bioRxiv preprint doi: https://doi.org/10.1101/664219; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Revealing microbial assemblage structure in the human gut microbiome using latent Dirichlet allocation

Shion Hosoda*[1,2], Suguru Nishijima[1,2,3], Tsukasa Fukunaga[1,4], Masahira Hattori[1,3,5], and Michiaki Hamada†[1,2,6,7,8]

[1]Graduate School of Advanced Science and Engineering, Waseda University
[2]Computational Bio Big-Data Open Innovation Laboratory (CBBD-OIL), National Institute of Advanced Industrial Science and Technology (AIST)
[3]Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo
[4]Department of Computer Science, Graduate School of Information Science and Engineering, The University of Tokyo
[5]RIKEN Center for Integrative Medical Sciences
[6]Artificial Intelligence Research Center (AIRC), National Institute of Advanced Industrial Science and Technology (AIST)
[7]Graduate School of Medicine, Nippon Medical School
[8]Center for Data Science, Waseda University

*shion_hosoda@asagi.waseda.jp
†mhamada@waseda.jp

bioRxiv preprint doi: https://doi.org/10.1101/664219; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Abstract

Recent research has revealed that there are various microbial species in the human gut microbiome. To clarify the structure of the human gut microbiome, many data mining methods have been applied to microbial composition data. Cluster analysis, one of the key data mining methods that have been used in human gut microbiome research, can classify the human gut microbiome into three clusters, called enterotypes. The human gut microbiome has been suggested to be composed of the microbial assemblages or groups of co-occurring microbes, and one human gut microbiome can contain several microbial assemblages. However, cluster analysis can cluster samples into groups without capturing minor assemblages. In addition, a reliable method of assemblage detection has not been established, and little is known about the distributions of microbial assemblages at a population-level scale. Accordingly, the purpose of this study was to clarify the microbial assemblages in the human gut microbiome. In this study, we detected gut microbiome assemblages using a latent Dirichlet allocation (LDA) method, which was first proposed for the classification of documents in natural language processing. We applied LDA to a large-scale human gut metagenome dataset and found that a four-assemblage LDA model can represent relationships between enterotypes and assemblages with high interpretability. This model indicates that each individual tends to have several assemblages, and each of three assemblages corresponded to each enterotype. However, the C-assemblage can exist in all enterotypes. Interestingly, the dominant genera of the C-assemblage (*Clostridium*, *Eubacterium*, *Faecalibacterium*, *Roseburia*, *Coprococcus*, and *Butyrivibrio*) included butyrate-producing species such as *Faecalibacterium prausnitzii*. Finally, we revealed that genera mainly appearing in the same assemblage were correlated to each other. We conducted an assemblage analysis on a large-scale human gut metagenome dataset using LDA, a powerful method for detection of microbial assemblages. This approach has the potential to reveal the structure of the human gut microbiome.

Keywords: Metagenomics, Latent Dirichlet allocation, Human gut microbiome, Enterotype, Microbial assemblage, Bayesian model, Machine learning

bioRxiv preprint doi: https://doi.org/10.1101/664219; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Introduction

The human gut microbiome varies greatly from person to person, depending on differences among human populations [1] and dietary habits [2]. The differences in gut microbial compositions affect host health and physiology [3], and in some cases, altered microbial compositions are associated with diseases such as inflammatory bowel disease [4], type-1 diabetes [5], colorectal cancer [6], and autism [7, 8]. Recent development of metagenome sequencing technologies have enabled investigations of gut microbial compositions of individuals with ease and rapidity, and many large-scale research projects focused on the human gut microbiome have been conducted [1, 9, 10, 11]. At present, by applying various data mining methods to these massive metagenomic datasets, the structure of the human gut microbiome and the relationship between a host's phenotype and its gut microbial profile can be revealed.

Cluster analysis of samples is one of the widely used data mining methods in metagenomic research. In this approach, individuals are clustered into groups based on similarities in their microbial profiles. For example, Arumugam *et al.* discovered that the gut microbial profiles of individuals can be classified into three types called enterotypes using the partitioning around medoids (PAM) clustering method [12]. In another example, Ding and Schloss reported that the human gut microbiome has considerable inter-individual variation, but the cluster type of an individual was almost unchanged in the sampling period, using the Dirichlet multinomial mixture (DMM) clustering method [13, 14]. Although cluster analysis is a powerful approach for revealing the overall structure of human gut microbiomes, this analysis is strongly affected by the dominant microbes in each individual. Therefore, cluster analyses of samples may ignore the existence of non-dominant but shared microbes among individuals (Fig. 1).

An alternative data mining method is microbial assemblage analysis, which clusters microbes into some assemblages. Here, following Boon *et al.* [15], we define assemblages as groups of microbes that are expected to co-occur. In this view, an individual can have several microbial assemblages. Therefore, this analysis can capture assemblages consisting of non-dominant microbes, unlike a cluster analysis of samples. Shafiei *et al.* developed BioMiCo, which is a Bayesian probabilistic model for microbial assemblage analysis, and discovered host-specific assemblages in human gut metagenomic time-series data [16]. Recently, Cai *et al.* also explored microbial assem-

bioRxiv preprint doi: https://doi.org/10.1101/664219; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

blages using non-negative matrix factorization methods and identified a shift of microbial assemblages in one individual [17]. Microbial assemblage analysis has also been used to track sources of contamination in metagenomic research [18] and to detect assemblage-level metabolic interactions [19], but microbial assemblage analysis based on a massive metagenome dataset has not yet been conducted. As such, the large-scale assemblage structure of human gut microbiomes and the relationship between microbial assemblages and enterotypes are still unknown.

In this study, we conducted a microbial assemblage analysis of a large-scale human gut metagenomic dataset in order to reveal the structures of microbial assemblages in the human gut microbiome. To detect assemblages, we used the latent Dirichlet allocation (LDA) method, which is an unsupervised probabilistic model [20]; LDA was first proposed for the classification of documents in natural language processing, and this method is now widely used in bioinformatics fields such as transcriptome analysis [21], pharmacology [22], and gene function prediction [23].LDA allows one microbe to be assigned to multiple clusters; this characteristic serves as an advantage for modeling microbiome clusters because dependency among microbes result from metabolic functions shared by several microbes, and microbes are interchangeable with other microbes having the same metabolic functions. Yan *et al.* developed MetaTopics [24] and applied it to the human oral metagenomic dataset and the human gut metagenomic dataset. However, they applied MetaTopics to small datasets including fewer than 200 samples and have neither performed detailed analyses nor discussed their findings in detail.

In the present study, we first considered the number of microbial assemblages based on the relationships between microbial assemblages and enterotypes. Next, we found that a four-assemblage model has high interpretability in the context of a large-scale human gut microbiome dataset and discovered that an individual may have not just one microbial assemblage but several assemblages in many cases. Then, we investigated the relationships between enterotypes and microbial assemblages and revealed that three assemblages correspond to each enterotype but that one assemblage can exist in all enterotypes. Thereafter, we examined the human population-level differences in microbial assemblages and found the existence of population-independent microbial assemblages. Finally, we estimated the functions of each assemblage by applying LDA to the functional profiles that with the same samples as the genus data. They are referred to as "functional assem-

bioRxiv preprint doi: https://doi.org/10.1101/664219; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

blages" in later analyses.

Materials and methods

Metagenomic dataset and preprocessing methods

We used the large-scale human gut metagenome dataset constructed by Nishijima *et al.* [25]. This dataset consisted of gut metagenomic data from 861 healthy adults from 12 countries. The taxon of each sequencing read was assigned by mapping the read to a reference genome dataset consisting of 6,149 microbial genomes.

We used genus as the taxonomic rank for each sequencing read because genus rank has been used in previous studies including enterotype analysis. We calculated the normalized number of occurrences of each genus in each individual so that the total number of occurrences of all genera per individual would be 10,000 (because LDA cannot be applied to datasets consisting of fractional values). After these preprocessing steps, the number of different genera included in the dataset became 252.

In functional assemblage analysis, we used the Kyoto Encyclopedia of Genes and Genomes (KEGG) [26] orthology annotated data as functional profiles. This dataset was also constructed by Nishijima *et al.* [25]. We calculated the normalized number of occurrences of each KEGG orthology in each individual, such that the total number of occurrences of all genera per individual would be 100,000. After these preprocessing steps, the number of different KEGG orthologies included in the dataset was 1790.

