HHS Public Access

Author manuscript

Nat Biotechnol. Author manuscript; available in PMC 2023 November 15.

Published in final edited form as:

Nat Biotechnol. 2023 October ; 41(10): 1424–1433. doi:10.1038/s41587-023-01674-2.

High-throughput microbial culturomics using automation and machine learning

Yiming Huang[1,#], **Ravi U. Sheth**[1,#], **Shijie Zhao**[1], **Lucas Cohen**[1], **Kendall Dabaghi**[1], **Thomas Moody**[1], **Yiwei Sun**[2], **Deirdre Ricaurte**[1], **Miles Richardson**[1], **Florencia Velez-Cortes**[1], **Tomasz Blazejewski**[1], **Andrew Kaufman**[1], **Carlotta Ronda**[1], **Harris H. Wang**[1,3,*]

[1]Department of Systems Biology, Columbia University, New York, NY, USA

[2]Department of Biomedical Informatics, Columbia University, New York, NY, USA

[3]Department of Pathology and Cell Biology, Columbia University, New York, NY, USA

Abstract

Pure bacterial cultures remain essential for detailed experimental and mechanistic studies in microbiome research and traditional methods to isolate individual bacteria from complex microbial ecosystems are labor-intensive, difficult-to-scale, and lack phenotype-genotype integration. Here, we describe an open-source high-throughput robotic strain isolation platform for the rapid generation of isolates on-demand. We develop a machine learning approach that leverages colony morphology and genomic data to maximize the diversity of microbes isolated and enable targeted picking of specific genera. Application of this platform on fecal samples from 20 humans yields personalized gut microbiome biobanks totaling 26,997 isolates that represented >80% of all abundant taxa. Spatial analysis on >100,000 visually captured colonies reveals co-growth patterns between *Ruminococcaceae*, *Bacteroidaceae*, *Coriobacteriaceae*, and *Bifidobacteriaceae* families that suggest important microbial interactions. Comparative analysis of 1,197 high-quality genomes from these biobanks shows interesting intra- and inter-personal strain evolution, selection, and horizontal gene transfer. This culturomics framework should empower new research efforts to systematize the collection and quantitative analysis of imaging-based phenotypes with high-resolution genomics data for many emerging microbiome studies.

Metagenomics offers the ability to broadly survey the composition of diverse microbial ecosystems ranging from soil communities to the gut microbiome. Yet microbes need to

*Correspondences should be addressed to hw2429@columbia.edu.
#These authors contributed equally.

Author Contributions Statement

Y.H., R.U.S. and H.H.W. developed the initial concept; Y.H. K.D. and T.B. developed morphology-based colony selection software; Y.H., R.U.S. and L.C. performed experiments and analyzed data with input from H.H.W.; T.M., D.R., Y.S., M.R., F.V.C., A.K., and C.R. assisted with colony isolation; Y.S. and S.Z. assisted with isolates whole genome sequencing; Y.H., R.U.S., S.Z. and H.H.W. wrote the manuscript. All other authors discussed results and approved the manuscript.

Competing Interests Statement

H.H.W. is a scientific advisor of SNIPR Biome, Kingdom Supercultures, Fitbiomics, Arranta Bio, VecX Biomedicines, Genus PLC, and a scientific co-founder of Aclid, all of whom are not involved in the study. R.U.S and K.D. are co-founders of Kingdom Supercultures. The authors declare no additional competing interests.

Code Availability

Scripts used to analyze plate images in this study can be accessed at https://github.com/hym0405/CAMII.

be isolated and cultured to mechanistically dissect their functional roles in a habitat and the myriad of interspecies processes that occur. Traditional cultivation methods relying on “brute force” random colony-picking are tedious and labor intensive[1–4]. Serial dilution-based isolation methods using 96 or 384-wells are resource-intensive and result in repeated isolation of the same dominant strains from the population[5]. Microfluidic systems enable growth in nanoliter reactors, but clonal isolates are difficult to extract[6, 7]. Given that a typical microbiome can contain hundreds to thousands of unique species exhibiting a long-tailed abundance distribution[8] (i.e., few dominate while most are rare), generating comprehensive strain collections via systematic culturomics remains an important and outstanding challenge.

Microbes can be distinguished based on their diverse phenotypes, whether by their ability to grow in certain media or the metabolites they produce[9–12]. Growth-based selection can enhance the isolation of rare species, for example with growth media containing different nutrients or antibiotics[1, 2, 13]. Mass spectrometry spectra can be used to differentiate between species[14, 15], but the approach is low-throughput and requires manual processing. Imaging-activated cell sorting has been developed to isolate eukaryotic cells based on multidimensional images, but this method requires sophisticated instrumentation and has not been implemented for bacteria[16]. With recent advances in artificial intelligence (AI) and deep learning models trained to discern nuanced features in multidimensional imaging and biological data[17], machine learning (ML) of combined phenotypic and genomic data streams is poised to transform next-generation microbial culturomics.

Here, we describe a ML-guided robotic strain isolation and genotyping platform that enables rapid and high-throughput generation of cultured biobanks on-demand. This system uses an intelligent imaging-based algorithm to increase the taxonomic diversity of culturomics compared to a random-picking method. We demonstrated the utility of this system by anaerobically generating personalized isolate biobanks for 20 human subjects, yielding a total of 26,997 isolates with 1,197 high-quality draft genomes, spanning 394 16S amplicon sequence variants (ASVs). Using the paired genomic and morphological information for each isolate, we trained a ML model that can predict taxonomic identity based only on colony morphology. Application of this ML model led to an improvement in targeted isolation of microbes-of-interest. Large-scale imaging analysis of all colonies grown on agar plates revealed interesting species-specific growth patterns and inter-species interactions. Whole genome analysis from personalized biobanks uncovered person-specific strain-level variation and signatures of horizontal gene transfer within major gut phyla. We further developed an open-access web-based database (http://microbial-culturomics.com/) containing searchable genotypic, morphologic, and phenotypic data of all isolates generated by automated culturomics as a unique and expanding community resource for the microbiome field.

RESULTS

Data-driven culturomics using phenotypes and automation

Colony picking is a classic microbiology method for clonally isolating bacterial strains. Colony growth on plates depends on many factors, including the composition of the media

(e.g., available nutrients), atmospheric conditions (e.g., level of oxygenation), presence of inhibitory molecules (e.g., antibiotics), pH, humidity, and effects of other diffusible metabolites derived from nearby colonies[18–20]. Different colony morphologies are observed based on strain-specific physiological differences, influenced by cell shape, rigidity, motility and growth kinetics, as well as production of pigmented molecules or extracellular matrices and surfactants[9–12]. Even though these colony traits are readily quantifiable, they are rarely documented during colony isolation. As a result, selective colony picking using visual features is generally qualitative and not standardized, and outcomes can vary substantially between experiments and experimentalists. To address these shortcomings, we devised a platform dubbed Culturomics by Automated Microbiome Imaging and Isolation (CAMII) to systematize culturomics with both morphologic and genotypic data for colony isolation and functional analysis.

The CAMII platform consists of four key elements (Figure 1A): (1) an imaging system that collects morphology data of colonies and an AI-guided colony selection algorithm, (2) an automated colony-picking robot for high-throughput isolation and arraying of isolates, (3) a cost-effective pipeline to rapidly generate genomic data for picked isolates, and (4) a physical isolate biobank and digital database with searchable colony morphology, phenotype, and genotype information. Thus, this end-to-end culturomics platform can produce isolates collections from diverse input microbiomes with minimized manual labor. The entire imaging and isolation system is built using off-the-shelf components housed in an anaerobic chamber that provides real-time control of temperature, humidity, and oxygen levels (Figure 1B, Table S1). The CAMII robot has an isolation throughput of 2,000 colonies per hour and can handle 12,000 colonies per run, which is >20 times higher capacity and faster than manual colony isolation by a person. To ensure that our genomic analysis capacity matches the robotic isolation throughput, we also developed a low-cost, high-throughput sequencing pipeline that leverages liquid handling automation to generate barcoded libraries for 16S rRNA sequencing or whole genome sequencing (WGS) (Methods). The cost per isolate in this pipeline is $0.45 for colony isolation and genomic DNA preparation, $0.46 for 16S rRNA sequencing, and $6.37 for WGS at a coverage of >60X on an Illumina HiSeq platform, which is substantially cheaper than commercial services (Table S2).

A key unique feature of the CAMII platform is the imaging system that collects and learns from morphological data of bacterial colonies (Figure 1C). Specifically, trans-illuminated images, which show height, radius, and circularity of a colony, and epi-illuminated images, which show color and complex morphological features such as wrinkling, are captured on CAMII to yield a multidimensional and quantifiable morphological dataset. We developed a custom colony analysis pipeline that segments colonies along diverse morphological features (Methods; Table S3, Figure S1). Area, perimeter, and mean radius reflect colony size, while circularity, convexity and inertia reveal colony shape. Pixel intensities and their variances in the red, green blue (RGB) channels highlight any density gradations and colors across a colony (Figure 1D). We next reasoned that morphologically distinct colonies are more likely to be phylogenetically diverse, which could be used to improve colony isolation. Thus, we developed an imaging-guided “smart picking” strategy to isolate more diverse isolates by embedding colonies to a multidimensional Euclidean space based on captured features

and selecting maximally distant points in this space representing the most morphologically distinct colonies (Figure S1; Methods). To further increase the diversity of bacteria that can be cultured and examined, CAMII also utilizes different antibiotic supplements to enrich the most unique and diverse subsets of microbes[1, 13] (Figure S2A, B). For instance, in a healthy human gut microbiome sample (H1t1), three antibiotics (Ciprofloxacin, Cip; Trimethoprim, Tmp; Vancomycin, Van) with different mechanisms of action elicited the most distinct enrichment cultures (Figure 1E, Figure S2C).

