frontiers in Microbiology

METHODS
published: 26 July 2018
doi: 10.3389/fmicb.2018.01683

Negative Binomial Mixed Models for Analyzing Longitudinal Microbiome Data

Xinyan Zhang[1†], Yu-Fang Pei[2†], Lei Zhang[2], Boyi Guo[3], Amanda H. Pendegraft[3], Wenzhuo Zhuang[4] and Nengjun Yi[3]*

[1] Department of Biostatistics, Jiann-Ping Hsu College of Public Health, Georgia Southern University, Statesboro, GA, United States, [2] Department of Epidemiology and Health Statistics, School of Public Health, Medical College of Soochow University, Suzhou, China, [3] Department of Biostatistics, School of Public Health, University of Alabama at Birmingham, Birmingham, AL, United States, [4] Department of Cell Biology, School of Biology & Basic Medical Science, Soochow University, Suzhou, China

OPEN ACCESS

Edited by:
*Michele Guindani,
University of California, Irvine,
United States*

Reviewed by:
*Ye Deng,
Research Center for
Eco-Environmental Sciences (CAS),
China
Qing Lu,
Michigan State University,
United States*

***Correspondence:**
*Nengjun Yi
nyi@uab.edu*

† These authors have contributed equally to this work.

Specialty section:
This article was submitted to Systems Microbiology, a section of the journal Frontiers in Microbiology

Received: *27 October 2017*
Accepted: *06 July 2018*
Published: *26 July 2018*

Citation:
Zhang X, Pei Y-F, Zhang L, Guo B, Pendegraft AH, Zhuang W and Yi N (2018) Negative Binomial Mixed Models for Analyzing Longitudinal Microbiome Data. Front. Microbiol. 9:1683. doi: 10.3389/fmicb.2018.01683

The metagenomics sequencing data provide valuable resources for investigating the associations between the microbiome and host environmental/clinical factors and the dynamic changes of microbial abundance over time. The distinct properties of microbiome measurements include varied total sequence reads across samples, over-dispersion and zero-inflation. Additionally, microbiome studies usually collect samples longitudinally, which introduces time-dependent and correlation structures among the samples and thus further complicates the analysis and interpretation of microbiome count data. In this article, we propose negative binomial mixed models (NBMMs) for longitudinal microbiome studies. The proposed NBMMs can efficiently handle over-dispersion and varying total reads, and can account for the dynamic trend and correlation among longitudinal samples. We develop an efficient and stable algorithm to fit the NBMMs. We evaluate and demonstrate the NBMMs method via extensive simulation studies and application to a longitudinal microbiome data. The results show that the proposed method has desirable properties and outperform the previously used methods in terms of flexible framework for modeling correlation structures and detecting dynamic effects. We have developed an R package NBZIMM to implement the proposed method, which is freely available from the public GitHub repository http://github.com//nyiuab//NBZIMM and provides a useful tool for analyzing longitudinal microbiome data.

Keywords: count data, longitudinal study, microbiome, metagenomics, negative binomial mixed model

INTRODUCTION

The human microbiome plays an important role in human health and disease. The complex microbiome is inherently dynamic and interacts with the host and the environmental factors over time (Gerber, 2014a). These complex dynamics start from the birth with increasingly richness in the communities of microbiota over time (Palmer et al., 2007; Koenig et al., 2011; Wu et al., 2011; De Muinck et al., 2013; Gerber, 2014a). Recent studies have found that the human microbiome in healthy adults can be altered by various host factors including genotype (Spor et al., 2011; Blekhman et al., 2015; Goodrich et al., 2016a,b), lifestyle such as dietary habit (De Filippo et al., 2010; Wu et al., 2011), physiological status such as aging (Biagi et al., 2010), pathophysiological status (Turnbaugh et al., 2009), and host environment (Dominguez-Bello et al., 2010). The dynamic shifts in compositional features of the microbiome can occur with human diseases such as obesity (Turnbaugh et al., 2006), diabetes (Samuel and Gordon, 2006), infections or inflammatory bowel disease (Frank et al., 2007), and cancers (Holmes et al., 2011). To decipher the relationship between the dynamic changes in microbiome and human diseases, high-throughput sequencing technologies, such as the 16S ribosome RNA (rRNA) gene sequencing or shotgun metagenomics sequencing, have been widely applied in longitudinal microbiome studies (Matsen et al., 2010; Ghodsi et al., 2011; Gilbert et al., 2011; La Rosa et al., 2014).

The metagenomics sequencing data provide valuable resources for investigating the dynamic changes of microbial abundance over time and the associations between the microbiome and host environmental/clinical factors. Multiple recent microbiome studies have employed the longitudinal study designs to address the crucial research question (La Rosa et al., 2014; DiGiulio et al., 2015; Zhou et al., 2015; Ward et al., 2016). Among them, La Rosa et al. (2014) utilized longitudinal analysis of repeated measures data to demonstrate that the dynamic shifts in dominating microbiota of the infant gut from *Bacilli* at birth, giving way to *Gammaproteobacteria*, then *Clostridia* at the end of the first month of life. In another recent published study, Ward et al. (2016) used longitudinal study to address the associations between the dynamic change of the early intestinal microbiome in preterm infants and the occurrence of Necrotizing enterocolitis (NEC) or NEC-associated deaths.

Despite our ability to generate large-scale metagenomics sequencing longitudinal data, many challenges exist in the development of robust and powerful statistical methods and computational tools for properly analyzing and interpreting longitudinal microbiome data. The metagenomics sequencing data has some properties that require tailored analytic tools; these include varied total sequence reads across samples, over-dispersion and zero-inflation. One common way to account for varying total reads is normalization, i.e., conversion of the sequence counts to the relative abundance (or proportion) using the total sum, mean, or median of representative OTUs across all samples (Anders and Huber, 2010; Robinson and Oshlack, 2010; Knights et al., 2011; Wagner et al., 2011; Kostic et al., 2012; Paulson et al., 2013). Several zero-inflated models were proposed to correct for excess zero counts in microbiome measurements, including zero-inflated Gaussian, lognormal, negative binomial, and beta models (Paulson et al., 2013; Peng et al., 2015; Sohn et al., 2015; Xu et al., 2015). On the other hand, the negative binomial regression, which is a standard statistical method for analyzing over-dispersed count observations, has been recently applied to microbiome data (White et al., 2009; Pookhao et al., 2015).

