RESOURCE
https://doi.org/10.1038/s41591-019-0559-3

nature medicine

A library of human gut bacterial isolates paired with longitudinal multiomics data enables mechanistic microbiome research

M. Poyet[1,2,3,8], M. Groussin[1,2,3,8], S. M. Gibbons[1,2,3,4,8], J. Avila-Pacheco[3], X. Jiang[1,2,3], S. M. Kearney[1,2,3], A. R. Perrotta[1,2], B. Berdy[1,2,3], S. Zhao[1,2], T. D. Lieberman[1,2,3], P. K. Swanson[1,5], M. Smith[5,6], S. Roesemann[3], J. E. Alexander[3], S. A. Rich[3], J. Livny[3], H. Vlamakis[3], C. Clish[3], K. Bullock[3], A. Deik[3], J. Scott[3], K. A. Pierce[3], R. J. Xavier[2,3,7]* and E. J. Alm[1,2,3,5,6]*

Our understanding of how the gut microbiome interacts with its human host has been restrained by limited access to longitudinal datasets to examine stability and dynamics, and by having only a few isolates to test mechanistic hypotheses. Here, we present the Broad Institute-OpenBiome Microbiome Library (BIO-ML), a comprehensive collection of 7,758 gut bacterial isolates paired with 3,632 genome sequences and longitudinal multi-omics data. We show that microbial species maintain stable population sizes within and across humans and that commonly used 'omics' survey methods are more reliable when using averages over multiple days of sampling. Variation of gut metabolites within people over time is associated with amino acid levels, and differences across people are associated with differences in bile acids. Finally, we show that genomic diversification can be used to infer eco-evolutionary dynamics and in vivo selection pressures for strains within individuals. The BIO-ML is a unique resource designed to enable hypothesis-driven microbiome research.

Engineering the gut microbiome to treat disease is an exciting new direction in medical science[1–3]. Fecal microbiota transplant (FMT) from a healthy donor into patients with recurrent *Clostridium difficile* infections is the first widely adopted microbiome-related therapy and has a ~90% success rate[4,5]. Investigational trials are underway in new disease areas, such as inflammatory bowel disease, liver disease, Parkinson's disease, severe acute malnutrition and infection by antibiotic-resistant pathogens[6–9] (see ongoing clinical trials at https://clinicaltrials.gov/). OpenBiome is a stool bank that has provided material for over 48,000 fecal transplants. Stool banks like OpenBiome represent an attractive opportunity for building a well-characterized culture collection because living biomass is preserved, allowing cultivation of isolated strains, and because dense longitudinal sampling (that is, several samples being collected per week) enables analysis of within-host dynamics. In addition, a resource of isolate genomes together with longitudinal dynamics can be useful in designing and analyzing future clinical trials. Finally, a comprehensive culture collection from successful donors could ultimately be used to replace FMT, which is a blunt tool for engineering the gut microbiome and may have long-term consequences due to the introduction of a wide variety of exogenous strains with unknown function[10–12].

While comprehensive strain collections are essential for mechanistic studies, culturing a diverse representation of gut bacteria has been challenging. Seminal work by several groups[13–17] has addressed many of the technological challenges of growing wide arrays of gut bacterial lineages, and two recent studies reported isolate and genome collections with broad phylogenetic representation[18,19]. However, existing isolate and genome collections are still limited, especially in strain-level diversity, for most of the bacterial species in the human gut. In addition, current collections are limited in examples of coexisting strain-level diversity from the same human host because the majority of strains were cultured from a large number of individuals or were targeted for maximizing phylogenetic diversity.

Recent work has shown that this within-host strain diversity is extensive in the human population[20] and within individual people[21–23]. New studies increasingly point to functional differences between strains of the same species that can impact human health[21,24,25]. For instance, strain-level differences can influence the metabolism of dietary compounds, such as galacto-oligosaccharides[26] or nondigestible fibers[27,28]. Bacteria-mediated metabolism of drugs can also differ across strains, influencing drug efficacy and toxicity[29,30]. In addition, genomic variation in virulence genes can alter pathogenicity among strains[31–33]. Finally, distinct strains can elicit different immune responses, such as cytokine production[25]. For these reasons, a large collection of isolates of multiple strains from many gut bacterial species, sampled both within and across people, is needed to better understand host–microbe interactions and to efficiently screen for candidate features that could ultimately be leveraged in rationally designed microbiome-based therapeutics.

Here, we introduce a comprehensive biobank of human gut bacteria: a library of 7,758 bacterial isolates obtained from healthy FMT donors recruited in the Boston area. This library covers most of the phylogenetic diversity found in the human gut, contains extensive

[1]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA. [2]Center for Microbiome Informatics and Therapeutics, Massachusetts Institute of Technology, Cambridge, MA, USA. [3]The Broad Institute of MIT and Harvard, Cambridge, MA, USA. [4]Institute for Systems Biology, Seattle, WA, USA. [5]Finch Therapeutics, Somerville, MA, USA. [6]OpenBiome, Somerville, MA, USA. [7]Gastrointestinal Unit and Center for Computational and Integrative Biology, Massachusetts General Hospital, Boston, MA, USA. [8]These authors contributed equally: M. Poyet, M. Groussin, S. M. Gibbons. *e-mail: xavier@molbio.mgh.harvard.edu; ejalm@mit.edu

strain diversity and is available to the research community. We report whole-genome sequences (WGSs) for 3,632 of these isolates that span a wide range of phyla and genera, to enable researchers to test and predict phenotypes in vitro and in vivo, such as metabolic capability or resistance to antibiotics. We also provide longitudinal 16S, metagenomic and metabolomic data for more than 80 FMT donors. Finally, we highlight examples that illustrate how these data can be used to better understand the eco-evolutionary dynamics of the gut microbiome within and between people.

Results

Isolation of an extensive collection of gut bacterial isolates for in vitro and in vivo testing of mechanistic hypotheses. Many strict anaerobes in the human gut were considered unculturable until recently[14–19,34,35]. As a result, densely sampled sets of strains from many anaerobic species are still not readily available. Here, we leverage recent advances in culturing techniques to isolate a large phylogenetic diversity of gut bacterial strains from healthy FMT donors.

Building a library of isolates that cover the diversity of gut bacteria from OpenBiome donors. We have designed and implemented protocols to culture, isolate and store a large diversity of anaerobic gut bacterial strains in pure culture. We used filtered stool extracts from 11 donors within our cohort, and we used 12 different media, combined with antibiotic, acid and ethanol treatments, resulting in 19 different culturing conditions (see Supplementary Methods). We used general media to obtain a wide phylogenetic diversity of bacterial species and selective media to grow specific clades of interest. This strategy allowed us to build a large and comprehensive open-access collection of human gut bacteria in pure culture (Fig. 1a and Extended Data Fig. 1). The BIO-ML currently contains 7,758 isolates belonging to the 6 dominant bacterial phyla in the human gut: Actinobacteria, Bacteroidetes, Firmicutes, Fusobacteria, Proteobacteria and Verrucomicrobia. We Sanger-sequenced the 16S rRNA gene to assign a taxonomy to each isolate. In total, 11 classes, 16 orders, 40 families and 133 genera are represented in our isolate library (Supplementary Table 1).

We next investigated whether the diversity of our cultured isolate collection overlaps with the diversity captured by culture-independent methods. We found that the BIO-ML comprehensively represents the in vivo bacterial genus-level diversity, weighted by abundance, found in the whole cohort of individuals (90 individuals; samples from only 11 were used to culture isolates) (Fig. 1c). In particular, we captured 99% of the diversity of Bacteroidetes genera, 96% of Firmicutes genera, 99% of Actinobacteria genera, 78% of Proteobacteria genera and 100% of Verrucomicrobia genera (represented by *Akkermansia*, the only Verrucomicrobia genus observed in the human gut; see 'The BIO-ML contains diverse taxa associated with human health' below for more details on this key genus). When looking at a range of operational taxonomic unit (OTU) resolutions, from 90% 16S similarity to amplicon sequence variants (ASVs, 100% 16S similarity OTUs), we confirmed that our library covers the diversity of high taxonomic ranks (Extended Data Fig. 2a). As expected, this coverage drops when considering more specific ranks, especially among Firmicutes, as each individual will tend to carry unique strains. Efforts to increase the taxonomic representation of missing Firmicutes strains are ongoing. Consistent with previous observations (Lau et al.[35] and Rettedal et al.[34]), we were able to isolate taxa that were present at very low average relative abundances across all donors (that is, <0.01%) or that were simply missed by 16S sequencing (Fig. 1d), such as strains from *Lactobacillus*, *Gardnerella*, *Clostridium* cluster XI or *Lactonifactor* genera. Overall, culture-based methodologies provide access to data that both overlap and complement sequencing surveys, enhancing our understanding of gut microbiome function and diversity.

Next, we asked whether relative abundance derived from culture-independent 16S data could provide meaningful information to guide the culturing and isolation of bacterial clades of interest. As selective media used to grow specific microbes were not available for the vast majority of gut bacteria, we tested this question using a non-selective culture medium (CGM medium, see Supplementary Methods). We compared the abundance of bacterial genera growing on CGM to their relative abundance in the 16S data. We observed no significant correlation between in vitro and culture-independent bacterial abundances across our four tested individuals ($P > 0.05$, Fig. 1d). In the absence of selective media, we caution that 16S relative abundance might not be a reliable predictor of which stool samples might yield bacterial species of interest.

We next tested whether the same bacteria are observed on the same medium across several individuals. We compared diversities with both a rich (CGM) medium and a selective culturing condition (Mmm + Ab4 media, see Methods), and we picked colonies randomly on plates with no morphological selection. The bacterial diversity captured varied extensively across individuals (Fig. 1e, $P < 0.001$) for both types of media, and differences in 16S relative abundances across individuals did not explain the variation in cultured diversity. This suggests that other factors, such as differences in dormancy states across individuals[36], might drive in vitro culturing outcomes.

The BIO-ML contains diverse taxa associated with human health. We isolated and sequenced strains from organisms that are strongly associated with human health. First, we cultured 159 of the 'Most Wanted' OTUs ($n = 485$) identified by the Human Microbiome Project (HMP) as both lacking cultured representatives and being associated with diseases[37] (Fig. 1b). We also biobanked bacteria that have been difficult to culture and isolate so far, such as *Akkermansia* and *Faecalibacterium*, and that have very few representatives in reference strain collections. *Akkermansia muciniphila* is a host mucin degrader[38], and has been associated with inflammatory bowel diseases and metabolic disorders[39,40]. We successfully isolated 132 different *Akkermansia* strains and sequenced the genomes of 45 strains of *A. muciniphila* and of 67 strains that, based on whole-genome information, belong to a previously unknown species within this genus (Extended Data Fig. 3). *Faecalibacterium prausnitzii* is a major butyrate producer[28] known to have anti-inflammatory effects[41]. The depletion of *F. prausnitzii* is correlated with Crohn's disease[41] and irritable bowel syndrome[42]. It is also the only characterized species within the *Faecalibacterium* genus. We cultured and isolated 75 *Faecalibacterium* strains. We sequenced the whole genome of 19 *F. prausnitzii* strains, as well as 4 additional strains that, based on whole-genome information, belong to unknown species in this genus (Extended Data Fig. 3).

Ecology and evolutionary dynamics inferred from isolate genomes. *Quality and diversity of BIO-ML isolate genomes.* To enable mechanistic studies with the BIO-ML isolates, we sequenced and assembled 3,632 bacterial genomes (Fig. 2a and Extended Data Fig. 2b). This genome collection consisted of 106 species clusters (defined by genomic similarity, see Methods) and 48 known genera across Actinobacteria, Bacteroidetes, Firmicutes, Proteobacteria and Verrucomicrobia (Extended Data Fig. 2b). We assigned a species taxonomy to 101 genome clusters. The five remaining clusters with unknown species affiliation were Firmicutes lineages that belong to the Ruminococcaceae and Peptostreptococcaceae families, and to the Clostridiales order. Among the 3,632 genomes, 1,337 genomes were from species that were longitudinally isolated from a single individual (individual am, Extended Data Fig. 1b). The overall quality of the genome assemblies was high: the median completeness level was 99.5%, the median positional coverage was 124×, the median scaffold *N50* (the minimum contig length needed to cover

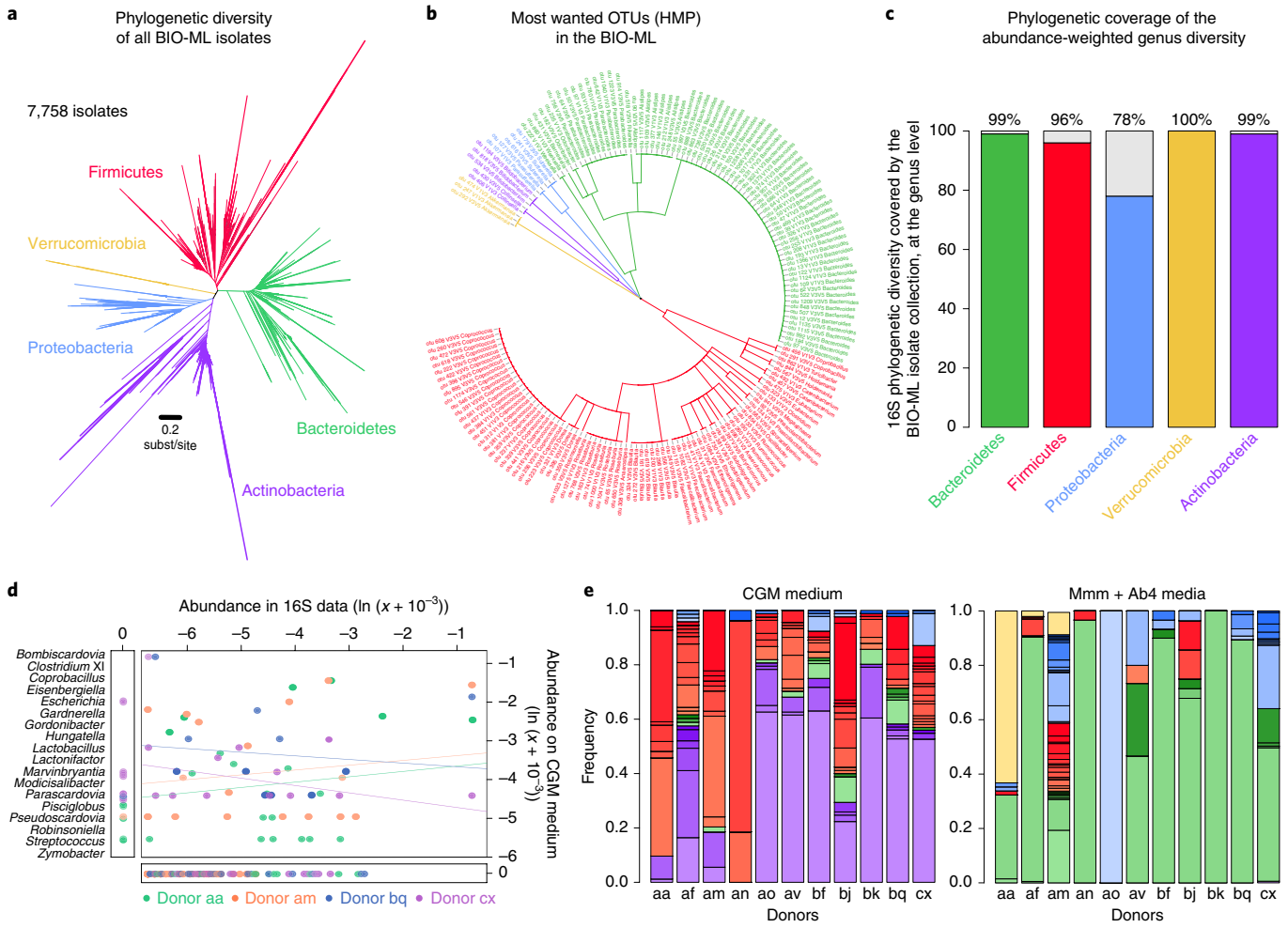


Fig. 1 | The BIO-ML library of human gut bacterial isolates. a, 16S phylogenetic tree of the 7,758 BIO-ML isolates. Lineages are colored by phylum. **b**, Cladogram showing the genus name and OTU ID of the Most Wanted OTUs identified by the HMP that have isolate representatives in the BIO-ML. **c**, Abundance-weighted taxonomic coverage of the library of bacterial isolates, compared with the diversity observed through culture-independent 16S amplicon sequencing. The library of isolates was built using 11 donors. The phylogenetic diversity of isolates was measured using 16S sanger sequencing, and this was compared with the total diversity observed in the 16S sequence data obtained from 1,168 samples from 90 individual donors of the BIO-ML. Taxonomic coverage was evaluated both at the genus levels (shown in **c**) and 97% OTU levels (Extended Data Fig. 2). Percentages and darker shades indicate diversity within each phylum captured by culture-dependent isolation methods. **d**, Culturing can sometimes capture bacterial taxa that are missed by culture-free methods. The relative abundance of bacterial 16S sequence variants of bacteria isolated on the general CGM medium was compared with culture-free 16S abundances. Relative abundances are on a log scale, and a pseudocount of 10^{-3} was added to represent sequence variants with null abundances, either on the CGM medium or in the culture-free 16S data. Each dot represents a bacterial genus. Dots below the plots show genera that were not obtained on CGM but were observed with culture-free sequencing. Dots on the left of plots show genera isolated on the CGM medium that were not seen in the culture-free sequencing data. For each individual, and on this general medium, the correlation between abundances is nonsignificant. **e**, The genus diversity captured by culturing approaches is inconsistent across individuals (Linear mixed-effects model, $P < 0.001$), with both a general and a selective medium. Each cell represents a genus, which is colored by phylum as in **a**.

