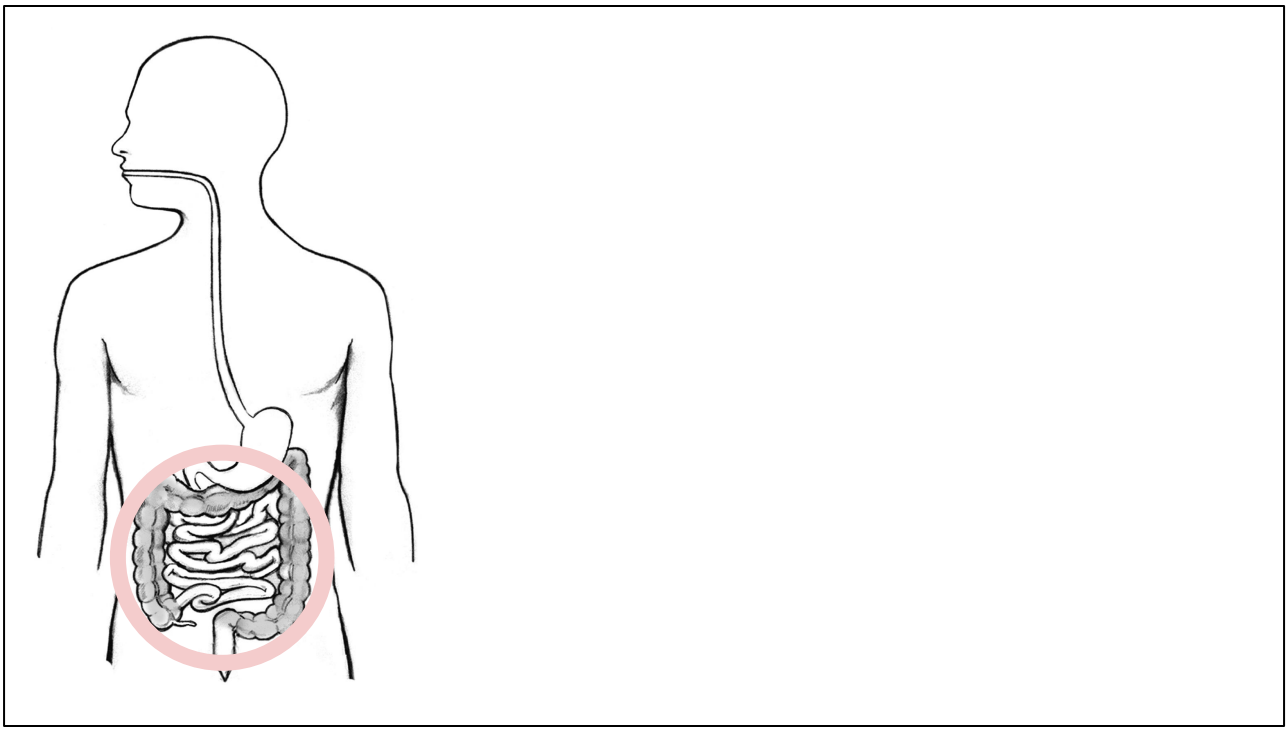
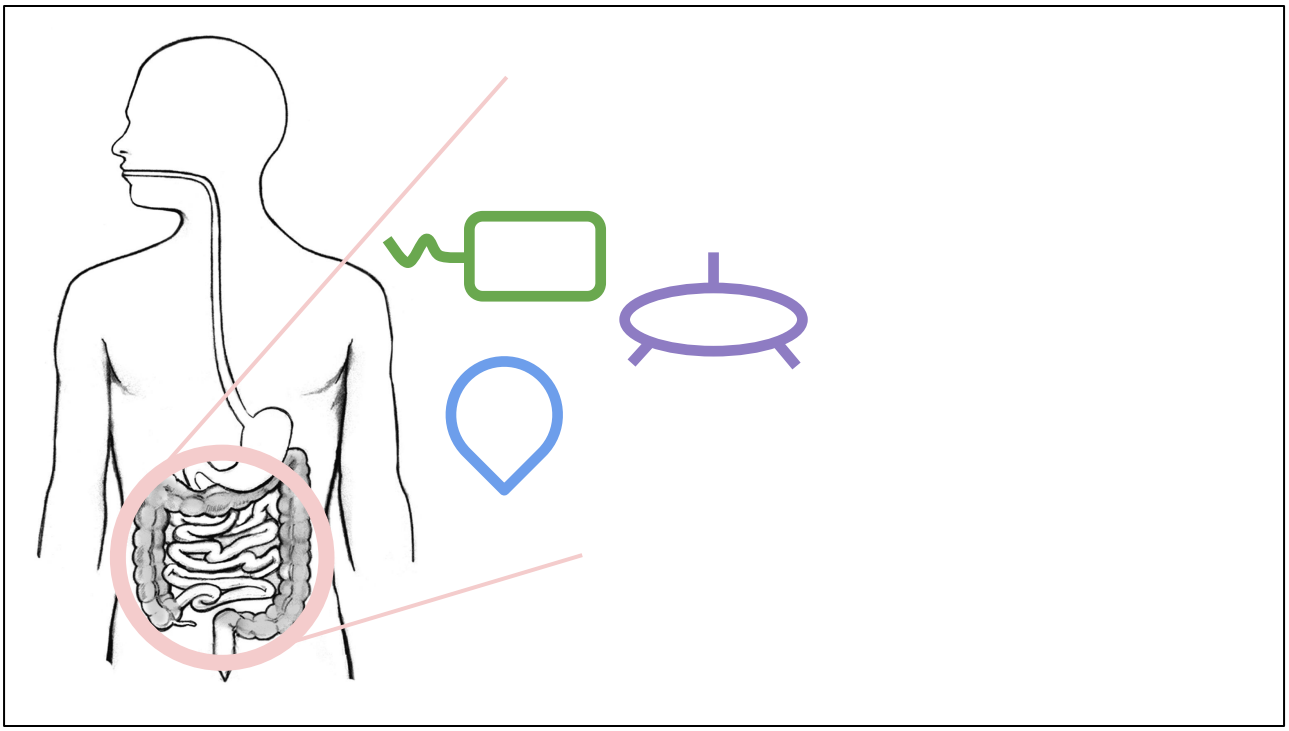
Improved methods for strain-specific genome reconstruction

