BMC Bioinformatics

RESEARCH

Open Access

Analyzing microbiome data with taxonomic misclassification using a zero-inflated Dirichlet-multinomial model

Matthew D. Koslovsky[1*]

*Correspondence:
matt.koslovsky@colostate.edu

[1] Department of Statistics, Colorado State University, Fort Collins, CO, USA

Abstract

The human microbiome is the collection of microorganisms living on and inside of our bodies. A major aim of microbiome research is understanding the role microbial communities play in human health with the goal of designing personalized interventions that modulate the microbiome to treat or prevent disease. Microbiome data are challenging to analyze due to their high-dimensionality, overdispersion, and zero-inflation. Analysis is further complicated by the steps taken to collect and process microbiome samples. For example, sequencing instruments have a fixed capacity for the total number of reads delivered. It is therefore essential to treat microbial samples as compositional. Another complicating factor of modeling microbiome data is that taxa counts are subject to measurement error introduced at various stages of the measurement protocol. Advances in sequencing technology and preprocessing pipelines coupled with our growing knowledge of the human microbiome have reduced, but not eliminated, measurement error. Ignoring measurement error during analysis, though common in practice, can then lead to biased inference and curb reproducibility. We propose a Dirichlet-multinomial modeling framework for microbiome data with excess zeros and potential taxonomic misclassification. We demonstrate how accommodating taxonomic misclassification improves estimation performance and investigate differences in gut microbial composition between healthy and obese children.

Keywords: Compositional, High-dimensional, Multivariate count data, Obesity

Introduction

Communities of microorganisms, referred to as microbiomes, are found almost everywhere on Earth, including on and inside our bodies, plants, animals, soil, oceans, and the atmosphere. Improving our understanding of the role of microbiota and their interactions with their hosts and other microbes has implications for a variety of fields, including human health and nutrition, medicine, ecology, agriculture, forensics, and exobiology. Microbiome datasets typically take the form of an $N \times T$ matrix of counts, where N represents the number of observations and T represents the number of unique microbial taxa. The conventional approach for obtaining taxa counts is to sequence the 16S rRNA gene, as it contains well-conserved and hypervariable regions to differentiate

different species [1]. Then, sequenced reads are clustered into operational taxonomic units, or OTUs, using 97% or 99% similarity thresholds and classified to a reference database (e.g., GreenGenes, SILVA, RDP, NCBI) using various methods [2–9]. More recently, researchers have promoted the use of amplicon sequence variants (ASVs) instead of OTUs as the unit of analysis in microbiome research [10–13]. These denoising approaches provide exact sequence variants instead of clustering sequences into OTUs to account for sequencing errors and can distinguish sequence variants differing by as little as one nucleotide. The strength of ASVs is that they provide a higher resolution to the data, are consistent labels which can be compared across studies, and have demonstrated equivalent or improved sensitivity and specificity as OTU-based methods [13]. Microbial count datasets are then used to generate inference for a variety of research aims, including estimating the relative abundances of microorganisms present in the host, determining whether species abundances vary across groups (latent or observed), and/or training models to predict phenotypic outcomes using information contained in the composition of the microbiome, among others [14–24].

Microbiome data are inherently challenging to analyze due to their high-dimensionality (i.e., 100s or even 1000s of taxa), overdispersion (i.e., large within- and between-subject variability), and zero-inflation (i.e., larger number of zero reads observed than expected under distributional assumptions). Analysis is further complicated by the steps taken to collect and process microbiome samples [25, 26]. For example, sequencing instruments have a fixed capacity for the total number of reads delivered. It is therefore essential to treat microbial samples as *compositional* multivariate count data, where the relative abundances of the taxa sum to one [27]. Another complicating factor of modeling microbiome data is that taxa counts are subject to measurement error introduced at various stages of the measurement protocol [28–30]. For example, clustered sequencing reads are subject to misclassification due to sequencing errors, and taxonomic allocation is sensitive to the clustering method and reference database used. Even in ASV studies, bacterial genomes may have multiple 16S rRNA genes that are not identical, which could lead to splitting a single genome into multiple clusters, and the 16S rRNA gene may not contain all necessary genetic variation underlying ecological and evolutionary differences to differentiate between species [13, 31, 32]. While it is now standard for microbiome analyses to accommodate high-dimensionality, overdispersion, zero-inflation, and the compositional structure characteristic of microbial data (though oftentimes not simultaneously), potential misclassification of the observed reads is typically ignored, which can bias downstream inference, underestimate parameter uncertainty, and curb reproducibility. In this work, we introduce a scalable Bayesian modeling framework that simultaneously accommodates the aforementioned challenges in microbiome data analysis.

In practice, zero reads in microbiome data can occur in two different ways: (1) the organism is not present in the sampling region and therefore the probability of occurrence is zero (i.e., structural zero) and (2) the organism is present but was not sampled (i.e., at-risk zero). A common technique for modeling zero-inflation in microbial count data is to construct a two-component mixture of a point mass at zero and a sampling distribution for the counts (e.g., Poisson or negative binomial distributions in the univariate setting), where a latent at-risk indicator is introduced to differentiate between

at-risk and structural zeros [16, 21, 23, 33–35]. To model zero-inflation in multivariate count data, researchers typically link zero-inflated univariate count models together by embedding latent parameters which control the dependence structure between counts [36, 37]. [38] recently introduced a zero-inflated Dirichlet-multinomial (ZIDM) model for handling excess zeros in multivariate compositional count data collected in microbiome research settings, which differs from previous methods for modeling zero-inflation by assuming a mixture distribution on the count probabilities as opposed to the sampling distribution.

While potential misclassification of microbial taxa is typically ignored in microbiome studies, researchers have investigated the effects of measurement error on inference [29]. For example, [28] demonstrate how measurement error in 16S rRNA gene studies can bias inference and diminish the replicability of measured differences between samples. [35] model microbial counts with a zero-inflated Poisson-gamma model, which introduces a multiplicative factor for the rate parameter in the Poisson distribution to accommodate the deviation of observed abundances to unobserved true abundances. [39] propose a log-error-in-variable regression model for handling measurement error in compositional regression settings (i.e., when microbiome data are treated as covariates).

Outside of microbiome research studies, there are numerous existing approaches for modeling misclassification in multivariate count data which have been applied in various settings [40–46]. The typical approach for modeling potential misclassification is to assume the observed classification of each observation follows a multinomial distribution given the true (latent) classification. Recently, this technique has been used to model potential misclassification in ecological multispecies occupancy models [44–46]. However, these methods are not designed to accommodate the compositional and/or high-dimensional structure of multivariate count data found in microbiome research settings.

A major challenge in modeling misclassification in count data is that the model is not identifiable without additional information about the misclassification process beyond the observed data [40]. In some research settings, validated or true classifications may be available for a subset of the data which can be used to inform misclassification rates in the model in a semi-supervised setting [47]. When validation data are not available, Bayesian methods are commonly preferred over frequentist alternatives to incorporate prior information regarding misclassification rates [40, 42, 47].

In this study, we investigate differences in the gut microbial composition of obese and healthy children and adolescents [48]. To this end, we propose a scalable zero-inflated Dirichlet-multinomial regression model which accommodates potential misclassification to investigate the relation between covariates and the true, unobserved microbial counts. Specifically, we take a hierarchical approach that assumes the true microbial abundances follow a zero-inflated Dirichlet-multinomial distribution which incorporates covariate associations with the true taxa counts and the at-risk probabilities. We then construct a confusion matrix to model the probability of the observed microbial taxa given the learned, true classifications. To accommodate high-dimensional settings, we extend the model with sparsity-inducing priors to identify covariates associated with the probability of an at-risk observation and the microbial taxa. In addition to providing relative abundance estimates that accommodate classification uncertainty for

downstream analysis, our analysis reveals key insights into the relation between obesity and microbial composition.

The rest of this work is organized as follows. In Sect. Model, Notation, and Inference we introduce our framework for modeling zero-inflation and misclassification in microbiome data and discuss recommendations for prior specification and posterior inference. Section Synthetic Data Evaluation provides a simulation study to evaluate how the proposed modeling framework performs in the presence of varying levels of misclassification, overdispersion, and sparsity. In Sect. The Effect of Obesity on the Composition of the Human Microbiome we investigate the relation between obesity and the human gut microbiome using our modeling framework. We provide concluding remarks in Sect. Conclusions.

Model, notation, and inference

In this section, we first introduce pertinent notation and then detail our approach for modeling potential misclassification of and zero-inflation in microbial count data, which we refer to as MicroMiss. Thereafter, we discuss posterior sampling and inference.