50% of the genome assembly) was 155,045 bp and the median estimated contamination was negligible (0.3% by CheckM analysis) (Fig. 2b–e). We next compared the genetic diversity of BIO-ML genomes to other isolate genome collections: National Center for Biotechnology Information (NCBI; comprising 79,226 human gut and non-human-associated genomes), HMP[13] (2,265 human-associated genomes, BioProject PRJNA28331), Cancer Genomics Research[18] (CGR; 1,520 human gut isolate genomes) and Human Gastrointestinal Bacteria Culture Collection[19] (HBC; 736 human gut isolate genomes). Of our genomes, 80–96% were closely related to at least one reference genome (measured by the Mash distance (≤0.05)), depending on the considered reference genome collection (Fig. 2g). This was expected, as both previous genome collections and BIO-ML genomes were sampled from industrialized populations. As such, the BIO-ML collection greatly increases the strain-level diversity in known species of human gut bacteria.

Nonetheless, for 17–39% of our genomes, protein similarity to their closest reference genome was lower than 95% (Fig. 2f), and 4–20% were part of species that have no representatives in the HMP, CGR and HBC collections (Fig. 2g). Finally, we evaluated diversity in gene content, focusing on two *Bifidobacterium* species: *B. adolescentis* and *B. longum*. We showed that strains within these two species have extensive variation in gene content, and that they greatly increased the diversity of gene repertoires as compared to reference *Bifidobacterium* species (Fig. 2h,i). Overall, our cross-sectional and longitudinal genome collection provides the necessary phylogenetic resolution to investigate long- and short-term genomic evolution at the levels of gene content and single-nucleotide polymorphisms

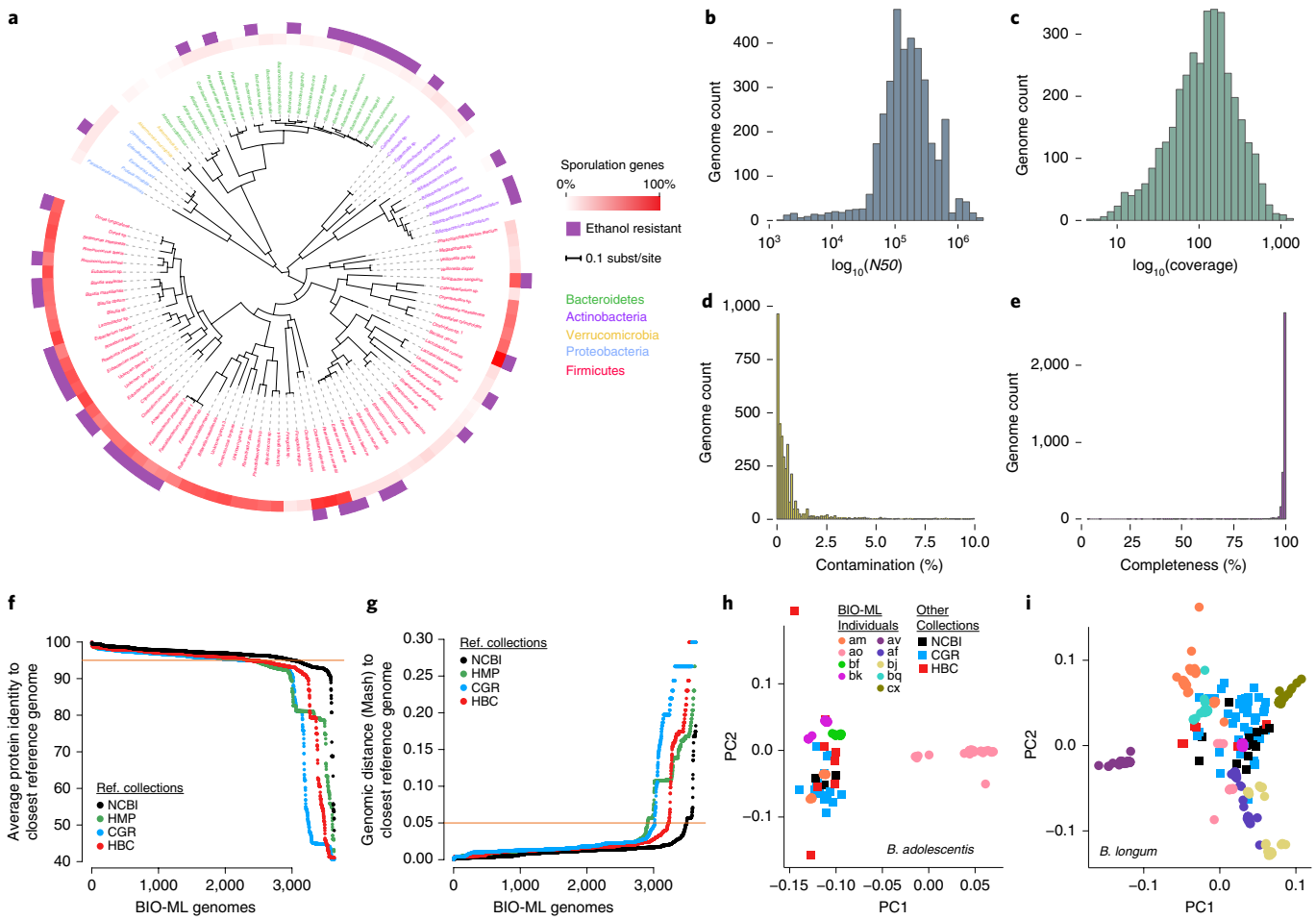


Fig. 2 | The BIO-ML of isolate genomes is large and diverse. a, Phylogenomic relationships of 106 bacterial species present in the isolate genome library. A single representative genome per species was selected out of the 3,632 genomes to reconstruct the multiple sequence alignment of ribosomal proteins. Bacterial species are colored by phylum. The inner circle represents the fraction of essential sporulation genes found in each species. Spores are dormant cellular states that allow bacteria to withstand environmental stress. The outer circle shows which species have representatives in our isolate library that were isolated after ethanol selection, which is commonly used to enrich for ethanol-resistant spores. **b**, Distribution of genome *N50s*, at $\log_{10}$ scale**. c**, Distribution of genome coverage values on a $\log_{10}$ scale. **d**, Distribution of genome contamination values. **e**, Distribution of genome completeness values. **f**, Average protein identity to the closest NCBI, HMP, CGR and HBC reference isolate genome collections. The horizontal line shows 95% protein identity. **g**, Genomic distance to the closest NCBI, HMP, CGR and HBC reference isolate genome collections, measured by Mash. The horizontal line shows a Mash distance (a genomic distance calculated by the Mash software) of 0.05, which is a threshold used to delineate species. **h**,**i**, Diversity of gene contents in BIO-ML *B. adolescentis* (**h**) and *B. longum* (**i**) strains, compared to reference NCBI, HMP, CGR and HBC genomes (squares).

within gut species (see 'Extensive sampling of isolate genomes reveals the long- and short-term evolution of gut commensal bacteria' below for such investigations in two *Bifidobacterium* species).

Resistance to ethanol is more widespread than previously thought and not restricted to spore-formers. In order to enrich for endospores when culturing our isolates, we treated samples with an equivalent volume of ethanol for 1 hour at room temperature, as described previously[16,36]. We show that, while ethanol treatment tends to enrich for organisms that have a set of shared endospore-forming genes[43], many organisms that do not possess genes involved in spore formation can be recovered by this method (Fig. 2a), suggesting that such organisms may possess cell walls that limit the diffusion of ethanol into the cell (in the phylum Actinobacteria or among the non-spore forming Firmicutes). Regardless, both endospores and other ethanol-resistant cell states appear frequently in the human fecal microbiota, suggesting that non-endospore environmental resistance and dormancy have a previously underappreciated role in this ecosystem[36].

Extensive sampling of isolate genomes reveals the long- and short-term evolution of gut commensal bacteria. The extensive gene content variation in *B. adolescentis* and *B. longum* prompted us to investigate the evolutionary dynamics of their gene repertoires within individuals. We observed that for both *Bifidobacterium* species, similarity in gene content did not necessarily match the phylogenetic history of the major lineages that had colonized each host (Fig. 3a,b), confirming that gene repertoires are plastic over evolutionary time[44]. However, it is unknown whether gene content can change within people after bacterial colonization. We observed that each individual carried a unique micro-diversity comprising very closely related strains. Even within these nearly identical descendants of a single ancestral cell, the diversification history (that is, the phylogeny) of these strains did not exactly match their similarity in gene content (Fig. 3a,b) suggesting multiple gene-gain and gene-loss events (Fig. 3c–f). As an illustration, this rapid turnover in gene repertoires can be observed in donor bk, with two different clades of *B. adolescentis* strains that experienced a convergent loss of a 50-kb gene cluster (Fig. 3c–f).

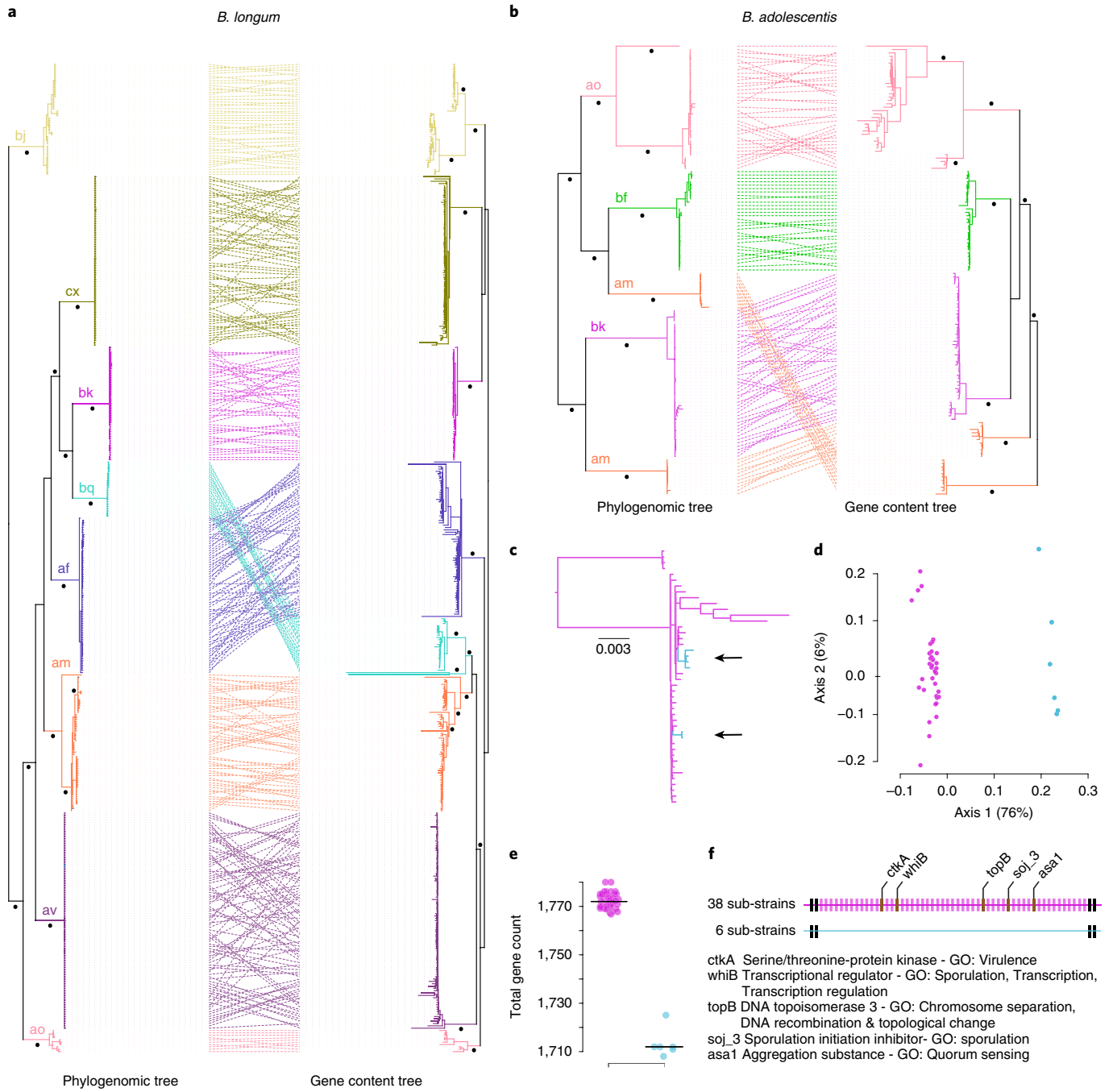


Fig. 3 | Rapid genomic evolution of gut commensal bacteria within people. a, Evolution of gene contents in *B. longum*. The tree on the left depicts phylogenetic relationships of 426 *B. longum* genomes sampled across 8 individuals. The tree on the right is a distance tree of gene contents. Dots represent branches with Bootstrap support ≥80. Dashed lines connect the same genomes in each tree. For the deepest (black) branches, similarity in gene contents does not recapitulate the phylogenetic history of lineages, indicating ancient and extensive gene turnovers. While all strains within an individual cluster by gene content, post-colonization gene turnovers that are not correlated to strain phylogeny can be observed. **b**, As in **a**, 248 *B. adolescentis* genomes sampled across 4 individuals show extensive gene content turnovers that occurred both generations ago and during individuals' lifetimes. **c**, Phylogenomic tree of the 44 strains colonizing individual bk. The tree reveals within-host diversification following bacterial colonization. Arrows show unrelated genomes that have similarly different gene contents compared with other genomes. All trees and genes in **c**–**f** are for individual 'bk', shown in pink. The blue represents two specific clades in bk that experienced the loss of genes shown in **f**. **d**, Multivariate analysis of gene contents in bk strains reveals rapid and convergent within-host dynamics of gene contents. The *x* axis explains 76% of the variance in gene content. Six strains, which group into two monophyletic clades (see **b**), have outlier gene contents (see arrows in **c**). **e**, The 6 substrains have independently lost about 60 genes within individual bk. **f**, Difference in gene content between strains is mostly explained by the loss of 53 genes that cluster into a 50-kb genomic region. Most of these genes have unknown functions. GO, gene ontology.

Thus, the genomic content, and presumably the functional capabilities, of strains can change during the lifespan of individuals, possibly in response to host-specific environmental factors or microbe–microbe interactions.

We next asked whether multiple distantly related strains of a given species that co-colonize the same host have gene contents that are more similar than expected by phylogeny, suggesting the occurrence of niche filtering by the host environment. We observed that

multiple distantly related strains of *B. adolescentis* had colonized individual am (Fig. 3b) and that these strains harbored remarkably similar gene content. This convergence in gene content suggests that these two distantly related strains stably thrived within similar niches. However, whether this convergence occurred within individual am due to adaptation via extreme gene loss or gene gain rates after colonization, or whether host niche filtering promotes the colonization of strains with similar pre-established functions, is unknown.

