doi:10.1038/nature11234

Contents

Supplementary Methods

Subject recruitment and microbiome sampling

Subjects were recruited and sampled one to three times at 15 (male) or 18 (female) body habitats as detailed elsewhere[32]. Briefly, 300 subjects between the ages of 18 and 40 years passed a screening for systemic health based on oral, cutaneous, and body mass exclusion criteria (see http://hmpdacc.org/micro_analysis/microbiome_sampling.php). An age range of 18 – 40 years was employed to minimize variability due to growth, development, and aging, resulting in a mean age of 27 sd. 5 and mean body mass index (BMI) of 24 sd. 4. Enrollments were approved by the Institutional Review Boards of the two recruitment centers (Baylor College of Medicine, Houston, TX and Washington University, St. Louis, MO), and a common sampling protocol (see http://hmpdacc.org/doc/HMP_Clinical_Protocol.pdf) was employed for nine oral samples (saliva, swabs from the buccal mucosa, tongue, keratinized gingiva, hard palate, tonsils, and throat, and sub- and supragingival plaque scraping). Four skin specimens were collected from the two (left and right) retroauricular creases and the two antecubital fossae. One specimen from the nares was collected by swabbing each anterior naris and pooling for extraction. One stool specimen was self-collected by participants to represent the microbiota of the lower gastrointestinal tract. Finally, three vaginal specimens were collected from the vaginal introitus, midpoint, and posterior fornix. In order to evaluate within-subject stability of the microbiome, 130 individuals included in these data were sampled at an additional time point (mean 219 sd. 69 days after first sampling, range 35-404 days). Subject phenotypic metadata was collected and coded by the EMMES corporation and is released as dbGaP accession phs000228. Genomic DNA from all samples was isolated using the MoBio PowerSoil DNA Isolation Kit (www.mobio.com) as detailed elsewhere[32]. After quality control, these specimens were used for 16S rRNA gene analysis via 454 pyrosequencing (mean 6,212 sd. 4,303 filtered sequences/sample of length 448 sd. 99, http://hmpdacc.org/doc/HMP_MDG_454_16S_Protocol_V4_2_102109.pdf), focusing on the V3-5 variable region as amplified by the 357F/926R barcoded primers (http://www.hmpdacc.org/HM16S). To assess function, 749 samples were sequenced using 101bp paired-end Illumina shotgun metagenomic reads (mean 29M sd. 21M sequences/sample, http://hmpdacc.org/doc/sops_2/manual_of_procedures_v11.pdf), which were extensively quality controlled prior to this analysis to generate a final metagenomic sample set[33].

16S data processing

For subsequent processing of 16S data, these files (sff/metadata) were demultiplexed using QIIME[34] and custom Python scripts (http://hmpdacc.org/tools_protocols/tools_protocols.php). Quality filtering included rejecting reads <200nt and >1000nt, excluding homopolymer runs >6nt, accepting <=1.5 barcode correction, 0 primer mismatches and 0 ambiguous bases, and requiring a minimum average quality of 25, parallelling criteria developed by Schloss et al[35]. OTU picking was performed for V3-5 region sequences using OTUPipe, an analysis pipeline based on UCHIME[36] for error correction, chimera checking, clustering (via UCLUST), and postprocessing by picking the optimal representative sequence centroid. Clustering both for error correction and for OTU centroid choice was performed at 97%, and reference-based chimera checking was performed against a set of trusted sequences from the "Gold" database (http://microbiomeutil.sourceforge.net). Taxonomy was assigned using the RDP classifier version 2.2[37], re-trained using the latest release (February 4, 2011) of the GreenGenes taxonomy[38].

The resulting OTU tables were checked for mislabeling and contamination as previously described[39]. Briefly, we added a new column indicating major body area (oral, skin including nares, gut, and urogenital). We then rarefied the OTU tables to 100 sequences/sample (as some individual runs of each sample had few sequences but valid data), and dropped OTUs present in only one sample. We then used a random forests classifier to estimate the probability of mislabeling of each sample (i.e. the probability that the sample came from a category other than that assigned). We dropped those samples with mislabeling probability >75% and applied SourceTracker[40] to estimate contamination levels for the remaining samples. Samples with estimated contamination >40% were again dropped and a final OTU table built by merging remaining multiple runs of the same sample. Finally, alpha and beta diversity for each sample/sample pair and Procrustes analysis were established using QIIME with default parameters. When applicable, inverse Simpson's distances were used for alpha diversities and weighted UniFrac for beta diversity comparisons. For comparison of individual subjects, UniFrac distances were used only from first-visit samples. A similar pipeline using mothur[41] was used to classify sequences to phylotypes using the RDP taxonomy as above (see Figure 4B, detailed elsewhere[33]).

