

Scaling microbial strain inference to thousands of metagenomes using fuzzy genotypes

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

1

1. Thank you to TODO for the generous introduction, and thank you to the organizers for inviting me to present today.
2. I'm excited to tell you today about some very recent work on scaling up strain inference in shotgun metagenomic data.

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

2

1. You can find the method and all of the results that I present today in a preprint that we uploaded last month to BioRxiv
2. I'll also direct you to StrainFacts on my GitHub
3. And you can tweet @ my handle shown here

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

Before I get started I should acknowledge colleagues, institutions, and funding sources without whom this project would not have been possible. In particular, i want to thank my co-authors, and in particular Katie Pollard for having been a wonderful mentor throughout my postdoc.

Outline

Intraspecific diversity in the microbiome

Strain inference

Metagenotype deconvolution

Application to large metagenome collections

4

My talk today is roughly broken up into four sections

1. First, Introducing intraspecific diversity in the gut microbiome and why you should care
2. Then I'll describe some of the existing methods for understanding strains using shotgun metagenomics (and their shortcomings)
3. Then I'll lay out metagenotype deconvolution and how StrainFacts scales strain inference to large numbers of samples
4. Finally I'll spend the last section of my talk showing a few exciting results from a large collection of publicly available metagenomes

Outline

Intraspecific diversity in the microbiome

Strain inference

Metagenotype deconvolution

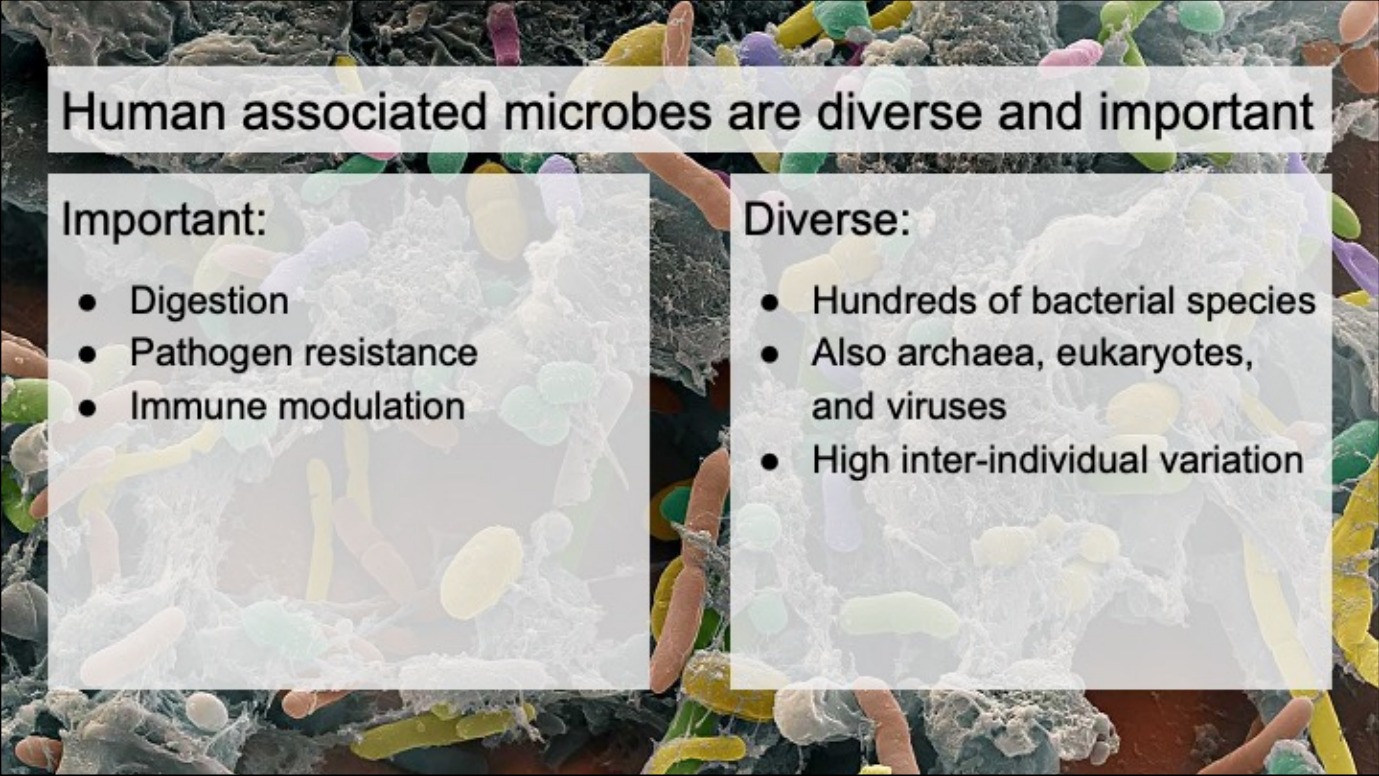
Application to large metagenome collections

Let's jump right in!

Human associated microbes are diverse and important



So before I tell you why you should care about strains in the human microbiome, let me start by very quickly telling you why you should care about the microbiome.

A scanning electron micrograph (SEM) showing a dense community of diverse human-associated microbes. The image displays various shapes and sizes of microorganisms, including rod-shaped bacteria, spherical cocci, and filamentous structures, all appearing to be part of a complex, interconnected microbial community. The background is dark, highlighting the intricate details of the microbial surfaces.

Human associated microbes are diverse and important

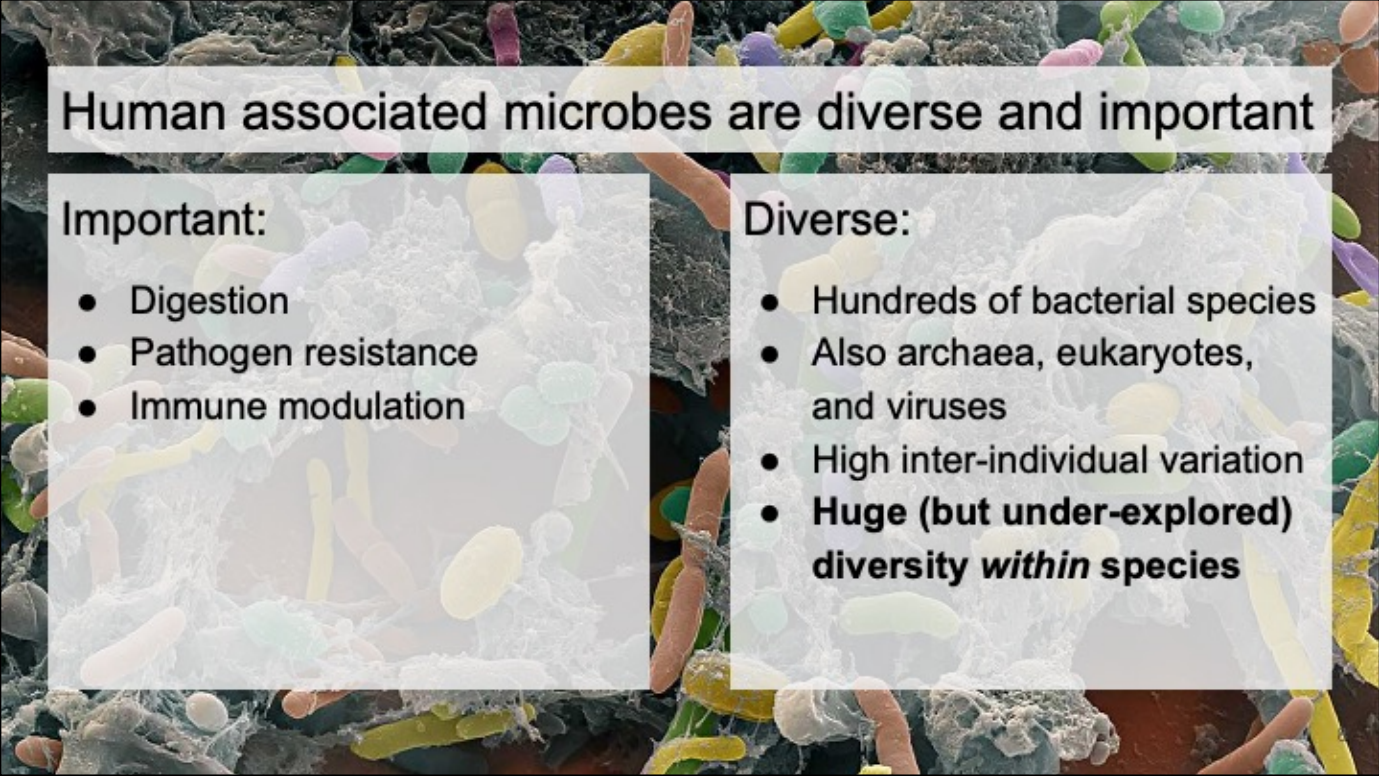
Important:

- Digestion
- Pathogen resistance
- Immune modulation

Diverse:

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

1. Microbes inhabiting the gut play important and well documented roles in host health.
2. Besides very obvious impacts like diseases, gut microbes also play a documented role in digestion of food, pathogen resistance, modulating the immune system, and more
3. I also want to stress how diverse the microbiome is, both within a single individual who can have hundreds of bacterial species (not to mention numerous archaea, protists, and viruses)
4. But also between individuals, which usually have more differences than similarities.

A scanning electron micrograph (SEM) showing a dense community of diverse human-associated microbes. The image displays various shapes and sizes of microorganisms, including rod-shaped bacteria, spherical cocci, and filamentous structures, all appearing to be part of a complex, interconnected microbial community. The background is dark, highlighting the intricate details of the microbial surfaces.

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**

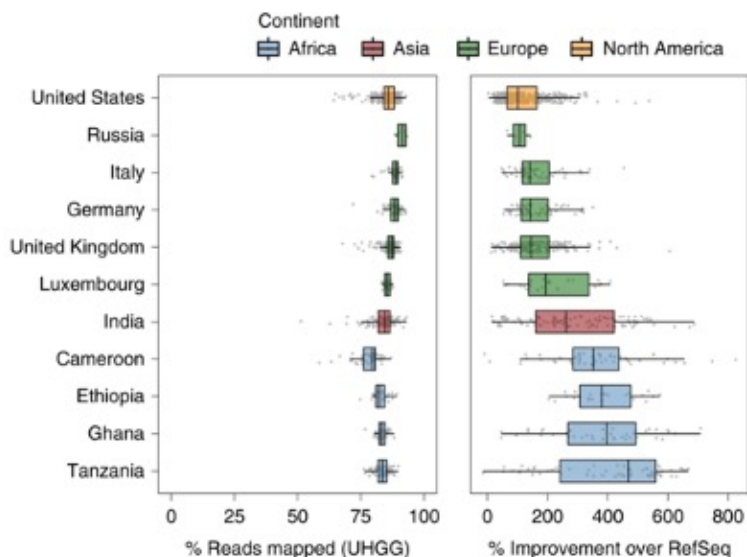
1. The focus of this presentation, however, will be on a different kind of diversity: within-species diversity.
2. It is becoming increasingly clear that when considering microbes at a strain-resolution, the intra-individual and inter-individual diversity is even higher, potentially with important impacts on the functions of the microbiome

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

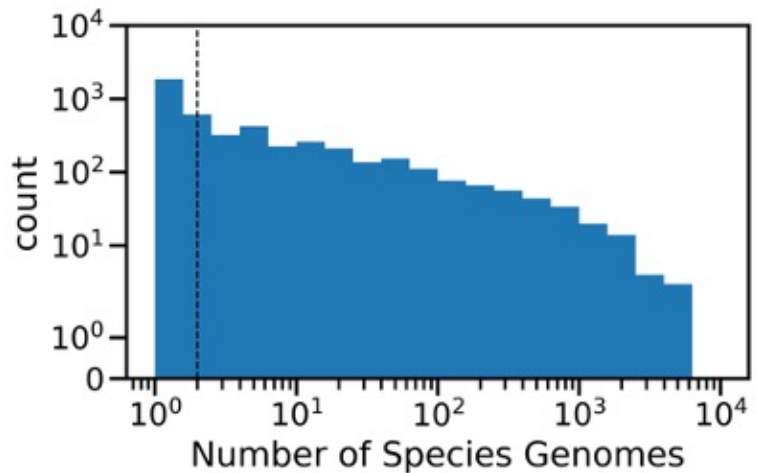


Almeida, A. et al. *Nat Biotechnol* 39, 105–114 (2021).
9

1. This understanding is being made possible by microbial genome references.
2. After 20+ years of studying the human gut microbiome, projects like the unified human gastrointestinal genome (UHGG)—which specifically harnesses culture-free genomes (MAGs)—are nearing a complete representation of species in the human gut.
3. I say this based on results like the figure on the right, which I've borrowed from the UHGG paper in 2021, which shows the fraction of shotgun metagenomic reads that map to that reference database, now finally >75-90%, which is a 2x or greater improvement over previous databases.
4. It's still important to point out that the degree of database coverage is lower in understudied populations, as we can see with samples from these four African nations. There's still more work needed to gather a truly representative database.

Strain diversity is not well documented for vast majority

75% of species have fewer than 10 representative genomes



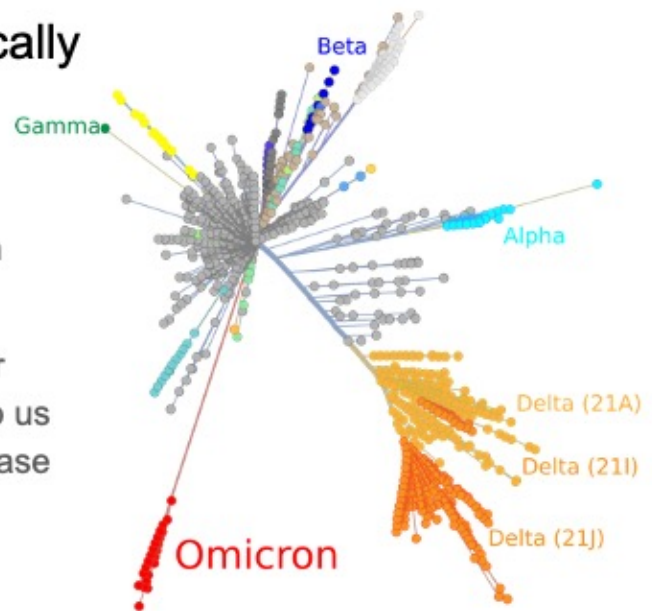
10

1. What's more, while we might now have a reference for the most common species
2. For the large majority of species, we don't have much strain diversity represented
3. The median number of representative genomes across all species in the UHGG is 2
4. And 75% of species have fewer than 10 genomes.
5. This means that for most species we know little or nothing about intraspecific diversity

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.



