Published in final edited form as:
Nature. 2017 October 05; 550(7674): 61–66. doi:10.1038/nature23889.

Strains, functions, and dynamics in the expanded Human Microbiome Project

Jason Lloyd-Price[1,2,*], **Anup Mahurkar**[3,*], **Gholamali Rahnavard**[1,2], **Jonathan Crabtree**[3], **Joshua Orvis**[3], **A. Brantley Hall**[2], **Arthur Brady**[3], **Heather H. Creasy**[3], **Carrie McCracken**[3], **Michelle G. Giglio**[3], **Daniel McDonald**[4], **Eric A. Franzosa**[1,2], **Rob Knight**[4,5], **Owen White**[3], and **Curtis Huttenhower**[1,2]

[1]Biostatistics Department, Harvard T. H. Chan School of Public Health, Boston, MA 02115, USA

[2]The Broad Institute, Cambridge, MA 02142, USA

[3]Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD 21201, USA

[4]Department of Pediatrics, University of California San Diego, La Jolla, CA 92093, USA

[5]Department of Computer Science & Engineering, University of California San Diego, La Jolla, CA 92093, USA

Summary

The characterization of baseline microbial and functional diversity in the human microbiome has enabled studies of microbiome-related disease, microbial population diversity, biogeography, and molecular function. The NIH Human Microbiome Project (HMP) has provided one of the broadest such characterizations to date. Here, we introduce an expanded second phase of the study, abbreviated HMP1-II, comprising 1,631 new metagenomic samples (2,355 total) targeting diverse body sites with multiple time points in 265 individuals. We applied updated profiling and assembly methods to these data to provide new characterizations of microbiome personalization. Strain identification revealed distinct subspecies clades specific to body sites; it also quantified species with phylogenetic diversity under-represented in isolate genomes. Body-wide functional profiling classified pathways into universal, human-enriched, and body site-enriched subsets. Finally, temporal analysis decomposed microbial variation into rapidly variable, moderately variable, and stable subsets. This study furthers our knowledge of baseline human microbial diversity, thus enabling an understanding of personalized microbiome function and dynamics.

Reprints and permissions information is available at www.nature.com/reprints.

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

[*]These authors contributed equally to this work.

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

Author Contributions
GR and AM collected and organized the data. GR generated and analyzed taxonomic and (with EAF) functional profiles and host metadata associations. JL-P designed and executed the Gaussian process models. AM, MGG, HHC, JC, JO, CM, and AB performed assembly and annotation. JL-P, GR, EAF, BH, AM, AB, MGG, OW, and CH drafted the manuscript. CH, OW, and RK designed the study. All authors discussed the results and commented on the paper.

Competing financial interests
The authors declare no competing financial interests.

Introduction

The human microbiome is an integral component in the maintenance of health[1,2] and of the immune system[3,4]. Population-scale studies have aided in understanding the functional consequences of its remarkable inter-individual diversity, the earliest of which include MetaHIT[5,6] and the Human Microbiome Project[1] (referred to here as HMP1). Studies continue to focus on the gut[7–9], with fewer population-scale cohorts investigating vaginal[10], oral[11], or skin[12] microbial communities. HMP1 remains the largest body-wide combined amplicon and metagenome survey of the healthy microbiome to date.

Here, we report on an expanded dataset from the HMP (HMP1-II), consisting of whole-metagenome sequencing (WMS) of 1,631 new samples from the HMP cohort[13] (for a total of 2,355; Extended Data Fig. 1a, Extended Data Table 1a, Table S1). New samples greatly expand the number of subjects with sequenced second and third visits, and primarily target six body sites (from 18 total sampled): anterior nares, buccal mucosa, supragingival plaque, tongue dorsum, stool, and posterior fornix. After quality control (**Methods**), the dataset consists of 2,103 unique metagenomes and 252 technical replicates, which were used in all following analyses. Profiles, raw data, and assemblies are publicly available at http://hmpdacc.org (Extended Data Table 1b).

Results

Body-wide strain diversity and ecology

The diversity and spatiotemporal distributions of strains were first investigated using StrainPhlAn[14] (Fig. 1), which identifies the dominant haplotype (“strain”) of each sufficiently abundant species in a metagenome (**Methods**; Table S2). Most previous culture-independent strain surveys have targeted only the gut[15,16], and body-wide phylogenetic distances (quantified using the Kimura two-parameter distance[17]) suggest that all other habitats possess significantly higher strain diversity (Fig. 1a). Consistent with previous observations[15,18], strain profiles were stable over time, with differences over time consistently lower than differences between people (Fig. 1a, 1b). Nevertheless, technical differences were even lower, indicating a baseline level of intra-individual strain variation over time (Extended Data Fig. 2b).

Several species exhibited differentiation into body site-specific subspecies clades (Fig. 1c; Extended Data Fig. 2e–u), defined here as discrete phylogenetically-related clusters of strains, according to a Silhouette-based score of niche-association (**Methods**). This is readily visible in extreme cases, such as *Haemophilus parainfluenzae* (Fig. 1d), where distinct subspecies clades are apparent in the supragingival plaque, buccal mucosa, and tongue dorsum. Other species with notable site-specific subspecies clades included *Rothia mucilaginosa, Neisseria flavescens*, and a *Propionibacterium* species. Other species did not sub-speciate within body sites, but instead specialized in clades differing among individuals (e.g. *Eubacterium siraeum*, Fig. 1e, or *Actinomyces johnsonii*, Extended Data Fig. 2d); others showed no discrete subspecies phylogenetic structure at all in this population (e.g. *Streptococcus sanguinis*, Extended Data Fig. 2u). Interestingly, no subspecies clades were found to be specific to either of the two cities in the study (Extended Data Fig. 2a), although geographically localized subspecies population structure has been observed in cohorts with greater geographic range[15].

Culture-independent strain profiling, in combination with the 16,903 NCBI isolate genomes used as references in this analysis[19], provided a new quantification[20] of how well-covered human microbial diversity is by these references (Fig. 1f). Well-sequenced species such as *Escherichia coli* (Extended Data Fig. 2c) and the lactobacilli showed little divergence from reference isolates. However, many prevalent and abundant species in the body-wide microbiome were strikingly diverged from the closest available reference genomes. Notable clades lacking isolate genomes representative of those in the microbiome included *Actinomyces* (Fig. 1b), *Haemophilus parainfluenzae* (Fig. 1d), *Eubacterium rectale*, several *Streptococcus* and *Bacteroides* species, which thus represent priority targets for isolation.

Due to improvements in methodology and reference genomes, new species-level taxonomic profiling included eukaryotes, viruses, archaea, and an additional 54 bacterial species in these metagenomes relative to HMP1 data[1]. The latter contained prevalent bacteria such as *Bacteroides dorei, B. fragilis, Alistipes finegoldii, A. onderdonkii*, and unclassified species of *Subdoligranulum* and *Oscillibacter*. The former included *Methanobrevibacter, Malassezia*, and *Candida* (Extended Data Fig. 1c, as well as several viruses: *Propionibacterium* phage in the anterior nares, *Streptococcus* phages in oral sites, and a *Lactococcus*-targeting C2-like virus in stool. Searching for co-occurrence patterns with non-bacterial species (Fisher’s exact test, presence/absence threshold of 0.1% relative abundance; Table S3), we found that *M. smithii* tended to co-occur with several Clostridiales species in the gut, including members of *Ruminococcus, Coprococcus, Eubacterium*, and *Dorea* (FDR < 0.1), reinforcing previous observations[21] and consistent with co-occurrence patterns of methanogens and clostridia in lean vs. obese individuals[22]. Prominent *Streptococcus* phages, which were the most abundant in the oral cavity, also co-occur with numerous *Streptococcus* species in oral sites, suggesting that the virus predominantly exists as a prophage, as observed previously[23].

Core pathways of the human microbiome

Strong prevalence (“coreness”) of a molecular function across niche-related microbial communities can be explained by 1) broad taxonomic distribution of the function (as in the case of essential housekeeping functions), or by 2) specific enrichment of the function among taxa inhabiting that niche (possibly because the function is selectively advantageous there). We investigated these mechanisms among core metabolic pathways of the human microbiome by functionally profiling all HMP1-II samples with HUMAnN2[24] (Fig. 2, Extended Data Fig. 3, Table S4; **Methods**). We focused on 1,087 metagenomes representing the first sequenced visit from each subject at the six targeted body sites. We considered a pathway to be “core” to a specific body site (niche) if it was confidently detected in >75% of individuals with strong taxonomic attribution and a taxonomic range consistent with the human microbiome. From a starting set of 857 quantifiable MetaCyc[25] pathways, we detected 950 instances of a pathway being core to a body site: 258 pathways were core to at least one body site, 176 were core to body sites from multiple body areas, and 28 were core to all six targeted body sites (Fig. 2a and Extended Data Fig. 3a). For convenience, we refer to these classes as “core” pathways, “multicore” pathways, and “supercore” pathways, respectively.

To distinguish between coreness resulting from broad taxonomic distribution versus niche-specific enrichment, we classified pathways according to their taxonomic range (quantified as the fraction of non-human-associated genera to which they were annotated in BioCyc). While the majority of pathways were annotated to <10% of genera, core pathways were annotated to 34% of genera, multicore pathways to 48%, and supercore pathways to 70% (median values; all enrichments over background had $P \ll 0.001$ by Wilcoxon rank-sum tests). Thus, coreness to a human body site is often associated with broad taxonomic distribution, and pathways that are core to more body sites tend to be more broadly distributed (Spearman’s r=0.40; $P \ll 0.001$; Extended Data Fig. 3b). Extreme examples included coenzyme A biosynthesis (see Fig. 2a) and adenosine nucleotides biosynthesis (Extended Data Fig. 3e): two “housekeeping” functions that are broadly distributed across not only the human microbiome, but all microbial life[26,27]. While we lack dispensability information for entire MetaCyc pathways, we found that individually essential gene families were considerably more prevalent than non-essential families across these samples (median 0.94 vs. 0.24; Wilcoxon rank-sum test, $P \ll 0.001$; **Methods**), consistent with essential functions being core to many body sites.

On the other hand, 19 of the 176 multicore pathways (including 2 supercore pathways) were confidently *not* broadly distributed, defined conservatively as 1) annotated to <10% of non-human-associated genera in BioCyc and 2) reconstructed from <10% of pangenomes in the HUMAnN2 database (Extended Data Fig. 3a and 4c). In these cases, coreness to multiple human body areas is better explained by enrichment among human-associated taxa, and may be indicative of functional adaptation to the human host as a broader niche. Notably, of these 19 pathways, 13 (68%) were more than two-fold enriched in human-associated genera relative to non-human-associated genera in BioCyc, though this was not required by their definition. Human microbiome-enriched pathways included vitamin B12 biosynthesis (adenosylcobalamin salvage from cobinamide), a process commonly performed by the microbiota that must be supplemented in germ-free mice (Fig. 2b). Vitamin B12 biosynthesis was also core in the oral cavity, where salivary haptocorrin may protect it for later absorption in the small intestine[28]. Fermentation to propionate (a short-chain fatty acid, SCFA), was also specifically enriched in the oral and gut environments (Extended Data Fig. 3f). SCFAs are noteworthy for their proposed role in the maintenance of gut health[29], while their role in the oral cavity is less well-studied.

Finally, a number of core pathways were specifically enriched in individual body sites. We identified a single site-enriched core pathway from the anterior nares, 7 from the oral body area (notably, there were few that were enriched for a single oral site), 10 from stool, and 3 from posterior fornix (Extended Data Fig. 3d). Examples of site-enriched pathways included nitrate reduction in the oral cavity (a known oral microbiome process related to nitrate accumulation in saliva[30]; Fig. 2c) and mannan degradation in the gut (mannan is a plant polysaccharide found in human diet[31]; Extended Data Fig. 3g). Such site-enriched pathways are suggestive of functional adaptation by the microbiota to a particular niche within the human body. Hence, while many core functions of the human microbiome reflect broadly distributed, globally essential metabolic processes, others are potentially indicative of microbial community adaptation to specific body sites or to the human host in general.

Characterization of temporal variability

The new availability of body-wide WMS samples at multiple time points per individual allowed us to further characterize the dynamics of microbial community composition at the species level (Fig. 3). Community-wide species retention rates were comparable to previous observations at all body sites but the posterior fornix[32,33] (Fig. 3a). To characterize the dynamics of individual species, we developed a Gaussian process model (**Methods**) that decomposed variability in abundance into four components: constitutive differences between subjects, time-varying dynamics (change measurable at a scale of several months), biological noise (true variation that appears instantaneous relative to our sampling), and technical noise (between technical replicates).

This analysis indicated which species at which body sites varied most between individuals, temporally, or rapidly (Fig. 3b; Table S5; Extended Data Fig. 4d–f). In the gut, Bacteroidetes species, and in particular the *Bacteroides* genus (Extended Data Fig. 5), exhibited primarily inter-individual variation, while Firmicutes were more temporally dynamic within individuals. Species abundances in the oral and skin microbiomes, meanwhile, exhibited greater time-varying dynamics and biological noise overall, and were less personalized, consistent with previous stability assessments[18]. A more detailed look (Extended Data Fig. 5) showed that some species possessed very similar dynamics when detected in multiple body sites (e.g. *Rothia dentocariosa*). Others, often those with site-specific subspecies clades analyzed above, possessed different dynamics between body sites (e.g. *Haemophilus parainfluenzae*). At a broad scale, these species dynamics are in agreement with a previous analysis of whole-community dynamics in the same cohort[34].

We repeated this Gaussian process analysis to characterize the dynamics of pathway abundances for all core pathways identified above (Fig. 3c; Table S5). Pathway abundances at all body sites but the posterior fornix were less personalized than the taxa that encoded them (farther from the inter-individual vertex), consistent with the hypothesis that community assembly is primarily mediated by functional niches rather than a requirement for specific organisms[35,36]. Time-varying pathways were enriched for amino-acid biosynthesis (P=0.00025; Wilcoxon rank-sum test), while inter-individual pathways were enriched for B-vitamin biosynthesis (P=0.00062). In contrast, the vaginal microbiome showed a large personal component, at both the species and pathway levels (all well-fit pathways near the inter-individual vertex), consistent with variation among stable community state types in the vaginal microbiome[37]. Functional dynamics in the gut were relatively slow, possibly reflecting trends in response to long-term factors such as dietary patterns. Conversely, dynamics in oral cavity sites were rapid, in particular in the buccal mucosa, concordant with the habitat's enrichment for fast energy harvest and much greater environmental exposure.

Gene family discovery by assembly

We next sought to establish an expanded gene catalog based on assembly of the expanded set of metagenomes. Based on extensive benchmarking, we chose a custom assembly protocol using IDBA-UD[38] (**Methods**). Compared to the 725 assemblies generated in HMP1[1,13], this protocol led to improvements in average assembly size, median contig length, and N50 (Table S6). Median metagenome assembly sizes ranged from 2.9 Mb for posterior fornix to 127.6 Mb for stool. To help detect new genes and improve overall assembly quality, we created additional co-assemblies from the combined set of reads from the same individual sampled at the same body site across multiple visits. In total, 406 and 240 co-assemblies were created by combining two and three visits, respectively (Table S6), and the assembly sizes were on average 86% larger than single assemblies: the median assembly size increased from 84.8 Mb to 158.4 Mb, and the median of the maximum contig size in each assembly increased from 152 Kb to 167 Kb (Fig. 4a–c). Gene finding was performed on contigs using MetaGeneMark[24] (Fig. 4d; Table S7). In co-assemblies, the average number of genes detected increased from 118,177 to 213,741 while the mean gene length remained similar (from 614 to 610 nucleotides). Functional assignments were made using Attributor (**Methods**) based on several sequence-based searches, and classified according to specificity. Approximately 35–45% of genes received specific functional annotations, while another ~30% received annotations at the domain, family, or motif level (Extended Data Fig. 6). In all cases, the number of genes in each specificity category increased in the co-assemblies, while the percentages remained similar. Thus, although more genes were predicted from the co-assemblies, their annotations are as specific as in single assemblies.