Let the C-dimensional vector $\boldsymbol{y}_{il}$ represent the observed taxon classification for the l^{th}, $l = 1, \ldots, \dot{z}_i$, organism of the i^{th}, $i = 1, \ldots, N$, host, where $y_{ilc} = 1$ indicates the observed organism was classified as the c^{th} taxon and 0 otherwise. Let the T-dimensional vector $\boldsymbol{z}_{il}$ represent the organism's true classification, where $z_{ilt} = 1$ indicates the organism truly belongs to the t^{th} taxon group and 0 otherwise. In this analysis, we assume $T = C$, and the ordering of the taxa is the same in $\boldsymbol{y}_{il}$ and $\boldsymbol{z}_{il}$.

To model the true (latent) classifications of each organism, we assume

$$\boldsymbol{z}_{il} | \boldsymbol{\Theta}_i \sim \text{Multinomial}(1, \boldsymbol{\Theta}_i), \tag{1}$$

where $\boldsymbol{\Theta}_i$ is a host-specific T-dimensional vector of true relative abundances. A standard approach for accommodating overdispersion in microbiome studies is to assume the relative abundances follow a Dirichlet distribution [15, 49]. Specifically, we let $\boldsymbol{\Theta}_i \sim \text{Dirichlet}(\boldsymbol{\gamma}_i)$, where $\boldsymbol{\gamma}_i$ represents a T-dimensional vector of concentration parameters. We introduce covariates into the model for the relative abundances by setting $\log(\gamma_{it}) = \boldsymbol{x}_i' \boldsymbol{\beta}_{\gamma_t}$ with $\beta_{\gamma_{tp}} \sim \text{Normal}(\mu_\gamma, \sigma^2_\gamma)$ and $\boldsymbol{x}_i$ a P-dimensional, $p = 1, \ldots, P$, host-specific set of covariates including an intercept term. When there is a large number of taxa and/or covariates, researchers typically induce sparsity on their relations [15, 38, 50]. To accommodate sparse settings, we alternatively place spike-and-slab priors on $\boldsymbol{\beta}_{\gamma_t}$, similar to [15, 18, 24, 38, 51]. Specifically, we let $\beta_{\gamma_{tp}} | \zeta_{\gamma_{tp}}, \sigma^2_\gamma \sim \zeta_{\gamma_{tp}} \cdot \text{Normal}(0, \sigma^2_\gamma) + (1 - \zeta_{\gamma_{tp}}) \cdot \delta_0(\beta_{\gamma_{tp}})$, where $\zeta_{\gamma_{tp}} \in \{0, 1\}$ is a latent inclusion indicator and $\delta_0(\cdot)$ is a Dirac delta function, or point mass, at 0. We then assume $\zeta_{\gamma_{tp}} \sim \text{Beta-Binomial}(a_\gamma, b_\gamma)$, where a_γ and b_γ can be set to impose various levels of sparsity in the model. Since the regression coefficients are taxon-specific, this modeling framework provides inference on the association between covariates and each taxon. However, inference on a covariate's association with a taxon's relative abundance is not straightforward, as it is indirectly modeled through the concentration parameters. We discuss this in more detail in Sect. Posterior Sampling and Inference.

In order to accommodate zero-inflation in the Dirichlet distribution, we reparameterize $\boldsymbol{\Theta}_i$ as a set of independent, zero-inflated gamma random variables, $\boldsymbol{\alpha}_i$, normalized by their sum (i.e., $\boldsymbol{z}_{il}|\boldsymbol{\alpha}_i \sim \text{Multinomial}(1, \frac{\boldsymbol{\alpha}_i}{\bar{\alpha}_i})$), where

$$\alpha_{it}|\zeta_{it}, \gamma_{it} \sim \zeta_{it}\ \text{Gamma}\ (\gamma_{it}, 1) + (1 - \zeta_{it})\delta_0(\alpha_{it}), \tag{2}$$

$\bar{\alpha}_i = \sum_{t=1}^{T} \alpha_{it}$, and ζ_{it} is a latent at-risk indicator that differentiates between at-risk observations and structural zeros, similar to [38]. We assume the latent at-risk indicators $\zeta_{it}|\boldsymbol{\beta}_{\eta_t}, \boldsymbol{x}_i \sim \text{Bernoulli}\ (\eta_{it})$, where $\text{logit}(\eta_{it}) = \boldsymbol{x}_i'\boldsymbol{\beta}_{\eta_t}$, $\boldsymbol{x}_i$ is a set of host-specific covariates including an intercept term, and $\boldsymbol{\beta}_{\eta_t}$ are the corresponding taxon-specific regression coefficients. Here, η_{it} is interpreted as the at-risk probability, and $\boldsymbol{\beta}_{\eta_t}$ represent log odds ratios for an at-risk observation which are taxon specific. We let $\beta_{\eta_{tp}} \sim \text{Normal}\ (\mu_\eta, \sigma_\eta^2)$. Similarly, we can induce sparsity in $\boldsymbol{\beta}_{\eta_t}$ by assuming $\beta_{\eta_{tp}}|\zeta_{\eta_{tp}}, \sigma_\eta^2 \sim \zeta_{\eta_{tp}} \cdot \text{Normal}(0, \sigma_\eta^2) + (1 - \zeta_{\eta_{tp}}) \cdot \delta_0(\beta_{\eta_{tp}})$, where $\zeta_{\eta_{tp}} \in \{0, 1\}$ is a latent inclusion indicator. We then assume $\zeta_{\eta_{tp}} \sim \text{Beta-Binomial}(a_\eta, b_\eta)$ and refer to the sparsity-induced version of the model as MicroMissS.

To model the observed reads, we assume

$$\boldsymbol{y}_{il}|\boldsymbol{\theta}_t, z_{ilt} = 1 \sim\ \text{Multinomial}\ (1, \boldsymbol{\theta}_t), \tag{3}$$

where $\boldsymbol{\theta}_t$ is a C-dimensional vector of observed taxa probabilities. We assume the observed classification probabilities depend on the true taxon classification of each organism, $z_{ilt} = 1$, with $\boldsymbol{\theta}_t \sim\ \text{Dirichlet}\ (\boldsymbol{\nu}_t)$, where $\boldsymbol{\nu}_t$ is a C-dimensional vector of taxon-specific concentration hyperparameters. For notational purposes, let $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \ldots, \boldsymbol{\theta}_T')'$ represent a confusion matrix which maps the potential misclassification in the data. Since read-level validation data are not available to inform $\boldsymbol{\theta}$, prior information regarding taxonomic misclassification will be incorporated through the specification of $\boldsymbol{\nu}_t$.

Prior specification for misclassification

In this section, we discuss different approaches for specifying the hyperparameters $\boldsymbol{\nu}_t$. Under the proposed modeling framework, the probability of correct classification of the t^{th} taxon a priori is $\nu_{tt}/\sum_{c=1}^{C} \nu_{tc}$. Thus, a simple and intuitive way of specifying $\boldsymbol{\nu}_t$ is to set each $\nu_{tc} = 1$ and increase ν_{tt} to the desired probability of correct classification. For example, if $C = 51$ and $\nu_{tt} = 50$, then the probably of correct classification would be 0.50, and the probability of misclassifying a t^{th} taxon as a c^{th} taxon for $t \neq c$ is 0.01. An alternative approach is to incorporate prior knowledge when assigning classification probabilities. For example, if the OTUs are aggregated at the genus level, one may place a very small probability of misclassification to genera that belong to different families. Caution is advised when taking this approach as it assumes some level of accuracy for OTU classification as well as the taxonomic structure, which is somewhat contradictory as the classifications are assumed to contain errors. Note that by assuming $\boldsymbol{\theta}_t$ is shared across all observations, the model borrows information across observations to inform the misclassification probabilities.

Posterior sampling and inference

For inference, we implement a Metropolis-Hastings within Gibbs algorithm to sample the resulting posterior distribution, which is outlined below in Algorithm 1 and detailed

in the Supplementary Information. The full joint distribution (see Fig. 1 for a graphical representation) is defined as

$$\prod_{i=1}^{N}\left[p(\mu_i|\bar{\alpha}_i)\prod_{l=1}^{\dot{z}_i}p(\boldsymbol{y}_{il}|\boldsymbol{\theta}_t,\boldsymbol{z}_{il})p(\boldsymbol{z}_{il}|\boldsymbol{\Theta}_i)\prod_{t=1}^{T}p(\alpha_{it}|\zeta_{it},\boldsymbol{\beta}_{\gamma_t},\boldsymbol{x}_i)p(\zeta_{it}|\boldsymbol{\beta}_{\eta_t},\boldsymbol{x}_i)p(\omega_{\zeta_{it}})\right] \times\prod_{t=1}^{T}\left[p(\boldsymbol{\beta}_{\eta_t})p(\boldsymbol{\beta}_{\gamma_t})p(u_t|\bar{a}_t)\prod_{c=1}^{C}p(a_{tc})\right]. \tag{4}$$

For efficient sampling, we introduce auxiliary parameters $\mu_i|\bar{\alpha}_i \sim$ Gamma $(\dot{z}_i, \bar{\alpha}_i)$ and a latent set of auxiliary parameters $\omega_{\zeta_{it}} \sim$ PG $(1,0)$, where PG represents a Pólya-gamma distribution, following [38] and [52], respectively. Additionally, we reparameterize $\theta_{tc} = a_{tc}/\bar{a}_t$ and assume $a_{tc} \sim$ Gamma $(\nu_{tc}, 1)$ with auxiliary parameter $u_t \sim$ Gamma $(\sum_{i=1}^{N}\sum_{l=1}^{\dot{z}_i} I(z_{ilt}=1), \bar{a}_t)$ and $\bar{a}_t = \sum_{c=1}^{C} a_{tc}$. This enables efficient sampling of $\boldsymbol{\theta}_t$ using a similar data augmentation approach as that introduced for $\boldsymbol{\Theta}_i$.

