Published in final edited form as:

Stat (Int Stat Inst). 2018 ; 7(1): . doi:10.1002/sta4.185.

A Zero-inflated Beta-binomial Model for Microbiome Data Analysis

Tao Hu^a, Paul Gallins^a, and Yi-Hui Zhou^{b,*}

^aBioinformatics Research Center, North Carolina State University, NC, 27695

^bDepartment of Biological Sciences and Bioinformatics Research Center, North Carolina State University, NC, 27695

Abstract

The microbiome is increasingly recognized as an important aspect of the health of host species, involved in many biological pathways and processes and potentially useful as health biomarkers. Taking advantage of high-throughput sequencing technologies, modern bacterial microbiome studies are metagenomic, interrogating thousands of taxa simultaneously. Several data analysis frameworks have been proposed for microbiome sequence read count data and determining the most significant features. However, there is still room for improvement. We introduce a zero-inflated beta-binomial (ZIBB) to model the distribution of microbiome count data and to determine association with a continuous or categorical phenotype of interest. The approach can exploit mean-variance relationships to improve power and adjust for covariates. The proposed method is a mixture model with two components: (i) a zero model accounting for excess zeros and (ii) a count model to capture the remaining component by beta-binomial regression, allowing for overdispersion effects. Simulation studies show that our proposed method effectively controls type I error and has higher power than competing methods to detect taxa associated with phenotype. An R package ZIBBSeqDiscovery is available on R CRAN.

Keywords

zero inflated beta binomial modeling; penalized generalized linear model; count data

1. Introduction

The human microbiome consists of the collection of all microbes at sites in or on the human body Cho & Blaser (2012), with 10 times more cells than the human host, and representing 1000-fold greater diversity of genes (Whitman et al., 1998; Consortium et al., 2012). Many studies have shown that microbial communities play an important role in maintaining health for the host species (Bäckhed et al., 2005; Turnbaugh et al., 2009; Clemente et al., 2012; Qin et al., 2012). Deciphering the composition of the microbiome and its association with phenotype (for example, disease status) is therefore critical in understanding biological mechanisms. In this paper, we focus on the bacterial microbiome.

* yihui_zhou@ncsu.edu.

Traditional methods for studying microbiomes have been culture-dependent and low-throughput, as the vast majority of microbes cannot be easily isolated and grown. With modern high-throughput sequencing and a metagenomic approach (Riesenfeld et al., 2004), investigators can now directly sequence the extracted DNA of an entire microbiomal community, revealing extensive diversity. A common sequencing strategy (Consortium et al., 2012) is to sequence only the 16S ribosomal RNA (rRNA) gene, which is highly specific to bacterial species. The 16S rRNA sequences can be grouped into different clusters called Operational Taxonomic Units (OTUs), or taxa. The data are represented by a sequence count matrix, organized here with columns representing samples and rows representing OTUs.

In the past decade, a number of methods have been developed for microbiome count data association analysis. Those methods aim to explore the associations between phenotypes of interest, which include observational outcomes or experimental conditions, and the microbiome counts. We refer to the implementation of these methods as *discovery studies*. In distance-based analysis (McArdle & Anderson, 2001; Lozupone & Knight, 2005; Chen et al., 2012), the basic strategy is to compute the pairwise distance matrix among samples using a specific metric, and then apply downstream analyses to these distances. For analyses by OTUs, a logistic normal multinomial regression model (LNM) (Xia et al., 2013) has been used, assuming the count data came from a multinomial distribution given the underlying bacterial composition, while the compositions are modeled with a logistic normal distribution. MiRKAT (Zhao et al., 2015) is a kernel-based regression model, where microbiome compositions are included in the kernels, and the phenotype is regressed on the kernels and other covariates. A more recent method employs a zero-inflated Gaussian distribution (ZIG) (Paulson et al., 2013) to accommodate the excessive zero counts in the microbiome count data.

Although the recently developed methods have appealing properties, it remains important to conduct basic analyses at the feature (OTU) level. Direct modeling of microbiome count data versus phenotype typically uses generalized linear models (GLMs), which must account for overdispersion (Anders & Huber, 2010; Xu et al., 2015; Weiss et al., 2015), and for which methods developed for RNA sequence (RNA-Seq) data may be appealing. For example, the negative binomial modeling in edgeR (Robinson et al., 2010), originally developed for transcriptomic analysis, has been applied on microbiome data with phyloseq (McMurdie & Holmes, 2014). BBSeq (Zhou et al., 2011b) employs a beta-binomial model for the count data and estimates the overdispersion parameter through a polynomial mean-overdispersion relationship. However, directly employing RNA-Seq data analysis methods may be problematic, because microbiome count data is usually sparse, with many zero counts (Weiss et al., 2015). Accordingly, zero-inflated negative binomial models (ZINB) Fang et al. (2014) have also been proposed, but do not allow for arbitrary covariates.

