Supporting Information

Lozupone *et al.* 10.1073/pnas.0807339105

SI Methods

Determining the Environmental Distribution of Sequenced Genomes. To obtain information on the lifestyle of the isolate and its source, we looked at descriptive information from NCBI (www.ncbi.nlm.nih.gov/genomes/lproks.cgi) and other related publications. We also determined which 16S rRNA-based environmental surveys of microbial assemblages deposited near-identical sequences in GenBank. We first downloaded the gbenv files from the NCBI ftp site on December 31, 2007, and used them to create a BLAST database. These files contain GenBank records for the ENV database, a component of the nonredundant nucleotide database (nt) where 16S rRNA environmental survey data are deposited. GenBank records for hits with >98% sequence identity over 400 bp to the 16S rRNA sequence of each of the 67 genomes were parsed to get a list of study titles associated with the hits. These study titles were used to evaluate the types of environments where close relatives of each of the 67 isolates had been found (Table S1 and Table S2).

Generation and Evaluation of the 16S rRNA Consensus Tree. 16S rRNA sequences were aligned with NAST (1) and hypervariable positions filtered out with lanemaskPH (2). This alignment was used to make 1,000 bootstrapped NJ trees using ClustalW and python code. The consensus tree was generated using the majority rule consensus method as implemented in PyCogent (3) and rooted between the Bacteria and the Archaea.

The 16S rRNA tree allowed us to evaluate the taxonomy of 21 newly sequenced genomes from the HGMI project. In accordance with previous studies that showed genera within the Firmicute class Clostridia to be extremely heterogeneous (4, 5), we found that some members of the genera *Clostridium*, *Ruminococcus*, and *Eubacterium* grouped with members of other named genera with high bootstrap support (Fig. 1A). One reported member of the Bacteroidetes (*Bacteroides capillosus*) grouped firmly within the Firmicutes. This taxonomic error was not surprising because gut isolates have often been classified as Bacteroides based on an obligate anaerobe, Gram-negative, nonsporulating phenotype alone (6, 7). A more recent 16S rRNA-based analysis of the genus Clostridium defined phylogenetically related clusters (4, 5), and these designations were supported in our phylogenetic analysis of the Clostridium species in the HGMI pipeline. We thus designated these Clostridium species, along with the species from other named genera that cluster with them in bootstrap supported nodes, as being within these clusters.

Annotation of GTs and GHs. NCBI accessions for GTs and GHs were retrieved from the CAZy website at www.cazy.org (8). GT and GH annotations for the 21 human gut-derived genomes from the HGMI were obtained with an automated procedure followed by manual curation. In the automated procedure, the protein models are compared to a library of ≈80,000 previously annotated GH and GT functional module sequences using a combination of BLAST (E <1e-6) (9) and HMMER (E <1e-6) with HMM models (10) of families derived from the same library. Distantly related sequences (those with borderline E-values and poor alignment coverage) were manually inspected to ascertain the preservation of key functional elements and active site residues, when known. This procedure provided consistent results with annotations at the CAZy website.

Note that the two gut Archaeal genomes were excluded from the GH-based genome clustering analysis because neither had any GH genes.