Byron J. Smith
Biohub Inter-lab Confab
2019-12-09

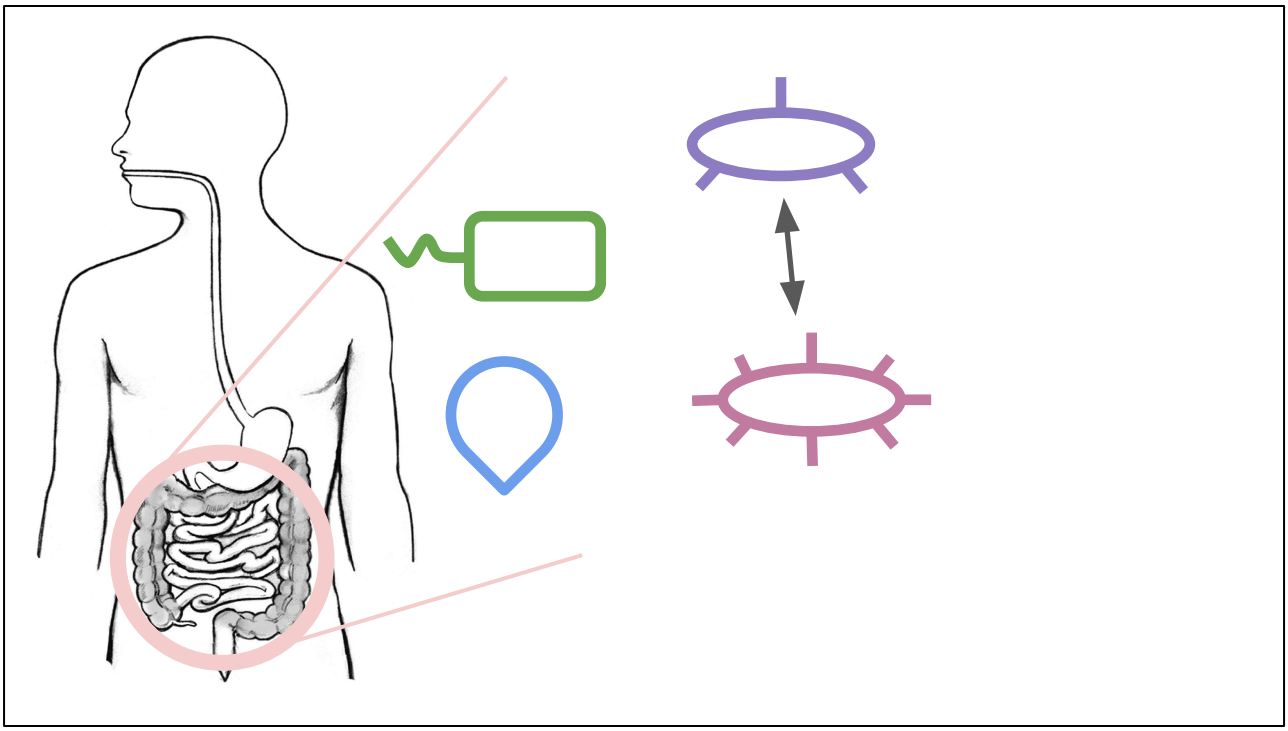
(Timing queues: https://docs.google.com/document/d/12lpg-SIbWsL2BW3xHw6avvxkDNA-stQwLDLI1SNYhRY/edit?usp=sharing)