In microbiome studies, it is common to observe millions of reads for a given sample. Since we model potential misclassification at the read level, our approach can quickly encounter storage and computing limitations when applied to even moderately sized datasets. To improve the scalability of the model, we propose block updates for the latent classifications with corresponding $y_{ilc} = 1$, which we denote $\boldsymbol{z}_{ic}$, instead of updating the latent classifications individually. Specifically, we update the latent indicators at the observation level from a Multinomial$(\sum_{l=1}^{\dot{z}_i} I(y_{ilc}=1), \boldsymbol{\Theta}_i \otimes \boldsymbol{\theta}_c)$, where $I(\cdot)$ is an indicator function and $\otimes$ represents an element-wise product. Let the vector of latent classifications $\boldsymbol{z}_i = \sum_{c=1}^{C} \boldsymbol{z}_{ic}$. In settings where the true relative abundances and/or

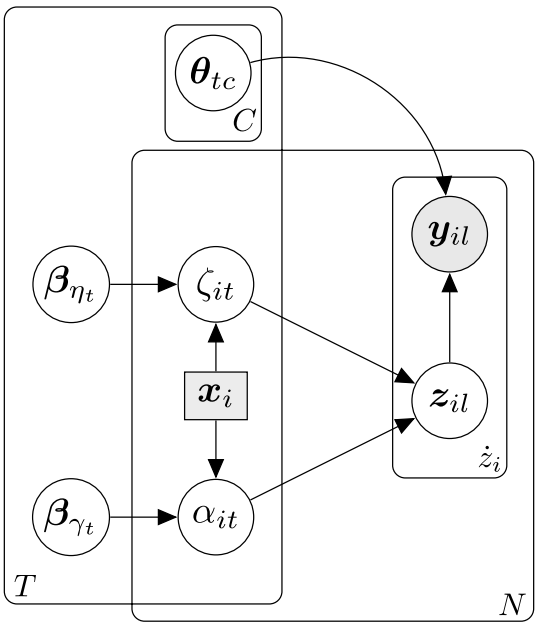

Fig. 1 Graphical representation of the proposed MicroMiss model. Note that auxiliary parameters and hyperparameters have been suppressed for clarity. ***N*** - total observations; $\dot{\boldsymbol{z}}_i$ - total reads per observation; ***T*** - true number of taxa; ***C*** - observed number of taxa.

the classification probabilities are modeled at the read level, it is no longer possible to aggregate the latent classifications without requiring substantial computing resources or resorting to approximate techniques for inference.

The main outcomes of interest in this analysis are the estimated relations between observed covariates and the probability of an at-risk observation as well as the true relative abundances. In the proposed model, $\beta_{\eta_{tp}}$ is interpreted as the expected change in log odds ratio of an at-risk observation for a one unit increase in the corresponding covariate holding all else constant, and $\exp(\beta_{\gamma_{tp}})$ is interpreted as the multiplicative change in the concentration parameter γ_t for a unit increase in x_p holding all else constant. While oftentimes overlooked in practice, the relation between covariates and relative abundances is more complicated, due to the fact that covariates are potentially related to each compositional element, as described in [53]. Thus the multiplicative effect on the t^{th} relative abundance for a one unit increase in the p^{th} covariate for the i^{th} observation is defined as

$$\pi_t(\boldsymbol{x}_{ip}) = \frac{\Theta_t(\boldsymbol{x}_i^{(p)})}{\Theta_t(\boldsymbol{x}_i)} = \exp(\beta_{\gamma_{tp}}) \frac{\sum_{s=1}^{T} \exp(\boldsymbol{x}_i' \boldsymbol{\beta}_{\gamma_s})}{\sum_{s=1}^{T} \exp(\boldsymbol{x}_i^{(p)\prime} \boldsymbol{\beta}_{\gamma_s})}, \tag{5}$$

where $\boldsymbol{x}_i^{(p)} = (x_{i1}, x_{i2}, \ldots, x_{ip} + 1, \ldots, x_{iP})'$. It is clear that the effect of the p^{th} covariate on the t^{th} relative abundance depends not only on its corresponding regression coefficient, $\beta_{\gamma_{tp}}$, but also its effect on the other relative abundances and their corresponding concentration parameters. As such, we may observe a decrease (increase) in the t^{th} relative abundance with an increase in x_p, even if $\beta_{\gamma_{tp}} > 0$ ($\beta_{\gamma_{tp}} < 0$). In sparse settings where x_p is associated with *only* the t^{th} compositional element, then the direction of $\beta_{\gamma_{tp}}$ matches the multiplicative effect on the corresponding relative abundance. For inference on these quantities, the posterior means of the MCMC samples are calculated and 95% credible intervals are constructed using the empirical quantiles. Estimates of the true relative abundances for each observation can be obtained by normalizing the vector $\boldsymbol{\alpha}_i$ over its sum for each MCMC iteration and the averaging over samples. Additionally, estimates of the confusion matrix $\boldsymbol{\theta}$ can be obtained similarly given $\boldsymbol{a}_t$.

Algorithm 1 MCMC sampler for the MicroMiss model

Input data $\boldsymbol{y}_{il}$, $\boldsymbol{x}_i$.
Initialize parameters: $\boldsymbol{z}_{il}$, $\boldsymbol{\beta}_{\eta_t}$, $\boldsymbol{\beta}_{\gamma_t}$, $\boldsymbol{\omega}_{\zeta_{it}}$, $\boldsymbol{\alpha}_i$, ζ_{it}, μ_i, $\boldsymbol{a}_t$, u_t.
Specify hyperparameters: $\mu_\eta, \mu_\gamma, \sigma^2_\eta, \sigma^2_\gamma, \boldsymbol{\nu}_t$.
for iteration $m = 1, \ldots, M$ **do**
  for $i = 1, \ldots, N$ **do**
    Update $\mu_i \sim \text{Gamma}(\dot{z}_i, \bar{\alpha}_i)$
    Update $\boldsymbol{z}_{ic} \sim \text{Multinomial}(\sum_{l=1}^{\dot{z}_i} I(y_{ilc} = 1), \boldsymbol{\Theta}_i \otimes \boldsymbol{\theta}_c)$
    for $t = 1, \ldots, T$ **do**
      Jointly update α_{it} and ζ_{it} with an Expand/Contract Step via [38]
      Update $\alpha_{it}|\zeta_{it} \sim \text{Gamma}\left(\sum_{l=1}^{\dot{z}_i} I(z_{ilt} = 1) + \gamma_{it}, 1 + \mu_i\right)$
      Update $\omega_{\zeta_{it}} \sim \text{PG}(1, \boldsymbol{x}_i' \boldsymbol{\beta}_{\eta_t})$
    end for
  end for
  for $t = 1, \ldots, T$ **do**
    Update $u_t \sim \text{Gamma}(\sum_{i=1}^{N} \sum_{l=1}^{\dot{z}_i} I(z_{ilt} = 1), \bar{a}_t)$
    Update $\boldsymbol{\beta}_{\eta_t} \sim N(\tilde{\boldsymbol{\beta}}_{\eta_t}, \tilde{\Sigma}_{\eta_t})$
    Update $\boldsymbol{\beta}_{\gamma_t}$ via Metropolis-Hastings step
    for $c = 1, \ldots, C$ **do**
      $a_{tc} \sim \text{Gamma}(\sum_{i=1}^{N} \sum_{l=1}^{\dot{z}_i} I(y_{ilc} = 1, z_{ilt} = 1\ ) + \nu_{tc}, u_t + 1)$
    end for
  end for
end for

Synthetic data evaluation

In this section, we investigate how ignoring misclassification in multivariate compositional count data can bias inference for the relation between observed covariates and relative abundances. Data are generated similar in structure to the application study with varying levels of misclassification, overdispersion, and sparsity. We compare the estimation results of our approach, MicroMiss, to those obtained with a naive zero-inflated Dirichlet-multinomial (ZIDM) regression model, which accommodates excess zeros but not misclassification [38], and a Dirichlet-multinomial regression model [54]. Additionally, we compare MicroMissS to three alternative Bayesian methods designed for sparse settings: ZIDMbvs [38], DMBVS [15], and a zero-inflated negative binomial regression model with sparsity-inducing priors (ZINB) [21]. The MicroMiss, ZIDM, and DM models with and without variable selection priors were implemented in `R` using `Rcpp` to improve computation time [55].

