Published in final edited form as:
Nature. ; 486(7402): 207–214. doi:10.1038/nature11234.

Structure, Function and Diversity of the Healthy Human Microbiome

The Human Microbiome Project Consortium

Abstract

Studies of the human microbiome have revealed that even healthy individuals differ remarkably in the microbes that occupy habitats such as the gut, skin, and vagina. Much of this diversity remains unexplained, although diet, environment, host genetics, and early microbial exposure have all been implicated. Accordingly, to characterize the ecology of human-associated microbial communities, the Human Microbiome Project has analyzed the largest cohort and set of distinct, clinically relevant body habitats to date. We found the diversity and abundance of each habitat's signature microbes to vary widely even among healthy subjects, with strong niche specialization both within and among individuals. The project encountered an estimated 81–99% of the genera, enzyme families, and community configurations occupied by the healthy Western microbiome. Metagenomic carriage of metabolic pathways was stable among individuals despite variation in community structure, and ethnic/racial background proved to be one of the strongest associations

Correspondence and requests for materials should be addressed to chuttenh@hsph.harvard.edu.

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Author Contributions

Principal investigators: BWB, RAG, SKH, BAM, KEN, JFP, GMW, OW, RKW Manuscript preparation: DG, CH, RK, OW
Funding agency management: CCB, TB, VB, JLC, SC, CD, VDF, CG, MYG, RDL, JM, PM, JP, LMP, JAS, LW, CW, KAW
Project leadership: SA, JHB, BWB, ATC, HHC, AME, MGF, RSF, DG, MGG, KH, SKH, CH, EAL, RM, VM, JCM, BAM, MM, DMM, KEN, JFP, EJS, JV, GMW, OW, AMW, KCW, JRW, SKY, QZ
Analysis preparation for manuscript: JCC, KF, DG, AG, KHH, CH, RK, DK, HHK, OK, KPL, REL, JR, JFS, PDS, NS
Data release: LA, TB, IAC, KC, HHC, NJD, DJD, AME, VMF, LF, JMG, SG, SKH, MEH, CJ, VJ, CK, AAM, VMM, TM, MM, DMM, JO, KP, JFP, CP, XQ, RKS, NS, IS, EJS, DVW, OW, KW, KCW, CY, BPY, QZ
Methods and research development: SA, HMA, MB, DMC, AME, RLE, MF, SF, MGF, DCF, DG, GG, BJH, SKH, MEH, WAK, NL, KL, VM, ERM, BAM, MM, DMM, CN, JFP, MEP, XQ, MCR, CR, EJS, SMS, DGT, DVW, GMW, YW, KMW, SY, BPY, SKY, QZ
DNA sequence production: SA, EA, TA, TB, CJB, DAB, KDD, SPD, AME, RLE, CNF, SF, CCF, LLF, RSF, BH, SKH, MEH, VJ, CLK, SLL, NL, LL, DMM, IN, CN, MO, JFP, XQ, JGR, YR, MCR, DVW, YW, BPY, YZ
Clinical sample collection: KMA, MAC, WMD, LLF, NG, HAH, ELH, JAK, WAK, TM, ALM, PM, SMP, JFP, GAS, JV, MAW, GMW
Body site experts: KMA, EAV, GA, LB, MJB, CCD, FED, LF, JI, JAK, SKH, HHK, KPL, PJM, JR, TMS, JAS, JDS, JV
Ethical, legal and social implications: RMF, DEH, WAK, NBK, CML, ALM, RR, PS, RRS, PS, LZ
Strain management: EAV, JHB, IAC, KC, SWC, HHC, TZD, ASD, AME, MGF, MGG, SKH, VJ, NCK, SLL, LL, KL, EAL, VMM, BAM, DMM, KEN, IN, IP, LS, EJS, CMT, MT, DVW, GMW, AMW, YW, KMW, BPY, LZ, YZ
16S data analysis: KMA, EJA, GLA, CAA, MB, BWB, JPB, GAB, SRC, SC, JC, TZD, FED, ED, AME, RCE, MF, AAF, JF, HG, DG, BJH, TAH, SMH, CH, JI, JKJ, STK, SKH, RK, HHK, OK, PSLR, REL, KL, CAL, DM, BAM, KAM, MM, MP, JFP, MP, KSP, XQ, KPR, MCR, BR, PDS, TMS, NS, JAS, WDS, TJS, CSS, EJS, RMT, JV, TAV, ZW, DVW, GMW, JRW, KMW, YY, SY, YZ
Shotgun data processing and alignments: CJB, JCC, ED, DG, AG, MEH, HJ, DK, KCK, CLK, YL, JCM, BAM, MM, DMM, JO, JFP, XQ, JGR, RKS, NUS, IS, EJS, GGS, SMS, JW, ZW, GMW, OW, KCW, TW, SKY, LZ
Assembly: HMA, CJB, PSC, LC, YD, SPD, MGF, MEH, HJ, SK, BL, YL, CL, JCM, JMM, JRM, PJM, MM, JFP, MP, MEP, XQ, MR, RKS, MS, DDS, GGS, SMS, CMT, TJT, WW, GMW, KCW, LY, YY, SKY, LZ
Annotation: OOA, VB, CJB, IAC, ATC, KC, HHC, ASD, MGG, JMG, JG, AG, SG, BJH, KH, SKH, CH, HJ, NCK, RM, VMM, KM, TM, MM, JO, KP, MP, XQ, NS, EJS, GGS, SMS, MT, GMW, KCW, JRW, CY, SKY, QZ, LZ
WGS Metabolic Reconstruction: SA, BLC, JG, CH, JI, BAM, MM, BR, AMS, NS, MT, GMW, SY, QZ, JDZ

Author Information

All data used in this study is available from the Human Microbiome Project Data Analysis and Coordination Center[1]. Reprints and permissions information is available at www.nature.com/reprints.

The authors declare no competing financial interests.

of both pathways and microbes with clinical metadata. These results thus delineate the range of structural and functional configurations normal in the microbial communities of a healthy population, enabling future characterization of the epidemiology, ecology, and translational applications of the human microbiome.

A total of 4,788 specimens from 242 screened and phenotyped adults[1–2] (129 males, 113 females) were available for this study, representing the majority of the target HMP cohort of 300 individuals. Adult subjects lacking evidence of disease were recruited based on a lengthy list of exclusion criteria; we will refer to them here as "healthy," as defined by the consortium clinical sampling criteria[2]. Women were sampled at 18 body habitats, men at 15 (excluding three vaginal sites), distributed among five major body areas. Nine specimens were collected from the oral cavity and oropharynx: saliva; buccal mucosa (cheek), keratinized gingiva (gums), palate, tonsils, throat, and tongue soft tissues; and supra- and subgingival dental plaque (tooth biofilm above and below the gum). Four skin specimens were collected from the two retroauricular creases (behind each ear) and the two antecubital fossae (inner elbows), and one specimen for the anterior nares (nostrils). A self-collected stool specimen represented the microbiota of the lower gastrointestinal tract, and three vaginal specimens were collected from the vaginal introitus, midpoint, and posterior fornix. In order to evaluate within-subject stability of the microbiome, 131 individuals in these data were sampled at an additional time point (mean 219 sd. 69 days after first sampling, range 35–404 days). After quality control, these specimens were used for 16S rRNA gene analysis via 454 pyrosequencing (abbreviated henceforth as 16S profiling, mean 5,408 sd. 4,605 filtered sequences/sample); to assess function, 681 samples were sequenced using paired-end Illumina shotgun metagenomic reads (mean 2.9Gb sd. 2.1 per sample)[1]. More details on data generation are provided in related HMP publications[1–2] and in Supplemental Methods.

