Communications in Information and Systems
Volume 18, Number 3, 173–191, 2018

A review on probabilistic models used in microbiome studies

Ahmed A. Metwally*, Hani Aldirawi*, and Jie Yang

In this paper, we first briefly review the background and significance of the microbiome, the technologies used for collecting microbiome data, and some public resources for downloading microbiome data. We then review the probabilistic models used in the literature in two categories: (1) for read counts from a specific feature, including Poisson, negative binomial, zero-inflated and hurdle models; (2) for read counts from multiple features, including Dirichlet-multinomial, generalized Dirichlet-multinomial, and zero-inflated models, as well as a nonparametric Bayesian model for a flexible number of features. We also review comprehensive comparisons among different probabilistic models.

1. Introduction

The microbiome, a dynamic ecosystem of microorganisms (bacteria, archaea, fungi, and viruses) that live in and on us, plays a vital role in host-immune responses resulting in significant effects on host health. Dysbiosis of the microbiome has been linked to diseases including asthma, obesity, diabetes, transplant rejection, and inflammatory bowel disease [7, 35, 37, 45, 53]. These observations suggest that modulation of the microbiome could become an important therapeutic modality for some diseases. For example, fecal transplants have been shown to alleviate diarrhea caused by *Clostridium difficile* infection and temporarily improve insulin sensitivity [16, 54]. Specifically, the gut microbiome, which has been the most extensively studied human microbiome ecosystem, is highly diverse and has been shown to include thousands of different bacterial species [10, 59]. This diverse community of bacteria is composed of a few species that are highly abundant and a large number of species that are found in trace amount [27]. The human

*These two authors contributed to this paper equally.

This work was partially supported by a UIC Chancellor's Graduate Research Fellowship, and UIC CCTS Pre-doctoral Education for Clinical and Translational Scientists fellowship (UL1TR002003), both awarded to AAM.

microbiome can be divided into the core microbiome and the variable microbiome [52]. The core microbiome is the set of taxa or genes that present in a given body location (gut, kidney, skin, oral, etc.) in almost all humans. The variable microbiome arises from various factors such as host physiological status, host environment, host genotype, host lifestyle, and host pathobiology. Moreover, given the strong association between microbiome and various diseases, computational models have been built to predict phenotypes from microbial profiles [9, 17, 40, 41].

The goal of this paper is to briefly introduce the microbiome data and the related probabilistic models to people who are interested in microbiome research and the corresponding analysis. In Section 2, we introduce the two technologies, 16S rRNA and metagenome shotgun sequencing, for obtaining microbial profiles of a biological sample. In Section 3, we introduce the typical format of microbiome data and some public resources for downloading microbiome data. In Section 4, we review the probabilistic models for modeling count data from each microbial feature independently. In Section 5, we review the probabilistic models for microbiome data from multiple features. In Section 6, we review a nonparametric Bayesian model for a flexible number of microbiome features. We conclude in Section 7.

2. Technologies used to study the microbiome

The number of studies investigating the microbiome has risen exponentially since the technological advances in high-throughput sequencing [19]. Sequencing technologies are able to identify the genetic content of microbial communities in the form of millions to billions of short DNA sequences. These technical advances have been paradigm shifting since the majority (>90%) of microbial species cannot be readily cultured using current laboratory culture techniques [32, 39, 48]. The most common sequencing technologies to analyze the microbiome are 16S rRNA gene sequencing and metagenome shotgun (MGS) sequencing [38].

In 16S rRNA sequencing, a 16S rRNA gene is amplified by polymerase chain reaction (PCR) with primers that recognize the highly conserved regions of the gene [44]. A limitation of this method is that the annotation is based on a putative association of the 16S rRNA gene with taxa defined as an operational taxonomic unit (OTU). In general, OTUs are annotated at higher levels, such as phylum to genus, and can be less precise at the species level. In 16S rRNA sequencing, other bacterial genes are not directly sequenced, but rather predicted based on the OTUs [22]. Due to horizontal gene transfer and the existence of numerous bacterial strains [18, 36],

the lack of direct gene identification potentially limits understanding of the microbiome.

An alternative sequencing approach is the metagenome shotgun (MGS) sequencing in which random fragments of a genome are sequenced. Compared to 16S rRNA, MGS is more expensive and requires more sophisticated data analysis methods [19, 24, 25, 47]. A major advantage of the MGS sequencing is that the sequence reads can be more accurately annotated at the species level and the microbial functional profile can be constructed from the sequence reads.

3. Microbiome data resources and data representation

There are several initiatives to store and manage data from microbiome studies in order to make them available and free for everyone to use. The major public servers are MG-RAST (`https://www.mg-rast.org/`) and QIITA (`https://qiita.ucsd.edu/`). Also, National Center of Biotechnology Information (NCBI, `https://www.ncbi.nlm.nih.gov/`) is one of the most comprehensive resources that have a database of curated and updated microbial genomes and taxonomic tree.