Specifically, we generated $N = 50$ observations of $\dot{z}_i = 10{,}000$ total reads to classify into $C = 50$ taxa groups. We assumed that the true number of compositional elements, T, matches the potentially observed number of taxa, C. Observation-specific at-risk indicators were sampled from a Bernoulli distribution with the at-risk probabilities set to $\exp(\beta_{\eta_{t0}} + \beta_{\eta_{t1}} x_{i1})/(1 + \exp(\beta_{\eta_{t0}} + \beta_{\eta_{t1}} x_{i1}))$, where x_{i1} was generated from a standard normal for each observation. The true classification of each read was generated from a $\text{Multinomial}(1, \boldsymbol{\psi}_i^*)$, where $\boldsymbol{\psi}_i^* \sim \text{Dirichlet}(\boldsymbol{\gamma}_i^*)$, $\gamma_{it}^* = \frac{\gamma_{it}}{\sum_{s=1}^{T} \gamma_{is}} \frac{1-d}{d}$, $\gamma_{it} = \exp(\beta_{\gamma_{t0}} + \beta_{\gamma_{t1}} x_{i1})$, and overdispersion parameter $d \in \{0.01, 0.05, 0.10\}$ so that the model assumptions did not match the true data generation process. We set 50% of the

covariate associations to zero in both levels of the model. When active, the corresponding regression coefficients $\beta_{\eta_{t1}}$ and $\beta_{\gamma_{t1}}$ were set to ± 1 with equal probability. The intercept terms $\beta_{\eta_{t0}} = \beta_{\gamma_{t0}} = 0$. To demonstrate how the model performs in sparse settings, we generated data similar to the above, but instead only allowed regression coefficients $\beta_{\eta_{t1}}$ and $\beta_{\gamma_{t1}}$ for $t = 1, \ldots, 10$ to be non-zero with 0.50 probability. The data were generated with $T = C = 50$ and $T = C = 250$ to further assess the scalability of the methods. In all settings, the observed classifications were generated from a Dirichlet-multinomial model with a similar overdispersion parameter as above. We set the off-diagonal elements of the concentration parameter matrix, $\nu_{tc} = 1$. The diagonal elements were then set to mimic varying levels of misclassification for each compositional element (i.e., $\nu_{tt} = p_{tt} * T/(1 - p_{tt})$), where p_{tt} is the assumed probability of correct classification. This assumes that the true counts for a compositional element were misclassified with equal probability.

All methods were run for 5000 iterations treating the first 2500 as burn-in and thinning to every 5^{th} iteration, providing 500 MCMC iterations for inference. We assumed non- or weakly-informative priors with $\mu_\eta = \mu_\gamma = 0$ and $\sigma^2_\eta = \sigma^2_\gamma = 5$. We set the hyperparameters $\nu_{tt} = 100$ and $\nu_{tc} = 1$, representing around a 0.70 probability of correct classification a priori when $C = 50$. To initialize each model, we set the true counts $\boldsymbol{z}_i$ to the observed counts $\boldsymbol{y}_i$. Regression coefficients were initialized at zero with auxiliary parameters $\omega_{\zeta_{it}}$ set to one. The at-risk indicators ζ_{it} and α_{tc} were also set to one. Auxiliary parameters u_t and μ_i were randomly initialized from a Gamma(1,1).

We evaluated the models' estimation performance in four settings. In the first, we assumed no misclassification, which we refer to as the *Null* setting. We then investigated the model with increasing levels of misclassification. Specifically, we assumed each taxon was misclassified with 0.05 probability (*Low*), 20% of the observations were misclassified with 0.25 or 0.15 probability and the remaining 60% with 0.05 probability (*Medium*), and 20% of the observations were misclassified with 0.15 or 0.05 probability and the remaining 60% with 0.25 probability (*High*). In each setting, the models were evaluated in terms of the average absolute value of the bias for the regression coefficients and 0.95 coverage probabilities in the at-risk portion of the model and the concentration parameters. Results were separated by active (i.e., $\beta_{\eta_{t1}}, \beta_{\gamma_{t1}} \neq 0$) and non-active (i.e., $\beta_{\eta_{t1}}, \beta_{\gamma_{t1}} = 0$) regression coefficients. Additionally, we evaluated the models' average absolute value of the bias and coverage probability for $\pi_t(\boldsymbol{x}_{ip})$ in Eq. 5 and computation time. The simulated average $\pi_t(\boldsymbol{x}_{ip})$ was 1.13, ranging from around 0.15 to 5. Results we report below were obtained by averaging over 50 replicated datasets for each setting (Table 1).

In the *Null* setting, we observed similar performance between MicroMiss and ZIDM, with the proposed approach having slightly better estimation performance for the effects on the concentration parameters and $\pi_t(\boldsymbol{x}_{ip})$. The DM regression model performed considerably worse when estimating the non-zero regression coefficients associated with the concentration parameters and $\pi_t(\boldsymbol{x}_{ip})$, since it ignores potential zero-inflation. As the misclassification probabilities increased, the estimation performance of the models decreased for non-zero regression coefficients in the at-risk portion of the model and concentration parameters, with the proposed method always outperforming ZIDM and DM. However, ZIDM's estimation performance for zero effects on the at-risk probabilities improved with higher misclassification. A similar

Table 1 Simulation results: estimation performance for $N = 50$ observations, $\dot{z}_i = 10{,}000$ reads, $T = 50$ compositional elements at varying levels of misclassification, and with overdispersion parameter $d = 0.01$

Null										
	$\beta_{\eta_{t1}} \neq 0$		$\beta_{\eta_{t1}} = 0$		$\beta_{\gamma_{t1}} \neq 0$		$\beta_{\gamma_{t1}} = 0$		$\pi_t(\boldsymbol{x}_{ip})$	
	ABS	COV	ABS	COV	ABS	COV	ABS	COV	ABS	COV
MicroMiss	0.352	0.945	0.271	0.929	0.115	0.685	0.144	0.761	0.115	0.886
ZIDM	0.350	0.946	0.270	0.934	0.261	0.543	0.209	0.621	0.151	0.565
DM	–	–	–	–	0.796	0.087	0.182	0.909	0.565	0.391
Low										
	$\beta_{\eta_{t1}} \neq 0$		$\beta_{\eta_{t1}} = 0$		$\beta_{\gamma_{t1}} \neq 0$		$\beta_{\gamma_{t1}} = 0$		$\pi_t(\boldsymbol{x}_{ip})$	
	ABS	COV	ABS	COV	ABS	COV	ABS	COV	ABS	COV
MicroMiss	0.354	0.941	0.288	0.927	0.273	0.610	0.164	0.851	0.216	0.776
ZIDM	0.720	0.847	0.468	0.956	0.911	0.000	0.124	0.915	0.625	0.216
DM	–	–	–	–	0.797	0.005	0.081	0.984	0.561	0.246
Medium										
	$\beta_{\eta_{t1}} \neq 0$		$\beta_{\eta_{t1}} = 0$		$\beta_{\gamma_{t1}} \neq 0$		$\beta_{\gamma_{t1}} = 0$		$\pi_t(\boldsymbol{x}_{ip})$	
	ABS	COV	ABS	COV	ABS	COV	ABS	COV	ABS	COV
MicroMiss	0.359	0.941	0.304	0.938	0.304	0.571	0.157	0.894	0.249	0.728
ZIDM	0.859	0.938	0.291	0.990	0.871	0.012	0.173	0.752	0.603	0.208
DM	–	–	–	–	0.774	0.005	0.076	0.983	0.547	0.234
High										
	$\beta_{\eta_{t1}} \neq 0$		$\beta_{\eta_{t1}} = 0$		$\beta_{\gamma_{t1}} \neq 0$		$\beta_{\gamma_{t1}} = 0$		$\pi_t(\boldsymbol{x}_{ip})$	
	ABS	COV	ABS	COV	ABS	COV	ABS	COV	ABS	COV
MicroMiss	0.392	0.942	0.358	0.924	0.344	0.485	0.155	0.931	0.278	0.691
ZIDM	0.974	0.974	0.181	0.998	0.789	0.007	0.126	0.830	0.563	0.211
DM	–	–	–	–	0.758	0.006	0.073	0.984	0.539	0.223

ABS, absolute value of the difference between the estimated and true parameters; COV, 0.95 coverage probabilities

trend was observed for all methods with respect to the zero effects on the concentration parameters. Intuitively when potential misclassification is present, the signal between covariates and taxa counts is biased towards the null for all associations, which reflects the improved performance for zero effects by the alternative models in some settings. The proposed method was able to maintain coverage for regression coefficients associated with the probability of an at-risk observation, obtaining roughly 93% coverage. Additionally, MicroMiss held above 70% coverage for $\pi_t(\boldsymbol{x}_{ip})$ regardless of the amount of misclassification. ZIDM's and DM's coverage were considerably lower for $\pi_t(\boldsymbol{x}_{ip})$ and largely affected by misclassification. With more overdispersion (Supplementary Tables S1 and S2), we observed a similar pattern in estimation performance. However, in the setting with the highest amount of overdispersion (i.e., $d = 0.10$), MicroMiss also obtained better estimation performance for zero effects on the at-risk probabilities with increased misclassification compared to ZIDM. On average, DM, ZIDM, and MicroMiss took 6, 16, and 70 seconds to run 5000 iterations on an Intel Xeon Bronze 3204 1.9 GHz processor with 16 GB RAM in all simulation settings, respectively.

