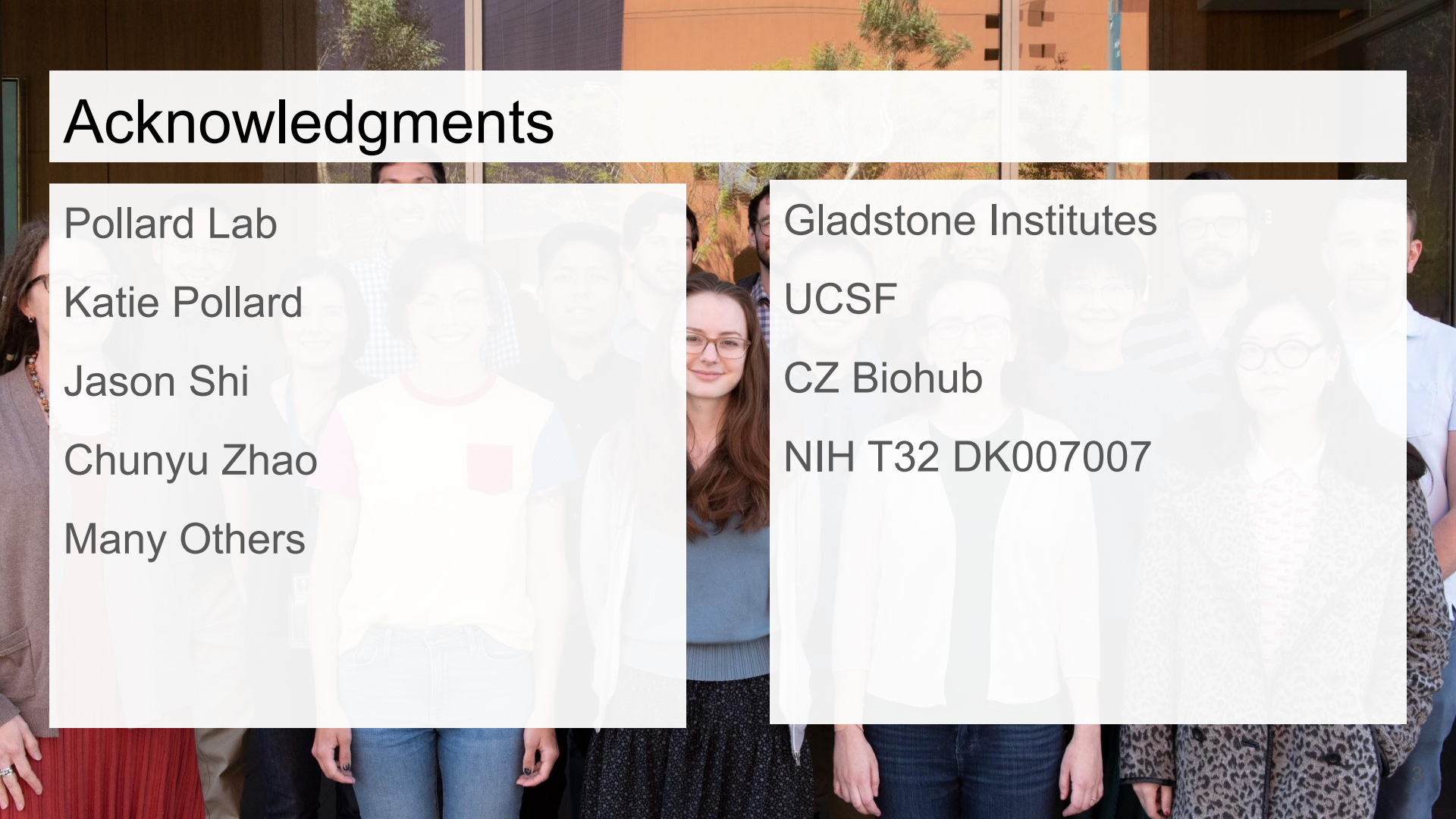
Identifying and tracking bacterial strains in metagenomic libraries