WGS data processing

Metagenomic data acquisition and processing are also detailed elsewhere[33] and are summarized here. Libraries were constructed for a subset of available samples (http://hmpdacc.org/tools_protocols/tools_protocols.php) and sequenced targeting ~10Gb of total sequence per sample using 101bp paired-end reads on the Illumina GAIIx platform. Human reads were identified and removed from each sample using 18-mer matches the Best Match Tagger BMTagger (http://cas-bioinfo.cas.unt.edu/sop/mediawiki/index.php/Bmtagger). Each resulting set of sequences was then subjected to three protocols: read-based metabolic reconstruction, assembly-based gene calling, and reference genome mapping.

Duplicate and low-quality reads (as defined by BWA[42] q=2) were removed from each sample, as were low-complexity reads and any resulting sequences of <60nt (http://hmpdacc.org/HMIWGS). For metabolic reconstruction, remaining 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[43], generating abundance and coverage results for each KEGG metabolic pathway and module of interest (http://hmpdacc.org/HMMRC). For gene cataloging, sequences from each sample were assembled using SOAPdenovo[44] v1.04 with parameters -K 25 -R -M 3 -d 1. Contigs were filtered using a custom Perl script fasta2apg.pl (http://hmpdacc.org/tools_protocols/tools_protocols.php) derived from AMOS[45] and scaffolds >300nt were retained. Coding sequences were predicted from the metagenome assemblies using MetaGeneMark[46] and annotated using the JCVI metagenome annotation pipeline[47] and are available in the HMP Gene Index (http://hmpdacc.org/HMGI). A non-redundant gene catalog was made from the Gene Index using USEARCH[48] with 95% identity and 90% coverage thresholds (http://hmpdacc.org/HMGC).

Microbial species abundance estimates

Species-level taxonomic abundances were inferred for all samples using MetaPhlAn 1.1[49] (http://hmpdacc.org/HMSMCP). Briefly, Starting from all 2,887 genomes available from IMG[50] as of November 2011, we systematically identified the 400,141 genes most representative of each taxonomic unit due of their intra-clade universality and inter-clade uniqueness, resulting in a marker catalog spanning 1,221 species. Metagenomic reads were searched against this unique sequence database and the relative abundance of taxa in each sample estimated using the 10% truncated (robust) mean copy number and length-normalized marker relative abundance per species.

Read mapping to microbial reference genomes

Filtered reads were aligned and assigned to reference genomes using BWA[42] version 5.9 with parameters -n 20 -t 4. The genome set consisted of 649 bacterial genomes selected nonredundantly as previously described[51]. Aligned reads were filtered at 80% identity to the best reference sequence, with an alignment length of least 80% of the read length. Normalized read counts for each feature (either gene or kb window), referred to as Reads Per Kilobases per Million mappable reads (RPKM), were calculated by 1000*(read counts/feature length)*(1e6/total mappable reads).

Microbial co-occurrence and co-exclusion network

Networks (Supp. Fig. 4) were constructed using a modification of previously described methods[52]. Briefly, 16S phylotype relative abundances from http://hmpdacc.org/HMMCP at all taxonomic levels were separated by body site and associated with each other within each clinical recruitment center independently using 1) an ensemble of four measures (Pearson correlation, Spearman correlation, Bray-Curtis dissimilarity, and Kullback-Leibler divergence) and 2) generalized boosted linear models of the form:

$$x_{tt,ts} = \mu_{tt,ts} + \Sigma_{st}\beta_{tt,ts,st,ss}x_{st,ss} + \varepsilon$$

That is, the relative abundance of each target taxon *tt* in a target body site *ts* was predicted as a linear function of its mean and all source taxa *st* in each source site *ss* (*ss* possibly equal to *ts*, in which case *st* excluded parent clades of *tt*). These models were sparsely feature selected using the gbm method in R and remaining features fit as adjusted R_2 scores using lm. p-values from all five total methods were computed to account for compositionality by bootstrapping the true values' confidence interval followed by randomly permuting the abundance matrix, renormalizing each sample's relative abundances, and comparing the resulting compositionality-corrected null distribution to the bootstrap using Welch's t-test. This process resulted in ten networks of inter-clade/body-site p-values, one per method per clinical center, which were combined to a final network using Fisher's method followed by False Discovery Rate q-value adjustment. Edges with $q<0.01$, consistent directionality of sign, and supported by at least two methods were retained in Supp. Fig. 4.

Gene and taxon saturation analysis

The number of reads per subject was normalized (1,000 reads per sample for 16S rRNA data) to ensure all subjects were sampled at the same depth of coverage. Samples were randomly selected from n=1 to the maximum number for each body site and the number of unique taxa or functions at a given level determined (OTU/genus/etc. through phylum, or fourth through second level Enzyme Class). Plots showing the difference in diversity with each additional subject were calculated by subtracting the diversity observed with (n-1) subjects from the diversity observed with n subjects. The sampling procedure was repeated 100 times to allow estimates of the variability: standard deviation errors of the mean were plotted. Final % saturations were then calculated by, without rarefaction, dividing the number of taxa/functions observed in all samples by the QIIME ACE estimation of total features.

