Microbial strain tracking and gene content reconstruction using metagenomic data

QBC Retreat
2022-11-08
Byron J. Smith

Acknowledgments

Acknowledgments

Pollard Lab

- Katie Pollard
- Chunyu Zhao
- Jason Shi

GLADSTONE INSTITUTES

UCSF

CZ BIOHUB

NIH National Institutes of Health

UC Noyce Initiative for Digital Transformation in Computational Biology & Health

@ByronJSmith

Human associated microbes are diverse and important

Human associated microbes are diverse and important

Important:

- Digestion
- Pathogen resistance
- Immune modulation
- Drug metabolism

Diverse:

- Hundreds of bacterial species
- Also archaea, eukaryotes, and viruses
- High inter-individual variation

Human associated microbes are diverse and important

Important:

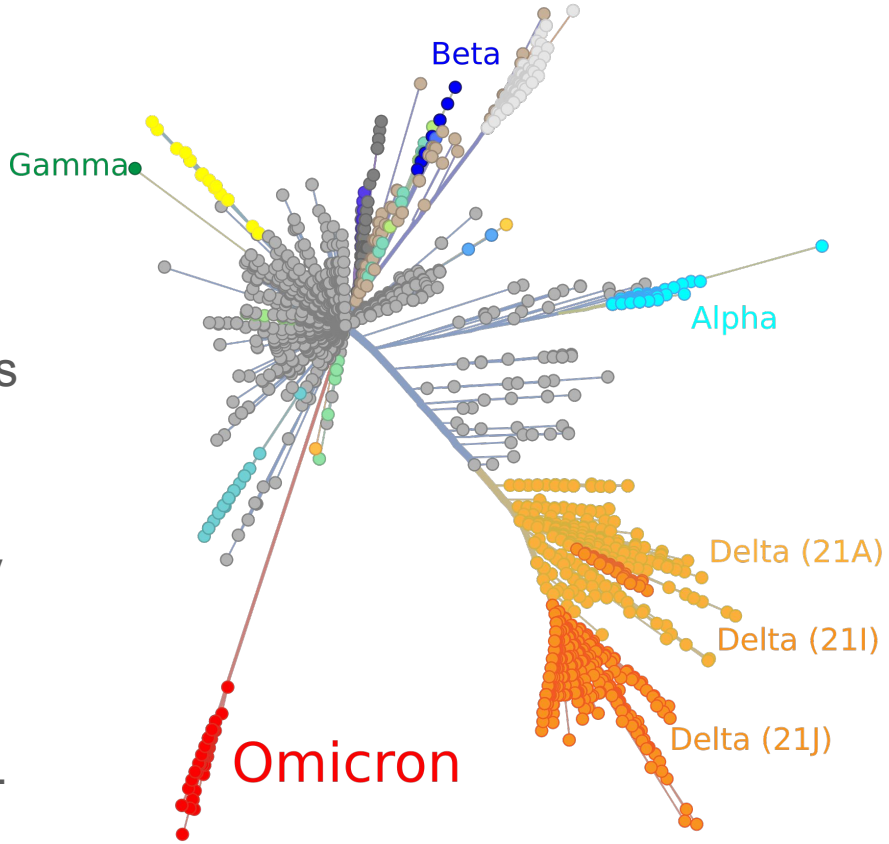
- Digestion
- Pathogen resistance
- Immune modulation
- Drug metabolism

Diverse:

- Hundreds of bacterial species
- Also archaea, eukaryotes, and viruses
- High inter-individual variation
- **Huge (but under-explored) diversity *within* species**

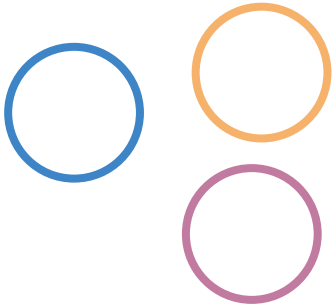
Microbial strain diversity is both biologically important and scientifically informative

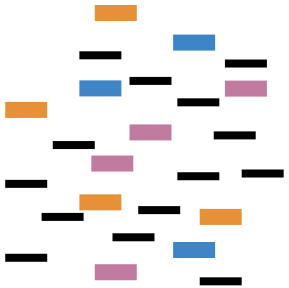
- Functional differences between strains
- Tracking strains between individuals, over time, or across global geography
- Transmission patterns, disease associations, selection pressures, etc.