1. DeSantis TZ, *et al.* (2006) NAST: A multiple sequence alignment server for comparative analysis of 16S rRNA genes. *Nucleic Acids Res* 34:394–399.
2. Hugenholtz P (2002) Exploring prokaryotic diversity in the genomic era. *Genome Biol* 3:1–8.
3. Knight R, *et al.* (2007) PyCogent: A toolkit for making sense from sequence. *Genome Biol* 8:R171.
4. Collins MD, *et al.* (1994) The phylogeny of the genus Clostridium: Proposal of five new genera and eleven new species combinations. *Int J Syst Bacteriol* 44:812–826.
5. Wiegel J, Tanner R, Rainey FA (2006) An introduction to the family Clostridiaceae. *Prokaryotes*, (Springer, New York), Vol 4, pp 654–678.
6. Farrow JAE, Lawson PA, Hippe H, Gauglitz U, Collins MD (1995) Phylogenetic evidence that the gram-negative nonsporulating bacterium *Tissierella* (Bacteroides) *preaeacuta* is a member of the Clostridium subphylum of the gram-positive bacteria and description of *Tissierella creatinini sp nov*. *Int J Syst Bacteriol* 45:436–440.
7. Jensen NS, Canale-Parola E (1985) Nutritionally limited pectinolytic bacteria from the human intestine. *Appl Environ Microbiol* 50:172–173.
8. Coutinho PM, Henrissat B (1999) Carbohydrate-active enzymes: An integrated database approach. *Recent Advances in Carbohydrate Bioengineering*, eds Gilbert HJ, Davies G, Henrissat B, Svensson B (The Royal Society of Chemistry, Cambridge), pp 3–12.
9. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215:403–410.
10. Eddy SR (1996) Hidden Markov models. *Curr Opin Struct Biol* 6:361–365.

Table S1. Sequenced genomes from the Human Gut Microbiome Initiative

Full Name	Lineage	Environmental Distribution (>98% ID hits to ncbi ENV database)	Description/Original Isolation	RefSeq
Bacteria: Bacteroidetes; Bacteroidetes (class); Bacteroidales				
Bacteroides caccae ATCC 43185	Bacteroidaceae	Human gut	Normal human gut microbiota and purported pathogenic factor in inflammatory bowel disease / Human feces	NZ_AAVM00000000
Bacteroides ovatus	Bacteroidaceae	Human gut	Normal human gut microbiota	NZ_AAXF00000000
Bacteroides stercoris	Bacteroidaceae	Human gut	Normal human gut microbiota / Human feces	NZ_ABFZ00000000
Bacteroides uniformis	Bacteroidaceae	None	Normal human gut microbiota	NZ_AAYH00000000
Parabacteroides merdae	Porphyromonadaceae	Human gut	Normal human gut microbiota / Human feces	NZ_AAXE00000000
Bacteria; Actinobacteria; Coriobacteridae				
Collinsella aerofaciens	Coriobacteriales; Coriobacteriaceae	Human gut	Normal human gut microbiota / Human feces	NZ_AAVN00000000
Bacteria; Firmicutes; Mollicutes				
Eubacterium dolichum DSM 3991		Human gut	Normal human gut microbiota / Human feces	NZ_ABAW00000000
Bacteria; Firmicutes; Clostridia; Clostridiales				
Clostridium bolteae ATCC BAA-613	Cluster XIVa	Human gut	Normal human gut microbiota / Human feces	NZ_ABCC00000000
Ruminococcus obeum ATCC 29174	Cluster XIVa	Human and pig gut	Normal human gut microbiota / Human feces	NZ_AAVO00000000
Ruminococcus gnavus ATCC 29149	Cluster XIVa	Human gut	Normal human gut microbiota / Human feces	NZ_AAYG00000000
Clostridium scindens ATCC 35704	Cluster XIVa	Human and mouse gut	Normal human gut microbiota / Human feces	NZ_ABFY00000000
Dorea longicatena DSM 13814	Cluster XIVa	Human gut	Normal human gut microbiota / Human feces	NZ_AAXB00000000
Ruminococcus torques ATCC 27756	Cluster XIVa	Human gut	Normal human gut microbiota / Human feces	NZ_AAVP00000000
Eubacterium ventriosum ATCC 27560	Cluster XIVa	Human gut	Normal human gut microbiota / Human feces	NZ_AAVL00000000
Coprococcus eutactus ATCC 27759	Cluster XIVa	Human gut	Normal human gut microbiota / Human feces	NZ_ABEY00000000
Clostridium sp. L2–50	Cluster XIVa	Human gut	Normal human gut microbiota / Human feces	NZ_AAYW00000000
Peptostreptococcus micros ATCC 33270	Peptostreptococcaceae	Human mouth, gut, and lung	Normal microbiota in human mouth and gut: pathogen / Lung infection	NZ_ABEE00000000
Faecalibacterium prausnitzii M21/2	Cluster IV	Human and pig gut	Normal human microbiota flora / Human feces	NZ_ABED00000000
Eubacterium siraeum DSM 15702	Cluster IV	Human gut	Normal human gut microbiota / Human feces	NZ_ABCA00000000
Anaerotruncus colihominis DSM 17241	Cluster IV	Human and avian gut	Normal human gut microbiota / Human feces	NZ_ABGD00000000
Bacteroides capillosus ATCC 29799	Cluster IV	Human and turkey gut	Normal human gut microbiota: human abscesses and wounds	NZ_AAXG00000000

