Abe *et al. BMC Genomics* 2019, **20**(Suppl 2):191
https://doi.org/10.1186/s12864-019-5476-9

BMC Genomics

RESEARCH Open Access

ENIGMA: an enterotype-like unigram mixture model for microbial association analysis

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

From The 17th Asia Pacific Bioinformatics Conference (APBC 2019)
Wuhan, China. 14-16 January 2019

Abstract

Background: One of the major challenges in microbial studies is detecting associations between microbial communities and a specific disease. A specialized feature of microbiome count data is that intestinal bacterial communities form clusters called as "enterotype", which are characterized by differences in specific bacterial taxa, making it difficult to analyze these data under health and disease conditions. Traditional probabilistic modeling cannot distinguish between the bacterial differences derived from enterotype and those related to a specific disease.

Results: We propose a new probabilistic model, named as ENIGMA (Enterotype-like uNIGram mixture model for Microbial Association analysis), which can be used to address these problems. ENIGMA enabled simultaneous estimation of enterotype-like clusters characterized by the abundances of signature bacterial genera and the parameters of environmental effects associated with the disease.

Conclusion: In the simulation study, we evaluated the accuracy of parameter estimation. Furthermore, by analyzing the real-world data, we detected the bacteria related to Parkinson's disease. ENIGMA is implemented in R and is available from GitHub (https://github.com/abikoushi/enigma).

Keywords: Enterotype, Topic model, Unigram mixture, Bayesian inference, Metagenomics

Background

More than 100 trillion microbes live on and within human beings and form of complex microbial communities (microbiota). Most microbes cannot be cultured in laboratories, making it difficult to understand how individual microorganisms mediate vital microbiome-host interactions under health and disease conditions. However, recent important advances in high-throughput sequencing technology have enabled observation of the composition of these intestinal microbes. For each sample drawn from an ecosystem, the number of occurrences of each operational taxonomic units (OTUs) is measured and the resulting OTU abundance can be summarized at any level of the bacterial phylogeny. Discovering recurrent microbial compositional patterns that are related to a specific disease is a significant challenge, as individuals with the same disease typically harbor different microbial community structures.

Recent large-scale sequencing surveys of the human intestinal microbiome, such as the US NIH Human Microbiome Project (HMP) and the European Metagenomics of the Human Intestinal Tract project (MetaHIT), have revealed considerable variations in microbiota composition among individuals [1, 2]. Particularly, community clusters characterized by differences in the abundance of signature taxa, referred to as enterotypes, were first reported in humans [3]. Later, other studies identified enterotype-like clusters that may reflect features of the host-microbial physiology and homeostasis in different species [4, 5] or at different human body sites [6–9]. This

*Correspondence: shimamura@med.nagoya-u.ac.jp
[4]Division of Systems Biology, Nagoya University Graduate School of Medicine, 65 Tsurumai-Cho, Showa-Ku, 466-8550 Nagoya, Japan
Full list of author information is available at the end of the article

microbial stratification has motivated the development of methods for examining unknown clusters of microbial communities.

Probabilistic modeling of microbial metagenomics data often provides a powerful framework for characterizing the microbial community structures [10–12]. For example, Knights et al. [10] applied a Dirichlet prior to a single-level hierarchy and proposed a Bayesian approach for estimating the proportion of microbial communities. Holmes et al. [11] extended the Dirichlet prior to Dirichlet multinomial mixtures to facilitate clustering of microbiome samples. Shafiei et al. [12] proposed a hierarchical model for Bayesian inference of microbial communities (BioMiCo) to identify clusters of OTUs related to environmental factors of interest.

However, such models are not suitable for discovering enterotype-like clusters of microbial communities and associations between microbes and a specific disease for the following two reasons. First, the frameworks of Knights et al. [10] and Holmes et al. [11] do not explicitly address the association between the microbial compositional patterns and environmental depend on the interest. Second, the framework of Shafiei et al. [12] models the structure of each sample using a hierarchical mixture of multinomial distributions that are depends on the factors of interest. Individual host properties such as body mass index, age, or gender cannot explain the observed enterotypes [3]. Thus, such enterotype-like clusters that describes interindividual variability among humans do not always to directly affect host probabilities such as diseases ranging from localized gastroenterologic disorders to neurologic, respiratory, metabolic hepatic, and cardiovascular illnesses.

Here, we introduce a novel probabilistic model of a microbial community structures, named as ENIGMA (Enterotype-like uNIGram mixture model for Microbial Association analysis), to address these problems. ENIGMA includes the following contributions:

1. ENIGMA uses OTU abundances as input and models each sample by the underlying unigram mixture whose parameters are represented by unknown group effects and known effects of interest. The group effects are represented by baseline parameters that change with a latent group of microbial communities. One of the most important features of our model is that the group effects are independent of the effects of interest. This enables the separation of interindividual variability and fixed effects of the host properties related to disease risk.
2. ENIGMA is regarded as Bayesian learning for detecting associations between a community structure and factors of interest. Our model can be used to simultaneously learn how enterotype-like clusters of OTUs contribute to the microbial structure and how microbial compositional patterns may be related to known features of the sample.
3. We provide an efficient learning procedure for ENIGMA by using a Laplace approximation to integrate latent variables and estimate the evidence of the complete model and credible intervals of the parameters. The software package that implements ENIGMA in the R environment is available from https://github.com/abikoushi/enigma.

We describe our proposed framework and algorithm in the "Methods" section. We evaluate the performance of ENIGMA using simulated data in terms of its accuracy to estimate parameters and identify clusters in the "Simulation study" section. We apply ENIGMA to clinical metagenomics data and demonstrate how ENIGMA simultaneously identifies enterotype-like clusters and gut microbiota related to Parkinson's disease (PD) in the "Results on real data" section.

Methods

The key idea of ENIGMA is to adjust for the effects of the enterotype and evaluate the increases and decreases of bacterial abundance associated with environmental factors. Figure 1 shows a conceptual view of ENIGMA. The formal definition of the model is described in the following Mode section. Here we introduce several notations.

Suppose that we observe microbiome count data of K taxa for N samples with M individual host properties, (y_{nk}, x_{nm}) $(n = 1,\dots,n; k = 1,\dots,K; m = 1,\dots,M)$ where $y_{nk} \in \mathbb{N}$ represents the abundance of the k-th taxa in the n-th sample and x_{nm} represents a binary variable such that $x_{nm} = 1$ if the n-th sample has the m-th host property and is otherwise $x_{nm} = 0$. Here the word taxa can represent any level of the bacterial phylogeny, e.g., species, genes, family, order, etc.

Model

Figure 2 shows a plate diagram of the proposed model for metagenome sequencing, where $\boldsymbol{y}_n$ is the read count vector of the n-th sample, $\boldsymbol{x}_n$ is the vector of the host properties of the n-th sample and $z_n \in \{1,\dots,L\}$ is a latent class of the n-th sample. Our model is a simple extension of the unigram mixture model. We assume that each sample is generated from a multinomial distribution with the parameter vector $\boldsymbol{p}_n = (p_{n1},\dots,p_{nK})^\top$. The elements of $\boldsymbol{p}_n$ and p_{nk} $(k = 1,\dots,K)$ are probabilities of the occurrence of the K taxa for the n-th sample. We also assume that p_{nk} can be influenced

