Supporting Online Material for

Metagenomic Analysis of the Human Distal Gut Microbiome

Steven R. Gill,* Mihai Pop, Robert T. DeBoy, Paul B. Eckburg, Peter Turnbaugh, Buck S. Samuel, Jeffrey I. Gordon, David A. Relman, Claire M. Fraser-Liggett, Karen E. Nelson

*To whom correspondence should be addressed. E-mail: srgill@buffalo.edu

Published 2 June 2006, *Science* **312**, 1355 (2006).
DOI: 10.1126/science.1124234

This PDF file includes

Science Supporting Online Material

Materials and Methods

Human subjects and extraction and purification of DNA. The use of human subjects was approved by the Stanford University Administrative Panel on Human Subjects in Medical Research; all participants signed informed consent. The two subjects from which samples were obtained, ages 28 and 37, female and male respectively, had not used antibiotics or any other medications during the year prior to specimen collection. Neither had been diagnosed with any significant medical condition, nor had been subjected to surgical procedures. One of these individuals followed a vegetarian diet and the other an unrestricted diet; and one had traveled to France and Brazil during the year prior to specimen collection. Approximately 0.3 g of fecal material from each of two healthy human subjects (referred to here as Subject-7 and Subject-8) was resuspended in 5 ml of ice-cold 0.05 M potassium phosphate buffer and divided into separate 1 ml aliquots. To isolate total fecal genomic DNA, 10 µl Proteinase K (20 mg/ml) and 50 µl 10% SDS was added to each aliquot, and the reaction mixtures were incubated for 1 hour at 55°C. Phenol (150 µl; pH 7.5) was added to the suspension that was then transferred to FastPrep lysing matrix D and agitated in a FastPrep FP120 instrument (Qbiogene, Carlsbad, CA) at 6.0 m/s for 40 s. The lysate was extracted twice with phenol/chloroform followed by one extraction in chloroform. DNA was precipitated with ethanol resuspended in TE buffer, and subjected to a final clean-up (Qiagen QIAamp mini spin columns; Qiagen Inc., Valencia, CA). The purified DNA was then used for construction of 16*S* rDNA libraries, and plasmid-based small insert (2-10 kb) libraries. Because the DNA was isolated using disruptive mechanical lysis with a FastPrep lysing matrix, the size of the purified DNA (2-10 kb) was not sufficient for construction of fosmid libraries.

Bacterial 16*S* rDNA amplification. Full-length 16*S* rDNA was amplified from broad-range bacterial primers Bact-8F (5'-AGAGTTTGATCCTGGCTCAG-3') and Bact-1510R (5'-CGGTTACCTTGTTACGACTT-3') (*S1*). Each 100-µl PCR reaction contained 100 ng DNA, 100 ng of each primer, 200 mM dNTPs, 1.5 mM $MgCl_2$ and 2.5 units of Platinum *Taq* DNA polymerase (Invitrogen, Carlsbad, CA) in the manufacturer's buffer. Conditions for PCR reactions were 94°C for 60 s, followed by 30 cycles of 94°C for 60 s, 55°C for 30 s, 72°C for 60 s and a final extension of 72°C for 9 min. Amplification of group- and species-specific 16*S* rDNA was performed using species-specific primers and reaction conditions described in Bartosch *et al.* (*S2*). Amplified PCR products were resolved by agarose gel electrophoresis, excised from the gel and extracted using QIAquick™ Gel Extraction Kit (Qiagen).

Cloning and sequencing. PCR products were cloned into pCR-4-TOPO vectors (Invitrogen) and electroporated into electrocompetent *E. coli* DH10B (Invitrogen). Transformants were selected by plating onto selective SOB/amp agar plates which were incubated overnight. Colonies were picked randomly and grown by overnight incubation in SOB/amp media. Plasmid template DNA from each transformant was prepared by a modified alkaline lysis method. The 16*S* rDNA nucleotide sequences of the clone inserts were determined by cycle sequencing using BigDye Terminator

(Applied Biosystems, Foster City, CA) and 3.2 pmol of M13F and M13R sequencing primers. Sequences were analyzed on ABI 3730xl sequencers (Applied Biosystems) and trimmed to remove vector sequence. After trimming and adjusting for quality values, the average single sequence read length was at least 900 nucleotides in length.

Construction and sequencing of human distal gut metagenome libraries. Total fecal genomic DNA was suspended in sucrose nebulization buffer (50% sucrose, 0.3M NaOAc, 1X TE) and nebulized to generate random fragments ranging in size from 2 to 8 kb. Fragments between 2 and 3 kb were isolated from the nebulized DNA by agarose gel purification and extracted using QIAquick™ Gel Extraction Kit (Qiagen). Purified DNA fragments were modified by addition of Bst XI linkers to their ends, ligated to Bst XI digested pHOS2 vector (*S3*) and electroporated into *E. coli* DH10B. Transformants were selected by plating onto selective SOB/amp agar plates. The percentage of colonies or plasmid clones with inserts was estimated by agarose gel analysis of pooled plasmid DNA from a lawn of *E. coli* transformants. Approximately 90% of the plasmid clones in the metagenome libraries contained random fecal DNA inserts. Sequencing reactions were carried out using M13F and M13R primers as described above.

Random sequence assembly. Sequence data were assembled with Celera Assembler (*S4*) as a combination of shotgun reads from Subject-7 and Subject-8 in order to overcome the extremely low coverage provided by the shotgun reads. The combined assembly resulted in increase coverage for the organisms present in both samples. The resulting assembly was separated into Subject-7– and Subject-8–specific assemblies using software developed for this project as well as components of the AMOS assembly package (http://amos.sourceforge.net). To generate a subject-specific assembly within every contig in the combined assembly we retained only the reads obtained from one of the subjects, then we regenerated the contig consensus by recomputing the multiple alignment of the remaining reads. Regions of the contig not supported by any reads from the chosen subject were replaced with Ns. Optimal combined assembly was attained by lowering the threshold for statistical repeat detection (unitigger parameter –j set to –20) so that the most abundant microbial species would not be characterized as repeats and the representation of low abundance microorganisms would be increased (*S4*). Furthermore, all singleton reads (reads not placed into contigs) were mapped back to the contigs using the nucmer and show-tiling programs from the MUMmer package (*S5*), in order to identify those reads that were incorrectly omitted by the assembler. These reads were excluded from further analysis to avoid duplications. The output of Celera Assembler consists of a collection of contiguous DNA pieces (contigs) linked together by paired end reads to form scaffolds. The depth of sequence coverage in the output of Celera Assembler was computed using the cvgChop program from the AMOS package.

Assembly validation. Validating the correctness of an assembly is still the subject of active research even when assembling a single organism (see for example the many arguments about the quality of the various assemblies of the human genome), let alone the case of environmental sequences. Currently, other than manual curation of the data (impractical for this project) the most reliable validation method is alignment to a previously sequenced reference genome. We performed such a test by aligning the contigs produced by Celera Assembler to the draft genome of *M. smithii*. Our assembly agreed well with the reference genome indicating no major assembly problems. While

we cannot rule out the possibility that some of the contigs are mis-assembled, representing chimeras between different organisms, we believe that the majority rule used in assigning a contig to a specific taxonomic unit (see below) is immune to such assembly problems. Furthermore, we believe more errors were introduced in our analysis by the absence of reliable sequence information in public databases than by assembly errors.

Database search parameters. The BLASTX and BLASTP programs were used to identify putative open reading frames (ORFs) and to assign putative functions to these ORFs. Searches were performed against AllGroup.niaa; a non-redundant, in-house repository of protein data obtained from GenBank, Uniprot, Protein Research Foundation (PRF), Protein Data Bank (PDB), and Omnium. Only matches of at least 50 amino acids in length, with a *P* value lower than $1e^{-15}$ and an identity greater than 35% were considered for further analysis.

Taxonomic assignment of random shotgun sequences. Two complementary analyses were performed to ascertain the representation of species from the bacterial, archaeal, and eukaryotic domains. Initially, the shotgun sequence assembly was mapped to a database of known 16*S* rDNA bacterial and archaeal sequences (RDP version 9.2 [bacterial], RDP version 8.1 (http://rdp.cme.msu.edu) [archaeal]) using BLASTN. The best matches (as defined by the product of percent identity and length) of rDNA sequences to the contigs and singletons were selected if they were longer than 200 base pairs (bp). Each contig or singleton that matched a known rDNA sequence was assigned to that organism.