Byron J. Smith
Tools&Tech RIPS
2020-12-07



Acknowledgments

Pollard Lab

Katie Pollard

Jason Shi

Chunyu Zhao

Many Others

Gladstone Institutes

UCSF

CZ Biohub

NIH T32 DK007007

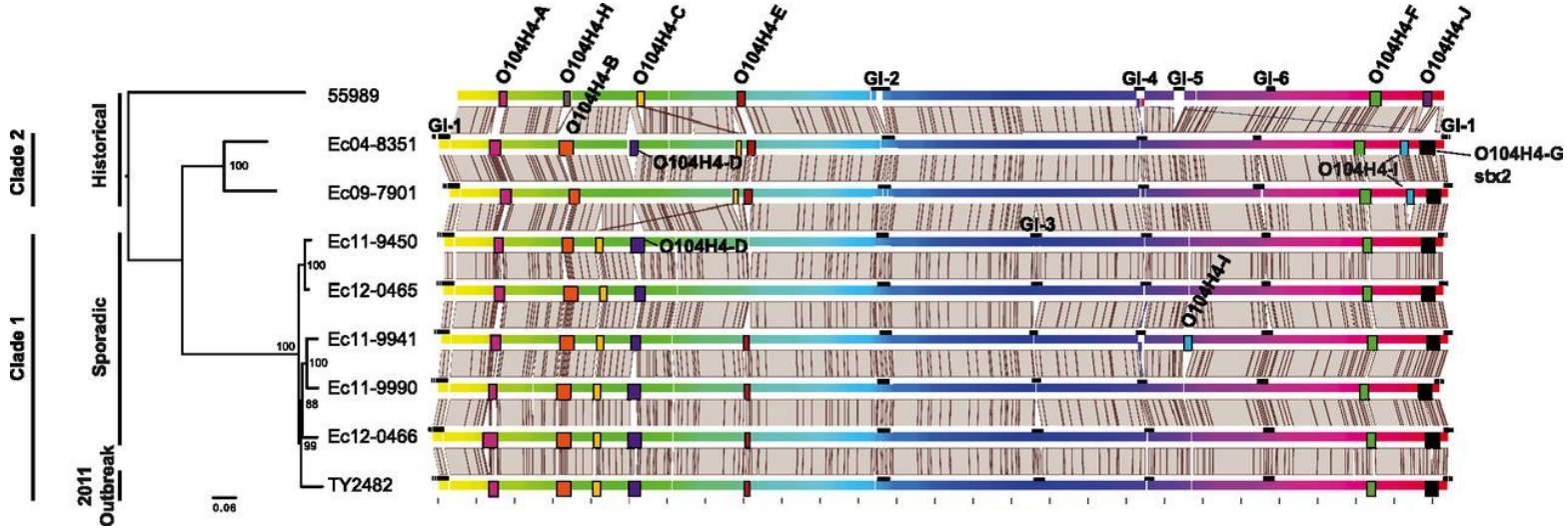
The human microbiome is important and diverse

- Involved in numerous physiological processes:
  - Digestion
  - Immune regulation
  - Pathogen resistance
- Complex community
  - Highly variable across individuals
  - Hundreds of bacterial species
  - Huge but under-explored diversity *within* species

Diversity within species (strains)

- Variation in gene sequences and content within species
- Relevant effects:

  e.g. Antibiotic resistance, Drug metabolism, Immune evasion, Toxin production



Grad et al. 2013 mBio

Diversity within species (strains)

- Variation in gene sequences and content within species
- Relevant effects:

  e.g. Antibiotic resistance, Drug metabolism, Immune evasion, Toxin production

- Intraspecific diversity is not captured by SOP taxonomic surveys
  - 16S rRNA gene evolves too slowly

How to survey bacterial strains?

- Culturing is the “Gold Standard”
- Enables phys. characterization
- Low-throughput, expensive, biased

- Metagenomics
- Easy, high-throughput, less biased
- Interpretation often difficult

Identifying strains in metagenomes

- Most strain diversity has not been previously documented
- Alignment and assembly based methods have been important tools, but require sufficient **sequencing depth** and **strain-pure samples**