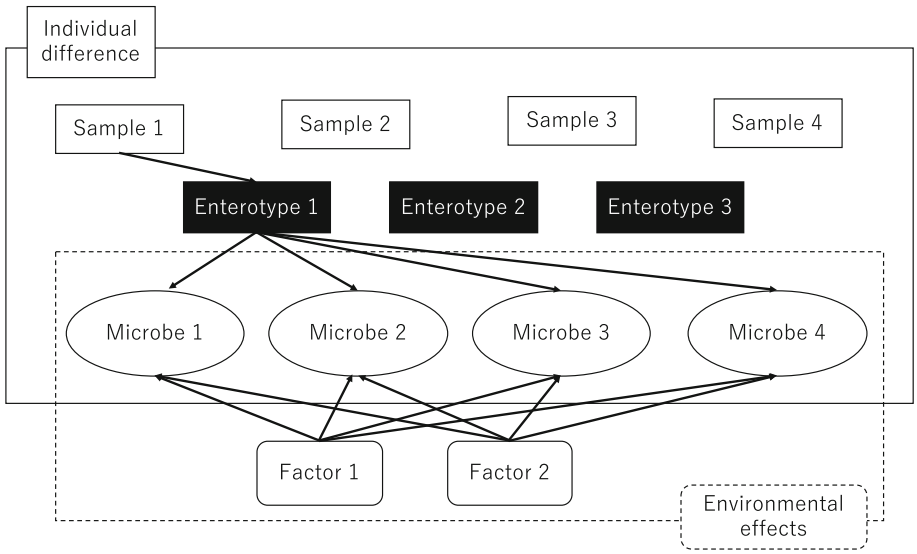

Fig. 1 A conceptual view of ENIGMA

independently by the environmental factor on the taxa that is common to all latent classes and the interindividual factor on the latent enterotype-like classes. More specifically, the generative process of ENIGMA is defined as follows:

$$\begin{aligned} \boldsymbol{y}_n|z_n, \boldsymbol{x}_n, \boldsymbol{\beta} &\sim \text{Multinomial}\left(\boldsymbol{p}_n\right) \\ \boldsymbol{p}_n &= \text{softmax}\left(\boldsymbol{\gamma}_{z_n} + \boldsymbol{x}_n\boldsymbol{B}\right) \\ z_n|\boldsymbol{\pi} &\sim \text{Categorical}(\boldsymbol{\pi}) \\ \boldsymbol{\pi}|\boldsymbol{\alpha} &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\ \boldsymbol{\beta}_m &\sim \text{Normal}_K\left(O_K, \sigma^2 I_K\right) \\ \boldsymbol{\gamma}_l &\sim \text{Normal}_K\left(O_K, \tau^2 I_K\right) \end{aligned} \tag{1}$$

where $\boldsymbol{\gamma}_l$ is baseline parameter (K-dimensional vector) that changes with the latent class, $M \times K$ matrix $\boldsymbol{B} = (\beta_{mk})$ is effect of a environmental factor common to all enterotype-like clusters, $\boldsymbol{\beta}_m$ is a m-th row-vector of $\boldsymbol{B}$, $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_L)$ is a mixing ratio of components, O_K is a K-dimensional zero matrix and I_K is K-dimensional identity matrix. Here, the softmax function is defined by $\text{softmax}(\boldsymbol{x}) = \frac{\exp(\boldsymbol{x})}{\sum_{k=1}^{K}\exp(x_k)}$ for a vector $\boldsymbol{x} = (x_1, \ldots, x_K)^\top$ using an element-wise exponential function and the probability function of categorical distribution is parameterized as $\Pr(z = l|\boldsymbol{\pi}) = \pi_l$, $l \in \{1, \ldots, L\}$. In a Bayesian approach, the prior distributions for $\boldsymbol{\pi}$, $\boldsymbol{\beta}$, and $\boldsymbol{\gamma}_l$ must be defiend. We set a prior based on the Dirichlet distribution for $\boldsymbol{\pi}$, and flat priors to the hyperparameters σ and τ for

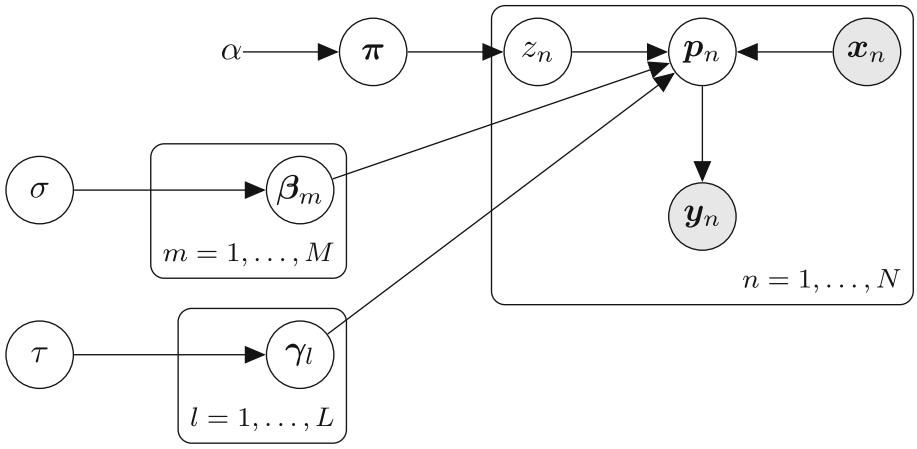

Fig. 2 Plate diagram of the model for ENIGMA $\boldsymbol{y}_n$ is affected from environmental factors $\boldsymbol{x}_n$ and latent variables $\boldsymbol{z}_n$

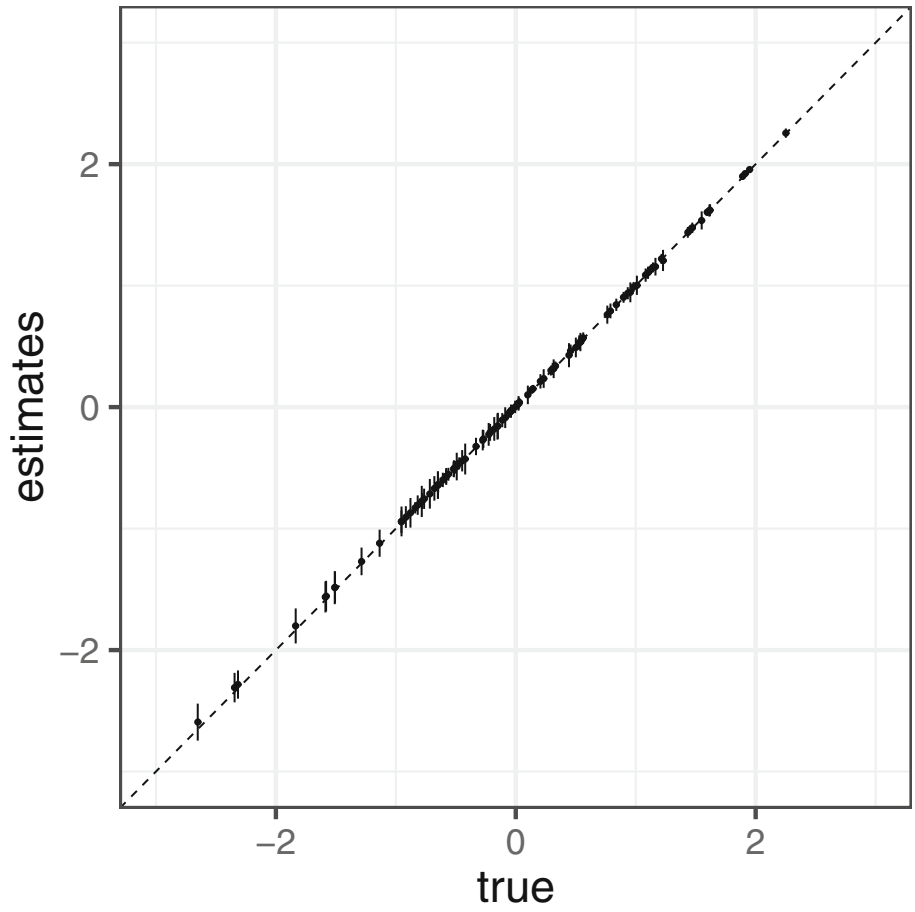

Fig. 3 Simulation result of $\boldsymbol{B}$ The comparison true $\boldsymbol{B}$ and the mean of $\hat{\boldsymbol{B}}$. The error bars indicates SE

$\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, respectively. For the convenience of later section, let $\boldsymbol{p}_l' = \mathrm{softmax}(\boldsymbol{\gamma}_l)$ be the probabilities of the occurrence of bacteria in the latent classes l.

