Strain-resolved inference of microbial gene content in fecal microbiota transplantation to treat ulcerative colitis

Byron J. Smith[1,2], Katherine S. Pollard[1,2,3]

Fecal microbiota transplantation (FMT) can be an effective therapy for the treatment of ulcerative colitis (UC) in some patients, but the importance of donor microbiota variation is unclear. Strains of a species often have dramatic differences in functional gene content.

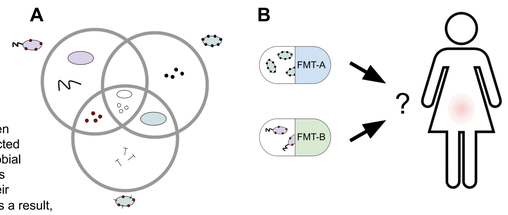


Figure 1: Expansive strain diversity within microbial species, adds an additional layer of uncertainty in attempts to replace or amend a dysbiotic microbiome through FMT. **(A)** Intraspecific variation is driven by pangenome dynamics—depicted here as a Venn diagram of microbial traits—where strains of a species may differ by 20% or more of their functional gene repertoire. **(B)** As a result, it is not always clear that the fecal microbiota of two different donors will have the same outcome for patients, even when the same species are present..

The Strain-Partitioned Gene Content (StrainPGC) method combines data across multiple metagenomic samples, harnessing correlations with species depth to confidently assign genes to each strain.

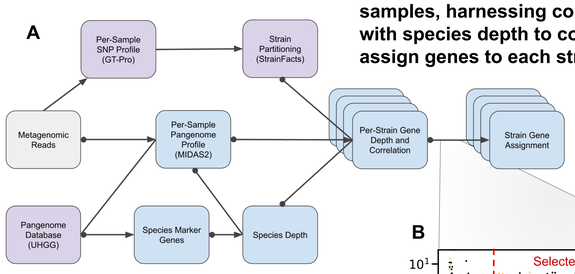


Figure 2: Graphical summary of the StrainPGC method illustrating **(A)** its inputs, key intermediates, and outputs. StrainFacts deconvolves GT-Pro SNP profiles in order to identify individual strains shared across multiple samples as well as the mixing of multiple strains in individual samples. StrainPGC then integrates pangenome profiles from one or more samples with the same strain, and infers a gene repertoire—visualized in **(B)**—based on both normalized depth (Y-axis) and correlations in depth across samples (X-axis). Subsequent quality control then selects strains with accurate inferences.

StrainPGC improves both the precision and recall of gene content inferences relative to comparable tools.

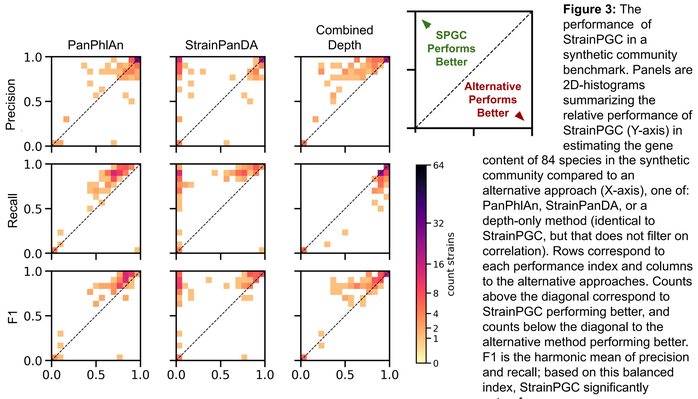


Figure 3: The performance of StrainPGC in a synthetic community benchmark. Panels are 2D-histograms summarizing the relative performance of StrainPGC (Y-axis) in estimating the gene content of 84 species in the synthetic community compared to an alternative approach (X-axis), one of: PanPhlAn, StrainPanDA, or a depth-only method (identical to StrainPGC, but that does not filter on correlation). Rows correspond to each performance index and columns to the alternative approaches. Counts above the diagonal correspond to StrainPGC performing better, and counts below the diagonal to the alternative method performing better. F1 is the harmonic mean of precision and recall; based on this balanced index, StrainPGC significantly outperforms the three alternatives (Wilcoxon signed-rank test, not shown) in predicting the true gene content of strains in the synthetic community.

In *Escherichia coli* (shown here), as well as many other species, gene content similarity decays quickly with core genome divergence, demonstrating the importance of strain-resolved analyses.

