PLOS ONE

RESEARCH ARTICLE

Zero-Inflated gaussian mixed models for analyzing longitudinal microbiome data

Xinyan Zhang[1], **Boyi Guo**[2], **Nengjun Yi**[2]*

1 Department of Statistics and Data Analytics, Kennesaw State University, Kennesaw, GA, United States of America, **2** Department of Biostatistics, University of Alabama at Birmingham, Birmingham, AL, United States of America

* nyi@uab.edu

Citation: Zhang X, Guo B, Yi N (2020) Zero-Inflated gaussian mixed models for analyzing longitudinal microbiome data. PLoS ONE 15(11): e0242073. https://doi.org/10.1371/journal.pone.0242073

Editor: Christopher Staley, University of Minnesota Twin Cities, UNITED STATES

Received: July 2, 2020

Accepted: October 26, 2020

Published: November 9, 2020

Data Availability Statement: The ZIGMMs is implemented in the R package NBZIMM, which is freely available from GitHub: http://github.com//nyiuab//NBZIMM.

Funding: The authors received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Motivation

The human microbiome is variable and dynamic in nature. Longitudinal studies could explain the mechanisms in maintaining the microbiome in health or causing dysbiosis in disease. However, it remains challenging to properly analyze the longitudinal microbiome data from either 16S rRNA or metagenome shotgun sequencing studies, output as proportions or counts. Most microbiome data are sparse, requiring statistical models to handle zero-inflation. Moreover, longitudinal design induces correlation among the samples and thus further complicates the analysis and interpretation of the microbiome data.

Results

In this article, we propose zero-inflated Gaussian mixed models (ZIGMMs) to analyze longitudinal microbiome data. ZIGMMs is a robust and flexible method which can be applicable for longitudinal microbiome proportion data or count data generated with either 16S rRNA or shotgun sequencing technologies. It can include various types of fixed effects and random effects and account for various within-subject correlation structures, and can effectively handle zero-inflation. We developed an efficient Expectation-Maximization (EM) algorithm to fit the ZIGMMs by taking advantage of the standard procedure for fitting linear mixed models. We demonstrate the computational efficiency of our EM algorithm by comparing with two other zero-inflated methods. We show that ZIGMMs outperform the previously used linear mixed models (LMMs), negative binomial mixed models (NBMMs) and zero-inflated Beta regression mixed model (ZIBR) in detecting associated effects in longitudinal microbiome data through extensive simulations. We also apply our method to two public longitudinal microbiome datasets and compare with LMMs and NBMMs in detecting dynamic effects of associated taxa.

1. Introduction

Since birth, the human body becomes host to millions of microbiota that influence health across whole lives and potentially over generations [1]. The combination of microbiota and the associated genomes (metagenome) interact with the host environment to form the human

microbiome [2]. Recent studies have investigated static associations between the human microbiome and many human diseases such as obesity, diabetes, inflammatory bowel disease, irritable bowel syndrome, vaginosis and even cancers [2–7]. However, the microbes could interact with the host and the environment over time [8]. Thus the human microbiome is variable and dynamic in nature, and the infant microbiome could possibly have subsequent implications in future health through the human host's early life and even adulthood [9]. Longitudinal studies could explain the mechanisms in maintaining the microbiome in health or causing dysbiosis in disease [10]. Recent microbiome studies have employed the longitudinal study design to investigate the dynamic changes of microbial abundance over time and the associations between the microbiome and host environmental/clinical factors [11–15].

As a result of the research interests and the development of high-throughput metagenomics, a large amount of longitudinal 16S rRNA data or metagenome shotgun sequencing data has been generated [16]. It is known that 16S rRNA data or metagenome shotgun sequencing data are both processed and output as number of fragments or reads (in terms of raw or relative abundance) in operational taxonomic units (OTUs) or functional units with various bioinformatics pipelines, such as QIIME and mothur for 16S rRNA data and MetaPhlAn, PhyloSift, and Kraken for shotgun libraries [16]. Although some of the pipelines output the microbiome data in raw counts, others, such as MetaPhlAn, output the relative abundance from shotgun data in proportions.

However, it remains challenging to properly analyze and interpret the longitudinal microbiome data, especially in terms of proportion. Due to both biological and technical reasons, microbiome sequencing data is sparse [17]. Moreover, longitudinal microbiome data possesses special features, for example, time-dependent effects and correlations among the samples within the subjects, for which tailored statistical methods are required [10]. La Rosa, Warner [12], as several previous studies, used linear mixed models (LMMs) to account for correlations in longitudinal microbiome studies [12,18–21]. However, using LMMs is not capable to correct for excess zeros in microbiome data. Recently, we have developed negative binomial mixed models (NBMMs) for analyzing longitudinal microbiome count data, but have not explicitly modeled zero-inflation [22,23]. Romero, Hassan [24] used zero-inflated negative binomial mixed-effects models to analyze longitudinal count data. Neither NBMMs nor the zero-inflated negative binomial mixed-effects models is applicable in analyzing longitudinal microbiome proportion data. Alternatively, Chen and Li [25] proposed a zero-inflated Beta regression model with random effects (ZIBR) for analyzing longitudinal microbiome proportions. However, according to the manual of R package **ZIBR** [26], ZIBR cannot handle missing data, which means each subject must have the same number of time points. Moreover, these two zero-inflated methods have not been developed to account for within-subject correlations and may be computationally sub-optimal for analyzing many OTUs. Thus, statistical models are needed to account for sample correlations over time as well as zero-inflation and other properties of microbiome data [25,27,28].

We here propose zero-inflated Gaussian mixed models (ZIGMMs) and an efficient algorithm to address the previous limitations. Our method is robust and flexible and can analyze longitudinal microbiome proportion data and count data generated with either 16S rRNA or shotgun sequencing technologies. The proposed model can effectively deal with zero-inflation and can include various types of fixed and random effects and within-subject correlation structures. We develop an efficient Expectation-Maximization (EM) algorithm to fit the ZIGMMs by taking advantage of the standard procedure for fitting LMMs. We show computational efficiency of ZIGMMs compared with the other two zero-inflated methods, ZIBR and zero-inflated negative binomial mixed models implemented in the R package **glmmTMB**. Extensive simulations demonstrate that our ZIGMMs outperform the various previously used methods

in detecting associated effects in longitudinal microbiome data. We also apply our method to a shotgun longitudinal microbiome proportion data and a 16S rRNA microbiome count data in detecting dynamic effects of associated taxa. We have implemented the ZIGMMs in the R package **NBZIMM**, which is freely available from the public GitHub repository http://github.com//nyiuab//NBZIMM.

2. Methods

2.1 Zero-Inflated Gaussian Mixed Models (ZIGMMs)

In a longitudinal microbiome study, we collect n subjects and measure each subject at multiple time points t_{ij}, $j = 1, \cdots, n_i$; $i = 1, \cdots, n$. For the j-th sample of the i-th subject, we denote c_{ijh} the observed count for the h-th taxon at certain taxonomic levels (OTU, e.g. species, genus, classes, etc.). As many previous methods, we analyze one taxon at a time. We first illustrate our model in analyzing the longitudinal microbiome proportion data. We transform the proportions of relative abundance with $\text{arcsine}(\sqrt{c_{ijh}/T_{ij}})$, where T_{ij} denotes the total sequence read. For notational simplification, we denote $y_{ij} = \text{arcsine}(\sqrt{c_{ijh}/T_{ij}})$ for any given taxon h. For taxa with excessive zeros, it can be assumed that transformed values y_{ij} may come from either a degenerate distribution having the point mass at zero (zero state) or a Gaussian (i.e., normal) distribution [17]. Thus, the transformed values y_{ij} can be modeled with the zero-inflated Gaussian distribution:

$$y_{ij} \sim \begin{cases} 0 & \text{with probability } p_{ij} \\ N(y_{ij}|\mu_{ij}, \sigma^2)\ y_{ij} \geq 0 & \text{with probability } 1 - p_{ij} \end{cases} \quad (1)$$

where μ_{ij} and σ are the mean and standard deviation parameters in normal distribution, respectively, and p_{ij} is the unknown probability that y_{ij} is from the zero state. The means μ_{ij} are expressed as:

$$\mu_{ij} = X_{ij}\beta + G_{ij}b_i \quad (2)$$

where X_{ij} is the vector of covariates for the j-th sample of the i-th subject; β is the vector of fixed effects (i.e. population-level effects), representing the average effects of the covariates over the subjects; b_i is the vector of subject-specific effects, or called random effects, and G_{ij} is the vector of group-level covariates, which is a subset of the population-level covariates X_{ij}. For longitudinal studies, X_{ij} could be $(1, X_i)$, $(1, X_i, t_{ij})$, or $(1, X_i, t_{ij}, X_i^s t_{ij})$, where X_i^s is the variable of interest in X_i, for example, an indicator variable for the case group and the control group. G_{ij} could be 1, i.e. only including the subject-specific intercept, or $(1, t_{ij})$, i.e. including the subject-specific intercept and time effect.

