

BAYESIAN MODELING OF INTERACTION BETWEEN FEATURES IN SPARSE MULTIVARIATE COUNT DATA WITH APPLICATION TO MICROBIOME STUDY

BY SHUANGJIE ZHANG^{1,*}, YUNING SHEN² IRENE A. CHEN² AND JUHEE LEE¹

¹*Department of Statistics, University of California Santa Cruz, *szhan209@ucsc.edu; juheelee@soe.ucsc.edu*

²*Department of Chemical and Biomolecular Engineering, University of California Los Angeles, yshen@chem.ucsbg.edu; ireneachchen@ucla.edu*

Many statistical methods have been developed for the analysis of microbial community profiles, but due to the complexity of typical microbiome measurements, inference of interactions between microbial features remains challenging. We develop a Bayesian zero-inflated rounded log-normal kernel method to model interaction between microbial features in a community using multivariate count data in the presence of covariates and excess zeros. The model carefully constructs the interaction structure by imposing joint sparsity on the covariance matrix of the kernel and obtains a reliable estimate of the structure with a small sample size. The model also includes zero inflation to account for excess zeros observed in data and infers differential abundance of microbial features associated with covariates through log-linear regression. We provide simulation studies and real data analysis examples to demonstrate the developed model. Comparison of the model to a simpler model and popular alternatives in simulation studies shows that in addition to an added and important insight on the feature interaction, it yields superior parameter estimates and model fit in various settings.

1. Introduction. High-throughput sequencing (HTS) technologies in microbial ecology generate multivariate count data to characterize and analyze microbial communities from a variety of habitats such as human body sites, soil and water. Widely used sequencing methods in microbiome research include 16S ribosomal RNA (rRNA) sequencing and shotgun metagenomic sequencing (Jovel *et al.*, 2016). 16S rRNA gene sequencing utilizes PCR to target and amplify some portions of the bacterial 16S rRNA subunit gene for sequencing. The sequence reads are then clustered based on their similarity into operational taxonomic units (OTUs), which represent bacteria types. Following some initial preprocessing procedures, 16S rRNA sequencing data is summarized into a large count matrix (referred to as an OTU table) for downstream analyses, where the columns represent samples, and the rows contain multivariate count vectors of sequences corresponding to OTUs in the samples. Different from marker gene-based community profiling, shotgun metagenomic sequencing sequences a sample's entire metagenome and offers finer resolution at a higher cost. After some bioinformatic preprocessing, it also produces multivariate count table data that has structure and properties similar to those of an OTU table for downstream analyses. 16S rRNA sequencing datasets are used for illustrations of the statistical method developed in this work, but it can be considered for analysis of the data generated by either sequencing technique. We note that their analysis units are different, and the resulting statistical inferences may have different biological interpretations. In the human gut microbiome data, one of our real data examples in § 4.2, 16S rRNA sequencing data was collected to study how the composition of the gut microbiome is associated with inflammatory bowel disease (IBD) such as Crohn's disease

Keywords and phrases: Covariance Matrix, Differential Abundance, Factor Model, Joint Sparsity, Multivariate Count Data, Rounded Kernel Model, Zero Inflation.

(CD) or ulcerative colitis (UC) (Lloyd-Price *et al.*, 2019). Understanding how the composition of the human gut microbiome is associated with covariates such as disease status and age is important to provide insights on its role in human health and disease. Also, detecting and investigating the structure of microbial interactions is critical to better characterize microbial communities. Accurately accounting for the interactions can further improve the quantification of covariate effects on microbial abundances.

HTS sequencing data in microbiome study presents various challenges for statistical analysis due to high dimensionality and some added complexity. Total OTU counts vary in samples due to experimental artifacts such as the sequencing depth, and raw counts do not reflect actual microbial abundances (called compositionality). Consequently, normalization of OTU counts is needed for meaningful comparison across samples. In addition, the high-dimensional structure with excess zeros and over-dispersion further complicates the analysis of an OTU table and calls for flexible statistical models. While various statistical models have been proposed for microbiome data analysis, most existing methods focus on either inference on the effects of environmental factors (i.e., covariate) on microbial abundances or their absence/presence or inference on associations between microbes. For studying associations with covariates, generalized regression models are popular. For example, Poisson or negative binomial (NB) regression models are one of the common approaches, where covariates are related to expected counts through a log-linear regression framework. Those models include sample size factors for normalization. Zero-inflated (ZI) Poisson or ZI-NB models are also utilized to address excess zeros. Under a ZI model, a count is distributed as a mixture, a component of which is the distribution with a point mass of one at zero. See Li *et al.* (2017), Zhang *et al.* (2017), Jiang *et al.* (2021), Shuler *et al.* (2021) among many others, for examples of using Poisson or NB regression models. Another common regression approach uses multinomial or ZI multinomial models, where a similar log-linear regression framework is used to relate covariates to (unconstrained) occurrence probability vectors, e.g., Xia *et al.* (2013), Wadsworth *et al.* (2017), Ren *et al.* (2017), Tang and Chen (2019) and Grantham *et al.* (2020) among many others. In particular, Grantham *et al.* (2020) proposed a Bayesian multinomial regression model that assumes a mixed effects model for unconstrained occurrence probabilities and uses a latent factor model for the covariance matrix of the prior distribution of the unconstrained probabilities. However, the implication of the covariance among unconstrained probabilities for microbial interactions is not clear due to the fixed total count constraint under the assumed multinomial distribution. Approaches of using a Dirichlet-tree multinomial model were also proposed to exploit the tree structure information via a phylogenetic tree, e.g., Wang and Zhao (2017), Mao, Chen and Ma (2020) and Wang, Mao and Ma (2021). They assume potential associations between microbes that have similar sequences but do not attempt to infer microbial interactions. Alternatively, Paulson *et al.* (2013) assumed a univariate log-normal distribution for individual counts after adding a pseudo count to observed counts and used regression to relate covariates to OTU abundances. For inferences on microbial interactions, correlations between pairs of microbes based on some transformed OTU counts are commonly used as a measure. The task of estimating correlations between microbes is complicated due to the aforementioned challenges. Centered-log-ratio (clr) transformation is usually applied to raw counts prior to analysis for compositionality, and small pseudo-counts are added to avoid numerical issues of excess zeros. To address high dimensionality, an additional structure such as sparsity through ℓ_1 penalty is often imposed on the covariance matrix or precision matrix for reliable inference. For example, SparCC in Friedman and Alm (2012) normalizes raw counts by sample total counts after adding pseudo counts and models log-transformed ratios of the normalized counts to infer correlations between OTUs. CCLasso in Fang *et al.* (2015) models log-transformed counts and provides a least

squares estimate of a correlation matrix with ℓ_1 penalty under some constraint for compositionality of microbiome data. SPIEC-EASI in [Kurtz et al. \(2015\)](#) builds an undirected graphical model for clr transformed data and yields inference on an association network between OTUs through a precision matrix. Sparsity is assumed for the underlying association network. [Schwager et al. \(2017\)](#) uses a Bayesian log-normal graphical model for unconstrained counts. A LASSO prior is used for the precision matrix. Similarly, [Prost, Gazut and Brüls \(2021\)](#) developed a likelihood-based zero-inflated log-normal graphical model (Zi-LN) that appropriately accounts for excess zeros in microbiome data. Graphical LASSO ([Friedman, Hastie and Tibshirani, 2008](#)) is used for estimation of the precision matrix. While existing methods can provide useful insights on microbial communities, methods that jointly infer associations between microbes and their associations with covariates are still lacking. Furthermore, statistical methods that carefully address excess zeros, compositionality and high dimensionality are needed for accurate inference on the associations.

To obtain a better understanding of the underlying biological processes, we develop a Bayesian rounded kernel regression model with zero inflation. The model enables a direct assessment of interrelationships between OTUs and their associations with covariates. The developed method directly models raw counts and simultaneously performs model-based normalization through random sample scale factors for compositionality. Specifically, we use a multivariate log-normal distribution as the kernel and define multivariate count responses $\mathbf{Y} = (Y_1, \dots, Y_J)$ of J OTUs in terms of multivariate log-normal latent variables $\mathbf{Y}^* = (Y_1^*, \dots, Y_J^*)$ using fixed thresholds. We then relate covariates \mathbf{x} to the mean vector $\boldsymbol{\mu}$ of the distribution of \mathbf{Y}^* through regression and use the covariance matrix Σ to characterize interrelationship among OTUs. $\boldsymbol{\mu}$ also includes sample size factors for normalization. For Σ , we assume joint sparsity to reliably learn a high dimensional covariance structure with a small sample size. Sparsity assumption is commonly used in the covariance matrix estimation when $p \gg n$ (e.g., [Cai, Ren and Zhou \(2016\)](#), [Pati et al. \(2014\)](#), [Gao and Zhou \(2015\)](#), [Xie et al. \(2018\)](#)). Specifically, we develop a joint sparse latent factor model for Σ , where we let the number of factors much smaller than the number of OTUs (features), and a majority of OTUs can have factor loadings close to zero, i.e., feature selection. The model greatly reduces the number of parameters to estimate and provides a simple interpretation of the interrelationship structure. The representation of the model with independent latent factors also allows introducing zero inflation in a convenient manner. The model appropriately accounts for excess zeros due to the absence of an OTU or the undersampling of a rare OTU, and Σ provides inferences on the interrelationship structure among OTUs present in a sample. In addition, overdispersion is accommodated through random effects, resulting in further improvement in the inference.

In the remainder of the paper, we describe the model and its applications. § 2 describes the zero-inflated multivariate log-normal kernel model (called “ZI-MLN”), and § 3 has results of simulation studies to evaluate the performance of our method. § 4 has results from the model applied to two real datasets, and § 5 concludes with some discussion of the results and areas of future research.

2. Statistical Model.

2.1. Sampling Distribution and Prior Specification. Consider multivariate count data obtained for J OTUs in a microbiome study. We let $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{iJ})$ denote a J -dimensional random count vector of OTU counts of sample $i = 1, \dots, N$ taken from subject $g_i \in \{1, \dots, M\}$, where $Y_{ij} \in \mathbb{N}^0$ is the count of OTU $j = 1, \dots, J$ in sample i . We let n_m be the number of samples taken from subject m and have $N = \sum_{m=1}^M n_m$. In addition, data may include a set of P covariates, $\mathbf{x}_i = (x_{i1}, \dots, x_{iP})$. Our skin microbiome dataset in

§ 4.1 consists of observed counts of 187 OTUs in 20 samples, one sample from each of 20 subjects. The dataset does not have covariates besides the subject factor. Human gut microbiome data in § 4.2 includes 67 samples collected from multiple biopsy sites of 37 patients. 107 OTUs are included with covariates such as disease phenotype and age for analysis. The model simultaneously infers the interaction structure of OTUs and the differential abundance of OTUs by covariates. It can also be easily simplified if no covariate is available, as we will show later.

We consider a Bayesian rounded multivariate log-normal kernel model for \mathbf{Y}_i in [Canale and Dunson \(2011\)](#). We first introduce continuous latent variables $\mathbf{Y}_i^* = (Y_{i1}^*, \dots, Y_{iJ}^*)$ with $Y_{ij}^* \in \mathbb{R}^+$, $i = 1, \dots, n$ and $j = 1, \dots, J$, and assume

$$(1) \quad \mathbf{Y}_i^* | \boldsymbol{\mu}_i, \Sigma \stackrel{\text{indep}}{\sim} \text{log-N}_J(\boldsymbol{\mu}_i, \Sigma),$$

where parameters $\boldsymbol{\mu}_i = (\mu_{i1}, \dots, \mu_{iJ})' \in \mathbb{R}^J$ and $\Sigma > 0$. In (1), we have the mean $E(Y_{ij}^* | \boldsymbol{\mu}_i, \Sigma) = \exp(\mu_{ij} + \frac{1}{2}\Sigma_{jj})$, the median $Q_{0.5} = \exp(\mu_{ij})$ and covariance $\text{Cov}(Y_{ij}^*, Y_{ij'}^*) = \exp\{\mu_{ij} + \mu_{ij'} + \frac{1}{2}(\Sigma_{jj} + \Sigma_{j'j'})\}\{\exp(\Sigma_{jj'}) - 1\} = E(Y_{ij}^*)E(Y_{ij'}^*)\{\exp(\Sigma_{jj'}) - 1\}$. We next use a threshold function to relate Y_{ij}^* to Y_{ij} by letting $Y_{ij} = y_j$ if $y_j \leq Y_{ij}^* < (y_j + 1)$. The multivariate log-normal density is zero for a vector with negative values, and the kernel defines a valid multivariate distribution for \mathbf{Y} . We further let $\tilde{\mathbf{Y}}_i^* = (\tilde{Y}_{i1}^*, \dots, \tilde{Y}_{iJ}^*)$ with $\tilde{Y}_{ij}^* = \log(Y_{ij}^*) \in \mathbb{R}$ and have

$$(2) \quad \begin{aligned} P(\mathbf{Y}_i = \mathbf{y}_i | \boldsymbol{\mu}_i, \Sigma) &= \int_{A(\mathbf{y}_i)} f_{\mathbf{y}^*}(\mathbf{y}^* | \boldsymbol{\mu}_i, \Sigma) d\mathbf{y}^* \\ &= \int_{\tilde{A}(\mathbf{y}_i)} \phi_J(\tilde{\mathbf{y}}^* | \boldsymbol{\mu}_i, \Sigma) d\tilde{\mathbf{y}}^*, \end{aligned}$$

where $f_{\mathbf{y}^*}$ represents the density function of the J -dimensional log-normal distribution with parameters $\boldsymbol{\mu}_i$ and Σ , and ϕ_J the density function of a J -dimensional normal distribution. The regions of integration are $A(\mathbf{y}_i) = \{\mathbf{y}^* | y_{i1} \leq y_1^* < y_{i1} + 1, \dots, y_{iJ} \leq y_J^* < y_{iJ} + 1\}$ and $\tilde{A}(\mathbf{y}_i) = \{\tilde{\mathbf{y}}^* | \log(y_{i1}) \leq \tilde{y}_1^* < \log(y_{i1} + 1), \dots, \log(y_{iJ}) \leq \tilde{y}_J^* < \log(y_{iJ} + 1)\}$. The properties of the distribution of Y_{ij} 's such as their means and covariances can be easily computed from (2). For example, we find $E(Y_{ij} | \mu_{ij}, \Sigma_{jj}) = \sum_{b=0}^{\infty} bP(Y_{ij} = b | \mu_{ij}, \Sigma_{jj})$ with $P(Y_{ij} = b | \mu_{ij}, \Sigma_{jj}) = \Phi_1(\log(b+1) | \mu_{ij}, \Sigma_{jj}) - \Phi_1(\log(b) | \mu_{ij}, \Sigma_{jj})$, where $\Phi_d(\cdot | a, \mathbf{B})$ is the cdf of the d -variate normal distribution with mean a and (co)variance \mathbf{B} . A large value of μ_{ij} thus implies high abundance of OTU j in sample i . We express μ_i as a function of covariates, sample-size factor and OTU-size factor. The factors account for differences in sample total counts and variability in baseline OTU abundances. We will give a regression model for $\boldsymbol{\mu}_i$ below. We can also compute variances and covariances of the counts. In particular, $\text{Cov}(Y_{ij}, Y_{ij'} | \boldsymbol{\mu}_i, \Sigma) = \sum_{b=0}^{\infty} \sum_{b'=0}^{\infty} bb'P(Y_{ij} = b, Y_{ij'} = b' | \boldsymbol{\mu}_i, \Sigma) - E(Y_{ij} | \mu_{ij}, \Sigma_{jj})E(Y_{ij'} | \mu_{ij'}, \Sigma_{j'j'})$. $P(Y_{ij} = b, Y_{ij'} = b' | \boldsymbol{\mu}_i, \Sigma)$ can be computed with a bivariate normal distribution in a way similar to $P(Y_{ij} = b | \mu_{ij}, \Sigma_{jj})$. Under (2), the counts of OTUs j and j' are dependent if $\Sigma_{jj'} \neq 0$. That is, Σ characterizes microbial interactions with a straightforward interpretation. In addition, overdispersion is known to be common in sequencing data and can be properly accommodated through heavy tails of a log-normal distribution.

We next build a prior distribution for Σ . The number of OTUs J is often much greater than the sample size N in microbiome studies, i.e., $J \gg N$. In a high-dimensional setting, the sample covariance matrix is singular and provides an unstable estimate for Σ . To overcome the difficulty, it is common that structural assumptions are imposed on Σ ([Cai, Ren and Zhou, 2016](#)). For example, [Friedman, Hastie and Tibshirani \(2008\)](#), [Bien and Tibshirani](#)

(2011) and Cai, Liu and Luo (2011) consider the sparsity assumption that most of the elements in Σ (or Σ^{-1}) are zero or negligible for marginal independencies between features (or conditional independencies). In particular, ℓ_1 penalty is used to shrink the elements of Σ (or Σ^{-1}) to zero. Alternatively, a low-rank structure is considered, sometimes jointly with the sparsity assumption (called joint sparsity). For example, see Cai, Ma and Wu (2015); Bhattacharya *et al.* (2015) and Xie *et al.* (2018). The joint sparsity structure allows to achieve good theoretical properties, such as faster minimax rate of convergence and tighter posterior contraction rate for estimating a covariance matrix (Cai, Ma and Wu, 2015; Xie *et al.*, 2018). Taking the latter approach, we first decompose Σ as

$$(3) \quad \Sigma = \Lambda \Lambda' + \sigma^2 \mathbf{I}_J,$$

where $\boldsymbol{\lambda}_j = [\lambda_{j1}, \dots, \lambda_{jk}]'$ and $\Lambda = [\boldsymbol{\lambda}_1', \dots, \boldsymbol{\lambda}_J']'$ is a $J \times K$ factor loading matrix with $K \ll J$. The model assumes most of the covariance structure between OTUs is explained by a small number of factors to obtain a more accurate and reliable estimate of Σ in the case of $N \ll J$. We assume an isotropic noise and consider a conditionally conjugate prior distribution $\sigma^2 \sim \text{inv-Ga}(a_\sigma, b_\sigma)$ with fixed a_σ and b_σ for easy computation. If needed, independent idiosyncratic noise can be considered by letting $\Sigma = \Lambda \Lambda' + \text{diag}(\sigma_j^2)$ and $\sigma_j^2 \stackrel{iid}{\sim} \text{inv-Ga}(a_\sigma, b_\sigma)$. We introduce joint sparsity on Σ by considering a Dirichlet-Laplace prior in Bhattacharya *et al.* (2015),

$$(4) \quad \begin{aligned} \tau_k | a_\tau, b_\tau &\stackrel{iid}{\sim} \text{Ga}(a_\tau, b_\tau), \\ \phi = (\phi_1, \dots, \phi_J) | a_\phi &\sim \text{Dir}(a_\phi, \dots, a_\phi), \\ \lambda_{jk} | \phi_j, \tau_k &\stackrel{indep}{\sim} \text{DE}(\phi_j \tau_k), \end{aligned}$$

where $\text{DE}(a)$ represents the double-exponential (Laplace) distribution with scale parameter a , and $\text{Ga}(a, b)$ is the gamma distribution with shape parameter a and scale parameter b (so mean a/b). Under the model in (4), a small value of ϕ_j shrinks λ_{jk} toward zero for all k , and $\Sigma_{jj'}$ tends to have small values for all j' . That is, ϕ_j induces joint sparsity for Σ together with K . OTUs with a small value of ϕ_j may be those less interacting with other OTUs. The model provides an easy interpretation of the interrelationships between OTUs and reliable inference even for cases with $N \ll J$. The double-exponential distribution for λ_{jk} has heavier tails and a more pointed center than the normal distribution that is a convenient choice, and facilitates sparsity in λ_{jk} , resulting in sparsity in Σ . Theorem 3.1 of Bhattacharya *et al.* (2015) proves that when a_ϕ is set to be $J^{-(1+b)}$ for any $b > 0$, the posterior contraction rate of λ_{jk} achieves the minimax rate. However, our simulation studies show that the model with $a_\phi = 1/J$ tends to overshrink λ_{jk} even when only a small number of OTUs interact, and we fix $a_\phi = 1/2$ with soften conditions for the contraction rate. We fix the factor dimension K at a reasonably large value for computational convenience. If desired, an exponentially decaying prior such as a Poisson distribution can be placed on K to attain optimal posterior contraction rate (Pati *et al.*, 2014). Pati *et al.* (2014) used the Dirichlet-Laplace prior for vectorized loadings $\text{vec}(\Lambda)$ in a Bayesian factor model for a multivariate normal outcome vector with mean zero and did not attempt to induce a joint sparsity structure. Xie *et al.* (2018) used a spike-and-slab prior for ϕ_j and developed a matrix spike-and-slab LASSO prior under the Gaussian sampling distribution assumption. However, placing spike-and-slab priors for individual matrix elements may cause computational difficulties, especially for large J . Similar to Bhattacharya and Dunson (2011) and Xie *et al.* (2018), we do not place any constraints on Λ such as orthogonality of the columns nor attempt to interpret latent factors since the primary interest of inference is on Σ .

We re-write the model in (1) and (3) by introducing a latent normal vector $\boldsymbol{\eta}_i \stackrel{iid}{\sim} N_K(0, I_K)$;

$$(5) \quad \tilde{Y}_{ij}^* | \mu_{ij}, \boldsymbol{\lambda}_j, \boldsymbol{\eta}_i, \sigma^2 \stackrel{indep}{\sim} N_1(\mu_{ij} + \boldsymbol{\lambda}'_j \boldsymbol{\eta}_i, \sigma^2).$$

By integrating over $\boldsymbol{\eta}_i$, we obtain the normal distribution with covariance matrix Σ in (3) for \tilde{Y}_{ij}^* . The conditional independence between \tilde{Y}_{ij}^* given $\boldsymbol{\eta}_i$ in (5) greatly facilitates the posterior computation. Furthermore, it enables easy implementation of a zero-inflated model. Excess zeros in microbiome data are very common. If excess zeros are not compatible with the distribution in (2), the resulting inferences can be distorted. For a zero-inflated model, we introduce binary indicators δ_{ij} that represent the absence/presence of OTUs, and assume $\delta_{ij} | \epsilon_{ij} \stackrel{indep}{\sim} Ber(\epsilon_{ij})$, where ϵ_{ij} is the probability of OTU j being absent in sample i . We let $\delta_{ij} = 1$ indicate the absence of OTU j in sample i , so $Y_{ij} = 0$. Given $\delta_{ij} = 0$, we assume, for $y = 0, 1, 2, \dots$,

$$(6) \quad P(Y_{ij} = y | \mu_{ij}, \boldsymbol{\lambda}_j, \boldsymbol{\eta}_i, \sigma^2, \delta_{ij} = 0) = \Phi_1(\log(y+1) | \mu_{ij} + \boldsymbol{\lambda}'_j \boldsymbol{\eta}_i, \sigma^2) - \Phi_1(\log(y) | \mu_{ij} + \boldsymbol{\lambda}'_j \boldsymbol{\eta}_i, \sigma^2).$$

Given the presence of an OTU, the model in (6) generates counts, some of which can be zero. Given $\boldsymbol{\delta}_i = (\delta_{i1}, \dots, \delta_{iJ})$, a vector of \tilde{Y}_{ij}^* with $\delta_{ij} = 0$ follows a multivariate normal distribution, and its mean vector and covariance matrix are a subvector of $\boldsymbol{\mu}_i$ omitting the elements with $\delta_{ij} = 1$ and a submatrix of Σ omitting the rows and columns with $\delta_{ij} = 1$, respectively. That is, $\boldsymbol{\mu}_i$ and Σ provide inferences on the mean abundance and interrelationship structure even when the zero inflation component is added to the model. We relate covariates \mathbf{x}_i to the probability of $\delta_{ij} = 1$ by using a probit link function,

$$(7) \quad \epsilon_{ij} = \Phi_1(\kappa_{j0} + \mathbf{x}'_i \boldsymbol{\kappa}_j | 0, 1),$$

where κ_{j0} and $\boldsymbol{\kappa}_j = (\kappa_{j1}, \dots, \kappa_{jP})'$ are parameters that quantify the effects of \mathbf{x}_i on ϵ_{ij} . We consider a normal distribution for the prior of κ_{jp} , $\kappa_{jp} \stackrel{iid}{\sim} N(\bar{\kappa}_p, u_\kappa^2)$, $p = 0, \dots, P$. With a high proportion of zero counts, adding subject specific random effects into ϵ_{ij} may produce unstable model fitting (Agarwal, Gelfand and Citron-Pousty, 2002). Thus, the model in (7) does not include subject specific random effects.

Lastly, we relate covariates \mathbf{x}_i and subject-specific group factors g_i to the mean OTU abundances through μ_{ij} ;

$$(8) \quad \mu_{ij} = r_i + \alpha_j + \mathbf{x}'_i \boldsymbol{\beta}_j + s_{g_i, j}.$$

r_i and α_j are sample size factors and OTU size factors, respectively. The observed OTU counts are a product of both the library size (total number of reads) and the OTU baseline abundance. r_i 's normalize OTU counts across samples, and α_j 's account for variability in OTU baseline abundances. We let r_i and α_j random. Thus, the model performs model-based normalization and addresses compositionality. We will specify priors of r_i and α_j below. In (8), regression coefficients β_{jp} quantify the change in the abundance of OTU j from the mean abundance by x_{ip} (so-called a factor effects model in an ANOVA setting). Under the formulation, choosing a reference category for a categorical covariate is not required, and an implicit assumption of the presence of an OTU under the arbitrarily chosen reference category is not needed to infer the effects of the other categories. When any covariate is categorical, \mathbf{x}_i in (8) is different from that in (7) due to a different parameterization of the covariate. An example will be illustrated in § 3.2. When no covariate is available as in Simulation 1 in § 3.1 and the skin microbiome data in § 4.1, we simply drop the regression terms $\mathbf{x}'_i \boldsymbol{\kappa}_j$ and $\mathbf{x}'_i \boldsymbol{\beta}_j$ from (7) and (8), respectively, and use the simplified model to infer OTU interaction

structure. $s_{g_i,j}$'s in (8) are random effects to account for between-subject heterogeneity and induce dependence among the samples collected from the same subject. We assume normal priors $\beta_{jp} \stackrel{iid}{\sim} N(0, u_\beta^2)$ with fixed u_β^2 . In addition, we place a sum-to-zero constraint on the prior of β_{jp} 's corresponding to the categories of a categorical covariate, and the model ensures meaningful inference on β_{jp} . If desired, a joint prior distribution of κ_j and β_j can be considered. For example, we assume $(\kappa'_j, \beta'_j)' \stackrel{iid}{\sim} N(\mathbf{0}, \mathbf{V})$, and \mathbf{V} accommodates potential association between covariates' effects on presence/absence of an OTU and their effects on the abundance of the OTU. We let $s_{g_i,j} | u_s^2 \stackrel{iid}{\sim} N(0, u_s^2)$ and $u_s^2 \sim Ga(a_s, b_s)$. Due to $s_{g_i,j}$, the marginal covariance matrix of $\tilde{\mathbf{Y}}_i^*$ is $\Omega = \Sigma + u_s^2 I_J$, and the marginal correlations between OTUs j and j' are $\rho_{jj'} = \{\Sigma_{jj'} + u_s^2 1(j=j')\} / \sqrt{(\Sigma_{jj} + u_s^2)(\Sigma_{j'j'} + u_s^2)} \in (-1, 1)$. While any of parameters, Σ , Ω and $\rho_{jj'}$, can be considered as a measure of dependence between OTUs, we use $\rho_{jj'}$ for easy interpretation in the simulation studies and real data analyses illustrated later.

Recall that the mean and median of Y_{ij}^* are proportional to $\exp(r_i + \alpha_j)$, implying that r_i and α_j are not identifiable. To circumvent potential identifiability issues, we follow [Li et al. \(2017\)](#) and use the mean-constrained prior with a mixture of mixture of normals on r_i and α_j :

$$(9) \quad \begin{aligned} r_i | \psi^r, \omega^r, \xi^r &\stackrel{iid}{\sim} \sum_{l=1}^{L^r} \psi_l^r \left\{ \omega_l^r N(\xi_l^r, u_r^2) + (1 - \omega_l^r) N\left(\frac{v_r - \omega_l^r \xi_l^r}{1 - \omega_l^r}, u_r^2\right) \right\}, \\ \alpha_j | \psi^\alpha, \omega^\alpha, \xi^\alpha &\stackrel{iid}{\sim} \sum_{l=1}^{L^\alpha} \psi_l^\alpha \left\{ \omega_l^\alpha N(\xi_l^\alpha, u_\alpha^2) + (1 - \omega_l^\alpha) N\left(\frac{v_\alpha - \omega_l^\alpha \xi_l^\alpha}{1 - \omega_l^\alpha}, u_\alpha^2\right) \right\}, \end{aligned}$$

where v_r and v_α are prespecified mean constraints for the distributions of r_i and α_j , respectively. u_r^2 and u_α^2 are fixed. Different from a multinomial model that conditions on sample total counts, our model assumes $E(Y_{ij}^* | \mu_{ij}, \Sigma) \propto \exp(\mu_{ij}) = \exp(r_i + \alpha_j + \mathbf{x}'_i \beta_j + s_{g_i,j})$ in (8), and simultaneously performs model-based normalization through random r_i 's. It flexibly accounts for compositionality in microbiome data and improves the inference on parameters of primary interest compared to a model using plug-in empirical estimates for normalizing factors ([Shuler et al., 2021](#)). To specify the value of v_r , we obtain sample scale factor estimates by the cumulative sum scaling (CSS) normalization method in [Paulson et al. \(2013\)](#), and fix v_r at the average of those estimates. Specifically, we let $v_r = \frac{1}{N} \sum_{i=1}^N \log(\sum_{j=1}^J Y_{ij})$, where q_i is set as the largest quantile such that the difference in quantiles across samples is small enough. Then we set $v_\alpha = \frac{1}{NJ} \sum_{i=1}^N \sum_{j=1}^J \log(Y_{ij} + 0.01) - v_r$. [Lee and Sison-Mangus \(2018\)](#) and [Shuler et al. \(2021\)](#) showed that overall means $r_i + \alpha_j$ can be well estimated under the mean-constrained prior and their posterior inference is not sensitive to the choice of v_r and v_α . To complete the specification of the mean-constrained prior, we place Dirichlet priors for $\psi^\chi = (\psi_1^\chi, \dots, \psi_{L^\chi}^\chi)$ and beta priors for ω_l^χ , $\chi \in \{r, \alpha\}$, $\psi^\chi \sim Dir(a_\psi^\chi, \dots, a_\psi^\chi)$ and $\omega_l^\chi \stackrel{iid}{\sim} Be(a_\omega^\chi, b_\omega^\chi)$, where the hyperparameters a_ψ^χ , a_ω^χ and b_ω^χ are fixed. Finally, we set $\xi_l^\chi \stackrel{iid}{\sim} N(\bar{\xi}^\chi, v_\chi^2)$ with fixed $\bar{\xi}^\chi$ and v_χ^2 . With random mixture weights, ω_l^χ and ψ_l^χ , and random locations ξ_l^χ , the mixture models in (9) flexibly capture various shapes of distributions, while keeping their means at v_χ and provide reasonable estimates of $r_i + \alpha_j$.

