Computational and Structural Biotechnology Journal 21 (2023) 1621–1629

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Method article

# Infants' gut microbiome data: A Bayesian Marginal Zero-inflated Negative Binomial regression model for multivariate analyses of count data

Morteza Hajihosseini [a], Payam Amini [b], Alireza Saidi-Mehrabad [c], Irina Dinu [d,*]

[a] Stanford Department of Urology, Center for Academic Medicine, Palo Alto, CA 94304
[b] Department of Biostatistics, School of public Health, IRAN University of Medical Sciences, Tehran, Iran
[c] Division of Hydrological Sciences, Desert Research Institute, Las Vegas, Nevada, USA
[d] School of Public Health, University of Alberta, Edmonton, Alberta, Canada

ARTICLE INFO

*Article history:*
Received 2 September 2022
Received in revised form 13 February 2023
Accepted 14 February 2023
Available online 15 February 2023

*Keywords:*
Infants
Gut Microbiome
Zero-inflation
Multivariate Structure

ABSTRACT

The infants' gut microbiome is dynamic in nature. Literature has shown high inter-individual variability of gut microbial composition in the early years of infancy compared to adulthood. Although next-generation sequencing technologies are rapidly evolving, several statistical analysis aspects need to be addressed to capture the variability and dynamic nature of the infants' gut microbiome. In this study, we proposed a Bayesian Marginal Zero-inflated Negative Binomial (BAMZINB) model, addressing complexities associated with zero-inflation and multivariate structure of the infants' gut microbiome data. Here, we simulated 32 scenarios to compare the performance of BAMZINB with glmFit and BhGLM as the two other widely similar methods in the literature in handling zero-inflation, over-dispersion, and multivariate structure of the infants' gut microbiome. Then, we showed the performance of the BAMZINB approach on a real dataset using SKOT cohort (I and II) studies. Our simulation results showed that the BAMZINB model performed as well as those two methods in estimating the average abundance difference and had a better fit for almost all scenarios when the signal and sample size were large. Applying BAMZINB on SKOT cohorts showed remarkable changes in the average absolute abundance of specific bacteria from 9 to 18 months for infants of healthy and obese mothers. In conclusion, we recommend using the BAMZINB approach for infants' gut microbiome data taking zero-inflation and over-dispersion properties into account in multivariate analysis when comparing the average abundance difference.



## 1. Introduction

The advent of high throughput sequencing in parallel with substantial advances in computational, molecular, and quantitative fields has opened new avenues in our understanding of trillions of microbes that call the human intestine home (termed as "human gut microbiome") [1–4]. In vitro and in vivo studies have shown that the human gut microbiome takes shape shortly after birth mainly via transmission from the maternal microbial pool (vagina, gut, skin, or breastmilk) and continues to develop until it becomes mature two to three years after the initial colonization [5–8]. Any changes that could disrupt the stability of the healthy gut microbiome, particularly at the early stages of life, could result in severe dysbiosis, which could pave the path for major health issues in adulthood [9–11].

For instance, dysbiosis in newly colonized microbes of an infant's gut could result in disorders such as failure to thrive, which have a negative impact on child growth [12]. Necrotizing enterocolitis is another common intestinal disease associated with early life dysbiosis, severe intestinal inflammation, and irritable bowel diseases [13]. Furthermore, a study conducted by Ivashkin and colleagues on patients with irritable bowel syndrome (IBS) showed a link between alteration in the gut microbial community and disruption of the pro-inflammatory, anti-inflammatory cytokines and tight junction proteins expression [14]. IBS and some diseases associated with intestinal inflammation are believed to be one of the long-term side

* Correspondence to: School of Public Health, University of Alberta, 3-278 Edmonton Clinic Health Academy, 11405 - 87 Ave NW, Edmonton, Alberta T6G 1C9, Canada.
*E-mail addresses:* mortezah@stanford.edu (M. Hajihosseini), payam.amini87@gmail.com (P. Amini), ali.saidi-mehrabad@dri.edu (A. Saidi-Mehrabad), idinu@ualberta.ca (I. Dinu).

https://doi.org/10.1016/j.csbj.2023.02.027

effects of early gut dysbiosis [15]. Another prime example of the effect of early gut dysbiosis, which appears later on in life, is obesity [16]. A study conducted by Kasai and colleagues show higher diversity in bacteria in obese individuals compared to non-obese individuals [17]. An increase in the abundance of specific microbial taxa has heavily influenced the biological activity of neutrophils, lymphocytes, antigen-presenting cells, and T and B cells [18]. As a result of such substantial influence of the gut microbiome over the human immune system, diseases such as atopy [19], rheumatoid arthritis [20], and nervous system's demyelination-related pathologies [21], or Crohn's disease and ulcerative colitis [22], have all been suspected to be triggered by changes in gut microbial community structure.

While innovations in next-generation sequencing (NGS) technologies decipher the relationship between the dynamic changes in the human microbiome and various diseases through 16 s ribosome RNA gene sequencing or shotgun metagenomics sequencing, the development of statistical methods in microbiome research has not kept up with the same pace. Recent NGS technologies offer a massive amount of sequence reads that provide information about species or bacteria with high resolution [23]. On the basis of this data, statistical techniques were used to examine the relationships between various types of bacteria and the characteristics of the subjects or the environment [24]. These circumstances or traits may have an impact on the absolute abundance of microorganisms, and must be taken into account for a more accurate differential abundance analysis [25]. Furthermore, there may be a clinical need to quantify the association between the microbiome and these confounders [26–29].

Current literature shows that the analysis of microbial data is complicated due to inherited characteristics of microbiome count data, such as over-dispersion, zero inflation, and fluctuating library size. Fortunately, some of these challenges have been widely studied in the context of microarray and single-cell RNA analysis. For example, library size is suggested to be controlled by implementing a complex normalization technique into the classic Negative Binomial (NB) model, a well-known model for handling over-dispersion. This can be found in R packages such as edgeR [30] and DESeq2 [31,32]. Another example is implementing a zero-inflated Gaussian mixture model [33] in the metagenomes package to accommodate zero-inflated data in the analysis. Although these tools seem useful for microbiome data analysis, the assurance of generating precise and unbiased results depends on the sample size and multivariate correlation structure of the microbiome data. The common solution for the dimensionality issue in such tools is the utilization of dimension reduction methods and dissimilarity matrices (i.e., principal component or partial least squares) [34]. However, relevant information could be lost by selecting a pre-specified number of eigenvalues or factors [35].

Another set of (single cell)-RNA-seq analysis methods has been recently adapted to microbiome studies. These methods address the dimensionality issue using Markov chain Monte Carlo (MCMC) algorithms [36]. Glmfit function in edgeR package and bglm function in BhGLM package [37] are two examples of such tools in genome data analysis. Glmfit function can generate effect sizes for each bacterium as well as the whole microbiome composition (for ecological studies), estimating one common over-dispersion term and adjusting for library size. On the contrary, bglm function is limited to only univariate analysis, generating the result for each bacterium at a time with the estimation of an over-dispersion term per bacterium and adjusting for library size. However, because both tools use the classic NB model, the analyses do not address the current zero-inflation issue.

Geert Molenberghs and Geert Verbeke discussed marginal models for discrete longitudinal data and their strength [38]. We proposed a Bayesian Marginal Zero-inflated Negative Binomial (BAMZINB) model, addressing complexities associated with zero-inflation and multivariate structure of gut microbiome data described above. Our modeling construction includes several advantages. First, BAMZINB can generate results for individual bacterium and the entire microbiome composition. Second, it incorporates zero inflation, over-dispersion, fluctuating library size, and multivariate correlation structure of the microbiome data using the generalized linear model framework. Furthermore, it models the heterogeneity among subjects via a random intercept component. More specifically, the BAMZINB model can perform the differential absolute abundance analysis of responses one by one, focusing on the relationship of interest and simultaneously controlling for the microbiome data's multivariate correlation structure.

