SUPPLEMENTARY INFORMATION

doi:10.1038/nature11209

Table of Contents

Supplementary Notes and Methods

This supplementary information follows the structure and layout of the main paper. Here we describe the methodology in detail, include additional analysis not covered in depth in the main text and summarize available Human Microbiome Project Consortium (HMP) resources.

Availability of HMP data and resources

All unprocessed HMP sequence data is available from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) (Supplementary Table 1) and NCBI Bioproject page (Supplementary Table 2) with accompanying metadata provided at the database of Genotypes and Phenotypes (dbGaP) Study Accession phs000228 (Supplementary Table 1). Authorized access to unfiltered data containing human sequence produced from this study can be requested via the authorized access system at dbGaP (https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?adddataset=phs000228&page=login) (Study= phs000228, Consent Group= HMP).

An overview of additional HMP resources is available in Supplementary Table 2 and an overview of processed phase I data available from the Data Analysis and Coordination Center (DACC) is listed in Supplementary Table 3. All data sets available at the DACC referred to in the main text and in this Supplementary Information document are denoted with the preface 'RES' followed by a short identifier representing the data set in question. Alternatively, these data sets are also available at the DACC using the general form 'http://hmpdacc.org/<XXXX>' where 'XXXX' denotes a specific data set identifier.

Metadata for all HMP 16S rRNA gene sequencing (16S) and Whole Genome Shotgun (WGS) sequencing is described using the Minimum Information about a Marker Sequence (MIMARKS) and Minimum Information about a Metagenomic Sequence (MIMS) [18] standards, respectively. Metadata for all HMP reference genomes are described using the 'Minimum Information about a Genome Sequence (MIGS) [19].

Donor recruitment and sampling

Adult subjects between the ages of 18 and 40 years who were defined as "healthy" based on a lengthy list of oral, cutaneous, vaginal, intestinal, and other exclusion criteria were recruited and sampled one to three times at 15 (male) or 18 (female) as detailed in [20]. Enrollments were approved by the Institutional Review Boards of the two recruitment centers (Baylor College of Medicine, Houston, TX and Washington University School of Medicine, St. Louis, MO), and a common sampling protocol (see http://hmpdacc.org/doc/HMP_Clinical_Protocol.pdf was employed for nine oral specimen types (saliva, swabs from the buccal mucosa, tongue, keratinized gingiva, hard palate, tonsils, and throat, and sub- and supragingival plaque scraping); three vaginal specimen types (swabs from the vaginal introitus, posterior fornix, and vaginal midpoint); four skin specimen types (bilateral retroauricular crease and anterior fossa swabs); the respiratory tract (both anterior nares, swabbed and pooled); and stool (self collected by

commode kit). Subject recruitment and enrollment was supervised by licensed physicians and health care providers including subspecialists for oral and vaginal sampling. Medical histories and physical examinations were conducted at Baylor College of Medicine and Washington University School of Medicine, and this information was extracted for metadata collection purposes. Subject phenotypic metadata was collected and coded by the EMMES Corporation. Genomic DNA from all samples was isolated using the Mo Bio PowerSoil DNA Isolation Kit (http://www.mobio.com) as detailed here http://hmpdacc.org/doc/sops_2/manual_of_procedures_v11.pdf.

A summary of the distribution of HMP donor samples by various donor metadata categories and the subsets of samples that were assayed by 16S and WGS sequencing are given in Supplementary Table 4. Metrics related to sequencing and data processing of the subset are also summarized in Table 1.

Description of synthetic mock communities

The organisms for the mock community (MC) included a variety of different genera commonly found on or within the human body (Supplementary Table 5). Bacteria and the yeast, *Candida albicans*, were cultivated on agar plates under appropriate growth conditions (temperature, aerobiosis or anaerobiosis), generally as recommended by ATCC or DSMZ. Cells were scraped from plates into TE buffer (10 mM Tris, pH 8, 1 mM EDTA) then genomic DNA was isolated using the E.Z.N.A. Bacterial DNA Kit (Omega Bio-Tek, Norcross, GA). DNA from the Archaeon *Methanobrevibacter smithii* ATCC 35061 was kindly provided by Jeffrey Gordon, Washington University, St. Louis, MO. The identity and purity of each bacterial DNA was verified by paired-end Sanger sequencing of 96 near full-length 16S clones generated by 16S PCR and analysis by blastn. DNA concentration was determined by Picogreen assay and the genomic DNA from each organism were then mixed, based on theoretical calculations of rDNA content, in grams per genome equivalent, to create the MCs. Mixes were formulated to contain 100,000 16S copies per organism or variable copies of the 16S gene, ranging from 1,000 to 1,000,000 copies per organism per aliquot (5 ng/µl) for 16S PCR. More information and related data are available at http://hmpdacc.org/HMMC.

Note: the qPCR for Deinococcus gave a very low value so the amount of DNA added was inflated more than 10-fold.

16S rRNA gene sequencing

Samples were amplified and sequenced using the Roche-454 FLX Titanium platform according to the "HMP 16S Protocol" (http://www.hmpdacc.org/doc/HMP_MDG_454_16S_Protocol.pdf) and in [21]. Amplification primers were designed with FLX Titanium adapters (A adapter sequence, 5' CCATCTCATCCCTGCGTGTCTCCGACTCAG 3'; B adapter sequence: 5' CCTATCCCCTGTGTGCCTTGGCAGTCTCAG 3') and a sample barcode sequence where applicable. Forward primers contained the B adapter and the reverse primers contained the A. The human filtered 16S data set is available from http://hmpdacc.org/HMR16S.

The 16S primer sequences used for phase I data generation are listed as follows:
V35 region primers
357F (V3 primer) 5'CCTACGGGAGGCAGCAG3'
926R (V5 primer) 5'CCGTCAATTCMTTTRAGT3'
V13 region primers
27F (V1 primer) 5'AGAGTTTGATCCTGGCTCAG3'
534R (V3 primer) 5'ATTACCGCGGCTGCTGG3'

For production runs, a minimum of 5,000 reads were attempted for each amplicon from a HMP donor sample. Amplicons that produced fewer than 3,000 reads passing QC were sequenced a second time to reach the deliverable of 3,000 reads passing QC. Minimal metrics for quality reads were a) > 300 nt (raw read) and b) minimum of 300 Q20 bases. A center could choose to use the same amplicon, or produce a new amplicon for the resequencing. If after two sequencing attempts from an amplicon, the minimum number of 3,000 reads passing QC was not achieved, no further sequencing was required to be completed. Any further sequencing undertaken for the sample in question was at the discretion of the sequencing center.

Choice of 16S variable regions

As part of the development of a suitable 16S protocol to be used by all participating centers in the HMP, comparisons of 16S variable regions were conducted using both the MC and HMP donor samples. We refer the reader to the Jumpstart Human Microbiome Project Data Generation Working Group paper [21] which describes these analyses and overall development of the HMP 16S protocol in detail. However, several of the most salient points are summarized here. The highest quality data generated from 454 sequencing in terms of diversity (as OTU estimations using the MC as a standard) taxonomic classifiability of reads, and lowest read error rates, were generated using the V13 and V35 regions. Overall, the V35 region was determined to provide the most precise representation of our MC. However, comparisons of our protocol in the V13 and V35 regions to identical HMP donor samples revealed different profiles of the microbial communities. Overall, it is not readily apparent that any one region or portion of the 16S gene can be selected as the most precise (or "best") for use in all scenarios and in some instances the use of multiple 16S regions may be most beneficial [21].

16S data processing

Two complementary 16S data processing pipelines were implemented using the mothur [22] and QIIME [23] software packages, respectively. The mothur pipeline included low and high stringency sequence processes, both allowing 1 unambiguous mismatch to the sample barcode and 2 mismatches to the adjacent PCR primer. Sequences with an ambiguous base call or a homopolymer longer than 8 nt were removed from subsequent analyses. For the high stringency pipeline, we then calculated the average quality score within a 50 bp window that was moved along the sequence. When the average quality score dropped below 35, the sequence was trimmed. For the low stringency pipeline, we trimmed the sequences at the position where the cumulative average quality score

dropped below 35. All sequences were then aligned using a NAST-based sequence aligner to a custom reference based on the SILVA alignment. Sequences that were shorter than 200 bp or that did not align to the anticipated region of the reference alignment were removed from further analysis. Chimeric sequences were then identified using the mothur implementation of the ChimeraSlayer algorithm trained to the "Gold" database (http://microbiomeutil.sourceforge.net) aligned to the SILVA reference alignment.

For the high stringency pipeline, we insured that all sequences overlapped in the same alignment space by trimming the ends of each sequence so all sequences began and ended at the same alignment coordinates; this was not performed for the low stringency pipeline. The high stringency sequences were then pre-clustered by merging sequence counts that were no more than 2 nt different from a more abundant sequence. Based on preliminary analyses using mock communities, we anticipated that the high stringency approach had an error rate of approximately 0.02% while the low stringency approach had an error rate of approximately 0.40%. Sequences processed by both approaches were then classified using a Bayesian classifier trained on the April 6, 2010 release of the Ribosomal Database Project (RDP) training set (http://sourceforge.net/projects/rdp-classifier). Definition of a sequence's taxonomy was determined using a pseudobootstrap threshold of 80%. As sequences of varying length classify differently, only sequences that could be classified to the genus-level were used from the low stringency analysis. Sequences from the high stringency pipeline where assigned to operational taxonomic units at a 3% distance cutoff using the average neighbor clustering algorithm. The genetic distance between all pairs of sequences was calculated assuming that insertions and deletions represented a single mutation. Outputs from both processes are available at http://hmpdacc.org/HMMCP (mothur) and http://hmpdacc.org/HMQCP (QIIME).

As a final note for completeness, an additional data set from the phase I 16S sequences was generated in the early stages of phase I analysis (prior to final development of the mothur and QIME processes). Outputs from the deconvolution, chimera filtering and RDP classifications of this early phase work is available at http://hmpdacc.org/HM16STR.

Whole-genome shotgun sequencing

A subset of samples was selected for metagenomic sequencing. After DNA extraction following the defined protocol, nucleic acid samples were quantified and checked for purity of the DNA, and only samples with a minimum of 50-100 ng of DNA were used.

Sequencing on Illumina GAIIx platform. Libraries were prepared following a standard protocol from Illumina with the following modifications. Library preparation was automated. The Agilent Bravo was employed for all reagent transfers, application of Agencourt AMPure XP bead clean-ups, and QPCR setup for enriched library quantification. DNA was sheared using the CovarisTM S2 or E210 System (Applied Biosystems) resulting in fragment insert sizes of on average 194 nt (sd. 27). Gel size selection was excluded in order to maximize yields and molecular diversity of low input samples. Cluster amplification was performed using the Illumina cBot Cluster Generation System prior to Flowcell loading on the GAIIx instrument employing the 101 bp, paired-end (PE) reads approach.

Sequencing on 454 FLX Titanium platform. Library construction, emPCR, enrichment and 454 sequencing were performed following the manufacturer's standard protocols with several modifications. Specifically, qPCR was used to estimate the number of molecules needed for emPCR. An automation system (BioMek FX, Beckman Coulter) was used to "break" the emulsions after emPCR and butanol was used to enable easier sample handling during the breaking process. The bead enrichment process was automated, employing the Robotic Enrichment Module (REM e, Roche).

Description of WGS read processing

To process WGS reads, we followed a series of steps to ensure quality and privacy of the datasets. The main steps in the process were: as follows: a) identify and mask human reads, b) remove duplicated reads, and c) trim low quality bases. Here we describe these steps in detail. Raw read data were first submitted to NCBI's SRA by the sequencing centers.