LDA for modeling the human gut microbiome

The probabilistic LDA model [20] can be used to estimate K microbial assemblages in whole human gut metagenomic datasets, where K is a given parameter. In the LDA model, each metagenome sample s_i ($i \in \{1, ..., 861\}$) has a multinomial distribution with parameter $\theta_i = \{\theta_{i,k}\}_{k=1}^{K}$ over microbial assemblages where $\theta_{i,k}$ is the occurrence probability of the k-th assemblage in the i-th sample; each microbial assemblage a_k ($k \in \{1, ..., K\}$) has a multinomial distribution with parameter $\phi_k = \{\phi_{k,j}\}_{j=1}^{252}$ over genera g_j ($j \in \{1, ..., 252\}$), where $\phi_{k,j}$ is the occurrence probability of the j-th microbe in the k-th assemblage.

bioRxiv preprint doi: https://doi.org/10.1101/664219; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

A microbial assemblage with high probability in an individual means that the individual tends to have that particular microbial assemblage in the gut microbiome, and a genus with high probability in a microbial assemblage means that the microbial assemblage tends to have that particular genus. In addition, the LDA model has prior distributions of θ_i and ϕ_k provided by the Dirichlet distribution whose hyperparameter is α and β, respectively. In this research, we used 0.1 and 0.05 as the initial values of all the elements of α and β, respectively.

The LDA parameters (θ and ϕ) can be learned from the dataset in an unsupervised manner. Various parameter inference methods for the LDA model have been proposed, and we used the variational Bayes (VB) method [20], which maximizes the approximation of a marginal likelihood, called the variational lower bound (VLB) score, by updating the parameter iteratively from random initial parameters. We concluded the iteration of the parameter update when the change in the VLB score between the previous and current step was less than 10^{-6}. Finally, we estimated each θ_i and ϕ_k as the expectation values of the distribution estimated by the VB method. We conducted 10 trials for each $K = 2, 3, 4, 5$ and adopted the estimated parameter with the highest VLB score among all trials for each K. In addition, we updated the hyperparameters α and β from the initial values using a fixed point iteration method in the parameter learning step [27]. Based on previous research about LDA hyperparameter settings [28], we estimated the parameter so that each element of α differed from the others but all elements of β have the same value. Furthermore, in the functional assemblage analysis, the experimental conditions and methods were the same as those explained above. In addition, K was set at the same number as the microbial assemblages.

Calculation method of entropy scores

To quantify the bias in the occurrence frequency for each element in the estimated probability distribution, we calculated the entropy scores of the occurrence distribution of the assemblage for each sample and for each genus. In a multinomial distribution, a high entropy score means that the distribution is similar to the uniform distribution, and a low score means that the distribution tends to take a specific value. The entropy score $H(a|s_i)$ of the

bioRxiv preprint doi: https://doi.org/10.1101/664219; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

assemblages $a = \{a_k\}_{k=1}^{K}$ for each sample s_i was calculated as follows:

$$H(a|s_i) = -\sum_{k=1}^{K} P(a_k|s_i) \log P(a_k|s_i). \tag{1}$$

As $P(a_k|s_i)$ is equal to $\theta_{i,k}$, we can directly calculate this score using the estimated LDA parameters. The entropy score $H(a|g_i)$ of the assemblages $a = \{a_k\}_{k=1}^{K}$ for each genus g_j was calculated as follows:

$$H(a|g_j) = -\sum_{k=1}^{K} P(a_k|g_j) \log P(a_k|g_j), \tag{2}$$

$$P(a_k|g_j) \propto P(g_j|a_k)P(a_k) \tag{3}$$

where $P(g_j|a_k)$ is equal to $\phi_{k,j}$ and the average probability of all $\theta_{i,k}$ was used as $P(a_k)$.

Results

Cluster analysis of the human gut microbiome enterotypes

In order to investigate the relationship between enterotypes and assemblages in the following analysis, we classified individual samples into three clusters using the PAM clustering method. We verified that the dominant genus in each identified cluster were *Bacteroides*, *Prevotella*, and *Blautia*, and these genera were specific to each cluster (Additional File 1, Figure S1). These results are consistent with the previous enterotype research, in which three enterotypes were identified in the human gut microbiome: *Bacteroides* dominant type, *Prevotella* dominant type, and *Ruminococcus* and *Blautia* dominant type [12]. Hence, we called these clusters the B-type, P-type, and R-type, respectively.

Analysis of the human gut microbial assemblage profiles estimated by LDA

We conducted estimates of the 2–5-assemblage LDA model parameters to find the model that has the highest interpretability of relationships between enterotypes and assemblages. Fig. 2 shows the assemblage distributions for

bioRxiv preprint doi: https://doi.org/10.1101/664219; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

each enterotype obtained by each model. The two-assemblage model identified the B-type specific assemblage and P- and R-type specific assemblage (IDs 1 and 2 in Fig. 2a). The three-assemblage model estimated the assemblages corresponding to each enterotype (Fig. 2b). In addition to these enterotype-specific assemblages, the four- and five-assemblage models estimated the general assemblage that appears in *all* the enterotypes (Fig. 2cd). The strength of LDA is that it is possible to obtain such an assemblage. The five-assemblage model estimated two general assemblages (IDs 4 and 5 in Fig. 2d), and it seems to be difficult to identify these assemblages because they are equivalent to enterotypes. In other words, the two assemblages of the five-assemblage model had the same occurrence pattern for enterotypes. Therefore, we used the four-assemblage model in this study. Note that the existence of a general assemblage is not trivial in models with four or more assemblages because there are not always genera that appear in all enterotypes. In the following analysis, we call the assemblages with IDs 1, 2, and 3 the "B-assemblage," "P-assemblage," and "R-assemblage," respectively, because these assemblages appeared specifically in the B-, P-, and R-type individuals, respectively (Fig. 2c). In addition, we call the assemblage with ID 4 the "C-assemblage."

Next, we investigated the kinds of genera that constituted each microbial assemblage. Fig. 3 shows the genus distribution of each microbial assemblage estimated by LDA (*i.e.*, $\phi_k = \{\phi_{k,j}\}_{j=1}^{252}$ in the previous section). B- and P-assemblages mainly consisted of one dominant genus, *Bacteroides* and *Prevotella*, and the relative frequencies were 71% and 66%, respectively. On the other hand, R- and C-assemblages consisted of genera with medium occurrence frequencies. The genera that constituted the R-assemblage were *Blautia* (22%), *Bifidobacterium* (20 %), and *Ruminococcus* (8.6 %), among others. The C-assemblage consisted of *Clostridium* (18%), *Eubacterium* (15 %), and unclassified *Firmicutes* (13 %), among others.

In the LDA model, a genus can appear in several microbial assemblages. We investigated whether genera occurred in just one specific assemblage or not using the entropy scores of the assemblage occurrence distributions for genera (Eq. 2). Fig. 4a shows a histogram of the entropy scores for all genera, and two peaks, at 0.00–0.125 and 0.50–0.75, were observed in the distribution. The former peak represents assemblage-specific genera, and *Bacteroides* and *Prevotella* belonged to this group (Additional File 1, Table S1). The latter peak represents a genus appearing in several but not all assemblages, and *Ruminococcus* and *Blautia* belong to this group (Additional File 1, Table

bioRxiv preprint doi: https://doi.org/10.1101/664219; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

S1). Several genera showed high entropy scores, which indicates they are universal genera among assemblages (Additional File 1, Figure S2). As such, occurrence tendencies of the assemblages for each genus varied (Additional File 1, Table S1).

Then, we calculated the entropy scores of the assemblage occurrence distributions (Eq. 1) for each individual (Fig. 4b). The distribution of the entropy score was unimodal, and the median was 0.7805. These results suggest that most individuals have multiple but not all microbial assemblages. In addition, we examined compositions of microbial assemblages for each individual (Additional File 1, Figure S3) and found that co-occurrence tendencies between microbial assemblages were not uniform. That is, P- and R-assemblages tended to exist exclusively in each sample, but B- and C-assemblages could coexist with other assemblages.

Relationships between microbial assemblages and enterotypes

We investigated the relationships between microbial assemblages and enterotypes to reveal how assemblages appear in enterotypes.

