A Bayesian Nonparametric Topic Model for Microbiome Data Using Subject Attributes

TASUKU OKUI^{1,a)}

Received: August 19, 2019, Accepted: October 19, 2019

Abstract: Microbiome data have been obtained relatively easily in recent years, and currently, various methods for analyzing microbiome data are being proposed. Latent Dirichlet allocation (LDA) models, which are frequently used to extract latent topics from words in documents, have also been proposed to extract information on microbial communities for microbiome data. To extract microbiome topics associated with a subject's attributes, LDA models that utilize supervisory information, including LDA with Dirichlet multinomial regression (DMR topic model) or supervised topic model (SLDA), can be applied. Further, a Bayesian nonparametric model is often used to automatically decide the number of latent classes for a latent variable model. An LDA can also be extended to a Bayesian nonparametric model using the hierarchical Dirichlet process. Although a Bayesian nonparametric DMR topic model has been previously proposed, it uses normalized gamma process for generating topic distribution, and it is unknown whether the number of topics can be automatically decided from data. It is expected that the total number of topics (with relatively large proportions) can be restricted to a smaller value using the stick-breaking process for generating topic distribution. Therefore, we propose a Bayesian nonparametric DMR topic model using a stick-breaking process and have compared it to existing models using two sets of real microbiome data. The results showed that the proposed model could extract topics that were more associated with attributes of a subject than existing methods, and it could automatically decide the number of topics from the data.

Keywords: microbiome data, 16S rRNA gene sequencing, Latent Dirichlet allocation model, Bayesian nonparametric model

1. Introduction

Microbiome compositional data generated by a next-generation sequencer have been obtained relatively easily in recent years. Among microbiome datasets, data using the 16S rRNA gene are often used to associate each microbiome species with a clinical outcome [1]. In these studies, DNA from microbiome samples is extracted from each subject. Subsequently, using a next-generation sequencer, a gene with taxonomic identity 16S rRNA is amplified and sequenced. The sequencing reads are clustered into operational taxonomic units (OTUs) based on sequence similarity. Further, each OTU can be classified into a taxonomic lineage using a reference database.

Statistically, OTU abundance data are multivariate count data, indicating the number of OTUs detected for each bacterial species in each subject. Various methods for analyzing microbiome data have been proposed, among which, the latent Dirichlet allocation (LDA) model [2], which is frequently used to extract latent topics from words in documents, has been proposed. Microbiomes are known to co-occur, and multivariate analysis methods are often applied to discover microbial clusters or communities. As a multivariate latent variable model, LDA might be used to discover microbial communities. Using an LDA in microbiome data analysis, a document in text data analysis corresponds to a subject, a

term corresponds to a bacterial species, a topic corresponds to a community, and each word corresponds to a sequencing read [3]. The details of correspondence between microbiome data and text data have been previously explained [3]. Further, the microbial environment of a subject differs depending on attributes such as age, time when data were measured, or sex [4], [5], [6]. Using LDAs that use extraneous variables (e.g., supervised topic model (SLDA) [7] or a topic model conditioned on arbitrary features using Dirichlet multinomial regression (DMR topic model) [8]), the attributes of subjects can be accounted for when extracting topics. Compared with SLDA, the DMR topic model has the advantage that it does not require specifying distributional assumptions for the extraneous variables.

When using an LDA for natural language processing or microbiome data, the number of topics of the LDA model must be pre-specified. If the pre-specified number of topics is changed, then the number of topics to be interpreted also changes. As a solution to the problem, Bayesian nonparametric models have been proposed, a series of methods that can mitigate the parametric assumption of a Bayesian model [9]. As one of the Bayesian nonparametric methods, the hierarchical Dirichlet process (HDP) has been proposed [10], which converts an LDA to a Bayesian nonparametric model. Further, a Bayesian nonparametric DMR topic model (Hierarchical Dirichlet scaling process: HDSP) has been previously proposed [11]. However, the method has not been applied to microbiome data, and it is unknown whether the associa-

¹ Medical Information Center, Kyushu University Hospital, Fukuoka 812–8582, Japan

^{a)} task10300@gmail.com

Fig. 1 Graphical model of ordinary LDA.

tion between extracted topics and attributes improves. Moreover, it has not been verified whether HDSP is robust against change in the pre-specified number of topics.

