REVIEW
published: 14 June 2022
doi: 10.3389/fams.2022.884810

A Survey of Statistical Methods for Microbiome Data Analysis

Kevin C. Lutz[1], *Shuang Jiang*[2,3], *Michael L. Neugent*[4], *Nicole J. De Nisco*[4], *Xiaowei Zhan*[3]* *and Qiwei Li*[1]*

[1] Department of Mathematical Sciences, The University of Texas at Dallas, Richardson, TX, United States, [2] Department of Statistical Science, Southern Methodist University, Dallas, TX, United States, [3] Department of Population and Data Sciences, The University of Texas Southwestern Medical Center, Dallas, TX, United States, [4] Department of Biological Sciences, The University of Texas at Dallas, Richardson, TX, United States

In the last decade, numerous statistical methods have been developed for analyzing microbiome data generated from high-throughput next-generation sequencing technology. Microbiome data are typically characterized by zero inflation, overdispersion, high dimensionality, and sample heterogeneity. Three popular areas of interest in microbiome research requiring statistical methods that can account for the characterizations of microbiome data include detecting differentially abundant taxa across phenotype groups, identifying associations between the microbiome and covariates, and constructing microbiome networks to characterize ecological associations of microbes. These three areas are referred to as differential abundance analysis, integrative analysis, and network analysis, respectively. In this review, we highlight available statistical methods for differential abundance analysis, integrative analysis, and network analysis that have greatly advanced microbiome research. In addition, we discuss each method's motivation, modeling framework, and application.

Keywords: microbiome data, metagenomics data, differential abundance analysis, integrative analysis, network analysis

OPEN ACCESS

Edited by:
Li Xing,
University of Saskatchewan, Canada

Reviewed by:
Padhmanand Sudhakar,
KU Leuven, Belgium
Jie Yang,
University of Illinois at Chicago,
United States

****Correspondence:***
Xiaowei Zhan
xiaowei.zhan@utsouthwestern.edu
Qiwei Li
qiwei.li@utdallas.edu

Specialty section:
This article was submitted to Statistics and Probability, a section of the journal Frontiers in Applied Mathematics and Statistics

Received: *27 February 2022*
Accepted: *16 May 2022*
Published: *14 June 2022*

Citation:
Lutz KC, Jiang S, Neugent ML, De Nisco NJ, Zhan X and Li Q (2022) A Survey of Statistical Methods for Microbiome Data Analysis. Front. Appl. Math. Stat. 8:884810. doi: 10.3389/fams.2022.884810

1. INTRODUCTION

Bacteria, viruses, fungi, and other microscopic living things are referred to as microorganisms or microbes. The term *microbiome* describes the collective genomes of the microorganisms or the microorganisms themselves [1]. The human microbiome plays a vital role in controlling vital functions in the body such as immune system development, protection against pathogens, and modulation of the central nervous system [2]. The microbiome is dynamic and changes with factors such as diet or the use of antibiotics [3]. Changes in the microbiome may affect host health and cause disease [2, 4]. In the last decade, many advances in sequencing technology and statistical methodology have made it possible to study and quantify the microbiome.

In quantitative microbiome research, there are three popular areas of interest that seek to detect and quantify (i) differentially abundant taxa across phenotype groups, (ii) associations between taxonomies and covariates, and (iii) associations between taxa in the whole microbiome network. These three areas are referred to as differential abundance analysis, integrative analysis, and network analysis, respectively. Many useful methods have been developed to perform these downstream analyses while taking multiple issues into consideration that arise from differences in sequencing technology such as 16S ribosomal RNA sequencing (16S rRNA) or metagenomic shotgun sequencing (MSS) [5], technical issues of sequencing technology [6], the complex nature of sequencing count data [7], and choice of data normalization techniques [8].

16S rRNA and MSS are commonly used high-throughput sequencing technologies that generate raw count data for microbiome statistical analysis. Both technologies have their advantages and disadvantages. In 16S rRNA, the 16S ribosomal gene sequence is useful for the identification and classification of bacteria and archaea [9] because it is conservative as well as found in most microbes [10] and contains multiple sequences of the gene within a single microbe [11]. 16S rRNA is a relatively short sequence in the bacterial genome. Their sequences can be clustered as operational taxonomic units (OTUs) or amplicon specific variants, which better classify bacteria at the phyla and genera levels but is less precise at the species level. Further, 16S rRNA has available reference genomes and pipelines to perform data analysis such as DADA2, Mothur, and QIIME [12–15]. In contrast, MSS targets entire genomes with greater resolution giving it the capability to efficiently classify bacteria at the species level as well as describe microbial communities and their functional differences [16]. MSS identifies far more species per read than 16S rRNA and is more advanced because it can also identify viruses, fungi, and protozoa [12]. As a result, MSS is more costly per sample than 16S rRNA and so sample sizes tend to be smaller in studies with MSS data [17]. Of course, these are not the only existing sequencing methods. RNA-Seq, ChIP-Seq, and MeDIP-Seq are some of the many sequencing technologies that are also available.

Microbiome count data have characteristics that pose numerous challenges to methodology such as zero inflation, overdispersion, high dimensionality, and sample heterogeneity [18, 19]. Further, when count data are transformed to compositional data (i.e., total sum scaling), the counts in each sample are only relative to each taxon and do not necessarily reflect absolute abundance [20] due to variable sequencing depth across samples [5]. Zero inflation is common where possibly up to 90% of all counts are zeros [20]. Further, MSS count data are typically much more sparse than 16S rRNA data [5]. Some of the zeros are true zeros and others are false zeros. False zeros result from technical variability and limitations in sequencing depth when taxa with low abundance are completely missed at random [8, 16]. Quality of DNA preparations such as inconsistencies in the DNA extraction or how samples are handled can also contribute to technical variability [16, 18]. Library size is the total number of reads per sample (i.e., the sum of all the counts in a sample). Different library sizes (i.e., sample heterogeneity) result as a consequence of technical variability, which make it difficult to compare the samples. Samples with greater library size could contain higher reads for non-differentially abundant features, which would lead to the spurious conclusion that those features are differentially abundant [8]. Batch effects are problematic and may lead to spurious conclusions especially with MSS data, which is generated over multiple sequencing runs [18]. Furthermore, biological, technical, and computational factors are probable sources of batch effects [21]. Normalization alone does not fully correct for batch effects [22]. Many statistical methods are available that correct for batch effects including linear mixed models *via* the LIMMA package in R [23], metagenomeSeq in the Bioconductor software for users in R [24], Bayesian Dirichlet-multinomial regression meta-analysis (BDMMA) [25], surrogate variable analysis (SVA) [26], and remove unwanted variation (RUV2 and RUV4) [27]. Other methods that help to remove batch effects include batch mean centering (BMC) [28], ComBat [29], or its extension ComBat-seq [22], removeBatchEffect [23], FAbatch [30], RUVIII [31], percentile normalization [32], and singular value decomposition (SVD) [33]. Assumptions such as whether or not the batch effect is known or if the design is balanced must be considered when selecting an appropriate method to correct for batch effects. Wang and LêCao [21] provide a detailed decision tree that is helpful for identifying appropriate statistical methods that correct for batch effects.

In this paper, we first briefly describe the data one would typically encounter in microbiome data analysis and introduce their notations in **Table 1**. The sources referenced in this paper vary in their data notations and so we present notations in a consistent manner throughout this paper. Then, we discuss available methods for differential abundance analysis, integrative analysis, and network analysis. Specifically, we introduce the methods, applications, motivations, normalization techniques, models, statistical tests, and provide a brief discussion. **Table 2** provides the classes and methods of statistical analyses discussed in this paper.

2. DIFFERENTIAL ABUNDANCE ANALYSIS

Microbial dysbiosis, or microbial imbalances, is related to disease. Microbiota have been implicated in the development of numerous diseases such as colorectal cancer [34], type 2 diabetes [35], liver cirrhosis [36], and inflammatory bowel disease [37]. The method of detecting differentially abundant taxa across phenotype groups is known as differential abundance analysis. Identifying differentially abundant taxa will help to understand the relationship between the symbiotic organism and human health as well as identify microbial biomarkers for disease screening. We discuss multiple methods for differential abundance analysis in this section including edgeR [38], metagenomeSeq [24], DESeq2 [39], analysis of compositions of microbiomes or ANCOM [40], a zero-inflated beta model or ZIBSeq [5], a zero-inflated generalized Dirichlet-multinomial model or ZIGDM [6], and count regression for correlated observations with a beta-binomial model or corncob [41]. The summary of these methods are found in **Table 3** and their implementations for users in R are in **Table 4**.

The common biological motivation of each method is to determine if any particular features of $\boldsymbol{Y}_{n\times p}$ are significantly different with respect to phenotype $\boldsymbol{z}_{n\times 1}$ in a high-dimensional setting where the number of features is much greater than the number of samples (i.e., $p \gg n$). Additionally, each method has its own statistical motivations. edgeR was motivated by the need to separate biological and technical variability in order to reduce bias when testing for significant phenotypic differences attributed to abundances of RNA-Seq data. Sparsity (i.e., zero inflation) is a common characteristic of MSS count data, which is one motivational factor for metagenomeSeq, ZIBSeq, and ZIGDM. For example, a particular bacterial species may be present in a small percentage of samples for both biological and technical

TABLE 1 | The notations and descriptions of typical microbiome data.

Label	Notation	Description
Count data	$\boldsymbol{Y}_{n\times p}$	A $n \times p$ matrix of count data where each element $y_{ij} \in \mathbb{N}$ is the abundance for sample $i = 1, \ldots, n$ and feature $j = 1, \ldots, p$. Denote $\boldsymbol{y}_{i\cdot} = (y_{i1}, \ldots, y_{ip})$ as the $1 \times p$ row vector of counts across all p features in sample i. Also, denote $\boldsymbol{y}_{\cdot j} = (y_{1j}, \ldots, y_{nj})^\top$ as the $n \times 1$ column vector of counts for feature j in all n samples. Denote $\tilde{y}_{ij}$ as relative abundance, $\check{y}_{ij} = \log_2(y_{ij} + c)$ as a log-transformed count with an added pseudo-value c, and library size $N_i = \sum_{j=1}^{p} y_{ij}$.
Covariates	$\boldsymbol{X}_{n\times q}$	A $n \times q$ matrix where each element $x_{ik} \in \mathbb{R}$ is a measure for covariate $k = 1, \ldots, q$ in sample i. Denote $\boldsymbol{x}_{i\cdot} = (x_{i1}, \ldots, x_{iq})$ as the $1 \times q$ row vector of covariate measures across all q covariates in sample i. Also, denote $\boldsymbol{x}_{\cdot k} = (x_{1k}, \ldots, x_{nk})^\top$ as the $n \times 1$ column vector of measures for covariate k in all n samples.
Phenotype	$\boldsymbol{z}_{n\times 1}$	A $n \times 1$ column vector for the phenotypic response, which is written as $(z_1, \ldots, z_n)^\top$ where each element z_i is the phenotypic response for sample i. The response is categorical* where $z_i = g$ indicates the phenotypic group of each sample for group $g = 1, \ldots, G$.

