frontiers in Genetics

ORIGINAL RESEARCH
published: 06 June 2019
doi: 10.3389/fgene.2019.00516

Microbial Networks in SPRING - Semi-parametric Rank-Based Correlation and Partial Correlation Estimation for Quantitative Microbiome Data

*Grace Yoon[1], Irina Gaynanova[1] and Christian L. Müller[2]**

[1] Department of Statistics, Texas A&M University, College Station, TX, United States, [2] Center for Computational Mathematics, Flatiron Institute, New York, NY, United States

OPEN ACCESS

Edited by:
Lingling An,
University of Arizona, United States

Reviewed by:
Yuan Jiang,
Oregon State University, United States
Michelle Lacey,
Tulane University, United States

***Correspondence:**
Christian L. Müller
cmueller@flatironinstitute.org

Specialty section:
This article was submitted to
Statistical Genetics and Methodology,
a section of the journal
Frontiers in Genetics

Received: 18 January 2019
Accepted: 13 May 2019
Published: 06 June 2019

Citation:
Yoon G, Gaynanova I and Müller CL
(2019) Microbial Networks
in SPRING - Semi-parametric
Rank-Based Correlation and Partial
Correlation Estimation for Quantitative
Microbiome Data.
Front. Genet. 10:516.
doi: 10.3389/fgene.2019.00516

High-throughput microbial sequencing techniques, such as targeted amplicon-based and metagenomic profiling, provide low-cost genomic survey data of microbial communities in their natural environment, ranging from marine ecosystems to host-associated habitats. While standard microbiome profiling data can provide sparse relative abundances of operational taxonomic units or genes, recent advances in experimental protocols give a more quantitative picture of microbial communities by pairing sequencing-based techniques with orthogonal measurements of microbial cell counts from the same sample. These tandem measurements provide absolute microbial count data albeit with a large excess of zeros due to limited sequencing depth. In this contribution we consider the fundamental statistical problem of estimating correlations and partial correlations from such quantitative microbiome data. To this end, we propose a semi-parametric rank-based approach to correlation estimation that can naturally deal with the excess zeros in the data. Combining this estimator with sparse graphical modeling techniques leads to the Semi-Parametric Rank-based approach for INference in Graphical model (SPRING). SPRING enables inference of statistical microbial association networks from quantitative microbiome data which can serve as high-level statistical summary of the underlying microbial ecosystem and can provide testable hypotheses for functional species-species interactions. Due to the absence of verified microbial associations we also introduce a novel quantitative microbiome data generation mechanism which mimics empirical marginal distributions of measured count data while simultaneously allowing user-specified dependencies among the variables. SPRING shows superior network recovery performance on a wide range of realistic benchmark problems with varying network topologies and is robust to misspecifications of the total cell count estimate. To highlight SPRING's broad applicability we infer taxon-taxon associations from the American Gut Project data and genus-genus associations from a recent quantitative gut microbiome dataset. We believe that, as quantitative microbiome profiling data will become increasingly available, the semi-parametric estimators for correlation and partial correlation estimation introduced here provide an important tool for reliable statistical analysis of quantitative microbiome data.

Keywords: absolute abundance, amplicon sequencing, association network, copula, graphical model, gut microbiome, zero inflation

1. INTRODUCTION

High-throughput sequencing techniques, including targeted amplicon-based sequencing (TAS) and metagenomic profiling, provide large-scale genomic survey data of microbial communities in their natural habitats. Collaborative efforts, such as the Human Microbiome Project (HMP) (Huttenhower et al., 2012), the Earth Microbiome Project (EMP) (Bahram et al., 2018), the TARA Ocean project (Sunagawa et al., 2015), and the American Gut Project (AGP) (McDonald et al., 2018) give an increasingly detailed picture of relative abundances of operational taxonomic units, their phylogenetic relationships, and gene abundances across diverse ecosystems, ranging from marine, soil, and fresh-water to human-associated habitats albeit at different scales and resolutions. Following the seminal work in Woese and Fox (1977), TAS protocols extract and amplify specific regions in marker genes, such as the 16S rRNA gene for bacteria and archea, the 18S rRNA gene for eukaryotes, and Internal Transcribed Spacer (ITS) regions for fungi, via universal primers followed by next-generation sequencing. These profiling efforts, together with elaborate bioinformatics processing and normalization work flows (Schloss et al., 2009; Caporaso et al., 2010; Edgar, 2013; Callahan et al., 2016; Lagkouvardos et al., 2017) allow low-cost determination of highly sparse relative counts of hundreds to thousands of operational taxonomic units (OTUs) or amplicon sequence variants (ASVs) (Edgar, 2016; Callahan et al., 2017) per sample across a large number of sample sites or participants. Metagenomic profiling (Handelsman, 2004) on the other hand provide unbiased samples of the majority of genes of the sampled habitat by high-throughput shotgun sequencing. Sophisticated reference-guided as well as reference-free metagenomic read assembly, binning, and taxonomic profiling pipelines (Alneberg et al., 2014; Sczyrba et al., 2017; Sedlar et al., 2017) can, under suitable conditions on read coverage, disentangle the complex mixture of sequencing reads into entire genomes of the underlying microbes and estimate, as a high-level by-product, relative microbial abundances.

Microbiome community-level analysis tasks, such as quantifying community composition shifts across conditions or associating high-dimensional species compositions and their taxonomic profiles to each other and to environmental or host-associated covariates, require statistical estimation procedures that can handle the restrictive nature of such sparse proportional (or compositional) microbiome datasets (Li, 2015). Important examples include differential abundance techniques (McMurdie and Holmes, 2014; Mandal et al., 2015), proportionality estimation (Quinn et al., 2017), regression models with compositional covariates (Holmes et al., 2012; Lin et al., 2014), composition-adjusted correlation estimation techniques (Friedman and Alm, 2012; Cao et al., 2018), and sparse graphical models for microbial association networks (Kurtz et al., 2015; Tipton et al., 2018).

Recent advancements in microbiome profiling protocols, however, promise to alleviate the experimental shortcomings of standard TAS or metagenomic experiments by enabling a more quantitative picture of microbial communities. The experimental protocols in Gifford et al. (2011) and Satinsky et al. (2013), originally introduced for marine microbiome profiling, establish quantitative count measurements of environmental metatranscriptomic or metagenomic data by adding orthogonal internal genomic mRNA or DNA standards (of known quantity) to the environmental sample prior to sequencing. A similar spike-in approach has been proposed for gut microbiome studies in Stämmler et al. (2016). Recent quantitative approaches combine TAS techniques with robust measurements of microbial cell counts, in particular flow cytometry (Props et al., 2017; Vandeputte et al., 2017). These tandem measurements provide absolute microbial count data albeit with a large number of zero measurements due to limited sequencing depth (see **Figure 2** for an overview). Thus far, however, statistical analysis methods for these novel quantitative microbiome data remain largely elusive.

In this contribution, we consider the statistical problem of correlation and partial correlation estimation for sparse quantitative microbiome count data. To this end, we first revisit a novel semi-parametric rank-based (SPR) approach to correlation estimation that can naturally deal with the large number of zeros in the data. The SPR estimator is easy to compute and can readily replace the naïve Pearson or rank-based sample correlation estimator which are often used as a first step in downstream statistical analysis tasks, including principal component analysis, principle coordinate analysis, discriminant analysis, or canonical correlation analysis (Yoon et al., 2018). Here we use the semi-parametric rank-based estimator as a starting point for sparse partial correlation estimation and introduce the Semi-Parametric Rank-based approach for INference in Graphical model (SPRING). SPRING follows the neighborhood selection methodology outlined in Meinshausen and Bühlmann (2006) to infer the conditional dependency graph and uses stability-based model selection (Liu et al., 2010; Müller et al., 2016) to identify a sparse set of stable partial correlation estimates from quantitative microbiome data (section 2). These partial correlations can be interpreted as direct (i.e., conditionally independent) statistical microbe-microbe associations and can serve as an initial community-level description of the underlying microbial ecosystem (Fuhrman et al., 2015; Sunagawa et al., 2015; Ruiz et al., 2017).

To evaluate our new methodology, we introduce a data generation mechanism that produces synthetic amplicon samples which exactly follow the empirical marginal cumulative distributions of measured amplicon count data while simultaneously obeying user-specified (partial) correlation dependencies among the variables and closely following user-defined total cell counts (see **Figure 2** for a summary). As ground-truth data for microbial associations remain largely elusive in current literature, our data generation mechanism might be of independent interest for testing other statistical inference schemes. We highlight SPRING's superior performance compared to standard sparse partial correlation estimation methods on a wide range of quantitative microbiome benchmark problems with varying prescribed network topologies. We also quantify, in the context of association network inference, the potential gains of quantitative over purely relative data even under misspecified totals. To showcase SPRING's broad applicability (see section 4), we first infer taxon-taxon associations from relative abundance data collected in the AGP using a pseudo-count-free log-ratio transform that can handle zero counts. Our key application is a genus-level analysis of the quantitative gut microbiome dataset put forward in Vandeputte et al. (2017). We discuss the inferred quantitative association network structure, compare it to published results, and assess, for the first time, the differences between inferred associations from measured absolute and relative abundance data in a consistent statistical framework. While we focus here on TAS-related applications, our methodology is broadly applicable to other data types with excess zeros, including quantitative metagenomics, single-cell RNA-seq, and mass spectrometry data, and thus provides a promising route toward a coherent statistical framework for correlation and partial correlation analysis of multi-omics biological data.

2. SEMI-PARAMETRIC RANK-BASED CORRELATION AND PARTIAL CORRELATION ESTIMATION

2.1. Rank-Based Estimation of Correlation Matrix for Zero-Inflated Data

A great number of multivariate statistical methods, such as principal component analysis, discriminant analysis, canonical and partial correlation analysis, to name a few, require the estimate of a covariance or correlation matrix of variables as one of the inputs. The overwhelming number of methods are based on the Pearson sample covariance matrix, which works well at capturing dependencies between variables that are normally distributed. One of the key challenges in analyzing TAS-based microbial abundance data is that it is far from normal: TAS-based measurements are inherently proportional, extremely right skewed, overdispersed, and comprise a large number of zero values. Furthermore, the zeros are not always indicative of the absence of the species, but rather a result of limited sequencing depth or primer bias. For these reasons, the sample covariance matrix is not appropriate for capturing dependencies present in microbiome data. Several methods use techniques from compositional data analysis (Aitchison, 1983), including log-ratio transforms, to adjust the data prior to any estimation, and enforce different structural constraints on the correlation or inverse correlation matrix (Friedman and Alm, 2012; Kurtz et al., 2015; Cao et al., 2018). The problem of excess zeros is typically dealt with by adding a small pseudo-count or, more recently, estimating pseudo-counts from multiple samples (Cao et al., 2017). For quantitative microbiome data, however, correlation and inverse correlation estimators are not yet available. In this work we propose to take a different approach relying on the recently proposed truncated Gaussian copula framework (Yoon et al., 2018).

First, we review the Gaussian copula model, which is sometimes referred to as non-paranormal (NPN) model (Liu et al., 2009).

Definition 1. *A random vector* $\mathbf{x} = (x_1, \ldots, x_p)^\top$ *satisfies the Gaussian copula model if there exists a set of monotonically increasing transformations* $f = (f_j)_{j=1}^{p}$ *satisfying* $f(\mathbf{x}) = \{f_1(x_1), \ldots, f_p(x_p)\}^\top \sim \mathrm{N}(\mathbf{0}, \mathbf{\Sigma})$ *with* $\sigma_{jj} = 1$. *We denote* $\mathbf{x} \sim \mathrm{NPN}(\mathbf{0}, \mathbf{\Sigma}, f)$.

The Gaussian copula model is commonly used in undirected graphical models (Liu et al., 2012; Fan et al., 2017) because it models the dependency between variables through the correlation matrix $\mathbf{\Sigma}$, and thus enjoys the mathematical simplicity of Gaussian multivariate distribution while relaxing the normality assumption. While the original model is only appropriate for modeling continuous variables, it has also been generalized to binary variables by adding an extra dichotomization step (Fan et al., 2017). The estimation of graphical models only requires the knowledge of the correlation matrix $\mathbf{\Sigma}$, and it has been shown (Fan et al., 2017) that consistent estimates of $\mathbf{\Sigma}$ could be easily obtained from sample Kendall's τ without the need to estimate unknown transformations f_j.

The Gaussian copula model is, however, not appropriate for quantitative microbiome data as (i) it does not take into account zero inflation, and (ii) it models continuous rather than count variables. To address (i), we take advantage of the model proposed in Yoon et al. (2018).

Definition 2 (Truncated Gaussian copula model of Yoon et al. (2018))**.** *A random vector* $\mathbf{x} = (x_1, \ldots, x_p)^\top$ *satisfies the truncated Gaussian copula model if there exists a p-dimensional random vector* $\mathbf{u} = (u_1, \ldots, u_p)^\top \sim \mathrm{NPN}(\mathbf{0}, \mathbf{\Sigma}, f)$ *such that*

$$x_j = I(u_j > c_j)u_j \quad (j = 1, \ldots, p),$$

where $I(\cdot)$ *is the indicator function and* $\mathbf{c} = (c_1, \ldots, c_p)$ *is a vector of positive constants.*

In other words, the model truncates a Gaussian copula variable so it is either zero or positive continuous. This model does not take into account that quantitative microbiome data have zeros or positive counts, but we found the continuous approximation to positive counts to work well in our simulation results (section 3).

