ANNUAL REVIEWS

Annual Review of Statistics and Its Application

Analysis of Microbiome Data

Christine B. Peterson, Satabdi Saha, and Kim-Anh Do

Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA; email: cbpeterson@mdanderson.org

Annu. Rev. Stat. Appl. 2024. 11:483–504

First published as a Review in Advance on October 13, 2023

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

https://doi.org/10.1146/annurev-statistics-040522-120734

Keywords

compositional data, differential abundance, network inference, ordination, regression modeling, zero inflation

Abstract

The microbiome represents a hidden world of tiny organisms populating not only our surroundings but also our own bodies. By enabling comprehensive profiling of these invisible creatures, modern genomic sequencing tools have given us an unprecedented ability to characterize these populations and uncover their outsize impact on our environment and health. Statistical analysis of microbiome data is critical to infer patterns from the observed abundances. The application and development of analytical methods in this area require careful consideration of the unique aspects of microbiome profiles. We begin this review with a brief overview of microbiome data collection and processing and describe the resulting data structure. We then provide an overview of statistical methods for key tasks in microbiome data analysis, including data visualization, comparison of microbial abundance across groups, regression modeling, and network inference. We conclude with a discussion and highlight interesting future directions.

1. INTRODUCTION

OTU: operational taxonomic unit

Microbiome studies seek to identify and characterize the functions of microorganisms, including bacteria, fungi, and viruses, present in a given habitat. Environmental microbiome research focuses on profiling the microorganisms found in settings such as the soil or ocean and understanding how these communities shape environmental and human health (Fierer 2017). There is a critical interest in analyzing microbiome data to address scientific questions such as how climate change will reshape the soil and ocean microbiomes (Jansson & Hofmockel 2020, Tara Ocean Found. et al. 2022). The human body itself serves as a host to various microbial populations that are distinct across body sites, with the gut harboring the most diverse set of organisms (Turnbaugh et al. 2007). The microbiome has been shown to influence conditions ranging from inflammatory bowel disease to neurological disorders (Lloyd-Price et al. 2019, Cryan et al. 2020) and also plays a key role in influencing risk, treatment response, and survival outcomes across a variety of cancer types (Gopalakrishnan et al. 2018, Riquelme et al. 2019). The human microbiome is dynamic and can be reshaped by diet, consumption of pre- or probiotics, or specially designed therapies with beneficial bacterial combinations. In this review, we discuss statistical methods for microbiome data analysis and identify future directions of research in the field. A visual summary of our review is provided in **Figure 1**.

1.1. Data Collection and Processing

Microbiome profiling begins with the collection of a physical sample from the target habitat. For human microbiome studies, this might consist of a stool sample or an oral or skin swab. For environmental microbiome studies, it may be a soil or water sample. The genetic material present in the sample is then sequenced. There are two main approaches for genetic sequencing of microbiome data: targeted sequencing of marker genes, most commonly the 16S rRNA gene, and untargeted sequencing of all the genetic material present in the sample, known as whole metagenome sequencing (WMS).

1.1.1. Marker gene sequencing data. The most popular microbiome profiling approach relies on sequencing of the 16S rRNA gene, which has stable regions that are shared across all bacteria and also variable regions that can be used to identify signatures of specific organisms. 16S sequencing has been a mainstay in the microbiome field, as it allows for efficient and cost-effective profiling. Since there may be errors in the sequencing process, bioinformatic tools are needed to identify biological features from the data. In the past, many studies relied on clustering of similar sequences into groupings known as operational taxonomic units (OTUs). However, since these are identified using clustering, they are specific to a given data set and are not stable units for comparison across studies. An alternative approach is to denoise the sequences to obtain

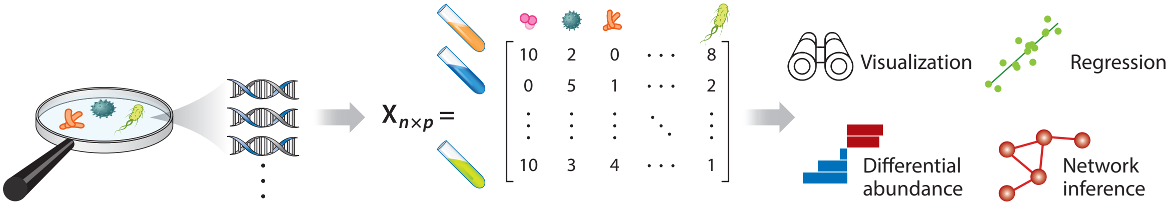

Figure 1

Overview illustration of microbiome analysis: Microorganisms present in a sample are first quantified using genomic sequencing to obtain an abundance table **X**, which quantifies *p* microbiome features across *n* samples and can be used for downstream statistical analyses.

amplicon sequence variants (ASVs), which correspond to specific marker gene sequences. Since these are of much finer resolution, they are sometimes referred to as zero-radius OTUs.

The observed features (either OTUs or ASVs) can be organized into a tree structure, with similar features appearing as nearby leaves in the tree. Two types of trees are typically used: taxonomic trees and phylogenetic trees. The first reflects the classical organization of living things into progressively finer groups, based on the ranks kingdom, phylum, class, order, family, genus, and species. The observed sequences can be assigned taxonomic labels through comparison to a reference database, using tools such as DADA2 or QIIME 2 (Callahan et al. 2016, Bolyen et al. 2019). In contrast, phylogenetic trees, which can be inferred directly from the observed marker gene sequences, represent potential evolutionary relationships among microorganisms in the sample. Such trees capture useful information for downstream analysis, as they summarize similarities of the genetic sequences; this offers a proxy for the functional similarity of the microorganisms represented in the data set.

ASV: amplicon sequence variant

WMS: whole metagenome sequencing

MAG: metagenome-assembled genome

SGB: species-level genome bin

Although it is a practical tool, 16S profiling has limitations: Only bacteria are quantified, so other microorganisms such as viruses are missed, and the focus is on identifying organisms, rather than characterizing their function. The first issue can be partially addressed by targeting additional marker genes, such as the internal transcribed spacer (ITS) region in fungi, while the second issue can be partially addressed through the use of bioinformatic tools that predict functional gene content based on observed marker gene sequences (Douglas et al. 2020). A more complete solution to these challenges is to rely on WMS, which may be used independently or as a complement to 16S profiling.

1.1.2. Whole metagenome sequencing data. WMS, also known as shotgun metagenomic sequencing, can be used to comprehensively profile all the DNA in a sample. Here, genome refers to the genetic material for an individual organism, while metagenome refers to the collection of genomes for all microorganisms in the sample. The observed sequences are typically short fragments; these can be assembled to reconstruct the genome of an individual organism, known as a metagenome-assembled genome (MAG). Although the underlying profiling approach is different, metagenomic sequencing data can also be used to quantify the abundance of taxonomic features using tools such as Kraken 2 or MetaPhlAn 4 (Lu et al. 2022, Blanco-Míguez et al. 2023). The resulting features in this context are known as species-level genome bins (SGBs) and correspond to both known bacterial species and unknown clusters of genomes related at roughly the species level. WMS also enables the characterization of nonbacterial organisms, including viruses, and the quantification of functional gene pathways by assigning the observed genetic sequences to biological roles (Beghini et al. 2021).

1.2. Microbiome Data Structure

The observed data from either 16S or WMS profiling can be summarized as an $n \times p$ matrix $\mathbf{X}$ of counts x_{ij} for each taxonomic feature (OTU, ASV, or SGB) or functional activity in each sample, where $i = 1, \ldots, n$ indexes the sample and $j = 1, \ldots, p$ indexes the feature. This poses several challenges for downstream analysis:

1. Microbiome data are count based, as $x_{ij} \in \{0, 1, 2, \ldots\}$, so many classical statistical methods based on the Gaussian distribution cannot be directly applied.
2. Microbiome data are compositional, which means that the counts within each sample (row of $\mathbf{X}$) have a fixed sum, $\sum_{j=1}^{p} x_{ij} = N_i$. Here, N_i represents the total number of reads for sample i, also known as the library size. Hence, the counts can only be interpreted on a relative scale.

3. Microbiome data are also zero-inflated, since features observed in one sample may not appear in samples from other subjects, resulting in exact zeros in the abundance table. In fact, many features may be rare, i.e., zero for nearly all samples. In practice, data are often filtered so that only features with sufficient prevalence (for example, present in at least 25% of samples) are considered in downstream analysis. This motivates the use of methods designed for zero-inflated and rare features.
4. Microbiome data are high dimensional, with thousands of features quantified in a given data set and $p \gg n$. This aspect of the data necessitates advanced methods for data visualization, multiplicity correction, and sparse modeling.
5. Microbiome data are tree-structured, in that the observed features can be meaningfully organized into a taxonomic or phylogenetic tree $\mathcal{T}$. In this context, phylogenetic trees may capture the most relevant information, as genomically similar features may play similar functional roles. Taxonomic trees are also meaningful, as they provide a scaffold for the interpretation and comparison of findings across studies.

Some aspects of microbiome data are shared with other high-throughput data types: In particular, bulk RNA-sequencing data are also compositional and count based, while single-cell RNA sequencing data have similarities with microbiome data, including the presence of many zero values. Methods developed for the analysis of these data types may sometimes be applicable to microbiome data, although microbiome-specific methods may be preferred. Since many microbiome features are rare or zero-inflated, taking the tree relations among features into account can be beneficial.

