JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 23, Number 2, 2016

Pp. 102–110
DOI: 10.1089/cmb.2015.0157

Zero-Inflated Beta Regression for Differential Abundance Analysis with Metagenomics Data

XIAOLING PENG,[1] GANG LI,[2] and ZHENQIU LIU[3]

ABSTRACT

Metagenomics data have been growing rapidly due to the advances in NGS technologies. One goal of human microbial studies is to detect abundance differences across clinical conditions. Besides small sample size and high dimension, metagenomics data are usually represented as compositions (proportions) with a large number of zeros and skewed distribution. Efficient tools for handling such compositional data need to be developed.

We propose a zero-inflated beta regression approach (ZIBSeq) for identifying differentially abundant features between multiple clinical conditions. The proposed method takes the sparse nature of metagenomics data into account and handle the compositional data efficiently. Compared with other available methods, the proposed approach demonstrates better performance with large AUC values for most simulation studies. When applied to a human metagenomics data, it also identifies biologically important taxa reported from previous studies. The software in R is available upon request from the first author.

Key words: algorithms, graphs and networks, machine learning, metagenomics, statistical models.

1. INTRODUCTION

LARGE VOLUMES OF METAGENOMIC SEQUENCING DATA have been generated in the last several years, due to the advances of next-generation sequencing technologies (Gilbert et al., 2011). Those shot-sequence reads are then clustered into operational taxonomic units (OTUs), or annotated against known taxonomic databases of reference sequences (Ghodsi et al., 2011; Wang et al., 2007). The resulting metagenomics read counts are then used for disease association studies (Liu et al., 2011, 2014). Microbiota have been known to be associated with various diseases including Crohn's disease, bacterial vaginosis, obesity, diabetes, and cancer (Morgan et al., 2012; Turnbaugh et al., 2009). It is critical to identify disease-associated pathogenic bacteria characterized by abundance differences across different clinical conditions.

There have been several methods developed for differential abundance analysis with RNA-seq data and gene expression analysis. Those analytical tools including edgeR and DESeq (Robinson et al., 2010; Anders and Huber, 2010; Li et al., 2012; Young et al., 2012; Mortazavi et al., 2008) may be applicable to metagenomic data analysis, because both data are from sequencing-based technologies.

[1]Division of Science and Technology, Beijing Normal University - Hong Kong Baptist University United International College, Zhuhai, China.
[2]Department of Biostatistics, School of Public Health, University of California at Los Angeles, Los Angeles, California.
[3]Samuel Oschin Comprehensive Cancer Institute, Cedars-Sinai Medical Center, Los Angeles, California.

However, metagenomics data have specific characteristics that need to be considered. Compared to RNA-seq data, metagenomic data are more sparse, with many zeros. In addition, metagenomics data need to be preprocessed into proportion (compositions) for adjusting biases in read-depth across different samples. Therefore, specific tools for dealing with such compositional data have to be developed. Recently, the problem of assessing differential expression in sparse high-throughput microbial marker-gene survey data has been addressed by introducing a cumulative-sum scaling (CSS) and zero-inflated Gaussian (ZIG) model (Paulson et al., 2013). R package ''metagenomeSeq'' was developed for the implementation of the proposed approach. It was shown that metagenomeSeq outperforms other tools currently used in this field. However, metagenomeSeq did not handle the compositional data directly. The data was first normalized with cumulative sum scaling, and then ZIG model was applied. The normalized data is sometimes hard to explain.

To overcome the limitations of existing methods for metagenomics analysis, we propose a zero-inflated beta regression approach to handle the proportion data directly. Beta regression is an extension of the generalized linear model (GLM) approach with the assumption that proportional dependent variable can be characterized by the beta distribution. Beta distribution is well known to be flexible for modeling proportional data (Johnson et al., 1993; Ospina et al., 2012). Zero-inflated beta regression incorporates the existing beta distribution with a degenerate distribution, allowing for modeling metagenomics compositional data efficiently (Stasinopoulos and Rigby, 2007). Under various simulations, we show that ZIBSeq performs well for all these types of distributions, especially that it significantly outperforms other methods for features with sparse counts. We also demonstrate the utility of our method on real human metegenomics data. Biologically significant taxa are identified with ZIBSeq. ZIBSeq can be directly applied to the comparison of more than two clinical conditions or continuous outcome variables without difficulty.

2. METHODS

We seek to construct an efficient method that takes the compositional nature of the data into account, deals with sparsity efficiently, and is flexible enough to perform well in various conditions.