In the second analysis, putative ORFs within the assembled contigs were identified using the long-orfs program from the Glimmer package (*S6*). The protein sequences corresponding to these ORFs were searched with BLASTP against AllGroup.niaa as described in Database search parameters. Only matches as defined above (in database search parameters) were considered for further analysis. For each ORF, the best BLASTP hit was used to assign the ORF to a taxonomic unit. The taxonomic assignments of the ORFs were used to assign entire contigs to specific taxonomic units. Due to the paucity of reliable phylogenetic anchors in public databases, the ORF data provides us with a reasonable approximation for the origin of each genomic fragment. In many situations multiple ORFs present in a single contig all agreed on a specific taxonomic assignment. In situations where the ORFs present in the same contig disagreed on the taxonomic assignment we chose the majority assignment if one existed; otherwise, we assigned the contig to the lowest taxonomic level where all assignments agreed (the later situation occurred in 697 of 6775 ORF-containing contigs from Subject-7 and in 1072 of 8154 ORF-containing contigs from subject-8). For the purposes of this analysis, a majority assignment is the assignment chosen by the largest number of ORFs (n) such that the second largest assignment is shared by at most n/2 ORFs. For the purposes of this study this simple procedure provides us with sufficiently reliable information, especially for assignments to taxonomic units well represented in public databases. In general, however, the specific assignments should be treated with caution due to the insufficient data available as well as to the previously reported limitations of BLAST as a tool for phylogenetic reconstruction (*S7*).

All contigs where the long-orfs program could not identify any putative ORFs, as well as all singletons, were searched using the same cut-off criteria as for BLASTP. The BLASTX matches represent either partial genes, or genes missed by the long-orfs program. All contigs and singletons containing reliable BLASTX matches were assigned to a specific taxonomic unit as described in the previous paragraph.

Organism reconstruction aided by phylogenetic markers. Partial chromosome assemblies from organisms that could be identified from their phylogenetic markers were constructed as follows. Seed contigs were generated by extracting all contigs associated with the selected phylogenetic markers, and then all contigs linked to the seed contigs were extracted by mate-pair information. Singleton reads were subsequently mapped to the selected contigs using the MUMmer alignment program (*S5*). Both reads contained in the selected contigs, and those later recruited through alignment and the use of mate-pair information, were assembled using Celera Assembler to obtain a larger assembly of the organisms of interest.

16*S* rDNA phylogenetic analysis and phylotype determination. Approximately 1000 PCR-amplified near-full length (1400-1500 bp), non-chimeric (see below) bacterial 16*S* rDNA sequences from each stool sample were subjected to detailed phylogenetic analysis (1024 sequences from subject-7 and 1038 from subject-8; total 2062 sequences). In addition, 132 partial-length bacterial 16*S* rDNA sequences (good quality sequences ≥500 bases in length) and 8 partial-length archaeal sequences (291-714 bases) from the random shotgun assembly were analyzed phylogenetically (73 bacterial and 4 archaeal sequences from subject-7; 59 bacterial and 4 archaeal sequences from subject-8). The 16*S* rDNA sequences were aligned to the small subunit rDNA Ribosomal Database Project II (RDP-II) (*S8*) running locally in the ARB package (*S9*) using FastAligner v1.03. Each sequence was manually edited in conjunction with its chromatogram and secondary structure information. Only unambiguous nucleotide positions were included in the analysis and primer sequences were excluded, totaling 1197 filter positions for all near-full-length bacterial 16*S* rDNA sequences, 341–680 positions for partial-length bacterial sequences, and 290–606 positions for partial-length archaeal sequences. The closest neighbors to our sequences among the RDP-II and NCBI GenBank databases were determined using the maximum likelihood algorithm, and such sequences were used to represent phylotypes where possible. Sequences were tested for possible chimeras using Chimera Check v2.7 (online analysis at RDP-II website, http://rdp.cme.msu.edu/cgis/chimera.cgi?su=SSU), and sequences without close RDP-II neighbors were subjected to a manual online BLAST analysis (NCBI website, http://www.ncbi.nlm.nih.gov/BLAST/). All chimeras, human sequences, vector sequences, and sequences of poor quality were deleted from further phylogenetic analysis. A phylogenetic tree with a representative from each phylotype was generated using a neighbor-joining algorithm from a Felsenstein-corrected distance matrix.

Sequences were grouped into phylotypes using Felsenstein-corrected similarity matrices such that the least similar pair within the phylotype shared at least 99% similarity for near-full-length 16*S* rDNA sequences and at least 97% similarity for partial-length sequences. Additionally, distance matrices were analyzed with the DOTUR program to identify phylotypes at every similarity cutoff value using the furthest-neighbor algorithm with 0.001 precision (*S10*). Sequences already present in

public databases (preferably from named organisms) were chosen to represent each phylotype whenever possible. If no such sequence met the criteria for inclusion in a phylotype, it was considered a novel phylotype, and a representative was arbitrarily selected. Novel phylotypes and sequences with uncultured GenBank neighbors were considered to represent uncultured species. Good's coverage estimation was calculated as [1-(n/N)] x 100, where n is the number of singletons and N is the total number of sequences. Sequence rarefaction curves were created in EstimateS version 7.50 (R.K. Colwell, http://viceroy.eeb.uconn.edu/EstimateS), using 100 randomizations, sampling without replacement, and bias-corrected Chao1 estimation.

Role category assignments. For all identified genes, translation start and stop sites were refined from the coordinates of the BLASTX alignments. These evidence-based genes were given non-hypothetical names and organism assignments using their BLAST results, and cellular role categories (*S11*) were assigned in an automated fashion. Frameshifts were identified and genes merged when two adjacent ORFs shared a common database accession.

Comparative assembly. The comparative assembly program AMOScmp (*S12*) was used to identify organisms closely related to previously sequenced species. We iteratively used as a reference the complete genomes of *B. longum* (*S13*), *Bacteroides thetaiotaomicron* (*S14*), as well as a draft assembly of *Methanobrevibacter smithii* (*S15*).

Orthologous groups analysis. The DNA sequences of contigs and singletons from Subject-7 and Subject-8 were searched using BLASTX and a protein database from NCBI containing a total of 4891 COGs (clusters of orthologous groups). Each COG consists of individual proteins or groups of paralogs from at least 3 major phylogenetic lineages. The BLASTX search results were filtered for matches to individual proteins from at least two distinct phylogenetic groups for each COG, and having e-values less than 1e-5.

The expected COG accumulation curve was computed analytically with the Mao Tau calculation in EstimateS version 7.50 (R.K. Colwell, http://viceroy.eeb.uconn.edu/EstimateS), using the number of individual COGs (from the total of 62,036 identified COGs) per unique COG function (n = 2407). A sampling without replacement algorithm (without randomizations) was used for rarefaction of the bias-corrected Chao1 and ACE estimators of COG richness in EstimateS.

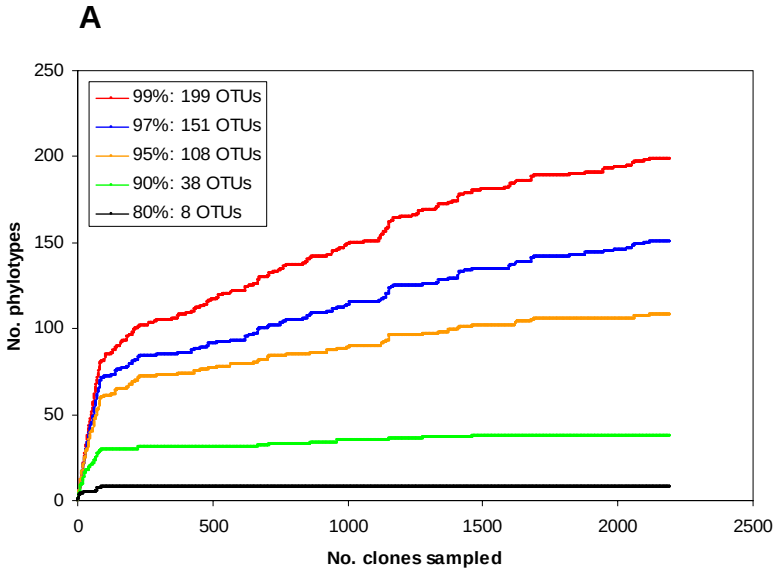
***In silico* reconstructions of the microbiome's metabolome**. Identified genes (BLASTX against ALLGroup.niaa with e<1e-15, open reading frames >50 amino acids with >35% identity to known sequences) were assigned enzyme commission numbers (ECs) and imported into the Washington University KEGG annotation viewer (WUBear; http://gordonlab.wustl.edu/supplemental/Gill/). For statistical analyses, highly redundant ECs (i.e. those with a dash) were removed, all ECs were converted to KEGG orthologous (KO) groups, and KEGG pathway hits were tallied using the latest release of KEGG (version 37). All predicted genes with Swiss-Prot accession numbers were imported into STRING, an extended COG database currently containing 179 microbial genomes, 163 of which are microbial (*S1*S*6*). These approaches allowed us to assign 17% and 31% of identified genes to KEGG maps and COG terms, respectively. Two metrics were then used to define whether a KEGG pathway or a COG term was enriched in the

