HARMONIES: A Hybrid Approach for Microbiome Networks Inference via Exploiting Sparsity — Source link

Shuang Jiang, Guanghua Xiao, Andrew Y. Koh, Bo Yao ...+2 more authors

Institutions: Southern Methodist University, University of Texas Southwestern Medical Center, University of Texas at Dallas

Related papers:

- HARMONIES: A Hybrid Approach for Microbiome Networks Inference via Exploiting Sparsity.

- Bayesian modeling of microbiome data for differential abundance analysis

- MHSNMF: multi-view hessian regularization based symmetric nonnegative matrix factorization for microbiome data analysis.

- MB-GAN: Microbiome Simulation via Generative Adversarial Network

- Bayesian approaches for flexible and informative clustering of microbiome data

Share this paper:

View more about this paper here: https://typeset.io/papers/harmonies-a-hybrid-approach-for-microbiome-networks-27w07cb323

Check for updates

HARMONIES: A Hybrid Approach for Microbiome Networks Inference via Exploiting Sparsity

Shuang Jiang [1,2], Guanghua Xiao [2], Andrew Y. Koh [3], Yingfei Chen [4], Bo Yao [2], Qiwei Li [5*] and Xiaowei Zhan [2*]

[1] Department of Statistical Science, Southern Methodist University, Dallas, TX, United States, [2] Quantitative Biomedical Research Center, Department of Population and Data Sciences, University of Texas Southwestern Medical Center, Dallas, TX, United States, [3] Departments of Pediatrics, Departments of Microbiology, University of Texas Southwestern Medical Center, Dallas, TX, United States, [4] Lyda Hill Department of Bioinformatics, Bioinformatics High Performance Computing, University of Texas Southwestern Medical Center, Dallas, TX, United States, [5] Department of Mathematical Sciences, The University of Texas at Dallas, Richardson, TX, United States

The human microbiome is a collection of microorganisms. They form complex communities and collectively affect host health. Recently, the advances in next-generation sequencing technology enable the high-throughput profiling of the human microbiome. This calls for a statistical model to construct microbial networks from the microbiome sequencing count data. As microbiome count data are high-dimensional and suffer from uneven sampling depth, over-dispersion, and zero-inflation, these characteristics can bias the network estimation and require specialized analytical tools. Here we propose a general framework, HARMONIES, Hybrid Approach foR MicrobiOme Network Inferences via Exploiting Sparsity, to infer a sparse microbiome network. HARMONIES first utilizes a zero-inflated negative binomial (ZINB) distribution to model the skewness and excess zeros in the microbiome data, as well as incorporates a stochastic process prior for sample-wise normalization. This approach infers a sparse and stable network by imposing non-trivial regularizations based on the Gaussian graphical model. In comprehensive simulation studies, HARMONIES outperformed four other commonly used methods. When using published microbiome data from a colorectal cancer study, it discovered a novel community with disease-enriched bacteria. In summary, HARMONIES is a novel and useful statistical framework for microbiome network inference, and it is available at https://github.com/shuangj00/HARMONIES.

Keywords: Bayesian statistics, microbiome network, Gaussian graphical model, Dirichlet process prior, hierarchical model

1. INTRODUCTION

Microbiota form complex community structures and collectively affect human health. Studying their relationship as a network can provide key insights into their biological mechanisms. The exponentially growing large datasets made available by next-generation sequencing (NGS) technology (Metzker, 2010), such as 16S rRNA gene and metagenomic profiling, motivate the development of statistic tools to quantitatively study the microbial organisms. While the number of discovered microbial taxa continues to increase, our knowledge of their interactive relationships is severely lacking. Understanding the structural organization of the human microbiome plays a vital role in revealing how the microbial taxa are collaborating or competing with each other under different physiologic conditions.

In sequencing-based microbial association studies, the enormous amount of NGS data can be summarized in a sample-by-taxon count table where each entry is a proxy to the underlying true abundance. However, there is no simple relationship between the true abundances and the observed counts. Additionally, microbiome sequencing data usually have an inflated amount of zeros, uneven sequencing depths across samples, and over-dispersion. Initial attempts at constructing microbial association networks with this type of data (Ban et al., 2015; Lo and Marculescu, 2017), first transformed the microbiome sequencing counts into their compositional formula. Specifically, a count was normalized to its proportion in the respective sample. Then, each sample was transformed by a choice of log-ratio transformations to remove the unit-sum constraint of the compositional data. While this type of normalization is simple to implement and preserves the original ordering of the counts in a sample, it fails to capture the sample to sample variation and it overlooks the excess zeros in the microbiome data. Note that these zeros can be attributed to biological or technical reasons: either certain taxa are not present among samples, or they are not sequenced due to insufficient sequencing depths. As the existing logarithmic transformation neglects the difference between these two types of zeros, it can lead to a biased estimation of the network structure. Thus, we propose a model-based normalization strategy for microbiome count data. Our normalization method simultaneously accounts for uneven sequencing depth, zero-inflation, over-dispersion, as well as the two types of zeros. Then we use the normalized abundances to estimate microbial abundance networks.

There are two major categories of statistical methods that are often used to infer microbial abundance networks. The first type is based on a taxa abundance covariance structure. For example, Faust and Raes (2016) and Weiss et al. (2016) used pairwise Pearson correlations to represent edge weights. This simple inference could be problematic since two variables (i.e., taxa) may be connected in the network due to their confounding variables (Gevers et al., 2014). The other type aims to estimate taxa abundance partial correlations, removing confounding effects. Kurtz et al. (2015) proposed a statistical model for inferring microbial ecological network, which is based on estimating the precision matrix (via exploiting sparsity) of a Gaussian multivariate model and relies on graphical lasso (Glasso) (Friedman et al., 2008). However, their data normalization step needs to be improved to account for unique characteristics observed in microbiome count data.

In this paper, we propose a general framework, HARMONIES (Hybrid Approach foR MicrobiOme Network Inferences via Exploiting Sparsity), to infer the microbiome networks. It consists of two major steps: (1) normalization of the microbiome count data by fitting a zero-inflated negative binomial (ZINB) model with the Dirichlet process prior (DPP), (2) application of Glasso to ensure sparsity and using a stability-based approach to select the tuning parameter in Glasso. The estimated network contains the information of both the degree and the direction of associations between taxa, which facilitates the biological interpretation. We demonstrated that HARMONIES could outperform other state-of-the-art tools on extensive simulated and synthetic data. Further, we used HARMONIES to uncover unique associations between disease-specific genera from microbiome profiling data generated from a colorectal cancer study. Based on these results, HARMONIES will be a valuable statistical model to understand the complex microbial associations in microbiome studies. The R package HARMONIES is freely available at https://github.com/shuangj00/HARMONIES.

2. METHODS

2.1. Microbiome Count Data Normalization

Let Y denote the n-by-p taxonomic count matrix obtained from either the 16S rRNA or the metagenomic shotgun sequencing (MSS) technology. Each entry $y_{ij}, i = 1, \ldots, n, j = 1, \ldots, p$ is a non-negative integer, indicating the total reads related to taxon j observed in sample i. It is recommended that all chosen taxa should be at the same taxonomic level (e.g., OTU for 16S rRNA or species for MSS) since that mixing different taxonomic levels in the proposed model could lead to improper biological interpretation. As the real microbiome data are characterized by zero-inflation and over-dispersion, we model y_{ij} through a zero-inflated negative binomial (ZINB) model as

$$y_{ij} \sim \pi_i \mathrm{I}(y_{ij} = 0) + (1 - \pi_i)\mathrm{NB}(\lambda_{ij}, \phi_j). \tag{1}$$

The first component in the Equation (1) models whether zeros come from a degenerate distribution with a point mass at zero. It can be interpreted as the "extra" zeros due to insufficient sequencing effort. We can assume there exists a true underlying abundance for the taxon in its sample, but we fail to observe it with the mixture probability π_i representing the proportion of "extra" zeros in sample i. The second component, $\mathrm{NB}(\lambda_{ij}, \phi_j)$, models the "true" zeros and all the nonzero observed counts. i.e., counts generated from a negative binomial (NB) distribution with the expectation of λ_{ij} and dispersion $1/\phi_j$. Here, "true" zero refers to a taxon that is truly absent in the corresponding sample. The variance of the random variable from NB distribution, under the current parameterization, equals to $\lambda_{ij} + \lambda_{ij}^2/\phi_j$. Smaller values of ϕ_j can lead to over-dispersion.

To avoid explicitly fixing the value of π_i's and ϕ_j's, we use a Bayesian hierarchical model for parameter inference. First, we rewrite the model (1) by introducing a binary indicator variable $\eta_{ij} \sim \mathrm{Bernoulli}(\pi_i)$, such that $y_{ij} = 0$ if $\eta_{ij} = 1$, and $y_{ij} \sim \mathrm{NB}(\lambda_{ij}, \phi_j)$ if $\eta_{ij} = 0$. Then, we formulate a beta-Bernoulli prior of η_{ij} by assuming $\pi_i \sim \mathrm{Beta}(a_\pi, b_\pi)$, and we let $a_\pi = b_\pi = 1$ to obtain a non-informative prior on η_{ij}. We specify independent Gamma prior $\mathrm{Ga}(a_\phi, b_\phi)$ for each dispersion parameter ϕ_j. Letting $a_\phi = b_\phi = 0.001$ results in a weakly informative gamma prior.