Data Structure: In a metagenomic data analysis, suppose n samples from two or more classes are collected and each sample measures expression levels of p genes; the data can be written as an $n \times p$ matrix C for which entry c_{ij} is the number of reads from sample i that mapped to taxa j. Since each sample may generate, a different total number of reads, count c_{ij} depends not only on the reads of taxa j but also on the total number of reads of sample $i (T_i = \sum_{j=1}^{p} c_{ij})$. We also denote the outcome measurement associated with sample i by y_i, $i = 1, \ldots, n$. The outcome variable could be an indictor variable indicating the class of the sample. A typical metagenomic data set can be described as follows (Table 1):

Data Preprocessing: *(i) Normalization.* It is very common to observe different levels of sampling across multiple individuals. To ensure observed counts are comparable across samples, various normalization approaches had been developed for metagenomics data analysis. In our method, we use a natural normalization by converting the raw abundance measure to a proportion representing the relative contribution of each feature to each of the samples. For sample $i, i = 1, \ldots, n$, the normalized feature abundance $x_i^{(j)}, j = 1, \ldots, p$ is the proportion of feature j reads in sample i that can be calculated by $x_i^{(j)} = c_{ij}/T_i$. We chose this simple normalization procedure because it provides a natural representation of the count data as a relative abundance measure. After the normalization, $x_i^{(j)}$ ranges between 0 and 1.

(ii) Transformation. When the distributions of the proportion are extremely left skewed, that is, most of the nonzero proportions are very small, the assumption of beta distribution may not be satisfied. In such

Table 1. Megagenomics Data Structure

	Feature 1	*Feature 2*	$\cdots$	*Feature* p	*Total*	*Outcome*
Sample 1	c_{11}	c_{12}	$\cdots$	c_{1p}	T_1	y_1
Sample 2	c_{21}	c_{22}	$\cdots$	c_{2p}	T_2	y_2
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
Sample n	c_{n1}	c_{n2}	$\cdots$	c_{np}	T_n	y_n

cases, some appropriate transformations such as square root transformation $\sqrt{x}$ or cube root transformation $\sqrt[3]{x}$ are suggested. After the transformation, the proportions still range between 0 and 1 but with a distribution better fitting a beta distribution.

2.1. Differential abundance analysis via zero-inflated beta regression

Since beta distribution has a wide range of different shapes depending on the values of two parameters, beta regression models (Ferrari and Cribari-Neto, 2004) are very useful when the response variables are continuous and restricted to the interval (0,1). Assume that x follows a beta distribution denoted as $x \sim \text{beta}(\mu,\phi)$, where $\mu(0 < \mu < 1)$ is the mean and ϕ ($\phi > 0$) is a precision parameter. The beta density then can be described as a function of μ and ϕ (Ferrari and Cribari-Neto, 2004):

$$f(\xi;\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)}\xi^{\mu\phi-1}(1-t)^{(1-\mu)\phi-1}, 0 < \xi < 1 \quad (1)$$

where $\Gamma(\cdot)$ is the gamma function. Under this parameterization,

$$\begin{cases} E(x) = \mu \\ Var(x) = \mu(1-\mu)/(\phi+1) \end{cases}$$

To deal with the case when interested variable contains zero or one, a more general class of zero-or-one inflated beta regression models for continuous proportions was recently proposed by Ospina and Ferrari (2012). Since zero instead of one is very frequently observed in metagenomics data, in this article, we only consider the zero-inflated beta regression, which assumes the response variable has a mixed continuous-discrete distribution with probability mass at zero. For zero-inflated beta distribution, a new parameter α is added to account for the probability of observations at zero. The subsequent mixture density is:

$$Bi(x;\alpha,\mu,\phi) = \begin{cases} \alpha & \text{if } x = 0 \\ (1-\alpha)f(x;\mu,\phi) & \text{if } 0 < x < 1 \end{cases} \quad (2)$$

where $f(x;\mu,\phi)$ is the beta density (1) and α is the probability of observing zero.