2. VISUALIZATION AND EXPLORATION

Visualization is one of the first steps in exploring a new data set and also is a key component of the presentation of study results in publications. Since microbiome data are high dimensional, classic plots for univariate data such as box plots or histograms cannot provide a useful overview, as they can only be used once specific taxa of interest have been identified. Stacked barplots, which break down the composition of each sample, are useful, but they become cumbersome when the number of samples is large. Ordination plots, which project the observations into the plane such that samples with more similar profiles are located closer together, are the most effective tool to provide an overview of variation across samples and visual cues regarding outliers or potential issues such as batch effects.

2.1. Microbiome Ordination Methods

The most commonly used ordination method for microbiome data is principal coordinates analysis (PCoA). PCoA is a classic multivariate analysis method and is also known as metric multidimensional scaling. Its goal is to translate pairwise distances or dissimilarities between data points, which are based on the original p input variables, into a lower-dimensional projection where points that are close in the high-dimensional space are similarly close in the resulting projection.

We now provide more details, following the groundbreaking work of Gower (1966). We begin with an $n \times n$ matrix $\mathbf{D}$, where entry d_{ij} summarizes the pairwise distance between samples i and j for each of the n samples. The PCoA projection is computed by the following steps:

1. Transform the pairwise distances d_{ij} to obtain a similarity matrix $\mathbf{A}$: $a_{ij} = -d_{ij}^2/2$.
2. Normalize $\mathbf{A}$ to get the centered matrix $\mathbf{G}$: $\mathbf{G} = (\mathbf{I} - \mathbf{1}\boldsymbol{s}')\mathbf{A}(\mathbf{I} - \mathbf{1}\boldsymbol{s}')$, where $\boldsymbol{s} = n^{-1}\mathbf{1}$, $\mathbf{I}$ is the identity matrix, and $\mathbf{1}$ is a column vector of ones.
3. Calculate the eigendecomposition of $\mathbf{G}$.

4. Project the n data points into the plane defined by the eigenvectors corresponding to the two leading eigenvalues.

The choice of distance metric to use in computing $\mathbf{D}$ is critical. Standard Euclidean distance is unsuitable for high-dimensional count-based data with many zeros, so a variety of distances have been proposed that are appropriate for count-based data and can integrate information on the presence versus absence or abundance of features based on their location within the phylogenetic tree. These distances are also known as β diversity metrics. Popular choices include the following:

- Jaccard dissimilarity, which focuses on the shared presence or absence of features (Jaccard 1900);
- Bray–Curtis dissimilarity, which considers the extent of overlap in the counts (Bray & Curtis 1957);
- unweighted UniFrac, which uses the fraction of shared branch lengths in the phylogenetic tree (Lozupone & Knight 2005); and
- weighted UniFrac, which reflects the relative abundances of features in the tree (Lozupone et al. 2007).

The development of distance metrics remains an active area of research, with recent proposals including double PCoA distance, which allows for a more general representation of the pairwise relations between features via a $p \times p$ matrix $\mathbf{Q}$ (Fukuyama 2019), and Wasserstein distance, which reflects statistical uncertainty in the observed counts (Wang et al. 2021). **Figure 2** shows a stacked barplot and corresponding PCoA plot generated using weighted UniFrac distances for an example data set.

In recent years, there have been several new proposals regarding microbiome ordination methods. For many studies, batch or site effects may introduce unwanted variation in the microbiome profiles; Shi et al. (2020) proposed adjusted PCoA (aPCoA) to allow for covariate adjustment in creating the data projection, assuming a linear dependence on covariates. Wang et al. (2022b) developed a more general method for covariate adjustment (adjustment for confounding factors using principal coordinates analysis, or AC-PCoA) which uses a kernel matrix to summarize the

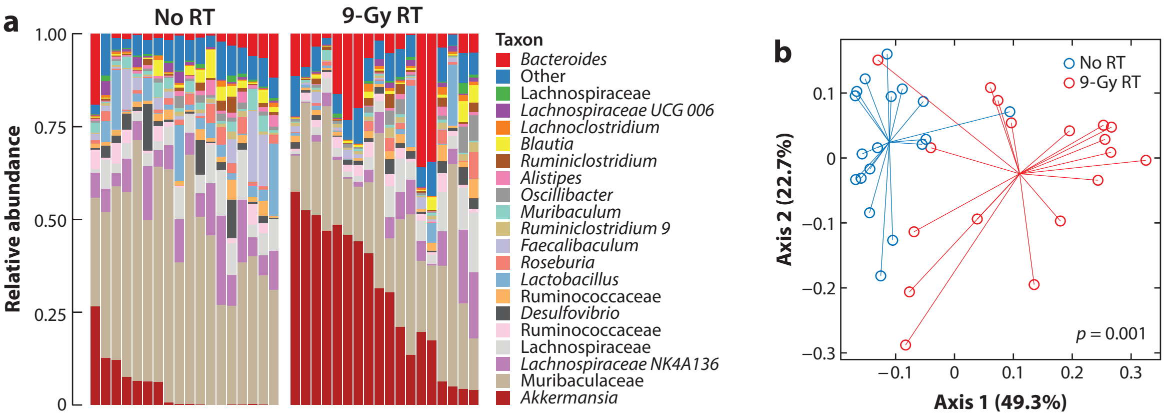

Figure 2

Example stacked barplot of genus-level abundances (*a*) and PCoA plot (*b*) contrasting the microbiome composition of mice that were exposed to radiation (9-Gy RT) to those that were not (no RT). The PCoA projection was obtained using weighted UniFrac. Figure adapted with permission from Schwabkey et al. (2022). Abbreviations: PCoA, principal coordinates analysis; RT, radiation therapy.

similarity in covariate values. The resulting covariate-adjusted distances obtained using either method can be used as input to downstream analysis methods such as permutation testing or distance-based clustering.

More broadly, there is an increasing interest in the application of machine learning methods for ordination. In particular, uniform manifold approximation and projection (UMAP), a nonlinear embedding technique popular for single-cell data, can be used to construct visually appealing plots (McInnes et al. 2018). Since UMAP does not preserve global structure, the relative distances between points cannot be directly interpreted, but it may be advantageous in highlighting clusters as a complement to PCoA (Armstrong et al. 2021).

2.2. Unsupervised Clustering Approaches

After visualization, a common next step in the microbiome analysis pipeline is unsupervised clustering, which aims to discover naturally occurring groups of samples in the data. These clusters may align with groups of points that are nearby in the PCoA projection, particularly when using distance-based clustering approaches.

2.2.1. Distance-based methods. Classic approaches for distance-based clustering include k-means, which aims to minimize the distances between points within a cluster and the cluster mean, and k-medoids (Kaufman & Rousseeuw 1990), also known as partitioning around medoids, which minimizes the distance to a central data point. Hierarchical clustering, which constructs a tree in which samples that are close in distance are grouped together, offers another distance-based clustering option.

As in the PCoA projection, the results of distance-based clustering depend heavily on the choice of distance metric. In practice, users may explore different distance metrics and select one that works well for their data. To better characterize aspects of the input data that might lead to a preference for one distance metric over another, Shi et al. (2022) reported that high-abundance features tend to drive the performance of clustering using Bray–Curtis dissimilarity, while differences in the presence of zeros in the leaf nodes shaped clustering results when using unweighted UniFrac. They recommend a weighted combination of these metrics to obtain a robust clustering that reflects patterns of similarity for both high- and low-abundance features.

2.2.2. Model-based methods. Parametric methods, which assume that the microbiome counts arise from a particular statistical distribution, may also be applied. In early work on model-based clustering of microbiome data, Holmes et al. (2012) relied on the assumption that the distribution of the observed counts within each cluster or community i followed a multinomial distribution with parameter $\boldsymbol{p}_i$, with entries p_{ij} that reflect the probability of a count belonging to feature j in community i. Their method, the Dirichlet multinomial mixture model, allows for the parameter $\boldsymbol{p}_i$ for each community to follow a Dirichlet prior distribution.

Shi et al. (2023) recently developed a method for sparse clustering of microbiome data, which offers some key advantages: It allows the number of clusters to be learned from the data, it integrates information encoded in the phylogenetic tree structure, and it incorporates feature selection to identify features that are relevant to the cluster assignments. To achieve the first objective, Shi et al. rely on a Bayesian approach known as a mixture of finite mixtures model. In this framework, the number of clusters M is a random variable, with a probability mass function defined on the counting numbers. Conditional on M, the vector $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_M)$, which sums to unity, represents the probability that a sample is allocated to a given cluster. The samples in each cluster are assumed to arise from a distribution with cluster-specific parameters $\theta_1, \ldots, \theta_M$.

To incorporate information on the phylogenetic tree structure, Shi et al. (2023) rely on the Dirichlet-tree distribution as the within-cluster density, so that the samples in cluster m arise from

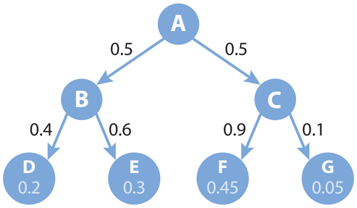

Figure 3

Illustrative example of the Dirichlet-tree distribution. Figure adapted with permission from Shi et al. (2023).

a Dirichlet-tree distribution with parameters θ_m specific to that cluster. A simple illustration of the Dirichlet-tree distribution is provided in **Figure 3**, which shows the allocation probabilities for a simple tree with three levels. The probabilities labeling each arrow are the branch probabilities b_{jk} leading from a parent node j to a child node k. For a single count, the probability of being assigned to a leaf node is the product of the b_{jk} values over the branches from the root node to that leaf. This distribution reflects the phylogenetic tree structure, with parameters capturing the sequential branch points in the tree, rather than focusing only on the leaf nodes.

