Supplementary Information

Supplementary Results

Taxonomic assignments of metagenomic reads: assessing the value of reference genomes - The International Human Microbiome Program has emphasized the importance of sequencing the genomes of a panel of reference microbial strains[1]. One fundamental parameter that governs the utility of reference genomes is the ability to accurately assign fragmentary reads from metagenomic datasets to these genomes. Therefore, we compared the filtered pyrosequencer reads to a custom database of 44 human gut-associated bacterial and archaeal genomes (**Supplementary Fig. 7** and Supplementary Information) using BLASTX, and validated these assignments independently against the NCBI non-redundant protein database. The relative abundance of sequences from the 18 individual microbiome datasets assigned to each reference genome was highly variable (see **Supplementary Fig. 7;** R^2=0.26±0.02 for all pairwise comparisons of taxonomic profiles), consistent with the considerable heterogeneity in microbial community structure among the fecal microbiomes that we had observed from sequencing 16S rRNA gene amplicons.

Our custom database of 44 reference genomes included 23 Firmicutes but only 14 Bacteroidetes. Since the Firmicutes dominate the gut microbiotas of our subjects (**Supplementary Fig. 6**), and the reference genome database, we might expect that reads assigned to Firmicutes would match the reference genomes more closely than reads assigned to Bacteroidetes. The opposite was true: on average, 46.3±2.6% of the pyrosequencing reads assigned to Bacteroidetes matched the reference genomes at 100% identity, as opposed to 16.7±1.1% of the reads assigned to Firmicutes ($p<10^{-4}$, Mann

Whitney; **Supplementary Figs. 8,9**). This observation underscores the high level of diversity within the gut-associated Firmicutes, indicates that the readily culturable sequenced gut Firmicutes are not closely related to the abundant gut genomes present in these 18 microbiomes, and suggests that future reference microbial genome sequencing efforts should be directed towards representatives of this dominant bacterial phylum.

Supplementary Fig. 10 summarizes the relative abundance of the major bacterial phyla present in these 18 microbiomes, as defined by six different approaches (sequencing full-length, V2 and V6 amplicons; BLAST comparisons of shotgun pyrosequencer reads with the NCBI non-redundant and the custom 44 gut genome databases, plus analysis of microbiome-derived 16S rRNA gene fragments). Pairwise comparisons of relative abundance data from 16S rRNA gene fragments generated from shotgun sequencing reads correlate most closely with V2 PCR data (see **Supplementary Fig. 10** and the next section).

Cross-comparison of taxonomic assignments - A frequently reported result from any 16S rRNA gene sequence-based survey is the relative abundance of bacterial phyla[6-8]. Given the broad nature of these phyla and the fact that relatively few phyla dominate the human distal gut microbiota, one might expect the relative abundance of each phylum to be consistent regardless of the amplification and sequencing methods used. However, we observed differences between methods (**Supplementary Fig. 10A-E**). Relative to the sampled gut microbiomes (defined by pyrosequencing of total community DNA), the full-length, V2, and V6 16S rRNA gene datasets were all significantly depleted for Bacteroidetes (paired Student's t-test, $p<0.001$), and significantly enriched for Firmicutes ($p<0.01$). One possible explanation for these differences is that the Bacteroidetes reference genomes are more closely related to those

in the microbiomes than the Firmicutes reference genomes, thereby inflating estimates of the relative abundance of this phylum (**Supplementary Fig. 8**). To address this potential confounding factor, we identified 16S rRNA gene fragments from all 18 microbiome datasets (see Supplementary Methods below) and classified them taxonomically. The results of this analysis confirmed that the three PCR-based methods underestimate the relative abundance of the Bacteroidetes (**Supplementary Fig. 10F**). However, results obtained from shotgun sequencing 16S rRNA gene fragments and PCR amplification of the V2 region showed the strongest correlation (**Supplementary Fig. 10G**).

Phylotype sampling model - We developed a sampling model that allows us to place bounds on the maximum abundance of any phylotype found across all samples. The principle here is that if a given phylotype made up not less than some proportion p of the fecal microbiome of all samples, we can calculate (i) the number of samples of a given size that we would expect to lack that phylotype due to sampling error, and (ii) the probability that we would observe an actual proportion (p-hat) as low as the minimum abundance observed in any sample.

The probability P of failing to observe a given microbe at proportion p in a sample of size n is given by Poisson statistics as simply e^{-pn}. For equal sample sizes, we can therefore calculate the probability of observing the phylotype in at least k samples using binomial sampling with Pr(success) = (1-P). We can then use the inverse binomial to ask what value of P, and therefore of p, gives a specified probability (say, 5%) of observing a given phylotype in as few samples as actually observed for the most abundant phylotype. This calculation yields an upper bound for p (i.e., the value of p at which we can reject the idea that we would have seen the phylotype in as few samples as actually observed at the 95% confidence level).

For unequal sizes, there is no analytical solution to the equivalent of the binomial in which Pr(success) differs for each trial (M. Lladser, personal communication). We must therefore solve for p using numerical optimization. Because the function relating p and the probability of observing the phylotype in at least a given number of samples is monotonic, we can use a bisection search (bounded by p=0 and p=1) to find the appropriate value of p for a desired confidence level. In practice, we calculate P for each sample, choose a vector of random numbers between 0 and 1, and count the number of times the random number at a given position was less than P. Repeating this procedure for a fixed number of iterations (100,000 for the reported values) gives sufficiently smooth values to approximate the monotonic function and to allow the bisection search to converge on the same value of p to three significant figures across repeated trials.

In the case where a phylotype is found in all samples, we can use a similar procedure to identify the maximum value of p consistent with the observed minimum abundance of the phylotype whose minimum abundance across all samples is highest. In this case, instead of calculating the fraction of samples in which the phylotype is absent, we can (i) use binomial sampling to randomly sample the number of observed counts of a phylotype given the parametric value of p and the sample size of each sample, (ii) measure the minimum abundance across all samples, and (iii) compare this minimum abundance to the minimum abundance actually observed. Again, an analytical solution using extreme-value statistics is possible if sample sizes are equal, but the solution must be obtained by numerical methods (in this case, the same type of bisection search used above). The sampling model was implemented in Python using PyCogent[34].

Our sampling model allows us to ask what level of abundance the most abundant phylotype could have in every individual before its absence from or limited

representation in some samples becomes surprising. For example, with 1,000 sequences/samples, we would be very surprised if we failed to recover a shared phylotype present at 50% abundance in each of 30 samples, but would not be surprised if we missed a phylotype at 0.00001% abundance.

Using this model we first analyzed the full-length 16S rRNA dataset. The most abundant 'species'-level phylotype in each sample made up 13% of that sample on average (range: 4.3%-45%), and the most abundant phylotype found across the combined dataset was found in 27 of the 30 fecal microbiotas (Bacteria; Firmicutes; Clostridia; Faecalibacterium). These data are consistent with no phylotype being present at more than 1.1% abundance in all samples.

The deeper pyrosequencing data confirmed this result. In the V6 dataset, using even sampling of 10,000 sequences/sample, the most abundant phylotype in each sample made up 12% of that sample on average (range: 5.0%-37%). The overall most abundant phylotype was found in all 33 samples (Bacteria; Firmicutes; Clostridia; Clostridiales; *Eubacterium rectale*). However, in some samples, this phylotype was present in frequencies as low as 0.01%.

1,000 randomly chosen V2 16S rRNA gene sequences per sample or all V2 sequences were analyzed using the sampling model. The most abundant phylotype in each sample made up 15% of that sample on average (range: 3.8%-47%; 7.6% on average using all sequences, range: 1.6%-26%). The phylotype observed in the most samples was present in 270 of 274 samples at this depth of coverage (Bacteria; Firmicutes; Clostridia; Clostridiales; Ruminococcus). The sampling model indicated that this frequency was consistent with a true abundance of no more than 0.53%.

These results were confirmed, with excellent agreement, by the V6 data: at 1,000 sequences/sample, the maximum abundance OTU is found in 32 of 33 samples, consistent with an abundance of no more than 0.66%. However, at a coverage depth of 10,000 sequences/sample, this OTU is found in all 33 samples but at a minimum observed abundance of 0.01%, consistent with a true abundance of no more than 0.1%. Using all the V6 data without controlling for sampling effort, the minimum observed abundance is consistent with a true abundance of no more than 0.07% (the estimate of the true abundance falls with increased sample size because it is less likely that the low frequency would be observed due to sampling error when more total sequences contribute to the result). Thus, we conclude, with 95% confidence, based on the even sampling used for the other analyses in this study (i.e., 1,000 sequences/sample from V2, 10,000 sequences/sample for V6) that the maximum abundance of any OTU across all samples cannot exceed the V2 result of 0.53%, although the true maximum abundance might be as much as an order of magnitude lower than this based on the greater depth of coverage in the V6 samples.

Sample Characteristics - The MOAFTS twin cohort, comprised of female like-sex twin pairs, were identified from Missouri birth records over the period 1994-1999, when the twins had a median age of 15 years. A total of 350 twins from the larger MOAFTS cohort completed screening interviews for the present study. We were able to take advantage of the wave five interview of the MOAFTS twin cohort (which has 90% retention of wave four participants) to identify pairs most likely to meet study criteria. Eligibility was then confirmed at screening interview.

Supplementary Table 15 summarizes BMI category by ethnicity for the entire MOAFTS wave 5 cohort, based on 3326 twins with complete data on height and weight.

DZ twins had a significantly higher mean BMI than MZ twins [25.8±6.5 vs. 24.8±5.9, $p<0.001$, mean±sd], and a higher prevalence of overweight (22.8 vs 20.9%) and obese (20.7 vs 16.1%; $\chi^2= 31.6$, $p<0.001$). This may reflect a higher dizygotic twinning rate among obese women (MZ twinning occurs randomly[35]). BMI was more highly correlated in MZ twins than in DZ twins, both in EA pairs (rMZ=0.80, rDZ=0.48) and in AA pairs (rMZ=0.73, rDZ=0.26), and this remained true when analysis was restricted to pairs concordant for obesity (EA: rMZ=0.61, rDZ=0.27; AA rMZ=0.62, rDZ=-0.11) or concordant for leanness (EA: rMZ=0.43, rDZ=0.14; AA: rMZ=0.55, rDZ=0.39). After age-adjustment, quantitative genetic modeling yielded an estimated additive genetic variance for BMI of 68% (95% Confidence Interval [CI]: 57-79%), shared environmental variance of 14% (95% CI: 2-24%), and non-shared environmental variance of 14% (95%CI: 17-21%). Twin pair correlations and heritability estimates are similar to those reported in studies of other twin cohorts of similar age[36]. Data from the Behavioral Risk Factor Surveillance System[37] for Missouri women of comparable age in 2006 yield higher rates of overweight and obesity in EA women (23.8% overweight and 25% obese) compared to rates observed in MOAFTS (19.6% overweight EA, 14.8% obese EA).

Lean and obese women selected for inclusion in the biospecimen collection project were representative of the entire cohort of lean and obese MOAFTS twins in terms of parity (nulliparous/parous), educational attainment (more than high school education/high school education or less) and marital status (married or living with someone as married/not married; $p>0.05$ for all comparisons), whether EA or AA. Obese EA and AA women providing biospecimens had a mean BMI at wave 5 of 36.9±4.7 and 39.8±7.8 (mean±sd), respectively, compared with a mean BMI of 21.4±1.5 (EA) and 21.1±1.2 (AA) among lean women. EA twins were selected as being stably lean across all waves

of data (self-reported BMI of 18.5-24.9 kg/m^2; baseline at median age 15, one-year follow-up, 5-year follow-up and seven year follow-up).

Supplementary Methods

Full-length 16S rRNA sequence-based surveys - Five replicate PCR reactions were performed for each fecal DNA sample. Each 25 μl reaction contained 100 ng of gel purified DNA (Qiaquick, Qiagen), 10 mM Tris (pH 8.3), 50 mM KCl, 2 mM $MgSO_4$, 0.16 μM dNTPs, 0.4 μM of the bacteria-specific primer 8F (5'-AGAGTTTGATCCTGGCTCAG-3'), 0.4 μM of the universal primer 1391R (5'-GACGGGCGGTGWGTRCA-3'), 0.4 M betaine, and 3 units of Taq polymerase (Invitrogen). Cycling conditions were 94°C for 2 min, followed by 25 cycles of 94°C for 1 min, 55°C for 45 sec, and 72°C for 2 min. Replicate PCRs were pooled and concentrated (Millipore; Montage PCR filter columns). Full-length 16S rRNA gene amplicons (1.3kb) were then gel-purified using the Qiaquick kit (Qiagen), subcloned into TOPO TA pCR4.0 (Invitrogen), and the ligated DNA transformed into *E. coli* TOP10 (Invitrogen). For each sample, 384 colonies containing cloned amplicons were processed for sequencing. Plasmid inserts were sequenced bi-directionally using vector-specific primers plus the internal primer 907R (5'-CCGTCAATTCCTTTRAGTTT-3').

16S rRNA gene sequences were edited and assembled into consensus sequences using the PHRED and PHRAP software packages within the Xplorseq program[6]. Sequences that did not assemble were discarded and bases with PHRED quality scores <20 were trimmed. Sequences were checked for chimeras using Bellerophon version 3 with the default parameters[38]. Alignments for reference genome 16S rRNA gene sequences were manually edited in ARB[39].

V2 16S rRNA sequence-based surveys - Four replicate PCR reactions were performed on the same fecal DNA samples used above. Each 20 μl reaction contained 100 ng of gel purified DNA (Qiaquick, Qiagen), 8 μl 2.5X HotMaster PCR Mix

(Eppendorf), 0.3 μM of a modified primer 8F [5'-GCCTTGCCAGCCCGCTCAG-**TC***AGAGTTTGATCCTGGCTCAG*-3'; composite of 454 primer B (underlined), linker nucleotides (TC), and the universal bacterial primer 8F (italics)], and 0.3 μM of a modified primer 338R [5'-GCCTCCCTCGCGCCATCAGNNNNNNNNNNNN**CA-***TGCTGCCTCCCGTAGGAGT*-3'; 454 Life Sciences primer A (underlined), a unique 12 base barcode[13] (Ns), linker nucleotides (CA), and the broad-range bacterial primer 338R (italics)]. Cycling conditions were 95°C for 2 min, followed by 30 cycles of 95°C for 20 sec, 52°C for 20 sec, and 65°C for 1 min. Replicate PCRs were pooled and amplicons purified using Ampure magnetic purification beads (Agencourt).

PCR products were quantified with the bisbenzimide H assay. An aliquot of each PCR product was incubated for 5 min at room temperature in TNE reagent [10 mM Trizma HCl pH 8.1, 100mM NaCl, 1 mM EDTA, and 50 ng/ml freshly prepared bisbenzimide H (Sigma)]. Samples were read on a flurometer or plate reader (excitation at 365nm, emission at 460nm) relative to a standard curve constructed using *E. coli* DNA (Sigma). Multiple pools, each containing equimolar amounts of PCR products, were assembled for 454 FLX amplicon pyrosequencing (n=33-100 barcoded[13] samples/pool). Technical replicates were analyzed from selected representatives of each pool across four different sequencing centers; results were highly reproducible, discriminating between individuals and between samples from the same individual over time (**Supplementary Fig. 5**).

V6 16S rRNA sequence-based surveys – PCR reactions targeting the V6 region were performed on the same fecal DNA samples used above. Each 32 μl reaction contained 100 ng of gel purified DNA (Qiaquick, Qiagen), PCR buffer (PurePeak DNA polymerization mix, Thermo-Fisher), 0.625 mM PurePeak dNTPs (Thermo-Scientific),

0.625 μM Fusion Primer A, 0.625 μM Fusion Primer B, and 5U Pfu polymerase (Stratagene). The primer set included 5 forward primers (Fusion A) and 4 reverse primers (Fusion B) fused to the 454 Life Sciences adaptors A and B, respectively[40]. Cycling conditions were 94°C for 3 min, followed by 30 cycles of 94°C for 30 sec, 57°C for 45 sec, and 72°C for 1 min, with a final extension period of 72°C for 2 min. PCR products were purified with MinElute columns (Qiagen), and DNA was quantitated using a Bioanalyzer (Agilent) and the PicoGreen assay (Invitrogen). Two pools of PCR products were constructed for 454 FLX amplicon pyrosequencing, composed of 18 and 20 samples, respectively [the second run contained 3 samples from the V2 region and 3 technical replicates; one additional sample (TS30) was sequenced in a third run, bringing the total number of V6 samples processed to 33]. Since technical replicates were highly reproducible (see above), datasets for a given individual's biospecimen were pooled for all subsequent analyses. Any sequences that did not have an exact match to the proximal primer or that contained one or more ambiguous bases were removed as 'low quality'. The proximal primer and any 'fuzzy' matches (identified with BLAST and fuzznuc) to the distal primer were then trimmed from the sequences. Finally, any trimmed sequences shorter than 50 nt were also removed as low-quality[41].

Database searches and metabolic reconstructions - The distributions of taxa, genes, orthologs, metabolic pathways, and high-level gene categories were tallied based on the corresponding annotation of the best-BLAST-hit sequence found in each reference database (BLASTX e-value<10^{-5}, %identity>50, and score>50). For KEGG analysis, the closest matching gene with an annotation was used, since many genes in the database remain unannotated, including all KEGG orthologous groups (KOs) assigned to genes with an identical e-value (commands -e 0.00001 -m 9 -b 100 were used to run NCBI

BLASTX). Custom Perl scripts were used for all KEGG[17] (v44), STRING[18] (v7), and NCBI NR (v11/24/07) analyses. Selected genes from the recently sequenced reference genomes were manually annotated using NCBI-BLASTP searches against the KEGG, STRING, and NCBI NR database. The 44 reference genome database includes predicted proteins from draft or complete assemblies of *Alistipes putredinis DSM17216, Bacteroides WH2, Bacteroides thetaiotaomicron 3731, Bacteroides thetaiotaomicron 7330, Bacteroides thetaiotaomicron VPI5482, Bacteroides fragilis NCTC9343/YCH46, Bacteroides caccae ATCC43185, Parabacteroides distasonis ATCC8503, Bacteroides ovatus ATCC8483, Bacteroides stercoris ATCC43183, Bacteroides uniformis ATCC8492, Bacteroides vulgatus ATCC8482, Parabacteroides merdae ATCC43184, Anaerostipes caccae DSM14662, Anaerotruncus colihominis DSM17241, Anaerofustis stercorihominis DSM17244, Bacteroides capillosus ATCC29799, Clostridium bartlettii DSM16795, Clostridium bolteae ATCCBAA-613, Coprococcus eutactus ATCC27759, Clostridium leptum DSM753, Clostridium ramosum DSM1402, Clostridium scindens ATCC35704, Clostridium sp. L2-50, Clostridium spiroforme DSM1552, Dorea longicatena DSM13814, Eubacterium dolichum DSM3991, Eubacterium eligens ATCC27750, Eubacterium rectale ATCC33656, Eubacterium siraeum DSM15702, Eubacterium ventriosum ATCC27560, Faecalibacterium prausnitzii M212, Peptostreptococcus micros ATCC33270, Ruminococcus gnavus ATCC29149, Ruminococcus obeum ATCC29174, Ruminococcus torques ATCC27756, Collinsella aerofaciens ATCC25986, Bifidobacterium adolescentis, Bifidobacterium longum DJO10A/NCC2705, Escherichia coli K12, Methanobrevibacter smithii ATCC35061,* and *Methanobrevibacter stadtmanae DSM3091* (see http://genome.wustl.edu/pub/ and NCBI GenBank). Draft assemblies of *Clostridium sp. SS2-1* and *Clostridium symbiosum*

ATCC14940 were also used for functional clustering and diversity analyses (http://genome.wustl.edu/pub/). Coverage plots (percent identity plots) were generated using nucmer and mummerplot (part of the MUMmer v3.19 package), and default parameters[42]. Full-length 16S sequences were obtained for each reference genome, likelihood parameters were determined using Modeltest[43], and a maximum-likelihood tree was generated using PAUP[44].

We validated our annotations with simulated datasets (**Supplementary Fig. 11**). To do so, the frequency of annotated genes from the KEGG database[17] (v44) was first tallied across the aggregate human gut microbiomes (n=18 datasets). The 1,000 most frequent microbial genes were then used to generate 'simulated reads' between 50 and 500 nt long. The simulated reads were subsequently annotated (BLASTX against the KEGG database), with self-hits excluded. As shown previously[45], this analysis revealed a low rate of false positives (i.e. high precision), but using very short sequences (e.g. 50-100 nt) increased the rate of false negatives (lower sensitivity) (**Supplementary Fig. 11**). Given the increased read-length of our libraries relative to 454 GS20 pyrosequencing data[3,22,45], simulated reads with an average length comparable to our data (200-250 nt), demonstrated robust assignments with an e-value$<10^{-5}$, %identity>50, and/or bit-score>50. Using all three cutoffs, sequences 200 nt in length returned 81.5% of the correct assignments, with a precision of 0.93 and sensitivity of 0.88, similar to what was observed by re-annotating the original full-length gene sequences after ignoring self-hits. The KEGG cutoff criteria were also applied to BLASTX analysis results for STRING-based predictions[18], given the similar size of the databases.

ABI 3730xl capillary sequencing reads from 9 previously published adult human gut microbiomes were obtained from the NCBI TraceArchive[20,21]. The full dataset from

each sample was annotated by BLASTX comparisons against the KEGG[17] and STRING[18] database (see above; BLASTX e-value<10^{-5}, %identity>50, and score>50). To allow quantitative comparisons between these datasets and our pyrosequencing data, we first extracted all forward sequencing reads and then generated one 'simulated pyrosequencer read' from each longer capillary read. Nucleotides spanning positions 100 to 322 were used from all capillary reads of suitable length, to avoid low quality regions that commonly occur at the beginning and end of the reads. These simulated reads were then annotated as described above.

16S rRNA gene fragments were identified in each microbiome through BLASTN searches of the RDP database[46] (version 9.33; e-value<10^{-5}; Bit-score>50; %identity>50; alignment length≥100). Putative 16S rRNA gene fragments were then aligned using the NAST multi-aligner[47] with a minimum template length of 100 bases and minimum identity of 75%. Taxonomy was assessed after insertion into an ARB neighbor-joining tree[39].

Microbiomes were clustered based on their profiles after normalizing across all sampled communities (z-score), using the Pearson's correlation distance metric, followed by single-linkage hierarchical clustering in addition to Principal Components Analysis[19] (Cluster3.0). Results were visualized using the Treeview Java applet[48]. Functional diversity (Shannon index and evenness) was calculated using the number of assignments in each microbiome to each of the 254 pathways present in the KEGG database[17] (EstimateS 8.0; ref. 49). The maximum possible index is the natural log of the total number of pathways: ln(254) or 5.54. Shannon evenness was calculated by dividing the Shannon index for a given microbiome by the maximum possible index (scale of 0 to 1, with 1 representing a microbiome with all pathways found at an equal abundance)[22].

Results were compared to simulated metagenomic reads generated from the 36 recently sequenced reference human gut-derived Bacteroidetes and Firmicutes genomes (http://genome.wustl.edu/pub/organism/). Reads were produced by Readsim v0.10 (ref. 50), using the following options: -n 10000 -modlr normal -meanlr 223 -stdlr 0.3. The mean and standard deviation for length of the simulated reads was based on the observed read-length distribution of the 18 fecal microbiome datasets (**Supplementary Table 5**).

Supplementary References

35. Bulmer, M.G. *The Biology of Twinning in Man* (Clarendon, Oxford, 1970).

36. Schousboe, K. *et al.* Sex differences in heritability of BMI: a comparative study of results from twin studies in eight countries. *Twin Research* **6**, 409-421 (2003).

37. Centers for Disease Control and Prevention (CDC). *Behavioral Risk Factor Surveillance System Survey Data*. Atlanta, Georgia: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, 2006.

38. Huber, T., Faulkner, G., and Hugenholtz, P. Bellerophon: a program to detect chimeric sequences in multiple sequence alignments. *Bioinformatics* **20**, 2317-2319 (2004).

39. Ludwig, W. *et al.* ARB: a software environment for sequence data. *Nucleic Acids Res.* **32**, 1363-1371 (2004).

40. Huber, J.A. *et al.* Microbial population structures in the deep marine biosphere. *Science* **318**, 97-100 (2007).

41. Huse, S.M., Huber, J.A., Morrison, H.G., Sogin, M.L., and Welch, D.M. Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol.* **8**, R143 (2007).

42. Kurtz, S., *et al.* Versatile and open software for comparing large genomes. *Genome Biology* **5**, R12 (2004).

43. Posada, D., and Crandall, K.A. Modeltest: testing the model of DNA substitution. *Bioinformatics* **14**, 817–818 (1998).

44. Swofford, D.L. (2003). PAUP*. Phylogenetic Analysis Using Parsimony (* and Other Methods). Version 4 (Sunderland, Massachusetts: Sinauer Associates).

45. Mou, X., Sun, S., Edwards, R.A., Hodson, R.E., and Moran, M.A. Bacterial carbon processing by generalist species in the coastal ocean. *Nature* **451**, 708-711 (2008).

46. Cole, J.R., *et al.* The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. *Nucleic Acids Res.* **33**, D294-296 (2005).

47. DeSantis, T.Z., *et al.* NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA genes. *Nucleic Acids Res.* **34**, W394-399 (2006).

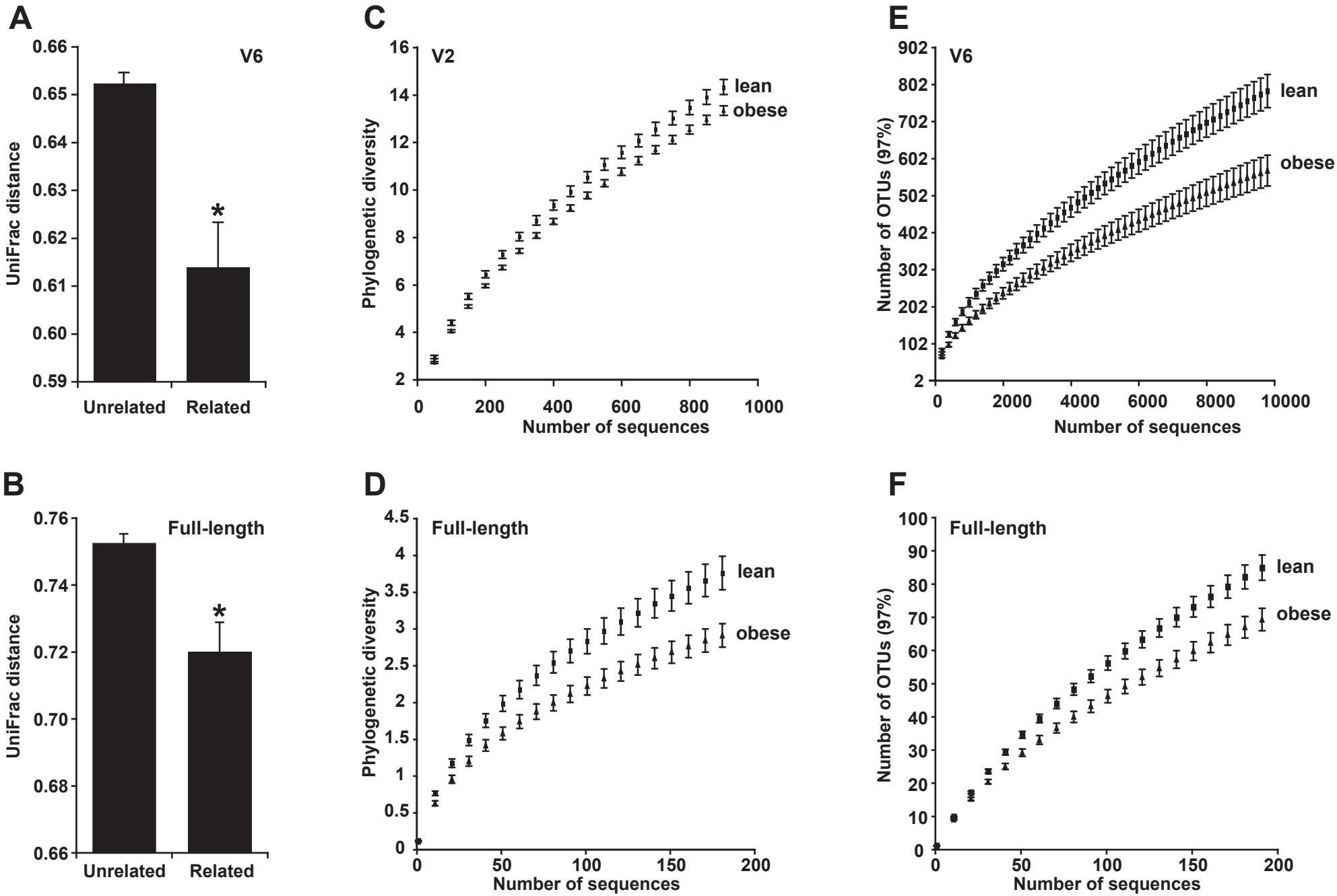
48. Saldanha, A.J. Java Treeview-extensible visualization of microarray data. *Bioinformatics* **20**, 3246-3248 (2004).