As we proceed, the material and method section will further describe the details of the BAMZINB method, the extensive simulation study, and the application of the SKOT Cohorts data [39–41]. Our simulation studies and a real data application will be presented in the results and discussion sections in terms of performance criteria such as absolute relative bias (ABR) and deviance.

## 2. Methods and materials

### 2.1. Bayesian Marginal Zero-inflated Negative Binomial (BAMZINB)

#### 2.1.1. The joint zero-inflated negative binomial distribution

Multivariate models are used to accommodate multiple correlated outcomes via statistical models that jointly represent relationships between outcomes and predictors. A broad objective of joint modeling is to provide a framework to ensure valid inferences by accounting for the correlation among the outcome variables. Let $Y_1$ and $Y_2$ be random variables representing correlated outcomes. While we restrict attention to the case of two response variables, an extension to a higher dimension is straightforward. In the zero-inflated negative binomial regression model, the objectives are to identify significant factors influencing the zero-inflation count of the bacteria and determine the extent of the effect of potential biological and environmental factors on the mean count of a specific bacteria in the presence of zero inflation and overdispersion. Let's assume that $Y_1$ follows the zero-inflated negative binomial distribution (Eq. 1):

$$\left\{\begin{array}{c} p(y=0)=\pi+(1+\pi)\left(\frac{\theta}{\mu+\theta}\right)^{\theta} y=0 \\ p(y>0)=(1-\pi)\left(\frac{\Gamma(y+\theta)}{\Gamma(\theta)\Gamma(y+1)}\right)\left(\frac{\theta}{\mu+\theta}\right)^{\theta}\left(1-\frac{\theta}{\mu+\theta}\right)^{y} y=1,2,3,\ldots \end{array}\right\} \tag{1}$$

where $0<\pi<1$ is the probability of an extra zero response, $\mu$ is the mean and $\theta^{-1}$ is the dispersion parameter in the negative binomial distribution. The joint distribution of $Y_1, Y_2, \ldots, Y_m$ across all $n$ samples based on Eq. 1 is [42]:

$$\begin{aligned} f(\boldsymbol{Y}|\pi.\theta)=&\prod_{i=1}^{m}\left(\prod_{j=1}^{n_i}\left[\pi_i+(1-\pi_i)\left(\frac{\theta_i}{\mu_i+\theta_i}\right)^{\theta_i}\right]I(y_{ij}\right. \\ &=0)\prod_{j=1}^{n_i}\left[(1-\pi_i)\left(\frac{\Gamma(y_{ij}+\theta_i)}{\Gamma(\theta_i)\Gamma(y_{ij}+1)}\right)\left(\frac{\theta_i}{\mu_i+\theta_i}\right)^{\theta_i}\right. \\ &\left.\left.\left(1-\frac{\theta_i}{\mu_i+\theta_i}\right)^{y_i}\right]I(y_{ij}>0)\right) \end{aligned} \tag{2}$$

where $\mathbf{Y}=(Y_1, Y_2, \ldots, Y_m)$, $\pi_i=(\pi_1, \pi_2, \ldots, \pi_m)$, $\mu_i=(\mu_1, \mu_2, \ldots, \mu_m)$ is the mean and $\theta_i=(\theta_1, \theta_2, \ldots, \theta_m)$. Lambert [43] and Mullay [44] proposed a model for zero inflated negative binomial model as follows:

$$logit(\pi_i)=\acute{Z}\alpha \text{ and } Log(\mu_i)=\acute{X}\beta;\quad i=1,2,\ldots,m \tag{3}$$

Where $\alpha = (\alpha_1, \alpha_2, ..., \alpha_p)$ and $\beta = (\beta_1, \beta_2, ..., \beta_q)$ are the regression coefficient vectors, $\boldsymbol{Z}$ and $\boldsymbol{X}$ are the covariates matrices with elements $z_{jk}$ and $x_{jk'}$ as the observed covariate for the j$^{th}$ individual of the $k^{th}$ and $k'^{th}$ predictor for each process, respectively. In practice, both processes often have the same vector of variables, but not necessarily all the time. The association between $Y_1, Y_2, ..., Y_m$ is considered using a working covariance matrix which contains the variance and covariance of model residuals. In the covariance matrix

$$\begin{bmatrix} \sigma_1^2 & \cdots & \sigma_{1m} \\ \vdots & \ddots & \vdots \\ \sigma_{m1} & \cdots & \sigma_m^2 \end{bmatrix}_{m\times m}$$

$\sigma_i^2$ is the variance of residuals for the i$^{th}$ outcome sub model and $\sigma_{ii'}$ is the covariance between the residuals of the outcomes $i$ and $i'$.

### 2.1.2. Marginalized Multivariate Zero-inflated Negative Binomial (MZINB) model

Marginalized zero-inflated count response models are useful when the overall mean of specific bacteria $\mu_i = E[Y_i]$ is of primary interest [42,45]. In this model, $\exp(\beta_j)$ represents the multiplicative increase in the mean count for bacteria in the overall population corresponding to a one-unit increase in the covariate $x_{ij}$ and $\exp(\alpha_j)$ represents the odds ratio of observing non-zero number of bacteria corresponding to a one-unit increase in the covariate $z_{ij}$. A marginalized Multivariate Zero-Inflated Negative Binomial (MZINB) regression model likelihood function is introduced within a likelihood framework of $m$ outcomes as follows:

$$\begin{aligned} (Y_i \mid \alpha, \beta, \theta_i) &= \prod_{y_i=0} p(Y_i = y_i \mid \alpha, \beta, \theta_i) \prod_{y_i>0} p(Y_i = y_i \mid \alpha, \beta, \theta_i) \\ &= \prod_{i=1}^{m} \left( \prod_{j=1}^{n_i} \left[ \frac{e^{Z'\alpha}}{1+e^{Z'\alpha}} + \left(\frac{1}{1+e^{Z'\alpha}}\right)\left(\frac{\theta_i}{e^{X'\beta}+\theta_i}\right)^{\theta_i} \right] I(y_{ij} \right. \\ &= 0) \prod_{j=1}^{n_i} \left[ \left(\frac{1}{1+e^{Z'\alpha}}\right)\left(\frac{\Gamma(y_{ij}+\theta_i)}{\Gamma(\theta_i)\Gamma(y_{ij}+1)}\right)\left(\frac{\theta_i}{e^{X'\beta}+\theta_i}\right)^{\theta_i} \right. \\ &\left. \left. \left(1-\frac{\theta_i}{e^{X'\beta}+\theta_i}\right)^{y_i} \right] I(y_{ij}>0) \right) \end{aligned} \tag{4}$$

we run simulations, including our proposed method, BAMZINB. Variance-covariance matrix is the responsible component of the model for capturing the association among the variables. This matrix can be of different covariance pattern models such as autoregressive, exchangeable, unstructured and etc. In the current study, we considered an unstructured covariance pattern model due to the Bayesian framework of analysis. In addition, we used Bayesian estimation methods to cover potentially problematic issues, such as over-parametrization and small sample sizes.

