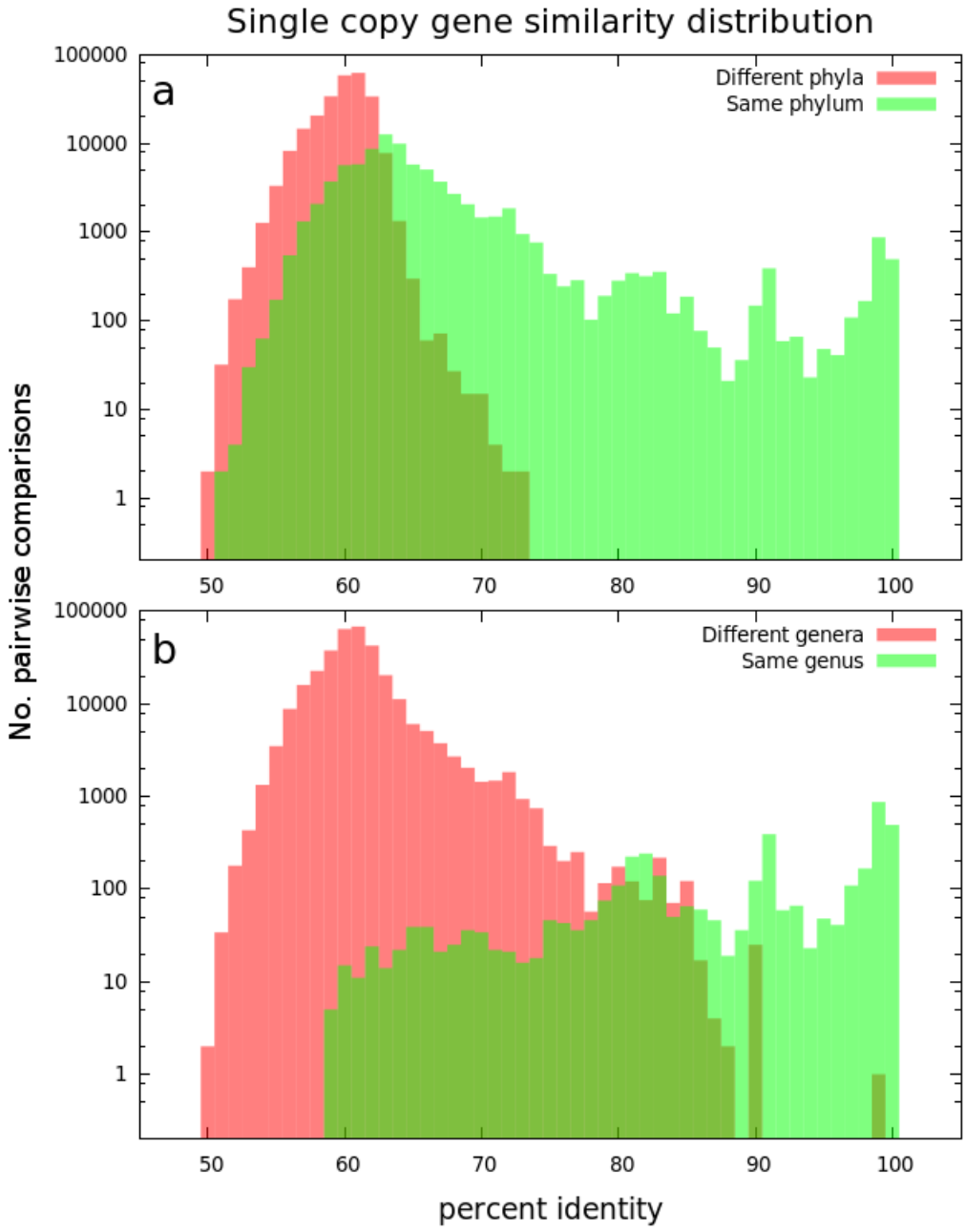
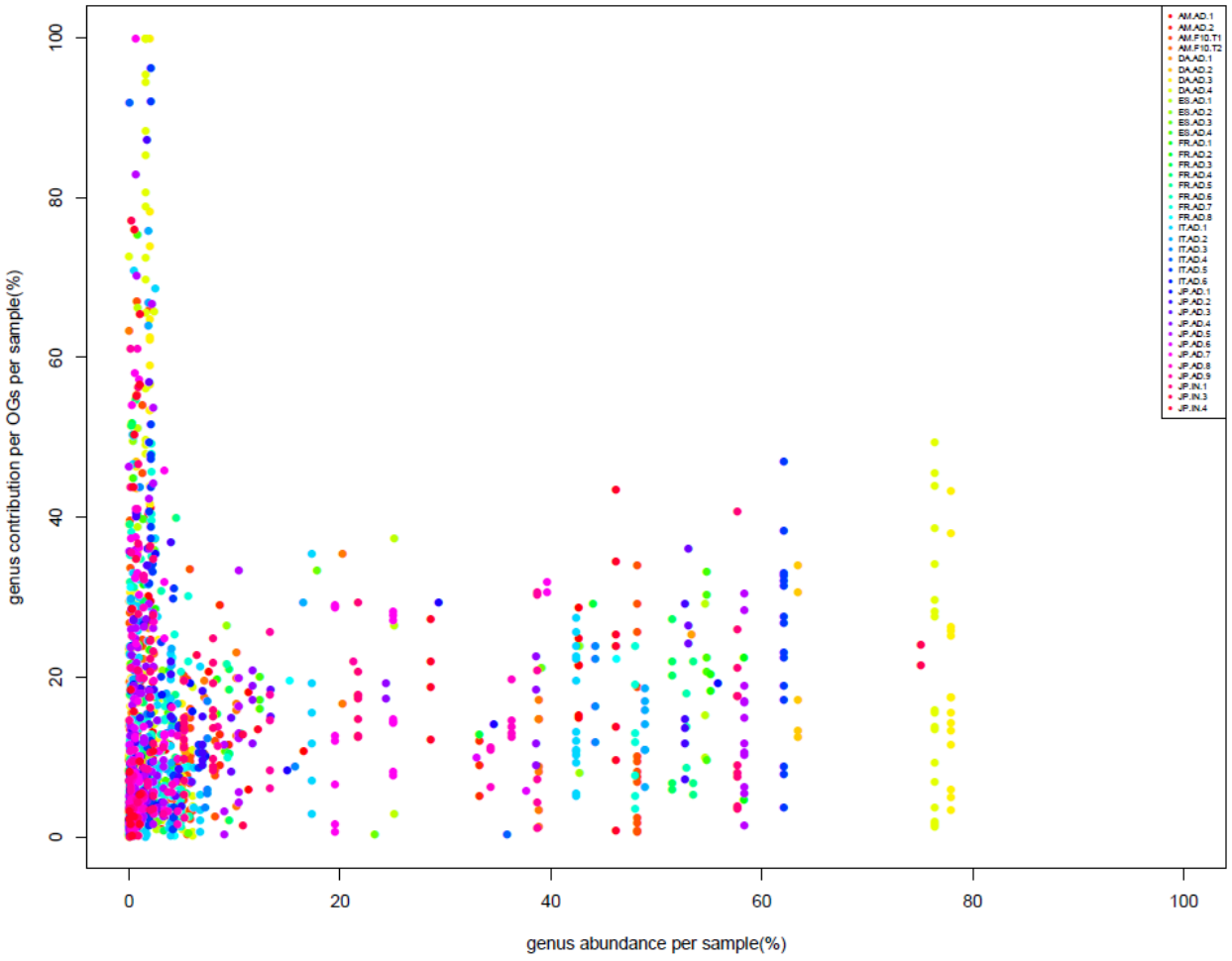
doi:10.1038/nature09944

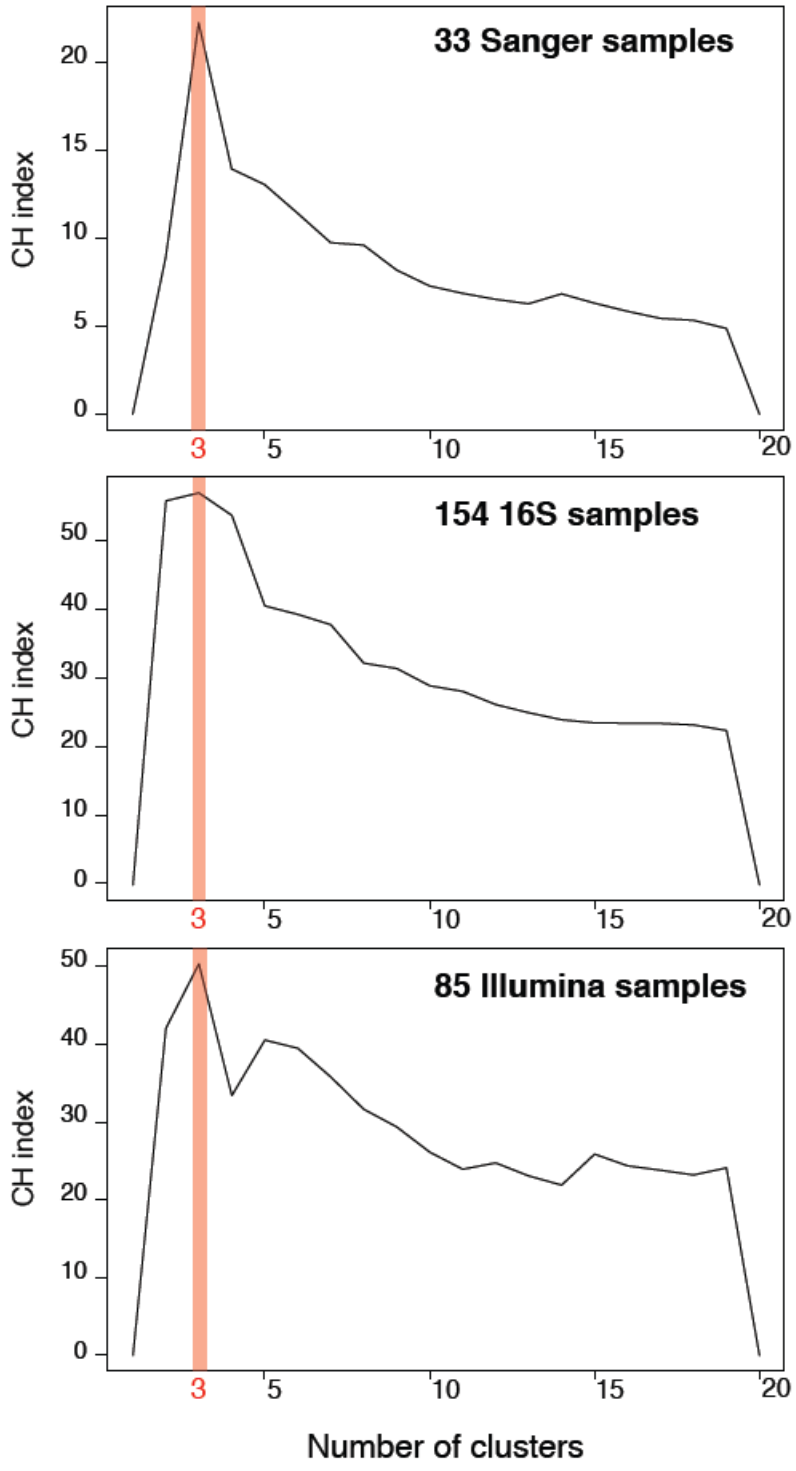


Supplementary Figure 1. Establishing DNA sequence similarity thresholds for phylum and genus levels

Sequence similarity distributions of pairwise alignments of 40 universal single copy genes from 835 microbial genomes reveal that (a) 65% DNA sequence similarity threshold accurately groups genomes within the same phylum (with 31.1% sensitivity and 0.77% false positive rate) and (b) 85% threshold accurately groups genomes within the same genera (with 63.23% sensitivity and 5.1% false positive rate). Pairwise comparisons of genomes within the same phylum (genus) are colored green and different phyla (genera) are colored red.

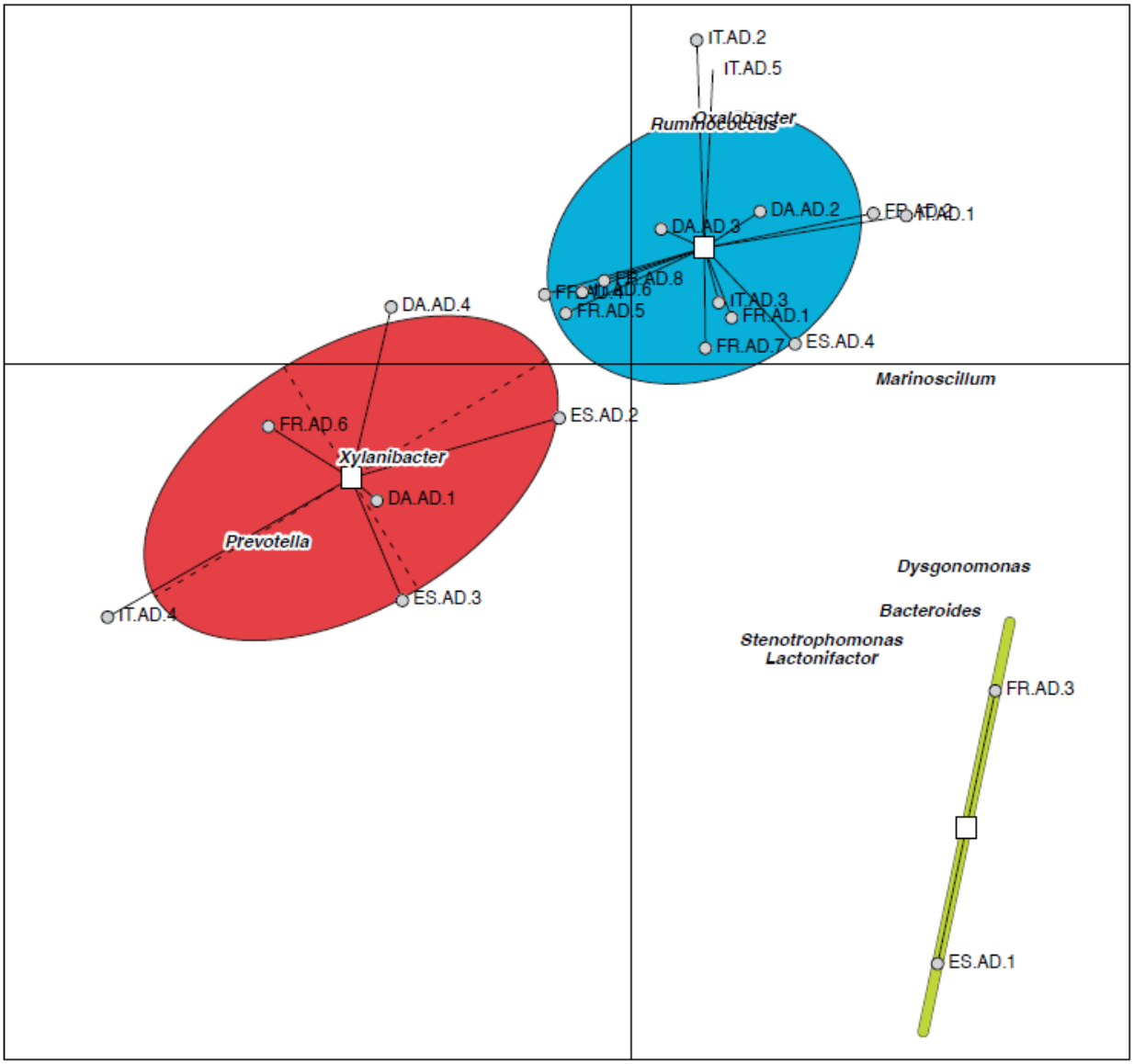


Supplementary Figure 2. Correlation between the abundance of a genus in the sample and its contribution to an abundant (top 20 percentile) orthologous group. Few functions (on the top left) are almost entirely contributed by genera that are low abundant in the samples.