It is even more challenging to analyze longitudinal microbiome count data. In addition to the special features of microbiome data, longitudinal studies possesses two fundamental time-dependent features: (a) time imposes an inherent and irreversible ordering on samples, and (b) samples exhibit statistical dependencies that are a function of time (Gerber, 2014b). Ignoring these properties of longitudinal data and applying statistical tools designed for analyzing static data can result in erroneous conclusions (Gerber, 2014a). Most of the previous studies resort to linear mixed models (LMMs) to account for time-dependent correlations in longitudinal microbiome study designs by treating transformed data as normally distributed responses (Benson et al., 2010; Srinivas et al., 2013; La Rosa et al., 2014; Leamy et al., 2014; Wang et al., 2015). However, using LMMs directly without addressing properties of microbiome data may result in lower power or potential inaccurate results to detect the dynamic effects of microbiota. Chen and Li (2016) developed zero-inflated beta mixed models for analyzing transformed proportions in microbiome longitudinal studies, but did not address time trends and within-subject correlations. Thus, statistical models to account for time series as well as properties of microbiome count data are required for analyzing microbiome data (Spor et al., 2011; Faust et al., 2015; Chen and Li, 2016).

Zhang et al. (2017) have recently developed negative binomial mixed models (NBMMs) for analyzing clustered microbiome data, but have not addressed longitudinal studies yet. We here extend negative binomial mixed models (NBMMs) proposed by Zhang et al. (2017) to analyze longitudinal microbiome count data. The extended NBMMs can include various types of fixed effects and random effects, and can incorporate various correlation structures among observations within the same subjects, thus fully addressing the special properties of longitudinal microbiome count data. We develop an efficient and stable IWLS (iterative weighted least squares) algorithm to fit the extended NBMMs by taking advantage of the standard procedure for fitting linear mixed models. Through extensive simulations, we show that the NBMMs outperform the previously used LMMs in terms of detecting dynamic effects in longitudinal microbiome count data. We also apply our method to a previously published microbiome data to detect significantly dynamic effects of associated taxa. We have implemented the proposed method in the R package NBZIMM, which is freely available from the public GitHub repository http://github.com//nyiuab//NBZIMM and provides a useful tool for longitudinal microbiome studies.

METHODS

Negative Binomial Mixed Models (NBMMS) for Longitudinal Microbiome Studies

Longitudinal studies collect multiple subjects and measure each subject at multiple time points (i.e., samples). Assume that there are n subjects, and subject i is measured at n_i time points t_{ij}; $j = 1, \cdots, n_i$; $i = 1, \cdots, n$. For each sample, microbiome data generated by the 16S rRNA gene sequencing or the shotgun metagenomics sequencing consist of counts for numerous taxa at certain taxonomic levels (OTU, species, genus, classes, etc.), c_{ijh}, $h = 1, \cdots, m$, and total sequence read T_{ij} (also referred to as depths of coverage or library size). We also measure some host clinical/environmental variables for each subject, X_i. **Table 1** summarizes the data structure for a longitudinal microbiome study. The goal of longitudinal microbiome studies is to detect associations between the microbiome counts and the

TABLE 1 | Longitudinal microbiome data structure.

Subject ID	Taxon 1	Taxon 2	···	Taxon *m*	Total reads	Host factors	Time variables
Subject 1	c_{111}	c_{112}	···	c_{11m}	T_{11}	X_1	t_{11}
Subject 1	c_{121}	c_{122}	···	c_{12m}	T_{12}	X_1	t_{12}
Subject 1	c_{131}	c_{132}	···	c_{13m}	T_{13}	X_1	t_{13}
Subject 2	c_{211}	c_{212}	···	c_{21m}	T_{21}	X_2	t_{21}
	·	·	·	·	·	·	·
	·	·	·	·	·	·	·
	·	·	·	·	·	·	·
Subject n	c_{n11}	c_{n12}	···	c_{n1m}	T_{n1}	X_n	t_{n1}

host variables, and characterize the time trends of microbiome abundance within subjects and between subjects.

We separately analyze each microbiome taxon, as most existing methods. For notational simplification, we denote $y_{ij} = c_{ijh}$ for any given taxon h. Since the microbiome count outcome is over-dispersed, we use negative binomial models. We extend negative binomial mixed models (NBMMs) proposed by Zhang et al. (2017) to analyze longitudinal microbiome data by including the time variable and its interaction with the host factor of interest in the model. In the next section, we will further extend NBMMs to account for within-subject correlation structures.

In our NBMMs, the counts y_{ij} are assumed to follow the negative binomial distribution:

$$y_{ij} \sim NB(y_{ij} \mid \mu_{ij}, \theta) = \frac{\Gamma(y_{ij}+\theta)}{\Gamma(\theta) y_{ij}!} \cdot \left(\frac{\theta}{\mu_{ij}+\theta}\right)^{\theta} \cdot \left(\frac{\mu_{ij}}{\mu_{ij}+\theta}\right)^{y_{ij}} \quad (1)$$

where θ is the dispersion parameter that controls the amount of over-dispersion, and μ_{ij} are the means. The means μ_{ij} are related to the host variables via the logarithm link function:

$$\log(\mu_{ij}) = \log(T_{ij}) + X_{ij}\beta + Z_{ij}b_i \quad (2)$$

where $\log(T_{ij})$ is the offset that corrects for the variation of the total sequence reads, $X_{ij} = (1, X_i, t_{ij}, X_i^s t_{ij})$, X_i^s is the variable of interest in X_i, for example, an indicator variable for the case group and the control group, and$Z_{ij} = (1, t_{ij})$; $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)^{\mathrm{T}}$ is the vector of fixed effects (i.e., population-level effects), including an intercept β_0, the effects β_1 of the host variables X_i, the overall time effect β_2, and the interaction β_3 betweenX_i^s andt_{ij}; $b_i = (b_{0i}, b_{1i})^{\mathrm{T}}$ is the vector of random effects (i.e., subject-level effects), including the random intercept b_{0i} and the random time effect b_{1i}. For simplicity, the above model only considers the linear function of t_{ij}. If sample size is large enough, however, we can extend the model to use polynomial functions, for example, (t_{ij}, ${t_{ij}}^2$), or B-spline functions, allowing us to detect arbitrary temporal trends.

The random effects are used to model multiple sources of variations and subject-specific effects, and thus avoid biased inference on the fixed effects. The vector of the random effects is usually assumed to follow a multivariate normal distribution (Pinheiro and Bates, 2000; McCulloch and Searle, 2001):

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

where Ψ is the variance-covariance matrix. Ψ can be a general positive-definite matrix that accounts for the correlation of the random covariates. In some applications, however, we can restrict Ψ to special forms of variance-covariance matrices that are parameterized by fewer parameters. For example, we may assume that the random effects are independent, in which case Ψ is a diagonal matrix.