49. Colwell, R.K. EstimateS: Statistical estimation of species richness and shared species from samples. Version 7.5. User's Guide and application published at: http://purl.oclc.org/estimates (2005).

50. Schmid, R., Schuster, S.C., Steel, M.A., and Huson, D.H. ReadSim- A simulator for Sanger and 454 sequencing, software freely available from www-ab.informatik.uni-tuebingen.de/software/readsim.

Supplementary Figures

Figure 1. 16S rRNA gene surveys reveal evidence for familial aggregation and reduced diversity in the obese gut microbiome. (A,B) Average unweighted UniFrac distance (a measure of differences in bacterial community structure) between related and unrelated individuals. 10,000 sequences were randomly sampled from each V6 dataset (Panel A) and 200 sequences were randomly sampled from each full-length dataset (Panel B), OTUs were chosen, a UniFrac tree was built from representative sequences, and random permutations were done on the resulting UniFrac distance matrix. Asterisks indicate significant differences between related and unrelated individuals [Student's t-test with Monte Carlo (1,000 permutations); *$p<0.001$]. (**C,D**) Phylogenetic diversity curves for the obese and lean gut microbiome. 1 to 1,000 sequences were randomly sampled from each V2 dataset (Panel C) and 1 to 200 sequences were randomly sampled from each full-length dataset (Panel D), and the average branch length leading to the sampled sequences was calculated. (**E,F**) Rarefaction curves for the obese and lean fecal microbiota. 1 to 10,000 sequences were randomly sampled from each V6 dataset (Panel E), and 1 to 200 sequences were randomly sampled from each full-length dataset (Panel F). The average number of OTUs in each sample was then calculated (mean±95%CI shown).


A
V6
UniFrac distance
0.66
0.65
0.64
0.63
0.62
0.61
0.60
0.59
*
Unrelated
Related
B
Full-length
UniFrac distance
0.76
0.74
0.72
0.70
0.68
0.66
*
Unrelated
Related
C
V2
Phylogenetic diversity
16
14
12
10
8
6
4
2
lean
obese
0
200
400
600
800
1000
Number of sequences
D
Full-length
Phylogenetic diversity
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
lean
obese
0
50
100
150
200
Number of sequences
E
V6
Number of OTUs (97%)
902
802
702
602
502
402
302
202
102
2
lean
obese
0
2000
4000
6000
8000
10000
Number of sequences
F
Full-length
Number of OTUs (97%)
100
90
80
70
60
50
40
30
20
10
0
lean
obese
0
50
100
150
200
Number of sequences

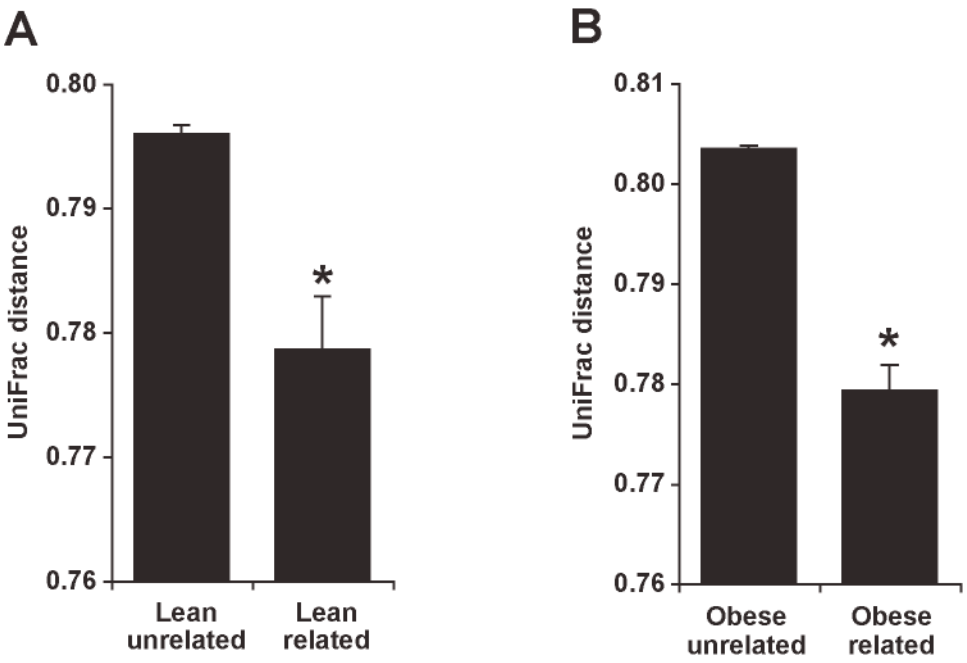


Figure 2. Stratification of related and unrelated individuals concordant for physiological states of obesity versus leanness confirms familial similarity. (A,B) Average UniFrac distance between related and unrelated individuals concordant for leanness (Panel A) or obesity (Panel B). Briefly, 1,000 sequences were randomly sampled from each V2 dataset, OTUs were chosen, a UniFrac tree was built from representative sequences, and random permutations were done on the resulting UniFrac distance matrix. Asterisks indicate significant differences between related and unrelated individuals [Student's t-test with Monte Carlo (1,000 permutations); *$p<10^{-5}$].

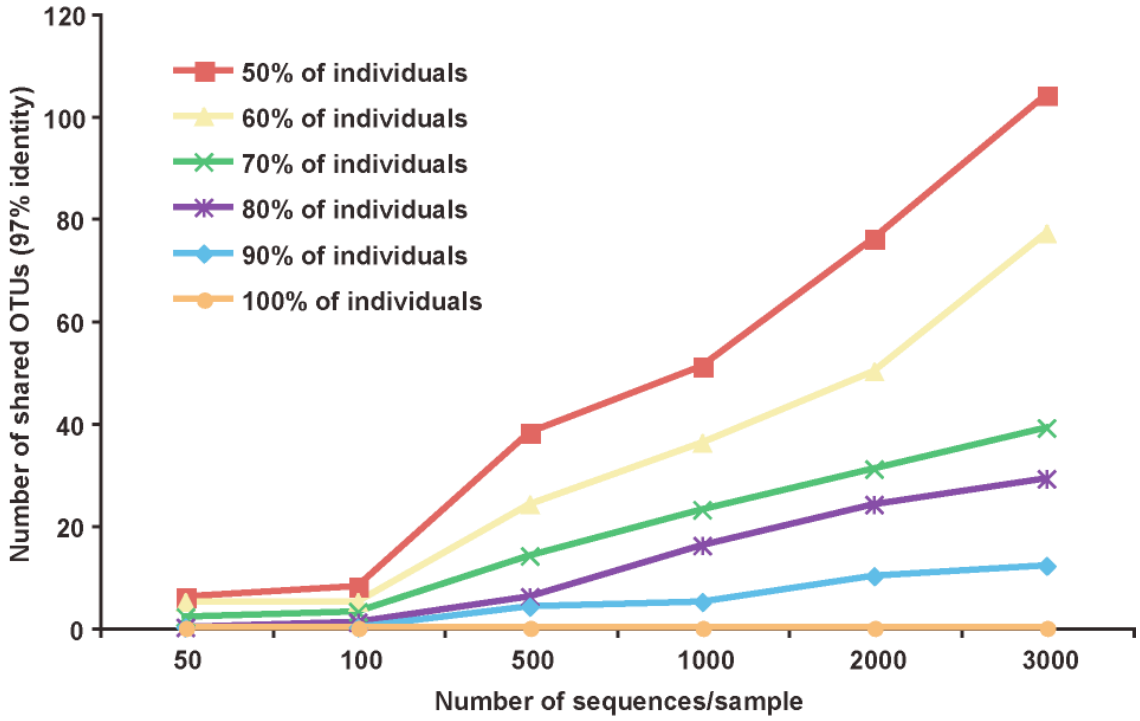


Figure 3. The number of shared phylotypes (OTUs) as a function of the number of sequences per sample. 50-3,000 sequences were randomly selected from each sample, obtained from 93 different individuals. All sequences were binned into 'species'-level phylotypes using a 97% identity threshold. Less stringent parameters were used for OTU binning at all levels of coverage to allow for analysis of 3,000 sequences per sample (density cutoff=0.65, maximum of 3000 nodes).

Figure 4. Clustering of the fecal microbiotas of twins and their mothers sampled at the beginning of the study and two months later. Unweighted UniFrac-based clustering. Colored boxes link samples from the same individual (also indicated by identical IDs followed by the number 1 or 2). 34 of the individuals were only sampled once. 1,000 randomly V2 16S rRNA gene sequences were analyzed per sample.

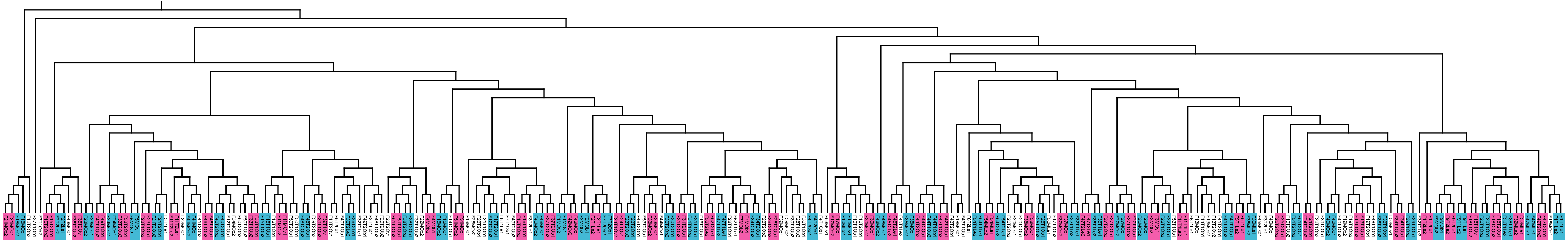


Figure 5. Technical replicates analyzed at four different sequencing centers cluster together. Fecal DNA samples were split and sequenced separately at four different sequencing centers. Abbreviations: usc, Environmental Genomics Core Facility, University of South Carolina; ok, Advanced Center for Genome Technology, University of Oklahoma, ct; 454 Life Sciences Branford, CT; and ma, Josephine Bay Paul Center, Marine Biological Laboratory, Woods Hole Massachusetts. Unweighted UniFrac-based clustering was performed on the combined dataset. Colored boxes enclose samples from the same individual (also indicated by identical IDs followed by the number 1 or 2). The location of the sequencing facility follows each sample ID. Randomly selected sequences were analyzed (≤500 per replicate).

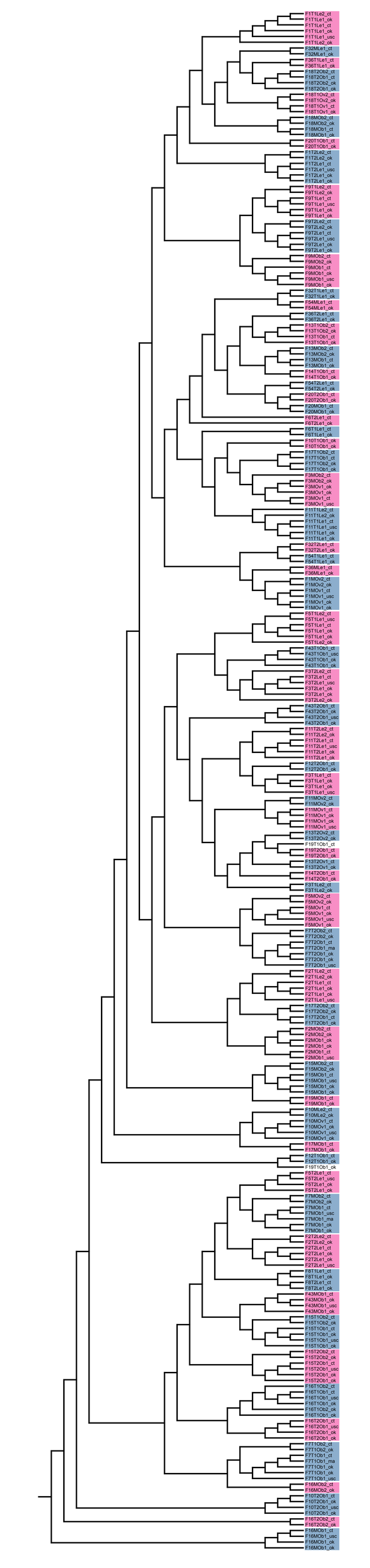

F1T1Le2_ct
F1T1Le1_ok
F1T1Le1_ct
F1T1Le1_ok
F1T1Le1_usc
F1T1Le2_ok
F32MLe1_ct
F32MLe1_ok
F36T1Le1_ct
F36T1Le1_ok
F18T2Ob2_ct
F18T2Ob1_ct
F18T2Ob2_ok
F18T2Ob1_ok
F18T1Ov2_ct
F18T1Ov2_ok
F18T1Ov1_ct
F18T1Ov1_ok
F18MOb2_ct
F18MOb2_ok
F18MOb1_ct
F18MOb1_ok
F20T1Ob1_ct
F20T1Ob1_ok
F1TLe2_ct
F1TLe2_ok
F1TLe1_ct
F1TLe1_usc
F1TLe1_ok
F1TLe1_ok
F9T1Le2_ct
F9T1Le2_ok
F9T1Le1_ct
F9T1Le1_usc
F9T1Le1_ok
F9T1Le1_ok
F9TLe2_ct
F9TLe2_ok
F9TLe1_ct
F9TLe1_usc
F9TLe1_ok
F9TLe1_ok
F9MOb2_ct
F9MOb2_ok
F9MOb1_ct
F9MOb1_ok
F9MOb1_usc
F9MOb1_ok
F32T1Le1_ct
F32T1Le1_ok
F54MLe1_ct
F54MLe1_ok
F36T2Le1_ct
F36T2Le1_ok
F13T1Ob2_ct
F13T1Ob2_ok
F13T1Ob1_ct
F13T1Ob1_ok
F13MOb2_ct
F13MOb2_ok
F13MOb1_ct
F13MOb1_ok
F14T1Ob1_ct
F14T1Ob1_ok
F54T2Le1_ct
F54T2Le1_ok
F20T2Ob1_ct
F20T2Ob1_ok
F20MOb1_ct
F20MOb1_ok
F6TLe1_ct
F6TLe1_ok
F6T1Le1_ct
F6T1Le1_ok
F10T1Ob1_ok
F10T1Ob1_ok
F17T1Ob2_ct
F17T1Ob1_ct
F17T1Ob2_ok
F17T1Ob1_ok
F3MOb2_ct
F3MOb2_ok
F3MOv1_ok
F3MOv1_ok
F3MOv1_ct
F3MOv1_usc
F11T1Le2_ct
F11T1Le2_ok
F11T1Le1_ct
F11T1Le1_usc
F11T1Le1_ok
F11T1Le1_ok
F32T2Le1_ct
F32T2Le1_ok
F54T1Le1_ct
F54T1Le1_ok
F36MLe1_ct
F36MLe1_ok
F1MOv2_ct
F1MOv2_ok
F1MOv1_ct
F1MOv1_usc
F1MOv1_ok
F1MOv1_ok
F5T1Le2_ct
F5T1Le1_usc
F5T1Le1_ct
F5T1Le1_ok
F5T1Le1_ok
F5T1Le2_ok
F43T1Ob1_ct
F43T1Ob1_usc
F43T1Ob1_ok
F43T1Ob1_ok
F3TLe2_ct
F3TLe1_ct
F3TLe1_ok
F3TLe1_ok
F3TLe1_usc
F3TLe2_ok
F11MOv2_ct
F11MOv2_ok
F11MOv1_ct
F11MOv1_ok
F11MOv1_ok
F11MOv1_usc
F13T2Ov2_ct
F13T2Ov2_ok
F19T1Ob1_ct
F19T2Ob1_ct
F19T2Ob1_ok
F13T2Ov1_ct
F13T2Ov1_ok
F14T2Ob1_ct
F14T2Ob1_ok
F3TLe2_ct
F3TLe2_ok
F5MOv2_ct
F5MOv2_ok
F5MOv1_ct
F5MOv1_ok
F5MOv1_usc
F5MOv1_ok
F7T2Ob2_ct
F7T2Ob2_ok
F7T2Ob1_ct
F7T2Ob1_ma
F7T2Ob1_ok
F7T2Ob1_ok
F7T2Ob1_usc
F2T1Le2_ct
F2T1Le2_ok
F2T1Le1_ct
F2T1Le1_ok
F2T1Le1_ok
F2T1Le1_usc
F17T2Ob2_ct
F17T2Ob2_ok
F17T2Ob1_ct
F17T2Ob1_ok
F2MOb2_ct
F2MOb2_ok
F2MOb1_ok
F2MOb1_ok
F2MOb1_ct
F2MOb1_usc
F15MOb2_ct
F15MOb2_ok
F15MOb1_ct
F15MOb1_usc
F15MOb1_ok
F15MOb1_ok
F19MOb1_ct
F19MOb1_ok
F10MLe2_ct
F10MLe2_ok
F10MOv1_ct
F10MOv1_ok
F10MOv1_usc
F10MOv1_ok
F17MOb1_ct
F17MOb1_ok
F12T1Ob1_ct
F12T1Ob1_ok
F19T1Ob1_ok
F5TLe1_ct
F5TLe1_usc
F5TLe1_ok
F5TLe1_ok
F7MOb2_ct
F7MOb2_ok
F7MOb1_ct
F7MOb1_usc
F7MOb1_ma
F7MOb1_ok
F7MOb1_ok
F2TLe2_ct
F2TLe2_ok
F2TLe1_ct
F2TLe1_ok
F2TLe1_ok
F2TLe1_usc
F8T1Le1_ct
F8T1Le1_ok
F8TLe1_ct
F8TLe1_ok
F43MOb1_ct
F43MOb1_ok
F43MOb1_usc
F43MOb1_ok
F15T1Ob2_ct
F15T1Ob2_ok
F15T1Ob1_ct
F15T1Ob1_ok
F15T1Ob1_usc
F15T1Ob1_ok
F15T2Ob2_ct
F15T2Ob2_ok
F15T2Ob1_ct
F15T2Ob1_usc
F15T2Ob1_ok
F15T2Ob1_ok
F16T1Ob2_ct
F16T1Ob1_ct
F16T1Ob1_usc
F16T1Ob1_ok
F16T1Ob2_ok
F16T1Ob1_ok
F16T2Ob1_ct
F16T2Ob1_usc
F16T2Ob1_ok
F16T2Ob1_ok
F7T1Ob2_ct
F7T1Ob2_ok
F7T1Ob1_ct
F7T1Ob1_ma
F7T1Ob1_ok
F7T1Ob1_ok
F7T1Ob1_usc
F16MOb2_ct
F16MOb2_ok
F10T2Ob1_ct
F10T2Ob1_ok
F10T2Ob1_usc
F10T2Ob1_ok
F16T2Ob2_ct
F16T2Ob2_ok
F16MOb1_ct
F16MOb1_usc
F16MOb1_ok
F16MOb1_ok

Figure 6. Relative abundance of the major gut bacterial phyla across 127 gut samples obtained at two different timepoints. Fecal samples were collected at the initial and second timepoints (on average 57±4 days between collections). The relative abundance of the major gut bacterial phyla is based on analysis of V2 16S rRNA gene sequences. Samples are organized based on the rank order abundance of Firmicutes in the initial timepoint.

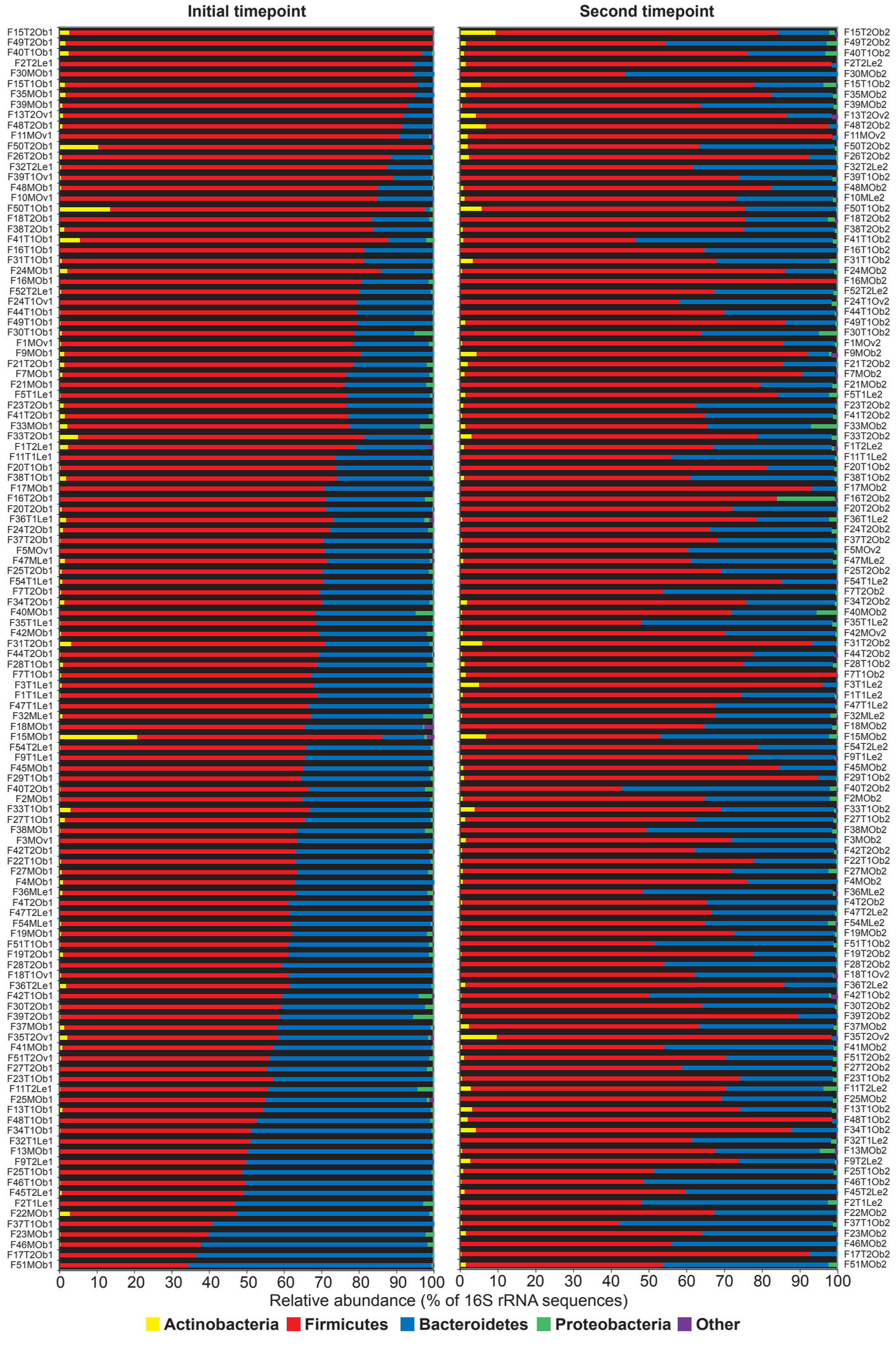
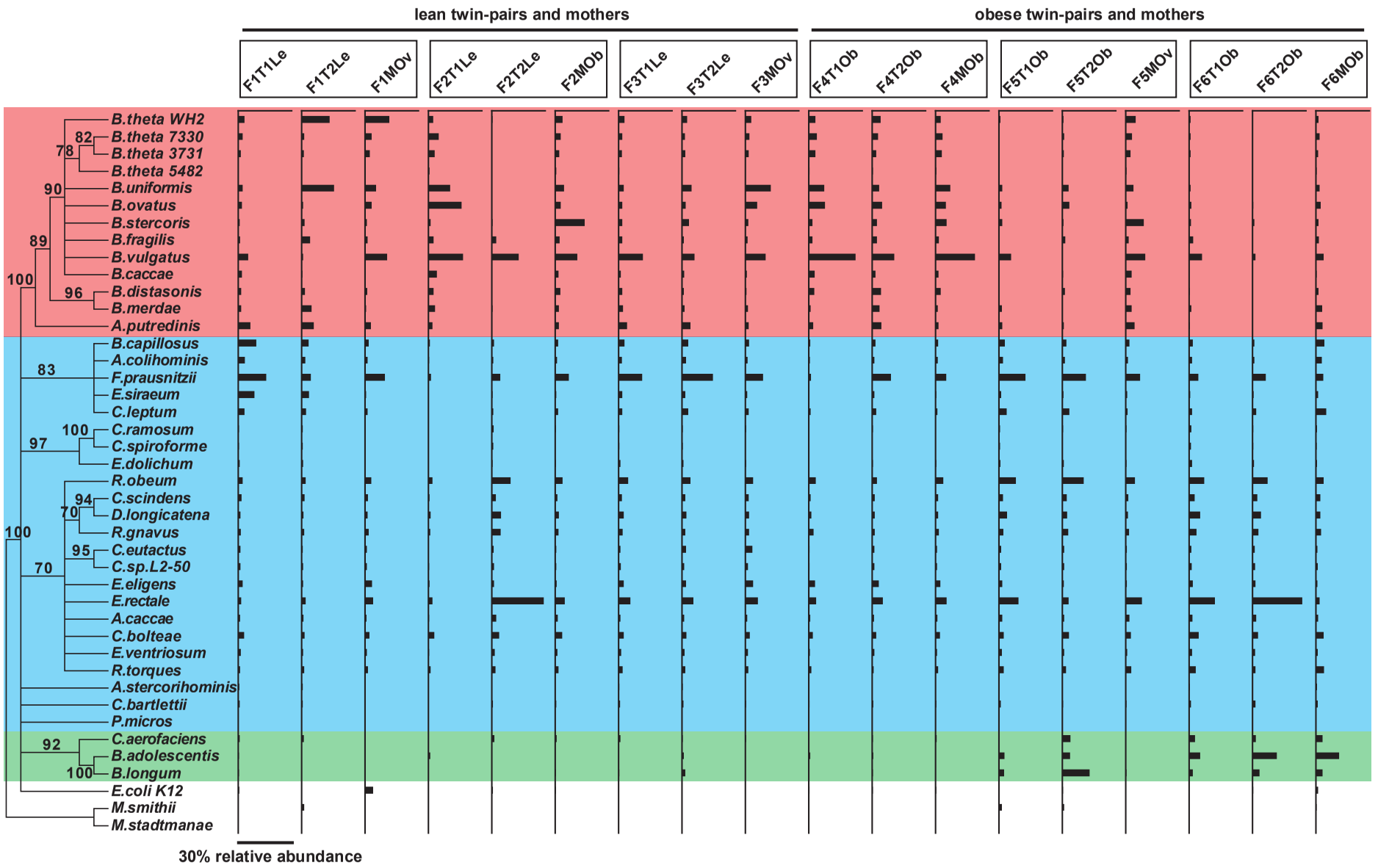

