Abe *et al. BMC Bioinformatics* 2018, **19**(Suppl 19):519
https://doi.org/10.1186/s12859-018-2530-6

BMC Bioinformatics

RESEARCH Open Access

A latent allocation model for the analysis of microbial composition and disease

Ko Abe[1], Masaaki Hirayama[2], Kinji Ohno[3] and Teppei Shimamura[4*]

From 29th International Conference on Genome Informatics
Yunnan, China. 3-5 December 2018

Abstract

Background: Establishing the relationship between microbiota and specific diseases is important but requires appropriate statistical methodology. A specialized feature of microbiome count data is the presence of a large number of zeros, which makes it difficult to analyze in case-control studies. Most existing approaches either add a small number called a pseudo-count or use probability models such as the multinomial and Dirichlet-multinomial distributions to explain the excess zero counts, which may produce unnecessary biases and impose a correlation structure taht is unsuitable for microbiome data.

Results: The purpose of this article is to develop a new probabilistic model, called BERnoulli and MUltinomial Distribution-based latent Allocation (BERMUDA), to address these problems. BERMUDA enables us to describe the differences in bacteria composition and a certain disease among samples. We also provide a simple and efficient learning procedure for the proposed model using an annealing EM algorithm.

Conclusion: We illustrate the performance of the proposed method both through both the simulation and real data analysis. BERMUDA is implemented with R and is available from GitHub (https://github.com/abikoushi/Bermuda).

Keywords: Latent allocation model, Mixture distribution, Metagenomics

Background

Low-cost metagenomic and amplicon-based sequencing has provided a snapshots of microbial communities and their surrounding environments. One of the goals for case-control studies using microbiome data is to investigate whether cases differ from controls in term of the microbiome composition of a particular body ecosystems (e.g., the gut) and which taxa are responsible for any differences observed [1]. (Here, we use the generic term "taxa" to denote a particular phylogenetic classification.) These studies present microbiome data are represented as count data using operational taxonomic units (OTUs). The number of occurrences of each OTU is measured for each sample drawn from an ecosystem, and the resulting OTU counts are summarized for any level of the bacterial phylogeny, e.g., species, genes, family, order, etc. An important feature of these microbiome count data is that it is highly sparse—i.e., a very high proportion of the data entries are zero—which makes analyzing these data difficult.

A common strategy to handle these excessive zeros is to add a small number called a pseudo-count. For example, Xia et al. (2013) [2] applied a logistic normal model to their data, adjusted by a pseudo-count. Although adding a pseudo-count is a simple and widely used strategy, it can add an unnecessary bias to the data. Further, Weiss et al. (2017) [3] noted that there is no clear consensus on how to choose that value. Another common strategy to mitigate the effects of these excessive zeros is to use non-parametric statistical tests. Wagner et al. (2011) [4] described a test statistic that combines the proportion of zeros in the data with the statistics on values other than 0.

*Correspondence: shimamura@med.nagoya-u.ac.jp
[4]Division of Systems Biology, Nagoya university Graduate School of Medicine, 65 Tsurumai-cho, Showa-ku, 4668550 Nagoya, Japan
Full list of author information is available at the end of the article

However, this statistical test can only be used for comparing two taxa. In addition, the test cannot evaluate the co-occurrence relationships between many taxa, or the effect of combination of taxa. Other strategies include modeling excess zeros using probability models [5, 6]. Such an approach is called "zero-inflated" modeling, and directly models the probability of producing excessive zeros. However, zero-inflated models make an implicit assumption that microbial composition is identical among individuals. Thus, such models cannot capture the effects of individual differences in microbial composition.

Contributions This article proposes a new probabilistic model, called BERnoulli and MUltinomial Distribution-based latent Allocation (BERMUDA), to address these problems. The contributions of our work are summarized below:

1. BERMUDA is a generative statistical model that allows a set of taxa to be explained by unobserved groups and can be used to find the inherent relationship between taxa and a specific disease and to generate microbiome count data through the model.
2. In BERMUDA, the abundance of each taxon can be viewed as a mixture of various groups, which enables us to describe the differences in bacteria composition between samples.
3. We provide a simple and efficient learning procedure for the proposed model using an annealing EM algorithm that reduces the local maxima problem inherent to the traditional EM algorithm. The software package that implements the proposed method in the R environment is available from GitHub (https://github.com/abikoushi/Bermuda).