Accounting for Within-Subject Correlations and IWLS Algorithm for Fitting the NBMMS

The IWLS (Iterative Weighted Least Squares) algorithm developed by Zhang et al. (2017) can be used to fit the above NBMMs. The basic idea of the IWLS algorithm is to iteratively approximate the negative binomial mixed model by a linear mixed model. However, Zhang et al. (2017) restricts the within-subject errors in the linear mixed model to be independent, and thus ignores special within-subject correlation structures. For longitudinal data, however, samples within the same subject are usually correlated. Thus, we extend the model by relaxing the assumption of independent within-subject errors to account for special within-subject correlation structures:

$$\begin{aligned} z_{ij} = \log(T_{ij}) + X_{ij}\beta + Z_{ij}b_i + w_{ij}^{-1/2} e_{ij},\ \ b_i \sim N(0, \Psi), \\ e_i = (e_{i1}, \cdots, e_{in_i})' \sim N(0, \sigma^2 R_i) \quad (4) \end{aligned}$$

where z_{ij} and w_{ij} are the pseudo-responses and the pseudo-weights, respectively, that depend on $\log(T_{ij}) + X_{ij}\hat{\beta} + Z_{ij}\hat{b}_i$ and $\hat{\theta}$as described in Zhang et al. (2017), and R_i is a correlation matrix, which describes dependence among observations, Pinheiro and Bates (2000) describes several ways to specify the correlation matrix R_i, all of which can be incorporated into our NBMMs. For longitudinal studies, a common choice of R_i is autoregressive of order 1, AR(1), or continuous-time AR(1).

We extend the IWLS algorithm developed by Zhang et al. (2017) to fit the proposed NBMMS with correlation structures. The algorithm alternatively updates the dispersion θ and the parameters in the linear mixed model (4). Given the estimates of β and b, we update the dispersion parameter θ by maximizing the negative binomial likelihood using the standard Newton-Raphson algorithm, and then calculate the pseudo-responses and the pseudo-weights. We then fit the linear mixed model (4) using the standard method as implemented in the core package **nmle** in R. At convergence of the algorithm, we get the maximum likelihood estimates of all the fixed effects β_k and their confidence intervals from the final linear mixed model. We then can test H_0: $\beta_k = 0$ following the linear mixed model framework.

R Package for Implementing the Proposed Method

We have created the function **glmm.nb** for setting up and fitting the proposed NBMMs, which is part of the R package **NBZIMM**. The function **glmm.nb** works by repeated calls to the function **lme** for fitting linear mixed models in the recommended package **nlme** in R, and allows for any types of random effects and within-subject correlation structures as described in the package **nlme**. The outputs from the function **glmm.nb** can be summarized by functions in **nlme**. The package **NBZIMM** is freely available from the public GitHub repository http://github.com//nyiuab//NBZIMM.

RESULTS

Simulation Studies

Simulation Designs

We performed extensive simulations to evaluate the proposed methods. We extended the simulation framework of Zhang et al. (2017) to simulate longitudinal microbiome counts from negative binomial distributions and incorporate time covariates, random effects and within-subject correlation structures.

Our simulation studies employed a case-control longitudinal study design with four different settings. All the four simulation settings followed a two-level longitudinal study, where all individuals (subjects) were from two groups (i.e., case or control) and multiple samples were measured at several time points for each individual. For all the settings, we simulated ($n =$) 50, 100 or 150 individuals, half of which were cases, and included three fixed covariates: a binary case-control indicator variable x_i, a continuous time variable t_{ij}, and their interaction. We denote the fixed effects of these three covariates by (β_1, β_2, β_3). The time points, random effects, and within-subject correlation structures were set as follows:

1) Setting A: 5 time points for each individual, only random intercept, and no within-subject correlation;
2) Setting B: 10 time points for each individual, only random intercept, and the within-subject correlation was autoregressive of order 1, AR(1);
3) Setting C: 5 time points for each individual, two random effects (i.e., random intercept and time effect), and no within-subject correlation;
4) Setting D: 4 or 5 different time points for individuals, only random intercept, and no within-subject correlation;

To minimize possible bias and yield reasonable count values that are similar to real microbiome data, we randomly generated the parameters in the model from reasonable ranges at each simulation replication (Zhang et al. 2017), which are described as follows:

1) The values, $\log(T_{ij}) + \beta_0$, control the means of simulated counts when all the effects are zero, where β_0 is the fixed intercept. We set $\beta_0 = -7$ and randomly sampled $\log(T_{ij})$ from the range [7.1, 10.5]. In this case, $\log(T_{ij}) + \beta_0$ were in the range [0.1, 3.5], which yield counts similar to real microbiome data;
2) The dispersion parameter θ were uniformly sampled from the range [0.1, 5], which yield highly or moderate over-dispersed counts;
3) To evaluate false positive rates, the fixed effects β_1, β_2 and β_3 were all set to be zero. To evaluate empirical powers, we considered four scenarios: a) β_1 and β_2 were set to 0, and β_3 was sampled from [0.2, 0.35]; b) β_1 and β_2 were set to 0, and β_3 was sampled from [0.35, 0.8]; c) β_1, β_2 and β_3 were all sampled from [0.2, 0.35]; d) β_1, β_2 and β_3 were all sampled from [0.35, 0.8];
4) The random effects b_{0i} and b_{1i} were generated from N(0, τ^2), where τ was randomly drawn from the range [0.5, 1];
5) The correlation coefficient ρ for AR(1) correlation was sampled from [0.1, 0.5], and the AR(1) correlation was generated by the function *arima.sim()* from R package *stats*;
6) The standard deviation σ was sampled from [0.1, 0.5];

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, β_2, β_3 (false positive rate)	0, 0, 0
Fixed effects β_1, β_2, β_3 (power of interaction)	0, 0, Unif(0.2, 0.35) or Unif(0.35, 0.8)
Fixed effects β_1, β_2, β_3 (power of both β_1 and β_3)	All from Unif(0.2, 0.35) or Unif(0.35, 0.8)
Standard deviation τ	Unif(0.5, 1)
Correlation ρ	Unif(0.1, 0.5)
Standard deviation σ	Unif(0.1, 0.5)

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

In all the four simulation settings, the procedure was repeated 10,000 times. At each replication, the parameters were sampled from the ranges described above. There were two hypotheses of interests to be tested, i.e., the group main effect $\beta_1 = 0$ and the group by time interaction $\beta_3 = 0$. Both empirical power and false positive rate for testing the hypotheses were calculated under significance level at 0.05. The empirical power and false positive rate were defined as the proportions of detecting non-zero and zero effects over the simulation replications, respectively. We compared the proposed NBMMs with the linear mixed model with the arcsine square root transformation, $arcsine\left(\sqrt{y_{ij}/T_{ij}}\right)$, as the response, denoted by LMM arcsin.