Microbial diversity of healthy humans

The diversity of microbes within a given body habitat can be defined as the number and abundance distribution of distinct types of organisms, which has been linked to several human diseases: low diversity in the gut to obesity and inflammatory bowel disease[3–4], for example, and high diversity in the vagina to bacterial vaginosis[5]. For this large study involving microbiome samples collected from healthy volunteers at two distinct geographic locations in the United States, we have defined the microbial communities at each body habitat, encountering 81–99% of predicted genera and saturating the range of overall community configurations (Fig. 1, Supp. Fig. 1, Supp. Table 1, see also Fig. 4). Oral and stool communities were especially diverse in terms of community membership, expanding prior observations[6], and vaginal sites harbored particularly simple communities (Fig. 1A). This study established that these patterns of alpha diversity (within samples) differed markedly from comparisons between samples from the same habitat among subjects (beta diversity, Fig. 1B). For example, the saliva had among the highest median alpha diversities of Operational Taxonomic Units (OTUs, roughly species level classification, see http://hmpdacc.org/HMQCP), but one of the lowest beta diversities - so although each individual's saliva was ecologically rich, members of the population shared similar organisms. Conversely, the antecubital fossae (skin) had the highest beta diversity but were intermediate in alpha diversity. The vagina had the lowest alpha diversity, with quite low beta diversity at the genus level but very high among OTUs due to the presence of distinct *Lactobacillus* spp. (Fig. 1B). The primary patterns of variation in community structure followed the major body habitat groups (oral, skin, gut, and vaginal), defining as a result the complete range of population-wide between-subject variation in human microbiome habitats (Fig. 1C). Within-subject variation over time was consistently lower than between-subject variation, both in organismal composition and in metabolic function (Fig. 1D). The uniqueness of each

individual's microbial community thus appear to be stable over time (relative to the population as a whole), which may be another feature of the human microbiome specifically associated with health.

No taxa were observed to be universally present among all body habitats and individuals at the sequencing depth employed here, unlike several pathways (Fig. 2 and Supp. Fig. 2, see below), although several clades demonstrated broad prevalence and relatively abundant carriage patterns[7–9]. Instead, as suggested by individually focused studies[3–4,6,10–11], each body habitat in almost every subject was characterized by one or a few signature taxa making up the plurality of the community (Fig. 3). Signature clades at the genus level formed on average anywhere from 17% to 84% of their respective body habitats, completely absent in some communities (0% at this level of detection) and representing the entire population (100%) in others. Strikingly, less dominant taxa were also highly personalized, both among individuals and body habitats; in the oral cavity, for example, most habitats are dominated by *Streptococcus*, but these are followed in abundance by *Haemophilus* in the buccal mucosa, *Actinomyces* in the supragingival plaque, and *Prevotella* in the immediately adjacent (but low oxygen) subgingival plaque[12].

Additional taxonomic detail of the human microbiome was provided by identifying unique marker sequences in metagenomic data[13] (Fig. 3A) to complement 16S profiling (Fig. 3B). These two profiles were typically in close agreement (Supp. Fig. 3), with the former in some cases offering more specific information on members of signature genera differentially present within habitats (e.g. vaginal *Prevotella amnii* and gut *P. copri*) or among individuals (e.g. vaginal *Lactobacillus* spp.) One application of this specificity was to confirm the absence of NIAID class A–C pathogens above 0.1% abundance (aside from *S. aureus* and *E. coli*) from the healthy microbiome, but the near-ubiquity and broad distribution of opportunistic "pathogens" as defined by PATRIC[14]. Canonical pathogens including *Vibrio cholerae*, *Mycobacterium avium*, *Campylobacter jejuni*, and *Salmonella enterica* were not detected at this level of sensitivity. *Helicobacter pylori* was found in only two gut samples, both at <0.01%, and *E. coli* was present at >0.1% abundance in 15% of stool microbiomes (>0% abundance in 61%). Similar species level observations were obtained for a small subset of stool samples with 454 pyrosequencing metagenomics data using PhylOTU[15–16]. In total 56 of 327 PATRIC "pathogens" were detected in the healthy microbiome (at >1% prevalence of >0.1% abundance, Supp. Table 2), all opportunistic and, strikingly, typically prevalent both among hosts and habitats. The latter is in contrast to many of the most abundant signature taxa, which were usually more habitat-specific and variable among hosts (Fig. 3A–B). This overall absence of particularly detrimental microbes supports the hypothesis that even given this cohort's high diversity, the microbiota tend to occupy a range of configurations in health distinct from many of the disease perturbations studied to date[4,17].

Carriage of specific microbes

Inter-individual variation in the microbiome proved to be specific, functionally relevant, and personalized. One example of this is illustrated by the *Streptococcus* spp. of the oral cavity. The genus dominates the oropharynx[18], with different species abundant within each sampled body habitat (see http://hmpdacc.org/HMSMCP) and, even at the species level, striking differences in carriage within each habitat among individuals (Fig. 4A). As the ratio of pan- to core-genomes is high in many human-associated microbes[19], this variation in abundance could be due to selective pressures acting on pathways differentially present among *Streptococcus* species or strains (Fig. 4B). Indeed, we observed extensive strain-level genomic variation within microbial species in this population, enriched for host-specific structural variants around genomic islands (Fig. 4C). Even with respect to the single

Streptococcus mitis strain B6, gene losses associated with these events were common, for example differentially eliminating *S. mitis* carriage of the V-type ATPase or choline binding proteins cbp6 and cbp12 among subsets of the host population (Fig. 4D). These losses were easily observable by comparison to reference isolate genomes, and these initial findings suggest that microbial strain- and host-specific gene gains and polymorphisms may be similarly ubiquitous.

Other examples of functionally relevant inter-individual variation at the species and strain levels occurred throughout the microbiome. In the gut, *Bacteroides fragilis* has been shown to prime T cell responses in animal models via the capsular polysaccharides PSA[20], and in the HMP stool samples this taxon was carried at a level of at least 0.1% in 16% of samples (over 1% abundance in 3%). *B. thetaiotaomicron* has been studied for its effect on host gastrointestinal metabolism[21] and was likewise common at 46% prevalence. On the skin, *Staphylococcus aureus*, of particular interest as the cause of methicillin-resistant *S. aureus* (MRSA) infections, had 29% nasal and 4% skin carriage rates, roughly as expected[22]. Close phylogenetic relatives such as *S. epidermidis* (itself considered commensal) were, in contrast, universal on the skin and present in 93% of nares samples, and at the opposite extreme *Pseudomonas aeruginosa* (a representative gram negative skin pathogen) was completely absent from both body habitats (0% at this level of detection). These and the data above suggest that the carriage pattern of some species in the human microbiome may be analogous to genetic traits, where recessive alleles of modest risk are maintained in a population. In the case of the human microbiome, high-risk pathogens remain absent, while species that pose a modest degree of risk also appear to be stably maintained in this ecological niche.

Finally, microorganisms within and among body habitats exhibited relationships suggestive of driving physical factors such as oxygen, moisture and pH, host immunological factors, and microbial interactions such as mutualism or competition[23] (Supp. Fig. 4). Both overall community similarity and microbial co-occurrence and co-exclusion across the human microbiome grouped the 18 body habitats together into four clusters corresponding to the five target body areas (Supp. Fig. 4A–B). There was little distinction among different vaginal sites, with *Lactobacillus* spp. dominating all three and correlating in abundance. However, *Lactobacillus* varied inversely with the Actinobacteria and Bacteroidetes (see Supp. Fig. 4C and Fig. 2–3), as also observed in the cohort of Ravel et al[11]. Gut microbiota relationships primarily comprised inverse associations with the *Bacteroides*, which ranged from dominant in some subjects to a minority in others who carried a greater diversity of Firmicutes. A similar progression was evident in the skin communities, dominated by one of *Staphylococcus* (phylum Firmicutes), *Propionibacterium*, or *Corynebacterium* (both phylum Actinobacteria), with a continuum of oral organisms (e.g. *Streptococcus*) appearing in nares communities (Supp. Fig. 4C). These observations suggest that microbial community structure in these individuals may sometimes occupy discrete configurations and under other circumstances vary continuously, a topic addressed in more detail by several HMP investigations[7,24–25]. An individual's location within such configurations is indicative of current microbial carriage (including pathogens) and of the community's ability to resist future pathogen acquisition or dysbiosis; it may thus prove to be associated with disease susceptibility or other phenotypic characteristics.

Microbiome metabolism and function

As the first study to include both marker gene and metagenomic data across body habitats from a large human population, we additionally assessed the ecology of microbial metabolic and functional pathways in these communities. We reconstructed the relative abundances of pathways in community metagenomes[26], which were much more constant and evenly

diverse than were organismal abundances (Fig. 2B, see also Fig. 1), confirming this as an ecological property of the entire human microbiome[3]. We were likewise able to determine for the first time that taxonomic and functional alpha diversity across microbial communities significantly correlate (Spearman of inverse Simpson's r=0.60, p=3.6e-67, n=661), the latter within a more proscribed range of community configurations (Supp. Fig. 5).