High-resolution genomic time series from FMT donors. To guide future in vitro and in vivo studies leveraging the library of isolates, we generated culture-independent cross-sectional and longitudinal sequencing and metabolomic data from a cohort of 90 FMT donors, including the donors used for culturing isolates. We provide longitudinal 16S data from 1,168 samples, producing 10 dense long-term time series (up to 1 sample every other day during 18 months; see Extended Data Fig. 1c). We generated longitudinal shotgun metagenomic data from 563 samples collected from 84 donors, producing 4 dense long-term time series (up to 1 sample every other day during 18 months; see Extended Data Fig. 1d). Finally, we conducted metabolite profiling on 179 stool samples from 83 donors that overlap with the 16S and metagenomic data, including several metabolomic time series (Extended Data Fig. 1e).

Time-series data improve abundance estimations and ecological inferences from metagenomic and 16S data. Averaging multiple timepoints may be optimal for precisely quantifying abundances of bacterial taxa and functions within individuals. However, there has not been a quantitative assessment of how much improvement is possible, or of how many samples are needed. Using our longitudinal dataset, we found that each person harbored a stable and unique microbiome structure, both in terms of taxa and broad functional categories (permutational multivariate analysis of variance (PERMANOVA), $P<0.0001$; Extended Data Figs. 4a and 5a). However, we found that the relative abundance of a given ASV (equivalent to 100% OTUs) and of a given clusters of orthologous groups (COG) category fluctuated substantially from day-to-day, but the median relative abundance remained relatively constant (Fig. 4a,d). We could predict the variance in our estimate of an ASV and COG median relative abundance for a given sample size by randomly subsampling the time series at different levels of temporal resolution (Fig. 4b,e). Overall, we found that the variance in our estimate was greatly reduced by collecting between five and nine timepoints (Fig. 4c,f). Collecting more than nine timepoints had a diminishing return for improving accuracy in the median abundance estimate (Fig. 4b,c,e,f). Consequently, to optimally estimate the abundance of a given BIO-ML isolate, we recommend calculating a median abundance by mapping isolate 16S or genomes to culturing-independent data on at least five longitudinal samples.

We next tested whether the increased accuracy in estimating abundance from averaging time points could help to identify species–species correlations. We generated a cross-sectional correlation matrix based on the median abundances of ASVs for the ten FMT donors with long, dense time series (Extended Data Fig. 6a). We identified all significant correlations between log-transformed median ASV relative abundances (Extended Data Fig. 6b) that were estimated from the full time series. We then recalculated the cross-sectional correlation matrix using differently sized subsets of each time series, by randomly drawing time points. We found that, when only collecting a single sample from each donor, we failed to identify ~60% of the significant edges that were found in the full network (Extended Data Fig. 6c). As the number of subsamples increased, the networks began to capture more of the edges from the full time series (Extended Data Fig. 6c). Thus, many taxon–taxon correlations can be missed if their abundance is only calculated from a single snapshot sample, rather than from a median abundance estimated from multiple timepoints.

In addition to identifying cross-sectional correlations, averaging across timepoints also revealed highly conserved relative abundances of bacterial taxa and functions across donors (Extended Data Figs. 4b and 5b). For every donor pair, there was a significant positive correlation between the log-median relative abundances of ASVs and COGs across different donors (Pearson's correlation test, $P<0.05$; Extended Data Figs. 4b and 5b). These correlations were weaker for single, randomly drawn time points (Extended Data Figs. 4b and 5b).

Bacterial genomic diversification within individuals and life-history traits are associated with ecological stability and disturbance of the gut ecosystem. Characterizing evolutionary and ecological dynamics in the microbiome has been limited by a dearth of longitudinal datasets. We found long-term ecological stability at the species level (Fig. 4), but this apparent stability might not reflect temporal dynamics at the strain level.

We jointly analyzed metagenomic and whole-genome time series from the same donor to characterize fine-grained within-host genomic diversification and genotype dynamics across three species. We focused on two abundant non-spore-forming species in individual am: *Bacteroides vulgatus* and *Bacteroides ovatus*. We also analyzed the dynamics of a spore-forming species, *Turicibacter sanguinis*, which is present at much lower abundance in the gut.

We observed that individual am was colonized by two distantly related *B. vulgatus* strains (Fig. 5a), suggesting that two independent colonization events had occurred and were followed by stable engraftment and very little diversification. Mapping of the metagenomic time-series data onto these genomes showed that these two primary strains stably coexisted within individual am over the sampling period (Fig. 5b,c). This stable coexistence of strains of the same species may indicate fine-scale niche partitioning in donor am's gut. *B. ovatus* also showed stable engraftment and post-colonization diversification within donor am. The clustering of *B. ovatus* strains into a single clade (Fig. 5d) and the number of SNPs observed among genotypes are consistent with a single colonization event. Following colonization, within-host genomic diversification occurred (Fig. 5e), which was not observed for *B. vulgatus* in the same individual. Three main *B. ovatus* substrains could be phylogenetically defined, and their abundances were tracked over time. The three substrains showed nonstationary dynamics, with strain 3 increasing in abundance relative to the 2 ancestral strains, from 2–5% shortly after the beginning of the sampling period to 60% by day 520 (Fig. 5f).

Finally, we show that donor am was serially colonized by multiple distantly related *T. sanguinis* strains (Fig. 5g), which rapidly displaced one another over the course of the sampling period. All sampled *T. sanguinis* strains clustered by culturing time points (Fig. 5h), and their data suggested that there were three independent colonization events, followed by full strain replacement (Fig. 5i). These strain turnovers may be the result of spore blooms from a pre-existing cocktail of distantly related strains. They could also result from successive colonization events followed by strain displacement. *T. sanguinis* was not abundant enough in the gut for accurate detection in the metagenome, and we were unable to track the abundance of strain genotypes at a high temporal resolution. Thus, we cannot completely rule out the possibility that alternative strains were present at lower abundance at each time point and were not captured by culturing. However, the extensive strain sampling (78 isolates) at the intermediate time point day 404 did not yield isolates closely related to strains 1 and 3, which supports the hypothesis of serial colonization events followed by strain replacement. At the intermediate time point (that is, day 404), some SNP diversity was observed (Fig. 5h), suggesting that *T. sanguinis* can rapidly accumulate

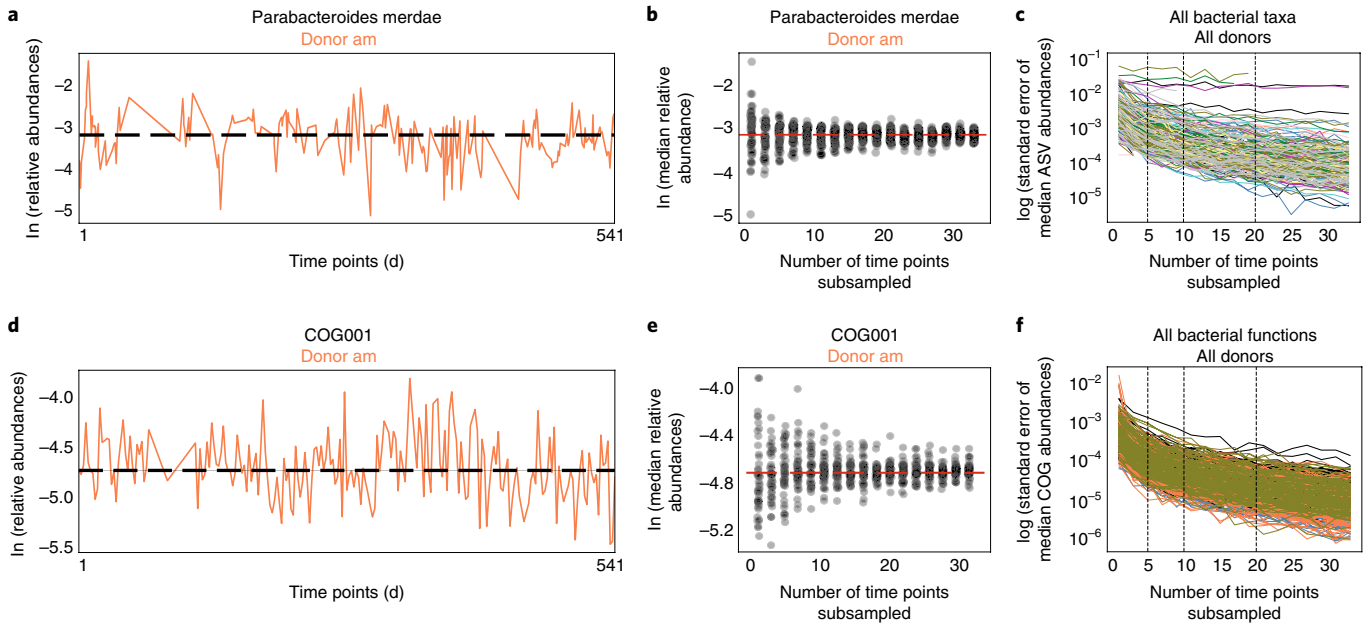


Fig. 4 | Densely sampled longitudinal data greatly improve ecological inferences. Dense longitudinal data are necessary to obtain accurate estimates of population size for both species and functions. **a** and **b** show examples for a single ASV, and **d** and **e** for a COG. **c** and **f** show that results in **a** and **b** and **d** and **e** replicate across all 100 of the most abundant ASVs and COGs. **a**, Longitudinal relative abundance (*y* axis) of a single ASV in donor am, annotated as *Parabacteroides merdae*. The abundance fluctuates over time (*x* axis), but continually returns to a conserved median abundance (dashed line). **b**, Estimation of median ASV relative abundance (*y* axis) depending on the number of time series samples used to calculate the median (*x* axis) (number of iterations = 50). The *P. merdae* ASV presented in **a** is shown as an example. The accuracy of this estimate is lower when considering only two samples, and improves substantially after collecting five samples. Gains in accuracy saturate at around nine samples. The red lines in **b** and **e** represent the median calculated across all time points. **c**, Estimates of median ASV abundances across the 100 most abundant ASVs become more accurate as more time series samples are collected. The elbow of the curve, where gains in accuracy begin to diminish, occurs at roughly five within-person samples. **d**, Longitudinal relative abundance (*y* axis) of COG001 in the donor am. The abundance fluctuates over time (*x* axis), but continually returns to a conserved median abundance, similar to what we see for ASVs (**a**–**c**). **e**, Estimation of median COG001 relative abundance (*y* axis) depending on the number of time series samples used to calculate the median (*x* axis) (number of iterations = 50). The accuracy of this estimate is lower when considering only two samples, and improves substantially after collecting only seven samples. Gains in accuracy saturate at around 11 samples. **f**, Estimates of median COG abundances across the 100 most abundant COGs become more accurate as more time series samples are collected. The elbow of the curve, where gains in accuracy begin to diminish, occurs at roughly seven within-person samples.

mutations following a colonization event that probably happened between days 168 and 404. Overall, our results support previous culture-independent reports indicating that spore-forming gut bacteria are more likely to turnover within a person and jump between hosts[36]. This strain-level analysis demonstrates that cross-host dissemination can be rapid and can occur multiple times within the span of several months, which influences the ecological stability of the gut microbiome on clinically relevant timescales.

Donor fecal metabolomes can be distinguished by their bile-acid profiles, while within-donor variation is driven largely by amino acids. We measured a total of 47,930 metabolomic features: 21,224 features in 7,021 non-redundant clusters, 26,706 unclustered features (no fragments or adducts detected) and 489 annotated compounds.

Unsupervised clustering of metabolomic data discriminates both time points and subjects (Extended Data Fig. 7). We focused our analyses on donors for which metabolomics data had been generated for more than six time points. The combination of principal components (PC) 1 and 2 clearly showed between-donor and between-time-point variation (Fig. 6a and Extended Data Fig. 7). We defined metabolites as varying across donors or across time points by their alignment in PC space: metabolites that aligned parallel with within-donor variance (Fig. 6b, red vectors) were associated with temporal variation, and metabolites perpendicular to these vectors were associated with cross-sectional variation (Fig. 6b, black vectors). Compounds contributing to cross-donor differences include saturated dicarboxylic acids, such as suberic, sebacic and azelaic acid, and polyunsaturated fatty, acids such as adrenic (C22:4), arachidonic (C20:4), eicosatrienoic (C20:3), docosahexahenoic (C22:6) and docosapentaenoic acid (C22:5). Likewise, conjugated and unconjugated primary bile acids (tauro- and glycocholate, tauro- and glycochenodeoxycholate), metanephrine, urobilin and GABA had donor-specific signatures (Fig. 6c). The significant clustering of annotated metabolite profiles by donor (PERMANOVA, $P < 0.0001$) supports prior work showing that the gut microbiome is unique to each person and relatively stable over time[45]. The metabolites associated with the temporal variation included several amino acids, such as serine, lysine, glutamine, tyrosine, and citrulline, as well as vitamins, such as nicotinate and pantothenate, and a few cholesteryl esters. These shifts in amino acids may be due to diet[46], inflammation[47] or cellular damage in the colon[48]. Despite the pronounced changes in their abundance in the stool of subjects through time, these metabolites are tightly correlated within subjects (Fig. 6d). The coupling of the dynamics of these various metabolites suggests that they are generated by a common, and as yet unknown, phenomenon in the gut. Individual bacterial taxa were correlated with certain dietary metabolites (for example carnitine, associated with red-meat consumption), bile acids (for example taurocholate, associated with spore germination) and a variety of lipids, which suggests that these factors are important for defining bacterial niches in the gut (Extended Data Fig. 8).

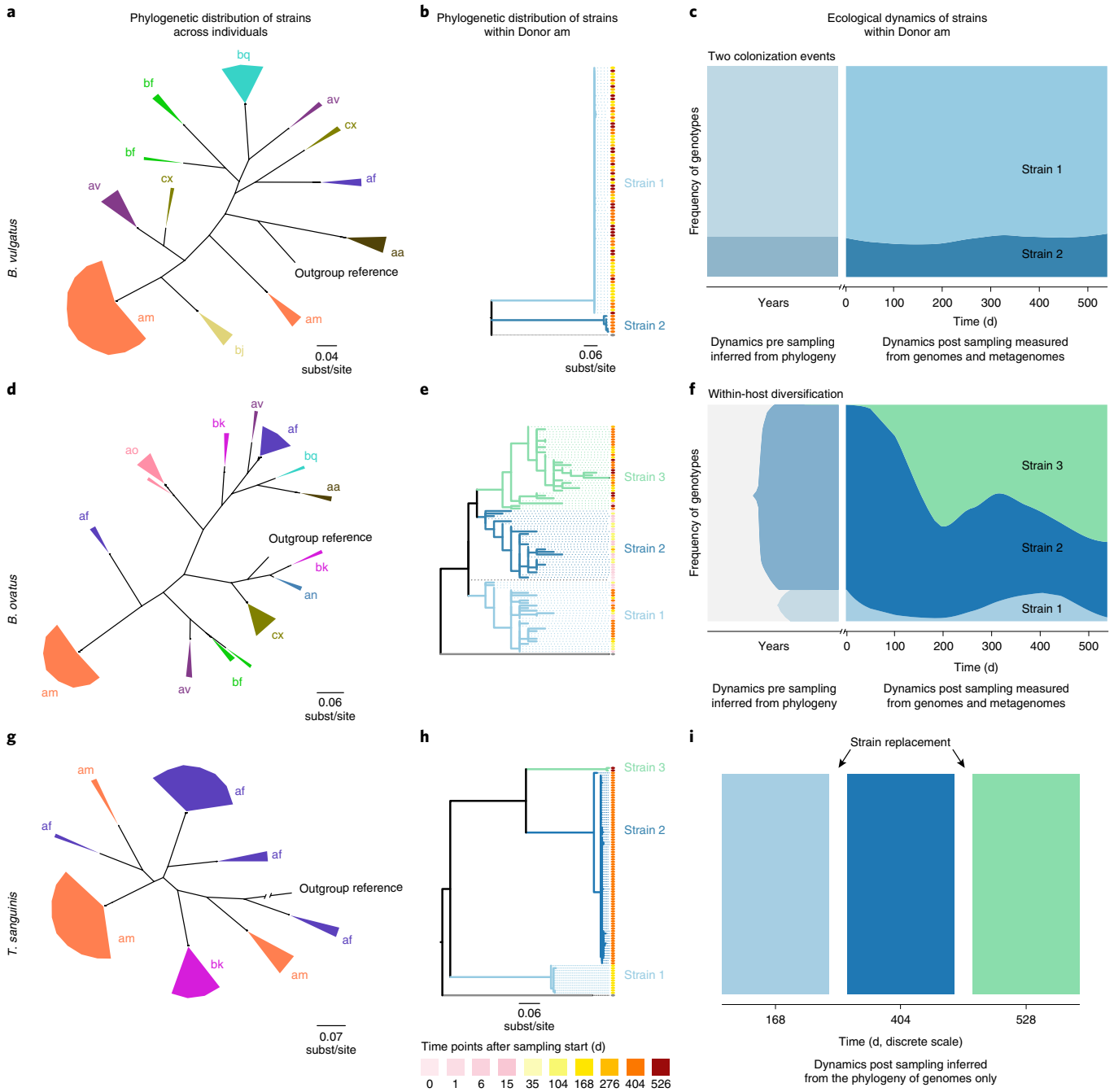


Fig. 5 | Eco-evolutionary dynamics of human gut bacterial strains and impact on community stability. Strain-level abundances were tracked over time in three species, showed by rows: *B. vulgatus* (**a**, **b** and **c**), *B. ovatus* (**d**, **e** and **f**) and *T. sanguinis* (**g**, **h** and **i**). The phylogenetic distribution of strains across individuals is shown in panels **a**, **d** and **g**, each clade being colored by individual. The size of clades is proportional to the number of strains. The phylogenetic relationships of strains colonizing individual am is shown in panels **b, e** and **h**, and were reconstructed from the alignment of SNPs that differentiate am strains between each other. Trees are rooted with the outgroup reference genome used to call SNPs. Sampling times (**d**) are color-coded and are represented in front of each isolate. In *B. vulgatus*, strains do not cluster by sampling date. In *B. ovatus*, isolates from the 'Strain 2' clade are mostly sampled in the beginning of the sampling period. 'Strain 1' is composed of isolates sampled at intermediary time points, while 'Strain 3' contains isolates sampled at the latest time points. In *T. sanguinis*, isolates perfectly cluster by sampling dates. Ecological dynamics of the main strain lineages within donor am is represented in panels **c**, **f** and **i**. For *B. vulgatus* and *B. ovatus* (**c** and **f**), metagenomes were mapped onto SNPs differentiating the main strain lineages to track abundance over time. For *T. sanguinis*, metagenomes were not used because this species is at too low abundance in individual am to obtain reliable estimates of abundance. Variations in abundance were inferred from the phylogenetic tree of isolate genomes (**h**) only. Gray areas in **c** and **f** represent precolonization abundance dynamics inferred from the phylogeny and the distribution of strains across sampling times.