To construct graphical models for the truncated Gaussian copula model, the estimation of the latent correlation matrix $\mathbf{\Sigma}$ is required. Yoon et al. (2018) develop a rank-based estimator for $\mathbf{\Sigma}$ by deriving the explicit form of the so-called bridge function F that connects the sample Kendall's τ estimates to the elements of $\mathbf{\Sigma}$. Given observed data $(x_{j1}, x_{k1}), \dots, (x_{jn}, x_{kn})$ for variables j and k, the sample Kendall's τ estimate is defined as

$$\widehat{\tau}_{jk} = \frac{2}{n(n-1)} \sum_{1 \le i < i' \le n} \text{sign}(x_{ji} - x_{ji'})\text{sign}(x_{ki} - x_{ki'}).$$

The bridge function F is defined so that $\mathbb{E}(\widehat{\tau}_{jk}) = F(\sigma_{jk})$, where σ_{jk} is the corresponding latent correlation between variables j and k. The explicit form of F for the truncated Gaussian copula model is given below.

Theorem 1 (Yoon et al. (2018)). *Let random variables x_j, x_k follow truncated Gaussian copula with corresponding latent correlation σ_{jk}. Then $\mathbb{E}(\widehat{\tau}_{jk}) = F(\sigma_{jk})$, where*

$$\begin{aligned} F(\sigma_{jk}) &= F(\sigma_{jk}; \delta_j, \delta_k) = -2\Phi_4(-\delta_j, -\delta_k, 0, 0; \mathbf{\Sigma}_{4a}) \\ &\quad + 2\Phi_4(-\delta_j, -\delta_k, 0, 0; \mathbf{\Sigma}_{4b}), \end{aligned}$$

$\delta_j = f_j(c_j)$, $\delta_k = f_k(c_k)$, $\Phi_4(\dots; \mathbf{\Sigma}_4)$ is the cumulative distribution function (cdf) of the four dimensional standard normal distribution with correlation matrix $\mathbf{\Sigma}_4$,

$$\mathbf{\Sigma}_{4a} = \begin{pmatrix} 1 & 0 & 1/\sqrt{2} & -\sigma_{jk}/\sqrt{2} \\ 0 & 1 & -\sigma_{jk}/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -\sigma_{jk}/\sqrt{2} & 1 & -\sigma_{jk} \\ -\sigma_{jk}/\sqrt{2} & 1/\sqrt{2} & -\sigma_{jk} & 1 \end{pmatrix}$$

and

$$\mathbf{\Sigma}_{4b} = \begin{pmatrix} 1 & \sigma_{jk} & 1/\sqrt{2} & \sigma_{jk}/\sqrt{2} \\ \sigma_{jk} & 1 & \sigma_{jk}/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & \sigma_{jk}/\sqrt{2} & 1 & \sigma_{jk} \\ \sigma_{jk}/\sqrt{2} & 1/\sqrt{2} & \sigma_{jk} & 1 \end{pmatrix}.$$

Moreover, $F(\sigma_{jk})$ is strictly increasing, so the inverse function $F^{-1}(\sigma_{jk})$ exists.

Remark 1. *To give more intuition for the form of the bridge function, we provide a brief summary of the underlying derivations here. The central part is the calculation of $\mathbb{E}\left\{\text{sign}(x_{ji} - x_{ji'})\text{sign}(x_{ki} - x_{ki'})\right\}$. Due to the effect of truncation, this calculation requires separation of events leading to zero or continuous realization of x_j before the equivalence $\text{sign}\{x_{ji} - x_{ji'}\} = \text{sign}\{f_1(x_{ji}) - f_1(x_{ji'})\}$ can be applied. This separation leads to the intersection of four events concerning normal variables (two events for continuous realization of x_j and x_k, and two events corresponding to each of the sign terms), thus explaining the appearance of the four-dimensional normal cdf in the form of the bridge function.*

Theorem 1 provides a closed-form expression of the bridge function F up to the values of thresholds δ_j, which we replace with moment-based estimators $\widehat{\delta}_j$. Let n_{0j} be the observed number of exact zeros across n realizations of variable x_j. By Definitions 1 and 2,

$$\mathbb{E}(n_{0j}/n) = P(x_j = 0) = P(u_j \le c_j) = P(f(u_j) \le \delta_j) = \Phi(\delta_j).$$

We use $\widehat{\delta}_j = \Phi^{-1}(n_{0j}/n)$ instead of δ_j and can thus calculate $\widehat{\sigma}_{jk} = F^{-1}(\widehat{\tau}_{jk})$. In practice, the inverse of the bridge function $F^{-1}(\widehat{\tau}_{jk})$ is determined numerically by finding the minimizer of the quadratic function $\{F(\sigma_{jk}) - \widehat{\tau}_{jk}\}^2$, which is unique due to the strict monotonicity of the function $F(\sigma_{jk})$.

The resulting $\widehat{\sigma}_{jk}$ are used to construct an element-wise estimator $\widehat{\Sigma}$. Since element-wise estimation does not guarantee positive semidefiniteness of $\widehat{\Sigma}$, we follow the suggestion of Fan et al. (2017) and replace $\widehat{\Sigma}$ with its projection onto the cone of positive semidefinite matrices. We use the `nearPD` function in `Matrix` R package to perform this projection. For numerical stability, we also include an additional shrinkage step of the form $\widetilde{\Sigma} = (1 - \rho)\widehat{\Sigma} + \rho I$ with $\rho = 0.01$, which guarantees strict positive definiteness of the final estimate. In simulations, we found that the method performs well across a wide range of small ρ values (see **Supplementary Material** for a sensitivity analysis of the parameter ρ). The described estimation procedure for Σ is implemented within the R package `mixedCCA` (Yoon and Gaynanova, 2018), and we refer the reader to Yoon et al. (2018) for more detailed derivations.

We refer to the proposed estimator $\widetilde{\mathbf{\Sigma}}$ of the correlation matrix $\mathbf{\Sigma}$ of truncated Gaussian copula variables as the Semi-Parametric Rank-based (SPR) correlation estimator. The SPR estimator forms the basis for the undirected graphical model framework outlined below.

2.2. Sparse Graphical Models and SPRING

We next introduce the Semi-Parametric Rank-based approach for INference in Graphical model (SPRING). SPRING relies on the estimation of an undirected graphical model from data. Undirected graphical models are typically used to represent the conditional independence relationship between the variables of random vector $\mathbf{x} \in \mathbb{R}^p$, so that

$$\text{no edge between } x_j \text{ and } x_k \iff x_j \perp x_k | \mathbf{x}_{-j,-k}\,,$$

where $\mathbf{x}_{-j,-k}$ means all components in $\mathbf{x}$ except component j and k. If the vector $\mathbf{x}$ follows a normal distribution, then conditional independence between x_j and x_k is equivalent to zero partial correlation between variables j and k. Therefore, sparse estimates of partial correlations lead to sparse conditional independence graphs. There is a rich literature on sparse estimation of partial correlations, with perhaps the most popular methods being the neighborhood selection of Meinshausen and Bühlmann (2006) (denoted by MB from here on) and the graphical lasso (Friedman et al., 2008). While the SPR estimator of the correlation matrix proposed in section 2.1 can be used in both approaches, we found the MB method to perform better than graphical lasso in numerical simulations and therefore focus on the MB method in the remainder of the paper.

The MB method takes advantage of the connection between partial correlations and regression coefficients and performs sparse estimation of partial correlations by regressing each of the p variables on the rest, thus finding each nodes' immediate neighbors by solving a lasso problem (Tibshirani, 1996). Given column-centered and scaled data matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$ with columns $\mathbf{x}^j$, the MB method solves for each variable j

$$\boldsymbol{\beta}^j = \underset{\boldsymbol{\beta}\in\mathbb{R}^p, \beta_j=0}{\operatorname{argmin}} \left\{ n^{-1}\|\mathbf{x}^j - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 \right\}.$$

Rewriting the objective function leads to

$$\begin{aligned}\boldsymbol{\beta}^j &= \underset{\boldsymbol{\beta}\in\mathbb{R}^p, \beta_j=0}{\operatorname{argmin}} \left\{ \boldsymbol{\beta}^\top n^{-1}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} - 2n^{-1}\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{x}^j + \lambda\|\boldsymbol{\beta}\|_1 \right\} \\ &= \underset{\boldsymbol{\beta}\in\mathbb{R}^p, \beta_j=0}{\operatorname{argmin}} \left\{ \boldsymbol{\beta}^\top\mathbf{S}\boldsymbol{\beta} - 2\boldsymbol{\beta}^\top\mathbf{s}^j + \lambda\|\boldsymbol{\beta}\|_1 \right\},\end{aligned}$$

where, given the centering and scaling of $\mathbf{X}$, $\mathbf{S} = n^{-1}\mathbf{X}^\top\mathbf{X}$ is the sample correlation matrix with columns $\mathbf{s}^j$. Since the standard sample correlation matrix is not suited for capturing dependencies in sparse quantitative microbiome data, SPRING replaces the sample correlation $\mathbf{S}$ in the MB method with the SPR estimator $\widetilde{\boldsymbol{\Sigma}}$ from section 2.1. The MB method comprises the regularization parameter λ which balances the trade-off between sparsity of the neighborhood and goodness of fit, and thus requires data-driven tuning. We here consider a stability-based model selection method, the Stability Approach to Regularization Selection (StARS) (Liu et al., 2010), which has been previously proven to be suitable for graphical model selection on microbiome data (Kurtz et al., 2015; Müller et al., 2016). The StARS method selects the optimal tuning parameter by repeatedly taking subsamples of the original data, estimating the graphical model for each subsample at each λ value along a prescribed regularization path, and then calculating empirical edge selection probabilities from the subsamples. The StARS edge stability criterion uses these probabilities to assess the sum of edge variabilities for each graph along the regularization path. The optimal λ is selected based on the supplied threshold t_S, with standard values being $t_S = 0.05$ and $t_S = 0.1$ (Liu et al., 2010; Kurtz et al., 2015). The threshold value represents a bound on the allowed overall edge variability over the entire graph. Lower thresholds lead to sparser, more robust graphs. Using the selected λ value, the final graphical model is refitted on the full dataset.

In summary, SPRING comprises three major components: (i) a semi-parametric rank-based correlation estimator for zero-inflated count data, (ii) the MB method to infer sparse conditional dependencies from the estimated correlation, and (iii) a stability-based approach (StARS) for sparse and robust neighborhood selection.

2.3. Extensions to Compositional Data

An important prerequisite for SPRING to be applicable to zero-inflated data is that individual count values across samples are comparable. For TAS-based microbial abundance data this condition is not satisfied because the total read count of a sample is not related to the total number of bacteria in the sample (Vandeputte et al., 2017), thus making the counts inherently proportional quantities. While this drawback is alleviated with the novel experimental techniques for quantitative microbiome data, as discussed earlier, a large number of available datasets, including the HMP and the AGP data, are only available as proportional (or compositional) data. To make SPRING amenable to statistical association inference from relative abundance data, we rely on a novel data transformation.

One of the key challenges in working with compositional data is the presence of unit-sum constraint. For correlation estimation, a common approach (see e.g., Aitchison, 1983; Kurtz et al., 2015; Cao et al., 2018) is to first apply the centered log-ratio transform (clr) to the compositional vector of each sample $\mathbf{x}_i \in \mathbb{S}^p$

$$\mathbf{z}_i = \operatorname{clr}(\mathbf{x}_i) = [\log\{x_{i1}/g(\mathbf{x}_i)\}, \log\{x_{i2}/g(\mathbf{x}_i)\}, \ldots, \log\{x_{ip}/g(\mathbf{x}_i)\}], \tag{1}$$

where $g(\mathbf{x}_i) = (\prod_{j=1}^p x_{ij})^{1/p}$ is the geometric mean of $\mathbf{x}_i$. A correlation matrix is then estimated based on the transformed $\mathbf{z}_i$, $i = 1, \ldots, n$, rather than directly on $\mathbf{x}_i$ (Aitchison, 1983). Since TAS-based microbiome profiling data have a large number of zeros, the addition of a large number of pseudo-counts is required to modify the vector of compositions to only have non-zero proportions. Adding such pseudo-counts changes the measured non-zero proportions and masks the zeros in the data, leading to zeros and non-zeros being treated equally in subsequent analysis. In addition, the choice of the actual value of the pseudo-count can influence downstream analysis results, and mere addition of extra zero components to the compositional vector would also change the transformation.

To avoid these drawbacks and to play on the strengths of SPRING in handling excess zeros, we propose a modified clr transform (mclr) that does not require the use of pseudo-counts. The key steps of the mclr transform are described below and visualized in **Figure 1**.