To analyze these massive amount of sequence data, metagenomic reads are processed for each microbiome sample to construct taxonomic and/or functional profiles [1, 2, 20, 29, 51, 56]. The taxonomic profiles, functional profiles, or both for all samples, are then combined into one count table (see Table 1 and Table 2 as an example of a toy taxonomic profile) with a dimension of $m \times n$, where m denotes the number of microbial features $F_1, \ldots, F_m$ and n denotes the number of metagenomic samples $S_1, \ldots, S_n$. The entry z_{ij} represents the number of reads from sample j that mapped to microbial feature i, while its capitalized version Z_{ij} represents the corresponding random variable. In the table, $N_j = \sum_{i=1}^{m} z_{ij}$ denotes the total number of reads for the m features in sample S_j, and $z_{i\cdot} = \sum_{j=1}^{n} z_{ij}$ denotes the total number of reads mapped to features F_i in all samples. Since metagenomic samples may have different sequencing depths, the aggregated metagenomic counts need to be normalized among samples [3]. There are several methods developed to tackle the normalization problem of a count table, such as centered log-ratio (CLR) transformation [11], cumulative sum scaling [33], median-of-ratios scaling factor [23], and trimmed mean of M values [43].

Feature/Sample	S_1	S_2	S_3	...	S_n	Total
F_1	z_{11}	z_{12}	z_{13}	...	z_{1n}	$z_{1.}$
F_2	z_{21}	z_{22}	z_{23}	...	z_{2n}	$z_{2.}$
...	...	...	...	...	...	...
F_m	z_{m1}	z_{m2}	z_{m3}	...	z_{mn}	$z_{m.}$
Total	N_1	N_2	N_3	...	N_n	$N_.$

Table 1: A Typical Microbial Count Table

Species/Sample	S_1	S_2	S_3	Total
Escherichia coli	587	102	3	692
Staphylococcus aureus	980	324	75	1379
Streptococcus pneumoniae	14	0	0	14
Total	1581	426	78	2085

Table 2: An Example of Taxonomic Profile Count Table

4. Probabilistic models for single feature

One of the objectives of the microbiome studies is to determine whether there is a particular microbial signature (e.g., taxa or genes) associated with a particular disease state and/or phenotype. These biomarkers can play an important role in the development of preventive and therapeutic strategies. Differential abundance tests have been developed to identify those microbial features that are significantly different between two phenotypic groups. In this section, we focus on probabilistic models built for sequence read counts from a single microbiome feature, that is, Z_{ij} for feature F_i and subject S_j. Assuming Z_{ij} follows a probabilistic model with a few unknown parameters, statistical inference can be made based on estimated parameters from the data. In practice, there are two types of experimental designs for microbiome studies: (1) snapshot studies, where each subject provides only one sample, (2) longitudinal studies, which include multiple samples per subject over time.

4.1. Models used in snapshot microbiome studies

4.1.1. Poisson model. Poisson distribution has been widely used for modeling non-negative outcomes as a count. If a random feature count Z_{ij} follows a Poisson distribution with mean $\theta > 0$, it assigns the probability

$$P(Z_{ij} = k) = \frac{\theta^k}{k!} e^{-\theta}$$

for $k = 0, 1, 2, \ldots$. As the mean count increases, the skewness diminishes, and the Poisson distribution becomes approximately a normal distribution. One property of Poisson distribution is that its variance equals the mean. For non-negative count outcomes, a model with Poisson distribution is usually more appropriate than an ordinary least squares linear model [4].

4.1.2. Negative binomial model. The negative binomial (NB) distribution is an alternative probabilistic model for count data [4]. It is especially useful when the sample variance exceeds the sample mean, known as overdispersion. Given a sequence of independent Bernoulli trials with probability p of success, Z_{ij} is the number of failures observed before the r^{th} success with the probability

$$P(Z_{ij} = k) = \binom{k + r - 1}{k} p^r (1 - p)^k$$

where $r > 0$ and $0 \leqslant p \leqslant 1$ are two parameters that can be estimated from the data. The negative binomial distribution looks similar to Poisson distribution but with a longer, fatter tail. If the observed outcome is suspected to have variance larger than mean, the negative binomial distribution would be more appropriate than either Poisson or normal distribution.

4.1.3. Zero-inflated models. For microbiome OTU counts, typically there are much more zeros than expected under the assumption of Poisson or negative binomial distributions. This phenomenon is known as zero-inflation. In order to solve this issue, zero-inflated models are used to model read counts that have an excess of zeros. A zero-inflated model assumes that the observed zeros are of two kinds: "sampling" or "structural". The sampling zeros come from a Poisson, negative binomial, or some other distribution due to chance. Other observed zeros are due to some specific structure in the data [14]. As a result, the combined probability under a zero-inflated

model is

$$P_{\mathrm{ZI}}(Z_{ij} = k) = \phi \mathbf{1}_{\{k=0\}} + (1 - \phi) P(Z_{ij} = k) \tag{1}$$

where $\phi > 0$ is a parameter estimated from the data, $P(Z_{ij} = k)$ stands for the probability determined by a Poisson, negative binomial, or other parametric distribution. Note that the zero-inflated model assigns the probability $\phi + (1 - \phi)P(Z_{ij} = 0)$ to zero, which is larger than $P(Z_{ij} = 0)$ itself. The corresponding distributions are known as zero-inflated Poisson (ZIP), zero-inflated negative binomial (ZINB), zero-inflated beta binomial (ZIBB, see Section 4.1.4), zero-inflated Gaussian (ZIG) distributions, etc. For example, Taylor, *et al.* [50] used a ZIP model for microbiome OTU counts.