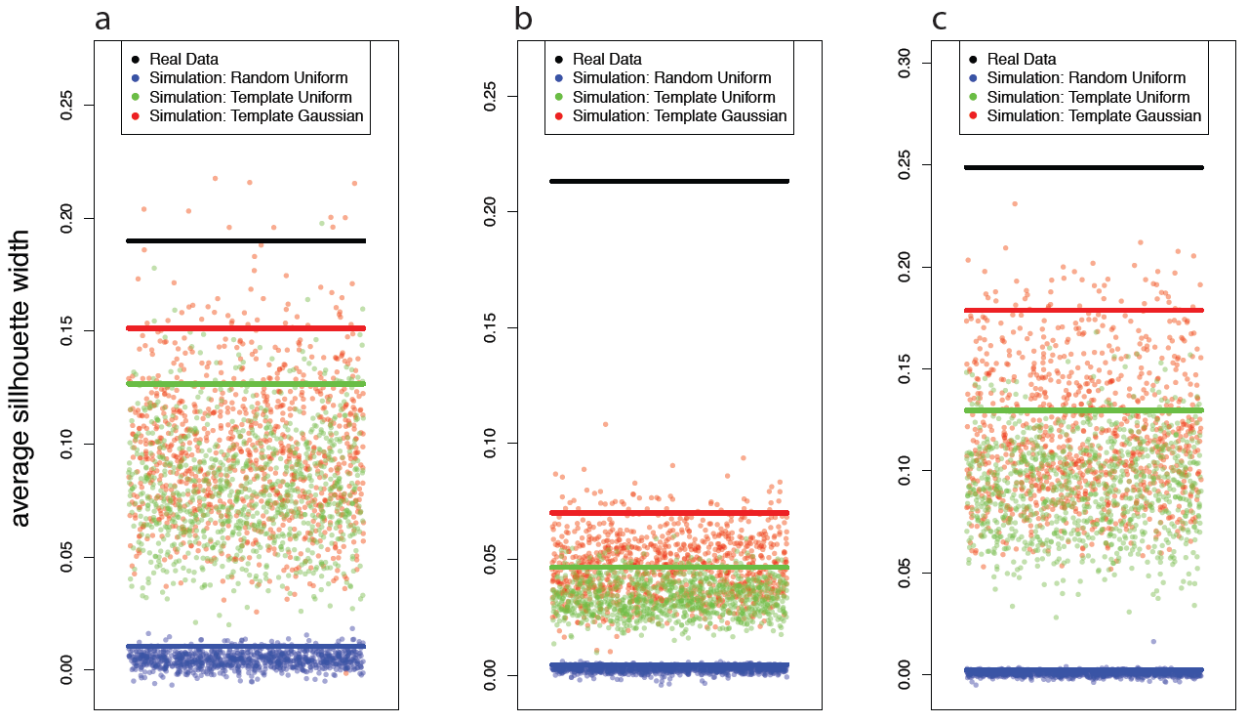


Supplementary Figure 3. Estimating the number of clusters in the gut metagenome datasets.

Calinski-Harabasz (CH) index predicts 3 clusters each for (a) Sanger-based metagenome, (b) pyrosequencing-based 16S and (c) Illumina-based metagenome datasets. For details on how the CH index is calculated see Supplementary Methods Section 7.3.

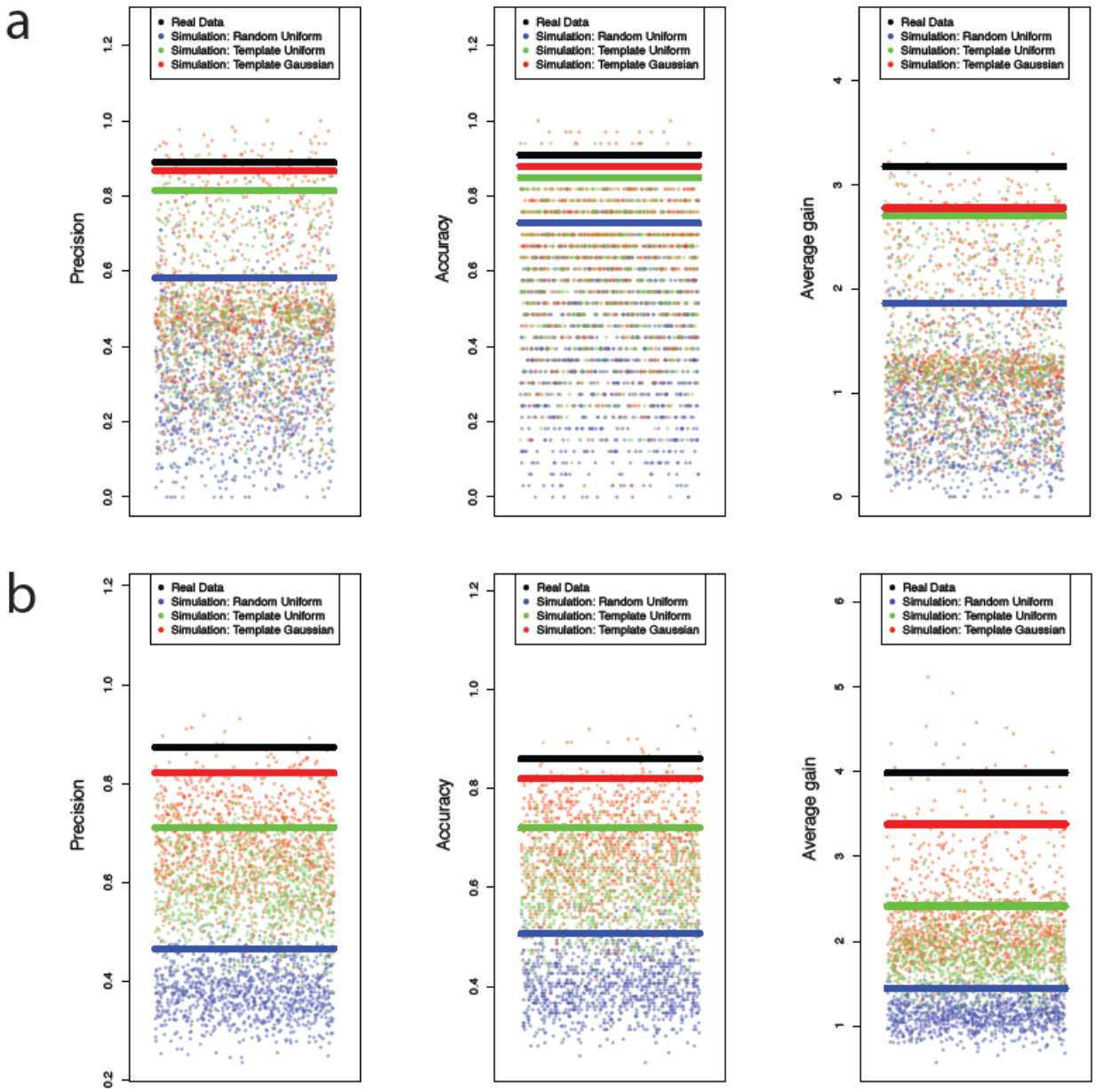


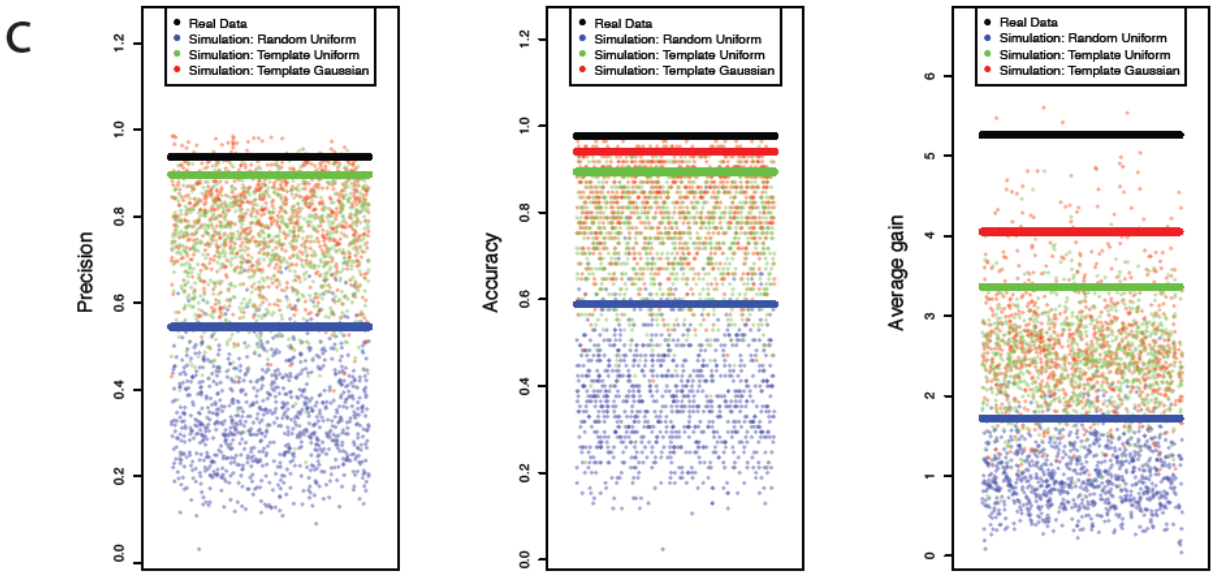
Supplementary Figure 4. Between class analysis of enterotype clusters on the HITChip data from 22 European individuals reveals the same drivers.



Supplementary Figure 5. Assessing the robustness of clusters.

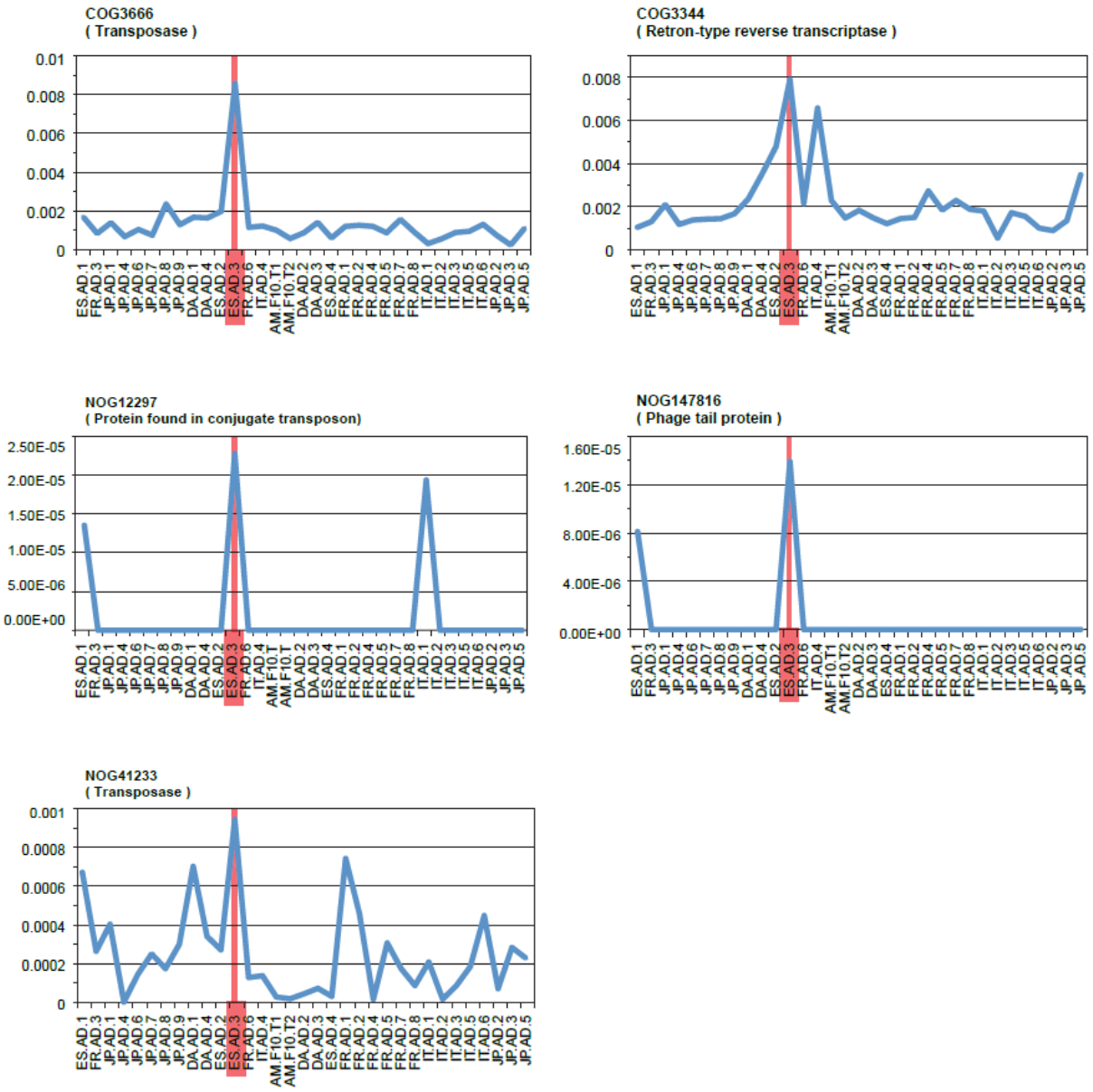
Clustering of real data (black bar) is more robust than three kinds of simulations (colored dots) based on silhouette coefficient for (a) Sanger-based metagenome, (b) pyrosequencing-based 16S and (c) Illumina-based metagenome datasets. See Supplementary Methods Sections 7.4 and 8 for details. Colored bars represent the 95% confidence intervals. P-values are listed in Supplementary Table 24.


a
Real Data
Simulation: Random Uniform
Simulation: Template Uniform
Simulation: Template Gaussian
Precision
Accuracy
Average gain
b
Real Data
Simulation: Random Uniform
Simulation: Template Uniform
Simulation: Template Gaussian
Precision
Accuracy
Average gain