Table S2. Previously sequenced genomes included in the analysis

Full Name	Lineage	Environmental Distribution (>98% ID hits to ncbi ENV database)	Description/Original Isolation	Gut(G) Non-Gut(N)	RefSeq
Bacteria;Proteobacteria					
Escherichia coli HS	Gammaproteobacteria; Enterobacteriales; Enterobacteriaceae	Cosmopolitan in free-living communities and abundant in vertebrate gut	Non-disease causing strain/Human gut	G	NC_009800
Pseudoalteromonas atlantica T6c	Gammaproteobacteria; Alteromonadales; Pseodoalteromonadaceae	Seawater	Associated with marine biofilms and shell disease in shellfish/ Diseased crab shells	N	NC_008228
Shewanella frigidimarina NCIMB 400	Gammaproteobacteria; Alteromonadales; Shewanellaceae	Marine water, ice, coral and saline lakes	Marine bacterium/Water from the North Sea	N	NC_008345
Hahella chejuensis KCTC 2396	Gammaproteobacteria; Oceanospirillales; Hahellaceae	Marine	Marine/ Marine sediment	N	NC_007645
Alcanivorax borkumensis SK2	Gammaproteobacteria; Oceanospirillales; Alcanivoracaceae	Marine sediment, water, and coral	Marine/ Marine sediment	N	NC_008260
Roseobacter denitrificans OCh 114	Alphaproteobacteria; Rhodobacterales; Rhodobacteraceae	Diverse Saline Environments (Marine Ice, bacterioplankton, and sediment)	Marine anaerobe/ Marine sediment	N	NC_008209
Mesorhizobium sp. BNC1	Alphaproteobacteria; Rhizobiales; Phyllobacteriaceae	Soil, waste-water treatment	Sewage / Sewage	N	NC_008254
Bacteria:Bacteroidetes					
Parabacteroides distasonis ATCC 8503	Bacteroidetes (class); Bacteroidales; Porphyromonadaceae	Vertebrate Gut	Human gut	G	NC_009615
Bacteroides vulgatus ATCC 8482	Bacteroidetes (class); Bacteroidales; Bacteroidaceae	Vertebrate Gut	Human gut / Human feces	G	NC_009614
Bacteroides thetaiotaomicron VPI-5482	Bacteroidetes (class); Bacteroidales; Bacteroidaceae	Vertebrate Gut	Human gut / Feces of a healthy adult	G	NC_004663
Bacteroides fragilis NCTC 9343	Bacteroidetes (class); Bacteroidales; Bacteroidaceae	Vertebrate Gut	Human gut: opportunistic pathogen	G	NC_003228
B. fragilis YCH46	Bacteroidetes (class); Bacteroidales; Bacteroidaceae	Vertebrate Gut	Human gut: opportunistic pathogen / patient with septicemia	G	NC_006347
Porphyromonas gingivalis W83	Bacteroidetes (class); Bacteroidales; Porphyromonadaceae	Human mouth	Human mouth: associated with periodontal disease / human infection	N	NC_002950
Flavobacterium johnsoniae UW101	Flavobacteria; Flavobacteriales; Flavobacteriaceae	Soil	Common in soil and freshwater / Soil	N	NC_009441
Flavobacterium psychrophilum JIP02/86	Flavobacteria; Flavobacteriales; Flavobacteriaceae	Fish-associated; Marine and Glacial Ice	Causes disease in salmonid fish / Kidney of a rainbow trout fry	N	NC_009613
Gramella forsetii KT0803	Flavobacteria; Flavobacteriales; Flavobacteriaceae	None	Seawater / North Sea during a phytoplankton bloom	N	NC_008571
Cytophaga hutchinsonii ATCC 33406	Sphingobacteria; Sphingobacteriales; Flexibacteraceae	Soil	Soil, freshwater, and marine habitats	N	NC_008255
Salinibacter ruber DSM 13855	Sphingobacteria; Sphingobacteriales	Salt crust; hypersaline lake; Saltern	Extremely halophilic aerobe / saltern crystallizer pond	N	NC_007677
Bacteria;Actinobacteria;Actinobacteridae					
Bifidobacterium adolescentis ATCC 15703	Bifidobacteriales; Bifidobacteriaceae	Human gut	Human gut commensal / infant feces	G	NC_0086181
Bifidobacterium longum NCC2705	Bifidobacteriales; Bifidobacteriaceae	Vertebrate gut and human vagina	Gut probiotic/ infant feces	G	NC_004307

