

Scaling microbial strain inference to thousands of metagenomes using fuzzy genotypes

Multiscale Microbial Communities
2022-02-21
Byron J. Smith

More information

bioRxiv

Scalable microbial strain inference in metagenomic data using StrainFacts. *bioRxiv* (2022)
doi: 10.1101/2022.02.01.478746



<https://github.com/bsmith89/StrainFacts>



@ByronJSmith

Acknowledgments

My Co-authors:

- Xiangpeng Li
- Jason Shi
- Adam Abate
- Katie Pollard

Pollard Lab

Gladstone Institute for Data
Science and Biotechnology

Chan Zuckerberg Biohub

NIH T32 DK007007

UCSF Initiative for Digital
Transformation in Computational
Biology & Health

Outline

Intraspecific diversity in the microbiome

Strain inference

Metagenotype deconvolution

Application to large metagenome collections

Outline

Intraspecific diversity in the microbiome

Strain inference

Metagenotype deconvolution

Application to large metagenome collections



Human associated microbes are diverse and important

Human associated microbes are diverse and important

Important:

- Digestion
- Pathogen resistance
- Immune modulation

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

Diverse:

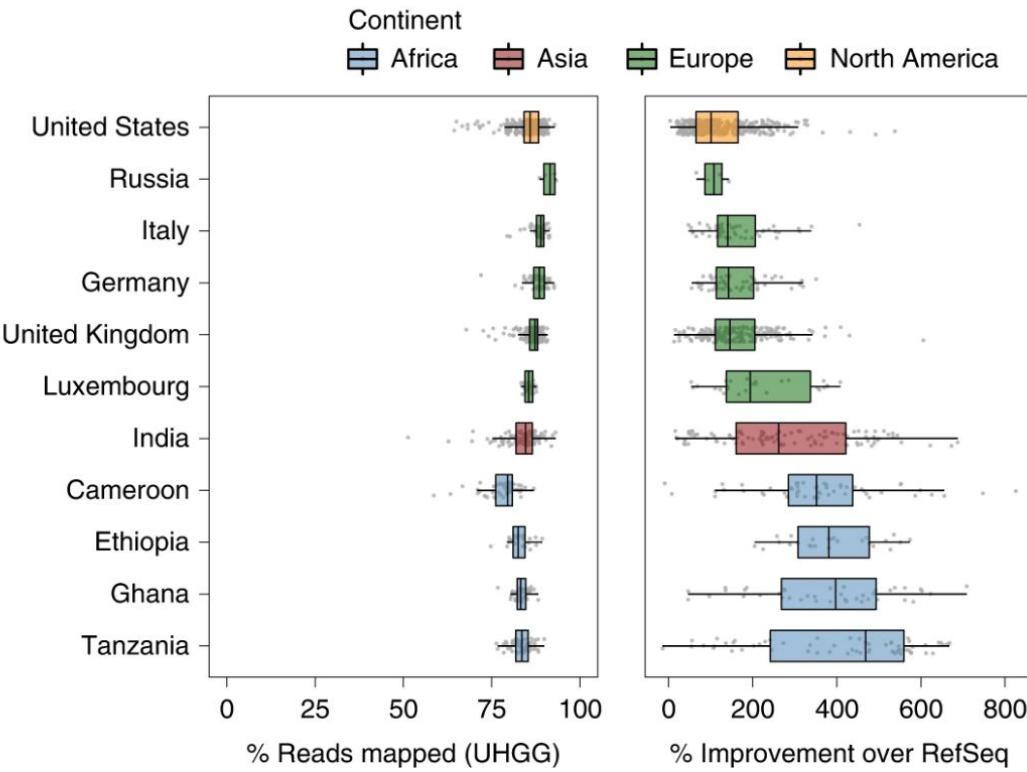
- Hundreds of bacterial species
- Also archaea, eukaryotes, and viruses
- High inter-individual variation
- **Huge (but under-explored) diversity *within* species**

Reference databases are approaching a complete catalog of species in the human gut

Key efforts include the Unified Human Gastrointestinal Genome (UHGG)

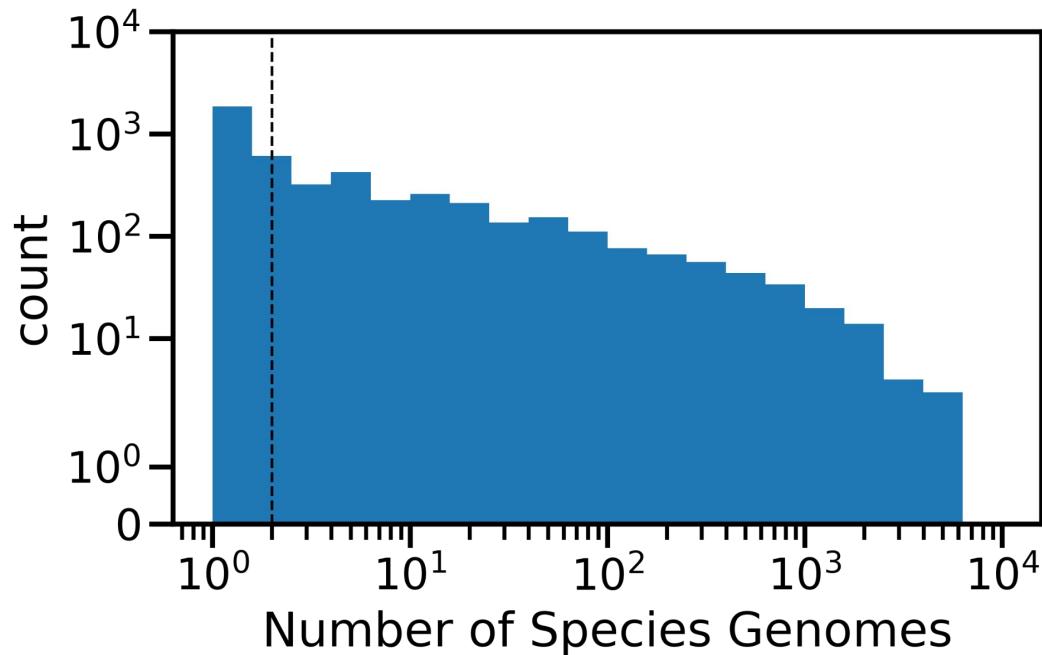
- Includes metagenome assembled genomes (MAGs)
- 204,938 genomes in 4,644 species

Remaining disparity from understudied human populations



Strain diversity is not well documented for vast majority

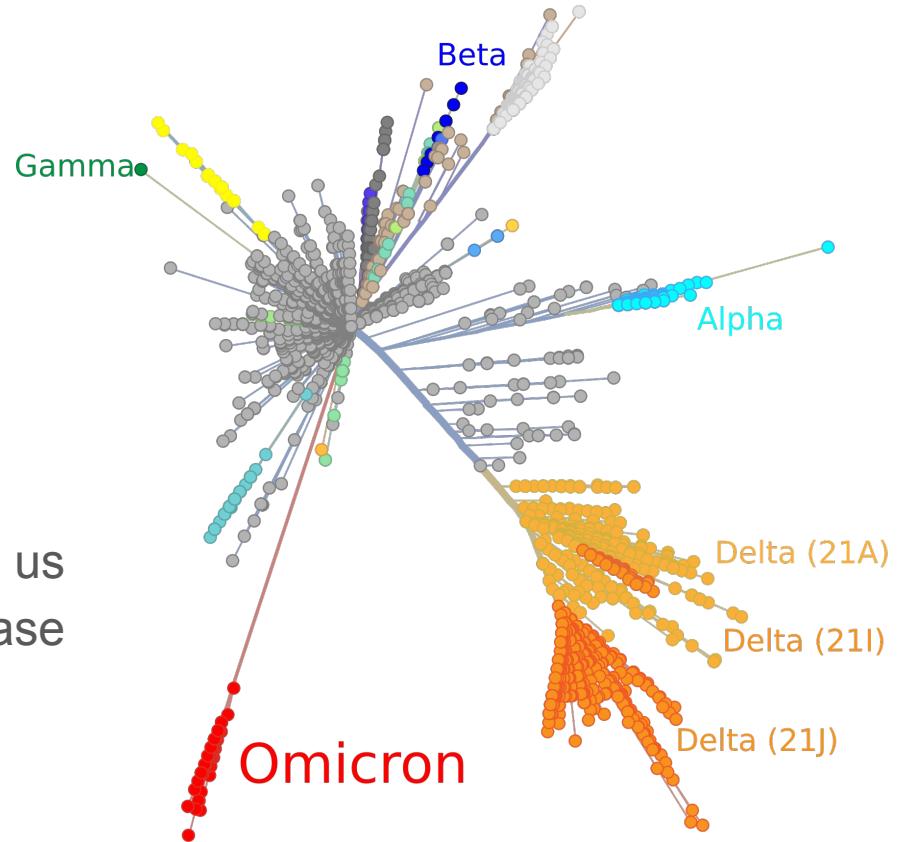
75% of species have fewer than 10 representative genomes



Strain diversity is both biologically important and scientifically informative

Differences between microbial strains can impact human health

Tracking strains between individuals, over time, or across global geography can help us to understand transmission patterns, disease associations, selection pressures, etc.



Strains of a species can have important differences

Intraspecific diversity has been appreciated for a long time

- Pathogenicity
- Antibiotic resistance
- Phage resistance
- Auxotrophy

Strains of a species can have important differences

Intraspecific diversity has been appreciated for a long time

- Pathogenicity
- Antibiotic resistance
- Phage resistance
- Auxotrophy



e.g. *E. coli*

Well studied, easy to culture

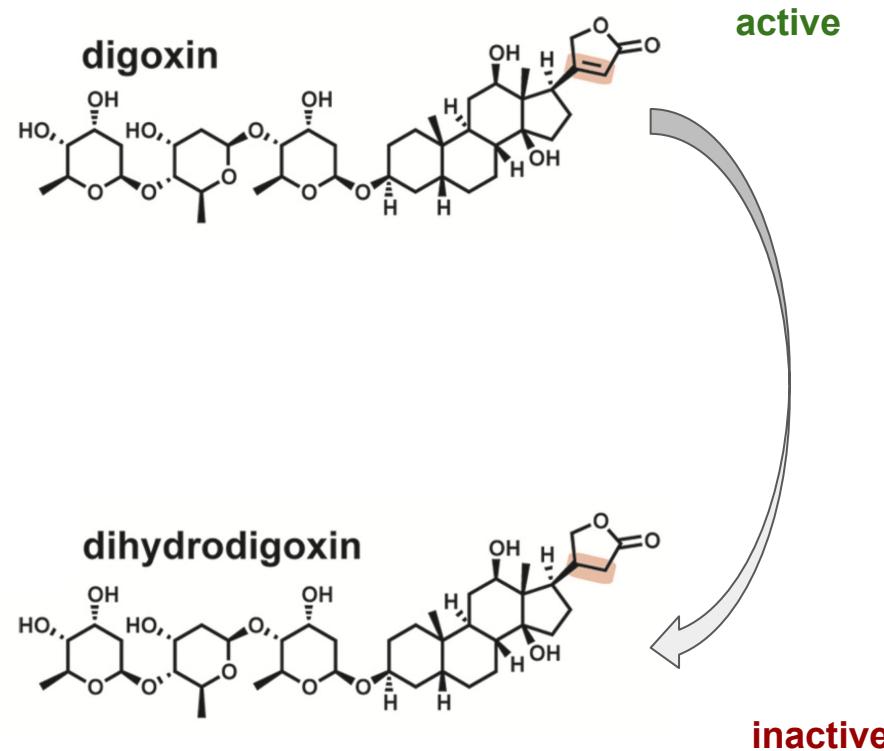
Strains of a species can have important differences

Intraspecific diversity has been appreciated for a long time

- Pathogenicity
- Antibiotic resistance
- Phage resistance
- Auxotrophy

e.g. *E. coli*

Well studied, easy to culture



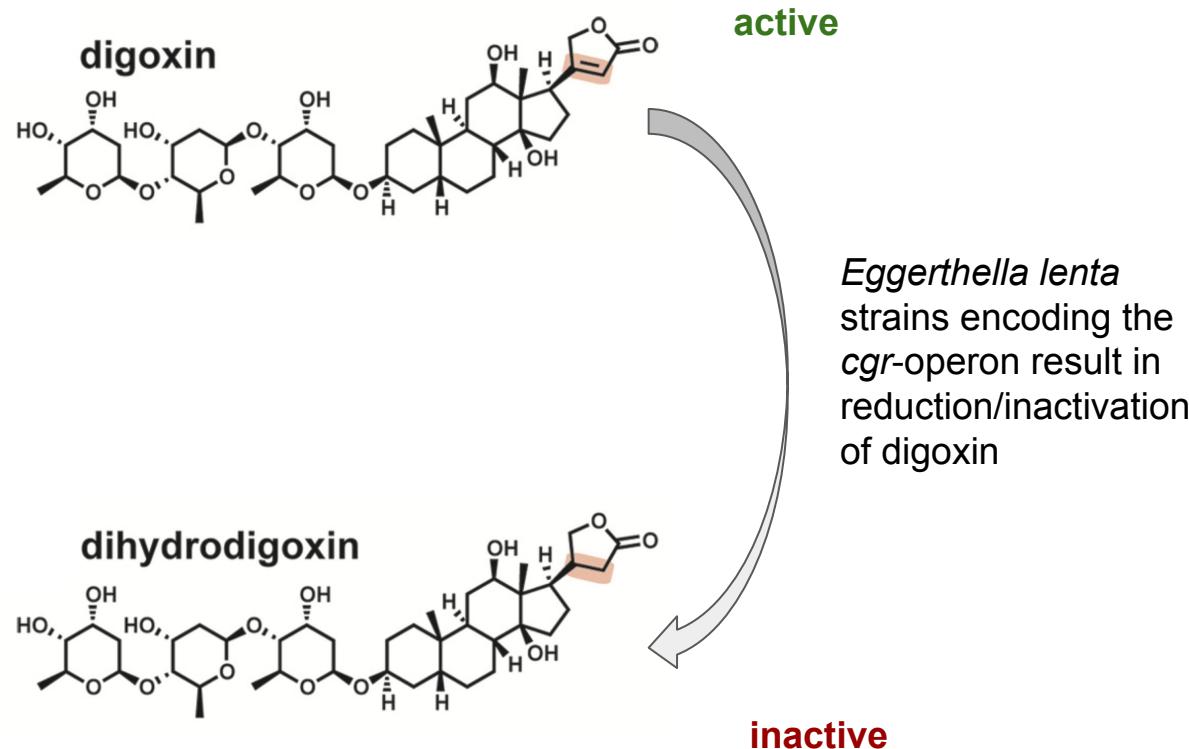
Strains of a species can have important differences