Parameter estimation

Let us denote observed matrix by $\boldsymbol{Y} = (y_{nk})$, $\boldsymbol{X} = (x_{nm})$, the unknown parameters by $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{B}, \boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_L, \sigma, \tau)$, and their prior by $\phi(\boldsymbol{\theta})$. The posterior distribution is represented as follows:

$$p(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{Y}) \propto \prod_{n=1}^{N} p(\boldsymbol{y}_n|z_n, \boldsymbol{x}_n, \boldsymbol{\theta}) p(z_n|\boldsymbol{\theta}) \phi(\boldsymbol{\theta}) \tag{2}$$

First, latent variable z_n must be marginalized. The likelihood is described by

$$\prod_{n=1}^{N} p(\boldsymbol{y}_n|\boldsymbol{x}_n, \boldsymbol{\theta}) = \prod_{n=1}^{N} \sum_{l=1}^{L} \pi_l p(\boldsymbol{y}_n|z_n = l, \boldsymbol{x}_n, \boldsymbol{\theta}). \tag{3}$$

The posterior distribution is proportional to the product of the likelihood and prior density:

$$p(\boldsymbol{\theta}|\boldsymbol{Y}) \propto \exp\left\{ \sum_{n=1}^{N} \log p\left(\boldsymbol{y}_n|\boldsymbol{x}_n, \boldsymbol{\theta}\right) + \log \phi(\boldsymbol{\theta}) \right\}$$

Let $\hat{\boldsymbol{\theta}}$ be the MAP estimator of $\boldsymbol{\theta}$, found by maximizing $\log p(\boldsymbol{\theta}, \boldsymbol{Y}, \boldsymbol{X})$.

We use a Laplace approximation [13] for parameter estimation. A Taylor expansion around $\hat{\boldsymbol{\theta}}$ gives

$$\log p(\boldsymbol{\theta}|\boldsymbol{Y}, \boldsymbol{X}) \approx \log p(\hat{\boldsymbol{\theta}}|\boldsymbol{Y}, \boldsymbol{X}) + \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^\top H(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \tag{4}$$

where $H(\hat{\boldsymbol{\theta}})$ is the Hessian of $\log p(\boldsymbol{\theta}|\boldsymbol{Y}, \boldsymbol{X})$ evaluated at $\hat{\boldsymbol{\theta}}$. Eq. 4 gives

$$p(\boldsymbol{\theta}|\boldsymbol{Y}, \boldsymbol{X}) \approx \frac{1}{C} \exp\left\{ \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^\top H\left(\hat{\boldsymbol{\theta}}\right)\left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right) \right\}$$

where C is a normalizing constant. This relationship shows that $p(\theta|\boldsymbol{Y},\boldsymbol{X})$ can be approximated by the normal distribution $N\left(\hat{\boldsymbol{\theta}}, H^{-1}\left(\hat{\boldsymbol{\theta}}\right)\right)$. Credible intervals can be calculated from this multivariate normal distribution.

We used the stochastic programming language Stan (http://mc-stan.org/) for its implementation. The MAP estimators were obtained by the L-BFGS method. Credible intervals were computed from the using a Stan function to compute the Hessian at the MAP estimates.

After fitting the model, the enterotype-like cluster of each sample must be classified. The conditional probability of $z_n = l$ is

$$\Pr(z_n = l) = \frac{\pi_l p(\boldsymbol{y}_n|\boldsymbol{\gamma}_l, \boldsymbol{\beta}, \boldsymbol{x}_n)}{\sum_{l=1}^{L} \pi_l p(\boldsymbol{y}_n|\boldsymbol{\gamma}_l, \boldsymbol{\beta}, \boldsymbol{x}_n)}. \quad (5)$$

This is the probability that the n-th sample belongs to cluster l. Next, the n-th sample is then classified into the l-th cluster that maximizes the conditional probability given by Eq. 5.

Model Selection

We also examined whether or not the whole set rather than individual bacteria is related to the environmental factors of interest. We compared between the two models when $\boldsymbol{B} \neq \boldsymbol{0}$ and $\boldsymbol{B} = \boldsymbol{0}$. We used the log marginal likelihood as the goodness of fit for model comparison. The marginal likelihood is given by

$$P(\boldsymbol{Y}|\boldsymbol{X}) = \int p(\boldsymbol{Y}, \boldsymbol{\theta}|\boldsymbol{X})\, d\boldsymbol{\theta}. \quad (6)$$

From Eq. 4, we have

$$\int p(\boldsymbol{\theta}, \boldsymbol{Y}|\boldsymbol{X})\, d\boldsymbol{\theta} \approx p\left(\hat{\boldsymbol{\theta}}|\boldsymbol{Y}, \boldsymbol{X}\right) \int \exp\left(\frac{1}{2}\left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right)^{\top} H\left(\hat{\boldsymbol{\theta}}\right)\left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right)\right) d\boldsymbol{\theta}. \quad (7)$$

Thus, the log marginal likelihood is approximated by the following formula:

$$\log P(\boldsymbol{Y}|\boldsymbol{X}) \approx \log p\left(\boldsymbol{Y}|\hat{\boldsymbol{\theta}}, \boldsymbol{X}\right) + \phi\left(\hat{\boldsymbol{\theta}}\right) + \frac{D}{2}\log 2\pi - \frac{1}{2}\log|H\left(\hat{\boldsymbol{\theta}}\right)| \quad (8)$$

where D is the number of free parameters. In model comparison, we choose the model showing larger log marginal likelihood.

Simulation study

To demonstrate the performance of ENIGMA, we conducted several simulation experiments. The synthetic data were naturally produced via our generative process

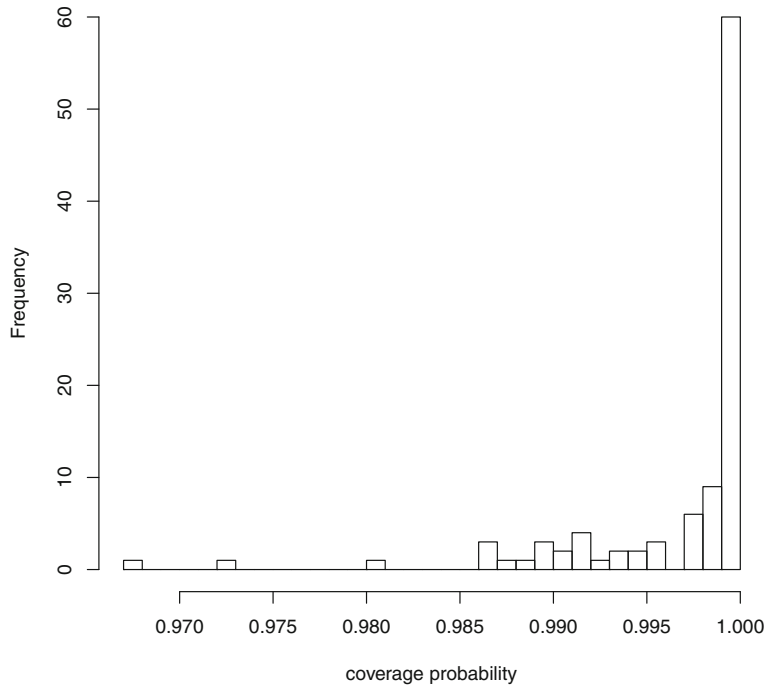

Fig. 4 Coverage probability of $\boldsymbol{B}$. The histogram of coverage probability of $\boldsymbol{B}$

given by Eq. 1. We set $M = 2000$, $L = 3$, $\pi_l = 1/3$, and $\boldsymbol{\alpha} = (1,1,1)^T$. We first generated $\boldsymbol{B}$ and $\boldsymbol{\gamma}_l$ from the standard normal distribution. The variables $\boldsymbol{x}_n$, z_n, and $\boldsymbol{y}_n$ are then sampled from the Bernoulli distribution with probability of 0.5, the categorical distribution, and the multinomial distribution, respectively. For the above parameter settings, we randomly generated a count dataset of 100 taxa for 100 samples for evaluation.