### 2.1.3. Bayesian parameter estimates

We needed to specify a prior distribution for parameters in the model to obtain a Bayesian estimation of the unknown parameters in the BAMZINB model. Formulation of an informative prior distribution results from providing good prior information [46]. In this study, we assumed Gamma (a=0.001, b=0.001) distribution for overdispersion parameter ($\theta$) and Normal (0, 10e+6) distribution for model parameter coefficients ($\alpha$, $\beta$). Therefore, the prior distribution for parameters $(\alpha, \beta, \theta)$ is written as the following:

$$\begin{aligned} &p(\alpha, \beta, \theta) \\ &= \prod_{y_i=0} \left( \frac{1}{\sigma_{\alpha_i}\sqrt{2\pi}} e^{\frac{-(\alpha_i-\mu_{\alpha_i})^2}{2\sigma^2_{\alpha_i}}} \right) \prod_{y_i>0} \left( \frac{1}{\sigma_{\beta_i}\sqrt{2\pi}} e^{\frac{-(\beta_i-\mu_{\beta_i})^2}{2\sigma^2_{\beta_i}}} \right) \frac{1}{b^a\Gamma(a)} \theta^{a-1} e^{-\frac{\theta}{b}} \end{aligned} \tag{5}$$

Using the combination of the prior (Eq. 5) and the likelihood (Eq. 4), the posterior for the parameters can be written as:

$$\begin{aligned} p(\alpha, \beta, \theta|, Y) = \prod_{i=1}^{m} \left( \prod_{j=1}^{n_i} \left[ \frac{e^{Z'\alpha}}{1+e^{Z'\alpha}} + \left(\frac{1}{1+e^{Z'\alpha}}\right)\left(\frac{\theta_i}{e^{X'\beta}+\theta_i}\right)^{\theta_i} \right] I\left(y_{ij}\right. \right. \\ \left. = 0\right) \prod_{j=1}^{n_i} \left[ \left(\frac{1}{1+e^{Z'\alpha}}\right) \left( \frac{\Gamma\left(y_{ij}+\theta_i\right)}{\Gamma(\theta_i)\Gamma\left(y_{ij}+1\right)} \right) \right. \\ \left. \left(\frac{\theta_i}{e^{X'\beta}+\theta_i}\right)^{\theta_i} \left(1-\frac{\theta_i}{e^{X'\beta}+\theta_i}\right)^{y_i} \right] I\left(y_{ij}>0\right) \\ \prod_{y_i=0} \left( \frac{1}{\sigma_{\alpha_i}\sqrt{2\pi}} e^{\frac{-(\alpha_i-\mu_{\alpha_i})^2}{2\sigma^2_{\alpha_i}}} \right) \prod_{y_i>0} \left( \frac{1}{\sigma_{\beta_i}\sqrt{2\pi}} e^{\frac{-(\beta_i-\mu_{\beta_i})^2}{2\sigma^2_{\beta_i}}} \right) \\ \left. \frac{1}{b^a\Gamma(a)} \theta^{a-1} e^{-\frac{\theta}{b}} \right) \end{aligned} \tag{6}$$

The posterior distributions were sampled by Monte Carlo Markov Chain (MCMC) [47,48] techniques available in *JAGS* [49] using *runjags* package [50] in R software. The Bayesian estimated average will be considered as the estimated effect size. The Deviance Information Criteria (DIC) [51] will be used as a goodness of fit criteria for model performance comparisons.

## 2.2. Simulation

We conducted a simulation study to compare the performance of BAMZINB with two alternative models, the Genewise Negative Binomial Generalized Linear Models (glmFit) [52] implemented in edgeR package [53] and the Bayesian hierarchical Generalized Linear Model (BhGLM) [37] implemented in BhGLM package in R software version 4.0.4 [54].

### 2.2.1. Data generation

We assumed a sample of 300 amplicon sequence variants (ASVs) for 50 and 100 subjects. We generated two sets of data for each simulation scenario with fixed treatment effects: $\beta = 0$, representing no signal, and $\beta = 2$, indicating a considerable large signal. There was only one binary covariate defined as an indicator of the exposed group, and the probability of a subject coming from the exposed

**Table 1**
Summary of simulation study parameters.

| Parameter | Ranges |
|---|---|
| Sample size, n | 50, 100 |
| Number of coefficients | One: 1, categorical (p = 0.5) |
| Effect size, β | Zero: 0, no signal<br>Two: 2, large signal |
| Over-dispersion | $\theta$: Low= 0.75, Moderate= 0.5 |
| Zero-inflation | $\alpha$: Low= 0.3, Moderate= 0.5 |
| Correlation | $\rho$: Low= 0.2, Moderate= 0.5 |

group was set at 50 %. Table 1 summarizes the simulation study parameters for 32 scenarios.

Each simulation scenario consisted of the following steps:

1. Take 50 or 100 random samples from each dataset with replacement.
2. Fit four models:
   a. BAMZINB without random intercept,
   b. BAMZINB with random intercept.
   c. glmFit [52],
   d. BhGLM [37].
3. Extract and store parameter estimates and deviances.
4. Compare parameter estimates with the true values.
5. Calculate the average absolute relative bias and deviance.

In step 2 and a BAMZINB with random intercept, adding a random intercept can be useful for detecting undefined heterogeneity in the dataset. We tried to evaluate the presence/absence of such a factor. In step 2 and a BAMZINB with random intercept, the random intercept can easily be added to the Eq. 3 as $logit(\pi_i) = \dot{Z}\alpha + b_i$ and$Log(\mu_i) = \dot{X}\beta + u_i$, where $b_i$ and $u_i$ can follow independent or bivariate normal distribution with mean zero and corresponding variances.

We set iteration parameters as follows, burning = 4000, sample = 10,000, adapt = 2000, and did not encounter any convergence issues. More information regarding the details of the model is available in the practical example explained in the Supplementary materials.

In this paper, we used Gaussian Copula [55] to accommodate the correlation among outcome variables and zero-inflated negative binomial distribution to generate each response variable (Eq. 7). The Gaussian copula conveniently describes a complex relationship [55]. The Gaussian copula function is

$$C(X, Y; \rho) = \phi(\phi^{-1}(x),\ \phi^{-1}(y); \rho) = \int_{-\infty}^{\phi^{-1}(x)} \int_{-\infty}^{\phi^{-1}(y)} \frac{1}{2\pi(1-\rho^2)^{1/2}} \left\{ \frac{-(s^2 - 2\rho st + t^2)}{2(1-\rho^2)} \right\} dsdt \quad (7)$$

where $\phi$ is the standard normal cumulative distribution function. The copula we consider here is extended for $p$ outcomes $C(Y_1. Y_2....Y_p; \rho) = \phi(\phi^{-1}(F(y_1)). \phi^{-1}(F(y_2)). ....\phi^{-1}(F(y_p)); \rho)$, where $F(y_j); j = 1. ....p$, are the zero-inflated negative binomial cumulative distribution functions.

Given the zero-inflation and over-dispersion in the gut microbiome count data, we reason that zero-inflated negative binomial distribution and Gaussian Copula to incorporate correlation represent the best choice for data generation.

#### 2.2.2. Performance comparison measures

We used the ARB as the average difference between the true values and estimated values across 100 bootstrap samples in each scenario. In addition, we used the Deviance Information Criteria (DIC) [51] to compare the goodness of fit of four models including BAMZINB without random intercept, BAMZINB with random intercept, glmFit [52], and BhGLM [37]. We reported the average ARB (SD) and average DIC (SD) among 300 simulated ASVs. Lower ARB and DIC values indicate better estimates and the preferred model.

### 2.3. Application of SKOT cohorts data

We used SKOT Cohorts (I and II) to show the application of BAMZINB on the real-life dataset. SKOT Cohorts include two studies, SKOT I and SKOT II. SKOT is the Danish abbreviation for "Dietary habits and wellbeing of young children." The main goal of SKOT studies was to investigate the relationship between obesity and metabolic syndrome with early diet and growth development. SKOT I included 311 single birth full-term infants with no chronic illness at nine months ± 2 weeks of age. All infants' fecal samples were taken at nine months and the second visit at 18 months. SKOT II included 184 infants from obese pregnant mothers who participated in the TOP study [56] (Treatment of Obese Pregnant Woman at Hvidovre Hospital in the Copenhagen area) with the same inclusion criteria as SKOT I study. Similar to SKOT I study, infants in SKOT II studies were examined at 9 ( ± 2 weeks) months and 18 ( ± 4 weeks) months [40]. In this study, Sequenced reads and infants' age were downloaded from the National Center for Biotechnology Information (NCBI) with the accession number SRP052851. We used 465 individuals in the SKOT I and SKOT II study to compare the abundance of specific bacteria at the Phylum and Class levels between 9 months and 18 months assessments in each cohort separately.