At NCBI, reads identified as human were masked using BMTagger (available at ftp://ftp.ncbi.nlm.nih.gov/pub/agarwala/bmtagger). The sequencing centers subsequently downloaded full SRA-formatted files from the SRA ftp for each Sequence Read Record (SRR) id. To assure download integrity, md5 sums (crypotographic hash functions, or digital fingerprints) were also downloaded and confirmed upon download completion. Once downloaded, fastq-dump v1.2.0 (part of the SRA Toolkit) was used to extract fastq files using the following options: fastq-dump -E -A $srr_id -D $srr_id/ -DB '@$sn/$ri' -DQ '+$sn/$ri' -O $sample/ >& $sample/$srr_id.fastq-dump. The –E option assured that all bases were written to the fastq file, while the –DB and –DQ options assured proper read naming. Once fastq files were created for each run, they were aggregated into a single fastq by sample (SRS id). This aggregated fastq file was then converted into BAM format using FastqToSam [24]. Duplicate reads were marked and removed using a modified version of EstimateLibraryComplexity, part of the Picard tool package for manipulating SAM and BAM formatted data (http://picard.sourceforge.net/index.shtml). This tool employs a method for identifying duplicate reads which are artifacts of the sequencing process without requiring alignment to a reference.

Lastly, we trimmed low quality sequence using a modified version of trimBWAstyle.pl (trimBWAstyle.pl. Fass, J., Unpublished, The Bioinformatics Core at UC Davis Genome Center), that worked directly with BAM files. This script was used to trim bases off the ends of reads which had a quality value of 2 or lower. This threshold was chosen to remove all bases of uncertain quality as defined by Illumina's End Anchored Max Scoring Segments (EAMMS) filter, and which were marked with quality values of 2. Reads trimmed to less than 60 bp were removed, and their partners, if longer than 60 bp were placed in a separate singletons file. This script produces three fastq files, one for each paired end read, and a third for singletons left after trimming. These processed read files are available from http://hmpdacc.org/HMIWGS (Illumina) and http://hmpdacc.org/HM4WGS (454).

Description of assembly

Data Assembly parameters. The processed sequences were assembled with SOAPdenovo V 1.04 [25] using the following parameters: -K 25 (k-mer size) –R –M 3 –d 1. Contig and scaffold sequences in FASTA format as well as scaffold information in AGP format were extracted from the .scafSeq file using a custom Perl script fasta2apg.pl, which is available for download from the HMP DACC. Only scaffolds larger than 300 bp were retained. All contigs contained in scaffolds longer than 300 bp were retained irrespective of their size. Contigs are available from http://hmpdacc.org/HMASM. The processed sequences from 454 and illumina hybrid data were assembled using Newbler [26] and contigs are available from http://hmpdacc.org/HMHASM.

Body site-specific assemblies of unassembled reads. Reads that did not assemble on a per sample basis, where pooled per body site and subjected to body site specific assembly. As SOAPdenovo does not directly report the placement of reads within the assembled contigs, read placement was reverse engineered through alignment of the input reads against the assembled contigs. The alignment was performed with bowtie (version 0.12.7) using parameters -v 1 -M 2 using only the first 25 bp of each read. The resulting alignments were post-processed using the get_singles.pl script distributed through metAMOS (available at http://www.cbcb.umd.edu/software/metamos/) to identify a set of unassembled reads. Unassembled reads were grouped together for each body site and subsequently assembled using SOAPdenovo with the same parameters used in the original assembly. The insert size was set to 180 bp for all body site specific assemblies. A single value was used for computational convenience (due to experimental variability each sample has a different library size) and this value was selected to reflect the average insert size of the whole HMP dataset).

Aggregate assembly statistics. Aggregate statistics were generated for all HMP assemblies, as well as for all of the assemblies downloaded from the MetaHit project [27] using the statistics.pl script in the metAMOS package. The script generates a collection of 'standard' aggregate measures of assembly quality, such as number of contigs, total size, maximum, average, and median contig sizes. It is important to note that the commonly reported N50 number (largest contig size c such that half of the genome is contained in contigs larger than c) cannot be used in the context of metagenomic assembly as the total genome size is unknown. Instead we report the size and number of contigs necessary to cover 1, 2, 4, and 10 Mb of DNA (Supplementary Fig. 5).

Assembly collector's curves. The shotgun reads were mapped to the assembled contigs using bowtie (version 0.12.7) and the first 25 bp of each read. Depth of coverage information was extrapolated from the alignment data and reported together with contig size information using the get_coverage.pl script in the metAMOS package. The coverage information was sub-sampled using the R script coverage.R (metAMOS) and aggregate assembly statistics computed on the resulting data. To simulate the effect of low coverage on assembly without requiring re-assembly of subsampled read sets, we eliminated a contig from consideration once the corresponding depth of coverage dropped below 3 (Supplementary Fig. 6).

Gene annotation

Annotation overview. Genes were called by use of the ab-initio gene finder MetageneMark [28] using a minimum length cutoff of 20 amino acids. Predicted genes that overlapped with possible ncRNA genes were identified (as described in [29]) were removed. Specifically, all 690 QC-passed metagenomic shotgun assemblies (composing some 76.5 million genes) were screened using BLAT against the SILVA database (SSURef/LSURef) [30]. Those genes that matched with >90% identity over 50% or more of their lengths were removed (n=153). Predicted peptides from metagenome and hybrid assemblies were functionally annotated using the JCVI metagenomic annotation pipeline [31]. This automated annotation process assigns functional (most probable biological role) annotation based on the evidence provided by a series of ranked homology searches. In this way, each putative protein is given an annotation that is both as reliable and as comprehensive as can be supported by the available collection of homology-based evidences.

The first evidence set used includes a BLASTP search against the UniRef 100 (release-2010_11) database which is a comprehensive set of clustered protein sequences (clustered at 100% identity) provided by the Uniprot Knowledgebase (UniProtKB). The second data collection component is the search against Hidden Markov Models (HMMs) comprising Pfam version 24.0 [32] and TIGRfam version 9.0 [33] models. In all cases, standard trusted cutoffs were used. The HMM hits were organized into ordered isology type classes ("isotypes"), each of which represents a different degree of confidence concerning the functional assignment. The final data collection component involves searches for lipoprotein motifs and transmembrane helices in the putative proteins. The former is accomplished using a regular expression search in the amino acid sequence, while the latter is performed using TMHMM, [34] a HMM-based search for transmembrane motifs. These two searches represent annotation states that fall well short of complete functional annotation (e.g., "putative lipoprotein"), but are more informative than the absence of any functional annotation. Annotation is then assigned using a value hierarchy scheme, established through a process of manual curation efforts. Putative proteins without any evidence, including those from TMHMM or lipoprotein motif searches, are classified as "hypothetical". The annotation summary includes gene common name, gene symbol, EC numbers and GO terms for each putative protein that are assigned using a value hierarchy scheme. Annotations referred to as a 'Gene Index', from the assemblies of Illumina only data (using SOAPdenovo) and the 454/Illumina hybrid data (using Newbler), are available in gff3 format at http://hmpdacc.org/HMGI and http://hmpdacc.org/HMHGI, respectively. A summary of annotation attributes for the Illumina only data (RES:HMGI) are given in Supplementary Table 7.

"GO slim" analysis. Gene Ontology (GO) annotations were generated for the predicted genes identified from the metagenomic shotgun assemblies using the JCVI metagenomic annotation pipeline [31]. A "GO Slim", or reduced set of more general terms into which more specific annotations can be collated, was used to find overall trends in the relative abundance of genes involved in various biological processes in different body sites. The GO Slim was developed at the DACC (http://www.geneontology.org/GO.slims.shtml). The slim contains all the first level children of the GO root "biological process" and additional specific terms under metabolism and regulation (DACCslim1). The number of

gene assignments into each slim term was calculated using the GO map2slim script (REF: http://search.cpan.org/~cmungall/go-perl/scripts/map2slim). Metastats was used to determine relative abundance of assignments to each term and to calculate statistical significance [35]. GO slim annotations are available at http://hmpdacc.org/HMGS and a description of the GO slim is given in Supplementary Table 6.

Quality control of metagenomic sequences. The WGS reads were screened using variables derived for assemblies, gene predictions, pathway reconstructions, and species abundance. Specifically, these variables for each sample were: 1) total number of reads, 2) total number of reads incorporated into each assembly, 3) combined length of all assemblies compared on the total amount of nucleotides used to generate each assembly, 4) total number of gene predictions derived from each assembly, 5) counts of a subset of 16S rRNA sequences compared to each assembly, 6) volume of human contamination, 7) lack of concordance for human contamination across multiple Illumina lanes, 8) inconsistent biochemical pathway composition, 9) inconsistent abundance profiles of each read compared to a reference data set and, 10) inconsistent kmer composition of reads. See below for calculation of values for 7-10. Independent distributions, for each body site, were then derived from variables 1-10. Samples scoring as outliers of two times the standard deviation from the mean for each distribution were eliminated from further processing. After manual review, variables 1-5 were considered to be non-independent and combined into a single class; variables 6-10 were treated as five additional independent classes. Samples were rejected if they had been flagged by at least two of the six independent classes.

Read-based metabolic reconstruction. Sequences were mapped using MBLASTX (MulticoreWare, St. Louis, MO) with default parameters against a functional sequence database including the KEGG orthology v54. Up to the 20 most significant hits at E<1 were provided as hits to HUMAnN generating abundance and coverage results for each KEGG metabolic pathway and module of interest Functional databases used are available at http://hmpdacc.org/HMFUNC and output of the metabolic reconstruction pipeline at http://hmpdacc.org/HMMRC.

As part of the overall phase I analyses completed by the HMP, selected functional categories of special interest were examined in more detail using metabolic reconstruction data. The output of theses analyses of 'genes of interest' are available at http://hmpdacc.org/HMGOI.

Shotgun based community profiling. Processed whole-genome shotgun reads were mapped onto reference genomes in order to calculate organism abundances. Fasta files containing a total of 38,691,635,796 microbial reads were subjected to a low-complexity screen using the 'dust' program (distributed with NCBI blast). Reads with fewer than 60 non-masked bases (not necessarily consecutive) were considered to be of low complexity and discarded from the final set. In cases where one end of a paired end set of mates was found to be of low complexity, and the other end was not, the orphaned (but good quality) read was removed from the paired end file, and moved into the fragment read file. The final set were aligned using the aligner clc_ref_assemble_long (CLC Assembly Cell package, CLC bio, http://www.clcbio.com/) with the parameters "-lengthfraction 0.75 -similarity 0.8 -p fb ss 180 250" (note: -p fb ss 180 250 sets paired end information,

'fb' indicates that the first read is in the 'forward' orientation, and the second is in the 'b'ackward' orientation (i.e. facing each other), the 'ss 180 250' part informs the program to expect the 's'tart to 's'tart (i.e. the far ends, since they are facing each other) distance between the reads to be from 180-250 bp in length). Both the paired end set and the fragment file were aligned in a single execution of the software.

The reference database used was comprised of all archaeal, bacterial, lower eukaryote and viral organisms available in GenBank. These sequences were downloaded via keyword search from the NCBI's GenBank on 11/10/2009, and were periodically updated over the course of the project. The archaeal, lower eukaryote and viral components were taken 'as-is' from the keyword searches "Archaea[ORGN]", "Virus[ORGN]" and "Eukaryota[ORGN] NOT Bilateria[ORGN] NOT Streptophyta[ORGN]" respectively. The bacterial component started with a similar keyword search, "Bacteria[ORGN] and complete" and "Bacteria[ORGN] and WGS", and was subject to special processing to remove highly redundant strains, while retaining all reference genomes sequenced as part of the HMP. Contigs were grouped into their respective genomes based on their GenBank ID ranges in random order. To assist with downstream identifications, all sequences from a given genome were tagged with a prefix id unique to that strain. This allows a hit to any contig in a draft genome to be easily related back to its parent genome, and was a required step to enable the creation of abundance metrics per genome.