The random effects are assumed to follow a multivariate normal distribution:

$$b_i \sim N(0, \Psi_b) \quad (3)$$

where Ψ_b is the variance-covariance matrix which can be defined as a general positive-definite matrix accounting for the correlation among the random covariates. In most applications we restrict Ψ_b to be a diagonal matrix for simplicity.

The zero-inflation probabilities p_{ij} are assumed to relate some covariates through the logit link function:

$$\text{logit}(p_{ij}) = Z_{ij}\alpha \quad (4)$$

where Z_{ij} includes some covariates that are potentially associated with the zero state. The simplest zero-inflation model includes only the intercept in Z_{ij}, resulting in the same probability of belonging to the zero state for all zeros. We can also add the random-effect terms into the above model:

$$\text{logit}(p_{ij}) = Z_{ij}\alpha + G_{ij}a_i \tag{5}$$

where the random effects a_i are assumed to follow a multivariate normal distribution:

$$a_i \sim N(0, \Psi_a) \tag{6}$$

As an alternative, for longitudinal microbiome count data, we transform the observed count data with $y_{ij} = \log_2(c_{ijh}+1)$, which equals zero if $c_{ijh} = 0$. We assume the y_{ij} can be modeled with the zero-inflated Gaussian distribution, with the means μ_{ij} being expressed as:

$$\mu_{ij} = \log(T_{ij}) + X_{ij}\beta + G_{ij}b_i \tag{7}$$

2.2 The EM algorithm for fitting the ZIGMMs

We propose an EM algorithm to fit the ZIGMMs. We introduce latent indicator variables $\xi = (\xi_{i1}, \cdots, \xi_{in_j})$ to distinguish the zero state and the Gaussian state, where $\xi_{ij} = 1$ when y_{ij} is from the zero state and $\xi_{ij} = 0$ when y_{ij} is from the normal distribution. The log-likelihood with the complete data (y, ξ) is given by:

$$L(\Phi; y, \xi) = \sum_{i=1}^{n}\sum_{j=1}^{n_i}(1-\xi_{ij})\log(N(y_{ij}|\mu_{ij}, \sigma^2)) + \sum_{i=1}^{n}\sum_{j=1}^{n_i}\log[p_{ij}^{\xi_{ij}}(1-p_{ij})^{1-\xi_{ij}}] \tag{8}$$

where Φ represents all the parameters (including random effects) in the ZIGMMs.

The EM algorithm replaces the indicator variables ξ_{ij} by their conditional expectations $\hat{\xi}_{ij}$ (E-step), and then updates the parameters by maximizing $L(\Phi; y, \hat{\xi})$ (M-step). The conditional expectation of ξ_{ij} can be calculated as:

$$\begin{aligned}\hat{\xi}_{ij} &= p(\xi_{ij} = 1|\Phi, y_{ij}) \\ &= \frac{p(y_{ij}|\mu_{ij}, \sigma^2, \xi_{ij} = 1)p(\xi_{ij} = 1|p_{ij})}{p(y_{ij}|\mu_{ij}, \sigma^2, \xi_{ij} = 0)p(\xi_{ij} = 0|p_{ij}) + p(y_{ij}|\mu_{ij}, \sigma^2, \xi_{ij} = 1)p(\xi_{ij} = 1|p_{ij})}\end{aligned} \tag{9}$$

If $y_i \neq 0$, we have $p(y_{ij}|\mu_{ij},\sigma^2,\xi_{ij} = 1) = 0$, and thus $\hat{\xi}_{ij} = 0$.

If $y_i = 0$, we have

$$\hat{\xi}_{ij} = \left[\frac{p(\xi_{ij} = 0|p_{ij})}{p(\xi_{ij} = 1|p_{ij})}p(y_{ij} = 0|\mu_{ij}, \sigma^2, \xi_{ij} = 0) + 1\right]^{-1} = \left[\frac{1-p_{ij}}{p_{ij}}N(y_{ij} = 0|\mu_{ij}, \sigma^2) + 1\right]^{-1}.$$

The parameters in the Gaussian distribution can be updated by fitting a weighted linear mixed model with $(1 - \hat{\xi}_{ij})$ as weights:

$$y_{ij} = X_{ij}\beta + G_{ij}b_i + (1-\hat{\xi}_{ij})^{-1/2}e_{ij},\ b_i \sim N_q(0, \Psi_b),\ e_{ij} \sim N(0, \sigma^2) \tag{10}$$

If the zero-inflation part does not include the random-effect term, the parameters can be updated by running a binomial logistic regression with $\hat{\xi}_{ij}$ as response:

$$\hat{\xi}_{ij} \sim \text{Bin}(1, p_{ij}), \text{logit}(p_{ij}) = Z_{ij}\alpha \tag{11}$$

Otherwise, we can fit the binomial logistic mixed model:

$$\hat{\xi}_{ij} \sim \text{Bin}(1, p_{ij}), \text{logit}(p_{ij}) = Z_{ij}\alpha + G_{ij}a_i, a_i \sim N(0, \Psi_a) \tag{12}$$

The EM algorithm starts from plausible values for the parameters and then updates the parameters as described above until convergence. We use the criterion $\sum_{i=1}^{n}\sum_{j=1}^{n_i}[(\eta_{ij}^{(t)} - \eta_{ij}^{(t-1)})^2 + (\gamma_{ij}^{(t)} - \gamma_{ij}^{(t-1)})^2] < \varepsilon\left(\sum_{i=1}^{n}\sum_{j=1}^{n_i}[(\eta_{ij}^{(t)})^2 + (\gamma_{ij}^{(t)})^2]\right)$ to assess convergence, where $\eta_{ij}^{(t)} = X_{ij}\beta^{(t)} + G_{ij}b_i^{(t)}$, $\gamma_{ij}^{(t)} = Z_{ij}\alpha^{(t)} + G_{ij}a_i^{(t)}$, and ε is a small value (say 10^{-5}). At convergence, we obtain the maximum likelihood estimates of the Gaussian-state fixed effects and the associated standard deviations from the final weighted LMM. We then can test H_0: $\beta_k = 0$ according to the LMM framework. We also obtain the estimates of the zero-state fixed effects and the associated standard deviations from the final binomial logistic (or mixed) model. Thus, we can test H_0: $\alpha_k = 0$ following the GLM or GLMM framework.

2.3 Accounting for within-subject correlations

The weighted linear mixed model (9) restricts the within-subject errors to be independent. We can relax the assumption of independent within-subject errors to account for special within-subject correlation structures:

$$e_i = (e_{i1}, \cdots, e_{in_i})' \sim N(0, \sigma^2 R_i) \tag{13}$$

where R_i is a correlation matrix. Pinheiro and Bates [29] described several ways to specify the correlation matrix R_i, for example, autoregressive of order 1, AR(1), or continuous-time AR (1), all of which can be incorporated into our ZIGMMs.

2.4 Software implementation

The proposed method has been implemented in the function lme.zig, which is part of the R package **NBZIMM**. In the E-step of the EM algorithm, the conditional expectation of ξ_{ij} can be calculated as in Eq (9). In the M-step, the parameters in the Gaussian distribution can be updated by repeated calls to the function lme in the R package **nlme** to fit the weighted linear mixed model with (1 - $\hat{\xi}_{ij}$) as weights. The other parameters can be updated by repeated calls to the functions glm or glmPQL in the package **MASS** to fit the binomial logistic or mixed logistic model. The function lme is the recommended tool for analyzing linear mixed models. The function lme.zig incorporates the nice features of lme, such as dealing with any types of random effects and within-subject correlation structures. Thus, it provides an efficient and flexible tool for analyzing zero-inflated longitudinal microbiome data. The package **NBZIMM** is freely available from the public GitHub repository http://github.com//nyiuab//NBZIMM.