<https://nextstrain.org>

11

1. And strain level diversity is *important*.
2. We've been talking about it for almost the last two years.
3. Just like SARS-coV-2, we can see important differences in the traits of different microbial strains
4. And like SARS-coV-2, sequence comparisons can also inform our understanding of microbial origins, transmission, ecology, and evolution

Strains of a species can have important differences

Intraspecific diversity has been appreciated for a long time

- Pathogenicity
- Antibiotic resistance
- Phage resistance
- Auxotrophy

12

1. If that's not enough motivation, we know a number of the ways that strain diversity can be biologically important in the gut microbiome
2. For instance, there are traits that are especially obvious, like pathogenicity and antibiotic resistance
3. Along with a few other traits like phage resistance and various metabolic auxotrophies.

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

1. One theme that I think is worth pointing out for these strain-specific traits is that they show up in *E. coli* and other model-microbes
2. These have been relatively easy to identify and study with a pure culture
3. But this is just the tip of the iceberg when it comes to strain-diversity

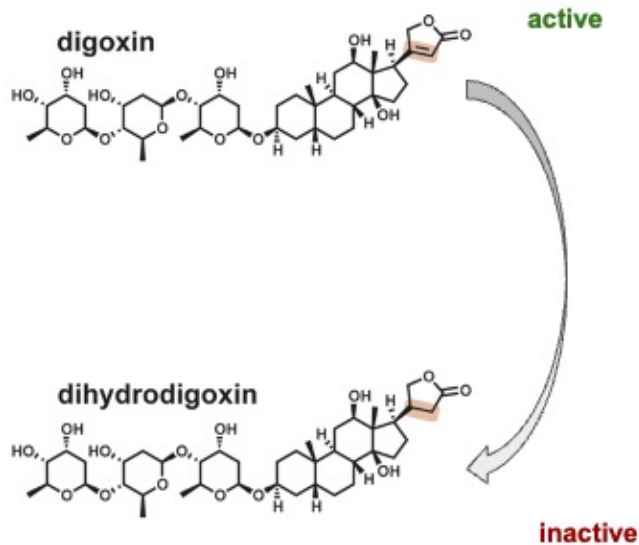
Strains of a species can have important differences

Intraspecific diversity has been appreciated for a long time

- Pathogenicity
- Antibiotic resistance
- Phage resistance
- Auxotrophy

e.g. *E. coli*

Well studied, easy to culture



Haiser, H.J., et al. *Gut Microbes* 5, 233–238 (2014).

14

1. For instance, here's a characteristic example: Digoxin is a cardiac glycoside, a commonly prescribed drug used to treat several heart conditions
2. However, when digoxin is reduced to dihydrodigoxin by some members of the gut microbiome it is inactivated,
3. changing its pharmacological profile in medically relevant ways

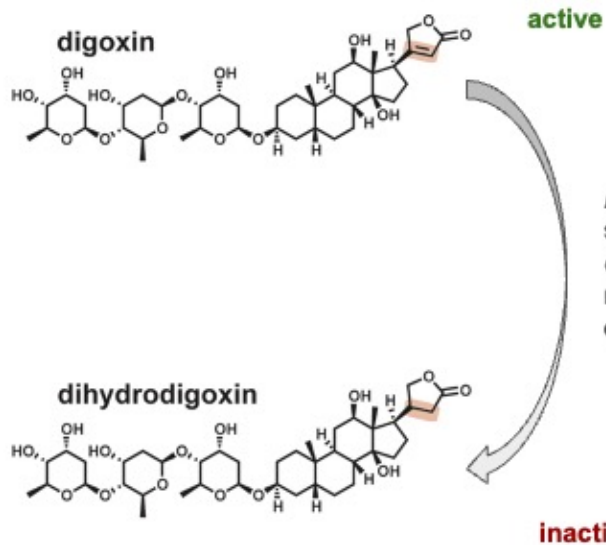
Strains of a species can have important differences

Intraspecific diversity has been appreciated for a long time

- Pathogenicity
- Antibiotic resistance
- Phage resistance
- Auxotrophy

e.g. *E. coli*

Well studied, easy to culture



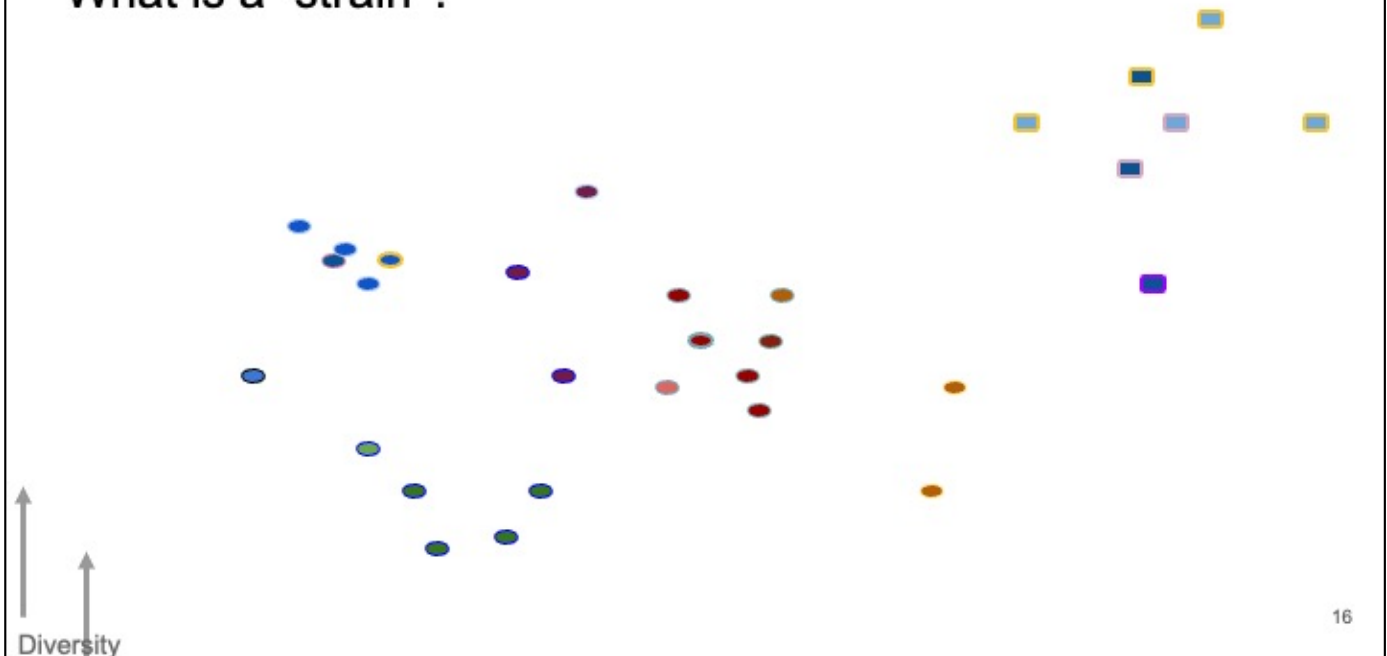
Eggerthella lenta strains encoding the *cgr*-operon result in reduction/inactivation of digoxin

Haiser, H.J., et al. *Gut Microbes* 5, 233–238 (2014).

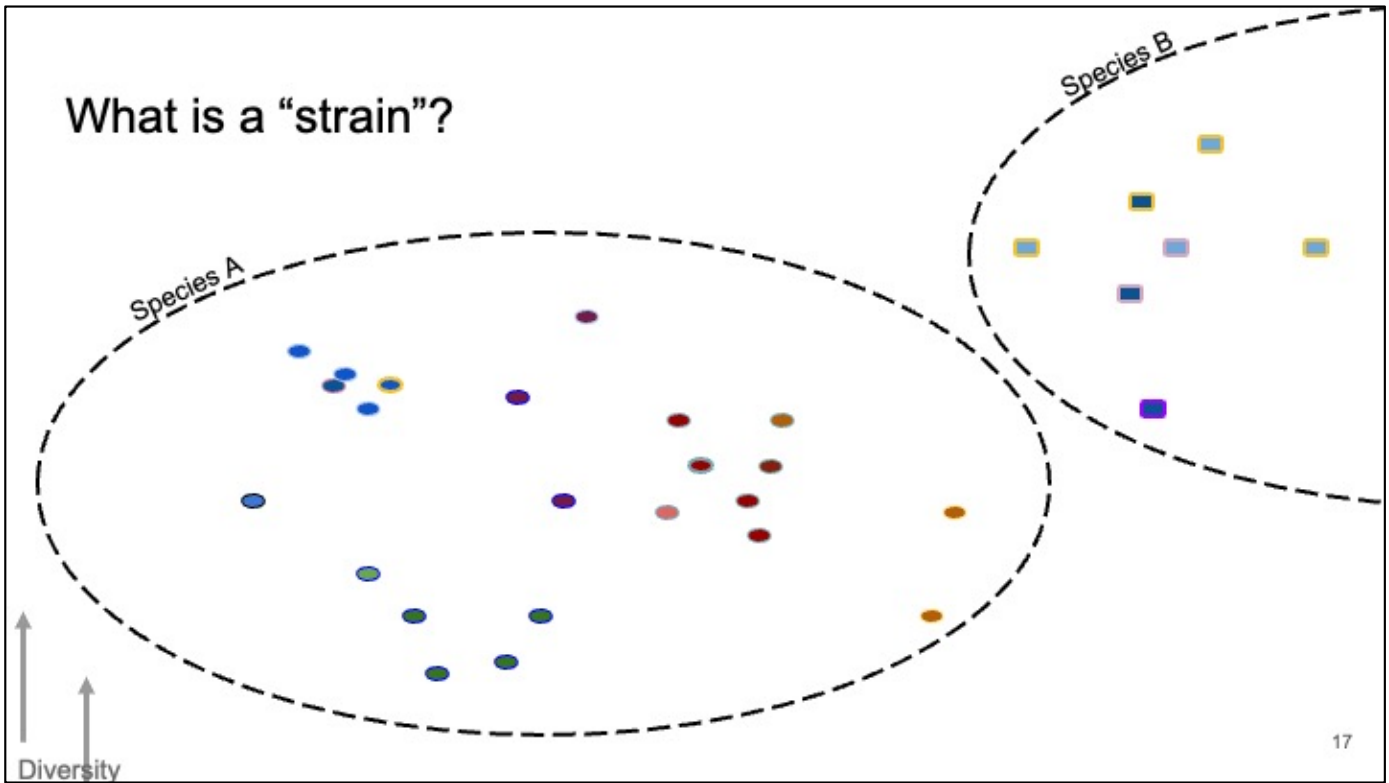
15

1. Importantly, while *Eggerthella lenta* is known to carry-out this process,
2. Only some strains have the *cgr*-operon that encodes this reduction
3. This motivates us to ask: “What other strain-specific traits are we missing when we only consider species-level taxonomy?”

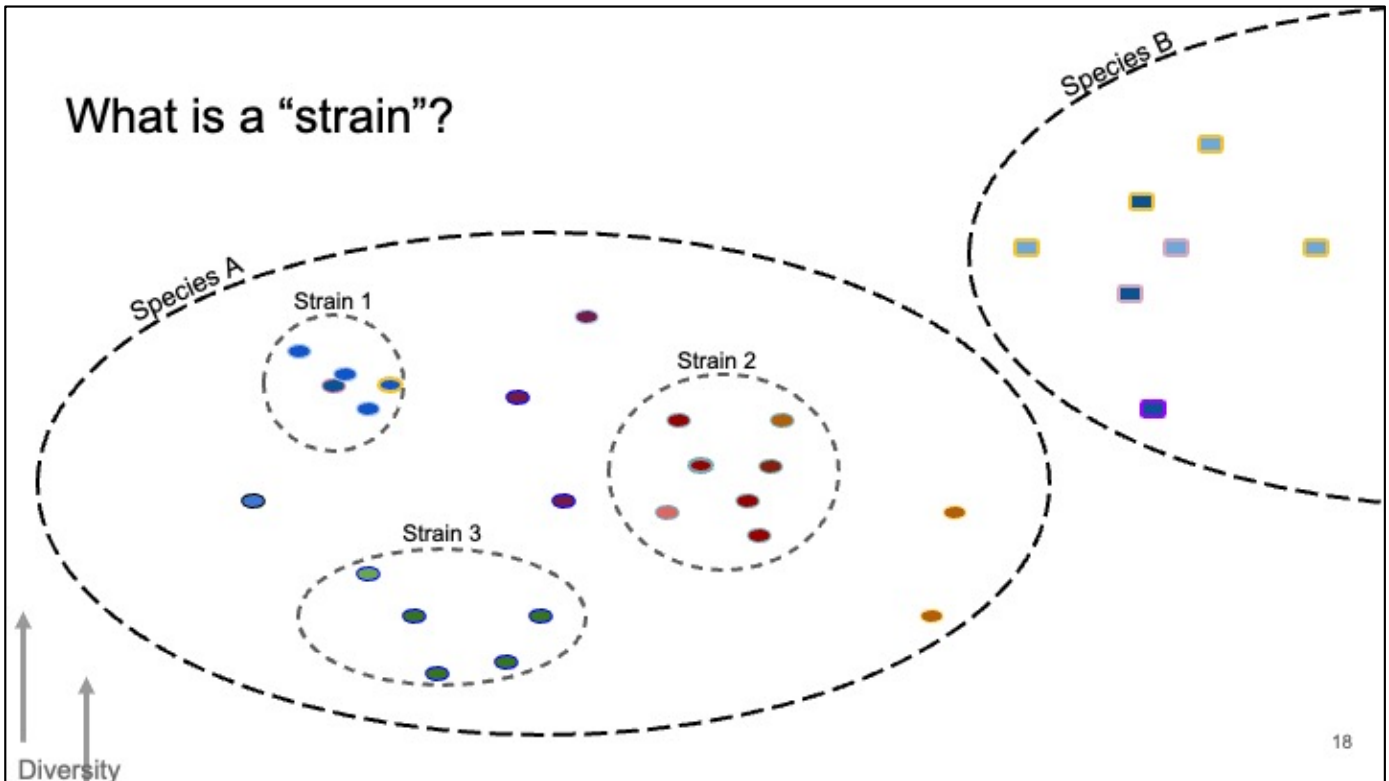
What is a “strain”?



