Big Data and Information Analytics

http://www.aimspress.com/journal/BDIA

BDIA, 4: 1–12.
DOI:10.3934/bdia.2019001
Received: 30 November 2019
Accepted: 23 December 2019
Published: 31 December 2019

*Research article*

# Statistical modeling on human microbiome sequencing data

**Dongyang Yang**[1] **and Wei Xu**[1,2,*]

[1] Department of Biostatistics, Dalla Lana School of Public Health, University of Toronto, Toronto, Canada

[2] Department of Biostatistics, Princess Margeret Cancer Centre, Toronto, Canada

* **Correspondence:** Email: Wei.Xu@uhnres.utoronto.ca.

**Abstract:** Research studies have shown that human microbiome is associated with many diseases through the linkage between bacterial taxa and environmental and genetic factors. Typical human microbiome sequencing data that obtained by next generation sequencing technologies of the 16S rRNA gene are high dimensional and sparse because most taxa are not shared among the samples. As a result, the data is often over-dispersed and with excess zeros. These features rise statistical challenges for compositional data analysis. We review the recent statistical methodology development for this setting. In particular, we summarize some current popular parametric probability models including the cases when repeated measurements of the microbiome are applicable. Multivariate analyses methods that are based on distance measurement for testing differences between microbes community are introduced. Statistical models which are developed to assess the association between genetic variants on X-chromosome and microbial components are highlighted. We discuss some applications on analysis of the association of host genome, microbial compositions and human diseases. Despite sophisticated approaches to statistical analysis of taxa count data, we suggest some future research directions on how to classify and predict clinical outcomes with microbial compositions.

**Keywords:** big data; statistical modeling; human microbiome; sequencing data; genomics; data mining

## 1. Introduction

The field of genomics has been developed to conduct metagenome of the microbiota over the last two decades [1–4]. Current microbiome studies are mostly motivated by the research topics in which aim to understand the relationship among microbiome, host, and genetic or environmental factors. A variety of studies has identified the association between microbiome and host [5, 6], and these works

have shown how human microbiota affects health and diseases. For example, microbial changes are proved to be linked with Parkinson's disease [7], inflammatory bowel disease [8, 9], diabetes [10], and cancers [11]. The other field of microbiome studies is to examine the association between microbiome and genetic/environmental variables, in particular, the effects of the interested genome [12] and environment covariates [13, 14] on specific microbiome composition. Research has been conducted on skin conditions [15, 16], obesity [17], and immunity system [18].

The microbiome data is quantified by amplified sequences using generic sequence similarity, produced via next-generation sequencing of the 16S ribosomal ribonucleic acid (rRNA) gene [19] and bioinformatic pipelines such as QIIME [20]. Classification on the sequencing reads based on phylogenetic levels (genus, family, suborder, order, subclass, class, phyla, kingdom, and domain) is referred to as operational taxonomic unit (OTU) counts (Figure 1). These OTU counts provide the

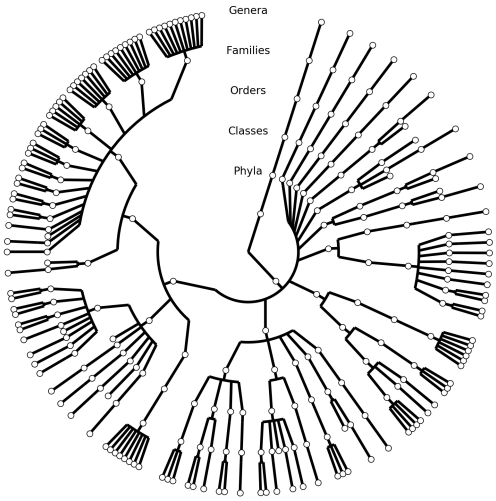


**Figure 1.** An illustration of the taxonomic structure of OTUs.

foundation and development of research on microbiome while generating challenges of statistical methods due to the features. The OTU counts are usually zero-inflated and over-dispersion since the particular host's microbiome taxonomy reads are context specific (Figure 2) [21]. In addition, both the hierarchical structure of microbiome and the sample collection of multiple measurements from related individuals both yields the correlation between different taxa. Another feature in many microbiome studies is that the number of subjects can be smaller than the number of taxa to be explored, which also imposes difficulties in statistical modeling [22–24].

