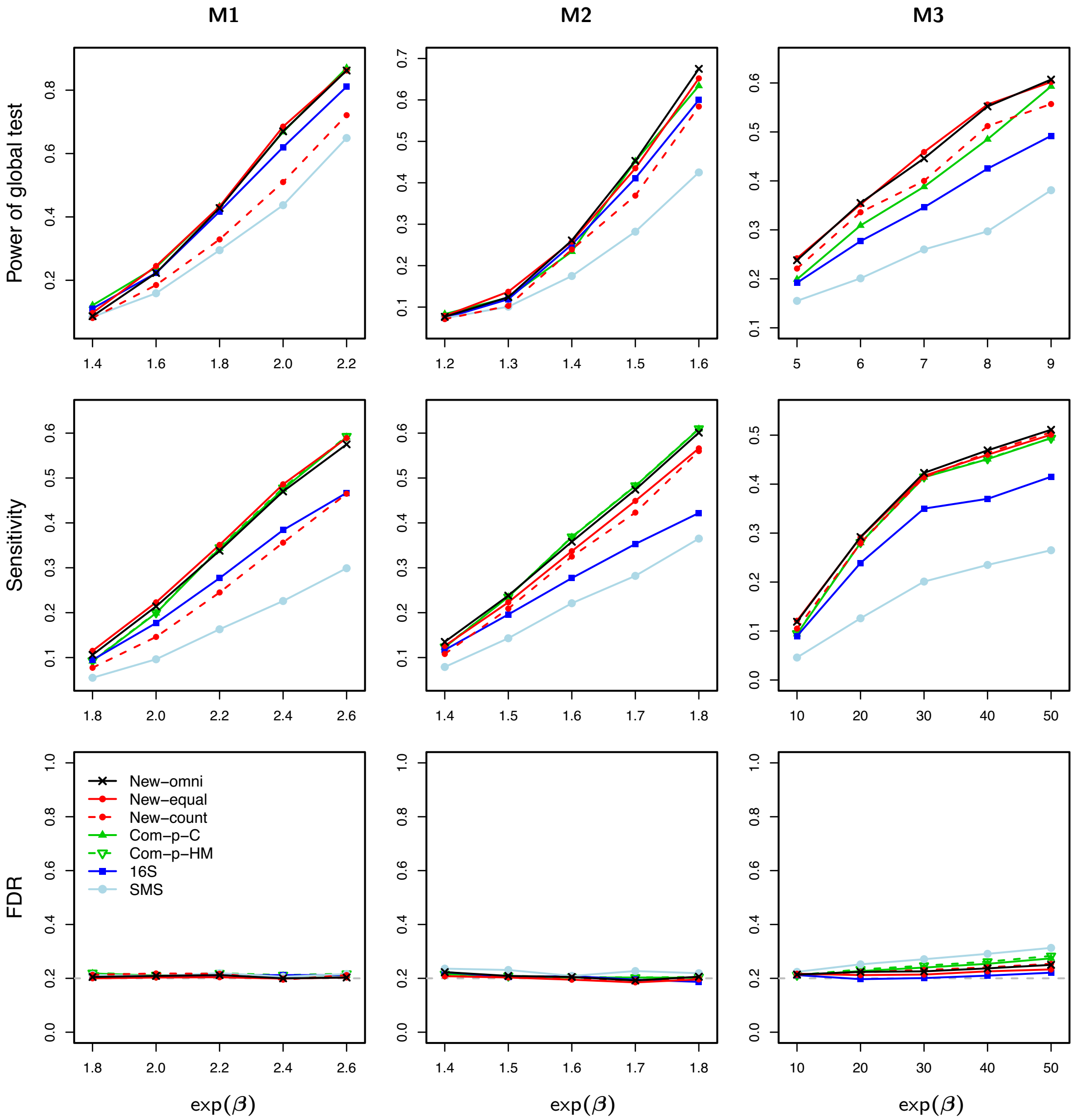


Figure 4.6: Power (upper panel) for testing the global association and sensitivity (middle panel) and empirical FDR (lower panel) for testing individual taxa, based on data simulated with partially overlapping samples, an overdispersion of $\tau = 0.01$, a depth ratio of 1:10.

analysis of the SMS data alone and Com-p-C were unable to detect any genera. For more detailed results of the detected genera, e.g., p-values, adjusted p-values, and observed relative abundances, refer to Table B2 and Figures B14–B15. Overall, the relative performance of these methods in this context aligns with their performance in the simulation studies.

We proceeded to analyze the full data for scientific discovery and summarized the results in Tables 4.1 (lower panel). Once again, New-omni, New-equal, New-count yielded significant global p-values, while the other methods did not. New-omni detected all three genera, *Butyrivibrio*, *Gemella*, and *Ignavigranum*, that were detected by either New-equal or New-count, whereas the analysis of the SMS data alone and Com-p-C failed to detect any genera. Note that discrepancies in the list of detected genera are expected, considering the large number of genera found unique to a single dataset. Further details about the detected genera can be found in Table B3 and Figure B16.

Specifically, *Butyrivibrio* was captured by both sequencing platforms and assigned non-significant, yet small, p-values in the analyses of individual datasets. As its abundance consistently appeared lower in prediabetic participants across both datasets, this trend was deemed significant by New-omni. Indeed, *Butyrivibrio* is known to produce short-chain fatty acids such as butyrate, which has been found to improve insulin sensitivity in mice (Gao et al., 2009). *Gemella* was not effectively captured by 16S sequencing and was subsequently excluded from the analysis of the 16S data. Nevertheless, the SMS data revealed that it was fairly abundant (1.5% on average) and significantly more abundant in prediabetic participants, resulting in its detection by New-omni. This finding is plausible, as *Gemella* have been linked to various infections, including those affecting heart valves (La Scola and Raoult, 1998), brain membranes (Ruoff, 2002), and bloodstreams (Woo et al., 2003). *Ignavigranum* was completely missed by 16S sequencing and its detection by New-omni was solely driven by its differential abundance in the SMS dataset. *Chelonobacter* exhibited significant differential abundance in the 16S dataset, but this signal was not replicated in the SMS dataset, leading to a non-significant result by New-omni.

4.2.4 Dietary data

We also analyzed data from Amato et al. (Amato et al., 2019) to test the influence of host dietary niche (folivore vs. non-folivore) on the gut microbiome of wild non-human primates, while controlling for host phylogeny (categorized as ages, lemurs, new world monkeys, and old world monkeys). One fecal sample was collected for each animal and sequenced by either 16S, SMS, or both. We downloaded both 16S and SMS taxa count tables with study ID 11212 from Qiita, the features

Table 4.1: Global test p-value and detected differentially abundant genera in the analysis of the ORIGINS data

Method	Global p-value	Detected genera
Data for overlapping samples and genera		
New-omni	0.0345	*Pseudoalteromonas*, *Actinomyces*, *Capnocytophaga*, *Lonepinella* *Campylobacter*, *Gemella*, *Kocuria*
New-equal	0.0436	*Pseudoalteromonas*, *Actinomyces*, *Capnocytophaga*, *Lonepinella*
New-count	0.0261	*Campylobacter*,*Gemella*, *Kocuria*
Com-count	0.0226	*Campylobacter*,*Gemella*, *Butyrivibrio*
Com-p-C	0.106	None
16S	0.0689	*Campylobacter*
SMS	0.170	None
Full data		
New-omni	0.0340	*Ignavigranum*, *Gemella*, *Butyrivibrio*
New-equal	0.0262	*Ignavigranum*, *Butyrivibrio*
New-count	0.0388	*Ignavigranum*, *Gemella*
Com-p-C	0.245	None
16S	0.0682	*Chelonobacter*
SMS	0.0866	None

Note: Com-count is invalid for analyzing data from partially overlapping samples and thus not applied to the full data. The nominal FDR level is 10%.

Table 4.2: Global test p-value and detected differentially abundant genera in the analysis of the dietary data

	Data for overlapping samples and genera		Full data	
Method	Global p-value	Detected genera	Global p-value	Detected genera
New-omni	0.0001	54	0.0003	353
New-equal	0.0001	46	0.0001	331
New-count	0.0001	30	0.0001	270
Com-count	0.0006	27	NA	NA
Com-p-C	0.00213	36	0.00196	195
16S	0.0001	24	0.0001	57
SMS	0.0002	23	0.0003	154

Note: See Note in Table 4.1.

of which are listed in Table B1. In particular, there are 172 distinct samples (94 folivore and 78 non-folivore) and 2062 distinct genera, among which 76 samples (40 folivore and 36 non-folivore) and 236 genera have data from both 16S and SMS. The ratio of 16S to SMS mean library sizes is 1:9.8.

Table 4.2 shows that all methods yielded highly significant global p-values and detected a large number of genera with differential abundance between folivore and non-folivore animals at the nominal FDR level of 10%, in both analyses of the data for overlapping samples and genera and the full data. In both cases, New-omni detected the most genera, exceeding the total detections by both 16S and SMS. Figure B17 displays the Venn diagrams of detected genera by various methods using the full data. The genera detected by New-omni nearly encompass the union of those detected by New-count and New-equal. New-omni detected 185 new genera that were missed by both 16S and SMS, while Com-p-C detected only 8 new genera.

4.3 Remarks

Integrative analysis of 16S and SMS data is a new problem that no one has addressed before. In this article, we have presented and compared several approaches to this problem, and New-omni consistently performed the best or nearly the best in all scenarios. Therefore, we choose New-omni as the preferred method and name it LOCOM-I. LOCOM-I inherits many features from LOCOM. The inference is based on permutation and thus valid for small sample sizes. It allows testing a trait that is binary, continuous, or multivariate (e.g., a categorical trait with more than two levels), permits testing of multiple traits simultaneously, and supports adjustment of confounding covariates.

It is worth noting that New-equal depends entirely on relative abundance data without need to know the original read count data. This is an important feature because some bioinformatics programs for processing shotgun metagenomic data, such as Kraken, output relative abundance data only. In this case, New-equal is applicable while New-count is not, and New-omni should take the results from New-equal.

We have implemented LOCOM-I in the existing R package LOCOM, which is available on GitHub at https://github.com/yijuanhu/LOCOM.

Appendices

Appendix A

Topic 1

Text A1

Because O_r is orthogonal to (Z, T), it follows from the inverse regression model (2.3) that

$$\mathrm{E}(M_j|Z,T) = \mathrm{E}_{O_r}\left\{\mathrm{E}(M_j|Z,T,O_r)\right\} = \beta_{0,j} + \beta_{\mathrm{Z},j}^{\mathrm{T}} Z + \beta_{1,j} T_r,$$

which is the mediator model (2.1) except that T_r is used in place of T. Since T_r is the residual of T after orthogonalizing against Z, the coefficients for T and T_r should be the same, i.e., $\beta_{1,j} = \alpha_{1,j}$.

If we assume that the mediators $M_1, M_2, \ldots, M_J$ are independent of each other conditional on (Z, T), then from the forward outcome model (2.2) that models the joint effects of all mediators, we obtain the forward outcome model that models the marginal effect of mediator M_j:

$$\mathrm{E}(O|Z,T,M_j) = \theta_0 + \theta_{\mathrm{Z}}^{\mathrm{T}} Z + \theta_1 T + \theta_{2,j} M_j. \tag{A1}$$

Comparing (A1) with the inverse regression model (2.3), we find that the positions of O (or O_r) and M_j are exchanged and it is well known that $\beta_{2,j} \neq \theta_{2,j}$ in this case. However, both $\beta_{2,j}$ in (2.3) and $\theta_{2,j}$ in (A1) capture the association between O and M_j conditional on (Z, T), so $\beta_{2,j} = 0$ and $\theta_{2,j} = 0$ coincide. This result easily extends to cases when the mediators $M_1, M_2, \ldots, M_J$ are correlated, because our approach focuses on testing *marginal* mediation effects instead of *conditional* mediation effects.

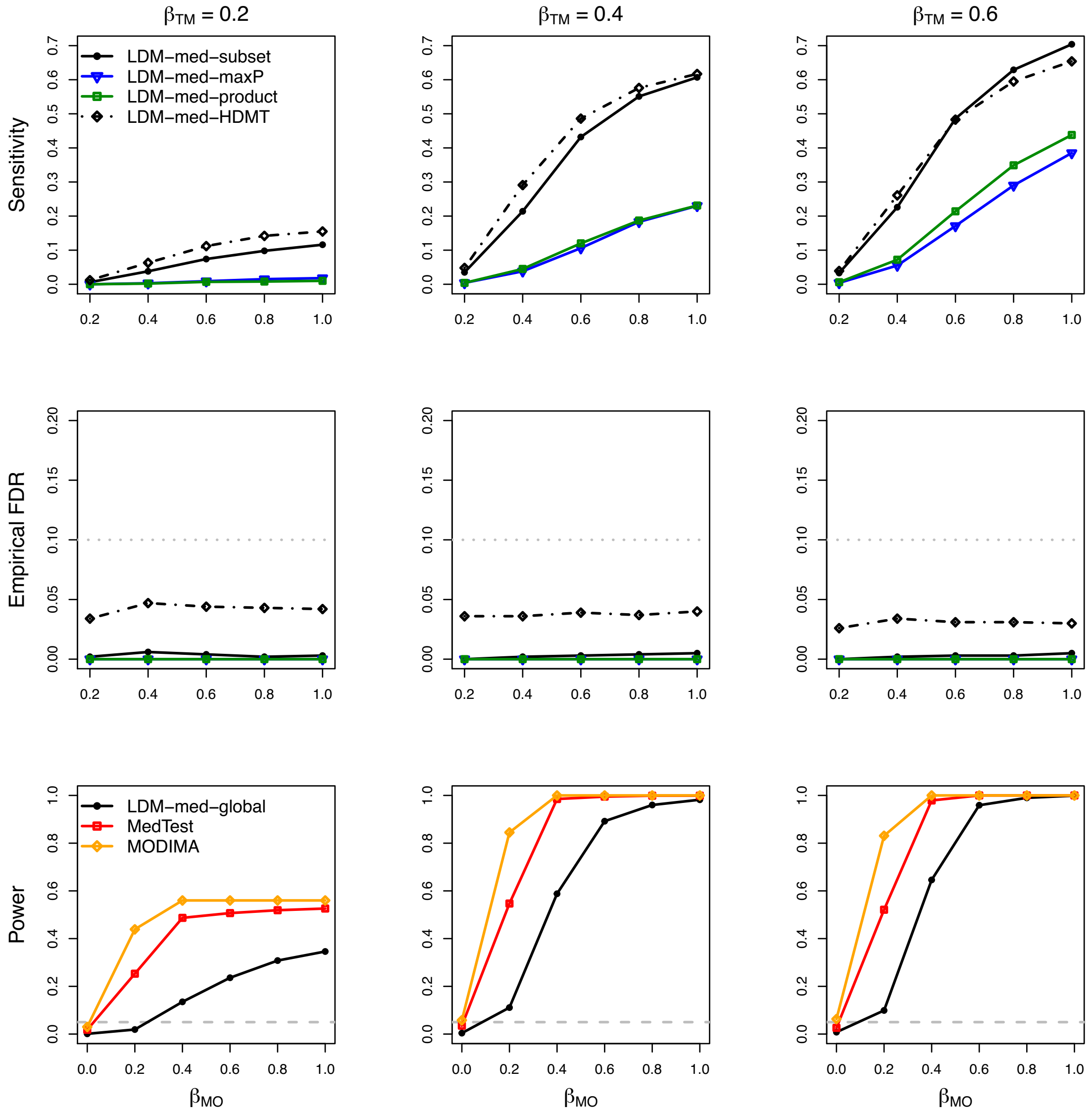


Figure A1: Simulation results in M-common with a continuous outcome and no confounder, in the absence of type-I and type-II null taxa.