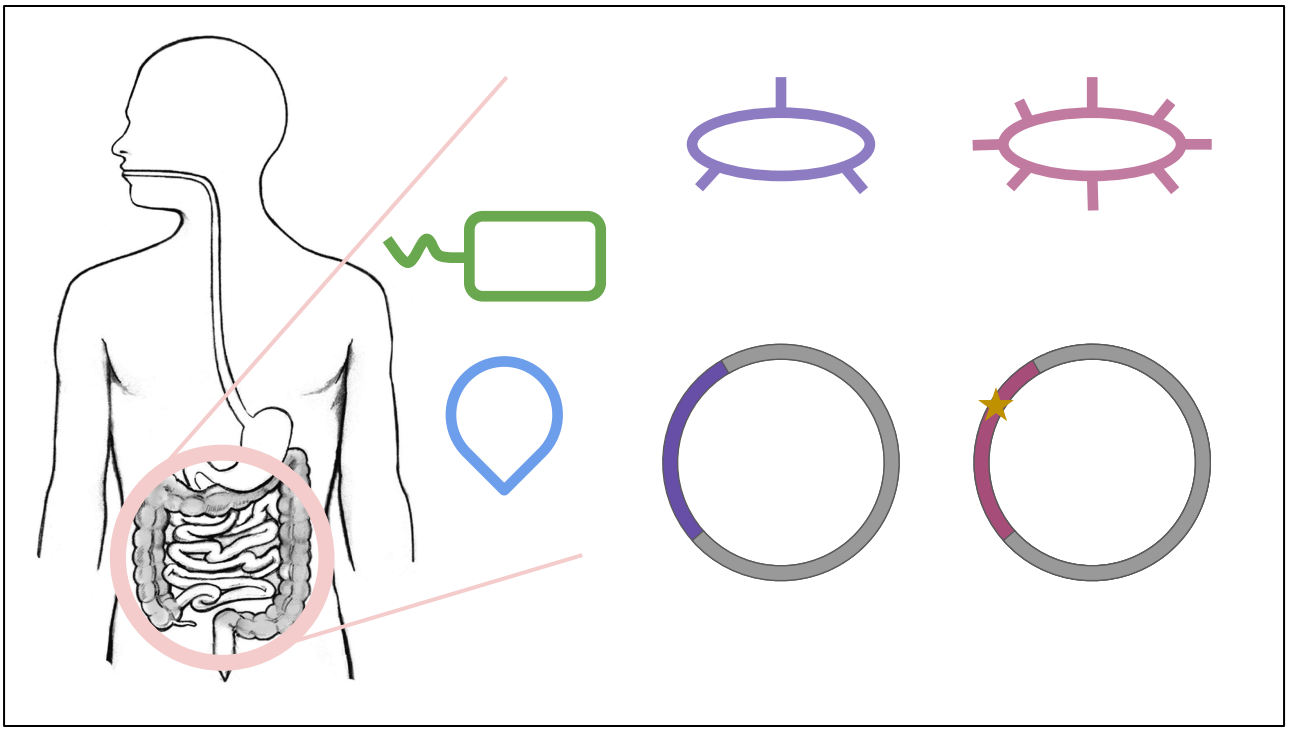
- The human microbiome has important impacts on our health and is on a long list of complex microbial communities that are not yet understood.



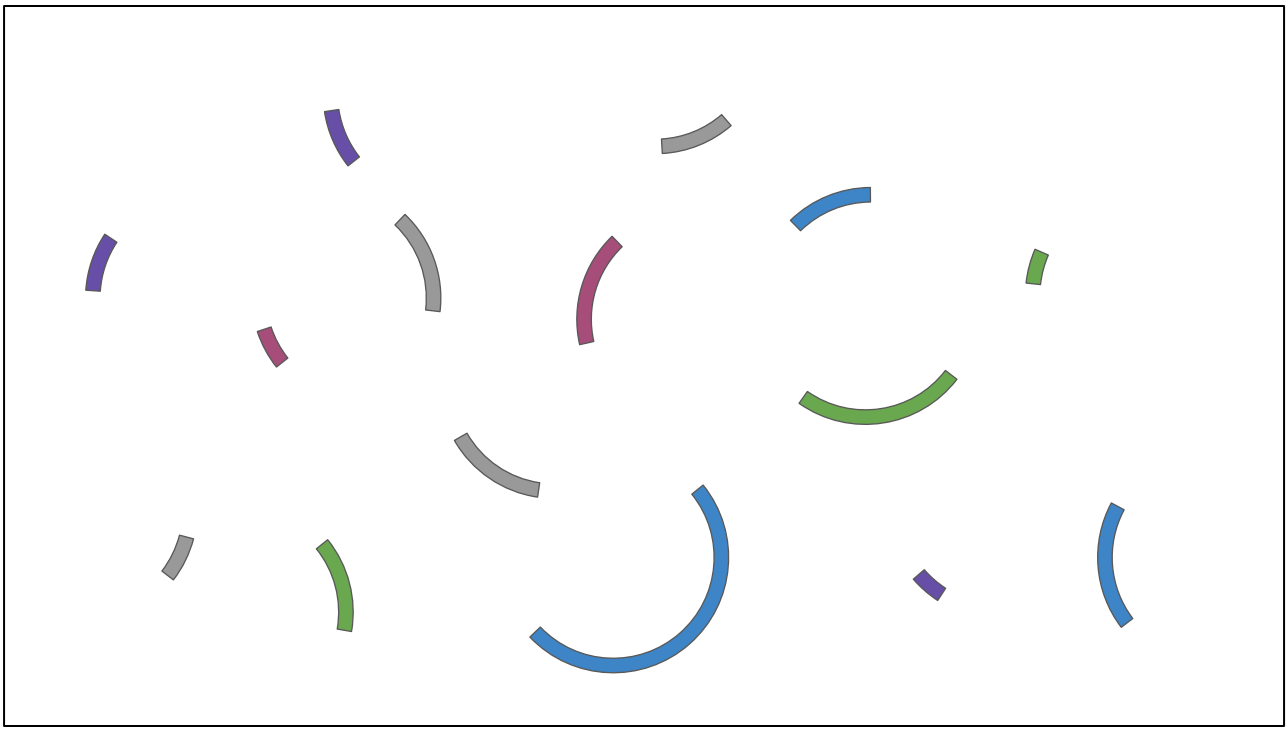
- The last decade of human microbiome research focused on identifying associations between community composition and host traits
- but has largely failed to identify molecular mechanisms that explain these associations or to accurately predict how perturbation of the microbiome will affect health.
- Advances in microbiome science have improved the taxonomic resolution of community surveys, potentially enabling a better understanding of bacterial function.