Some models may have a continuous phenotypic response where each $z_i \in \mathbb{R}$.

TABLE 2 | Classes and alphabetized methods of statistical analyses discussed in this paper.

Differential abundance	Longitudinal differential abundance	Integrative analysis	Network analysis
ANCOM	maSigPro	DMBVS	CCLasso
corncob	MetaDprof	DMLMbvs	HARMONIES
DESeq2	MetaLonDA	DMR	REBACCA
edgeR	MetaSplines	IntegrativeBayes	SparCC
metagenomeSeq	mixMC		SpiecEasi
mixMC	MetaLonDA		SPRING
ZIBSeq	NBMM		
ZIGDM	NBZIMM		
ZINB-DPP			

TABLE 3 | Summary of methods for differential abundance analysis in microbiome studies.

Method	Model assumption	Normalization	References	Availability
edgeR*	Negative binomial	TMM	[42]	Bioconductor
metagenomeSeq	Zero-inflated normal or log-normal	CSS	[24]	Bioconductor
DESeq2*	Negative binomial	RLE	[39]	Bioconductor
ANCOM	ANOVA	ALR	[40]	GitHub
ZIBseq	Zero-inflated beta	TSS	[5]	CRAN
ZIGDM	Zero-inflated generalized Dirichlet-multinomial	None†	[6]	CRAN
corncob	Beta-binomial	None†	[41]	GitHub
mixMC	PCA/sPLS-DA‡	CSS/TSS+CLR	[43]	Bioconductor
maSigPro*	Generalized linear models	User specified††	[44]	Bioconductor
NBME*	Negative binomial mixed effects	User specified††	[45]	CRAN
MetaSplines	Gaussian + SS-ANOVA	CSS	[46]	Bioconductor
MetaDprof	Gaussian + SS-ANOVA	TMM	[47]	Online
MetaLonDA	Negative binomial + SS-ANOVA	TMM/CSS‡‡	[48]	CRAN
NBZIMM	Negative binomial or Gaussian mixed effects	See below**	[49]	GitHub

Developed for RNA-Seq data analysis.

**For these zero-inflated models, the negative binomial mixed effects model can incorporate the counts directly so normalization is not required; whereas, the Gaussian mixed effects model requires the arcsine square root transformation of the compositional data.*

†*These models do not require data normalization.*

††*These models require the user to normalize the data beforehand.*

‡*Principal components analysis (PCA) and sparse partial least squares discriminant analysis (sPLS-DA).*

‡‡*Median-of-ratios scaling factor is a third normalization technique available in MetaLonDA.*

TABLE 4 | Implementation in R for differential abundance analysis in microbiome studies.

Method	Implementation	Updated
edgeR	https://bioconductor.org/packages/release/bioc/html/edgeR.html	2021
metagenomeSeq	https://rdrr.io/bioc/metagenomeSeq/	2021
DESeq2	https://bioconductor.org/packages/release/bioc/html/DESeq2.html	2021
ANCOM	https://github.com/FrederickHuangLin/ANCOM	2020
ZIBseq	https://cran.r-project.org/web/packages/ZIBseq/index.html	2017
ZIGDM	https://www.rdocumentation.org/packages/miLineage/versions/2.1	2017
corncob	https://github.com/bryandmartin/corncob	2021
mixMC	https://www.bioconductor.org/packages/release/bioc/html/mixOmics.html	2022
maSigPro	https://www.bioconductor.org/packages/release/bioc/html/maSigPro.html	2021
NBME	https://cran.r-project.org/web/packages/timeSeq/index.html	2019
MetaSplines	https://rdrr.io/bioc/metagenomeSeq/	2019
MetaDprof	https://cals.arizona.edu/~anling/sbg/software.htm	2016
MetaLonDA	https://cran.r-project.org/web/packages/MetaLonDA/index.html	2020
NBZIMM	https://github.com/nyiuab/NBZIMM	2022

reasons. Biologically, this particular bacterial species may be found in only a small percentage of samples. Limitations in technology such as sequencing depth may miss a particular bacterial species with low abundance completely at random. As a result, the number of zero counts becomes inflated. DESeq2 wanted a model that can account for the presence of outliers and small replicate sizes while producing interpretable results. Other methods including ZIBSeq, ANCOM, ZIGDM, and corncob were motivated by the need for models that can also account for the compositional nature of the count data. ZIGDM was further motivated by the need to account for correlation structure and dispersion patterns amongst features.

Normalization is a transformation needed for robust analysis that accounts for the challenges of microbiome data and technical variability of sequencing technology [8, 16, 18, 20]. Measurable, informative, and direct comparisons of samples are only possible after normalization [8]. Some methods for normalization are based on either sample-specific scaling of the raw counts or replacing the raw counts with normalized counts [16]. Popular normalization methods include but are not limited to total sum scaling (TSS), cumulative sum scaling (CSS), variance stabilizing transformation (VST), relative log expression (RLE), Aitchison's centered log-ratio (CLR) or log ratio (ALR) of compositions, trimmed mean of M-values (TMM), and upper-quartile (Q75). The default normalization methods for edgeR, metagenomeSeq, DESeq2, ANCOM, and ZIBSeq are TMM, CSS, RLE, ALR, and TSS, respectively. Further, TMM, RLE, Q75, and TSS can be applied by both edgeR and DESeq2. Both ZIGDM and corncob apply model-based normalization. Model-based normalization estimates the normalized abundances *via* a distribution parameter rather than using a separate normalization step. The choice of normalization method produces more precise and biologically interpretable results when it is chosen appropriately. For instance, the CSS normalization technique used in metagenomeSeq scales the data with cumulative count sums up to a certain quantile. Further, CSS was shown to produce optimal model performance particularly for MSS count data. CSS is also helpful when there is zero inflation in the count data. Zero counts do not imply the nonexistence of a feature but could be the result of undersampling. On the other hand, TMM in edgeR helped to minimize the false discovery rate (FDR). Additionally, TMM attempts to trim away unwanted undersampling and oversampling effects in the log-fold changes. Normalization in edgeR was later updated to account for outliers by applying weights to the normalized counts. Similarly, DESeq2 accounts for outliers in the count data by taking advantage of the median-of-ratios or RLE normalization techniques, which could result in a more robust model. ANCOM normalizes the data *via* ALR to map the data from the simplex $\mathbb{S}$ to $\mathbb{R}$, which allows for the use of classical tests such as ANOVA or Kruskal-Wallis to detect differential abundance. Normalizing the counts using TSS offers a way to model the compositional data directly as is done in ZIBSeq.

The microbiome data are assumed to follow a particular probabilistic model or distribution, which accounts for the noise in the count data. The probability density functions (pdf) or probability mass functions (pmf) and additional information for the parameters of each of the models are listed in **Table 5**. In general, the count data are sampled from

$$y_{ij}|\cdot \sim \mathcal{M}(\cdot) \tag{1}$$

where $\mathcal{M}(\cdot)$ is a probabilistic model that depends on either the normalized or non-normalized abundances as well as other parameters such as mean or dispersion when applicable. The models discussed in this section choose candidates for $\mathcal{M}$ to deal with at least one of the characterizations of count data. The negative binomial distribution (NB) is the candidate for $\mathcal{M}$ for both edgeR and DESeq2 and can be generalized as $y_{ij}|\cdot \sim \text{NB}(\lambda_{ij}, \phi_j)$ where λ_{ij} is the mean and ϕ_j is the feature-specific dispersion parameter. The mean λ_{ij} is parameterized as the product of a normalization factor s_i and a parameter related to the count data μ_{ij}, which is expressed as $\lambda_{ij} = s_i\mu_{ij}$.

TABLE 5 | Summary of the count data models that appear throughout this paper.

Method	Model	Additional information
edgeR	$y_{ij}\|\cdot \sim \text{NB}(\lambda_{ij}, \phi_j)^{*}$	The mean parameter $\lambda_{ij} = N_i\tilde{y}_{ij}$ accounts for variation in library size and relative abundance.
DESeq2	$y_{ij}\|\cdot \sim \text{NB}(\lambda_{ij}, \phi_j)$	$\lambda_{ij} = s_i q_{ij}$ where s_i is estimated by the median-of-ratios method; q_{ij} is proportional to the amount of feature-wise cDNA fragments in a sample.
metagenomeSeq	$\check{y}_{ij}\|\cdot \sim \text{N}(\mu_j, \sigma_j^2)^{**}$	In this zero-inflated model, $\pi_i(C_i)$ is the probability that an observed count is zero and is estimated *via* $\text{logit}(\pi_i) = \beta_0 + \beta_1 \log C_i$, where C_i is the normalized abundance *via* CSS, μ_j and σ_j^2 are feature-specific Gaussian mean and variance.
ANCOM	Not applicable	Uses standard ANOVA to model the ALR-transformed relative abundances.
ZIBSeq	$\tilde{y}_{ij}\|\cdot \sim \text{Beta}(\mu_{ij}, \phi_{ij})^{\dagger}$	In this zero-inflated model, π_i is the probability that a relative abundance is zero, μ_{ij} is the mean and ϕ_{ij} is precision. This parametrized beta distribution has shape parameters $\mu_{ij}\phi_{ij}$ and $(1-\mu_{ij})\phi_{ij}$.
ZIGDM	$y_{ij}\|\cdot \sim \text{GDM}(\boldsymbol{\omega}_{i\cdot}, \boldsymbol{a}_{i\cdot}, \boldsymbol{b}_{i\cdot})^{\ddagger}$	In this zero-inflated model, π_{ij} is the probability that a count is zero.
corncob	$y_{ij}\|\cdot \sim \text{Binomial}(N_i, \tilde{y}_{ij})^{\dagger\dagger}$	The prior on $\tilde{y}_{ij}$ is $p(\tilde{y}_{ij}) = \text{Beta}(a_{1j}, a_{2j})^{\ddagger\ddagger}$ with expected value $\mu_{ij} = a_{1j}/(a_{1j}+a_{2j})$, which allows the use of the logit function to model the mean of the compositions.
DMR, DMBVS, DMLMbvs	$\boldsymbol{y}_{i\cdot}\|\cdot \sim \text{DM}(\boldsymbol{\alpha}_{i\cdot})^{***}$	This zero-inflated model depends on a single parameter, $\boldsymbol{\alpha}_{i\cdot}$ which can be interpreted as the model-based normalized abundances of sample i. Each $\alpha_{ij} \in \mathbb{R}^{+}$.
IntegrativeBayes	$y_{ij}\|\cdot \sim \text{NB}(\lambda_{ij}, \phi_j)$	In this zero-inflated model, the mean parameter is $\lambda_{ij} = s_i\alpha_{ijg}$ where size factor s_i is sequencing depth and α_{ijg} is the CSS normalized abundance of feature j in sample i and phenotypic group g. The feature-wise dispersion parameter is ϕ_j.

