Strain-resolved inference of microbial gene content in large metagenomic datasets

Byron J. Smith
Gladstone Institutes, San Francisco
CSHL Microbiome (2022-10-27)

Acknowledgments

Acknowledgments

Pollard Lab

- Katie Pollard
- Chunyu Zhao
- Jason Shi

GLADSTONE INSTITUTES

UCSF

CZ BIOHUB

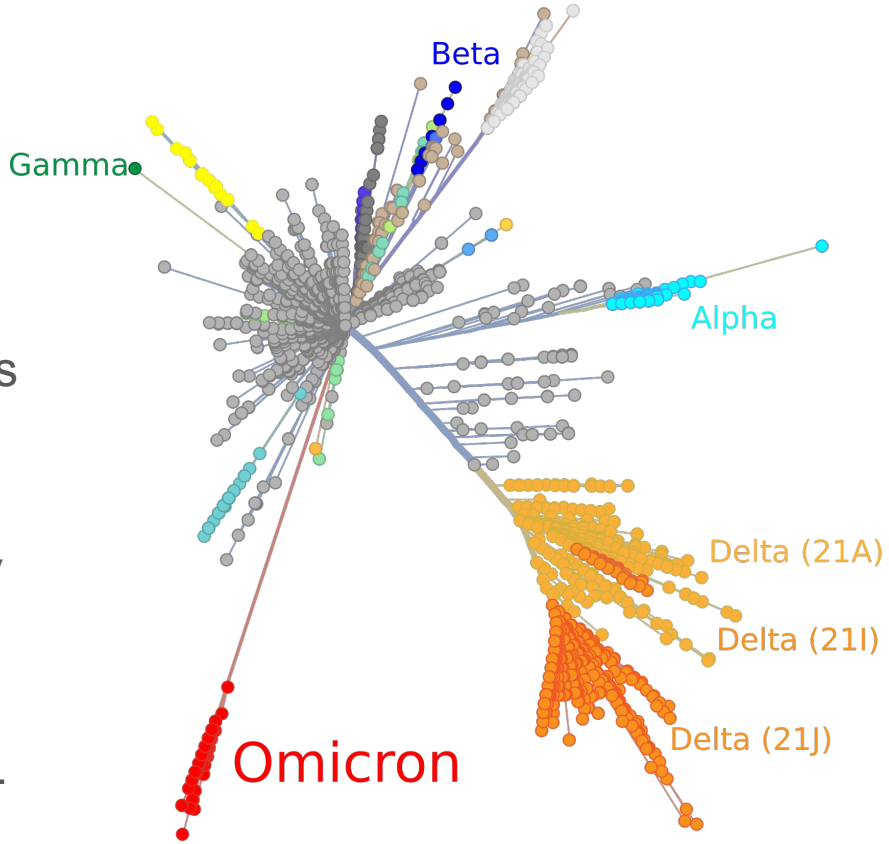
NIH National Institutes of Health

UC Noyce Initiative for Digital Transformation in Computational Biology & Health

@ByronJSmith

Strain diversity is both biologically important and scientifically informative

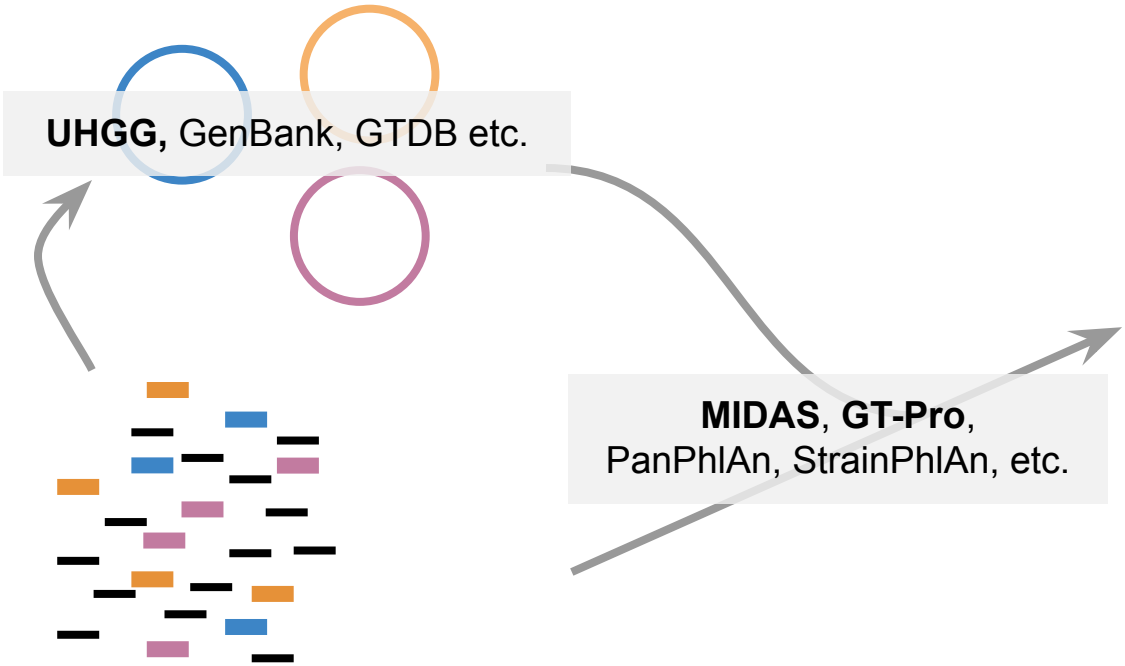
- Functional differences between strains
- Tracking strains between individuals, over time, or across global geography
- Transmission patterns, disease associations, selection pressures, etc.



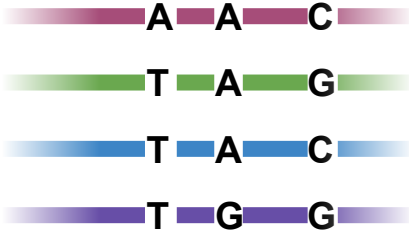
https://nextstrain.org

Reference based methods for metagenomic profiling

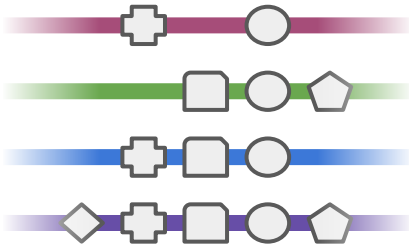
Reference based methods for metagenomic profiling



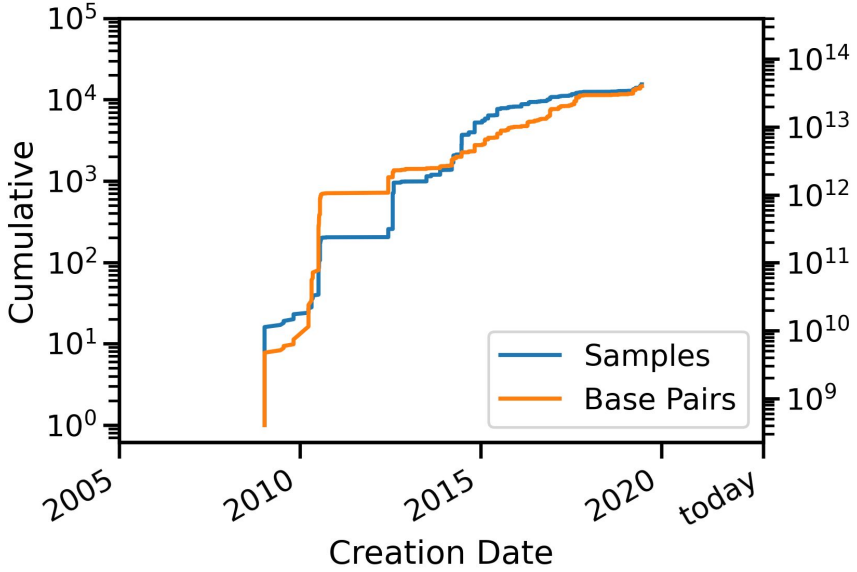
SNP Profiles



Gene Profiles

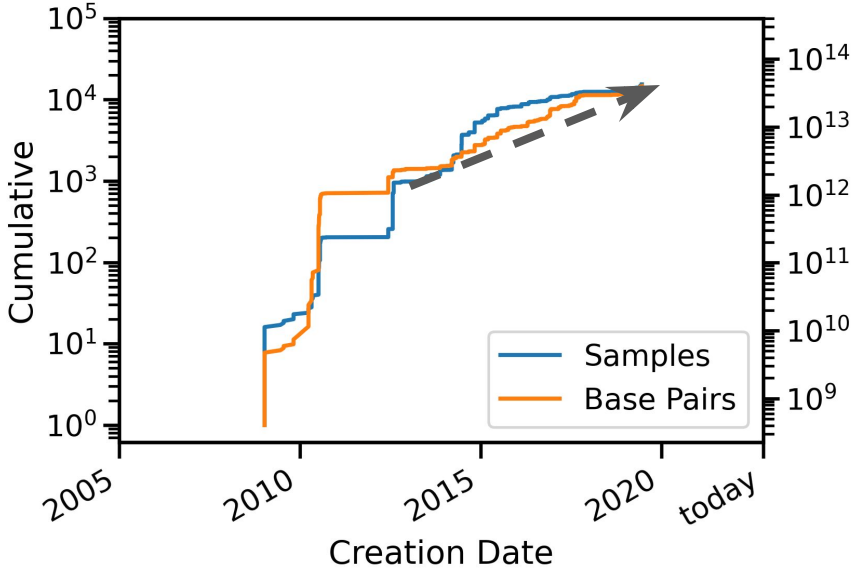


Metagenomic datasets are growing rapidly



Kasmanas et al., HumanMetagenomeDB, *Nucleic Acids Res*. 2021

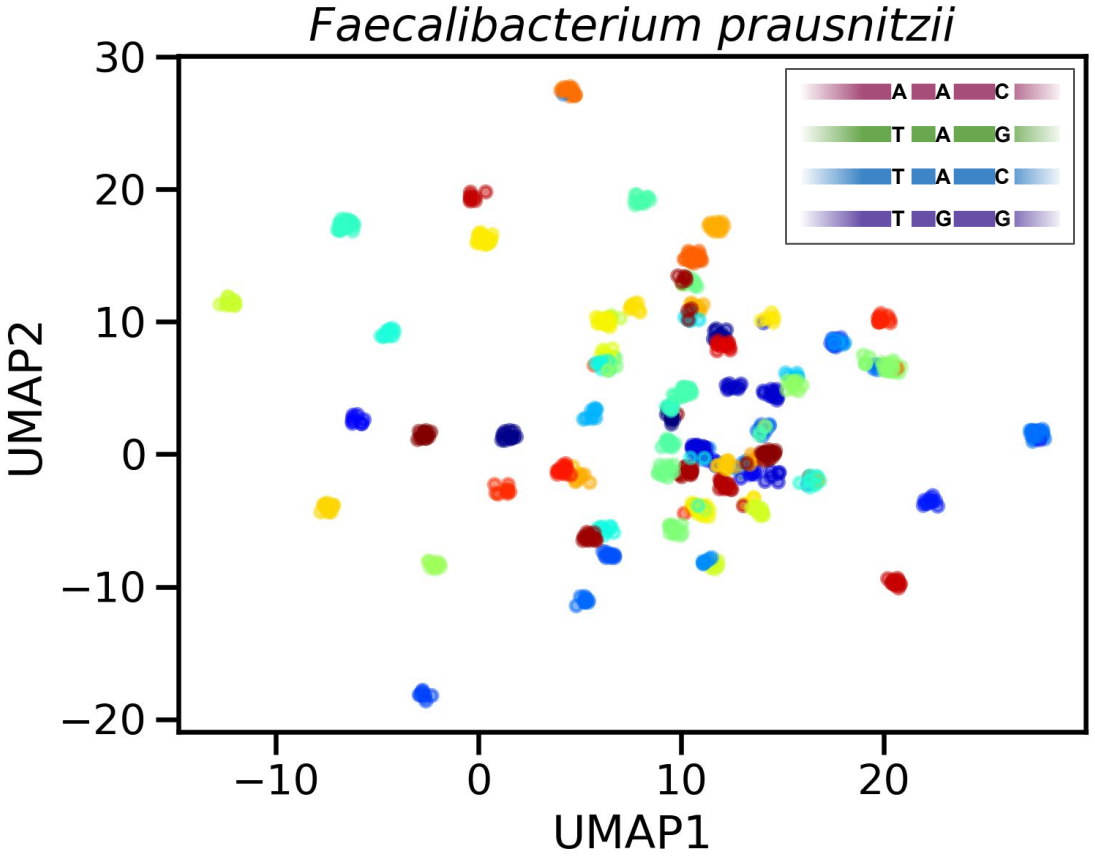
Metagenomic datasets are growing rapidly