Metadata association

Metadata were obtained from dbGaP accession phs000228.v3.p1 and tabulated using v2.5 of the processing script at http://huttenhower.sph.harvard.edu/webfm_send/39. Ethnicity was recoded as a categorical variable, missing values removed, and the resulting tables merged with the OTUs and metabolic reconstructions described above. All counts were converted to relative abundances, and relative abundances for each clade in the RDP taxonomy were inferred by summation of children. Subsequently, in addition to the whole-sample QC described above and elsewhere[33], samples were discarded when the abundance of the most frequent genome or module fell below 1% of its maximum value in order to remove samples with sequence coverage below threshold for reliable association detection. Left/right information for bilateral body habitats (antecubital fossa and retroauricular crease) was subsequently discarded before further processing and these locations processed as a single (not dual) habitat.

For each body habitat, 15 linear models of the form:

$$feature = \beta_0 + \Sigma\beta_i x_i + \beta_j x_j$$

were assessed, where *feature* included each reconstructed pathway (metagenomic data) or clade (16S genera or metagenomically identified species); potential confounds *i* comprised sample DNA concentration, sequencing center, total quality bases sequenced, clinical center, percent of detected human reads, and study day processed; and the metadata *j* included subject age, subject diet, BMI, pulse, introitus pH, posterior fornix pH, temperature, gender, ethnicity, breastfeeding, subjects having given birth, diastolic and systolic blood pressure, weight, or height. Feature relative abundances were arcsin square root transformed, missing values imputed by the gam package, the significance of each association parameter β was assessed using R, factors with multiple levels Bonferroni corrected, and, if below a globally FDR-corrected 0.2 threshold, included in Supplemental Table 3.

Supplementary Discussion

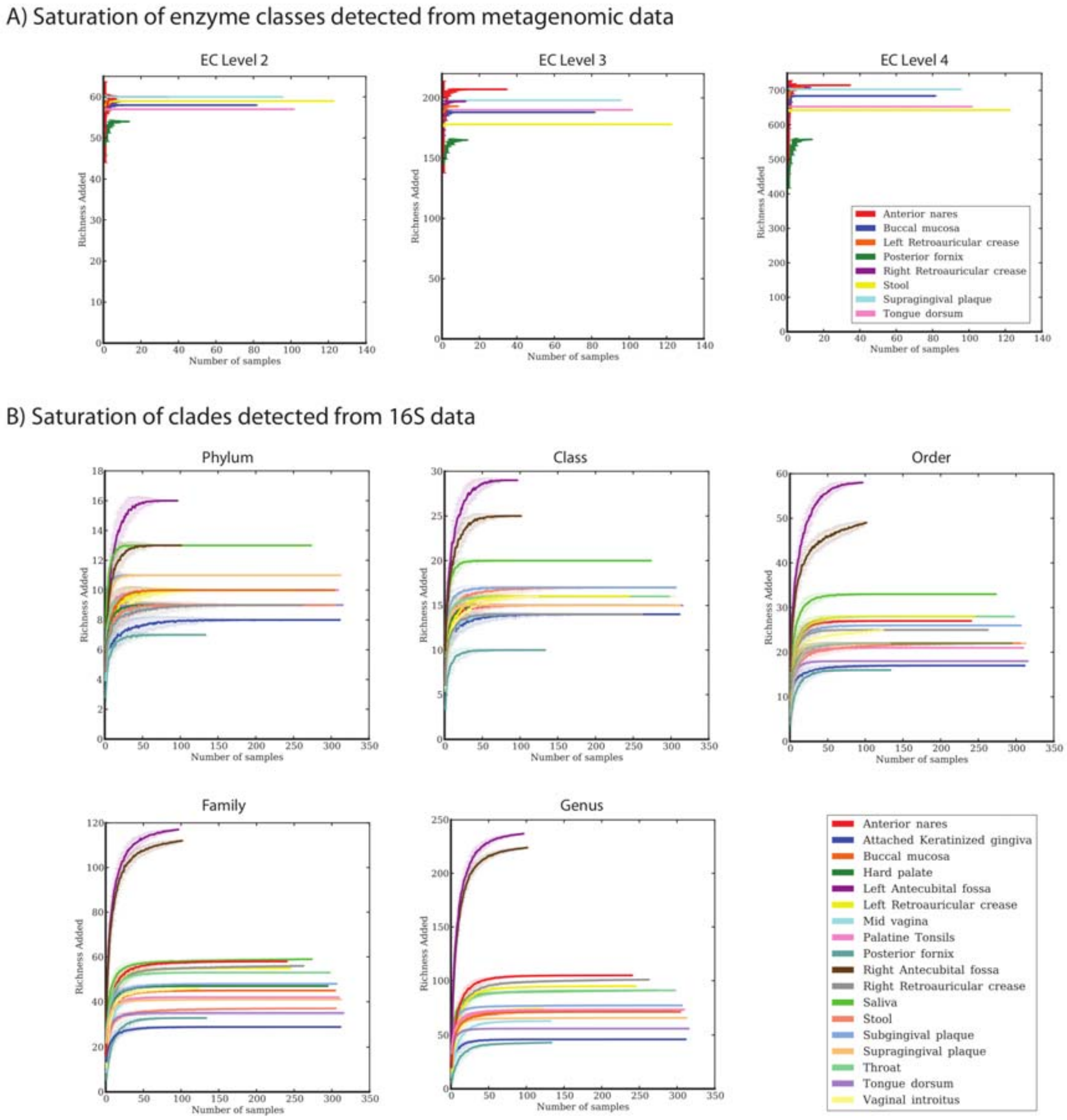
Relationship between microbial carriage and metagenomic function in the nares