human distal gut microbiome: (i) an odds ratio and (ii) a probability derived from a binomial distribution. The odds ratio can be thought of as the relative risk of observing a given group in the sample relative to the comparison dataset. We calculated the odds ratios using $(A/B)/(C/D)$ where A is the number of hits to a given category in the human gut microbiome, B is the number of hits to all other categories in the human gut microbiome, C is the number of hits to a given category in the comparison dataset, and D is the number of hits to all other categories in the comparison dataset. The binomial distribution was used to allow sampling with replacement, since the same gene could be present multiple times in each gut dataset. We used the same input values for the binomial to calculate the probability of observed A hits given $A + B$ total chances. This calculation assumes that the genes are normally distributed across metabolic groups and that the expected frequency is $C/(C + D)$ (the frequency observed in the comparison dataset).

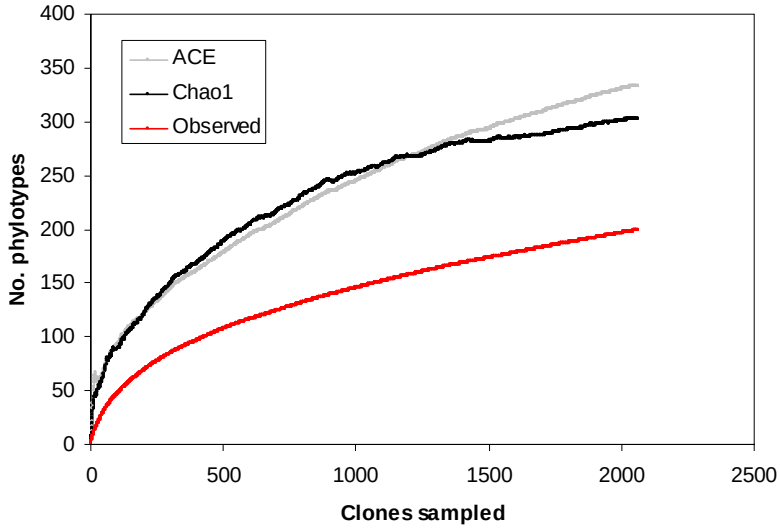
To minimize false negatives, no corrections were made for multiple testing. For COG analysis, the distribution of genes in the STRING database (163 microbial genomes) was used to calculate the odds ratio and binomial probability of obtaining the observed number of hits in each group. For KEGG analysis, the statistical calculations were performed against three separate comparison datasets: all bacteria in KEGG (total of 202 genomes), the *Homo sapiens* genome, and all archaea in KEGG (21 genomes). These comparisons allowed us to identify metabolic properties that were characteristic of the gut microbiome, that were characteristic of the bacterial members of the community, and that could potentially endow the human host with additional physiological traits. Groups with an odds ratio > 1 and $P < 0.05$ were defined as enriched and groups with an odds ratio < 1 and $P < 0.05$ as under-represented.

Supporting Figures

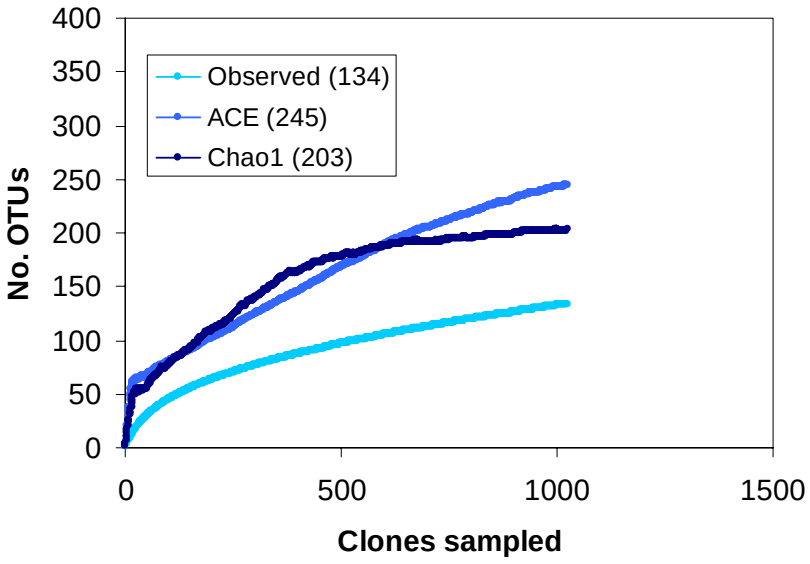
Figure S1. (**A**) Rarefaction curves for the combined 16*S* rDNA stool sequences at multiple phylotype cutoff levels. The curves appear to flatten as the phylotype cutoffs are relaxed below 95%, and every clone has been sampled more than once at the 80% cutoff. The total numbers of phylotypes (OTUs) are listed in the inset. (**B**) Collector's curves of observed and estimated phylotype richness for the combined subject data set. Each curve represents phylotype richness as clones are selected in a random order. Richness was estimated by two different calculations, abundance-based coverage estimation (ACE) and Chao1 richness estimation. Final richness estimations were 334 (ACE) and 303 (Chao1) phylotypes; however, the lack of plateaued curves indicates that both observed and estimated richness will increase with continued clone sequencing. (**C**) and (**D**) Collector's curves of observed and estimated phylotype richness for stools 7 and 8 are shown separately in (C) and (D), respectively. Numbers of phylotypes are presented in parentheses



B


ACE
Chao1
Observed
No. phylotypes
Clones sampled
0
50
100
150
200
250
300
350
400
0
500
1000
1500
2000
2500

C


Observed (134)
ACE (245)
Chao1 (203)
No. OTUs
Clones sampled
0
50
100
150
200
250
300
350
400
0
500
1000
1500