We describe our proposed model and algorithm in the "Methods" section. We also provide the efficiency of BERMUDA using synthetic and real data in "A Simulation Study" and "Results" sections, respectively.

Methods

Proposed model

Suppose that we observe a microbial count dataset with disease labels, $\{(w_{nk}, y_n); n = 1, \ldots, N, k = 1, \ldots, K)\}$, where w_{nk} is the abundance of the k-th taxon and y_n is a binary outcome such that $y_n = 1$ if the n-th sample has a certain disease and $y_n = 0$ otherwise. Let w_n be the k-th row of matrix $W = (w_{nk})$ and $M_n = \sum_{k=1}^{K} w_{nk}$ be the total reads count of the n-th sample.

We extract the associations between microbial composition and a specific disease by also supposing that there exist L latent clusters that vary with microbial composition and the disease risk. Let $z_n = (z_{n1}, \ldots, z_{nL})^T$ be an indicator vector such that $z_{nl} = 1$ if the n-th sample is in the l-th class and $z_{nl} = 0$ otherwise. We then consider the following generative model:

$$\begin{cases} y_n|z_n, \rho & \sim \text{Bernoulli}\left(\rho_1^{z_{n1}} \cdots \rho_L^{z_{nL}}\right), \\ w_n|M_n, z_n, P & \sim \text{Multinomial}\left(M_n, z_n^T P\right), \\ z_n|\phi & \sim \text{Multinomial}(1, \phi), \\ p_l|\alpha & \sim \text{Dirichlet}(\alpha), \end{cases} \tag{1}$$

where $\rho = (\rho_1, \ldots, \rho_L)^T$ is the probability of developing a certain disease, $P = (p_{lk})$ $(l = 1, \ldots L)$ is an $L \times K$ matrix of the appearance probability of taxa, p_l is the l-th row vector of matrix P, $\phi = (\phi_1 \ldots, \phi_L)^T$ is a vector of each component's mixing ratios, and $\alpha = (\alpha_1, \ldots, \alpha_K)^T$ is a vector of the hyperparameters of the Dirichlet prior distribution. Figure 1 displays the plate notation for the proposed model. The gray node represents an observed variable and the white node represents an unobserved variable; the latent variable z_n affects both y_n and w_n.

If the latent variable z_n is given, the complete likelihood of this model is represented by the following formula:

$$\begin{aligned} &\prod_{n=1}^{N} f(y_n, w_n, z_n|P, \rho, \phi) \\ &= \prod_{n=1}^{N}\prod_{l=1}^{L} \phi_l^{z_{nl}} \left\{\rho_l^{y_n}(1-\rho_l)^{1-y_n}\right\}^{z_{nl}} \left(\frac{\left(\sum_{k=1}^{K} w_{nk}\right)!}{w_{n1}! \cdots w_{nK}!} \prod_{k=1}^{K} \left(p_{lk}^{w_{nk}}\right)^{z_{nl}}\right). \end{aligned} \tag{2}$$

The posterior distribution is then proportional to:

$$\exp\left(\sum_{n=1}^{N} \log f(y_n, w_n, z_n|P, \rho, \phi) + \sum_{l=1}^{L}\sum_{k=1}^{K} (\alpha_k - 1)\log p_{lk}\right). \tag{3}$$

Parameter estimation

We find the maximum a posteriori probability (MAP) estimators, using an annealing EM (AEM) algorithm [7]. One advantage of using an AEM algorithm is that it reduces the local maxima problem from which the traditional EM algorithm suffers.