The complete and draft genomes were categorized on per species level, resulting in categories including single strain up to over 50 strains per species (e.g. *Escherichia coli* and *Bacillus anthracis*). Redundancy removal was implemented to exclude strains with nearly identical sequences. For selecting representatives among multiple strains within a species, we used the Mauve program [36]. The mauveAligner module of Mauve program was wrapped into custom-built PERL scripts to automate most of the process. Our criteria were simple, if there was more than 90% similarity between two genomes, we would pick the longer one. Mauve worked well for the smaller number of genomes that were in one or very few sequences. However the challenges grew when the number of sequences increased and as the homology decreased among greater numbers of genomic pieces. In some of cases many pair-wise alignments were done and the sequences were eliminated progressively. In cases of large numbers of strains, a slightly relaxed homology (as low as 82-83%) was used. An additional filter was used to verify if a given strain was known to have originated from human (i.e. is part of the HMP project). Since our focus is the study of the human microbiome, human originated references were excluded from the removal process. Finally, plasmids corresponding to the non-redundant genomes that were selected through the above analysis were added in. The final reference database that was used in the analysis for this paper contained 1742 bacterial strains, 131 Archaea, 326 strains of lower eukaryotes, and 3683 strains of virus. The process of removing highly redundant bacterial strains resulted in the elimination of 2265 complete genomes, draft genomes, and plasmid sequences. The WGS alignments are available at http://hmpdacc.org/HMSCP and the reference genome set is available at http://hmpdacc.org/HMREFG.

Comparative metagenomics. The JCVI METAREP software [37] (an open source tool to query, browse, and compare extremely large volumes of metagenomic annotations) processed the annotation output for over 1,000 samples (498 read-based metabolic

reconstructions from the HUMAnN pipeline, 15 hybrid assemblies and 690 Illumina only assemblies) to facilitate comparative genomic analyses. The dedicated HMP instance of the software and output is accessible at http://www.jcvi.org/hmp-metarep.

Correspondence between 16S and WGS datasets

Additional methodology regarding the comparison of accumulation curves for 16S OTU and gene predictions from WGS data presented in Figure 1 of the main paper.

Collector's curves. The median and 95% confidence intervals of the number of taxa or genes discovered (Fig. 1a and b, Supplementary Fig. 9) were computed using a bootstrapping approach by resampling among the available donors for each site. The data set used for taxonomic analysis was the V35 OTU set derived from the mothur pipeline (RES:HMMCP) and for genes is given below (see 'Gene accumulation'). To improve the visual clarity of the multiple body site (habitat) collectors' curves, confidence intervals were plotted periodically to display the degree of overlap between the body habitats. As expected, body habitats with fewer available donors, such as from the vaginal region, tended to have much broader CIs. The posterior fornix does not appear to be nearing saturation, most likely due to the high variation in the samples resulting from the smaller donor count and low taxonomic diversity within the site. To plot all body habitats in a single figure, the discovered taxa or genes (y-axis) were logarithm based 10 transformed. This was necessary due to the significant difference of discovery counts between stool and remaining body habitats which made the relative differences between non-stool body habitats visually indistinguishable using a linear scale.

Gene accumulation. Gene accumulation values from HMP data used for plotting the accumulation curves (Fig. 1b) were obtained as follows. The predicted protein sequence of ORFs generated from WGS assemblies (Illumina only) by the annotation pipeline from each body habitat, were reduced into non-redundant gene sets (per sample within each body site) using USEARCH [38] with thresholds of 95% identity and the aligned length covering over 90% of the shorter gene. Genes were clustered using the Markov Cluster (MCL) algorithm [39] with an inflation factor of 1.1. Clusters were counted as sample number increased using 100 permutations and random selection of samples.

Calculation of Log(Gene/Taxa). An estimate of the gene-to-taxa discovery ratio was computed for anterior nares, buccal mucosa, posterior fornix, stool, and supragingival plaque (Fig. 1c). This estimate was made by taking the median number of genes discovered and dividing this by the median number of taxa discovered for each sample collected. The based-10 logarithm of this ratio was then plotted for each body site (y-axis) for each of the number of samples taken (x-axis). The labeled ratios for each curve represent the average number of unique genes contributed per unique OTU at the final sample count.

If sampling were to approach infinity, the number of genes and taxa would both stabilize and the gene-to-taxa ratio would become the actual average number of unique genes contributed per unique taxa. As a proportion, the gene-to-taxa discovery ratio would never approach zero as this would imply that the number of taxa discovered is infinitely greater than the number of genes discovered. A comparison of the labeled ratios indicates

that the genetic diversity per organism is greatest in decreasing order approximately as follows: stool, buccal mucosa, posterior fornix or anterior nares, and supragingival plaque.

Comparisons of HMP and MetaHIT WGS data from stool

Comparisons of Open Reading Frames (ORFs) (as predicted proteins) from HMP stool samples, to those generated by the MetaHIT project [27] (Fig. 1b, Supplementary Fig. 10, Supplementary Fig. 11, Supplementary Table 8) were completed as follows. The predicted protein sequence of ORFs generated from WGS assemblies (Illumina only) by the annotation pipeline from all stool samples (n=138) (http://hmpdacc.org/HMGI) were combined. From this pooled set, the genes were reduced into a non-redundant gene set using USEARCH [38] with thresholds of 95% identity and the aligned length covering over 90% of the shorter gene. This non-redundant set consisted of 5,140,472 genes. A non-redundant gene set generated previously by the MetaHIT project using the same procedure followed by the HMP from a pooled set of genes from 124 stool samples was used for further evaluation [27]. This set as reported by MetaHIT consisted of 3,299,822 ORFs [30].

The union of the HMP and MetaHIT non-redundant gene sets was subsequently analyzed as follows. First, the protein predictions from the combined set were searched against EggNOG [40] using BLASTP with cutoff thresholds of an e-value < $1e^{-6}$ and bits per position < 1. These results represented matches to known functional annotation. Next, those genes without any assignment to an orthologous group (3,401,774 for HMP and 2,144,077 MetaHIT data, respectively) and therefore representing novel genes, were clustered using USEARCH at 80% protein identity and MCL using a 1.1 inflation factor. The distribution of genes from HMP and MetaHIT into these clusters was then examined.

Comparison of HMP WGS reads to HMP reference genomes

An analysis to elucidate the contribution of HMP WGS reads that could be aligned to reference genomes contributed by the HMP was conducted as follows. First, alignments of Illumina only reads to an all inclusive database of genomes (RES:HMREFG) was completed as described earlier in the supplemental text ('Shotgun based community profiling' P. S9). In this analysis, HMP WGS reads were aligned using match criteria of 75% nucleotide identity over 80% of the read length to the 'all genome' (RES:HMREFG) database. To obtain the results of those reads mapping to the subset of genomes that were contributed by the HMP the following steps were taken. The WGS alignments passing a low complexity filter (http://hmpdacc.org/HMSCP) from 754 samples were extracted and further screened using a list of 25,758 unique contig identifiers representing 223 HMP reference genomes. Sequence Alignment/Map (SAM) tools were used to convert the binary BAM alignments into tab-delimited SAM files and the extraction of the desired HMP data was completed using custom Perl and shell scripts.

From a total of 38,237,669,683 reads, 58% (22,034,286,362/38,237,669,683) could be aligned to any reference genome. From WGS read alignments to the total of 25,758 contigs that were grouped as 223 HMP reference genomes, it was determined that ~26%,

or 10,102,664,871 reads could be aligned to HMP reference genomes from the total set of HMP WGS reads. Further, ~46% of the HMP WGS reads that could be aligned to any reference genome, were found to match the HMP reference genome set (10,102,664,871 reads aligned to HMP reference genomes from 22,034,286,362 reads which could be aligned to any genome in the set of reference genomes).

The range of aligned reads to the HMP reference data set revealed a minimum of 19,157 reads aligned to *Lactobacillus hilgardii* and a maximum of 636,327,876 reads aligned to *Corynebacterium matruchotii* with an average of 48,805,144 and a median of 9,232,831 reads aligned, respectively. Only 16 of the 223 HMP genomes (7%) yielded no matches. The details of these results can be viewed in the downloadable Supplementary Data file "**HMP_WGS_MAPPED_TO_HMP_REF_GENOME.xls**."

Supplementary References

18 Yilmaz, P. et al. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nat Biotechnol* **29** (5), 415-420 (2011).

19 Field, D. et al. The minimum information about a genome sequence (MIGS) specification. *Nat. Biotechnol* **5**, 541-547 (2008).

20 Aagaard, K. et al. A Comprehensive Strategy for Sampling the Human Microbiome. (in review).

21 Jumpstart Human Microbiome Project Data Generation Working Group, Evaluation of 16S rDNA-based community profiling for human microbiome research *PLoS One*. http://dx.plos.org/10.1371/journal.pone.0039315 (in the press).

22 Schloss, P.D. et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* **75** (23), 7537-7541 (2009).

23 Caporaso, J.G. et al., QIIME allows analysis of high-throughput community sequencing data. *Nature methods* **7** (5), 335-336 (2010).

24 Li, H. The Sequence Alignment/Map format and SAMtools. *Bioinform* **25** (16), 2078-2079 (2009).

25 Li, R. et al. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res* **20** (2), 265-272 (2010).

26 Margulies, M. et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437** (7057), 376-380 (2005).

27 Qin, J. et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464** (7285),59-65 (2010).

28 Zhu, W., Lomsadze, A., Borodovsky M. Ab initio gene identification in metagenomic sequences. *Nucleic Acids Res* **38** (12), e132 (2010).

29 Tripp, H.J, Hewson, I., Boyarsky, S., Stuart, J.M., Zehr, J.P. Misannotations of rRNA can now generate 90% false positive protein matches in metatranscriptomic studies. *Nucleic Acids Res* **39**, 8792-8802 (2011).

30 Pruesse, E. et al. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* **35** (21), 7188-7196 (2007).

31 Tanenbaum, D. M. et al. *Standards in Genomic Sciences* **2** (2), 229 (2010).

32 Finn, R.D. et al. The Pfam protein families database. *Nucleic Acids Res* **38**, 211 (2010).

33 Haft, D.H., Selengut, J., White O. The TIGRFAMs database of protein families. *Nucleic Acids Res* **31** (1), 371-373 (2003).

34 Sonnhammer, E.L., von Heijne, G., Krogh, A. A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc In Conf Intell syst Mol Biol.* **6**, 175-182 (1998).

35 White, J.R., Nagarajan, N., Pop, M. Statistical methods for detecting differentially abundant features in clinical metagenomic samples. *PLoS Comput Biol* **5** (4), e1000352 (2009).

36 Darling, A.C., Mau, B., Blattner, F.R., Perna, N.T. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res* **14** (7), 1394-1403 (2004).

37 Goll, J., Thiagarajan, M., Abubucker, S., Huttenhower, C., Yooseph, S, Methé, B.A. A case study for large-scale human microbiome analysis using JCVI's Metagenomics Reports (METAREP). *PloS One*. http://dx.doi.org/10.1371/journal.pone.002904 (in the press).

38 Edgar, R.C. Search and clustering orders of magnitude faster than BLAST. *Bioinform* **26** (19), 2460-2461 (2010).

39 Lattimore, S.B, van Dongen, S., Crabbe, M.J. GeneMCL in microarray analysis. *Comput Biol Chem* **29**(5), 354-359 (2005).

40 Muller, J., et al. eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. *Nucleic Acids Res* **38**, 190-195 (2010).