This paper seeks to summarize the recent development regarding statistical methods to analyze human microbiome sequencing data, including assessment and comparison between current popular models and advanced technique for longitudinal microbiome involving serial correlations and hierarchical structure. Diversity among microbiome communities is identified using non-parametric methods. Furthermore, the association between microbial components and genome has been explored, especially including genetic variants on X-chromosome. Finally, potential directions are discussed to consider for further research within this field, such as prediction using microbiome data and classification of OTUs based on the research outcome of interest.

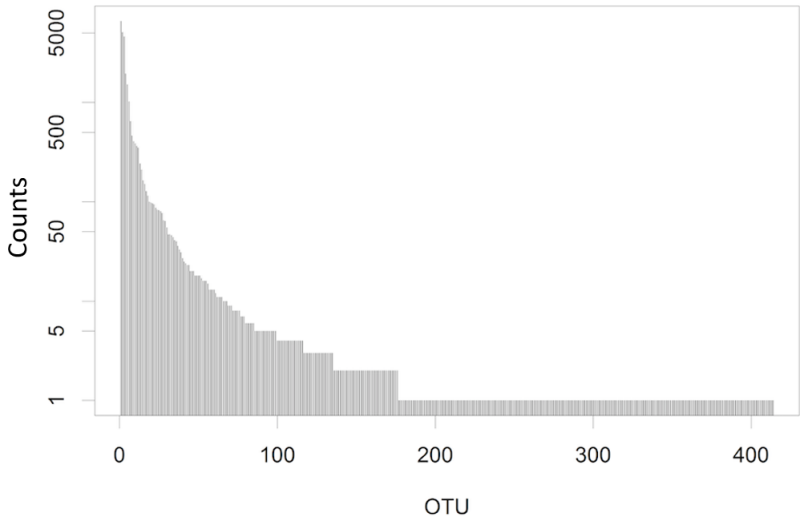


**Figure 2.** A histogram of the counts of a single OTU to exemplify the distribution of microbial data.

## 2. Methods

The relationship between the abundance of single or multiple OTUs and environmental or genetic factors has been investigated extensively yet without standard statistical methods [12–14]. Classical models such as linear regression and logistic regression models are the most popular approaches, while facing the risk of the violation of normality and constant variance assumptions for linear models, and the loss of information on zero parts hence lower statistical power on (generalized) logistic regression models [25]. Non-parametric models such as Wilcoxon rank sum (WRS) test can be used as an alternative approach without the normality assumption, but it cannot adjust for covariate effects [26]. Standard t-tests are used for comparison between two relative abundance datasets [27–29]. When the number of communities is more than two, the one-way analysis of variance (ANOVA) or the Kruskal-Wallis test become appropriate. However, none of these methods above can account for the excess zeros, which is a key feature of microbiome components data, yet currently are still widely used for microbiome studies.

### *2.1. Parametric models for Zero-inflated data*

In order to deal with excess zeros, mixture models are proposed such as zero inflated (ZI) models and hurdle models (also called two-part models). Typical microbiome data contains OTU counts which can be referred to as the variable $Y_{ij}$ for OTU $j$ in sample $i$. Since the mixture models deal with single OTU, we can ignore the OTU index $j$ in this section. ZI models are usually a mixture of a Poisson (ZIP) or Negative Binomial (ZINB) model with a point mass at zero. These models process data with excessive zeros in two steps. First, structure zeros are distinguished from counts data using a Bernoulli trial with probability $\phi_i$ for each $y_i$. Then the counts data is assumed to follow a Poisson or NB distribution. Specifically, the probability mass function (PMF) of a zero-inflated model for $y_i$ can

be written as:

$$f_{ZI}(y_i) = \begin{cases} \phi_i + (1-\phi_i)f(0), & \text{for } y_i = 0 \\ (1-\phi_i)f(y_i), & \text{for } y_i > 0 \end{cases} \tag{1}$$