2.2. Posterior Computation. Let $\boldsymbol{\theta} = \{\lambda_{jk}, \phi_j, \tau_k, \kappa_{jp}, \delta_{ij}, \eta_i, \sigma^2, r_i, \alpha_j, \beta_{jp}, s_{g_i,j}, u_s^2, \omega_l^\alpha, \psi_l^\alpha, \xi_l^\alpha, \omega_l^r, \psi_l^r, \xi_l^r\}$ be a vector of all random parameters. We use Markov chain Monte Carlo (MCMC) methods to draw samples from the posterior distribution of $\boldsymbol{\theta}$. We write a Laplace

distribution in (4) as a normal scale mixture to facilitate the posterior computation, and introduce latent mixture indicators for easy computation in updating ω_i^χ , ψ_i^χ and ξ_i^χ , $\chi \in \{r, \alpha\}$. Given the latent variables, all parameters except for ϕ_j are in standard conjugate forms and can be easily updated through a data augmented Gibbs step. Details of the posterior computation are given in Supp. §1 ([Shuangjie et al., 2022](#)). We examined the mixing and convergence of the Markov chains using trace plots and autocorrelation plots and did not find evidence of poor mixing or bad convergence for both the upcoming simulation examples and the real data analyses. The open-source code that implements the model is available online at <https://github.com/Zsj950708/ZI-MLN>.

3. Simulation Studies.

3.1. Simulation 1. We performed simulation studies and assessed the performance of the zero-inflated multivariate log-normal kernel model (ZI-MLN). For Simulation 1, we considered a case where no covariate is included, and each subject has one sample. We fitted a simplified model that has $\mu_{ij} = r_i + \alpha_j + s_{g_i,j}$ and $\epsilon_{ij} = \Phi_1(\kappa_{j0} | 0, 1)$. The simplified model is useful in estimating the interactions between OTUs for data without covariates. We let $J = 150$ OTUs and $N = 20$ samples, a sample from each of $M = 20$ subjects. For joint sparsity, we set $K^{\text{tr}} = 5$ and generated $e_{jk} \stackrel{iid}{\sim} \text{Ber}(g)$ with sparsity level $g = 0.8$. We then let $\lambda_{jk}^{\text{tr}} = 0$ if $e_{jk} = 1$ and otherwise, simulated $\lambda_{jk}^{\text{tr}} \stackrel{iid}{\sim} \text{Unif}(-3, 3)$. We let $\Sigma^{\text{tr}} = \Lambda^{\text{tr}} \Lambda^{\text{tr},\top} + \sigma^{2,\text{tr}} \mathbf{I}_J$ with $\sigma^{2,\text{tr}} = 1$. We also simulated random effects $s_{g_i,j}^{\text{tr}} \stackrel{iid}{\sim} \mathcal{N}(0, u_s^{2,\text{tr}})$ with $u_s^{2,\text{tr}} = 1$, sample size factors $r_i^{\text{tr}} \stackrel{iid}{\sim} \text{Unif}(3, 7)$ and OTU size factors $\alpha_j^{\text{tr}} \stackrel{iid}{\sim} \text{Unif}(0, 2)$. We then simulated $\mathbf{Y}_i^{\star,\text{tr}} \stackrel{iid}{\sim} \text{log-N}_J(r_i^{\text{tr}} \mathbf{1}_J + \boldsymbol{\alpha}^{\text{tr}} + \mathbf{s}_i^{\text{tr}}, \Sigma^{\text{tr}})$. For excess zeros, we generated $\kappa_{j0}^{\text{tr}} \stackrel{iid}{\sim} \text{Unif}(-1, 0)$ and simulated $\delta_{ij}^{\text{tr}} \mid \epsilon_j^{\text{tr}} \stackrel{iid}{\sim} \text{Ber}(\epsilon_j^{\text{tr}})$ with $\epsilon_j^{\text{tr}} = \Phi_1(\kappa_{j0}^{\text{tr}} | 0, 1)$. We then let $Y_{ij} = 0$ if $\delta_{ij}^{\text{tr}} = 1$ and otherwise, let $Y_{ij} = \lfloor Y_{ij}^{\star,\text{tr}} \rfloor$. It yielded approximately 40% of Y_{ij} being 0. The lower left triangle of the heatmap in Fig 1(a) illustrates the true marginal correlation matrix $\rho_{jj'}^{\text{tr}} = \{\Sigma_{jj'}^{\text{tr}} + u_s^{2,\text{tr}} \mathbf{1}(j = j')\} / \sqrt{(\Sigma_{jj}^{\text{tr}} + u_s^{2,\text{tr}})(\Sigma_{j'j'}^{\text{tr}} + u_s^{2,\text{tr}})}$. Empirical correlation estimates $\rho_{jj'}^{\text{em}}$ are computed using transformed raw counts and illustrated in Supp. Fig 2(a) ([Shuangjie et al., 2022](#)). It shows that naive correlation estimates are noisy and do not capture the true interrelationship between OTUs.

To fit the model, we set the fixed hyperparameters as follows; For the mean-constrained priors of r_i and α_j , we let $L^r = 5$, $L^\alpha = 10$, $a_\psi^r = a_\psi^\alpha = 1$, and $a_\omega^r = b_\omega^r = a_\omega^\alpha = b_\omega^\alpha = 5$. The values of the mean constraints v^r and v^α were specified through the empirical approach described in § 2.1. We set the prior mean and variance of κ_{j0} , $\bar{\kappa}_0 = 0$ and $u_\kappa^2 = 3$. Also, we set $a_\sigma = b_\sigma = 3$ and $a_s = b_s = 1$. Lastly, we set $K = 10$, $a_\phi = 1/2$, $a_\tau = 1$ and $b_\tau = 1/50$. We simulated posterior samples through MCMC described in § 2.2. We discarded the first 15,000 draws for burn-in and kept the next 15,000 draws for posterior inference. It took 25 minutes for every 5,000 iterations on a M1 Mac. Assessment of MCMC simulation convergence is discussed in Supp. §2.1 ([Shuangjie et al., 2022](#)). We also checked the posterior distributions of τ_k to examine if a greater value of K is needed. The posterior distributions of some τ_k 's are greatly concentrated close to zero, indicating that $K = 10$ is sufficiently large for the dataset. We also performed sensitivity analyses to the specification of a_ϕ and b_τ to examine the robustness of the model in estimating Σ .

Posterior inference on the marginal correlations $\rho_{jj'}$ is illustrated in Fig 1. The heatmap in panel (a) compares posterior mean estimates $\hat{\rho}_{jj'}$ in the upper right triangle to their truth $\rho_{jj'}^{\text{tr}}$ in the lower left triangle. Panel (b) shows a histogram of the differences $\hat{\rho}_{jj'} - \rho_{jj'}^{\text{tr}}$, $j < j'$. In the histogram, the differences are tightly centered around 0, indicating that the method

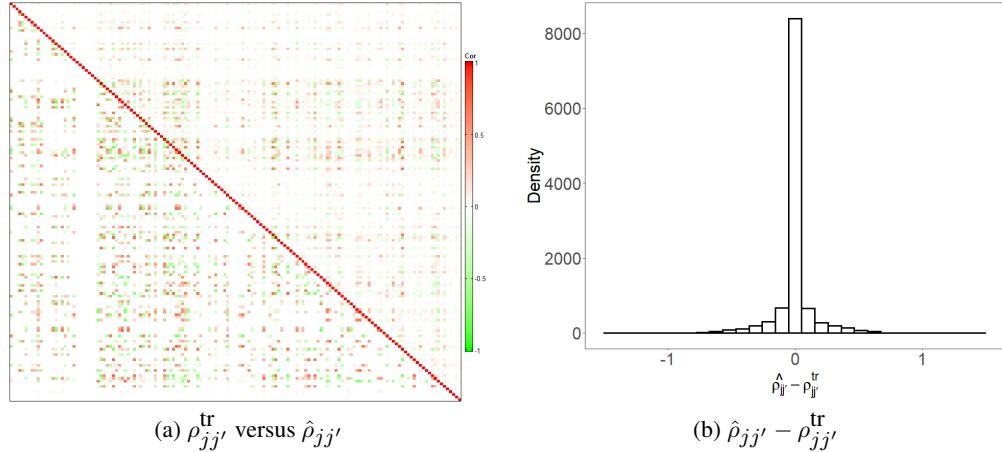


FIG 1. [Simulation 1] The upper right and lower left triangles of the heatmap in panel (a) illustrate posterior estimates of correlations $\hat{\rho}_{jj'}$ and their true values $\rho_{jj'}^{\text{tr}}$, respectively. Panel (b) has a histogram of differences between $\hat{\rho}_{jj'}$ and $\rho_{jj'}^{\text{tr}}$.

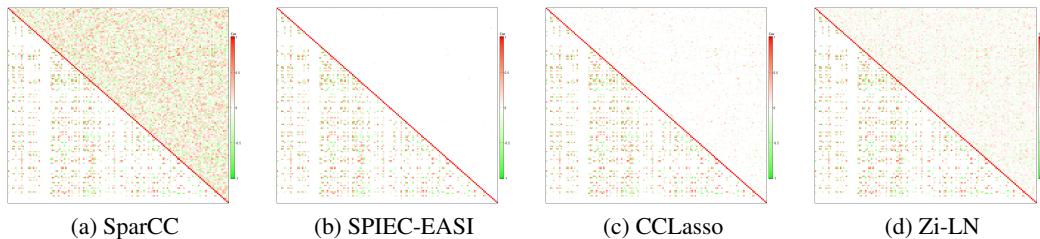


FIG 2. [Simulation 1: Comparison] The upper right and lower left triangles of each heatmap illustrate estimates $\hat{\rho}_{jj'}$ of correlations between OTUs and their true values $\rho_{jj'}^{\text{tr}}$, respectively. Panels (a)-(d) are for SparCC, SPIEC-EASI, CCLasso and Zi-LN, respectively.

provides good estimates of the correlations. Our method identifies the truly inactive OTUs successfully, and the true OTU interrelationship structure is reasonably well captured even when the sample size is much smaller than the number of OTUs ($N = 20$ and $J = 150$), and excess zeros are present. Supp. Fig 3 (Shuangjie *et al.*, 2022) compares posterior mean estimates of baseline abundances $r_i + \alpha_j$ and probabilities ϵ_{ij} of an OTU being absent to their truth. In the figure, the absence/presence of OTUs and OTU baseline abundances are well estimated, which provides a crucial basis for the estimation of the parameters of primary interest, such as Σ . We performed posterior predictive checking to examine model fit under ZI-MLN. Fig 4(a) compares posterior predictive median estimates $\hat{y}_{ij}^{\text{pred}}$ of OTU counts to the observed counts y_{ij} and shows that our model provides a good model fit to the data.

For comparison, we applied SparCC (Friedman and Alm, 2012), SPIEC-EASI (Kurtz *et al.*, 2015), CCLasso (Fang *et al.*, 2015) and Zi-LN (Prost, Gazut and Brüls, 2021) that are briefly described in § 1. The comparators infer dependence structure between OTUs through the estimation of covariance or precision matrix under some sparsity assumptions and yield correlation estimates $\hat{\rho}_{jj'}$. The tuning parameter for sparsity in SparCC, SPIEC-EASI and Zi-LN is chosen by cross-validation. $\hat{\rho}_{jj'}$ under the comparators are compared to the true values $\rho_{jj'}^{\text{tr}}$ in Fig 2. Fig 3 illustrates histograms of differences $\hat{\rho}_{jj'} - \rho_{jj'}^{\text{tr}}$. Root mean square error (RMSE) for $\rho_{jj'}, j < j'$ for the models including ZI-MLN is shown in Tab 1(a). ZI-MLN outperforms in recovering the dependence structure between OTUs. Poor performance of the comparators can be because they do not account for overdispersion and/or excess zeros and/or they lack flexible normalization for compositionality. In addition, we compare our

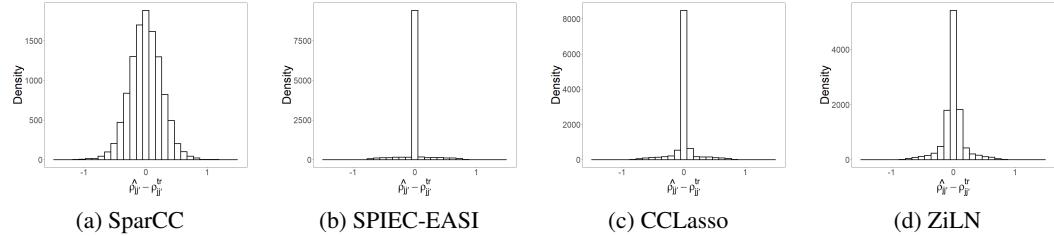


FIG 3. [Simulation 1: Comparison] A histogram of differences between $\hat{\rho}_{jj'}$ under SparCC, SPIEC-EASI, CCLasso and Zi-LN and $\rho_{jj'}^{tr}$, in panels (a)-(d), respectively.

TABLE 1

[Simulation 1: Comparison] RMSEs are computed for correlations $\rho_{jj'}$, $j < j'$, binary indicator δ_{ij} of an OTU being absent in a sample and mean abundance μ_{ij} under ZI-MLN and comparators.

Model	$\rho_{jj'}$		
Model	δ_{ij}	μ_{ij}	
ZI-MLN	0.129		
SparCC	0.258		
SPIEC-EASI	0.167		
CCLasso	0.166		
Zi-LN	0.173		

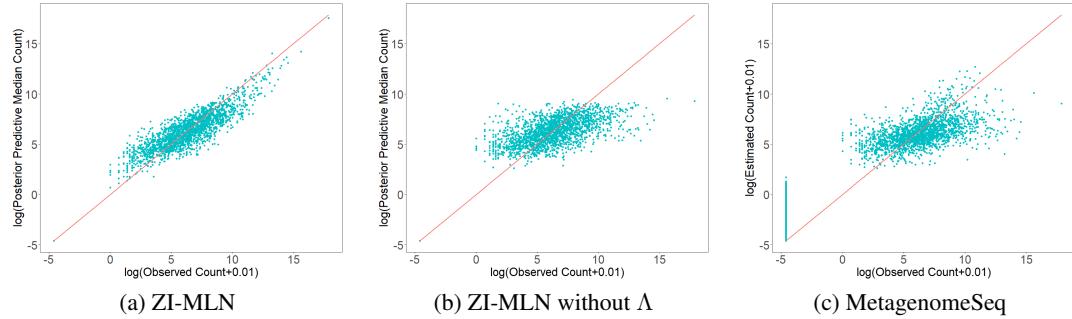


FIG 4. [Simulation 1] Scatter plots of observed $\log(y_{ij} + 0.01)$ versus $\log(\hat{y}_{ij}^{\text{pred}} + 0.01)$ estimated by ZI-MLN with Λ and ZI-MLN without Λ are shown in panels (a) and (b), respectively. $\hat{y}_{ij}^{\text{pred}}$ is the median estimate of the posterior predictive distribution. Panel (c) is the scatter plots of observed $\log(y_{ij} + 0.01)$ versus $\log(\hat{\mu}_{ij} + 0.01)$, where $\hat{\mu}_{ij}$ are mean abundances of OTUs estimated by metagenomeSeq.

method to ZI-MLN without Λ , a simpler version of our ZI-MLN, and metagenomeSeq in Paulson *et al.* (2013) for comparison of the estimation of μ_{ij} and δ_{ij} . We simplified our ZI-MLN by letting $\Sigma = \sigma^2 I_J$ and kept the remaining model components including zero-inflation and subject-specific random effects the same. We call it “ZI-MLN without Λ .” MetagenomeSeq is a likelihood-based model that uses transformed counts $\log_2(y_{ij} + 1)$ and assumes a zero-inflated normal mixture model separately for individual OTUs, where the mean has a regression function of covariates, a sample size factor fixed at estimates by CSS normalization method and an OTU size factor similar to ZI-MLN. Under metagenomeSeq, the zero inflation probabilities of y are common for all OTUs in a sample and regressed on the sample total counts through a logit link. An EM algorithm is used to estimate unknown parameters.

The additional comparators do not account for the interrelationships between OTUs and do not provide any inference on OTU interaction. We compared parameter estimates of μ_{ij} and δ_{ij} under each of the three models, including ZI-MLN, to the truth and computed RMSE for the parameters, summarized in Tab 1(b). The table shows that our model outperforms the comparators in the estimation of OTU mean abundances and absence/presence. Especially, comparison to ZI-MLN without Λ indicates that ignoring the dependence structure among counts when it is present can deteriorate the inference on the other parameters, including μ_{ij} . It is also indicated from posterior predictive checking under ZI-MLN without Λ shown in Fig 4(b). Comparison of mean abundance estimates $\hat{\mu}_{ij}$ by metegenomSeq to observed counts in Fig 4(c) also shows potential model misfit under metagenomeSeq.

3.2. Simulation 2. We conducted Simulation 2 for a case having covariates. We examined the estimation of covariate effects on OTU abundances and their presence/absence in addition to the estimation of Σ . We set the number of OTUs $J = 150$ and assumed two samples from each of $M = 35$ subjects under two experimental conditions. We thus have the number of samples $N = 70$ and $g_i \in \{1, \dots, M\}$ with $n_{g_i} = 2$ for all g_i . The remaining setup is similar to that of Simulation 1. We set $K^{\text{tr}} = 5$, $\sigma^{2,\text{tr}} = 1$ and $u_s^{2,\text{tr}} = 1$, and simulated λ_{jk}^{tr} , r_i^{tr} , α_j^{tr} and $s_{g_i,j}^{\text{tr}}$, as done in Simulation 1. We included a binary covariate that represents the experimental conditions using a pair of dummy variables $(x_{i1}, x_{i2}) \in \{(1, 0), (0, 1)\}$. The corresponding coefficients β_{j1} and β_{j2} thus quantify changes in mean abundance by a condition compared to the overall mean abundance $r_i^{\text{tr}} + \alpha_j^{\text{tr}}$. In addition, we included a continuous covariate, x_{i3} generated from $N(0, 1)$, so we have $\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3})'$ with $P = 3$. For the coefficients, we set $\beta_{jp}^{\text{tr}} \stackrel{iid}{\sim} N(0, 1)$ for $p = 1, \dots, P$. For ϵ_{ij} , we let $\tilde{\mathbf{x}}_i = (x_{i2}, x_{i3})'$ with $P_\kappa = 2$ using x_{i1} as a reference category, and simulated $\kappa_{jp}^{\text{tr}} \stackrel{iid}{\sim} \text{Unif}(-0.5, 0)$, $p = 0, \dots, P_\kappa$. We finally generated counts Y_{ij} as follows; we simulated $\mathbf{Y}_i^{\star, \text{tr}} \stackrel{\text{indep}}{\sim} \log\text{-}N_J(r_i^{\text{tr}} \mathbf{1}_J + \boldsymbol{\alpha}^{\text{tr}} + \mathbf{x}_i' \boldsymbol{\beta}^{\text{tr}} + s_i^{\text{tr}}, \Sigma^{\text{tr}})$, with $\Sigma^{\text{tr}} = \Lambda^{\text{tr}} \Lambda^{\text{tr}, \prime} + \sigma^{2,\text{tr}} \mathbf{I}_J$ and $\boldsymbol{\beta}^{\text{tr}}$ being a $J \times P$ matrix of β_{jp}^{tr} . We also generated binary indicators $\delta_{ij}^{\text{tr}} | \epsilon_{ij}^{\text{tr}} \stackrel{\text{indep}}{\sim} \text{Ber}(\epsilon_j^{\text{tr}})$ with $\epsilon_{ij}^{\text{tr}} = \Phi(\kappa_{j0}^{\text{tr}} + \kappa_j^{\text{tr}, \prime} \tilde{\mathbf{x}}_i | 0, 1)$. We then let $Y_{ij} = 0$ if $\delta_{ij}^{\text{tr}} = 1$, and let $Y_{ij} = \lfloor Y_{ij}^{\star, \text{tr}} \rfloor$, otherwise. The simulated dataset has approximately 40% of counts being zero. Fig 5(a) and Supp. Fig 5(a) ([Shuangjie et al., 2022](#)) illustrate the true marginal correlations ρ_{jj}' and their naive empirical estimates $\rho_{jj}'^{\text{em}}$ using transformed counts after the normalization, respectively.

We specified the fixed hyperparameter values similar to those in Simulation 1. We set $L^r = 8$ due to a larger sample size. We set $u_\beta^2 = 25$ for the prior of β_{jp} and placed the sum-to-zero constraint for β_{j1} and β_{j2} for identifiability. We set $\bar{\kappa}_p = 0$ for all p and $u_\kappa^2 = 3$. The MCMC simulation was run over 30,000 iterations, with the first 15,000 iterations discarded as burn-in. A discussion on the chain's convergence and mixing is in Supp. §2.2 ([Shuangjie et al., 2022](#)). It took 0.7 hours on average for every 5,000 iterations on a M1 Mac.

Fig 5 illustrates posterior mean estimates $\hat{\rho}_{jj'}$ of marginal correlations between OTUs j and j' , $j \neq j'$. The figure shows that the underlying interrelationships between OTUs are well captured even with small sample size and excess zero counts. The histogram in panel (b) shows the differences $\hat{\rho}_{jj'} - \rho_{jj'}^{\text{tr}}$ are close to zero. Figs 6(a)-(b) and Supp. Figs 6(a)-(c) ([Shuangjie et al., 2022](#)) compare regression coefficient estimates, $\hat{\beta}_{jp}$ and $\hat{\kappa}_{jp}$ to their true values. From Figs 6(a)-(b), posterior mean estimates of $\beta_{j1} - \beta_{j2}$ and β_{j3} are close to the true values. Here, $\beta_{j1} - \beta_{j2}$ quantifies the difference in the mean abundances between two categories of the binary covariate. Their posterior 95% credible intervals capture the truth well. Supp. Fig 7 ([Shuangjie et al., 2022](#)) shows that posterior estimates $\hat{r}_i + \hat{\alpha}_j$ and $\hat{\epsilon}_{ij}$ are also close to their true values. To check the model fit, we compare median estimates $\hat{y}_{ij}^{\text{pred}}$ of

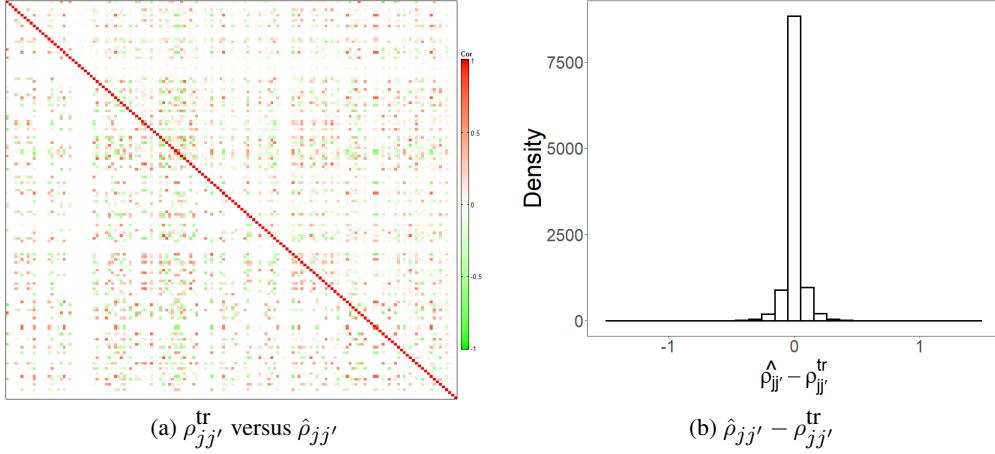


FIG 5. [Simulation 2] The upper right and lower left triangles of the heatmap in panel (a) illustrate posterior estimates of correlations $\hat{\rho}_{jj'}$ and the true values of the correlations $\rho_{jj'}^{\text{tr}}$, respectively. Panel (b) has a histogram of differences between $\hat{\rho}_{jj'}$ and $\rho_{jj'}^{\text{tr}}$.

TABLE 2
[Simulation 2: Comparison] RMSEs are computed for $\rho_{jj'}, j < j'$, δ_{ij} , μ_{ij} , $\beta_{j2} - \beta_{j1}$, β_{j3} and κ_{jp} under ZI-MLN and comparators.

Model	$\rho_{jj'}$	Model	δ_{ij}	μ_{ij}	$\beta_{j2} - \beta_{j1}$	β_{j3}	κ_{j0}	κ_{j1}	κ_{j2}
ZI-MLN	0.063	ZI-MLN	0.096	1.084	0.570	0.359	0.214	0.183	0.335
SparCC	0.176	ZI-MLN without Λ	0.123	1.172	0.750	0.426	0.234	0.191	0.361
SPIEC-EASI	0.158	MetagenomeSeq	0.130	1.962	1.409	0.843	-	-	-
CCLasso	0.155	EdgeR	-	2.205	0.902	0.585	-	-	-
Zi-LN	0.157								

(a) $\rho_{jj'}$

(b) $\delta_{ij}, \mu_{ij}, \beta_{j2} - \beta_{j1}, \beta_{j3}$ and κ_{jp}

the posterior predictive distributions to the observed counts. Fig 6(c) provides evidence for a good model fit under ZI-MLN.

For comparison, we applied the four comparators that provide estimates of associations between OTUs, SparCC, SPIEC-EASI, CCLasso and Zi-LN, to the simulated data. The heatmaps in Fig 7 and histograms in Fig 8 compare their estimates $\hat{\rho}_{jj'}$ to the truth $\rho_{jj'}^{\text{tr}}$. RMSE for $\rho_{jj'}$ are computed for comparison between the models including ZI-MLN. Tab 2(a) shows that ZI-MLN outperforms the comparators in estimating the dependencies between OTUs. Note that the comparators do not account for covariate effects, potentially resulting in poor performance. Also, we applied three other comparators, ZI-MLN without Λ , metagenomeSeq and edgeR (Robinson, McCarthy and Smyth, 2010) and compared the abundance and absence/presence related model parameters. EdgeR is a likelihood-based method that uses a negative binomial generalized linear regression approach for the analysis of HTS data. It uses the normalization factors estimated by an empirical Bayes strategy and does not account for excess zeros. Similar to ZI-MLN without Λ and metagenomeSeq, edgeR does not account for the dependence structure among OTUs and does not provide inferences on the relationship among OTUs. MetagenomeSeq and edgeR require selecting a category of a discrete covariate as a reference category, and their β_{jp} 's estimate changes in the mean abundance relative to that in the reference category. We chose x_{i1} as the reference for those

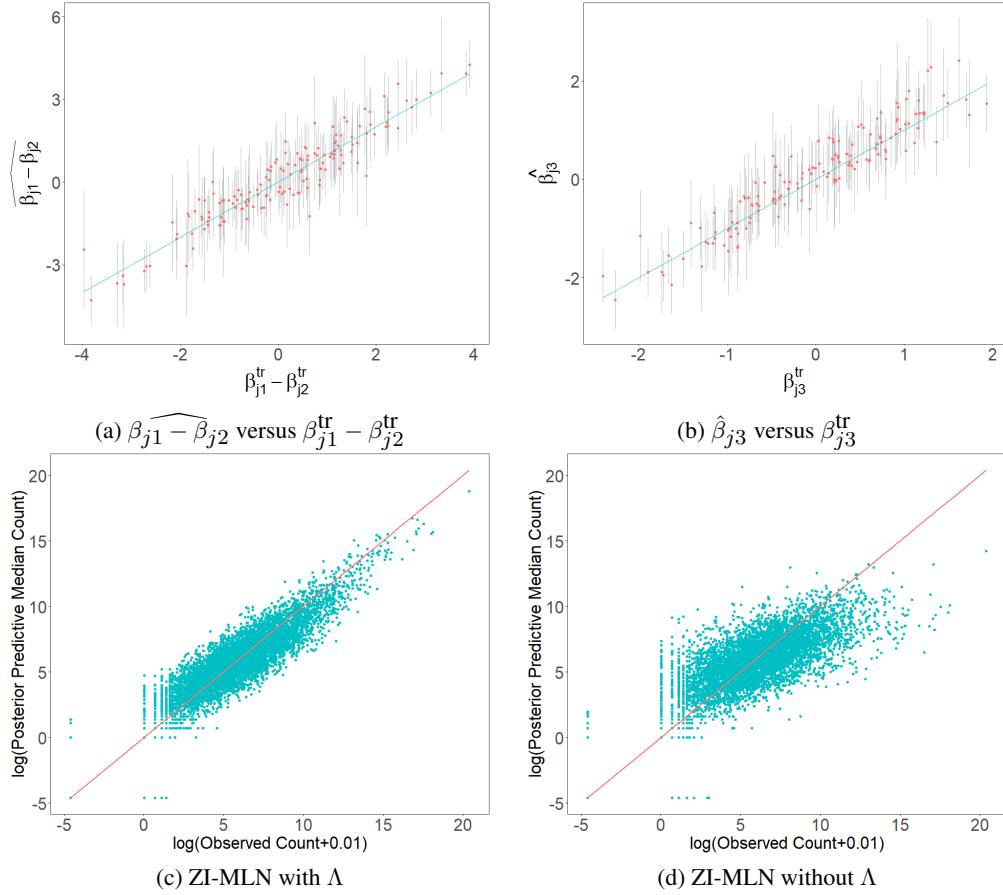


FIG 6. [Simulation 2] Panels (a) and (b) compare posterior estimates of regression coefficients $\widehat{\beta}_{j1} - \widehat{\beta}_{j2}$ and $\widehat{\beta}_{j3}$ to the truth $\beta_{j1}^{\text{tr}} - \beta_{j2}^{\text{tr}}$ and β_{j3}^{tr} , respectively, where the vertical lines represent 95% credible intervals. Panels (c) and (d) compare posterior predictive median count estimates to their observed counts on the logarithm scale, $\log(y_{ij} + 0.01)$ versus $\log(\hat{y}_{ij}^{\text{pred}} + 0.01)$. ZI-MLN with Λ and ZI-MLN without Λ are used in panels (c) and (d), respectively.