Initial timepoint
Second timepoint
F15T2Ob1
F49T2Ob1
F40T1Ob1
F2T2Le1
F30MOb1
F15T1Ob1
F35MOb1
F39MOb1
F13T2Ov1
F48T2Ob1
F11MOv1
F50T2Ob1
F26T2Ob1
F32T2Le1
F39T1Ov1
F48MOb1
F10MOv1
F50T1Ob1
F18T2Ob1
F38T2Ob1
F41T1Ob1
F16T1Ob1
F31T1Ob1
F24MOb1
F16MOb1
F52T2Le1
F24T1Ov1
F44T1Ob1
F49T1Ob1
F30T1Ob1
F1MOv1
F9MOb1
F21T2Ob1
F7MOb1
F21MOb1
F5T1Le1
F23T2Ob1
F41T2Ob1
F33MOb1
F33T2Ob1
F1T2Le1
F11T1Le1
F20T1Ob1
F38T1Ob1
F17MOb1
F16T2Ob1
F20T2Ob1
F36T1Le1
F24T2Ob1
F37T2Ob1
F5MOv1
F47MLe1
F25T2Ob1
F54T1Le1
F7T2Ob1
F34T2Ob1
F40MOb1
F35T1Le1
F42MOb1
F31T2Ob1
F44T2Ob1
F28T1Ob1
F7T1Ob1
F3T1Le1
F1T1Le1
F47T1Le1
F32MLe1
F18MOb1
F15MOb1
F54T2Le1
F9T1Le1
F45MOb1
F29T1Ob1
F40T2Ob1
F2MOb1
F33T1Ob1
F27T1Ob1
F38MOb1
F3MOv1
F42T2Ob1
F22T1Ob1
F27MOb1
F4MOb1
F36MLe1
F4T2Ob1
F47T2Le1
F54MLe1
F19MOb1
F51T1Ob1
F19T2Ob1
F28T2Ob1
F18T1Ov1
F36T2Le1
F42T1Ob1
F30T2Ob1
F39T2Ob1
F37MOb1
F35T2Ov1
F41MOb1
F51T2Ov1
F27T2Ob1
F23T1Ob1
F11T2Le1
F25MOb1
F13T1Ob1
F48T1Ob1
F34T1Ob1
F32T1Le1
F13MOb1
F9T2Le1
F25T1Ob1
F46T1Ob1
F45T2Le1
F2T1Le1
F22MOb1
F37T1Ob1
F23MOb1
F46MOb1
F17T2Ob1
F51MOb1
F15T2Ob2
F49T2Ob2
F40T1Ob2
F2T2Le2
F30MOb2
F15T1Ob2
F35MOb2
F39MOb2
F13T2Ov2
F48T2Ob2
F11MOv2
F50T2Ob2
F26T2Ob2
F32T2Le2
F39T1Ob2
F48MOb2
F10MLe2
F50T1Ob2
F18T2Ob2
F38T2Ob2
F41T1Ob2
F16T1Ob2
F31T1Ob2
F24MOb2
F16MOb2
F52T2Le2
F24T1Ov2
F44T1Ob2
F49T1Ob2
F30T1Ob2
F1MOv2
F9MOb2
F21T2Ob2
F7MOb2
F21MOb2
F5T1Le2
F23T2Ob2
F41T2Ob2
F33MOb2
F33T2Ob2
F1T2Le2
F11T1Le2
F20T1Ob2
F38T1Ob2
F17MOb2
F16T2Ob2
F20T2Ob2
F36T1Le2
F24T2Ob2
F37T2Ob2
F5MOv2
F47MLe2
F25T2Ob2
F54T1Le2
F7T2Ob2
F34T2Ob2
F40MOb2
F35T1Le2
F42MOv2
F31T2Ob2
F44T2Ob2
F28T1Ob2
F7T1Ob2
F3T1Le2
F1T1Le2
F47T1Le2
F32MLe2
F18MOb2
F15MOb2
F54T2Le2
F9T1Le2
F45MOb2
F29T1Ob2
F40T2Ob2
F2MOb2
F33T1Ob2
F27T1Ob2
F38MOb2
F3MOb2
F42T2Ob2
F22T1Ob2
F27MOb2
F4MOb2
F36MLe2
F4T2Ob2
F47T2Le2
F54MLe2
F19MOb2
F51T1Ob2
F19T2Ob2
F28T2Ob2
F18T1Ov2
F36T2Le2
F42T1Ob2
F30T2Ob2
F39T2Ob2
F37MOb2
F35T2Ov2
F41MOb2
F51T2Ob2
F27T2Ob2
F23T1Ob2
F11T2Le2
F25MOb2
F13T1Ob2
F48T1Ob2
F34T1Ob2
F32T1Le2
F13MOb2
F9T2Le2
F25T1Ob2
F46T1Ob2
F45T2Le2
F2T1Le2
F22MOb2
F37T1Ob2
F23MOb2
F46MOb2
F17T2Ob2
F51MOb2
0 10 20 30 40 50 60 70 80 90 100
Relative abundance (% of 16S rRNA sequences)
Actinobacteria
Firmicutes
Bacteroidetes
Proteobacteria
Other

Figure 7. Taxonomic profiles of microbial gene content in the human gut (fecal) microbiome. Branches and distributions are colored by phylum: Bacteroidetes (red), Firmicutes (blue), and Actinobacteria (green). Proteobacteria (*E.coli*) and Archaea (*M.smithii* and *M.stadtmanae*) are uncolored. The relative abundance of sequences homologous to each genome is depicted on a scale of 0 to 30% (BLASTX comparisons of microbiome datasets to reference genomes).


lean twin-pairs and mothers
obese twin-pairs and mothers
F1T1Le
F1T2Le
F1MOv
F2T1Le
F2T2Le
F2MOb
F3T1Le
F3T2Le
F3MOv
F4T1Ob
F4T2Ob
F4MOb
F5T1Ob
F5T2Ob
F5MOv
F6T1Ob
F6T2Ob
F6MOb
B.theta WH2
B.theta 7330
B.theta 3731
B.theta 5482
B.uniformis
B.ovatus
B.stercoris
B.fragilis
B.vulgatus
B.caccae
B.distasonis
B.merdae
A.putredinis
B.capillosus
A.colihominis
F.prausnitzii
E.siraeum
C.leptum
C.ramosum
C.spiroforme
E.dolichum
R.obeum
C.scindens
D.longicatena
R.gnavus
C.eutactus
C.sp.L2-50
E.eligens
E.rectale
A.caccae
C.bolteae
E.ventriosum
R.torques
A.stercorihominis
C.bartlettii
P.micros
C.aerofaciens
B.adolescentis
B.longum
E.coli K12
M.smithii
M.stadtmanae
82
78
90
89
100
96
83
100
97
94
70
100
95
70
92
100
30% relative abundance

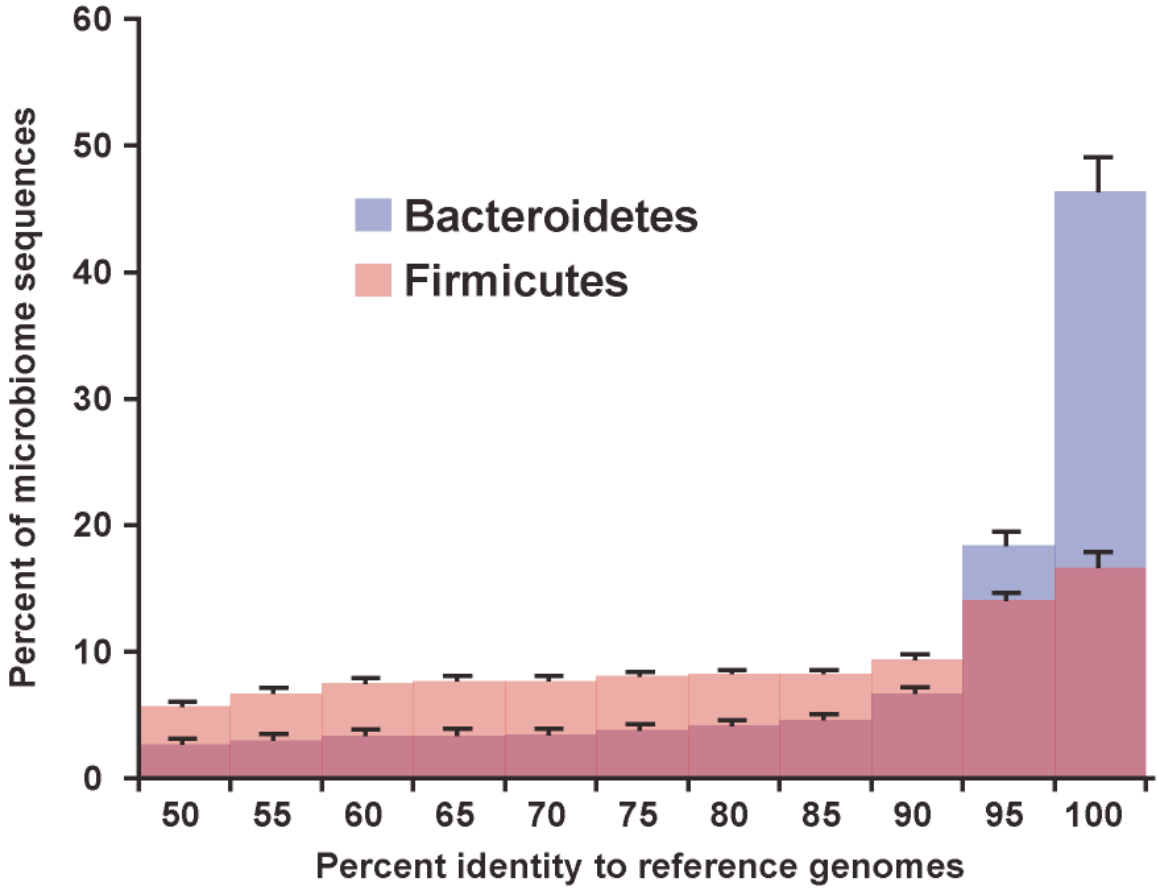
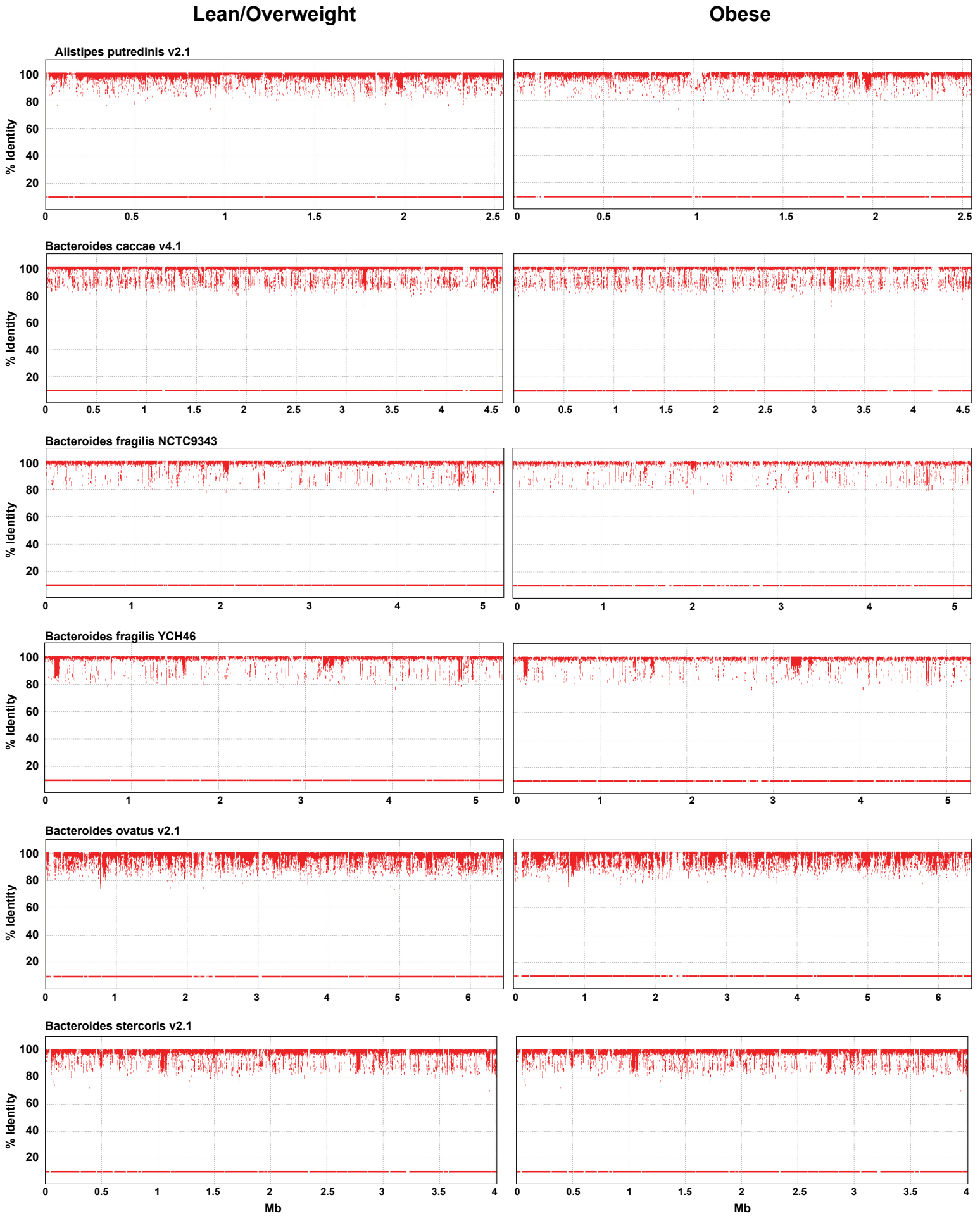
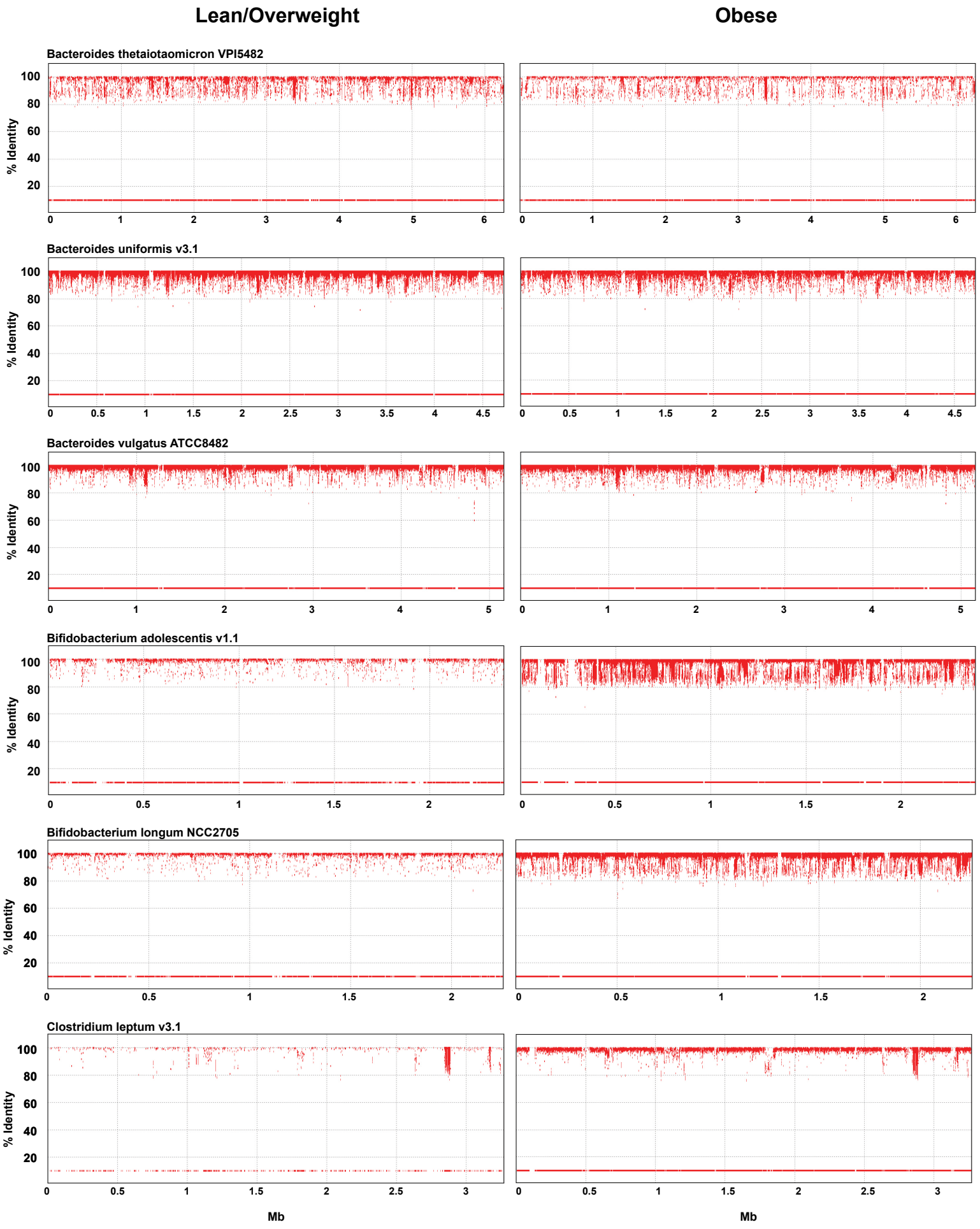
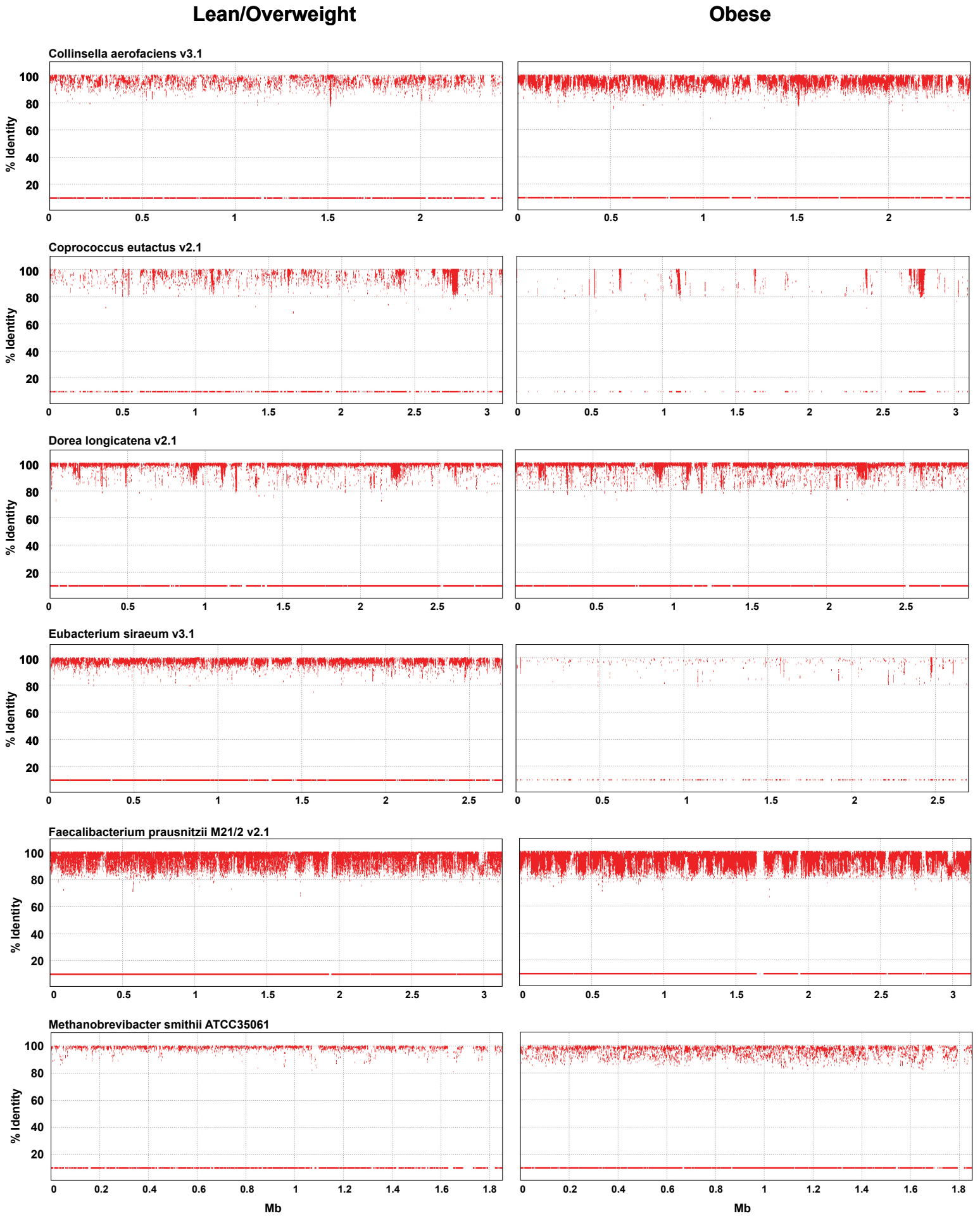


Figure 8. Assignment of fecal microbiome reads to sequenced reference human gut-derived Bacteroidetes and Firmicutes genomes. Histogram of the percent identity (mean±SEM) obtained from sequence alignments between gut microbiome reads (n=18 datasets) and Firmicutes or Bacteroidetes reference genomes.

Figure 9. Percent identity plots of the fecal microbiomes versus reference genomes. Each row (x-axis) represents a different genome. The y-axis shows the percent identity to microbiome sequences (red dots). The combined data from lean/overweight individuals are in the left column while the combined data from obese individuals is displayed in the right column. Supercontigs were used for draft genomes; the assembly version (v) can be found after the strain name. The lines found at 10% identity on each plot depict the sum of all sequences mapped across each genome.


Lean/Overweight
Obese
Alistipes putredinis v2.1
Bacteroides caccae v4.1
Bacteroides fragilis NCTC9343
Bacteroides fragilis YCH46
Bacteroides ovatus v2.1
Bacteroides stercoris v2.1
% Identity
Mb


Lean/Overweight
Obese
Bacteroides thetaiotaomicron VPI5482
Bacteroides uniformis v3.1
Bacteroides vulgatus ATCC8482
Bifidobacterium adolescentis v1.1
Bifidobacterium longum NCC2705
Clostridium leptum v3.1
% Identity
Mb


Lean/Overweight
Obese
Collinsella aerofaciens v3.1
Coprococcus eutactus v2.1
Dorea longicatena v2.1
Eubacterium siraeum v3.1
Faecalibacterium prausnitzii M21/2 v2.1
Methanobrevibacter smithii ATCC35061
% Identity
Mb

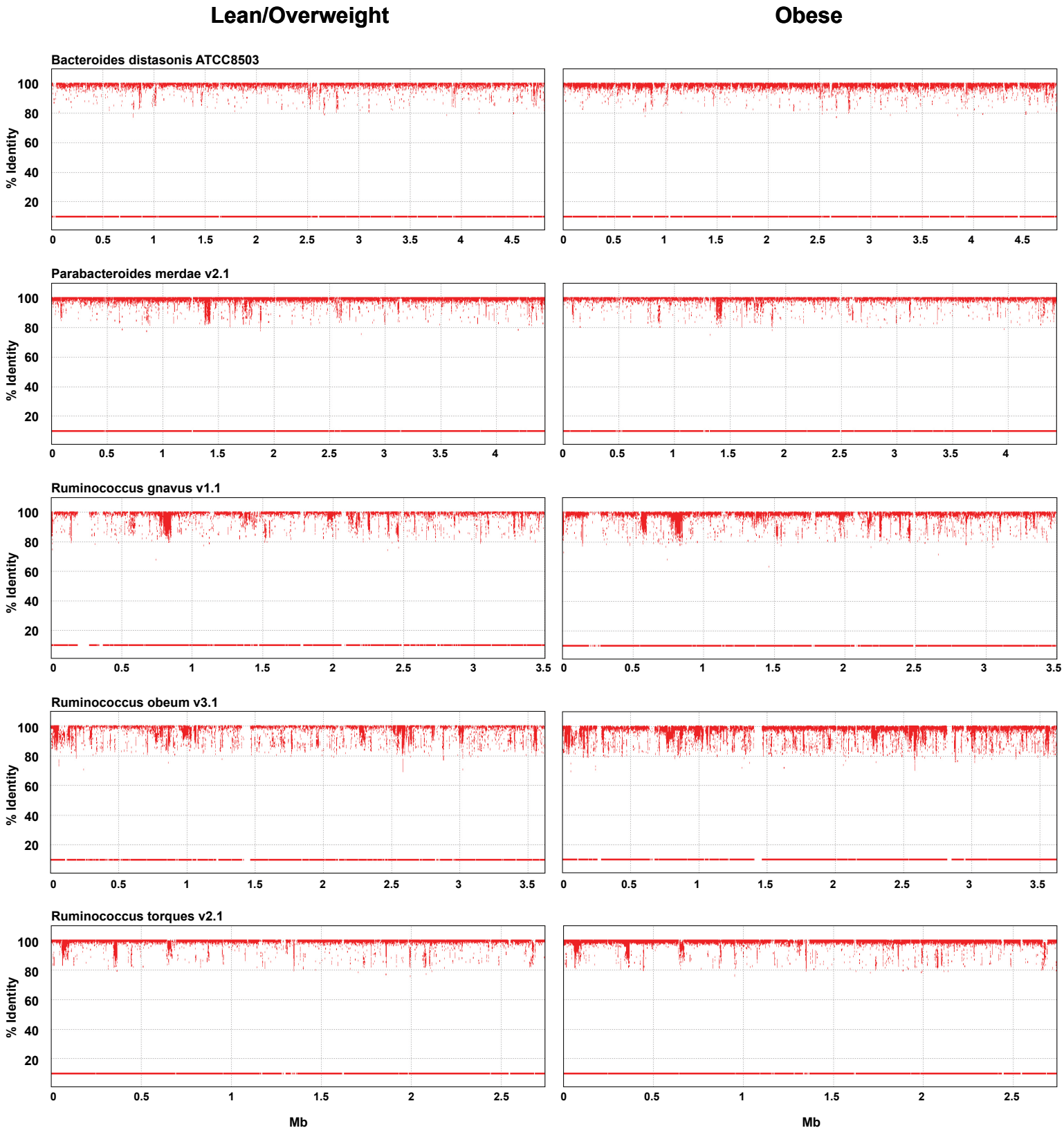
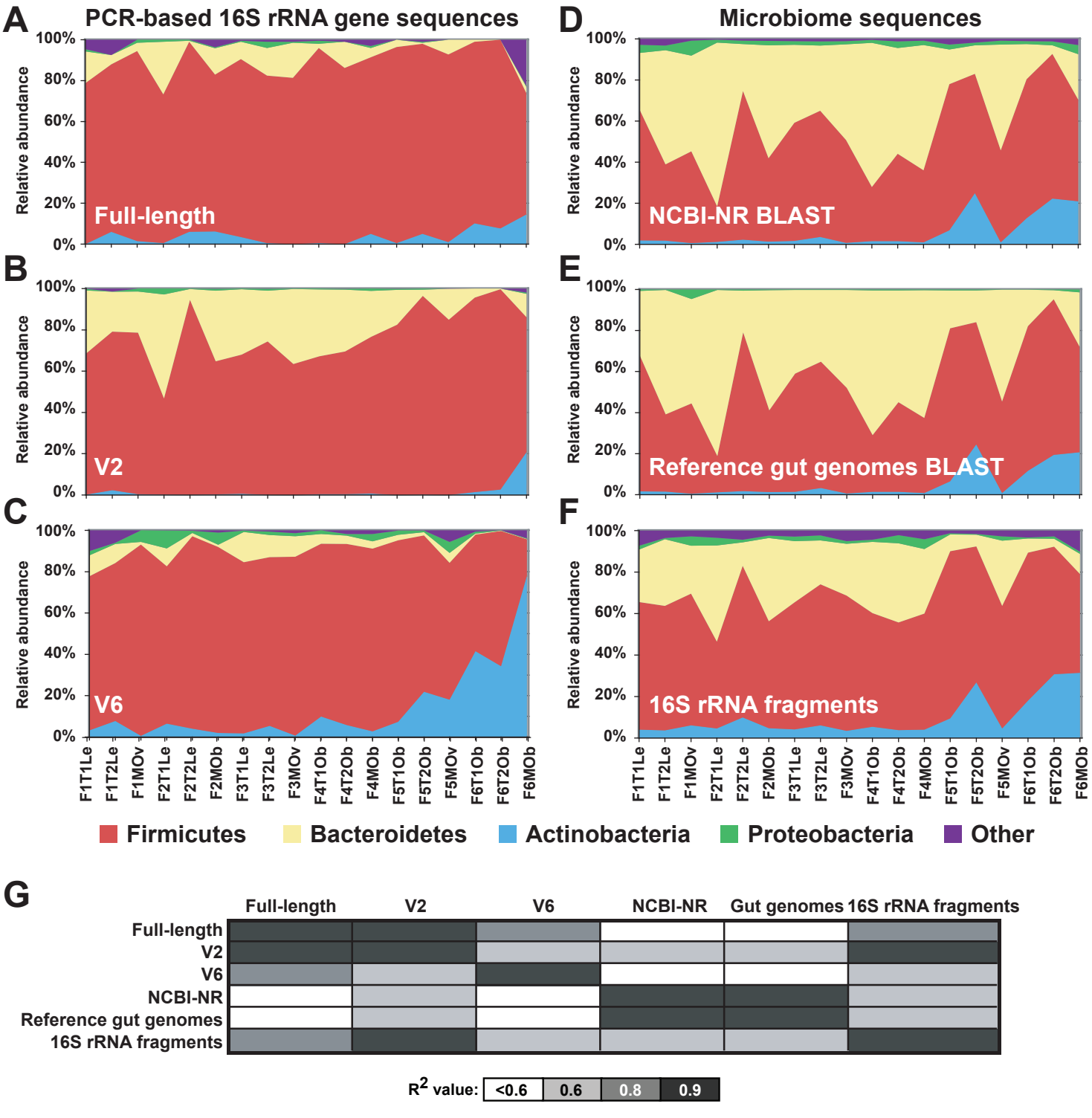

Lean/Overweight
Obese
Bacteroides distasonis ATCC8503
Parabacteroides merdae v2.1
Ruminococcus gnavus v1.1
Ruminococcus obeum v3.1
Ruminococcus torques v2.1
% Identity
100
80
60
40
20
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
Mb

Figure 10. Relative abundance of bacterial phyla in 18 human gut microbiomes. (**A-C**) PCR-based 16S rRNA gene sequences [(**A**) full-length, (**B**) V2 region, and (**C**) V6]. (**D-E**) Microbiome data analyzed by BLAST comparisons [(**D**) NCBI non-redundant database and (**E**) a custom 44 gut genome database]. (**F**) Analysis of 16S rRNA gene fragments identified in each microbiome. (**G**) Correlation matrix based on all pairwise comparisons (R^2) of the relative abundance of the four major phyla (Actinobacteria, Firmicutes, Bacteroidetes, and Proteobacteria) across all six methods.