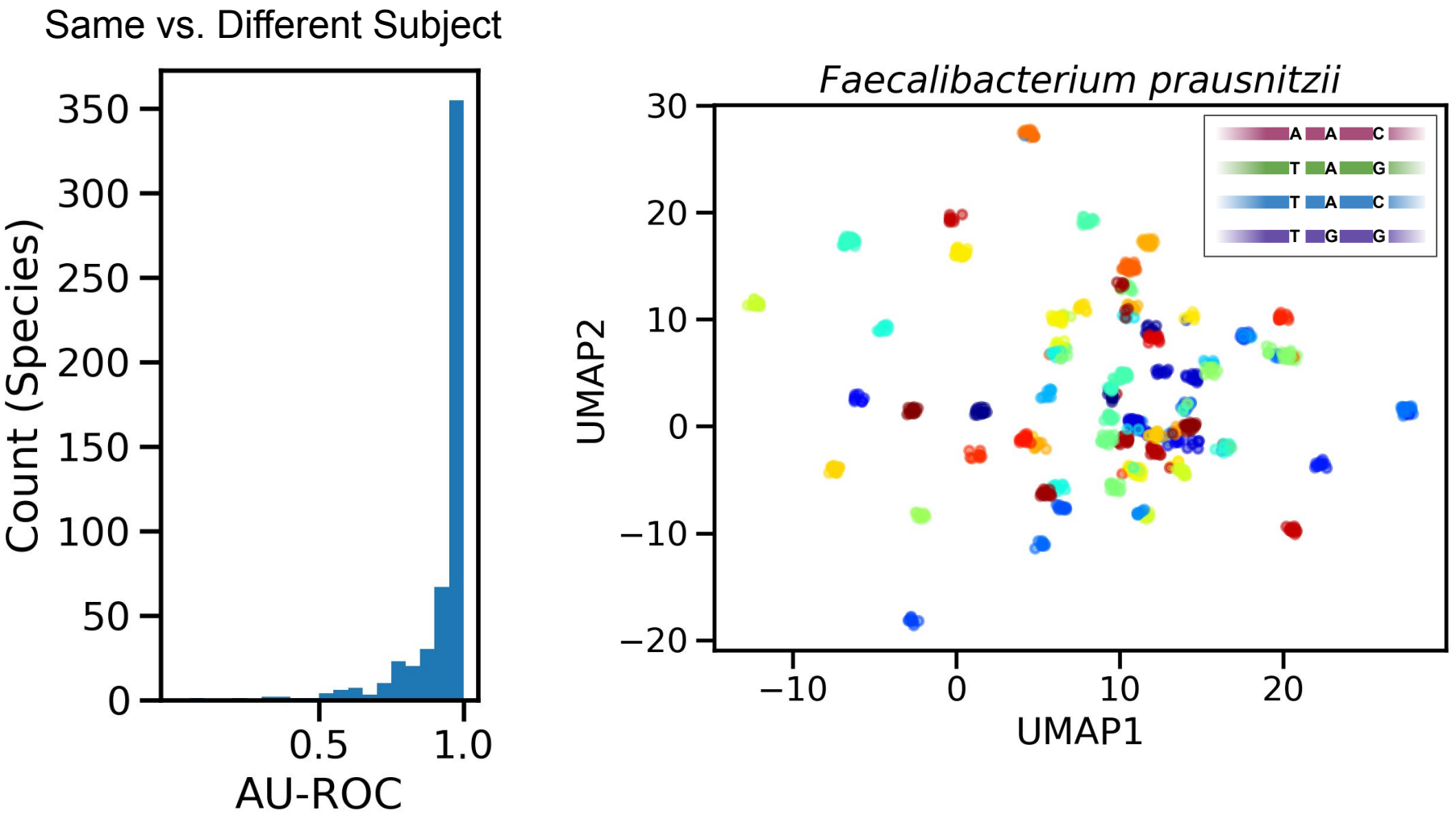


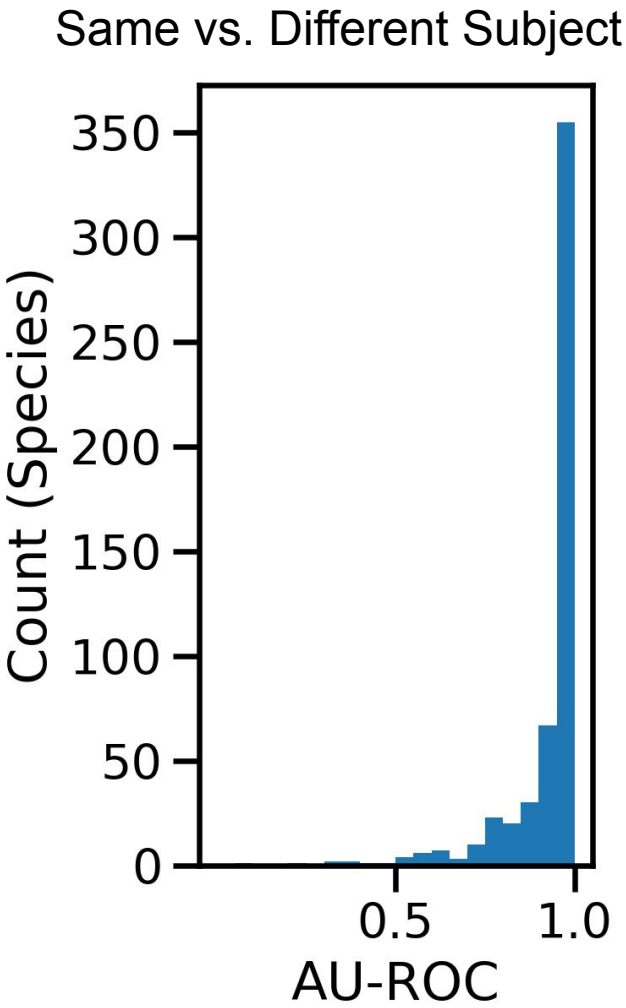
- 10^5 stool metagenomes publicly available (or soon)
- Median depth of ~10M reads
- One notable example: HMP2 composed of
  - ~1300 samples
  - ~100 subjects
  - Longitudinal sampling

Kasmanas et al., HumanMetagenomeDB, *Nucleic Acids Res*. 2021

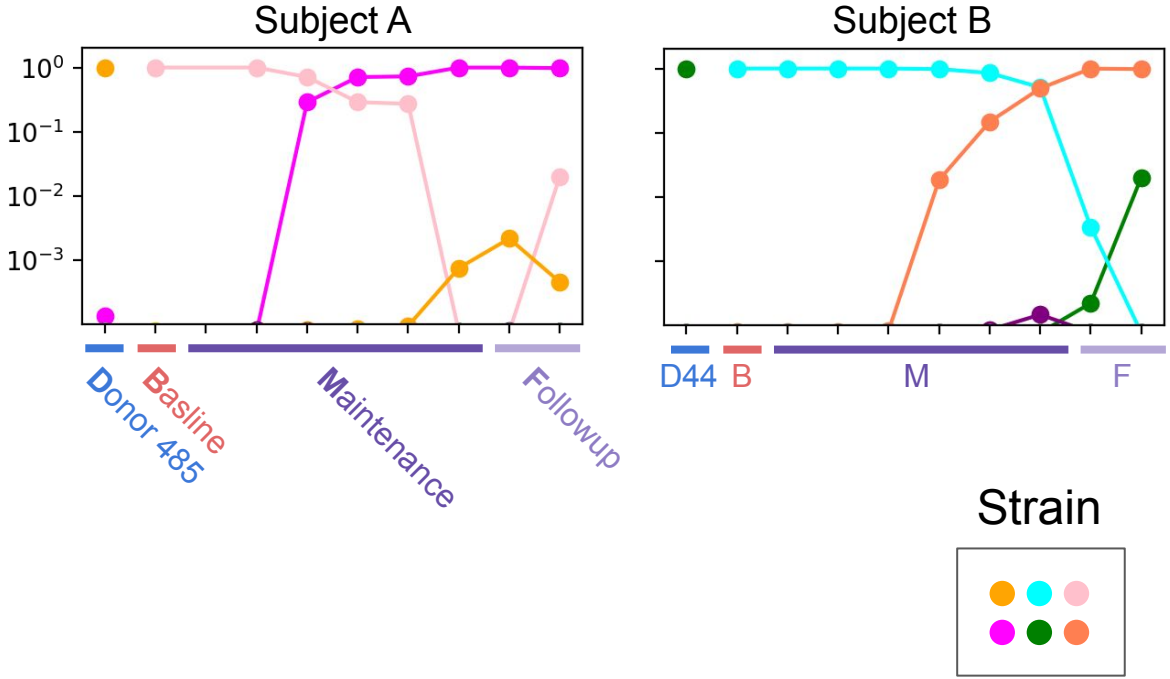
Beyond species diversity:
100 or 1000s of distinct strains across subjects



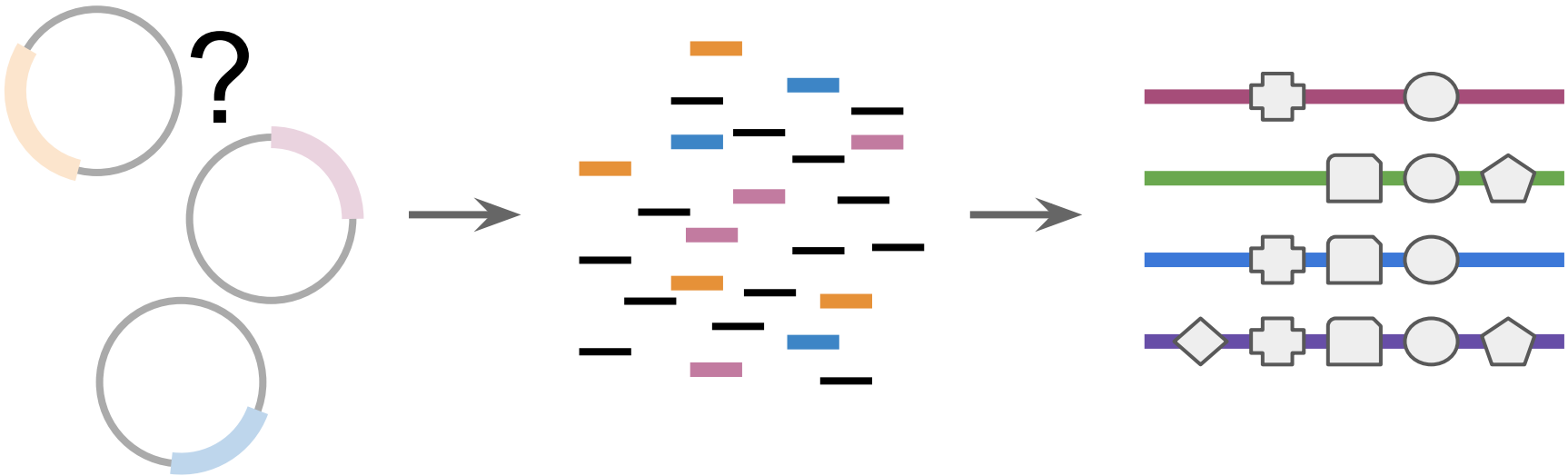

Same vs. Different Subject
Count (Species)
AU-ROC
0
50
100
150
200
250
300
350
0.5
1.0
Faecalibacterium prausnitzii
UMAP2
UMAP1
30
20
10
0
−10
−20
−10
0
10
20
A A C
T A G
T A C
T G G