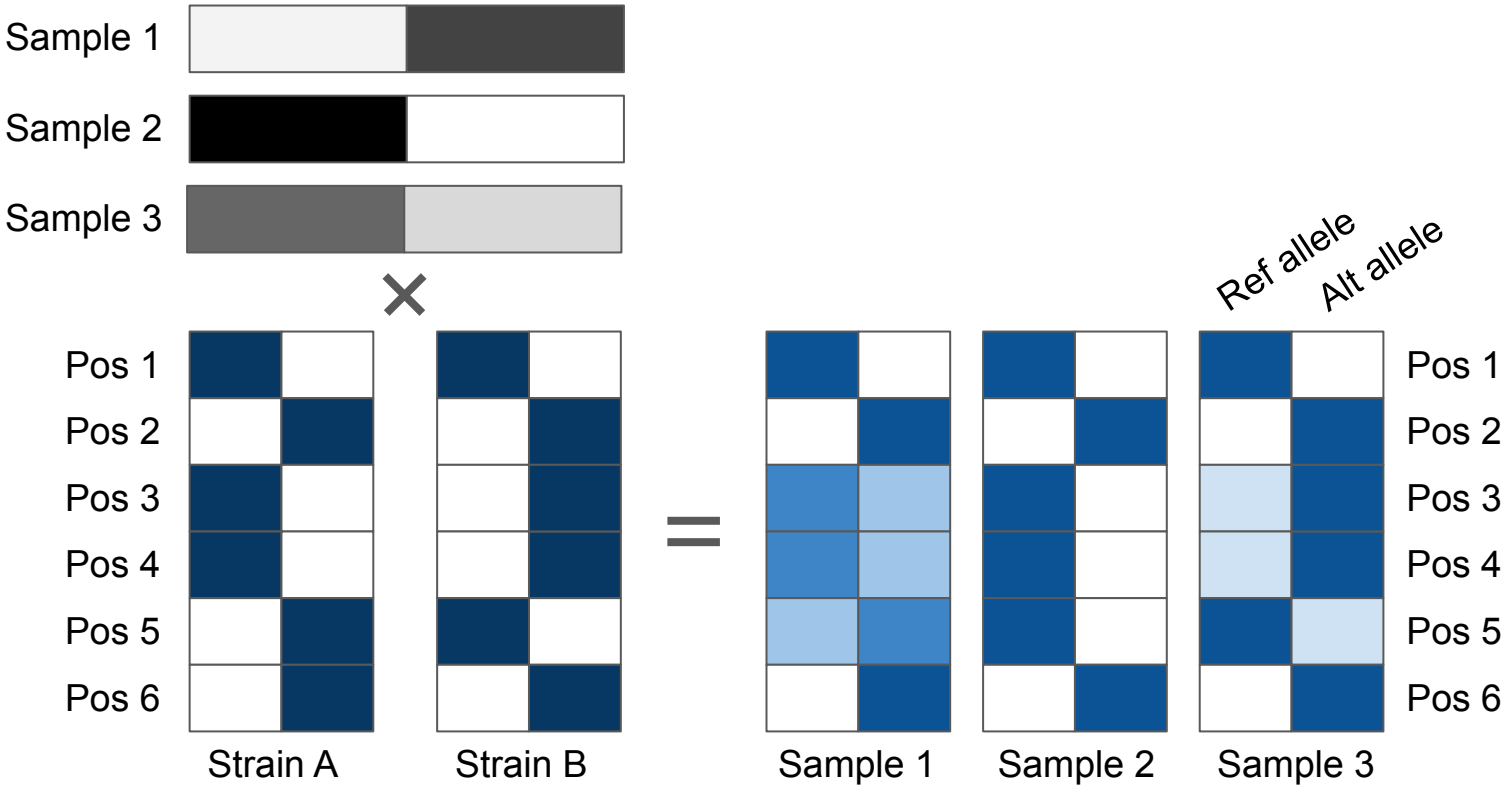
- Tools are available for high-throughput
- “metagenotyping”
  - e.g. GT-PRO