Fig. 5a shows the relationship between microbial assemblages and enterotypes. While B-, P-, and R-assemblages correspond to three enterotypes as mentioned above, the C-assemblage was observed in all three enterotypes. This result suggests that the dominant microbial assemblages of each enterotype differ from each other but there is an assemblage of non-dominant genera that can exist in all enterotypes. To confirm this interpretation, we investigated the occurrence tendency of the genera mainly appearing in each assemblage. Here, each genus was regarded as mainly appearing in the assemblage of the highest $P(a_k|g_i)$ (Eq. 3). Fig. 6 shows the relative abundance of genera in each individual, which was constructed by Nishijima *et al.* [25]. We found that the genera mainly appearing in the C-assemblage occurred in all enterotypes and that the genera mainly appearing in the R-assemblage occurred in R-type individuals. These results were supported by the estimated parameters shown in Fig. 5a. In addition, the genera mainly appearing in the B-assemblage, other than *Bacteroides* occurring in the B-type, was 2.05 times as frequent as in the other enterotypes. The genera mainly appearing in the P-assemblage, other than *Prevotella* occurring in P-type, was 3.00 times as frequent as in the other enterotypes.

bioRxiv preprint doi: https://doi.org/10.1101/664219; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Correlations between microbes in the same or different assemblages

We examined the relationship between microbes in the same or different assemblages to determine the significance of the assemblages. Fig. 7a shows the Spearman's correlation coefficients between the major genera of the B-, P-, R-, and C-assemblages, and suggests that the genera of the same assemblage are correlated with each other. To verify this suggestion, we conducted a correlation test, and most genera from the same assemblage were significantly correlated (Fig. 7b) at $p < 0.01$ (two-sided test, after Benjamini–Hochberg correction[29]). Some genera (*i.e.*, *Eubacterium*, *Faecalibacterium*, *Dorea*, *Ruminococcus*, *Streptococcus*, and *Catenibacterium*) were significantly correlated with many genera in other assemblages. These results are consistent with the fact that their $P(a_k|g_i)$ (in Eq.3) is high for multiple assemblages. For example, *Ruminococcus* has a positive correlation with the genera mainly appearing in the R-assemblage. Indeed, *Ruminococcus* has a high association with the R-assemblage even though its main assemblage is the C-assemblage (Additional File 1, Figure S4). These results indicate that the LDA model can capture the assemblages as groups of correlated genera. Fig. 7a also shows high correlation coefficients between the genera mainly appearing in the R-assemblage. This observation may be affected by many Japanese samples, which have high R-assemblage abundance as mentioned below.

Relationships between microbial assemblages and countries

We investigated the relationships between microbial assemblages and countries to observe tendencies of the assemblages among host countries.

Fig. 5b shows the average assemblage distributions of individuals for each country. We discovered that the occurrence distributions of microbial assemblages vary from country to county; for example, Japan and Austria tend to have R-assemblages while Malawi and Venezuela tend to have P-assemblages. On the other hand, the C-assemblage was frequently found in all countries except Japan. Note that Nishijima *et al.* reported that the Japanese gut microbiome was characterized by low abundance of *Clostridium* and unclassified *Firmicutes*, which are main components of the C-assemblage (Table 2) based on the same dataset [25]. Incidentally, *Eubacterium* and *Faecalibacterium*, which are abundant genera in the C-assemblage, were not less abundant in

bioRxiv preprint doi: https://doi.org/10.1101/664219; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

the Japanese population in comparison with other countries (Additional File 1: Figure S5).

Correlations between microbial assemblages and butyrate-producing functions

Dominant genera in the C-assemblage included butyrate-producing bacteria. Thus, we examined correlations between microbial assemblages and butyrate-producing functions (K00929: butyrate kinase, K01034: acetate CoA/acetoacetate CoA-transferase alpha subunit, and K01035: acetate CoA/acetoacetate CoA-transferase beta subunit). Fig. 8 indicates the Pearson's correlation coefficients between microbial assemblages and butyrate-producing functions, showing that the C-assemblage is positively correlated with all three functions ($p < 0.01$, two-sided test, after Benjamini–Hochberg correction). However, the P- and R-assemblages were negatively correlated with some functions, and B-assemblage was positively correlated with only K00929, concurrent with the finding that *Bacteroides fragilis* has only K00929 of three functions[30].

Functional profiles of each microbial assemblage

To discuss the functional profiles of the microbial assemblages, we applied LDA for individual functional profiles, using the same K number as microbial assemblages, referring to each of them as *functional assemblages*. We first investigated the correlation between microbial assemblages and functional assemblages. Supplementary Figure S6 (Additional File 1) shows that functional assemblages have a one-to-one correspondence with microbial assemblages. Therefore, we regarded functional assemblages as functional profiles of microbial assemblages in further analyses. We determined the abundances of functional categories for each assemblage (Additional File 1: Figure S7) and investigated the assemblages with the largest relative abundance for each functional category (Table 1). This table shows that metabolic functions of glycan/lipid, terpenoid/nucleotide, and vitamin/amino acid are abundant in B-, P-, and R-assemblages, respectively. In addition, no metabolic functions were abundant in the C-assemblage; however, general functional categories including the immune system and translation are abundant.

bioRxiv preprint doi: https://doi.org/10.1101/664219; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Discussion

In this study, we used LDA for the detection of microbial assemblages in population-scale human gut microbiome data and discovered four microbial assemblages. Among these assemblages, while three assemblages (B-, P-, and R-assemblages) specifically emerged in the corresponding enterotypes (B-, P-, and R-types), the C-assemblage was frequently observed in every enterotype. As conventional cluster analysis of the sample focuses on the dominant genus of a cluster and the differences among clusters, the existence of non-dominant but shared microbial assemblages among individuals may have been overlooked. The detection of the C-assemblage suggested that LDA is a powerful approach for revealing the assemblage structure in massive metagenomic datasets. In addition, we found that LDA can estimate assemblages as groups of correlated microbes from the correlation analysis.

To determine K, i.e., the number of assemblages, we used the ad-hoc method to capture characteristics for each enterotype with high interpretability. This task, called "model selection," is typically difficult for mixed models. Some methods for this task have been suggested [31, 32]. Yan *et al.* used cross-validation, one of the methods selecting the model with the highest likelihood against the test data. However, these methods tend to overestimate K, leading to presenting difficulties in clarifying the association between enterotypes and assemblages. Indeed, Yan *et al.* estimated $K = 60$, although the number of samples was less than that of the samples used in this study.

As mentioned above, the genera mainly appearing in the B- and P-assemblages tend to occur in the B- and P-types, respectively. The genera specifically appearing in the B- and P-types were reported to have functions for metabolizing protein/animal fat and carbohydrates, respectively [33], and the genera mainly appearing in the B- and P-assemblages may consequently have functions for metabolizing protein/animal fat and carbohydrates, respectively. We could confirm that lipid metabolism functions were abundant in the B-assemblage through functional assemblage analysis. This result suggests that the B-assemblage in the human gut becomes dominant through a fat-rich diet. In the same way, the genera mainly appearing in the C-assemblage may have functions that do not correspond with dietary habits because they appeared in all enterotypes. This suggestion is concurrent with the finding that immune cells and translation are abundant in the C-assemblage.

There are two notable issues about the C-assemblage. First, the C-

bioRxiv preprint doi: https://doi.org/10.1101/664219; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

assemblage can coexist with all of the other three assemblages, which was found in almost all countries. In other words, genera mainly appearing in the C-assemblage were generalists in the human gut environment [34, 35]. While generalists can adapt to diverse environments, they were not specialized to particular environments unlike specialists. This difference in survival strategy may be the reason the genera mainly appearing in the C-assemblage were not dominant genera in the human gut microbiome. Second, it is therefore possible that the C-assemblage is the core gut microbiome [9, 36]. However, C-assemblage abundance is not consistent from person to person; as such, what determines the existence of C-assemblage in the gut microbiome? The dominant genera of the C-assemblage (*i.e.*, *Clostridium*, *Eubacterium*, *Faecalibacterium*, *Roseburia*, *Coprococcus*, and *Butyrivibrio*) include representative butyrate-producing species (Table 2) [37, 38]. In addition, we found that the C-assemblage had correlations with the three butyrate-producing functions. Butyrate is known to have anti-inflammatory effects [39] and to be associated with IBD, type-2 diabetes, and colorectal cancer [40, 41, 42]. Therefore, C-assemblage abundance may indicate the health of hosts, although the dataset used in this study contains only healthy individuals. In addition, we found that ages and BMI did not relate to the presence of the C-assemblage (Additional File 1: Figure S8). Further research is accordingly required, such as comparisons of C-assemblage abundance between individuals with and without a disease.

