Biostatistics (2018) **00**, 00, *pp.* 1–16
doi:10.1093/biostatistics/kxy025

Zero-inflated generalized Dirichlet multinomial regression model for microbiome compositional data analysis

ZHENG-ZHENG TANG*

Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI 53715, USA and Wisconsin Institute for Discovery, Madison, WI 53715, USA

tang@biostat.wisc.edu

GUANHUA CHEN

Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI 53715, USA

Summary

There is heightened interest in using high-throughput sequencing technologies to quantify abundances of microbial taxa and linking the abundance to human diseases and traits. Proper modeling of multivariate taxon counts is essential to the power of detecting this association. Existing models are limited in handling excessive zero observations in taxon counts and in flexibly accommodating complex correlation structures and dispersion patterns among taxa. In this article, we develop a new probability distribution, zero-inflated generalized Dirichlet multinomial (ZIGDM), that overcomes these limitations in modeling multivariate taxon counts. Based on this distribution, we propose a ZIGDM regression model to link microbial abundances to covariates (e.g. disease status) and develop a fast expectation–maximization algorithm to efficiently estimate parameters in the model. The derived tests enable us to reveal rich patterns of variation in microbial compositions including differential mean and dispersion. The advantages of the proposed methods are demonstrated through simulation studies and an analysis of a gut microbiome dataset.

Keywords: Compositional data analysis; Differential abundance; Hierarchical model; Microbiome; Score test; Zero-inflated model.

1. Introduction

The human microbiome is the microbe population living in and on the human body. The importance of microbiome in human health and disease has been increasingly recognized. Research in this area has begun to discover relationships between human microbiome and many complex diseases such as cancer, diabetes, psoriasis, and obesity (Cho and Blaser, 2012; Cho *and others*, 2012; Qin *and others*, 2012; Ahn *and others*, 2013; Alekseyenko *and others*, 2013). Advances in high-throughput sequencing technologies

*To whom correspondence should be addressed.

allow scientists to accurately quantify a wide variety of microbes. In most microbiome studies, the data come from high-throughput sequencing of the 16S ribosomal ribonucleic acid (rRNA) gene. This gene is the target in sequencing experiments because it is ubiquitous in the bacterial kingdom, comprised of conserved and variable regions, and evolves at relative constant rates (Kuczynski *and others*, 2012). These features allow us to determine the types and abundances of different microbes in a sample with high sensitivity. By comparing sequencing reads to a reference 16S rRNA database, each sequence can be assigned to a series of taxonomic identities (i.e. lineage) at the levels of kingdom, phylum, class, order, family, and genus (DeSantis *and others*, 2006; Cole *and others*, 2007; Liu *and others*, 2008; Jovel *and others*, 2016). Therefore, for each taxonomic level, the final data from one sample can be summarized as a vector of read counts, referred to as taxon counts, with each component corresponding to a taxon at that particular level. The total taxon counts for each sample is bounded by the sequencing depth that does not reflect the actual microbial load in the sample. Thus, the taxon counts are compositional data and only reflect the relative abundances (i.e. proportions) of different microorganisms.

Linking multivariate taxon counts to covariates of interest (as a special case, detecting differential abundance between groups) is the common practice in microbiome studies of human disease. The quasi-conditional association test (QCAT) does not make any distributional assumption on taxon counts (Tang *and others*, 2017). This non-parametric test is very robust to the underlying correlation structures, but can be less powerful than parametric tests based on suitable distributional assumptions. In addition, the test is underpowered in the presence of excessive zeros in taxon counts. In the microbiome data, most taxa exhibit zero-inflation with zero observations representing the absence of the taxa in the samples (often referred to as structural zeros) or the undersampling of the rare taxa (often referred to as sampling zeros) (Li, 2015). The two-part version of QCAT models positive counts and zero counts separately and combines association evidence from both parts (Tang *and others*, 2017). This test is more powerful than regular QCAT if taxon counts are zero-inflated. However, the two-part test does not distinguish different types of zeros. Indeed, it is not possible to disentangle structural zeros and sampling zeros unless fitting a good parametric model to the data.

The power of parametric tests heavily depends on how well the adopted probability distribution fits the data. The Dirichlet multinomial (DM) distribution is commonly used to model taxon counts. In the DM, a vector of counts follows the multinomial distribution with underlying proportion parameters sampled from a Dirichlet distribution. The DM has been shown to fit microbiome data better than the multinomial distribution and many statistical methods have been developed based on the DM (La Rosa *and others*, 2012; Chen and Li, 2013; Wadsworth *and others*, 2017; Wang and Zhao, 2017). In particular, two DM-based association tests exist: one tests for differential mean and the other tests for both differential mean and dispersion (La Rosa *and others*, 2012). The test for differential mean assumes homogeneous dispersion levels between groups, but many scenarios can lead to deviations from this assumption. It is well-known that microbiome compositions are very dynamic—the abundance of one microbe can be altered by the presence/absence of other microbes that cooperate or compete with that microbe, as well as by changes in environmental factors such as temperature, host diet, and lifestyle (Gilbert *and others*, 2016). Therefore, the disease-microbe association can be moderated by many factors, resulting in heterogeneous dispersion levels between the disease and control groups. A taxon with heterogeneous dispersion could represent a previously undetected interaction with other taxa or environmental factors. Consequently, ignoring the differential dispersion misses the opportunity of identifying these important taxa.

Several recent studies have shown that the DM may not be adequate for microbiome data (O'Brien *and others*, 2016; Sankaran and Holmes, 2017; Shi and Li, 2017; Tang *and others*, 2017). Specifically, the DM model intrinsically imposes a negative correlation among taxon counts, whereas the actual data display both positive and negative correlations (Mandal *and others*, 2015). In addition, the DM has only one dispersion parameter so that it cannot flexibly handle various dispersion patterns and zero-inflation levels among multiple taxa.

Generalized DM (GDM) distribution addresses one key limitation of the DM by allowing more general covariance structure. Comparing to the DM, the GDM chooses a more flexible distribution, the Generalized Dirichlet (GD) (Connor and Mosimann, 1969), as a prior for the multinomial. The GDM model has not been applied to microbiome data. It is not clear if the additional parameters in the GDM are necessary for modeling microbiome data and if the GDM can handle excessive zeros in taxon counts. In addition, the complex form of the GDM probability density function renders numerical challenges in statistical inference (Zhang *and others*, 2017).

In this article, we develop a new probability distribution—zero-inflated GDM (ZIGDM)—for modeling microbiome compositional data that includes the GDM as a special case. In contrast to the DM model, the ZIGDM has additional parameters to flexibly accommodate the over-dispersion and zero-inflation of the data. We focus on demonstrating the usefulness of this additional flexibility in microbiome association analysis. We propose to use the ZIGDM regression model to link the mean and dispersion levels of the microbial abundance to the covariates of interest. Based on the ZIGDM regression, we derive association tests for detecting differential mean and dispersion. Like the GDM, the ZIGDM model does not belong to the natural exponential family and the parameter estimation is not simple. To meet this challenge, we develop a fast expectation–maximization (EM) algorithm. We use extensive simulation studies to evaluate the performance of these tests. An application to a gut microbiome data leads to a discovery of differential microbial sub-compositions between the body mass index (BMI) groups. The ZIGDM and GDM models fit the gut microbiome data significantly better than the DM model.