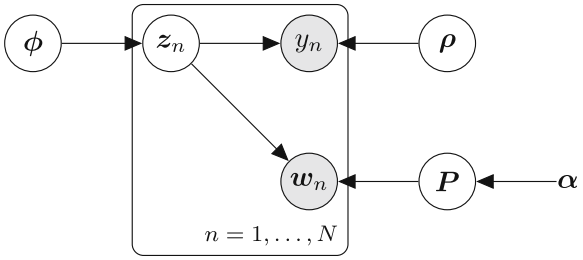

Fig. 1 The plate notation for the proposed model

In the E-step, using the inverse temperature $0 < \beta \leq 1$, we calculate

$$z_{nl}^{(i+1)} = \frac{f\left(y_n, w_n, z_{nl}|P^{(i)}, \rho^{(i)}, \phi^{(i)}\right)^{\beta}}{\sum_{z_{nl}} f\left(y_n, w_n, z_{nl}|P^{(i)}, \rho^{(i)}, \phi^{(i)}\right)^{\beta}}. \quad (4)$$

To simplify the explanation, we set $\gamma = \alpha_k - 1$. From the logarithm of (3), in the M-step, we update the parameters using:

$$\phi_l^{(i+1)} = \frac{1}{N}\sum_{n=1}^{N} z_{nl}^{(i+1)}, \quad (5)$$

$$\rho_l^{(i+1)} = \frac{\sum_{n=1}^{N} z_{nl}^{(i+1)} y_n}{\sum_{n=1}^{N} z_{nl}^{(i+1)}}, \quad (6)$$

$$p_{lk}^{(i+1)} = \frac{\sum_{n=1}^{N} z_{nl}^{(i+1)} w_{nk} + \gamma}{\sum_{n=1}^{N} z_{nl}^{(i+1)} M_n + K\gamma}. \quad (7)$$

If $\gamma = 0$, MAP estimators are equivalent to maximum likelihood estimatos (MLEs).

The procedure of BERMUDA is then summarized as follows:

1 Set β.
2 Arbitrarily choose an initial estimate $P^{(0)}$, $\phi^{(0)}$ and $\rho^{(0)}$. Set $i \leftarrow 0$.
3 Iterate the following two steps until convergence:
   (a) E-step: Compute $z_{nl}^{(i+1)}$ from (4).
   (b) M-step: Compute $P^{(i+1)}$, $\phi^{(i+1)}$ and $\rho^{(i+1)}$ from (5), (6) and (7). Set $i \leftarrow i = i + 1$.
4 Increase β.
5 If $\beta < 1$, repeat from step 3; otherwise stop.

Let $\hat{\phi}$, $\hat{\rho}$ and $\hat{P}$ be MAP estimators of ϕ, ρ and P. If given w_n and the estimators, we can evaluate the probability that the n-th sample has the target disease. The conditional probability is given by

$$\begin{aligned}\tilde{\rho}_n &= \Pr\left(y_n = 1|w_n, \hat{P}, \hat{\rho}, \hat{\phi}\right) \\ &= \frac{\Pr\left(y_n = 1, w_n|\hat{P}, \hat{\rho}, \hat{\phi}\right)}{\Pr\left(w_n|\hat{P}, \hat{\rho}, \hat{\phi}\right)} \\ &= \frac{\sum_{z_{nl}} f\left(y_n = 1, w_n, z_{nl}|\hat{P}, \hat{\rho}, \hat{\phi}\right)}{\sum_{z_{nl}}\sum_{y_n} f\left(y_n, w_n, z_{nl}|\hat{P}, \hat{\rho}, \hat{\phi}\right)}. \end{aligned} \quad (8)$$

The advantage of using the Dirichlet prior distribution is that we can evaluate the abundance of the taxa whose abundance is exactly zero.

The n-th sample is then classified into the l-th class that maximizes the conditional probability given by

$$\hat{z}_{nl} = \frac{f\left(y_n, w_n, z_{nl}|P^{(i)}, \rho^{(i)}, \phi^{(i)}\right)}{\sum_{z_{nl}} f\left(y_n, w_n, z_{nl}|P^{(i)}, \rho^{(i)}, \phi^{(i)}\right)}. \quad (9)$$

In fitting the model, it is important to choose an appropriate number for L. In this article, we use cross-validation to choose L. From (8), we can evaluate the probability that the n-th sample has the target disease. We can then evaluate the log-loss function represented by:

$$LL = -\sum_{j=1}^{J}\left(y_j \log\left(\tilde{\rho}_j\right) + \left(1 - y_j\right)\log\left(1 - \tilde{\rho}_j\right)\right), \quad (10)$$

where J is an arbitrarily chosen subsample size for the validation data. We then select an L which minimizes (10) in this analysis.

