Strain-resolved inference of microbial gene content in 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]University of California, San Francisco, Department of Epidemiology and Biostatistics, San Francisco, CA [3]Chan Zuckerberg Biohub, San Francisco, CA

Among members of the human microbiome, strains of a species may differ in their resistance to antimicrobial therapies, metabolism of pharmaceutical products, and interactions with the host immune system, among many other relevant traits. Evaluating the functional potential encoded in each genome is a first step in predicting the ecological, evolutionary, and health impacts of strains. Given the availability of extensive, publicly available, metagenomic datasets, methods that accurately infer the gene content of strains in high-throughput sequencing data have the potential to greatly improve our understanding of the extent and consequences of intraspecific diversity.

Unfortunately, most existing approaches—including those based on *de novo* assembly or genomic references—are limited by imprecise alignment of reads to reference sequence, low coverage of less abundant genomes, and extensive strain diversity both across and within biological samples. In this work we overcome all three of these limitations by developing a novel, reference-based method that leverages strain abundances inferred by scalable metagenotype deconvolution paired with pangenome profiling to identify gene families that can be confidently attributed to each strain across multiple samples. As a proof of concept, we apply our approach to the HMP2 metagenomic dataset, which includes more than 1300 samples from more than 100 subjects, and reconstruct gene family content for 17 uncultured strains of *Escherichia coli*, an important, prevalent, but often low-abundance species in the human gut microbiome.

As expected, among inferred genomes, some gene families are consistent and others are variable across strains. We find that the prevalence of gene families in our inferences and those observed in reference genome collections are highly concordant. We also identify functional annotations that are more abundant in the core and variable sets. Gene families without functional annotations are heavily enriched in the variable fraction, suggesting that biochemical study of common lab strains may not be sufficient to predict the physiology of *E. coli* found in the wild. Our method is designed to be applied to large collections of metagenomes, and can be easily extended to a diverse set of species. Validation of our method on simulated data is ongoing. This work establishes the value of strain-resolved genomic inference with metagenomic data and enables future studies on the physiological impact of gene content variation in microbial populations.