Accurate estimation of intraspecific microbial gene content variation in metagenomic data with MIDAS v3 and StrainPGC

Byron J. Smith[1], Chunyu Zhao[2], Veronika Dubinkina[1], Xiaofan Jin[1], Jacqueline Moltzau-Anderson[2], Katherine S. Pollard[1,3,4]

[1] Gladstone Institute for Data Science and Biotechnology, San Francisco, USA

[2] Department of Gastroenterology, University of California, San Francisco, USA

[3] Chan Zuckerberg Biohub San Francisco, San Francisco, USA

[4] Department of Epidemiology and Biostatistics, University of California, San Francisco, USA

ORCID:

- BJS: 0000-0002-0182-404X
- CZ: 0000-0001-9589-2416
- VD: 0000-0002-4844-6795
- XJ: 0000-0002-0802-7692
- JA: 0000-0003-1398-5980
- KP: 0000-0002-9870-6196

Running Title: Microbial gene content with MIDAS and StrainPGC

Abstract:

Metagenomics has greatly expanded our understanding of the gut microbiome by revealing vast diversity within and across human hosts. Even within a single species, different strains can have highly divergent gene content, affecting traits such as antibiotic resistance, metabolism, and virulence. Methods that harness metagenomic data to resolve strain-level differences in functional potential are crucial for understanding the causes and consequences of this intraspecific diversity. The enormous size of pangenome references, strain mixing within samples, and inconsistent sequencing depth present challenges for existing tools that analyze samples one at a time. To address this gap, we updated the MIDAS pangenome profiler and developed StrainPGC, an approach to strain-specific gene content estimation that combines strain tracking and correlations across multiple samples. We validate our tool using a synthetic community and find that it outperforms existing approaches. Analyzing a large, publicly available metagenome collection from inflammatory bowel disease patients and healthy controls, we catalog the functional repertoires of thousands of strains across hundreds of species, capturing extensive diversity missing from reference databases. Finally, we apply StrainPGC to metagenomes from a clinical trial of fecal microbiota transplantation for the treatment of ulcerative colitis. We identify two *Escherichia coli* strains from two different donors that are both frequently transmitted to patients, but have notable differences in functional potential. StrainPGC and MIDAS v3 together enable precise, intraspecific pangenomic investigations using large collections of metagenomic data without microbial isolation or de novo assembly.

Introduction

In both diseased and healthy individuals, distinct strains of the same microbial species can differ in medically relevant traits, including metabolic capacity [@Joglekar2018], immunological interactions [@Yang2020;@Carrow2020], antimicrobial resistance [@Ray2017], and pathogenic potential [@Pakbin2021]. Evaluating the functional potential encoded in each genome is the first step in predicting strain-specific impacts on human health, and methods that accurately determine gene content from metagenomic data can greatly improve our understanding of the extent and importance of this intraspecific diversity. Widely used tools for analyzing metagenomic data can accurately quantify the abundance of species present in a microbial community, but often fall short in characterizing variation in gene content between strains [@Onate2018]. As a result, it is challenging to study the functional consequences of strain-level variation in the gut microbiome.

The most common way to study intraspecific variation *in situ* is to quantify the gene families present in shotgun metagenomes—an approach referred to as “pangenome profiling”. Pangenome profiling estimates the mean sequencing depth—sometimes called vertical coverage—of a gene family as the mean number of reads aligning to each base of a representative sequence [@Milanese2019] (Fig. 1A). (For brevity, we use “gene” as short-hand for gene family and “depth” for mean sequencing depth throughout this paper.) Several existing tools, including PanPhlAn [@Beghini2021] and MIDAS [@Zhao2022;@Nayfach2016] perform pangenome profiling. However, due to several sources of error in quantifying gene depth, a second algorithm is needed to infer which genes are actually present in a specific strain’s genome, a step that we call gene content estimation. Since this strain is never directly observed in isolation—indeed, it is only a hypothesis—we refer to it as an inferred strain. Tools for gene content estimation are often based on the assumption that all encoded genes will be at a similar depth: the same as the overall species depth [@Onate2018], which can be directly estimated from the depth of species marker genes [@Blanco-Miguez2023;@Milanese2019]. Therefore, the depth ratio—the ratio of a given gene’s depth and the overall species depth—can be used as the key criterion for the selection of genes [@Nayfach2016].

However, gene content estimation using pangenome profiles faces four key challenges (Fig. 1A, B):

1. an incomplete set of representative gene sequences in pangenome reference databases,
2. ambiguous alignment of short-reads to multiple sequences both within and across species ("cross-mapping"),
3. poor discrimination between present and absent genes at low depth due to high variance of the depth ratio, and
4. a low depth ratio for strain-specific genes when other strains of the same species are also abundant ("strain mixing").

Significant progress towards (1) has been recently achieved by expanding pangenome reference databases to include metagenome assembled genomes (MAGs), substantially improving their completeness [@Almeida2020]. However, cross-species contamination and genome assembly errors like gene fragmentation, which are common in MAGs, can exacerbate cross-mapping (2), reducing the accuracy of pangenome profiling [@Zhao2023]. Careful curation of the pangenome database is needed to reduce the impact of these issues. One promising approach for dealing with low depth (3), is to combine data across multiple samples (Fig. 1C) [@Carr2013;@Onate2018], taking advantage of increased depth from pooling reads. As a bonus, the correlation between the species depth and gene depth can be used as an additional criterion to better exclude genes with cross-mapping (2; Fig 1D). However, combining samples can exacerbate the impacts of strain mixing (4). Methods are needed for strain-aware gene content estimation that benefit from the increased sensitivity and specificity of multiple samples while also accounting for intraspecific variation.

Here we introduce StrainPGC, a computational method that leverages modern strain tracking tools to separate samples into strain-pure subsets in order to accurately estimate gene content based on multi-sample pangenome profiling (Fig. 1C, D). We also describe changes in MIDAS v3, including updates to its pangenome database and profiling pipeline to reduce cross-mapping, improve quantification, and facilitate the interpretation of strain-specific gene content. As part of a complete workflow (Fig. 1E), our method requires only shotgun metagenomes as input and outputs estimates of the gene content of individual strains. We apply this workflow to explore strains across a diverse collection of publicly available metagenomes from the human gut microbiome and find gene content variation with potential clinical relevance.

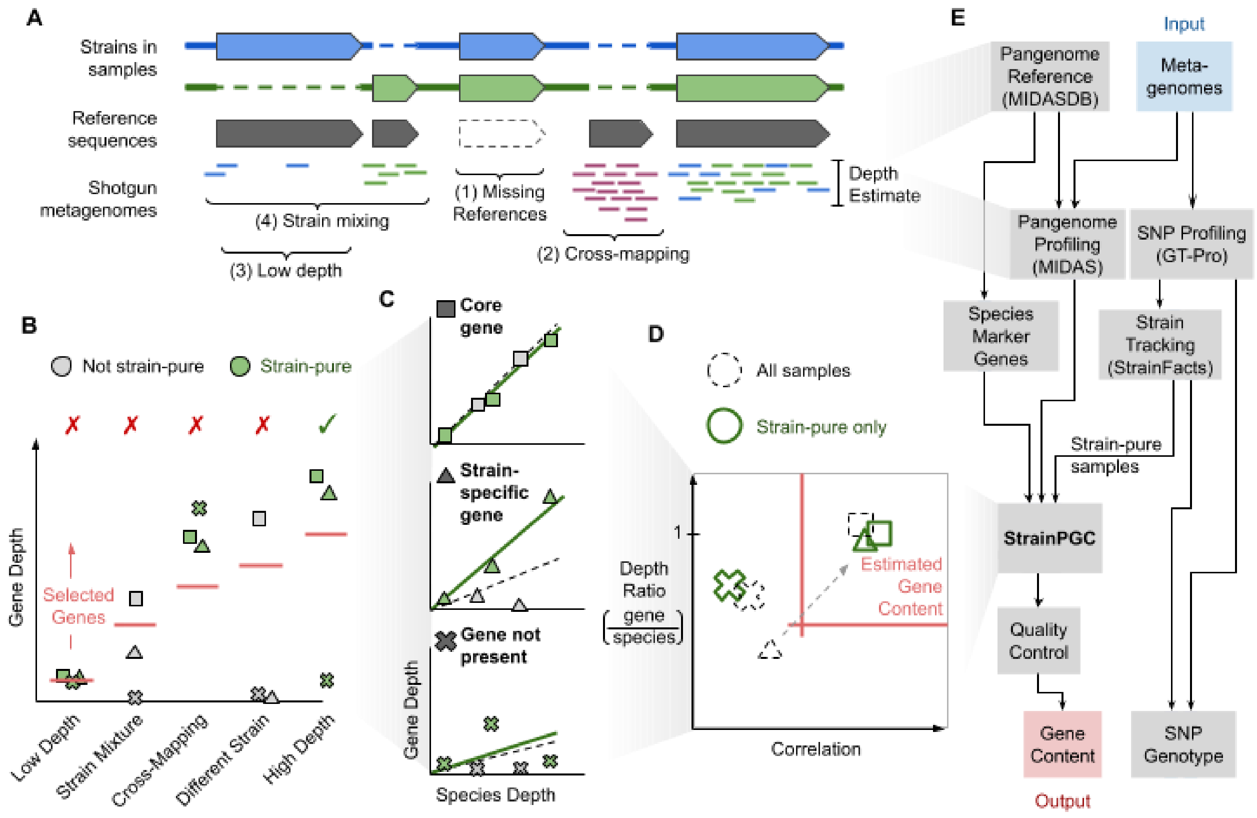

Figure 1: Conceptual overview of strain-resolved gene content reconstruction using StrainPGC. **(A)** Schematic representation of pangenome profiling, which estimates gene depth based on short-read alignment. The illustration represents profiling of a hypothetical microbial population harboring two strains of the same species (blue and green), each with both shared

and strain-specific gene content. Four key challenges for pangenome profiling and gene content estimation are highlighted (brackets, see Introduction). **(B)** Limitations of gene content estimation using single samples. Depth is shown across five samples (scenarios described along the x-axis) for three genes: one gene is ubiquitous across strains (“core”, square), another found in only the strain of interest (“strain-specific”, triangle), and a third not present in the strain of interest but subject to cross-mapping (“not present”, x-shape). Colors distinguish between strain-pure samples (green markers) and samples with a different strain or a mixture of more than one strain (gray markers). Traditional, single-sample analysis estimates gene content by selecting genes with a minimum depth (red, horizontal line, which is chosen based on the species's depth). As a result, samples with low depth, cross-mapping, and strain mixing all lead to decreased accuracy (indicated with red x’s). Only gene content estimation in a strain-pure, high-depth sample without cross-mapping (green check) accurately reflects the strain of interest. **(C)** Relationship between gene depth and species depth for each of the three genes (panels) across the five samples (marker shape and color as in B). For each, the linear relationship is shown between species depth and gene depth in the set of strain-pure samples (solid green line). We contrast this fit with the linear relationship across all five samples without considering strain variation (dashed line). **(D)** Schematic depiction of how StrainPGC estimates gene content based on both correlation and depth ratio. The red lines indicate the thresholds of depth ratio and correlation used by StrainPGC to select genes. With all samples combined (dashed markers), the “not-present” gene is correctly excluded due to low correlation, and the core gene is correctly included, but the strain-specific gene is lost due to its low depth ratio and correlation. Analyzing the strain-pure set separately moves the strain-specific gene into the selection region (dashed arrow), increasing accuracy. **(E)** Schematic depiction of our integrated workflow to infer gene content across strains using only shotgun metagenomic reads as input.