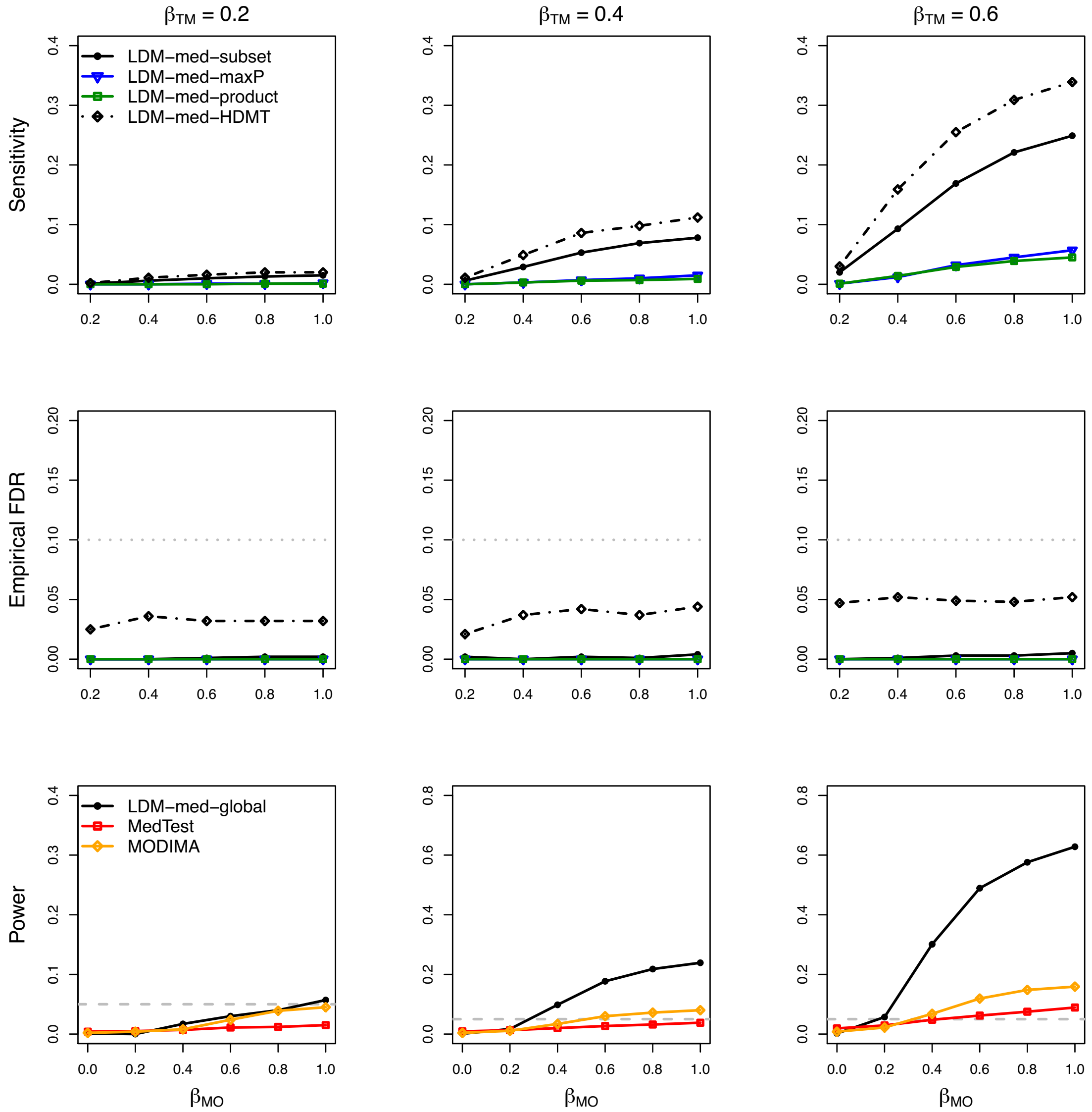


Figure A2: Simulation results in M-rare with a continuous outcome and no confounder, in the absence of type-I and type-II null taxa.

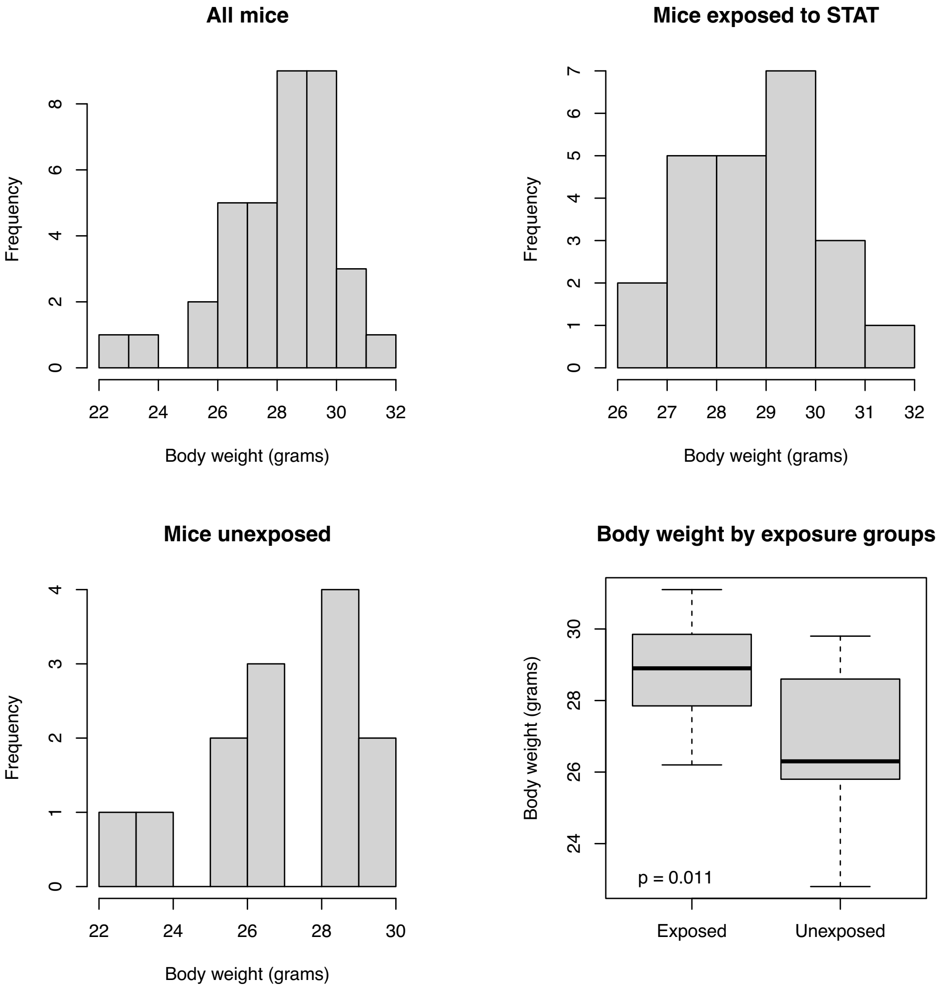


Figure A3: Distribution of the body weight values in the murine microbiome dataset.

Table A1: Type I error (at level 0.05) of the global tests in M-common and M-rare with a continuous outcome and no confounder, in 12 scenarios under the global null

	β_{TM}	β_{MO}	α_{TM}	α_{MO}	Type(s) of null	LDM-med-global	MedTest	MODIMA
M-common	0.0	0.4	0.0	0.0	II	0.004	0.028	0.048
				0.4	II	0.005	0.028	0.048
			0.6	0.0	I, II	0.010	0.577	0.995
				0.4	I, II	0.009	0.610	0.997
	0.6	0.0	0.0	0.0	I	0.008	0.026	0.063
				0.4	I, II	0.010	0.102	0.767
			0.6	0.0	I	0.012	0.030	0.059
				0.4	I, II	0.014	0.092	0.750
	0.0	0.0	0.0	0.0	III	0.000	0.004	0.003
				0.4	II	0.005	0.018	0.039
			0.6	0.0	I	0.006	0.030	0.053
				0.4	I, II	0.009	0.317	0.813
M-rare	0.0	0.4	0.0	0.0	II	0.002	0.010	0.013
				0.4	II	0.003	0.020	0.039
			0.6	0.0	I, II	0.009	0.085	0.233
				0.4	I, II	0.009	0.297	0.827
	0.6	0.0	0.0	0.0	I	0.002	0.019	0.008
				0.4	I, II	0.006	0.055	0.139
			0.6	0.0	I	0.008	0.044	0.051
				0.4	I, II	0.011	0.332	0.807
	0.0	0.0	0.0	0.0	III	0.000	0.004	0.003
				0.4	II	0.005	0.018	0.039
			0.6	0.0	I	0.006	0.030	0.053
				0.4	I, II	0.009	0.317	0.813

Note: see the Note to Table 2.1.

Table A2: Type I error (at level 0.05) of the global tests in M-mixed with a confounder and a continuous outcome, in 3 scenarios under the global null

	β_{TM}	β_{MO}	LDM-med-global	LDM-med-global*	MedTest
Adjusting for the confounder	0.0	0.4	0.007	0.073	0.026
	0.6	0.0	0.004	0.069	0.020
	0.0	0.0	0.001	0.042	0.005
Not adjusting for the confounder	0.0	0.4	0.023	0.119	0.034
	0.6	0.0	0.016	0.108	0.032
	0.0	0.0	0.001	0.024	0.006

Note: we set $\alpha_{\text{TM}} = 0.0$ and $\alpha_{\text{MO}} = 0.0$. LDM-med-global* is a variant of LDM-med-global that uses the information on the type of null for each taxa (only available in simulations). The type I error rates 0.073 and 0.069 after adjusting for the confounder were slightly inflated, due to the small sample size 100, and was reduced to 0.067 and 0.055 when the sample size was increased to 200.

Table A3: Bivariate association analyses of the murine microbiome dataset

	Day 28	Days 21 & 28
Exposure–microbiome		
Detected taxa (FDR = 20%)	*Candidatus Arthromitus*	*Candidatus Arthromitus*
	Turicibacter	*Turicibacter*
	Clostridium.1	*Clostridium.1*
	RF39	*RF39*
	Dehalobacterium	*Dehalobacterium*
	Clostridiales	*Clostridiales*
	Ruminococcus	*Ruminococcus*
	Clostridiaceae	*Clostridiaceae*
	rc4-4	*rc4-4*
	Oscillospira	*Oscillospira*
	Dorea	*Dorea*
	[Ruminococcus]	*[Ruminococcus]*
	Allobaculum	*Allobaculum*
	Enterococcus	*Enterococcus*
	Lactobacillus	
	[Mogibacteriaceae]	*[Mogibacteriaceae]*
	Rikenellaceae	
	Erysipelotrichaceae	*Erysipelotrichaceae*
	Anaeroplasma	
	Clostridium	*Clostridium*
	Adlercreutzia	*Adlercreutzia*
	Coprococcus	
	Akkermansia	*Akkermansia*
	Ruminococcaceae	
	Coriobacteriaceae	
		Anaerostipes
		Enterobacteriaceae
Microbiome–outcome | exposure		
Detected taxa (FDR = 20%)	*[Ruminococcus]*	*[Ruminococcus]*
	Clostridium	
	Candidatus.Arthromitus	
	Ruminococcus	
	Clostridiales	

Appendix B

Topic 3

Bias correction for generalized estimating equations

To simplify, we omit the index j for taxon j and introduce the notation $\mathbb{U} = \mathbb{U}_j$, $\theta = \left(\eta_{1,j}, \eta_{2,j}, \beta_j^{\mathrm{T}}\right)^{\mathrm{T}}$, $X_{ik} = \left(\mathbb{I}(k=1), \mathbb{I}(k=2), Z_i^{\mathrm{T}}\right)^{\mathrm{T}}$, and $M_{ik} = Y_{ik,j} + Y_{ik,J}$. With this notation, we rewrite the estimation equation (4.2) as follows:

$$\mathbb{U}(\theta) = \sum_{i=1,\dots n, k=1,2} \omega_{ik} \left(\frac{Y_{ik}}{M_{ik}} - \mu_{ik} \right) X_{ik} = 0. \tag{B1}$$