A
PCR-based 16S rRNA gene sequences
Full-length
B
V2
C
V6
D
Microbiome sequences
NCBI-NR BLAST
E
Reference gut genomes BLAST
F
16S rRNA fragments
Relative abundance
100%
80%
60%
40%
20%
0%
F1T1Le
F1T2Le
F1MOv
F2T1Le
F2T2Le
F2MOb
F3T1Le
F3T2Le
F3MOv
F4T1Ob
F4T2Ob
F4MOb
F5T1Ob
F5T2Ob
F5MOv
F6T1Ob
F6T2Ob
F6MOb
Firmicutes
Bacteroidetes
Actinobacteria
Proteobacteria
Other

G
Full-length
V2
V6
NCBI-NR
Gut genomes
16S rRNA fragments
Full-length
V2
V6
NCBI-NR
Reference gut genomes
16S rRNA fragments
R2 value:
<0.6
0.6
0.8
0.9

Figure 11. Validation of annotation parameters using control datasets. (**A-C**) Percent of randomly fragmented annotated genes (KEGG v44) assigned to the correct KEGG orthologous group as a function of the (**A**) e-value, (**B**) % identity, or (**C**) bit-score cutoff used. (**D-F**) Sensitivity [true positives (TP) divided by true positives plus false negatives (FN)] as a function of the (**D**) e-value, (**E**) % identity, or (**F**) bit-score cutoff used. (**G-I**) Precision [true positives divided by true positives plus false positives (FP)] as a function of the (**G**) e-value, (**H**) % identity, or (**I**) bit-score cutoff used. The vertical gray line and circle indicates the cutoff values used in this analysis.

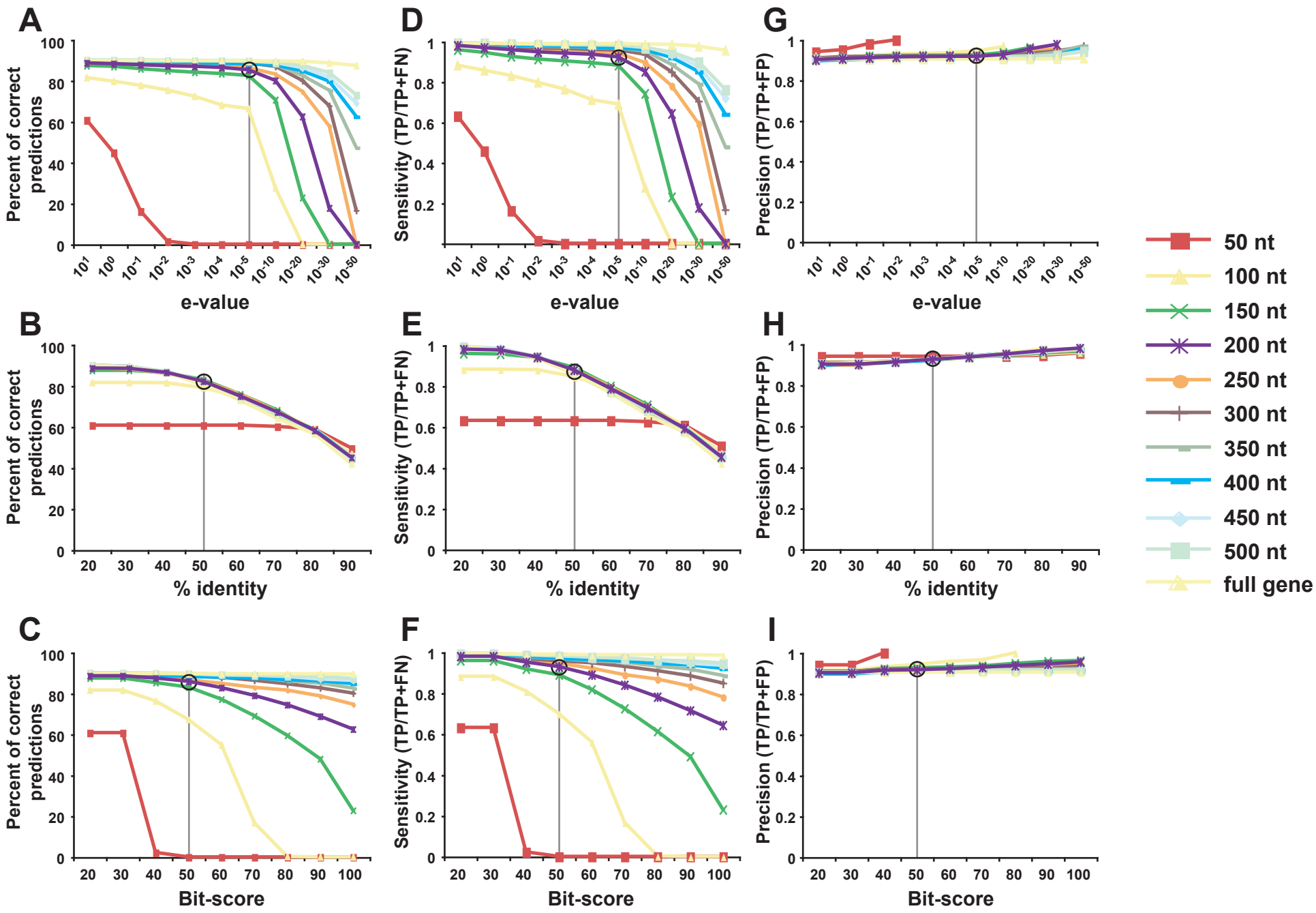
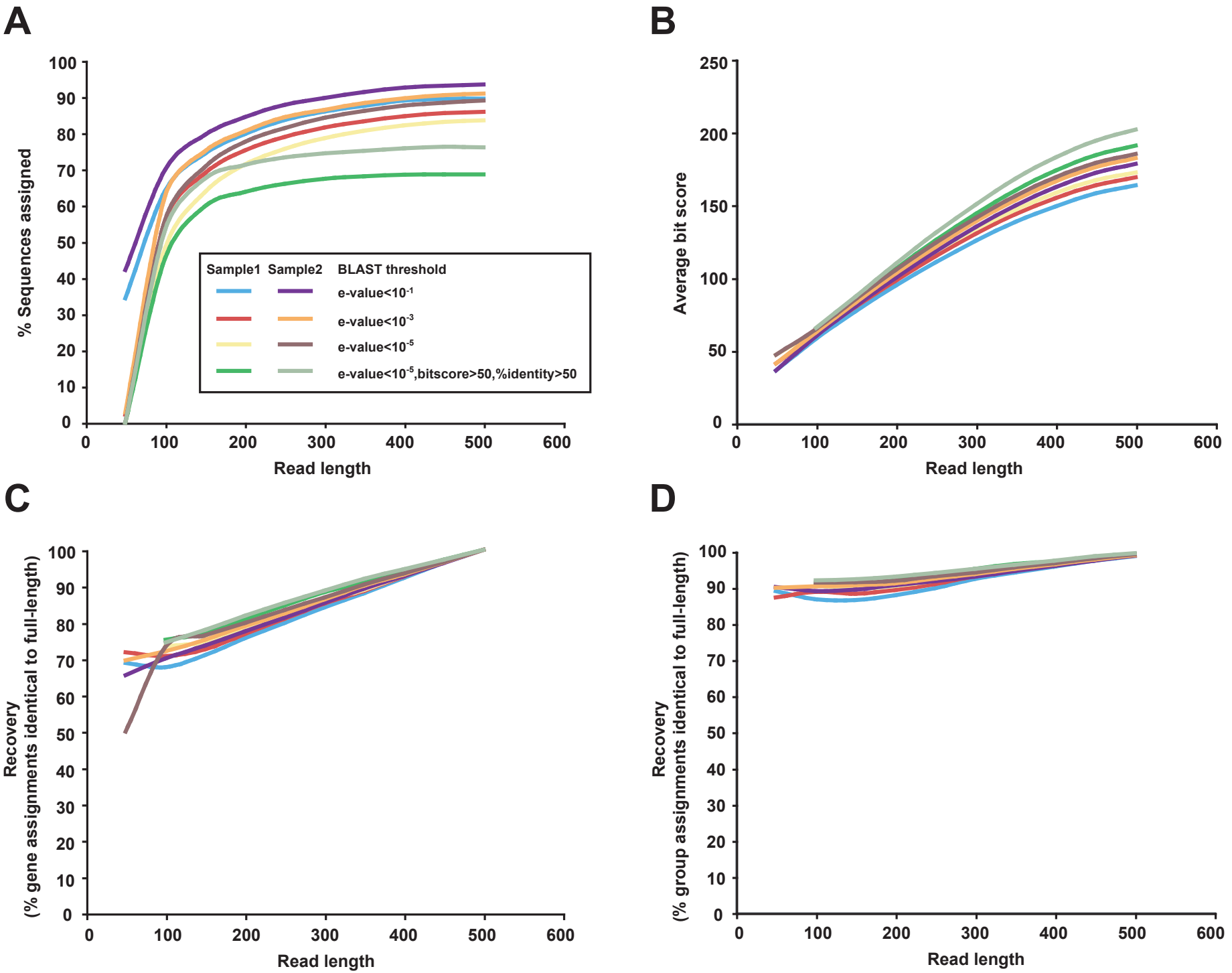

A
B
C
D
E
F
G
H
I
Percent of correct predictions
Sensitivity (TP/TP+FN)
Precision (TP/TP+FP)
e-value
% identity
Bit-score
50 nt
100 nt
150 nt
200 nt
250 nt
300 nt
350 nt
400 nt
450 nt
500 nt
full gene

Figure 12. Dependence of percentage (A), quality (B), and accuracy (C-D) of sequence assignments on read-length. Two fecal samples were processed using extra-long read pyrosequencing (454 FLX Titanium kit; samples TS28 and TS29). 10,000 sequences from the maximum of each read-length distribution (between 490 and 505 nt) were randomly selected from each sample. Simulated reads were created by sampling the first 50-500 nt of each of these 10,000 sequences, and each simulated read was compared using NCBI-BLASTX against our custom gut genome database. Multiple BLAST thresholds were used (see key in panel **A**). (**A**) Percent of sequences assigned to the reference genomes as a function of read-length. (**B**) Average BLAST bit score as a function of read-length. (**C**) Percent of gene assignments (from the gut genome database) identical to full-length sequence as a function of read-length. (**D**) Percent of group assignments (same assigned COG as the full-length sequence) as a function of read-length.


A
% Sequences assigned
Read length
Sample1 Sample2 BLAST threshold
e-value<10-1
e-value<10-3
e-value<10-5
e-value<10-5,bitscore>50,%identity>50
B
Average bit score
Read length
C
Recovery
(% gene assignments identical to full-length)
Read length
D
Recovery
(% group assignments identical to full-length)
Read length

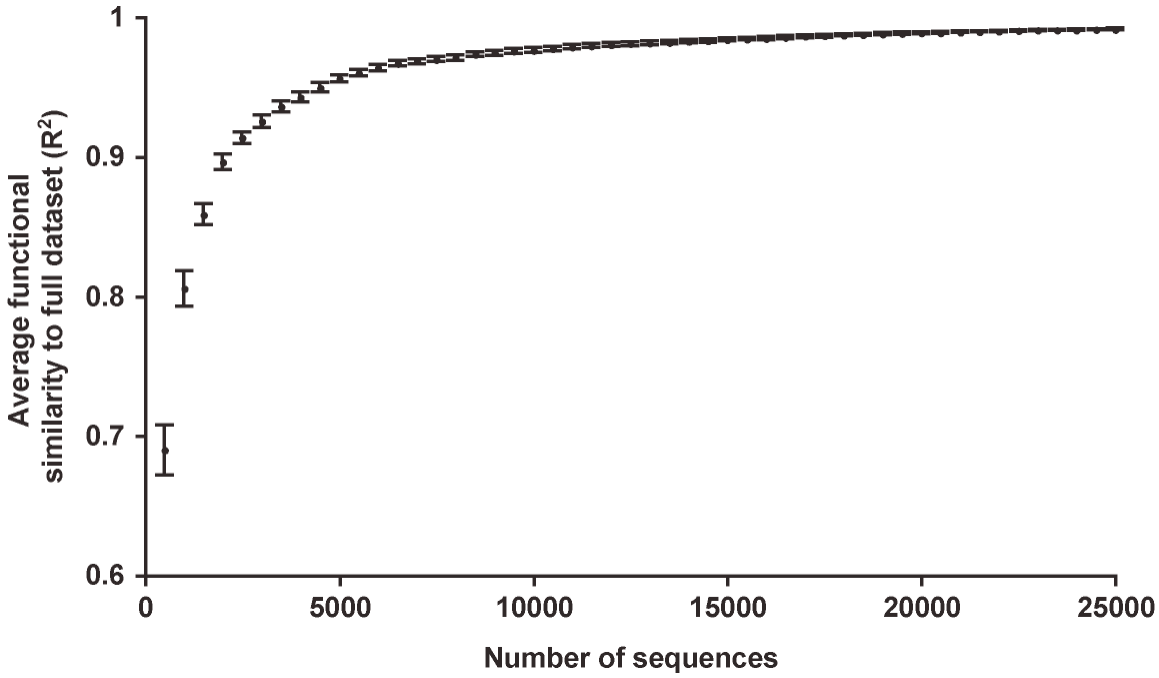
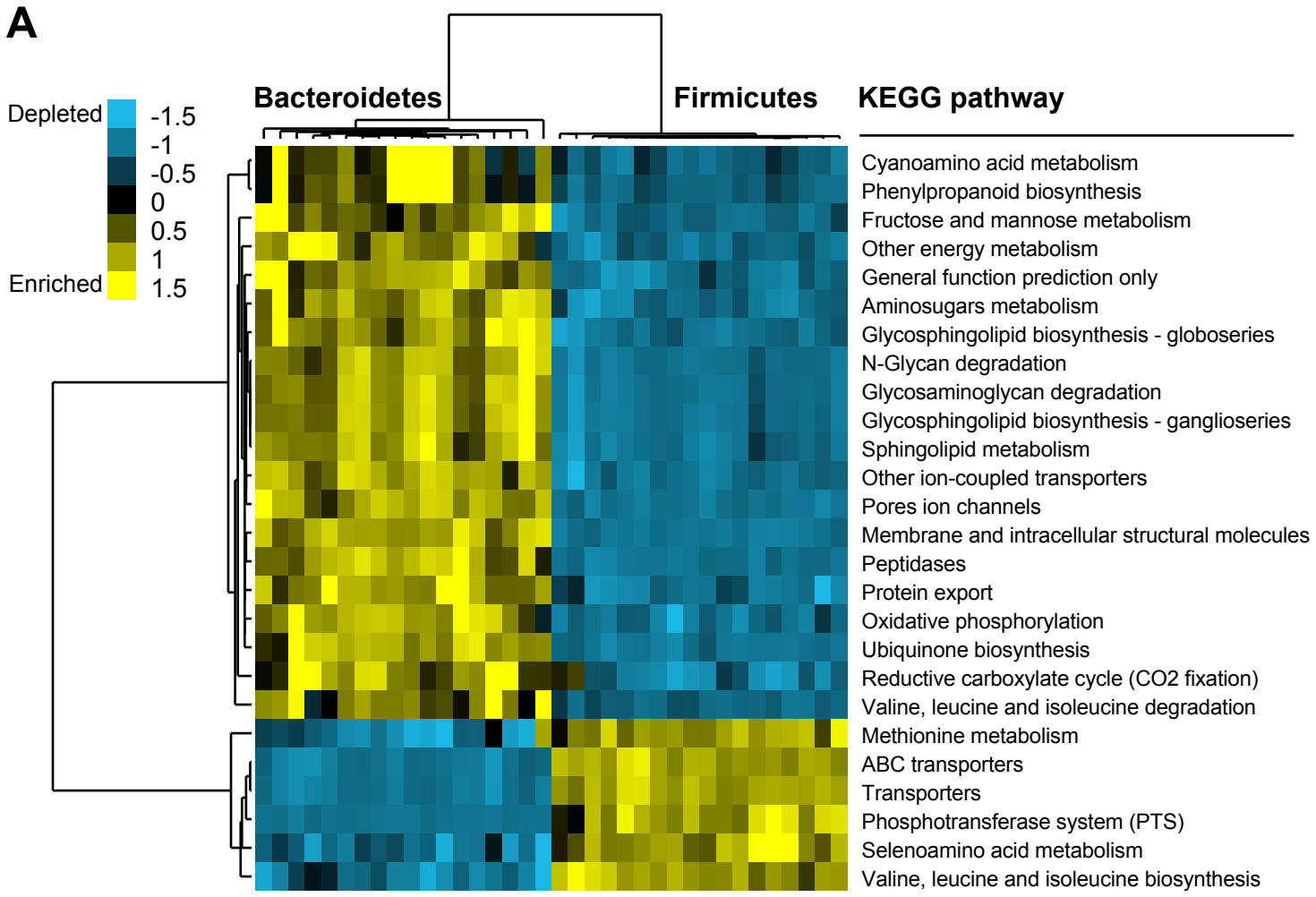
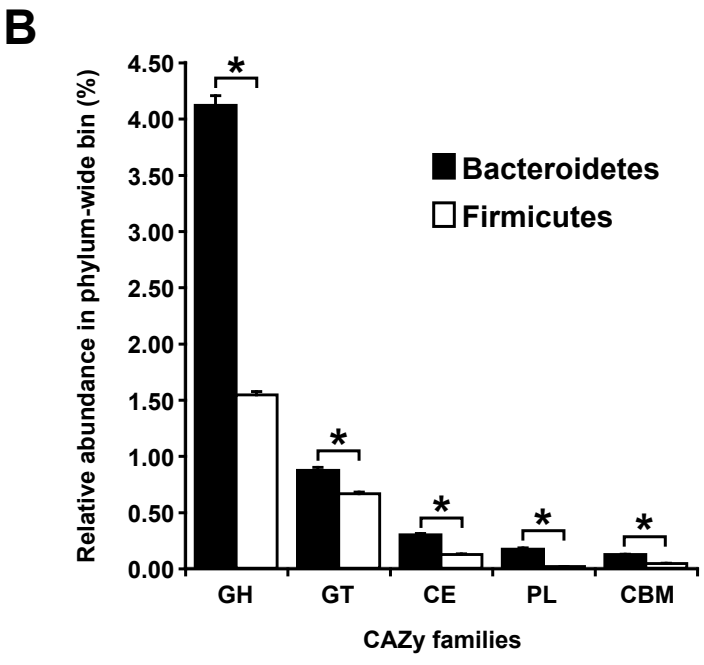


Figure 13. Functional profiles of MZ fecal microbiomes, based on the relative abundance of KEGG pathways, stabilize after ~20,000 sequences are collected for a given sample. Datasets were randomly subsampled between 500 and 25,000 sequences. The average functional similarity (R^2) between the subsampled dataset and the full dataset is shown as a function of sequencing effort.

Figure 14. KEGG pathways and CAZy families whose representation is significantly different between Firmicutes and Bacteroidetes bins. Sequences from each of the 18 fecal microbiomes were binned based on sequence homology to the custom 44-member reference human gut genome database. **(A)** The frequency of each KEGG pathway was tallied for each bin and significantly different pathways were identified using a bootstrap re-sampling analysis[23] (Xipe v2.4). Significantly different pathways reaching at least 0.6% relative abundance in at least two microbiomes were clustered using single-linkage hierarchical clustering and the Pearson's correlation distance metric. **(B)** The relative abundance of CAZy families in the Bacteroidetes and Firmicutes sequence bins. Asterisks indicate significant differences (Mann-Whitney test, $p<0.0001$).


A
Depleted
-1.5
-1
-0.5
0
0.5
1
Enriched
1.5
Bacteroidetes
Firmicutes
KEGG pathway
Cyanoamino acid metabolism
Phenylpropanoid biosynthesis
Fructose and mannose metabolism
Other energy metabolism
General function prediction only
Aminosugars metabolism
Glycosphingolipid biosynthesis - globoseries
N-Glycan degradation
Glycosaminoglycan degradation
Glycosphingolipid biosynthesis - ganglioseries
Sphingolipid metabolism
Other ion-coupled transporters
Pores ion channels
Membrane and intracellular structural molecules
Peptidases
Protein export
Oxidative phosphorylation
Ubiquinone biosynthesis
Reductive carboxylate cycle (CO2 fixation)
Valine, leucine and isoleucine degradation
Methionine metabolism
ABC transporters
Transporters
Phosphotransferase system (PTS)
Selenoamino acid metabolism
Valine, leucine and isoleucine biosynthesis


B
Relative abundance in phylum-wide bin (%)
4.50
4.00
3.50
3.00
2.50
2.00
1.50
1.00
0.50
0.00
*
Bacteroidetes
Firmicutes
GH
GT
CE
PL
CBM
CAZy families

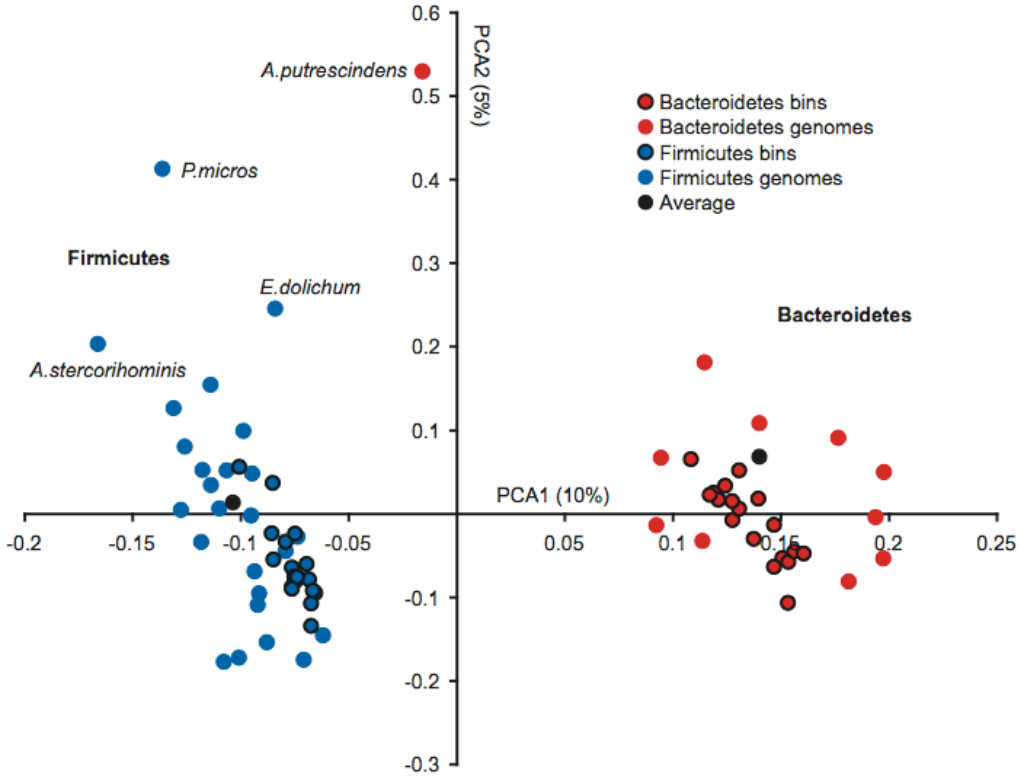
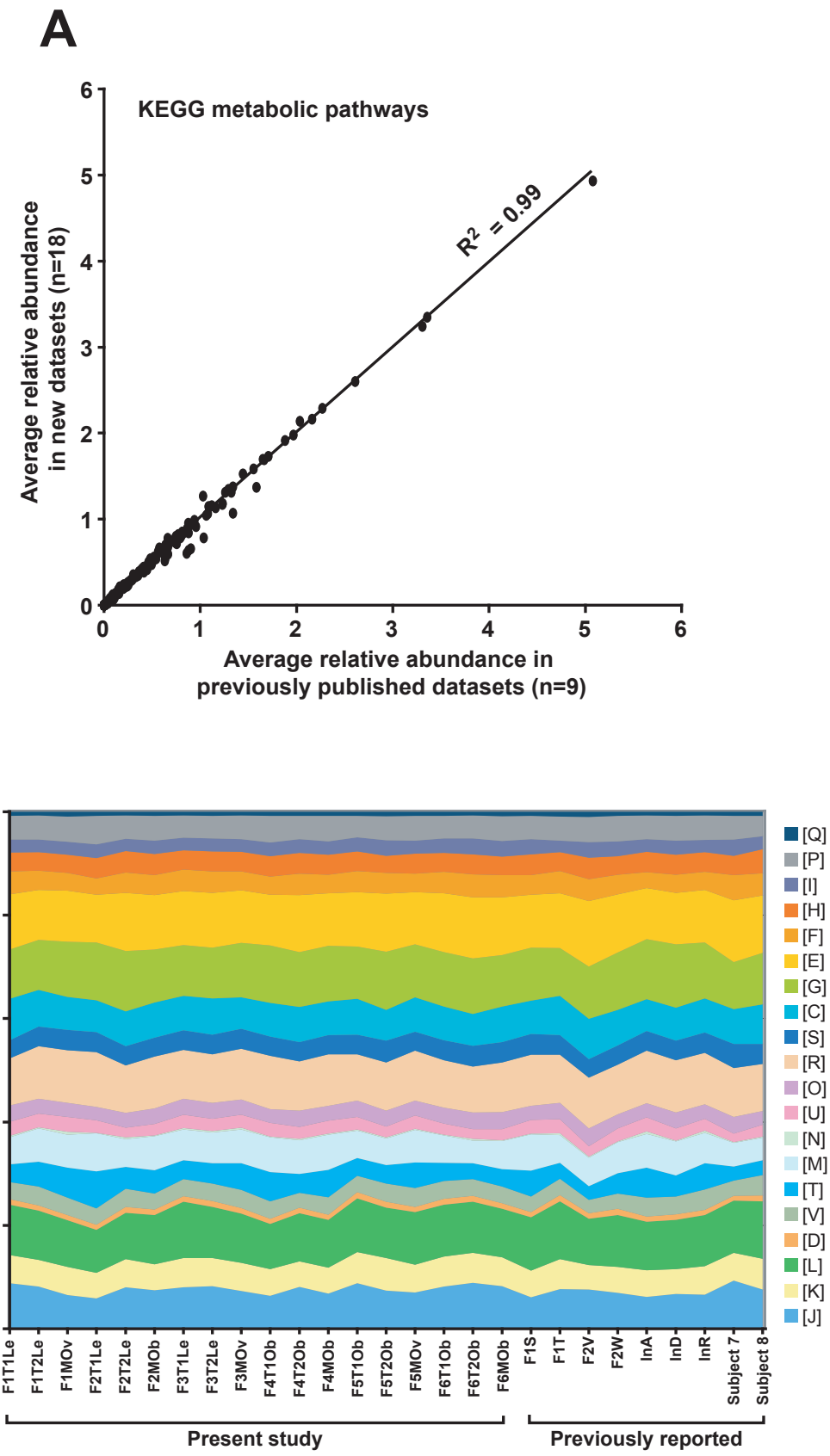


Figure 15. Functional clustering of phylum-wide sequence bins and reference genomes from 36 human gut-derived Bacteroidetes and Firmicutes. The frequency of each KEGG pathway in phylum-wide sequence bins, and in 10,000 'simulated reads' generated from each of the reference genomes (Readsim v0.10; ref. 50), was tallied and pathways reaching at least 0.6% relative abundance in at least two fecal microbiomes were clustered using Principal Components Analysis (PCA). An 'average' Firmicutes and Bacteroidetes genome was generated by pooling all reads generated from genomes within each phylum.

Figure 16. Relative abundance of KEGG pathways and COG categories in the gut microbiomes of 18 individuals (6 MZ twin pairs and their mothers), plus 9 previously published adult microbiomes[20,21]. 'Simulated reads' were generated from each of the 9 previously published microbiomes datasets obtained by capillary sequencing to mimic pyrosequencing reads, then re-annotated using the KEGG[17] and STRING-extended COG databases[18]. (**A**) The average relative abundance of KEGG pathways in MZ twin pairs and their mothers graphed as a function of the average relative abundance of KEGG pathways in the 9 previously published adult gut microbiome datasets. (**B**) The distribution of COG categories across all 27 datasets.