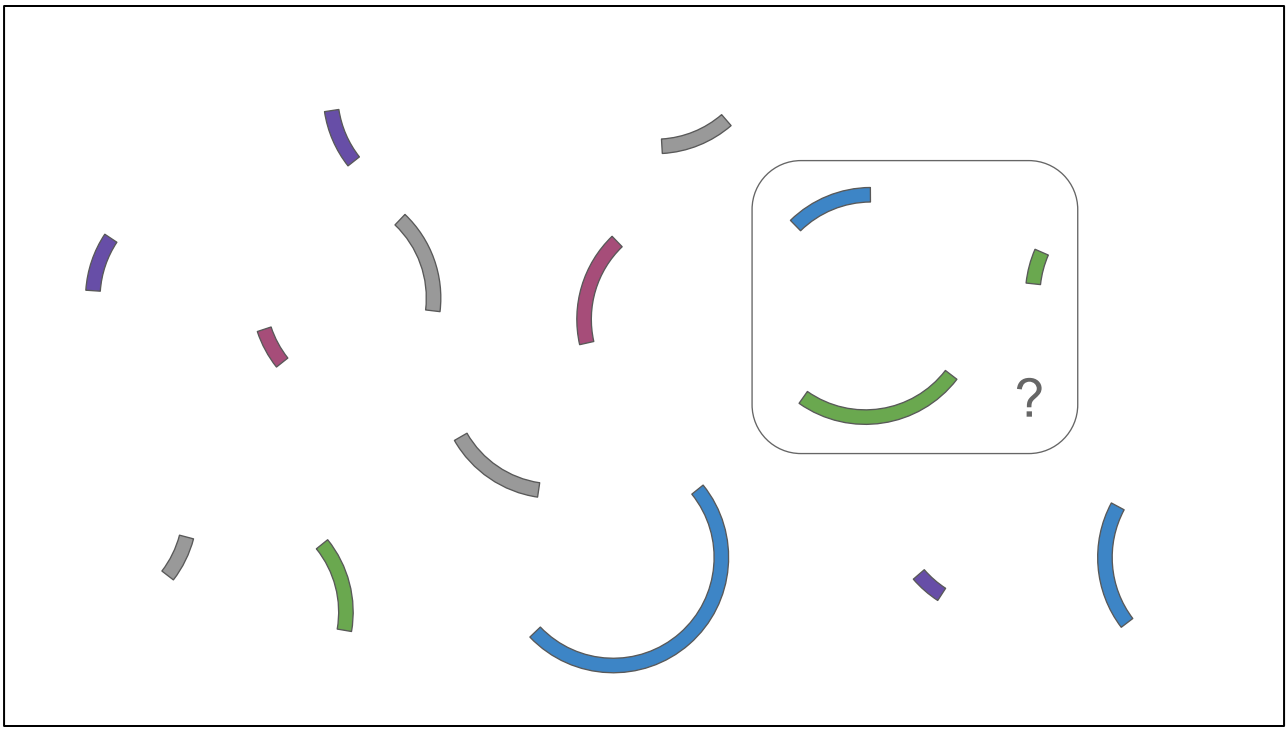
- Unfortunately, classification, even at the species level, may be blind to functional differences between closely related bacteria



- For many bacterial lineages, only a small number of genes are found in the “core genome” (here in grey), possessed by all members of a species.
- while, genes controlling things like toxin production, antibiotic resistance, metabolism, and immune system evasion are often found to be part of the “variable genome”, and can be gained and lost in surprisingly short evolutionary timescales.
- There are undoubtedly many more, medically important examples of this phenomenon to be discovered.