A simulation study

In this section, we generated synthetic data and evaluated the performance of our method in order to gain insights into the accuracy of the parameters estimated using the proposed model. A simulation study was conducted as follows. An i.i.d. sample is generated by (1) where we set $N = 700$, $M_n = 10000$, $L = 7$, $\gamma = 10^{-9}$, $\phi = (1/7, \ldots, 1/7)^T$, and $\rho = (0, 3, 0.4, \ldots, 0.9)^T$. P is chosen by a standard Dirichlet random number. We estimated the parameters from 10,000 replicates of the experiment.

Table 1 shows the mean and standard error (se) of the estimates for ρ and ϕ using the proposed method. It can be observed that the estimates are unbiased to the order of 1/100. Figure 2 shows the relationship between estimates and true P in this simulation. In this figure, the points are arranged diagonally, implies that the estimator is unbiased. The overall accuracy of classification by $\hat{z}_{nl}$ (9) is 0.87.

Results

Parkinson's disease data

We first seek to identify the gut dysbiosis in relation to the development of Parkinson's disease (PD), which is thought to be associated with intestinal microbiota. We analyzed

Table 1 The mean and se of $\hat{\rho}$ and $\hat{\phi}$

Cluster	1	2	3	4	5	6	7
ρ	0.30	0.40	0.50	0.60	0.70	0.80	0.90
Mean	0.30	0.40	0.50	0.60	0.70	0.80	0.90
se	0.05	0.05	0.05	0.05	0.05	0.04	0.03
ϕ	0.14	0.14	0.14	0.14	0.14	0.14	0.14
Mean	0.14	0.14	0.14	0.14	0.14	0.14	0.14
se	0.01	0.01	0.01	0.01	0.01	0.01	0.01

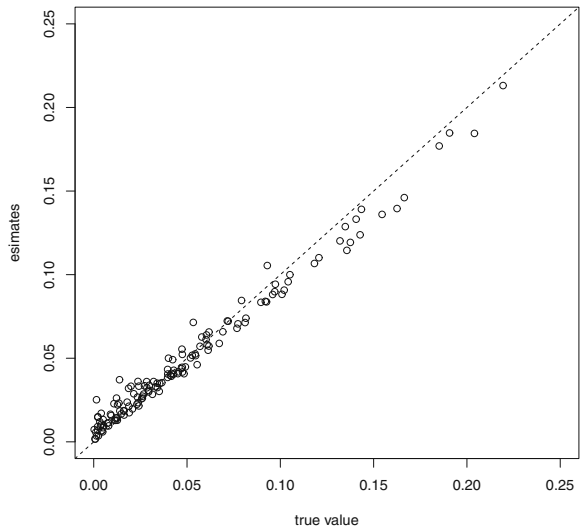

Fig. 2 The comparison of true$P$ and mean of $\hat{P}$

intestinal microbial data for PD cases and controls in three different countries. Scheperjans et al. (2015) [8] , Hill-Burns et al. (2017), [9], Hopfner et al. (2017) [10] and Heintz et al. (2018) [11] conducted case-control studies by sequencing the bacterial 16S ribosomal RNA gene in Finland, the USA, and Germany, respectively.

The OTUs are then mapped to the SILVA taxonomic reference, version 132 (https://www.arb-silva.de/) and the abundances of genus-level taxa are calculated. We focused on the top 20 genera in terms of sample mean of normalized abundance w_{nk}/M_n for 336 PD cases and 277 controls.

We set $\gamma = 10^{-9}$, which is equivalent to giving a weakly informative prior. The number of components $L = 6$ is selected using 10-fold cross-validation (Fig. 3). To ensure the stability, we iterated the cross validation 10,000 times and used the mean of log-loss functions. Figure 3 shows the log-loss functions for different numbers of the components L.