The nares harbored a particularly mixed community that varied substantially among individuals, with *Corynebacterium* (mean 31% sd. 21%), *Propionibacterium* (24% sd. 17%), and *Staphylococcus* (17% sd. 18%) all dominating (>50%) subsets of samples. This provided a unique opportunity to examine the metabolic processes uniquely enriched in these three alternative community states (*Corynebacterium*, *Propionibacterium*, or *Staphylococcus* dominant). Metagenomically encoded enzymes and pathways correlated with these single-genus-dominated nares communities fell into three classes. The first type were correlated with known physiological features of the dominant genus, e.g. corynomycolic acids in *Corynebacterium* (positively correlated gene family K12437, pks13[53-54]) or the *Staphylococcus* requirement for nicotinic acid or nicotinamide (negatively correlated metabolic module M00115, NAD biosynthesis[55]). The second type of positive association with dominant clades included pathways not necessarily encoded by those clades, but which provide them with a selective advantage when present in the community. This might be the case for the *Propionibacterium*-associated genes involved in biosynthesis of cobalamin (vitamin B12), a cobalt-containing compound, which included the cob regulon (K03394, K00595, K05936) and the cobalt transport module (M00245[56]). The third class of associations pointed to functional specialization of subsets of species within these signature genera; for example, the negative association of fatty acid biosynthesis/elongation (M00083) with *Corynebacterium*-dominated communities is consistent with the majority of *Corynebacterium* in the nares being “lipophilic” (i.e. lipid requiring) species[57-60].

Subdominance patterns suggest competition among OTUs within genera

Also supporting a high degree of within-community niche specialization, all signature genera consistently colonized their target habitats in dichotomous patterns when binned into specific OTUs. Each abundant genus was represented by a small number of majority OTUs (1-4 each at ≥5% mean abundance in all five body areas) accompanied by several highly prevalent but low-abundance OTUs within the same genus. Since neither strain-level identification nor genome-wide protein function annotations can yet be assigned to 16S-based OTUs, possible hypotheses include, among others, widespread niche specialization at the sub-genus level (since these OTUs regularly co-occur within the same samples at low abundance[61]) or biogeographic localization of individual OTUs within samples. This is the case in the vaginal communities, for example, where different *Lactobacillus* reference genomes encode distinct mechanisms for metabolism of glycogen to lactic acid or maintenance of community structure[62-63]. In the oral cavity, signature *Streptococcus* spp. function variously as metabolizers of simple sugars to lactate (supporting energy production by anaerobes such as *Veillonella*[64]) or as biofilm first colonizers (providing structure to co-aggregators such as *Fusobacterium*[65]). Sequenced reference genomes from the skin and nares signature genera (*Propionibacterium*, *Corynebacterium*, and *Staphylococcus*) indicate that all three rely specifically on degradation of host products in their relatively nutrient-limited environment, by way of lipases, proteases, and lyases[55-56,66]. Coupled with our observed patterns of OTU carriage, these examples might suggest functional competition among dominant organisms and complementarity among subdominant taxa, which will require targeted studies of specific community function to validate.