In this study, we propose a Bayesian nonparametric DMR topic model using the stick-breaking process for generating topic distribution. Using the stick-breaking process, it is expected that the number of topics can be automatically decided from the data because a large amount of topic proportion can be concentrated on a few topics based on an estimated hyperparameter value from the data. As a result, the number of topics whose proportions are relatively large would be restricted to a smaller value than that for other models. Further, topics that are more associated with a subject's attributes might be extracted because the proportions of topics are directly affected by the values of a subject's attributes in the proposed model. Therefore, we propose a Bayesian nonparametric DMR topic model using the stick-breaking process for microbiome data and have compared it to existing methods using real microbiome data.

2. Method

In this section, we explain the ordinary LDA, HDSP, and the proposed Bayesian nonparametric DMR topic model.

2.1 Latent Dirichlet Allocation Model (LDA)

In this subsection, we explain the ordinary LDA [2], which is the method mainly used for microbiome data [3], [12], [13]. **Figure 1** shows a graphical illustration of the model. z_{dn} represent the n^{th} ($n=1, \dots, N_d$) sequencing read of subject d ($d=1, \dots, D$), w_{dn} represent a bacterial species for the n^{th} ($n=1, \dots, N_d$) sequencing read of a subject d , π_d represent topic distribution of each subject d , ϕ_k represent the species distribution of a topic k ($k=1, \dots, K$), and η represents the hyper parameter of ϕ_k . In LDA, word distribution in text data analysis represents species distribution in microbiome data analysis.

Let D as the number of subjects, K as the number of topics, and N_d as the sum of sequencing reads for a subject d , then the frames in Fig. 1 indicate that the numbers of parameters or variables in the same frame are identical. The number of each parameter or variable in a frame is the value of the foot of the frame. That is, z_{dn} and w_{dn} exist for every sequence read of each subject. π_d exist for each subject, and ϕ_k exist for each topic. Additionally, z_{dn} fall in the range $1, \dots, K$, and w_{dn} fall in the range $1, \dots, W$. Furthermore, π_d are vectors with lengths of $K \times 1$, and ϕ_k and η are vectors with lengths of $W \times 1$.

From the result of π_d , we can grasp the proportions of topics each subject d has, and from species distribution ϕ_k , the contents of each topic k can be grasped.

The following is the data generation process for the LDA

model:

- (1) For each topic k ,
 - (a) Draw $\phi_k \sim \text{Dirichlet}(\eta)$
- (2) For each subject d ,
 - (a) Draw $\pi_d \sim \text{Dirichlet}(\alpha)$
 - (b) For each n^{th} sequencing read of subject d ,
 - (i) Draw $z_{dn} \sim \text{Multinomial}(1, \pi_d)$
 - (ii) Draw $w_{dn} \sim \text{Multinomial}(1, \phi_{z_{dn}})$

2.2 Bayesian Nonparametric Topic Model

In this subsection, we explain HDP and the proposed method that uses HDP, taking into account a subject's attributes.

2.2.1 Hierarchical Dirichlet Process (HDP)

The building block of HDP is that the random base measure G_0 for a Dirichlet process $G_j \sim \text{DP}(\xi, G_0)$ is itself a draw from a Dirichlet process $G_0 \sim \text{DP}(\gamma, H)$ [10].

$$G_0 \mid \gamma, H \sim \text{DP}(\gamma, H),$$

$$G_j \mid \xi, G_0 \sim \text{DP}(\xi, G_0)$$

for $j \in J$

G_j are group (j) specific random measures, and H is the base measure of G_0 . J is the index set. γ is a hyperparameter of the first level of the Dirichlet process, and ξ is a hyperparameter for beta distribution of the second-level Dirichlet process.

Here is the stick-breaking representation for the random base measure G_0 . The random base measure G_0 can be represented in the following equation.

$$G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\theta_k^{**}},$$

$$\delta_{\theta_k^{**}} \mid H \sim H.$$

θ_k^{**} represent independent random variables distributed according to G_0 , where $\delta_{\theta_k^{**}}$ is an atom at θ_k^{**} . The random variables β_k are generated based on the stick-breaking process.

$$v_k \mid \gamma \sim \text{Beta}(1, \gamma),$$

$$\beta_k = v_k \prod_{l=1}^{k-1} (1 - v_l)$$

for $k = 1, \dots, \infty$

The stick-breaking process postulates an infinite number of latent classes to a model, and by using this method, only the proportions of a few latent classes tend to become large.