Results

A workflow integrating StrainPGC for strain-specific gene content estimation from metagenomes

To discover and characterize strains in large metagenome collections, we developed an integrated pipeline that takes as input shotgun metagenomes and outputs a holistic picture of strains and their gene content. In order to improve the completeness, curation, and interpretability of pangenome profiles, we made major updates to the pangenome database build process as well as the profiling algorithm and released these as MIDAS v3 (see Methods). Our overall workflow (Fig. 1E) is divided into four major stages: First, we profile pangenomes with MIDAS v3. Second, we identify and track strains with GT-Pro [@Shi2022] and StrainFacts [@Smith2023]. Third, we assign genes to strains with the core StrainPGC algorithm. Finally, we quality control strains, identifying and removing those likely to be of low accuracy.

The key novel contribution of our workflow is StrainPGC, an algorithm for strain-aware gene content estimation that integrates data from multiple samples to overcome the limitations of pangenome profiling and intraspecific variation (Fig. 1A-C). For each species, StrainPGC requires three inputs: (1) pangenome profiles, (2) a list of marker genes, which are used to estimate the species depth, and (3) the list of "strain-pure" samples for each strain, which was determined by StrainFacts. StrainPGC estimates the overall species depth across samples using the provided marker genes; "species-free" samples, those where the species is below the detection limit, are identified in this way. Then, working separately for each strain of a species, StrainPGC calculates two statistics for each gene (Fig. 1C): (1) the depth ratio in strain-pure samples—the total gene depth across samples divided by the total species depth—and (2) the Pearson correlation coefficient relating that gene's depth to the overall species depth across both the strain-pure and species-free samples. Genes with a sufficiently high correlation and depth ratio are estimated to be present in that strain's genome (Fig. 1D).

StrainPGC is open source and freely available at <https://github.com/bsmith89/StrainPGC>. While the work presented here uses MIDAS v3 and the comprehensive UHGG genome collection [@Almeida2020], the core StrainPGC software is designed to also work with pangenome profiling and strain tracking from alternative tools. Our integrated analysis workflow is implemented with Snakemake [@Molder2021] and is available at <https://github.com/bsmith89/StrainPGC-manuscript>.

StrainPGC accurately estimates gene content of strains in a complex synthetic community

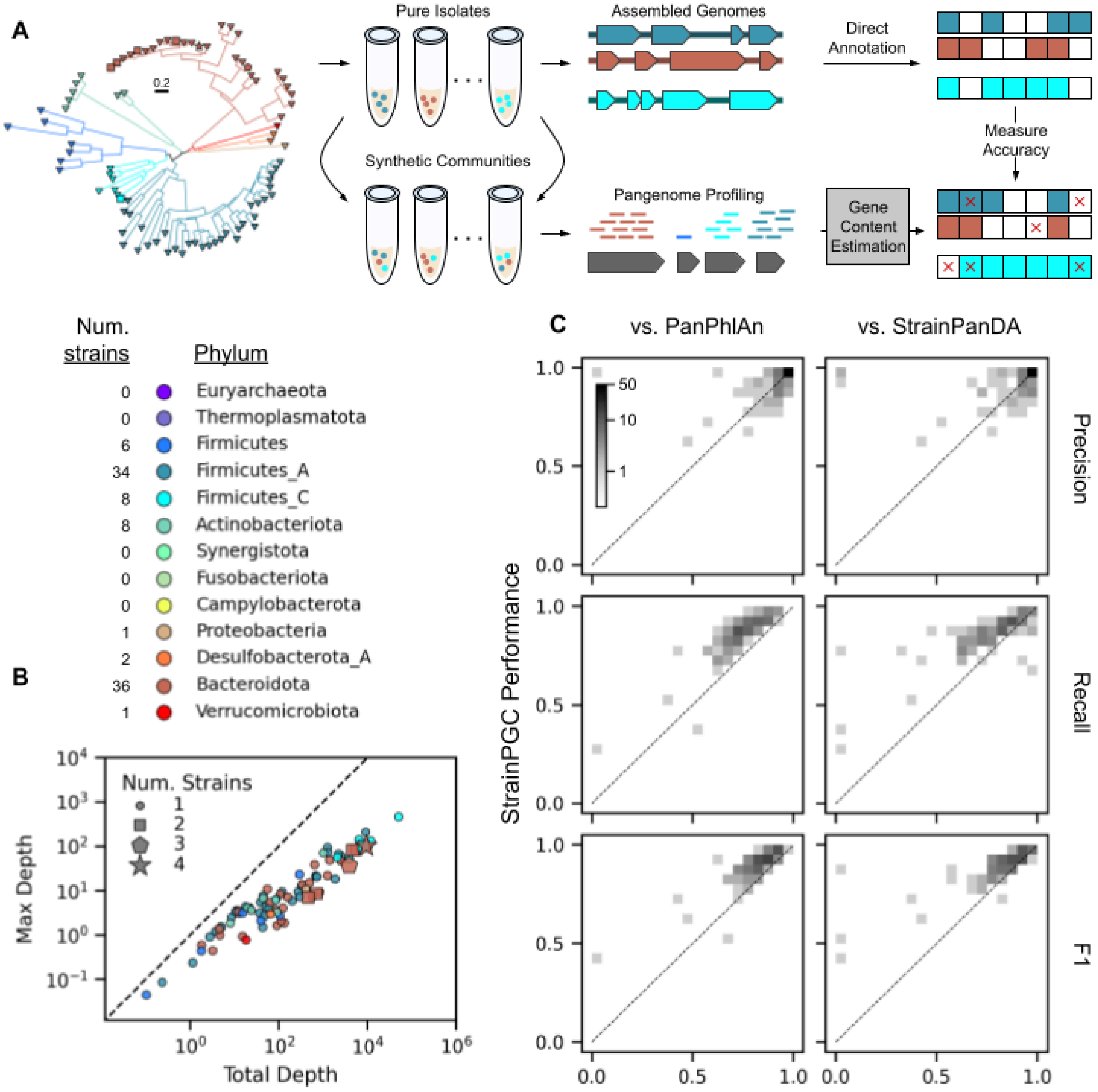

Figure 2: Evaluation of StrainPGC's gene content estimation performance on a highly diverse, synthetic community [@Jin2023]. (A) Schematic diagram of our procedure for benchmarking gene content estimates using a synthetic community constructed to reflect the species and strain diversity found in human gut microbiomes [@Cheng2022]. StrainPGC and alternative tools were applied to pangenome profiles from different samples derived from the

synthetic community, and estimates of gene content were compared to high-quality reference genomes for 105 strains. Strains were drawn from 95 species across 8 phyla (phylogenetic tree on the left, colored by phylum, scale bar in units of substitutions per position). **(B)** Core genome depths of 87 detectable benchmarking species span more than two orders of magnitude. Points represent individual species, are colored by phylum, and are placed based on that species's maximum depth across samples (x-axis) and total depth summed over all samples combined (y-axis). Species are closer to the 1-to-1 diagonal (dashed line) when the sample with the highest depth contributes more of their total depth. Some species are represented by more than one strain (marker shape). **(C)** Accuracy of gene content estimates by StrainPGC (y-axis) compared to PanPhlAn [@Beghini2021] and StrainPanDA [@Hu2022] (x-axes), as measured by precision, recall, and F1. All three indices range between 0 and 1, and higher values reflect better performance. The data are represented as two-dimensional histograms using a gray density scale to represent the number of strains falling in each (x, y) bin; density above the 1-to-1 diagonal (dotted line) indicates strains where StrainPGC outperformed the alternative on that index.

In order to evaluate StrainPGC's performance, we ran our workflow on publicly available metagenomes from a diverse, synthetic bacterial community grown *in vitro* under five different experimental conditions (276 metagenomes in total) [@Jin2023]. The initial inoculum of this community was composed of 117 bacterial isolates spanning 8 phyla, each with a high-quality genome assembly, which we refer to as ground truth genomes (Fig. 2A). Most species were represented by a single strain, some by 2 or 3 strains, and one by 4 (Fig. 2B). We refer to the collection of ground-truth genomes and experimental metagenomes as the benchmarking dataset. We annotated predicted protein-coding genes in the ground truth genomes with EggNOG OGs (Fig. 2A). After removing species that could not be genotyped by GT-Pro, or that were undetected in metagenomes, the benchmarking task amounted to 87 species encompassing 97 strains and with highly disparate depths (estimated maximum sample depth interquartile range of 2.7–22.4x) (Fig. 2B). We applied StrainPGC to estimate gene content across inferred strains, matched each ground truth strain to a single inferred strain based on SNP genotypes, and compared the EggNOG OGs annotations between these. In this benchmark, StrainPGC had a median precision of 0.96 (IQR: 0.90–0.98; Fig. 1C), a recall of 0.88 (0.82–0.93), and an F1 score of 0.91 (0.87–0.94).