To systematically evaluate the capacity and fidelity of imaging-guided colony isolation, we applied CAMII to gut microbiome samples from three human volunteers (H1t4, H5t1 and H6t1; Table S4). Morphological data from plated colonies were analyzed by Principal Component Analysis (PCA) to assess the most informative visual features (Figure 1C, Figure S1C; Methods). Interestingly, colony density and size were the most dominant signatures (Principal Component 1 and 2, respectively) that together accounted for 72.0% of the morphological variance (Figure S3). We then used the CAMII robot to isolate 6,144 colonies, roughly a half of them are randomly picked from mGAM plates and another half by using our imaging-guided “smart picking” strategy and antibiotic selection. Isolates were grown in 384-wells and subjected to 16S rRNA sequencing for taxonomy identification. Unique 16S-V4 Sequences were then clustered into amplicon sequence variants (ASVs, 100% identity cutoff) that provide approximate species-level identity[21]. Remarkably, colony isolation informed by phenotypic data yielded a substantially more diverse set of ASVs than compared to random isolation for all three microbiome samples (Figure 1F). For example, to obtain 30 unique ASVs, we require only 85 ± 11 colonies to be isolated using our imaging selective strategy compared to 410 ± 218 colonies needed by random selection. Importantly, this enhanced isolation efficiency was maintained throughout picking, implying that there is a sustained advantage in using our strategy at a range of desired isolation depth (Figure S4A), and the generated isolate collection better represented the underlying input microbial diversity and were substantially more even in composition as measured by Shannon’s equitability (Figure S4B). Phylogenetic analysis of isolates showed that CAMII-optimized colony picking significantly improved the diversity of obtained microbes (Figure S5). This advantage is particularly evident given that finding unique ASVs becomes asymptotically more difficult with an increasing number of isolates. Altogether, these results demonstrated our AI-guided data-driven isolation framework in the CAMII platform can significantly increase the efficiency of culturomics and lessened the labor to isolate especially rare species.

Rapid generation of personalized gut isolate biobanks

While microbiome from different people may share similar sets of bacterial species, the strains belonging to these species are highly unique to the individual and may co-colonize the same host for many years[22, 23]. We sought to showcase the utility of CAMII to generate personalized gut isolate collections for 20 healthy people (Table S4, Figure S6A, B). A total of 102,071 colonies were visually analyzed and 26,997 colonies were picked and taxonomically identified by 16S rRNA sequencing (Figure 2A), yielding 394 unique ASVs that cover a broad diversity of healthy commensal gut microbiome (Figure 2B, C; Table S5).

To assess the comprehensiveness of this isolate collection, we calculated the abundance of isolated ASVs in the corresponding fecal samples by bulk 16S rRNA sequencing (Figure 2D). Remarkably, for each individual, 80.9 ± 9.4% of the ASVs by abundance are represented at least once in the entire isolate collection. Isolates derived from each person constituted on average 45.6 ± 21.6% of the total bacterial ASV abundance within that individual (Figure 2D). Moreover, comparison of isolate collections and bulk feces samples showed most of the highly abundant and prevalent ASVs are isolated at least once in the collection (Figure S6C–E). Moreover, each personalized isolate collection mimics the bulk feces sample with comparable microbiome profiles and Shannon's diversity index (Figure S6F, G).

In all, we demonstrated the use of CAMII to build a deep human gut isolate collection containing 26,997 isolates spanning 394 ASVs, with a rich set of linked morphologic, phenotypic, taxonomic, and WGS data. To increase its utility for the research community, we further developed a searchable online resource (http://microbial-culturomics.com) to house all CAMII-enabled biobank data including genomes, phenotypes, and images. We envision this portal will facilitate further genotype-to-phenotype analyses and lead to more shared isolate collections from other environments.

Identifying under-cultured "dark matter" gut microbiome

Previous studies have observed that many microbes from different environments are difficult to culture in the laboratory[24, 25]. We therefore leveraged our systematically generated isolate biobanks to assess the culturability of the human gut microbiome and to identify bacterial ASVs that remain recalcitrant to isolation in our experimental setting. Across all 20 personalized isolate collections, we determined whether abundant ASVs in the bulk fecal matter (average relative abundance > 0.1%) are found in the biobank. Notably, a substantial fraction of the uncultured gut bacteria belonged to the *Ruminococcaceae* and *Lachnospiraceae* families (Figure 2E, Table S6), which has also been previously documented as "unculturable"[24]. For each ASV, we compared the number of isolates generated in our total isolate collection versus their average abundance in the bulk feces (Figure 2F), which appeared to be positively correlated. Still, we identified a set of abundant yet difficult-to-culture bacteria, including *Faecalibacterium* ASV-58, *Prevatella* ASV-470 and ASV-324, *Oscillibacter* ASV-215 and *Clostridium XlVa* ASV-287 (Figure 2G). Interestingly, *Faecalibacterium* ASV-58, from which we obtained one isolate and performed WGS, matched with >98% genome-wide average nucleotide identity (ANI) to the metagenome-assembled genome (MAG) of *Candidatus cibiobacter qucibialis*. This strain in our collection was previously reported as the most abundant uncultured species in human gut[25] and is highly depleted in IBD patients as are other *Faecalibacterium* strains[26].

We further compared isolates in our biobanks to existing database[1, 3, 22, 25] by WGS and identified 11 additional species that had not been cultivated in any reference collections (BIO-ML, CGR, HMP) but are only associated with MAGs in the SGB collection (Figure S7; Table S7). For example, besides *Faecalibacterium* ASV-58, we isolated another abundant species *Faecalibacterium sp.* ASV-76 that represents >3% relative abundance on average in the bulk fecal matter, which further expands the collection of culturable gut microbiome.

Together, these results highlight cultured isolates and the remaining missing diversity based on our current media and growth conditions, and offer directions to guide future culturomics efforts focused on these “dark matter” gut microbiome (Table S6).

Taxonomy prediction from morphology enables targeted isolation

Focused cultivation of bacteria of interests from a microbiome sample can be crucial for mechanistic studies. Unfortunately, we lack the capacity to selectively culture most bacterial species in a specific manner. Consequently, picking a large number of colonies and relying on statistical probability is the only practical solution for obtaining the bacteria of interest. This strategy, however, is often too resource-consuming as it may require manually picking thousands of colonies. CAMII offers a machine learning-guided and automated colony selection method based on linking taxonomical identity to colony morphology and thus could in theory enhance targeted isolation. To test this, we systematically probed our deep gut isolate collection to analyze the relationship between morphologic and genotypic data. Interestingly, colonies of different genera exhibited diverse morphological patterns (Figure 3A, B). For example, colonies of *Dorea, Bacteroides* and *Collinsella* are generally large and dense but show different circularity (*Collinsella* > *Bacteroides* > *Dorea*), reflecting differences in their growth characteristics. On the other hand, colonies of *Faecalibacterium* are smaller and fainter, in line with our earlier results of their poor culturability. Furthermore, colony morphologies are significantly clustered according to their phylogeny ($p = 0.008$ by PERMANOVA test in Figure 3C). For instance, most genera of *Clostridia* are closer to each other by morphology-based ordination (Figure 3C). Therefore, colony morphologies may embed a substantial amount of information that could be linked to taxonomic identities.

We assessed whether taxonomic identity of colonies could be uniquely predicted by only incorporating their morphologic information on plates. We trained a random forest classification model using morphology and taxonomy data from randomly selected subsets of isolates (70% of total) (Methods). The model performance was evaluated on the remaining 30% of isolates. Remarkably, our model achieved ~70% precision for most genera that had more than 100 isolates in the training dataset (Figure 3D). The recall rate at the genus level varied more widely, highlighting open opportunities to use more sophisticated models to learn additional unique colony features[27–29]. Some genera such as *Eggerthella* had high precision and recall, indicating that highly conserved and unique colony morphologies could be specifically leveraged for taxonomic predictions. When analyzing isolates from the same ASV, we found that colony morphology was highly conserved for isolates within the same person but was much more variable between isolates from different people (Figure S8). Given that different people usually carry distinct strains of the same species, our results suggested a high degree of strain-level variation in colony morphology.

To assess whether AI-informed colony features can improve targeted microbe isolation, we next trained random forest models on our biobank isolates data from 3 different people separately (H12, H13, H14). The models were used to predict colonies of *Bifidobacterium*, *Parabacteroides* and *Eggerthella* from new plates derived from the same fecal samples,

and the colonies were then isolated by CAMII and 16S rRNA sequenced to confirm taxonomic identity (Methods). Notably, morphology-guided picking substantially improved the isolation efficiency for these targeted genera by up to 8-fold on average (Figure 3E), largely increasing the precision of picking and mitigating the need to screen many colonies to find the desired microbes. These results emphasize the value of our biobank datasets that link phenotype to genotype and demonstrate taxonomic predictions from visual colony features alone, which can greatly enhance targeted microbial isolation.

Inter-bacterial growth associations between gut microbiota

Bacterial colonies can influence the growth of their neighbors through species interactions such as competing for nutrients or cross-feeding essential metabolites. Previous studies suggest that neighboring cells can critically affect the size of colonies in a predictable manner[19]. Since CAMII can track the kinetic growth of colonies continuously, we systematically probed the co-growth associations between gut isolates on agar plates. A fecal sample (H1t5; Table S4) was plated and imaged daily, and all colonies were subsequently isolated at Day 6 and their taxonomical identity were determined with 16S sequencing (Figure 4A). For each ASV, the cumulative area of colonies on agar plates correlated with their abundances in the original fecal sample (Figure S9), indicating that our *in vitro* conditions generally fostered growth to the same degree as in the gut. Interestingly, colonies belonging to the *Faecalibacterium* genus exhibited slower initial growth and only began to emerge in the presence of other nearby growing colonies (Figure 4B; Methods). This observation suggests that commensal or mutualistic interactions may be at play between *Faecalibacterium* and other species.

To more systematically study species interactions enabled by CAMII, we analyzed the colony morphology, taxonomic identity, and colony neighborhood data together. We aggregated morphology data and physical coordinates from 102,071 visually captured colonies (26,997 isolated) and assessed whether a colony's growth is affected by neighboring cells. Surprisingly, we observed a number of interesting co-growth patterns that may reflect interspecies interactions (Table S8). For example, the colony size of *Phocaeicola vulgatus* ASV-6 is negatively correlated with the number of neighboring cells, consistent with a scenario that there are general negative interactions mediated by competition or antagonism between *P. vulgatus* and other bacteria in the gut[30] (Figure 4C). On the other hand, *Faecalibactrium prausnitzii* ASV-39, one of the species associated with slower initial growth in colony kinetics (Figure 4B), grew larger colonies with more neighbors reflective of a positive species interaction (Figure 4C).