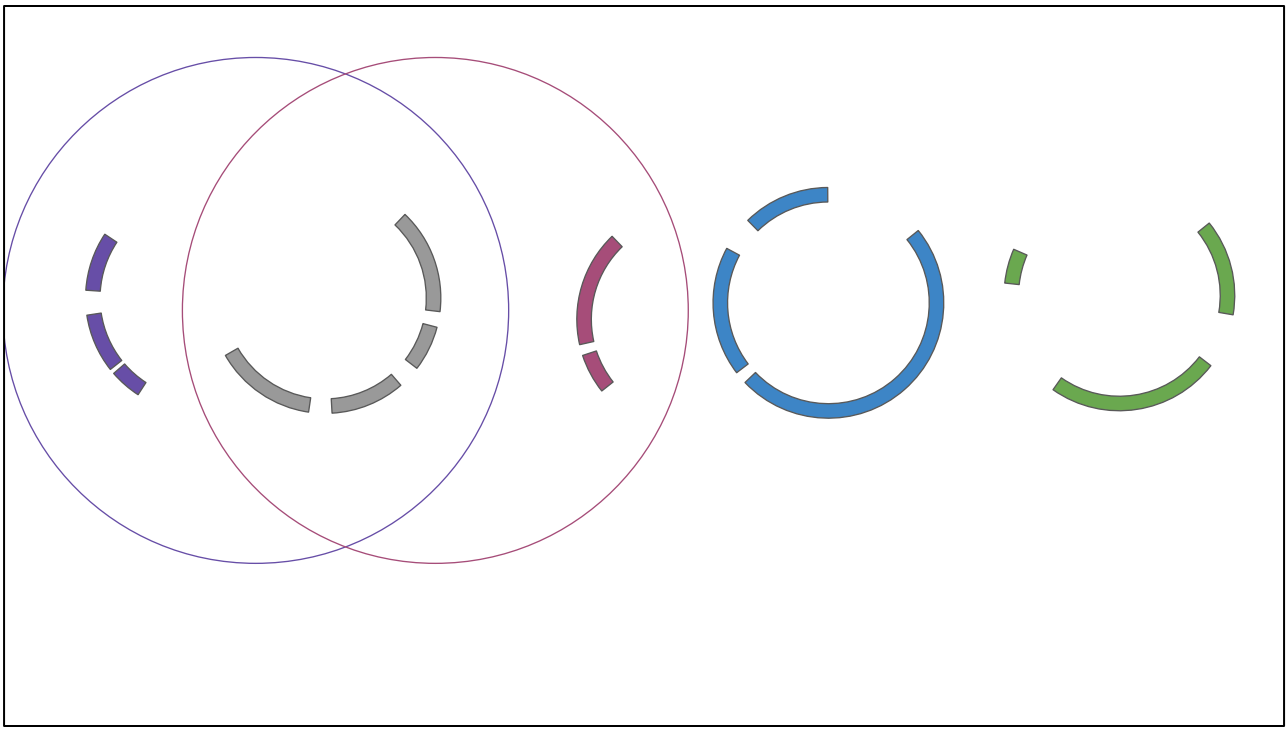


- Recently it became possible to reconstruct genomes directly from metagenomic reads by first assembling, and then combining these assembled fragments into genomic “bins” based on correlated mapping coverage across samples.

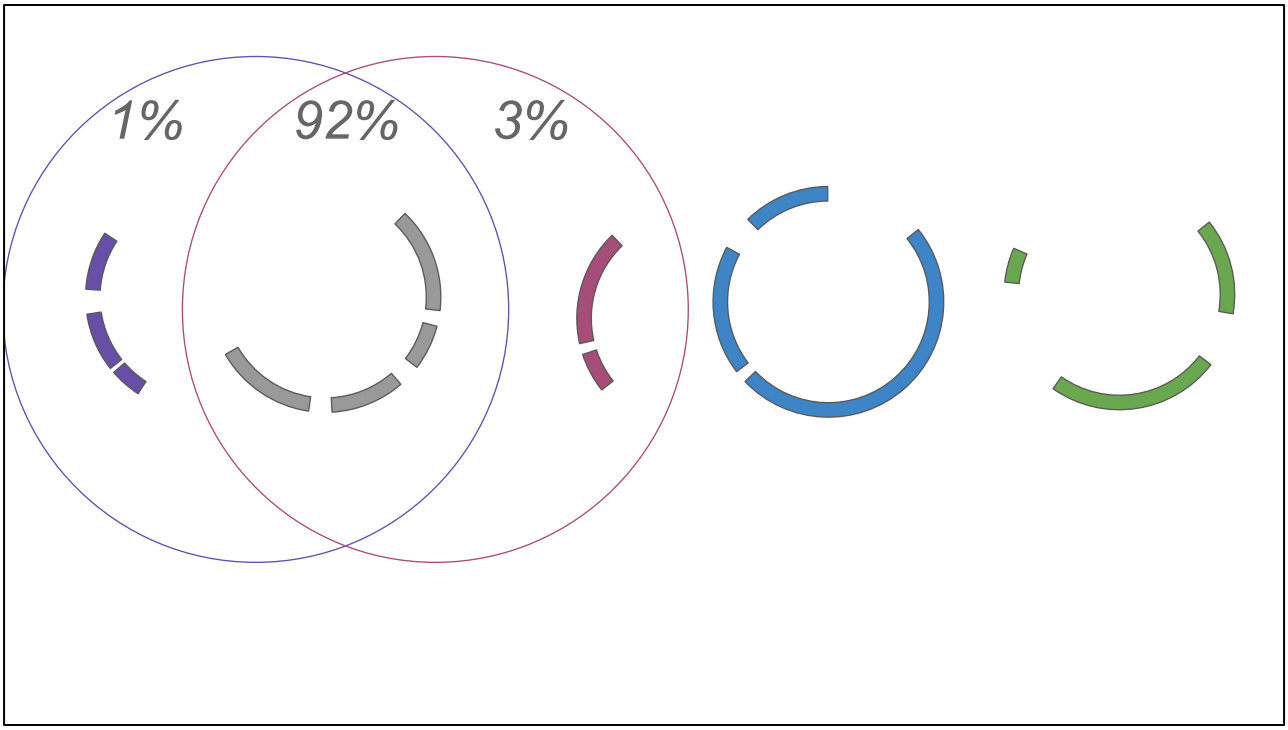


- Unfortunately, binning requires either careful, manual refinement or can easily generate erroneous genomes.