The HMP sequence resources available at SRA	
Project ID	**Title**
SRP004306	16S rRNA 454 Protocol Development-Mock
SRP004311	Metagenomes Mock Pilot (454 & Illumina)
SRP002012	16S rRNA 454 Clinical Production Pilot
SRP002395	16S rRNA 454 Clinical Production Phase I
SRP002163	Metagenomes Production Phase
The HMP sample metadata available at dbGaP	
Study Accession	**Title**
phs000228	HMP- Core Microbiome Sampling Protocol A (HMP-A)

Supplementary Table 1. HMP unprocessed sequence and sample metadata available from the SRA and dbGaP.

Additional HMP Resources	
Short Description of Resource	**Resource Location**
NCBI BioProject page -HMP unprocessed 16S, WGS and reference genome sequence	http://www.ncbi.nlm.nih.gov/bioproject/43021
HMP demonstration projects (microbiome correlations to disease)	http://commonfund.nih.gov/hmp/fundedresearch.aspx
	http://hmpdacc.org/impacts_health/impact_health.php
HMP projects (developing tools and technologies for microbiome research)	http://commonfund.nih.gov/hmp/fundedresearch.aspx
	http://hmpdacc.org/tech_development/tools.php
HMP projects (ethical, legal & social implications of microbiome research)	http://commonfund.nih.gov/hmp/fundedresearch.aspx
	http://hmpdacc.org/ethical/ethical.php
List of publications from HMP supported projects	http://commonfund.nih.gov/hmp/fundedresearch.aspx
	http://hmpdacc.org/pubs/publications.php
HMP core microbiome sampling (Protocol A)	http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/GetPdf.cgi?id=phd002854.2
	http://hmpdacc.org/doc/HMP_Clinical_Protocol.pdf
HMP manual of procedures	http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/GetPdf.cgi?id=phd003190.2
	http://hmpdacc.org/doc/sops_2/manual_of_procedures_v11.pdf
HMP clinical sampling overview	http://hmpdacc.org/micro_analysis/microbiome_sampling.php
HMP tools and protocols (reference genomes, 16S, WGS, other analyses)	http://hmpdacc.org/tools_protocols/tools_protocols.php
HMP Reference genomes data	http://hmpdacc.org/HMRGD/
HMP "most wanted" list of strains	http://hmpdacc.org/most_wanted/
HMP project catalogue	http://www.hmpdacc-resources.org/hmp_catalog/main.cgi
IMG data warehouse and analytical system	http://www.hmpdacc-resources.org/cgi-bin/imgm_hmp/main.cgi
BEI- strains repository (HMP collection)	http://www.beiresources.org/tabid/1901/stabid/1901/CollectionLinkID/4/Default.aspx
METAREP- Comparisons of annotations from HMP WGS samples	http://www.jcvi.org/hmp-metarep

Supplementary Table 2. Additional HMP resources. Links to additional resources developed by the HMP are listed. In some instances, there are multiple links to a resource. Mock Community cells (BEI:HM-280, HM-281) and DNA extracts (BEI:HM-278D, HM-279D) are available from the BEI HMP Collection.

Processed Phase I Data Sets Available at the DACC		
Data Set Identifier	**Short Description of Data Set**	**Data Location**
RES:HMMC	16S and WGS reads generated from Mock Community	http://hmpdacc.org/HMMC
RES:HMR16S	16S human filtered reads (V1V3, V3V5) & library metadata	http://hmpdacc.org/HMR16S
RES:HMMCP	16S data from mothur pipeline	http://hmpdacc.org/HMMCP
RES:HMQCP	16S data from QIME	http://hmpdacc.org/HMQCP
RES:HM16STR	16S data from early Phase I processing	http://hmpdacc.org/HM16STR
RES:HMIWGS	WGS Illumina GAIIx reads (human and low quality reads removed)	http://hmpdacc.org/HMIWGS
RES:HM4WGS	WGS 454 sequence reads (human and low quality reads removed)	http://hmpdacc.org/HM4WGS
RES:HMASM	Assemblies of WGS (Illumina) data	http://hmpdacc.org/HMASM
RES:HMBSA	Body site specific assembly data	http://hmpdacc.org/HMBSA
RES:HMHASM	Hybrid WGS (Illumina/454) assemblies	http://hmpdacc.org/HMHASM
RES:HMREFG	Reference genomes used for WGS alignments	http://hmpdacc.org/HMREFG
RES:HMSCP	WGS alignments to reference genomes	http://hmpdacc.org/HMSCP
RES:HMFUNC	Metabolic reconstruction database	http://hmpdacc.org/HMFUNC
RES:HMMRC	Metabolic reconstruction pipeline output	http://hmpdacc.org/HMMRC
RES:HMGI	'Gene Index'- annotations from metagenomic assemblies	http://hmpdacc.org/HMGI
RES:HMGS	'GO Slim' Annotations	http://hmpdacc.org/HMGS
RES:HMHGI	'Gene Index'- annotations from hybrid metagenomic assemblies	http://hmpdacc.org/HMHGI
RES:HMGC	Clustered 'Gene Index' (gene accumulation data)	http://hmpdacc.org/HMGC
RES:HMGOI	Genes of interest	http://hmpdacc.org/HMGOI

Supplementary Table 3. An overview of HMP processed Phase I data available from the DACC.

		V13 Samples		V35 Samples		WGS Samples	
	total	2971		4879		681	
Age	18-20	250	8.4%	283	5.8%	50	7.3%
	21-25	1119	37.7%	1901	39.0%	232	34.1%
	26-30	910	30.6%	1662	34.1%	277	40.7%
	31-35	393	13.2%	626	12.8%	65	9.5%
	36-40	299	10.1%	407	8.3%	57	8.4%
BMI	Normal	1605	54.0%	2719	55.7%	347	51.0%
	Obese Class1	364	12.3%	667	13.7%	87	12.8%
	Overweight	1002	33.7%	1493	30.6%	247	36.3%
Body Site	Stool	193	6.5%	328	6.7%	139	20.4%
	Anterior Nares	169	5.7%	283	5.8%	84	12.3%
	Buccal Mucosa	184	6.2%	314	6.4%	107	15.7%
	Hard Palate	179	6.0%	310	6.4%	1	0.1%
	Keratinized gingiva	183	6.2%	319	6.5%	6	0.9%
	Palatine Tonsils	189	6.4%	315	6.5%	6	0.9%
	Saliva	166	5.6%	292	6.0%	5	0.7%
	Subgingival plaque	186	6.3%	314	6.4%	7	1.0%
	Supragingival plaque	192	6.5%	316	6.5%	115	16.9%
	Throat	176	5.9%	312	6.4%	7	1.0%
	Tongue Dorsum	193	6.5%	320	6.6%	122	17.9%
	Left Antecubital Fossa	158	5.3%	221	4.5%	0	0.0%
	Left Retroauricular Crease	188	6.3%	295	6.0%	9	1.3%
	Right Antecubital Fossa	158	5.3%	229	4.7%	0	0.0%
	Right Retroauricular Creas	190	6.4%	304	6.2%	15	2.2%
	Mid Vagina	91	3.1%	140	2.9%	2	0.3%
	Posterior Fornix	89	3.0%	136	2.8%	53	7.8%
	Vaginal Introitus	87	2.9%	131	2.7%	3	0.4%
Breast Fed	Don't know/remember	287	9.7%	567	11.6%	112	16.4%
	No	474	16.0%	672	13.8%	124	18.2%
	Yes	2099	70.6%	3405	69.8%	421	61.8%
	na	111	3.7%	235	4.8%	24	3.5%
Collection Site	Baylor	1214	40.9%	3248	66.6%	319	46.8%
	Washington University	1757	59.1%	1631	33.4%	362	53.2%
Gender	Male	1414	47.6%	2609	53.5%	363	53.3%
	Female	1557	52.4%	2270	46.5%	318	46.7%
Race	Asian	215	7.2%	579	11.9%	57	8.4%
	Asian/White	48	1.6%	76	1.6%	0	0.0%
	Black	154	5.2%	175	3.6%	38	5.6%
	Black/White	28	0.9%	29	0.6%	0	0.0%
	White	2526	85.0%	4020	82.4%	586	86.0%
Tobacco Usage	No	2757	92.8%	4605	94.4%	636	93.4%
	Yes	214	7.2%	274	5.6%	45	6.6%
Visit Number	1	1687	56.8%	2884	59.1%	403	59.2%
	2	1272	42.8%	1968	40.3%	265	38.9%
	3	12	0.4%	27	0.6%	13	1.9%

Supplementary Table 4. The Human Microbiome Project reference population. The distributions of HMP samples assayed by 16S and WGS are given by donor metadata categories.

Genus Species	Repository Number	16S rRNA Operons per genome	Grams of genomic DNA per 16S copy	Even Mixture 16S copies*	Staggered Mixture 16S copies*
Acinetobacter baumannii	ATCC 17978	5	8.16E-16	100,000	100,000
Actinomyces odontolyticus	ATCC 17982	3	1.00E-15	100,000	1,000
Bacillus cereus	ATCC 10987	12	4.47E-16	100,000	100,000
Bacteroides vulgatus	ATCC 8482	7	7.57E-16	100,000	1,000
Candida albicans	ATCC MYA-2876	NA	2.92E-14	1,120[b]	1,000[b]
Clostridium beijerinckii	ATCC 51743	14	4.40E-16	100,000	100000
Deinococcus radiodurans	DSM 20539	3	1.05E-15	100,000	1,000
Enterococcus faecalis	ATCC 47077	4	8.25E-16	100,000	1,000
Escherichia coli	ATCC 700926	7	6.81E-16	100,000	1,000,000
Helicobacter pylori	ATCC 700392	2	8.55E-16	100,000	100,000
Lactobacillus gasseri	DSM 20243	6	3.25E-16	100,000	100,000
Listeria monocytogenes	ATCC BAA-679	6	5.03E-16	100,000	100,000
Methanobrevibacter smithii	ATCC 35061	2	9.50E-16	100,000	1,000,000
Neisseria meningitidis	ATCC BAA-335	4	5.83E-16	100,000	10,000
Propionibacterium acnes	DSM 16379	3	8.76E-16	100,000	10,000
Pseudomonas aeruginosa	ATCC 47085	4	1.61E-15	100,000	100,000
Rhodobacter sphaeroides	ATCC 17023	3	1.41E-15	100,000	100,000
Staphylococcus aureus	ATCC BAA-1718	5	5.89E-16	100,000	1000000
Staphylococcus epidermidis	ATCC 12228	5	5.13E-16	100,000	100,000
Streptococcus agalactiae	ATCC BAA-611	7	3.17E-16	100,000	1,000,000
Streptococcus mutans	ATCC 700610	5	4.17E-16	100,000	1,000
Streptococcus pneumoniae	ATCC BAA-334	4	5.54E-16	100,000	100,000

*Mixtures contain the listed number of 16S rRNA gene copies per organism in a total 5 ng mixture resuspended in 10 µl.
[a]NA, not applicable
[b]18S rRNA gene copies

Supplementary Table 5. 16S composition of the Even and Staggered Mock DNA communities.