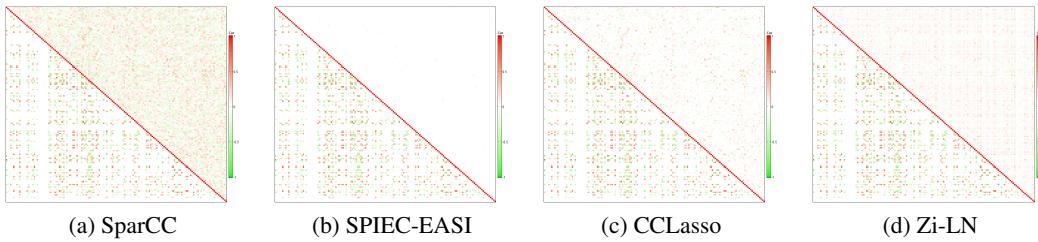


FIG 7. [Simulation 2: Comparison] The upper right and lower left triangles of each heatmap illustrate estimates $\hat{\rho}_{jj'}$ of correlations between OTUs and their true values $\rho_{jj'}^{\text{tr}}$, respectively. Panels (a)-(d) are for SparCC, SPIEC-EASI, CCLasso and Zi-LN, respectively.

methods. Supp. Figs 6(d)-(f) and 8 ([Shuangjie et al., 2022](#)) compare estimates of β_{jp} and κ_{jp} under the comparators to the truth. RMSE for each of the four models, including ZI-MLN, is computed and summarized in Tab 2(b). RMSE of κ_{jp} is not computed for metagenomeSeq since it has a logit regression of ϵ_{ij} on the total sample count, but not on covariates. The results show that our model outperforms the comparators in the estimation of the parameters, δ_{ij} , μ_{ij} , β_{jp} and κ_{jp} . We also performed posterior predictive checking for ZI-MLN without

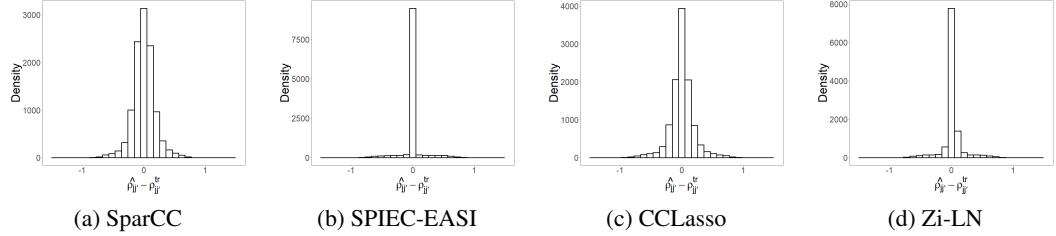


FIG 8. [Simulation 2: Comparison] A histogram of differences between $\hat{\rho}_{jj'}$ under SparCC, SPIEC-EASI, CCLasso and ZiLN and $\rho_{jj'}^{\text{tr}}$, in panels (a)-(d), respectively.

Λ by comparing $\hat{y}_{ij}^{\text{pred}}$ under the model to the observed counts. As shown in Fig 6(d), ZI-MLN without Λ provides a poor fit to the data. Their posterior mean estimates of σ^2 and u_s^2 are greatly inflated compared to their true value. Estimates $\hat{\sigma}^2$ and \hat{u}_s^2 are 3.86 and 0.77, respectively, while their true values are $\sigma^2, \text{tr} = 1$ and $u_s^2, \text{tr} = 1$. The comparison of the inference under ZI-MLN to that under ZI-MLN without Λ shows the necessity of modeling the dependence structure between OTUs to enhance the inference on the other parameters such as covariate effects when the interactions between OTUs are present. Estimates of the mean abundances under metagenomeSeq and edgeR are compared to the observed counts in Supp. Fig 9 ([Shuangjie et al., 2022](#)).

Additional Simulations. We conducted additional simulation studies, Simulations 3-5 to examine the performance of our model under various settings. In Simulation 3, we first generated correlated mean vectors $\tilde{\mu}_i^{\text{tr}} = (\tilde{\mu}_{i1}^{\text{tr}}, \dots, \tilde{\mu}_{iJ}^{\text{tr}})$ from a multivariate normal distribution and simulated OTU counts from zero-inflated Poisson distributions with means $\exp(\tilde{\mu}_{ij}^{\text{tr}})$. The simulation results show that our model provides reasonable estimates of the parameters even when the simulation truth is different from the assumed model, showing the robustness of the model. Importantly, the OTU interaction structure is also reasonably well reconstructed even when the dependency is embedded in the mean abundances, and the sampling distribution is incorrectly specified. In Simulation 4, we kept the simulation setup the same as in Simulation 2, but let $\Sigma^{\text{tr}} = \sigma^{2,\text{tr}} I_J$, i.e., OTU counts are independent given the mean parameters. Although the simulation truth is closer to the assumption made under ZI-MLN without Λ , the results show that ZI-MLN performs almost the same as well. For Simulation 5, we used SparseDOSSA in [Ma et al. \(2021\)](#) to simulate a dataset. SparseDOSSA takes a real microbiome dataset as an input, estimates some input parameters of their data-generating model, and generates a realistic microbiome dataset that has a dependence structure between OTUs using the estimates. We used the skin microbiome dataset in § 4.1 as an input dataset. An open-source software, *SparseDOSSA2* is provided by the authors. SparseDOSSA estimates a precision matrix, one of the input parameters, with ℓ_1 penalty for sparsity. The sparsity assumption is similar to that under some of the comparators, SPIEC-EASI and CCLasso. It simulates count vectors from a multinomial distribution conditioning random total counts. The dataset in the scenario was thus simulated from a model significantly different from ZI-MLN. The results greatly demonstrate the robustness of ZI-MLN. The model-based normalization appropriately accounts for differences in total counts. More importantly, the model does a good job of capturing the dependence between OTUs in the truth and recovers the truly underlying between-OTU structure reasonably well. In all simulation studies, the results also show that our model compares very favorably relative to the comparators for estimation of covariate effects and of dependence structure between OTUs. More details of Simulations 3- 5 are in Supp. §2.3-2.5 ([Shuangjie et al., 2022](#)), respectively. In addition, we assumed a different sparsity level for Λ^{tr} by generating $e_{jk} \stackrel{iid}{\sim} \text{Ber}(g)$ with $g = 0.5$, and reran analyses under the settings of Simulations 1-4. The results show that ZI-MLN recovers the truth

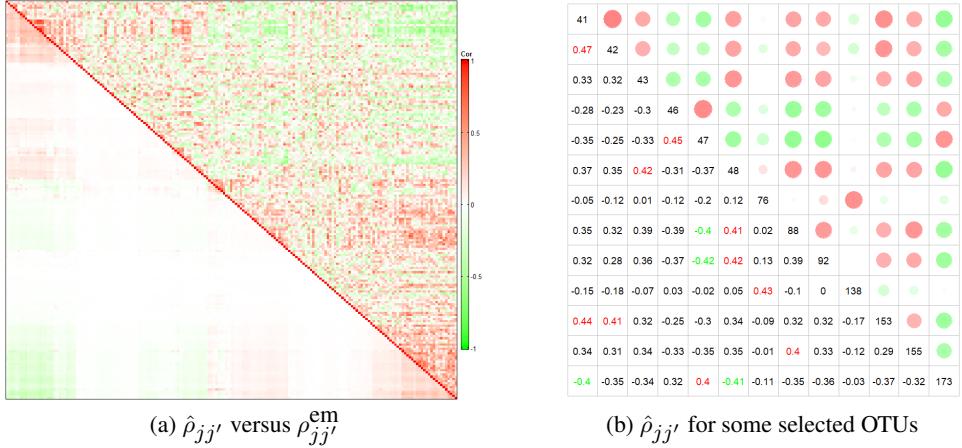


FIG 9. [Skin Microbiome Data] Posterior correlation estimates $\hat{\rho}_{jj'}$ (lower left triangle) and empirical correlation estimates $\rho_{jj'}^{\text{em}}$ (upper right triangle) are shown in panel (a). Panel (b) have the OTUs having $|\hat{\rho}_{j,j'}| \geq 0.40$ for any $j' \neq j$.

well with a lower sparsity level and works better than the comparators under the comparison metrics.

4. Real Data Analyses.

4.1. Skin Microbiome Data. We applied our ZI-MLN to a subset of the chronic wound microbiome data in [Verbanic et al. \(2020\)](#). The study was conducted to investigate the effect of debridement on the wound microbial community. Skin swab samples were collected under three conditions, healthy skin, pre-debridement, and post-debridement conditions. The skin microbiome dataset was analyzed [Shuler et al. \(2021\)](#), which showed changes in the community-level microbial richness and abundance diversity by the experimental conditions. For an illustration of ZI-MLN without covariates, we used a subset of the data that consists of $N = 20$ healthy skin samples collected from $M = 20$ subjects and investigated the interaction structure between OTUs in the healthy skin samples. We removed OTUs that have zero counts in more than 50% of the samples, leaving $J = 187$ OTUs for analysis. The threshold of 50% was chosen so that each OTU has at least 10 nonzero counts, and the model parameters such as α_j can be reliably estimated. Manual inspection of the curated OTU list indicated that the threshold chosen did not eliminate OTUs of major biological importance. In addition, we performed sensitivity analysis to the specification of the threshold. We found that any reasonable choice has little impact on the posterior inference, showing robustness of our model. Details of the sensitivity analysis are summarized in Supp. §3.1 ([Shuangjie et al., 2022](#)). Fig 9(a) shows empirical correlation estimates $\rho_{jj'}^{\text{em}}$ computed using $\log(y_{ij} + 0.01)$ after normalization with CSS sample size factor estimates. To fit ZI-MLN, the values of the fixed hyperparameter values were set similar to those of Simulation 1 in § 3.1. The MCMC simulation was run over 30,000 iterations, with the first 15,000 iterations discarded as burn-in. It took 25 minutes for every 5,000 iterations on a M1 Mac.

Fig 9(a) illustrates posterior mean estimates $\hat{\rho}_{jj'}$ of the marginal correlations for all OTUs. From panel (a), correlation estimates are overall small for most of (j, j') , implying weak interactions between OTUs. Compared to $\rho_{jj'}^{\text{em}}$, $\hat{\rho}_{jj'}$'s are shrunk toward zero for many OTUs. The overall weak correlations among OTUs in the skin samples are consistent with previous analysis. Specifically, [Bashan et al. \(2016\)](#) analyzed data from the Human Microbiome Project and the Student Microbiome Project, and compared samples from the gut and

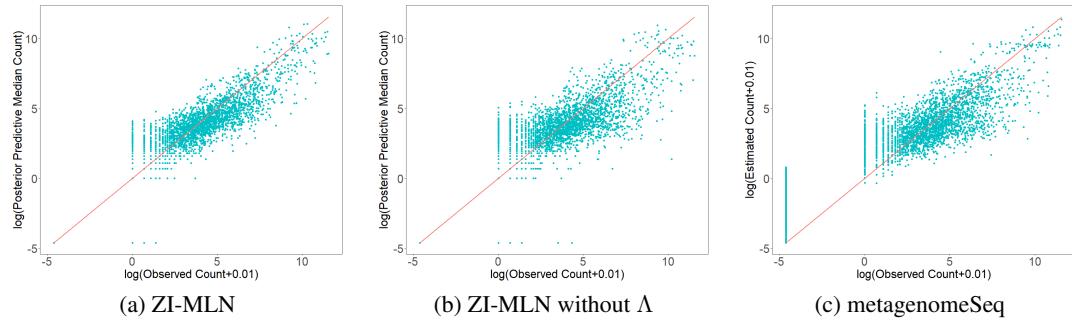


FIG 10. [Skin Microbiome Data: Comparison] Panels (a) and (b) have scatter plots of observed $\log(y_{ij} + 0.01)$ versus $\log(\hat{y}_{ij}^{pred} + 0.01)$ under ZI-MLN and ZI-MLN without Λ , respectively. Panel (c) is the scatter plots of observed $\log(y_{ij} + 0.01)$ versus mean abundance estimates $\log(\hat{\mu}_{ij} + 0.01)$ by metagenomeSeq.

oral microbiome to those from the skin microbiome. They reported that, while the gut and mouth microbiome samples appeared to exhibit universal dynamics of inter-species interactions, the extent of such interactions in the skin microbiome samples was relatively low. Fig 9(b) presents $\hat{\rho}_{jj'}$ for the OTUs that have $|\hat{\rho}_{jj'}| \geq 0.40$ for any $j' \neq j$, where the value of 0.4 is arbitrarily chosen to make the estimates readable. Taxonomic information of the OTUs in Fig 9(b) is given in Supp. Tab 4 ([Shuangjie et al., 2022](#)). From panel (b) and the supp. table 4 ([Shuangjie et al., 2022](#)), OTUs 43 and 88 belonging to genera *Porphyromonas* and *Peptoniphilus*, respectively, are estimated to be positively correlated with $\hat{\rho} = 0.39$. Interestingly, they were found to co-occur in a large sample of genitourinary microbiome samples ([Qin et al., 2021](#)) as well as vaginal samples ([Xiaoming et al., 2021](#)) and were suggested to be ‘keystone’ species, i.e., strongly interacting species that help define their ecological system. These species are also found to co-occur in skin samples ([Chattopadhyay et al., 2021](#)), where they are more abundant in patients with diabetic foot ulcers ([Park et al., 2019](#)). OTUs 43 and 48 having correlation estimate $\hat{\rho} = 0.42$ belong to genera *Porphyromonas* and *Campylobacter*, respectively, that are both potentially pathogenic. *Porphyromonas* is a known pathogenic genus in periodontitis and is a risk factor in inflammatory bowel disease, while *Campylobacteri* is a known gut and oral pathogen with a role in inflammatory bowel disease. Their positive correlation estimate may reflect a tendency to co-occur, as both are observed in inflammatory bowel disease ([Cai et al., 2021](#)). From Supp. Tab 4 ([Shuangjie et al., 2022](#)), OTUs that have a large positive value of $\hat{\rho}_{j,j'}$ tend to be phylogenetically closely related. For example, OTUs 41 and 42 having $\hat{\rho} = 0.47$ belong to the same order *Micrococcales*. Similarly, OTUs 46 and 47 with $\hat{\rho} = 0.45$ having are in family *Chitinophagaceae*. On the other hand, some OTUs are estimated to have a positive association with phylogenetically distant OTUs. For example, the correlation estimates between OTU 153 and OTUs 41 and 42 are $\hat{\rho} = 0.44$ and 0.41, respectively, but OTU 153 is not phylogenetically closely related to OTUs 41 and 42. Interestingly, OTU 153 has similar interaction patterns with OTUs 41 and 42 in the same genus. Fig 10(a) has a scatter plot comparing the posterior predictive median estimates $\hat{y}_{ij}^{\text{pred}}$ to the observed counts. The posterior predictive checking indicates a good model fit by ZI-MLN.

We also applied the comparators, SparCC, SPIEC-EASI, CCLasso and Zi-LN to the skin microbiome data for comparison. Their correlation estimates $\hat{\rho}_{jj'}$ are illustrated in Fig 11 with the naive estimates of the correlations. SPIEC-EASI and CCLasso produce $\hat{\rho}_{jj'}$ very close or equal to zero for most OTU pairs, while SparCC has nonzero estimates for a majority of $\rho_{jj'}$. In addition, ZI-MLN without Λ and metagenomeSeq are applied for further comparison. In Fig 10(b), the posterior predictive median estimates $\hat{y}_{ij}^{\text{pred}}$ under ZI-MLN

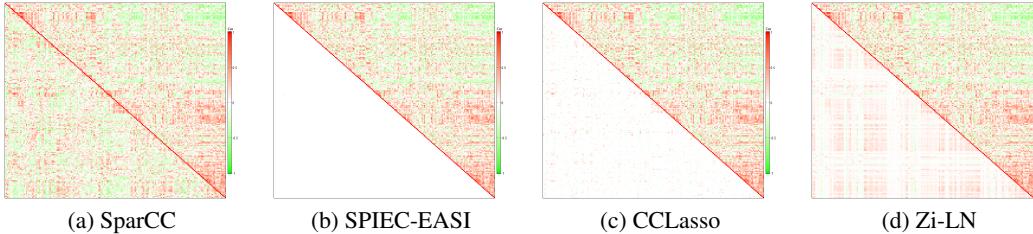


FIG 11. [Skin Microbiome Data: Comparison] Correlation estimates $\hat{\rho}_{jj'}$ (lower left triangle) and empirical correlation estimates $\rho_{jj'}^{\text{em}}$ (upper right triangle) are shown. The estimates in panel (a)-(d) are obtained by SparCC, SPIEC-EASI, CCLasso and Zi-LN, respectively.

without Λ are compared to the observed counts. In panel (c), mean abundance estimates under metagenomeSeq are compared to the observed counts. A comparison of those plots to that in panel (a) indicates that our ZI-MLN provides a better model fit, possibly because our model accounts for microbial interactions.

4.2. Human Gut Microbiome Data. We analyzed the microbiome dataset available from the inflammatory bowel disease (IBD) multi-omics database (<https://ibdmdb.org/>) with our ZI-MLN. Crohn's disease (CD) and ulcerative colitis (UC) are the most prevalent forms of IBD and are characterized by chronic inflammation of the gastrointestinal tract. As part of the Integrative Human Microbiome Project (iHMP), [Lloyd-Price et al. \(2019\)](#) conducted an integrated study of multiple molecular features of the gut microbiome to investigate host- and microbiome-specific taxonomic and molecular features related to IBD and how they vary over time. In the study, biopsies were taken during the initial screening colonoscopy from the participants who were recruited from multiple medical centers and sequenced using 16S rRNA gene amplicon sequencing. For an illustration of our statistical model, we used part of their 16S rRNA sequencing data. In particular, we included the samples obtained from 37 pediatric participants from two recruitment sites, Cincinnati Children's Hospital and Massachusetts General Hospital (MGH) Pediatrics. For some subjects, two samples were collected from different biopsy locations, resulting in a total of 67 samples. In addition to biopsy locations, we included one continuous covariate, age and five categorical covariates such as sex, race, recruitment site and disease phenotype. Disease phenotype is a trinary covariate taking a value of UC, CD or non-IBD, and the others are binary, resulting in $P = 12$ after adding dummy variables to indicate the categories of the discrete covariates. Supp. Tab 5 ([Shuangjie et al., 2022](#)) lists all covariates with their supports. We removed OTUs having zero count in more than 80% of the samples or average counts smaller than five. $J = 107$ OTUs are left after the preprocessing. With the threshold of 80%, each OTU has approximately 13.4 nonzero counts, similar to that in the skin microbiome data analysis, to ensure reliable estimates of κ_{jp} , β_{jp} and Σ . We specified hyperparameters similar to those in § 3.2. The MCMC simulation was run over 30,000 iterations, with the first 15,000 iterations discarded as burn-in. It took 0.75 hours for every 5,000 iterations on a M1 Mac.

Posterior mean estimates $\hat{\rho}_{jj'}$ of the marginal correlations (lower left triangle) are illustrated with naive empirical correlation estimates $\rho_{jj'}^{\text{em}}$ (upper right triangle) in Fig 12(a). The figure shows relatively rich microbial interactions in the gut microbiome samples as reported in [Bashan et al. \(2016\)](#). Fig 12(b) reports $\hat{\rho}_{jj'}$ for the OTUs having $|\hat{\rho}_{jj'}| > 0.5$ for any $j' \neq j$, where the value of 0.5 is chosen to make the estimates in the figure readable. Taxonomic information of the OTUs in panel (b) is in Supp. Tab 6 ([Shuangjie et al., 2022](#)). In panel (b), a group of OTUs 4, 31, 37, 39, 44, 56, 93 and 96 that are positively correlated with each other, are taxa that are found to indicate dysbiotic microbiota from gastrointestinal diseases. For example, OTUs 31 and 39 that belong to family *Erysipelotrichaceae* are observed to

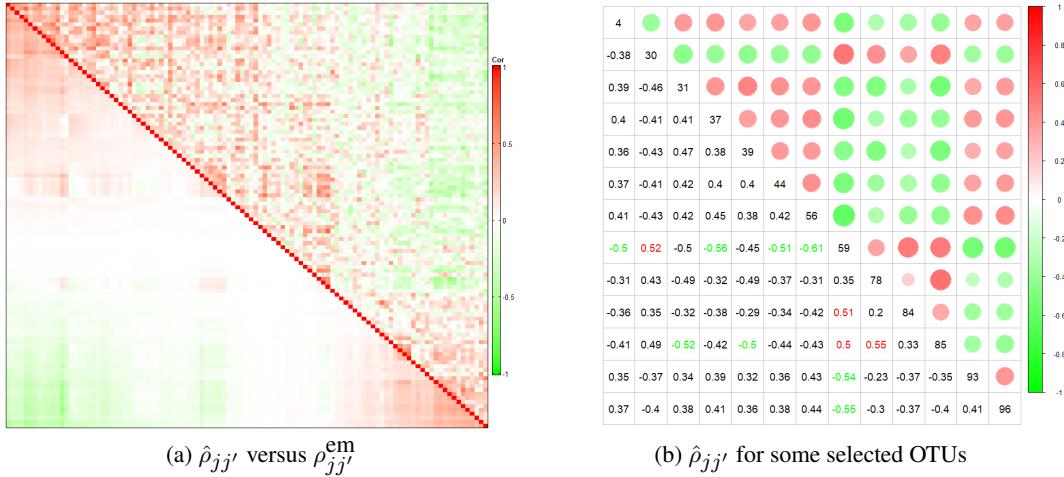


FIG 12. [Human Gut Microbiome Data]: Posterior marginal correlation estimates $\hat{\rho}_{jj'}$ (lower left triangle) and empirical correlation estimates $\rho_{jj'}^{\text{em}}$ (upper right triangle) are shown in panel (a). Panel (b) illustrates the OTUs having $|\hat{\rho}_{jj'}| > 0.5$ for any $j' \neq j$.

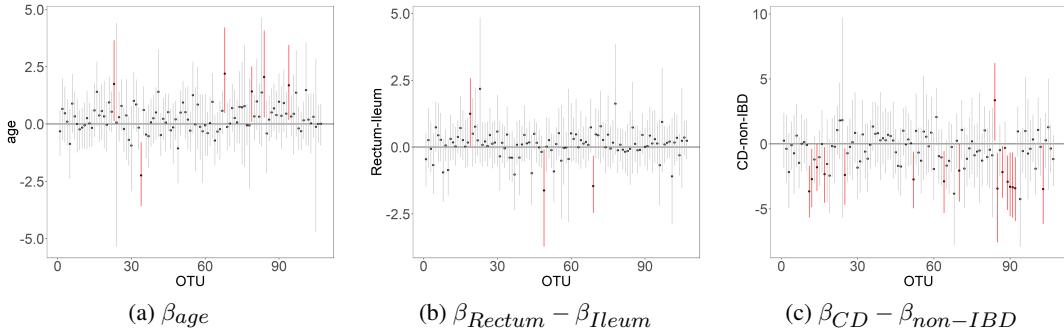


FIG 13. [Human Gut Microbiome Data] Posterior inference of regression coefficients β_{age} , $\beta_{\text{Rectum}} - \beta_{\text{Ileum}}$, and $\beta_{CD} - \beta_{\text{non-IBD}}$, where the posterior mean estimates are denoted by dots, and the 95% credible estimates with vertical lines. The intervals that do not contain zero are marked.

be related to gastrointestinal inflammatory disorders (Kaakoush, 2015). And some species in *Escherichia* (OTU 93) (e.g., E. Coli (Mirsepasi-Lauridsen *et al.*, 2019)) and *Clostridium* (OTUs 31 and 96) (e.g., C. difficile (Nitzan *et al.*, 2013)) are known to be related to the development of IBD. Another group of OTUs that are positively associated with each other includes genera, *Bacteroides* (OTU 59), *Faecalibacterium* (OTU 30), *Lachnospiraceae* (OTU 84) and *Ruminococcaceae* (OTU 85). The group of those genera contains species that were found active in metabolic processes and can produce short-chain fatty acids (Parada Venegas *et al.*, 2019). These species might interact through exchanging metabolic products; for example, *Bacteroides thetaiotaomicron* and *Faecalibacterium prausnitzii* were found metabolically complementary, where the former is an acetate producer, and the latter is acetate consumer and butyrate producer (Wrzosek *et al.*, 2013). Furthermore, such metabolic functions might be part of a complex interplay between the microbiota and disease states. For example, butyrate is an anti-inflammation promoter, and the decrease of butyrate producers might also indicate dysbiotic gut microbiota (Andrade *et al.*, 2020). Interestingly, the OTUs in those two groups are negatively associated. The correlation patterns between the groups indicate how gut microbiota may shift from dysbiosis and may suggest further investigation through experiments. From taxonomic information in Supp. Tab 6 (Shuangjie *et al.*, 2022), the OTUs

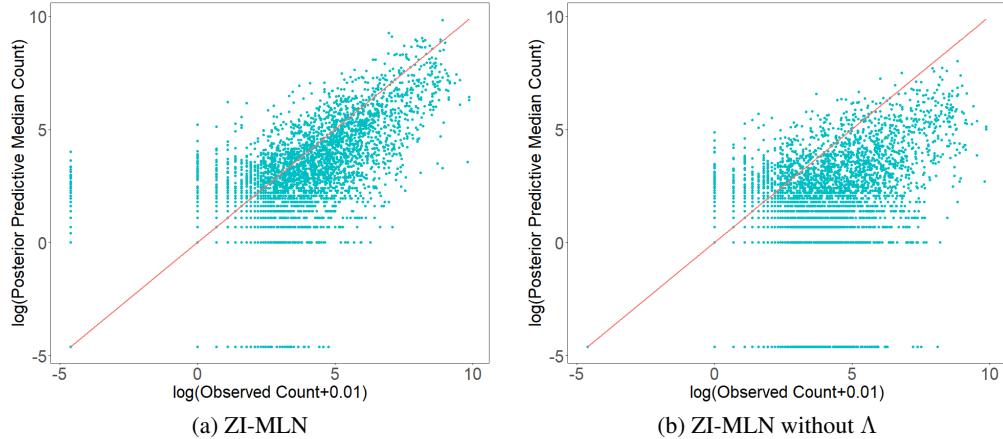


FIG 14. [Human Gut Microbiome Data: Comparison]: Panels (a) and (b) have scatter plots of observed $\log(y_{ij} + 0.01)$ versus $\log(\hat{y}_{ij}^{pred} + 0.01)$ under ZI-MLN and ZI-MLN without Λ , respectively.

in the groups belong to different families and orders, indicating that phylogenetically distant OTUs interact in gut microbiota.

Fig 13 and Supp. Fig 39(a)-(b) ([Shuangjie et al., 2022](#)) illustrate posterior mean estimates $\hat{\beta}_{jp}$ and $\hat{\kappa}_{jp}$ of the regression coefficients, respectively, with their 95% credible intervals for some selected covariates. Dots represent point estimates and vertical lines interval estimates. In the figures, β_{jp} and κ_{jp} that do not contain zero in their 95% credible interval are marked. In addition, Supp. Tabs 7 and 8 ([Shuangjie et al., 2022](#)) provide taxonomic information of the OTUs whose abundance or presence/absence is statistically associated with change in covariates. Overall, the covariate effects are statistically significant for a small number of OTUs. From panel (c), the effect of having condition CD compared to non-IBD $\beta_{CD} - \beta_{non-IBD}$ is statistically significant for 16 OTUs. The effect estimates are negative for those except for OTU 84, which implies that their abundance is lower for a subject with CD than for a subject with non-IBD. Also, among those, 14 OTUs belong to phylum *Firmicutes* and order *Clostridiales*. Significant decrease in abundance of phylum *Firmicutes* (*Clostridium leptum* and *Clostridium coccoides* groups) in active IBD subjects compared to that in non-IBD subjects is reported in [Sokol et al. \(2009\)](#), [Vester-Andersen et al. \(2019\)](#) and [Alam et al. \(2020\)](#). [Lloyd-Price et al. \(2019\)](#) also reported a statistically significant decrease in abundance of *Clostridium leptum* in active IBD subjects. We compare posterior predictive median estimates of OTU counts to the observed data in Fig 14(a) to access the model fit. The figure shows that the model fits the data well.