Intraspecific diversity has been appreciated for a long time

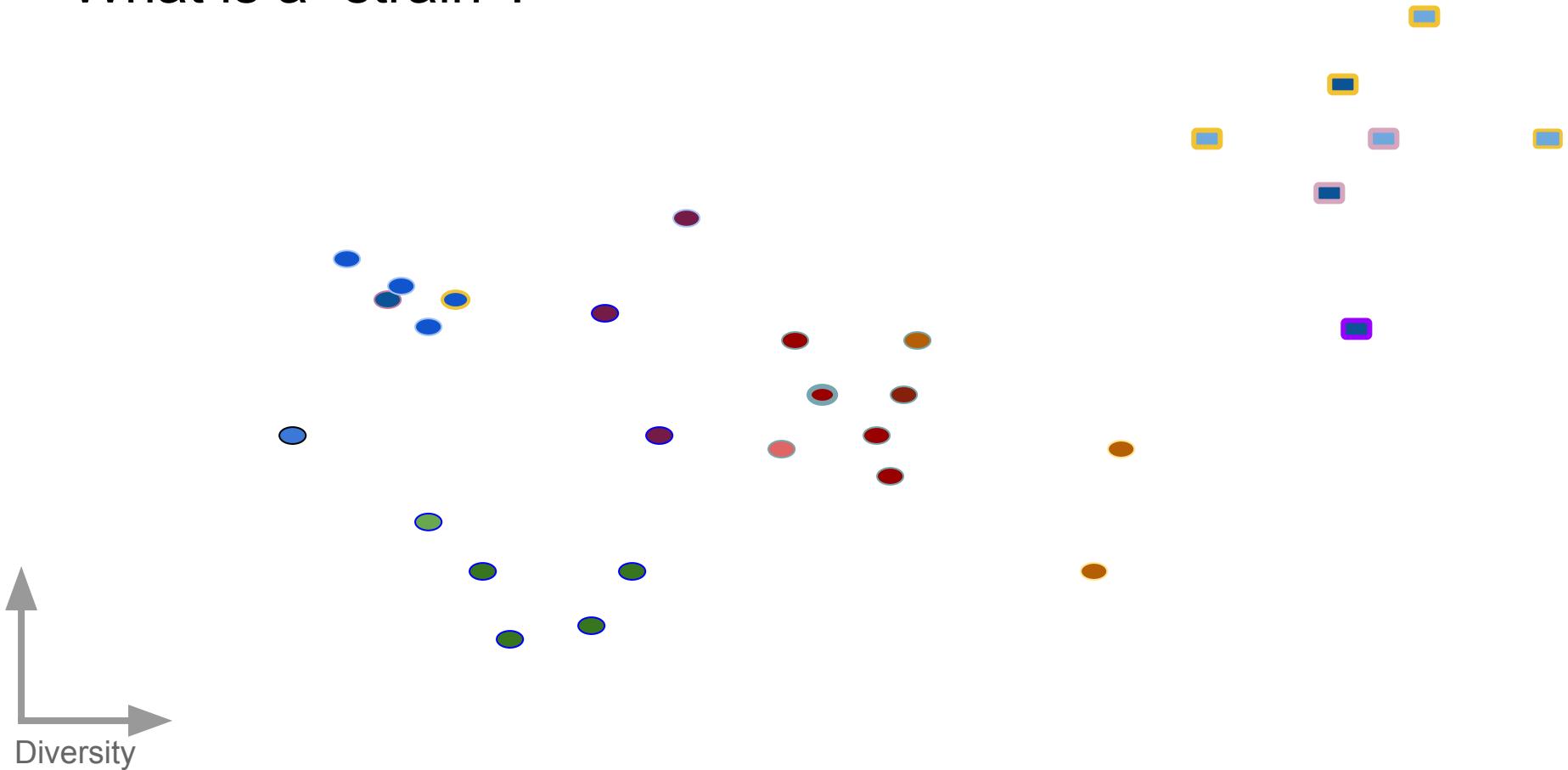
- Pathogenicity
- Antibiotic resistance
- Phage resistance
- Auxotrophy

e.g. *E. coli*

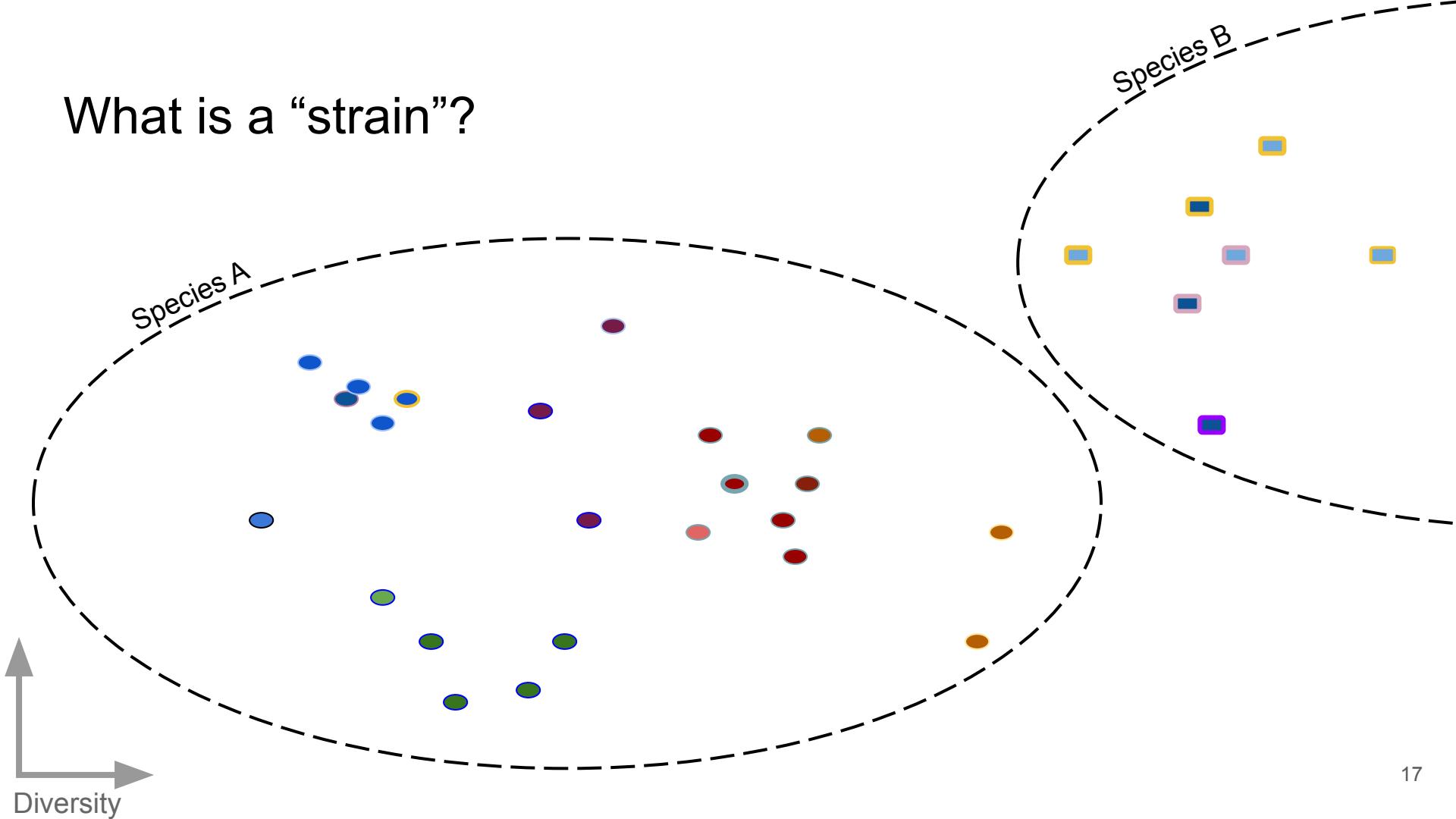
Well studied, easy to culture



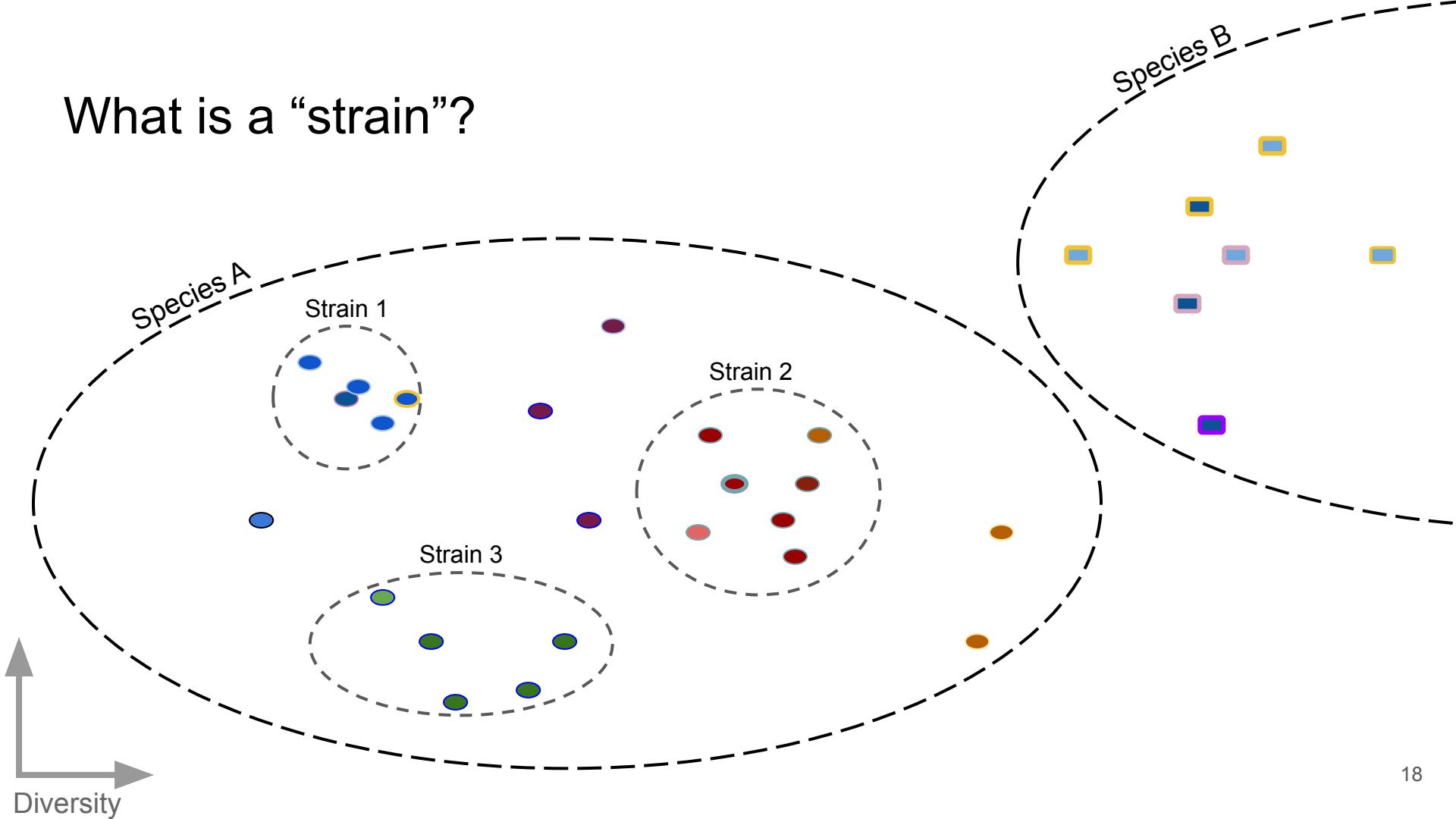
What is a “strain”?



What is a “strain”?



What is a “strain”?



What differentiates strains?

Gene content



Genotype

A A C

T A G

T A C

T G G

What differentiates strains?

Gene content



Genotype

A A C

T A G

T A C

T G G



Outline

Intraspecific diversity in the microbiome

Strain inference

Metagenotype deconvolution

Application to large metagenome collections

Existing methods lack taxonomic resolution

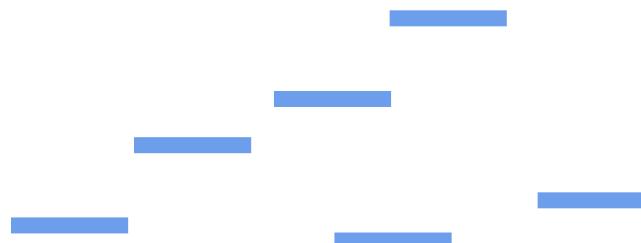
Marker genes (e.g. 16S) are too conserved

Shotgun metagenomic data is increasingly available

Marker genes (e.g. 16S) are too conserved



Standard methods for taxonomic surveys assign reads to species based on read mapping



Metagenomic reads encode strain-level information

Marker genes (e.g. 16S) are too conserved



Standard methods for taxonomic surveys assign reads to species based on read mapping



But shotgun reads also cover single-nucleotide variants that encode much finer taxonomic detail



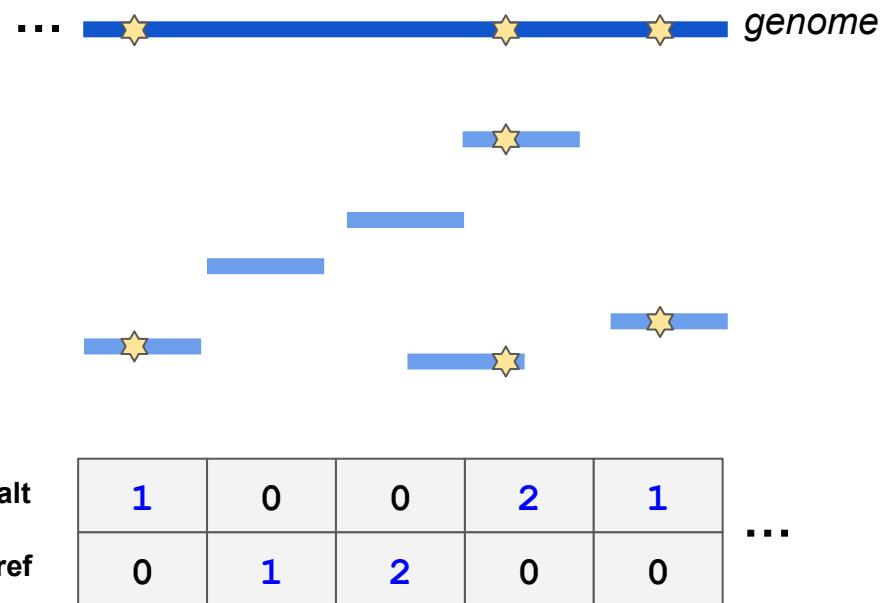
“Metagenotyping”

Marker genes (e.g. 16S) are too conserved

Standard methods for taxonomic surveys assign reads to species based on read mapping

But shotgun reads also cover single-nucleotide variants that encode much finer taxonomic detail

Metagenotypers tally variants at polymorphic positions (SNPs)



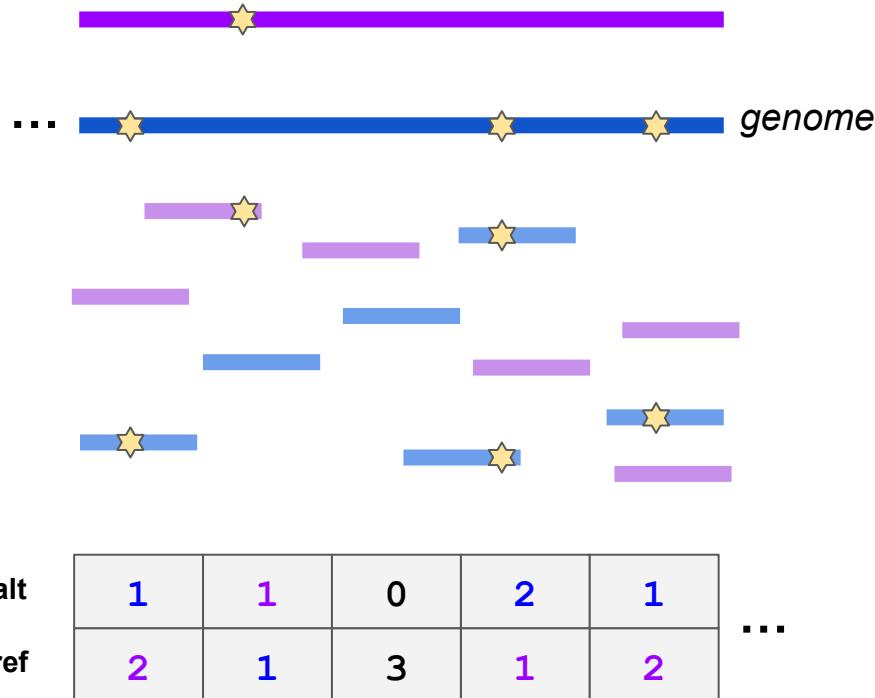
“Metagenotyping”

Marker genes (e.g. 16S) are too conserved

Standard methods for taxonomic surveys assign reads to species based on read mapping

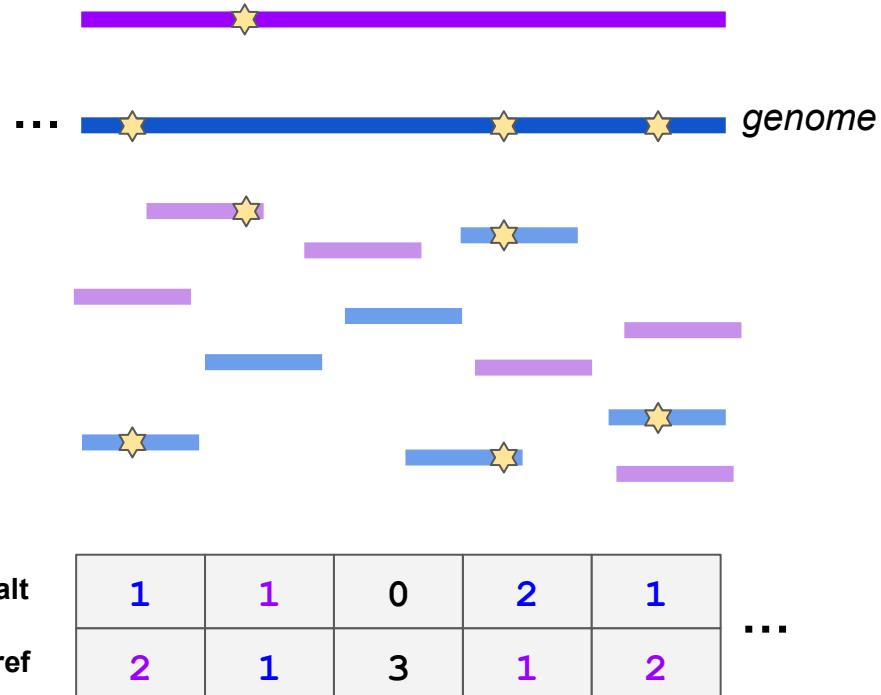
But shotgun reads also cover single-nucleotide variants that encode much finer taxonomic detail

