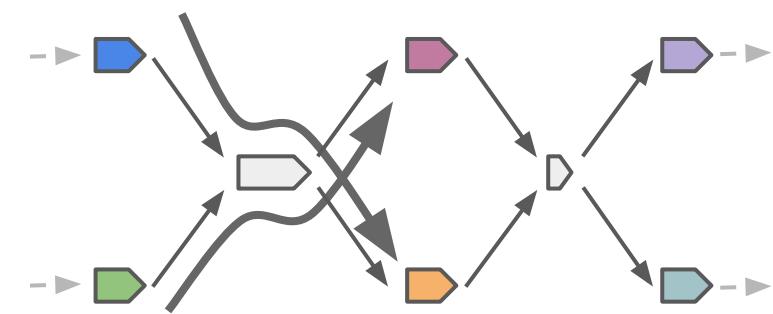


Unzipping the metagenome: strain-level discovery in the gut microbiome

Byron J. Smith

Bhatt Lab Computational Subgroup
2024-09-10



$$\begin{array}{c} \text{Legend: } \\ \text{Blue arrow: } \begin{matrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{matrix} \\ \text{Green arrow: } \begin{matrix} p_{1,1} & p_{1,2} & p_{1,3} \\ p_{2,1} & p_{2,2} & p_{2,3} \\ p_{3,1} & p_{3,2} & p_{3,3} \\ p_{4,1} & p_{4,2} & p_{4,3} \end{matrix} \\ \times \\ \approx \\ \begin{matrix} e_{1,1} & e_{1,2} & e_{1,3} \\ e_{2,1} & e_{2,2} & e_{2,3} \\ e_{3,1} & e_{3,2} & e_{3,3} \\ e_{4,1} & e_{4,2} & e_{4,3} \end{matrix} \end{array}$$

First Thing: Thank You!