We next compared StrainPGC's performance to two alternative, state-of-the-art methods: PanPhlAn [@Beghini2021], which is widely used and operates on single samples, and

StrainPanDA [@Hu2022], a recently published tool that harnesses information across multiple samples and applies non-negative matrix factorization to jointly estimate gene content and strain depth (Fig. 2C). For all three methods, we used the same reference database and pangenome profiles as input, thereby comparing the core gene content estimation approaches on an equal basis. However, since strains inferred using PanPhlAn and StrainPanDA do not have SNP genotypes to be used for matching, for each benchmark genome, we instead selected the inferred strain with the highest F1 score, giving these two methods an advantage. Nonetheless, StrainPGC performed better on average than either alternative: a median increase of 0.069 in F1 score compared to PanPhlAn (IQR: 0.038–0.093; $p < 1e-10$ by Wilcoxon, non-parametric, paired, t-test) and 0.042 relative to StrainPanDA (IQR: 0.022–0.079; $p<1e-10$). All three tools had high precision, and the superior performance of StrainPGC was driven primarily by the recall: 0.12 greater than PanPhlAn and 0.08 greater than StranPanDA ($p < 1e-10$ for both). For all three tools, species with higher estimated depth had better performance on this benchmark (Spearman's correlation between maximum species depth across samples and F1 score: Spearman's $\rho = 0.30$, 0.55, and 0.33 for StrainPGC, PanPhlAn, and StrainPanDA, respectively; Supplementary Results 1). However, StrainPGC had a weaker relationship between precision and depth than PanPhlAn and StrainPanDA did ($\rho = 0.19$, 0.53, and 0.56, respectively). Since we controlled for other steps of the gene content estimation process, these findings support the idea that StrainPGC's use of correlation across strain-pure samples allows us to maintain high precision even while increasing recall. In particular, we find our approach maintains this specificity even at low depths more effectively than existing methods.

In real-world applications—where ground-truth gene content is not known a priori—it is beneficial to understand the confidence of StrainPGC estimates. We, therefore, calculated two scores to serve as proxies for accuracy and compared these to the performance we measured on the benchmarking datasets (Supplementary Results 1). First, we hypothesize that the fraction of high-prevalence, species marker genes assigned to a given inferred strain reflects the overall completeness of the estimated gene content for that strain. Indeed, across benchmark genomes, we found a strong correlation between the fraction of species marker genes and the F1 score ($\rho = 0.60$, $p < 1e-10$). As expected, this appears to be driven primarily by a strong association with the recall ($\rho = 0.63$, $p < 1e-10$); a weaker correlation was found with the precision ($\rho = 0.34$, $p < 1e-3$). Second, for strains suffering from low signal-to-noise, such as those at low sequencing depths, the depth ratio of assigned genes will be more variable. We, therefore, calculated a noise index reflecting: the standard deviation across all assigned genes of the log10-transformed depth ratio. For this score, we found a negative correlation with the F1 score ($\rho = -0.68$, $p < 1e-10$), this time driven by an association with the precision ($\rho = -0.58$, $p < 1e-9$) as well as recall ($\rho = -0.53$, $p < 1e-8$). In our benchmark, the 22 strains with < 95% species marker genes or a noise index > 0.25 had substantially lower F1 scores than those that passed this quality control (median of 0.83 versus 0.92, $p < 1e-5$ by MWU test; Supplementary Results 1). We propose using these two criteria together in order to exclude inferred strains with lower accuracy gene content estimates.

Inferred strains in publicly available metagenomes substantially expand the catalog of intraspecific diversity

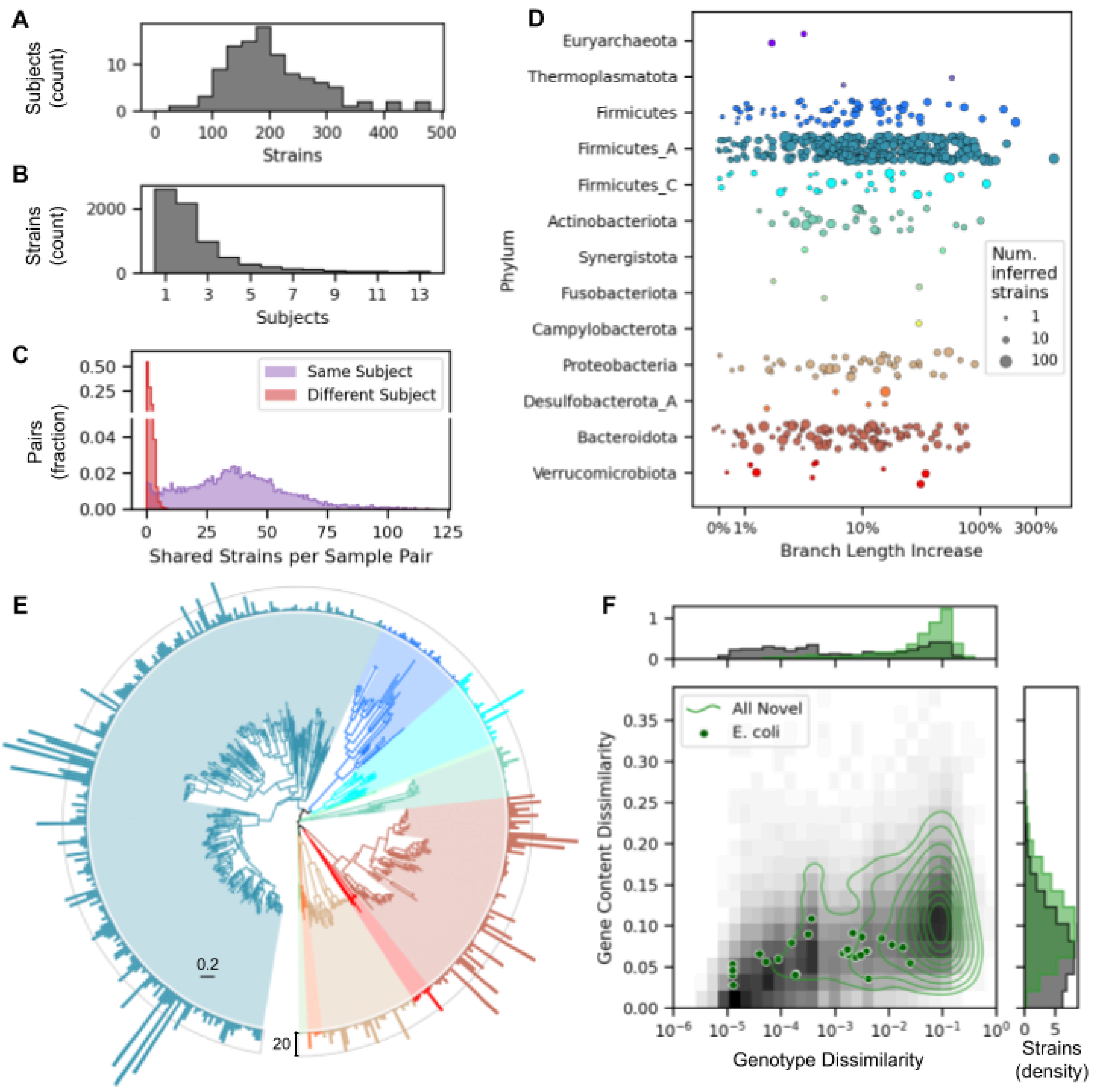

Figure 3: Strain diversity in the HMP2 metagenome collection. (A–B) Histograms reflecting the distribution of inferred strains of any species across subjects in the HMP2 metagenome collection. **(A)** Number of strains for 106 subjects, summed over all samples (median of 11 samples per subject; IQR: 9–14). Most subjects harbor between 100 and 300 inferred strains (median of 191.5). **(B)** Number of subjects where each strain was detected. Only strains found in two or more samples are tallied. Most strains (67%) were found in just one or two subjects. **(C)** Number of strains shared in any pair of samples from the same (purple) or different (red) subjects. Pairs of samples from different subjects shared a mean of just 0.7 strains. **(D)** A

substantial increase in strain diversity was captured when including inferred strains. Diversity was quantified based on total branch length in a hierarchical clustering (UPGMA) of all SNP genotypes, and the increase was measured as the change in branch length relative to a tree with only reference strains. Points represent individual species, are colored by phylum, and increasing size reflects a larger number of inferred strains. Five species with fewer than 3 inferred strains had a small decrease in branch length when inferred strains were included; one of these is excluded from the plot, left of the x-axis limit. **(E)** Taxonomic diversity of 3504 inferred strains of Bacteria. The species tree is colored by phylum as in (D). Species that had no strains with estimated gene content were omitted, and bars around the outer ring indicate the number of inferred strains (outer ring indicates 20 strains). The branch length scale bar (interior) is in units of substitutions per position. **(F)** Estimated genotype and gene content dissimilarity from the closest reference genome. Joint (main panel) and marginal distributions (panels above and to the right) are plotted for all high-quality reference (gray background) and inferred (green contours) strains of all species. Gene content dissimilarity of inferred strains is calculated after batch correction (see Methods). Points reflecting each of 28 inferred *E. coli* strains are also shown. Green contours in the main panel reflect deciles in the 2D kernel density estimator.

We applied our workflow to the 106 subjects and 1338 samples of the HMP2 metagenome collection—which we refer to as simply the HMP2 throughout this paper.

First, to explore the strain-level diversity that might be discovered in publicly available datasets, we used StrainFacts to identify and estimate the distribution of strains based on SNP profiles. We defined detection as an estimated depth of ≥ 0.1x, a threshold chosen to balance false positives with the sensitivity of strain tracking (Supplementary Results 2). All species combined, a median of 59 strains were detected in each metagenomic sample (not shown) and 191.5 across all samples from each subject (Fig. 3A). This strain-level diversity was highly subject-specific; among inferred strains detected in two or more samples, 36% were detected in just one subject, and only 34% were detected in three or more (Fig. 3B). Strain sharing was dramatically more common in pairs of samples from the same subject than in pairs of samples from different subjects (mean of 36.7 shared, detected strains from same subject vs. 0.7 from different subjects, p<1e-10 by MWU; Fig. 3C), consistent with prior studies of the HMP2 and other cohorts [@Lloyd-Price2017].

