# StrainPGC: Resolving strain-level gene content variation from large, metagenomic datasets

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

## 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.

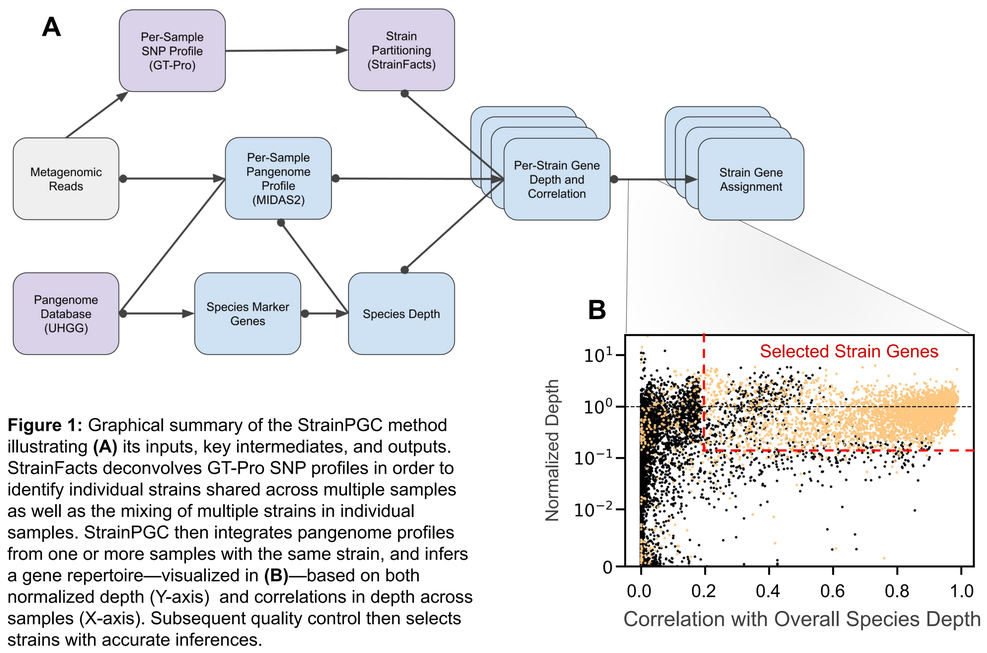



**Figure 1:** Graphical summary of the StrainPGC method illustrating **(A)** its inputs, key intermediates, and outputs. StrainFacts deconvolves GT-Pro SNP profiles in order to identify individual strains shared across multiple samples as well as the mixing of multiple strains in individual samples. 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.

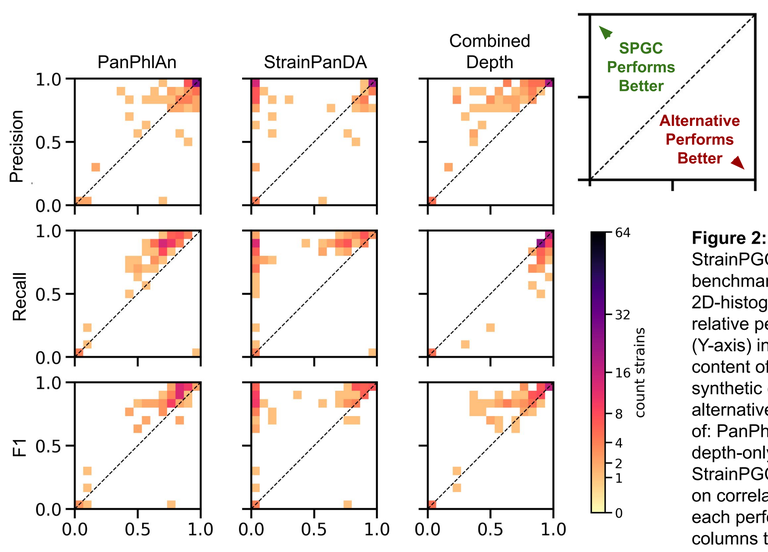



**Figure 2:** The performance of StrainPGC in a synthetic community benchmark. Panels are 2D-histograms summarizing the relative performance of StrainPGC (Y-axis) in estimating the gene content of 84 species in the synthetic community compared to an alternative approach (X-axis), one of: PanPhlAn, StrainPanDA, 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.

## Applying StrainPGC to the HMP2 metagenomic dataset reveals strain-specific gene repertoires corresponding to thousands of novel genomes.

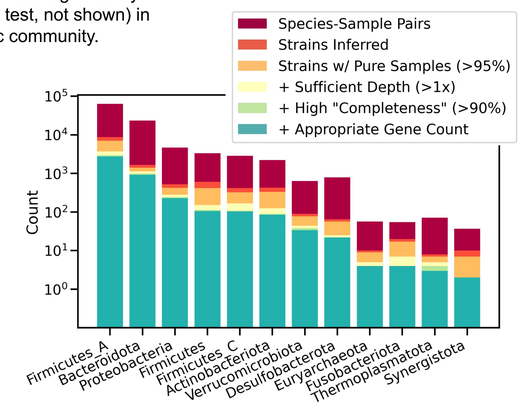



**Figure 3:** Potential and realized strain diversity inferred by StrainPGC in the HMP2 collection. Among more than 1300 metagenomes, each with tens or hundreds of species, StrainPGC ultimately identifies 4292 strains that make it past final quality filters. These include representatives of 1 archaeal and 11 bacterial phyla,