We summarize the major challenges of analyzing microbiome count data as: 1) microbiome count data has excessive zero counts; 2) structure in the data (Zhou et al., 2011b) should be exploited to improve power and false positive control; 3) a lack of flexible software to conduct the analysis, including covariate control and handling both discrete and continuous phenotypes. Here we propose a framework to handle all of these challenges, based on a zero-inflated beta-binomial model (ZIBB) of association between phenotypes/covariates and

microbiome count values. The model assumes the count data distribution consists of a mixture of a point mass probability at zero, and a beta-binomial distribution (which may have an appreciable additional mass at zero). Our proposed method is an extension of the beta-binomial model of BBSeq (Zhou et al., 2011b). A Wald statistic is used to test statistical significance of the phenotype while accounting for covariates. The proposed framework has the following advantages over canonical methods:

- (i) ZIBB can be envisioned in two stages, whereby each count follows a binomial distribution for a varying success probability that itself is a beta random variable, lending itself to clear interpretation.
- (ii) ZIBB employs logistic regression to model the zero point mass probability as a function of covariates, providing additional flexibility in modeling.
- (iii) Using the same strategy as BBSeq (Zhou et al., 2011b), ZIBB considers a constrained approach to model the over-dispersion as a polynomial function of the systematic (mean) component of the GLM. The constraint follows the observation that sample overdispersion estimates are often strongly correlated with the sample mean for real count data, including microbiome data. Simulations will show that the constrained approach indeed increases the power of our method compared to competing methods.
- (v) An R package, ZIBBSeqDiscovery, is implemented. This package aims to provide a user friendly pipeline to analyzing microbiome count data, with both simple default settings and embedded functions that may be useful for experienced bioinformaticians.

Simulations and real data analysis demonstrate that the proposed ZIBB approach performs well for microbiome data analysis.

2. Methods

2.1. The Need for Alternate Modeling

We illustrate some of the modeling aspects by showing behavior of three different datasets. The dataset from Zhou et al. (2011a) comes from measurements of soil microbial communities. After data cleaning, a total of 56 samples and 16825 OTUs are reported in this dataset. Analysis of the dataset is presented further below, but here we focus on only the relationship of sample variance vs. same mean across OTUs (Figure 1). The figure illustrates a strong relationship, including overdispersion (i.e., the variance exceeds the mean), a pattern that is observed in every microbiome dataset we have encountered.

Microbiome sequencing count data also typically contain a high proportion of zeros. The occurrence of zeros is not necessarily “zero-inflation”, as some zeros would be expected to be encountered for rare microbiome taxa (i.e. with low mean). Using the kostic dataset (Kostic et al., 2012) (so named for the first author) of $n = 95$ tumor vs. $n = 95$ normal colon microbiota, we demonstrate that the zero counts are in excess of that predicted by simple overdispersion modeling. We started with a standard likelihood ratio test (LRT) of two nested negative binomial models applied to each OTU, using $\log(\text{library size})$ as a covariate,

and including zero-inflation in the larger model as described further below, compared to a null model with no zero inflation. P-values for the likelihood ratio test acknowledge a boundary condition under the null, so the asymptotic null distribution is a mixture of a point mass at zero and X_1^2 (Self & Liang, 1987). We performed the test within each of tumor vs. normal separately, after dropping OTUs that are all zeros within each subgroup, yielding about 2000 OTUs per analysis. After computing the likelihood ratio p-values, we used the R qvalue package with default settings to estimate the proportion of alternatives (i.e. requiring zero-inflation) to be 72% for each of tumor and normal. In other words, zero inflation is pervasive, and must be considered for accurate modeling.

The two major improvements of our proposed framework are 1) employing zero-inflated modeling and 2) considering a constrained approach to estimate the over-dispersion parameters. To illustrate how these improvements appear at the level of an individual OTU/taxon, we display the count data of *Lactobacillus vaginalis* from the ravel vaginal microbiome dataset (Ravel et al., 2013). Figure 2(a) shows a histogram of the original data for taxon *Lactobacillus vaginalis* (values over 30 are truncated), while the remaining panels redisplay the model fits in terms of expected counts. Panels (b)-(d) show the predicted distributions from fitted models Poisson, zero-inflated Poisson, and zero-inflated negative binomial. Panels (e) and (f) show the distributions under our proposed methods of a zero-inflated beta-binomial and a “constrained” zero-inflated beta-binomial as we will describe below. The zero-inflated models evidently provide a better general fit. Below we conduct more simulations and real data analysis to explore the performance of our proposed methods.