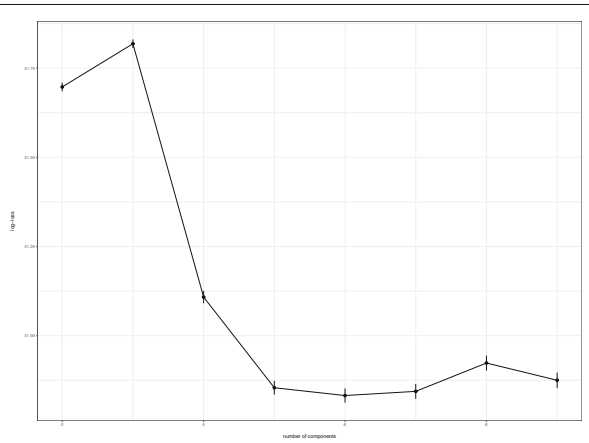

Fig. 3 The behavior of the log-loss functions given by different numbers of components L. The error-bars indicate standard error

Figure 4 presents the estimated appearance probabilities of the 20 genera. The clusters are sorted by estimated PD risk $\hat{\rho}$ (Table 2). As displayed Fig. 4, the distribution of *Prevotella* is quite distinctive, being concentrated in the low-risk cluster of PD. *Faecalibacterium* also tends to be higher in the low-risk cluster. In contrast, *Akkermansia* is concentrated in the high-risk cluster.

Zeller's colorectal carcinoma data

Next, we investigate the identification of gut dysbiosis associated with the development of colorectal cancer (CRC). Zeller et al. (2014) [12] studied gut metagenomes extracted from 157 persons, 91 of whom are CRC patients and 66 are controls. The data are available as an R package "curatedMetagenomicData" (https://github.com/waldronlab/curatedMetagenomicData). In the analysis, we used the abundance of order-level taxa.

While training the model, we set $\gamma = 10^{-9}$. The number of components $L = 3$ was selected using 10-fold cross-validation. To ensure the stability, we iterated the cross validation 10,000 times and used the mean of log-loss functions. Figure 5 shows that log-loss functions for different numbers of the components, L. The clusters are sorted by the estimated CRC risk $\hat{\rho}$.

Figure 6 presents the estimated appearance probabilities for each cluster. Previous studies showed that *Fusobacterium* flourishes in colorectal cancer cells [13]. Figure 6 shows that the abundance of *Fusobacteriales* is positively correlated $\hat{\rho}$. We also observe bacteria, such as *Bacteroides* and *Chlamydiales* with monotonically increasing abundance. This result indicates that BERMUDA can be a valuable tool for discovering new disease-related bacteria.

Discussion

We evaluated the accuracy of parameter estimation using the simulated data. Table 1 and Fig. 2 shows that the proposed method can produce reasonable estimates and classify samples into true groups.

We also applied BERMUDA to the real metagenomic sequencing data in order to identify the associations between the gut microbiota and PD. We compared the results of BERMUDA with those of previous studies. Petrov et al. (2016) [14] reported that the gut microbiota of PD patients contained high levels of *Christensenella*, *Catabacter*, *Lactobacillus*, *Oscillospira*, and *Bifidobacteriumm*, and the control cluster was characterized by increased content of *Dorea*, *Bacteroides*, *Prevotella*, and *Faecalibacterium*. In family level analysis, Hill-Burns et al. (2017) [9] reported PD patients contained high levels of *Bifidobacteriaceae*, *Lactobacillaceae*,