ASVs were generated from raw archived sequences with the aid of DADA2 ("High-resolution sample inference from Illumina amplicon data") implemented in the Quantitative Insights into Microbial Ecology (QIIME 2™) pipeline [57,58]. ASVs were assigned to taxonomy via a Naive Bayes classification algorithm using Silva (version 132) as the reference database from the 515F/806R region of the reference sequences [59]. The quality filtering threshold based on expected errors was set to 5, and the reverse sequence read length was truncated to 110. The maximum number of the reads was used for training the error model. We reported the average abundance difference for Actinobacteria, Bacteroidetes, Firmicutes, and Proteobacteria at the Phylum level. Laursen et al. study on SKOT Cohorts showed that these four phyla categories were the most dominant groups covering 95 % of the data [40]. In addition, we extended our investigation to include Actinobacteria, Bacteroidia, Bacilli, Gammaproteobacteria, and Alphaproteobacteria at the Class-level for infants born to healthy mothers and obese mothers from 9 months to 18 months.

## 3. Results

### 3.1. Simulation results

First, we compared the performance of four models for each scenario presented in Tables 2 and 3. Then we compared the performance of the different scenarios for each model in Figs. S1-S4.

The average ARB and average deviance results for each model in different scenarios can be found in Supplementary material.

Table 2 shows the average ARB of the estimated simulated effect size for BAMZINB with and without random intercept, BhGLM, and glmFit. For all scenarios, the maximum difference between the average ARB of the BAMZINB models and BhGLM or glmFit was less than 0.22. This result shows that the BAMZINB model performed as well as BhGLM and glmFit.

Table 3 shows the average deviance of the models in the simulation study. Average deviance showed the goodness of fit of BAMZINB models (with- and without- random intercept) were better for almost all cases when the signal and sample size were large ($\beta = 2$, $n = 100$), except for two scenarios when zero-inflation and correlation were high at both levels of over-dispersion. In other scenarios, the BhGLM or glmFit had better deviance when there was no signal, and the sample size was low ($\beta = 0$, $n = 50$), except for five scenarios when zero-inflation was low, and over-dispersion was high at both levels of correlation, when zero-inflation and correlation were low at both levels of over-dispersion, and when all properties were at the highest level. One of the BAMZINB models performed better in these five scenarios, as mentioned earlier.

### 3.2. SKOT cohorts data results

We showed the application of BAMZINB with a random intercept on the SKOT Cohorts data. Figs. 1 and 2 show the Bayesian mean (SD)

**Table 2**
The average and standard deviation of absolute relative bias among 300 simulated AVSs in the simulation study.

| Zero-inflation | Over-dispersion | Correlation | Effect size, β | Sample size, N | Average Absolute Relative Bias (SD) | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | BAMZINB_ No Random Intercept | BAMZINB+ Random Intercept | BhGLM | glmFit |
| 0.3 | 0.75 | 0.2 | Zero | 50 | 0.417 | **0.394** | 0.455 | 0.423 |
| | | | | | (0.374) | **(0.328)** | (0.407) | (0.363) |
| | | | | 100 | 0.328 | 0.331 | **0.254** | 0.268 |
| | | | | | (0.256) | (0.268) | **(0.202)** | (0.224) |
| | | | Two | 50 | 2.06 | 2.041 | 1.996 | **1.981** |
| | | | | | (0.505) | (0.591) | (0.532) | **(0.462)** |
| | | | | 100 | 2.008 | 2 | 1.969 | **1.949** |
| | | | | | (0.391) | (0.34) | (0.463) | **(0.401)** |
| 0.3 | 0.75 | 0.5 | Zero | 50 | 0.309 | 0.309 | **0.303** | 0.335 |
| | | | | | (0.258) | (0.262) | **(0.255)** | (0.28) |
| | | | | 100 | 0.224 | 0.213 | 0.224 | **0.212** |
| | | | | | (0.182) | (0.174) | (0.183) | **(0.179)** |
| | | | Two | 50 | 1.659 | 1.674 | 1.715 | **1.604** |
| | | | | | (0.766) | (0.765) | (0.768) | **(0.761)** |
| | | | | 100 | **1.564** | 1.615 | 1.62 | 1.593 |
| | | | | | **(0.736)** | (0.737) | (0.749) | (0.737) |
| 0.3 | 0.5 | 0.2 | Zero | 50 | 0.522 | **0.492** | 0.506 | 0.542 |
| | | | | | (0.489) | **(0.385)** | (0.421) | (0.513) |
| | | | | 100 | **0.273** | 0.31 | 0.284 | 0.33 |
| | | | | | **(0.209)** | (0.241) | (0.231) | (0.273) |
| | | | Two | 50 | 1.721 | 1.667 | 1.649 | **1.597** |
| | | | | | (0.777) | (0.81) | (0.849) | **(0.776)** |
| | | | | 100 | 2.047 | 1.984 | **1.933** | 2.001 |
| | | | | | (0.458) | (0.393) | **(0.424)** | (0.516) |
| 0.3 | 0.5 | 0.5 | Zero | 50 | 0.303 | 0.299 | **0.283** | 0.318 |
| | | | | | (0.309) | (0.225) | **(0.245)** | (0.23) |
| | | | | 100 | 0.317 | **0.306** | 0.354 | 0.364 |
| | | | | | (0.268) | **(0.269)** | (0.275) | (0.319) |
| | | | Two | 50 | 2.081 | 2.079 | **1.919** | 1.975 |
| | | | | | (0.551) | (0.51) | **(0.574)** | (0.552) |
| | | | | 100 | **1.608** | 1.655 | 1.621 | 1.625 |
| | | | | | **(0.764)** | (0.784) | (0.772) | (0.776) |
| 0.5 | 0.75 | 0.2 | Zero | 50 | 0.633 | 0.545 | **0.537** | 0.539 |
| | | | | | (0.598) | (0.445) | **(0.61)** | (0.507) |
| | | | | 100 | 0.358 | 0.34 | **0.331** | 0.365 |
| | | | | | (0.308) | (0.269) | **(0.265)** | (0.306) |
| | | | Two | 50 | 1.618 | 1.691 | 1.643 | **1.575** |
| | | | | | (0.821) | (0.824) | (0.777) | **(0.79)** |
| | | | | 100 | 1.646 | 1.6 | 1.62 | **1.539** |
| | | | | | (0.776) | (0.763) | (0.777) | **(0.795)** |
| 0.5 | 0.75 | 0.5 | Zero | 50 | 0.363 | **0.343** | 0.366 | 0.451 |
| | | | | | (0.317) | **(0.319)** | (0.364) | (0.401) |
| | | | | 100 | 0.295 | **0.269** | 0.271 | 0.325 |
| | | | | | (0.22) | **(0.217)** | (0.216) | (0.266) |
| | | | Two | 50 | **1.955** | 2.035 | 1.986 | 2.013 |
| | | | | | **(0.434)** | (0.464) | (0.429) | (0.395) |
| | | | | 100 | 1.486 | 1.564 | **1.444** | 1.539 |
| | | | | | (0.803) | (0.849) | **(0.814)** | (0.834) |
| 0.5 | 0.5 | 0.2 | Zero | 50 | 0.603 | 0.549 | 0.524 | **0.435** |
| | | | | | (0.547) | (0.552) | (0.487) | **(0.481)** |
| | | | | 100 | 0.459 | 0.443 | **0.41** | 0.436 |
| | | | | | (0.399) | (0.36) | **(0.377)** | (0.384) |
| | | | Two | 50 | 1.773 | 1.869 | 1.822 | **1.69** |
| | | | | | (0.95) | (1.028) | (0.892) | **(0.926)** |
| | | | | 100 | **1.575** | 1.625 | 1.648 | 1.632 |
| | | | | | **(0.772)** | (0.773) | (0.79) | (0.802) |
| 0.5 | 0.5 | 0.5 | Zero | 50 | 0.481 | 0.457 | **0.405** | 0.468 |
| | | | | | (0.417) | (0.422) | **(0.383)** | (0.408) |
| | | | | 100 | 0.359 | 0.403 | **0.297** | 0.338 |
| | | | | | (0.358) | (0.322) | **(0.254)** | (0.32) |
| | | | Two | 50 | 1.801 | 1.749 | 1.745 | **1.594** |
| | | | | | (0.791) | (0.839) | (0.774) | **(0.822)** |
| | | | | 100 | 1.727 | 1.674 | 1.715 | **1.583** |
| | | | | | (0.806) | (0.783) | (0.829) | **(0.76)** |

abundance difference from 9 months to 18 months for infants from healthy and obese mothers at the Phylum and Class levels.