We envision two future directions for applications of LDA to metagenomic data. The first is the application to more diverse datasets. Metagenomic data have been sampled from not only human guts but also various environments such as the atmosphere [43], ocean [44], and soil [45]. Application of LDA to these data should help reveal the structure of microbial assemblages at a global-scale [46]. The second is the extension of the LDA model because LDA has high model extensibility. Indeed, many extended LDA models have been proposed for natural language processing [47, 48, 49, 50]. The application of these extended LDA models to metagenomic analysis is a fascinating research focus for further elucidation of microbial assemblage structure. For example, applying supervised topic models [51], which utilize label information to estimate assemblage structures, to patient metagenomic data could detect microbial assemblages related to disease. The pachinko allocation model [47], which models hierarchical assemblage structures, may be useful for revealing sub-assemblages within an assemblage. A transition in assemblage composition could be estimated from time-series data from human gut

bioRxiv preprint doi: https://doi.org/10.1101/664219; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

microbiomes[52] using the topic tracking model[50].

Conclusions

In this study, we conducted an assemblage analysis on a large-scale human gut metagenome dataset using LDA. While three assemblages specifically emerged that corresponded to enterotypes, the C-assemblage was frequently observed in all three enterotypes. Interestingly, the dominant genera of the C-assemblage include representative butyrate-producing species. LDA is a powerful method for detecting microbial assemblages, and it has the potential to reveal the structure of the human gut microbiome.

bioRxiv preprint doi: https://doi.org/10.1101/664219; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

List of abbreviations

LDA: latent Dirichlet allocation **IBD:** inflammatory bowel disease **PAM:** partitioning around medoids **DMM:** Dirichlet multinomial mixture **VB:** variational Bayes **VLB:** variational lower-bound

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

Supplementary material is available from the journal website.

Funding

This work was supported by the Ministry of Education, Culture, Sports, Science and Technology (KAKENHI) (grant numbers JP16H05879, JP16H01318, JP16H02484, and 17K20032 to MH)

Competing interests

The authors declare that they have no competing interests.

Author's contributions

M. Hamada and TF conceived the study. M. Hamada supervised this study. SN and M. Hattori processed the data. SH implemented the method and performed all the computational experiments. SH, SN, TF, and M. Hamada analyzed the results. TF, SH, and M. Hamada wrote the draft manuscript,

bioRxiv preprint doi: https://doi.org/10.1101/664219; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

and SN and M. Hattori revised it critically. All authors read and approved the final manuscript.

Acknowledgements

The computations in this research were performed using the supercomputing facilities at the National Institute of Genetics in Research Organization of Information and Systems.

References

[1] Ken Kurokawa, Takehiko Itoh, Tomomi Kuwahara, Kenshiro Oshima, Hidehiro Toh, Atsushi Toyoda, Hideto Takami, Hidetoshi Morita, Vineet K. Sharma, Tulika P. Srivastava, Todd D. Taylor, Hideki Noguchi, Hiroshi Mori, Yoshitoshi Ogura, Dusko S. Ehrlich, Kikuji Itoh, Toshihisa Takagi, Yoshiyuki Sakaki, Tetsuya Hayashi, and Masahira Hattori. Comparative Metagenomics Revealed Commonly Enriched Gene Sets in Human Gut Microbiomes. DNA Research, 14(4):169–181, January 2007.

[2] Volker Mai, Quintece M. McCrary, Rashmi Sinha, and Michael Glei. Associations between dietary habits and body mass index with gut microbiota composition and fecal water genotoxicity: An observational study in African American and Caucasian American volunteers. Nutrition Journal, 8:49, October 2009.

[3] Ilseung Cho and Martin J. Blaser. The human microbiome: At the interface of health and disease. Nature Reviews Genetics, 13(4):260–270, April 2012.

[4] R. J. Xavier and D. K. Podolsky. Unravelling the pathogenesis of inflammatory bowel disease. Nature, 448(7152):427–434, July 2007.

[5] Adriana Giongo, Kelsey A. Gano, David B. Crabb, Nabanita Mukherjee, Luis L. Novelo, George Casella, Jennifer C. Drew, Jorma Ilonen, Mikael Knip, Heikki Hyöty, Riitta Veijola, Tuula Simell, Olli Simell, Josef Neu, Clive H. Wasserfall, Desmond Schatz, Mark A. Atkinson, and Eric W. Triplett. Toward defining the autoimmune microbiome for type 1 diabetes. The ISME journal, 5(1):82–91, January 2011.

bioRxiv preprint doi: https://doi.org/10.1101/664219; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[6] Aleksandar D. Kostic, Dirk Gevers, Chandra Sekhar Pedamallu, Monia Michaud, Fujiko Duke, Ashlee M. Earl, Akinyemi I. Ojesina, Joonil Jung, Adam J. Bass, Josep Tabernero, José Baselga, Chen Liu, Ramesh A. Shivdasani, Shuji Ogino, Bruce W. Birren, Curtis Huttenhower, Wendy S. Garrett, and Matthew Meyerson. Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. Genome Research, 22(2):292–298, February 2012.

[7] Jennifer G. Mulle, William G. Sharp, and Joseph F. Cubells. The Gut Microbiome: A New Frontier in Autism Research. Current psychiatry reports, 15(2):337, February 2013.

[8] Jose C. Clemente, Luke K. Ursell, Laura Wegener Parfrey, and Rob Knight. The impact of the gut microbiota on human health: An integrative view. Cell, 148(6):1258–1270, March 2012.

[9] Peter J. Turnbaugh, Micah Hamady, Tanya Yatsunenko, Brandi L. Cantarel, Alexis Duncan, Ruth E. Ley, Mitchell L. Sogin, William J. Jones, Bruce A. Roe, Jason P. Affourtit, Michael Egholm, Bernard Henrissat, Andrew C. Heath, Rob Knight, and Jeffrey I. Gordon. A core gut microbiome in obese and lean twins. Nature, 457(7228):480–484, January 2009.

[10] Junjie Qin, Ruiqiang Li, Jeroen Raes, Manimozhiyan Arumugam, Kristoffer Solvsten Burgdorf, Chaysavanh Manichanh, Trine Nielsen, Nicolas Pons, Florence Levenez, Takuji Yamada, Daniel R. Mende, Junhua Li, Junming Xu, Shaochuan Li, Dongfang Li, Jianjun Cao, Bo Wang, Huiqing Liang, Huisong Zheng, Yinlong Xie, Julien Tap, Patricia Lepage, Marcelo Bertalan, Jean-Michel Batto, Torben Hansen, Denis Le Paslier, Allan Linneberg, H. Bjørn Nielsen, Eric Pelletier, Pierre Renault, Thomas Sicheritz-Ponten, Keith Turner, Hongmei Zhu, Chang Yu, Shengting Li, Min Jian, Yan Zhou, Yingrui Li, Xiuqing Zhang, Songgang Li, Nan Qin, Huanming Yang, Jian Wang, Søren Brunak, Joel Doré, Francisco Guarner, Karsten Kristiansen, Oluf Pedersen, Julian Parkhill, Jean Weissenbach, MetaHIT Consortium, Maria Antolin, François Artiguenave, Hervé Blottiere, Natalia Borruel, Thomas Bruls, Francesc Casellas, Christian Chervaux, Antonella Cultrone, Christine Delorme, Gérard Denariaz, Rozenn Dervyn, Miguel Forte, Carsten

bioRxiv preprint doi: https://doi.org/10.1101/664219; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Friss, Maarten van de Guchte, Eric Guedon, Florence Haimet, Alexandre Jamet, Catherine Juste, Ghalia Kaci, Michiel Kleerebezem, Jan Knol, Michel Kristensen, Severine Layec, Karine Le Roux, Marion Leclerc, Emmanuelle Maguin, Raquel Melo Minardi, Raish Oozeer, Maria Rescigno, Nicolas Sanchez, Sebastian Tims, Toni Torrejon, Encarna Varela, Willem de Vos, Yohanan Winogradsky, Erwin Zoetendal, Peer Bork, S. Dusko Ehrlich, and Jun Wang. A human gut microbial gene catalogue established by metagenomic sequencing. Nature, 464(7285):59–65, March 2010.