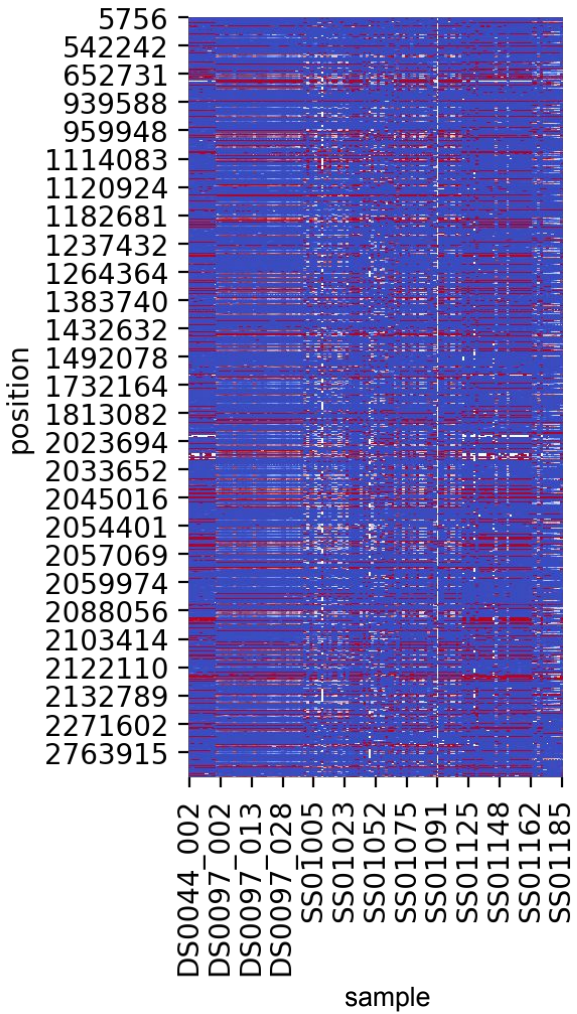
Identifying strains in metagenomes

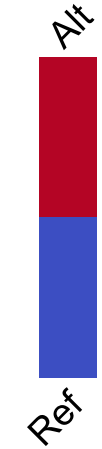
- Tools are available for high-throughput “metagenotyping”
- **How to recover strain genotype/abundance from metagenotype?**

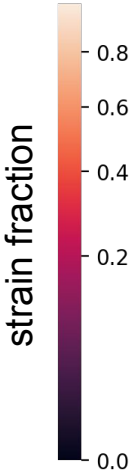
Species A: Sample 001		
Genome Position	Ref (Count)	Alt (Count)
0001	A (8)	C (1)
0114	T (11)	C (0)
1005	C (0)	G (7)
1202	C (12)	T (5)
...	...	...

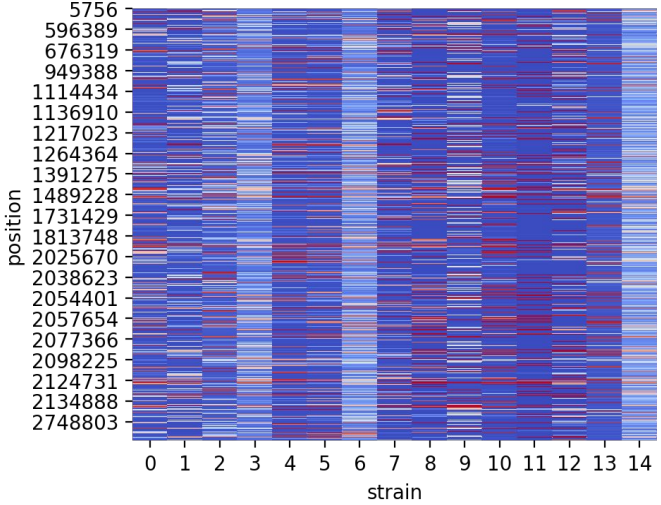
Resolving strains from metagenotype matrix

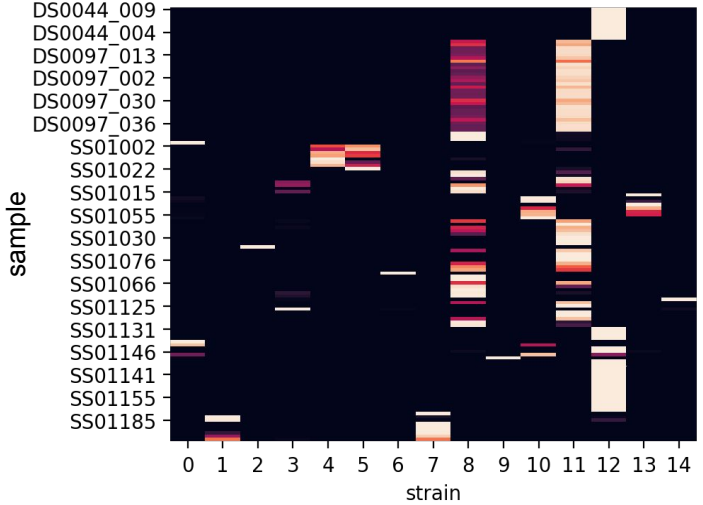



position
5756
542242
652731
939588
959948
1114083
1120924
1182681
1237432
1264364
1383740
1432632
1492078
1732164
1813082
2023694
2033652
2045016
2054401
2057069
2059974
2088056
2103414
2122110
2132789
2271602
2763915
DS0044_002
DS0097_002
DS0097_013
DS0097_028
SS01005
SS01023
SS01052
SS01075
SS01091
SS01125
SS01148
SS01162
SS01185
sample