We next incorporated taxonomic information of nearby colonies and looked at how colony size of specific genus could be affected by other genera. Briefly, for each pair of genera, we compared the colony sizes of one genus with the other genus present in the neighborhood and without any colonies present (Methods). Remarkably, we identified isolates from two genera, *Faecalibacterium* and *Clostridium IV*, that grow into larger sizes when the isolates were close to *Bifidobacterium*, *Phocaeicola* and *Bacteroides* (Figure 4D). *Faecalibacterium* and *Clostridium IV* have been reported to be major butyrate producing bacteria in the gut and could benefit from co-culture growth with *Bifidobacterium* and

Bacteroides species[31–33], which is consistent with our findings. On the other hand, we observed that *Phocaeicola* isolates are smaller with *Faecalibacterium* isolates as neighbors (Figure 4D), indicating that the co-growth interaction might be beneficial to only one side. Furthermore, consistent with our previous correlation analysis that examined neighboring isolate numbers without the consideration of neighbors' identity, we observed that growth of *Phocaeicola* and *Bacteroides* could be inhibited by multiple other genera, suggesting further investigations to better understand the underlying mechanism of these positive and negative interactions between gut microbiota. Together, our results highlight that CAMII can reveal colony co-growth patterns governed by interspecies interactions, which may help identify growth-promoting microbes and their diffusible metabolites that stimulate *in vitro* growth of fastidious species.

Intra- and inter-personal genomic diversity of gut strains

Mapping the strain-level genome-wide diversity of gut bacteria within a person is important for understanding the dynamics of gut colonization and the drivers of bacterial selection and adaptation specific to each human host[1, 2, 34]. A key advantage of the CAMII system is the ability to isolate and perform WGS for a large number of isolates to help investigate the inter- and intra-personal genomic variations. As such, we selected isolates covering the most unique and prevalent ASVs from our 20-person microbiome biobank and performed WGS that yielded 1,197 high-quality draft genomes (Figure S10; Table S9). Genome assemblies were further analyzed to determine the accurate species-level taxonomy of isolates (Methods).

We first explored the inter-personal strain-level genomic variations across our isolate collection (Methods). Consistent with previous reports[1, 35], most isolates within the same individuals had very few genomic variations (i.e., less than 10^2 SNPs) while isolates between people differed by 10^3 to 10^5 genome-wide SNPs (Figure 5A). Interestingly, some phylogenetically distinct isolates (i.e., more than 10^4 SNPs) of the same species were observed to co-exist within the same person (Figure 5A). For instance, two distinct strains of *P. vulgatus* were isolated from the H4 individual and two distinct strains of *B. uniformis* were found in the H2 individual (Figure S11).

We next sought to assess the strain-level diversity within a single person by analyzing 408 isolate genomes derived from the H1 individual (Figure S10; Methods). Since abundant species in the gut are expected to undergo more cell divisions, we hypothesized that they may accumulate more SNPs across their genomes, assuming approximately the same duration of gut colonization. Indeed, the number of genome-wide SNPs within each taxon is generally correlated with its abundance in the original microbiome (Figure 5B). *B. fragilis* shows a higher proportion (56.0%) of leaf SNPs (i.e., present in only one genotype) while other species show much lower proportions, including *P. goldsteinii* (20.5%), *B. stercoris* (22.4%) and *B. xylanisolvens* (25.6%), which suggests differential population bottlenecks and selective sweeps at the species level. At the gene level, we also observed evidence of convergent adaptive evolution. For instance, between different *P. dorei* isolate lineages, we identified two coding variants in gene *TodS* (Figure S12), which encodes a two-component kinase sensor regulating toluene metabolism in bacteria[36]. Toluene and other aromatic

hydrocarbons are found in foods and are also used as industrial feedstocks that could contaminate foods and thus drive evolution in the gut[37].

Another major driver of within-person gut microbiome evolution is horizontal gene transfer (HGT). Accordingly, we used all whole-genome sequenced H1 isolates to reconstruct a HGT network of shared DNA elements >2 kb in length (Methods). Consistent with recent reports[38, 39], we observed that HGT events were strongly linked to the phylogeny of the isolates, i.e., most HGT events occurred within the same phyla but were also quite prevalent across different families and between distinct species (Figure 5C; Table S10). Interestingly, we observed that HGT were predominantly enriched between isolates with the same gram-staining, with gram-negative species showing more prevalent HGTs than gram-positive species ($p = 0.0005$ by Pearson's Chi-squared test). This result is consistent with recent finding[39] and suggests different cell wall structures may play an important role in HGTs. Notably, HGTs between gram-positive and negative species were also observed in our dataset, inspiring future studies to study the effect of cell wall structures on HGTs and engineer these HGT elements into a microbiome editing tool. Next, to examine whether these HGTs occurred recently, we calculated the mean HGT frequency between all species pairs (Methods). We hypothesized that if HGTs occurred recently between two species, they would be only associated with a small proportion of isolates, resulting in a low frequency between species, while if HGTs occurred earlier and provided growth benefits, they would be enriched and vertically inherited by later generation, resulting in a high frequency. Interestingly, we found most HGT elements were frequently present across isolates (71.5% HGTs with >50% frequency), especially for ones within *Bacteroidaceae* species (Figure 5C), suggesting that they occurred in the distant past and were enriched under strong selection within the gut environment.

Given the high prevalence and frequency of within-individual HGT, we next annotated the protein-coding sequences of most widespread HGT elements to probe their potential functions (Methods). Interestingly, we identified multiple antibiotics resistance genes (ARGs) with different mechanisms of action as well as secretion system genes (Figure S13). For example, the top four most widespread HGT sequences are found surprisingly in at least 13 different species of *Bacteroidaceae*, *Porphyromonadaceae*, *Odoribacteraceae* and *Rikenellaceae*, and contained multiple ARGs including ribosomal protectors and antibiotic efflux pumps, as well as Type-III and Type-IV secretion systems. While ARGs and secretion systems shared through HGT may confer clear evolutionary advantages[40, 41], there were numerous widespread elements across different species with genes of unknown function (Figure S13), hinting at unexplored mechanisms that drive their long-term persistence in the gut. Taken together, these results highlight that isolates within and across people have genomic diversity that can be systematically characterized using CAMII-enabled deep strain biobanking and genomic analysis to study person-specific gut microbiome colonization, adaptation, and ecology.

DISCUSSION

Strain isolation from the gut microbiome has historically been performed in an *ad hoc* manner where important phenotypic features are inadequately captured and poorly

documented alongside genomic data. Here, we described the CAMII platform to industrialize the generation of isolate biobanks by leveraging automation, machine vision, supervised learning, and genomics. When combined with low-cost 16S and whole-genome sequencing, the systematically generated phenotypic and genomic data produced from the pipeline forms a rich resource to study microbial colony morphology, diversity, and evolution. Using the gut microbiome as a showcase example, CAMII-enabled isolation yielded extensive isolate biobanks from 20 healthy individuals that in aggregate covered >80% of all microbiota by abundance present. This isolate collection covers a majority of microbial diversity in the healthy gut and is one of the most extensive personalized isolate biobanks described to date. Using this resource, we demonstrated that quantitative analysis of colony morphologies can predictive taxonomy, enhance the isolation of targeted genera, and reveal potential interactions between microbes. Systematic analysis of genomic differences between isolates within and across people revealed interesting patterns of population selection, adaptation, and horizontal gene transfer.

The majority of the data presented here relied on a common mGAM rich media for strain isolation and characterization in the context of the human gut microbiome. Exploration of alternative media formulations, other micro- and macro-nutrients, and host or environmentally associated biochemical perturbations (e.g., bile acids, xenobiotic compounds) could yield morphologic and growth profile changes that inform unexplored physiologies and characteristics of the gut microbiome. The interspecies interactions derived from CAMII datasets could be further utilized to systematically map out the drivers of microbiome dynamics. We envision that these interactions could facilitate cultivation of recalcitrant “dark matter” microbiome by helping to identify unknown microbially-derived molecules that promote cooperative growth observed in this and other studies.

The CAMII system uses commercially available off-the-shelf components and open-source code that can be readily replicated by other researchers (see Table S1 for list of components). We envision that the searchable online portal will facilitate sharing of standardized phenotypic and genomic data, which is poised to grow over time. The CAMII hardware could be further expanded to integrate mass spectrometry measurements to gain additional colony characteristic profiles that can improve species and metabolite identification. Onboard automated microscopy could further introduce orthogonal data streams to visualize microbial cells at micrometer resolution across different spectral channels. Improved machine vision and ML algorithms could yield even better strain predictions and enhance isolation performance.

Since individual strains are the unit of action within a complex community, more complete strain collections are needed. Such comprehensive biobanks can be used to recreate a more holistic context that take into account the composition, interspecies interactions, and metabolic capacity of the entire community, which will improve studies of microbiome function, dynamics, and stability. Beyond the human gut, CAMII can be useful for other microbiomes such as those from soil, aquatic or agricultural settings, including further isolation and analysis of phages, fungi and protozoa. The robotic automation system can also help generate systematic strain libraries such as arrayed transposon insertion knock-out

collections[42] or functional genomics expression libraries[43] as well as improve screening for tractable microbial chassis for genetic engineering[44].

Methods

Ethical review.

This study was approved and conducted under Columbia University Medical Center Institutional Review Board protocol AAAR0753. Written informed consent was obtained from subjects in the study.

Fecal sample collection and storage.