Pollard Lab

Katie Pollard
Veronika Dubinkina
and everyone

Collaborators

Archit Verma
Dylan Cable

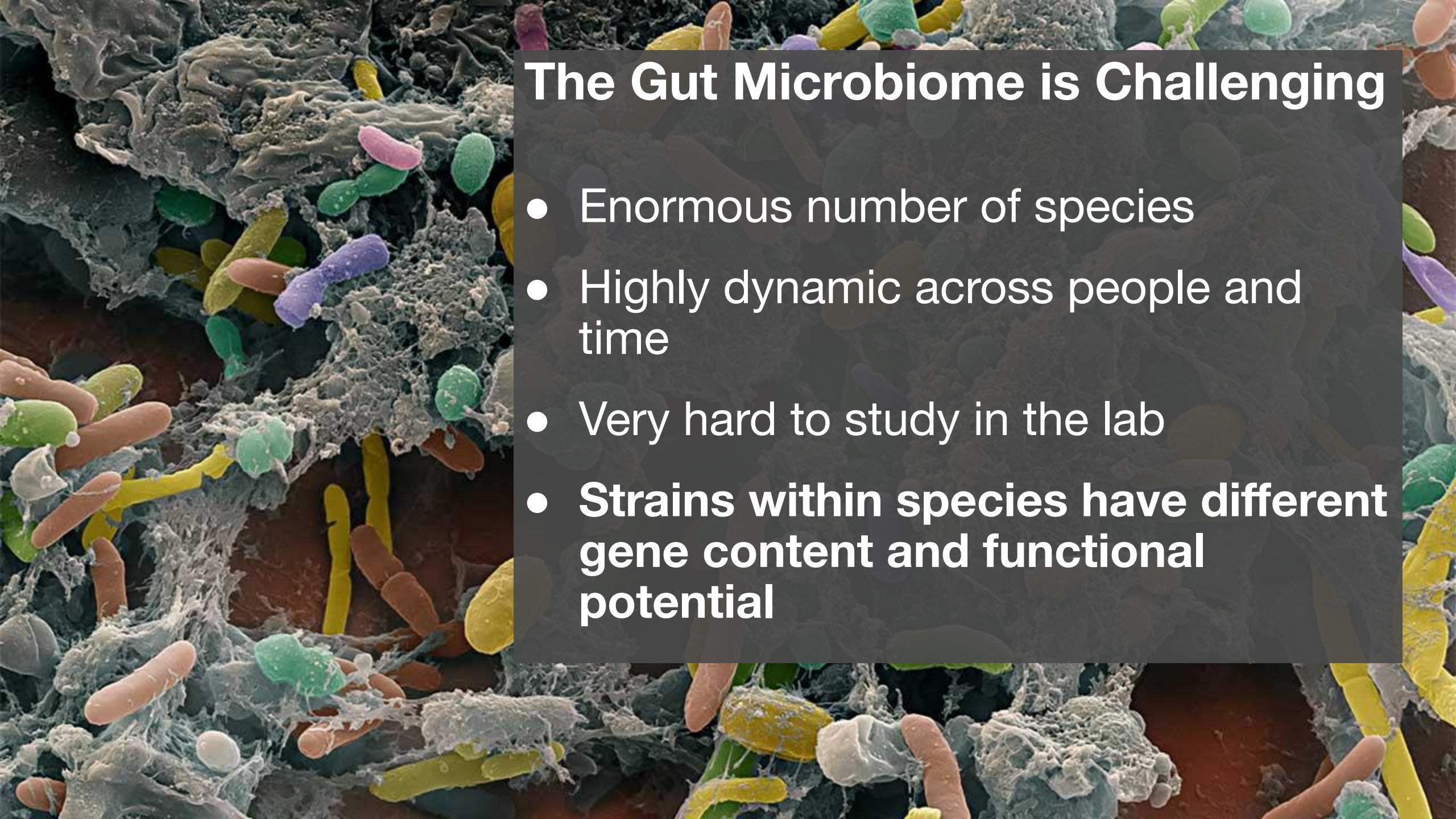
Funders

Gladstone Institutes
NIH
CZ Biohub
UC Noyce Initiative
Helmsley Charitable Trust



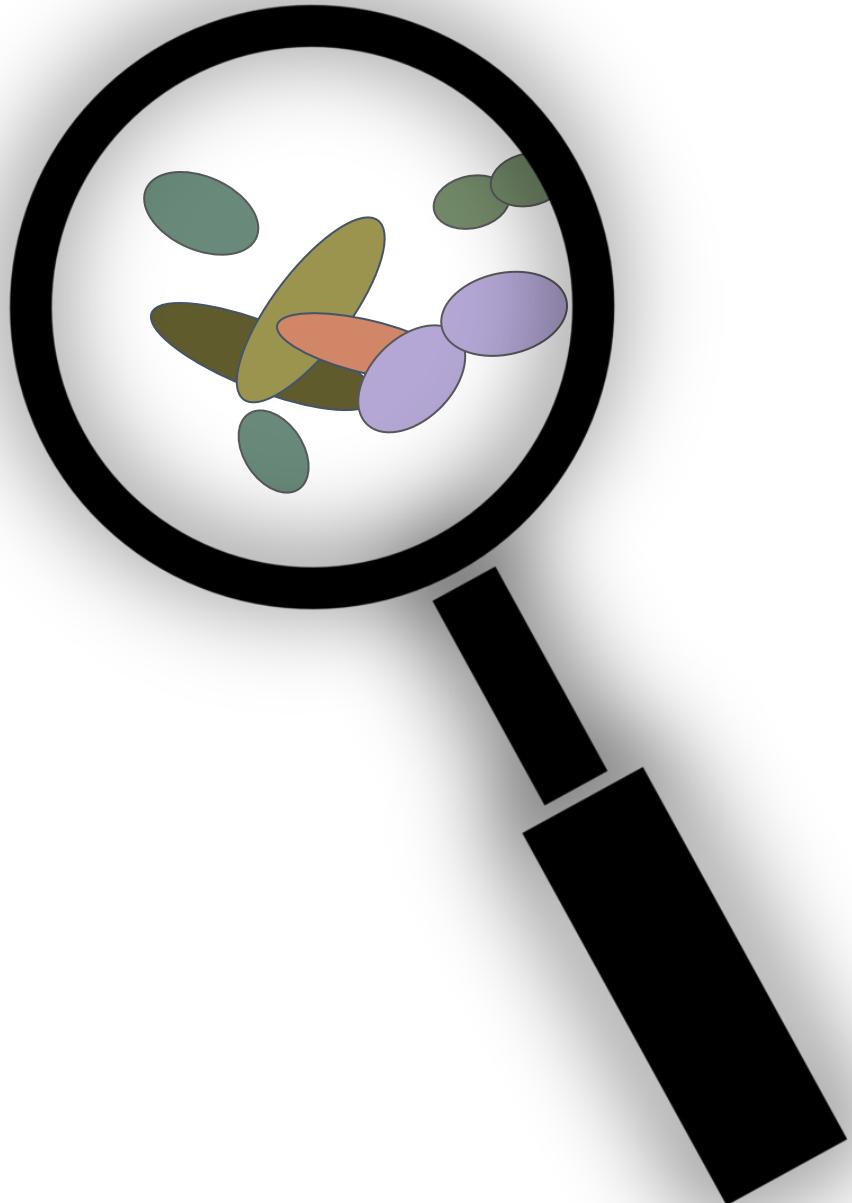
Introduction:

The gut microbiome and
shotgun metagenomics

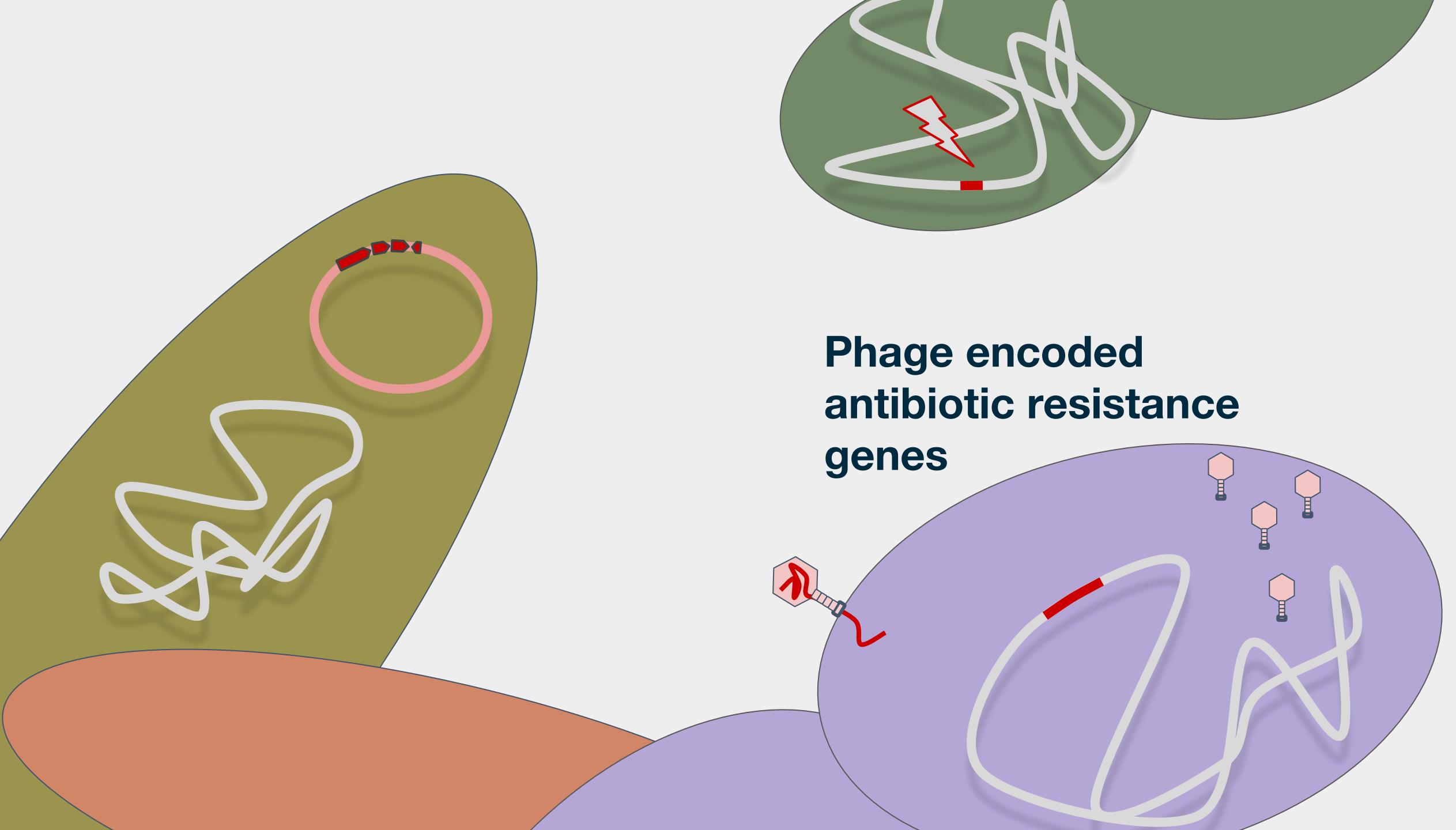


The Gut Microbiome is Challenging

- Enormous number of species
- Highly dynamic across people and time
- Very hard to study in the lab
- **Strains within species have different gene content and functional potential**