3. Results

3.1 Simulation studies

3.1.1 Assess the ZIGMMs in analyzing microbiome proportion data. 3.1.1.1 Simulation design. To evaluate the proposed ZIGMMs, we performed extensive simulations. We first

evaluated the ZIGMMs in analyzing microbiome proportion data. We compared ZIGMMs with ZIBR proposed by Chen and Li [25]. We used the function simulate_zero_inflated_beta_random_effect_data in the R package **ZIBR** [25] to simulate longitudinal microbiome proportion data from zero-inflated beta distribution:

$$y_{ij} \sim \begin{cases} 0 & \text{with probability } p_{ij} \\ Beta(y_{ij}|u_{ij}\phi, (1-u_{ij})\phi) & \text{with probability } 1-p_{ij} \end{cases}$$

with the link functions $\text{logit}(p_{ij}) = Z_{ij}\alpha+G_{ij}a_i$ and $\text{logit}(u_{ij}) = X_{ij}\beta+G_{ij}b_i$. We employed a case-control longitudinal design with the following settings: 5 time points for each subject, fixed effects in both parts, random intercepts in both parts (i.e. $G_{ij} = 1$)). We also considered three numbers of subjects: $n = 50$, 100 and 150, half of which were designated to be cases. We set the regression coefficients as $\alpha = (\alpha_0, \alpha_1) = (-0.5, 0)$, $\beta = (\beta_0, \beta_1) = (-0.5, 0)$ to test for false positive rate; while $\alpha = (\alpha_0, \alpha_1) = (-0.5, 0.3)$, $\beta = (\beta_0, \beta_1) = (-0.5, 0.3)$ to test for power at a low effect setting and $\alpha = (\alpha_0, \alpha_1) = (-0.5, 0.5)$, $\beta = (\beta_0, \beta_1) = (-0.5, 0.5)$ to test for power at a high effect setting. The variance of the random effects to control a_i and b_i were set to be 1. The dispersion parameter ϕ was set to be 5.

Each simulation was repeated 10000 times. We tested for the hypothesis of $\beta_1 = 0$. Empirical power and false positive rate were summarized at the significance level of 0.05. We compared zero-inflated Beta regression mixed model, denoted by ZIBR, and the proposed ZIGMMs with the arcsine square root transformation for proportion data, $\arcsin(\sqrt{y_{ij}})$, denoted by ZIGMMs(arcsine), the transformed data was standardized by its standard deviation before model fitting.

3.1.1.2 Simulation results. Table 1 shows the comparison of empirical power and false positive rates between ZIGMMs and ZIBR in analyzing the longitudinal microbiome proportion data. ZIGMMs and ZIBR controlled the false positive rates similarly close to the significance level under all three different sample sizes. Although the proportion data were simulated under the zero-inflated beta distribution, ZIGMMs lead to a higher empirical power to detect the group effect than ZIBR.

3.1.2 Assess the ZIGMMs in analyzing microbiome count data. 3.1.2.1 Simulation design. We then assessed the ZIGMMs in analyzing microbiome count data. We employed the function sim in **NBZIMM** to simulate zero-inflated longitudinal microbiome count data c_{ij} as follows. We used the latent-data formulation of the logistic regression to simulate zero-state indicators; the logistic model, $p(\xi_{ij} = 1) = \text{logit}^{-1}(\mu+Z_{ij}\alpha+G_{ij}a_i)$, is approximately equivalent to the model, $u_{ij} \sim N(Z_{ij}\alpha+G_{ij}a_i, 1.6^2)$, $u_{ij} > h \Leftrightarrow \xi_{ij} = 1$ [30], where h is a constant determined by the preset overall zero-inflation proportion p. Thus, we first simulated latent normal variables u_{ij} and then set samples with the $100p\%$ largest u_{ij} as from zero state. This method can easily control the overall zero-inflation proportion and also allow for the sample-specific zero-inflation

Table 1. False positive rate and power for testing H0: $\beta_1 = 0$ based on ZIGMMs and ZIBR for significance level at 0.05 for various sample sizes.

	False Positive Rate		Power (Low Effect Setting)		Power (High Effect Setting)	
Sample Size	**ZIGMMs (arcsine)†**	**ZIBR‡**	**ZIGMMs (arcsine)**	**ZIBR**	**ZIGMMs (arcsine)**	**ZIBR**
n = 50	0.0681	0.0577	0.1937	0.1438	0.4100	0.3022
n = 100	0.0554	0.0578	0.3025	0.2218	0.6592	0.5135
n = 150	0.0563	0.0533	0.4308	0.3031	0.8296	0.6906

ZIBR‡: Zero-inflated beta mixed model.

ZIGMMs(arcsine)†: Zero-inflated Gaussian mixed models with arcsine transformation.

https://doi.org/10.1371/journal.pone.0242073.t001

probabilities p_{ij}. For the samples from nonzero state, we simulated counts c_{ij} from the negative binomial distribution $NB(c_{ij}|\mu_{ij},\theta)$, where $\mu_{ij} = \log(T_{ij})+X_{ij}\beta+G_{ij}b_i$.

We adopted a longitudinal design and utilized four different simulation settings. In all the settings, we generated subjects from two groups (i.e. case or control) and simulated samples at multiple time points for each subject. We considered three numbers of subjects: n = 50, 100 and 150, half of which were designated to be cases. Each subject was measured at 5 time points. The random effects, and within-subject correlation structures were set as follows:

1. Setting A: a group variable (β_1) is included as fixed effect in the count part, no fixed effect in the zero-inflation part (i.e. Z_{ij} = 1), random intercepts in both count and zero-inflation parts (i.e. G_{ij} = 1)), and no within-subject correlation;
2. Setting B: a group variable is included as fixed effects in both parts, random intercept in the count part only, and no within-subject correlation;
3. Setting C: a group variable is included as fixed effects in both parts, random intercepts in both parts (i.e. G_{ij} = 1)), and no within-subject correlation;
4. Setting D: a group variable is included as fixed effect in the count part, no fixed effect in the zero-inflation part, random intercept in the count part only, and the within-subject correlation was autoregressive of order 1, AR(1), in the count part;
5. Setting E: a group variable (β_1), a time variable (β_2), and a time by main effect interaction term (β_3) are included as fixed effects in both parts, random intercept in the count part only, and no within-subject correlation;

We randomly generated the parameters in the models from reasonable ranges. The parameters to simulate the counts from negative binomial distribution were set by following the work of [31]. This can largely reduce the combinations of parameter values and minimize possible bias from setting inappropriate values for parameters. The ranges were described as follows:

1. To simulate counts similar to real microbiome data, we controlled the means of simulated counts through $\log(T_{ij}) + \beta_0$, where β_0 is the fixed intercept. We set β_0 = -7 and randomly sampled $\log(T_{ij})$ from the range [7.1, 10.5];
2. For settings A-D, the dispersion parameter θ was uniformly sampled from the range [0.1, 5], which yielded highly or moderate over-dispersed counts; for setting E, the dispersion parameter θ was set to be 5.
3. To evaluate false positive rates, the fixed effects β_1 was set to be zero. To evaluate empirical powers, we considered two scenarios: a) low effect scenario: β_1 was sampled from [0.2, 0.3]; b) high effect scenario: β_1 was sampled from [0.3, 0.4]; fixed effects in the zero-inflation part were considered in setting B and C, where α_1 was set to be the same as β_1; for setting E, β_1 was set to be equal to β_3. And β_2 was set to be 0 in all scenarios.
4. The random effects b_i and a_i were generated from N(0, τ^2), for settings A-D, where τ was randomly drawn from the range [0.5, 1]; for setting E, τ was set to be 0.5.
5. For settings A-D, the overall zero-inflation proportion was set to be chosen from three levels, that is [0, 0.2], [0.2, 0.4] and [0.4, 0.6]; for setting E, the proportion was set to be chosen from [0, 0.5].
6. The correlation coefficient ρ and the standard deviation σ for AR(1) correlation were both sampled from [0.1, 0.5], and the AR(1) correlation was generated by the function arima.sim from R package **stats**;

Table 2. Parameter ranges in simulation studies.