Metagenotypers tally variants at polymorphic positions (SNPs)



GT-Pro scales to tens-of-thousands of samples

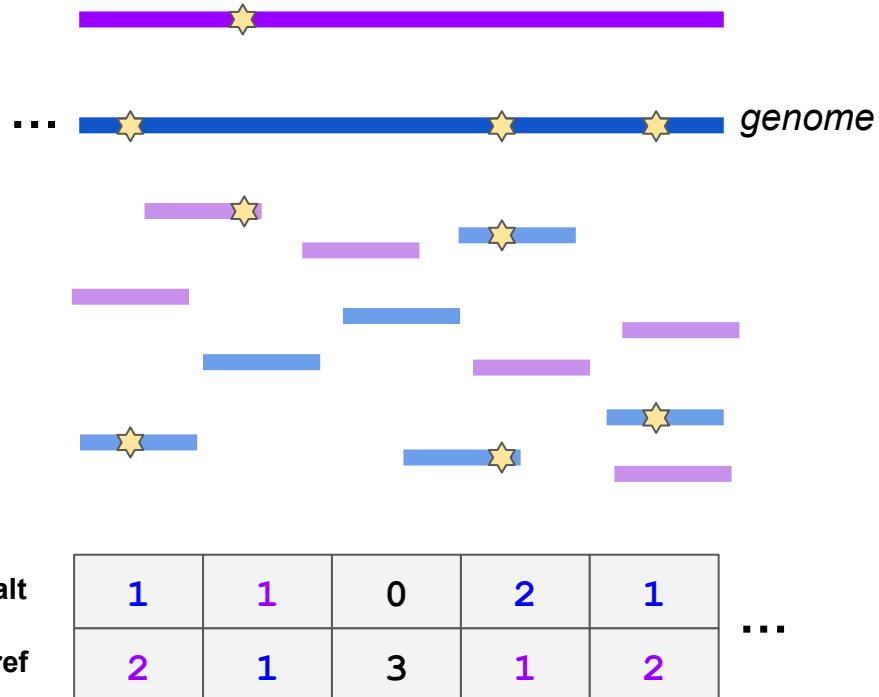
Uses exact k-mer matching to accelerate
metagenotyping



GT-Pro scales to tens-of-thousands of samples

Uses exact k-mer matching to accelerate metagenotyping

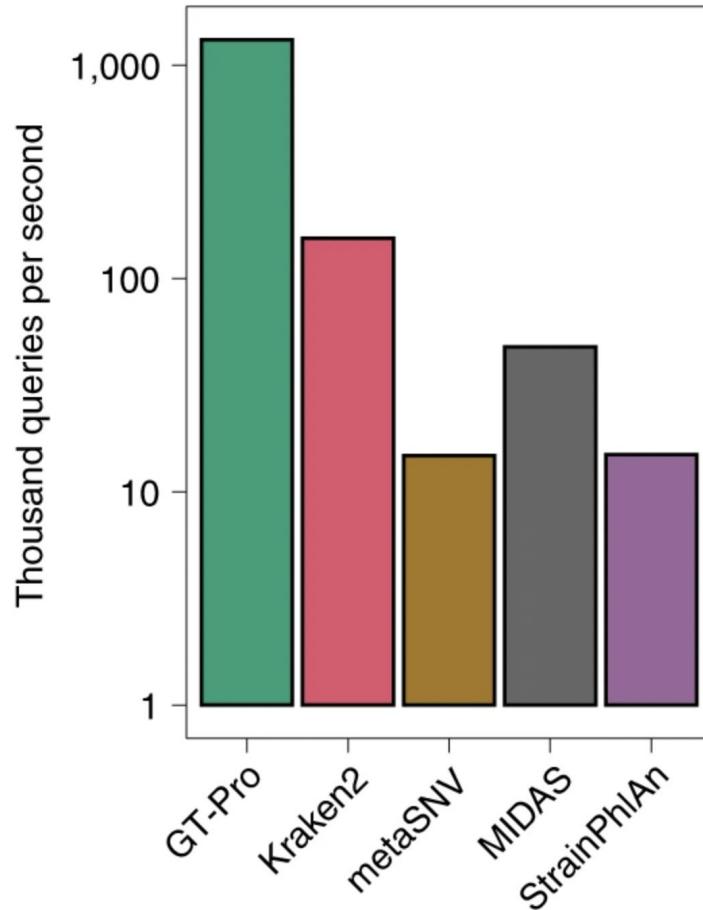
Tallies variants at known, bi-allelic SNPs in the core genome using a database of known SNVs



GT-Pro scales to tens-of-thousands of samples

Uses exact k-mer matching to accelerate metagenotyping

Tallies variants at known, bi-allelic SNPs in the core genome using a database of known SNVs



Low
sequencing
coverage

Mixtures
of strains

Interpreting metagenotypes is challenging

Closely
related
strains

Novel
strains

Low
sequencing
coverage

Consensus Genotypes

Mixtures
of strains

Interpreting metagenotypes is challenging

Closely
related
strains

Novel
strains

Low
sequencing
coverage

Interpreting metagenotypes is challenging

Closely
related
strains

**Mixtures
of strains**

**Reference
Database**

**Novel
strains**

Low
sequencing
coverage

Dissimilarity
Methods /
Ordination

Closely
related
strains

Mixtures
of strains

Novel
strains

Interpreting metagenotypes is challenging

Outline

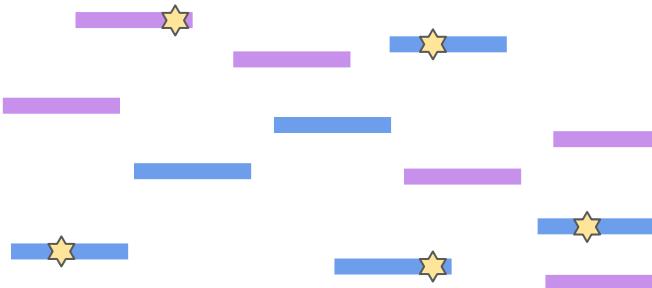
Intraspecific diversity in the microbiome

Strain inference

Metagenotype deconvolution

Application to large metagenome collections

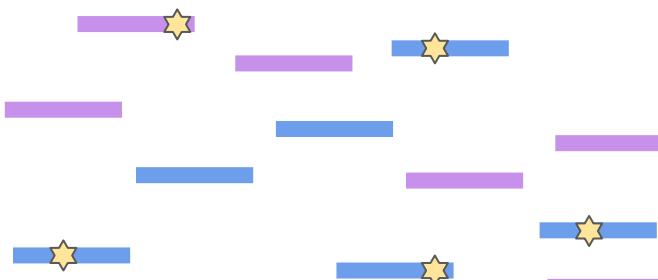
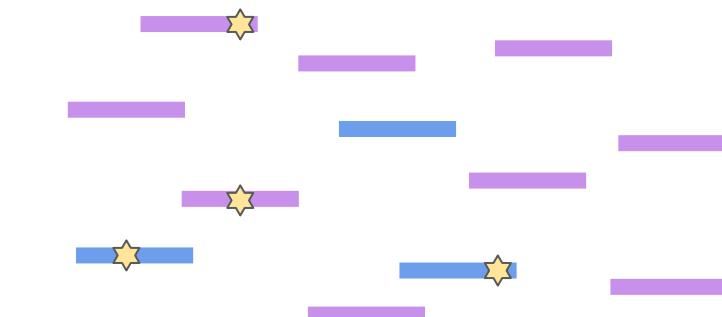
Strain deconvolution



1	1	0	2	1
2	1	3	1	2

...

Strain deconvolution harnesses SNV covariance

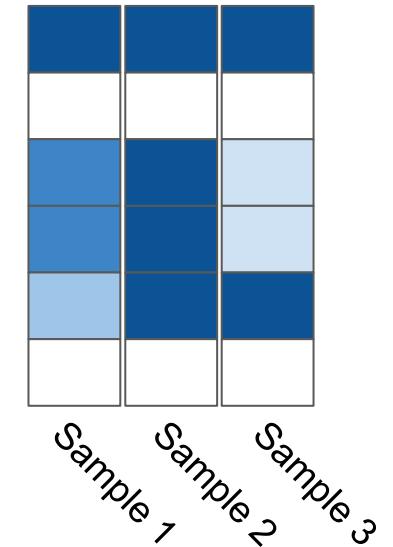


alt	1	2	0	1	0
ref	2	0	4	2	2

1	1	0	2	1
2	1	3	1	2

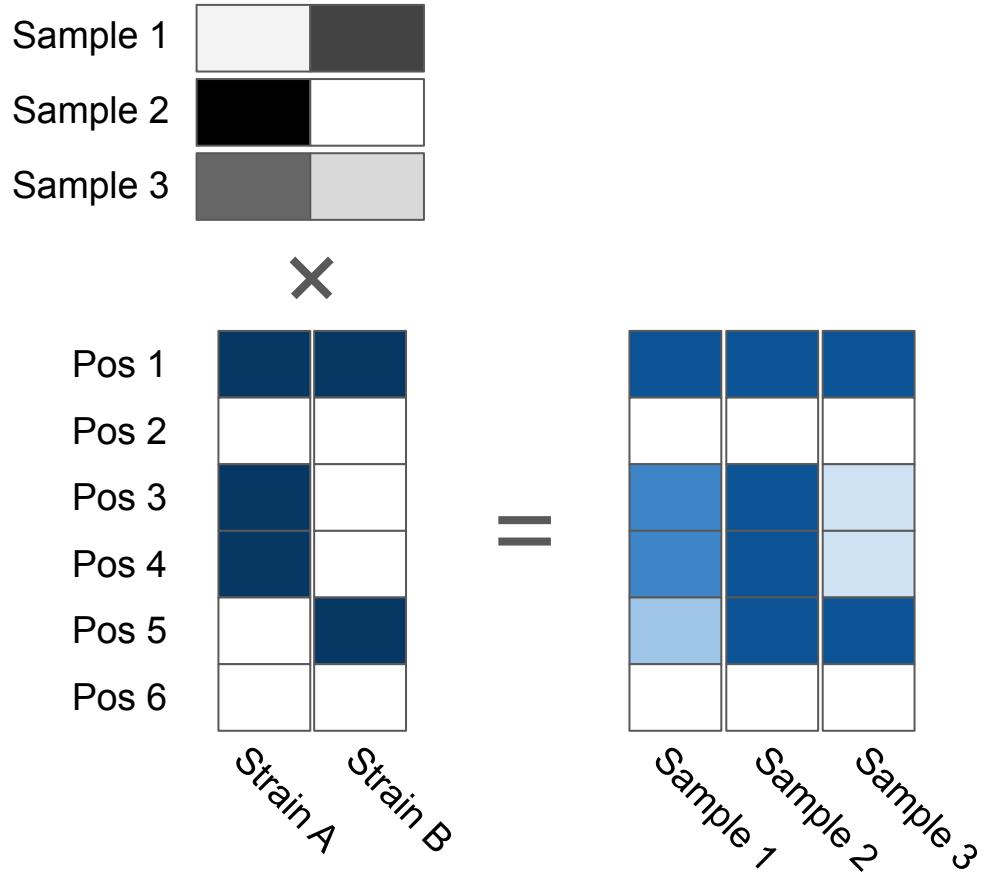
...

Non-negative matrix factorization for metagenotypes



Non-negative matrix factorization for metagenotypes

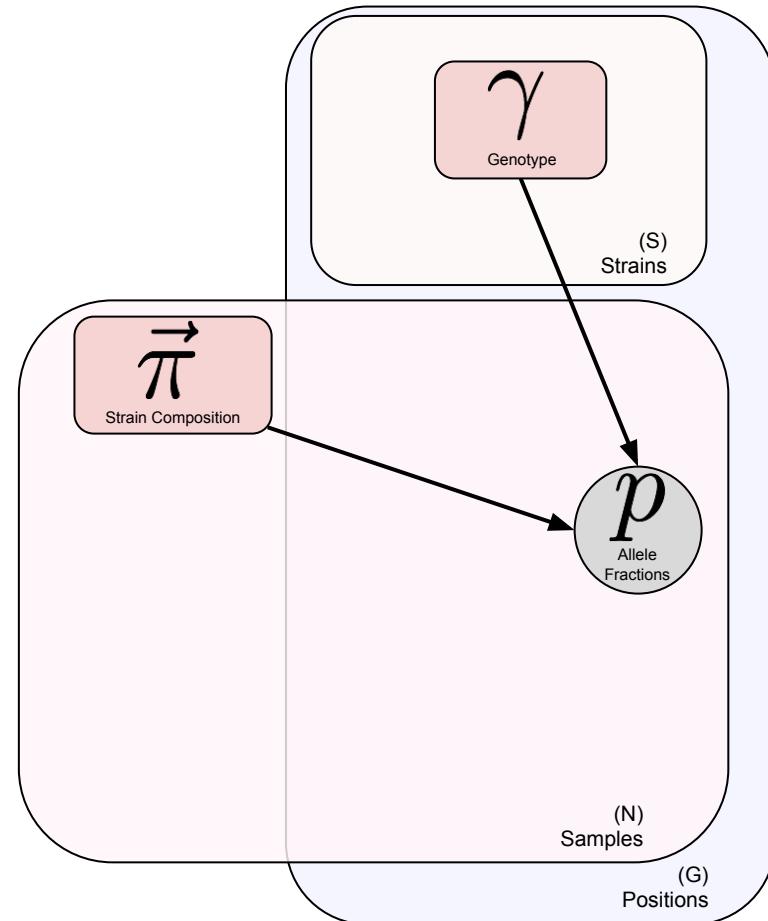
Like NMF: model allele fractions as linear combinations of strain genotypes



Strain deconvolution as model-based inference

Like NMF: model allele fractions as linear combinations of strain genotypes

Except:

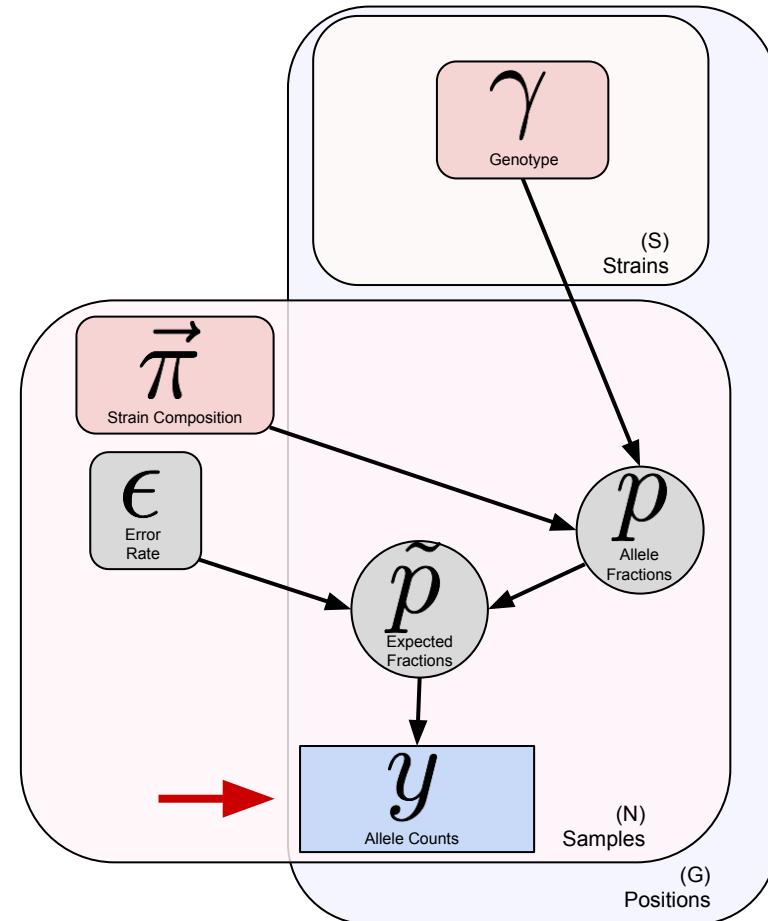


Strain deconvolution as model-based inference

Like NMF: model allele fractions as linear combinations of strain genotypes

Except:

- Counts observed (sequencing error, binomial likelihood)