GO ID	GO Term
GO:0000003	reproduction
GO:0000746	conjugation
GO:0001906	cell killing
GO:0002376	immune system process
GO:0005976	polysaccharide metabolism
GO:0006091	generation of precursor metabolites and energy
GO:0006139	nucleobase, nucleoside, nucleotide and nucleic acid metabolism
GO:0006276	plasmid maintenance
GO:0006281	DNA repair
GO:0006520	cellular amino acid metabolic process
GO:0006791	sulfur utilization
GO:0006794	phosphorus utilization
GO:0006805	xenobiotic metabolic process
GO:0006950	response to stress
GO:0007587	sugar utilization
GO:0008150	biological process
GO:0008152	metabolic process
GO:0008283	cell proliferation
GO:0009237	siderophore metabolism
GO:0009306	protein secretion
GO:0009307	DNA restriction-modification system
GO:0009404	toxin metabolic process
GO:0009405	pathogenesis
GO:0009758	carbohydrate utilization
GO:0009987	cellular process
GO:0015948	methanogenesis
GO:0015976	carbon utilization
GO:0016032	viral reproduction
GO:0016265	death
GO:0017144	drug metabolic process
GO:0019740	nitrogen utilization
GO:0022414	reproductive process
GO:0022610	biological adhesion
GO:0023052	signaling
GO:0030030	cell projection organization and biogenesis
GO:0030436	asexual sporulation
GO:0032196	transposition
GO:0032501	multicellular organismal process
GO:0032502	developmental process
GO:0040007	growth
GO:0040011	locomotion
GO:0042710	biofilm formation
GO:0043473	pigmentation
GO:0044237	cellular metabolic process

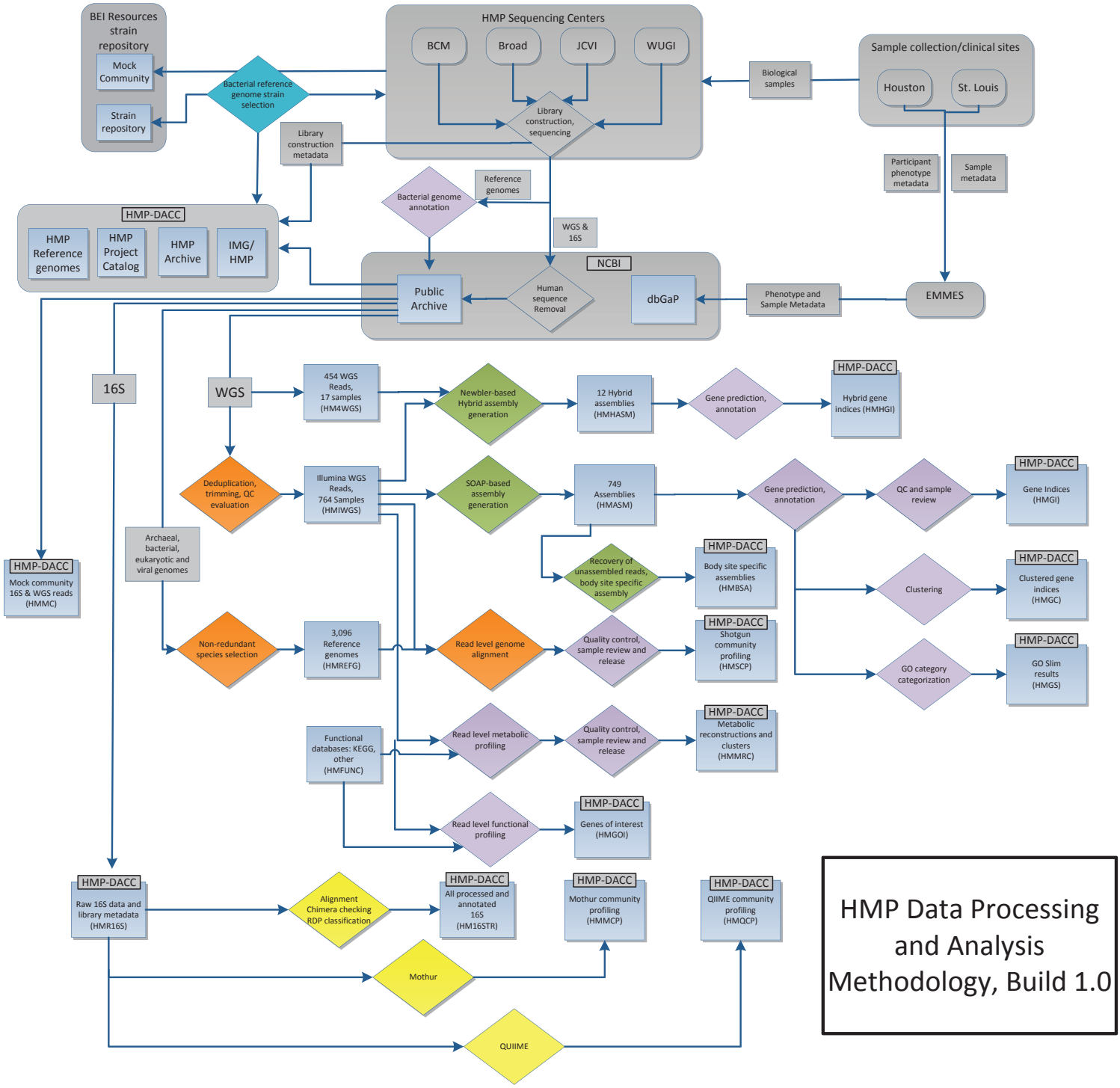
Supplementary Table 6. Gene Ontology (GO) 'slim' controlled vocabulary used for analysis of human microbiome annotation ('Gene Index'). This GO slim includes all direct children of the GO Biological Process ontology as well as selected additional, and more specific, terms in the categories of metabolism and regulation.

BODY_SITE	SAMPLE_COUNT	GENE_COUNT	FUNCTIONAL_ASSIGNMENTS	GENE_SYMBOLS	GO_TERM_ASSIGNED	GO_PROCESS_TERMS	GO_COMPONENT_TERMS	GO_FUNCTION_TERMS	EC_NUMBER_ASSIGNED	CAZY_ID_ASSGINED	TAXON_ASSIGNMENTS
Stool	139	26,572,444	70%	28%	34%	39%	20%	49%	22%	0.47%	87%
Tongue_dorsum	128	21,187,585	63%	23%	26%	29%	14%	37%	18%	0.39%	83%
Supragingival_plaque	118	21,409,084	66%	23%	26%	30%	16%	38%	18%	0.45%	84%
Buccal_mucosa	107	3,264,223	61%	21%	25%	28%	13%	35%	18%	0.30%	83%
Anterior_nares	87	449,840	74%	29%	33%	37%	19%	46%	20%	0.16%	93%
Posterior_fornix	51	186,169	48%	14%	25%	27%	11%	34%	14%	0.41%	83%
R_Retroauricular_creas	17	430,190	75%	34%	40%	44%	22%	53%	23%	0.19%	94%
L_Retroauricular_creas	9	215,231	62%	22%	25%	29%	15%	37%	17%	0.32%	82%
Subgingival_plaque	7	936,348	63%	22%	25%	29%	14%	36%	18%	0.26%	84%
Throat	7	578,676	76%	30%	39%	42%	22%	51%	23%	0.25%	94%
Keratinized_gingiva	6	314,611	69%	31%	33%	40%	20%	52%	26%	0.32%	86%
Palatine_Tonsils	6	454,544	71%	28%	31%	37%	21%	48%	21%	0.66%	86%
Saliva	3	138,135	51%	22%	29%	38%	22%	45%	18%	0.33%	76%
Vaginal_introitus	3	12,174	63%	22%	26%	30%	14%	37%	18%	0.30%	84%
Mid_vagina	2	7,947	54%	24%	30%	39%	22%	48%	20%	0.39%	81%

Supplementary Table 7. Annotation Summary Attributes. Metagenome shotgun annotation gff3 files (RES:HMGI) were parsed and counts generated for the following annotation attributes for each sample: non-hypothetical gene product name, gene symbol, Gene Ontology (GO) term, Enzyme Commission (EC) number, Carbohydrate-Active Enzymes (CAZY) id, and taxonomy. GO term counts were subdivided into the three GO ontologies, biological process, cellular component and molecular function. Samples were then categorized by body site to provide a summary of annotation attributes by site.

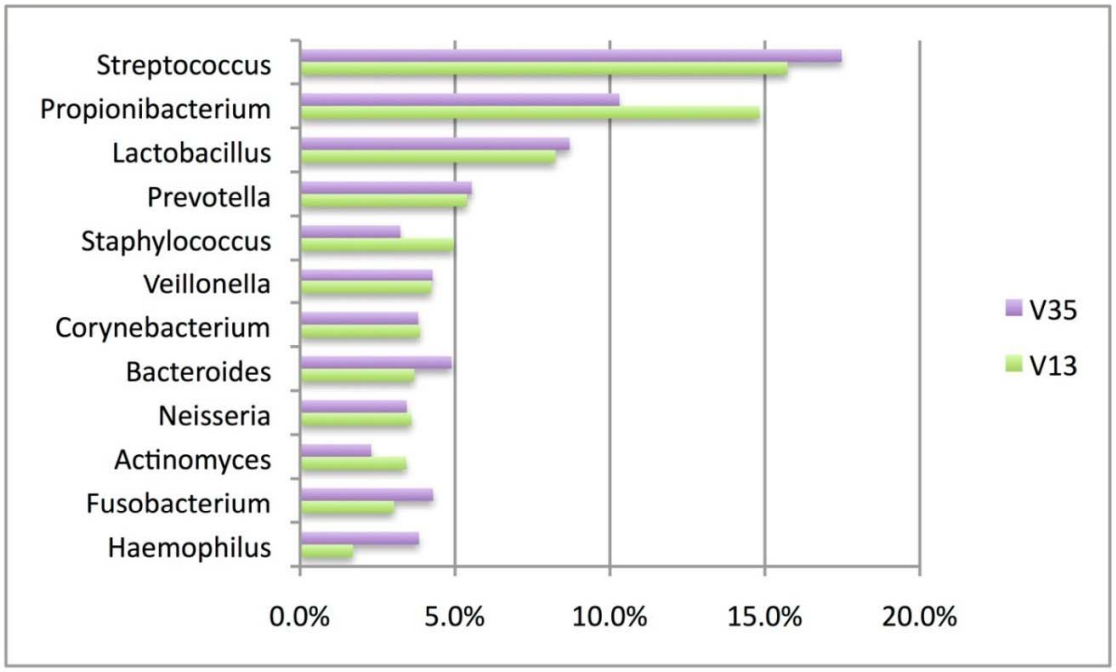
	HMP	MetaHIT	# Shared	% Shared
Total # of Non-redundant ORFs from Stool Samples	5,140,472	3,299,822		
# of NOGs	14,849	10,868	9,286	57
# of Non-redundant ORFs in NOGs	1,738,698	1,155,745		
% of Non-redundant ORFs in NOGs	34	35		
# of ORFs without NOG Assignment (novel)	3,401,774	2,144,077		
# of Novel Clusters Generated from Combined (HMP+MetaHIT) Non-redundant	931,715	816,991	769,411	79
Number of Singletons	1,422,482	778,090		
% Singletons (from total non-redundant ORFs)	28	24		

Supplementary Table 8. Comparisons of HMP and MetaHIT gene catalogues from stool. The total non-redundant ORFs recovered from stool samples (n=138) representing HMP data was compared to an analogous set generated by the MetaHIT project (n=124). First, a combined data set of HMP and MetaHIT ORFs were compared by grouping ORFs into Non-supervised Orthologous Groups (NOGs) by matches to the eggNOG database. Next, ORFs that did not receive a NOG assignment (novel ORFs) were subsequently clustered using USEARCH at 80% protein identity.

