Published in final edited form as:

Nat Microbiol. 2020 February ; 5(2): 343–353. doi:10.1038/s41564-019-0625-0.

Tracking microbial evolution in the human gut using Hi-C reveals extensive horizontal gene transfer, persistence, and adaptation

Eitan Yaffe[1], **David A. Relman**[1,2,3,*]

[1]Department of Medicine, Stanford University School of Medicine, Stanford, CA 94305-5107, USA

[2]Department of Microbiology & Immunology, Stanford University School of Medicine, Stanford, CA 94305-5124, USA

[3]Infectious Diseases Section, Veteran Affairs Palo Alto Health Care System, Palo Alto, CA 94304-1207, USA

Abstract

Despite the importance of horizontal gene transfer for rapid bacterial evolution, reliable assignment of mobile genetic elements to their microbial hosts in natural communities such as the human gut microbiota remains elusive. We used Hi-C (High-throughput chromosomal conformation capture), coupled with probabilistic modeling of experimental noise, to resolve 88 strain-level metagenome-assembled genomes of distal gut bacteria from two subjects, including 12,251 accessory elements. Comparisons of 2 samples collected 10 years apart for each of the subjects revealed extensive *in situ* exchange of accessory elements, as well as evidence of adaptive evolution in core genomes. Accessory elements were predominantly promiscuous and prevalent in the distal gut metagenomes of 218 adult subjects. This work provides a foundation and approach for studying microbial evolution in natural environments.

One of the major forces shaping the genomic landscape of microbial communities is horizontal gene transfer (HGT)[1]. HGT is of particular importance for the human gut microbiome, where it is involved in the emergence of antibiotic-resistant bacterial strains and mobilization of virulence factors[2,3]. In comparison to other microbial communities,

[*]Corresponding author. relman@stanford.edu.

Author Contributions
E.Y. and D.A.R. designed the study. E.Y. developed the methodology, performed and supervised the experiments, and performed the analysis. E.Y. and D.A.R. reviewed the analysis and wrote the manuscript.

Ethics Statement
Written, informed consent was obtained from each of the human subjects prior to fecal sample collection in compliance with a human subjects protocol (IRB-25268) approved by the Stanford University Administrative Panels on Human Subjects in Medical Research.

Data availability
Unprocessed DNA sequence reads and recovered MAGs are available in the NCBI database, under project PRJNA505354. MAGs can be downloaded from https://purl.stanford.edu/fd871xp9063.

Code availability
HPIPE is available for download as an open-source tool at https://github.com/eitanyaffe/hpipe.

Competing interests
The authors declare no competing interests.

human and other animal gut microbiotas show evidence of especially widespread HGT among bacterial members[4]. Moreover, there is mounting evidence of HGT between bacterial pathogens and commensals, based on *in vitro* experiments[5] and animal models[6–8]. Because strains can persist for decades within the same subject[9], the human gut microbiota has the potential to reveal quantitative and time-resolved aspects of HGT in a natural setting, with implications for both microbial evolution and human health.

The genome of any specific microbe is a mosaic of components that follow distinct evolutionary paths, ranging from tightly coupled, co-evolving house-keeping genes, to a collection of loosely associated mobile elements, including bacteriophages, transposons, plasmids, and other non-essential genes[10]. Comparisons of closely related genomes for most generalist microbial species (representing strains of the same species) identify a set of genes that are shared by all strains (‘core’), and a remaining set that are present in only a subset of strains (‘accessory’). These accessory genes contribute to the genetic diversity of the species and the capacity for adaptation to new environmental challenges and conditions[11]. Computational methods based on gene co-occurrence patterns across individuals have identified core genomes from human gut metagenomic data; however, linkage of accessory elements with their hosts has been limited to simple cases of species-specific elements, such as narrow-host-range bacteriophages[12].

De novo genotyping of microbial communities with a complex population structure, such as the human gut microbiota, is challenging for several reasons. First, a community may contain multiple conspecific strains[13]. Second, promiscuous mobile elements may be harbored by multiple microbial hosts in the same community[14,15]. These features of the genomic landscape prevent robust recovery of metagenome-assembled genomes (MAGs) from complex communities using standard approaches, such as metagenomic binning[16]. Thus, while core genomes can be inferred from metagenomic data with current methods, characterization of mobile elements and their linkage to host species in natural settings remains elusive.

Hi-C is a fixation-based method for estimating the probability of close physical proximity between DNA fragments[17]. A single Hi-C assay typically produces millions of ‘contacts’, where each contact reflects two sequence fragments that were adjacent in three-dimensional space at the time of fixation. Hi-C maps have revealed large-scale chromatin structures involved in genome regulation in eukaryotes[18,19]. More broadly, the technique has been used to study DNA folding across the tree of life, from bacteria to mammals[20–22], and to perform *de novo* genome assembly of isolated species[23–26]. When applied to microbial communities (‘metagenomic Hi-C’), the global nature of Hi-C enables the study of multiple genomes simultaneously[27–29]. Hi-C has enhanced genome co-assembly, as shown with synthetic bacterial communities[28], and has facilitated the association of extra-chromosomal DNA with the chromosomes of their microbial hosts[29]. Hi-C has provided insights into virus-host interactions in the mouse gut[30] and resolved diverse microbial genomes in the human gut[31]. However, direct modelling of noise, in the form of spurious inter-cellular contacts, has not been performed thus far for metagenomic Hi-C[32], confounding the interpretation of the data.

Here we applied Hi-C to genotype two subjects, recovering 88 MAGs with species-level reference genomes. Accessory elements, defined as groups of adjacent genes that lack homologs in the reference genome, comprised on average a quarter of each MAG. Samples collected ten years apart from each of the subjects yielded a total of 12 MAGs that corresponded to strains that persisted during the entire period. MAG analysis revealed dynamic accessory genomes, along with evidence for adaptive evolution in core genomes. Finally, the majority of the accessory elements identified in the two subjects were prevalent in gut metagenomes of 218 additional adult subjects, where they showed promiscuous associations with diverse microbial species.

Results

Stool was collected from a healthy adult (subject A); DNA was extracted, paired-end sequenced, and the resulting 202M (million) paired reads were compiled into a metagenome assembly (N50 measure of 4.7Kb), composed of 308K (thousand) contigs (consensus DNA regions) that collectively spanned 648Mb. The same sample was assayed in triplicate using the Hi-C protocol as described in Marbouty *et al.*[27], with minor adaptations (Methods). Briefly, stool was treated with formaldehyde, and cells were lysed. DNA was digested using the restriction enzyme DpnII, ligated under dilute conditions using T4 ligase, sheared and size-selected (>500bp), and paired-end sequenced with 1.4B (billion) Hi-C read pairs in total. After quality filtering, 797M read pairs were mapped successfully back onto the assembly. Within contigs, the density of mapped reads varied inversely with the genomic distance between the two paired ends, confirming that the global and stochastic nature of Hi-C data was recapitulated in our system (Extended Data Fig. 1). Technical replicates were correlated (Spearman coefficient between inter-contig read count matrices was >0.72) and were therefore united. Downstream analysis was limited to 37.5M inter-contig read pairs (5.6% of total reads). By locating nearby DpnII restriction sites, each read pair was converted into a *contact*, which is a pair of restriction fragment ends that were inferred to have been ligated during the procedure. The resulting contact map contained 10.3M unique inter-contig contacts.

Genotyping microbial communities using Hi-C

We use the term *genome configuration* to refer to a set of contigs that represent the genomic capacity (including extra-chromosomal DNA) of a clonal strain (Supplementary Note 1). In a community composed of distantly related strains that do not exchange genes, there is a one-to-one correspondence between strains, configurations and genomes, as contigs are unambiguously related to a single population and genome. The relationship is more complex when the community contains conspecific strains, or when mobile genetic elements are shared between species (Fig. 1a). In such cases, near-identical DNA sequences that belong to distinct strains are implicitly merged during the assembly process, resulting in partially overlapping configurations. To address this problem, we focus on finding clusters of contigs we call *anchors*, where (1) each anchor is a subset of the intersection of one or more overlapping configurations, and (2) no configuration contains contigs belonging to two distinct anchors. Anchor are operationally defined contig sets that provide a species-level representation of a potentially complex configuration space.

We developed HPIPE, an algorithm that recovers anchors and infers a model that predicts the probability of an inter-cellular contact between two restriction fragments, as a function of fragment lengths and abundances (Methods). The model and anchors are co-optimized such that upon convergence each anchor is enriched for intra-anchor contacts relative to the model, and the contact enrichment between two different anchors matches the level predicted by the background model. In a final step, each anchor is extended into a *genome union*, by adding to it contigs that are enriched for anchor-specific contacts using multiple criteria. Each genome union (referred to interchangeably as a metagenome-assembled genome (MAG) throughout this work) represents the combined genome capacity of one or more conspecific strains that are associated with an anchor, potentially including shared genetic elements (Fig. 1b). The reduced representation of the genomic landscape using anchor-union pairs creates a unique opportunity to characterize genome structure in complex communities, which we exploited here to study HGT.

Application of the method to the human gut microbiome

First, we tested our approach on two simple datasets. Application of the method to a simulated contact map generated for a community composed of 55 common gut bacteria, with varying degrees of relatedness and abundance (GOLD database[33], Supplementary Table 1), resulted in 32 anchor-union pairs. Importantly, the probability of detecting a community member was associated with its abundance, confirming the non-biased nature of the method (Extended Data Fig. 2). Application of the method to published Hi-C data, generated from a synthetic microbial community composed of 5 strains[29], resulted in the recovery of all species-level genomes, while merging two conspecific strains into a single anchor-union pair, confirming the ability of the method to work with real data (Extended Data Fig. 3).

We then applied the method to the contact map of subject A, resulting in 83 anchors (1.2Mb median anchor length). Thousands of spurious contacts between pairs of anchors were detected, yet the inferred background model was accurate in predicting this noise (Pearson=0.96, Fig. 2a). Each anchor was extended to a matching MAG, using stringent criteria (≥10-fold contact enrichment and ≥8 contacts, Extended Data Fig. 4). A sensitivity test showed that varying the threshold parameters in this range had only a minor effect on MAG size and quality (Supplementary Fig. 1). Contigs that were not associated with any anchor were discarded from downstream analyses. The resulting 83 MAGs (2.7Mb median per MAG) accounted for 75% of the estimated DNA mass in the sample. The remaining 25% came primarily from low abundant members of the community (Fig. 2b).

Genome completeness and contamination were estimated for all 83 MAGs using the presence of universal single-copy genes[34]. Completeness was correlated with MAG abundance (Spearman=0.36), and not with median contig length (representing assembly fragmentation, Spearman=−0.09), indicating that the major limiting factor for genome recovery in our community was sequencing depth. We examined 53 MAGs that were draft-quality or better (>50% complete and <10% contaminated, Fig. 2c), and for each sought a single reference genome within the same species. We selected the most-closely related publicly-available genome, which was defined as the reference genome with the most conserved sequence (Methods). Nine of the 53 MAGs lacked a species-level reference altogether, underscoring the still-incomplete characterization of the human gut microbiota, despite extensive study (Extended Data Fig. 5). Downstream analysis was limited to the remaining 44 MAGs with a species-level reference.

Our results were comparable, in terms of the number of MAGs and their quality, to a state-of-the-art metagenomic binning method[35], and a recently published Hi-C binning method[36] (Extended Data Fig. 6). However, the anchor-MAG approach we have implemented is unique in its ability to recover genetic overlaps between MAGs.

Characterization of core and accessory genes

For each MAG, we defined the *core genome* to be the portion of the MAG with >90% nucleotide sequence identity to the reference, and the *accessory genome* to be the remaining portion of the MAG (Fig. 3). Cores were on average 35% larger than their matching anchor, due to stringent anchor criteria (Extended Data Fig. 7). The accessory component was 25% (+/− 8.6%) of each MAG, and accounted for 24,147 genes in total, grouped by adjacency into 6391 accessory elements. Most cores showed high sequence conservation (>99% nucleotide identity) with respect to their reference, while accessory components diverged by hundreds of genes, highlighting the contribution of HGT to strain diversification (Fig. 4a). We reasoned that if within-subject HGT is ongoing in these subjects then it may be manifest by the sharing of mobile genetic elements between microbial hosts (i.e., donor and recipient strains). Indeed, using Hi-C, a total of 264 elements (1086 genes) were robustly associated with multiple host MAGs in a single community at one point in time. Sharing was associated with sequence similarity but extended across family-level boundaries (Fig. 4b). The fraction of host pairs that shared elements increased from 4% to 84% as the host amino acid identity varied from 50% to 60%, confirming phylogenetic relatedness as a major determinant of HGT compatibility (Fig. 4c). Strikingly, 96 elements (307 genes) were shared by 3 or more microbial hosts, and some by as many as 6 hosts (Fig. 4d).

To explore HGT dynamics and gut colonization history in greater depth, we estimated the within-subject polymorphism levels of cores, by mapping metagenomic reads back onto the assembly and computing the densities of intermediate SNPs (single nucleotide polymorphisms with allele frequencies ranging from 20%–80%) (Methods). As shown in Fig. 4e, the majority of cores had low polymorphism levels ($<10^{-4}$ SNPs/bp), consistent with a dominant clonal population that has experienced a recent within-subject bottleneck (based on mutation accumulation rates in the range of 10^{-8} to 10^{-5} substitutions/bp per year, measured across diverse bacteria[37]). At the tail of the distribution, the most highly polymorphic cores likely represent distinct colonization events of conspecific strains, as they have polymorphism levels close to those that are typical for unrelated strains. Polymorphism levels were also estimated for 9 shared elements (out of 264), for which sufficient data were available (>10× coverage and >10kb in length). Strikingly, all 9 elements were highly clonal ($<2*10^{-4}$ SNPs/bp), indicating they were likely spreading *in situ* (within the gut). To directly quantify HGT rates, we next took a direct approach by using stool collected from the same person 10 years prior.

Within-subject evolution over a 10-year period

We analyzed temporal changes in gene sequence and gene content, via metagenomic sequencing of a sample collected from the same subject 10 years prior to the genotyped sample. DNA was extracted and sequenced (320M reads), and reads were mapped to the 44 MAGs described above. An investigation of mapped reads allowed us to detect genetic changes ranging from single nucleotide substitutions to the turnover of entire genes (Fig. 5a). A total of 18 genome cores were not detected in the sample collected 10 years prior. The read coverage for 24 of the remaining 26 MAGs was sufficiently high (>10×) to compute the core distance between the contemporary and past samples (Fig. 5b). A total of 3 MAGs accumulated low-level mutations (using a threshold of 10^{-4} substitutions/bp, based on empirical data[37]) and were classified as 'persistent', while the remaining 21 were classified as 'replaced'.

