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

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

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

Metagenomics has greatly expanded our understanding of microbiomes by revealing the vast diversity within and across microbial communities. Even within a single species, different strains can have highly divergent gene content, affecting traits such as antibiotic resistance, cofactor synthesis, and biofilm formation. However, it is challenging to identify variable gene content in species with few or no cultured representatives. Methods that harness metagenomic data to resolve strain-level differences in functional potential will be crucial for understanding the causes and consequences of intraspecific diversity.

To overcome this limitation, here we describe a novel computational method that integrates pangenome profiling with strain tracking based on single-nucleotide polymorphisms. By combining data from multiple samples, we identify robust correlations in sequence abundance in order to confidently assign genes to individual genomes. Crucially, by incorporating tracking of individual strains, we expand this approach to also capture variable gene content, which would otherwise be obscured by inconsistency across samples. Our approach enables the reconstruction of gene content at a strain-level resolution and can be applied to large collections of metagenomes.

We validate our method using a realistic, synthetic community and find that it outperformed standard approaches for gene content determination based on single samples and sequence abundance alone. Applying this approach to a large collection of stool metagenomes from inflammatory bowel disease patients and healthy controls, we catalog extensive gene content variation across hundreds of species and thousands of strains. Notably, many of the strains we identify are substantially different from any previously included in reference databases. In a focused survey of *Escherichia coli* strains, we find that, relative to the core genome, variable genes are enriched in functions relevant to their niche in the human gut, such as motility, extracellular structures, and defense mechanisms.

By incorporating strain tracking and integrating information across multiple samples, our approach improves the accuracy of gene content determination from metagenomes, and enhances its resolution to the level of individual strains. We demonstrate that variation in functional potential is ubiquitous across bacterial species residing in the human gut. These findings motivate and facilitate the continued exploration of strain diversity across complex, uncultured, microbial communities.