The number of distinct, well-covered Pfam[39] domains detected by reference- versus assembly-based profiling tended to correlate strongly within-sample (Spearman's r=0.92; Extended Data Fig. 7d), suggesting that the two methods provide similar relative rankings of community functional diversity. In addition, the two methods tended to co-detect most Pfam domains that were core to a body site (>75% prevalent; Extended Data Fig. 7e). While reference-based profiles called Pfam domain presence based on the annotations of characterized proteins, Pfam domains could be directly detected in assemblies through profile alignment, thus capturing novel sequence diversity. Indeed, assembly tended to detect (median) 19% more Pfam domains per sample than the reference-based approach, which conversely tended to detect established Pfam domains with greater sensitivity. This effect was particularly notable in the anterior nares site, where reduced microbial sequencing depth limited the sensitivity of assembly relative to reference-based profiling.

Compared to external datasets, total non-redundant gene clusters were similar to MetaHIT in the stool[6] (HMP1-II contained 7,780,363 gene clusters, MetaHIT 9,879,896); relative to existing moist skin site metagenomes[12], HMP1-II represented a 780% increase (170,206 gene clusters to 1,326,693). However, even with thousands of deeply sequenced human microbiomes in this study, microbial gene family space is not yet saturated for any of the seven examined body sites (Fig. 4e).

Conclusions

Here, we provide and analyze the largest body-wide metagenomic profile of the human microbiome to date. The associated deep, longitudinal shotgun sequencing has enabled a broad-scale characterization of new aspects of the personalized microbiome. New strain profiling techniques[14] distinguished temporally-stable subspecies population structures for a number of species, some unique to individuals and others associated with particular body sites. Species with human microbiome strain diversity under-represented in isolate genomes were identified, to be prioritized for isolation and sequencing. New taxonomic profiling resolved co-occurrence patterns between bacterial abundances and several archaea, eukaryotes, and viruses. New functional profiling methods[24] identified pathways required for microbial colonization of the human body, differentiating those enriched for the human habitat from those universal to microbial life. Gaussian process models characterized microbial and functional variation over time, and identified gut community composition (Bacteroidetes species in particular) as highly personalized compared to other sites. This example implies that the gut Bacteroidetes/Firmicutes balance may not be a defining attribute of an individual's gut microbiome; individuals instead carry a "personal equilibrium" among Bacteroidetes, with a group of phylogenetically diverse, temporally variable Firmicutes fluctuating atop this core.

A number of key properties of the human microbiome remain to be characterized even in healthy cohorts, in addition to microbiome contributions to disease. Further investigation will be required to determine the functional origins and consequences of subspecies structures identified here. Such structures must also be investigated comprehensively across populations, including variations in geography, genetic background, ethnicity, and environment (e.g. outside of the HMP1-II's North American focus). Notably, the evidence in this study suggests that even in this relatively homogeneous population with extensive metagenomic sampling, the full complement of extant microbial genes has not yet been sequenced. Likewise, although an updated covariation analysis between metadata and microbial features (Supplementary Note, Extended Data Figs. 8 and 9) revealed several novel associations, most variance in the microbiome is not explained by measured covariates. The HMP1-II, for example, did not measure transit time[8], immune status, or the participants' detailed diet and pharmaceutical history, limiting our ability to assess these important factors. Finally, our understanding of the dynamics and responses of microbial communities must be expanded from the descriptive models here to include the rapid effects of acute perturbations. For this, studies with longer, more densely-sampled time courses in the presence of controlled perturbations will be required, beyond the three time points used here. In order to rationally repair a dysbiotic microbiome, it is thus necessary to deepen our understanding of the personalized microbiome in human health.

Methods

HMP1-II samples and metagenomic sequencing

Sample collection, storage, handling, and WMS sequencing were performed as in the HMP1[1]. Details on subject exclusion criteria, the sampling protocols, and timeline can be found in previous publications[1,13,40]. WMS reads and accompanying metadata are available at the SRA and dbGaP under two studies: SRP002163 (BioProject PRJNA48479), containing 11,245 runs and a total of 22.8 Tbp, and SRP056641 (BioProject PRJNA275349), with 312 runs and 1.2 Tbp. All metagenomes analyzed here were obtained from the SRA after human DNA removal by the SRA using BMTagger (Extended Data Fig. 7a). All SRA native format read files were converted to FASTQ for further analysis using the fastq-dump utility from the SRA SDK toolkit[19].

Quality control of nucleotides, reads, and samples

One or more SRA read files from each sample were concatenated per read direction to create a single pair of FASTQ files for each sample. These FASTQs were converted to unaligned BAM using Picard (http://picard.sourceforge.net) and exact duplicates were removed with a modified version of Picard's EstimateLibraryComplexity module. Finally, all reads were trimmed and length filtered (−q2 −l60) using the trimBWAstyle.usingBam.pl script from the Bioinformatics Core at UC Davis Genome Center[1].

After taxonomic profiling (below), ecologically abnormal WMS samples were identified for further per-sample quality control based on median species-level Bray-Curtis dissimilarity to other samples from the same body site. If a sample's median dissimilarity exceeded the upper inner fence (1.5 times the interquartile range above the third quartile) for all median dissimilarities from its body site, the sample was labeled an outlier and discarded. This process removed 86 (3.6%) WMS samples that were highly atypical for their respective body sites. Downstream analyses utilized the remaining 2,355 samples.

Taxonomic and strain profiling

Taxonomic profiling of the metagenomic samples was performed using MetaPhlAn2[20], which utilizes a library of clade-specific markers to provide pan-microbial (bacterial, archaeal, viral, and eukaryotic) profiling (http://huttenhower.sph.harvard.edu/metaphlan2). MetaPhlAn2 profiles recapitulated observed ecological patterns from HMP1 (Extended Data Fig. 1b), and agreed with direct read mapping to reference genomes. Mapped reads covered an average of 81.7% (median 92.8%) of the reference genomic sequence of each modestly-dominant strain (comprising at least 5% of the community) across all samples. Mean coverage depth (total base pairs in aligned reads divided by total base pairs in reference genome) for these strains over all samples was 3.9×, with depth-of-coverage means varying widely by body site from 0.04× (right antecubital fossa) to 11.1× (tongue dorsum) (Table S8). Batch effects were not visible in the first two axes of variation within each body site (Extended Data Fig. 1d).

Strain characterization was performed using StrainPhlAn[14]. StrainPhlAn characterizes SNVs in the MetaPhlAn2 marker genes for an organism. For a given sample, we required minimum of 80% of markers for a given species to have a minimum mean read depth of 10×, to ensure sufficient data to perform haplotype calling. In total, 151 species satisfied these requirements in at least two WMS samples (Table S2). Distances between strains were

[1] https://github.com/genome/genome/blob/master/lib/perl/Genome/Site/TGI/Hmp/HmpSraProcess/trimBWAstyle.usingBam.pl

assessed using the Kimura 2-parameter distance[17] (available from Extended Data Table 1b). Both MetaPhlAn2 and StrainPhlAn were used with their default settings.

Reference genome coverage was scored by the complement of the asymmetric phylogenetic distance (1 - Unifrac G[42]) between HMP1-II strains and reference genomes. All coverage estimates are presented in Table S2.

Niche-association score

Species with niche-associated subspecies clades were detected by a measure similar to the silhouette score, which compares the mean phylogenetic divergence of strains within each body site to the divergence of strains (within the same species) spanning body sites. Specifically, we first define a body site dissimilarity score $D(u, v)$ for a given species at body sites u and v as:

$$D(u,v)=\frac{1}{|S_u|}\sum_{i\in S_u}\frac{\beta_{i,v}-\beta_{i,u}}{\max(\beta_{i,u},\beta_{i,v})},\beta_{i,v}=\frac{1}{|S_v\backslash i|}\sum_{j\in S_v,i\neq j}d(i,j)$$

where S_x is the set of samples which pass the StrainPhlAn coverage requirements in body site x, and $d(i, j)$ is the Kimura 2-parameter distance between dominant haplotypes in samples i and j. The niche-association score A for each species (Fig. 1b) was then defined as the maximum observed $D(u, v)$ over all directed pairs of body sites u and v where the StrainPhlAn coverage requirements were met for at least 5 samples in both sites. That is, for a set of body sites B:

$$A=\max\{D(u,v),u\in B,v\in B,u\neq v,|S_u|\geq 5,|S_v|\geq 5\}$$

One concern with this score is that greater technical difficulty in SNV calling in one site may result in apparent niche-association where there is none. This is not a concern here, however, since all sites where the niche-association score was calculated were oral sites with similar technical variability (Fig. 1a). This is a by-product of the limitation that the species were required to have a significant presence (5 samples with sufficient coverage) at multiple sites, which was not possible outside of the ecologically more similar set of oral sites.

Functional profiling

Functional profiling was performed using HUMAnN2[24] (http://huttenhower.sph.harvard.edu/humann2). Briefly, for a given sample, HUMAnN2 constructs a sample-specific reference database from the pangenomes of the subset of species detected in the sample by MetaPhlAn2 (pangenomes are precomputed representations of the open reading frames of a given species[43]). HUMAnN2 then maps sample reads against this database to quantify gene presence and abundance on a per-species basis. Remaining unmapped reads are further mapped by translated search against a UniRef-based protein sequence catalog[44]. Finally, for gene families quantified at both the nucleotide and protein levels, HUMAnN2 reconstructs pathways from the functionally characterized subset and assesses community total, species-resolved, and unclassified pathway abundances based on the MetaCyc pathway database[45].

Analyses of metabolic pathway coreness were focused on 1,087 HMP1-II metagenomes representing the first sequenced visit from each subject at the six targeted body sites. Follow-up samples and technical replicates for a given (subject, body site) combination were excluded to avoid biasing population estimates in their direction. We defined a "core" pathway at a particular body site as one that was detected with relative abundance $>10^{-4}$ in at least 75% of subject-unique samples. We further filtered these highly prevalent pathways to ensure sensible taxonomic range and confident taxonomic attribution. Specifically, a potential core pathway was excluded 1) if its BioCyc[45]-annotated taxonomic range did not include any human-associated microbial genera (defined as genera detected in at least 5 HMP subjects with relative abundance $>10^{-3}$) or 2) if >50% of pathway copies had "unclassified" taxonomic attribution in >25% of samples. These filtering criteria yielded a total of 950 core (pathway, body site) associations covering 258 unique MetaCyc pathways. Notably, these numbers were reasonably insensitive to the exact parameter settings described above, provided that the overall definition of "coreness" encompassed 1) a majority population prevalence (i.e. >50%), 2) a non-extreme detection threshold [i.e. below (number of pathways)$^{-1}$], and 3) some form of taxonomic filtering to limit false positives (e.g. to rare variants of otherwise common pathways; Table S9).

We quantified the taxonomic range of a pathway as the fraction of unique genera to which it was annotated in BioCyc. We subdivided this measure into ranges over "human-associated" and "non-human-associated" genera (as defined above), and focused on the latter measure to avoid circular reasoning (a function that is broadly distributed across human-associated taxa would be enriched in the human microbiome by definition). As a further control, we also directly applied HUMAnN2 to its underlying pangenome database to associate pathways with >4,000 microbial species. To conservatively define core pathways as "enriched to the human microbiome," we required them to be annotated to <10% of non-human-associated genera in BioCyc, and also directly annotated to <10% of non-human-associated pangenomes. The second criterion further reduced cases of rare variants of common pathways (as defined by MetaCyc) being called as enriched in metagenomes due to cross-detection of the common pathway.

We defined a core pathway to be strongly enriched in a particular body site if the first quartile of the pathway's abundance at that site was >2× larger than the third quartile of abundance at sites from all other body areas (i.e. the focal and background abundance distributions must be very well separated, as opposed to just significantly different). Notably, this definition only requires core pathways at oral body sites to separate well from non-oral sites, and not other oral sites (very few core pathways at oral body sites were strongly enriched relative to other oral sites).

We investigated the relationship between coreness and essentiality of functions using a dataset of ~300 essential COG[46] gene families determined in *E. coli*[47] (the "Keio collection"). We computed COG abundance across the 1,087 metagenomes introduced above by summing the abundance of individual UniRef gene families (as computed by

HUMAnN2) according to UniProt-derived COG annotations[48]. We considered a COG to be confidently detected in a sample if its relative abundance exceeded 10^{-4}. Among detected COGs, essential COGs (n=272) were both more globally prevalent than non-essential COGs (n=3,629; median 0.94 vs. 0.24) and core to more body sites (mean 4.7 vs. 1.2; “core” defined here as >75% prevalent within-site); both trends were highly statistically significant ($P \ll 0.001$) by Wilcoxon signed-rank tests and robust to a smaller detection threshold (10^{-6}).

Gaussian Process dynamics modeling

Gaussian Processes (GPs) are a versatile nonparametric probabilistic model for performing inferences about sampled continuous functions. This section covers the justification of the specific GP model used to model microbial and functional abundances (referred to here as “features”) in the microbiome, and discusses its assumptions, advantages and drawbacks. Implementation details are presented in the following section.

In a GP, the joint distribution of the modeled function at any finite set of points follows a multivariate-normal distribution. Without loss of generality, GPs can be parameterized solely by their covariance function or kernel, defining the covariance of the output between any pair of sample points. This pairwise definition permits the use of the irregular temporal sampling present in the HMP1-II dataset (Extended Data Fig. 4a). The shape of the GP’s covariance function determines a number of properties of the modeled function, such as its smoothness, how quickly it changes, and which features of the input vector it is sensitive to. Our first goal here was thus to assess the strength of the evidence for several common covariance functions describing biologically meaningful behaviors, and to determine what components should be included in a parsimonious model that captures the observable dynamics for the majority of features. The candidate set of covariance functions we considered includes: fast variation (“biological noise”), inter-individual differences, an Ornstein-Uhlenbeck process (OU), a squared-exponential covariance function (SE), and seasonal dynamics with a period of 1 year (formulas can be found in Table S10).

All candidate covariance functions describe stationary processes, given the inherently limited state space of relative abundances, though they differ significantly in their temporal dynamics and in their implications for the biological systems that generate these behaviors. Fast variation, i.e. variation on a timescale faster than measurable, is represented by a Gaussian white noise process. Inter-individual differences are modelled by constant covariance between samples from the same person. The two time-varying components, OU and SE both describe monotonically decreasing covariance as the difference in time between two samples increases, i.e. time points closer to one another will be more similar than those farther apart. These two functions primarily differ in the smoothness of the underlying function. The OU process is the only stationary Markovian GP with non-trivial covariance over time, and produces functions which are not differentiable, and thus very jagged, resembling Brownian motion. For example, this covariance function is expected for a slowly-changing feature’s abundance under continuous stochastic perturbation from the environment (e.g. random day-to-day dietary choices). Meanwhile, the SE function describes functions which are infinitely differentiable, and are thus extremely smooth. This function implies a significant amount of latent state relevant to the process generating the feature's abundance. Both of these time-varying covariance functions are parameterized by their lengthscale, the characteristic timescale at which the function changes. Lastly, the seasonal component is represented by the canonical periodic covariance function from GP literature, with its period fixed at 1 year, but with an unknown lengthscale. Here, a model refers to a combination of these covariance functions.

