Strain-resolved bacterial genome reconstruction in large, metagenomic datasets

Byron J. Smith, University of California, San Francisco

Noyce Symposium 2022

Project Statement

- Disease-causing bacteria (e.g. enterohemorrhagic *E. coli*) may differ by just one or a few genes and are difficult or impossible to tell apart from benign members of the same species when using common tools for microbiome analysis.
- No bacterial pathogens have yet been found that lead to inflammatory bowel disease (IBD) and other common conditions.
- **Might a subset of strains within species be contributing to these diseases?**
- Publicly available collections of human microbiome data, in particular shotgun metagenomic sequence, are quickly growing, and will enable new searches for microbial drivers of disease.

Project Goals

- I seek to **characterize bacterial, strain-level diversity within and across the human population.**
- I will develop tools to track strains and to identify differences in gene content—and therefore functional potential—between strains.
- These will enable a search for **previously undetected roles of gut bacteria in health and disease**.

Methods

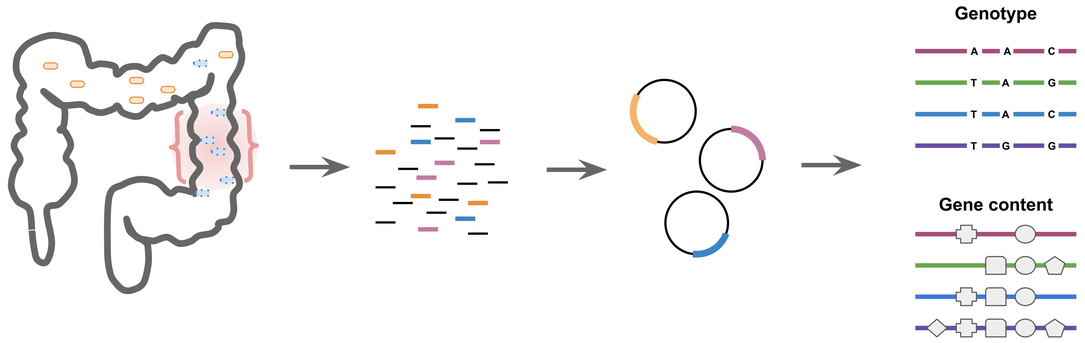
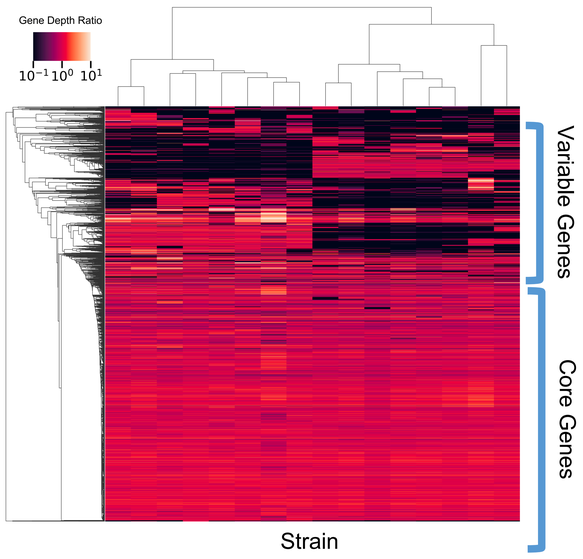


Figure 1: Using a large collection of microbiome samples from patients with IBD, shotgun metagenomic reads are matched to existing reference databases. By combining data from multiple samples that share a strain-specific genotype "signature", I can overcome previous limitations and identify gene content for each strain.

Results & Impact

Figure 2: My pipeline successfully identifies both core and variable gene content (here shown for *Escherichia coli*. Variable genes have the potential to drive different effects on the host, which would have gone undetected in traditional microbiome analysis.



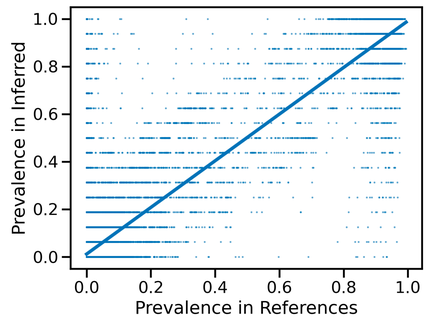


Figure 3: For this well studied species, gene prevalence is very close to what is seen in reference genomes (Pearson r = 0.92), supporting the validity of our method.

Variable gene content is enriched with genes that affect bacterial physiology in ways that may be relevant to human health: e.g. motility, cell wall composition, anti-phage defense, etc.

Challenges & Opportunities

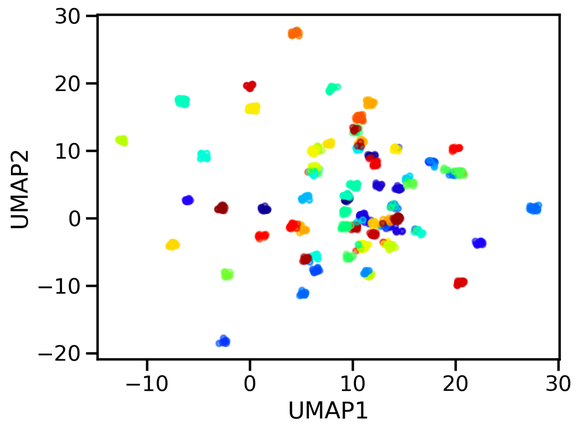


Figure 4: For many bacteria, intraspecific diversity is so high that we see a unique strain for each subject in the study. This UMAP ordination—points are metagenotypes for *Faecalibacterium prausnitzii* colored by which study subject contributed the sample—demonstrates that each subject can be distinguished by the genotype fingerprint of just one species. This remarkable diversity limits our ability to identify disease associations, because strains are highly confounded with inter-subject variability.

While this level of diversity makes direct association studies challenging, now that I can identify the gene content of strains, searching this functional potential for associations is now possible.

Acknowledgements

- Advisor: Dr. Katie S. Pollard
- Pollard Lab, especially Chunyu Zhao, Jason Shi
- University of California, San Francisco, Department of Epidemiology and Biostatistics
- The Gladstone Institute for Data Science and Biotechnology