Hurdle models contain two parts: One part is a binomial model to indicate whether a zero or non-zero outcome occurs, and the other part is a truncated-at-zero model for count parts only. In particular, the PMF of the hurdle model for Possion or NB model is written as:

$$f_H(y_i) = \begin{cases} \phi_i, & \text{for } y_i = 0 \\ (1-\phi_i)\frac{f(y_i)}{1-f(0)}, & \text{for } y_i > 0 \end{cases} \tag{2}$$

and the $\phi_i$ can be linked to covariates through logit link

$$logit(\phi_i) = log(\frac{\phi_i}{1-\phi_i}) = \beta_0 + \boldsymbol{W_i^T\beta} \tag{3}$$

where $\boldsymbol{W_i}$ denotes the vector of covariates for $\phi_i$. The two sets of models treat zero part distinctly. ZI models allow structural zeros in the point mass and leave sampling zeros in the parametric model part, while hurdle models handle structural zeros and sampling zeros together in the binomial model part.

Xu. et al. [30] performed a comprehensive comparison of different model fitting performance for zero inflated data in human microbiome studies, especially in the aspects of type I error and statistical power of the tests. The performance of parameter estimations such as accuracy and efficiency on count parts and zero parts as well as the goodness of fit of the models were also evaluated. Simulations revealed that both hurdle models and ZI models provide a better model fit than standard one part models, resulting in less biased and more efficient parameter estimations while controlling for type I errors and maximizing power.

### *2.2. Bayesian latent variable models for hierarchical clustered data with repeated measures*

The model introduced in the last section follows independent data assumption, which does not always hold for general microbiome data due to the hierarchical clusters and repeated measurements. Xu et al. [31] proposed a Bayesian latent variable (BLV) model, which is joint modeling of multiple taxa within a single taxonomic cluster. This model can not only make inference on the genetic and environmental risk factors effects within the cluster, but also account for repeated measurements of the microbiome from related family members.

The BLV model can incorporate multiple response variables to account for underlying correlations among the multiple taxa within the taxonomic cluster. Let $\boldsymbol{Y}_{cit} = (y_{cit1}, ..., y_{citJ})^T$ be the $J \times 1$ vector of outcomes measured at the $t$th time point on the $i$th individual from the $c$th family, for $t = 1, ..., T_{ci}$, $i = 1, ..., N_c$, $c = 1, ..., C$. $C$ denotes the total number of families; $N_c$ denotes the number of family members in the $c$th family; $T_{ci}$ denotes the total number of repeated measurements for the $i$th individual from the $c$th family. Let $\boldsymbol{U}_{ci} = (U_{ci1}, ..., U_{ciT_{ci}})^T$ and $\boldsymbol{U}_{z,ci} = (U_{z,ci1}, ..., U_{z,ciT_{ci}})^T$ be the two vectors for the longitudinal latent trait underlying the count components and the structural zero parts.

Note that the model can accommodate a mixture of distributions such as NB and ZINB. This is important in microbiome research since the multivariate OTU outcomes can be mixed counts with and without a zero-inflated feature. Consequently, a portion of $\boldsymbol{Y}_{cit}$ follows $NB(\tau_j, \mu_{citj})$ and the other part

of the outcomes follows ZINB with PMF

$$f_{ZINB}(y_i) = \begin{cases} \phi_i + (1-\phi_i) f_{NB}(0; \tau_j, \mu_{citj}), & \text{for } y_i = 0 \\ (1-\phi_i) f(y_i; \tau_j, \mu_{citj}), & \text{for } y_i > 0 \end{cases} \tag{4}$$

The BLV model consists of two parts. The first part is the measurement model, which models the latent traits and a portion of the covariates effects directly on the responses using a generalized linear mixed model. In particular, the parameters depend on both latent traits $\boldsymbol{U}$ and $\boldsymbol{U_z}$ and some other covariates $\boldsymbol{W}$ and $\boldsymbol{W_z}$ through canonical links $log(\mu_{citj}) = \gamma_{0j} + \boldsymbol{W}_{cit}^T \gamma_j + \lambda_j U_{cit} + b_{cij}$ and $log(\frac{\phi_{citj}}{1-\phi_{citj}}) = \beta_{0j} + \boldsymbol{W}_{z,cit}^T \beta_j + \lambda_{z,j} U_{z,cit} + b_{z,cij}$. $b_{(z),cij}$ is the family-specific, within-subject random effect and both $b_{cij}$ and $b_{z,cij}$ are assumed follow normal distribution with mean zero and variance $\eta_j^2$ and $\eta_{z,j}^2$, respectively. The second part, known as the structural model, shows the relations between the other portion of covariates and the latent traits with a linear mixed model. Specifically, the second part is written as