A set of group (j) specific random measures G_j is represented by the following equation.

$$G_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\theta_k^{**}}$$

Fig. 2 Graphical model of Bayesian nonparametric DMR topic model.

A set of group (j) specific random measures G_j can be constructed in multiple ways. The following equation applies if using the normalized gamma process [11].

$$\pi'_{jk} \sim \text{Gamma}(\xi\beta_k, 1),$$

$$\pi_{jk} = \frac{\sum_{m=1}^{\infty} \pi'_{jm}}{\pi'_{jk}}$$

for $k = 1, \dots, \infty$

Parameters of each group π_{jk} are also generated based on beta distributions for the stick-breaking process [10].

$$\pi'_{jk} \sim \text{Beta}\left(\xi\beta_k, \xi\left(1 - \sum_{l=1}^k \beta_k\right)\right),$$

$$\pi_{jk} = \pi'_{jk} \prod_{l=1}^{k-1} (1 - \pi'_{jl})$$

for $k = 1, \dots, \infty$

π_j and π'_j are group specific vectors.

2.2.2 Hierarchical Dirichlet Scaling Process (HDSP)

In this section, we explain HDSP [11]. **Figure 2** is the graphical model of HDSP. $\mathbf{x}_d$ represent covariate vectors of a subject d , and λ_k represent regression coefficients. β_k represent the hyper parameter of π_d . γ is the hyperparameter of β_k , and ξ is the hyperparameter of π_k .

$\mathbf{x}_d$ exist for each subject, and β_k and λ_k exist for each topic. If the number of covariates is P , $\mathbf{x}_d$ and λ_k are vectors of $P \times 1$. Although we postulate an infinite number for the number of topics K , we set a large finite number in this process because we cannot set an infinite number actually.

HDSP uses the normalized gamma process for the second level of the Dirichlet process. By using a method of dependent Dirichlet process, HDSP generates topic distribution based on the normalized gamma process using information derived from the covariates. The following is the data generation process for the model shown in Fig. 2:

- (1) For each topic k ,
 - (a) Draw $\lambda_k \sim \text{Normal}(0, \sigma^2 \mathbf{I})$
 - (b) Draw $v_k \sim \text{Beta}(1, \gamma)$
 - (c) Let $\beta_k = v_k \prod_{i=1}^{k-1} (1 - v_i)$
 - (d) Draw $\phi_k \sim \text{Dirichlet}(\eta)$
- (2) For each subject d ,

- (a) Let $\alpha_{dk} = \text{expit}(\lambda'_k \mathbf{x}_d)$.
- (b) Draw $\pi'_{dk} \sim \text{Gamma}(\xi\beta_k, \alpha_{dk})$
- (c) Let $\pi_{dk} = \frac{\pi'_{dk}}{\sum_{k=1}^K (\pi'_{dk})}$
- (d) For each n^{th} sequencing read of d^{th} subject,
 - (i) Draw $z_{dn} \sim \text{Multinomial}(1, \pi_d)$
 - (ii) Draw $w_{dn} \sim \text{Multinomial}(1, \phi_{z_{dn}})$

2.2.3 The Proposed Method

To establish a Bayesian nonparametric DMR topic model that can automatically decide the number of topics from data, we proposed a model using the stick-breaking process. The proposed method uses the stick-breaking process for generating topic distribution instead of the normalized Gaussian process. A graphical illustration of the proposed model can be represented in the same way as HDSP in Fig. 2.