Finally, Shi et al. (2023) incorporate feature selection within this framework, using a Bayesian variable selection approach to identify features that are useful in discriminating between the clusters. The authors let the latent variable $\gamma_j \in \{0, 1\}$ represent the binary inclusion of feature j. Nonselected nodes correspond to features that are not informative in clustering, and have the same vector of probabilities leading to their child nodes across all clusters, while selected nodes have cluster-specific allocation probabilities. This modeling framework allows the identification of the level of the tree where the sample composition begins to differentiate between clusters. In closely related work, Mao & Ma (2022) proposed a Bayesian clustering approach using the Dirichlet-tree multinomial distribution as the within-cluster density, but with a Dirichlet process prior for sample clustering.

To allow more flexible clustering, where features may be differential across some but not all clusters, Zhou et al. (2022) propose a biclustering model, which enables clustering of both microbiome features and samples, i.e., clustering of both the rows and columns in the data matrix. Their method is based on a hierarchical Bayesian matrix factorization approach, where the observed counts are assumed to come from a Dirichlet multinomial distribution, with parameters that depend on the latent clusters.

3. DIFFERENTIAL ABUNDANCE ANALYSIS

The methods discussed in the previous section allow for unsupervised exploration of a data set. In many scientific studies, there may be an interest in identifying features that differ between known groups, such as treatment versus control. This is known as differential abundance testing.

3.1. Differential Abundance Tests

The microbiome field has not reached a consensus on the best approach for differential abundance testing, and a variety of methods are in active use. The most basic is to rarefy the observed counts, or downsample the data so that each observation has the same number of sequences, and then carry out cross-group comparisons using familiar methods such as the t or Mann–Whitney test. While simple, rarefaction is suboptimal since there is a loss of information on the actual number of sequences observed, potentially reducing statistical power.

A variety of more advanced methods have been proposed to avoid this loss of information and to better reflect aspects of microbiome data such as compositionality and zero inflation.

Biclustering: simultaneous clustering of the rows and columns of a matrix

Rarefaction: subsampling the observed counts in each sample without replacement

alr: additive log ratio

clr: centered log ratio

Microbiome-specific differential abundance testing approaches include metagenomeSeq (Paulson et al. 2013), ALDEx2 (Fernandes et al. 2014), ANCOM (analysis of compositions of microbiomes) and ANCOM-II (Mandal et al. 2015, Kaul et al. 2017), ANCOM-BC (ANCOM with bias correction) (Lin & Peddada 2020), and MaAsLin2 (microbiome multivariable associations with linear models) (Mallick et al. 2021). Methods originally designed for RNA-seq data, such as DESeq2 (Love et al. 2014), are also widely applied to microbiome data.

Unfortunately, different methods will yield different hits when applied to the same data set. Recently, Nearing et al. (2022) systemically compared the existing toolbox of methods and found that the choice of preprocessing, filtering, and testing approaches resulted in a wide variation in findings. They also noted that intrinsic features of a data set, such as sample size, the extent of zero inflation, and sequencing depth, influence method performance. Although no single approach was preferred in every scenario, ALDEx2 and ANCOM-II, while on the conservative side, generally produced the most consistent results. We discuss these methods in more detail in the next few paragraphs. Nonetheless, the authors recommend adopting a consensus approach based on the results from multiple methods to ensure robust feature identification, rather than relying on a single preferred method.

Both ALDEx2 and ANCOM-II adopt the compositional data analysis framework. More specifically, they rely on log ratio transformations, originally proposed in the foundational work of Aitchison (1982, 1986) to map values from a constrained space to the real space. Log ratio transformations include the additive log ratio, centered log ratio, and isometric log ratio (Egozcue et al. 2003). Given a p-vector $\boldsymbol{x}$ with a compositional constraint, the additive log ratio (alr) transform maps $\boldsymbol{x}$ to an unconstrained vector in $\mathbb{R}^{p-1}$:

$$\mathrm{alr}(\boldsymbol{x}) = \left\{\log\frac{x_1}{x_p}, \ldots, \log\frac{x_{p-1}}{x_p}\right\},$$

where x_p, the last component of the vector, is arbitrarily taken as the reference component. The centered log ratio (clr) transform takes the geometric mean of $\boldsymbol{x}$ as its reference value and is defined as

$$\mathrm{clr}(\boldsymbol{x}) = \left\{\log\frac{x_1}{g(\boldsymbol{x})}, \ldots, \log\frac{x_p}{g(\boldsymbol{x})}\right\},$$

where $g(\boldsymbol{x}) = (\prod_{j=1}^{p} x_{ij})^{1/p}$. Finally, the isometric log ratio satisfies the property of isometry, which means that it preserves the distances between the pairwise log ratios.

We now summarize the ALDEx2 procedure (Fernandes et al. 2014), which entails the following steps: performing Monte Carlo sampling from a Dirichlet distribution to obtain a vector of posterior probabilities for the relative abundance of each feature in each sample, applying a centered log ratio transform to these probabilities, applying a t or Mann–Whitney test to compare these values across conditions, and adjusting the resulting p-values using the Benjamini–Hochberg procedure (Benjamini & Hochberg 1995). ALDEx2 acknowledges the compositionality of the data by focusing on relative abundances, or proportions within a sample; uncertainty regarding these proportions given the feature counts and sequencing depth is reflected through the Monte Carlo sampling.

Instead of relying on the centered log ratio transform, ANCOM (Mandal et al. 2015) adopts the additive log ratio transform to handle compositionality, which they argue helps focus on meaningful cross-group differences relative to a reference background value. ANCOM-II (Kaul et al. 2017) additionally integrates explicit handling of zero values; this version of the procedure seeks to differentiate between different sources of zero values and identify features as having differential abundance if the proportion of zero values differs across groups. Recently, ANCOM-BC

(Lin & Peddada 2020) extended ANCOM-II to allow for differences in the sampling fraction across samples.

Although these methods were developed primarily with 16S data in mind, they are also applicable to feature tables generated from WMS. As WMS data may be even sparser than 16S data, the choice of a method that can handle zero inflation may be beneficial. Finally, in real-world studies, it may be important to control for covariates; methods such as ANCOM-BC and MaAsLin2 that allow for covariate adjustment would be preferred.

False discovery rate (FDR): the proportion of significant hits that are truly null

3.2. Controlling False Discoveries

Given the large number of features in microbiome data sets, careful consideration of multiplicity is critical to avoid too many false discoveries. As a basic approach, controlling the false discovery rate (FDR) using the Benjamini–Hochberg method (Benjamini & Hochberg 1995) is popular. Filtering out low-prevalence features before testing can help avoid overpenalizing in the multiple testing adjustment.

FDR-controlling methods that can account for the structure of hypotheses under consideration have also been developed, including the p-filter and TreeBH procedures. The p-filter (Barber & Ramdas 2017, Ramdas et al. 2019) allows the incorporation of preferences on the importance of different hypotheses, the prior probability of hypotheses being nonnull, grouping of hypotheses, and the dependence between hypotheses. The TreeBH procedure (Bogomolov et al. 2021) focuses specifically on the setting with a hierarchical organization of hypotheses and aims to control a selective version of the FDR that accounts for sequential testing steps to zoom in on interesting branches of the tree.

The p-filter and TreeBH methods can be applied directly to the p-values resulting from differential abundance testing. FDR-controlling methods have also been proposed for the multivariate regression setting (discussed in more detail in Section 4 below). In particular, methods based on the knockoff filter (Barber & Candès 2015) have been proposed; in this procedure, knockoff variables that resemble the original features in terms of their correlation but have no association with the outcome are generated. These artificial variables serve as a negative control and allow for the selection of a data-adaptive threshold to control the FDR. The original knockoff filter assumed a single set of variables with $n > p$; more recent work has enabled the application of this framework to the microbiome context. Specifically, the multilayer knockoff filter (Katsevich & Sabatti 2019) enables control of the FDR at multiple levels of resolution, while the compositional knockoff filter (Srinivasan et al. 2021) offers FDR control for high-dimensional compositional predictors.

To assess the robustness of cross-group differences both globally and for individual features, L. Zhang et al. (2021a) proposed progressively permuting the group labels and repeating the differential abundance test on the partially and fully permuted data sets. Features with robust signal will remain significant on partially permuted data, while more fragile hits will quickly drop out. Globally, data sets with strong signal will exhibit a much higher number of hits relative to those in fully permuted data. This approach can be applied as a sanity check on the differential abundance results from any of the methods described in Section 3.1.

4. REGRESSION MODELING

In this section, we discuss statistical approaches to address three key goals in microbiome research: characterizing the relationship between microbiome features and biological or clinical outcomes, identifying biological or environmental factors that influence microbiome composition, and discovering the role of the microbiome in shaping causal relationships between an exposure and an outcome. These scientific questions can be framed as regression models where microbiome features serve as the predictor, response, or mediator.

4.1. Microbiome as Predictor

The primary challenge in regression modeling with microbiome features as the covariates lies in the compositionality; since each sample has a unit-sum constraint, the full set of p features will exhibit perfect multicollinearity, violating one of the basic assumptions of ordinary linear regression. In addition to handling this constraint, another key objective for this class of regression models is to leverage information on the phylogenetic or taxonomic relationships among the covariates. This motivates the development of models that utilize the tree information to inform either joint selection or aggregation of related features.