Fresh fecal samples were collected from 20 healthy human donors and processed within 3 hours of defecation. Briefly, feces were collected using the Commode Specimen Collection System (Fisher 02-544-208). An inverted sterile 200-μl pipette tip (Rainin RT-L200F) was used to core out a small sample from the stool specimen, which was then immediately placed in a sterile cryovial (Fisher NC9347001). The collected fecal samples were then transferred to an anaerobic chamber (Coy Laboratory) and homogenized in 5mL of pre-reduced Phosphate Buffered Saline (PBS) by thorough vortexing. Homogenized samples were further passed through a 40-micron filter (Fisher 22363547) to remove dietary debris, aliquoted into multiple cryovials with glycerol (20% final concentration) and transferred to a −80 °C freezer for long-term storage.

Plate preparation and bacterial culture.

All gut microbiota were grown in Gifu Anaerobic Medium Broth, Modified (mGAM) (HyServe 05433) under anaerobic conditions (5% H_2, 10% CO_2, 85% N_2) in an anaerobic chamber. Briefly, 1.5% agar plates (Thermo 242811) with mGAM media were made using a peristaltic pump (New Era Pump Systems NE-9000) and labelled with unique barcodes. For plates supplemented with Ciprofloxacin (10μg/mL), Trimethoprim (50μg/mL) or Vancomycin (50μg/mL), antibiotics were added during plate preparation. All plates were then transferred to the anaerobic chamber and pre-reduced for ~24 hours prior to plating. Frozen fecal samples were thawed in the anaerobic chamber and diluted to 10^3 CFU/mL for each culturing conditions. Optimal dilutions were determined by sample-specific serial dilution experiments. 200μL of diluted fecal samples was then dispensed onto the plate and spread using sterile glass beads. Plates were sealed in Ziploc bags to reduce desiccation and incubated at 37°C for 5-day of colony growth.

Strain imaging and isolation.

Strain imaging and isolation was performed using a custom automated imaging and colony-picking system (CAMII). After 5 days of growth, agar plates were imaged automatically on the CAMII system (Figure 1C). Briefly, plates were first placed on a carousel stacker. A robotic arm gripper carried individual plates past a barcode scanner to an illuminated imaging platform on the colony picker where they were imaged under two lighting conditions (epi-illumination and trans-illumination) by the Hudson RapidPick control software. The plate labels are linked to the captured images and imaged plates were automatically re-stacked by the robotic arm. Following completion of the imaging process,

plates were sealed in Ziploc bags to avoid desiccation and a custom script was used to segment different colonies and identify morphologically unique colonies for subsequent picking based on plate images (Figure S1A). Morphologic features include area, perimeter, mean radius, circularity, convexity, inertia, and mean and variances along grey channel (trans-illuminated images) and red-green-blue (RGB) channels (epi-illuminated images). Raw images of all colonies on the plate are also collected.

For *random picking* performed in this study, a random subset of a given number of colonies were generated from all detected colonies by the script and the automatic isolation was performed on these colonies. For *phenotype-guided picking*, all detected colonies were firstly subjected to optimized selection based on their morphology and a subset of a given number of colonies with maximized morphological diversity were isolated by CAMII. Detailed algorithm of optimized colony selection can be found in Figure S1B, and Scripts used to analyze plate images and colony morphologies can be accessed at https://github.com/hym0405/CAMII.

After analyzing plate images and generating a list of colonies to pick, a similar robotic protocol was executed to isolate these colonies. Firstly, plates were re-stacked for picking and a multichannel media dispenser was used to dispense 50μL of mGAM liquid media into each well of two barcoded sterile 384-well optical plates (Fisher 12-566-2) (duplicate 'A' and 'B'), which were then moved to the colony picker. Next, an agar plate was transferred to the colony picker and heat-sterilized needles picked individual colonies into the duplicate optical plates. Plates were automatically switched out when all targeted colonies were picked (agar plate) or all wells were inoculated (optical plate). After colony picking, inoculated optical plates were transferred to a plate sealer (Brandel 9795), sealed and re-stacked. Optical plates were then incubated at 37°C for ~5 days for bacteria culturing. After bacterial growth, 'A' plates were subjected to downstream genomic DNA extraction and 30uL of 40% glycerol was added to each well of 'B' plates, which were transferred to −80 °C for long-term storage.

Colony morphology analysis.

To achieve morphology-guide colony selection, colony morphological features extracted in raw image processing were centralized and scaled to unit variance and then embedded by principal component analysis (PCA). An optimized colony selection algorithm was further applied to embedded features to search a set of colonies with most morphological diversity (Figure S1B).

To evaluate how different ASVs respond to nearby colonies (Figure 4C), number of nearby colonies were calculated for isolates on plates and "nearby colony" pair was defined as two colonies with distance between their X-Y coordinates shorter than 30 pixels plus the sum of their radii. To avoid potential impact of antibiotics on colony morphology, only colonies grown on mGAM-only plates were used for morphology analysis.

To evaluate how the growth of specific genus could be affected by other genera (Figure 4D), we firstly identified "nearby colony" pairs as described above, and the growth impact of genus-A on genus-B is quantified by comparing the colony sizes of genus-B with genus-A

present in the neighborhood, i.e., as nearby colony, to the colony sizes of genus-B without any nearby colonies: the effect size was defined as the fold-change of average colony size with genus-A present to average colony size without any nearby colonies, and the P-values were calculated by Mann-Whitney U test on the size distributions between with genus-A present in the neighborhood and without any nearby colonies and FDR correction was performed using Bonferroni–Holm methods. To avoid potential impact of antibiotics on colony morphology, only colonies grown on mGAM-only plates were used for morphology analysis.

Taxonomy prediction and targeted isolation.

To test whether colony morphology on plates could help predict taxonomic identity, colonies of data-rich genus (>100 isolates across all 20 individuals) were subjected to model training and testing. Considering the potential impact of antibiotics perturbation and neighbor colonies, a multi-label random forest model was trained on 70% of isolates which was randomly sampled using 14 colony morphological features, antibiotics condition and number of nearby colonies, and the performance (precision and recall) of the model was evaluated on the remaining 30% of isolates. The procedure of model training and evaluating was bootstrapped 20 times with different randomization settings to minimize bias and background performance of the model were calculated by null model (prediction based on number of isolates). To perform targeted microbial isolation, a multi-label random forest model was trained on colonies of data-rich genus (>15 isolates from same individual) from individual H12, H13 and H14 separately as described above. The same fecal samples were then plated out and the model was applied to the new plates after bacterial growth to screen all colonies on plates and predict colonies of targeted genus-level taxonomy. All colonies of the plates were then isolated on CAMII and subjected to 16SV4 sequencing to identify their taxonomy and evaluate the performance of targeted isolation.

Daily kinetic growth analysis of colonies on plate.

To monitor the growth kinetics on a daily basis, fecal sample H1t5 was plated out on mGAM-only plates and the plates were imaged every day during 6 days of growth. Colony detection and segmentation was performed on images and colony morphology features on different days were matched based on their X-Y coordinates (Figure 4A). All colonies on plates were then isolated on CAMII at day-6 and subjected to taxonomy identification by 16S rRNA sequencing. To quantify differential initial growth of genera, the number of detectable colonies of each genus on each day (tracked by X-Y coordinates) were normalized to their total number of colonies at day-6 to calculate the proportion of detectable colonies (Figure 4B) at each day.

Genomic DNA extraction.

Genomic DNA (gDNA) of picked isolates were extracted in 384-well format using a silica bead beating-based protocol adapted from a prior study[1]. Firstly, 40μL 0.1mm Zirconia Silica beads (Biospec 11079101Z) and 120μL lysis solution (50mM Tris-HCl pH7.5 and 0.2mM EDTA) were added to each well of 384-well deep-well plates (Fisher 07-202-505). Next, 40μL culture solutions of isolates were added to each well and the plates were centrifuged for 1 minute at 4500g and affixed with a sealing mat (Axygen AM-384-DW-

SQ). To avoid overheating during bead beating, the plates were vortexed for 5 second and incubated at −20°C for 10 minutes prior to beating. Then, plates were fixed on a bead beater (Biospec 1001) and subjected to bead beating for 5 minutes, followed by a 10-minute cooling period. The bead beating cycle was repeated once and plates were centrifuged at 4500g for 5 minutes to spin down cell debris. Next, 10μL cell lysate was transferred to a 384-well PCR plate (Bio-Rad HSP3801) and 2μL proteinase K solution (50mM Tris-HCl pH7.5 and 1μg/μL proteinase K [Lucigen MPRK092]) was added using a Formulatrix Mantis. Finally, cell lysate was subjected to proteinase K digestion on a thermal cycler (65°C 30min, 95°C 30min, 4°C infinite) and transferred to −20°C for long-term storage. gDNA extraction for bulk feces samples were performed using the same protocol with scale-up reaction volumes in 96-well format.

16S rRNA amplicon sequencing.

16S sequencing of the V4 region for isolates taxonomy identification was performed in 384-well format using a set of dual-indexing sequencing primers. Briefly, barcoded 16SV4 amplicon primers were designed based on universal 16SV4 primers and synthesized by Integrated DNA Technologies. Next, 1μL of each unique combination of barcoded forward primer 16SV4f_5xx and reverse primer 16SV4r_7xx were transferred to a 384-well PCR plate using Labcyte Echo to make unique dual-indexed primer plates. Then, ~130nL of gDNA was transferred to a primer plate by a 384-well pin replicator (Scinomix SCI-6010.OS) and 2μL NEBNext Q5 PCR master mix (NEB M0543L) was added to each well using Formulatrix Mantis. The samples were then subjected to 16SV4 amplification on a thermal cycler (98°C 30s, 40 cycles: 98°C 10s, 55°C 20s, 65°C 60s; 65°C 5min; 4°C infinite). The resulting amplicon libraries were manually pooled and subjected to gel electrophoresis on E-Gel™ EX Agarose Gels, 2% (ThermoFisher G402002). Expected DNA bands (~390bp) were excised from gel and extracted by Wizard™ SV Gel and PCR Cleanup System (Promega A9282) following the manufacturer's instructions to remove PCR primers and adapter dimers. Gel-purified libraries were quantified by Qubit dsDNA HS assay (Thermo Q32851) and sequenced on Illumina MiSeq platform (reagent kits: v2 300-cycles, paired-end mode) at 8 pM loading concentration with 20% PhiX spike-in (Illumina FC-110-3001) along with custom sequencing primers spiked into Miseq reagent cartridge (6 μl of 100 μM stock; well 12: 16SV4_read1, well 13: 16SV4_index1, well 14: 16SV4_read2) following the manufacturer's instructions. Sequences of all primers used in library preparation and sequencing are provided in the Table S11. 16S-V4 sequencing of the bulk samples were performed using similar protocol with scale-up reaction volumes in 96-well format. Moreover, SYBR Green I (final concentration: 0.2x) (Thermo S7563) was added to the PCR reaction and a quantitative 16SV4 amplification was performed and stopped during exponential phase (typically 13-17 cycles) and the reaction was advanced to the final extension step.