We applied the same analysis to the 6391 accessory elements, classifying 3226 (51%) as 'not-detected', 1265 (19.8%) as 'replaced', and 1188 (18.6%) as 'persistent'. The remaining 675 elements (10.6%) were detected 10 years prior but had low read coverage (<10×), confounding the differentiation between 'replaced' and 'persistent'. Compared to elements associated with a single microbial host, shared elements were enriched for persistence and replacement (Fig. 5c). Analysis of element class, stratified by the associated host class, showed that elements did not always share the same history as their identified host (Fig. 5d). For example, out of 434 elements associated with persistent hosts, 341 (78.6%) were classified as persistent, while 83 (19.1%) were classified as 'not-detected' or 'replaced', suggesting that gene flux had occurred during that time period. Surprisingly, we also observed the reverse scenario, in which an accessory element seemingly predated its host in the gut: out of 2137 elements that were associated with 'not-detected' hosts, 45 (2.1%) were classified as 'persistent'. These 45 elements provide direct evidence for dissemination of mobile elements within a single gut community, and a contrasting view to the idea of mobile elements as highly transient.

These intriguing findings led us to study a second individual (subject B), in an attempt to develop a more general understanding of HGT in the gut. In the case of subject B, we genotyped an early sample using Hi-C (650M Hi-C reads) and used a second sample collected 10 years later in order to track genetic changes (the reverse strategy to that used in subject A). The early sample of subject B generated 87 partial MAGs, 44 of which were draft-quality or better and had a species-level reference (Extended Data Fig. 8). The MAGs of subject B contained 25,327 accessory genes in total, grouped by synteny into 5860 elements; these genes accounted for 24% (+/− 10%) of each MAG on average. DNA was extracted from the later sample of subject B and sequenced with 100M reads. Polymorphism levels and element classification distributions were remarkably similar between subjects (Extended Data Fig. 9). However, the gut community of subject B displayed greater levels of stability compared to subject A, with 9 bacterial hosts that were classified as persistent (Fig. 5e). By considering the 12 persistent MAGs identified in both subjects, we could directly compare the contribution of accessory gene turnover and nucleotide-level changes (Extended Data Fig. 10). HGT rates for the persistent MAGs were 4–19 genes/year (median 12 genes/year), and exceeded core site substitution rates for all but one MAG (Fig. 5f). The HGT rates also superseded by an order of magnitude previous estimates that were computed using long evolutionary branches[38]. These rapid HGT rates are in agreement with previous work that has shown that mutation accumulation rates are inversely correlated with the sampling time[37].

To characterize whether selection was driving these rapid genetic changes, we performed the McDonald-Kreitman test[39] for each MAG, by comparing within-subject polymorphism levels and divergence from the 10-year distant sample (Extended Data Fig. 10). The test indicated that some of the bacteria were evolving under strong adaptive (positive) selection during the 10-year period, while for others, the data were consistent with evolution in equilibrium (Fig. 5g). Although the test was significant for only 2 MAGs, pooling across all 12 MAGs boosted the significance dramatically (χ^2 test $P<10^{-7}$). Together, the data suggested that gut bacteria evolve under a combination of varying levels of adaptive selection and extensive HGT.

Species specificity and prevalence of accessory genes

To extend the results obtained from the 2 subjects and gain a population-based perspective on accessory genes, we used publicly available human gut metagenomes from 218 individuals (Supplementary Table 2). Reads were mapped using an efficient k-mer based approach to the assemblies from subjects A and B, and coverage vectors that spanned the 218 individuals were generated for all cores and elements (Methods). Each vector reflected the presence (>97% nucleotide identity) of either a core or an element across the cohort. The relationship between vectors of elements and of cores indicated the population-wide specificity of elements for their hosts, beyond the particular host-element associations observed in the MAGs recovered from the two local subjects. At one extreme, a narrow-range element (for example, a species-specific bacteriophage) is expected to be present only when its host species is present, while at the other extreme, the presence of a broad-range element will be uncorrelated with the presence of the subject-specific host. Following this approach, we classified 12.4% and 15.1% of the elements of subjects A and B, respectively, as narrow-range, while the majority of the elements (69.6% for A and 73.8% for B) were classified as broad-range (Fig. 6a). When considering the contribution to any specific MAG, broad-range elements accounted for an average of 12.3% and 13.8% of each MAG of subject A and B, respectively, compared to only 3.5% and 4.7% for narrow-range elements (Fig. 6b). To obtain a more refined understanding of the host-specificity of broad-range elements we computed a species specificity score, defined as the Pearson correlation between the element vector and the vector of the host MAG of that element in subjects A and B (or union of all host vectors, in the case of a shared element). Specificity scores ranged between 0 and 1, suggesting that a substantial portion of broad-range elements were decoupled from their locally inferred hosts (Fig. 6c). Unlike narrow-range elements which were rare, broad-range elements were found to be highly prevalent across the population, in levels comparable to microbial hosts (Fig. 6d). Compared to narrow-range elements, broad-range elements were enriched for persistence and replacement during the ten-year period (Fig. 6e). Together, the systematic analysis of hundreds of healthy individuals indicated that accessory elements are predominantly promiscuous and prevalent in human gut microbiotas.

Finally, a functional analysis based on gene ontology (Supplementary Note 2) highlighted conjugative elements (extrachromosomal and integrated) and to a lesser extent, bacteriophages, as the drivers of HGT within and between subjects.

Discussion

We developed a culture-free genotyping method to characterize genome dynamics in microbial gut communities. Analysis produced estimates of HGT rates and evidence for adaptive evolution acting on some members of the community. The approach presented here combines Hi-C with a probabilistic framework to represent complex population structures, and is well poised to make significant inroads towards an understanding of complex microbial community structures and dynamics, such as those found in soil, which routinely defy standard binning and other approaches. While there are alternative genotyping approaches based on long-reads[40,41], Hi-C is notable for its ability to provide proximity information across millions of base-pairs of contiguous sequence, including inter-molecular contacts, as demonstrated by the association of plasmids with their respective host chromosomes. The limitations of the method include possible strain interference (i.e., fragmented assemblies due to the presence of conspecific strains) and possible differing experimental efficiencies (e.g., differential lysis of cell walls or resistance to restriction enzymes). However, a more obvious limiting factor is sequencing depth; a back-of-the-envelope calculation suggests that the allocation of 1 billion reads results in an abundance detection limit of 0.1%, and the detection limit is expected to drop linearly with sequencing depth.

Recent attention to microbial evolution *in situ*, long appreciated as a primary ecological process underpinning community assembly and diversification, has provided an unprecedented view of genome dynamics in natural environments, in real time, and with implications for human health. Other recent work provides independent evidence for HGT and adaptive evolution in the human gut, using an isolate-based approach focused on *Bacteroides fragilis*[42] and a reference-based approach using the pangenomes of common gut species[43]. The culture-independent and reference-free approach presented here opens the door to studying fundamental aspects of microbial evolution in complex and poorly characterized environments.

Methods

Sample collection and shotgun procedure.

Subjects A and B are healthy Western adult males who had not used antibiotics for at least 6 months prior to sampling. Fresh stool was collected and stored at −80C until processing. To generate standard DNA libraries (for the metagenomic assembly and for the temporal comparison), DNA was extracted using the AllPrep DNA/RNA Mini Kit (Qiagen), sheared and size-selected (>300bp), and paired-end sequenced using Illumina HiSeq 2500.

Hi-C procedure.

To generate the Hi-C DNA libraries, 50–100mg of stool was suspended in 10ml cold PBS, vortexed for 20min at RT, and spun down at 20g for 10m at 4C. The supernatant was centrifuged at 5000g for 10min, the resulting pellet was washed 2 more times in cold PBS, and the final microbial pellet weight **W** (in mg) was recorded. The pellet was suspended in 5.5ml PBS, fixated with 2.5ml formaldehyde 16% (final 5%) for 30min at RT and 30m on ice. The reaction was quenched with 1525ul glycine 2.5M (final 0.4M) for 5min at RT and 15min on ice. Fixated cells were washed twice with 10ml cold PBS, suspended with 4×**W**ul of H_2O (4 times the recorded microbial pellet weight **W**), and 50ul aliquots of the fixated cell pellet were stored at −80C. For lysis, 10ul fixated input (~2mg of microbial pellet) were suspended in 190ul TE and 1.1ul Ready-Lysozyme 36KU/ul (final 200U/ul), and incubated 15min at RT with occasional pipetting. Next, 10ul SDS 10% (final 0.5%) was added and samples were incubated for 10min at RT (total reaction volume, 200ul). For digestion, 150ul H_2O, 50ul 10× DpnII buffer, 50ul Triton 10% (final 1%), and 50ul DpnII restriction enzyme (final 5U/ul) were added, and samples were incubated at 37C for 3hrs (final reaction volume, 500ul). Samples were incubated 10min with 25ul SDS 10% (final 0.5%) at RT. For ligation, 800ul Triton 10% (final 1%), 800ul 10× T4 buffer, 80ul 10 mg/ml BSA and 5800ul H_2O and 20ul T4 ligase (final 2000U/ul) were added, and the sample was incubated for 4 hours at 16C (final reaction volume, 8ml). Following ligation, 100ul Proteinase K 20ug/ul (final 250ug/ml) was added and samples were incubated overnight at 65C. DNA was then cleaned with phenol-chloroform, precipitated in ethanol, suspended in 500ul TE, transferred to 1.5ml tubes, and incubated 1hr at 37C with RNase 0.5ug/ul (final 30ug/ml). DNA was cleaned with 2 more rounds of phenol-chloroform, ethanol precipitated, washed twice with 70% ethanol, and eluted in TE. DNA was sonicated, size-selecting for fragments 500–800bp and paired-end sequenced using Illumina HiSeq 2500.

Preprocessing raw reads.

Identical duplicate reads were removed, reads were quality-trimmed using Sickle[44] with default parameters, adaptor sequences were removed using SeqPrep[45] (min length of 60nt), and human sequences were removed using DeconSeq[46] (alignment coverage threshold 10%, identity threshold 80%), resulting in unique high-quality non-human paired reads.

Metagenomic assembly.

De novo metagenome assembly was performed using MEGAHIT[47] with parameters "--min-contig-len 300 --k-min 21 --k-max 141 --k-step 12 --merge-level 20,0.95", and filtering out contigs shorter than 1kb. For mapping reads onto the assembly, the first 10nt of each read were trimmed, and the following 40nt were mapped using BWA-MEM[48] with default parameters. Low quality or non-unique reads (>0 mismatches, <30nt match length or mapping score <30) were filtered out.

Hi-C contacts.

Contigs were pairwise aligned using Mummer[49], identifying identical stretches of sequence (>=20nt long) shared between pairs of contigs. If the two sides of an inter-contig Hi-C paired read mapped up to 2000bp away from a perfect alignment region, the read was filtered out. The restriction enzyme that was used (DpnII) induces a partitioning of all contigs into restriction fragments. Every Hi-C ligation event ('contact') occurs between two fragment ends. To infer a contact from a mapped read pair, the contig was scanned from the mapped read coordinate, in the direction of the mapped read strand, until the first DpnII restriction site was reached, separately for both sides of each read pair. To minimize sequencing amplification noise, contact multiplicity was ignored, i.e. only unique contacts were considered.

Inference of anchor-union pairs.

We defined the *abundance* of a contig c to be the normalized read-coverage $H(c) = \frac{L(M)R(c)}{L(c)R_{total}}$, where $R(c)$ is the number of Hi-C reads that mapped to c, R_{total} is the total number of reads in the library, M is the set of all contigs in the metagenome assembly, and $L(X)$ is the total length in base pairs of a contig set $X \subseteq M$. We defined the *weighted mean abundance* of a contig set $C \subseteq M$ to be $H_{\mu}(C) = \frac{\sum_{c \in C} L(c)H(c)}{\sum_{c \in C} L(c)}$, the *weighted standard deviation* to be $H_{\sigma}(C) = \sqrt{\frac{\sum_{c \in C} L(c)\left(H(c) - H_{\mu}(C)\right)^2}{\sum_{c \in C} L(c)}}$, and the *abundance z-score* of a contig $c \in C$ to be $Z_C(c) = \frac{H(c) - H_{\mu}(C)}{H_{\sigma}(C)}$.

We modelled the probability of a spurious contact between two fragment ends x, y as:

$$P(x, y) = N \bullet H(x) \bullet H(y) \bullet F_{len}\left(B_{len}(x), B_{len}(y)\right)$$

where N is a normalizing constant, $H(x)$ and $H(y)$ are the abundances of the contigs on which the fragments with ends x and y reside (respectively), and F_{len} is a function that transforms a pair of binned values $B_{len}(x)$,$B_{len}(y)$ of fragment lengths into a single empirical correction factor.

Given a spurious model P and constants $\alpha, \beta \in \mathbb{R}$, we denoted two disjoint contig sets $X, Y \subseteq M$ as *(α,β) -associated* if (1) X and Y were connected by at least α contacts, (2) the number of connecting contacts was at least β-fold enriched over the spurious contacts predicted by the model P, and (3) the false positive binomial probability for the observed contacts was below 10^{-6}. The inferred *anchors* were a disjoint collection of contig sets $\mathbb{A}$, for which each anchor

$A \in \mathbb{A}$ satisfied these five conditions:

(A1) **Clique**: Over 90% of pairs of contigs $a, b \in A$ were associated by one or more contacts.

(A2) **Association**: Every contig $a \in A$ and $A \setminus a$ were associated, with α=5,β=1.6.

(A3) **Uniqueness**: No contig $a \in A$ and $A' \in \mathbb{A} \setminus A$ were associated, with α=5,β=1.6.

(A4) **Size**: Every contig $a \in A$ was ≥10kb, and the total length of contigs in A was ≥200kb.

(A5) **Abundance**: $H_{\sigma}(A)$ ≤0.2, and for all $a \in A$ the z-score $Z_A(a)$ ≤1.5.