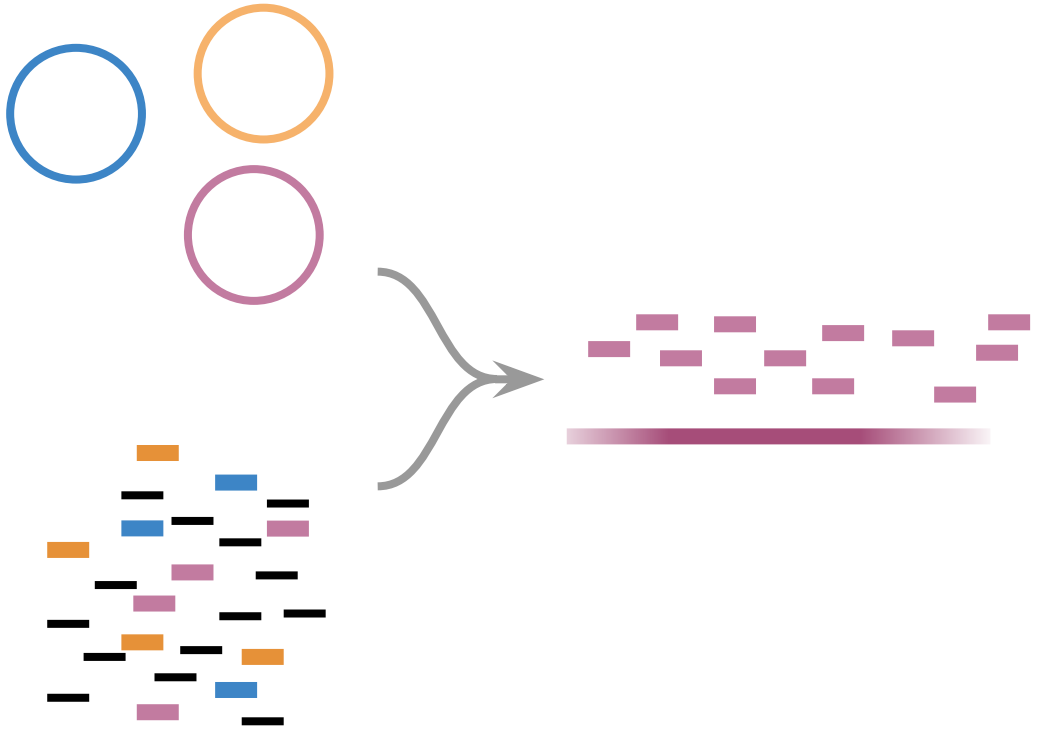
https://nextstrain.org

Metagenomic and reference genome collections have grown quickly

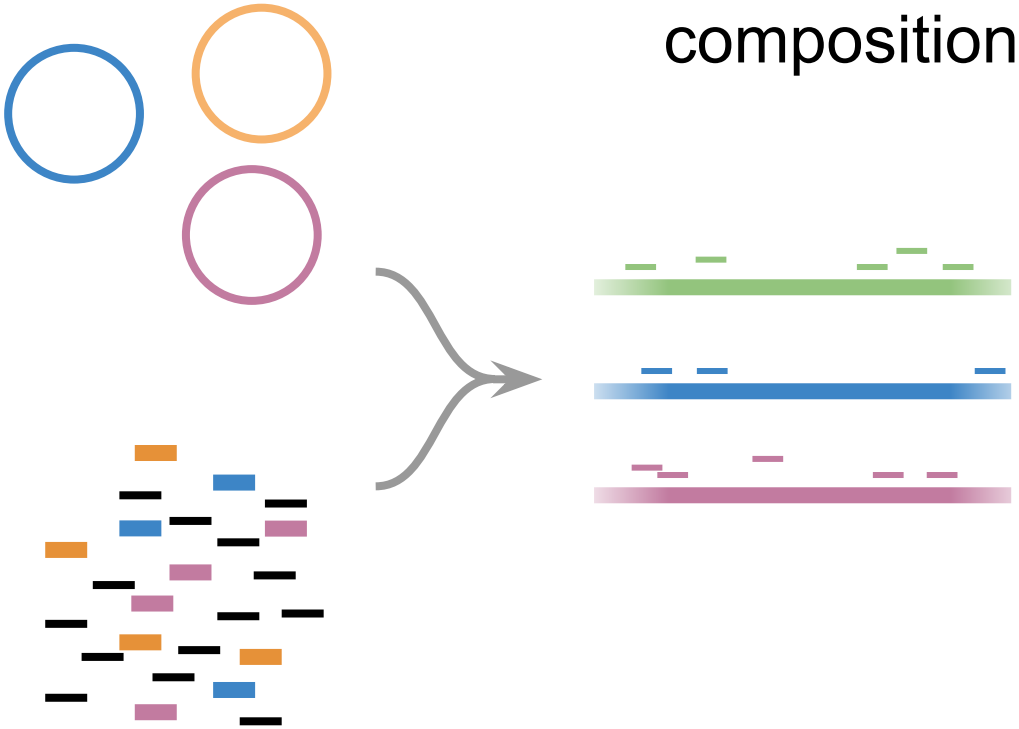




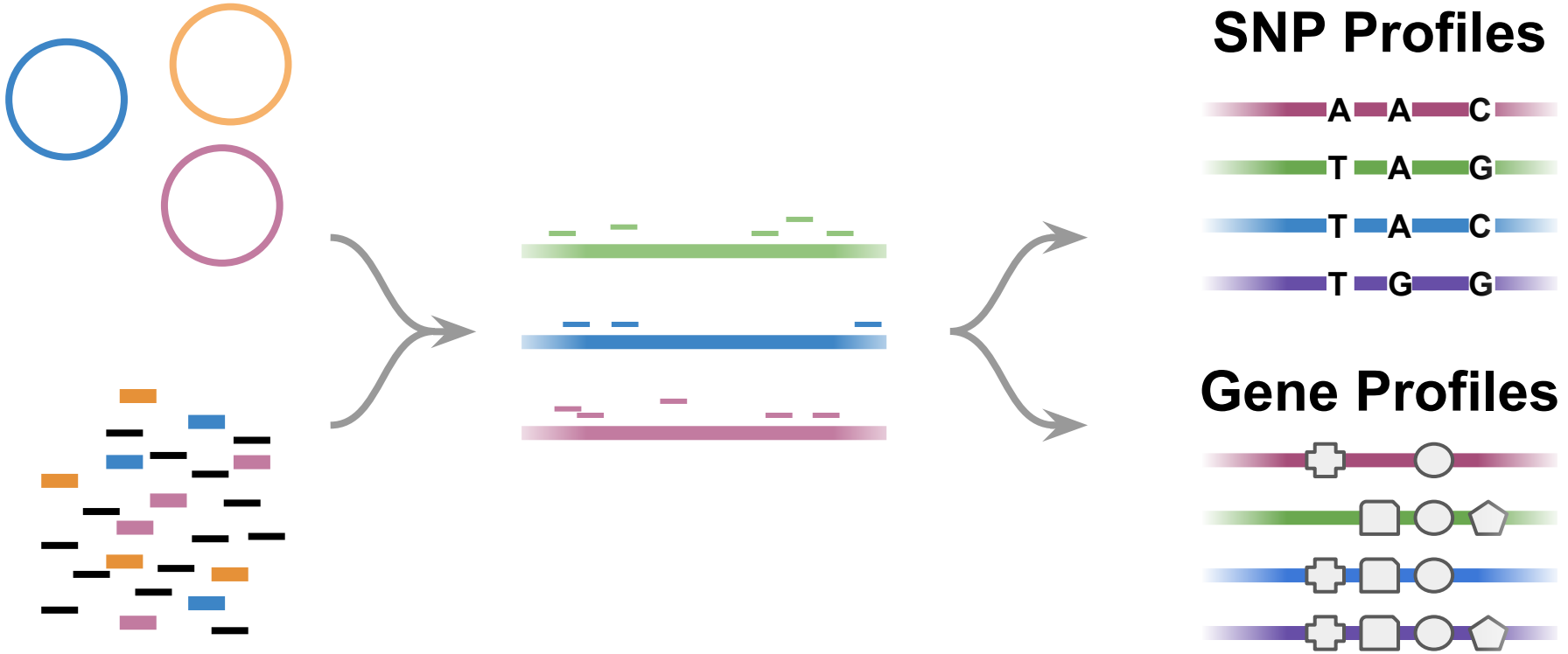
Reference-based tools map reads to species-specific genome sequence



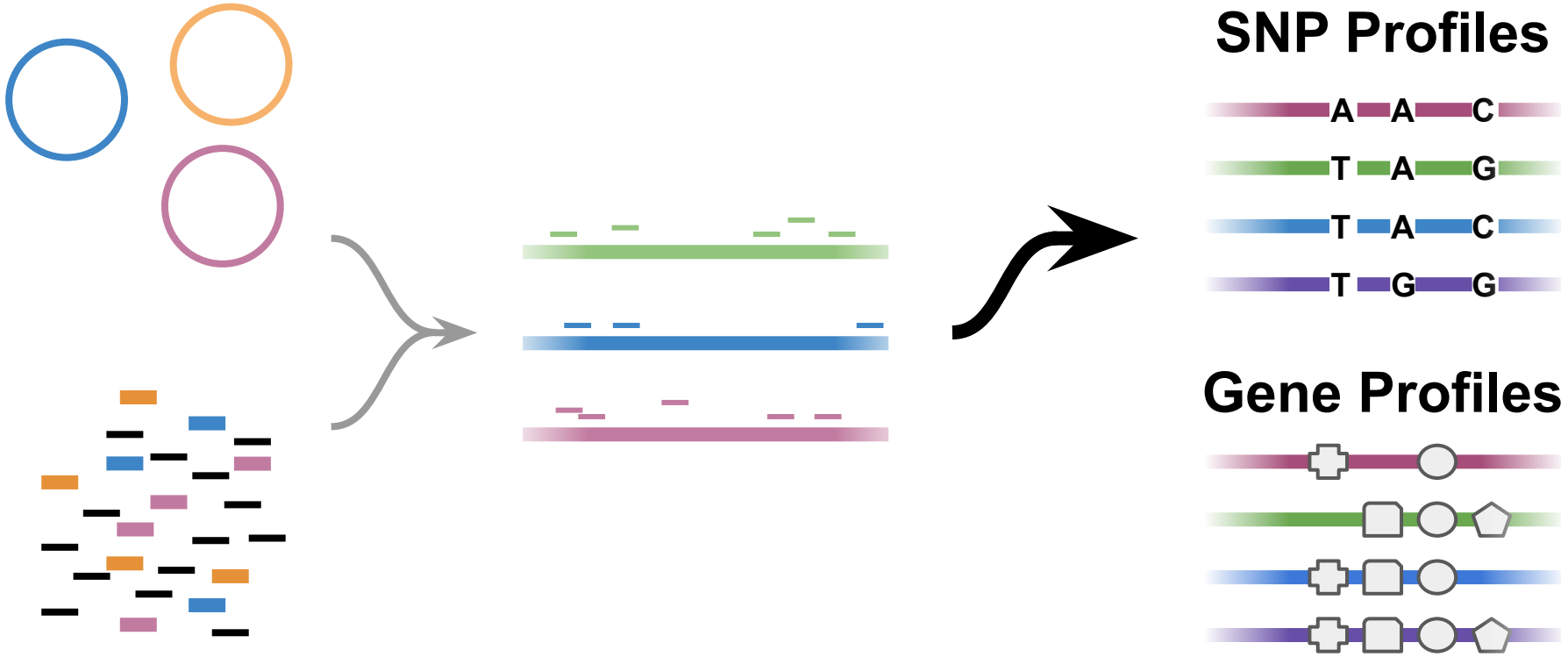
Taxonomic profiling allows us to understand the species-level composition of the microbiome