Simulation Results

Figure 1 and Figure A.1 show the empirical power to detect the group by time interaction under the four different simulation settings, when the group main effect was set to zero. The power was affected by the sample size. It can be clearly seen that the proposed method performed consistently better than the LMM arcsin method across almost all the scenarios. The second setting was set to represent time-series structure in longitudinal data with 10 measurements for each individual, and thus had the largest power among all the four settings. It was shown that the first setting had higher power than the third setting, on the other hand, a similar performance in power compared with the fourth setting.

It is of interest to detect both the group main effect and the group by time interaction. Therefore, in another set of parameter settings, we targeted to detect both the group main effect and the group by time interaction. **Figure 2** and Figure A.2 show the

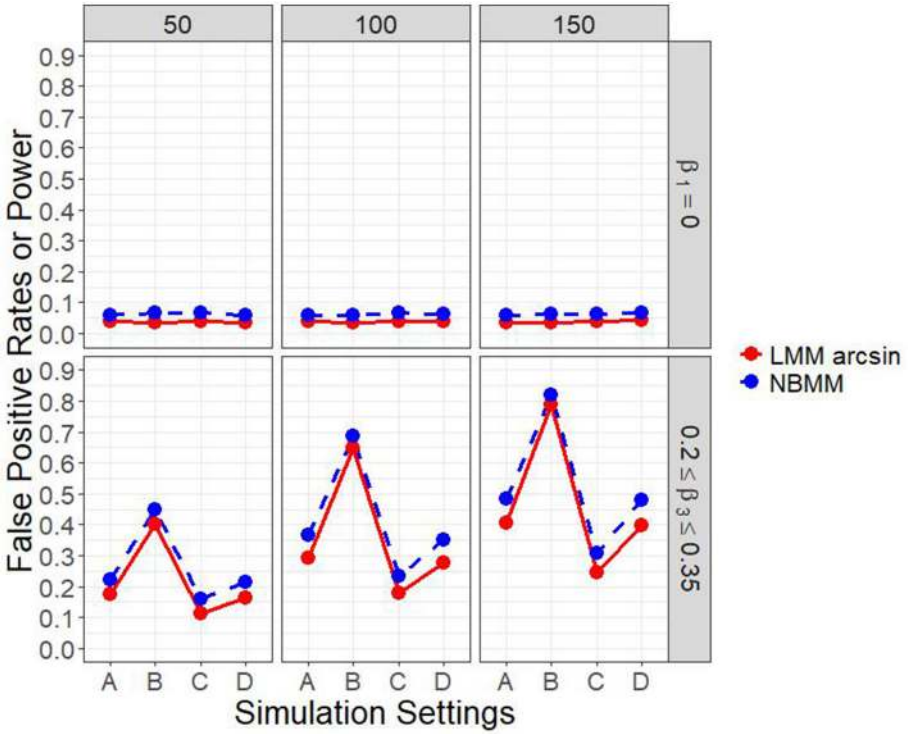

FIGURE 1 | Empirical power of interaction term and false positive rates of main effect in all four simulation settings.

empirical power to detect both the group main effect and the group by time interaction under the four different simulation settings. The results showed that the LMM arcsine method resulted in a slightly higher power in detecting interaction term than our proposed method across all the scenarios. However, it showed an extreme low power close to alpha level in detecting the group main effect across all the scenarios. It inferred that LMM arcsine method is not an appropriate approach to be used when the group main effect and the group by time interaction effect are both nonzero. **Figure 3** displays the false positive rates for detecting both the group main effect and interaction effect. For all the four simulation settings, the false positive rates were well controlled under all the scenarios.

Application to Temporal and Spatial Pregnant Data

We applied our method to a public microbiome data from a longitudinal study to investigate the bacterial taxonomic composition for pregnant and postpartum women by DiGiulio et al. (2015). This case-control longitudinal study included 49 pregnant women, 15 of whom delivered preterm. The discovery data was consisted with 40 of those women. Among those 40 women, they collected 3,767 specimens prospectively and weekly during gestation and monthly after delivery from the vagina, distal gut, saliva, and tooth/gum. The specimens were analyzed for bacterial taxonomic composition. The final dataset contained a total of 1271 taxa from 3432 specimens which were identified for pregnant women delivered at term and preterm. Detailed information about population and material is available in DiGiulio et al. (2015). Clinical data included race, weeks/days when the samples were obtained, way of delivery, and household income level were acquired. The public processed OTU data available from the study is from species level. The clinical data for the validation dataset for the rest of 9 pregnant women is not available.

We used the proposed NBMMs and the linear mixed models (LMMs) with the arcsine square root transformations to detect associations between delivery term and vaginal bacteria taxa composition during pregnancy. The host factor in the analysis was defined as two groups with patients who delivered at preterm vs. term. The patients who delivered at marginal term were excluded from the analysis. Only specimens collected in vaginal during pregnancy were included in the analysis. Meanwhile, according to the original paper, the samples could be divided to 5 Vaginal Community State Types. Only samples with community state type 4 were analyzed in the original paper. To be consistent, we followed the same criteria for sample filtering. The sample size in the final analysis was 103. We included 58 taxa with zero proportion greater than 0.25 for 103 samples in our analysis. The real data and the R code for our analysis are available from the GitHub page: https://abbyyan3.github.io//NBZIMM-tutorial/NBZIMM_NBMMs_Longitudinal.html.