Parameter	Range
$\log(T_{ij}) + \beta_0$	Unif(0.1, 3.5)
dispersion parameter θ	Unif(0.1, 5)
Fixed effects β_1 (false positive rate)	0
Fixed effects β_1 (power)	Unif(0.2, 0.3)
	Unif(0.3, 0.4)
Fixed effect α_1 (Setting B and C only)	Unif(0.2, 0.3)
	Unif(0.3, 0.4)
standard deviation τ	Unif(0.5, 1)
correlation ρ	Unif(0.1, 0.5)
standard deviation σ	Unif(0.1, 0.5)
Overall zero-inflation proportion	Unif(0.0, 0.2)
	Unif(0.2, 0.4)
	Unif(0.4, 0.6)

https://doi.org/10.1371/journal.pone.0242073.t002

The ranges of all the parameters used in the simulation are summarized in Table 2.

We repeated the procedure 10000 times for each combination of the parameters. The hypothesis of interest is the fixed effect H_0: $\beta_1 = 0$. Empirical power and false positive rate for testing the hypothesis were calculated at the significance level of 0.05. We compared the proposed ZIGMMs, denoted by ZIGMMs(log), with a previously developed negative binomial mixed model, denoted by NBMMs, and the linear mixed model with the arcsine square root transformed response, $\arcsin(\sqrt{y_{ij}/T_{ij}})$, denoted by LMMs.

3.1.2.2 Simulation results. Fig 1 showed empirical power to detect the group effect for settings A, B, C and D at the low effect scenario. It can be clearly seen that the proposed method performed consistently better than NBMMs and LMMs in all the scenarios. Under setting B and C, we simulated fixed effects in the zero-inflation part. ZIGMMs performed extremely remarkable than NBMMs and LMMs in those two settings, inferring ignoring the association between zero-inflation and any covariate could lead to a significant decrease in power. The power was largely affected by the sample size and the zero-inflation probability. The difference in power among ZIGMMs and NBMMs and LMMs increased significantly as the zero-inflation probability increased. With the zero-inflation proportion less than 20%, ZIGMMs performed similarly as NBMMs but still better than LMMs. ZIGMMs had a more noteworthy higher power than NBMMs and LMMs to detect the fixed effect especially when the data was highly zero-inflated. We also summarized the empirical power to detect the binary group effect for the settings A, B, C and D with the high effect scenario in S1 Fig. In the high effect scenario, ZIGMMs outperformed NBMMs and LMMs more significantly when the zero-inflation probability was higher and the sample size was smaller. Fig 2 displays false positive rates for detecting the group effect. For all the four settings, ZIGMMs controlled the false positive rates close to the significance level under all the combinations of parameters. As expected, the increase in sample size n led to the decrease in false positive rates in ZIGMMs.

Table 3 summarized empirical power and false positive rates for setting E comparing LMMs, NBMMs and ZIGMMs. In this setting, we included group variable, time variable and a time by group interaction term in the simulation and reported empirical power and false positive rates for group variable and time by group interaction term. ZIGMMs had a higher power than LMMs and NBMMs for both group effect and interaction term under various sample

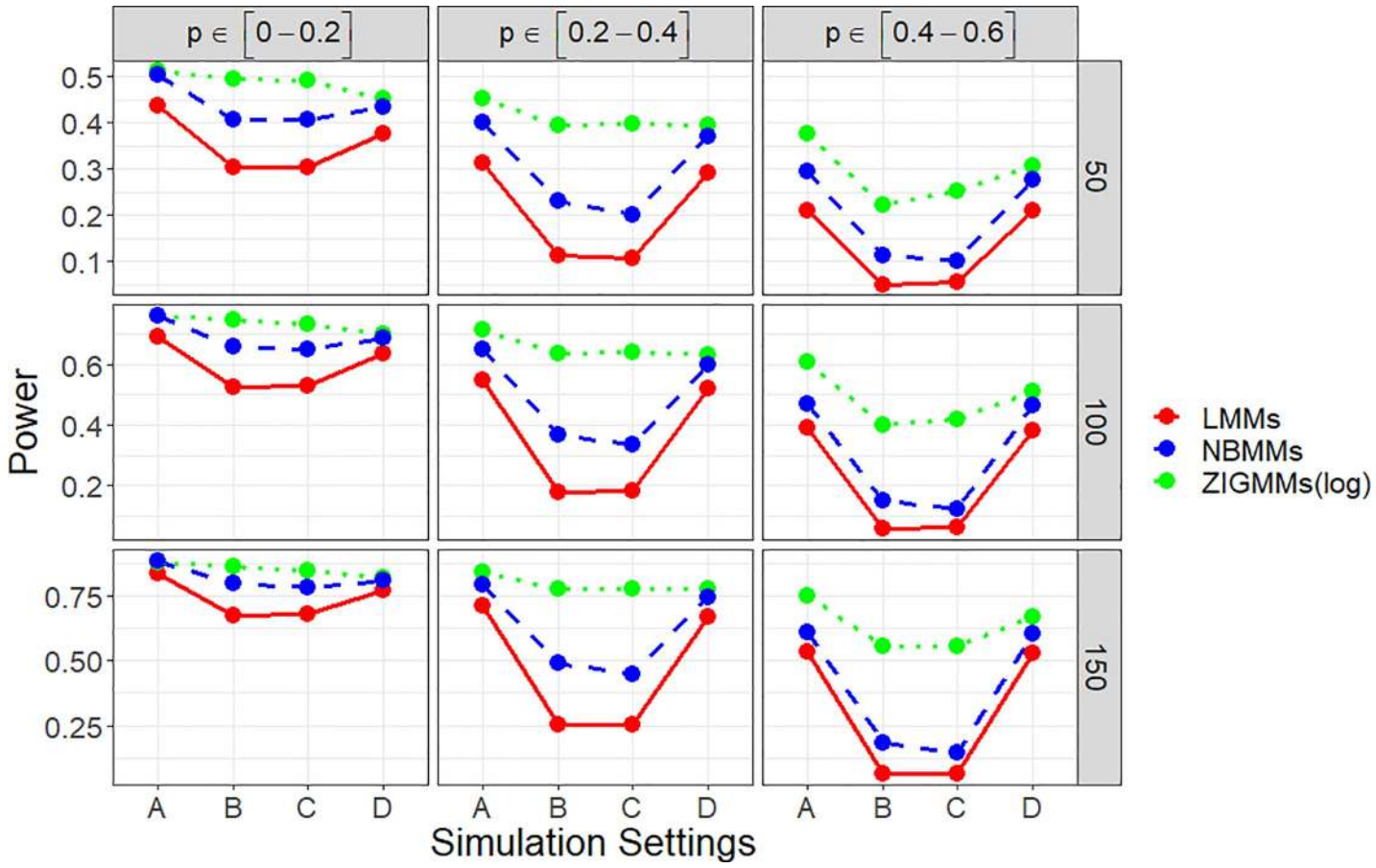

Fig 1. Empirical powers in four simulation settings under low effect scenario.

https://doi.org/10.1371/journal.pone.0242073.g001

sizes however ZIGMMs had inflated the false positive rates compared to LMMs and NBMMs especially for the interaction term.

3.1.3 Assess the computational efficiency of ZIGMMs. To evaluate the computational efficiency of ZIGMMs, we recorded the computation time for ZIGMMs and two other zero-inflated methods in one simulation when sample size is set to be 100. First, we compared ZIGMMs and ZIBR in analyzing the longitudinal microbiome proportion data. We found that the computation time for ZIGMMs and ZIBR in one simulation was 0.011 and 0.023 minutes, respectively. Besides, we compared ZIGMMs and a zero-inflated negative binomial mixed model which was implemented in the R package **glmmTMB** in analyzing the longitudinal microbiome count data, and found that the computation time for ZIGMMs and the zero-inflated negative binomial mixed model in one simulation was 0.009 and 0.041 minutes, respectively. ZIGMMs remarkably outperformed in computational efficiency than the other two zero-inflated methods.

3.2 Application to 16S rRNA and shotgun sequencing microbiome data

In our real data analysis, there are two major purposes, one is to evaluate the performances of ZIGMMs in analyzing 16S rRNA data in raw counts, the other is to evaluate the performances of ZIGMMs in analyzing shotgun sequencing data in proportions. So that, we applied our ZIGMMs in two publicly available datasets from Romero, Hassan [24] and Vincent, Miller [32].