Discussion

Cross-sectional and longitudinal surveys of the human gut microbiome have generated hypotheses of how bacteria influence our health. The next phase in microbiome research requires that we begin to test these hypotheses directly with isolates. Here, we describe a biobank of human gut bacteria, and a corresponding genomic dataset

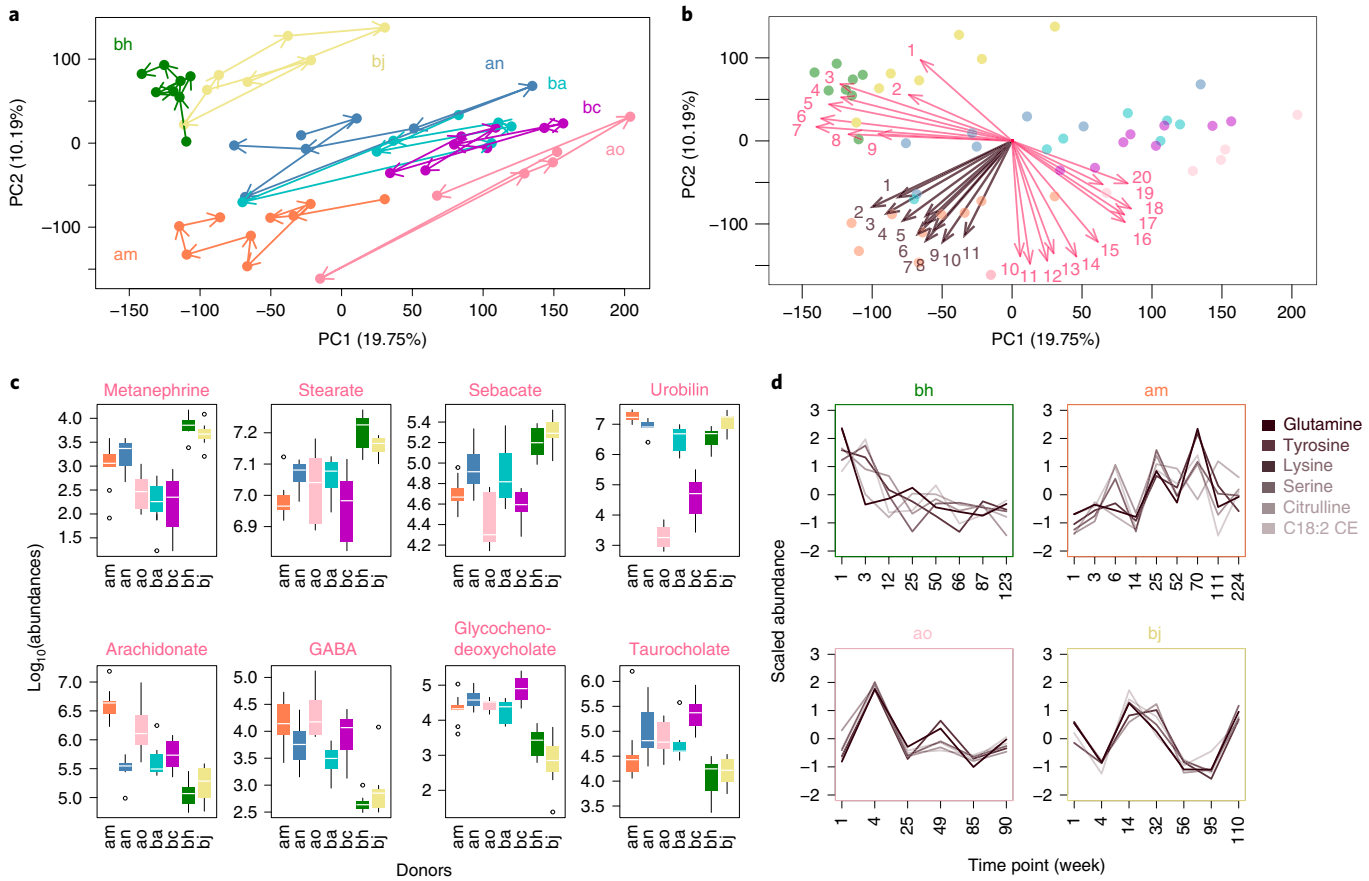


Fig. 6 | Gut metabolome profiles are highly specific to individual people, and this is mostly driven by differences in bile-acid concentrations. **a**, Despite this person-specific signature (different colors), metabolomes also vary within people over time due to fluctuations in amino acid concentrations. **b**, Principal component (PC) plot of metabolomic time series for seven donors in ordination space (proximity between points indicates similarity in metabolic profiles). Donors appear to vary along two axes. Within-donor variation follows a diagonal axis going from the bottom left of the plot to the top right (pink arrows: 1: stearate; 2: oleanate; 3: sebacate; 4: azelate; 5: metanephrine; 6: suberate; 7:adipate; 8:urobilin; 9:hydrocinnamate; 10: adrenate; 11: arachidonate; 12: eicosatrienoate; 13: DHA; 14: DPA; 15: GABA; 16: glycochenodeoxycholate; 17: glycocholate; 18: taurochenodeoxycholate; 19: xanthurenate; 20: taurocholate. Cross-donor variation follows a perpendicular axis going from the top left to the bottom right (brown arrows: 1: citrulline; 2: C18:2 CE; 3: pantothenate; 4: nicotinate; 5: tryptophan; 6: serine; 7: lysine; 8: histidine; 9: tyrosine; 10: threonine; 11: glutamine. **c**, Bile acids and saturated and unsaturated fatty acids are among the dominant metabolites defining differences between donors. **d**, Amino acid scaled (standard deviation scaled mean-centered) abundances co-vary with one another in some, but not all donor time series.

that greatly expands the existing collections of isolates currently available[13–19]. These isolates cover a large phylogenetic diversity (Fig. 1 and 2), and are available for research (see Methods).

Culture-based work can provide rich phenotypic information about gut bacteria, including nutritional preferences[49], drug metabolism[50] or host immune response[51–53]. For example, we found that many taxa that do not harbor sporulation genes were nonetheless able to survive ethanol treatment (a common technique for isolating spores; Fig. 2a). We also demonstrated how the genomes from closely related strains isolated from the same host can be used to track evolutionary dynamics. High-resolution multi-omic time-series data provide an additional layer of context to the BIO-ML gut bacterial isolates and genomes, enable detailed study of within-person strain dynamics, and signal averaging across timepoints for greater accuracy. Identifying within-person turnover in ecological-niche occupancy could be translated into personalized probiotic treatments, for example following antibiotics or gastrointestinal infections. The BIO-ML data are particularly relevant to ongoing clinical studies using OpenBiome donors, as they can be used to track engraftment of strains, and the genomes of those strains can be correlated with clinical outcomes.

In addition to the relatively simple analyses described here, we anticipate that the BIO-ML isolate collection will enable new and more powerful experimental designs. In particular, complex synthetic communities can be grown reproducibly in vitro using strains isolated from a single donor, and their dynamics can be compared with those of the same strains in vivo. Synthetic isolate communities can be designed on the basis of genomic information to efficiently perform a given function relevant for health, such as short-chain fatty acid production. The integration of previously underrepresented clades, such as *Turicibacter* and *Akkermansia*, into these experimental designs will enable new mechanistic studies on these key gut bacteria.

The BIO-ML collection is a unique resource, providing open access to thousands of clinically relevant, and in some cases underrepresented, strains and their accompanying omics data. With available cultivable isolates, this comprehensive resource has the potential to elucidate complex dynamics of the human gut microbiome and enable unprecedented hypothesis-driven research.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of code and data availability and

associated accession codes are available at https://doi.org/10.1038/s41591-019-0559-3.

Received: 11 March 2019; Accepted: 23 July 2019;
Published online: 2 September 2019

References

1. Shen, T.-C. D. et al. Engineering the gut microbiota to treat hyperammonemia. *J. Clin. Invest.* **125**, 2841–2850 (2015).
2. Ronda, C., Chen, S. P., Cabral, V., Yaung, S. J. & Wang, H. H. Metagenomic engineering of the mammalian gut microbiome in situ. *Nat. Methods* **16**, 167–170 (2019).
3. Holmes, E. et al. Therapeutic modulation of microbiota-host metabolic interactions. *Sci. Transl. Med.* **4**, 137rv6 (2012).
4. van Nood, E. et al. Duodenal infusion of donor feces for recurrent *Clostridium difficile*. *N. Engl. J. Med.* **368**, 407–415 (2013).
5. Kassam, Z., Lee, C. H., Yuan, Y. & Hunt, R. H. Fecal microbiota transplantation for *Clostridium difficile* infection: systematic review and meta-analysis. *Am. J. Gastroenterol.* **108**, 500–508 (2013).
6. Moayyedi, P. et al. Fecal microbiota transplantation induces remission in patients with active ulcerative colitis in a randomized controlled trial. *Gastroenterology* **149**, 102–109.e6 (2015).
7. Ratner, M. Microbial cocktails join fecal transplants in IBD treatment trials. *Nat. Biotechnol.* **33**, 787–788 (2015).
8. Mullish, B. H., McDonald, J. A. K., Thursz, M. R. & Marchesi, J. R. Fecal microbiota transplant from a rational stool donor improves hepatic encephalopathy: a randomized clinical trial. *Hepatology* **66**, 1354–1355 (2017).
9. Flameling, I. A. & Rijkers, G. T. Fecal Microbiota Transplants as a Treatment Option for Parkinson's Disease. *Gut Microbiota - Brain Axis* https://doi.org/10.5772/intechopen.78666(2018).
10. Fischer, M., Bittar, M., Papa, E., Kassam, Z. & Smith, M. Can you cause inflammatory bowel disease with fecal transplantation? A 31-patient case-series of fecal transplantation using stool from a donor who later developed Crohn's disease. *Gut Microbes* **8**, 205–207 (2017).
11. Smillie, C. S. et al. Strain tracking reveals the determinants of bacterial engraftment in the human gut following fecal microbiota transplantation. *Cell Host Microbe* **23**, 229–240.e5 (2018).
12. Li, S. S. et al. Durable coexistence of donor and recipient strains after fecal microbiota transplantation. *Science* **352**, 586–589 (2016).
13. Human Microbiome Jumpstart Reference Strains Consortium. et al. A catalog of reference genomes from the human microbiome. *Science* **328**, 994–999 (2010).
14. Faith, J. J. et al. The long-term stability of the human gut microbiota. *Science* **341**, 1237439–1237439 (2013).
15. Goodman, A. L. et al. Extensive personal human gut microbiota culture collections characterized and manipulated in gnotobiotic mice. *Proc. Natl Acad. Sci. USA* **108**, 6252–6257 (2011).
16. Browne, H. P. et al. Culturing of 'unculturable' human microbiota reveals novel taxa and extensive sporulation. *Nature* **533**, 543–546 (2016).
17. Lagier, J.-C. et al. Culture of previously uncultured members of the human gut microbiota by culturomics. *Nat. Microbiol.* **1**, 16203 (2016).
18. Zou, Y. et al. 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. *Nat. Biotechnol.* **37**, 179–185 (2019).
19. Forster, S. C. et al. A human gut bacterial genome and culture collection for improved metagenomic analyses. *Nat. Biotechnol.* **37**, 186–192 (2019).
20. Truong, D. T., Tett, A., Pasolli, E., Huttenhower, C. & Segata, N. Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Res.* **27**, 626–638 (2017).
21. Zhao, S. et al. Adaptive evolution within the gut microbiome of individual people. Preprint at https://doi.org/10.1101/208009 (2017).
22. Greenblum, S., Carr, R. & Borenstein, E. Extensive strain-level copy-number variation across human gut microbiome species. *Cell* **160**, 583–594 (2015).
23. Garud, N. R., Good, B. H., Hallatschek, O. & Pollard, K. S. Evolutionary dynamics of bacteria in the gut microbiome within and across hosts. Preprint at https://doi.org/10.1101/210955 (2017).
24. Ahern, P. P., Faith, J. J. & Gordon, J. I. Mining the human gut microbiota for effector strains that shape the immune system. *Immunity* **40**, 815–823 (2014).
25. Bron, P. A., van Baarlen, P. & Kleerebezem, M. Emerging molecular insights into the interaction between probiotics and the host intestinal mucosa. *Nat. Rev. Microbiol.* **10**, 66–78 (2011).
26. Barboza, M. et al. Glycoprofiling bifidobacterial consumption of galacto-oligosaccharides by mass spectrometry reveals strain-specific, preferential consumption of glycans. *Appl. Environ. Microbiol.* **75**, 7319–7325 (2009).
27. Rossi, M. et al. Fermentation of fructooligosaccharides and inulin by bifidobacteria: a comparative study of pure and fecal cultures. *Appl. Environ. Microbiol.* **71**, 6150–6158 (2005).
28. Lopez-Siles, M. et al. Cultured representatives of two major phylogroups of human colonic *Faecalibacterium prausnitzii* can utilize pectin, uronic acids, and host-derived substrates for growth. *Appl. Environ. Microbiol.* **78**, 420–428 (2012).
29. Haiser, H. J. et al. Predicting and manipulating cardiac drug inactivation by the human gut bacterium *Eggerthella lenta*. *Science* **341**, 295–298 (2013).
30. Wilson, I. D. & Nicholson, J. K. Gut microbiome interactions with drug metabolism, efficacy, and toxicity. *Transl. Res.* **179**, 204–222 (2017).
31. Cover, T. L. *Helicobacter pylori* diversity and gastric cancer risk. *MBio* **7**, e01869–15 (2016).
32. Arthur, J. C. et al. Intestinal inflammation targets cancer-inducing activity of the microbiota. *Science* **338**, 120–123 (2012).
33. Conway, T. & Cohen, P. S. Commensal and pathogenic *Escherichia coli* metabolism in the gut. *Microbiol Spectr* **3**, https://doi.org/10.1128/microbiolspec.MBP-0006-2014 (2015).
34. Rettedal, E. A., Gumpert, H. & Sommer, M. O. A. Cultivation-based multiplex phenotyping of human gut microbiota allows targeted recovery of previously uncultured bacteria. *Nat. Commun.* **5**, 4714 (2014).
35. Lau, J. T. et al. Capturing the diversity of the human gut microbiota through culture-enriched molecular profiling. *Genome Med.* **8**, 72 (2016).
36. Kearney, S. M. *et al*. Endospores and other lysis-resistant bacteria comprise a widely shared core community within the human microbiota. *ISME J.* **12**, 2403-2416 (2018).
37. Fodor, A. A. et al. The 'Most Wanted' taxa from the human microbiome for whole genome sequencing. *PLoS ONE* **7**, e41294 (2012).
38. Derrien, M., Vaughan, E. E., Plugge, C. M. & de Vos, W. M. *Akkermansia muciniphila* gen. nov., sp. nov., a human intestinal mucin-degrading bacterium. *Int. J. Syst. Evol. Microbiol.* **54**, 1469–1476 (2004).
39. Schneeberger, M. et al. *Akkermansia muciniphila* inversely correlates with the onset of inflammation, altered adipose tissue metabolism and metabolic disorders during obesity in mice. *Sci. Rep.* **5**, 16643 (2015).
40. Dao, M. C. et al. *Akkermansia muciniphila* and improved metabolic health during a dietary intervention in obesity: relationship with gut microbiome richness and ecology. *Gut* **65**, 426–436 (2016).
41. Sokol, H. et al. *Faecalibacterium prausnitzii* is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients. *Proc. Natl Acad. Sci. USA* **105**, 16731–16736 (2008).
42. Miquel, S. et al. *Faecalibacterium prausnitzii* and human intestinal health. *Curr. Opin. Microbiol.* **16**, 255–261 (2013).
43. Galperin, M. Y. et al. Genomic determinants of sporulation in *Bacilli* and *Clostridia*: towards the minimal set of sporulation-specific genes. *Environ. Microbiol.* **14**, 2870–2890 (2012).
44. Daubin, V., Moran, N. A. & Ochman, H. Phylogenetics and the cohesion of bacterial genomes. *Science* **301**, 829–832 (2003).
45. Lozupone, C. A., Stombaugh, J. I., Gordon, J. I., Jansson, J. K. & Knight, R. Diversity, stability and resilience of the human gut microbiota. *Nature* **489**, 220–230 (2012).
46. Windey, K., De Preter, V. & Verbeke, K. Relevance of protein fermentation to gut health. *Mol. Nutr. Food Res.* **56**, 184–196 (2012).
47. Jansson, J. et al. Metabolomics reveals metabolic biomarkers of Crohn's disease. *PLoS One* **4**, e6386 (2009).
48. Weir, T. L. et al. Stool microbiome and metabolome differences between colorectal cancer patients and healthy adults. *PLoS One* **8**, e70803 (2013).
49. Tramontano, M. et al. Nutritional preferences of human gut bacteria reveal their metabolic idiosyncrasies. *Nat. Microbiol.* **3**, 514–522 (2018).
50. Maier, L. et al. Extensive impact of non-antibiotic drugs on human gut bacteria. *Nature* **555**, 623–628 (2018).
51. Atarashi, K. et al. T_{reg} induction by a rationally selected mixture of *Clostridia* strains from the human microbiota. *Nature* **500**, 232–236 (2013).
52. Wlodarska, M. et al. Indoleacrylic acid produced by commensal peptostreptococcus species suppresses inflammation. *Cell Host Microbe* **22**, 25–37.e6 (2017).
53. Tanoue, T. et al. A defined commensal consortium elicits CD8 T cells and anti-cancer immunity. *Nature* **565**, 600–605 (2019).