4.1.1. Linear log-contrast model.

Let $\boldsymbol{y} \in \mathbb{R}^n$ be n observations of a variable we wish to predict, and let $\mathbf{X} \in \mathbb{R}_+^{n\times p}$ be a matrix in which entry X_{ij} gives the relative abundance of feature j in sample i. Due to the sum constraint, the components of each observation must lie in a $(p-1)$-dimensional simplex $\mathcal{S}^{p-1} = \{(x_1, \dots, x_p) : x_j > 0 \text{ for } j = 1, \dots, p, \sum_{j=1}^{p} x_j = 1\}$. Since the predictors cannot be interpreted independently, standard linear regression models are inappropriate for modeling the relationship between microbial features and outcomes of interest. A natural way to handle this is to transform the predictors to a Euclidean space to allow for the application of standard statistical models. Regression modeling is done by the following steps:

1. Transform the predictor matrix $\mathbf{X}$ using any linear transformation that transforms the compositional simplex to a real Euclidean space. Popular choices include the additive log ratio, centered log ratio, and isometric log ratio transformations. We use the centered log ratio transform for further demonstration.
2. For observation i, we can define the regression model as

   $$\begin{aligned} y_i &= \operatorname{clr}(x_{i1})\beta_1 + \operatorname{clr}(x_{i2})\beta_2 + \cdots + \operatorname{clr}(x_{ip})\beta_p + \varepsilon_{ij} \\ &= \beta_1 \log(x_{i1}) + \beta_2 \log(x_{i2}) + \cdots + \beta_p \log(x_{ip}) - (1/p)\sum_{j=1}^{p}\beta_j \sum_{j=1}^{p}\log(x_{ij}) + \varepsilon_{ij}. \end{aligned}$$

   In vector form, this can be written

   $$\boldsymbol{y} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \text{subject to } \sum_{j=1}^{p}\beta_j = 0, \tag{1}$$

   where $\mathbf{Z} = \log \mathbf{X}$, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$ is the corresponding p-vector of regression coefficients, and $\boldsymbol{\varepsilon}$ is an n-vector of independent noise distributed as $\mathcal{N}(0, \sigma^2)$. The model in Equation 1 can be solved as a constrained linear regression problem.

4.1.2. Penalized compositional regression model.

Although the linear log-contrast model is useful in addressing the compositionality constraint, it was developed in the context of experiments involving a limited number of predictors. In the context of microbiome data, the large number of features $p \gg n$ necessitates the use of sparse modeling approaches to identify relevant features. To address the high dimensionality of the predictors, Lin et al. (2014) propose to solve the l_1 regularized, constrained convex optimization problem

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta}\left(\frac{1}{2n}||\boldsymbol{y} - \mathbf{Z}\boldsymbol{\beta}||_2^2 + \lambda||\boldsymbol{\beta}||_1\right), \qquad \text{subject to } \sum_{j=1}^{p}\beta_j = 0, \tag{2}$$

where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$, $\lambda > 0$ is a regularization parameter, and $||.||_1$ denotes the l_1 norm. This estimator has attractive properties including permutation invariance, which means that it does not depend on the ordering of the predictors, and selection invariance, which means that it is unaffected by the omission of predictors that are not relevant to the outcome.

Although Lin et al. (2014) address the challenges of compositionality and high dimensionality, their method does not reflect the tree-structured organization of the predictors. To fill this gap, Shi et al. (2016) propose a variable selection procedure based on the compositional regression model described in Equation 2, but modified to allow for subcompositions within the taxonomic tree, which reflect the relative abundances of child nodes for a given parent node. For example, the abundances of the species that belong to a genus in the tree form a subcomposition with their own sum constraint. Shi et al. propose a linear regression model that links the response $\boldsymbol{y}$ to the subcompositions and impose linear constraints to reflect the sum constraint within each subcomposition.

4.1.3. Bayesian variable selection in the compositional setting.

The regression models developed by Aitchison (1982) and Aitchison & Bacon-Shone (1984) rely on transformations of the compositional variables that generally require one of the original variables to be dropped. L. Zhang et al. (2021b) designed a Bayesian modeling approach that offers some key advantages: It incorporates variable selection to identify features that are relevant for outcome prediction; it proposes a generalized transformation, which avoids dropping variables while simultaneously satisfying the permutation and selection invariance properties; and it incorporates information on the phylogenetic similarity of the predictors within the feature selection framework. To fulfill the first objective, the authors define a latent indicator $\gamma_i \in \{0, 1\}$ that represents the inclusion of the ith covariate in the model, with the p-vector $\boldsymbol{\gamma}$ used to represent the feature selection. Conditional on $\boldsymbol{\gamma}$, the response vector $\boldsymbol{y}$ is assumed to follow a multivariate normal distribution

$$\boldsymbol{y} \mid \boldsymbol{\gamma}, \boldsymbol{\beta}_{\gamma}, \sigma^2 \sim \mathcal{N}_n(\mathbf{Z}_{\gamma}\boldsymbol{\beta}_{\gamma}, \sigma^2\mathbf{I}_n),$$

where $\mathbf{Z}_{\gamma}$ denotes a modified version of the $\mathbf{Z}$ matrix, defined in Equation 2, including only those columns corresponding to nonzero entries in $\boldsymbol{\gamma}$, and $\boldsymbol{\beta}_{\gamma}$ represents the corresponding coefficients for the selected covariates. The coefficient vector $\boldsymbol{\beta}_{\gamma}$ has length $p_{\gamma} = \sum_i \gamma_i$. To incorporate the zero-sum constraint on the parameters, they propose a novel z-prior for $\boldsymbol{\beta}_{\gamma}$ conditional on $\boldsymbol{\gamma}$:

$$\boldsymbol{\beta}_{\gamma} \mid \boldsymbol{\gamma}, \sigma^2, \tau^2 \sim \mathcal{N}(\mathbf{0}, \sigma^2\tau^2(\mathbf{T}_{\gamma}^{'}\mathbf{T}_{\gamma})^{-1}),$$

where the $(p + 1) \times p$ transformation matrix $\mathbf{T} = [\mathbf{I}_p\, c\mathbf{1}_p]'$ represents a contrast transformation matrix that satisfies the condition that the sum of columns in the matrix $(\mathbf{T}'\mathbf{T})^{-1}$ is zero, c is a constant, and $\mathbf{1}_p$ is a p-dimensional column vector of ones. The matrix $\mathbf{T}_{\gamma}$ consists of the columns of the generalized transformation $\mathbf{T}$ corresponding to the selected variables, that is, the nonzero entries of $\boldsymbol{\gamma}$. The idea behind the multivariate prior on the parameters is to ensure that the sum of a random draw of $\boldsymbol{\beta}_{\gamma}$ is zero. This is attained as $\mathrm{var}(\sum_{j\in\gamma} \beta_j)$ goes to zero as c becomes large, which implies that more shrinkage is imposed on $\sum_{j\in\gamma} \beta_j$. Finally, to incorporate the relatedness of the covariates within the Bayesian variable selection framework, the authors place an Ising prior on the latent variable selection indicator $\boldsymbol{\gamma}$:

$$\mathrm{P}(\boldsymbol{\gamma}) \propto \exp(\boldsymbol{a}'\boldsymbol{\gamma} + \boldsymbol{\gamma}'\mathbf{Q}\boldsymbol{\gamma}).$$

The shrinkage parameters $\boldsymbol{a}$, which take negative values, influence the sparsity of $\boldsymbol{\gamma}$. The entries in the structural parameter $\mathbf{Q}$ influence the strength of association between the selection of features i and j. The larger q_{ij} is, the more likely it is under the prior that the ith and jth covariates will be jointly selected. In sum, this model offers a nice framework for handling both the compositionality and structural relations among the predictors. However, it does not directly address the challenge of zero-inflated or rare features.

4.1.4. Tree aggregated models.

As noted in Section 1.2, a high degree of zero inflation is an intrinsic aspect of microbiome data. A common strategy to reduce zero inflation is to collapse or aggregate rare features (leaf nodes in the tree) and focus instead on internal nodes. A basic approach is to group finer-resolution strains into higher levels of taxonomic resolution by summing over all the features that belong to the corresponding classification in a taxonomic tree $\mathcal{T}$. For example, feature counts might be summed to the genus level prior to analysis. Suppose we obtain a new aggregated feature for the ith subject $x_{i,a} = x_{i,1} + \cdots + x_{i,k}$, where k denotes the number of leaf nodes descending from the ancestor node a. As noted by Yan & Bien (2021), in the linear model setting, $x_{i,a}\beta = (x_{i,1} + \cdots + x_{i,k})\beta = x_{i,1}\beta + \cdots + x_{i,k}\beta$. Effectively, this means that learning a model where some features have exactly equal coefficients β corresponds to aggregating the original features into less zero-inflated groupings. The question then becomes how to estimate these βs in a flexible manner, to allow grouping of rare features when the data support their having equivalent effects on the outcome. Our above insight on aggregation over subtrees of a taxonomic or phylogenetic tree tells us that if our estimate of β is constant within the subtrees of $\mathcal{T}$, then that corresponds to a regression model with tree-aggregated features. Bien et al. (2021) propose the new aggregated features to be the log of the geometric mean of counts within subtrees having constant β values. The corresponding linear log-contrast model with collapsed features reduces to solving the optimization problem

$$\min_{\alpha \in \mathbb{R}^{|\mathcal{T}|-1}} \left(\frac{1}{2n} ||y - \log(\text{geom}(X; \mathcal{T}))\boldsymbol{\alpha}||_2^2 + \lambda \sum_{u \in \mathcal{T}-\{r\}} w_u |\alpha_u| \right), \text{subject to} \sum_{m=1}^{|\mathcal{T}|-1} \alpha_m = 0, \qquad 3.$$

where $\text{geom}(X; \mathcal{T}) \in \mathbb{R}^{n\times(|\mathcal{T}|-1)}$ is a matrix in which each column corresponds to the geometric mean of all base-level taxa counts within the subtree rooted at u, which is a nonroot node of $\mathcal{T}$, and $\boldsymbol{\alpha}$ represents the vector of regression coefficients for the aggregated features. Equation 3 expresses a constrained linear regression problem with a weighted l_1 norm penalty in which the regularization parameter λ determines the overall trade-off between prediction error on the training data and the optimal level of feature aggregation. In summary, this model allows for simultaneous data-adaptive aggregation and regression modeling, thus directly addressing the challenge of predictive modeling with rare features.