16S rRNA amplicon analysis and ASV clustering.

Raw sequencing reads of 16SV4 amplicon were analyzed by USEARCH v11.0.667[2]. Specifically, paired-end reads were merged using "-fastq_mergepairs" mode with default setting. Merged reads were then subjected to quality filtering using "-fastq_filter" mode with the option "-fastq_maxee 1.0 -fastq_minlen 240" to only keep reads with less

than 1 expected error base and greater than 240bp. Remaining reads were deduplicated (-fastx_uniques) and clustered into ASVs (-unoise3) at 100% identity, and merged reads were then searched against ASV sequences (-otutab) to generate ASV count table. Taxonomy of ASVs were assigned using Ribosomal Database Project classifier v2.13 trained with 16S rRNA training set 18[3]. Relative abundance of ASVs in bulk samples are defined as reads count of ASVs normalized by total number of mapped reads.

Isolate taxonomy identification and 16S phylogeny analysis.

After ASV clustering, ASV count table were parsed to calculate the following metrics for each isolate: total reads count, the ASVs with the highest reads count and purity of that ASV. Isolates with insufficient reads or poor purity (reads counts < 5 or purity < 0.5) were filtered and the taxonomy of remaining isolates were defined as the ASVs with the highest reads count. To construct the phylogeny of isolates, multi-sequence alignment was performed on ASV sequences of the isolates using MUSCLE v5[4] and aligned ASV sequences was subsequently analyzed by MEGA v11.0.11[5] to calculate neighbor joining tree with default setting for phylogeny reconstruction.

Isolates whole-genome sequencing and reads processing.

The same gDNA used for 16SV4 amplicon sequencing was subjected to whole-genome sequencing for isolates. Paired-end libraries were constructed following a published protocol of low-volume Nextera library preparation[6] and sequenced on Illumina Nextseq 500/550 platform (2x75bp) and HiSeq platform (2×150bp). Raw reads were then processed by Cutadapt v2.1 with following parameters “--minimum-length 25:25 -u 10 -u -5 -U 10 -U -5 -q 15 --max-n 0 --pair-filter=any” to remove low-quality bases and Nextera adapters. Coverage was 1.42 ± 2.86 million paired-end reads per isolate and PacBio long-read sequencing was performed for some isolates by SNPsaurus to improve the performance of *de novo* genome assembling.

De novo genome assembling and SNP variation analysis.

Illumina reads passing quality filtering and PacBio long reads were assembled by Unicycler v0.4.4[7] with default setting to generate draft genomes of isolates, and the quality and species-level taxonomy of draft genomes were then assessed by QUAST v4.6.3[8], CheckM v1.0.13[9] and GTDB-Tk v0.2.2[10]. Among all 3,271 isolates assemblies, 1,197 were defined as high-quality draft genomes (Coverage > 20X; N50 > 5,000bp; Completeness > 80%; Contamination < 5%) and used for downstream genomic variation and HGT analysis.
To identify strain-level genomic variation of gut microbiota isolates within and between individuals, draft assemblies with highest Completeness and N50 of each species were selected as reference genome for reads alignment, and processed Illumina reads of isolates were aligned to reference genomes of same species by Bowtie2 v2.3.4[11] in paired-end mode with “--very-sensitive” setting. Resulting reads alignments were then processed by SAMtools v1.9 and BCFtools v1.9[12] with “--ploidy 1” setting to call genomic variation (SNPs and Indels). Resulting variations was then subjected to quality filtering to identify “reliable” genotypes (covered by ≥ 5 reads; with ≥ 0.9 haploidy) and only SNP variations with more than 90% “reliable” genotypes across all isolates were used for downstream analysis. To construct SNP-based phylogeny, base profiles of isolates at SNP sites were

concatenated together and UPGMA tree was then calculated by MEGA v11.0.11 with default setting.

Genome-wide average nucleotide identity (ANI) calculation.

To identify species isolated in our biobank that not been cultivated previously, the average nucleotide identity between draft genomes obtained in this study and MAGs or isolates genomes from public available databases were calculated by FastANI v1.0[13], and genomes with >95% ANI were considered to be the same species.

Horizontal gene transfer identification and annotation.

To identity HGT occurring between species within H1 isolates, we compared all genomes pairs of different species by BLASTN v2.7.1[14] with "-evalue 0.1 -perc_identity 99" setting to systematically screen blocks of genomic regions with high sequence identity. The P-value of candidate HGTs were then calculated based on the genome-wide ANI between isolates and further adjusted by Benjamini-Hochberg procedure. Blast hits with adjusted P-value < 1e-5 and larger than 2000bp in length were considered as HGT event between isolates of different species. The frequency of HGTs between species were quantified using a previously published method[15, 16], defined as the number of between-species genome pairs that share at least one HGT divided by the total number of between-species genome pairs. To annotate antibiotics resistance genes (ARGs) and secretion systems in HGT elements, sequences of HGT elements were annotated by Prokka v1.12[17] in metagenome mode and resulting CDS were searched against CARD database v3.1.4[18] by BLASTP v2.7.1 to identify ARG hits with e-value < 1e-5, identity > 20 and query coverage > 50. Secretion systems were also predicted on CDS of HGT elements by EffectiveDB[19] with default setting.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank members of the Wang laboratory for advice and comments on the manuscript. H.H.W. acknowledges relevant funding support from the NSF (MCB-2025515), NIH (1R01AI132403, 1R01DK118044, 1R21AI146817), ONR (N00014-18-1-2237, N00014-17-1-2353), Burroughs Wellcome Fund (1016691), Irma T. Hirschl Trust, and Schaefer Research Award. R.U.S. was supported by a Fannie and John Hertz Foundation Fellowship and an NSF Graduate Research Fellowship (DGE-1644869). T.M. is supported by NIH Medical Scientist Training Program (T32GM007367). M.R. and F.V.C. were supported by NSF Graduate Research Fellowships (DGE-1644869). C.R. was supported by a Junior Fellows Scholarship from the Simons Society of Fellows.

Data Availability

The sequencing data generated in this study have been submitted to the NCBI BioProject database (http://www.ncbi.nlm.nih.gov/bioproject/) under accession number PRJNA745993. Other associated data of the isolate collection, including morphological features and raw images, can be accessed at http://microbial-culturomics.com. Taxonomy of ASVs were assigned based on 16S rRNA training set 18 provided by Ribosomal Database Project. The

annotation of ARG genes and secretion systems in HGT elements was based on CARD database v3.1.4 and EffectiveDB database respectively.

References

1. Poyet M et al. A library of human gut bacterial isolates paired with longitudinal multiomics data enables mechanistic microbiome research. Nat Med 25, 1442–1452 (2019). [PubMed: 31477907]
2. Zhao SJ et al. Adaptive Evolution within Gut Microbiomes of Healthy People. Cell Host Microbe 25, 656-+ (2019). [PubMed: 31028005]
3. Zou YQ et al. 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. Nat Biotechnol 37, 179-+ (2019). [PubMed: 30718868]
4. Browne HP et al. Culturing of 'unculturable' human microbiota reveals novel taxa and extensive sporulation. Nature 533, 543-+ (2016). [PubMed: 27144353]
5. Goodman AL et al. Extensive personal human gut microbiota culture collections characterized and manipulated in gnotobiotic mice. P Natl Acad Sci USA 108, 6252–6257 (2011).
6. Villa MM et al. Interindividual Variation in Dietary Carbohydrate Metabolism by Gut Bacteria Revealed with Droplet Microfluidic Culture. Msystems 5 (2020).
7. Watterson WJ et al. Droplet-based high-throughput cultivation for accurate screening of antibiotic resistant gut microbes. Elife 9 (2020).
8. Ji BW, Sheth RU, Dixit PD, Tchourine K & Vitkup D Macroecological dynamics of gut microbiota. Nature Microbiology 5, 768-+ (2020).
9. Qamer S, Sandoe JAT & Kerr KG Use of colony morphology to distinguish different enterococcal strains and species in mixed culture from clinical specimens. Journal of Clinical Microbiology 41, 2644–2646 (2003). [PubMed: 12791893]
10. Allegrucci M & Sauer K Characterization of colony morphology variants isolated from Streptococcus pneumoniae biofilms. Journal of Bacteriology 189, 2030–2038 (2007). [PubMed: 17189375]
11. Cabeen MT, Leiman SA & Losick R Colony-morphology screening uncovers a role for the Pseudomonas aeruginosa nitrogen-related phosphotransferase system in biofilm formation. Molecular Microbiology 99, 557–570 (2016). [PubMed: 26483285]
12. Martin-Rodriguez AJ et al. Regulation of colony morphology and biofilm formation in Shewanella algae. Microbial Biotechnology 14, 1183–1200 (2021). [PubMed: 33764668]
13. Rettedal EA, Gumpert H & Sommer MOA Cultivation-based multiplex phenotyping of human gut microbiota allows targeted recovery of previously uncultured bacteria. Nature Communications 5 (2014).
14. Strittmatter N et al. Analysis of intact bacteria using rapid evaporative ionisation mass spectrometry. Chemical Communications 49, 6188–6190 (2013). [PubMed: 23736664]
15. Fang JS & Dorrestein PC Emerging mass spectrometry techniques for the direct analysis of microbial colonies. Current Opinion in Microbiology 19, 120–129 (2014). [PubMed: 25064218]
16. Isozaki A et al. A practical guide to intelligent image-activated cell sorting. Nature Protocols 14, 2370–2415 (2019). [PubMed: 31278398]
17. Hosny A, Parmar C, Quackenbush J, Schwartz LH & Aerts HJWL Artificial intelligence in radiology. Nature Reviews Cancer 18, 500–510 (2018). [PubMed: 29777175]
18. Cole JA, Kohler L, Hedhli J & Luthey-Schulten Z Spatially-resolved metabolic cooperativity within dense bacterial colonies. Bmc Systems Biology 9 (2015).
19. Chacon JM, Mobius W & Harcombe WR The spatial and metabolic basis of colony size variation. Isme Journal 12, 669–680 (2018). [PubMed: 29367665]
20. Ratzke C & Gore J Modifying and reacting to the environmental pH can drive bacterial interactions. Plos Biology 16 (2018).
21. Edgar RC Updating the 97% identity threshold for 16S ribosomal RNA OTUs. Bioinformatics 34, 2371–2375 (2018). [PubMed: 29506021]
22. Lloyd-Price J et al. Strains, functions and dynamics in the expanded Human Microbiome Project. Nature 550, 61-+ (2017). [PubMed: 28953883]