D

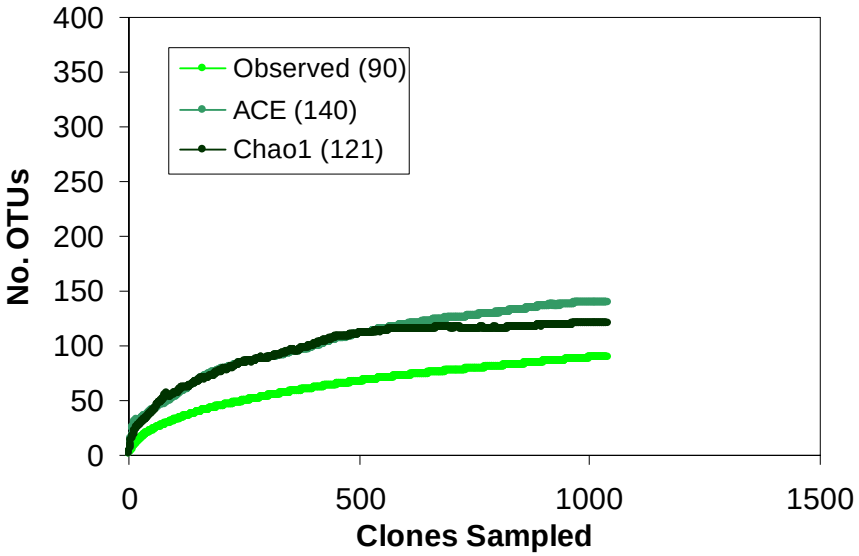
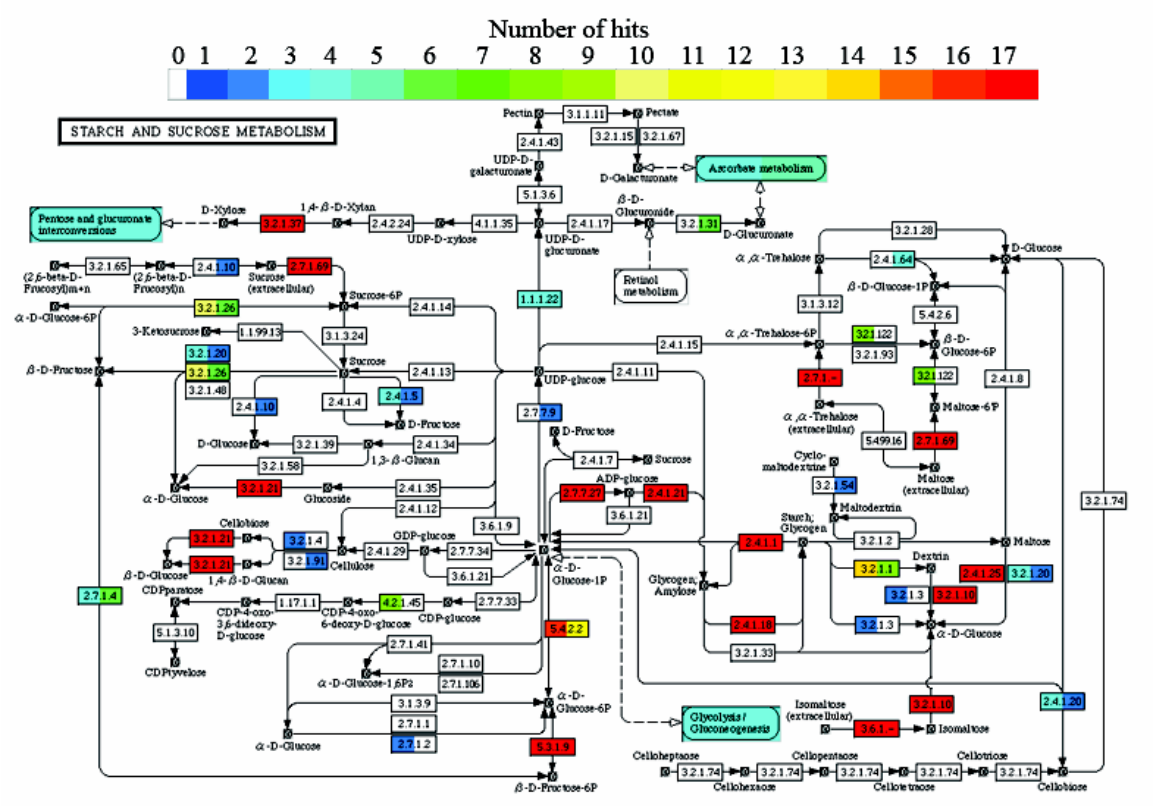

400
350
300
250
200
150
100
50
0
No. OTUs
Observed (90)
ACE (140)
Chao1 (121)
0
500
1000
1500
Clones Sampled

Fig. S2. Enrichment for genes in the starch metabolism pathway in the human distal gut microbiome. The left and right sides of each boxed Enzyme Commission (EC) number indicate whether the microbial gene product is present in Subjects-7 and 8, respectively, and to what extent (color scale: white no hits; red ≥17 hits). A total of 301 out of 8625 predicted genes are found in this metabolic network ($p<0.001$).



Supporting Tables

Table S1. Assignment of shotgun data to specific orders using two complementary methods: shotgun rDNA-contigs and singletons that matched a 16*S* rDNA operon from the specified order; best hit BLASTX contigs-contigs that contain complete genes (as identified by Glimmer long-orfs) or partial genes (as identified by BLASTX) whose best database match belong to the specific order; best match BLASTX singletons-singletons that contain partial gene (identified by BLASTX) whose best database hit belongs to the specific order. All results separated into the two human subjects (7 and 8). The numbers represent the number of bases from the shotgun data that can be associated with a specific order. When interpreting this Table it is important to take into account the reliability of each source of information. Assignments based on the 16*S* rDNA are the most reliable due to the large amount of data available for this marker, however this method can only be applied to the relatively few contigs or singletons that contain a portion of the 16*S* rDNA. At the other end of the scale are the assignments of singleton reads which represent regions of extremely low coverage and contain only partial gene fragments. In addition, for any specific genome in our samples, the number of singletons is inversely proportional to the quality of the assembly. One example of this phenomenon is the order *Methanobacteriales* which appears to be poorly represented in the "best hit BLASTX singletons" column even though this order is well represented in our samples. The singleton data should therefore be used cautiously with its main role being to supplement the contig data for the low-abundance organisms that could not be properly assembled.