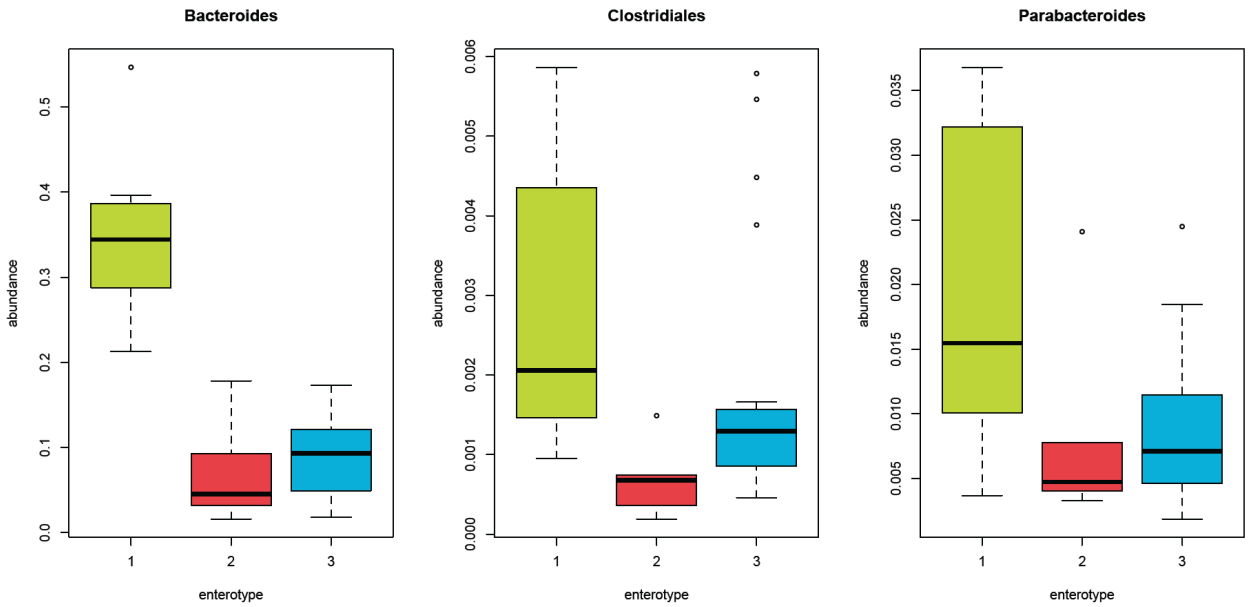
Supplementary Figure 6. Assessing the accuracy of classification.

Predictive power of the clusters from real data (black bar) is stronger than those from three kinds of simulations (colored dots) based on precision, accuracy and average gain for (a) Sanger-based metagenome, (b) pyrosequencing-based 16S and (c) Illumina-based metagenome datasets. See Supplementary Methods Sections 8 and 9 for details. Colored bars represent the 95% confidence intervals. P-values are listed in Supplementary Table 24.

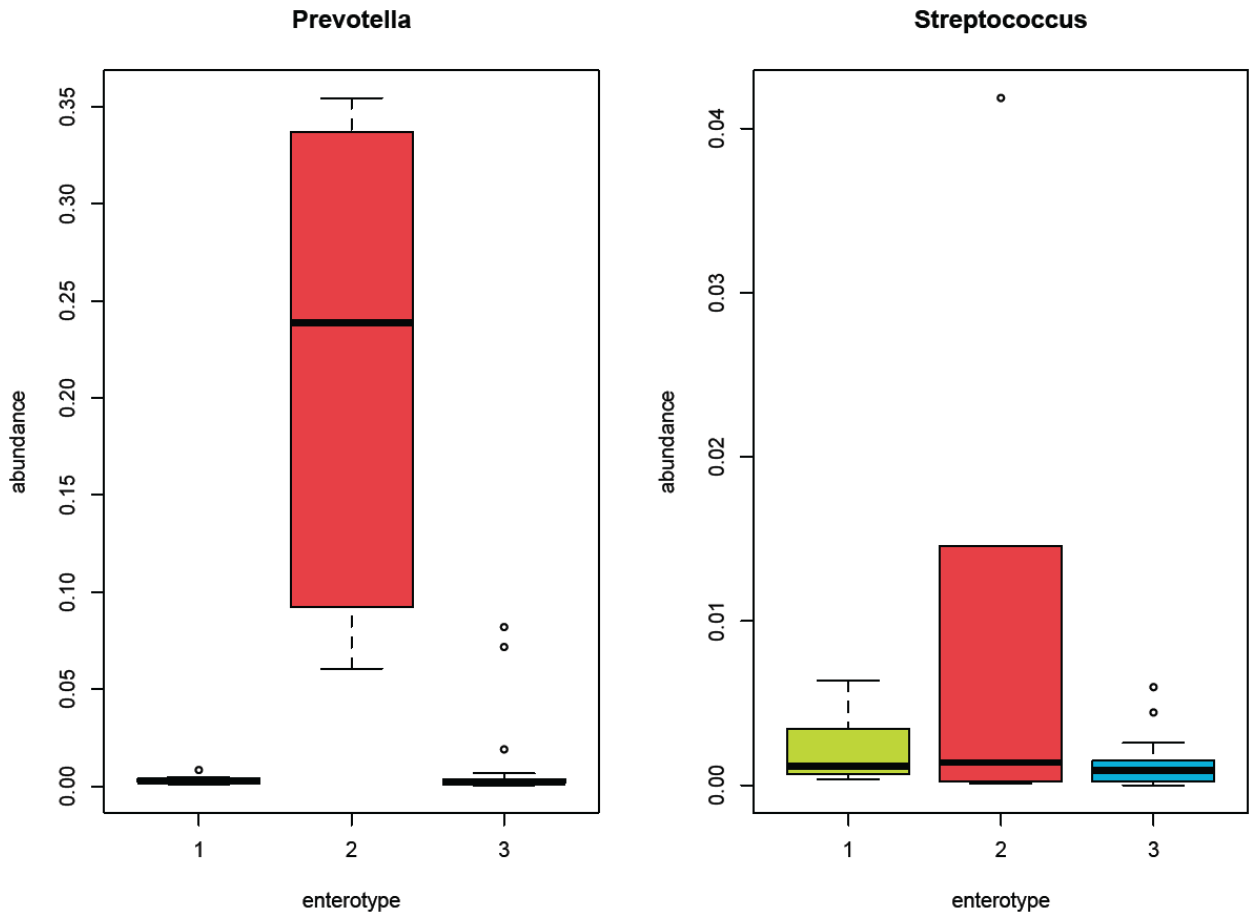


Supplementary Figure 7. Elevated levels of phage-related proteins in Spanish individual ES-AD-3.

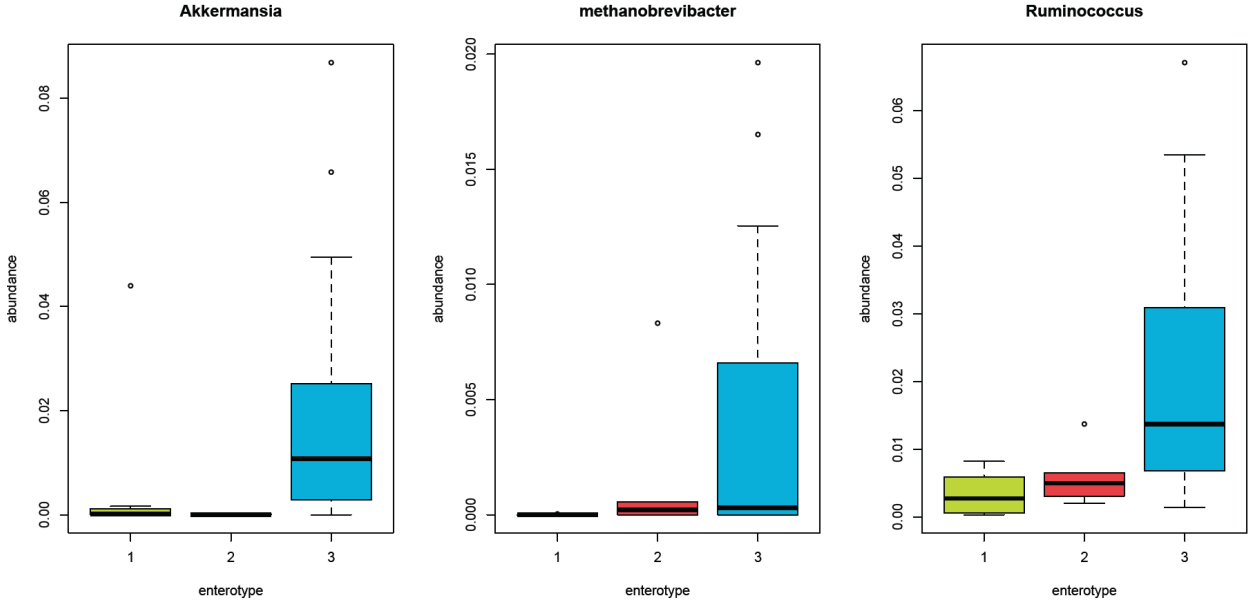
ES-AD-3 has more than five times higher abundance than the rest of the samples in the five orthologous groups shown.



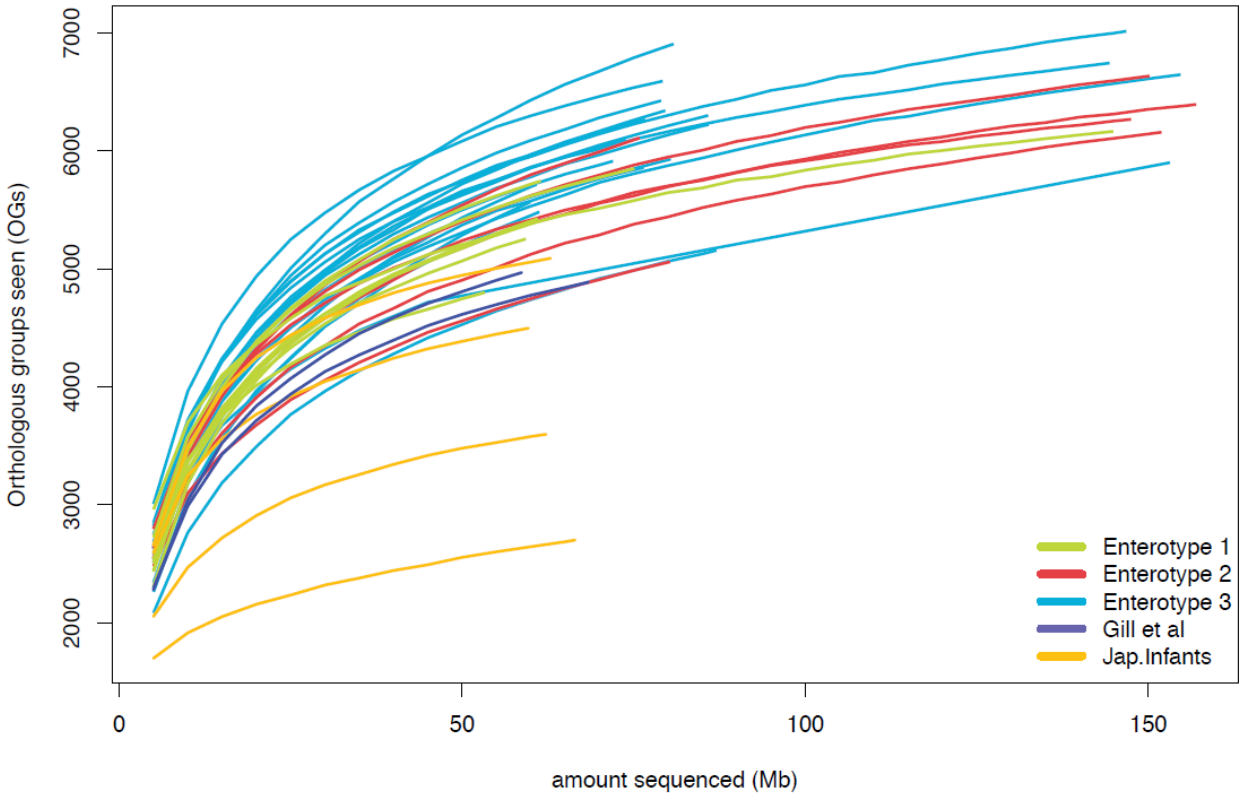
Supplementary Figure 8. Enrichment of *Bacteroides, Clostridiales* and *Parabacteroides* in enterotype 1.



Supplementary Figure 9. Enrichment of *Prevotella* and *Streptococcus* in enterotype 2.

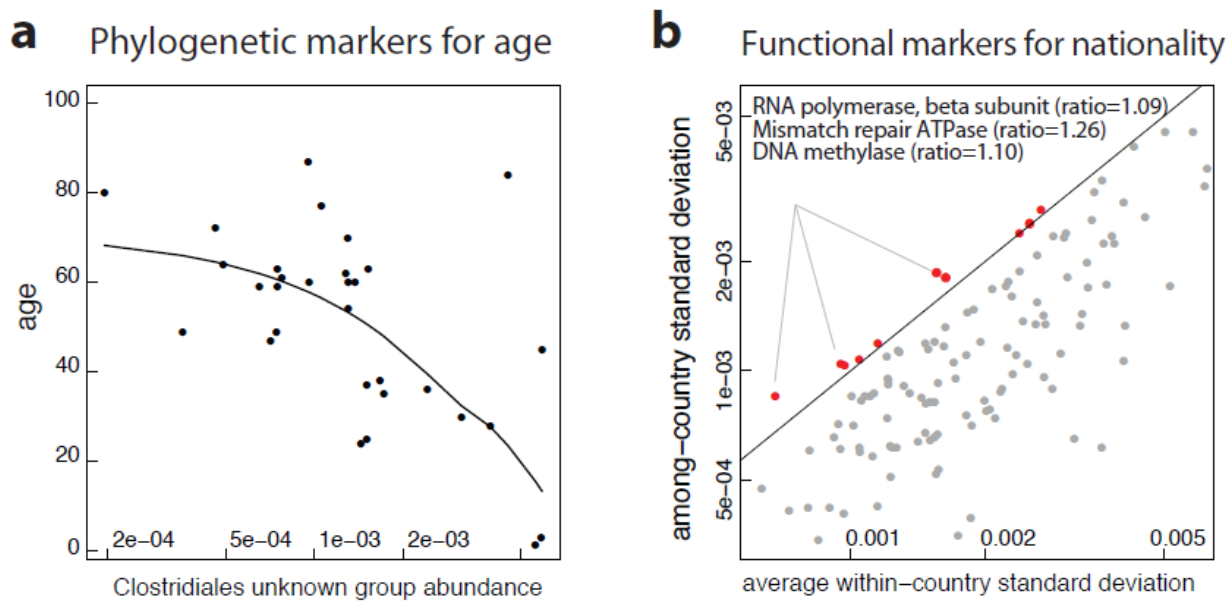


Supplementary Figure 10. Enrichment of *Akkermansia, Methanobrevibacter* and *Ruminococcus* in enterotype 3.