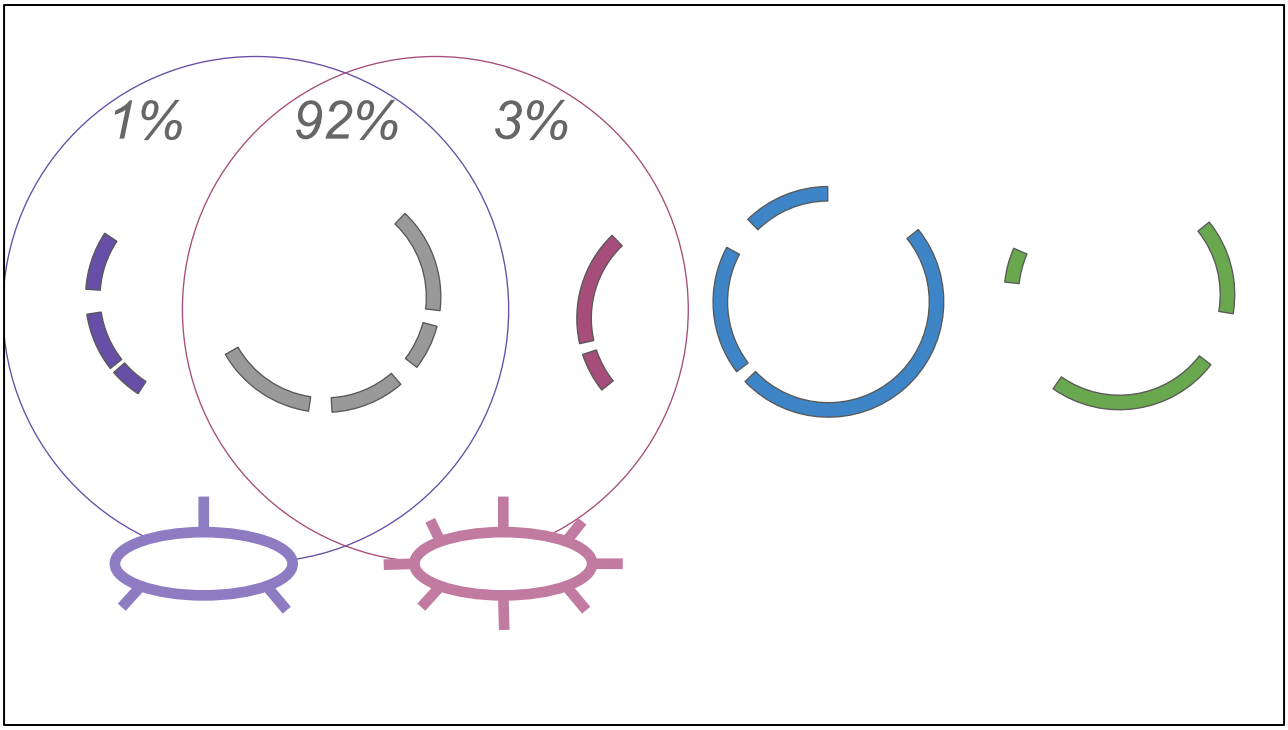
- Taking a conservative approach, our lab and others have generated large databases of reconstructed genomes from publicly available metagenomic data, providing an important resource for future studies.



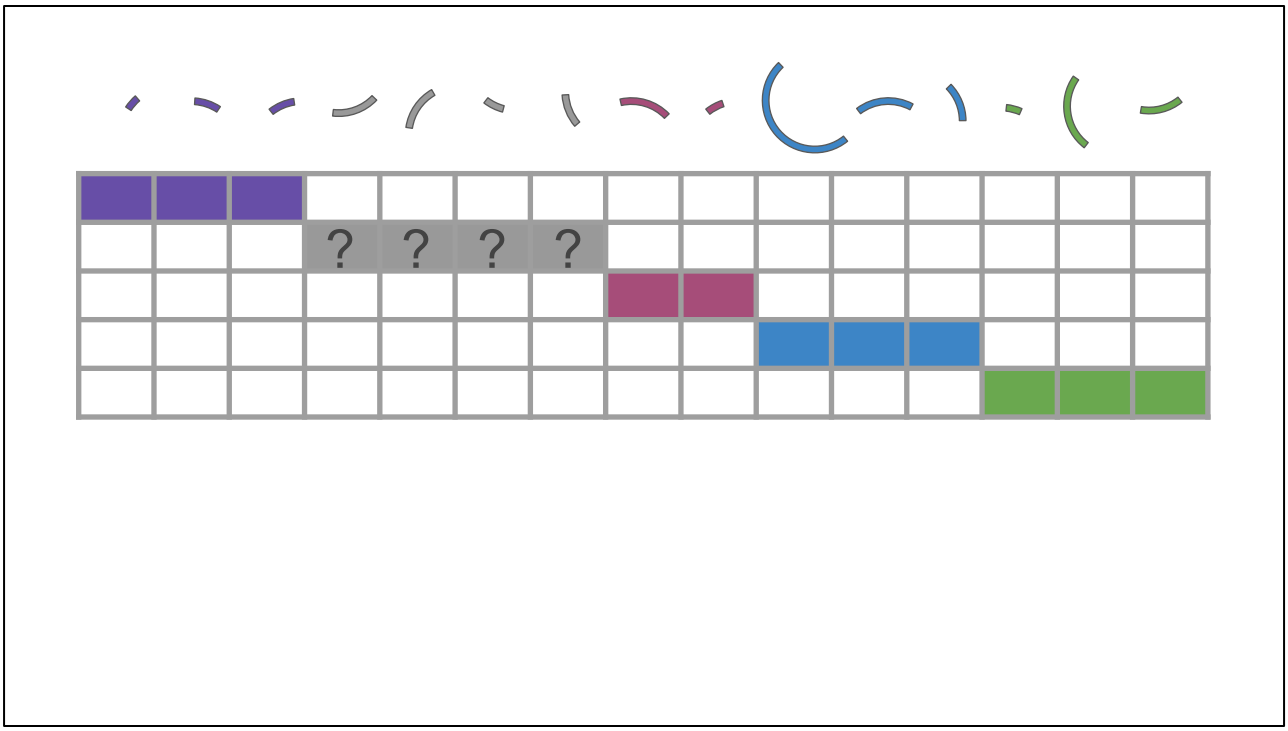
- However, this approach likely partitions core and variable genes into different bins for several reasons:
    - For one, variables genes often fail to cluster with the core genome, since their coverage across samples does not match.
    - In addition, the statistical models used to cluster sequences—such as mixtures of multivariate gaussians (GMM), hierarchical clustering, and k-means—assume that contigs only belong to one bin.