The model P and anchors $\mathbb{A}$ were inferred simultaneously. Briefly, seed anchors were computed using hierarchical clustering. A seed model was inferred over the seed anchors using maximum likelihood. Contigs that were associated with multiple anchors were discarded sequentially until convergence. Finally, anchors that were small or had a large abundance variance were discarded.

The matching genome union $A \subseteq G$ was generated by including any contig $c \in M$ that satisfied:

(G1) **Association**: The contig c and $A|c$ were associated, with α=8,β=10.

(G2) **Anchor support**: The association was supported by at least 2 anchor contigs.

(G3) **Contig suppor**t: The association was supported by least 50% of the fragment ends within the contig c.

Default HPIPE parameters were tuned to favor precision over sensitivity, and are customizable. See the SI methods section for a complete description of the algorithm.

Validation on simulated communities.

Reference genomes for 55 common gut bacteria (GOLD database[33]) were downloaded from NCBI (Supplementary Table 1). The contigs of each reference genome were concatenated into a single circular pseudo contig. Genomes were ordered randomly and assigned an × fold-coverage value that ranged between 1 and 1000 following a geometric progression. To generate the assembly library, random read pairs (2×150bp) were generated, given the assigned x-coverage for all genomes, resulting in a total of 120M read pairs. The distribution of the distances between read pairs was a Gaussian with an offset: 200bp + N (mean=800bp, sd=200bp). To generate the Hi-C library, 100M random read pairs were generated as follows. A total of 1% of reads were assigned as spurious, and were associated with two independently selected genome coordinates chosen according to genome abundance. The remaining 99% reads were assigned to genomes according to their abundance. Within each genome the distance between 50% of reads was uniformly distributed and the distance between the remaining 50% was distributed following a power law with an exponent of −1. HPIPE was run on the assembly and Hi-C library using default parameters.

Validation on a synthetic community.

Raw Hi-C sequencing data were downloaded for a clonal synthetic community[29], which was composed of 5 microbial strains: *Pediococcus pentosaceus* (ATCC 25745), *Lactobacillus brevis* (ATCC 367), *Burkholderia thailandensis* (E264) and two strains of *Escherichia coli* (BL21 and K-12). Matching reference genomes were downloaded from NCBI. An assembly library with an x-coverage of 100 was simulated, as described for the simulated community above. HPIPE was run on the simulated assembly library and downloaded Hi-C data using default parameters.

Comparison to alternative methods.

MetaBAT2 (version 2.12.1) was applied to the metagenomic assembly and the supporting reads of the assembly of Subject A, using default parameters. Bin3C (downloaded from GitHub on March 2019) was run on the metagenomic assembly and the raw Hi-C DNA library of Subject A, following the guidelines supplied by the bin3C authors, and using default parameters.

Genome sequence similarity.

Genes were predicted on all contigs using MetaGeneMark[50] and were self-aligned using DIAMOND[51] (sensitive mode, E<0.001). For all pairs of genome unions, if there were at least 12 aligned gene pairs (>30% identity and >70% coverage), the average amino acid identity (AAI) was computed by averaging the alignment identities (correcting identity for partial gene coverage, to reflect alignment over all of the gene), and otherwise it was set to 0. To generate the sequence similarity matrix (Fig. 4B), genome unions were clustered using hierarchical clustering, using AAI as the similarity metric and merging clusters using the ‘average’ method.

Taxonomic affiliation.

Single-copy gene analysis was performed using CheckM[34]. Genomes which were less than 50% complete or more than 10% contaminated were discarded from downstream analysis. Predicted genes were blast-aligned to UniRef100 (downloaded in December 2015) using DIAMOND (sensitive mode, E<0.001). For each genome union, UniRef homolog genes (>30% identity and >70% coverage) were converted into one or more corresponding NCBI taxonomic Entrez entries, and organized on a taxon tree. The number of homolog genes was propagated up the tree. A *species taxon* was determined to be the species-level tree node that (1) had the maximal gene count among all species-level nodes, and (2) had one or more available reference genomes in the GenBank database[52] (downloaded in May 2018).

Species-level reference genomes.

For each genome union, all reference genomes of the species taxon, as defined by the GenBank database, were downloaded from NCBI. For every candidate reference genome, a bi-directional mapping was performed by splitting the genome union and the reference genome into overlapping 100bp windows (sliding 1bp along the genome), and mapping in both directions using BWA-MEM[48] with default parameters. For both the genome union and the reference genome, each coordinate was assigned the maximal sequence identity of all windows that contained it, producing an *identity track* for both directions of mapping. The *alignable fraction* was defined as the portion of the genome union that was successfully mapped, averaged over both directions of mapping. The *nearest reference genome* was selected to be the reference genome for which the alignable fraction was maximal.

Core and accessory fractions.

For each genome that had a nearest reference genome, a gene-level nucleotide identity vector, was computed by averaging the mapping identity over entire genes. Genes for which the identity was 90% or more were defined as core genes, and the remaining were defined as accessory genes. A genome was classified as 'no-reference' if (1) there the assigned species taxon had no reference genomes in GenBank, or (2) the fraction of core genes was <50%. This resulted in 9 genomes for subject A and 13 genomes for subject B that lacked a species-level reference. Accessory genes were grouped into accessory elements according to synteny, i.e. if they appeared sequentially within a contig. Elements for which the gene x-coverage z-score distribution had a high standard deviation (>4) were removed from downstream analysis (in total <2.5% of elements were removed in this manner).

Polymorphism levels.

Complete assembly read sides were mapped onto the assembly using BWA-MEM[48] with default parameters. Only matches that were 100bp or more, with a maximal edit distance of 2 and a score of 30 were used. A nucleotide-level vector with the allele frequency for all 4 nucleotides was computed by parsing the SAM alignment result. A nucleotide coordinate was called *intermediate* if (1) the allele frequency f satisfied 20%<f<80%, and (2) there were at least 3 supporting reads for the allele. The *polymorphism level* (i.e. the standing variation) for a gene-set (core or element gene-set), which had a sufficient read coverage (>10×), was defined to be the mean density of intermediate SNPs over the gene-set, discarding a 200bp margin near contig edges.

10-year core and element classification.

The secondary sample, taken 10 years apart, was mapped onto the assembly using BWA-MEM, and generating a nucleotide-level vector with the allele frequencies as for the standing variation. A nucleotide coordinate was called *fixed* if (1) the dominant nucleotide was different from the assembly reference nucleotide, (2) the allele frequency was at least 95%, and (3) there were at least 3 supporting reads for the allele. The *substitution density* for a gene-set (core or element gene-set), was defined to be the density of fixed coordinates over the gene-set, discarding a 200bp margin near contig edges. A gene-set (core or element) was classified as *detected* if >90% of the genes had a median read coverage of 1× or more, and it was classified as *not-detected* otherwise. A detected gene-set was further classified as *high-detected* if (1) the median read coverage over the entire gene-set was at least 10×, and was classified as *low-coverage* otherwise. High-detected gene-sets were further classified as *persistent* if the substitution density over the gene-set was $<D_t$, and classified as *replaced* otherwise. The threshold D_t was set to 10^{-4}, based on empirical estimates of mutation accumulation rates in bacteria, that range between 10^{-8} and 10^{-5} substitutions/bp per year[37]. The *accessory divergence* of a genome was the total number of accessory genes associated with the genome that were on elements classified as not-detected or replaced.

McDonald-Kreitman test.

Test values were computed for each of the 12 genomes that were classified as persistent across both subjects. Synonymous and non-synonymous sites were determined using Translation Table 11 (NCBI). The number of synonymous (#*Ps*) and non-synonymous (#*Pn*) polymorphic sites were computed per core using intermediate SNPs in the genotyped sample ('base sample'). The number of synonymous (#*Ds*) and non-synonymous (#*Dn*) divergent sites were computed per core using fixed SNPs. Matching densities (*Ps, Pn, Ds, Dn*) were computed from raw count (adding 1 pseudo-count) by normalizing for the total number sites of each type (synonymous and non-synonymous). P-values for the McDonald-Kreitman test were generated using the χ^2 test over (#*Ps*, #*Pn*, #*Ds*, #*Dn*), while adding 1 pseudo-count to the raw numbers.

Gene ontology enrichments.

Enrichments for GO (Gene ontology) categories were computed as follows: All Uniref100 hits were transformed into GO categories, using the Uniparc and Uniprot databases as intermediates. To generate the P-values reported for a given GO category and a selected set of predicted genes, an exact Fisher test (single-tailed) was performed by comparing the selected set to a background set composed of all predicted genes that were associated with a MAG. A category was deemed significant in a subject if (1) P<0.05, (2) the enrichment over background was at least 2-fold, and (3) the number of supporting genes was at least 2. False discovery rates (Q-values) were computed by generating 1000 random gene sets with a size matching the selected gene set. Each Q-value reflects the fraction of false 'discoveries' (i.e., reported categories) that are false for the matching P-value.

Population presence analysis.

218 human gut metagenomic DNA libraries collected from distinct subjects were downloaded from the HMP and the EMBL-EBI repositories (Supplementary Table 2). Each of the 218 subject libraries was converted to a *k-table* (k=16), by counting the frequency of all k-mers across the library reads. The following analysis was performed separately for subjects A and B. Each k-table was projected on each predicted gene, generating a 1-bp vector of k-mer frequencies. The *gene coverage* was defined as the median k-mer frequency over the entire gene vector. The *gene fraction* was defined as the fraction of the gene vector that was covered by segments of hits that were at least q=30 long. The value of the parameter q was selected to balance between false positives and the detection limit, that was estimated to be 96.66% (100–100/q), assuming substitutions are disturbed uniformly. A gene g was called *present* in the library of subject i if (1) the gene fraction in library i was at least 80%, and (2) the gene coverage in the library was at least 2. The *presence value* v_g^i was set to be the gene coverage if the gene was called as present in the subject library, and set to zero otherwise, resulting for each gene g in a *gene presence vector* $v_g = \left(v_g^i\right)_{i=1}^n$ that spanned all 218 subjects.

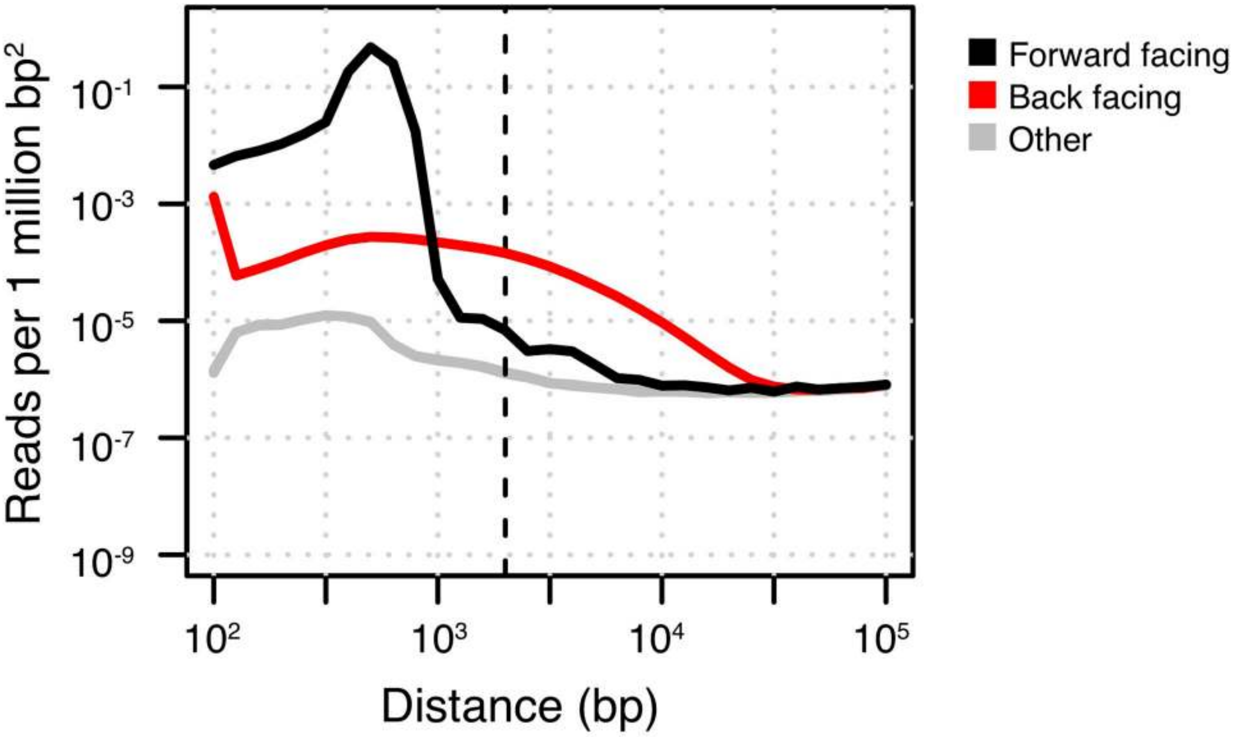
For a gene-set x (either a core or an element), the *set presence vector* v_x was defined to be a per-coordinate median over the presence vector of the genes in the gene set: $v_x = \left(v_x^i\right)_{i=1}^n = \left(median\left\{v_g^i : g \in x\right\}\right)_{i=1}^n$. In this manner presence vectors for all elements and their associated cores were computed. The *detected subject set* $s(v)$ of a presence vector v was defined to be $s(v)=\{i: v^i>0\}$. For each element e and its matching set of host cores H_e (one or more hosts), the *element host presence vector* was defined to be $v_{H_e} = \sum_{h \in H_e} v_h$. The element was classified as *rare* if the detected subject set $\left|s\left(v_e\right)\right| < 2$, as *narrow* if

$s(v_e) \subseteq s(v_{H_e})$, and as *broad* otherwise. The *element-host specificity score* was defined to be the Pearson correlation between the presence vectors $\rho(v_e, v_{H_e})$.

Association between 10-year classification and population classification.