2. GD AND ZIGD

2.1. *GD model for random proportions*

We consider $K+1$ taxa in the microbial composition with a total of N sequencing reads. We denote the vector of counts as $\mathbf{Y} = (Y_1, \ldots, Y_K)$ with $Y_{K+1} = N - \sum_{j=1}^{K} Y_j$, and the underlying unobserved proportions as $\mathbf{P} = (P_1, \ldots, P_K)$ with $P_{K+1} = 1 - \sum_{j=1}^{K} P_j$. To flexibly model random proportions with a complex correlation structure, Connor and Mosimann (1969) proposed the Generalized Dirichlet (GD) distribution. Specifically, the GD random proportions $\mathbf{P}$ can be constructed from a set of mutually independent Beta variables $\mathbf{Z} = (Z_1, \ldots, Z_K)$ as

$$P_1 = Z_1, \;\; P_j = Z_j \prod_{i=1}^{j-1} (1 - Z_i), \;\; j = 2, \ldots, K. \tag{2.1}$$

This GD construction in (2.1) resembles the stick-breaking process for constructing a truncated Dirichlet process (Ishwaran and James, 2001). The density function for Z_j can be written as $f(Z_j) = \frac{1}{\mathcal{B}(a_j, b_j)} Z_j^{a_j - 1} (1 - Z_j)^{b_j - 1}$, where $a_j > 0$ and $b_j > 0$ are two parameters in the Beta distribution and $\mathcal{B}(\cdot, \cdot)$ is the Beta function. Applying the transformation, we obtain the density function of $\mathbf{P}$ as

$$f(\mathbf{P}) = \prod_{j=1}^{K} \frac{1}{\mathcal{B}(a_j, b_j)} P_j^{a_j - 1} (1 - P_1 - \ldots - P_j)^{c_j}, \tag{2.2}$$

where $c_j = b_j - a_{j+1} - b_{j+1}$ for $j = 1, \ldots, K-1$ and $c_K = b_K - 1$. This distribution is referred to as GD($\mathbf{a}, \mathbf{b}$), where $\mathbf{a} = (a_1, \ldots, a_K)$ and $\mathbf{b} = (b_1, \ldots, b_K)$. Conversely, a set of mutually independent Beta variables can be derived from a GD random proportions $\mathbf{P}$ by the transformation

$$Z_1 = P_1, \;\; Z_j = P_j \Bigg/ \left(1 - \sum_{i=1}^{j-1} P_i\right), \;\; j = 2, \ldots, K. \tag{2.3}$$

The GDM is given by using the GD as a prior for the multinomial distribution. Wong (1998) has shown that the GD is a conjugate prior for the multinomial. Specifically, suppose $\mathbf{Y}$ follows the multinomial distribution with a GD($\mathbf{a}, \mathbf{b}$) prior on the proportion parameters $\mathbf{P}$, the posterior probability of ($\mathbf{P} \mid \mathbf{Y}$) is a GD($\mathbf{a}^*, \mathbf{b}^*$), where $\mathbf{a}^* = (a_1^*, \ldots, a_K^*)$, $\mathbf{b}^* = (b_1^*, \ldots, b_K^*)$, $a_j^* = a_j + Y_j$ and $b_j^* = b_j + Y_{j+1} + \ldots + Y_{K+1}$, $j = 1, \ldots, K$. The GDM has been applied to model multivariate count data with complex correlation structures such as in RNA-seq data analysis (Zhang *and others*, 2017). However, its usefulness in analyzing microbiome data has not been evaluated.

2.2. *ZIGD model for random proportions with zero components for absent taxa*

The GD model assumes all taxa have positive proportions (i.e. are present in the sample) and the observed zeros in $\mathbf{Y}$ are sampling zeros. To model absent taxa (i.e. structural zeros), we assume Z_j follows zero-inflated Beta (ZIB) distribution with parameters (π_j, a_j, b_j), where π_j represents the probability of $Z_j = 0$. We, then, transform these Z's to P's using (2.1), and refer to the resulting distribution of $\mathbf{P}$ as ZIGD($\boldsymbol{\pi}, \mathbf{a}, \mathbf{b}$), where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$. Clearly, $Z_j = 0$ is equivalent to $P_j = 0$ based on their relationship (2.1). We let $I(\cdot)$ denote the indicator function. The variable $\Delta_j = I(Z_j = 0) = I(P_j = 0)$ follows Bernoulli(π_j) and indicates the absence and presence of the taxon j by values of 1 and 0, respectively. If we assume all taxa are present ($\Delta_1 = \ldots = \Delta_K = 0$), the ZIGD becomes the GD. Suppose we have L taxa present in the sample. We let $\mathcal{U} = (u_1, \ldots, u_L)$ denote the set of indexes for these taxa (i.e. $\Delta_{u_1} = \ldots = \Delta_{u_L} = 0$). The complement set $\overline{\mathcal{U}}$ includes the indexes for the absent taxa. Suppose we observe M taxa with zero counts. We let $\mathcal{V} = (v_1, \ldots, v_M)$ denote the set of indexes for these taxa (i.e. $Y_{v_1} = \ldots = Y_{v_M} = 0$) and $\overline{\mathcal{V}}$ denote the complement of set $\mathcal{V}$. Note that the two sets $\mathcal{U}$ and $\mathcal{V}$ are not exclusive: their intersection $\mathcal{U} \cap \mathcal{V}$ indexed taxa that are present in the sample but have zero counts due to the undersampling in the sequencing experiment (i.e. sampling zeros).

We can use the ZIGD as a prior for the multinomial and term the resulting marginal distribution for the counts as ZIGDM. In the remaining content of this section, we show that the ZIGD is a conjugate prior for the multinomial. In the derivation, we let $X_{\mathcal{A}}$ denote the sub-vector of vector X defined by the index set $\mathcal{A}$. Suppose $\mathbf{Y}$ follows a multinomial with a ZIGD($\boldsymbol{\pi}, \mathbf{a}, \mathbf{b}$) prior on the proportion parameters $\mathbf{P}$. We let $\boldsymbol{\Delta} = (\Delta_1, \ldots, \Delta_K)$. Because $\Delta_j = I(P_j = 0)$, we have $f(\mathbf{P}, \boldsymbol{\Delta}) = f(\mathbf{P})$, hence, the posterior probability of the proportions given observed counts can be expressed as

$$f(\mathbf{P} \mid \mathbf{Y}) = f(\mathbf{P}, \boldsymbol{\Delta} \mid \mathbf{Y}) = f(\mathbf{P} \mid \boldsymbol{\Delta}, \mathbf{Y}) f(\boldsymbol{\Delta} \mid \mathbf{Y}). \tag{2.4}$$

Because $P_j = 0$ when $\Delta_j = 1$ (a taxon has the proportion of 0 if it is absent from a sample), we have $f(\mathbf{P} \mid \boldsymbol{\Delta}) = I(\mathbf{P}_{\overline{\mathcal{U}}} = \mathbf{0}) f(\mathbf{P}_{\mathcal{U}} \mid \boldsymbol{\Delta}_{\mathcal{U}} = \mathbf{0}, \boldsymbol{\Delta}_{\overline{\mathcal{U}}} = \mathbf{1})$. Because $f(\mathbf{P}_{\mathcal{U}} \mid \boldsymbol{\Delta}_{\mathcal{U}} = \mathbf{0}, \boldsymbol{\Delta}_{\overline{\mathcal{U}}} = \mathbf{1})$ follows the GD($\mathbf{a}_{\mathcal{U}}, \mathbf{b}_{\mathcal{U}}$) and the GD is a conjugate prior for the multinomial, the posterior probability $f(\mathbf{P}_{\mathcal{U}} \mid \boldsymbol{\Delta}_{\mathcal{U}} = \mathbf{0}, \boldsymbol{\Delta}_{\overline{\mathcal{U}}} = \mathbf{1}, \mathbf{Y})$ follows the GD($\mathbf{a}_{\mathcal{U}}^*, \mathbf{b}_{\mathcal{U}}^*$). Because $\Delta_j = 0$ when $Y_j > 0$ (a taxon is present in the sample if its count is positive), we have $f(\boldsymbol{\Delta} \mid \mathbf{Y}) = I(\boldsymbol{\Delta}_{\overline{\mathcal{V}}} = \mathbf{0}) f(\boldsymbol{\Delta}_{\mathcal{V}} \mid \mathbf{Y}_{\mathcal{V}} = \mathbf{0}, \mathbf{Y}_{\overline{\mathcal{V}}} > \mathbf{0})$. The mass function for $\boldsymbol{\Delta}_{\mathcal{V}}$ given $\mathbf{Y}_{\mathcal{V}} = \mathbf{0}, \mathbf{Y}_{\overline{\mathcal{V}}} > \mathbf{0}$ can be derived as follows