[11] Tanya Yatsunenko, Federico E. Rey, Mark J. Manary, Indi Trehan, Maria Gloria Dominguez-Bello, Monica Contreras, Magda Magris, Glida Hidalgo, Robert N. Baldassano, Andrey P. Anokhin, Andrew C. Heath, Barbara Warner, Jens Reeder, Justin Kuczynski, J. Gregory Caporaso, Catherine A. Lozupone, Christian Lauber, Jose Carlos Clemente, Dan Knights, Rob Knight, and Jeffrey I. Gordon. Human gut microbiome viewed across age and geography. Nature, 486(7402):222–227, June 2012.

[12] Manimozhiyan Arumugam, Jeroen Raes, Eric Pelletier, Denis Le Paslier, Takuji Yamada, Daniel R. Mende, Gabriel R. Fernandes, Julien Tap, Thomas Bruls, Jean-Michel Batto, Marcelo Bertalan, Natalia Borruel, Francesc Casellas, Leyden Fernandez, Laurent Gautier, Torben Hansen, Masahira Hattori, Tetsuya Hayashi, Michiel Kleerebezem, Ken Kurokawa, Marion Leclerc, Florence Levenez, Chaysavanh Manichanh, H. Bjørn Nielsen, Trine Nielsen, Nicolas Pons, Julie Poulain, Junjie Qin, Thomas Sicheritz-Ponten, Sebastian Tims, David Torrents, Edgardo Ugarte, Erwin G. Zoetendal, Jun Wang, Francisco Guarner, Oluf Pedersen, Willem M. de Vos, Søren Brunak, Joel Doré, MetaHIT Consortium (additional Members), María Antolín, François Artiguenave, Hervé M. Blottiere, Mathieu Almeida, Christian Brechot, Carlos Cara, Christian Chervaux, Antonella Cultrone, Christine Delorme, Gérard Denariaz, Rozenn Dervyn, Konrad U. Foerstner, Carsten Friss, Maarten van de Guchte, Eric Guedon, Florence Haimet, Wolfgang Huber, Johan van Hylckama-Vlieg, Alexandre Jamet, Catherine Juste, Ghalia Kaci, Jan Knol, Karsten Kristiansen, Omar Lakhdari, Severine Layec, Karine Le Roux, Emmanuelle Maguin, Alexandre Mérieux, Raquel Melo Minardi, Christine M'rini, Jean Muller, Raish Oozeer, Julian Parkhill, Pierre Renault, Maria Rescigno, Nicolas Sanchez, Shinichi

bioRxiv preprint doi: https://doi.org/10.1101/664219; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Sunagawa, Antonio Torrejon, Keith Turner, Gaetana Vandemeulebrouck, Encarna Varela, Yohanan Winogradsky, Georg Zeller, Jean Weissenbach, S. Dusko Ehrlich, and Peer Bork. Enterotypes of the human gut microbiome. Nature, 473(7346):174–180, May 2011.

[13] Tao Ding and Patrick D. Schloss. Dynamics and associations of microbial community types across the human body. Nature, 509(7500):357–360, May 2014.

[14] Ian Holmes, Keith Harris, and Christopher Quince. Dirichlet Multinomial Mixtures: Generative Models for Microbial Metagenomics. PLOS ONE, 7(2):e30126, February 2012.

[15] Eva Boon, Conor J. Meehan, Chris Whidden, Dennis H.-J. Wong, Morgan G. I. Langille, and Robert G. Beiko. Interactions in the microbiome: Communities of organisms and communities of genes. FEMS microbiology reviews, 38(1):90–118, January 2014.

[16] Mahdi Shafiei, Katherine A. Dunn, Eva Boon, Shelley M. MacDonald, David A. Walsh, Hong Gu, and Joseph P. Bielawski. BioMiCo: A supervised Bayesian model for inference of microbial community structure. Microbiome, 3:8, 2015.

[17] Yun Cai, Hong Gu, and Toby Kenney. Learning Microbial Community Structures with Supervised and Unsupervised Non-negative Matrix Factorization. Microbiome, 5:110, August 2017.

[18] Dan Knights, Justin Kuczynski, Emily S. Charlson, Jesse Zaneveld, Michael C. Mozer, Ronald G. Collman, Frederic D. Bushman, Rob Knight, and Scott T. Kelley. Bayesian community-wide culture-independent microbial source tracking. Nature Methods, 8(9):761–763, September 2011.

[19] Mahdi Shafiei, Katherine A. Dunn, Hugh Chipman, Hong Gu, and Joseph P. Bielawski. BiomeNet: A Bayesian Model for Inference of Metabolic Divergence among Microbial Communities. PLOS Computational Biology, 10(11):e1003918, November 2014.

[20] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. Journal of Machine Learning Research, 3(Jan):993–1022, 2003.

bioRxiv preprint doi: https://doi.org/10.1101/664219; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[21] Bing Liu, Lin Liu, Anna Tsykin, Gregory J. Goodall, Jeffrey E. Green, Min Zhu, Chang Hee Kim, and Jiuyong Li. Identifying functional miRNA–mRNA regulatory modules with correspondence latent dirichlet allocation. Bioinformatics, 26(24):3105–3111, December 2010.

[22] Yonghui Wu, Mei Liu, W. Jim Zheng, Zhongming Zhao, and Hua Xu. Ranking gene-drug relationships in biomedical literature using latent dirichlet allocation, 2012.

[23] P. Pinoli, D. Chicco, and M. Masseroli. Latent Dirichlet Allocation based on Gibbs Sampling for gene function prediction. In 2014 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology, pages 1–8, May 2014.

[24] Jifang Yan, Guohui Chuai, Tao Qi, Fangyang Shao, Chi Zhou, Chenyu Zhu, Jing Yang, Yifei Yu, Cong Shi, Ning Kang, Yuan He, and Qi Liu. MetaTopics: An integration tool to analyze microbial community profile by topic model. BMC Genomics, 18(1):962, January 2017.

[25] Suguru Nishijima, Wataru Suda, Kenshiro Oshima, Seok-Won Kim, Yuu Hirose, Hidetoshi Morita, and Masahira Hattori. The gut microbiome of healthy Japanese and its microbial and functional uniqueness. DNA research: an international journal for rapid publication of reports on genes and genomes, 23(2):125–133, April 2016.

[26] Minoru Kanehisa and Susumu Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research, 28(1):27–30, January 2000.

[27] Thomas Minka. Estimating a Dirichlet distribution, 2000.

[28] Hanna M. Wallach, David M. Mimno, and Andrew McCallum. Rethinking LDA: Why Priors Matter. Advances in Neural Information Processing Systems 22, pages 1973–1981, 2009.

[29] Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B (Methodological), 57(1):289–300, 1995.

bioRxiv preprint doi: https://doi.org/10.1101/664219; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[30] KO (KEGG ORTHOLOGY) Database. https://www.kegg.jp/kegg/ko.html Accessed 28 Feb. 2019.

[31] Adrian Corduneanu and Christopher M. Bishop. Variational Bayesian model selection for mixture distributions, 2001.

[32] Ryohei Fujimaki and Satoshi Morinaga. Factorized asymptotic bayesian inference for mixture modeling. In Artificial Intelligence and Statistics, pages 400–408, 2012.

[33] Gary D. Wu, Jun Chen, Christian Hoffmann, Kyle Bittinger, Ying-Yu Chen, Sue A. Keilbaugh, Meenakshi Bewtra, Dan Knights, William A. Walters, Rob Knight, Rohini Sinha, Erin Gilroy, Kernika Gupta, Robert Baldassano, Lisa Nessel, Hongzhe Li, Frederic D. Bushman, and James D. Lewis. Linking Long-Term Dietary Patterns with Gut Microbial Enterotypes. Science, 334(6052):105–108, October 2011.

[34] Shubha N. Pandit, Kolasa Jurek, and Cottenie Karl. Contrasts between habitat generalists and specialists: An empirical extension to the basic metacommunity framework. Ecology, 90(8):2253–2262, August 2009.

[35] Sira Sriswasdi, Ching-chia Yang, and Wataru Iwasaki. Generalist species drive microbial dispersion and evolution. Nature Communications, 8(1):1162, October 2017.