To evaluate the models' variable selection performance in sparse settings, we calculated the sensitivity (1 - false negative rate) and specificity (1 - false positive rate) for

$\beta_{\eta_{t1}}$ and $\beta_{\gamma_{t1}}$, defined as Sensitivity $= \frac{TP}{FN+TP}$ and Specificity $= \frac{TN}{FP+TN}$, where TN, TP, FN, and FP represent the true negatives, true positives, false negatives, and false positives, respectively. Additionally, we evaluated the models' average absolute value of the bias and coverage probability for $\pi_t(\boldsymbol{x}_{ip})$ and computation time. With $T = C = 50$ (Supplementary Table S3), we found that ZIDMbvs was unable to identify any non-zero $\beta_{\eta_{t1}}$ in the presence of misclassification. Whereas MicroMissS was able to obtain a sensitivity around 0.75 and specificity above 0.70 in all settings. Note that DMBVS and ZINB do not perform selection on covariates potentially associated with the probability of an at-risk observation. For $\beta_{\gamma_{t1}}$, we observed a decrease in sensitivity with more misclassification for MicroMissS. Similar results were observed for ZINB. However, as the amount of misclassification increased, MicroMissS was able to obtain a higher specificity than ZINB. ZIDMbvs and DMBVS tended to underselect $\beta_{\gamma_{t1}}$ with misclassification present. Given the improved selection performance, the proposed MicroMissS model also obtained the best estimation and coverage performance for $\pi_t(\boldsymbol{x}_{ip})$. In settings with $T = C = 250$, we observed a similar pattern in selection performance for $\beta_{\eta_{t1}}$ (Supplementary Table S4). For $\beta_{\gamma_{t1}}$, the proposed method was outperformed by ZINB in terms of sensitivity as it tended to overselect. As a result, ZINB had the lowest specificity among all models. We also observed that despite the improved selection performance of MicroMissS compared to ZIDMbvs and DMBVS, the two alternative methods were able to obtain better estimation results for $\pi_t(\boldsymbol{x}_{ip})$. We attribute this result to the alternative methods obtaining better specificity levels in the presence of misclassification as the regression coefficients are biased towards the null when misclassification is ignored. With $T = C = 50$, we observed similar computation times in the sparse and non-sparse settings for all methods. In the high-dimensional settings (i.e., $T = C = 250$), ZINB, DMBVS, ZIDMbvs, and MicroMissS took roughly 120, 25, 60, and 900 seconds to run 5000 MCMC iterations, respectively.

Sensitivity analysis

In this section, we evaluate the sensitivity of the proposed model to hyperparameter specification. To assess the model's sensitivity to hyperparameter settings, we set each of the hyperparameters to default values and then evaluated the effect of manipulating each term on parameter estimation using data simulated similar to the *Medium* setting. The model was evaluated with different hyperparameter settings for ν_{tt} and the variance of the regression coefficients (σ^2_η and σ^2_γ).

We observed similar performance for the proposed method with $\nu_{tt} = 10$ and 100 in terms of absolute error and coverage probability for non-zero effects on the at-risk probabilities (Table 2). With larger ν_{tt}, the results between MicroMiss and ZIDM should agree, as the probability of misclassification is negligible. This occurred with $\nu_{tt} = 1{,}000$ in this setting. We also observed that the average absolute bias increased with ν_{tt} for zero effects on the at-risk probabilities and relative abundances with coverage probabilities remaining relatively unaffected. Lastly, we found a slight reduction in performance as the variance of the regression coefficients increased for all metrics.

Table 2 Sensitivity results for *Medium* scenario: estimation performance for $N = 50$ observations, $\dot{z}_i = 10{,}000$ reads, and $T = 50$ compositional elements at varying levels of misclassification

Hyperparameter	$\beta_{\eta t1} \neq 0$		$\beta_{\eta t1} = 0$		$\pi_t(x_{ip})$	
Setting	ABS	COV	ABS	COV	ABS	COV
$\nu_{tt} = 10$	0.466	0.893	0.302	0.947	0.630	0.560
$\nu_{tt} = 1000$	0.904	0.793	0.446	0.982	0.623	0.559
$\sigma^2_\eta = \sigma^2_\gamma = 1$	0.440	0.853	0.268	0.954	0.620	0.554
$\sigma^2_\eta = \sigma^2_\gamma = 10$	0.514	0.877	0.330	0.930	0.629	0.569

ABS, absolute value of the difference between the estimated and true parameters; COV, 0.95 coverage probabilities

The effect of obesity on the composition of the human microbiome

Maintaining a healthy gut microbiota can potentially help prevent or alleviate obesity and other metabolic diseases [56, 57]. However, the relation between the composition of the gut microbiome and obesity is often inconsistent across studies [58]. The goal of this analysis is to investigate the effect of obesity on the composition of the microbiome in a cohort of children and adolescents while accounting for potential zero-inflation and misclassification of microbial counts. The data investigated in this study were first published in [48] and made available at [59]. Prior to analysis, the microbial samples were processed similar to [60], where samples with less than 100 reads and OTUs with less than 10 reads were removed. Additionally, OTUs with less than 1% non-zero reads were excluded from this analysis. For analysis, the taxa counts were then aggregated at the genus level, with any OTUs that were not annotated removed prior to analysis. After processing the data, there were $N = 41$ individuals (16 healthy and 25 obese) with $C = 97$ different compositional elements. Of these individuals, 18 were female (44%), and the average (SD) age was 13.4 (2.8) years old. This dataset contained up to 50% zero reads at the genus-level.

To analyze these data, we set $\nu_{tt} = 500$ with $\nu_{tc} = 1$, and $\sigma^2_\gamma = \sigma^2_\eta = 1$. Since there were 97 compositional elements, this assumes that the probability of correct classification for each taxon was roughly 0.85 a priori, with the probability of misclassification equally spread across the other OTUs. The model was initialized similar to the simulation study. Disease status (Obese = 1, Healthy = 0) was included in both levels of the model (i.e., at-risk probability and the true (latent) relative abundances). The MCMC algorithm was run for 10,000 iterations, treating the first 5000 as burn-in and thinning to every 5th iteration, leaving 1000 MCMC samples for inference. Convergence and mixing of the models was visually inspected using traceplots. A random subset of these are found in the Supplementary Information for reference (Figure S1).

Figure 2 presents the estimated average relative abundances at the family level for obese and healthy participants using the proposed MicroMiss model. Table 3 presents the estimated average relative abundances at the genus level for obese and healthy participants using the proposed MicroMiss model as well as the ZIDM model (which does not accommodate potential misclassification) for comparison. For ease of exposition, only the genera with an estimated relative abundance of more than 0.01 in either the obese or healthy groups are included. A typical downstream analysis of estimated relative abundances is to perform differential abundance (DA) testing to

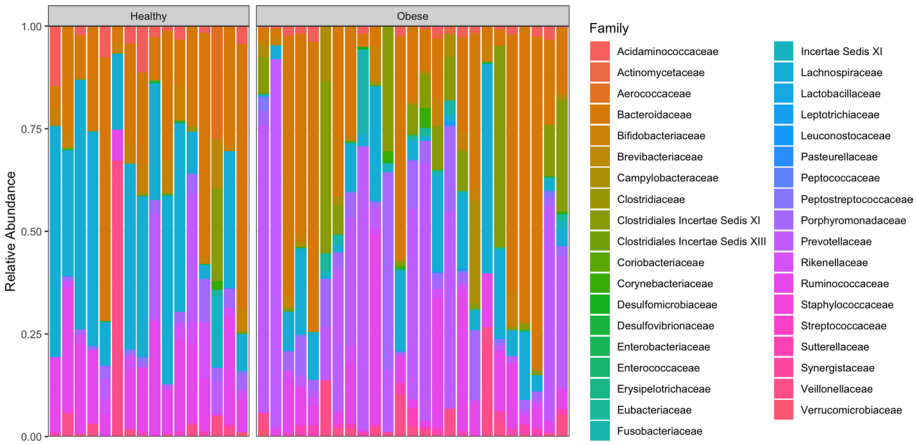

Fig. 2 Posterior relative abundance estimates for obese and healthy participants aggregated at the family level obtained with MicroMiss