Contrary to recent efforts in data-driven inference of pseudo-counts (see e.g., Cao et al., 2017; de la Cruz and Kreft, 2018 and references therein), we compute the geometric mean of each sample from positive proportions only, normalize and log-transform all non-zero proportions by using that geometric mean, and apply an identical shift operation to all non-zero components in the dataset. Specifically, let $\mathbf{x}_i \in \mathbb{S}^p$ be the vector of compositions for sample i, and for simplicity of illustration, assume that the first q elements of $\mathbf{x}_i$ are zero, and the other elements are non-zero. Then we propose to apply

$$\begin{aligned}\mathbf{z}_i &= \operatorname{mclr}_\varepsilon(\mathbf{x}_i) = [0, \ldots, 0, \log\{x_{i(q+1)}/\tilde{g}(\mathbf{x}_i)\} + \varepsilon, \ldots, \\ &\quad \log\{x_{ip}/\tilde{g}(\mathbf{x}_i)\} + \varepsilon],\end{aligned} \tag{2}$$

where $\tilde{g}(\mathbf{x}_i) = (\prod_{j=q+1}^p x_{ij})^{1/(p-q)}$ is the geometric mean of the non-zero elements of $\mathbf{x}_i$. When $\varepsilon = 0$, mclr_0 corresponds to clr transform applied to non-zero proportions only (**Figure 1**, middle panel). When $\varepsilon > 0$, $\operatorname{mclr}_\varepsilon$ applies a positive shift to all non-zero compositions. To make all non-zero values strictly positive, we use the data-driven shift $\varepsilon = |z_{\min}| + c$, where $z_{\min} = \min_{ij} \log\{x_{ij}/\tilde{g}(\mathbf{x}_i)\}$ and c a positive constant with the default value $c = 1$. Alternative choices are discussed in the **Supplementary Material**. The ultimate rationale for the shift is to preserve the original ordering of the entries of the compositional vector $\mathbf{x}_i$ (with zeros being the smallest) in the transformed vector $\mathbf{z}_i$. The constraint $\varepsilon > |z_{\min}|$ ensures that $z_{i(q+1)}, \ldots, z_{ip}$ are

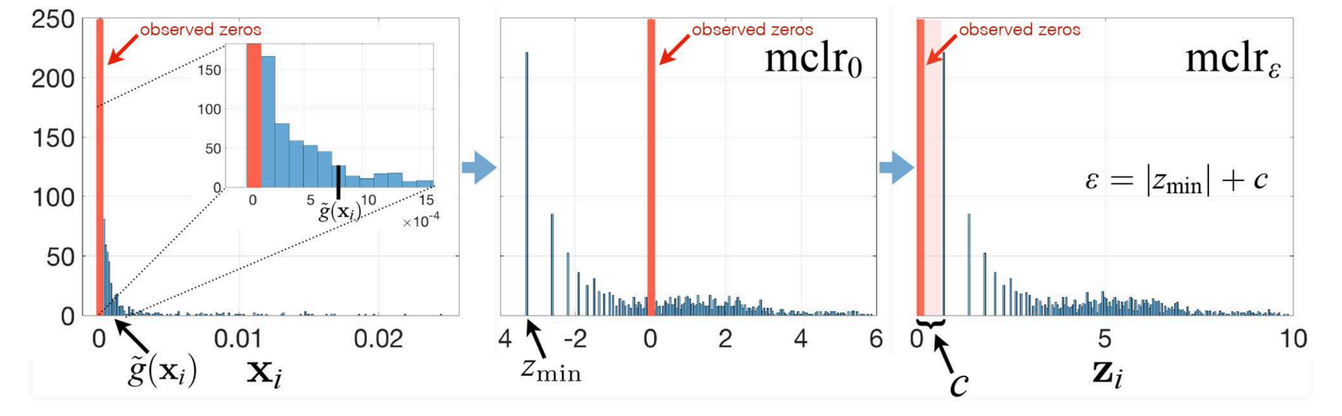

FIGURE 1 | Summary of the $\mathrm{mclr}_\varepsilon$ transform for a relative abundance sample $\mathbf{x}_i$. See main text for a description of the different steps.

strictly positive for all i. The modified clr transform is invariant to the addition of extra zero components, preserves the original zero measurements, and is overall rank-preserving.

If a practitioner intends to infer microbial associations from relative abundance data using SPRING, we suggest to first use the $\mathrm{mclr}_\varepsilon$ transform on relative abundance data and then apply SPRING to the transformed data. While SPRING is completely invariant to the choice of ε in $\mathrm{mclr}_\varepsilon$ for any value of ε within the constraint due to the rank-based estimation of correlation, it does not take into the account the compositional nature of the data. Alternative ways of measuring associations between compositional components include Aitchison's variation (Aitchison, 2003), linear compositional associations (Egozcue et al., 2018), and proportionality (Quinn et al., 2017), which take the compositional constraints directly into account. Here, we will focus on correlation-based approaches and present an application of SPRING to the compositional AGP data in section 4.1.

3. SIMULATION STUDIES

3.1. Generation of Synthetic Quantitative Microbial Abundance Data

We first describe generating mechanisms for synthetic microbial abundance data with prescribed correlation or inverse correlation matrices that emulates as close as possible quantitative microbial abundance data. We closely follow ideas presented in Kurtz et al. (2015) for synthetic data generation with several important differences. The work flow of our data generation mechanism is summarized in **Figure 2**.

We propose two constructions for correlation matrices. The first construction takes directly into account the covariance of measured quantitative microbial abundance data. Given a set of n quantitative abundance samples on p taxa $\mathbf{X} \in \mathbb{R}^{n\times p}$, we compute the SPR estimator $\widehat{\boldsymbol{\Sigma}}$ proposed in section 2.1 from the data and consider the resulting correlation matrix as the ground truth correlation matrix $\boldsymbol{\Sigma}$. The generation of synthetic samples given this correlation matrix estimate is then outlined below. Note that we do not impose any particular properties on the correlation matrix estimate, such as bounded condition number or sparsity. This construction is thus only useful for benchmarking different correlation estimation techniques.

An alternative way of generating a correlation matrix $\boldsymbol{\Sigma}$ is through explicitly controlling certain properties of the inverse correlation matrix. Let p be the number of nodes, i.e., the number of taxa or OTUs, and let $\boldsymbol{\Theta}$ be the p by p symmetric adjacency matrix such that $\theta_{ij} = 1$ if there is an edge between nodes i and j, $i \neq j$, and $\theta_{ij} = 0$ otherwise. We assume that the induced graph has no self-loops, i.e., $\theta_{ii} = 0$. We control the topology of the graph by considering three types of graph topologies: band graphs, cluster graphs, and scale-free graphs. The number of edges in the graph is denoted by e. The default value considered here is equal to twice the number of nodes ($e = 2p$), resulting in sparse graphs. Given this fixed sparsity level and the graph type, we use the R package `SpiecEasi` (Kurtz et al., 2017) to generate a precision matrix $\boldsymbol{\Omega}$ with the pattern of zeros corresponding to $\boldsymbol{\Theta}$. The non-zero entries of the lower triangular elements of $\boldsymbol{\Omega}$, ω_{ij} with $i > j$, are sampled uniformly at random from the intervals $[-3, -2]$ and $[2, 3]$, and the upper triangular elements are set to $\omega_{ji} = \omega_{ij}$. The diagonal elements are set to a constant such that the final precision matrix $\boldsymbol{\Omega}$ has a default condition number $\kappa = 100$. Using $\boldsymbol{\Omega}$, we generate the correlation matrix $\boldsymbol{\Sigma}$ by taking the inverse of the precision matrix, followed by scaling. This construction thus allows to benchmark different sparse inverse or partial correlation estimation techniques.

Given a correlation matrix $\boldsymbol{\Sigma}$ from either of the two constructions, we follow Kurtz et al. (2015) and use the "Normal to Anything" (NorTA) approach to generate synthetic abundance data. The NorTA method allows to generate variables with arbitrary marginal distributions from multivariate normal variables with given correlation structure. Specifically, we first generate $n \times p$ matrix $\mathbf{Z}$ with independent normal rows $\mathbf{z}_i \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$ with given correlation matrix $\boldsymbol{\Sigma}$, then get uniform random vectors by applying standard normal cdf transformation to each column of $\mathbf{Z}$, $\mathbf{u}^j = \Phi(\mathbf{z}^j)$ element-wise, and then apply the quantile functions of the target marginal distributions to each $\mathbf{u}^j$. In Kurtz et al. (2015), the zero-inflated negative binomial distribution (zinegbin) from `VGAM` package (Yee, 2010) is used, where the marginal distributional parameters are estimated from measured amplicon data. However, we

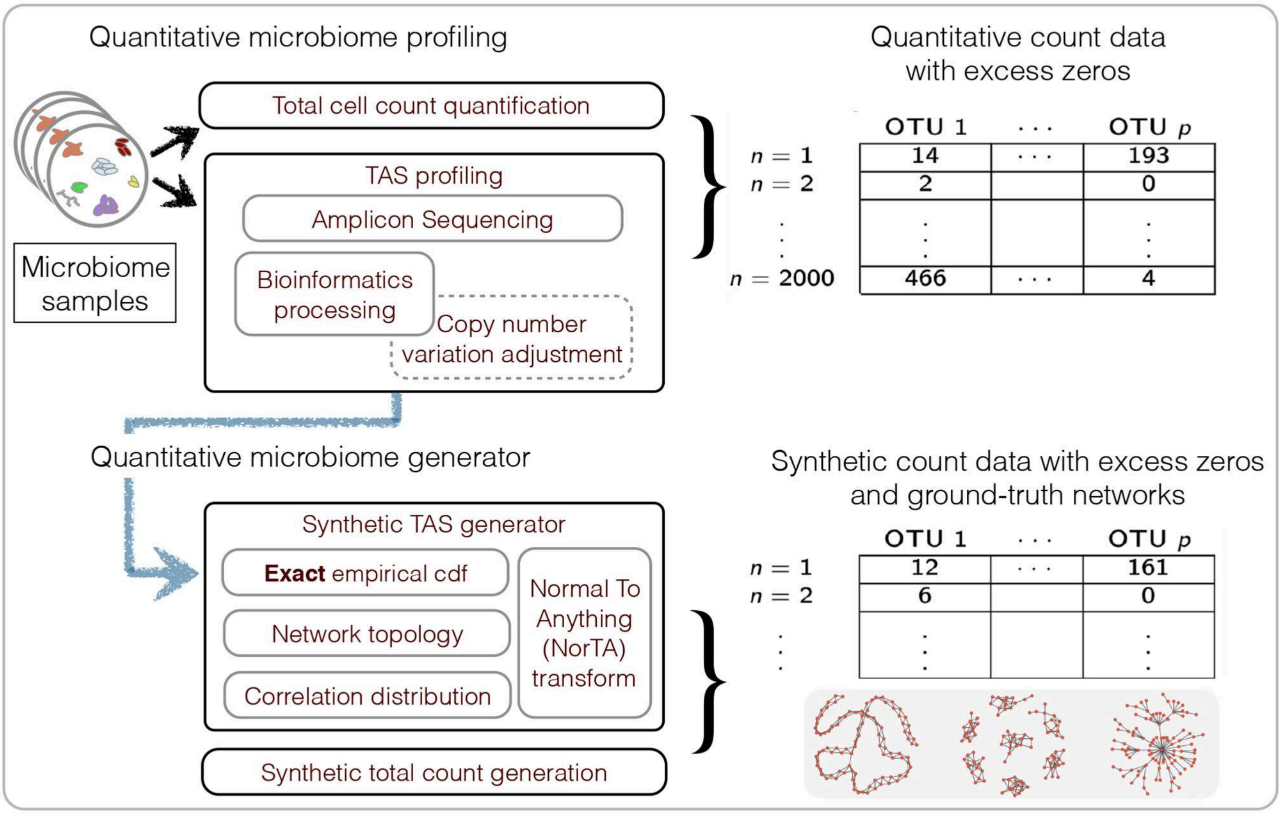
Quantitative microbiome profiling
Microbiome samples
Total cell count quantification
TAS profiling
Amplicon Sequencing
Bioinformatics processing
Copy number variation adjustment
Quantitative count data with excess zeros
OTU 1
...
OTU p
n = 1
n = 2
n = 2000
14
193
2
0
466
4
Quantitative microbiome generator
Synthetic TAS generator
Exact empirical cdf
Network topology
Correlation distribution
Normal To Anything (NorTA) transform
Synthetic total count generation
Synthetic count data with excess zeros and ground-truth networks
12
161
6
0

FIGURE 2 | Summary of the workflow for quantitative microbiome profiling (QMP) and the associated synthetic data generation mechanism.

found that the zinegbin distribution does not emulate well the overdispersion and skewness present in real data. This is evident by comparing the summary statistics between, e.g., the AGP data and corresponding synthetic data generated using the zinegbin, as shown in **Table 1**. To better match real amplicon data, we propose to take a different approach by using the inverse of the *empirical* cumulative distribution function (ecdf) of each OTU. This inverse can be calculated numerically by using the `uniroot.all` function in `rootSolve` package in `R` (Soetaert, 2009). As is evident from **Table 1**, the ecdf approach works well in mimicking the summary statistics of real TAS-based data. The match across all counts is considerably better than the match across sample abundances since the ecdf transformation is applied separately to each OTU. Although the within-sample counts are affected by the imposed correlation structure $\mathbf{\Sigma}$, the values of the sample total abundance of synthetic data with the ecdf are much closer to the measured ones than those with zinegbin. In terms of count summary statistics, the synthetic data is nearly indistinguishable from the measured data.