$$\boldsymbol{U_{cit}} = \boldsymbol{X_{cit}}\boldsymbol{\alpha} + g_c + \boldsymbol{Z_{ci}^T}\boldsymbol{\alpha_c} + \boldsymbol{R_{cit}^T}\boldsymbol{d_{ci}} + \epsilon_{cit}, \tag{5}$$

and

$$\boldsymbol{U_{z,cit}} = \boldsymbol{X_{z,cit}}\boldsymbol{\alpha_z} + g_{z,c} + \boldsymbol{Z_{ci}^T}\boldsymbol{\alpha_{z,c}} + \boldsymbol{R_{z,cit}^T}\boldsymbol{d_{z,ci}} + \epsilon_{z,cit}, \tag{6}$$

where $\boldsymbol{X_{cit}}$ and $\boldsymbol{X_{z,cit}}$ are the covariates of interest on the latent traits. $g_c$ and $g_{z,c}$ are the environmental random effects for $\boldsymbol{U_{cit}}$ and $\boldsymbol{U_{z,cit}}$ within a family. $\boldsymbol{\alpha_c}$ and $\boldsymbol{\alpha_{z,c}}$ are the additive genetic random effects. $\boldsymbol{d_{ci}}$ and $\boldsymbol{d_{z,ci}}$ are the serial random effects. Hence, the risk factors in this model have indirect effects on the response through the latent variable. This Bayesian method is applied for parameter inference on the complicated form of the posterior distribution. Polya-Gamma data augmentation (PGDA) technique is used as a part of the Markov chain Monte Carlo (MCMC) algorithm for the model.

Simulation studies showed that both the direct and indirect effect parameters have small bias and the root mean square errors (RMSEs), suggesting that the proposed model performs well with controlled type I error and reasonable power of tests. Random effects estimations mostly are unbiased, except for the variance of serial random effects on the probability of structural zeros and the LVs.

### *2.3. Distance based method to access differences between communities*

Parametric models introduced in the previous sections can provide parameter estimations and statistical inference on the risk factor association. However, it is difficult to identify and select the optimal distribution of the model, and sometimes the model assumptions may be violated due to complex microbiome data structure. Other than parametric models, distance-based non-parametric models provide another approach to analyze differences between microbiome communities as a whole rather than only incorporate a univariate outcome. In Shestopaloff's paper [32], a mixture model has been proposed to model OTUs generated from Poisson models with subject-specific underlying rates. The observed counts are assumed to follow Poisson distribution $X \sim Poisson(r_i)$ and $r_i = q_i N_i$, in which $q_i$ is individual specific relative abundance sampled from underlying population distribution $G_q$, and $N_i$ is the total reads in the sample.

Due to the complexity of the observed data, a set of distribution components is proposed to model the underlying population distribution, including a zero point mass, a set of left-skewed distribution for low rates, a set of Gamma distributions based on the posterior of the Poisson rate $\lambda|n \sim \Gamma(n+1, 1)$,

and a truncation point mass $P(X > C) = 1$ for a sufficiently large number C to account for sparsity. The final mixture distribution depends on the specified M gamma distribution $(\alpha_m, \beta_m)$ with weights $\boldsymbol{w} = (w_z, w_1, ..., w_M, w_{C+})$ respectively. The least square optimization function is applied to estimate the optimal weights by taking the difference between the observed and expected aggregate counts.

The choices of parameters for low rate structure are decided by selecting the model among a set of nested models using a nonparametric bootstrap based on the minimum distance between the expected aggregate counts and the observed counts. This model selection technique is similar to cross-validation as a way to fit the model with a portion of data to avoid overfitting. The joint mixture distribution estimate $\vec{w} = \sum_l v(l)\vec{w}_l$ can be calculated using the weight $v(l)$ of each model obtained from the bootstrap. Then the probability of observing $n_i$ from the $m^{th}$ Gamma mixture conditioning on the estimated weight and resolution $t_i$ is

$$p_{im} = P(i \in G_m | n_i, t_i, \boldsymbol{w}) = w_m \frac{\Gamma(n_i + \alpha_m)}{\Gamma(n_i + 1)\Gamma(\alpha_m)} \left(\frac{\beta_m}{t_i + \beta_m}\right)^{\alpha_m} \left(1 - \frac{\beta_m}{t_i + \beta_m}\right)^{n_i}. \tag{7}$$

And the probabilities of the two point masses are $P(i \in G_z) = I(n_i = 0)$ and $P(i \in G_{C+} = I(n_i > C))$.