The mean parameter of the NB distribution, λ_{ij}, contains the key information of the true underlying abundance of the corresponding count. As λ_{ij} is affected by the varying sequencing effort across samples, we use a multiplicative characterization of the NB mean to justify the latent heterogeneity in microbiome sequencing data. Specifically, we assume $\lambda_{ij} = s_i \alpha_{ij}$. Here, s_i is the sample-specific size factor that captures the variation

in sequencing depth across samples, and α_{ij} is the normalized abundance of taxon j in sample i.

In parameter estimation, one needs to ensure identifiability between s_i and α_{ij}. For example, s_i can be the reciprocal of the total number of reads in sample i. The resulted α_{ij} is often called relative abundance, which represents the proportion of taxon j in sample i. In this setting, the relative abundances of all the taxa in one sample always sum up to 1. Similarly, other methods have been proposed with different constraints for normalizing the sequencing data (Anders and Huber, 2010; Bullard et al., 2010; Robinson and Oshlack, 2010; Paulson et al., 2013). Some normalization methods can perform better than the others in the downstream analysis (e.g., the differential abundance analysis) in certain settings. From a Bayesian perspective, fixing the values of s_i's imposes a strongly informative prior in model inference. Hence, all these methods could bias the estimations of other model parameters and degrade the performance of downstream analyses. We thus propose a regularizing prior with a stochastic constraint for estimating s_i's. Our method can simultaneously infer the size factor and other model parameters. In particular, we adopt the following mixture model for s_i,

$$\log s_i \sim \sum_{m=1}^{M} \psi_m \left[t_m \mathrm{N}(\nu_m, \sigma_s^2) + (1 - t_m)\mathrm{N}\left(-\frac{t_m \nu_m}{1 - t_m}, \sigma_s^2\right) \right],$$

(2)

where ψ_m is the weight for outer mixtures of the mth component. The inner mixture of the mth component consists of two Gaussian distributions with t_m and $1 - t_m$ as weights, respectively. It is straightforward to see that the inner mixture has a mean of zero and thus ensuring the stochastic constraint of $\mathrm{E}(\log s_i) = 0$. For the outer mixtures, M is an arbitrarily large positive integer. Letting $M \to \infty$ and defining the weight ψ_m by the stick-breaking procedure (i.e., $\psi_1 = V_1, \psi_m = V_m \prod_{u=1}^{m-1}(1 - V_u), m = 1, 2, \ldots$) makes model (2) a special case of Dirichlet process mixture models. This class of Bayesian nonparametric infinite mixtures is widely used in quantifying the model uncertainty and allowing for flexibility in parameter estimation (Kyung et al., 2011; Taddy and Kottas, 2012). In particular, this Dirichlet process prior (DPP) has been used to account for sample heterogeneity since it is able to capture multi-modality and skewness in a distribution (Li et al., 2017; Lee and Sison-Mangus, 2018). In practice, we set M to be a large positive integer, and adopt the following hyper-prior distributions for the parameters in (2) such that $\nu_m \sim \mathrm{N}(0, \tau_\nu), t_m \sim \mathrm{Beta}(a_t, b_t),$ and $V_m \sim \mathrm{Beta}(a_m, b_m)$. We further set $\sigma_s^2 = 1$ to complete the parameter specification in the DPP prior.

In our model, the normalized abundance matrix $\boldsymbol{A} = \{\alpha_{ij}\}$ represents the true underlying abundance of the original count matrix. We further assume $\log \alpha_{ij} \sim \mathrm{N}(\mu_j, \sigma_j^2)$. This variance-stabilizing transformation on each α_{ij} not only reduces the skewness of the normalized abundance, but converts the non-negative α_{ij} to a real number. We apply the following conjugate setting to specify the priors for μ_j and $\sigma_j^2, j = 1, \ldots, p$. We let $\mu_j \sim \mathrm{N}(0, h_0\sigma_0^2)$ and $\sigma_j^2 \sim \text{inverse-gamma}(a_0, b_0)$. After integrating out μ_j and σ_j^2, the prior of the normalized

abundances of taxon j follows a non-standardized Student's t-distribution, i.e.,

$$p(\boldsymbol{\alpha}_{\cdot j}) = (nh_0 + 1)^{-\frac{1}{2}} \frac{\Gamma\left(a_0 + \frac{n}{2}\right)}{\Gamma(a_0)}$$

$$\frac{b_0^{a_0}}{\left\{ b_0 + \frac{1}{2}\left[\sum_{i=1}^{n} \log \alpha_{ij}^2 - \frac{(\sum_{i=1}^{n} \log \alpha_{ij})^2}{n + \frac{1}{h_0}} \right] \right\}^{a_0 + \frac{n}{2}}}.$$

(3)

As for the fixed parameters $a_0, b_0, h_0,$ and σ_0^2, we follow Li et al. (2019) and set $a_0 = 2, b_0 = 1$ to obtain a weakly informative prior for σ_j^2. We fix $\sigma_0^2 = 1$ and let $h_0 = 10$ such that the normal prior on μ_j is fairly flat. We adopt the following prior specification for the rest model parameters. First, we assume an noninformative prior for each π_i by letting $a_\pi = b_\pi = 1$. Next, we specify $a_\phi = b_\phi = 0.001$ in the Gamma prior distribution for all ϕ_j's. Then, we apply the following prior setting for the DPP: $M = n/2, \sigma_s = 1, \tau_\nu = 1, a_t = b_t = 1,$ and $a_m = b_m = 1$.

The logarithmic scale of $\boldsymbol{A}$, denoted as $\boldsymbol{Z} = \log \boldsymbol{A}$, represents the normalized microbiome abundances on the log scale. We use Markov chain Monte Carlo (MCMC) algorithm for model parameter estimation (see details in the **Supplementary Material**), and calculate the posterior mean of $\boldsymbol{Z}$ to fit the Gaussian graphical model in the next step. Since the observed zero counts may not always represent the absence of taxa in the samples, we treat these zeros differently in the matrix $\boldsymbol{Z}$. We categorize the two types of zeros ("extra" and "true" zeros) based on the estimated η_{ij} for each observed $y_{ij} = 0$ in the data. In particular, suppose that we observe L zeros in total. We calculate the marginal posterior probability of being 1 for each $\eta_l, l = 1, \ldots, L$ as $p_l = \sum_{b=1}^{B} \mathrm{I}(\eta_l = 1)/B$, where $\mathrm{I}(\cdot)$ is the indicator function, and B is the number of MCMC iteration after burn-in. This marginal posterior probability p_l represents the proportion of MCMC iterations in which the lth 0 is essentially a missing value rather than the lowest count in the corresponding sample. Then, the observed zeros can be dichotomized by thresholding the L probabilities. The zeros with p_l greater than the threshold are considered as "true" zeros in the data, whereas the rest are imputed by the corresponding posterior mean of $\log \boldsymbol{\alpha}_{\cdot j}$. We used the method proposed by Newton et al. (2004) to determine the threshold that controls the Bayesian false discovery rate (FDR) to be smaller than c_η. Specifically, we first specify a small number c_η, which is analog to the significance level in the frequentist setting. Then we compute the threshold following Equation (4), which guarantees the imputed zeros have a Bayesian FDR to be smaller than c_η,

$$\text{Bayesian FDR} = \frac{\sum_{l=1}^{L}(1 - p_l)\,\mathrm{I}(1 - p_l < c_\eta)}{\sum_{l=1}^{L}(1 - p_l < c_\eta)}.$$

(4)

In practice, a choice of $c_\eta = 0.01$ guarantees that the Bayesian FDR to be at most 0.01. We set $c_\eta = 0.05$ for the simulation study and $c_\eta = 0.01$ for the real data analysis.

2.2. Graphical Model for Inferring Taxa-Taxa Association

Based on the normalized microbial abundances, we estimate their partial correlation matrix in order to construct the microbiome network under the Gaussian graphical model (GGM) framework. An undirected graph $G = (V, E)$ is used to illustrate the associations among vertices $V = \{1, \ldots, p\}$, representing the p microbial taxa. $E = \{e_{mk}\}$ is the collection of (undirected) edges, which is equivalently represented via a p-by-p adjacency matrix with $e_{mk} = 1$ or 0 according to whether vertices m and k are directly connected in G or not. GGM assumes that the joint distribution of p vertices is multivariate Gaussian $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, yielding the following relationship between the dependency structure and the network: a zero entry in the precision matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ indicates the corresponding vertices are conditional independent, and there is no edge between them in the graph G. Hence, a GGM can be defined in terms of the pairwise conditional independence. If $X \sim N(\boldsymbol{\mu}, \boldsymbol{\Omega})$, then

$$\omega_{mk} = 0 \Leftrightarrow X_m \perp X_k | X_{V \setminus \{m,k\}} \Leftrightarrow \rho_{mk} = 0,$$

where $\rho_{mk} = -\omega_{mk}/\sqrt{\omega_{mm}\omega_{kk}}$ is the partial correlation between vertices m and k, representing the degree and direction of association between two vertices, conditional on the rest variables. Consequently, learning the network is equivalent to estimating the precision matrix $\boldsymbol{\Omega}$. For real microbiome data, we set the taxa (on the same taxonomic level) as vertices. Hence, a zero partial correlation in the precision matrix can be interpreted as no association between the corresponding pair of taxa, while a nonzero partial correlation can be interpreted as cooperative or competing associations between that taxa pair.