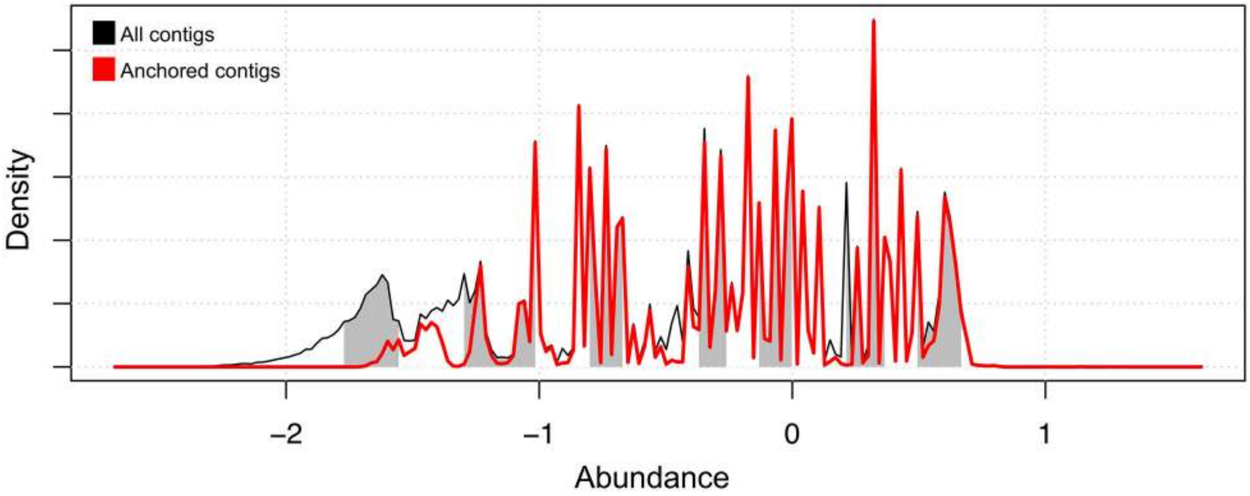
Each element was classified into 3 classes using the 10-year dataset (not-detected/low-coverage/replaced/persistent), and into 3 classes using the population dataset (rare/broad-range/narrow-range). The observed number of elements classified under all 12 combinations of classification pairs was counted. To generate Fig. 6E the observed number of elements was compared to the expected number of elements was estimated using a generalized Bernoulli distribution.

Extended Data



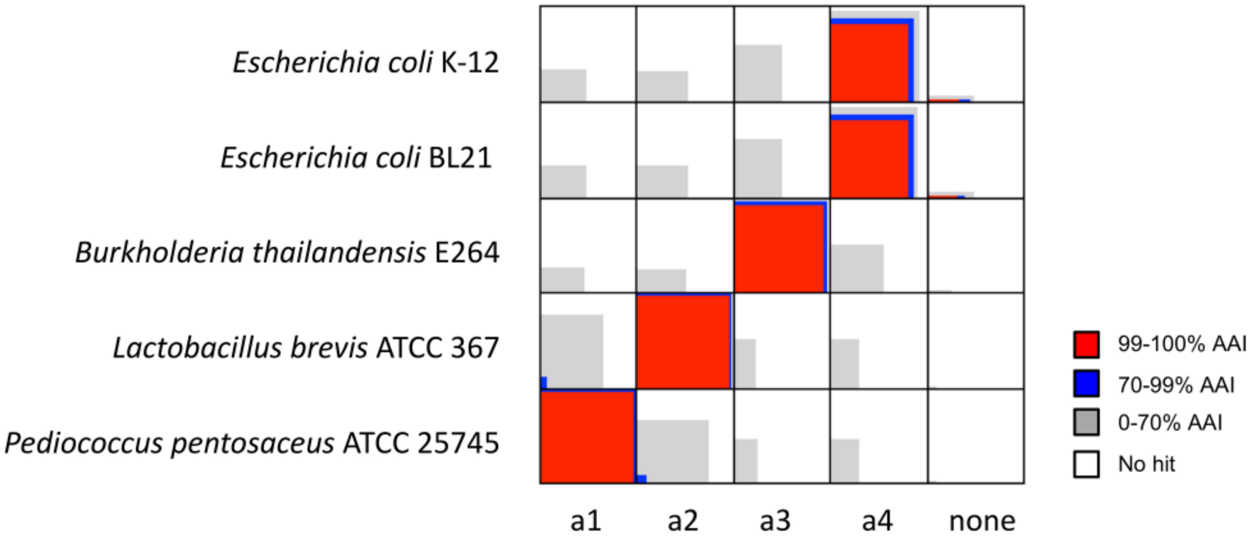
Extended Data Fig. 1. Hi-C contact density as a function of linear distance.

Intra-contig read density as a function of the distance between mapped read sides, colored according to the relative strand orientation of the two read sides.



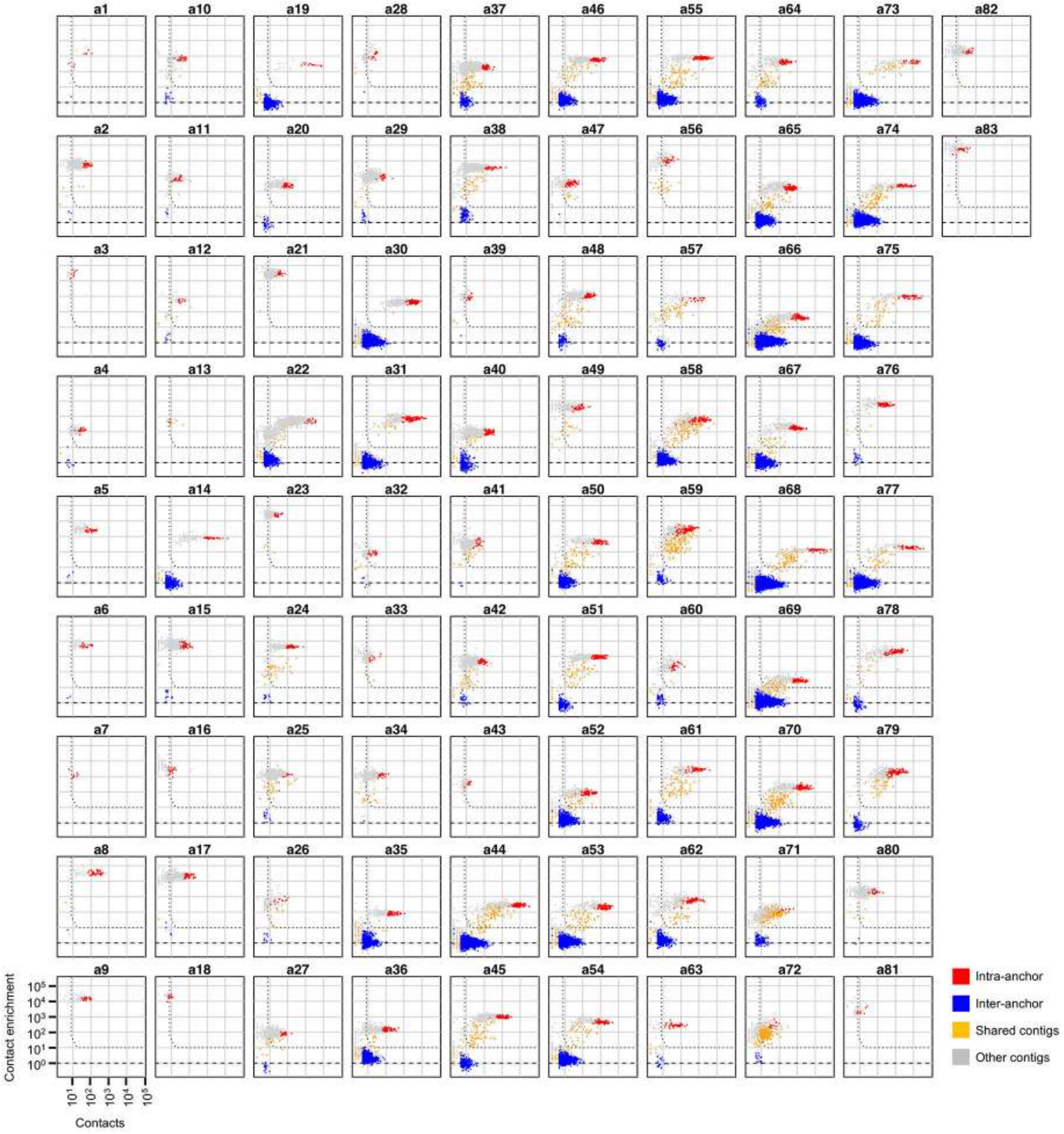
Extended Data Fig. 2. Validation on a simulated microbial community.

The genomes of 55 common gut microbes (GOLD database) were downloaded and 120M simulated shotgun reads and 100M simulated Hi-C reads were generated, with relative representation ranging from 1 to 1000. HPIPE identified 32 MAGs. Shown is the density plot of the relative abundance of the entire metagenomic assembly (contigs >1k), as in Figure 1d. The abundance is the enrichment of the read coverage over a uniform distribution of reads. White/gray stripes denote chunks of 10Mb. The fraction of the assembly that was included in any recovered MAG (‘anchored contigs’) is depicted with a red line.



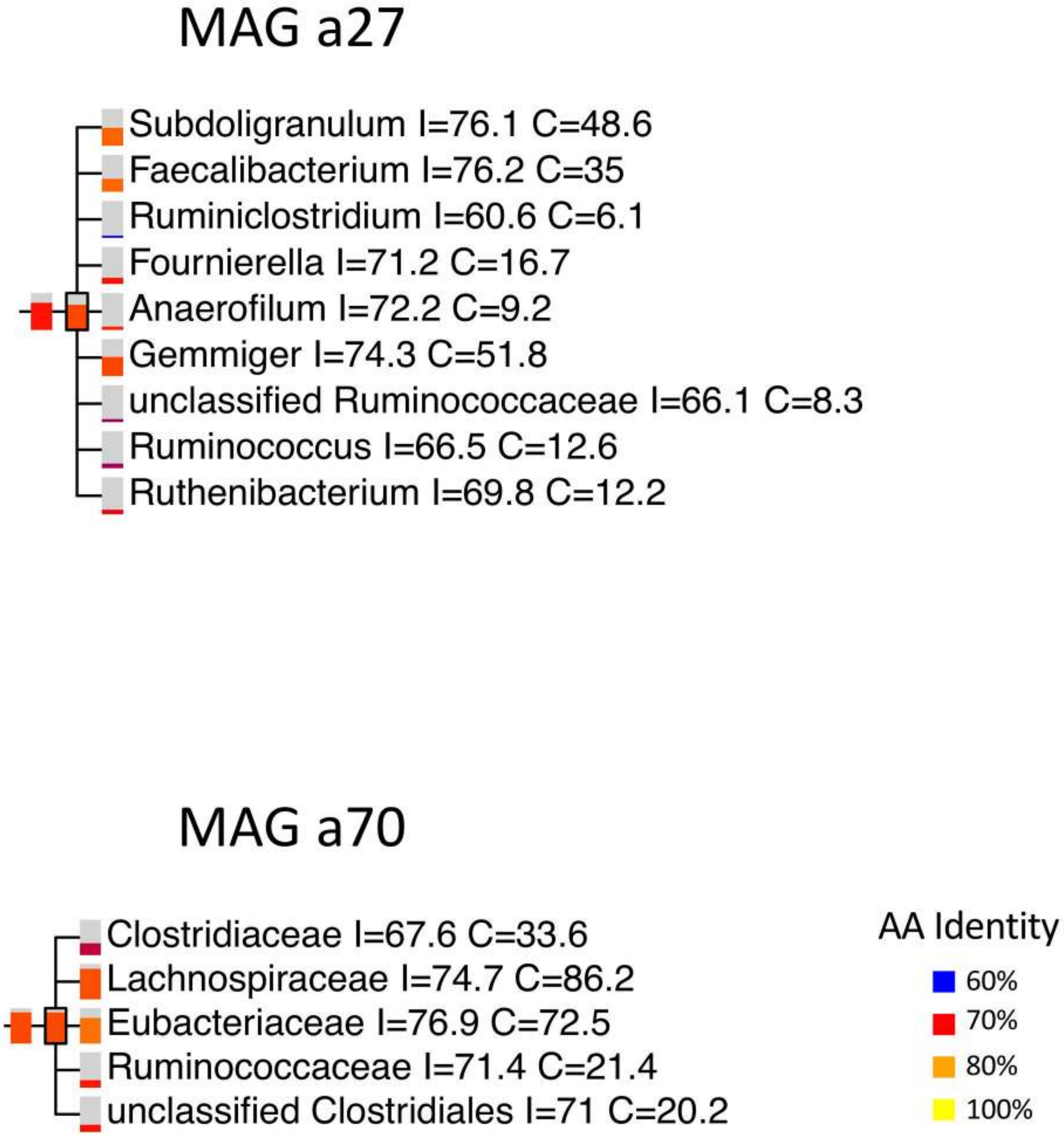
Extended Data Fig. 3. Validation on a synthetic microbial community.

The community was composed of *Pediococcus pentosaceus* (ATCC 25745), *Lactobacillus brevis* (ATCC 367), *Burkholderia thailandensis* (E264) and two strains of *Escherichia coli* (BL21 and K-12), as described in Beitel *et al.*, 2014. The pipeline recovered 4 anchor/union pairs. Shown is a pairwise gene alignment between the 4 inferred MAGs (genome unions) and the 5 reference genomes.



