

# Supplementary Materials for “Revealing the microbial assemblage structure in the human gut microbiome using latent Dirichlet allocation”

Shion Hosoda, Suguru Nishijima, Tsukasa Fukunaga, Masahira Hattori  
and Michiaki Hamada\*

## S1 Functional assemblage analysis

We investigated the human population-level differences in microbial assemblages and their compositions. We also estimated the functions of each assemblage by applying LDA to the functional profiles of identical samples as the genus data. These assemblages are referred to as “functional assemblages” in later analyses.

### S1.1 Functional profile dataset and preprocessing methods

We normalized the functional profiles in the same manner as for the taxonomic profiles in the main text. We confirmed that the estimated parameters do not depend on the constants (Figure A1). After these preprocessing steps, the number of KEGG orthology types included in the dataset was 4,840.

### S1.2 LDA for modeling the human gut functional profiles

The inference method for functional assemblages is almost the same as that for microbial assemblages (the details are described in the “Materials and methods” section in the main text), with differences in the following two points. First, we used the same number of functional assemblages with the microbial assemblages, *i.e.*  $K = 4$ . Second, we calculated the correlation score (the details are described in the next paragraph) for each trial and adopted the trial with the highest correlation score (see the next paragraph) among all 10 trials.

We defined the correlation score to quantify appropriateness of treating estimated functional assemblages as the relative abundance of functions of microbial assemblages. The correlation score is defined by

$$\max_p \sum_{k=1}^4 C_{k,p_k}, \quad (1)$$

where  $C_{k,k'}$  is a Pearson’s correlation coefficient between the  $k$ -th microbial assemblage and  $k'$ -th functional assemblage with respect to samples,  $p$  is a permutation of functional assemblage indices  $\{1, 2, 3, 4\}$ , and  $p_i$  is the  $i$ -th element of the permutation,  $p$  (that is, 1, 2, 3, or 4).

As a result of 10 trials, the third trial, which had the highest correlation score, was adopted (Figure A2). We named the functional assemblages based on the corresponding microbial assemblage with (ko) appended (that is, “B-assemblage(ko),” “P-assemblage(ko),” “R-assemblage(ko),” and “C-assemblage(ko)”).

---

\*To whom correspondence should be addressed. Department of Electrical Engineering and Bioscience Faculty of Science and Engineering, Waseda University 55N-06-10, 3-4-1, Okubo Shinjuku-ku, Tokyo 169-8555, Japan. Tel: +81 3 5286 3130; Fax: +81 3 5286 3130; Email: mhamada@waseda.jp

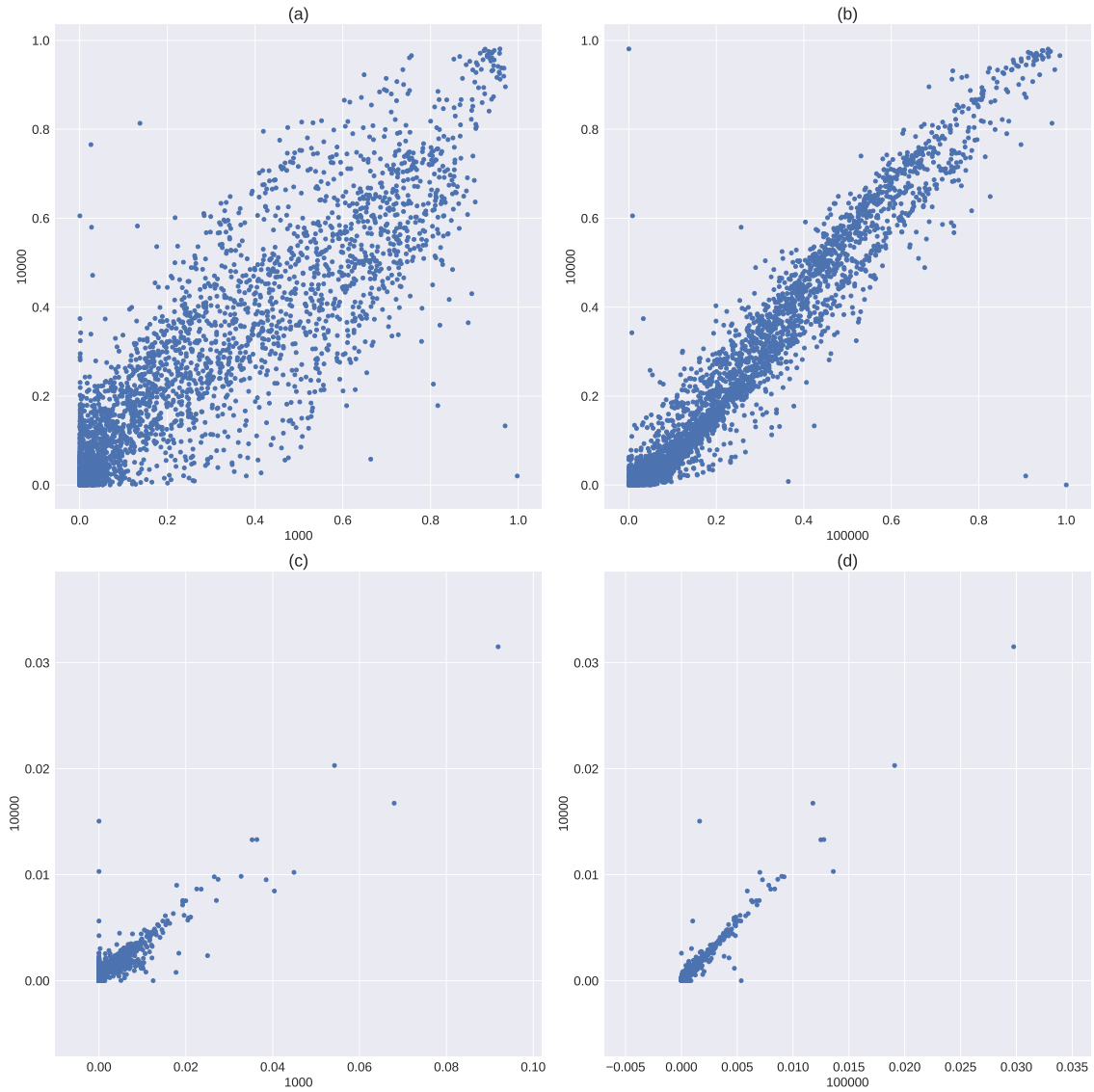


Figure A1: Scatter charts between estimated parameters of the constants (described in the main text) of 10,000 and 1,000/100,000. (a)(b) Parameters of the distributions of samples over functional assemblages. (a) 10,000 versus 1,000. (b) 10,000 versus 100,000. (c)(d) Parameters of the distributions of functional assemblage over functions. (c) 10,000 versus 1,000. (d) 10,000 versus 100,000. The  $x$ - and  $y$ -axes represent the estimated values of the parameters in the case of 1,000/100,000 and 10,000, respectively.

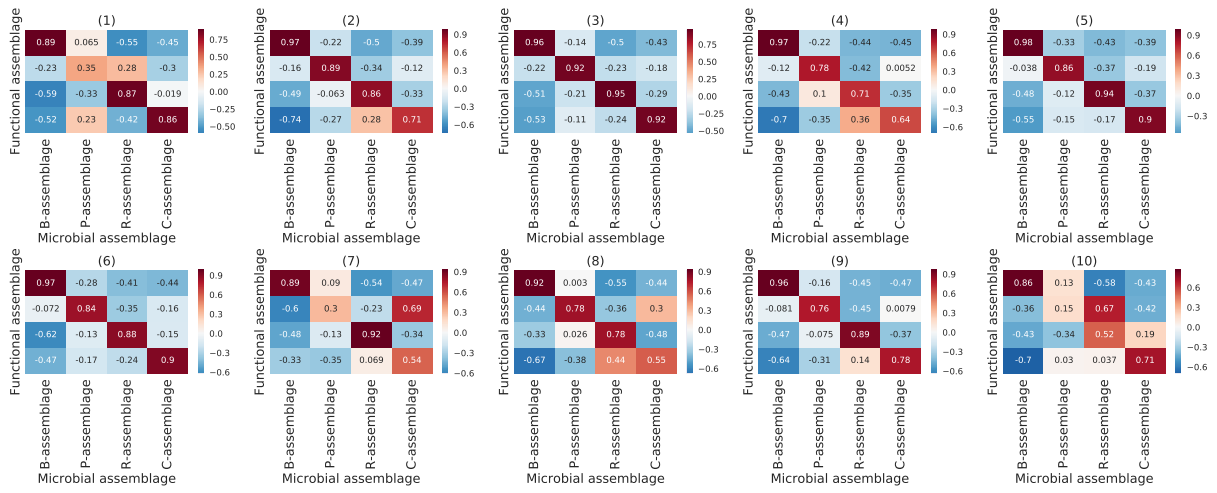


Figure A2: Pearson's correlation coefficients among microbial assemblages and functional assemblages for each trial ( $C$  in section S1). Both the  $x$ - and  $y$ -axes represent assemblages.