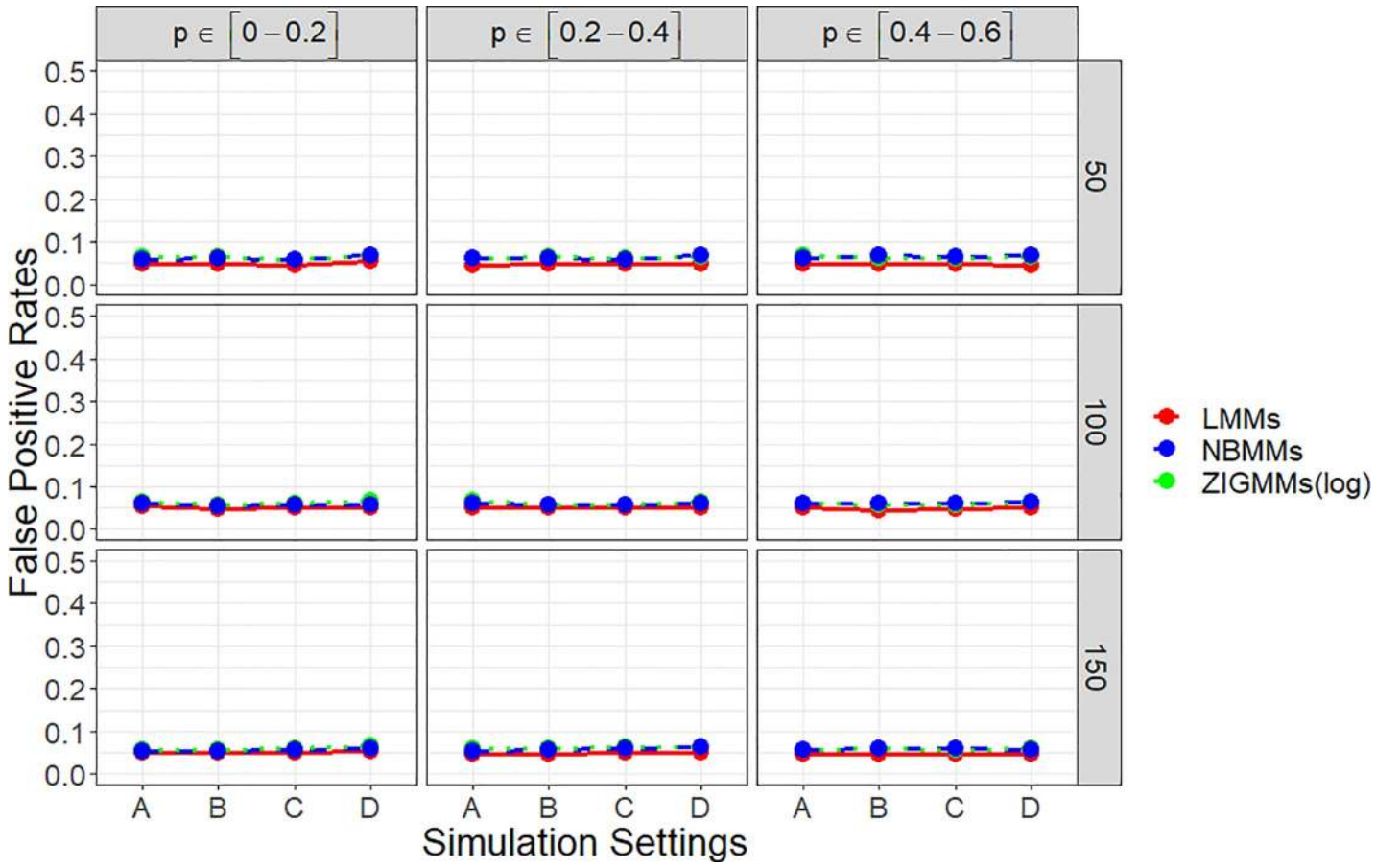

Fig 2. False positive rates in all four simulation settings.

https://doi.org/10.1371/journal.pone.0242073.g002

Romero, Hassan [24] employed a retrospective case-control longitudinal study to investigate the difference of composition and stability of vaginal microbiota between pregnant and non-pregnant women. They conducted a 16S rRNA gene sequence-based survey among 22 normal pregnant women who delivered at term (38–40 weeks) and 32 non-pregnant women. Vaginal fluid samples were collected every two to four weeks apart for the pregnant group and twice per week for 16 weeks in the non-pregnant group. We analyzed the 16S rRNA sequencing data from Romero, Hassan [24] in terms of counts to evaluate the performances of ZIGMMs(log).

Vincent, Miller [32] used metagenome shotgun sequencing to examine the diversity and composition of the fecal microbiota from 98 hospitalized patients. The prospective cohort study was carried out among 8 patients who were either *Clostridium difficile* infected or colonized and other 90 patients. Clinical data included gender, age, and days from first collection of the fecal samples. The clinical data and shotgun sequencing microbiome relative abundance data were downloaded by R package **curatedMetagenomicData** [33]. The shotgun sequencing data is normally output as proportion data. So, here, we illustrated our ZIGMMs(arcsine) to analyze this shogun sequencing microbiome data from Vincent, Miller [32] in proportions. According to the manual of R package **ZIBR** [26], ZIBR cannot handle missing data. Therefore, we could not compare with ZIBR in our real data example.

We used the following eight different models to compare the performances of LMMs, NBMMs, and ZIGMMs in detecting the dynamic association between host factor and

Table 3. False positive rate and power for testing H0: $\beta_1 = 0$ and H0: $\beta_3 = 0$ from setting E for significance level at 0.05 for various sample sizes.

	False Positive Rate					
	Test of β_1			**Test of β_3**		
Sample Size	LMMs[§]	NBMMs[¶]	ZIGMMs(log)[!]	LMMs[§]	NBMMs[¶]	ZIGMMs(log)[!]
n = 50	0.045	0.053	0.065	0.045	0.064	0.084
n = 100	0.050	0.061	0.067	0.054	0.072	0.082
n = 150	0.047	0.061	0.071	0.050	0.068	0.082
	Power (Low Effect Setting)					
	Test of β_1			**Test of β_3**		
Sample Size	LMMs[§]	NBMMs[¶]	ZIGMMs(log)[!]	LMMs[§]	NBMMs[¶]	ZIGMMs(log)[!]
n = 50	0.082	0.158	0.187	0.172	0.251	0.334
n = 100	0.148	0.265	0.325	0.295	0.425	0.563
n = 150	0.204	0.360	0.439	0.405	0.562	0.720
	Power (High Effect Setting)					
	Test of β_1			**Test of β_3**		
Sample Size	LMMs[§]	NBMMs[¶]	ZIGMMs(log)[!]	LMMs[§]	NBMMs[¶]	ZIGMMs(log)[!]
n = 50	0.121	0.252	0.304	0.303	0.418	0.558
n = 100	0.224	0.439	0.522	0.507	0.654	0.815
n = 150	0.340	0.602	0.699	0.628	0.769	0.920

LMMs[§]: Linear mixed models.

NBMMs[¶]: Negative Binomial mixed models.

ZIGMMs(log)[!]: Zero-inflated Gaussian mixed models with log transformation.

https://doi.org/10.1371/journal.pone.0242073.t003

microbiota composition. Models A-D were used in all LMMs, NBMMs and ZIGMMs while models E-G were only used in ZIGMMs:

1. Model A: host factor and time as fixed effects in Gaussian part, random intercept in Gaussian part;
2. Model B: host factor, time, host factor and time interaction term as fixed effects in Gaussian part, random intercept in Gaussian part;
3. Model C: host factor, time, host factor and time interaction term as fixed effects in Gaussian part, random intercept and the within-subject correlation was autoregressive of order 1, AR (1) in Gaussian part;
4. Model D: host factor, time, host factor and time interaction term as fixed effects in Gaussian part, two random effects (i.e., random intercept and time effect) in Gaussian part;
5. Model E: host factor and time as fixed effects only in both zero-inflation part and Gaussian part, random intercept in Gaussian part;
6. Model F: host factor, time, host factor and time interaction term as fixed effects in both zero-inflation part and Gaussian part, random intercept in Gaussian part;
7. Model G: host factor, time, host factor and time interaction term as fixed effects in both zero inflation part and Gaussian part, random intercept and the within-subject correlation was autoregressive of order 1, AR(1) in Gaussian part;
8. Model H: host factor, time, host factor and time interaction term as fixed effects in both zero-inflation part and Gaussian part, two random effects (i.e., random intercept and time effect) in Gaussian part;

The real data and the R code for our analysis are available from the GitHub page: https://abbyyan3.github.io//NBZIMM-tutorial/ZIGMMs-longitudinal.html.