3.2. Estimation of Pairwise Correlations

3.2.1. Synthetic Data Generation and Methods for Comparison

We first benchmark estimation of pairwise correlations from synthetic quantitative microbial abundance data. For this purpose, we generate synthetic count data based on the quantitative microbiome profiling data, put forward in Vandeputte et al. (2017) and referred to as QMP data, and consider genus-level correlations. As the processed data used in Vandeputte et al. (2017) are not publicly available, we apply the work flow outlined in **Figure 2**. We reprocessed the available amplicon sequencing data using the standard QIIME protocol with closed-reference OTU picking (Caporaso et al., 2010), adjusted for copy number variations of the 16S rRNA gene using PICRUSt (Langille et al., 2013), filtered the data using the following three steps: (i) exclude samples whose sequencing depths (total read abundances) are $\leq$ 10000; (ii) exclude all taxa present in $<$30% of samples; and (iii) exclude samples whose abundance is less than the first percentile of all sequencing depths. We then combined the resulting samples with the corresponding measured total cell counts (Vandeputte et al., 2017). We next pooled $n = 106$ healthy subjects from the two available cohorts and merged all OTUs on the genus level, resulting in $p = 91$ genera. To generate synthetic data based on the QMP data with realistic correlation structure, we use the first construction method of the correlation matrix, outlined in section 3.3.1, thus considering the SPR correlation estimate on the QMP data as the ground-truth correlation matrix $\mathbf{\Sigma}$. We then generate $n = 91$ synthetic genus-level quantitative microbial abundance data that mimic the original QMP data both in terms of marginal genus distributions and correlation structure.

TABLE 1 | Comparison of summary statistics for all the counts and sample total abundance values between AGP data and two synthetic data generators.

Data	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Count data						
American Gut Project	0.0	0.0	3.0	144.9	34.0	176673.0
Synthetic (zinegbin)	0.0	0.0	33.0	170.0	125.0	54704.0
Synthetic (ecdf)	0.0	0.0	3.0	144.3	34.0	176673.0
Sample total abundance						
American Gut Project	10002.0	15149.2	20715.5	28989.0	32964.8	341632.0
Synthetic (zinegbin)	21696.0	30119.2	32543.5	33995.7	36068.0	86033.0
Synthetic (ecdf)	7354.0	18860.5	25095.5	28854.7	34285.2	196732.0

The sample size is n = 2000, the number of OTUs is p = 200, and the synthetic data is based on scale-free graph type.

In addition to the SPR correlation estimation (section 2.1) on the quantitative data, we consider three compositional correlation estimation approaches: (i) Pearson sample correlation on clr-transformed data with pseudocount addition [as used in SPIEC-EASI (Kurtz et al., 2015)], (ii) SparCC estimation from log-transformed compositions with pseudocount addition (Friedman and Alm, 2012), and (iii) SPR estimation on mclr$_\varepsilon$-transformed data (as described in section 2.3).

3.2.2. Results

We measure the performance of the different estimators in terms of absolute differences $|\sigma_{jk} - \widehat{\sigma}_{jk}|$, where σ_{jk} is the ground-truth correlation between genera j and k, and $\widehat{\sigma}_{jk}$ is the estimated correlation for each of the four methods. **Figure 3** shows box plots of absolute differences for the different methods. We observe that the SPR correlation estimates from the synthetic quantitative data outperform all other estimates, closely followed by the SPR estimates from mclr$_\varepsilon$-transformed data. SparCC and Pearson correlation on clr-transformed compositions are considerably outperformed by the SPR-type methods. The superiority of SPR-type methods is likely due to the preservation of the zero counts as zeros, thus avoiding distortions through the use of pseudo-counts, and the effective handling of the non-normality of the samples (as visible in the histogram of mclr$_\varepsilon$-transformed data in **Figure 1**). **Figure 4** shows the corresponding scatter plots of estimated and true pairwise correlations. We observe that SPR estimates on quantitative data are unbiased and have the smallest variance among all methods. SPR estimates on mclr$_\varepsilon$-transformed data have a slight downward bias and higher variance. SparCC and Pearson correlation on clr-transformed data have the worst performance both in terms of bias and variance.

3.3. Estimation of Microbial Association Networks

3.3.1. Synthetic Data Generation and Methods for Comparison

We next consider the estimation of microbial association networks. For this purpose, we generate synthetic counts from a large subset of the American Gut Project (AGP) data (McDonald et al., 2018), which comprises $p = 27116$ taxa across $n = 8440$ samples. The high dimensionality and the large sample size of the AGP data enable a more comprehensive and realistic investigation of the effects of dimensionality and sample size on the estimation of microbial associations than the QMP data. We consider the same data filtering steps as used in section 3.2.1: we (i) exclude samples whose sequencing depths (total read abundances) are ≤ 10000; (ii) exclude all taxa present in $<30\%$ of samples; and (iii) exclude samples whose abundance is less than the first percentile of all sequencing depths. This leads to a reduced dataset with $p = 481$ taxa across $n = 6482$. We consider two scenarios for the simulation studies: a large and a small sample size setting. For the large sample size setting, we randomly pick $n = 2000$ samples with total abundance at least 10,000, and then select $p = 100$ OTUs with largest abundances leading to 2000×100 matrix of synthetic counts. For the small sample size setting, we use the same strategy with $n = 500$ and $p = 200$. In the synthetic benchmarks, we treat the total observed read abundances as quantitative microbiome profiling abundances and impose sparse conditional dependencies on these counts by using the second correlation construction method, outlined in section 3.1. We refer to these samples as "True data" in the simulations. To investigate the robustness of SPRING to misspecifications of the assumed total, we also generate "Distorted data" by multiplying counts in every sample with an individual scale factor chosen uniformly at random from the interval $[0.5, 3]$. The scale factor does not affect a sample's compositional data but does distort the total abundances. The scale factor interval $[0.5, 3]$ represents a realistic distortion scenario in gut microbiome samples (see e.g., in Vandeputte et al., 2017, **Figure 2**) and is on the same order as typical fold changes of observed image-based total species counts in marine ecosystems (Ducklow, 2000). We study the performance of SPRING both on the "True" and "Distorted" synthetic data in order to assess how strongly a misspecification of the total affects association network inference.

Along with SPRING, we consider three methods for comparison. To study the influence of the sample correlation estimation, we consider the standard MB method using the Pearson sample correlation (Meinshausen and Bühlmann, 2006) [implemented in the R package huge (Zhao et al., 2012)]. We also consider two popular methods for microbial association inference from relative abundance data: SPIEC-EASI in the MB mode (Kurtz et al., 2015) and SparCC (Friedman and Alm, 2012)

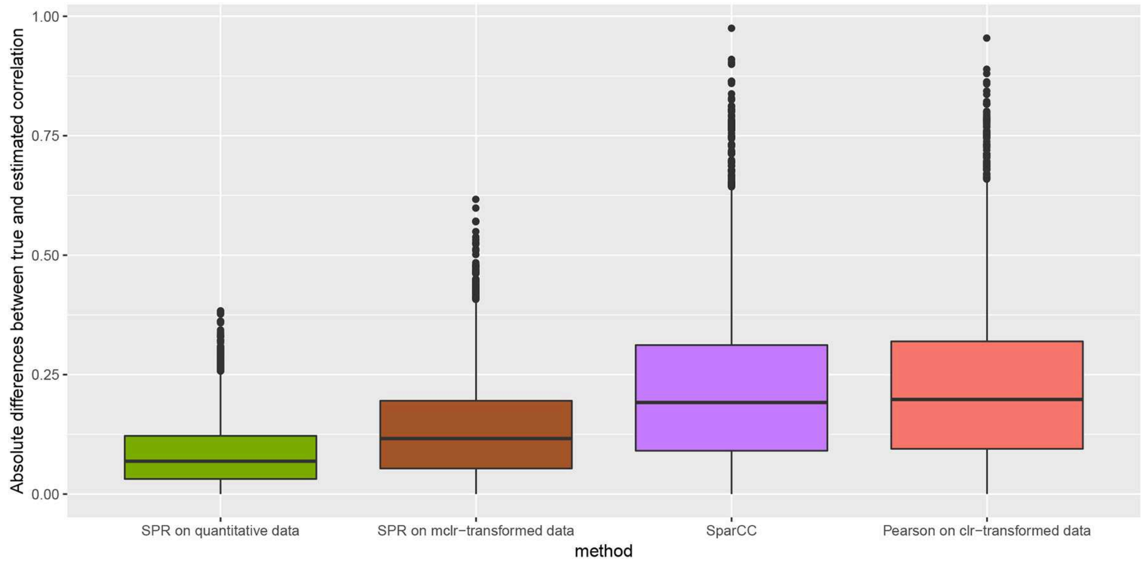

FIGURE 3 | Absolute differences between true and estimated correlation coefficients, $|\sigma_{jk} - \widehat{\sigma}_{jk}|$, for four methods: SPR correlation estimation on quantitative data (green), SPR correlation estimation on mclr$_{\varepsilon}$-transformed compositional data (brown), SparCC estimation (Friedman and Alm, 2012) (purple), and Pearson sample correlation on clr-transformed data (red) [as used in SPIEC-EASI (Kurtz et al., 2015)].

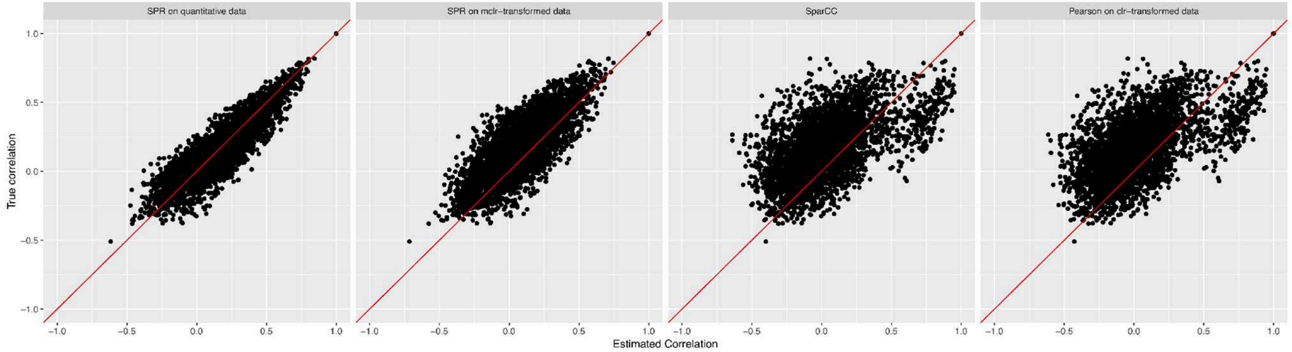

FIGURE 4 | True pairwise correlation values σ_{jk} (y-axis) vs. estimated values $\widehat{\sigma}_{jk}$ (x-axis) for four methods (from left to right): SPR correlation estimation on quantitative data, SPR correlation estimation on mclr$_{\varepsilon}$-transformed compositional data, SparCC estimation (Friedman and Alm, 2012), and Pearson sample correlation on clr-transformed data [as used in SPIEC-EASI (Kurtz et al., 2015)].

(both implemented in the R package `SpiecEasi`). The original SparCC method, however, is used for inferring marginal rather than conditional dependencies. For fair comparison with the other methods, we therefore introduce a modification of SparCC, termed invSparCC. The invSparCC method estimates the correlation matrix using the default SparCC method (as implemented in the R package `SpiecEasi`), and then uses the SparCC correlation estimator as input to the MB method, described in section 2.2. All considered methods use the neighborhood selection principle to derive a sparse graphical model, see **Table 2** for summary of all methods. The inferred adjacency and coefficient matrices are thus not guaranteed to be symmetric. We use the "or" rule and the "maxabs" rule to symmetrize the estimated adjacency and coefficient matrices, respectively. The "or" rule assigns an edge between nodes i and j if either node i is selected as a neighbor of j or node j is selected as a neighbor of i. The "maxabs" rule symmetrizes the coefficient matrix by taking the coefficient with maximum absolute value. For tuning parameter λ selection, we use the R package `pulsar` with "StARS" edge stability criterion and use 50 subsamples with subsampling ratio being fixed at $10\sqrt{n}/n$, where n is the sample size.