Taxonomy				shotgun rDNA				best hit blastx contigs				best hit blastx singletons			
				7	%	8	%	7	%	8	%	7	%	8	%
BACTERIA	Firmicutes	Bacilli	Bacillales	1,201	0.81		0.00	994,084	9.94	1,597,352	13.67	1,475,166	15.02	1,270,429	14.96
			Lactobacillales	853	0.58	930	0.68	973,729	9.74	1,450,336	12.41	1,005,572	10.24	1,016,490	11.97
		Clostridia	Clostridiales	70,055	47.38	102,140	74.16	1,749,529	17.50	3,167,301	27.10	2,646,766	26.96	2,394,773	28.20
			Thermoanaerobacteriales		0.00		0.00	625,953	6.26	1,027,484	8.79	1,003,757	10.22	978,474	11.52
			Halanaerobiales		0.00		0.00	0	0.00	0	0.00	0	0.00	953	0.01
		Mollicutes	Acholeplasmatales		0.00		0.00	2,020	0.02	5,158	0.04	15,933	0.16	11,032	0.13
			Entomoplasmatales		0.00		0.00	8,820	0.09	2,153	0.02	23,798	0.24	10,316	0.12
			Anaeroplasmatales		0.00		0.00	3,620	0.04	2,456	0.02	1,762	0.02	2,616	0.03
			Mycoplasmatales		0.00		0.00	17,627	0.18	32,047	0.27	35,840	0.37	21,683	0.26
	Actinobacteria	Actinobacteria	Actinomycetales	332	0.22	793	0.58	474,820	4.75	297,985	2.55	345,799	3.52	250,739	2.95
			Bifidobacteriales	31,443	21.27	5,101	3.70	2,183,231	21.84	615,749	5.27	699,336	7.12	235,529	2.77
			Coriobacteriales	25,781	17.44	10,804	7.84	0	0.00	0	0.00	0	0.00	0	0.00
	Proteobacteria	Alpha proteobacteria	Caulobacterales		0.00		0.00	10,349	0.10	12,194	0.10	9,752	0.10	13,102	0.15
			Rhizobiales		0.00		0.00	115,127	1.15	151,347	1.30	139,897	1.42	122,856	1.45
			Rhodobacterales		0.00		0.00	978	0.01	1,512	0.01	3,402	0.03	0	0.00
			Rhodospirillales		0.00		0.00	924	0.01	2,074	0.02	937	0.01	2,472	0.03
			Rickettsiales		0.00		0.00	3,750	0.04	9,356	0.08	10,339	0.11	5,982	0.07
			Sphingomonadales		0.00		0.00	10,815	0.11	0	0.00	4,244	0.04	2,676	0.03
		Beta proteobacteria	Burkholderiales		0.00		0.00	73,843	0.74	38,134	0.33	65,670	0.67	44,742	0.53
			Methylophilales		0.00		0.00	0	0.00	0	0.00	850	0.01	0	0.00
			Neisseriales		0.00		0.00	26,515	0.27	34,357	0.29	42,403	0.43	31,306	0.37
			Nitrosomonadales		0.00		0.00	10,099	0.10	11,362	0.10	18,515	0.19	10,067	0.12
			Rhodocyclales		0.00		0.00	0	0.00	0	0.00	1,458	0.01	3,600	0.04
		Delta proteobacteria	Bdellovibrionales		0.00		0.00	9,301	0.09	20,224	0.17	18,032	0.18	8,128	0.10
			Desulfobacterales		0.00		0.00	51,124	0.51	45,737	0.39	58,507	0.60	50,195	0.59
			Desulfovibrionales		0.00		0.00	42,501	0.43	67,324	0.58	71,580	0.73	51,430	0.61
			Desulfuromonadales		0.00		0.00	119,409	1.19	116,716	1.00	144,283	1.47	97,783	1.15
			Myxococcales		0.00		0.00	4,088	0.04	7,018	0.06	9,933	0.10	2,725	0.03
		Epsilon proteobacteria	Campylobacterales		0.00		0.00	64,310	0.64	71,518	0.61	62,378	0.64	76,007	0.89
		Gamma proteobacteria	Acidithiobacillales		0.00		0.00	0	0.00	0	0.00	2,585	0.03	0	0.00
			Alteromonadales		0.00		0.00	15,087	0.15	17,281	0.15	15,703	0.16	24,933	0.29
			Aeromonadales		0.00		0.00	2,496	0.02	0	0.00	2,507	0.03	914	0.01
			Cardiobacteriales		0.00		0.00	3,208	0.03	2,379	0.02	1,782	0.02	1,736	0.02
			Chromatiales		0.00		0.00	0	0.00	0	0.00	578	0.01	887	0.01
			Enterobacteriales		0.00		0.00	140,297	1.40	176,066	1.51	191,099	1.95	137,205	1.62
			Legionellales		0.00		0.00	17,326	0.17	16,622	0.14	19,392	0.20	15,601	0.18
			Methylococcales		0.00		0.00	13,954	0.14	15,227	0.13	27,182	0.28	15,253	0.18
			Pasteurellales		0.00		0.00	61,969	0.62	90,921	0.78	89,367	0.91	64,313	0.76
			Pseudomonadales		0.00		0.00	47,911	0.48	43,324	0.37	36,858	0.38	41,051	0.48
			Thiotrichales		0.00		0.00	0	0.00	0	0.00	861	0.01	0	0.00
			Vibrionales		0.00		0.00	52,010	0.52	72,984	0.62	76,030	0.77	79,910	0.94
			Xanthomonadales		0.00		0.00	17,469	0.17	30,631	0.26	22,488	0.23	26,694	0.31
	Fusobacteria	Fusobacteria	Fusobacterales		0.00		0.00	173,171	1.73	236,794	2.03	190,353	1.94	205,932	2.42
	Fibrobacteres /Acidobacteria	Fibrobacteres	Fibrobacterales		0.00		0.00	0	0.00	0	0.00	0	0.00	869	0.01
	Bacteroidetes	Bacteroides	Bacteroidales		0.00		0.00	273,673	2.74	395,831	3.39	332,370	3.38	333,432	3.93
		Flavobacteria	Flavobacteriales		0.00		0.00	2,899	0.03	4,000	0.03	2,482	0.03	2,799	0.03
		Sphingobacteria	Sphingobacteriales		0.00		0.00	0	0.00	0	0.00	1,753	0.02	777	0.01
	Spirochaetes	Spirochaetes	Spirochaetales		0.00		0.00	218,073	2.18	251,060	2.15	272,717	2.78	216,296	2.55
	Cyanobacteria	Chroococcales			0.00		0.00	47,969	0.48	45,520	0.39	47,036	0.48	50,965	0.60
		Nostocales			0.00		0.00	29,726	0.30	37,954	0.32	26,804	0.27	40,408	0.48
		Prochlorales			0.00		0.00	3,820	0.04	5,988	0.05	13,634	0.14	10,048	0.12
		Stigonematales			0.00		0.00	0	0.00	0	0.00	0	0.00	0	0.00
		Oscillatoriales			0.00		0.00	0	0.00	2,552	0.02	821	0.01	1,810	0.02
		Gloeobacteria	Gloeobacterales		0.00		0.00	27,658	0.28	19,582	0.17	12,269	0.12	17,167	0.20
	Chlamydiae /Verrucomicrobia	Chlamydiae	Chlamydiales		0.00		0.00	24,682	0.25	15,658	0.13	15,610	0.16	9,697	0.11
	Dictyoglomi	Dictyoglomi	Dictyoglomales		0.00		0.00	0	0.00	0	0.00	0	0.00	0	0.00
	Deinococci	Deinococci	Deinococcales		0.00		0.00	22,803	0.23	32,526	0.28	18,393	0.19	11,179	0.13
			Thermales		0.00		0.00	15,105	0.15	22,305	0.19	34,622	0.35	20,258	0.24
	Chloroflexi	Chloroflexi	Chloroflexales		0.00		0.00	754	0.01	0	0.00	859	0.01	796	0.01
	Thermotogae	Thermotogae	Thermotogales		0.00		0.00	39,088	0.39	68,798	0.59	75,729	0.77	86,961	1.02
	Planctomycetes	Planctomycetacia	Planctomycetales		0.00		0.00	23,524	0.24	39,945	0.34	30,686	0.31	21,728	0.26
	Chlorobi	Chlorobia	Chlorobiales		0.00		0.00	33,531	0.34	28,895	0.25	28,792	0.29	30,862	0.36
	Thermomicrobia	Thermomicrobia	Thermomicrobiales		0.00		0.00	0	0.00	0	0.00	859	0.01	0	0.00
	Aquificae	Aquificae	Aquificales		0.00		0.00	16,124	0.16	26,172	0.22	24,646	0.25	14,535	0.17
ARCHAEA	Crenarchaeota	Thermoprotei	Desulfurococcales		0.00		0.00	0	0.00	1,922	0.02	936	0.01	3,451	0.04
			Cenarchaeales		0.00		0.00	0	0.00	1,487	0.01	0	0.00	0	0.00
			Sulfolobales		0.00		0.00	9,190	0.09	8,760	0.07	6,654	0.07	3,336	0.04
			Thermoproteales		0.00		0.00	0	0.00	2,424	0.02	1,626	0.02	1,628	0.02
	Euryarchaeota	Archaeoglobi	Archaeoglobales		0.00		0.00	15,239	0.15	24,045	0.21	18,983	0.19	27,526	0.32
		Halobacteria	Halobacteriales		0.00		0.00	2,616	0.03	995	0.01	6,000	0.06	3,291	0.04
		Methanomicrobia	Methanosarcinales		0.00		0.00	78,135	0.78	134,022	1.15	109,236	1.11	111,141	1.31
		Methanobacteria	Methanobacteriales	18,188	12.30	17,970	13.05	906,553	9.07	916,899	7.85	36,703	0.37	29,430	0.35
		Methanococci	Methanococcales		0.00		0.00	44,413	0.44	71,172	0.61	58,126	0.59	61,272	0.72
		Methanopyri	Methanopyrales		0.00		0.00	8,053	0.08	6,504	0.06	6,486	0.07	12,837	0.15
		Thermococci	Thermococcales		0.00		0.00	24,248	0.24	27,264	0.23	34,841	0.35	28,160	0.33
		Thermoplasmata	Thermoplasmatales		0.00		0.00	6,910	0.07	5,060	0.04	6,810	0.07	9,932	0.12
			Totals:	147,853	100.00	137,738	100.00	9,996,250	100.00	11,686,463	100.00	9,819,177	100.00	8,493,125	100.00

Table S2. Distribution of genes among KEGG pathways and COGs.

	Sample 7	Sample 8	Samples 7,8
Total number of predicted genes	25077	25087	50164
Genes assigned to COGs	7563	8097	15660
Enriched COGs (*P*<0.05)[1]	733 (5544)	718 (5986)	788 (11847)
Genes assigned to KEGG pathways	4468	4157	8625
Enriched KEGG pathways vs. all bacteria (*P*<0.05)[2]	26(2054)	23(1748)	28 (4194)
Enriched KEGG pathways vs. all archaea (*P*<0.05)[2]	32(1845)	30(1784)	37 (4257)
Enriched KEGG pathways vs. H.sapiens (*P*<0.05)[2]	41 (2626)	39 (2462)	45 (5351)

[1]Number of COGs (number of genes assigned to COGs) enriched in the human distal gut microbiome compared to all microbial genomes in the STRING extended database (163).

[2]Number of KEGG pathways (number of genes assigned to KEGG pathways) enriched in human distal gut microbiome compared to the respective dataset.