1. Before we go any further I want to quickly cover the semantics of the term “strain”.
2. If microbial diversity is conceptually flattened into two dimensions
3. we might see that every single isolate has *something* that distinguishes it from others

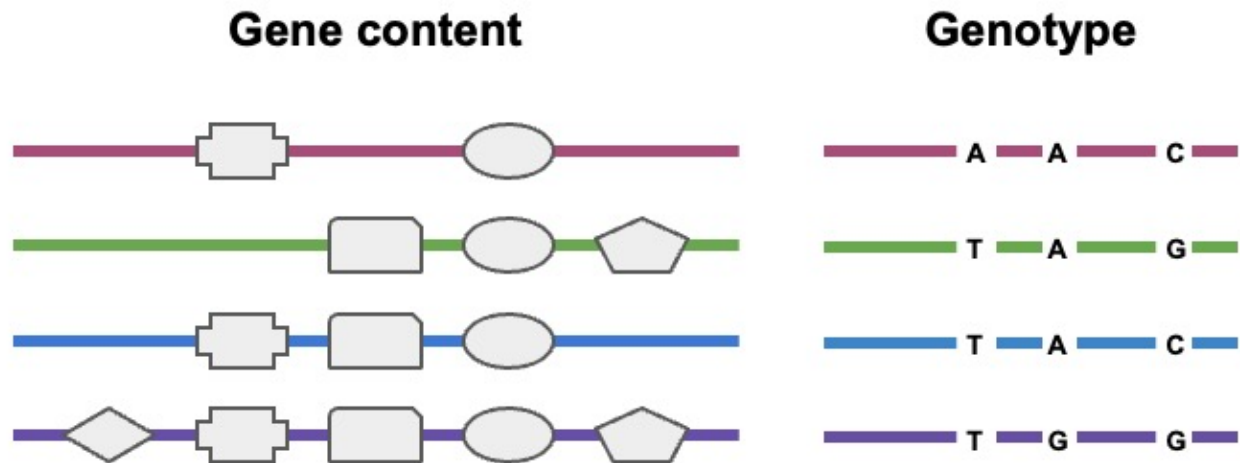


1. We divide this diversity space at the largest level into species



1. For this presentation, strains are groups of very similar genomes clustered within species
2. When we talk about "strain inference" I want to be clear that I'm using this operational definition of strains based on what we have the technical ability to differentiate.

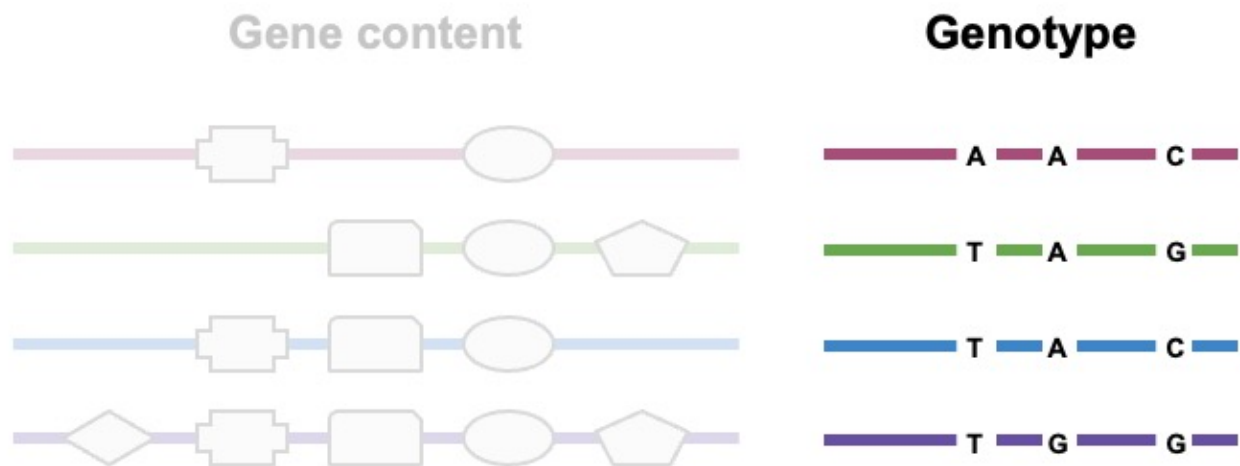
What differentiates strains?



19

1. And when I say similar genomes, I'm broadly referring to two different, but related things.
2. Strains can differ in
 - a. the set of genes encoded
 - b. Or the sequence of single-nucleotide variants at polymorphic sites within these genes

What differentiates strains?



20

1. While the first is clearly very important, today I'm talking exclusively about the second
2. Differentiating strains based on the sequence of variants at SNP sites in the core genome

Outline

Intraspecific diversity in the microbiome

Strain inference

Metagenotype deconvolution

Application to large metagenome collections

1. This is where strain inference comes in

Existing methods lack taxonomic resolution

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

1. 16S rRNA gene is highly conserved, and therefore has only limited phylogenetic resolution

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



23

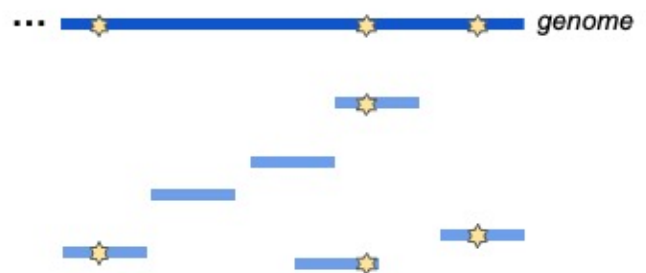
1. On the other hand, shotgun metagenomic data is increasingly available, raising many exciting possibilities in microbiome science.
2. However, standard methods for taxonomic characterization using metagenomic data are mostly limited to mapping reads to reference genome databases
3. and mostly only resolve taxonomy at the species level

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



1. Nonetheless, shotgun metagenomic reads mapping across SNP sites
2. include single-nucleotide variants that characterize strains

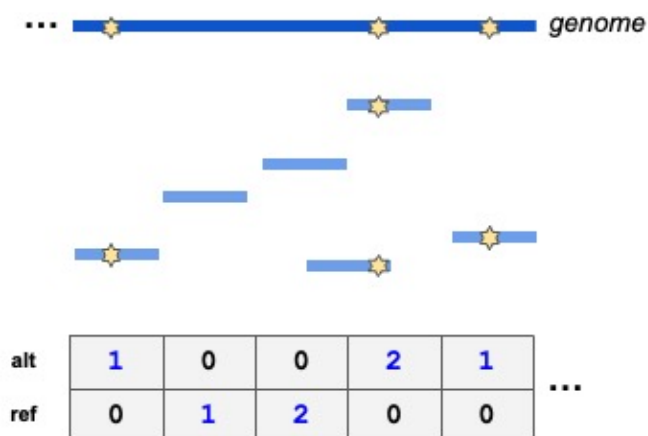
“Metagenotyping”

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)



1. We call the process of tallying up all the alleles seen in reads covering each of these SNP sites: “metagenotyping”.
2. The “Genotyping” just like what we do for individual single organisms

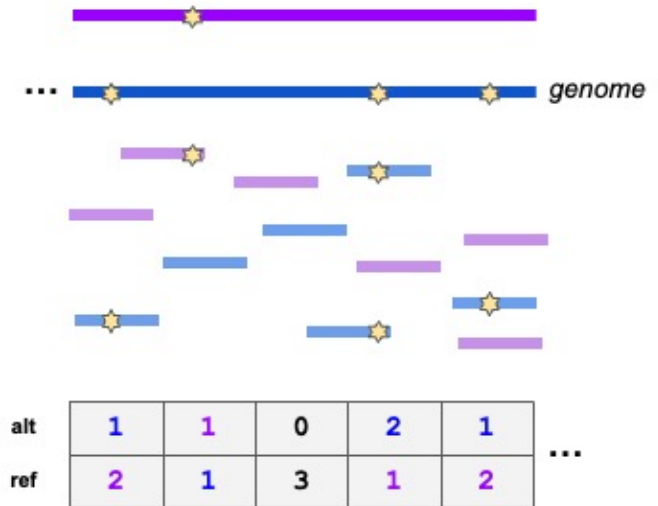
“Metagenotyping”

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

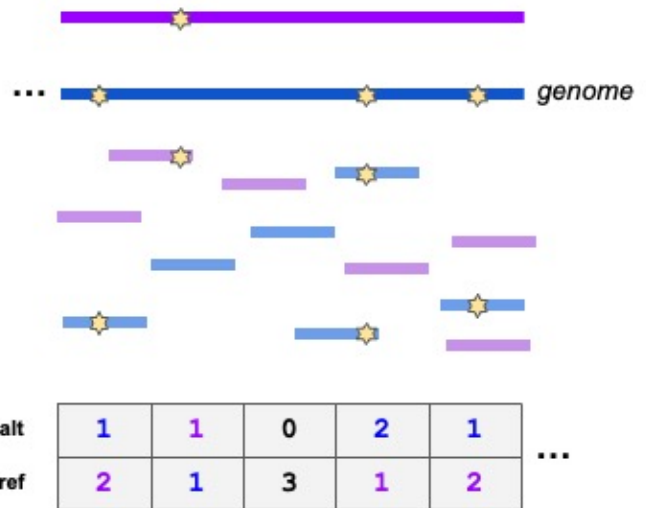
Metagenotypers tally variants at polymorphic positions (SNPs)



1. But “meta”-genotyping because multiple strains can co-exist in the same sample and may be sequenced

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

Uses exact k-mer matching to accelerate metagenotyping

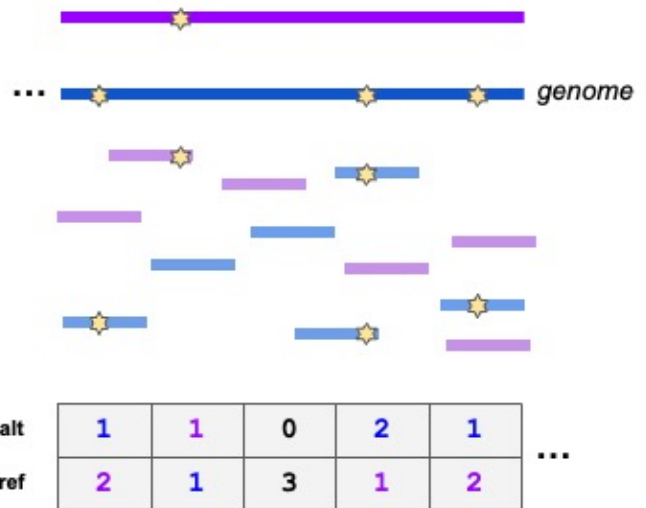


1. While a multitude of tools that map reads and count alleles at SNP sites have been created in the past few years
2. The work that I'm presenting today was really motivated by a metagenotyper built by my colleagues in the Pollard Lab
3. GT-Pro is a very very FAST metagenotyper, which trades read alignment for exact k-mer matching
4. This depends on a pre-computed database of SNVs

GT-Pro scales to tens-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

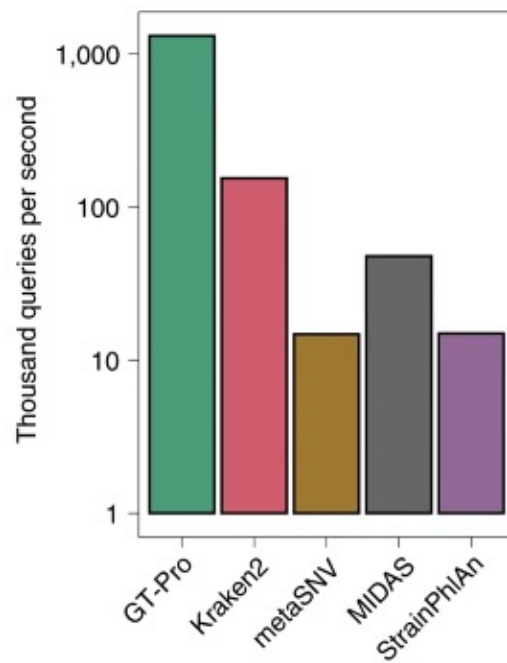


1. GT-Pro focuses on bi-allelic sites, which make up >90% of all SNPs
2. And also focuses on sites in the core genome
3. While simultaneously tallying SNVs for almost a thousand species in the default database

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



29

1. And it really is fast. GT-Pro is about an order of magnitude faster than other metagenotypers
2. Making it possible to metagenotype hundreds of species in tens of thousands of human microbiome samples in publicly available sequence databases
3. However, while metagenotyping is now easy and fast...