2.2. Notation and models

2.2.1. A zero-inflated model for discovery analysis, with a mean-variance relationship.—We assume the count data is an $m \times n$ matrix arising from 16S rRNA gene profiling or other sequencing strategy. Let $Y = (y_{ij}) \in \mathbb{Z}^{m \times n}$ be the count matrix, and each element y_{ij} represents the count of OTU i in sample j ($i = 1, \dots, m, j = 1, \dots, n$). Let

$S_j = \sum_{i=1}^m y_{ij}$ be the library size for sample j .

The canonical beta-binomial model assumes that y_{ij} follows a two-stage model. Namely, y_{ij} follows a binomial distribution $\text{Bin}(S_j, \mu_{ij})$, and $\mu_{ij} \sim \text{Beta}(a_{1ij}, a_{2ij})$. Let $f_{\text{count}}(\cdot | a_{1ij}, a_{2ij})$ be the probability mass function (pmf) of the beta-binomial distribution with parameter (S_j, a_{1ij}, a_{2ij}) , or

$$f_{\text{count}}(y_{ij} | a_{1ij}, a_{2ij}) = \binom{S_j}{y_{ij}} \frac{B(y_{ij} + a_{1ij}, S_j - y_{ij} + a_{2ij})}{B(a_{1ij}, a_{2ij})}, \quad (1)$$

where the dependence on S_j is implied. Because $\mu_{ij} \sim \text{Beta}(a_{1ij}, a_{2ij})$, we have $E(\mu_{ij}) = a_{1ij} / (a_{1ij} + a_{2ij})$. Noting that the variance of $\text{Beta}(a_{1ij}, a_{2ij})$ is $\phi_i E(\mu_{ij})(1 - E(\mu_{ij}))$, it is trivial to solve that $a_{1ij} = E(\mu_{ij})(1 - \phi_i) / \phi_i$, and $a_{2ij} = (1 - E(\mu_{ij}))(1 - \phi_i) / \phi_i$. Therefore, the beta distribution using parameter (a_{1ij}, a_{2ij}) can be reparameterized with $(E(\mu_{ij}), \phi_i)$. The

reparameterized form has a clear interpretation: $\phi_i = 0$ is the overdispersion parameter. In the rest of this paper, we will focus on this reparameterized form.

As we have seen, an important characteristic of the data is the presence of excessive zeros, *i.e.* larger than predicted from a standard count model, even with overdispersion. We propose a zero-inflated beta-binomial (ZIBB) model to account for this zero inflation. The zero-inflated beta-binomial model assumes that the density of count y_{ij} is a mixture of a point mass at zero (zero model) and a beta-binomial distribution (count model). Let the density of y_{ij} be

$$f(y_{ij} | \alpha_{1ij}, \alpha_{2ij}, \pi_{ij}) = \pi_{ij} I_{y_{ij}=0} + (1 - \pi_{ij}) f_{\text{count}}(y_{ij} | \alpha_{1ij}, \alpha_{2ij}), \quad (2)$$

where π_{ij} is the point mass at zero, $f_{\text{count}}(\cdot)$ is the pmf of the canonical beta-binomial distribution as in (1). To include the effects of phenotype and covariates in our modeling, we use the following link functions:

- (i) For the *zero model*, we model π_{ij} as

$$\text{logit}(\pi_{ij}) = \log\left(\frac{\pi_{ij}}{1 - \pi_{ij}}\right) = z_j^T \eta_i, \quad (3)$$

where $z_j = (z_{0,j}, \dots, z_{q-1,j})^T \in \mathbb{R}^q$ is the vector of zero-inflation related covariates (including the intercept, thus $z_{0,j} = 1$) for sample j , and $\eta_i = (\eta_{0,i}, \dots, \eta_{q-1,i})^T \in \mathbb{R}^q$ is the vector of corresponding coefficients for OTU i . An example for the choice of zero-inflation-related covariates is $z_j^T = (1, \log S_j)$ if $q = 2$. We denote $Z = (z_1, \dots, z_n)^T \in \mathbb{R}^{n \times q}$, and refer to Z as the design matrix for the zero model.

- (ii) For the *count model*, we model $E(\mu_{ij})$ as

$$\text{logit}(E(\mu_{ij})) = \log\left(\frac{E(\mu_{ij})}{1 - E(\mu_{ij})}\right) = x_j^T \beta_i, \quad (4)$$

where $x_j = (x_{0,j}, \dots, x_{p-1,j})^T \in \mathbb{R}^p$ is the vector of phenotypes of interest (the design matrix includes the intercept, thus $x_{0,j} = 1$) for sample j , and $\beta_i = (\beta_{0,i}, \dots, \beta_{p-1,i})^T \in \mathbb{R}^p$ is the vector of corresponding coefficients for OTU i . We denote $X = (x_1, \dots, x_n)^T \in \mathbb{R}^{n \times p}$, and refer to X as the design matrix for the count model.