Let $\widehat{\theta}$ be the estimate of θ that solves Equation (B1). By taking a Taylor series expansion up to the 2nd order and using $\otimes$ to denote outer product, we obtain

$$\mathbb{U}(\widehat{\theta}) = 0 \approx \mathbb{U}(\theta) + \mathbb{J}(\theta)(\widehat{\theta} - \theta) + \frac{1}{2}(\widehat{\theta} - \theta)^{\mathrm{T}} \mathbb{K}(\theta)(\widehat{\theta} - \theta), \tag{B2}$$

where

$$\mathbb{J}(\theta) = \frac{\partial \mathbb{U}(\theta)}{\partial \theta} = -\sum_{i,k} \omega_{ik} \mu_{ik} \big(1 - \mu_{ik}\big) X_{ik} \otimes X_{ik}$$

and

$$\mathbb{K}(\theta) = \frac{\partial \mathbb{J}(\theta)}{\partial \theta} = -\sum_{i,k} \omega_{ik} \mu_{ik} \big(1 - \mu_{ik}\big)\big(1 - 2\mu_{ik}\big) X_{ik} \otimes X_{ik} \otimes X_{ik}.$$

We calculate the expected value of the right hand side of (B2), taking into account that $\mathrm{E}\big[\mathbb{U}(\theta)\big] = 0$ and that $\mathbb{J}(\theta)$ and $\mathbb{K}(\theta)$ are not functions of the data Y_{ik}, to obtain

$$0 = \mathbb{J}(\theta)\mathrm{E}(\widehat{\theta} - \theta) + \frac{1}{2}\mathrm{E}\Big\{(\widehat{\theta} - \theta)^{\mathrm{T}} \mathbb{K}(\theta)(\widehat{\theta} - \theta)\Big\}.$$

Let $b(\theta) = \mathrm{E}(\widehat{\theta} - \theta)$ be the asymptotic bias in $\widehat{\theta}$, and let $\Sigma(\theta) = \mathrm{E}(\widehat{\theta} - \theta)(\widehat{\theta} - \theta)^{\mathrm{T}}$ be the variance-covariance matrix of $\widehat{\theta}$. Follow from Firth (Firth, 1993), we obtain the bias-corrected estimating equation as

$$\mathbb{U}^*(\theta) = \mathbb{U}(\theta) + \mathbb{J}(\theta) b(\theta) = \mathbb{U}(\theta) - \frac{1}{2}\mathrm{trace}\Big[\mathbb{K}(\theta)\Sigma(\theta)\Big]. \tag{B3}$$

When the weight $\omega_{ik} = M_{ik}$, $\mathbb{U}(\theta)$ resembles the score function for read count data that follow the Binomial distribution. In this case, the model-based variance-covariance estimator $-\mathbb{J}(\theta)^{-1}$ is a reasonable estimator for $\Sigma(\theta)$, as adopted by LOCOM. When $\omega_{ik} = 1$, $\mathbb{U}(\theta)$ significantly deviates from a score function, and we estimate $\Sigma(\theta)$ using the robust sandwich estimator

$$\mathbb{J}(\theta)^{-1} \left[\sum_{i,k} \omega_{ik}^2 \left(\frac{Y_{ik}}{M_{ik}} - \mu_{ik} \right)^2 X_{ik} \otimes X_{ik} \right] \mathbb{J}(\theta)^{-1}.$$

The model-based estimator might not be consistent in the presence of overdispersion in the read count data, and the sandwich estimator may not perform well with finite samples. However, it is important to note that we use these estimators solely for the purpose of bias correction. In the end, we depend on permutation replicates to make valid inferences.

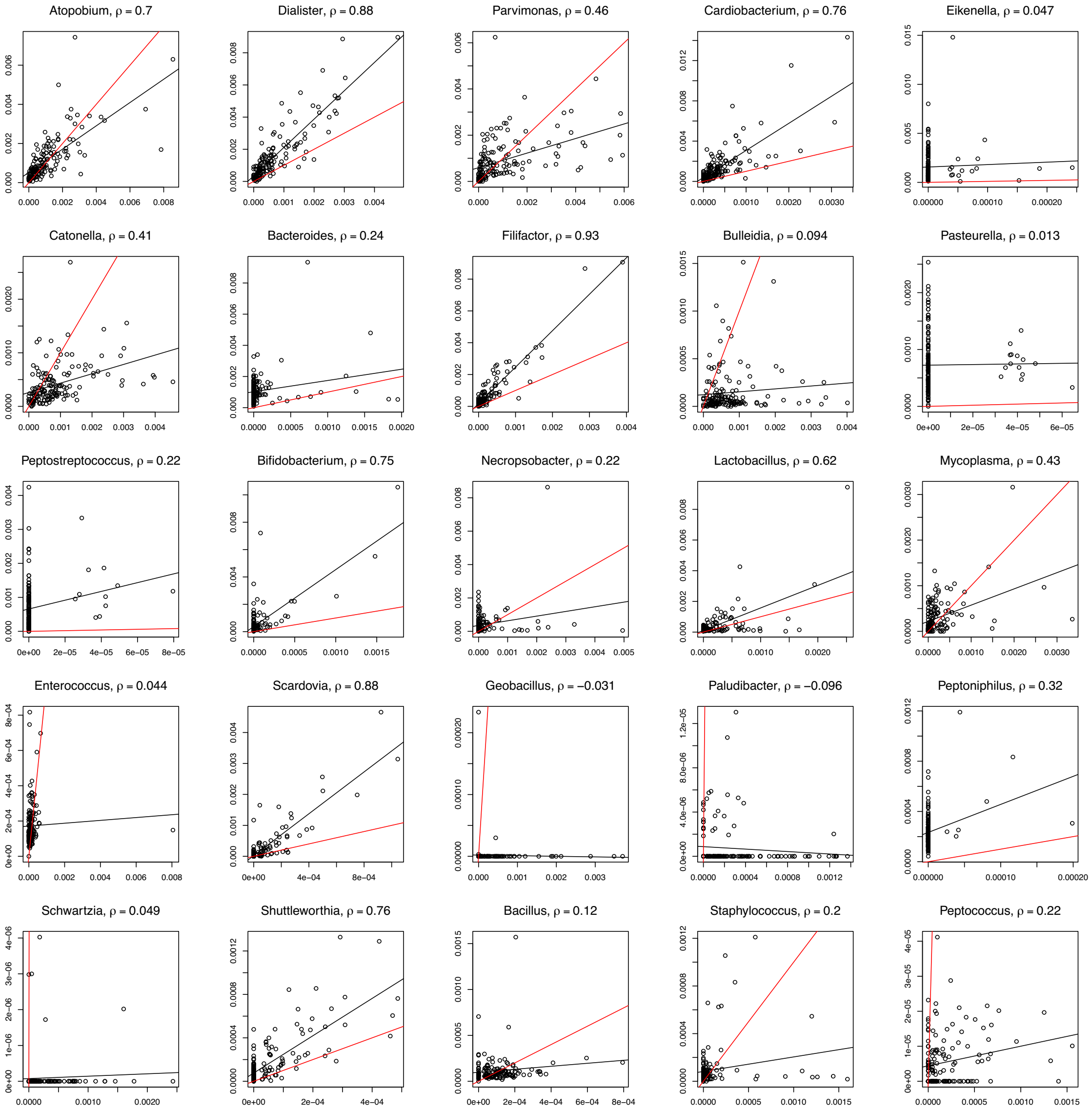


Figure B1: Continuation of Figure 4.2 with the top 26–50 most abundant genera in the ORIGINS data.

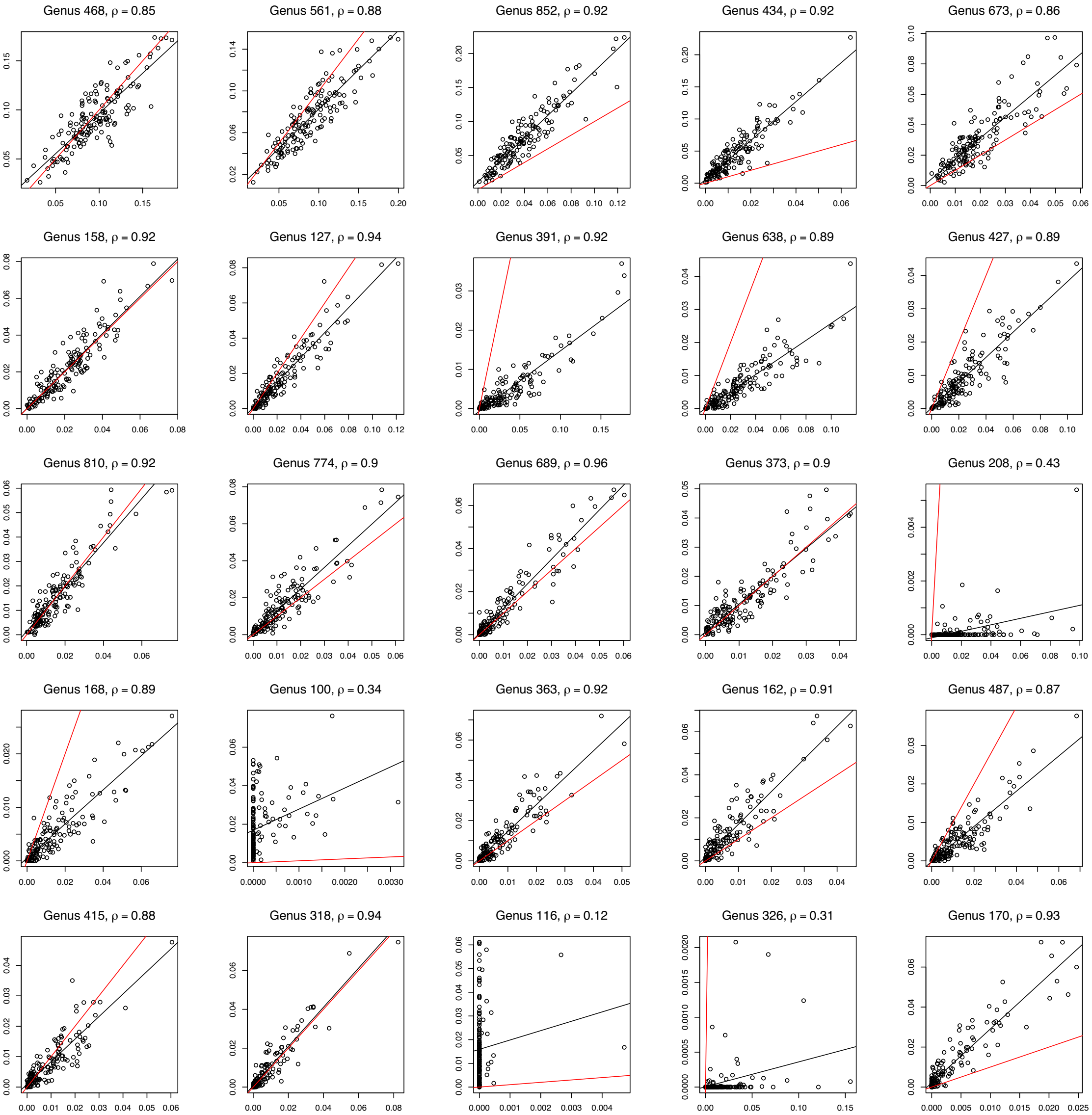


Figure B2: Scatter plot of observed relative abundances from 16S (x-axis) and SMS (y-axis) for the top 1–25 most abundant genera (ordered by decreasing abundance) in the data simulated with 152 completely overlapping samples (the sample size as in Figure 4.2), an overdispersion of $\tau = 0.001$, and a depth ratio of 1:10. Find additional information in the caption of Figure 4.2.

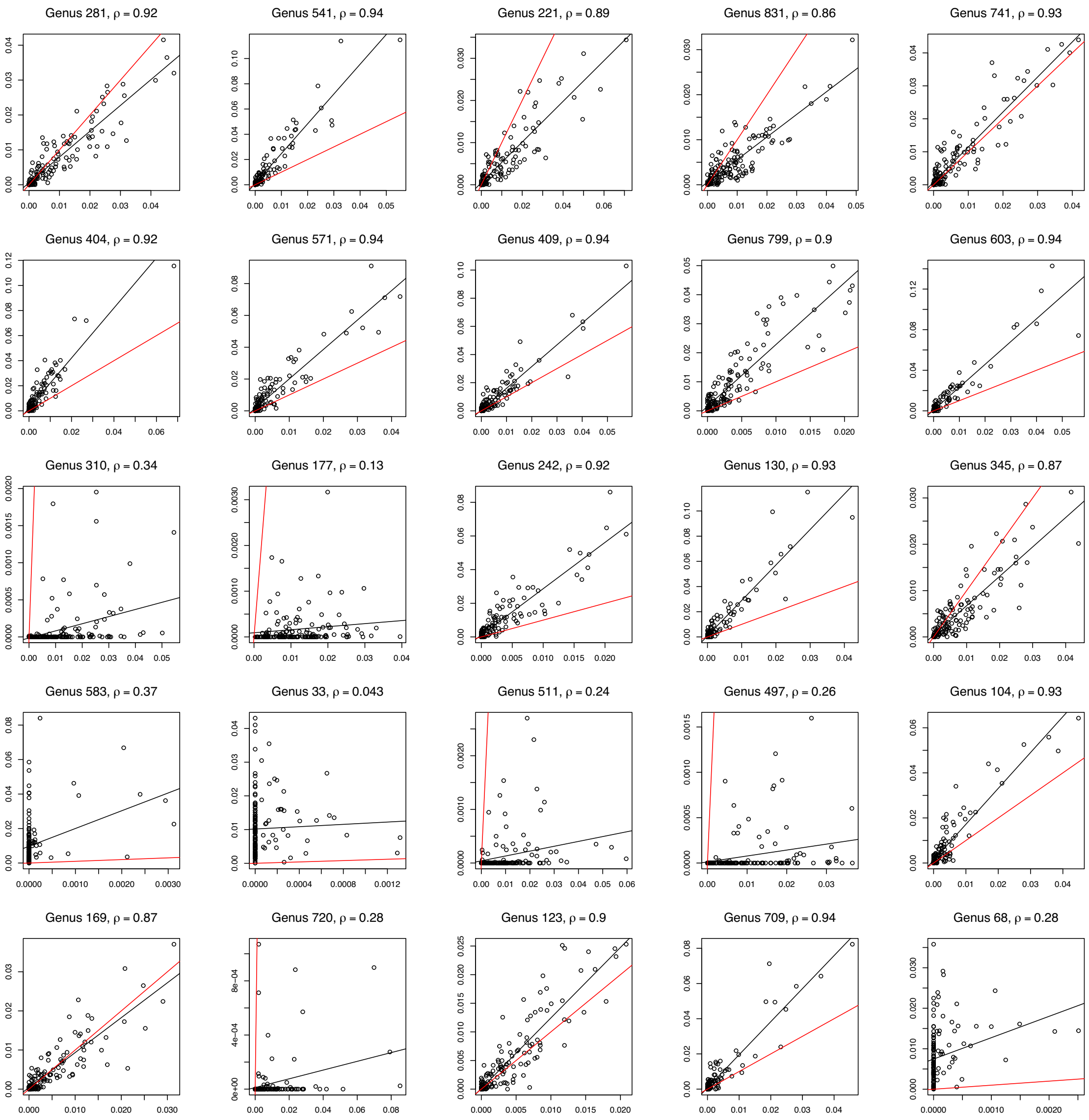


Figure B3: Continuation of Figure B2 with the top 26–50 most abundant genera in the simulated data.