Fig. 1 shows that the average abundance for Actinobacteria, Bacteroidetes, and Firmicutes increased, and the abundance of Proteobacteria decreased over time for infants with healthy mothers. Infants with obese mothers showed the same pattern for Actinobacteria and Proteobacteria. Compared to infants born to healthy mothers, the average abundance of Bacteroidetes had a remarkable lower positive difference for infants from obese mothers from 9 months to 18 months. In addition, the average abundance of Firmicutes decreased over time for infants from obese mothers.

Fig. 2 shows an increase in the average abundance of Actinobacteria, Bacteroidia, Bacilli, Gammaproteobacteria and a decrease in average abundance for Alphaproteobacteria over time for infants

**Table 3**
The average and standard deviation of deviance among 300 simulated AVSs in the simulation study.

| Zero-inflation | Overdispersion | Correlation | Effect size, β | Sample size, N | Average Deviance (SD) | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | BAMZINB_ No Random Intercept | BAMZINB+ Random Intercept | BhGLM | glmFit |
| 0.3 | 0.75 | 0.2 | Zero | 50 | 23.21 (23.267) | 22.541 (21.151) | 22.52 (21.327) | **22.47 (21.284)** |
| | | | | 100 | 45.145 (43.623) | **44.866 (43.42)** | 44.95 (43.482) | 44.93 (43.471) |
| | | | Two | 50 | 62.029 (78.834) | 73.346 (114.514) | **54.635 (59.92)** | 57.597 (61.56) |
| | | | | 100 | **98.65 (9.765)** | 105.146 (70.941) | 99.511 (10.662) | 110.248 (101.446) |
| 0.3 | 0.75 | 0.5 | Zero | 50 | 25.183 (23.313) | 25.512 (23.52) | **25.151 (23.337)** | 25.265 (23.525) |
| | | | | 100 | **41.884 (40.551)** | 42.79 (41.531) | 42.337 (41.138) | 43.372 (42.089) |
| | | | Two | 50 | 27.774 (35.505) | 26.195 (32.027) | **24.763 (23.374)** | 27.51 (38.83) |
| | | | | 100 | 55.58 (57.911) | **53.066 (50.362)** | 53.596 (50.981) | 54.18 (51.984) |
| 0.3 | 0.5 | 0.2 | Zero | 50 | 22.183 (23.241) | 21.81 (22.653) | 21.006 (21.245) | **20.759 (20.242)** |
| | | | | 100 | **43.231 (41.666)** | 43.701 (44.011) | 44.084 (43.516) | 43.27 (41.73) |
| | | | Two | 50 | **31.225 (62.697)** | 35.13 (66.232) | 33.363 (63.579) | 33.345 (68.432) |
| | | | | 100 | **97.125 (106.473)** | 120.361 (180.038) | 100.118 (108.267) | 116.785 (175.243) |
| 0.3 | 0.5 | 0.5 | Zero | 50 | **21.587 (20.283)** | 22.631 (21.745) | 21.731 (20.497) | 21.68 (20.566) |
| | | | | 100 | 53.711 (38.314) | 53.527 (37.583) | 53.543 (37.538) | **53.461 (37.952)** |
| | | | Two | 50 | 55.409 (51.182) | **52.019 (42.644)** | 56.483 (73.472) | 59.813 (75.561) |
| | | | | 100 | 53.716 (77.988) | **46.595 (46.451)** | 48.324 (53.733) | 47.397 (47.013) |
| 0.5 | 0.75 | 0.2 | Zero | 50 | **17.345 (18.581)** | 17.372 (19.064) | 18.156 (20.12) | 17.48 (18.211) |
| | | | | 100 | 37.404 (36.321) | **37.291 (36.239)** | 38.004 (37.619) | 37.679 (38.046) |
| | | | Two | 50 | 45.557 (105.697) | 36.61 (76.06) | **33.501 (67.864)** | 40.655 (90.208) |
| | | | | 100 | 46.224 (82.145) | **44.261 (72.867)** | 52.69 (96.289) | 56.871 (126.317) |
| 0.5 | 0.75 | 0.5 | Zero | 50 | 30.234 (21.422) | 30.005 (20.994) | **29.262 (21.336)** | 30.699 (21.67) |
| | | | | 100 | 84.994 (11.443) | 83.364 (13.431) | 81.791 (12.893) | **81.762 (11.512)** |
| | | | Two | 50 | 30.234 (21.422) | 30.005 (20.994) | **29.262 (21.336)** | 30.699 (21.67) |
| | | | | 100 | 68.984 (103.569) | 58.118 (60.645) | 64.185 (90.921) | **58.008 (56.018)** |
| 0.5 | 0.5 | 0.2 | Zero | 50 | **34.438 (15.757)** | 39.587 (33.979) | 38.566 (28.874) | 35.499 (19.657) |
| | | | | 100 | **43.866 (31.102)** | 45.853 (36.111) | 49.673 (48.2) | 47.02 (39.599) |
| | | | Two | 50 | **35.283 (76.951)** | 37.75 (84.385) | 40.819 (89.961) | 45.822 (102.464) |
| | | | | 100 | 56.574 (114.541) | **50.559 (104.925)** | 59.399 (129.781) | 65.171 (146.331) |
| 0.5 | 0.5 | 0.5 | Zero | 50 | 17.934 (14.447) | **16.983 (14.903)** | 17.053 (14.494) | 17.221 (14.9) |
| | | | | 100 | 42.926 (31.228) | **41.626 (29.832)** | 42.681 (30.294) | 42.741 (30.506) |
| | | | Two | 50 | 29.753 (51.233) | **25.159 (39.436)** | 26.767 (56.673) | 29.69 (54.827) |
| | | | | 100 | 43.263 (61.705) | 51.5 (103.458) | **41.772 (56.621)** | 43.915 (64.84) |

from healthy mothers. Although Actinobacteria, Bacilli, and Alphaproteobacteria showed the same pattern for infants born to obese mothers, Bacteroidia and Gammaproteobacteria decreased from 9 months to 18 months in infants with obese mothers. In addition, the average abundance change of Alphaproteobacteria was remarkably higher for infants born to obese mothers that infants of a healthy mothers over time.

Firmicutes phyla play an essential role in breaking down the carbohydrates in the infants' gut. Significance changes in Firmicutes and Bacteroidetes abundance over time could be associated with childhood obesity [60–64]. Recently, much attention was given to the Firmicutes/Bacteroidetes ratio as a relevant marker of gut microbiome-related diseases [65–68], especially in relation to obesity and inflammatory bowel disease [66,69]. Evidence shows that

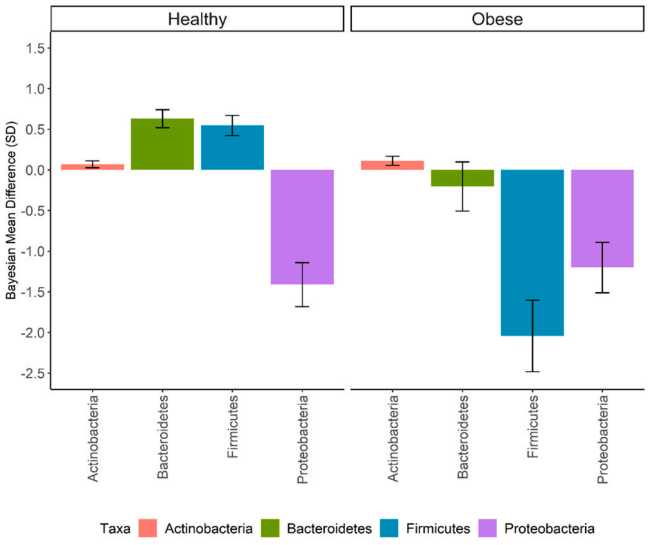


**Fig. 1.** This figure shows the average abundance difference for Actinobacteria, Bacteroidetes, Firmicutes, and Proteobacteria at the phylum-level for infants from healthy mothers and obese mothers from 9 months to 18 months.