Unlike microbial taxa, several pathways were ubiquitous among individuals and body habitats. The most abundant of these "core" pathways include the ribosome and translational machinery, nucleotide charging and ATP synthesis, and glycolysis, and reflect the basics of host-associated microbial life. Also in contrast to taxa, few pathways were highly variable among subjects within any body habitat; exceptions included the Sec (orally, sd. 0.0052, mean oral sd. 0.0011 sd. 0.0016) and Tat (globally, sd 0.0055, mean global sd. 0.0023 sd. 0.0033) secretion systems, indicating a high degree of host-microbe and microbe-microbe interactions in the healthy human microbiota. This high variability was particularly present in the oral cavity, for phosphate, mono- and di-saccharide, and amino acid transport in the mucosa, as well as LPS biosynthesis and spermidine/putrescine synthesis and transport on the plaque and tongue (http://hmpdacc.org/HMMRC). The stability and high metagenomic abundance of this housekeeping "core" contrasts with the greater variability and lower abundance of niche-specific functionality in rare but consistently present pathways, e.g. spermidine biosynthesis, methionine degradation, and hydrogen sulfide production, all examples highly prevalent in gastrointestinal body sites (nonzero in >92% of samples) but at very low abundance (median rel. abd. <0.0052). This "long tail" of low-abundance genes and pathways also likely encodes much of the uncharacterized biomolecular function and metabolism of these metagenomes, the expression levels of which remain to be explored in future metatranscriptomic studies.

Protein families showed diversity and prevalence trends similar to those of full pathways, ranging from maxima of only ~16,000 unique families per community in the vagina to almost 400,000 in the oral cavity (Fig. 1A–B, http://hmpdacc.org/HMGI). A striking fraction of these families were indeed functionally uncharacterized, including those detected by read mapping, with a minimum in the oral cavity (mean 58% sd. 6.8%) and maximum in the nares (mean 77% sd. 11%). Likewise, many genes annotated from assemblies could not be assigned a metabolic function, with a minimum in the vagina (mean 78% sd. 3.4%) and maximum in the gut (mean 86% sd. 0.9%). The latter range did not differ substantially by body habitat and is in close agreement with previous comprehensive gene catalogs of the gut metagenome[4]. Taken together with the microbial variation observed above throughout the human microbiome, functional variation among individuals might indicate pathways of particular importance in maintaining community structure in the face of personalized immune, environmental, or dietary exposures among these subjects. Determining the functions of uncharacterized core and variable protein families will be especially essential in understanding the microbiota's role in health and disease.

Correlations with host phenotype

We finally examined relationships associating both clades and metabolism in the microbiota with host properties such as age, gender, BMI, and other available clinical metadata (Fig. 5; Supp. Table 3). Using a sparse multivariate model, 960 microbial, enzymatic, or pathway abundances were significantly associated with one or more of 15 subject phenotype and sample metadata features. A wide variety of taxa, gene families, and metabolic pathways were differentially distributed with subject ethnicity at every body habitat (Fig. 5A), representing the phenotype with the greatest number 266 at FDR q<0.2) of total associations with the microbiome. Vaginal pH has also been observed to correlate with microbiome composition[11], and we detected in this population both the expected reduction in

Lactobacillus at high pH and a corresponding increase in metabolic diversity (Fig. 5B). Intriguingly, and not previously observed, subject age was most associated with a collection of highly differential metagenomically encoded pathways on the skin (Fig. 5C), as well as shifts in skin clades including retroauricular Firmicutes (p=1.0e-4, q=0.033). The examples of associations with ethnicity and vaginal pH are among the strongest associations with the microbiome, however, and most correlates (e.g. with subject BMI, Fig. 5D) are more representatively modest. This lower degree of correlation held for most available biometrics (gender, temperature, blood pressure, etc.), with even the most significant associations possessing generally low effect sizes and considerable unexplained variance. We conclude that most variation in the human microbiome is not well-explained by these phenotypic metadata, and other potentially important factors such as short- and long-term diet, daily cycles, founder effects such as mode of delivery, and host genetics should be considered in future analyses.

Conclusions

This extensive sampling of the human microbiome across many subjects and body habitats provides an initial characterization of the normal microbiota of healthy adults in a Western population. The large sample size and consistent sampling of many sites from the same individuals allows for the first time an understanding of the relationships among microbes, and between the microbiome and clinical parameters, that underpin the basis for individual variation -- variation that may ultimately be critical for understanding microbiome-based disorders. Clinical studies of the microbiome will be able to leverage the resulting extensive catalogs of taxa, pathways, and genes[1], although they must also still include carefully matched internal controls. The uniqueness of each individual's microbiome even in this reference population argues for future studies to consider prospective within-subjects designs where possible. The HMP's unique combination of organismal and functional data across body habitats, encompassing both 16S and metagenomic profiling, together with detailed characterization of each subject, has allowed us and subsequent studies to move beyond the observation of variability in the human microbiome to ask how and why these microbial communities vary so extensively.

Many details remain for further work to fill in, building on this reference study. How do early colonization and life-long change vary among body habitats? Do epidemiological patterns of transmission of beneficial or harmless microbes mirror patterns of transmission of pathogens? Which co-occurrences among microbes reflect shared response to the environment, as opposed to competitive or mutualistic interactions? How large a role does host immunity or genetics play in shaping patterns of diversity, and how do the patterns observed in this North American population compare to those around the world? Future studies building on the gene and organism catalogs established by the Human Microbiome Project, including increasingly detailed investigations of metatranscriptomes and metaproteomes, will help to unravel these open questions and allow us to more fully understand the links between the human microbiome, health, and disease.

Methods Summary

Microbiome samples were collected from up to 18 body sites at one or two time points from 242 individuals clinically screened for absence of disease[2]. Samples were subjected to 16S rRNA gene pyrosequencing (454 Life Sciences), and a subset were shotgun sequenced for metagenomics using the Illumina GAIIx platform[1]. 16S data processing and diversity estimates were performed using QIIME[27], and metagenomic data were taxonomically profiled using MetaPhlAn[13], metabolically profiled by HUMAnN[26], and assembled for gene annotation and clustering into a unique catalog[1]. Potential pathogens were identified using

the PATRIC database[14], isolate reference genome annotations drawn from KEGG[28], and reference genome mapping performed by BWA[29] to a reduced set of genomes to which short reads could be matched[30]. Microbial associations were assessed by similarity measures accounting for compositionality[23], and phenotypic association testing was performed in R. All data and additional protocol details are available at http://hmpdacc.org. Full methods accompany this paper.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The Consortium would like to thank our external scientific advisory board: Richard Blumberg, Julian Davies, Robert Holt, Pilar Ossorio, Francis Ouellette, Gary Schoolnik, and Alan Williamson. We would also like to thank our collaborators throughout the International Human Microbiome Consortium, particularly the investigators of the MetaHIT project, for advancing human microbiome research. Data repository management was provided by the National Center for Biotechnology Information and the Intramural Research Program of the NIH National Library of Medicine. We especially appreciate the generous participation of the individuals from the Saint Louis, MO, and Houston, TX areas who made this study possible. This research was supported in part by National Institutes of Health grants U54HG004969 to B.W.B.; U54HG003273 to R.A.G.; U54HG004973 to R.A.G., S.K.H. and J.F.P.; U54HG003067 to E.S.L.; U54AI084844 to K.E.N.; N01AI30071 to R.L.S.; U54HG004968 to G.M.W.; U01HG004866 to O.R.W.; U54HG003079 to R.K.W.; R01HG005969 to C.H.; R01HG004872 to R.K.; R01HG004885 to M.P.; R01HG005975 to P.D.S.; R01HG004908 to Y.Y.; R01HG004900 to M.K.C. and P.L.S.; R01HG005171 to D.E.H.; R01HG004853 to A.L.M.; R01HG004856 to R.R.; R01HG004877 to R.R.S. and R.F.; R01HG005172 to P.G.S.; R01HG004857 to M.P.; R01HG004906 to T.M.S.; R21HG005811 to E.A.V.; M.J.B. was supported by UH2AR057506; G.A.B. was supported by UH2AI083263 and UH3AI083263 (G.A.B., Cynthia N. Cornelissen, Lindon K. Eaves and Jerome F. Strauss); S.M.H. was supported by UH3DK083993 (Vincent B. Young, Eugene B. Chang, Folker Meyer, Thomas M. Schmidt, Mitchell L. Sogin, James M. Tiedje); K.P.R. was supported by UH2DK083990 (James Versalovic); J.A.S. and H.H.K. were supported by UH2AR057504 and UH3AR057504 (J.A.S.); DP2OD001500 to K.M.A.; N01HG62088 to the Coriell Institute for Medical Research; U01DE016937 to F.E.D.; S.K.H. was supported by RC1DE202098 and R01DE021574 (S.K.H. and Huiying Li); J.G.I. was supported by R21CA139193 (J.G.I. and Dominique S. Michaud); K.P.L. was supported by P30DE020751 (Daniel J. Smith); Army Research Office grant W911NF-11-1-0473 to C.H.; National Science Foundation grants NSF DBI-1053486 to C.H. and NSF IIS-0812111 to M.P.; The Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 for P.S.G.C.; LANL Laboratory-Directed Research and Development grant 20100034DR and the U.S. Defense Threat Reduction Agency grants B104153I and B084531I to P.S.G.C.; Research Foundation -Flanders (FWO) grant to K.F. and J.R.; R.K. is an HHMI Early Career Scientist; Gordon & Betty Moore Foundation funding and institutional funding from the J. David Gladstone Institutes to K.S.P.; A.M.S was supported by fellowships provided by the Rackham Graduate School and the NIH Molecular Mechanisms in Microbial Pathogenesis Training Grant T32AI007528; a Crohn's and Colitis Foundation of Canada Grant in Aid of Research to E.A.V.; 2010 IBM Faculty Award to K.C.W.; Analysis of the HMP data was performed using National Energy Research Scientific Computing resources; the BluBioU Computational Resource at Rice University.