Full Name	Lineage	Environmental Distribution (>98% ID hits to ncbi ENV database)	Description/Original Isolation	Gut(G) Non-Gut(N)	RefSeq
Leifsonia xyli subsp. xyli str. CTCB07	Actinomycetales; Micrococcineae; Microbacteriaceae	Cosmopolitan free-living/plant-associated	Causes sugar-cane disease	N	NC_006087
Propionibacterium acnes KPA171202	Actinomycetales; Propionibacterineae; Propionibacteriaceae	Cosmopolitan in free-living communities	Causes acne/ human skin	N	NC_006085
Frankia sp. Ccl3	Actinomycetales; Frankineae; Frankiaceae	Plant-associated	nitrogen-fixing plant commensal/root nodules	N	NC_007777
Bacteria;Firmicutes;Clostridia;Clostridiales					
Clostridium thermocellum ATCC 27405	Clostridiaceae [Cluster IV]	Compost	Thermophilic anaerobe that degrades cellulose	N	NC_009012
Clostridium phytofermentans ISDg	Clostridiaceae [Cluster XIVa]	None	Anaerobe that degrades cellulose/ forest soil	N	NC_010001
Clostridium acetobutylicum ATCC 824	Clostridiaceae [Cluster I]	Industrial	Industrially important soil bacterium/garden soil	N	NC_003030
Clostridium beijerinckii NCIMB 8052	Clostridiaceae [Cluster I]	Soils and non-saline sediments; many contaminated	Industrial production of solvents	N	NC_009617
Alkaliphilus metalliredigens QYMF	Clostridiaceae [Peptostreptococcaceae]	None	alkiphilic, moderately halophilic metal-reducing bacterium/ borax leachate ponds	N	NC_009633
Bacteria;Firmicutes;Mollicutes					
Acholeplasma laidlawii PG-8A	Acholeplasmatales; Acholeplasmataceae	None	Ubiquitous species of mycoplasmas, in both free-living and animal-associated communities	N	NC_010163
Bacteria; Firmicutes;Lactobacillales;					
Lactobacillus acidophilus NCFM	Lactobacillaceae	Abundant in the vertebrate gut and human vagina	Used for infant formula, yogurt, and fluid milk production/Human gut	G	NC_006814
Lactobacillus johnsonii NCC 533	Lactobacillaceae	Predominantly in the verebrate gut but also the human vagina	Human gut bacteria that inhibits pathogens; probiotic / Human gut	G	NC_005362
Lactobacillus gasseri ATCC 33323	Lactobacillaceae	Predominantly in the verebrate gut but also the human vagina	Major homofermentative Lactobacillus in human intestinal tract.	G	NC_008530
Lactobacillus reuteri F275	Lactobacillaceae	Vertebrate gut	Human gut bacteria that inhibits pathogens; probiotic / Human gut	G	NC_009513
Lactobacillus salivarius UCC118	Lactobacillaceae	Vertebrate gut	Human gut bacteria that inhibits pathogens; probiotic / Human gut	G	NC_007929
Lactobacillus brevis ATCC 367	Lactobacillaceae	Fermented food and household waste	Starter culture for beer, sourdough, and silage.	N	NC_008497
Oenococcus oeni PSU-1		None	Marshes; Used for secondary fermentation in wine production.	N	NC_008528
Bacteria; Firmicutes;Bacillales					
Geobacillus kaustophilus HTA426	Bacillaceae	Deep sea sediments, soil, petroleum reservoir	Moderate thermophile/ deep sea sediment	N	NC_006510
Bacillus cereus ATCC 14579	Bacillaceae	Mostly soil	Soil organism that can cause food poisoning	N	NC_004722
Bacillus amyloliquefaciens FZB42	Bacillaceae	Air, sediments, soil, subsurface	Plant-growth-promoting rhizosphere bacterium / soil	N	NC_009725
Bacillus clausii KSM-K16	Bacillaceae	Air, termite gut	Used on industrial production of alkaline enzymes/ soil	N	NC_006582
Bacillus halodurans C-125	Bacillaceae	None	Alkaliphilic organism used in the industrial production of β-galactosidase and xylanase	N	NC_002570