In biological applications, we often require a sparse and stable estimation of the precision matrix $\boldsymbol{\Omega}$. The sparsity can be achieved by imposing l_1-penalized log-likelihood,

$$\hat{\boldsymbol{\Omega}} = \underset{\boldsymbol{\Omega} \succ \mathbf{0}}{\operatorname{argmin}} \log \det \boldsymbol{\Omega} - \operatorname{trace}(\boldsymbol{S}\boldsymbol{\Omega}) - \lambda \|\boldsymbol{\Omega}\|_1, \qquad (5)$$

where $\boldsymbol{S}$ is the sample covariance matrix. The coordinate descent algorithm can iteratively solve p. The estimated precision matrix is sparsistent (i.e., all the parameters that are zeros would be estimated as zero with probability one) (Lam and Fan, 2009), as Glasso theoretically guarantees a consistent recovery of the sparse graph for the p vertices. When $p >> n$, the computational efficiency is often satisfactory, and thus Glasso is widely used in studying large-scale biological networks (Menéndez et al., 2010; Oh and Deasy, 2014; Zhao and Duan, 2019). We employ a stability-based approach to select the tuning parameter in the Glasso, which is named Stability Approach to Regularization Selection (StARS) (Liu et al., 2010). This method is an improved algorithm for estimating the tuning parameter λ in (5). The StARS selects the optimal sparsity parameter according to the graph reproducibility under the subsampling of the original data. In general, for each λ along the sparsity parameter path, we first obtain random subsamples from the original data. Then we estimate the graph for each subsample using the Glasso. Next, for each sparsity parameter, we calculate the overall edge selection

instability from all the graphs constructed by the subsamples. Finally, the optimal sparsity parameter λ^* is chosen such that it corresponds to the smallest amount of regularization and still results in a graph instability to be lower than the pre-specified tolerance level. Liu et al. (2010) showed that StARS could provide the "sparsistent" network estimation that includes all the true associations with probability one. Further, the StARS has been widely used in biological network studies (Kurtz et al., 2015; Tipton et al., 2018; Zhao and Duan, 2019). Due to its excellent performance, here we adopt the StARS to select the tuning parameter for Glasso. In summary, we use the normalized abundances (on the log scale) as inputs, calculate the sparse estimation of the precision matrix using the Glasso, and use the StARS method to select λ in problem (5) to obtain the estimated graph that represents the microbiome network.

2.3. Simulation Scenarios

We compare the performance of the HARMONIES and several widely used methods for inferring microbiome networks. These methods include SPIEC-EASI (Kurtz et al., 2015), CClasso (Fang et al., 2015), and correlation-based network estimation used in Faust and Raes (2016) and Weiss et al. (2016). While the proposed model and SPIEC-EASI infer the network structure from sparse precision matrices, CClasso, and the correlation-based method utilize sparse correlation matrices to represent the network. We generated both simulated and synthetic datasets that mimic the real microbiome sequencing count data. We use $\boldsymbol{Y}_{n \times p}$ to denote the generated count matrix. For a comprehensive comparison, we varied the sample size and the number of taxa as $n \in \{60, 100, 200, 500\}$, and the number of taxa $p \in \{40, 60\}$.

2.3.1. Generating Simulated Data

We generated the simulated datasets from a Dirichlet-multinomial (DM) model using the following steps: (1) to generate the binary adjacency matrix; (2) to simulate the precision matrix and the corresponding covariance matrix; (3) to generate n multivariate Gaussian variables based on the covariance matrix to represent the true $n \times p$ underlying taxonomic abundances, denoted as $\boldsymbol{D}$; (4) to simulate the count table $\boldsymbol{Y}_{n \times p}$ from a DM model, with its parameters being $\exp(\boldsymbol{D})$; (5) to mimic the zero-inflation in real microbiome data by randomly setting part of entries in the count table to zeros. Note that the data generative scheme is different from the model assumption, which is given in Equation (1). The detailed generative models are described below.

We began with simulating a p-by-p adjacency matrix for the p taxa in the network. Here, the adjacency matrix was generated according to an Erdős–Rényi (ER) model. An ER model ER(p, ρ) generates each edge in a graph G with probability ρ independently from every other edge. Therefore, all graphs with p nodes and M edges have an equal probability of $\rho^M (1-\rho)^{\binom{p}{2}-M}$. All the edges in graph G correspond to the 1's in the resulted binary adjacency matrix. Next, we simulated the precision matrix $\boldsymbol{\Omega}$ following Peng et al. (2009). We started by setting all the diagonal elements of $\boldsymbol{\Omega}$ to be 1. Then, for the rest elements that correspond to the 1s in the adjacency matrix, we sampled their values independently from a uniform distribution

Unif($[-0.1, 0] \cup [0, 0.1]$). To ensure positive definiteness of the precision matrix, we followed Peng et al. (2009) by dividing each off-diagonal element by 1.5 times the sum of the absolute value of all the elements in its row. Finally, we averaged the rescaled precision matrix with its transpose and set the diagonal elements to 1. This process ensured the preceding matrix was positive definite and symmetric. The corresponding covariance matrix was set as $\boldsymbol{\Sigma} = \boldsymbol{\Omega}^{-1}$.

Next, we simulated n multivariate Gaussian variables from MN($\boldsymbol{\mu}, \boldsymbol{\Sigma}$) to represent the true underlying abundances $\boldsymbol{D}$. To obtain a count matrix that fully mimics the microbiome sequencing data, we generated counts from a DM model with parameter exp($\boldsymbol{D}$). Specifically, we first sampled the underlying fractional abundances for the ith sample from a Dirichlet distribution. The ith underlying fractional abundance was then denoted as $\boldsymbol{\psi}_i \sim$ Dirichlet(exp($\boldsymbol{D}_{i\cdot}$)). Next, the counts in the ith sample were generated from Multinomial($N_i, \boldsymbol{\psi}_i$). Finally, we randomly selected $\pi_0\%$ out of $n \times p$ counts and set them to zeros to mimic the zero-inflation observed in the real microbiome data. In general, the generative process had different assumptions from the proposed method. Under the appropriate choice of parameters, the simulated count data was zero-inflated, overdispersed, and the total reads varied largely between samples. In practice, we let $\rho = 0.1$ in the ER model. The mean parameter $\boldsymbol{\mu}$ of the underlying multivariate Gaussian variable was randomly sampled from a uniform distribution Unif[0, 10]. The number of total counts across samples $N_i, i = 1, \ldots, n$ was sampled from a discrete uniform distribution with range [50, 000, 100, 000]. Under each combination of n, p, and π_0, we generated 50 replicated datasets by repeating the process above.

2.3.2. Generating Synthetic Data
We generated synthetic data following the Normal-to-Anything (NorTA) approach proposed in Kurtz et al. (2015). NorTA was designed to generate multivariate random variables with an arbitrary marginal distribution from a pre-specified correlation structure (Cario and Nelson, 1997). Given the observations of p taxa from a real microbiome dataset, the NorTA generates the synthetic data with n samples as follows: (1) to calculate the p-by-p covariance matrix $\boldsymbol{\Sigma}_0$ from the input real dataset; (2) to generate an n-by-p matrix, denoted by $\boldsymbol{Z}_0$, from a multivariate Gaussian distribution with a mean of $\boldsymbol{0}_{1 \times p}$ and the covariance matrix of $\boldsymbol{\Sigma}_0$; (3) to use standard normal cumulative distribution function to scale values in each column of $\boldsymbol{Z}_0$ within [0, 1]; (4) to apply the quantile function of a ZINB distribution to generate count data from those scaled values in each column of $\boldsymbol{Z}_0$. In practice, we used R package `SPIEC-EASI` to implement the above data generative scheme, where the real data were from those healthy control subjects in our case study presented in section 3.2. Under each combination of n and p, we generated 50 replicated datasets.

2.4. Model Performance
2.4.1. Alternative Methods in Network Learning
We considered the four commonly used network learning methods. The first two methods, SPIEC-EASI-Glasso and SPIEC-EASI-mb, use the transformed microbiome abundances which

are different from the normalized abundances estimated by HARMONIES. Both infer the microbial network by estimating a sparse precision matrix. The former method (SPIEC-EASI-Glasso) measures the dependency among microbiota by their partial correlation coefficients, and the latter method (SPIEC-EASI-mb) uses the "neighborhood selection" introduced by Meinshausen and Bühlmann (2006) to construct the network. The third method, denoted as Pearson-corr, calculates Pearson's correlation coefficients between all pairs of taxa. In its estimated network, the edges correspond to large correlation coefficients. To avoid arbitrarily thresholding the correlation coefficients, the fourth method, CClasso (Fang et al., 2015), directly infers a sparse correlation matrix with l_1 regularization. However, as discussed in section 1, representing the dependency structure by the correlation matrix may lead to the detection of spurious associations.