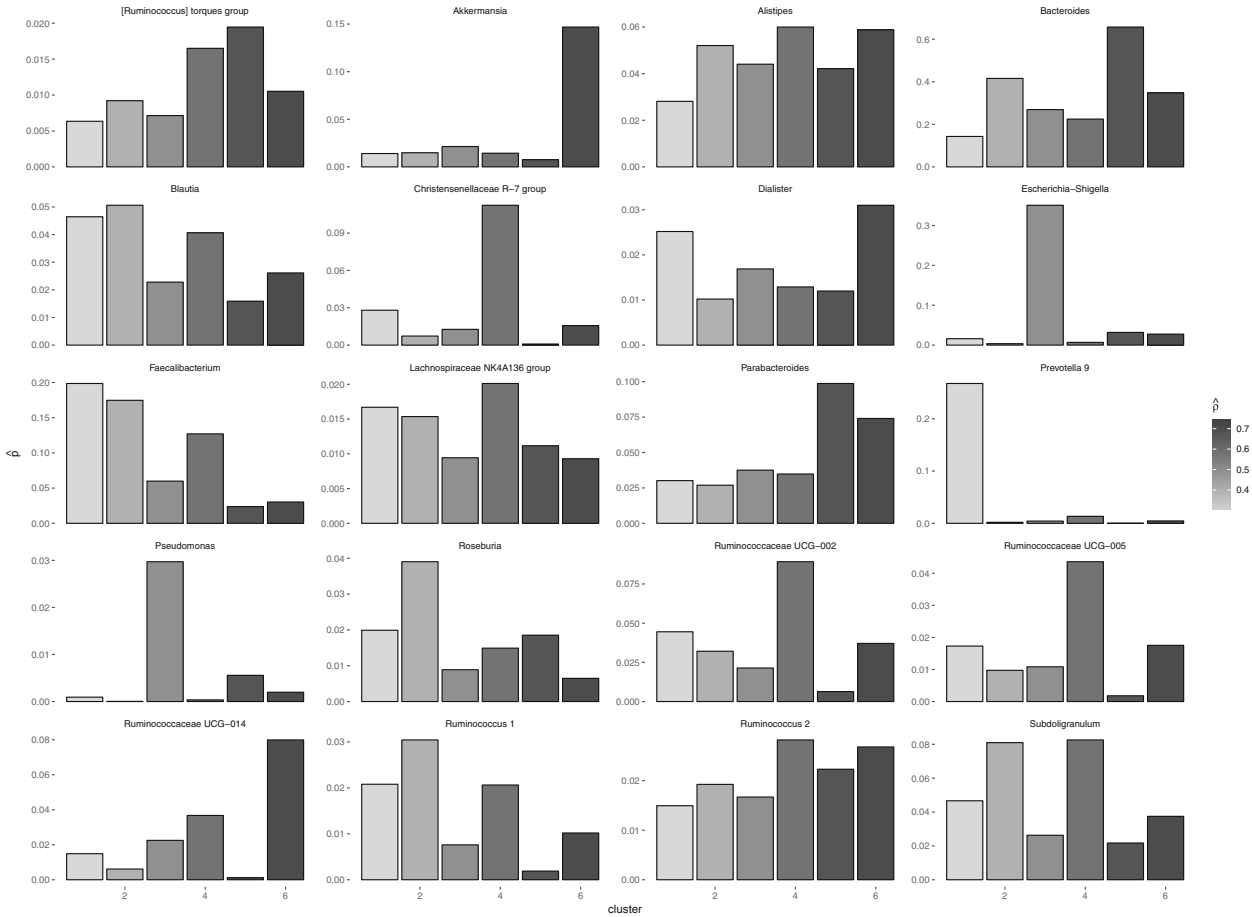

Fig. 4 The appearance probability of the 20 genera ($\hat{P}$)

Tissierellaceae, *Christensenellaceae* and *Verrucomicrobiaceae* and low levels of *Lachnospiraceae*, *Pasteurellaceae*. *Scheperjans* et al.(2015) [8] reported PD patients contained high levels of *Lactobacillaceae*, *Verrucomicrobiaceae*, *Bradyrhizobiaceae* and *Ruminococcaceae* and low levels of *Prevotellaceae* and *Clostridiales Incertae Sedis IV*. *Akkermansia* belongs in *Verrucomicrobiaceae*. Of the *Verrucomicrobiaceae*, it has been suggested that *Akkarmansia* may be related to PD. BERMUDA revealed *Prevotella*, *Faecalibacterium*, and *Akkermansia* associated with PD, which were commonly found in several studies. Thus, the analysis with real data demonstrates that the proposed method can identify the connection between the gut microbiota and PD, with the results are strongly supported by the previous PD research.

Table 2 The estimated disease risk ($\hat{\rho}_l$) within each cluster

	1	2	3	4	5	6
1	0.31	0.43	0.54	0.62	0.71	0.73

Conclusion

We proposed the new probabilistic model BERMUDA to analyze the relationship between microbiota and a specific diseases. Although the existing approaches tend to underestimate individual differences in microbial composition,

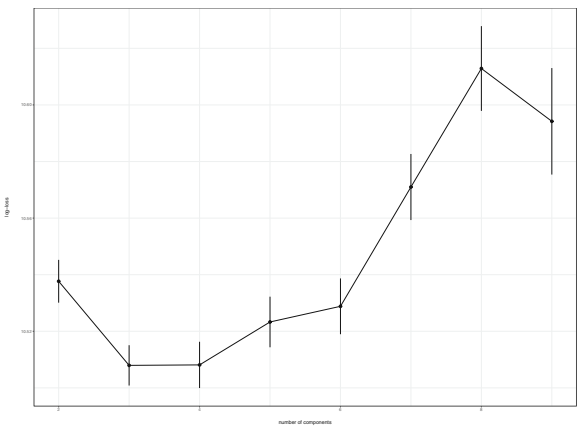

Fig. 5 The behavior of the log-loss functions given by different numbers of components *L*. The error-bars indicate standard error