Profiling single-nucleotide and gene content variants



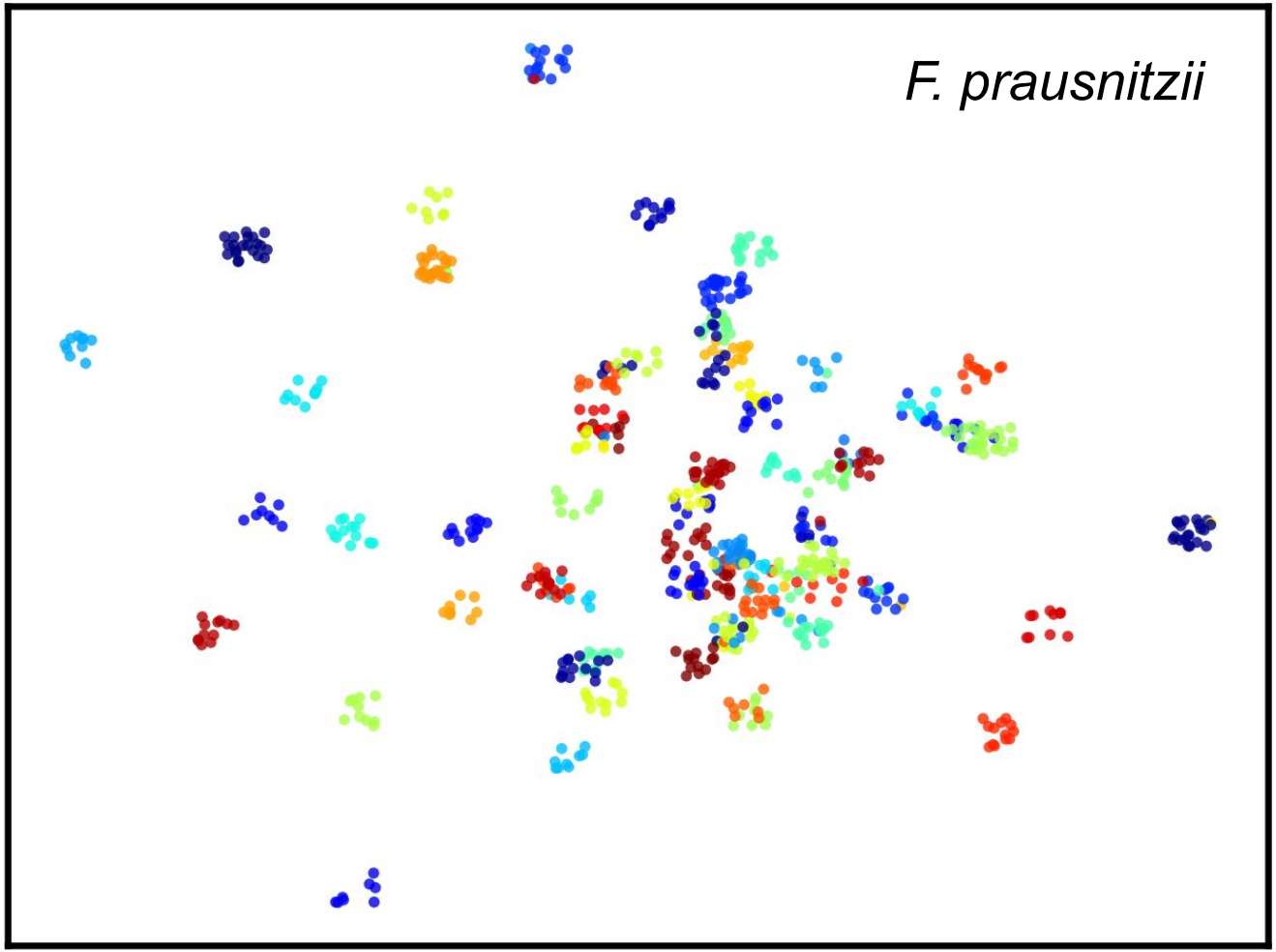
SNP profiles allow us to identify distinct strains



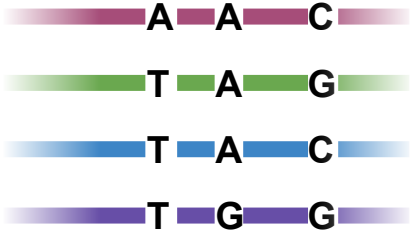
Strain identification using SNP profiles