2.4.2. Evaluation Criteria
We quantified the model performances on the simulated data by computing their receiver operating characteristic (ROC) curves and area under the ROC curve (AUC). For the HARMONIES or SPIEC-EASI, the network inference was based on the precision matrix. Hence, under each tuning parameter of Glasso, we calculated the number of edges being true positive (TP) by directly comparing the estimated precision matrix against the true one. More specifically, we considered an edge between taxon m and taxon k to be true positive if $\omega_{mk} \neq 0$, $\hat{\omega}_{mk} \neq 0$, and $\hat{\omega}_{mk}$ shared the same sign with ω_{mk}. We calculated the number of true negative (TN), false positive (FP), and false negative (FN) in a similar manner. Therefore, each tuning parameter defined a point on a ROC curve. As for the correlation-based methods, we started with ranking the absolute values in the estimated correlation matrices, denoted as $\hat{\boldsymbol{C}}$. Next, we used each value as a threshold and set all the entries in $\hat{\boldsymbol{C}}$ having their absolute values smaller than the current threshold to be zeros. Then, the number of TP, TN, FP, or FN was obtained by comparing the sparse $\hat{\boldsymbol{C}}$ against the true partial correlation matrix. Therefore, each unique absolute value in the original estimated correlation matrix defined a point on the ROC curve.

We further used the Matthew's correlation coefficient (MCC) to evaluate results from the simulated data. The MCC is defined as

$$\frac{(\text{TP} \times \text{TN} - \text{FP} \times \text{FN})}{\sqrt{(\text{TP+FP})(\text{TP+FN})(\text{TN+FP})(\text{TN+FN})}}.$$

Here, the MCC was particularly suitable for evaluating network models. As the number of conditionally independent taxa pairs was assumed to be much greater than the number of dependent pairs in a sparse network, MCC was preferable to quantify the performances under such an imbalanced situation. Note that MCC ranges from $[-1, 1]$, with a value close to 1 suggesting a better performance. Since each value of MCC was calculated using a given set of TP, TN, FP, and FN, we adopted the optimal choice of tuning parameter for the HARMONIES or SPIEC-EASI (with either Glasso or MB for network inference), given by StARS. As for the correlation-based methods, CClasso outputted a sparse correlation matrix. We used the result to calculate TP,

TN, FP, and FN directly. For Pearson-corr, we set the threshold such that the resulted number of nonzero entries in the sparse correlation matrix was the same as the number of non-zero entries in the true sparse partial correlation matrix. In fact, this choice could favor the performance of Pearson-corr for larger sample size, as shown in section 3.1.

To assess model performances on the synthetic datasets, we followed Kurtz et al. (2015) to use a metric called area under the precision-recall (AUPR) curves, in addition to AUC. Briefly speaking, the AUPR and AUC were calculated as follows: (1) to rank all possible edges according to their confidence values; (2) to generate the precision-recall curve and the ROC curve by comparing edge inclusions against the true sparse precision matrix; (3) to calculate the area under the precision-recall curve or the ROC curve. Note that the confidence values were chosen as the edge stabilities under the optimal choice of the tuning parameter selected by StARS for HARMONIES, SPIEC-EASI-Glasso, and SPIEC-EASI-mb, while for CClasso and Pearson-corr, p-values were used.

3. RESULTS

3.1. Simulation Results

Figures 1, **2** compare the AUCs and MCCs on the simulated data under various scenarios, including varying sample sizes ($n = 60, 100, 200,$ or 500), total numbers of taxa ($p = 40$ or 60), extra percentages of zeros added ($\pi_0 = 10,$ or 20%). In each subfigure, the HARMONIES outperformed the alternative methods in terms of both AUC and MCC, and it maintained this advantage even with the number of sample size greatly increases. Further, a smaller sample size, a larger proportion of extra zeros added ($\pi_0 = 20\%$), as well as a larger number of taxa in the network ($p = 60$), would hamper the performance of all the methods, as we expected. Two modes of SPIEC-EASI, SPIEC-EASI-Glasso, and SPIEC-EASI-mb, showed very similar performances under all the scenarios, with SPIEC-EASI-Glasso having only a marginal advantage over the other. Further, we observed that the Pearson-corr method yielded higher AUCs even than the precision matrix based methods, especially when there was a lager proportion of extra zeros or larger number of taxa in the network. This result suggested that the Pearson-corr could capture the overall rank of the signal strength in the actual network. However, under a fixed cut-off value that gave a sparse correlation network, the MCCs from the Pearson-corr were always smaller than the precision matrix based methods. Note that the cut-off value we specified for Pearson's correlation method indeed favored its performance. In general, the alternative methods considered here were able to reflect the overall rank of the signal strength by showing reasonable AUCs. However, they failed to give an accurate estimation of the network under a fixed cut-off value.

Figure 3 demonstrates that our model outperformed all others on the synthetic datasets. The performances in terms of AUC under different scenarios are summarized in **Figures 3A,B**, while those in terms of AUPR are displayed in (**Figures 3C,D**). As we can see, either increasing the sample size n or decreasing the number of features p would improve the performance of all methods and lead to greater disparity between

partial and pairwise correlation-based methods. In general, our HARMONIES maintained the best in all simulation and evaluation settings except for one case, where the SPIEC-EASI-mb only showed a marginal advantage (see $n = 60$ in **Figure 3C**). Interestingly, our observation confirmed a finding mentioned by Kurtz et al. (2015), that is, the SPIEC-EASI-mb was slightly better than SPIEC-EASI-Glasso in terms of AUPR under the optimal choice of the tuning parameter. As for the two correlation-based methods, we found that Pearson-corr outperformed CClasso in most of the scenarios.
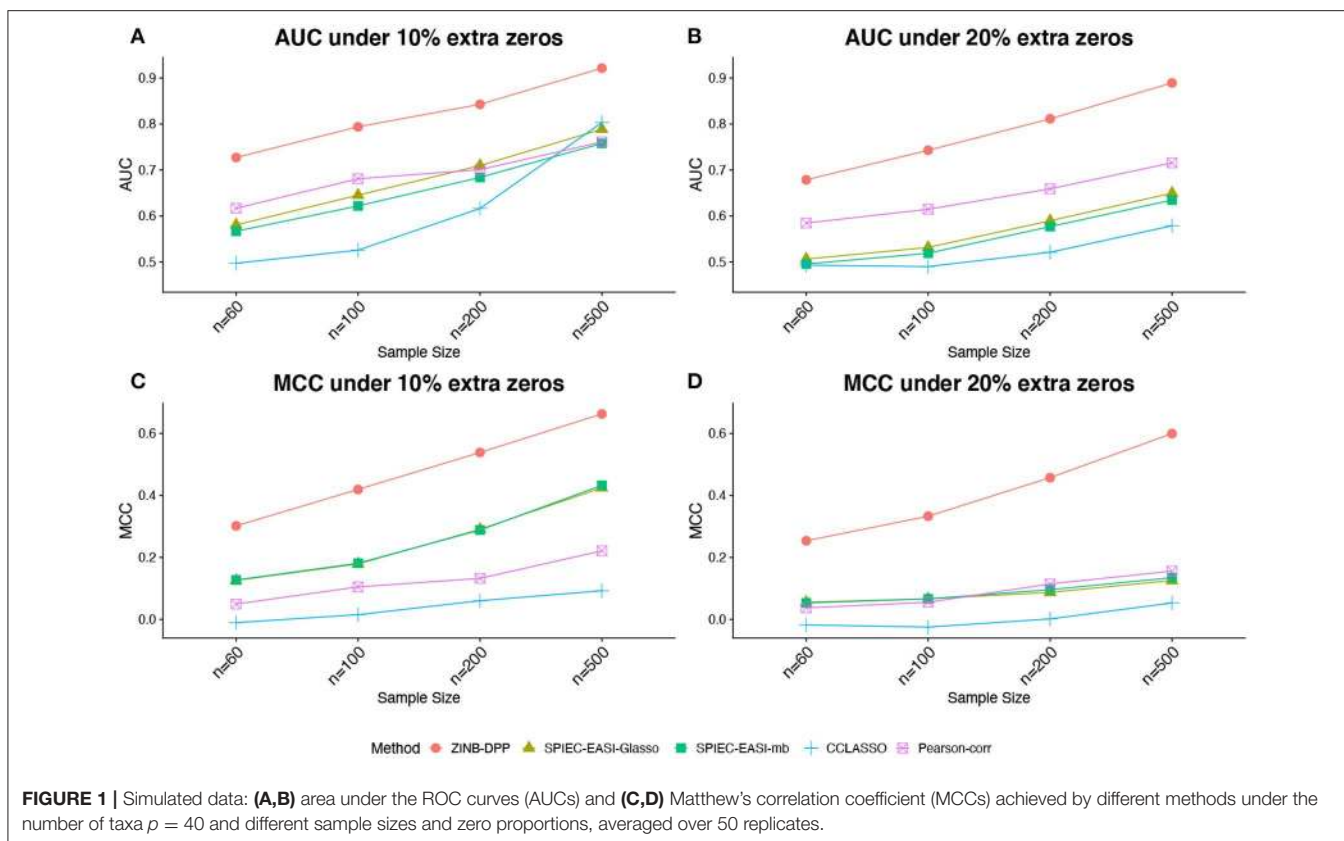
3.2. Analysis of Microbiome Data From Colorectal Cancer Patients