To compare the abilities of LMMs and NBMMs in detecting the static and dynamic association between host factor and vaginal bacterial taxa composition, we used the following four different models:

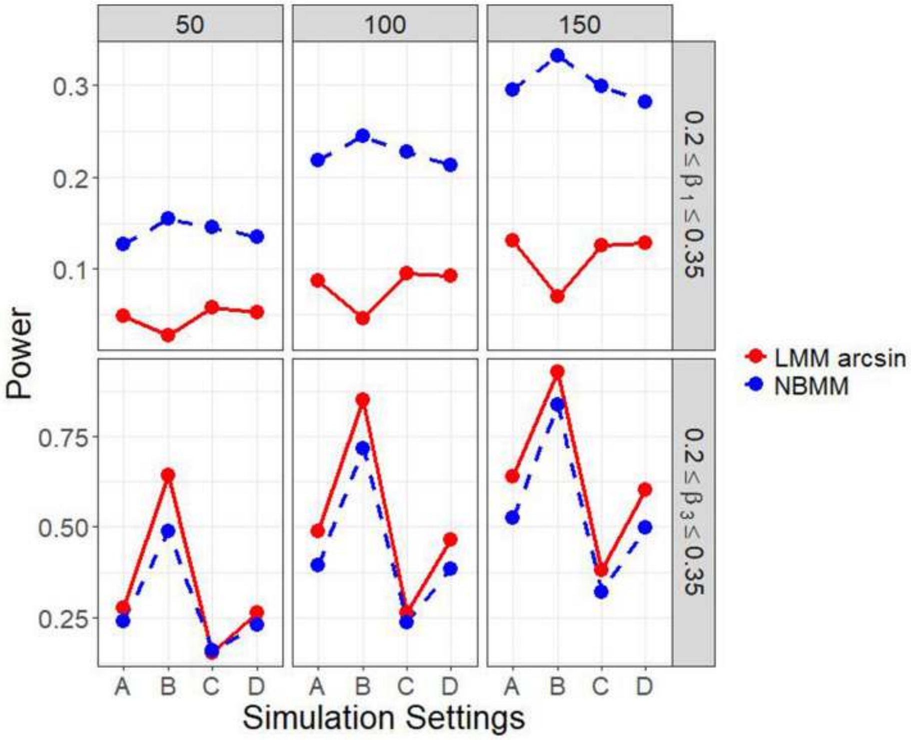

FIGURE 2 | Empirical power of both interaction term and main effect in all four simulation settings.

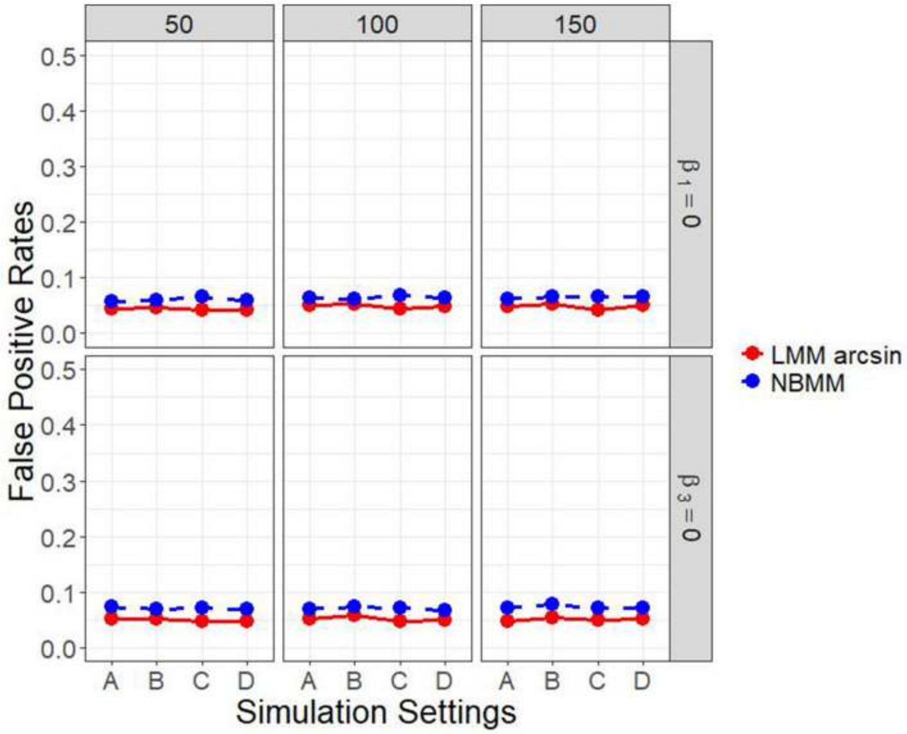

FIGURE 3 | False positive rates of both interaction term and main effect in all four simulation settings.

TABLE 3 | Significant taxa rates detected in four models with LMMs and NBMMs.

		Alpha Level	0.05
Model 1	Test of β_1	LMMs	0.034483
		NBMMs	0.068966
Model 2	Test of β_1	LMMs	0.034483
		NBMMs	0.12069
Model 3	Test of β_1	LMMs	0.12069
		NBMMs	0.224138
	Test of β_3	LMMs	0.137931
		NBMMs	0.275862
Model 4	Test of β_1	LMMs	0.137931
		NBMMs	0.206897
	Test of β_3	LMMs	0.137931
		NBMMs	0.293103

1) Model A: the host factor as fixed effect only, no host factor and time interaction term, only random intercept;
2) Model B: the host factor as fixed effect only, no host factor and time interaction term, two random effects (i.e., random intercept and time effect);
3) Model C: the host factor, time, host factor and time interaction term as fixed effects, only random intercept;
4) Model D: the host factor, time, host factor and time interaction term as fixed effects, two random effects (i.e., random intercept and time effect);

We summarized the number of significant taxa and calculated the rate of significant taxa detected by LMMs and NBMM each using Model A-D at alpha level at 0.05 (**Table 3).** In model A and model B, the numbers of detected significant taxa were substantially less than the numbers from model C and model D. It inferred that failing to incorporate the host factor and time interaction term as fixed effect in the model will largely affect our ability to detect shifts in microbiome studies. Meanwhile, it showed that our NBMMs is capable in detecting more significant taxa than LMMs. Consistent differences have also been found at different significance levels, like 0.01 and 0.001.

Figure 4 shows the significant features of species level in the model with the host factor and the host factor and time interaction term both at the 5% significance threshold and their minus log transformed *p*-values for NBMMs and LMMs. It showed that NBMMs could discover more species than LMMs in detecting both static association (with host factor term) and dynamic association (with host factor and time interaction term). To compare our analysis results with the published results in DiGiulio et al. (2015), we found that the original paper made two extreme assumptions to the longitudinal study as completely independent or averaged over samples for each subject. The top identified taxa overlapped between our NBMMs with the original paper included *Gardnerella_137183, Lactobacillus jensenii_31171, Staphylococcus aureus_4446058, Lactobacillus crispatus_4447432, Prevotella_760967, Dialister_1105876*. In summary, our NBMMs method is not only a statistical valid method without making extreme assumptions and data transformation, but also detected more significant taxa and yielded much smaller *p*-values than the LMMs, showing that the proposed method could be more powerful than the conventional LMMs.