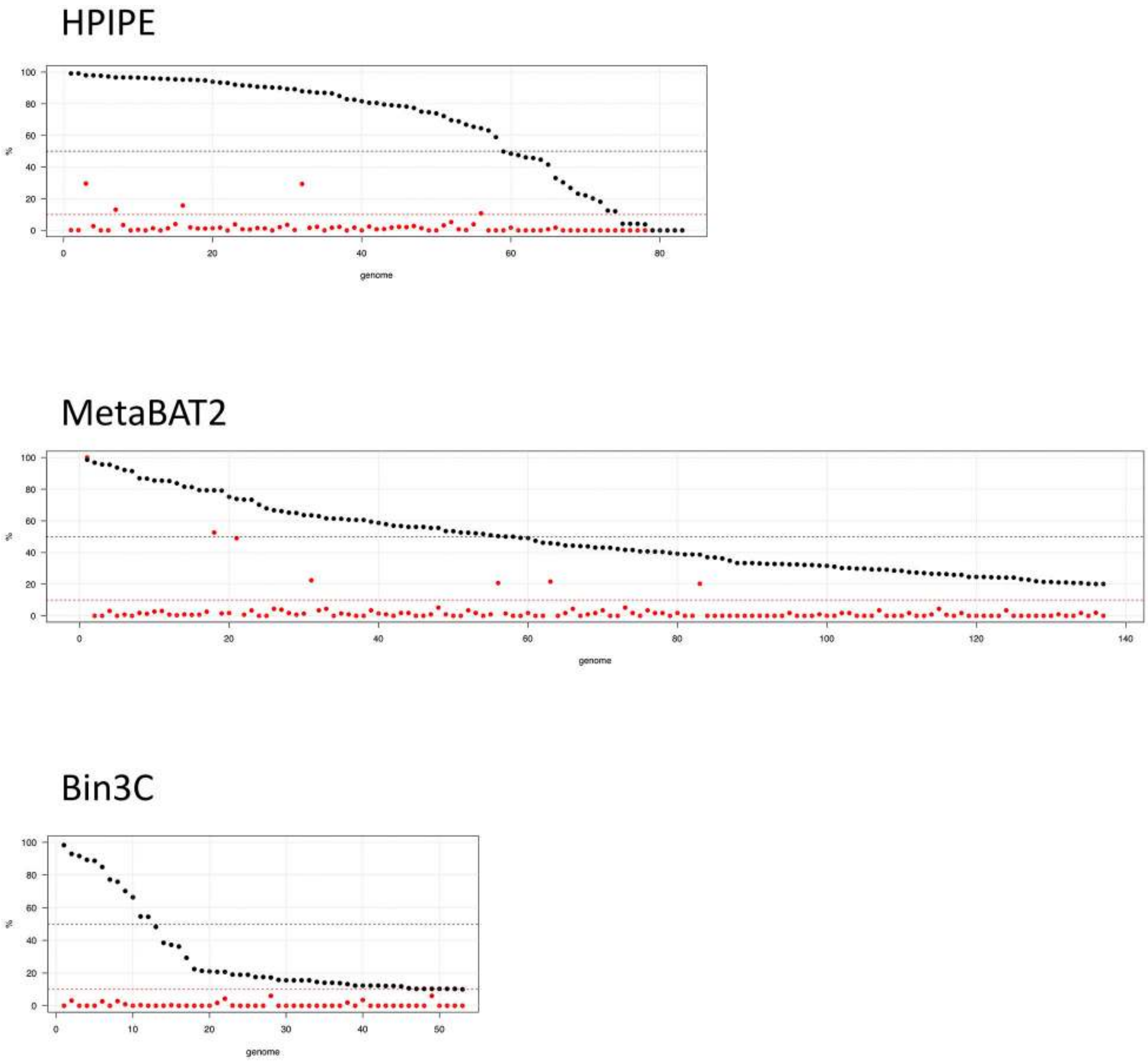
Extended Data Fig. 4. Contig-anchor contact enrichments over all anchors.

On the x-axis is the observed number of contacts between the contig and the anchor, and on the y-axis is the enrichment score over the background model. Anchor contigs are colored red, contigs belonging to other anchors are colored blue, and all other contigs are colored gray. Anchors are extended into MAGs (genome unions) by including contigs with >=10-fold contact enrichment (dashed horizontal line), >=8 contacts (dashed vertical line), and a false positive probability of 10-6 assuming a binomial distribution (transition between vertical and horizontal line).


MAG a27
Subdoligranulum I=76.1 C=48.6
Faecalibacterium I=76.2 C=35
Ruminiclostridium I=60.6 C=6.1
Fournierella I=71.2 C=16.7
Anaerofilum I=72.2 C=9.2
Gemmiger I=74.3 C=51.8
unclassified Ruminococcaceae I=66.1 C=8.3
Ruminococcus I=66.5 C=12.6
Ruthenibacterium I=69.8 C=12.2
MAG a70
Clostridiaceae I=67.6 C=33.6
Lachnospiraceae I=74.7 C=86.2
Eubacteriaceae I=76.9 C=72.5
Ruminococcaceae I=71.4 C=21.4
unclassified Clostridiales I=71 C=20.2
AA Identity
60%
70%
80%
100%

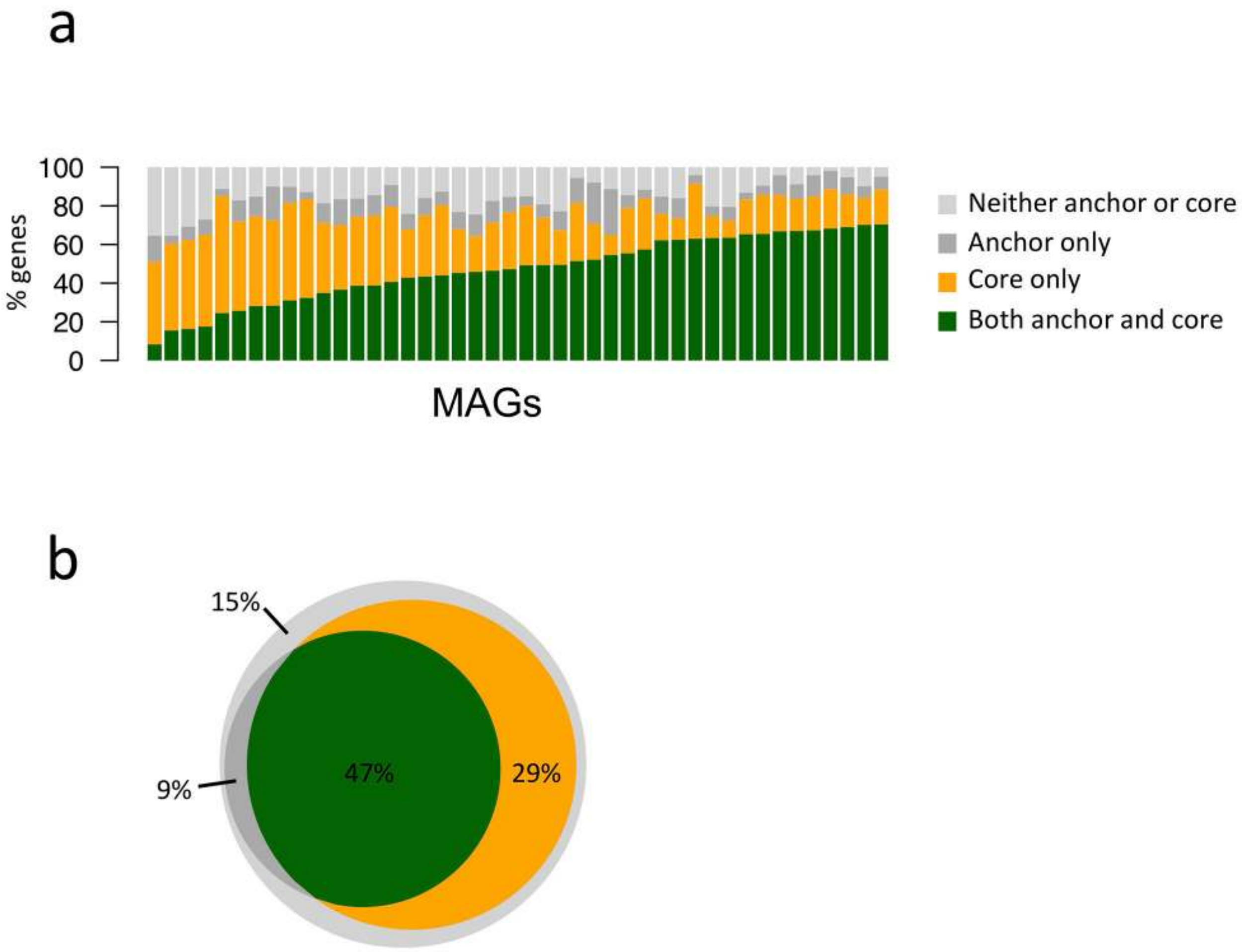
Extended Data Fig. 5. Examples of 2 putative novel MAGs.

On top, 68% of the genes of MAG a27 align to the *Ruminococcaceae* family (mean identity 74.3%), suggesting it is a novel species in that family. On the bottom, 88% of the genes of MAG a70 align to the *Clostridiales* order (mean identity 74.5%), indicating it is a novel genome within *Lachnospiraceae* or *Eubacteriaceae*. Each taxon is colored according to the mean amino acid identity, and the colored fraction of each rectangle represents the percentage of the aligned genes.



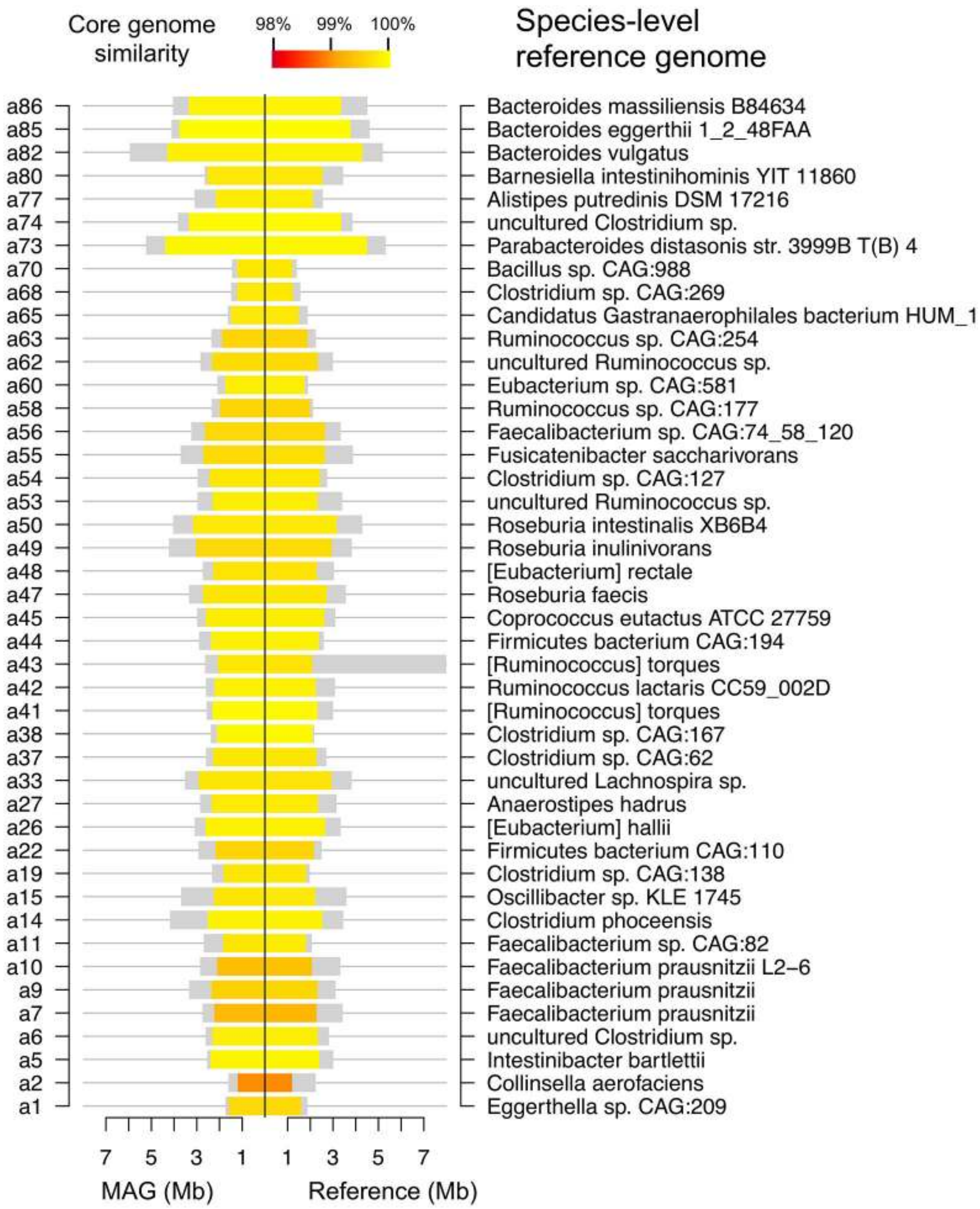
Extended Data Fig. 6. Comparison of HPIPE to alternative metagenomic binning methods.

Single-copy gene estimates of genome completeness percentage (in black) and contamination percentage (in red) with HPIPE, metaBAT2, and bin3C, sorted according to completeness. Minimal completeness (50%) and maximal contamination (10%) thresholds are depicted with dashed horizontal lines. Our results (HPIPE, as in Figure 2c), compared to metaBAT2 (tool based on abundance and tetranucleotide frequency), and bin3C (tool based on clustering of Hi-C data).



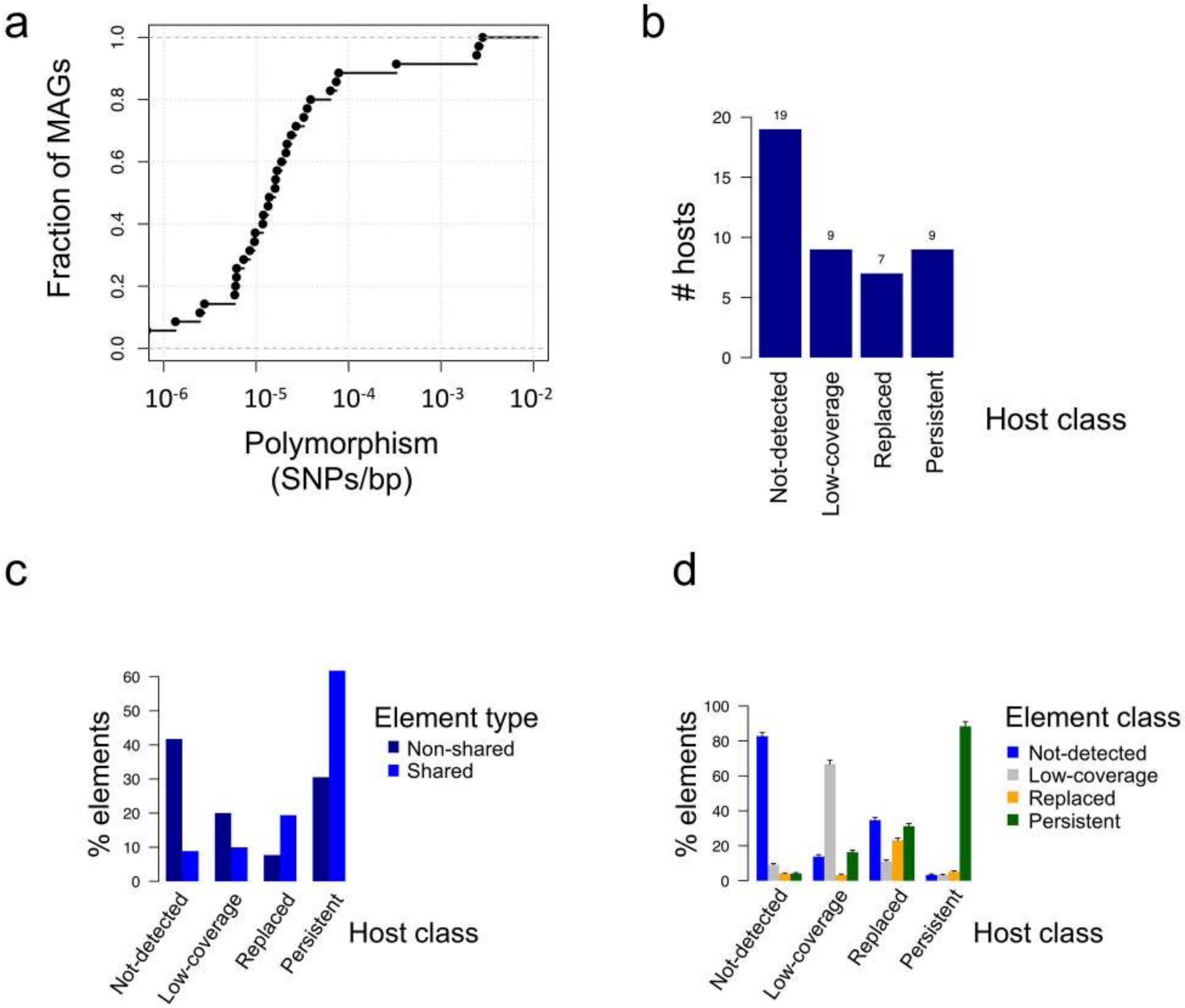
Extended Data Fig. 7. Comparison of anchors and cores.