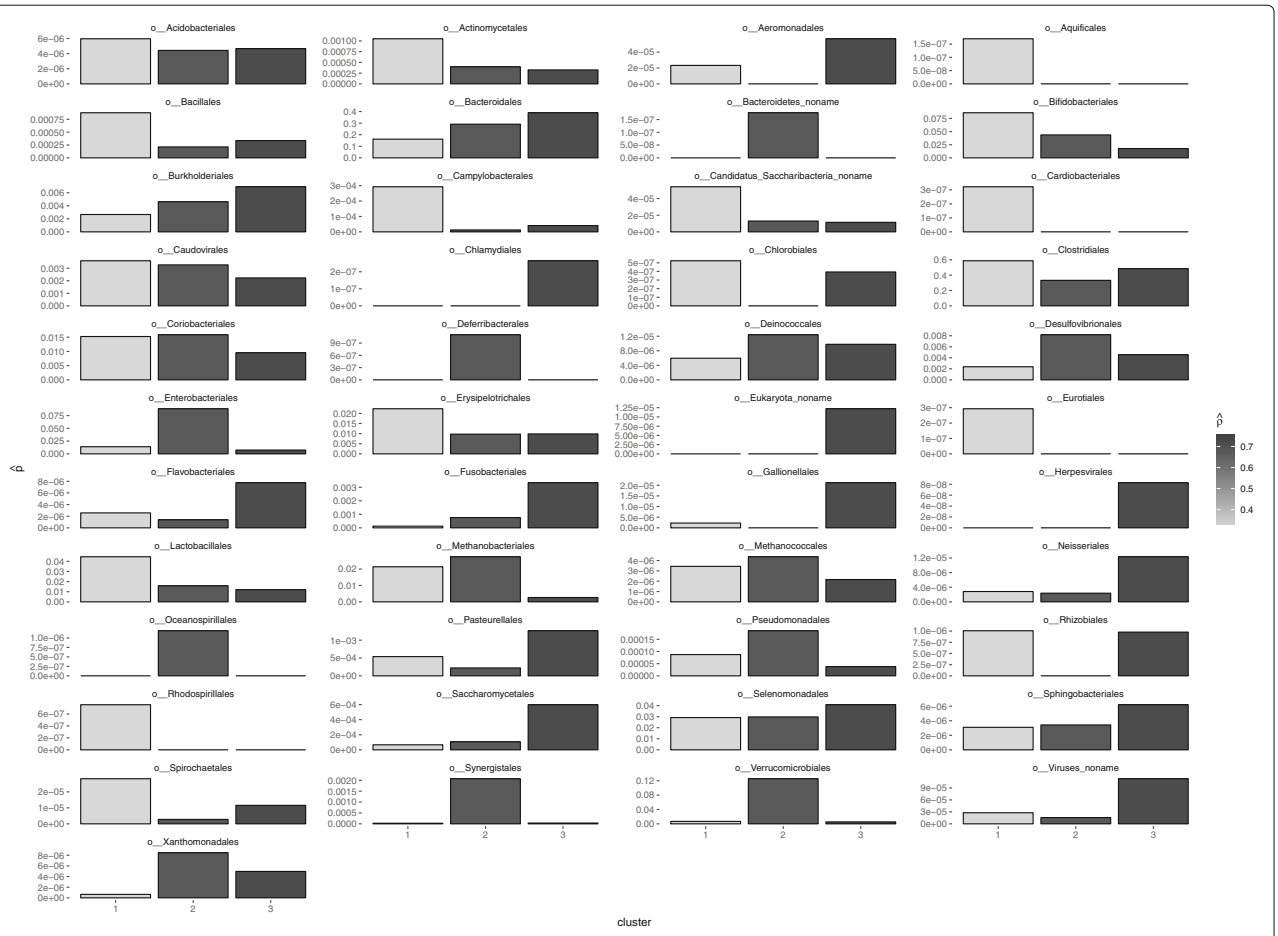

Fig. 6 The appearance probability of the 45 orders ($\hat{P}$)

BERMUDA can take into account these differences and identify combinations of taxa rather than single taxa in the analysis of association with a specific disease risk. We demonstrated the applicability of BERMUDA to microbial analyses with simulation and real data. We expect that BERMUDA can be efficiently applied to studies that seek for an association between gut dysbiosis and a specific disease.