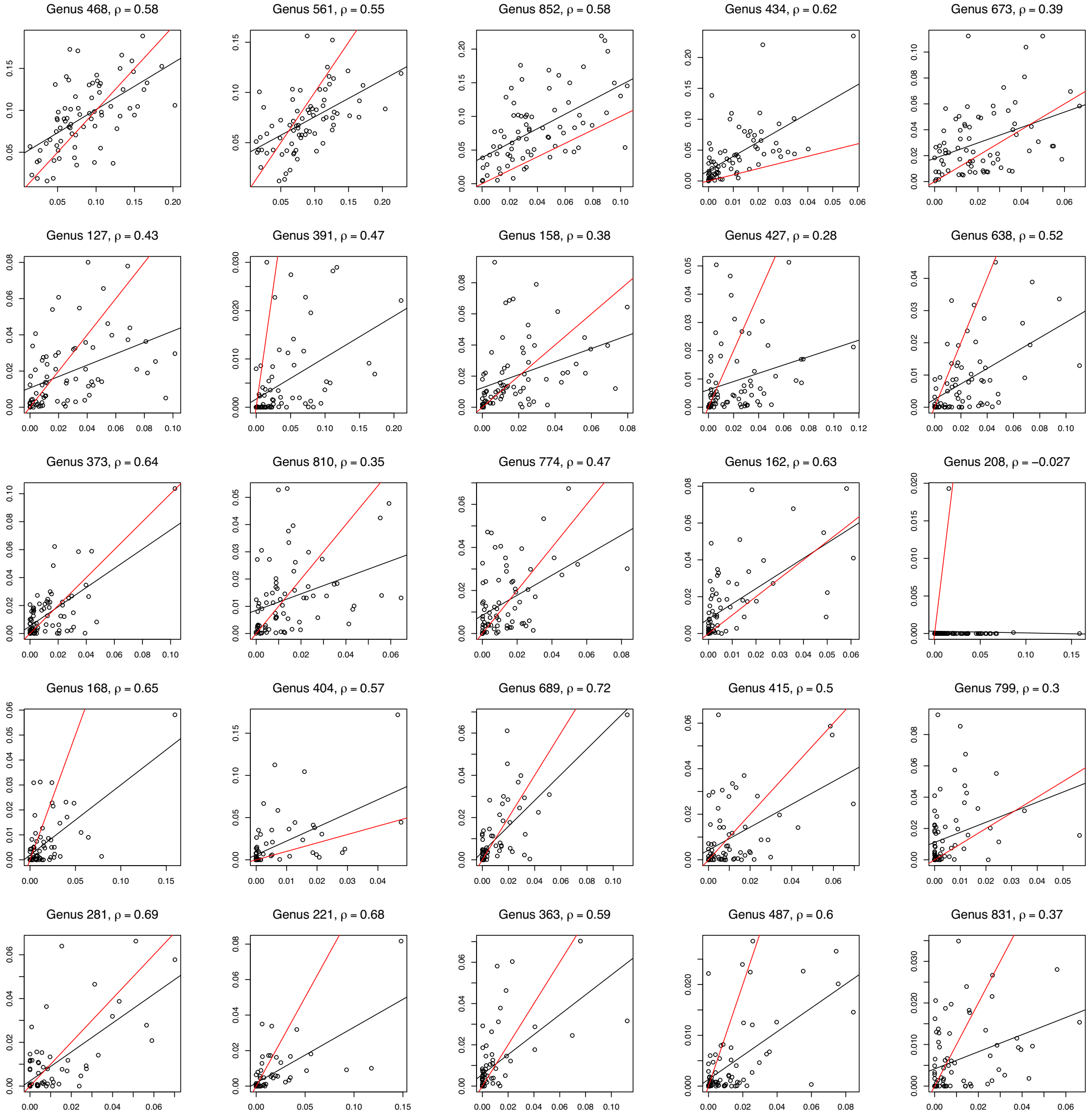


Figure B4: Scatter plot of observed relative abundances from 16S (x-axis) and SMS (y-axis) for the top 1–25 most abundant genera (ordered by decreasing abundance) in the data simulated with 76 completely overlapping samples (the sample size as in Figure B6), an overdispersion of $\tau = 0.01$, and a depth ratio of 1:10. Find additional information in the caption of Figure 4.2.

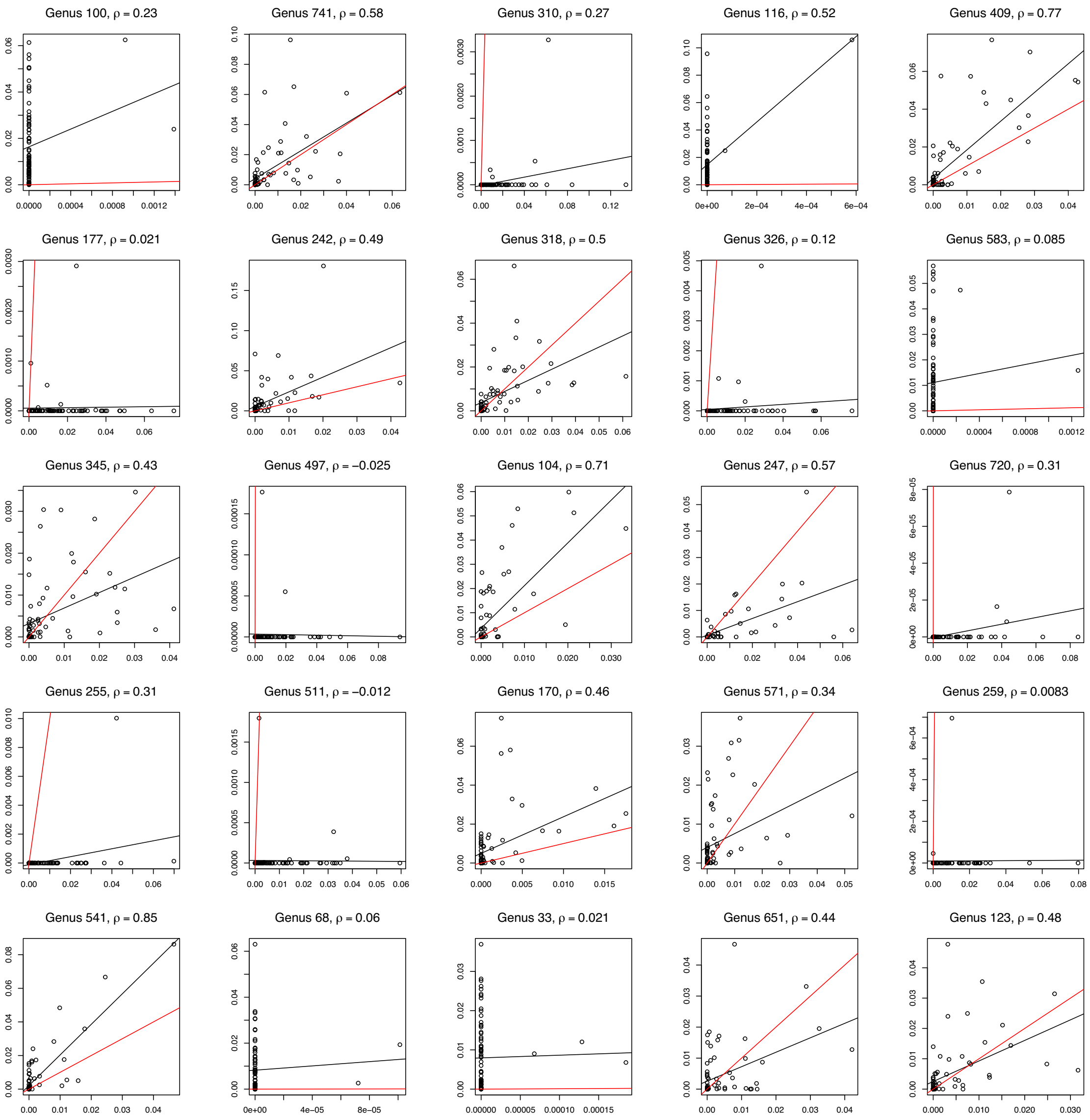


Figure B5: Continuation of Figure B4 with the top 26–50 most abundant genera in the simulated data.

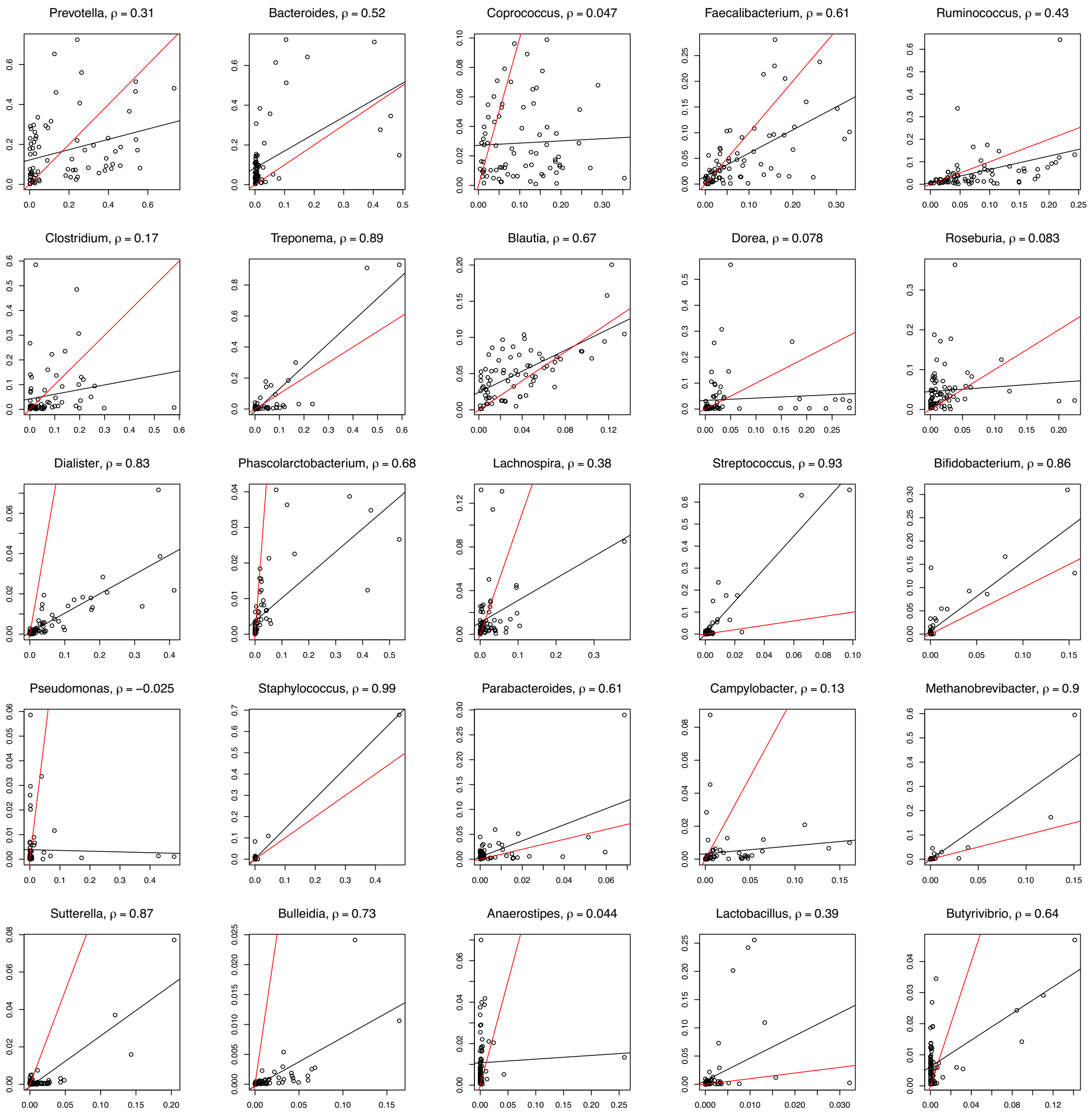


Figure B6: Scatter plot of observed relative abundances from 16S (x-axis) and SMS (y-axis) for the top 1–25 most abundant genera (ordered by decreasing abundance) in the dietary data. Find additional information in the caption of Figure 4.2.