(a) Shown for all 44 MAGs (genome unions), is the breakdown of genes into 'core-only, 'anchor-only', 'both' or 'neither', sorted according to the 'both' fraction. (b) The fraction of the 4 gene classifications, colored as in (a), averaged over all 44 MAGs. Core-only genes (29%) are present due to the stringent selection of anchors, which considers only long contigs (>10k).



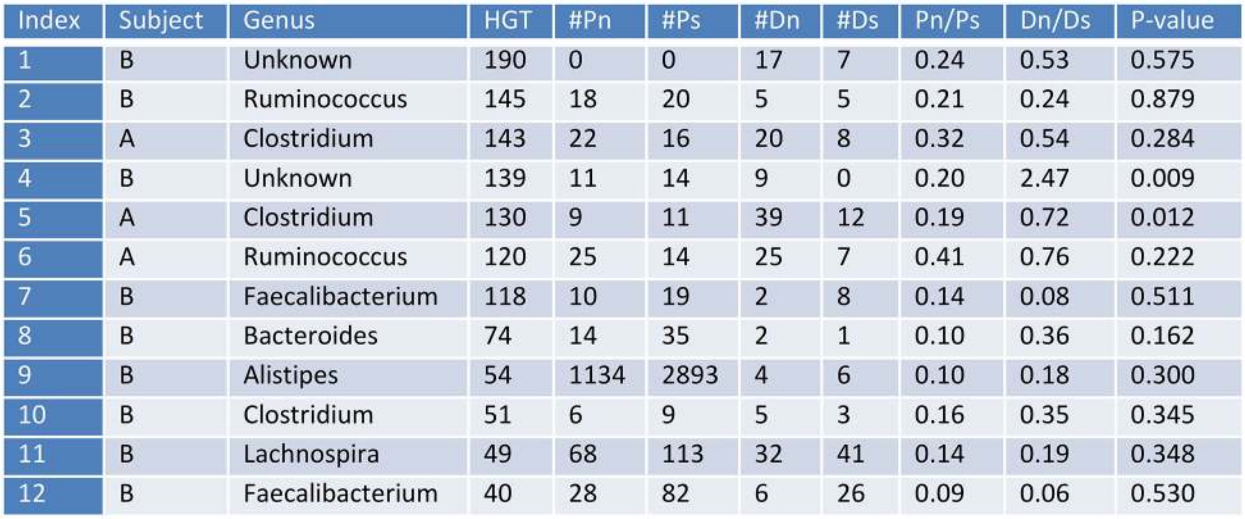
Extended Data Fig. 8. Species-level reference genomes for Subject B.

Shown are the core and accessory fractions for the 44 MAGs that had a species-level reference for Subject B. For both the recovered MAGs (left) and the matching species-level reference genomes (right), the core fraction is depicted using a colored rectangle, and the accessory fraction (i.e., strain-specific genes) is depicted using a gray rectangle. Cores are colored according to genome similarity (nucleotide sequence identity) between MAG cores and matching reference cores.



Extended Data Fig. 9. Polymorphism and 10-year divergence patterns for Subject B.

(a) Polymorphism levels, estimated using the density of intermediate alleles (SNPs with a frequency in the range 20%–80%), shown for 35 MAGs of Subject B that had at least 10x coverage. (b) Host classification for the 44 MAGs of Subject B. (c) The distribution among element classes, stratified according to element type (shared and non-shared). Data are normalized so that each type sums to 100%. (d) The distribution among element classes, stratified according to host class. Data are normalized so that each host class sums to 100%. Standard deviations are depicted using error bars.



Index	Subject	Genus	HGT	#Pn	#Ps	#Dn	#Ds	Pn/Ps	Dn/Ds	P-value
1	B	Unknown	190	0	0	17	7	0.24	0.53	0.575
2	B	Ruminococcus	145	18	20	5	5	0.21	0.24	0.879
3	A	Clostridium	143	22	16	20	8	0.32	0.54	0.284
4	B	Unknown	139	11	14	9	0	0.20	2.47	0.009
5	A	Clostridium	130	9	11	39	12	0.19	0.72	0.012
6	A	Ruminococcus	120	25	14	25	7	0.41	0.76	0.222
7	B	Faecalibacterium	118	10	19	2	8	0.14	0.08	0.511
8	B	Bacteroides	74	14	35	2	1	0.10	0.36	0.162
9	B	Alistipes	54	1134	2893	4	6	0.10	0.18	0.300
10	B	Clostridium	51	6	9	5	3	0.16	0.35	0.345
11	B	Lachnospira	49	68	113	32	41	0.14	0.19	0.348
12	B	Faecalibacterium	40	28	82	6	26	0.09	0.06	0.530

Extended Data Fig. 10. Attributes of the 12 MAGs classified as persistent over the 10-year period.

Columns indicate the Subject in whom the MAG was found, the number of non-persistent accessory genes (HGT column), the number of non-synonymous (#Pn) and synonymous (#Ps) sites that were polymorphic within the genotyped sample, and the number of non-synonymous (#Dn) and synonymous (#Ds) sites that were divergent between the genotyped sample and the 10-year sample. Matching site densities (Pn, Ps, Dn and Ds) equal the number of sites divided by the total number of sites of each type (synonymous or non-synonymous). P-values are for the McDonald-Kreitman test (χ^2), which examines whether the ratios, Pn/Ps and Dn/Ds differ significantly.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank Megan Kennedy for help in processing clinical samples and the members of the Relman and Holmes laboratories for invaluable discussion and feedback. **Funding:** This work was supported by NIH R01AI112401 (D.A.R. and Susan Holmes), EMBO Long-Term Fellowship ALTF 772-2014 (E.Y.), the Chan Zuckerberg Biohub Microbiome Initiative (D.A.R.), and the Thomas C. and Joan M. Merigan Endowment at Stanford University (D.A.R.).

References

1. Soucy SM, Huang J & Gogarten JP Horizontal gene transfer: building the web of life. Nat. Rev. Genet 16, 472–482 (2015). [PubMed: 26184597]
2. Wintersdorff, von CJH et al. Dissemination of antimicrobial resistance in microbial ecosystems through horizontal gene transfer. Front. Microbio.l 7, 173 (2016).
3. Allen HK et al. Call of the wild: antibiotic resistance genes in natural environments. Nat. Rev. Microbiol 8, 251–259 (2010). [PubMed: 20190823]
4. Smillie CS et al. Ecology drives a global network of gene exchange connecting the human microbiome. Nature 480, 241–244 (2011). [PubMed: 22037308]

5. Maiques E et al. beta-lactam antibiotics induce the SOS response and horizontal transfer of virulence factors in Staphylococcus aureus. J. Bacteriol 188, 2726–2729 (2006). [PubMed: 16547063]
6. Zhang X et al. Quinolone antibiotics induce Shiga toxin-encoding bacteriophages, toxin production, and death in mice. J. Infect. Dis 181, 664–670 (2000). [PubMed: 10669353]
7. Modi SR, Lee HH, Spina CS & Collins JJ Antibiotic treatment expands the resistance reservoir and ecological network of the phage metagenome. Nature 499, 219–222 (2013). [PubMed: 23748443]
8. Stecher B et al. Gut inflammation can boost horizontal gene transfer between pathogenic and commensal Enterobacteriaceae. Proc. Natl. Acad. Sci. U.S.A 109, 1269–1274 (2012). [PubMed: 22232693]
9. Faith JJ et al. The long-term stability of the human gut microbiota. Science 341, 1237439 (2013). [PubMed: 23828941]
10. Koonin EV & Wolf YI Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. Nucleic Acids Res. 36, 6688–6719 (2008). [PubMed: 18948295]
11. Tettelin H, Riley D, Cattuto C & Medini D Comparative genomics: the bacterial pan-genome. Curr. Opin. Microbiol 11, 472–477 (2008). [PubMed: 19086349]
12. Nielsen HB et al. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nat. Biotechnol 32, 822–828 (2014). [PubMed: 24997787]
13. Truong DT, Tett A, Pasolli E, Huttenhower C & Segata N Microbial strain-level population structure and genetic diversity from metagenomes. Genome Res. 27, 626–638 (2017). [PubMed: 28167665]
14. Brown Kav A et al. Insights into the bovine rumen plasmidome. Proc. Natl. Acad. Sci. U.S.A 109, 5452–5457 (2012). [PubMed: 22431592]
15. Jørgensen TS, Xu Z, Hansen MA, Sørensen SJ & Hansen LH Hundreds of circular novel plasmids and DNA elements identified in a rat cecum metamobilome. PLoS ONE 9, e87924 (2014). [PubMed: 24503942]
16. Alneberg J et al. Binning metagenomic contigs by coverage and composition. Nat. Methods 11, 1144–1146 (2014). [PubMed: 25218180]
17. Lieberman-Aiden E et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 326, 289–293 (2009). [PubMed: 19815776]
18. Dixon JR et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 485, 376–380 (2012). [PubMed: 22495300]
19. Sexton T et al. Three-dimensional folding and functional organization principles of the Drosophila genome. Cell 148, 458–472 (2012). [PubMed: 22265598]
20. Le TBK, Imakaev MV, Mirny LA & Laub MT High-resolution mapping of the spatial organization of a bacterial chromosome. Science 342, 731–734 (2013). [PubMed: 24158908]
21. Duan Z et al. A three-dimensional model of the yeast genome. Nature 465, 363–367 (2010). [PubMed: 20436457]
22. Rao SSP et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell 159, 1665–1680 (2014). [PubMed: 25497547]
23. Burton JN et al. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. Nat. Biotechnol 31, 1119–1125 (2013). [PubMed: 24185095]
24. Dudchenko O et al. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. Science 356, 92–95 (2017). [PubMed: 28336562]
25. Marie-Nelly H et al. High-quality genome (re)assembly using chromosomal contact data. Nature Commun 5, 5695 (2014). [PubMed: 25517223]
26. Putnam NH et al. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. Genome Res. 26, 342–350 (2016). [PubMed: 26848124]
27. Marbouty M et al. Metagenomic chromosome conformation capture (meta3C) unveils the diversity of chromosome organization in microorganisms. eLife 3, 533–19 (2014).

28. Burton JN, Liachko I, Dunham MJ & Shendure J Species-level deconvolution of metagenome assemblies with Hi-C-based contact probability maps. G3 (Bethesda) 4, 1339–1346 (2014). [PubMed: 24855317]

29. Beitel CW et al. Strain- and plasmid-level deconvolution of a synthetic metagenome by sequencing proximity ligation products. PeerJ 2, e415–19 (2014). [PubMed: 24918035]

30. Marbouty M, Baudry L, Cournac A & Koszul R Scaffolding bacterial genomes and probing host-virus interactions in gut microbiome by proximity ligation (chromosome capture) assay. Sci Adv 3, e1602105 (2017). [PubMed: 28232956]

31. Press MO et al. Hi-C deconvolution of a human gut microbiome yields high-quality draft genomes and reveals plasmid-genome interactions. Preprint at https://www.biorxiv.org/content/10.1101/198713v1 (2017).

32. Stalder T, Press MO, Sullivan S, Liachko I & Top EM Linking the resistome and plasmidome to the microbiome. ISME J. 9, e00105–10 (2019).

33. Mukherjee S et al. Genomes OnLine Database (GOLD) v.6: data updates and feature enhancements. Nucleic Acids Res. 45, D446–D456 (2017). [PubMed: 27794040]

34. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P & Tyson GW CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 25, 1043–1055 (2015). [PubMed: 25977477]

35. Kang DD, Froula J, Egan R & Wang Z MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ 3, e1165 (2015). [PubMed: 26336640]

36. DeMaere MZ & Darling AE bin3C: exploiting Hi-C sequencing data to accurately resolve metagenome-assembled genomes. Genome Biol. 20, 46 (2019). [PubMed: 30808380]

37. Duchêne S et al. Genome-scale rates of evolutionary change in bacteria. Microb. Genom 2, e000094 (2016). [PubMed: 28348834]

38. Puigbò P, Lobkovsky AE, Kristensen DM, Wolf YI & Koonin EV Genomes in turmoil: quantification of genome dynamics in prokaryote supergenomes. BMC Biol. 12, 66 (2014). [PubMed: 25141959]

39. McDonald JH & Kreitman M Adaptive protein evolution at the Adh locus in Drosophila. Nature 351, 652–654 (1991). [PubMed: 1904993]

40. Bishara A et al. High-quality genome sequences of uncultured microbes by assembly of read clouds. Nat. Biotechnol 486, 207 (2018).