References

1. The Human Microbiome Project Consortium. A framework for human microbiome research. (in review)
2. Aagaard, K., et al. A Comprehensive Strategy for Sampling the Human Microbiome. (in review)
3. Turnbaugh PJ, et al. A core gut microbiome in obese and lean twins. Nature. 2009; 457:480–484. nature07540 [pii]. 10.1038/nature07540 [PubMed: 19043404]
4. Qin J, et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature. 2010; 464:59–65. nature08821 [pii]. 10.1038/nature08821 [PubMed: 20203603]
5. Fredricks DN, Fiedler TL, Marrazzo JM. Molecular identification of bacteria associated with bacterial vaginosis. The New England journal of medicine. 2005; 353:1899–1911.10.1056/NEJMoa043802 [PubMed: 16267321]
6. Costello EK, et al. Bacterial community variation in human body habitats across space and time. Science. 2009; 326:1694–1697.10.1126/science.1177486 [PubMed: 19892944]

7. Huse, S.; Ye, Y.; Zhou, Y.; Fodor, A. A Core Human Microbiome as Viewed Through 16S rRNA Sequences Clusters. (in press)
8. Zhou, Y.; Gao, H.; Mihindukulasuriya, K.; Sodergen, E.; Weinstock, GM. Defining Core Microbiomes in Healthy Humans. (in preparation)
9. Li, K.; Bihan, M.; Yooseph, S.; Methe, BA. Analyses of the Microbial Diversity across the Human Microbiome. (in press)
10. Grice EA, et al. Topographical and temporal diversity of the human skin microbiome. Science. 2009; 324:1190–1192. 324/5931/1190 [pii]. 10.1126/science.1171700 [PubMed: 19478181]
11. Ravel J, et al. Vaginal microbiome of reproductive-age women. Proc Natl Acad Sci U S A. 2011; 108(Suppl 1):4680–4687. 1002611107 [pii]. 10.1073/pnas.1002611107 [PubMed: 20534435]
12. Segata, N., et al. Composition of the Adult Digestive Tract Microbiome Based on Seven Mouth Surfaces, Tonsils, Throat and Stool Samples. (in review)
13. Segata, N., et al. Efficient metagenomic microbial community profiling using unique clade-specific marker genes. (in press)
14. Gillespie JJ, et al. PATRIC: The Comprehensive Bacterial Bioinformatics Resource with a Focus on Human Pathogenic Species. Infection and immunity. 201110.1128/IAI.00207-11
15. Sharpton TJ, et al. PhylOTU: a high-throughput procedure quantifies microbial community diversity and resolves novel taxa from metagenomic data. PLoS Comput Biol. 2011; 7:e1001061.10.1371/journal.pcbi.1001061 [PubMed: 21283775]
16. Wylie KM, et al. Novel Bacterial Taxa in the Human Microbiome. PLoS ONE. (in press).
17. Sokol H, et al. Faecalibacterium prausnitzii is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients. Proc Natl Acad Sci U S A. 2008; 105:16731–16736. 0804812105 [pii]. 10.1073/pnas.0804812105 [PubMed: 18936492]
18. Aas JA, Paster BJ, Stokes LN, Olsen I, Dewhirst FE. Defining the normal bacterial flora of the oral cavity. J Clin Microbiol. 2005; 43:5721–5732. 43/11/5721 [pii]. 10.1128/JCM.43.11.5721–5732.2005 [PubMed: 16272510]
19. Medini D, et al. Microbiology in the post-genomic era. Nat Rev Microbiol. 2008; 6:419–430. nrmicro1901 [pii]. 10.1038/nrmicro1901 [PubMed: 18475305]
20. Mazmanian SK, Round JL, Kasper DL. A microbial symbiosis factor prevents intestinal inflammatory disease. Nature. 2008; 453:620–625. nature07008 [pii]. 10.1038/nature07008 [PubMed: 18509436]
21. Goodman AL, et al. Identifying genetic determinants needed to establish a human gut symbiont in its habitat. Cell host & microbe. 2009; 6:279–289.10.1016/j.chom.2009.08.003 [PubMed: 19748469]
22. Kuehnert MJ, et al. Prevalence of Staphylococcus aureus nasal colonization in the United States, 2001–2002. J Infect Dis. 2006; 193:172–179. JID35385 [pii]. 10.1086/499632 [PubMed: 16362880]
23. Faust, K., et al. Microbial co-occurrence relationships in the human microbiome. (in review)
24. Koren, O., et al. Enterotypes Lost in Gradients: A Meta-Analysis of Human Microbiome Project and Community 16S rRNA Data. (in review)
25. Zhou, Y., et al. Community states and variability of the human microbiome. (in preparation)
26. Abubucker, S., et al. Metabolic reconstruction for metagenomic data and its application to the human microbiome. (in press)
27. Caporaso JG, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Methods. 2010; 7:335–336. nmeth.f.303 [pii]. 10.1038/nmeth.f.303 [PubMed: 20383131]
28. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M. KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res. 2010; 38:D355–360. gkp896 [pii]. 10.1093/nar/gkp896 [PubMed: 19880382]
29. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010; 26:589–595. btp698 [pii]. 10.1093/bioinformatics/btp698 [PubMed: 20080505]

30. Giannoukos G, et al. Efficient and robust RNA-seq process for cultured bacteria and complex community transcriptomes. Genome biology. 2012; 13:R23.10.1186/gb-2012-13-3-r23 [PubMed: 22455878]

31. Langille MG, Brinkman FS. IslandViewer: an integrated interface for computational identification and visualization of genomic islands. Bioinformatics. 2009; 25:664–665.10.1093/bioinformatics/btp030 [PubMed: 19151094]