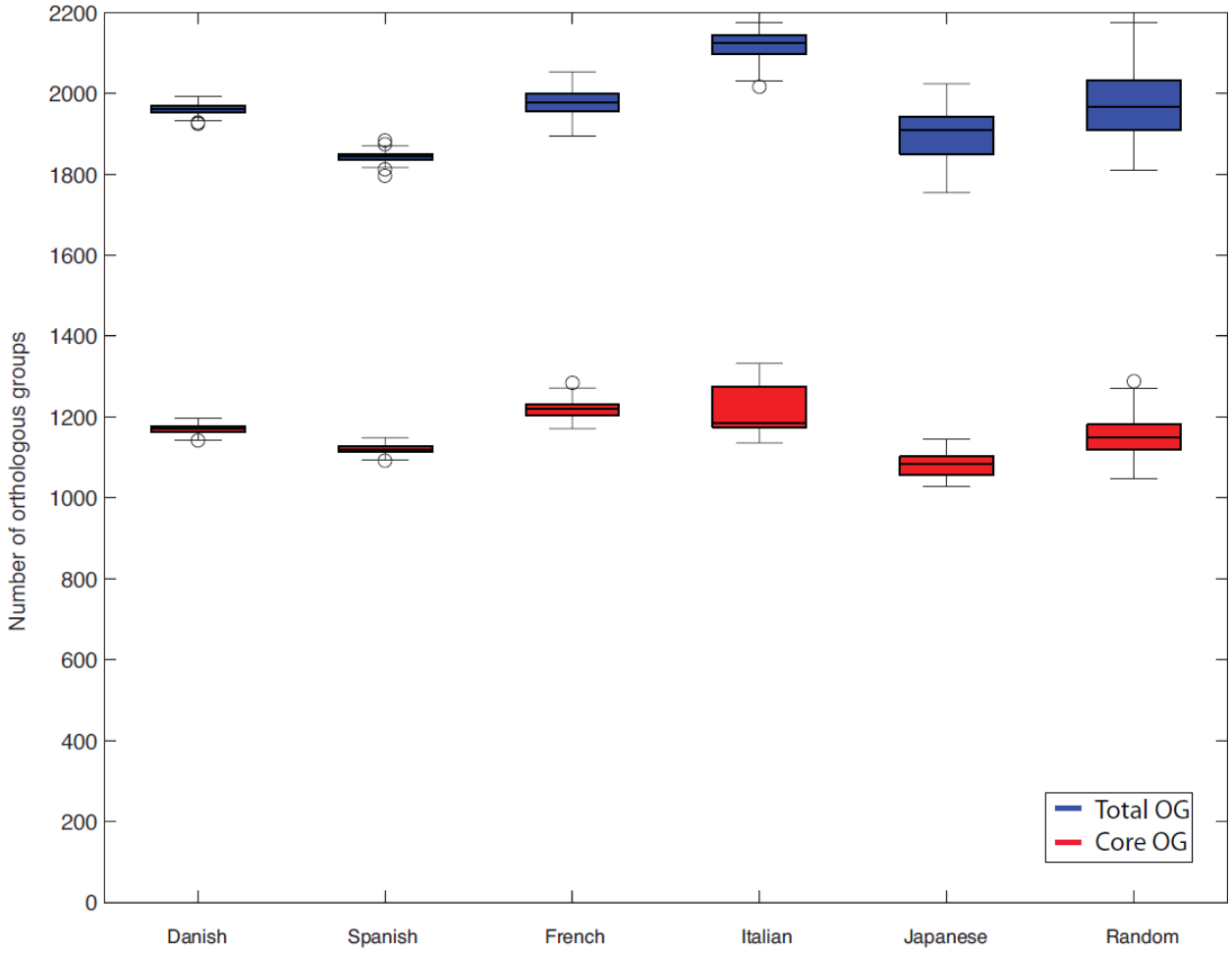


Supplementary Figure 11. Effect of sequencing depth on function retrieval.

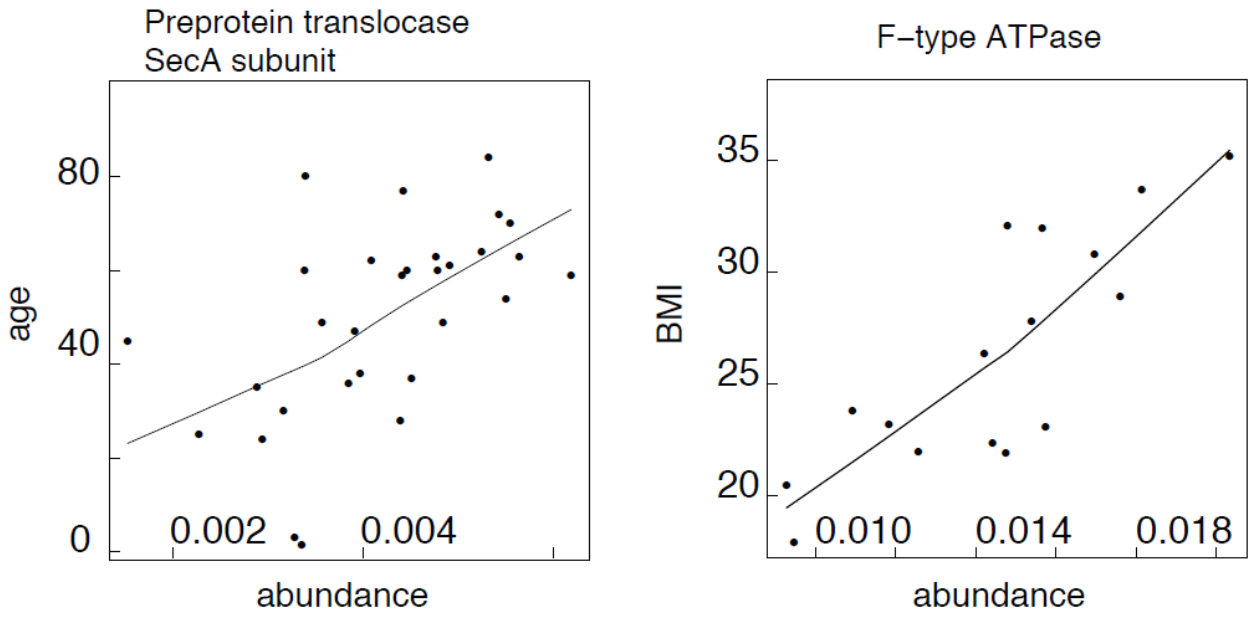
Number of unique orthologous groups identified in 39 samples relative to the sequencing depth. Each sample is colored according to the enterotype it belongs to. Six samples that were removed before the enterotype clustering (four Japanese infants and two American individuals) are also shown. There is no difference between the enterotypes based on functional alpha diversity.



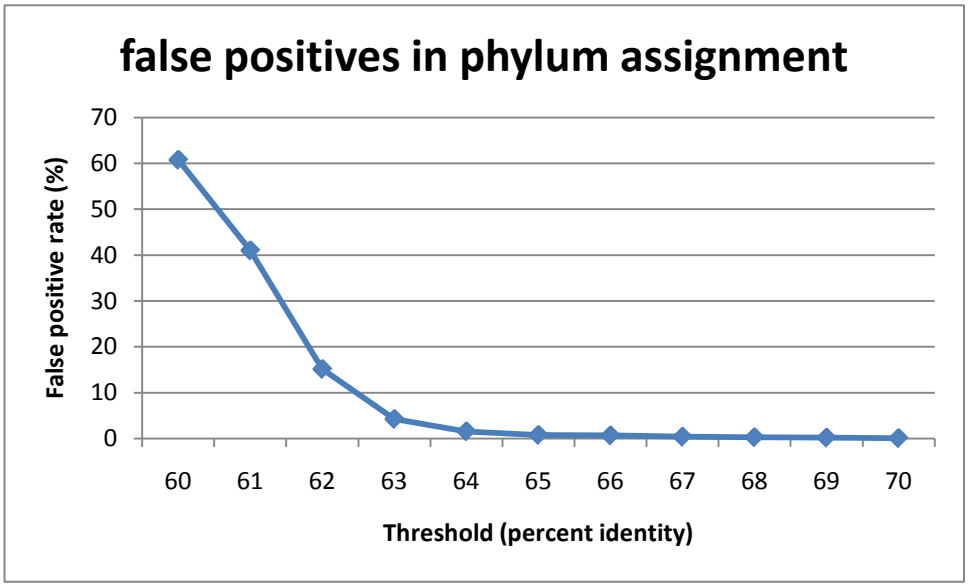
Supplementary Figure 12. Correlations with host properties. (a) Negative correlation of abundance of an unknown Clostridiales genus with age (pairwise correlation p=0.02, rho=-0.55). (b) Variation of orthologous groups (OGs) with nationality. The plot compares the among-nationality standard deviation (SD) with the within-nationality SD. Points above the diagonal (red, discussed in text) represent OGs whose abundance varies more among than within nationalities. See Supplementary Table 14 for a full list.

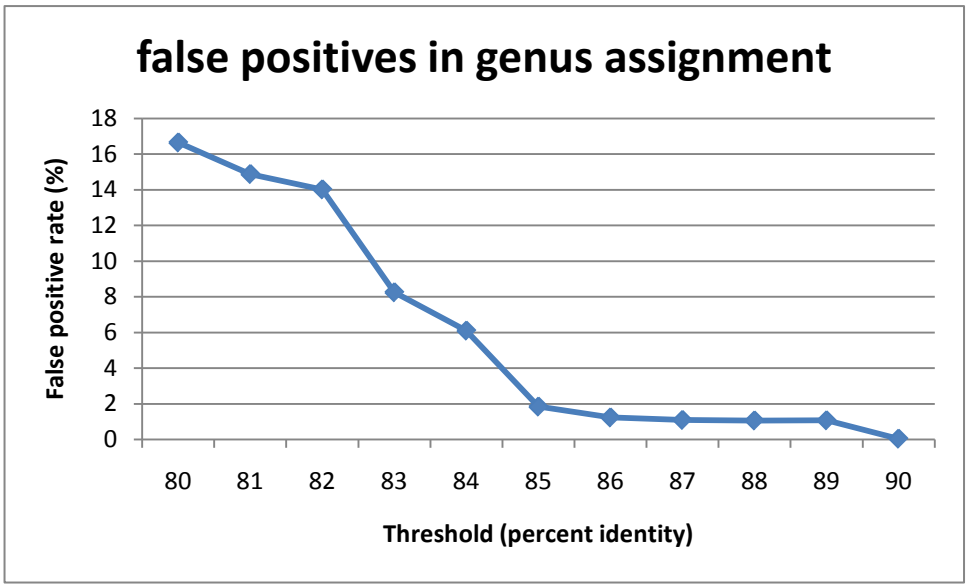


Supplementary Figure 13. Core metagenome sizes in metagenomes from different nations are similar.

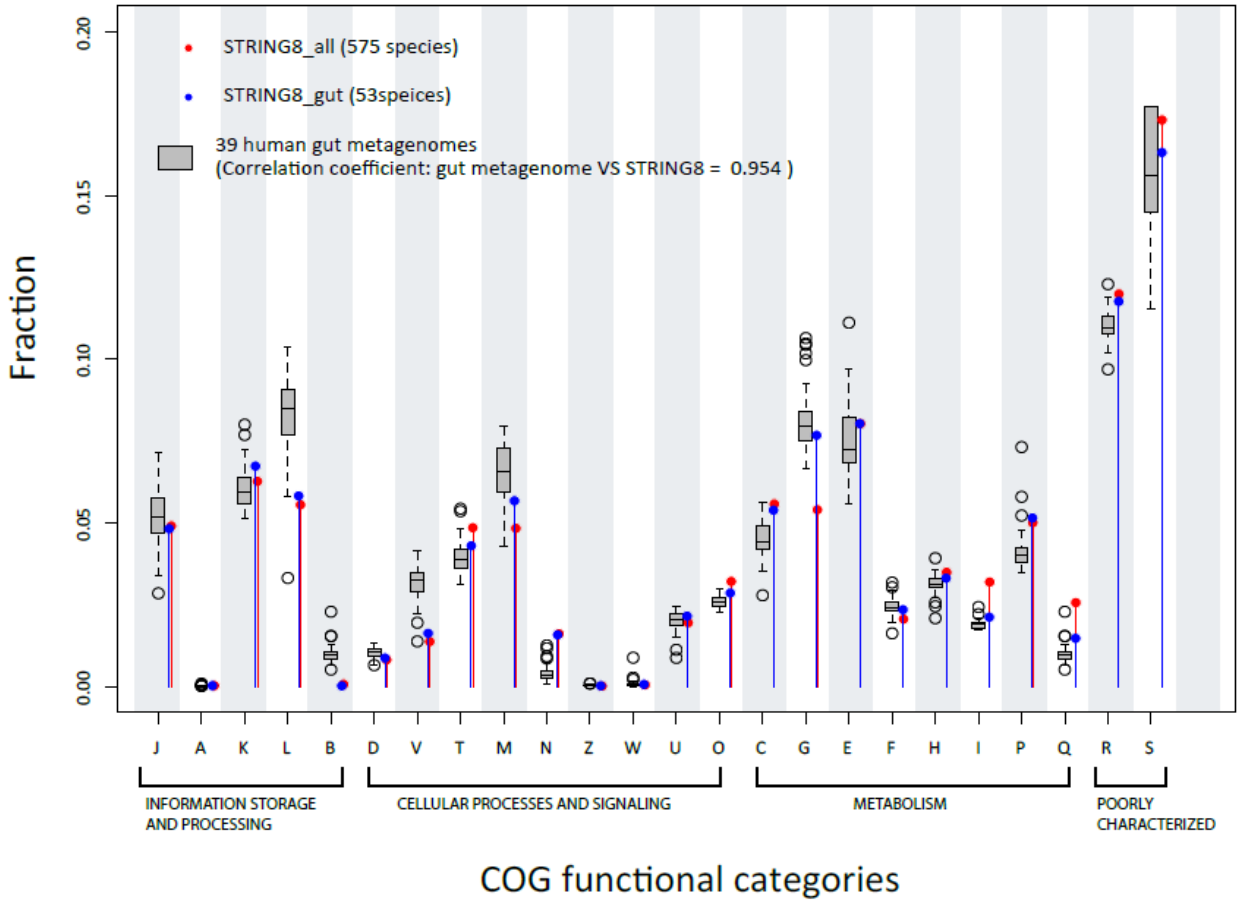


Supplementary Figure 14. Correlations of functions with host properties. Left: pairwise correlation of preprotein translocase secA subunit (COG0653) with host age (p=0.0008, rho=0.57). Right: pairwise correlation of F-type ATPase (KEGG module M00286) with host body mass index (p=0.04, rho=0.78).