3.2.1 Application in 16S rRNA longitudinal pregnancy data. We first applied our ZIGMMs to the data of Romero, Hassan [24]. We explored the abilities of ZIGMMs in detecting the dynamic associations between vaginal bacteria taxa composition and two groups (pregnancy vs non-pregnancy) controlled by possible confounding effects of the covariates. We analyzed 16S rRNA sequencing microbiome count data with log transformation (ZIGMMs (log)). In all the eight models, the binary case-control indicator for pregnancy vs non-pregnancy was the host factor of interest (β_1), and the collection time (GA_days) was the time variable. An interaction term between host factor and time variable (β_3) was included in model B, C, D, F, G and H. We also included age and race as confounding covariates. The sample size was 897 in the final analysis. We included 59 taxa which has a proportion of zeros greater than 0.3 but smaller than 0.9 in our analysis.

Table 4 shows the proportions of significant taxa detected by LMMs, NBMMs and ZIGMMs(log) at the alpha level at 0.05, respectively. The significance of the taxa was evaluated at the alpha level of 0.05 (p-value <0.05) for Models A-H. Test of β_1 in Table 4 summarized the proportions of taxa which is significantly differentiated presented between pregnancy group vs non-pregnancy group. Test of β_3 in Table 4 summarized the proportions of taxa which is significantly differentiated presented between pregnancy group vs non-pregnancy group over the collection time. The proportions of detected significant taxa in model B, C, D, F, G and H were substantially less than the rates from models A and E. It inferred that the majority of taxa existing in the vaginal microbiome did not possess a time-dependent association between the pregnant and non-pregnant groups. Moreover, it showed that ZIGMMs(log) detected more associated taxa than NBMMs and LMMs. We also found ZIGMMs with fixed effects in zero-inflation and Gaussian part in models E-H decrease slightly in the number of significant taxa detected than ZIGMMs with fixed effects in Gaussian part from models A-D. It implied that those taxa did not possess a strong association between the host factors and the zero-inflation.

To compare the differences in detecting significant taxa for both host factor and interaction term between LMMs, NBMMs, and ZIGMMs(log), we presented model C in Fig 3 and S2 Fig. Fig 3 shows significant taxa in model C at the 5% significance threshold and minus log transformed p-values for LMMs, NBMMs, and ZIGMMs(log). S2 Fig presents three heatmaps of p-values between the taxa and each variable from model C using LMMs, NBMMs, and ZIGMMs (log). We found that ZIGMMs(log) discovered more taxa than NBMMs and LMMs consistently, and yielded smaller p-values. In model C, we were interested in both the host factor and the interaction effect between time and host factor. ZIGMMs(log) identified not only the same taxa which were detected by LMMs and NBMMs but also more taxa for both effects. For the host factor, several taxa were only identified with ZIGMMs(log), including *Clostridiales*, *Streptococcus*, *Proteobacteria*, *BVAB1* and *Lactobacillales*. For the interaction effect between time and host factor, *Prevotella genogroup 3*, *Gemella*, *Lactobacillus gasseri*, *Megasphaera sp type 1* and *Firmicutes* were identified both by NBMMs and ZIGMMs(log). *BVAB1*, and *Sneathia Sanguinegens* were only identified by ZIGMMs(log). Among them, *bacterial vaginosis associated bacteria 1 (BVAB1)* has been previously reported as a highly specific novel bacteria for bacterial vaginosis in the *Clostridiales* order [34]. Also, the abundance of *Gemella*, *BVAB1*, and *Sneathia sanguinegens* have been reported to change within the duration of pregnancy from another study by Romero, Hassan [35].

3.2.2 Application in shotgun sequencing longitudinal intestinal microbiome data. We then applied our ZIGMMs to the shotgun sequencing microbiome proportion data from Vincent, Miller [32]. In this case, we only compared our ZIGMMs with LMMs. We explored the

Table 4. Proportions of significant taxa detected in four models with LMMs, NBMMs and ZIGMMs.

	Model A	Model B		Model C		Model D	
	Test of β_1	Test of β_1	Test of β_3	Test of β_1	Test of β_3	Test of β_1	Test of β_3
LMMs[§]	0.29	0.03	0.15	0.03	0.12	0.07	0.10
NBMMs[¶]	0.49	0.12	0.25	0.12	0.25	0.12	0.25
ZIGMMs(log)[!]	0.63	0.34	0.24	0.39	0.27	0.36	0.24
	Model E	Model F		Model G		Model H	
	Test of β_1	Test of β_1	Test of β_3	Test of β_1	Test of β_3	Test of β_1	Test of β_3
ZIGMMs(log)	0.54	0.19	0.31	0.20	0.24	0.20	0.20

LMMs[§]: Linear mixed models.

NBMMs[¶]: Negative Binomial mixed models.

ZIGMMs(log)[!]: Zero-inflated Gaussian mixed models with log transformation.

https://doi.org/10.1371/journal.pone.0242073.t004

abilities of ZIGMMs in detecting the dynamic associations between fecal microbiome composition and *Clostridium difficile* colonization or infection. We adapted ZIGMMs in analyzing microbiome proportion data with arcsine transformation (ZIGMMs(arcsine)). In all the eight models, the binary case-control indicator for *Clostridium difficile* colonization or infection vs