3.3.2. Results

We first compare the methods in terms of the Hamming distance between the true and the estimated graph. The Hamming distance is calculated as the number of edges that disagree with the true graph at each value of tuning parameter λ. The comparison of Hamming distance curves across the values of λ allows us to check the best achievable Hamming distance value that is agnostic to tuning parameter selection scheme. We consider 50 values of λ for all methods equally spaced on a logarithmic scale, with $\lambda_{\max}$ corresponding to no edges in the estimated graph, and $\lambda_{\min} = 0.01\lambda_{\max}$. For more accurate comparison, we consider 50 replications of the data generating process for each specified combination of n and p. The mean Hamming distance values over 50 replications as functions of λ are plotted in **Figure 5**, with bands corresponding to ± two standard errors. The MB method is uniformly outperformed by all methods, confirming that standard sample correlation is not suitable for capturing dependencies in sparse quantitative microbiome data. SPIEC-EASI and invSparCC have comparable performance, with SPIEC-EASI achieving smaller mean values. SPRING performs best in all cases considered here. The most challenging scenario is the scale-free graph with low sample size, with SPRING, SPIEC-EASI, and invSparCC having comparable performance. As expected, the distortion of total abundances has no effect on the compositional methods SPIEC-EASI and invSparCC, but decreases the performance of MB and SPRING. Nevertheless, the minimum Hamming distance achieved by SPRING on distorted data is still comparable or better than the minimum distances achieved by other methods, thus suggesting that SPRING is robust to misspecification of total abundance values.

To gain further insights into the edge selection performance of the different methods, we analyze the overlapping sets of selected edges for all methods. We here focus on the cluster graph type in the low sample size regime ($n = 500$, $p = 200$). For each method we select the tuning parameter λ using StARS at $t_S = 0.1$ and repeat the experiment over 50 replications. **Figure 6** shows the average number of edges that overlap across all methods as well as average proportions of true edges among the selected ones. Among all sets uniquely identified by an individual method, SPRING shows the highest true positive rate (0.72), followed by SPIEC-EASI (0.42), invSparCC (0.12), and MB (0.01). The edge set that is jointly selected by SPRING, SPIEC-EASI, and invSparCC shows the highest true positive rate (0.95) and highest number of selected edges ($\approx$ 246), followed by the edge set jointly selected by all four methods (true positive rate 0.94 and $\approx$ 54 edges). This suggests that a promising strategy for a practitioner screening for true statistical associations is to apply SPRING, SPIEC-EASI, and invSparCC independently and select the overlapping edge set.

Next, we consider one data replication and compare the Hamming distances achieved by selecting the tuning parameter λ using StARS. The results are shown in **Figure 7** with two StARS thresholds considered (stars indicating 0.1 and circles indicating 0.05). As expected, smaller threshold corresponds to larger tuning parameter leading to sparser graph. At the same time, based on numerical results, the threshold of 0.1 tends to reach smaller Hamming distances for all methods except MB. In general, both thresholds lead to reasonable values of λ in terms of Hamming distance. As in the previous comparison, SPRING leads to smaller Hamming distance values for "True" data and is robust to misspecified total abundance values.

Finally, we compare the estimated graphs from all methods in terms of precision and recall curves, where

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}};$$

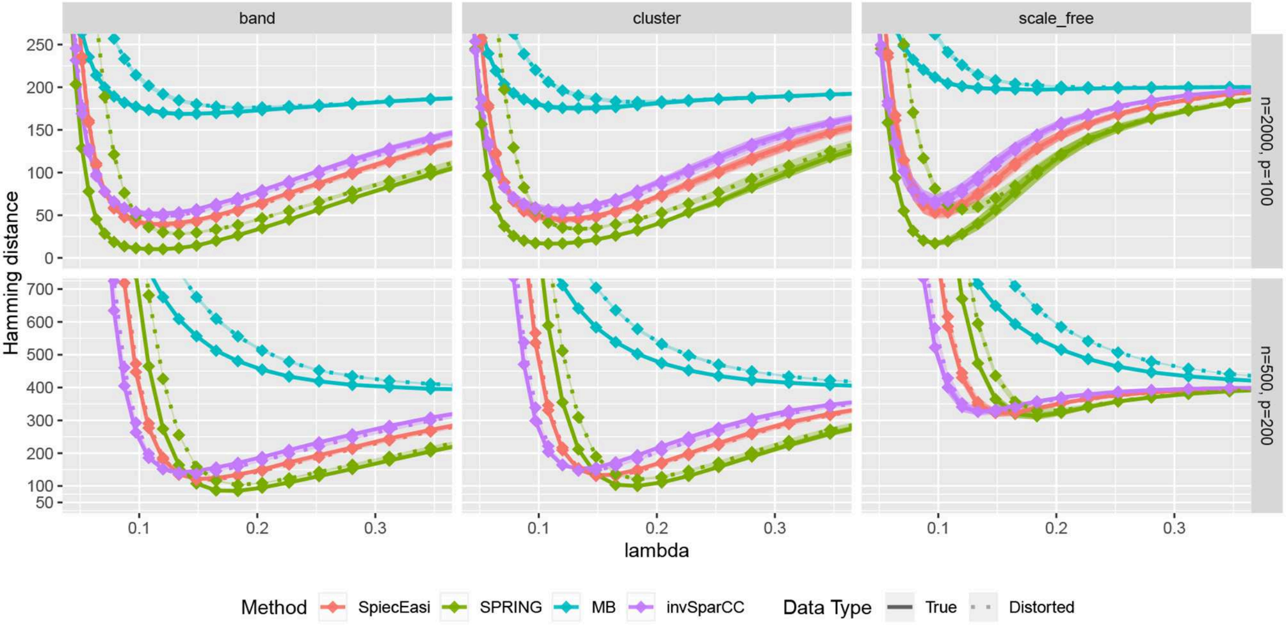

FIGURE 5 | Hamming distance as a function of tuning parameter λ. The lines correspond to mean values across 50 replications, and the bands show ± two standard errors. True abundance and distorted abundance data are distinguished by the transparency level and the line type: true data are less transparent and have solid lines; distorted data are more transparent and have dotted lines.

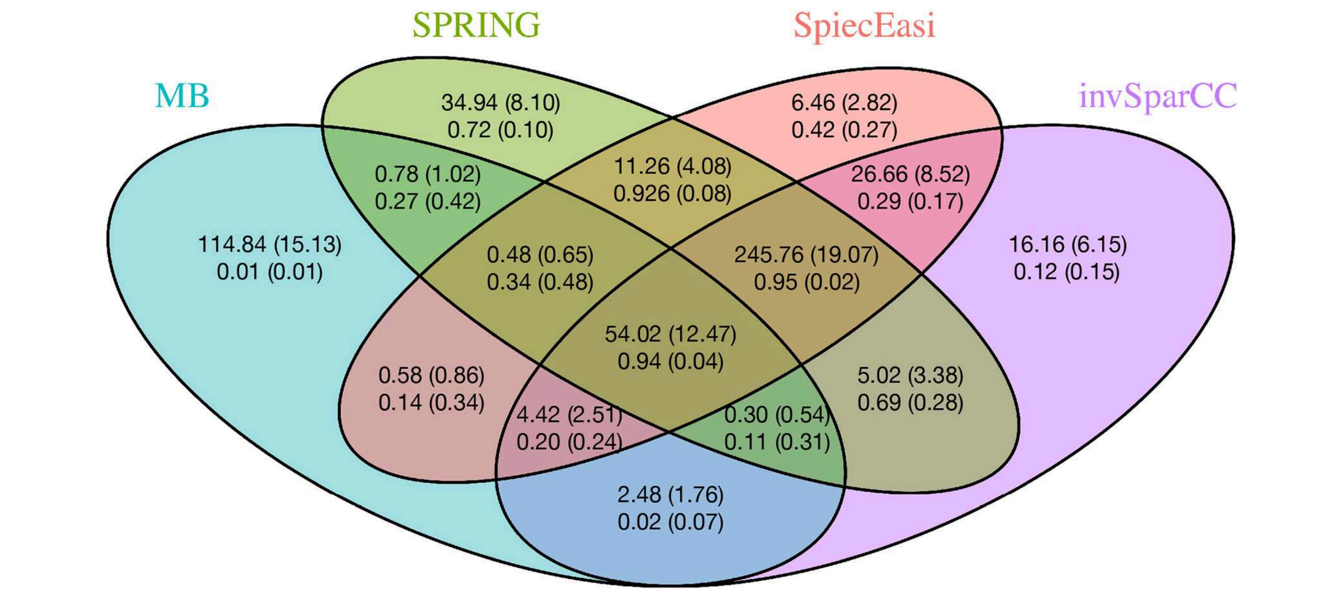

FIGURE 6 | Average number of overlapping edges (top row) and the average proportion of true edges in each corresponding overlap (bottom row) for four methods over 50 replications with $n = 500$, $p = 200$, and cluster-type graph. Corresponding standard deviations are given in parentheses.

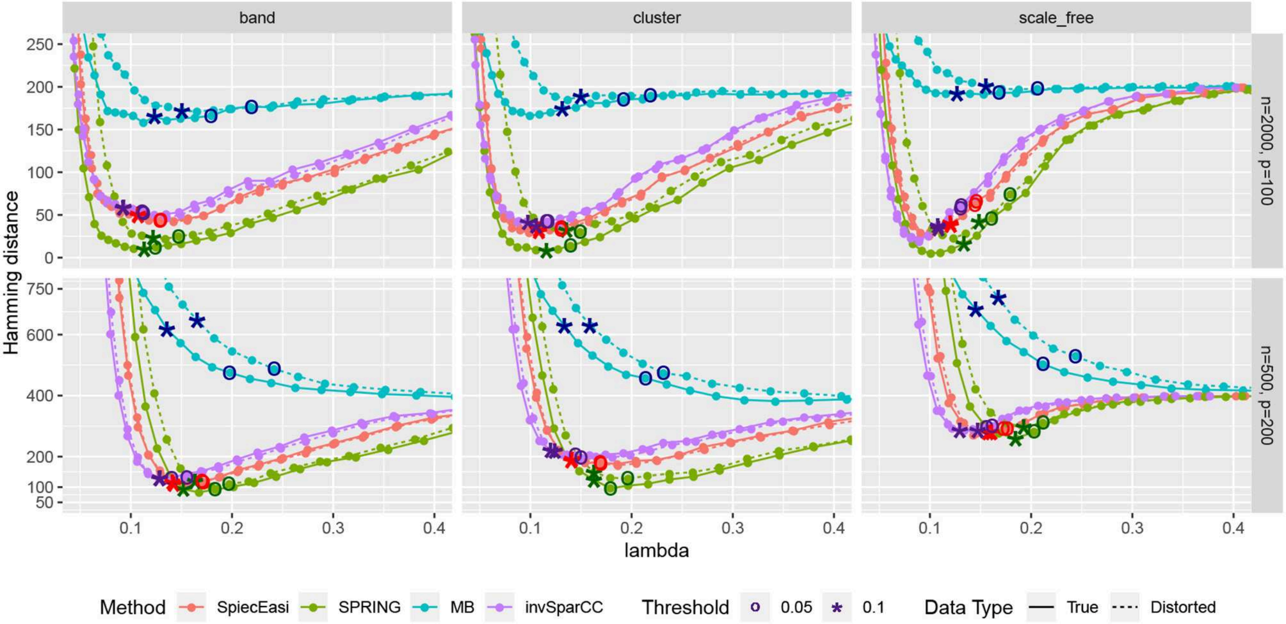

FIGURE 7 | Hamming distance as a function of tuning parameter λ. The distances at the tuning parameters selected by StARS are marked with star-shaped points ($t_S = 0.1$) and circle-shaped points ($t_S = 0.05$). True data are plotted with solid lines and distorted data are plotted with dotted lines.