Colorectal cancer (CRC) is the third most common cancer diagnosed in both men and women in the United States (Arnold et al., 2017). Increasing evidence from recent studies highlights a vital role for the intestinal microbiota in malignant gastrointestinal diseases including CRC (Louis et al., 2014; Sears and Garrett, 2014; Drewes et al., 2016). In particular, studies have reported that dysbiosis of specific microbiota is directly associated with CRC (Marchesi et al., 2011; Kostic et al., 2013; Flynn et al., 2016). The current microbiome research interests have gone beyond the discovery of disease-related microbiota, with a growing number of studies investigating the interactive associations among the microbial taxa. Using the proposed model, we interrogated the microbiome profiling data of a CRC study to determine the microbiome network structures.

We analyzed the gut microbiome dataset of a CRC study published by Feng et al. (2015). We extracted from the original cohort[1] the 43 CRC patients and the 58 healthy controls. The original sequencing data at the genus level were quantified using curatedMetagenomicData (Pasolli et al., 2017). We had $p = 187$ genera for both the 43 CRC patients and the 58 healthy controls. We implemented the HARMONIES as follows. For the CRC group, we first applied the ZINB model to obtain the normalized abundance matrix A, utilizing the specifications detailed in section 2.1. We then took the logarithmic transformation of the normalized abundance and imputed the missing values. Before implementing the proposed method, we filtered out the low abundant genera with zeros occurring more than half samples. Removing low abundant taxa is a common step in microbiome research (see e.g., Qin et al., 2014; Zeller et al., 2014; Kostic et al., 2015; Kurtz et al., 2015; Wadsworth et al., 2017; Yilmaz et al., 2019). The rationale being that these "zero-abundant" taxa may be less important in a network, which was also confirmed by our simulation study. This filtering process left 51 and 36 genera in the CRC and control group, respectively. The result from using a more relaxed filtering threshold is available in the supplement, where we kept the genera that had at least 10% nonzero observations across the samples.

Figures 4A,B display the estimated networks for the CRC and the control group, respectively. Each node, corresponding to a genus, was named after its phylum level. All the genera
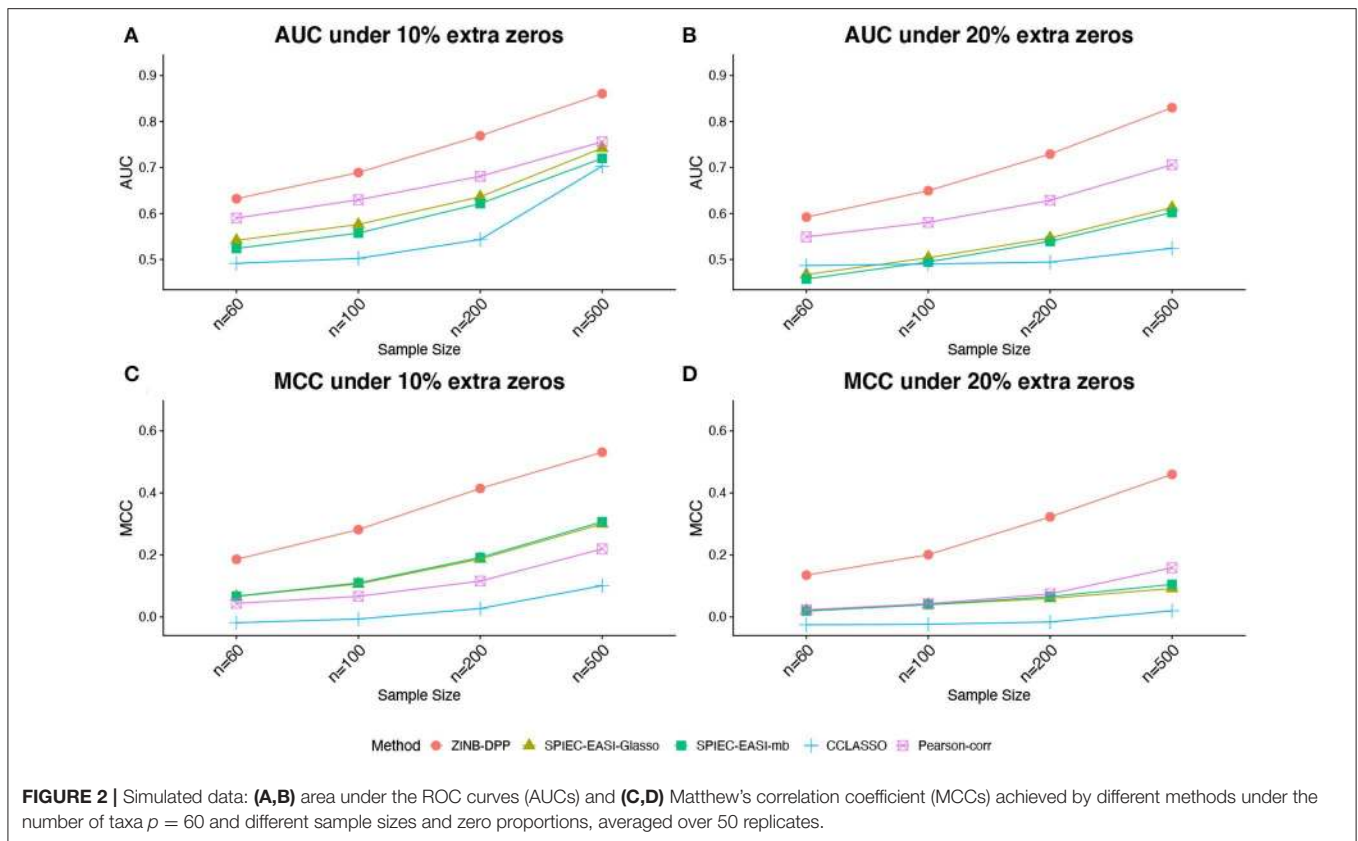
---

[1]The original metagenomic shotgun sequencing data from the fecal samples are available in the European Bioinformatics Institute Database (accession number ERP008729).

FIGURE 1 | Simulated data: **(A,B)** area under the ROC curves (AUCs) and **(C,D)** Matthew's correlation coefficient (MCCs) achieved by different methods under the number of taxa $p = 40$ and different sample sizes and zero proportions, averaged over 50 replicates.

shown in **Figure 4** belong to six phyla in total. By using their phylum name to further categorize these distinct genera, we aimed at exploring interesting patterns among them at a higher taxonomic level. **Figure S1** displays the same network using the actual genus name on each node. The node sizes are proportional to its normalized abundances in the logarithmic scale. The green or red edge indicates a positive or a negative partial correlation, respectively. And the width of an edge is proportional to the absolute value of the partial correlation coefficient. To make a clear comparison, we intentionally kept the nodes and their positions to be consistent between the two subfigures. In either of the two groups, we included a node in the current plot if there exists an edge between it with any nodes in at least one group. In general, the two groups share several edges with the same direction of partial correlations, but the majority of edges are unique within each group.