## S2 Supplementary Figures and Table

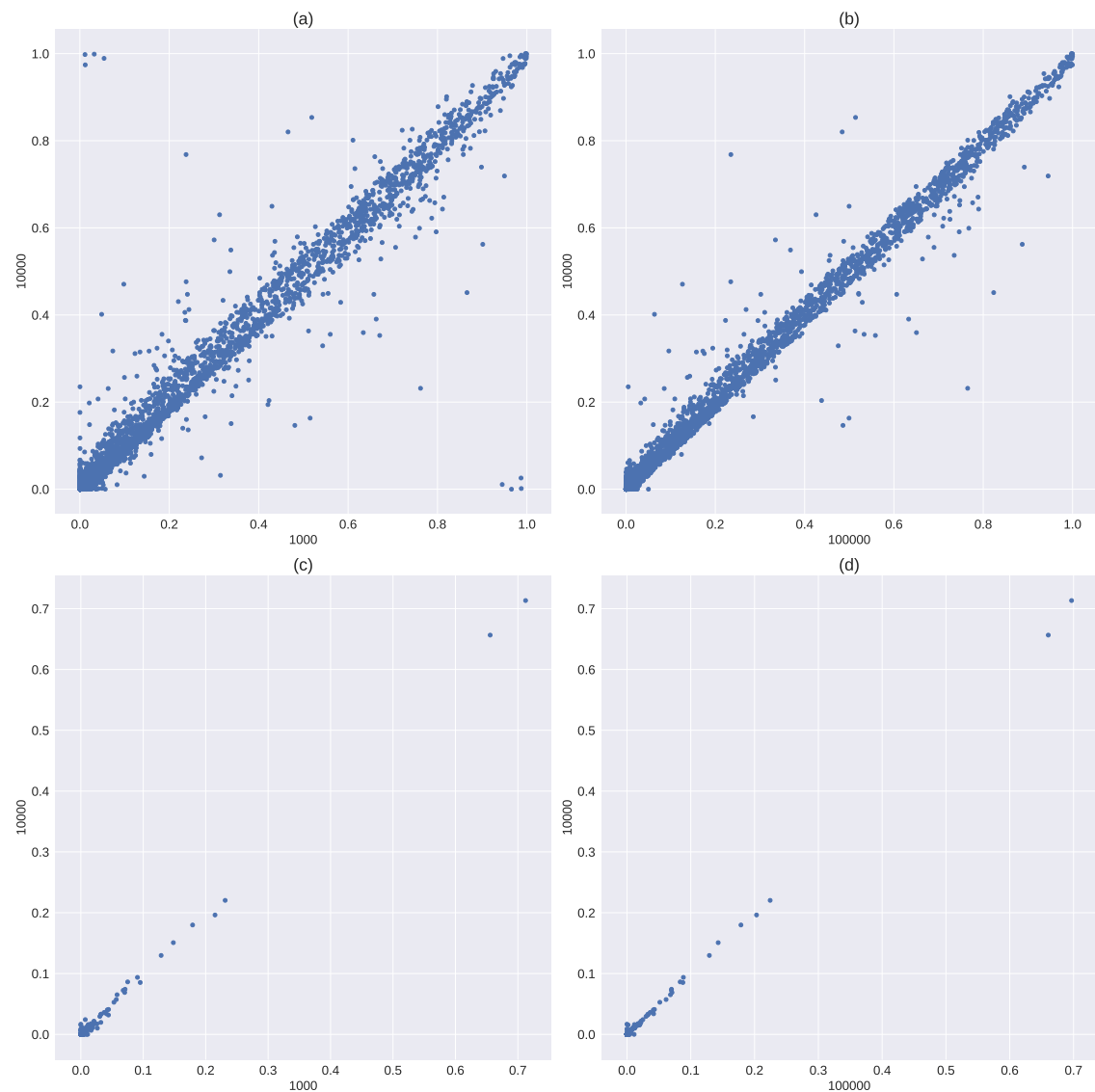


Figure S1: Scatter charts for estimated parameters with respect to different constants. (a,b) Comparison of estimated parameters of distributions of samples over microbial assemblages (indicated as  $\theta_{i,k}$  in the main text): (a) 10,000 versus 1,000 and (b) 10,000 versus 100,000. (c,d) Comparison of estimated parameters of distributions of of microbial assemblages over genera (indicated as  $\phi_{k,j}$  in the main text): (c) 10,000 versus 1,000 and (d) 10,000 versus 100,000. The  $x$ - and  $y$ -axes represent the estimated values of parameters in the cases of 1,000/10,000 and 10,000, respectively.

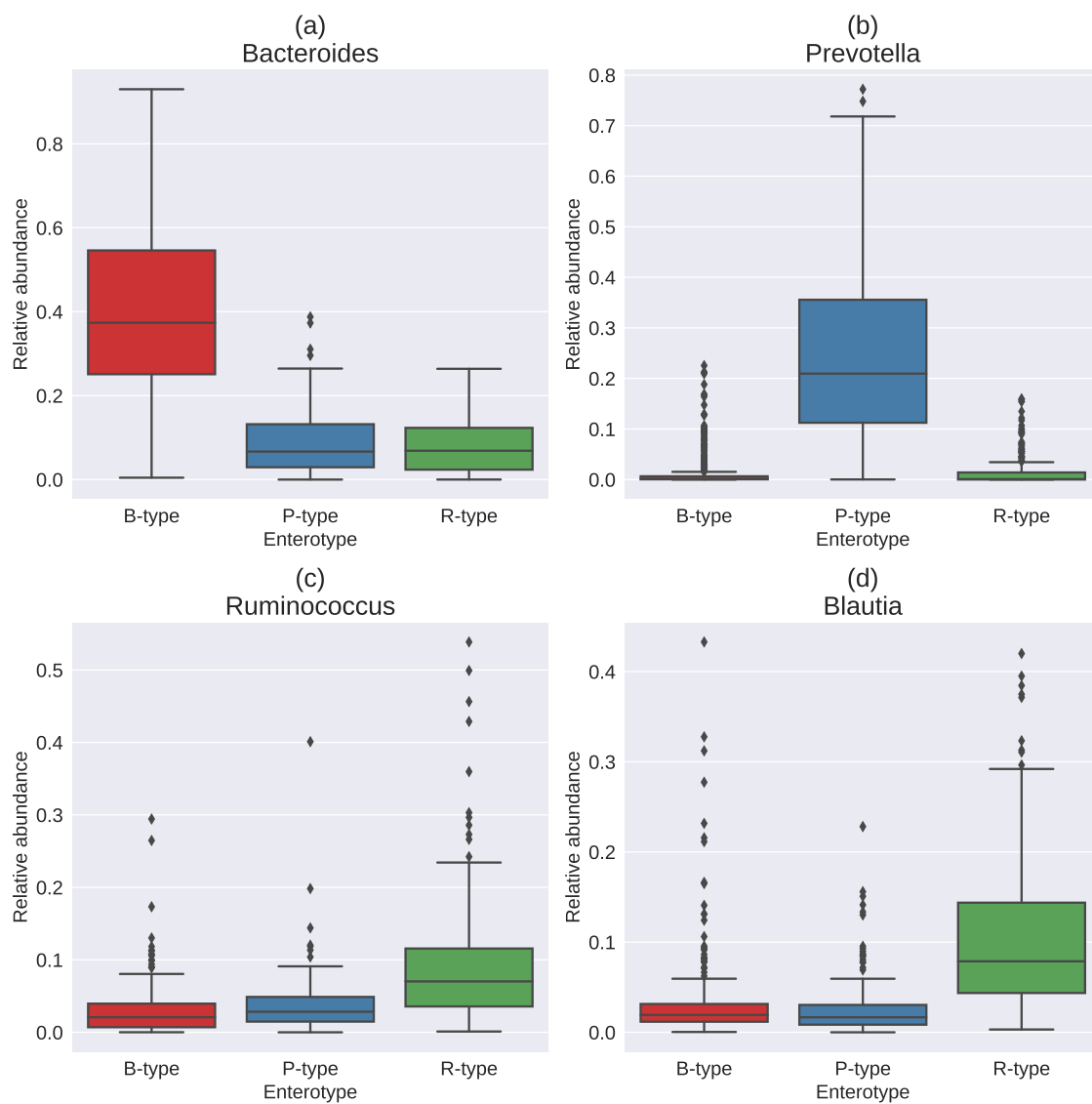


Figure S2: The relative abundance of four major genera ((a) *Bacteroides*, (b) *Prevotella*, (c) *Ruminococcus*, (d) *Blautia*) for each enterotype. The  $x$ - and  $y$ -axes represent the enterotype and relative abundance of the genus, respectively.