Let $x_i^{(j)}$ denote the normalized feature abundance, that is, the proportion of feature j reads in sample i, which is calculated by $x_i^{(j)} = c_{ij}/T_i$. Then $x_1^{(j)}, x_2^{(j)}, \ldots, x_n^{(j)}$ are proportions of feature j, $j = 1, \ldots, p$ on n samples. By assuming that each $x_i^{(j)}$ has a probability density function (2) with parameters $\alpha = \alpha_i^{(j)}$, $\mu = \mu_i^{(j)}$, and $\phi = \phi_i^{(j)}$, we define the following zero-inflated beta regression model (ZIBSeq) to fit the parameters in mixture distribution (2) for feature j:

$$\text{logit}(\alpha_i^{(j)}) = \rho_0^{(j)}, \quad \text{logit}(\mu_i^{(j)}) = \beta_0^{(j)} + \beta_1^{(j)} y_i, \quad \phi_i^{(j)} = T_i - 1, \quad (3)$$

where $\rho_0^{(j)}$, $\beta_0^{(j)}$, and $\beta_1^{(j)}$ are unknown regression parameters to be estimated, y_i is an outcome measurement indicating the class of sample i, while T_i is the ith sample depth. If feature j is not associated with the output class (for example, there is no significant difference in relative abundance of taxa i between two conditions) then $\beta_1^{(j)}$ is zero. For large samples, a chi-squared distribution can be used as an approximation to the true null distributions in testing the significance of $\beta_1^{(j)}$. Numerically finding the maximum likelihood estimates of the parameters in model (3) is implemented by R package GAMLSS (Stasinopoulos and Rigby, 2007). For testing the significance of $\beta_1^{(j)}$, t statistic and corresponding p-value are calculated with GAMLSS. When there are more than two classes, we recode **y** into multiple binary vectors with the popular one-vs.-rest or one-vs.-one scheme, and then apply zero-inflated beta regression to the recoded **y**'s.

Approximation of Dispersion Parameter ϕ: Let c denote the number of reads on a particular feature; assume c follows a binomial distribution, that is, $c \sim \text{Bin}(T,\mu)$, where T is the sample depth and μ is the probability that the sample reads in this feature. Let $x = \frac{c}{T} \sim N(\mu, \frac{\mu(1-\mu)}{T})$. Then the variance of proportion x is

$$Var(x) = \frac{\mu(1-\mu)}{T}. \quad (4)$$

On the other hand, if the proportion x follows a beta distribution under parameterization in (1), then we have

$$Var(x) = \frac{\mu(1-\mu)}{\phi+1}. \tag{5}$$

From (4) and (5) it is easy to obtain an approximation of the dispersion parameter ϕ:

$$\phi \approx T - 1. \tag{6}$$

We implemented this result in Equation (3).

Multiple Hypothesis Testing Correction: In metagenomics studies, it is very common to have thousands of features tested against the null hypothesis. The purpose of these tests is to identify many truly alternative features without including too many false positives. However, the direct application of the unadjusted chi-squared statistic can lead to large numbers of false positives; therefore, multiple hypothesis testing correction must be considered.

Aims to control the false discovery rate (FDR), q-value has been widely accepted as an alternative approach for multiple hypothesis testing correction in recent years (Benjamini, 2010). In this article, the significance of a differential test is measured by a q-value calculated using the algorithm developed by Storey and Tibshirani (2003). Under the assumption that p-values are uniformly distributed, this method will yield conservative q-value estimates. Our method can be summarized as below:

Step 1 Feature screening. Remove the features with total counts less than $2 \times n$, where n is the sample size.
Step 2 Data normalization. Convert the raw abundance c_{ij} to a proportion c_{ij}/T_i, where T_i is the total counts of sample i.
Step 3 Zero-inflated beta regression. Perform zero-inflated beta regression between each normalized feature (response variable) and outcome (explanatory variable); obtain the p-value of the regression coefficient in each regression.
Step 4 Multiple hypothesis testing correction. Calculate q-values based on p-values obtained in step 3; features with q-values less than or equal to significant levels are chosen.

3. RESULTS

We first present the results of four simulation studies to evaluate and compare ZIBSeq (zib), ZIBSeq with square root transformed response variable (zib_sqrt), and metagenomeSeq (zig) discussed in section 2. Then, one case study on real data will show how ZIBSeq is employed to find meaningful features in genomic studies.

3.1. Simulation

In order to compare the abilities of zib, zib_sqrt, and zig in detecting differentially expressed features, simulations were designed to generate data from four common types of distributions: zero-inflated Poisson distribution (ZIP), zero-inflated negative binomial distribution (ZINBI), multinomial distribution (MN), and binomial distribution (BI). All methods use standard approaches of false discovery rate (FDR) and q-value (Benjamini and Hochberg, 1995) for multiple hypothesis correction. Differentially abundant features were determined at FDR <0.05. To evaluate the ability of ZIBSeq and metagenomeSeq in identifying differential features, R package ROCR (Sing et. al, 2005) is employed to perform the ROC analysis. In each simulation, ROC curves and area under the curve (AUC) values were averaged by 100 random experiments.