4.1.4. Zero-inflated beta binomial model. As a special kind of zero-inflated models introduced in Section 4.1.3, the zero-inflated beta binomial (ZIBB) model provides a flexible option for modeling Z_{ij} [15]. In a ZIBB model, the probability $P(Z_{ij} = k)$ in Equation (1) is formulated by a beta-binomial distribution. It has two folds: (1) Given a probability p_{ij}, Z_{ij} follows a binomial distribution with parameters N_j and p_{ij}; (2) In order to make the model flexible, the probability p_{ij} itself is also random, which follows a beta distribution with parameters $\alpha_1 > 0$, $\alpha_2 > 0$. As a result, the probability based on the beta-binomial distribution is

$$P(Z_{ij} = k) = \binom{N_j}{k} \frac{\mathrm{Beta}(k + \alpha_1, N_j - k + \alpha_2)}{\mathrm{Beta}(\alpha_1, \alpha_2)} \tag{2}$$

The probability $P_{\mathrm{ZI}}(Z_{ij} = k)$ based on the ZIBB model takes the same form as in Equation (1).

A relevant R package *ZIBBSeqDiscovery* is available at the Comprehensive R Archive Network (CRAN, `https://CRAN.R-project.org/package=ZIBBSeqDiscovery`). Hu, *et al.* [15] compared ZIBB with ZINB and a few other models using the Gevers microbiome data and concluded that ZIBB shows the highest number of significantly enriched genera.

4.1.5. Hurdle models. Hurdle models, also known as zero-altered models, provide another way of dealing with the excess zeros in OTU counts [4]. A hurdle model consists of two components, one generating the zeros and one generating the positive values. In contrast to zero-inflated models, a hurdle model assumes that all zeros are from the "structural" source. In order to make the comparison clearly, we define the hurdle models using a

similar formula as in Equation (1):

$$P_{\mathrm{ZA}}(Z_{ij}=k)=\phi\mathbf{1}_{\{k=0\}}+(1-\phi)P_{\mathrm{tr}}(Z_{ij}=k) \tag{3}$$

where $P_{\mathrm{tr}}(Z_{ij}=k)$ is a truncated version of $P(Z_{ij}=k)$ determined by $P_{\mathrm{tr}}(Z_{ij}=0)=0$ and $P_{\mathrm{tr}}(Z_{ij}=k)=P(Z_{ij}=k)/[1-P(Z_{ij}=0)]$ for $k>0$. For example, if $P(Z_{ij}=k)$ comes from a Poisson distribution, then $P_{\mathrm{tr}}(Z_{ij}=k)$ is known as a zero-truncated Poisson distribution [58].

The hurdle model $P_{\mathrm{ZA}}(Z_{ij}=k)$ collapses to the standard model $P(Z_{ij}=k)$ if $\phi=P(Z_{ij}=0)$. It clearly allows for excess zeros when $\phi>P(Z_{ij}=0)$. Different from zero-inflated models, in principle, hurdle models can also model too few zeros when $\phi<P(Z_{ij}=0)$. In other words, hurdle models are more flexible than zero-inflated models.

Similar to zero-inflated models, hurdle models include zero-altered Poisson (ZAP) or Poisson hurdle (PH), zero-altered negative binomial (ZANB) or negative binomial hurdle (NBH) models, etc.

4.1.6. Comparisons between probabilistic models used in snapshot studies. Xu, *et al.* [57] classified the competing methods for modeling microbiome data into three categories based on how the excess zeros are treated: standard, zero-inflated (ZI), and hurdle models. Standard models do not consider the excess zeros and model the data using a standard distribution, for examples, Poisson or negative binomial distributions. ZI and hurdle models are reviewed in Sections 4.1.3 and 4.1.5, respectively. Xu, *et al.* [57] compared the performance of different models, including Poisson, ZIP, PH, NB, ZINB, and NBH, through simulation studies and real microbiome data. Their comparison was from different perspectives, including type I error, power, model selection, and goodness of fit. They concluded that: (1) Poisson regression has inflated type I error and may not be appropriate for data with excess zeros; (2) ZI or hurdle models perform consistently well in all scenarios examined in terms of test power; (3) ZINB is more robust than ZIP; (4) In many situations, hurdle models (PH or NBH) produce identical fitting results as their corresponding ZI models (ZIP or ZINB); (5) In terms of Akaike Information Criterion (AIC), NBH and ZINB models perform the best among all fitted models.

4.2. Models used in longitudinal microbiome studies

The recent advances in DNA sequencing technologies and rapid reduction in costs have fostered longitudinal analyses, which include multiple samples per

subject over time. These longitudinal studies provide increased insights into the underlying biological mechanisms of the microbiome role in health and disease. In addition to identifying differentially abundant features, detecting the time intervals where these features exhibit changes in their abundance between two phenotypes in longitudinal studies adds insights into disease pathogenesis.