[36] Peter J. Turnbaugh and Jeffrey I. Gordon. The core gut microbiome, energy balance and obesity. The Journal of Physiology, 587(Pt 17):4153–4158, September 2009.

[37] Susan E. Pryde, Sylvia H. Duncan, Georgina L. Hold, Colin S. Stewart, and Harry J. Flint. The microbiology of butyrate formation in the human colon. FEMS microbiology letters, 217(2):133–139, December 2002.

[38] Petra Louis and Harry J. Flint. Diversity, metabolism and microbial ecology of butyrate-producing bacteria from the human large intestine. FEMS microbiology letters, 294(1):1–8, May 2009.

[39] Harry Sokol, Bénédicte Pigneur, Laurie Watterlot, Omar Lakhdari, Luis G. Bermúdez-Humarán, Jean-Jacques Gratadoux, Sébastien Blugeon, Chantal Bridonneau, Jean-Pierre Furet, Gérard Corthier, Corinne Grangette, Nadia Vasquez, Philippe Pochart, Germain Trugnan, Ginette

bioRxiv preprint doi: https://doi.org/10.1101/664219; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Thomas, Hervé M. Blottière, Joël Doré, Philippe Marteau, Philippe Seksik, and Philippe Langella. Faecalibacterium prausnitzii is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients. Proceedings of the National Academy of Sciences of the United States of America, 105(43):16731–16736, October 2008.

[40] Yukihiro Furusawa, Yuuki Obata, Shinji Fukuda, Takaho A. Endo, Gaku Nakato, Daisuke Takahashi, Yumiko Nakanishi, Chikako Uetake, Keiko Kato, Tamotsu Kato, Masumi Takahashi, Noriko N. Fukuda, Shinnosuke Murakami, Eiji Miyauchi, Shingo Hino, Koji Atarashi, Satoshi Onawa, Yumiko Fujimura, Trevor Lockett, Julie M. Clarke, David L. Topping, Masaru Tomita, Shohei Hori, Osamu Ohara, Tatsuya Morita, Haruhiko Koseki, Jun Kikuchi, Kenya Honda, Koji Hase, and Hiroshi Ohno. Commensal microbe-derived butyrate induces the differentiation of colonic regulatory T cells. Nature, 504(7480):446–450, December 2013.

[41] Jeremy K. Nicholson, Elaine Holmes, James Kinross, Remy Burcelin, Glenn Gibson, Wei Jia, and Sven Pettersson. Host-Gut Microbiota Metabolic Interactions. Science, 336(6086):1262–1267, June 2012.

[42] Valentina Tremaroli and Fredrik Bäckhed. Functional interactions between the gut microbiota and host metabolism. Nature, 489(7415):242–249, September 2012.

[43] Susannah G. Tringe, Tao Zhang, Xuguo Liu, Yiting Yu, Wah Heng Lee, Jennifer Yap, Fei Yao, Sim Tiow Suan, Seah Keng Ing, Matthew Haynes, Forest Rohwer, Chia Lin Wei, Patrick Tan, James Bristow, Edward M. Rubin, and Yijun Ruan. The Airborne Metagenome in an Indoor Urban Environment. PLOS ONE, 3(4):e1862, April 2008.

[44] J. Craig Venter, Karin Remington, John F. Heidelberg, Aaron L. Halpern, Doug Rusch, Jonathan A. Eisen, Dongying Wu, Ian Paulsen, Karen E. Nelson, William Nelson, Derrick E. Fouts, Samuel Levy, Anthony H. Knap, Michael W. Lomas, Ken Nealson, Owen White, Jeremy Peterson, Jeff Hoffman, Rachel Parsons, Holly Baden-Tillson, Cynthia Pfannkoch, Yu-Hui Rogers, and Hamilton O. Smith. Environmental Genome Shotgun Sequencing of the Sargasso Sea. Science, 304(5667):66–74, April 2004.

bioRxiv preprint doi: https://doi.org/10.1101/664219; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[45] Noah Fierer, Jonathan W. Leff, Byron J. Adams, Uffe N. Nielsen, Scott Thomas Bates, Christian L. Lauber, Sarah Owens, Jack A. Gilbert, Diana H. Wall, and J. Gregory Caporaso. Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. Proceedings of the National Academy of Sciences, 109(52):21390–21395, December 2012.

[46] Samuel Chaffron, Hubert Rehrauer, Jakob Pernthaler, and Christian von Mering. A global network of coexisting microbes from environmental and whole-genome sequence data. Genome Research, 20(7):947–959, July 2010.

[47] Wei Li and Andrew McCallum. Pachinko Allocation: DAG-structured Mixture Models of Topic Correlations. In Proceedings of the 23rd International Conference on Machine Learning, ICML '06, pages 577–584, New York, NY, USA, 2006. ACM.

[48] Simon Lacoste-Julien, Fei Sha, and Michael I. Jordan. DiscLDA: Discriminative Learning for Dimensionality Reduction and Classification. Advances in Neural Information Processing Systems 21, pages 897–904, 2009.

[49] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled LDA: A Supervised Topic Model for Credit Attribution in Multi-labeled Corpora. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1, EMNLP '09, pages 248–256, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[50] Tomoharu Iwata, Shinji Watanabe, Takeshi Yamada, and Naonori Ueda. Topic Tracking Model for Analyzing Consumer Purchase Behavior. In Proceedings of the 21st International Jont Conference on Artifical Intelligence, IJCAI'09, pages 1427–1432, San Francisco, CA, USA, 2009. Morgan Kaufmann Publishers Inc.

[51] Jon D. Mcauliffe and David M. Blei. Supervised Topic Models. Advances in Neural Information Processing Systems 20, pages 121–128, 2008.

[52] Lawrence A. David, Arne C. Materna, Jonathan Friedman, Maria I. Campos-Baptista, Matthew C. Blackburn, Allison Perrotta, Susan E.

bioRxiv preprint doi: https://doi.org/10.1101/664219; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Erdman, and Eric J. Alm. Host lifestyle affects human microbiota on daily timescales. Genome Biology, 15:R89, July 2014.

Figures

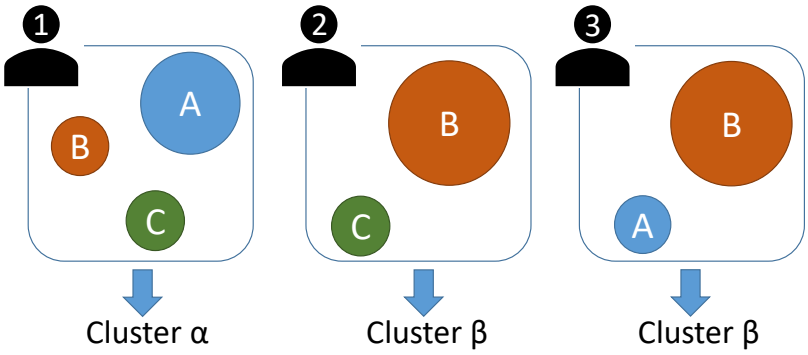

Figure 1: Schematic illustration of microbial assemblage and cluster analysis for the human gut microbiome, where A, B, and C show microbial assemblages with circle size indicating abundance. The cluster of each individual was determined by the dominant assemblage. However, a cluster analysis cannot capture the non-dominant but shared microbes among samples like those comprising assemblage C.

bioRxiv preprint doi: https://doi.org/10.1101/664219; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

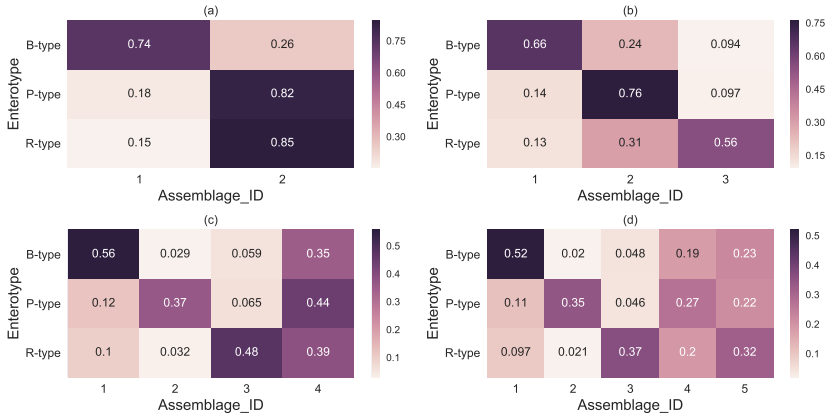

Figure 2: Assemblage distributions for each enterotype. Each row shows a distribution obtained by averaging the estimated assemblage distributions of individuals corresponding to each enterotype. The x- and y-axes represent the microbial assemblages and enterotypes, respectively. Darker colors indicate higher probabilities, and each number inside the partition indicates a different probability. (a), (b), (c), and (d) indicate 2–5-assemblage LDA models, respectively.