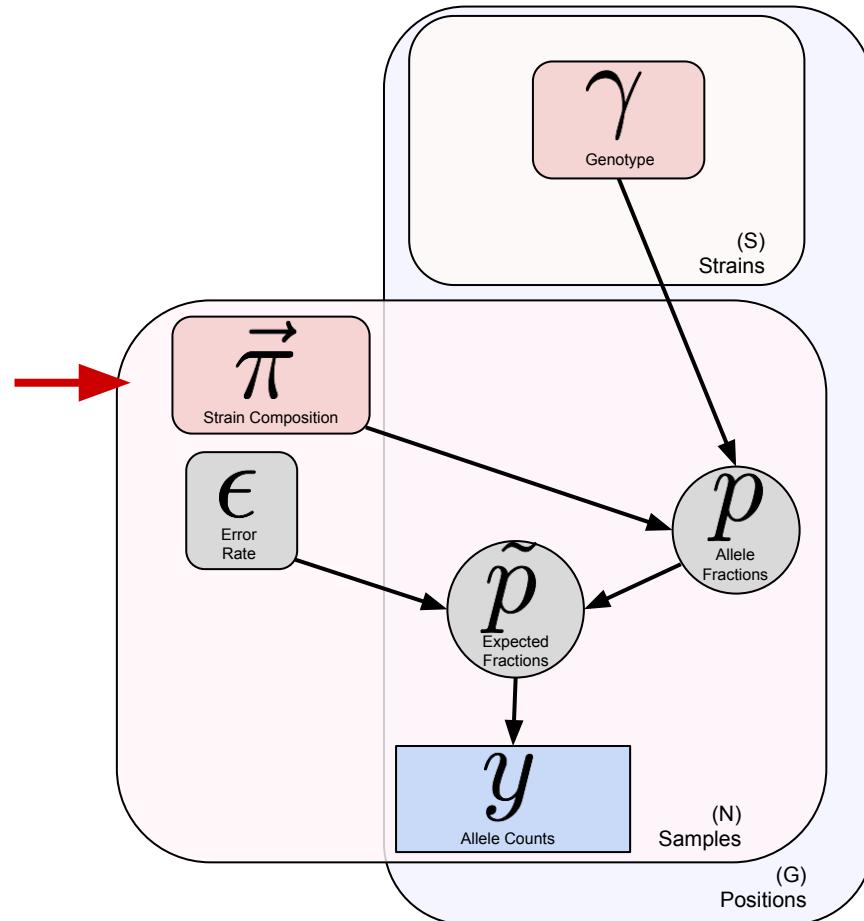


Strain deconvolution as model-based inference

Like NMF: model allele fractions as linear combinations of strain genotypes

Except:

- Counts observed (sequencing error, binomial likelihood)
- Strain composition sums-to-1

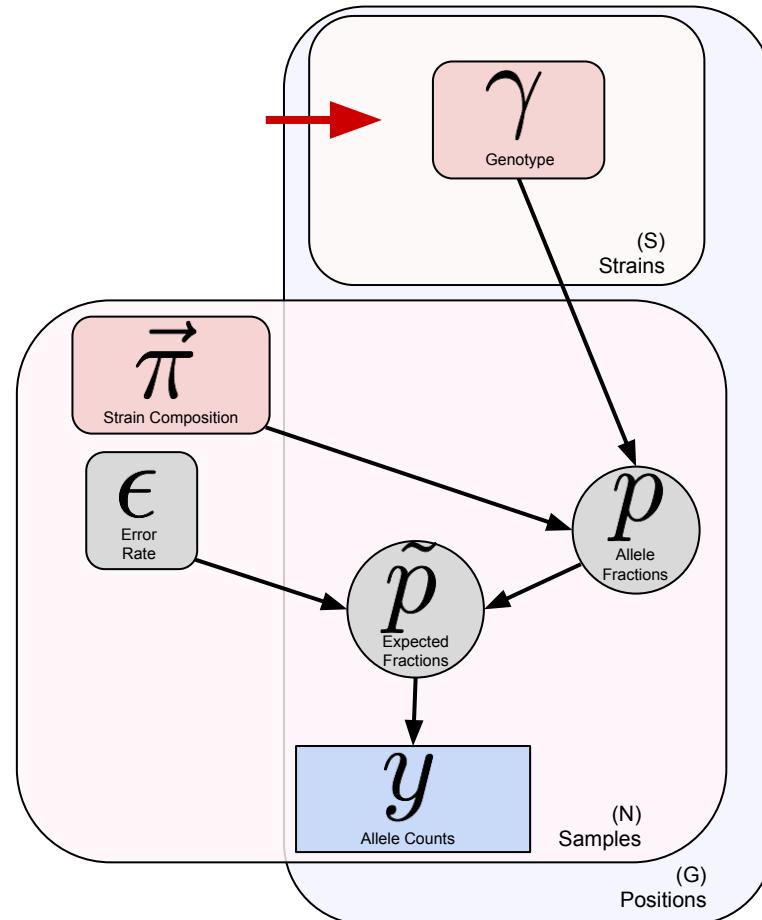


Strain deconvolution as model-based inference

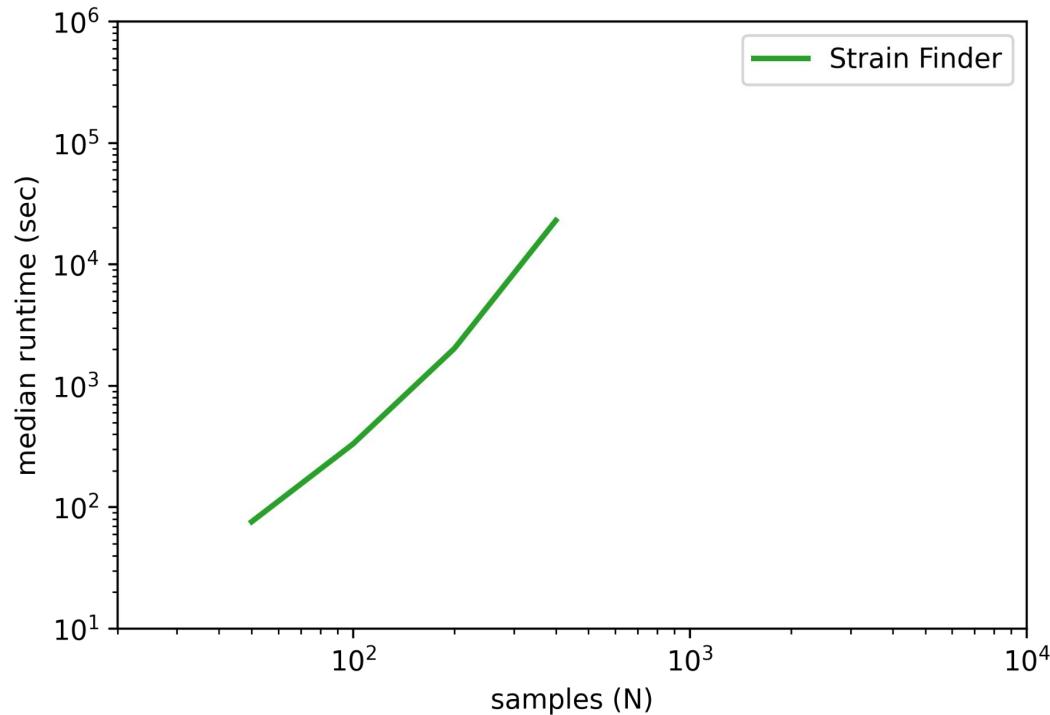
Like NMF: model allele fractions as linear combinations of strain genotypes

Except:

- Counts observed (sequencing error, binomial likelihood)
- Strain composition sums-to-1
- Genotypes at each position are binary, either 0 (reference) or 1 (alternative) allele



Computational scalability of existing deconvolution tools

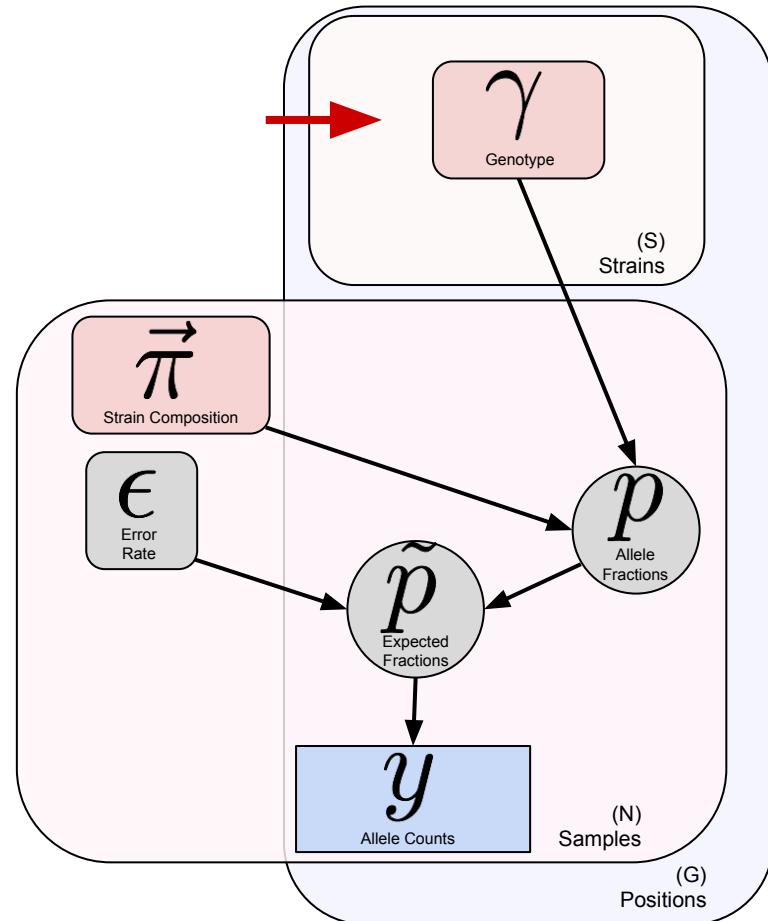


Discrete genotypes are computationally challenging

Like NMF: model allele fractions as linear combinations of strain genotypes

Except:

- Counts observed (sequencing error, binomial likelihood)
- Strain composition sums-to-1
- Genotypes at each position are binary, either 0 (reference) or 1 (alternative) allele

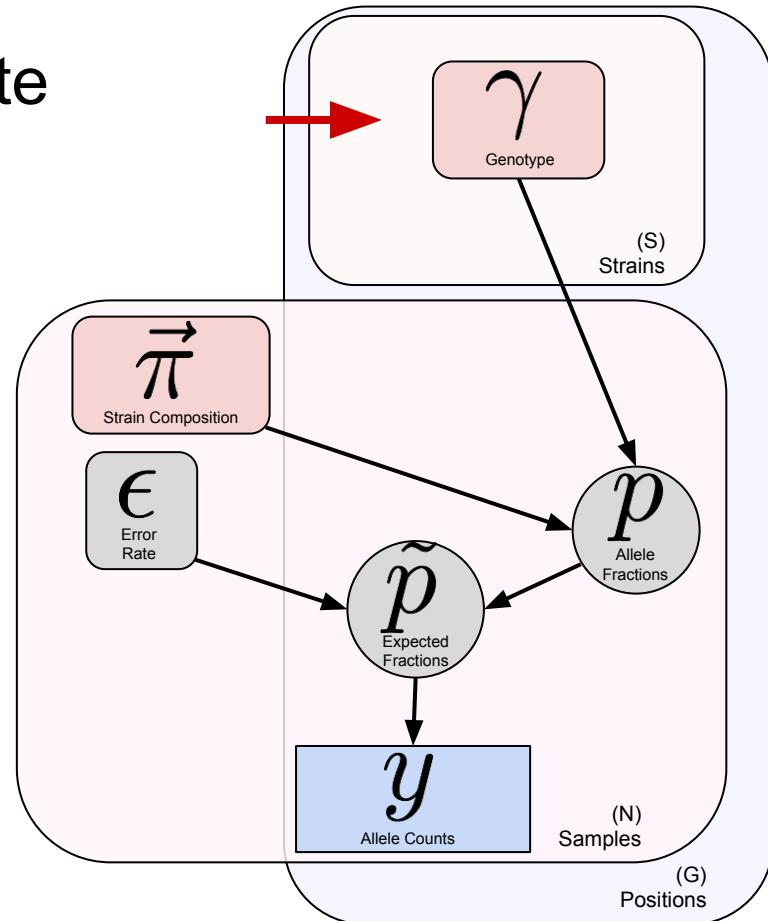


StrainFacts relaxes the discrete allele constraint

Like NMF: model allele fractions as linear combinations of strain genotypes

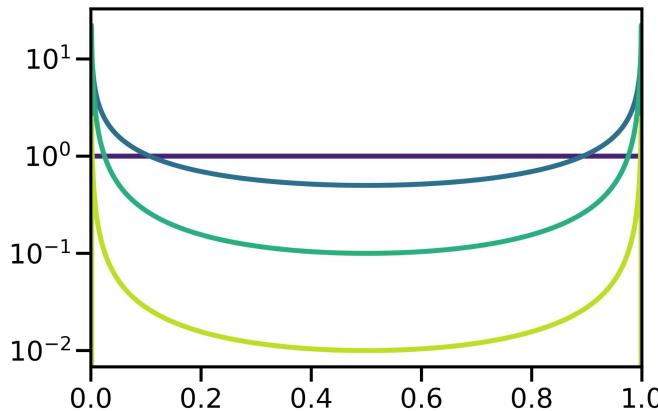
Except:

- Counts observed (sequencing error, binomial likelihood)
- Strain composition sums-to-1
- ~~Genotypes at each position are binary, either 0 (reference) or 1 (alternative) allele~~
- Genotypes at each position are ***between*** 0 (reference) and 1 (alternative) alleles

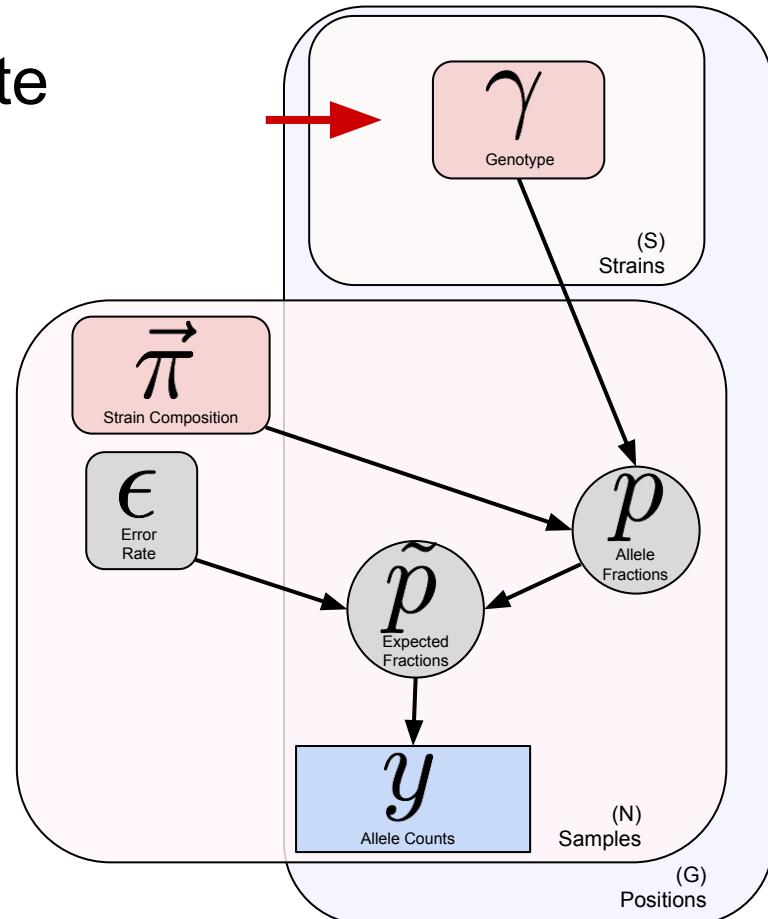


StrainFacts relaxes the discrete allele constraint

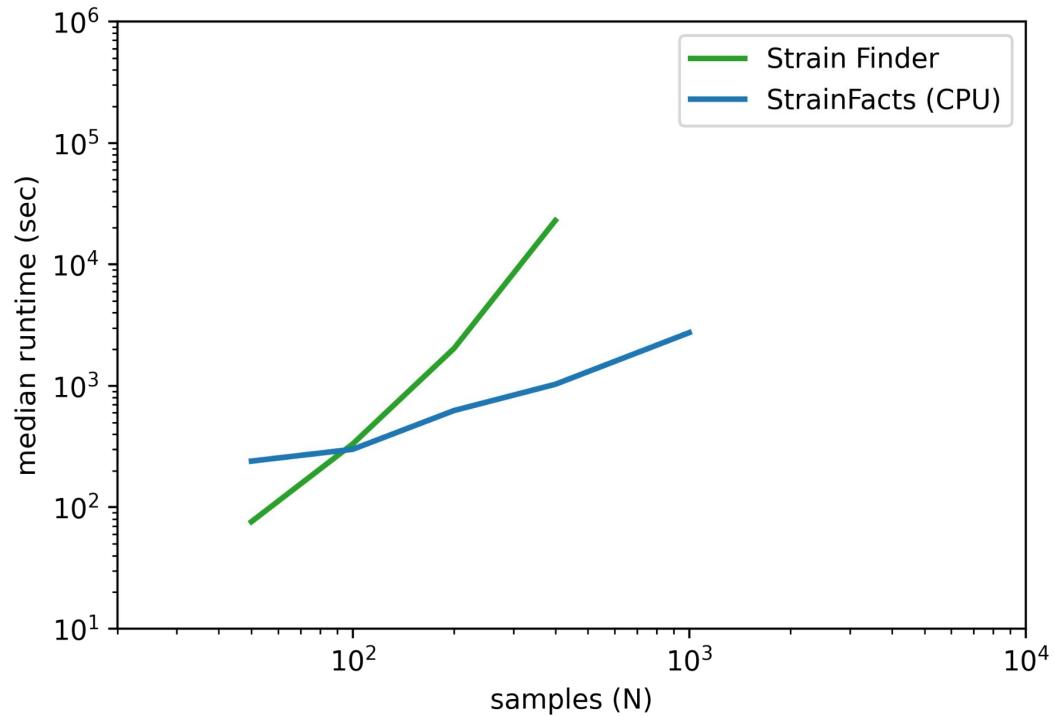
By putting a strong prior on γ , we encourage fuzzy genotypes to be closer to 0 or 1