Concordant with this level of strain diversity, estimated genotypes for inferred strains were often distinct from the closest reference strain (Supplementary Results 3). Using SNP profiles in strain-pure samples, we estimated each inferred strain's genotypes as the consensus allele, masking ambiguous positions. Among inferred strains with ≥ 100 genotyped positions, 68% had a genotype dissimilarity of greater than 0.05 to the closest reference. Representing the strain diversity of each species with a UPGMA tree, we calculated the increase in total branch length when including inferred strains relative to only references (Fig. 3D). For many species, a substantial increase in total branch length was observed: more than 10% for 288 species, more than 20% for 183 species, and more than 50% for 63 species when inferred strains were included. Overall, these findings suggest that inferring strains from publicly available metagenome collections will reveal novel intraspecific diversity not already found in reference databases.

We next applied StrainPGC to estimate gene content for these strains. After quality control, we estimated gene content for 3511 inferred strains in 443 species across 12 phyla (Fig. 3E, Supplementary Results 3). While these were primarily Bacteria, we were also able to estimate gene content for strains in three species of Archaea. The largest number of inferred strains were classified in the phylum Firmicutes_A (2232 strains; an additional 80 and 141 strains were also in "Firmicutes", and "Firmicutes_B", respectively, which are classified as separate phyla in the GTDB taxonomy), followed by Bacteroidota (727), and Proteobacteria (189). Hence, StrainPGC resolved gene content for myriad strains across a diverse set of species found in the human gut (Supplementary Table 1).

Just like SNP genotypes, for most inferred strains, the estimated gene content was quite distinct from the closest reference. Measuring dissimilarity using the cosine dissimilarity after batch correction (see Methods), inferred strains were a median of 0.18 from the closest, high-quality reference genome (Fig. 3F). As would be expected, strains with more dissimilar SNP genotypes were often those with dissimilar gene content as well. For instance, across the 28 inferred strains of *E. coli*, we found a significant correlation between the gene content dissimilarity and the genotype dissimilarity (Spearman's $\rho = 0.44$, $p = 0.018$; Fig. 3F). This suggests that the increased diversity captured by StrainPGC facilitates expanded analyses of intraspecific gene content variation in the gut microbiome.

Estimated gene content enables pangenome analyses in prevalent human gut microbes

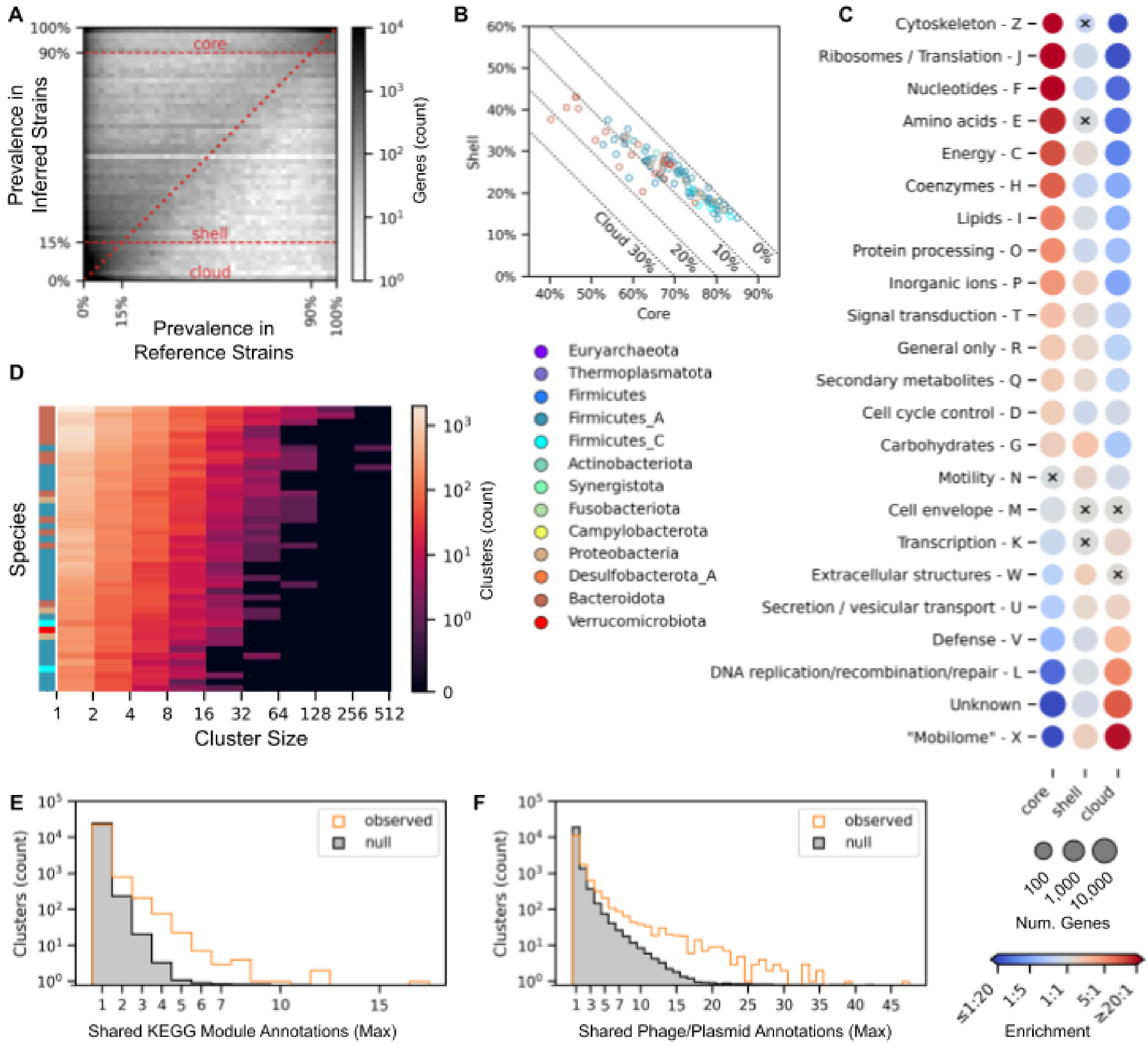

Figure 4: StrainPGC reveals patterns of gene content variation across dozens of species. (A) Gene prevalence across inferred strains from HMP2 is very similar to prevalence in reference genomes. Combining genes from all species, the 2D histogram shows the joint distribution of prevalence estimated from reference genomes (x-axis) and inferred strains (y-axis). These independent estimates are highly concordant, with higher density along the 1-to-1 diagonal (dotted line). Dashed horizontal lines represent the thresholds defining core, shell, and cloud prevalence classes based on inferred strains. **(B)** Fraction of shell versus core genes in inferred strains. For each species (circle), x and y values are the median gene content

in the core and shell classes, respectively. The remaining gene content is composed of cloud genes and is indicated by the dotted diagonal lines. Markers are colored by phylum. **(C)** Enrichment (red) or depletion (blue) in genes of various functional categories in each of the core, shell, and cloud prevalence classes. Dots representing each COG category (rows) and prevalence class (columns) are colored by odds ratio, with red and blue indicating enrichment and depletion, respectively. Dot size reflects the number of genes in that prevalence class that are in the given functional category. All enrichments/depletions shown are significant (Two-tailed Fisher Exact Test; $p < 0.05$), except for those marked with a black cross. COG categories A, B, and Y are omitted, as these had very few members (173, 74, and 0 genes, respectively). **(D)** Gene co-occurrence clusters based on estimated gene content. The heatmap depicts histograms for each of 44 species (rows) of cluster sizes (columns). Colors indicate the number of clusters in each interval, and labels along the x-axis indicate the bounds of the intervals (left exclusive, right inclusive). Colors on the left indicate phylum as elsewhere. **(E, F)** The maximum number of related annotations in each co-occurrence cluster. The orange histogram represents the observed distribution, while the gray region is the mean in each bin across 100 random permutations of cluster labels (i.e. the null distribution). The higher number of clusters with multiple, shared annotations in the observed data compared to the null suggests clumping of **(E)** KEGG module and **(F)** phage or plasmid genes into co-occurrence clusters.

To demonstrate the value of gene content estimates derived from the HMP2 for pangenome analysis, we focused on the 99 species with estimated gene content for 10 or more inferred strains (Median: 17 inferred strains per species, IQR: 12–28, 7 phyla). For each species, we calculated the prevalence and distribution of genes across strains. Gene prevalence estimates based on inferred strains were highly correlated with the prevalence observed in high-quality reference genomes ($r = 0.84$, $p < 1e-10$; Fig. 4A), supporting the consistency of our estimates with the existing reference database.

Based on these de novo prevalence estimates, we assigned genes to the “core” (≥ 90% prevalence), “shell” (< 90% and ≥ 15%), or “cloud” (< 15%) pangenome fractions. We then calculated the portion of estimated gene content that fell into each prevalence class for each inferred strain (Fig. 4B). Computing the median first within and then across species, genes in the core fraction made up 70% (IQR: 63–76%) of each strain’s estimated gene content, shell fraction 25% (19–28%), and cloud fraction 5% (4–9%), in general agreement with reference genomes (Supplementary Results 4). Certain categories of functional annotations were more common in each fraction (Fig. 4C). Core genes were enriched for COG categories with housekeeping functions, such as translation, cytoskeleton, and the transport and metabolism of nucleotides, amino acids, coenzymes, and lipids. The shell pangenome, on the other hand, was enriched in functional categories including carbohydrate and inorganic ion transport and metabolism, and extracellular structures. Finally, the cloud pangenome was enriched in functional categories including the mobilome, DNA replication, recombination and repair, and defense mechanisms, as well as genes without a COG category. Broadly, these patterns of enrichment confirm our expectations that core genes perform obligate functions and make up a plurality of genes for most strains.