DISCUSSION

The main research interest in longitudinal microbiome study is to detect the associations between host clinical/environmental factors and the dynamic shifts in microbiome composition while accounting for sources of heterogeneity and dependence in microbiome measurements. To study the dynamic composition of microbiome, many studies collect samples with temporal structures (Hill et al., 2010; Morrow et al., 2013; Srinivas et al., 2013; La Rosa et al., 2014; Leamy et al., 2014; Faust et al., 2015; Wang et al., 2015; Zhou et al., 2015). These longitudinal studies enable us to study the inherent dynamic properties in microbiome data which have provided extraordinary opportunities to elucidate the true roles of the microbiome in health and disease states and to develop new diagnostics and therapeutic targets (Knights et al., 2011; Segata et al., 2011; Virgin and Todd, 2011; Collison et al., 2012). Accurately identifying and understanding these associations is critical to further predict the probabilities of disease with the identified taxa or biomarkers. However, the traditional methods of using LMMs to model longitudinal data fail to address the count data features in microbiome data. Our simulation studies revealed the impact of the specific features on the microbiome data, showing that ignoring those features can substantially reduce the power for detecting the effects of host clinical/environmental factors with dynamic effects, thus leading to biased and false inferences. We extended our previously proposed negative binomial mixed model (NBMMs) specifically to directly analyze longitudinal microbiome count data without data transformation.

The previously proposed NBMMs (Zhang et al., 2017) have demonstrated its superior ability in family structured clustered microbiome count data. The proposed NBMMs directly model microbiome counts generated by the 16S rRNA gene sequencing or the shotgun sequencing with an efficient IWLS algorithm (Schall, 1991; Breslow and Clayton, 1993; McCulloch and Searle, 2001; Venables and Ripley, 2002). It not only addresses statistical challenges of over-dispersion and varied total reads in microbiome count data, but also accounts for correlation among the observations. Our simulations and real data analysis also show that our algorithm is stable and efficient (Zhang et al., 2017). Meanwhile, the IWLS algorithm is an extension of a commonly used procedure for fitting GLMs and GLMMs which allows us to model non-constant variances or special correlation structures. Therefore, by extending the NBMMs to analyze longitudinal microbiome count data, we illustrated the capability of our proposed NBMMs to handle complex longitudinal study design, such as to include time in the random slope model or to account for the auto-regressive residual correlation in time-series data.

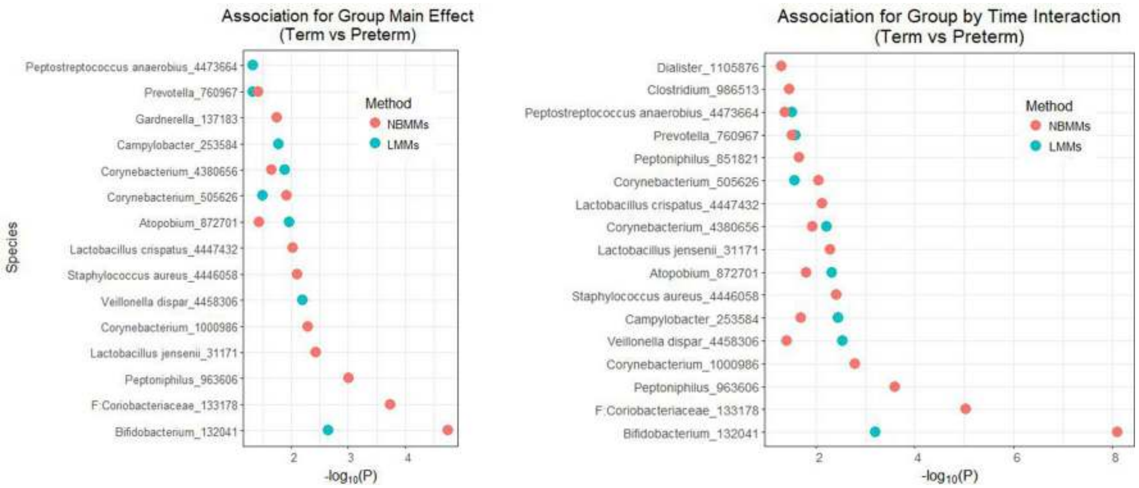

FIGURE 4 | The analyses of NBMMs and LMM: minus log transformed *p*-values for the significant differentially abundant taxa at the 5% significance threshold between term and preterm groups for species level; the left panel shows the minus log transformed *p*-values for association of group main effect, and the right panel shows the minus log transformed *p*-values for association for group by time interaction.

Our simulations indicate that our proposed approach is flexible to handle complex structured longitudinal data, allowing for incorporating any types of random effects and within-subject correlation structures (Pinheiro and Bates, 2000; McCulloch and Searle, 2001). In the simulations, our proposed approach outperformed LMMs consistently.

We also applied our method to a previously published data set. The purpose of the real data is to detect host factors that associated with dynamic compositional features of the microbiome (Leamy et al., 2014). Notably, by applying our NBMMs to the temporal and spatial dataset from DiGiulio et al. (2015), the goal of our analysis was to detect taxa that are significantly associated with dynamic change in compositional microbiome between termed and preterm pregnancy. Our proposed method detected the same species *Gardnerella_137183, Lactobacillus jensenii_31171, Staphylococcus aureus_4446058, Lactobacillus crispatus_4447432, Prevotella_760967, Dialister_1105876,* as in the original paper. In the original paper, they made two extreme assumptions to the longitudinal study as completely independent or averaged over samples for each subject. Our NBMMs, on the other hand, does not make any extreme assumption and is more statistically valid. Nevertheless, we still identified overlapped species as in the original paper, showing NBMMs picked out the significant species under extremes as well. Our NBMMs method detected more significant taxa and yielded much smaller *p*-values than the LMMs, showing that the proposed method could be more powerful than the conventional LMMs. Furthermore, comparing the species identified in the real data using LMMs and NBMMs, we found that the species identified by NBMMs only are mostly overlapped with the original paper. It inferred that the transformation of count data could potentially lead to misleading information and interpretation. One potential limitation of our NBMMs is that it is not designed to explicitly handle zero-inflation and we recommend it as future work. Even though, our NBMMs has shown it outperformed LMMs in longitudinal microbiome study in terms of power and accurate interpretation. It is also directly applicable to be used as an analytic tool in longitudinal RNA-seq study.