The following is the data generation process for the proposed model:

- (1) For each topic k ,
 - (a) Draw $\lambda_k \sim \text{Normal}(0, \sigma^2 \mathbf{I})$
 - (b) Draw $v_k \sim \text{Beta}(1, \gamma)$
 - (c) Let $\beta_k = v_k \prod_{i=1}^{k-1} (1 - v_i)$
 - (d) Draw $\phi_k \sim \text{Dirichlet}(\eta)$
- (2) For each subject d ,
 - (a) Let $\alpha_{dk} = \text{expit}(\lambda'_k \mathbf{x}_d)$.
 - (b) Draw $\pi'_{dk} \sim \text{Beta}(\xi\beta_k, \xi(1 - \sum_{l=1}^k \beta_k))$
 - (c) Let $\pi_{dk} = \pi'_{dk} \alpha_{dk} \prod_{l=1}^{k-1} (1 - \pi'_{dl} \alpha_{dl})$
 - (d) For each n^{th} sequencing read of d^{th} subject,
 - (i) Draw $z_{dn} \sim \text{Multinomial}(1, \pi_d)$
 - (ii) Draw $w_{dn} \sim \text{Multinomial}(1, \phi_{z_{dn}})$

expit represents sigmoid function. In the proposed method, we combined π'_{dk} with the sigmoid function of the covariates in the stick-breaking process. Because π'_{dk} are in the range of 0 to 1, sigmoid function was used for scaling the function of the covariates in the range of 0 to 1. This method of incorporating the attributes into the stick-breaking process is called the kernel stick-breaking process [14], [15]. By this method, proportions of topics are directly related to the values of a subject's attributes. Further, by using the stick-breaking process for generating topic distribution, a large amount of topic proportion can be concentrated on a few topics based on the value of the hyperparameter ξ . If ξ is large, the topic proportion of each topic would be relatively even. Conversely, if ξ is small, the topic proportion of each subject would be concentrated on a few topics. ξ is estimated from the data. As

Table 1 Attributes of the data.

	Smoking data	Pregnant data
Number of subjects	60	54
Types of species	99	53
Number of sequencing reads for a subject†	1426.1(534.7)	3491.5(1124.0)
Smokers*	28	-
Men*	39	-
Pregnant*	-	22
Bacterial vaginosis*	-	12

* Number of cases

† Mean (Standard deviation)

Table 2 The result of AUC.

Data	Attributes	Method	Pre-specified number of topics				
			10	20	30	40	50
Pregnant data	Pregnancy	LDA	0.930	0.802	0.648	0.651	0.618
		HDSP	0.912	0.908	0.932	0.783	0.949
		Proposed model	0.922	0.954	0.951	0.958	0.982
	Bacterial vaginosis status	LDA	0.831	0.887	0.670	0.591	0.648
		HDSP	0.824	0.876	0.911	0.771	0.864
		Proposed model	0.898	0.939	0.977	0.994	0.991
Smoking data	Smoking status	LDA	0.840	0.934	0.903	0.859	0.618
		HDSP	0.807	0.900	0.916	0.950	0.970
		Proposed model	0.803	0.849	0.937	0.957	0.929
		LDA	0.795	0.854	0.861	0.839	0.648
	Sex	HDSP	0.785	0.809	0.868	0.862	0.909
		Proposed model	0.847	0.963	0.978	0.985	0.988

a result, it is expected that the number of topics whose proportions are relatively large can be restricted to a smaller value than that for existing methods.

2.3 Numerical Implementation

There are several numerical computing methods for LDAs, and the EM algorithm and Gibbs sampling are frequently used. We used collapsed Gibbs sampling for estimating parameters of the ordinary LDA. The hyperparameters were estimated based on the empirical Bayesian method. For the proposed model and HDSP, Stan was used for estimating parameters [16]. Stan optimizes each parameter using the Hamiltonian Monte Carlo method for maximizing likelihood. We used R3.5.1 [17] for the analysis. The code for estimation of parameters for the ordinary LDA was written by Rcpp [18], and those of the proposed method and HDSP were written by rstan [16]. Ridge regression was conducted using R package glmnet [19], and AUC (Area under the receiver operating characteristics curve) was calculated by pROC [20]. Stan and Rcpp code for the ordinary LDA, HDSP and the proposed method are available under MIT license at <https://github.com/tokui174/BNPDMRtopicmodel>.

3. Performance Evaluation

We evaluated the proposed method using real microbiome data. Taking into account subject attributes for estimating the parameters of a topic model, it is thought that topics obtained by the LDA would be more associated with the attributes. Therefore, we evaluated the predictive ability of the topics for a subject's attributes. Specifically, we calculated AUC for the attributes using a ridge regression model, whose explanatory variables were the topic proportion. After estimating species distribution in the training data using the LDA models, we estimated topic distribution and calculated the AUC for the attributes in the test data.