As an assembly-free approach, gene content estimation lacks synteny information, which can be useful for understanding biological phenomena such as operonic co-regulation and horizontal gene transfer. To get around this limitation, we clustered genes based on the Pearson correlation of their presence and absence across inferred strains in the HMP2. For the 44 species with more than 20 high-quality inferred strains, we identified 36,208 co-occurring gene clusters with 2 or more members, a median of 681.5 per species (Fig. 4D, Supplementary Results 5). Genes in the same cluster were more likely to have related annotations; clusters having three or more genes in the same KEGG module were 12.7x more common than expected by random chance (n = 100 permutations of cluster labels within species, $p < 1e-2$; Fig. 4E). Likewise, phage- or plasmid-associated genes were more frequently found in the same clusters than expected by chance (three or more shared annotations 2.4x more common, $p < 1e-2$; Fig. 4F). This supports our interpretation of StrainPGC–enabled gene co-occurrence clustering across genomes as evidence of related biochemical function or linked transmission, which may help to generate testable hypotheses about relationships between genes in a species' pangenome.

Overall, large surveys of gene content estimated by StrainPGC have the potential to vastly expand the coverage and diversity of pangenome analyses.

Integrative analysis of *E. coli* strain gene content can inform the selection of donors for fecal microbiota transplantation

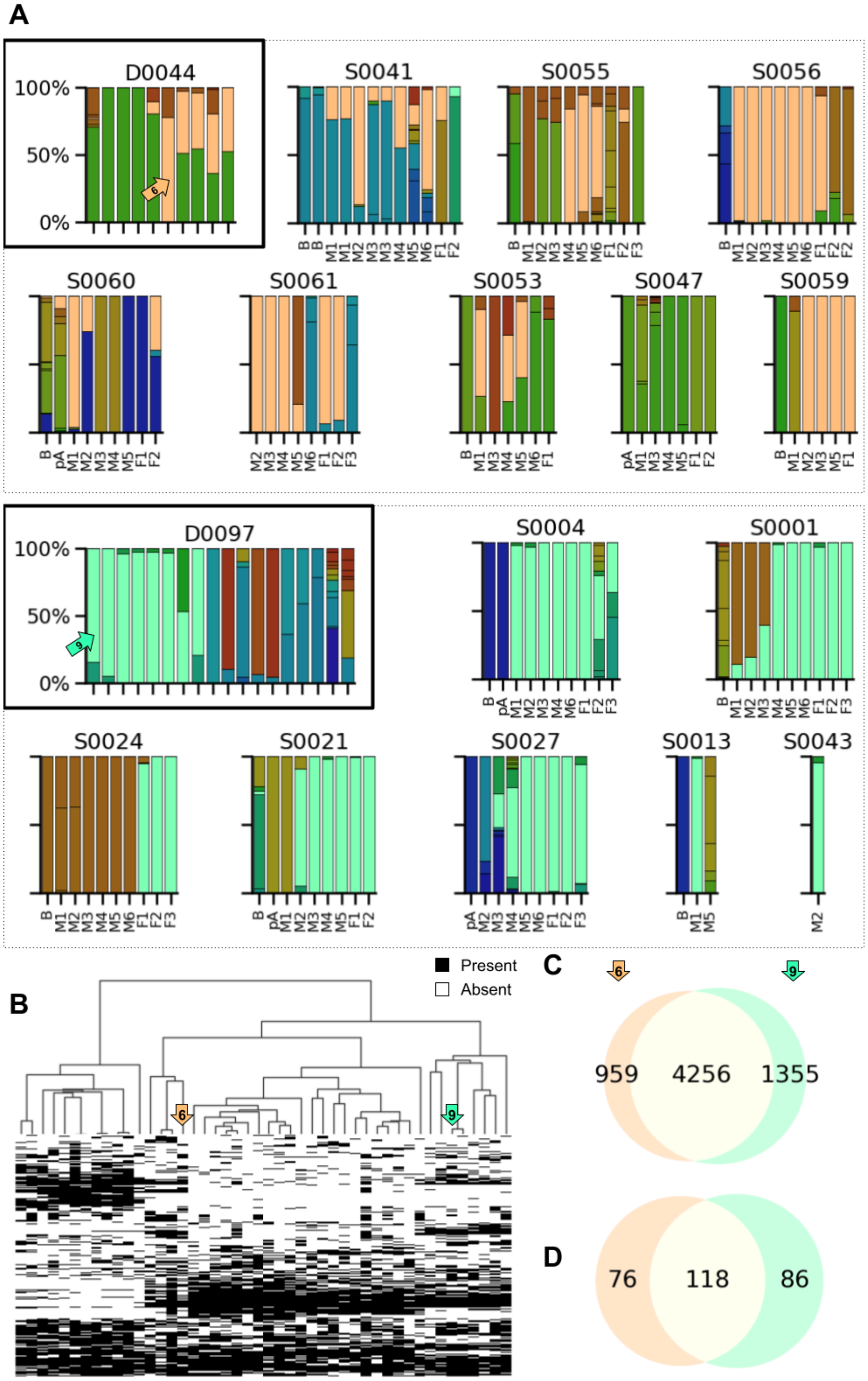

Figure 5: Different donors in a fecal microbiota transplant (FMT) trial [@Smith2022] have engrafting *E. coli* strains that differ in their functional potential. (A) *E. coli* strains found in repeated sampling of two independent donors' fecal materials (boxed panels) and in the fecal time series of their respective recipients. Columns in each panel represent individual samples, colors represent *E. coli* strains inferred from StrainFacts, and the height of colored bars indicates strain abundance normalized to total *E. coli* abundance in the sample. For donors, samples are ordered arbitrarily. Recipient samples are ordered by collection day and include samples at baseline (labeled "B") collected before initial FMT treatment, samples collected before each of up to six maintenance FMT doses (labeled "M1" to "M6"), and up to three follow-up samples (labeled "F1" to "F3"). For a subset of recipients, samples were also collected after antibiotic treatment and before FMT (labeled "pA", post-antibiotics). For each donor, one strain (tan in D44, aqua in D97) showed a high rate of engraftment in recipients at follow-up. **(B)** Comparison of shell gene content between inferred strains from the FMT experiment (18 strains) and *E. coli* strains from the HMP2 (28). Heatmap indicates the presence and absence of genes (rows) across inferred strains (columns). Strains are ordered by UPGMA tree of estimated SNP genotype dissimilarity. Genes are filtered to only the 3,134 genes in the shell pangenome fraction. Arrows (tan and aqua) highlight the high-engraftment strains from panel (A). **(C, D)** Estimated gene content that is shared and distinct between the two high-engraftment strains. Venn diagrams depict the intersection of **(C)** genes and **(D)** gene co-occurrence clusters.

We next sought to assess the potential utility of StrainPGC gene content estimates for optimizing microbial therapies such as FMT. Current donor screening protocols focus on detection of known pathogens and do little to match donors to recipients or optimize for transmission and engraftment of particular microbial functions. To assess the sensitivity of our approach for comparing donor strains, we re-analyzed metagenomes from a previously published study of FMT for the treatment of ulcerative colitis [@Smith2022]. We refer to these metagenomes as the UCFMT dataset. As a proof-of-concept, we focused on strains of *Escherichia coli*, a well-studied and highly prevalent member of the human gut microbiome with well-documented examples of not only pathogenic but also commensal and even probiotic strains [@Blount2015].

Using 231 samples collected longitudinally from patients (189 samples) and donors (42 samples) in the UCFMT study, we identified and tracked strains using StrainFacts. For two donors in particular, D44 and D97, we observed repeated transmission of strains during FMT

(Fig. 5A). Next, with StrainPGC, we obtained gene content estimates for inferred strains of *E. coli*; 18 passed quality control. In order to examine their genetic relatedness—and to put them in the context of the earlier pangenome analysis—we combined inferred strains from the UCFMT and HMP2 metagenomes and generated a UPGMA tree based on their SNP genotype dissimilarity (Fig. 5B). As before, genotype and gene content were related (Fig. 5B): for the combined set of inferred strains, we found a robust correlation between the cosine dissimilarity of the shell pangenome fraction—defined above using the HMP2 strains—and genotype dissimilarity (r = 0.88, Fig. 5C).

In recipients of donor D44, one strain, strain-6, stood out as frequently present both during six weeks of maintenance dosing and in subsequent follow-up sampling (Fig. 5A). Likewise, strain-9 engrafted frequently for recipients of D97. These two strains had a SNP genotype dissimilarity of 0.23, while the median dissimilarity across all pairs of UCFMT strains was 0.25 (IQR: 0.13 – 0.31; Supplementary Results 6). Approximately 80% of each strain’s gene content was shared with the other, while 18% and 24% was private to strain-6 and strain-9, respectively (Fig. 5C; Supplementary Table 2). Cross-referencing co-occurrence clusters with the estimated gene content of these strains, about 60% of clusters in each were shared, with 39% and 42% private, respectively (Fig. 5D). Of the 118 shared clusters, 12 were found in no more than two additional UCFMT strains. We hypothesize that these might indicate important physiological similarities that distinguish high-engraftment strains from the others. Among 85 genes in these shared clusters, the most common COG category annotation was X ("Mobilome") reinforcing that phage, plasmids, and other mobile genetic elements are an important source of shared gene content across distantly related strains.

Next we sought to understand functional gene differences between the two high-engraftment strains, in particular any that might result in disparate impacts on host health. We therefore examined the unshared gene content in order to identify plausible physiological differences (Supplementary Table 2). Strikingly, strain-9 had 12 genes annotated as related to antimicrobial resistance, suggesting potential resistance to 17 different antibiotics, while strain-6 had none. Among gene co-occurrence clusters, one (labeled clust-861) is also found only in strain-9, and includes genes with homology to components of a type VI secretion system (T6SS). Most T6SSs are involved in inter-microbial competition, although a role in pathogenesis has also been described [@Navarro-Garcia2019]. Another cluster private only in strain-9, labeled clust-37, includes genes with homology to many components of a type IV secretion system (T4SS), other secretion systems, a helicase, and a component of a toxin/anti-toxin system.