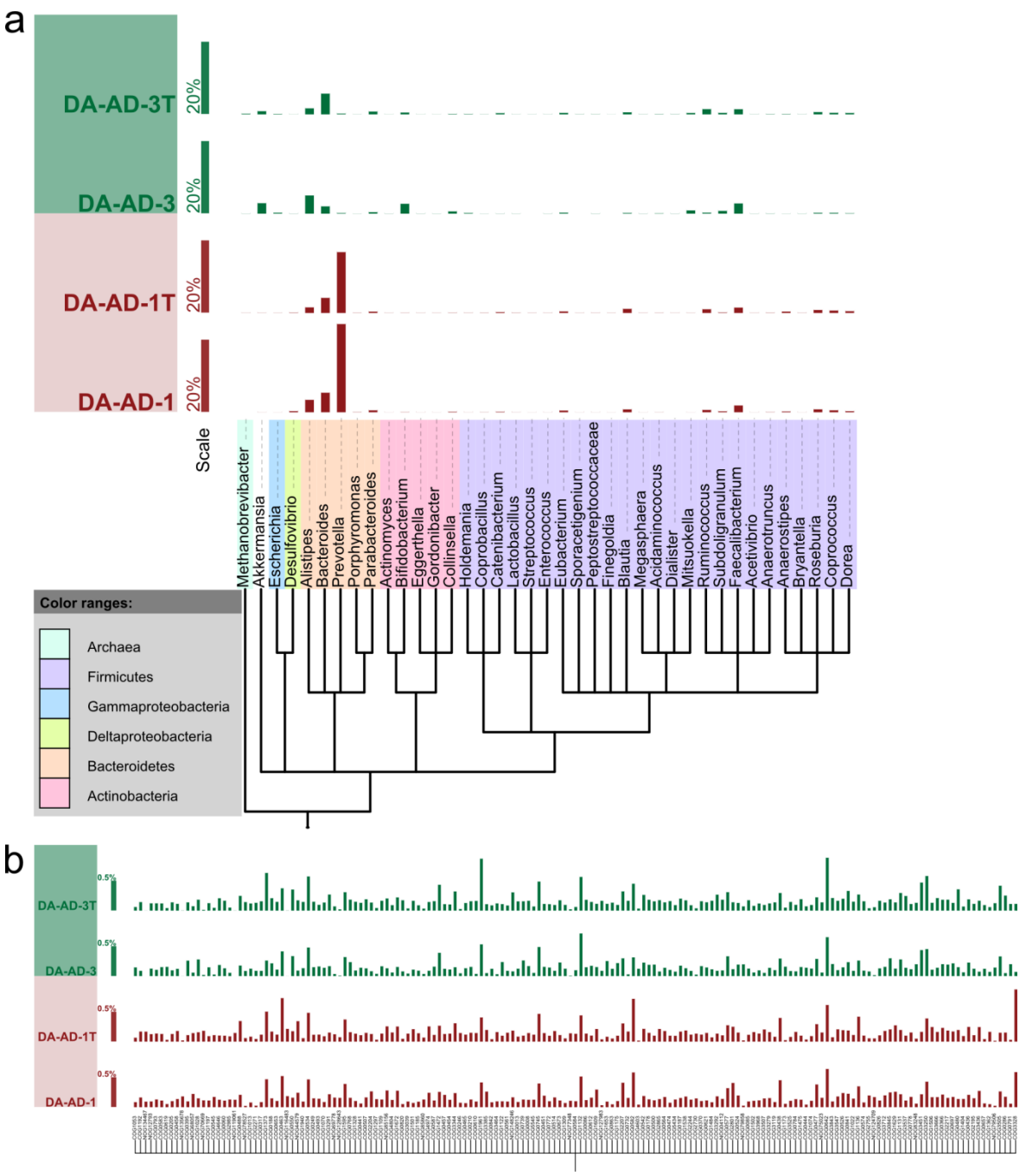
Supplementary Figure 15. False positive rates at the phylum and genus levels estimated by pairwise comparisons of 40 marker genes for different sequence similarity thresholds.



Supplementary Figure 16. Abundance distribution of COG functional categories

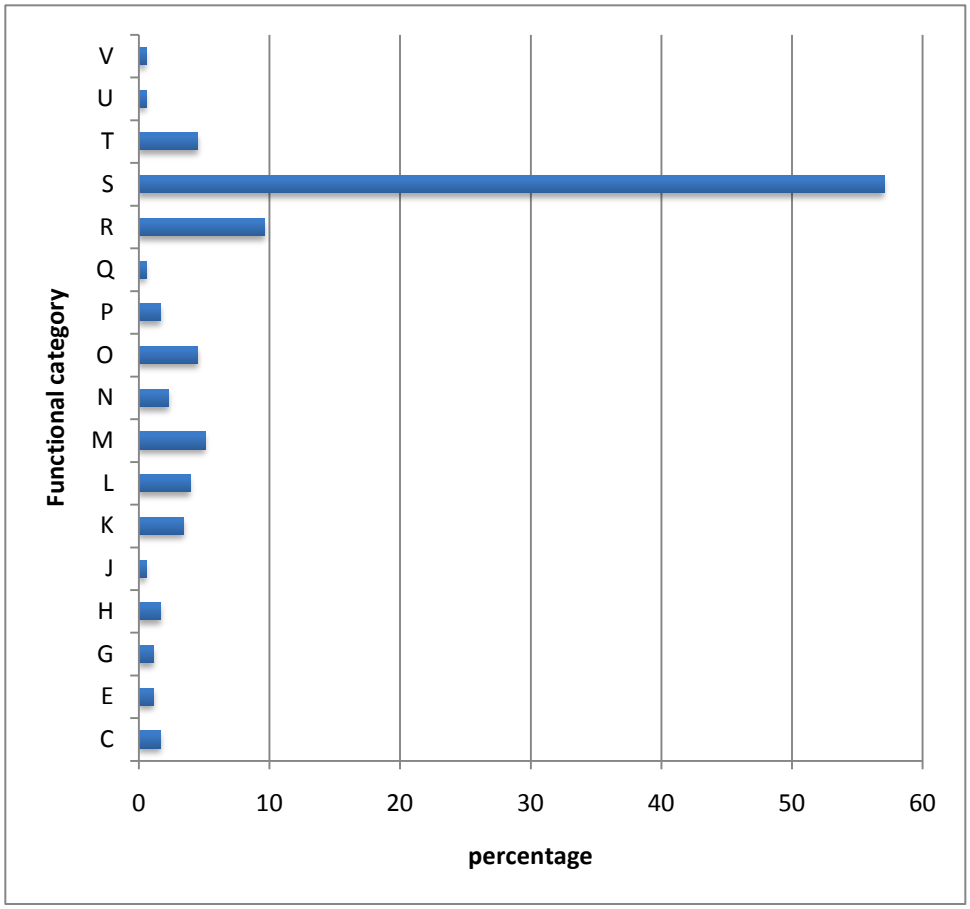
Abundance distribution of COG functional categories in 575 bacterial species in STRING v8.0 (red), 53 gut-associated species stored in the STRING v8.0 (blue) and 39 metagenomic samples (gray boxes). Distributions of metagenomic samples are similar to gut-associated bacteria.

Functional category L (replication, recombination and repair), V (defense mechanisms), M (cell wall/membrane/envelope biogenesis) and G (carbohydrate transport and metabolism) are enriched in the gut metagenomes. In particular, enrichment of L and V is not supported by gut specific bacteria in STRING (blue). This implies metagenomic data includes several bacteria which are not in STRING, and have higher ratio of these functional categories.



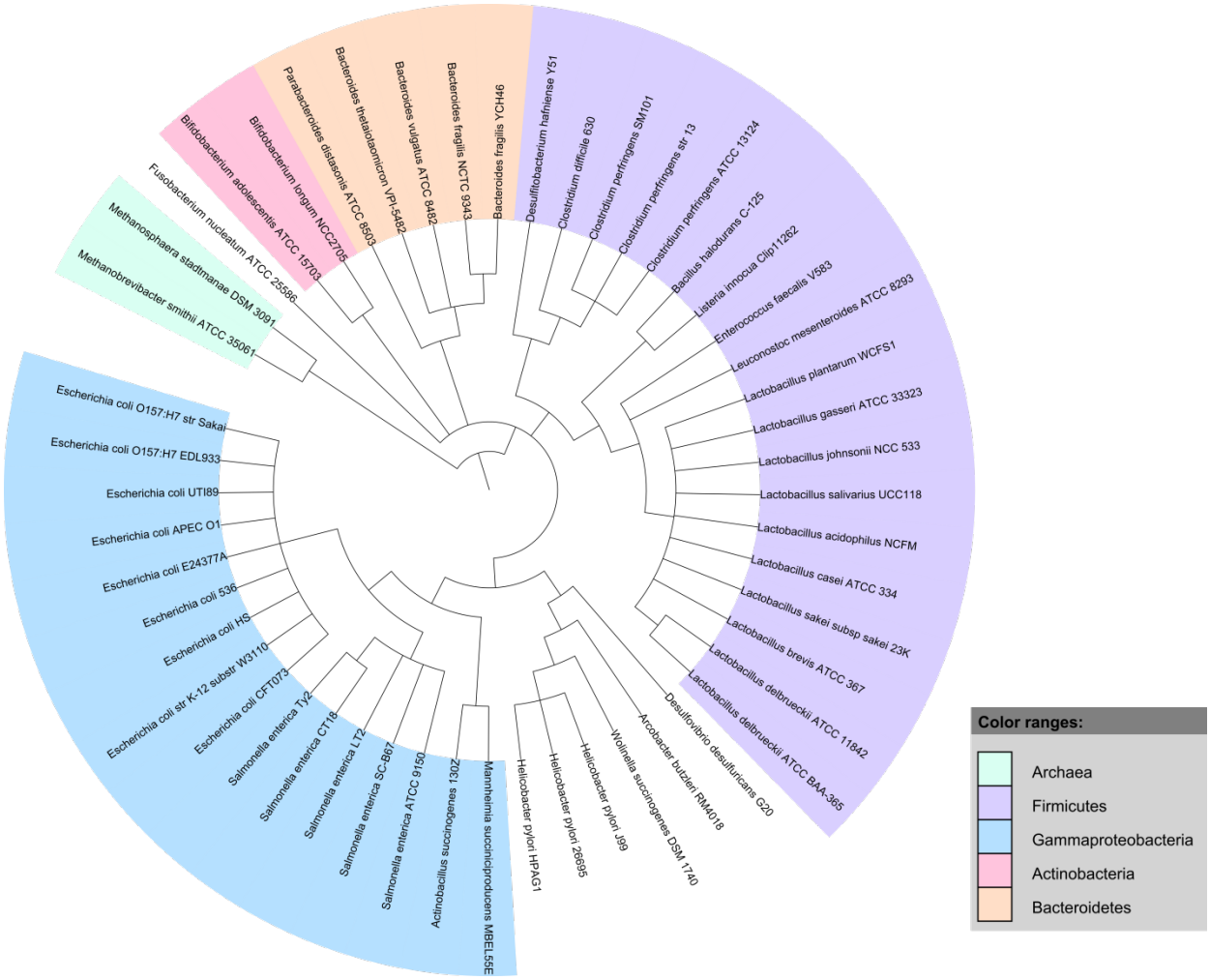
Supplementary Figure 17. Genus and eggNOG orthologous group (OG) abundance distributions of Sanger and 454 Titanium based sequences from the same samples are similar.

DA-AD-1 and DA-AD-1T are Sanger and 454 Titanium based sequences from MH6. DA-AD-3 and DA-AD-3T are from MH12. a) Genus abundance. Pearson correlation coefficients between DA-AD-1/DA-AD-1T is 0.9852, between DA-AD-3/DA-AD-3T is 0.9968. (Background for DA-AD-1/DA-AD-3 is 0.9037; DA-AD-1T/DA-AD-3T is 0.9624). b) Abundance of 50 most abundant orthologous groups. Pearson correlation coefficients between DA-AD-1/DA-AD-1T is 0.9482, between DA-AD-3/DA-AD-3T is 0.9153. (Background for DA-AD-1/DA-AD-3 is 0.8408; DA-AD-1T/DA-AD-3T is 0.8436).

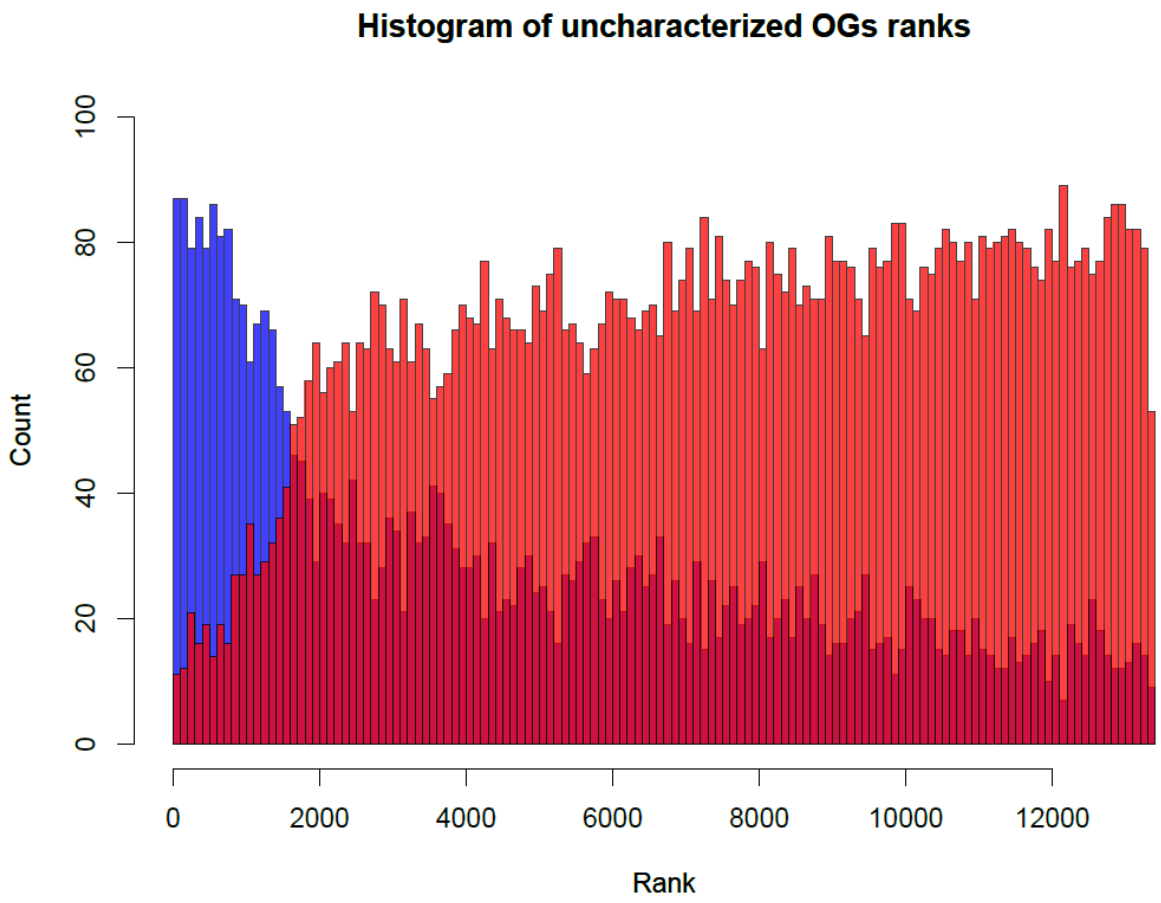


Supplementary Figure 18. Functional composition of most differing OGs between Sanger and 454 datasets.