For comparison, we applied the comparators, SparCC, SPIEC-EASI, CCLasso and Zi-LN to the gut microbiome data. Fig 15 illustrates $\hat{\rho}_{jj'}$ under the comparators. Also, additional comparators, ZI-MLN without Λ , metagenomeSeq and edgeR were applied. The first set of the comparators does not account for covariate effects, and the second set does not infer the dependence structure between OTUs. SPIEC-EASI yields a very sparse estimate, whereas the other comparators produce very dense estimates. Supp. Figs 39(c)-(d) and 40 ([Shuangjie et al., 2022](#)) illustrate posterior estimates of regression coefficients β_{jp} and κ_{jp} obtained by the second set of the comparators. While ZI-MLN without Λ yields similar estimates, the estimates under metagenomeSeq and edgeR are greatly different from those under ZI-MLN. Specifically, under metagenomeSeq, the effects of covariate *age* are positive and statistically significant for most OTUs. A similar pattern is also observed from edgeR. For ZI-MLN without Λ , we further examine posterior predictive distributions of OTU counts (shown in

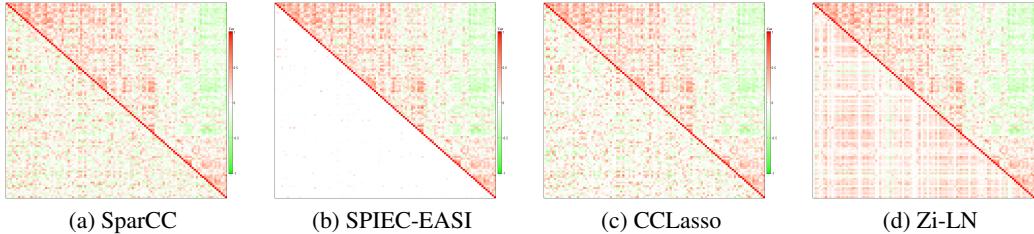


FIG 15. [Human Gut Microbiome Data: Comparison] Correlation estimates $\hat{\rho}_{jj'}$ by SparCC, SPIEC-EASI, CCLasso and Zi-LN (lower left triangle) and empirical correlation estimates $\rho_{jj'}^{em}$ (upper right triangle) are shown in panel (a)-(d), respectively.

Fig 14(b)). Compared to the fit under ZI-MLN, ZI-MLN without Λ yields a poor fit, especially for large counts. Supp. Fig 41 (Shuangjie *et al.*, 2022) compares mean abundance estimates under edgeR and metagenomeSeq to the observed counts and indicates poor model fit under those models.

5. Discussion. We have presented a Bayesian zero-inflated rounded log-normal kernel model to analyze multivariate count data with excess zeros. Different from most existing models, the model directly infers interrelationships between counts and produces reliable inference on microbial interaction with a small sample size. It offers a straightforward interpretation of microbial dependence structures. Furthermore, the model simultaneously incorporates covariates and accounts for excess zeros. The simulations showed that the developed model compares very favorably in parameter estimation and model fit to a model that ignores between-OTUs' dependence structure and some popular alternatives that do not model covariate effects and/or dependence structure.

ZI-MLN can be further extended to accommodate more complex data structures. Specifically, Lloyd-Price *et al.* (2019) collected multi-omics data to obtain a comprehensive understanding of the IBD microbial ecosystem. Multi-omic measurements from the same subject may be interrelated, and joint analysis of bacterial sequencing data with other types of sequencing data such as viral sequencing data can be useful. In general, latent factor models provide a convenient way to model complex interrelationship structures in multivariate data and can be extended to accommodate multiple coupled observation matrices, e.g., a group factor model (Zhao *et al.*, 2016). In that vein, our ZI-MLN can be extended to jointly analyze multiple correlated count matrices from a multi-omics study using an approach of a group factor model. Another possible extension is to incorporate phylogenetic information into the model. Investigating potential interactions between phylogenetically related microbes is biologically interesting, e.g., see Faust *et al.* (2012); Connor, Barberán and Clauzel (2017); Kamneva (2017). Similar to Lo and Marculescu (2018), phylogenetic information can be utilized in building a prior model of Σ .

Acknowledgements. This work was supported by NIH: DP2 GM123457-01 to IAC (Irene Chen) and NSF grant DMS-1662427 (Juhee Lee).

SUPPLEMENTARY MATERIAL

Supplementary FileSupplementary material: Additional Results

The Supplementary Material consists of four sections. In Supp. § 1, we provide details of the posterior computation. Supp. § 2 has additional results from the simulation studies. Supp. § 3 contains additional results of the real data analyses.

Supplementary material: ZI-MLN Code

The supplementary source code contains the implementation of the models described in the paper. It is also available from <https://github.com/Zsj950708/ZI-MLN>.

REFERENCES

- AGARWAL, D. K., GELFAND, A. E. and CITRON-POUSTY, S. (2002). Zero-inflated models with application to spatial count data. *Environmental and Ecological statistics* **9** 341–355.
- ALAM, M. T., AMOS, G. C., MURPHY, A. R., MURCH, S., WELLINGTON, E. M. and ARASARADNAM, R. P. (2020). Microbial imbalance in inflammatory bowel disease patients at different taxonomic levels. *Gut pathogens* **12** 1–8.
- ANDRADE, J. C., ALMEIDA, D., DOMINGOS, M., SEABRA, C. L., MACHADO, D., FREITAS, A. C. and GOMES, A. M. (2020). Commensal obligate anaerobic bacteria and health: production, storage, and delivery strategies. *Frontiers in Bioengineering and Biotechnology* **8** 550.
- BASHAN, A., GIBSON, T. E., FRIEDMAN, J., CAREY, V. J., WEISS, S. T., HOHMANN, E. L. and LIU, Y.-Y. (2016). Universality of human microbial dynamics. *Nature* **534** 259–262.
- BHATTACHARYA, A. and DUNSON, D. B. (2011). Sparse Bayesian infinite factor models. *Biometrika* 291–306.
- BHATTACHARYA, A., PATI, D., PILLAI, N. S. and DUNSON, D. B. (2015). Dirichlet-Laplace priors for optimal shrinkage. *Journal of the American Statistical Association* **110** 1479–1490.
- BIEN, J. and TIBSHIRANI, R. J. (2011). Sparse estimation of a covariance matrix. *Biometrika* **98** 807–820.
- CAI, T., LIU, W. and LUO, X. (2011). A constrained L1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association* **106** 594–607.
- CAI, T., MA, Z. and WU, Y. (2015). Optimal estimation and rank detection for sparse spiked covariance matrices. *Probability theory and related fields* **161** 781–815.
- CAI, T. T., REN, Z. and ZHOU, H. H. (2016). Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation. *Electronic Journal of Statistics* **10** 1–59.
- CAI, Z., ZHU, T., LIU, F., ZHUANG, Z. and ZHAO, L. (2021). Co-pathogens in Periodontitis and Inflammatory Bowel Disease. *Frontiers in Medicine* **8**.
- CANALE, A. and DUNSON, D. B. (2011). Bayesian kernel mixtures for counts. *Journal of the American Statistical Association* **106** 1528–1539.
- CHATTOPADHYAY, S., ARNOLD, J. D., MALAYIL, L., HITTLE, L., MONGODIN, E. F., MARATHE, K. S., GOMEZ-LOBO, V. and SAPKOTA, A. R. (2021). Potential role of the skin and gut microbiota in premenarchal vulvar lichen sclerosus: A pilot case-control study. *PloS one* **16** e0245243.
- CONNOR, N., BARBERÁN, A. and CLAUSET, A. (2017). Using null models to infer microbial co-occurrence networks. *PloS one* **12** e0176751.
- FANG, H., HUANG, C., ZHAO, H. and DENG, M. (2015). CCLasso: correlation inference for compositional data through Lasso. *Bioinformatics* **31** 3172–3180.
- FAUST, K., SATHIRAPONGSASUTI, J. F., IZARD, J., SEGATA, N., GEVERS, D., RAES, J. and HUTTENHOWER, C. (2012). Microbial co-occurrence relationships in the human microbiome. *PLoS computational biology* **8** e1002606.
- FRIEDMAN, J. and ALM, E. J. (2012). Inferring correlation networks from genomic survey data.
- FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441.
- GAO, C. and ZHOU, H. H. (2015). Rate-optimal posterior contraction for sparse PCA. *The Annals of Statistics* **43** 785–818.
- GRANTHAM, N. S., GUAN, Y., REICH, B. J., BORER, E. T. and GROSS, K. (2020). Mimix: A bayesian mixed-effects model for microbiome data from designed experiments. *Journal of the American Statistical Association* **115** 599–609.
- JIANG, S., XIAO, G., KOH, A. Y., KIM, J., LI, Q. and ZHAN, X. (2021). A Bayesian zero-inflated negative binomial regression model for the integrative analysis of microbiome data. *Biostatistics* **22** 522–540.
- JOVEL, J., PATTERSON, J., WANG, W., HOTTE, N., O’KEEFE, S., MITCHEL, T., PERRY, T., KAO, D., MASSON, A. L., MADSEN, K. L. et al. (2016). Characterization of the gut microbiome using 16S or shotgun metagenomics. *Frontiers in microbiology* **7** 459.
- KAAKOUSH, N. O. (2015). Insights into the Role of Erysipelotrichaceae in the Human Host. *Frontiers in Cellular and Infection Microbiology* **5** 84.
- KAMNEVA, O. K. (2017). Genome composition and phylogeny of microbes predict their co-occurrence in the environment. *PLoS computational biology* **13** e1005366.
- 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 computational biology* **11** e1004226.

- LEE, J. and SISON-MANGUS, M. (2018). A Bayesian semiparametric regression model for joint analysis of microbiome data. *Frontiers in microbiology* **9** 522.
- 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. *Statistical Analysis and Data Mining: The ASA Data Science Journal* **10** 393–409.
- LLOYD-PRICE, J., ARZE, C., ANANTHAKRISHNAN, A. N., SCHIRMER, M., AVILA-PACHECO, J., POON, T. W., ANDREWS, E., AJAMI, N. J., BONHAM, K. S., BRISLAWN, C. J. et al. (2019). Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* **569** 655–662.
- LO, C. and MARCULESCU, R. (2018). PGLasso: Microbial Community Detection through Phylogenetic Graphical Lasso. <https://arxiv.org/abs/1807.08039v1>.
- MA, S., REN, B., MALLICK, H., MOON, Y. S., SCHWAGER, E., MAHARJAN, S., TICKLE, T. L., LU, Y., CARMODY, R. N., FRANZOSA, E. A. et al. (2021). A statistical model for describing and simulating microbial community profiles. *PLoS computational biology* **17** e1008913.
- MAO, J., CHEN, Y. and MA, L. (2020). Bayesian graphical compositional regression for microbiome data. *Journal of the American Statistical Association* **115** 610–624.
- MIRSEPAEI-LAURIDSEN, H. C., VALLANCE, B. A., KROGFELT, K. A. and PETERSEN, A. M. (2019). Escherichia coli pathobionts associated with inflammatory bowel disease. *Clinical microbiology reviews* **32** e00060–18.
- NITZAN, O., ELIAS, M., CHAZAN, B., RAZ, R. and SALIBA, W. (2013). Clostridium difficile and inflammatory bowel disease: role in pathogenesis and implications in treatment. *World journal of gastroenterology: WJG* **19** 7577.
- PARADA VENEGAS, D., DE LA FUENTE, M. K., LANDSKRON, G., GONZÁLEZ, M. J., QUERA, R., DIJKSTRA, G., HARMSEN, H. J., FABER, K. N. and HERMOSO, M. A. (2019). Short chain fatty acids (SCFAs)-mediated gut epithelial and immune regulation and its relevance for inflammatory bowel diseases. *Frontiers in immunology* **277**.
- PARK, J.-U., OH, B., LEE, J. P., CHOI, M.-H., LEE, M.-J. and KIM, B.-S. (2019). Influence of microbiota on diabetic foot wound in comparison with adjacent normal skin based on the clinical features. *BioMed research international* **2019**.
- PATI, D., BHATTACHARYA, A., PILLAI, N. S., DUNSON, D. et al. (2014). Posterior contraction in sparse Bayesian factor models for massive covariance matrices. *Annals of Statistics* **42** 1102–1130.
- PAULSON, J. N., STINE, O. C., BRAVO, H. C. and POP, M. (2013). Differential abundance analysis for microbial marker-gene surveys. *Nature methods* **10** 1200–1202.
- PROST, V., GAZUT, S. and BRÜLS, T. (2021). A zero inflated log-normal model for inference of sparse microbial association networks. *PLoS Computational Biology* **17** e1009089.
- QIN, J., SHI, X., XU, J., YUAN, S., ZHENG, B., ZHANG, E., HUANG, G., LI, G., JIANG, G., GAO, S. et al. (2021). Characterization of the genitourinary microbiome of 1,165 middle-aged and elderly healthy individuals. *Frontiers in Microbiology* **12**.
- REN, B., BACALLADO, S., FAVARO, S., VATANEN, T., HUTTENHOWER, C. and TRIPPA, L. (2017). Bayesian nonparametric mixed effects models in microbiome data analysis. *arXiv preprint arXiv:1711.01241*.
- ROBINSON, M. D., MCCARTHY, D. J. and SMYTH, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26** 139–140.
- SCHWAGER, E., MALLICK, H., VENTZ, S. and HUTTENHOWER, C. (2017). A Bayesian method for detecting pairwise associations in compositional data. *PLoS computational biology* **13** e1005852.
- SHUANGJIE, Z., YUNING, S., IRENE A., C. and JUHEE, L. (2022). Supplement to "Bayesian Modeling of Interaction between Features in Sparse Multivariate Count Data with Application to Microbiome Study".
- SHULER, K., VERBANIC, S., CHEN, I. A. and LEE, J. (2021). A Bayesian nonparametric analysis for zero-inflated multivariate count data with application to microbiome study. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*.
- SOKOL, H., SEKSIK, P., FURET, J., FIRMEsse, O., NION-LARMURIER, I., BEAUGERIE, L., COSNES, J., COR-THIER, G., MARTEAU, P. and DORÉ, J. (2009). Low counts of *Faecalibacterium prausnitzii* in colitis microbiota. *Inflammatory bowel diseases* **15** 1183–1189.
- TANG, Z.-Z. and CHEN, G. (2019). Zero-inflated generalized Dirichlet multinomial regression model for microbiome compositional data analysis. *Biostatistics* **20** 698–713.
- VERBANIC, S., SHEN, Y., LEE, J., DEACON, J. M. and CHEN, I. A. (2020). Microbial predictors of healing and short-term effect of debridement on the microbiome of chronic wounds. *NPJ biofilms and microbiomes* **6** 1–11.
- VESTER-ANDERSEN, M., MIRSEPAEI-LAURIDSEN, H., PROSBERG, M., MORTENSEN, C., TRÄGER, C., SKOVSEN, K., THORKILGAARD, T., NØJGAARD, C., VIND, I., KROGFELT, K. A. et al. (2019). Increased abundance of proteobacteria in aggressive Crohn's disease seven years after diagnosis. *Scientific reports* **9** 1–10.

- 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** 1–12.
- WANG, Z., MAO, J. and MA, L. (2021). Logistic-tree normal model for microbiome compositions. *arXiv preprint arXiv:2106.15051*.
- WANG, T. and ZHAO, H. (2017). A Dirichlet-tree multinomial regression model for associating dietary nutrients with gut microorganisms. *Biometrics* **73** 792–801.
- WRZOSEK, L., MIQUEL, S., NOORDINE, M.-L., BOUET, S., CHEVALIER-CURT, M. J., ROBERT, V., PHILIPPE, C., BRIDONNEAU, C., CHERBUY, C., ROBBE-MASSELOT, C. et al. (2013). Bacteroides thetaiotomicron and Faecalibacterium prausnitzii influence the production of mucus glycans and the development of goblet cells in the colonic epithelium of a gnotobiotic model rodent. *BMC biology* **11** 1–13.
- XIA, F., CHEN, J., FUNG, W. K. and LI, H. (2013). A logistic normal multinomial regression model for microbiome compositional data analysis. *Biometrics* **69** 1053–1063.
- XIAOMING, W., JING, L., YUCHEN, P., HUILI, L., MIAO, Z. and JING, S. (2021). Characteristics of the vaginal microbiomes in prepubertal girls with and without vulvovaginitis. *European Journal of Clinical Microbiology & Infectious Diseases* **40** 1253–1261.
- XIE, F., XU, Y., PRIEBE, C. E. and CAPE, J. (2018). Bayesian estimation of sparse spiked covariance matrices in high dimensions. *arXiv preprint arXiv:1808.07433*.
- ZHANG, X., MALLICK, H., TANG, Z., ZHANG, L., CUI, X., BENSON, A. K. and YI, N. (2017). Negative binomial mixed models for analyzing microbiome count data. *BMC bioinformatics* **18** 1–10.
- ZHAO, S., GAO, C., MUKHERJEE, S. and ENGELHARDT, B. E. (2016). Bayesian group factor analysis with structured sparsity. *The Journal of Machine Learning Research* **17** 6868–6914.

**SUPPLEMENTARY FOR BAYESIAN MODELING OF INTERACTION
BETWEEN FEATURES IN SPARSE MULTIVARIATE COUNT DATA WITH
APPLICATION TO MICROBIOME STUDY**

BY SHUANGJIE ZHANG^{1,*}, YUNING SHEN² IRENE A. CHEN² AND JUHEE LEE¹

¹*Department of Statistics, University of California Santa Cruz, *szhan209@ucsc.edu; juheelee@soe.ucsc.edu*

²*Department of Chemical and Biomolecular Engineering, University of California Los Angeles, yshen@chem.ucsbg.edu; ireneachen@ucla.edu*

1. Details of Posterior Computation. We use Markov chain Monte Carlo (MCMC) to draw samples of the random parameters from their posterior distribution. Recall that $Y_{ij} \in \mathbb{N}^0$, $i = 1, \dots, N$ and $j = 1, \dots, J$ denotes the count of OTU j in sample i , and the model assumes $Y_{ij} = \lfloor Y_{ij}^* \rfloor$ with $Y_{ij}^* \in \mathbb{R}^+$. The distribution of $\mathbf{Y}_i^* = (Y_{i1}^*, \dots, Y_{iJ}^*)$ is specified in (1) of the main text. We also let \mathbf{X} represent a $N \times P$ covariate matrix whose rows have a P -dim covariate vector \mathbf{x}_i . The probit regression for the probabilities of an OTU being absent may have a different covariate vector $\tilde{\mathbf{x}}_i$ with P_κ even for the same set of covariates due to different parameterizations. We let $\tilde{\mathbf{X}}$ be a $N \times P_\kappa$ covariate matrix having $\tilde{\mathbf{x}}_i$ in rows and $\tilde{\mathbf{X}}_* = [1, \tilde{\mathbf{X}}]$ a $N \times (P_\kappa + 1)$ covariate matrix with the first column being a column of 1.

To facilitate updating the parameters related to zero inflation such as δ_{ij} and $\boldsymbol{\kappa}_j$, we introduce a continuous real valued latent variable $z_{ij} \sim N(\kappa_{j0} + \tilde{\mathbf{x}}_i' \boldsymbol{\kappa}_j, 1)$ and let $\delta_{ij} = 1$ if $z_{ij} < 0$ and otherwise $\delta_{ij} = 0$. We then have $\epsilon_{ij} = \Pr(\delta_{ij} = 1) = \Pr(z_{ij} < 0 | \boldsymbol{\kappa}_j) = \Phi(\kappa_{j0} + \tilde{\mathbf{x}}_i' \boldsymbol{\kappa}_j)$. The MCMC steps of updating parameters $\boldsymbol{\kappa}_j$ and latent variables δ_{ij} and z_{ij} can be summarized as below;

- $\boldsymbol{\kappa}_j$: The full conditional distribution of $\boldsymbol{\kappa}_j$ is

$$\boldsymbol{\kappa}_j | \boldsymbol{\mu}_\kappa, \Sigma_\kappa \sim N_{(P_\kappa+1)}((\tilde{\mathbf{X}}_*' \tilde{\mathbf{X}}_* + u_\kappa \mathbf{I}_{P+1})^{-1} ((u_\kappa \mathbf{I}_{P+1})^{-1} \boldsymbol{\mu}_\kappa + \tilde{\mathbf{X}}' \mathbf{z}_j), (\tilde{\mathbf{X}}_*' \tilde{\mathbf{X}}_* + u_\kappa \mathbf{I}_{P+1})^{-1}).$$

Draw $\boldsymbol{\kappa}_j$ from its full conditional.

- z_{ij} : The full conditional distribution is

$$z_{ij} \sim \begin{cases} N_-(\kappa_{j0} + \tilde{\mathbf{x}}_i' \boldsymbol{\kappa}_j, 1) & \text{if } \delta_{ij} = 1, \\ N_+(\kappa_{j0} + \tilde{\mathbf{x}}_i' \boldsymbol{\kappa}_j, 1) & \text{if } \delta_{ij} = 0, \end{cases}$$

where N_+ and N_- represent normal distributions truncated below and above at zero, respectively.

- δ_{ij} : For (i, j) with $Y_{ij} = 0$, update δ_{ij} using the full conditional;

$$\Pr(\delta_{ij} = 1 | -) \propto \epsilon_{ij},$$

$$\Pr(\delta_{ij} = 0 | -) \propto (1 - \epsilon_{ij}) \int_{-\infty}^0 \phi(\tilde{y}_{ij}^* | r_i + \alpha_j + s_{g_i,j} + \lambda_j' \eta_i + \mathbf{x}_i' \beta_j, \sigma^2) d\tilde{y}_{ij}^*,$$

where $\epsilon_{ij} = \Phi(\kappa_{j0} + \tilde{\mathbf{x}}_i' \boldsymbol{\kappa}_j, 1)$ and $\phi(\cdot | a, b^2)$ represents the probability density function of the normal distribution with mean a and variance b^2 . If $Y_{ij} > 0$, $\delta_{ij} = 0$ with probability 1.

Updating of the random parameters related to μ_{ij} and Σ can be more convenient with latent continuous variables $\tilde{Y}_{ij}^* = \log(Y_{ij}^*)$ for (i, j) having $\delta_{ij} = 0$ imputed as follows;

$$\tilde{Y}_{ij}^* \sim N(r_i + \alpha_j + s_{g_i,j} + \lambda_j' \eta_i + \mathbf{x}_i' \beta_j, \sigma^2) \mathbf{1}(\log(y_{ij}) \leq \tilde{Y}_{ij}^* < \log(y_{ij})),$$

that is, a truncated normal distribution, where the support is determined by the observed count y_{ij} . Given \tilde{Y}_{ij}^* , the full conditionals of the parameters in Σ except for ϕ_j have a standard form, and the Gibbs sampler can be used to update τ_k , λ_{jk} and σ^2 . Specifically, we re-write the Laplace distribution as a normal scale mixture to facilitate the step of updating λ_{jk} from its full conditional; $\lambda_{jk} | \zeta_{jk}, \phi_j, \tau_k \stackrel{\text{indep}}{\sim} N(0, \zeta_{jk}\phi_j^2\tau_k^2)$ and $\zeta_{jk} \stackrel{iid}{\sim} \text{Exp}(1/2)$. Then λ_{jk} can be easily obtained through a data augmented Gibbs step. The full conditional distribution of ζ_{jk}^{-1} and τ_k can be sampled from the inverse Gaussian and generalized inverse Gaussian sampling distribution (Park and Casella, 2008). We update ϕ_j using a Metropolis-Hastings step. We let $\phi_j^* \stackrel{iid}{\sim} \text{Ga}(a_\phi, 1)$ and have $\phi = (\phi_1, \dots, \phi_J) \sim \text{Dir}(a_\phi, \dots, a_\phi)$ with $\phi_j = \phi_j^*/\sum_{j'} \phi_{j'}^*$. The full conditional of ϕ is given by

$$\begin{aligned} p(\phi | -) &\propto p(\boldsymbol{\lambda} | \boldsymbol{\tau}, \phi, \boldsymbol{\zeta}) p(\phi) \\ &\propto \prod_{j=1}^J \prod_{k=1}^K N(\lambda_{jk} | 0, \zeta_{jk}\phi_j^2\tau_k^2) \prod_{j=1}^J \text{Ga}(\phi_j^* | a_\phi, 1). \end{aligned}$$

In order to explore the posterior distribution of ϕ_j efficiently, the adaptive MH algorithm (Haario, Saksman and Tamminen, 2001) is used. We adjust the MH step size according to the acceptance ratio, and the convergence rate is accelerated.

Recall that we have parameters, r_i , α_j , β_j , $s_{g_i,j}$ and u_s^2 , for μ_{ij} in (8) of the main text, and parameters, ψ^χ , ω_l^χ and ξ_l^χ , $\chi \in \{\alpha, r\}$, in (9) of the main text. The full conditional distributions of the parameters β_j , $s_{g_i,j}$ and u_s^2 have a standard distribution, and their samples are easily drawn through a usual Bayesian Normal-Gamma model update. Size factors r_i and α_j have a mixture of mixtures as their prior. To facilitate computation, we introduce a pair of auxiliary variables for each $\chi \in \{\alpha, r\}$ that specifies the mixture component from which each particular χ is, i.e., (S_{i1}^r, S_{i2}^r) for r_i , where $S_{i1}^r \in \{1, \dots, L^r\}$ and $S_{i2}^r \in \{0, 1\}$, and $(S_{j1}^\alpha, S_{j2}^\alpha)$ for α_j , where $S_{j1}^\alpha \in \{1, \dots, L^\alpha\}$ and $S_{j2}^\alpha \in \{0, 1\}$. We then assume $P(S_{i1}^r = l) = \psi_l^r$ and $P(S_{i2}^r = 0 | S_{i1}^r = l) = \omega_l^r$, and similarly, assume $P(S_{j1}^\alpha = l) = \psi_l^\alpha$ and $P(S_{j2}^\alpha = 0 | S_{j1}^\alpha = l) = \omega_l^\alpha$. The conditional prior distributions of r_i and α_j are

$$\begin{aligned} r_i | \psi^r, \omega^r, \xi^r, S_{i1}^r = l, S_{i2}^r &\sim \begin{cases} N(\xi_l^r, u_r^2) & \text{if } S_{i2}^r = 0, \\ N\left(\frac{v_r - \omega_l^r \xi_l^r}{1 - \omega_l^r}, u_r^2\right) & \text{if } S_{i2}^r = 1, \end{cases} \\ \alpha_j | \psi^\alpha, \omega^\alpha, \xi^\alpha, S_{j1}^\alpha = l, S_{j2}^\alpha &\sim \begin{cases} N(\xi_l^\alpha, u_\alpha^2) & \text{if } S_{j2}^\alpha = 0, \\ N\left(\frac{v_\alpha - \omega_l^\alpha \xi_l^\alpha}{1 - \omega_l^\alpha}, u_\alpha^2\right) & \text{if } S_{j2}^\alpha = 1, \end{cases} \end{aligned}$$

Conditional on those indicators, ψ^χ can be drawn through a traditional Multinomial-Dirichlet model update and ω_l^χ through Beta-Binomial update. Also, given the indicators, the full conditional distributions of r_i , α_j and ξ_l^χ , $\chi \in \{r, \alpha\}$ have a Gaussian distribution.

2. Additional Simulation Studies.

2.1. Additional Results of Simulation 1. In this subsection, we present additional results from Simulation 1 in § 3.1 of the main text. We examined the convergence of the MCMC simulation using trace plots of the log-likelihood. The model was run under different initializations and random seeds. Traceplots of the log-likelihood in Fig 1 suggest that the model converged to a similar state under these different initializations. The figure provides practical evidence of the model's convergence. Fig 2 illustrates empirical estimates of the marginal correlations $\rho_{jj'}^{\text{em}}$ of the logarithm transformed counts, $\log(y_{ij} + 0.01)$ after normalization with sample size factors estimated by CSS. The true values $\rho_{jj'}^{\text{tr}}$ of the correlations are shown

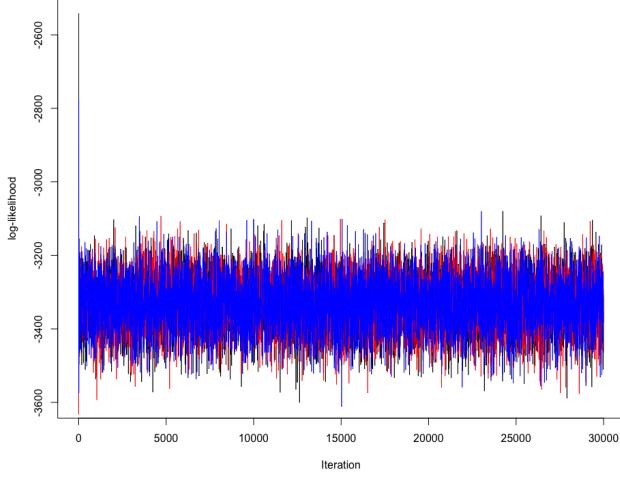


FIG 1. [Simulation 1] Traceplots of log likelihood under three different initializations are presented in black, red and blue, respectively.

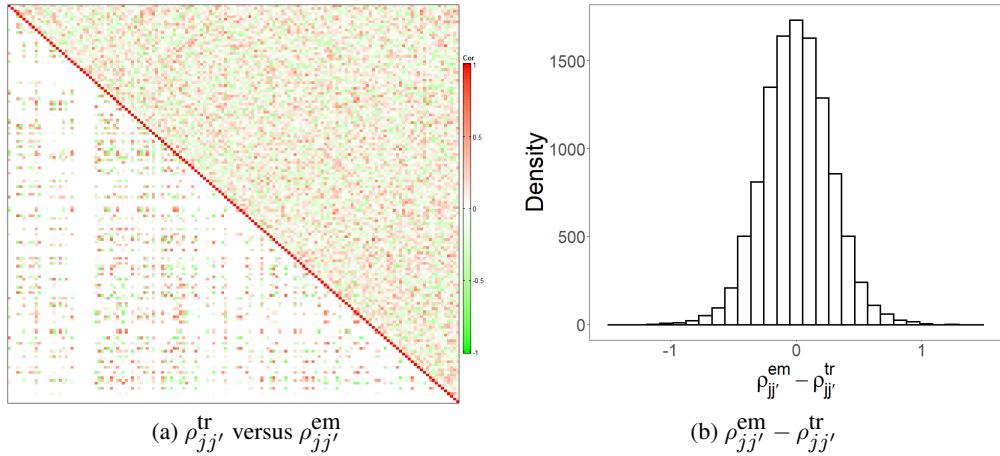


FIG 2. [Simulation 1] The upper right and lower left triangles of the heatmap in panel (a) illustrates empirical correlation estimates $\rho_{jj'}^{em}$ of $\log(Y_{ij} + 0.01)$ scaled with CSS and the true values of the correlations $\rho_{jj'}^{tr}$, respectively. Panel (b) has a histogram of differences between $\rho_{jj'}^{em}$ and $\rho_{jj'}^{tr}$.

in the lower left triangle of the heatmap in panel (a). In panel (b), a histogram of the differences $\rho_{jj'}^{em} - \rho_{jj'}^{tr}$ is shown. Fig 3(a) shows posterior mean estimates of the baseline abundance $r_i + \alpha_j$ of OTU j in sample i compared to the simulation truth $r_i^{tr} + \alpha_j^{tr}$. The figure indicates that the mean abundances $r_i + \alpha_j$ are identifiable although r_i and α_j are not individually identifiable, and our model provides good estimates of the mean abundance. Furthermore, it provides a good basis for the estimation of parameters of our main interest such as Σ . We also examined the performance of estimating probabilities ϵ_{ij} of OTUs being absent in samples. Under the setup without covariates, $\epsilon_{ij} = \Phi_1(\kappa_{j0} | 0, 1)$ is identical for all i , i.e., $\epsilon_{ij} = \epsilon_j$ for all i . Fig 3(b) shows that ϵ_j 's are well estimated even with a small sample size $N = 20$.