Combined, these annotations suggest that the cluster may primarily reflect a mobilizable plasmid in strain-9 that is missing in strain-6. Similarly, related annotations in several clusters (clust-351, clust-352, and clust-353) have homology to genes in the *pdu*-operon. This operon encodes components of catabolic bacterial microcompartments, which are involved in various catabolic pathways, including 1,2-propanediol utilization. These co-occurrence clusters are found only in strain-9, and strain-6 is missing homology to most of the genes in the *pdu*-operon. Microcompartments and 1,2-propanediol utilization have been associated with pathogenicity in *E. coli* and other species of *Enterobacteriaceae* [@Prentice2021].

Given the presence of AMR genes and the plausible association between several co-occurrence clusters and pathogenesis, we speculate that the engraftment of *E. coli* strain-9, found in FMT samples donated by D97, could result in a less beneficial or even detrimental treatment for recipients. Similarly, the engraftment of strain-6 from D44 might contribute to the competitive exclusion of related pathogenic strains. While the previously published study found no difference in outcomes between recipients of the two donors [@Smith2022], this study may have been underpowered (n = 8 recipients for each of D44 and D97). Our computational predictions could be tested *in vitro* with isolates obtainable from archived donor materials.

Discussion

Here we have described updates to the MIDAS v3 pangenome database and profiling software, as well as StrainPGC, a novel tool for accurate, strain-specific gene content estimation using metagenomic data. The key innovations of StrainPGC are the use of depth correlation information and selection of strain-pure samples. Together, these innovations enable StrainPGC to outperform PanPhlAn and StrainPanDA in a benchmark based on a complex community modeled after the human gut microbiome. Combining the updated MIDAS v3 and StrainPGC in our workflow, we estimated gene content for thousands of strains in the HMP2 metagenome collection, substantially expanding on the diversity found in reference genome collections and enabling analyses of intraspecific variation without isolation or assembly. Finally, we used StrainPGC to compare the functional potential of two different strains of *E. coli* that were successfully transferred from two different donors in a clinical trial of FMT.

StrainPGC is an assembly-free method, and complements high-quality genome sequences enabled by laboratory isolation and culturing, as well as modern, long-read sequencing and de novo assembly from metagenomes, which remain the gold standard for comparative genomics.

However, these methods are labor intensive, expensive, and often fail to capture low-abundance organisms [@Chen2020]. Interestingly, DESMAN [@Quince2017], while based on de novo assembly, takes a conceptually similar approach to StrainPGC, combining strain tracking with gene content estimation. StrainPGC identified extensive, underexplored diversity in the well-studied HMP2, further suggesting that many strains are missed by culturing and assembly-based methods. Nonetheless, these technologies are important complements to our approach and contribute to the completeness of reference databases.

Given the enormous diversity of strains found across subjects in the HMP2, the StrainPGC approach may be most useful for analyzing FMT, longitudinal, or other study designs where the same strains are expected to be found in multiple samples. While StrainPGC is specifically designed to overcome the limitations of short-read, alignment-based pangenome profiling, in particular ambiguous mapping to homologous sequences both within and across species, systematic false positive and false negative gene assignments may still occur. As a result, we caution against over-interpreting analyses that rely on directly comparing the gene content of inferred strain with reference strains. Another major barrier to interpreting gene content estimates by StrainPGC or other methods is the sparsity of robust genetic, biochemical, structural, and experimental characterization of gene products [@Zhou2019]. While we augmented available annotations by leveraging co-occurrence clusters to investigate epistatic and evolutionary relationships between genes—as others have done previously [@Minot2019]—laboratory-based characterization is still vital.

Packaged as stand-alone software tools and integrated into an automated workflow, MIDAS v3 and StrainPGC together facilitate the broad exploration of strain-specific gene content in metagenome collections. Future studies can expand surveys across additional metagenomic datasets, look for associations between microbial strains and disease, and identify determinants of success for FMT.

Methods

MIDAS v3 update

Here we describe updates in MIDAS v3, including changes to the pangenome reference database construction procedure and the pangenome profiling method. Together, these updates clean, functionally annotate, and expand the phylogenetic coverage of MIDAS pangenome profiling, providing a foundation for accurately estimating and interpreting gene content across species. MIDAS v3 is available at <https://github.com/czbiohub-sf/MIDAS2> and can be installed using conda or Docker. Compatible, pre-built MIDAS databases based on UHGG [@Almeida2021] v2.0 and GTDB [@Parks2022] r202 will be available in the near future. We use the UHGG database throughout this work.

Pangenome database curation and clustering

A MIDAS v3 pangenome database can be constructed from any reference genome collection, and is composed, for each species, of predicted gene sequences from all example genomes clustered into operational gene families (OGFs) at a series of average nucleotide identity (ANI) thresholds. For clarity, we have referred to these OGFs simply as genes in the main text. In order to minimize the impacts of inter- and intra-specific cross-mapping on pangenome profiling, which can be major problems for gene databases constructed with MAGs, we made major changes to the clustering and curation pipeline. In this MIDAS update, described below, we sought to minimize the impact of fragmented gene sequences, spurious gene calls, chimeric assemblies, and redundant OGFs resulting from these errors [@Li2022;@Hyatt2012;@Dimonaco2022].

For each species, for each reference genome in the source genome collection, genes were predicted by Prokka v1.14.6 [@Seemann2014], wrapping Prodigal v2.6.3 [@Hyatt2010]. Gene sequences less than 200 bp or with ambiguous bases (anything but A, C, G, or T) were removed. Then, the remaining sequences were dereplicated by clustering at a 99% ANI threshold using VSEARCH v2.23.0 [@Rognes2016], with the longest sequence initially assigned as the representative sequence for the cluster. Next, in order to identify and remove additional cases of fragmented genes, we applied CD-HIT v4.8.1 [@Fu2012] (using options `-c 1 -aS 0.9 -G 0 -g 1 -AS 180`); when a shorter representative sequence had perfect identity over ≥ 90% of length to a longer sequence, the two clusters were merged, and the longer sequence was assigned as representative. Short gene sequences predicted on the opposite strand, a known complication [@Trimble2012], were also merged in this way.

Having dereplicated and cleaned gene sequences, we further clustered representative sequences into OGFs using VSEARCH, defining final OGF clusters at thresholds between 95% and 75% ANI.

Pangenome database annotation

Next, we annotated sequences using a variety of tools. We ran EggNOG mapper v2.1.12 [@Cantalapiedra2021] on dereplicated genes to identify homology relative to several commonly used gene orthologies: COGs, EggNOG OGs, and KOs. ResFinder v4.4.2 [@FerrerFlorensa2022], geNomad v1.7.4 [@PedroCamargo2023] and MobileElementFinder v1.1.2 [@Johansson2021] were run directly on contigs of each reference genome to identify AMR, phage, plasmid, and mobile element associated regions, and these annotations were transferred onto predicted genes based on overlapping coordinates.

While annotations are performed on genomic sequences or dereplicated gene sequences, interpretation of estimated gene content requires annotations at the OGF level. We therefore implemented a voting procedure intended to enable the transfer of annotations from gene sequences to gene clusters. For OGFs at each ANI level, we calculated the fraction of genes in each cluster annotated as an AMR gene, phage-associated, plasmid-associated, or mobile element-associated. In this way, users can identify annotations robustly associated with genes of interest.

Alignment and gene depth estimation

For pangenome profiling, the MIDASDB representative gene sequences from selected species are compiled into an index for alignment and quantification. At this stage, we apply additional filtering to the set of representative sequences, which we refer to as “pruning”, with the goal of speeding up alignment and improving quantification by reducing the rate of cross-mapping within and between species. First, we remove representative sequences that are less than 50% of the median length in the 95% ANI cluster, as these are more likely to be truncated genes resulting from assembly fragmentation. Second, for species with more than 10 reference genomes, we remove representative sequences where their 75% ANI clusters had only one member, as these are more likely to be spurious gene calls or contamination resulting from chimeric assembly. Finally, an alignment index is constructed from the remaining representative sequences, and reads are mapped using Bowtie2 [@Langmead2012].

Pangenome profiling with MIDAS v3 proceeds through four stages: (1) building a reference index as described above, (2) alignment of reads to the reference index, (3) calculation of the mean depth across the length of the representative sequence, and then (4) summation of representative sequence depths into clusters in order to estimate the total depth of the OGF at the chosen ANI threshold.

Shotgun metagenomes

All shotgun metagenomes analyzed in this work are publicly available as SRA BioProjects, including the HMP2 (PRJNA398089), UCFMT (PRJNA737472), and the synthetic community data (PRJNA885585) used for benchmarking. Downloaded HMP2 metagenomes had human reads removal and quality control procedures previously applied. UCFMT metagenomes were filtered for human reads, deduplicated, adapter trimmed, and quality trimmed, as described in [@Smith2022]. Benchmark metagenomes were processed in the same way, except that human read removal was skipped because the data was collected *in vitro*.

Integrated analysis pipeline

Pangenome profiling

For the work presented here, we ran MIDAS v3 as follows. Using Bowtie2 v2.5.1 throughout, a single reference index was built for 627 species using `midas2 build_bowtie2db --prune_centroids --remove_singleton`. Paired-end reads for each sample were aligned to this index using ` midas2 run_genes --aln_speed sensitive --aln_extra_flags '--mm --ignore-quals' --total_depth 0`. Mean mapping depth was calculated using `samtools depth` and summed up at the 75% ANI OGF level.

Reference genomes and species marker genes

High-quality reference genomes in the UHGG were defined as those with estimated completeness of > 90% and contamination of < 5%. OGFs found in > 95% of high-quality reference genomes were selected as species marker genes and were used for species depth estimation, quality control, and downstream analyses.

SNP profiling

SNP profiles were obtained from metagenomes using GT-Pro v1.0.1 [@Shi2022] and the default database, which was built using UHGG v1.0. GT-Pro was run on preprocessed reads, and counts from forward and reverse reads were summed. The resulting SNP profile matrix, a three-dimensional array of counts indexed by sample, genotyped position, and allele (reference or alternative), is the core input for StrainFacts [@Smith2023].