Further, to evaluate the ability to automatically decide the num-

ber of topics from data, we aggregated the number of topics with the largest or second largest topic proportions for any subject. We also calculated the mean topic proportion of the topics. If a method could automatically decide the number of topics from the data, the number and proportion of the high ranking topics would not change significantly, regardless of the pre-specified number of topics. To evaluate these measures, we varied the pre-specified number of topics from 10 to 50.

We used two sets of microbiome data to perform the evaluation. One data set was retrieved from a study that investigated differences between vaginal microbiomes of pregnant and non-pregnant women [21]. The data set was obtained from R package:NBZIMM [22], and this data set is referred to as pregnant data. As the attributes, the Nugent score and a dummy variable of whether a subject was pregnant were used. The Nugent score is used in diagnosis of bacterial vaginosis, with a score of 4 or higher indicative of bacterial vaginosis. Subsequently, we divided the score into binary based on whether the score was 4 or higher. The other data set was derived from a study that investigated differences in microbiome environment between smokers and nonsmokers [23]. The data were obtained from R package:Gunifrac [24] and are referred to as smoking data. As subject attributes, sex and smoking status were used. The data sets contained many microbiome species variables with limited information. Therefore, we used species variables whose sum of counts for all subjects was larger than 100. Because the AUC could change during every training of the model, we repeated training of each model with each pre-specified number of topics five times and calculated the mean of the AUCs. **Table 1** shows the attributes of the data.

The AUC results are shown in **Table 2**. Although the superiority of the AUCs of the proposed method and other methods differed by case, that of the proposed method was higher than those of the others in many cases, regardless of the data. In a few cases

Table 3 The number of topics with significant proportions for any subject and the mean of the proportions.

Data	Topics	Method	Pre-specified number of topics				
			10	20	30	40	50
Pregnant data	The largest topics	LDA	10(0.56)	20(0.47)	25(0.43)	35(0.44)	42(0.48)
		HDSP	10(0.77)	16(0.72)	15(0.72)	16(0.72)	18(0.71)
		Proposed model	3(0.82)	5(0.63)	3(0.66)	4(0.66)	4(0.70)
	The largest or second largest topics	LDA	10(0.77)	20(0.69)	30(0.62)	39(0.64)	50(0.67)
		HDSP	10(0.91)	19(0.87)	24(0.85)	29(0.86)	32(0.85)
		Proposed model	8(0.97)	6(0.89)	6(0.90)	6(0.91)	9(0.89)
Smoking data	The largest topics	LDA	8(0.86)	16(0.80)	24(0.79)	22(0.74)	20(0.76)
		HDSP	9(0.29)	16(0.23)	15(0.23)	17(0.23)	19(0.21)
		Proposed model	6(0.82)	7(0.77)	9(0.76)	7(0.79)	7(0.43)
	The largest or second largest topics	LDA	10(0.97)	19(0.94)	29(0.93)	32(0.93)	40(0.91)
		HDSP	10(0.49)	19(0.38)	22(0.37)	28(0.37)	26(0.35)
		Proposed model	10(0.95)	12(0.92)	14(0.91)	14(0.93)	9(0.63)

where LDA outperformed the proposed method, the pre-specified numbers of topics were 10 or 20. However, if we increased the pre-specified number of topics, the AUCs of the proposed method were higher than those of LDA. Further, when HDSP was compared with LDA, there was a tendency for the AUCs of LDA to be higher when the pre-specified number of topics was low, but lower when the pre-specified number of topics was high.

Table 3 shows the number of topics with significant proportions for any subject and the mean of the proportions. The values between parentheses represent the mean of the proportions. In many cases, the means of the proportions were large, and it is evident that the significant topics accounted for a large percentage of topics in a subject. Additionally, the numbers were relatively stable for the proposed model, and the numbers were relatively smaller than LDA and HDSP. For LDA and HDSP, the numbers tended to increase according to the increase in the pre-specified number of topics.

4. Discussion

Regarding the result of AUCs, in many cases, the proposed method was superior to LDA and HDSP. One reason is thought to be that by generating topic distribution based on attributes, the topics became more associated with the attributes than those of LDA. Further, it is thought that, by employing the dependent Dirichlet process using the stick-breaking process, more topics were generated in which the proportions depended on the values of the attributes than those of HDSP.