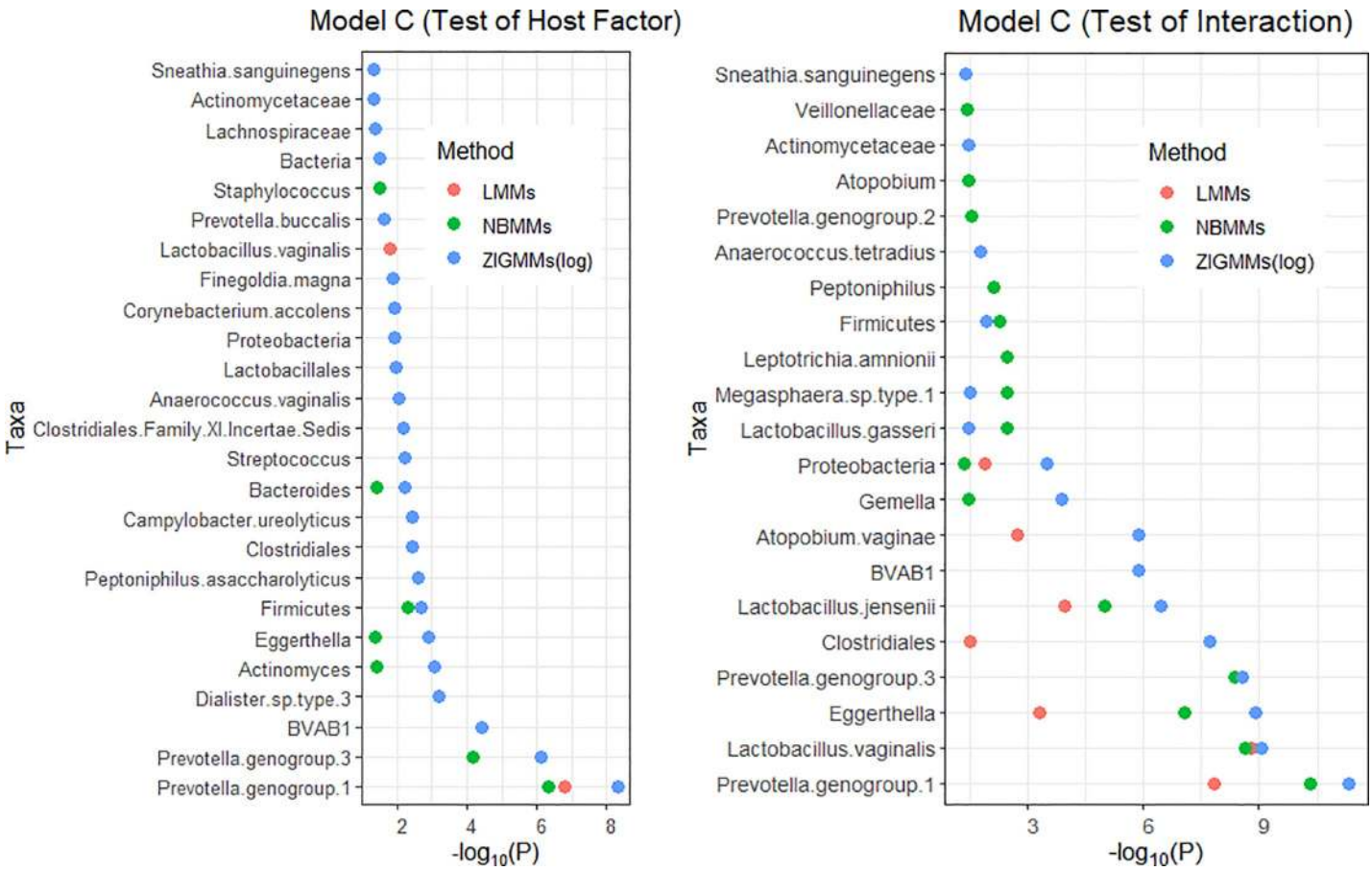

Fig 3. The analyses of ZIGMMs(log), NBMMs and LMMs: minus log transformed p-values for the significant differentially abundant taxa at the 5% significance threshold between pregnancy and non-pregnancy groups for host factor effect (left panel) and interaction effect (right panel) from Model C.

https://doi.org/10.1371/journal.pone.0242073.g003

Table 5. Proportions of significant taxa detected in four models with LMMs and ZIGMMs.

	Model A	Model B		Model C		Model D	
	Test of β_1	Test of β_1	Test of β_3	Test of β_1	Test of β_3	Test of β_1	Test of β_3
LMMs[$]	0.11	0.13	0.12	0.11	0.11	0.10	0.06
ZIGMMs (arcsine)[†]	0.12	0.12	0.19	0.17	0.18	0.11	0.10
	Model E	Model F		Model G		Model H	
	Test of β_1	Test of β_1	Test of β_3	Test of β_1	Test of β_3	Test of β_1	Test of β_3
ZIGMMs (arcsine)	0.15	0.14	0.21	0.14	0.23	0.14	0.10

ZIGMMs(arcsine)[†]: Zero-inflated Gaussian mixed models with arcsine transformation.

LMMs[$]: Linear mixed models.

https://doi.org/10.1371/journal.pone.0242073.t005

control was the host factor of interest (β_1), and the collection time (days from the first collection) was the time variable. An interaction term between host factor and time variable (β_3) was included in models B, C, D, F, G and H. We also included age and gender as confounding covariates. The sample size was 229 in the final analysis. We included 357 taxa which has a proportion of zeros greater than 0.3 but smaller than 0.9 in our analysis.

Table 5 shows the proportions of significant taxa detected by LMMs and ZIGMMs(arcsine) at the alpha level at 0.05, respectively. The significance of the taxa was evaluated at the alpha level of 0.05 (p-value <0.05) for Models A-H. Test of β_1 in Table 5 summarized the proportions of taxa which is significantly differentiated presented between *Clostridium difficile* colonization or infection group vs control group. Test of β_3 in Table 5 summarized the proportions of taxa which is significantly differentiated presented between *Clostridium difficile* colonization or infection group vs control group over the collection time. We found that our ZIGMMs(arcsine) detected more associated taxa than LMMs in most scenarios. We also found ZIGMMs (arcsine) with fixed effects in zero-inflation and Gaussian part in models E-H increase slightly in the number of significant taxa detected than ZIGMMs(arcsine) with fixed effects in Gaussian part from models A-D. It implied that there is a significant association between the host factors and the zero-inflation in those taxa.

4. Discussion

With the emergence of longitudinal microbiome studies, more understandings about the dynamic shifts of the microbiota have been unraveled [8]. It is of interest in studying the dynamic associations between the microbiota and various host factors [8,36]. To realize these research interests, powerful analytic methods are necessary to account for sources of heterogeneity and dependence in microbiome measurements. However, previous methods have not fully addressed the properties of longitudinal microbiome data and are not computationally feasible for analyzing many taxa.

Here, we propose ZIGMMs to model longitudinal microbiome proportion and count data. The method is robust in performance when applied to both 16S rRNA gene sequencing and genome shotgun sequencing data, in terms of proportion or count data. The proportions data, mostly from genome shotgun sequencing data, should be transformed with arcsine square root transformation. For count data, mostly from 16S rRNA platforms, log transformation is more appropriate because if converting those count data to proportion data will lead to very small proportions. The proposed ZIGMMs can effectively handle excessive zeros observed in microbiome data, and can incorporate various types of random effects and within-subject correlation structures [29,37]. We have developed an EM algorithm to fit the proposed ZIGMMs by extending a commonly used procedure for fitting LMMs [37–40]. This allows us to

integrate the well-established procedures for analyzing longitudinal data into our ZIGMMs. Our analyses show that our algorithm is efficient and stable for most of the scenarios. We showed the computational efficiency of our EM algorithm by comparing with the other two zero-inflated methods. In the simulations, ZIGMMs outperform LMMs, NBMMs and ZIBR consistently. We have also shown that ZIGMMs can efficiently deal with various fixed and random effects in both normal distribution and zero-inflation models, moreover, and account for the auto-regressive correlation among samples. However, we found ZIGMMs had inflated false positive rates especially in detecting interaction terms, suggesting potential fitting issues. According to Weiss, Xu [41] and Hawinkel, Mattiello [42], most of the parametric methods, such as edgeR, limma–voom and metagenomeSeq, fail to control the false positive rate at the nominal level. A possible reason could be the p-value distributions tend to be smaller than uniform distribution especially when taxa is highly inflated [42]. Thus, in current analysis of a real microbial data, researchers normally focus on the top abundant taxa with less zero-inflation rates.

Moreover, we applied our method to two previously published datasets and compared the performances of LMMs, NBMMs and ZIGMMs in detecting the dynamic association between host factor and taxa composition. We could not apply the ZIBR in the real data since according to the manual of R package ZIBR, it could only deal with subjects measured at the same number of time points [26]. We found that our ZIGMMs was capable to detect more significant taxa than LMMs and NBMMs. The differences between our ZIGMMs and the other two methods were more substantial when analyzing the taxa with high zero rates. Notably, we found that several taxa from Romero, Hassan [24], which have only been identified by ZIGMMs, have been previously reported for the associations between pregnancy and vaginal bacterial composition by Romero, Hassan [35]. However, we still encounter the fitting issues similarly as other parametric methods to control false positive rates under nominal level, especially when analyzing complex microbiome/metagenomics data. A future plan is to develop analyzing methods under Bayesian framework using MCMC algorithm to possibly address the current fitting issues.

Supporting information

S1 Fig. Empirical power of hypothesis in four simulation settings under high effect scenario.
(PDF)

S2 Fig. Heat map for p-values between the taxa and each variable from Model C using LMMs (left panel), NBMMs (middle panel) and ZIGMMs (right panel). The sign "+" indicates the positive effect.
(PDF)

Acknowledgments

We thank two reviewers and the associate editor for their constructive suggestions and comments that have improved the manuscript.

Author Contributions

Conceptualization: Nengjun Yi.

Formal analysis: Xinyan Zhang.

Methodology: Nengjun Yi.

Software: Nengjun Yi.

Writing – original draft: Xinyan Zhang.

Writing – review & editing: Xinyan Zhang, Boyi Guo, Nengjun Yi.