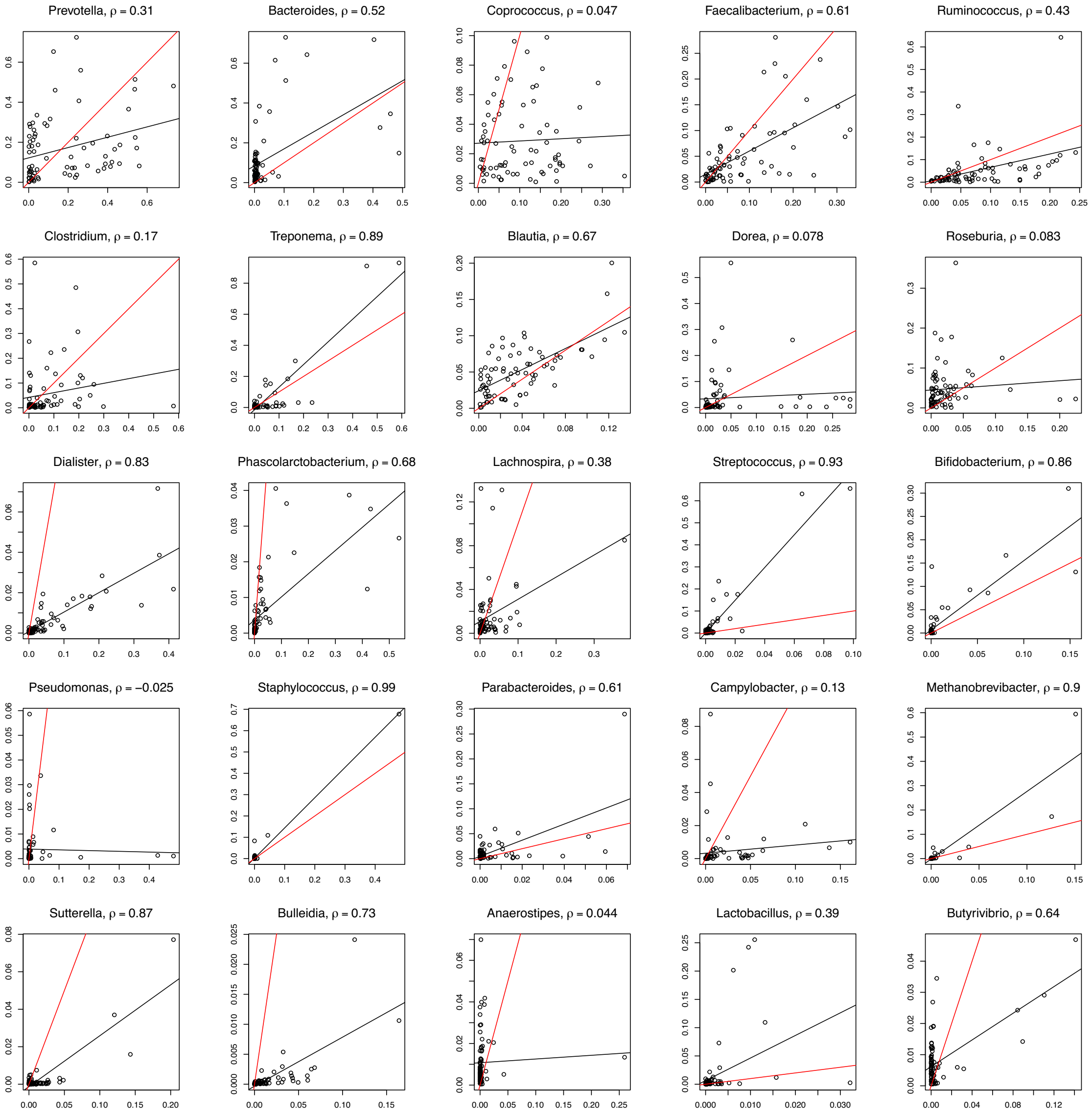


Figure B7: Continuation of Figure B6 with the top 26–50 most abundant genera in the dietary data.

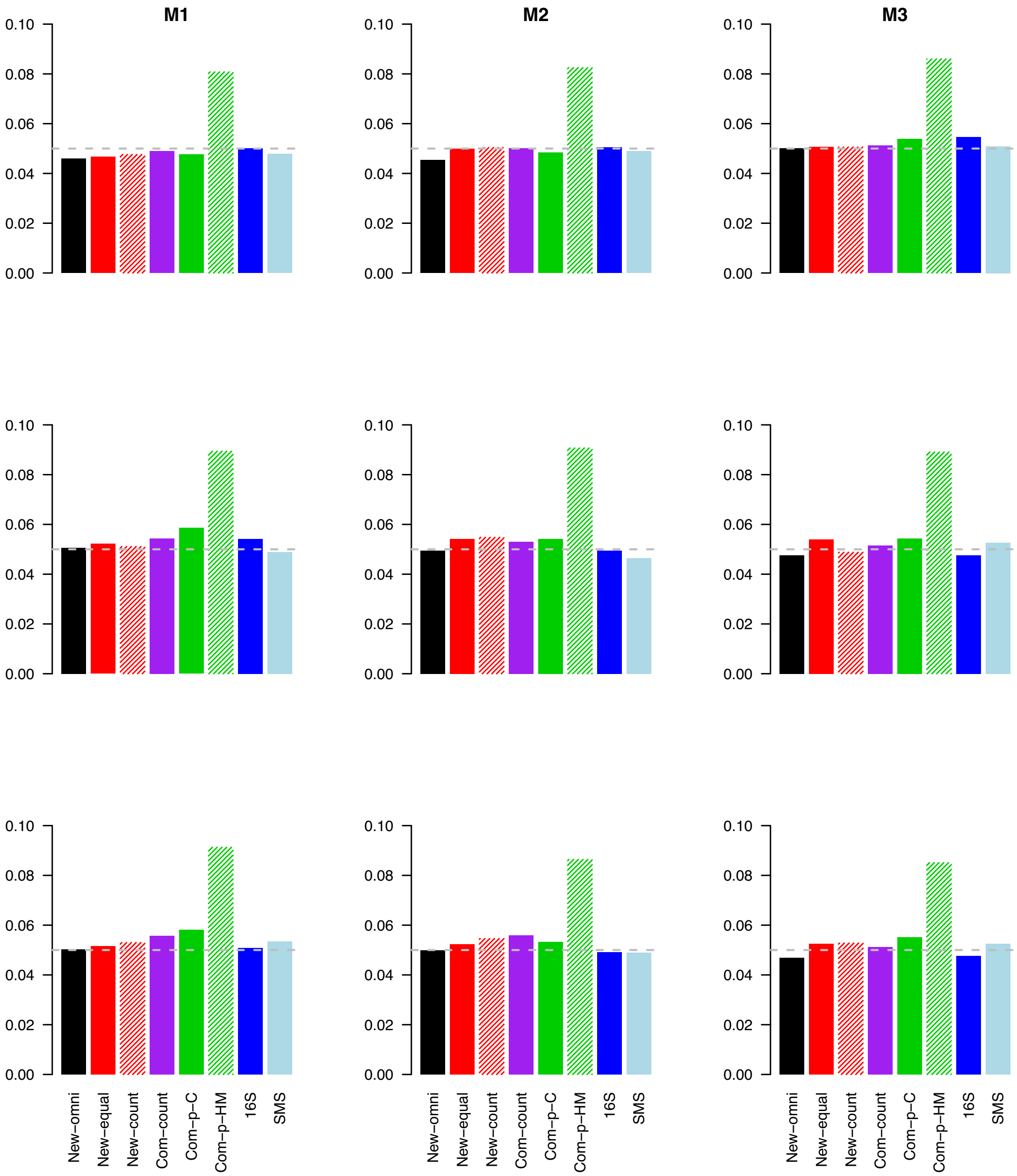


Figure B8: Type I error rate (y-axis) of global tests at the nominal level of 0.05 (gray dashed line), based on data simulated with completely overlapping samples, (upper panel) an overdispersion of $\tau = 0.01$ and a depth ratio of 1:1, (middle panel) $\tau = 0.001$ and a depth ratio of 1:10, and (lower panel) $\tau = 0.001$ and a depth ratio of 1:1.

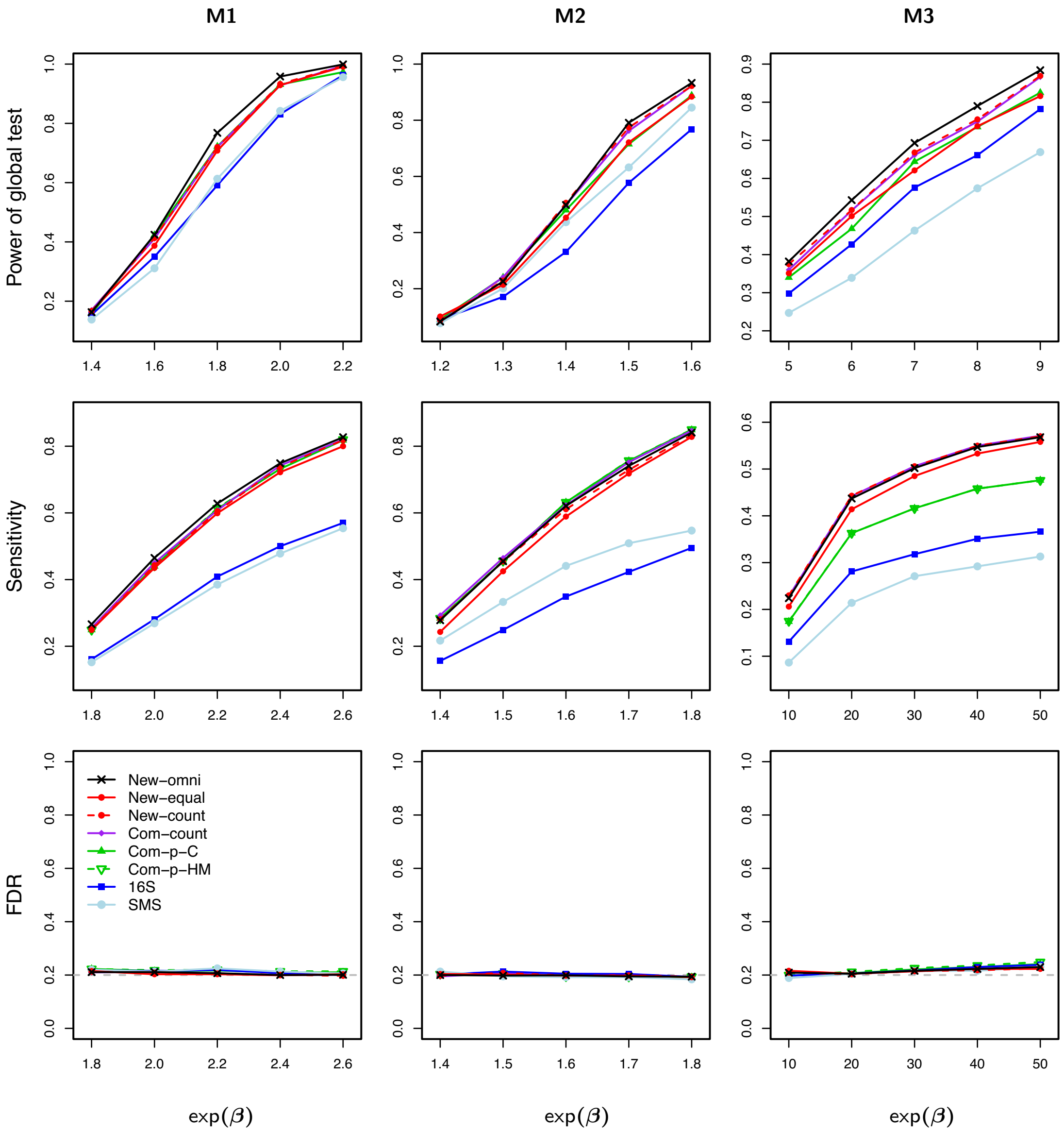


Figure B9: Power (upper panel) for testing the global association and sensitivity (middle panel) and empirical FDR (lower panel) for testing individual taxa, based on data simulated with completely overlapping samples, an overdispersion of $\tau = 0.01$, and a depth ratio of 1:1.

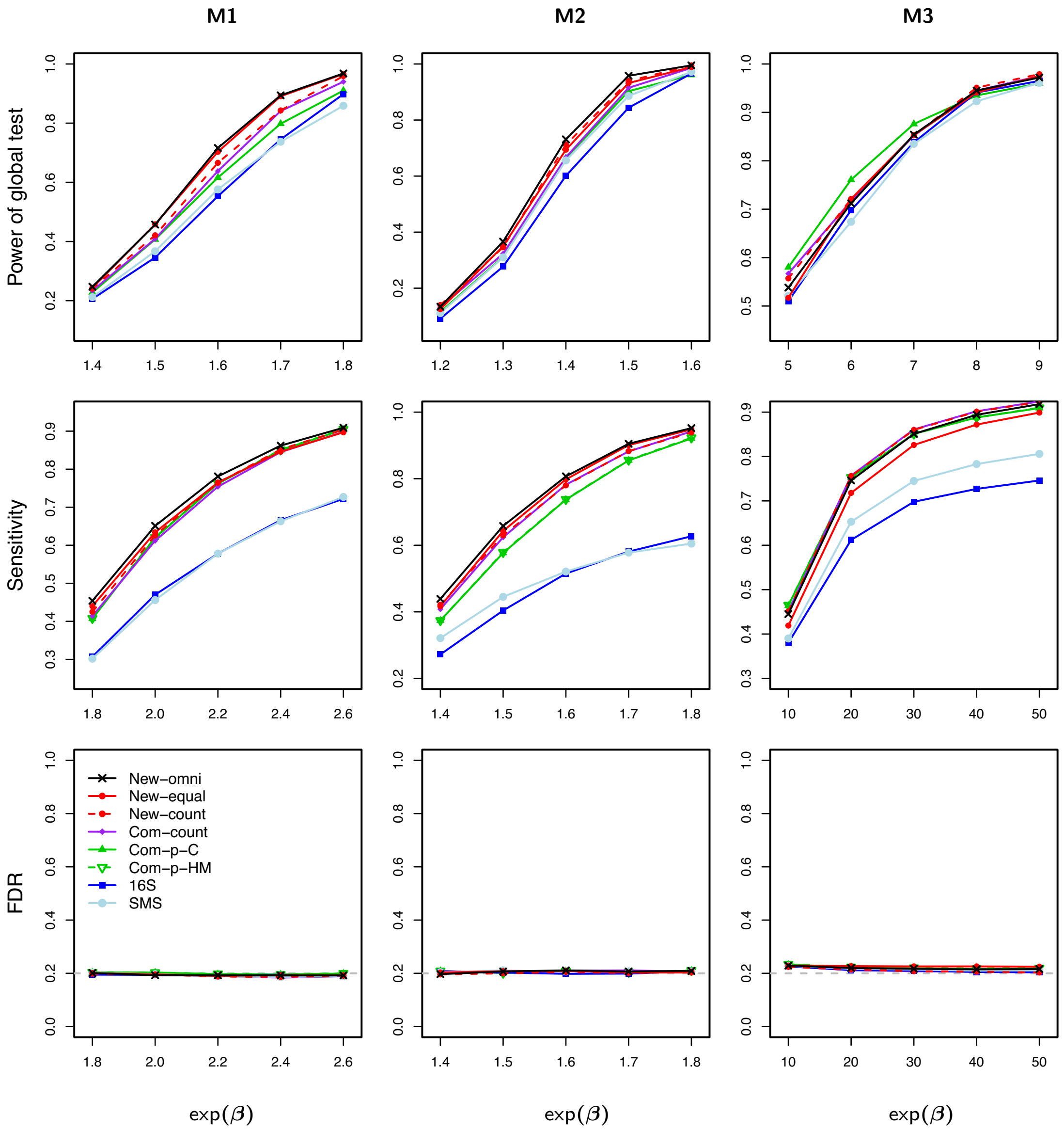


Figure B10: Power (upper panel) for testing the global association and sensitivity (middle panel) and empirical FDR (lower panel) for testing individual taxa, based on data simulated with completely overlapping samples, an overdispersion of $\tau = 0.001$, and a depth ratio of 1:10.