## 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.

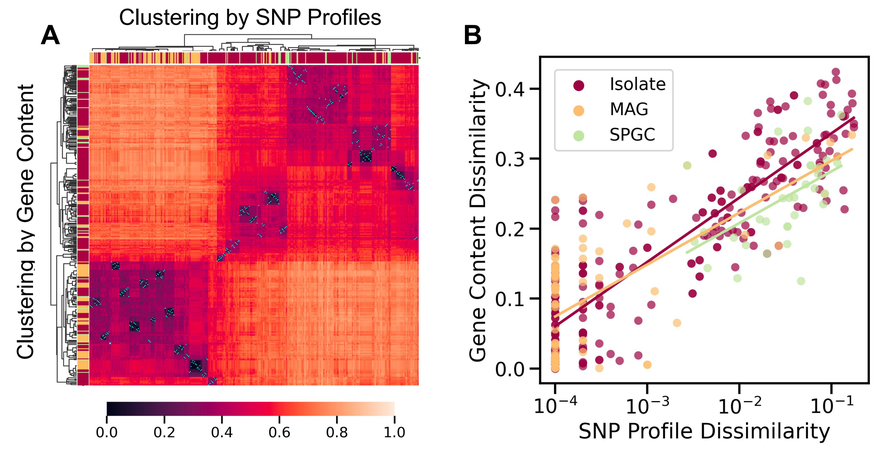



**Figure 4:** Pairwise relationship between SNP profile and gene content dissimilarity in both reference and StrainPGC-inferred *E. coli* genomes. **(A)** Heatmap depicting the degree of gene content dissimilarity, calculated as a weighted cosine-dissimilarity. Columns are ordered based on the average-neighbor, 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. **(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.

## Inferred genomes recapitulate and expand on our understanding of pangenome dynamics: variable gene content is enriched in functional annotations with potential relevance to human health and does not always coincide with core genome SNPs.

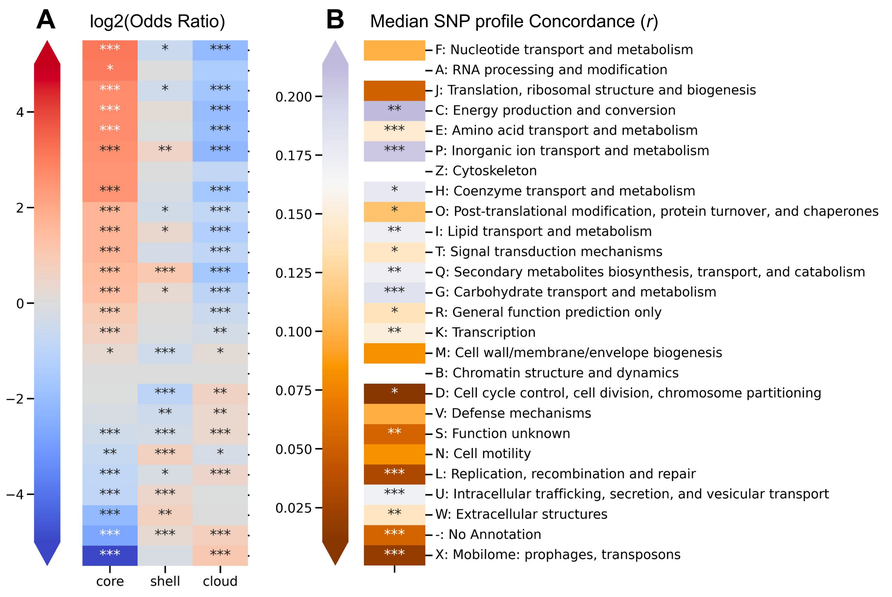
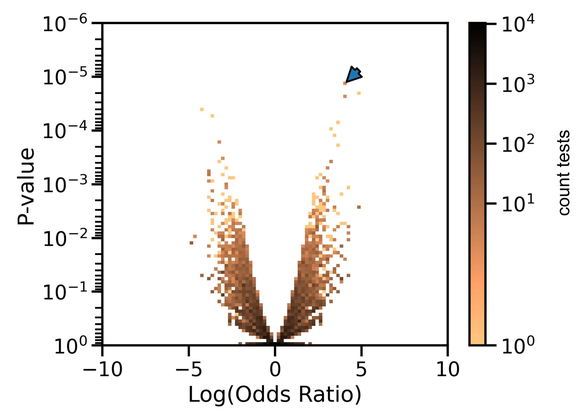



**Figure 5:** Functional category enrichment and SNP profile concordance in StrainPGC-inferred *E. coli* genomes. **(A)** Heatmap of COG category enrichment colored by the log odds ratio of a gene being found in that pangenome fraction given the functional category of its EggNOG annotation. Genes are partitioned into core, shell, and cloud fractions based on their prevalence in 33 novel *E. coli* genomes—core: >95%, shell: 10-95%, cloud: <10% prevalence. Red and blue colors indicate categories enriched and depleted, respectively, in a particular pangenome fraction. Markers indicate the significance of the result (Fisher's exact test). **(B)** Heatmap of SNP profile concordance colored by the median score for shell pangenome genes in each COG category. Concordance scores for each gene are calculated as the Pearson's correlation coefficient of all pairwise SNP dissimilarities versus difference in that genes presence. Purple and orange colors indicate categories with shell genes that are generally more concordant or discordant, respectively, than the overall gene set. Markers indicate the significance of this result (Mann-Whitney U test) **(A, B)** The "No Annotation" category indicates genes that are either not annotated by EggNOG mapper or annotated but without a COG category. Markers indicate significance level (*: 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 6:** 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 ulcerative colitis, Crohn's disease, or non-IBD control patients (Fisher's exact test). This 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".

**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

**@ByronJSmith**

**Byron.Smith@gladstone.ucsf.edu**

**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.

CZ BIOHUB

GLADSTONE INSTITUTES
SCIENCE OVERCOMING DISEASE

UCSF
University of California San Francisco

**Poster PDF**