Low
sequencing
coverage

Mixtures
of strains

**Interpreting
metagenotypes
is challenging**

Closely
related
strains

Novel
strains

30

1. Interpreting metagenotypes is still hard
2. Specifically, four key challenges exist...
 - a. Metagenotypes for low abundance species can be sparse
 - b. Closely related strains may not be well differentiated by their metagenotype
 - c. Most strains may not have been previously characterized
 - d. And, crucially, as I've mentioned before, a single sample may have multiple, co-existing strains

**Low
sequencing
coverage**

Consensus Genotypes

**Mixtures
of strains**

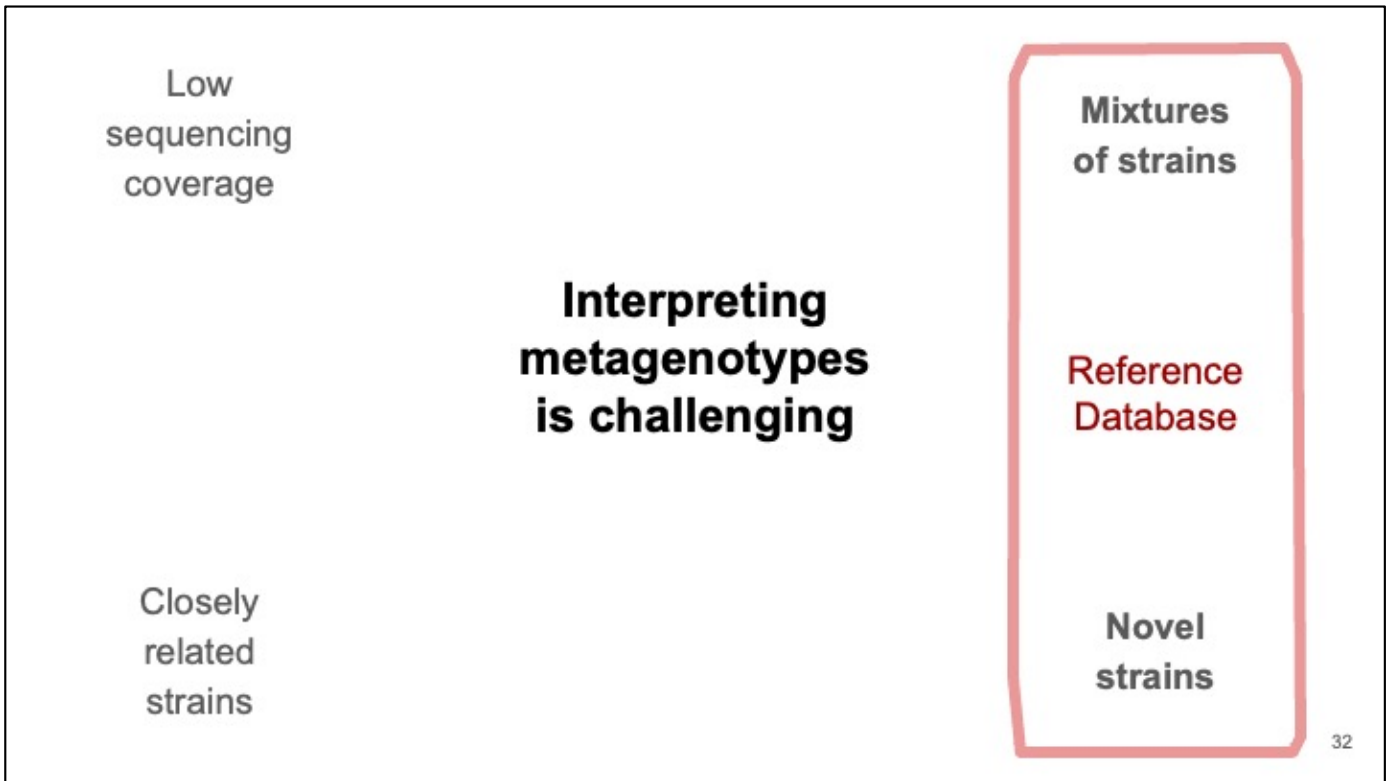
**Interpreting
metagenotypes
is challenging**

Closely
related
strains

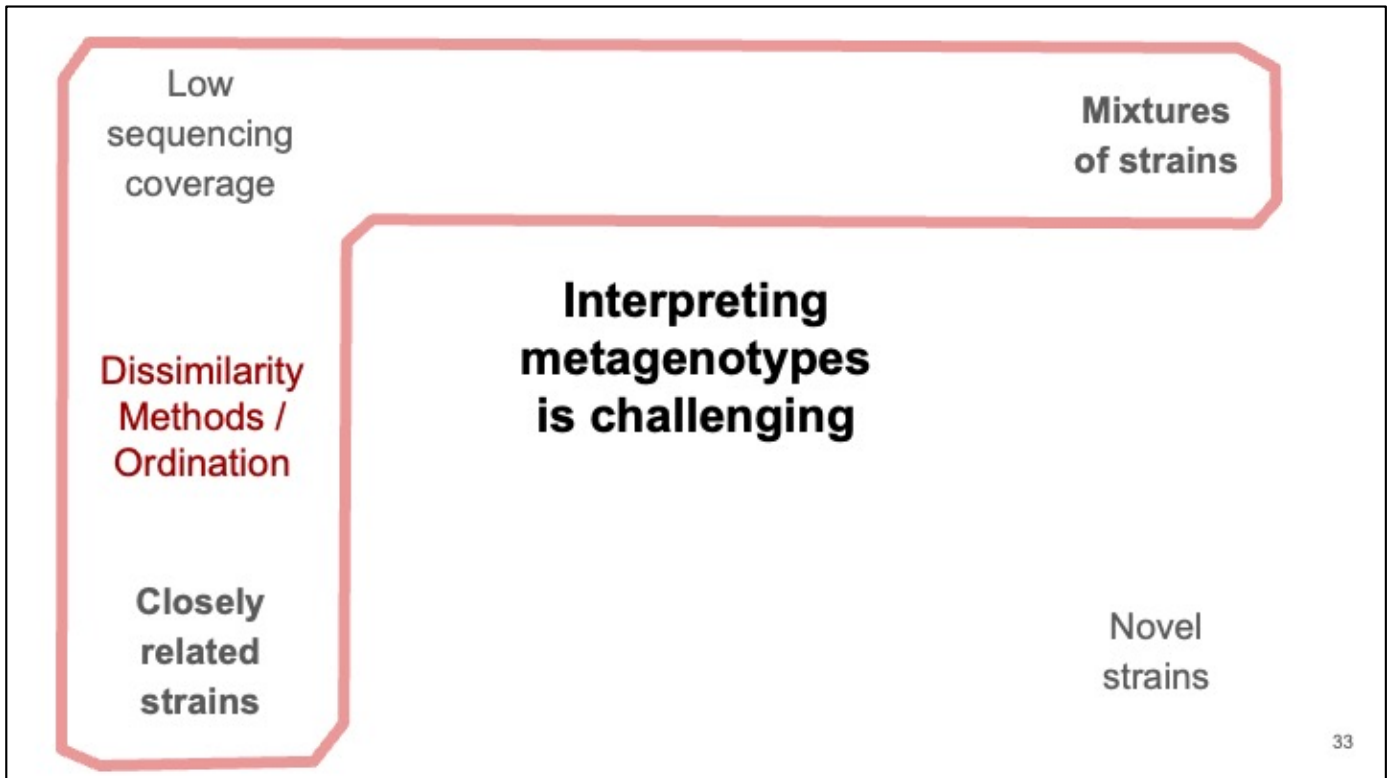
Novel
strains

31

1. As a result, naive approaches to strain-inference are each handicapped by some combination of these challenges.
2. For instance, taking the consensus allele at each position in the metagenotype from one sample will only recover a dominant strain
3. And even then might be sparse



1. Alternatively, using the observed SNVs as a “fingerprint” and matching to a database of known strains, fails when strain diversity has not already been well characterized.



1. One especially popular approach has been to consider the the dissimilarity between metagenotypes using, for example, the cosine distance.
2. A cutoff dissimilarity is chosen below which samples are considered to have the “same strain”
3. Unfortunately, this approach doesn’t have a principled way to differentiate between shared strains in a mixture and genotypically similar strains.

Outline

Intraspecific diversity in the microbiome

Strain inference

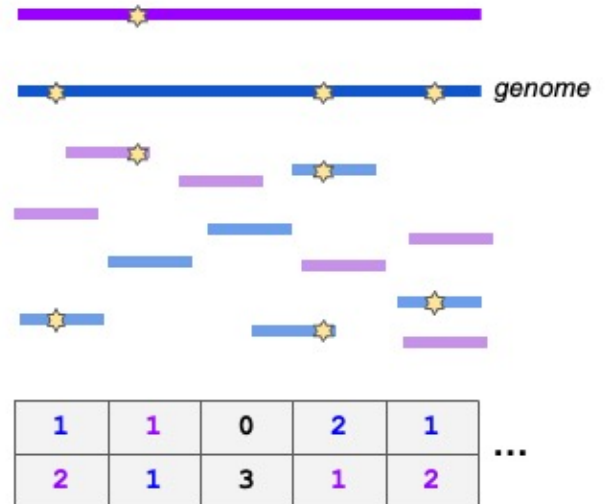
Metagenotype deconvolution

Application to large metagenome collections

34

1. So, if none of the other approaches is well suited to interpreting metagenotype data,
2. Let's talk about an approach to strain inference—which I'm calling strain deconvolution—that gracefully handles all four of these challenges

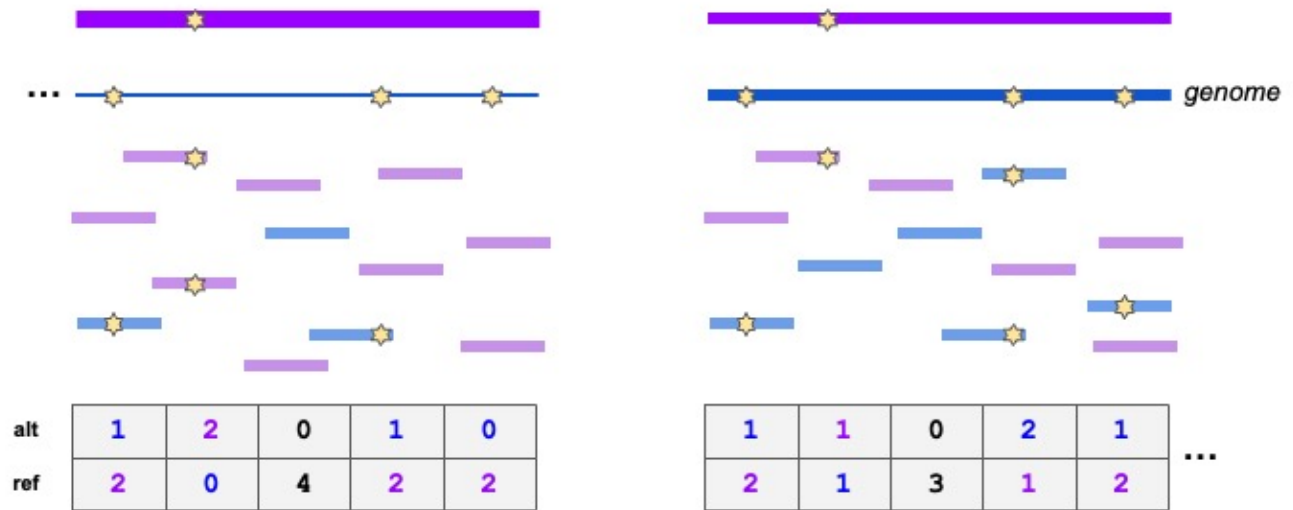
Strain deconvolution



35

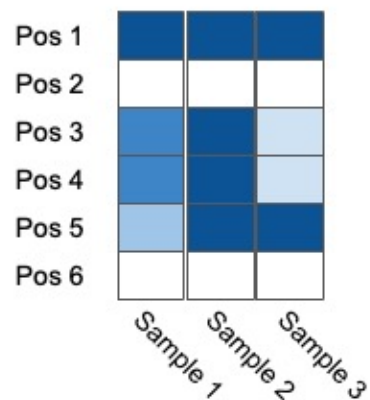
1. The key concept of the strain deconvolution approach is that we combine metagenotypes from multiple samples...

Strain deconvolution harnesses SNV covariance



1. To fill in sparse genotypes
2. And allowing us to use use covariance in the observed frequency of alleles at SNP positions
3. To disentangle strain mixtures

Non-negative matrix factorization for metagenotypes

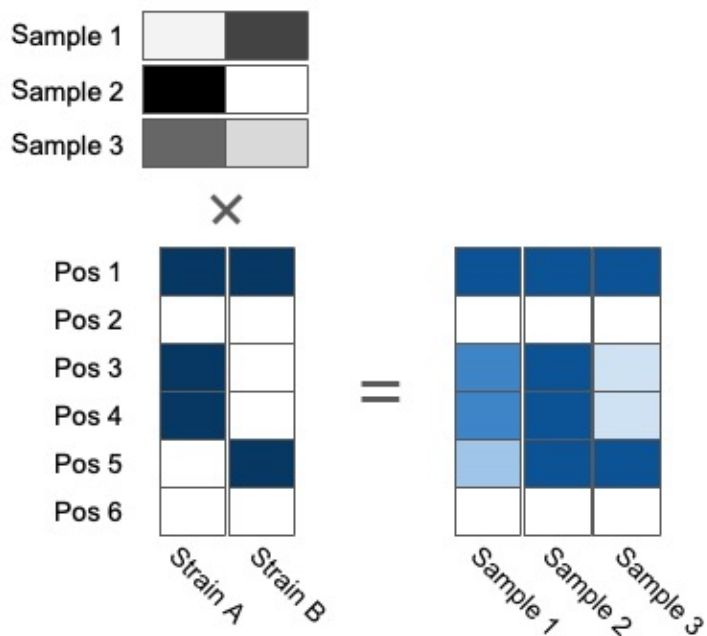


37

1. Strain deconvolution is analogous to non-negative matrix factorization (or NMF)
2. Here, multiple metagenotypes have been stacked into a matrix, where I'm depicting a higher frequency of the alternative allele with darker colors
3. (in this cartoonized version I'm showing just three samples, but we might consider many more)

Non-negative matrix factorization for metagenotypes

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



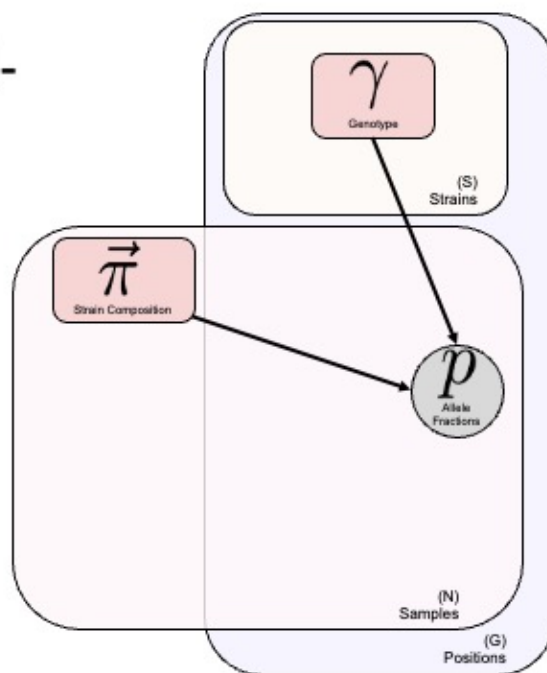
38

1. This metagenotype matrix can be decomposed into
2. a matrix of strain genotypes (here a binary matrix of genotypes with each column a strain)
3. And a matrix of strain relative abundances across each of the samples
4. We can model the metagenotype matrix as a linear combination of the strain genotypes, i.e. the matrix product of the relative abundance matrix and the genotype matrix.