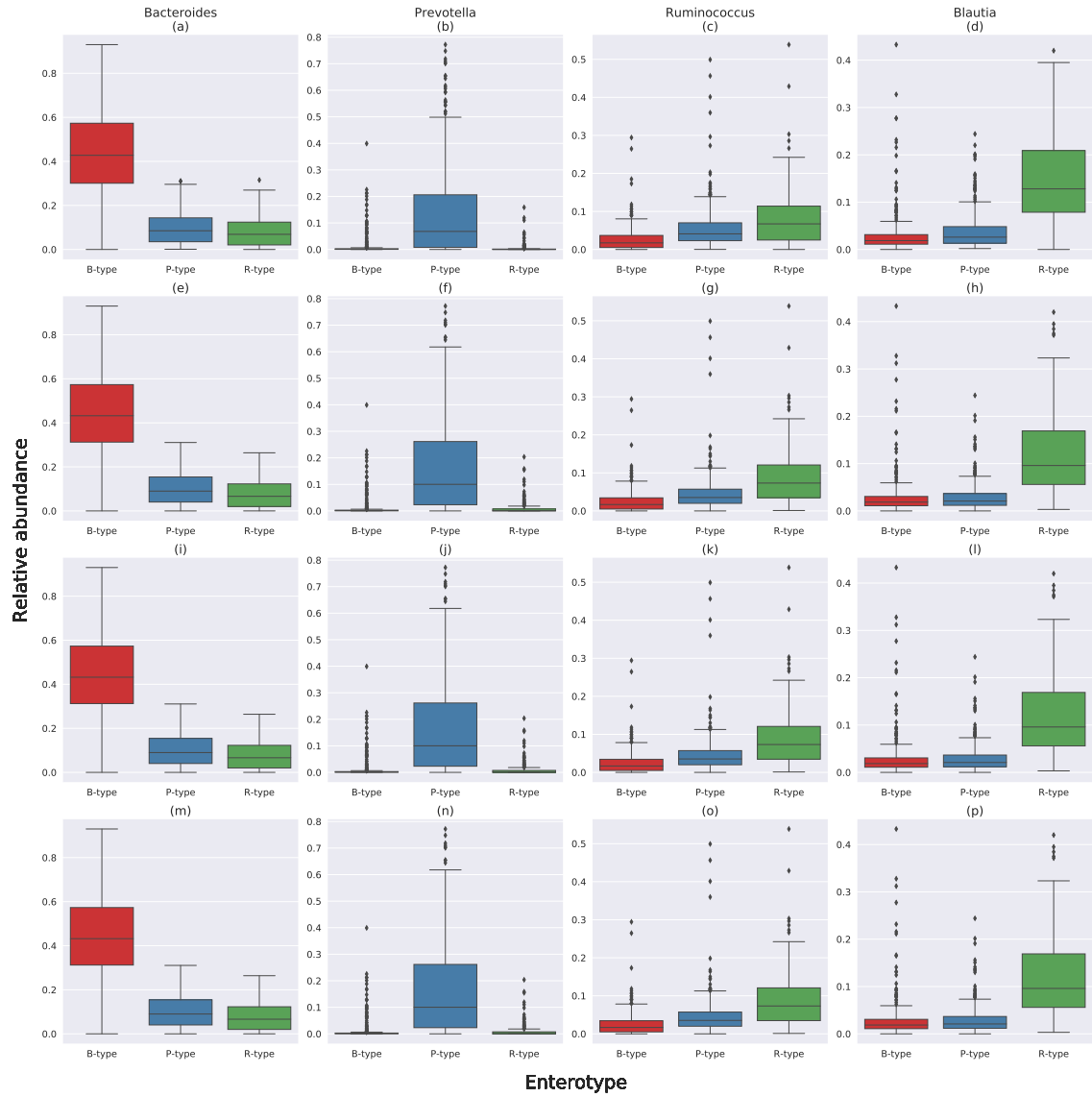


Figure S3: Relative abundance of four major genera ( (a, e, i, m)*Bacteroides*, (b, f, j, n)*Prevotella*, (c, g, k, o)*Ruminococcus*, (d, h, l, p)*Blautia*) for each enterotype obtained by four trials with the (a-d) second, (e-h) third, (i-l) fourth, and (m-p) fifth highest silhouette coefficients. The  $x$ - and  $y$ -axes represent enterotype and relative abundance of the genus, respectively.

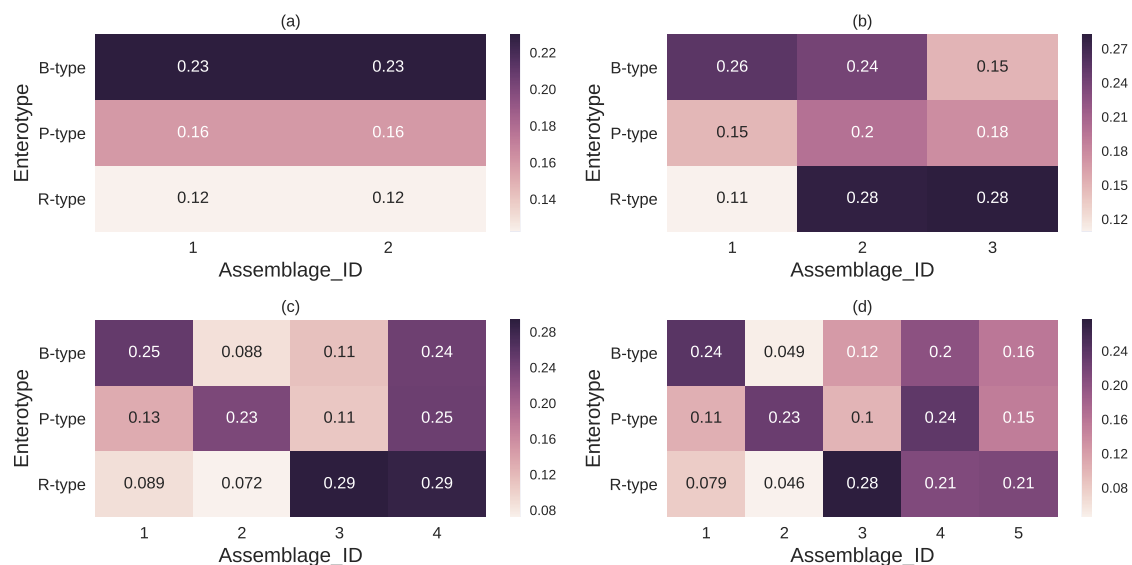


Figure S4: Standard deviations of the assemblage distributions for each enterotype. The  $x$ - and  $y$ -axes represent the microbial assemblages and enterotypes, respectively. Darker colors indicate higher standard deviations, and each number inside the partition indicates a different standard deviation. (a), (b), (c), and (d) indicate  $K$ -assemblage LDA models with  $K = 2, 3, 4$ , and  $5$ , respectively.

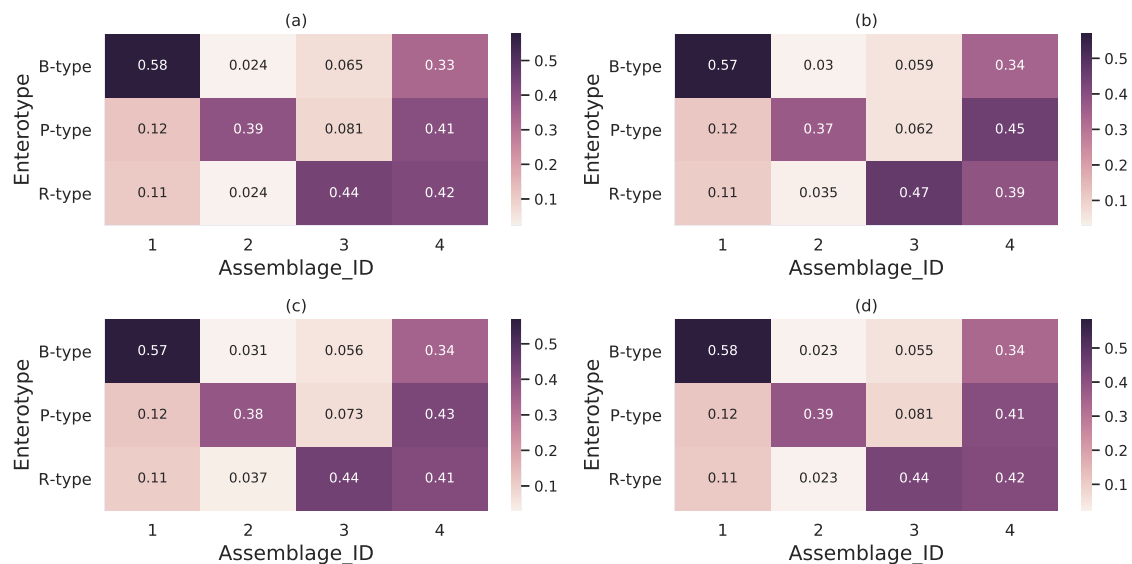


Figure S5: Assemblage distributions of the other four trials with the second, third, fourth, and fifth highest VLB scores for each enterotype. Each row shows a distribution obtained by averaging the estimated assemblage distributions across individuals corresponding to each enterotype. The  $x$ - and  $y$ -axes represent the microbial assemblages and enterotypes, respectively. Darker colors indicate higher probabilities, and each number inside the partition indicates a different probability, where the sum of the values over each row is 1. (a), (b), (c), and (d) indicate the trials with the second, third, fourth, and fifth highest VLB scores, respectively.