As we have seen, there often exists a strong correlation between the mean and variance, as well as between the mean and overdispersion. To model this relationship, we use the following polynomial fit as the constraint between overdispersion parameter ϕ_i and coefficients β_i in the count model

$$\text{logit}(\phi_i) = \sum_{k=0}^k \gamma_k \{\text{mean}(X\beta_i)\}^k. \quad (5)$$

In practice, $K = 3$ is typically sufficient to effectively model the relationship. To distinguish this modeling for mean overdispersion relationship from fitting the relationship for each OTU separately, we call this model and associated estimation approach the *constrained* model. Otherwise, if ϕ_i is estimated separately for each i , it is called the *free* model.

For a small proportion of OTUs, the ZIBB methods may fail to converge. This typically occurs when OTUs have few (e.g. four or fewer) nonzero counts, and for such OTUs the power to detect association with experimental variables is small, and we filter out these OTUs. For the remaining small number of OTUs which fail to converge (usually 2% or fewer), we substitute p-values from the MCC package, which mimics permutation trend-testing (Zhou & Wright, 2015).

2.2.2. Parameter estimation—The unknown parameters are $\{\eta_i, \beta_i, \phi_i\}_{i=1,\dots,m}$ for free modeling, and $\{\eta_i, \beta_i, \gamma_k\}_{i=1,\dots,m, k=1,\dots,K}$ for constrained modeling (ϕ_i 's being implied by γ_k 's). Parameter estimation for free modeling is handled by maximum likelihood, and in our modeling, the parameters for different OTUs are separate. Thus, we can treat OTUs as separate for parameter estimation, and a parallel computing strategy can therefore be applied to accommodate large number of OTUs in modern datasets. Within each OTU, the number of unknown parameters is small, so problematic issues of high dimensional testing (number of features exceeding the sample size) are avoided.

To estimate $\{\gamma_k\}_{k=1,\dots,K}$ in constrained modeling, we use the estimates of β_i and ϕ_i from free modeling and fit the polynomial model according to Equation (5), and use least squares solutions to estimate the γ_k 's. Estimations for other parameters are the same as in free modeling. However, the constrained modeling technically ties together the OTUs, so that the likelihood approach for estimation may be viewed as a form of composite likelihood (Lindsay, 1988).

2.2.3. Testing procedure—The main purpose of a discovery study is to test associations of OTUs with a phenotype of interest, which is included in the design matrix X . Below, for simplicity, we assume only one phenotype is included in X . Thus, we $p = 2$ in this case. Our primary interest is to test the statistical significance of β_{1i} for each OTU i , $i = 1, \dots, m$, and we use a Wald statistic $\hat{\beta}_{1i}/SE(\hat{\beta}_{1i})$ for testing. In real microbiome count datasets, the included phenotype can either be discrete (for example, group indicator), or continuous (for example, body mass index values). For the discrete case, we use a t distribution with degrees of freedom $n - 2$ to approximate the distribution of the Wald statistic under the null hypothesis. For the continuous case, we approximate the Wald statistic's null distribution by a standard normal.

3. Simulations

3.1. Type I error and power

This simulation compared the performances of various statistical methods for detecting association between microbiome composition and experimental phenotypes. Two types of phenotypes were considered: a discrete phenotype and a continuous phenotype, with no additional covariates. Thus, the effect coefficient β_i for OTU i was a vector with length $p = 2$, $\beta_i = (\beta_{0i}, \beta_{1i})^T$, where β_{0i} corresponded to the intercept. Thus we performed the test with $H_0 : \beta_{1i} = 0$. For the discrete case, our proposed ZIBB methods were compared with BBSseq, ZINB and edgeR. For the continuous case, only edgeR was chosen because other candidates were designed for handling discrete phenotypes. In summary, this simulation compares the type I errors and powers among competing methods for testing $H_0 : \beta_{1i} = 0$.

In this simulation, we try to simulate data which mimics real data as much as possible. The simulations were based on analyses of the kostic microbiome dataset, with 2505 OTUs and 190 samples (Kostic et al., 2012). Although the original data were paired, the pairing was not considered relevant for these analyses, which were designed merely to use the realistic counts as a basis for phenotype simulation. To obtain reasonable parameter values, we fit the kostic data with the free and constrained ZIBB models, and treated the estimated $\hat{\beta}_{0i}$ and $\hat{\gamma}_k$ as true parameter values. For a randomly chosen 10% of OTUs, we specified an effect size $r = e^{\beta_{1i}}(1 + e^{\beta_{0i}})/(1 + e^{\beta_{0i}} + e^{\beta_{1i}})$ to determine β_{1i} given β_{0i} and the specified r . The remaining 90% (Macklaim et al., 2013; Paulson et al., 2013; Zhou et al., 2011b) of the OTUs had β_{1i} 's chosen to be zero.