In simulations on zero-inflated Poisson distribution ZIP(α,μ) and zero-inflated negative binomial distribution ZINBI(α,μ,σ) (Johnson et al., 1993), 5 out of 100 features were set to be differentially expressed, with means $\mu = \{5, 5, 50, 300, 500\}$ and $\{6, 8, 60, 350, 600\}$, respectively. The remaining 95 genes were generated from the same distribution with μ varied from 5 to 750. The probability of zero α is chosen to be 0.05 in both simulations, while the dispersion parameter is $\sigma = 0.1$ in zero-inflated negative binomial distribution.

To get reasonable values in simulations on multinomial distribution MN($s, p_1, p_2, \ldots, p_{50}$) and binomial distribution BI(s, p_1), $\ldots$, BI(s, p_{50}), the sampling depth s of each sample was determined by random sampling from 2000 and 20000. Eight out of 50 features were differentially expressed in two sample classes

with probabilities p_1 to p_8 set to be {.000268, .000868, .00388, .00288, .0156, .0756, .134, .066904} and {.000368, .000268, .00288, .00888, .0256, .0156, .184, .062404}, respectively. Same $\{p_9, \ldots, p_{50}\}$ were used to generate the remaining 42 features values in two sample classes; these probabilities varied from 0.0005 to 0.17.

ROC curves were shown in Figures 1 and 2 to illustrate the performance of the two methods in detecting significant features for sample size $n = 200$ and $n = 50$, respectively. Corresponding AUC values were calculated and listed in Table 2 to compare the performance of ZIBSeq and metagenomeSeq on data from these four distributions.

As shown from Figure 1 and 2 and Table 2, all methods (zib, zib_sqrt, and zig) performed similarly on ZIP and ZINBI data. With the sample size 200, the average AUCs of zib and zib_sqrt methods for ZIP data were 0.983, and for ZINBI data were 0.944 and 0.945, respectively, while the average AUCs of zig model for ZIP and ZINBI data were 0.991 and 0.959, respectively. Similar results were achieved with the sample size 50, indicating that zig performed slightly better than zib and zib_sqrt with larger AUC values at different sample sizes. However, it is clear that zib and zib_sqrt were more effective than zig on MN and BI data. With the sample size 200, both zib and zib_sqrt achieved the AUC of 0.999 for MN data and achieved the AUC of 0.998 for MN data respectively, while zig only had the average AUCs of 0.897 and 0.873, respectively. The advantages of zib and zib_sqrt are more significant with a smaller sample size of $n = 50$. Both zib and zib_sqrt had the average AUCs of 0.980 for ZIP data, and 0.955 for ZINBI data, while zig