A
KEGG metabolic pathways
R^2 = 0.99
Average relative abundance in new datasets (n=18)
Average relative abundance in previously published datasets (n=9)
0
1
2
3
4
5
6
B
COG category relative abundance
100%
80%
60%
40%
20%
0%
[Q]
[P]
[I]
[H]
[F]
[E]
[G]
[C]
[S]
[R]
[O]
[U]
[N]
[M]
[T]
[V]
[D]
[L]
[K]
[J]
F1T1Le
F1T2Le
F1MOv
F2T1Le
F2T2Le
F2MOb
F3T1Le
F3T2Le
F3MOv
F4T1Ob
F4T2Ob
F4MOb
F5T1Ob
F5T2Ob
F5MOv
F6T1Ob
F6T2Ob
F6MOb
F1S
F1T
F2V
F2W
InA
InD
InR
Subject 7
Subject 8
Present study
Previously reported

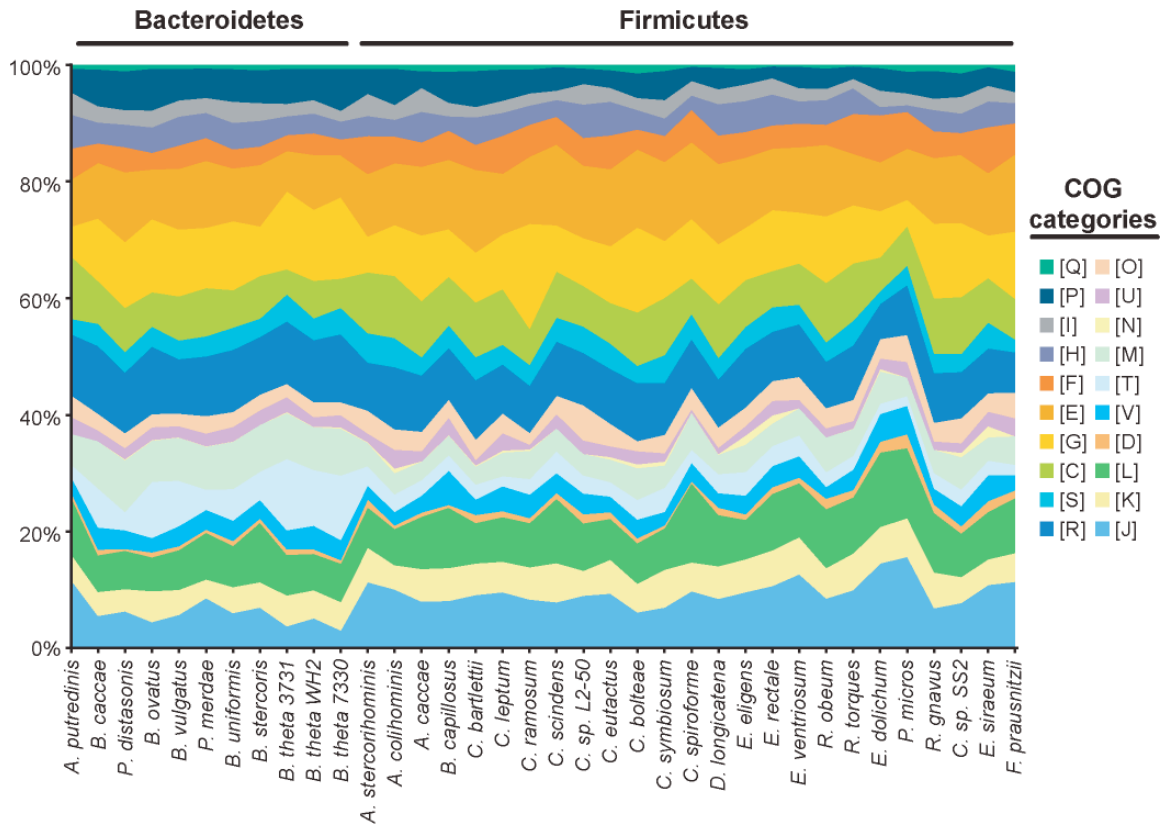


Figure 17. Relative abundance of COG categories in 36 sequenced reference human gut-derived Firmicutes and Bacteroidetes genomes. 10,000 'simulated reads', generated from each of the reference genomes (Readsim v0.10; ref. 50), were annotated using the STRING-extended COG database[18].

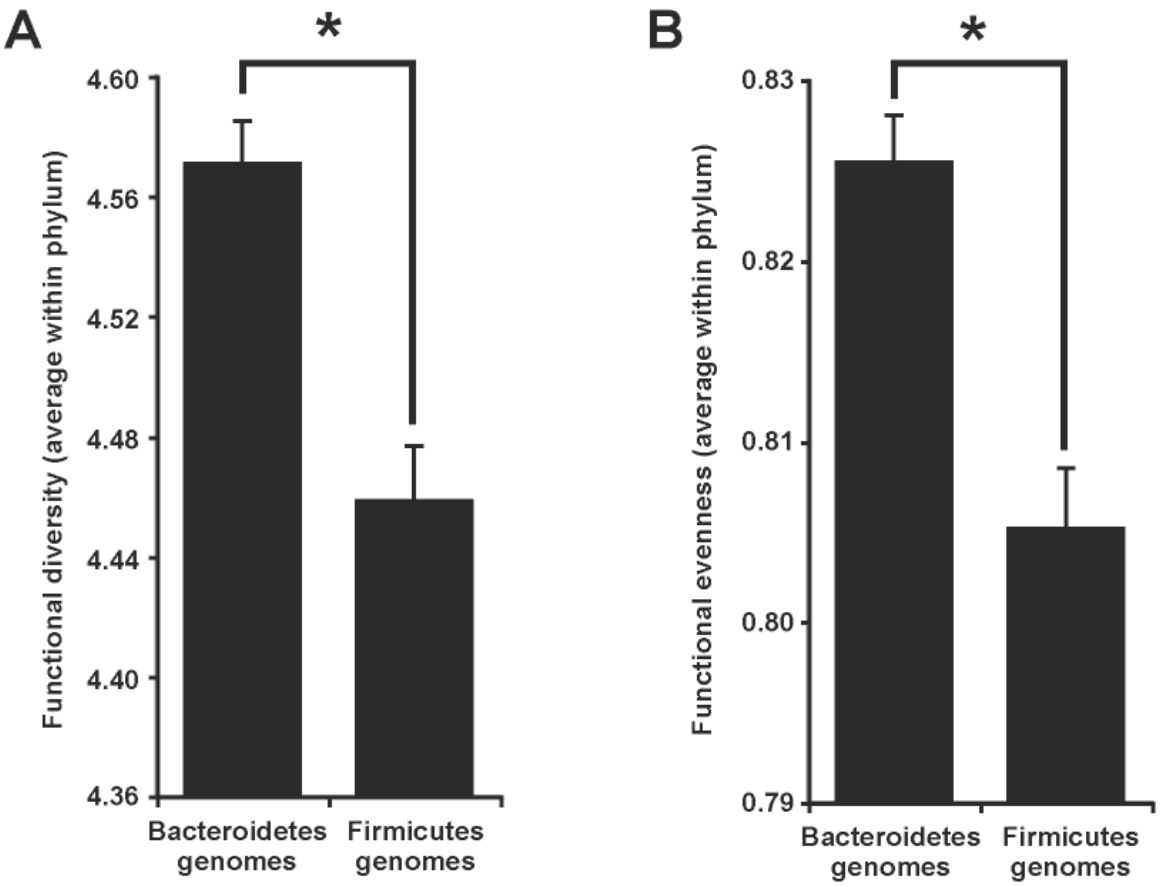


Figure 18. Average functional diversity and evenness of 'simulated reads' generated from reference gut Firmicute or Bacteroidetes genomes. (A) Functional diversity was calculated in EstimateS[49] (v8.0), based on the abundance of each metabolic pathway across 10,000 'simulated reads' generated from each of the 36 reference genomes (Readsim v0.10; ref. 50). **(B)** Shannon evenness. Asterisks indicate significant differences (Mann-Whitney test, $p<0.01$).

Figure 19. 'Enzyme'-level functional groups shared between all or a subset of the sampled gut microbiomes. Sequences from each of the 18 microbiomes characterized in this study were assigned to (**A**) KEGG groups, (**B**) CAZy families, and (**C**) STRING annotations (see Supplementary Methods). The proportion of functional groups (inner circle), and the proportion of sequences assigned to each functional group (outer circle) were tallied based on the co-occurrence of each functional group in any combination of 1 to 18 microbiomes. For example, the outer aqua-colored segment (labeled with the number 18) in Panel A shows that 96.2% of the total sequences generated from all 18 samples were assigned to functional groups that were common to all 18 microbiomes, while the outer blue-colored segment indicates that 1.4% of the total sequences were assigned to groups common to 17 microbiomes. The inner aqua-colored segment (number 18) in Panel A demonstrates that 39.9% of the functional groups found across the total dataset were common to all 18 samples, while the inner light blue-colored segment (number 1) indicates groups only found in one sample. **(D)** Rarefaction curves of the increasing number of 'core' or 'variable' functional groups obtained with additional metagenomic sequencing, as defined by the number of STRING orthologous groups that are or are not shared across all 18 microbiomes after randomly sampling each microbiome in 1,000 sequence intervals.

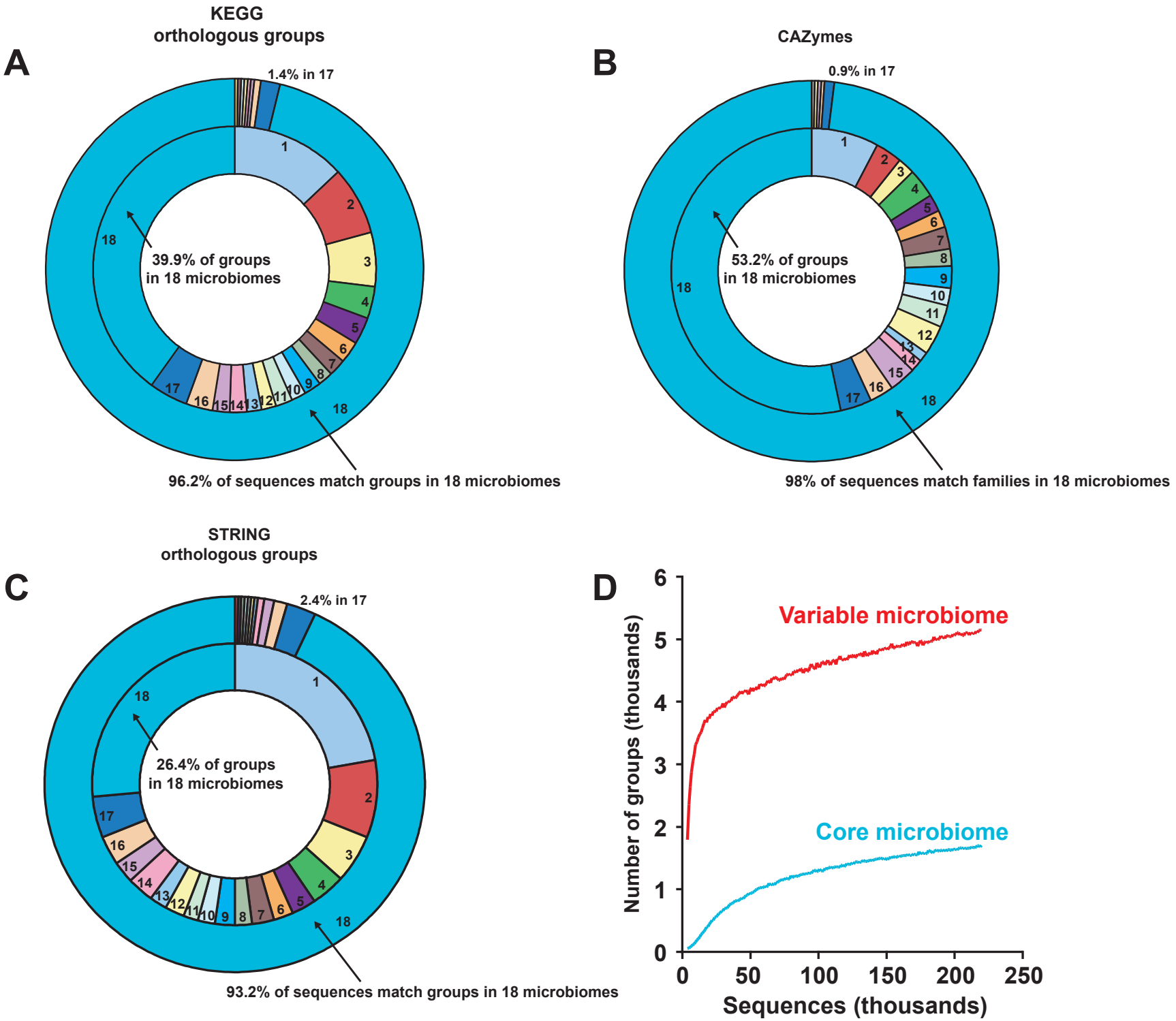
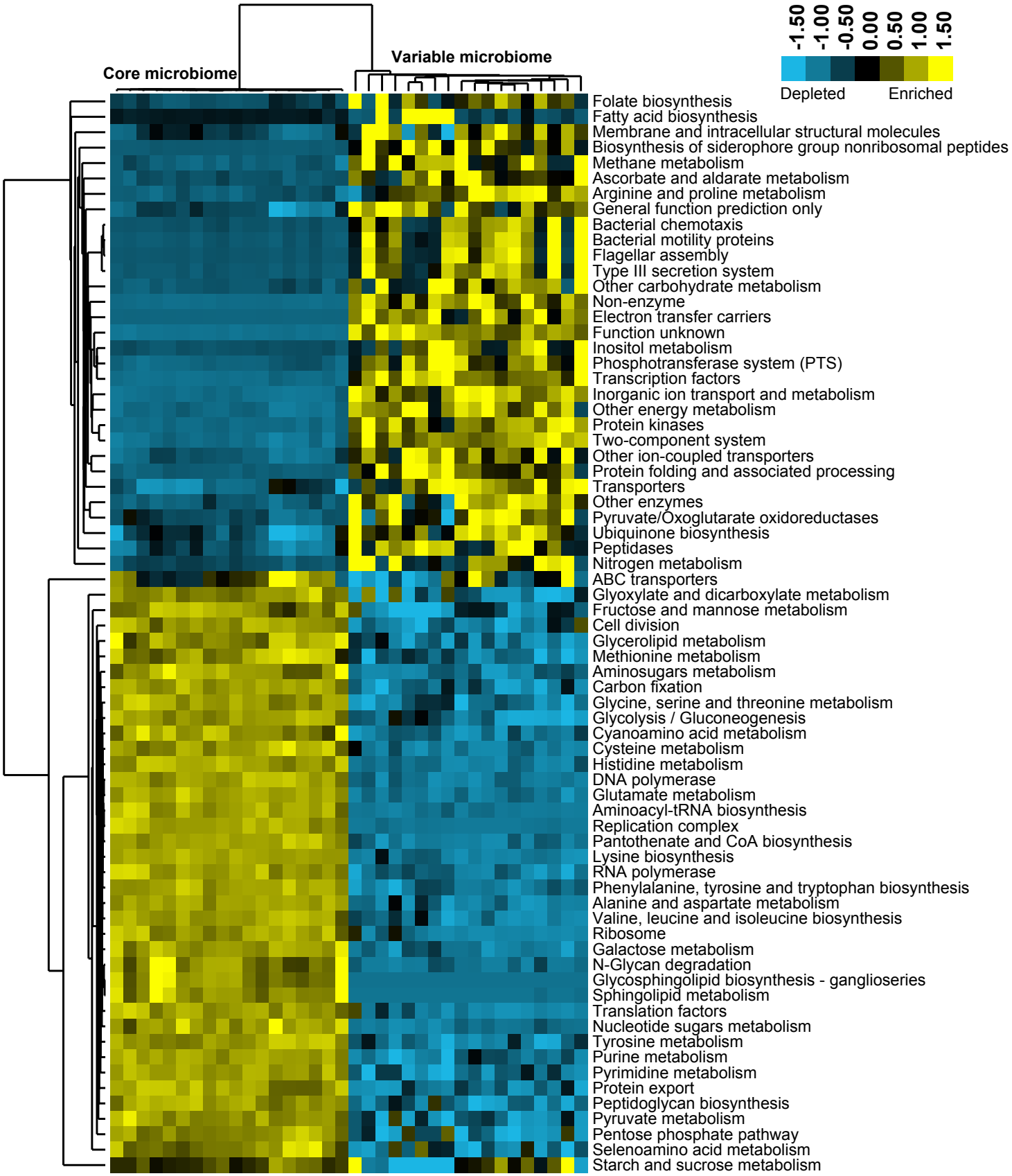

A
KEGG
orthologous groups
1.4% in 17
39.9% of groups
in 18 microbiomes
96.2% of sequences match groups in 18 microbiomes
B
CAZymes
0.9% in 17
53.2% of groups
in 18 microbiomes
98% of sequences match families in 18 microbiomes
C
STRING
orthologous groups
2.4% in 17
26.4% of groups
in 18 microbiomes
93.2% of sequences match groups in 18 microbiomes
D
Variable microbiome
Core microbiome
Number of groups (thousands)
Sequences (thousands)
0
1
2
3
4
5
6
0
50
100
150
200
250

Figure 20. Clustering of pathways enriched or depleted in the core microbiome. Sequences from each of the 18 distal gut microbiomes were binned into the 'core' or 'variable' microbiome based on the co-occurrence of KEGG orthologous groups [core groups were found in all 18 microbiomes while variable groups were present in fewer (<18) microbiomes; see **Supplementary Figure 19A**]. The frequency of each KEGG pathway was tallied for each bin and significantly different pathways were identified using a bootstrap re-sampling analysis[23] (Xipe v2.4). Pathways significantly enriched (yellow) or depleted (blue), reaching at least 0.6% relative abundance in at least two microbiomes, were clustered using single-linkage hierarchical clustering and the Pearson's correlation distance metric.


Core microbiome
Variable microbiome
-1.50
-1.00
-0.50
0.00
0.50
1.00
1.50
Depleted
Enriched
Folate biosynthesis
Fatty acid biosynthesis
Membrane and intracellular structural molecules
Biosynthesis of siderophore group nonribosomal peptides
Methane metabolism
Ascorbate and aldarate metabolism
Arginine and proline metabolism
General function prediction only
Bacterial chemotaxis
Bacterial motility proteins
Flagellar assembly
Type III secretion system
Other carbohydrate metabolism
Non-enzyme
Electron transfer carriers
Function unknown
Inositol metabolism
Phosphotransferase system (PTS)
Transcription factors
Inorganic ion transport and metabolism
Other energy metabolism
Protein kinases
Two-component system
Other ion-coupled transporters
Protein folding and associated processing
Transporters
Other enzymes
Pyruvate/Oxoglutarate oxidoreductases
Ubiquinone biosynthesis
Peptidases
Nitrogen metabolism
ABC transporters
Glyoxylate and dicarboxylate metabolism
Fructose and mannose metabolism
Cell division
Glycerolipid metabolism
Methionine metabolism
Aminosugars metabolism
Carbon fixation
Glycine, serine and threonine metabolism
Glycolysis / Gluconeogenesis
Cyanoamino acid metabolism
Cysteine metabolism
Histidine metabolism
DNA polymerase
Glutamate metabolism
Aminoacyl-tRNA biosynthesis
Replication complex
Pantothenate and CoA biosynthesis
Lysine biosynthesis
RNA polymerase
Phenylalanine, tyrosine and tryptophan biosynthesis
Alanine and aspartate metabolism
Valine, leucine and isoleucine biosynthesis
Ribosome
Galactose metabolism
N-Glycan degradation
Glycosphingolipid biosynthesis - ganglioseries
Sphingolipid metabolism
Translation factors
Nucleotide sugars metabolism
Tyrosine metabolism
Purine metabolism
Pyrimidine metabolism
Protein export
Peptidoglycan biosynthesis
Pyruvate metabolism
Pentose phosphate pathway
Selenoamino acid metabolism
Starch and sucrose metabolism

Supplementary Tables

Supplementary Table 1: V2 16S rRNA gene sequencing statistics								
Subject ID[a]	Data ID/timepoint	Family number	Twin/Mom	Ancestry	Zygosity	BMI category	Months without Antibiotics[b]	Total sequences
F1T1Le1	TS1	1	Twin	EA	MZ	Lean	>6	6415
F1T1Le2	TS1.2	1	Twin	EA	MZ	Lean	>6	1627
F1T2Le1	TS2	1	Twin	EA	MZ	Lean	NA	15495
F1T2Le2	TS2.2	1	Twin	EA	MZ	Lean	>6	1957
F1MOv1	TS3	1	Mom	EA	NA	Overweight	>6	7870
F1MOv2	TS3.2	1	Mom	EA	NA	Overweight	>6	1799
F2T1Le1	TS4	2	Twin	EA	MZ	Lean	>6	9343
F2T1Le2	TS4.2	2	Twin	EA	MZ	Lean	>6	2886
F2T2Le1	TS5	2	Twin	EA	MZ	Lean	>6	13991
F2T2Le2	TS5.2	2	Twin	EA	MZ	Lean	>6	3606
F2MOb1	TS6	2	Mom	EA	NA	Obese	>6	7717
F2MOb2	TS6.2	2	Mom	EA	NA	Obese	>6	4325
F3T1Le1	TS7	3	Twin	EA	MZ	Lean	>6	11808
F3T1Le2	TS7.2	3	Twin	EA	MZ	Lean	>6	2962
F3T2Le1	TS8	3	Twin	EA	MZ	Lean	>6	16793
F3T2Le2	TS8.2	3	Twin	EA	MZ	Lean	>6	632
F3MOv1	TS9	3	Mom	EA	NA	Overweight	>6	11291
F3MOb2	TS9.2	3	Mom	EA	NA	Obese	>6	2965
F4T1Ob1	TS10	4	Twin	AA	MZ	Obese	>6	2280
F4T1Ob2	TS10.2	4	Twin	AA	MZ	Obese	>6	979
F4T2Ob1	TS11	4	Twin	AA	MZ	Obese	>6	2458
F4T2Ob2	TS11.2	4	Twin	AA	MZ	Obese	>6	2437
F4MOb1	TS12	4	Mom	AA	NA	Obese	>1	2086
F4MOb2	TS12.2	4	Mom	AA	NA	Obese	>2	1692
F5T1Le1	TS13	5	Twin	EA	MZ	Lean	>6	8509
F5T1Le2	TS13.2	5	Twin	EA	MZ	Lean	>6	1689
F5T2Le1	TS14	5	Twin	EA	MZ	Lean	>6	15903
F5MOv1	TS15	5	Mom	EA	NA	Overweight	>6	15690
F5MOv2	TS15.2	5	Mom	EA	NA	Overweight	>6	3967
F6T1Le1	TS16	6	Twin	EA	MZ	Lean	NA	5975
F6T2Le1	TS17	6	Twin	EA	MZ	Lean	>6	1182
F7T1Ob1	TS19	7	Twin	EA	MZ	Obese	>6	21469
F7T1Ob2	TS19.2	7	Twin	EA	MZ	Obese	>6	3953
F7T2Ob1	TS20	7	Twin	EA	MZ	Obese	>6	32871
F7T2Ob2	TS20.2	7	Twin	EA	MZ	Obese	>6	5045
F7MOb1	TS21	7	Mom	EA	NA	Obese	>6	26781
F7MOb2	TS21.2	7	Mom	EA	NA	Obese	>6	4752
F8T1Le1	TS22	8	Twin	EA	MZ	Lean	>6	5110
F8T2Le1	TS23	8	Twin	EA	MZ	Lean	>6	1978
F9T1Le1	TS25	9	Twin	EA	MZ	Lean	>6	10017
F9T1Le2	TS25.2	9	Twin	EA	MZ	Lean	>6	4626
F9T2Le1	TS26	9	Twin	EA	MZ	Lean	>6	16757
F9T2Le2	TS26.2	9	Twin	EA	MZ	Lean	>6	5111
F9MOb1	TS27	9	Mom	EA	NA	Obese	>6	11885
F9MOb2	TS27.2	9	Mom	EA	NA	Obese	>6	2068
F10T1Ob1	TS28	10	Twin	EA	MZ	Obese	>6	6694
F10T2Ob1	TS29	10	Twin	EA	MZ	Obese	>6	2411
F10MOv1	TS30	10	Mom	EA	NA	Overweight	>6	8273
F10MLe2	TS30.2	10	Mom	EA	NA	Lean	>6	3280
F11T1Le1	TS31	11	Twin	EA	MZ	Lean	>6	18941
F11T1Le2	TS31.2	11	Twin	EA	MZ	Lean	>6	5842
F11T2Le1	TS32	11	Twin	EA	MZ	Lean	>6	9773
F11T2Le2	TS32.2	11	Twin	EA	MZ	Lean	>6	6178
F11MOv1	TS33	11	Mom	EA	NA	Overweight	>6	18037
F11MOv2	TS33.2	11	Mom	EA	NA	Overweight	>6	1593
F12T1Ob1	TS34	12	Twin	EA	MZ	Obese	>6	1730
F12T2Ob1	TS35	12	Twin	EA	MZ	Obese	>6	3887
F13T1Ob1	TS37	13	Twin	EA	MZ	Obese	>6	3534
F13T1Ob2	TS37.2	13	Twin	EA	MZ	Obese	>6	4458
F13T2Ov1	TS38	13	Twin	EA	MZ	Overweight	>6	3043
F13T2Ov2	TS38.2	13	Twin	EA	MZ	Overweight	>6	2566

F13MOb1	TS39	13	Mom	EA	NA	Obese	>6	5848
F13MOb2	TS39.2	13	Mom	EA	NA	Obese	>6	2146
F14T1Ob1	TS43	14	Twin	EA	MZ	Obese	>6	2905
F14T2Ob1	TS44	14	Twin	EA	MZ	Obese	>6	1621
F15T1Ob1	TS49	15	Twin	EA	MZ	Obese	>6	11936
F15T1Ob2	TS49.2	15	Twin	EA	MZ	Obese	>6	4220
F15T2Ob1	TS50	15	Twin	EA	MZ	Obese	>6	12672
F15T2Ob2	TS50.2	15	Twin	EA	MZ	Obese	>6	4603
F15MOb1	TS51	15	Mom	EA	NA	Obese	>6	13789
F15MOb2	TS51.2	15	Mom	EA	NA	Obese	>6	3284
F16T1Ob1	TS55	16	Twin	EA	DZ	Obese	>6	3817
F16T1Ob2	TS55.2	16	Twin	EA	DZ	Obese	>6	5210
F16T2Ob1	TS56	16	Twin	EA	DZ	Obese	>6	5147
F16T2Ob2	TS56.2	16	Twin	EA	DZ	Obese	>6	4490
F16MOb1	TS57	16	Mom	EA	NA	Obese	>0	8440
F16MOb2	TS57.2	16	Mom	EA	NA	Obese	>1	2365
F17T1Ob1	TS61	17	Twin	EA	DZ	Obese	>6	672
F17T1Ob2	TS61.2	17	Twin	EA	DZ	Obese	>6	3738
F17T2Ob1	TS62	17	Twin	EA	DZ	Obese	>6	2311
F17T2Ob2	TS62.2	17	Twin	EA	DZ	Obese	>6	3821
F17MOb1	TS63	17	Mom	EA	NA	Obese	>6	2132
F17MOb2	TS63.2	17	Mom	EA	NA	Obese	>6	1853
F18T1Ov1	TS64	18	Twin	EA	MZ	Overweight	>6	4571
F18T1Ov2	TS64.2	18	Twin	EA	MZ	Overweight	>6	4523
F18T2Ob1	TS65	18	Twin	EA	MZ	Obese	>6	2502
F18T2Ob2	TS65.2	18	Twin	EA	MZ	Obese	>6	3943
F18MOb1	TS66	18	Mom	EA	NA	Obese	>6	3491
F18MOb2	TS66.2	18	Mom	EA	NA	Obese	>6	6187
F19T1Ob1	TS67	19	Twin	EA	DZ	Obese	NA	988
F19T1Ob2	TS67.2	19	Twin	EA	DZ	Obese	NA	1861
F19T2Ob1	TS68	19	Twin	EA	DZ	Obese	>6	3870
F19T2Ob2	TS68.2	19	Twin	EA	DZ	Obese	>6	2242
F19MOb1	TS69	19	Mom	EA	NA	Obese	>6	5290
F19MOb2	TS69.2	19	Mom	EA	NA	Obese	>0	2305
F20T1Ob1	TS70	20	Twin	EA	DZ	Obese	>6	2139
F20T1Ob2	TS70.2	20	Twin	EA	DZ	Obese	>6	2166
F20T2Ob1	TS71	20	Twin	EA	DZ	Obese	>6	3130
F20T2Ob2	TS71.2	20	Twin	EA	DZ	Obese	>6	2293
F20MOb1	TS72	20	Mom	EA	NA	Obese	>6	1674
F20MOb2	TS72.2	20	Mom	EA	NA	Obese	>6	379
F21T1Ob1	TS73	21	Twin	EA	DZ	Obese	>6	2963
F21T2Ob1	TS74	21	Twin	EA	DZ	Obese	>6	2177
F21T2Ob2	TS74.2	21	Twin	EA	DZ	Obese	>6	1791
F21MOb1	TS75	21	Mom	EA	NA	Obese	>6	1434
F21MOb2	TS75.2	21	Mom	EA	NA	Obese	>6	1887
F22T1Ob1	TS76	22	Twin	AA	MZ	Obese	>6	2977
F22T1Ob2	TS76.2	22	Twin	AA	MZ	Obese	>6	1962
F22T2Ov1	TS77	22	Twin	AA	MZ	Overweight	>6	2168
F22MOb1	TS78	22	Mom	AA	NA	Obese	>6	1460
F22MOb2	TS78.2	22	Mom	AA	NA	Obese	>6	2482
F23T1Ob1	TS82	23	Twin	AA	MZ	Obese	>6	1628
F23T1Ob2	TS82.2	23	Twin	AA	MZ	Obese	>6	1673
F23T2Ob1	TS83	23	Twin	AA	MZ	Obese	>6	1572
F23T2Ob2	TS83.2	23	Twin	AA	MZ	Obese	>6	3349
F23MOb1	TS84	23	Mom	AA	NA	Obese	>6	2215
F23MOb2	TS84.2	23	Mom	AA	NA	Obese	>6	2033
F24T1Ov1	TS85	24	Twin	EA	DZ	Overweight	>3	2385
F24T1Ov2	TS85.2	24	Twin	EA	DZ	Overweight	>6	2122
F24T2Ob1	TS86	24	Twin	EA	DZ	Obese	>1	4107
F24T2Ob2	TS86.2	24	Twin	EA	DZ	Obese	>3	1704
F24MOb1	TS87	24	Mom	EA	NA	Obese	>6	2605
F24MOb2	TS87.2	24	Mom	EA	NA	Obese	>6	1587
F25T1Ob1	TS88	25	Twin	EA	DZ	Obese	>4	2497
F25T1Ob2	TS88.2	25	Twin	EA	DZ	Obese	>6	2129
F25T2Ob1	TS89	25	Twin	EA	DZ	Obese	>6	2108