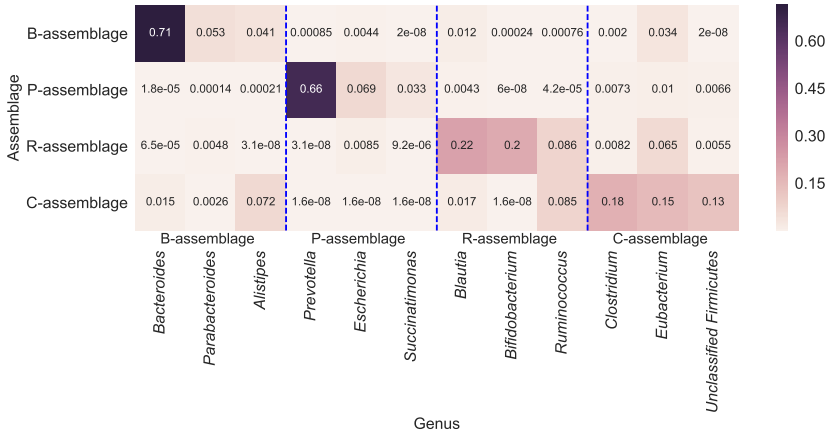

Figure 3: The estimated genus distribution of each microbial assemblage. The x- and y-axes represent genera and assemblages, respectively. We displayed only the three genera with the highest probability in each assemblage.

bioRxiv preprint doi: https://doi.org/10.1101/664219; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

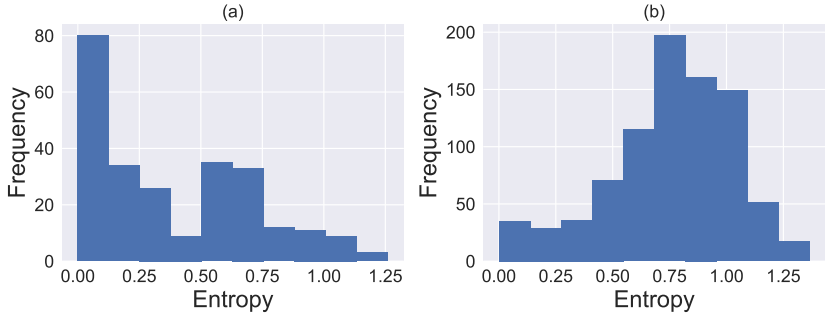

Figure 4: Histograms of the entropy scores of the microbial assemblage (a) for each genus (Eq. 2) and (b) for each individual (Eq. 2). The x- and y-axes represent entropy and the number of samples, respectively.

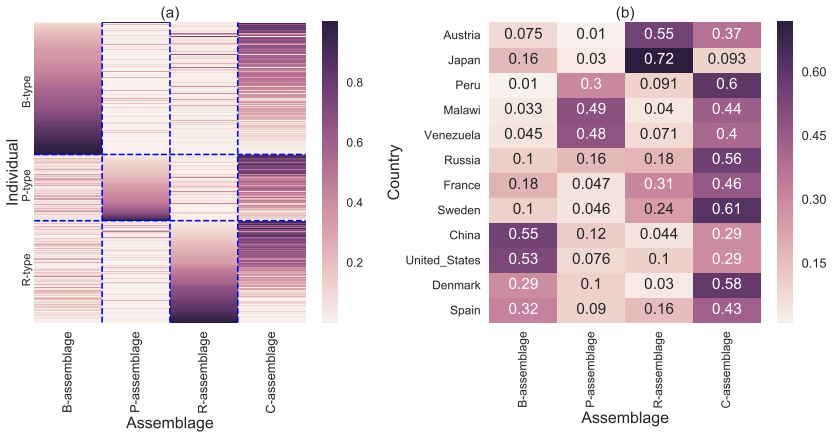

Figure 5: (a) Estimated microbial assemblage distribution for each individual. The x- and y-axes represent microbial assemblages and individual samples, respectively. Individuals are segregated by the enterotype and sorted by the B-assemblage, the P-assemblage, and the R-assemblage, respectively. (b) Average assemblage distributions for each country. Each row shows a distribution obtained by averaging the estimated assemblage distributions of individuals corresponding to each country. The x- and y-axis represent the microbial assemblage and the country of the individual, respectively.

bioRxiv preprint doi: https://doi.org/10.1101/664219; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

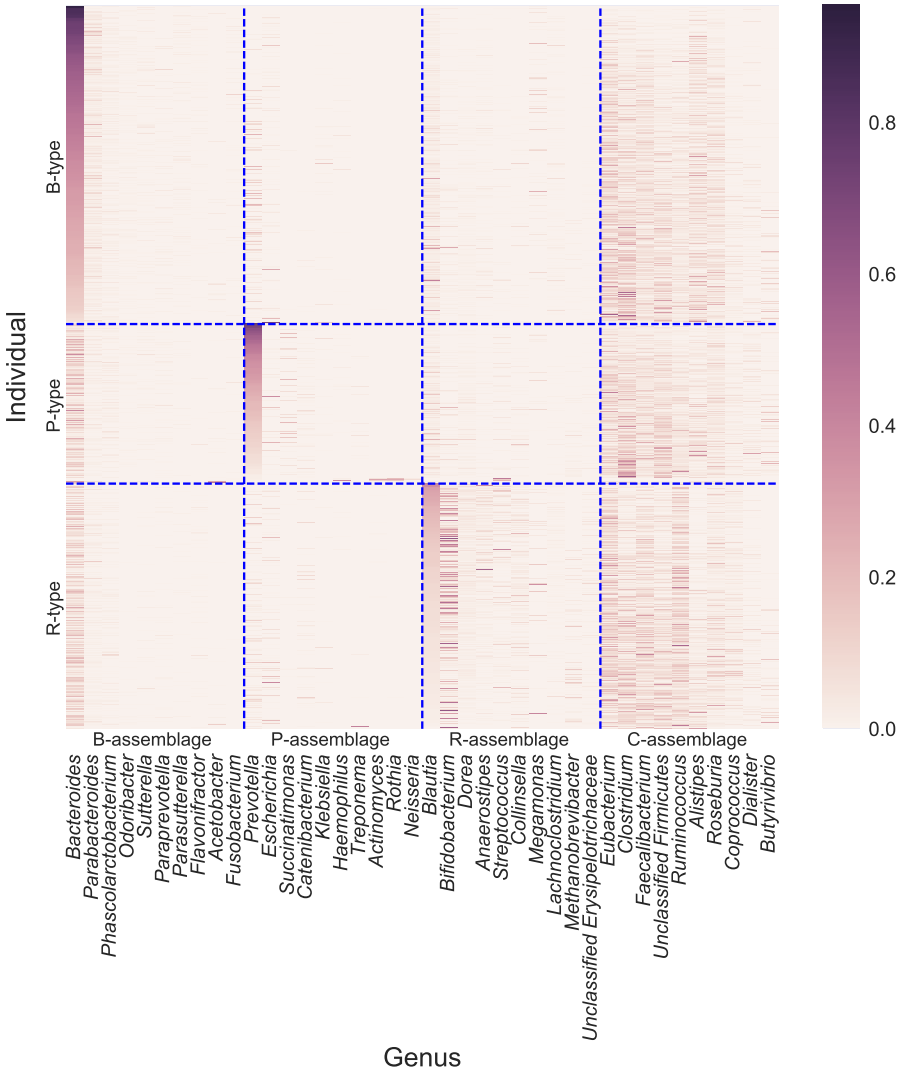

Figure 6: Relative abundance of genera in each individual. The x- and y-axes represent the genus and individuals, respectively. Individuals are divided by the enterotype and sorted by *Bacteroides*, *Prevotella*, and *Blautia*, respectively. Genera are divided by the assemblage that they mainly appear in and are sorted by the abundance of each genus. Each genus was regarded as mainly appearing in the assemblage of the highest $P(a_k|g_i)$, where a_k and g_i are the k-th assemblage and the i-th genus, respectively (Eq. 3 in the main text).