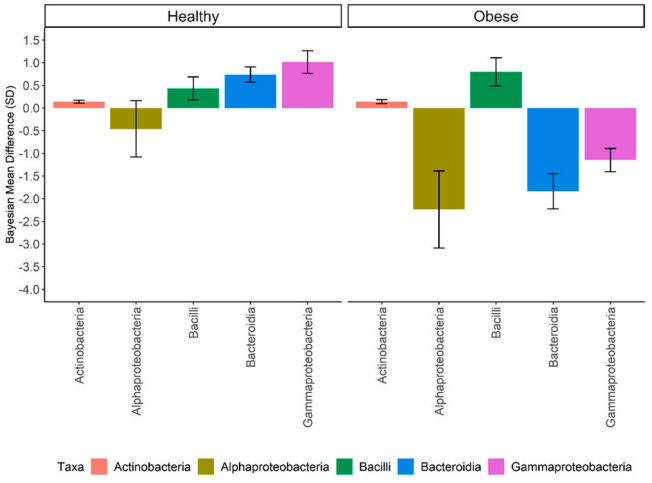


**Fig. 2.** This figure shows the average abundance difference for Actinobacteria, Bacteroidia, Bacilli, Gammaproteobacteria, and Alphaproteobacteria at the Class-level for infants from healthy mothers and obese mothers from 9 months to 18 months.

maternal microbiota is an initial provider of infants' gut microbiota, and this transfer process impacts the newborn's overall physiological condition [70,71]. Therefore, maternal obesity could be a potential risk factor for overweight or childhood obesity [72]. Muller et al. study on infants of overweight/obese mothers showed a reduction in Proteobacteria, suggesting the changes in the gram-negative bacteria such as Gammaproteobacteria may be caused by the vertical transition of maternal microbiota [73]. Several studies have found that women who had obesity prior to pregnancy or gained weight during the pregnancy had significantly different gut microbiome than normal-weight pregnant women [74,75].

## 4. Discussion

We proposed a Bayesian Marginal Zero-inflated Negative Binomial (BAMZINB) model for gut microbiome count data. BAMZINB estimates the effects of covariates on microbial composition and each taxon while allowing for incorporating correlations of residuals. Other methods in the literature, such as glmFit [52], designed to analyze microbiome count data by considering taxon one by one, have heavily depended on the utilization of normalization methods using the negative binomial distribution [76]. Tang et al. proposed BhGLM to address this issue by using raw counts and incorporating library size as an offset [37]. However, zero-inflation is one of the proven properties of gut microbiome data, and both BhGLM and glmFit models ignore it by using negative binomial distribution for microbiome count data [77,78].

In addition to taking properties of zero-inflated negative binomial distribution into account, BAMZINB is capable of considering random intercept and library size as modeling parameters. Using offset features in the BAMZINB model is similar to modeling the taxa abundance (each raw count divided by the total sequence reads for each sample), therefore accounting for the differences in the library size of the microbiome data. This also allows for an analysis of microbiome data without normalization which preserves the original nature of the data and makes the differential abundance results more interpretable [79].

Literature has shown high inter-individual variability of gut microbial composition in the early years of infancy compared to adulthood [80,81]. Generally, the microbial composition tends to be the same for everyone in adulthood. In BAMZINB, we incorporated random intercept to take inter-individual variability into account when analyzing infants' microbiome count data. Although BAMZINB has more advantages than the other two models (theoretically), simulation results showed that the BAMZINB model performed as well as BhGLM and glmFit models with respect to the ARB. One reason for this result could be the difference in the initial values for model parameters. We explained the priors for BAMZINB in Section 2.1.3. BhGLM offers three priors: Student-t (default), Double-exponential, and mixture Student-t [37]. In this paper, we used the default priors as Student-t for comparison with BAMZINB and glmFit. Further studies could focus on comparing different priors and comparing them to the other existing models. Deviance was different among 32 scenarios depending on the dataset's properties and sample size.

The real data application on SKOT Cohorts showed a different pattern over time for the average abundance of specific bacteria between infants of healthy mothers and infants of obese mothers. The structure of the SKOT Cohort data required using a random intercept due to changes in gut microbiome composition of infants in early life from 9 to 18 months, given that interpersonal changes are significantly higher in childhood than in adulthood [82]. In addition, the acquisition of zero-inflated negative binomial distribution and unstructured variance-covariance matrix in BAMZINB helped with zero-inflation, over-dispersion, and within-sample correlation issues in infants' gut microbiome data.

Future studies are needed to develop the BAMZINB method to analyze studies with random slopes and use different variance-covariance structures for multivariate analyses. There are several gaps in the literature for microbiome data analyses that need to be addressed or developed in the future. Machine learning methods are still developing in microbiome literature as they are well-known for feature selection and high-dimension data analyses [83,84].

Sample size calculation for microbiome studies is one of the common challenges. Current methods that exist in the literature are not well developed due to a lack of established metrics to define a suitable magnitude of reasonable and clinically meaningful effect size in microbiome studies [85]. Jiang et al. study explained three strategies to extract effect size for sample size and power calculations in microbiome studies, including pilot studies, data from prior studies, and simulation studies [86].

Another gap in the microbiome studies is the lack of advanced statistical tools for longitudinal analyses. Bokulich et al. proposed plugin software for the QIIME2 platform that provides various tools, including mixed models and interactive plots for microbiome longitudinal analysis [87,88]. Zhang et al. introduced a negative binomial mixed model to handle over-dispersion and variability in total reads

and dynamic trend and correlation structure among longitudinal samples [89]. More advanced longitudinal methods are needed to elucidate the relationship between taxa, environmental factors, and the host over time. In addition, one of the advantages of our proposed BAMZINB model is the ability to include multiple time measurements in the analysis.

## 5. Conclusions

It has been shown that the gut microbiome data analyses can be affected by the choice of statistical analysis method. This study proposed the BAMZINB method to account for over-dispersion, zero-inflation, multivariate correlation structure, and dimensionality issues in the infants' gut microbiome data. We know from previous studies that the consequences of ignoring these features of gut microbiome data cause a lack of precision in estimating effect sizes and loss of statistical power. In this study, we compared the performance of several statistical methods in 32 scenarios and showed the application of BAMZINB on a real data set. The findings of this study could help other research groups compare the properties of their dataset with one of the 32 scenarios and make a better decision when choosing a statistical analysis method.

## CRediT authorship contribution statement

MH identified the need for simulation studies of microbiota data analysis methods. MH, ID, PA, and AS developed the theory and performed the computations. All the authors discussed the results and contributed to the final manuscript.

## Conflict of Interest

The authors have no conflict of interest to declare.

## Acknowledgement

We would like to thank the reviewers for their thoughtful comments and efforts towards improving our manuscript. We would also thank the following people for their insightful comments and support: Dr. Benjamin I. Chung, Dr. Duane Moser, Dr. Brian Lanoil.

## Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2023.02.027.