- **Coverage probability (CP)**: The coverage probability is the proportion of the time over which the interval contains the true value. A discrepancy between the coverage probability and the nominal coverage probability frequently occurs. When the actual coverage is greater than the nominal coverage, the interval is referred to as conservative. If the interval is conservative, there is no inconsistency in interpretation.
- **Bias**: The bias of $\boldsymbol{B}$ is defined by the difference between true value and estimated value $E[\hat{\boldsymbol{B}}] - \boldsymbol{B}$.
- **Standard error (SE)**: The standard error is the standard deviation from the estimate. A smaller standard error indicates the higher accuracy of estimation.
- **Root mean squared error (RMSE)**: The RMSE is defined by $\sqrt{E[\left(\hat{\boldsymbol{B}} - \boldsymbol{B}\right)^2]}$. A smaller RMSE indicates the higher accuracy of the estimation.
- **Accuracy**: The accuracy is the percentage of samples correctly classified into original group.

To calcurate these metrics, we detrmined that we calculated the sample means and standard deviations of $\hat{\boldsymbol{B}}$ and $\left(\hat{\boldsymbol{B}} - \boldsymbol{B}\right)^2$ from the 10,000 synthetic datasets.

Figure 3 shows a comparison of the true $\boldsymbol{B}$ and the mean and standard deviation of estimates $\hat{\boldsymbol{B}}$ obtained from the 10,000 simulations. We observed that the points were arranged diagonally, indicating that the estimator of ENIGMA was unbiased. We also calculated the proportion of the time for which the 95% credible interval contains the true value of $\boldsymbol{B}$. We found that this proportion was greater than nominal value of 0.95 for all $\boldsymbol{B}$ in Fig. 4. Table 1 shows the coverage probability (CP), bias, standard error (SE), and RMSE of $\hat{\boldsymbol{B}}$, respectively. We observed that the bias and standard error decreased when β_{mk} was large (i.e. the corresponding abundance was large). We also found that the accuracy of classification given by Eq. 5 was exactly 100%. Thus, these results indicate that ENIGMA can produce reasonable estimates.

Table 1 Coverage probability (CP), bias, standard error (SE), and RMSE of $\hat{\boldsymbol{B}}$

$\boldsymbol{\beta}$	CP	Bias	SE	RMSE	$\boldsymbol{\beta}$	CP	Bias	SE	RMSE
-3.40	0.97	0.08	0.15	0.17	-0.04	1.00	0.01	0.05	0.05
-2.65	0.97	0.06	0.15	0.16	-0.04	1.00	0.01	0.05	0.05
-2.34	0.99	0.04	0.12	0.13	-0.01	1.00	0.01	0.05	0.05
-2.32	0.99	0.03	0.12	0.12	0.01	1.00	0.01	0.04	0.04
-1.83	0.98	0.03	0.14	0.15	0.02	1.00	0.01	0.06	0.06
-1.59	0.99	0.02	0.13	0.13	0.02	1.00	0.01	0.04	0.05
-1.58	0.99	0.03	0.13	0.13	0.03	1.00	0.01	0.04	0.04
-1.51	0.99	0.02	0.14	0.14	0.10	1.00	-0.00	0.08	0.08
-1.51	0.99	0.02	0.13	0.13	0.13	1.00	0.01	0.03	0.03
-1.29	0.99	0.02	0.11	0.11	0.14	1.00	0.01	0.03	0.03
-1.14	0.99	0.01	0.11	0.11	0.21	1.00	0.01	0.06	0.06
-0.95	1.00	0.01	0.09	0.09	0.23	1.00	0.00	0.08	0.08
-0.95	0.99	0.01	0.12	0.12	0.29	1.00	0.01	0.04	0.04
-0.92	1.00	0.01	0.09	0.09	0.31	1.00	0.01	0.05	0.05
-0.88	0.99	0.01	0.12	0.12	0.32	1.00	0.00	0.08	0.08
-0.84	1.00	0.01	0.05	0.05	0.33	1.00	0.01	0.04	0.04
-0.82	1.00	0.01	0.08	0.08	0.44	0.99	-0.02	0.10	0.10
-0.78	0.99	0.01	0.13	0.13	0.46	1.00	0.01	0.05	0.05
-0.78	1.00	0.01	0.07	0.07	0.50	1.00	-0.01	0.08	0.08
-0.76	1.00	0.01	0.08	0.08	0.53	1.00	0.00	0.06	0.06
-0.72	0.99	0.00	0.12	0.12	0.54	1.00	-0.00	0.08	0.08
-0.68	1.00	0.01	0.10	0.10	0.55	1.00	0.01	0.04	0.04
-0.65	0.99	0.01	0.11	0.11	0.55	1.00	0.01	0.03	0.03
-0.65	0.99	0.01	0.11	0.11	0.56	1.00	0.01	0.05	0.05
-0.65	1.00	0.01	0.06	0.06	0.76	1.00	-0.00	0.07	0.07
-0.61	1.00	0.01	0.06	0.06	0.79	1.00	0.00	0.06	0.06
-0.58	1.00	0.01	0.06	0.06	0.84	1.00	0.00	0.05	0.05
-0.58	1.00	0.01	0.07	0.07	0.90	1.00	0.01	0.04	0.04
-0.56	1.00	0.01	0.05	0.05	0.93	1.00	0.00	0.05	0.05
-0.52	1.00	0.01	0.06	0.06	0.96	1.00	-0.01	0.08	0.08
-0.52	1.00	0.01	0.07	0.07	0.98	1.00	0.01	0.04	0.04
-0.51	1.00	0.01	0.04	0.05	1.01	1.00	-0.01	0.08	0.08
-0.50	1.00	0.01	0.05	0.05	1.08	1.00	0.00	0.05	0.06
-0.50	1.00	0.01	0.04	0.04	1.10	1.00	0.00	0.05	0.05
-0.49	0.99	0.00	0.11	0.11	1.13	1.00	0.01	0.04	0.04
-0.47	1.00	0.01	0.05	0.05	1.14	1.00	0.01	0.04	0.04
-0.45	1.00	0.01	0.09	0.09	1.16	1.00	-0.01	0.07	0.07
-0.42	0.99	-0.01	0.13	0.13	1.22	1.00	0.01	0.04	0.04
-0.33	1.00	0.01	0.07	0.07	1.23	1.00	-0.02	0.09	0.09
-0.28	1.00	0.00	0.09	0.09	1.43	1.00	0.00	0.04	0.04
-0.27	1.00	0.01	0.07	0.07	1.45	1.00	0.01	0.04	0.04
-0.23	1.00	0.00	0.09	0.09	1.47	1.00	0.00	0.04	0.04
-0.21	1.00	0.01	0.07	0.07	1.55	1.00	-0.01	0.07	0.08
-0.18	1.00	0.00	0.10	0.10	1.60	1.00	0.01	0.03	0.03
-0.15	0.99	-0.01	0.11	0.11	1.61	1.00	0.00	0.05	0.05
-0.11	1.00	0.01	0.06	0.06	1.89	1.00	0.01	0.03	0.03
-0.09	1.00	0.00	0.09	0.09	1.91	1.00	0.01	0.03	0.03
-0.05	1.00	0.01	0.04	0.04	1.95	1.00	0.01	0.02	0.02
-0.05	1.00	0.01	0.04	0.04	2.25	1.00	0.00	0.04	0.04