bioRxiv preprint doi: https://doi.org/10.1101/664219; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

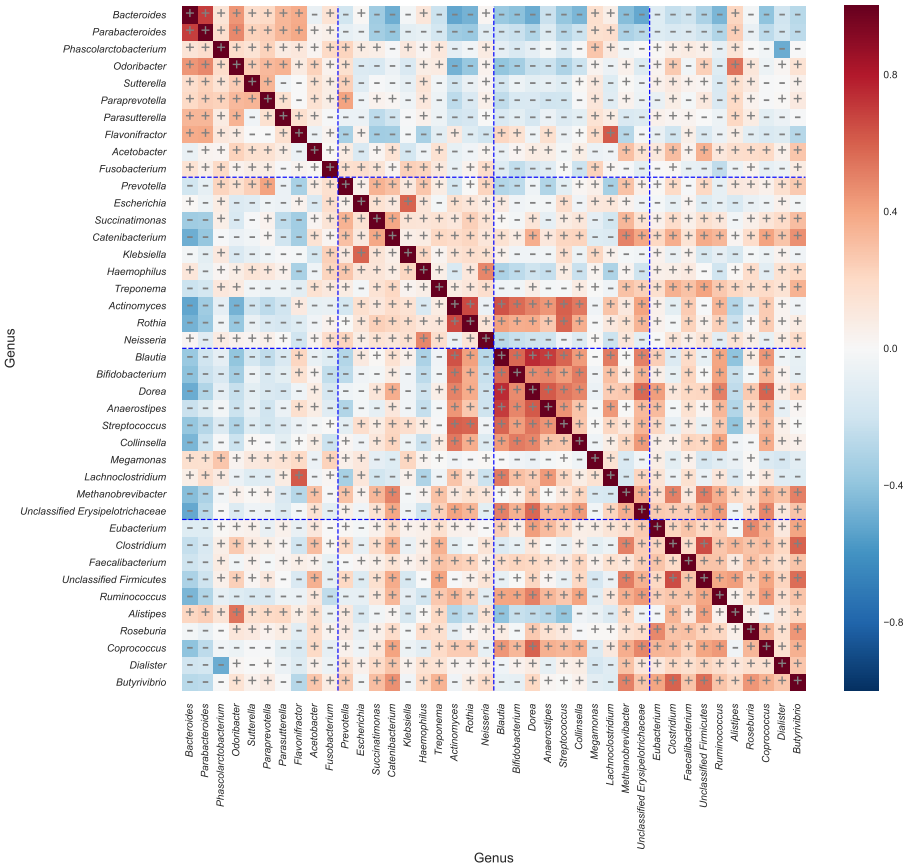

Figure 7: The Spearman's correlation coefficients among the 20 genera that are major in each enterotype. Both the x- and y-axes represent genera, which are divided and sorted in the same way as in Fig. 6. Plus and minus signs indicate significant positive and negative correlations, respectively. Significance was determined at $p < 0.01$ (two-sided test, after Benjamini–Hochberg correction).

bioRxiv preprint doi: https://doi.org/10.1101/664219; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

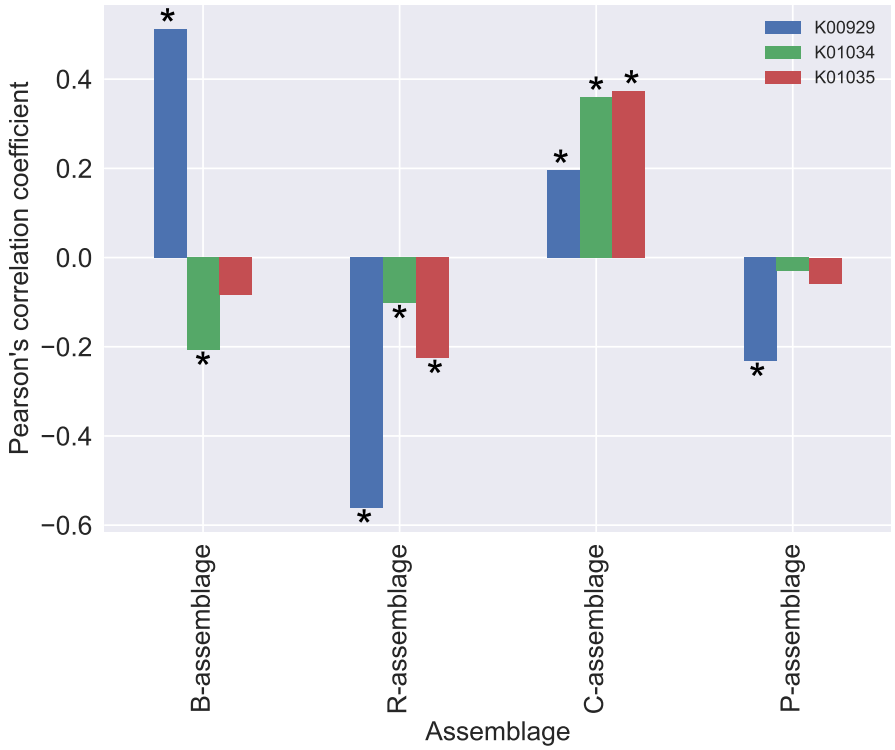

Figure 8: The Pearson's correlation coefficients among the 4 assemblages and 3 butyrate-producing functions. The x- and y-axes represent the assemblages and Pearson's correlation coefficients, respectively. Each bar of each assemblage indicates K01034, K00929, and K01035 from the left, respectively. Asterisks indicate significant differences. Significance was determined at $p < 0.01$ (two-sided test, after Benjamini–Hochberg correction).

bioRxiv preprint doi: https://doi.org/10.1101/664219; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Table 1: The assemblage having the largest relative abundance for each functional category

Functional category	**Functional assemblage**
Biosynthesis of Other Secondary Metabolites	B-assemblage(ko)
Carbohydrate Metabolism	B-assemblage(ko)
Digestive System	B-assemblage(ko)
Endocrine System	B-assemblage(ko)
Endocrine and Metabolic Diseases	B-assemblage(ko)
Environmental Adaptation	B-assemblage(ko)
Glycan Biosynthesis and Metabolism	B-assemblage(ko)
Lipid Metabolism	B-assemblage(ko)
Nervous System	B-assemblage(ko)
Transport and Catabolism	B-assemblage(ko)
Cell Growth and Death	P-assemblage(ko)
Folding, Sorting, and Degradation	P-assemblage(ko)
Infectious Diseases	P-assemblage(ko)
Metabolism of Terpenoids and Polyketides	P-assemblage(ko)
Nucleotide Metabolism	P-assemblage(ko)
Replication and Repair	P-assemblage(ko)
Membrane Transport	R-assemblage(ko)
Metabolism of Cofactors and Vitamins	R-assemblage(ko)
Metabolism of Other Amino Acids	R-assemblage(ko)
Transcription	R-assemblage(ko)
Xenobiotic Biodegradation and Metabolism	R-assemblage(ko)
Immune Diseases	R-assemblage(ko)
Energy Metabolism	R-assemblage(ko)
Amino Acid Metabolism	R-assemblage(ko)
Cancers	C-assemblage(ko)
Cell Motility	C-assemblage(ko)
Immune System	C-assemblage(ko)
Neurodegenerative Diseases	C-assemblage(ko)
Signal Transduction	C-assemblage(ko)
Translation	C-assemblage(ko)

bioRxiv preprint doi: https://doi.org/10.1101/664219; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Table 2: Dominant genera of the C-assemblage and the probability of the C-assemblage as estimated by LDA.

Genus	**Probability**
Clostridium	0.179865
Eubacterium	0.150802
Unclassified Firmicutes	0.129783
Faecalibacterium	0.093720
Ruminococcus	0.085272
Roseburia	0.074214
Alistipes	0.072359
Coprococcus	0.029497
Butyrivibrio	0.021738

bioRxiv preprint doi: https://doi.org/10.1101/664219; this version posted June 12, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Additional Files

Additional File 1 — supplementary.pdf

This file includes Figures S1, S2, S3, S4, S5, S6, S7, S8, and Table S1.