The Human Microbiome Project Consortium

Curtis Huttenhower[1,2,*], Dirk Gevers[2,*], Rob Knight[3,4], Sahar Abubucker[5], Jonathan H Badger[6], Asif T Chinwalla[5], Heather H Creasy[7], Ashlee M Earl[2], Michael G FitzGerald[2], Robert S Fulton[5], Michelle G Giglio[7], Kymberlie Hallsworth-Pepin[5], Elizabeth A Lobos[5], Ramana Madupu[6], Vincent Magrini[5], John C Martin[5], Makedonka Mitreva[5], Donna M Muzny[8], Erica J Sodergren[5], James Versalovic[9,10], Aye M Wollam[5], Kim C Worley[8], Jennifer R Wortman[2], Sarah K Young[2], Qiandong Zeng[2], Kjersti M Aagaard[11], Olukemi O Abolude[7], Emma Allen-Vercoe[12], Eric J Alm[13,2], Lucia Alvarado[2], Gary L Andersen[14], Scott Anderson[2], Elizabeth Appelbaum[5], Harindra M Arachchi[2], Gary Armitage[15], Cesar A Arze[7], Tulin Ayvaz[16], Carl C Baker[17], Lisa Begg[18], Tsegahiwot Belachew[19], Veena Bhonagiri[5], Monika Bihan[6], Martin J Blaser[20], Toby Bloom[2], Vivien Bonazzi[21], J Paul Brooks[22,23], Gregory A Buck[23,24], Christian J Buhay[8], Dana A Busam[6], Joseph L Campbell[21,19], Shane R Canon[25], Brandi L Cantarel[7], Patrick S G Chain[26,27], I-Min A Chen[28], Lei Chen[5], Shaila Chhibba[21], Ken Chu[28], Dawn M Ciulla[2], Jose C Clemente[3], Sandra W Clifton[5], Sean Conlan[21], Jonathan Crabtree[7], Mary A Cutting[29], Noam J Davidovics[7], Catherine C Davis[30], Todd Z DeSantis[31], Carolyn Deal[19], Kimberley D Delehaunty[5], Floyd E Dewhirst[32,33], Elena Deych[34], Yan Ding[8], David J Dooling[5],

[1]Biostatistics, Harvard School of Public Health, Boston MA
[2]The Broad Institute of MIT and Harvard, Cambridge MA
[*]Equal contribution
[3]Department of Chemistry and Biochemistry, University of Colorado, Boulder CO
[4]Howard Hughes Medical Institute, Boulder CO
[5]The Genome Institute, Washington University School of Medicine, St. Louis MO
[6]J. Craig Venter Institute, Rockville MD
[7]Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore MD
[8]Human Genome Sequencing Center, Baylor College of Medicine, Houston TX
[9]Department of Pathology & Immunology, Baylor College of Medicine, Houston TX
[10]Department of Pathology, Texas Children's Hospital, Houston TX
[11]Department of Obstetrics & Gynecology, Division of Maternal-Fetal Medicine, Baylor College of Medicine, Houston TX
[12]Molecular and Cellular Biology, University of Guelph, Guleph, Canada
[13]Department of Civil & Environmental Engineering, Massachusetts Institute of Technology, Cambridge MA
[14]Center for Environmental Biotechnology, Lawrence Berkeley National Laboratory, Berkeley CA
[15]School of Dentistry, University of California, San Francisco, San Francisco CA
[16]Molecular Virology and Microbiology, Baylor College of Medicine, Houston TX
[17]National Institute of Arthritis and Musculoskeletal and Skin, National Institutes of Health, Bethesda MD
[18]Office of Research on Women's Health, National Institutes of Health, Bethesda MD
[19]National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda MD
[20]Department of Medicine, New York University Langone Medical Center, New York NY
[21]National Human Genome Research Institute, National Institutes of Health, Bethesda MD
[22]Department of Statistical Sciences and Operations Research, Virginia Commonwealth University, Richmond VA
[23]Center for the Study of Biological Complexity, Virginia Commonwealth University, Richmond VA
[24]Department of Biology, Virginia Commonwealth University, Richmond VA
[25]Technology Integration Group, National Energy Research Scientific Computing Center, Lawrence Berkeley National Laboratory, Berkeley CA
[26]Genome Science Group, Bioscience Division, Los Alamos National Laboratory, Los Alamos NM
[27]Joint Genome Institute, Walnut Creek CA
[28]Biological Data Management and Technology Center, Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley CA
[29]National Institute of Dental and Craniofacial Research (NIDCR), National Institutes of Health, Bethesda MD
[30]FemCare Product Safety and Regulatory Affairs, The Procter & Gamble Company, Cincinnati OH
[31]Bioinformatics Department, Second Genome, Inc., San Bruno CA
[32]Department of Molecular Genetics, Forsyth Institute, Cambridge MA

Shannon P Dugan[8], Wm Michael Dunne[35,36], A Scott Durkin[6], Robert C Edgar[37], Rachel L Erlich[2], Candace N Farmer[5], Ruth M Farrell[38], Karoline Faust[39,40], Michael Feldgarden[2], Victor M Felix[7], Sheila Fisher[2], Anthony A Fodor[41], Larry J Forney[42], Leslie Foster[6], Valentina Di Francesco[19], Jonathan Friedman[43], Dennis C Friedrich[2], Catrina C Fronick[5], Lucinda L Fulton[5], Hongyu Gao[5], Nathalia Garcia[44], Georgia Giannoukos[2], Christina Giblin[19], Maria Y Giovanni[19], Jonathan M Goldberg[2], Johannes Goll[6], Antonio Gonzalez[45], Allison Griggs[2], Sharvari Gujja[2], Susan Kinder Haake[46], Brian J Haas[2], Holli A Hamilton[29], Emily L Harris[29], Theresa A Hepburn[2], Brandi Herter[5], Diane E Hoffmann[47], Michael E Holder[8], Clinton Howarth[2], Katherine H Huang[2], Susan M Huse[48], Jacques Izard[32,33], Janet K Jansson[49], Huaiyang Jiang[8], Catherine Jordan[7], Vandita Joshi[8], James A Katancik[50], Wendy A Keitel[16], Scott T Kelley[51], Cristyn Kells[2], Nicholas B King[52], Dan Knights[45], Heidi H Kong[53], Omry Koren[54], Sergey Koren[55], Karthik C Kota[5], Christie L Kovar[8], Nikos C Kyrpides[27], Patricio S La Rosa[34], Sandra L Lee[8], Katherine P Lemon[32,56], Niall Lennon[2], Cecil M Lewis[57], Lora Lewis[8], Ruth E Ley[54], Kelvin Li[6], Konstantinos Liolios[27], Bo Liu[55], Yue Liu[8], Chien-Chi Lo[26], Catherine A Lozupone[3], R Dwayne Lunsford[29], Tessa Madden[58], Anup A Mahurkar[7], Peter J Mannon[59], Elaine R Mardis[5], Victor M Markowitz[27,28], Konstantinos Mavromatis[27], Jamison M McCorrison[6], Daniel McDonald[3], Jean McEwen[21], Amy L McGuire[60], Pamela McInnes[29], Teena Mehta[2], Kathie A Mihindukulasuriya[5], Jason R Miller[6], Patrick J Minx[5], Irene Newsham[8], Chad Nusbaum[2], Michelle O'Laughlin[5], Joshua Orvis[7], Ioanna Pagani[27], Krishna Palaniappan[28], Shital M Patel[61], Matthew Pearson[2], Jane Peterson[21], Mircea Podar[62], Craig Pohl[5], Katherine S Pollard[63,64,65], Mihai Pop[55,66], Margaret E Priest[2], Lita M Proctor[21], Xiang Qin[8], Jeroen Raes[39,40], Jacques Ravel[7], Jeffrey G Reid[8], Mina Rho[67], Rosamond Rhodes[68], Kevin P Riehle[69], Maria C Rivera[23,24], Beltran Rodriguez-

[33]Department of Oral Medicine, Infection and Immunity, Harvard School of Dental Medicine, Boston MA
[34]Department of Medicine, Division of General Medical Science, Washington University School of Medicine, St. Louis MO
[35]Department of Pathology & Immunology, Washington University School of Medicine, St. Louis MO
[36]bioMerieux, Inc., Durham NC
[37]drive5.com, Tiburon CA
[38]Center for Ethics, Humanities and Spiritual Care, Cleveland Clinic, Cleveland OH
[39]Department of Structural Biology, VIB, Belgium, Brussels, Belgium
[40]Department of Applied Biological Sciences (DBIT), Vrije Universiteit Brussel, Brussels, Belgium
[41]Department of Bioinformatics and Genomics, University of North Carolina - Charlotte, Charlotte NC
[42]Department of Biological Sciences, University of Idaho, Moscow ID
[43]Computational and Systems Biology, Massachusetts Institute of Technology, Cambridge MA
[44]Center for Advanced Dental Education, Saint Louis University, St. Louis MO
[45]Department of Computer Science, University of Colorado, Boulder CO
[46]Division of Associated Clinical Specialties and Dental Research Institute, UCLA School of Dentistry, Los Angeles CA
[47]University of Maryland Francis King Carey School of Law, Baltimore MD
[48]Josephine Bay Paul Center, Marine Biological Laboratory, Woods Hole MA
[49]Ecology Department, Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley CA
[50]Department of Periodontics, University of Texas Health Science Center School of Dentistry, Houston TX
[51]Department of Biology, San Diego State University, San Diego CA
[52]Faculty of Medicine, McGill University, Montreal, Canada
[53]Dermatology Branch, CCR, National Cancer Institute, Bethesda MD
[54]Department of Microbiology, Cornell University, Ithaca NY
[55]Center for Bioinformatics and Computational Biology, University of Maryland, College Park MD
[56]Division of Infectious Diseases, Children's Hospital Boston, Harvard Medical School, Boston MA
[57]Department of Anthropology, University of Oklahoma, Norman OK
[58]Department of Obstetrics and Gynecology, Washington University School of Medicine, Saint Louis MO
[59]Division of Gastroenterology and Hepatology, University of Alabama at Birmingham, Birmingham AL
[60]Center for Medical Ethics and Health Policy, Baylor College of Medicine, Houston TX
[61]Medicine-Infectious Disease, Baylor College of Medicine, Houston TX
[62]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge TN
[63]Gladstone Institutes, University of California, San Francisco, San Francisco CA
[64]Institute for Human Genetics, University of California, San Francisco, San Francisco CA
[65]Division of Biostatistics, University of California, San Francisco, San Francisco CA
[66]Department of Computer Science, University of Maryland, College Park MD
[67]School of Informatics and Computing, Indiana University, Bloomington IN
[68]Mount Sinai School of Medicine, New York NY