Regarding the number of topics, as Table 3 shows, greater numbers of topics were assigned large topic proportions with an increase in the pre-specified number of topics for LDA and HDSP. However, the reason for using topic models is that they could summarize multivariate data into smaller numbers of topics. As such, interpreting the results of these models may be difficult due to interpreting an increased number of topics if the pre-specified number of topics is increased. Further, the reason for using Bayesian nonparametric model is that we could automatically decide the number of topics from data. The results showed that HDSP did not automatically decide the number of topics, and it is difficult to decide an appropriate number of topics based on the data. As Table 3 shows, the number of topics with the largest proportion for any subject did not change significantly by the pre-specified number of topics for the proposed model, suggesting

that only proportions of topics that were minimally essential for data became large in the proposed model.

The proposed method can also be used for classifying each subject into an enterotype, as well as extracting microbiome communities. Enterotypes are often used for classifying subjects to a few types based on the composition of bacterial species [25]. Enterotypes are often determined by cluster analysis methods, such as the mixture model. In this case, we have to pre-specify the number of types [26]. It is difficult to decide an appropriate number of types from data, and a method to automatically decide the number of types has not been proposed. As Table 3 shows, the number of topics that showed the largest proportions for any subject was unchanged by the pre-specified number of topics for the proposed model. Based on these topics, we could classify subjects into enterotypes by taking into account the attributes.

The limitation of this study is that we evaluated the methods with only two sets of microbiome data. In the future, it is important to evaluate the methods using more varied data. Finally, it is thought that the proposed method can be applied to other data, such as text data or other metagenomics data. Therefore, it might be meaningful to evaluate the model using other kinds of data.

5. Conclusion

We proposed a Bayesian nonparametric topic model for microbiome data and evaluated the method using actual microbiome data. We could extract microbial topics that were more associated with subject attributes and also automatically decide the number of topics from the data.

Acknowledgments We would like to thank the two referees for their thorough review of the manuscript and appropriate comments.

References

- [1] Kim, M., Lee, K.H., Yoon, S.W., Kim, B.S., Chun, J. and Yi, H.: Analytical tools and databases for metagenomics in the next-generation sequencing era, *Genomics & Informatics*, Vol.11, No.3, pp.102–113 (2013).
- [2] Blei, D., Ng, A.Y. and Jordan, M.I.: Latent dirichlet allocation, *Journal of Machine Learning Research*, Vol.3, pp.993–1022 (2003).
- [3] Sankaran, K. and Holmes, S.P.: Latent variable modeling for the microbiome, *Biostatistics* (online), DOI: <https://doi.org/10.1093/biostatistics/kxy018> (2018).
- [4] Āijō, T., Müller, C.L. and Bonneau, R.: Temporal probabilistic modeling of bacterial compositions derived from 16S rRNA sequencing, *Bioinformatics*, Vol.34, No.3, pp.372–380 (2018).
- [5] Yatsunenkov, T., Rey, F.E., Manary, M.J., Trehan, I., Dominguez-Bello,