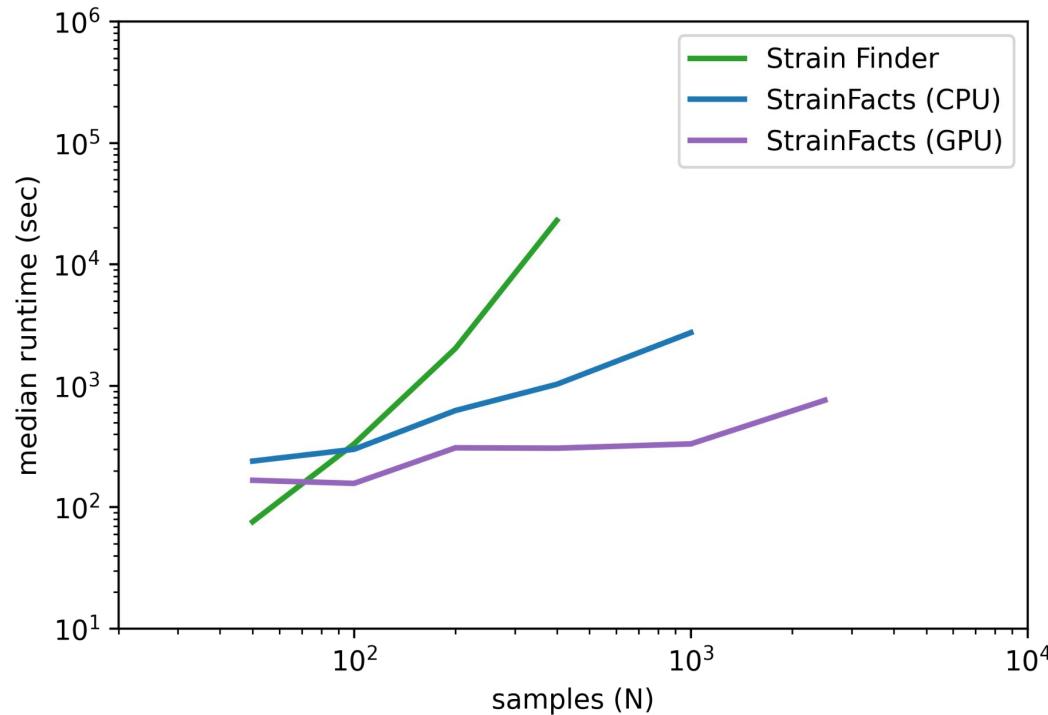
Shifted-scaled Dirichlet distribution
(SSD; similar to the Dirichlet/Beta)



Fuzzy genotypes and gradient descent can scale to thousands of samples



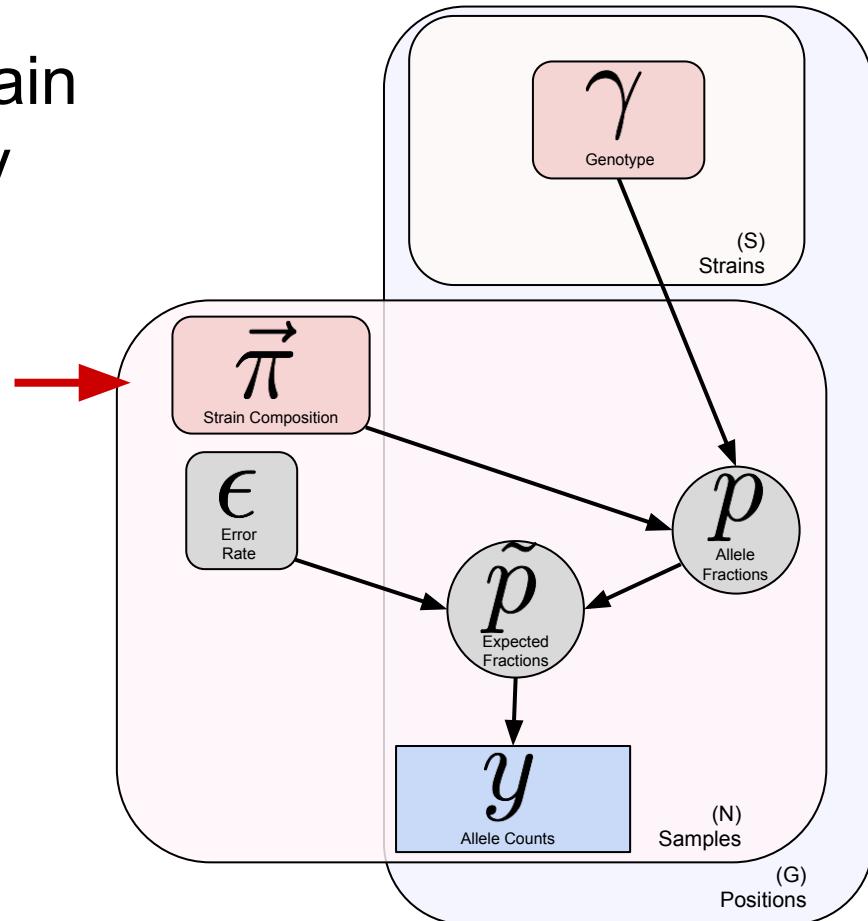
Fuzzy genotypes and gradient descent can scale to thousands of samples



StrainFacts regularizes strain heterogeneity and diversity

Additionally, we put a hierarchical prior on π

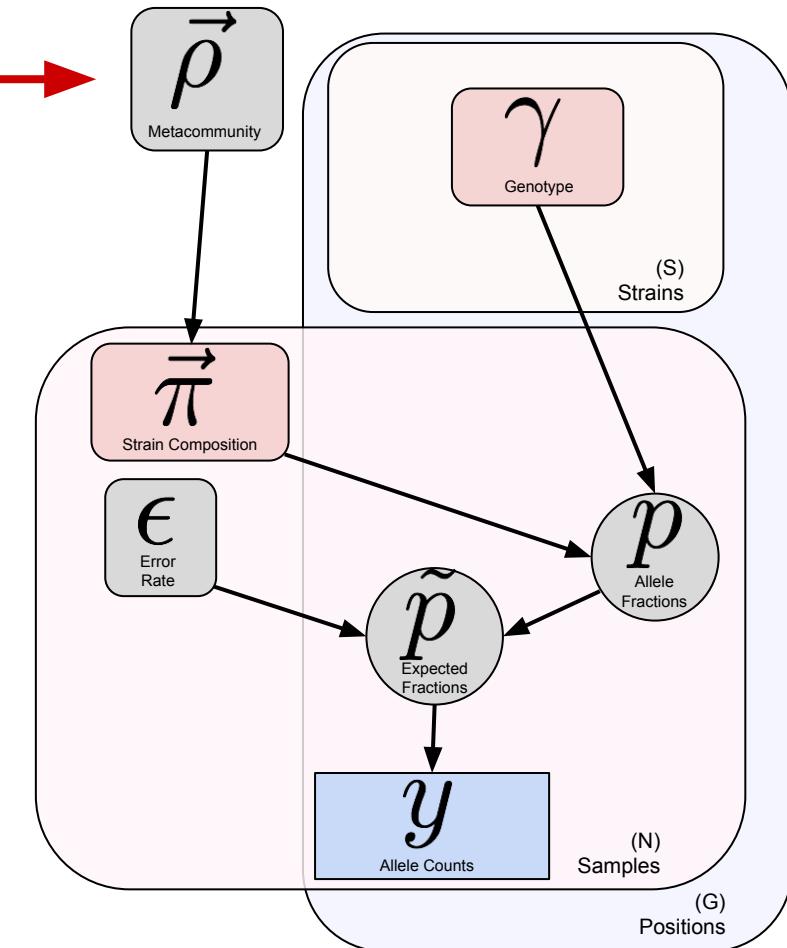
- Strain heterogeneity regularization
(strains per sample)



StrainFacts regularizes strain heterogeneity and diversity

Additionally, we put a hierarchical prior on π

- Strain heterogeneity regularization (strains per sample)
- Overall strain diversity regularization



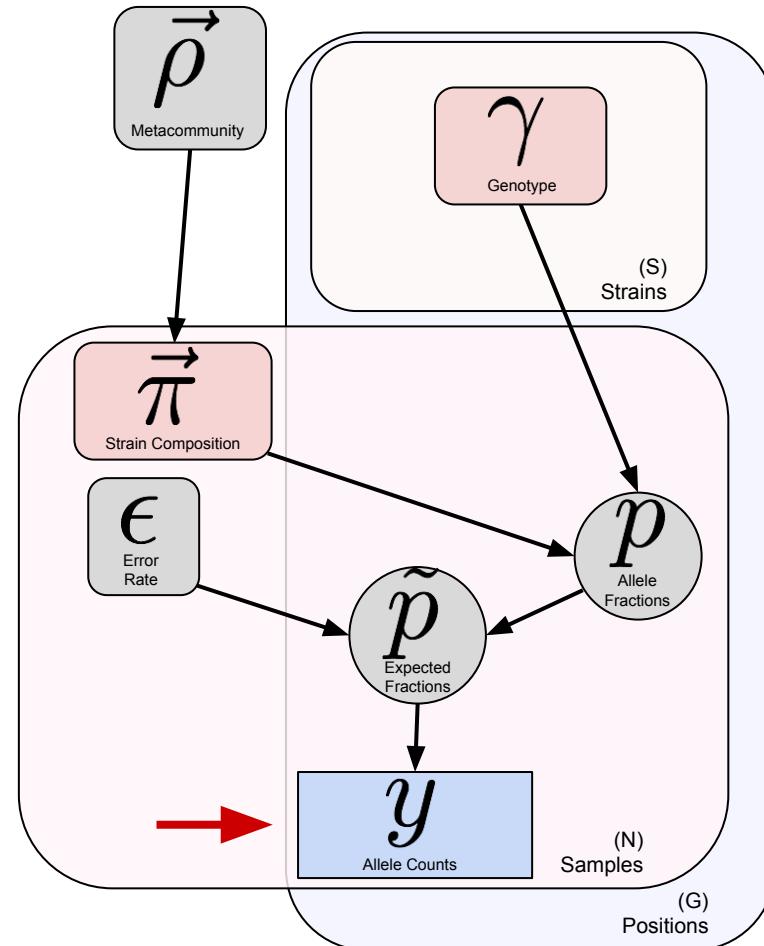
Strain deconvolution as model-based inference

Additionally, we put a hierarchical prior on π

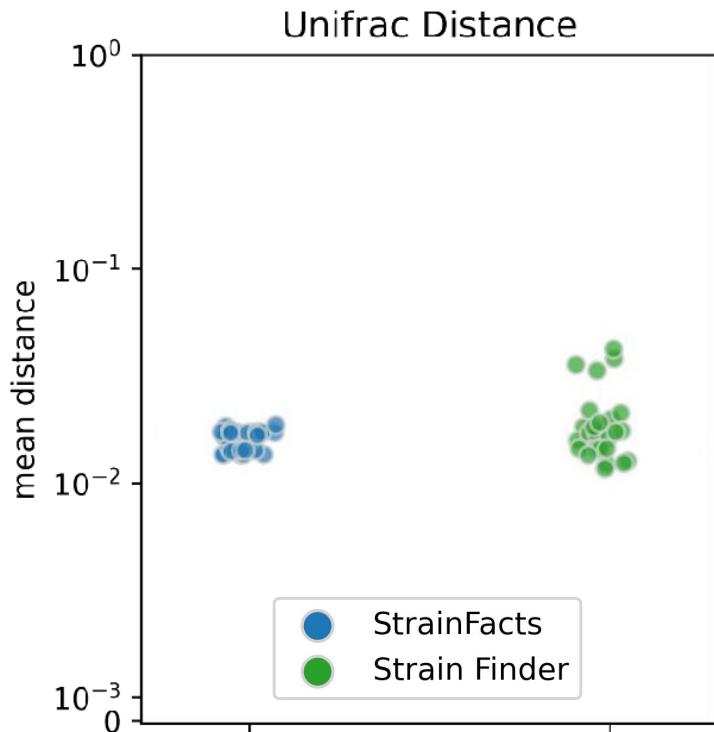
- Strain heterogeneity regularization (strains per sample)
- Overall strain diversity regularization

Full model also includes count over-dispersion (Beta-Binomial likelihood)

Maximum a posteriori (MAP) estimation for inference



StrainFacts is accurate

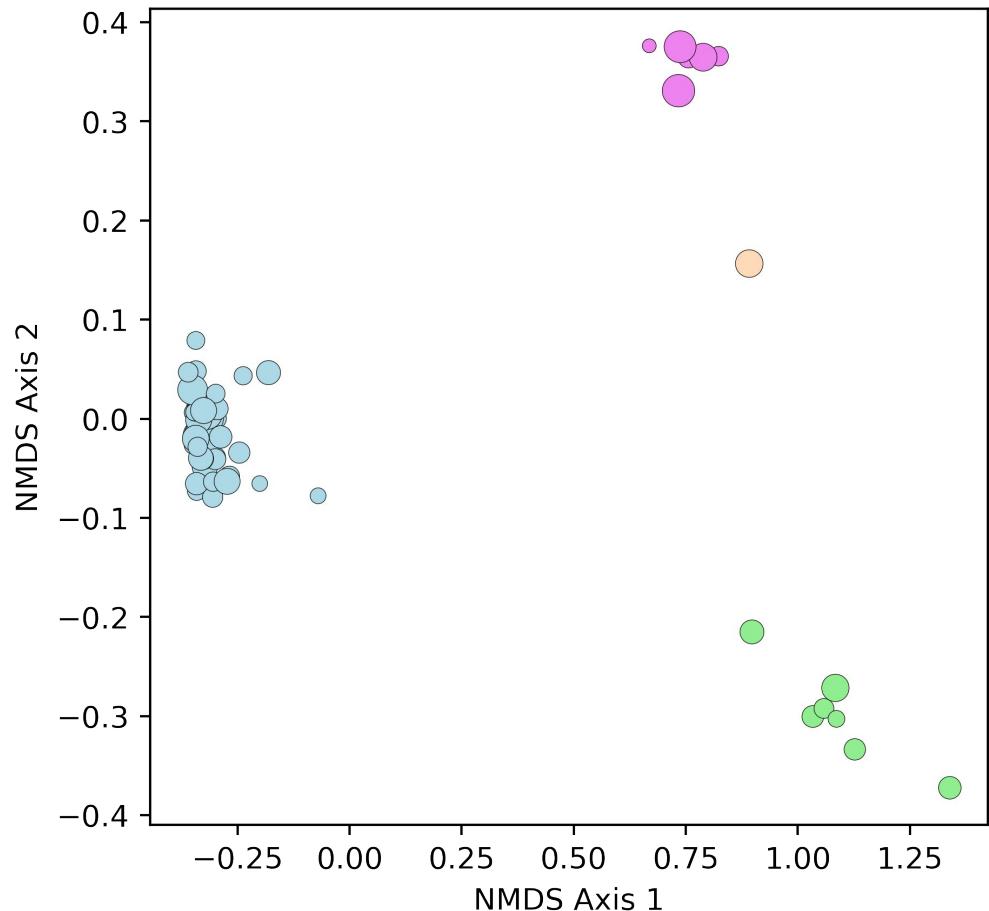


Based on simulations, StrainFacts is similarly accurate to an existing deconvolution tool