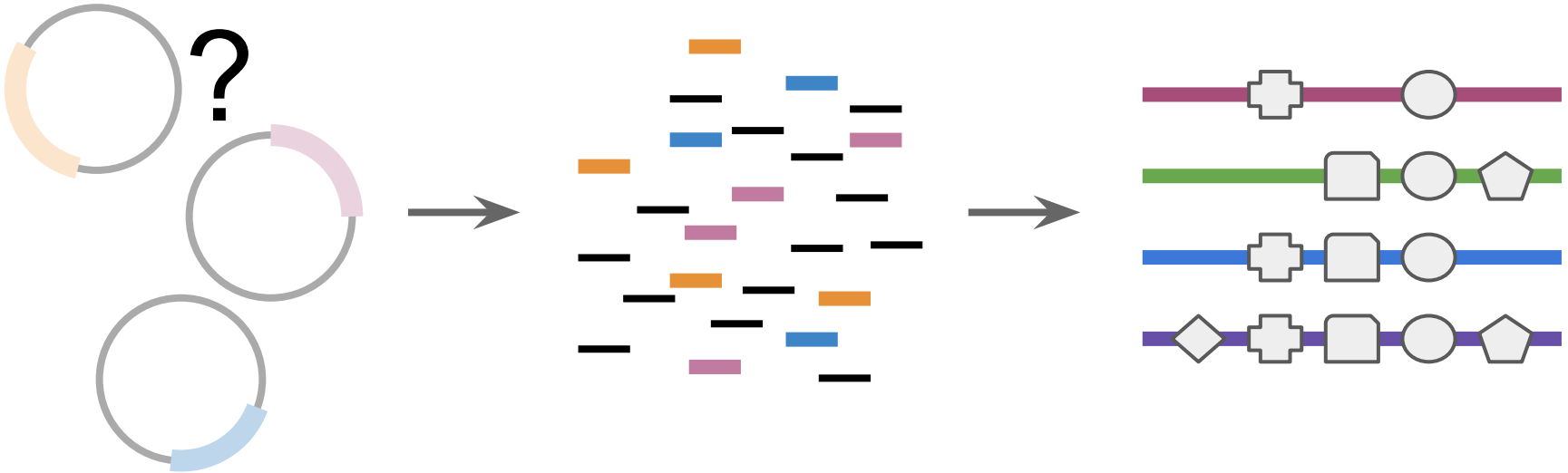
Strain diversity enables tracking of transmission between microbiomes



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

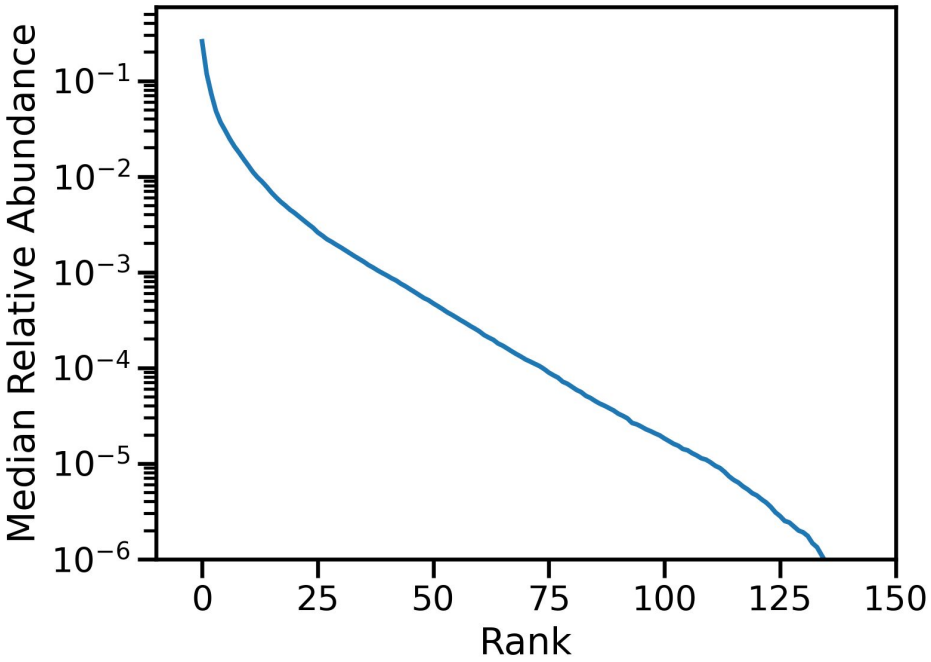


How do we accurately reconstruct strains from metagenomes?



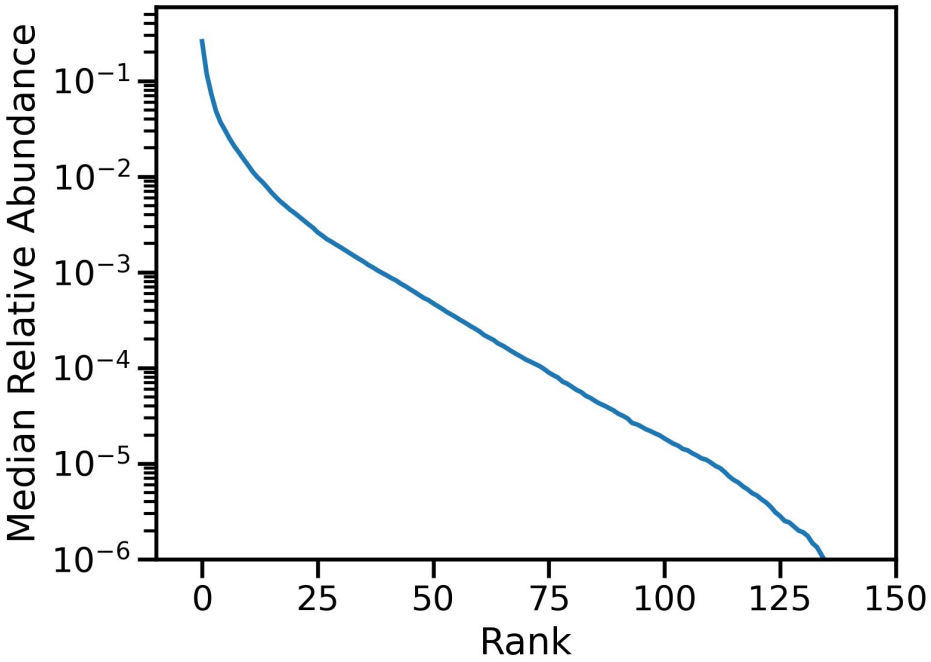
Inferring gene content accurately is difficult.

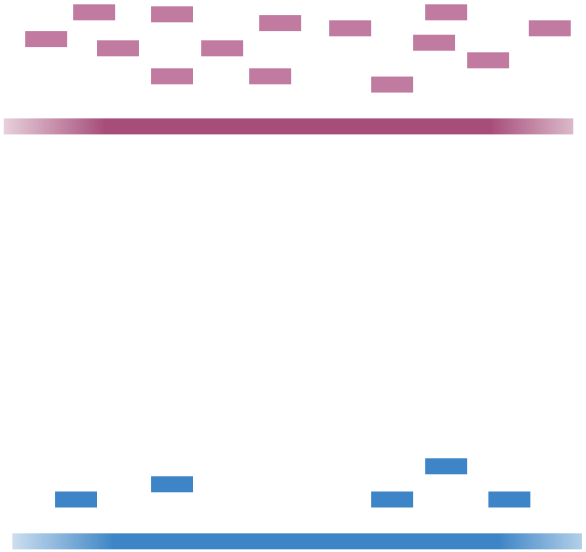
Challenge: Long tail of species abundance



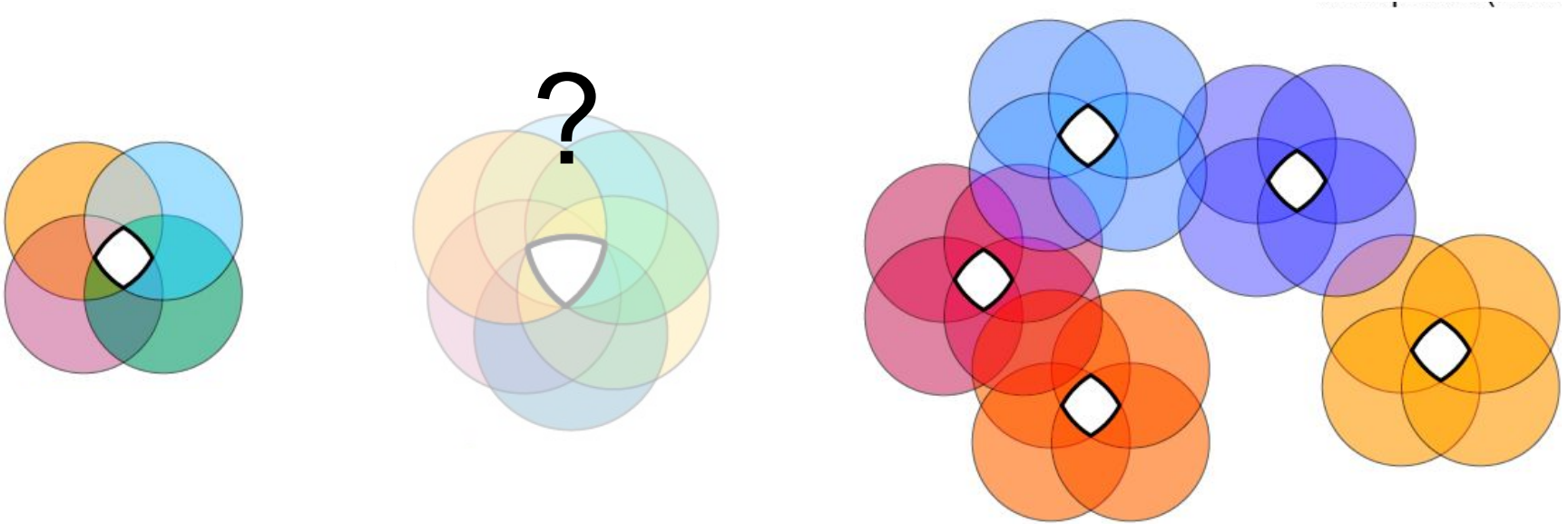
Inferring gene content accurately is difficult.

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





Challenge: Pangenomes are large, incomplete, and overlapping



wikipedia.org

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

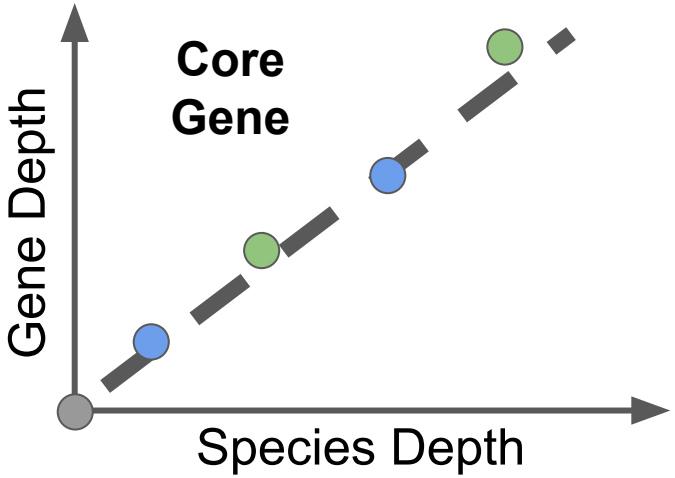
- Low abundance (sparsity)
- Missing references
- Cross-mapping from other species

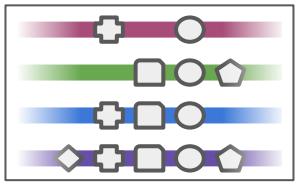
Strain-resolved gene content reconstruction: major challenges