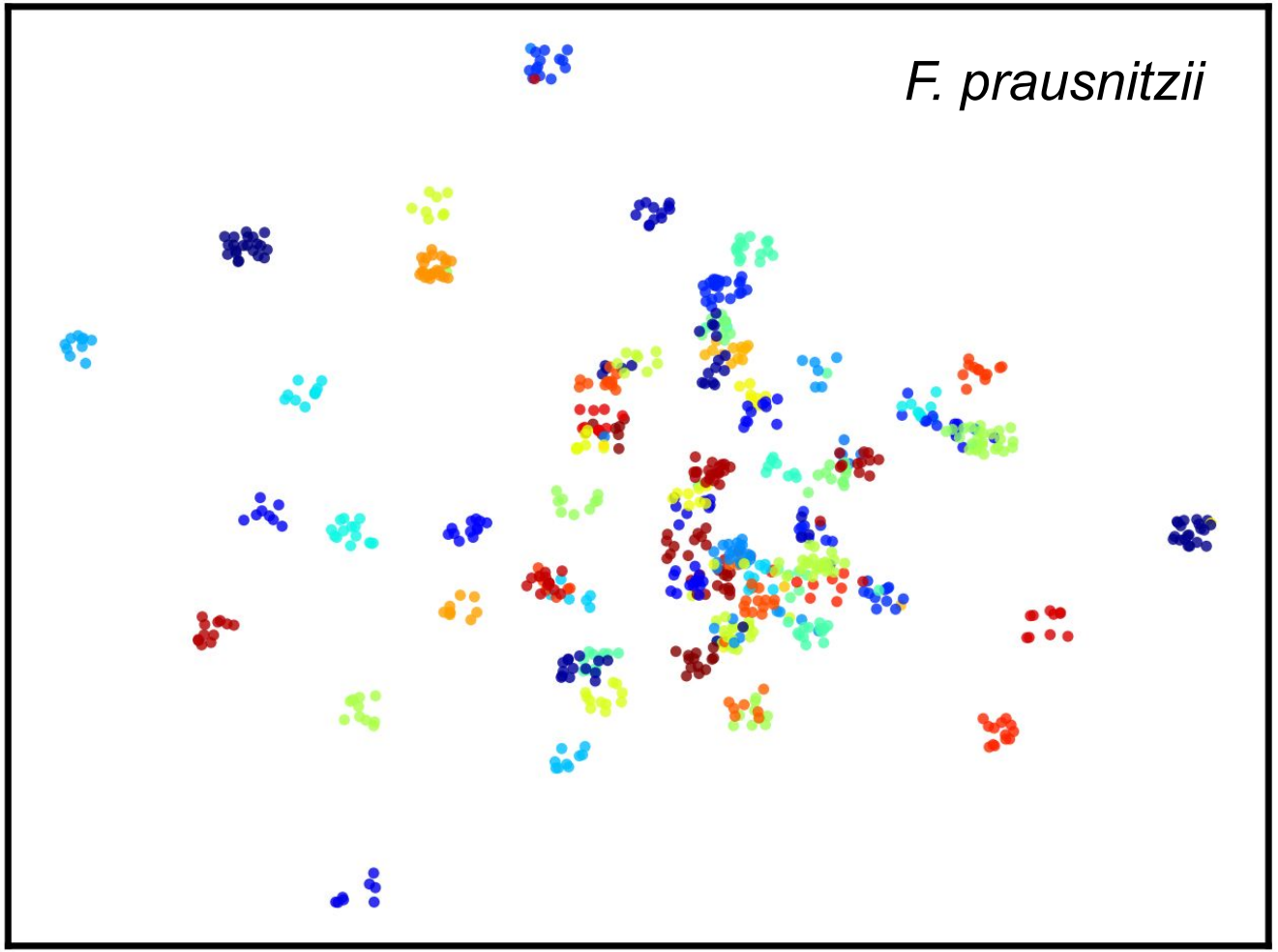
In this talk: HMP2 dataset

composed of
- ~1300 samples
- ~100 subjects

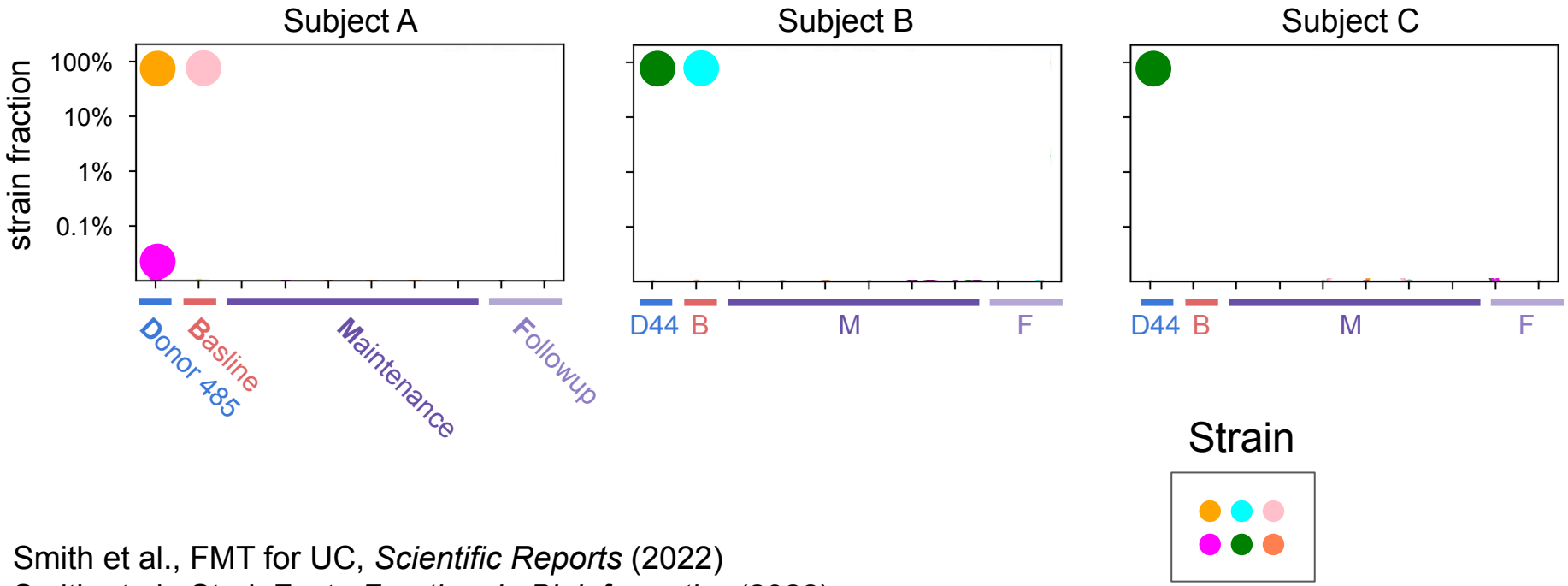


Strain identification using SNP profiles



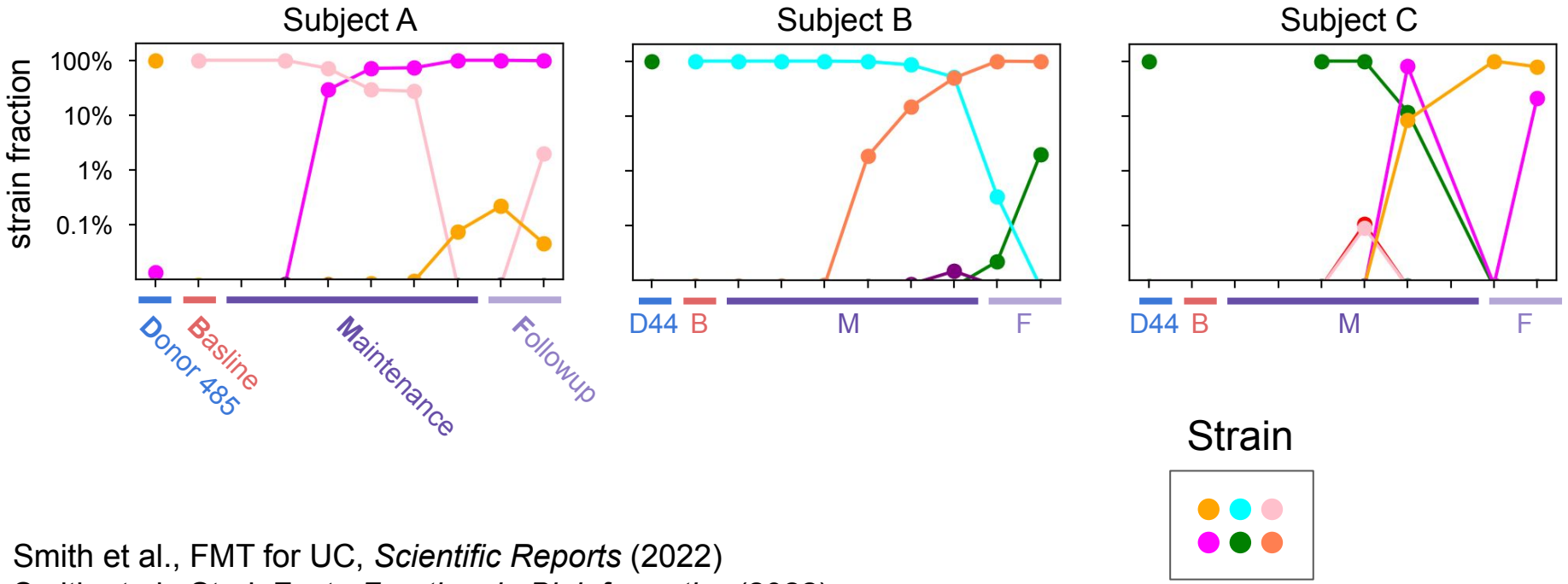


Strain diversity enables tracking of transmission between microbiomes



Smith et al., FMT for UC, *Scientific Reports* (2022)
Smith et al., StrainFacts *Frontiers in Bioinformatics* (2022)

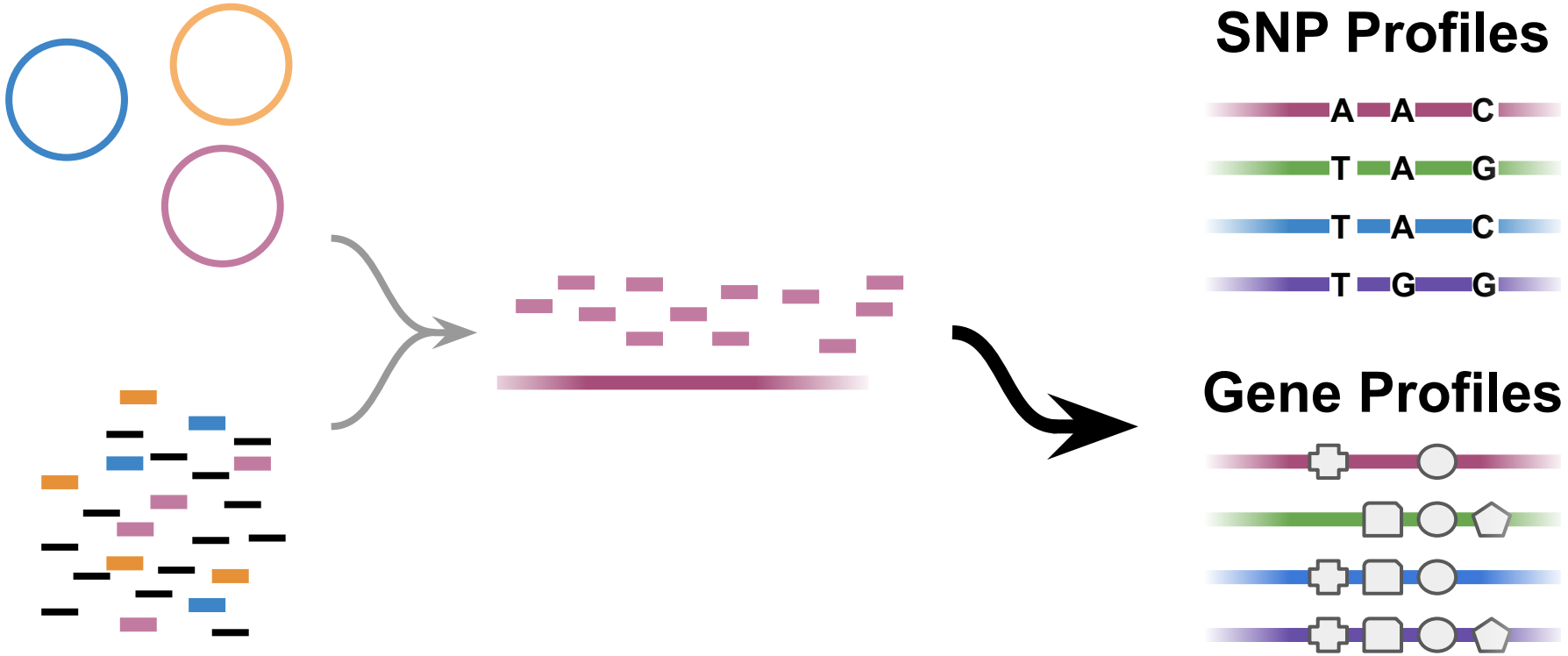
Strain diversity enables tracking of transmission between microbiomes



Smith et al., FMT for UC, *Scientific Reports* (2022)
Smith et al., StrainFacts *Frontiers in Bioinformatics* (2022)

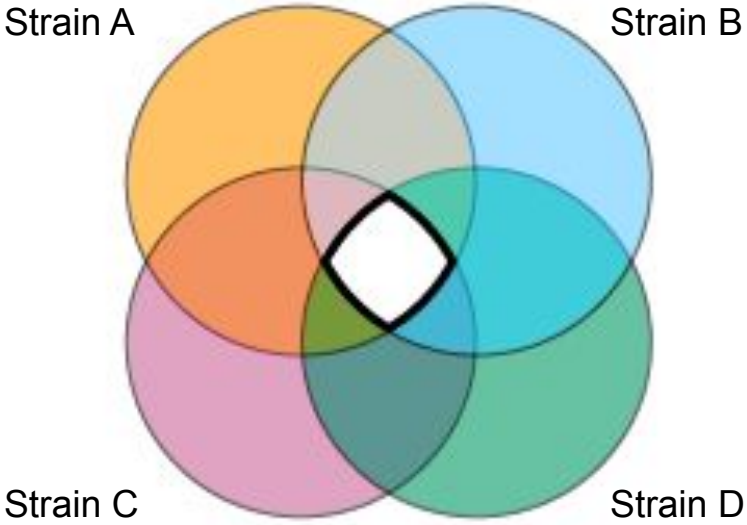
What are the impacts of this strain diversity on the microbiome and human health?