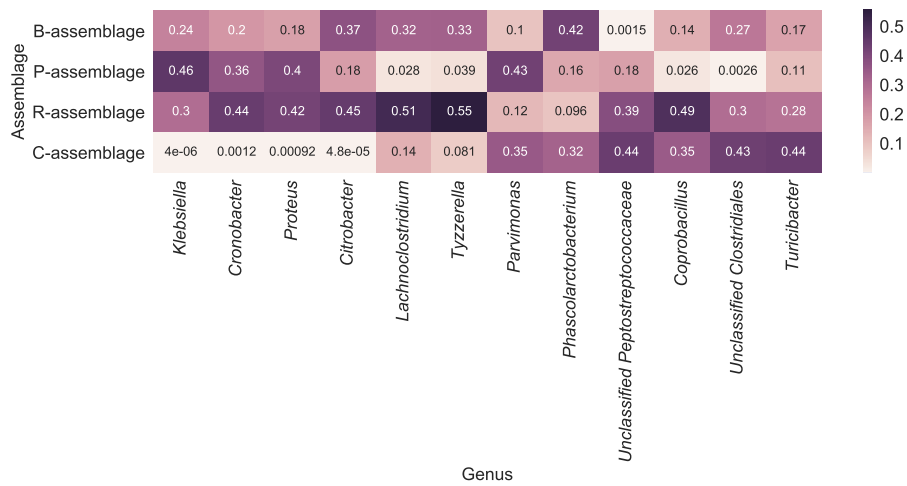


Figure S6: The  $P(k|j)$  (Eq. 3 in the main text) of the genera with high entropy scores ( $> 1.0$ ). The  $x$ - and  $y$ -axes represent the genus and microbial assemblage, respectively.

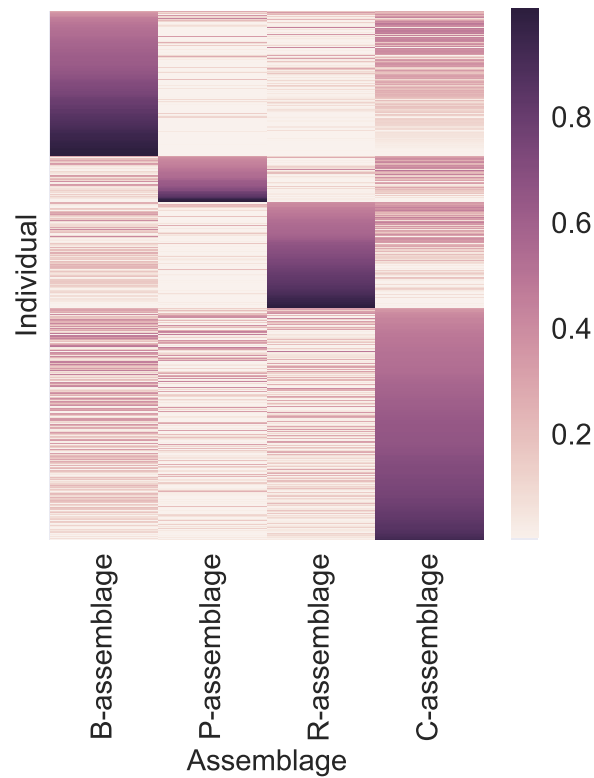


Figure S7: Estimated microbial assemblage distribution for each individual. The  $x$ - and  $y$ -axes represent microbial assemblages and individual samples, respectively. Individuals are ordered by the probability of the dominant assemblage.



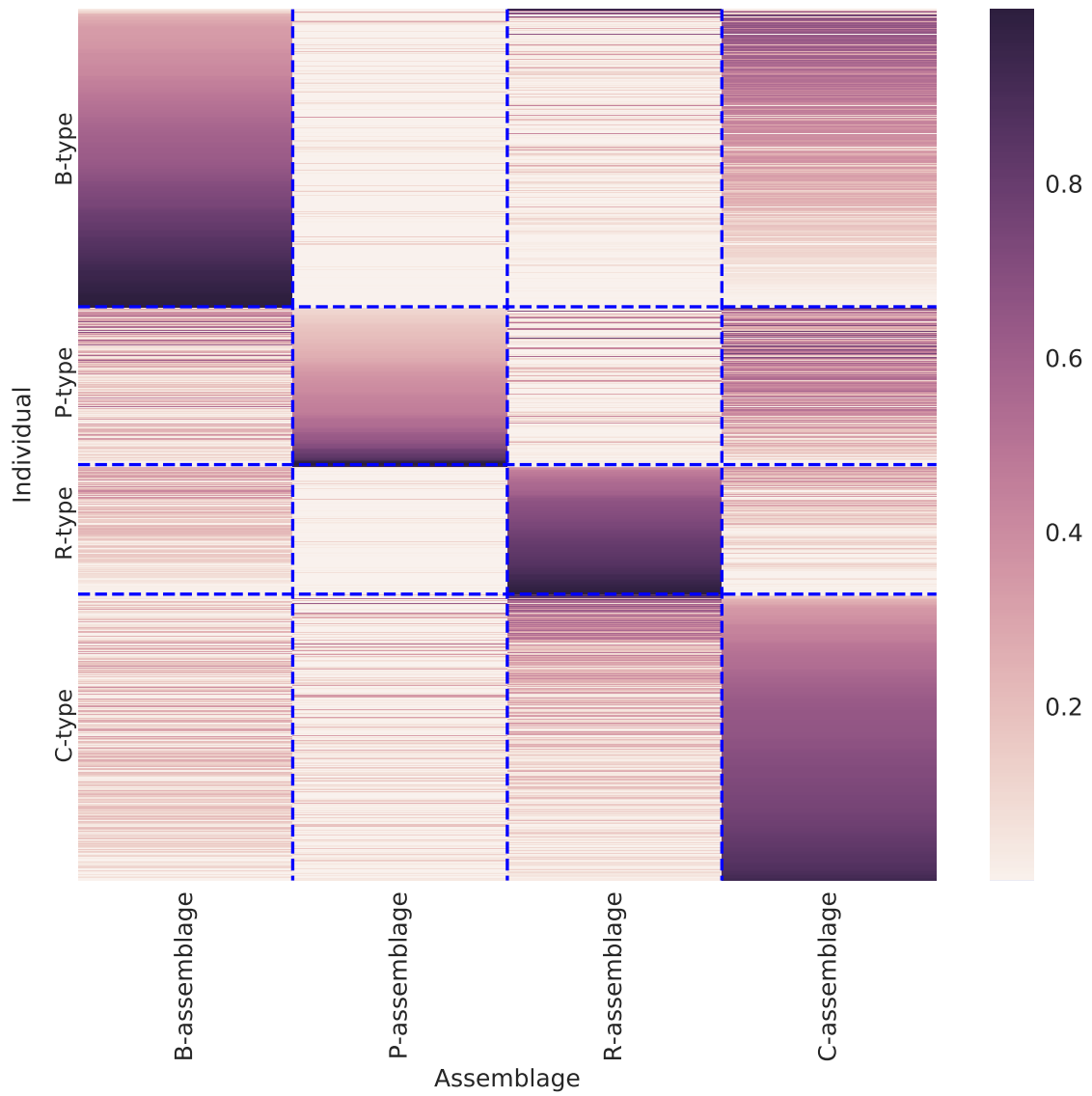


Figure S8: Estimated microbial assemblage distribution for each individual. The  $x$ - and  $y$ -axes represent microbial assemblages and individual samples, respectively. Individuals are segregated by the types obtained by the 4-type PAM method and sorted by the B-assemblage, P-assemblage, R-assemblage, and C-assemblage, respectively. The fourth type is referred to as the “C-type.”