Several orders of magnitude faster

Single-cell genomics validates inferences in complex strain mixtures

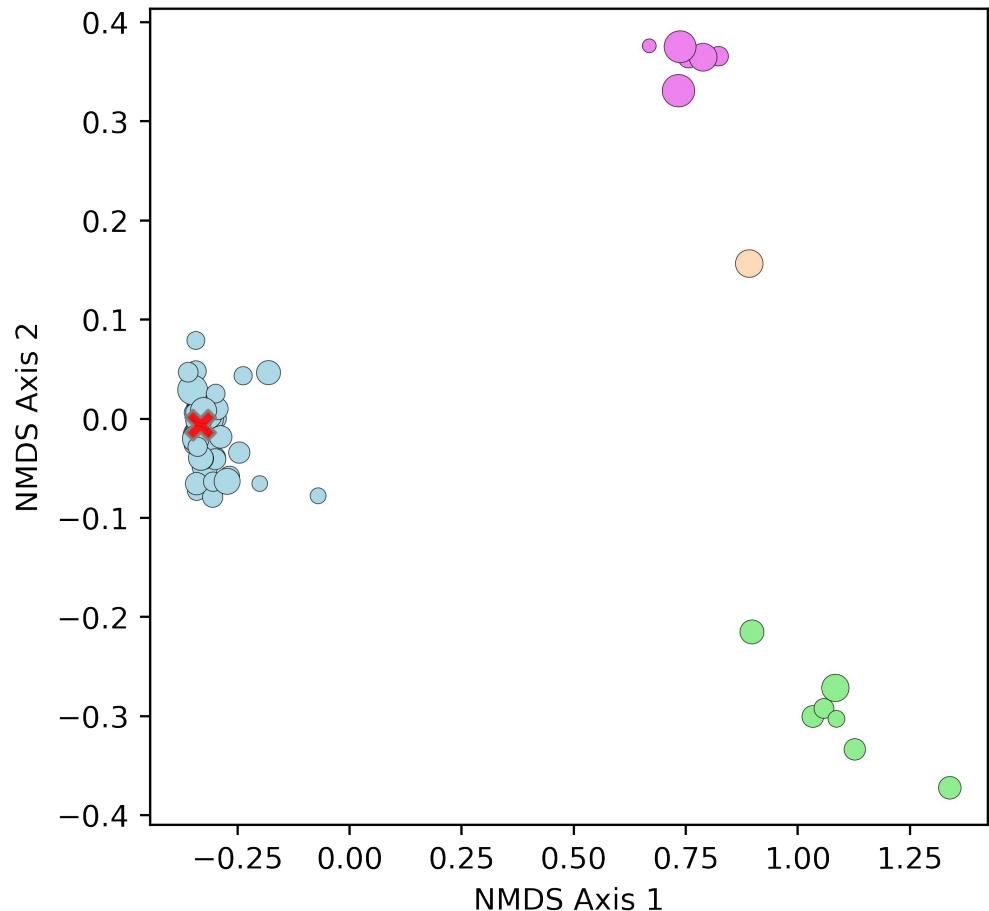
Streptococcus thermophilus
single-cell genotypes cluster into four
groups



Single-cell genomics validates inferences in complex strain mixtures

Streptococcus thermophilus
single-cell genotypes cluster into four
groups

Consensus metagenotype only
reflects dominant strain

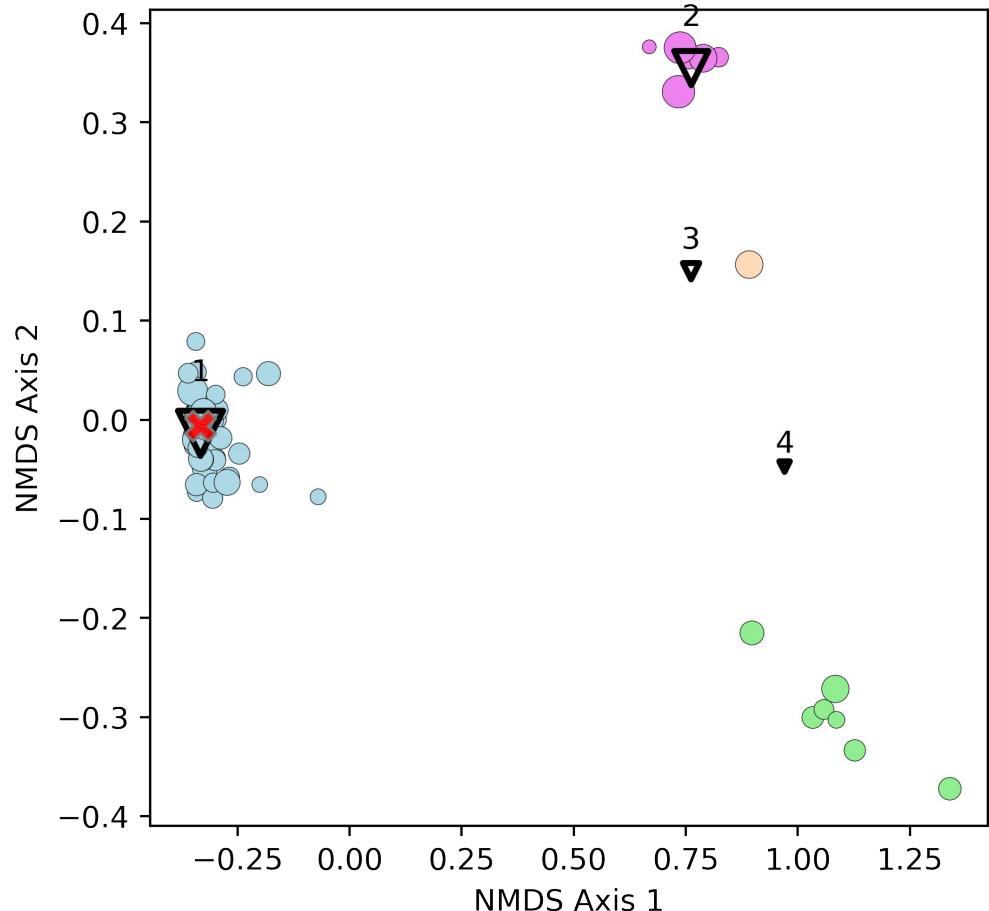


Single-cell genomics validates inferences in complex strain mixtures

Streptococcus thermophilus
single-cell genotypes cluster into four
groups

Consensus metagenotype only
reflects dominant strain

StrainFacts identifies four strains,
three of which match the single-cell
genotypes



Outline

Intraspecific diversity in the microbiome

Strain inference

Metagenotype deconvolution

Application to large metagenome collections

What can we do with strain inference across thousands of metagenomic samples?

What can we do with strain inference across thousands of metagenomic samples?

Two examples:

- Biogeography
- Population Genetics

Strain biogeography

Agathobacter rectalis is a prevalent
and abundant gut bacterium

20 countries

→ 33 studies

→ 9,224 stool samples

→ 198 inferred strains

Strain biogeography

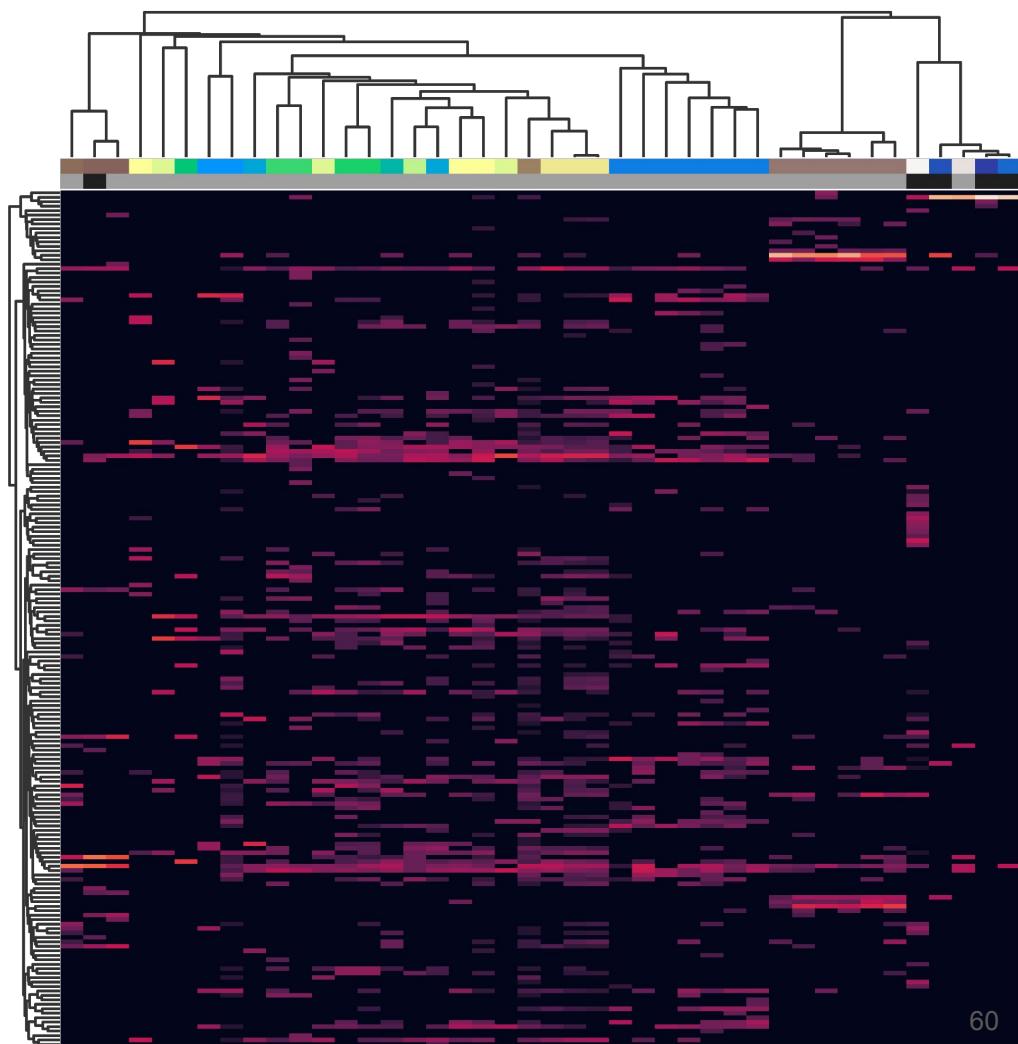
Agathobacter rectalis is a prevalent and abundant gut bacterium

20 countries

→ 33 studies

→ 9,224 stool samples

→ 198 inferred strains



Strain biogeography

Agathobacter rectalis is a prevalent and abundant gut bacterium

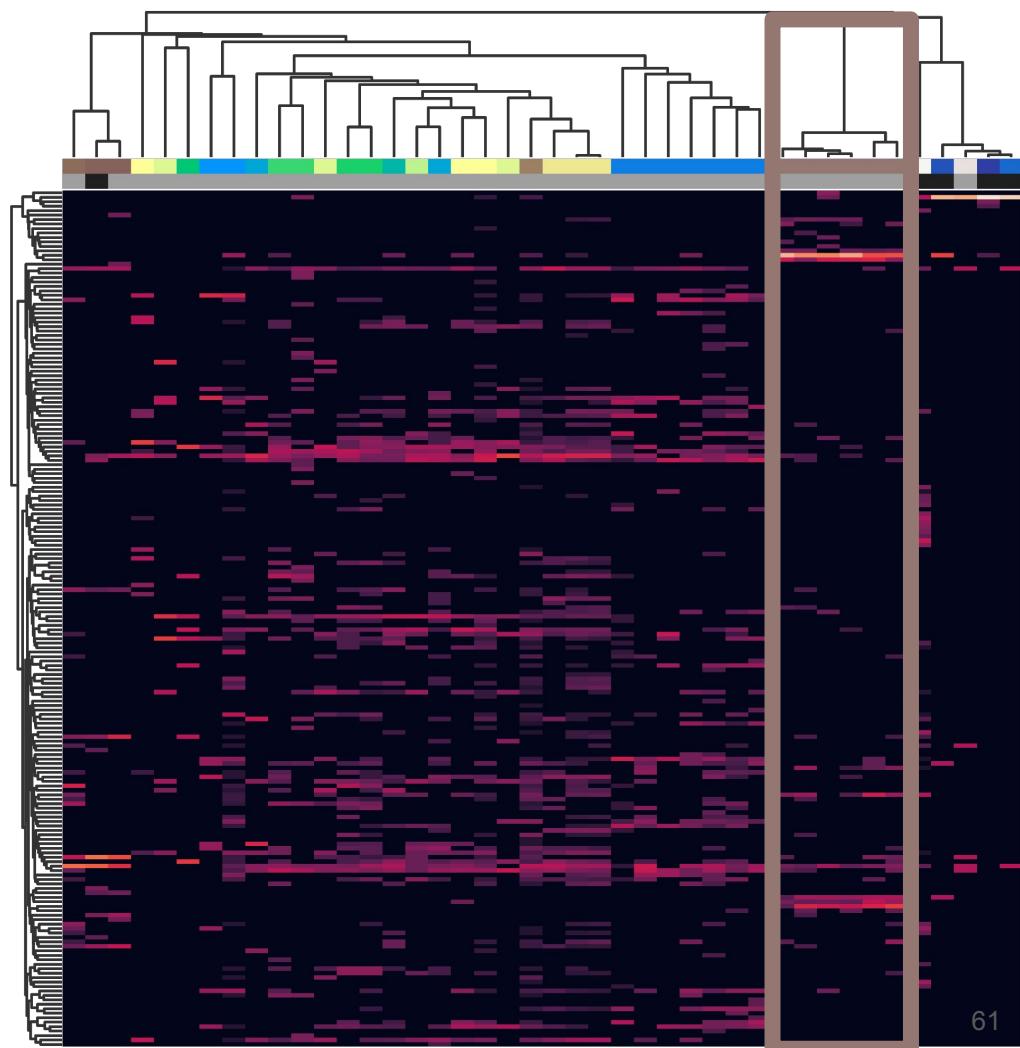
20 countries

→ 33 studies

→ 9,224 stool samples

→ 198 inferred strains

Strains reflect which country samples were collected in across independent studies



Strain biogeography

Agathobacter rectalis is a prevalent and abundant gut bacterium

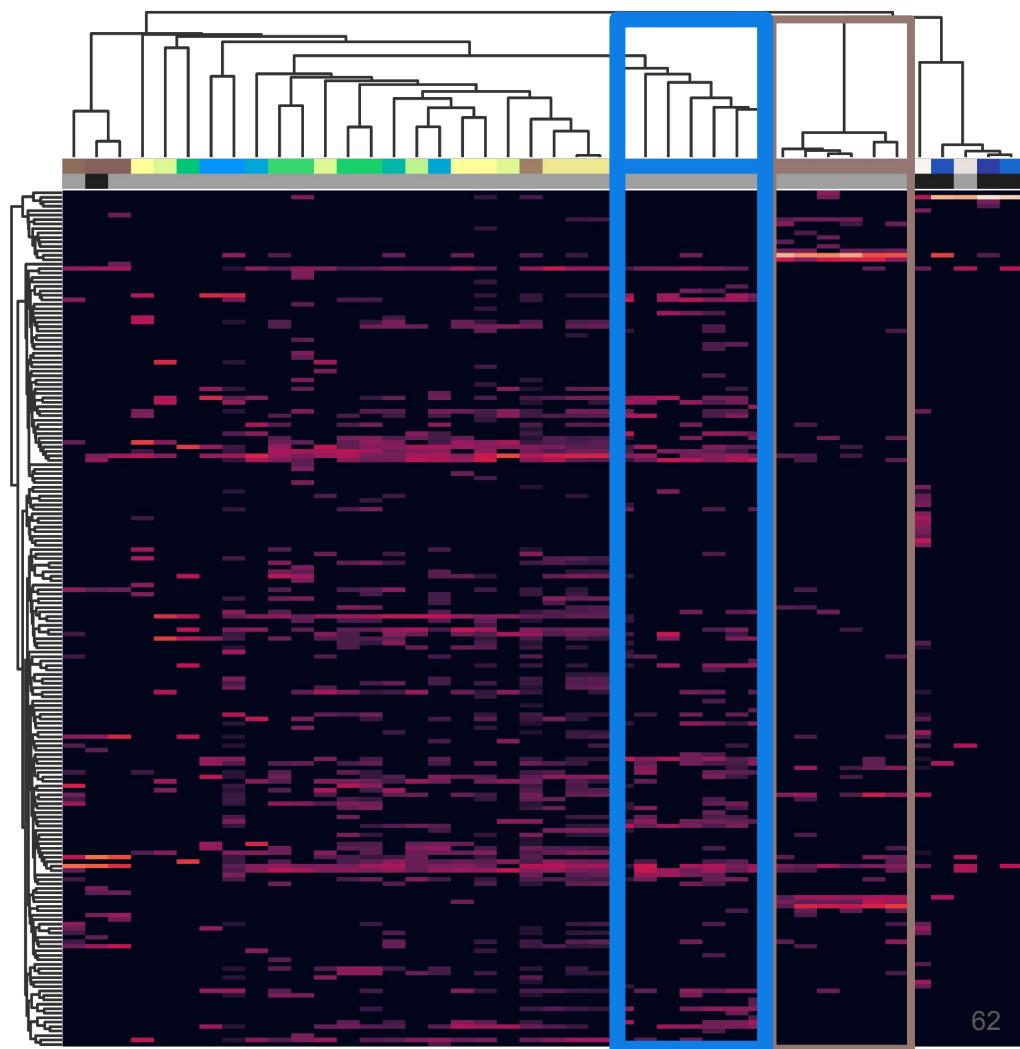
20 countries

→ 33 studies

→ 9,224 stool samples

→ 198 inferred strains

Strains reflect which country samples were collected in across independent studies