Network estimation of the CRC group demonstrated several microbial communities. For example, three genera: *Fusobacterium*, *Peptostreptococcus*, and *Parvimonas* consisted of a unique subnetwork as highlighted in **Figure 4A**. These three genera were isolated in the control group's network, as shown in **Figure 4B**. Interestingly, specific species under these three genera have been reported as enriched taxa in CRC and related to worse clinical outcome (Mima et al., 2016; Yu et al., 2017; Long et al., 2019). A previous CRC study by Kostic et al. (2013) supported the causal role of species *Fusobacterium nucleatum*
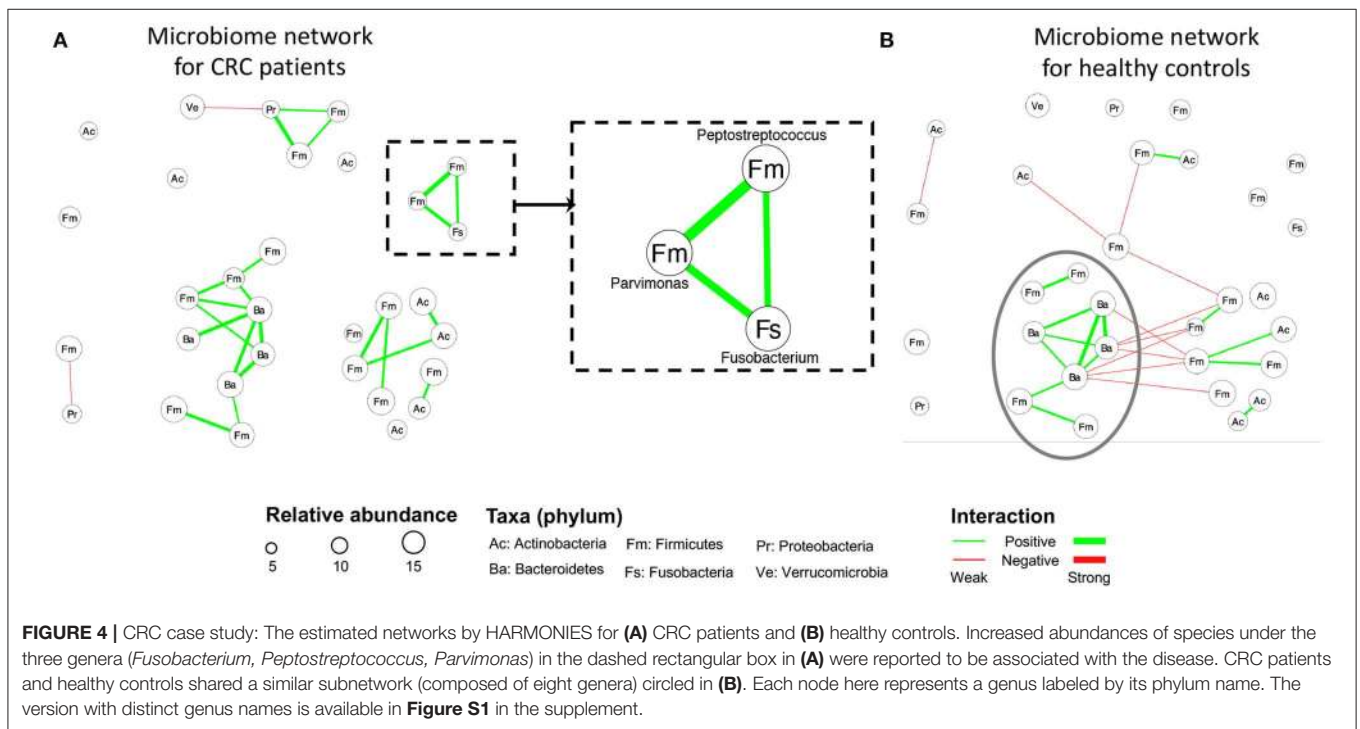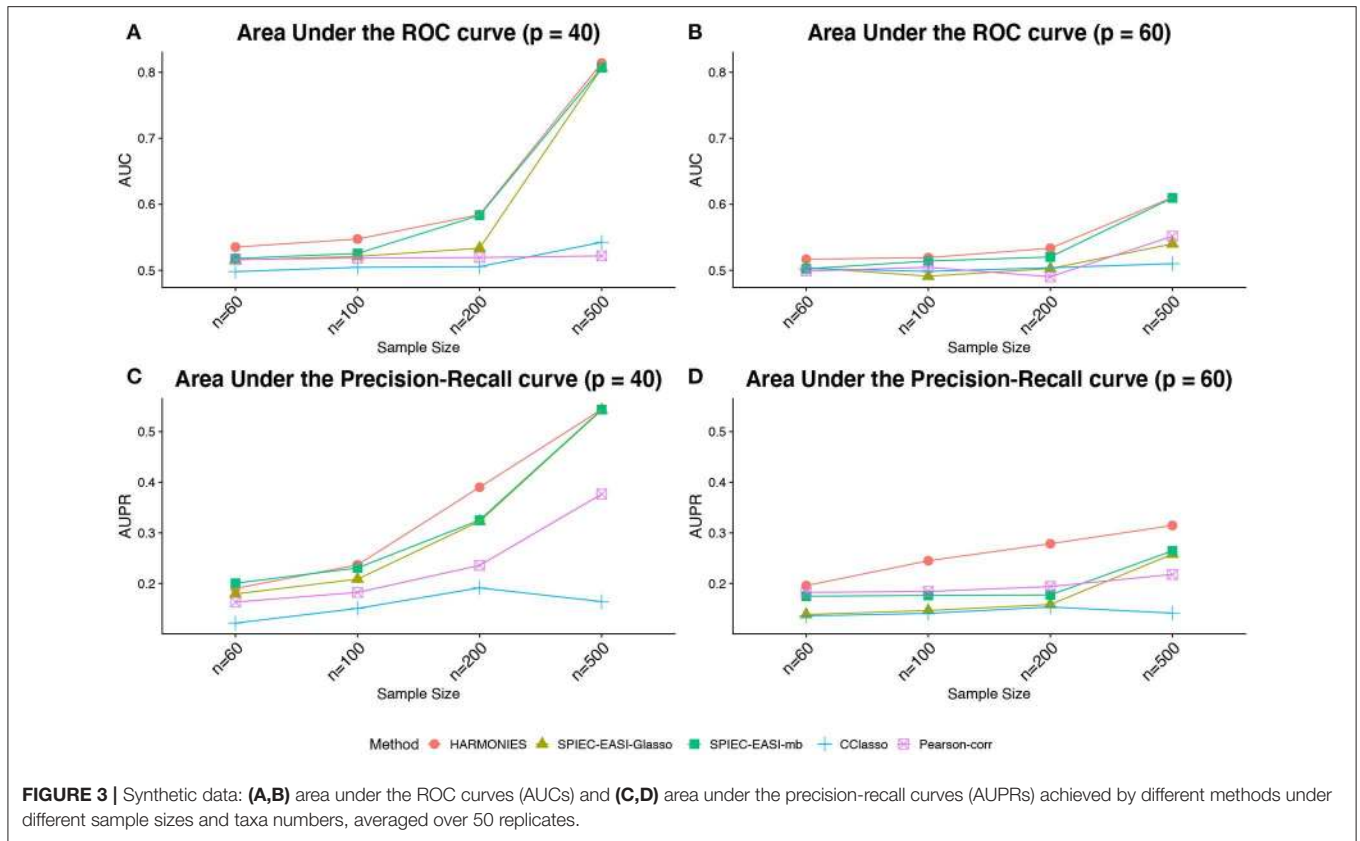
by showing that *F. nucleatum* promotes tumor progression by increasing both tumor multiplicity and tumor-infiltrating myeloid cells in a preclinical CRC model. Further, a recent study (Long et al., 2019) demonstrated that *Peptostreptococcus anaerobius* accelerated colorectal tumorigenesis in a murine CRC model. This study suggested that *P. anaerobius* directly interacted with colonic epithelial cells and also promoted CRC by modifying the tumor immune microenvironment. While the causal role of the species *Parvimonas micra* has not been biologically validated, multiple clinical studies reported an elevated level of *P. micra* in CRC patients (Purcell et al., 2017; Yu et al., 2017; Dai et al., 2018). Of interest, *Parvimonas* were closely associated with animal-based diets, which have previously been shown to be significantly associated with increased risk for CRC (Chan et al., 2011). The previous studies only investigated those CRC-related taxa individually, whereas a novel finding by HARMONIES analysis suggested that all the three genera were co-aggregating in CRC patients as their pairwise associations are all positive. Interestingly, in a prior study direct positive associations between *Fusobacterium* and *Peptostreptococcus*, as well as *Peptostreptococcus* and *Parvimonas*, were identified (Hibberd et al., 2017). However, there was no direct association between *Fusobacterium* and *Parvimonas*. Similarly, another study (Drewes et al., 2017) found a direct co-occurrence pattern between two species: *F. nucleatum* and *P. micra*. Using HARMONIES, we could jointly identify the relationship among each pair of the three genera,

FIGURE 2 | Simulated data: **(A,B)** area under the ROC curves (AUCs) and **(C,D)** Matthew's correlation coefficient (MCCs) achieved by different methods under the number of taxa $p = 60$ and different sample sizes and zero proportions, averaged over 50 replicates.

conditional on all other genera. This novel subcommunity of three CRC-enriched genera formulated a recurring module and may function as a cooperative group in CRC patients. A closer investigation of their co-occurrence pattern could potentially elucidate both their contributions to CRC and the basic biology under their relationships. Two additional novel taxa interactions were identified by HARMONIES analysis: *Streptococcus* and *Veillonella*, and *Streptococcus* and *Haemophilus*. In fact, previous CRC studies showed enrichment of these three genera or their species in CRC patients (see e.g., Geng et al., 2014; Ugai et al., 2014; Kumar et al., 2017; Koliarakis et al., 2019), but had not detected these novel interactions. In conclusion, HARMONIES may reveal how multiple CRC-related taxa could potentially promote disease progression together.