The generation procedures for the design matrix X were different for the discrete and continuous cases. For the discrete case, we assumed that the phenotype was an experimental group indicator, and X is a n -by-2 matrix. The first column of X were 1's which correspond to intercept, and the second column indicators for the two groups, of sizes n_1 and n_2 . For the continuous case, we assume the phenotype can be any real values. To generate the continuous phenotype, we first drew a random subsample Y' of the kostic data with n samples, row-standardized the resulting subsample and obtained the first principal component of the subsample. The phenotype was then set as 0.05 times the first principal component, plus a standard normal random noise term.

Next, the overdispersion parameters ϕ_i 's were generated according to Equation (5). The count data Y were then simulated based on the beta-binomial distribution (1). To add a zero inflation effect, we fit a logistic regression model between the proportion of zeros of simulated count data and corresponding library size according to Equation (3), and then treated predicted $\hat{\pi}_{ij}$'s as the true point mass at zero. Finally, the simulated beta-binomial count data was set to zero with probability $\hat{\pi}_{ij}$. Any OTUs with zero counts across all samples were removed. In the discrete case, OTUs with zero counts across any one of groups were also removed.

For each value of effect size r , 100 datasets were generated using the procedures above. Each dataset included 3000 OTUs.

3.2. Standard error accuracy

Because the test statistic we use is a Wald statistic, the standard deviation $SE(\hat{\beta}_{1i})$ of the estimated parameter $\hat{\beta}_{1i}$ is critical to obtain accurate corresponding p -values. To check the accuracy of $SE(\hat{\beta}_{1i})$ by our proposed ZIBB method, we compared it with the one calculated using a computationally intensive bootstrap strategy. We emphasize that the bootstrap estimates, where both X and Y vary, are conditional on the data in a somewhat different manner than the likelihood-based approach, and so are not expected to match precisely.

Using the same data generation strategy as in Section 3.1, we are able to obtain a simulated microbiome count dataset (Y, X, Z) with m OTUs and n samples. Applying ZIBB method on this dataset, we can calculate corresponding standard deviations $\{SE(\hat{\beta}_{1i})\}_{i=1,\dots,m}$ for parameter β_{1i} . Then, we construct $B = 100$ bootstrap datasets $\{(Y^{(b)}, X^{(b)}, Z^{(b)})\}_{b=1,\dots,B}$. Each bootstrap dataset $(Y^{(b)}, X^{(b)}, Z^{(b)})$ is generated by re-sampling n samples with replacement from the original dataset (Y, X, Z) . For each bootstrap dataset, we calculate the estimated parameter $\{\hat{\beta}_{1i}^{(b)}\}_{i=1,\dots,m}$. For a specific OTU i , the standard deviation of estimated $\hat{\beta}_{1i}^{(b)}$ across all bootstrap datasets is denoted it as $SE(\hat{\beta}_{1i}^B)$. We then compare the standard deviations $\{SE(\hat{\beta}_{1i})\}_{i=1,\dots,m}$ by the ZIBB method to the standard deviations $\{SE(\hat{\beta}_{1i}^B)\}_{i=1,\dots,m}$ by bootstrapping.

We assessed the accuracy of estimated standard deviations by using both the free modeling approach and constrained modeling approach. For the discrete case (*i.e.*, phenotype is discrete), the sample size was $n_1 = 50$ versus $n_2 = 50$. For the continuous case, the sample size is $n = 100$. The number of OTUs is $m = 3000$. To determine the true value for β_{1i} , three effect size r values were chosen, $r = \{0.5, 1, 2\}$.

4. Results

4.1. Type I errors

The type I errors for the different methods and sample sizes are listed in Table 1 (discrete case) and Table 2 (continuous). For the discrete scenario, free modeling of ZIBB has inflated type I error, while constrained modeling of ZIBB is able to control the type I error well. We observe the same patterns for BBSeq, and further indicating that adding constraints on the overdispersion is helpful. For both approaches under the ZIBB modeling, the type I errors are slightly increasing as the sample size increases. The type I errors for ZINB are slightly inflated, while edgeR is conservative. For the continuous case, again the constrained ZIBB approach controls the type I error the best among the methods, although they are slightly inflated compared to nominal. For edgeR, the type I error is about the same (a little bit larger) as constrained modeling of ZIBB at a small sample size, and is anticonservative and similar to free modeling of ZIBB at larger sample sizes.