**The NB pmf in general is* $f(y|\cdot) = \frac{\Gamma(y+\phi)}{y!\Gamma(\phi)}\left(\frac{\phi}{\lambda+\phi}\right)^{\phi}\left(\frac{\lambda}{\lambda+\phi}\right)^{y}$.

***The normal pdf in general is* $f(y|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2\right)$.

†*The beta pdf in general can be parametrized as* $f(y|\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)}y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}$.

†*The GDM pmf in general can be written as.*

††*The binomial pmf in general is* $f(y|\cdot) = \binom{N}{y}p^{x}(1-p)^{N-y}$.

‡‡*This beta pdf is non-parametrized and written generally as* $f(y|\cdot) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}y^{\alpha-1}(1-y)^{\beta-1}$.

****The DM pmf in general is* $f(\boldsymbol{y}_{i\cdot}|\boldsymbol{\alpha}_{i\cdot}) = \frac{\Gamma(\sum_{j=1}^{p} y_{ij}+1)\Gamma(\sum_{j=1}^{p}\alpha_{ij})}{\Gamma(\sum_{j=1}^{p} y_{ij}+\sum_{j=1}^{p}\alpha_{ij})}\prod_{j=1}^{p}\frac{\Gamma(y_{ij}+\alpha_{ij})}{\Gamma(y_{ij}+1)\Gamma(\alpha_{ij})}$.

The choices for these two parameters for edgeR and DESeq2 are provided in **Table 5**. The variance of y_{ij} is $\lambda_{ij} + \lambda_{ij}^2/\phi_j$. The variance increases as ϕ_j tends toward small values, which accounts for overdispersion in the count data [19]. However, the NB does not account for zero inflation. Next, a generalized linear model (GLM) using a log-link to model the mean abundance of feature *j* in sample *i* is fitted *via* maximum-likelihood estimation. The GLM can be written as $\log(\mu_{ij}) = \boldsymbol{\beta}_{j\cdot}(1, z_i, \boldsymbol{x}_{i\cdot})^{\top}$ where $\boldsymbol{\beta}_{j\cdot} = (\beta_{j0}, \beta_{j1}, \ldots, \beta_{j,k+2})$ are the regression coefficients for the intercept, phenotype, and covariates, respectively. Testing for differential abundance here is the equivalent of testing $H_0 : \beta_{j1} = 0$. edgeR uses the modified Fisher's exact test by [50] where NB replaces the hypergeometric distribution; whereas, DESeq2 used the Wald test. Originally, edgeR was designed for only a binary phenotype but has since been updated to handle multiple groups.

Zero-inflated models help to significantly reduce the bias of sparse count data resulting from the undersampling effect of limited sequencing depth. The zero-inflated model is a mixture distribution with a continuous component and spike-mass at zero. An attractive feature of the zero-inflated model is the ability to estimate the probability that a zero count is a true zero (i.e., true absence of feature *j*) or false zero (i.e., undersampling) by the use of a discrete spike-mass at zero [51]. Zero-inflated models are employed by metagenomeSeq, ZIBSeq, and the ZIGDM. Generally, $\mathcal{M}$ as a zero-inflated model can be expressed as $y_{ij}|\cdot \sim \pi_i I_0(y_{ij} = 0) + (1-\pi_i)\mathcal{M}(\cdot)$. An equivalent way to model $y_{ij}|\cdot$ is by the use of a latent binary variable r_{ij} such that

$$y_{ij}\begin{cases} = 0 & \text{when } r_{ij} = 1 \\ \sim \mathcal{M}(\cdot) & \text{when } r_{ij} = 0 \end{cases} \tag{2}$$

where π_i is the probability that y_{ij} is zero [24] and r_{ij} is sampled from a Bernoulli distribution with parameter π_i [19]. A discrete spike assumes $y_{ij} = 0$ with positive probability [51]; whereas, a continuous spike assumes $y_{ij} = 0$ with zero probability [52]. The candidates for a zero-inflated $\mathcal{M}$ are Gaussian or log-Gaussian, beta, and generalized Dirichlet-multinomial for metagenomeSeq, ZIBSeq, and the ZIGDM, respectively. Specifically, metagenomeSeq models the $\log_2$ continuity-corrected count data as $\check{y}_{ij} = \log_2(y_{ij} + 1)$ along with the CSS-normalized scaling factor s_i for two population groups. Pseudo counts are created by adding a positive value to each y_{ij} to avoid taking a logarithm at zero. The mean model here is written as $\mu_{ij}|\cdot = \boldsymbol{\beta}_{j\cdot}[1, z_i, \boldsymbol{x}_{i\cdot}, \log_2(C_i + 1)]^{\top}$ where $\boldsymbol{\beta}_{j\cdot} = (\beta_{j0}, \beta_{j1}, \ldots, \beta_{j,k+3})$ are the regression coefficients for the intercept, phenotype, and covariates. The term $\log_2(C_i + 1)$ is the main contribution of metagenomeSeq, which includes the CSS-normalized value, C_i, that helps to remove bias due to extremely large counts in any sample. The expectation-maximization (EM) algorithm is used to fit the model and estimate the parameters. Testing for differential abundance here is the equivalent of testing $H_0 : \beta_{j1} = 0$. Then, metagenomeSeq computes *q*-values for multiple testing from a modified t-statistic calculated *via* Empirical Bayes. One issue here is that FDR increases when either sample size or effect size increases [20].

ZIBSeq models the relative abundances *via* a parametrized beta distribution [53] where $\tilde{y}_{ij} \sim \text{Beta}(\mu_{ij}, \phi_{ij})$ has mean μ_{ij}, precision ϕ_{ij}, and variance $\mu_{ij}(1 - \mu_{ij})/(\phi_{ij} + 1)$. Then, the mean is modeled by GLM binomial regression with a logit-link, $\text{logit}(\mu_{ij}) = \boldsymbol{\beta}_{j\cdot}(1, z_i)^\top$ where $\boldsymbol{\beta}_{j\cdot} = (\beta_{j0}, \beta_{j1})$ are regression coefficients for the intercept and phenotype. The model parameters are estimated using maximum likelihood estimation *via* the `R` package `GAMLSS`. Testing for differential abundance here is the equivalent of testing $H_0 : \beta_{j1} = 0$. ZIBSeq computes q-values for multiple testing from either a Chi-square or t-distribution depending on sample size.

The relative abundances are modeled by ZIBSeq one feature at a time; whereas, the ZIGDM models the abundances directly under a multivariate setting which can better capture the compositional effects. Further, the Dirichlet-multinomial (DM) distribution can account for overdispersion. But, the ZIGDM is a more flexible model than the DM. The novelty of the ZIGDM is two-fold: the use of the GDM to model count data and account for zero inflation with additional parameters had not been done before. The ZIGDM models the counts of each sample as $\boldsymbol{y}_{i\cdot} \sim \text{ZIGDM}(\boldsymbol{\omega}_{i\cdot}, \boldsymbol{a}_{i\cdot}, \boldsymbol{b}_{i\cdot})$ where π_{ij} is a Bernoulli parameter that controls the absence probability, and a_{ij} and b_{ij} are beta distribution parameters that control the presence of feature j in sample i. The beta mean of the proportion of feature j in sample i is $a_{ij}/(a_{ij} + b_{ij})$. By letting the dispersion parameter be $\sigma_{ij} = 1/(1 + a_{ij} + b_{ij})$, the beta variance can be expressed as $\mu_{ij}(1 - \mu_{ij})\sigma_{ij}$ which accounts for overdispersion. Similar to ZIBSeq, the ZIGDM also uses GLMs to model the mean parameter μ_{ij}, but then uses score statistics and permutation p-values to test for differential abundance. The mean model is written as $\text{logit}(\mu_{ij}) = \boldsymbol{\beta}_{j\cdot}(1, z_i, \boldsymbol{x}_{i\cdot})^\top$ where $\boldsymbol{\beta}_{j\cdot} = (\beta_{j0}, \beta_{j1}, \ldots, \beta_{j,q+2})$ are the regression coefficients for the intercept, phenotype, and covariates, respectively. Testing for differential abundance here is the equivalent of testing $H_0 : \beta_{j1} = 0$.

The beta-binomial distribution is the candidate for $\mathcal{M}$ in corncob. In particular, corncob models the abundances as $y_{ij} \sim \text{Binomial}(N_i, \tilde{y}_{ij})$ to account for library size and relative abundance. Further, corncob assumes a beta prior on the probability parameter, which is expressed as $\tilde{y}_{ij} \sim \text{Beta}(a_{1j}, a_{2j})$, for model flexibility. Under this setting, the expected relative abundance of feature j in sample i is estimated by $\mu_{ij} = E(\tilde{y}_{ij}) = a_{1j}/(a_{1j} + a_{2j})$ and provides convenient support on (0,1) for modeling compositions *via* binomial regression. The binomial variance of $y_{ij}|N_i$ is $N_i\mu_{ij}(1 - \mu_{ij}) \times [1 + \phi_{ij}(N_i - 1)]$ with an inflation factor of $1 + \phi_{ij}(N_i - 1)$ where $\phi_{ij} = 1/(1 + a_{1j} + a_{2j})$. The flexibility of the prior allows corncob to account for library size and overdispersion when ϕ_{ij} is large. The mean model is written as $\text{logit}(\mu_{ij}) = \boldsymbol{\beta}_{j\cdot}(1, z_i, \boldsymbol{x}_{i\cdot})^\top$ and fitted using the trust region optimization algorithm for more efficient computation. The regression coefficients are given by $\boldsymbol{\beta}_{j\cdot} = (\beta_{j0}, \beta_{j1}, \ldots, \beta_{j,q+2})$ for the intercept, phenotype, and covariates, respectively. Testing for differential abundance here is the equivalent of testing $H_0 : \beta_{j1} = 0$. The parametric bootstrap Wald test is used to test for differential abundance.

ANCOM uses standard ANOVA to model the ALR-transformed relative abundances. ALR is based on Aitchison's methodology of log ratios of compositional data [54]. The ALR transformation overcomes the unit-sum constraint on the relative abundances [18] and creates a map from the simplex $\mathbb{S}$ to $\mathbb{R}$, which then allows for the use of classical statistical methods such as ANOVA [20]. Each feature is used as a reference feature one at a time which produces $p(p - 1)$ regression models [20]. Each model is written as

$$\log \frac{\tilde{y}_{ijg}}{\tilde{y}_{ij'g}} = \alpha_{jj'} + \beta_{jj'g} + \epsilon_{ijj'g} \qquad (3)$$

where $\alpha_{jj'}$ is the mean, $\beta_{jj'g}$ captures the phenotypic group effect, and $\epsilon_{ijj'g}$ is the error term for sample i, taxa $j \neq j'$, and phenotype $z_i = g$. Covariates can also be included in the linear model when applicable. Testing for differential abundance is the equivalent of testing $H_0 : \beta_{jj'1} = \cdots = \beta_{jj'G}$ using classical tests such as ANOVA, t-test, Wilcoxon, or Kruskal–Wallis depending on the number of groups and linear assumptions. The p-values are adjusted using the Benjamini–Hochberg procedure for multiple testing.

