Scalable strain reconstruction and tracking in metagenomic data with fuzzy genotypes

Byron J. Smith[1,2], Xiangpeng Li[3], Zhou Jason Shi[1,4], Adam R. Abate[3,4], Katherine S. Pollard[1,2,4]

[1]Gladstone Institute of Data Science and Biotechnology; [2]Department of Epidemiology and Biostatistics; [3]Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco; [4]Chan Zuckerberg Biohub

Strain-level variation in microbial traits such as antibiotic resistance and drug metabolism have profound impacts on their hosts, yet standard methods for studying the microbiome are often limited to species-level taxonomic resolution. What's more, while genome databases are nearing a complete catalog of species commonly inhabiting the human gut, their representation of intraspecific diversity is lacking for all but the most abundant and frequently studied taxa. Shotgun metagenomic data can in principle be used to characterize and track strains, but admixture and low sequence coverage present significant challenges. One solution is statistical deconvolution of metagenotypes from many samples into strain genotypes and their relative abundances. Existing implementations of this approach are promising, but cannot scale to the enormous quantities of data in publicly available shotgun metagenomes.

Here we introduce StrainFacts (https://github.com/bsmith89/StrainFacts), a method for strain deconvolution that enables inference across thousands of metagenomes. The key advantage of this approach is a "fuzzy" genotype approximation that makes the model fully differentiable so that it can be optimized by gradient descent. By incorporating sparsity-inducing priors on latent parameters, our inferences of strain genotypes and abundances are accurate and interpretable.. We validate StrainFacts with extensive simulation as well as single-cell genomic sequencing from human stool samples, and show that it is often an order of magnitude faster than previous methods. We infer *Agathobacter rectalis* strain genotypes from a collection of more than 10,000 publicly available shotgun metagenomes, and find numerous strain clusters that segregate dramatically by geography, suggesting that the physiology of this species may not be uniform across human populations. Likewise, in an analysis of *Escherichia coli* genotypes, we find that linkage disequilibrium between polymorphic sites falls off sharply with genomic distance, consistent with a high rate of recombination in the species.