Finally, permutation tests can be performed to evaluate the significance of differences between communities after the calculation of pairwise distances. These distances can be treated as the pairwise $L^2$-PDF norms $D_{L^2}(i_1, i_2) = \|P_{i_1} - P_{i_2}\|$ for $i_1$,$i_2 = 1, ..., I$ and $i_1 \neq i_2$, where the PDF for each sample $i$ is $P_i = [P_i(Z), P_i(0), ..., P_i(C), P_i(C+)]$. The permutation method is introduced by Anderson [33].

The simulation study showed that the mixture model correctly estimated the true underlying rate distribution and the proportion of structure zeros. When the true model followed certain parametric models, the corresponding parametric methods such as 2P-LOLS and NBH performed well. Non-parametric Kruskal-Wallis tests gave a relatively robust estimation, too. However, these methods can incorporate only a single outcome. When the comparison of multiple OTUs is of interest, distance-based methods become the only applicable approach, and the proposed mixture model performed better than the Manhattan distance method for most scenarios.

### *2.4. X-chromosome association with microbial composition*

Besides the research works examined the association between microbiome and diseases, many researchers have interests in the relationship between host factors and intestinal microbial composition. The host factors include age, gender, dietary, and genetic variants. However, among all the studies proved the linkage of the host genome and human microbiome [12, 34–36], none of the research investigated the impact of genetic variants on X-chromosome on the microbiota. Espin-Garcia et al. [37] proposed a finite mixture model (FMM) for the analysis of the X-chromosome and the composition of the microbiota. The method incorporates multiple unknown underlying X-chromosome mechanisms (XCMs), including random X-chromosome inactivation (XCI), skewed XCI (XCI-S), and escape of XCI (XCI-E).

This regression-based model consists of two parts: a distribution function and a mixing proportion of XCMs. Let $y_i$ be the OTU counts or the indicator function for the presence/absence of the OTU for subject $i$ depending on the outcome of interest. Let $x_i^k = (1, s_i, g_i^{s,k}, w_i')$ be the covariates vector for subject $i$ under $k$th mechanism, where $s_i$ be the sex indicator; $g_i^{s,k}$ is the coded SNPs; and $w_i$ is the vector for additional covariates. Then the likelihood for the assumed observed data with a pseudo

mechanism is

$$L(\beta, p) = \prod_{i=1}^{n} \sum_{k=1}^{K} f(y_i|x_i^k;\beta)p_k, \tag{8}$$

where $p_k$ specifies the mixing proportion for the $k$th mechanism. $f(\cdot)$ is one of the two distribution functions - either a zero-inflated probability distribution using Poisson or negative binomial distributions for count parts, or a two-part model treating zero and count parts separately.

Comprehensive simulations were constructed to compare the performance of the proposed X-chromosome model with some existing ones such as a Clayton-like model [38] and a PLINK-like model [39], as well as strategies which assume the same mechanisms for all subjects. Expectation-maximization (EM) algorithm was used to estimate the potential genetic effects, and a score statistic was computed to evaluate hypothesis testing. In conclusion, the FMM provides relatively less biased estimates and competitively higher power while controlling for type I error in the comparison of other methods.

### *2.5. Association of host genome, microbial composition, and intestinal permeability*

As implementation of the statistical models on real human microbiome studies, the heritability of microbial components are explored using a log-normal model with a generalized estimating equation (GEE) algorithm among 270 related individuals, and successfully identified 94 of 249 OTUs with significant additive genetic components with high heritability [12]. This result suggested that host genetics is strongly associated with the intestinal microbial composition. In addition, Genome-wide association studies (GWASs) were conducted to identify an association between genetic variants and bacterial taxa. GEE framework was adopted to identify the association between genetic polymorphisms and the relative abundance of heritable taxa, controlling for age, sex, and the top three genetic principal components. The two-part log-normal model was fitted on zero counts and nonzero counts separately by using logistic regression and log-normal model, respectively. External validation confirmed four specific OTUs associated with the host genetic variations.