The above models are univariate with the exception of the ZIGDM. Multivariate models for analyzing microbiome data are also available. However, the multivariate methods do not perform better than the univariate methods for differential abundance analysis [43, 55]. An advantage of multivariate methods are the useful plots and numerical summaries of dimension reduction techniques including, but not limited to, principal components analysis (PCA), principal coordinates analysis (PCoA), canonical correlation analysis (CCA), and partial least squares discriminant analysis (PLS-DA) [43, 56]. For example, mixMC [43] is a multivariate method available in the `mixOmics` package in Bioconductor. Using CLR-transformed compositions, mixMC applies PCA to visualize diversity patterns and multivariate regression using sparse PLS-DA *via* a lasso penalty to select the most differentially abundant features. Lastly, mixMC also provides a multivariate approach for longitudinal differential abundance analysis.

The literature for longitudinal differential abundance analysis had been lacking until recently. The cost for sequencing has decreased over time allowing for the production of more longitudinal data [46]. Some earlier methods include Next maSigPro (microarray Significant Profiles) available in Bioconductor [44] and a negative binomial mixed effects model (NBME) available in the `R` package `timeSeq` [45]. Both Next maSigPro (the updated version of maSigPro) and NBME do not have normalization methods built into the software and so the user must normalize the data beforehand. Three methods developed around the same time for longitudinal differential abundance analysis include MetaSplines [46], MetaDprof [47], and MetaLonDA (Metagenomic Longitudinal Differential Abundance) [48, 57]. MetaSplines is available in metagenomeSeq, MetaDprof has no `R` package available, and MetaLonDA can be accessed through CRAN. MetaSplines, MetaDprof, and MetaLonDA apply a semi-parametric method known as smoothing spline ANOVA or SS-ANOVA to detect longitudinal differential abundance. MetaLonDA models the count data using the negative binomial distribution; whereas, MetaSplines and MetaDprof use the Gaussian distribution. MetaLonDA is designed to handle inconsistencies in time points, different number of samples per subject, and different number of subjects per phenotypic group. Further information regarding MetaSplines, MetaDprof, and MetaLonDA is provided in detail by [58]. More recently, NBZIMM [49] allows for the implementation of a negative binomial mixed effects model, zero-inflated negative binomial model, and Gaussian mixed effects model. NBZIMM is available for users in R *via* GitHub.

All of the above methods are useful. Appropriate model selection should be determined by a statistical procedure rather than by user choice. For example, a statistical procedure for identifying zero-inflated and hurdle distributions (iZID) was recently developed to appropriately model MSS data [59] and can be implemented in R *via* the iZID package [60]. Hurdle models, introduced by [61], are also referred to as zero-altered (ZA) models. Generally, ZA consists of one process that generates zeros and a second process truncated at zero that generates positive counts [62]. Unlike zero-inflated models, the slab of a zero-altered model cannot generate zeros. Wang et al. [60] provide the details of multiple zero-inflated and hurdle models along with any existing R packages. Keep in mind that there are other zero-inflated models not discussed in this paper that are available for microbiome data analysis. For example, a recent zero-inflated negative binomial model with a Dirichlet-process prior (ZINB-DPP) offers a Bayesian approach for differential abundance analysis [63].

The count data for edgeR and DESeq2 were sequenced *via* RNA-Seq. Other methods in this section such as metagenomeSeq and ZIBSeq generated the count data *via* 16S rRNA sequencing technology. An implication here is that these models were built for specific types of count data. Consequently, one must also consider the type of data when choosing a model. Further, corncob used soil microbiome data with three treatments; where as, human microbiome data were analyzed by the other methods. Of course, a zero-inflated model should be preferred when sparsity in the count data is high. Compared to metagenomeSeq, ZIBSeq is better suited for larger sample sizes and sparse count data based on its reported area under the receiver operating characteristic curve (AUC) *via* simulations. For smaller sample sizes, ZIBSeq is well-suited for multinomial and binomial data but not for zero-inflated Poisson or zero-inflated negative binomial data. Similar to ZIBSeq, corncob uses a single-feature modeling approach which does not take the compositional nature of the count data into consideration. So, a multivariate version of both of these two methods ought to be considered. Interestingly, ANCOM outperformed metagenomeSeq by significantly reducing FDR and increasing power even though it is not a zero-inflated model. However, ANCOM accounts for the compositionality of the count data using ALR normalization. ANCOM can also perform longitudinal analysis to test differential abundance at different time points. Later in 2020, Lin and Peddada [20] released ANCOM-BC which is an improved version of ANCOM that includes bias correction (BC). Most of the models above assume that abundance dispersions between groups are homogeneous. Both edgeR and DESeq2 use shrinkage to estimate dispersion and log-fold changes. Shrinkage helped to improve consistency and interpretation of results. The method of shrinkage is what sets edgeR and DESeq2 apart. Shrinkage in edgeR is determined by a user-adjusted parameter that depends on the prior degrees of freedom to impose weight on individual gene estimates and dispersions, which creates a weighted likelihood conditional on the count data. The conditional likelihood is based on the assumption that features with similar observed abundances have similar variances. The assumption of homogeneity may be problematic since phenotype level abundances can be influenced by multiple factors (e.g., other features, covariates, host, environment, etc.). The ZIGDM and corncob take differential dispersion into account, which is the assumption that dispersions between phenotype groups are heterogeneous. Also, tests for differential dispersion are available for the ZIGDM and corncob to determine if there is a significant association between dispersion and covariates. Notably, corncob was the first method to develop a test for differential dispersion. Finally, these models can be adjusted to include covariates. The ability to incorporate covariates into a model brings us to integrative analysis in the next section.

3. INTEGRATIVE ANALYSIS

There is an association between the microbiome and covariates including but not limited to metabolites, antibiotic usage, environmental factors, and host genetics that can influence host health [64, 65]. Recently, numerous associations between dietary covariates and taxa were implicated in the development of chronic diseases such as obesity [66]. The goal of integrative analysis is to identify and quantify associations between the microbiome and covariates. We discuss four methods for integrative analysis in this section including Dirichlet-multinomial regression or DMR [67], Dirichlet-multinomial Bayesian variable selection or DMBVS [68], a Bayesian zero-inflated negative binomial (ZINB) also referred to as IntegrativeBayes [19], and a Dirichlet-multinomial linear model with Bayesian variable selection or DMLMbvs [66]. **Table 6** provides a summary of the methods discussed in this section and **Table 7** summarizes the implementation of these methods in R.

The common biological motivation of each method is to determine if associations exist between any of the p features from $\boldsymbol{Y}_{n\times p}$ and q covariates from $\boldsymbol{X}_{n\times q}$ while controlling for the phenotypic response $\boldsymbol{z}_{n\times 1}$. Statistical motivations here are similar to the differential abundance methods. DMR and DMLMbvs model the count data directly to account for issues arising from compositionality as well as overdispersion. DMBVS was highly interested in the connection between disease development and the association of the microbiome with other covariates. The lack of available models for integrative analysis motivated IntegrativeBayes to construct a model that could account for zero inflation and overdispersion.

Three of the models employ the DM, which is one of the distributions useful for model-based normalization. Other

TABLE 6 | Summary of methods for integrative analysis in microbiome studies.

Method	Model assumption	Normalization	Data type	Covariate type	References
DMR	Dirichlet-multinomial	None*	16S rRNA	Nutrient intake	[67]
DMBVS	Dirichlet-multinomial	None*	16S rRNA	KEGG pathways	[68]
IntegrativeBayes	Zero-inflated negative binomial	CSS	MSS	KEGG pathways; Metabolomics	[19]
DMLMbvs	Dirichlet-multinomial	None*	16S rRNA	Dietary	[66]

**The Dirichlet-multinomial model does not require data normalization.*

TABLE 7 | Implementation of methods for integrative analysis in microbiome studies.

Method	Implementation	Updated
DMR	http://statgene.med.upenn.edu/software.html	2013
DMBVS	https://github.com/duncanwadsworth/dmbvs	2017
IntegrativeBayes	https://github.com/shuangj00/IntegrativeBayes	2019
DMLMbvs	https://github.com/mkoslovsky/DMLMbvs	2020

techniques of normalization decreases the power of the DM due to some loss of variation when using compositions. Thus, using the count data directly in the DM instead of the compositions results in better model performance [67]. The ZINB is most robust if the counts are first normalized using CSS when compared to other normalization methods. Further, the results of metagenomeSeq showed that CSS is highly beneficial for analyzing zero-inflated count data.

The first step of integrative analysis is to model the count data using Equation (1). The DM is the candidate for $\mathcal{M}$ to denoise the count data in DMR, DMBVS, and DMLMbvs. The model learns the compositionality while also accounting for uncertainty and overdispersion in the count data. The count data are modeled as $\boldsymbol{y}_{i\cdot}|\boldsymbol{\alpha}_{i\cdot} \sim \mathrm{DM}(\boldsymbol{\alpha}_{i\cdot})$ where $\boldsymbol{\alpha}_{i\cdot} = (\alpha_{i1}, \ldots, \alpha_{ip})$ is the i-th sample row vector of normalized abundances estimated by the model. The DM parameter is strictly positive where each $\alpha_{ij} \in \mathbb{R}^{+}$ and the unit-sum constraint has been removed.

The DM is derived by each model as follows [66–68]. The counts are assumed to follow a multinomial (Multi) distribution $\boldsymbol{y}_{i\cdot}|N_i, \boldsymbol{\psi}_{i\cdot} \sim \mathrm{Multi}(N_i, \boldsymbol{\psi}_{i\cdot})$. Then, a Dirichlet (Dir) prior is placed on the multinomial parameter $\boldsymbol{\psi}_{i\cdot}|\boldsymbol{\alpha}_{i\cdot} \sim \mathrm{Dir}(\boldsymbol{\alpha}_{i\cdot})$. The DM is the result of integrating out the multinomial parameter, $\boldsymbol{\psi}_{i\cdot}$ which is expressed as $f_{\mathrm{DM}}(\boldsymbol{y}_{i\cdot}|\boldsymbol{\alpha}_{i\cdot}) = \int p(\boldsymbol{y}_{i\cdot}|N_i, \boldsymbol{\psi}_{i\cdot})p(\boldsymbol{\psi}_{i\cdot}|\boldsymbol{\alpha}_{i\cdot})d\boldsymbol{\psi}_{i\cdot}$. Integrating out this parameter makes the model more efficient by having one less parameter. The variance of the DM is $\mathrm{Var}(y_{ij}) = (N_i + A_i)/(1 + A_i)E(\psi_{ij})[1 - E(\psi_{ij})]N_i$, where $A_i = \sum_{j=1}^{p} \alpha_{ij}$. The variance is inflated by a factor of $(N_i + A_i)/(1 + A_i)$ relative to the variance of the multinomial distribution. As a result, the DM model accounts for overdispersion in the count data. Letting A_i tend toward zero will result in large overdispersion. If $A_i \to \infty$, the DM model reduces to a multinomial model. DMBVS and DMLMbvs are Bayesian adaptations of the DM.