4.2. Power

We plot the power versus effective size r for different methods in Figure 3. For the discrete scenarios, panel (a) shows the power plot for the discrete case when the sample sizes are $n_1 = 15$ versus $n_2 = 30$, and panel (b) $n_1 = 30$ versus $n_2 = 30$. Results for $n_1 = n_2 = 50$ are similar (not shown). Panels (c) and (d) show the power plots for the continuous scenarios with $n = 20$ and $n = 60$. In all scenarios, constrained modeling for ZIBB performs the best, as expected. When the sample size increases, the differences between constrained modeling for ZIBB and other competing methods become smaller, which we attribute to the fact that for large sample sizes, constrained modeling does not offer a benefit, as overdispersion may be well-estimated directly per OTU.

We also observe differing patterns in discrete or continuous cases. For the discrete case, ZINB and free ZIBB perform similarly. The constrained BBSeq performs better than edgeR and converge under larger effect sizes. Free BBSeq performs the worst at all three sample sizes. For free BBSeq, it neither considers the mean-overdispersion relationship, nor takes the excessive zero counts into consideration (*i.e.*, does not include zero inflation portion), and it may not be surprising that it performs the worst. For the continuous case, free ZIBB performs better than edgeR under smaller effect sizes while edgeR performs better than free ZIBB under larger effect sizes.

4.2.1. Standard Error Accuracy—In the Supplement, we plot the estimated standard deviation of β_{ji} versus the standard deviation of β_{ji} obtained by bootstrapping strategy for both the discrete and continuous scenarios. We checked the results obtained by free modeling approach and constrained modeling approach, and varied the true value of β_{ji} at the three different levels. In general, the bootstrap standard deviations show good correlation with our proposed ZIBB likelihood method, and the relationship is better for larger sample sizes. The results for continuous case follow a similar pattern. We emphasize that the type I error for ZIBB was shown earlier to be accurate.

5. Real Data Analysis and Discussion

Finally, we analyzed the dataset from Gevers, which compared 352 mucosal tissue biopsies (terminal ileum and rectum) from pediatric Crohn's disease (CD) cases (pre-treatment) with 212 control samples, with 4218 OTUs. The authors had found several taxa associated with CD, and we applied the ZIBB methods, edgeR, BBSeq (free and constrained) and ZINB. Table 3 shows the results for the top 10 taxa using each method, and the overlap for the top findings is considerable across the methods. For real data, it is difficult to know true underlying state of nature. We reasoned that the methods could be compared in terms of consistency of ranking taxa at a higher level (genera) than individual OTUs, in comparisons that are similar in spirit to pathway analysis. A total of 58 genera were represented by at least 5 taxa, and for each genus and each method we compared the p -value ranks for OTUs in the genus to the remaining OTUs using a one-sided Wilcoxon rank-sum test for an enrichment p -value for the genus. Then across the 58 genera we computed the number with Benjamini-Hochberg FDR $q < 0.01$ for enrichment. The result is shown in Figure 4, showing that ZIBB constrained showed the greatest number of significantly enriched genera,

followed by the two BBSeq models. We conclude that our proposed ZIBB framework performs well in real data analysis, and simulations and real analysis suggested that both zero inflation and the constraint (5) are vital for accurate and powerful analysis.

In this paper, we have described a zero-inflated beta-binomial model (ZIBB) for the distribution of microbiome count data. Simulations indicate that ZIBB modeling with the constrained approach is preferred among several competing methods. We created an R package ZIBBSeqDiscovery on R CRAN. ZIBBSeqDiscovery provides a user friendly pipeline to analyze microbiome count data.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