References

1. Yang I., et al., The Infant Microbiome: Implications for Infant Health and Neurocognitive Development. Nurs Res, 2016. 65(1): p. 76–88. https://doi.org/10.1097/NNR.0000000000000133 PMID: 26657483
2. Cho I. and Blaser M.J., The human microbiome: at the interface of health and disease. Nat Rev Genet, 2012. 13(4): p. 260–70. https://doi.org/10.1038/nrg3182 PMID: 22411464
3. Plottel C.S. and Blaser M.J., Microbiome and malignancy. Cell Host Microbe, 2011. 10(4): p. 324–35. https://doi.org/10.1016/j.chom.2011.10.003 PMID: 22018233
4. Pflughoeft K.J. and Versalovic J., Human microbiome in health and disease. Annu Rev Pathol, 2012. 7: p. 99–122. https://doi.org/10.1146/annurev-pathol-011811-132421 PMID: 21910623
5. Honda K. and Littman D.R., The microbiome in infectious disease and inflammation. Annu Rev Immunol, 2012. 30: p. 759–95. https://doi.org/10.1146/annurev-immunol-020711-074937 PMID: 22224764
6. Holmes E., et al., Understanding the role of gut microbiome-host metabolic signal disruption in health and disease. Trends Microbiol, 2011. 19(7): p. 349–59. https://doi.org/10.1016/j.tim.2011.05.006 PMID: 21684749
7. Kinross J.M., Darzi A.W., and Nicholson J.K., Gut microbiome-host interactions in health and disease. Genome Med, 2011. 3(3): p. 14. https://doi.org/10.1186/gm228 PMID: 21392406
8. Gerber G.K., The dynamic microbiome. FEBS Lett, 2014. 588(22): p. 4131–9. https://doi.org/10.1016/j.febslet.2014.02.037 PMID: 24583074
9. McGeachie M.J., et al., Longitudinal Prediction of the Infant Gut Microbiome with Dynamic Bayesian Networks. Sci Rep, 2016. 6: p. 20359. https://doi.org/10.1038/srep20359 PMID: 26853461
10. Gerber G.K., Longitudinal Microbiome Data Analysis, in *Metagenomics for Microbiology*. 2015, Elsevier. p. 97–111.
11. Ward D.V., et al., Metagenomic Sequencing with Strain-Level Resolution Implicates Uropathogenic E. coli in Necrotizing Enterocolitis and Mortality in Preterm Infants. Cell Rep, 2016. 14(12): p. 2912–24. https://doi.org/10.1016/j.celrep.2016.03.015 PMID: 26997279
12. La Rosa P.S., et al., Patterned progression of bacterial populations in the premature infant gut. Proc Natl Acad Sci U S A, 2014. 111(34): p. 12522–7. https://doi.org/10.1073/pnas.1409497111 PMID: 25114261
13. Zhou Y., et al., Longitudinal analysis of the premature infant intestinal microbiome prior to necrotizing enterocolitis: a case-control study. PLoS One, 2015. 10(3): p. e0118632. https://doi.org/10.1371/journal.pone.0118632 PMID: 25741698
14. DiGiulio D.B., et al., Temporal and spatial variation of the human microbiota during pregnancy. Proc Natl Acad Sci U S A, 2015. 112(35): p. 11060–5. https://doi.org/10.1073/pnas.1502875112 PMID: 26283357
15. Morris A., et al., Longitudinal analysis of the lung microbiota of cynomolgous macaques during long-term SHIV infection. Microbiome, 2016. 4(1): p. 38. https://doi.org/10.1186/s40168-016-0183-0 PMID: 27391224
16. Jovel J., et al., Characterization of the Gut Microbiome Using 16S or Shotgun Metagenomics. Frontiers in Microbiology, 2016. 7. https://doi.org/10.3389/fmicb.2016.00459 PMID: 27148170
17. Paulson J.N., et al., Differential abundance analysis for microbial marker-gene surveys. Nat Methods, 2013. 10(12): p. 1200–2. https://doi.org/10.1038/nmeth.2658 PMID: 24076764
18. Leamy L.J., et al., Host genetics and diet, but not immunoglobulin A expression, converge to shape compositional features of the gut microbiome in an advanced intercross population of mice. Genome Biol, 2014. 15(12): p. 552. https://doi.org/10.1186/s13059-014-0552-6 PMID: 25516416
19. Benson A.K., et al., Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors. Proc Natl Acad Sci U S A, 2010. 107(44): p. 18933–8. https://doi.org/10.1073/pnas.1007028107 PMID: 20937875
20. Srinivas G., et al., Genome-wide mapping of gene-microbiota interactions in susceptibility to autoimmune skin blistering. Nat Commun, 2013. 4: p. 2462. https://doi.org/10.1038/ncomms3462 PMID: 24042968

21. Wang J., et al., Analysis of intestinal microbiota in hybrid house mice reveals evolutionary divergence in a vertebrate hologenome. Nat Commun, 2015. 6: p. 6440. https://doi.org/10.1038/ncomms7440 PMID: 25737238

22. Zhang X., et al., Negative Binomial Mixed Models for Analyzing Microbiome Count Data. BMC Bioinformatics, 2017. 18: p. 4. https://doi.org/10.1186/s12859-016-1441-7 PMID: 28049409

23. Zhang X., et al., Negative Binomial Mixed Models for Analyzing Longitudinal Microbiome Data. Frontiers in Microbiology 2018. https://doi.org/10.3389/fmicb.2018.01683 PMID: 30093893

24. Romero R., et al., The composition and stability of the vaginal microbiota of normal pregnant women is different from that of non-pregnant women. Microbiome, 2014. 2(1): p. 4. https://doi.org/10.1186/2049-2618-2-4 PMID: 24484853

25. Chen E.Z. and Li H., A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. Bioinformatics, 2016. 32(17): p. 2611–7. https://doi.org/10.1093/bioinformatics/btw308 PMID: 27187200

26. Chen E.Z. and Li H. *ZIBR (Zero-Inflated Beta Random Effect model)*. 2019; Available from: https://github.com/chvlyl/ZIBR.

27. Spor A., Koren O., and Ley R., Unravelling the effects of the environment and host genotype on the gut microbiome. Nat Rev Microbiol, 2011. 9(4): p. 279–90. https://doi.org/10.1038/nrmicro2540 PMID: 21407244

28. Faust K., et al., Metagenomics meets time series analysis: unraveling microbial community dynamics. Curr Opin Microbiol, 2015. 25: p. 56–66. https://doi.org/10.1016/j.mib.2015.04.004 PMID: 26005845

29. Pinheiro J.C. and Bates D.C., *Mixed-Effects Models in S and S-PLUS*. 2000: Springer Verlag New York.

30. Gelman A. and Hill J., *Data Analysis Using Regression and Multilevel/Hierarchical Models*. 2007, New York: Cambridge University Press.

31. Sohn M.B., Du R., and An L., A robust approach for identifying differentially abundant features in metagenomic samples. Bioinformatics, 2015. 31(14): p. 2269–75. https://doi.org/10.1093/bioinformatics/btv165 PMID: 25792553

32. Vincent C., et al., Bloom and bust: intestinal microbiota dynamics in response to hospital exposures and Clostridium difficile colonization or infection. Microbiome, 2016. 4: p. 12. https://doi.org/10.1186/s40168-016-0156-3 PMID: 26975510

33. Pasolli E., et al., Accessible, curated metagenomic data through ExperimentHub. Nat Methods, 2017. 14(11): p. 1023–1024. https://doi.org/10.1038/nmeth.4468 PMID: 29088129

34. Srinivasan S., et al., Bacterial communities in women with bacterial vaginosis: high resolution phylogenetic analyses reveal relationships of microbiota to clinical criteria. PLoS One, 2012. 7(6): p. e37818. https://doi.org/10.1371/journal.pone.0037818 PMID: 22719852

35. Romero R., et al., The vaginal microbiota of pregnant women who subsequently have spontaneous preterm labor and delivery and those with a normal delivery at term. Microbiome, 2014b. 2: p. 18. https://doi.org/10.1186/2049-2618-2-18 PMID: 24987521

36. Biagi E., et al., Through ageing, and beyond: gut microbiota and inflammatory status in seniors and centenarians. PLoS One, 2010. 5(5): p. e10667. https://doi.org/10.1371/journal.pone.0010667 PMID: 20498852

37. McCulloch C.E. and Searle S.R., *Generalized, Linear, and Mixed Models*. 2001: John Wiley & Sons, Inc.

38. Schall R., Estimation in generalized linear models with random effects. Biometrika, 1991(78): p. 719–727.

39. Breslow N.E. and Clayton D.C., Approximate inference in generalized linear mixed models. Journal of American Statistical Association, 1993(88): p. 9–25.

40. Venables W.N. and Ripley B.D., *Modern Applied Statistics with S*. 2002: Springer-Verlag New York.

41. Weiss S., et al., Normalization and microbial differential abundance strategies depend upon data characteristics. Microbiome, 2017. 5(1): p. 27. https://doi.org/10.1186/s40168-017-0237-y PMID: 28253908

42. Hawinkel S., et al., A broken promise: microbiome differential abundance methods do not control the false discovery rate. Brief Bioinform, 2019. 20(1): p. 210–221. https://doi.org/10.1093/bib/bbx104 PMID: 28968702