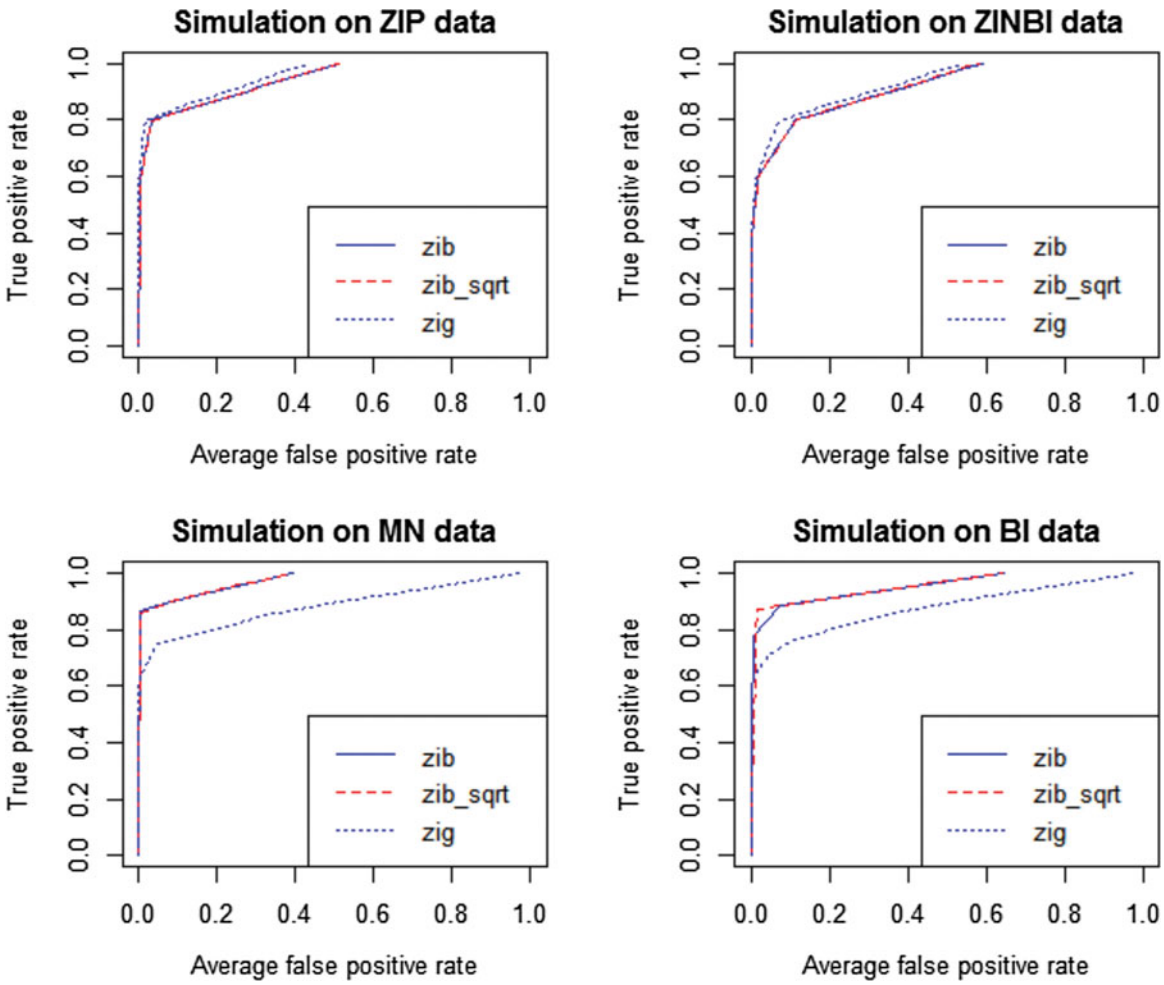

FIG. 1. Comparison of ZIBSeq (zib), ZIBSeq with square root transformation (zib_sqrt), and metagenomeSeq (zip) based on simulated data from zero-inflated Poisson distribution (ZIP), zero-inflated negative binomial distribution (ZINBI), multinomial distribution (MN), binomial distribution (BI), and $n = 200$.