IntegrativeBayes employs the NB as the candidate for the zero-inflated $\mathcal{M}$ from Equation (2). If $r_{ij} = 0$ in Equation (2), then the ZINB models the count data and their uncertainty while accounting for sample-wise sequencing depth s_i by reparameterizing each sample taxon-specific negative binomial mean as the product of s_i and the CSS normalized abundances, which are expressed as α_{ijg}. Further, the model also introduces a latent binary vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)$ where $\gamma_j = 1$ indicates the j-th feature is differentially abundant among the G groups. A beta-Bernoulli prior is placed on each γ_j to quantify the proportion of features that are believed to be discriminatory. Then, the ZINB conditional on $r_{ij} = 0$ is expressed as

$$y_{ij}|r_{ij} = 0, \gamma_j, s_i, \alpha_{ijg}, \alpha_{ij0} \sim \begin{cases} \mathrm{NB}(s_i\alpha_{ij0}, \phi_j) & \text{if } \gamma_j = 0 \\ \mathrm{NB}(s_i\alpha_{ijg}, \phi_j) & \text{if } \gamma_j = 1, z_i = g \end{cases} \tag{4}$$

where ϕ_j is the feature-specific dispersion parameter with a Gamma prior. The parameter ϕ_j captures overdispersion in the same manner as edgeR and DESeq2 described earlier.

Next, log-linear regression is employed by all four methods to identify any feature-covariate associations. The general framework of log-linear regression can be expressed as

$$\log \alpha_{ijg} = \boldsymbol{\beta}_{j\cdot}(1, \boldsymbol{x}_{i\cdot})^{\top} \tag{5}$$

where the abundance-related response is $\log \alpha_{ijg}$ for the covariates in sample i and group g. The feature-covariate regression coefficients are $\boldsymbol{\beta}_{j\cdot} = (\beta_{j0}, \beta_{j1}, \ldots, \beta_{jq})$ where the feature-specific intercept term is β_{j0} and $\beta_{j1}, \ldots, \beta_{jq}$ estimate the associations between the j-th feature and the q covariates in the i-th sample. DMR, DMBVS, and DMLMbvs do not deviate from this specification. However, IntegrativeBayes expresses their log-linear regression model as

$$\begin{cases} \log \alpha_{ij0} = \mu_{0j} + \boldsymbol{x}_{i\cdot}^{\top}\boldsymbol{\beta}_{j\cdot} & \text{if } \gamma_j = 0 \\ \log \alpha_{ijg} = \mu_{0j} + \mu_{gj} + \boldsymbol{x}_{i\cdot}^{\top}\boldsymbol{\beta}_{j\cdot} & \text{if } \gamma_j = 1, z_i = g \end{cases} \tag{6}$$

where μ_{0j} is the feature-specific intercept term or baseline and μ_{gj} captures the baseline shift between the g-th group and reference group. Both μ_{0j} and μ_{gj} are assigned zero-mean

Gaussian priors. The regression coefficients $\boldsymbol{\beta}_{j\cdot} = (\beta_{j1}, \dots, \beta_{jq})$ estimate the associations between the j-th feature and the q covariates in the i-th sample. Equations (5) and (6) do not require error terms because the uncertainty in the count data is accounted for by Equation (1) before regression is applied. DMLMbvs further calculates the mathematical balances $B(\boldsymbol{\psi}_{i\cdot})$ of the taxa proportions $\boldsymbol{\psi}_{i\cdot}$ estimated by $\mathcal{M}$ to predict a continuous phenotypic response *via*

$$z_i = (1, \boldsymbol{B}(\boldsymbol{\psi})_{i\cdot})^{\top} \boldsymbol{\beta}_m + \epsilon_i \tag{7}$$

where $\boldsymbol{\beta}_m = (\beta_0, \beta_1, \dots, \beta_M)$ contains the regression coefficients for the balances. Further, β_0 is the intercept term with a zero-mean Gaussian prior, β_m is the coefficient of balance m or $B(\psi)_{im}$, and the error term ϵ_i has a zero-mean Gaussian prior. Equation (7) requires an error term since the uncertainty in the phenotypic response is not accounted for in any previous steps. The multinomial prior $\boldsymbol{\psi}_{i\cdot}$ estimates the compositions and is usually integrated out for model efficiency; however, DMLMbvs retains this parameter because the phenotypic response depends on the estimated compositions in Equation (7). Mathematical balances are constructed *via* random sequential binary partitioning of $\boldsymbol{\psi}$, which results in a total of $M = p - 1$ partitions. Mathematical balances help to identify a subset of features having the greatest association with the response rather than using a single feature selection approach.

Equations (5), (6), and (7) are high-dimensional with $q \times (p + 1)$, $q \times (p + 2)$, and p parameters, respectively. Multiple testing results in a loss of power in a high-dimensional setting and so regularization would help to increase the power [67]. Thus, regularization of Equations (5), (6), and (7) is necessary to optimize the model by reducing the parameter space [19]. Regularization for DMR, the only frequentist method here, includes both group and individual ℓ_1 penalties and is optimized using an efficient block-coordinate descent algorithm, which utilizes a quadratic approximation of the log-likelihood function. Sparse ℓ_1 regularization encourages sparsity in the regression coefficients, which is useful for identifying significant taxa-covariate associations whose regression coefficients are non-zero. Then, DMR uses the likelihood ratio test to identify significant feature-covariate associations. DMBVS, DMLMbvs, and IntegrativeBayes employ spike-and-slab priors to reduce the parameter space and for identifying significant feature-covariate associations. Spike-and-slab priors are conventional for Bayesian variable selection [19, 66, 68] and are defined as

$$\beta_{jk} \sim (1 - \delta_{jk}) I(\beta_{jk} = 0) + \delta_{jk} N(0, \sigma_{\beta}^2) \tag{8}$$

where $\delta_{jk} = 1$ indicates that feature j and covariate k are associated (i.e., $\beta_{jk} \neq 0$) and $\delta_{jk} = 0$ otherwise. A beta-binomial prior is imposed on the latent binary variable δ_{jk} to control the number of significant associations selected by the model. The variance term σ_{β}^2 is assigned a conjugate prior (inverse-gamma) for model efficiency. Model fitting and sampling of non-zero regression coefficients is implemented using the Markov Chain Monte Carlo (MCMC) Metropolis-Hastings algorithm within a Gibbs sampler. The proportion of posterior samples with non-zero coefficients, or where $\delta_{jk} = 1$, is called the posterior probability of inclusion (PPI) and makes a parsimonious quantification of uncertainty in variable selection. PPI above a certain threshold indicates the significance of any feature-covariate association, which is equivalent to testing $H_0 : \beta_{jk} = 0$. The null hypothesis is retained if PPI is below the chosen threshold and otherwise rejected. IntegrativeBayes uses a threshold that controls FDR; whereas, DMBVS and DMLMbvs use median PPI (i.e., 0.5) as the threshold.

Each of the four models here included numerous covariates. For instance, DMR included dietary intake of nutrients. DMBVS and IntegrativeBayes included molecular function covariates from the Kyoto Encyclopedia of Genes and Genomes (KEGG). IntegrativeBayes focused primarily on metabolomic covariates. DMLMbvs included multiple dietary covariates. The three DM-based models also included 16S rRNA count data as the features; whereas, IntegrativeBayes applied their model to MSS count data. IntegrativeBayes is the only model presented here that accounts for zero inflation in the count data, estimates the effect size for the discriminating features, and identifies differentially abundant features while simultaneously quantifying feature-covariate associations. Because IntegrativeBayes accounts for zero inflation, it is a well-suited model for MSS data. However, the ZINB is not the only zero-inflated model available for integrative analysis. A comprehensive review of other practical zero-inflated models is provided by [60, 74, 75]. All four models here account for overdispersion, compositionality, and high dimensionality of the data. DMLMbvs was the only model that had a continuous phenotype (body mass index or BMI); however, the authors noted that the model can be adjusted to include a categorical response. DMLMbvs uses the estimated compositional data to simultaneously identify feature-covariate associations and predict a continuous phenotypic outcome. The Bayesian approach is used by three of the four methods because it has several advantages over frequentist methods. Bayesian models can incorporate prior knowledge, quantify the uncertainty of model parameters, offer efficient model fitting *via* the MCMC algorithm, and calculate parsimonious inferential summaries such as PPI in variable selection.

4. NETWORK ANALYSIS

Microbial ecological interactions affect microbiome function and host health *via* the formation of complex communities with various symbiotic relationships where microbes coexist. Findings of a study implicated pH as a main factor for the networking of microbial communities in arctic soil [76]. The soil study found a two-cluster microbial network where one cluster was correlated to pH and the second was uncorrelated to pH. The goal of network analysis is to construct microbiome networks that characterize microbial ecological associations (i.e., taxa-taxa dependencies), which may help discover fundamental properties and mechanisms of microbial ecosystems [73]. Graphical models consist of nodes and edges, which are used to visualize the estimated microbial network. Each node corresponds to a taxon and an existing edge represents a direct association between any two nodes. Current statistical methods for network analysis estimate the correlation or partial correlation structure of the normalized count data to construct a network of nodes and edges [73]. Correlation-based methods include SparCC (Sparse Correlations for Compositional data) [69], CCLasso (Correlation inference for Compositional data through Lasso) [70], and REBACCA (Regularized Estimation of the BAsis Covariance based on Compositional dAta) [71]. Partial correlation-based methods include SpiecEasi (SParse InversE Covariance Estimation for Ecological Association Inference) [72] and HARMONIES (Hybrid Approach foR MicrobiOme Network Inferences *via* Exploiting Sparsity) [73]. SPRING (Semi-Parametric Rank-based approach for INference in Graphical model) [7] employs both correlation and partial correlation methods under a semi-parametric setting. Semi-parametric rank (SPR) correlation can be used as an alternative to correlation measures such as Pearson or Spearman, can be extended to partial correlation, and can account for zero inflation [7]. **Table 8** provides a summary of the methods discussed in this section and **Table 9** summarizes the implementation of these methods in R.

The main biological motivation of each of the above methods is the exigency to estimate correlation or partial correlation networks using normalized count data to make inferences about microbial interactions. Since the dawn of high-throughput next-generation sequencing technology, we have had the quantitative ability to characterize microbial communities [71] and learn how they associate with environmental conditions such as host health, metabolism, etc. [72]. The number of spurious taxa-taxa associations tend to be about three times the number of true associations and miss about 60% of the true associations when using conventional methods such as Pearson or Spearman correlation on compositional data [69]. Small sample sizes are common in microbiome studies, which can result in lower power for network inference [72]. Many taxa-taxa associations have not been verified in the literature and so benchmarking tools to assess model quality are needed [7, 73]. Statistical motivations are no different here than in previous sections. Microbiome data tend to be zero-inflated [7]. Further, microbiome data have high dimensionality, overdispersion, and sample heterogeneity [73]. Thus, network analysis models that consider the characteristics of microbiome data are most appropriate.