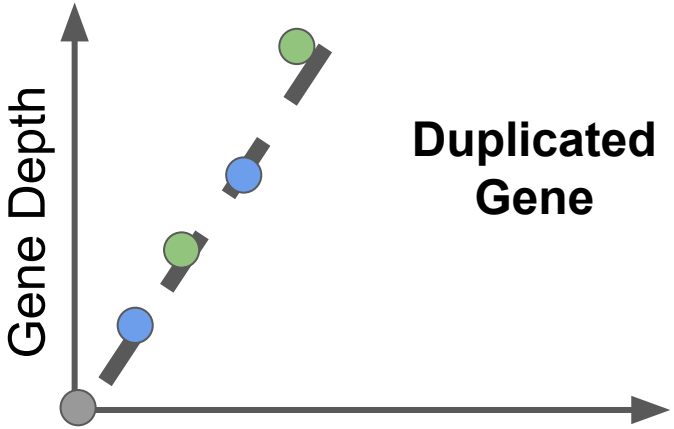
- Low abundance (sparsity)
- Missing references
- Cross-mapping from other species

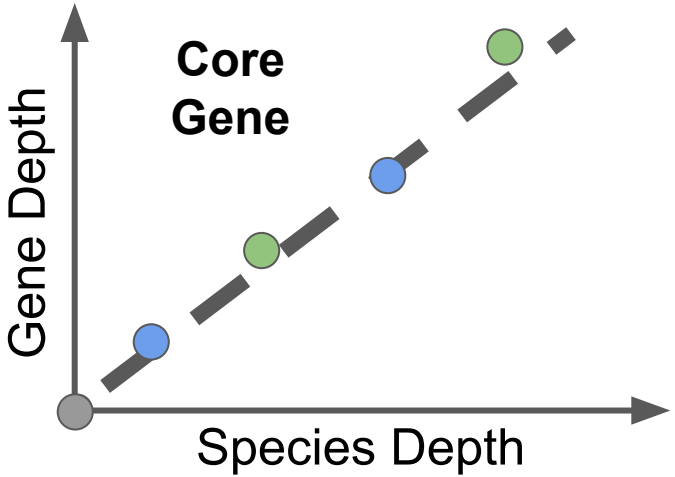
How to overcome these limitations?

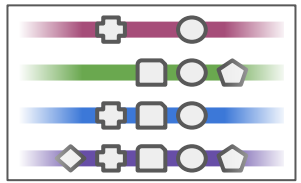
Solution: Look for correlations across multiple samples, instead of depth alone

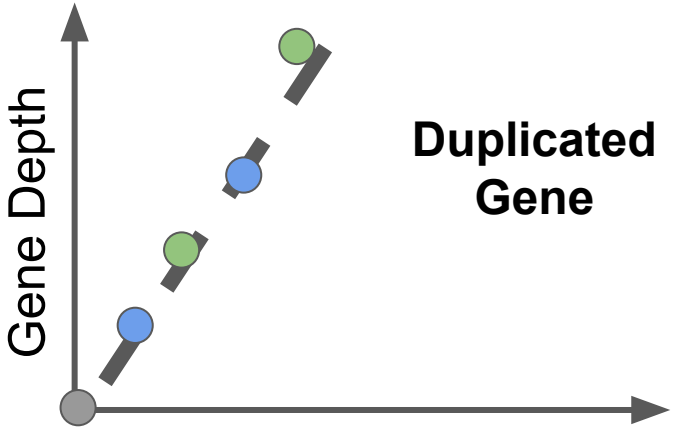


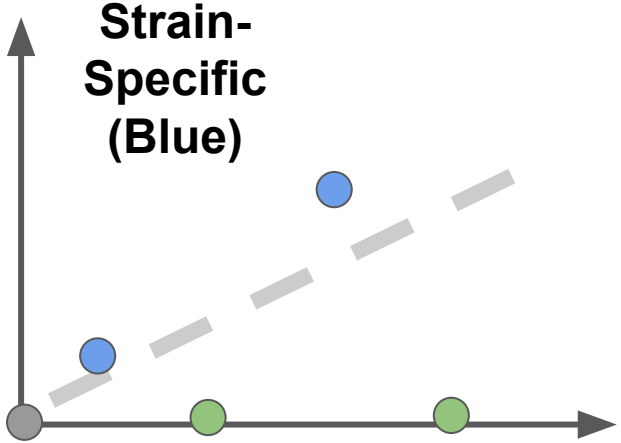


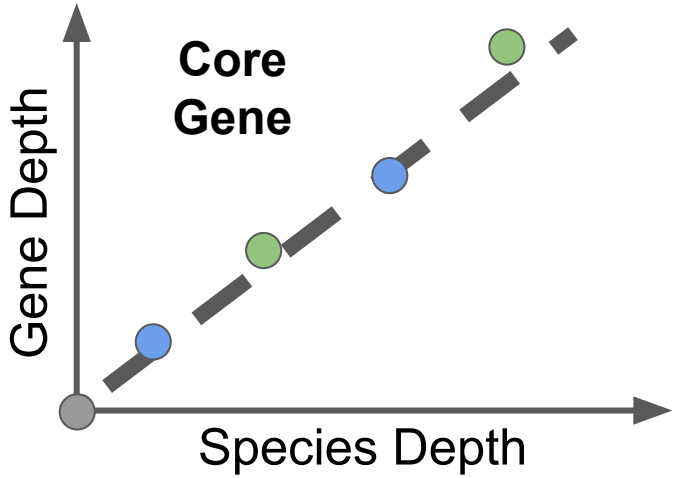

Gene Depth
Duplicated
Gene

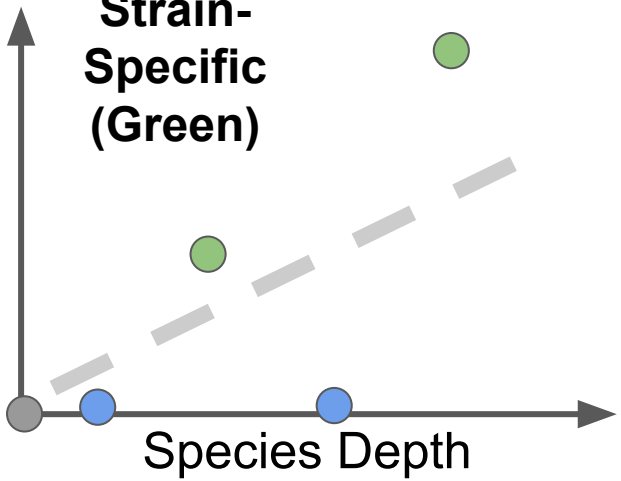

Core
Gene
Gene Depth
Species Depth





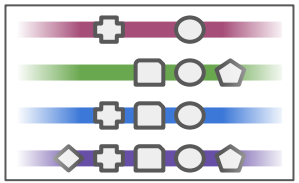






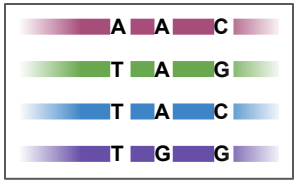
Challenge:
Strain variation

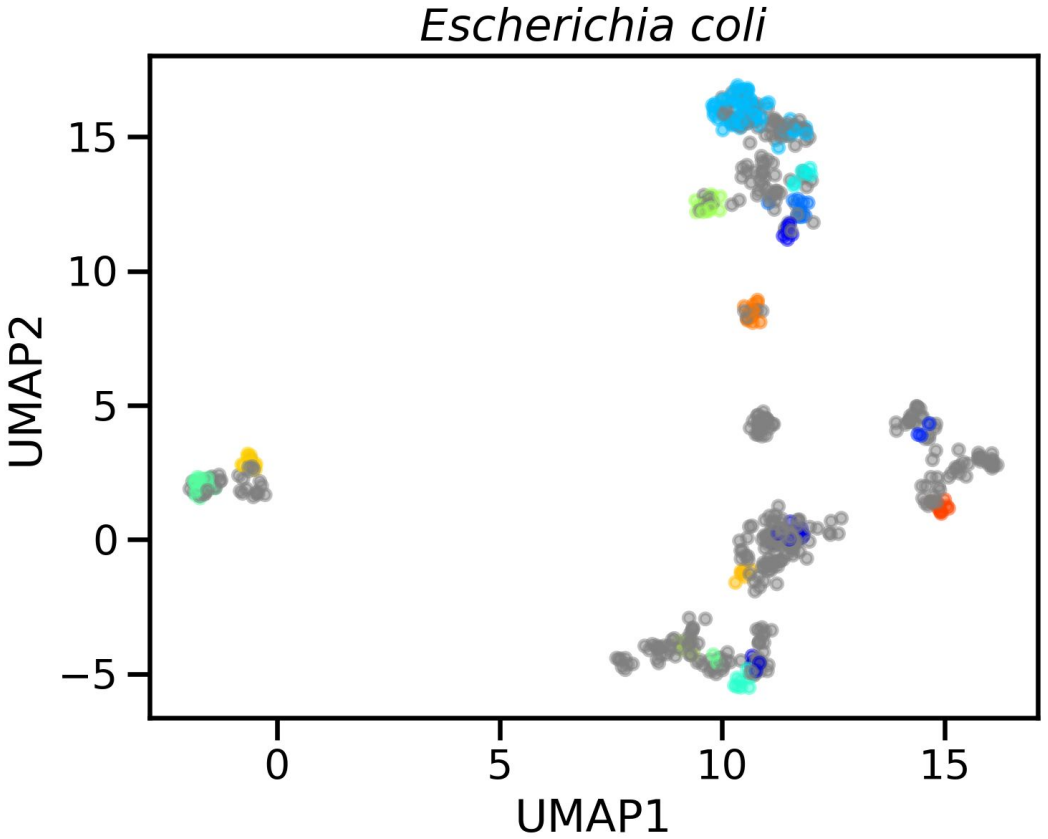
Correlations are weakened and strain-specific genes are lost due to inconsistent depth.