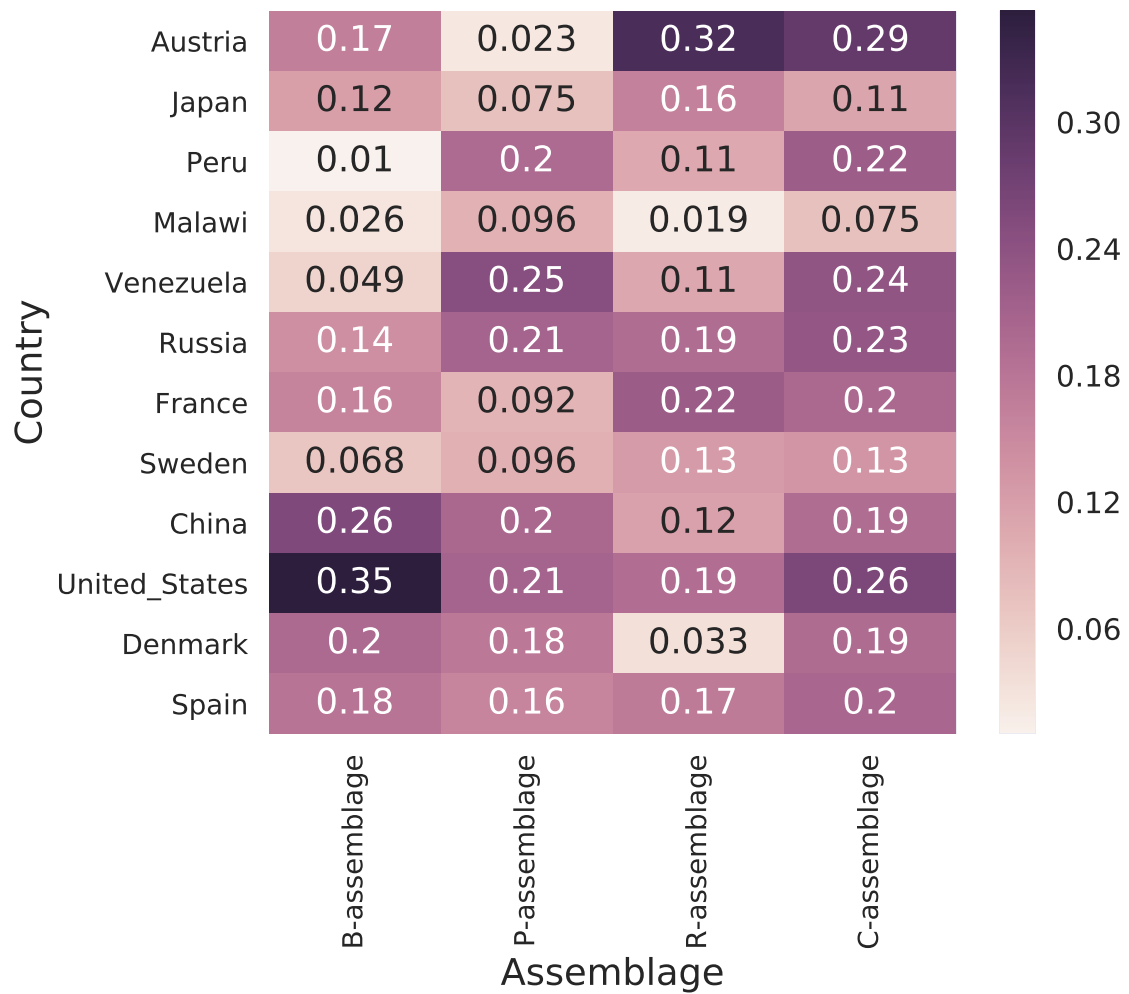


Figure S9: Standard deviations of the assemblage distributions for each country. The  $x$ - and  $y$ -axes represent the microbial assemblages and countries, respectively. Darker colors indicate higher standard deviations, and each number inside the partition indicates a different standard deviation. (a), (b), (c), and (d) indicate  $K$ -assemblage LDA models with  $K = 2, 3, 4,$  and  $5,$  respectively.

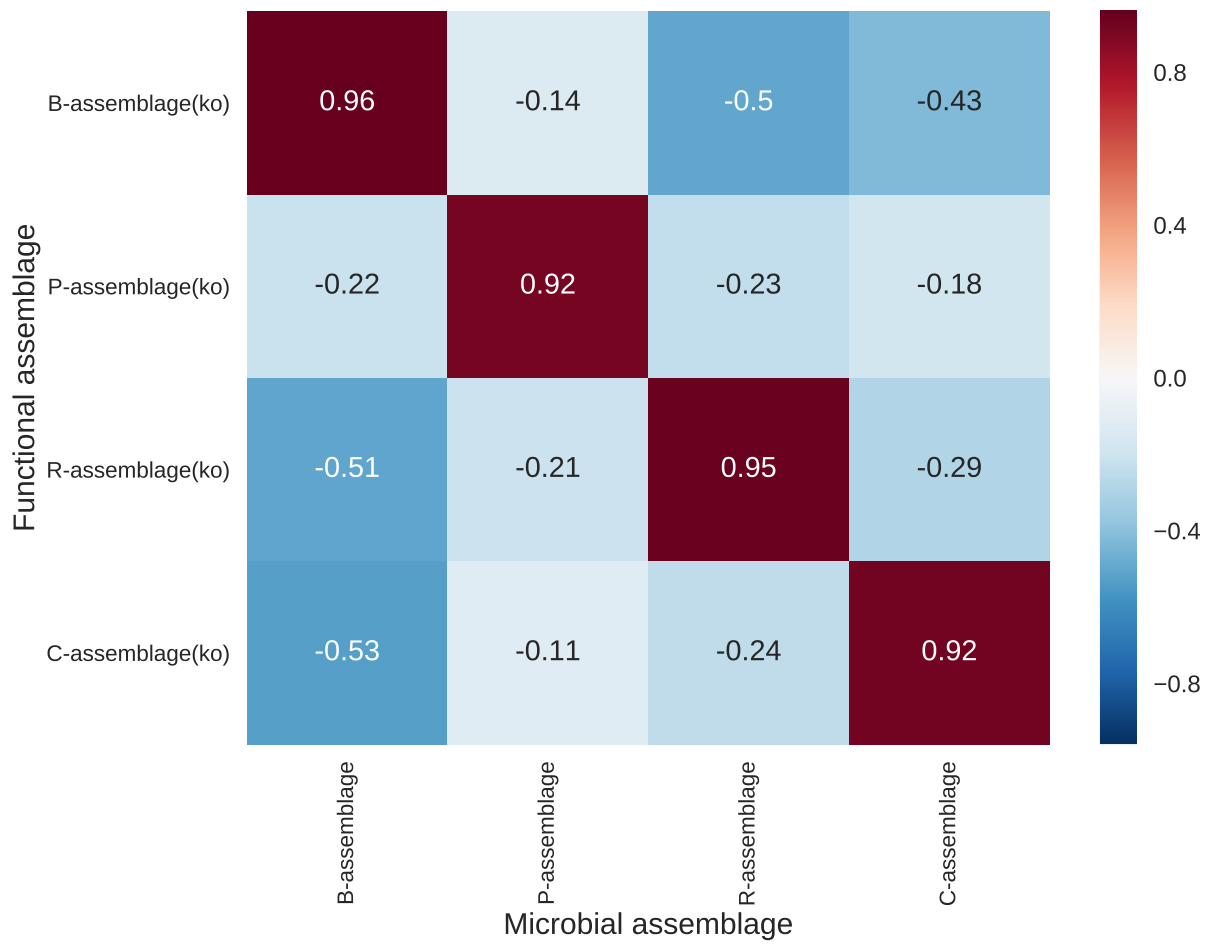


Figure S10: Pearson's correlation coefficients among microbial and functional assemblages (the same figure as Fig. A2(3)). Both the  $x$ - and  $y$ -axes represent assemblages.

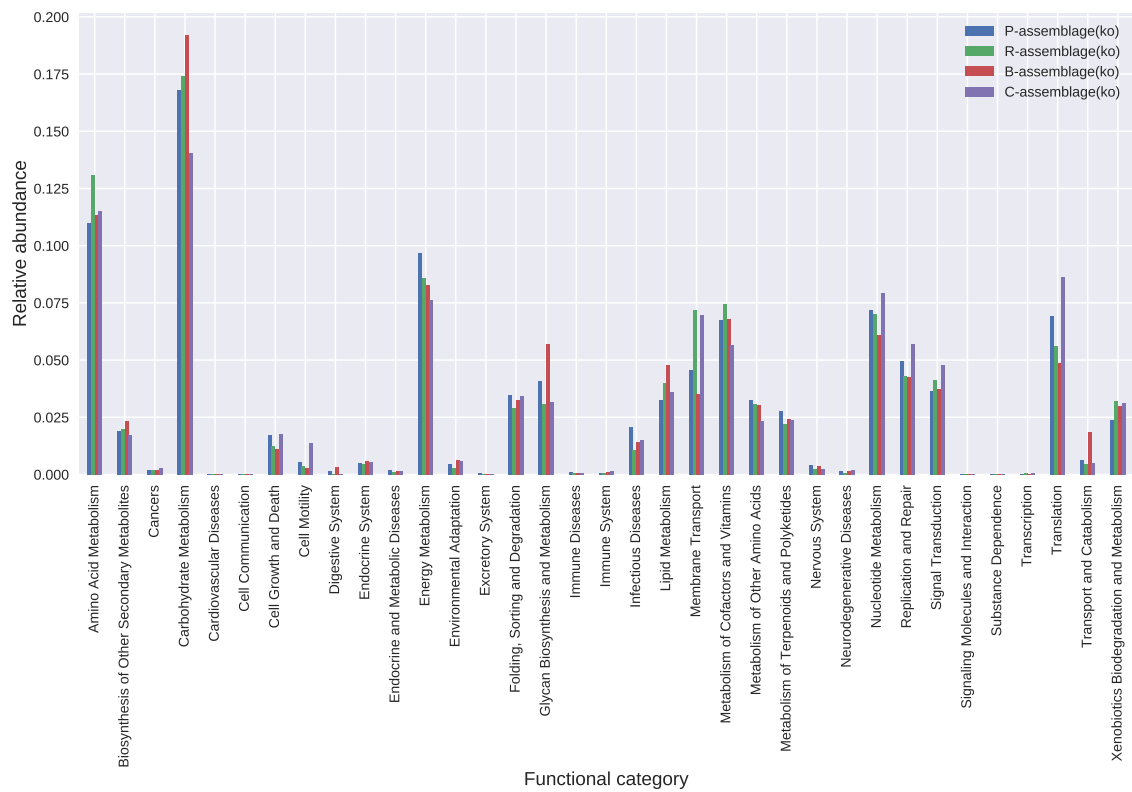


Figure S11: Relative abundance of functional categories for each assemblage. The  $x$ - and  $y$ -axes represent the function category and relative abundance, respectively.