Models were compared based on their marginal likelihoods (also termed "evidence"), reported in bits (i.e. $\log_2$ ratio of marginal likelihoods = $\log_2$ Bayes factors) of evidence against a given model when compared to the best model for a feature (Table S10). More than 3.3 bits is considered strong evidence against a model, while >6.6 bits is considered decisive. Marginal likelihoods were estimated from MCMC samples of the posterior distribution by a truncated harmonic mean of the un-normalized posterior distribution at the sampled points. Truncation was performed since this estimator is known to have poor convergence characteristics since MCMC samples with very low likelihoods have an unreasonable influence on the harmonic mean. Comparisons were performed for models fit to the abundances of the top 10 most prevalent species (with at least 70% non-zero abundances) and top 5 most abundant pathways at each targeted body site (Table S10). Comparisons were also performed for a set of simulated features with known dynamics ("controls"), which were sampled from the corresponding GP with 5% of variance due to technical noise and remaining variance distributed evenly between components.

To determine which of these components have statistical support in the data, we employed a standard greedy search through the space of possible models, which starts from the simplest model (all variation is technical), and iteratively rejects simpler models in favor of a more complex one if the evidence against the simpler model exceeded 6 bits. The set of more complex models considered at each iteration are those with only one more parameter, and contain the simpler model as a special case (pseudocode presented in Table S6). This procedure selected models which included the two simplest components, biological noise and inter-individual differences, 47 and 53 times among the 72 features tested, respectively. Among more complex components, the OU component was selected 13 times, while neither the SE nor the seasonal component were selected for a single tested feature. These trends were robust to increases in the model rejection threshold, with the evidence for the OU component remaining significant to at least 10 bits, while the SE and seasonal components are only selected for more lenient thresholds (≤4 bits). We note, however, that this procedure had difficulty identifying SE and seasonal components in control samples which included other components (in particular, biological noise), indicating that these components are difficult to distinguish given the available temporal sampling pattern. Thus, while the data clearly currently prefers OU over SE and does not support the inclusion of a seasonal component, we are not sufficiently powered to eliminate these as potentially significant contributors to the dynamics of the microbiome. Finally, the null model with only technical noise was rejected for 71 of the 73 features, often with very high evidence (median 69.6 bits).

For the remainder of the analysis, we thus converged on a model with four components: inter-individual differences, an Ornstein-Uhlenbeck process, biological noise, and technical noise. Let U, T, B, and N be the respective magnitudes of these components, and I be the timescale of the OU process. Estimation of these parameters (hyperparameters in GP nomenclature) was performed by fitting a GP with the following covariance function to all features (species and pathways) with at least 75% prevalence within a site (Fig. 3, Table S5):

$$k(i,j) = \left[U + \mathrm{T}e^{\frac{-|t_i - t_j|}{l}} + B \cdot (t_i = t_j)\right](s_i = s_j) + N \cdot (i = j)$$

All 4 parameters were fit simultaneously by MCMC (next section). Since the three magnitude components must sum to the variability of the population, this can be seen as a decomposition of variance into sources of variability which differ in their temporal signature. Since we are interested only in the three biological components here, we therefore normalize out the estimated technical noise component (i.e. [*U, T, B*] / *N*) before visualizing the decomposition on a standard ternary plot (Fig. 3b–c). For illustration, we show three examples which illustrate the three types of dynamics on a plot designed to allow a direct comparison between the data and the fit GPs (Extended Data Fig. 4d–f).

The identifiability of any component of a time-dependent model is limited by the temporal sampling pattern available. The current dataset contains only up to 3 time points per person, with the time between samples roughly evenly distributed between 1 month and 1 year for each body site (Extended Data Fig. 4a). Processes too fast to measure will contribute to the biological noise component, while processes much slower than the maximum time intervals available contribute to the inter-individual component. We tested what timescales would be detected by the OU component, and which would contribute to the inter-individual or biological noise components, by simulating data from OU processes of varying lengthscales and performing parameter fits (Extended Data Fig. 4b). These show that the time-varying component is sensitive to processes with characteristic lengthscales of ~3–24mo.

We note that the resolution of the time-varying component is only possible due to the large spread in the time differences between samples available in the HMP1-II dataset (Extended Data Fig. 4a). In another common longitudinal study design, where a small number of samples are gathered per person with a fixed time interval between them, this would not be possible, though this design may make the analysis simpler (samples can be grouped by time point and a method like GPs would not be necessary). Likewise, richer longitudinal data in the form of longer time series would allow even more to be inferred about the dynamics of the microbiome. Of particular interest, this would enable differences in the temporal component(s) to be resolved between people. Here, with only up to 3 time points per person, the fit model parameters describing temporal changes (*B, T,* and *l*) are only a best-fit over the population. Such a sampling pattern would also provide the opportunity to more conclusively differentiate between the Markovian OU process and other possible non-Markovian processes (such as described by the SE covariance function, or an intermediate like the Matérn covariance functions), indicative of latent state or time-delayed events in the microbiome.

The HMP1-II dataset also includes a number of technical replicates (252 in total), which were instrumental in distinguishing the two fast-varying components (“biological” and technical noise). We encourage the addition of a non-trivial number of technical replicates in future longitudinal studies, not simply for validation but also to allow a quantitative characterization of diversity that is not captured in the remainder of the experiment due to limited sampling rates. Since technical noise is also estimated with the other variance components, estimates of the relative magnitude of the technical noise are also reported (Table S5). The proportion of variance due to technical noise was generally lower for species abundances (median of 5.4%, 90th percentile of 19.3%) than for pathways (median of 16%, 90th percentile of 44%), consistent with the observation that true biological variation between pathway abundances is lower than between species abundances[1]. Noise levels in pathways were predominantly influenced by body site, with pathways in the anterior nares having the greatest noise (median of 40%).

We assessed how accurate the parameter fitting process was under these noise conditions by simulating samples from mixtures of the three components and performing parameter fits for each targeted body site (Extended Data Fig. 4c). For all noise levels, pure components were always inferred with high confidence, with inter-individual differences being the most identifiable. Mixtures of inter-individual dynamics with biological noise were also confidently recovered, while mixtures of inter-individual and biological noise were more variable, and mixtures of inter-individual and time-varying dynamics were biased towards a greater influence of time-varying dynamics. Thus, when the time-varying component is present, parameter estimates should be considered biased away from the inter-individual corner of the ternary diagram. Mixtures of all three components had the greatest uncertainty. Among body sites, inferences at the anterior nares and posterior fornix sampling distributions were the most unreliable, due to the relatively limited number of samples at these sites (Extended Data Fig. 4a), reflected as a large number of highly uncertain features at these sites (Fig. 3). At 20% technical noise (the 90th percentile of the noise distribution for species), parameter estimates degrade noticeably, and tend towards the mean of the prior (an even mixture of all components). This thus results in the low-confidence species and pathways tending to locate towards the center of the ternary diagrams (Fig. 3).

We note that for a particular feature (microbe or pathway abundance), each of the non-technical components represents the sum of all processes with that temporal signature which affect that feature, and these do not necessarily reflect intrinsic properties of the feature. Examples of extrinsic processes likely to produce biological noise include, among others, day-to-day dietary differences, the timing of sample collection relative to meals, tooth-brushing and other personal hygiene, spatial variation of the microbiome within subjects (e.g. gradients across the stool), and weekend/workday differences. Extrinsic sources of inter-individual differences may arise from culture/ethnicity (ethnicity is strongly associated with the abundances of several microbes[1]), differences in habits (e.g. habitual vs. infrequent tooth brushers and flossers), and long-term dietary differences, among others. Finally, time-varying processes may include properties such as weight or slowly changing preferences in diet.

Gaussian Process parameter optimization details

All parameter fits and model comparisons were performed by MCMC sampling with the GPstuff toolbox version 4.6 in MATLAB. Before fitting, relative abundances were first arcsine-square root transformed, filtered for outliers using the Grubbs outlier test (significance threshold 0.05), and standardized to have zero mean and unit variance. A Gamma-distributed prior with shape 3.1 and mean 10 months was imposed on the lengthscale parameter of all time-varying components. These parameters for l were selected based on the intervals between samples, and guarantee that the model is identifiable when the biological noise and/or inter-individual difference components are included by ensuring that l is neither too short nor too long. All parameters of all models were fit simultaneously. All models were fit using a Gaussian likelihood. This function performs poorly for significantly non-Gaussian distributions, which frequently occur in microbiome data in the form of zero-inflated abundance distributions. For this reason, the dynamics analysis was performed for highly-prevalent features (species with ≥75% prevalence within a site, and core pathways). One exception was made for this: species with mean abundance when present ≥2% and non-zero in at least 50 samples were also included, so as to include important species such as *Prevotella copri* with lower prevalence but exceptional abundance when present. Other models specifically accounting for zero-inflation (both technical and real) will be needed to study the dynamics of the rarer microbiome.

Evidence presented in Table S5 was calculated from 5 MCMC chains per model, with 150 samples after 20 sample burn-in, which were started from a random point in the prior distribution. Parameter estimates presented in Fig. 3 and Table S5 were fit with the additional constraint that $U + T + B + N = 1$, to eliminate an additional degree of freedom from the model. A Dirichlet(1, 1, 1, 1) prior was imposed on $[U, T, B, N]$. For each feature tested here, a more thorough MCMC sampling was performed than for the model selection, consisting of 10 chains with 200 samples each (after 30 burn-in and thinning every other sample), starting from a random point from the prior distribution. In all cases, all parameters were fit simultaneously. Convergence was assessed with the $\hat{R}$ statistic[49]. Over all 196 species and 950 pathways tested, 97% of $\hat{R}$ statistics were <1.1 for all parameters (median 1.01, max 1.17), indicating good convergence.

Association testing between microbiome features and phenotypic covariates

Associations between microbial and pathway abundances and metadata were determined using MaAsLin[1,50]. MaAsLin tests a sparse multivariate generalized linear model against each feature independently. Relative abundances were first arcsine square-root transformed for variance stabilization, and the Grubbs test was used (significance level 0.05) to remove outliers. A univariate prescreen was applied using boosting to identify potentially associated features, and significantly associated covariates among the remaining features were identified with a multivariate linear model without zero-inflation. Unless otherwise stated, a final false discovery rate <0.1 (Benjamini-Hochberg controlled across feature tests) was used as a significance threshold.

The same model was applied to all features (microbial and pathway) during this analysis and included the following covariates: broad dietary characterization, whether the subject was breastfed, temperature, introitus pH, posterior fornix pH, gender, age, ethnicity, study day processed, sequencing center, clinical center, number of quality bases, percent of human reads, systolic blood pressure, diastolic blood pressure, pulse, whether the subject had given birth, HMP1/HMP1-II, and BMI. A summary of these metadata can be found in Extended Data Table 1a. Of note, several recently identified confounders such as transit time[8] for stool samples were not collected during sampling.

Assembly

Benchmarking and assembly protocol design—We benchmarked several assemblers including IDBA-UD[38], MetaVelvet[51], SOAPDenovo2[52], Newbler (Roche, Basel, Switzerland), Ray[53], SPAdes[54], and Velvet[55] using eight samples (SRS017820, SRS014126, SRS052668, SRS017820, SRS048870, SRS020220, SRS057205, SRS017820) across five body sites that represented a range of metagenomic complexity. Based on the assembly size, median length, fragmentation level, and N50, we chose IDBA-UD to process all HMP1-II samples.

Digital normalization—Following quality control, sequence reads for each sample were run through a "digital normalization" pipeline prior to assembly. This process was designed to reduce, as much as possible, the volume of information from the most dominant source taxa (without sacrificing the ability to assemble what remains) so that lower-abundance taxa could be assembled more evenly, instead of having their reads discarded by the assembler software as not being sufficiently covered (compared to the dominant taxa).

Median k-mer coverage was first estimated for all reads using the khmer Python library[56]. These data were then used to filter input reads so as to normalize k-mer coverage within preselected bounds: for each k-mer of length 20 in each read, the total number of observations of the k-mer was used as a proxy for coverage. Reads where median k-mer coverage was already greater than 20 were discarded. Remaining reads were then trimmed at the first instance of a single-copy k-mer (representing putative error sequences). Reads whose post-trim length was less than the k-mer length (20) were also discarded. Surviving reads were trimmed again, this time at the first instance of a high-abundance (> 50×) k-mer; again, reads whose post-trim length was less than 20 nt were discarded. For remaining reads, we re-normalized (based on median k-mer coverage as in the first step) to remove all reads whose median k-mer coverage was >5×. This is a more aggressive filter on putatively redundant sequences, after elimination of initial reads with highly-overrepresented (redundant) k-mers or severely underrepresented (error) k-mers.

For subsequent assembly after this quality control and normalization, we increased k to 32 (to maximize sensitivity on the remaining reads) and built an overlap graph from all remaining reads. This graph was then partitioned into groups of reads with a high likelihood of internal overlap, separating components at precomputed “stoptags”: k-mer sequences automatically identified by khmer in its initial profiling scan as unreliable assembly-traversal nodes. Reads were then extracted from each such partition into separate FASTA files. Each partition was tested for more coherent subgroups, beginning with the least consistent (ranked in order of graph separability). Re-partitioning was carried out as above, but with more aggressive parameters: stoptags in the initially-computed overlap graph were explicitly detected and removed before re-partitioning (which included the generation of new stoptags from the remainder of the graph after removal of the earlier ones). The least-consistent read group was broken into sub-partitions exactly once in this way: further iteration risks overfitting and is not guaranteed to converge to a meaningful result.

IDBA-UD assembly and post-processing—Following digital normalization, each final partition was assembled independently of the others with IDBA-UD. For values of k in (20, 30, 40, …, 80), IDBA-UD will attempt to assemble its partition (via de Bruijn graph methods) using k-mers of size k, and will then merge and extend results from all passes to produce a final assembly of that partition (requiring a minimum contig length of 100 nt). For each sample (or pool), all (independently-produced) partition assemblies were then concatenated. As a final step to reduce any redundancy present in the final concatenated assembly, we merged and extended all assembled contigs (across all partitions), based on overlaps of 40 nt or more, to produce a final "consolidated" sequence collection.

Quality assessment—To assess assembly quality we undertook a number of post assembly QC checks including an examination of the rate at which reads aligned to assemblies as well as identifying chimeras, which are a potential problem caused by mis-assemblies.

To check for what portion of the reads were incorporated into the assembly, sample reads were aligned against their assembly using Bowtie v1, resulting in counts for reads with at least one alignment and for those that failed to align. Total reads include reads from the human host. Because human reads were masked as all Ns by SRA using BMTagger, the human reads would affect the portion of unaligned reads. To assess the effect, we counted the number of masked reads to get a count for human reads. These are summarized by body site in Extended Data Fig. 7c.