F25T2Ob2	TS89.2	25	Twin	EA	DZ	Obese	>6	3549
F25MOb1	TS90	25	Mom	EA	NA	Obese	>6	2615
F25MOb2	TS90.2	25	Mom	EA	NA	Obese	>6	2725
F26T1Ob1	TS91	26	Twin	AA	MZ	Obese	>0	675
F26T1Ob2	TS91.2	26	Twin	AA	MZ	Obese	>6	2307
F26T2Ob1	TS92	26	Twin	AA	MZ	Obese	>6	2036
F26T2Ob2	TS92.2	26	Twin	AA	MZ	Obese	>6	2335
F27T1Ob1	TS94	27	Twin	AA	MZ	Obese	>6	1861
F27T1Ob2	TS94.2	27	Twin	AA	MZ	Obese	>6	2511
F27T2Ob1	TS95	27	Twin	AA	MZ	Obese	>6	2842
F27T2Ob2	TS95.2	27	Twin	AA	MZ	Obese	>6	2550
F27MOb1	TS96	27	Mom	AA	NA	Obese	>6	1516
F27MOb2	TS96.2	27	Mom	AA	NA	Obese	>6	2909
F28T1Ob1	TS97	28	Twin	AA	DZ	Obese	>6	2326
F28T1Ob2	TS97.2	28	Twin	AA	DZ	Obese	>6	2944
F28T2Ob1	TS98	28	Twin	AA	DZ	Obese	>6	2970
F28T2Ob2	TS98.2	28	Twin	AA	DZ	Obese	>6	2851
F28MOv2	TS99.2	28	Mom	AA	NA	Overweight	>6	3136
F29T1Ob1	TS100	29	Twin	AA	MZ	Obese	>6	3504
F29T1Ob2	TS100.2	29	Twin	AA	MZ	Obese	>6	2616
F29T2Ob2	TS101.2	29	Twin	AA	MZ	Obese	>6	2387
F30T1Ob1	TS103	30	Twin	AA	MZ	Obese	>6	1473
F30T1Ob2	TS103.2	30	Twin	AA	MZ	Obese	>6	3012
F30T2Ob1	TS104	30	Twin	AA	MZ	Obese	>6	1970
F30T2Ob2	TS104.2	30	Twin	AA	MZ	Obese	>6	2895
F30MOb1	TS105	30	Mom	AA	NA	Obese	>6	1864
F30MOb2	TS105.2	30	Mom	AA	NA	Obese	>6	2096
F31T1Ob1	TS106	31	Twin	AA	MZ	Obese	>6	2698
F31T1Ob2	TS106.2	31	Twin	AA	MZ	Obese	>6	2250
F31T2Ob1	TS107	31	Twin	AA	MZ	Obese	>6	3132
F31T2Ob2	TS107.2	31	Twin	AA	MZ	Obese	>6	4521
F32T1Le1	TS109	32	Twin	EA	DZ	Lean	>6	2583
F32T1Le2	TS109.2	32	Twin	EA	DZ	Lean	>6	1682
F32T2Le1	TS110	32	Twin	EA	DZ	Lean	>6	2286
F32T2Le2	TS110.2	32	Twin	EA	DZ	Lean	>6	4408
F32MLe1	TS111	32	Mom	EA	NA	Lean	>6	3822
F32MLe2	TS111.2	32	Mom	EA	NA	Lean	>6	2597
F33T1Ob1	TS115	33	Twin	AA	MZ	Obese	>6	2619
F33T1Ob2	TS115.2	33	Twin	AA	MZ	Obese	>6	2017
F33T2Ob1	TS116	33	Twin	AA	MZ	Obese	>6	5558
F33T2Ob2	TS116.2	33	Twin	AA	MZ	Obese	>6	2440
F33MOb1	TS117	33	Mom	AA	NA	Obese	>6	3430
F33MOb2	TS117.2	33	Mom	AA	NA	Obese	>6	2932
F34T1Ob1	TS118	34	Twin	AA	DZ	Obese	>0	2209
F34T1Ob2	TS118.2	34	Twin	AA	DZ	Obese	>6	3030
F34T2Ob1	TS119	34	Twin	AA	DZ	Obese	>6	2791
F34T2Ob2	TS119.2	34	Twin	AA	DZ	Obese	>0	3828
F34MOb1	TS120	34	Mom	AA	NA	Obese	>6	97
F34MOb2	TS120.2	34	Mom	AA	NA	Obese	>6	3015
F35T1Le1	TS124	35	Twin	EA	DZ	Lean	>6	2336
F35T1Le2	TS124.2	35	Twin	EA	DZ	Lean	>6	2102
F35T2Ov1	TS125	35	Twin	EA	DZ	Overweight	>6	2381
F35T2Ov2	TS125.2	35	Twin	EA	DZ	Overweight	>6	1889
F35MOb1	TS126	35	Mom	EA	NA	Obese	>6	1733
F35MOb2	TS126.2	35	Mom	EA	NA	Obese	>6	2676
F36T1Le1	TS127	36	Twin	EA	DZ	Lean	>6	4119
F36T1Le2	TS127.2	36	Twin	EA	DZ	Lean	>6	1929
F36T2Le1	TS128	36	Twin	EA	DZ	Lean	>6	4698
F36T2Le2	TS128.2	36	Twin	EA	DZ	Lean	>6	2857
F36MLe1	TS129	36	Mom	EA	NA	Lean	>6	2628
F36MLe2	TS129.2	36	Mom	EA	NA	Lean	>6	2247
F37T1Ob1	TS130	37	Twin	AA	MZ	Obese	>6	3121
F37T1Ob2	TS130.2	37	Twin	AA	MZ	Obese	>1	3391
F37T2Ob1	TS131	37	Twin	AA	MZ	Obese	>6	3338
F37T2Ob2	TS131.2	37	Twin	AA	MZ	Obese	NA	3186

F37MOb1	TS132	37	Mom	AA	NA	Obese	>1	2586
F37MOb2	TS132.2	37	Mom	AA	NA	Obese	NA	4130
F38T1Ob1	TS133	38	Twin	AA	MZ	Obese	>6	2355
F38T1Ob2	TS133.2	38	Twin	AA	MZ	Obese	>6	3902
F38T2Ob1	TS134	38	Twin	AA	MZ	Obese	>3	1378
F38T2Ob2	TS134.2	38	Twin	AA	MZ	Obese	>5	2656
F38MOb1	TS135	38	Mom	AA	NA	Obese	>6	3068
F38MOb2	TS135.2	38	Mom	AA	NA	Obese	>6	2436
F39T1Ov1	TS136	39	Twin	AA	DZ	Overweight	>6	2962
F39T1Ob2	TS136.2	39	Twin	AA	DZ	Obese	>6	4164
F39T2Ob1	TS137	39	Twin	AA	DZ	Obese	>6	3748
F39T2Ob2	TS137.2	39	Twin	AA	DZ	Obese	>0	2902
F39MOb1	TS138	39	Mom	AA	NA	Obese	>6	3289
F39MOb2	TS138.2	39	Mom	AA	NA	Obese	>6	1369
F40T1Ob1	TS139	40	Twin	AA	DZ	Obese	>6	2756
F40T1Ob2	TS139.2	40	Twin	AA	DZ	Obese	>6	3195
F40T2Ob1	TS140	40	Twin	AA	DZ	Obese	>6	2698
F40T2Ob2	TS140.2	40	Twin	AA	DZ	Obese	>6	2851
F40MOb1	TS141	40	Mom	AA	NA	Obese	>6	2083
F40MOb2	TS141.2	40	Mom	AA	NA	Obese	>6	3125
F41T1Ob1	TS142	41	Twin	AA	DZ	Obese	>6	2432
F41T1Ob2	TS142.2	41	Twin	AA	DZ	Obese	>0	3466
F41T2Ob1	TS143	41	Twin	AA	DZ	Obese	>6	3944
F41T2Ob2	TS143.2	41	Twin	AA	DZ	Obese	>6	3721
F41MOb1	TS144	41	Mom	AA	NA	Obese	>6	2804
F41MOb2	TS144.2	41	Mom	AA	NA	Obese	>6	4354
F42T1Ob1	TS145	42	Twin	AA	DZ	Obese	>0	2738
F42T1Ob2	TS145.2	42	Twin	AA	DZ	Obese	>1	3633
F42T2Ob1	TS146	42	Twin	AA	DZ	Obese	>0	3214
F42T2Ob2	TS146.2	42	Twin	AA	DZ	Obese	>1	3380
F42MOb1	TS147	42	Mom	AA	NA	Obese	>2	3513
F42MOv2	TS147.2	42	Mom	AA	NA	Overweight	>4	4957
F43T1Ob1	TS148	43	Twin	EA	MZ	Obese	>6	6128
F43T2Ob1	TS149	43	Twin	EA	MZ	Obese	>5	11555
F43MOb1	TS150	43	Mom	EA	NA	Obese	>6	8045
F44T1Ob1	TS151	44	Twin	AA	DZ	Obese	>6	3800
F44T1Ob2	TS151.2	44	Twin	AA	DZ	Obese	>6	3210
F44T2Ob1	TS152	44	Twin	AA	DZ	Obese	>6	3326
F44T2Ob2	TS152.2	44	Twin	AA	DZ	Obese	>6	2742
F44MOv1	TS153	44	Mom	AA	NA	Overweight	>6	4118
F45T1Le2	TS154.2	45	Twin	AA	MZ	Lean	>6	1466
F45T2Le1	TS155	45	Twin	AA	MZ	Lean	>6	2267
F45T2Le2	TS155.2	45	Twin	AA	MZ	Lean	>6	2361
F45MOb1	TS156	45	Mom	AA	NA	Obese	>2	1694
F45MOb2	TS156.2	45	Mom	AA	NA	Obese	>6	1906
F46T1Ob1	TS160	46	Twin	AA	DZ	Obese	>6	2367
F46T1Ob2	TS160.2	46	Twin	AA	DZ	Obese	>6	2049
F46T2Ob1	TS161	46	Twin	AA	DZ	Obese	>6	2185
F46MOb1	TS162	46	Mom	AA	NA	Obese	>6	3564
F46MOb2	TS162.2	46	Mom	AA	NA	Obese	>6	4041
F47T1Le1	TS163	47	Twin	AA	MZ	Lean	>2	1624
F47T1Le2	TS163.2	47	Twin	AA	MZ	Lean	>3	2495
F47T2Le1	TS164	47	Twin	AA	MZ	Lean	>6	2651
F47T2Le2	TS164.2	47	Twin	AA	MZ	Lean	>6	3018
F47MLe1	TS165	47	Mom	AA	NA	Lean	>6	2767
F47MLe2	TS165.2	47	Mom	AA	NA	Lean	>6	2839
F48T1Ob1	TS166	48	Twin	AA	DZ	Obese	>2	3628
F48T1Ob2	TS166.2	48	Twin	AA	DZ	Obese	>6	3252
F48T2Ob1	TS167	48	Twin	AA	DZ	Obese	>6	2822
F48T2Ob2	TS167.2	48	Twin	AA	DZ	Obese	>6	4538
F48MOb1	TS168	48	Mom	AA	NA	Obese	>6	2882
F48MOb2	TS168.2	48	Mom	AA	NA	Obese	>6	4569
F49T1Ob1	TS169	49	Twin	AA	DZ	Obese	>6	4217
F49T1Ob2	TS169.2	49	Twin	AA	DZ	Obese	>6	3644
F49T2Ob1	TS170	49	Twin	AA	DZ	Obese	>3	2117

F49T2Ob2	TS170.2	49	Twin	AA	DZ	Obese	>6	2785
F50T1Ob1	TS178	50	Twin	AA	DZ	Obese	>6	2378
F50T1Ob2	TS178.2	50	Twin	AA	DZ	Obese	>6	2894
F50T2Ob1	TS179	50	Twin	AA	DZ	Obese	>6	2122
F50T2Ob2	TS179.2	50	Twin	AA	DZ	Obese	>6	3189
F50MLe1	TS180	50	Mom	AA	NA	Lean	>6	2132
F51T1Ob1	TS181	51	Twin	AA	DZ	Obese	>3	3455
F51T1Ob2	TS181.2	51	Twin	AA	DZ	Obese	>6	2812
F51T2Ov1	TS182	51	Twin	AA	DZ	Overweight	>6	7014
F51T2Ob2	TS182.2	51	Twin	AA	DZ	Obese	>6	6903
F51MOb1	TS183	51	Mom	AA	NA	Obese	>2	3243
F51MOb2	TS183.2	51	Mom	AA	NA	Obese	>6	2884
F52T1Le1	TS184	52	Twin	AA	MZ	Lean	>6	1925
F52T2Le1	TS185	52	Twin	AA	MZ	Lean	>6	2545
F52T2Le2	TS185.2	52	Twin	AA	MZ	Lean	>2	2538
F52MOv1	TS186	52	Mom	AA	NA	Overweight	>6	1735
F53T1Ob1	TS190	53	Twin	AA	MZ	Obese	NA	3165
F53T2Ob1	TS191	53	Twin	AA	MZ	Obese	>6	2720
F53MOv1	TS192	53	Mom	AA	NA	Overweight	>6	5067
F54T1Le1	TS193	54	Twin	EA	DZ	Lean	>6	1799
F54T1Le2	TS193.2	54	Twin	EA	DZ	Lean	>6	1739
F54T2Le1	TS194	54	Twin	EA	DZ	Lean	>6	2291
F54T2Le2	TS194.2	54	Twin	EA	DZ	Lean	>6	1612
F54MLe1	TS195	54	Mom	EA	NA	Lean	>6	2782
F54MLe2	TS195.2	54	Mom	EA	NA	Lean	>6	2462
						TOTAL		**1119519**

[a]ID nomenclature: Family number, Twin number or mother, and BMI category (Le=lean, Ov=overweight, Ob=obese; e.g. F1T1Le stands for family 1, twin 1, lean)

[b]Minimum number of months between sampling date and last administration of antibiotics

Supplementary Table 2: V6 16S rRNA gene sequencing statistics					
Subject ID[a]	**Data ID**	**Twin/Mom**	**Family**	**BMI**	**Sequences**
F1T1Le1	TS1	Twin	1	Lean	25,140
F1T2Le1	TS2	Twin	1	Lean	42,186
F1MOv1	TS3	Mom	1	Overweight	17,726
F2T1Le1	TS4	Twin	2	Lean	25,705
F2T2Le1	TS5	Twin	2	Lean	26,608
F2MOb1	TS6	Mom	2	Obese	27,007
F3T1Le1	TS7	Twin	3	Lean	17,469
F3T2Le1	TS8	Twin	3	Lean	17,170
F3MOv1	TS9	Mom	3	Overweight	14,787
F5T1Le1	TS13	Twin	5	Lean	15,296
F5T2Le1	TS14	Twin	5	Lean	14,220
F5MOv1	TS15	Mom	5	Overweight	14,244
F7T1Ob1	TS19	Twin	7	Obese	43,635
F7T2Ob1	TS20	Twin	7	Obese	13,476
F7MOb1	TS21	Mom	7	Obese	23,714
F9T1Le1	TS25	Twin	9	Lean	20,491
F9T2Le1	TS26	Twin	9	Lean	27,626
F9MOb1	TS27	Mom	9	Obese	25,494
F10T1Ob1	TS28	Twin	10	Obese	20,905
F10T2Ob1	TS29	Twin	10	Obese	15,698
F10MOv1	TS30	Mom	10	Overweight	32,083
F11T1Le1	TS31	Twin	11	Lean	16,530
F11T2Le1	TS32	Twin	11	Lean	31,690
F11MOv1	TS33	Mom	11	Overweight	28,962
F15T1Ob1	TS49	Twin	15	Obese	22,201
F15T2Ob1	TS50	Twin	15	Obese	30,498
F15MOb1	TS51	Mom	15	Obese	22,691
F16T1Ob1	TS55	Twin	16	Obese	37,027
F16T2Ob1	TS56	Twin	16	Obese	31,512
F16MOb1	TS57	Mom	16	Obese	30,392
F43T1Ob1	TS148	Twin	43	Obese	26,458
F43T2Ob1	TS149	Twin	43	Obese	35,838
F43MOb1	TS150	Mom	43	Obese	23,463
				TOTAL	817,942

[a]ID nomenclature: Family number, Twin number or mother, and BMI category (Le=lean, Ov=overweight, Ob=obese; e.g. F1T1Le stands for family 1, twin 1, lean)

Supplementary Table 3: Full-length 16S rRNA gene sequencing statistics					
Subject ID[a]	**Data ID**	**Twin/Mom**	**Family**	**BMI**	**Sequences**
F1T1Le1	TS1	Twin	1	Lean	349
F1T2Le1	TS2	Twin	1	Lean	351
F1MOv1	TS3	Mom	1	Overweight	331
F2T1Le1	TS4	Twin	2	Lean	351
F2T2Le1	TS5	Twin	2	Lean	345
F2MOb1	TS6	Mom	2	Obese	348
F3T1Le1	TS7	Twin	3	Lean	237
F3T2Le1	TS8	Twin	3	Lean	354
F3MOv1	TS9	Mom	3	Overweight	357
F5T1Le1	TS13	Twin	5	Lean	337
F5T2Le1	TS14	Twin	5	Lean	350
F5MOv1	TS15	Mom	5	Overweight	338
F7T1Ob1	TS19	Twin	7	Obese	333
F7T2Ob1	TS20	Twin	7	Obese	340
F7MOb1	TS21	Mom	7	Obese	332
F9T1Le1	TS25	Twin	9	Lean	351
F9T2Le1	TS26	Twin	9	Lean	252
F9MOb1	TS27	Mom	9	Obese	343
F10T1Ob1	TS28	Twin	10	Obese	344
F10T2Ob1	TS29	Twin	10	Obese	337
F10MOv1	TS30	Mom	10	Overweight	261
F15T1Ob1	TS49	Twin	15	Obese	338
F15T2Ob1	TS50	Twin	15	Obese	319
F15MOb1	TS51	Mom	15	Obese	331
F16T1Ob1	TS55	Twin	16	Obese	353
F16T2Ob1	TS56	Twin	16	Obese	278
F16MOb1	TS57	Mom	16	Obese	348
F43T1Ob1	TS148	Twin	43	Obese	323
F43T2Ob1	TS149	Twin	43	Obese	340
F43MOb1	TS150	Mom	43	Obese	349
				TOTAL	9,920

[a]ID nomenclature: Family number, Twin number or mother, and BMI category (Le=lean, Ov=overweight, Ob=obese; e.g. F1T1Le stands for family 1, twin 1, lean)

Supplementary Table 4: Microbiome sequencing statistics									
Subject ID[a]	Data ID	Twin/Mom	Family	BMI	Platform	Total nt	Number reads	Filtered reads[b]	16S rRNA gene fragments[c]
F1T1Le1	TS1	Twin	1	Lean	FLX	60,016,519	254,044	217,386	439
F1T2Le1	TS2	Twin	1	Lean	FLX	90,271,969	514,022	443,640	512
F1MOv1	TS3	Mom	1	Overweight	FLX	113,506,401	571,301	510,972	723
F2T1Le1	TS4	Twin	2	Lean	FLX	107,008,761	472,154	414,754	626
F2T2Le1	TS5	Twin	2	Lean	FLX	112,835,879	553,142	490,776	928
F2MOb1	TS6	Mom	2	Obese	FLX	135,976,476	623,027	535,763	1,039
F3T1Le1	TS7	Twin	3	Lean	FLX	146,946,832	607,386	555,853	1,188
F3T2Le1	TS8	Twin	3	Lean	FLX	113,177,766	468,769	414,497	976
F3MOv1	TS9	Mom	3	Overweight	FLX	137,564,473	552,870	499,499	934
F7T1Ob1	TS19	Twin	7	Obese	FLX	95,538,760	583,989	498,880	569
F7T2Ob1	TS20	Twin	7	Obese	FLX	108,342,331	550,695	495,040	829
F7MOb1	TS21	Mom	7	Obese	FLX	95,960,723	451,177	413,772	774
F10T1Ob1	TS28	Twin	10	Obese	Titanium	138,364,927	399,717	302,780	652
F10T2Ob1	TS29	Twin	10	Obese	Titanium	239,971,702	672,196	502,399	1,190
F10MOv1	TS30	Mom	10	Overweight	FLX	105,932,316	564,184	495,865	791
F15T1Ob1	TS49	Twin	15	Obese	FLX	104,449,087	596,149	519,072	769
F15T2Ob1	TS50	Twin	15	Obese	FLX	129,037,456	642,191	549,700	1,209
F15MOb1	TS51	Mom	15	Obese	FLX	101,531,105	557,165	434,187	582
					SUM	2,136,433,483	9,634,178	8,294,835	14,730

[a]ID nomenclature: Family number, Twin number or mom, and BMI category (Le=lean, Ov=overweight, Ob=obese; e.g. F1T1Le stands for family 1, twin 1, lean)

[b]Sequences used after removing low quality, duplicate, and human sequences

[c]16S rRNA gene fragments identified in microbiome sequencing reads

Supplementary Table 5: Microbiome BLAST statistics[a]												
Subject ID	**Data ID**	**Raw Reads**	**Reads Used**	**%Sequences used**	**Nucleotides used**	**Mean read-length**	**%Hsa**	**%RDP**	**%KEGG**	**%STRING**	**%NR**	**%Gut**
F1T1Le1	TS1	254,044	217,386	85.6	51,708,794	237.9	0.42	0.21	29.1	34.5	54.9	57.9
F1T2Le1	TS2	514,022	443,640	86.3	78,853,892	177.7	0.08	0.12	20.3	28.7	46.9	51.7
F1MOv1	TS3	571,301	510,972	89.4	102,717,417	201.0	0.16	0.15	23.8	33.6	56.5	61.2
F2T1Le1	TS4	472,154	414,754	87.8	95,003,113	229.1	0.14	0.15	26.2	44.5	72.3	74.9
F2T2Le1	TS5	553,142	490,776	88.7	100,599,979	205.0	0.22	0.19	23.0	27.8	54.1	62.1
F2MOb1	TS6	623,027	535,763	86.0	118,207,161	220.6	0.62	0.20	26.9	37.2	58.9	62.1
F3T1Le1	TS7	607,386	555,853	91.5	134,889,015	242.7	0.13	0.22	26.9	34.0	58.4	61.7
F3T2Le1	TS8	468,769	414,497	88.4	100,520,072	242.5	0.20	0.24	28.5	35.7	61.1	64.4
F3MOv1	TS9	552,870	499,499	90.3	124,768,172	249.8	0.14	0.19	26.8	36.6	63.2	66.3
F7T1Ob1	TS19	583,989	498,880	85.4	82,117,565	164.6	0.06	0.12	19.1	30.6	52.9	57.1
F7T2Ob1	TS20	550,695	495,040	89.9	98,053,098	198.1	0.32	0.17	22.3	29.3	47.2	49.9
F7MOb1	TS21	451,177	413,772	91.7	88,786,017	214.6	0.09	0.19	25.5	37.6	62.8	66.3
F10T1Ob1	TS28	399,717	302,780	75.7	101,434,082	335.0	0.06	0.36	24.5	28.4	53.2	55.5
F10T2Ob1	TS29	672,196	502,399	74.7	173,386,030	345.1	0.11	0.29	27.5	34.8	63.2	63.9
F10MOv1	TS30	564,184	495,865	87.9	94,405,318	190.4	0.21	0.16	22.4	32.0	54.7	60.7
F15T1Ob1	TS49	596,149	519,072	87.1	91,987,878	177.2	0.29	0.15	18.6	23.0	43.7	46.4
F15T2Ob1	TS50	642,191	549,700	85.6	111,999,603	203.7	0.24	0.22	24.6	29.4	51.9	57.9
F15MOb1	TS51	557,165	434,187	77.9	81,330,211	187.3	0.40	0.14	21.0	26.3	44.2	43.9
Average		**535,232**	**460,824**	**86.1**	**101,709,301**	**223.5**	**0.22**	**0.19**	**24.3**	**32.5**	**55.6**	**59.1**
Sum		9,634,178	8,294,835	-	1,830,767,417	-	-	-	-	-	-	-

[a]Key: %Sequences used=percentage of sequences remaining after removing low quality, duplicate, and human sequences; Hsa=reads matching the H.sapiens genome; %RDP=percentage of reads matching the RDP 16S rRNA database; %KEGG, %STRING, %NR=percentage of reads that were assignable to entries in these various databases; %Gut=percentage of reads assigned to the database of 44 reference genomes

Supplementary Table 6: Phylotypes shared across ≥70% of all individuals (V2 dataset; 1,000 random sequences/individual)[a]							
Phylotype ID	Individuals with phylotype	% of individuals with phylotype	Number of reads grouped into phylotype	Highest relative abundance across all individuals	Lowest relative abundance across all individuals	Mean±sem % of 16S rRNA gene sequences across all individuals	Taxonomic classification[b]
1	151	98.1	7942	28.7	0	6.53 ± 0.41	Bacteria; Firmicutes; Clostridia; Faecalibacterium
2	151	98.1	5375	25.5	0	4.41 ± 0.34	Bacteria; Firmicutes; Clostridia; Clostridiales; Ruminococcus
3	144	93.5	2518	14.7	0	2.06 ± 0.16	Bacteria; Firmicutes; Clostridia; Clostridiales
4	143	92.9	5606	30.5	0	4.56 ± 0.41	Bacteria; Firmicutes; Clostridia; Clostridiales; Eubacterium_rectale
5	140	90.9	1629	8.1	0	1.34 ± 0.11	Bacteria; Firmicutes; Clostridia; Clostridiales; Clostridium_clostridioforme
6	134	87.0	757	12.7	0	0.62 ± 0.09	Bacteria; Firmicutes; Clostridia; Clostridiales; Ruminococcus; Ruminococcus_schinkii
7	133	86.4	1485	12.2	0	1.23 ± 0.14	Bacteria; Firmicutes; Clostridia; Clostridiales; Coprococcus
8	133	86.4	1392	6.5	0	1.14 ± 0.10	Bacteria; Firmicutes; Clostridia; Clostridiales
9	133	86.4	1201	10.5	0	0.99 ± 0.12	Bacteria; Firmicutes; Clostridia; Clostridiales; Ruminococcus
10	128	83.1	819	5.2	0	0.68 ± 0.06	Bacteria; Firmicutes; Clostridia; Clostridiales
11	127	82.5	747	3.7	0	0.62 ± 0.05	Bacteria; Firmicutes; Clostridia; Faecalibacterium
12	126	81.8	11598	51.6	0	9.39 ± 0.79	Bacteria; Bacteroidetes; Bacteroidales; Bacteroidaceae
13	125	81.2	2585	34.3	0	2.15 ± 0.31	Bacteria; Firmicutes; Clostridia; Faecalibacterium
14	123	79.9	3512	15.3	0	2.89 ± 0.25	Bacteria; Firmicutes; Clostridia; Faecalibacterium
15	120	77.9	792	8.4	0	0.66 ± 0.08	Bacteria; Firmicutes; Clostridia; Clostridiales; Clostridium_nexile
16	118	76.6	632	2.7	0	0.52 ± 0.05	Bacteria; Firmicutes; Clostridia; Faecalibacterium
17	115	74.7	3422	43.3	0	2.79 ± 0.41	Bacteria; Bacteroidetes; Bacteroidales; Bacteroidaceae
18	113	73.4	441	2.3	0	0.37 ± 0.03	Bacteria; Firmicutes; Clostridia; Clostridiales; Clostridium_nexile
19	112	72.7	1168	17.4	0	0.98 ± 0.16	Bacteria; Firmicutes; Clostridia; Clostridiales; Ruminococcus
20	111	72.1	749	5.2	0	0.61 ± 0.07	Bacteria; Firmicutes; Clostridia; Clostridiales
21	108	70.1	640	3.5	0	0.53 ± 0.06	Bacteria; Firmicutes; Clostridia; Clostridiales; Ruminococcus