The 1% OGs that differ most in abundance between DA-AD-1/DA-AD-1T (1% with highest DA-AD-1/DA-AD-1T ratio + 1% with lowest) and DA-AD-3/DA-AD-3T (likewise) were mapped on COG functional categories, excluding OGs with zero abundance in any sample to avoid artifacts. These results indicate that mostly unknown functions (from smaller, low abundant OGs, data not shown) differ between these two datatypes.

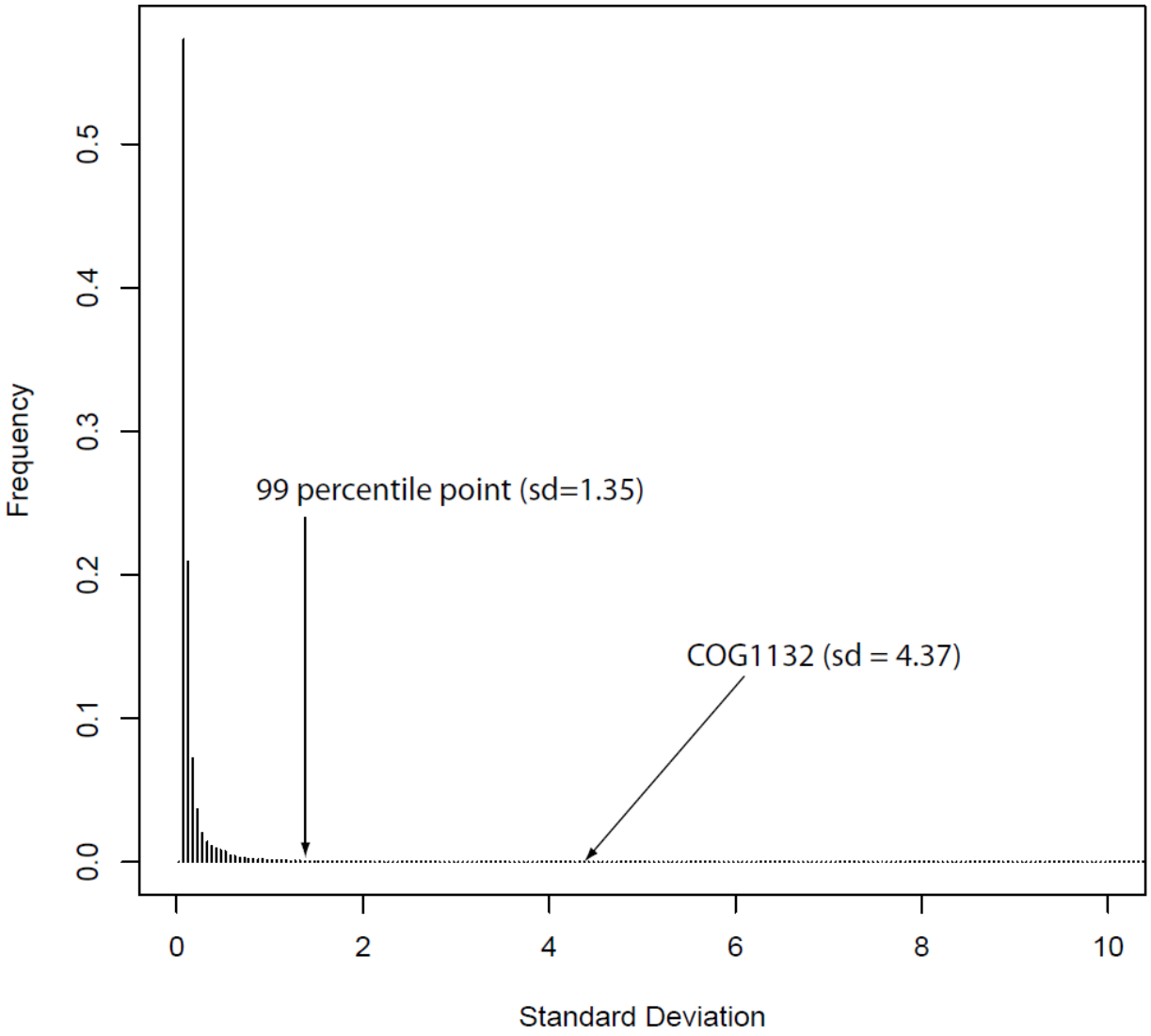


Supplementary Figure 19. Phylogenetic tree of the 53 gut-specific genomes out of the 575 prokaryotic genomes in STRING.



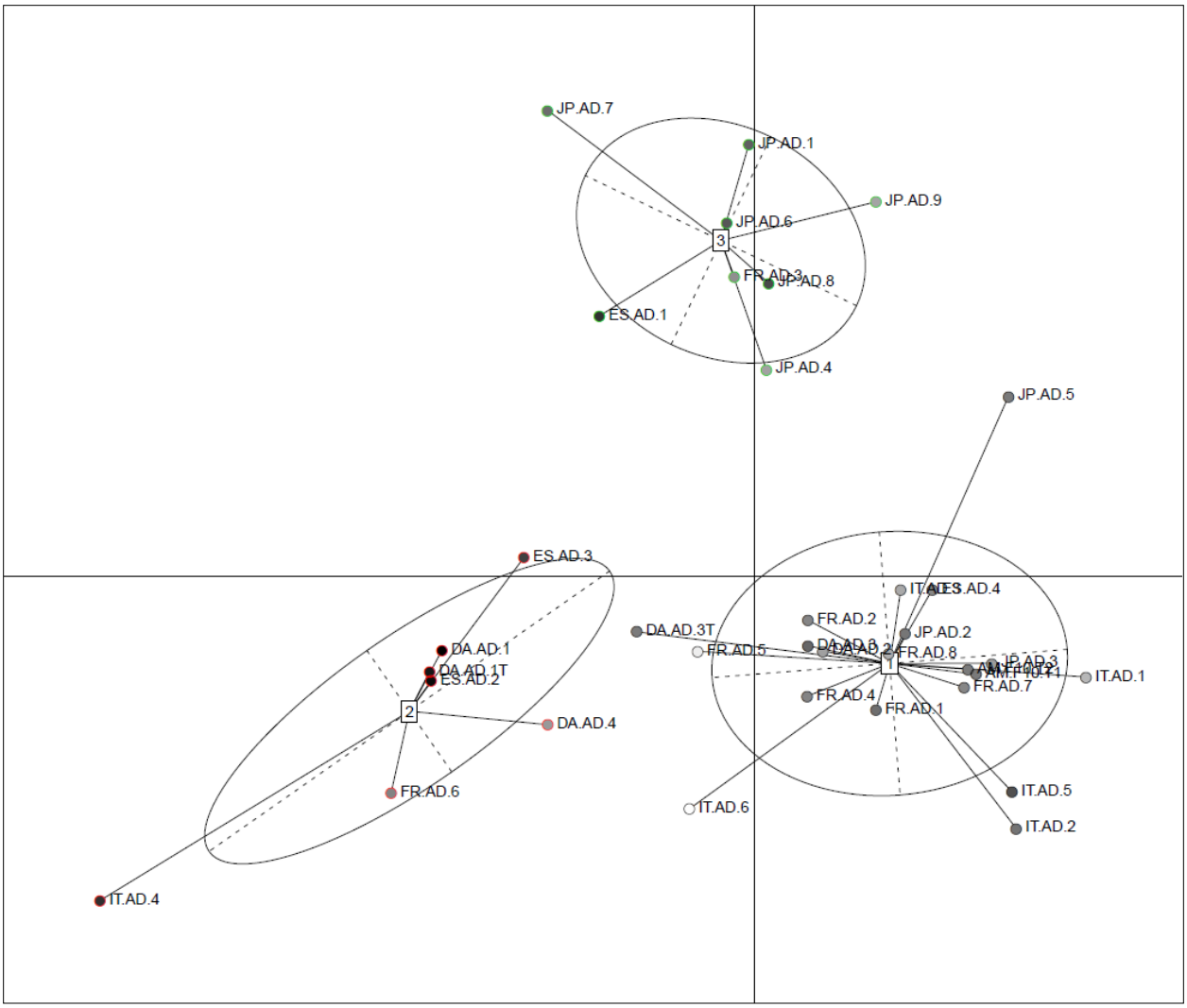
Supplementary Figure 20. Lower ranked functions are enriched in uncharacterized orthologous groups (OGs), agreeing with the findings of ref. 79.

A histogram of the number of functionally characterized (blue) and uncharacterized (red) OGs ranked by their average abundance in 35 metagenomes shows that uncharacterized proteins usually form small OGs (hence are predominantly ranked lower in abundance). In contrast, functionally characterized OGs are large with many genes and are usually ranked higher in abundance.

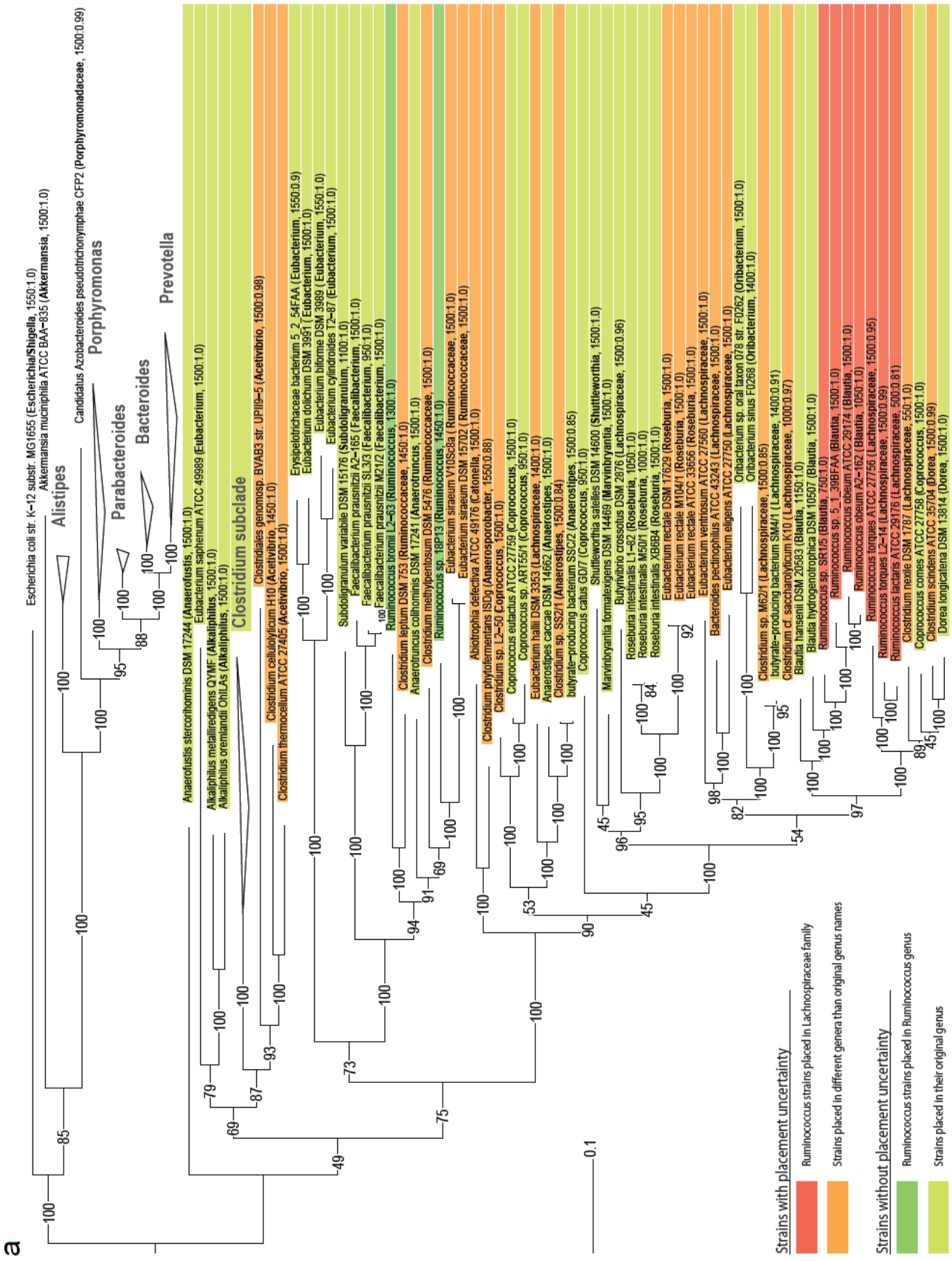


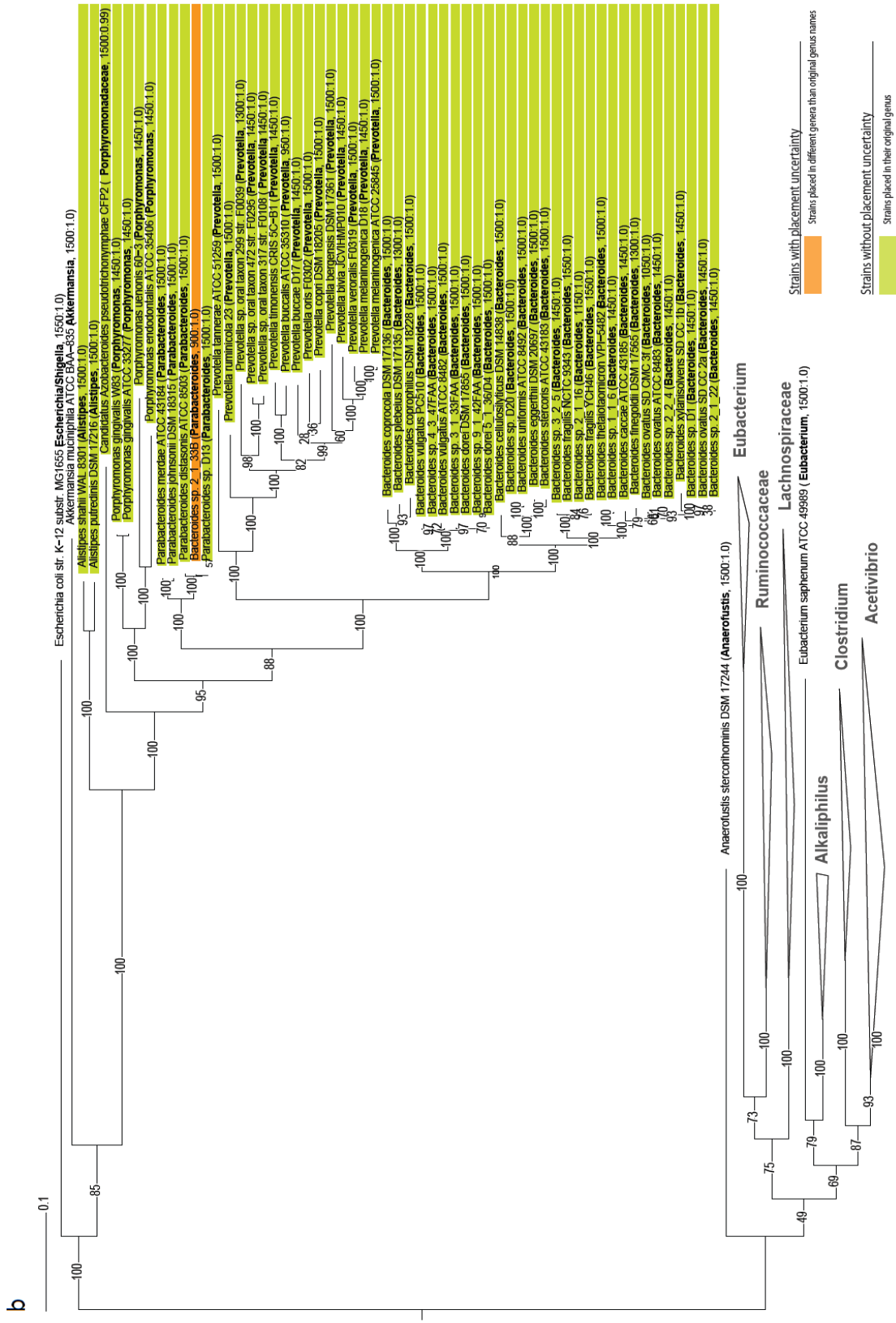
Supplementary Figure 21. Distribution of standard deviation of the number of genes in orthologous groups in the STRING database.