An analogous approach was used to obtain SNP genotypes for genomic sequence. Specifically, for both reference and benchmarking genomes, contigs were fragmented into 500 bp tiles with 31 bp of overlap and used as input to GT-Pro. We filtered out tallies for SNP sites that did not match the expected species.

Strain tracking and genotyping

For each species, SNP profiles obtained from GT-Pro were filtered to remove low-depth samples (those with <5% of positions observed). For the HMP2 and UCFMT datasets, low polymorphism positions (minority allele observed in <5% of samples) were also removed. However, this latter filter was not applied to the synthetic community since many species had only one strain. Strain genotypes and proportions were estimated with StrainFacts v0.6.0, using the updated Model 4 (Supplementary Methods 1) and a number of strains set as $n^{0.85}$ where n is the number of samples. For the vast majority of species, this model was fit using a single, standardized set of hyperparameters: `--optimizer-learning-rate 0.05 --min-optimizer-learning-rate 1e-2 --hyperparameters gamma_hyper=1e-15 pi_hyper=0.01 pi_hyper2=0.01 rho_hyper=1.0 rho_hyper2=1.0 --anneal-hyperparameters gamma_hyper=0.999 --anneal-steps 120000`. However, for seven species (species IDs: sp-100076, sp-101302, sp-101306, sp-101704, sp-102478, sp-103456, sp-103683), amended hyperparameters were found to perform better: `gamma_hyper=1e-10 pi_hyper=1e-3 pi_hyper2=1e-3 gamma_hyper=0.1 rho_hyper=10.0 rho_hyper2=10.0 --anneal-steps 20000`.

Each strain-pure set was defined as those samples where StrainFacts estimated it to be > 95% of the species. For analyses requiring estimated genotypes, we used a consensus genotype for each strain, pooling all samples in the strain-pure set. Based on this pooling, the consensus genotype for each strain was the majority allele at each position. Positions with unexpectedly high counts of the minor allele (≥ 10%)—which suggests issues with genotyping—were masked. Similarly, positions without any observed alleles were also masked in subsequent comparisons. Likewise, SNPs in reference and benchmark genotypes where neither allele was observed were masked in downstream analyses. We selected this as a more conservative approach compared to directly using the genotypes estimated by StrainFacts. All pairwise dissimilarities between inferred strain, reference, and benchmark genotypes were calculated as the masked Hamming distance, with a pseudocount of 1 added, i.e.:

$d(Gi, Gj) = \frac{P_{\Delta} + 1}{P_{*} + 1}$ where P_{Δ} is the number of positions with different allele and P_{*} is the number of unmasked positions.

Note that this measure of genetic distance is related to but not equivalent to the complement of the core genome average nucleotide identity (“ANI dissimilarity”: $1 - \mathrm{ANI}$), since it is based on only known polymorphic sites in the core genome, and the actual ANI dissimilarity—which includes many non-polymorphic sites in the denominator, as well—is likely to be much smaller.

StrainPGC

We estimated gene content for each strain with StrainPGC v0.1.0, providing the list of species marker genes from the MIDASDB, the strain pure sets derived from StrainFacts, and pangenome profiles from MIDAS as the three inputs.

StrainPGC estimates the depth of each species in each sample as the 15%-trimmed mean depth across all species marker genes, i.e., the mean depth of species marker genes excluding those genes with the 15% highest and lowest depth. Species-free samples were defined as those with an estimated species depth of < 0.0001x. Genes were selected using a depth-ratio threshold of 0.2 and a correlation threshold of 0.4 in order to strike a balance between sensitivity and specificity.

Gene family annotation

To facilitate functional interpretation, we extended the voting procedure used for the MIDASDB to EggNOG mapper annotations, which include COGs, COG categories, EggNOG OGs, KOs, and KEGG Modules. We augmented the COG categories assigned by EggNOG mapper with additional categories available from <https://ftp.ncbi.nlm.nih.gov/pub/COG/COG2020/data/>. Since annotations were performed on representative sequences for each dereplicated gene (99% ANI cluster), we first transferred specific annotations to all cluster members. Annotations within each gene (75% ANI cluster) were then counted as votes. Any annotations possessed by > 50% of member sequences were assigned to the gene family as a whole. Note that while the annotation voting for the MIDASDB, described above, operates on binary annotations (e.g., it is or is not a phage gene), this additional voting procedure was performed for individual annotations (e.g., a specific COG or AMR reference accession).

Downstream analysis

Performance benchmarking

For benchmarking, strains were excluded where the genome did not match GT-Pro SNPs from primarily a single species, or where the species's marker genes were never detected in metagenomes.

We identified gene sequences in these genomes with Prodigal v2.6.3 [@Hyatt2010] (masking ambiguous bases and using the `meta` procedure), translated them with codon table 11, and annotated them with EggNOG mapper version 2.1.10. The ground-truth annotations used to assess performance were defined as the complete set of all EggNOG OGs assigned to all genes in the ground-truth genome. These were compared to the complete set of OG annotations in each inferred strain's estimated gene content.

In order to select which inferred strain to compare to each benchmark genome, the GT-Pro genotype of the ground-truth genome was compared to all strain-pure sample consensus genotypes, and the best match was identified based on the smallest masked hamming distance. For each benchmark genome, we calculated the precision, recall, and F1 score for this best match.

Both PanPhlAn and StrainPanDA are packaged with their own pangenome databases and profiling scripts. However, in order to compare the core algorithms directly, the same MIDAS pangenome profiles were provided as input to all three tools. Both alternative tools have several parameters that control when they fail to run on low sequencing depth datasets. Since, for some species, the use of default parameter values results in a runtime exception, we adjusted these parameters to be much more lenient. For PanPhlAn, we used the flags: `--left_max 1000000 --right_min 0 --min_coverage 0`. For StrainPanDA, we made modifications to the code (see <https://github.com/bsmith89/StrainPanDA>) and used the flags `--mincov 10 --minfrac 0.9 --minreads 1e6 --minsamples 1`. We also fixed the number of latent strains to 6 using `--max_rank 6 --rank 6` for all runs. For PanPhlAn and StrainPanDA, the inferred strain with the highest F1 score was used for performance comparisons.

Inferred strain quality filtering

For analysis of the HMP2 and UCFMT datasets—but not performance benchmarking—strains were filtered to remove those likely to be low accuracy. Strains with fewer than 100 unmasked positions in their consensus genotype were included in benchmarking but excluded from all other analyses. This criterion *a priori* excludes 19 of the 627 species profiled in this work. For analyses of gene content, strains with an estimated depth of < 1x across all strain-pure samples were also excluded. Finally, strains with < 90% of species genes or with a standard deviation in the log10-transformed depth-ratio across selected genes of > 0.75 were flagged as low quality and removed.

Analysis of species and strain diversity

The species phylogeny in Fig. 2A and Fig. 3E was obtained directly from the UHGG (https://ftp.ebi.ac.uk/pub/databases/metagenomics/mgnify_genomes/human-gut/v2.0.2/phylogenies/bac120_iqtree.nwk).

For the analysis of strain distribution in the HMP2, strain depth was estimated as the product of the estimated species depth and estimated strain fraction. All strains with depth >0.1x were considered to be “present” in a sample. The number of strains in each subject was calculated as the total number of strains present in any of that subject’s samples. For shared-strain analysis (Fig. 3C), samples with fewer than 10 strains present of any species were excluded from analysis, as this removed several samples with anomalously low diversity.

Gene content was compared using the cosine dissimilarity. For comparisons between inferred strains and references, the inferred strains' gene content was first batch corrected by subtracting the difference in means (i.e., the difference in prevalence).

Pangenome Analyses

To calculate the correlation between gene prevalence in reference genomes and inferred strains we first removed genes that were very rare (<1%) in both.

Genes found in no more than one or missing from no more than one genome were excluded from clustering analysis. The remaining genes were then hierarchically clustered based on their correlation across inferred strains using the average-neighbor method at a correlation threshold of 0.9. Only clusters with more than one member were kept.

To analyze the clumping of related genes in co-occurrence clusters, we considered annotations of (1) individual KEGG modules and (2) binary classification of genes as phage and/or plasmid. For each co-occurrence cluster, we took the maximum count for any one annotation. To estimate a distribution under the null, we permuted cluster labels within species before again collecting the maximum counts across clusters. Significance was tested by comparing the number of clusters with ≥3 related annotations to the null.

For analysis of the UCFMT *E. coli* strains, shell genes and co-occurrence clusters were defined using the HMP2 inferred strains, not *de novo*.

Availability of software, code, metadata, and compute environments

Code and metadata needed to replicate our analyses and plots are available at <https://github.com/bsmith89/StrainPGC-manuscript>. Jupyter notebooks with these and extended analyses are also rendered as Supplementary Results 1-6.

Additional Details

Data Access

Reference genomes and metagenomic data analyzed for this study are available in public repositories as described in the methods.

Competing Interests Statement

The authors declare no competing interests.

Acknowledgments

This work was funded by a NHLBI grant #HL160862, the Chan Zuckerberg Biohub San Francisco, Gladstone Institutes, and the Sam Simeon Fund. BJS was supported by a Computational Innovation Postdoctoral Fellowship from the Noyce Initiative for Digital Transformation in Computational Biology and Health Data Science. JA was supported by funding from the Kenneth Rainin Foundation and the Crohn's and Colitis Foundation. The authors thank Françoise Chanut for extensive editorial support.

Author Contributions

- BJS: Conceptualization, Methodology, Software, Formal Analysis, Investigation, Writing – Original Draft, Writing – Review & Editing, Visualization
- CZ: Conceptualization, Methodology, Software, Writing – Original Draft, Writing – Review & Editing
- VD: Writing – Review & Editing, Visualization
- XJ: Writing – Original Draft, Writing – Review & Editing, Visualization
- JA: Data Curation, Writing – Review & Editing
- KP: Conceptualization, Methodology, Investigation, Resources, Writing – Original Draft, Writing – Review & Editing, Supervision, Funding Acquisition

References

Almeida A, Nayfach S, Boland M, Strozzi F, Beracochea M, Shi ZJ, Pollard KS, Sakharova E, Parks DH, Hugenholtz P, et al. 2021. A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat Biotechnol* **39**: 105–114.