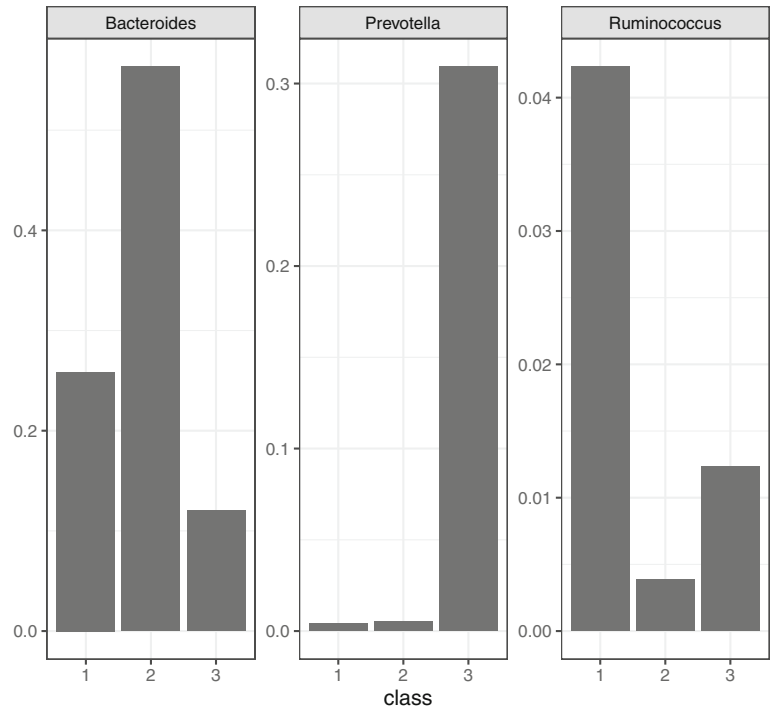

Fig. 5 Probability of occurrence in three bacteria

Results on real data

Arumugam et al. (2011)'s data

We demonstrated that the enterotype-like cluster can be estimated using the data of Arumugam et al. [3]. This data is $N = 33$, $K = 55$. The data of Arumugam et al. [3] does not disclose the total read count. Thus, We used the relative abundance multiplied by 10,000 as y_{nk}. Based on the result of Arumugam et al. [3], the number of latent classes in ENIGMA was chosen to be $L = 3$. We estimated the parameters using the ENIGMA and setting all $\beta_{mk} = 0$ in Eq. 1. We set the hyperparameters of Dirichlet prior $\boldsymbol{\alpha} = (1, \ldots, 1)^\top$, which is equivalent to a noninformative prior.

Arumugam et al. [3] showed that the enterotype is characterized by the differences in the abundance of *Bacteroides*, *Prevotella*, and *Ruminococcus*. Estimates of the probability of occurrence of those bacteria in three clusters are shown in the Fig. 5. Class 1 contains high-level *Ruminococcus*, class 2 contains high-level *Bacteroides*, and class 3 contains high-level *Prevotella*. This result is consistent with that of Arumugam et al. (2011) [3].

Parkinson's disease data

To validate the performance of ENIGMA in discovering clusters of microbial communities and associations between microbes and a specific disease, we applied ENIGMA to the real metagenomic sequencing data from Scheperjans et al. [14], Hill-Burns et al. [15], Heintz-Buschart et al. [16] and Hopfner et al. [17]. The data was analyzed by sequencing the bacterial 16S ribosomal RNA genes sampled from patients with Parkinson's disease (PD) and controls in Finland, USA, and Germany. Table 2 shows the summary statistics of the data. The OTUs were mapped to the SILVA taxonomic reference, version 132 (https://www.arb-silva.de/) and the abundances of family-level taxa were calculated.

To assess the optimal number of clusters, we used the Calinski-Harabasz (CH) Index. It is defined as:

$$\mathrm{CH}_l = \frac{\mathrm{BC}_l/(l-1)}{\mathrm{WC}_l/(n-l)} \tag{9}$$

Table 2 Data summary

	PD	CO
Finland	74	74
German	55	64
USA	207	139

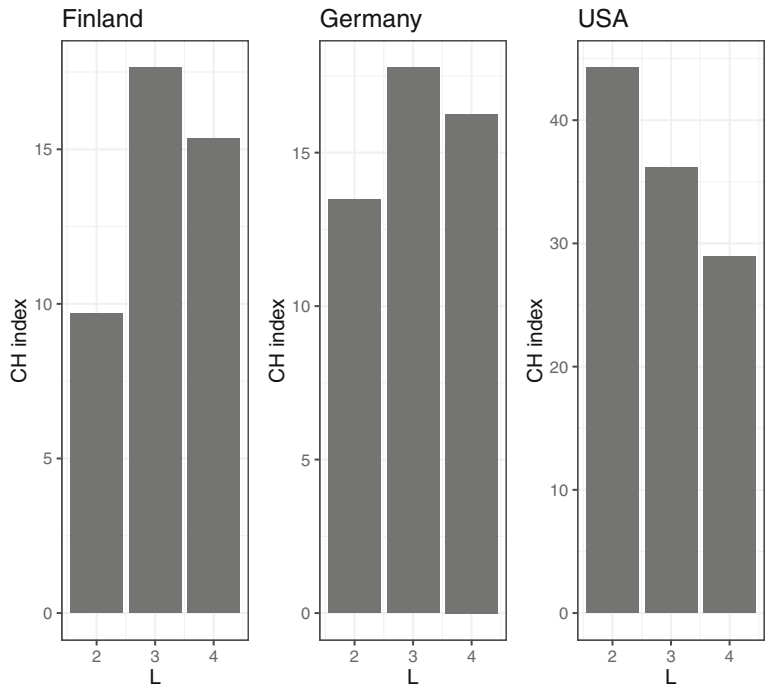

Fig. 6 The CH indexes in each country

where BC_l is the between-cluster sum of squares (i.e. the squared distances between all points i and j, for which i and j are not in the same cluster) and WC_l is the within-clusters sum of squares (i.e. the squared distances between all points i and j, for which i and j are in the same cluster). Here, we used Jensen-Shannon divergence (JSD) as the distance. The JSD between samples $\boldsymbol{a} = (a_1, \ldots, a_K)$ and $\boldsymbol{b} = (b_1, \ldots, b_K)$ is defined as follows:

$$\mathrm{JSD}(\boldsymbol{a}, \boldsymbol{b}) = \frac{1}{2}\left(\sum_{k=1}^{K} a_k \log(a_k/b_k)) + \sum_{k=1}^{K} b_k \log(b_k/a_k))\right). \tag{10}$$

When calculating the JSD, we used the normalized abundance obtained by dividing y_{nk} by the total read count, and 0 was replaced with pseudo count 10^{-6}. We chose the number of clusters L such that CH_l was maximal. To evaluate the CH Index, we use the function `index.G1()` from the R library clusterSim. The number of latent classes in ENIGMA was chosen to be $L = 3$ in Finland and Germany and $L = 2$ in USA by the CH indexes (Fig. 6).

First, we evaluated whether the model assumption was satisfied when using this data. According to Arumugam et al. (2011) [3], the gender of the host is not related with the enterotype. The genders of the subjects were published in the Finland study. We examined the relationship between gender and enterotype-like cluster using the data from Finland. Table 3 shows there was no correlation between them. We conducted a Chi-squared test for independence as shown in Table 3 and the p-value was 0.66.