COG1132 is at 99.8 percentile, meaning only 0.2% of the orthologous groups in STRING have a higher variation in the number of genes than COG1132.

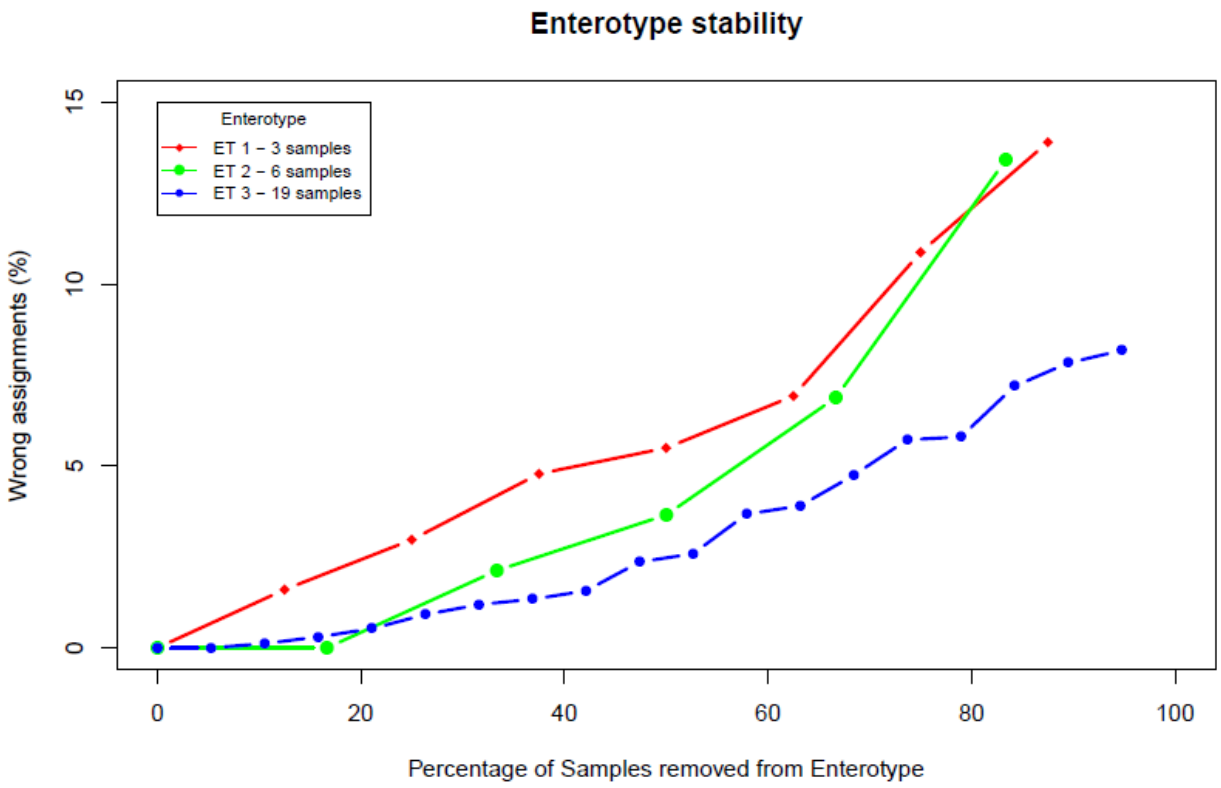


Supplementary Figure 22. Pyrosequencing-based metagenomes from two Danish samples (DA-AD-1T and DA-AD-3T) cluster in the same enterotypes along with the Sanger-based metagenomes from the same samples (DA-AD-1 and DA-AD-3).


a
Escherichia coli str. K-12 substr. MG1655 (Escherichia/Shigella, 1550:1.0)
Akkermansia muciniphila ATCC BAA-835 (Akkermansia, 1500:1.0)
Alistipes
Candidatus Azobacteroides pseudotrichonymphae CFP2 (Porphyromonadaceae, 1500:0.99)
Porphyromonas
Parabacteroides
Bacteroides
Prevotella
Clostridium subclade
Anaerofustis stercorihominis DSM 17244 (Anaerofustis, 1500:1.0)
Eubacterium saphenum ATCC 49989 (Eubacterium, 1500:1.0)
Alkaliphilus metalliredigens QYMF (Alkaliphilus, 1500:1.0)
Alkaliphilus oremlandii OhILAs (Alkaliphilus, 1500:1.0)
Clostridiales genomosp. BVAB3 str. UPII9-5 (Acetivibrio, 1500:0.98)
Clostridium cellulolyticum H10 (Acetivibrio, 1450:1.0)
Clostridium thermocellum ATCC 27405 (Acetivibrio, 1500:1.0)
Erysipelotrichaceae bacterium 5_2_54FAA (Eubacterium, 1550:0.9)
Eubacterium dolichum DSM 3991 (Eubacterium, 1500:1.0)
Eubacterium biforme DSM 3989 (Eubacterium, 1550:1.0)
Eubacterium cylindroides T2-87 (Eubacterium, 1500:1.0)
Subdoligranulum variabile DSM 15176 (Subdoligranulum, 1100:1.0)
Faecalibacterium prausnitzii A2-165 (Faecalibacterium, 1500:1.0)
Faecalibacterium prausnitzii SL3/3 (Faecalibacterium, 950:1.0)
Faecalibacterium prausnitzii M21/2 (Faecalibacterium, 1500:1.0)
Ruminococcus bromii L2-63 (Ruminococcus, 1300:1.0)
Clostridium leptum DSM 753 (Ruminococcaceae, 1450:1.0)
Anaerotruncus colihominis DSM 17241 (Anaerotruncus, 1500:1.0)
Clostridium methylpentosum DSM 5476 (Ruminococcaceae, 1500:1.0)
Ruminococcus sp. 18P13 (Ruminococcus, 1450:1.0)
Eubacterium siraeum V10Sc8a (Ruminococcaceae, 1500:1.0)
Eubacterium siraeum DSM 15702 (Ruminococcaceae, 1500:1.0)
Abiotrophia defectiva ATCC 49176 (Catonella, 1500:1.0)
Clostridium phytofermentans ISDg (Anaerosporobacter, 1550:0.88)
Clostridium sp. L2-50 (Coprococcus, 1500:1.0)
Coprococcus eutactus ATCC 27759 (Coprococcus, 1500:1.0)
Coprococcus sp. ART55/1 (Coprococcus, 950:1.0)
Eubacterium hallii DSM 3353 (Lachnospiraceae, 1400:1.0)
Anaerostipes caccae DSM 14662 (Anaerostipes, 1500:1.0)
Clostridium sp. SS2/1 (Anaerostipes, 1500:0.84)
butyrate-producing bacterium SSC/2 (Anaerostipes, 1500:0.85)
Coprococcus catus GD/7 (Coprococcus, 950:1.0)
Shuttleworthia satelles DSM 14600 (Shuttleworthia, 1500:1.0)
Marvinbryantia formatexigens DSM 14469 (Marvinbryantia, 1500:1.0)
Butyrivibrio crossotus DSM 2876 (Lachnospiraceae, 1500:0.96)
Roseburia intestinalis L1-82 (Roseburia, 1050:1.0)
Roseburia intestinalis M50/1 (Roseburia, 1000:1.0)
Roseburia intestinalis XB6B4 (Roseburia, 1500:1.0)
Eubacterium rectale DSM 17629 (Roseburia, 1500:1.0)
Eubacterium rectale M104/1 (Roseburia, 1500:1.0)
Eubacterium rectale ATCC 33656 (Roseburia, 1500:1.0)
Eubacterium ventriosum ATCC 27560 (Lachnospiraceae, 1500:1.0)
Bacteroides pectinophilus ATCC 43243 (Lachnospiraceae, 1500:1.0)
Eubacterium eligens ATCC 27750 (Lachnospiraceae, 1500:1.0)
Oribacterium sp. oral taxon 078 str. F0262 (Oribacterium, 1500:1.0)
Oribacterium sinus F0268 (Oribacterium, 1400:1.0)
Clostridium sp. M62/1 (Lachnospiraceae, 1500:0.85)
butyrate-producing bacterium SM4/1 (Lachnospiraceae, 1400:0.91)
Clostridium cf. saccharolyticum K10 (Lachnospiraceae, 1000:0.97)
Blautia hansenii DSM 20583 (Blautia, 1150:1.0)
Blautia hydrogenotrophica DSM 10507 (Blautia, 1500:1.0)
Ruminococcus sp. SR1/5 (Blautia, 750:1.0)
Ruminococcus sp. 5_1_39BFAA (Blautia, 1500:1.0)
Ruminococcus obeum ATCC 29174 (Blautia, 1500:1.0)
Ruminococcus obeum A2-162 (Blautia, 1050:1.0)
Ruminococcus torques ATCC 27756 (Lachnospiraceae, 1500:0.95)
Ruminococcus torques L2-14 (Lachnospiraceae, 1500:0.99)
Ruminococcus lactaris ATCC 29176 (Lachnospiraceae, 500:0.81)
Clostridium nexile DSM 1787 (Lachnospiraceae, 550:1.0)
Coprococcus comes ATCC 27758 (Coprococcus, 1500:1.0)
Clostridium scindens ATCC 35704 (Dorea, 1500:0.99)
Dorea longicatena DSM 13814 (Dorea, 1500:1.0)
0.1
Strains with placement uncertainty
Ruminococcus strains placed in Lachnospiraceae family
Strains placed in different genera than original genus names
Strains without placement uncertainty
Ruminococcus strains placed in Ruminococcus genus
Strains placed in their original genus