Mueller[51], Yu-Hui Rogers[6], Matthew C Ross[16], Carsten Russ[2], Ravi K Sanka[6], Pamela Sankar[70], J Fah Sathirapongsasuti[1], Jeffery A Schloss[21], Patrick D Schloss[71], Thomas M Schmidt[72], Matthew Scholz[26], Lynn Schriml[7], Alyxandria M Schubert[71], Nicola Segata[1], Julia A Segre[21], William D Shannon[34], Richard R Sharp[38], Thomas J Sharpton[63], Narmada Shenoy[2], Nihar U Sheth[23], Gina A Simone [73], Indresh Singh[6], Christopher S Smillie[43], Jack D Sobel[74], Daniel D Sommer[55], Paul Spicer[57], Granger G Sutton[6], Sean M Sykes[2], Diana G Tabbaa[2], Mathangi Thiagarajan[6], Chad M Tomlinson[5], Manolito Torralba[6], Todd J Treangen[75], Rebecca M Truty[63], Tatiana A Vishnivetskaya[62], Jason Walker[5], Lu Wang[21], Zhengyuan Wang[5], Doyle V Ward[2], Wesley Warren[5], Mark A Watson[35], Christopher Wellington[21], Kris A Wetterstrand[21], James R White[7], Katarzyna Wilczek-Boney[8], YuanQing Wu[8], Kristine M Wylie[5], Todd Wylie[5], Chandri Yandava[2], Liang Ye[5], Yuzhen Ye[67], Shibu Yooseph[76], Bonnie P Youmans[16], Lan Zhang[8], Yanjiao Zhou[5], Yiming Zhu[8], Laurie Zoloth[77], Jeremy D Zucker[2], Bruce W Birren[2], Richard A Gibbs[8], Sarah K Highlander[8,16], Barbara A Methé [6], Karen E Nelson[6], Joseph F Petrosino[8,78,16], George M Weinstock[5], Richard K Wilson[5], Owen White[7]

[69]Molecular & Human Genetics, Baylor College of Medicine, Houston TX
[70]Center for Bioethics and Department of Medical Ethics, University of Pennsylvania, Philadelphia PA
[71]Department of Microbiology & Immunology, University of Michigan, Ann Arbor MI
[72]Department of Microbiology and Molecular Genetics, Michigan State University, East Lansing MI
[73]The EMMES Corporation, Rockville MD
[74]Harper University Hospital, Wayne State University School of Medicine, Detroit MI, Detroit MI
[75]McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore MD
[76]J. Craig Venter Institute, San Diego CA
[77]Feinberg School of Medicine, Northwestern University, Chicago IL
[78]Alkek Center for Metagenomics and Microbiome Research, Baylor College of Medicine, Houston TX

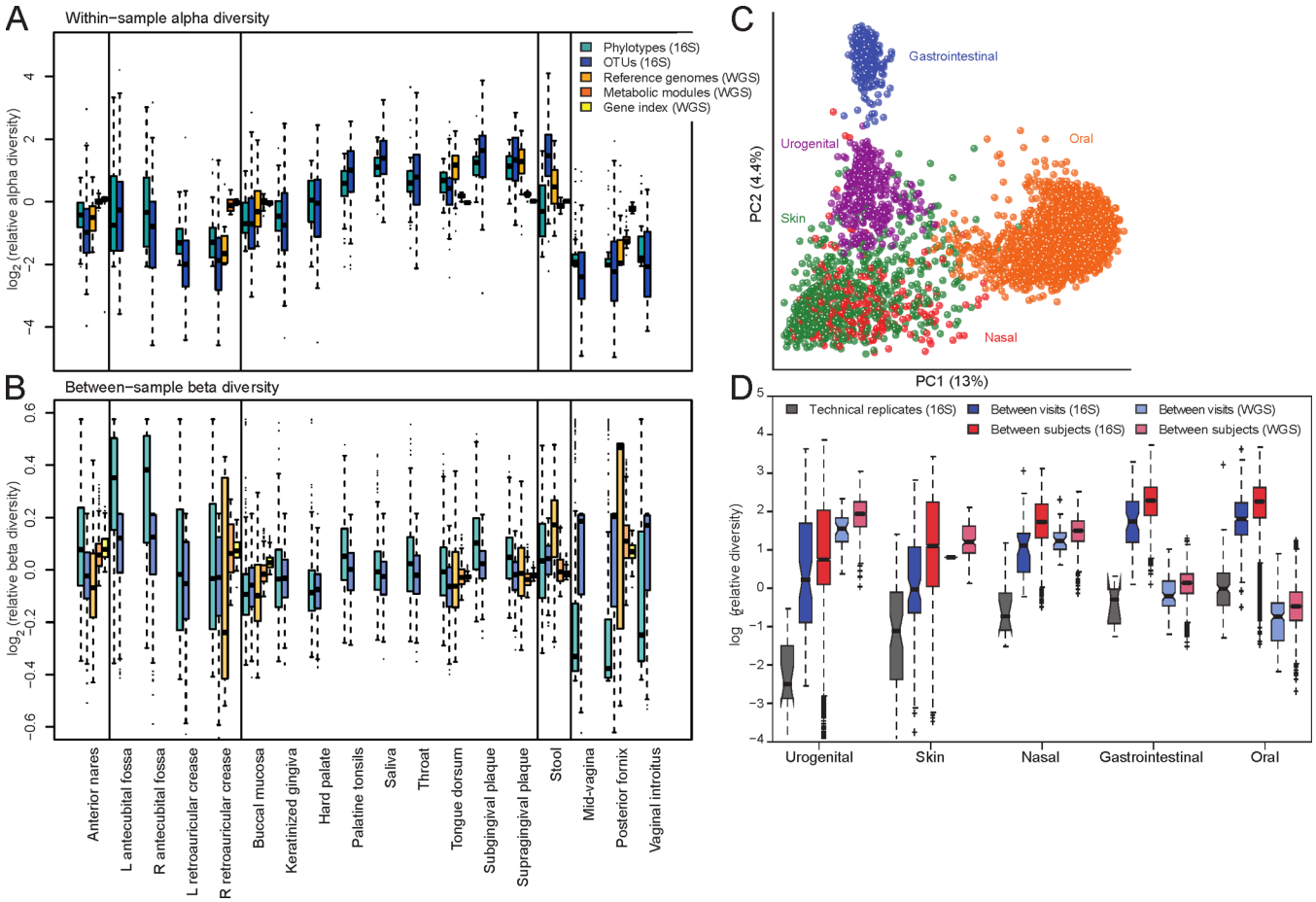

Figure 1. Diversity of the human microbiome is concordant among measures, unique to each individual, and strongly determined by microbial habitat

A) Alpha diversity within subjects by body habitat, as measured using the relative inverse Simpson index of 16S rRNA gene OTUs (red), genus-level phylotypes (blue), shotgun metagenomic reads matched to reference genomes (green), functional modules (yellow), and enzyme families (white). The mouth generally shows high within-subject diversity and the vagina low diversity, with other habitats intermediate; variation among individuals often exceeds variation among body habitats. B) Bray-Curtis beta diversity among subjects by body habitat, colors as for A. Skin differs most between subjects, with oral habitats and vaginal genera more stable. Although alpha- and beta-diversity are not directly comparable, changes in structure among communities (A) occupy a wider dynamic range than do changes within communities among individuals (B). C) Principal coordinates plot showing variation among samples demonstrates that primary clustering is by body area, with the oral, gastrointestinal, skin, and urogenital habitats separate; the nares bridge oral and skin habitats. D) Repeated samples from the same subject (red) are more similar than microbiomes from different subjects (yellow). Technical replicates (green) are in turn more similar; these patterns are consistent for all body habitats and for both phylogenetic and metabolic community composition. See previously described sample counts[1] for all comparisons.