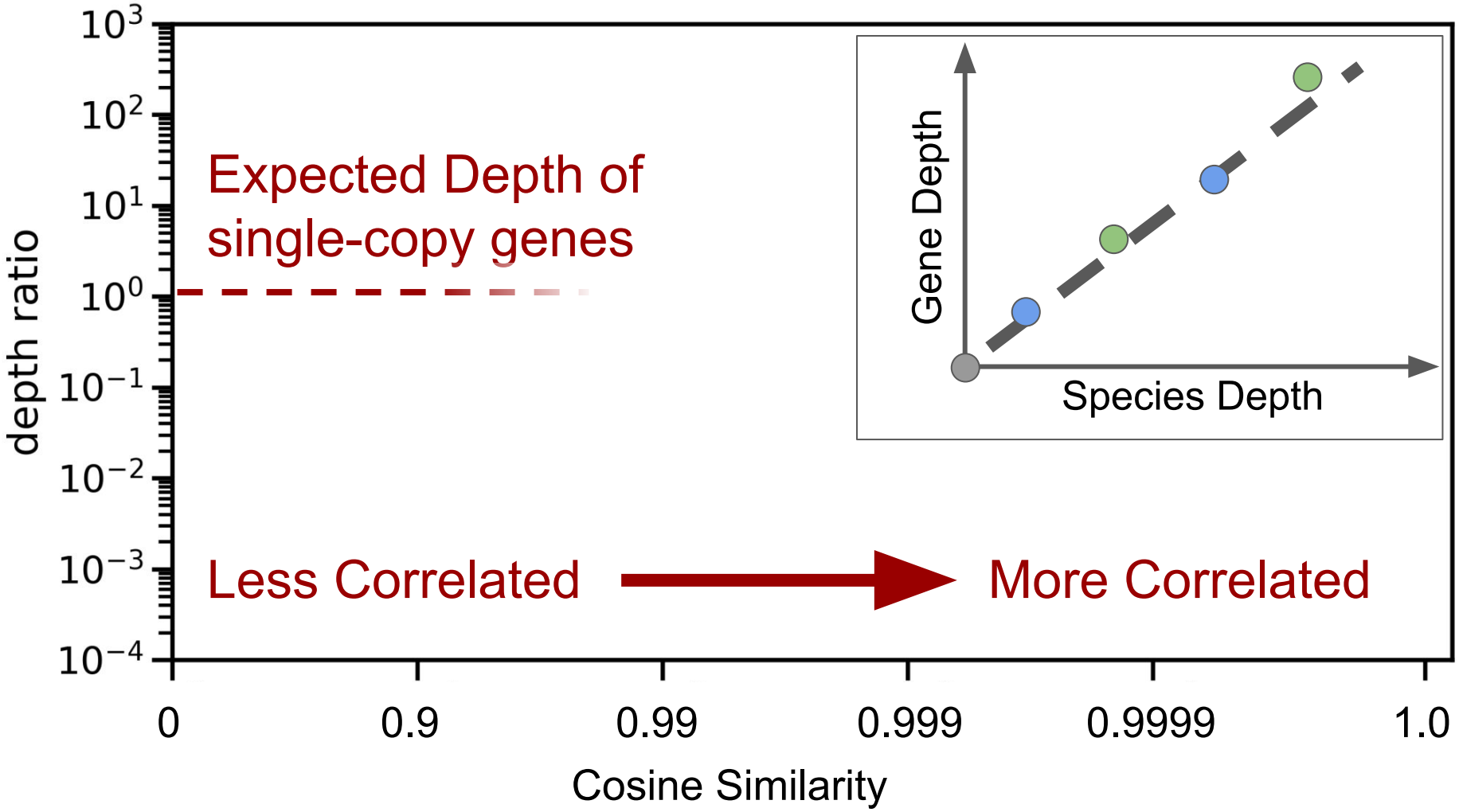


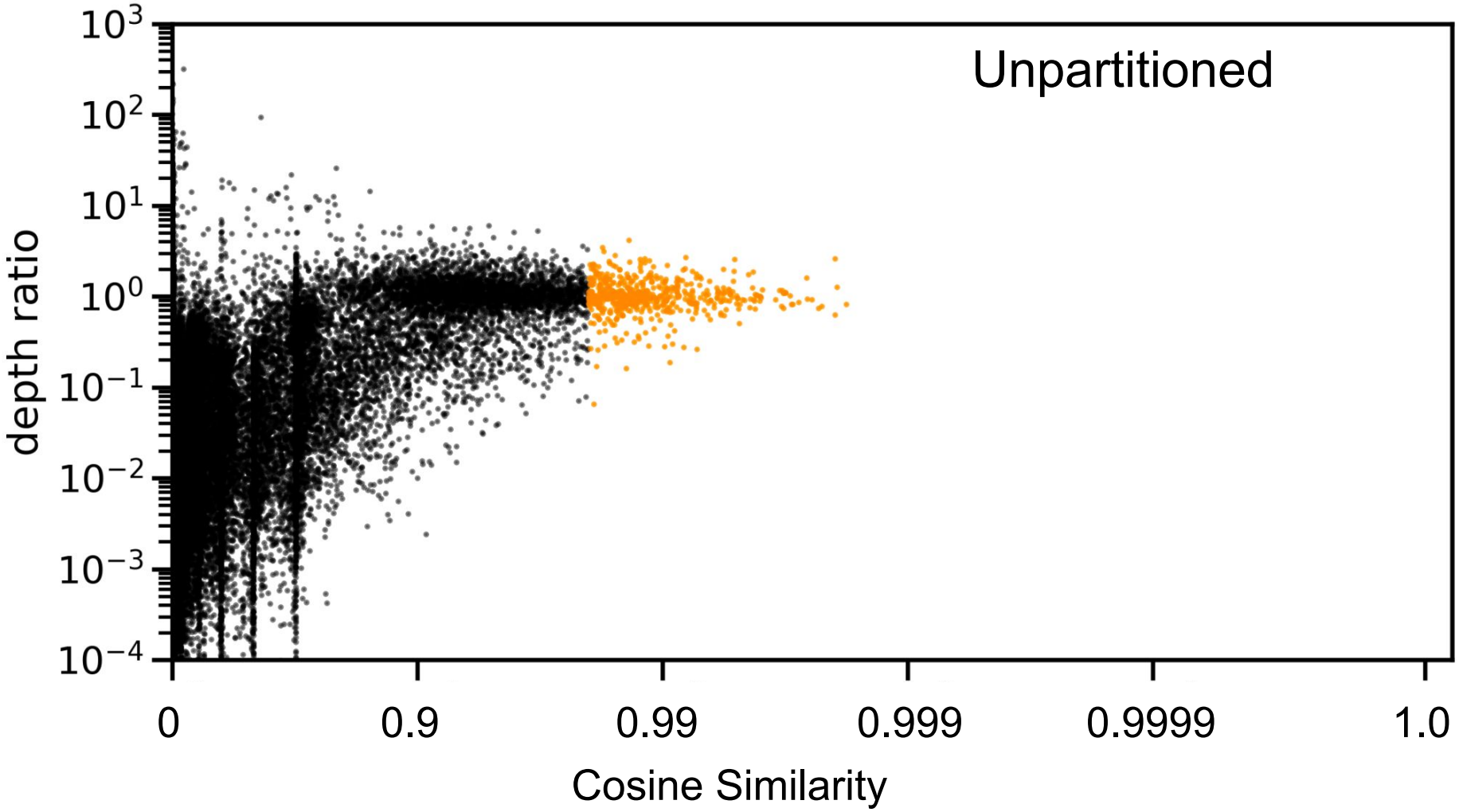
Solution: Partition samples by strain

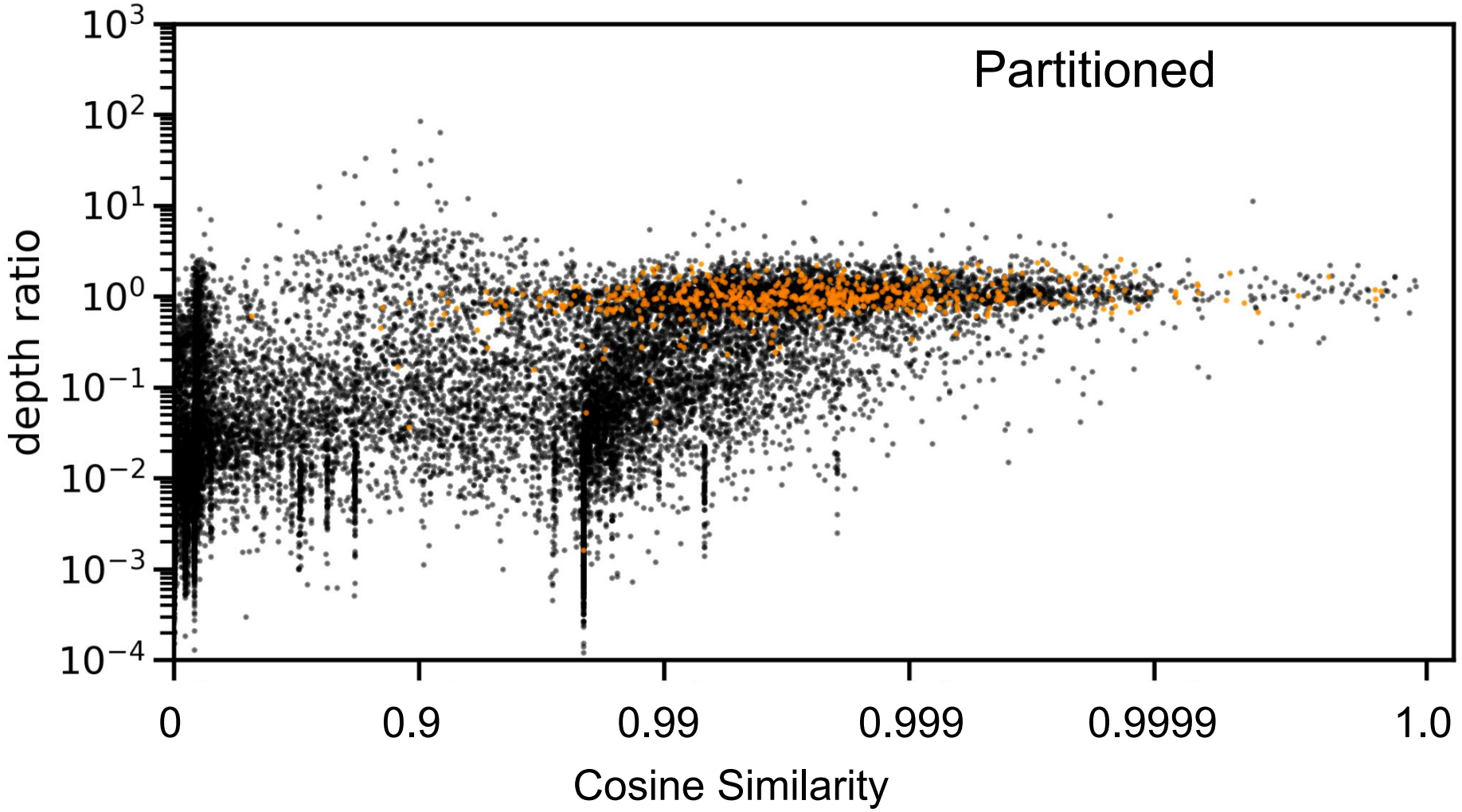
Use strain SNP profiles to select pure samples of each strain.