- Anders S & Huber W (2010), 'Differential expression analysis for sequence count data,' *Genome Biology*, 11(R106).
- Bäckhed F, Ley RE, Sonnenburg JL, Peterson DA & Gordon JI (2005), 'Host-bacterial mutualism in the human intestine,' *science*, 307(5717), pp. 1915–1920. [PubMed: 15790844]
- Chen J, Bittinger K, Charlson ES, Hoffmann C, Lewis J, Wu GD, Collman RG, Bushman FD & Li H (2012), 'Associating microbiome composition with environmental covariates using generalized unifracs distances,' *Bioinformatics*, 28(16), pp. 2106–2113. [PubMed: 22711789]
- Cho I & Blaser MJ (2012), 'The human microbiome: at the interface of health and disease,' *Nature Reviews Genetics*, 13(4), pp. 260–270.
- Clemente JC, Ursell LK, Parfrey LW & Knight R (2012), 'The impact of the gut microbiota on human health: an integrative view,' *Cell*, 148(6), pp. 1258–1270. [PubMed: 22424233]
- Consortium HMP et al. (2012), 'Structure, function and diversity of the healthy human microbiome,' *Nature*, 486(7402), pp. 207–214. [PubMed: 22699609]
- Fang R, Wagner B, Harris JK & Fillon SA (2014), 'Application of zero-inflated negative binomial mixed model to human microbiota sequence data,' *Tech. rep., PeerJ PrePrints*.
- Kostic AD, Gevers D, Pedamallu CS, Michaud M, Duke F, Earl AM, Ojesina AI, Jung J, Bass AJ, Tabernero J et al. (2012), 'Genomic analysis identifies association of fusobacterium with colorectal carcinoma,' *Genome research*, 22(2), pp. 292–298. [PubMed: 22009990]
- Lindsay BG (1988), 'Composite likelihood methods,' *Contemporary mathematics*, 80(1), pp. 220–239.
- Lozupone C & Knight R (2005), 'Unifrac: a new phylogenetic method for comparing microbial communities,' *Applied and environmental microbiology*, 71(12), pp. 8228–8235. [PubMed: 16332807]
- Macklaim JM, Fernandes AD, Di Bella JM, Hammond JA, Reid G & Gloor GB (2013), 'Comparative meta-rna-seq of the vaginal microbiota and differential expression by lactobacillus iners in health and dysbiosis,' *Microbiome*, 1(1), p. 1. [PubMed: 24467924]
- McArdle BH & Anderson MJ (2001), 'Fitting multivariate models to community data: a comment on distance-based redundancy analysis,' *Ecology*, 82(1), pp. 290–297.
- McMurdie PJ & Holmes S (2014), 'Waste not, want not: why rarefying microbiome data is inadmissible,' *PLoS Comput Biol*, 10(4), p. e1003531. [PubMed: 24699258]
- Paulson JN, Stine OC, Bravo HC & Pop M (2013), 'Differential abundance analysis for microbial marker-gene surveys,' *Nature methods*, 10(12), pp. 1200–1202. [PubMed: 24076764]
- Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F, Liang S, Zhang W, Guan Y, Shen D et al. (2012), 'A metagenome-wide association study of gut microbiota in type 2 diabetes,' *Nature*, 490(7418), pp. 55–60. [PubMed: 23023125]

- Ravel J, Brotman RM, Gajer P, Ma B, Nandy M, Fadrosch DW, Sakamoto J, Koenig SS, Fu L, Zhou X et al. (2013), 'Daily temporal dynamics of vaginal microbiota before, during and after episodes of bacterial vaginosis,' *Microbiome*, 1(1), p. 1. [PubMed: 24467924]
- Riesenfeld CS, Schloss PD & Handelsman J (2004), 'Metagenomics: genomic analysis of microbial communities,' *Annu. Rev. Genet.*, 38, pp. 525–552. [PubMed: 15568985]
- Robinson MD, McCarthy DJ & Smyth GK (2010), 'edgeR: a bioconductor package for differential expression analysis of digital gene expression data,' *Bioinformatics*, 26(1), pp. 139–140. [PubMed: 19910308]
- Self SG & Liang KY (1987), 'Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions,' *Journal of the American Statistical Association*, 82(398), pp. 605–610.
- Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, Sogin ML, Jones WJ, Roe BA, Affourtit JP et al. (2009), 'A core gut microbiome in obese and lean twins,' *nature*, 457(7228), pp. 480–484. [PubMed: 19043404]
- Weiss SJ, Xu Z, Amir A, Peddada S, Bittinger K, Gonzalez A, Lozupone C, Zaneveld JR, Vazquez-Baeza Y, Birmingham A et al. (2015), 'Effects of library size variance, sparsity, and compositionality on the analysis of microbiome data,' *Tech. rep.*, PeerJ PrePrints.
- Whitman WB, Coleman DC & Wiebe WJ (1998), 'Prokaryotes: the unseen majority,' *Proceedings of the National Academy of Sciences*, 95(12), pp. 6578–6583.
- Xia F, Chen J, Fung WK & Li H (2013), 'A logistic normal multinomial regression model for microbiome compositional data analysis,' *Biometrics*, 69(4), pp. 1053–1063. [PubMed: 24128059]
- Xu L, Paterson AD, Turpin W & Xu W (2015), 'Assessment and selection of competing models for zero-inflated microbiome data,' *PloS one*, 10(7), p. e0129606. [PubMed: 26148172]
- Zhao N, Chen J, Carroll IM, Ringel-Kulka T, Epstein MP, Zhou H, Zhou JJ, Ringel Y, Li H & Wu MC (2015), 'Testing in microbiome-profiling studies with mirkat, the microbiome regression-based kernel association test,' *The American Journal of Human Genetics*, 96(5), pp. 797–807. [PubMed: 25957468]
- Zhou J, Wu L, Deng Y, Zhi X, Jiang YH, Tu Q, Xie J, Van Nostrand JD, He Z & Yang Y (2011a), 'Reproducibility and quantitation of amplicon sequencing-based detection,' *The ISMEjournal*, 5(8), pp. 1303–1313.
- Zhou YH & Wright FA (2015), 'Hypothesis testing at the extremes: fast and robust association for high-throughput data,' *Biostatistics*, 16(3), pp. 611–625, doi:10.1093/biostatistics/kxv007. [PubMed: 25792622]
- Zhou YH, Xia K & Wright FA (2011b), 'A powerful and flexible approach to the analysis of rna sequence count data,' *Bioinformatics*, 27(19), pp. 2672–2678. [PubMed: 21810900]