$$\begin{aligned}
&f(\boldsymbol{\Delta}_{\mathcal{V}} \mid \mathbf{Y}_{\mathcal{V}} = \mathbf{0}, \mathbf{Y}_{\overline{\mathcal{V}}} > \mathbf{0}) \\
&\propto f(\boldsymbol{\Delta}_{\mathcal{V}}) f(\mathbf{Y}_{\mathcal{V}} = \mathbf{0}, \mathbf{Y}_{\overline{\mathcal{V}}} > \mathbf{0} \mid \boldsymbol{\Delta}_{\mathcal{V}}, \boldsymbol{\Delta}_{\overline{\mathcal{V}}} = \mathbf{0}) \\
&= f(\boldsymbol{\Delta}_{\mathcal{V}}) \int_{\mathbf{P}} f(\mathbf{Y}_{\mathcal{V}} = \mathbf{0}, \mathbf{Y}_{\overline{\mathcal{V}}} > \mathbf{0} \mid \mathbf{P}, \boldsymbol{\Delta}_{\mathcal{V}}, \boldsymbol{\Delta}_{\overline{\mathcal{V}}} = \mathbf{0}) f(\mathbf{P} \mid \boldsymbol{\Delta}_{\mathcal{V}}, \boldsymbol{\Delta}_{\overline{\mathcal{V}}} = \mathbf{0}) \mathrm{d}\mathbf{P} \\
&\propto \prod_{j \in \mathcal{V}} \left\{ \pi_j^{\Delta_j} (1 - \pi_j)^{(1 - \Delta_j)} \right\} \prod_{j \in \mathcal{U} \cap \mathcal{V}} \left\{ \frac{\mathcal{B}(a_j^*, b_j^*)}{\mathcal{B}(a_j, b_j)} \right\}
\end{aligned}$$

$$= \prod_{j \in \mathcal{V}} \left\{ \pi_j^{\Delta_j} \left[(1-\pi_j) \frac{\mathcal{B}(a_j^*, b_j^*)}{\mathcal{B}(a_j, b_j)} \right]^{(1-\Delta_j)} \right\}. \tag{2.5}$$

The last equality holds because $\Delta_j = 0$ if and only if $j \in \mathcal{U}$ by definition. Hence, the posterior probability $f(\mathbf{P} \mid \mathbf{Y})$ follows a ZIGD with zero-inflation on the taxa having observed zero counts and the probability of the observed zero being structural zero is $\frac{\pi_j}{\pi_j + (1-\pi_j)\frac{\mathcal{B}(a_j^*, b_j^*)}{\mathcal{B}(a_j, b_j)}}$ $(j \in \mathcal{V})$.

3. ZIGDM REGRESSION MODEL

Suppose we have n subjects measured on $K+1$ taxa. We let Y_{ij} and P_{ij} denote the observed count and the underlying true proportion for taxon j in subject i, and $\mathbf{X}_i$ denote the d-dimensional vector including a unit component for the intercept, covariates of interest, and confounding variables. We assume the count vector $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iK})$ follows the ZIGDM$(\boldsymbol{\pi}_i, \mathbf{a}_i, \mathbf{b}_i)$, where $\boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{iK})$, $\mathbf{a}_i = (a_{i1}, \ldots, a_{iK})$, and $\mathbf{b}_i = (b_{i1}, \ldots, b_{iK})$. As described in the previous section, the equivalent hierarchical model can be expressed as

$$\begin{gathered}
\Delta_{ij} \sim \text{Bernoulli}(\pi_{ij}), \;\; j = 1, \ldots, K, \\
Z_{ij} = 0 \text{ if } \Delta_{ij} = 1, \;\; Z_{ij} \mid \Delta_{ij} = 0 \sim \text{Beta}(a_{ij}, b_{ij}), \;\; j = 1, \ldots, K, \\
P_{i1} = Z_{i1}, \;\; P_{ij} = Z_{ij} \prod_{k=1}^{j-1} (1 - Z_{ik}), \;\; j = 2, \ldots, K, \\
\mathbf{Y}_i \mid \mathbf{P}_i \sim \text{Multinomial}(\mathbf{P}_i, N_i), \text{ where } \mathbf{P}_i = (P_{i1}, \ldots, P_{iK}) \text{ and } N_i = \sum_{j=1}^{K+1} Y_{ij}.
\end{gathered} \tag{3.1}$$

Under this model, $\boldsymbol{\pi}_i$, $\mathbf{a}_i$, and $\mathbf{b}_i$ are three possible sets of parameters that can be linked to $\mathbf{X}_i$. For each taxon j in subject i, the π_{ij} controls the probability of absence and the a_{ij} and b_{ij} control the abundance distribution at the presence. To facilitate the interpretation, we model $\mu_{ij} = a_{ij}/(a_{ij} + b_{ij})$ and $\sigma_{ij} = 1/(1 + a_{ij} + b_{ij})$ as opposed to a_{ij} and b_{ij}, where μ_{ij} pertains to the mean of the Beta variable and σ_{ij} can be viewed as the dispersion parameter because the variance of the Beta variable takes the form $\mu_{ij}(1-\mu_{ij})\sigma_{ij}$. It is natural to use logit link functions because π_{ij}'s, μ_{ij}'s and σ_{ij}'s take values between 0 and 1:

$$\pi_{ij} = \frac{e^{\boldsymbol{\gamma}_j^{\mathrm{T}} \mathbf{x}_i}}{1 + e^{\boldsymbol{\gamma}_j^{\mathrm{T}} \mathbf{x}_i}}, \;\; \mu_{ij} = \frac{e^{\boldsymbol{\alpha}_j^{\mathrm{T}} \mathbf{x}_i}}{1 + e^{\boldsymbol{\alpha}_j^{\mathrm{T}} \mathbf{x}_i}}, \;\; \text{and} \;\; \sigma_{ij} = \frac{e^{\boldsymbol{\beta}_j^{\mathrm{T}} \mathbf{x}_i}}{1 + e^{\boldsymbol{\beta}_j^{\mathrm{T}} \mathbf{x}_i}}, \;\; j = 1, \ldots, K, \tag{3.2}$$

where $\boldsymbol{\gamma}_j = (\gamma_{1j}, \ldots, \gamma_{dj})$, $\boldsymbol{\alpha}_j = (\alpha_{1j}, \ldots, \alpha_{dj})$, and $\boldsymbol{\beta}_j = (\beta_{1j}, \ldots, \beta_{dj})$ are regression coefficients for taxon j. The design matrices are not necessarily the same for the three features of abundance distribution (i.e. absence probability, mean, and dispersion). For ease of presentation, we keep them the same in describing the regression model.

We write the complete set of parameters as $\boldsymbol{\theta} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_K, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_K, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K)$. The likelihood-based inference on $\boldsymbol{\theta}$ is not simple because the observed log-likelihood function is complicated. We describe below an efficient EM algorithm for fitting the model and estimating parameters. Formula (3.3)

gives the complete data log-likelihood expressed in terms of Z's:

$$\begin{aligned} l(\boldsymbol{\theta}) &= \log\left[\prod_{i=1}^{n}\left\{f(\mathbf{Y}_i \mid \mathbf{Z}_i)\prod_{j=1}^{K} f(Z_{ij})\right\}\right] \\ &= \sum_{i=1}^{n}\log\{f(\mathbf{Y}_i \mid \mathbf{Z}_i)\} \\ &\quad + \sum_{j=1}^{K}\sum_{i=1}^{n}\Big\{\Delta_{ij}\log\pi_{ij} + (1-\Delta_{ij})\log(1-\pi_{ij}) + \\ &\quad (1-\Delta_{ij})\left[-\log(\mathcal{B}(a_{ij},b_{ij})) + (a_{ij}-1)\log(Z_{ij}) + (b_{ij}-1)\log(1-Z_{ij})\right]\Big\}, \end{aligned} \tag{3.3}$$

where $a_{ij} = \mu_{ij}(1/\sigma_{ij} - 1)$ and $b_{ij} = (1-\mu_{ij})(1/\sigma_{ij} - 1)$. Using Z's instead of P's allows us to derive the explicit form of posterior expectations in the E-step and estimate parameters for each taxon independently in the M-step.