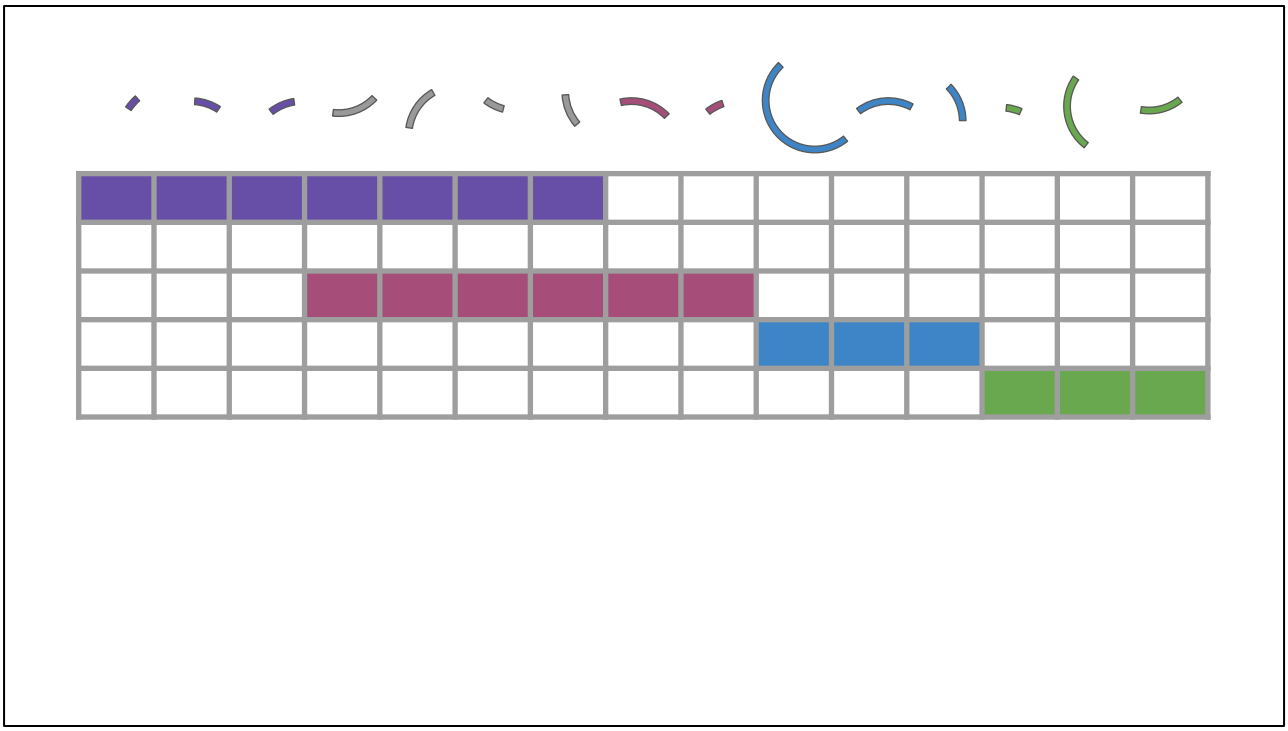
- What's more, the tools used to assess the quality of genome reconstruction count only conserved, single-copy genes, and therefore also overestimate the completeness of bins representing the core genome.