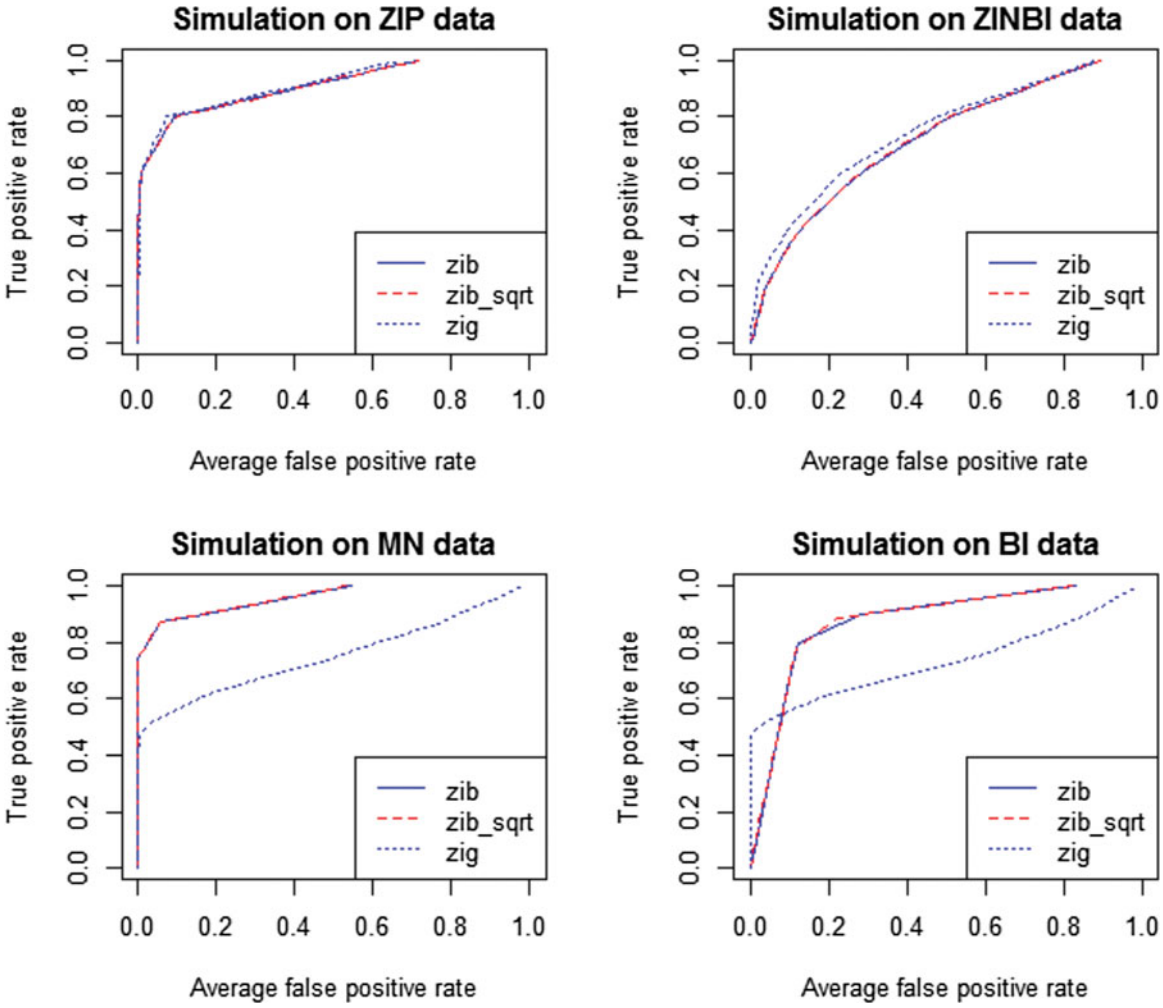

FIG. 2. Comparison of ZIBSeq (zib), ZIBSeq with square root transformation (zib_sqrt), and metagenomeSeq (zip) based on simulated data from zero-inflated Poisson distribution (ZIP), zero-inflated negative binomial distribution (ZINBI), multinomial distribution (MN), binomial distribution (BI), and $n = 50$.

only achieved the average AUCs of 0.761 and 0.740 for MN and BI data, respectively. In all simulations, zib and zib_sqrt performed almost the same.

3.2. *Real metagenomic data*

The metagenomic dataset was downloaded from dbGaP under study ID phs000258. The data and analytical results were first reported by Zupancic et al. (2012). There were a total of 310 Amish adult samples with 112 males and 198 females. After aligning the 16S rRNA sequences of gut microbiota to reference sequences and taxonomy databases, there were a total of 240 taxa at the genus level. The clinical

TABLE 2. AUC VALUES WITH ZIP, ZINBI, MN, AND BI SIMULATED DATA AND DIFFERENT SAMPLE SIZES

Sample Size	n = *200*				n = *50*			
Models	*ZIP*	*ZINBI*	*MN*	*BI*	*ZIP*	*ZINBI*	*MN*	*BI*
zib	0.983	0.944	0.999	0.998	0.938	0.734	0.980	0.955
zib_sqrt	0.983	0.945	0.999	0.998	0.938	0.736	0.980	0.955
zig	0.991	0.959	0.897	0.873	0.943	0.765	0.761	0.740