In the t-th E-step, we need to compute the expected complete data log-likelihood,

$$\begin{aligned} Q_\theta^* = \sum_{j=1}^{K}\sum_{i=1}^{n}\mathrm{E}\Big\{&\Delta_{ij}\log\pi_{ij} + (1-\Delta_{ij})\log(1-\pi_{ij}) + \\ &(1-\Delta_{ij})\left[-\log(\mathcal{B}(a_{ij},b_{ij})) + (a_{ij}-1)\log Z_{ij} + (b_{ij}-1)\log(1-Z_{ij})\right]\Big\}, \end{aligned} \tag{3.4}$$

where the expectation is with respect to the posterior distributions of $(\boldsymbol{\Delta}_i \mid \mathbf{Y}_i; \boldsymbol{\theta}^{(t-1)})$ and $(\mathbf{Z}_i \mid \boldsymbol{\Delta}_i, \mathbf{Y}_i; \boldsymbol{\theta}^{(t-1)})$ with $\boldsymbol{\theta}^{(t-1)}$ being the parameter estimates in the $(t-1)$-th M-step. Based on the results from the previous section, we have

$$\begin{aligned} \Delta_{ij}^* &= \mathrm{E}\left(\Delta_{ij} \mid \mathbf{Y}_i\right) = \begin{cases} 0 & \text{if } Y_{ij} > 0 \\ \dfrac{\pi_{ij}}{\pi_{ij} + (1-\pi_{ij})\dfrac{\mathcal{B}\left(a_{ij}^*, b_{ij}^*\right)}{\mathcal{B}\left(a_{ij}, b_{ij}\right)}} & \text{if } Y_{ij} = 0 \end{cases}, \\ A_{ij}^* &= \mathrm{E}\left(\log Z_{ij} \mid \mathbf{Y}_i, \Delta_{ij} = 0\right) = \psi\left(a_{ij}^*\right) - \psi\left(a_{ij}^* + b_{ij}^*\right), \\ B_{ij}^* &= \mathrm{E}\left(\log(1-Z_{ij}) \mid \mathbf{Y}_i, \Delta_{ij} = 0\right) = \psi\left(b_{ij}^*\right) - \psi\left(a_{ij}^* + b_{ij}^*\right), \end{aligned} \tag{3.5}$$

where $a_{ij}^* = a_{ij} + Y_{ij}$, $b_{ij}^* = b_{ij} + Y_{i(j+1)} + \ldots + Y_{i(K+1)}$, and $\psi(\cdot)$ is the digamma function.

Thus, Q_θ^* can be rewritten as

$$Q_\theta^* = \sum_{j=1}^{K} Q_{\gamma_j}^* + \sum_{j=1}^{K} Q_{\alpha_j,\beta_j}^*, \tag{3.6}$$

where $Q_{\gamma_j}^* = \sum_{i=1}^{n}\{\Delta_{ij}^*\log\pi_{ij} + (1-\Delta_{ij}^*)\log(1-\pi_{ij})\}$ and $Q_{\alpha_j,\beta_j}^* = \sum_{i=1}^{n}(1-\Delta_{ij}^*)\{-\log(\mathcal{B}(a_{ij},b_{ij})) + (a_{ij}-1)A_{ij}^* + (b_{ij}-1)B_{ij}^*\}$.

In the t-th M-step, for each taxon j, we obtain $\boldsymbol{\gamma}_j^{(t)}$ from maximizing the function $Q_{\gamma_j}^*$ and obtain $\boldsymbol{\alpha}_j^{(t)}$ and $\boldsymbol{\beta}_j^{(t)}$ from maximizing the function Q_{α_j,β_j}^*. The computation burden for the optimization is the same as a logistic and a weighted Beta regressions. Because the parameters for individual taxa are updated

independently, we can estimate parameters for all taxa in parallel by distributing the optimization jobs to multiple computing cores. In summary, the EM algorithm is computationally efficient because of the simple calculation of posterior expectations and the ability to update parameters in a taxon-by-taxon manner.

4. Association tests

By testing the different sets of parameters in the ZIGDM regression model, we can reveal a variety of patterns of variation in microbial communities. In this article, we focus on testing the null hypothesis that the covariates are not associated with mean ($H_0 : \boldsymbol{\alpha}_{*1} = \ldots = \boldsymbol{\alpha}_{*K} = 0$) or dispersion ($H_0 : \boldsymbol{\beta}_{*1} = \ldots = \boldsymbol{\beta}_{*K} = 0$), where $\boldsymbol{\alpha}_{*j}$ and $\boldsymbol{\beta}_{*j}$ are subsets of $\boldsymbol{\alpha}_j$ and $\boldsymbol{\beta}_j$ corresponding to the covariates of interest. We may test the null hypotheses by using the score, Wald, or likelihood ratio (LR) statistics. We have adopted score statistics, which are computationally faster and more stable than Wald and LR statistics (Lin and Tang, 2011). The derivation of the score statistic is included in the Appendix of supplementary material available at *Biostatistics* online. In our numerical studies, we evaluate the performance of score tests based on the ZIGDM and GDM regression models. When performing the association test for one feature of the abundance distribution (e.g. mean), we include only the intercept in models for the other features (e.g. absence probability, dispersion).

The asymptotic approximation of the test statistics may not be accurate when most of the observations are zeros, especially when the sample size is small. Therefore, we need to use permutation techniques to obtain *P*-values. It is computationally efficient to obtain the permutation *P*-values for score statistics because the null model (without confounding variables) would only include the intercept term and needs to be fit only once in the permutation procedure. In particular, we permute the covariate of interest and calculate the score test statistic in each permutation. The permutation *P*-value is the proportion of the permuted test statistics that are greater than the observed statistic.

5. Simulation evaluations

5.1. *Simulation strategy*

We conducted extensive simulation studies to investigate the performance of the proposed and existing methods. The ZIGDM_1 and ZIGDM_2 are tests based on the ZIGDM distribution for detecting differential mean and dispersion, respectively. The GDM_1 and GDM_2 are the GDM counterparts. The DM_1 and DM_2 are two tests based on the DM distribution. The DM_1 is a test for detecting differential mean and the DM_2 is an omnibus test for jointly detecting differential mean and dispersion. In the current implementation of these DM tests, the DM_1 employs the Wald statistic and the DM_2 employs the LR statistic (La Rosa *and others*, 2012). We used the HMP package in R for the DM tests in our numerical studies (La Rosa *and others*, 2016). The QCAT_1 and QCAT_2 are distribution-free tests based on generalized score statistics for detecting differential mean. The QCAT_2 employs the two-part model with one part modeling zero observations and the other part modeling positive abundance. We used the miLineage package in R for the QCAT tests in our numerical studies.

We simulated six taxon counts for two groups with same sample sizes and tested differential abundance in the six taxa between the two groups. The number of taxa is chosen to be six in order to reflect the fact that testing sub-composition on a taxonomic tree usually involves less than 10 taxa (see details in Section 6). We considered the sample sizes of 100 and 200 in all simulation studies. The total sequencing reads for each sample was simulated from Poisson(1000). In the power evaluation, we perturbed one taxon by changing either its mean abundance or its dispersion level in one group. We used 5000 simulated datasets to evaluate Type I error and power of the tests at the 0.05 significance level.

In the first set of simulations, taxon counts were generated from the DM model. In particular, the vector of proportion parameters for the multinomial distribution was simulated from a Dirichlet distribution with equal mean parameters of 1/6 and the dispersion parameter of 0.3. To evaluate power, we randomly selected a taxon and sampled its mean parameter from Uniform (0, 0.5) or its dispersion parameter from Uniform(0, 0.5).

In the second set of simulations, taxon counts were generated from the GDM or ZIGDM model. In particular, we first simulated five independent Beta or ZIB variables $(Z_{i1}, \ldots, Z_{i5})$ with equal Beta mean parameters of 0.2 and equal Beta dispersion parameters of 0.2. For the ZIGDM, the zero-inflation levels were sampled without replacement from $(0.1, 0.2, 0.4, 0.6, 0.8)$. We then transformed the Z's into a vector of proportions $(P_{i1}, \ldots, P_{i5})$ according to (2.1). The proportion of the sixth taxon was determined by $1 - \sum_{j=1}^{5} P_{ij}$. To evaluate power, we changed the Beta mean or dispersion parameter of a taxon: the Beta mean parameter for the differential taxon was sampled from Uniform(0, 0.5); the Beta dispersion parameter for the differential taxon was sampled from Uniform(0, 1). For the GDM, the differential taxon was randomly picked; for the ZIGDM, the differential taxon was the one with a certain level of zero-inflation. For example, when we consider the zero-inflation level of 0.2, we only perturb the taxon with that particular level of zero-inflation.