41. Kuleshov V et al. Synthetic long-read sequencing reveals intraspecies diversity in the human microbiome. Nat. Biotechnol 34, 64–69 (2016). [PubMed: 26655498]

42. Zhao S et al. Adaptive Evolution within Gut Microbiomes of Healthy People. Cell Host Microbe 25, 656–667.e8 (2019). [PubMed: 31028005]

43. Garud NR, Good BH, Hallatschek O & Pollard KS Evolutionary dynamics of bacteria in the gut microbiome within and across hosts. PLoS Biol. 17, e3000102 (2019). [PubMed: 30673701]

44. Joshi NA & Fass JN Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software] (2011). Available at https://github.com/najoshi/sickle

45. St. John J SeqPrep. Available at https://github.com/jstjohn/SeqPrep (2011).

46. Schmieder R & Edwards R Fast identification and removal of sequence contamination from genomic and metagenomic datasets. PLoS ONE 6, e17288 (2011). [PubMed: 21408061]

47. Li D et al. MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. Methods 102, 3–11 (2016). [PubMed: 27012178]

48. Li H & Durbin R Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26, 589–595 (2010). [PubMed: 20080505]

49. Kurtz S et al. Versatile and open software for comparing large genomes. Genome Biol. 5, R12 (2004). [PubMed: 14759262]

50. Zhu W, Lomsadze A & Borodovsky M Ab initio gene identification in metagenomic sequences. Nucleic Acids Res. 38, e132–e132 (2010). [PubMed: 20403810]

51. Buchfink B, Xie C & Huson DH Fast and sensitive protein alignment using DIAMOND. Nat. Methods 12, 59–60 (2015). [PubMed: 25402007]

52. Benson DA et al. GenBank. Nucleic Acids Res. 41, D36–42 (2013). [PubMed: 23193287]

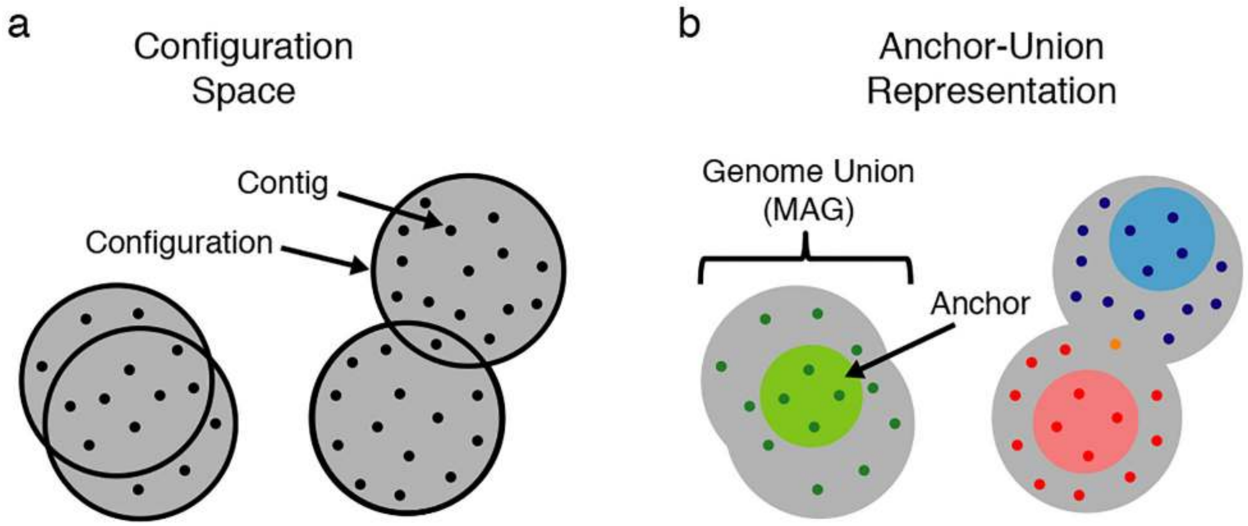


Fig. 1 |. Genomic configuration space and an anchor-union representation.

(a) Example with 4 configurations (large gray circles), each composed of contigs (black dots). Two related strains are represented by partially overlapping configurations. **(b)** Possible anchor-union pairs for the configurations in (a). There are 3 anchors (contigs within light-shade colored circles) and 3 matching genome unions (or MAGs), colored according to the anchor (dark shades). One contig is shared by two unions (colored orange), representing a shared element, such as a plasmid. The two conspecific strains are represented by a single anchor-union pair.

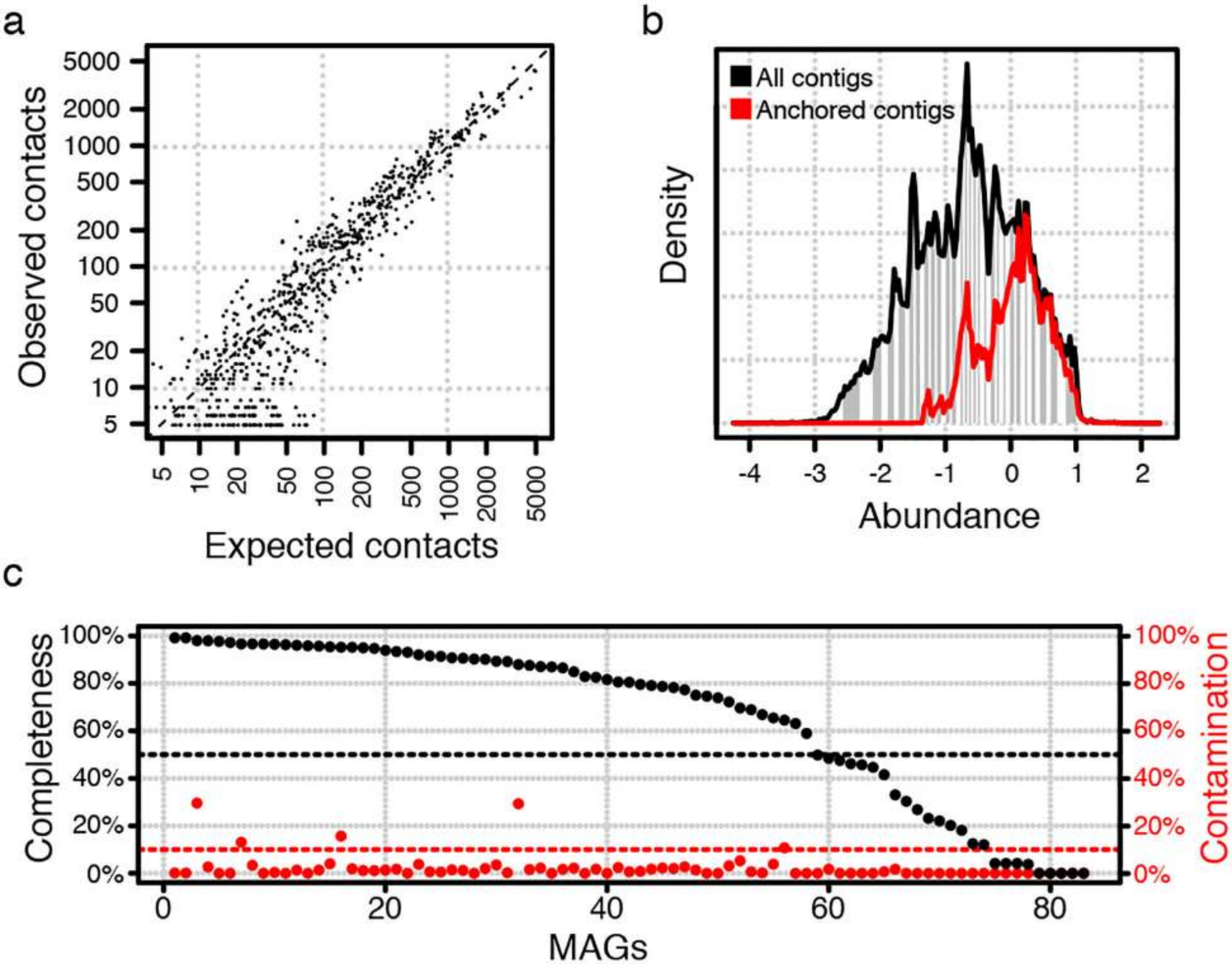


Fig. 2 |. Genotyping complex microbial communities using Hi-C.

(a) 83 anchor-union pairs were recovered for subject A. Shown is the expected number of inter-anchor spurious contacts (predicted by model, x-axis) vs. the observed number of inter-anchor contacts (y-axis). **(b)** A density plot of the relative abundance of all contigs from the metagenomic assembly (contigs >1k). The abundance (x-axis) is the enrichment of the contig read coverage over a uniform distribution of reads. The fraction of the assembly that was included in any recovered genome (‘anchored contigs’) is shown using a red line. White/gray stripes denote 10Mb bins. **(c)** Single-copy gene estimates of MAG completeness (in black) and contamination (in red), sorted according to completeness. Minimal completeness (50%) and maximal contamination (10%) thresholds depicted with dashed horizontal lines.

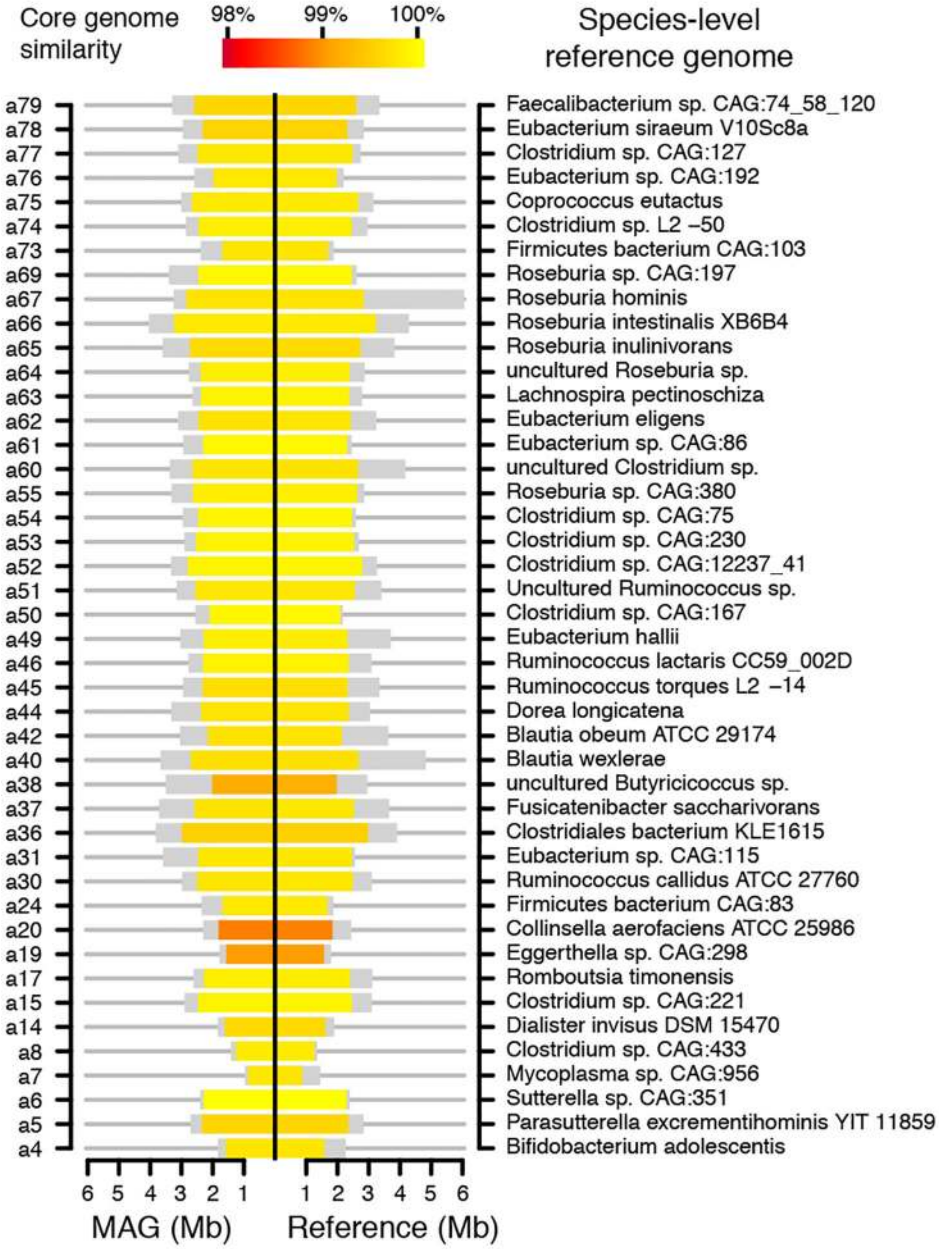


Fig. 3 |. Core and accessory divergence from species-level reference genomes.

Shown are the core and accessory fractions for the 44 MAGs that had a species-level reference. For both the recovered MAGs (left) and the matching species-level reference genomes (right), the core fraction is depicted using a colored rectangle, and the accessory fraction (i.e., strain-specific genes) is depicted using a gray rectangle. Cores are colored according to genome similarity (nucleotide sequence identity) between MAG cores and matching reference cores.