Strain deconvolution as model-based inference

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

Except:



39

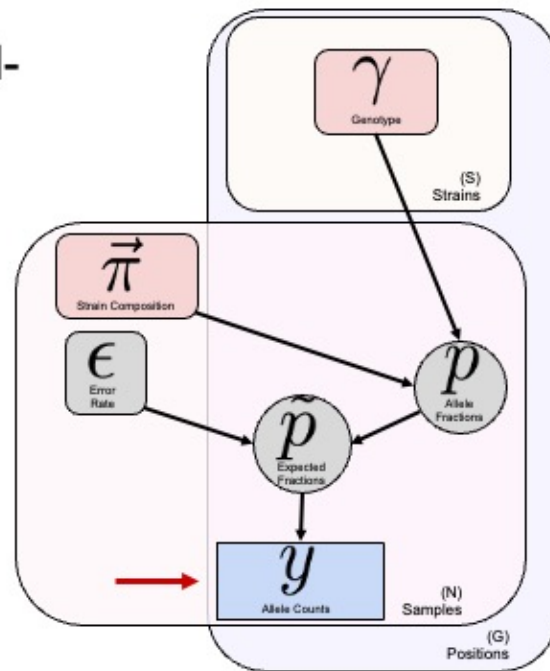
1. Alternatively, we can describe this deconvolution as a probabilistic graphic model
2. which I am showing here in an informal plate notation
3. (plates in the background indicate the indexing for each variable)
4. In this simplest model, “gamma”, the variable I’ll use throughout to refer to the strain genotypes
5. And “pi”, a vector of strain relative abundances
6. Combine (implicitly through matrix multiplication) to determine the allele frequencies: “p”
7. This model-based approach to deconvolution has a few key differences from canonical NMF...

Strain deconvolution as model-based inference

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

Except:

- Counts observed (sequencing error, binomial likelihood)



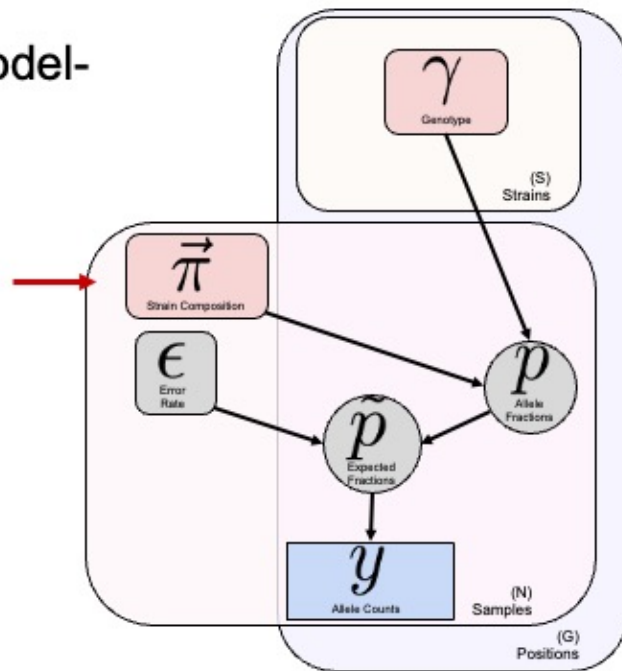
1. First, we can model the observed metagenotype counts, e.g. by explicitly modeling sequencing error and the discrete allele counts with a binomial likelihood

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



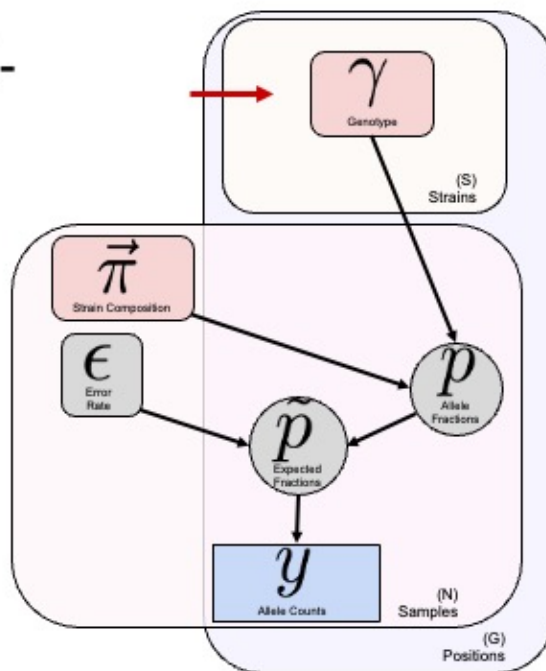
1. Second, unlike NMF where the only constraint is non-negativity, here we constrain the rows of "pi" to sum-to-1 (since they're relative abundances)

Strain deconvolution as model-based inference

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

Except:

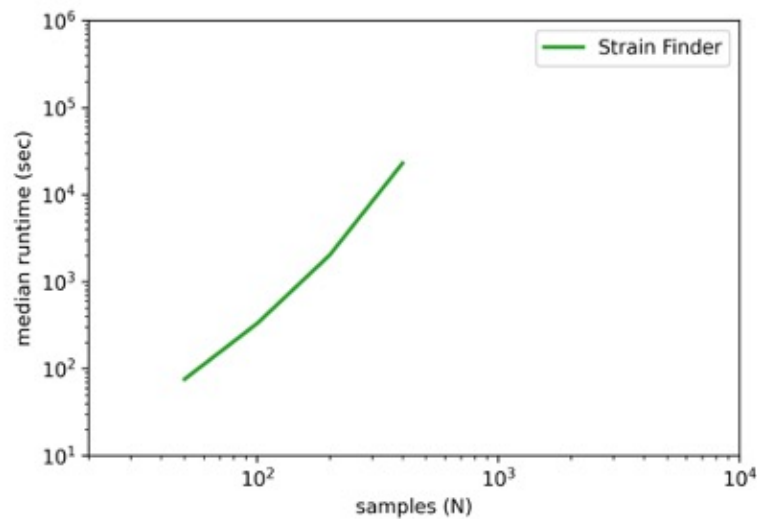
- Counts observed (sequencing error, binomial likelihood)
- Strain composition sums-to-1
- Genotypes at each position are binary, either 0 (reference) or 1 (alternative) allele



42

1. And third, “gamma” can be modeled as binary: either the reference allele (0) or the alternative allele (1).
2. We can then use constrained optimization (e.g. maximum likelihood) to obtain estimates of gamma and pi, the strain genotypes and their relative abundances across samples.
3. I am aware of two existing tools that already implement this approach.
4. Unfortunately, there’s a shortcoming: discrete optimization is HARD

Computational scalability of existing deconvolution tools



43

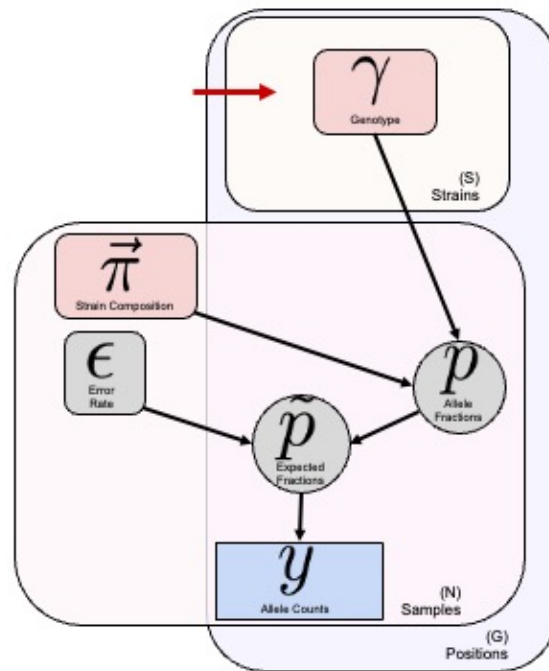
1. Here I'm showing the runtime of a tool called "Strain Finder", which uses expectation maximization approach to optimizing the parameters of the model.
2. What you see is that over a 1 order-of-magnitude increase in the number of samples and strains, Strain Finder takes almost 3 orders-of-magnitude more time to estimate parameters.
3. For 400 samples and 120 strains, this requires a median of 6.4 hours on simulated data
4. Increasing the size of our model, e.g. by an additional order of magnitude, is simply not feasible within a reasonable runtime

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



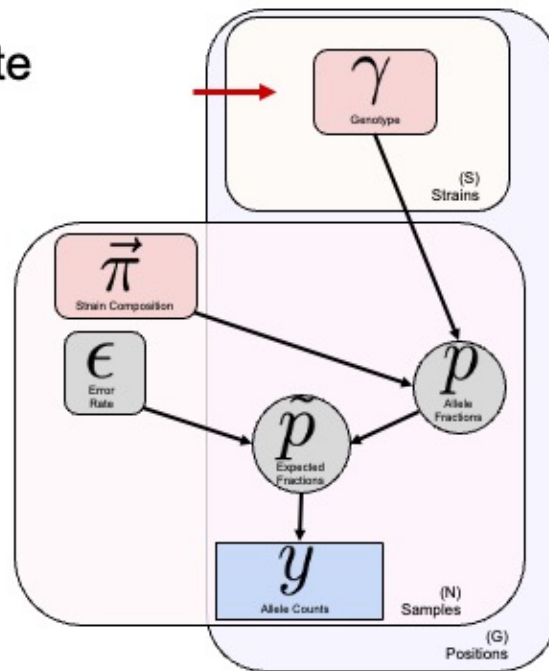
1. I believe that a major reason for this poor scaling is the discreteness of the genotypes, gamma
2. Strain Finder is limited by the computational scaling of its expectation maximization algorithm

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

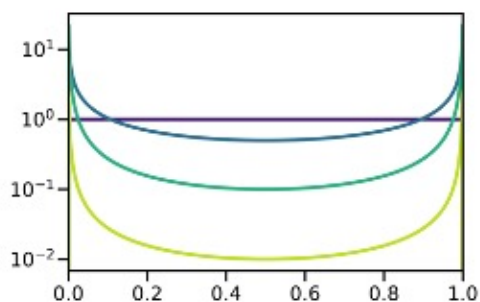


45

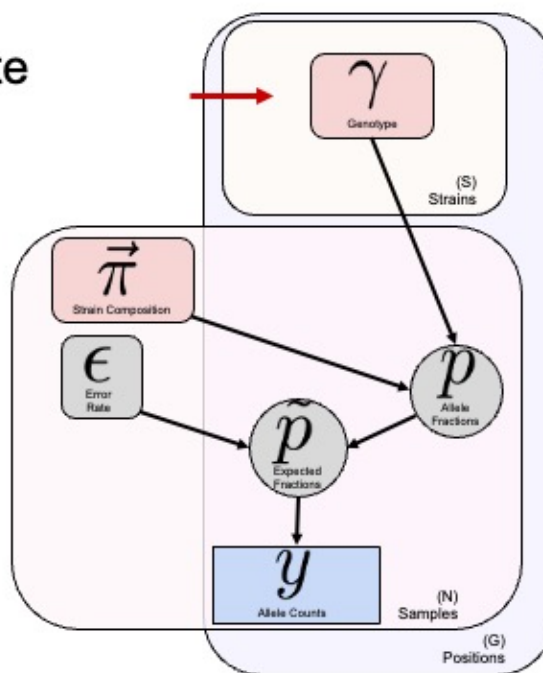
1. Which brings us to the key approach taken by my tool:
2. In StrainFacts, which stands for Strain Factorization...
3. We replace the binary genotype constraint with a fuzzy genotype: on the unit interval between 0 and 1.
4. This TRANSFORMS our model from discrete to fully differentiable, and allows us to apply efficient, gradient-based methods for parameter optimization.

StrainFacts relaxes the discrete allele constraint

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

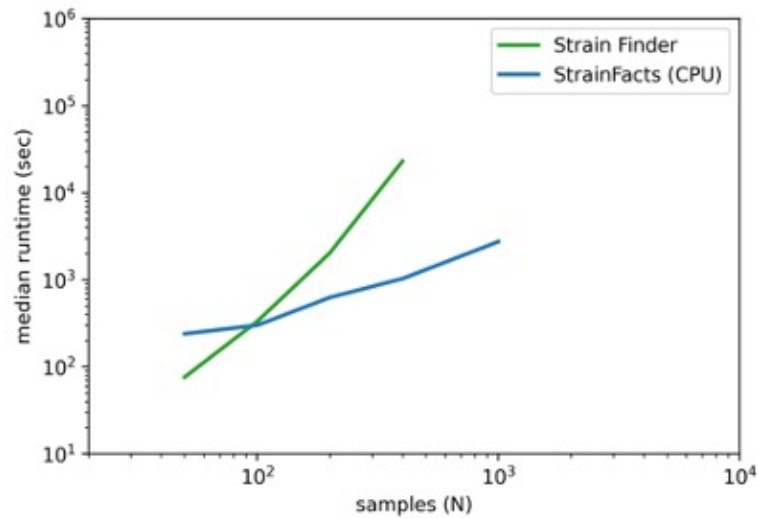


Shifted-scaled Dirichlet distribution (SSD; similar to the Dirichlet/Beta)



1. You might notice that fuzzy genotypes are not biologically realistic, since alleles are discrete.
2. To deal with this issue without losing differentiability, we take a regularization approach, putting a prior on gamma that keeps genotypes close to 0 or 1, despite their fuzziness
3. Specifically, we use the shift-scaled dirichlet distribution for this fuzzy approximation, tuning the scaling parameter to regularize our estimates towards binary values.