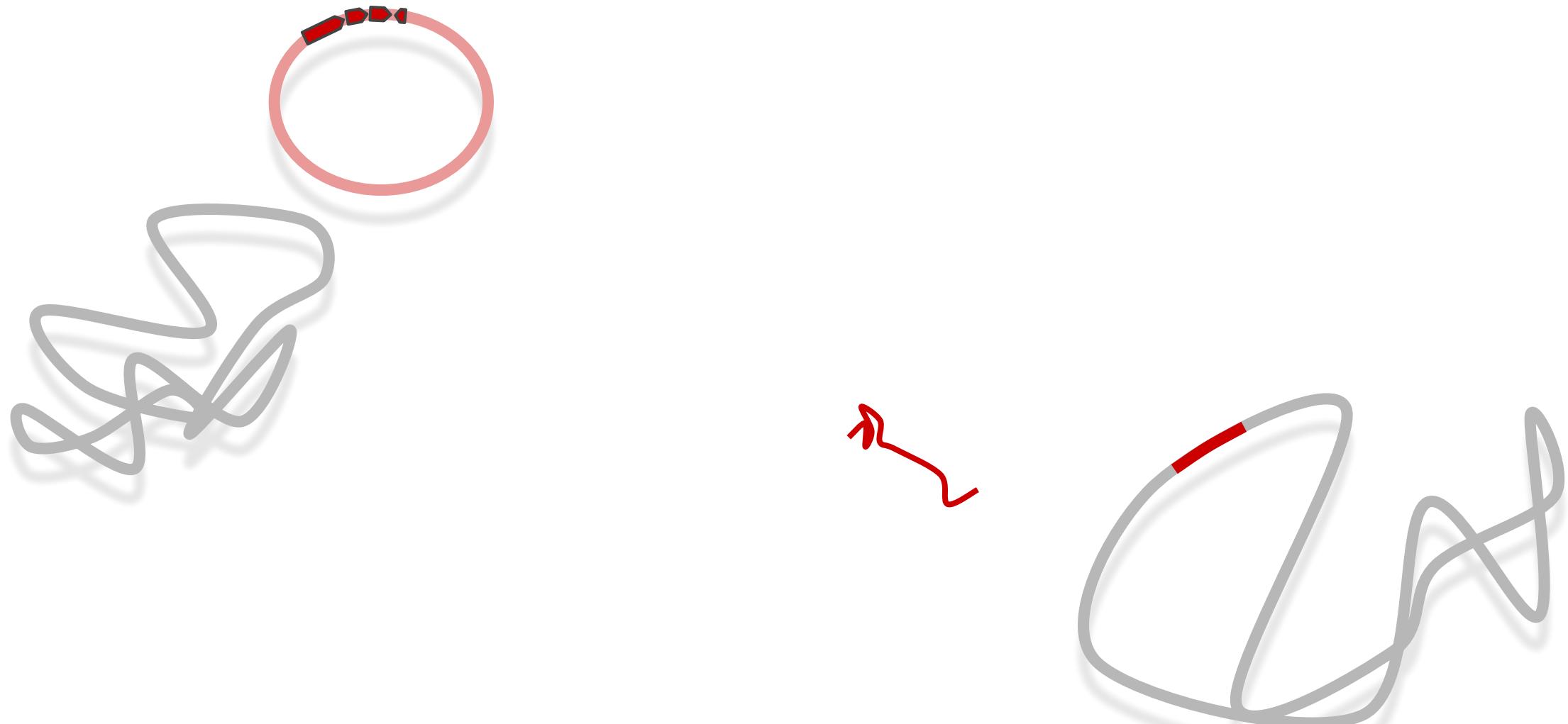


Bacterial genomes are
key to understanding
strain diversity



**Phage encoded
antibiotic resistance
genes**

Metagenomic sequencing surveys all genomes



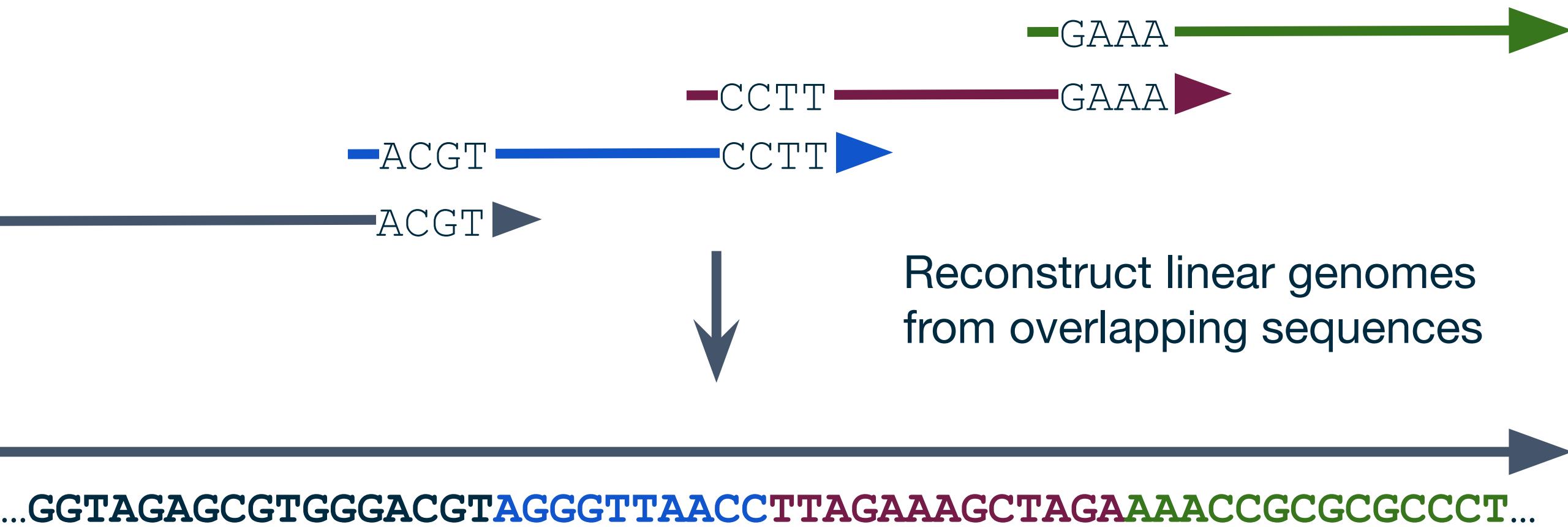
Short-read, shotgun metagenomes enable modern microbiome science