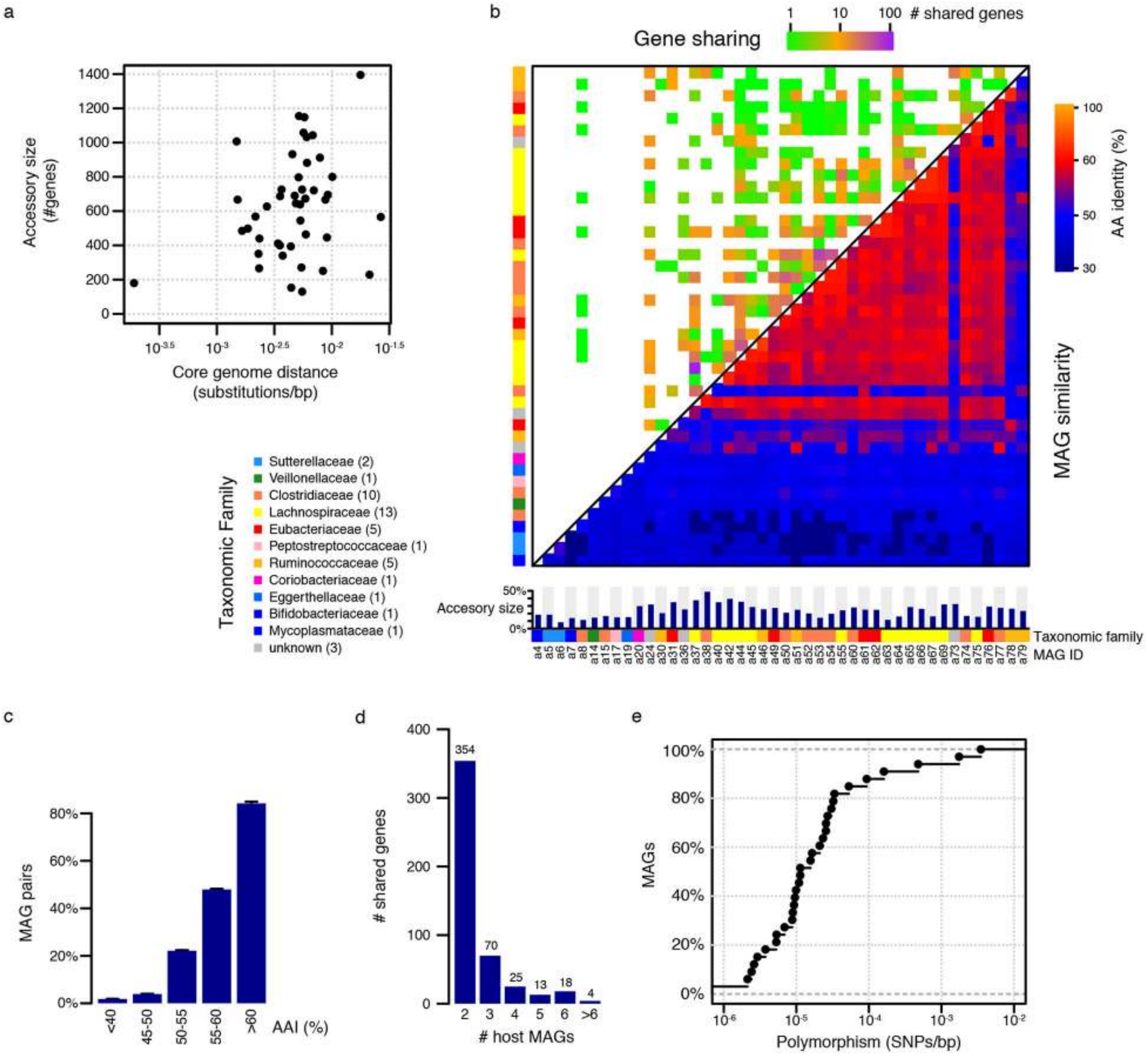


Fig. 4 |. Attributes of accessory genes.

(a) The substitution density within core genomes (x-axis) vs. the number of accessory genes (y-axis, genes that belonged to a recovered MAG and were missing in the matching reference genome), for all 44 MAGs that had a species-level reference. **(b)** Top left section of the matrix shows the number of shared genes and bottom right shows the mean amino acid identity (AAI). MAGs are sorted according to a hierarchical clustering based on AAI. Shown below the matrix is the size of the accessory fraction, and the family taxonomic assignment for each MAG (colored rectangles). The taxonomic family legend is shown with the number of MAGs written in parenthesis. **(c)** The percentage of pairs of MAGs that shared at least one gene, stratified by the sequence similarity (AAI) between the MAG pair. Standard deviations are depicted using error bars. **(d)** The number of shared genes, stratified according to the number of host MAGs with which they were associated. **(e)** The densities of intermediate SNPs (with allele frequency in the range 20–80%) within core genomes is plotted as an empirical distribution function, for 33 MAGs that had a read coverage of 10× or more.

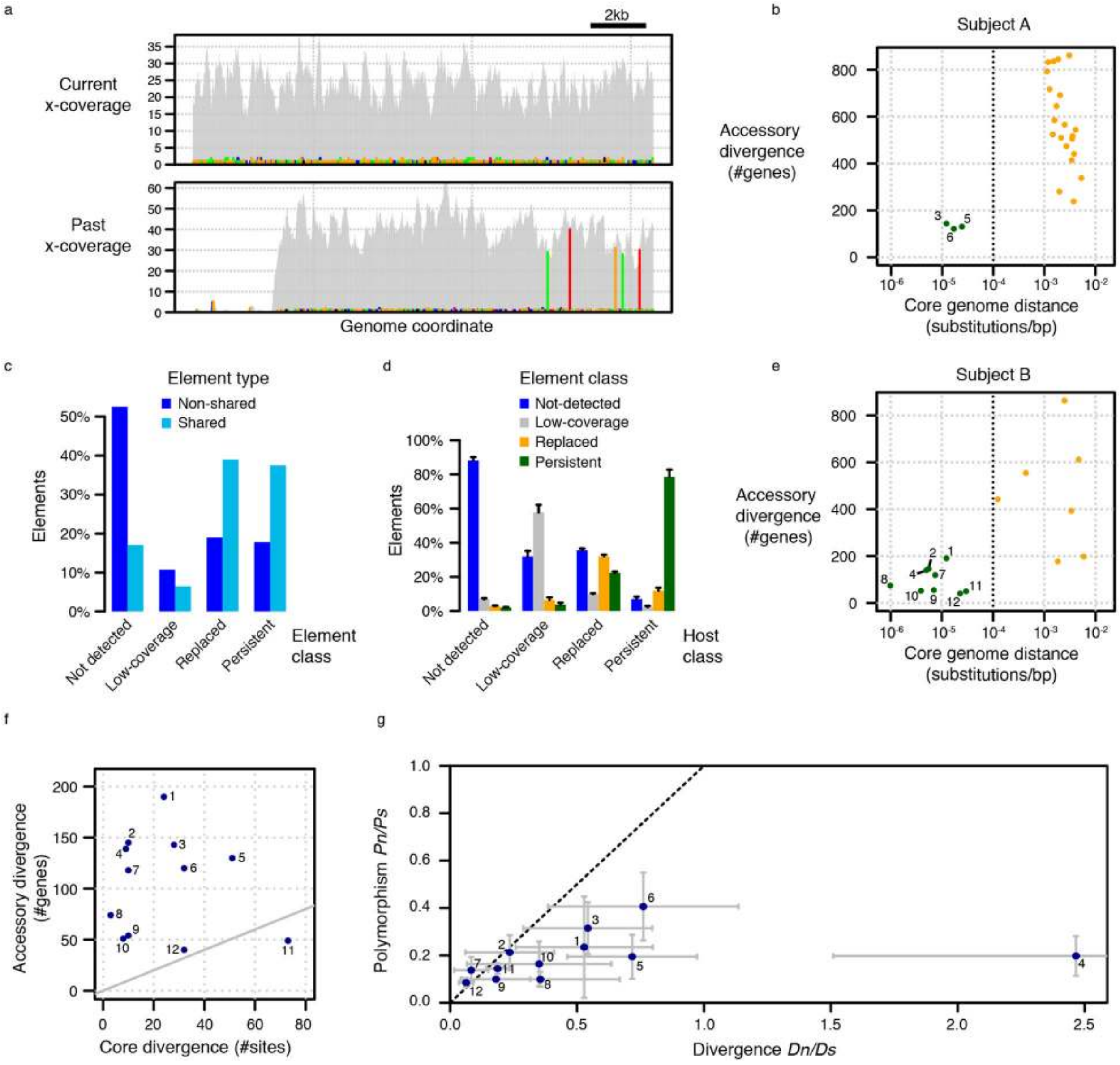


Fig. 5 |. 10-year community evolution.

(a) Genetic changes along a 15kb segment (x-axis). Shown for the genotyped sample (top) and the sample collected from the same subject 10 years prior (bottom), is the number of read supporting each SNP (y-axis). SNPs that agree with the assembly are colored gray, and deviating SNPs are colored by nucleotide (A/C/G/T are colored red/blue/green/orange). Note in the 10-year profile the region on the left that has low read coverage (reflecting gene-content change), and the 5 divergent SNPs on the right (reflecting nucleotide-level changes). **(b)** Shown for 24 MAGs that had >10× read coverage in the 10-year sample, is the core divergence (x-axis, substitutions/bp within cores) vs. the accessory divergence (y-axis, number of accessory genes classified as not-detected or replaced) over the 10-year period. MAGs are colored according to classification (persistent: green, replaced: orange), and the classification threshold (10^{-4}) is depicted with a dashed vertical line. Persistent MAG indices are numbered on the plot as in Extended Data Figure 10. **(c)** The distribution among element classes, stratified according to element type (shared and non-shared). Data are normalized so that each type sums to 100%. **(d)** The distribution among element classes, stratified according to host class. Data are normalized so that each host class sums to 100%. Standard deviations are depicted using error bars. **(e)** Same as panel b, for Subject B. **(f)** The number of core sites with substitutions (x-axis) vs. the number of accessory genes that were gained/lost (y-axis), for the 12 MAGs classified as persistent, with indices as in Extended Data Figure 10. **(g)** For the cores of the 12 persistent MAGs, shown is the ratio between the density of non-synonymous (*Dn*) and synonymous (*Ds*) divergent sites (x-axis), vs. the ratio between the density of non-synonymous (*Pn*) and synonymous (*Pn*) polymorphic sites (y-axis), with indices as in Extended Data Figure 10. Error bars (standard deviation) are plotted using gray whiskers.

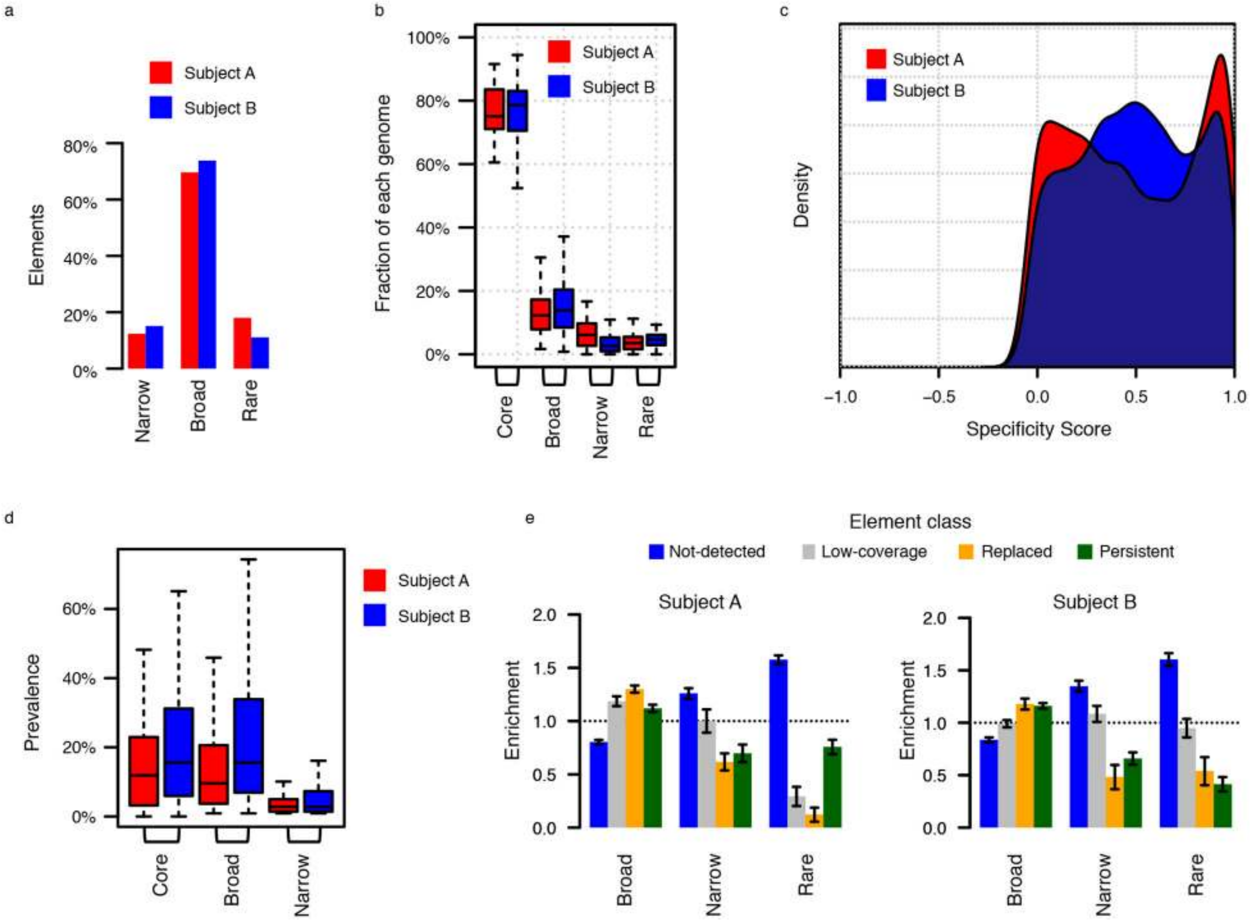


Fig. 6 |. Population based perspective on accessory genes for the two subjects.

(a) Elements were classified according to their distribution across 218 public gut metagenomic DNA libraries obtained from 218 individuals. The percentage of elements in each class for each of subjects A and B is shown. A ‘rare’ element was defined as an element detected in 0–2 individuals, and a ‘narrow-range’ element was defined as an element detected only in individuals in which one of its associated microbial hosts was also detected. All other elements were defined as ‘broad-range’. **(b)** A boxplot representation (horizontal line, median; box, 20%−75% percentiles; whiskers, 5%−95% percentiles; outliers removed) depicts the distribution across all 44 MAGs of the fraction of each genome (y-axis, percentage of genes out of the entire genome) that corresponds to core and to broad/narrow/rare accessory genes. **(c)** Population coverage vectors, spanning all 218 individuals, were computed for all accessory elements and core genes. Shown is the density plot of element specificity scores, defined as the Pearson coefficient between the vectors of broad-range elements and the vectors of their matching cores, colored by subjects. **(d)** A boxplot representation of the distribution of the prevalence of subject A and B core genes, broad-range elements and narrow-range elements. **(e)** The enrichment of all combinations of population-based element classifications and evolution-based element classifications, over a null-model that assumes both classifications are independent. Standard deviations are depicted using error bars.