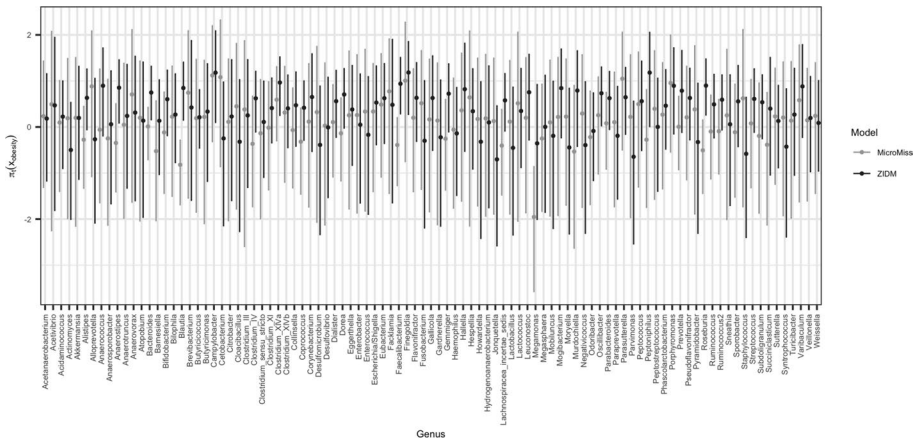

Fig. 3 Posterior estimates of the log multiplicative effect of being obese versus healthy for the relative abundance of each taxon using the proposed MicroMiss and ZIDM models. Dot represents the posterior mean with error bars capturing the 95% credible intervals

determine if the relative abundances of certain microorganisms are different across groups [61]. For reference, differential abundance testing was performed using a rank sum test after applying a centered log ratio transformation to the estimated relative abundances [62]. Supplementary Table S5 presents similar findings as Table 3 but with the centered log ratio transformed relative abundances. We observe stark differences in the relative abundances of the two groups, specifically for *Lachnospiraceae* and *Prevotellaceae*. Genus *Bacteriodies* was most abundant in both healthy (25.1%) and obese (27.1%) participants, and *Prevotella* was enriched for those with obesity. We also observed a higher relative abundance of *Blautia* in healthy (17.3%) participants versus obese (4.2%).

Of interest in this analysis is the estimated multiplicative effect of being obese versus healthy on the relative abundance of each taxon. Figure 3 presents the estimated effects using MicroMiss and ZIDM and corresponding 95% credible intervals on the log scale. As mentioned previously, the effect of disease status on a given OTU depends on its

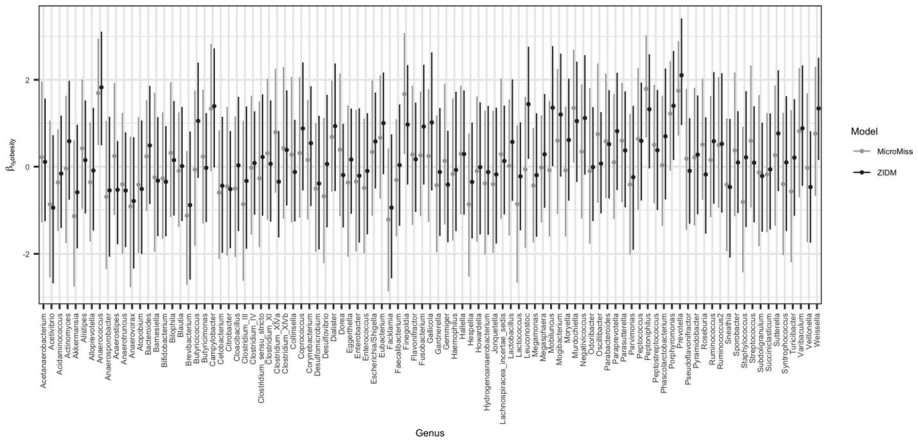

Fig. 4 Posterior estimates of at-risk log odds ratios for obese versus healthy participants for each taxon using the proposed MicroMiss and ZIDM models. Dot represents the posterior mean with error bars capturing the 95% credible intervals

Table 3 Average posterior relative abundances estimated with the proposed MicroMiss and ZIDM models for healthy and obese participants

Family	Genus	MicroMiss			ZIDM		
		Obese	Healthy	*p* value	Obese	Healthy	*p* value
Acidaminococcaceae	*Acidaminococcus*	0.0059	0.0129	0.789	0.0056	0.0122	0.103
	Succiniclasticum	0.0018	0.0106	0.012	0.0017	0.0099	0.454
Aerococcaceae	*Facklamia*	0.0000	0.0172	< 0.001	0.0000	0.0160	< 0.001
Bacteroidaceae	*Bacteroides*	0.2709	0.2510	0.316	0.2632	0.2447	0.019
Bifidobacteriaceae	*Bifidobacterium*	0.0183	0.0307	0.109	0.0178	0.0300	0.008
Campylobacteraceae	*Campylobacter*	0.0217	0.0006	< 0.001	0.0213	0.0007	< 0.001
Clostridiales Incertae Sedis XI	*Anaerococcus*	0.0202	0.0057	< 0.001	0.0197	0.0056	< 0.001
	Finegoldia	0.0271	0.0002	< 0.001	0.0264	0.0004	< 0.001
	Peptoniphilus	0.0426	0.0077	< 0.001	0.0417	0.0077	< 0.001
Incertae Sedis XI	*Murdochiella*	0.0034	0.0106	< 0.001	0.0034	0.0101	0.003
Lachnospiraceae	*Anaerostipes*	0.0081	0.0230	0.030	0.0092	0.0249	< 0.001
	Blautia	0.0417	0.1728	< 0.001	0.0382	0.1548	< 0.001
	Coprococcus	0.0047	0.0105	0.019	0.0060	0.0116	0.004
	Lachnospiracea Incertae Sedis	0.0061	0.0164	0.005	0.0099	0.0265	< 0.001
	Roseburia	0.0209	0.0558	0.019	0.0210	0.0556	< 0.001
	Ruminococcus	0.0079	0.0156	0.033	0.0093	0.0177	< 0.001
Porphyromonadaceae	*Parabacteroides*	0.0192	0.0169	0.612	0.0191	0.0170	0.262
	Porphyromonas	0.0434	0.0024	< 0.001	0.0423	0.0024	0.001
Prevotellaceae	*Prevotella*	0.2189	0.0412	0.001	0.2189	0.0412	0.001
Rikenellaceae	*Alistipes*	0.0080	0.0240	0.045	0.0083	0.0238	0.003
Ruminococcaceae	*Clostridium IV*	0.0031	0.0126	0.010	0.0038	0.0141	< 0.001
	Faecalibacterium	0.0597	0.1091	0.004	0.0588	0.1057	< 0.001
	Ruminococcus	0.0185	0.0219	0.250	0.0179	0.0212	0.012
Veillonellaceae	*Dialister*	0.020	0.013	0.310	0.020	0.013	0.621
	Megamonas	0.0003	0.0314	< 0.001	0.0003	0.0295	0.128
	Megasphaera	0.0123	0.0097	0.336	0.0115	0.0092	0.209

Only genera with relative abundances above 0.01 for obese or healthy groups are presented. *P*-values were obtained from a rank sum test on the centered log ratio transformations of relative abundances at the genus level

effect on the other OTUs in the model as well as their relative abundances. Thus, there may not be a one-to-one relationship between the estimated regression coefficients associated with the concentration parameters (available in Supplementary Information Figure S2) and the true relative abundances. In this analysis, we observed 2 genera-obesity associations in which the corresponding 95% credible intervals did not contain a multiplicative effect of 1; *Blautia* and *Megamonas*. Recently, *Blautia* was found to be inversely related to visceral fat accumulation in adults [63] and depleted in obese children [64]. *Megamonas* was found to be positively associated with obesity in Chinese [65] and Taiwanese adults [66]. Our post-hoc DA found *Megamonas* depleted in obese participants. With the ZIDM model, which does not accommodate classification uncertainty, we observed associations between *Anaerostipes*, *Bacteroides*, *Blautia*, *Campylobacter*, *Clostridium XlV*, *Faecalibacterium*, *Finegoldia*, *Oscillibacter*, *Peptoniphilus*, and *Roseburia* and obesity. While a majority of the estimated relations between different genera and obesity were similar with both models, a few flipped directions. Interestingly, obesity was positively associated with the relative abundance of *Blautia* using the ZIDM model, but we observed a negative relation with MicroMiss. Recall that we also observed *Blautia* enriched for the healthy group. These seemingly conflicting results highlight the importance of considering potential misclassification as well as the compositional structure of the microbiome data when performing analysis, especially in regression settings where covariates are potentially associated with multiple taxa. Said differently, when evaluating the effect of a covariate on a particular relative abundance, it is often not possible to "hold all else constant" as other relative abundances may also depend on the given covariate.