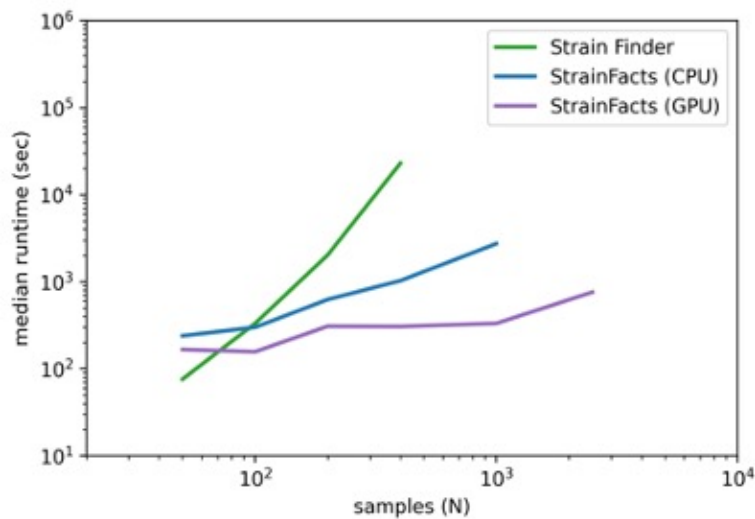
Fuzzy genotypes and gradient descent can scale to thousands of samples



47

1. And it works!
2. We get much better scaling with StrainFacts than Strain Finder, a nearly two order of magnitude decrease in runtime for larger models.

Fuzzy genotypes and gradient descent can scale to thousands of samples



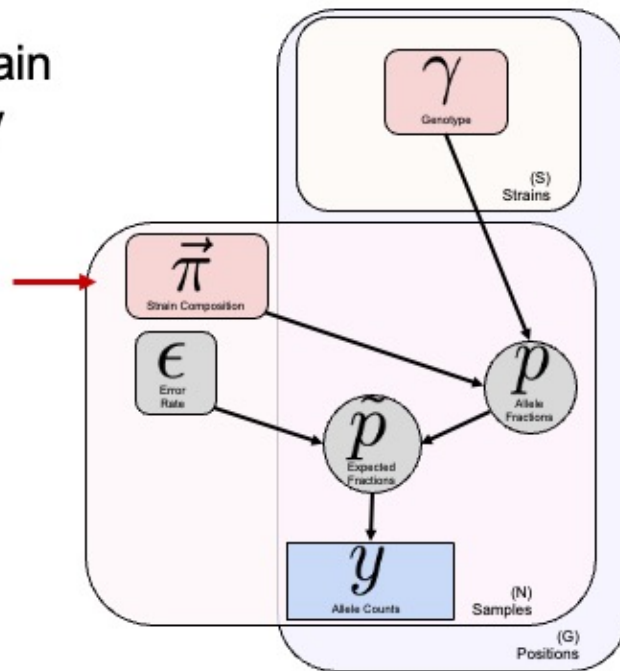
48

1. What's more, we can run StrainFacts on a GPU, further improving our runtime for large models
2. StrainFacts is therefore capable of fitting very large models with thousands or tens-of-thousands of samples and hundreds of strains

StrainFacts regularizes strain heterogeneity and diversity

Additionally, we put a hierarchical prior on π

- Strain heterogeneity regularization (strains per sample)



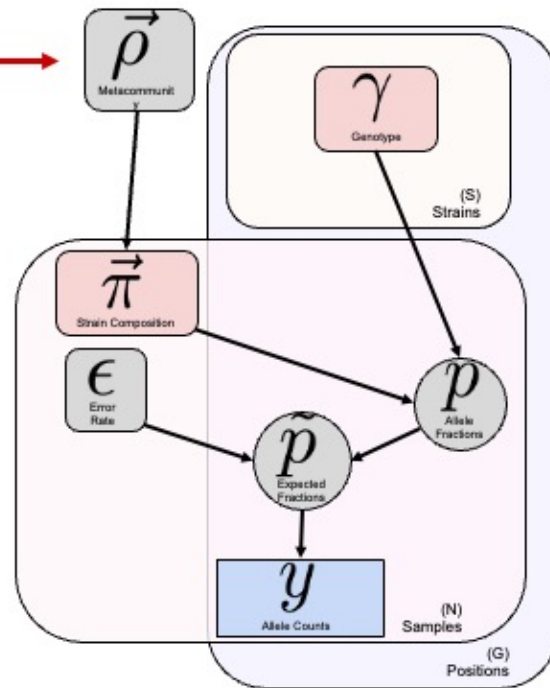
49

1. Besides these fuzzy genotypes, StrainFacts also applies regularization to strain relative abundances
2. We push estimates towards a smaller number of active strains in each 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



50

1. And, with a hierarchical prior, we also regularize the overall diversity towards fewer strains
2. This regularization reflects our preference for greater parsimony
3. And, since we don't need a secondary model selection step to choose a strain number,
4. The computational demands of our approach is further reduced

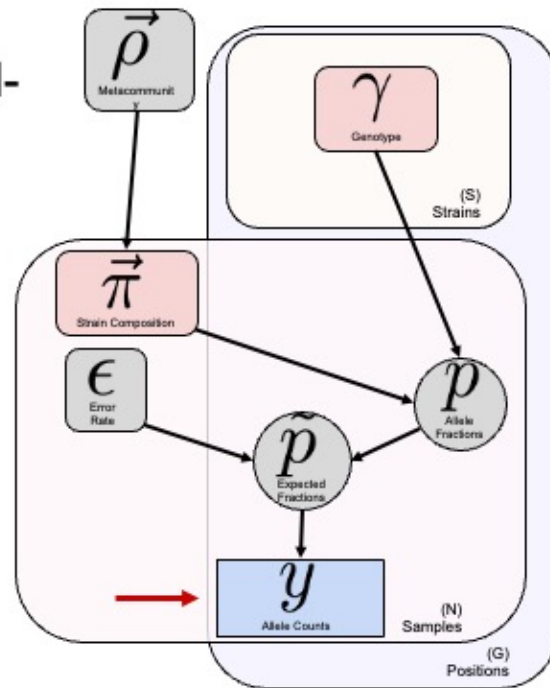
Strain deconvolution as model-based inference

Additionally, we put a hierarchical prior on π

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

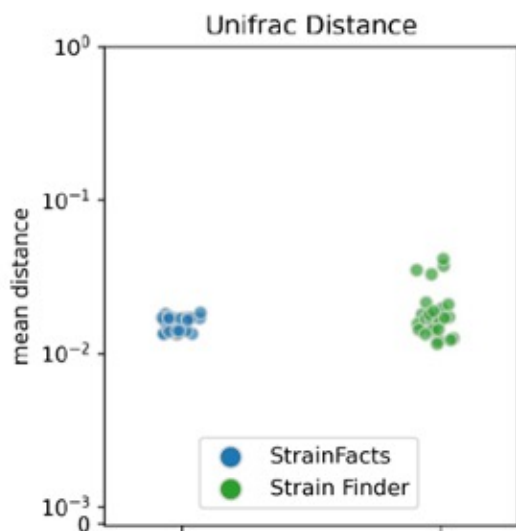
Full model also includes count over-dispersion (Beta-Binomial likelihood)

Maximum a posteriori (MAP) estimation for inference



1. Finally, to complete this description of the StrainFacts model
2. I'll add that we also use a Beta-Binomial likelihood, to model count overdispersion

StrainFacts is accurate



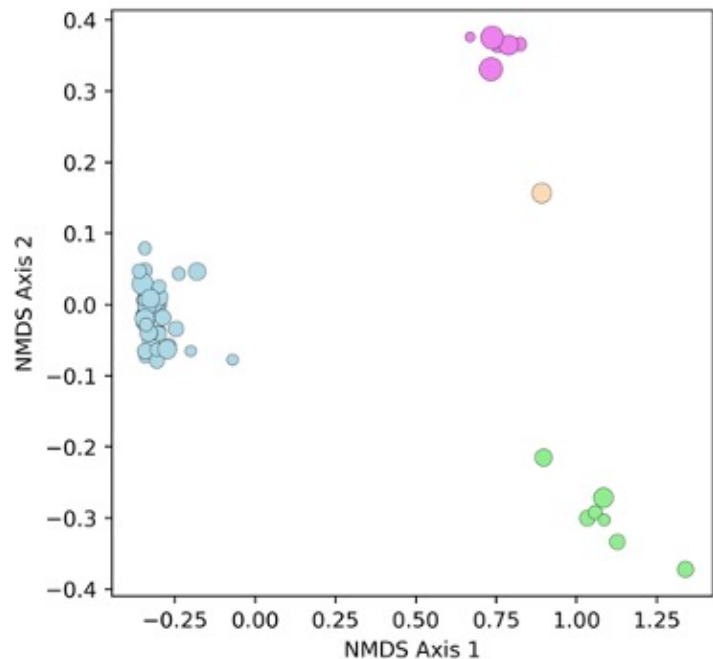
Based on simulations, StrainFacts is similarly accurate to an existing deconvolution tool

Several orders of magnitude faster

1. The accuracy of our inferences is, of course, important
2. We use simulations to compare our estimates to a ground truth
3. And find that StrainFacts estimates are comparable to those obtained by Strain Finder
4. (while being several orders of magnitude faster)

Single-cell genomics validates inferences in complex strain mixtures

Streptococcus thermophilus single-cell genotypes cluster into four groups



Xiangpeng Li

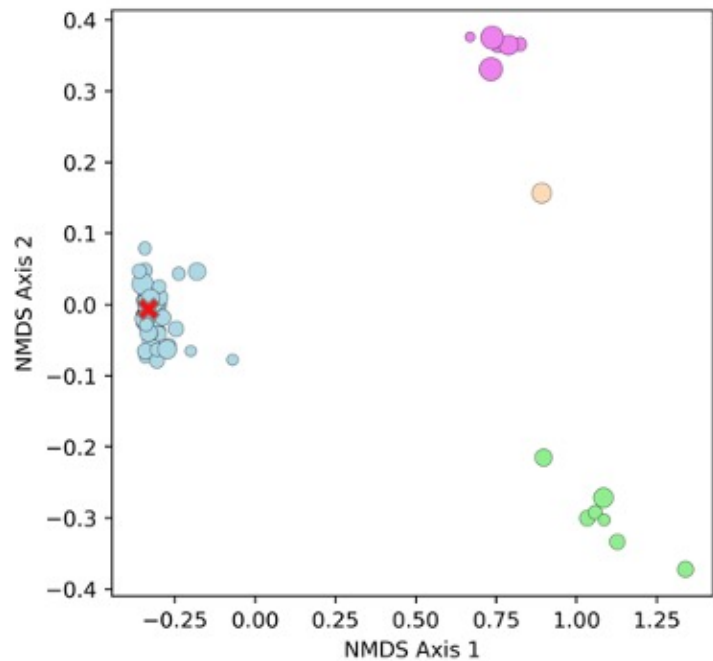
53

1. But simulations don't necessarily capture all of the features of real data
2. To validate our inferences we applied single-cell genomics to one sample from a larger study of ulcerative colitis patients
3. Single-cell genomics captures a sparse representation of the genotypes of individual cells in a sample, and therefore reflects individual strains
4. On this slide I'm showing an ordination of these single-cell genotypes for one species, *Streptococcus thermophilus*.
5. Closer points reflect more similar genotypes, based on the Hamming distance
6. What you can see is that in this sample, *S. thermophilus* has a remarkable amount of strain diversity: with genotypes clustering into four distinct types, here highlighted by colors.

Single-cell genomics validates inferences in complex strain mixtures

Streptococcus thermophilus single-
cell genotypes cluster into four groups

Consensus metagenotype only
reflects dominant strain



Xiangpeng Li

54

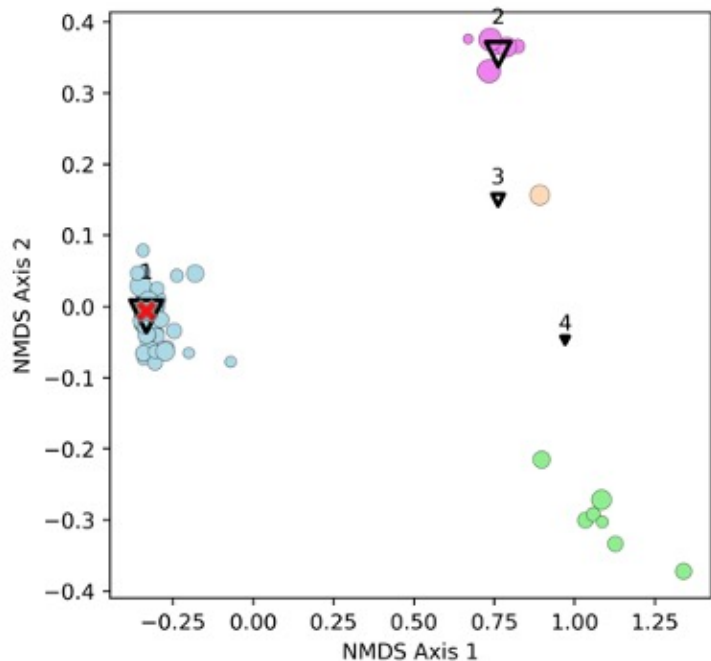
1. Adding the consensus metagenotype to this plot (red “x”), you can see that the majority-vote genotype reflects the majority strain, while completely missing the three other clusters.

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



Xiangpeng Li

55

1. On the other hand, StrainFacts correctly infers that there are four distinct strains, here shown as black triangles
2. For three of the four inferred strains, these inferred strains match the single-cell genomes closely, suggesting that StrainFacts is capable of accurate inference in real biological data even for species with a high degree of strain heterogeneity.
3. And, while for the fourth cluster the similarity is not so clear, this strain was estimated to be at the lowest relative abundance, suggesting that there may not have been enough information to infer this genotype.

Outline

Intraspecific diversity in the microbiome

Strain inference

Metagenotype deconvolution

Application to large metagenome collections

56

1. For the remainder of this talk I want to show you some of what we found when we applied StrainFacts to real data

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

57

1. Because of rapid metagenotyping thanks to GT-Pro, we now have access to metagenotypes for tens of thousands of metagenomic samples across numerous independent human microbiome studies.