Figure 1.
Mean-Variance plot on the log scale of the dataset from Zhou et al. (2011a).

Figure 2.

Fittings of microbiome count data for taxa *Lactobacillus.vaginalis* under different models in “ravel” dataset. Panel (a) is the histogram for original data. Panel (b)-(f) are histograms of fittings under corresponding models. Counts that exceed 30 are truncated at 30 for display purposes.

Figure 3.

(a) Powers for discrete case: $n_1 = 15$ vs $n_2 = 30$. (b) Powers for discrete case: $n_1 = 30$ vs $n_2 = 30$. (c) Powers for continuous case: $n = 20$. (d) Powers for continuous case: $n = 60$.

Figure 4. For the Gevers data, number of significantly over-represented genera (FDR $q < 0.01$) from among 58 genera when ordered by taxon-level p-values, for the various analysis approaches

Table 1.

Type I errors for the discrete case at level $\alpha = 0.05$. “cstr” is an abbreviation for constrained modeling.

Sample size	ZIBB.free	ZIBB.cstr	BBSeq.free	BBSeq.cstr	ZINB	edgeR
15 vs 30	0.05381039	0.04075553	0.05532779	0.02620762	0.06062154	0.02792048
30 vs 30	0.06030656	0.04036705	0.06089566	0.02821479	0.0507933	0.0287378
50 vs 50	0.07498005	0.03851371	0.06413183	0.02709413	0.05120469	0.03098385

Table 2.

Type I errors for the continuous case at level $\alpha = 0.05$.

Sample size	ZIBB free	ZIBB constrained	edgeR
20	0.09262155	0.06105356	0.067535
60	0.09250315	0.05443886	0.09042197

BBSseq cstr	ZINB	edgeR	p-value
OTU	OTU	OTU	p-value
Clostridiales/Lachnospiraceae/Roseburia	Fusobacteriales/Fusobacteriaceae/Fusobacterium	Fusobacteriales/Fusobacteriaceae/Fusobacterium	6.36e-51(1.34e-47)
Clostridiales/Ruminococcaceae/Faecalibacterium	Clostridiales/Lachnospiraceae/Coproccoccus	Bacteroidales/Bacteroidaceae/Bacteroides	3.04e-45(3.20e-42)
Bacteroidales/Bacteroidaceae/Bacteroides	Clostridiales/Lachnospiraceae/Roseburia	Bacteroidales/Prevotellaceae/Prevotella	3.61e-39(2.54e-36)
Pasteurellales/Pasteurellaceae/Haemophilus	Clostridiales/Lachnospiraceae/Blautia	Bacteroidales/Porphyromonadaceae/Porphyromonas	5.10e-39(3.07e-36)
Bacteroidales/Bacteroidaceae/Bacteroides	Fusobacteriales/Fusobacteriaceae/Fusobacterium	Bacteroidales/Porphyromonadaceae/Parabacteroides	1.38e-35(4.15e-33)
Pasteurellales/Pasteurellaceae/Aggregatibacter	Clostridiales/Lachnospiraceae/Roseburia	Clostridiales/Lachnospiraceae/Roseburia	1.53e-35(4.29e-33)
Clostridiales/Lachnospiraceae/Roseburia	Clostridiales/Lachnospiraceae/Roseburia	Campylobacteriales/Helicobacteraceae/Helicobacter	4.16e-35(1.10e-32)
Clostridiales/Veillonellaceae/Dialister	Pasteurellales/Pasteurellaceae/Haemophilus	Lactobacillales/Leuconostocaceae/Leuconostoc	1.05e-34(2.61e-32)
Turicibacteriales/Turicibacteraceae/Turicibacter	Pasteurellales/Pasteurellaceae/Aggregatibacter	Lactobacillales/Lactobacillaceae/Lactobacillus	2.86e-34(6.43e-32)
Enterobacteriales/Enterobacteriaceae/Morganella	Pasteurellales/Pasteurellaceae/Haemophilus	Bacteroidales/Prevotellaceae/Prevotella	2.90e-34(6.43e-32)

Author manuscript; available in PMC 2019 June 19.