In the third set of simulations, taxon counts were generated from the Log-Normal (LN) or zero-inflated LN (ZILN) model. This set of simulations are designed to evaluate the robustness of different tests to the underlying distributions. In particular, we first simulated $(W_{i1}, \ldots, W_{i5})$ from a multivariate normal distribution with means of zero, variances of one, and a polynomial decay correlation matrix Σ given by $\Sigma_{\rho\rho'} = 0.5^{|\rho-\rho'|}$. We then transformed the W's into a vector of proportions $(P_{i1}, \ldots, P_{i5}) = \left(\frac{e^{W_{i1}}}{\sum_{j=1}^{5} e^{W_{ij}}+1}, \ldots, \frac{e^{W_{i5}}}{\sum_{j=1}^{5} e^{W_{ij}}+1}\right)$. For the ZILN, we randomly changed the observed proportions for each taxon to zero according to the zero-inflation levels sampled without replacement from $(0.1, 0.2, 0.4, 0.6, 0.8)$. The proportion of the sixth taxon was determined by $1 - \sum_{j=1}^{5} P_{ij}$. To evaluate power under the LN, we changed the Normal mean or variance for a randomly-picked taxon: its Normal mean parameter was sampled from Uniform(0, 1), and its Normal variance parameter was sampled from Uniform(1, 6). To evaluate power under the ZILN, we changed the Normal mean or variance for the taxon with a certain level of zero-inflation: its Normal mean parameter was sampled from Uniform(0, 2); its Normal variance parameter was sampled from Uniform(1, 8).

5.2. *Simulation results*

The empirical Type I errors for asymptotic and permutation tests are presented in Table 1. The Type I errors are properly controlled in the permutation tests. The asymptotic distribution-free tests QCAT_1 and QCAT_2 preserve the Type I error in all scenarios. The rest of the asymptotic tests tend to be too liberal under zero-inflated models (i.e. ZIGDM and ZILN models), especially when the data are not generated from the distribution the test assumes. The DM_2 asymptotic test is overly conservative under non-zero-inflated models (i.e. DM, GDM, and LN models). For a fair comparison, we only report permutation results in the power evaluation.

The power of permutation tests under non-zero-inflated models are presented in Table 2. The results for ZIGDM and GDM are identical under the LN model because almost all simulated taxa counts are positive. When the mean differs, the differential-mean tests (ZIGDM_1, GDM_1, DM_1, QCAT_1, and QCAT_2) outperform the differential-dispersion tests (ZIGDM_2, GDM_2). As the DM_2 jointly tests the mean and dispersion, we view it as both the differential-mean and differential-dispersion tests. When the dispersion differs, differential-dispersion tests are more powerful than their differential-mean counterparts. The differential-mean test QCAT_2 yields decent power for detecting differential dispersion if the change in dispersion also alters the percentages of zero counts in the data (see settings in the DM and GDM models).

Table 1. *Type I error of the tests*

	Model	n	$ZIGDM_1$	$ZIGDM_2$	GDM_1	GDM_2	DM_1	DM_2	$QCAT_1$	$QCAT_2$
Asymptotic	DM	100	0.058	0.041	0.064	0.057	0.051	0.012	0.048	0.042
		200	0.058	0.058	0.054	0.054	0.050	0.009	0.049	0.051
	GDM	100	0.047	0.033	0.060	0.057	0.042	0.022	0.042	0.028
		200	0.053	0.041	0.059	0.050	0.056	0.018	0.055	0.039
	LN	100	0.072	0.041	0.072	0.041	0.058	0.027	0.047	0.048
		200	0.057	0.037	0.057	0.037	0.055	0.024	0.047	0.049
	ZIGDM	100	0.110	0.120	0.130	0.099	0.061	0.100	0.046	0.043
		200	0.069	0.062	0.110	0.092	0.060	0.100	0.045	0.049
	ZILN	100	0.085	0.072	0.180	0.190	0.064	0.230	0.047	0.050
		200	0.057	0.049	0.180	0.190	0.062	0.210	0.044	0.050
Permutation	DM	100	0.048	0.048	0.054	0.050	0.051	0.050	0.050	0.049
		200	0.046	0.054	0.046	0.051	0.051	0.047	0.049	0.050
	GDM	100	0.046	0.046	0.046	0.043	0.042	0.050	0.047	0.050
		200	0.050	0.054	0.050	0.050	0.052	0.050	0.055	0.052
	LN	100	0.053	0.046	0.053	0.046	0.054	0.056	0.051	0.052
		200	0.052	0.047	0.052	0.047	0.046	0.049	0.047	0.050
	ZIGDM	100	0.041	0.052	0.048	0.056	0.051	0.047	0.054	0.046
		200	0.048	0.050	0.051	0.049	0.045	0.050	0.048	0.050
	ZILN	100	0.046	0.050	0.052	0.051	0.047	0.050	0.048	0.052
		200	0.048	0.049	0.049	0.049	0.049	0.049	0.044	0.048

Table 2. *Power of the permutation tests under non-zero-inflated models*

Model	Difference	n	$ZIGDM_1$	$ZIGDM_2$	GDM_1	GDM_2	DM_1	DM_2	$QCAT_1$	$QCAT_2$
DM	Mean	100	0.52	0.33	0.67	0.47	0.60	0.66	0.58	0.59
		200	0.67	0.54	0.76	0.62	0.71	0.76	0.70	0.72
	Disp	100	0.17	0.72	0.42	0.77	0.05	0.76	0.06	0.64
		200	0.52	0.84	0.60	0.84	0.05	0.81	0.05	0.75
GDM	Mean	100	0.60	0.27	0.66	0.34	0.59	0.60	0.60	0.59
		200	0.70	0.39	0.76	0.48	0.70	0.71	0.71	0.70
	Disp	100	0.17	0.47	0.56	0.79	0.07	0.55	0.07	0.59
		200	0.23	0.54	0.67	0.86	0.07	0.62	0.08	0.66
LN	Mean	100	0.48	0.06	0.48	0.06	0.50	0.51	0.52	0.52
		200	0.63	0.06	0.63	0.05	0.65	0.66	0.66	0.66
	Disp	100	0.05	0.70	0.05	0.70	0.24	0.39	0.17	0.18
		200	0.06	0.83	0.06	0.83	0.47	0.61	0.40	0.41

When the mean differs, the performances of differential-mean tests are largely similar. In contrast, when the dispersion differs, the performances of differential-dispersion tests are diverse: the GDM_2 is more powerful and robust than the $ZIGDM_2$ and DM_2 tests. In practice, we can use statistical model selection criteria (e.g. Bayesian information criterion, Akaike information criterion) to choose between ZIGDM and GDM models.