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

Two examples:

- Biogeography
- Population Genetics

58

1. I'd like to share two vignettes that demonstrate the value of StrainFacts for:
 - a. Understanding microbial biogeography
 - b. Studying microbial population genetics

Strain biogeography

Agathobacter rectalis is a prevalent and abundant gut bacterium

20 countries

→ 33 studies

→ 9,224 stool samples

→ 198 inferred strains

59

1. *Agathobacter rectalis* (previously known as *Eubacterium rectale*) is a prevalent and abundant member of the human gut microbiome.
2. Across dozens of studies, about 10,000 samples had sufficient coverage of *A. rectalis* SNPs to deconvolve strains.
3. This resulted in 198 inferred strains
4. Access to sample metadata for many of the available metagenomic samples means that we can also ask about the association between strains and the country where that sample was collected

Strain biogeography

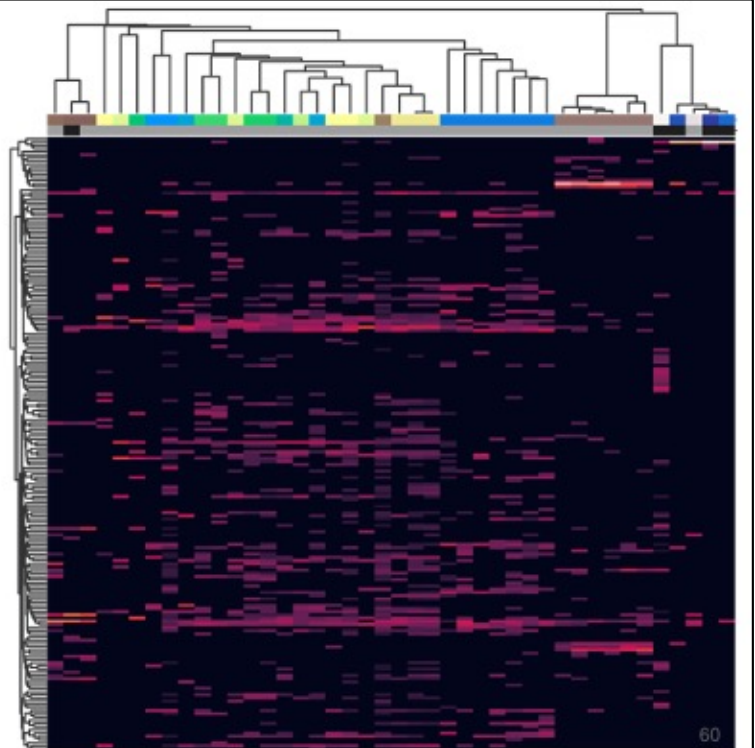
Agathobacter rectalis is a prevalent and abundant gut bacterium

20 countries

→ 33 studies

→ 9,224 stool samples

→ 198 inferred strains



1. This heatmap summarizes the distribution of the 198 inferred strains (rows) across 33 studies (columns) with a minimum of 10 human stool metagenomes.
2. Each column summarizes a single study, with the brighter colors reflecting a larger fraction of samples dominated by each strain (row)
3. Columns are sorted based on similarities in this dominance profile
4. You'll also see that columns have been colored by the study *country*
5. You also might notice that these study profiles seem to cluster:
 - a. While several strains are found across many studies,
 - b. Some are much more prevalent in a subset of these

Strain biogeography

Agathobacter rectalis is a prevalent and abundant gut bacterium

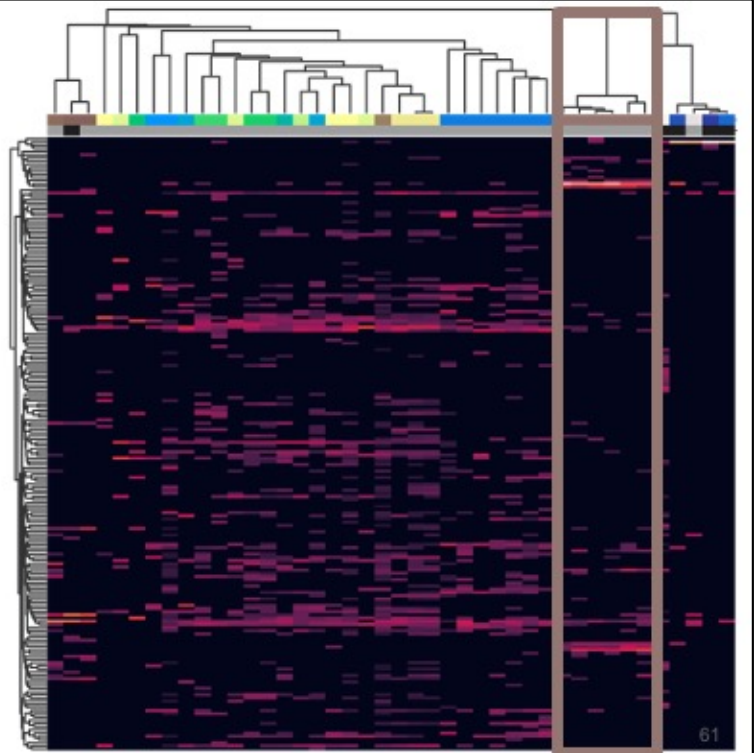
20 countries

→ 33 studies

→ 9,224 stool samples

→ 198 inferred strains

Strains reflect which country samples were collected in across independent studies



1. Since these colors can be a bit hard to parse, here I've highlighting all six studies performed in China
2. You'll see that these cluster closely together, and are characterized by a couple of distinct strain groups
3. This is consistent with previous reports of a subspecies of *A. rectalis* highly enriched in Chinese metagenomes.

Strain biogeography

Agathobacter rectalis is a prevalent and abundant gut bacterium

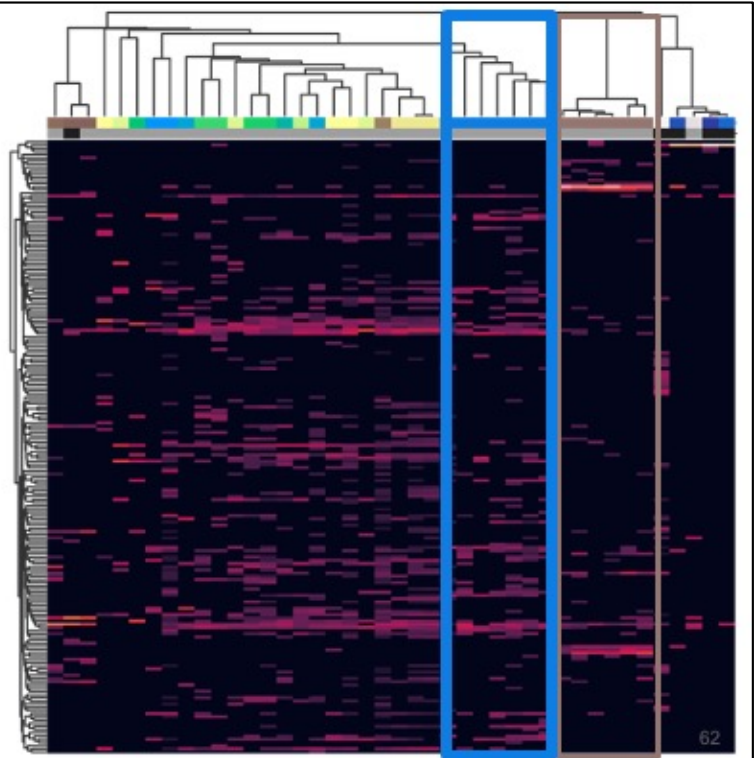
20 countries

→ 33 studies

→ 9,224 stool samples

→ 198 inferred strains

Strains reflect which country samples were collected in across independent studies



1. But China is not the only country where independent studies cluster together.
2. This box highlights 7 studies performed in the US
3. While these studies are not as distinctly/visibly different as the studies from China, they still cluster together
4. What I think is particularly exciting about this is that each column represents a completely INDEPENDENT study of human stool samples, with different subjects and protocols
5. But they nonetheless clearly reflect the geographic origin based on the dominant *A. rectalis* strains

Strain biogeography

Agathobacter rectalis is a prevalent and abundant gut bacterium

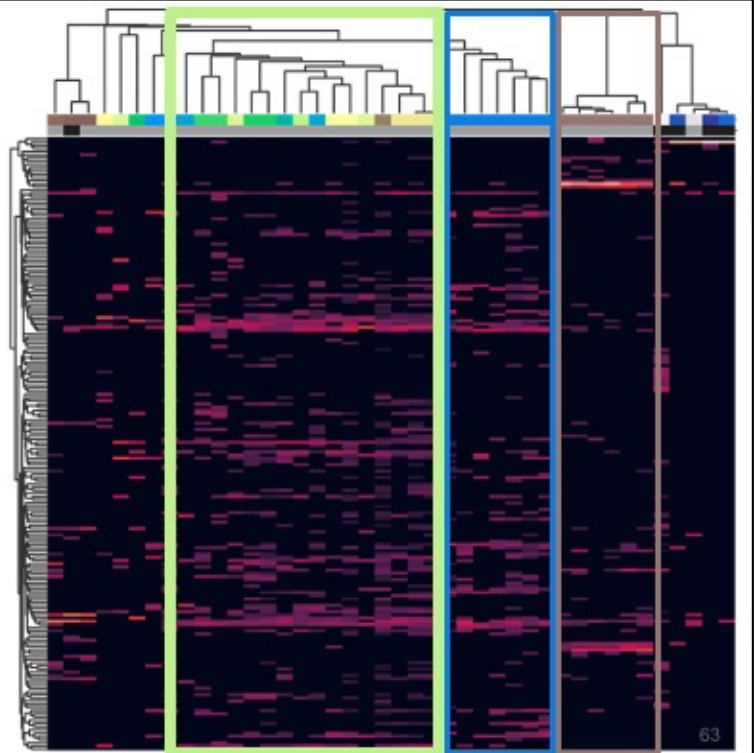
20 countries

→ 33 studies

→ 9,224 stool samples

→ 198 inferred strains

Strains reflect which country samples were collected in across independent studies

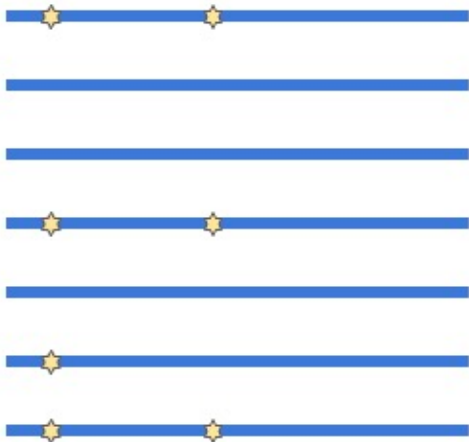


1. And, to a less dramatic extent, this geographic clustering also extends to many of the remaining studies from across Europe and Canada
2. (Although this clustering is not perfect)
3. I'm really excited by the potential for tracking strains across human populations to inform our understanding of how microbes are transmitted between individuals globally.
4. Not to mention identifying associations with human physiology, diet, and disease

Population genetics

1. Finally, I want to talk about the potential for strain inferences to inform our understanding of microbial population structure and evolution

Population genetics



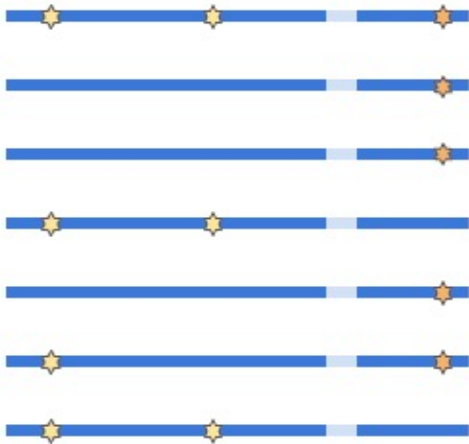
Linkage disequilibrium

Correlations between SNPs reflect shared descent

In asexual organisms, SNPs stay correlated

1. But first, a quick refresher on linkage disequilibrium
2. When we look across polymorphic sites in the genome, we see that we can use the variants that we see at one position to predict the variants at other positions
3. This correlation between alleles at SNP sites is called linkage, and reflect the shared inheritance of the sites together, as a unit
4. In perfectly asexual organisms, this linkage is only broken by subsequent mutations at one of the sites.