- M.G., Contreras, M., et al.: Human gut microbiome viewed across age and geography, *Nature*, Vol.486, pp.222–227 (2012).
- [6] Gajer, P., Brotman, R.M., Bai, G., Sakamoto, J., Schütte, U.M., Zhong, X., Koenig, S.S., Fu, L., Ma, Z.S., Zhou, X., Abdo, Z., Forney, L.J. and Ravel, J.: Temporal dynamics of the human vaginal microbiota, *Science Translational Medicine*, Vol.4, No.132, pp.307–323 (2012).
- [7] Blei, D. and McAuliffe, J.D.: Supervised topic models, *Advances in Neural Information Processing Systems (NIPS)*, Platt, J.C., Koller, D., Singer, Y. and Roweis, S. (Eds.) (2007).
- [8] Mimno, D. and McCallum, A.: Topic models conditioned on arbitrary features with Dirichlet multinomial regression, *Proc. Intl. Conf. Uncertainty in Artificial Intelligence in Helsinki*, pp.411–418 (2008).
- [9] Gershman, S.J. and Blei, D.: A tutorial on Bayesian nonparametric models, *Journal of Mathematical Psychology*, Vol.56, No.1, pp.1–12 (2012).
- [10] Teh, Y.W., Jordan, M.I., Beal, M.J. and Blei, D.M.: Hierarchical Dirichlet Processes, *Journal of the American Statistical Association*, Vol.101, No.476, pp.1566–1581 (2006).
- [11] Kim, D. and Oh, A.: Hierarchical Dirichlet scaling process, *Machine Learning*, Vol.106, No.3, pp.387–418 (2017).
- [12] Hosoda, S., Nishijima, S., Fukunaga, T., Hattori, M. and Hamada, M.: Revealing microbial assemblage structure in the human gut microbiome using latent Dirichlet allocation (online), DOI: <https://doi.org/10.1101/664219> (2019).
- [13] Yan, J., Chuai, G., Qi, T., Shao, F., Zhou, C., Zhu, C., Yang, J., Yu, Y., Shi, C., Kang, N. and He, Y.: MetaTopics: An integration tool to analyze microbial community profile by topic model, *BMC Genomics*, Vol.18, No.1, p.962 (2018).
- [14] Dunson, D.B. and Park, J.H.: Kernel stick-breaking process, *Biometrika*, Vol.95, No.2, pp.307–323 (2008).
- [15] Hossain, M.M., Lawson, A.B., Cai, B., Choi, J., Liu, J. and Kirby, R.S.: Space-time stick-breaking processes for small area disease cluster estimation, *Environ. Ecol.*, Vol.20, No.1, pp.91–107 (2013).
- [16] Stan Development Team.: RStan: The R interface to Stan. R package version 2.19.2 (2019) (online), available from <http://mc-stan.org/> (accessed 2019-07-31).
- [17] R Core Team. Vienna, Austria (2019) (online), available from <http://www.R-project.org/> (accessed 2019-07-31).
- [18] Eddelbuettel, D. and Francois, R.: Rcpp: Seamless R and C++ Integration, *Journal of Statistical Software*, Vol.40, No.8, pp.1–18 (2011), available from <http://www.jstatsoft.org/v40/i08/>.
- [19] Friedman, R., Hastie, T. and Tibshirani, R.: Regularization Paths for Generalized Linear Models via Coordinate Descent, *Journal of Statistical Software*, Vol.33, No.1, pp.1–22 (2010), available from <http://www.jstatsoft.org/v33/i01/> (accessed 2019-07-31).
- [20] Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.C., Muller, M. and Siegest, S: pROC: An open-source package for R and S+ to analyze and compare ROC curves, *BMC Bioinformatics*, Vol.12, p.77 (2011).
- [21] Romero, R., Hassan, S.S., Gajer, P., Tarca, A.L., Fadrosch, D.W., Nikita, L., Galuppi, M., Lamont, R.F., Chaemsaitong, P., Miranda, J., Chaiworapongsa, T. and Ravel, J.: The composition and stability of the vaginal microbiota of normal pregnant women is different from that of non-pregnant women, *Microbiome*, Vol.2, No.1 (2014).
- [22] NBZIMM (online), available from <https://github.com/nyiuab/NBZIMM> (accessed 2019-07-31).
- [23] Charlson, E.S., Chen, J., Custers-Allen, R., Bittinger, K., Li, H., Sinha, R., Hwang, J., Bushman, F.D. and Collman, R.G.: Disordered Microbial Communities in the Upper Respiratory Tract of Cigarette Smokers, *PLoS One*, Vol.5, No.12, pp.993–1022 (2010).
- [24] Chen, J.: Package ‘GUniFrac’ Generalized UniFrac distances for comparing microbial communities, Permutational multivariate analysis of variance using multiple distance matrices (online), available from <https://cran.r-project.org/web/packages/GUniFrac/index.html> (accessed 2019-07-31).
- [25] Arumugam, M., Raes, J., Pelletier, E., Paslier, D.L., Yamada, T., et al.: Enterotypes of the human gut microbiome, *Nature*, Vol.473, pp.174–180 (2011).
- [26] Holmes, I., Harris, K. and Quince, C.: Dirichlet multinomial mixtures: Generative models for microbial metagenomics, *PloS One*, Vol.7, No.2 (2012).

Tasuku Okui was born in 1990. He received his Ph.D. degree from Tokyo University in 2018. He is currently an assistant professor at Kyushu University. His research interest is biostatistics and bioinformatics.

(Communicated by Yuki Kato)