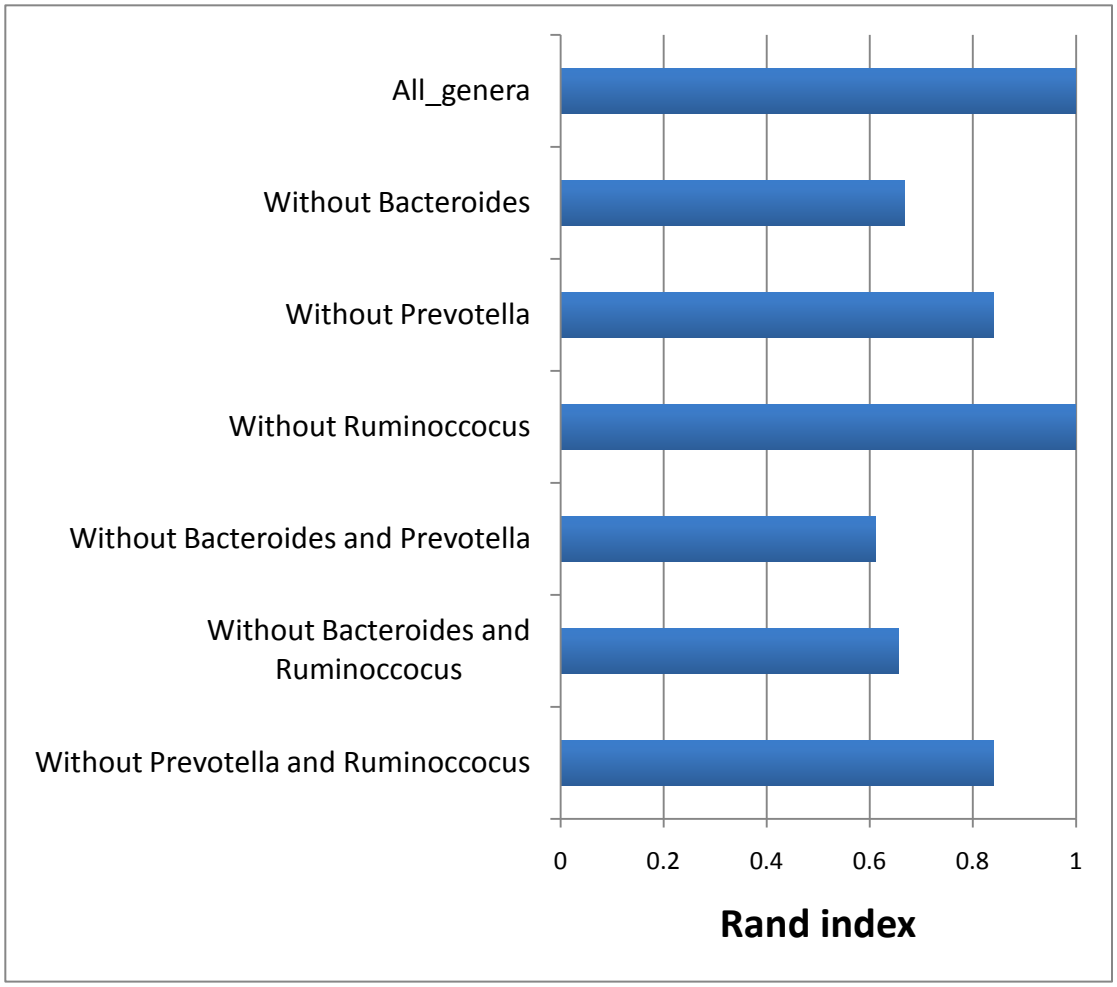

b
0.1
Escherichia coli str. K–12 substr. MG1655 Escherichia/Shigella, 1550:1.0)
Akkermansia muciniphila ATCC BAA–835 Akkermansia, 1500:1.0)
Alistipes shahii WAL 8301 (Alistipes, 1500:1.0)
Alistipes putredinis DSM 17216 (Alistipes, 1500:1.0)
Candidatus Azobacteroides pseudotrichonymphae CFP2 (Porphyromonadaceae, 1500:0.99)
Porphyromonas gingivalis W83 (Porphyromonas, 1450:1.0)
Porphyromonas gingivalis ATCC 33277 (Porphyromonas, 1450:1.0)
Porphyromonas uenonis 60–3 (Porphyromonas, 1450:1.0)
Porphyromonas endodontalis ATCC 35406 (Porphyromonas, 1450:1.0)
Parabacteroides merdae ATCC 43184 (Parabacteroides, 1500:1.0)
Parabacteroides johnsonii DSM 18315 (Parabacteroides, 1500:1.0)
Parabacteroides distasonis ATCC 8503 (Parabacteroides, 1500:1.0)
Bacteroides sp. 2_1_33B (Parabacteroides, 900:1.0)
Parabacteroides sp. D13 (Parabacteroides, 1500:1.0)
Prevotella tannerae ATCC 51259 (Prevotella, 1500:1.0)
Prevotella ruminicola 23 (Prevotella, 1500:1.0)
Prevotella sp. oral taxon 472 str. F0039 (Prevotella, 1300:1.0)
Prevotella sp. oral taxon 472 str. F0295 (Prevotella, 1450:1.0)
Prevotella sp. oral taxon 472 str. F0109 (Prevotella, 1450:1.0)
Prevotella buccalis ATCC 35310 (Prevotella, 1450:1.0)
Prevotella buccae D17 (Prevotella, 1450:1.0)
Prevotella oris F0302 (Prevotella, 1500:1.0)
Prevotella copri DSM 18205 (Prevotella, 1500:1.0)
Prevotella bergensis DSM 17361 (Prevotella, 1500:1.0)
Prevotella bivia JCVIHMP010 (Prevotella, 1450:1.0)
Prevotella veroralis F0319 (Prevotella, 1500:1.0)
Prevotella melaninogenica D18 (Prevotella, 1450:1.0)
Prevotella melaninogenica ATCC 25845 (Prevotella, 1500:1.0)
Bacteroides coprocola DSM 17136 (Bacteroides, 1500:1.0)
Bacteroides plebeius DSM 17135 (Bacteroides, 1300:1.0)
Bacteroides coprophilus DSM 18228 (Bacteroides, 1500:1.0)
Bacteroides vulgatus PC510 (Bacteroides, 1500:1.0)
Bacteroides sp_4_3_47FAA (Bacteroides, 1500:1.0)
Bacteroides vulgatus ATCC 8482 (Bacteroides, 1500:1.0)
Bacteroides sp_3_1_33FAA (Bacteroides, 1500:1.0)
Bacteroides dorei DSM 17855 (Bacteroides, 1500:1.0)
Bacteroides sp_9_1_42FAA (Bacteroides, 1500:1.0)
Bacteroides dorei 5_1_36/D4 (Bacteroides, 1500:1.0)
Bacteroides cellulosilyticus DSM 14838 (Bacteroides, 1500:1.0)
Bacteroides sp_D20 (Bacteroides, 1500:1.0)
Bacteroides uniformis ATCC 8492 (Bacteroides, 1500:1.0)
Bacteroides eggerthii DSM 20697 (Bacteroides, 1500:1.0)
Bacteroides sp_3_2_5 (Bacteroides, 1450:1.0)
Bacteroides fragilis NCTC 9343 (Bacteroides, 1550:1.0)
Bacteroides sp_2_1_16 (Bacteroides, 1150:1.0)
Bacteroides fragilis YCH46 (Bacteroides, 1550:1.0)
Bacteroides thetaiotaomicron VPI–5482 (Bacteroides, 1500:1.0)
Bacteroides sp_1_1_6 (Bacteroides, 1450:1.0)
Bacteroides caccae ATCC 43185 (Bacteroides, 1450:1.0)
Bacteroides finegoldii DSM 17565 (Bacteroides, 1300:1.0)
Bacteroides ovatus SD CMC 3f (Bacteroides, 1050:1.0)
Bacteroides ovatus ATCC 8483 (Bacteroides, 1450:1.0)
Bacteroides sp_2_2_4 (Bacteroides, 1450:1.0)
Bacteroides xylanisolvens SD CC 1b (Bacteroides, 1450:1.0)
Bacteroides sp_D1 (Bacteroides, 1450:1.0)
Bacteroides ovatus SD CC 2a (Bacteroides, 1450:1.0)
Bacteroides sp_2_1_22 (Bacteroides, 1450:1.0)
Anaerofustis stercorihominis DSM 17244 (Anaerofustis, 1500:1.0)
Eubacterium
Ruminococcaceae
Lachnospiraceae
Eubacterium saphenum ATCC 49989 (Eubacterium, 1500:1.0)
Alkaliphilus
Clostridium
Acetivibrio
Strains with placement uncertainty
Strains placed in different genera than original genus names
Strains without placement uncertainty
Strains placed in their original genus

Supplementary Figure 23. Taxonomic uncertainties in Clostridiales observed through Maximum likelihood tree of 129 strains from Firmicutes (a) and Bacteroidetes (b), constructed from 40 single copy marker genes[51,85] using RAxML[86] v 7.04. Phylogenetic classification of these strains by RDP classifier based on 16S rRNA gene sequences are given in parenthesis, along with the confidence threshold and approximate length (minimum length 500bp, minimum confidence threshold 0.8). Bootstrap proportion values from 100 replicates are shown on each branch. *E. coli* and *A. muciniphila* were included as outgroups. For clarity, Bacteroidetes subclades are collapsed in (a) and Firmicutes subclades are collapsed in (b). 7 out of 9 *Ruminococcus* (driver of enterotype 3 in Sanger dataset, Fig. 2a) strains are placed in Lachnospiraceae, very close to *Blautia* (driver of enterotype 3 in 16S dataset, Fig. 2c) and other unclassified Lachnospiraceae (driver of enterotype 3 in Illumina dataset, Fig. 2b). 22 strains out of 40 in Lachnospiraceae are placed in different genera than their original genus names suggest. Similarly, four strains out of 11 in Ruminococcaceae are placed in different genera. In contrast, only one strain out of 57 in Bacteroidetes is placed in different genus than its genus name suggests, showing consistency of taxonomy in this clade that contains the drivers for enterotypes 1 and 2 in all three datasets.

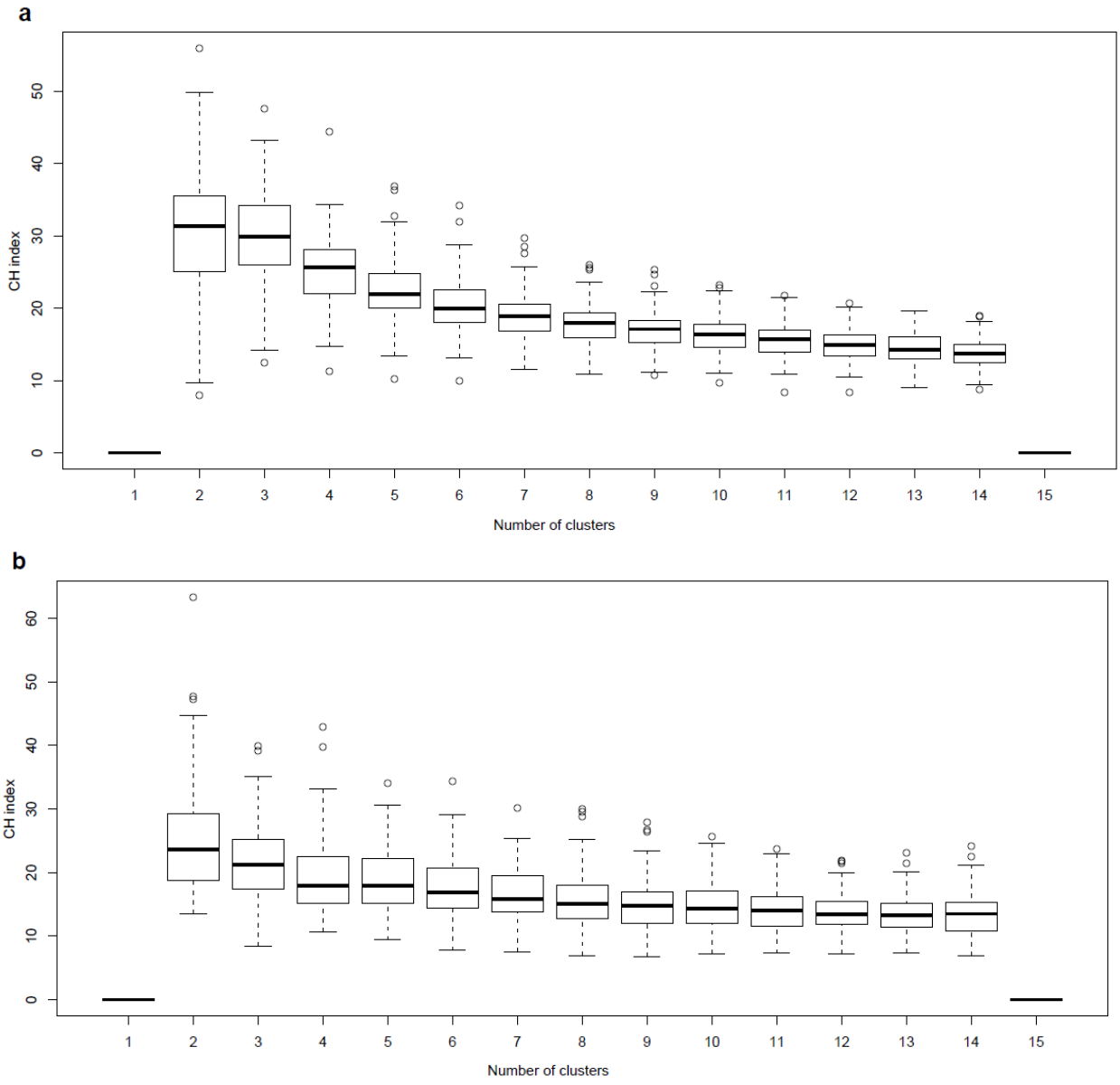


Supplementary Figure 24. Effect of removing samples from an enterotype on the clustering behavior.

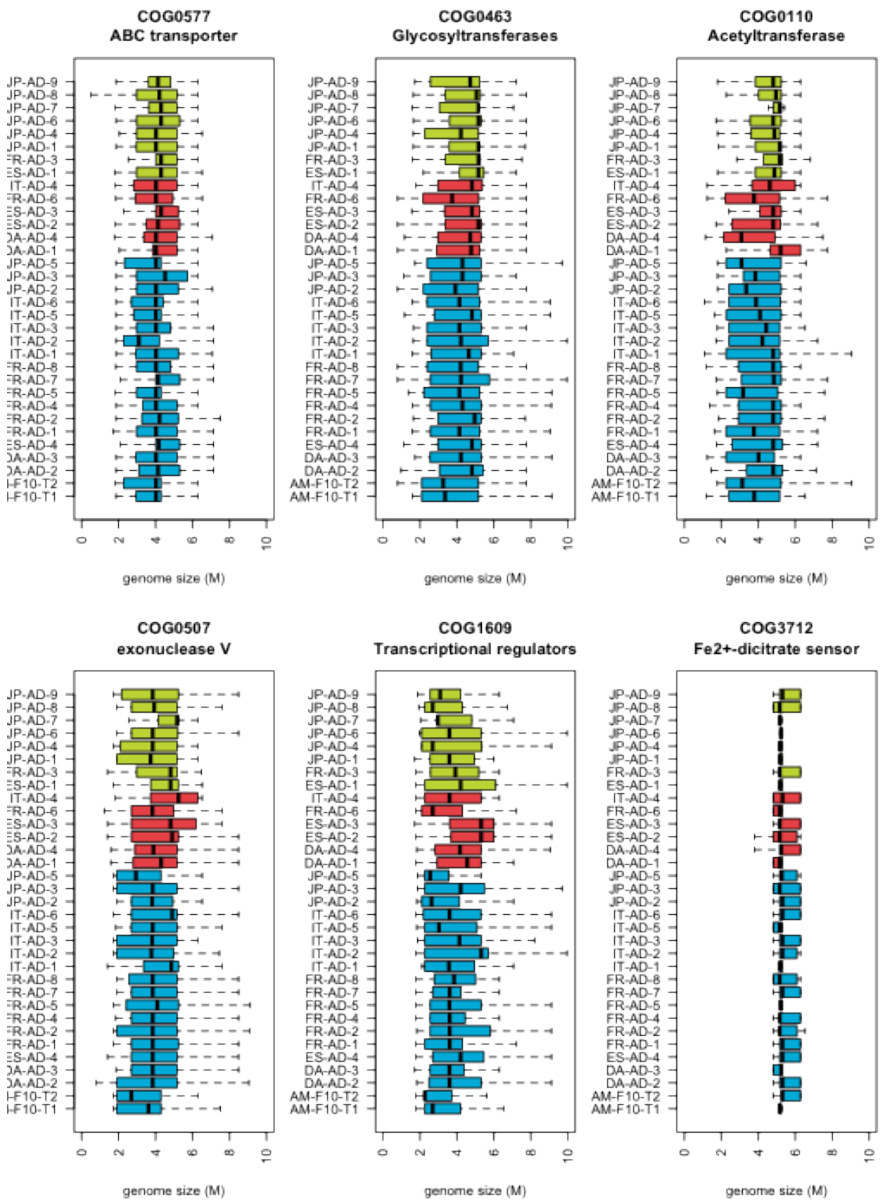
We randomly removed a certain percentage of samples (x-axis) from each enterotype 1000 times and re-clustered, and estimated the average misassignment percentage (y-axis) over all three enterotypes. Clusters generally stay intact – even when half the samples from a cluster, less than 6% of the samples are assigned to the wrong cluster.



Supplementary Figure 25. Effects of removing major drivers of enterotypes. Enterotype clusters can be recovered when two out of the three major drivers (except *Bacteroides*) are removed either individually or altogether.



Supplementary Figure 26. Estimating the optimal number of clusters using half the samples. (a) 16S dataset and (b) Illumina dataset. Using only half the samples in each enterotype and estimating the optimal number of clusters using CH-index shows that two clusters are preferred over 3 clusters. CH index curves for the two datasets from Supplementary Figure 3 are not flatter than the curves here. Boxplots correspond to the CH index values from 1000 replicates of 50% subsampled datasets.



Supplementary Figure 27. Effect of genome size in abundance of orthologous groups.

We checked the genome size distributions for 2 COGs which are over-represented in each enterotype (left:1, center:2, right:3). Horizontal axis represents the estimated size of genome encoding the sequence read assigned to the COG. All COGs have a relatively stable average, suggesting that genome size does not have a major effect on functional abundance.