Requirements:

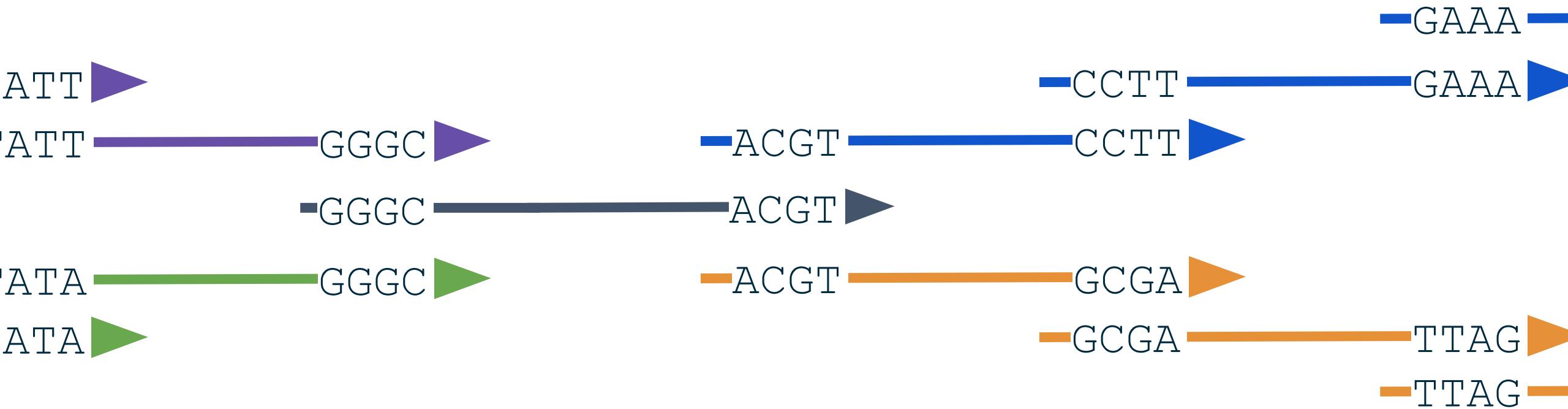
- strain-resolved genome sequences ➤ high accuracy
- capture low-abundance organisms ➤ very deep sequencing
- longitudinal designs and lots of samples ➤ cheap
- long sequences ➤ ...



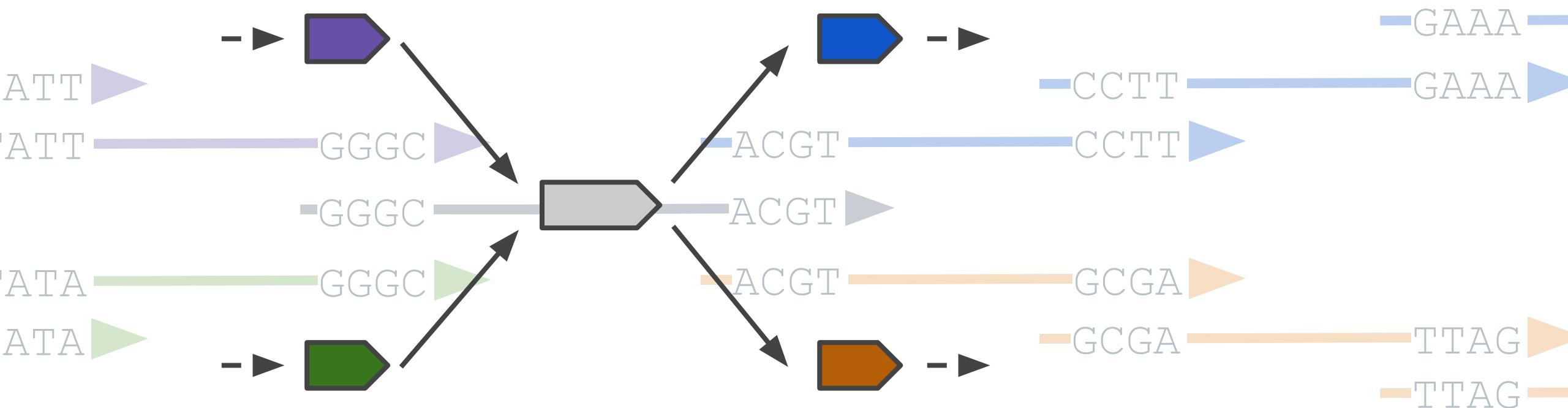
Turning short reads into long sequences



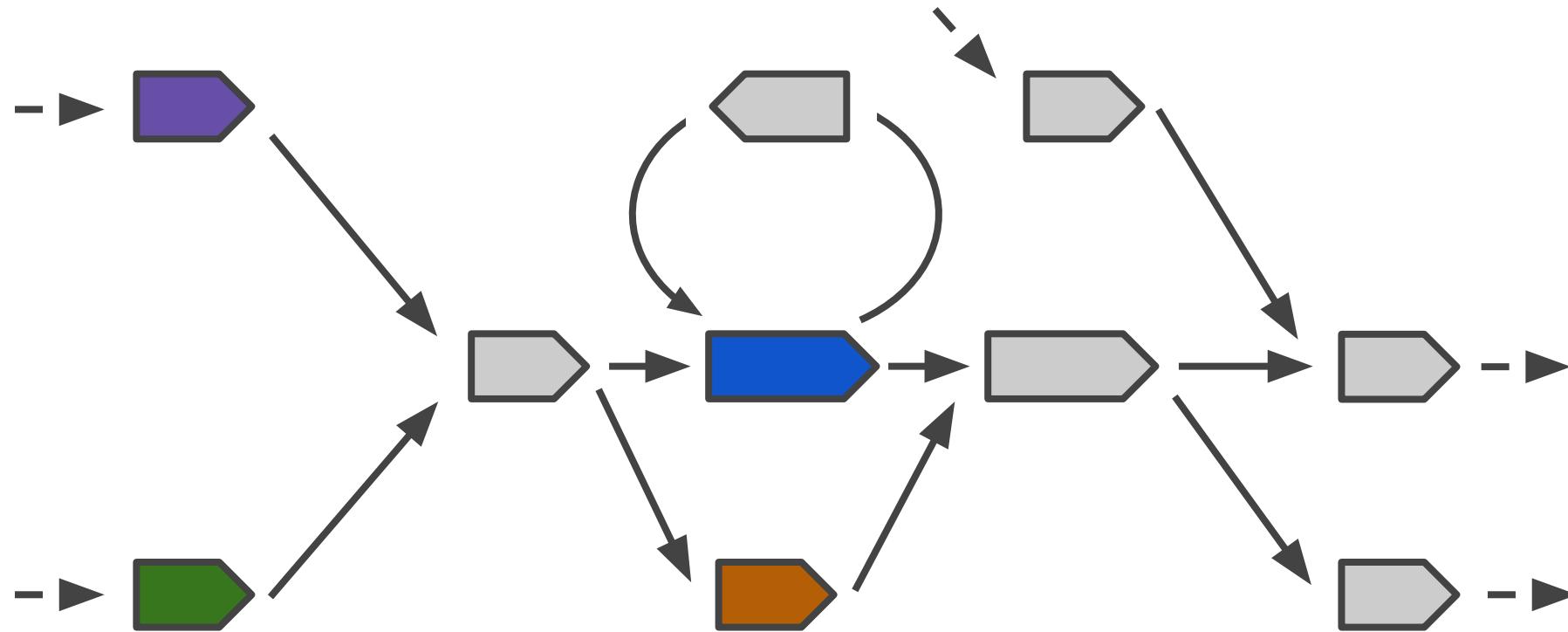
Problem: Closely related strains make read-chaining ambiguous