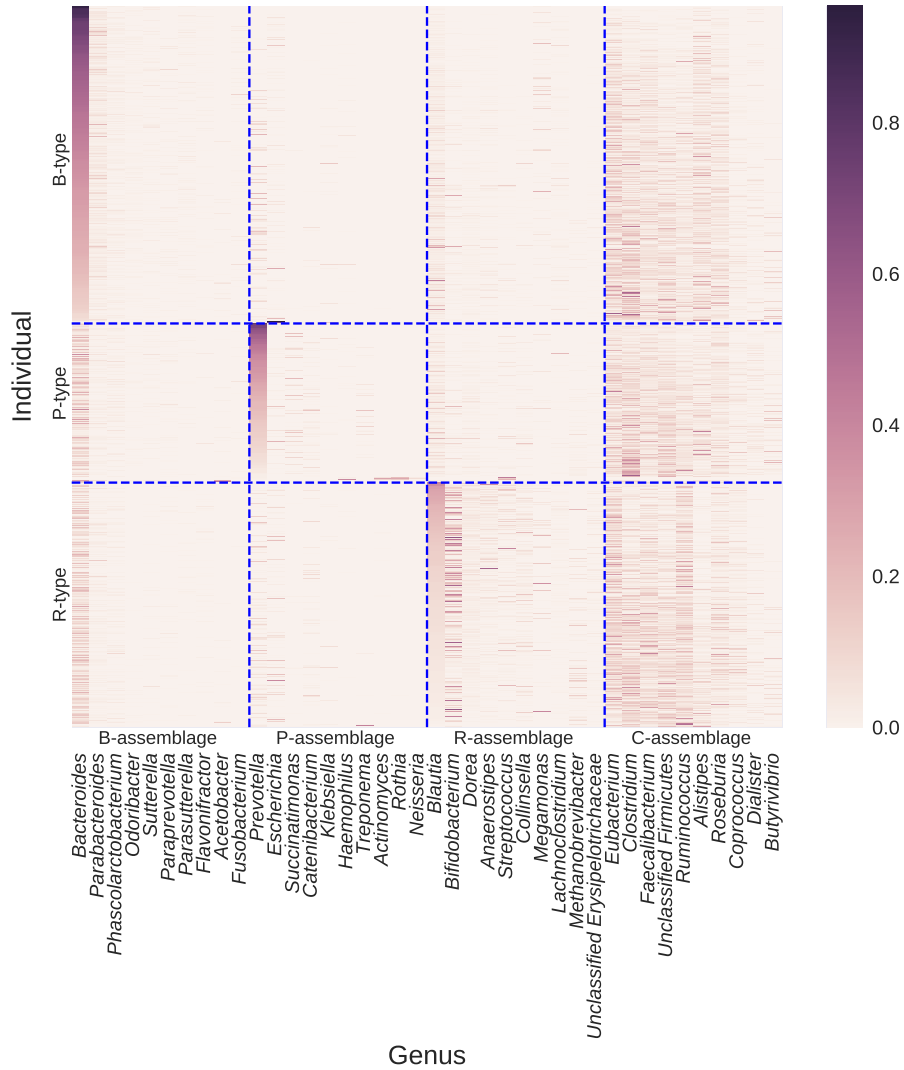


Figure S12: Relative abundance of genera in each individual. The  $x$ - and  $y$ -axes represent the genus and individuals, respectively. Individuals are divided by the enterotype and sorted by *Bacteroides*, *Prevotella*, and *Blautia*. Genera are divided by the assemblage in which they mainly appear and are sorted by the abundance of each genus. Each genus was regarded as mainly appearing in the assemblage of the highest  $P(k|j)$ , which is a probability of the  $k$ -th assemblage given the  $j$ -th genus (Eq. 3 in the main text).

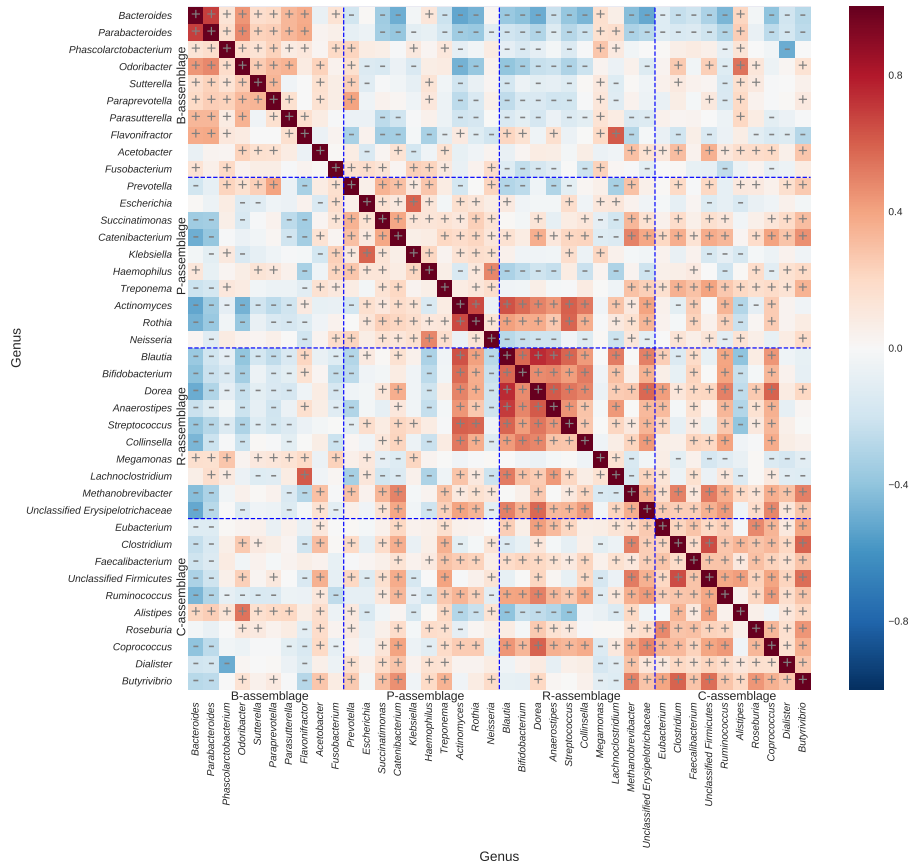


Figure S13: The Spearman's correlation coefficients among the 20 genera that are predominant in each enterotype. Both the  $x$ - and  $y$ -axes represent genera, which are divided and sorted similarly as in Fig. S12. Plus and minus signs indicate significant positive and negative correlations, respectively. Significance was determined at  $p < 0.01$  (two-sided test, after Benjamini–Hochberg correction).

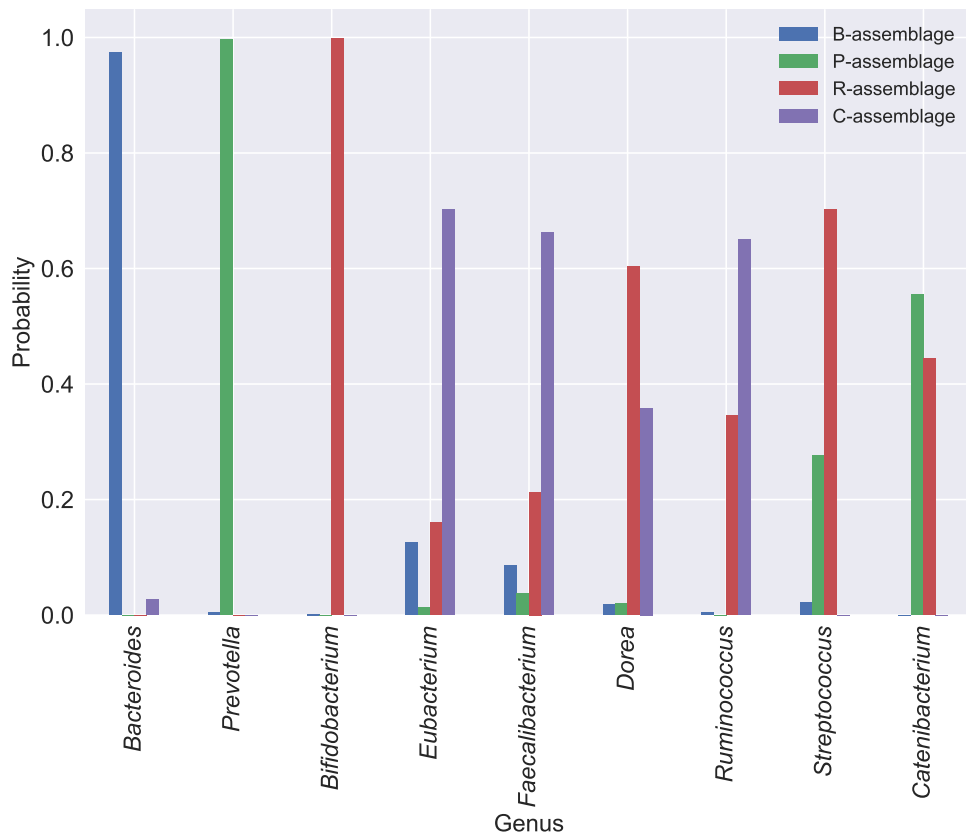


Figure S14:  $P(k|j)$  of *Bacteroides*, *Prevotella*, *Bifidobacterium*, and the genera that have high Spearman's correlation coefficients with the other assemblage genera in Fig. S13. The  $x$ - and  $y$ -axes represent the genus and  $P(k|j)$ , respectively.