23. Lozupone CA, Stombaugh JI, Gordon JI, Jansson JK & Knight R Diversity, stability and resilience of the human gut microbiota. Nature 489, 220–230 (2012). [PubMed: 22972295]
24. Almeida A et al. A new genomic blueprint of the human gut microbiota. Nature 568, 499-+ (2019). [PubMed: 30745586]
25. Pasolli E et al. Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. Cell 176, 649-+ (2019). [PubMed: 30661755]
26. Franzosa EA et al. Gut microbiome structure and metabolic activity in inflammatory bowel disease. Nature Microbiology 4, 293–305 (2019).
27. Huang L & Wu T Novel neural network application for bacterial colony classification. Theoretical Biology and Medical Modelling 15 (2018).
28. Qu KY, Guo F, Liu XR, Lin Y & Zou Q Application of Machine Learning in Microbiology. Frontiers in Microbiology 10 (2019).
29. Wang HD et al. Early detection and classification of live bacteria using time-lapse coherent imaging and deep learning. Light-Science & Applications 9 (2020).
30. Venturelli OS et al. Deciphering microbial interactions in synthetic human gut microbiome communities. Molecular Systems Biology 14 (2018).
31. Kim H, Jeong Y, Kang SN, You HJ & Ji GE Co-Culture with Bifidobacterium catenulatum Improves the Growth, Gut Colonization, and Butyrate Production of Faecalibacterium prausnitzii: In Vitro and In Vivo Studies. Microorganisms 8 (2020).
32. Lindstad LJ et al. Human Gut Faecalibacterium prausnitzii Deploys a Highly Efficient Conserved System To Cross-Feed on beta-Mannan-Derived Oligosaccharides. Mbio 12 (2021).
33. Louis P, Young P, Holtrop G & Flint HJ Diversity of human colonic butyrate-producing bacteria revealed by analysis of the butyryl-CoA:acetate CoA-transferase gene. Environmental Microbiology 12, 304–314 (2010). [PubMed: 19807780]
34. Tierney BT et al. The Landscape of Genetic Content in the Gut and Oral Human Microbiome. Cell Host Microbe 26, 283-+ (2019). [PubMed: 31415755]
35. Chen L et al. The long-term genetic stability and individual specificity of the human gut microbiome. Cell 184, 2302–2315 e2312 (2021). [PubMed: 33838112]
36. Lau PCK et al. A bacterial basic region leucine zipper histidine kinase regulating toluene degradation. P Natl Acad Sci USA 94, 1453–1458 (1997).
37. Defois C et al. Food Chemicals Disrupt Human Gut Microbiota Activity And Impact Intestinal Homeostasis As Revealed By In Vitro Systems. Scientific Reports 8 (2018).
38. Jeong H, Arif B, Caetano-Anolles G, Kim KM & Nasir A Horizontal gene transfer in human-associated microorganisms inferred by phylogenetic reconstruction and reconciliation. Scientific Reports 9 (2019).
39. Groussin M et al. Elevated rates of horizontal gene transfer in the industrialized human microbiome. Cell 184, 2053-+ (2021). [PubMed: 33794144]
40. Juhas M, Crook DW & Hood DW Type IV secretion systems: tools of bacterial horizontal gene transfer and virulence. Cell Microbiol 10, 2377–2386 (2008). [PubMed: 18549454]
41. Woods LC et al. Horizontal gene transfer potentiates adaptation by reducing selective constraints on the spread of genetic variation. Proc Natl Acad Sci U S A 117, 26868–26875 (2020). [PubMed: 33055207]
42. Goryshin IY, Jendrisak J, Hoffman LM, Meis R & Reznikoff WS Insertional transposon mutagenesis by electroporation of released Tn5 transposition complexes. Nat Biotechnol 18, 97–100 (2000). [PubMed: 10625401]
43. Mutalik VK et al. Dual-barcoded shotgun expression library sequencing for high-throughput characterization of functional traits in bacteria. Nat Commun 10, 308 (2019). [PubMed: 30659179]
44. Ronda C, Chen SP, Cabral V, Yaung SJ & Wang HH Metagenomic engineering of the mammalian gut microbiome in situ. Nat Methods 16, 167–170 (2019). [PubMed: 30643213]

Methods-only References

1. Ji BW et al. Quantifying spatiotemporal variability and noise in absolute microbiota abundances using replicate sampling. Nat Methods 16, 731–736 (2019). [PubMed: 31308552]
2. Edgar RC Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26, 2460–2461 (2010). [PubMed: 20709691]
3. Wang Q, Garrity GM, Tiedje JM & Cole JR Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol 73, 5261–5267 (2007). [PubMed: 17586664]
4. Edgar RC MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32, 1792–1797 (2004). [PubMed: 15034147]
5. Kumar S, Stecher G, Li M, Knyaz C & Tamura K MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Mol Biol Evol 35, 1547–1549 (2018). [PubMed: 29722887]
6. Baym M et al. Inexpensive multiplexed library preparation for megabase-sized genomes. PLoS One 10, e0128036 (2015). [PubMed: 26000737]
7. Wick RR, Judd LM, Gorrie CL & Holt KE Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol 13, e1005595 (2017). [PubMed: 28594827]
8. Gurevich A, Saveliev V, Vyahhi N & Tesler G QUAST: quality assessment tool for genome assemblies. Bioinformatics 29, 1072–1075 (2013). [PubMed: 23422339]
9. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P & Tyson GW CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res 25, 1043–1055 (2015). [PubMed: 25977477]
10. Chaumeil PA, Mussig AJ, Hugenholtz P & Parks DH GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics (2019).
11. Langmead B & Salzberg SL Fast gapped-read alignment with Bowtie 2. Nat Methods 9, 357–359 (2012). [PubMed: 22388286]
12. Li H et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079 (2009). [PubMed: 19505943]
13. Jain C, Rodriguez RL, Phillippy AM, Konstantinidis KT & Aluru S High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun 9, 5114 (2018). [PubMed: 30504855]
14. Camacho C et al. BLAST+: architecture and applications. BMC Bioinformatics 10, 421 (2009). [PubMed: 20003500]
15. Groussin M et al. Elevated rates of horizontal gene transfer in the industrialized human microbiome. Cell 184, 2053–2067 e2018 (2021). [PubMed: 33794144]
16. Smillie CS et al. Ecology drives a global network of gene exchange connecting the human microbiome. Nature 480, 241–244 (2011). [PubMed: 22037308]
17. Seemann T Prokka: rapid prokaryotic genome annotation. Bioinformatics 30, 2068–2069 (2014). [PubMed: 24642063]
18. Alcock BP et al. CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. Nucleic Acids Res 48, D517–D525 (2020). [PubMed: 31665441]
19. Eichinger V et al. EffectiveDB--updates and novel features for a better annotation of bacterial secreted proteins and Type III, IV, VI secretion systems. Nucleic Acids Res 44, D669–674 (2016). [PubMed: 26590402]