One of the main issues of microbiome data is the unit-sum constraint of compositional data. As stated earlier, Aitchison's log-ratios are a useful normalization technique for compositional data. Methods such as SparCC, CCLasso, and REBACCA apply ALR normalization. SpieceEasi and SPRING apply CLR normalization. Unlike ALR, the CLR-transformed compositions are p-dimensional since no feature is used as a baseline. Both ALR and CRL map compositions from $\mathbb{S}$ to $\mathbb{R}$. However, ALR can be more advantageous because these ratios are equivalent to the ratio of absolute abundances and have the subcompositional coherence property where the ratio of two features' compositions is independent of other features [69]. HARMONIES applies model-based normalization where the count data are modeled directly *via* the parameters of the ZINB, which account for zero inflation, sample heterogeneity, overdispersion, and high dimensionality.

The methods discussed in this section use different approaches to estimate networks. Without loss of generality, we write

$$\rho_{jj'} \sim \mathcal{L}(\cdot) \tag{9}$$

where $\mathcal{L}(\cdot)$ is some process that estimates the covariance structure of the data, which is expressed as $\tilde{\boldsymbol{\Sigma}}$, to infer the correlation $\rho_{jj'}$ between features j and j'. Depending on the transformation, the process estimates either a log-basis covariance (e.g., models that use ALR or CLR transformations) or model-basis covariance (e.g., probabilistic models such as ZINB). The various candidates for $\mathcal{L}(\cdot)$ used by each method are discussed below.

SparCC makes an iterative estimation of the correlation matrix based on the Aitchison log-ratio transformation of relative abundances of any two features with the assumption of sparsity, which can be loosely summarized in two steps. First, SparCC estimates the correlation matrix of the log-ratio transformation. Secondly through iteration, the pair with the greatest correlation is removed and the correlations are estimated again until certain criteria are met. The log transformation contains information of the true absolute abundances, or basis abundances, which means that the ratio of the relative abundances of any two features is equal to the ratio of their two basis abundances. Further, the ratio of any two relative abundances is independent of any ratio of other features. The relative abundances are estimated under a Bayesian framework where they are treated as being not fixed. While SparCC has advantages such as good overall performance, robustness to sparsity, and does not depend on any underlying distributions, it can be computationally intense and so it is not the most efficient model. Further, SparCC does not account for the influence of errors in the estimating equations that are used to estimate the correlation matrix (e.g., some of the estimated correlations can fall outside of $[-1, 1]$).

CCLasso accounts for the influence of error through the use of a loss function and for sparsity with an ℓ_1 penalty term under a least squares framework to infer correlation structure. Similar to SparCC, CCLasso considers the compositional nature of microbiome data in estimating the correlation matrix with consistent accuracy. CCLasso has several advantages over SparCC. First, it produces a more accurate and positive definite correlation matrix. Second, all elements of the correlation matrix are contained in $[-1, 1]$. Third, CCLasso shrinks small correlations to zero, unlike SparCC, especially in the case of a shuffled microbiome dataset where correlations are actually zero.

REBACCA was published around the same time as CCLasso and so its performance was compared to only SparCC. While SparCC estimates the correlation matrix through an iterative procedure, REBACCA estimates the pairwise correlations by solving a linear system with deficient rank that is equivalent to the log-ratio transformations using ℓ_1-norm shrinkage to induce sparsity in the network. The advantages of REBACCA include a more efficient algorithm, higher power, lower false positive rate (FPR), and more consistency than SparCC. Finally, it is worth noting that SparCC, CCLasso, and REBACCA are based on linear methods of correlation.

Partial correlation is a measure of association of two random variables after removing the effects of confounding variables.

TABLE 8 | Summary of methods for network analysis in microbiome studies.

Method	Network type	Method	Normalization	References	Application
SparCC	Correlation	Iterative estimation of correlation	ALR	[69]	GitHub
CCLasso	Correlation	Least squares with l_1 penalty	ALR	[70]	GitHub
REBACCA	Correlation	Fast l_1-norm shrinkage	ALR	[71]	Online
SpiecEasi	Partial correlation	Gaussian graphical model	CLR	[72]	GitHub
SPRING	Partial correlation, SPR correlation	Truncated Gaussian copula model	Modified CLR	[7]	CRAN
HARMONIES	Partial correlation	Gaussian graphical model	DPP	[73]	GitHub

TABLE 9 | Implementation of methods for network analysis in microbiome studies.

Method	Implementation	Updated
SparCC	https://www.rdocumentation.org/packages/SpiecEasi/versions/1.0.7/topics/sparcc	2012
CCLasso	https://github.com/huayingfang/CCLasso	2016
REBECCA	https://faculty.wcas.northwestern.edu/hji403/REBACCA.htm	2015
SpiecEasi	https://github.com/zdk123/SpiecEasi	2021
SPRING	https://rdrr.io/github/GraceYoon/SPRING/man/SPRING.html	2020
HARMONIES	https://github.com/shuangj00/HARMONIES	2020

Controlling for confounding variables helps to reduce or eliminate spurious results. Further, partial correlation can be used to test for conditional independence of two random variables U and V given another random variable W [77]. Two features conditioned on abundances of all other features are conditionally independent if one of the two features provides no information about the abundance of the other feature. This implies that there is no direct association between the two features. The methods below estimate the partial correlation between features using several different approaches.

SPIEC-EASI employs two types of commonly used inference methods that utilize conditional independence for inferring sparse graphical models for CLR-transformed microbiome data. In short, covariance estimation and neighborhood selection are the two employed inference methods. Network sparsity is inferred through the Stability Approach to Regularization Selection (StARS), which uses subsampling to estimate a sparse network using minimal regularization [78]. For the first method, the graphical network is inferred through the estimation of the sparse inverse covariance matrix, which is regularized using graphical lasso (or glasso) so that indirect associations shrink to zero and only direct associations are selected. Glasso employs a penalized maximum likelihood approach with a global optimal solution for the reconstruction of the entire network. A Gaussian graphical framework is employed specifically because the inverse covariance matrix (or precision matrix) has the nice property of revealing conditionally dependent variables in the off-diagonal entries. For the second method, neighborhood selection is based on the known framework of Meinshausen and Bühlmann (MB method) to estimate local conditional independence one node at a time *via* Lasso [79]. Both inferential methods are advantageous because their formulations are convex optimizations, are useful for high-dimensional settings, can incorporate prior information regarding network topology, and outperformed SparCC overall.

SPRING is based on the use of novel SPR estimators of correlation and partial correlation for relative abundance data with a modified CLR transformation that can handle zero inflation. The modified transformation is rank-preserving and does not add a pseudo-value to zero counts. The semi-parametric model combines a truncated Gaussian copula graphical model with rank-based partial correlation to construct a sparse network using MB for neighborhood selection and StARS for model selection. SPRING outperformed existing methods such as SparCC for correlation inference and SPIEC-EASI for partial correlation inference. Also, the SPRING authors showed that Pearson correlation is not useful for identifying sparse partial correlations amongst features in microbiome data.

HARMONIES offers a competitive hybrid approach that includes a Bayesian zero-inflated negative binomial model with a Dirichlet process prior (ZINB-DPP). Further, glasso is applied to induce sparsity in the Gaussian graphical model that is regularized *via* StARS. The count data are normalized using a model-based approach *via* the DPP on the size factor s_i. ZINB-DPP accounts for overdispersion, zero inflation, sample heterogeneity, and high dimensionality making it a suitable model for estimating the true underlying abundances *via* the normalized abundances. Next, the posterior means of the normalized abundances on the log scale are used to fit a Gaussian graphical model to estimate the precision matrix for inferring the network with robust edge selection. HARMONIES demonstrated superior performance over existing methodologies including SPIEC-EASI and CCLasso in almost every scenario because it is designed to handle multiple challenges of microbiome data. Also, HARMONIES ensures proper biological interpretation of detected taxa-taxa associations by suggesting that all associated nodes are taxa of the same taxonomic level.

As an added bonus, SPIEC-EASI, SPRING, and HARMONIES each have their own novel synthetic data-generating tools that incorporate various network topologies to be used as a benchmark for assessing model quality. Synthetic data mimics real microbiome data and is currently a well-adapted way of assessing model quality due to the lack of a validated gold-standard network. MB-GAN (Microbiome Simulation *via* Generative Adversarial Network) [80] is another data synthesis tool useful for assessing model quality. MB-GAN addresses the challenges of simulating realistic microbiome data by learning from the given count data. The simulated count data are indistinguishable from the observed count data due to their similar properties such as sparsity, diversity, and taxa-taxa correlations. Other interesting features of MB-GAN include the use of real data as input without requiring model assumptions and efficient convergence.

Co-occurrence research within microbiomes has often focused on taxa-taxa associations. This is especially true in studies using amplicon sequencing techniques, such as 16S rRNA sequencing. However, recent work has begun to focus on reconstructing functional associations in metagenomic data [17, 81]. Genome-scale metabolic models (GEM), also known as Stoichiometric Metabolic Network models (SMN), computationally reconstruct and describe these associations [82]. The detailed information regarding the workflow, modeling, simulation, computational tools, and applications of GEMs can be found in [82–84]. These analyses used either the inferred or directly observed genetic content within the sampled metagenome to explore the microbial metabolic landscape. In addition, microbiome functional profiling of metagenomic data provides insight into what the microbial community has the potential to do at a molecular level [85]. With this in mind, research efforts have endeavored to model the community-scale metabolic potential encoded within metagenomes [83, 84]. These models not only need to consider the genetic content of the metagenome, but also must attempt to model fundamental arrangements of the microbial community, such as compartmentalization and availability of metabolites or nutrients, static or dynamic time constraints, and environmental sharing [84]. For example, Roume et al. [86] used a comparative multi-omic approach to reconstruct community-level metabolic networks within the microbial community present in wastewater. This type of analysis could aggregate the whole genetic content of the metagenome into one large community-scale organism. Taken together, metabolic modeling of functional metagenomic data allows for a closer mechanistic scrutinization of microbial co-occurrence within communities.

5. CONCLUSION AND OUTLOOK

In summary, multiple statistical methods for three major areas of microbiome research are available. Differential abundance analysis seeks to identify features that are discriminatory between phenotype groups. Since multiple diseases develop as a result of microbial dysbiosis, the results of differential abundance analysis may help find new and better ways to treat disease. Integrative analysis quantifies associations between taxa and covariates that potentially create an environment that enables the host to be more prone to disease. Understanding these associations between the microbiome and its environment can provide new insight to the cause, diagnosis and treatment of disease. Network analysis detects and quantifies taxa-taxa associations. Microbes commune with one another, which can modulate microbiome functions and host health. The methods discussed in this paper have been developed over the last decade due to the demand for statistical models that can handle the challenging characteristics of count data generated by high-throughput next-generation sequencing technology. At the very least, those challenges include zero inflation, overdispersion, sample heterogeneity, high dimensionality, correlation or partial correlation structure, technological variability, biological variability, normalization technique, and the compositional unit-sum constraint.