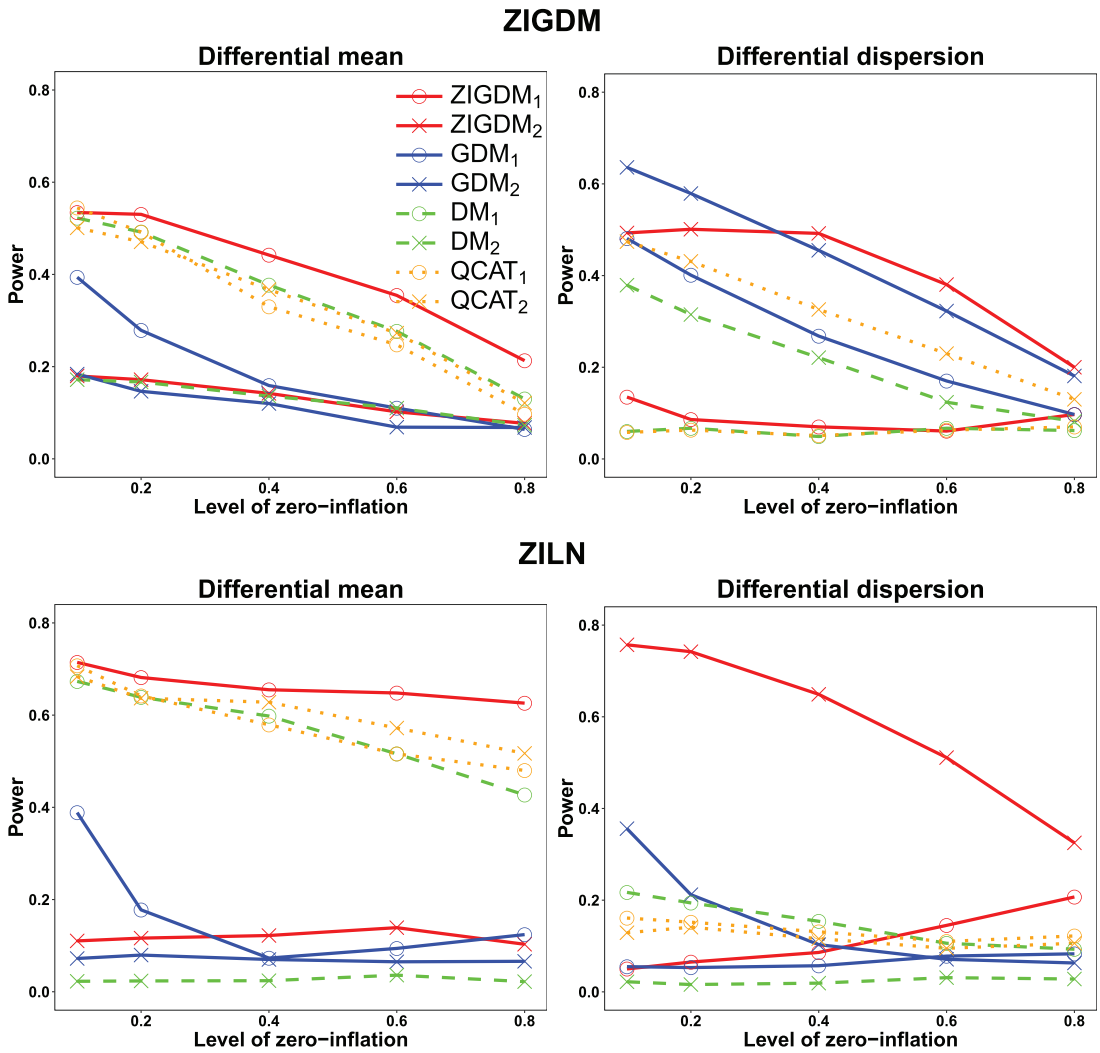

Fig. 1. Power of the permutation tests under ZIGDM and ZILN models when the sample size is 100. The pattern of variation is indicated above each graph.

The power of permutation tests under zero-inflated models when $n = 100$ are displayed in Figure 1 (see Figure 2 of supplementary material available at *Biostatistics* online for the results when $n = 200$). Each plot demonstrates the power curve as a function of the zero-inflation levels. When the mean differs, the ZIGDM_1 outperforms the other tests. When the dispersion differs, under the ZIGDM model, the ZIGDM_2, and GDM_2 perform substantially better than the other tests; under the ZILN model, the ZIGDM_2 is the most powerful test and the other tests have dramatic power loss.

In summary, comparing to the DM tests, the GDM and ZIGDM tests are more powerful to detect differential mean/dispersion and are more robust to the underlying distributions. If the data are zero-inflated, the ZIGDM tests are more desirable than the GDM tests. If the data are not zero-inflated, the GDM tests should be preferred over the ZIGDM tests.

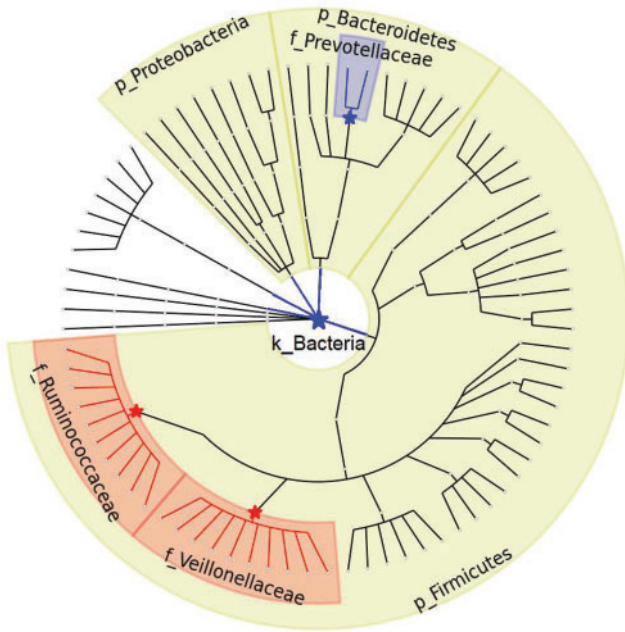

Fig. 2. Taxonomic tree for the gut microbiome data. The root nodes of the discovered lineages are marked with stars and annotated by the names of the taxa on the root nodes. The lineages under family *Prevotellaceae* and kingdom *Bacteria* were identified exclusively by the ZIGDM test. Three most abundant phyla (*Bacteroidetes*, *Firmicutes*, and *Proteobacteria*) are also annotated in the tree.

6. Gut microbiome and BMI

Gut microbiome plays an important role in obesity by contributing to nutrient digestion and absorption in humans. Wu *and others* (2011) investigated the relationship between micronutrients and gut microbiome composition. Fecal samples from 98 healthy volunteers were collected, along with their demographic data and diet information. The sample DNA was analyzed by sequencing the V1–V2 region of the 16S rRNA genes in the fecal samples. The sequencing reads were taxonomically classified to the genus level via QIIME (Caporaso *and others*, 2010). Taxa that appear in less than two samples were removed resulting in 80 genera.

These genera can be mapped to a taxonomic tree according to their taxonomic identities at the higher taxonomic levels. Figure 2 displays the tree with six levels from genus to kingdom. The outer circle of the tree represents lower taxonomic level. Each node on the tree represents a taxon. At each internal node, counts of the sequencing reads that assigned to that node are distributed to multiple taxa at the lower taxonomic levels (i.e. child nodes). As described in the previous literature (Shi and Li, 2017; Tang *and others*, 2017), in order to identify all differential lineages on the tree, we visited every internal node; for each node, we applied the tests to the sub-composition defined as the vector of taxon counts on its immediate child nodes. We then employed the Benjamini–Hochberg procedure (Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001) to control the false discovery rate (FDR). Alternative FDR control methods that respect the hierarchical structure of the taxonomy can potentially increase discovery power (Bogomolov *and others*, 2017; Lei *and others*, 2017). Testing sub-compositions on the lineage leverages taxonomic structure to define the unit of the tests in a biologically meaningful way and essentially reduces the dimension in each test. In our analysis, the number of taxa in each test is usually less than 10.

Because dysbiosis of the gut microbiome has been shown to be associated with obesity (Sanderson *and others*, 2006), we were interested in identifying differential bacterial lineages in high vs. normal BMI groups. We dichotomized the BMI value in our analysis because the current implementation of the DM tests can only handle the group-wise comparison (La Rosa *and others*, 2016). The BMI $\geq$25 is the commonly accepted range for overweight. We performed the ZIGDM, GDM, DM, and QCAT permutation tests to compare the mean or dispersion level of the microbial abundance between the high BMI ($\geq$25) and normal BMI ($<$25) groups over all lineages on the taxonomic tree at the family, order, class, phylum, and kingdom levels. The *P*-values for all the discovered lineages (FDR = 5%) are listed in Table 1 of supplementary material available at *Biostatistics* online. The DM and QCAT tests identify two BMI-associated lineages on the tree: family *Ruminococcaceae* (DM_1 *P*-value = 0.0013, $QCAT_1$ *P*-value = 0.00014) and family *Veillonellaceae* (DM_2 *P*-value = 0.0012, $QCAT_2$ *P*-value = 0.00080). Besides family *Veillonellaceae*, the GDM and ZIGDM tests identified another two BMI-associated lineages: family *Prevotellaceae* ($ZIGDM_2$ *P*-value = 0.0014 and GDM_2 *P*-value = 0.0014) and kingdom *Bacteria* ($ZIGDM_2$ *P*-value = 0.0016). The test for family *Prevotellaceae* involves two genera. The test for kingdom *Bacteria* involves eight phyla, among which the three most abundant taxa (phyla *Bacteroidetes*, *Firmicutes*, and *Proteobacteria*) dominate the community and the remaining five taxa are rare and the majority of their counts are zeros. In Figure 2, all the differential lineages and the three most abundant phyla are annotated in the tree. Figure 3 of supplementary material available at *Biostatistics* online displays the distributions of relative abundance for the three phyla and demonstrates significant dispersion differences between BMI groups in phyla *Bacteroidetes* and *Firmicutes*. The dispersion difference for taxa at such high taxonomic level is probably due to the aggregation of many low-level taxa that interact with each other and/or with environments. On the other hand, the difference in mean abundance would usually be difficult to detect at such high level because the differences are diluted by aggregating many taxa with no difference in mean.