Table S3. KEGG pathways enriched or under-represented in the human gut microbiome relative to all bacterial genomes in KEGG, the human genome, and all archaeal genomes in KEGG.[+]

KEGG Category	Pathway	Hits	Bacteria	H.sapiens	Archaea
Carbohydrate Metabolism	Nucleotide sugars metabolism	97	1.02	12.52	0.71
	Galactose metabolism	271	1.29	11.05	3.58
	Aminosugars metabolism	245	1.27	9.97	8.96
	Starch and sucrose metabolism	485	1.81	8.90	3.84
	Pentose phosphate pathway	260	1.11	8.06	2.02
	Fructose and mannose metabolism	235	0.98	6.37	2.08
	Propanoate metabolism	204	0.90	6.30	0.67
	Glycolysis / Gluconeogenesis	429	1.01	5.87	1.48
	Pyruvate metabolism	283	0.87	5.76	0.80
	Butanoate metabolism	152	0.59	5.46	0.53
	Glyoxylate and dicarboxylate metabolism	61	0.41	4.36	0.55
	Inositol metabolism	32	0.69	4.11	1.02
	C5-Branched dibasic acid metabolism	11	0.18	3.53	0.15
	Ascorbate and aldarate metabolism	20	0.56	3.21	1.19
	Pentose and glucuronate interconversions	91	0.95	2.66	2.50
	Citrate cycle (TCA cycle)	72	0.37	2.57	0.27
	Inositol phosphate metabolism	6	0.26	0.16	0.22
Energy Metabolism	Carbon fixation	197	0.96	7.10	1.06

	ATP synthesis	73	0.60	5.88	0.52
	Nitrogen metabolism	155	0.73	4.36	0.75
	Sulfur metabolism	35	0.35	3.75	0.62
	Methane metabolism	40	0.56	3.67	0.36
	Reductive carboxylate cycle (CO2 fixation)	42	0.24	2.70	0.12
	ATPases	70	2.06	2.65	2.67
Lipid Metabolism	Biosynthesis of steroids	164	1.65	9.66	2.31
	Glycerolipid metabolism	133	1.03	2.26	1.59
	Fatty acid metabolism	46	0.27	0.64	0.20
	Prostaglandin and leukotriene metabolism	2	0.15	0.04	0.30
	Androgen and estrogen metabolism	1	0.22	0.02	0.46
Nucleotide Metabolism	Pyrimidine metabolism	799	1.34	6.98	1.09
	Purine metabolism	762	1.23	4.28	1.08
Amino Acid Metabolism	Lysine biosynthesis	220	1.56	143.21	1.67
	Phenylalanine, tyrosine and tryptophan biosynthesis	335	1.13	21.96	0.79
	Valine, leucine and isoleucine biosynthesis	262	1.41	21.37	1.04
	Alanine and aspartate metabolism	389	1.77	17.05	1.36
	Glutamate metabolism	356	1.49	11.12	1.11
	Histidine metabolism	242	1.85	8.75	1.68
	Methionine metabolism	195	1.94	8.44	2.01
	Urea cycle and metabolism of amino groups	274	1.70	7.77	1.59
	Glycine, serine and threonine metabolism	286	0.89	7.17	0.86
	Cysteine metabolism	68	0.53	4.38	0.65
	Arginine and proline metabolism	239	1.01	4.08	0.82
	Phenylalanine metabolism	46	0.51	2.96	0.55
	Lysine degradation	79	0.60	2.42	0.85
	Tyrosine metabolism	49	0.43	1.57	0.33
	Valine, leucine and isoleucine degradation	46	0.24	1.41	0.22
	Tryptophan metabolism	43	0.29	0.63	0.40
Non-peptidal Amino Acid Metabolism	D-Glutamine and D-glutamate metabolism	91	2.03	14.68	3.91
	Cyanoamino acid metabolism	85	1.25	6.85	1.32
	Selenoamino acid metabolism	149	1.11	5.35	1.31
	beta-Alanine metabolism	77	0.69	3.54	0.65
	Taurine and hypotaurine metabolism	44	0.90	3.54	2.74
	Glutathione metabolism	22	0.23	0.38	0.75
Glycan Biosynthesis and Metabolism	Peptidoglycan biosynthesis	239	1.83	77.88	7.06
	Glycosaminoglycan degradation	58	2.87	5.33	4.19
	N-Glycan degradation	58	2.34	2.49	3.18
	Glycosphingolipid metabolism	78	3.88	1.93	4.66
	Globoside metabolism	8	0.52	0.51	10.96
	N-Glycan biosynthesis	2	0.56	0.10	0.11
Polyketides/	Biosynthesis of ansamycins	49	2.27	10.51	2.59

Nonribosomal Peptides					
	Polyketide sugar unit biosynthesis	11	0.47	7.06	0.18
Metabolism of Cofactors and Vitamins	Thiamine metabolism	39	1.15	12.54	0.67
	One carbon pool by folate	113	1.08	10.43	1.69
	Pantothenate and CoA biosynthesis	119	0.75	8.54	1.04
	Vitamin B6 metabolism	39	0.89	6.27	1.03
	Biotin metabolism	9	0.20	5.78	0.29
	Folate biosynthesis*	74	0.85	5.30	0.34
	Riboflavin metabolism	42	0.83	3.37	0.76
	Nicotinate and nicotinamide metabolism	91	0.88	3.26	0.94
	Porphyrin and chlorophyll metabolism	117	0.61	2.36	0.46
	Ubiquinone biosynthesis	1	0.01	0.08	0.01
Biosynthesis of secondary metabolites	Novobiocin biosynthesis	46	0.83	29.59	0.54
	Stilbene, coumarine and lignin biosynthesis	35	2.02	4.50	6.86
	Limonene and pinene degradation	22	0.38	3.53	0.46
	Alkaloid biosynthesis II	16	1.17	3.42	2.19
	Alkaloid biosynthesis I	9	0.27	2.89	0.15
	Streptomycin biosynthesis	29	0.39	2.33	0.26
Xenobiotic Metabolism	Benzoate degradation via CoA ligation	43	0.34	5.53	0.38
	Tetrachloroethene degradation	9	1.29	2.89	0.40
	1,2-Dichloroethane degradation	17	0.63	2.73	0.93
	Caprolactam degradation	7	0.15	2.25	0.15
	gamma-Hexachlorocyclohexane degradation	17	0.50	0.39	0.37
	1- and 2-Methylnaphthalene degradation	2	0.08	0.21	0.05

* - contains methanogenesis pathway.

[+]Odds ratios are shown for the human distal gut microbiome against all bacterial genomes in KEGG, the human genome, and all archaeal genomes in KEGG. All pathways shown are enriched (odds ratio > 1) or under-represented (odds ratio < 1) when compared to the human genome ($P < 0.05$). NA indicates that a ratio could not be calculated, because there are no hits to the pathway in the comparison dataset. The odds ratios are a measure of relative gene content based on the number of independent hits to enzymes in each pathway; future studies will be necessary to experimentally validate these metabolic predictions. See http://gordonlab.wustl.edu/supplemental/Gill/ for each of these maps

Table S4. Human distal gut microbiome shows enrichment (relative to all microbial genomes in STRING) for glycan degradation COGs

Category	COG	Annotation	Total hits in microbiome	Odds Ratio ($p<0.05$)
General	COG1070	Sugar (pentulose and hexulose) kinases	30	4.25
	COG0061	Predicted sugar kinase	22	3.26
	COG0366	Glycosidases	49	2.97
	COG1082	Sugar phosphate isomerases/epimerases	12	1.56
Arabinose	COG3534	Alpha-L-arabinofuranosidase	11	12.18
	COG2160	L-arabinose isomerase	7	9.44
Fructose	COG1621	Beta-fructosidases (levanase/invertase)	15	6.74
	COG0205	6-phosphofructokinase	31	4.90
	COG1105	Fructose-1-phosphate kinase (PfkB)	8	2.05
Fucose	COG2407	L-fucose isomerase and related proteins	23	20.98
	COG4154	Fucose dissimilation pathway protein FucU	4	4.28
	COG3669	Alpha-L-fucosidase	5	3.04
Galactose	COG4468	Galactose-1-phosphate uridyltransferase	19	32.73
	COG1486	Alpha-galactosidases/6-phospho-beta-glucosidases	23	8.70
	COG3250	Beta-galactosidase/beta-glucuronidase	32	5.42
	COG3345	Alpha-galactosidase	4	4.00
	COG0153	Galactokinase	11	3.63
	COG2723	Beta-glucosidase/6-phospho-beta-glucosidase/beta-galactosidase	34	3.47
Glucuronose	COG3661	Alpha-glucuronidase	4	17.72
	COG1904	Glucuronate isomerase	10	7.95
	COG3250	Beta-galactosidase/beta-glucuronidase	32	5.42
Glucosamine	COG1820	N-acetylglucosamine-6-phosphate deacetylase	11	2.60
	COG0363	6-phosphogluconolactonase/Glucosamine-6-phosphate isomerase/deaminase	16	2.28
Glucose	COG3405	Endoglucanase Y	7	11.43
	COG3459	Cellobiose phosphorylase	7	8.68
	COG0297	Glycogen synthase	24	7.09
	COG0296	1,4-alpha-glucan branching enzyme	27	5.70
	COG1523	Pullulanase PulA and related glycosidases	20	5.13
	COG2723	Beta-glucosidase/6-phospho-beta-glucosidase/beta-galactosidase	34	3.47
	COG0166	Glucose-6-phosphate isomerase	18	3.05
	COG1501	Alpha-glucosidases, family 31 of glycosyl hydrolases	17	2.66
	COG1472	Beta-glucosidase-related glycosidases	18	2.65
Mannose	COG1312	D-mannonate dehydratase	18	12.69
	COG0246	Mannitol-1-phosphate/altronate dehydrogenases	17	4.47
	COG1482	Phosphomannose isomerase	9	2.49
	COG1109	Phosphomannomutase	23	1.72