Population genetics



Linkage disequilibrium

Correlations between SNPs reflect shared descent

In asexual organisms, SNPs stay correlated

SNPs that are *not* correlated reflect recombination

1. However, this correlation can also be broken due to recombination.

Population genetics



Linkage disequilibrium

Correlations between SNPs reflect shared descent

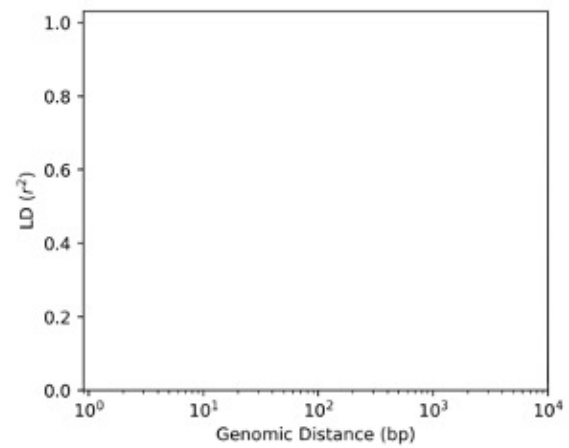
In asexual organisms, SNPs stay correlated

SNPs that are *not* correlated reflect recombination

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

1. And, since this is a spatial process, SNPs that are further apart are more likely to be separated by recombination

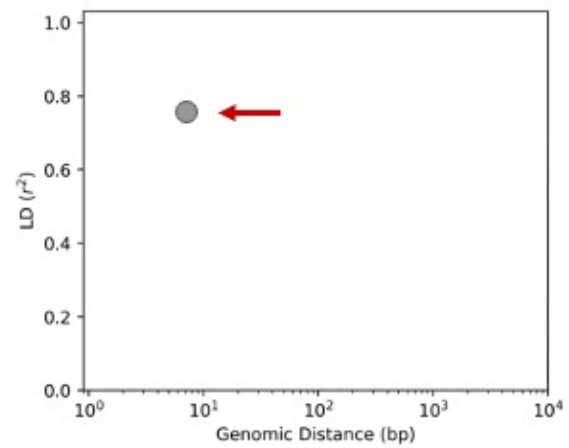
Population genetics



68

1. As a result, we can use the relationship between genomic distance (here on the x-axis) and linkage disequilibrium (on the y) to understand recombination in microbes

Population genetics

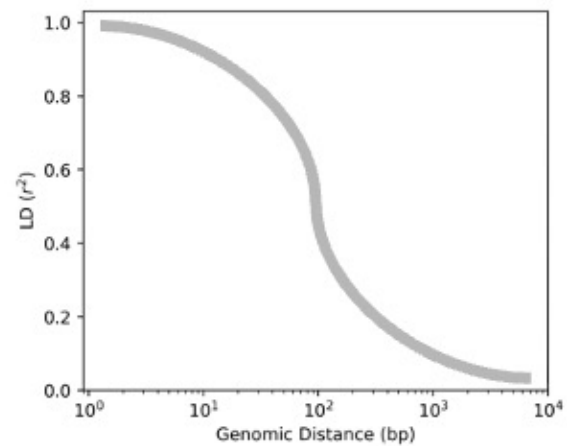


69

1. LD is a property of pairs of SNPs
2. and we can compare the pairwise LD and distance for *all* pairs of SNPs

Population genetics

Decay of LD for pairs at larger distances reflects recombination



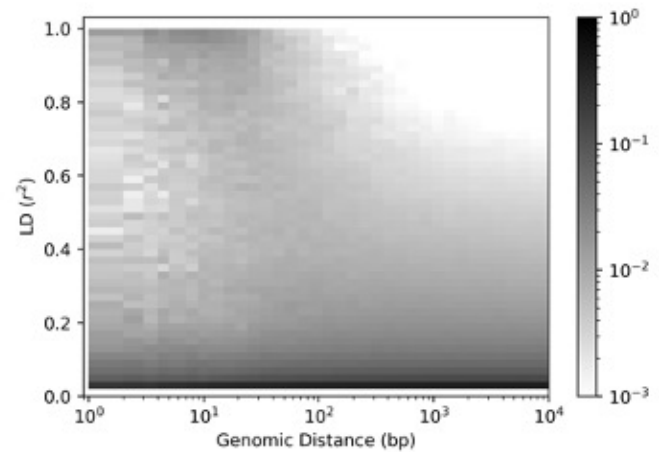
70

1. The expectation is that, in the presence of recombination, LD will drop quickly with distance.
2. And this LD decay is one way to detect recombination in microbes.

Population genetics

Decay of LD for pairs at larger distances reflects recombination

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



71

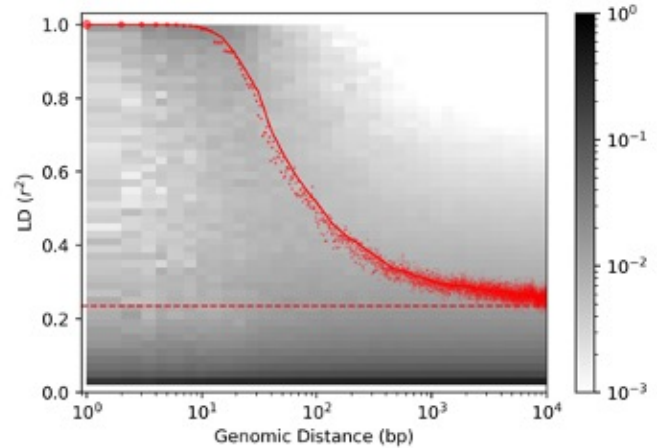
1. For Escherechia coli, this is exactly what we see.
2. Here I'm showing a two dimensional histogram, where each column reflects the distribution of pairs at that distance
3. And the darker shades indicate more pairs with that LD

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



72

1. And here I've plotted the 90th-percentile LD for SNP pairs at a range of distances in the genome
2. You can see that LD starts very high for neighboring SNPs and then quickly decays, before leveling out at larger distances
3. This suggests both that recombination has occurred frequently in the *E. coli* population
4. And that we're still seeing some population structure (i.e. *E. coli* is not a panmictic species)

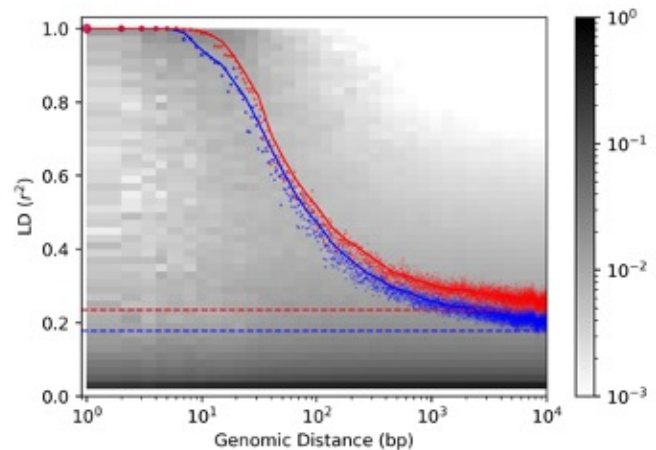
Population genetics

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



73

1. What's more, because *E. coli* is *very* well-studied, we can confirm this finding using reference strains
2. which I've now added to this plot as a blue profile.
3. You can see that these results are both qualitatively and quantitatively very similar to the de novo estimates

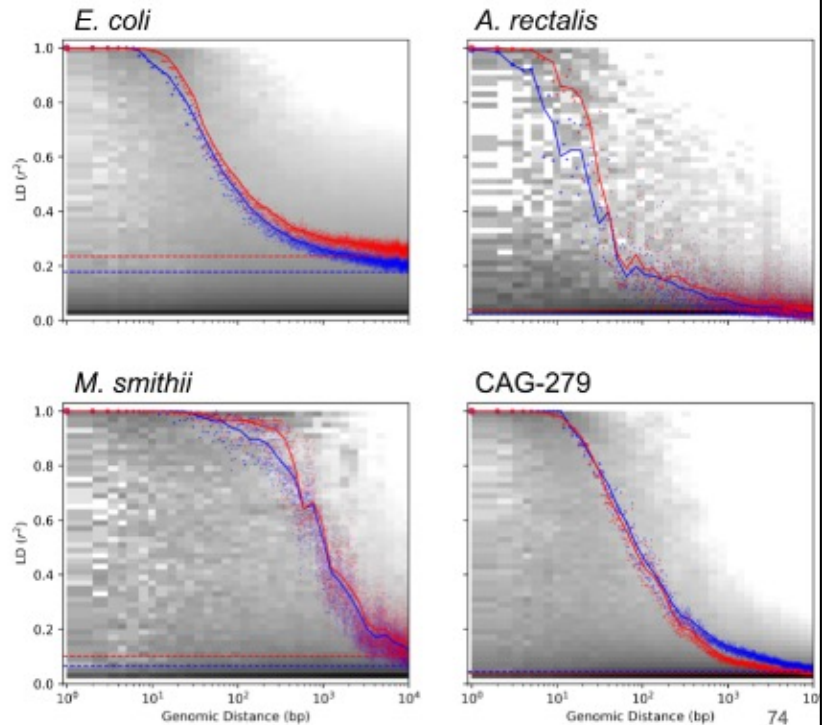
Population genetics

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



1. And this result isn't limited to *E. coli*
2. Here I've extend this analysis to *A. rectalis*, as well as *M. smithii*, the most abundant archaeon in the human gut, as well as CAG-279, a unnamed and never-isolated species of bacterium.
3. We observe distinctly different LD-decay profiles for each
 - a. Different decay rates, and different long-distance LD
4. Comparative population genetics is potentially a valuable tool for understanding population structure and recombination across species

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

75

1. So, in summary, deconvolution leverages metagenotypes to quantify strain abundance and reconstruct genotypes
2. StrainFacts scales this approach to datasets the size of publicly available data from many studies
3. We were able to validate our tool using simulations and single-cell genomics
4. And we find that it has biologically interesting applications across a number of fields

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

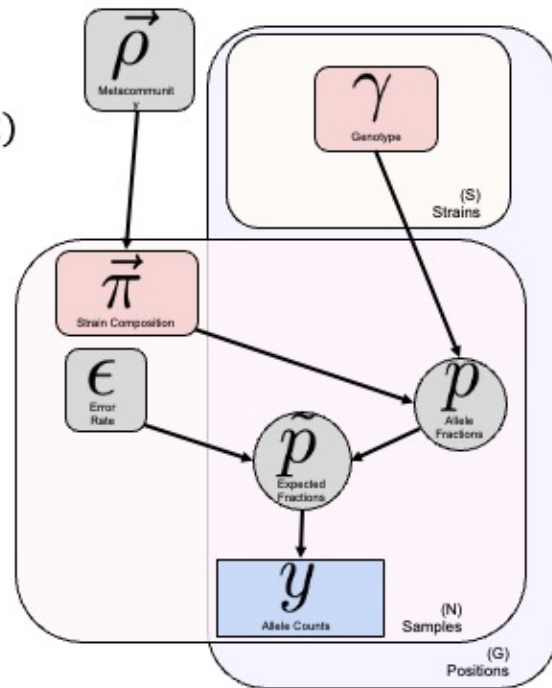
Questions?

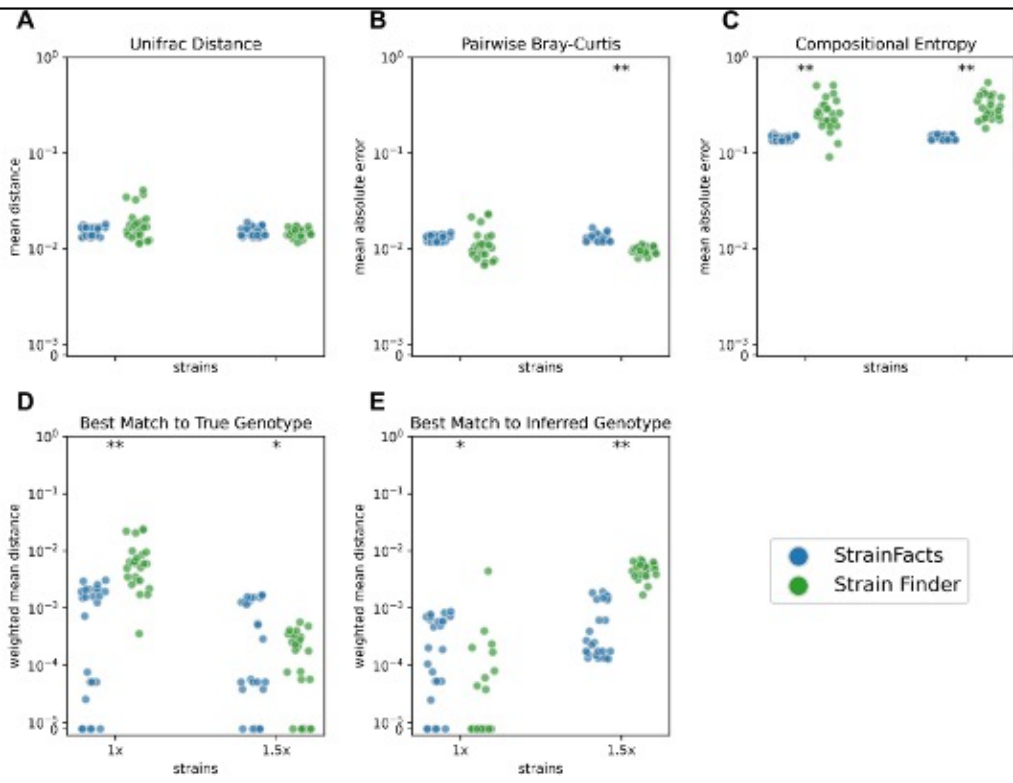
76

1. Thank you very much for your attention and I'd love to take any questions

Pocket Slides

$$\begin{aligned}
 y_{ig} &\sim \text{BetaBinom}(\tilde{p}_{ig}, \alpha^* \mid m_{ig}) \\
 \tilde{p}_{ig} &= p_{ig}(1 - \epsilon_i/2) + (1 - p_{ig})(\epsilon_i/2) \\
 p_{ig} &= \sum_s \pi_{is} \gamma_{sg} \\
 \gamma_{sg} &\sim \text{SSD}_0\left(\mathbf{1}, \mathbf{1}, \frac{1}{\gamma^*}\right) \\
 \tilde{\pi}_i &\sim \text{SSD}\left(\mathbf{1}, \vec{\rho}, \frac{1}{\pi^*}\right) \\
 \vec{\rho} &\sim \text{SSD}\left(\mathbf{1}, \mathbf{1}, \frac{1}{\rho^*}\right) \\
 \epsilon &\sim \text{Beta}\left(\epsilon_a^*, \frac{\epsilon_a^*}{\epsilon_b^*}\right)
 \end{aligned}$$



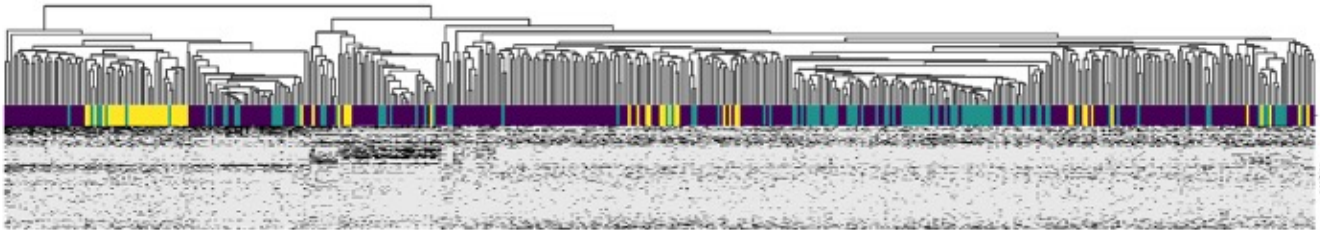


Cataloging strain diversity

■ 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