2.2. Additional Results of Simulation 2 . In this subsection, we include additional results from Simulation 2, described in § 3.2 of the main text. We ran the model on the dataset

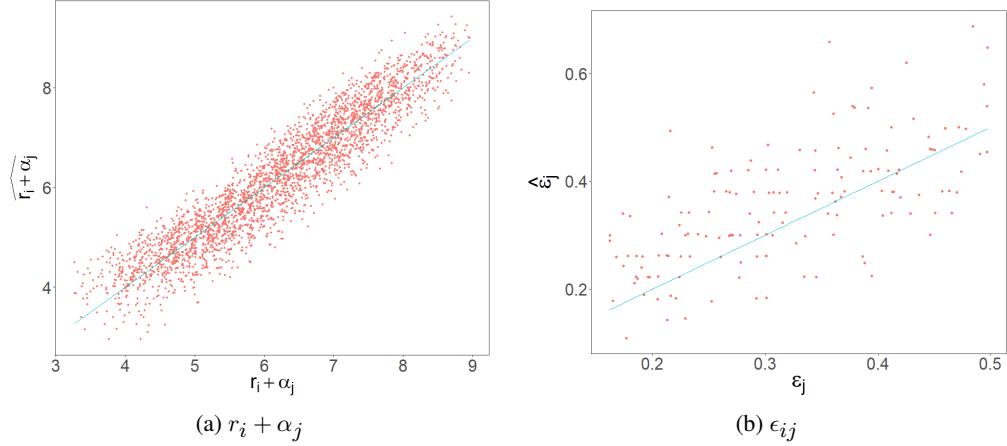


FIG 3. [Simulation 1] In panel (a), posterior mean estimates of the mean abundance $r_i + \alpha_j$ are plotted against the simulation truth $r_i^{tr} + \alpha_j^{tr}$. Panel (b) has a histogram of the differences in posterior mean estimates $\hat{\epsilon}_{ij}$ of probabilities of an OTU being absent and their true values ϵ_{ij}^{tr} .

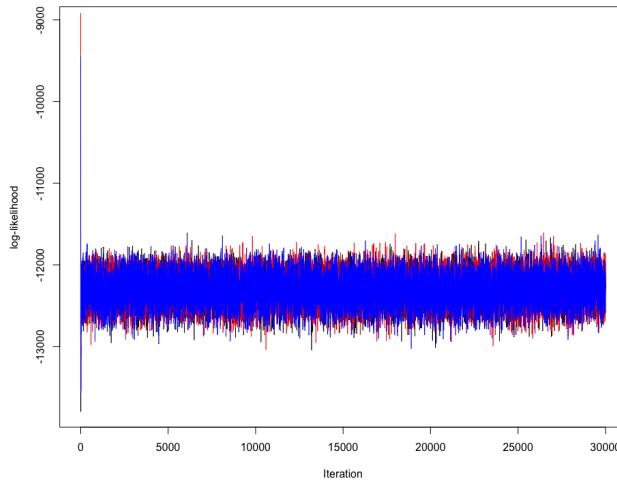


FIG 4. [Simulation 2] Traceplots of log likelihood under three different initializations are presented in black, red and blue, respectively.

with different initializations and random seeds for the MCMC chain. Traceplots of the log-likelihood under the different random seeds and initializations are shown in Fig 4. The figure shows the MCMC converges to similar log-likelihood ranges under these different specifications, and we did not find evidence suggesting the Markov chain failed to converge. Fig 5 compares empirical estimates $\rho_{jj'}^{\text{em}}$ of the marginal correlations to the true values $\rho_{jj'}^{\text{tr}}$ of the correlations. $\rho_{jj'}^{\text{em}}$ and $\rho_{jj'}^{\text{tr}}$ are shown in the upper right and lower left triangles of the heatmap in panel (a), respectively. A histogram of differences $\rho_{jj'}^{\text{em}} - \rho_{jj'}^{\text{tr}}$ is in panel (b). Figs 6(a)-(c) compare posterior mean estimates $\hat{\kappa}_{jp}$ of probit regression coefficients on ϵ_{ij} to their true values under ZI-MLN. In Fig 7, we examine the estimation of mean abundances $r_i + \alpha_j$ and the probabilities ϵ_{ij} of an OTU being absent in a sample. The figure shows that posterior mean estimates of $r_i + \alpha_j$ and of ϵ_{ij} are close to their simulation truth, and the model reasonably well recovers the simulation truth.

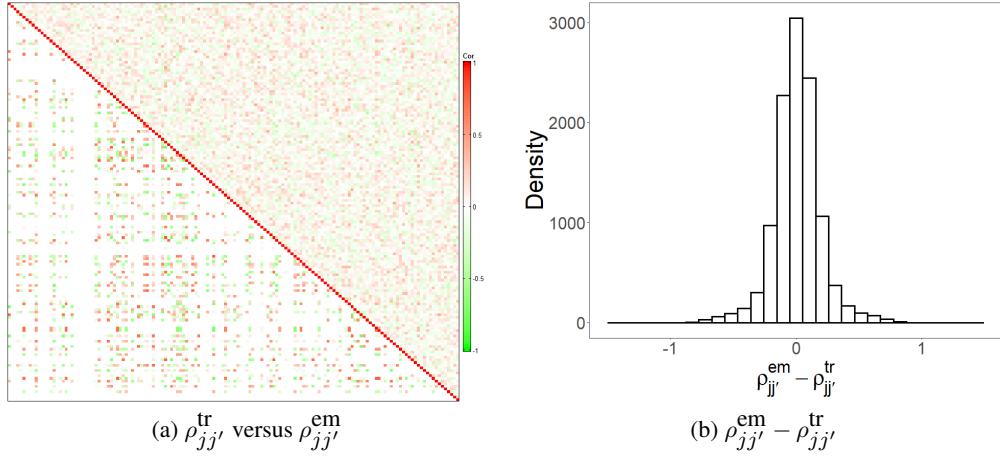


FIG 5. [Simulation 2] The upper right and lower left triangles of the heatmap in panel (a) illustrate empirical correlation estimates $\rho_{jj'}^{em}$ of $\log(Y_{ij} + 0.01)$ scaled with CSS and the true values of the correlations $\rho_{jj'}^{tr}$, respectively. Panel (b) has a histogram of differences between $\rho_{jj'}^{em}$ and $\rho_{jj'}^{tr}$.

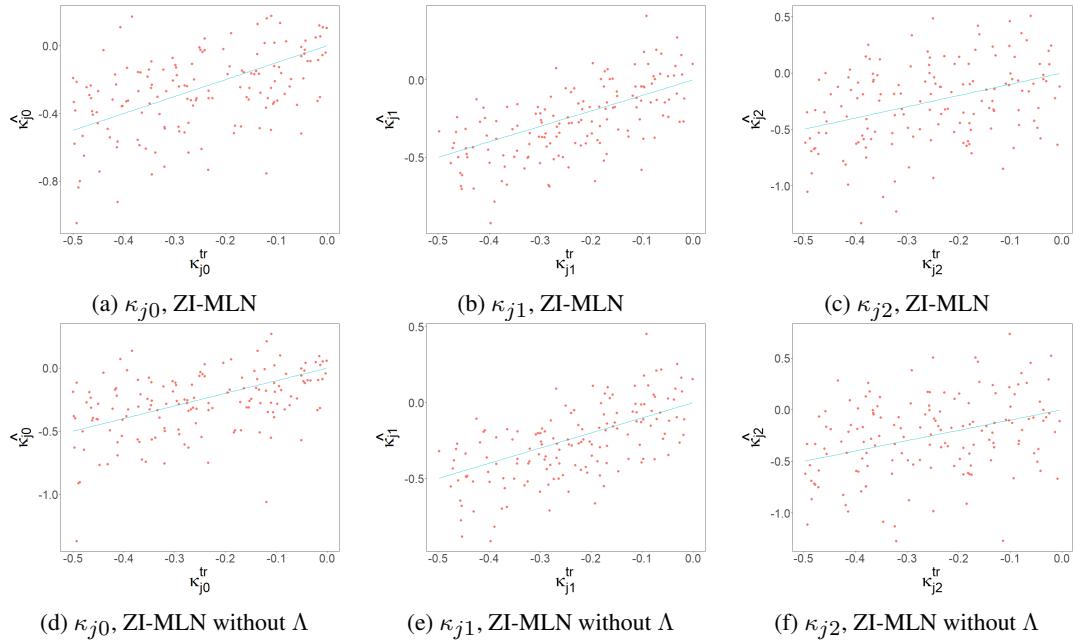


FIG 6. [Simulation 2] Posterior mean estimates $\hat{\kappa}_{jp}$ of coefficients on ϵ_{ij} are plotted against the truth for $p = 0, 1, 2$ in columns 1-3, respectively. The top and bottom rows are for ZI-MLN and ZI-MLN without Λ , respectively.

Fig 8 illustrates posterior mean estimates of regression coefficients β_{jp} under comparators, ZI-MLN without Λ , metagenomeSeq and edgeR. Figs 6(d)-(f) compare posterior mean estimates of κ_{jp} under ZI-MLN without Λ to their truth. Estimates of OTU mean abundances under metagenomeSeq and edgeR are compared to the observed counts in Fig 9.

2.3. Simulation 3. We performed an additional simulation study, Simulation 3, where OTU counts are generated from Poisson distributions with correlated means and examined the robustness of our ZI-MLN. We kept most of the setup of Simulation 2 the same for Simulation 3; we assumed the number of OTUs $J = 150$ and the number of subjects $M = 35$ assuming two samples from each subject (so the number of samples $N = 70$). We simulated

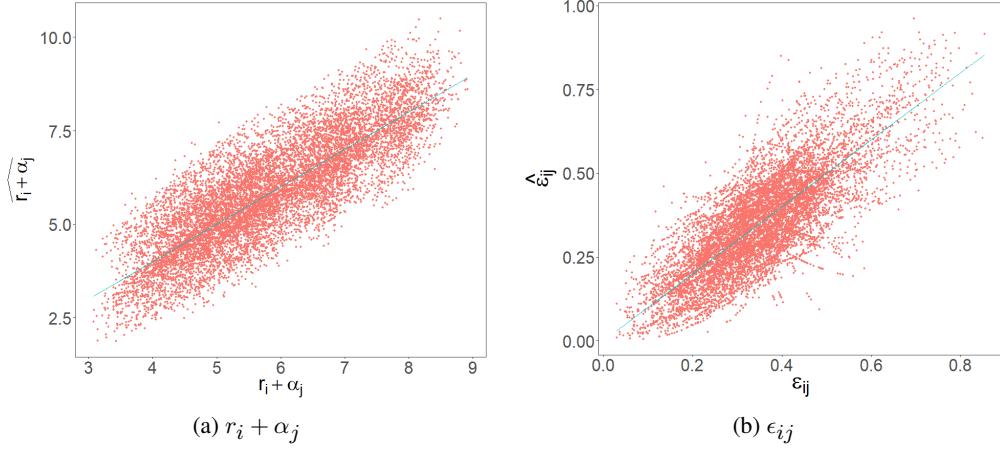


FIG 7. [Simulation 2] In panel (a), posterior mean estimates of the mean abundance $r_i + \alpha_j$ are plotted against the simulation truth $r_i^{tr} + \alpha_j^{tr}$. Panel (b) has a histogram of the differences in posterior mean estimates $\hat{\epsilon}_{ij}$ of probabilities of an OTU being absent and their true values ϵ_{ij}^{tr} .

TABLE 1
[Simulation 3: Comparison] RMSEs are computed for $\rho_{jj'}$, $j < j'$, δ_{ij} , μ_{ij} , $\beta_{j2} - \beta_{j1}$, β_{j3} and κ_{jp} under ZI-MLN and comparators.

Model	$\rho_{jj'}$	Model	δ_{ij}	μ_{ij}	$\beta_{j2} - \beta_{j1}$	β_{j3}	κ_{j0}	κ_{j1}	κ_{j2}
ZI-MLN	0.063	ZI-MLN	0.095	1.696	0.464	0.339	0.221	0.202	0.339
SparCC	0.178	ZI-MLN without Λ	0.115	1.731	0.601	0.397	0.242	0.222	0.350
SPIEC-EASI	0.158	MetagenomeSeq	0.113	1.913	1.245	0.729	-	-	-
CCLasso	0.154	EdgeR	-	3.400	0.952	0.595	-	-	-
Zi-LN	0.160								

(a) $\rho_{jj'}$ (b) δ_{ij} , μ_{ij} , $\beta_{j2} - \beta_{j1}$, β_{j3} and κ_{jp}

λ_{jk}^{tr} with $K^{\text{tr}} = 5$ assuming sparsity level $g = 0.8$ for joint sparsity, and set $\sigma^{2,\text{tr}} = 1$ and $v_s^{2,\text{tr}} = 1$. We included a pair of dummy variables $(x_{i1}, x_{i2}) \in \{(1, 0), (0, 1)\}$ to represent a binary covariate, and a continuous covariate, x_{i3} generated from $N(0, 1)$. We have $\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3})$ for the mean abundance μ_{ij} and $\tilde{\mathbf{x}}_i = (x_{i2}, x_{i3})$ for the probability ϵ_{ij} of an OTU being absent. We then simulated sample size factors r_i^{tr} , OTU size factors α_j^{tr} , subject-specific random effects $s_{g_i,j}^{\text{tr}}$, regression coefficients for mean abundances β_{jp}^{tr} and regression coefficients for zero inflation κ_{jp}^{tr} , the same as done in Simulation 2. We finally generated counts Y_{ij} for Poisson distributions as follows;

$$\delta_{ij}^{\text{tr}} | \epsilon_{ij}^{\text{tr}} \stackrel{\text{indep}}{\sim} \text{Ber}(\epsilon_{ij}^{\text{tr}}), \text{ where } \epsilon_{ij}^{\text{tr}} = \Phi_1(\kappa_{j0}^{\text{tr}} + \tilde{\mathbf{x}}_i' \boldsymbol{\kappa}_j^{\text{tr}} | 0, 1),$$

$$\tilde{\boldsymbol{\mu}}_i^{\text{tr}} \stackrel{iid}{\sim} N_J(r_i^{\text{tr}} \mathbf{1}_J + \boldsymbol{\alpha}^{\text{tr}} + \mathbf{s}_{g_i}^{\text{tr}} + \boldsymbol{\beta}^{\text{tr}} \mathbf{x}_i, \Sigma^{\text{tr}}),$$

$$\begin{cases} y_{ij} | \tilde{\mu}_{ij}^{\text{tr}} \stackrel{\text{indep}}{\sim} \text{Poi}(\exp(\tilde{\mu}_{ij}^{\text{tr}})) & \text{if } \delta_{ij}^{\text{tr}} = 0, \\ y_{ij} = 0 & \text{if } \delta_{ij}^{\text{tr}} = 1, \end{cases}$$

where $\Sigma^{\text{tr}} = \Lambda^{\text{tr}} (\Lambda^{\text{tr}})' + \sigma^{2,\text{tr}} \mathbf{I}_J$. Here, $\boldsymbol{\beta}^{\text{tr}}$ is the $J \times P$ matrix of the true β_{jp}^{tr} . To fit the model, we specified the fixed hyperparameter values similar to those in Simulation 2 and approxi-

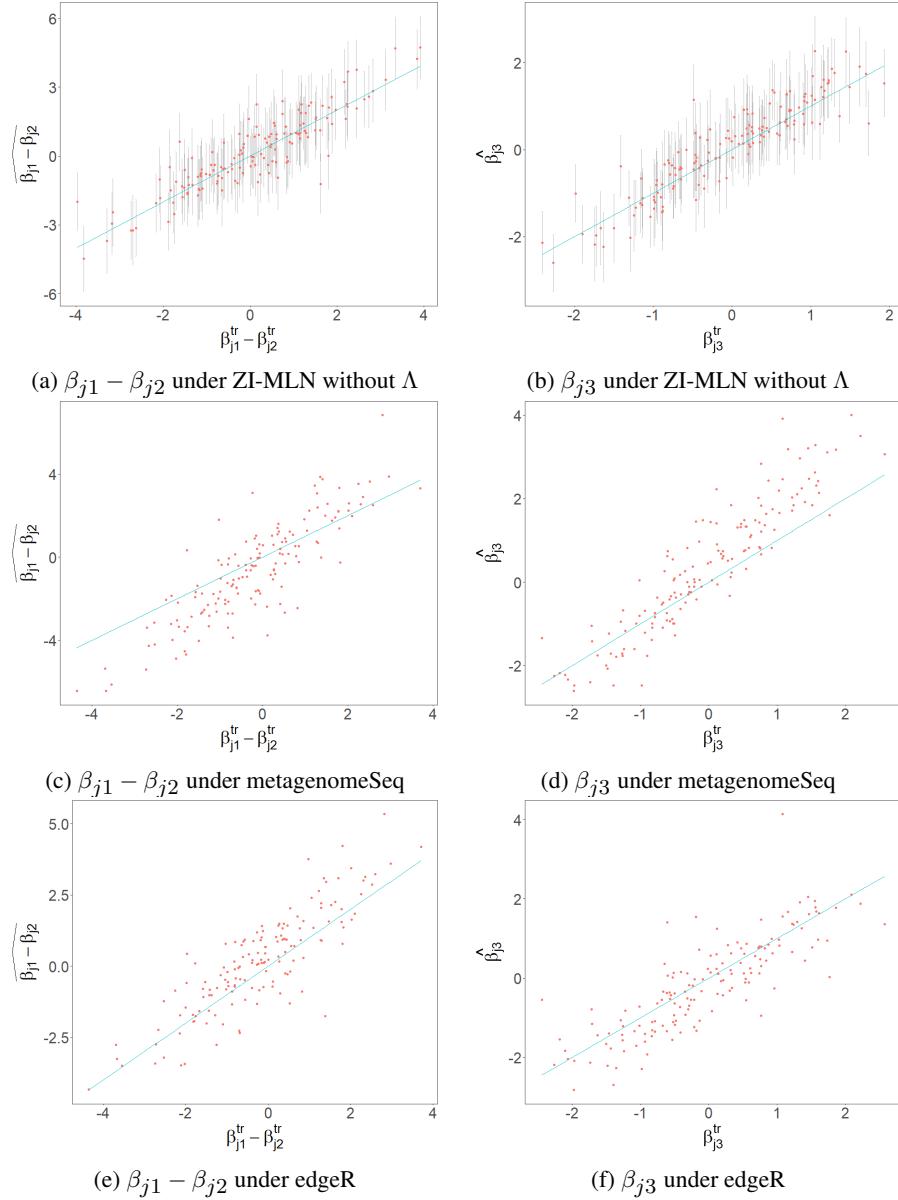


FIG 8. [Simulation 2: Comparison] Estimates $\widehat{\beta_{j1} - \beta_{j2}}$ and $\widehat{\beta_{j3}}$ of regression coefficients are compared to the truth, $\beta_{j1}^{tr} - \beta_{j2}^{tr}$ and β_{j3}^{tr} . The estimates in rows 1-3 are produced by ZI-MLN without Λ , metagenomeSeq and edgeR, respectively.

mated the posterior distribution using MCMC. The MCMC simulation was run for 30,000 iterations, discarding the first 15,000 iterations. Reasonable convergence was achieved and the chain mixed well from checking traceplots and auto-correlation plots. Fig 10 shows traceplots of the log-likelihood from MCMC chains under different random seeds and initializations. The MCMC converges to similar log-likelihood ranges, showing no evidence of poor mixing or convergence problem.

The results are shown in Figs 11-16. Fig 11 compares posterior mean estimates of the marginal correlation $\rho_{jj'}$ to the truth. The figure shows that the dependency structure is well recovered. Fig 12(a) shows a scatter plot of posterior mean estimates $r_i + \alpha_j$ of mean abundances compared to the truth $r_i^{tr} + \alpha_j^{tr}$. In panel (b) of the figure, posterior estimates of ϵ_{ij}

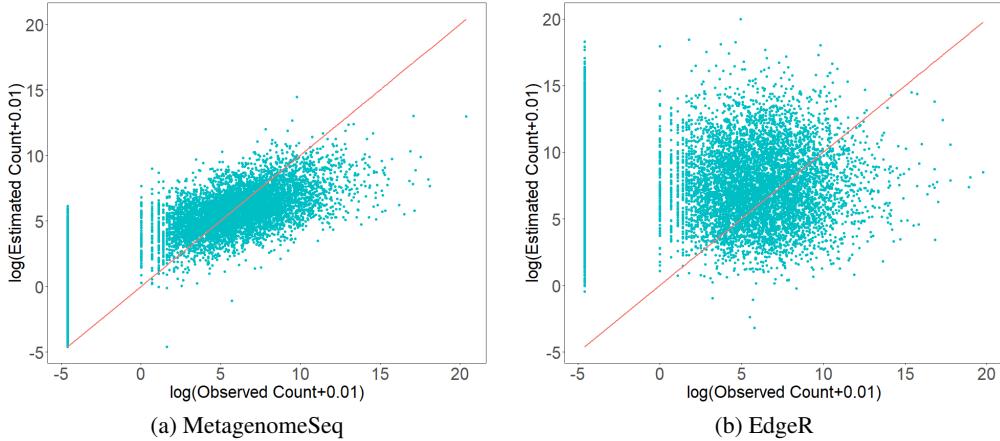


FIG 9. [Simulation 2: Comparison] Panels (a) and (b) present scatter plots of observed $\log(y_{ij} + 0.01)$ versus $\log(\hat{\mu}_{ij} + 0.01)$ by metagenomeSeq and edgerR. $\hat{\mu}_{ij}$ are estimated mean abundances of OTUs in samples.

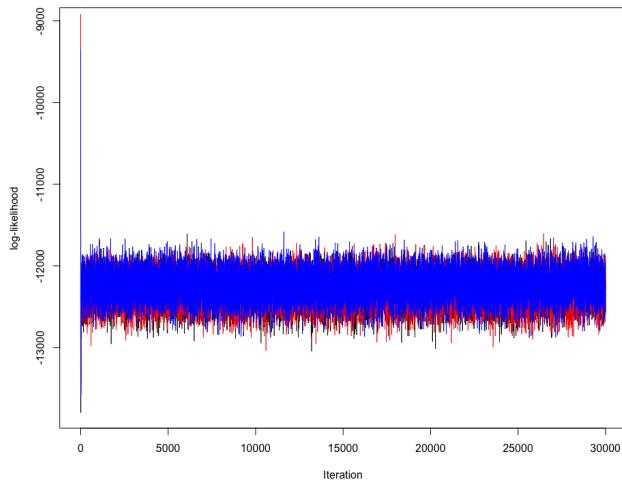


FIG 10. [Simulation 3] Traceplots of log likelihood under three different initializations are presented in black, red and blue, respectively.

are compared to the truth $\epsilon_{ij}^{\text{tr}}$. Posterior mean estimates of regression coefficients $\beta_{j1} - \beta_{j2}$ and β_{j3} are compared to their truth in Fig 13(a) and (b). Posterior mean estimates of probit regression coefficients κ_{jp} are compared to their truth in Fig 15(a)-(c). In Fig 16(a), posterior predictive median estimates of OTU counts are plotted against the observed counts for model checking. Overall, the underlying data generation structure is reasonably well approximated although the simulation truth is greatly different from the assumption that ZI-NNL makes, and the model provides a reasonable fit to the data.

We also applied our comparators, SparCC, SPIEC-EASI, CCLasso and Zi-LN, to the simulated dataset. Estimates $\hat{\rho}_{jj'}$ of the correlations under the comparators are compared to the truth $\rho_{jj'}^{\text{tr}}$ in Figs 17 and 18. The comparators fail to recover the true dependence structure between OTUs. The RMSE computed for $\rho_{jj'}$ is shown in Tab 1(a) for all methods in comparison including ZI-MLN. The RMSE under ZI-MLN is much smaller than those under the comparators. It is possibly because the comparators do not attempt to estimate covariate effects on OTU abundances. Additional comparators, ZI-MLN without Λ , metagenomeSeq

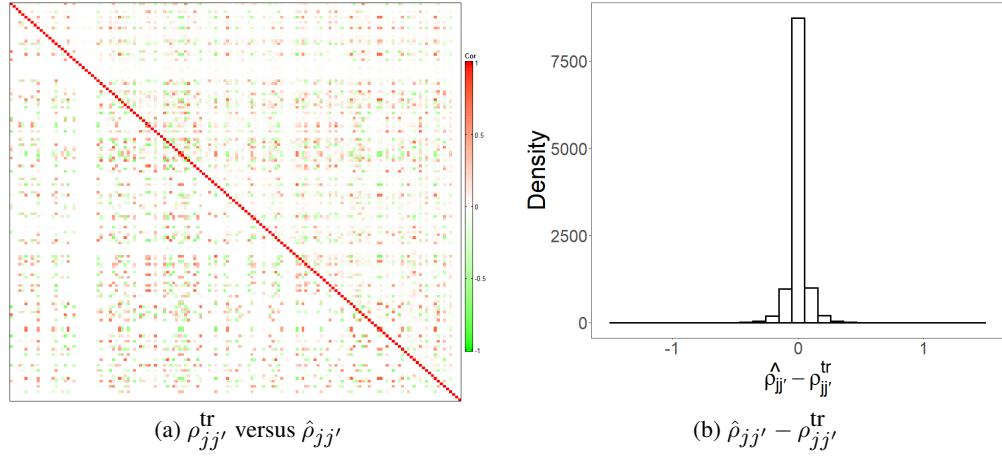


FIG 11. [Simulation 3] The upper right and lower left triangles of the heatmap in panel (a) illustrate posterior estimates of correlations $\hat{\rho}_{jj'}$ and their true values $\rho_{jj'}^{tr}$, respectively. Panel (b) has a histogram of differences between $\hat{\rho}_{jj'}$ and $\rho_{jj'}^{tr}$.

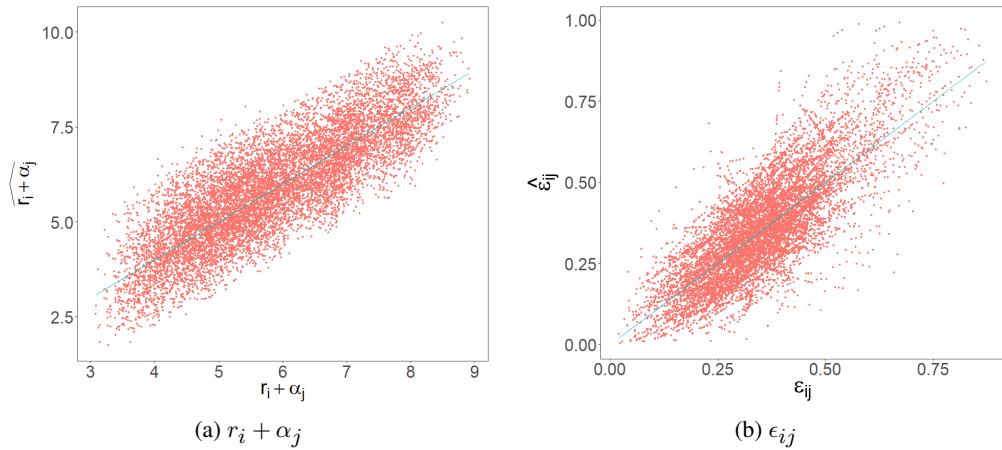


FIG 12. [Simulation 3] In panel (a), posterior mean estimates of the mean abundance $r_i + \alpha_j$ are plotted against the simulation truth $r_i^{tr} + \alpha_j^{tr}$. Panel (b) has a histogram of the differences in posterior mean estimates $\hat{\epsilon}_{ij}$ of probabilities of an OTU being absent and their true values ϵ_{ij}^{tr} .

and edgeR were also applied to compare estimates of the covariate effects. The RMSEs are computed for the parameters δ_{ij} , μ_{ij} , $\beta_{j2} - \beta_{j1}$, β_{j3} and κ_{jp} and summarized in Tab 1(b). Our ZI-MLN outperforms the comparators even when the counts are generated from Poisson distributions. Estimates of regression coefficients $\beta_{j1} - \beta_{j2}$ and β_{j3} obtained from the comparators are compared to their truth in Figs 13(c)-(h). Figs 15(d)-(f) illustrate a comparison of posterior mean estimates of probit regression coefficients κ_{jp} under ZI-MLN without Λ to the truth. Predictive model checking for ZI-MLN without Λ is reported in Fig 16(b) by comparing their posterior predictive median estimates $\hat{y}_{ij}^{\text{pred}}$ of OTU counts to the observed counts. Estimates of mean abundance levels under metagenomeSeq and edgeR are plotted against the observed counts in Figs 16(c) and (d), respectively. The comparison indicates that ignoring the interrelationship between OTUs may distort inferences on mean abundances and the absence/presence of OTUs. Also, comparison of ZI-MLN to edgeR shows that ignoring excess zeros may lead to poor estimation of mean abundance levels.