Reconstructing gene content from metagenomes



Inferring gene content *accurately* is difficult.

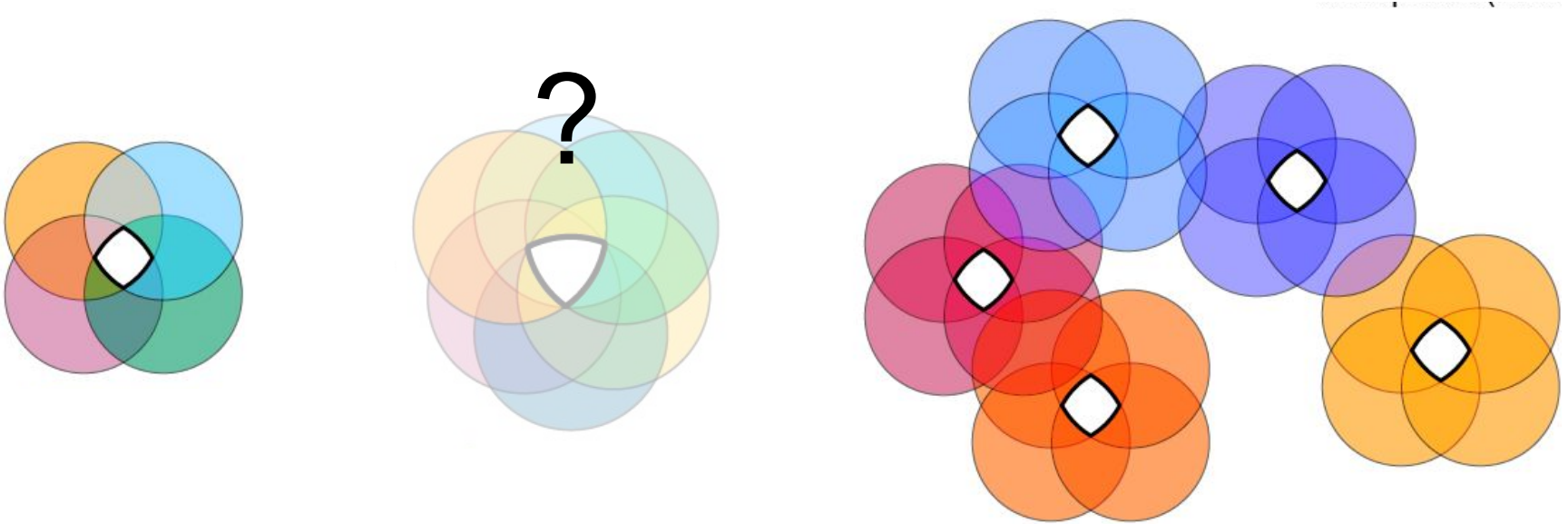
Challenge: Pangenomes



Inferring gene content *accurately* is difficult.

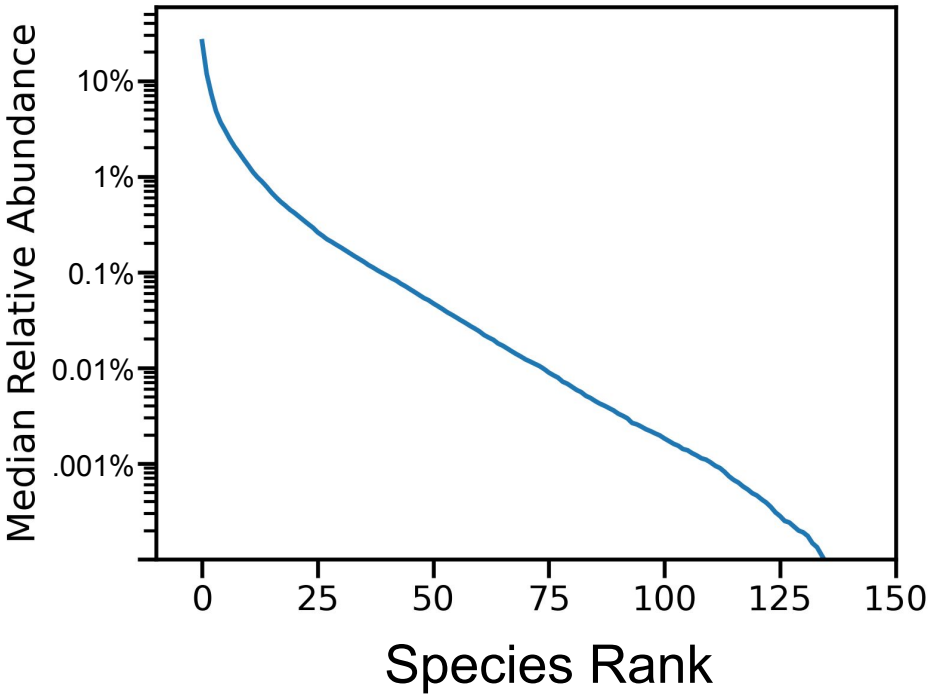
wikipedia.org

Challenge: Pangenomes are large, incomplete, and overlapping

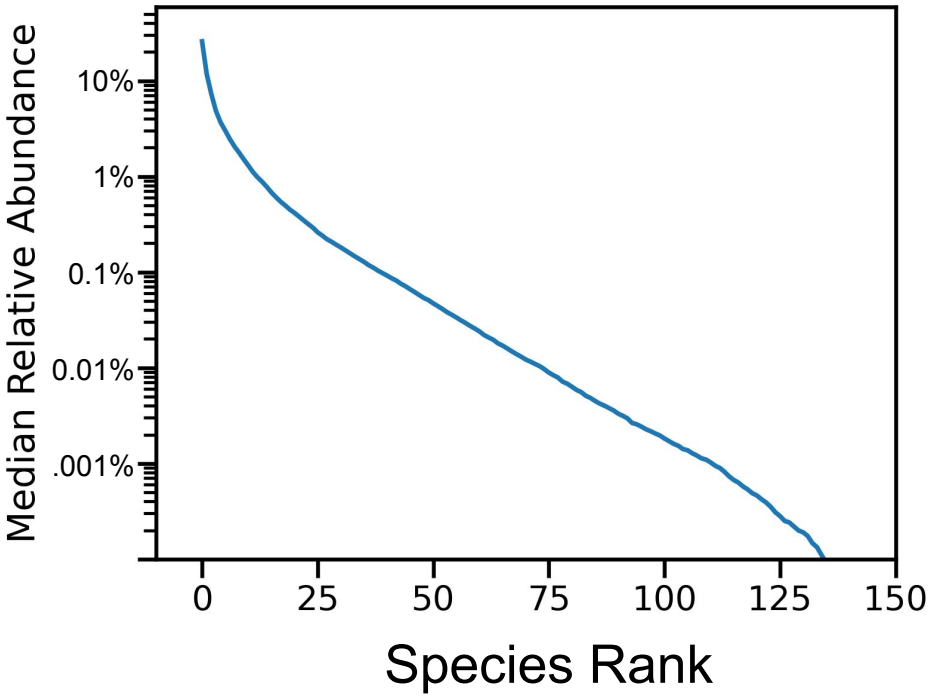


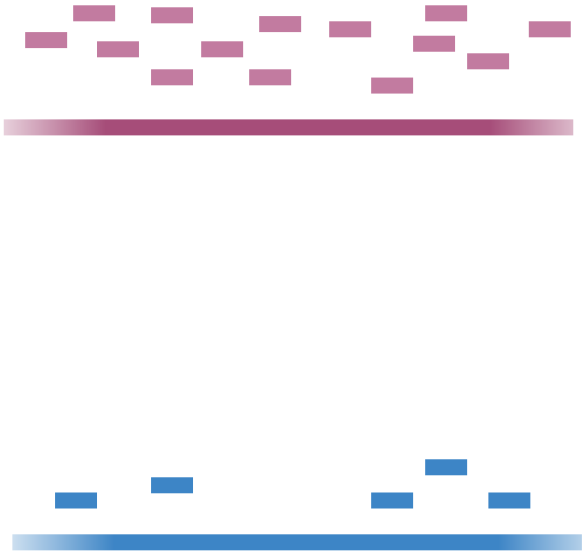
wikipedia.org

Challenge: Long tail of species diversity



High levels of diversity results in insufficient sequencing depth for low-abundance species