Acknowledgements

The authors are thankful to M. Sovie, C. Kim, W. Kelley, E. Lee, W. Pettee, J. Watson and P. Panchal from OpenBiome for their assistance in processing materials and donor metadata used in this study. This work was funded by a grant from the Broad Institute (Broad Next 10 grant 4000017).

Author contributions

M.P., M.G., S.M.G, R.J.X. and E.J.A. designed the project. M.P. and M.G. built the library of bacterial isolates and whole genomes. M.P. and M.G analyzed whole-genome

sequence data. S.M.K., M.G. and M.P. analyzed the sporulation and ethanol-resistance data. S.M.G. analyzed the 16S data. S.M.G. and X.J. analyzed the metagenomics data. J.A.-P. analyzed the metabolomics data. M.P., S.M.K. and A.R.P. designed the culturing protocols. M.P. and B.B. curate the library of isolates. S.Z. and T.D.L. provided technical advice for WGS library preparation and analysis. P.K.S. and M.S. provided OpenBiome samples and associated metadata. S.R., J.E.A, S.A.R., J.L. and H.V. generated the 16s and metagenomics data. C.C., K.B., A.D., J.S. and K.A.P. generated the metabolomics data. M.P., M.G., S.M.G. and E.J.A. wrote the paper, with input from all authors. E.J.A. and R.J.X. obtained funding and supervised the project.

Competing interests

M.S. and E.J.A. are co-founders and shareholders of Finch Therapeutics, a company that specializes in microbiome-targeted therapeutics.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41591-019-0559-3.

Supplementary information is available for this paper at https://doi.org/10.1038/s41591-019-0559-3.

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

Correspondence and requests for materials should be addressed to R.J.X. or E.J.A.

Peer review information: Alison Farrell is the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Methods

Study cohort and sample collection. Stool samples were obtained from OpenBiome (https://www.openbiome.org/), a non-profit stool bank, under a protocol approved by the institutional review boards (IRBs) at MIT and the Broad Institute (IRB protocol ID no. 1603506899). Subjects were healthy people screened by OpenBiome to minimize the potential for carrying pathogens. They were 19–45 years of age (28 years old on average) and had body mass indexes of 17.5–29.8 (23.4 on average) at initial sampling. Individuals could be called healthy with confidence at recruitment based on the absence of ongoing symptoms, pain and medication, and based on past history of gastrointestinal conditions, autoimmune and inflammatory diseases, cardiovascular and metabolic conditions, neurological conditions, psychological conditions, cancer and infectious diseases. Supplementary Table 2 contains health, disease and social-history metadata on each subject. Subjects were deidentified before receipt of samples.

Raw stool samples were diluted 1:10 in 12.5% glycerol buffer and 0.9% NaCl, homogenized and filtered through a 330-μm filter. A total of 1,207 stool samples were obtained from 90 subjects from July 2014 to May 2016. Detailed information about samples used for 16S, metagenomic, metabolomic and/or isolation is in Supplementary Table 3.

Culture and isolation of bacterial strains. To culture and isolate bacteria, we used 11 OpenBiome stool samples collected across 11 healthy donors. In addition, we used 10 additional samples from one donor (am), longitudinally collected between December 2014 and May 2016.

To culture an exhaustive representation of the diversity of human gut bacteria, human fecal samples were processed anaerobically at every step in a chamber using gas monitors to constantly control physicochemical conditions (5% Hydrogen, 20% Carbon dioxide, balanced with nitrogen). Human fecal samples were diluted in prereduced (anaerobic) PBS (with 0.1 % L-cysteine hydrochloride hydrate). Diluted samples were then plated onto prereduced agar plates and incubated anaerobically at 37 °C for 7 d.

Nonselective media were first used to culture diverse groups of organisms from the gut microbiota. Then a series of selective culturing methods were used to isolate additional members of more specific taxonomic groups. The media used were comprised of commercially available components, without the need of complex ingredients like rumen fluid or fecal extracts. After incubation, bacteria were isolated by picking individual colonies with an inoculation loop and streaking them onto a second prereduced agar plate. After 2 d of incubation at 37 °C, 1 colony from each individual was re-streaked again onto another agar plate for 2 more days of incubation, increasing the purity of individual colonies. One colony from each individual streak was then inoculated in liquid medium in a 96-well culture plate. After 2 d of anaerobic incubation at 37 °C, the taxonomy of the isolate was identified using 16S rRNA gene Sanger sequencing (starting at the V4 region). We first amplified the full 16S rRNA gene by PCR (27F 5′-AGAGTTTGATCMTGGCTCAG-3′; 1492R 5′-GGTTACCTTGTTACGACTT-3′) and then generated a ~1-kb-long sequence by Sanger reaction (U515 5′-GTGCCAGCMGCCGCGGTAA-3′). All isolates were stored in −80 °C freezers with prereduced glycerol solution as a cryoprotectant. Detailed information about isolates is in Supplementary Table 1.

DNA extraction, library construction and Illumina sequencing of whole genomes. To extract the whole-genome DNA of each individual colony, we used the DNeasy UltraClean96 MicrobioalKit (Qiagen) and the PureLinkPro96_gDNAkit (Invitrogen). Genomic DNA libraries were constructed from 1.2 ng of DNA using the Nextera DNA Library Preparation kit (Illumina), with reaction volumes scaled accordingly. Prior to sequencing, libraries were pooled by collecting an equal quantity of each library from batches of 250 samples. Insert sizes and concentrations of each pooled library were measured using an Agilent Bioanalyzer DNA 1000 kit (Agilent Technologies). Paired-end 150-bp read sequencing was performed using an Illumina NextSeq 500 instrument (Illumina Inc) at the Broad Institute.

Draft assembly and annotation of whole-genome sequences. Reads were first demultiplexed using in-house scripts. We used Trimmomatic v0.36 (ref. [54]) for the quality filtering of data (with parameters PE -phred33 LEADING:3 TRAILING:3 SLIDINGWINDOW:5:20 MINLEN:50) and to remove barcodes and Illumina adapters. Reads were assembled de novo into contigs using SPAdes v.3.9.1 (ref. [55]) (with parameter—careful). To iteratively improve genomic assemblies, we used SSPACE v3.0 and GapFiller v1-10 (ref. [56]) to scaffold contigs and to fill sequence gaps (with default parameters). Scaffolds smaller than 1 kb were removed from genome assemblies. We aligned all reads back to the assembly to compute genome coverage using BBmap v37.68 (https://jgi.doe.gov/data-and-tools/bbtools/) and the covstats option (with default parameters). The final assemblies were annotated using Prokka v1.12 (ref. [57]) (with default parameters).

Assessing assembly quality. We measured genome assembly statistics using CheckM v1.0.7 (ref. [58]) (with parameters lineage_wf —tab_table -x fna Prokka_annotations/). Although we implemented many sanity checks in the culturing and isolation protocols used to build the isolate library, final isolate stocks might still have contained mixtures of multiple strains, or sometimes even different species. Consequently, we included several contamination-removal steps in our genome assembly pipeline. Small contigs with extreme coverage and similarity with different taxonomic groups are often a signature of contamination and impurity of the original colonies. Assemblies with contamination levels higher than 10 (as measured by CheckM, only 2% of original genomes) were cleaned using the following conservative approach: we sorted contigs by coverage, and we used the Strucchange R package to detect breakpoints in the distribution of coverage across contigs (with cov defined as a sorted vector of contig converages, the function br eakpoints(log(cov)~seq(1,length(cov))) was used to calculate the breakpoints. If multiple jumps in coverage data were detected, the contig with the highest coverage was selected as the breakpoint. Then, all contigs with higher or equal coverage to the breakpoint contig are excluded from the assembly file. We re-run CheckM on each filtered genome to measure contamination again. We chose to exclude all assemblies that still exhibited contamination levels higher than 10. Finally, after calling for single-nucleotide polymorphisms (SNPs) within all bacterial species of the library (see 'SNP calling'), we built phylogenetic trees to detect genomes that were obvious contaminants and removed those from the library. The final median contamination is 0.3%. We further removed all assemblies that had genome completeness lower than 90%. All summary and quality statistics can be found in Supplementary Table 4. Reads for isolate genomes of the HBC collection[19] were assembled and checked for contamination using the same assembly pipeline as described above.

Taxonomy calling. We used whole-genome information to call for taxonomies at the species level. We used an approach similar to the open-reference method used to cluster sequences and assign taxonomies from amplicon (usually 16S) sequencing data. We used the Mash distance[59] (with default parameters) to compute the pairwise distances across all 3,632 genomes. Mash computes a distance between two genomes on the basis of the Jaccard Index, which accounts for both mutation and gene-content differences. It was recently shown that clustering genomes using a Mash distance threshold of ≤0.05 is equivalent to using an average nucleotide identity threshold of ≥95%, and reconstructs groups of genomes that are in good agreement with the NCBI bacterial species taxonomy[59] (an ANI of ≥95% has historically been used to define bacterial species). We used an unsupervised hierarchical clustering approach to group genomes that had Mash distances ≤0.05 into taxonomic units using the bClust function from the micropan R package. We then measured the genetic distance between the representative genome of each species cluster (defined as the genome with the highest *N50*) and a reference set of 79,226 non-contaminated complete and draft genomes downloaded from the NCBI FTP repository (ftp://ftp.ncbi.nlm.nih.gov/genomes/) on 27 March 2017. As species names can be incomplete or incorrect for NCBI draft genomes, we manually curated Mash results to assign a taxonomy to each cluster. We assigned 'Unknown_genus' and 'Genus_sp.' names to clusters that had no hit in the NCBI genome collection or that are closely related to a characterized genus but had no hit to a known species within this genus, respectively. Further phylogenetic and comparative genomics analyses will be needed in the future to refine the taxonomic assignments of these genomes. All genome taxonomies are compiled in Supplementary Tables 1 and 4.

SNP calling. We aligned reads against reference genomes using Bowtie2 v2.2.6 (ref. [60]) (with parameters—n-ceil 0,0.01 —dovetail —no-mixed —very-sensitive -X 2000). Potential single-nucleotide variants were called with Samtools v1.2 (ref. [61]). We then used a series of functions from the PicardTools (v2.6.0) and GATK (v.3.7) packages with a set of very conservative filters to improve read alignments and remove false positive polymorphisms. The objective is to filter out variants that are either caused from sequencing errors or from systematic errors at particular genomic positions (for example, misalignment near insertion/deletion regions).

Briefly, we used the MarkDuplicates (with parameters REMOVE_DUPLICATES = true) and AddOrReplaceReadGroups (with default parameters) functions from the PicardTools package to mask regions with very high coverage, which may reflect duplication events. We then recalibrated base-quality scores using a set of functions from the GATK package. Base-quality score recalibration (BQSR) is a two-step process that models sequencing errors and adjusts the quality scores accordingly. First, haplotypes are called (HaplotypeCaller function with parameters —sample_ploidy 1 -mmq 40 —genotyping_mode DISCOVERY) on each individual sample, and variants are filtered (VariantFiltration function, with parameters described below). The BQSR is run (BaseRecalibrator function) to produce a recalibration table using the filtered SNPs as a reference. The reads with recalibrated quality scores are then used in a second phase of haplotype calling, variant filtering and base recalibrating. After these two steps, convergence of quality scores is checked (AnalyzeCovariates function). All individual gVCF files are then merged together (CombineGVCFs function) to jointly genotype all samples (GenotypeGVCFs function). We only conserved variants that fulfill all of the following criteria:

- A minimum read-mapping quality required to consider a read for calling higher than 40.
- A quality by depth (QD) higher than 2.0. QD is the Phred-scaled probability that a polymorphism exists at this site given sequencing data, normalized by allele depth.

- A strand bias (Phred-scaled probability that there is strand bias at the site) estimated using Fisher's exact test (FS) lower than 60.
- A strand bias estimated by the symmetric odds ratio test (SOR) lower than 4.0 (this is another way to control for strand bias).
- A root mean square mapping quality over all the reads at the site (MQ) higher than 40.
- A rank sum test comparing mapping qualities of reads supporting the reference allele versus the alternate allele (MQRankSum) higher than −4.0.
- A rank sum test comparing the relative positioning of reference versus alternate alleles within reads (ReadPosRankSum) higher than −2.0.
- A rank sum test comparing the reference versus alternate base-quality scores (BaseQRankSum) higher than −2.0.
- A rank sum test for hard-clipped bases on reference versus alternate reads (ClippingRankSum) higher than −2.0.

When a variant in a given sample did not pass these filters, the allele at this position was assigned a 'N'. In addition, when a variant was not supported by more than 10 reads, the allele is assigned a 'N'. When an ALT variant has a normalized Phred-scaled genotype likelihood (PL) lower than 50, the variant is also assigned a 'N'. When a given site contained more than 20% of Ns, the polymorphic site was discarded. Also, all positions that had a low average read depth across all samples were removed (all sites with a ratio between total read count across all samples and the total number of samples lower than 30 were discarded).

Finally, we removed variants that were likely to result from recombination and not de novo mutations using a sliding-window approach. When more than 4 SNPs occurred within a region of 4 kb, all polymorphic positions were removed from the SNP alignment.