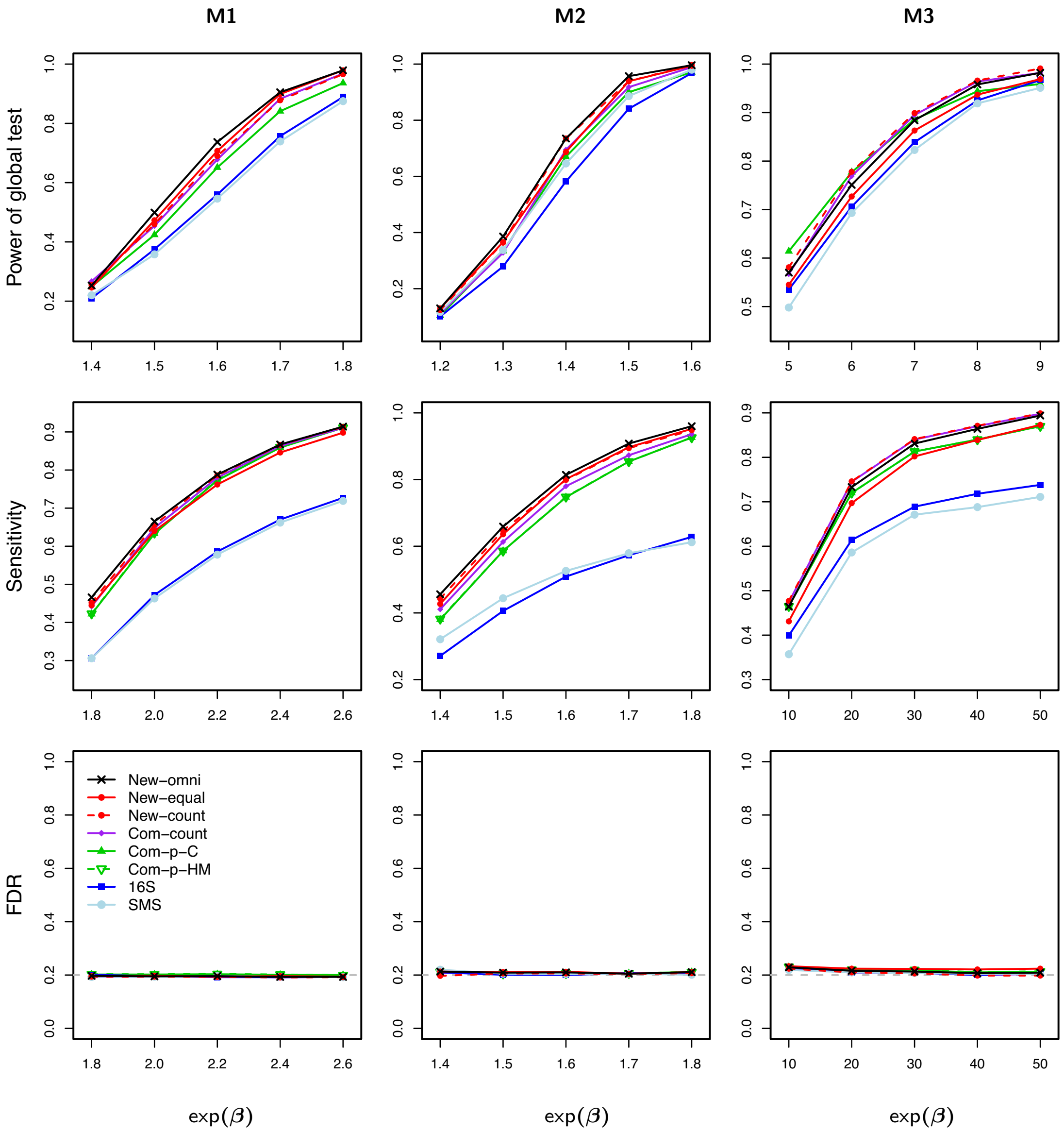


Figure B11: Power (upper panel) for testing the global association and sensitivity (middle panel) and empirical FDR (lower panel) for testing individual taxa, based on data simulated with completely overlapping samples, an overdispersion of $\tau = 0.001$, and a depth ratio of 1:1.

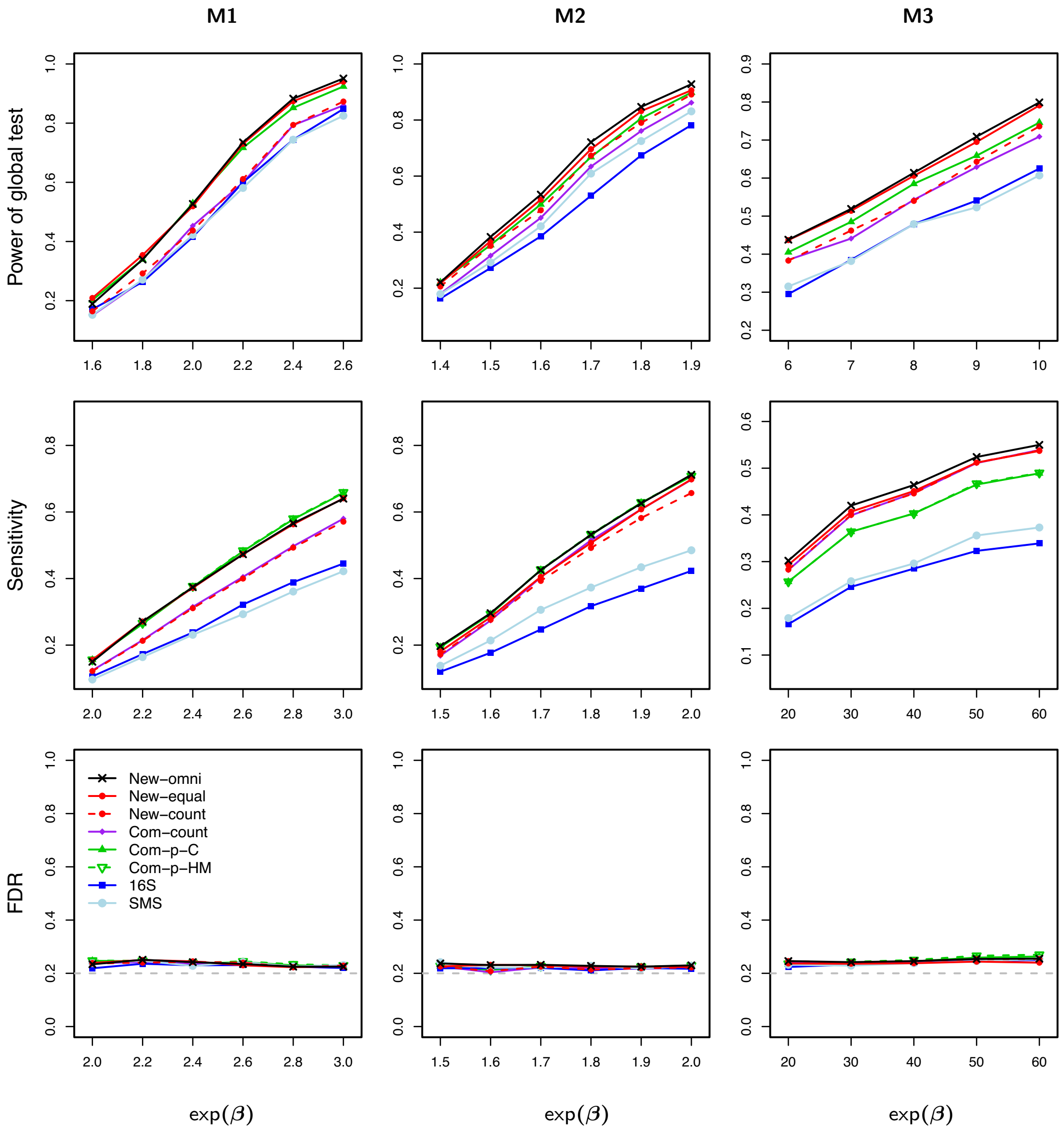


Figure B12: Power (upper panel) for testing the global association and sensitivity (middle panel) and empirical FDR (lower panel) for testing individual taxa, based on data simulated with completely overlapping samples, a confounder, an overdispersion of $\tau = 0.01$, and a depth ratio of 1:10.

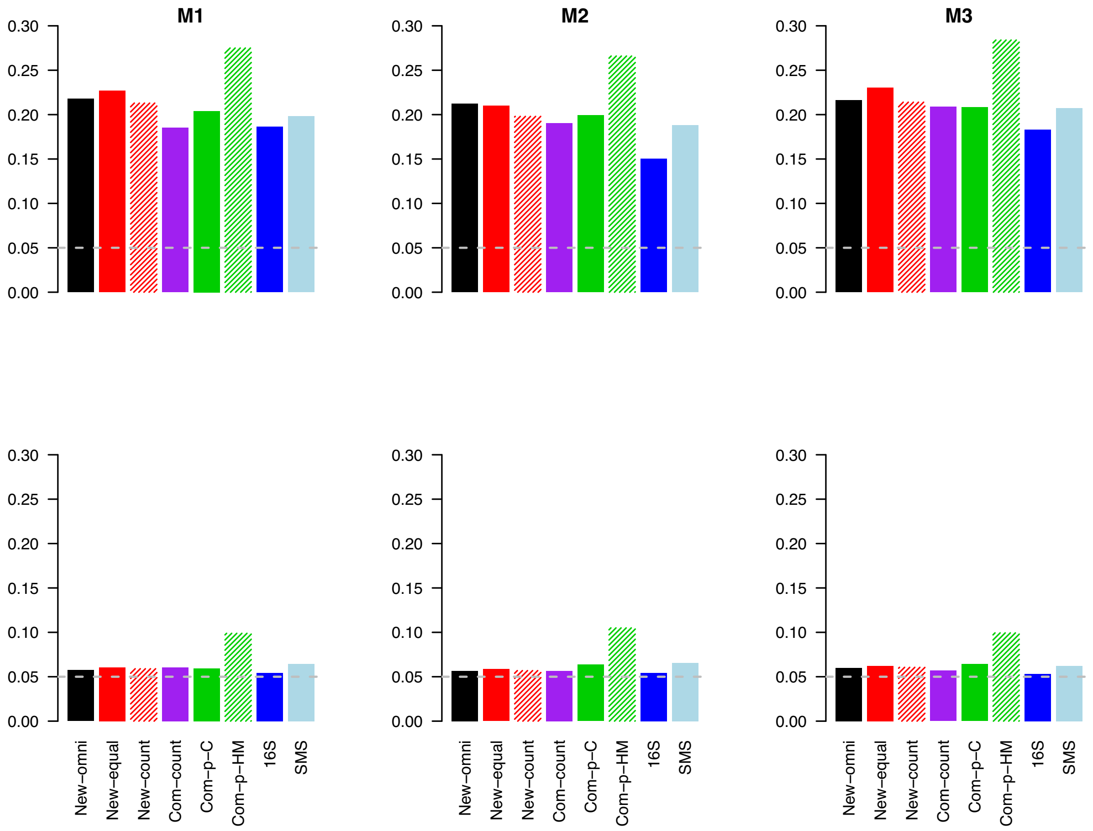


Figure B13: Type I error rate (y-axis) of global tests (at the nominal level of 0.05) without adjusting for the confounder (upper panel) and with adjustment (lower panel), based on data simulated with completely overlapping samples, a confounder, an overdispersion of $\tau = 0.01$, and a depth ratio of 1:10.