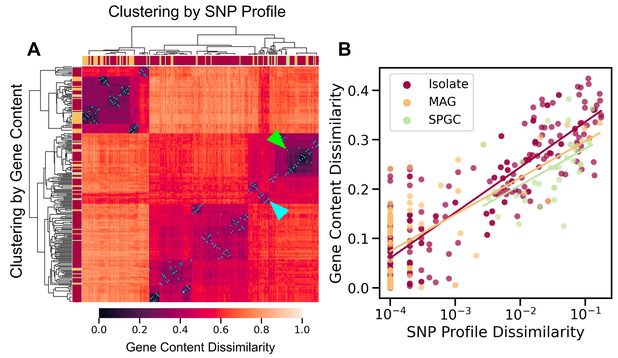
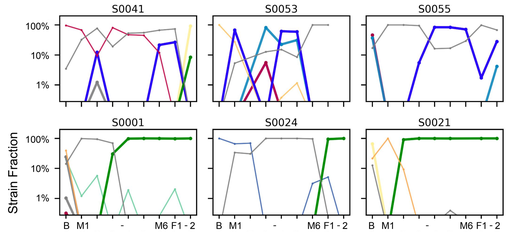


Figure 4: Pairwise relationship between SNP profile and gene content dissimilarity in both reference and StrainPGC-inferred *E. coli* genomes. **(A)** Heatmap depicting the degree of gene content dissimilarity, calculated as a weighted cosine-dissimilarity. Columns are ordered based on the average-neighbor, agglomerative clustering tree calculated on SNP profile dissimilarities drawn above. Rows are ordered by a parallel agglomerative clustering, but based on gene dissimilarity, instead. Colors along columns and rows match the legend in (B) and indicate the source of the genome (isolates and MAGs derived from the UHGG and annotated in the MIDAS2 reference database). Entries for genomes compared to themselves are highlighted in bright teal. As a result, clusters of teal points in dark foci suggest clusters of genomes where SNP profiles and gene content are both very similar, and larger dark patches demonstrate higher-level population structure. StrainPGC genomes recapitulate the known diversity of *E. coli* strains based on reference genomes, with one or more newly inferred strains found in most of the intraspecific sub-groups. The bright green and bright blue arrows indicate strains identified in donors D44 and D97, respectively, which were transmitted to FMT recipients as an experimental treatment for UC. **(B)** Scatter plot showing the minimum SNP dissimilarity to a reference sequence (X-axis), as well as the corresponding gene content dissimilarity (Y-axis). Points correspond to an individual reference or StrainPGC-inferred genome, and are colored to indicate their source. Lines depict a multiple linear regression predicting gene content dissimilarity based on SNP dissimilarity (log transformed in the regression), along with terms for genome source and the interaction between the two. Given the generally higher SNP dissimilarity between StrainPGC genomes and existing references, it is not surprising that their gene content also differs markedly. That novel strains were identified with highly divergent gene content, even in a species as well-studied as *E. coli*, emphasizes the critical importance of strain-level understanding in the analysis of microbiomes.

Donors contribute strains with distinct functional repertoires to patients in a clinical trial of FMT for UC.

Figure 5: Longitudinal tracking of *E. coli* strain composition in patient feces. Panels reflect time-series collected from a subset of FMT recipients. Samples were collected at baseline and before initial FMT (B on X-axis), before weekly maintenance doses (M1 - M6), and at two follow up times (F1 - 2). Colored lines reflect individual strains, and grey lines a sum of minor strains. Thick lines indicate strains also detected in donor samples. The two rows of panels each come from different donors, D44 and D97.



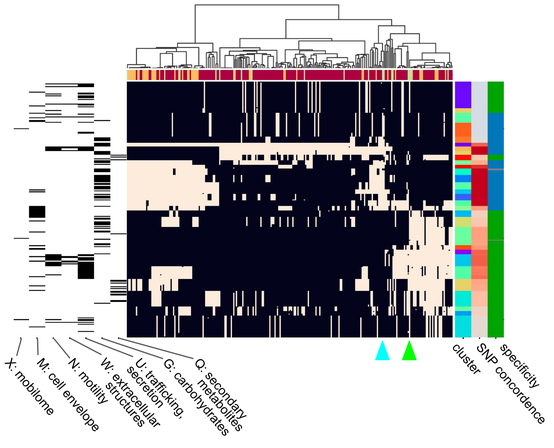


Figure 6: Variable gene content differentiating donor strains. The central panel depicts the presence and absence (light and dark colors, respectively) for a selection of genes in both reference and inferred *E. coli* genomes. Identically to Figure 4A, columns are ordered by SNP profile clustering and the top row of colors indicates the genome source (see legend in Figure 4B) The bright green and bright blue arrows indicate columns corresponding to the dominant strain identified in donors D44 and D97, respectively. Columns to the left and right of the heatmap depict several features of the genes. Black bars in columns to the left of the heatmap indicate gene annotations to a subset of COG categories, as labeled. On the right, the column of bright colors labeled “cluster” denote distinct groups of genes that are nearly always present in the same genomes, (thereby suggesting a functional interaction). The subsequent column, labeled, “SNP concordence”, indicates whether the distribution of each gene across genomes is concordant with SNP profile similarity (brighter red colors are shown for genes where similar genomes are more likely to both share or both lack it). Finally, blue and green in the last column indicates strain specificity: genes found in the strain from D44 and missing in D97 are blue, and green for the converse.