Our modeling framework simultaneously provides inference on the relation between obesity status and at-risk probabilities (Fig. 4). We observed a positive association between obesity and at-risk observations in *Anaerococcus*, *Finegoldia*, Murdochiella, *Peptoniphilus*, and *Prevotella*. *Prevotella* is commonly found to be associated with obesity and has shown a positive association in weight loss studies [65, 67]. We estimated the relative abundance of *Prevotella* in the obese group as 21.9%, whereas it was only 4.1% in the healthy group. With the ZIDM model, we observed associations between obesity and the at-risk probability of *Anaerococcus*, *Leuconostoc*, *Mobiluncus*, *Porphyromonas*, *Prevotella*, and *Weissella*. The proposed model's estimated credible intervals were typically larger than the ZIDM model, as expected from the results on simulated data.

We further analyzed the application data with the sparsity-induced version of our model, MicroMissS. Using non-informative prior probabilities of inclusion (i.e., $a_\eta = b_\eta = a_\gamma = b_\gamma = 1$), MicroMissS found that obesity status was associated with the probability of an at-risk observation for *Acidaminococcus*, *Akkermansia*, *Anaerosporobacter*, *Bacteroides*, *Barnesiella*, *Bifidobacterium*, *Blautia*, *Clostridium sensu stricto*, *Coprococcus*, *Gemmiger*, *Haemophilus*, *Odoribacter*, *Prevotella*, *Pseudoflavonifractor*, and *Ruminococcus*. The proposed model also identified associations between obesity status and the concentration parameters for *Anaerosporobacter*, *Blautia*, *Clostridium IV*, *Faecalibacterium*, *Gemmiger*, *Lachnospiracea incertae sedis*, and *Megamonas*. Applying the ZINB model to the data, over half of the taxa were found to be associated with obesity status using a non-informative prior probability of inclusion. Reducing the prior

probability of inclusion to 0.001, the model similarly identified *Blautia*, *Clostridium IV*, *Faecalibacterium*, *Gemmiger*, and *Lachnospiracea incertae sedis* as associated with obesity status, in addition to 17 other taxa. These results were not surprising as ZINB tended to identify more associations than MicroMissS in simulation. In terms of computation time, the proposed MicroMiss and MicroMissS models took roughly 5 minutes to generate the 10,000 MCMC samples, whereas ZIDM and ZINB took 1 min and 30 s, respectively.

Lastly, we performed a sensitivity analysis in order to assess how the results may change when $\boldsymbol{\nu}$ is specified using phylogenetic information and with higher levels of misclassification artifically introduced into the observed counts. See section S3 in the Supplementary Information for more details.

Conclusions

In this work, we proposed a Bayesian zero-inflated Dirichlet-multinomial regression model for microbiome data with potential taxonomic misclassification. Our framework treats zero-inflation as a model selection problem and accommodates uncertainty in observed taxa counts by assuming they are realizations from the true (unobserved) classifications through a taxon-specific Dirichlet-multinomial model. Our approach is scalable, handles the complex structure of microbial count data, and allows covariates to be associated with the at-risk probabilities as well as the concentration parameters for the latent relative abundances. Additionally, it is agnostic to the procedures used to collect and process the observed taxa counts and can flexibly incorporate information regarding the structure of the data that may inform potential misclassification patterns. Through simulation and real data analysis, we demonstrate how ignoring misclassification can affect inference on covariates' associations.

As a first attempt at modeling misclassification in microbiome data, we assume that misclassification rates are shared across individuals, borrowing information across observations for inference. An extension of this work is to modify the misclassification probabilities to vary at the host, or even read level. For example, the model could be specified so that reads with similar sequences may be misclassified with a larger probability. While this would provide more nuanced inference, it would also greatly increase an already large parameter space. Relatedly, future extensions could explore incorporating covariate information to inform potential misclassification probabilities when available. While developed for microbiome data analysis, the proposed modeling framework is generally applicable to other classification settings in which zero-inflation and potential misclassification are present. Lastly, an active area in microbiome research is developing personalized interventions to help moderate the composition of the microbiome to improve health outcomes. Given the high variability of microbiome samples within and between individuals, a future next step would be to adapt the model to handle time-varying and individual-level effects following [68].

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-025-06078-4.

Supplementary file 1.

Supplementary file 2.

Supplementary file 3.

Supplementary file 4.

Acknowledgements
The author gratefully acknowledges Dr. Duvallet for their help accessing the application data.

Author contributions
MDK developed the model and accompanying `R` package, performed all analyses, and wrote the manuscript.

Funding
The author graciously acknowledges the support of NSF grant DMS-2245492. The opinions, findings, and conclusions expressed are those of the author and do not necessarily reflect the views of the NSF.

Availability of data and materials
Code to implement the method (MicroMiss.zip), replicate the study findings (DataAnalysis.R), and simulate data in the simulation study (Simulation_code.R) as well as details of the MCMC algorithm and supplemental figures are found in the Supplementary Information. Data used in the analysis are found at: https://zenodo.org/records/569601.

Declarations

Ethical approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
Not applicable.

Received: 8 October 2024 Accepted: 10 February 2025
Published online: 27 February 2025

References

1. Gwak HJ, Rho M. Data-driven modeling for species-level taxonomic assignment from 16S rRNA: application to human microbiomes. Front Microbiol. 2020;11:570825.
2. Huson DH, Auch AF, Qi J, Schuster SC. MEGAN analysis of metagenomic data. Genome Res. 2007;17(3):377–86.
3. Wang Q, Garrity GM, Tiedje JM, Cole JR. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol. 2007;73(16):5261–7.
4. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014;15(3):1–12.
5. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010;26(19):2460–1.
6. Edgar RC. UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nat Methods. 2013;10(10):996.
7. Allard G, Ryan FJ, Jeffery IB, Claesson MJ. SPINGO: a rapid species-classifier for microbial amplicon sequences. BMC Bioinform. 2015;16(1):1–8.
8. Shah N, Meisel JS, Pop M. Embracing ambiguity in the taxonomic classification of microbiome sequencing data. Front Genet. 2019;10:1022.
9. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol. 2019;37(8):852–7.
10. Eren AM, Maignien L, Sul WJ, Murphy LG, Grim SL, Morrison HG, et al. Oligotyping: differentiating between closely related microbial taxa using 16S rRNA gene data. Methods Ecol Evol. 2013;4(12):1111–9.
11. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: high-resolution sample inference from Illumina amplicon data. Nat Methods. 2016;13(7):581–3.
12. Amir A, McDonald D, Navas-Molina JA, Kopylova E, Morton JT, Zech XuZ, et al. Deblur rapidly resolves single-nucleotide community sequence patterns. MSystems. 2017;2(2):10–1128.
13. Callahan BJ, McMurdie PJ, Holmes SP. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. ISME J. 2017;11(12):2639–43.
14. Shi P, Zhang A, Li H. Regression analysis for microbiome compositional data. Ann Appl Stat. 2016;10(2):1019–40.
15. Wadsworth WD, Argiento R, Guindani M, Galloway-Pena J, Shelburne SA, Vannucci M. An integrative Bayesian Dirichlet-multinomial regression model for the analysis of taxonomic abundances in microbiome data. BMC Bioinform. 2017;18(1):1–12.
16. Zhang X, Yi N. NBZIMM: negative binomial and zero-inflated mixed models, with application to microbiome/metagenomics data analysis. BMC Bioinform. 2020;21(1):1–19.
17. Okui T. A Bayesian nonparametric topic model for microbiome data using subject attributes. IPSJ Trans Bioinform. 2020;13:1–6.
18. Koslovsky MD, Hoffman KL, Daniel CR, Vannucci M. A Bayesian model of microbiome data for simultaneous identification of covariate associations and prediction of phenotypic outcomes. Ann Appl Stat. 2020;14(3):1471–92.