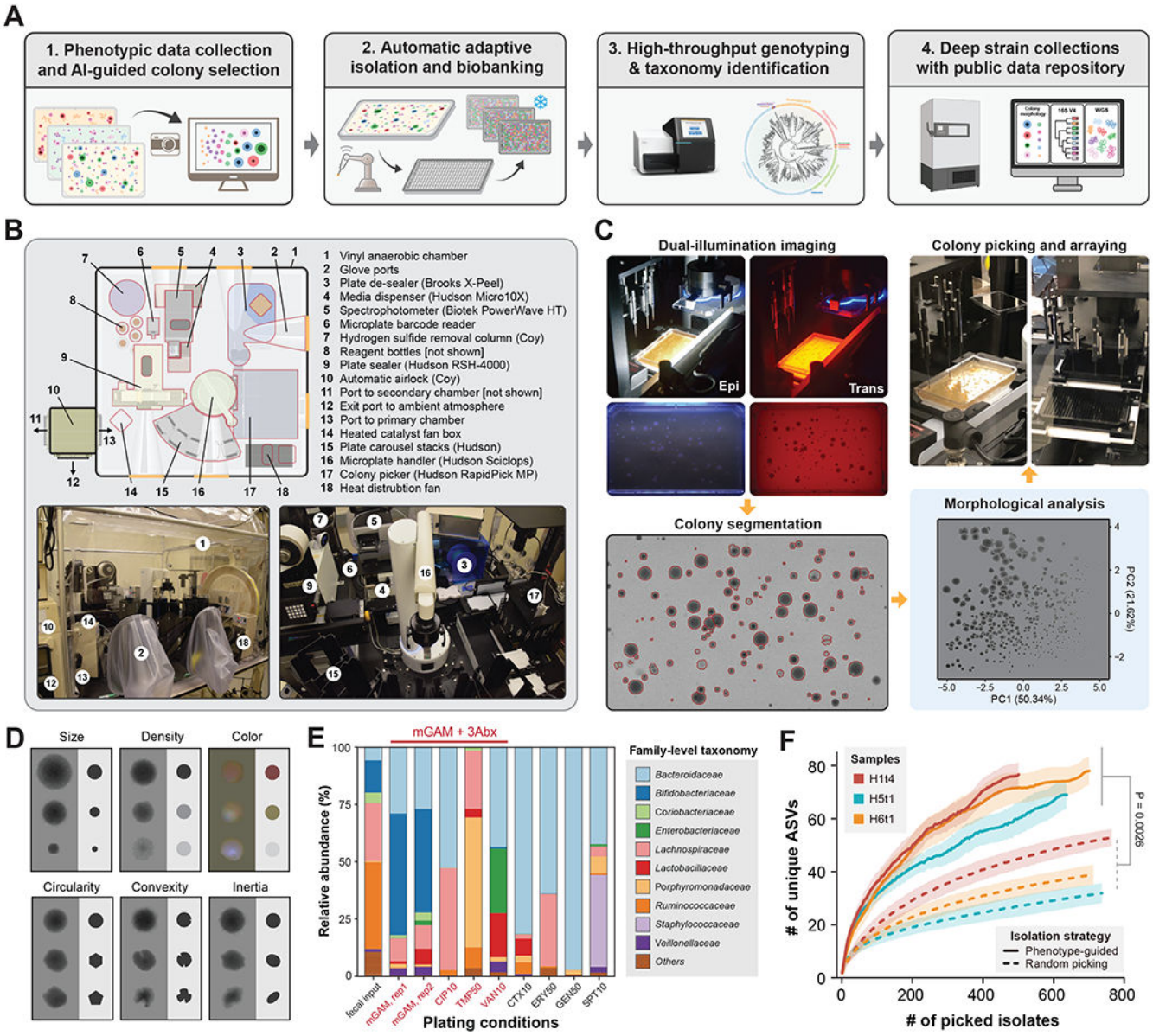

Figure 1. A data-driven microbial isolation strategy using phenotypic and morphologic features.

(A) Framework of phenotype and morphology-driven strain isolation and data collection of the human gut microbiome. Human fecal samples were plated and cultured under different antibiotics selection and morphologically diverse colonies were then isolated, biobanked and analyzed by downstream sequencing.

(B) Setup of the automated anaerobic microbial isolation and cultivation system CAMII.

(C) Illustration of morphology-guided colony isolation on CAMII. Colonies grown on plates are imaged under trans- and epi-illumination and subjected to contour segmentation and morphologic features extraction. Data are analyzed by PCA to identify the set of most morphologically diverse colonies that are then isolated by an integrated colony picker.

(D) Illustration of diverse colony morphology on plates. Colony size and shape features were extracted from trans-illuminated images and colony color features were extracted from epi-illuminated images.

(E) Fecal sample H1t1 were cultured with seven different antibiotics to evaluate their capacity to yield the most unique and diverse bacteria by 16S analysis at the family level. Ciprofloxacin, Trimethoprim and Vancomycin were selected for subsequent colony isolations.

(F) Number of unique ASVs obtained from phenotype-guided isolation compared to random isolation of three human fecal samples H1t4, H5t1 and H6t1. Isolation was performed by CAMII; Random isolation was performed on a random subset of all detected colonies on the plates and phenotype-guided isolation was performed on morphology-selected colonies by the algorithm (Figure S1B). P-value is calculated by a two-sided paired t-test on area under the curve. Ribbons on the curves represent the standard deviations of the number of obtained unique ASVs by the algorithm.

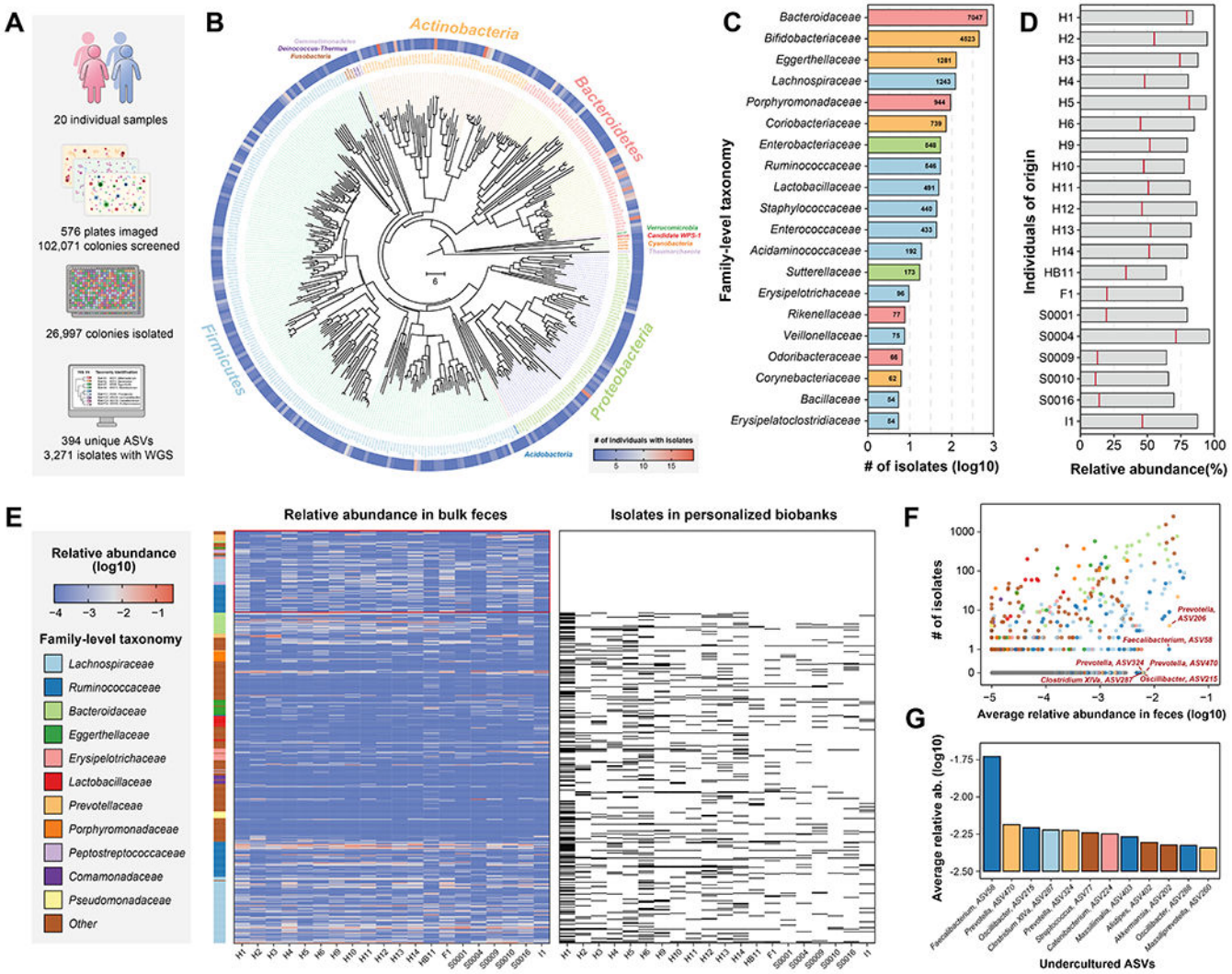

Figure 2. Generation of personalized gut isolate biobanks for 20 individuals.

(A) Statistics of 20 personalized gut isolate biobanks.

(B) Phylogenetic tree of 394 ASVs covered by 26,997 gut microbiome isolates in this study. Neighbor-joining tree of phylogeny was constructed based on 16S V4 sequences. Branch color distinguishes bacterial phylum, and the outer circle shows the prevalence of isolated ASVs in the 20 biobanks.

(C) Number of isolates for top 20 family-level taxonomy.

(D) Accumulated relative abundance of the ASVs represented by isolates from personalized biobanks in original fecal samples. The bars show isolates from any individual in the entire collection and the red lines show isolates derived from the same person.

(E) Heatmaps for relative abundance of abundant ASVs in original fecal samples and presence or absence in the biobanks. ASVs with average relative abundance > 0.1% are shown and the side bar on the left represents their family-level taxonomy. ASVs found in personalized biobanks are shown as black bar in the right heatmap and uncultured ASVs not found in any biobank are highlighted.

(F) Correlation of average relative abundance in original feces sample and number of isolates in entire collection for ASVs. Highly abundant ASVs that are difficult-to-culture, i.e., with fewer isolates, are highlighted.

(G) Average relative abundance of top abundant ASVs but with no more than 2 isolates in the entire collection. Color of bars represents family-level taxonomy.

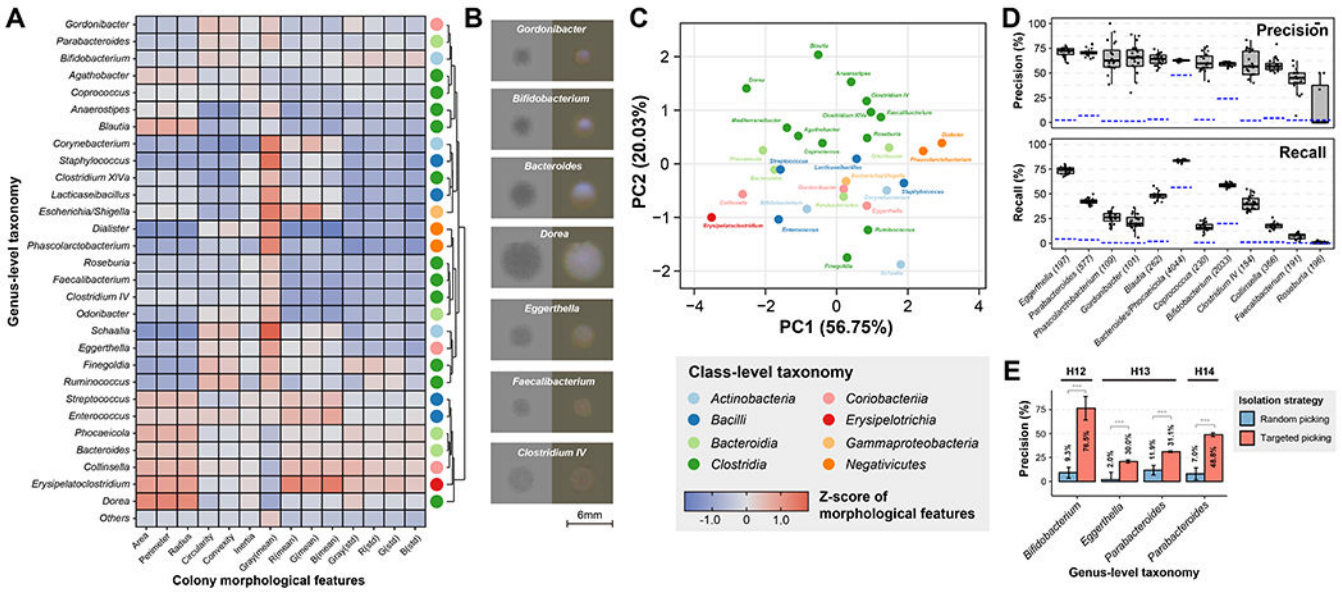

Figure 3. Using colony morphology to predict taxonomic identity enhances targeted isolation.