## References

[1] Hatzenpichler R, Krukenberg V, Spietz RL, Jay ZJ. Next-generation physiology approaches to study microbiome function at single cell level. Nat Rev Microbiol 2020;18(4):241–56.
[2] Dominguez-Bello MG, Godoy-Vitorino F, Knight R, Blaser MJ. Role of the microbiome in human development. Gut 2019;68(6):1108–14.
[3] Aagaard K, Ma J, Antony KM, Ganu R, Petrosino J, Versalovic J. The placenta harbors a unique microbiome. Sci Transl Med 2014;6(237). 237ra65-ra65.
[4] Fan Y, Pedersen O. Gut microbiota in human metabolic health and disease. Nat Rev Microbiol 2021;19(1):55–71.
[5] Milani C, Duranti S, Bottacini F, Casey E, Turroni F, Mahony J, et al. The first microbial colonizers of the human gut: composition, activities, and health implications of the infant gut microbiota. Microbiol Mol Biol Rev 2017;81(4):e00036–17.
[6] Wang S, Ryan CA, Boyaval P, Dempsey EM, Ross RP, Stanton C. Maternal vertical transmission affecting early-life microbiota development. Trends Microbiol 2020;28(1):28–45.
[7] Van Daele E, Knol J, Belzer C. Microbial transmission from mother to child: improving infant intestinal microbiota development by identifying the obstacles. Crit Rev Microbiol 2019;45(5–6):613–48.
[8] Koo H, Crossman DK, Morrow CD. Strain tracking to identify individualized patterns of microbial strain stability in the developing infant gut ecosystem. Front Pediatr 2020:8.
[9] Dogra SK, Doré J, Damak S. Gut microbiota resilience: definition, link to health and strategies for intervention. Front Microbiol 2020;11:2245.
[10] Sanders DJ, Inniss S, Sebepos-Rogers G, Rahman FZ, Smith AM. The role of the microbiome in gastrointestinal inflammation. Biosci Rep 2021;41(6). BSR20203850.
[11] Cheng H-Y, Ning M-X, Chen D-K, Ma W-T. Interactions between the gut microbiota and the host innate immune response against pathogens. Front Immunol 2019;10:607.
[12] Robertson RC, Manges AR, Finlay BB, Prendergast AJ. The human microbiome and child growth–first 1000 days and beyond. Trends Microbiol 2019;27(2):131–47.
[13] Ivashkin V, Poluektov Y, Kogan E, Shifrin O, Sheptulin A, Kovaleva A, et al. Disruption of the pro-inflammatory, anti-inflammatory cytokines and tight junction proteins expression, associated with changes of the composition of the gut microbiota in patients with irritable bowel syndrome. Plos One 2021;16(6):e0252930.
[14] Castaner O, Goday A, Park Y-M, Lee S-H, Magkos F, Shiow S-ATE, et al. The gut microbiome profile in obesity: a systematic review. Int J Endocrinol 2018;2018.
[15] Turroni F, Milani C, Ventura M, van Sinderen D. The human gut microbiota during the initial stages of life: Insights from bifidobacteria. Curr Opin Biotechnol 2022;73:81–7.
[16] Baothman OA, Zamzami MA, Taher I, Abubaker J, Abu-Farha M. The role of gut microbiota in the development of obesity and diabetes. Lipids Health Dis 2016;15(1):1–8.
[17] Kasai C, Sugimoto K, Moritani I, Tanaka J, Oya Y, Inoue H, et al. Comparison of the gut microbiota composition between obese and non-obese individuals in a Japanese population, as analyzed by terminal restriction fragment length polymorphism and next-generation sequencing. BMC Gastroenterol 2015;15(1):1–10.
[18] Wu H-J, Wu E. The role of gut microbiota in immune homeostasis and autoimmunity. Gut Microbes 2012;3(1):4–14.
[19] Petersen C, Dai DL, Boutin RC, Sbihi H, Sears MR, Moraes TJ, et al. A rich meconium metabolome in human infants is associated with early-life gut microbiota composition and reduced allergic sensitization. Cell Rep Med 2021;2(5):100260.
[20] Manasson J, Blank RB, Scher JU. The microbiome in rheumatology: where are we and where should we go? Ann Rheum Dis 2020;79(6):727–33.
[21] Ma Q, Xing C, Long W, Wang HY, Liu Q, Wang R-F. Impact of microbiota on central nervous system and neurological diseases: the gut-brain axis. J Neuroinflamm 2019;16(1):1–14.
[22] Liu S, Zhao W, Lan P, Mou X. The microbiome in inflammatory bowel diseases: from pathogenesis to therapy. Protein Cell 2021;12(5):331–45.
[23] Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. Metagenomic microbial community profiling using unique clade-specific marker genes. Nat Methods 2012;9(8):811–4.
[24] Xia Y, Sun J. Hypothesis testing and statistical analysis of microbiome. Genes Dis 2017;4(3):138–48.
[25] Vujkovic-Cvijin I, Sklar J, Jiang L, Natarajan L, Knight R, Belkaid Y. Host variables confound gut microbiota studies of human disease. Nature 2020;587(7834):448–54.
[26] Kinross JM, Darzi AW, Nicholson JK. Gut microbiome-host interactions in health and disease. Genome Med 2011;3(3):1–12.
[27] Maier L, Pruteanu M, Kuhn M, Zeller G, Telzerow A, Anderson EE, et al. Extensive impact of non-antibiotic drugs on human gut bacteria. Nature 2018;555(7698):623–8.
[28] Zhu W, Winter MG, Byndloss MX, Spiga L, Duerkop BA, Hughes ER, et al. Precision editing of the gut microbiota ameliorates colitis. Nature 2018;553(7687). 208-11.
[29] Chen YY, Zhao X, Moeder W, Tun HM, Simons E, Mandhane PJ, et al. Impact of maternal intrapartum antibiotics, and caesarean section with and without labour on bifidobacterium and other infant gut microbiota. Microorganisms 2021;9(9):1847.
[30] Chen Y, McCarthy D, Robinson M, Smyth GK. edgeR: differential expression analysis of digital gene expression data User's Guide. Bioconductor User's Guide 2014.
[31] Anders S, Huber W. Differential expression analysis for sequence count data. Nat Preced 2010. 1-.
[32] Anders S, McCarthy DJ, Chen Y, Okoniewski M, Smyth GK, Huber W, et al. Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. Nat Protoc 2013;8(9):1765–86.
[33] Zhang X, Guo B, Yi N. Zero-Inflated gaussian mixed models for analyzing longitudinal microbiome data. Plos One 2020;15(11). e0242073.
[34] Meng C, Zeleznik OA, Thallinger GG, Kuster B, Gholami AM, Culhane AC. Dimension reduction techniques for the integrative analysis of multi-omics data. Brief Bioinforma 2016;17(4):628–41.
[35] Jiang D, Armour CR, Hu C, Mei M, Tian C, Sharpton TJ, et al. Microbiome multi-omics network analysis: statistical considerations, limitations, and opportunities. Front Genet 2019;10:995.
[36] Brooks S, Gelman A, Jones G, Meng X-L. Handbook of markov chain monte carlo. CRC press; 2011.
[37] Yi N, Tang Z, Zhang X, Guo B. BhGLM: Bayesian hierarchical GLMs and survival models, with applications to genomics and epidemiology. Bioinformatics 2019;35(8):1419–21.
[38] Geert Molenberghs GV. Models for Discrete Longitudinal Data. New York: Springer; 2010.
[39] Laursen MF, Laursen RP, Larnkjær A, Mølgaard C, Michaelsen KF, Frøkiær H, et al. Faecalibacterium gut colonization is accelerated by presence of older siblings. mSphere 2017;2(6):e00448–17.
[40] Laursen MF, Andersen LBB, Michaelsen KF, Mølgaard C, Trolle E, Bahl MI, et al. Infant gut microbiota development is driven by transition to family foods independent of maternal obesity. mSphere 2016;1(1):e00069–15.