Strain biogeography

Agathobacter rectalis is a prevalent and abundant gut bacterium

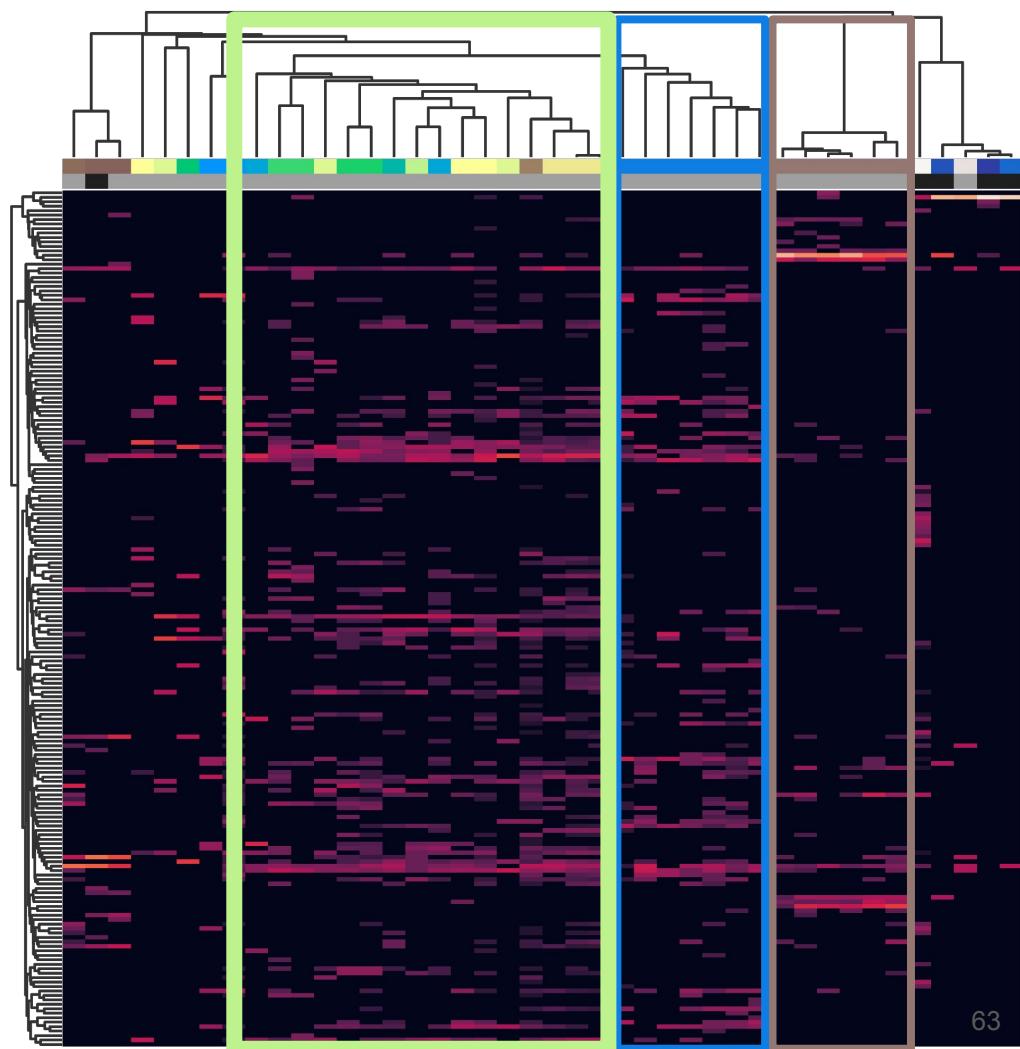
20 countries

→ 33 studies

→ 9,224 stool samples

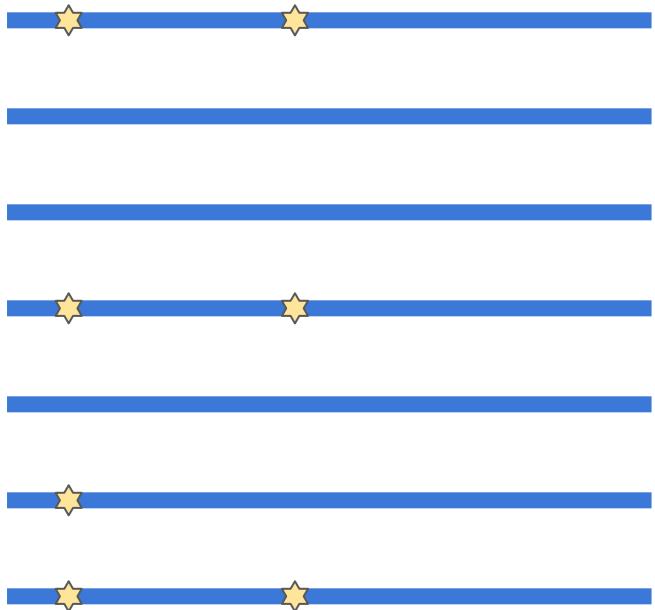
→ 198 inferred strains

Strains reflect which country samples were collected in across independent studies



Population genetics

Population genetics

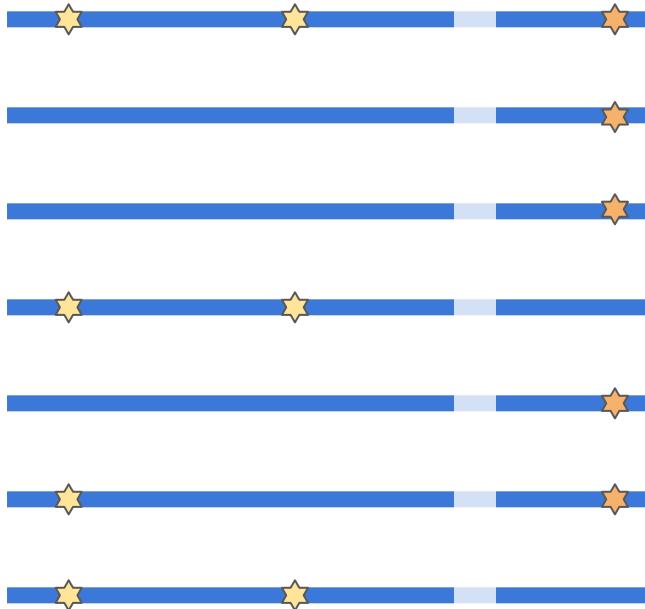


Linkage disequilibrium

Correlations between SNPs reflect shared descent

In asexual organisms, SNPs stay correlated

Population genetics



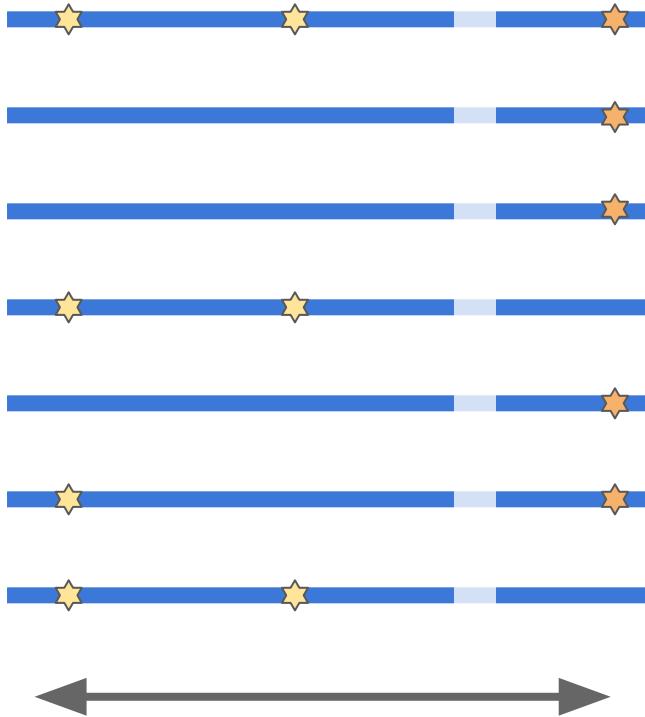
Linkage disequilibrium

Correlations between SNPs reflect shared descent

In asexual organisms, SNPs stay correlated

SNPs that are *not* correlated reflect recombination

Population genetics



Linkage disequilibrium

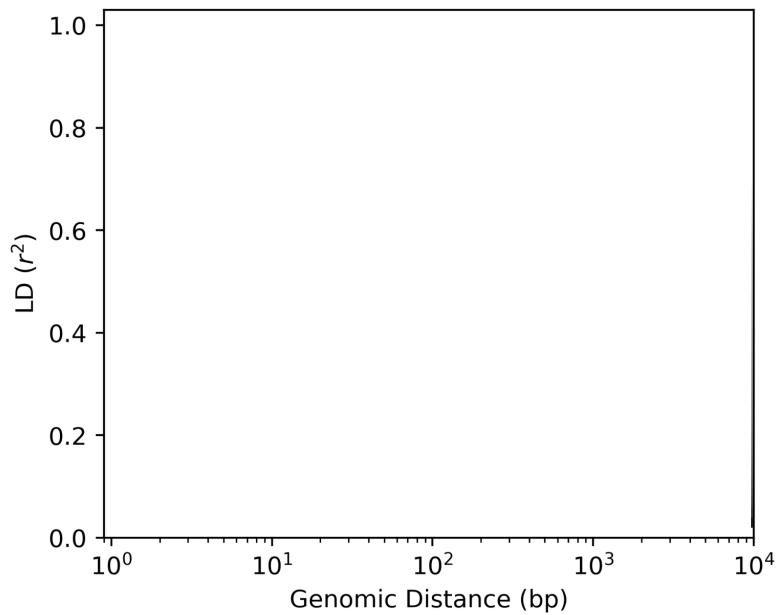
Correlations between SNPs reflect shared descent

In asexual organisms, SNPs stay correlated

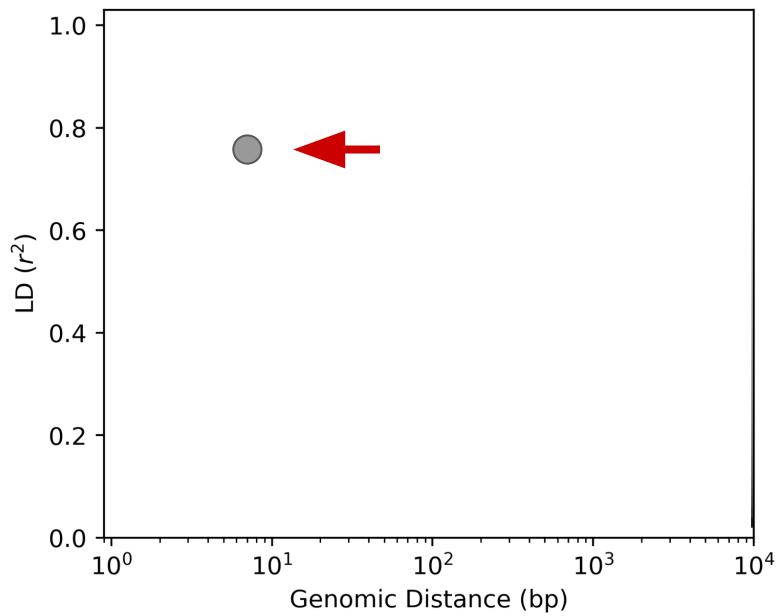
SNPs that are *not* correlated reflect recombination

We can see greater chance of recombination occurring between SNPs that are farther apart in the genome

Population genetics

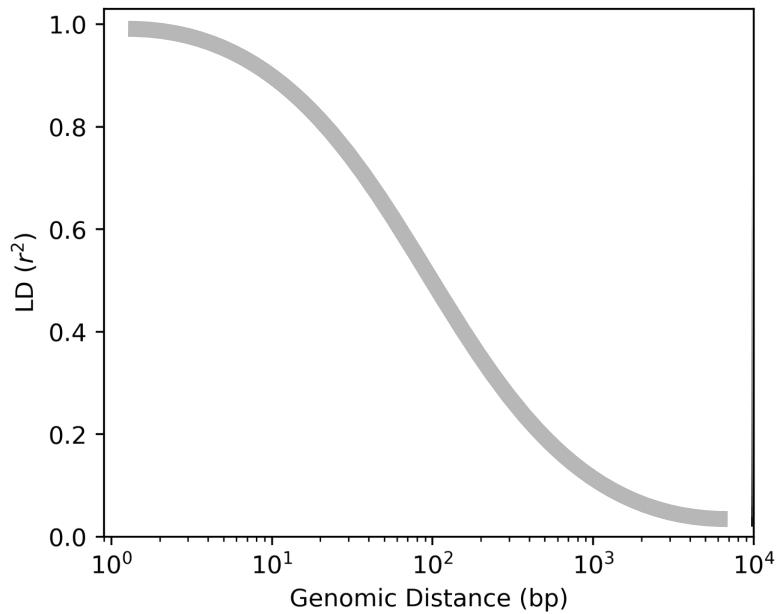


Population genetics



Population genetics

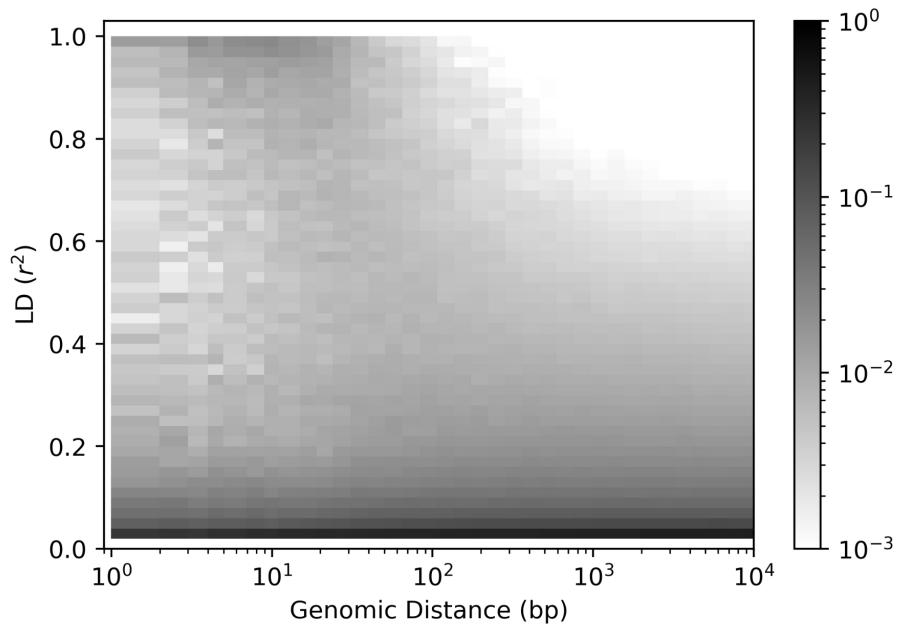
Decay of LD for pairs at larger distances reflects recombination