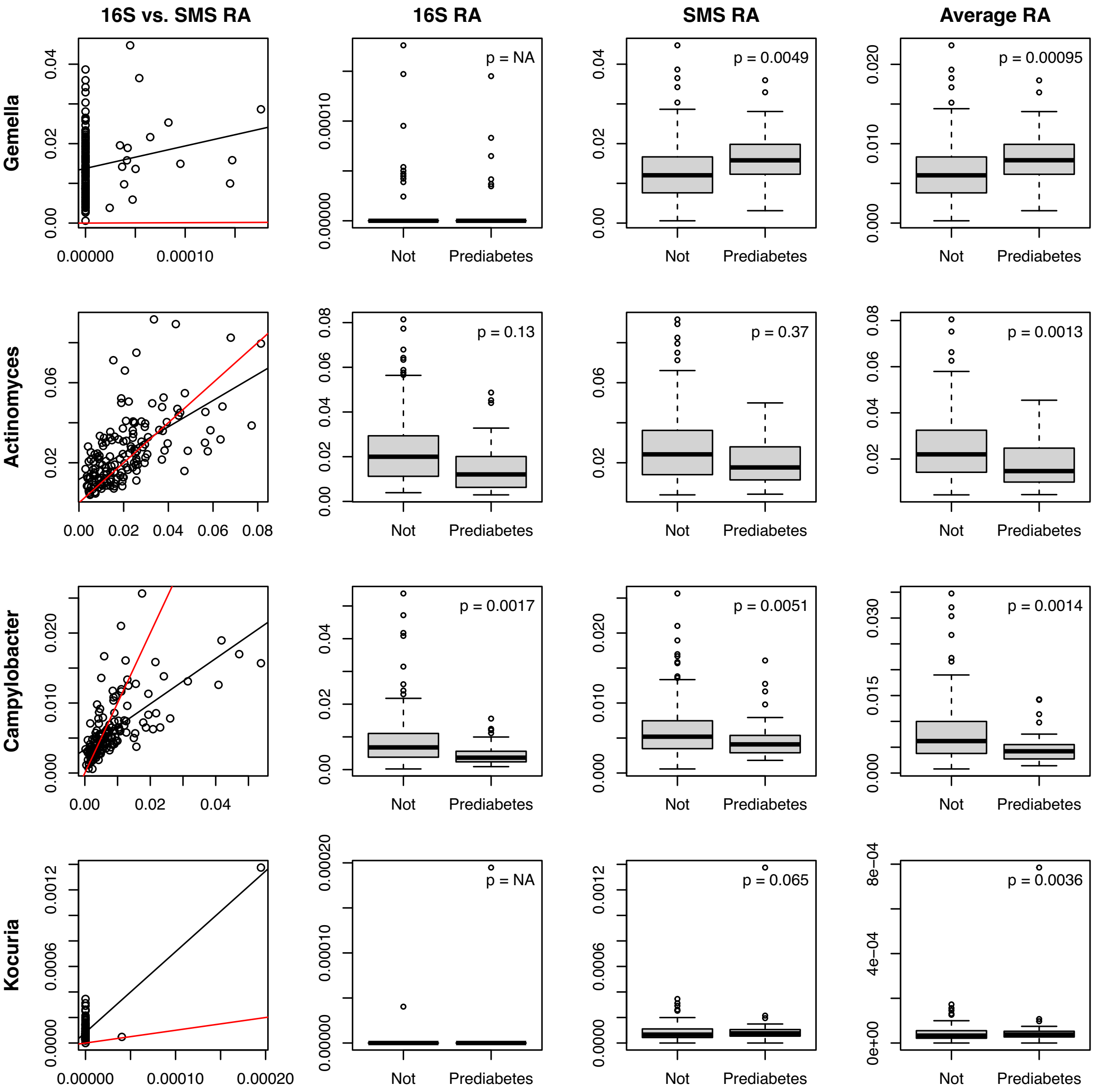


Figure B14: Observed relative abundances (RA) for the detected genera in the analysis of the ORIGINS data for overlapping samples and genera. The first column displays the same scatter plots as those in Figures 4.2 and B1. In the second and third columns, the *p*-values are from the analysis of individual 16S and SMS datasets, as given in Table B2. The last column shows the average of observed relative abundances from 16 and SMS, along with the *p*-values generated by New-omni.

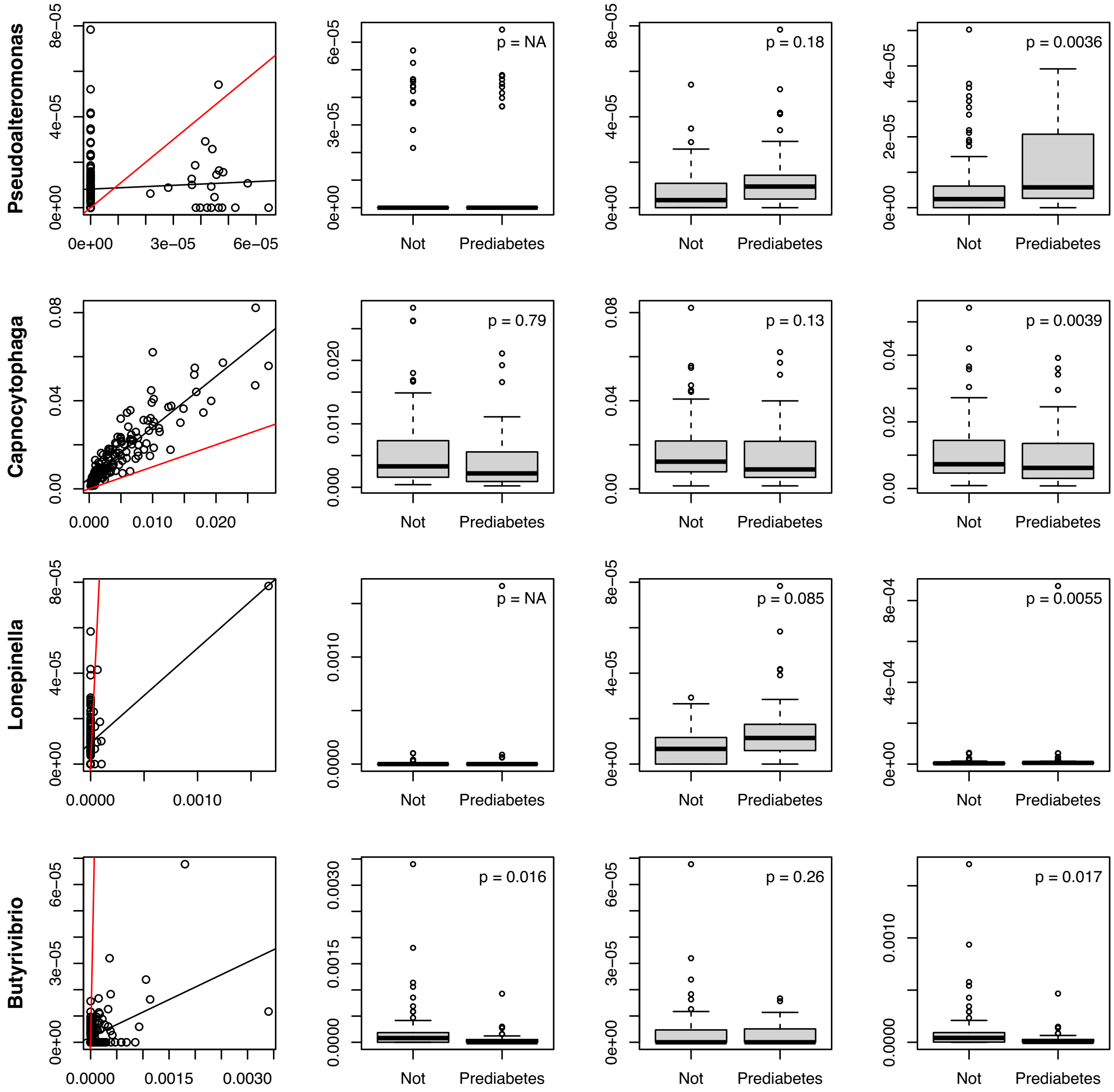


Figure B15: Continuation of Figure B14 for the remaining detected genera.

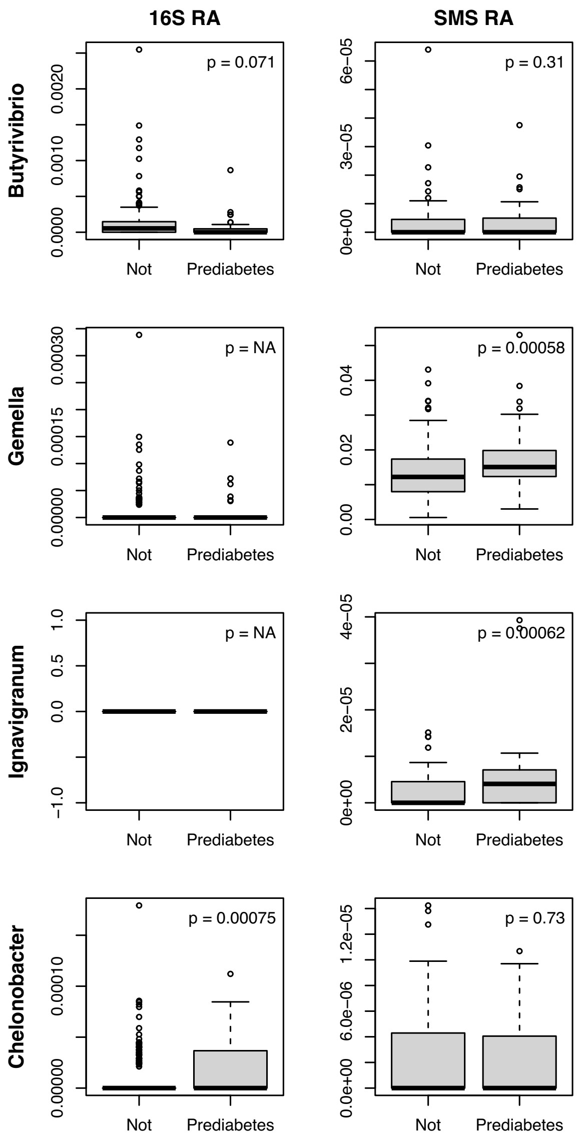


Figure B16: Observed relative abundances (RA) for the detected genera in the analysis of the full ORIGINS data. The observed relative abundances were calculated based on the full 16S or SMS taxa count table. The *p*-values are from the analysis of individual 16S and SMS datasets, as given in Table B3.

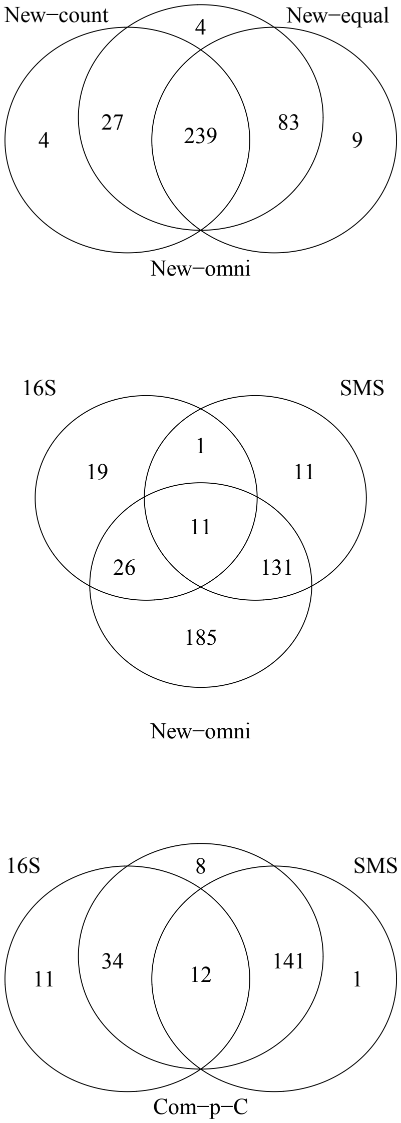


Figure B17: Venn diagram of detected genera at the nominal FDR level of 10% in the analysis of the full dietary data.

Table B1: Studies in Qiita that have both 16S and SMS datasets.

Qiita	16S data				SMS data				Overlapping		
Study	n of	n of	n of	Mean	n of	n of	n of	Mean	n of	n of	Depth
ID	sam	OTU	genus	depth	sam	species	genus	depth	sam	genus	ratio
11479	654	5786	403	19305	1145	3899	1412	26346	638	200	1 : 1.4
10285	26	3217	468	30879	14	3716	1165	46094	14	203	1 : 1.5
10768	23	1386	253	2348	12	795	309	5445	12	77	1 : 2.3
*11808	271	1853	234	26950	183	3346	756	176321	152	125	1 : 6.5
11841	279	4919	497	37290	81	1712	832	259345	54	132	1 : 7.0
†11212	153	9464	524	17072	95	8391	1778	167100	76	236	1 : 9.8
11405	2467	7503	807	39813	1389	7090	1954	466774	1379	404	1 : 11.7
11896	93	1363	273	44588	96	4446	1271	536085	90	155	1 : 12.0
12201	571	17206	1487	22011	382	7925	2076	309328	95	609	1 : 14.1
13114	474	30621	1998	13349	758	11623	2896	195518	364	866	1 : 14.6
11926	94	2525	266	18510	96	3234	1181	294455	94	125	1 : 15.9
10394	1401	2960	621	44342	768	6217	1784	765228	709	275	1 : 17.3
11624	573	5891	790	24035	191	4710	1299	431808	57	337	1 : 18.0
11358	955	21070	1430	23690	40	4233	1324	773541	40	472	1 : 32.7
11444	40	3053	210	40430	40	3024	1047	1345136	40	125	1 : 33.3
11326	576	6335	400	17142	655	6666	1929	589432	119	235	1 : 34.4
11484	96	9486	580	313924	150	8283	2290	11044777	31	327	1 : 35.2
11166	1539	34564	1896	89076	90	11240	2778	6222788	79	913	1 : 70.0
13241	96	3254	547	31007	96	5784	1578	2170533	96	271	1 : 70.0
11149	61	2313	233	29507	84	4544	1375	3320517	24	139	1 : 112.5
13692	205	5427	439	45287	222	9458	2354	7315192	202	263	1 : 161.5
11673	288	8720	997	25068	96	5307	1606	4567862	77	415	1 : 182.2
2338	192	6589	968	82125	6	7704	1913	15932826	6	448	1 : 194.0
11549	70	2447	196	12168	40	3931	1259	4709817	38	112	1 : 387.1
10283	102	4712	376	54484	50	3960	1341	23960223	47	224	1 : 439.8
11546	360	5770	526	21289	382	5628	1766	33312236	306	285	1 : 1564.8

Note: n–number. "sam"–sample. *–the ORIGINS study. †–the dietary study. The studies are ordered by depth ratio, which is the ratio of mean depths in the 16S and SMS data.