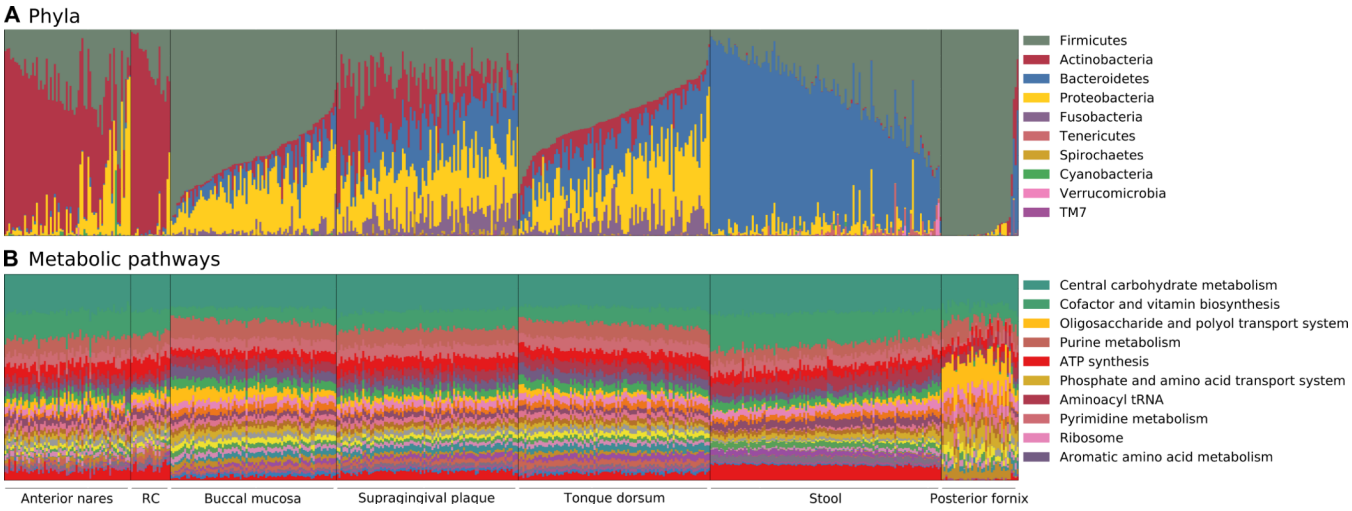

Figure 2. Carriage of microbial taxa varies while metabolic pathways remain stable within a healthy population

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 phyla from binned OTUs and B) metabolic modules. Legend indicates most abundant phyla/pathways by average within one or more body habitats; RC = retroauricular crease. A plurality of most communities' memberships consists of a single dominant phylum (and often genus; see Supp. Fig. 2), 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.

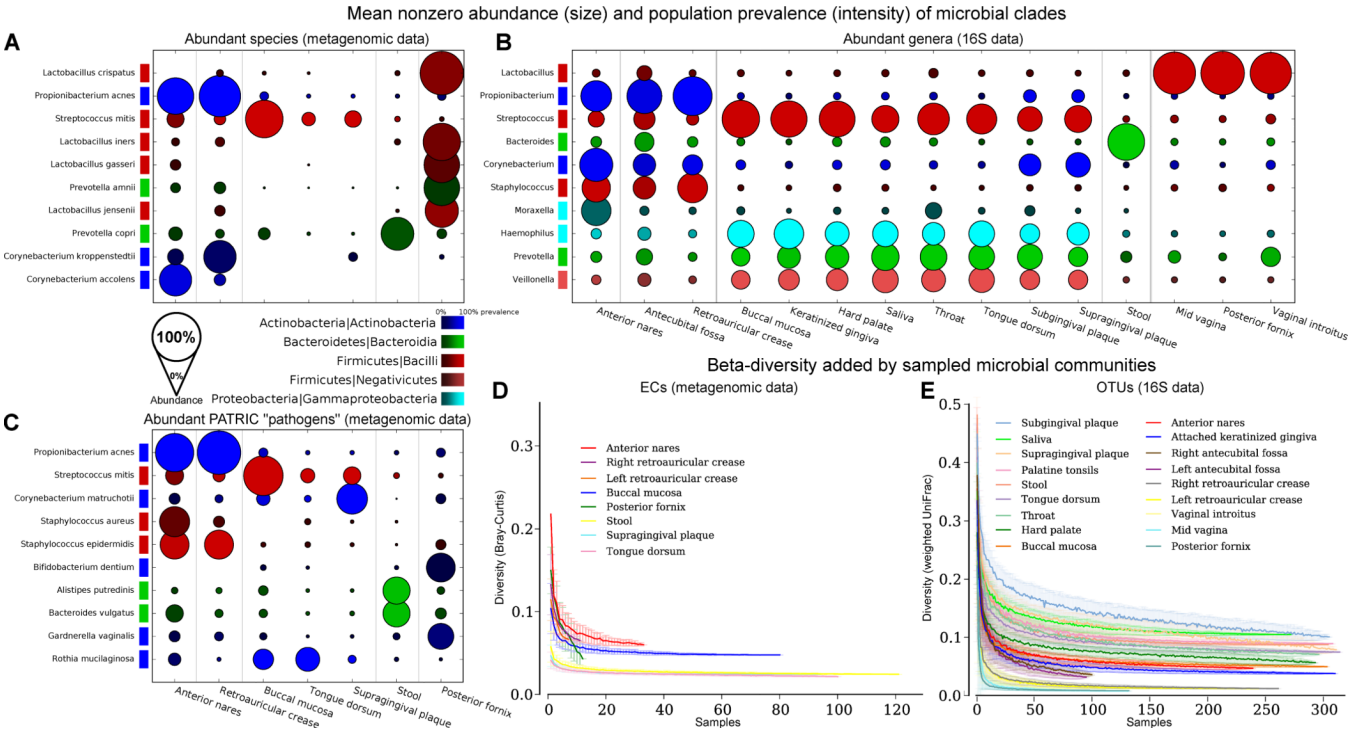

Figure 3. Abundant taxa in the human microbiome, which has been metagenomically and taxonomically well-defined in the HMP population

A–C) Prevalence (intensity, color denoting phylum/class) and abundance when present (size) of clades in the healthy microbiome. The most abundant A) metagenomically-identified species, B) 16S-identified genera, and C) PATRIC[14] “pathogens” (metagenomic). The population size and sequencing depths of the HMP have well-defined the microbiome at all assayed body sites, as assessed by saturation of added community D) metabolic configurations (rarefaction of minimum Bray-Curtis β-diversity of metagenomic EC abundances to nearest neighbor, inter-quartile range over 100 samples) and E) phylogenetic configurations (min. 16S OTU weighted UniFrac distance to nearest neighbor).