Assembly protocol validation—To examine the rates of chimeric contigs and mis-assemblies, we undertook an assembly assessment of two mock data sets generated during the HMP, one where the community was created with all 21 organisms in equal abundance (“even”), and one with staggered abundances. We assembled these mock communities using the same protocol and aligned assembled contigs for both sets against all 21 input genomes. We found that 94.21% and 96.84%, respectively, of all assembled contigs aligned uniquely to a single reference genome for the even- and staggered-coverage mock communities ("aligned" here means aligned with ≥ 95% sequence identity over ≥ 95% of their length). Contigs aligned to closely-related *Staphylococcus* and *Streptococcus* strains exhibited slightly more non-exclusive matching (or cross-matching) than contigs aligned to other strains. For the even set, an average of 97.85% of all *Staphylococcus*- and *Streptococcus*-aligned contigs were uniquely aligned to their reference strain, with an average of 92.98% for the staggered set, as compared to averages across all other strains of 99.89% (even) and 98.98% (staggered), neatly reflecting the inherent genetic ambiguity of these taxonomically narrow subgroups yet showing a very strong ability to distinguish between related strains.

Recovery statistics do not correlate well with input coverage in the staggered set, implying that our pipeline (given a minimum of 4× coverage) is robust against differences in relative abundance of up to 3 orders of magnitude at these scales. Phylogenetic proximity, in this case, seems to show a greater influence on uniqueness of assembly (albeit still a very weak one) than does coverage. Fractions of contigs not aligning to any of the 21 reference strains (over ≥ 95% of their length at ≥ 95% identity) were 5.6% and 3.0% for the even and staggered sets respectively; we can thus postulate these proportions to be upper bounds on the combined rates of chimeras and mis-assemblies produced by our pipeline, consistent with other chimera assembly metrics[57].

Annotation

Annotation of open reading frames (ORFs) within assembled contigs was performed using Metagenemark-3.25[58]. The resulting ORF sequences were used as input for searches against a) Uniref100[59] using RAPSearch2[60], b) Pfam[61] and TIGRfam[62] HMM models using hmmer-3.0[63], c) TMHMM[64] for the identification of transmembrane helices, and d) a regular expression search for membrane lipoprotein lipid attachment sites for the identification of putative signal peptides. The latter three searches were run as implemented in the Ergatis workflow monitoring system[65].

Annotation was assigned by Attributor (github.com/jorvis/Attributor) using a hierarchical scheme developed out of the IGS Prokaryotic Annotation Pipeline[66]. Attributor assigns common names, gene symbols, EC numbers and GO terms, as applicable, based on a hierarchy of evidence including hits to HMM models, Uniref100 sequences, TMHMM predicted helical spans, and lipoprotein motifs. Assignments are exclusive, meaning that for each ORF, Attributor takes the strongest piece of evidence available and assigns all attributes possible based on that evidence. Attributes are not assigned from multiple sources to ensure that annotation attributes assigned to a single ORF do not conflict. Attributor annotation was output as gff3 and FASTA files (Extended Data Table 1b).

Rarefaction curves

Rarefaction curves were generated by extracting predicted polypeptides from the MetaGeneMark output for each sample, and estimating a “unique gene family” count for rarefied sample size *n* as follows, using usearch v.8.1.1861 ×64[67]:

1. Concatenate the MetaGeneMark predicted polypeptides from a random sampling of *n* samples that were not technical replicates, eliminating duplicates.
2. Sort sequences by decreasing length
3. Cluster sequences at 90% identity (using usearch cluster_fast)
4. Retrieve the “unique gene family” count from the results

The number of unique clusters was estimated from 50 random subsets for each *n*. This procedure was repeated for each body site for *n* = 1,10,20,… until the number of unique samples available at the body site.

Read mapping to assemblies and reference genomes

Mapping reads to reference genomes—In addition to taxonomic and functional profiling as above, all samples’ individual raw reads were aligned directly to MetaRef[43] reference genomes. Before alignment, all reads with 80% or higher percentage of Ns were discarded using the Biocode fastq::filter_fastq_by_N_content utility[2]. Bowtie2[68] (v2.2.4) was then used to align reads to reference genomes using the default, paired-end alignment options and including the singleton reads. The resulting SAM files were converted to BAM, sorted, and then partitioned into two separate files per sample - one of only matching reads and the other of unaligned reads. This entire pipeline is encapsulated in the Biocode generate_read_to_metaref_seed_alignment.py pipeline script[3].

Mapping reads to assembled contigs—The quality-trimmed reads from each sample were mapped back onto the assembled contigs from that same sample using Bowtie (v0.12.9) with a 512MB max best-first search frames value, Phred33 score quality setting, 21bp seed length, and limit of 2 mismatches per seed. All alignments per read were reported (unless there were more than 20 for a given read) with hits guaranteed best stratum and ties broken by quality. Hits in sub-optimal strata were not reported.

Code Availability Statement

Code for the annotation pipeline and the Gaussian Process analysis are available from Extended Data Table 1b.

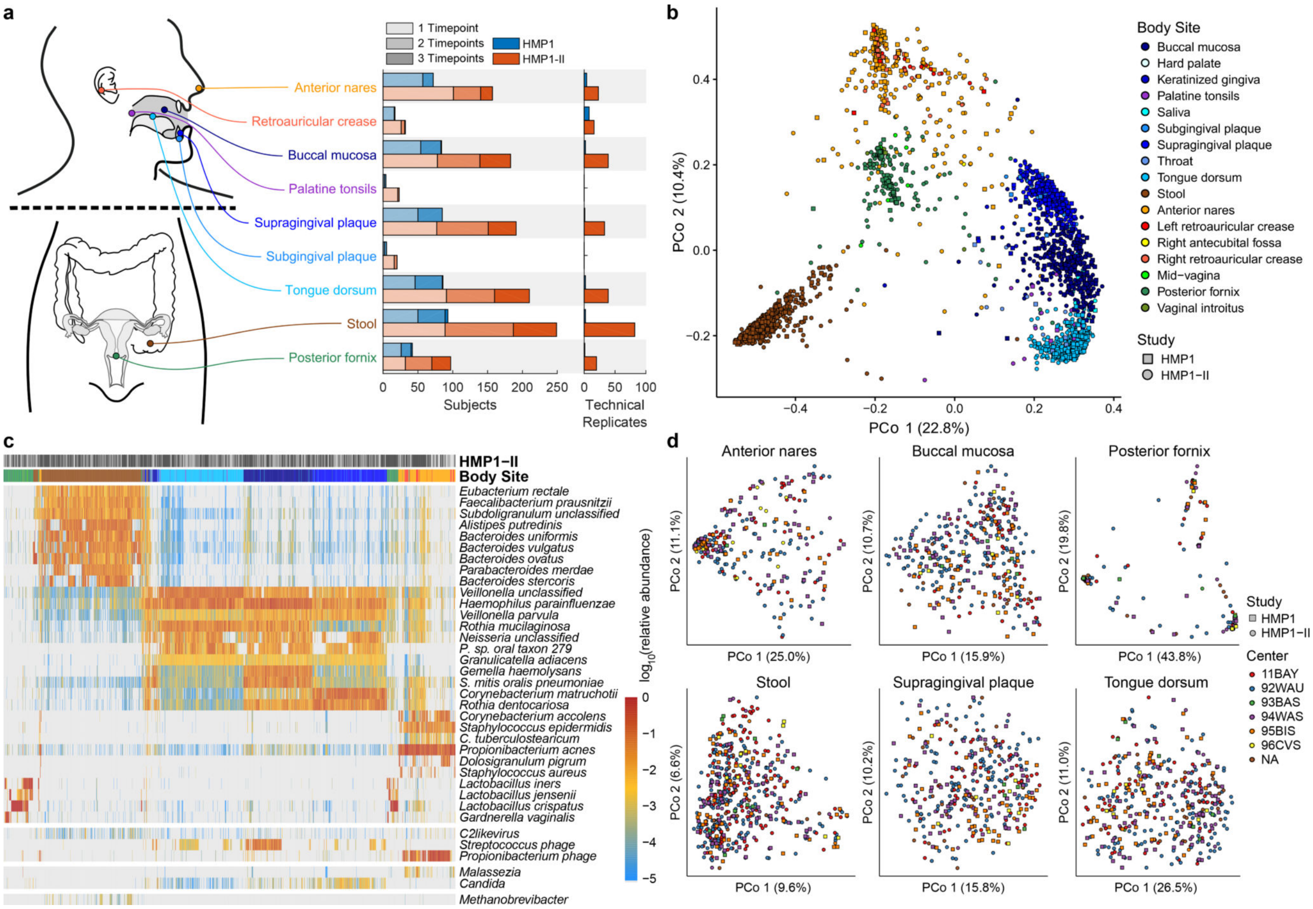
Data Availability Statement

Sequence data are available from the HMP DACC (http://hmpdacc.org) and from SRA BioProjects PRJNA48479 and PRJNA275349. Taxonomic and functional profiles, as well as assembled metagenomes, are available from the HMP DACC (Extended Data Table 1b).

[2] https://github.com/jorvis/biocode/blob/master/fastq/filter_fastq_by_N_content.py

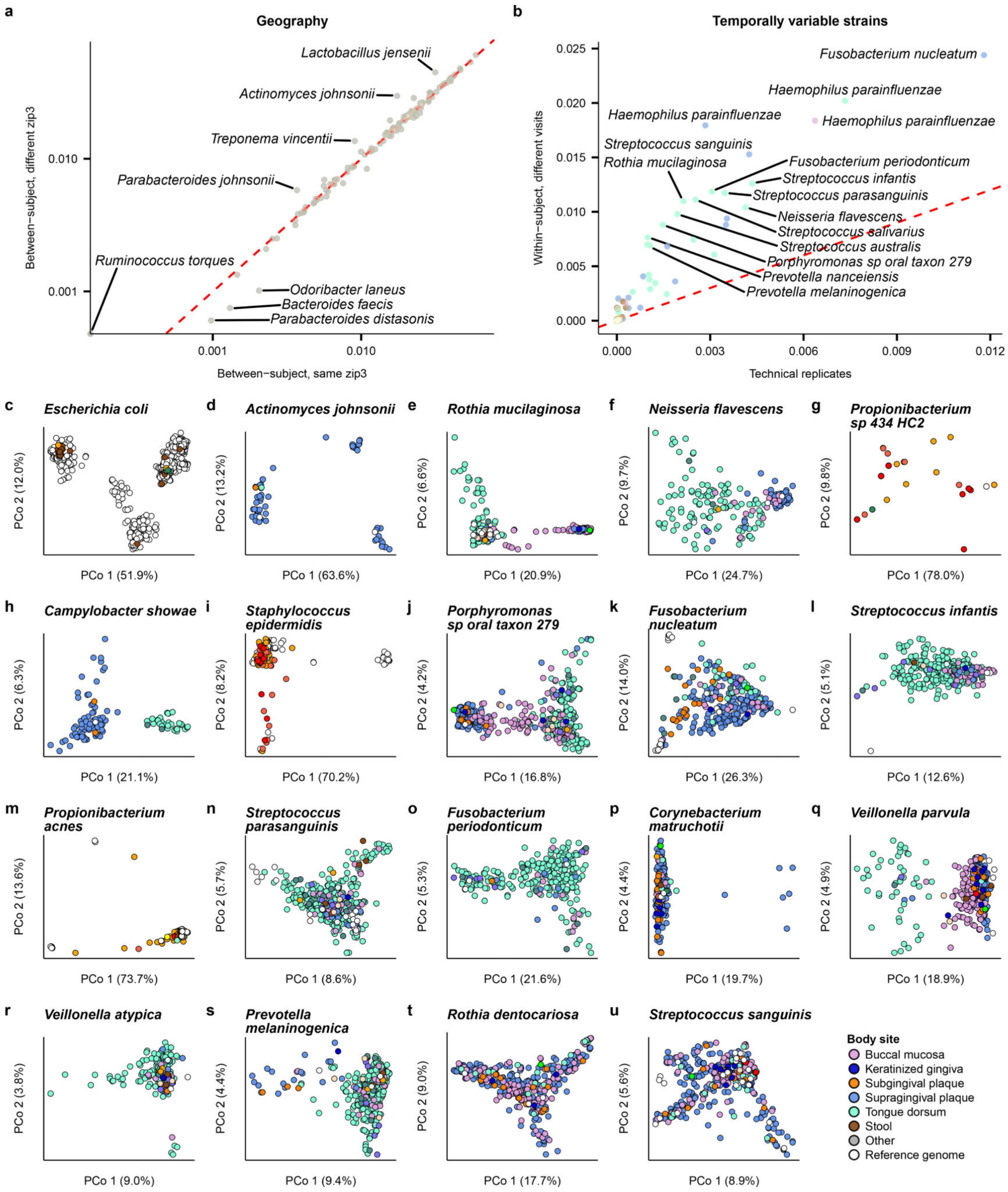
[3] https://github.com/jorvis/biocode/blob/master/sandbox/jorvis/generate_read_to_metaref_seed_alignment.py

Extended Data



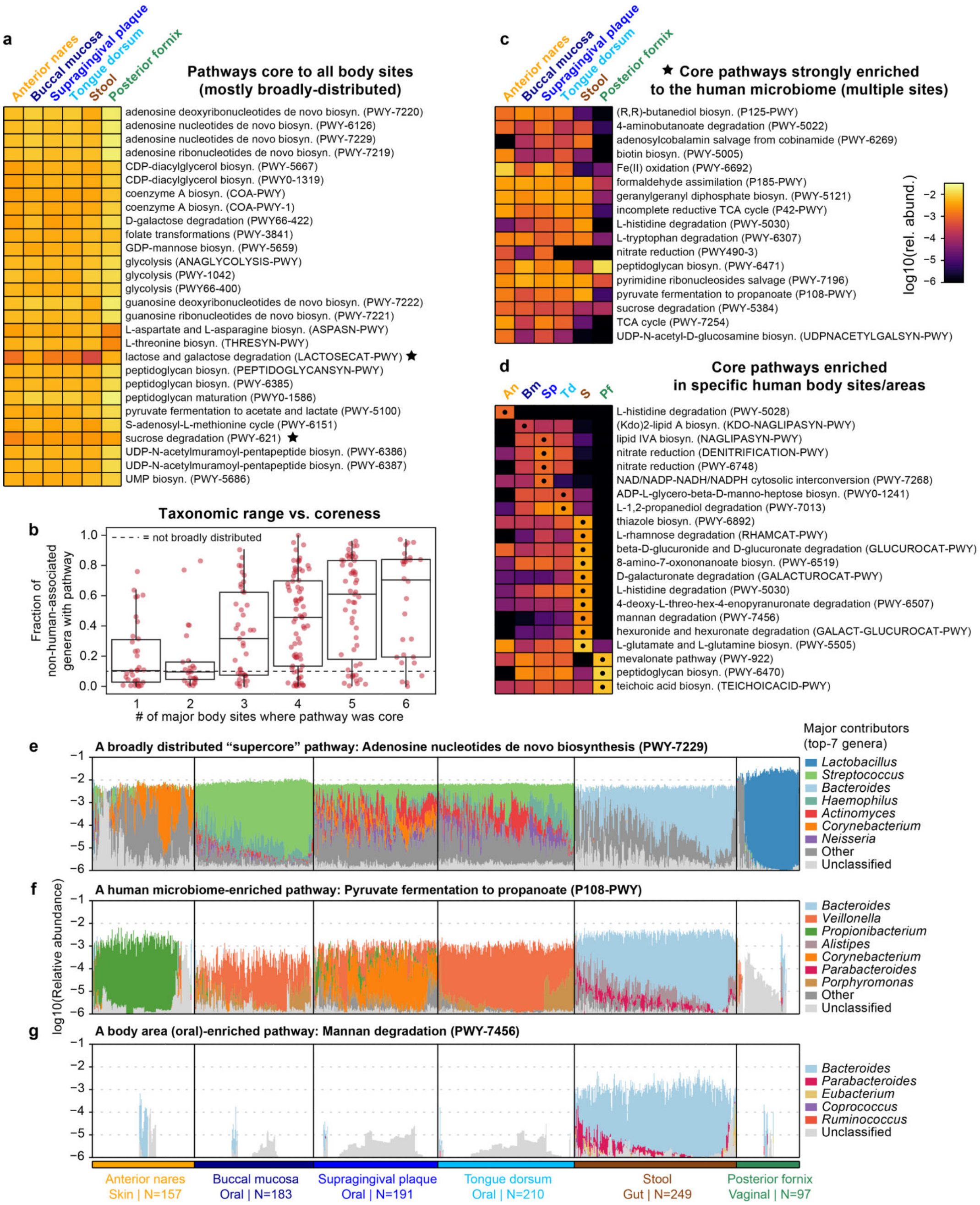
Extended Data Figure 1. Extended body-wide metagenomic taxonomic profiles in HMP1-II

a, The combined HMP1-II datasets include a total of 2,355 metagenomes (724 previously published and 1,631 new, including 252 technical replicates). These span the project's six targeted body sites (anterior nares, buccal mucosa, supragingival plaque, tongue dorsum, stool, and posterior fornix) in addition to at least 20 samples each from three additional sites, of the 18 total sampled sites: Retroauricular crease, Palatine tonsils, and Subgingival plaque. Metagenomes are now available for at least one body site for a total of 265 individuals. **b,** Principal coordinates ordination using Bray-Curtis distances among all microbes at the species level. **c,** Relative abundances of the most prevalent and abundant microbes (bacterial, viral, eukaryotic, and archaeal) among all body sites, as profiled by MetaPhlAn2[20]. Prevalent eukaryotic microbes are shown at the genus level. **d,** Taxonomic profiles do not vary more between sequencing centers, batches, or clinical centers than they do among individuals within body sites. Ordinations show Bray-Curtis principal coordinates of species-level abundances at each body site. Within-site ecological structure is as expected[1], with no divergence associated with technical variables along the first two ordination axes.