We evaluated whether the model showing that bacteria were associated with PD is better than the model without the associations in terms of marginal likelihood. Marginal

Table 3 Cross-tabulation of gender and cluster

Class	1	2	3
Female	22	31	21
Male	21	27	26

Table 4 Comparison marginal likelihood

	Finland	Germany	USA
$\mathcal{M}_0$	-442734.62	-5913441.14	-3010279.35
$\mathcal{M}_1$	-355079.50	-3807297.76	-2063932.02

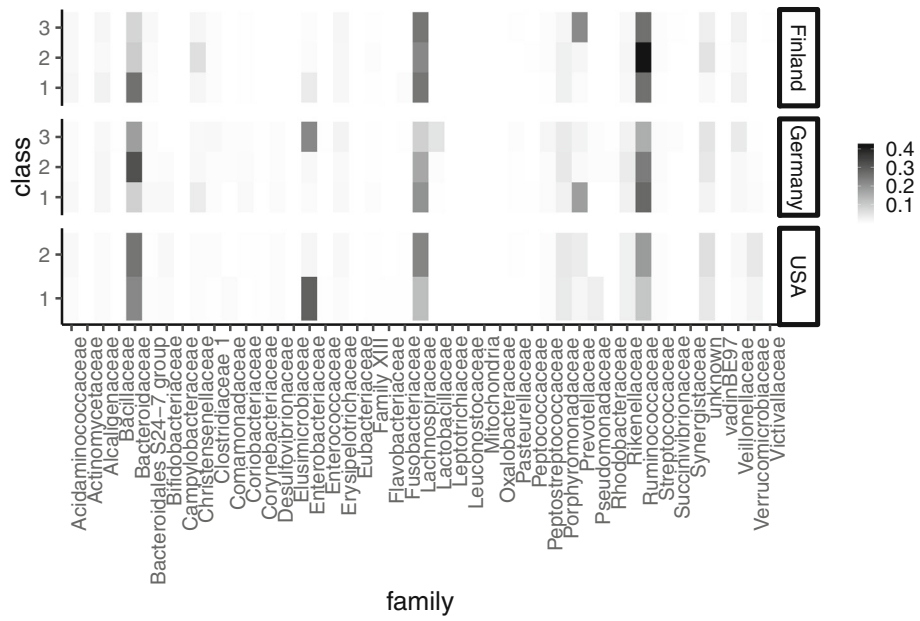

Fig. 7 Heatmap showing $(\hat{\boldsymbol{p}}_l')$. These quantities correspond to the probabilities of the occurrences of bacteria for the three latent classes

likelihood represents the model evidence expressing the preference of the data for different models. Let $\mathcal{M}_1$ be the model which is described by Eq. 1 and $\mathcal{M}_0$ be the model setting all $\beta_{mk} = 0$ in Eq. 1. Table 4 shows that the marginal likelihood of $\mathcal{M}_1$ was greater than $\mathcal{M}_0$. It is preferrd to explain the data by considering the association between the microbiota and PD.

Figure 7 shows the estimated probabilities of the occurrences of bacteria for the three latent classes, $\boldsymbol{p}_l'$, $(l = 1, 2, 3)$. Bacteria detected in fewer three countries were removed. Arumugam et al. [3] showed that enterotype is characterized by the differences in the abundance of *Bacteroides*, *Prevotella*, and *Ruminococcus*. The results of ENIGMA showed the same tendency as the previous survey. Figure 8 shows the $(\hat{\boldsymbol{\gamma}}_l)'$ values and their credible intervals. The top three microbes in each enterotype-like cluster are shown in excerpts for this plot. According to the results of ENIGMA, the abundance of *Enterobacteriaceae* and *Lachnospiraceae* also differed greatly among clusters. Bacterial abundance differed between countries. In the USA, there was a high abundance of *Verrucomicrobiaceae*, while in Finland, few of these bacteria were detected. In contrast, Finland showed more *Prevotellaceae*, with fewer in the in USA it is less.

Table 5 shows the coefficients whose 95% credible intervals did not contain zero in more than two countries. The microbes with these coefficients indicates that the corresponding microbial composition patterns were significantly related to PD. We found that at the family levels, *Clostridiaceae*, *Comamonadaceae*, *Prevotellaceae*, *Actinomycetaceae*, *Bifidobacteriaceae*, *Enterococcaceae*, *Synergistaceae*, *Verrucomicrobiaceae* and *Victivallaceae*, the signs of the coefficients matched in all countries. These results are consistent with those of previous studies. Hill-Burns et al. [15] reported that patients with PD contained high levels of *Bifidobacteriaceae* and *Verrucomicrobiaceae*. Scheperjans et al. [14] reported PD patients contained high levels of *Verrucomicrobiaceae* and low levels of *Prevotellaceae*. Hopfner et al. reported that patients with PD have high levels of *Enterococcaceae*.

We compared ENIGMA to the Wilcoxon rank sum test, a classical methods for identifying bacteria related with an environmental factor of interest [16]. Table 6 shows the bacteria significantly related to PD with p-value < 0.05 in more than two countries. We observed that the bacteria detected by the Wilcoxon test were mostly included in those of ENIGMA (Table 5). Notably, all of the corrected p-values in Table 6 are larger than 0.05. This result shows that ENIGMA was superior to the Wilcoxon rank sum test in terms of identifying a larger number of associations between microbiota and PD.

Finally, we combined the results of ENIGMA to those of PICRUSt (version 1.1.3) [18] in order to

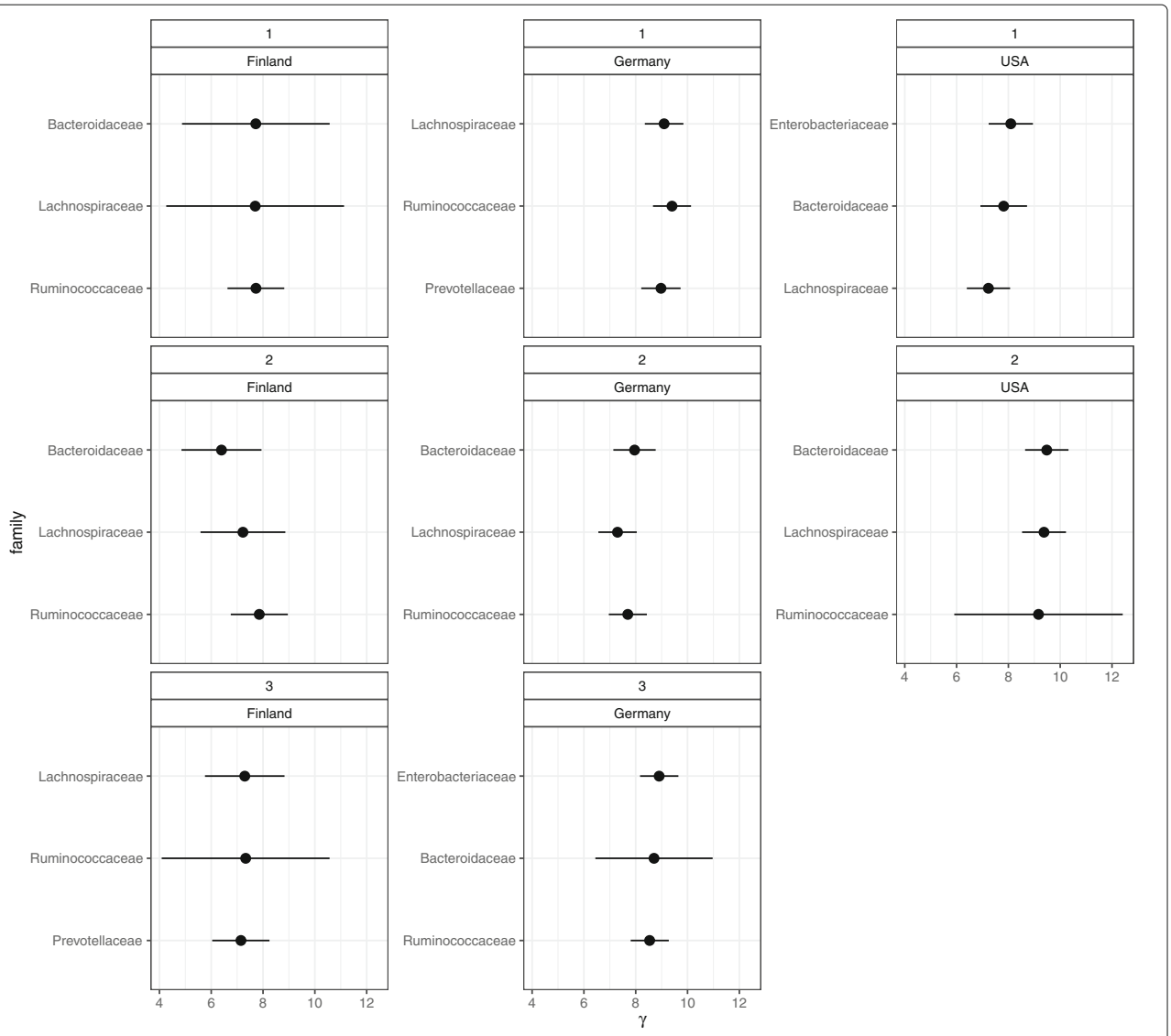

Fig. 8 Top three $\left(\hat{\boldsymbol{\gamma}}_{l}\right)'$ in each country

evaluate which functions are related to PD. In the present study, PICRUSt was performed using the default settings. The Fig. 9 shows the functions exhibiting an increase and decrease from the median, which matched in all countries and clusters with respect to PD and control (CO). This result indicates that ENIGMA is a valuable tool for discovering new disease-related functions.