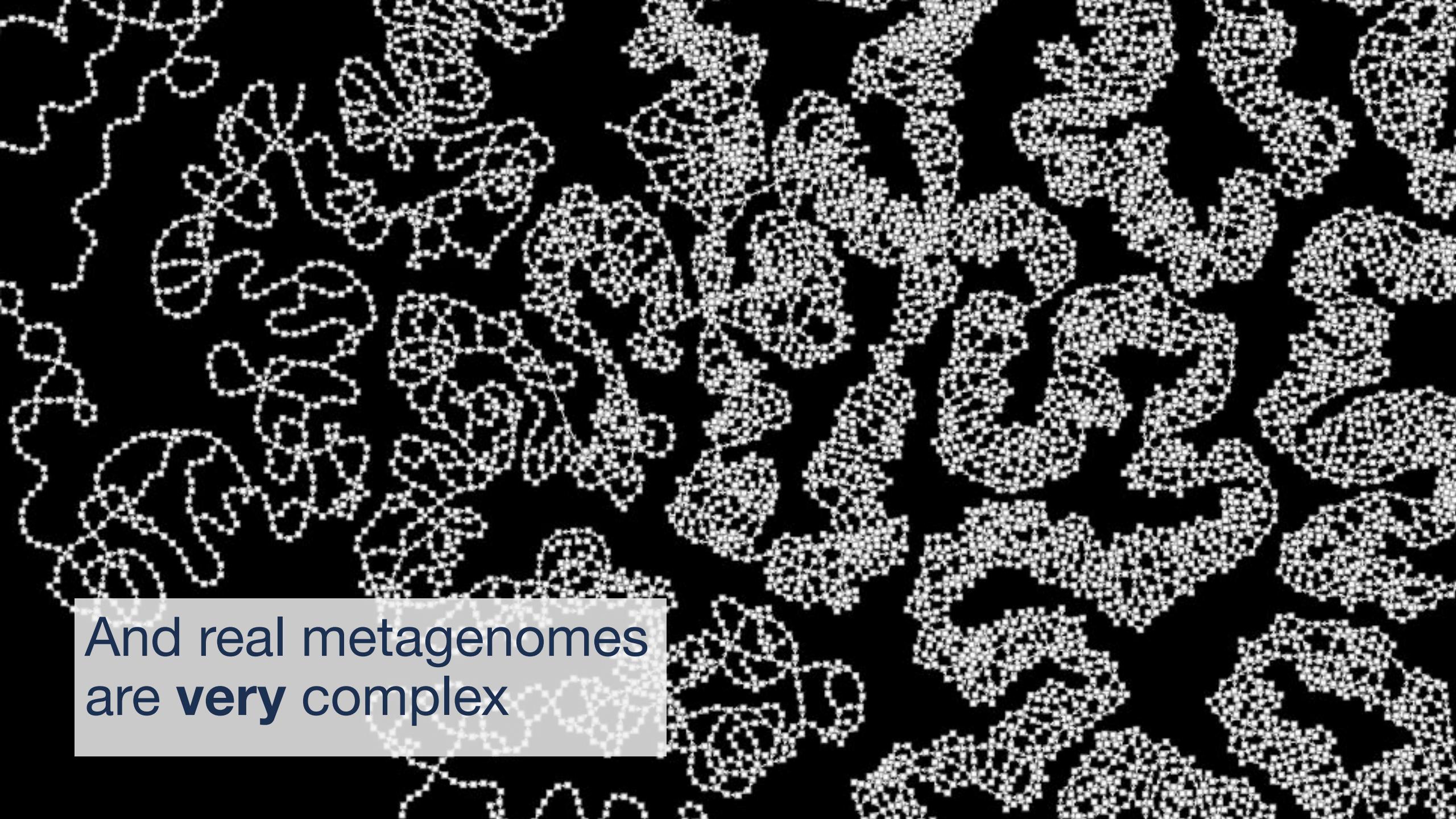
Can be represented as
a graph of sequences
linked by their overlaps