Category	COG	Annotation	Total hits in microbiome	Odds Ratio ($p<0.05$)
Rhamnose	COG4806	L-rhamnose isomerase	5	7.05
Xylose	COG3507	Beta-xylosidase	12	8.09
Transferase	COG1640	4-alpha-glucanotransferase	9	3.36
Transport	COG4209	ABC-type polysaccharide transport system, permease component	26	25.20
	COG0395	ABC-type sugar transport system, permease component	56	3.63
	COG2190	Phosphotransferase system IIA components	21	3.48
	COG1263	Phosphotransferase system IIC components, glucose/maltose/GlcNAc-specific	30	2.90
	COG1129	ABC-type sugar transport system, ATPase component	31	2.85
	COG1264	Phosphotransferase system IIB components	30	2.83
	COG1175	ABC-type sugar transport systems, permease components	39	2.36
	COG1299	Phosphotransferase system, fructose-specific IIC component	11	2.03
	COG1593	TRAP-type C4-dicarboxylate transport system, large permease component	12	2.01
	COG1638	TRAP-type C4-dicarboxylate transport system, periplasmic component	10	1.82
	COG1080	Phosphoenolpyruvate-protein kinase (PTS system EI component in bacteria)	8	1.75
	COG1445	Phosphotransferase system fructose-specific component IIB	9	1.74
	COG1653	ABC-type sugar transport system, periplasmic component	27	1.65
	COG1879	ABC-type sugar transport system, periplasmic component	19	1.58

Table S5. KEGG analysis shows presence of at least 81 glycoside hydrolase families in the distal gut microbiome. Only families that are not present in the human genome are shown.

Glycoside Hydrolase Family	Family members	Hits in the distal gut microbiome
3	beta-glucosidase xylan 1,4-beta-xylosidase	108
4	Maltose-6-phosphate glucosidase alpha-glucosidase	62
5	Chitosanase beta-mannosidase Cellulase Glucan	42
7	endoglucanase cellobiohydrolase The cellobiohydrolases	35
8	Chitosanase Cellulase Licheninase Endo-1,4-beta-xylanase	6
6	endoglucanase cellobiohydrolase The cellobiohydrolases	2
57	alpha-amylase 4-alpha-glucanotransferase	67
43	beta-xylosidase alpha-L-arabinofuranosidase	65
51	alpha-L-arabinofuranosidase endoglucanase	61
32	invertase inulinase levanase exo-inulinase	56
42	beta-galactosidase	54
10	xylanase endo-1,3-beta-xylanase cellobiohydrolase	45
77	amylomaltase or 4-alpha-glucanotransferase	45
73	endo-beta-N-acetylglucosaminidase beta-1,4-N-acetylmuramoylhydrolase	40
16	Xyloglucan:xyloglucosyl transferase Keratan-sulfate	38
12	endoglucanase xyloglucan hydrolase beta-1,3-1,4-glucanase	34
26	mannanase beta-1,3-xylanase	34
28	polygalacturonase exo-polygalacturonase	34

64	beta-1,3-glucanase	33
82	i-carrageenase	33
88	Deltalpha-4,5 unsaturated glucuronyl hydrolases	33
90	Endorhamnosidases	33
92	alpha-1,2-mannosidase	33
93	exo-arabinanase	33
96	alpha-agarase	33
98	endo-beta-galactosidase	33
102	peptidoglycan lytic transglycosylase	33
103	peptidoglycan lytic transglycosylase	33
104	peptidoglycan lytic transglycosylase	33
105	unsaturated rhamnogalacturonyl hydrolase	33
54	alpha-L-arabinofuranosidase beta-xylosidase	27
52	beta-xylosidase	21
70	Dextransucrase Alternansucrase	18
94	cellobiose phosphorylase cellodextrin phosphorylase	18
66	cycloisomaltooligosaccharide glucanotransferase	16
17	Glucan endo-1,3-beta-D-glucosidase Glucan	15
72	beta-1,3-glucanosyltransglycosylase	15
14	beta-amylase	11
62	alpha-L-arabinofuranosidase	11
68	levansucrase beta-fructofuranosidase	9
36	alpha-galactosidase alpha-N-acetylgalactosaminidase	8
100	alkaline and neutral invertase	8
11	xylanase	5
78	alpha-L-rhamnosidase	5
106	alpha-L-rhamnosidase	5
67	alpha-glucuronidase	3
97	alpha-glucosidase	3
19	chitinase	2
25	lysozyme	2
48	endoglucanase cellobiohydrolase	2
15	glucoamylase glucodextranase	1
34	sialidase or neuraminidase	1
44	endoglucanase	1
45	endoglucanase	1
49	Dextranase Isopullulanase Dextran 1,6-alpha-isomaltotriosidase	1
61	endoglucanase	1
74	endoglucanase oligoxyloglucan reducing end-specific	1
83	hemagglutinin-neuraminidase	1
95	alpha-L-fucosidase	1

Table S6. Human distal gut microbiome shows enrichment (relative to all microbial genomes in STRING) for genes involved in fermentation of carbohydrates.

End Product	COG	Annotation	Total hits in microbiome	Odds Ratio ($p<0.05$)
Butyrate	COG3426	Butyrate kinase	6	9.30
Acetate	COG0282	Acetate kinase	16	3.31
Lactate	COG2055	Malate/L-lactate dehydrogenases	8	2.85
	COG0039	Malate/lactate dehydrogenases	26	2.70
	COG1052	Lactate dehydrogenase and related dehydrogenases	17	2.14
CO2/Acetyl-CoA	COG1014	Pyruvate:ferredoxin oxidoreductase, gamma subunit	13	2.08
	COG1013	Pyruvate:ferredoxin oxidoreductase, beta subunit	10	2.05
	COG0674	Pyruvate:ferredoxin oxidoreductase, alpha subunit	9	1.78
Succinate	COG1053	Succinate dehydrogenase/fumarate reductase, flavoprotein subunit	15	1.60

Table S7. Human distal gut microbiome shows enrichment (relative to all microbial genomes in STRING) for vitamin biosynthetic COGs.

Vitamin	COG	Annotation	Total hits in microbiome	Odds Ratio ($p<0.05$)
Folate	COG0720	6-pyruvoyl-tetrahydropterin synthase	8	1.87
	COG0190	5,10-methylene-tetrahydrofolate dehydrogenase/Methenyl tetrahydrofolate cyclohydrolase	12	1.71
Isoprenoid	COG0743	1-deoxy-D-xylulose 5-phosphate reductoisomerase, dxr	37	10.09
	COG1154	Deoxyxylulose-5-phosphate synthase, dxs	35	8.63
	COG0821	Enzyme involved in the deoxyxylulose pathway of isoprenoid biosynthesis, isG	29	7.70
	COG1947	4-diphosphocytidyl-2C-methyl-D-erythritol 2-phosphate synthase, ispE	20	4.89
	COG0761	Penicillin tolerance protein, ispH	17	4.40
	COG1211	4-diphosphocytidyl-2-methyl-D-erithritol synthase, ispD	16	3.47
	COG0245	2C-methyl-D-erythritol 2,4-cyclodiphosphate synthase, ispF	12	3.16
Thiamine	COG0301	Thiamine biosynthesis ATP pyrophosphatase	29	9.78
	COG0422	Thiamine biosynthesis protein, ThiC	29	8.25
	COG2022	Thiamine biosynthesis protein, ThiG	14	4.94
	COG0352	Thiamine monophosphate synthase, ThiE	24	4.43
	COG1060	Thiamine biosynthesis enzyme, ThiH	17	4.22
Vitamin B6	COG0214	Pyridoxine biosynthesis enzyme	15	7.05
	COG0311	Predicted glutamine amidotransferase involved in pyridoxine biosynthesis	5	2.63
	COG2240	Pyridoxal/pyridoxine/pyridoxamine kinase	6	2.09

Vitamin B12	COG1903	Cobalamin biosynthesis protein CbiD	15	9.69
	COG1492	Cobyric acid synthase	12	4.71
	COG1797	Cobyrinic acid a,c-diamide synthase	10	4.49
	COG1010	Precorrin-3B methylase	7	3.39
	COG2099	Precorrin-6x reductase	4	2.88

Table S8. Human distal gut microbiome shows enrichment (relative to all microbial genomes in STRING) for amino acid biosynthetic COGs.