Full Name	Lineage	Environmental Distribution (>98% ID hits to ncbi ENV database)	Description/Original Isolation	Gut(G) Non-Gut(N)	RefSeq
Archaea; Euryarchaeota;					
Methanobrevibacter smithii ATCC 35061	Methanobacteria; Methanobacteriales; Methanobacteriaceae	Sheep, cattle, and reindeer rumen; sediment	anaerobic soil and sediment, and the gut of humans, ruminants and other animals/ sewage	G	NC_009515
Methanosphaera stadtmanae DSM 3091	Methanobacteria; Methanobacteriales; Methanobacteriaceae	Vertebrate feces; sediment	Methanogenic / human feces	G	NC_007681
Methanothermobacter thermautotrophicus str. Delta H	Methanobacteria; Methanobacteriales; Methanobacteriaceae	Sludge digestors; petroleum reservoirs; deep crustal fluid; anoxic soil, compost	thermophilic, anaerobic sewage sludge digestors / sewage	N	NC_000916
Methanocaldococcus jannaschii DSM 2661	Methanococci; Methanococcales; Methanocaldococcaceae	Hydrothermal vents; deep –sea sediment	obligately anaerobic methane-producing thermophile/ deep sea vent	N	NC_000909
Methanococcoides burtonii DSM 6242	Methanomicrobia; Methanosarcinales; Methanosarcinaceae	Cold sediment from the ocean and saline lakes	Psychrotolerant, methanogenic / anoxic water in Ace Lake, Antarctica	N	NC_007955

Table S5. Comparison of GT and GH UPGMA clustering results to the 16S rRNA phylogeny

Cluster type	Fraction shared nodes	Cluster parsimony count	16S rRNA parsimony count	Probability better than phylogeny	Probability better than chance
GT UniFrac UW	0.46	11	13	≤0.001	≤0.001
GT UniFrac W	0.46	10	13	0.001	≤0.001
GT Count UW	0.09	14	13	0.99	≤0.001
GT Count W	0.03	16	13	1.0	0.003
GH UniFrac UW	0.40	9	11	≤0.001	≤0.001
GH UniFrac W	0.31	10	11	0.22	≤0.001
GH Count UW	0.10	12	11	0.98	≤0.001
GH Count W	0.10	13	11	1.0	0.001

UW, unweighted; W, weighted.

Other Supporting Information Files

Table S3

Table S4