Beghini F, McIver LJ, Blanco-Míguez A, Dubois L, Asnicar F, Maharjan S, Mailyan A, Manghi P, Scholz M, Thomas AM, et al. 2021. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3 eds. P. Turnbaugh, E. Franco, and C.T. Brown. *eLife* **10**: e65088.

Blanco-Míguez A, Beghini F, Cumbo F, McIver LJ, Thompson KN, Zolfo M, Manghi P, Dubois L, Huang KD, Thomas AM, et al. 2023. Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nat Biotechnol* **41**: 1633–1644.

Camargo AP, Roux S, Schulz F, Babinski M, Xu Y, Hu B, Chain PSG, Nayfach S, Kyrpides NC. 2023. Identification of mobile genetic elements with geNomad. *Nat Biotechnol* 1–10.

Carr R, Shen-Orr SS, Borenstein E. 2013. Reconstructing the Genomic Content of Microbiome Taxa through Shotgun Metagenomic Deconvolution. *PLOS Computational Biology* **9**: e1003292.

Carrow HC, Batachari LE, Chu H. 2020. Strain diversity in the microbiome: Lessons from Bacteroides fragilis. *PLOS Pathogens* **16**: e1009056.

Dimonaco NJ, Aubrey W, Kenobi K, Clare A, Creevey CJ. 2022. No one tool to rule them all: prokaryotic gene prediction tool annotations are highly dependent on the organism of study. *Bioinformatics* **38**: 1198–1207.

Florensa AF, Kaas RS, Clausen PTLC, Aytan-Aktug D, Aarestrup FM. 2022. ResFinder – an open online resource for identification of antimicrobial resistance genes in next-generation sequencing data and prediction of phenotypes from genotypes. *Microb Genom* **8**: 000748.

Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**: 3150–3152.

Gao J, Newberry M. 2024. Fractal scaling and the aesthetics of trees. http://arxiv.org/abs/2402.13520 (Accessed February 26, 2024).

Henderson G, Gudys A, Baharav T, Sundaramurthy P, Kokot M, Wang PL, Deorowicz S, Carey AF, Salzman J. 2024. Ultra-efficient, unified discovery from microbial sequencing with SPLASH and precise statistical assembly. 2024.01.18.576133. https://www.biorxiv.org/content/10.1101/2024.01.18.576133v1 (Accessed February 15, 2024).

Hu H, Tan Y, Li C, Chen J, Kou Y, Xu ZZ, Liu Y-Y, Tan Y, Dai L. 2022. StrainPanDA: Linked reconstruction of strain composition and gene content profiles via pangenome-based decomposition of metagenomic data. *iMeta* **1**: e41.

Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**: 119.

Hyatt D, LoCascio PF, Hauser LJ, Uberbacher EC. 2012. Gene and translation initiation site prediction in metagenomic sequences. *Bioinformatics* **28**: 2223–2230.

Jin X, Yu FB, Yan J, Weakley AM, Dubinkina V, Meng X, Pollard KS. 2023. Culturing of a complex gut microbial community in mucin-hydrogel carriers reveals strain- and gene-associated spatial organization. *Nat Commun* **14**: 3510.

Joglekar P, Sonnenburg ED, Higginbottom SK, Earle KA, Morland C, Shapiro-Ward S, Bolam DN, Sonnenburg JL. 2018. Genetic Variation of the SusC/SusD Homologs from a Polysaccharide Utilization Locus Underlies Divergent Fructan Specificities and Functional Adaptation in Bacteroides thetaiotaomicron Strains. *mSphere* **3**: 10.1128/mspheredirect.00185-18.

Johansson MHK, Bortolaia V, Tansirichaiya S, Aarestrup FM, Roberts AP, Petersen TN. 2021. Detection of mobile genetic elements associated with antibiotic resistance in Salmonella enterica using a newly developed web tool: MobileElementFinder. *J Antimicrob Chemother* **76**: 101–109.

Li J, Jia H, Cai X, Zhong H, Feng Q, Sunagawa S, Arumugam M, Kultima JR, Prifti E, Nielsen T, et al. 2014. An integrated catalog of reference genes in the human gut microbiome. *Nat Biotechnol* **32**: 834–841.

Lloyd-Price J, Mahurkar A, Rahnavard G, Crabtree J, Orvis J, Hall AB, Brady A, Creasy HH, McCracken C, Giglio MG, et al. 2017. Strains, functions and dynamics in the expanded Human Microbiome Project. *Nature* **550**: 61–66.

Milanese A, Mende DR, Paoli L, Salazar G, Ruscheweyh H-J, Cuenca M, Hingamp P, Alves R, Costea PI, Coelho LP, et al. 2019. Microbial abundance, activity and population genomic profiling with mOTUs2. *Nat Commun* **10**: 1014.

Miller CS, Baker BJ, Thomas BC, Singer SW, Banfield JF. 2011. EMIRGE: reconstruction of full-length ribosomal genes from microbial community short read sequencing data. *Genome Biol* **12**: R44.

Minot SS, Barry KC, Kasman C, Golob JL, Willis AD. 2021. geneshot: gene-level metagenomics identifies genome islands associated with immunotherapy response. *Genome Biology* **22**: 135.

Minot SS, Willis AD. 2019. Clustering co-abundant genes identifies components of the gut microbiome that are reproducibly associated with colorectal cancer and inflammatory bowel disease. *Microbiome* **7**: 110.

Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, Forster J, Lee S, Twardziok SO, Kanitz A, et al. 2021. Sustainable data analysis with Snakemake. https://f1000research.com/articles/10-33 (Accessed February 15, 2024).

Navarro-Garcia F, Ruiz-Perez F, Cataldi Á, Larzábal M. 2019. Type VI Secretion System in Pathogenic Escherichia coli: Structure, Role in Virulence, and Acquisition. *Front Microbiol* **10**: 1965.

Nayfach S, Rodriguez-Mueller B, Garud N, Pollard KS. 2016. An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. *Genome Res* **26**: 1612–1625.

Pakbin B, Brück WM, Rossen JWA. 2021. Virulence Factors of Enteric Pathogenic Escherichia coli: A Review. *International Journal of Molecular Sciences* **22**: 9922.

Parks DH, Chuvochina M, Rinke C, Mussig AJ, Chaumeil P-A, Hugenholtz P. 2022. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Research* **50**: D785–D794.

Plaza Oñate F, Le Chatelier E, Almeida M, Cervino ACL, Gauthier F, Magoulès F, Ehrlich SD, Pichaud M. 2019. MSPminer: abundance-based reconstitution of microbial pan-genomes from shotgun metagenomic data. *Bioinformatics* **35**: 1544–1552.

Ray S, Das S, Suar M. 2017. Molecular Mechanism of Drug Resistance. In *Drug Resistance in Bacteria, Fungi, Malaria, and Cancer* (eds. G. Arora, A. Sajid, and V.C. Kalia), pp. 47–110, Springer International Publishing, Cham https://doi.org/10.1007/978-3-319-48683-3_3 (Accessed February 16, 2024).

Rognes T, Flouri T, Nichols B, Quince C, Mahé F. 2016. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**: e2584.

Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**: 2068–2069.

Smith BJ, Piceno Y, Zydek M, Zhang B, Syriani LA, Terdiman JP, Kassam Z, Ma A, Lynch SV, Pollard KS, et al. 2022. Strain-resolved analysis in a randomized trial of antibiotic pretreatment and maintenance dose delivery mode with fecal microbiota transplant for ulcerative colitis. *Sci Rep* **12**: 5517.

Trimble WL, Keegan KP, D'Souza M, Wilke A, Wilkening J, Gilbert J, Meyer F. 2012. Short-read reading-frame predictors are not created equal: sequence error causes loss of signal. *BMC Bioinformatics* **13**: 183.

Yang C, Mogno I, Contijoch EJ, Borgerding JN, Aggarwala V, Li Z, Siu S, Grasset EK, Helmus DS, Dubinsky MC, et al. 2020. Fecal IgA Levels Are Determined by Strain-Level Differences in Bacteroides ovatus and Are Modifiable by Gut Microbiota Manipulation. *Cell Host & Microbe* **27**: 467-475.e6.

Zhang Y, Zhang H, Zhang Z, Qian Q, Zhang Z, Xiao J. 2023. ProPan: a comprehensive database for profiling prokaryotic pan-genome dynamics. *Nucleic Acids Research* **51**: D767–D776.

Zhong C, Chen C, Wang L, Ning K. 2021. Integrating pan-genome with metagenome for microbial community profiling. *Computational and Structural Biotechnology Journal* **19**: 1458–1466.

Zhou N, Jiang Y, Bergquist TR, Lee AJ, Kacsoh BZ, Crocker AW, Lewis KA, Georghiou G, Nguyen HN, Hamid MN, et al. 2019. The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome Biology* **20**: 244.

Supplementary Materials

Supplementary results and code to reproduce our analyses and figures are integrated into analysis notebooks available at <https://github.com/bsmith89/StrainPGC-manuscript>.

- Supplementary Results 1: Performance benchmarking in a synthetic community
  - `nb/analyze_benchmarking_results.ipynb`
- Supplementary Results 2: Distribution of strains in HMP2
  - `nb/analyze_distribution_of_hmp2_strains.ipynb`
- Supplementary Results 3: Diversity of inferred strains
  - `nb/analyze_hmp2_strain_diversity.ipynb`
- Supplementary Results 4: Gene prevalence in inferred strains
  - `nb/analyze_pangenome_fractions_hmp2_strains.ipynb`

- Supplementary Results 5: Co-occurrence clustering
  - `nb/analyze_gene_clusters_in_hmp2_strains.ipynb`
- Supplementary Results 6: Tracking and comparison of *E. coli* strain gene content in UCFMT
  - `nb/analyze_ucfmt_donor_strains_102506.ipynb`
- Supplementary Table 1: Details about all inferred strains in HMP2
  - `hmp2_inferred_strains_supplementary_table1.tsv`
- Supplementary Table 2: Details about gene content of E. coli strain-6 vs. strain-9 in UCFMT
  - `ucfmt_focal_strain_genes_supplementary_table2.tsv`