Strain-resolved gene content reconstruction: major challenges

- Low abundance (sparsity)

Strain-resolved gene content reconstruction: major challenges

- Low abundance (sparsity)
- Missing references

Strain-resolved gene content reconstruction: major challenges

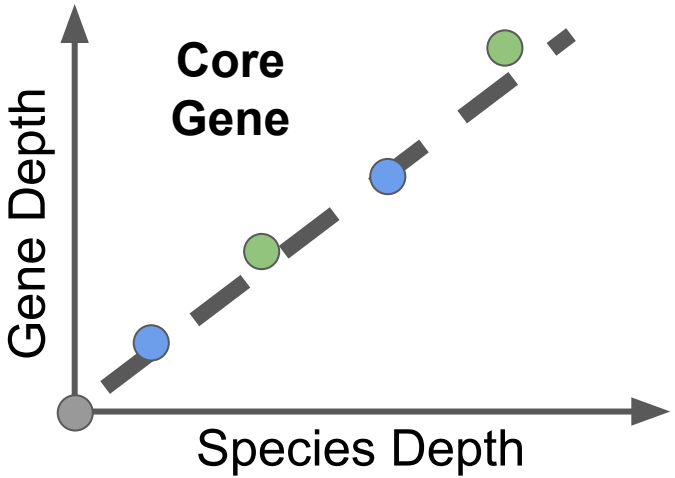
- Low abundance (sparsity)
- Missing references
- Cross-mapping from other species

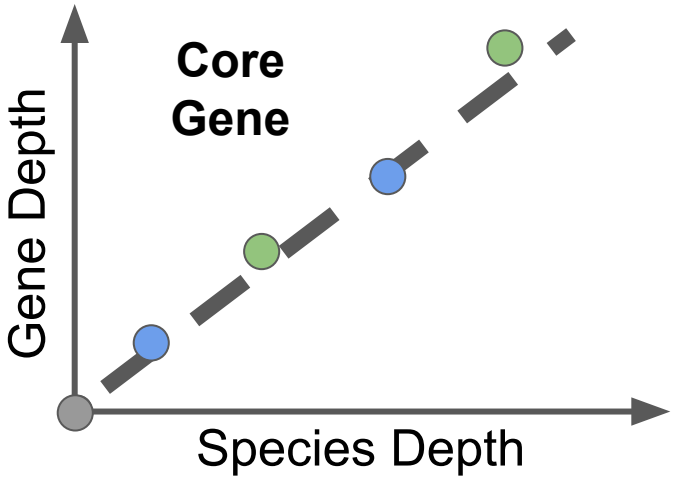
Strain-resolved gene content reconstruction: major challenges

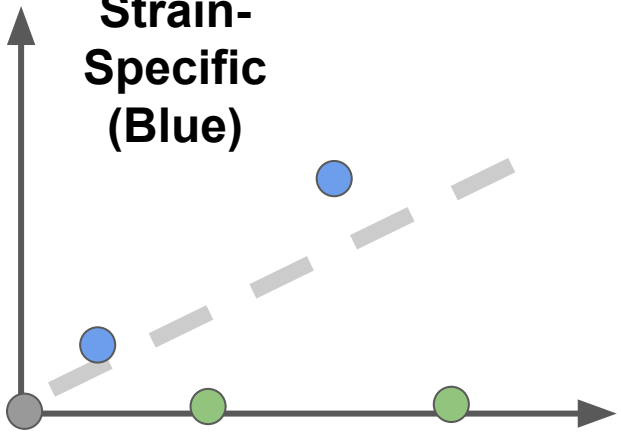
- Low abundance (sparsity)
- Missing references
- Cross-mapping from other species

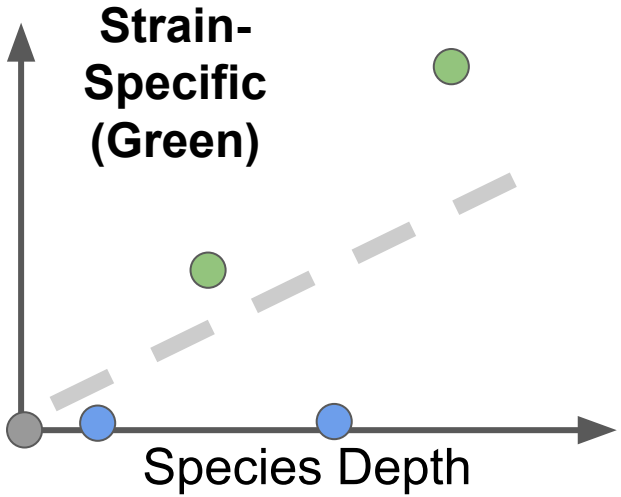
How to overcome these limitations?

Solution: Look for correlations across multiple samples, instead of depth alone



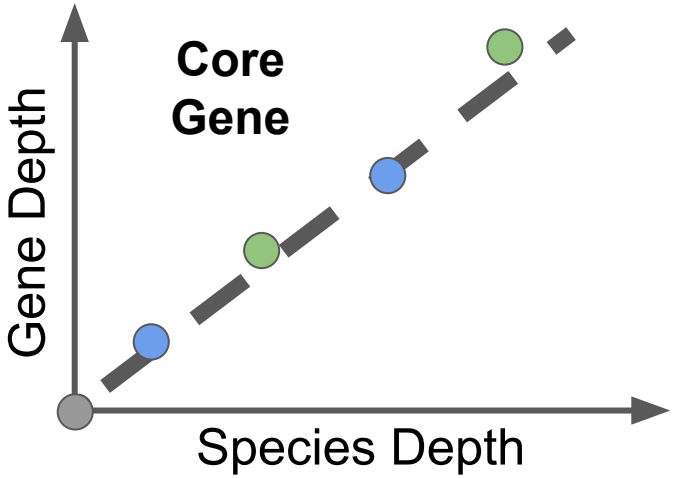


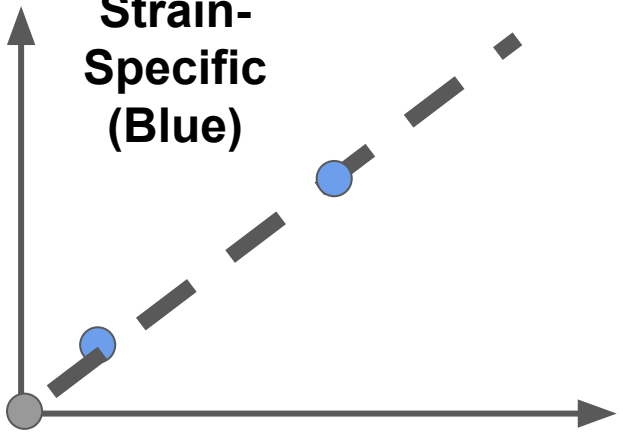


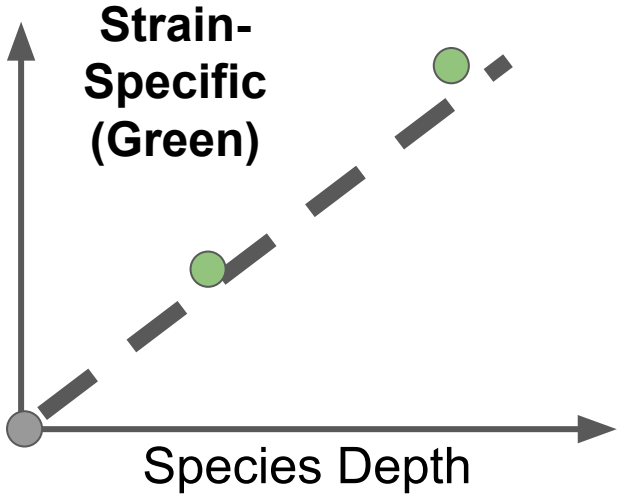


Challenge:
Strain variation

Correlations are weakened and strain-specific genes are lost due to inconsistent depth.