AUTHOR CONTRIBUTIONS

NY design the study, develop the method and the software, and participate in writing the paper; XZ simulation study, real analysis, and draft the manuscript; Y-FP design the study, real data analysis, and participate in writing the paper; LZ design the study, and participate in writing the paper; BG design the simulation and real data analysis; AP participate in revising the manuscript; WZ real data analysis, and participate in revising the manuscript.

ACKNOWLEDGMENTS

This work was supported in part by research grants from USA National Institutes of Health (R03-DE024198, R03-DE025646), National Natural Science Foundation of China (31771417), National Natural Science Foundation of China (81673448), and Natural Science Foundation of Jiangsu Province China (BK 20161218).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2018.01683/full#supplementary-material

REFERENCES

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

Benson, A. K., Kelly, S. A., Legge, R., Ma, F., Low, S. J., Kim, J., et al. (2010). 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.* 107, 18933–18938. doi: 10.1073/pnas.1007028107

Biagi, E., Nylund, L., Candela, M., Ostan, R., Bucci, L., Pini, E., et al. (2010). Through ageing, and beyond: gut microbiota and inflammatory status in seniors and centenarians. *PLoS ONE* 5:e10667. doi: 10.1371/annotation/df45912f-d15c-44ab-8312-e7ec0607604d

Blekhman, R., Goodrich, J. K., Huang, K., Sun, Q., Bukowski, R., Bell, J. T., et al. (2015). Host genetic variation impacts microbiome composition across human body sites. *Genome Biol.* 16:191. doi: 10.1186/s13059-015-0759-1

Breslow, N. E., and Clayton, D. C. (1993). Approximate inference in generalized linear mixed models. *J. Am. Stat. Assoc.* 88, 9–25.

Chen, E. Z., and Li, H. (2016). A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. *Bioinformatics* 32, 2611–2617. doi: 10.1093/bioinformatics/btw308

Collison, M., Hirt, R. P., Wipat, A., Nakjang, S., Sanseau, P., and Brown, J. R. (2012). Data mining the human gut microbiota for therapeutic targets. *Brief Bioinformatics* 13, 751–768. doi: 10.1093/bib/bbs002

De Filippo, C., Cavalieri, D., Di Paola, M., Ramazzotti, M., Poullet, J. B., Massart, S., et al. (2010). Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. *Proc. Natl. Acad. Sci. U.S.A.* 107, 14691–14696. doi: 10.1073/pnas.1005963107

De Muinck, E. J., Lagesen, K., Afset, J. E., Didelot, X., Ronningen, K. S., Rudi, K., et al. (2013). Comparisons of infant *Escherichia coli* isolates link genomic profiles with adaptation to the ecological niche. *BMC Genomics* 14:81. doi: 10.1186/1471-2164-14-81

DiGiulio, D. B., Callahan, B. J., Mcmurdie, P. J., Costello, E. K., Lyell, D. J., Robaczewska, A., et al. (2015). Temporal and spatial variation of the human microbiota during pregnancy. *Proc. Natl. Acad. Sci. U.S.A.* 112, 11060–11065. doi: 10.1073/pnas.1502875112

Dominguez-Bello, M. G., Costello, E. K., Contreras, M., Magris, M., Hidalgo, G., Fierer, N., et al. (2010). Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proc. Natl. Acad. Sci. U.S.A.* 107, 11971–11975. doi: 10.1073/pnas.1002601107

Faust, K., Lahti, L., Gonze, D., De Vos, W. M., and Raes, J. (2015). Metagenomics meets time series analysis: unraveling microbial community dynamics. *Curr. Opin. Microbiol.* 25, 56–66. doi: 10.1016/j.mib.2015.04.004

Frank, D. N., St Amand, A. L., Feldman, R. A., Boedeker, E. C., Harpaz, N., and Pace, N. R. (2007). Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proc. Natl. Acad. Sci. U.S.A.* 104, 13780–13785. doi: 10.1073/pnas.0706625104

Gerber, G. K. (2014a). The dynamic microbiome. *FEBS Lett.* 588, 4131–4139. doi: 10.1016/j.febslet.2014.02.037

Gerber, G. K. (2014b). "Longitudinal Microbiome Data Analysis," in *Metagenomics for Microbiology*, eds J. Izardm (Cambridge, MA: Academic Press).

Ghodsi, M., Liu, B., and Pop, M. (2011). DNACLUST: accurate and efficient clustering of phylogenetic marker genes. *BMC Bioinformatics* 12:271. doi: 10.1186/1471-2105-12-271

Gilbert, J. A., Meyer, F., and Bailey, M. J. (2011). The future of microbial metagenomics (or is ignorance bliss?). *ISME J.* 5, 777–779. doi: 10.1038/ismej.2010.178

Goodrich, J. K., Davenport, E. R., Beaumont, M., Jackson, M. A., Knight, R., Ober, C., et al. (2016a). Genetic determinants of the gut microbiome in UK twins. *Cell Host Microbe* 19, 731–743. doi: 10.1016/j.chom.2016.04.017

Goodrich, J. K., Davenport, E. R., Waters, J. L., Clark, A. G., and Ley, R. E. (2016b). Cross-species comparisons of host genetic associations with the microbiome. *Science* 352, 532–535. doi: 10.1126/science.aad9379

Hill, D. A., Hoffmann, C., Abt, M. C., Du, Y., Kobuley, D., Kirn, T. J., et al. (2010). Metagenomic analyses reveal antibiotic-induced temporal and spatial changes in intestinal microbiota with associated alterations in immune cell homeostasis. *Mucosal Immunol.* 3, 148–158. doi: 10.1038/mi.2009.132

Holmes, E., Li, J. V., Athanasiou, T., Ashrafian, H., and Nicholson, J. K. (2011). Understanding the role of gut microbiome-host metabolic signal disruption in health and disease. *Trends Microbiol.* 19, 349–359. doi: 10.1016/j.tim.2011.05.006

Knights, D., Parfrey, L. W., Zaneveld, J., Lozupone, C., and Knight, R. (2011). Human-associated microbial signatures: examining their predictive value. *Cell Host Microbe* 10, 292–296. doi: 10.1016/j.chom.2011.09.003

Koenig, J. E., Spor, A., Scalfone, N., Fricker, A. D., Stombaugh, J., Knight, R., et al. (2011). Succession of microbial consortia in the developing infant gut microbiome. *Proc. Natl. Acad. Sci. U.S.A.* 108, (Suppl. 1), 4578–4585. doi: 10.1073/pnas.1000081107