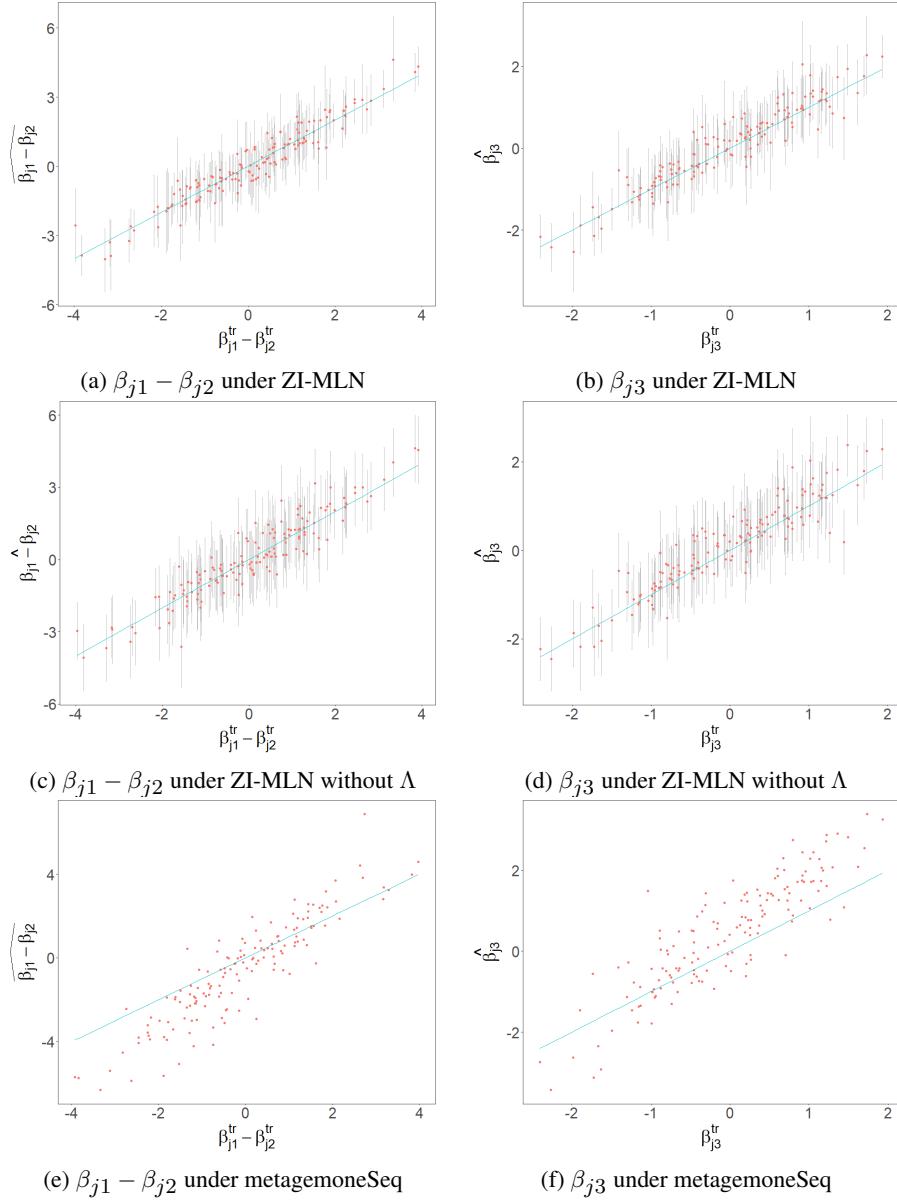


FIG 13. [Simulation 3] Estimates $\widehat{\beta_{j1} - \beta_{j2}}$ and $\hat{\beta}_{j3}$ of regression coefficients are compared to the truth $\beta_{j1}^{tr} - \beta_{j2}^{tr}$ and β_{j3}^{tr} . Estimates in the four rows are obtained by ZI-MLN, ZI-MLN without Λ , metagenomeSeq and edgeR, respectively. The left and right columns are for $\widehat{\beta_{j1} - \beta_{j2}}$ and $\hat{\beta}_{j3}$, respectively.

2.4. Simulation 4. In Simulation 4, we considered a dataset without any dependency structure between OTUs. We let $\lambda_{jk}^{tr} = 0$ for all (j, k) and had $\Sigma^{tr} = \sigma^{2,tr} \mathbf{I}_J$ with $\sigma^{2,tr} = 1$. We kept the remaining simulation setup the same as in Simulation 2. The fixed hyperparameters are set the same as in Simulation 2 to fit the model, and the posterior samples were drawn from the posterior distribution via MCMC. We discarded the first 15,000 iterates for burn-in and kept the next 15,000 iterates for posterior inference. We examined the mixing and convergence of the Markov chains using trace plots and did not find evidence of poor mixing or bad convergence. For example, Fig 19 shows traceplots of the log-likelihood from MCMC chains under different random seeds and initializations and does not indicate evidence of poor mixing or convergence problem.

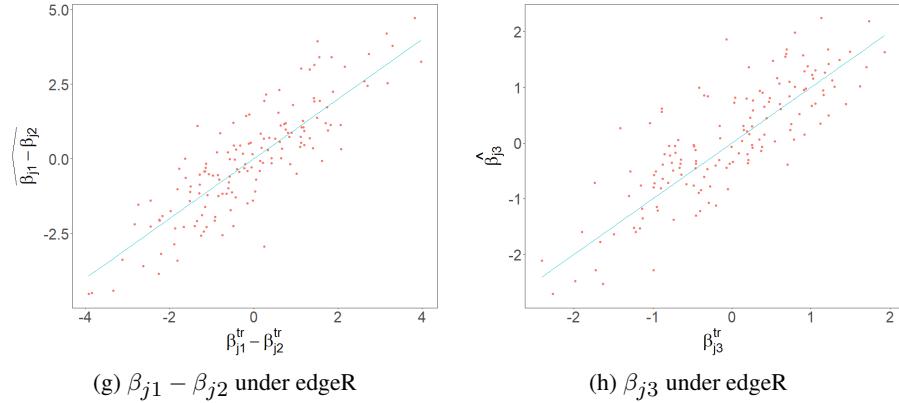


FIG 14. *

Fig 13 continued [Simulation 3] Estimates of regression coefficients $\widehat{\beta_{j1} - \beta_{j2}}$ and $\widehat{\beta_{j3}}$ are compared to the truth $\beta_{j1}^{\text{tr}} - \beta_{j2}^{\text{tr}}$ and β_{j3}^{tr} . Estimates in the four rows are obtained by ZI-MLN, ZI-MLN without Λ , metagenomeSeq and edgeR, respectively. The left and right columns are for $\widehat{\beta_{j1} - \beta_{j2}}$ and $\widehat{\beta_{j3}}$, respectively.

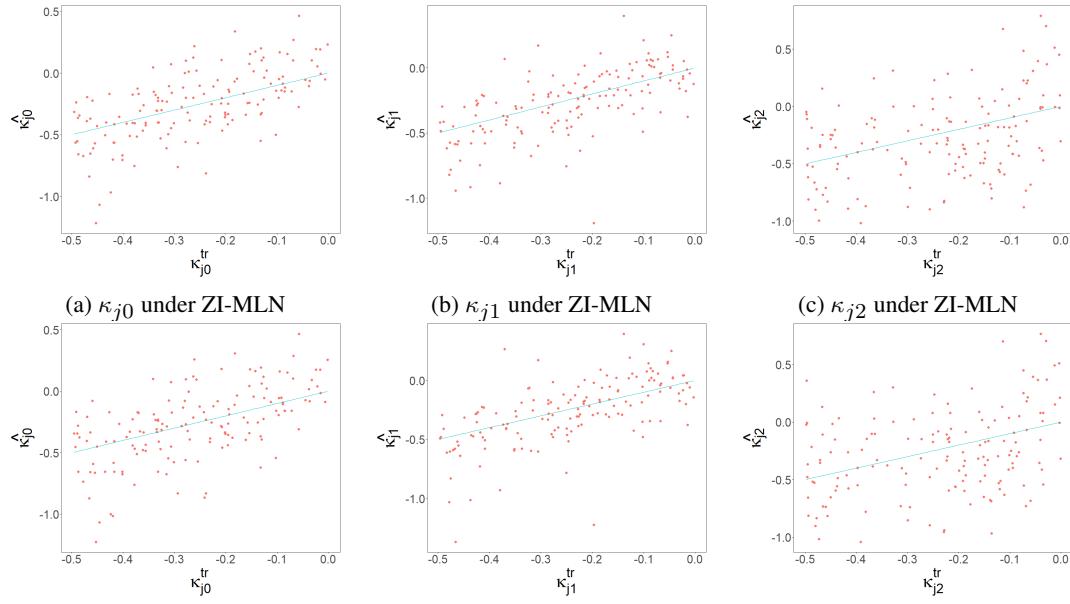


FIG 15. [Simulation 3] Posterior mean estimates $\hat{\kappa}_{j\eta}$ of coefficients on ϵ_{ij} are plotted against the truth for

Posterior inference is summarized in Figs 20-25. Fig 20 shows that our posterior mean estimates of marginal correlations are close to the truth, $\rho_{jj}^{\text{tr}} = 1$ and $\rho_{jj'}^{\text{tr}} = 0$, $j \neq j'$. Posterior mean estimates of σ^2 and u_s^2 are (1.002, 0.922), which are close to their true values of 1. Also, we check the estimation of the mean abundances. Specifically, from Fig 21(a) posterior mean estimates $\widehat{r_i + \alpha_j}$ are tightly around $r_i^{\text{tr}} + \alpha_j^{\text{tr}}$. Fig 21(b) compares posterior estimates of the probabilities of OTUs being absent, ϵ_{ij} to the truth, indicating reasonable inferences on the absence/presence of OTUs. Figs 22(a)-(b) and 24(a)-(c) show that estimates of the regression coefficients are reasonably well estimated. Posterior predictive checking is illustrated in Fig 25(a). The plot shows that our model provides a good fit even when there is no dependence structure assumed in the truth.

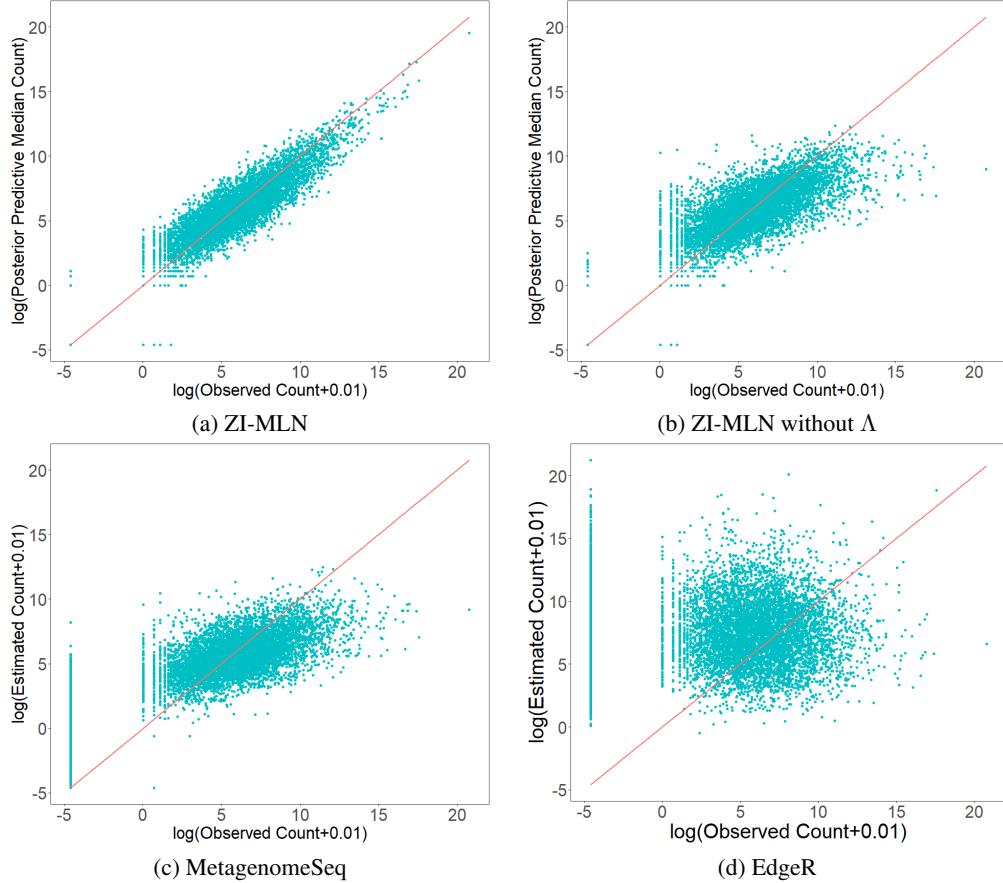


FIG 16. [Simulation 3] Panels (a) and (b) compare posterior predictive median counts to their observed counts on the logarithm scale, $\log(y_{ij} + 0.01)$ versus $\log(\hat{y}_{ij}^{\text{pred}} + 0.01)$. $\hat{y}_{ij}^{\text{pred}}$ are estimated with ZI-MLN with Λ in (a) and without Λ in (b). Panels (c) and (d) present scatter plots of observed $\log(y_{ij} + 0.01)$ versus $\log(\hat{\mu}_{ij} + 0.01)$ by metagenomeSeq and edgeR, respectively, where $\hat{\mu}_{ij}$ are estimated mean abundances of OTUs in samples.

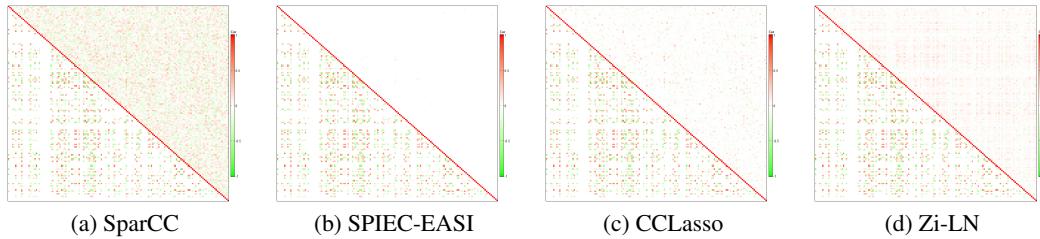


FIG 17. [Simulation 3: Comparison] The upper right and lower left triangles of each heatmap illustrate estimates $\hat{\rho}_{jj'}$ of correlations between OTUs and their true values $\rho_{jj'}^{\text{tr}}$. Panels (a)-(d) are for SparCC, SPIEC-EASI, CCLasso and Zi-LN, respectively.

We applied the comparators, SparCC, SPIEC-EASI, CCLasso and Zi-LN, to the simulated data and compared their performance of estimating $\rho_{jj'}$ to that of our ZI-MLN. Figs 26 and 27 compare the estimates of $\rho_{jj'}$ under the comparators to the truth. The RMSE of $\rho_{jj'}$ is computed and given in Tab 2(a). ZI-MLN outperforms the other methods in comparison for estimating $\rho_{jj'}$. SparCC yields estimates not close to zero for many $\rho_{jj'}$ and yields the largest RMSE for this simulated dataset. The additional comparators, ZI-MLN without Λ , metagenomeSeq and edgeR were applied to the dataset for further comparison. RMSEs

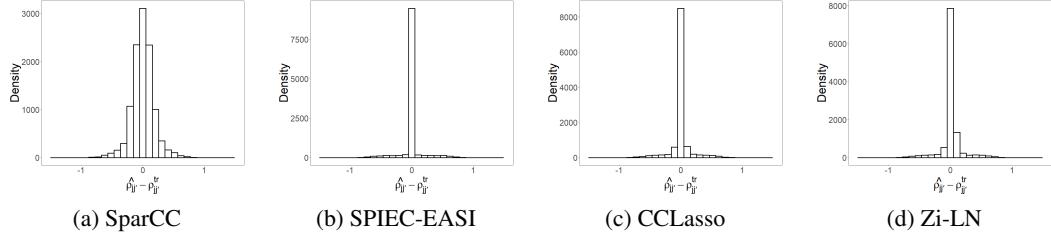


FIG 18. [Simulation 3: Comparison] A histogram of differences between $\hat{\rho}_{jj'}$ under SparCC, SPIEC-EASI, CCLasso and ZiLN and $\rho_{jj'}^{tr}$, in panels (a)-(d), respectively.

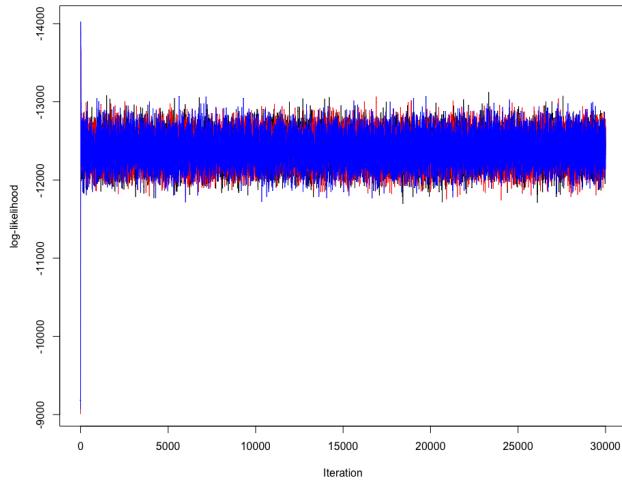


FIG 19. [Simulation 4] Traceplots of log likelihood under three different initializations are presented in black, red and blue, respectively.

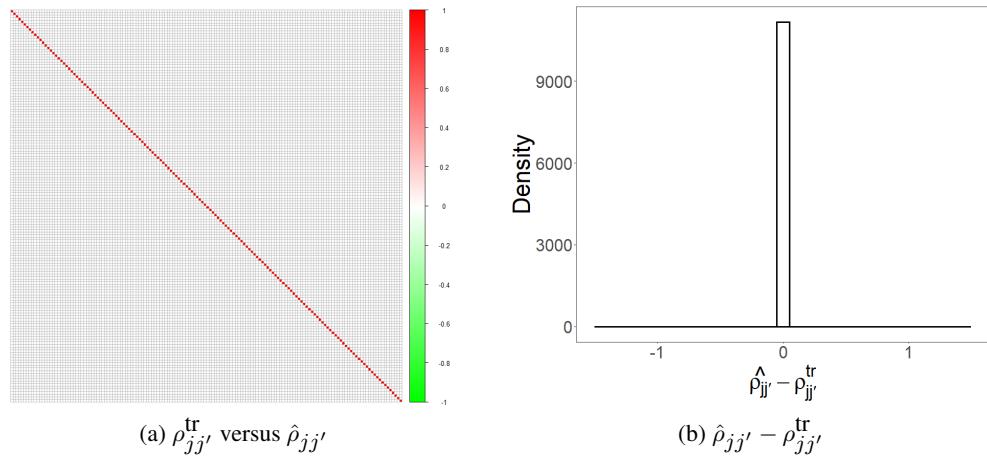


FIG 20. [Simulation 4] The upper right and lower left triangles of the heatmap in panel (a) illustrate posterior estimates of correlations $\hat{\rho}_{jj'}$ and their true values $\rho_{jj'}^{tr}$, respectively. Panel (b) has a histogram of differences between $\hat{\rho}_{jj'}$ and $\rho_{jj'}^{tr}$.

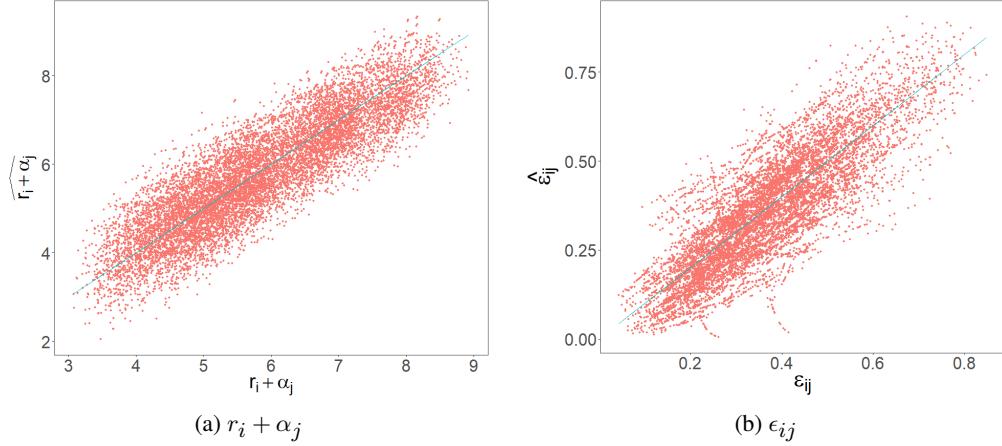


FIG 21. [Simulation 4] In panel (a), posterior mean estimates of the mean abundance $r_i + \alpha_j$ are plotted against the simulation truth $r_i^{tr} + \alpha_j^{tr}$. Panel (b) has a histogram of the differences in posterior mean estimates $\hat{\epsilon}_{ij}$ of probabilities of an OTU being absent and their true values ϵ_{ij}^{tr} .

TABLE 2

[Simulation 4: Comparison] RMSEs are computed for $\rho_{jj'}$, $j < j'$, δ_{ij} , μ_{ij} , $\beta_{j2} - \beta_{j1}$, β_{j3} and κ_{jp} under ZI-MLN and comparators.

Model	$\rho_{jj'}$
ZI-MLN	0.0011
SparCC	0.128
SPIEC-EASI	0.0042
CCLasso	0.053
Zi-LN	0.051

(a) $\rho_{jj'}$

Model	δ_{ij}	μ_{ij}	$\beta_{j2} - \beta_{j1}$	β_{j3}	κ_{j0}	κ_{j1}	κ_{j2}
ZI-MLN	0.052	0.783	0.325	0.235	0.223	0.170	0.170
ZI-MLN without Λ	0.052	0.783	0.370	0.241	0.223	0.170	0.170
metagenomeSeq	0.078	2.161	1.092	0.711	-	-	-
edgeR	-	1.491	0.741	0.452	-	-	-

(b) δ_{ij} , μ_{ij} , $\beta_{j2} - \beta_{j1}$, β_{j3} and κ_{jp}

for the parameters, δ_{ij} , μ_{ij} , $\beta_{j2} - \beta_{j1}$, β_{j3} and κ_{jp} , are computed and listed in Tab 2(b). ZI-MLN performs the best, very closely followed by ZI-MLN without Λ , or the same as ZI-MLN without Λ although the simulation truth is closer to the assumption made under ZI-MLN without Λ . Also, the truth is close to the assumption that metagenomeSeq, but RMSE of mean abundances μ_{ij} is large compared to those under the other methods. Figs 22(c)-(h) compare estimates of the regression coefficients on OTU abundances under the comparators. Figs 24(d)-(f) show posterior mean estimates of κ_{jp} under ZI-MLN without Λ . Posterior predictive checking is shown in Fig 25(b) for ZI-MLN without Λ . Estimates of mean abundances under metagenomeSeq and edgeR are shown in Figs 25(c) and (d), respectively. The results indicate that there is no degradation in the performance of ZI-MLN in a case where there is no dependence in the data generating process. Note that Simulations 1-3 indicate the additional flexibility of ZI-MLN allows to outperform the comparators when data has a dependence structure.

2.5. *Simulation 5.* In this simulation study, we used SparseDOSSA (Ma *et al.*, 2021) to generate a dataset. An open-source software *SparseDOSSA2* is available from the authors' webpage. SparseDOSSA takes a real microbiome dataset as an input and generates a realistic microbiome dataset. Generated OTU counts in a sample are constrained to sum up to

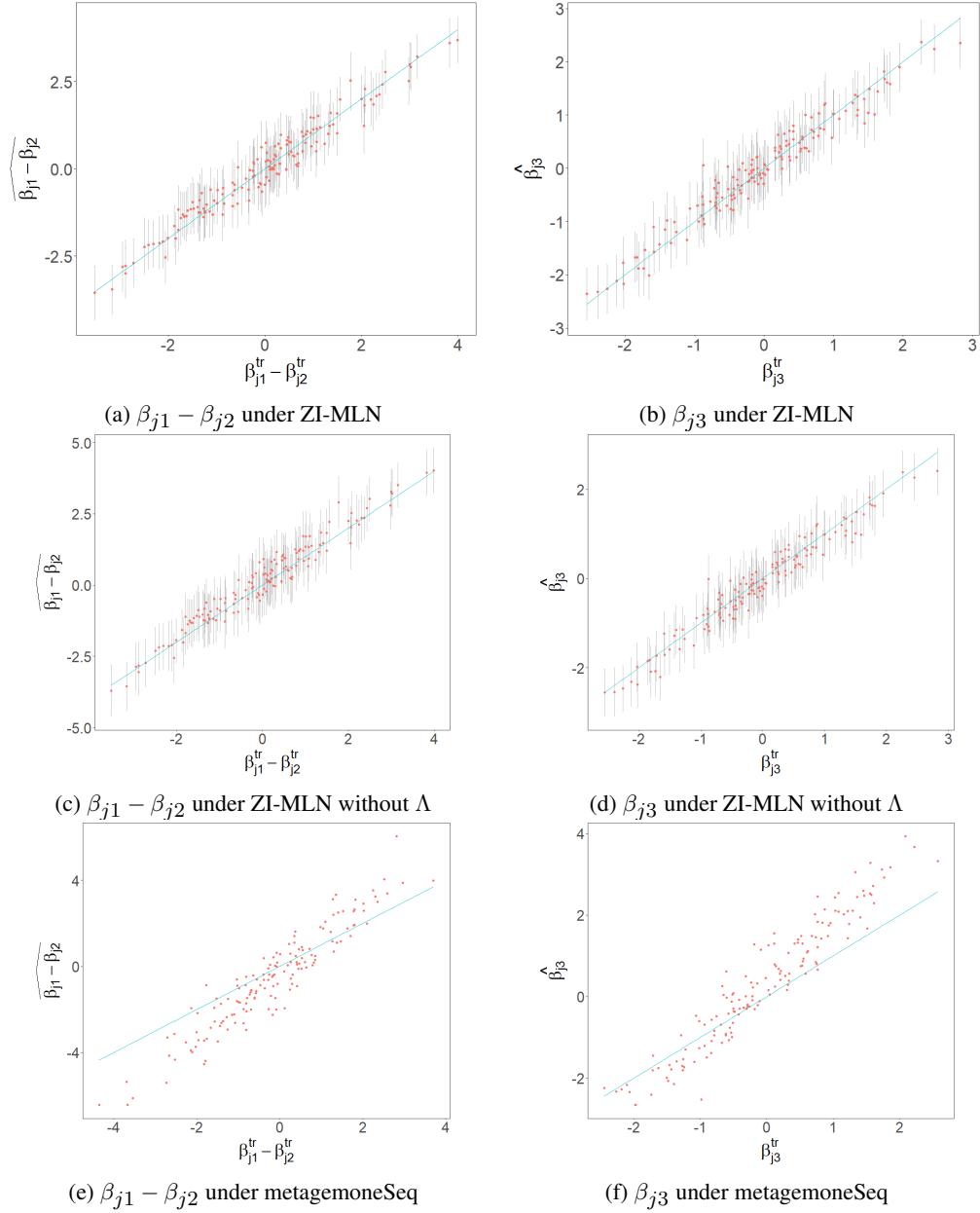


FIG 22. [Simulation 4] Estimates $\widehat{\beta}_{j1} - \widehat{\beta}_{j2}$ and $\widehat{\beta}_{j3}$ of regression coefficients are compared to the truth $\beta_{j1}^{\text{tr}} - \beta_{j2}^{\text{tr}}$ and β_{j3}^{tr} . Estimates in the four rows are obtained by ZI-MLN, ZI-MLN without Λ , metagenomeSeq and edgeR, respectively. The left and right columns are for $\widehat{\beta}_{j1} - \widehat{\beta}_{j2}$ and $\widehat{\beta}_{j3}$, respectively.

a constant (compositionality), are enriched for zero counts (zero-inflated), and are dependent due to microbial interactions. SparseDOSSA assumes a Gaussian copula model with a zero-inflated log-normal distribution for latent absolute (unnormalized) OTU abundances and generates OTU count vectors from a multinomial distribution with the relative abundances normalized from the absolute abundances. For the multinomial distribution, sample total counts are independently simulated from a log-normal distribution. SparseDOSSA does not include random effects, and the same relative abundance vector is assumed for the samples. Thus, a simulated dataset may not exhibit overdispersion that ZI-MLN accounts for

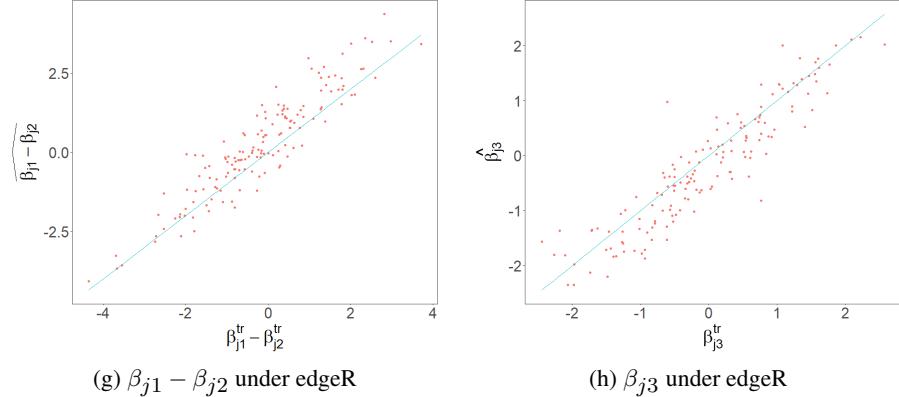


FIG 23. *

Fig 22 continued [Simulation 4] Estimates of regression coefficients $\widehat{\beta_{j1} - \beta_{j2}}$ and $\widehat{\beta_{j3}}$ are compared to the truth $\beta_{j1}^{\text{tr}} - \beta_{j2}^{\text{tr}}$ and β_{j3}^{tr} . Estimates in the four rows are obtained by ZI-MLN, ZI-MLN without Λ , metagenomeSeq and edgeR, respectively. The left and right columns are for $\widehat{\beta_{j1} - \beta_{j2}}$ and $\widehat{\beta_{j3}}$, respectively.