Supplementary Figures and Legends

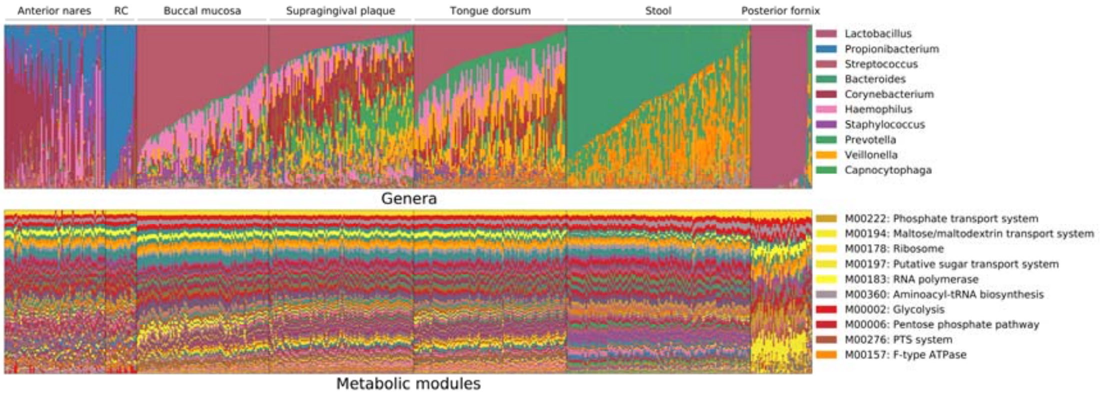


A) Saturation of enzyme classes detected from metagenomic data

B) Saturation of clades detected from 16S data

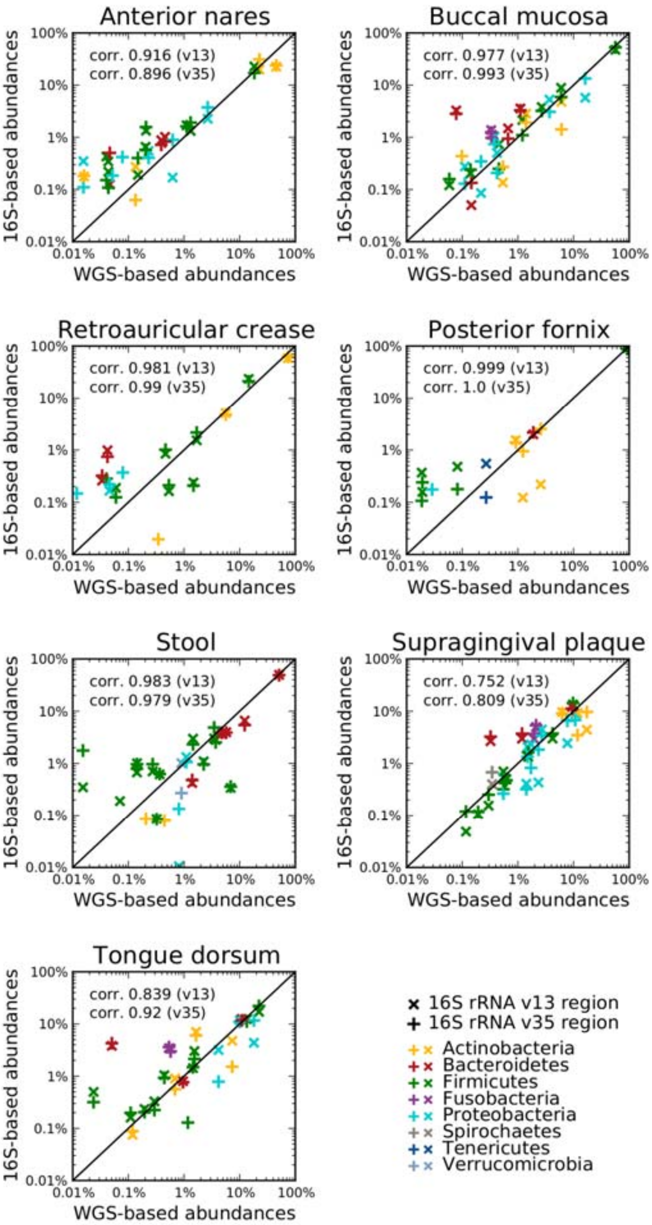
Supplementary Figure 1: Saturation of clade and enzyme class diversity in the HMP cohort.

Rarefaction curves of enzyme classes from the fourth to second levels in HMP metagenomes (A) and of clades from the genus to phylum levels in HMP 16S samples (B), inter-quartile range over 100 samples. See the HMP framework manuscript[33] for OTU and gene family rarefactions. At all non-leaf clade levels, both taxonomically and functionally, the healthy Western human mcirobiome has been well-explored by the depth of sequencing and population size employed by the Human Microbiome Project.