The analyses using real-world data thus show that ENIGMA can identify enterotype-like clusters and the associations between the gut microbiota and PD. Some of the results were strongly supported by those of previous studies.

Conclusion

We proposed a novel hierarchical Bayesian model, ENIGMA, for discovering the underlying microbial community structures and associations between microbiota and their environmental factors from microbial metagenome data. ENIGMA is based on a probabilistic model of a microbial community structures and supplied with labels for one or more environmental factors of interest for each sample. The structures of each sample are modeled by a multinomial distribution whose parameters are represented independently by group and environmental effects of each sample, which prevents mixing of individual differences and the effects of interest. This

Table 5 Bacteria significantly associated with PD in more than two countries

Family	Finland			Germany			USA		
	$\hat{\beta}$	Lower bound	Upper bound	$\hat{\beta}$	Lower bound	Upper bound	$\hat{\beta}$	Lower bound	Upper bound
Anaeroplasmataceae	-0.87	-1.28	-0.45	-1.69	-2.03	-1.35	-	-	-
Bacteroidales S24-7 group	-0.52	-0.93	-0.11	0.22	-0.12	0.56	-0.80	-1.16	-0.44
Bradyrhizobiaceae	-	-	-	-0.82	-1.17	-0.47	-1.44	-2.21	-0.66
Brevibacteriaceae	-	-	-	-1.02	-1.38	-0.66	-0.65	-1.05	-0.25
Brucellaceae	-	-	-	-1.69	-2.50	-0.87	-1.34	-1.75	-0.92
Clostridiaceae 1	-0.54	-0.96	-0.13	-0.08	-0.42	0.26	-0.52	-0.88	-0.16
Comamonadaceae	-0.85	-1.35	-0.35	-1.27	-1.61	-0.93	-0.21	-0.57	0.15
Elusimicrobiaceae	-4.17	-5.60	-2.74	-2.11	-2.54	-1.68	2.52	1.03	4.01
Intrasporangiaceae	-	-	-	-3.47	-4.86	-2.07	-3.00	-4.72	-1.28
Leuconostocaceae	-2.66	-4.30	-1.02	0.50	0.13	0.86	-1.74	-2.22	-1.25
Moraxellaceae	-	-	-	-1.58	-1.92	-1.24	-0.92	-1.28	-0.56
Pasteurellaceae	-1.62	-2.07	-1.17	0.30	-0.04	0.64	-1.88	-2.25	-1.51
Prevotellaceae	-2.46	-2.87	-2.05	-0.03	-0.37	0.30	-0.53	-0.89	-0.17
Rhodocyclaceae	-	-	-	-3.53	-4.93	-2.13	-0.75	-1.18	-0.32
Actinomycetaceae	0.11	-0.78	1.01	0.42	0.07	0.78	0.91	0.54	1.28
Bacillaceae	1.72	0.34	3.11	-2.35	-2.72	-1.99	0.80	0.43	1.17
Bdellovibrionaceae	-	-	-	1.43	0.40	2.46	3.07	1.78	4.36
Bifidobacteriaceae	1.34	0.82	1.86	0.54	0.20	0.88	0.01	-0.35	0.37
Campylobacteraceae	0.36	-0.31	1.03	4.90	4.48	5.33	0.83	0.46	1.21
Cytophagaceae	-	-	-	2.45	1.56	3.34	1.70	0.27	3.13
Enterococcaceae	3.87	2.70	5.05	0.74	0.40	1.08	0.09	-0.28	0.45
Lactobacillaceae	3.00	2.56	3.43	-0.51	-0.85	-0.18	1.73	1.36	2.09
Leptotrichiaceae	-0.90	-1.89	0.09	2.57	1.88	3.26	0.82	0.36	1.27
Methanobacteriaceae	-	-	-	0.93	0.59	1.27	0.67	0.30	1.04
Mitochondria	0.60	-1.27	2.46	0.73	0.11	1.36	1.57	0.95	2.20
Paenibacillaceae	-	-	-	2.19	1.28	3.10	1.71	1.30	2.12
Planococcaceae	-	-	-	1.06	0.72	1.41	3.26	2.67	3.85
Rhizobiaceae	-	-	-	0.64	0.24	1.03	1.52	1.08	1.95
Streptococcaceae	0.44	0.03	0.86	0.84	0.50	1.17	0.26	-0.10	0.62
Succinivibrionaceae	-0.32	-0.76	0.11	0.74	0.40	1.08	4.31	3.76	4.86
Synergistaceae	1.26	0.80	1.71	0.25	-0.10	0.61	1.44	1.06	1.82
Verrucomicrobiaceae	1.71	1.23	2.19	1.62	1.29	1.96	-0.06	-0.42	0.30
Victivallaceae	0.42	-0.00	0.85	0.68	0.34	1.02	0.93	0.54	1.32

The "-" notation indicates the bacteria undetected in that country

framework enables the model to simultaneously learn (*i*) how microbes contribute to an underlying community structures (cluster) and (*ii*) how microbial compositional patterns are explained by environmental factors of interest. The effectiveness of ENIGMA was evaluated through experiments involving both synthetic and read-world datasets. These newly discovered clusters and associations estimated using ENIGMA can provide insight into the the mechanisms of a microbial communities.

Table 6 *p*-Value of Wilcoxon test

	Finland	Germany	USA
Lachnospiraceae	0.009371	0.719014	0.002839
Lactobacillaceae	0.030404	0.077771	0.000002
Pasteurellaceae	0.006493	0.495315	0.004232
Prevotellaceae	0.001303	0.030892	0.194592