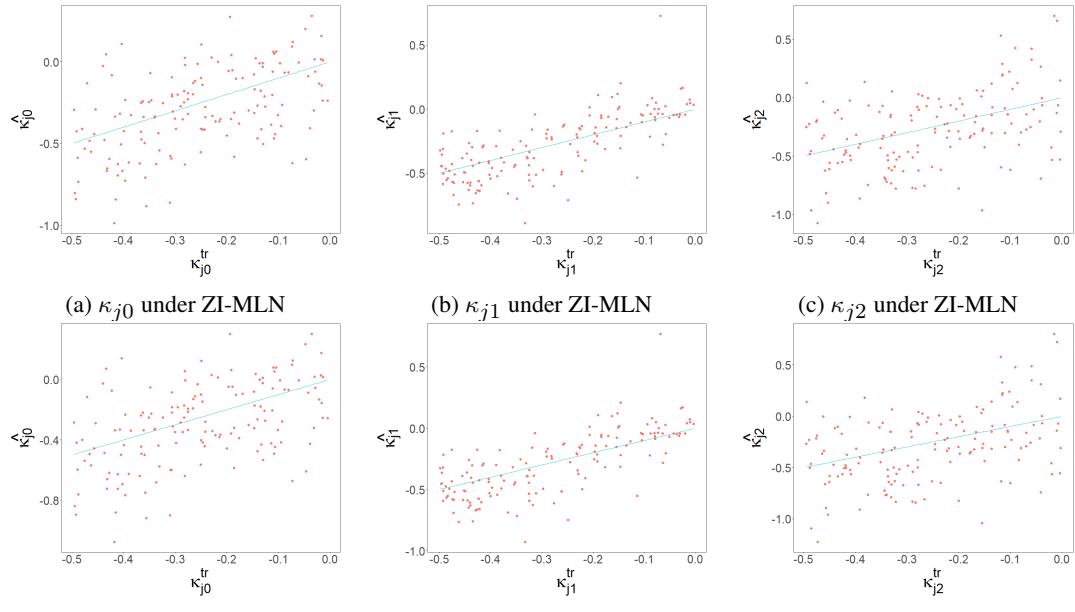


FIG 24. [Simulation 4] Posterior mean estimates $\hat{\kappa}_{jp}$ of coefficients on ϵ_{ij} are plotted against the truth for

through subject group factor effects s_{g_i} . SparseDOSSA estimates the model parameters using an input dataset, e.g., the mean vector and precision matrix of the absolute abundance vector, and then sets the true input parameter values at its estimates. The parameters are estimated by an EM algorithm. In particular, its precision matrix is estimated with a ℓ_1 penalty function for sparsity. The correlation matrix estimated from SparseDOSSA is shown in the lower triangles of the heatmaps in Fig 29. The data-generating process under SparseDOSSA is greatly different from that assumed for ZI-MLN. We used the skin microbiome dataset in §4.1 of the main text as an input dataset. The dataset has $N = 20$ samples and $J = 187$ OTUs. Similar to that of the skin microbiome data, the simulated dataset has zeros for 30% of the counts.

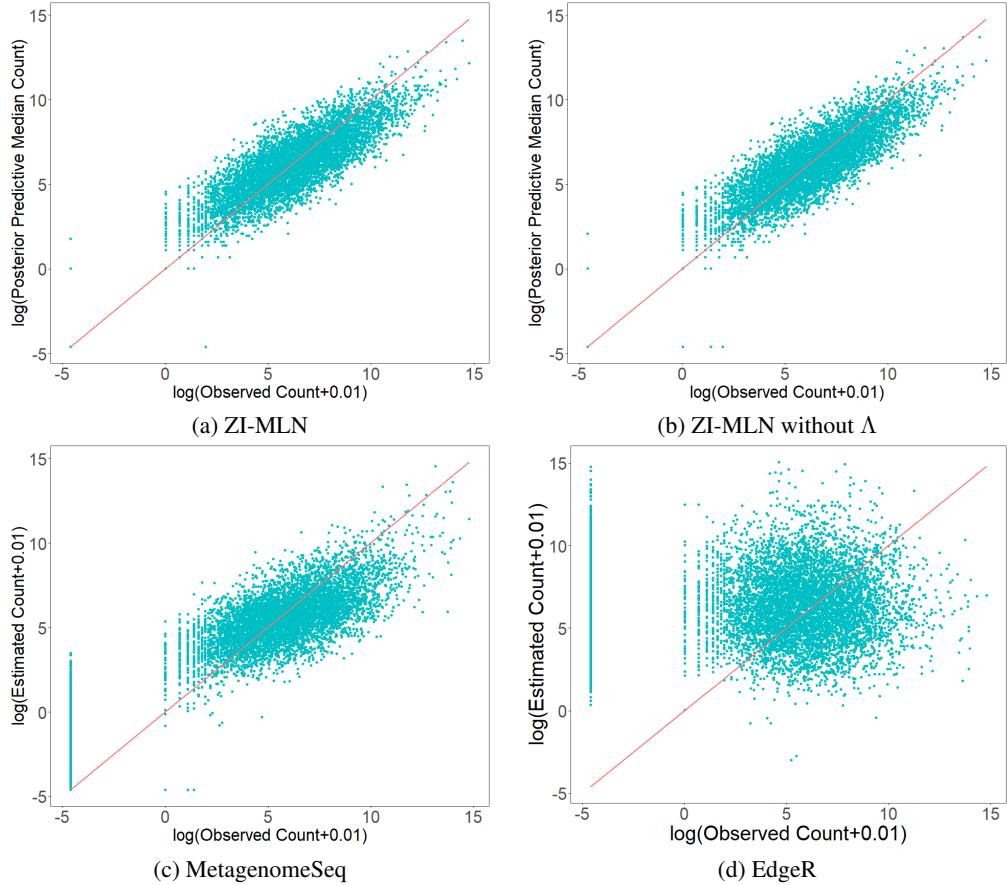


FIG 25. [Simulation 4] Panels (a) and (b) compare posterior predictive median counts to their observed counts on the logarithm scale, $\log(y_{ij} + 0.01)$ versus $\log(\hat{y}_{ij}^{\text{pred}} + 0.01)$. $\hat{y}_{ij}^{\text{pred}}$ are estimated with ZI-MLN with Λ in (a) and without Λ in (b). Panels (c) and (d) present scatter plots of observed $\log(y_{ij} + 0.01)$ versus $\log(\hat{\mu}_{ij} + 0.01)$ by metagenomeSeq and edgeR, respectively, where $\hat{\mu}_{ij}$ are estimated mean abundances of OTUs in samples.

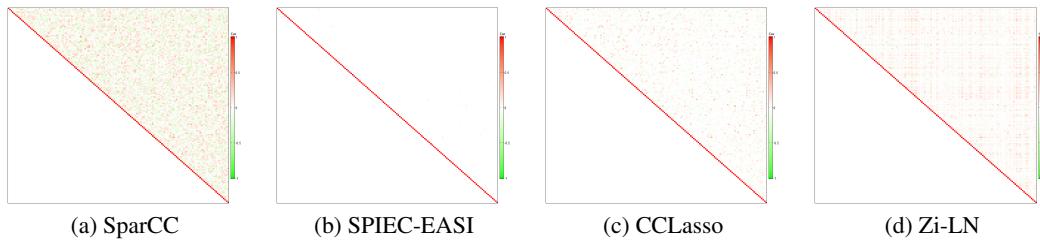


FIG 26. [Simulation 4: Comparison] The upper right and lower left triangles of each heatmap illustrate estimates $\hat{\rho}_{jj'}$ of correlations between OTUs and their true values $\rho_{jj'}^{\text{tr}}$. Panels (a)-(d) are for SparCC, SPIEC-EASI, CCLasso and Zi-LN, respectively.

To fit ZI-MLN, we specified the fixed hyperparameter values similar to those in the previous simulation studies. For this simulation study, we set $a_\phi = 1/10$ for greater sparsity. We ran MCMC for 30,000 iterations with the first 15,000 iterations to burn-in. To examine mixing and convergence, we ran multiple chains under different initial values and random seeds. Fig 28 shows the chains with different initial values and random seeds converge to similar log-likelihood ranges, indicating no empirical evidence of bad mixing or convergence.

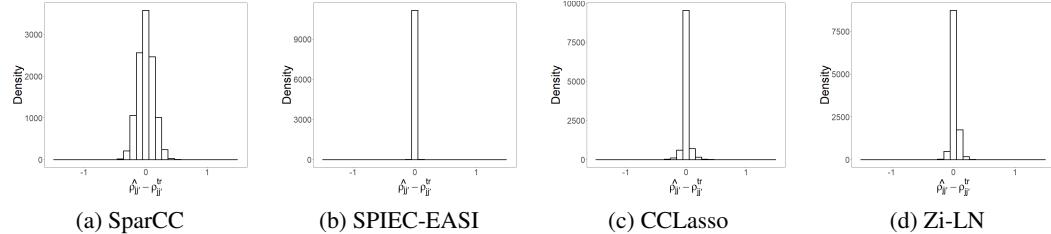


FIG 27. [Simulation 4: Comparison] A histogram of differences between $\hat{\rho}_{jj'}$ under SparCC, SPIEC-EASI, CCLasso and ZiLN and $\rho_{jj'}^{tr}$, in panels (a)-(d), respectively.

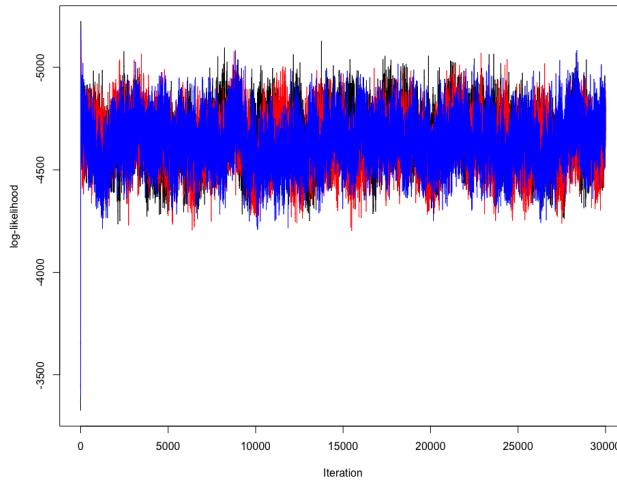


FIG 28. [Simulation 5] Traceplots of log likelihood under three different initializations are presented in black, red and blue, respectively.

TABLE 3

[Simulation 5: Comparison] RMSEs are computed for $\rho_{jj'}$, $j < j'$, δ_{ij} and $\tilde{\mu}_{ij}$ under ZI-MLN and comparators. $\tilde{\mu}_{ij}$ is the mean abundance adjusted by a sample total count.

Model	$\rho_{jj'}$
ZI-MLN	0.034
SparCC	0.253
SPIEC-EASI	0.034
CCLasso	0.142
Zi-LN	0.048

(a) $\rho_{jj'}$

Model	δ_{ij}	$\tilde{\mu}_{ij}$
ZI-MLN	0.071	0.917
ZI-MLN without Λ	0.071	0.918
metagenomeSeq	0.091	1.922

(b) δ_{ij} and $\tilde{\mu}_{ij}$

Posterior inferences under ZI-MLN are summarized in Figs 29(a), 30(a) and 31(a). Figs 29(a) and 30(a) show that the true underlying between-OTU dependence structure is well recovered although the dataset was generated from a very different model. Especially, the true data-generating process assumes a multimodal distribution that conditions on sample total counts. The model-based normalization through sample size factors r_i under ZI-MLN accounts for compositionality reasonably well, and the model provides reasonable estimates

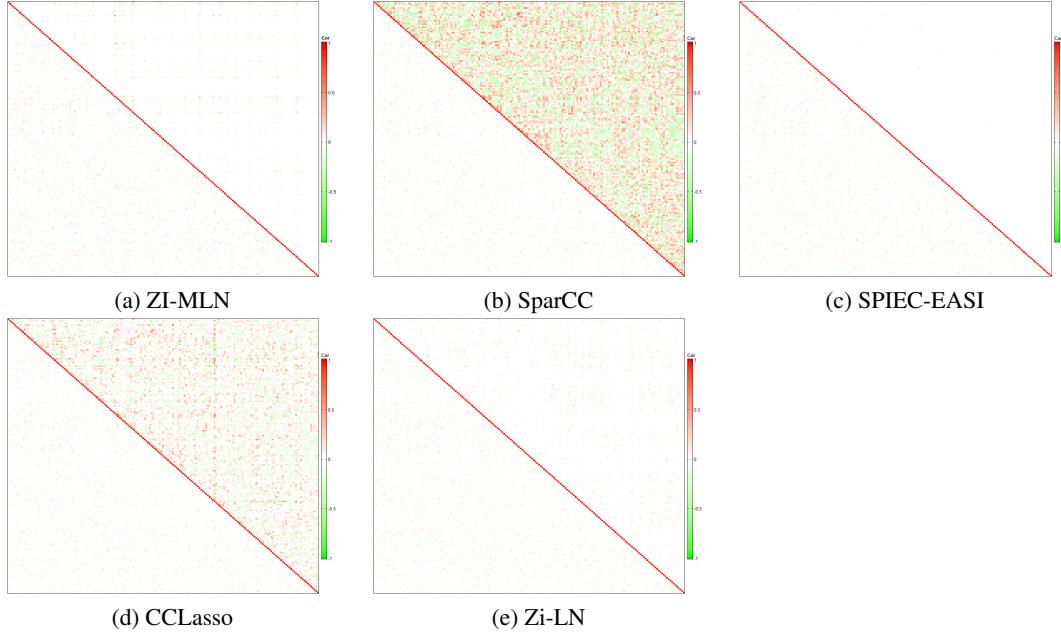


FIG 29. [Simulation 5] The upper right and lower left triangles of each heatmap illustrate estimates $\hat{\rho}_{jj'}$ of correlations between OTUs and their true values $\rho_{jj'}^{tr}$. Panels (a)-(e) are for ZI-MLN, SparCC, SPIEC-EASI, CCLasso and Zi-LN, respectively.

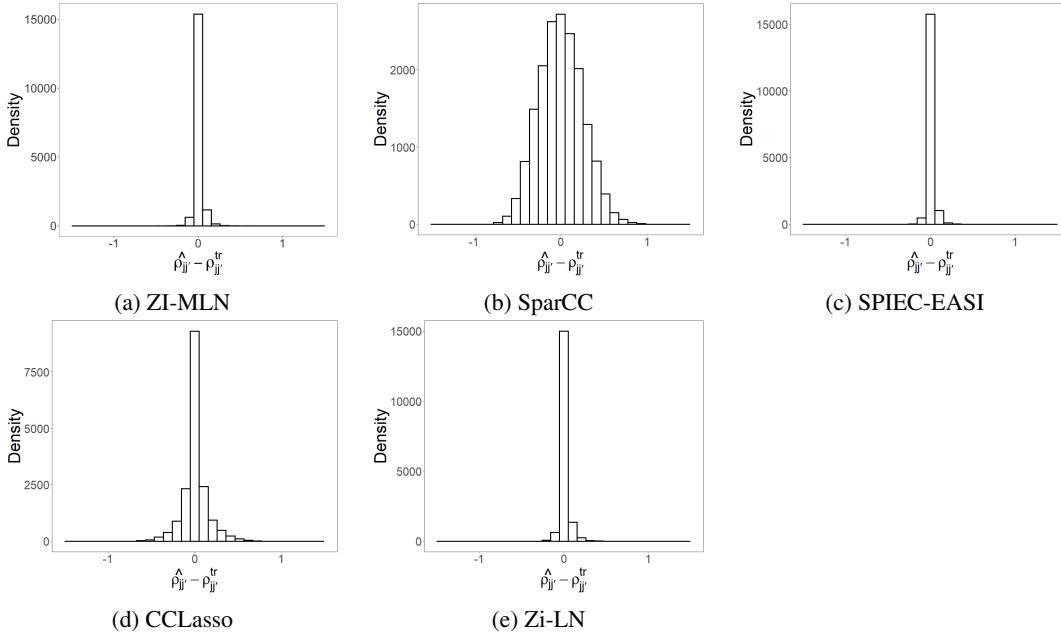


FIG 30. [Simulation 5] A histogram of differences between $\hat{\rho}_{jj'}$ under ZI-MLN, SparCC, SPIEC-EASI, CCLasso and ZiLN and $\rho_{jj'}^{tr}$, in panels (a)-(e), respectively.

of $\rho_{jj'}$. From Fig 31(a), the absence/presence of OTUs is also well estimated. Posterior predicted mean counts are compared to observed counts in Fig 32(a). The plot indicates that ZI-MLN fits the data well.

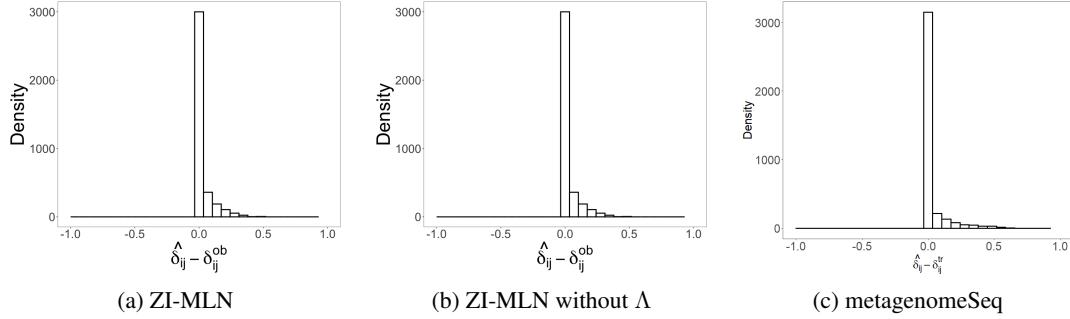


FIG 31. [Simulation 5] Histograms of the differences between $\hat{\delta}_{ij}$ and the observed zero indicator $1(Y_{ij} = 0)$ under ZI-MLN, ZI-MLN without Λ and metagenomeSeq, respectively. Posterior mean estimates are used for ZI-MLN and ZI-MLN without Λ .

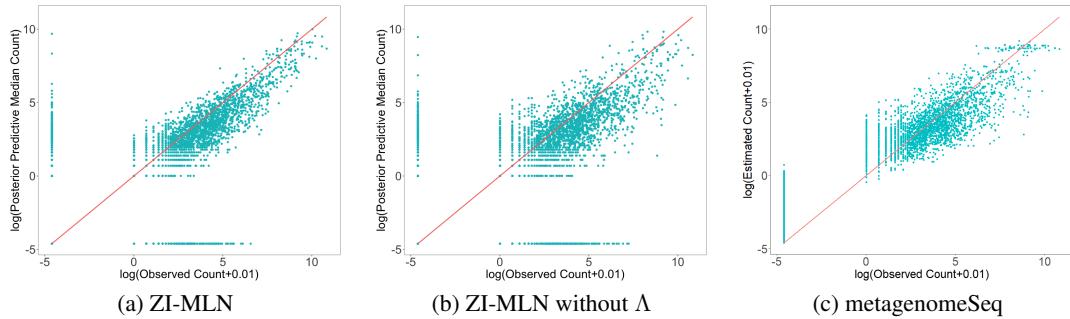


FIG 32. [Simulation 5] Panels (a) and (b) have scatter plots of observed $\log(y_{ij} + 0.01)$ versus $\log(\hat{y}_{ij}^{pred} + 0.01)$ under ZI-MLN and ZI-MLN without Λ , respectively. Panel (c) is the scatter plots of observed $\log(y_{ij} + 0.01)$ versus mean abundance estimates $\log(\hat{\mu}_{ij} + 0.01)$ by metagenomeSeq.

For comparison, we applied SparCC, SPIEC-EASI, CCLasso and Zi-LN to the dataset simulated by SparseDOSSA. Note that SPIEC-EASI and CCLasso use ℓ_1 penalty to estimate dependence structure, similar to SparseDOSSA. Estimates $\hat{\rho}_{jj'}$ of the correlations obtained under the comparators are compared to their true values $\rho_{jj'}^{\text{tr}}$ in Figs Figs 29(b)-(e) and 30(b)-(e). We also computed the RMSE of $\rho_{jj'}$ under all methods in comparison including ZI-MLN. From Tab 3(a), ZI-MLN and SPIEC-EASI produce the smallest value of RMSE. It is noticeable that correlation matrix estimates under SparCC and CCLasso are very dense, resulting in very large values of RMSE. Also, we applied the additional comparators, ZI-MLN without Λ and metagenomeSeq, to the dataset. RMSE of δ_{ij} and μ_{ij} are computed and summarized in Tab 3(b). Since the counts were generated from a multinomial distribution, we adjusted estimates of μ_{ij} by the total sample counts, $\tilde{\mu}_{ij} = \hat{\mu}_{ij} - \log(\sum_j Y_{ij})$, and compared $\tilde{\mu}_{ij}$ to the true normalized abundance of SparseDOSSA. ZI-MLN outperforms ZI-MLN without Λ and metagenomeSeq in estimating the presence/absence of OTUs and their mean abundances. $\hat{\delta}_{ij}$ under ZI-MLN without Λ and metagenomeSeq are compared to the observed zero indicators $1(Y_{ij} = 0)$ in Fig 31(b) and (c), respectively. Posterior predictive mean counts under ZI-MLN without Λ are plotted against the observed counts in Fig 32(b). ZI-MLN without Λ yielded a poorer fit to the data than ZI-MLN. The mean abundance estimates under metagenomeSeq are compared to transformed observed counts in Fig 32(c).

3. Additional Results for Real Data Analyses.

3.1. Additional Results for Skin Microbiome Data Analysis. Multiple MCMC chains were run with different initial values and random seeds to examine the mixing and conver-

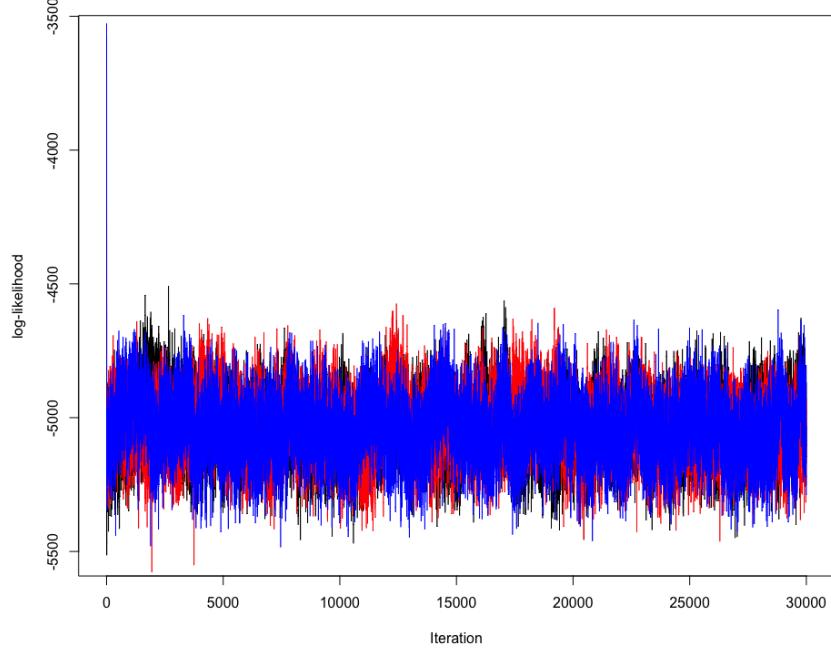


FIG 33. [Skin Microbiome Data] Traceplots of log likelihood under three different initializations are presented in black, red and blue, respectively.

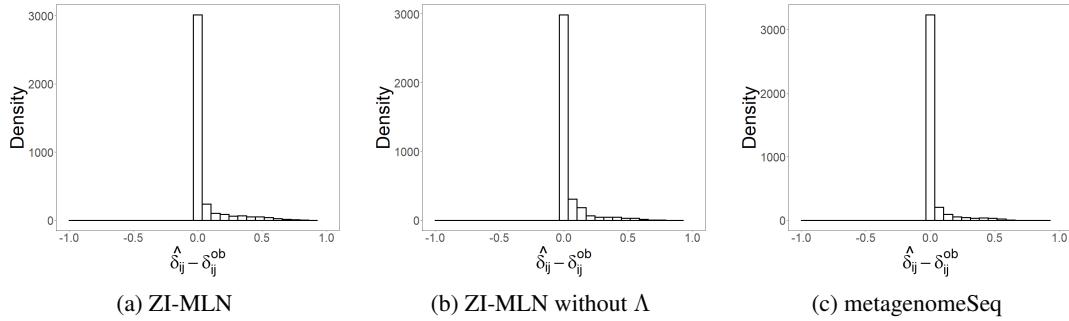


FIG 34. [Skin Microbiome Data] Histograms of the differences between $\hat{\delta}_{ij}$ and the observed zero indicators $1(Y_{ij} = 0)$ under ZI-MLN, ZI-MLN without Λ and metagenomeSeq, respectively. Posterior mean estimates are used for ZI-MLN and ZI-MLN without Λ .

gence of the MCMC. Fig 33 illustrates traceplot of the log-likelihood of the MCMC runs and shows no evidence of poor mixing or convergence issues. Tab 4 has taxonomic information for the OTUs illustrated in Fig 9(b) of the main text. We include the comparison of the observed zero inflation rate against the observed zero indicators $1(Y_{ij} = 0)$ under ZI-MLN, ZI-MLN without Λ and metagenomeSeq in Fig 34.

To examine robustness to the specification of the threshold used for data preprocessing, we performed a sensitivity analysis. We used five different values of the threshold to remove OTUs that have zeros in too many samples from analysis. In particular, OTUs that have zero counts in more than $b\%$ of the samples were removed, where 40%, 45%, 50%, 55% and 60% are used for b . Note that $b=50\%$ is used for the analysis in §4.1 of the main text. The skin microbiome data has a total of 20 samples, and those cutoff values remove OTUs who

TABLE 4
[Skin Microbiome Data] Taxonomic information for the OTUs illustrated in Fig 9(b) of the main text.

OTU	Taxonomic information (Kingdom/ Phylum/ Class/ Order/ Family / Genus)
41	Bacteria - Actinobacteria - Actinobacteria - Micrococcales - NA - NA
42	Bacteria - Actinobacteria - Actinobacteria - Micrococcales - Micrococcaceae - Glutamicibacter
43	Bacteria - Proteobacteria - Epsilonproteobacteria - Campylobacterales - Campylobacteraceae - Campylobacter
46	Bacteria - Bacteroidetes - Sphingobacteriia - Sphingobacteriales - Chitinophagaceae - uncultured
47	Bacteria - Bacteroidetes - Sphingobacteriia - Sphingobacteriales - Chitinophagaceae - Segetibacter
48	Bacteria - Bacteroidetes - Bacteroidia - Bacteroidales - Porphyromonadaceae - Porphyromonas
76	Bacteria - Proteobacteria - Gammaproteobacteria - Pseudomonadales - Moraxellaceae - Enhydrobacter
88	Bacteria - Firmicutes - Clostridia - Clostridiales - Family XI - Peptoniphilus
92	Bacteria - Firmicutes - Clostridia - Clostridiales - Family XIII - uncultured
138	Bacteria - Proteobacteria - Alphaproteobacteria - Caulobacterales - Caulobacteraceae - Brevundimonas
153	Bacteria - Firmicutes - Bacilli - Lactobacillales - Aerococcaceae - uncultured
155	Bacteria - Firmicutes - Bacilli - Bacillales - Staphylococcaceae - Staphylococcus
173	Bacteria - Actinobacteria - Actinobacteria - Streptomycetales - Streptomycetaceae - Streptomyces

have zero counts in more than 8, 9, 10, 11 and 12 samples, resulting that 147, 163, 187, 213 and 238 OTUs are included for analysis. We fitted the model to each of the preprocessed datasets and compared posterior inferences and model fit. We used the same hyperparameter values. From the posterior predictive checking illustrated in Fig 35, we observe that the model provides a good fit to all datasets the mode fit does not change much by the value of b . We also examined correlation estimates for the OTUs that are included in all five preprocessed datasets and compared. Fig 36 shows the posterior mean estimates $\hat{\rho}_{jj'}$ for seven OTUs that are arbitrarily chosen among the OTUs included in all datasets for illustration. The figure shows that the correlation estimates remain almost unchanged by the value of b , indicating the robustness of the model to the specification of b for preprocessing.

3.2. Additional Results from Human Gut Microbiome Data Analysis. Tab 5 presents the names of the covariates included for human gut microbiome data analysis and their support. The dataset has 37 children subjects collected from two different recruitment sites. The biopsy samples were taken from either of two biopsy locations, ileum or rectum or both locations. The model was run for the dataset three times under different initializations and random seeds. Traceplots of the log-likelihood shown in Fig 37. The plot suggests that the model converged to a similar state under these alternative specifications, and provides practical evidence of the chain's convergence. Fig 39 illustrates posterior mean estimates of κ_{jp} for two selected covariates, age and binary indicator of a subject being white, where posterior mean estimates are represented with black dots, and 95% credible intervals with vertical lines. κ_{jp} whose credible interval does not contain zero are in red.

Tab 6 has taxonomic information of the OTUs in Fig 12(b) of the main text. Fig 38 compares posterior mean estimates $\hat{\delta}_{ij}$ to the observed indicator $1(Y_{ij} = 0)$. Fig 39 presents posterior estimates of coefficients κ_{jp} of the probit regression for two selected covariates, age and race. $\hat{\kappa}_{jp}$ for age is larger than 0 and $\hat{\kappa}_{jp}$ for race smaller than 0 for many OTUs, although they are not statistically significant. Tabs 7 and 8 provide taxonomic information of the OTUs whose abundance and absence/presence are statistically significantly associated with the covariates, respectively.