Variable gene content is enriched in functions relevant to human health, and this was reflected in differences between donor strains in this study.

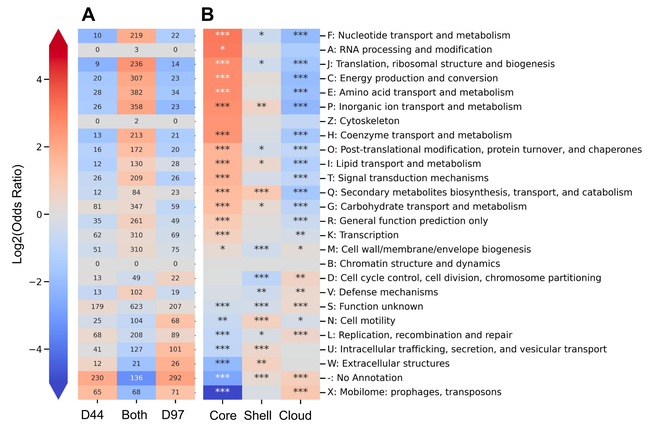
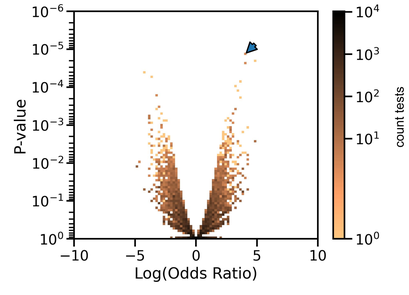


Figure 7: Functional category enrichment across pangenome fractions and specific donor strains in StrainPGC-inferred *E. coli* genomes. In both heatmaps **(A and B)** red and blue colors indicate categories enriched and depleted (log odds ratio), respectively, in various gene subsets (columns) given the functional category of its EggNOG annotation (rows). In **(A)** genes are partitioned to the 3400 genes shared versus the 779 and 988 genes specific to the the dominant *E. coli* strain from D44 and D97, respectively, two donors in an experimental study of FMT for UC. Numbers in cells indicate the number of genes in that partition with the specified COG category. In **(B)** genes are partitioned into core, shell, and cloud fractions based on their prevalence in all 33 novel *E. coli* genomes identified in the HMP2 dataset—core: >95%, shell: 10-95%, cloud: <10% prevalence. Markers indicate the significance of the result (Fisher’s exact test; *: p<0.05, **: p<1e-3, ***: p<1e-5).

Expanding microbiome-wide association studies (MWAS) to incorporate strain-level resolution, has the potential to reveal key functional links to human health and disease.

Figure 8: Strain-informed, microbiome-wide association study on inflammatory bowel disease phenotypes across 234 species and 131,470 genes. For each species, genes are assigned across HMP2 study subjects based on whether they posses a strain inferred to encode that gene. Genes with between 25% and 75% prevalence across subjects were tested for enrichment in UC, Crohn’s disease, or non-IBD control patients (Fisher’s exact test). The volcano plot, visualized as a 2D-histogram, summarizes the distribution of effect sizes (log odds ratio) and P-values across all gene-by-diagnosis pairs. A blue arrow indicates the two most significant hits, both genes in *Bacteroides xylanisolvens*. Notably, one of these is annotated as “Carbohydrate esterase, sialic acid-specific acetylesterase”, suggesting a potential role in mucin degradation.



Affiliations:

[1]The Gladstone Institute of Data Science and Biotechnology [2]Chan Zuckerberg Biohub
[3]University of California, San Francisco, Department of Epidemiology and Biostatistics

Acknowledgements:

This work was supported by funding from CZ Biohub and a Computational Innovation Post-doctoral Fellowship from the UC Noyce Initiative for Digital Transformation in Computational Biology & Health. Xiaofan Jin contributed the synthetic community metagenomes. Chunyu Zhao assisted with MIDAS2 references and pangenome profiling.

@ByronJSmith

Byron.Smith @gladstone.ucsf.edu

Poster PDF