Analysis of intestinal permeability (IP) is also conducted in healthy first-degree relatives of individuals with Crohn's disease [40]. A generalized least squares model was applied to evaluate the association between fecal microbiota and IP. Potential clustering within families was accounted by a compound symmetry correlation matrix. No significant correlation between several levels of bacterial taxa (phylum, family and genus) and IP was found, adjusted by clinical factors such as age, age squared, gender, and the province of origin.

## 3. Future research works

In the last few years, the introduced statistical methods have been performed more frequently to model the zero-inflated, over-dispersion, and clustered microbiome data, and investigate the association of microbiota and phenotypes or genotypes. In the following session, some potential directions are introduced for prediction and classification purposes using microbial composition.

### 3.1. Prediction of phenotypes based on human microbiome data

Statistical testing methods which are traditionally used for identifying associated OTUs can be performed for prediction purposes, but they are limited by the correctness of regression models. Particularly for microbiome data with multiple measurements for each subject, generalized linear mixed models (GLMMs) are an effective way to estimate the OTUs' effects on clinical outcomes while accounting for both fixed and random effects. By setting certain thresholds for the p-values controlled by false discovery rate (FDR), GLMMs for OTU counts given a phenotype and other covariates can investigate not only the association between microbial composition and the phenotype, but also the potential predictors in which we are interested. Two-part models and zero-inflated models can be adopted to GLMMs to account for excessive zeros.

However, disagreements often occur between statistical significance and predictivity. These limitations lead to the usage of statistical and machine learning methods such as the least absolute shrinkage and selection operator (LASSO) for prediction purposes. LASSO serves feature selection goal by adding the $L_1$ of the regression coefficients as a penalty term to the log likelihood function to achieve the intent of shrinkage. The objective log likelihood function for LASSO given observations $(y_i, x_i), i = 1, ..., n$ and a tuning parameter $\lambda$ is

$$l_{LASSO}(\beta) = -\sum_{i=1}^{n} log(P(y_i|x_i, \beta)) + \lambda \sum_{k=1}^{K} ||\beta_k||_1, \tag{9}$$

where $||\beta_k||_1$ is the $L_1$ penalty of $\beta_k$. Therefore, fitting the model for a phenotype given OTU composition with the LASSO method could be performed to predict the presence/absence of the phenotype. As both LASSO multinomial logistic regression and GLMMs can select OTUs as potential predictors, a combination of the two models becomes a promising method for prediction of phenotypes by first using GLMMs to screen on all the given OTUs then applying LASSO to the subset of OTUs selected previously.

### 3.2. Classification on microbiome data

Classification methods as a supervised learning technique, provide another angle to predict the outcome of interest by building a classifier based on the training dataset and predicting for the unknown observations. Some existing classification methods have been developed extensively, such as logistic regression models, discriminant analysis, classification decision trees, and k-nearest neighbor methods. Besides those methods, the distance-based mixture models introduced previously can be extended for a new classification method on microbiome sequencing data.

### 3.3. High-dimensional GLM for graphic structure

Generalized linear models (GLMs) are another way of classification under the high-dimensional setting in which the number of predictors exceeds the number of observations. In particular, logistic regression models are popular for classification in microbiome data. Matson et al. [41] showed that relevant microbial compositions are naturally correlated, and thus these components within a cluster reflected the association with an interesting clinical outcome simultaneously. Therefore, consideration of such structure within the neighborhood that can incorporate sparsity pattern would boost the

correctness of prediction. A combination of the graph structure and GLM through a node-wise penalty which is able to account for neighborhood sparsity can be applied for classification.

## 4. Conclusions

In this paper, we review and summarize current statistical methods to analyze zero-inflated and over-dispersion microbiome sequencing data. Other than the traditional linear models, zero-inflated models and hurdle models are recommended which have advantages of handling excess zeros. In addition, repeated measurements for individuals with their relatives are commonly encountered in microbiome studies. A Bayesian latent variable model along with the PGDA technique is proposed to handle the hierarchical clusters and the longitudinal correlation. Meanwhile, the parametric models can only accommodate single OTU as the outcome, but distance-based methods test the differences between communities across the population with multiple microbial compositions. The other difference between the parametric models and non-parametric approach is that covariates can be adjusted in parametric regression models, while distance-based methods are limited to randomized clinical trials due to the lack of ability to handle the potential confounding variables.