Longitudinal differential abundance is to identify time intervals of differentially abundant microbial features. To date, three methods have been proposed, *MetaSplines* [34], *MetaDprof* [26], and *MetaLonDA* [30, 31]. *MetaSplines* and *MetaDprof* are both based on the Gaussian Smoothing Spline ANOVA (SS-ANOVA) approach [12, 13, 55]. *MetaSplines* has a higher sensitivity of detecting time intervals of differentially abundant features than *MetaDprof*, while *MetaDprof* has higher specificity [26, 31]. *MetaDprof* has a major drawback in its implementation since it assumes consistency in longitudinal microbial samples. It is only able to perform the analysis on an equivalent number of subjects per phenotypic group, the same number of samples from each subject, and the same elapsed time between adjacent time points, which are rarely fulfilled in human microbiome longitudinal studies.

MetaLonDA (Metagenomic Longitudinal Differential Abundant) is a recently developed method that can identify significant time intervals of differentially abundant microbial features such as taxonomies, genes, or pathways. *MetaLonDA* is flexible such that it can perform differential abundance tests on longitudinal samples with different numbers of subjects per phenotypic group, different numbers of samples per subject, and samples that are not collected at consistent time points. These inconsistencies are often the case for samples collected from human subjects in translational studies. Inconsistencies increase with the complexity of the procedure utilized to obtain the samples.

MetaLonDA relies on two modeling components: the NB distribution for modeling the mapped read counts for each feature and the semi-parametric SS-ANOVA technique for modeling longitudinal profiles associated with different phenotypes. Specific significant time intervals of microbial features can then be utilized to establish targeted timely screening or prevention of individual features and facilitate timely interventions, such as the use of antibiotics or probiotics. Unlike with cross-sectional methods that are incapable of identifying significant time intervals associated with differentially abundant features, significant time intervals of differentially abundant features identified through *MetaLonDA* may lead to the reconstitution of the microbiome and reestablishment of homeostasis prior to the onset of

overt disease. *MetaLonDA* is publicly available on the CRAN repository (`https://CRAN.R-project.org/package=MetaLonDA`).

5. Probabilistic models for a group of features in the count table

Modeling multivariate feature counts is becoming important in the microbiome research community because these models are able to retain more information contained in the data. For example, researchers are interested in testing multivariate hypotheses concerning the effects of treatments or experimental factors on the whole assemblages of bacterial taxa, so that they may know the impact of the microbiome on human health and on characterizing the microbial diversity in general [21]. In addition, they may be able to identify relevant microbes that play essential roles in a microbial network by connecting a variety of key features.

5.1. Multinomial and Dirichlet multinomial (DM) models

Suppose we have m bacterial taxa, and their counts $\mathbf{Z}_j = (Z_{1j}, \ldots, Z_{mj})^T$ from the j^{th} subject. With a given column sum N_j and a probability measure $\mathbf{p} = (p_1, \ldots, p_m)^T$, a multinomial distribution assigns the following probability to an observed column $\mathbf{z}_j = (z_{1j}, \ldots, z_{mj})^T$ of the OTU table

$$P(\mathbf{Z}_j = \mathbf{z}_j) = \frac{N_j!}{z_{1j}! \cdots z_{mj}!} p_1^{z_{1j}} \cdots p_m^{z_{mj}} \tag{4}$$

where $p_i \geq 0$, $\sum_{i=1}^m p_i = 1$, and $\sum_{i=1}^m z_{ij} = N_j$.

In order to make the model more flexible and thus can fit the data better, the probability vector $\mathbf{p} = (p_1, \ldots, p_m)^T$ is assumed to follow a Dirichlet distribution with a probability density function

$$f(\mathbf{p}) = \frac{\Gamma(\sum_{i=1}^m \gamma_i)}{\Gamma(\gamma_1) \cdots \Gamma(\gamma_m)} p_1^{\gamma_1 - 1} \cdots p_m^{\gamma_m - 1} \tag{5}$$

where $\gamma_i > 0$, $i = 1, \ldots, m$ are parameters of the Dirichlet distribution.

The aggregated distribution is known as a Dirichlet multinomial (DM) distribution, which assigns the following probability

$$P_{DM}(\mathbf{Z}_j = \mathbf{z}_j) = \frac{N_j!\, \Gamma(\sum_{i=1}^m \gamma_i)}{\Gamma(N_j + \sum_{i=1}^m \gamma_i)} \prod_{i=1}^m \frac{\Gamma(z_{ij} + \gamma_i)}{z_{ij}!\, \Gamma(\gamma_i)} \tag{6}$$

The DM distribution is commonly used to model taxon counts. For example, Rosa, *et al.* [21] used DM to calculate the powers and sample sizes for experimental designs, perform tests of hypotheses (for example, comparison of microbiomes across groups), and estimate parameters describing microbiome properties. Chen and Li [6] used a DM model for developing a penalized likelihood approach to estimate the regression parameters. Nevertheless, several recent studies showed that a multinomial or DM model might not be appropriate for microbiome data [28, 46] since those models assume a negative correlation among all paired OTUs. Actually, Mandal, *et al.* [28] showed that the correlation between a pair of OTUs could be positive as well.

5.2. Generalized Dirichlet multinomial (GDM) and zero-inflated generalized Dirichlet multinomial (ZIGDM) models

Dirichlet distribution has been widely used as a conjugate prior for the parameters of a multinomial distribution since the calculation on its posterior distribution is easier. Nevertheless, if a random vector follows a Dirichlet distribution, all its components must have the same variance and sum to one. Motivated by this limitation, a generalized Dirichlet (GD) distribution was introduced by Connor and Mosimann [8] to allow more general covariance structure. The multinomial model with a GD prior is called a generalized Dirichlet multinomial (GDM) model [49].