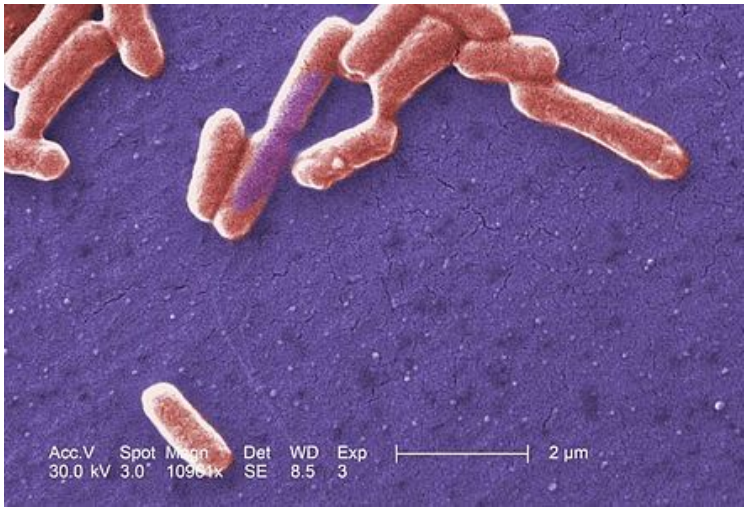


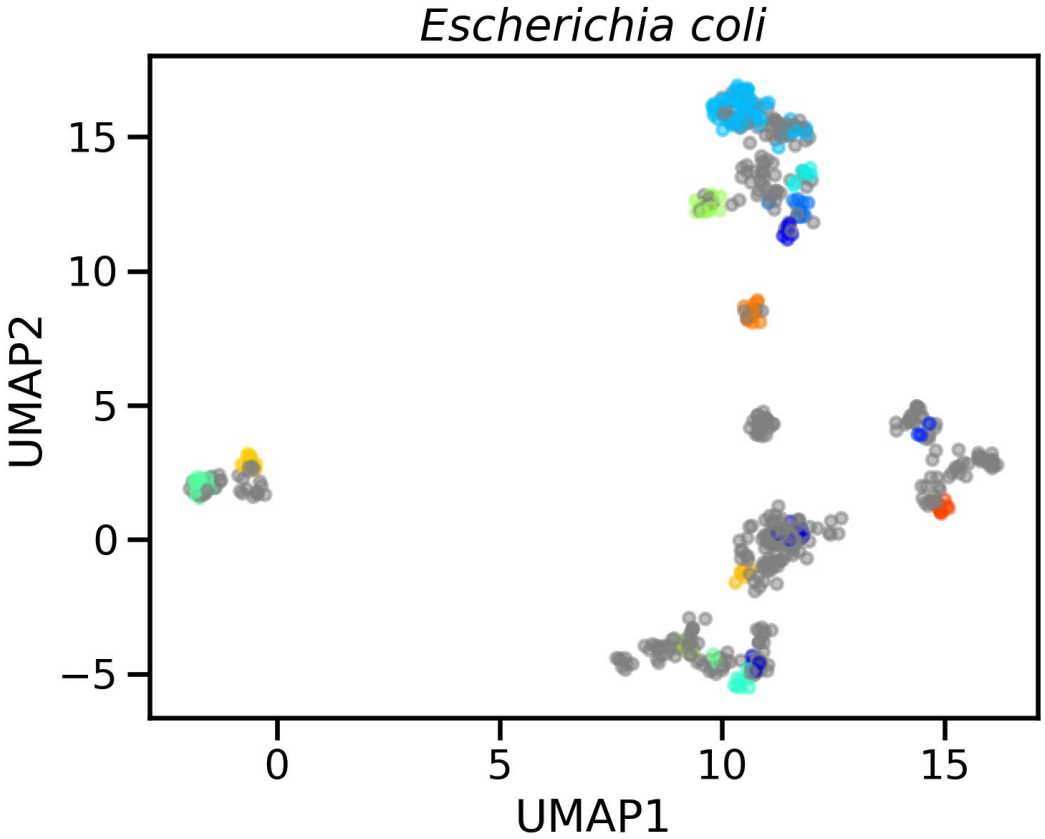
Challenge:
Strain variation

Correlations are weakened and strain-specific genes are lost due to inconsistent depth.

Solution:
Partition samples by strain

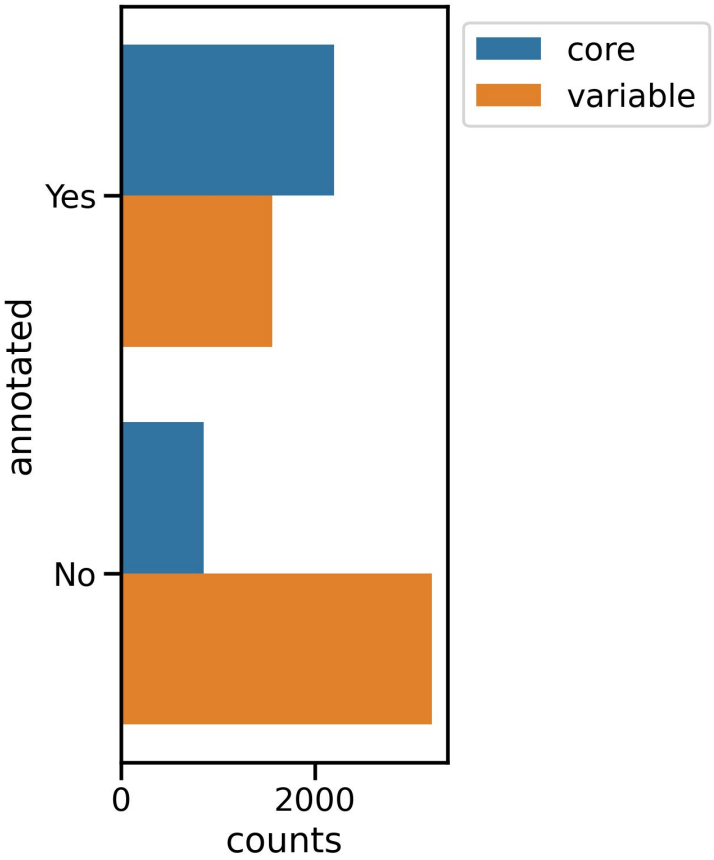
Solution: Partition samples by strain



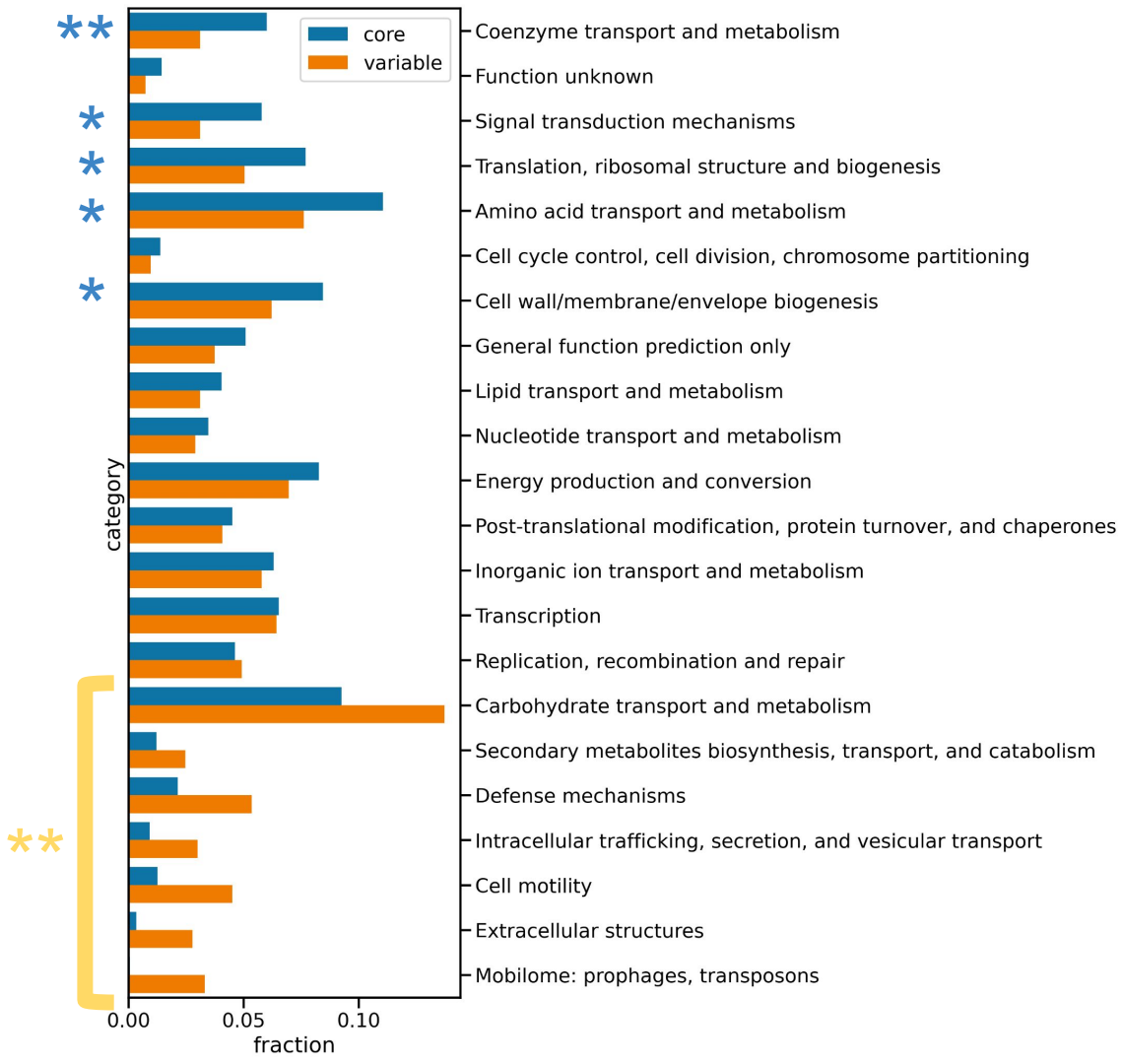


Inferred genes for 16 distinct *E. coli* strains

The variable fraction is enriched with un-annotated genes.