Having shared edges between the two networks suggests that the HARMONIES is robust to the edge selection. We observed that the shared edges tended to appear for those more abundant genera. For example, we circled eight genera in **Figure 4B**, and the HARMONIES suggested multiple positive partial correlations among them. For these eight genera, we observed six shared edges between the CRC and healthy control networks. Notice that all the shared edges were consistent in the association directions, and they also corresponded to the relatively stronger association in both networks (wider in the edge width). We found these shared edges tend to connect those more abundant genera (node with larger size). Indeed,

the eight genera considered here belong to phyla *Bacteroidetes* and *Firmicutes*, both were in the top three most abundant phyla for CRC patients and healthy controls reported by Gao et al. (2015) and Mori et al. (2018). Therefore, it was more likely that the highly abundant genera shared similar association patterns between the two groups, and the HARMONIES demonstrated its robustness by preserving these relatively stronger partial correlations among these genera. On the other hand, the network of the control group contained more negative partial correlations as shown in **Figure 4B**. Furthermore, the two edges linked to *Streptococcus* were different from the CRC group. Here, *Streptococcus* had a negative association with *Subdoligranulum* and a positive association with *Rothia*. There has been no evidence suggesting these two genera are CRC-related. Hence a further investigation is merited. Additionally, the CRC group has another distinct small subnetwork formed by the four genera, two from *Firmicutes*, one from *Proteobacteria*, and one from *Verrucomicrobia*. These group-specific associations were never reported. Lastly, we observed several interesting patterns between the two groups when summarizing the genera to their phylum levels. Genera in Firmicutes (labeled as "Fm" in **Figure 4**) showed more positive associations in the case group than in the control group, whereas negative associations between Firmicutes and Bacteroidetes (labeled as "Ba" in **Figure 4**) were more common in the control group. Again, these novel patterns still need further biological validations to elucidate their functions.

FIGURE 3 | Synthetic data: **(A,B)** area under the ROC curves (AUCs) and **(C,D)** area under the precision-recall curves (AUPRs) achieved by different methods under different sample sizes and taxa numbers, averaged over 50 replicates.

FIGURE 4 | CRC case study: The estimated networks by HARMONIES for **(A)** CRC patients and **(B)** healthy controls. Increased abundances of species under the three genera (*Fusobacterium, Peptostreptococcus, Parvimonas*) in the dashed rectangular box in **(A)** were reported to be associated with the disease. CRC patients and healthy controls shared a similar subnetwork (composed of eight genera) circled in **(B)**. Each node here represents a genus labeled by its phylum name. The version with distinct genus names is available in **Figure S1** in the supplement.

4. DISCUSSION

With the advent of next-generation sequencing technology, microbiome research now has the opportunity to explore microbial community structure and to characterize the microbial ecological association for different populations or physiology conditions (Kurtz et al., 2015). In this paper, we introduce HARMONIES as a statistical framework to infer sparse networks using microbiome sequencing data. It models the original count data by a zero-inflated negative binomial distribution to capture the large amount for zeros and over-dispersion, and it further implements Dirichlet process priors to account for sample heterogeneity. In contrast, current methods for microbiome network analyses rely on the compositional data, which could cause information loss due to ignoring the unique characteristics of the microbiome sequencing count data. Following the data normalization step, the HARMONIES explores the direct connections in the network by estimating the partial correlations. The results from the simulation study have demonstrated the advantage of the HARMONIES over alternative approaches under various conditions. When applied to an actual microbiome dataset, the HARMONIES suggests all the nodes to be taxa at the same taxonomic level, such as species, genus, family, etc. This ensures proper biologically interpretations of those detected associations. When applied to a real CRC study, the HARMONIES revealed an intriguing community among three CRC-enriched genera. Further, shared patterns between the CRC and the control networks suggest a common community pattern of disease neutral genera. Additional studies validating the biological relevance of these microbial associations, however, will need to be conducted.

Both the simulated and synthetic data showed that a larger sample size improved the performance of all the network learning methods. In practice, many disease-related microbiome studies, especially those studying rare diseases, always have small sample sizes. This limitation directly affects the estimation of the normalized matrix A from the ZINB model. Notice that for a taxon j, a small sample size could result in a large variance in the posterior distribution of $\log \alpha_{\cdot j}$. However, many disease studies include reference groups where the measurements on the same taxonomic features are available. The additional information from the subjects in the reference group can potentially help improve the posterior inference of the normalized abundances. We generalized the proposed ZINB model to handle two groups, with the goal of borrowing information between groups in estimating the normalized abundances. These detailed model formula and implementation were included in the supplement (see **Supplementary Material**: section 2).

Our hybrid approach for microbiome network inference can be extended. One future direction is to incorporate the differential network analysis into the existing framework. It jointly considers the association strengths between each pair of taxa from different groups, and it compares the estimated individual networks to capture the significantly different connectivities. Our current method can infer the normalized abundances for two groups, and we provided the details steps in the supplement. However, an integrated differential network can be expected to better study the differential microbial community structure and link the communities to human health status.

SOFEWARE AVAILABILITY STATEMENT

We developed the R package HARMONIES that is freely available at https://github.com/shuangj00/HARMONIES. We also released a webtool that allows users to upload micorbiome datasets and run HARMONIES for microbiome network analysis. The online webtool is available at http://lce.biohpc.swmed.edu/harmonies.

DATA AVAILABILITY STATEMENT

The datasets generated for the simulation study can be found in the author's Github page—https://github.com/shuangj00/HARMONIES. The original metagenomic shotgun sequencing data analyzed in the real data study are available in the European Bioinformatics Institute Database (accession number ERP008729).

AUTHOR CONTRIBUTIONS

SJ performed the experiments. GX and AK provided resources and helpful discussions. SJ, QL, and XZ designed the experiment, performed data analysis, wrote the software, and the manuscript. SJ, YC, BY, and XZ developed the website for online implementation of HARMONIES.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00445/full#supplementary-material

REFERENCES

Anders, S, and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol*. 11:R106. doi: 10.1186/gb-2010-11-10-r106

Arnold, M., Sierra, M. S., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F. (2017). Global patterns and trends in colorectal cancer incidence and mortality. *Gut* 66, 683–691. doi: 10.1136/gutjnl-2015-310912

Ban, Y., An, L., and Jiang, H. (2015). Investigating microbial co-occurrence patterns based on metagenomic compositional

data. *Bioinformatics* 31, 3322–3329. doi: 10.1093/bioinformatics/btv364

Bullard, J. H., Purdom, E., Hansen, K. D., and Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* 11:94. doi: 10.1186/1471-2105-11-94

Cario, M. C., and Nelson, B. L. (1997). *Modeling and Generating Random Vectors With Arbitrary Marginal Distributions and Correlation Matrix*. Technical Report, Citeseer.

Chan, D. S., Lau, R., Aune, D., Vieira, R., Greenwood, D. C., Kampman, E., et al. (2011). Red and processed meat and colorectal cancer incidence: meta-analysis of prospective studies. *PLoS ONE* 6:e20456. doi: 10.1371/journal.pone.0020456

Dai, Z., Coker, O. O., Nakatsu, G., Wu, W. K., Zhao, L., Chen, Z., et al. (2018). Multi-cohort analysis of colorectal cancer metagenome identified altered bacteria across populations and universal bacterial markers. *Microbiome* 6:70. doi: 10.1186/s40168-018-0451-2

Drewes, J. L., Housseau, F., and Sears, C. L. (2016). Sporadic colorectal cancer: microbial contributors to disease prevention, development and therapy. *Br. J. Cancer* 115:273. doi: 10.1038/bjc.2016.189

Drewes, J. L., White, J. R., Dejea, C. M., Fathi, P., Iyadorai, T., Vadivelu, J., et al. (2017). High-resolution bacterial 16S rRNA gene profile meta-analysis and biofilm status reveal common colorectal cancer consortia. *NPJ Biofilms Microb.* 3, 1–12. doi: 10.1038/s41522-017-0040-3

Fang, H., Huang, C., Zhao, H., and Deng, M. (2015). CClasso: correlation inference for compositional data through lasso. *Bioinformatics* 31, 3172–3180. doi: 10.1093/bioinformatics/btv349

Faust, K., and Raes, J. (2016). CoNet app: inference of biological association networks using Cytoscape. *F1000Res.* 5:1519. doi: 10.12688/f1000research.9050.2

Feng, Q., Liang, S., Jia, H., Stadlmayr, A., Tang, L., Lan, Z., et al. (2015). Gut microbiome development along the colorectal adenoma-carcinoma sequence. *Nat. Commun.* 6:6528. doi: 10.1038/ncomms7528

Flynn, K. J., Baxter, N. T., and Schloss, P. D. (2016). Metabolic and community synergy of oral bacteria in colorectal cancer. *Msphere* 1: e00102-16. doi: 10.1128/mSphere.00102-16

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

Gao, Z., Guo, B., Gao, R., Zhu, Q., and Qin, H. (2015). Microbiota disbiosis is associated with colorectal cancer. *Front. Microbiol.* 6:20. doi: 10.3389/fmicb.2015.00020

Geng, J., Song, Q., Tang, X., Liang, X., Fan, H., Peng, H., et al. (2014). Co-occurrence of driver and passenger bacteria in human colorectal cancer. *Gut Pathogens* 6:26. doi: 10.1186/1757-4749-6-26

Gevers, D., Kugathasan, S., Denson, L. A., Vázquez-Baeza, Y., Van Treuren, W., Ren, B., et al. (2014). The treatment-naive microbiome in new-onset Crohn's disease. *Cell Host Microbe* 15, 382–392. doi: 10.1016/j.chom.2014.02.005