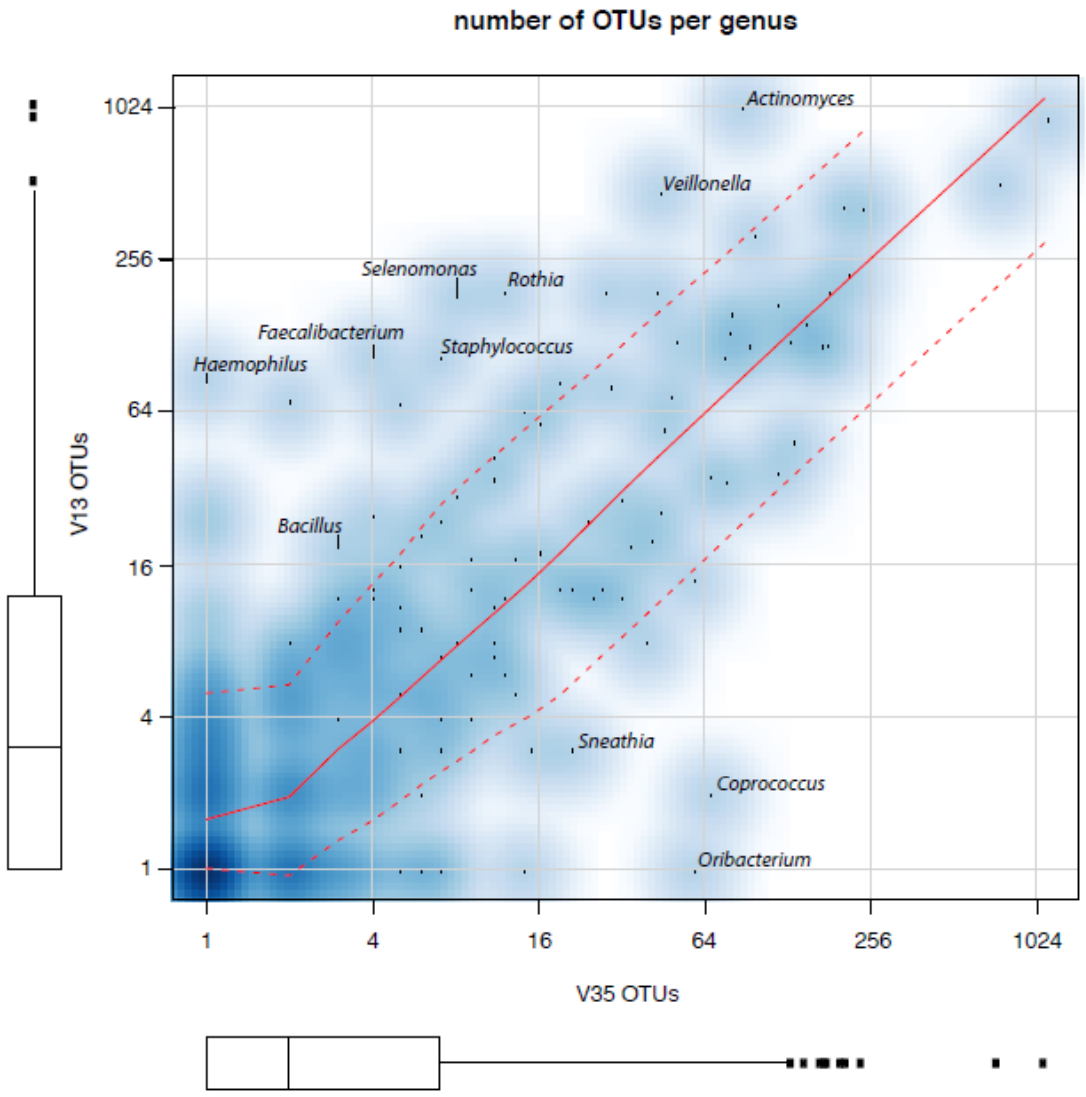


Supplementary Figure 1. Protocols, quality control, data processing, analyses, and resources of the Human Microbiome Project. The HMP Data Analysis Working Group (DAWG) consisted of a multi-institutional working group comprised of genome sequencing centers, several independent research groups, clinical collection sites as well as data management and coordination sites. Samples were collected from participants at

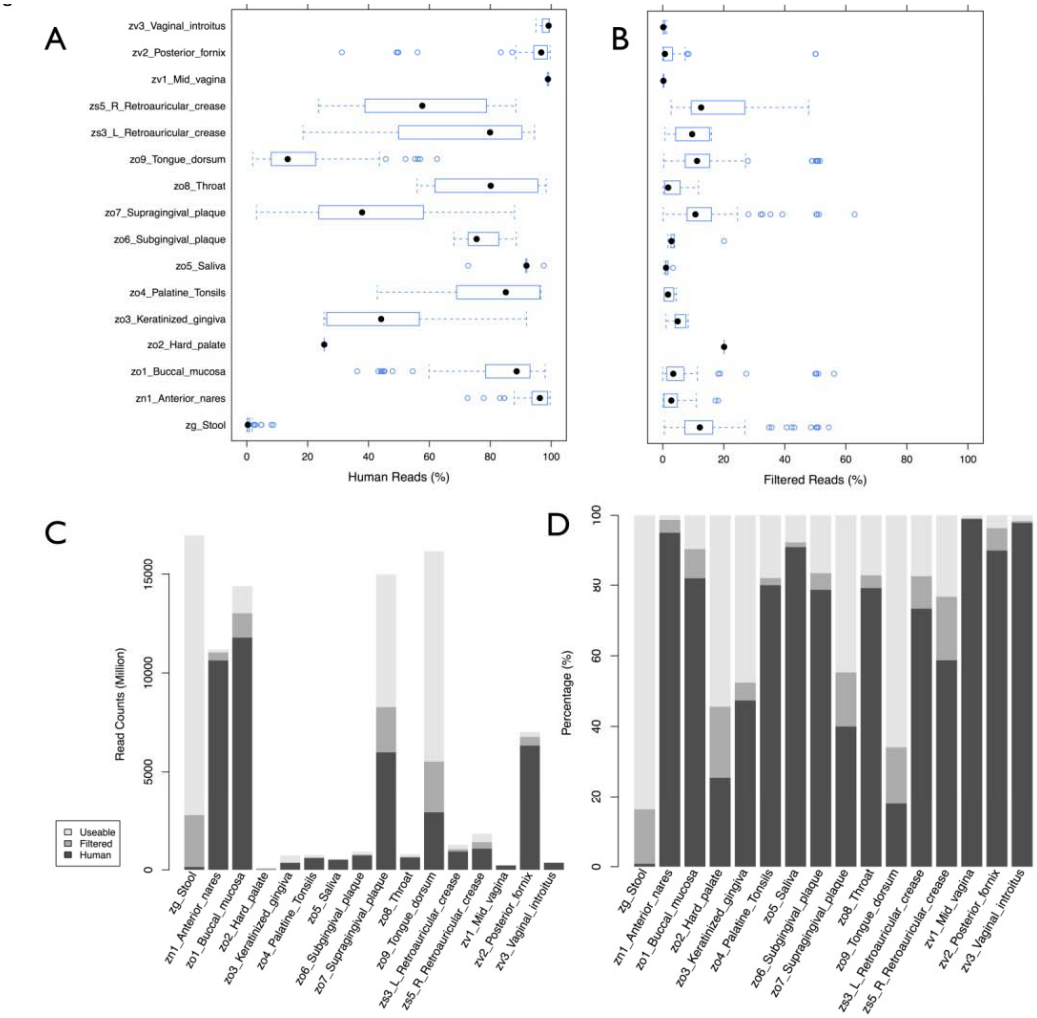
two clinical recruitment centers located at the Baylor College of Medicine (BCM) (Houston) and Washington University School of Medicine. Nucleic acids were extracted from collected specimens and sequencing operations were performed at the following institutions: BCM Human Genome Sequencing Center, Broad Institute of MIT and Harvard, J. Craig Venter Institute and The Genome Institute at Washington University School of Medicine. Tracking of biological samples and participant phenotype was performed by the EMMES Corporation. The Biodefense and Emerging Infections Research Resources Repository served as the repository for all HMP bacterial reference strains. The National Center for Biotechnology Information houses the resulting sequence information at the Sequence Read Archive and BioProject page, while participant phenotype information is housed at dbGaP. All remaining operations were performed by sub-groups of the DAWG listed in the HMP consortium membership of this publication. For each process shown in the diagram blue rectangles represent data sets that are publically available. Processes performed by each working group are depicted with colored diamonds, with colors corresponding to the following HMP working groups: strains (teal), annotation (purple), data processing (orange), WGS assembly (green) and 16S rDNA processing and analysis (yellow). All processes descriptions and datasets are hotlinked and available for download at http://hmpdacc.org/. Publically available data sets and resources are summarized in Supplementary Tables 1-3.



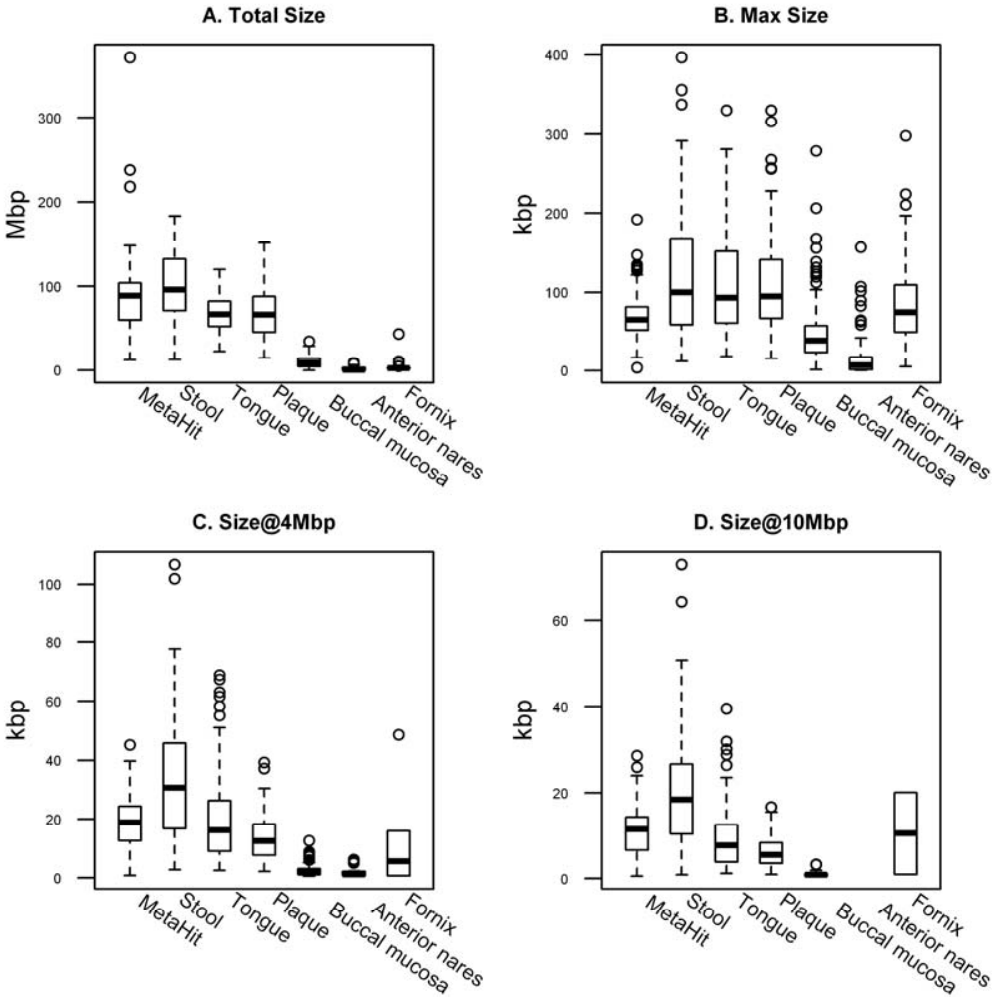
Supplementary Figure 2. Top 10 genera by 16S. For each of the two 16S windows, the 10 most frequent genera across all 18 body sites and samples are shown. The top four are identical for both windows, an additional four genera are shared by both windows although in a different order, and each window has two genera in their respective top 10 abundances not present in the other window.