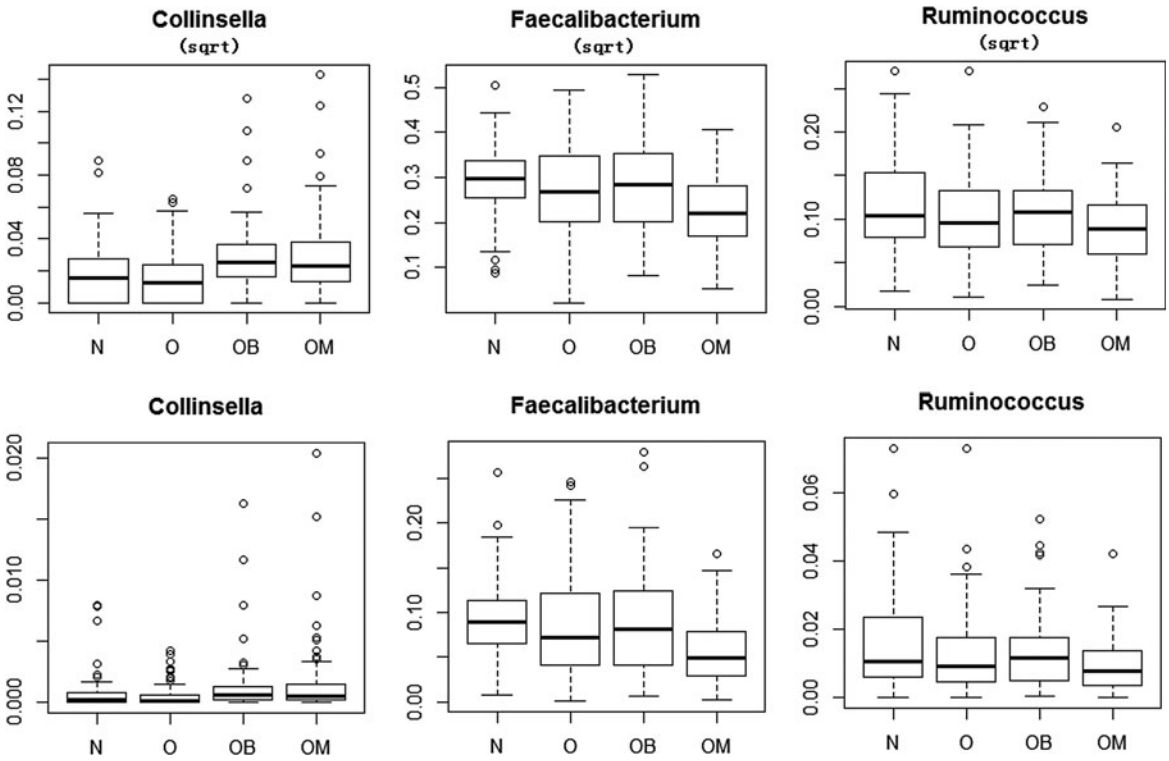

FIG. 3. Boxplots of Collinsella, Faecalibacterium, and Ruminococcus as square root transformations (first row) and proportion (second row) for four BMI groups.

phenotype we study is the body mass indices (BMI). The average BMIs are 27.2 and 30.3 for male and female, respectively. We try to identify BMI-associated genera with the proposed approach. We first drop the taxa with average reads below 2, because taxa with a small number of reads are not reliable and subject to noises and measurement errors. There are 62 out of 240 features that average more than 2 reads left for further studies. BMI can be treated as either an ordinal or categorical variable with normal (N: $BMI < 25$), overweight (OW: $25 \leq BMI < 30$), obese (OB: $30 \leq BMI \leq 40$), and morbidly obee (OM: $BMI > 40$). We first treated BMI as a categorical variable and applied zib and zib_sqrt to $N + OW + OB$ vs OM. Two taxa including *Faecalibacterium* and *Ruminococcus* were identified with both zib and zib_sqrt with $FDR < 0.05$. The same two taxa were selected with zib, but three taxa including an additional, *Collinsella* (besides *Faecalibacterium* and *Ruminococcus*), are chosen when considering the BMI group as ordinal variables as shown in Figure 3.

Also shown in Figure 3, both *Faecalibacterium* and *Ruminococcus* have lower relative abundance in the morbid obesity group, while *Collinsella* has higher relative abundance in both obese and morbidly obee groups. It is well known that intestinal microbiota composition varies between healthy and diseased individuals for numerous diseases including obesity. It has been shown in the literature that *Faecalibacterium* was significantly increased in response to dietary factors and weight loss (Remely et al., 2015) and was suggested as a target for intervention (Thomas et al., 2014). Both *Ruminococcus* and *Collinsella* are less studied, but there was evidence indicating that the relative abundance of *Ruminococcus* varies between obese and nonobese individuals (Kasaiet al., 2015). The biological importance of *Collinsella* needs to be further explored.

4. DISCUSSION

We proposed ZIBSeq tools to handle the metagenomics compositional data with many zeros efficiently. When compared with the popular zig model, our approach performed well with various kinds of data, especially with data simulated from multinomial and binomial distributions. In addition, even though

negative q-values were sometimes observed in simulations based on MN and BI data for all methods, zig approach is more likely to obtain negative q-values since it tends to produce more small p-values, which violates the assumption of uniform distribution in the q-value calculation algorithm proposed by Storey and Tibshirani (2003). Benefitting from the flexibility of beta distribution, our method is more stable. It can also identify biologically important taxa in a real application.

ACKNOWLEDGMENTS

This work was partially supported by the National Science Foundation (NSF) grant DMS-222381 (Z.L.), NSFC grant 10726054 (X.P.), and NIH grants 5P30CA-16042 and 8UL1TR000124 (G.L.). The funders had no role in the preparation of this article.

AUTHOR DISCLOSURE STATEMENT

The authors declare that no competing financial interests exist.