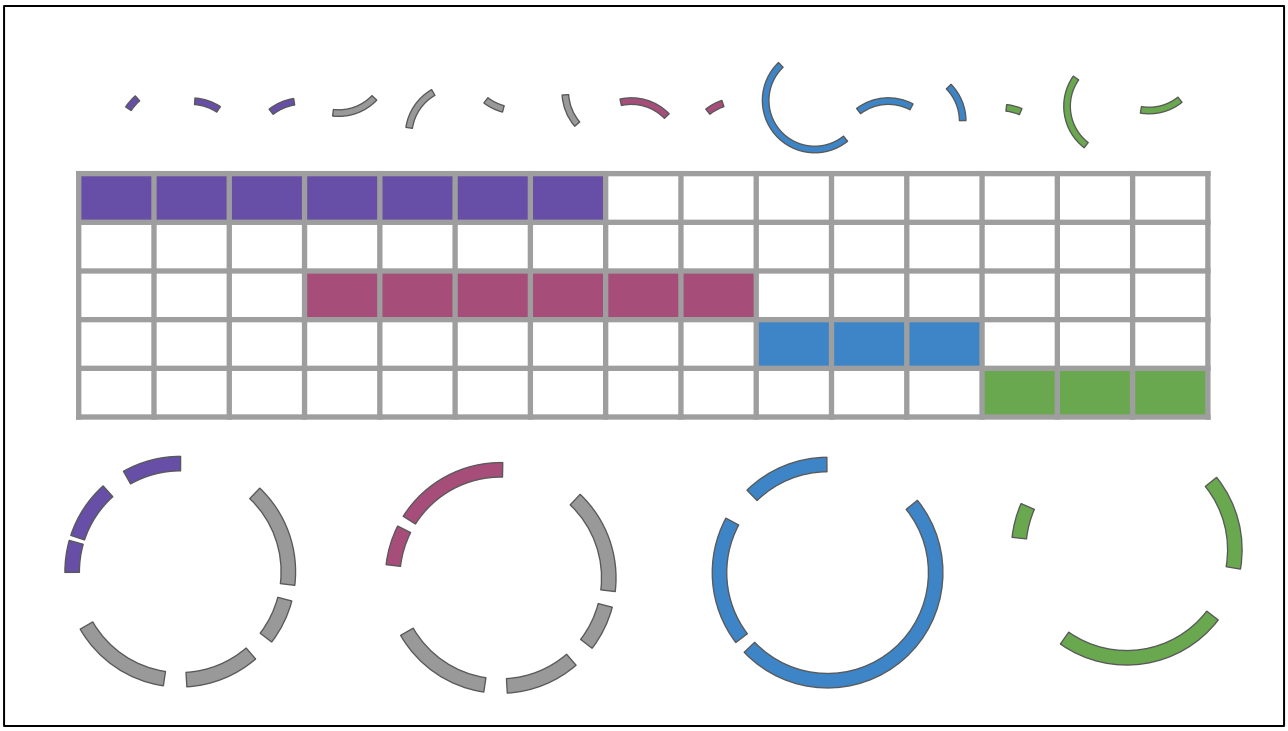
- How, then, do we accurately capture not only the core genome but the variable genome in order to track related strains and compare their functional potential?



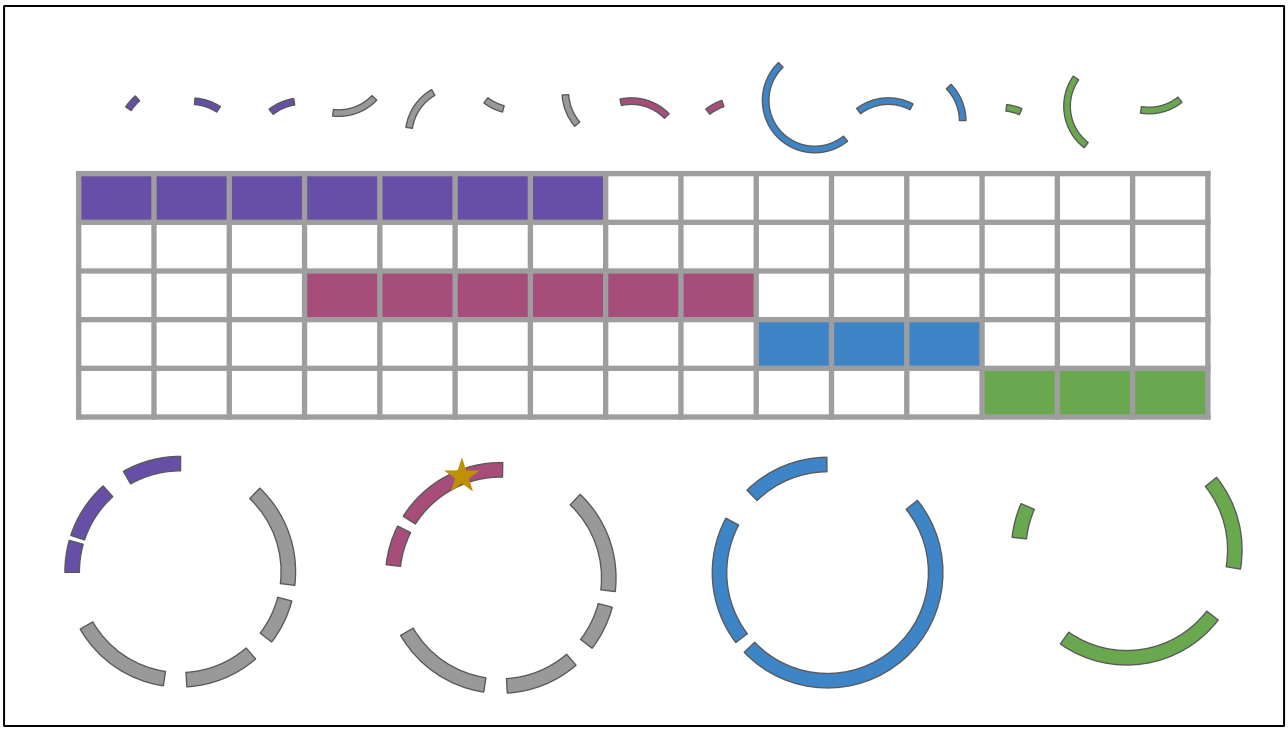
- The approach that I am taking is to select more biologically realistic statistical models for genome binning.



- By accounting for the possibility that some sequences are present in more than one organism, we're able to capture both core and variable genes where multiple strains are present.
- Specifically non-negative matrix factorization (NMF) offers a well developed toolbox of methods whose assumptions better reflect metagenomic data



- Using this approach we are able to accurately reconstruct genomes from metagenomes, without manual refinement, enabling automated, accurate assembly of massive datasets.



- Going forward, I aim to identify functions from variable genomes that are important for host health.