Can be represented as
a graph of sequences
linked by their overlaps

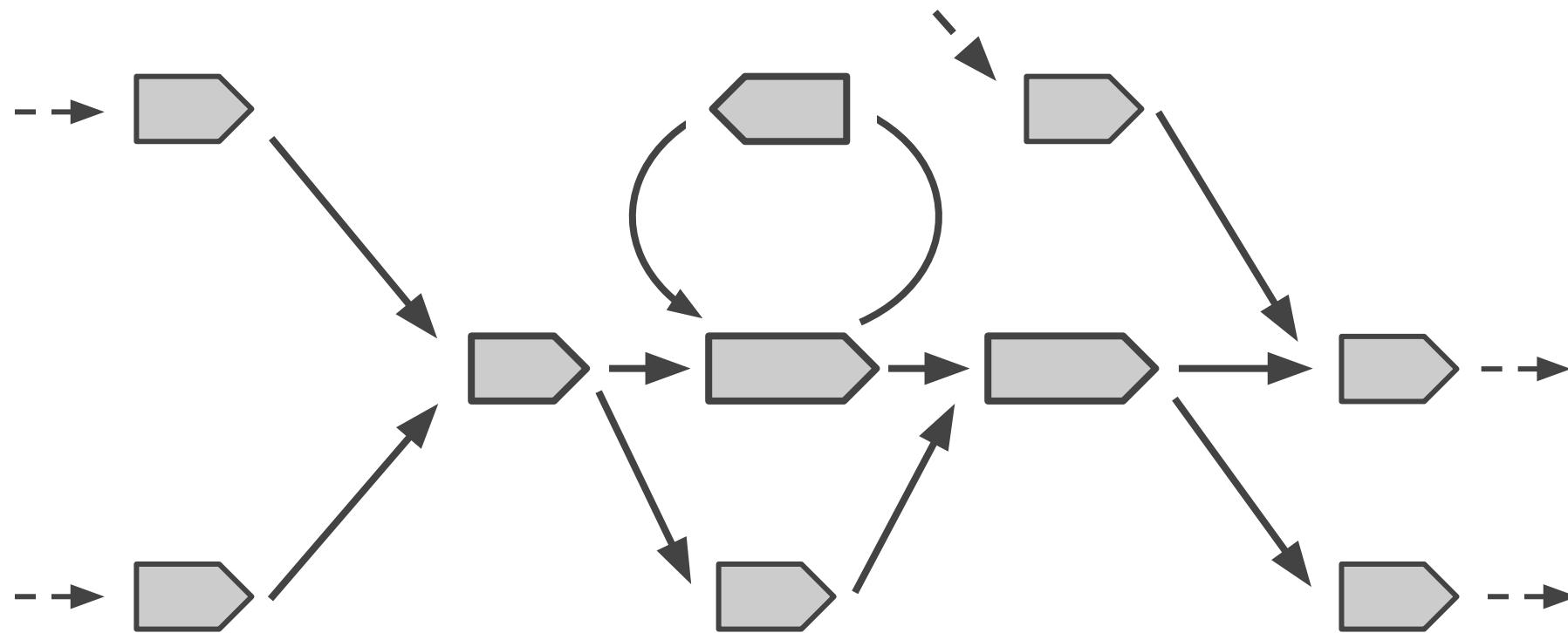


(This problem also comes up for mRNA alternative splicing)