Furthermore, we used the DM, GDM, and ZIGDM models to fit individual sub-compositions defined at internal nodes of the tree. In each model, we linked the mean, dispersion, and absence probability (in the ZIGDM model only) to the binary BMI and obtained the maximum likelihood estimates for the coefficients of the intercept and the binary BMI. We then generated synthetic data from each distribution with the estimated parameters. The GDM and ZIGDM models provide superior fit for most of the sub-compositions. For example, Figure 3 shows quantile-quantile plots of synthetic and observed relative abundances for the families *Prevotellaceae* and *Porphyromonadaceae* under order *Bacteroidales*. *Prevotellaceae* has many observed zeros (58%) and *Porphyromonadaceae* has very few (3.1%). The GDM and ZIGDM fit the data much better than the DM for these families and the ZIGDM has the best fit at the lower tail for *Prevotellaceae*. Another example for the sub-composition of three most abundant phyla under kingdom *Bacteria* is shown in Figure 4 of supplementary material available at *Biostatistics* online. Our analysis demonstrates that the additional parameters in GDM/ZIGDM are well-spent to provide better fit to the data, resulting in powerful association tests.

7. DISCUSSION

In this article, we develop the ZIGDM distribution for modeling microbiome compositional data with excess zeros, and propose score tests based on the ZIGDM regression model to detect differential mean or dispersion level of microbial composition. In contrast to the commonly-used DM model, the ZIGDM model provides a more flexible way of accommodating excess zeros and handling the complex correlation structure and dispersion patterns in taxon count data. We develop an EM algorithm to efficiently estimate parameters in the ZIGDM regression. The EM provides explicit forms of the posterior expectations in the E-step and updates the parameters for individual taxa independently in the M-step. Extensive simulation studies have been conducted to compare the proposed tests to existing ones. The results have demonstrated

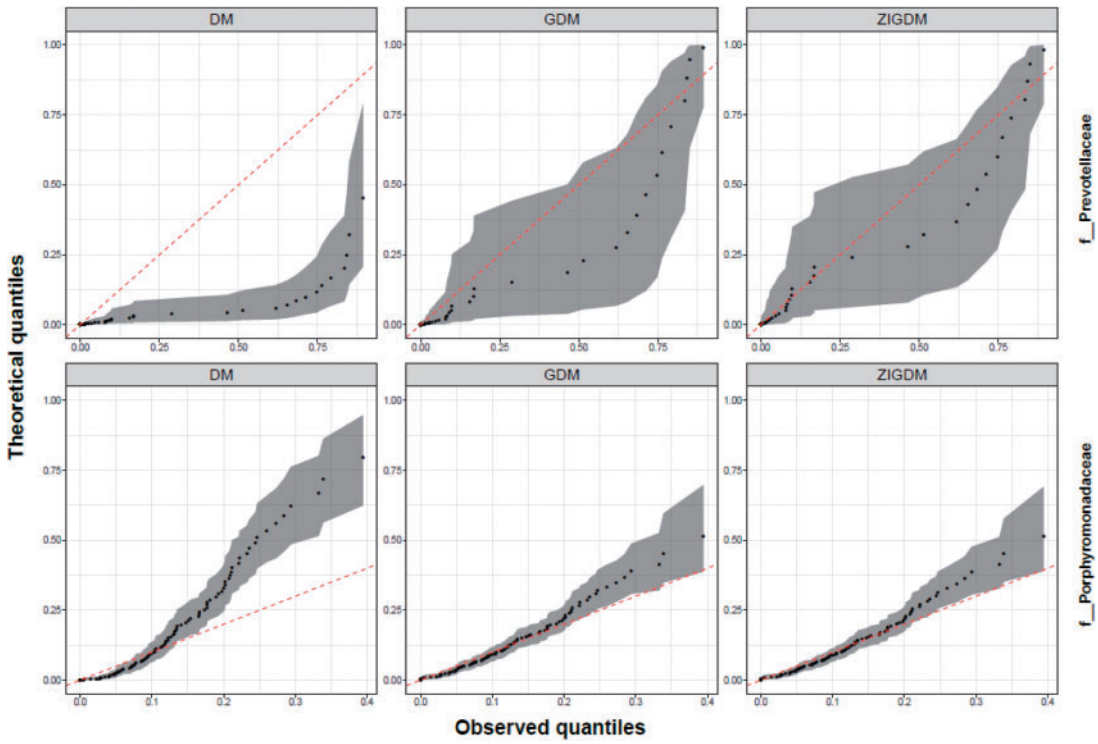

Fig. 3. Quantile-quantile plots from fitting three distributions to the sub-composition of two families under the order *Bacteroidales*. A thousand synthetic datasets were randomly generated from each distribution with the estimated parameters. The medians were used to draw points and the 2.5 and 97.5 percentiles were used to draw envelopes.

that the ZIGDM tests are more powerful to detect differential mean/dispersion and are more robust to the underlying distribution if the taxon counts are zero-inflated. If the taxon counts are not zero-inflated, the GDM tests are more desirable. In the analysis of a gut microbiome dataset, the proposed methods identify additional lineages with differential dispersion between BMI groups. We have demonstrated that the GDM provides a superior fit to taxon counts compared to the DM and the ZIGDM can further improve the goodness-of-fit for taxa with many zero counts.

We can potentially develop an omnibus test by combining the mean and dispersion tests. Moreover, besides testing the mean and dispersion, we can incorporate the test for differential presence-absence frequencies. However, the high degrees of freedom of the omnibus test will compromise its statistical power. To reduce the degree of the freedom, we can aggregate the abundances of multiple taxa or use the maximum statistic in the association test (Lin and Tang, 2011). Alternatively, we can construct a test statistic from the minimal P-value among tests for different features of the abundance distribution and use permutation to obtain a uniform P-value (Tang *and others*, 2016). The performance of these tests in analyzing microbiome data requires further study.

The multivariate association tests cannot handle high-dimensional microbial taxa, especially when the number of taxa is larger than the sample size. This similar problem has been investigated under the DM regression (Chen and Li, 2013; Wang and Zhao, 2017). The ZIGDM model is able to incorporate standard regularization approaches to deal with high dimensionality. Depending on the need in practice, the model can produce sparse estimates with the lasso penalty (Tibshirani, 1996) and the group lasso penalty

(Yuan and Lin, 2006). In addition, a phylogenetic structure-constrained penalty function can be used to incorporate important prior knowledge on evolutionary relationships among microbial taxa (Chen *and others*, 2012). These penalized ZIGDM regressions would enable us to identify microbes with differential mean, dispersion level, and presence-absence frequency.

8. SOFTWARE

R codes to implement the methods have been incorporated into the software miLineage, which is available at https://tangzheng1.github.io/tanglab/software.html.

SUPPLEMENTARY MATERIAL

Supplementary material is available online at http://biostatistics.oxfordjournals.org.

ACKNOWLEDGMENTS

We are grateful to the two anonymous reviewers for their helpful comments. We thank Dr. Hongzhe Li for providing the BMI data. *Conflict of Interest*: None declared.

REFERENCES

AHN, J., SINHA, R., PEI, Z., DOMINIANNI, C., WU, J., SHI, J., GOEDERT, J. J., HAYES, R. B. AND YANG, L. (2013). Human gut microbiome and risk for colorectal cancer. *Journal of the National Cancer Institute* **105**, 1907–1911.