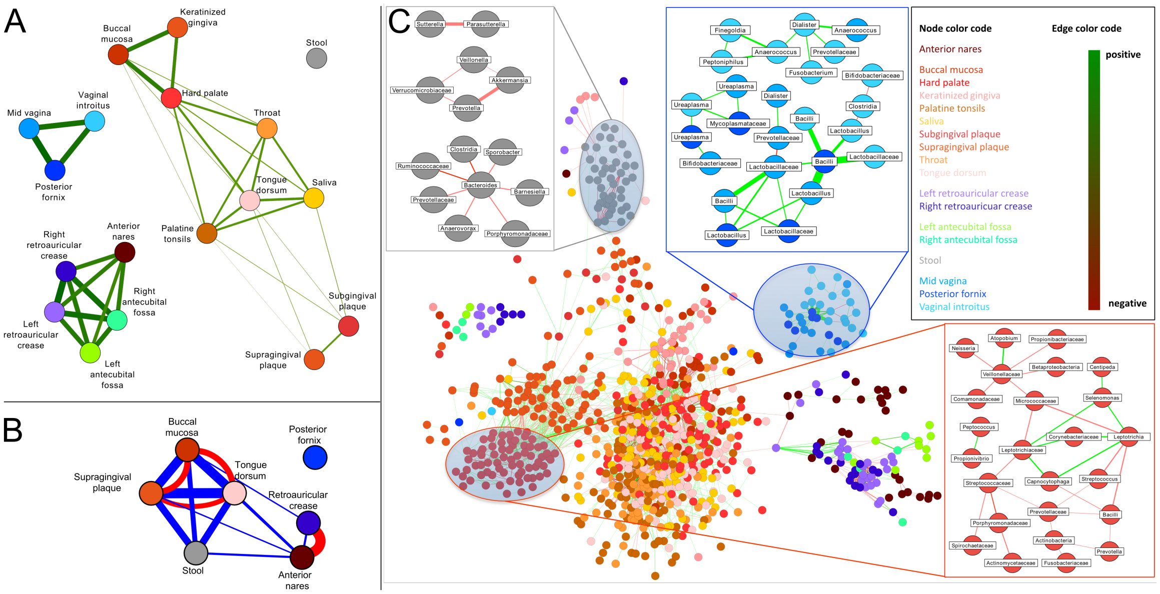
Supplementary Figure 2: Carriage of microbial genera varies while metabolic pathways remain stable within a healthy population.

Vertical bars represent microbiome samples by body habitat in the seven locations with both shotgun and 16S data; bars indicate relative abundances colored by A) microbial genera from binned OTUs and B) metabolic modules. Legend indicates most abundant genera/pathways by average within one or more body habitats. A plurality of most communities' memberships consists of a single dominant genus, but this is universal neither to all body habitats nor to all individuals. Conversely, most metabolic pathways are evenly distributed and prevalent across both individuals and body habitats.



Supplementary Figure 3: 16S and metagenomic taxonomic profiles of the human microbiome provide consistent genus level estimations of bacterial abundance.

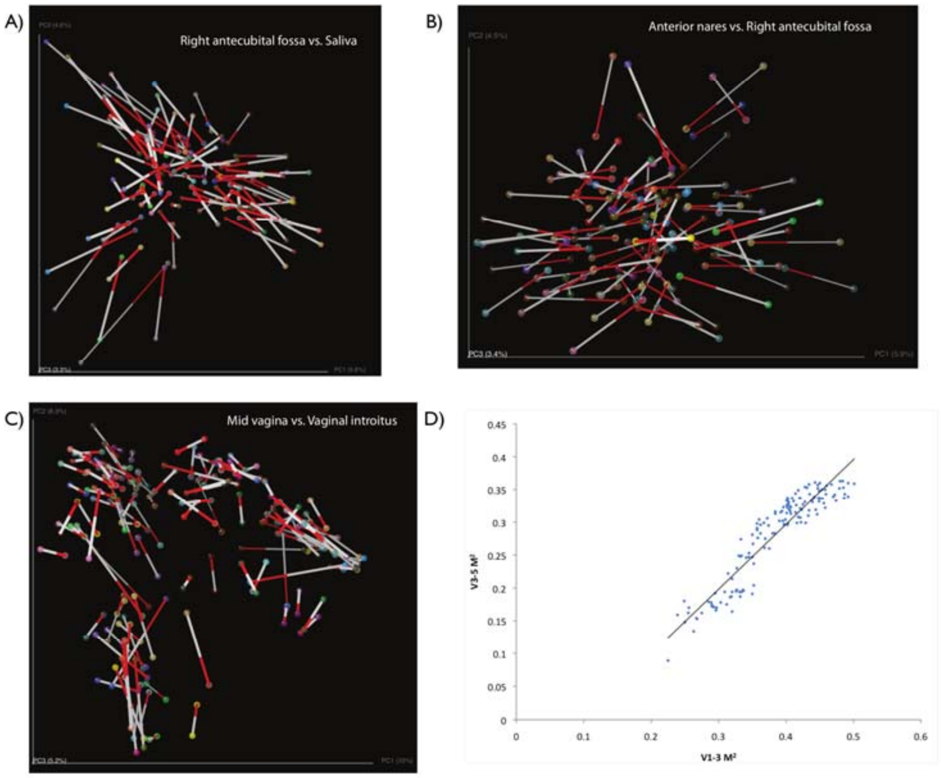
Plot includes each genus' average abundance estimated from 16S rRNA gene pyrosequencing and from shotgun metagenomic data over all body sites for which both data were available. Spearman correlation between the two approaches was >0.97 in four of the seven habitats (for both the V13 and v35 16S regions) and slightly smaller in anterior nares (0.916/0.896 in V13/V35), supragingival plaque (0.752/0.809), and tongue dorsum (0.839/0.920). The absence of a single highly dominant genus in these habitats induces more variability in the estimations. Aside from somewhat higher sensitivity due to much greater sequencing depth, no remarkable general biases were observed for any phyla in these data.



Supplementary Figure 4: Body habitat is a strong determinant of microbial co-occurrence and co-exclusion patterns.