Kostic, A. D., Gevers, D., Pedamallu, C. S., Michaud, M., Duke, F., Earl, A. M., et al. (2012). Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. *Genome Res.* 22, 292–298. doi: 10.1101/gr.126573.111

La Rosa, P. S., Warner, B. B., Zhou, Y., Weinstock, G. M., Sodergren, E., Hall-Moore, C. M., et al. (2014). Patterned progression of bacterial populations in the premature infant gut. *Proc. Natl. Acad. Sci. U.S.A.* 111, 12522–12527. doi: 10.1073/pnas.1409497111

Leamy, L. J., Kelly, S. A., Nietfeldt, J., Legge, R. M., Ma, F., Hua, K., et al. (2014). 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.* 15:552. doi: 10.1186/s13059-014-0552-6

Matsen, F. A., Kodner, R. B., and Armbrust, E. V. (2010). Pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics* 11:538. doi: 10.1186/1471-2105-11-538

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

Morrow, A. L., Lagomarcino, A. J., Schibler, K. R., Taft, D. H., Yu, Z., Wang, B., et al. (2013). Early microbial and metabolomic signatures predict later onset of necrotizing enterocolitis in preterm infants. *Microbiome* 1:13. doi: 10.1186/2049-2618-1-13

Palmer, C., Bik, E. M., Digiulio, D. B., Relman, D. A., and Brown, P. O. (2007). Development of the human infant intestinal microbiota. *PLoS Biol.* 5:e177. doi: 10.1371/journal.pbio.0050177

Paulson, J. N., Stine, O. C., Bravo, H. C., and Pop, M. (2013). Differential abundance analysis for microbial marker-gene surveys. *Nat. Methods* 10, 1200–1202. doi: 10.1038/nmeth.2658

Peng, X., Li, G., and Liu, Z. (2015). Zero-inflated beta regression for differential abundance analysis with metagenomics data. *J. Comput. Biol.* 23, 102–110 doi: 10.1089/cmb.2015.0157

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

Pookhao, N., Sohn, M. B., Li, Q., Jenkins, I., Du, R., Jiang, H., et al. (2015). A two-stage statistical procedure for feature selection and comparison in functional analysis of metagenomes. *Bioinformatics* 31, 158–165. doi: 10.1093/bioinformatics/btu635

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

Samuel, B. S., and Gordon, J. I. (2006). A humanized gnotobiotic mouse model of host-archaeal-bacterial mutualism. *Proc. Natl. Acad. Sci. U.S.A.* 103, 10011–10016. doi: 10.1073/pnas.0602187103

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika* 78, 719–727. doi: 10.1093/biomet/78.4.719

Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W. S., et al. (2011). Metagenomic biomarker discovery and explanation. *Genome Biol.* 12:R60. doi: 10.1186/gb-2011-12-6-r60

Sohn, M. B., Du, R., and An, L. (2015). A robust approach for identifying differentially abundant features in metagenomic samples. *Bioinformatics* 31, 2269–2275. doi: 10.1093/bioinformatics/btv165

Spor, A., Koren, O., and Ley, R. (2011). Unravelling the effects of the environment and host genotype on the gut microbiome. *Nat. Rev. Microbiol.* 9, 279–290. doi: 10.1038/nrmicro2540

Srinivas, G., Moller, S., Wang, J., Kunzel, S., Zillikens, D., Baines, J. F., et al. (2013). Genome-wide mapping of gene-microbiota interactions in susceptibility to autoimmune skin blistering. *Nat. Commun.* 4:2462. doi: 10.1038/ncomms3462

Turnbaugh, P. J., Hamady, M., Yatsunenko, T., Cantarel, B. L., Duncan, A., Ley, R. E., et al. (2009). A core gut microbiome in obese and lean twins. *Nature* 457, 480–484. doi: 10.1038/nature07540

Turnbaugh, P. J., Ley, R. E., Mahowald, M. A., Magrini, V., Mardis, E. R., and Gordon, J. I. (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* 444, 1027–1031. doi: 10.1038/nature05414

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

Virgin, H. W., and Todd, J. A. (2011). Metagenomics and personalized medicine. *Cell* 147, 44–56. doi: 10.1016/j.cell.2011.09.009

Wagner, B. D., Robertson, C. E., and Harris, J. K. (2011). Application of two-part statistics for comparison of sequence variant counts. *PLoS* ONE 6:e20296. doi: 10.1371/journal.pone.0020296

Wang, J., Kalyan, S., Steck, N., Turner, L. M., Harr, B., Kunzel, S., et al. (2015). Analysis of intestinal microbiota in hybrid house mice reveals evolutionary divergence in a vertebrate hologenome. *Nat. Commun.* 6:6440. doi: 10.1038/ncomms7440

Ward, D. V., Scholz, M., Zolfo, M., Taft, D. H., Schibler, K. R., Tett, A., et al. (2016). Metagenomic sequencing with strain-level resolution implicates uropathogenic *E.* coli in necrotizing enterocolitis and mortality in preterm infants. *Cell Rep.* 14, 2912–2924. doi: 10.1016/j.celrep.2016.03.015

White, J. R., Nagarajan, N., and Pop, M. (2009). Statistical methods for detecting differentially abundant features in clinical metagenomic samples. *PLoS Comput. Biol.* 5:e1000352. doi: 10.1371/journal.pcbi.1000352

Wu, G. D., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y. Y., Keilbaugh, S. A., et al. (2011). Linking long-term dietary patterns with gut microbial enterotypes. *Science* 334, 105–108. doi: 10.1126/science.1208344

Xu, L., Paterson, A. D., Turpin, W., and Xu, W. (2015). Assessment and selection of competing models for zero-inflated microbiome data. *PLoS ONE* 10:e0129606. doi: 10.1371/journal.pone.0129606

Zhang, X., Mallick, H., Tang, Z., Zhang, L., Cui, X., Benson, A. K., et al. (2017). Negative binomial mixed models for analyzing microbiome count data. *BMC Bioinformatics* 18:4. doi: 10.1186/s12859-016-1441-7

Zhou, Y., Shan, G., Sodergren, E., Weinstock, G., Walker, W. A., and Gregory, K. E. (2015). Longitudinal analysis of the premature infant intestinal microbiome prior to necrotizing enterocolitis: a case-control study. *PLoS ONE* 10:e0118632. doi: 10.1371/journal.pone.0118632

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.