Alt
Ref


strain fraction
0.8
0.6
0.4
0.2
0.0


position
5756
596389
676319
949388
1114434
1136910
1217023
1264364
1391275
1489228
1731429
1813748
2025670
2038623
2054401
2057654
2077366
2098225
2124731
2134888
2748803
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14
strain


sample
DS0044_009
DS0044_004
DS0097_013
DS0097_002
DS0097_030
DS0097_036
SS01002
SS01022
SS01015
SS01055
SS01030
SS01076
SS01066
SS01125
SS01131
SS01146
SS01141
SS01155
SS01185
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14
strain

GT-PRO metagenotypes work great

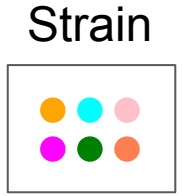
Pros:

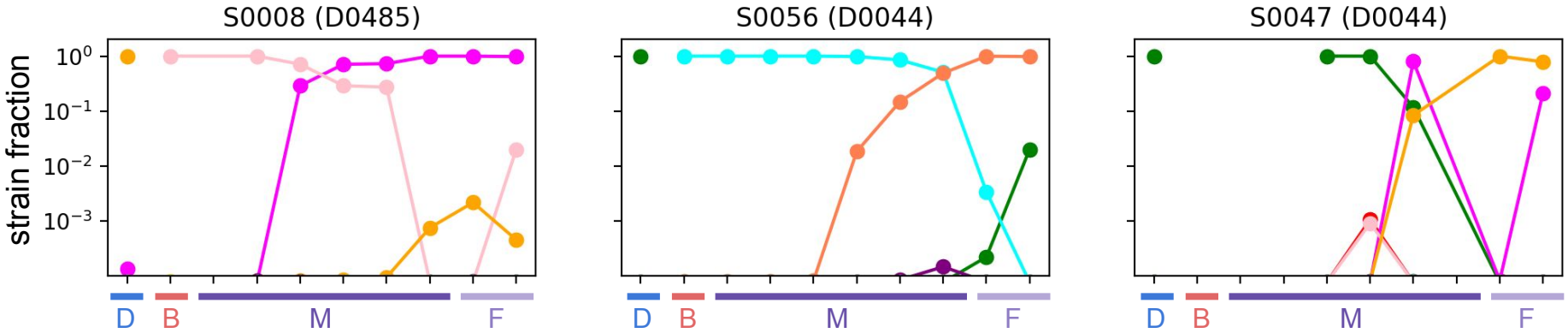
- Preconstructed DB of core genome SNPs
- Fast
- Pre-computed results for >20,000 publicly available metagenomes

Cons:

- Can only use known alleles

We can track strains in an FMT experiment





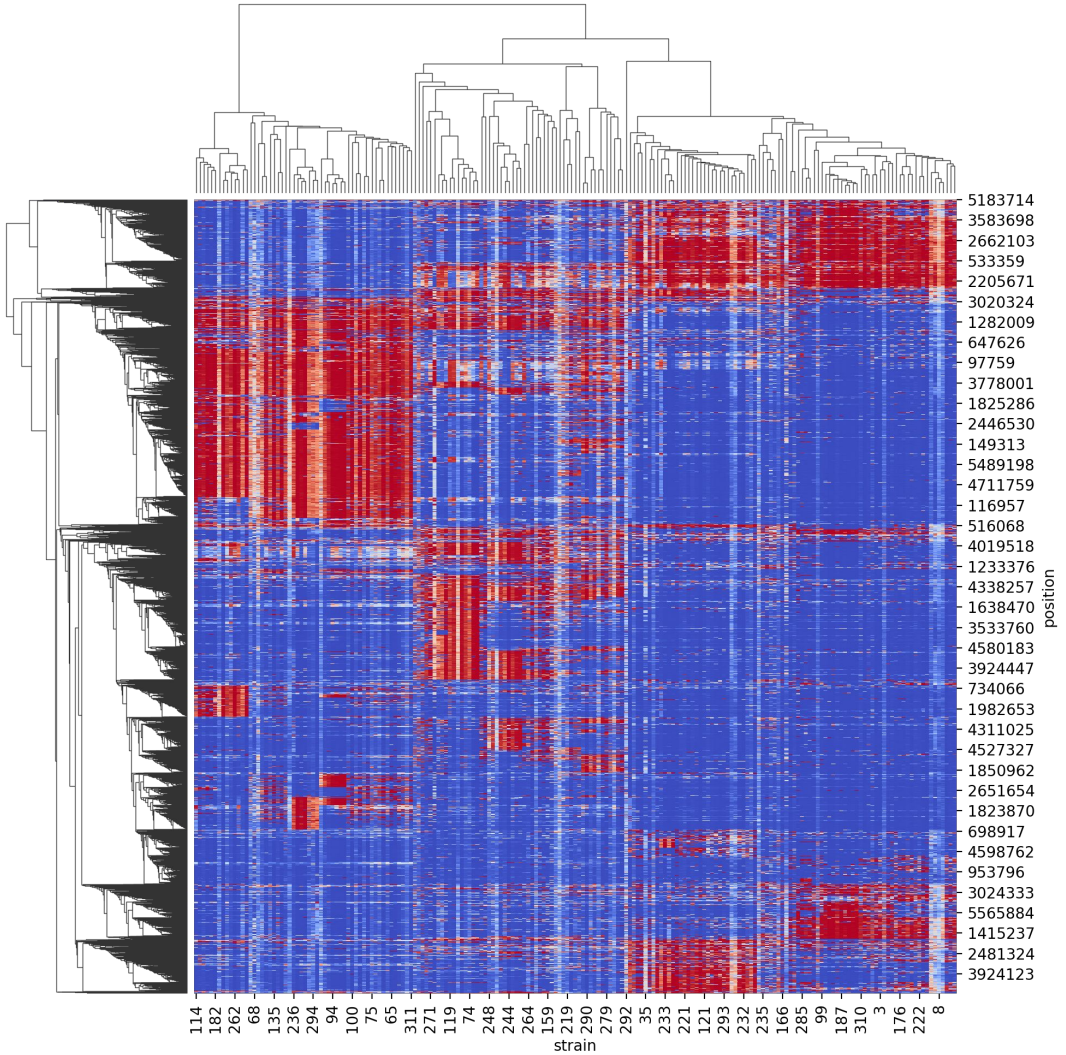
What can we do with this tool?

- Computation (time and memory) scale approximately linearly with dimensions
- Can be greatly accelerated on GPUs, making it possible to run on thousands of positions, samples, strains
- GT-PRO metagenotypes were previously collected for 25,133 human metagenomes

- *What strains exist?* • *How are they related?* • *How do they evolve?* •

We can identify hundreds of distinct *E. coli* genotypes

- Genotypes cluster into separate sub-types, suggesting incomplete recombination and/or shared evolutionary history



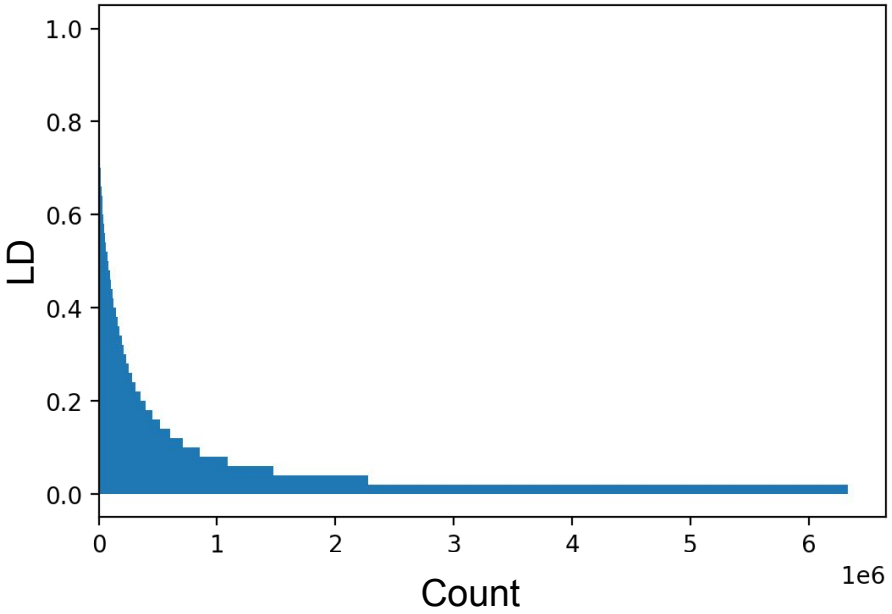
Linkage disequilibrium (LD) among *E. coli* strains