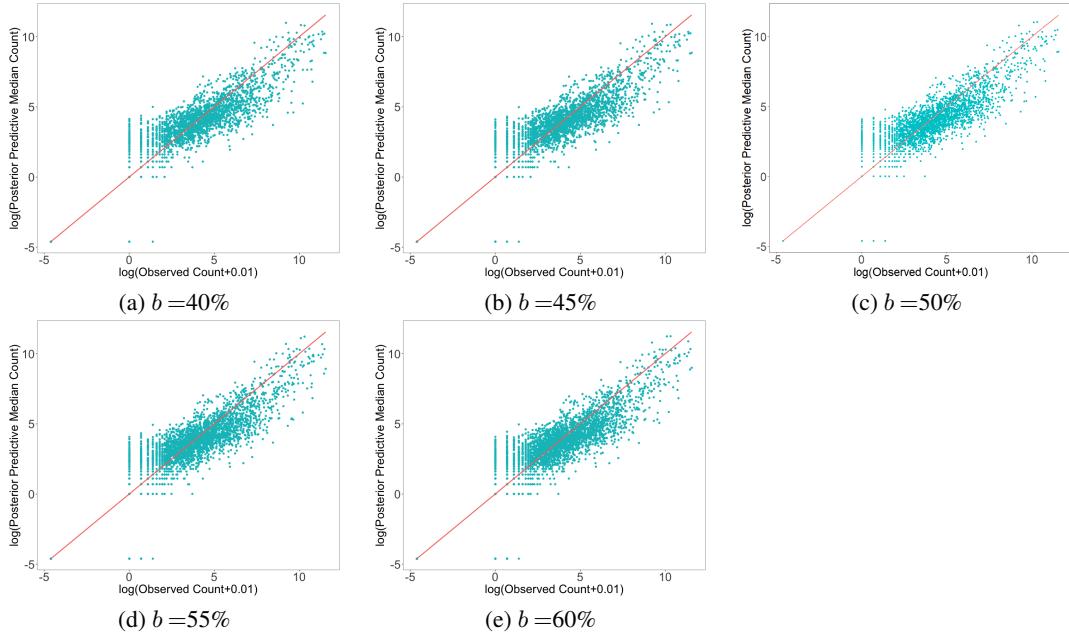


FIG 35. [Sensitivity Analysis for the Skin Microbiome Data] Scatter plots of observed $\log(y_{ij} + 0.01)$ versus posterior predictive log count $\log(\hat{y}_{ij}^{\text{pred}} + 0.01)$ estimated by ZI-MLN. Different threshold values are used for data preprocessing. $b=40\%$, 45% , 50% , 55% and 60% are used for panels (a)-(e), respectively.

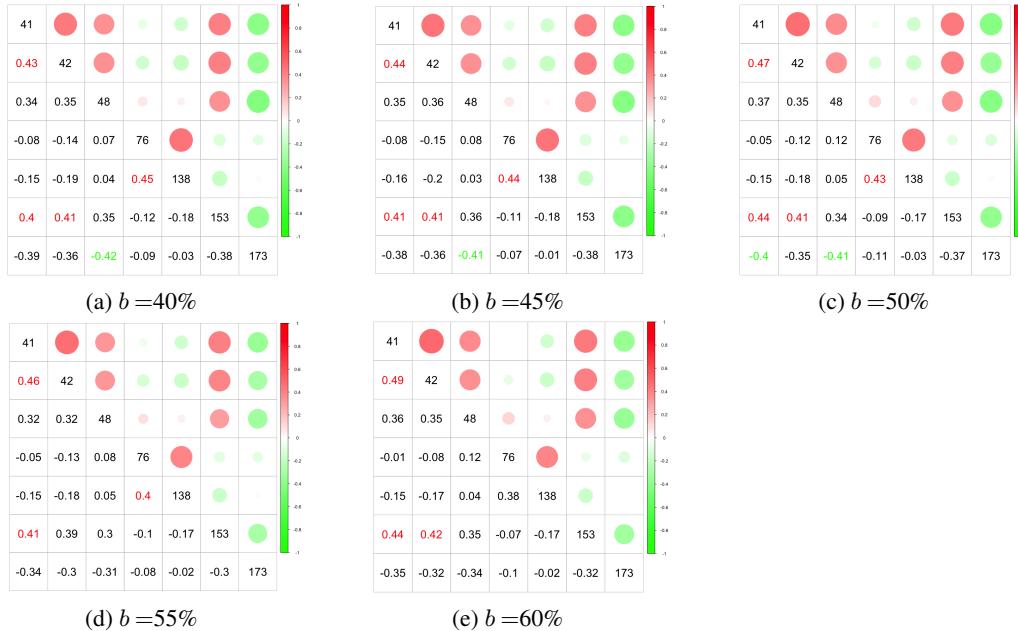


FIG 36. [Sensitivity Analysis for the Skin Microbiome Data] The posterior mean estimates $\hat{\rho}_{jj'}$ of correlations for seven OTUs. The OTUs are arbitrarily chosen for illustration among the OTUs that are included in datasets preprocessed with different threshold values. The value of a preprocessor threshold, $b=40\%$, 45% , 50% , 55% and 60% are used for panels (a)-(e), respectively.

Fig 40 provides point estimates for β_{jp} for some selected covariates under the comparators. In panel (d), age under metagenomeSeq has statistically significant positive effects for most

TABLE 5
[Human Gut Microbiome Data] Covariates names with their support

Covariate Name	Support
Age	6 to 17
Gender	Male or Female
Race	White or non-white
Site Name	Cincinnati Children's Hospital or Massachusetts General Hospital (MGH) Pediatrics
Biopsy location	Ileum or Rectum
Disease phenotype	UC, CD or non-IBD

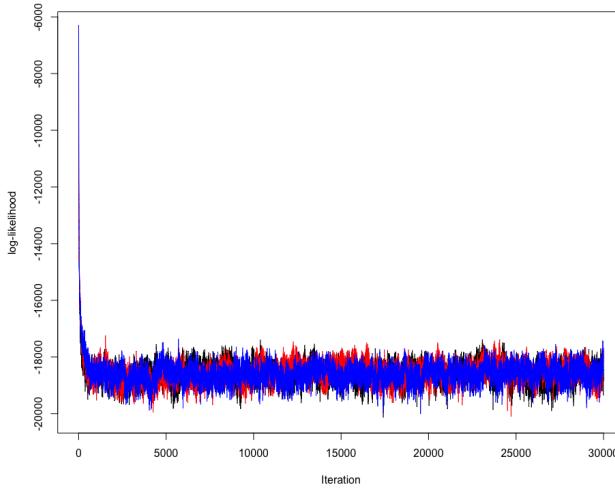


FIG 37. [Human Gut Microbiome Data] Traceplots of log likelihood under three different initializations are presented in black, red and blue, respectively.

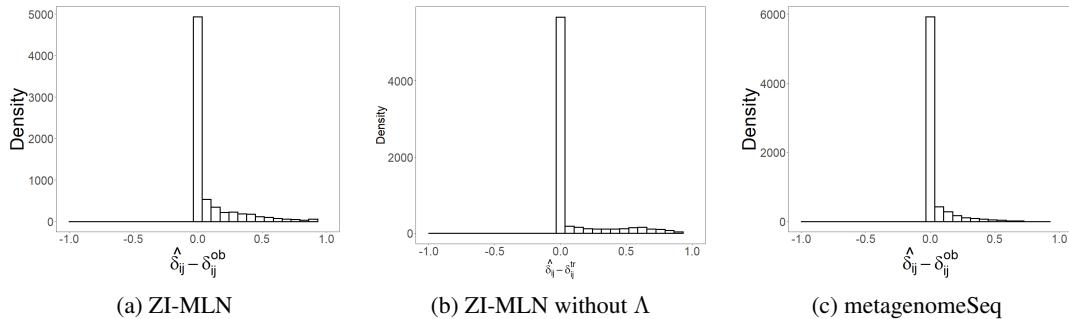


FIG 38. [Human Gut Microbiome Data] Histograms of the differences between $\hat{\delta}_{ij}$ and the observed zero indicators $1(Y_{ij} = 0)$ under ZI-MLN, ZI-MLN without Λ and metagenomeSeq, respectively. Posterior mean estimates are used for ZI-MLN and ZI-MLN without Λ .

OTUs. EdgeR does not provide interval estimates, and Fig 40 (g)-(i) illustrate point estimates only. Fig 41 compares posterior predictive median estimates under ZI-MLN without Λ and mean abundance estimates under metagenomeSeq and edgeR to the observed counts.

REFERENCES

HAARIO, H., SAKSMAN, E. and TAMMINEN, J. (2001). An adaptive Metropolis algorithm. *Bernoulli* 223–242.

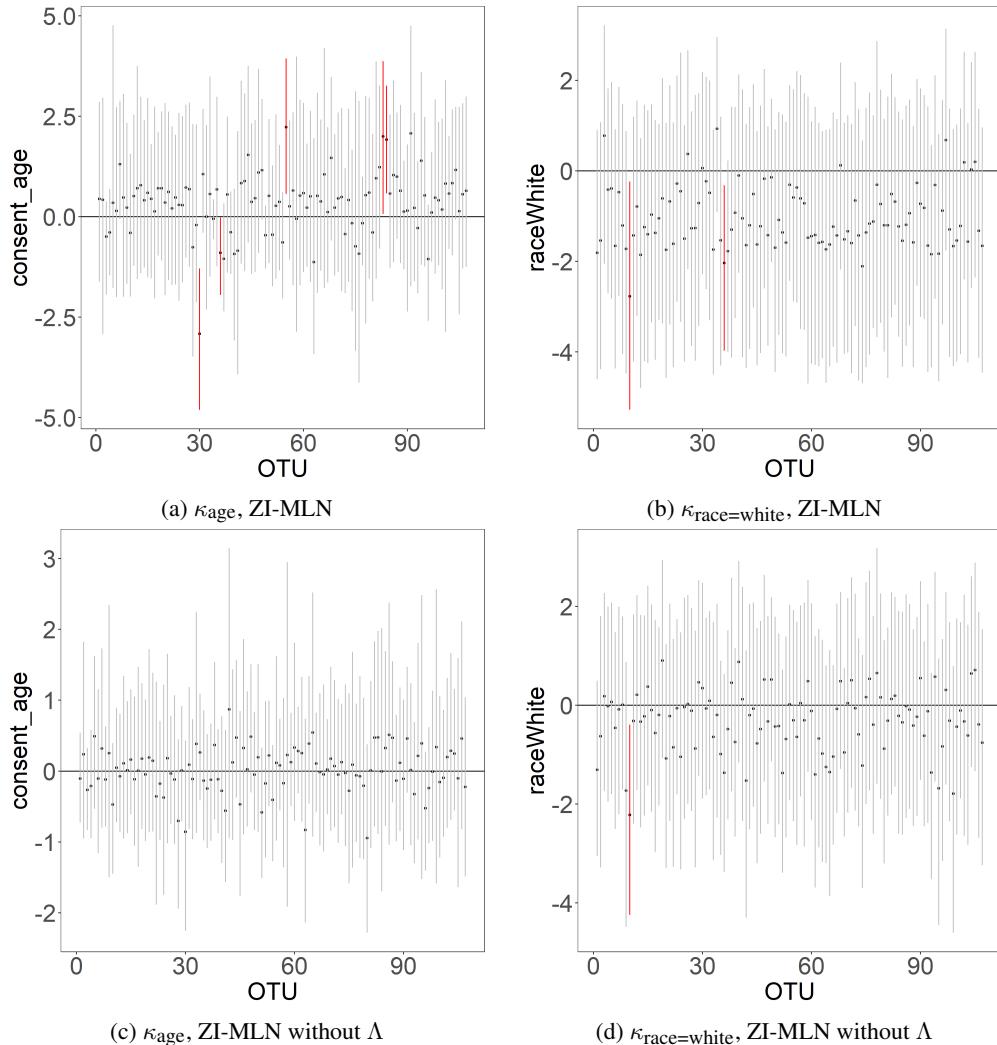


FIG 39. [Human Gut Microbiome Data] Posterior estimates of regression coefficients κ_{age} and $\kappa_{race=white}$ under ZI-MLN and ZI-MLN without Λ for two selected covariates, where black dots are posterior mean estimates with vertical lines for 95% credible intervals. The intervals that do not contain zero are marked in red.

MA, S., REN, B., MALLICK, H., MOON, Y. S., SCHWAGER, E., MAHARJAN, S., TICKLE, T. L., LU, Y., CARMODY, R. N., FRANZOSA, E. A. et al. (2021). A statistical model for describing and simulating microbial community profiles. *PLoS computational biology* **17** e1008913.

PARK, T. and CASELLA, G. (2008). The bayesian lasso. *Journal of the American Statistical Association* **103** 681–686.

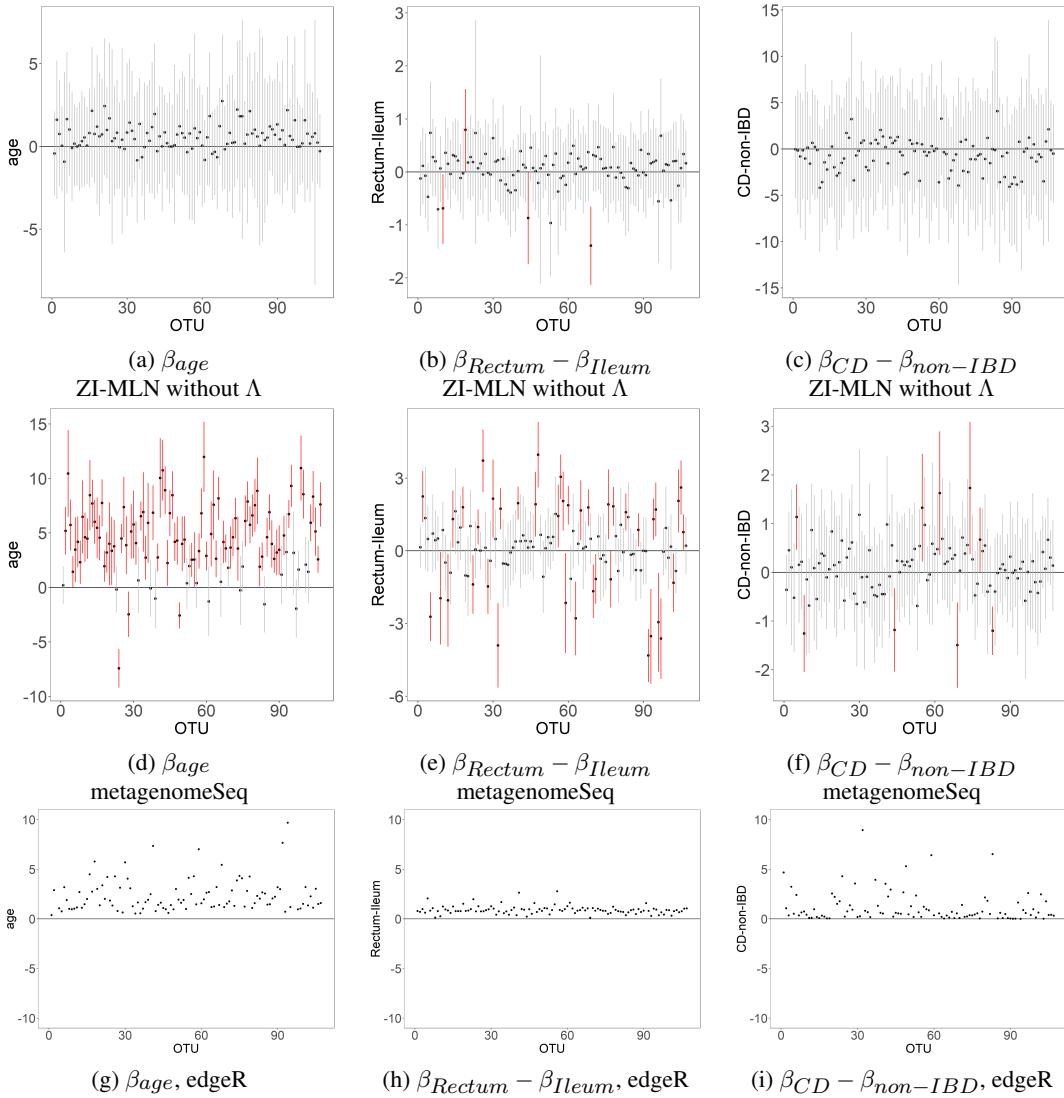


FIG 40. [Human Gut Microbiome Data: Comparison] Posterior mean estimates of β_{jp} under the comparators for some selected covariates. Rows 1-3 are for ZI-MLN without Λ , metagenomeSeq and edgeR, respectively. Black dots and vertical lines represent point estimates and 95% confidence intervals. The intervals that do not contain zero are marked in red.

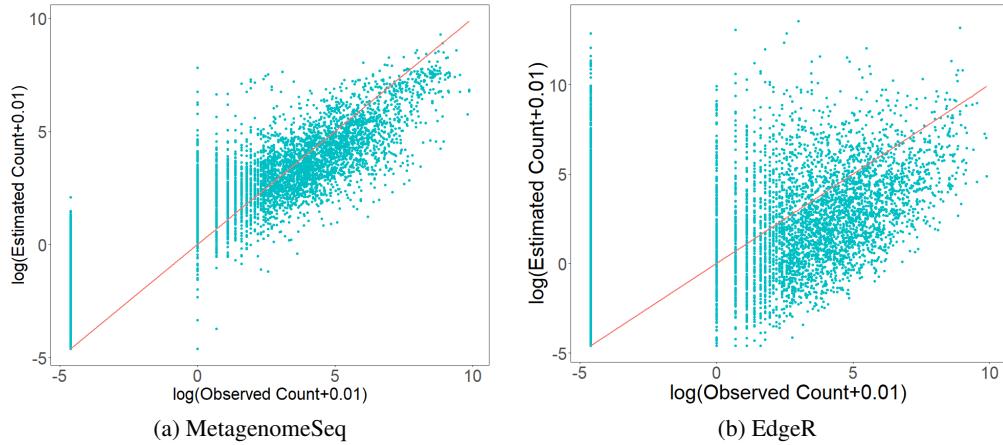


FIG 41. [Human Gut Microbiome Data: Comparison] Panels (a) and (b) present scatter plots of observed $\log(y_{ij} + 0.01)$ versus mean estimated $\log(\hat{y}_{ij} + 0.01)$ by metagenomeSeq and edgeR, respectively.

TABLE 6
[Human Gut Microbiome Data] Taxonomic information for the OTUs illustrated in Fig 12(b) of the main text.

OTU	Taxonomic information (Kingdom/ Phylum/ Class/ Order/ Family / Genus)
4	Bacteria - Firmicutes - Negativicutes -Selenomonadales - Veillonellaceae - Veillonella
30	Bacteria - Firmicutes - Clostridia - Clostridiales - Ruminococcaceae - Faecalibacterium
31	Bacteria - Firmicutes - Erysipelotrichia - Erysipelotrichales - Erysipelotrichaceae - Clostridium innocuum group
37	Bacteria - Proteobacteria - Betaproteobacteria - Neisseriales - Neisseriaceae - Eikenella
39	Bacteria - Firmicutes - Erysipelotrichia - Erysipelotrichales - Erysipelotrichaceae - Erysipelaclclostridium
44	Bacteria - Firmicutes - Clostridia - Clostridiales - FamilyXI - Parvimonas
56	Bacteria - Firmicutes - Clostridia - Clostridiales - FamilyXI - Anaerococcus
59	Bacteria - Bacteroidetes - Bacteroidia - Bacteroidales - Bacteroidaceae - Bacteroides
78	Bacteria - Bacteroidetes - Bacteroidia - Bacteroidales - Rikenellaceae - Alistipes
84	Bacteria - Firmicutes - Clostridia - Clostridiales - Lachnospiraceae - LachnospiraceaeUCG010
85	Bacteria - Firmicutes - Clostridia - Clostridiales - Ruminococcaceae - RuminococcaceaeNK4A214group
93	Bacteria - Proteobacteria - Gammaproteobacteria - Enterobacteriales - Enterobacteriaceae - Escherichia Shigella
96	Bacteria - Firmicutes - Clostridia - Clostridiales - Clostridiaceae1 - Clostridium Sensu Stricto 1

TABLE 7

[Human Gut Microbiome Data] Taxonomic information for the OTUs, for which a 95% posterior credible interval estimate of κ_{jp} does not contain zero for covariates.

OTU	Covariate	Pos mean	95% credible interval	Taxonomic information (Kingdom/ Phylum/ Class/ Order / Family / Genus)
10	white–non-white	-2.77	[-5.28, -0.23]	Bacteria - Firmicutes - Bacilli - Lactobacillales - Streptococcaceae - Streptococcus
26	male–female	-2.52	[-4.86, -0.18]	Bacteria - Firmicutes - Clostridia - Clostridiales - Lachnospiraceae - uncultured
30	age	-2.91	[-4.80, -1.29]	Bacteria - Firmicutes - Clostridia - Clostridiales - Ruminococcaceae - Faecalibacterium
36	age white–non-white	-0.89 -2.03	[-1.95, -0.03] [-3.97, -0.32]	Bacteria - Fusobacteria - Fusobacteriia - Fusobacteriales - Fusobacteriaceae - Fusobacterium
49	Rectum–ileum	-2.32	[-4.25, -0.74]	Bacteria - Bacteroidetes - Flavobacteriia - Flavobacteriales - Flavobacteriaceae - Cloacibacterium
55	age	2.23	[0.57, 3.93]	Bacteria - Firmicutes - Clostridia - Clostridiales - Lachnospiraceae - uncultured
68	MGH –Cincinnati	-2.43	[-4.98, -0.04]	Bacteria - Firmicutes - Erysipelotrichia - Erysipelotrichales - Erysipelotrichaceae - Erysipelatoclostridium
83	age	1.99	[0.07, 3.87]	Bacteria - Firmicutes - Clostridia - Clostridiales - Lachnospiraceae - Coprococcus 1
84	age male–female	1.91 2.29	[0.58, 3.25] [0.65, 4.16]	Bacteria - Firmicutes - Clostridia - Clostridiales - Lachnospiraceae - Lachnospiraceae UCG010
104	male–female	-2.64	[-4.91, -0.53]	Bacteria - Firmicutes - Erysipelotrichia - Erysipelotrichales - Erysipelotrichaceae - Erysipelotrichaceae UCG003

TABLE 8

[Human Gut Microbiome Data] Taxonomic information for the OTUs, for which a 95% posterior credible interval estimate of β_{jp} does not contain zero for covariates

OTU	Covariate	Pos mean	95% credible interval	Taxonomic information (Kingdom/ Phylum/ Class/ Order / Family / Genus)
9	white–non-white	3.06	[0.28, 5.61]	Bacteria - Bacteroidetes - Bacteroidia - Bacteroidales - Bacteroidaceae - Bacteroides
11	CD–non-IBD	-3.65	[-5.66, -1.69]	Bacteria - Firmicutes - Clostridia - Clostridiales - Peptostreptococcaceae - Intestinibacter
12	CD–non-IBD	-2.71	[-4.86, -0.38]	Bacteria - Bacteroidetes - Bacteroidia - Bacteroidales - Bacteroidaceae - Bacteroides
14	CD–non-IBD	-1.80	[-3.61, -0.02]	Bacteria - Actinobacteria - Actinobacteria - Bifidobacteriales - Bifidobacteriaceae - Bifidobacterium
17	CD–non-IBD	-2.32	[-4.53, -0.25]	Bacteria - Firmicutes - Clostridia - Clostridiales - Lachnospiraceae - Roseburia
19	Rectum–ileum	1.24	[0.04, 2.58]	Bacteria - Firmicutes - Clostridia - Clostridiales - Lachnospiraceae - Eubacterium Xylanophilum Group
23	age	1.74	[0.13, 3.65]	Bacteria - Actinobacteria - Actinobacteria - Corynebacteriales - Corynebacteriaceae- Corynebacterium
25	CD–non-IBD	-2.39	[-4.70, -0.33]	Bacteria - Firmicutes - Clostridia - Clostridiales - Lachnospiraceae - Coprococcus1
26	male–female	-2.40	[-4.56, -0.43]	Bacteria - Firmicutes - Clostridia - Clostridiales - Lachnospiraceae - uncultured
32	MGH –Cincinnati	-2.98	[-5.35, -0.47]	Bacteria - Firmicutes - Clostridia - Clostridiales - Lachnospiraceae - Tyzzerella
34	age	-2.24	[-3.57, -0.79]	Bacteria - Bacteroidetes - Bacteroidia - Bacteroidales
	UC–non-IBD	5.57	[1.01, 9.48]	- Prevotellaceae - Prevotella9
37	male–female	2.74	[0.18, 5.06]	Bacteria - Proteobacteria - Betaproteobacteria - Neisseriales - Neisseriaceae - Eikenella
40	MGH –Cincinnati	1.97	[0.19, 3.67]	Bacteria - Firmicutes - Clostridia - Clostridiales - FamilyXIII - FamilyXIIAD3011 group
43	white–non-white	-3.27	[-6.07, -0.35]	Bacteria - Firmicutes - Clostridia - Clostridiales - Lachnospiraceae - NK4A136group
48	MGH –Cincinnati	3.44	[1.74, 5.19]	Bacteria - Firmicutes - Clostridia - Clostridiales - Lachnospiraceae - Lachnoclostridium
49	male–female	3.37	[1.00, 5.87]	
	MGH –Cincinnati	3.04	[0.38, 5.71]	Bacteria - Bacteroidetes - Flavobacteriia - Flavobacteriales
	Rectum–ileum	-1.62	[-3.71, -0.09]	- Flavobacteriaceae - Cloacibacterium
	UC–non-IBD	3.78	[0.72, 7.24]	
52	CD–non-IBD	-2.74	[-4.96, -0.59]	Bacteria - Firmicutes - Clostridia - Clostridiales - Ruminococcaceae - uncultured
55	male–female	-3.51	[-5.84, -1.21]	Bacteria - Firmicutes - Clostridia - Clostridiales - Lachnospiraceae - uncultured
59	male–female	3.92	[1.07, 7.06]	Bacteria - Bacteroidetes - Bacteroidia - Bacteroidales
	MGH –Cincinnati	-4.66	[-9.48, -0.46]	- Bacteroidaceae - Bacteroides
61	UC–non-IBD	2.76	[0.14, 5.31]	Bacteria - Proteobacteria - Gammaproteobacteria - Enterobacteriales - Enterobacteriaceae - Citrobacter
63	MGH –Cincinnati	-2.75	[-4.81, -0.70]	Bacteria - Bacteroidetes - Bacteroidia - Bacteroidales - Porphyromonadaceae - Parabacteroides
64	CD–non-IBD	-2.88	[-5.31, -0.55]	Bacteria - Firmicutes - Clostridia - Clostridiales - Ruminococcaceae - Ruminococcaceae UCG 013
66	UC–non-IBD	-2.54	[-5.22, -0.22]	Bacteria - Verrucomicrobia - Verrucomicrobiae - Verrucomicrobiales - Verrucomicrobiaceae- Akkermansia
68	age	2.19	[0.48, 4.22]	Bacteria - Firmicutes - Erysipelotrichia - Erysipelotrichales - Erysipelotrichaceae - Erysipelatoclostridium
69	Rectum–ileum	-1.46	[-2.46, -0.33]	Bacteria - Firmicutes - Bacilli - Lactobacillales - Carnobacteriaceae - Granulicatella

TABLE 9

*

Tab 8 continued [Human Gut Microbiome Data] Taxonomic information for the OTUs, for which a 95% posterior credible interval estimate of β_{jp} does not contain zero for covariates

OTU	Covariate	Pos mean	95% credible interval	Taxonomic information (Kingdom/ Phylum/ Class/ Order / Family / Genus)
70	CD–non-IBD	-2.05	[-4.43, -0.01]	Bacteria - Firmicutes - Clostridia - Clostridiales - Lachnospiraceae - ND3007group
71	white–non-white	-3.44	[-6.66, -0.56]	Bacteria - Proteobacteria - Deltaproteobacteria - Desulfovibrionales- Desulfovibrionaceae- Bilophila
79	age male–female	1.42 -1.87	[0.09, 2.50] [-3.68, -0.14]	Bacteria - Firmicutes - Clostridia - Clostridiales - Ruminococcaceae - Subdoligranulum
84	age CD–non-IBD	2.05 3.36	[0.29, 4.08] [0.23, 6.22]	Bacteria - Firmicutes - Clostridia - Clostridiales - Lachnospiraceae - Lachnospiraceae UCG 010
85	CD–non-IBD	-3.43	[-7.57, -0.69]	Bacteria - Firmicutes - Clostridia - Clostridiales - Ruminococcaceae - NK4A214group
86	white–non-white	-3.23	[-6.05, -0.24]	Bacteria - Firmicutes - Clostridia - Clostridiales - Ruminococcaceae - Ruminiclostridium5
87	CD–non-IBD	-2.17	[-4.14, -0.15]	Bacteria - Firmicutes - Clostridia - Clostridiales - Ruminococcaceae - Subdoligranulum
89	CD–non-IBD	-2.90	[-5.08, -0.90]	Bacteria - Firmicutes - Clostridia - Clostridiales - Lachnospiraceae - Eubacterium Ventriosum Group
90	CD–non-IBD	-3.30	[-5.55, -0.68]	Bacteria - Firmicutes - Clostridia - Clostridiales - Lachnospiraceae - Lachnospira
91	white–non-white CD–non-IBD	3.99 -3.33	[0.07, 8.68] [-5.68, -0.67]	Bacteria - Firmicutes - Clostridia - Clostridiales - Clostridiaceae1 - Clostridium Sensu Stricto 1
92	CD–non-IBD	-3.41	[-5.94, -1.02]	Bacteria - Firmicutes - Clostridia - Clostridiales - Ruminococcaceae - Ruminiclostridium 6
93	MGH –Cincinnati	-2.76	[-4.70, -0.77]	Bacteria - Proteobacteria - Gammaproteobacteria - Enterobacteriales - Enterobacteriaceae - Escherichia Shigella
94	age white–non-white	1.68 -4.38	[0.17, 3.44] [-7.75, -1.18]	Bacteria - Firmicutes - Clostridia - Clostridiales - Ruminococcaceae - Ruminococcus 1
96	MGH –Cincinnati	-4.83	[-8.79, -1.15]	Bacteria - Firmicutes - Clostridia - Clostridiales - Clostridiaceae1 - Clostridium Sensu Stricto 1
103	CD–non-IBD	-3.47	[-6.14, -0.06]	Bacteria - Firmicutes - Clostridia - Clostridiales - Lachnospiraceae - Eubacterium eligens group
104	male–female MGH –Cincinnati	-2.36 1.92	[-4.28, -0.51] [0.07, 3.99]	Bacteria - Firmicutes - Erysipelotrichia - Erysipelotrichales - Erysipelotrichaceae - Erysipelotrichaceae UCG 003