TP, FP, and FN indicate the number of True Positives, False Positives, and False Negatives, respectively. To construct the curves, we extract the edge selection probabilities based on 50 subsamples from `pulsar` corresponding to tuning parameter with $t_S = 0.1$. We calculate precision and recall values by changing the threshold for edge selection probability from 1 to 0, interpolating the precision-recall values at the edges for no selection (recall = 0, precision = 1) and complete selection [recall = 1, precision = $4/(p-1)$]. Here $4/(p-1)$ is the probability of choosing true edges ($e = 2p$) at random among all possible edges ($p(p-1)/2$). The resulting curves are shown in **Figure 8**. For True data, SPRING achieves the highest precision-recall curves across all scenarios. The Area Under the Precision-Recall curve (AUPR) values are reported in **Table 3**. For the distorted data,

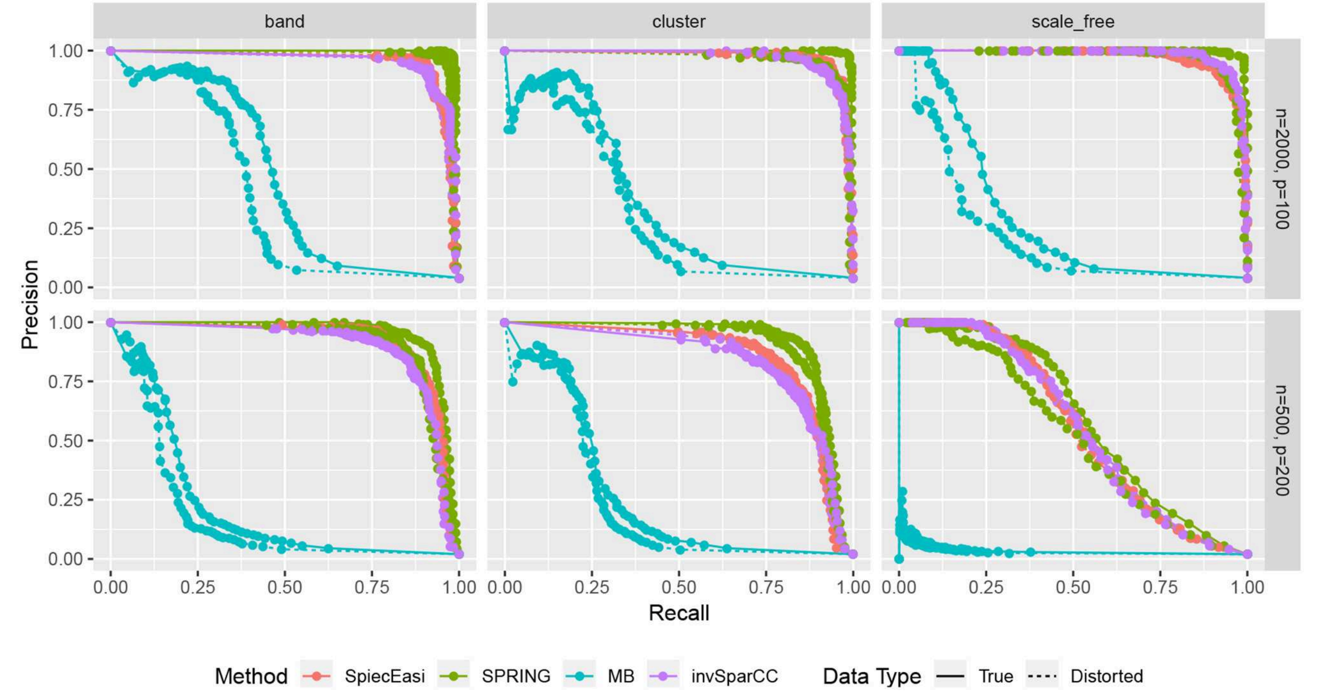

FIGURE 8 | Precision-Recall curves based on edge selection probabilities from pulsar with $t_S = 0.1$. True data are plotted with solid lines and distorted data are plotted with dotted lines.

SPRING is still best or among the best methods for band and cluster graph types, and is outperformed by the compositional methods for scale-free graph type in the low sample size regime.

In conclusion, SPRING exhibits considerably better graph recovery performance than existing methods, and is robust to misspecification of total sample abundance. This suggests that incorporating quantitative abundance information in the analysis leads to more reliable graphical model inference.

4. STATISTICAL MICROBIAL ASSOCIATIONS IN GUT MICROBIOME DATA

We provide two applications of SPRING to TAS-based microbial abundance data: a subset of the relative abundance data from the American Gut Project (AGP) (McDonald et al., 2018) and the QMP data from Vandeputte et al. (2017).

4.1. Taxon-Taxon Associations From the American Gut Project Data

We first use SPRING to infer taxon-taxon associations from the relative abundance AGP data. After the pruning and filtering steps described in section 3.3.1, we arrive at $p = 481$ OTUs from $n = 6482$ samples. Prior to applying SPRING, we transform the compositions $\mathbf{X} \in \mathbb{S}^{n\times p}$ using the mclr$_\epsilon$ transform introduced in Equation (2). The minimum value of the mclr$_0$-transformed data across all samples is $z_{\min} = -4.8142$. To make all non-zero values strictly positive, we add an arbitrary constant $c = 1$ to $|z_{\min}|$ and use the shift $\varepsilon = |z_{\min}| + c = 5.8142$ in the final mclr$_\varepsilon$ transform. We also consider SPIEC-EASI, MB, and invSparCC (see **Table 2**) for comparison. All four methods use the same parameterization for the regularization path and StARS model selection: 50 subsamples with the same seed number, subsampling ratio ($10\sqrt{n}/n = 0.1242$) and 50 tuning parameter values with the same ratio of the smallest to largest λ value ($\lambda_{\min}/\lambda_{\max} = 0.01$). For each method, $\lambda_{\max}$ is set to the maximum value of the off-diagonal elements of the respective correlation matrix. All computations were performed in R using the R packages `pulsar`, `SpiecEasi`, `huge`, and `mixedCCA`, respectively.

We report summary statistics of the estimated association networks for two StARS stability thresholds: 0.05 (the standard setting in `SpiecEasi`) and 0.1 (the standard setting in Liu et al., 2010) in **Table 4**. For both stability thresholds, the MB method estimates the sparsest networks with highest percentage of positive edges (PEP) while invSparCC estimates the densest networks with the lowest percentage of positive edges. SPRING and SPIEC-EASI's association networks have similar edge densities while SPRING has a considerably higher percentage of positive partial correlation edges.

To get a bird's eye view of the topologies of the different association networks we visualize the four different networks at StARS threshold 0.05 in **Figure 9A**. The force-directed layout of all networks follows the optimal layout of the SPRING network. At the selected StARS threshold, all networks have one connected component. The overall network structure suggests a dense core with two peripheral network modules, similar to previous analysis (Müller et al., 2016). The networks of the compositionally-adjusted methods SPIEC-EASI and invSparCC connect the core and one of the modules by a large number of positive (shown in green) and negative (shown in red) associations. SPRING considerably sparsifies these connections, leaving only few positive and negative edges between the modules, and MB does not infer any negative associations. We assess the similarity among the estimated networks by analyzing their edge set overlap in **Figure 9B**. All methods share common core of 601 edges. As expected, SPIEC-EASI and invSparCC share the largest unique two-set overlap with 637. SPRING's network takes an intermediate role between MB and the compositionally-adjusted methods. It shares 833 edges with SPIEC-EASI and invSparCC, and 112 edges exclusively with MB. Each method by itself also comprises a considerable set of exclusive edges, ranging from 418 for SPIEC-EASI to 767 for SPRING.

4.2. Genus-Genus Associations From Quantitative Gut Microbiome Profiling Data

We next analyze the quantitative gut microbiome data put forward in Vandeputte et al. (2017). We focus on estimating genus-genus associations both from the quantitative and the relative microbiome profiles, referred to as QMP and RMP, and analyze the consistency among the inferred networks. We follow the processing steps outlined in section 3.2.1 leading to $n = 106$ subjects and $p = 91$ genera. To infer statistical genus-genus associations we use SPRING for the QMP data (without transformation), and SPIEC-EASI for the corresponding RMP data (using the standard clr transformation) with the same computational protocol as detailed in the previous section.

We first show the agreement of signed edges between the two association networks at StARS stability level 0.1 in **Table 5**. Overall, out of the 4095 possible genus-genus associations, SPRING infers a set of 237 stable edges with a PEP of 98%. SPIEC-EASI infers 220 edges with a PEP of 66%. From the quantitative data, SPRING is able to detect considerably more positive associations, 140 of which are missed by SPIEC-EASI from the relative abundance data. SPRING detects only four negative associations three of which are missed by SPIEC-EASI despite having a considerable larger set of negative edges (74 overall). However, both methods do agree on a set of 93 edges, 92 positive and one negative edge. Importantly, we do not observe any sign flips among the different inferred edge sets. Missed positive or negative edges are simply absent in the other method.

We next focus on the induced genus-genus sub-network which only includes genera that have an assigned taxonomy and have at least one strong association $\geq |0.2|$ in either the SPRING-inferred or SPIEC-EASI-inferred association network. The weighted adjacency of this sub-network includes 32 genera and is shown in **Figure 10**. Among the 14 genera with highest total abundance across all samples (Bacteroides to Odoribacter), we observe 50% agreement between the two estimated networks (six edges are the same across all networks, three edges are different in SPIEC-EASI, four are different in SPRING). Both networks include a strong negative association between Phascolarctobacterium and Dialister and exactly four positive associations of Bacteroides with Parabacteroides, Holdemania, Bilophila, and Odoribacter (first row and column in **Figure 10**). We also observe the absence of a negative association between Bacteroides and Prevotella genera in the quantitative data which is often reported in the literature and also present in the SPIEC-EASI network (see also Vandeputte et al., 2017 for a discussion).

5. DISCUSSION

Advances in experimental microbiome profiling protocols have combined high-throughput environmental sequencing techniques with robust measurements of microbial cell counts

TABLE 2 | Summary of methods considered for comparison.

Method	Type of data	transformation	Correlation estimation
MB	Absolute abundance	None	Sample correlation
SPIEC-EASI	Relative abundance	clr	Sample correlation
invSparCC	Relative abundance	log	SparCC
SPRING	Absolute/relative abundance*	None/mclr*	SPR correlation

*For all methods, the final graphical model is estimated based on combining neighborhood selection approach with pulsar tuning parameter selection. *When absolute abundance data is not available, SPRING can be applied to relative abundance data following mclr transform described in section 2.3.*

TABLE 4 | AGP data: total number of partial correlation edges and percentage of positive partial correlation edges (PEP) (Faust et al., 2015) as estimated by MB, SPRING, SPIEC-EASI, and invSparCC for StARS stability thresholds $t_S = 0.05$ and 0.1.

	MB	SPRING	SPIEC-EASI	invSparCC
StARS threshold, t_S	Number of stable edges			
0.05	1621	2725	2702	3099
0.1	2970	4004	4008	4681
StARS threshold, t_S	Percentage of positive edges (PEP)			
0.05	1.0000	0.9835	0.8531	0.8341
0.1	0.9798	0.9515	0.7867	0.7584

TABLE 3 | Area under the Precision-Recall curves (AUPR) of **Figure 8**.

Dimension (*n*, *p*)	Graph type	SPIEC-EASI	SPRING	MB	invSparCC
(2000, 100)	Band	0.91 (0.92)	0.95 (0.94)	0.42 (0.34)	0.91 (0.91)
	Cluster	0.93 (0.93)	0.95 (0.92)	0.32 (0.27)	0.93 (0.93)
	Scale-free	0.93 (0.93)	0.96 (0.93)	0.26 (0.18)	0.94 (0.94)
(500, 200)	Band	0.89 (0.90)	0.93 (0.89)	0.20 (0.16)	0.87 (0.88)
	Cluster	0.83 (0.84)	0.90 (0.88)	0.25 (0.22)	0.81 (0.82)
	Scale-free	0.55 (0.54)	0.58 (0.50)	0.01 (0.01)	0.54 (0.54)

In each cell, AUPR of the True data and the Distorted data (given in parenthesis) are reported. AUPR value is based on edge selection probabilities using StARS with $t_S = 0.1$.

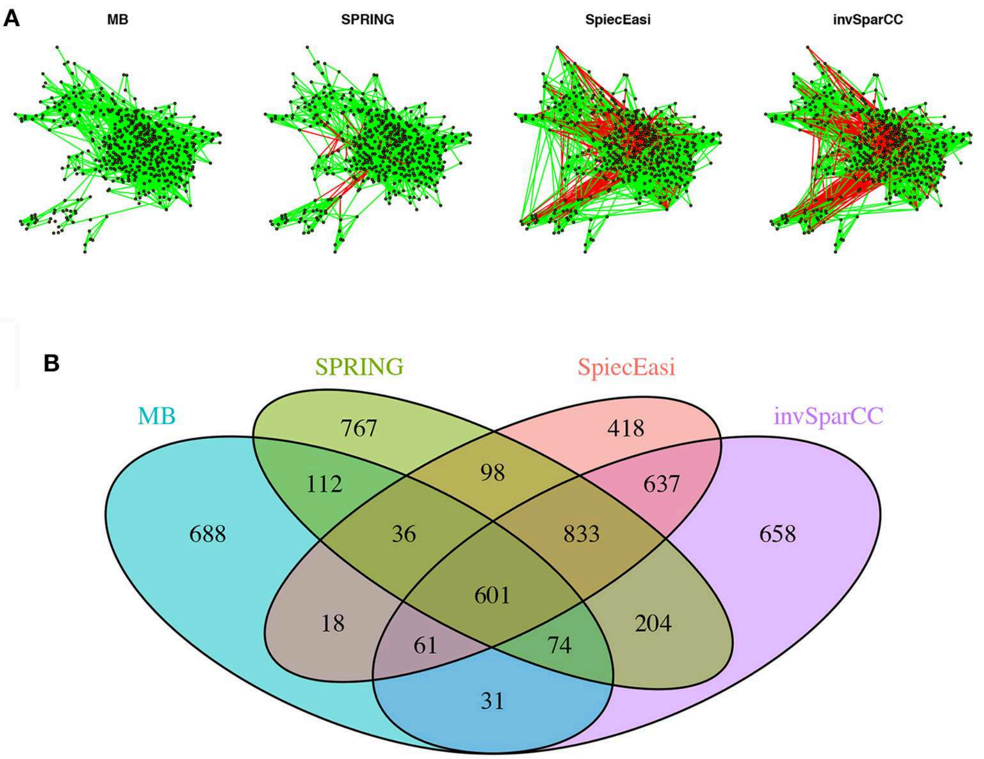

FIGURE 9 | Analysis of AGP data. **(A)** Force-directed layout (using `igraph` package in R) of the microbial association networks inferred by MB, SPRING, SPIEC-EASI, and invSparCC at StARS stability level 0.05. Green edges represent positive associations, red edges negative ones. **(B)** The four set Venn diagram showing the overlap of edge sets of the different association networks.