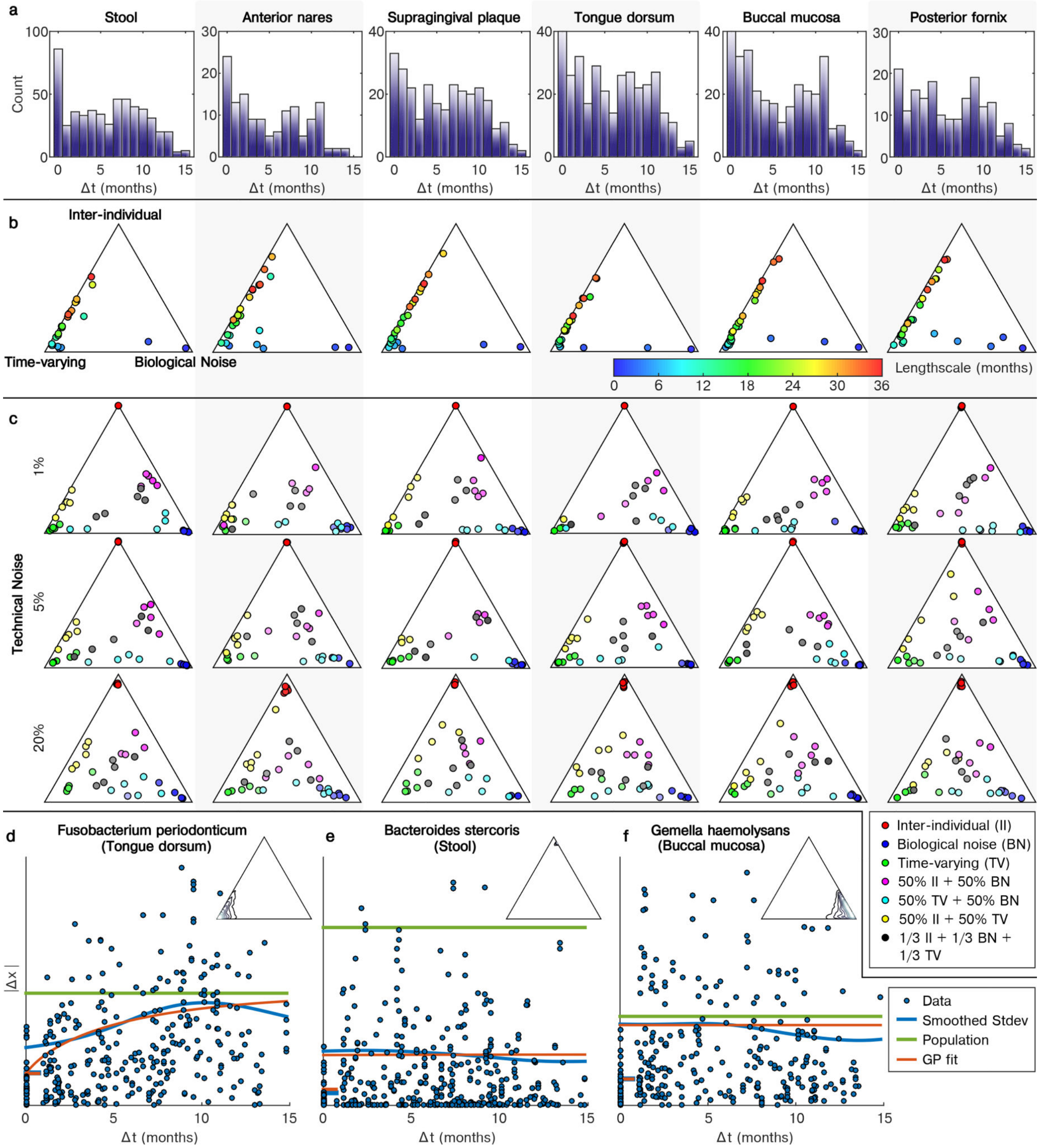
Extended Data Figure 2. Geographic, temporal, and biogeographic strain variation

a, Mean distance (Kimura two-parameter) among strains from subjects either within or between three-digit zip codes (the finest degree of geographic information available). Data and samples sizes in Table S2. **b,** Mean strain divergences between different visits for the same subject and body site compared to the mean distance between the same visits for the same subject and body site (technical replicates) for each species. **c–u,** PCoA plots based on the Kimura two-parameter distance[17] are shown for (**c**) *Escherichia coli*, (**d**) *Actinomyces johnsonii*, and (**e–u**) all species (i.e. those shown Fig. 1b), sorted in descending order of their niche-association score (**Methods**). Distance matrices used to generate these PCoAs are publicly available (Extended Data Table 1b).



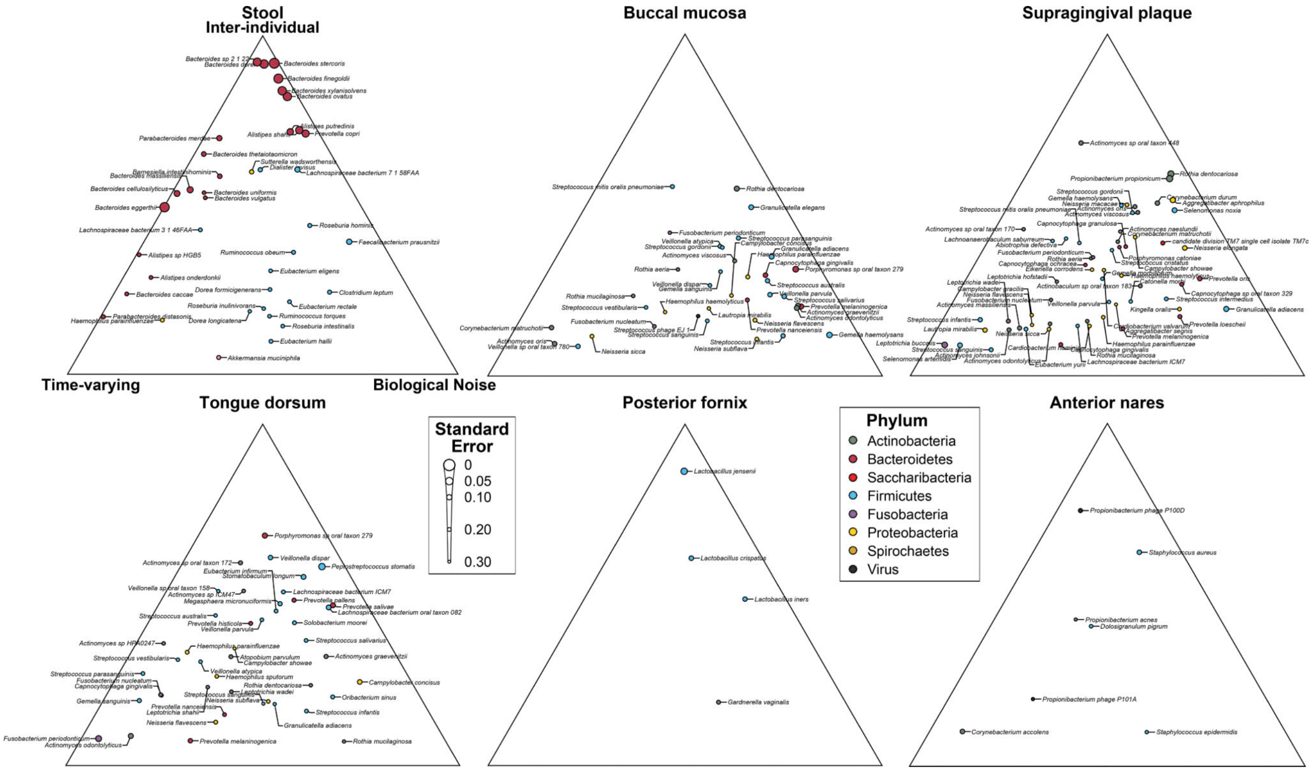
Extended Data Figure 3. Core and distinguishing functions of human body site microbiomes (additional details and examples)

This figure extends Figure 3 from the main text. **a,** 28 metabolic pathways were core (>75% prevalent) at all major body sites. We refer to these as "supercore" pathways. **b,** Pathways core to greater numbers of body sites tended to have broader taxonomic ranges, with supercore pathways among the most broadly distributed (Tukey boxplots). **c,** 19 pathways (including two supercore pathways, starred in A) were core in multiple body areas and specifically enriched among taxa inhabiting the human microbiome (annotated to <10% of non-human-associated genera). Human-microbiome-enriched pathways include specific MetaCyc-defined variants of more broadly defined/distributed processes, e.g. peptidoglycan biosynthesis (PWY-6471). **d,** “Site-enriched” pathways are considerably more abundant at one body site compared with sites from all other body areas. Black dots indicate the site where each site-enriched pathway achieved its peak abundance. Heatmap values reflect the first quartile of relative abundance in a particular body site (coordinated with the percentile cutoff for a core pathway). **e, f**, and **g** highlight additional examples of the three pathway classes enumerated in panels **a, c**, and **d**. In each example, total (community) abundance is log-scaled, while the contributions of the top seven genera are proportionally scaled within the community total. “Other” encompasses pathway contributions from genera outside of the top seven, and “unclassified” encompasses pathway contributions from unidentified members of the community.



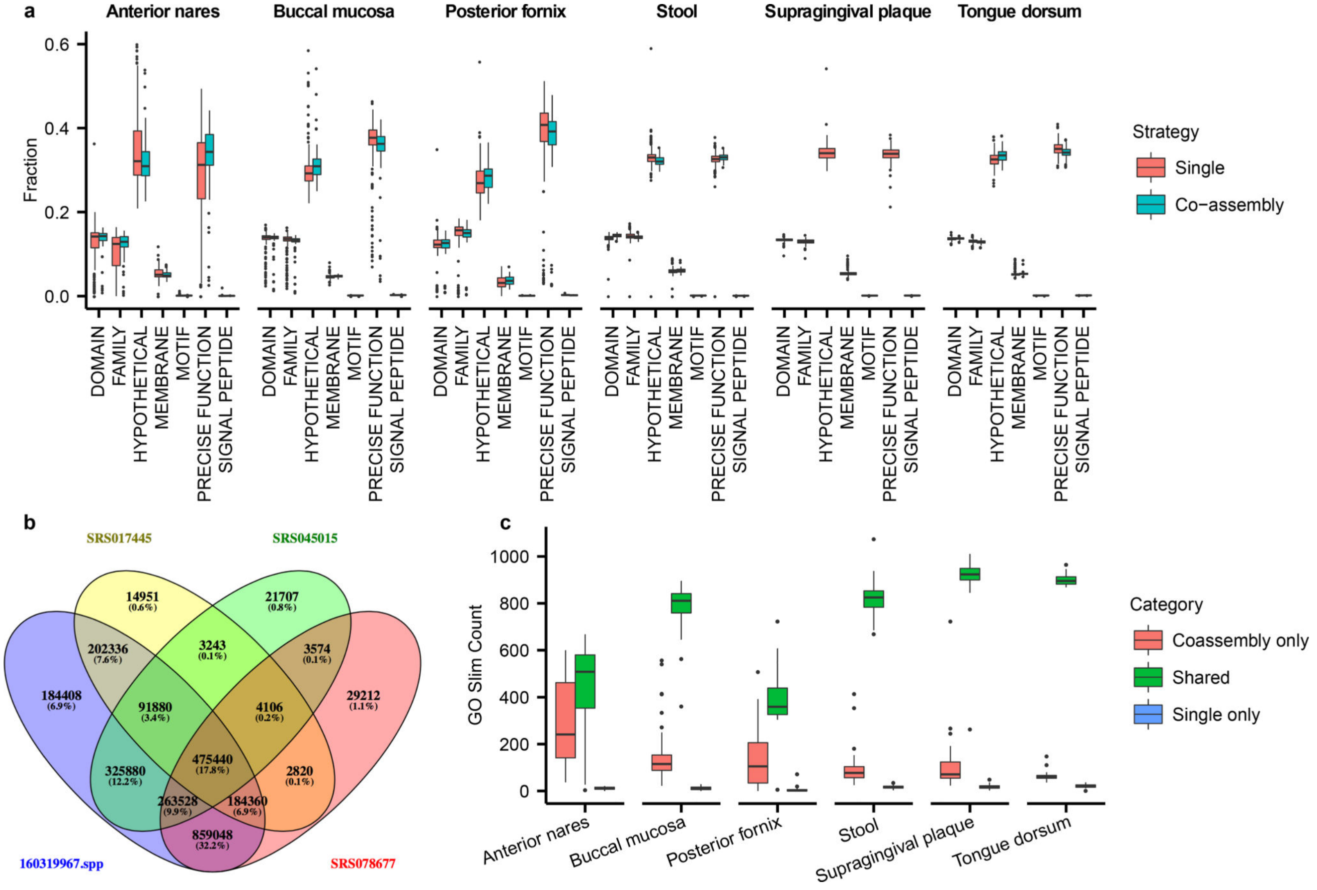
Extended Data Figure 4. Sampling interval distribution, parameter fits for simulated samples, and examples of microbial species abundance dynamics and corresponding Gaussian process fits

a, Distributions of differences in time between samples at each targeted body site. Technical replicates are shown as Δt = 0. **b,** Parameter fits for simulated samples with $U = 0$, $B = 0$, $T = 0.95$, $N = 0.05$, and varying l. Simulated samples were drawn with the real sample distribution and count from each site, to show how limitations in sampling at certain sites alter the fidelity of the fits. **c,** Parameter fits for 5 simulated samples with each of the three pure components (colored red, green and blue), as well as all even mixtures of pairs of them (e.g. yellow points are even mixtures of U and T), and even mixtures of all 3 (black), for differing levels of technical noise (*N*) and fixed *l* = 0.5. Uncertain inferences are more desaturated. **d–f,** Three examples of taxonomic profiles fit with the GP model are shown on plots designed to allow a direct comparison between the data and the fit GP, and allow the different dynamics to be visualized despite the limit of only up to 3 time points per person. Each example was chosen as an exemplar of one of the three non-technical components in the model. **Insets:** Confidence deciles of the MCMC samples. The abundance of *Fusobacterium periodonticum* in tongue dorsum shows strong time-varying behaviour (**d**), *Bacteroides stercoris* in stool shows mostly inter-individual differences (**e**), and *Gemella haemolysans* in stool is dominated by biological noise (**f**). The plots show the absolute difference in arcsin-sqrt-transformed microbial abundance (|Δx|) between pairs of samples from the same person against the difference in time between samples (points). A Gaussian-smoothed estimate of the standard deviation of the points is also shown (blue line, bandwidth 3 months), along with the expected difference from the fit GP (red line). The standard deviation of differences between technical replicates (points with Δt = 0 months) is also shown as the line stub at the origin, directly visualizing the level of technical noise. Biological noise is visible here as the difference between technical noise and the variance of the remaining points extrapolated to the origin. The time-varying component is visible as a gradual increase in the variance of the differences over time (i.e. gradually increasing red/blue lines). Finally, inter-individual differences are visible by comparing the limit of the variance of the data with the variance of differences between subjects (green line).