The available methods for analyzing microbiome data have greatly advanced metagenomic research. Great efforts have been made to account for the challenges of microbiome count data, to determine the normalization technique best suited for each model, and to assess model performance *via* available reference databases or synthetic data-generating tools. While some models do not account for the compositional nature of the data, it is suggested that this characteristic ought to be taken into consideration because the unit-sum constraint invalidates the assumption of independence of the data [87]. Future considerations should include methods for longitudinal studies, causal mediation analysis, and stochastic blocking, which are currently limited in the microbiome literature compared to other methods. Causal mediation analysis estimates the direct and indirect effects of predictor and mediating variables on the response variable [88]. An indirect effect is a relationship where there exists a pathway from a predictor variable to the response variable through a mediating variable; whereas, a direct effect is the relationship between only the predictor and response variables. For example, causal mediation could separate the effect of excessive alcohol consumption (predictor) on blood pressure (response) through a pathway such as BMI (mediator) [89]. The stochastic block model is an extension of network analysis where unsupervised learning is used to cluster the nodes of a network based on similar connectivity patterns [90, 91]. Unsupervised clustering methods can infer the number of clusters (or communities) that make up an entire network and the structure of taxa-taxa interactions within each community, which would then require scientific interpretation. Lastly, genomic reference databases need to be improved and updated because they are inadequate for the current needs of metagenomic research [92].

Microbiome research will continue to expand and present many complex statistical, scientific, and computing challenges. Future research must address these challenges in a collaborative effort of experts in statistics, science, and technology while building on the ideas from previous peer-reviewed research to offer reliable and interpretable solutions to the many important quests of microbiome research.

AUTHOR CONTRIBUTIONS

KL, SJ, XZ, and QL: conceptualization. KL, SJ, MN, ND, XZ, and QL: resources. KL and SJ: writing—original draft preparation. KL, MN, ND, XZ, and QL: writing—review and

editing. XZ and QL: supervision and project administration. All authors contributed to the article and approved the submitted version.

FUNDING

This study was partially supported by the National Institutes of Health (NIH) [1R01DK131267, 1R01GM140012, 1R01GM141519, 1R56HG011035, 5R01GM126479, P30CA142543, and P50CA070907], the Cancer Prevention and Research Institute of Texas [RP180319], and the Welch Foundation [AT-2030-20200401].

ACKNOWLEDGMENTS

The authors thank the chief editor, guest editor, and reviewer for their suggestions and feedback, which improved the paper significantly.

REFERENCES

1. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. The human microbiome project. *Nature*. (2007) 449:804–10. doi: 10.1038/nature06244
2. Amon P, Sanderson I. What is the microbiome? *Arch Dis Childhood Educ Pract*. (2017) 102:257–60. doi: 10.1136/archdischild-2016-311643
3. Zheng D, Liwinski T, Elinav E. Interaction between microbiota and immunity in health and disease. *Cell Res*. (2020) 30:492–506. doi: 10.1038/s41422-020-0332-7
4. Marchesi JR, Adams DH, Fava F, Hermes GD, Hirschfield GM, Hold G, et al. The gut microbiota and host health: a new clinical frontier. *Gut*. (2016) 65:330–9. doi: 10.1136/gutjnl-2015-309990
5. Peng X, Li G, Liu Z. Zero-inflated beta regression for differential abundance analysis with metagenomics data. *J Comput Biol*. (2016) 23:102–10. doi: 10.1089/cmb.2015.0157
6. Tang ZZ, Chen G. Zero-inflated generalized Dirichlet multinomial regression model for microbiome compositional data analysis. *Biostatistics*. (2019) 20:698–13. doi: 10.1093/biostatistics/kxy025
7. Yoon G, Gaynanova I, Müller CL. Microbial networks in SPRING-Semi-parametric rank-based correlation and partial correlation estimation for quantitative microbiome data. *Front Genet*. (2019) 10:516. doi: 10.3389/fgene.2019.00516
8. Evans C, Hardin J, Stoebel DM. Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. *Brief Bioinformatics*. (2018) 19:776–92. doi: 10.1093/bib/bbx008
9. Kim M, Chun J. 16S rRNA gene-based identification of bacteria and archaea using the EzTaxon server. *Methods Microbiol*. (2014) 41:61–74. doi: 10.1016/bs.mim.2014.08.001
10. Yarza P, Yilmaz P, Pruesse E, Glöckner FO, Ludwig W, Schleifer KH, et al. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nat Rev Microbiol*. (2014) 12:635–45. doi: 10.1038/nrmicro3330
11. Case RJ, Boucher Y, Dahllöf I, Holmström C, Doolittle WF, Kjelleberg S. Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies. *Appl Environ Microbiol*. (2007) 73:278–88. doi: 10.1128/AEM.01177-06
12. Ranjan R, Rani A, Metwally A, McGee HS, Perkins DL. Analysis of the microbiome: advantages of whole genome shotgun versus 16S amplicon sequencing. *Biochem Biophys Res Commun*. (2016) 469:967–77. doi: 10.1016/j.bbrc.2015.12.083
13. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: high-resolution sample inference from Illumina amplicon data. *Nat Methods*. (2016) 13:581–3. doi: 10.1038/nmeth.3869
14. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. *Nat Methods*. (2010) 7:335–6. doi: 10.1038/nmeth.f.303
15. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol*. (2009) 75:7537–41. doi: 10.1128/AEM.01541-09
16. Pereira MB, Wallroth M, Jonsson V, Kristiansson E. Comparison of normalization methods for the analysis of metagenomic gene abundance data. *BMC Genomics*. (2018) 19:274. doi: 10.1186/s12864-018-4637-6
17. Langille MG, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA, et al. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat Biotechnol*. (2013) 31:814–21. doi: 10.1038/nbt.2676
18. Badri M, Kurtz ZD, Müller CL, Bonneau R. Normalization methods for microbial abundance data strongly affect correlation estimates. *bioRxiv*. (2018) 2018:406264. doi: 10.1101/406264
19. Jiang S, Xiao G, Koh AY, Kim J, Li Q, Zhan X. A Bayesian zero-inflated negative binomial regression model for the integrative analysis of microbiome data. *Biostatistics*. (2019) 2019:kxz050. doi: 10.1093/biostatistics/kxz050
20. Lin H, Peddada SD. Analysis of microbial compositions: a review of normalization and differential abundance analysis. *NPJ Biofilms Microbiomes*. (2020) 6:1–13. doi: 10.1038/s41522-020-00160-w
21. Wang Y, LêCao KA. Managing batch effects in microbiome data. *Brief Bioinformatics*. (2020) 21:1954–70. doi: 10.1093/bib/bbz105
22. Zhang Y, Parmigiani G, Johnson WE. ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR Genomics Bioinformatics*. (2020) 2:lqaa078. doi: 10.1093/nargab/lqaa078
23. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. (2015) 43:e47. doi: 10.1093/nar/gkv007
24. Paulson JN, Stine OC, Bravo HC, Pop M. Differential abundance analysis for microbial marker-gene surveys. *Nat Methods*. (2013) 10:1200. doi: 10.1038/nmeth.2658
25. Dai Z, Wong SH, Yu J, Wei Y. Batch effects correction for microbiome data with Dirichlet-multinomial regression. *Bioinformatics*. (2019) 35:807–14. doi: 10.1093/bioinformatics/bty729
26. Leek JT. Svaseq: removing batch effects and other unwanted noise from sequencing data. *Nucl Acids Res*. (2014) 42:e161. doi: 10.1093/nar/gku864
27. Gagnon-Bartsch JA, Speed TP. Using control genes to correct for unwanted variation in microarray data. *Biostatistics*. (2012) 13:539–52. doi: 10.1093/biostatistics/kxr034
28. Sims AH, Smethurst GJ, Hey Y, Okoniewski MJ, Pepper SD, Howell A, et al. The removal of multiplicative, systematic bias allows integration of breast cancer gene expression datasets-improving meta-analysis and prediction of prognosis. *BMC Med Genomics*. (2008) 1:42. doi: 10.1186/1755-8794-1-42
29. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*. (2007) 8:118–27. doi: 10.1093/biostatistics/kxj037
30. Hornung R, Boulesteix AL, Causeur D. Combining location-and-scale batch effect adjustment with data cleaning by latent factor adjustment. *BMC Bioinformatics*. (2016) 17:27. doi: 10.1186/s12859-015-0870-z
31. Jacob L, Gagnon-Bartsch JA, Speed TP. Correcting gene expression data when neither the unwanted variation nor the factor of interest are observed. *Biostatistics*. (2016) 17:16–28. doi: 10.1093/biostatistics/kxv026
32. Gibbons SM, Duvallet C, Alm EJ. Correcting for batch effects in case-control microbiome studies. *PLoS Comput Biol*. (2018) 14:e1006102. doi: 10.1371/journal.pcbi.1006102
33. Alter O, Brown PO, Botstein D. Singular value decomposition for genome-wide expression data processing and modeling. *Proc Natl Acad Sci USA*. (2000) 97:10101–6. doi: 10.1073/pnas.97.18.10101
34. Marchesi JR, Dutilh BE, Hall N, Peters WH, Roelofs R, Boleij A, et al. Towards the human colorectal cancer microbiome. *PLoS ONE*. (2011) 6:e20447. doi: 10.1371/journal.pone.0020447
35. Karlsson FH, Tremaroli V, Nookaew I, Bergström G, Behre CJ, Fagerberg B, et al. Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature*. (2013) 498:99–103. doi: 10.1038/nature12198