Hibberd, A. A., Lyra, A., Ouwehand, A. C., Rolny, P., Lindegren, H., Cedgåård, L., et al. (2017). Intestinal microbiota is altered in patients with colon cancer and modified by probiotic intervention. *BMJ Open Gastroenterol.* 4:e000145. doi: 10.1136/bmjgast-2017-000145

Koliarakis, I., Messaritakis, I., Nikolouzakis, T. K., Hamilos, G., Souglakos, J., and Tsiaoussis, J. (2019). Oral bacteria and intestinal dysbiosis in colorectal cancer. *Int. J. Mol. Sci.* 20:4146. doi: 10.3390/ijms20174146

Kostic, A. D., Chun, E., Robertson, L., Glickman, J. N., Gallini, C. A., Michaud, M., et al. (2013). *Fusobacterium nucleatum* potentiates intestinal tumorigenesis and modulates the tumor-immune microenvironment. *Cell Host Microbe* 14, 207–215. doi: 10.1016/j.chom.2013.07.007

Kostic, A. D., Gevers, D., Siljander, H., Vatanen, T., Hyötyläinen, T., Hämäläinen, A.-M., et al. (2015). The dynamics of the human infant gut microbiome in development and in progression toward type 1 diabetes. *Cell Host Microbe* 17, 260–273. doi: 10.1016/j.chom.2015.01.001

Kumar, R., Herold, J. L., Schady, D., Davis, J., Kopetz, S., Martinez-Moczygemba, M., et al. (2017). *Streptococcus gallolyticus* subsp. gallolyticus promotes colorectal tumor development. *PLoS Pathogens* 13:e1006440. doi: 10.1371/journal.ppat.1006440

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

Kyung, M., Gill, J., and Casella, G. (2011). Sampling schemes for generalized linear Dirichlet process random effects models. *Stat. Methods Appl.* 20, 259–290. doi: 10.1007/s10260-011-0168-x

Lam, C., and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Stat.* 37:4254. doi: 10.1214/09-AOS720

Lee, J., and Sison-Mangus, M. (2018). A Bayesian semiparametric regression model for joint analysis of microbiome data. *Front. Microbiol.* 9:522. doi: 10.3389/fmicb.2018.00522

Li, Q., Cassese, A., Guindani, M., and Vannucci, M. (2019). Bayesian negative binomial mixture regression models for the analysis of sequence count and methylation data. *Biometrics* 75, 183–192. doi: 10.1111/biom.12962

Li, Q., Guindani, M., Reich, B. J., Bondell, H. D., and Vannucci, M. (2017). A Bayesian mixture model for clustering and selection of feature occurrence rates under mean constraints. *Stat. Anal. Data Min.* 10, 393–409. doi: 10.1002/sam.11350

Liu, H., Roeder, K., and Wasserman, L. (2010). "Stability approach to regularization selection (StARS) for high dimensional graphical models," in *Advances in Neural Information Processing Systems* (Vancouver, CA), 1432–1440.

Lo, C., and Marculescu, R. (2017). MPLasso: Inferring microbial association networks using prior microbial knowledge. *PLoS Comput. Biol.* 13:e1005915. doi: 10.1371/journal.pcbi.1005915

Long, X., Wong, C. C., Tong, L., Chu, E. S., Szeto, C. H., Go, M. Y., et al. (2019). Peptostreptococcus anaerobius promotes colorectal carcinogenesis and modulates tumour immunity. *Nat. Microbiol.* 4, 2319–2330. doi: 10.1038/s41564-019-0541-3

Louis, P., Hold, G. L., and Flint, H. J. (2014). The gut microbiota, bacterial metabolites and colorectal cancer. *Nat. Rev. Microbiol.* 12, 661–672. doi: 10.1038/nrmicro3344

Marchesi, J. R., Dutilh, B. E., Hall, N., Peters, W. H., Roelofs, R., Boleij, A., et al. (2011). Towards the human colorectal cancer microbiome. *PLoS ONE* 6:e20447. doi: 10.1371/journal.pone.0020447

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

Menéndez, P., Kourmpetis, Y. A., ter Braak, C. J., and van Eeuwijk, F. A. (2010). Gene regulatory networks from multifactorial perturbations using Graphical Lasso: application to the DREAM4 challenge. *PLoS ONE* 5:e14147. doi: 10.1371/journal.pone.0014147

Metzker, M. L. (2010). Sequencing technologies–the next generation. *Nat. Rev. Genet.* 11:31. doi: 10.1038/nrg2626

Mima, K., Nishihara, R., Qian, Z. R., Cao, Y., Sukawa, Y., Nowak, J. A., et al. (2016). *Fusobacterium nucleatum* in colorectal carcinoma tissue and patient prognosis. *Gut* 65, 1973–1980. doi: 10.1136/gutjnl-2015-310101

Mori, G., Rampelli, S., Orena, B. S., Rengucci, C., De Maio, G., Barbieri, G., et al. (2018). Shifts of faecal microbiota during sporadic colorectal carcinogenesis. *Sci. Rep.* 8:10329. doi: 10.1038/s41598-018-28671-9

Newton, M. A., Noueiry, A., Sarkar, D., and Ahlquist, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* 5, 155–176. doi: 10.1093/biostatistics/5.2.155

Oh, J. H., and Deasy, J. O. (2014). Inference of radio-responsive gene regulatory networks using the graphical lasso algorithm. *BMC Bioinformatics* 15:S5. doi: 10.1186/1471-2105-15-S7-S5

Pasolli, E., Schiffer, L., Manghi, P., Renson, A., Obenchain, V., Truong, D. T., et al. (2017). Accessible, curated metagenomic data through experimenthub. *Nat. Methods* 14:1023. doi: 10.1038/nmeth.4468

Paulson, J. N., Stine, O. C., Bravo, H. C., and Pop, M. (2013). Differential abundance analysis for microbial marker-gene surveys. *Nat. Methods* 10:1200. doi: 10.1038/nmeth.2658

Peng, J., Wang, P., Zhou, N., and Zhu, J. (2009). Partial correlation estimation by joint sparse regression models. *J. Am. Stat. Assoc.* 104, 735–746. doi: 10.1198/jasa.2009.0126

Purcell, R. V., Visnovska, M., Biggs, P. J., Schmeier, S., and Frizelle, F. A. (2017). Distinct gut microbiome patterns associate with consensus molecular

subtypes of colorectal cancer. *Sci. Rep.* 7:11590. doi: 10.1038/s41598-017-11237-6

Qin, N., Yang, F., Li, A., Prifti, E., Chen, Y., Shao, L., et al. (2014). Alterations of the human gut microbiome in liver cirrhosis. *Nature* 513, 59–64. doi: 10.1038/nature13568

Robinson, M. D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* 11:R25. doi: 10.1186/gb-2010-11-3-r25

Sears, C. L., and Garrett, W. S. (2014). Microbes, microbiota, and colon cancer. *Cell Host Microbe* 15, 317–328. doi: 10.1016/j.chom.2014.02.007

Taddy, M. A., and Kottas, A. (2012). Mixture modeling for marked Poisson processes. *Bayesian Anal.* 7, 335–362. doi: 10.1214/12-BA711

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

Ugai, T., Norizuki, M., Mikawa, T., Ohji, G., and Yaegashi, M. (2014). Necrotizing fasciitis caused by haemophilus influenzae type b in a patient with rectal cancer treated with combined bevacizumab and chemotherapy: a case report. *BMC Infect. Dis.* 14:198. doi: 10.1186/1471-2334-14-198

Wadsworth, W. D., Argiento, R., Guindani, M., Galloway-Pena, J., Shelburne, S. A., and Vannucci, M. (2017). An integrative bayesian dirichlet-multinomial regression model for the analysis of taxonomic abundances in microbiome data. *BMC Bioinformatics* 18:94. doi: 10.1186/s12859-017-1516-0

Weiss, S., Van Treuren, W., Lozupone, C., Faust, K., Friedman, J., Deng, Y., et al. (2016). Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *ISME J.* 10:1669. doi: 10.1038/ismej.2015.235

Yilmaz, B., Juillerat, P., Øyås, O., Ramon, C., Bravo, F. D., Franc, Y., et al. (2019). Microbial network disturbances in relapsing refractory Crohn's disease. *Nat. Med.* 25, 323–336. doi: 10.1038/s41591-018-0308-z

Yu, J., Feng, Q., Wong, S. H., Zhang, D., Liang, Q., Qin, Y., et al. (2017). Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. *Gut* 66, 70–78. doi: 10.1136/gutjnl-2015-309800

Zeller, G., Tap, J., Voigt, A. Y., Sunagawa, S., Kultima, J. R., Costea, P. I., et al. (2014). Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol. Syst. Biol.* 10:766. doi: 10.15252/msb.20145645

Zhao, H., and Duan, Z.-H. (2019). Cancer genetic network inference using Gaussian graphical models. *Bioinform. Biol. Insights* 13:1177932219839402. doi: 10.1177/1177932219839402