- Two loci are “linked” if the allele at one position is correlated with the allele at the other across strains

		Locus 1	
		A	a
Locus 2	B	0.75	0.25
	b	0.25	0.75

Most alleles are only weakly linked in *E. coli*

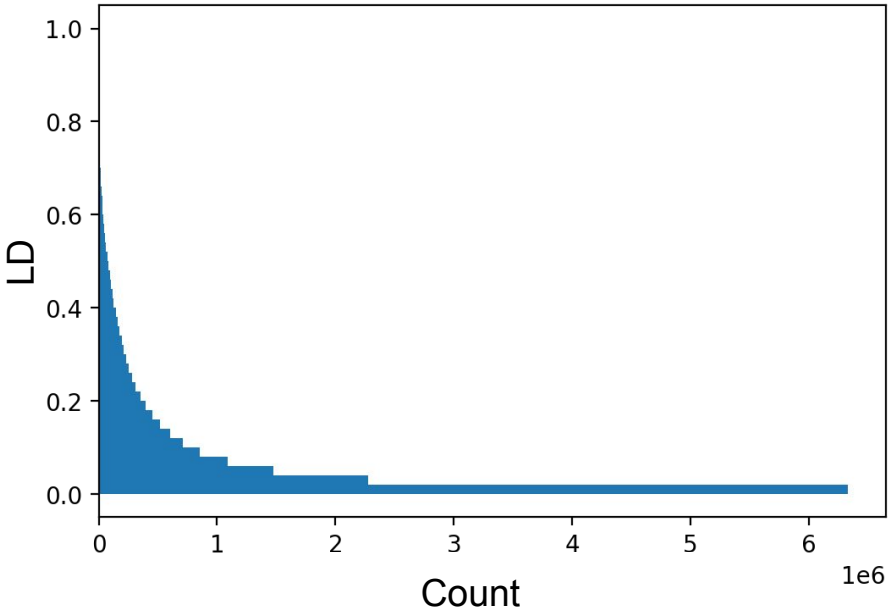
- Two loci are “linked” if the allele at one position is correlated with the allele at the other across strains