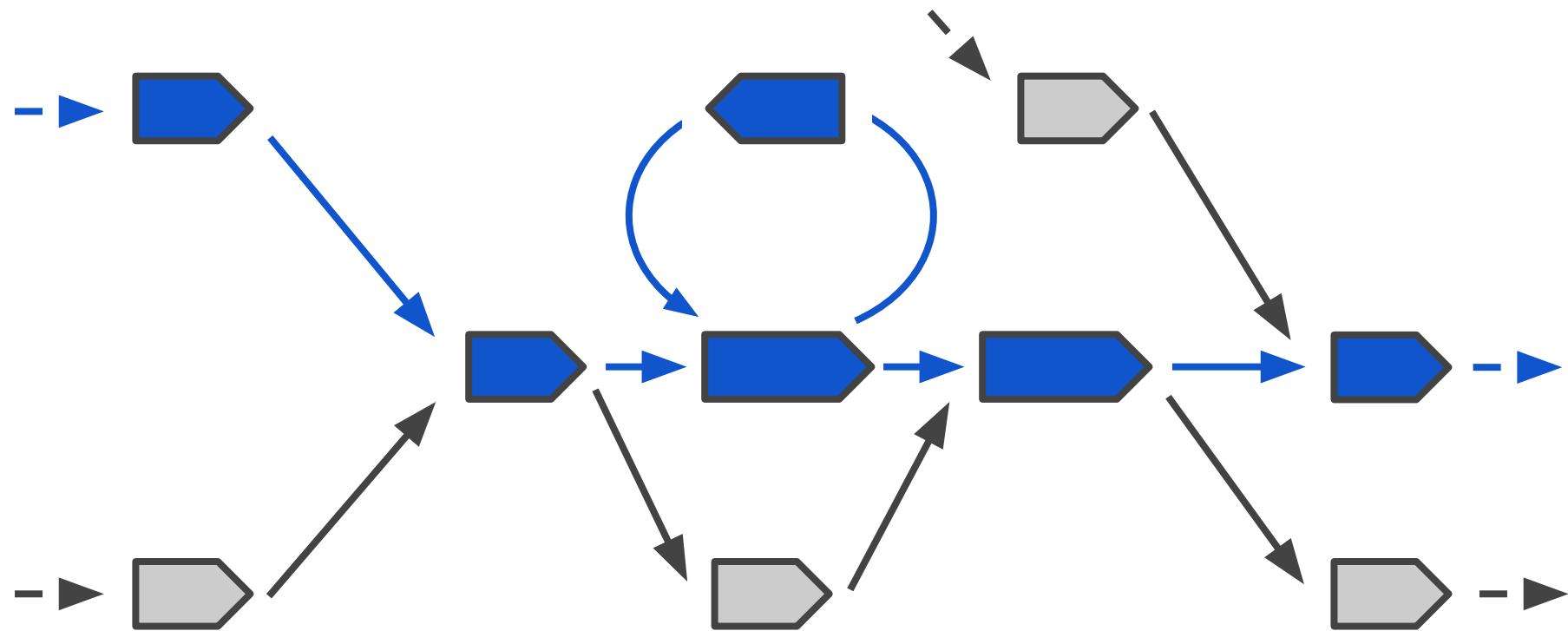


And real metagenomes
are **very** complex

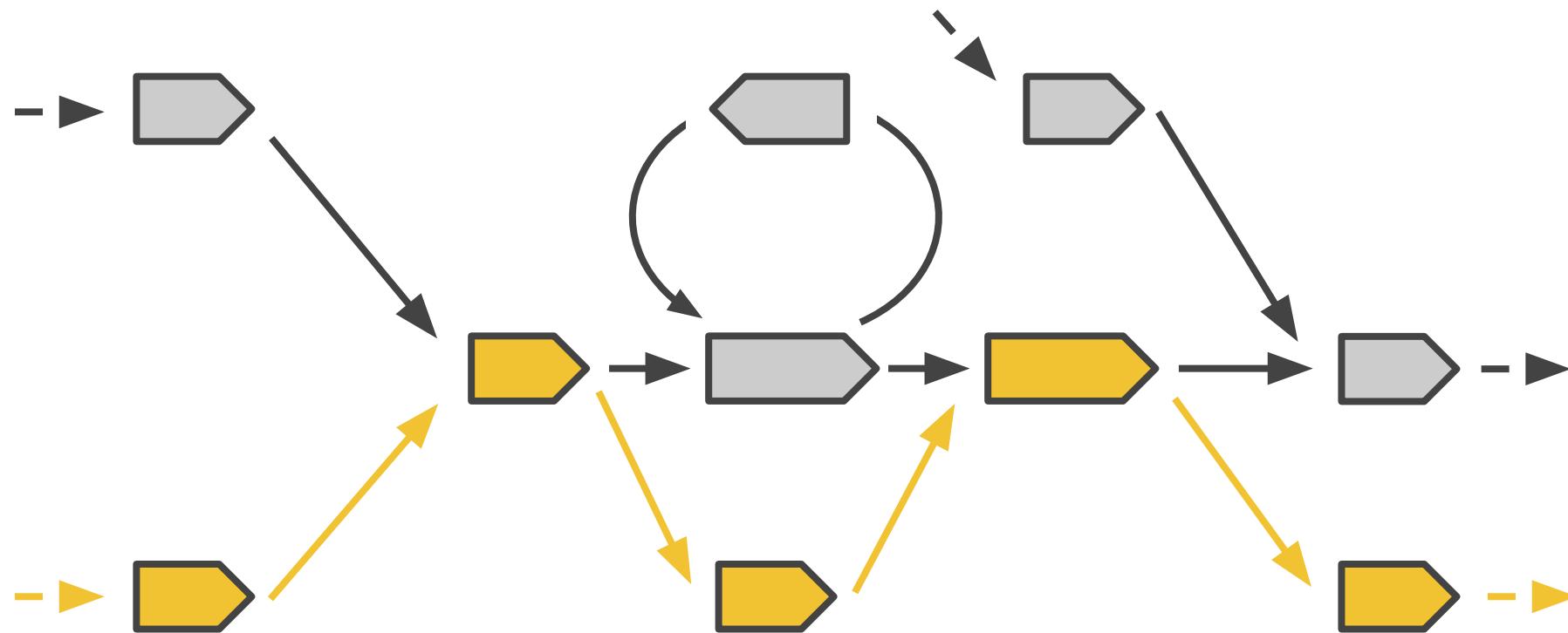
Real genomic
sequences are
paths on the graph



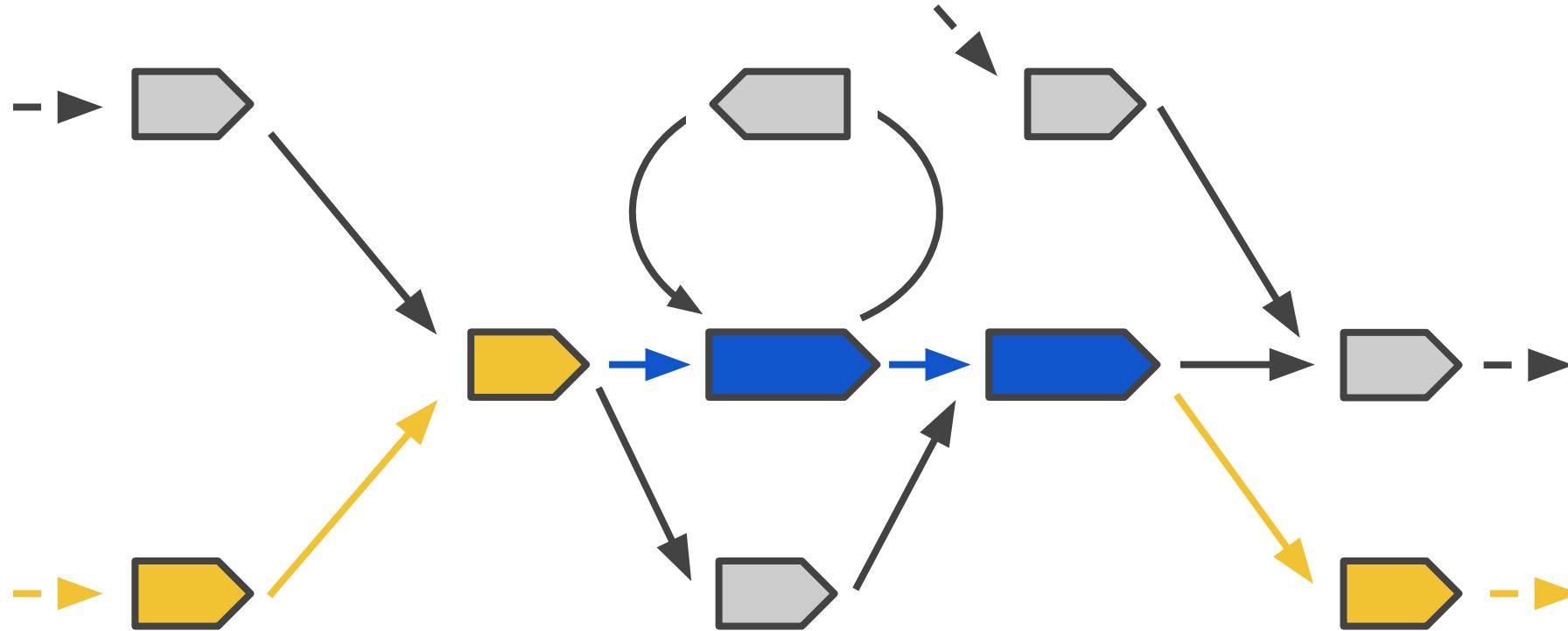
Real genomic sequences are paths on the graph