4.1.5. Kernel regression methods. The models discussed in the previous sections assume a linear relationship between the compositional covariates and the outcomes of interest. In addition, variable selection and subcompositional coherence are the major focus of these constrained log-linear models. In a different approach, Zhao et al. (2015) propose a kernel association test, MiRKAT (microbiome regression-based kernel association test), that employs a semiparametric kernel machine regression framework for regressing the outcome on the microbiome profiles. Kernel machine regression is a nonparametric method that relates similarity of outcome values to similarity of predictor profiles. MiRKAT presents a hypothesis-testing framework for assessing the global association between microbiome composition and a continuous outcome, via kernels that incorporate phylogenetic distances between individuals' microbiome profiles. The semiparametric kernel regression model is formulated as

$$\boldsymbol{y} = \mathbf{U}\boldsymbol{\eta} + f(\mathbf{X}) + \boldsymbol{\varepsilon},$$

where $\mathbf{U} \in \mathbb{R}^{n\times q}$ represents the matrix of additional covariates of interest, such as age, sex, or other clinical variables, and $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_q)'$ is the corresponding vector of regression coefficients for the q covariates. The relationship between the microbiome and the outcome variable is characterized by the function f. Testing that there is no association between microbiome composition and the outcome is equivalent to testing $f(\mathbf{X}) = 0$. Under the kernel machine regression framework, $f(\boldsymbol{x}_i)$ lies in a reproducing kernel Hilbert space $\mathcal{H}_k$, generated from a positive definite kernel function K such that $f(\boldsymbol{x}_i) = \sum_{i'=1}^{n} \alpha_{i'} K(\boldsymbol{x}_i, \boldsymbol{x}_{i'})$, for some weights $\alpha_1, \ldots, \alpha_n$. The kernel matrix $\mathbf{K}_{n\times n}$ defines the pairwise similarities between samples, with the element (i, i') set to $K(\boldsymbol{x}_i, \boldsymbol{x}_{i'})$. For microbiome applications, a kernel matrix incorporating distances appropriate for microbiome data (see Section 2.1) can be defined as

$$\mathbf{K} = -\frac{1}{2}\left(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}'}{n}\right)\mathbf{D}^2\left(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}'}{n}\right),$$

where $\mathbf{D} = \{d_{ij}\}$ is the pairwise distance matrix (e.g., weighted or unweighted UniFrac distance or the Bray–Curtis dissimilarity), $\mathbf{I}$ is the identity matrix, and $\mathbf{1}$ is a vector of ones. In a seminal paper, Liu et al. (2007) showed that there is a strong connection between kernel regression and linear mixed effects models, whereby $f(\mathbf{X})$ can be written as a subject-specific random effect that follows a distribution with mean 0 and variance $\tau\mathbf{K}$. Therefore, testing for $f(\mathbf{X}) = 0$ is equivalent to testing the null hypothesis that $H_0 : \tau = 0$. Under the mixed model framework, this can be done using the score statistic

$$Q = \frac{1}{2\hat{\sigma}_0^2}(\boldsymbol{y} - \hat{\boldsymbol{y}}_0)\mathbf{K}(\boldsymbol{y} - \hat{\boldsymbol{y}}_0),$$

where $\hat{\boldsymbol{y}}_0$ and $\hat{\sigma}_0^2$ are the predicted mean of $\boldsymbol{y}$ and the estimated residual variance under the null model. The variable Q can be shown to asymptotically follow a weighted mixture of χ^2 distributions under H_0, so the p-value can be obtained analytically. The power of MiRKAT depends heavily on the choice of an appropriate kernel. To avoid the burden of choosing a perfect kernel, which requires a priori knowledge of the microbiome data association, the authors define optimal MiRKAT, a kernel machine regression framework capable of incorporating multiple kernel choices. Optimal MiRKAT performs the score test individually for each kernel, selects the minimum p-value, and then adjusts this using a multiple comparison technique. In recent years, MiRKAT has been further extended to incorporate binary, survival, multivariate, and several other outcome types; a detailed overview of these methods is provided by Wilson et al. (2021). To better deal with irregular or noisy outcome measurements (e.g., excess zeros or outliers), Wang et al. (2022a) recently proposed a new association analysis tool named MiRKAT-IQ within the MiRKAT framework using integrated quantile regression, which employs quantile-based methods in place of mean-based association analysis methods.

4.2. Microbiome as Response

In the previous section, we focused on the goal of predicting a response variable from microbiome profiling data; in that setting, the primary statistical challenge lay in handling compositionality and rare features and integrating structural information on relations among the predictors. In contrast, models that aim to predict microbiome profiles as the response need to grapple with the choice of an appropriate distribution to model the observed microbiome count data. As in the model-based clustering methods discussed in Section 2.2.2, the Dirichlet multinomial model offers one practical option.

4.2.1. Dirichlet multinomial models. Let $X = (X_1, \ldots, X_p)'$ denote the random vector of counts on p bacterial taxa, with $M = \sum_{j=1}^{p} X_j$. The most natural model for describing multivariate count data is the multinomial probability mass function,

$$f(\boldsymbol{x}; \boldsymbol{\pi}, m) = \frac{m!}{\prod_{j=1}^{p} x_j!}\prod_{j=1}^{p}\pi_j^{x_j},$$

where $\boldsymbol{x} = (x_1, \ldots, x_p), m = \sum_{j=1}^{p} x_j$, and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_p)$ is the vector of probabilities with $\sum_{j=1}^{p}\pi_j = 1$. One of the main issues of microbiome composition data is overdispersion, i.e., the increased variation arising due to heterogeneity of the microbiome samples. Since the multinomial distribution fails to account for overdispersion owing to its assumption of fixed underlying proportions $\boldsymbol{\pi}$, the standard convention is to assume $\boldsymbol{\pi}$ to be a random vector having a Dirichlet probability density function

$$f(\boldsymbol{\pi}, \boldsymbol{\tau}) = \frac{\Gamma(\sum_{j=1}^{p} \tau_j)}{\prod_{j=1}^{p} \Gamma(\tau_j)} \prod_{j=1}^{p} \pi_j^{\tau_j - 1},$$

where $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_p)$ is a vector of positive scalars, $\tau_j > 0$. The posterior distribution of $\boldsymbol{x}$ has the Dirichlet multinomial mass function

$$f(\boldsymbol{x}; \boldsymbol{\tau}, m) = \frac{\Gamma(m+1)\Gamma(\sum_{j=1}^{p} \tau_j)}{\Gamma(m + \sum_{j=1}^{p} \tau_j)} \prod_{j=1}^{p} \frac{\Gamma(x_j + \sum_{j=1}^{p} \tau_j)}{\Gamma(x_j + 1)\Gamma(\tau_j)}. \qquad 4.$$

To understand the effect of confounding covariates, Chen & Li (2013) linked the parameters τ_j to an $n \times q$-dimensional covariate matrix $\mathbf{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_q)$ via a log linear representation

$$\log\left(\tau_j(\boldsymbol{u}_i)\right) = \eta_j + \sum_{l=1}^{q} \beta_{jl} u_{il}, \qquad 5.$$

where $\boldsymbol{u}_i$ represents the ith row vector of $\mathbf{U}$ and η_j is an intercept term. When the number of covariates q is large, Chen & Li propose a penalized regression model based on Equation 4 to improve model interpretability. Since β_{jl} measures the effect of the lth covariate on the jth taxa, determining the nonzero β_{jl} parameters is equivalent to identifying the significant associations between taxa and covariates in the models in Equations 4 and 5. Using the log-linear Dirichlet multinomial setup, Wadsworth et al. (2017) placed spike and slab mixture priors on the $\boldsymbol{\beta}$ parameters to allow for variable selection.

Several studies have shown that the Dirichlet multinomial setup is inadequate for modeling microbiome data with complex dependence structures (O'Brien et al. 2016, Tang et al. 2017), since it imposes a negative correlation among taxon counts when these correlations could be positive. Furthermore, it ignores the fact that microbial compositions are associated via a phylogenetic tree. To address these limitations, Wang & Zhao (2017) proposed a Dirichlet multinomial tree distribution that enables modeling of phylogenetic tree-based covariance structures. Ostner et al. (2021) proposed a Bayesian model for tree-aggregated amplicon and single-cell compositional data analysis that integrates phylogenetic information and experimental covariate information into the generative modeling of microbiome data. Another major concern is the presence of a large number of zero-inflated taxa. Tang & Chen (2019) proposed the zero-inflated generalized Dirichlet multinomial, which extends the Dirichlet multinomial model for handling excess zeros while dealing with the complex dispersion patterns found in microbiome data sets. Koslovsky (2023) proposed a Bayesian zero-inflated Dirichlet multinomial regression model with sparsity-inducing priors that allows variable selection for high-dimensional covariates.

4.3. Microbiome as Mediator

Very often in biomedical studies, the relationship between two important covariates may be due to the indirect effect of an intermediary variable that is influenced by the independent variable. For example, it is well known that diet and obesity are interlinked; this raises the question of whether intake of fat and other macronutrients affects BMI via the composition of the gut microbiome. In statistics, mediation analysis seeks to understand the unobserved relationship between independent and dependent variables that may have been caused by the influence of a third, mediator, variable. Mediation theory proposes that the independent variable influences a mediator, which in turn affects the dependent variable; therefore, understanding the role of mediation clarifies the exact nature of the relationship between the independent and dependent variables.