Population genetics

Decay of LD for pairs at larger distances reflects recombination

In *E. coli*, this is exactly what we see

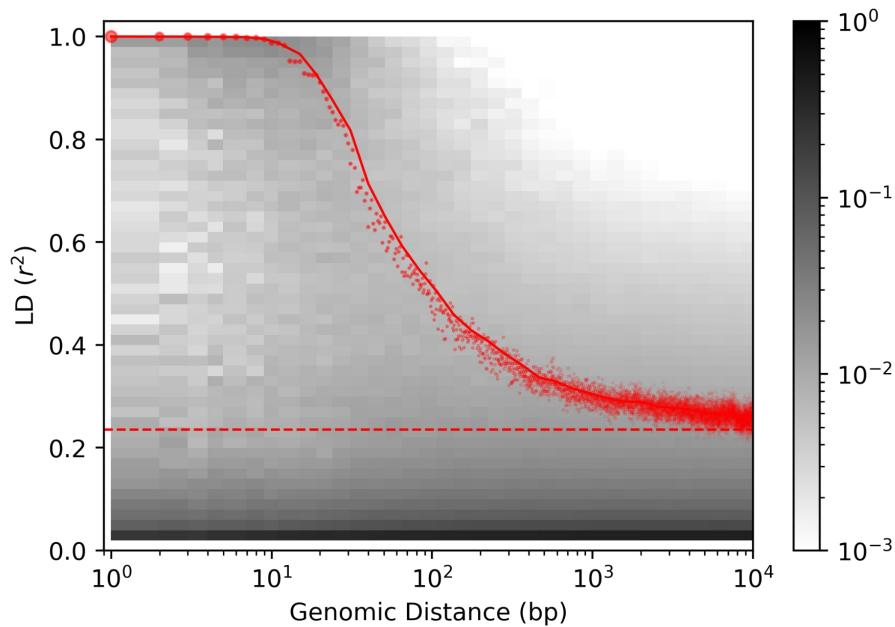


Population genetics

Decay of LD for pairs at larger distances reflects recombination

In *E. coli*, this is exactly what we see

Linkage disequilibrium quickly decays with distance, reflecting extensive recombination



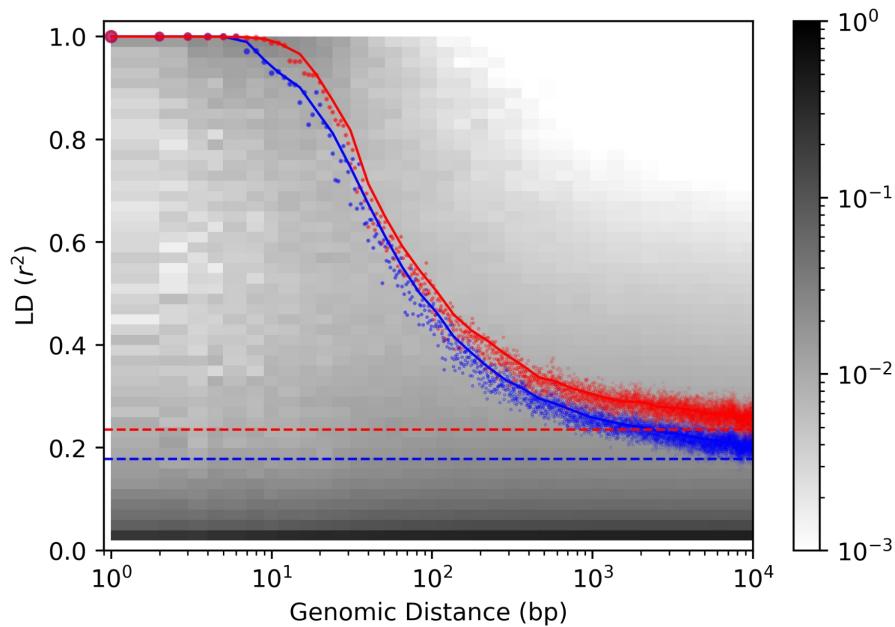
Population genetics

Decay of LD for pairs at larger distances reflects recombination

In *E. coli*, this is exactly what we see

Linkage disequilibrium quickly decays with distance, reflecting extensive recombination

This is consistent with what we independently calculate using reference genomes



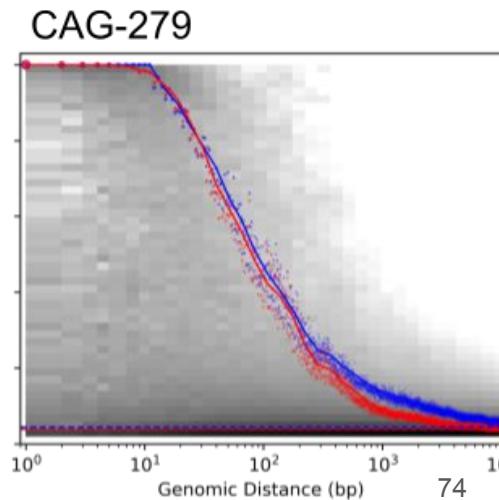
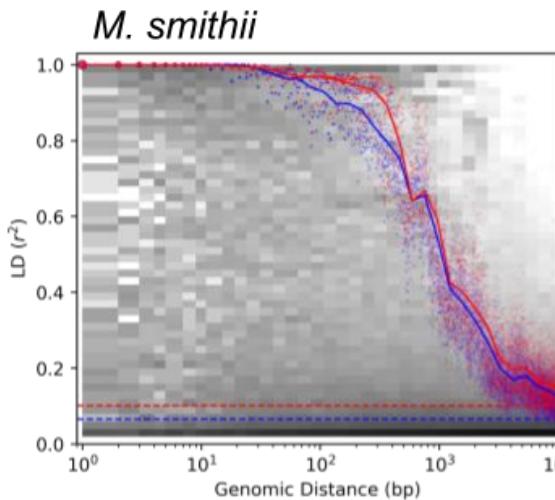
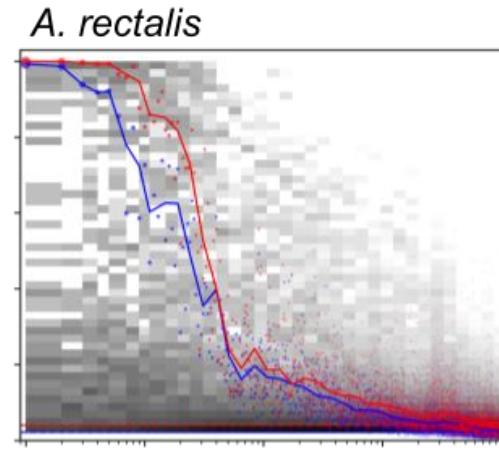
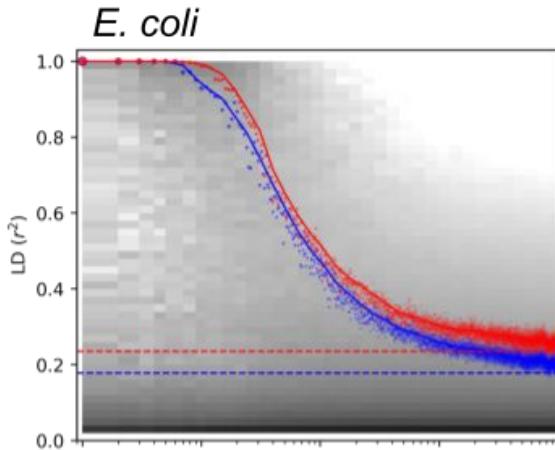
Population genetics

Decay of LD for pairs at larger distances reflects recombination

In *E. coli*, this is exactly what we see

Linkage disequilibrium quickly decays with distance, reflecting extensive recombination

This is consistent with what we independently calculate using reference genomes



Summary and Conclusions

Deconvolution integrates strain quantification and genotype reconstruction

StrainFacts scales the approach to tens-of-thousands of metagenomes

Validated with simulations and single-cell genomics

Enables microbial biogeography, population genetics, and more at a global scale

Summary and Conclusions

Deconvolution integrates strain quantification and genotype reconstruction

StrainFacts scales the approach to tens-of-thousands of metagenomes

Validated with simulations and single-cell genomics

Enables microbial biogeography, population genetics, and more at a global scale

Questions?

Pocket Slides

$$y_{ig} \sim \text{BetaBinom}(\tilde{p}_{ig}, \alpha^* | m_{ig})$$

$$\tilde{p}_{ig} = p_{ig}(1 - \epsilon_i/2) + (1 - p_{ig})(\epsilon_i/2)$$

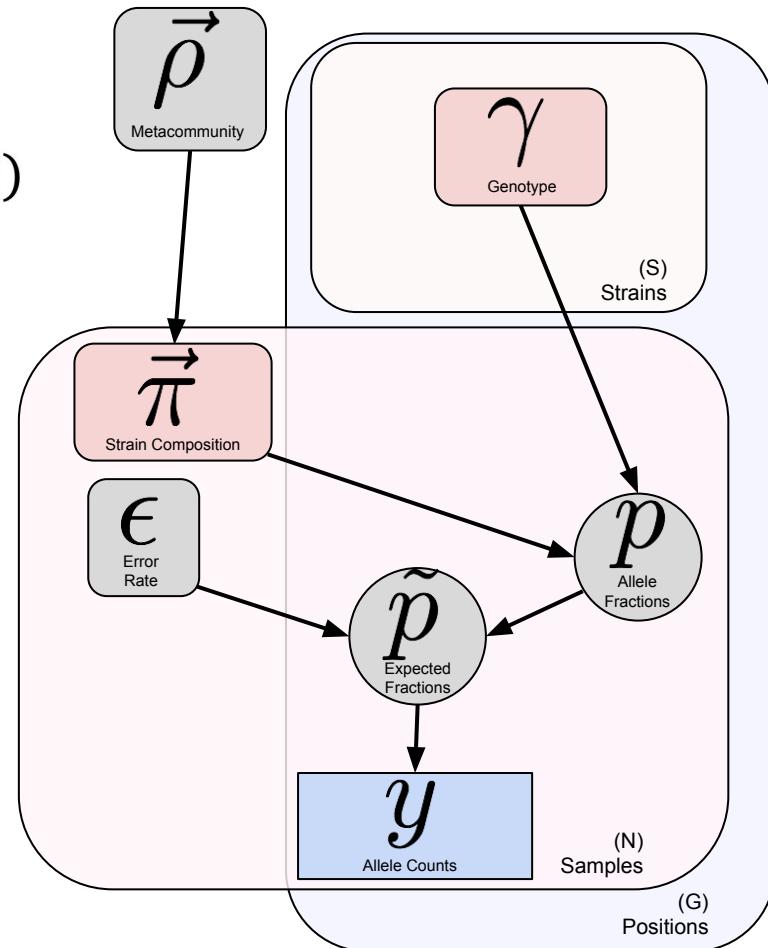
$$p_{ig} = \sum_s \pi_{is} \gamma_{sg}$$

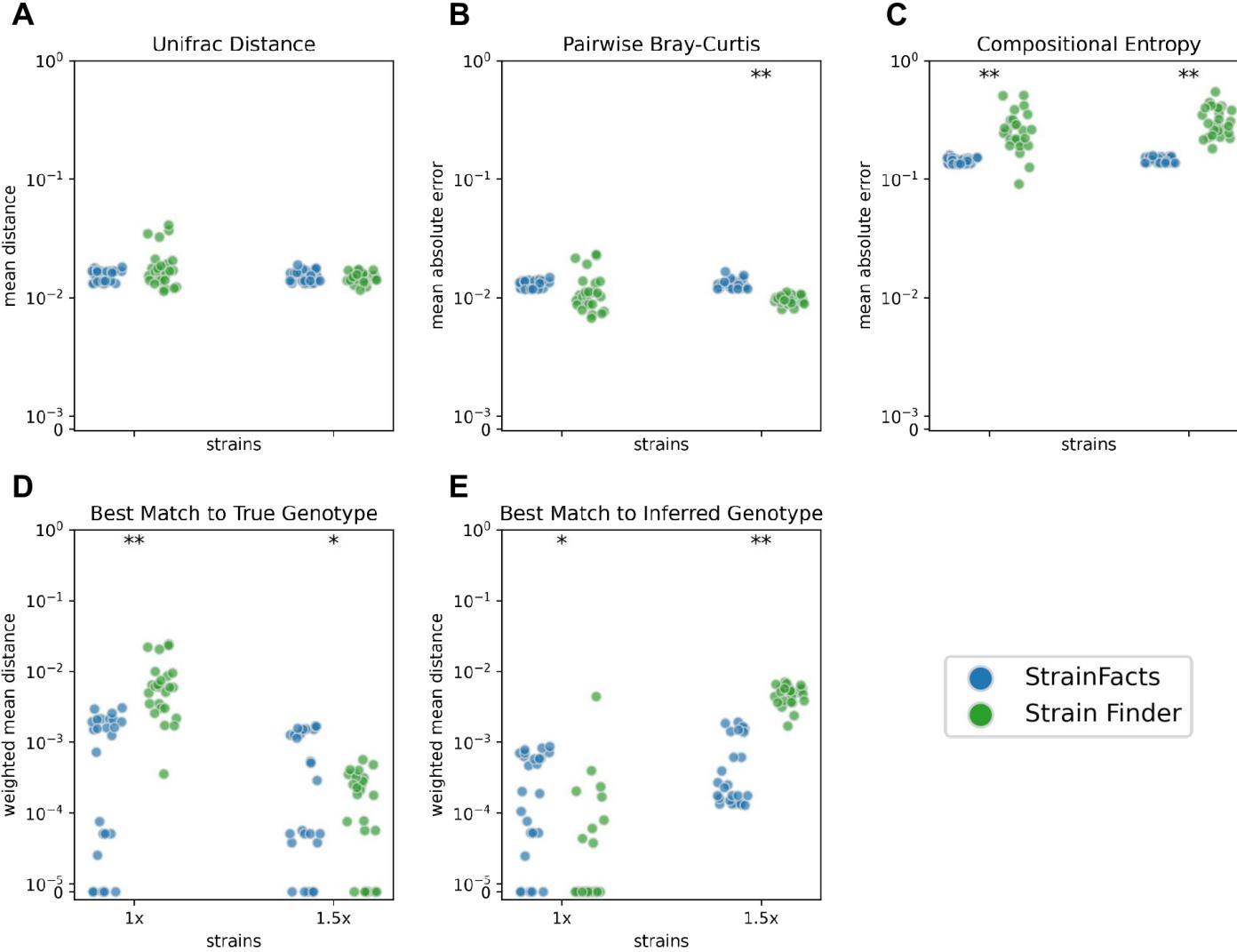
$$\gamma_{sg} \sim \text{SSD}_0 \left(\mathbf{1}, \mathbf{1}, \frac{1}{\gamma^*} \right)$$

$$\vec{\pi}_i \sim \text{SSD} \left(\mathbf{1}, \vec{\rho}, \frac{1}{\pi^*} \right)$$

$$\vec{\rho} \sim \text{SSD} \left(\mathbf{1}, \mathbf{1}, \frac{1}{\rho^*} \right)$$

$$\epsilon \sim \text{Beta} \left(\epsilon_a^*, \frac{\epsilon_a^*}{\epsilon_b^*} \right)$$



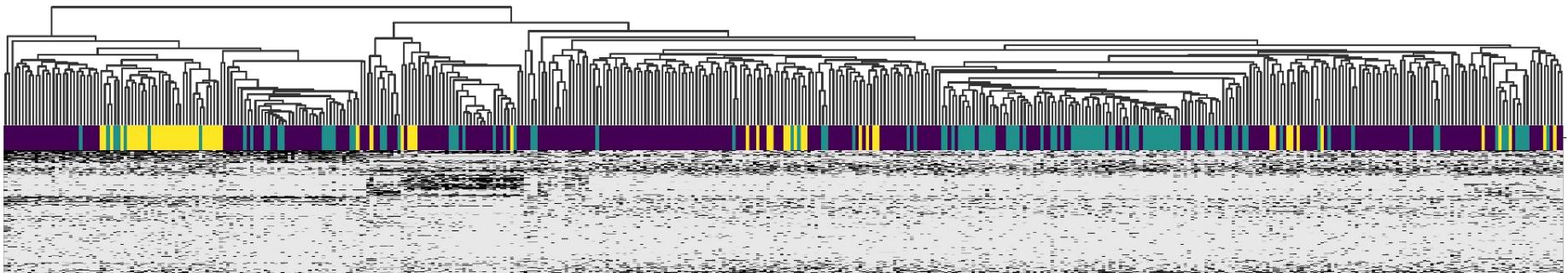


Cataloging strain diversity

■ Reference ■ Both ■ Inferred

Agathobacter rectalis is a prevalent and abundant gut bacterium

Dozens of studies → 11,860 samples → 198 inferred strains / 752 references



De novo strains reflect much of the diversity previously seen in references