TABLE 5 | QMP data: summary of agreement of signed genus-genus partial correlations, inferred by SPRING and SPIEC-EASI at StARS stability threshold $t_S = 0.1$.

		SPRING		
Sign of estimated edges		**Positive**	**Zero**	**Negative**
SPIEC-EASI	Positive	92	54	0
	Zero	140	3731	4
	Negative	0	73	1

(Gifford et al., 2011; Satinsky et al., 2013; Stämmler et al., 2016; Props et al., 2017; Vandeputte et al., 2017; Tkacz et al., 2018), providing, for the first time, a more quantitative picture of the underlying microbial ecosystems in their natural habitat. To facilitate a high-level summary of the complex interplay between the constituents of the ecosystem, an important first exploratory analysis step is the estimation of statistical association networks between the identified operational taxonomic units or gene sets (Faust and Raes, 2012; Fuhrman et al., 2015; Sunagawa et al., 2015; Ruiz et al., 2017). In order to learn such association networks from sparse quantitative microbiome data, we have introduced the Semi-Parametric Rank-based approach for INference in Graphical model (SPRING). SPRING combines neighborhood selection (Meinshausen and Bühlmann, 2006) to infer the conditional dependency graph with stability-based model selection (Liu et al., 2010; Müller et al., 2016) to identify a sparse set of partial correlation estimates. The resulting network of partial correlations represents direct (i.e., conditionally independent) microbe-microbe associations and provides a statistical community-level description of the underlying microbial ecosystem. As ground truth microbial association networks are largely elusive in the literature, we have based our numerical simulation benchmarks on a novel synthetic quantitative microbiome data generation mechanism which might be of independent interest to researchers who want to test novel statistical techniques on such data.

Our benchmark test cases revealed a number of interesting observations. Firstly, we showed that, on synthetic quantitative microbiome data with prescribed ground-truth correlation structure, the SPR-type correlation estimates are considerably more accurate than SparCC and naive Pearson sample correlation on clr-transformed compositional data. Secondly, we showed that Pearson sample correlation estimation cannot be used to identify sparse partial correlations in quantitative microbiome data. Thirdly, SPRING outperformed sparse graphical modeling techniques that were designed with compositional data in mind, namely SPIEC-EASI (Kurtz et al., 2015) and the invSparCC estimator introduced here, which uses neighborhood selection with SparCC correlation estimation (Friedman and Alm, 2012). SPRING compared favorably to the other methods both in terms of achievability,

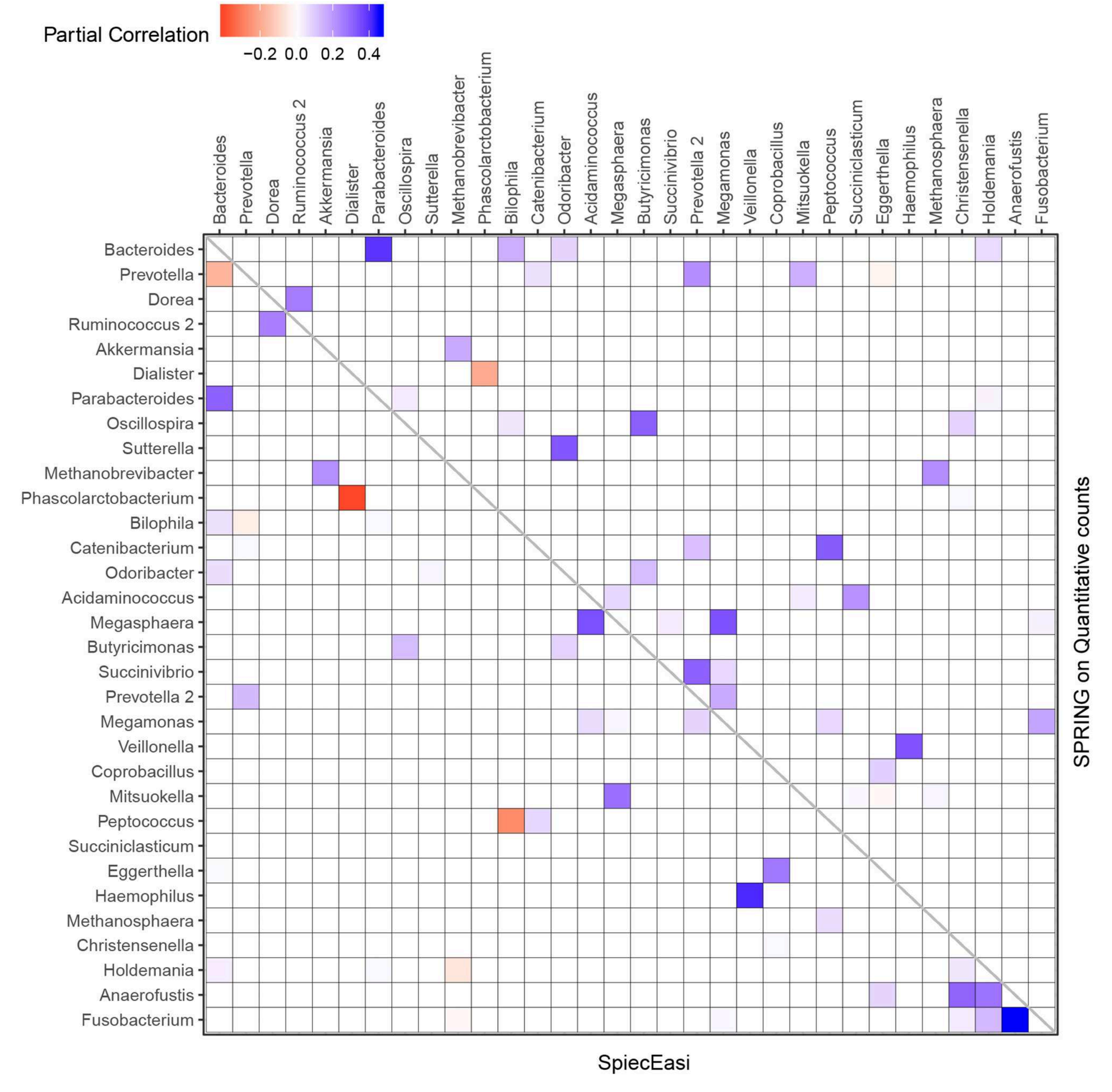

FIGURE 10 | Genus-genus association network using relative (lower triangular part) vs. quantitative count data (upper triangular part). Only genera with at least one strong association ≥ |0.2| in either SPIEC-EASI or SPRING are shown. The genera are ordered by the total quantitative abundance over healthy subjects (n = 106).

that is, in terms of minimum Hamming distance to the true underlying network achieved across the regularization path (see **Figure 7**), and in combination with stability-based model selection in terms of Precision-Recall (see **Figure 8**). We also quantified the robustness of SPRING to misspecification of the total by randomly distorting the counts of each sample up to a 6-fold change which represents a realistic distortion scenario in gut microbiome samples (see e.g., in Vandeputte et al., 2017, **Figure 2**) and is on the same order as typical fold changes of observed image-based total species counts in marine ecosystems (Ducklow, 2000). Even under these distortions SPRING's performance was on par or superior to SPIEC-EASI and invSparCC (which are scale-invariant by design). SPRING's robustness to total count misspecifications thus suggested to include an application of association inference from relative microbiome profiling data. In order to apply

SPRING to relative abundance data we introduced a modified centered log-ratio (clr) transform that can seamlessly handle excess zeros without pseudo-count addition. Contrary to recent efforts in data-driven pseudo-count inference (see de la Cruz and Kreft, 2018 and references therein) we computed the geometric mean of each sample from positive proportions only, normalized and log-transformed all non-zero proportions by using that geometric mean, and applied an identical shift operation to all non-zero variables in the dataset. This transformation is rank-preserving while leaving the original zero proportions unchanged, thus enabling the application of the SPRING methodology without further modification to relative abundance data.

We applied SPRING to two prominent gut microbial datasets, the relative abundance data collected in the American Gut Project (AGP) (McDonald et al., 2018) and the quantitative gut microbiome profiling (QMP) data from Vandeputte et al. (2017). As the processed data from Vandeputte et al. (2017) was not publicly available, a reprocessing of the amplicon sequencing reads was necessary.

From the AGP data, we inferred taxon-taxon association networks across $p = 481$ taxa from $n = 6482$ samples using neighborhood selection (MB), SPIEC-EASI, invSparCC, and SPRING. In line with previous findings (Faust et al., 2015), the percentage of positive edges in the networks is $> 75\%$, with MB and SPRING having even higher percentages than SPIEC-EASI and invSparCC. At both StARS stability levels 0.05 and 0.1 reported here, SPRING and MB tended to infer slightly sparser association networks than SPIEC-EASI and invSparCC. At StARS stability level 0.05, we analyzed the overlap of edge sets among the different methods (**Figure 9**). All methods share a common core of 601 edges. In addition, SPRING, SPIEC-EASI, and invSparCC shared the largest common edge set of size 833 among all three-set overlaps. As expected, the two compositionally-adjusted methods SPIEC-EASI and invSparCC shared the largest common two-set overlap of 637 edges. In the absence of verified taxon-taxon associations, our analysis suggests that a practitioner screening for coherent statistical associations among taxa can apply SPRING, SPIEC-EASI, and invSparCC independently and select the set of strongest edges out of the edge set these three methods inferred. This strategy is also supported by our synthetic benchmark results where the joint edge set of the three methods achieved a true positive rate of 0.95 for cluster graphs. For the analysis on the AGP data, this strategy would result in an edge set of size 1434, an average of about three associations per taxon. This core network can then be further studied in terms of modularity, network stability, and node centrality measures, as shown, e.g., in Ruiz et al. (2017); (Tipton et al., 2018).

For the QMP data, we used SPRING and SPIEC-EASI to estimate the genus-genus associations from the quantitative and the relative microbiome profiles, respectively. Our analysis revealed considerable differences to the published results in Vandeputte et al. (2017). The original study described dramatic differences between significant marginal genus-genus correlations from 66 healthy control samples in the QMP disease cohort when applying Spearman's ρ correlation to the relative and quantitative microbiome profiling data (see e.g., **Figure 3** in Vandeputte et al., 2017). Our results here showed more coherence of the statistical associations inferred from relative and absolute abundance data. Overall, 92 positive, 1 negative, as well as 3731 zero associations were in common among both association networks, while both networks differed in 280 associations (**Table 5**). Our analysis on the genus sub-network that comprised all genera with at least one strong association $\geq |0.2|$, shown in **Figure 10**, verified a strong negative association between Phascolarctobacterium and Dialister inferred from both data types, as well as the absence of a negative association between Bacteroides and Prevotella genera in the quantitative data, both in agreement with published results. However, we recovered, for both data types, exactly four positive associations for Bacteroides, namely with Parabacteroides, Holdemania, Bilophila, and Odoribacter (First row and column in **Figure 10**). The latter two associations were previously reported only to be present in the quantitative data. Overall, more than 30% of the edges in the sub-network agreed which is in marked contrast to the results reported in Vandeputte et al. (2017). The higher network consistency reported here can be attributed to several factors. Firstly, our amplicon data processing framework may result in slight differences in terms of OTU picking and avoids a rarefaction step which was included previously. Secondly, we considered partial rather than marginal correlations among the genera to avoid any influence of indirect associations. Thirdly, we analyzed both data types within the same coherent statistical learning framework: sparse learning of partial correlations via neighborhood selection followed by stability-based model selection with the identical stability threshold (here 0.1). Finally, we considered a larger sample size of $n = 106$ representing healthy subjects from two different cohorts available in the QMP data as opposed to the $n = 66$ samples used in the original study. We conclude that differences in association networks from relative and absolute abundance data are not only attributable to the data themselves but also highly method-dependent.

In summary, we believe that, as quantitative microbiome profiling will become increasingly available, the semi-parametric rank-based estimators for correlation and partial correlation estimation discussed here provide an important tool for reliable statistical analysis of quantitative microbiome data. While we have focused here on targeted amplicon-based sequencing datasets, our methodology is broadly applicable to other biological high-throughput data with large excess of zero counts, including quantitative metagenomics (Satinsky et al., 2013), single-cell RNA-Seq data (see Risso et al., 2018 for a recent statistical analysis framework), and mass spectrometry proteomics data (Drew et al., 2017). Moreover, the concept of SPR-type correlation employed in SPRING can naturally generalize to joint analysis of multi-omics dataset when, on the same sample, several zero-inflated data types are measured in tandem. The approach in Yoon et al. (2018) already exploits this idea for RNA-seq and micro-RNA data in the context of canonical correlation analysis. Extending SPRING in a similar way to joint graphical modeling of mixed data types is a promising next step toward a consistent and coherent statistical analysis framework for sparse high-throughput biological datasets.