4.3.1. Compositional mediation model.

We begin with a description of the single mediator model. For subject i, let t_i be a treatment, x_i a single mediator, y_i an outcome, and $\boldsymbol{u}_i$ a set of pretreatment variables that may affect the treatment, mediator, and outcome. Mathematically, mediation can be formulated and implemented within the framework of linear structural equation models (LSEMs):

$$x_i = a_0 + at_i + \boldsymbol{u}_i'\boldsymbol{b} + \eta_{1i}, \qquad 6.$$

$$y_i = c_0 + ct_i + bx_i + \boldsymbol{u}_i'\boldsymbol{g} + \eta_{2i},$$

where η_{1i} and η_{2i} denote the respective disturbance variables for x_i and y_i. As illustrated in **Figure 4**, under this model, the effect of the treatment on the outcome, transmitted through the mediator x_i, is called the direct effect and is quantified by the path coefficient c. There also exist two indirect effects: the effect of the treatment on the mediator quantified by a and the effect of the mediator on the outcome quantified by the path coefficient b. The purpose of this model is to estimate these causal direct and indirect effects, and to test whether the mediator variable plays a role in defining the relationship between the treatment and the outcome. Due to the compositionality and high dimensionality of the microbiome mediators, the standard single mediation model cannot be directly used in microbiome mediation analysis. Sohn & Li (2019) proposed to restructure the model in Equation 6, incorporating multiple mediators $\boldsymbol{x}_i$ and using the additive log ratio transformation (Aitchison 1982). For a composition vector $\boldsymbol{\zeta}$, the additive log ratio transformation of $\boldsymbol{\zeta}$ is defined as $\text{alt}(\boldsymbol{\zeta}) = \big(\log(\zeta_1/\zeta_p), \log(\zeta_2/\zeta_p), \ldots, \log(\zeta_{p-1}/\zeta_p)\big)'$. Extending the model in Equation 6 for a vector of compositional mediators $\boldsymbol{x_i}$ and taking the additive log ratio on both sides of the model in Equation 6, we get

$$\text{alt}(\boldsymbol{x}_i) = \text{alt}(\boldsymbol{a}_0) + t_i\text{alt}(\boldsymbol{a}) + \sum_{r=1}^{q} u_{ir}\text{alt}(\boldsymbol{b}) + \text{alt}(\boldsymbol{\eta}_{1i}), \qquad 7.$$

$$y_i = c_0 + ct_i + \text{alt}(\boldsymbol{x}_i)'\boldsymbol{b}_{-k} + \boldsymbol{u}_i'\boldsymbol{g} + \eta_{2i},$$

where $\boldsymbol{b}_{-k} = (b_1, \ldots, b_{k-1})$. The model for y_i is adjusted to accommodate the linear constraint $\boldsymbol{b}'\mathbf{1}_k = 0$ in order to account for the compositional nature of $\boldsymbol{X}_i$. This constrained form of the compositional regression model has already been explained in Section 4.1.2. For clarity, in this section we have described the basic version of the compositional mediation model with a single predictor and mediator; however, interested readers can find the more general version of the model with multiple mediators and pretreatment covariates described in Sohn & Li (2019). Estimation of the model components follows the ideas of Lin et al. (2014), and the estimated regression parameters are then used to test total and component-wise mediation effects. In recent work by Sohn et al. (2022), the mediation model has been extended to allow binary response variables.

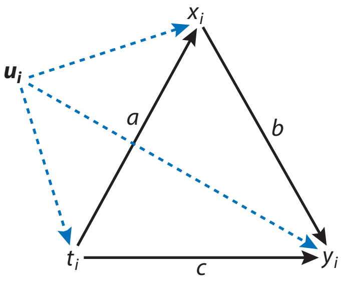

Figure 4

Illustrative example of a single-mediator model where a, b, and c are the path coefficients. Figure adapted from Sohn & Li (2019) with the permission of the Institute of Mathematical Statistics.

The sparse compositional regression model for microbiome mediation analysis is primarily designed to test the overall mediation effect and does not focus on selecting individual taxa that mediate the path between the treatment and the outcome. To allow for effective variable selection, H. Zhang et al. (2021) proposed a screening algorithm within the LSEM framework using a closed testing-based selection procedure. Their model uses the isometric log ratio transform on the compositional mediators, and the LSEM framework is then used to test for the joint mediation effects.

5. NETWORK INFERENCE

The methods discussed in Section 4 adopt a regression-based framework, in which one set of variables plays the role of predictor and another the response. In this section, we discuss network inference approaches that aim to characterize the interdependencies among a set of variables. In the context of microbiome data, these dependence networks seek to capture ecological relationships among the microbiome features, such as mutualism (interactions where each species has a net benefit) or competition (interactions where species compete for the same resources).

5.1. Correlation-Based Approaches

A naïve approach to assessing the dependence among features, by computing the Pearson or Spearman correlation, has two key problems. First, since these measures of correlation do not account for compositionality, spurious negative associations may be observed. Second, given the high-dimensional nature of the data, a large number of nonzero correlations will be identified, making visualization and interpretation of any relationships discovered challenging.

The first issue can be addressed within the compositional data framework; for example, SparCC (sparse correlations for compositional data) (Friedman & Alm 2012) seeks to estimate correlations between log-transformed features, leveraging estimates of the variance in the log ratio between two features to approximate their correlation. CCLasso (Fang et al. 2015) incorporates an ℓ_1 penalty on the sum of the off-diagonal elements of the correlation matrix to encourage sparsity in the solution and ensures that the resulting correlation matrix is positive definite with entries in the interval $[-1, 1]$. By encouraging a sparse solution, they partially address the large number of nonzero correlations; nonetheless, the correlations may still represent indirect relationships. The graphical modeling approaches discussed in the next subsection offer an advantage in this regard, as they seek to focus on direct connections.

5.2. Graphical Models

Graphical models seek to identify direct dependence between features using a graph structure $G = (V, E)$, where V represents the set of vertices and E represents the set of edges. In a statistical setting, the vertices V correspond to random variables, and an undirected edge (i, j) represents that variables i and j are dependent, conditional on the other observed features in the data set. Much of the work on graphical models is centered on the Gaussian setting due to its computational convenience. Suppose our data matrix $\mathbf{X}$ follows the multivariate normal distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}^{-1})$. Here, $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ is the inverse covariance matrix, also known as the precision matrix. The multivariate normal distribution has a very nice property, which is that two variables i and j are conditionally independent if and only if ω_{ij}, the corresponding entry in $\boldsymbol{\Omega}$, equals 0. Popular methods for graphical modeling such as the graphical lasso (Friedman et al. 2008) seek to identify a graph structure by imposing sparsity on the precision matrix; the pattern of zero and nonzero values in $\hat{\boldsymbol{\Omega}}$ can be translated into a graph structure, with nonzero elements corresponding to selected edges. By focusing on conditional dependence, rather than marginal or pairwise correlation, relations between two features that correspond to an indirect association should drop out.

Microbiome count data are non-Gaussian, but many graphical modeling approaches for microbiome data build on the Gaussian graphical model framework by employing data transformations or assuming latent variables that are normally distributed. The most popular approach for microbiome network inference is SPIEC-EASI (sparse inverse covariance estimation for ecological association inference) (Kurtz et al. 2015), which adopts a centered log ratio transform followed by inference of a Gaussian graphical model on the transformed variables using ℓ_1 penalized methods. SPIEC-EASI deals with the presence of zeros in the data by adding a small pseudocount prior to the centered log ratio transform.

Given the high degree of zero inflation in many microbiome data sets, methods that directly model the excess zeros may be more appropriate. Recently proposed methods that explicitly handle the presence of zeros include HARMONIES (hybrid approach for microbiome network inferences via exploiting sparsity) (Jiang et al. 2020) and COZINE (compositional zero-inflated network estimation) (Ha et al. 2020). The first accounts for zeros by assuming a zero-inflated negative binomial distribution on the observed count data; a network is then inferred using a Gaussian graphical model on the log of the normalized abundance estimates. COZINE adopts a hurdle model to account for excess zeros and applies a centered log ratio transform to the nonzero abundances. In estimating the final network, the authors allow for multiple types of relations: binary-binary (i.e., conditional dependence in the presence/absence patterns), binary-continuous, and continuous-continuous interactions. The authors illustrate the proposed method with an application to oral microbiome profiles for patients starting cancer treatment. The resulting ecological network is shown in **Figure 5**. The inferred dependence relationships are of interest as chemotherapy treatment places patients at risk of infection, sores, and inflammation in the mouth, and these may be related to the composition of the oral microbiome.

5.3. Integrative Graphical Models

Finally, there is an interest in understanding the relationships between microbiome profiles and high-dimensional covariates. Osborne et al. (2022) propose an integrative network model, which estimates a network among the microbiome features and performs variable selection to identify covariates that influence the microbial abundances. Incorporating covariates in the model not only allows for the identification of key factors that influence microbial populations but also enables estimation of a sparser network. In particular, if the abundances of two microbial features are correlated due to their shared dependence on an external covariate, a graphical model learned on the microbial features only would infer an edge between these two nodes, while the joint model including both microbiome and covariate data would be able to identify that these features have shared dependence on another variable, rather than a direct connection. This idea is illustrated schematically in **Figure 6**.

6. DISCUSSION

Microbiome research is an exciting and actively evolving field in terms of the scientific questions being posed, the scale and types of data being collected, and the statistical tools used to analyze the resulting data. Future issues of interest including spatial mapping, single-cell data, and the design of microbiome interventions are highlighted below.