Reference genomes used to call for SNPs in this study were *B. adolescentis* American Type Culture Collection (ATCC) 15703, *B. longum* ATCC 15697 *B. ovatus* ATCC 8483, *B. vulgatus* ATCC 8482, *T. sanguinis* PC909 and *A. muciniphila* 54 46.

Core- and pan-genome analyses. We used Roary v3.9.1 (ref. [62]) to reconstruct the core- and pan-genome of *B. longum* and *B. adolescentis* strains (with parameters -e -n -i 90 -cd 95 -r). To reconstruct gene families, we set the minimum percentage identity between protein sequences to 90%, and the minimum frequency of isolates in which a gene must be present to belong to the core genome to 95%. To include outgroups, all complete genomes for these two *Bifidobacterium* species present on the NCBI FTP repository (ftp://ftp.ncbi.nlm.nih.gov/genomes/) were also input to Roary. Outgroup B. adolescentis genomes are 22L, ATCC 15703 and BBMN23. Outgroup B. longum genomes are ATCC 15697, JDM301, BBMN68, JCM 1217, 157F, KACC 91563, BXY01, GT15, 105-A, BT1, BG7, NCIMB 8809, CCUG 30698, 35624 and AH1206.

Phylogenetic reconstructions. We used DNAPARS (parsimony) from the PHYLIP package v3.6 (with default parameters) and RAxML v8.0.0 (maximum likelihood) (with parameters -m ASC_GTRGAMMA —asc-corr=lewis) to reconstruct phylogenetic trees from the reconstructed SNP alignments of each bacterial species analyzed in this study. Because SNP alignments do not contain invariable positions, we corrected for the ascertainment bias (using the Lewis correction) when reconstructing trees with RAxML to correct likelihoods and branch-length estimates.

To reconstruct the phylogenomic tree of all 3,632 genomes, we first built a concatenated alignment of 47 nearly universal and single-copy ribosomal protein families. We used Diamond v0.8.22.84 (ref. [63]) (with parameters blastx —more-sensitive -e 0.000001 —id 35 —query-cover 80) to BLAST all 3,632 proteomes against the RiboDB database (v1.4.1) of bacterial ribosomal protein genes[64]. We excluded proteins bL17, bS16, bS21, uL22, uS3 and uS4, as they were not sufficiently distributed across all genomes. In each RiboDB gene family, we excluded genomes that contained gene duplicates. Then, we aligned all protein families individually with Mafft v7.310 (with parameter —auto). We filtered out misaligned sites using BMGE v1.12 (with parameters -t AA -g 0.95 -m BLOSUM30) and concatenated all individual alignments using Seaview v4.7. We reconstructed the phylogenomic tree using FastTree v2.1.10 (with parameters -lg –gamma).

DNA extraction from raw stool samples. For DNA extraction, the MoBio Powersoil 96 kit (now Qiagen Cat No./ID: 12955-4) was used with minor modifications. All samples were thawed on ice, and between 625 µL and 1 mL homogenized stool was transferred to the MoBio High Throughput PowerSoil plate (12955-4-BP) and centrifuged at 4,000*g* for 10 min. Supernatant was removed, and 750 µL of bead solution was added along with 60 µL of C1 solution. Samples were bead beaten on the TissueLyzer at 20 Hz for 10 minutes. The plate was then rotated 180 degrees and beaten for another 10 minutes at 20 Hz to ensure even beating across all wells. Samples were then centrifuged at 4,500*g* for 6 min and 850 µL of supernatant was transferred to a clean 1-mL collection plate with the remainder of the protocol, as per the manufacturer's instructions.

16S library preparation and sequencing. 16S rRNA gene libraries targeting the V4 region of the 16S rRNA gene were prepared by first normalizing template concentrations and determining optimal cycle number by way of qPCR. Two 25 µL reactions for each sample were amplified with 0.5 units of Phusion with 1X High Fidelity buffer, 200 µM of each dNTP, 0.3 µM of 515 F (5′-AATGATACGGCGACCACCGAGATCTACACTATGGTAATTGTG-TGCCAGCMGCCGCGGTAA-3′) and 806rcbc0 (5′-CAAGCAGAAGACGGC ATACGAGATTCCCTTGTCTCCAGTCAGTCAGCCGGACTACHVGGGTW TCTAAT-3′). 0.25 µL 100x SYBR was added to each reaction, and samples were quantified using the formula $1.75^{(\Delta Ct)}$. To ensure minimal overamplification, each sample was diluted to the lowest concentration sample, amplifying with this sample optimal cycle number for the library construction PCR. Four 25-µL reactions were prepared per sample with master mix conditions listed above, without SYBR. Each sample was given a unique reverse barcode primer from the Golay primer set[65]. Replicates were then pooled and cleaned via Agencourt AMPure XP-PCR purification system. Purified libraries were diluted 1:100 and quantified again via qPCR (two 25-µL reactions, 2× iQ SYBR SUPERMix (Bio-Rad, ref no. 1708880) with Read 1 (5′-TATGGTAATTGTGTGYCAGCMGCCGCGGTAA-3′), Read 2 (5′-AGTCAGTCAGCCGGACTACNVGGGTWTCTAAT-3′)). Undiluted samples were normalized by way of pooling using the formula mentioned above. Pools were quantified by Qubit (Life Technologies, Inc.) and normalized into a final pool by Qubit concentration and number of samples. Final pools were sequenced on an Illumina MiSeq 300 using custom index 5′-ATTAGAWACCCBDGTAGTCCGGCTGACTGACT-3′ and custom Read 1 and Read 2, mentioned above.

ASV and taxonomy calling of the 16S sequencing data. 16S amplicon sequence data was split into separate, forward and reverse, demultiplexed fastq files for each sample. These paired-end fastq files were used as input for DADA2 (ref. [66]) in R v3.4.3, run using default parameters. Amplicon sequence variants (ASVs) were estimated by DADA2 and summarized in a ASV-by-sample abundance matrix. Taxonomic identities were assigned using the RDP classifier and the RDP trainset 16 (https://zenodo.org/record/801828).

Metagenomic library preparation and sequencing. Whole-genome fragment libraries were prepared as follows. Metagenomic DNA samples were quantified by Quant-iT PicoGreen dsDNA Assay (Life Technologies) and normalized to a concentration of 50 pg/µL. Illumina sequencing libraries were prepared from 100–250 pg of DNA using the Nextera XT DNA Library Preparation kit (Illumina), according to the manufacturer's recommended protocol, with reaction volumes scaled accordingly. Prior to sequencing, libraries were pooled by collecting equal volumes (200 nl) of each library from batches of 96 samples. Insert sizes and concentrations for each pooled library were determined using an Agilent Bioanalyzer DNA 1000 kit (Agilent Technologies). Libraries were sequenced on HiSeq 2×101 to yield ~10 million PE reads.

Post-sequencing de-multiplexing and BAM and Fastq files are generated using the Picard suite (https://broadinstitute.github.io/picard/command-line-overview.html).

COG genes construction and abundance estimation. Shotgun metagenomic sequencing data contained 1.5×10^7 reads on average per sample. These data were quality-trimmed with trimmomatic with parameters 'LEADING:20 TRAILING:20 MINLEN:50'. We removed reads that align to the human reference genome (hg19) using BWA and default parameters. PCR duplicate sequences were removed with fastuniq. Post filtering, samples contain 9.8×10^6 reads on average per sample.

For each sample, the metagenomic data were assembled with metaSPAdes[67]. Protein-coding genes were predicted from each assembly with Prodigal[68] and then combined and clustered with CD-HIT ('-d 0 -n 10 -l 100 -p 1 -G 0 -c 0.95 -aS 0.8 -M 0 -T 0') to create a nonredundant gene set[69]. The gene set was then annotated with COG terms by rps-blast search[70]. Metagenomic reads from each sample were then aligned to the CDSs of the nonredundant gene set with bowtie2 (ref. [60]). The mean coverage of each gene in each sample was calculated. The sum of the mean coverages for all genes of a given COG family was used to estimate the abundance of the COG family. The relative abundance was obtained by dividing the absolute COG abundance by the total coverage of the COG-annotated genes for the sample. The coverage for each COG class in each sample was calculated by summing the relative abundance of all COG families belonging to the same class.

Metagenomics data, whole genomes and strain dynamics. We tracked longitudinal variation in abundance between the main strains of *B. vulgatus* and *B. ovatus* within individual am. SNPs were identified by mapping isolated genome sequencing data to NCBI reference genomes (see 'SNP calling'). Nucleotide alleles at SNP positions were extracted and concatenated into individual sequences for each strain. The sequences were then joined together to make a multiple sequence alignment for each species, and trees were reconstructed by parsimony (see 'Phylogenetic reconstructions'). Metagenomics sequencing reads were aligned to each isolate genome of *B. vulgatus* and *B. ovatus*, assembled from the same donor with Bowtie2 (ref. [60]), using default parameters. The counts of reads mapped to each allele at the SNP position were calculated. The mean frequency of the SNPs specific to each main strain was used to estimate their strain relative abundances.

Metabolomics. Stool metabolites were profiled using four complimentary liquid chromatography tandem mass spectrometry (LC–MS) methods designed to measure a broad range of metabolites, as described previously[71]. Briefly, two hydrophilic interaction liquid chromatography (HILIC) methods were used for the analysis of water soluble polar metabolites in positive (HILIC-pos) or negative (HILIC-neg) ion mode, and two reverse-phase chromatography methods for measuring lipids in positive ionization mode (C8-pos) or metabolites of intermediate polarity, such as bile acids and free fatty acids using a C18 column in negative ionization mode (C18-neg). For a detailed description of the methods see Supplementary Methods. Raw data were processed using TraceFinder 3.1 software (Thermo Fisher Scientific; Waltham, MA) and Progenesis QI (Nonlinear Dynamics; Newcastle upon Tyne, UK). Pooled plasma samples were analyzed after intervals of approximately 20 participant samples to enable standardizing temporal drift in instrument response over time and between batches. For each method, metabolite identities were confirmed using authentic reference standards or reference samples.

Owing to differences in stool water content, stool metabolomic data were median scaled. For this purpose, the medians of all feature intensities in each sample (per LC–MS method) were computed. The median of these values was then used to calculate a scalar for each sample that, when multiplied by each metabolite intensity for each sample, yields a data set where the median intensities across all samples are equal. Analyses were conducted using the data obtained from all four LC-MS methods after removal of features observed in <95% of the samples and imputing missing values with half of the minimum observed measurement for each feature. All analyses and figures were done using R (version 3.4.3). Dendrograms were generated using the Spearman correlation coefficient as the distance metric, and the Ward D clustering method using function in the stats package on samples from subjects for whom data were available from at least six time points. Dendrogram visualizations were generated using the dendextend package[72]. PCA and biplots were computed on log-transformed, scaled and centered data, using the PCA implementation in the prcomp function in the stats package, and functions available in the factoextra (v.1.0.5) package[73].

Statistics. Statistical analyses were run in R and scikit-bio. For Fig. 1d, we run Pearson correlation tests (donor aa: $n=20$, $t=0.68$, df$=18$, $r=0.16$, $P=0.51$; donor am: $n=18$, $t=0.79$, df$=16$, $r=0.19$, $P=0.44$; donor bq: $n=11$, $t=-0.49$, df$=9$, $r=-0.16$, $P=0.64$; donor cx: $n=13$, $t=$-1.59, df$=11$, $r=-0.43$, $P=0.14$). For Fig. 1e, we run a linear mixed-effects model using the lmer function in the lmerTest R package (CGM medium: $F=8.3006$, df$=60$, $P<2.2\times10^{-16}$; Mmm + Ab4 media: $F=15.039$, df$=48$, $P<2.2\times10^{-16}$). Individuals were considered as the fixed effect, and genus counts as the random effect. For the PERMANOVA test run on 16S data (Extended Data Fig. 4a), 10,000 permutations were run (pseudo $F=38.2454$, $P<0.0001$). For the PERMANOVA test on metabolite data (Fig. 6a and Extended Data Fig. 7), 10,000 permutations were performed (pseudo $F=2.40656$, $P<0.0001$). Pearson correlation tests were also run on species abundances (Extended Data Fig. 4b; single abundances: red dots, $n=193{,}500$, df$=193{,}498$, $r=0.46$, $P=0$; median abundances: black dots, $n=18{,}206$, df$=18{,}204$, $r=0.5$, $P=0$) and function abundances (Extended Data Fig. 5b; single abundances: red dots, $r=0.88$, $P=0$; median abundances: black dots, $r=0.94$, $P=0$).

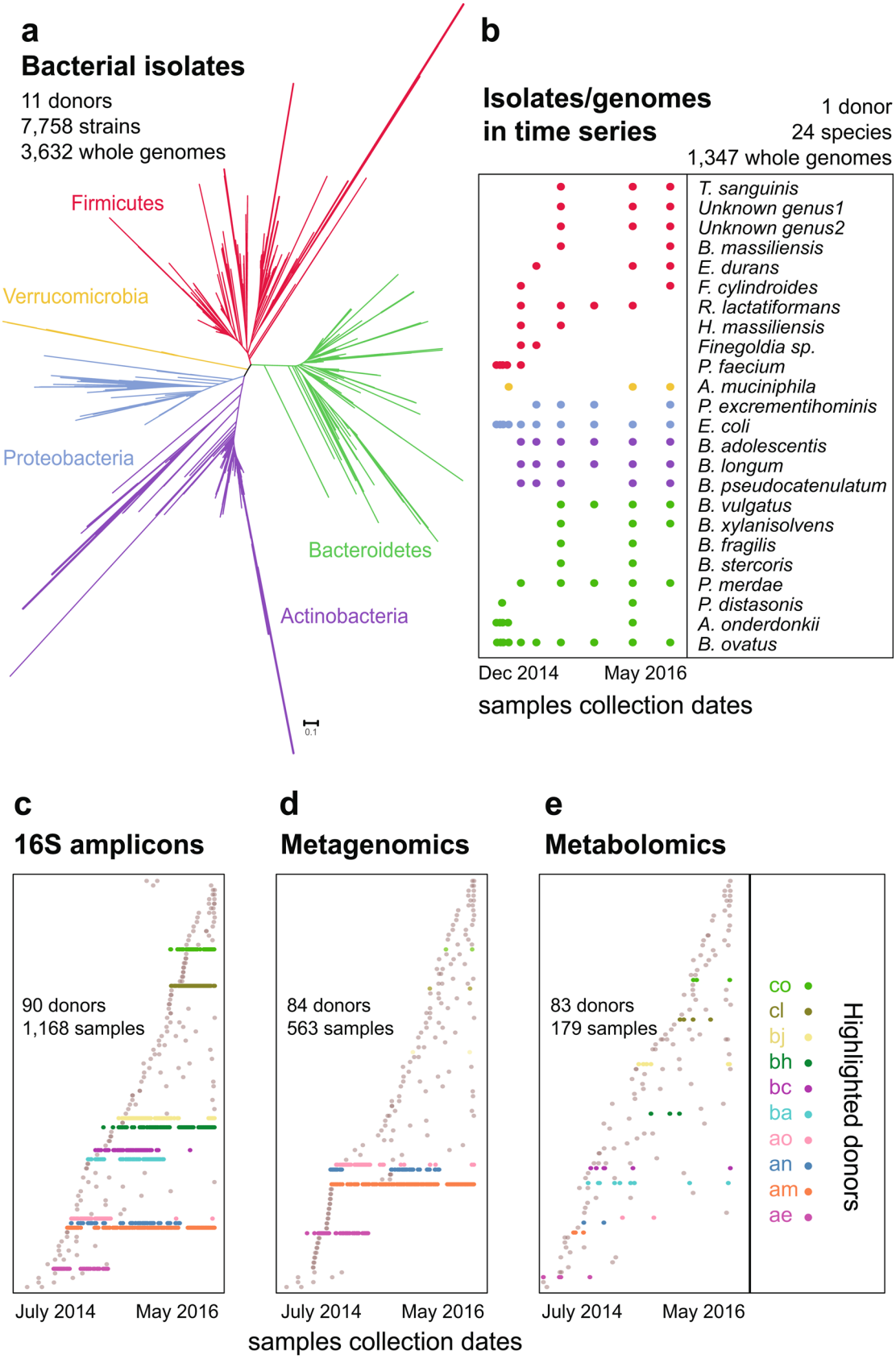
Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data Availability