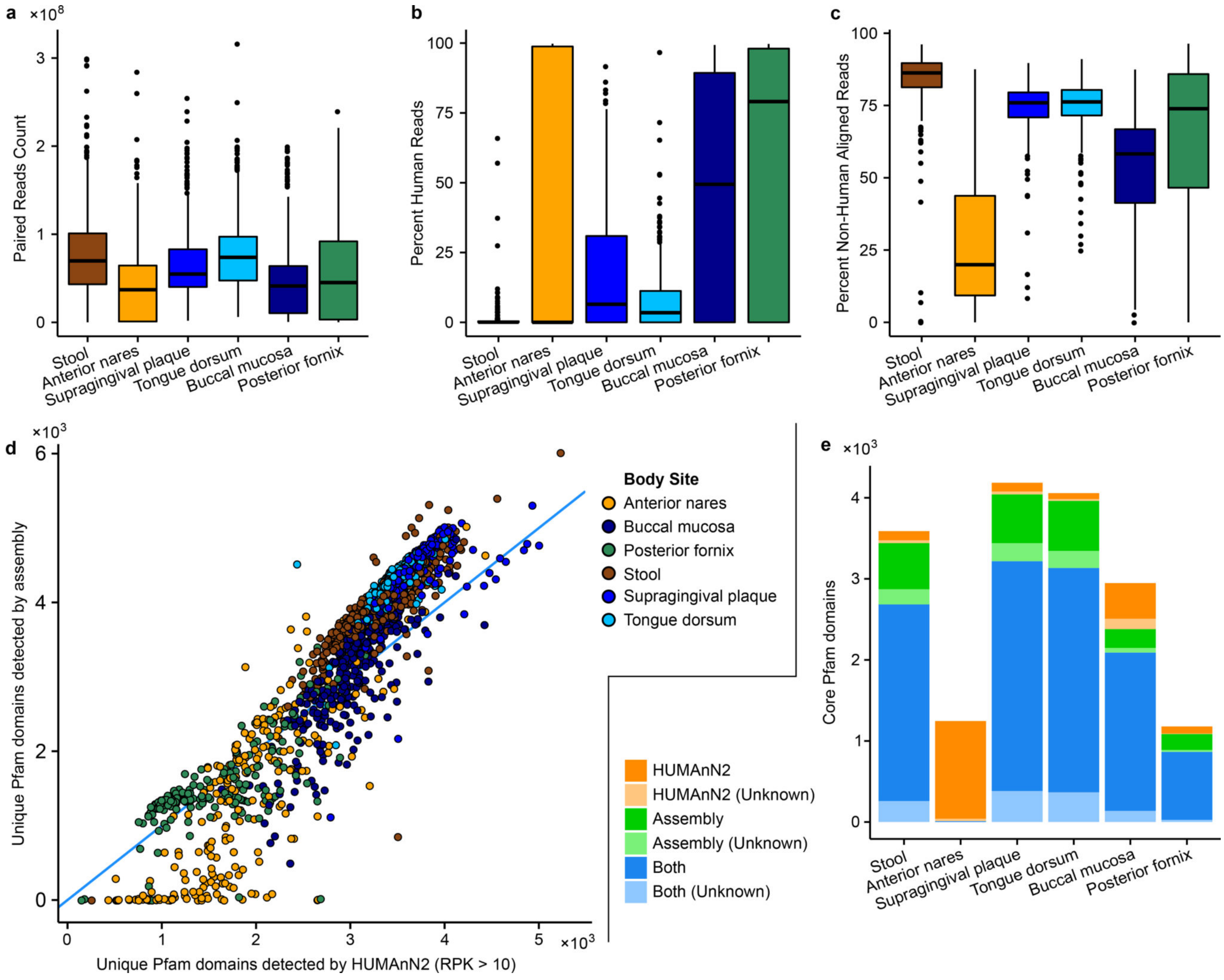
Extended Data Figure 5. Gaussian process decomposition of temporal variance for metagenomic species abundances

The posterior mean of the decomposition of variance (**Methods**) is shown for each species (Table S5), colored by phylum. Uncertainty in the estimate was assessed by the square root of the mean squared distance on the ternary plot of MCMC samples from the posterior mean, and is codified with larger points indicating more certain estimates.



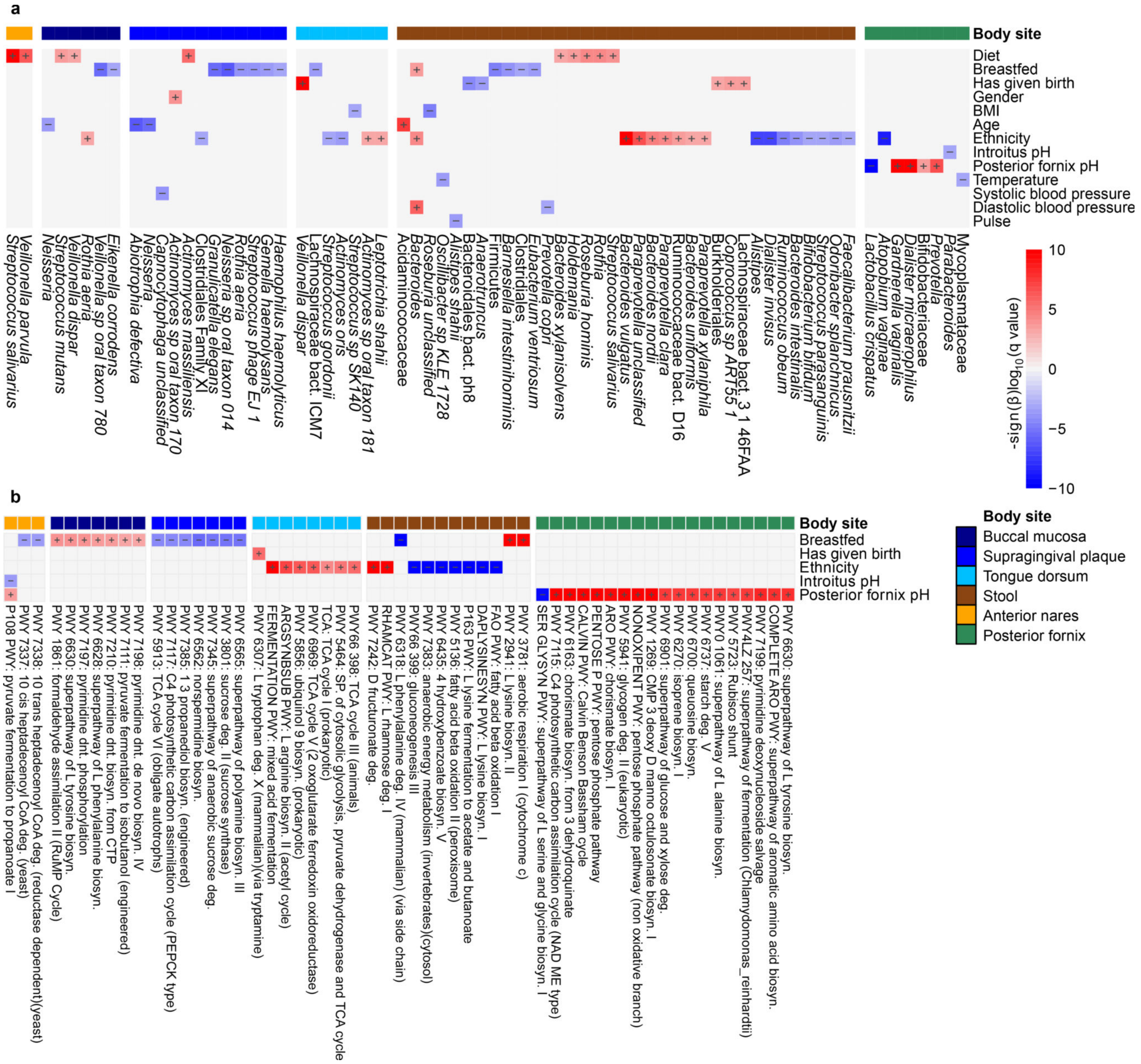
Extended Data Figure 6. Assembly annotation specificity for single and co-assemblies

a, Tukey boxplot of the percentage of proteins in each functional specificity category. **b,** An example Venn diagram for one set of single-sample supragingival plaque assemblies and their combined co-assembly (co-assembly on lower left), showing counts of shared genes (computed via strict alignment) between all combinations of assemblies; the co-assembly by itself contains 96.9% of all detected genes. **c,** Tukey boxplot of the number of GO terms (generated using a GO Slim with ~1,700 terms) shared between single and co-assemblies, unique to the co-assembly, or unique to one of the single assemblies, generated from a random selection of 250 assemblies across six body sites. Co-assemblies capture GO terms that are not in individual assemblies.



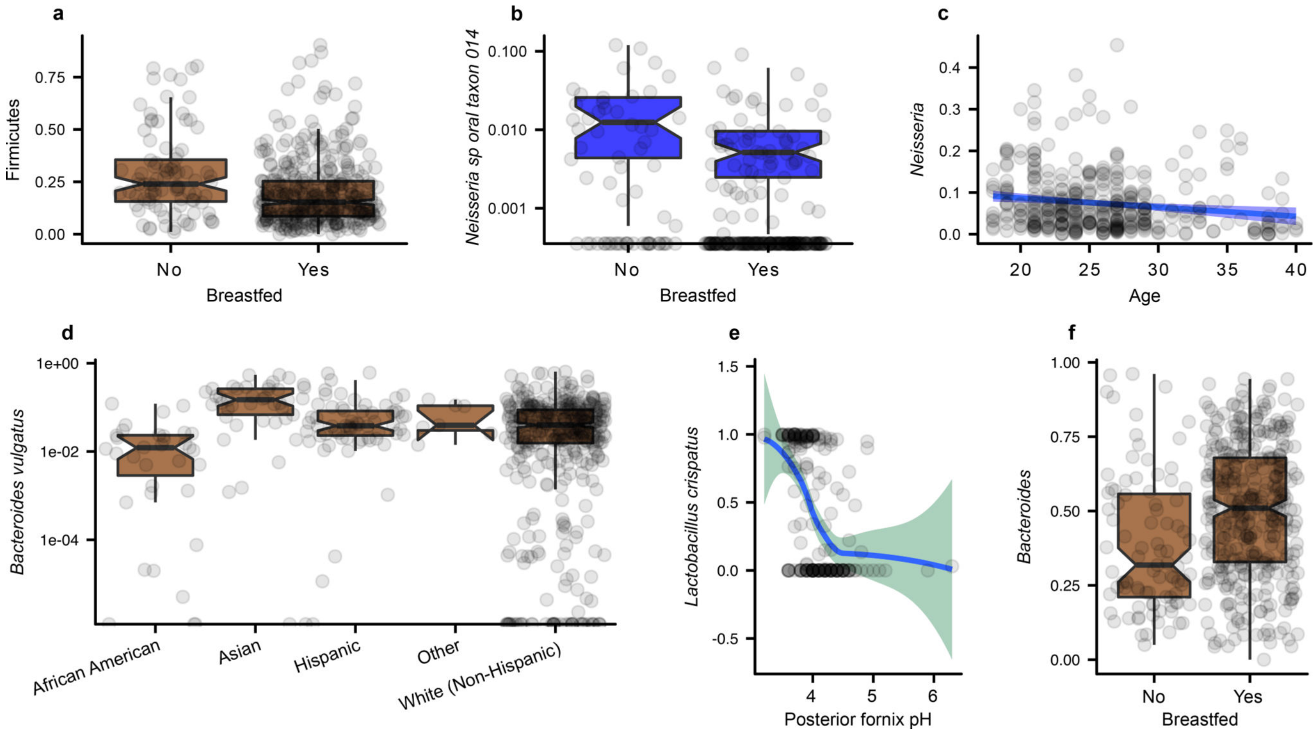
Extended Data Figure 7. Sequencing statistics and assembly quality assessments

a, Tukey boxplots of total raw reads per sample among body sites upon retrieval from the SRA. **b,** Percentages of human reads marked by BMTagger per body site. **c,** Percentages of non-human (bacterial) reads aligned to assemblies showing assembly effectiveness (**Methods, Read and contig mapping to assemblies and reference genomes**). **d,** Comparison of the number of unique Pfam domains detected in each sample by HUMAnN2 and in the assemblies, colored by body site. Pfam domains in HUMAnN2 were considered “detected” if UniRef50 sequences annotated with the domain were present in a sample at >10 RPK (~1× coverage). Pfam domains were detected in the assemblies if they were found on a single contig by Attributor (**Methods**). **e,** Number of Pfam domains that were detected in at least 75% of samples (“core” domains) by each method, for each targeted body site. Pfams domains are stratified by unknown function.



Extended Data Figure 8. Metagenomic features abundances significantly associated with host phenotype

a,b, Significant associations of nontrivial effect size (FDR < 0.1 and |*β*| > 0.01) in a multivariate linear model (significance and coefficients in Table S5) between taxon abundances (**a**) and pathway abundances (**b**). All detected associations are independent of all other metadata, including whether the subject was breastfed, the subject's broad dietary characterization, temperature, introitus pH, posterior fornix pH, gender, age, ethnicity, study day processed, sequencing center, clinical center, number of quality bases, percent of human reads, systolic blood pressure, diastolic blood pressure, pulse, whether the subject had given birth, HMP1/HMP1-II, and BMI (group sizes in Extended Data Table 1; see **Methods**). Non-significant associations here should not be considered evidence of no association.