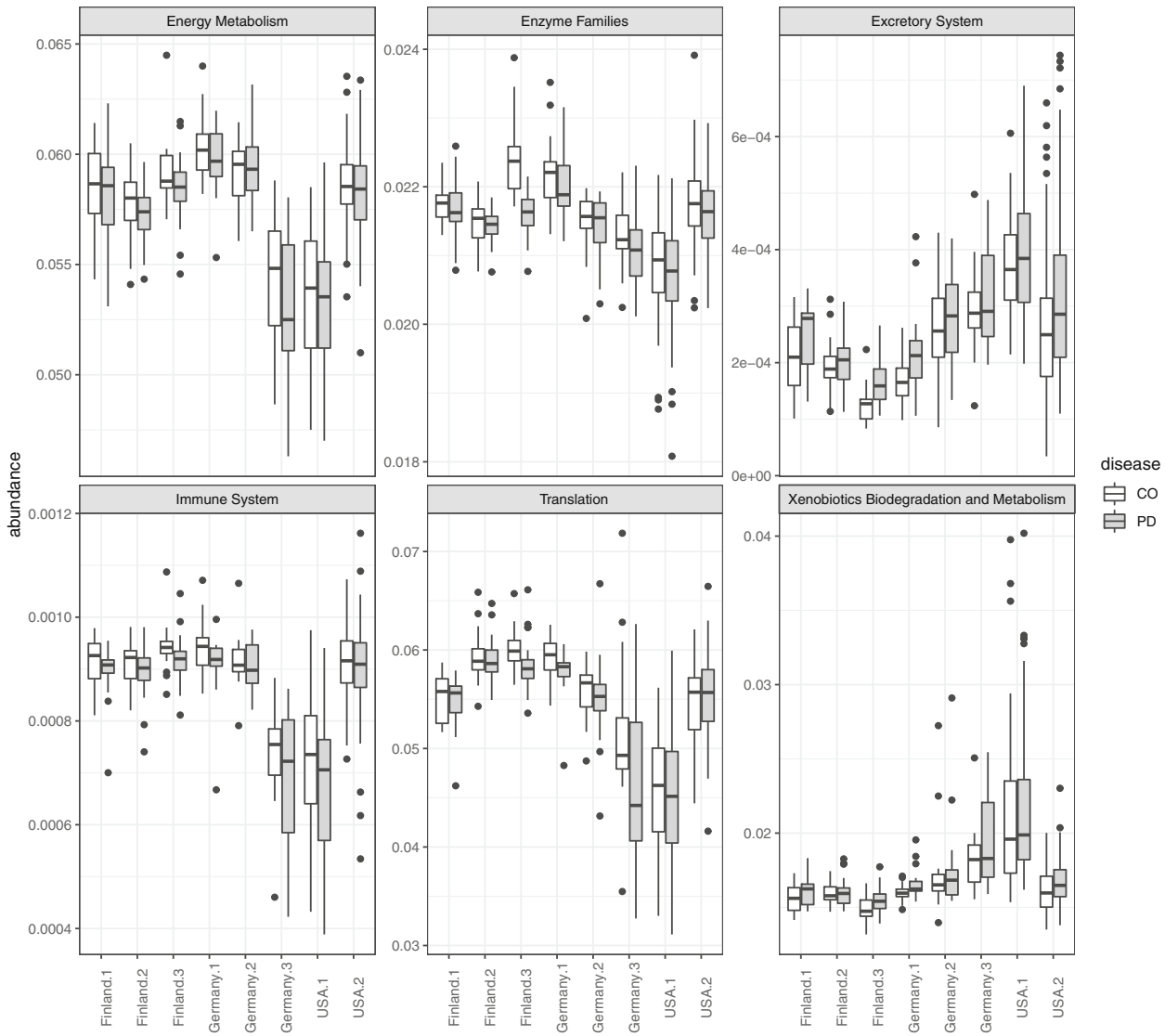

Fig. 9 Predicted functional trait abundance in each country and cluster

The major limitation of ENIGMA is its scalability and efficiency, as the number of the parameters in the model increase proportionally with the number of taxa when the number of environmental factors of interest is large. Further studies should focus on developing a scalable probabilistic model of microbial compositions to analyze underlying microbial structures with a large number of these effects by using sparse parameter estimation [19]. We are also interested in developing a dynamic probabilistic model similar to that reported by Blei and Lafferty [20] for analyzing time-varying bacteria compositions during disease progression.

Abbreviations

CH: Calinski-Harabasz; CO: Control; CP: Coverage probability; JSD: Jensen-Shannon divergence; MAP: Maximum a posteriori; OTU: Operational taxonomic units; PD: Parkinson's disease; RMSE: Root mean squared error; SE: Standard error

Acknowledgements

Not applicable.

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 of this article was sponsored by AMED CREST JP18gm1010002.

Availability of materials

ENIGMA is implemented with R and is available from GitHub (https://github.com/abikoushi/enigma).

About this supplement

This article has been published as part of *BMC Genomics Volume 20 Supplement 2, 2019: Selected articles from the 17th Asia Pacific Bioinformatics Conference*

(APBC 2019): genomics. The full contents of the supplement are available online at https://bmcgenomics.biomedcentral.com/articles/supplements/volume-20-supplement-2.

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, 466-8550 Nagoya, Japan. [2]School of Health Sciences, Nagoya University Graduate School of Medicine, 1-1-20 Daiko-Minami, Higashi-Ku, 461-8873 Nagoya, Japan. [3]Division of Neurogenetics, Center for Neurological Diseases and Cancer, Nagoya University Graduate School of Medicine, 65 Tsurumai-Cho, Showa-Ku, 466-8550 Nagoya, Japan. [4]Division of Systems Biology, Nagoya University Graduate School of Medicine, 65 Tsurumai-Cho, Showa-Ku, 466-8550 Nagoya, Japan.

Published: 10 April 2019

References
1. Huttenhower C, Gevers D, Knight R, Abubucker S, Badger JH, Chinwalla AT, et al. Human Mircrobiome Project Consortium. Structure, function and diversity of the healthy human microbiome. Nature. 2012;486:207–14.
2. Le Chatelier E, Nielsen T, Qin J, Prifti E, Hildebrand F, Falony G, et al. Richness of human gut microbiome correlates with metabolic markers. Nature. 2013;500:541-6.
3. Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, et al. Enterotypes of the human gut microbiome. Nature. 2011;473(7346):174.
4. Moeller AH, Degnan PH, Pusey AE, Wilson ML, Hahn BH, Ochman H. Chimpanzees and humans harbour compositionally similar gut enterotypes. Nat Commun. 3:1179.
5. Hildebrand F, Nguyen TL, Brinkman B, Yunta RG, Cauwe B, Vandenabeele P, et al. Inflammation-associated enterotypes, host genotype, cage and inter-individual effects drive gut microbiota variation in common laboratory mice. Genome Biol. 2013;14(1):R4.
6. Ravel J. Vaginal microbiome of reproductive-age women. Proc Natl Acad Sci USA. 2011;108.Supplement 1:4680-7.
7. Koren O, Knights D, Gonzalez A, Waldron L, Segata N, Huttenhower C, et al. A guide to enterotypes across the human body: meta-analysis of microbial community structures in human microbiome datasets. PLoS Comput Biol. 2013;9(1):e1002863.
8. Ding T, Schloss PD. Dynamics and associations of microbial community types across the human body. Nature. 2014;509:357–60.
9. Zhou Y, Mihindukulasuriya KA, Gao H, La Rose PS, Wylie KM, Martin JC, et al. Exploration of bacterial community classes in major human habitats. Genome Biol. 2014;15:R66.
10. Knights D1, Costello EK, Knight R. Supervised classification of human microbiota. FEMS Microbiol Rev. 2011;35(2):343–59.
11. Holmws I, Harris K, Quince C. Dirichlet multinomial mixtures: generative models for microbial metagenomics. PloS ONE. 2012;7(2):e30126.
12. Shafiei M, Dunn KA, Boon E, MacDonald SM, Walsh DA, Gu H, et al. BioMiCo: a supervised Bayesian model for inference of microbial community structure. Microbiome. 2015;3(1):8.
13. Bishop C. Pattern recognition and machine learning. New York: Springer-Verlag; 2006.
14. Scheperjans F, Aho V, Pereira PA, Koskinen K, Paulin L, Pekkonen E, et al. Gut microbiota are related to Parkinson's disease and clinical phenotype. Mov Disord. 2015;30(3):350–8.
15. Hill-Burns EM, Debelius JW, Morton JT, Wissemann WT, Lewis MR, Wallen ZD, et al. Parkinson's disease and Parkinson's disease medications have distinct signatures of the gut microbiome. Mov Disord. 2017;32(5):739–49.
16. Heintz-Buschart A, Pandey U, Wicke T, Sixel-Döring F, Janzen A, Sittig-Wiegand E, 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.
17. Hopfner F, Künstner A, Müller SH, Künzel S, Zeuner KE, Margraf NG, et al. Gut microbiota in Parkinson disease in a northern German cohort. Brain Res. 1667:41–5.
18. Langille MG, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA, et al. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat Biotechnol. 2013;31(9):814–21.
19. Yang Y, Chen N, Chen T. mLDM: a new hierarchical Bayesian statistical model for sparse microbioal association discovery. bioRxiv. 2016;042630. https://doi.org/10.1101/042630.
20. Blei DM, Lafferty JD. Dynamic topic models. In: Proceedings of the 23rd international conference on Machine learning. New York: ACM; 2006. p. 113–20.