[a]1,000 sequences were randomly sampled from a single timepoint for each individual

[b]Based on the consensus taxonomy of ≥90% sequences within each phylotype (best-BLAST-hit against the Greengenes database)

Supplementary Table 7: Phylotypes shared across >90% of all individuals (V6 dataset; 10,000 random sequences/individual)									
Phylotype ID	Individuals with phylotype	% of individuals with phylotype	Number of reads grouped into phylotype	Highest relative abundance across all individuals	Lowest relative abundance across all individuals	Mean±sem % of 16S rRNA gene sequences across all individuals			Taxonomic classification[a]
1	33	100.0	10400	9.7	0.011	3.40	±	0.45	Bacteria; Firmicutes; Clostridia; Clostridiales; Clostridium_nexile
2	33	100.0	5161	5.9	0.011	1.67	±	0.23	Bacteria; Firmicutes; Clostridia; Clostridiales; Clostridium_nexile; Clostridium_fusiformis
3	33	100.0	6077	6.7	0.021	1.97	±	0.32	Bacteria; Firmicutes; Clostridia; Clostridiales; Ruminococcus
4	33	100.0	16600	26.8	0.011	5.36	±	1.02	Bacteria; Firmicutes; Clostridia; Clostridiales; Eubacterium_rectale
5	33	100.0	11654	12.5	0.011	3.78	±	0.58	Bacteria; Firmicutes; Clostridia; Clostridiales; Ruminococcus
6	32	97.0	3113	5.8	0.000	1.01	±	0.23	Bacteria; Firmicutes; Clostridia; Clostridiales; Clostridium_nexile
7	32	97.0	2908	4.2	0.000	0.96	±	0.21	Bacteria; Bacteroidetes; Bacteroidales; Bacteroidaceae
8	32	97.0	2382	3.7	0.000	0.78	±	0.13	Bacteria; Firmicutes; Clostridia; Clostridiales; Ruminococcus
9	32	97.0	1712	4.4	0.000	0.56	±	0.14	Bacteria; Firmicutes; Clostridia; Clostridiales; Ruminococcus; Ruminococcus_schinkii
10	31	93.9	3940	6.6	0.000	1.29	±	0.26	Bacteria; Firmicutes; Clostridia; Faecalibacterium
11	31	93.9	3729	4.9	0.000	1.21	±	0.18	Bacteria; Firmicutes; Clostridia; Clostridiales; Clostridium_nexile
12	30	90.9	454	0.7	0.000	0.15	±	0.03	Bacteria; Firmicutes; Clostridia; Clostridiales; Ruminococcus
13	30	90.9	687	1.1	0.000	0.23	±	0.04	Bacteria; Firmicutes; Clostridia
14	30	90.9	999	2.3	0.000	0.33	±	0.08	Bacteria; Firmicutes; Clostridia; Peptostreptococcaceae; Peptostreptococcus_anaerobius; Clostridium_bifermentans
15	30	90.9	1241	5.3	0.000	0.40	±	0.16	Bacteria; Firmicutes; Clostridia; Clostridiales; Clostridium_bolteae
16	30	90.9	160	0.2	0.000	0.05	±	0.01	Bacteria; Actinobacteria; Actinobacteridae; Actinomycineae
17	30	90.9	1417	2.0	0.000	0.46	±	0.09	Bacteria; Firmicutes; Clostridia; Clostridiales; Ruminococcus
18	30	90.9	1014	1.2	0.000	0.33	±	0.06	Bacteria; Firmicutes; Clostridia; Clostridiales
19	30	90.9	1353	1.6	0.000	0.44	±	0.08	Bacteria; Firmicutes; Clostridia; Clostridiales; Ruminococcus; Ruminococcus_luti
20	30	90.9	2686	6.0	0.000	0.88	±	0.22	Bacteria; Firmicutes; Clostridia; Clostridiales; Clostridium_clostridioforme
21	30	90.9	7454	12.2	0.000	2.43	±	0.63	Bacteria; Firmicutes; Clostridia; Faecalibacterium

[a]Based on the consensus taxonomy of ≥90% sequences within each phylotype (best-BLAST-hit against the Greengenes database)

Supplementary Table 8: Phylotypes shared across ≥70% of all individuals (Full-length dataset; 200 random sequences/individual)							
Phylotype ID	Individuals with phylotype	% of individuals with phylotype	Number of reads grouped into phylotype	Highest relative abundance across all individuals	Lowest relative abundance across all individuals	Mean±sem % of 16S rRNA gene sequences across all individuals	Taxonomic classification[a]
1	28	93.3	378	17.9	0.0	7.81 ± 1.04	Bacteria; Firmicutes; Clostridia; Faecalibacterium
2	27	90.0	347	25.0	0.0	6.90 ± 1.20	Bacteria; Firmicutes; Clostridia; Clostridiales; Ruminococcus
3	26	86.7	128	9.9	0.0	2.62 ± 0.47	Bacteria; Firmicutes; Clostridia; Clostridiales
4	26	86.7	298	23.1	0.0	6.00 ± 1.14	Bacteria; Firmicutes; Clostridia; Clostridiales; Eubacterium_rectale
5	26	86.7	127	12.0	0.0	2.64 ± 0.49	Bacteria; Firmicutes; Clostridia; Clostridiales; Clostridium_clostridioforme
6	22	73.3	110	10.9	0.0	2.33 ± 0.55	Bacteria; Bacteroidetes; Bacteroidales; Bacteroidaceae
7	22	73.3	87	5.7	0.0	1.76 ± 0.29	Bacteria; Firmicutes; Clostridia; Clostridiales; Clostridium_nexile; Clostridium_fusiformis
8	21	70.0	112	11.9	0.0	2.32 ± 0.49	Bacteria; Firmicutes; Clostridia; Clostridiales; Coprococcus
9	21	70.0	75	6.9	0.0	1.53 ± 0.32	Bacteria; Firmicutes; Clostridia; Clostridiales; Clostridium_nexile
10	21	70.0	54	5.7	0.0	1.14 ± 0.23	Bacteria; Firmicutes; Clostridia; Clostridiales; Clostridium_nexile

[a]Based on the consensus taxonomy of ≥90% sequences within each phylotype (best-BLAST-hit against the Greengenes database)

Supplementary Table 9: Phylum-level taxonomic assignments[a]							
		lean			obese		
		mean	sem	N	mean	sem	N
V2 (EA)	%Bacteroidetes	26.76	2.46	26	24.39	1.89	42
	%Firmicutes	71.48	2.50	26	72.57	1.92	42
	%Actinobacteria	0.72	0.14	26	1.70	0.58	42
V2 (AA)[b]	%Bacteroidetes	37.52	3.05	8	29.41	1.49	62
	%Firmicutes	60.74	3.04	8	68.14	1.42	62
	%Actinobacteria	0.97	0.40	8	1.27	0.21	62
V6 (EA)	%Bacteroidetes	6.85	1.25	12	3.15	0.93	16
	%Firmicutes	81.72	2.41	12	75.99	4.60	16
	%Actinobacteria	7.14	1.76	12	17.91	5.01	16
Full-length (EA)	%Bacteroidetes	11.44	2.77	10	7.58	2.35	16
	%Firmicutes	83.50	2.28	10	84.60	3.03	16
	%Actinobacteria	2.78	0.78	10	4.41	1.14	16
BLAST (EA)[c]	%Bacteroidetes	42.60	8.75	6	34.69	8.16	9
	%Firmicutes	51.54	8.35	6	51.25	5.47	9
	%Actinobacteria	2.07	0.33	6	10.34	3.35	9

[a]A subset of each dataset was included in the analysis: 10,000 sequences/sample (V6), 1,000 sequences/sample (V2) and 200 sequences/sample (full-length). Sequences from the same individual across both timepoints were pooled.

[b]The AA lean individuals surveyed have significantly more Bacteroidetes and less Firmicutes than the lean EA individuals ($p<0.05$, Student's t-test)

[c]BLASTX comparisons between microbiomes and NCBI non-redundant database

Supplementary Table 10: Relative abundance of CAZymes across 18 gut microbiomes (% of sequence assignments across all identified CAZymes)[a]																		
Subject ID[b]	**F1T1Le**	**F1T2Le**	**F1MOv**	**F2T1Le**	**F2T2Le**	**F2MOb**	**F3T1Le**	**F3T2Le**	**F3MOv**	**F4T1Ob**	**F4T2Ob**	**F4MOb**	**F5T1Ob**	**F5T2Ob**	**F5MOv**	**F6T1Ob**	**F6T2Ob**	**F6MOb**
Glycoside hydrolases	**70.56**	**73.96**	**72.14**	**72.40**	**68.38**	**67.37**	**68.69**	**67.84**	**69.92**	**73.46**	**70.45**	**71.57**	**64.19**	**69.11**	**69.96**	**68.15**	**69.61**	**71.50**
GH13	8.96	6.31	6.37	3.97	10.78	8.04	8.63	9.97	8.02	4.68	8.36	6.37	11.17	11.80	7.05	12.34	16.84	11.19
GH2	7.40	7.10	7.01	6.51	5.13	5.49	5.81	6.02	5.94	6.43	6.53	6.53	5.52	5.40	5.93	5.69	5.64	6.21
GH43	3.48	5.78	5.63	6.61	4.39	4.69	5.05	4.14	5.75	5.80	6.49	5.00	4.34	6.57	5.04	5.05	5.59	4.56
GH92	3.44	6.25	5.00	7.70	3.25	5.47	3.28	2.65	4.50	7.66	4.36	6.72	1.71	1.73	5.70	1.93	0.60	3.59
GH3	5.72	5.37	4.31	4.47	3.20	3.94	4.03	4.70	4.09	3.46	3.77	4.27	3.89	5.07	3.75	3.75	4.29	3.41
GH97	1.97	5.45	4.01	4.67	1.18	3.38	3.51	2.23	3.91	4.06	3.95	3.62	0.96	1.25	3.96	1.22	0.28	1.87
GH31	2.98	2.48	2.53	2.41	3.84	2.11	2.16	3.04	2.13	2.67	2.06	2.49	2.86	3.37	2.52	2.81	3.99	2.79
GH20	2.40	2.30	2.35	3.34	1.93	2.93	1.99	1.92	2.19	3.33	2.45	3.32	1.09	1.17	3.12	1.66	0.92	3.18
GH29	1.99	1.51	2.12	2.54	2.94	2.52	2.53	2.19	1.83	3.93	1.53	3.31	1.80	1.47	2.59	1.51	0.93	1.81
GH77	2.13	1.39	1.43	0.86	2.18	2.18	2.18	2.45	1.99	1.32	1.95	1.49	2.87	2.95	1.62	2.64	3.47	2.04
GH28	1.58	2.44	3.71	3.07	1.46	2.24	2.25	1.79	2.00	2.63	1.99	2.49	1.64	1.01	2.31	1.44	0.54	1.11
GH51	1.18	1.51	1.38	1.44	2.12	1.58	1.73	1.68	1.31	1.73	2.29	1.51	1.80	2.74	1.40	1.71	2.34	1.60
GH36	1.62	1.12	1.19	0.99	1.80	1.23	1.64	2.02	1.37	1.24	1.79	1.39	1.52	1.92	1.28	2.20	2.63	2.37
GH1	1.51	0.87	1.02	0.34	2.90	1.08	1.50	1.50	1.67	0.72	0.79	0.71	2.01	2.50	1.35	3.74	2.29	2.25
GH5	1.95	2.41	1.75	1.53	1.07	0.98	2.62	1.45	1.95	1.37	2.56	1.30	1.29	1.37	0.90	0.84	1.22	0.95
GH42	0.91	0.49	0.83	0.90	2.43	0.62	1.09	1.10	1.03	0.94	0.44	0.98	1.80	2.82	0.93	2.26	3.87	2.06
GH105	1.56	1.65	2.07	2.07	1.01	1.38	1.46	1.27	1.83	1.77	0.83	1.63	0.95	0.50	1.65	0.98	0.39	0.83
GH95	1.56	1.18	1.36	1.24	0.91	1.21	1.22	1.04	0.99	1.33	1.90	1.12	0.68	0.75	1.35	1.01	0.48	1.44
GH32	0.91	0.61	0.70	0.75	2.12	1.18	1.05	0.91	0.84	0.99	1.15	0.82	1.15	1.52	0.99	1.47	2.04	1.00
GH78	1.91	1.09	1.22	1.61	0.60	0.70	1.05	0.89	1.25	1.43	1.45	0.98	1.03	1.39	0.80	0.90	0.58	1.21
Glycosyltransferases	**20.25**	**17.20**	**17.49**	**16.26**	**23.34**	**21.64**	**22.09**	**22.78**	**19.66**	**16.68**	**20.34**	**18.24**	**26.36**	**23.15**	**19.53**	**23.54**	**23.99**	**21.50**
GT2	5.66	6.26	6.31	5.58	7.68	7.91	7.14	7.48	7.39	6.19	6.80	6.97	9.41	9.80	6.74	7.98	7.14	6.78
GT4	3.55	3.76	3.96	4.44	4.93	4.43	4.64	4.60	4.20	4.17	3.99	4.08	5.62	4.43	4.50	4.42	4.18	4.80
GT35	4.75	2.47	2.07	1.62	4.75	2.85	3.58	3.91	2.90	1.81	2.76	2.13	4.50	3.78	2.59	4.42	5.25	3.66
GT28	1.51	0.85	0.89	0.53	1.51	1.00	1.34	1.48	1.00	0.58	0.94	0.83	1.31	1.00	1.01	1.48	2.12	1.33
GT5	1.74	0.77	0.79	0.33	1.72	0.81	1.38	1.62	1.15	0.46	0.83	0.65	1.54	1.24	0.96	1.74	1.90	0.96
GT51	0.77	0.78	0.75	0.74	0.99	1.08	0.92	1.17	0.80	0.68	1.06	0.72	1.82	1.27	0.88	1.06	1.63	1.02
Carbohydrate binding modules	**1.76**	**2.40**	**2.15**	**2.02**	**2.05**	**2.22**	**2.38**	**2.25**	**2.11**	**1.90**	**2.06**	**2.15**	**2.66**	**2.88**	**2.08**	**2.22**	**2.28**	**1.98**
Carbohydrate esterases	**5.89**	**4.70**	**5.45**	**5.53**	**5.00**	**5.81**	**5.64**	**5.36**	**6.04**	**5.19**	**5.19**	**5.02**	**5.24**	**3.94**	**6.01**	**4.68**	**3.84**	**4.15**
CE4	1.53	1.01	1.03	0.78	1.41	1.04	1.16	1.27	1.20	0.73	0.84	0.92	1.35	0.96	1.04	1.31	1.51	0.91
Polysaccharide lyases	**1.55**	**1.74**	**2.77**	**3.79**	**1.22**	**2.95**	**1.20**	**1.78**	**2.27**	**2.78**	**1.95**	**3.02**	**1.55**	**0.93**	**2.43**	**1.43**	**0.28**	**0.87**

[a]Groups found at an average relative abundance >1% are shown

[b]ID nomenclature: Family number, Twin number or mother, and BMI category (Le=lean, Ov=overweight, Ob=obese; e.g. F1T1Le stands for family 1, twin 1, lean)

Supplementary Table 11: Relative abundance of metabolic pathways in the gut microbiome (% of KEGG assignments)[a]			
KEGG metabolic pathway	**Mean±sem across all 18 microbiomes**		
Transporters	4.93	±	0.21
Other replication, recombination and repair proteins	3.35	±	0.04
ABC transporters	3.24	±	0.13
General function prediction only	2.60	±	0.06
Purine metabolism	2.29	±	0.02
Other enzymes	2.16	±	0.03
Aminoacyl-tRNA biosynthesis	2.14	±	0.05
Glutamate metabolism	1.98	±	0.03
Starch and sucrose metabolism	1.92	±	0.03
Pyruvate metabolism	1.73	±	0.02
Pyrimidine metabolism	1.70	±	0.02
Peptidases	1.69	±	0.05
Alanine and aspartate metabolism	1.58	±	0.02
Glycine, serine and threonine metabolism	1.53	±	0.02
Other translation proteins	1.37	±	0.02
Galactose metabolism	1.37	±	0.03
Glycolysis / Gluconeogenesis	1.35	±	0.02
Other ion-coupled transporters	1.34	±	0.06
Fructose and mannose metabolism	1.31	±	0.03
Two-component system	1.31	±	0.03
Ribosome	1.27	±	0.03
Replication complex	1.18	±	0.02
Phenylalanine, tyrosine and tryptophan biosynthesis	1.17	±	0.02
Valine, leucine and isoleucine biosynthesis	1.15	±	0.02
Carbon fixation	1.15	±	0.01
Nitrogen metabolism	1.13	±	0.02
Glycerolipid metabolism	1.07	±	0.02
Oxidative phosphorylation	1.07	±	0.03
Butanoate metabolism	1.05	±	0.02
Chaperones and folding catalysts	0.99	±	0.01
Pentose phosphate pathway	0.95	±	0.01
Tyrosine metabolism	0.95	±	0.02
Histidine metabolism	0.92	±	0.02
Cell division	0.91	±	0.01
Aminosugars metabolism	0.89	±	0.03
Arginine and proline metabolism	0.85	±	0.01
Citrate cycle (TCA cycle)	0.84	±	0.02
Methionine metabolism	0.83	±	0.02
Lysine biosynthesis	0.82	±	0.01
RNA polymerase	0.81	±	0.02
Reductive carboxylate cycle (CO2 fixation)	0.80	±	0.03
Propanoate metabolism	0.80	±	0.01
Peptidoglycan biosynthesis	0.79	±	0.01
N-Glycan degradation	0.78	±	0.05
Urea cycle and metabolism of amino groups	0.78	±	0.01
Translation factors	0.78	±	0.02
Selenoamino acid metabolism	0.77	±	0.02
Glyoxylate and dicarboxylate metabolism	0.73	±	0.01
DNA polymerase	0.72	±	0.01
Pentose and glucuronate interconversions	0.70	±	0.02
Cysteine metabolism	0.68	±	0.02
Pantothenate and CoA biosynthesis	0.67	±	0.01
Nucleotide sugars metabolism	0.67	±	0.02
Glycosaminoglycan degradation	0.66	±	0.04
Function unknown	0.66	±	0.01
One carbon pool by folate	0.65	±	0.01
Sphingolipid metabolism	0.64	±	0.03
Protein export	0.62	±	0.01

[a]Pathways with an average relative abundance of >0.6% are shown

Supplementary Table 12: KEGG Pathways enriched or depleted in the distal gut microbiome of obese twins[a]	
Enriched	Fatty acid biosynthesis
	Nicotinate and nicotinamide metabolism
	Other ion-coupled transporters
	Other transporters
	Pentose and glucuronate interconversions
	Phosphotransferase system (PTS)
	Protein folding and associated processing
	Signal transduction mechanisms
	Transcription factors
Depleted	Bacterial chemotaxis
	Bacterial motility proteins
	Benzoate degradation via CoA ligation
	Butanoate metabolism
	Citrate cycle (TCA cycle)
	Glycosaminoglycan degradation
	Other enzymes
	Oxidative phosphorylation
	Pyruvate/Oxoglutarate oxidoreductases
	Starch and sucrose metabolism
	Tryptophan metabolism

[a]Identified by comparing KEGG pathway abundance in the variable component of the fecal microbiomes of obese versus lean MZ twins using a bootstrap algorithm ($p<0.01$)

Supplementary Table 13: Bacterial genes enriched in the gut microbiomes of obese MZ twins				
Genome and NCBI proteinID*	**Annotation**	**COG**	**COG Categories**	**KEGG orthologous groups**
Bifidobacterium_adolescentis_154486403	tRNA-ribosyltransferase	COG0343	J	K00773
Bifidobacterium_longum_23465114	Transcriptional regulators	COG1609	K	
Bifidobacterium_longum_23466186	ABC-type sugar transport system, periplasmic component	COG1653	G	
Bifidobacterium_adolescentis_154488903	Superfamily I DNA and RNA helicases	COG3973	R	
Bifidobacterium_adolescentis_154486727	DNA polymerase IV	COG0389	L	K02346
Bifidobacterium_adolescentis_154488882	peptide/nickel transport system ATP-binding protein	COG1123	R	K02031/2
Bifidobacterium_adolescentis_154488633	Trk-type K+ transport systems	COG0168	P	
Bifidobacterium_adolescentis_154488131	Asp-tRNAAsn/Glu-tRNAGln amidotransferase B subunit	COG0064	J	K02434
Bifidobacterium_adolescentis_154487571	Threonine dehydratase	COG1171	E	K01754
Bifidobacterium_adolescentis_154486641	Glucose-6-phosphate isomerase	COG0166	G	K01810
Bifidobacterium_adolescentis_154488790	ATP-dependent helicase Lhr and Lhr-like helicase	COG1201	R	K03724
Bifidobacterium_adolescentis_119025482	Predicted ATPase involved in cell division	COG2884	D	K09812
Bifidobacterium_adolescentis_154486531	Predicted phosphohydrolases	COG1409	R	
Bifidobacterium_adolescentis_154486606	tRNA-(guanine-N1)-methyltransferase	COG0336	J	K00554
Bifidobacterium_adolescentis_154486895	IMP dehydrogenase/GMP reductase	COG0516/7	FR	K00088
Bifidobacterium_adolescentis_154486720	Aspartate/tyrosine/aromatic aminotransferase	COG0436	E	K00812
Bifidobacterium_adolescentis_119026599	Cation transport ATPase	COG0474	P	K01529
Bifidobacterium_adolescentis_154486334	hypothetical protein			
Bifidobacterium_adolescentis_119025743	NAD/NADP transhydrogenase alpha subunit	COG3288	C	K00324
Bifidobacterium_longum_23336617	UspA and related nucleotide-binding proteins	COG0589	T	
Bifidobacterium_adolescentis_154486937	ABC-type sugar transport system	COG1653	G	K02027
Bifidobacterium_longum_23465912	hypothetical protein			
Bifidobacterium_longum_23335963	K+ transporter	COG3158	P	K03549
Bifidobacterium_adolescentis_119025729	ABC-type transport system, Fe-S cluster assembly	COG0719	O	
Bifidobacterium_adolescentis_154487396	Glutamine synthetase adenylyltransferase	COG1391	OT	K00982
Bifidobacterium_adolescentis_154488156	hypothetical protein			
Bifidobacterium_adolescentis_154486668	Acetyl/propionyl-CoA carboxylase	COG4770	I	K01946
Bifidobacterium_adolescentis_154487299	Nuclease subunit of the excinuclease complex	COG0322	L	K03703
Bifidobacterium_longum_23465540	Acetate kinase	COG0282	C	K00925
Clostridium_bartlettii_164687465	putative conjugative transposon protein	NOG13238		
Bifidobacterium_longum_23465037	Dipeptidase	COG4690	E	K08659
Bifidobacterium_adolescentis_154488210	Predicted hydrolase of the metallo-beta-lactamase superfamily	COG0595	R	K07021
Bifidobacterium_adolescentis_154487598	tRNA/rRNA methyltransferase protein			K00599
Bifidobacterium_adolescentis_119025149	hypothetical protein			
Bifidobacterium_adolescentis_154487052	hypothetical protein	NOG07592		
Bifidobacterium_adolescentis_154486554	PTS system, enzyme I			K00935
Bifidobacterium_longum_23335005	Selenocysteine lyase	COG0520	E	K01763
Bifidobacterium_longum_23465294	Branched-chain amino acid permeases	COG1114	E	K03311
Bifidobacterium_adolescentis_119025432	Acyl-CoA thioesterase	COG1946	I	K01076
Bifidobacterium_adolescentis_154486528	Aspartate-semialdehyde dehydrogenase	COG0136	E	K00133
Bifidobacterium_adolescentis_154487076	Predicted ATPase with chaperone activity	COG0606	O	K07391
Bifidobacterium_longum_23466221	Alcohol dehydrogenase, class IV	COG1454	C	K00048
Bifidobacterium_adolescentis_119025541	Phosphoribosylformylglycinamidine synthase	COG0046/7	F	K01952
Bifidobacterium_adolescentis_119026031	Geranylgeranyl pyrophosphate synthase	COG0142	H	
Bifidobacterium_longum_23465502	Signal transduction histidine kinase	COG4585	T	
Bifidobacterium_adolescentis_154486631	Predicted metal-binding, possibly nucleic acid-binding protein	COG1399	R	
Bifidobacterium_adolescentis_154488013	Sugar (pentulose and hexulose) kinases	COG1070	G	K00853
Bifidobacterium_adolescentis_119025777	Aspartate carbamoyltransferase	COG0540	F	K00609
Bifidobacterium_adolescentis_119025510	Superfamily II DNA helicase	COG0514	L	K03654
Bifidobacterium_adolescentis_119026360	Protease II	COG1770	E	K01354
Bifidobacterium_adolescentis_119025672	Signal transduction histidine kinase	COG3920	T	
Bifidobacterium_adolescentis_154487392	Orotidine-5'-phosphate decarboxylase	COG0284	F	K01591
Bifidobacterium_adolescentis_154487114	Permeases of the major facilitator superfamily	COG0477	GEPR	
Bifidobacterium_adolescentis_119025804	Predicted Fe-S-cluster redox enzyme	COG0820	R	K06941
Bifidobacterium_longum_23465197	Permeases of the major facilitator superfamily	COG0477	GEPR	
Bifidobacterium_adolescentis_154487064	Superfamily II RNA helicase	COG4581	L	K01529
Bifidobacterium_longum_23465727	ABC-type dipeptide transport system	COG0747	E	K02035
Bifidobacterium_adolescentis_154486507	hypothetical protein			
Bifidobacterium_longum_23465472	Predicted transcriptional regulator	COG2865	K	
Bifidobacterium_adolescentis_154486695	ABC-type phosphate transport system	COG0226	P	K02040
Bifidobacterium_longum_23466332	Dihydroxyacid dehydratase/phosphogluconate dehydratase	COG0129	EG	K01687
Bifidobacterium_adolescentis_154489143	Predicted phosphatase/phosphohexomutase	COG0637	R	
Bifidobacterium_adolescentis_154486988	Phosphoribosylaminoimidazole carboxylase	COG0026	F	K01589
Bifidobacterium_adolescentis_154486732	glycoside hydrolase family 77	COG1640	G	K00705
Bifidobacterium_adolescentis_154487590	Uncharacterized conserved protein	COG3247	S	
Bifidobacterium_adolescentis_154486669	Acetyl-CoA carboxylase	COG4799	I	K01966
Bifidobacterium_adolescentis_154488016	Homoserine kinase	COG0083	E	K00872
Bifidobacterium_adolescentis_119026221	glycoside hydrolase family 43			
Bifidobacterium_adolescentis_119025727	CTP synthase (UTP-ammonia lyase)	COG0504	F	K01937
Bifidobacterium_adolescentis_154486325	Uncharacterized protein conserved in bacteria	COG3583	S	
Bifidobacterium_adolescentis_119025371	Transcription elongation factor	COG0195	K	K02600
Bifidobacterium_adolescentis_154486867	Sugar (pentulose and hexulose) kinases	COG1070	G	K00854
Bifidobacterium_adolescentis_154487511	putative cell division protein			
Bifidobacterium_adolescentis_154487124	hypothetical protein			
Bifidobacterium_adolescentis_119025212	hypothetical protein			
Bifidobacterium_adolescentis_154487481	hypothetical protein			
Bifidobacterium_adolescentis_154488824	putative two-component sensor kinase			