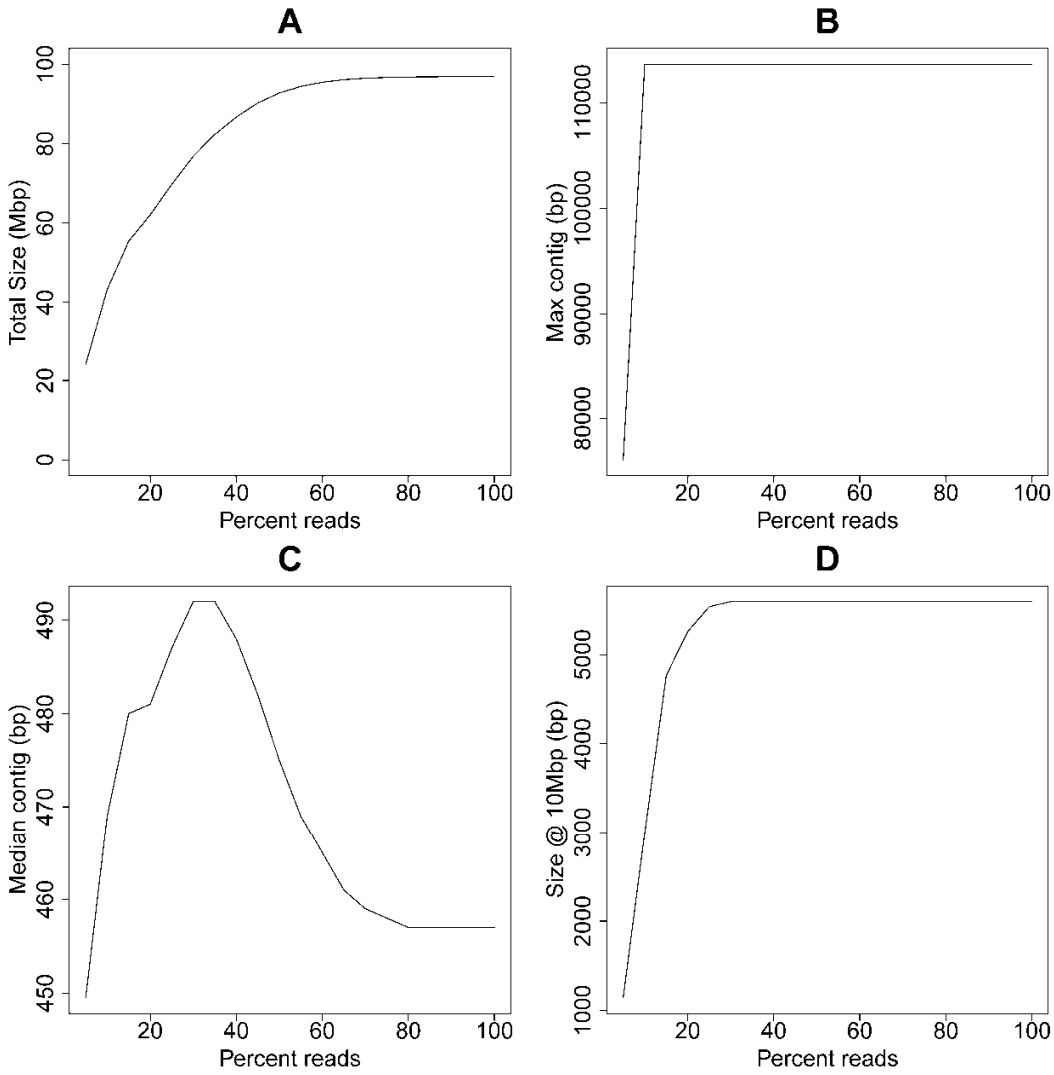
Supplementary Figure 3. Comparison of the number of OTUs per genus between 16S windows. For each of the genera found in samples for which data from both 16S windows (V13 and V35) was available, the number of OTUs within that genus were compared between V13 and V35 (log base$_2$ is plotted). The majority of the data (density representation of data is indicated in a gradient of blue) falls at low OTU/genus level. A number of selected exceptions to this observation are indicated on the figure. The box plots represent the distribution within each of the windows individually. A LOESS curve was fitted to the data (red), and the root-mean-square positive and negative residuals from the LOESS curve are shown (dotted red lines).



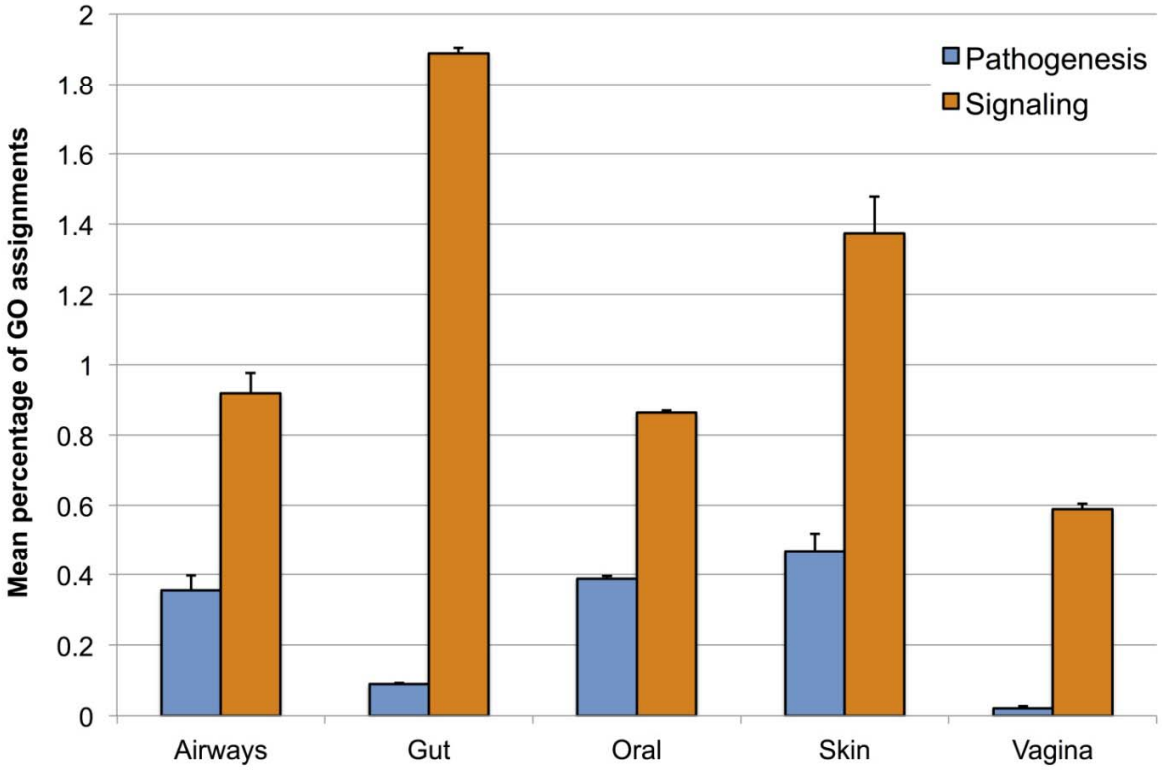
Supplementary Figure 4. Impact of quality and human filtering of the shotgun metagenomic dataset. Thorough quality filtering and removal of reads resulting from human DNA contamination was performed on all shotgun metagenomic data. The variation in fraction of reads per sample removed across the 18 body sites is shown by boxplots for % human (A) and quality filtered reads (B). A similar amount of raw data was generated for each of the selected samples across different sample types (i.e. two lanes of Illumina GAIIx were generated per sample, on average 13Gb/sample). However, as shown in panel C, the total amount of usable data (white) per site varied significantly due to (i) the different number of samples per site, (ii) the differential impact of human contamination (dark grey), and (iii) the differential impact of quality filtering (light grey). Panel D provides a summary view of the fractions usable, versus human and quality filtered data, per body site.



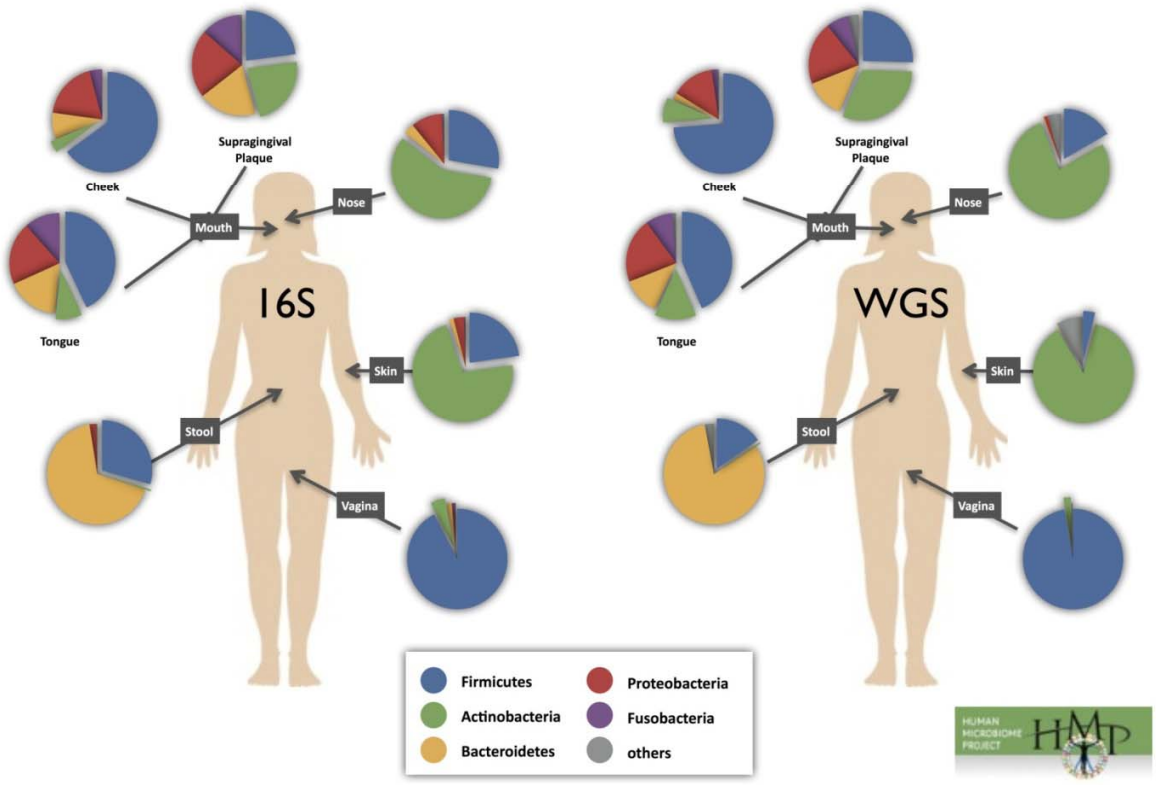
SupplementaryFigure 5. Assembly sizes across body sites. Both MetaHIT and HMP shotgun genomics datasets were assembled using SOAPdenovo with similar parameters (see Supplemental Methods), allowing a detailed comparison of contig size distributions. (A) Cumulative size of assembled contigs longer than 300 bp (total microbiome size); (B) Size of largest contig in the assembly; (C) Size of the smallest contig c, such that the cumulative length of contigs longer than c exceeds 4 Mb (contiguity of the top 4Mb in the assembly); (D) Same as (C) but for contigs adding up to 10 Mb.