19. Ha MJ, Kim J, Galloway-Peña J, Do KA, Peterson CB. Compositional zero-inflated network estimation for microbiome data. BMC Bioinform. 2020;21(21):1–20.
20. Ren B, Bacallado S, Favaro S, Vatanen T, Huttenhower C, Trippa L. Bayesian mixed effects models for zero-inflated compositions in microbiome data analysis. Ann Appl Stat. 2020;14(1):494–517.
21. Jiang S, Xiao G, Koh AY, Kim J, Li Q, Zhan X. A Bayesian zero-inflated negative binomial regression model for the integrative analysis of microbiome data. Biostatistics. 2021;22(3):522–40.
22. Zhou C, Zhao H, Wang T. Transformation and differential abundance analysis of microbiome data incorporating phylogeny. Bioinformatics. 2021;37(24):4652–60.
23. Shuler K, Verbanic S, Chen IA, Lee J. A Bayesian nonparametric analysis for zero-inflated multivariate count data with application to microbiome study. J R Stat Soc Ser C Appl Stat. 2021;70(4):961–79.
24. Osborne N, Peterson CB, Vannucci M. Latent network estimation and variable selection for compositional data via variational EM. J Comput Graph Stat. 2022;31(1):163–75.
25. Bandoy DDR, Huang BC, Weimer BC. Misclassification of a whole genome sequence reference defined by the human microbiome project: a detrimental carryover effect to microbiome studies. medRxiv. 2019;p. 19000489.
26. Cao Q, Sun X, Rajesh K, Chalasani N, Gelow K, Katz B, et al. Effects of rare microbiome taxa filtering on statistical analysis. Front Microbiol. 2021;11:607325.
27. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome datasets are compositional and this is not optional. Front Microbiol. 2017;8:2224.
28. Clausen DS, Willis AD. Evaluating replicability in microbiome data. Biostatistics. 2022;23(4):1099–114.
29. Pollock J, Glendinning L, Wisedchanwet T, Watson M. The madness of microbiome: Attempting to find consensus "best practice" for 16S microbiome studies. Appl Environ Microbiol. 2018;84(7):e02627.
30. Di Cecco D, Tancredi A. Estimating the number of sequencing errors in microbial diversity studies. Environ Ecol Stat. 2024;31:485–507.
31. Berry MA, White JD, Davis TW, Jain S, Johengen TH, Dick GJ, et al. Are oligotypes meaningful ecological and phylogenetic units? A case study of Microcystis in freshwater lakes. Front Microbiol. 2017;8:365.
32. Schloss PD. Amplicon sequence variants artificially split bacterial genomes into separate clusters. Msphere. 2021;6(4):10–1128.
33. Neelon B. Bayesian zero-inflated negative binomial regression based on Pólya-gamma mixtures. Bayesian Anal. 2019;14(3):829.
34. Xu L, Paterson AD, Turpin W, Xu W. Assessment and selection of competing models for zero-inflated microbiome data. PloS One. 2015;10(7):e0129606.
35. Jiang R, Zhan X, Wang T. A flexible zero-inflated poisson-gamma model with application to microbiome sequence count data. J Am Stat Assoc. 2023;118(542):792–804.
36. Aitchison J, Ho C. The multivariate Poisson-log normal distribution. Biometrika. 1989;76(4):643–53.
37. Chiquet J, Mariadassou M, Robin S. The poisson-lognormal model as a versatile framework for the joint analysis of species abundances. Front Ecol Evol. 2021;9:188.
38. Koslovsky MD. A Bayesian zero-inflated Dirichlet-multinomial regression model for multivariate compositional count data. Biometrics. 2023;79(4):3239–51.
39. Shi P, Zhou Y, Zhang AR. High-dimensional log-error-in-variable regression with applications to microbial compositional data analysis. Biometrika. 2022;109(2):405–20.
40. Swartz TB, Haitovsky Y, Vexler A, Yang TY. Bayesian identifiability and misclassification in multinomial data. Can J Stat. 2004;32(3):285–302.
41. Wang S, Wang L, Swartz TB. Inference for misclassified multinomial data with covariates. Can J Stat. 2020;48(4):655–69.
42. Pérez CJ, Girón FJ, Martín J, Ruiz M, Rojano C. Misclassified multinomial data: a Bayesian approach. RACSAM. 2007;101(1):71–80.
43. Frénay B, Verleysen M. Classification in the presence of label noise: a survey. IEEE Trans Neural Netw Learn Syst. 2013;25(5):845–69.
44. Wright WJ, Irvine KM, Almberg ES, Litt AR. Modelling misclassification in multi-species acoustic data when estimating occupancy and relative activity. Methods Ecol Evol. 2020;11(1):71–81.
45. Spiers AI, Royle JA, Torrens CL, Joseph MB. Estimating species misclassification with occupancy dynamics and encounter rates: a semi-supervised, individual-level approach. Methods Ecol Evol. 2022;13(7):1528–39.
46. Koslovsky MD, Kaplan A, Terranova VA, Hooten MB. A unified Bayesian framework for modeling measurement error in multinomial data. Bayesian Anal. 2024;1(1):1–31.
47. Stratton C, Irvine KM, Banner KM, Wright WJ, Lausen C, Rae J. Coupling validation effort with in situ bioacoustic data improves estimating relative activity and occupancy for multiple species with cross-species misclassifications. Methods Ecol Evol. 2022;13(6):1288–303.
48. Zhu L, Baker SS, Gill C, Liu W, Alkhouri R, Baker RD, et al. Characterization of gut microbiomes in nonalcoholic steatohepatitis (NASH) patients: a connection between endogenous alcohol and NASH. Hepatology. 2013;57(2):601–9.
49. Wang T, Zhao H. A Dirichlet-tree multinomial regression model for associating dietary nutrients with gut microorganisms. Biometrics. 2017;73(3):792–801.
50. Chen J, Li H. Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis. Ann Appl Stat. 2013;7(1):418–42.
51. Koslovsky MD, Vannucci M. MicroBVS: Dirichlet-tree multinomial regression models with Bayesian variable selection-an R package. BMC Bioinform. 2020;21(1):1–10.
52. Polson NG, Scott JG, Windle J. Bayesian inference for logistic models using Pólya-Gamma latent variables. J Am Stat Assoc. 2013;108(504):1339–49.
53. Dai Z, Wong SH, Yu J, Wei Y. Batch effects correction for microbiome data with Dirichlet-multinomial regression. Bioinformatics. 2019;35(5):807–14.
54. Zhang Y, Zhou H, Zhou J, Sun W. Regression models for multivariate count data. J Comput Graph Stat. 2017;26(1):1–13.

55. Eddelbuettel D, François R. Rcpp: seamless R and C++ integration. J Stat Softw. 2011;40:1–18.
56. John GK, Mullin GE. The gut microbiome and obesity. Curr Oncol Rep. 2016;18:1–7.
57. Liu BN, Liu XT, Liang ZH, Wang JH. Gut microbiota in obesity. World J Gastroenterol. 2021;27(25):3837.
58. Castaner O, Goday A, Park YM, Lee SH, Magkos F, Shiow SATE, et al. The gut microbiome profile in obesity: a systematic review. Int J Endocrinol. 2018;2018:4095789.
59. Duvallet C, Gibbons S, Gurry T, Irizarry R, Alm E. MicrobiomeHD: the human gut microbiome in health and disease [Data set]. Zenodo. 2017. https://doi.org/10.5281/zenodo.569601.
60. Duvallet C, Gibbons SM, Gurry T, Irizarry RA, Alm EJ. Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. Nat Commun. 2017;8(1):1784.
61. Wallen ZD. Comparison study of differential abundance testing methods using two large Parkinson disease gut microbiome datasets derived from 16S amplicon sequencing. BMC Bioinform. 2021;22(1):265.
62. Aitchison J. The statistical analysis of compositional data. J R Stat Soc Ser B Methodol. 1982;44(2):139–60.
63. Ozato N, Saito S, Yamaguchi T, Katashima M, Tokuda I, Sawada K, et al. Blautia genus associated with visceral fat accumulation in adults 20–76 years of age. NPJ Biofilms Microbiomes. 2019;5(1):28.
64. Benítez-Páez A, Gómez del Pugar EM, López-Almela I, Moya-Pérez Á, Codoñer-Franch P, Sanz Y. Depletion of Blautia species in the microbiota of obese children relates to intestinal inflammation and metabolic phenotype worsening. Msystems. 2020;5(2):10–1128.
65. Duan M, Wang Y, Zhang Q, Zou R, Guo M, Zheng H. Characteristics of gut microbiota in people with obesity. Plos One. 2021;16(8):e0255446.
66. Chiu CM, Huang WC, Weng SL, Tseng HC, Liang C, Wang WC, et al. Systematic analysis of the association between gut flora and obesity through high-throughput sequencing and bioinformatics approaches. BioMed Res Int. 2014;2014:906168.
67. Christensen L, Vuholm S, Roager HM, Nielsen DS, Krych L, Kristensen M, et al. Prevotella abundance predicts weight loss success in healthy, overweight adults consuming a whole-grain diet ad libitum: A post hoc analysis of a 6-wk randomized controlled trial. J Nutr. 2019;149(12):2174–81.
68. Pedone M, Amedei A, Stingo FC. Subject-specific Dirichlet-multinomial regression for multi-district microbiota data analysis. Ann Appl Stat. 2023;17(1):539–59.

Publisher's Note