A-B) Body habitats (nodes) connected by average similarity, A) 16S samples by weighted UniFrac beta diversity and B) metagenomic samples by Spearman correlation of metabolic pathways and Bray-Curtis similarity of reference genome abundances. Both views summarize three major clusters within the oral cavity[67], the distinctiveness of the gut and similarity of the vaginal communities, and the grouping of the anterior nares with skin habitats; habitats are again more similar functionally than phylogenetically. C) Predicted bacterial relationships within and across body habitats. Taxa per body habitat are connected by significant positive (green) or negative (red) co-occurrence relationships. Callouts include genera, families, and classes only and show Bacteroidetes/Firmicutes interactions in the gut (grey), negative relationships between early (e.g. *Streptococcus*) and late (e.g. Porphyromonadaceae) colonizers in the dental plaque (red), and the tightly associated *Lactobacilli* in most vaginal sites contrasted with the alternate state of diverse Actinobacteria, Tenericutes, Bacteroidetes, and other Firmicutes (blue).

Supplemental Figure 2



Supplementary Figure 5: Microbial community structure co-variation within subject across body sites.

Most pairs of body sites within individuals show significantly correlated covariance by Procrustes analysis, with p-values assigned by 1000-fold permutation testing. Significant site pairs include A) less related sites such as skin and saliva, and also B) distal skin sites (nares and fossae), and C) highly related vaginal sites. D) These associations are robust to analyses performed on OTUs derived from either the V1-3 or V3-5 variable regions, with site similarity highly correlated between both regions.

Body habitat	Genera	OTUs	ECs	Genes
Antecubital fossa	0.993479	0.780187		
Anterior nares	0.967648	0.726138	0.95098	0.455495
Buccal mucosa	0.899372	0.867807	0.94761	0.500891
Hard palate	0.911657	0.85866		
Keratinized gingiva	0.833763	0.851855		
Mid vagina	0.869688	0.825001		
Palatine tonsils	0.903041	0.900424		
Posterior fornix	0.812009	0.847724	0.93188	0.495431
Retroauricular crease	0.962619	0.740688	0.95557	
Saliva	0.9177	0.888413		
Stool	0.92208	0.969282	0.96416	0.402833
Subgingival plaque	0.910602	0.87107		
Supragingival plaque	0.845638	0.890782	0.95843	0.824122
Throat	0.932003	0.871024		
Tongue dorsum	0.913735	0.875002	0.95844	1
Vaginal introitus	0.884948	0.792583		

Supplementary Table 1: Approximate saturation percentages of clades and gene families cataloged by the Human Microbiome Project.

Values indicate the number of observed features (genera and OTUs from 16S V35 sequences, Enzyme Classes and gene families from metagenomic contig annotations) divided by the QIIME ACE estimator run on all samples available for each data type. While both gene families and OTUs were well-explored by the sequencing depth and cohort size of the HMP, substantial fractions of taxa likely remain to be explored, particularly in populations outside of this healthy Western sample. However, the general phylogenetic space of genera and functional enzyme classes have been well-saturated by this sample.