Bifidobacterium_adolescentis_154488224	serine_threonine protein kinase			
Bifidobacterium_adolescentis_154487149	carbohydrate esterase family 1			
Bifidobacterium_adolescentis_154488135	rRNA methylases	COG0566	J	K00599
Bifidobacterium_adolescentis_154489172	glycoside hydrolase family 77	COG1640	G	K00705
Bifidobacterium_adolescentis_154487327	Superfamily II RNA helicase	COG4581	L	K03727
Bifidobacterium_adolescentis_119025670	Transcription elongation factor	COG0782	K	K03624
Bifidobacterium_adolescentis_154486326	Dimethyladenosine transferase	COG0030	J	K02528
Bifidobacterium_longum_23465077	glycosyl-transferase family 51	COG0744	M	K03693
Bifidobacterium_longum_23464647	hypothetical protein	NOG25707		
Bifidobacterium_adolescentis_154486363	hypothetical protein			
Bifidobacterium_adolescentis_154486438	Permeases of the major facilitator superfamily	COG0477	GEPR	
Bifidobacterium_longum_23335686	ABC-type antimicrobial peptide transport system	COG0577	V	K02004
Bifidobacterium_adolescentis_154486327	4-diphosphocytidyl-2C-methyl-D-erythritol 2-phosphate synthase	COG1947	I	K00919
Bifidobacterium_adolescentis_154488959	twitching motility protein PilT			K02669
Bifidobacterium_adolescentis_154486273	Leucyl-tRNA synthetase	COG0495	J	K01869
Bifidobacterium_adolescentis_154486329	tRNA nucleotidyltransferase/poly(A) polymerase	COG0617	J	K00970
Bifidobacterium_adolescentis_154487191	putative phage protein			
Bifidobacterium_adolescentis_154486270	DNA polymerase III, delta subunit	COG1466	L	K02340
Bifidobacterium_adolescentis_154486380	hypothetical protein			
Anaerostipes_caccae_167747544	Non-ribosomal peptide synthetase modules and related proteins	COG1020	Q	
Bifidobacterium_adolescentis_154486501	Predicted unusual protein kinase	COG0661	R	
Bifidobacterium_adolescentis_154486855	LacI-family transcriptional regulator			
Bifidobacterium_adolescentis_154486358	Hemolysins and related proteins	COG1253	R	K03699
Bifidobacterium_adolescentis_154486649	Acetylornithine deacetylase/Succinyl-diaminopimelate desuccinylase	COG0624	E	K01439
Bifidobacterium_adolescentis_119025555	Orotidine-5'-phosphate decarboxylase	COG0284	F	K01591
Bifidobacterium_longum_23465600	Gamma-glutamyl phosphate reductase	COG0014	E	K00147
Bifidobacterium_adolescentis_154486786	FAD synthase/riboflavin kinase/FMN adenylyltransferase	COG0196	H	K00861/0953
Bifidobacterium_adolescentis_154488712	Ribonuclease D	COG0349	J	K03684
Bifidobacterium_adolescentis_154488649	N-acetylglutamate synthase (N-acetylornithine aminotransferase)	COG1364	E	K00620/0642
Bifidobacterium_adolescentis_154489082	Ribonucleoside-triphosphate reductase	COG1328	F	K00527
Bifidobacterium_adolescentis_154487141	transcriptional regulator, AraC family			
Bifidobacterium_longum_23335562	Acetyltransferase (isoleucine patch superfamily)	COG0110	R	K00680
Bifidobacterium_adolescentis_119025600	ABC-type amino acid transport system, permease component	COG0765	E	
Bifidobacterium_adolescentis_154486349	Recombinational DNA repair ATPase (RecF pathway)	COG1195	L	K03629
Bifidobacterium_adolescentis_154487341	Succinyl-CoA synthetase	COG0045	C	K01903
Bifidobacterium_adolescentis_154486419	Adenylosuccinate synthase	COG0104	F	K01939
Bifidobacterium_adolescentis_154486323	transcriptional regulator, AraC family			
Bifidobacterium_adolescentis_119025197	3-isopropylmalate dehydratase large subunit	COG0065	E	K01702/3
Bifidobacterium_adolescentis_154489094	Predicted dehydrogenases and related proteins	COG0673	R	
Bifidobacterium_longum_23336262	O-acetylhomoserine sulfhydrylase	COG2873	E	K01740
Bifidobacterium_longum_23465907	ABC-type dipeptide/oligopeptide/nickel transport systems	COG0601	EP	K02033
Bifidobacterium_adolescentis_154487000	Threonine aldolase	COG2008	E	K01620
Bifidobacterium_adolescentis_154487167	Sortase and related acyltransferases	COG1247	M	K03823
Bifidobacterium_longum_23465198	Thioredoxin reductase	COG0492/0526	OC	K00384
Bifidobacterium_adolescentis_154488926	Arabinose efflux permease	COG2814	G	
Bifidobacterium_longum_23465931	ABC-type antimicrobial peptide transport system, ATPase component	COG1136	V	K02003/4
Bifidobacterium_adolescentis_154486352	Type IIA topoisomerase (DNA gyrase/topo II, topoisomerase IV)	COG0188	L	K01863/2469
Bifidobacterium_adolescentis_119026009	Pyruvate-formate lyase-activating enzyme	COG1180	O	K04069
Bifidobacterium_adolescentis_154487279	Methionine synthase II (cobalamin-independent)	COG0620	E	K00549
Bifidobacterium_adolescentis_119025238	Acetolactate synthase	COG0440	E	K01653
Bifidobacterium_adolescentis_119025129	Signal recognition particle GTPase	COG0552	U	K03110
Bifidobacterium_adolescentis_154488132	Asp-tRNAAsn/Glu-tRNAGln amidotransferase	COG0154	J	K02433
Bifidobacterium_adolescentis_154486940	ABC-type dipeptide transport system	COG0747	E	K02035
Bifidobacterium_adolescentis_154488789	Type IIA topoisomerase (DNA gyrase/topo II, topoisomerase IV)	COG0188	L	K01863/2469
Bifidobacterium_adolescentis_154487377	Long-chain acyl-CoA synthetases	COG1022	I	K01897
Bifidobacterium_adolescentis_154488794	DNA-directed RNA polymerase, sigma subunit	COG0568	K	K03086
Bifidobacterium_adolescentis_154488989	Superfamily I DNA and RNA helicases	COG0210	L	K01529
Bifidobacterium_adolescentis_154486903	Prolyl-tRNA synthetase	COG0442	J	K01881
Bifidobacterium_adolescentis_154488684	putative helicase			
Bifidobacterium_adolescentis_154486399	Lysophospholipase	COG2267	I	
Bifidobacterium_adolescentis_119026611	ABC-type sugar transport systems, ATPase components	COG3839	G	K05816
Bifidobacterium_adolescentis_154486670	Putative fatty acid synthase/reductase	COG0304/0331/2030/4981/4982	IQ	K00059/209/665/666/680
Bifidobacterium_adolescentis_154488852	ABC-type oligopeptide transport system	COG4166	E	K02035
Bifidobacterium_adolescentis_154486664	putative ABC-type sugar transport system			
Bifidobacterium_adolescentis_119025257	Ribonucleases G and E	COG1530	J	K01128
Bifidobacterium_adolescentis_154486472	ABC-type antimicrobial peptide transport system	COG0577	V	K02004
Bifidobacterium_adolescentis_154487036	hypothetical protein			
Bifidobacterium_adolescentis_154487636	glycoside hydrolase family 2	COG3250	G	K01190
Eubacterium_dolichum_160915695	glycoside hydrolase family 31			
Bifidobacterium_adolescentis_154489092	Aspartate/tyrosine/aromatic aminotransferase	COG0436	E	K00812
Bifidobacterium_adolescentis_119026440	hypothetical protein	NOG21350		
Bifidobacterium_adolescentis_119025397	Myosin-crossreactive antigen	COG4716	S	
Bifidobacterium_adolescentis_119026143	Glutamine amidotransferase	COG0118	E	K02501
Bifidobacterium_adolescentis_154487050	Universal stress protein UspA	COG0589	T	
Bifidobacterium_adolescentis_154486729	Phosphoglycerate dehydrogenase	COG0111	HE	
Bifidobacterium_adolescentis_154488261	Predicted hydrolases or acyltransferases	COG0596	R	
Bifidobacterium_adolescentis_154489101	hypothetical protein			

Bifidobacterium_adolescentis_154487476	Phosphotransacetylase	COG0280/0857	CR	K00625
Bifidobacterium_adolescentis_154488788	Uncharacterized proteins of the AP superfamily	COG1524	R	
Ruminococcus_obeum_153809835	putative ketose-bisphosphate aldolase			
Clostridium_leptum_160933115	hypothetical protein			
Bifidobacterium_adolescentis_119026429	Ribulose-5-phosphate 4-epimerase	COG0235	G	K03080
Bifidobacterium_adolescentis_154487579	glycoside hydrolase family 36	COG3345	G	K07407
Bifidobacterium_longum_23464678	hypothetical protein			
Bifidobacterium_adolescentis_154486391	Serine/threonine protein phosphatase	COG0631	T	K01090
Bifidobacterium_adolescentis_154486962	ABC-type amino acid transport/signal transduction systems	COG0834	ET	K02030
Bifidobacterium_adolescentis_154486954	DNA primase	COG0358	L	K02316
Bifidobacterium_adolescentis_154486993	Glutamine phosphoribosylpyrophosphate amidotransferase	COG0034	F	K00764
Bifidobacterium_adolescentis_154488913	HrpA-like helicases	COG1643	L	K03578
Bifidobacterium_adolescentis_154486787	Predicted ATP-dependent serine protease	COG1066	O	K04485
Bifidobacterium_adolescentis_154486493	Ammonia permease	COG0004	P	K03320
Bifidobacterium_adolescentis_154487494	Methenyl tetrahydrofolate cyclohydrolase	COG0190	H	K00288/1491
Bifidobacterium_adolescentis_119025196	Transcriptional regulator	COG1414	K	
Dorea_longicatena_153853202	hypothetical protein			
Bifidobacterium_adolescentis_154487329	putative transcriptional regulator			
Bifidobacterium_adolescentis_154487591	LacI-family transcriptional regulator			
Bifidobacterium_adolescentis_154486321	glycoside hydrolase family 3			
Bifidobacterium_adolescentis_119025741	GTPase	COG1159	R	K03595
Clostridium_scindens_167758922	dUTPase	COG0756	F	K01520
Bifidobacterium_adolescentis_119025587	Signal transduction histidine kinase	COG0642	T	
Bifidobacterium_adolescentis_154486470	Predicted membrane protein	COG4393	S	
Clostridium_scindens_167760262	putative sporulation protein			
Bacteroides_stercoris_167763769	hypothetical protein			
Anaerostipes_caccae_167746872	putative ABC transporter			
Bifidobacterium_adolescentis_154486920	ABC-type amino acid transport/signal transduction systems	COG0834	ET	K02030
Bifidobacterium_adolescentis_154487063	Uncharacterized conserved protein	COG2326	S	
Bifidobacterium_adolescentis_119025989	glycoside hydrolase family 13	COG0366	G	K01187
Clostridium_bartlettii_164687864	Lactoylglutathione lyase	COG0346	E	K01759
Bifidobacterium_adolescentis_154486443	ABC-type antimicrobial peptide transport system	COG0577	V	K02004
Bifidobacterium_adolescentis_154488245	NADH:flavin oxidoreductases/NADPH2 dehydrogenase	COG1902	C	K00354
Bifidobacterium_longum_23465963	atypical histidine kinase sensor of two-component system	NOG21560		
Bifidobacterium_adolescentis_154488949	hypothetical protein			
Bifidobacterium_adolescentis_154486865	maltose O-acetyltransferase			
Clostridium_scindens_167759009	cytidylate kinase			K00945
Bifidobacterium_adolescentis_154486901	ATP-dependent exoDNAse	COG0507	L	
Ruminococcus_torques_153814251	hypothetical protein			
Bifidobacterium_adolescentis_119025327	Ribosomal protein L13	COG0102	J	K02871
Bifidobacterium_adolescentis_154488916	ABC-type antimicrobial peptide transport system	COG1136	V	
Bifidobacterium_adolescentis_119025389	putative histidine kinase sensor of two component system			
Ruminococcus_gnavus_154504598	Translation elongation factor P (EF-P)/initiation factor 5A (eIF-5A)	COG0231	J	K02356
Bifidobacterium_adolescentis_119026648	ribonuclease P	NOG21633		K03536
Clostridium_scindens_167760715	hypothetical protein			
Bifidobacterium_adolescentis_119026098	Uncharacterized conserved protein	COG2606	S	
Clostridium_scindens_167761320	ABC-type antimicrobial peptide transport system	COG1136	V	K02003
Bacteroides_stercoris_167762249	hypothetical protein			
Anaerostipes_caccae_167746530	putative ion channel			
Bifidobacterium_adolescentis_119025057	Serine/threonine protein kinase	COG0515	RTKL	
Clostridium_bartlettii_164686672	Molybdopterin biosynthesis enzymes	COG0521	H	K03638
Ruminococcus_obeum_153811887	hypothetical protein			
Clostridium_spiroforme_169349879	protein-Np-phosphohistidine-sugar phosphotransferase			K00890
Clostridium_ramosum_167756439	type I restriction enzyme, S subunit			K01154
Bifidobacterium_adolescentis_119025640	Short-chain alcohol dehydrogenase of unknown specificity	COG4221	R	
Eubacterium_ventriosum_154483925	Uncharacterized conserved protein	COG2501	S	
Bifidobacterium_adolescentis_154487477	Phosphoketolase	COG3957	G	K01621/32/36
Bifidobacterium_adolescentis_154489149	Putative molecular chaperone	COG0443	O	K01529/4043/8070
Bifidobacterium_adolescentis_119025585	hypothetical protein			
Clostridium_scindens_167759334	ABC-type antimicrobial peptide transport system	COG1136	V	K02003
Anaerostipes_caccae_167748732	Serine-pyruvate aminotransferase/archaeal aspartate aminotransferase	COG0075	E	K03430
Ruminococcus_gnavus_154505702	Putative phage replication protein RstA	COG2946	L	K07467
Bifidobacterium_adolescentis_154486389	Cell division protein FtsI	COG0768	M	
Bifidobacterium_adolescentis_154488668	ABC-type cobalt transport system	COG1122	P	K02006
Bifidobacterium_adolescentis_154486277	Fructose-2,6-bisphosphatase/phosphoglycerate mutase	COG0406	G	K01834
Clostridium_scindens_167758556	hypothetical protein			
Dorea_longicatena_153855715	putative acetyltransferase			
Eubacterium_dolichum_160915136	ABC-type antimicrobial peptide transport system	COG1136	V	K02003
Bifidobacterium_adolescentis_119026205	Isoleucyl-tRNA synthetase	COG0060	J	K01870
Ruminococcus_obeum_153810514	glycoside hydrolase family 23	COG0741/91	M	
Eubacterium_eligens_Contig2011.538	putative phosphohydrolase			
Bifidobacterium_adolescentis_154487387	Transcriptional regulator	COG0583	K	
Ruminococcus_obeum_153812199	putative flavodoxin			
Bifidobacterium_adolescentis_154486996	Phosphoribosylformylglycinamidine (FGAM) synthase	COG0046/7	F	K01952
Dorea_longicatena_153854194	Ornithine/acetylornithine aminotransferase	COG4992	E	K00818
Ruminococcus_gnavus_154505209	Predicted GTPases	COG1160	R	
Dorea_longicatena_153853531	Predicted transcriptional regulators	COG1695	K	
Ruminococcus_torques_153814203	Acetyltransferases	COG0456	R	K03826

Clostridium_scindens_167761371	putative ABC-type transport system			
Bifidobacterium_longum_38906105	F0F1-type ATP synthase	COG0055	C	K02112
Collinsella_aerofaciens_139439837	hypothetical protein			
Clostridium_leptum_160933570	ABC-type antimicrobial peptide transport system	COG0577/1136	V	K02003
Eubacterium_rectale_2731	putative sensor histidine kinase			
Bifidobacterium_adolescentis_154489126	ABC-type multidrug transport system	COG1132	V	K06147
Ruminococcus_obeum_153812105	putative conjugative transposon protein	NOG05968		
Dorea_longicatena_153853999	hypothetical protein			
Clostridium_bolteae_160937390	hypothetical protein			
Ruminococcus_torques_153814809	cytidylate kinase			K00945
Ruminococcus_obeum_153810530	hypothetical protein			
Clostridium_scindens_167758273	putative alanine racemase			
Clostridium_scindens_167760222	putative ABC transporter			
Dorea_longicatena_153854759	Sporulation protein	COG2088	M	K06412
Bifidobacterium_adolescentis_119025414	glycosyl-transferase family 4			
Ruminococcus_obeum_153813075	hypothetical protein			
Eubacterium_ventriosum_154482695	Queuine/archaeosine tRNA-ribosyltransferase	COG0343	J	K00773
Ruminococcus_obeum_153811892	hypothetical protein			
Ruminococcus_obeum_153810246	Type IV secretory pathway, VirB4 components	COG3451	U	
Dorea_longicatena_153854838	Ribosomal protein S16	COG0228	J	K02959
Dorea_longicatena_153855241	putative DNA gyrase, subunit A			
Collinsella_aerofaciens_139438412	putative transcriptional regulator			
Clostridium_leptum_160934853	putative ribosomal-protein-alanine acetyltransferase			
Eubacterium_rectale_3602	Type IV secretory pathway, VirD4 components	COG3505	U	
Bifidobacterium_adolescentis_154486460	ABC-type multidrug transport system	COG1132	V	K06147
Anaerostipes_caccae_167746203	exonuclease SbcC			K03546
Ruminococcus_obeum_153813732	hypothetical protein			
Eubacterium_ventriosum_154484729	protein-Np-phosphohistidine-sugar phosphotransferase			K00890
Eubacterium_rectale_3363	putative ABC transporter			
Ruminococcus_obeum_153809913	hypothetical protein			
Anaerostipes_caccae_167748861	putative arylsulfate sulfotransferase			
Eubacterium_eligens_Contig2011.154	Uncharacterized conserved protein	COG4283	S	
Clostridium_scindens_167759418	putative competence protein ComEA			
Eubacterium_rectale_3439	putative RNA-directed DNA polymerase			
Clostridium_bolteae_160940954	SAM-dependent methyltransferases	COG0500	QR	K00599
Ruminococcus_obeum_153811726	putative DNA topoisomerase			
Ruminococcus_obeum_153813044	putative transposase			
Eubacterium_rectale_2410	type I restriction enzyme, R subunit			K01152/3
Clostridium_bolteae_160941795	putative recombination protein			
Bifidobacterium_adolescentis_154486724	putative esterase			
Collinsella_aerofaciens_139438485	putative amidohydrolase			

*Protein sequences from *E.rectale, E.eligens, B.theta3731, B.theta7330, and B.WH2* can be found at http://gordonlab.wustl.edu/SuppData.html

Supplementary Table 14: Bacterial genes enriched in the gut microbiomes of lean MZ twins				
Genome and NCBI proteinID*	**Annotation**	**COG**	**COG Categories**	**KEGG orthologous groups**
Bacteroides_capillosus_154500567	putative amidohydrolase			
Clostridium_leptum_160934848	putative acetyltransferase			
Ruminococcus_obeum_153810033	phosphocarrier protein HPr			K02784
Eubacterium_siraeum_167749283	putative ABC transporter related protein			
Bacteroides_capillosus_154497054	Polyribonucleotide nucleotidyltransferase	COG1185	J	K00962
Eubacterium_siraeum_167749675	Isoleucyl-tRNA synthetase	COG0060	J	K01870
Eubacterium_rectale_3617	hypothetical protein			
Bacteroides_capillosus_154498345	putative sporulation protein			
Parabacteroides_merdae_154490921	hypothetical protein			
Bacteroides_capillosus_154500960	putative chromosome segregation protein			
Ruminococcus_torques_153814925	putative sporulation protein			
Clostridium_scindens_167758815	glycosyl-transferase family 4			
Clostridium_sp._L2_50_160893842	Protease subunit of ATP-dependent Clp proteases	COG0740	OU	K01358
B_WH2_000545	putative type I restriction enzyme EcoAI specificity protein			
Bacteroides_capillosus_154500843	trk system potassium uptake protein TrkA			K03499
Clostridium_bolteae_160936948	putative two-component transcriptional regulator			
Bacteroides_capillosus_154498005	ATP-dependent serine protease/cysteine S-methyltransferase	COG1066	O	K00567
Parabacteroides_merdae_154492394	hypothetical protein			
Bacteroides_capillosus_154498009	Fructose/tagatose bisphosphate aldolase	COG0191	G	K01622
B_theta_3731_000845	hypothetical protein			
Anaerotruncus_colihominis_167769594	Predicted ATPase (AAA+ superfamily)	COG1373	R	
Bacteroides_capillosus_154500228	putative translation protein			
Anaerofustis_stercorihominis_169334667	putative DNA recombinase			
B_theta_3731_003400	hypothetical protein			
Parabacteroides_distasonis_150008749	hypothetical protein			
Bacteroides_fragilis_19068109	mobilization protein BmgA	NOG11714		
Eubacterium_dolichum_160914154	glycoside hydrolase family 20	COG3525	G	K01207
Bacteroides_capillosus_154497125	RNA methyltransferase, TrmH family			K03218
Clostridium_sp._L2_50_160894658	NTP pyrophosphohydrolases	COG0494/3323	LRS	K03574
Parabacteroides_merdae_154494925	Glyceraldehyde-3-phosphate dehydrogenase	COG0057	G	K00134
Bacteroides_capillosus_154496139	Type IIA topoisomerase (DNA gyrase/topo II, topoisomerase IV)	COG0188	L	K01863/2469
Clostridium_ramosum_167755346	MoxR-like ATPase			K03924
Bacteroides_uniformis_160888848	hypothetical protein			
Ruminococcus_gnavus_154504651	Putative translation initiation inhibitor	COG0251	J	K07567
Bacteroides_uniformis_160890270	putative phage protein			
Bacteroides_capillosus_154500164	putative DNA recombinase			
B_WH2_000807	sulfotransferase/FAD synthetase	COG0175	EH	K00957
Bacteroides_uniformis_160892052	carbohydrate esterase family 4 and 12			
Clostridium_sp._L2_50_160893671	hypothetical protein			
Bacteroides_capillosus_154500952	hypothetical protein			K09710
Clostridium_scindens_167759293	putative ribonucleoside-triphosphate reductase activating protein			
Bacteroides_capillosus_154498134	Predicted GTPases	COG1160	R	K03977
Bacteroides_capillosus_154500412	ribosomal protein			
Bacteroides_fragilis_60683403	Imidazolonepropionase and related amidohydrolases	COG1228	Q	K01468
Peptostreptococcus_micros_160946111	hypothetical protein	NOG15344		
B_theta_7330_001524	putative transposase			
Bacteroides_capillosus_154500229	putative peptidase			
Bacteroides_vulgatus_150006208	Integrase	COG0582	L	
Bacteroides_capillosus_154501540	hypothetical protein			
Bacteroides_stercoris_167762500	Site-specific recombinase XerD	COG4974	L	
Bacteroides_fragilis_60679880	glycoside hydrolase family 38	COG0383	G	K01191
Bacteroides_capillosus_154497979	putative replication protein			
Bacteroides_capillosus_154500160	putative helicase			
Bacteroides_stercoris_167752230	Retron-type reverse transcriptase	COG3344	L	
B_WH2_003792	hypothetical protein	NOG14996		
Bacteroides_capillosus_154497731	hypothetical protein			
Parabacteroides_merdae_154494117	UDP-N-acetyl-D-mannosaminuronate dehydrogenase	COG0677	M	K02472
Bacteroides_caccae_153807847	2-succinyl-6-hydroxy-2,4-cyclohexadiene-1-carboxylate synthase	COG1165	H	K02551
Anaerotruncus_colihominis_167771309	N-acetylglutamate synthase (N-acetylornithine aminotransferase)	COG1364	E	K00618
B_WH2_003808	putative outer membrane protein			
Eubacterium_dolichum_160914195	putative copper-translocating P-type ATPase			K01529
Bacteroides_fragilis_53715551	Predicted ATPase	COG1373	R	
Clostridium_bolteae_160937654	putative phage protein			
Bacteroides_fragilis_53712550	Alkyl hydroperoxide reductase	COG3634	O	K03387
Parabacteroides_merdae_154492101	hypothetical protein			
Clostridium_bolteae_160936352	Uncharacterized conserved protein	COG2606	S	
Bacteroides_uniformis_160889340	TraM			
B_theta_7330_002089	Adenine-specific DNA methylase	COG0827/4646	KL	
B_WH2_003982	putative outer membrane protein			
Bacteroides_capillosus_154496743	hypothetical protein			
Clostridium_bolteae_160941240	putative citrate lyase			
Bacteroides_capillosus_154496327	putative v-type ATPase			
Bacteroides_capillosus_154496839	putative cobalamin biosynthesis protein			
Bacteroides_fragilis_60683742	Small-conductance mechanosensitive channel	COG0668	M	

Eubacterium_siraeum_167749611	putative transcriptional regulator			
Parabacteroides_distasonis_150007998	Cobyric acid synthase	COG1492	H	K02232
Parabacteroides_distasonis_150008480	putative pyruvate formate-lyase 3 activating enzyme			
Bacteroides_capillosus_154496329	Na+-transporting two-sector ATPase/ATP synthase			K01549/50
Bacteroides_capillosus_154496850	hypothetical protein			
Bacteroides_capillosus_154496749	putative spore maturation protein			
Bacteroides_capillosus_154496148	putative spore protease			
Clostridium_bolteae_160937655	DNA polymerase			K00961
Bacteroides_fragilis_60683107	Putative copper/silver efflux pump	COG3696	P	K07239/7787
Bacteroides_capillosus_154496295	putative short-chain dehydrogenase/reductase			
Anaerotruncus_colihominis_167771023	stage V sporulation protein AC			K06405
B_WH2_004992	ABC-type multidrug transport system	COG0842	V	K09686
Bacteroides_capillosus_154500409	Transcription antiterminator	COG0250	K	K02601
B_theta_3731_003445	putative tyrosine type site-specific recombinase	NOG36763		
B_WH2_003671	putative 3-oxoacyl-[acyl-carrier-protein] synthase			
Parabacteroides_distasonis_150010457	hypothetical protein			
Bacteroides_fragilis_60681723	putative hydrolase lipoprotein	NOG09493		
Clostridium_scindens_167758928	putative transcriptional regulator			
Bacteroides_capillosus_154498046	Exonuclease VII small subunit	COG1722	L	K03602
Ruminococcus_gnavus_154504691	putative phage protein			
Anaerotruncus_colihominis_167772969	hypothetical protein			
Bacteroides_caccae_153808785	Predicted nucleoside-diphosphate sugar epimerases	COG1086	MG	
Alistipes_putredinis_167751920	phosphoglycolate phosphatase			K01091
Anaerotruncus_colihominis_167772790	hypothetical protein			
Parabacteroides_merdae_154494124	putative transcriptional regulator			
Bacteroides_caccae_153809523	glycoside hydrolase family 29	COG3669	G	K01206
Bacteroides_fragilis_46242778	TraO conjugation protein			
Bacteroides_capillosus_154499075	putative site-specific recombinase			
Anaerotruncus_colihominis_163816273	putative DNA helicase			
Bacteroides_capillosus_154495881	Pentose-5-phosphate-3-epimerase	COG0036	G	K01783
Bacteroides_uniformis_160887913	hypothetical protein			
Dorea_longicatena_153853397	putative phage protein			
Bacteroides_vulgatus_150003721	putative outer membrane protein			
B_WH2_002145	putative outer membrane protein			
Bacteroides_capillosus_154500525	hypothetical protein			
Alistipes_putredinis_167752229	putative DNA primase	NOG22337		

*Protein sequences from *E.rectale, E.eligens, B.theta3731, B.theta7330, and B.WH2* can be found at http://gordonlab.wustl.edu/SuppData.html

Supplementary Table 15: BMI category in the Missouri Adolescent Female Twin Study[a].						
	Underweight (n=138)	**Lean (n=1893)**	**Overweight (n=711)**	**Obese I (n=309)**	**Obese II (n=174)**	**Obese III (n=113)**
EA (n=2860)	4.79	60.87	19.58	8.08	4.27	2.41
AA (n=478)	0.21	31.80	31.59	16.32	10.88	9.21

[a]All numbers are percentages. Underweight: <18.5 kg/m^2; Lean 18.5-24.9 kg/m^2; Overweight: 25-29.9 kg/m^2; Obese I: 30-34.9 kg/m^2; Obese II: 35-39.9 kg/m^2; Obese III: ≥40 kg/m^2.