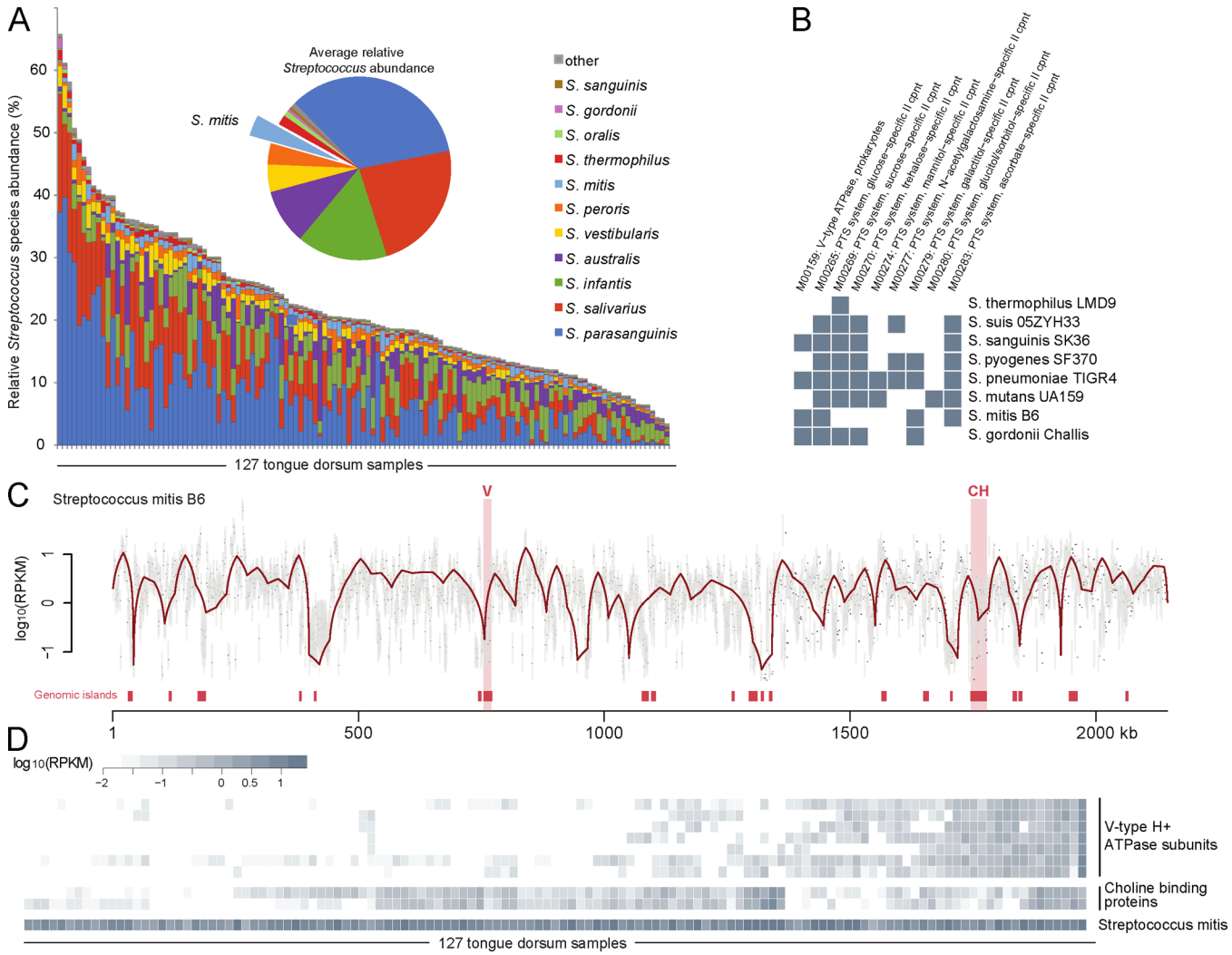

Figure 4. Microbial carriage varies between subjects down to the species and strain level

Metagenomic reads from 127 tongue samples spanning 90 subjects were processed with MetaPhlAn to determine relative abundances for each species. A) Relative abundances of 11 distinct *Streptococcus* spp. In addition to variation in broader clades (see Fig. 2), individual species within a single habitat demonstrate a wide range of compositional variation. Inset illustrates average tongue sample composition. B) Metabolic modules present/absent (grey/white) in KEGG[28] reference genomes of tongue streptococci denote selected areas of strain-specific functional differentiation. C) Comparative genomic coverage for the single *Streptococcus mitis* B6 strain. Grey dots are median Reads Per Kilobase per Million reads (RPKM) for 1kb windows, gray bars are the 25th to 75th percentiles across all samples, red line the lowess smoothed average. Red bars at the bottom highlight predicted genomic islands[31]. Large, discrete, and highly variable islands are commonly under-represented. D) Two islands highlighted, V = V-type H+ ATPase subunits I,K,E,C,F,A & B, and CH = Choline binding proteins cbp6 and cbp12, indicating functional cohesion of strain-specific gene loss within individual human hosts.

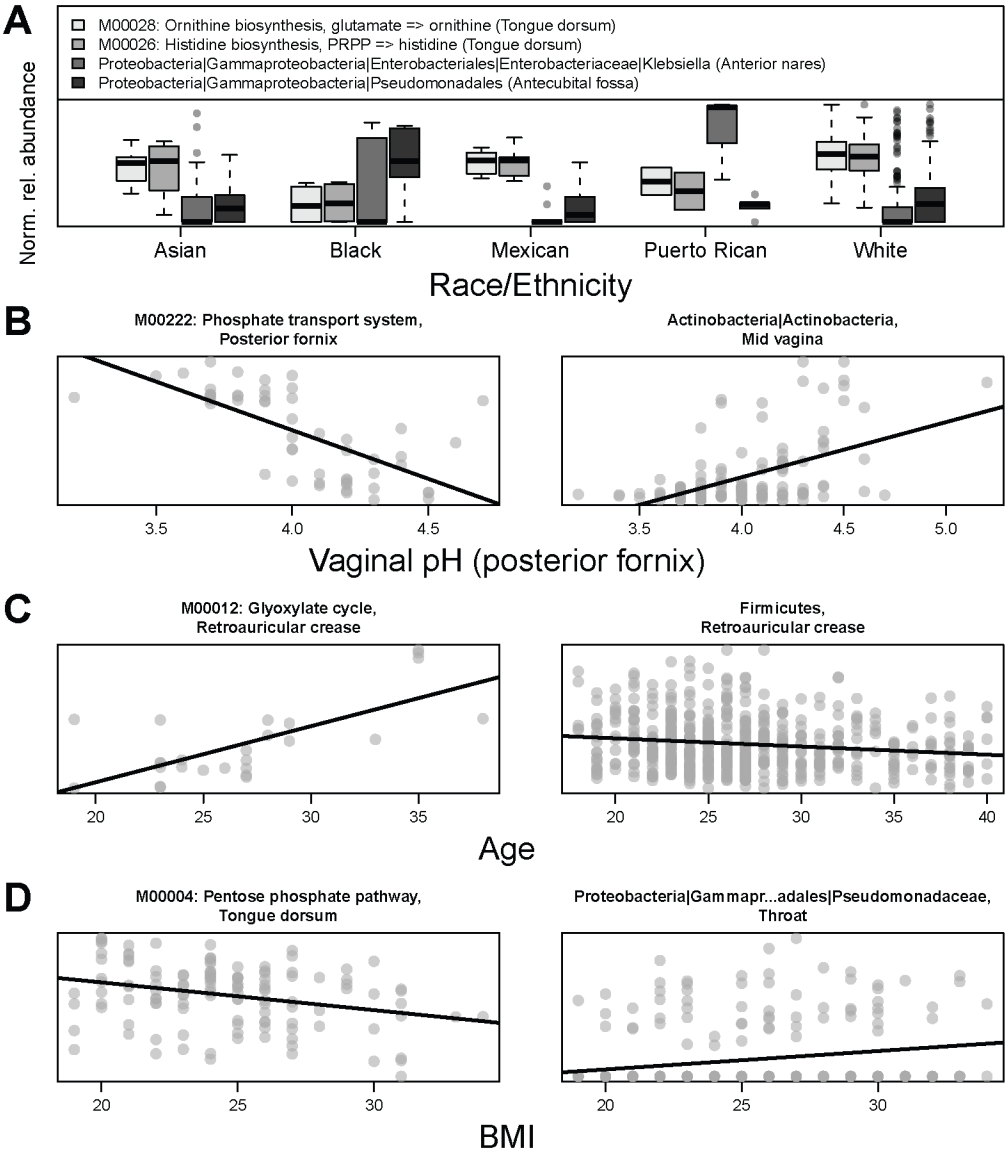

Figure 5. Microbial community membership and function correlates with host phenotype and sample metadata

The pathway and clade abundances most significantly associated (all FDR q<0.2) using a multivariate linear model with A) subject race or ethnicity, B) vaginal posterior fornix pH, C) subject age, and D) BMI. Samples' scatter plots are shown with lines indicating best simple linear fit. Race/ethnicity and vaginal pH are particularly strong associations; age and BMI are more representative of typically modest phenotypic associations (Sup. Table 3), suggesting that variation in the healthy microbiota may correspond to other host or environmental factors.