(A) Heatmap of average Z-scores of morphological features across different bacterial genera. Different genera exhibiting diverse morphological patterns were classified into different groups by hierarchical clustering and the colored dot on the right represents their class-level taxonomy.

(B) Examples of colony images. Trans-illuminated images are on the left side and epi-illuminated images are on the right side.

(C) PCA ordination of genera based on their colony morphological features. Colors indicate class-level taxonomy.

(D) Performance of bacterial genus prediction based on morphological features by a random forest classifer. The numbers in brackets represent the number of isolates for each genus. Model training and evaluation was bootstrapped 20 times and the box plots show the variance of performance (N = 20). Blue line represents the performance of null model. Definition of box-plot elements: center line: median; box limits: upper and lower 25th quartiles; whiskers: 1.5x interquartile range.

(E) Performance of model-based targeted isolation. Bars represent the mean of prediction precision by individual-specific models that were bootstrapped 20 times and error bars represent the standard deviations. P-values were calculated by two-sided Student's t-test on precisions from N = 20 randomly initialized model bootstrapping.

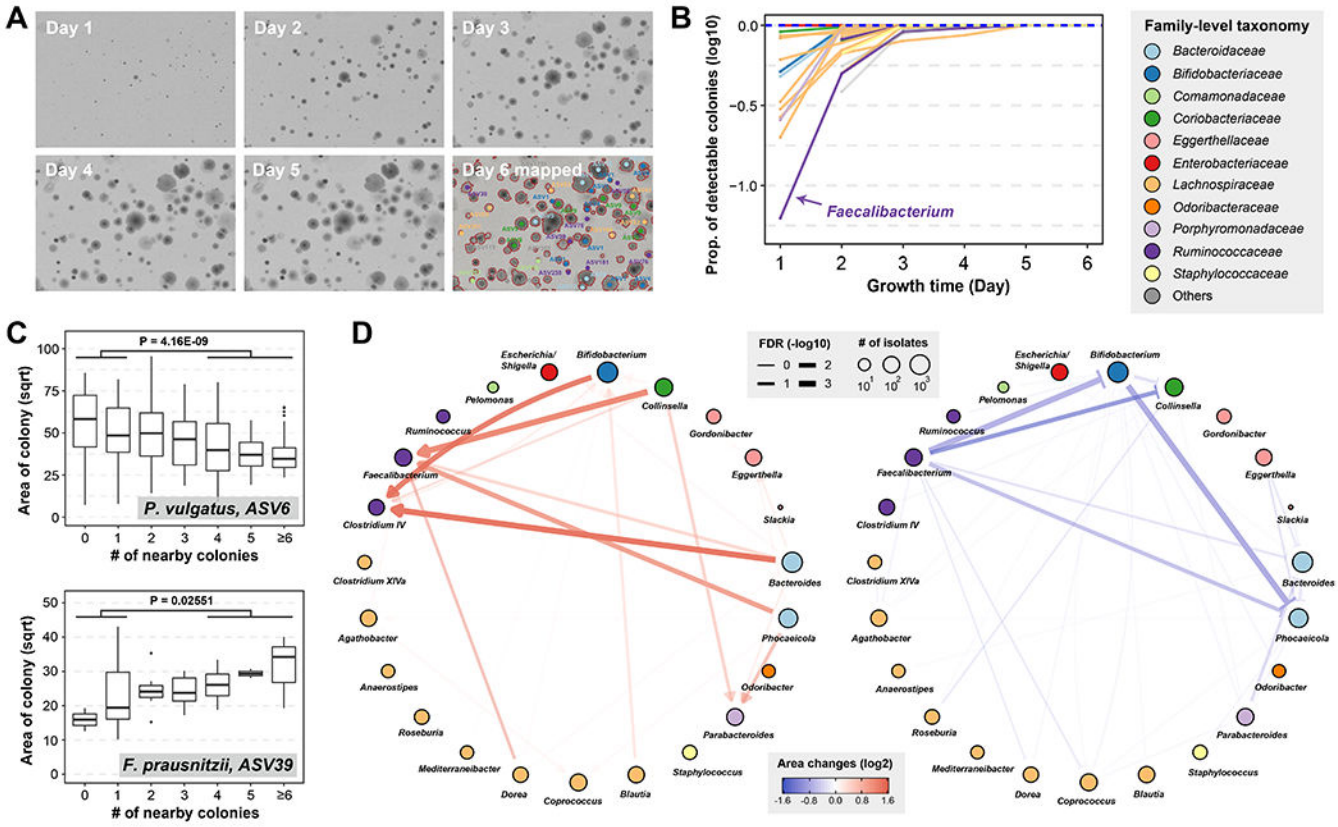

Figure 4. Mapping interaction between gut microbiota by colony morphology analysis.

(A) Images of an example plate during 6 days of growth and colony identities on the plate by 16S sequencing.

(B) Proportion of detectable colonies at different time points compared to day-6 for each genus. Colors indicate the family-level taxonomy.

(C) Correlation of colony size and number of nearby colonies for two representative ASVs. A full list of correlations is provided in Table S8. P-values are calculated by one-sided Mann–Whitney U test on area of colonies with no more than 1 nearby colonies or no less than 4 nearby colonies (N = 101 vs. 82 for ASV-6 and 17 vs. 9 for ASV-39). Definition of box-plot elements: center line: median; box limits: upper and lower 25th quartiles; whiskers: 1.5x interquartile range.

(D) Pairwise growth promoting and inhibiting networks between genera. Directional growth promoting effects are shown in red sharp arrow and directional growth inhibiting effects are shown in blue blunt arrows. Nodes represent bacterial genera and are colored by family. Node sizes are proportional to the number of isolates used in this analysis and edges widths are proportional to the significance of the interactions.

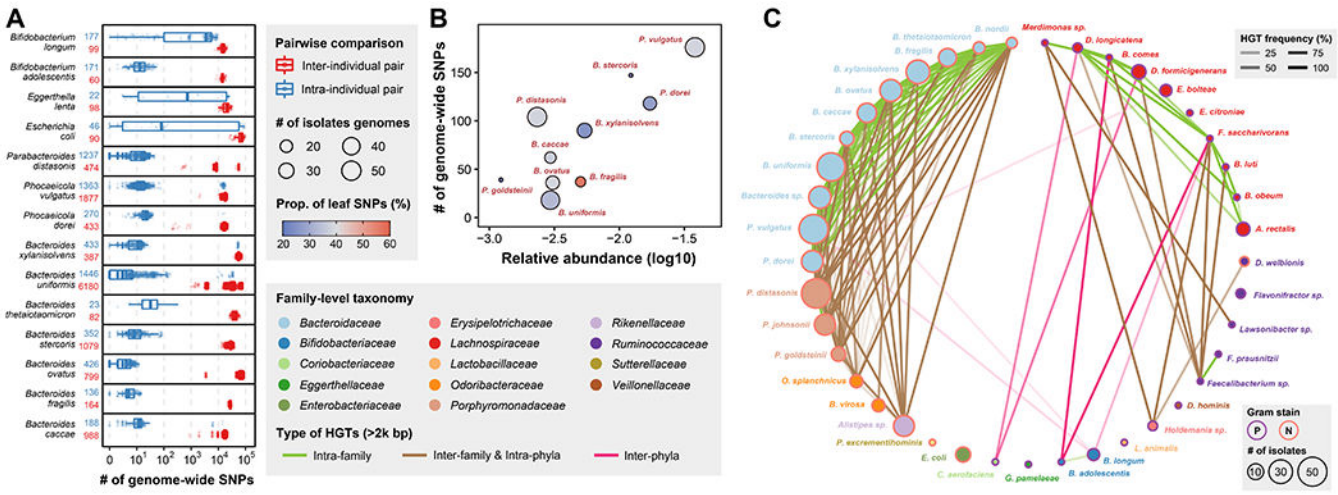

Figure 5. Strain-level genomic diversity of the gut microbiome within and between individuals.

(A) Number of genome-wide SNPs between isolates from the same or different individuals for 14 isolates-rich species. The numbers after the species name represent the number of inter-individual (red) and intra-individual (blue) pairs. Definition of box-plot elements: center line: median; box limits: upper and lower 25th quartiles; whiskers: 1.5x interquartile range.

(B) Correlation between number of genome-wide SNPs and relative abundance in original fecal sample for isolates-rich species in individual H1. The size of dot represents number of isolates used in this analysis and the color represents proportion of SNPs present in only 1 genotype.

(C) Network of 2kb+ HGT frequency based on 409 isolates from H1. Nodes represent bacterial species and are colored by family. Node sizes are proportional to the number of isolates in the H1 collection and edges opacity are proportional to the HGT frequency between the two connected species. Edge color represents different types of HGT, i.e., inter-phyla, intra-phyla & inter-family, or intra-family HGT and node outline color represents gram staining of the species.