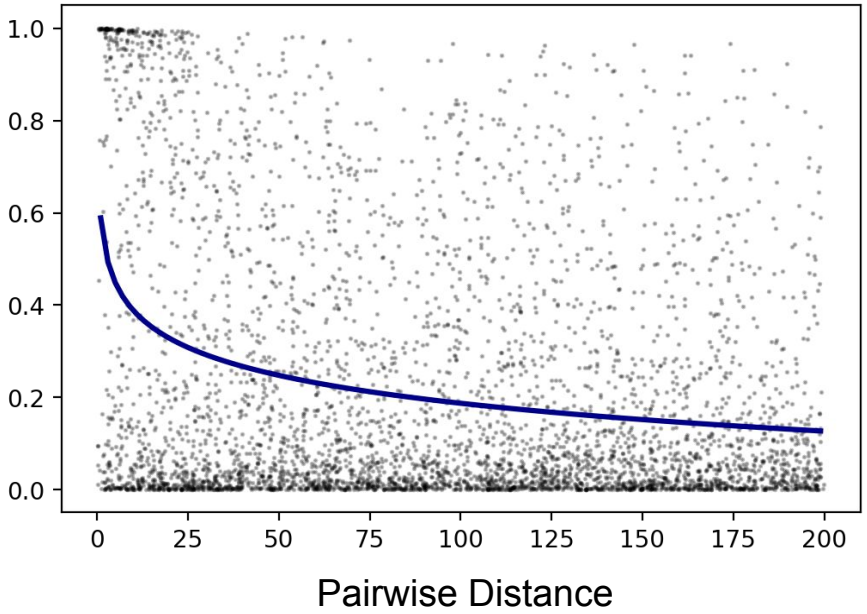


		Locus 1	
		A	a
Locus 2	B	0.75	0.25
	b	0.25	0.75

Much higher linkage between nearby alleles

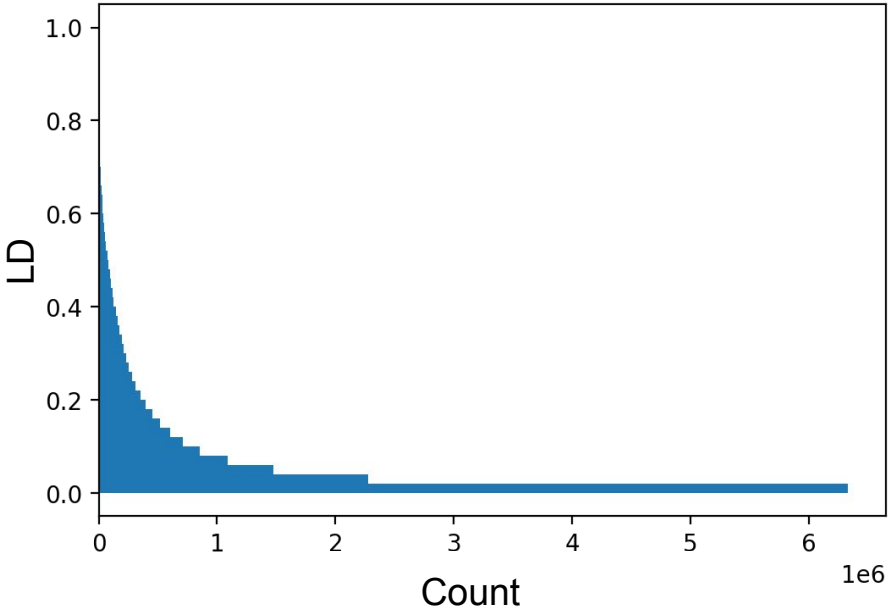
- Two loci are “linked” if the allele at one position is correlated with the allele at the other across strains

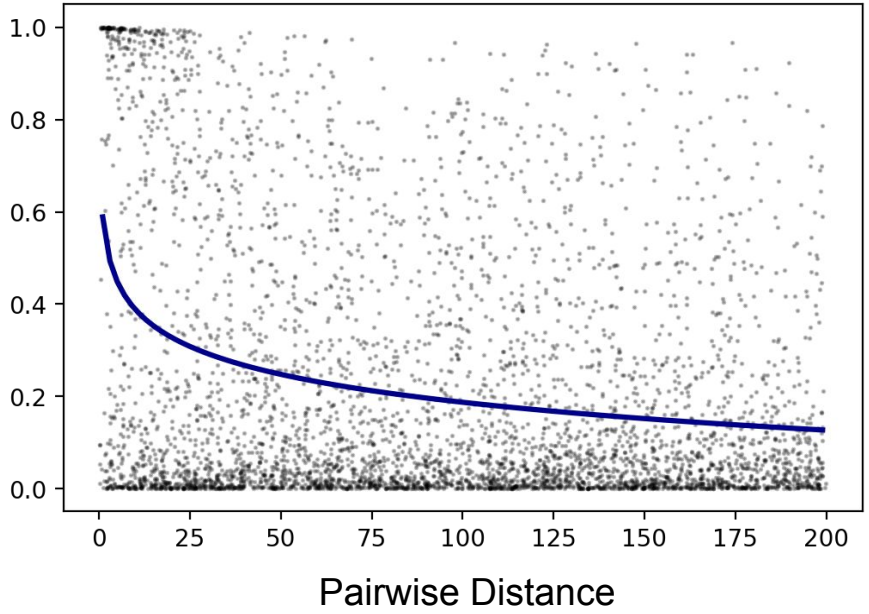




E. coli LD patterns consistent with high recombination

- Theoretically: with no recombination, linkage is nearly 100% across all pairs
- As recombination increases, linkage decreases overall, but less for proximate pairs





Conclusion

- Strain deconvolution from metagenotypes greatly improves surveys of within-species diversity
- Enables precise strain-tracking in e.g. FMT studies
- Scales well to very large datasets
- Allows us to explore bacterial population structure and evolution