Extended Data Figure 9. Updated associations in HMP1-II

a, HMP cohort subjects reported whether they were breastfed as infants. Remarkably, overall phylum Firmicutes abundances were lower even during adulthood (subjects' current ages were 18–40) in individuals who had been historically breastfed. **b,** Differences associated with infant breastfeeding persisted in other clades and body sites, e.g. oral *Neisseria*, although **c,** age-linked associations differed among taxa (e.g. overall oral *Neisseria* decrease with age). Examples of associations significant in the original HMP1 metagenome set[1] that were retained in the larger HMP1-II dataset include: **d,** *Bacteroides vulgatus* in stool is significantly more abundant in Asians compared to other ethnicities. **e,** *Lactobacillus crispatus* in the posterior fornix is negatively associated vaginal pH. **f,** *Bacteroides* are significantly more abundant in individuals who have been breastfed as infants. Boxplot whiskers are defined by Tukey's method.

a	Metadata	N, mean ± std [min, max]
	Gender	Female (n=128), Male (n=137)
	Age (years)	26.4 ± 5.1 [18, 40]
	Ethnicity	Hispanic (n=29), Non-Hispanic (n=236)
	Diet (meat consumption)	3+ days/week (n=224), 1–2 days/week (n=15), eggs/dairy but no meat (n=11), no animal products (n=1), N/A (n=14)
	Breastfed	Yes (n=179), No (n=40), Don't know/remember (n=32), N/A (n=14)
	Given Birth	Never (n=101), Once (n=16), More than once (n=7), N/A (n=4)
	Recruitment center	Baylor College of Medicine (n=116), Washington University (n=133)
	Pulse (bpm)	71.3 ± 11.3 [42, 100]

a	Metadata	N, mean ± std [min, max]
	Systolic blood pressure (mmHg)	119.8 ± 12.5 [91, 151]
	Diastolic blood pressure (mmHg)	71.8 ± 9.3 [50, 98]
	Weight (lb)	161.6 ± 31.9 [95.4, 271.8]
	Height (in)	67.8 ± 4.1 [58, 78.5]
	BMI	24.3 ± 3.5 [19, 34]
	Per-visit	
	Temperature (°F)	97.9 ± 0.7 [95.4, 99.7] (n=2271, 84 N/A)
	Introitus pH	4.5 ± 0.5 [3.4, 6.5] (n=1181, 85 N/A)
	Posterior fornix pH	4.1 ± 0.5 [3.2, 7] (n=1181, 85 N/A)

b	Data	URL
	Metagenomic taxon abundances (MetaPhlAn2)	hmpdacc.org/hmsmcp2
	Gene family abundances (UniRef50, from HUMAnN2)	hmpdacc.org/hmmrc2
	Pathway abundances and coverages (HUMAnN2)	hmpdacc.org/hmmrp2
	Kimura 2-parameter strain distance matrices	hmpdacc.org/hmsdm2
	Assembled metagenomes	hmpdacc.org/hmasm2
	Co-assembled metagenomes	hmpdacc.org/hmcasm2
	Assembly annotations	hmpdacc.org/hmgi2
	Individual assembly clustered gene index	hmpdacc.org/hmgc2
	Co-assembly annotations	hmpdacc.org/hmcgi2
	Co-assembly clustered gene index	hmpdacc.org/hmcgc2
	Read alignments to reference genomes	hmpdacc.org/hmrarg2
	Assembly-contig alignments to reference genomes	hmpdacc.org/hmcarg2
	Read alignments to assembly contigs	hmpdacc.org/hmrac2
	Gaussian Process models Matlab source	hmpdacc.org/hmgps2
	Assembly pipeline	https://github.com/IGS/IMA

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We wish to acknowledge Jain A. Aluvathingal, Gautam Shankar and Kent Shefchek for their contributions in running analyses for this project. Victor Felix is responsible for making the data available through the DACC web portal. We are also grateful to the original HMP external scientific advisory board (Richard Blumberg, Julian Davies, Robert Holt, Pilar Ossorio, Francis Ouellette, Gary Schoolnik, and Alan Williamson), collaborators throughout the International Human Microbiome Consortium, and the generous participation of the individuals from the Saint Louis, MO, and Houston, TX areas who made the Human Microbiome Project possible. This work was supported by National Institutes of Health grants U54DK102557 to CH and OW, U01HG006537 to OW, U01HG004866 to OW, U54HG004969 to Bruce W Birren, U54HG004973 to Richard A Gibbs, Sarah K Highlander, Joseph F Petrosino, U54AI084844 to Karen E Nelson, U54HG004968 to George M Weinstock, U54HG003079 to Richard K Wilson, and Crohn's & Colitis Foundation of America (CCFA) award #3162 to Jonathan Braun with subaward to CH and Ramnik Xavier.

References

1. The Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. Nature. 2012; 486:207–214. DOI: 10.1038/nature11234 [PubMed: 22699609]

2. Lloyd-Price J, Abu-Ali G, Huttenhower C. The healthy human microbiome. Genome Med. 2016; 8:51. [PubMed: 27122046]

3. Gensollen T, Iyer SS, Kasper DL, Blumberg RS. How colonization by microbiota in early life shapes the immune system. Science. 2016; 352:539–544. DOI: 10.1126/science.aad9378 [PubMed: 27126036]

4. Honda K, Littman DR. The microbiota in adaptive immune homeostasis and disease. Nature. 2016; 535:75–84. DOI: 10.1038/nature18848 [PubMed: 27383982]

5. Qin J, et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature. 2010; 464:59–65. DOI: 10.1038/nature08821 [PubMed: 20203603]

6. Li J, et al. An integrated catalog of reference genes in the human gut microbiome. Nat Biotechnol. 2014; 32:834–841. DOI: 10.1038/nbt.2942 [PubMed: 24997786]

7. Beaumont M, et al. Heritable components of the human fecal microbiome are associated with visceral fat. Genome Biol. 2016; 17:189. [PubMed: 27666579]

8. Falony G, et al. Population-level analysis of gut microbiome variation. Science. 2016; 352:560–564. DOI: 10.1126/science.aad3503 [PubMed: 27126039]

9. Zhernakova A, et al. Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. Science. 2016; 352:565–569. DOI: 10.1126/science.aad3369 [PubMed: 27126040]

10. Si J, You HJ, Yu J, Sung J, Ko G. Prevotella as a Hub for Vaginal Microbiota under the Influence of Host Genetics and Their Association with Obesity. Cell Host Microbe. 2017; 21:97–105. DOI: 10.1016/j.chom.2016.11.010 [PubMed: 28017660]

11. Gonzalez A, et al. Migraines Are Correlated with Higher Levels of Nitrate-, Nitrite-, and Nitric Oxide-Reducing Oral Microbes in the American Gut Project Cohort. mSystems. 2016; 1

12. Oh J, et al. Biogeography and individuality shape function in the human skin metagenome. Nature. 2014; 514:59–64. DOI: 10.1038/nature13786 [PubMed: 25279917]

13. The Human Microbiome Project Consortium. A framework for human microbiome research. Nature. 2012; 486:215–221. DOI: 10.1038/nature11209 [PubMed: 22699610]

14. Truong DT, Tett A, Pasolli E, Huttenhower C, Segata N. Microbial strain-level population structure and genetic diversity from metagenomes. Genome Res. 2017; 27:626–638. DOI: 10.1101/gr.216242.116 [PubMed: 28167665]

15. Schloissnig S, et al. Genomic variation landscape of the human gut microbiome. Nature. 2013; 493:45–50. DOI: 10.1038/nature11711 [PubMed: 23222524]

16. Luo C, et al. ConStrains identifies microbial strains in metagenomic datasets. Nat Biotechnol. 2015; 33:1045–1052. DOI: 10.1038/nbt.3319 [PubMed: 26344404]

17. Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol. 1980; 16:111–120. [PubMed: 7463489]

18. Franzosa EA, et al. Identifying personal microbiomes using metagenomic codes. Proc Natl Acad Sci U S A. 2015; 112:E2930–2938. DOI: 10.1073/pnas.1423854112 [PubMed: 25964341]

19. Leinonen R, Sugawara H, Shumway M. International Nucleotide Sequence Database, C. The sequence read archive. Nucleic Acids Res. 2011; 39:D19–21. DOI: 10.1093/nar/gkq1019 [PubMed: 21062823]

20. Truong DT, et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat Methods. 2015; 12:902–903. DOI: 10.1038/nmeth.3589 [PubMed: 26418763]

21. Hoffmann C, et al. Archaea and fungi of the human gut microbiome: correlations with diet and bacterial residents. PLoS One. 2013; 8:e66019. [PubMed: 23799070]

22. Schwiertz A, et al. Microbiota and SCFA in lean and overweight healthy subjects. Obesity (Silver Spring). 2010; 18:190–195. DOI: 10.1038/oby.2009.167 [PubMed: 19498350]

23. Pride DT, et al. Evidence of a robust resident bacteriophage population revealed through analysis of the human salivary virome. ISME J. 2012; 6:915–926. DOI: 10.1038/ismej.2011.169 [PubMed: 22158393]

24. Abubucker S, et al. Metabolic reconstruction for metagenomic data and its application to the human microbiome. PLoS Comput Biol. 2012; 8:e1002358. [PubMed: 22719234]

25. Caspi R, et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic Acids Res. 2014; 42:D459–471. DOI: 10.1093/nar/gkt1103 [PubMed: 24225315]

26. Leonardi R, Zhang YM, Rock CO, Jackowski S. Coenzyme A: back in action. Prog Lipid Res. 2005; 44:125–153. DOI: 10.1016/j.plipres.2005.04.001 [PubMed: 15893380]

27. Khakh BS, Burnstock G. The double life of ATP. Sci Am. 2009; 301:84–90. 92. [PubMed: 20058644]

28. Morkbak AL, Poulsen SS, Nexo E. Haptocorrin in humans. Clin Chem Lab Med. 2007; 45:1751–1759. DOI: 10.1515/CCLM.2007.343 [PubMed: 17990953]

29. Roy CC, Kien CL, Bouthillier L, Levy E. Short-chain fatty acids: ready for prime time? Nutr Clin Pract. 2006; 21:351–366. [PubMed: 16870803]

30. Schreiber F, et al. Denitrification in human dental plaque. BMC Biol. 2010; 8:24. [PubMed: 20307293]

31. Flint HJ, Scott KP, Duncan SH, Louis P, Forano E. Microbial degradation of complex carbohydrates in the gut. Gut Microbes. 2012; 3:289–306. DOI: 10.4161/gmic.19897 [PubMed: 22572875]

32. Faith JJ, et al. The long-term stability of the human gut microbiota. Science. 2013; 341:1237439. [PubMed: 23828941]

33. Flores GE, et al. Temporal variability is a personalized feature of the human microbiome. Genome Biol. 2014; 15:531. [PubMed: 25517225]

34. Ding T, Schloss PD. Dynamics and associations of microbial community types across the human body. Nature. 2014; 509:357–360. DOI: 10.1038/nature13178 [PubMed: 24739969]

35. Turnbaugh PJ, et al. The human microbiome project. Nature. 2007; 449:804–810. DOI: 10.1038/nature06244 [PubMed: 17943116]

36. Shafquat A, Joice R, Simmons SL, Huttenhower C. Functional and phylogenetic assembly of microbial communities in the human microbiome. Trends Microbiol. 2014; 22:261–266. DOI: 10.1016/j.tim.2014.01.011 [PubMed: 24618403]

37. Gajer P, et al. Temporal dynamics of the human vaginal microbiota. Sci Transl Med. 2012; 4:132ra152.

38. Peng Y, Leung HC, Yiu SM, Chin FY. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics. 2012; 28:1420–1428. DOI: 10.1093/bioinformatics/bts174 [PubMed: 22495754]

39. Finn RD, et al. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 2016; 44:D279–285. DOI: 10.1093/nar/gkv1344 [PubMed: 26673716]

40. Aagaard K, et al. The Human Microbiome Project strategy for comprehensive sampling of the human microbiome and why it matters. FASEB J. 2013; 27:1012–1022. DOI: 10.1096/fj.12-220806 [PubMed: 23165986]

41. Roager HM, et al. Colonic transit time is related to bacterial metabolism and mucosal turnover in the gut. Nat Microbiol. 2016; 1:16093. [PubMed: 27562254]

42. Caporaso JG, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Methods. 2010; 7:335–336. DOI: 10.1038/nmeth.f.303 [PubMed: 20383131]

43. Huang K, et al. MetaRef: a pan-genomic database for comparative and community microbial genomics. Nucleic Acids Res. 2014; 42:D617–624. DOI: 10.1093/nar/gkt1078 [PubMed: 24203705]

44. Suzek BE, et al. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics. 2015; 31:926–932. DOI: 10.1093/bioinformatics/btu739 [PubMed: 25398609]

45. Caspi R, et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res. 2016; 44:D471–480. DOI: 10.1093/nar/gkv1164 [PubMed: 26527732]

46. Galperin MY, Makarova KS, Wolf YI, Koonin EV. Expanded microbial genome coverage and improved protein family annotation in the COG database. Nucleic Acids Res. 2015; 43:D261–269. DOI: 10.1093/nar/gku1223 [PubMed: 25428365]

47. Baba T, et al. Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol Syst Biol. 2006; 2 2006 0008.

48. The UniProt Consortium. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 2017; 45:D158–D169. DOI: 10.1093/nar/gkw1099 [PubMed: 27899622]

49. Gelman A, Rubin DB. Inference from Iterative Simulation Using Multiple Sequences. Statistical Science. 1992; 7:457–511.

50. Morgan XC, et al. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. Genome Biol. 2012; 13:R79. [PubMed: 23013615]

51. Namiki T, Hachiya T, Tanaka H, Sakakibara Y. MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. Nucleic Acids Res. 2012; 40:e155. [PubMed: 22821567]

52. Luo R, et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience. 2012; 1:18. [PubMed: 23587118]

53. Boisvert S, Raymond F, Godzaridis E, Laviolette F, Corbeil J. Ray Meta: scalable de novo metagenome assembly and profiling. Genome Biol. 2012; 13:R122. [PubMed: 23259615]

54. Bankevich A, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012; 19:455–477. DOI: 10.1089/cmb.2012.0021 [PubMed: 22506599]

55. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008; 18:821–829. DOI: 10.1101/gr.074492.107 [PubMed: 18349386]

56. Pell J, et al. Scaling metagenome sequence assembly with probabilistic de Bruijn graphs. Proc Natl Acad Sci U S A. 2012; 109:13272–13277. DOI: 10.1073/pnas.1121464109 [PubMed: 22847406]

57. Mende DR, et al. Assessment of metagenomic assembly using simulated next generation sequencing data. PLoS One. 2012; 7:e31386. [PubMed: 22384016]

58. Zhu W, Lomsadze A, Borodovsky M. Ab initio gene identification in metagenomic sequences. Nucleic Acids Res. 2010; 38:e132. [PubMed: 20403810]

59. Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics. 2007; 23:1282–1288. DOI: 10.1093/bioinformatics/btm098 [PubMed: 17379688]

60. Zhao Y, Tang H, Ye Y. RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. Bioinformatics. 2012; 28:125–126. DOI: 10.1093/bioinformatics/btr595 [PubMed: 22039206]

61. Finn RD, et al. Pfam: the protein families database. Nucleic Acids Res. 2014; 42:D222–230. DOI: 10.1093/nar/gkt1223 [PubMed: 24288371]

62. Haft DH, et al. TIGRFAMs and Genome Properties in 2013. Nucleic Acids Res. 2013; 41:D387–395. DOI: 10.1093/nar/gks1234 [PubMed: 23197656]

63. Eddy SR. Accelerated Profile HMM Searches. PLoS Comput Biol. 2011; 7:e1002195. [PubMed: 22039361]

64. Sonnhammer EL, von Heijne G, Krogh A. A hidden Markov model for predicting transmembrane helices in protein sequences. Proc Int Conf Intell Syst Mol Biol. 1998; 6:175–182. [PubMed: 9783223]