Under the GDM model, the j^{th} column $\mathbf{Z}_j$ of OTUs with m observed features $(Z_{1j}, \ldots, Z_{mj})$ is modeled as $m+1$ features with one additional unobserved feature $Z_{m+1,j}$. Let the extended column vector $\mathbf{Z}_j^+ = (Z_{1j}, \cdots, Z_{mj}, Z_{m+1,j})^T$ with $\sum_{i=1}^{m} Z_{ij} = N_j$ being the total number of reads in the original j^{th} column. The corresponding proportions for the $m+1$ features are $p_1, \ldots, p_m, p_{m+1}$ with $\sum_{i=1}^{m+1} p_i = 1$. The random parameter vector $\mathbf{p} = (p_1, \ldots, p_m)^T$ follows a GD distribution with the density function

$$f(\mathbf{p}) = \prod_{j=1}^{m} \frac{1}{\text{Beta}(\alpha_j, \beta_j)} p_j^{\alpha_j - 1} (1 - p_1 - \cdots - p_j)^{c_j} \tag{7}$$

where $\alpha_j > 0$, $\beta_j > 0$, $c_j = \beta_j - \alpha_{j+1} - \beta_{j+1}$ for $j = 1, \ldots, m-1$ and $c_m = \beta_m - 1$ (see [49] for more details).

The zero-inflated generalized Dirichlet (ZIGD) distribution can be produced by adding a zero-inflated component to the GD distribution [49], and the zero-inflated generalized Dirichlet (ZIGDM) model is constructed by using the ZIGD as a prior for the multinomial parameters. Tang and Chen [49]

used real gut microbiome data to compare three models, DM, GDM, and ZIGDM. Their conclusions include: (1) ZIGDM is more flexible for handling the excess zeros and a complex correlation structure; (2) ZIGDM is more robust when the counts are zero-inflated; (3) GDM fits the data better than DM, while ZIGDM can further improve the goodness of fit for zero-inflated counts; (4) ZIGDM seems more appropriate for handling high-dimensional microbiome taxa.

6. A nonparametric Bayesian model for microbiome data

In a typical OTU count table, the number of rows m is fixed. However, from the collection procedure of microbiome data, the number of potential features is typically unknown, and the number of observed features by the experiments is not predetermined.

Following a nonparametric Bayesian model [42], we may denote $\mathcal{F} = \{F_1, F_2, \ldots\}$ as the collection of all potential features in the n biological samples. Different from the models in Section 5, the number of features in $\mathcal{F}$ could be infinity. For biological sample j, a discrete probability measure $P^j = (p_{1j}, p_{2j}, \ldots)$ on $\mathcal{F}$ determines the distribution of the frequencies $(Z_{1j}, Z_{2j}, \ldots)$. In other words, $P^j(\{F_i\}) = p_{ij}$, $\sum_i p_{ij} = 1$, and $E(Z_{ij}) \propto p_{ij}$.

The goal of the nonparametric Bayesian model [42] is to model the distribution of P^j's and the variation among them. It assigns the probability mass to any subset A of the potential features in $\mathcal{F}$:

$$P^j(A) = \frac{M^j(A)}{M^j(\mathcal{F})} \tag{8}$$

where $M^j(A) = \sum_{i=1}^{\infty} \mathbf{1}_{(F_i \in A)} \sigma_i \langle \mathbf{Z}_{(i)}, \mathbf{Z}_j \rangle^{+2}$, $\sigma_i \in (0, 1)$ is the average abundance of feature F_i across all biological samples, $\mathbf{Z}_{(i)}$ is the row of OTU table associated with feature F_i, $\mathbf{Z}_j = (Z_{1j}, Z_{2j}, \ldots)^T$ is the j^{th} column, $\langle \cdot \rangle$ stands for the usual inner product (see [42] for more details).

Ren, *et al.* [42] constructed a Bayesian framework with dependent Dirichlet process for microbiome analysis. They applied their model to two microbiome datasets. Based on their Bayesian nonparametric method, they obtained the posterior probabilities of any two biological samples being clustered together. Their result are consistent with Caporaso, *et al.* [5]'s conclusion.

7. Conclusion

In this paper, we review three types of probabilistic models based on the number of features under discussion, single, a fixed number, and a flexible number of features. Since the microbiome data is often sparse, it may contain much more than expected zeros, which makes the construction of a suitable probabilistic model challenging. Zero-inflated or hurdle model seems more appropriate than a standard model. Among different models for analyzing the differential abundance of a single OTU, ZINB was recommended by Xu, *et al.* [57], while ZIBB was preferred by Hu, *et al.* [15].

Moreover, we may model multiple OTUs or features simultaneously. This kind of models may foster our understanding of interactions between species, or even build a network among species. Quite a few models have been proposed for this purpose, including multinomial, Dirichlet multinomial (DM), generalized Dirichlet multinomial (GDM), and zero-inflated generalized Dirichlet multinomial (ZIGDM) models. Although multinomial and Dirichlet multinomial (DM) were commonly used to model multiple OTUs, several recent studies suggested that those two models may not be appropriate for microbiome data [28, 46] due to their negative correlations between all paired OTUs. The comparison study was done by Tang and Chen [49] concluded that ZIGDM is more appropriate than the others.