AUTHOR CONTRIBUTIONS

GY, IG, and CM developed the methodology. GY lead the numerical analysis. IG assisted GY in simulation studies. CM assisted GY in data analyses. GY, IG, and CM prepared the manuscript.

FUNDING

GY was supported by the National Cancer Institute grant T32-CA090301. IG was supported by the National Science Foundation grant DMS-1712943. CM work was supported by the Simons Foundation.

ACKNOWLEDGMENTS

We are grateful to Dr. Zachary D. Kurtz, Lodo Therapeutics, for providing the processed American Gut data and the QMP data.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00516/full#supplementary-material

REFERENCES

Aitchison, J. (1983). Principal component analysis of compositional data. *Biometrika* 70, 57–65.

Aitchison, J. (2003). "A concise guide to compositional data analysis," in *2nd Compositional Data Analysis Workshop* (Girona).

Alneberg, J., Bjarnason, B. S., De Bruijn, I., Schirmer, M., Quick, J., Ijaz, U. Z., et al. (2014). Binning metagenomic contigs by coverage and composition. *Nat. Methods* 11, 1144–1146. doi: 10.1038/nmeth.3103

Bahram, M., Hildebrand, F., Forslund, S. K., Anderson, J. L., Soudzilovskaia, N. A., Bodegom, P. M., et al. (2018). Structure and function of the global topsoil microbiome. *Nature* 560, 233–237. doi: 10.1038/s41586-018-0386-6

Callahan, B. J., McMurdie, P. J., and Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J.* 11, 2639–2643. doi: 10.1038/ismej.2017.119

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., and Holmes, S. P. (2016). DADA2: high-resolution sample inference from Illumina amplicon data. *Nat. Methods* 13:581. doi: 10.1038/nmeth.3869

Cao, Y., Lin, W., and Li, H. (2018). Large covariance estimation for compositional data via composition-adjusted thresholding. *J. Am. Stat. Assoc.* 1–14. doi: 10.1080/01621459.2018.1442340

Cao, Y., Zhang, A., and Li, H. (2017). Microbial composition estimation from sparse count data. *ArXiv e-prints*. Available online at: http://arxiv.org/abs/1706.02380 (accessed May 27, 2019).

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., et al. (2010). QIIME allows analysis of high- throughput community sequencing data. *Nature* 7, 335–336. doi: 10.1038/nmeth.f.303

de la Cruz, R., and Kreft, J.-U. (2018). Geometric mean extension for data sets with zeros. *ArXiv e-prints*. Available online at: https://arxiv.org/abs/1806.06403 (accessed May 27, 2019).

Drew, K., Müller, C. L., Bonneau, R., and Marcotte, E. M. (2017). Identifying direct contacts between protein complex subunits from their conditional dependence in proteomics datasets. *PLoS Comput. Biol.*, 13:e1005625. doi: 10.1371/journal.pcbi.1005625

Ducklow, H. W. (2000). "Bacterial production and biomass in the oceans," in *Microbial Ecology of the Oceans*, ed D. L. Kirchman (New York, NY: Wiley-Liss), 85–120.

Edgar, R. C. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat. Methods* 10, 996–998. doi: 10.1038/nmeth.2604

Edgar, R. C. (2016). UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. *bioRxiv*. Available online at: https://www.biorxiv.org/content/early/2016/10/15/081257 (accessed May 27, 2019).

Egozcue, J. J., Pawlowsky-Glahn, V., and Gloor, G. B. (2018). Linear association in compositional data analysis. *Aust. J. Stat.* 47:3. doi: 10.17713/ajs.v47i1.689

Fan, J., Liu, H., Ning, Y., and Zou, H. (2017). High dimensional semiparametric latent graphical model for mixed data. *J. R. Stat. Soc. B* 79, 405–421. doi: 10.1111/rssb.12168

Faust, K., Lima-Mendez, G., Lerat, J.-S., Sathirapongsasuti, J. F., Knight, R., Huttenhower, C., et al. (2015). Cross-biome comparison of microbial association networks. *Front. Microbiol.* 6:1200. doi: 10.3389/fmicb.2015.01200

Faust, K., and Raes, J. (2012). Microbial interactions: from networks to models. *Nat. Rev. Microbiol.* 10, 538–550. doi: 10.1038/nrmicro2832

Friedman, J., and Alm, E. J. (2012). Inferring correlation networks from genomic survey data. *PLoS Comput. Biol.* 8:e1002687. doi: 10.1371/journal.pcbi.1002687

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* 9, 432–441. doi: 10.1093/biostatistics/kxm045

Fuhrman, J., Cram, J., and Needham, D. M. (2015). Marine microbial community dynamics and their ecological interpretation. *Nat. Rev. Microbiol.* 13, 133–146. doi: 10.1038/nrmicro3417

Gifford, S. M., Sharma, S., Rinta-Kanto, J. M., and Moran, M. A. (2011). Quantitative analysis of a deeply sequenced marine microbial metatranscriptome. *ISME J.* 5, 461–472. doi: 10.1038/ismej.2010.141

Handelsman, J. (2004). Metagenomics: application of genomics to uncultured microorganisms. *Microbiol. Mol. Biol. Rev.* 68, 669–685. doi: 10.1128/MMBR.68.4.669-685.2004

Holmes, I., Harris, K., and Quince, C. (2012). Dirichlet multinomial mixtures: Generative models for microbial metagenomics. *PLoS ONE* 7:e30126. doi: 10.1371/journal.pone.0030126

Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J. H., Chinwalla, A. T., et al. (2012). Structure, function and diversity of the healthy human microbiome. *Nature* 486, 207–214. doi: 10.1038/nature11234

Kurtz, Z., Mueller, C., Miraldi, E., and Bonneau, R. (2017). *SpiecEasi: Sparse Inverse Covariance for Ecological Statistical Inference*. R package version 1.0.2.

Kurtz, Z. D., Müller, C. L., Miraldi, E. R., Littman, D. R., Blaser, M. J., and Bonneau, R. A. (2015). Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput. Biol.* 11:e1004226. doi: 10.1371/journal.pcbi.1004226

Lagkouvardos, I., Fischer, S., Kumar, N., and Clavel, T. (2017). Rhea: a transparent and modular R pipeline for microbial profiling based on 16S rRNA gene amplicons. *PeerJ* 5:e2836. doi: 10.7717/peerj.2836

Langille, M. G., Zaneveld, J., Caporaso, J. G., McDonald, D., Knights, D., Reyes, J. A., et al. (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat. Biotechnol.* 31, 814–821. doi: 10.1038/nbt.2676

Li, H. (2015). Microbiome, metagenomics and high-dimensional compositional data analysis. *Annu. Rev. Stat. Appl.* 2, 73–94. doi: 10.1146/annurev-statistics-010814-020351

Lin, W., Shi, P., Feng, R., and Li, H. (2014). Variable selection in regression with compositional covariates. *Biometrika* 101, 785–797. doi: 10.1093/biomet/asu031

Liu, H., Han, F., Yuan, M., Lafferty, J. D., and Wasserman, L. (2012). High-dimensional semiparametric Gaussian copula graphical models. *Ann. Stat.* 40, 2293–2326. doi: 10.1214/12-AOS1037

Liu, H., Lafferty, J., and Wasserman, L. (2009). The nonparanormal: semiparametric estimation of high dimensional undirected graphs. *J. Mach. Learn. Res.* 10, 2295–2328. doi: 10.1145/1577069.1755863

Liu, H., Roeder, K., and Wasserman, L. (2010). "Stability approach to regularization selection (stars) for high dimensional graphical models," in *Proceedings of the Twenty-Third Annual Conference on Neural Information Processing Systems (NIPS)* (Vancouver, BC), 1432–1440.

Mandal, S., Van Treuren, W., White, R. A., Eggesbø, M., Knight, R., and Peddada, S. D. (2015). Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microb. Ecol. Health Dis.* 26, 1–7. doi: 10.3402/mehd.v26.27663

McDonald, D., Hyde, E., Debelius, J. W., Morton, J. T., Gonzalez, A., Ackermann, G., et al. (2018). American gut: an open platform for citizen science microbiome research. *mSystems* 3:e00031-18. doi: 10.1128/mSystems.00031-18

McMurdie, P. J., and Holmes, S. (2014). Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput. Biol.* 10:e1003531. doi: 10.1371/journal.pcbi.1003531

Meinshausen, N., and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Ann. Stat.* 34, 1436–1462. doi: 10.1214/009053606000000281

Müller, C. L., Bonneau, R., and Kurtz, Z. (2016). Generalized stability approach for regularized graphical models. *ArXiv e-prints*. Available online at: https://arxiv.org/abs/1605.07072 (accessed May 27, 2019).

Props, R., Kerckhof, F.-M., Rubbens, P., De Vrieze, J., Hernandez Sanabria, E., Waegeman, W., et al. (2017). Absolute quantification of microbial taxon abundances. *ISME J.* 11, 584–587. doi: 10.1038/ismej.2016.117

Quinn, T. P., Richardson, M. F., Lovell, D., and Crowley, T. M. (2017). Propr: an R-package for identifying proportionally abundant features using compositional data analysis. *Sci. Rep.* 7, 1–9. doi: 10.1038/s41598-017-16520-0

Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S., and Vert, J. P. (2018). A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat. Commun.* 9:284. doi: 10.1038/s41467-017-02554-5

Ruiz, V. E., Battaglia, T., Kurtz, Z. D., Bijnens, L., Ou, A., Engstrand, I., et al. (2017). A single early-in-life macrolide course has lasting effects on murine microbial network topology and immunity. *Nat. Commun.* 8:518. doi: 10.1038/s41467-017-00531-6

Satinsky, B. M., Gifford, S. M., Crump, B. C., and Moran, M. A. (2013). Use of internal standards for quantitative metatranscriptome and metagenome analysis. *Methods Enzymol.* 531, 237–250. doi: 10.1016/B978-0-12-407863-5.00012-5

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., et al. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* 75, 7537–7541. doi: 10.1128/AEM.01541-09

Sczyrba, A., Hofmann, P., Belmann, P., Koslicki, D., Janssen, S., Dröge, J., et al. (2017). Critical assessment of metagenome interpretation–a benchmark of metagenomics software. *Nat. Methods* 14, 1063–1071. doi: 10.1038/nmeth.4458

Sedlar, K., Kupkova, K., and Provaznik, I. (2017). Bioinformatics strategies for taxonomy independent binning and visualization of sequences in shotgun metagenomics. *Comput. Struct. Biotechnol. J.* 15, 48–55. doi: 10.1016/j.csbj.2016.11.005

Soetaert, K. (2009). *rootSolve: Nonlinear Root Finding, Equilibrium and Steady-State Analysis of Ordinary Differential Equations*. R package version 1.6.

Stämmler, F., Gläsner, J., Hiergeist, A., Holler, E., Weber, D., Oefner, P. J., et al. (2016). Adjusting microbiome profiles for differences in microbial load by spike-in bacteria. *Microbiome* 4:28. doi: 10.1186/s40168-016-0175-0

Sunagawa, S., Coelho, L. P., Chaffron, S., Kultima, J. R., Labadie, K., Salazar, G., et al. (2015). Structure and function of the global ocean microbiome. *Science* 348:1261359. doi: 10.1126/science.1261359

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B* 58, 267–288.

Tipton, L., Müller, C. L., Kurtz, Z. D., Huang, L., Kleerup, E., Morris, A., et al. (2018). Fungi stabilize connectivity in the lung and skin microbial ecosystems. *Microbiome* 6:12. doi: 10.1186/s40168-017-0393-0

Tkacz, A., Hortala, M., and Poole, P. S. (2018). Absolute quantitation of microbiota abundance in environmental samples. *Microbiome* 6, 1–13. doi: 10.1186/s40168-018-0491-7

Vandeputte, D., Kathagen, G., D'hoe, K., Vieira-Silva, S., Valles-Colomer, M., Sabino, J., et al. (2017). Quantitative microbiome profiling links gut community variation to microbial load. *Nature* 551, 507–511. doi: 10.1038/nature24460

Woese, C., and Fox, G. (1977). Phylogenetic structure of the prokaryotic domain. *PNAS* 74, 5088–5090.

Yee, T. W. (2010). The VGAM package for categorical data analysis. *J. Stat. Softw.* 32, 1–34. doi: 10.18637/jss.v032.i10

Yoon, G., Carroll, R. J., and Gaynanova, I. (2018). Sparse semiparametric canonical correlation analysis for data of mixed types. *ArXiv e-prints*. Available online at: https://arxiv.org/abs/1807.05274 (accessed May 27, 2019).

Yoon, G., and Gaynanova, I. (2018). *mixedCCA: Sparse CCA for High-Dimensional Mixed Data*. R package version 1.0.1.

Zhao, T., Liu, H., Roeder, K., Lafferty, J., and Wasserman, L. (2012). The huge package for high-dimensional undirected graph estimation in R. *J. Mach. Learn. Res.* 13, 1059–1062. Available online at: http://www.jmlr.org/papers/v13/zhao12a.html

Conflict of Interest Statement: 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.