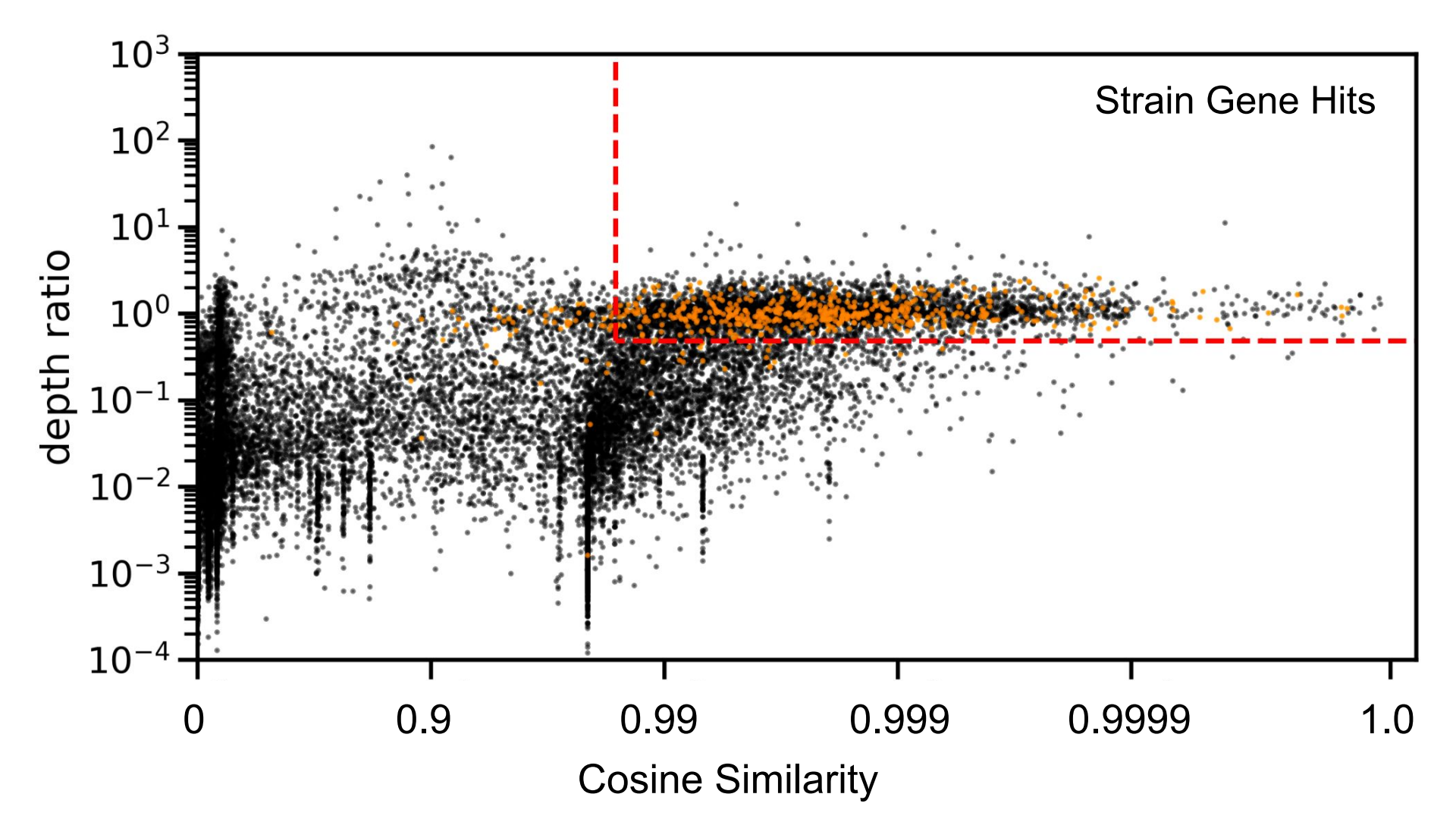




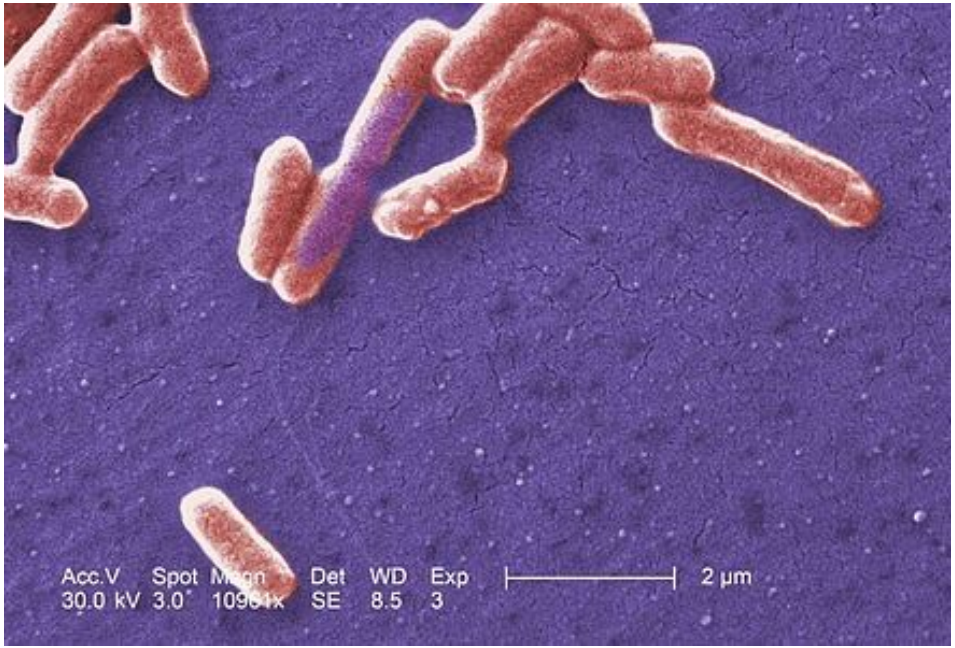

depth ratio
10^{3}
10^{2}
10^{1}
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
Expected Depth of single-copy genes
Less Correlated
More Correlated
Gene Depth
Species Depth
0
0.9
0.99
0.999
0.9999
1.0
Cosine Similarity


Unpartitioned
depth ratio
10^{3}
10^{2}
10^{1}
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
0
0.9
0.99
0.999
0.9999
1.0
Cosine Similarity


Partitioned
depth ratio
10^{3}
10^{2}
10^{1}
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
0
0.9
0.99
0.999
0.9999
1.0
Cosine Similarity

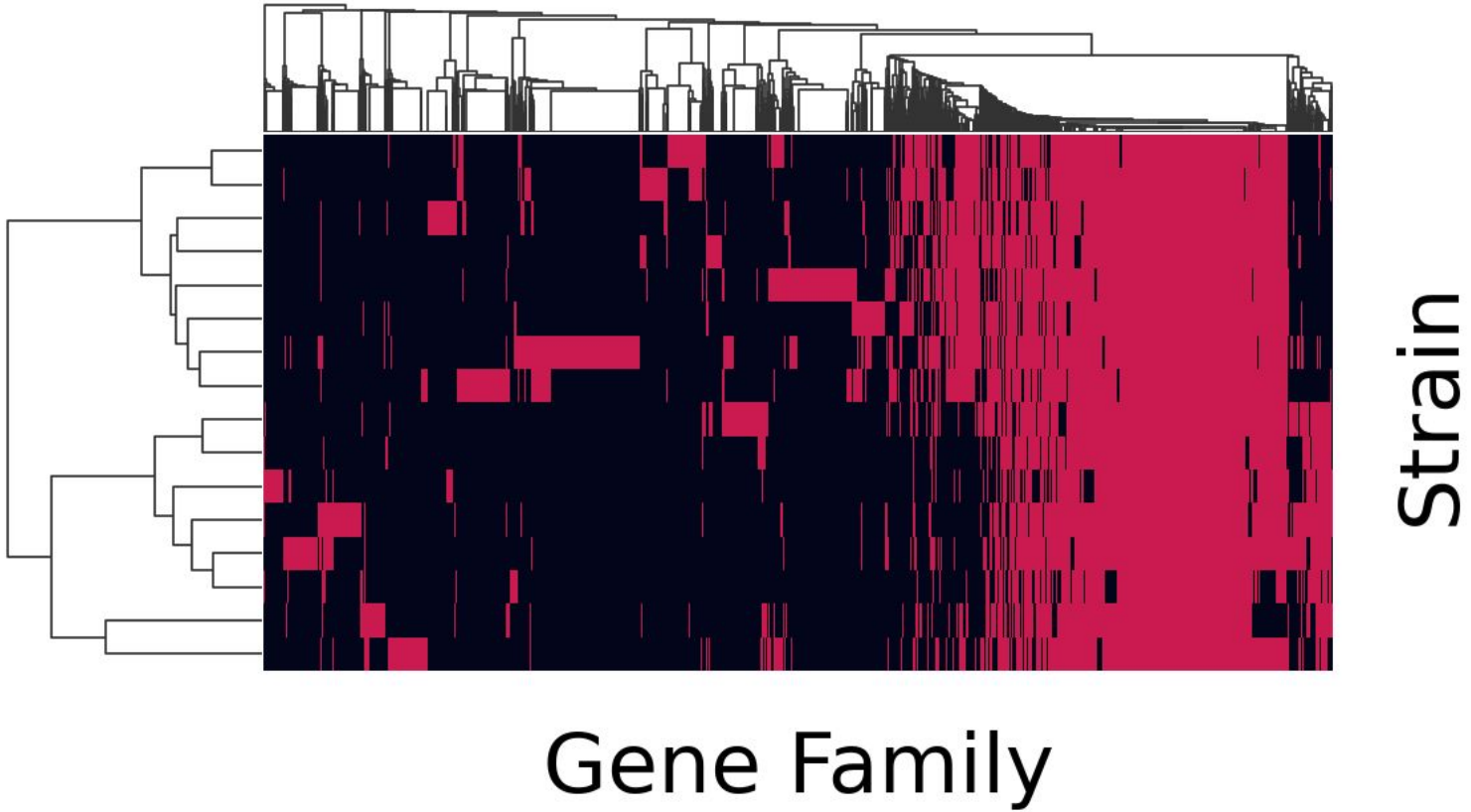

Strain Gene Hits
depth ratio
10^{3}
10^{2}
10^{1}
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
0
0.9
0.99
0.999
0.9999
1.0
Cosine Similarity

Genes inferred for 16 distinct *E. coli* strains

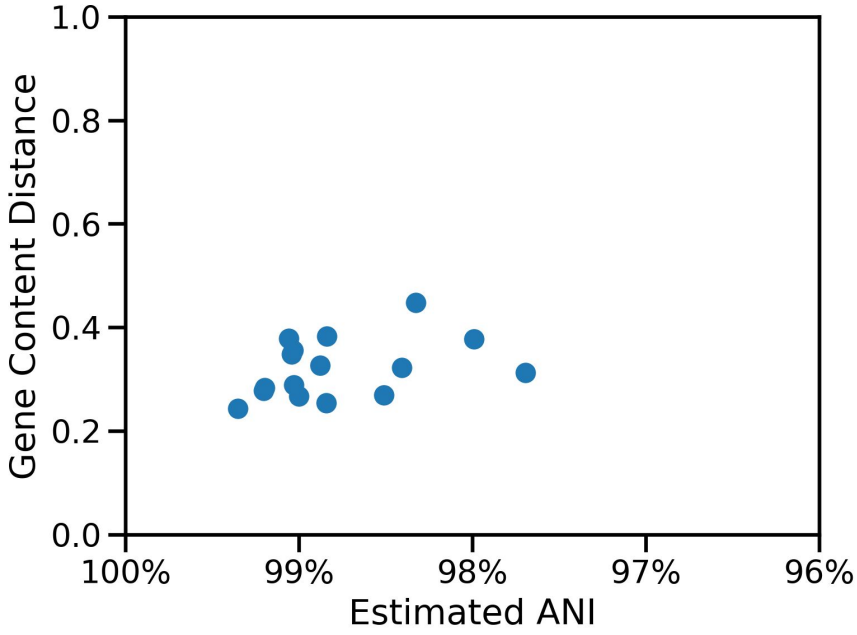


wikipedia.org

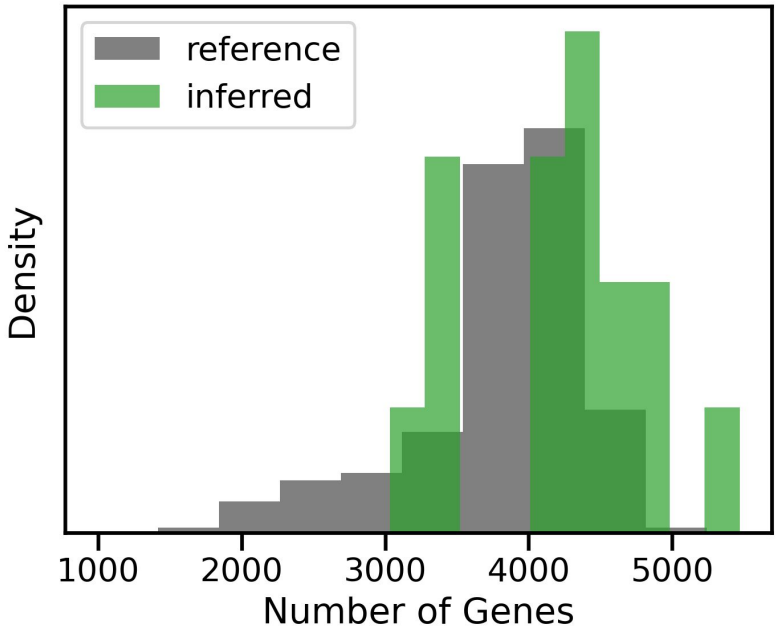
Inferred genes for 16 distinct *E. coli* strains