On the other hand, nonparametric Bayesian approach for modeling microbiome data has not been commonly used yet. Recently, Ren, *et al.* [42]'s work provided a nonparametric Bayesian framework for modeling a flexible number of OTUs. It requires more complicated probabilistic models, such as dependent Dirichlet process.

References

[1] S. Abubucker, N. Segata, J. Goll, A. M. Schubert, J. Izard, B. L. Cantarel, B. Rodriguez-Mueller, J. Zucker, M. Thiagarajan, B. Henrissat, et al., *Metabolic reconstruction for metagenomic data and its application to the human microbiome*, PLoS computational biology **8** (2012), no. 6, e1002358.

[2] A. Asem and A. A. Metwally, *Cloud-based solution for improving usability and interactivity of metagenomic ensemble taxonomic classification methods*, IEEE EMBS International Conference on Biomedical and Health Informatics (BHI) (2018), 198–201.

[3] J. P. Brooks, D. J. Edwards, M. D. Harwich, M. C. Rivera, J. M. Fettweis, M. G. Serrano, R. A. Reris, N. U. Sheth, B. Huang, P. Girerd, J. F. Strauss, K. K. Jefferson, and G. A. Buck, *The truth about metagenomics: Quantifying and counteracting bias in 16S rRNA studies Ecological and evolutionary microbiology*, BMC Microbiology **15** (2015), no. 1, 1–14.

[4] A. C. Cameron, Regression Analysis of Count Data, Cambridge University Press (2013), ISBN 9781107014169.

[5] J. G. Caporaso, C. L. Lauber, W. A. Walters, D. Berg-Lyons, C. A. Lozupone, P. J. Turnbaugh, N. Fierer, and R. Knight, *Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample*, Proceedings of the National Academy of Sciences **108** (2011), no. Supplement_1, 4516–4522.

[6] J. Chen and H. Li, *Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis*, Annals of Applied Statistics **7** (2013), no. 1, 418–442.

[7] I. Cho and M. J. Blaser, *The human microbiome: At the interface of health and disease*, Nature Reviews Genetics **13** (2012), no. 4, 260–270.

[8] R. J. Connor and J. E. Mosimann, *Concepts of Independence for Proportions with a Generalization of the Dirichlet Distribution*, American Statistical Association **64** (1969), no. 325, 194–206.

[9] G. Ditzler, R. Polikar, and G. Rosen, *Multi-Layer and Recursive Neural Networks for Metagenomic Classification*, IEEE Transactions on Nanobioscience **14** (2015), no. 6, 608–616.

[10] S. Gill, M. Pop, R. DeBoy, and P. Eckburg, *Metagenomic analysis of the human distal gut microbiome*, Science **312** (2006), no. 5778, 1355–1359.

[11] G. B. Gloor, J. M. Macklaim, V. Pawlowsky-Glahn, and J. J. Egozcue, *Microbiome datasets are compositional: And this is not optional*, Frontiers in Microbiology **8** (2017), no. NOV, 1–6.

[12] C. Gu,Smoothing Spline ANOVA Models, Springer Science , and Business Media **297** (2013).

[13] C. Gu, *Smoothing spline ANOVA models: R package gss*, Journal of Statistical Software **58** (2014), no. 5.

[14] M. C. Hu, M. Pavlicova, and E. V. Nunes, *Zero-inflated and hurdle models of count data with extra zeros: Examples from an HIV-risk reduction intervention trial*, American Journal of Drug and Alcohol Abuse **37** (2011), no. 5, 367–375.

[15] T. Hu, P. Gallins, and Y.-h. Zhou, *A zero-inflated beta-binomial model for microbiome data analysis Stat*, Stat **7** (2018), no. 1.

[16] C. P. Kelly, *Fecal microbiota transplantation — An old therapy comes of age*, New England Journal of Medicine **368** (2013), no. 5, 474–475.

[17] D. Knights, L. W. Parfrey, J. Zaneveld, C. Lozupone, and R. Knight, *Human-associated microbial signatures: Examining their predictive value*, Cell Host and Microbe **10** (2011), no. 4, 292–296.

[18] K. T. Konstantinidis and J. M. Tiedje, *Prokaryotic taxonomy and phylogeny in the genomic era: advancements and challenges ahead*, Current Opinion in Microbiology **10** (2007), no. 5, 504–509.

[19] J. Kuczynski, C. L. Lauber, W. A. Walters, L. W. Parfrey, J. C. Clemente, D. Gevers, and R. Knight, *Experimental and analytical tools for studying the human microbiome*, Nature Reviews Genetics **13** (2012), no. 1, 47–58.

[20] J. Kuczynski, J. Stombaugh, W. A. Walters, A. González, J. G. Caporaso, and R. Knight, *Using QIIME to analyze 16S rRNA gene sequences from microbial communities*, Current protocols in microbiology **27** (2012), no. 1, 1E–5.

[21] P. S. la Rosa, J. P. Brooks, E. Deych, E. L. Boone, D. J. Edwards, Q. Wang, E. Sodergren, G. Weinstock, and W. D. Shannon, *Hypothesis testing and power calculations for taxonomic-based human microbiome data*, PLoS ONE **7** (2012), no. 12, 1–13.