Table B2: P-values and adjusted p-values for the detected genera in the analysis of the ORIGINS data for overlapping samples and genera

Method	*Gemella*	*Actinomyces*	*Campylobacter*	*Kocuria*	*Pseudoalteromonas*	*Capnocytophaga*	*Lonepinella*	*Butyrivibrio*
				p-value				
New-omni	0.000947	0.00132	0.00137	0.00358	0.00363	0.00389	0.00553	0.0175
New-equal	0.557	0.000684	0.573	0.283	0.00205	0.00195	0.00311	0.0207
New-count	0.000474	0.187	0.000684	0.00195	0.0105	0.263	0.00542	0.00963
Com-count	0.000500	0.374	0.00090	0.0133	0.0501	0.248	0.0165	0.0005
Com-p-C	0.00491	0.201	0.00257	0.0647	0.183	0.347	0.0852	0.0300
16S	NA	0.131	0.00172	NA	NA	0.794	NA	0.0157
SMS	0.00491	0.368	0.00509	0.0647	0.183	0.127	0.0852	0.261
				Adjusted p-value				
New-omni	0.0447	0.0447	0.0447	0.0636	0.0636	0.0636	0.0774	0.1900
New-equal	0.9080	0.0670	0.9080	0.9000	0.0670	0.0670	0.0761	0.3380
New-count	0.0335	0.5970	0.0335	0.0636	0.1470	0.6290	0.1330	0.1470
Com-count	0.0245	0.7200	0.0294	0.2020	0.3510	0.6570	0.2020	0.0245
Com-p-C	0.236	0.682	0.236	0.415	0.652	0.818	0.481	0.288
16S	NA	0.634	0.093	NA	NA	0.967	NA	0.212
SMS	0.229	0.862	0.229	0.531	0.660	0.660	0.565	0.730

Note: "NA" means that the genus failed to pass the LOCOM filter. The nominal FDR level is 10%.

Table B3: P-values and adjusted p-values for the detected genera in the analysis of the full ORIGINS data

Method	*Butyrivibrio*	*Gemella*	*Ignavigranum*	*Chelonobacter*
		p-value		
New-omni	0.00042	0.00052	0.00014	0.189
New-equal	0.00022	0.260	0.00010	0.112
New-count	0.00590	0.00026	0.00018	0.135
Com-p-C	0.120	0.00058	0.00062	0.0015
16S	0.0711	NA	NA	0.00075
SMS	0.313	0.00058	0.00062	0.727
		Adjusted p-value		
New-omni	0.0763	0.0763	0.0616	0.617
New-equal	0.0484	0.6800	0.0440	0.559
New-count	0.1620	0.0572	0.0572	0.521
Com-p-C	0.499	0.136	0.136	0.191
16S	0.465	NA	NA	0.0637
SMS	0.741	0.125	0.125	0.945

Note: See Note in Table B2.

Bibliography

Alekseyenko, A. V. (2016), 'Multivariate welch t-test on distances', Bioinformatics **32**(23), 3552–3558.

Amato, K. R., G. Sanders, J., Song, S. J., Nute, M., Metcalf, J. L., Thompson, L. R., Morton, J. T., Amir, A., J. McKenzie, V., Humphrey, G. et al. (2019), 'Evolutionary trends in host physiology outweigh dietary niche in structuring primate gut microbiomes', The ISME journal **13**(3), 576–587.

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.

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.

Baron, R. M. and Kenny, D. A. (1986), 'The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations.', Journal of personality and social psychology **51**(6), 1173.

Beghini, F., McIver, L. J., Blanco-Míguez, A., Dubois, L., Asnicar, F., Maharjan, S., Mailyan, A., Manghi, P., Scholz, M., Thomas, A. M. et al. (2021), 'Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with biobakery 3', elife **10**, e65088.

Benjamini, Y. and Hochberg, Y. (1995), '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.

Biegert, G., El Alam, M. B., Karpinets, T., Wu, X., Sims, T. T., Yoshida-Court, K., Lynn, E. J., Yue, J., Medrano, A. D., Petrosino, J. et al. (2021), ‘Diversity and composition of gut microbiome of cervical cancer patients: Do results of 16s rrna sequencing and whole genome sequencing approaches align?’, Journal of microbiological methods **185**, 106213.

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.

Bokulich, N. A., Kaehler, B. D., Rideout, J. R., Dillon, M., Bolyen, E., Knight, R., Huttley, G. A. and Gregory Caporaso, J. (2018), ‘Optimizing taxonomic classification of marker-gene amplicon sequences with qiime 2’s q2-feature-classifier plugin’, Microbiome **6**(1), 1–17.

Bray, J. R. and Curtis, J. T. (1957), ‘An ordination of the upland forest communities of southern wisconsin’, Ecological monographs **27**(4), 326–349.

Brill, B., Amir, A. and Heller, R. (2020), ‘Testing for differential abundance in compositional counts data, with application to microbiome studies’.

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.

Chen, J., Bittinger, K., Charlson, E. S., Hoffmann, C., Lewis, J., Wu, G. D., Collman, R. G., Bushman, F. D. and Li, H. (2012), ‘Associating microbiome composition with environmental covariates using generalized unifrac distances’, Bioinformatics **28**(16), 2106–2113.

Clooney, A. G., Fouhy, F., Sleator, R. D., O’Driscoll, A., Stanton, C., Cotter, P. D. and Claesson, M. J. (2016), ‘Comparing apples and oranges?: next generation sequencing and its impact on microbiome analysis’, PloS one **11**(2), e0148028.

Costea, P., Zeller, G., Sunagawa, S. and Bork, P. (2014), ‘A fair comparison’, Nature methods **11**, 359.

Dai, J. Y., Stanford, J. L. and LeBlanc, M. (2022), 'A multiple-testing procedure for high-dimensional mediation hypotheses', Journal of the American Statistical Association **117**(537), 198–213.

de Vries, J., Saleem, F., Li, E., Chan, A. W. Y., Naphtali, J., Naphtali, P., Paschos, A. and Schellhorn, H. E. (2023), 'Comparative analysis of metagenomic (amplicon and shotgun) dna sequencing to characterize microbial communities in household on-site wastewater treatment systems', Water **15**(2), 271.

Demmer, R., Jacobs Jr, D., Singh, R., Zuk, A., Rosenbaum, M., Papapanou, P. and Desvarieux, M. (2015), 'Periodontal bacteria and prediabetes prevalence in origins: the oral infections, glucose intolerance, and insulin resistance study', Journal of dental research **94**(9_suppl), 201S–211S.

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

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**.

Durazzi, F., Sala, C., Castellani, G., Manfreda, G., Remondini, D. and De Cesare, A. (2021), 'Comparison between 16s rrna and shotgun sequencing data for the taxonomic characterization of the gut microbiota', Scientific reports **11**(1), 1–10.

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), 15.

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.

Gao, Z., Yin, J., Zhang, J., Ward, R. E., Martin, R. J., Lefevre, M., Cefalu, W. T. and Ye, J. (2009), 'Butyrate improves insulin sensitivity and increases energy expenditure in mice', Diabetes **58**(7), 1509–1517.

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

Gloor, G. 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**.

Gonzalez, A., Navas-Molina, J. A., Kosciolek, T., McDonald, D., Vázquez-Baeza, Y., Ackermann, G., DeReus, J., Janssen, S., Swafford, A. D., Orchanian, S. B. et al. (2018), 'Qiita: rapid, web-enabled microbiome meta-analysis', Nature methods **15**(10), 796–798.

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

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.

Hartstra, A. V., Bouter, K. E., Bäckhed, F. and Nieuwdorp, M. (2015), 'Insights into the role of the microbiome in obesity and type 2 diabetes', Diabetes care **38**(1), 159–165.

Hillmann, B., Al-Ghalith, G. A., Shields-Cutler, R. R., Zhu, Q., Gohl, D. M., Beckman, K. B., Knight, R. and Knights, D. (2018), 'Evaluating the information content of shallow shotgun metagenomics', Msystems **3**(6), e00069–18.

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. and Satten, G. A. (2020), 'Testing hypotheses about the microbiome using the linear decomposition model (LDM)', Bioinformatics pp. bbtaa260, https://doi.org/10.1093/bioinformatics/btaa260.

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., Li, Y., Satten, G. A. and Hu, Y.-J. (2022), 'Testing microbiome associations with survival times at both the community and individual taxon levels', PLoS computational biology p. https://doi.org/110.1371/journal.pcbi.1010509.

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

Jaccard, P. (1912), 'The distribution of the flora in the alpine zone. 1', New phytologist **11**(2), 37–50.

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.

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.

Knight, R., Vrbanac, A., Taylor, B. C., Aksenov, A., Callewaert, C., Debelius, J., Gonzalez, A., Kosciolek, T., McCall, L.-I., McDonald, D. et al. (2018), 'Best practices for analysing microbiomes', Nature Reviews Microbiology **16**(7), 410–422.

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

La Scola, B. and Raoult, D. (1998), 'Molecular identification of gemella species from three patients with endocarditis', Journal of clinical microbiology **36**(4), 866–871.

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

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.

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.

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.

Maier, L., Pruteanu, M., Kuhn, M., Zeller, G., Telzerow, A., Anderson, E. E., Brochado, A. R., Fernandez, K. C., Dose, H., Mori, H. et al. (2018), ‘Extensive impact of non-antibiotic drugs on human gut bacteria’, Nature **555**(7698), 623–628.

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.

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.

Marchesi, J. R., Dutilh, B. E., Hall, N., Peters, W. H., Roelofs, R., Boleij, A. and Tjalsma, H. (2011), ‘Towards the human colorectal cancer microbiome’, PloS one **6**(5), e20447.

Mas-Lloret, J., Obón-Santacana, M., Ibáñez-Sanz, G., Guinó, E., Pato, M. L., Rodriguez-Moranta, F., Mata, A., García-Rodríguez, A., Moreno, V. and Pimenoff, V. N. (2020), ‘Gut microbiome diversity detected by high-coverage 16s and shotgun sequencing of paired stool and colon sample’, Scientific data **7**(1), 92.

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.

McDonald, D., Hyde, E., Debelius, J. W., Morton, J. T., Gonzalez, A., Ackermann, G., Aksenov, A. A., Behsaz, B., Brennan, C., Chen, Y. et al. (2018), ‘American gut: an open platform for citizen science microbiome research’, Msystems **3**(3), e00031–18.

McLaren, M. R., Willis, A. D. and Callahan, B. J. (2019), ‘Consistent and correctable bias in metagenomic sequencing experiments’, Elife **8**, e46923.

Nearing, J. T., Comeau, A. M. and Langille, M. G. (2021), 'Identifying biases and their potential solutions in human microbiome studies', Microbiome **9**(1), 1–22.

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).

Paulson, J. N., Bravo, H. C. and Pop, M. (2014), 'Reply to: "a fair comparison"', Nature methods **11**, 359.

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.

Peterson, D., Bonham, K. S., Rowland, S., Pattanayak, C. W., Consortium, R. and Klepac-Ceraj, V. (2021), 'Comparative analysis of 16s rrna gene and metagenome sequencing in pediatric gut microbiomes', Frontiers in microbiology **12**, 670336.

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.

Quigley, E. M. and Gajula, P. (2020), 'Recent advances in modulating the microbiome', F1000Research **9**.

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.

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.

Ruoff, K. L. (2002), 'Miscellaneous catalase-negative, gram-positive cocci: emerging opportunists', Journal of Clinical Microbiology **40**(4), 1129–1133.

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.

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.

Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O. and Huttenhower, C. (2012), ‘Metagenomic microbial community profiling using unique clade-specific marker genes’, Nature methods **9**(8), 811–814.

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.

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. (2022), ‘A compositional mediation model for a binary outcome: Application to microbiome studies’, Bioinformatics **38**(1), 16–21.

Spencer, C. N., McQuade, J. L., Gopalakrishnan, V., McCulloch, J. A., Vetizou, M., Cogdill, A. P., Khan, M. A. W., Zhang, X., White, M. G., Peterson, C. B. et al. (2021), ‘Dietary fiber and probiotics influence the gut microbiome and melanoma immunotherapy response’, Science **374**(6575), 1632–1640.

Székely, G. J. and Rizzo, M. L. (2009), ‘Brownian distance covariance’, The annals of applied statistics **3**(4), 1236–1265.

Székely, G. J. and Rizzo, M. L. (2014), ‘Partial distance correlation with methods for dissimilarities’, The Annals of Statistics **42**(6), 2382–2412.

Székely, G. J., Rizzo, M. L. and Bakirov, N. K. (2007), ‘Measuring and testing dependence by correlation of distances’, The annals of statistics **35**(6), 2769–2794.

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.

Truong, D. T., Franzosa, E. A., Tickle, T. L., Scholz, M., Weingart, G., Pasolli, E., Tett, A., Huttenhower, C. and Segata, N. (2015), ‘Metaphlan2 for enhanced metagenomic taxonomic profiling’, Nature methods **12**(10), 902–903.

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.

Wang, C., Hu, J., Blaser, M. J. and Li, H. (2020), 'Estimating and testing the microbial causal mediation effect with high-dimensional and compositional microbiome data', Bioinformatics **36**(2), 347–355.

Weinstock, G. M. (2012), 'Genomic approaches to studying the human microbiota', Nature **489**(7415), 250–256.

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

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

Woo, P., Lau, S., Fung, A., Chiu, S., Yung, R. and Yuen, K. (2003), 'Gemella bacteraemia characterised by 16s ribosomal rna gene sequencing', Journal of clinical pathology **56**(9), 690–693.

Wood, D. E., Lu, J. and Langmead, B. (2019), 'Improved metagenomic analysis with kraken 2', Genome biology **20**, 1–13.

Wood, D. E. and Salzberg, S. L. (2014), 'Kraken: ultrafast metagenomic sequence classification using exact alignments', Genome biology **15**(3), 1–12.

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.

Zhang, H., Chen, J., Li, Z. and Liu, L. (2021), 'Testing for mediation effect with application to human microbiome data', Statistics in Biosciences **13**(2), 313–328.

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, 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. PMCID: PMC4570290.

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

Zhu, Q., Huang, S., Gonzalez, A., McGrath, I., McDonald, D., Haiminen, N., Armstrong, G., Vázquez-Baeza, Y., Yu, J., Kuczynski, J. et al. (2022), 'Phylogeny-aware analysis of metagenome community ecology based on matched reference genomes while bypassing taxonomy', Msystems **7**(2), e00167–22.

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 p. https://doi.org/10.1093/bioinformatics/btac181.

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.

Zuo, W., Wang, B., Bai, X., Luan, Y., Fan, Y., Michail, S. and Sun, F. (2022), '16s rrna and metagenomic shotgun sequencing data revealed consistent patterns of gut microbiome signature in pediatric ulcerative colitis', Scientific Reports **12**(1), 6421.