[41] Laursen MF, Zachariassen G, Bahl MI, Bergström A, Høst A, Michaelsen KF, et al. Having older siblings is associated with gut microbiota development during early childhood. BMC Microbiol 2015;15. 154-.
[42] Preisser JS, Das K, Long DL, Divaris K. Marginalized zero-inflated negative binomial regression with application to dental caries. Stat Med 2016;35(10):1722–35.
[43] Lambert D. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics 1992;34(1):1–14.
[44] Mullahy J. Specification and testing of some modified count data models. J Econ 1986;33(3). 341-65.
[45] Martin J, Hall DB. Marginal zero-inflated regression models for count data. J Appl Stat 2017;44(10):1807–26.
[46] Park ES, Park J, Lomax TJ. A fully Bayesian multivariate approach to before–after safety evaluation. Accid Anal Prev 2010;42(4):1118–27.
[47] Gamerman D, Lopes HF. Markov chain Monte Carlo: stochastic simulation for Bayesian inference. CRC Press; 2006.
[48] El-Basyouny K, Sayed T. Full Bayes approach to before-and-after safety evaluation with matched comparisons: case study of stop-sign in-fill program. Transp Res Rec 2010;2148(1):1–8.
[49] Hornik K., Leisch F., Zeileis A., editors. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. Proceedings of DSC; 2003.
[50] Denwood MJ runjags. An R package providing interface utilities, model templates, parallel computing methods and additional distributions for MCMC models in JAGS. J Stat Softw 2016;71(9):1–25.
[51] Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A. Bayesian measures of model complexity and fit. J R Stat Soc: Ser B (Stat Methodol) 2002;64(4):583–639.
[52] McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res 2012;40(10):4288–97.
[53] Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 2010;26(1):139–40.
[54] Team RCR. A Lang Environ Stat Comput 2013.
[55] Andersen LBB, Pipper CB, Trolle E, Bro R, Larnkjær A, Carlsen E, et al. Maternal obesity and offspring dietary patterns at 9 months of age. Eur J Clin Nutr 2015;69(6):668.
[56] Renault KM, Nørgaard K, Nilas L, Carlsen EM, Cortes D, Pryds O, et al. The Treatment of Obese Pregnant Women (TOP) study: a randomized controlled trial of the effect of physical activity intervention assessed by pedometer with or without dietary intervention in obese pregnant women. Am J Obstet Gynecol 2014;210(2). 134. e1-. e9.
[57] 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.
[58] 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.
[59] Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res 2012;41(D1):D590–6.
[60] CMdSP Indiani, Rizzardi KF, Castelo PM, Ferraz LFC, Darrieux M, Parisotto TM. Childhood obesity and Firmicutes/Bacteroidetes ratio in the gut microbiota: a systematic review. Child Obes 2018;14(8):501–9.
[61] Bergström A, Skov TH, Bahl MI, Roager HM, Christensen LB, Ejlerskov KT, et al. Establishment of intestinal microbiota during early life: a longitudinal, explorative study of a large cohort of Danish infants. Appl Environ Microbiol 2014;80(9):2889–900.
[62] Scheepers L, Penders J, Mbakwa C, Thijs C, Mommers M, Arts I. The intestinal microbiota composition and weight development in children: the KOALA Birth Cohort Study. Int J Obes 2015;39(1):16–25.
[63] Xu P, Li M, Zhang J, Zhang T. Correlation of intestinal microbiota with overweight and obesity in Kazakh school children. BMC Microbiol 2012;12(1):1–6.
[64] Borgo F, Verduci E, Riva A, Lassandro C, Riva E, Morace G, et al. Relative abundance in bacterial and fungal gut microbes in obese children: a case control study. Child Obes 2017;13(1):78–84.
[65] Magne F, Gotteland M, Gauthier L, Zazueta A, Pesoa S, Navarrete P, et al. The firmicutes/bacteroidetes ratio: a relevant marker of gut dysbiosis in obese patients? Nutrients 2020;12(5):1474.
[66] Stojanov S, Berlec A, Štrukelj B. The influence of probiotics on the firmicutes/bacteroidetes ratio in the treatment of obesity and inflammatory bowel disease. Microorganisms 2020;8(11):1715.
[67] Takezawa K, Fujita K, Matsushita M, Motooka D, Hatano K, Banno E, et al. The Firmicutes/Bacteroidetes ratio of the human gut microbiota is associated with prostate enlargement. Prostate 2021;81(16):1287–93.
[68] Houtman TA, Eckermann HA, Smidt H, de Weerth C. Gut microbiota and BMI throughout childhood: the role of firmicutes, bacteroidetes, and short-chain fatty acid producers. Sci Rep 2022;12(1):1–13.
[69] Sutoyo DA, Atmaka DR, Sidabutar LMG. Dietary factors affecting firmicutes and bacteroidetes ratio in solving obesity problem: a literature review. Media Gizi Indones 2020;15(2):94–109.
[70] Galley JD, Bailey M, Kamp Dush C, Schoppe-Sullivan S, Christian LM. Maternal obesity is associated with alterations in the gut microbiome in toddlers. PLoS One 2014;9(11):e113026.
[71] Kozyrskyj A, Kalu R, Koleva P, Bridgman S. Fetal programming of overweight through the microbiome: boys are disproportionately affected. J Dev Orig Health Dis 2016;7(1):25–34.
[72] Trandafir L, Temneanu O. Pre and post-natal risk and determination of factors for child obesity. J Med Life 2016;9(4):386.
[73] Mueller N, Shin H, Pizoni A, Werlang I, Matte U, Goldani M, et al. Birth mode-dependent association between pre-pregnancy maternal weight status and the neonatal intestinal microbiome. Sci Rep 2016;6:23133.
[74] Collado M, Isolauri E, Laitinen K, Salminen S. Distinct composition of gut microbiota during pregnancy in overweight and normal-weight women. Am J Clin Nutr 2008;88:894–9.
[75] Santacruz A, Collado M, García-Valdés L, Segura M, Martín-Lagos J, Anjos T, et al. Gut microbiota composition is associated with body weight, weight gain and biochemical parameters in pregnant women. Br J Nutr 2010;104:83–92.
[76] Lin H, Peddada SD. Analysis of microbial compositions: a review of normalization and differential abundance analysis. npj Biofilms Micro 2020;6(1):60.
[77] Silverman JD, Roche K, Mukherjee S, David LA. Naught all zeros in sequence count data are the same. Comput Struct Biotechnol J 2020;18:2789–98.
[78] Kaul A, Mandal S, Davidov O, Peddada SD. Analysis of microbiome data in the presence of excess zeros. Front Microbiol 2017:8.
[79] Zhang X, Mallick H, Yi N. Zero-inflated negative binomial regression for differential abundance testing in microbiome studies. J Bioinforma Genom 2016;2(2).
[80] Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, et al. Human gut microbiome viewed across age and geography. Nature 2012;486(7402):222–7.
[81] Dominguez-Bello MG, Costello EK, Contreras M, Magris M, Hidalgo G, Fierer N, et al. Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. Proc Natl Acad Sci 2010;107(26):11971–5.
[82] Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, et al. Human gut microbiome viewed across age and geography. Nature 2012;486(7402):222–7.
[83] Pasolli E, Truong DT, Malik F, Waldron L, Segata N. Machine learning meta-analysis of large metagenomic datasets: tools and biological insights. PLoS Comput Biol 2016;12(7):e1004977.
[84] Knight R, Vrbanac A, Taylor BC, Aksenov A, Callewaert C, Debelius J, et al. Best practices for analysing microbiomes. Nat Rev Microbiol 2018;16(7):410–22.
[85] Galloway-Peña J, Hanson B. Tools for analysis of the microbiome. Dig Dis Sci 2020;65(3):674–85.
[86] Jiang L, Ferdous T, Dinu I, Groizeleau J, Danska J, McCoy KD, et al. Beta-diversity distance matrices for microbiome sample size and power calculations—how to obtain good estimates. Comput Struct Biotechnol J 2022.
[87] Bokulich NA, Dillon MR, Zhang Y, Rideout JR, Bolyen E, Li H, et al. q2-longitudinal: longitudinal and paired-sample analyses of microbiome data. MSystems 2018;3(6):e00219–18.
[88] Bokulich NA, Zhang Y, Dillon M, Rideout JR, Bolyen E, Li H, et al. q2-longitudinal: a QIIME 2 plugin for longitudinal and paired-sample analyses of microbiome data. BioRxiv 2017:223974.
[89] Zhang X, Pei Y-F, Zhang L, Guo B, Pendegraft AH, Zhuang W, et al. Negative binomial mixed models for analyzing longitudinal microbiome data. Front Microbiol 2018;9:1683.