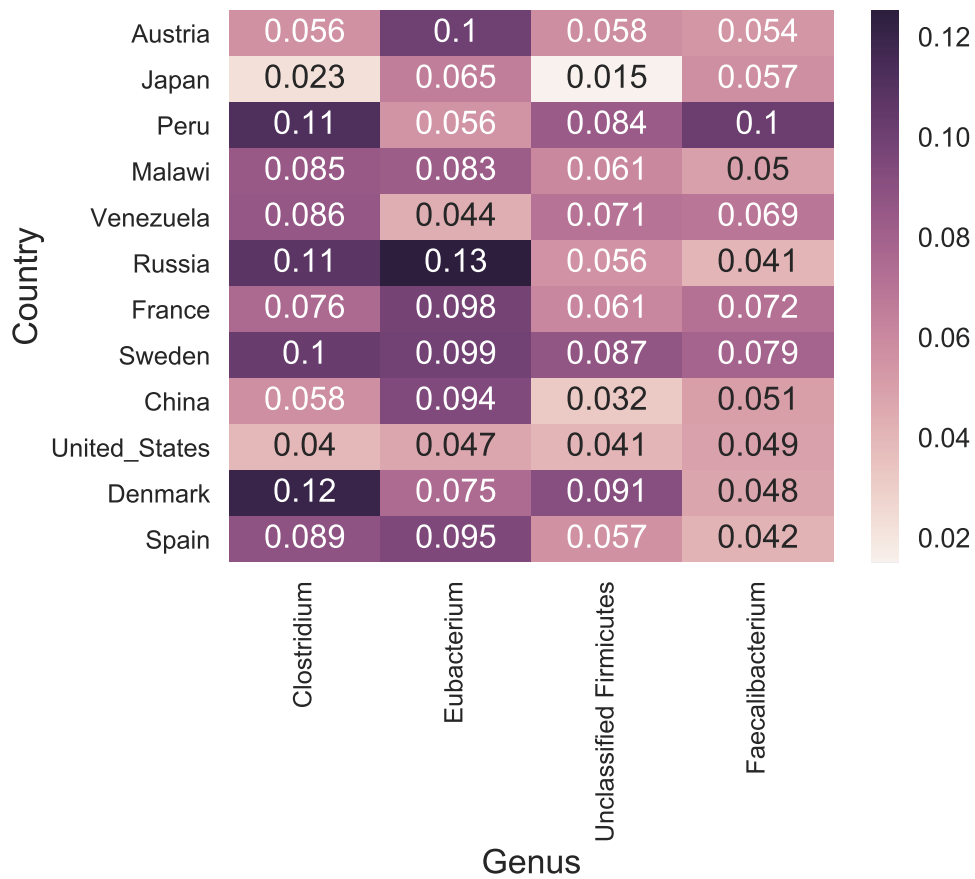


Figure S15: Average relative abundance of *Clostridium*, *Eubacterium*, unclassified *Firmicutes*, and *Faecalibacterium* for each country. The  $x$ - and  $y$ -axes represent the genus and country of the individual, respectively.

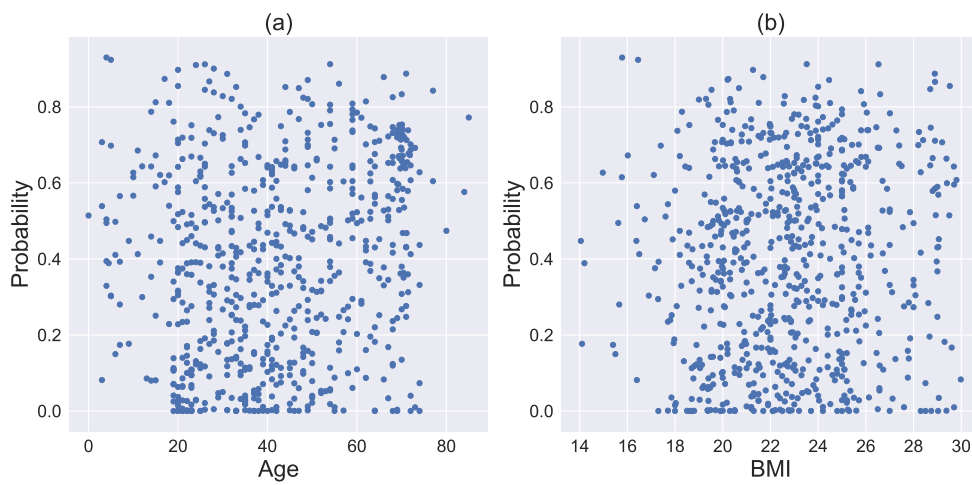


Figure S16: Relationship between C-assembly probability and (a) age and (b) BMI metadata. The  $x$ - and  $y$ -axes represent the metadata and C-assembly probability, respectively.



Table S1: The entropy score of  $P(k|j)$  for all genera used.

Genus	entropy
<i>Turicibacter</i>	1.260681
<i>Phascolarctobacterium</i>	1.245466
<i>Parvimonas</i>	1.220488
<i>Unclassified Clostridiales</i>	1.092140
<i>Coprobacillus</i>	1.088414
<i>Lachnoclostridium</i>	1.082888
<i>Klebsiella</i>	1.061892
<i>Cronobacter</i>	1.056325
<i>Unclassified Peptostreptococaceae</i>	1.044608
<i>Proteus</i>	1.042833
<i>Citrobacter</i>	1.037434
<i>Tyzzereella</i>	1.021686
<i>Veillonella</i>	0.999699
<i>Holdemania</i>	0.983926
<i>Yersinia</i>	0.980842
<i>Bilophila</i>	0.951916
<i>Faecalibacterium</i>	0.937553
<i>Unclassified Lachnospiraceae</i>	0.933686
<i>Paraprevotella</i>	0.926816
<i>Enterobacter</i>	0.924527
<i>Unclassified Bacteria</i>	0.906336
<i>Subdoligranulum</i>	0.893402
<i>Pseudomonas</i>	0.891144
<i>Peptoclostridium</i>	0.880690
<i>Flavonifractor</i>	0.878429
<i>Eubacterium</i>	0.857729
<i>Lachnospira</i>	0.840097
<i>Dorea</i>	0.822750
<i>Unclassified Erysipelotrichaceae</i>	0.816711
<i>Escherichia</i>	0.814104
<i>Megasphaera</i>	0.812753
<i>Roseburia</i>	0.788690
<i>Erysipelatoclostridium</i>	0.784843
<i>Acidaminococcus</i>	0.773040
<i>Dielma</i>	0.756564
<i>Intestinibacter</i>	0.744818
<i>Mycobacterium</i>	0.736164
<i>Sutterella</i>	0.725791
<i>Scardovia</i>	0.718660
<i>Olsenella</i>	0.713303
<i>Providencia</i>	0.709675
<i>Stomatobaculum</i>	0.709085
<i>Odoribacter</i>	0.704806
<i>Unclassified Clostridiales Family XIII. Incerta...</i>	0.703168
<i>Peptostreptococcus</i>	0.697437
<i>Kandleria</i>	0.696656
<i>Marvinbryantia</i>	0.695373
<i>Cloacibacillus</i>	0.695198
<i>Sharpea</i>	0.695112
<i>Atopobium</i>	0.694374
<i>Barnesiella</i>	0.693038
<i>Catenibacterium</i>	0.687024
<i>Tannerella</i>	0.686389
<i>Streptococcus</i>	0.685115
<i>Oenococcus</i>	0.684101
<i>Unclassified Bacteroidales</i>	0.683936

Table S1: The entropy score of  $P(k|j)$  for all genera used. – *Continued from previous page*