An ongoing challenge in research on statistical methodology for the analysis of microbiome data is how to best evaluate and compare the performance of methods. Real-world applications are important but, aside from tightly controlled synthetic experiments, lack a ground truth for performance evaluation. Simulation is a standard tool for the evaluation of novel methods, but the

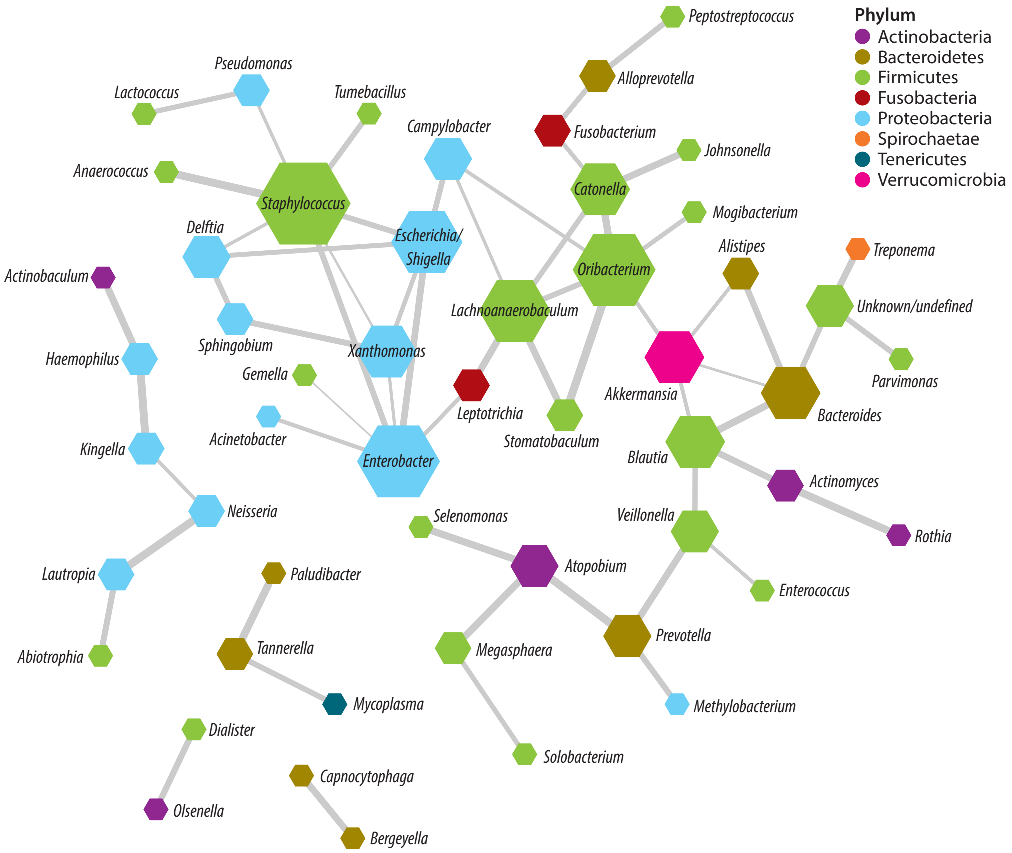

Figure 5

COZINE (compositional zero-inflated network estimation) network inference result. Node size is proportional to the node degree, and edge width is proportional to the edge stability. Figure adapted with permission from Ha et al. (2020).

simulation of realistic microbiome data, especially with interesting dependence structure, remains challenging. Recently, there have been several proposals for improving such simulations, including parametric methods that rely on reference data sets to generate realistic data (Ma et al. 2021) and methods based on machine learning approaches such as generative adversarial networks (Rong et al. 2021).

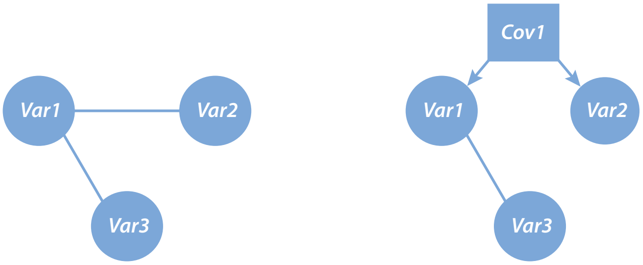

Figure 6

Schematic illustration of dependence on an external covariate.

FUTURE ISSUES

1. The spatial distribution of microorganisms within an environment can highlight ecological niches and improve understanding of the potential interactions between microbes within a community based on their spatial colocalization. Recent breakthroughs in spatial profiling technology, including improvements in techniques for imaging and sequencing-based spatial profiling, unlock new opportunities for biomedical and statistical research in this area.
2. Zooming in more closely, single-cell data may provide the opportunity to better study bacteria with very low abundance; for example, single-cell profiling can support the identification of antibiotic-resistant strains within a larger population, while dual profiling of host and microbiome single-cell RNA abundance can be applied to address questions such as how gene expression differs for infected versus noninfected cells. Interestingly, single-cell data share many structural similarities with bulk metagenomic data, with zero inflation as a hallmark of both data types.
3. The design of microbiome therapeutics is another exciting area of research. Rational combination therapies, which combine specially selected beneficial microorganisms, may also include metabolites or nutrients to support the engraftment of the strains included. Statistical methods are needed to identify promising combinations based on functional characterization, clinical associations, and ecological relations among microorganisms.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

C.B.P. and S.S. are partially supported by National Institutes of Health/National Heart, Lung, and Blood Institute (NIH/NHLBI) R01HL158796. K.A.D. is partially supported by National Institutes of Health/National Cancer Institute Cancer Center Support Grant (NIH/NCI CCSG) P30CA016672, MD Anderson Moon Shot Programs, Prostate Cancer Specialized Program of Research Excellence (SPORE) P50CA140388, Center for Clinical and Translational Sciences (CCTS) 5UL1TR000371, and Cancer Prevention and Research Institute of Texas (CPRIT) RP160693.

LITERATURE CITED

Aitchison J. 1982. The statistical analysis of compositional data. *J. R. Stat. Soc. Ser. B* 44(2):139–60

Aitchison J. 1986. *The Statistical Analysis of Compositional Data*. London: Chapman and Hall

Aitchison J, Bacon-Shone J. 1984. Log contrast models for experiments with mixtures. *Biometrika* 71(2):323–30

Armstrong G, Martino C, Rahman G, Gonzalez A, Vázquez-Baeza Y, et al. 2021. Uniform manifold approximation and projection (UMAP) reveals composite patterns and resolves visualization artifacts in microbiome data. *mSystems* 6(5):e0069121

Barber RF, Candès EJ. 2015. Controlling the false discovery rate via knockoffs. *Ann. Stat.* 43(5):2055–85

Barber RF, Ramdas A. 2017. The *p*-filter: multilayer false discovery rate control for grouped hypotheses. *J. R. Stat. Soc. Ser. B* 79(4):1247–68

Beghini F, McIver LJ, Blanco-Míguez A, Dubois L, Asnicar F, et al. 2021. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *eLife* 10:e65088

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* 57(1):289–300

Bien J, Yan X, Simpson L, Müller CL. 2021. Tree-aggregated predictive modeling of microbiome data. *Sci. Rep.* 11:14505

Blanco-Míguez A, Beghini F, Cumbo F, McIver LJ, Thompson KN, et al. 2023. Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nat. Biotechnol.* 41(11):1633–44

Bogomolov M, Peterson CB, Benjamini Y, Sabatti C. 2021. Hypotheses on a tree: new error rates and testing strategies. *Biometrika* 108(3):575–90

Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, et al. 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* 37(8):852–57

Bray JR, Curtis JT. 1957. An ordination of the upland forest communities of southern Wisconsin. *Ecol. Monogr.* 27(4):326–49

Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. 2016. DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* 13(7):581–83

Chen J, Li H. 2013. Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis. *Ann. Appl. Stat.* 7(1):418–42

Cryan JF, O'Riordan KJ, Sandhu K, Peterson V, Dinan TG. 2020. The gut microbiome in neurological disorders. *Lancet Neurol.* 19(2):179–94

Douglas GM, Maffei VJ, Zaneveld JR, Yurgel SN, Brown JR, et al. 2020. PICRUSt2 for prediction of metagenome functions. *Nat. Biotechnol.* 38(6):685–88

Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, Barcelo-Vidal C. 2003. Isometric logratio transformations for compositional data analysis. *Math. Geol.* 35(3):279–300

Fang H, Huang C, Zhao H, Deng M. 2015. CCLasso: correlation inference for compositional data through Lasso. *Bioinformatics* 31(19):3172–80

Fernandes AD, Reid JN, Macklaim JM, McMurrough TA, Edgell DR, Gloor GB. 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* 2:15

Fierer N. 2017. Embracing the unknown: disentangling the complexities of the soil microbiome. *Nat. Rev. Microbiol.* 15(10):579–90

Friedman J, Alm EJ. 2012. Inferring correlation networks from genomic survey data. *PLOS Comput. Biol.* 8(9):e1002687

Friedman JH, Hastie TJ, Tibshirani RJ. 2008. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* 9(3):432–41

Fukuyama J. 2019. Emphasis on the deep or shallow parts of the tree provides a new characterization of phylogenetic distances. *Genome Biol.* 20(1):131

Gopalakrishnan V, Spencer CN, Nezi L, Reuben A, Andrews M, et al. 2018. Gut microbiome modulates response to anti–PD-1 immunotherapy in melanoma patients. *Science* 359(6371):97–103

Gower JC. 1966. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* 53(3–4):325–38

Ha M, Kim J, Galloway-Peña J, Do K, Peterson CB. 2020. Compositional zero-inflated network estimation for microbiome data. *BMC Bioinformatics* 21:581