36. Qin N, Yang F, Li A, Prifti E, Chen Y, Shao L, et al. Alterations of the human gut microbiome in liver cirrhosis. *Nature*. (2014) 513:59–64. doi: 10.1038/nature13568
37. Halfvarson J, Brislawn CJ, Lamendella R, Vázquez-Baeza Y, Walters WA, Bramer LM, et al. Dynamics of the human gut microbiome in inflammatory bowel disease. *Nat Microbiol*. (2017) 2:1–7. doi: 10.1038/nmicrobiol.2017.4
38. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. (2010) 26:139–40. doi: 10.1093/bioinformatics/btp616
39. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. (2014) 15:1–21. doi: 10.1186/s13059-014-0550-8
40. Mandal S, Van Treuren W, White RA, EggesbøM, Knight R, Peddada SD. Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microb Ecol Health Dis*. (2015) 26:27663. doi: 10.3402/mehd.v26.27663
41. Martin BD, Witten D, Willis AD. Modeling microbial abundances and dysbiosis with beta-binomial regression. *Ann Appl Stat*. (2020) 14:94. doi: 10.1214/19-AOAS1283
42. McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res*. (2012) 40:4288–97. doi: 10.1093/nar/gks042
43. Lê Cao KA, Costello ME, Lakis VA, Bartolo F, Chua XY, Brazeilles R, et al. MixMC: a multivariate statistical framework to gain insight into microbial communities. *PLoS ONE*. (2016) 11:e0160169. doi: 10.1371/journal.pone.0160169
44. Nueda MJ, Tarazona S, Conesa A. Next maSigPro: updating maSigPro bioconductor package for RNA-seq time series. *Bioinformatics*. (2014) 30:2598–602. doi: 10.1093/bioinformatics/btu333
45. Sun X, Dalpiaz D, Wu D, S Liu J, Zhong W, Ma P. Statistical inference for time course RNA-Seq data using a negative binomial mixed-effect model. *BMC Bioinformatics*. (2016) 17:324. doi: 10.1186/s12859-016-1180-9
46. Paulson JN, Talukder H, Bravo HC. Longitudinal differential abundance analysis of microbial marker-gene surveys using smoothing splines. *BioRxiv*. (2017) 2017:099457. doi: 10.1101/099457
47. Luo D, Ziebell S, An L. An informative approach on differential abundance analysis for time-course metagenomic sequencing data. *Bioinformatics*. (2017) 33:1286–92. doi: 10.1093/bioinformatics/btw828
48. Metwally AA, Yang J, Ascoli C, Dai Y, Finn PW, Perkins DL. MetaLonDA: a flexible R package for identifying time intervals of differentially abundant features in metagenomic longitudinal studies. *Microbiome*. (2018) 6:1–12. doi: 10.1186/s40168-018-0402-y
49. Zhang X, Yi N. NBZIMM: negative binomial and zero-inflated mixed models, with application to microbiome/metagenomics data analysis. *BMC Bioinformatics*. (2020) 21:488. doi: 10.1186/s12859-020-03803-z
50. Robinson MD, Smyth GK. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics*. (2008) 9:321–32. doi: 10.1093/biostatistics/kxm030
51. Kuo L, Mallick B. Variable selection for regression models. *Sankhyā: The Indian Journal of Statistics, Series B*. (1998) 65–81.
52. George EI, McCulloch RE. Variable selection *via* Gibbs sampling. *J Am Stat Assoc*. (1993) 88:881–9. doi: 10.1080/01621459.1993.10476353
53. Ferrari S, Cribari-Neto F. Beta regression for modelling rates and proportions. *J Appl Stat*. (2004) 31:799–815. doi: 10.1080/0266476042000214501
54. Aitchison J. The statistical analysis of compositional data. *J R Stat Soc Ser B*. (1982) 44:139–60. doi: 10.1111/j.2517-6161.1982.tb01195.x
55. Calgaro M, Romualdi C, Waldron L, Risso D, Vitulo N. Assessment of statistical methods from single cell, bulk RNA-seq, and metagenomics applied to microbiome data. *Genome Biol*. (2020) 21:1–31. doi: 10.1186/s13059-020-02104-1
56. Sánchez A, Fernández-Real J, Vegas E, Carmona F, Amar J, Burcelin R, et al. Multivariate methods for the integration and visualization of omics data. In: *Spanish Symposium on Bioinformatics*. Berlin; Heidelberg: Springer. (2010). p. 29–41. doi: 10.1007/978-3-642-28062-7_4
57. Metwally AA, Finn PW, Dai Y, Perkins DL. Detection of differential abundance intervals in longitudinal metagenomic data using negative binomial smoothing spline ANOVA. In: *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. Boston, MA (2017). p. 295–304. doi: 10.1145/3107411.3107429
58. Metwally AA, Aldirawi H, Yang J. A review on probabilistic models used in microbiome studies. *Commun Inform Syst*. (2018) 18:173–91. doi: 10.4310/CIS.2018.v18.n3.a3
59. Aldirawi H, Yang J, Metwally AA. Identifying appropriate probabilistic models for sparse discrete omics data. In: *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. Chicago, IL (2019). p. 1–4. doi: 10.1109/BHI.2019.8834661
60. Wang L, Aldirawi H, Yang J. Identifying zero-inflated distributions with a new R package iZID. *Commun Inform Syst*. (2020) 20:23–44. doi: 10.4310/CIS.2020.v20.n1.a2
61. Cragg JG. Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica*. (1971) 829–44. doi: 10.2307/1909582
62. Aldirawi H, Yang J. Modeling sparse data using MLE with applications to microbiome data. *J Stat Theory Pract*. (2022) 16:1–16. doi: 10.1007/s42519-021-00230-y
63. Li Q, Jiang S, Koh AY, Xiao G, Zhan X. Bayesian modeling of microbiome data for differential abundance analysis. *arXiv[Preprint].arXiv:190208741*. (2019). doi: 10.48550/arXiv.1902.08741
64. Levy M, Thaiss CA, Elinav E. Metabolites: messengers between the microbiota and the immune system. *Genes Dev*. (2016) 30:1589–97. doi: 10.1101/gad.284091.116
65. Visconti A, Le Roy CI, Rosa F, Rossi N, Martin TC, Mohney RP, et al. Interplay between the human gut microbiome and host metabolism. *Nat Commun*. (2019) 10:1–10. doi: 10.1038/s41467-019-12476-z
66. Koslovsky MD, Hoffman KL, Daniel CR, Vannucci M, et al. A Bayesian model of microbiome data for simultaneous identification of covariate associations and prediction of phenotypic outcomes. *Ann Appl Stat*. (2020) 14:1471–92. doi: 10.1214/20-AOAS1354
67. Chen J, Li H. Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis. *Ann Appl Stat*. (2013) 7:418–42. doi: 10.1214/12-AOAS592
68. Wadsworth WD, Argiento R, Guindani M, Galloway-Pena J, Shelburne SA, Vannucci M. An integrative Bayesian Dirichlet-multinomial regression model for the analysis of taxonomic abundances in microbiome data. *BMC Bioinformatics*. (2017) 18:94. doi: 10.1186/s12859-017-1516-0
69. Friedman J, Alm EJ. Inferring correlation networks from genomic survey data. *PLoS Comput Biol*. (2012) 8:e1002687. doi: 10.1371/journal.pcbi.1002687
70. Fang H, Huang C, Zhao H, Deng M. CCLasso: correlation inference for compositional data through Lasso. *Bioinformatics*. (2015) 31:3172–80. doi: 10.1093/bioinformatics/btv349
71. Ban Y, An L, Jiang H. Investigating microbial co-occurrence patterns based on metagenomic compositional data. *Bioinformatics*. (2015) 31:3322–9. doi: 10.1093/bioinformatics/btv364
72. Kurtz ZD, Müller CL, Miraldi ER, Littman DR, Blaser MJ, Bonneau RA. Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput Biol*. (2015) 11:e1004226. doi: 10.1371/journal.pcbi.1004226
73. Jiang S, Xiao G, Koh AY, Chen Y, Yao B, Li Q, et al. HARMONIES: a hybrid approach for microbiome networks inference *via* exploiting sparsity. *Front Genet*. (2020) 11:445. doi: 10.3389/fgene.2020.00445
74. Xia Y, Sun J, Chen DG. *Statistical Analysis of Microbiome Data With R*. Vol. 847. Singapore: Springer (2018). doi: 10.1007/978-981-13-1534-3
75. Liu L, Shih YCT, Strawderman RL, Zhang D, Johnson BA, Chai H. Statistical analysis of zero-inflated nonnegative continuous data: a review. *Stat Sci*. (2019) 34:253–79. doi: 10.1214/18-STS681
76. Faust K, Raes J. CoNet app: inference of biological association networks using Cytoscape. *F1000Research*. (2016) 5:1519. doi: 10.12688/f1000research.9050.2
77. Baba K, Shibata R, Sibuya M. Partial correlation and conditional correlation as measures of conditional independence. *Austr N Z J Stat*. (2004) 46:657–64. doi: 10.1111/j.1467-842X.2004.00360.x
78. Liu H, Roeder K, Wasserman L. Stability approach to regularization selection (StARS) for high dimensional graphical models. *Adv Neural Information Process Syst*. (2010) 24:1432. doi: 10.48550/arXiv.1006.3316
79. Meinshausen N, Bühlmann P. High-dimensional graphs and variable selection with the lasso. *Ann Stat*. (2006) 34:1436–62. doi: 10.1214/009053606000000281

80. Rong R, Jiang S, Xu L, Xiao G, Xie Y, Liu DJ, et al. MB-GAN: microbiome simulation *via* generative adversarial network. *GigaScience*. (2021) 10:giab005. doi: 10.1093/gigascience/giab005
81. Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. Shotgun metagenomics, from sampling to analysis. *Nat Biotechnol*. (2017) 35:833–44. doi: 10.1038/nbt.3935
82. Gu C, Kim GB, Kim WJ, Kim HU, Lee SY. Current status and applications of genome-scale metabolic models. *Genome Biol*. (2019) 20:1–18. doi: 10.1186/s13059-019-1730-3
83. Perez-Garcia O, Lear G, Singhal N. Metabolic network modeling of microbial interactions in natural and engineered environmental systems. *Front Microbiol*. (2016) 7:673. doi: 10.3389/fmicb.2016.00673
84. Dillard LR, Payne DD, Papin JA. Mechanistic models of microbial community metabolism. *Mol Omics*. (2021) 17:365–75. doi: 10.1039/D0MO00154F
85. Franzosa EA, McIver LJ, Rahnavard G, Thompson LR, Schirmer M, Weingart G, et al. Species-level functional profiling of metagenomes and metatranscriptomes. *Nat Methods*. (2018) 15:962–8. doi: 10.1038/s41592-018-0176-y
86. Roume H, Heintz-Buschart A, Muller EE, May P, Satagopam VP, Laczny CC, et al. Comparative integrated omics: identification of key functionalities in microbial community-wide metabolic networks. *NPJ Biofilms Microbiomes*. (2015) 1:1–11. doi: 10.1038/npjbiofilms.2015.7
87. Xia Y, Sun J. Hypothesis testing and statistical analysis of microbiome. *Genes Dis*. (2017) 4:138–48. doi: 10.1016/j.gendis.2017.06.001
88. Hicks R, Tingley D. Causal mediation analysis. *Stata J*. (2011) 11:605–19. doi: 10.1177/1536867X1201100407
89. Daniel RM, De Stavola BL, Cousens S, Vansteelandt S. Causal mediation analysis with multiple mediators. *Biometrics*. (2015) 71:1–14. doi: 10.1111/biom.12248
90. McDaid AF, Murphy TB, Friel N, Hurley NJ. Improved Bayesian inference for the stochastic block model with application to large networks. *Comput Stat Data Anal*. (2013) 60:12–31. doi: 10.1016/j.csda.2012.10.021
91. Aicher C, Jacobs AZ, Clauset A. Learning latent block structure in weighted networks. *J Complex Netw*. (2015) 3:221–48. doi: 10.1093/comnet/cnu026
92. Loeffler C, Karlsberg A, Martin LS, Eskin E, Koslicki D, Mangul S. Improving the usability and comprehensiveness of microbial databases. *BMC Biol*. (2020) 18:37. doi: 10.1186/s12915-020-0756-z

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.