References

32 Aagaard, K. *et al.* A Comprehensive Strategy for Sampling the Human Microbiome. (in review).
33 The Human Microbiome Project Consortium. A framework for human microbiome research. (in review).
34 Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* **7**, 335-336, doi:nmeth.f.303 [pii]
10.1038/nmeth.f.303 (2010).
35 Schloss, P. D., Gevers, D. & Westcott, S. L. Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PLoS One* **6**, e27310, doi:10.1371/journal.pone.0027310 (2011).
36 Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C. & Knight, R. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27**, 2194-2200, doi:10.1093/bioinformatics/btr381 (2011).
37 Wang, Q., Garrity, G. M., Tiedje, J. M. & Cole, J. R. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* **73**, 5261-5267, doi:AEM.00062-07 [pii]
10.1128/AEM.00062-07 (2007).
38 McDonald, D. *et al.* An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *The ISME journal* **6**, 610-618, doi:10.1038/ismej.2011.139 (2012).
39 Knights, D. *et al.* Supervised classification of microbiota mitigates mislabeling errors. *The ISME journal* **5**, 570-573, doi:10.1038/ismej.2010.148 (2011).
40 Knights, D. *et al.* Bayesian community-wide culture-independent microbial source tracking. *Nat Methods* **8**, 761-763, doi:10.1038/nmeth.1650 (2011).
41 Schloss, P. D. *et al.* Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* **75**, 7537-7541, doi:AEM.01541-09 [pii]
10.1128/AEM.01541-09 (2009).
42 Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589-595, doi:btp698 [pii]
10.1093/bioinformatics/btp698 (2010).
43 Abubucker, S. *et al.* Metabolic reconstruction for metagenomic data and its application to the human microbiome. (in press).
44 Li, R. *et al.* De novo assembly of human genomes with massively parallel short read sequencing. *Genome research* **20**, 265-272, doi:10.1101/gr.097261.109 (2010).
45 Treangen, T. J., Sommer, D. D., Angly, F. E., Koren, S. & Pop, M. Next generation sequence assembly with AMOS. *Curr Protoc Bioinformatics* **Chapter 11**, Unit 11 18, doi:10.1002/0471250953.bi1108s33 (2011).
46 Zhu, W., Lomsadze, A. & Borodovsky, M. Ab initio gene identification in metagenomic sequences. *Nucleic Acids Research* **38**, e132, doi:10.1093/nar/gkq275 (2010).
47 Tanenbaum, D. M. *et al.* The JCVI standard operating procedure for annotating prokaryotic metagenomic shotgun sequencing data. *Standards in Genomic Sciences* **2**, 229-237, doi:10.4056/sigs.651139 (2010).
48 Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460-2461, doi:10.1093/bioinformatics/btq461 (2010).
49 Segata, N. *et al.* Efficient metagenomic microbial community profiling using unique clade-specific marker genes. (in press).
50 Markowitz, V. M. *et al.* The integrated microbial genomes system: an expanding comparative analysis resource. *Nucleic Acids Res* **38**, D382-390, doi:gkp887 [pii]
10.1093/nar/gkp887 (2010).
51 Giannoukos, G. *et al.* Efficient and robust RNA-seq process for cultured bacteria and complex community transcriptomes. *Genome biology* **13**, R23, doi:10.1186/gb-2012-13-3-r23 (2012).
52 Faust, K. *et al.* Microbial co-occurrence relationships in the human microbiome. (in review).
53 Portevin, D. *et al.* A polyketide synthase catalyzes the last condensation step of mycolic acid biosynthesis in mycobacteria and related organisms. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 314-319, doi:10.1073/pnas.0305439101 (2004).
54 Tauch, A. *et al.* Complete genome sequence and analysis of the multiresistant nosocomial pathogen Corynebacterium jeikeium K411, a lipid-requiring bacterium of the human skin flora. *Journal of bacteriology* **187**, 4671-4682, doi:10.1128/JB.187.13.4671-4682.2005 (2005).
55 Kuroda, M. *et al.* Whole genome sequencing of meticillin-resistant Staphylococcus aureus. *Lancet* **357**, 1225-1240 (2001).
56 Bruggemann, H. *et al.* The complete genome sequence of Propionibacterium acnes, a commensal of human skin. *Science* **305**, 671-673, doi:10.1126/science.1100330 (2004).
57 McGinley, K. J. *et al.* Analysis of cellular components, biochemical reactions, and habitat of human cutaneous lipophilic diphtheroids. *J Invest Dermatol* **85**, 374-377 (1985).
58 Wos-Oxley, M. L. *et al.* A poke into the diversity and associations within human anterior nare microbial communities. *The ISME journal* **4**, 839-851, doi:10.1038/ismej.2010.15 (2010).
59 Frank, D. N. *et al.* The human nasal microbiota and Staphylococcus aureus carriage. *PLoS One* **5**, e10598, doi:10.1371/journal.pone.0010598 (2010).
60 Lemon, K. P. *et al.* Comparative analyses of the bacterial microbiota of the human nostril and oropharynx. *MBio* **1**, doi:10.1128/mBio.00129-10 (2010).
61 Huse, S., Ye, Y., Zhou, Y. & Fodor, A. A Core Human Microbiome as Viewed Through 16S rRNA Sequences Clusters. (in press).
62 O'Hanlon, D. E., Moench, T. R. & Cone, R. A. In vaginal fluid, bacteria associated with bacterial vaginosis can be suppressed with lactic acid but not hydrogen peroxide. *BMC Infect Dis* **11**, 200, doi:10.1186/1471-2334-11-200 (2011).

63 Ravel, J. *et al.* Vaginal microbiome of reproductive-age women. *Proc Natl Acad Sci U S A* **108 Suppl 1**, 4680-4687, doi:1002611107 [pii]
10.1073/pnas.1002611107 (2011).

64 Nobbs, A. H., Lamont, R. J. & Jenkinson, H. F. Streptococcus adherence and colonization. *Microbiology and molecular biology reviews : MMBR* **73**, 407-450, Table of Contents, doi:10.1128/MMBR.00014-09 (2009).

65 Merritt, J., Niu, G., Okinaga, T. & Qi, F. Autoaggregation response of Fusobacterium nucleatum. *Appl Environ Microbiol* **75**, 7725-7733, doi:AEM.00916-09 [pii]
10.1128/AEM.00916-09 (2009).

66 Ruiz, J. C. *et al.* Evidence for reductive genome evolution and lateral acquisition of virulence functions in two Corynebacterium pseudotuberculosis strains. *PLoS One* **6**, e18551, doi:10.1371/journal.pone.0018551 (2011).

67 Segata, N. *et al.* Composition of the Adult Digestive Tract Microbiome Based on Seven Mouth Surfaces, Tonsils, Throat and Stool Samples. (in review).