Holmes I, Harris K, Quince C. 2012. Dirichlet multinomial mixtures: generative models for microbial metagenomics. *PLOS ONE* 7(2):e30126

Jaccard P. 1900. Contribution au problème de l'immigration post-glaciaire de la flore alpine. *Bull. Soc. Vaudoise Sci. Nat.* 36:87–130

Jansson JK, Hofmockel KS. 2020. Soil microbiomes and climate change. *Nat. Rev. Microbiol.* 18(1):35–46

Jiang S, Xiao G, Koh AY, Chen Y, Yao B, et al. 2020. HARMONIES: a hybrid approach for microbiome networks inference via exploiting sparsity. *Front. Genet.* 11:445

Katsevich E, Sabatti C. 2019. Multilayer knockoff filter: Controlled variable selection at multiple resolutions. *Ann. Appl. Stat.* 13(1):1–33

Kaufman L, Rousseeuw P. 1990. *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: Wiley

Kaul A, Mandal S, Davidov O, Peddada SD. 2017. Analysis of microbiome data in the presence of excess zeros. *Front. Microbiol.* 8:2114

Koslovsky MD. 2023. A Bayesian zero-inflated Dirichlet-multinomial regression model for multivariate compositional count data. *Biometrics*. **https://doi.org/10.1111/biom.13853**

Kurtz ZD, Müller CL, Miraldi ER, Littman DR, Blaser MJ, Bonneau RA. 2015. Sparse and compositionally robust inference of microbial ecological networks. *PLOS Comput. Biol.* 11(5):e1004226

Lin H, Peddada SD. 2020. Analysis of compositions of microbiomes with bias correction. *Nat. Commun.* 11:3514

Lin W, Shi P, Feng R, Li H. 2014. Variable selection in regression with compositional covariates. *Biometrika* 101(4):785–97

Liu D, Lin X, Ghosh D. 2007. Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. *Biometrics* 63(4):1079–88

Lloyd-Price J, Arze C, Ananthakrishnan AN, Schirmer M, Avila-Pacheco J, et al. 2019. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* 569(7758):655–62

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15:550

Lozupone C, Hamady M, Kelley S, Knight R. 2007. Quantitative and qualitative diversity measures lead to different insights into factors that structure microbial communities. *Appl. Environ. Microbiol.* 73(5):1576–85

Lozupone C, Knight R. 2005. UniFrac: a new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.* 71(12):8228–35

Lu J, Rincon N, Wood DE, Breitwieser FP, Pockrandt C, et al. 2022. Metagenome analysis using the Kraken software suite. *Nat. Protoc.* 17(12):2815–39

Ma S, Ren B, Mallick H, Moon YS, Schwager E, et al. 2021. A statistical model for describing and simulating microbial community profiles. *PLOS Comput. Biol.* 17(9):e1008913

Mallick H, Rahnavard A, McIver LJ, Ma S, Zhang Y, et al. 2021. Multivariable association discovery in population-scale meta-omics studies. *PLOS Comput. Biol.* 17(11):e1009442

Mandal S, Van Treuren W, White RA, Eggesbø M, Knight R, Peddada SD. 2015. Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microb. Ecol. Health Dis.* 26:27663

Mao J, Ma L. 2022. Dirichlet-tree multinomial mixtures for clustering microbiome compositions. *Ann. Appl. Stat.* 16(3):1476–99

McInnes L, Healy J, Melville J. 2018. UMAP: uniform manifold approximation and projection for dimension reduction. arXiv:1802.03426 [stat.ML]

Nearing JT, Douglas GM, Hayes MG, MacDonald J, Desai DK, et al. 2022. Microbiome differential abundance methods produce different results across 38 datasets. *Nat. Commun.* 13(1):342

O'Brien JD, Record NR, Countway P. 2016. The power and pitfalls of Dirichlet-multinomial mixture models for ecological count data. bioRxiv 045468. **https://doi.org/10.1101/045468**

Osborne N, Peterson CB, Vannucci M. 2022. Latent network estimation and variable selection for compositional data via variational EM. *J. Comput. Graph. Stat.* 31(1):163–75

Ostner J, Carcy S, Müller CL. 2021. tascCODA: Bayesian tree-aggregated analysis of compositional amplicon and single-cell data. *Front. Genet.* 12:766405

Paulson JN, Stine OC, Bravo HC, Pop M. 2013. Differential abundance analysis for microbial marker-gene surveys. *Nat. Methods* 10(12):1200–2

Ramdas AK, Barber RF, Wainwright MJ, Jordan MI. 2019. A unified treatment of multiple testing with prior knowledge using the *p*-filter. *Ann. Stat.* 47(5):2790–821

Riquelme E, Zhang Y, Zhang L, Montiel M, Zoltan M, et al. 2019. Tumor microbiome diversity and composition influence pancreatic cancer outcomes. *Cell* 178(4):795–806

Rong R, Jiang S, Xu L, Xiao G, Xie Y, et al. 2021. MB-GAN: microbiome simulation via generative adversarial network. *GigaScience* 10(2):giab005

Schwabkey ZI, Wiesnoski DH, Chang CC, Tsai WB, Pham D, et al. 2022. Diet-derived metabolites and mucus link the gut microbiome to fever after cytotoxic cancer treatment. *Sci. Transl. Med.* 14(671):eabo3445

Shi P, Zhang A, Li H. 2016. Regression analysis for microbiome compositional data. *Ann. Appl. Stat.* 10(2):1019–40

Shi Y, Zhang L, Do K, Jenq RR, Peterson CB. 2023. Sparse tree-based clustering of microbiome data to characterize microbiome heterogeneity in pancreatic cancer. *J. R. Stat. Soc. Ser. C* 72(1):20–36

Shi Y, Zhang L, Do K, Peterson CB, Jenq RR. 2020. aPCoA: covariate adjusted principal coordinates analysis. *Bioinformatics* 36(13):4099–101

Shi Y, Zhang L, Peterson CB, Do K, Jenq RR. 2022. Performance determinants of unsupervised clustering methods for microbiome data. *Microbiome* 10:25

Sohn MB, Li H. 2019. Compositional mediation analysis for microbiome studies. *Ann. Appl. Stat.* 13(1):661–81

Sohn MB, Lu J, Li H. 2022. A compositional mediation model for a binary outcome: application to microbiome studies. *Bioinformatics* 38(1):16–21

Srinivasan A, Xue L, Zhan X. 2021. Compositional knockoff filter for high-dimensional regression analysis of microbiome data. *Biometrics* 77(3):984–95

Tang ZZ, Chen G. 2019. Zero-inflated generalized Dirichlet multinomial regression model for microbiome compositional data analysis. *Biostatistics* 20(4):698–713

Tang ZZ, Chen G, Alekseyenko AV, Li H. 2017. A general framework for association analysis of microbial communities on a taxonomic tree. *Bioinformatics* 33(9):1278–85

Tara Ocean Found., Tara Oceans, Eur. Mol. Biol. Lab. (EMBL), Eur. Marine Biol. Resour. Cent. Eur. Res. Infrastruct. Consort. (EMBRC-ERIC). 2022. Priorities for ocean microbiome research. *Nat. Microbiol.* 7(7):937–47

Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. 2007. The human microbiome project. *Nature* 449(7164):804–10

Wadsworth WD, Argiento R, Guindani M, Galloway-Peña J, Shelburne SA, Vannucci M. 2017. An integrative Bayesian Dirichlet-multinomial regression model for the analysis of taxonomic abundances in microbiome data. *BMC Bioinformatics* 18:94

Wang S, Cai TT, Li H. 2021. Optimal estimation of Wasserstein distance on a tree with an application to microbiome studies. *J. Am. Stat. Assoc.* 116(535):1237–53

Wang T, Ling W, Plantinga AM, Wu MC, Zhan X. 2022a. Testing microbiome association using integrated quantile regression models. *Bioinformatics* 38(2):419–25

Wang T, Zhao H. 2017. A Dirichlet-tree multinomial regression model for associating dietary nutrients with gut microorganisms. *Biometrics* 73(3):792–801

Wang Y, Sun F, Lin W, Zhang S. 2022b. AC-PCoA: adjustment for confounding factors using principal coordinate analysis. *PLOS Comput. Biol.* 18(7):e1010184

Wilson N, Zhao N, Zhan X, Koh H, Fu W, et al. 2021. MiRKAT: kernel machine regression-based global association tests for the microbiome. *Bioinformatics* 37(11):1595–97

Yan X, Bien J. 2021. Rare feature selection in high dimensions. *J. Am. Stat. Assoc.* 116(534):887–900

Zhang H, Chen J, Feng Y, Wang C, Li H, Liu L. 2021. Mediation effect selection in high-dimensional and compositional microbiome data. *Stat. Med.* 40(4):885–96

Zhang L, Shi Y, Do K, Peterson CB, Jenq RR. 2021a. ProgPerm: progressive permutation for a dynamic representation of the robustness of microbiome discoveries. *BMC Bioinformatics* 22:126

Zhang L, Shi Y, Jenq R, Do K, Peterson C. 2021b. Bayesian compositional regression with structured priors for microbiome feature selection. *Biometrics* 77(3):824–38

Zhao N, Chen J, Carroll IM, Ringel-Kulka T, Epstein M, et al. 2015. Testing in microbiome-profiling studies with MiRKAT, the microbiome regression-based kernel association test. *Am. J. Hum. Genet.* 96(5):797–807

Zhou F, He K, Li Q, Chapkin RS, Ni Y. 2022. Bayesian biclustering for microbial metagenomic sequencing data via multinomial matrix factorization. *Biostatistics* 23(3):891–909