65. Orvis J, et al. Ergatis: a web interface and scalable software system for bioinformatics workflows. Bioinformatics. 2010; 26:1488–1492. DOI: 10.1093/bioinformatics/btq167 [PubMed: 20413634]

66. Galens K, et al. The IGS Standard Operating Procedure for Automated Prokaryotic Annotation. Stand Genomic Sci. 2011; 4:244–251. DOI: 10.4056/sigs.1223234 [PubMed: 21677861]

67. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010; 26:2460–2461. DOI: 10.1093/bioinformatics/btq461 [PubMed: 20709691]

68. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012; 9:357–359. DOI: 10.1038/nmeth.1923 [PubMed: 22388286]

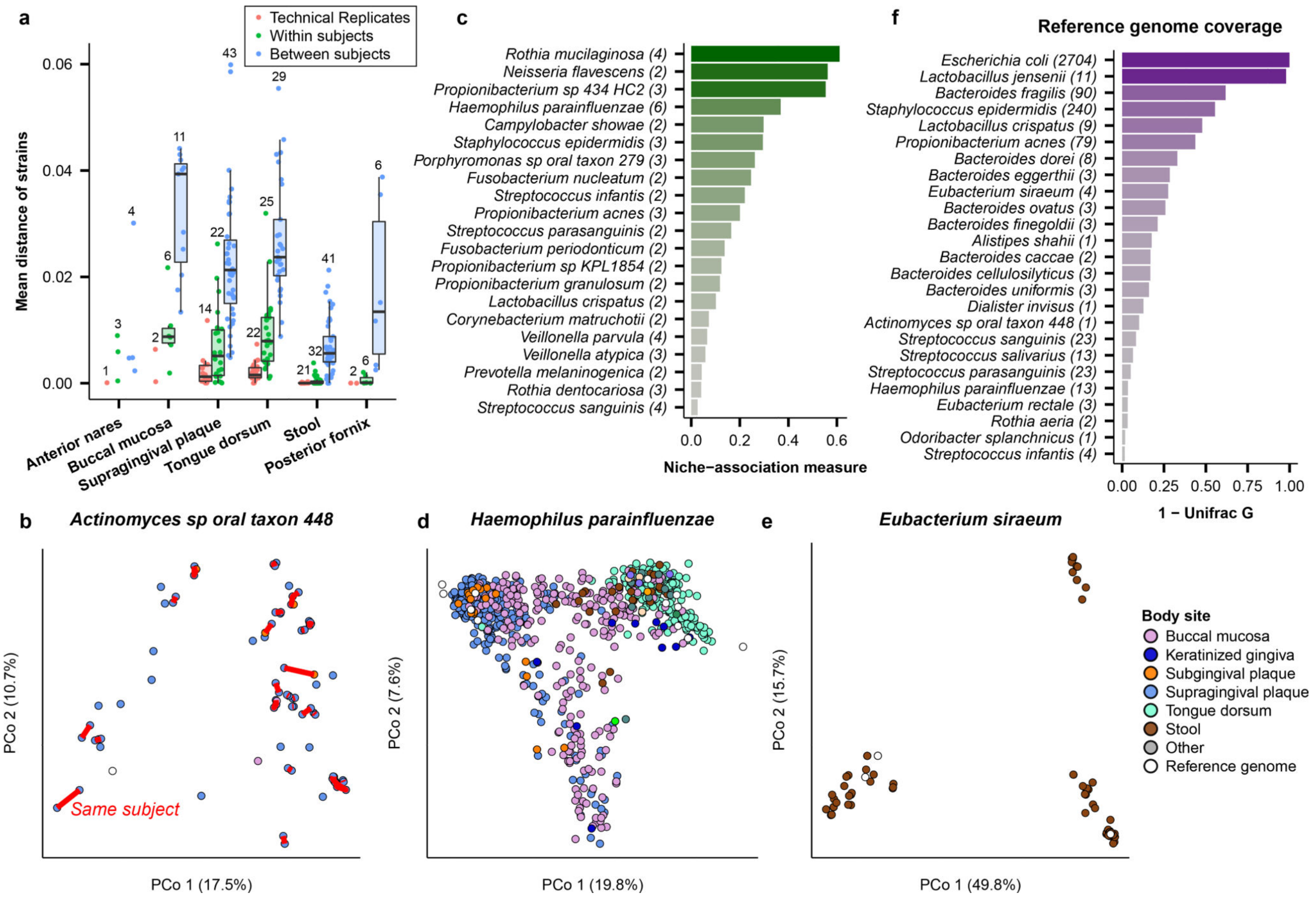


Figure 1. Personalization, niche-association, and reference genome coverage in strain-level metagenomic profiles

a, Mean phylogenetic divergences[17] between strains of species with sufficient coverage at each targeted body site (minimum 2 strain pairs). **b,** Individuals tended to retain personalized strains, as visualized by a Principal Coordinates Analysis (PCoA) plot for *Actinomyces sp oral taxon 448*, where lines connect samples from the same individual. **c,** Quantification of niche-association (**Methods**; only species with sufficient coverage in ≥5 samples at two or more body sites). Higher values indicate greater phylogenetic separation between body sites. **d,** PCoA showing niche-association of *Haemophilus parainfluenzae*, showing subspecies specialization to three different body sites. **e,** PCoA for *Eubacterium siraeum*. **f,** Coverage of human-associated strains by the current 16,903 reference genome set (**Methods**). Top 25 species by mean relative abundance when present (>0.1% relative abundance) are shown (minimum prevalence of 50 samples). Sample counts in Table S2, distance matrices available from Extended Data Table 1b.

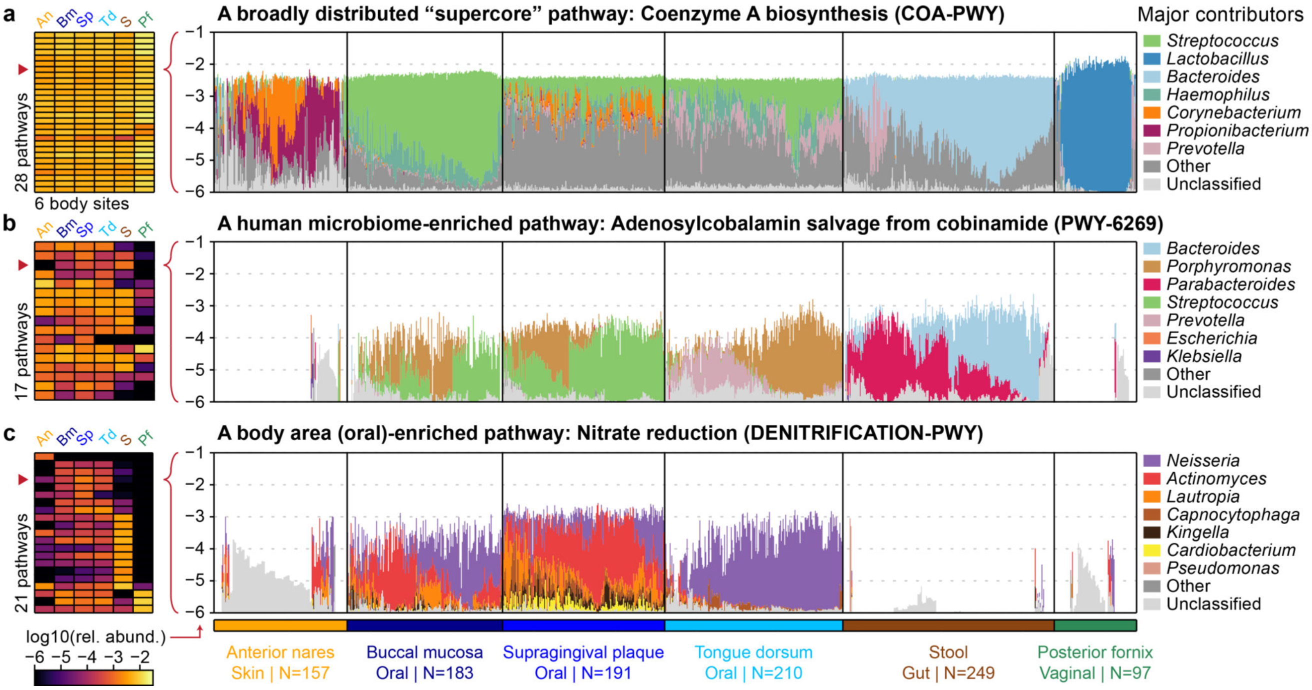


Figure 2. Core and distinguishing functions of human body site microbiomes

a, 28 metabolic pathways were core at all six major body sites ("supercore" pathways). Two supercore pathways and **b,** 17 additional pathways were core in multiple body areas and enriched among human-associated taxa ("human microbiome-enriched" pathways). **c,** 21 pathways were considerably more abundant at one body site compared with sites from all other body areas ("body site-enriched" pathways). Heatmap values reflect the first quartile of relative abundance (heatmaps are expanded in Extended Data Fig. 3). In pathway barplots, total (community) abundance is log-scaled, while the contributions of the top-seven genera are proportionally scaled within the total. "Other" encompasses contributions from additional, known genera; "unclassified" encompasses contributions of unknown taxonomy.

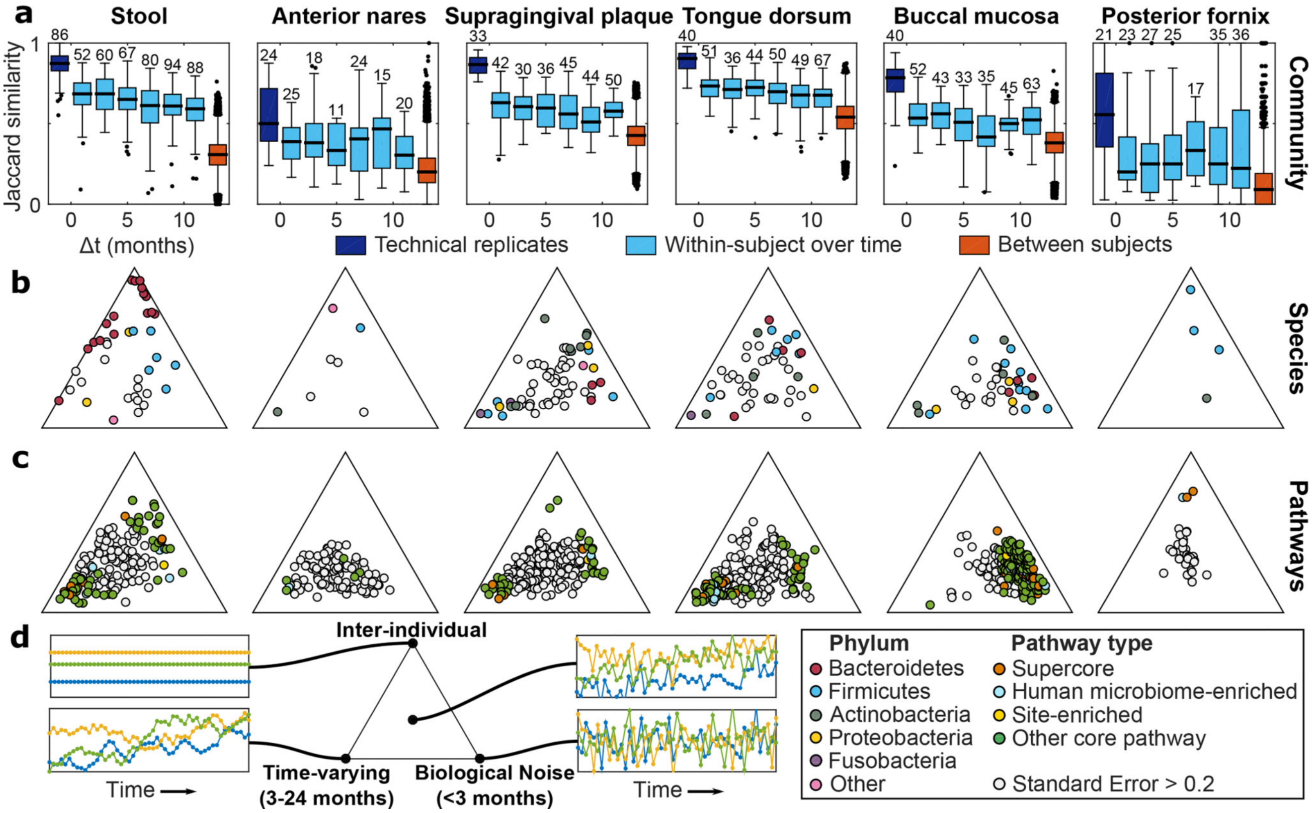


Figure 3. Temporal dynamics of individual species and microbial pathways at each targeted body site

a, Jaccard similarity is maximal between technical replicates and decreases with time, though within-subject similarity always exceeds between-subject similarity. **b,** Gaussian process decomposition of the variance in species abundances (each point is one species; filtering criteria in **Methods**) into three biologically-relevant components based on their characteristic timescales (**Methods**). Technical noise was estimated (Table S5) but not visualized. Species with high inference uncertainty (standard error on the ternary diagram > 0.2) are gray and the inference is biased towards the center of the diagrams (**Methods**). Labeled version in Extended Data Fig. 5. **c,** Same as (**b**), but for abundances of all core pathways. **d,** Illustrative time series showing dynamics at different locations within the ternary plots (Extended Data Fig. 4d–f for concrete examples). Sample counts in Table S1.

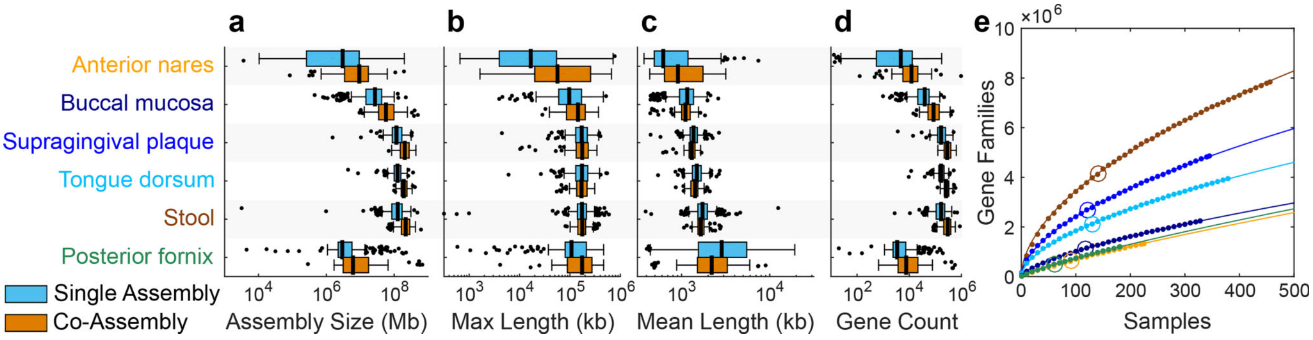


Figure 4. Assembly and annotation of body-wide human microbiomes

a–d, Tukey boxplots of total assembly size, maximum and average contig lengths, and gene counts for single and co-assemblies (sample sizes in Table S1). **e,** Rarefaction curves of gene families (ORF clusters at 90% sequence similarity) from predicted genes generated from single assemblies for the targeted body sites (points), with a power-law fit (lines). The size of the HMP1 WMS dataset at each body site is also shown (open circles). The rarefaction trajectory was robust to changes in the sequence similarity threshold (for the 188 posterior fornix samples, the number of gene families ranged between only 1,131,796 to 1,271,891 for similarities 70–95%). Coloring is as in axis labels of (**a**). Sample counts in Table S1.