Genus	entropy
<i>Dialister</i>	0.683460
<i>Azospirillum</i>	0.680177
<i>Weissella</i>	0.677471
<i>Ruminococcus</i>	0.673173
<i>Megamonas</i>	0.669763
<i>Leuconostoc</i>	0.668794
<i>Unclassified Burkholderiales</i>	0.665163
<i>Methanosphaera</i>	0.654763
<i>Blautia</i>	0.646975
<i>Solobacterium</i>	0.644240
<i>Enterococcus</i>	0.644030
<i>Akkermansia</i>	0.631433
<i>Alistipes</i>	0.624308
<i>Actinomyces</i>	0.622316
<i>Unclassified Ruminococcaceae</i>	0.620387
<i>Peptoniphilus</i>	0.616613
<i>Senegalimassilia</i>	0.595117
<i>Pseudoflavonifractor</i>	0.594075
<i>Unclassified Proteobacteria</i>	0.591897
<i>Haemophilus</i>	0.588196
<i>Desulfovibrio</i>	0.585395
<i>Alloprevotella</i>	0.582315
<i>Actinobaculum</i>	0.579283
<i>Propionibacterium</i>	0.578085
<i>Oxalobacter</i>	0.576018
<i>Mitsuokella</i>	0.571258
<i>Coprococcus</i>	0.562809
<i>Fusobacterium</i>	0.561115
<i>Adlercreutzia</i>	0.546762
<i>Butyricicoccus</i>	0.543292
<i>Anaerostipes</i>	0.539088
<i>Thauera</i>	0.535747
<i>Pedobacter</i>	0.535627
<i>Luteimonas</i>	0.535625
<i>Allobaculum</i>	0.535624
<i>Sphaerochaeta</i>	0.535624
<i>Kallipyga</i>	0.535624
<i>Lamprocystis</i>	0.535565
<i>Sphingobacterium</i>	0.535565
<i>Sphingopyxis</i>	0.535565
<i>Unclassified Aminicenantes</i>	0.535499
<i>Brevibacillus</i>	0.535499
<i>Campylobacter</i>	0.521678
<i>Methanobrevibacter</i>	0.520127
<i>Cryptobacterium</i>	0.513993
<i>Ruminiclostridium</i>	0.510680
<i>Eggerthella</i>	0.509646
<i>Lactococcus</i>	0.504261
<i>Butyricimonas</i>	0.491595
<i>Unclassified Bacteroidetes</i>	0.474846
<i>Shewanella</i>	0.459807
<i>Porphyromonas</i>	0.434124
<i>Parabacteroides</i>	0.427271
<i>Johnsonella</i>	0.423664
<i>Corynebacterium</i>	0.420475

Table S1: The entropy score of  $P(k|j)$  for all genera used. – *Continued from previous page*

Genus	entropy
<i>Brachyspira</i>	0.393297
<i>Gemella</i>	0.377228
<i>Gordonibacter</i>	0.369317
<i>Lactobacillus</i>	0.347828
<i>Unclassified Xanthomonadaceae</i>	0.339638
<i>Dickeya</i>	0.339637
<i>Kosakonia</i>	0.339551
<i>Moraxella</i>	0.339550
<i>Arthrospira</i>	0.339549
<i>Anaeroglobus</i>	0.339549
<i>Carnobacterium</i>	0.339549
<i>Shinella</i>	0.339507
<i>Stenotrophomonas</i>	0.339459
<i>Allofustis</i>	0.339458
<i>Lelliottia</i>	0.339458
<i>Bacillus j_bacteriumj</i>	0.336655
<i>Hafnia</i>	0.325478
<i>Granulicatella</i>	0.298182
<i>Collinsella</i>	0.294826
<i>Actinobacillus</i>	0.253157
<i>Mobiluncus</i>	0.253157
<i>Cupriavidus</i>	0.253089
<i>Unclassified Propionibacteriaceae</i>	0.253089
<i>Shuttleworthia</i>	0.253055
<i>Ethanoligenens</i>	0.253055
<i>Finegoldia</i>	0.253018
<i>Shigella</i>	0.252294
<i>Coprobacter</i>	0.251584
<i>Rikenella</i>	0.240003
<i>Parasutterella</i>	0.239172
<i>Lachnoanaerobaculum</i>	0.235800
<i>Clostridium</i>	0.216568
<i>Sinorhizobium</i>	0.203630
<i>Aeromicrobium</i>	0.203630
<i>Tetrasphaera</i>	0.203574
<i>Janthinobacterium</i>	0.203572
<i>Dysgonomonas</i>	0.203572
<i>Microbunatus</i>	0.203546
<i>Pantoea</i>	0.203546
<i>Synergistes</i>	0.203514
<i>Unclassified Firmicutes</i>	0.174778
<i>Vibrio</i>	0.171243
<i>Anaerovibrio</i>	0.171171
<i>Bulleidia</i>	0.171171
<i>Unclassified Oxalobacteraceae</i>	0.171171
<i>Lachnobacterium</i>	0.171144
<i>Raoultella</i>	0.169913
<i>Abiotrophia</i>	0.169097
<i>Anaerotruncus</i>	0.168921
<i>Streptomyces</i>	0.157833
<i>Thermus</i>	0.148239
<i>Trueperella</i>	0.148238
<i>Serratia</i>	0.148238
<i>Macrococcus</i>	0.148238
<i>Terrisporobacter</i>	0.148238

Table S1: The entropy score of  $P(k|j)$  for all genera used. – *Continued from previous page*

Genus	entropy
<i>Parascardovia</i>	0.148238
<i>Paenibacillus</i>	0.136984
<i>Riemerella</i>	0.131086
<i>Sphingomonas</i>	0.131048
<i>Candidatus Arthromitus</i>	0.131048
<i>Negativicoccus</i>	0.131030
<i>Bacteroides</i>	0.122743
<i>Candidatus Stoquefichus</i>	0.117656
<i>Aggregatibacter</i>	0.112499
<i>Pseudoalteromonas</i>	0.106937
<i>Unclassified Betaproteobacteria</i>	0.106937
<i>Robinsoniella</i>	0.106905
<i>Filifactor</i>	0.106891
<i>Brevibacterium</i>	0.106891
<i>Kocuria</i>	0.098099
<i>Unclassified Alphaproteobacteria</i>	0.098099
<i>Georgenia</i>	0.098099
<i>Mogibacterium</i>	0.098071
<i>Paracoccus</i>	0.098056
<i>Morganella</i>	0.090671
<i>Rothia</i>	0.089105
<i>Nocardioides</i>	0.084365
<i>Staphylococcus</i>	0.082676
<i>Unclassified Coriobacteriaceae</i>	0.078953
<i>Anaerofustis</i>	0.078953
<i>Slackia</i>	0.078174
<i>Pasteurella</i>	0.070096
<i>Aerococcus</i>	0.070047
<i>Anaerococcus</i>	0.066380
<i>Anaerosalibacter</i>	0.057431
<i>Unclassified Candidatus Saccharibacteria</i>	0.057431
<i>Gardnerella</i>	0.052776
<i>Plesiomonas</i>	0.050727
<i>Xanthomonas</i>	0.048869
<i>Elizabethkingia</i>	0.048838
<i>Succinimonas</i>	0.047110
<i>Selenomonas</i>	0.046316
<i>Oribacterium</i>	0.044003
<i>Beggiatoa</i>	0.041308
<i>Unclassified Bacteroidaceae</i>	0.040074
<i>Acidovorax</i>	0.035891
<i>Bifidobacterium</i>	0.035527
<i>Micrococcus</i>	0.029757
<i>Catonella</i>	0.029757
<i>Aeromonas</i>	0.028547
<i>Leclercia</i>	0.026936
<i>Kluyvera</i>	0.026434
<i>Burkholderia</i>	0.025968
<i>Geobacillus</i>	0.025960
<i>Alloscardovia</i>	0.025499
<i>Rhodococcus</i>	0.025052
<i>Prevotella</i>	0.024904
<i>Microbacterium</i>	0.024628
<i>Acinetobacter</i>	0.023983
<i>Varibaculum</i>	0.023434

Table S1: The entropy score of  $P(k|j)$  for all genera used. – *Continued from previous page*

Genus	entropy
<i>Listeria</i>	0.022350
<i>Alcanivorax</i>	0.022160
<i>Eikenella</i>	0.020773
<i>Oscillibacter</i>	0.020362
<i>Enterorhabdus</i>	0.018686
<i>Helicobacter</i>	0.017187
<i>Bifidobacterium</i>	0.013555
<i>Methylobacterium</i>	0.012762
<i>Leptotrichia</i>	0.012559
<i>Pyramidobacter</i>	0.010146
<i>Lysinibacillus</i>	0.009544
<i>Kingella</i>	0.008453
<i>Bordetella</i>	0.007775
<i>Cardiobacterium</i>	0.007596
<i>Methanomassiliicoccus</i>	0.006476
<i>Lautropia</i>	0.005972
<i>Succinatimonas</i>	0.004608
<i>Spiroplasma</i>	0.004588
<i>Acidiphilium</i>	0.003029
<i>Corallococcus</i>	0.002747
<i>Achromobacter</i>	0.002647
<i>Capnocytophaga</i>	0.002549
<i>Cetobacterium</i>	0.002347
<i>Butyrivibrio</i>	0.002281
<i>Coraliomargarita</i>	0.002105
<i>Enorma</i>	0.001869
<i>Acholeplasma</i>	0.000901
<i>Pediococcus</i>	0.000880
<i>Neisseria</i>	0.000365
<i>Treponema</i>	0.000171
<i>Acetobacter</i>	0.000142