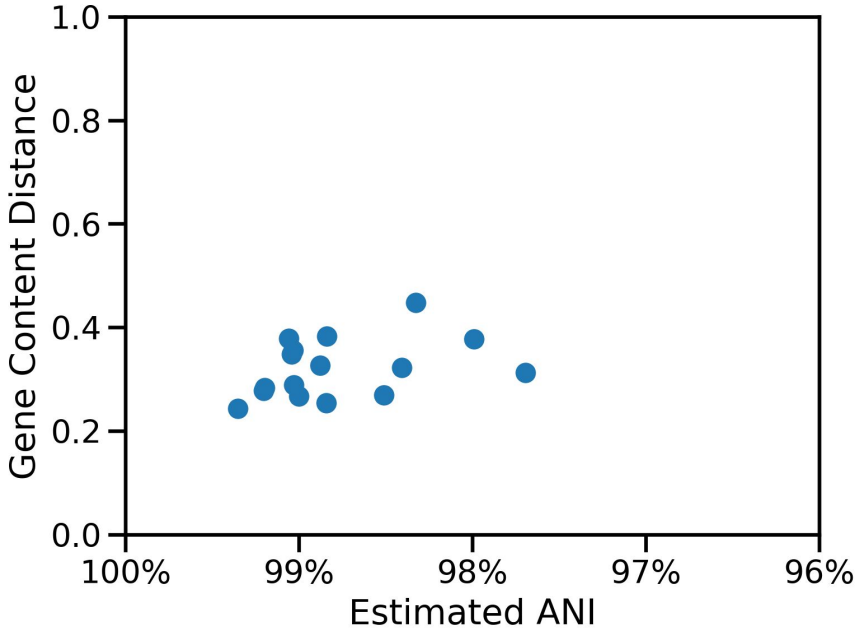


Strain genotypes and gene content are both different from existing references

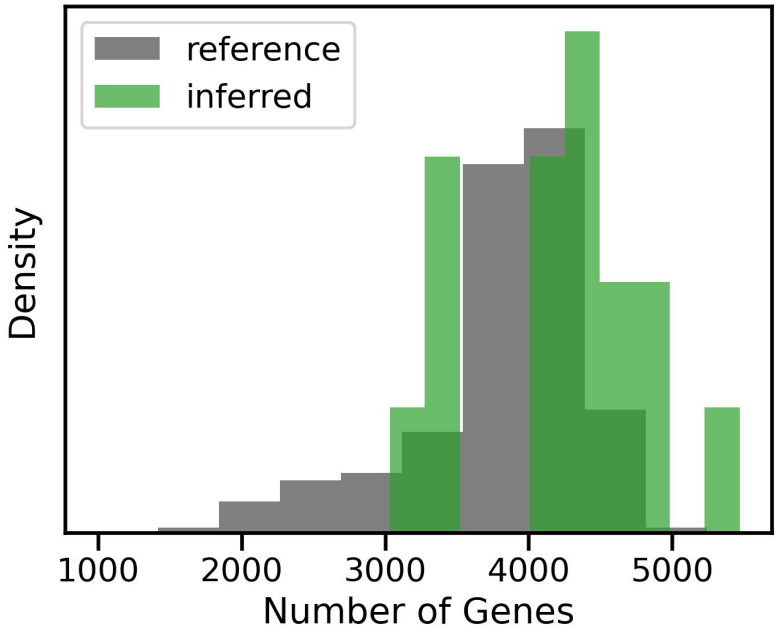


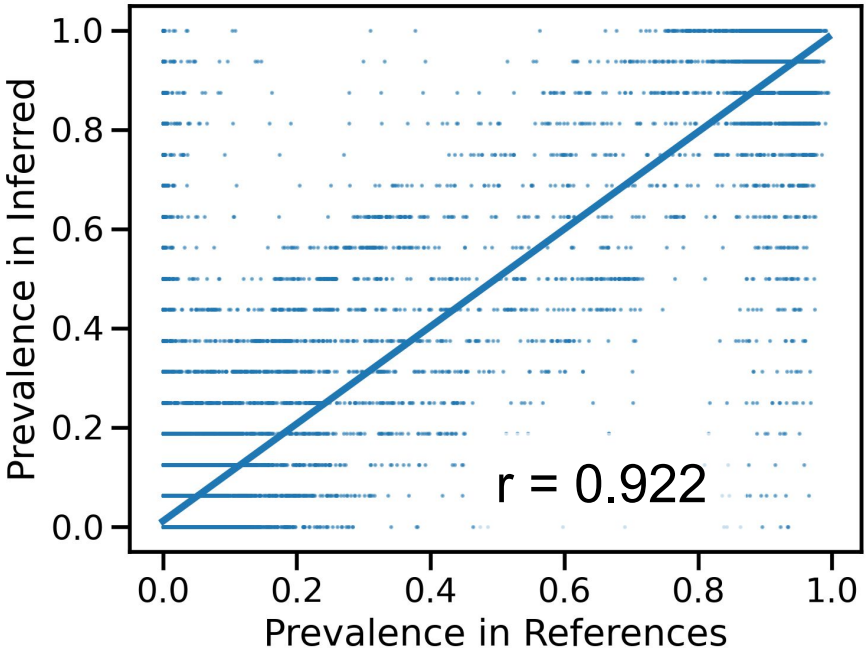
Genome size and gene prevalence are consistent with reference databases



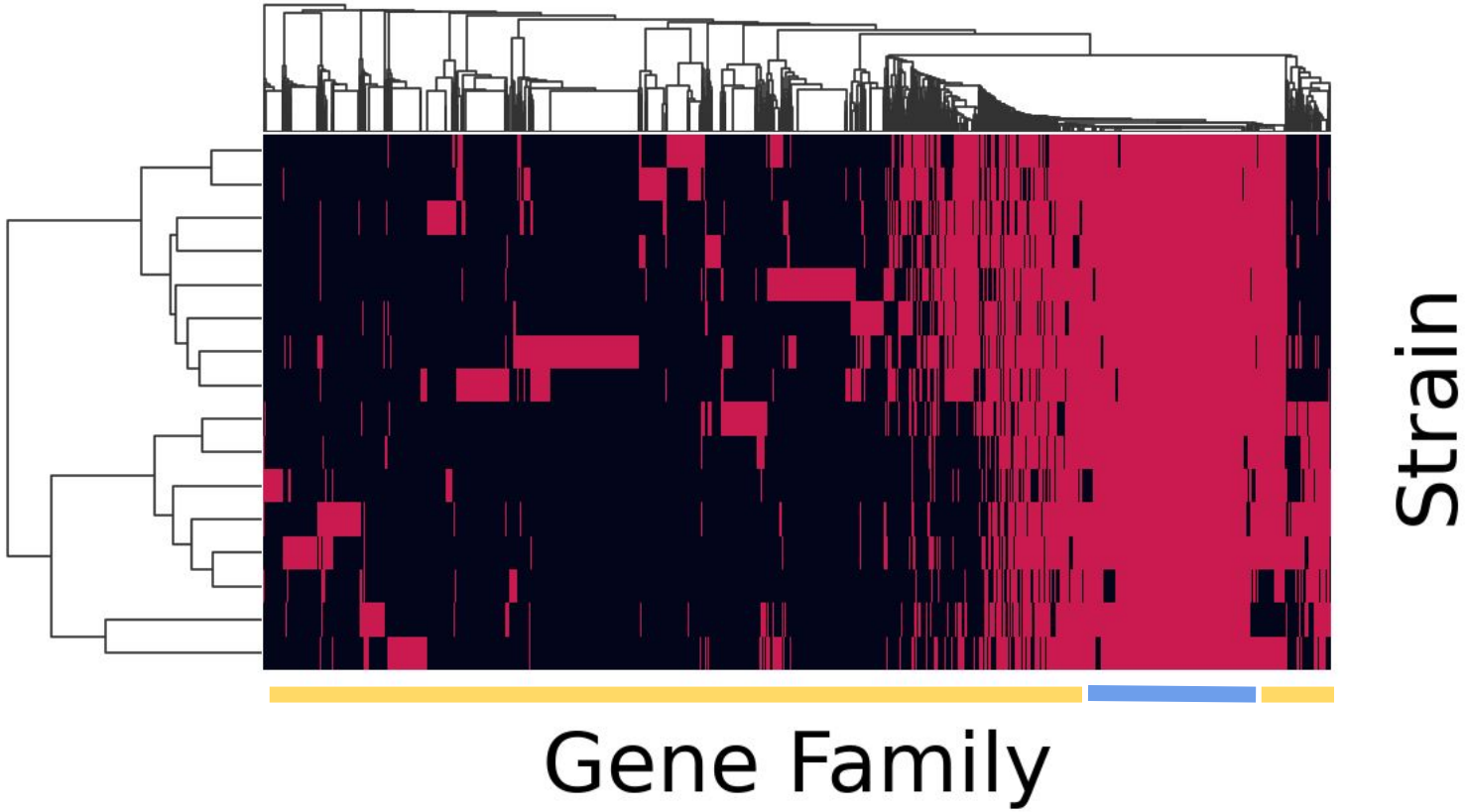


Genome size and gene prevalence are consistent with reference databases

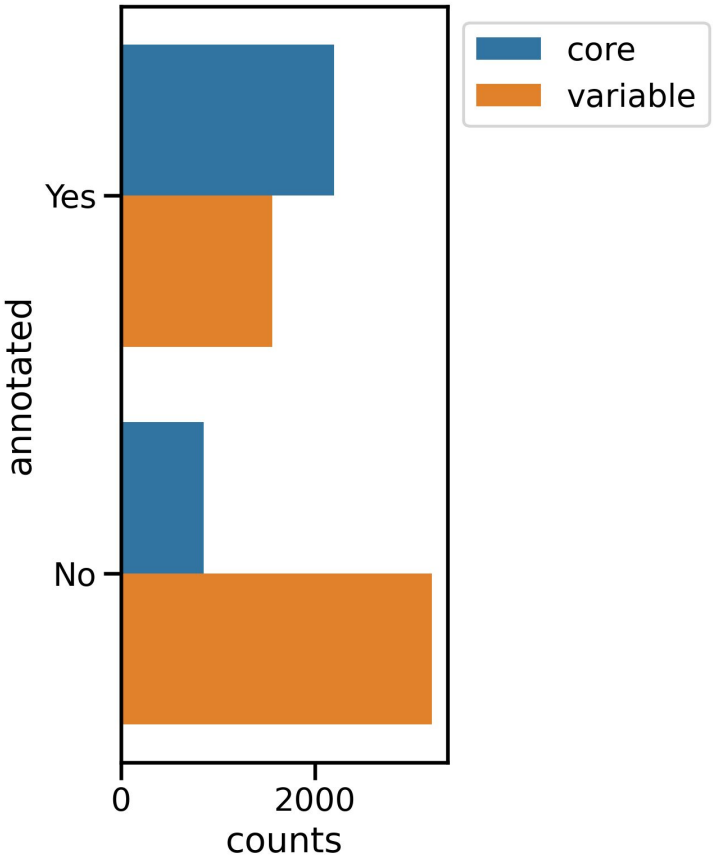




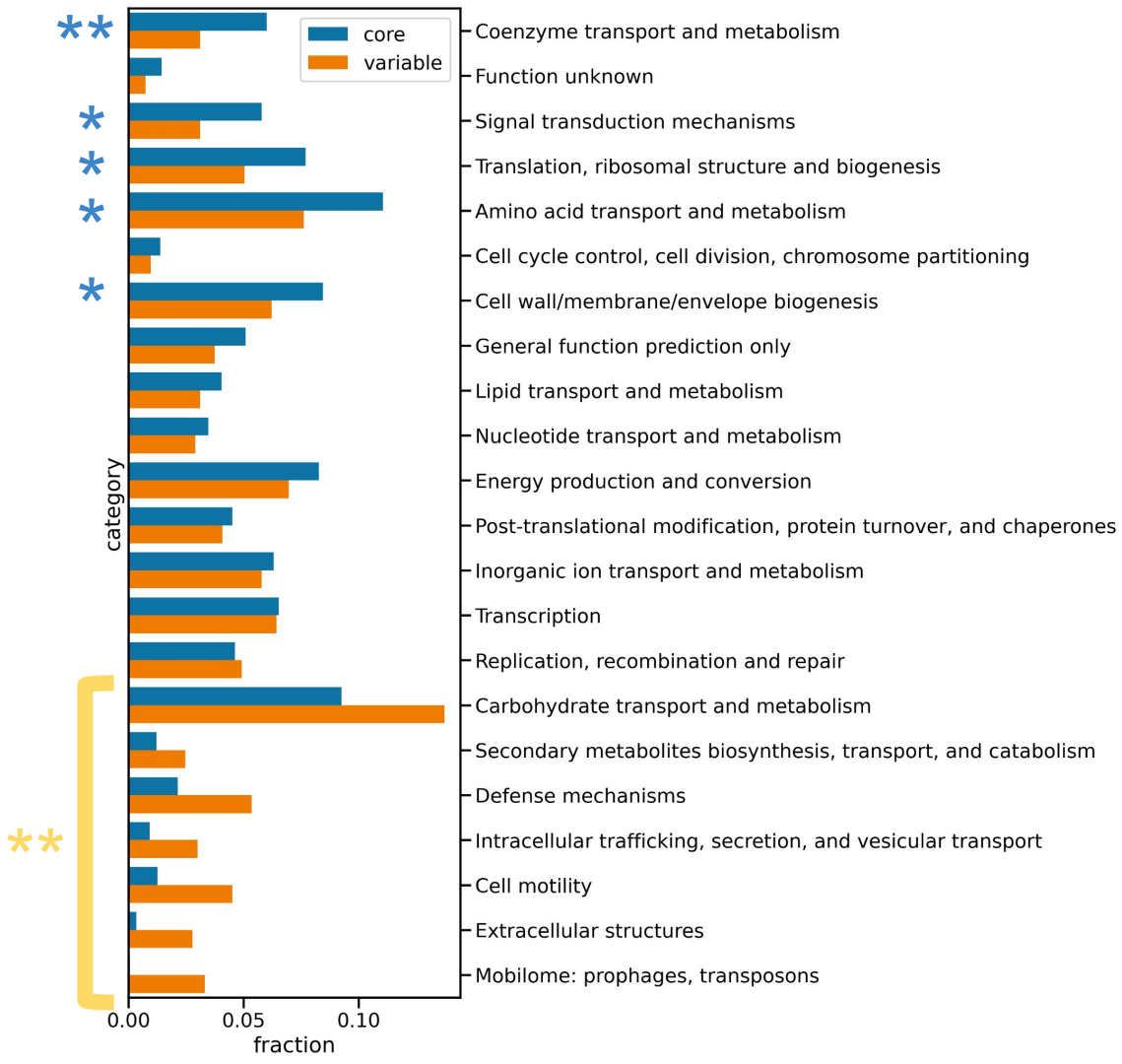
Inferred genes for 16 distinct *E. coli* strains



The variable fraction is enriched with un-annotated genes.