Besides treating OTUs as the outcome of some interested clinical variables, some researchers also investigate the association between genome and microbiome. Although not much work had been done to explore the genetic effects of sex-chromosome on the microbiome, one of the papers which we review focuses on the relationship with X-chromosome genetic variants and microbiota, by incorporating the underlying X-chromosome architecture to the finite mixture model. Some applied analysis has also been conducted, specifically on the association of host genome, microbiome, and intestinal permeability.

Prediction for human health using microbial components is a promising direction. In particular, GLMM and LASSO are the models that have good performance when a subset of OTUs is selected to predict the disease status. Furthermore, classification of microbial communities based on disease outcomes also has a potential impact in the medical field. Distance-based classification algorithm and high-dimensional GLM graphic structure models are worth to investigate further to classify the related OTUs given the presence-absence status.

## Acknowledgements

W.X. was funded by Canadian Institutes of Health Research (CIHR Grant 145546) Natural Sciences and Engineering Research Council of Canada (NSERC Grant RGPIN201706672), Crohns and Colitis Canada (CCC Grant CCCGEMIII), and Helmsley Charitable Trust. D.Y. was supported by NSERC Grant RGPIN201706672, CCC Grant CCCGEMIII, and Edwin S.H. Leong Scholarship.

## Conflict of interest

The authors declare that there are no conflicts of interest.

## References


1. Whiteside SA, Razvi H, Dave S, et al. (2015) The microbiome of the urinary tracta role beyond infection. *Nat Rev Urol* 12: 81–90.
2. Cho I and Blaser MJ, (2012) The human microbiome: At the interface of health and disease. *Nat Rev Genet* 13: 260–270.
3. HMP Integrative, (2014) The integrative human microbiome project: Dynamic analysis of microbiome-host omics profiles during periods of human health and disease. *Cell Host Microbe* 16: 276–289.
4. Young VB, (2017) The role of the microbiome in human health and disease: An introduction for clinicians. *BMJ* 356: j831.
5. Singh RK , Chang HW, Yan D, et al. (2017) Influence of diet on the gut microbiome and implications for human health. *J Trans Med e* 15: 73.
6. Hollister EB, Gao C and Versalovic J, (2014) Compositional and functional features of the gastrointestinal microbiome and their effects on human health. *Gastroenterology* 146: 1449–1458.
7. Sampson TR, Debelius JW, Thron T, et al. (2016) Gut microbiota regulate motor deficits and neuroinflammation in a model of parkinsons disease. *Cell* 167: 1469–1480.
8. Greenblum S, Turnbaugh PT and Borenstein E, (2012) Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. *Proc Natl Acad Sci* 109: 594–599.
9. Morgan XC, Tickle TL, Sokol H, et al. (2012) Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biol* 13: R79.
10. Samuel BS and Gordon JI, (2006) A humanized gnotobiotic mouse model of host–archaeal–bacterial mutualism. *Proc Natl Acad Sci* 103: 10011–10016.
11. Holmes E, Li JV, Athanasiou T, et al. (2011) Understanding the role of gut microbiome–host metabolic signal disruption in health and disease. *Trends Microbiol* 19: 349–359.
12. Turpin W, Espin-Garcia O, Xu W, et al. (2016) Association of host genome with intestinal microbial composition in a large healthy cohort. *Nat Genet* 48: 1413–1417.
13. Schloissnig S, Arumugam M, Sunagawa S, et al. (2013) Genomic variation landscape of the human gut microbiome. *Nature* 493: 45–50.
14. Chase J, Fouquier J, Zare M, et al. (2016) Geography and location are the primary drivers of office microbiome composition. *mSystems* 1: e00022-16.
15. Kong HH, Oh J, Deming C, et al. (2012) Temporal shifts in the skin microbiome associated with disease flares and treatment in children with atopic dermatitis. *Genome Res* 22: 850–859.
16. Grice EA, Kong HH, Conlan S, et al. (2009) Topographical and temporal diversity of the human skin microbiome. *Science* 324: 1190–1192.
17. Turnbaugh PJ, Ley RE, Mahowald MA, et al. (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* 444: 1027–1031.