Amino Acid Metabolism	COG	Annotation	Total hits in microbiome	Odds Ratio ($p<0.05$)
Alanine, aspartate and asparagines	COG2502	Asparagine synthetase A	24	19.08
	COG0367	Asparagine synthase (glutamine-hydrolyzing)	40	6.97
	COG1027	Aspartate ammonia-lyase	6	2.24
	COG0520	Selenocysteine lyase	14	1.70
Arginine	COG4187	Arginine degradation protein	5	15.51
	COG4992	Ornithine/acetylornithine aminotransferase*	32	4.59
	COG0137	Argininosuccinate synthase*	19	3.95
	COG0078	Ornithine carbamoyltransferase*	20	3.45
	COG0165	Argininosuccinate lyase*	16	3.45
	COG3869	Arginine kinase	6	2.27
Beta-Alanine	COG0421	Spermidine synthase**	14	2.65
Glutamate	COG1364	N-acetylglutamate synthase (N-acetylornithine aminotransferase)	29	10.84
	COG0014	Gamma-glutamyl phosphate reductase	30	7.38
	COG0458	Carbamoylphosphate synthase large subunit	40	5.91
	COG0263	Glutamate 5-kinase	24	5.81
	COG0548	Acetylglutamate kinase	27	4.98
	COG0070	Glutamate synthase domain 3	13	3.45
	COG0067	Glutamate synthase domain 1	13	3.20
	COG0549	Carbamate kinase	8	3.10
	COG0505	Carbamoylphosphate synthase small subunit	18	2.97
	COG0069	Glutamate synthase domain 2	13	2.52
	COG0002	Acetylglutamate semialdehyde dehydrogenase	9	2.31
Glycine, serine and threonine	COG1897	Homoserine trans-succinylase	18	11.39
	COG3048	D-serine dehydratase	3	4.23
	COG2008	Threonine aldolase	10	3.16
	COG0112	Glycine/serine hydroxymethyltransferase	22	3.12
	COG0498	Threonine synthase	17	2.85
	COG1063	Threonine dehydrogenase and related Zn-dependent dehydrogenases	20	1.52
Histidine	COG3705	ATP phosphoribosyltransferase (histidine biosynthesis)	14	9.24

	COG0141	Histidinol dehydrogenase	35	8.90
	COG0131	Imidazoleglycerol-phosphate dehydratase	31	8.15
	COG0040	ATP phosphoribosyltransferase	22	5.78
	COG0107	Imidazoleglycerol-phosphate synthase	19	4.68
	COG0079	Histidinol-phosphate/aromatic aminotransferase/cobyric acid decarboxylase	27	3.43
Leucine	COG2309	Leucyl aminopeptidase (aminopeptidase T)	5	2.63

Amino Acid Metabolism	**COG**	**Annotation**	**Total hits in microbiome**	**Odds Ratio ($p<0.05$)**
Lysine	COG0289	Dihydrodipicolinate reductase	12	2.84
	COG0019	Diaminopimelate decarboxylase	19	2.28
	COG0527	Aspartokinases	17	2.21
	COG0329	Dihydrodipicolinate synthase/N-acetylneuraminate lyase	21	1.94
	COG0136	Aspartate-semialdehyde dehydrogenase	10	1.69
Methionine	COG0620	Methionine synthase II (cobalamin-independent)	21	4.23
Phenylalanine, tyrosine and tryptophan* **	COG1685	Archaeal shikimate kinase	4	7.75
	COG0128	5-enolpyruvylshikimate-3-phosphate synthase	32	6.05
	COG0710	3-dehydroquinate dehydratase	9	4.43
	COG0169	Shikimate 5-dehydrogenase	28	4.22
	COG0337	3-dehydroquinate synthetase	19	3.90
	COG0082	Chorismate synthase	19	3.88
	COG0722	3-deoxy-D-arabino-heptulosonate 7-phosphate (DAHP) synthase	14	3.34
	COG2876	3-deoxy-D-arabino-heptulosonate 7-phosphate (DAHP) synthase	6	2.91
	COG0133	Tryptophan synthase beta chain	12	2.84
	COG0757	3-dehydroquinate dehydratase II	8	2.67
	COG0547	Anthranilate phosphoribosyltransferase	11	2.16
	COG0135	Phosphoribosylanthranilate isomerase	8	1.98
	COG0512	Anthranilate/para-aminobenzoate synthases component II	11	1.66
Valine, leucine and isoleucine	COG0065	3-isopropylmalate dehydratase large subunit	44	9.48
	COG0129	Dihydroxyacid dehydratase/phosphogluconate dehydratase	29	4.37
	COG0066	3-isopropylmalate dehydratase small subunit	18	3.77
	COG0059	Ketol-acid reductoisomerase	13	3.08
	COG0119	Isopropylmalate/homocitrate/citramalate synthases	16	1.72
Peptidases	COG1362	Aspartyl aminopeptidase	25	16.50
	COG2195	Di- and tripeptidases	18	4.20
Transferases	COG0118	Glutamine amidotransferase	17	4.18
	COG0115	Branched-chain amino acid aminotransferase/4-amino-4-deoxychorismate lyase	27	2.51
	COG0436	Aspartate/tyrosine/aromatic aminotransferase	25	1.30

Transporters	COG1687	Predicted branched-chain amino acid permeases (azaleucine resistance)	4	7.30
	COG0747	ABC-type dipeptide transport system, periplasmic component	55	2.97
	COG1126	ABC-type polar amino acid transport system, ATPase component	32	2.46
	COG0444	ABC-type dipeptide/oligopeptide/nickel transport system, ATPase component	28	2.31
	COG4608	ABC-type oligopeptide transport system, ATPase component	22	2.13
	COG0765	ABC-type amino acid transport system, permease component	23	1.28

* - enzyme in the urea cycle
** - also involved in synthesis of biogenic amines
*** - includes shikamate pathway

Supporting References

S1. G. C. Baker, J. J. Smith, D. A. Cowan, *J Microbiol Methods* **55**, 541 (Dec, 2003).

S2. S. Bartosch, A. Fite, G. T. Macfarlane, M. E. McMurdo, *Appl Environ Microbiol* **70**, 3575 (Jun, 2004).

S3. J. C. Venter *et al.*, *Science* **304**, 66 (Apr 2, 2004).

S4. E. W. Myers *et al.*, *Science* **287**, 2196 (Mar 24, 2000).

S5. S. Kurtz *et al.*, *Genome Biol* **5**, R12 (2004).

S6. A. L. Delcher, D. Harmon, S. Kasif, O. White, S. L. Salzberg, *Nucleic Acids Res* **27**, 4636 (Dec 1, 1999).

S7. L. B. Koski, G. B. Golding, *J Mol Evol* **52**, 540 (Jun, 2001).

S8. B. L. Maidak *et al.*, *Nucleic Acids Res* **29**, 173 (Jan 1, 2001).

S9. W. Ludwig *et al.*, *Nucleic Acids Res* **32**, 1363 (2004).

S10. P. D. Schloss, J. Handelsman, *Appl Environ Microbiol* **71**, 1501 (Mar, 2005).

S11. M. Riley, *Microbiol Rev* **57**, 862 (Dec, 1993).

S12. M. Pop, A. Phillippy, A. L. Delcher, S. L. Salzberg, *Brief Bioinform* **5**, 237 (Sep, 2004).

S13. M. A. Schell *et al.*, *Proc Natl Acad Sci U S A* **99**, 14422 (Oct 29, 2002).

S14. J. Xu *et al.*, *Science* **299**, 2074 (Mar 28, 2003).

S15. http://gordonlab.wustl.edu/supplemental/Gill/Msmithii/draftgenome/

S16. C. von Mering *et al.*, *Nucleic Acids Res* **33**, D433 (Jan 1, 2005).