[22] M. G. Langille, J. Zaneveld, J. G. Caporaso, D. McDonald, D. Knights, J. A. Reyes, J. C. Clemente, D. E. Burkepile, R. L. V. Thurber, R. Knight, et al., *Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences*, Nature biotechnology **31** (2013), no. 9, 814.

[23] M. I. Love, W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2*, Genome Biology **15** (2014), no. 12, 1–21.

[24] C. Luo, L. M. Rodriguez-R, and K. T. Konstantinidis, *A user's guide to quantitative and comparative analysis of metagenomic datasets*, Vol. 531, Elsevier Inc., 1 edition (2013). ISBN 9780124078635.

[25] C. Luo, L. M. Rodriguez-R, and K. T. Konstantinidis, *MyTaxa: An advanced taxonomic classifier for genomic and metagenomic sequences*, Nucleic Acids Research **42** (2014), no. 8, 1–12.

[26] D. Luo, S. Ziebell, and L. An, *An informative approach on differential abundance analysis for time-course metagenomic sequencing data*, Bioinformatics **33** (2017), no. 9, 1286–1292.

[27] M. D. Lynch and J. D. Neufeld, *Ecology and exploration of the rare biosphere*, Nature Reviews Microbiology **13** (2015), no. 4, 217–229.

[28] S. Mandal, W. Van Treuren, R. A. White, M. Eggesbø, R. Knight, and S. D. Peddada, *Analysis of composition of microbiomes: a novel method for studying microbial composition*, Microbial Ecology in Health & Disease **26** (2015), no. 0, 1–7.

[29] A. A. Metwally, Y. Dai, P. W. Finn, and D. L. Perkins, *WEVOTE: Weighted voting taxonomic identification method of microbial sequences*, PLoS ONE **11** (2016), no. 9, 1–18.

[30] A. A. Metwally, P. W. Finn, Y. Dai, and D. L. Perkins., *Detection of differential abundance intervals in longitudinal metagenomic data using negative binomial smoothing spline ANOVA*, Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics (2017), 295–304.

[31] A. A. Metwally, J. Yang, C. Ascoli, Y. Dai, P. W. Finn, and D. L. Perkins, *MetaLonDA: A flexible R package for identifying time intervals of differentially abundant features in metagenomic longitudinal studies*, Microbiome **6** (2018), no. 1, 1–12.

[32] D. Nichols, N. Cahoon, E. M. Trakhtenberg, L. Pham, A. Mehta, A. Belanger, T. Kanigan, K. Lewis, and S. S. Epstein, *Use of ichip for high-throughput in situ cultivation of "uncultivable" microbial species*, Applied and Environmental Microbiology **76** (2010), no. 8, 2445–2450.

[33] J. N. Paulson, O. Colin Stine, H. C. Bravo, and M. Pop, *Differential abundance analysis for microbial marker-gene surveys*, Nature Methods **10** (2013), no. 12, 1200–1202.

[34] J. N. Paulson, H. Talukder, and H. C. Bravo, *Longitudinal differential abundance analysis of microbial marker-gene surveys using smoothing splines*, bioRxiv (2017), 099457.

[35] K. J. Pflughoeft and J. Versalovic, *Human Microbiome in Health and Disease*, Annual Review of Pathology: Mechanisms of Disease **7** (2012), no. 1, 99–122.

[36] R. Poretsky, L. M. Rodriguez-R, C. Luo, D. Tsementzi, and K. T. Konstantinidis, *Strengths and limitations of 16S rRNA gene amplicon sequencing in revealing temporal microbial community dynamics*, PLoS ONE **9** (2014), no. 4.

[37] A. Rani, R. Ranjan, H. S. McGee, A. A. Metwally, Z. Hajjiri, D. C. Brennan, P. W. Finn, and D. L. Perkins, *A diverse virome in kidney transplant patients contains multiple viral subtypes with distinct polymorphisms*, Scientific Reports **6** (2016), no. September, 1–13.

[38] R. Ranjan, A. Rani, A. A. Metwally, H. S. McGee, and D. L. Perkins, *Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing*, Biochemical and Biophysical Research Communications **469** (2016), no. 4, 967–977.

[39] M. S. Rappé and S. J. Giovannoni, *The Uncultured Microbial Majority*, Annual Review of Microbiology **57** (2003), no. 1, 369–394.

[40] D. Reiman, A. A. Metwally, and Y. Dai, *Using convolutional neural networks to explore the microbiome*, 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (2017), no. November, 4269–4272.

[41] D. Reiman, A. A. Metwally, and Y. Dai, *PopPhy-CNN: A Phylogenetic Tree Embedded Architecture for Convolution Neural Networks for Metagenomic Data*, bioRxiv (2018), 257931.

[42] B. Ren, S. Bacallado, S. Favaro, S. Holmes, and L. Trippa, *Bayesian nonparametric ordination for the analysis of microbial communities*, American Statistical Association **112** (2017), no. 520, 1430–1442.

[43] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, *edgeR: A Bioconductor package for differential expression analysis of digital gene expression data*, Bioinformatics **26** (2009), no. 1, 139–140.

[44] S. Sanschagrin and E. Yergeau, *Next-generation sequencing of 16S ribosomal RNA gene amplicons*, Journal of Visualized Experiments (2014), no. 90, 3–8.