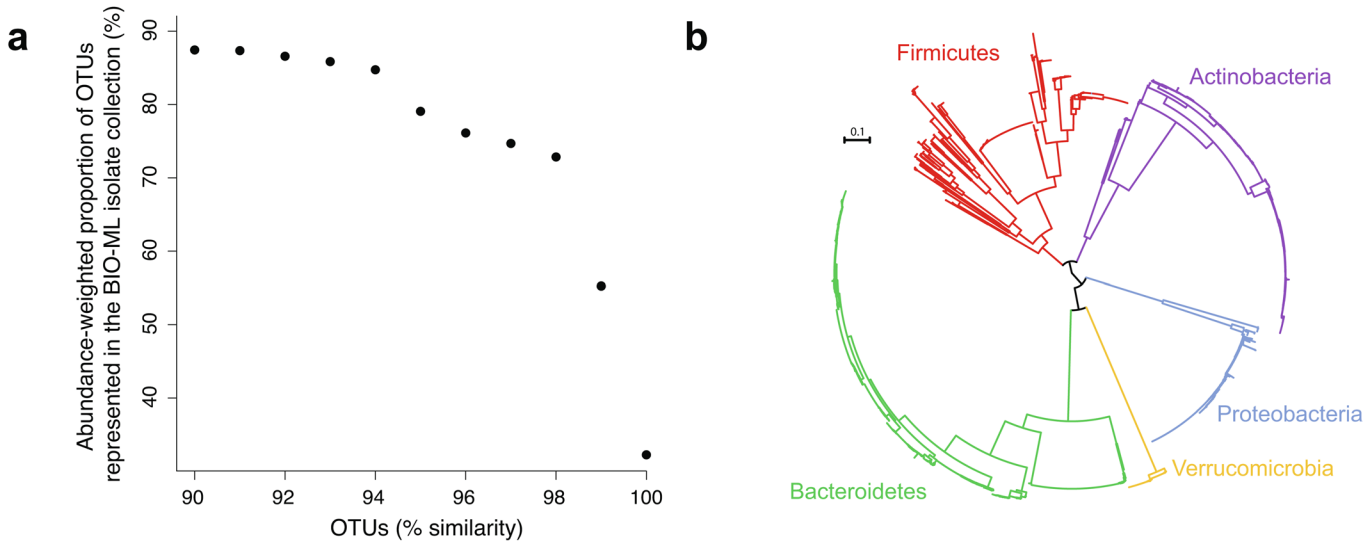
Sequencing and genomic data were deposited on the NCBI, under BioProject PRJNA544527.
BioSample accession numbers for raw sequencing data of isolate genomes: SAMN11846030-SAMN11847029; SAMN11847047-SAMN11848046; SAMN11848055–SAMN11849054; SAMN11849056-SAMN11849687.
BioSample accession numbers for isolate genome assemblies: SAMN11943001–SAMN11944000; SAMN11944002-SAMN11945001; SAMN11945004-SAMN11946003; SAMN11946038-SAMN11946669.
BioSample accession numbers for raw 16S data: SAMN11941243–SAMN11942242; SAMN11942243–SAMN11942410.
BioSample accession numbers for metagenomic data: SAMN11950000–SAMN11950562.
The processed metabolomics data is available at the NIH Common Fund's Metabolomics Data Repository and Coordinating Center (supported by NIH grant, U01-DK097430) website, the Metabolomics Workbench, http://www.metabolomicsworkbench.org, where it has been assigned Project ID PR000804.
Scripts and command lines used to analyze the sequencing and genomic data are available at https://github.com/almlab/BIO-ML.
The library of isolates is maintained and stored at the Broad Institute and strains will be made available for purchase upon request by researchers through a Broad Institute online platform: https://www.broadinstitute.org/bio-ml

References

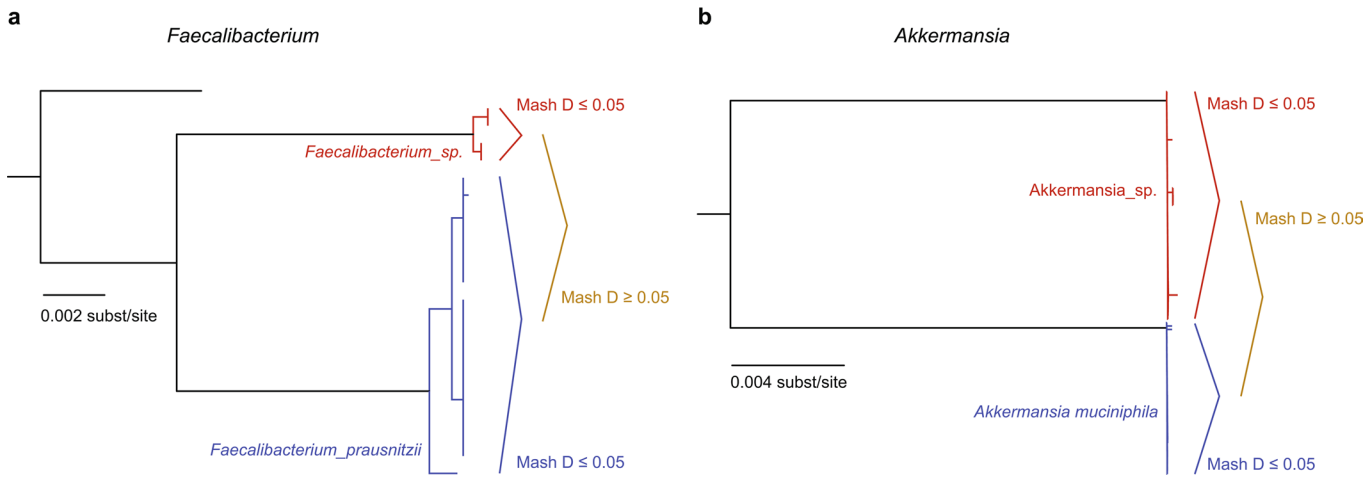
54. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
55. Bankevich, A. et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
56. Nadalin, F., Vezzi, F. & Policriti, A. GapFiller: a de novo assembly approach to fill the gap within paired reads. *BMC Bioinforma.* **13**, S8 (2012).
57. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
58. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
59. Ondov, B. D. et al. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**, 132 (2016).
60. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
61. Li, H. et al. The sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
62. Page, A. J. et al. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691–3693 (2015).
63. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
64. Jauffrit, F. et al. RiboDB Database: a comprehensive resource for prokaryotic systematics. *Mol. Biol. Evol.* **33**, 2170–2172 (2016).
65. Caporaso, J. G. et al. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J.* **6**, 1621–1624 (2012).
66. Callahan, B. J. et al. DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**, 581–583 (2016).
67. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).
68. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinforma.* **11**, 119 (2010).
69. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
70. Tatusov, R. L., Galperin, M. Y., Natale, D. A. & Koonin, E. V. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res* **28**, 33–36 (2000).
71. O'Sullivan, J. F. et al. Dimethylguanidino valeric acid is a marker of liver fat and predicts diabetes. *J. Clin. Invest* **127**, 4394–4402 (2017).
72. Galili, T. dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics* **31**, 3718–3720 (2015).
73. Kassambara, A. *Practical Guide To Principal Component Methods in R: PCA, M(CA), FAMD, MFA, HCPC, factoextra* (STHDA, 2017).



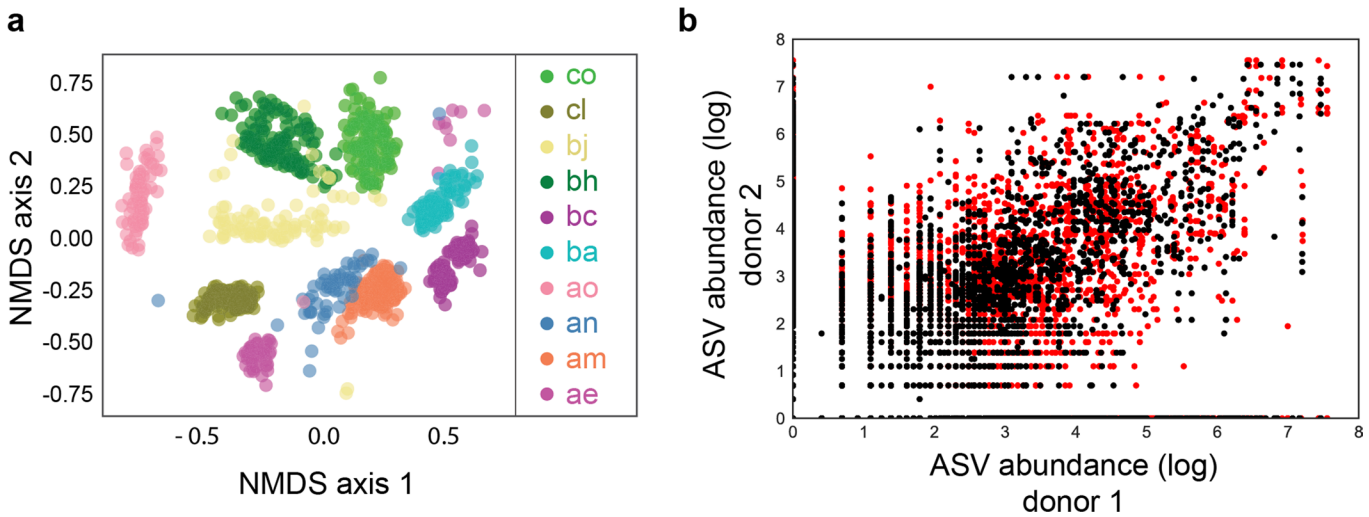
Extended Data Fig. 1 | Description of the BIO-ML. a, 16S phylogenetic tree of the 7,758 isolates in the BIO-ML. Lineages are colored by phylum. **b**, Depiction of the distribution of 1,347 isolates across 24 bacterial species (*y* axis) over time (*x* axis) that were whole-genome sequenced. **c**, Depiction of the distribution of 1,168 samples across individuals (*y* axis) and over time (*x* axis) that were processed for 16S amplicon sequencing. **d**, Depiction of the distribution of 563 samples across individuals and over time that were processed for shotgun metagenomic sequencing. **e**, Depiction of the distribution of 179 samples across individuals and over time that were processed for metabolomic study.



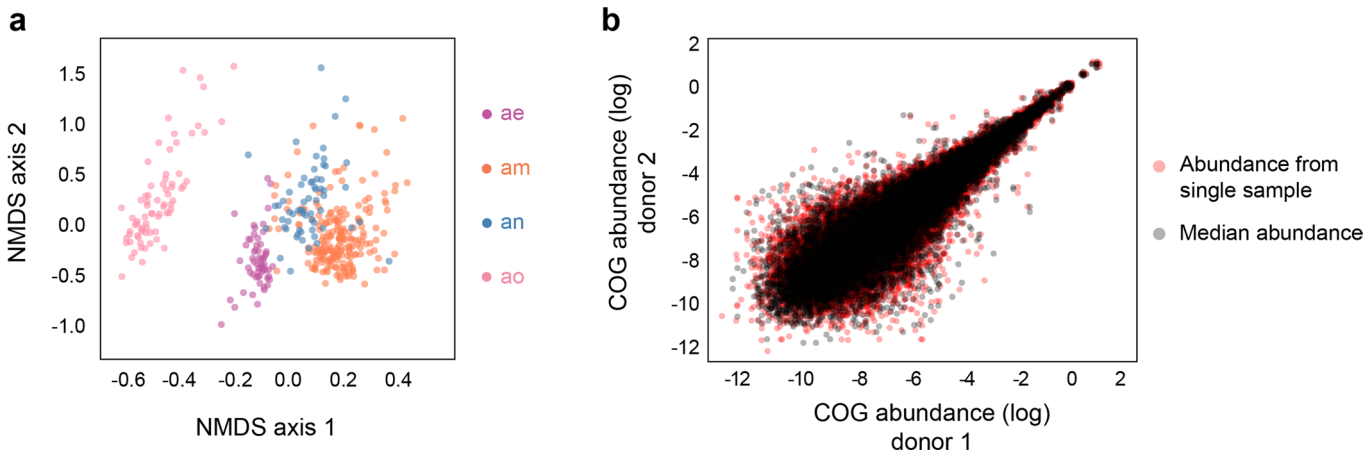
Extended Data Fig. 2 | Taxonomic coverage and composition of the BIO-ML of isolates and genomes. a, Abundance-weighted taxonomic coverage of the library of bacterial isolates (7,758 isolates) (*y* axis), compared to the diversity observed through culture-independent 16S amplicon sequencing (*x* axis). Eleven donors were used to build the library of isolates. The phylogenetic diversity of isolates was measured with 16S sanger sequencing, and this was compared to the total diversity observed in the 16S sequence data obtained from 1,168 samples from 90 individual donors of the BIO-ML. Taxonomic coverage was evaluated using different 16S OTU clustering thresholds, from 90% to 100% (ASV) similarity. **b**, Phylogenomic tree of the 3,632 genomes of the BIO-ML. Branches are colored by phylum.



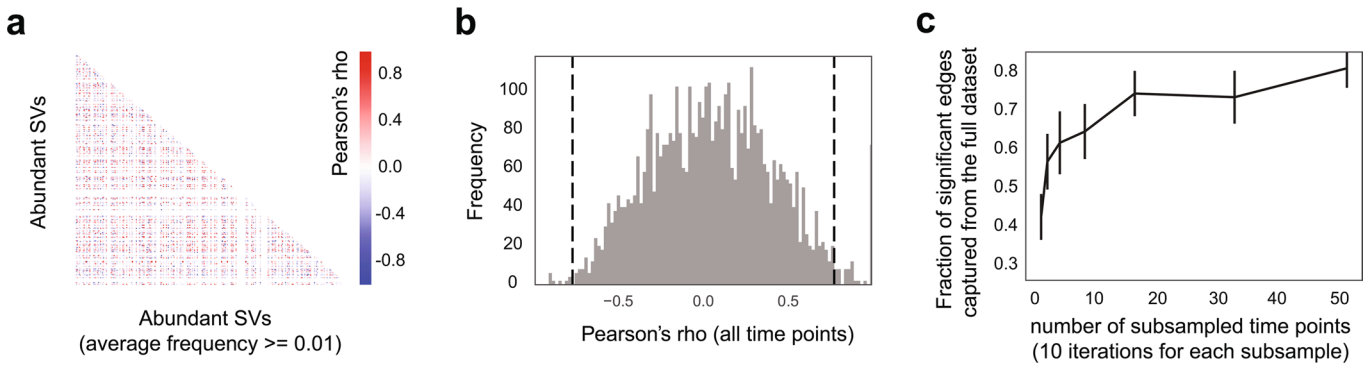
Extended Data Fig. 3 | The library of genomes contain multiple species within the *Faecalibacterium* and *Akkermansia* genera. Phylogenetic trees of *Faecalibacterium* (**a**) and *Akkermansia* (**b**) genomes were reconstructed using the concatenate alignment of ribosomal proteins (see Methods). We used RAxML to reconstruct the tree, using the PROTGAMMALGF substitution model. Pairwise Mash distances are represented on the right of each tree. Within each major clade, pairwise Mash distances were lower than 0.05, the threshold used to define species taxonomic units. Between clades, pairwise distances were higher than 0.05. Genomes in the *F. prausnitzii* and *A. muciniphila* clades have Mash distances with corresponding NCBI reference genomes that were lower than 0.05. Two different *Akkermansia* species are present in our genome library. At least two different *Faecalibacterium* species are present in the genome library.