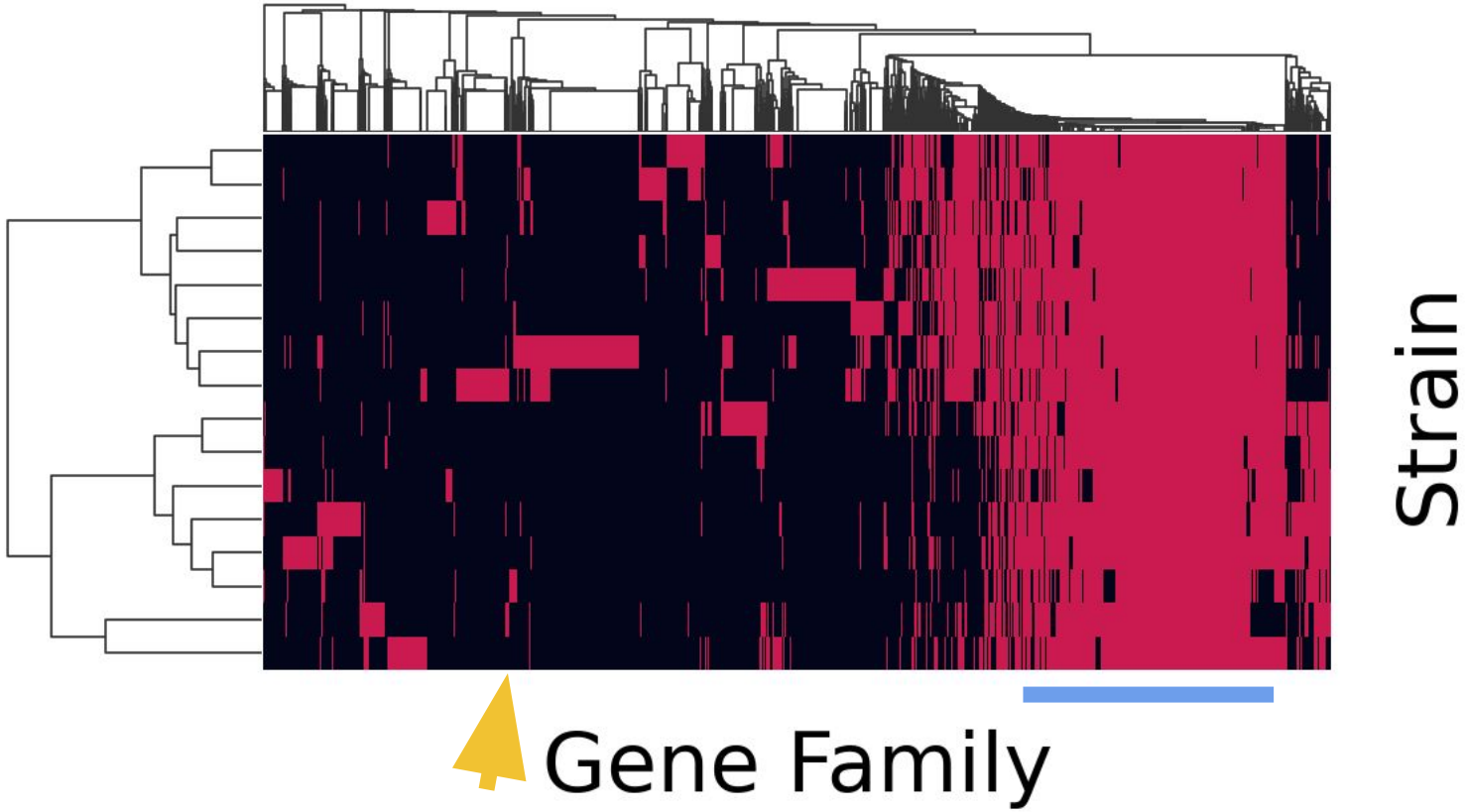
Model lab strains and other isolates may be insufficient for understanding physiology in the gut microbiome.



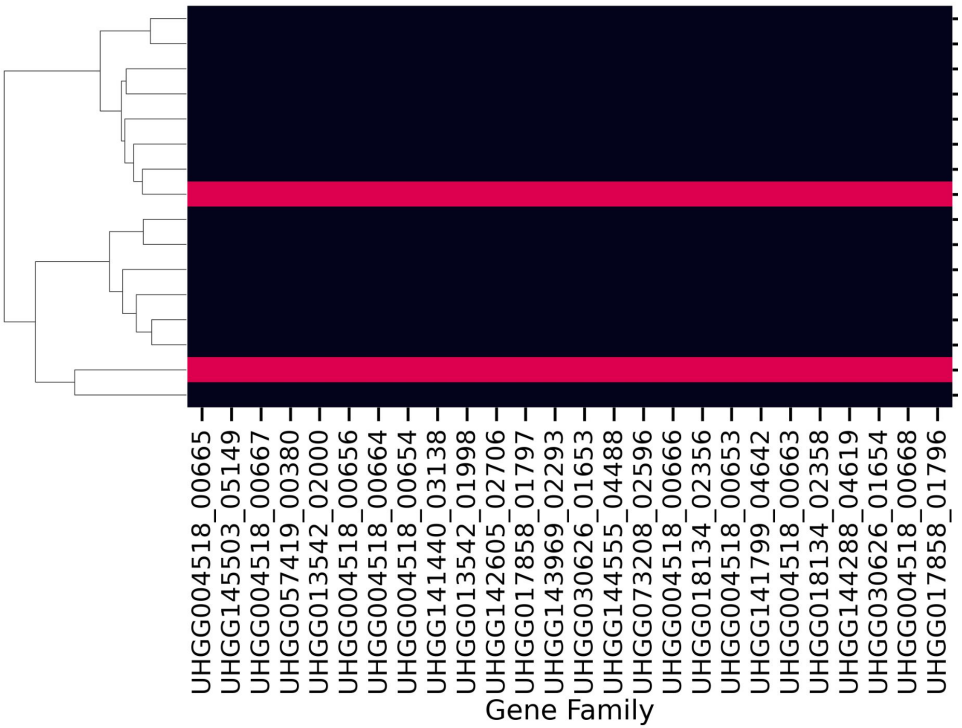
Among COG-annotated genes, variable genome is enriched with important functional categories, e.g.:

- 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

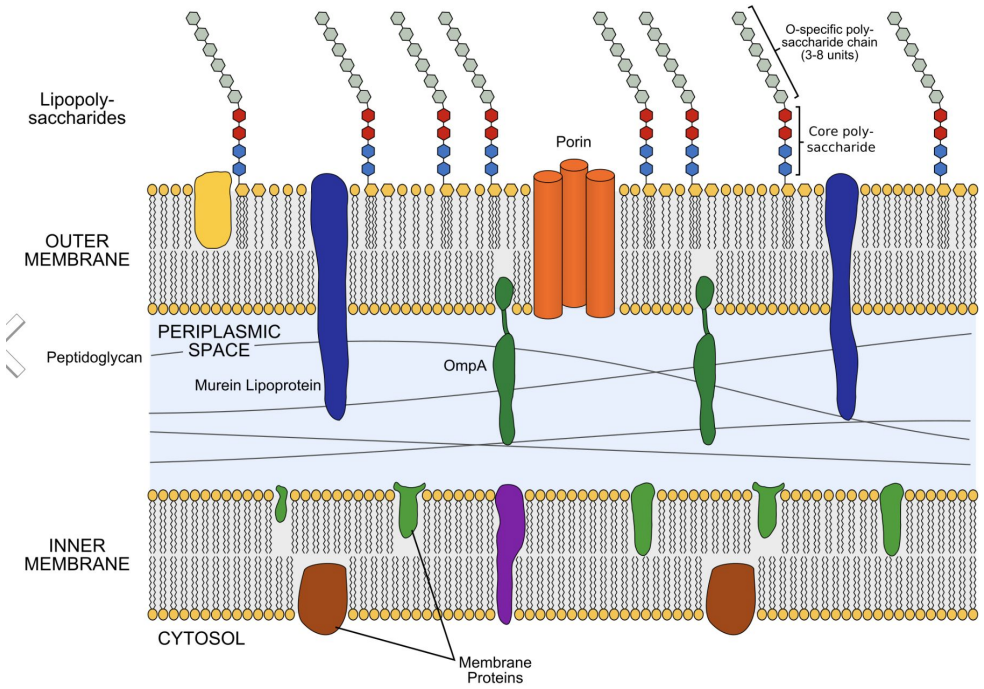
- fdtC

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)

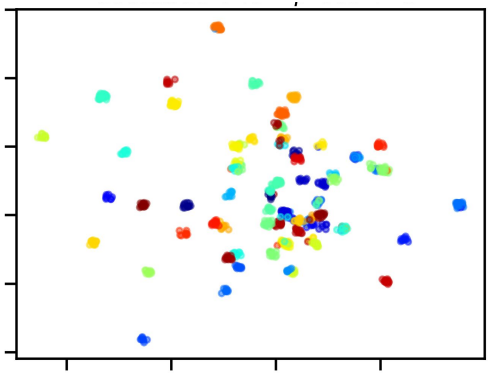
- rfbB/C/D (COG1088, COG1898, COG1091)
- rmlA (COG1209)

S-layer glycoprotein synthesis

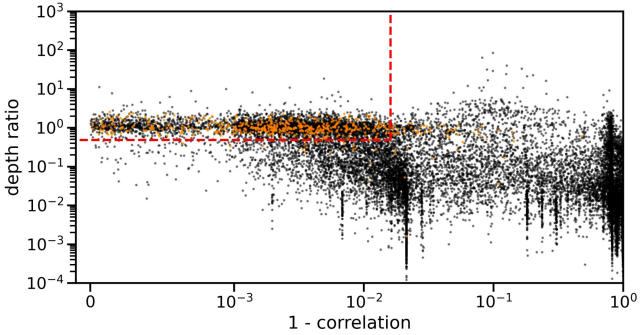
- fdtC

18 un-annotated proteins

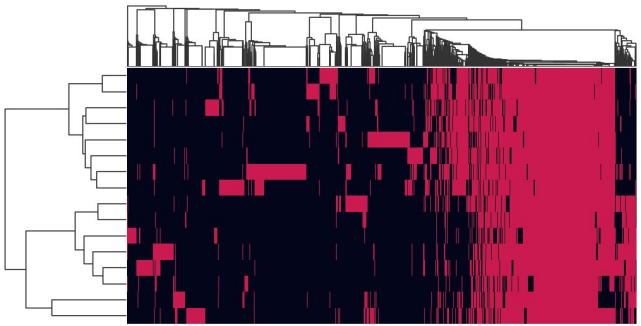
Enormous Strain Diversity



Strain-informed Gene Inference



Core and Variable Gene Content



Functional Enrichment in Variable Fraction