[45] C. Schott, S. S. Weigt, B. A. Turturice, A. A. Metwally, J. Belperio, P. W. Finn, and D. L. Perkins, *Bronchiolitis obliterans syndrome susceptibility and the pulmonary microbiome*, The Journal of Heart and Lung Transplantation (2018), 1–10.

[46] P. Shi and H. Li, *A model for paired-multinomial data and its application to analysis of data on a taxonomic tree*, Biometrics **73** (2017), no. 4, 1266–1278.

[47] D. Sims, I. Sudbery, N. E. Ilott, A. Heger, and C. P. Ponting, *Sequencing depth and coverage: Key considerations in genomic analyses*, Nature Reviews Genetics **15** (2014), no. 2, 121–132.

[48] E. J. Stewart, *Growing unculturable bacteria*, Journal of Bacteriology **194** (2012), no. 16, 4151–4160.

[49] Z.-Z. Tang and G. Chen, *Zero-inflated generalized Dirichlet multinomial regression model for microbiome compositional data*, Biostatistics (2018), no. June, 1–16.

[50] P. Taylor, D. Lambert, T. B. Laboratories, and M. Hill, *Zero-Inflated Poisson Regression , With an Application to Defects in Manufacturing Zero-Inflated Poisson Regression , With an Application to Defects in Manufacturing* **34** (2012), no. February 2015, 37–41.

[51] D. T. Truong, E. A. Franzosa, T. L. Tickle, M. Scholz, G. Weingart, E. Pasolli, A. Tett, C. Huttenhower, and N. Segata, *MetaPhlAn2 for enhanced metagenomic taxonomic profiling*, Nature Methods **12** (2015), no. 10, 902–903.

[52] P. J. Turnbaugh, R. E. Ley, M. Hamady, C. Fraser-liggett, R. Knight, and J. I. Gordon, *The human microbiome project: exploring the microbial part of ourselves in a changing world*, Nature **449** (2007), no. 7164, 804–810.

[53] T. Vatanen, A. D. Kostic, E. D'Hennezel, H. Siljander, E. A. Franzosa, M. Yassour, R. Kolde, H. Vlamakis, T. D. Arthur, A. M. Hämäläinen, A. Peet, V. Tillmann, R. Uibo, S. Mokurov, N. Dorshakova, J. Ilonen, S. M. Virtanen, S. J. Szabo, J. A. Porter, H. Lähdesmäki, C. Huttenhower, D. Gevers, T. W. Cullen, M. Knip, and R. J. Xavier, *Variation in Microbiome LPS Immunogenicity Contributes to Autoimmunity in Humans*, Cell **165** (2016), no. 4, 842–853.

[54] A. Vrieze, E. Van Nood, F. Holleman, J. Salojärvi, R. S. Kootte, J. F. Bartelsman, G. M. Dallinga-Thie, M. T. Ackermans, M. J. Serlie,

R. Oozeer, M. Derrien, A. Druesne, J. E. Van Hylckama Vlieg, V. W. Bloks, A. K. Groen, H. G. Heilig, E. G. Zoetendal, E. S. Stroes, W. M. De Vos, J. B. Hoekstra, and M. Nieuwdorp, *Transfer of intestinal microbiota from lean donors increases insulin sensitivity in individuals with metabolic syndrome*, Gastroenterology **143** (2012), no. 4, 913–916.

[55] G. Wahba, Y. Wang, C. Gu, R. Klein, and B. Klein, *Smoothing spline ANOVA for exponential families, with application to the Wisconsin Epidemiological Study of Diabetic Retinopathy*, The Annals of Statistics **23** (1995), no. 6, 1865–1895.

[56] D. E. Wood and S. L. Salzberg, *Kraken: Ultrafast metagenomic sequence classification using exact alignments*, Genome Biology **15** (2014), no. 3.

[57] L. Xu, A. D. Paterson, W. Turpin, and W. Xu, *Assessment and selection of competing models for zero-inflated microbiome data*, PLoS ONE **10** (2015), no. 7, 1–30.

[58] S. Yang, L. L. Harlow, G. Puggioni, and C. A. Redding, *A comparison of different methods of zero-inflated data analysis and an application in health surveys*, Journal of Modern Applied Statistical Methods **16** (2017), no. 1, 518–543.

[59] T. Yatsunenko, F. E. Rey, M. J. Manary, I. Trehan, M. G. Dominguez-Bello, M. Contreras, M. Magris, G. Hidalgo, R. N. Baldassano, A. P. Anokhin, A. C. Heath, B. Warner, J. Reeder, J. Kuczynski, J. G. Caporaso, C. A. Lozupone, C. Lauber, J. C. Clemente, D. Knights, R. Knight, and J. I. Gordon, *Human gut microbiome viewed across age and geography*, Nature **486** (2012), no. 7402, 222–227.

Departments of Bioengineering and Computer Science
University of Illinois at Chicago
Chicago, IL 60607, USA
E-mail address: ametwa2@uic.edu

Department of Mathematics, Statistics, and Computer Science
University of Illinois at Chicago
Chicago, IL 60607, USA
E-mail address: haldir2@uic.edu

Department of Mathematics, Statistics, and Computer Science
University of Illinois at Chicago
Chicago, IL 60607, USA
E-mail address: jyang06@uic.edu