18. Kau AL, Ahern PP, Griffin NW, et al. (2011) Human nutrition, the gut microbiome and the immune system. *Nature* 474: 327–336.

19. Tringe SG and Rubin EM, (2005) Metagenomics: Dna sequencing of environmental samples. *Nat Rev Genet* 6: 805.

20. Caporaso JG, Kuczynski J, Stombaugh J, et al. (2010) Qiime allows analysis of high-throughput community sequencing data. *Nat Methods* 7: 335–336.

21. Blaxter M, Mann J, Chapman T, et al. (2005) Defining operational taxonomic units using dna barcode data. *Philos Trans R Soc, B* 360: 1935–1943.

22. Lin W, Shi P, Feng R, et al. (2014) Variable selection in regression with compositional covariates. *Biometrika* 101: 785–797.

23. Hongzhe Li, (2015) Microbiome, metagenomics, and high-dimensional compositional data analysis. *Annu Rev Stat Its Appl* 2: 73–94.

24. Shankar J, Szpakowski S, Solis NV, et al. (2015) A systematic evaluation of high-dimensional, ensemble-based regression for exploring large model spaces in microbiome analyses. *BMC Bioinf* 16: 31.

25. McMurdie PJ and Holmes S, (2014) Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput Biol* 10: e1003531.

26. La Rosa PS, Brooks JP, Deych E, et al. (2012) Hypothesis testing and power calculations for taxonomic-based human microbiome data. *PloS One* 7: e52078.

27. Chen W, Liu F, Ling Z, et al. (2012) Human intestinal lumen and mucosa-associated microbiota in patients with colorectal cancer. *PloS One* 7: e39743.

28. Iwai S, Fei M, Huang D, et al. (2012) Oral and airway microbiota in hiv-infected pneumonia patients. *J Clin Microbiol* 50: 2995–3002.

29. Kim KA, Jung IH, Park SH, et al. (2013) Comparative analysis of the gut microbiota in people with different levels of ginsenoside rb1 degradation to compound k. *PLoS One* 8: e62409.

30. Xu L, Paterson AD, Turpin W, et al. (2015) Assessment and selection of competing models for zero-inflated microbiome data. *PloS One* 10: e0129606.

31. Xu L, Paterson AD and Xu W, (2017) Bayesian latent variable models for hierarchical clustered count outcomes with repeated measures in microbiome studies. *Genet Epidemiol* 41: 221–232.

32. Shestopaloff K, Escobar MD and Xu W, (2018) Analyzing differences between microbiome communities using mixture distributions. *Stat Med* 37: 4036–4053.

33. Anderson MJ, (2001) A new method for non-parametric multivariate analysis of variance. *Aust Ecol* 26: 32–46.

34. Jostins L, Ripke S, Weersma RK, et al. (2012) Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* 491: 119–124.

35. Goodrich JK, Waters JL, Poole AC, et al. (2014) Human genetics shape the gut microbiome. *Cell* 159: 789–799.

36. Blekhman R, Goodrich JK, Huang K, et al. (2015) Host genetic variation impacts microbiome composition across human body sites. *Genome Biol* 16: 191.

37. Espin-Garcia O, Croitoru K and Xu W, (2019) A finite mixture model for x-chromosome association with an emphasis on microbiome data analysis. *Genet Epidemiol* 43: 427–439.

38. David Clayton, (2008) Testing for association on the x chromosome. *Biostatistics* 9: 593–600.

39. Zheng G, Joo J, Zhang C, et al. (2007) Testing association for markers on the x chromosome. *Genet Epidemiol* 31: 834–843.

40. Kevans D, Turpin W, Madsen K, et al. (2015) Determinants of intestinal permeability in healthy first-degree relatives of individuals with crohn's disease. *Inflammatory Bowel Dis* 21: 879–887.

41. Matson V, Fessler J, Bao R, et al. (2018) The commensal microbiome is associated with anti–pd-1 efficacy in metastatic melanoma patients. *Science* 359: 104–108.


AIMS Press