ALEKSEYENKO, A. V., PEREZ-PEREZ, G. I., DE SOUZA, A., STROBER, B., GAO, Z., BIHAN, M., LI, K., METHÉ, B. A. AND BLASER, M. J. (2013). Community differentiation of the cutaneous microbiota in psoriasis. *Microbiome* **1**, 31.

BENJAMINI, Y. AND HOCHBERG, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **57**, 289–300.

BENJAMINI, Y. AND YEKUTIELI, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics* **29**, 1165–1188.

BOGOMOLOV, M., PETERSON, C. B., BENJAMINI, Y. AND SABATTI, C. (2017). Testing hypotheses on a tree: new error rates and controlling strategies. *arXiv preprint arXiv:1705.07529*.

CAPORASO, J. G., KUCZYNSKI, J., STOMBAUGH, J., BITTINGER, K., BUSHMAN, F. D., COSTELLO, E. K., FIERER, N., PENA, A. G., GOODRICH, J. K., GORDON, J. I. AND OTHERS. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* **7**, 335–336.

CHEN, J., BUSHMAN, F. D., LEWIS, J. D., WU, G. D. AND LI, H. (2012). Structure-constrained sparse canonical correlation analysis with an application to microbiome data analysis. *Biostatistics* **14**, 244–258.

CHEN, J. AND LI, H. (2013). Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis. *The Annals of Applied Statistics* **7**, 418–442.

CHO, I. AND BLASER, M. J. (2012). The human microbiome: at the interface of health and disease. *Nature Reviews Genetics* **13**, 260–270.

CHO, I., YAMANISHI, S., COX, L., METHÉ, B. A., ZAVADIL, J., LI, K., GAO, Z., MAHANA, D., RAJU, K., TEITLER, I. AND OTHERS. (2012). Antibiotics in early life alter the murine colonic microbiome and adiposity. *Nature* **488**, 621–626.

COLE, J. R., CHAI, B., FARRIS, R. J., WANG, Q., KULAM-SYED-MOHIDEEN, A. S., MCGARRELL, D. M., BANDELA, A. M., CARDENAS, E., GARRITY, G. M. AND TIEDJE, J. M. (2007). The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data. *Nucleic Acids Research* **35**, 169–172.

CONNOR, R. J. AND MOSIMANN, J. E. (1969). Concepts of independence for proportions with a generalization of the Dirichlet distribution. *Journal of the American Statistical Association* **64**, 194–206.

DESANTIS, T. Z., HUGENHOLTZ, P., LARSEN, N., ROJAS, M., BRODIE, E. L., KELLER, K., HUBER, T., DALEVI, D., HU, P. AND ANDERSEN, G. L. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and Environmental Microbiology* **72**, 5069–5072.

GILBERT, J. A., QUINN, R. A., DEBELIUS, J., XU, Z. Z., MORTON, J., GARG, N., JANSSON, J. K., DORRESTEIN, P. C. AND KNIGHT, R. (2016). Microbiome-wide association studies link dynamic microbial consortia to disease. *Nature* **535**, 94–103.

ISHWARAN, H. AND JAMES, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association* **96**, 161–173.

JOVEL, J., PATTERSON, J., WANG, W., HOTTE, N., O'KEEFE, S., MITCHEL, T., PERRY, T., KAO, D., MASON, A. L., MADSEN, K. L. AND OTHERS. (2016). Characterization of the gut microbiome using 16S or shotgun metagenomics. *Frontiers in Microbiology* **7**, 459.

KUCZYNSKI, J., LAUBER, C. L., WALTERS, W. A., PARFREY, L. W., CLEMENTE, J. C., GEVERS, D. AND KNIGHT, R. (2012). Experimental and analytical tools for studying the human microbiome. *Nature Reviews Genetics* **13**, 47–58.

LA ROSA, P. S., BROOKS, J. P., DEYCH, E., BOONE, E. L., EDWARDS, D. J., WANG, Q., SODERGREN, E., WEINSTOCK, G. AND SHANNON, W. D. (2012). Hypothesis testing and power calculations for taxonomic-based human microbiome data. *PLoS One* **7**, e52078.

LA ROSA, P. S., DEYCH, E., SHANDS, B. AND SHANNON, W. D. (2016). *HMP: Hypothesis Testing and Power Calculations for Comparing Metagenomic Samples from HMP*. R package version 1.4.3, https://CRAN.R-project.org/package=HMP.

LEI, L., RAMDAS, A. AND FITHIAN, W. (2017). Star: a general interactive framework for FDR control under structural constraints. *arXiv preprint arXiv:1710.02776*.

LI, H. (2015). Microbiome, metagenomics, and high-dimensional compositional data analysis. *Annual Review of Statistics and Its Application* **2**, 73–94.

LIN, D.-Y. AND TANG, Z.-Z. (2011). A general framework for detecting disease associations with rare variants in sequencing studies. *The American Journal of Human Genetics* **89**, 354–367.

LIU, Z., DESANTIS, T. Z., ANDERSEN, G. L. AND KNIGHT, R. (2008). Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. *Nucleic Acids Research* **36**, e120.

MANDAL, S., VAN TREUREN, W., WHITE, R. A., EGGESBØ, M., KNIGHT, R. AND PEDDADA, S. D. (2015). Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microbial Ecology in Health and Disease* **26**, 27663.

O'BRIEN, J. D., RECORD, N. AND COUNTWAY, P. (2016). The power and pitfalls of Dirichlet-multinomial mixture models for ecological count data. *bioRxiv*. https://doi.org/10.1101/045468.

QIN, J., LI, Y., CAI, Z., LI, S., ZHU, J., ZHANG, F., LIANG, S., ZHANG, W., GUAN, Y., SHEN, D. AND OTHERS. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490**, 55–60.

SANDERSON, S., BOARDMAN, W., CIOFI, C. AND GIBSON, R. (2006). Human gut microbes associated with obesity. *Nature* **444**, 1022–1023.

SANKARAN, K. AND HOLMES, S. (2017). Latent variable modeling for the microbiome. *arXiv* preprint arXiv: 1706.04969.

SHI, P. AND LI, H. (2017). A model for paired-multinomial data and its application to analysis of data on a taxonomic tree. *Biometrics* **73**, 1266–1278.

TANG, Z.-Z., CHEN, G. AND ALEKSEYENKO, A. V. (2016). PERMANOVA-S: association test for microbial community composition that accommodates confounders and multiple distances. *Bioinformatics* **32**, 2618–2625.

TANG, Z.-Z., CHEN, G., ALEKSEYENKO, A. V. AND LI, H. (2017). A general framework for association analysis of microbial communities on a taxonomic tree. *Bioinformatics* **33**, 1278–1285.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **58**, 267–288.

WADSWORTH, W. D., ARGIENTO, R., GUINDANI, M., GALLOWAY-PENA, J., SHELBOURNE, S. A. AND VANNUCCI, M. (2017). An integrative Bayesian Dirichlet-multinomial regression model for the analysis of taxonomic abundances in microbiome data. *BMC Bioinformatics* **18**, 94.

WANG, T. AND ZHAO, H. (2017). A Dirichlet-tree multinomial regression model for associating dietary nutrients with gut microorganisms. *Biometrics* **73**, 792–801.

WONG, T.-T. (1998). Generalized Dirichlet distribution in Bayesian analysis. *Applied Mathematics and Computation* **97**, 165–181.

WU, G. D., CHEN, J., HOFFMANN, C., BITTINGER, K., CHEN, Y.-Y., KEILBAUGH, S. A., BEWTRA, M., KNIGHTS, D., WALTERS, W. A., KNIGHT, R. AND OTHERS. (2011). Linking long-term dietary patterns with gut microbial enterotypes. *Science* **334**, 105–108.

YUAN, M. AND LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**, 49–67.

ZHANG, Y., ZHOU, H., ZHOU, J. AND SUN, W. (2017). Regression models for multivariate count data. *Journal of Computational and Graphical Statistics* **26**, 1–13.

[*Received December 20, 2017; revised April 26, 2018; accepted for publication May 6, 2018*]