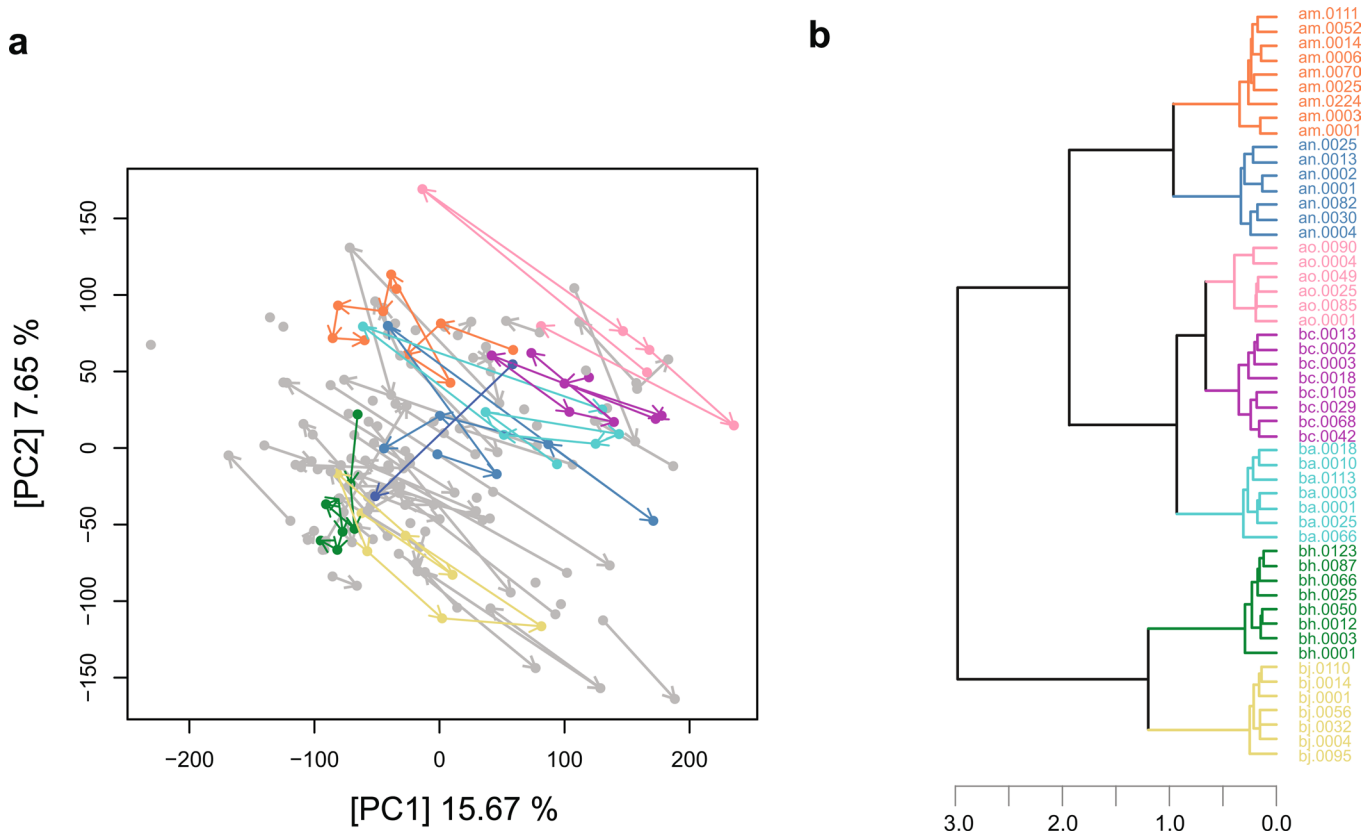
Extended Data Fig. 4 | Stability and conservation of microbiome species over time within and across people. a, Non-metric multidimensional scaling (NMDS) plot showing 16S community structure (Bray–Curtis distances) across long-term time series from ten stool donors. Samples are colored by donors (right). Donors maintain unique microbial signatures over many months to years (ANOSIM, $P < 0.0001$). **b**, The black points show the median abundance comparisons, and the red points show the results for a single, randomly drawn sample. Species abundances are conserved across donor pairs. The spread in the red points is larger than for the black points, indicating the median abundances show a tighter correlation across donors (black points Pearson's $R^2 = 0.25$; red points Pearson's $R^2 = 0.19$).



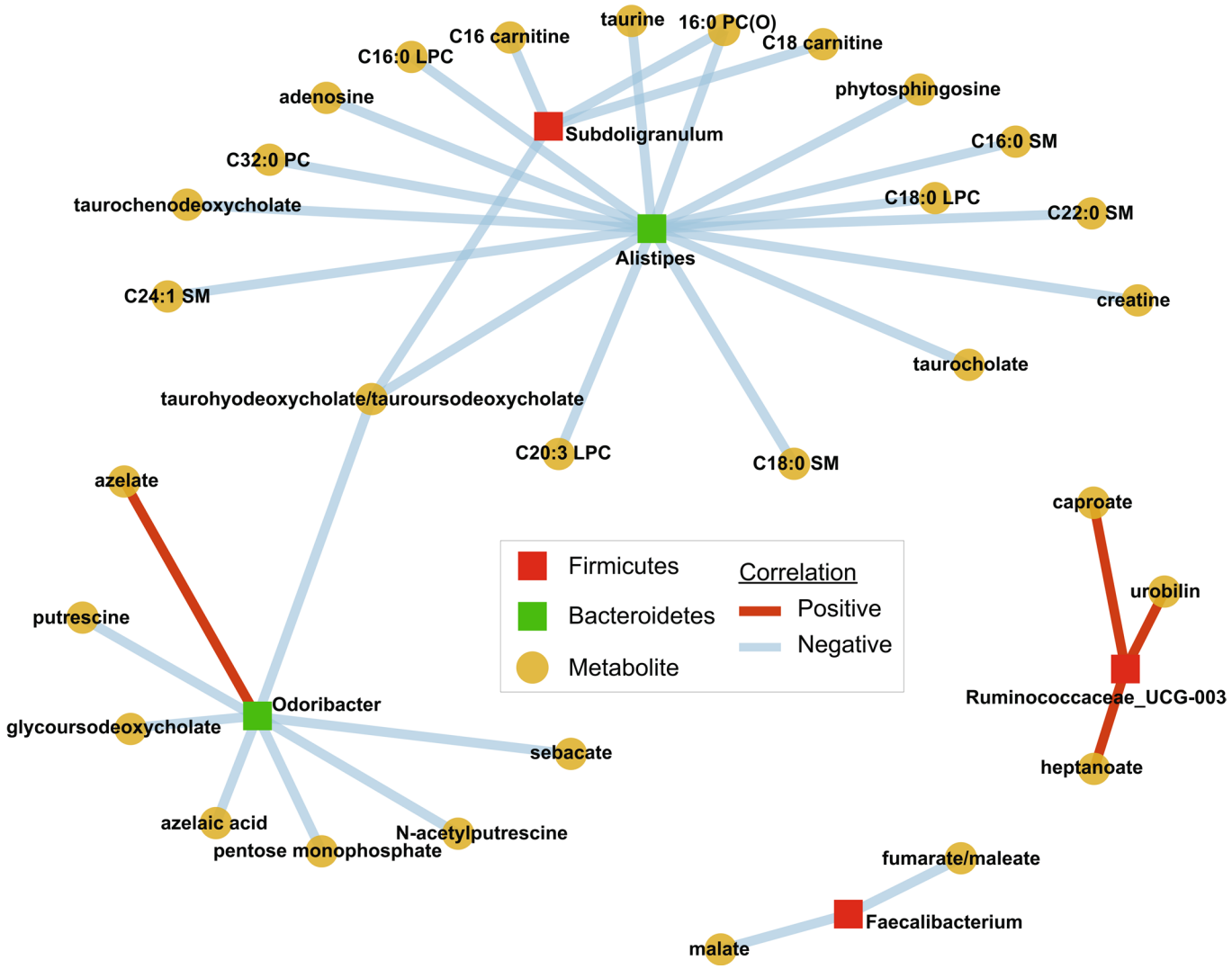
Extended Data Fig. 5 | Stability and conservation of microbiome functions over time within and across people. a, NMDS plot showing functional structure (Bray–Curtis distances) across long-term time series from four stool donors. Donors maintain unique functional signatures over many months-to-years. **b**, COG abundances are conserved across donor pairs. The black points show the median abundance comparisons, and the red points show the results for a single, randomly drawn sample. The spread in the red points is larger than that for the black points, indicating the median abundances show a tighter correlation across donors (black points Pearson's $R^2 = 0.88$; red points Pearson's $R^2 = 0.77$).



Extended Data Fig. 6 | Averaging taxa abundances across time points improves the identification of species–species correlations. a, Correlation matrix of log median ASV relative abundances across ten donors with long, dense time series (that is cross-sectional correlations) filtered to only look at abundant SVs with average frequencies of ≥0.01 across the dataset. **b**, Distribution of correlation coefficients from panel **a**. Dashed lines show the significance threshold ($P < 0.05$). Correlations beyond this threshold were used to infer a cross-sectional correlation network from the full dataset. **c**, The fraction of edges from the cross-sectional correlation network inferred from the full dataset that are captured by random subsampling of donor time series. Choosing a single sample from each donor only captures ~40% of 'true' network edges (number of iterations = 10).



Extended Data Fig. 7 | Metabolomics data capture crossdonor variation as well as within-donor variation through time. a, PC scores plot of all 179 samples for which metabolomic data were generated. Samples colored in gray correspond to subjects for which metabolomics data had been generated for less than six time points. Arrows connecting samples reflect the chronological order in which samples were collected. **b**, Dendrogram for donors for which metabolomics data had been generated for more than six time points. Metabolomes are colored by subject, as in **a**. The first two letters indicate the donor ID.



Extended Data Fig. 8 | Bacterial taxa–metabolites correlation network reveals strong functional associations in the human gut. Significant correlations between bacterial taxa and metabolite abundances (|Spearman's rho| > 0.7, $P < 0.01$) suggest a link between eating meat and bacterial community composition. *Alistipes* and *Subdoligranulum* are strongly associated with the bile acid taurocholate and its derivatives. *Subdoligranulum* is also associated with carnitine, which has been linked to eating meat. Other taxa are associated with acids and lipids common to the gut environment.

nature research

Corresponding author(s): Eric Alm

Last updated by author(s): 2019/07/13

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	N/A
Data analysis	Publicly available softwares for analyzing next-generation sequencing data were used: e.g. Dada2, Kraken, Trimmomatic, Spades, CheckM, Prokka, Mash. All softwares used, along with parameters, are listed and described in the Methods.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The datasets generated during the current study will be made available to the public upon acceptance of the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Samples were provided by OpenBiome. Sample sizes for 16S, metagenomics and metabolomics were determined to obtain densely-sampled time series within several individuals, which provided almost daily data over the course of several months. Samples from additional individuals with fewer within-subject time points were also included to obtain sufficient population-level data. 7,758 isolates were cultured from samples collected from 11 individuals. A large number of isolates were cultured and biobanked for several bacterial species that are easy to grow in vitro to obtain densely sampled strain diversity within subjects. Our culturing approach was also designed to maximize the phylogenetic diversity of isolates across our 11 individuals. We isolated multiple strains from longitudinal samples within one subject to investigate the within-person functional, ecological and evolutionary dynamics of strains. Finally, a set of 3,632 whole genome sequences were generated from the library of strains. Isolates chosen for whole genome sequencing represent the phylogenetic diversity of the original collection of isolates.
Data exclusions	A few isolates with apparent cross contamination were excluded from the collection of genomes. This is described in the Methods.
Replication	N/A
Randomization	N/A
Blinding	N/A

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☐ Antibodies
☐	☐ Eukaryotic cell lines
☐	☐ Palaeontology
☐	☐ Animals and other organisms
☐	☒ Human research participants
☐	☐ Clinical data

Methods

n/a	Involved in the study
☐	☐ ChIP-seq
☐	☐ Flow cytometry
☐	☐ MRI-based neuroimaging

Antibodies

Antibodies used	*Describe all antibodies used in the study; as applicable, provide supplier name, catalog number, clone name, and lot number.*
Validation	*Describe the validation of each primary antibody for the species and application, noting any validation statements on the manufacturer's website, relevant citations, antibody profiles in online databases, or data provided in the manuscript.*

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	*State the source of each cell line used.*
Authentication	*Describe the authentication procedures for each cell line used OR declare that none of the cell lines used were authenticated.*
Mycoplasma contamination	*Confirm that all cell lines tested negative for mycoplasma contamination OR describe the results of the testing for mycoplasma contamination OR declare that the cell lines were not tested for mycoplasma contamination.*
Commonly misidentified lines (See ICLAC register)	*Name any commonly misidentified cell lines used in the study and provide a rationale for their use.*

Palaeontology

Specimen provenance	*Provide provenance information for specimens and describe permits that were obtained for the work (including the name of the issuing authority, the date of issue, and any identifying information).*

Specimen deposition	*Indicate where the specimens have been deposited to permit free access by other researchers.*
Dating methods	*If new dates are provided, describe how they were obtained (e.g. collection, storage, sample pretreatment and measurement), where they were obtained (i.e. lab name), the calibration program and the protocol for quality assurance OR state that no new dates are provided.*

☐ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	*For laboratory animals, report species, strain, sex and age OR state that the study did not involve laboratory animals.*
Wild animals	*Provide details on animals observed in or captured in the field; report species, sex and age where possible. Describe how animals were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.*
Field-collected samples	*For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.*
Ethics oversight	*Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Human research participants

Policy information about studies involving human research participants

Population characteristics	Individuals live in the Boston area, and are Fecal Microbiota Transplant donors who donated samples to OpenBiome. Metadata on these individuals as provided by OpenBiome are available in Supplementary Table 1.
Recruitment	Participants were originally recruited by OpenBiome. Participants are healthy individuals, who were recruited following a strict health and lifestyle survey to be enrolled in the FMT program. Samples with potential presence of bacterial and eukaryotic pathogens were originally screened out by OpenBiome.
Ethics oversight	Stool samples were obtained from OpenBiome under a protocol approved by the institutional review boards at MIT and the Broad Institute (IRB protocol ID #1603506899)

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	*Provide the trial registration number from ClinicalTrials.gov or an equivalent agency.*
Study protocol	*Note where the full trial protocol can be accessed OR if not available, explain why.*
Data collection	*Describe the settings and locales of data collection, noting the time periods of recruitment and data collection.*
Outcomes	*Describe how you pre-defined primary and secondary outcome measures and how you assessed these measures.*

ChIP-seq

Data deposition

☐ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☐ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links *May remain private before publication.*	*For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.*
Files in database submission	*Provide a list of all files available in the database submission.*
Genome browser session (e.g. UCSC)	*Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.*

Methodology

Replicates	*Describe the experimental replicates, specifying number, type and replicate agreement.*
Sequencing depth	*Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end.*
Antibodies	*Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.*
Peak calling parameters	*Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.*
Data quality	*Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.*
Software	*Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.*

Flow Cytometry

Plots

Confirm that:

☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☐ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☐ All plots are contour plots with outliers or pseudocolor plots.

☐ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	*Describe the sample preparation, detailing the biological source of the cells and any tissue processing steps used.*
Instrument	*Identify the instrument used for data collection, specifying make and model number.*
Software	*Describe the software used to collect and analyze the flow cytometry data. For custom code that has been deposited into a community repository, provide accession details.*
Cell population abundance	*Describe the abundance of the relevant cell populations within post-sort fractions, providing details on the purity of the samples and how it was determined.*
Gating strategy	*Describe the gating strategy used for all relevant experiments, specifying the preliminary FSC/SSC gates of the starting cell population, indicating where boundaries between "positive" and "negative" staining cell populations are defined.*

☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

Design type	*Indicate task or resting state; event-related or block design.*
Design specifications	*Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials.*
Behavioral performance measures	*State number and/or type of variables recorded (e.g. correct button press, response time) and what statistics were used to establish that the subjects were performing the task as expected (e.g. mean, range, and/or standard deviation across subjects).*

Acquisition

Imaging type(s)	*Specify: functional, structural, diffusion, perfusion.*
Field strength	*Specify in Tesla*
Sequence & imaging parameters	*Specify the pulse sequence type (gradient echo, spin echo, etc.), imaging type (EPI, spiral, etc.), field of view, matrix size, slice thickness, orientation and TE/TR/flip angle.*
Area of acquisition	*State whether a whole brain scan was used OR define the area of acquisition, describing how the region was determined.*

Diffusion MRI ☐ Used ☐ Not used

Preprocessing

Preprocessing software	*Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.).*
Normalization	*If data were normalized/standardized, describe the approach(es): specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization.*
Normalization template	*Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized.*
Noise and artifact removal	*Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration).*
Volume censoring	*Define your software and/or method and criteria for volume censoring, and state the extent of such censoring.*

Statistical modeling & inference

Model type and settings	*Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation).*
Effect(s) tested	*Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used.*

Specify type of analysis: ☐ Whole brain ☐ ROI-based ☐ Both

Statistic type for inference (See Eklund et al. 2016)	*Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods.*
Correction	*Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo).*

Models & analysis

n/a	Involved in the study	
☐	☐	Functional and/or effective connectivity
☐	☐	Graph analysis
☐	☐	Multivariate modeling or predictive analysis

Functional and/or effective connectivity	*Report the measures of dependence used and the model details (e.g. Pearson correlation, partial correlation, mutual information).*
Graph analysis	*Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph, subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.).*
Multivariate modeling and predictive analysis	*Specify independent variables, features extraction and dimension reduction, model, training and evaluation metrics.*