REFERENCES

Anders, S., and Huber, W. 2010. Differential expression analysis for sequence count data. *Genome Biol.* 11, R106.

Benjamini, Y. 2010. Simultaneous and selective inference: Current successes and future challenges. *Biom. J.* 52, 708–721.

Benjamini, Y., and Hochberg, Y. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Statist. Soc. B* 57, 289–300.

Ferrari, S.L.P., and Cribari-Neto, F. 2004. Beta regression for modelling rates and proportions. *J. Appl. Stat.* 31, 799–815.

Ghodsi, M., Liu, B., and Pop, M. 2011. DNACLUST: Accurate and efficient clustering of phylogenetic marker genes. *BMC Bioinform.* 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.

Johnson, N.L., Kotz, S., and Kemp, A.W. 1993. *Univariate Discrete Distributions, Vol. 1*, 1st ed. John Wiley and Sons, New York.

Kasai, C., Sugimoto, K., Moritani, I., et al. 2015. Comparison of the gut microbiota composition 11 between obese and non-obese individuals in a Japanese population, as analyzed by terminal restriction fragment length polymorphism and next-generation sequencing. *BMC Gastroenterol.* 15, 100.

Li, J., Witten, D.M., Johnstone, I.M., et al. 2012. Normalization, testing, and false discovery rate estimation for RNA-sequencing data. *Biostatistics* 13, 523–538.

Liu, Z., Hsiao, W., Cantarel, B.L., et al. 2011. Sparse distance based learning for simultaneous multiclass classification and feature selection of metagenomic data. *Bioinformatics* 27, 3242–3249.

Liu, Z., Sun, F., Braun, J., et al. 2014. Multilevel regularized regression for simultaneous taxa selection and network construction with metagenomic count data. *Bioinformatics* 31, 1067–1074.

Morgan, X.C., Tickle, T.L., Sokol, H., et al. 2012. Dysfunction of the intestinal microbiome in inammatory bowel disease and treatment. *Genome Biol.* 13, R79.

Mortazavi, A., Williams, B.A., McCue, K., et al. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* 5, 621–628.

Ospina, R., and Ferrari, S.L.P. 2012. A general class of zero-or-one inated beta regression models. *Comput. Stat. Data Anal.* 56, 1609–1623.

Paulson, J.N., Stine, O.C., Bravo, H.C., et al. 2013. Differential abundance analysis for microbial marker-gene surveys. *Nat. Methods* 10, 1200–1202.

Remely, M., Tesar, I., Hippe, B., et al. 2015. Gut microbiota composition correlates with changes in body fat content due to weight loss. *Benef. Microbes* 6, 431–439.

Robinson, M., McCarthy, D.J., and Smyth, G.K. 2010. edgeR: A bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140.

Sing, T., Sander, O., Beerenwinkel, N., et al. 2005. ROCR: Visualizing classifier performance in R. *Bioinformatics* 21, 3940–3941.

Stasinopoulos, D.M., and Rigby, R.A. 2007. Generalized additive models for location scale and shape (GAMLSS) in R. *J. Stat. Softw.* 23, 1–43.

Storey, J.D., and Tibshirani, R. 2003. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U. S. A.* 100, 9440–9445.

Thomas, L.V., Ockhuizen, T., and Suzuki, K. 2014. Exploring the inuence of the gut microbiota and probiotics on health: A symposium report. *Br. J. Nutr.* 112 Suppl 1, S1–S18.

Turnbaugh, P.J., Hamady, M., Yatsunenko, T., et al. 2009. A core gut microbiome in obese and lean twins. *Nature* 457, 480–484.

Wang, Q., Garrity, G.M., Tiedje, J.M., et al. 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* 73, 5261–5267.

Young, M.D., McCarthy, D.J., Wakefield, M.J., et al. 2012. Differential expression for RNA sequencing (RNA-Seq) data: Mapping, summarization, statistical analysis, and experimental design, 169–190. In *Bioinformatics for High Throughput Sequencing*. Springer, New York.

Zupancic, M.L., Cantarel, B.L., Liu, Z.E.F., et al. 2012. Analysis of the gut microbiota in the old order Amish and its relation to the metabolic syndrome. *PLoS One* 7, e43052.

Address correspondence to:
Prof. Zhenqiu Liu
Department of Medicine/Biomedical Sciences
Cedars-Sinai Medical Center
116 N. Robertson Blvd, PACT, Suite 900C
Los Angeles, CA 90048

E-mail: zhenqiu.liu@cshs.org