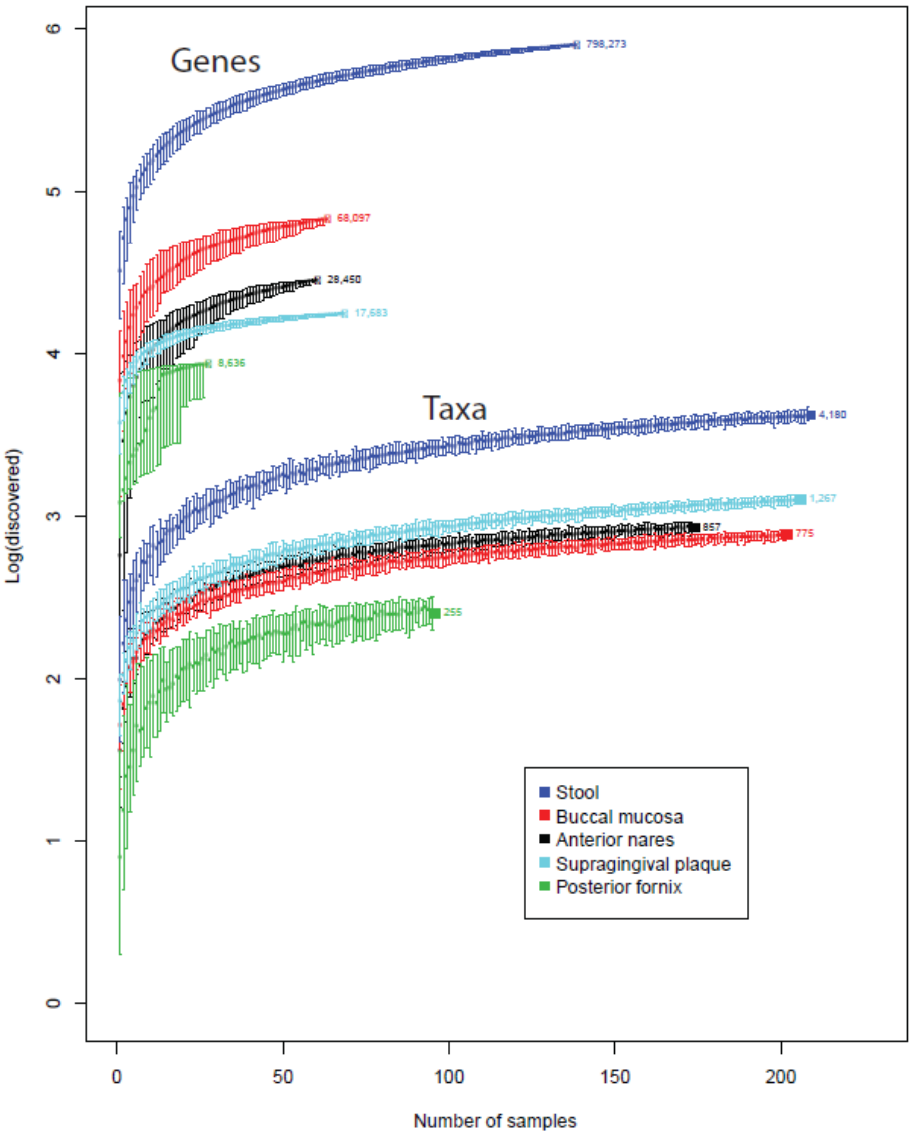
Supplementary Figure 6. Assembly collectors curves. The quality of the assembly of a stool sample (SRS04817) changes as the amount of sequencing increases. The results indicate that we reached a plateau in terms of being able to improve the assembly by increasing coverage. Coverage information (x axis) is represented as a fraction of the total number of reads generated for this sample. **A.** The total size of the assembly (sum of contig sizes) does not change significantly beyond ~75% coverage. **B.** The maximum contig size reaches a plateau at ~40% coverage. **C.** The median contig size peaks at ~40% coverage (where maximal contiguity is achieved), drops as the number (but not size) of contigs grows as the coverage increases, then stabilizes at ~80% coverage when the majority of additional reads are added to existing contigs rather than growing the number or size of contigs produced (also reflected in the plateau in size of the assembly seen in panel A.). **D.** The contiguity within the top 10 Mbp of the assembly (size *c* such that all contigs with size > *c* add up to more than 10 Mbp) does not improve beyond ~30% coverage.



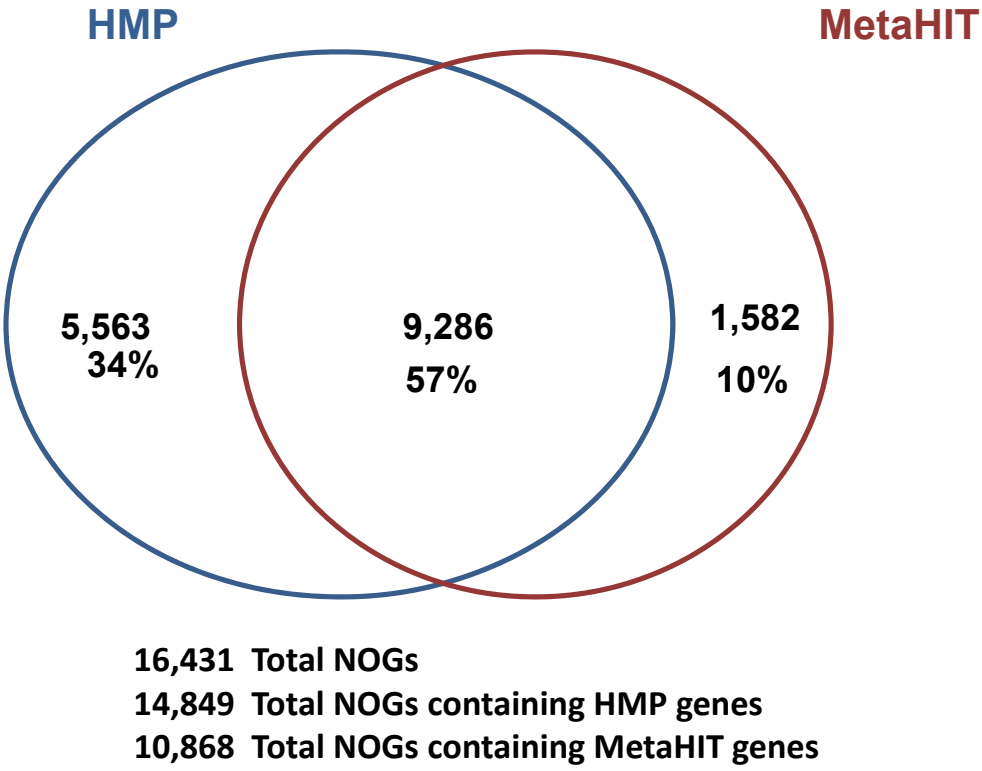
Supplementary Figure 7. Gene index GO slim analysis. An example of the use of the GO slim controlled vocabulary with HMP data is presented here. In this example, the mean percentage of GO assignments for each main body region (with standard error measure) to the GO Slim terms “signaling” and “pathogenesis” are presented. The abundances of the GO term “signaling” is highest in the skin and gut. Non significant differences (P-values >0.05) for the GO term "pathogenesis“ were determined for the comparisons between airways, oral, and skin. All other pairwise comparisons were significant (P values <0.0002). For the GO term "signaling“ the airways and oral are not significantly different from each other. For all other pairwise comparisons P-values are significant (P <0.0004).



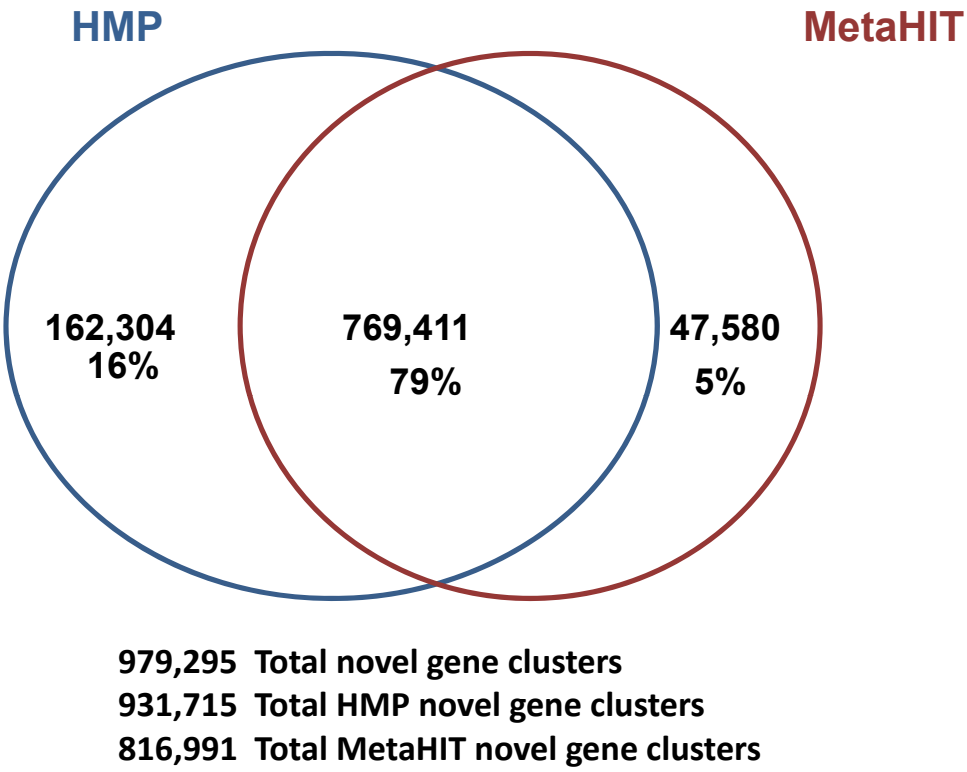
Supplementary Figure 8. Phylum abundances per body site. For each of the body sites studied by both 16S rRNA gene sequencing (A) and whole-genome shotgun sequencing (B) the five most abundant phyla are shown. The small remaining fraction of the data is collapsed and labeled as other phyla (grey).



Supplementary Figure 9. Rates of gene and OTU discovery from five body sites using HMP taxonomic and metagenomic data. For improved ease of viewing, the accumulation curves for genes ('Genes' top) and OTUs ('Taxa' bottom) for five applicable body sites sampled by the HMP (and shown in Fig. 1a and 1b) are compared in this figure. The values given for each curve are final median bootstrap values.



Supplementary Figure 10. Comparison of HMP and MetaHIT orthologous gene clusters from stool. The Venn diagram shows the results of grouping the combined HMP and MetaHIT non-redundant stool gene set by matches to Non-supervised Orthologous Groups (NOGs) in the eggNOG database. Approximately 34% (1,738,698) of genes from HMP data and 35% (1,155,745) of genes from the MetaHIT data were placed into a total of 16,431 NOGs. From this total, 14,849 NOGs contain at least one gene from HMP data and 10,868 NOGs contain at least one gene from MetaHIT data. The intersection reveals a total of 9,286 (~57%) of NOGs that possess representation from both HMP and MetaHIT data. A total of 5,563 (~34%) and 1,582 (~10%) of NOGs are unique to either HMP or MetaHIT data, respectively.



Supplementary Figure 11. Comparison of HMP and MetaHIT novel gene clusters from stool. The Venn diagram shows the results of clustering the combined HMP and MetaHIT non-redundant stool gene set using USEARCH at 80% protein identity after removal of genes which received orthologous group assignments. This resulted in a total of 979,295 novel gene clusters (cluster size ≥2 genes). From this total, 931,715 clusters contain at least one gene from HMP data and 816,991 clusters contain at least one gene from MetaHIT data. The intersection reveals a total of 769,411 gene clusters (~79%) that possess representation from both HMP and MetaHIT data,. A total of 162,304 (~16%) and 47,580 (~5%) novel gene clusters are unique to either HMP or MetaHIT data, respectively.