Funding
This work was supported by Grants-in-Aid from the Ministry of Education, Culture, Sports, Science and Technology of Japan (MEXT); Ministry of Health, Labour and Welfare of Japan (MHLW); Japan Agency for Medical Research and Development (AMED), and the Hori Sciences and Arts Foundation. Publication costs are funded by AMED CREST JP18gm1010002.

Availability of data and materials
BERMUDA is implemented with R and is available from GitHub (https://github.com/abikoushi/Bermuda).

About this supplement
This article has been published as part of BMC Bioinformatics Volume 19 Supplement 19, 2018: Proceedings of the 29th International Conference on Genome Informatics (GIW 2018): bioinformatics. The full contents of the supplement are available online at https://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-19-supplement-19.

Authors' contributions
KA and TS designed the proposed algorithm. KO and MH designed the experiments. All authors have read and approved the final manuscript.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details
[1]Division of Systems Biology, Nagoya University Graduate School of Medicine, 65 Tsurumai-cho, Showa-ku, 4668550 Nagoya, Japan. [2]School of Health Sciences, Nagoya University Graduate School of Medicine, 1-1-20 Daiko-Minami, Higashi-ku, 61-8873 Nagoya, Japan. [3]Division of Neurogenetics, Center for Neurological Diseases and Cancer, Nagoya University Graduate School of Medicine, 65 Tsurumai-cho, Showa-ku, 4668550 Nagoya, Japan. [4]Division of Systems Biology, Nagoya university Graduate School of Medicine, 65 Tsurumai-cho, Showa-ku, 4668550 Nagoya, Japan.

Published: 31 December 2018

References

1. Brooks JP. Challenges for case-control studies with microbiome data. Ann Epidemiol. 2016;26(5):336–41.
2. Xia F, Chen J, Fung WK, Li H. A logistic normal multinomial regression model for microbiome compositional data analysi. Biometrics. 2013;69(4): 1053–63.
3. WEISS S, et al. Normalization and microbial differential abundance strategies depend upon data characteristics. Microbiome. 2017;5.1:27.
4. Wagner BD, Robertson CE, Harris JK. Application of two-part statistics for comparison of sequence variant counts. PloS ONE. 2011;6.5:e20296.
5. Chen EZ, Li H. A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. Bioinformatics. 2016;32(17):2611–7.
6. Paulson JN, Stine OC, Bravo HC, Pop M. Differential abundance analysis for microbial marker-gene surveys. Nat Methods. 2013;10(12):1200–2.
7. Ueda N, Nakano R. Deterministic annealing EM algorithm, Neural Networks. Adv Neural Inf Process Syst. 1998;11(2):271–282.
8. Scheperjans F, et al. Gut microbiota are related to Parkinson's disease and clinical phenotype. Mov Disord. 2015;30(3):350–8.
9. Hill-Burns EM, et al. Parkinson's disease and Parkinson's disease medications have distinct signatures of the gut microbiome. Mov Disord. 2017;32(5):739–49.
10. Hopfner F, et al. Gut microbiota in Parkinson disease in a northern German cohort. Brain Res. 2017;1667:41–5.
11. Heintz-Buschart A, et al. The nasal and gut microbiome in Parkinson's disease and idiopathic rapid eye movement sleep behavior disorder. Mov Disord. 2018;33(1):88–98.
12. ZELLER G, et al. Potential of fecal microbiota for early- stage detection of colorectal cancer. Mol Syst Biol. 2014;10.11:766.
13. ZHU Q, et al. The role of gut microbiota in the pathogenesis of colorectal cancer. Tumor Biol. 2013;34.3:1285–300.
14. Petrov VA, et al. Analysis of Gut Microbiota in Patients with Parkinson's Disease. Bull Exp Biol Med. 2016;162(6):734-7.