Model lab strains and other isolates may be insufficient for understanding physiology in the gut microbiome.

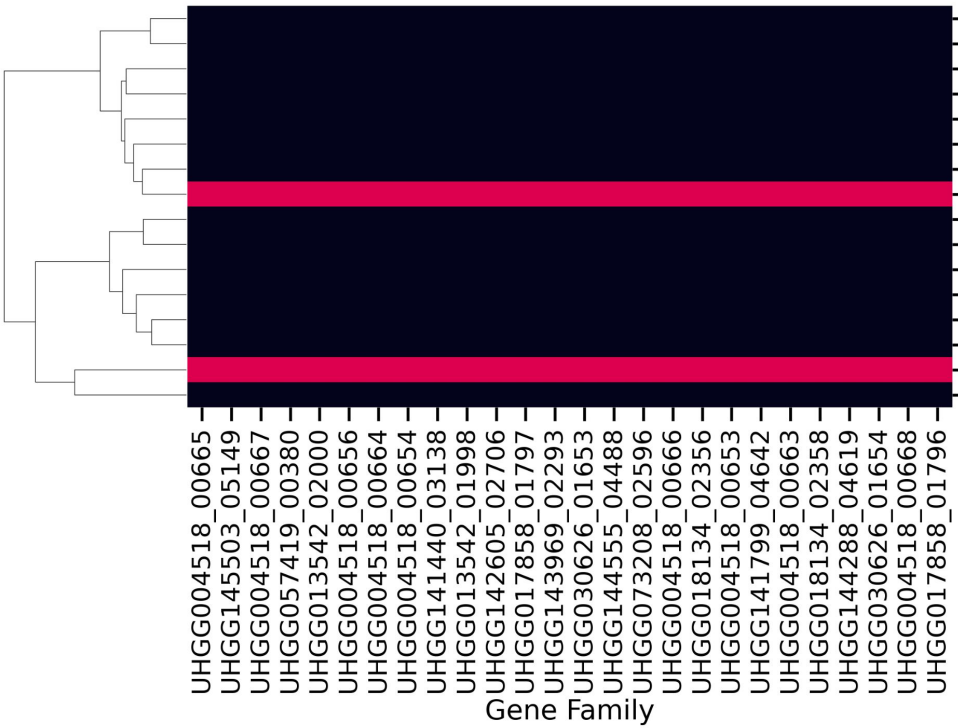


Among annotated genes, variable genome is enriched with important functional categories, e.g.:

- Motility
- Carbohydrate and secondary metabolism
- Defense
- Etc.

Distantly related strains can share an entire suite of genes

Distantly related strains can share an entire suite of genes



Transporter for capsular polysaccharide:

- kpsD/M
  (COG1596, COG1682)

Rhamnose synthesis (component of O-antigen)

- rfbB/C/D
  (COG1088, COG1898, COG1091)
- rmlA (COG1209)

S-layer glycoprotein synthesis

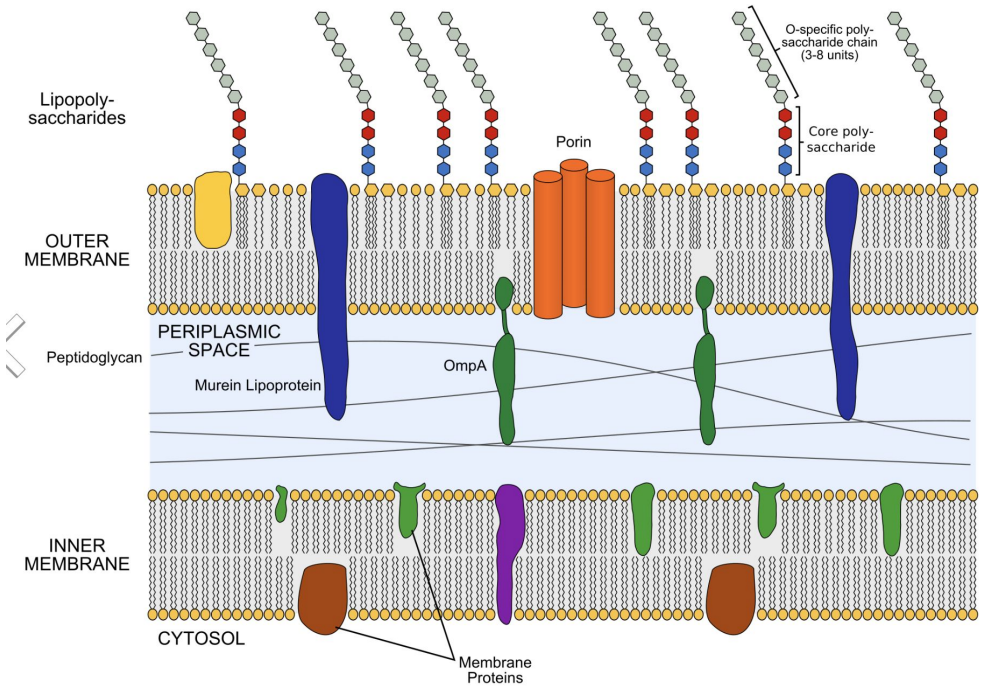
- fdtC

Prophage integrase

- intA (COG0582)

18 un-annotated proteins

Distantly related strains can share an entire suite of genes



Prophage integrase

- intA (COG0582)

Transporter for capsular polysaccharide:

- kpsD/M (COG1596, COG1682)

Rhamnose synthesis (component of O-antigen)

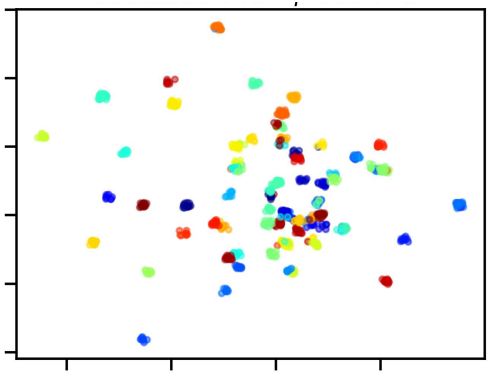
- rfbB/C/D (COG1088, COG1898, COG1091)
- rmlA (COG1209)

S-layer glycoprotein synthesis

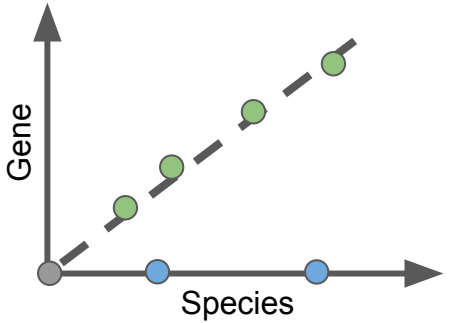
- fdtC

18 un-annotated proteins

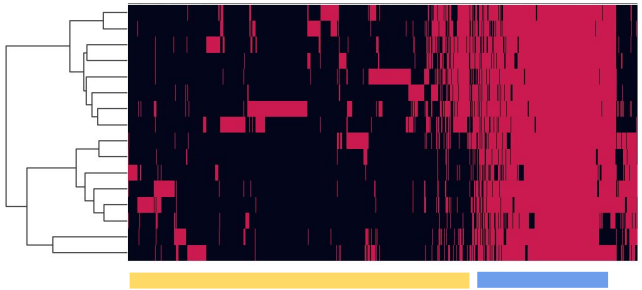
Enormous Strain Diversity



Strain-informed Gene Inference



Core and Variable Gene Content



Functional Enrichment in Variable Fraction