Real genomic
sequences are
paths on the graph

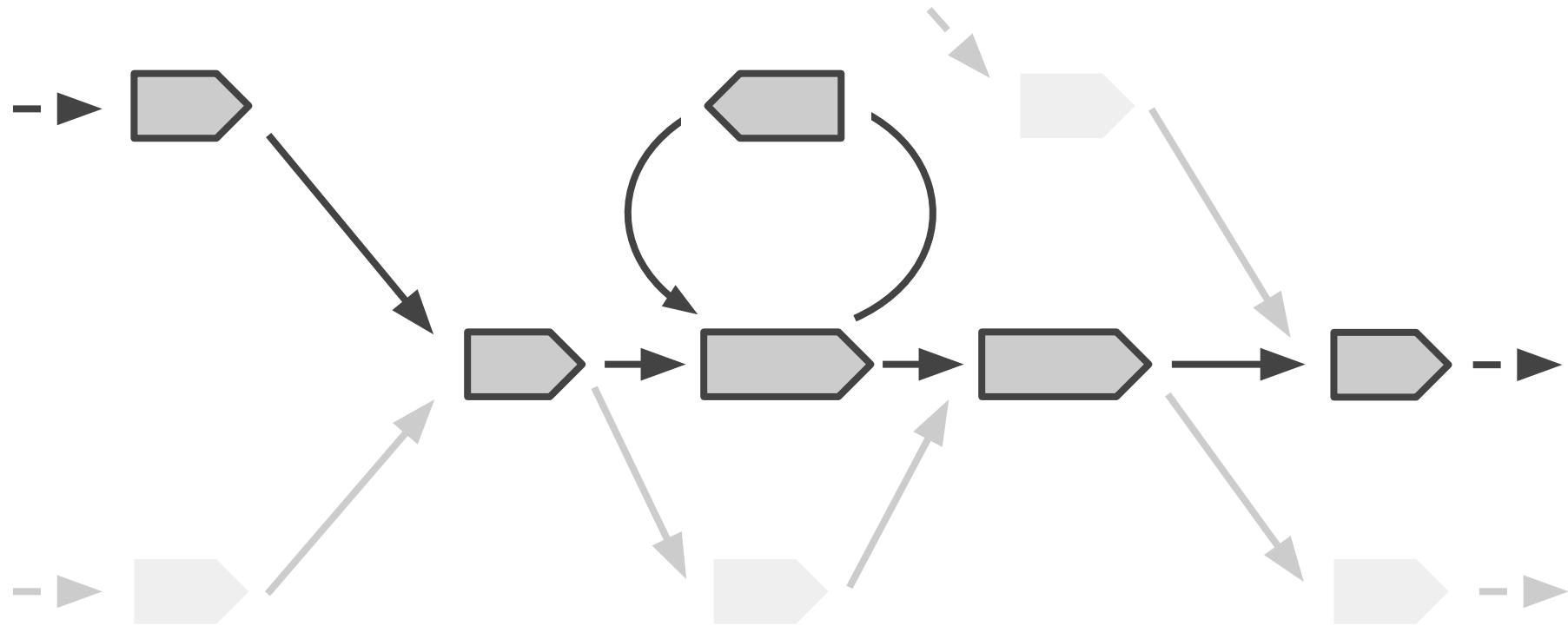


Lots of incorrect paths also exist...
How do we avoid these?

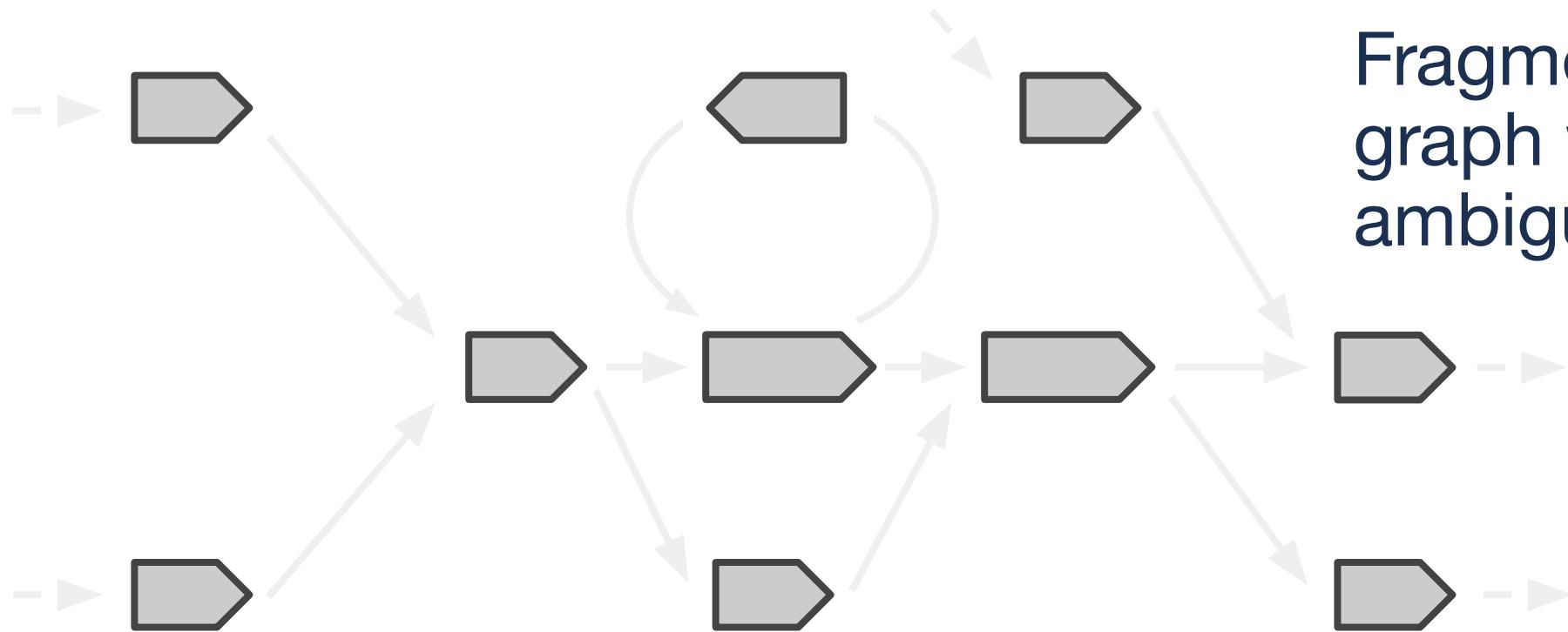


Lots of incorrect paths also exist...
How do we avoid these?

Standard Tools:
Filter out
low-abundance
sequences

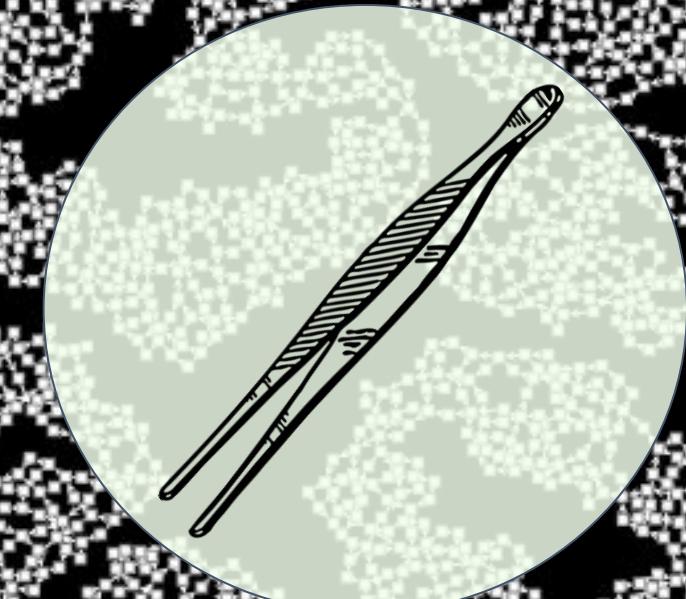


Lots of incorrect paths also exist...
How do we avoid these?



- Standard Tools:
 - Filter out low-abundance sequences
 - Fragment the graph when it's ambiguous

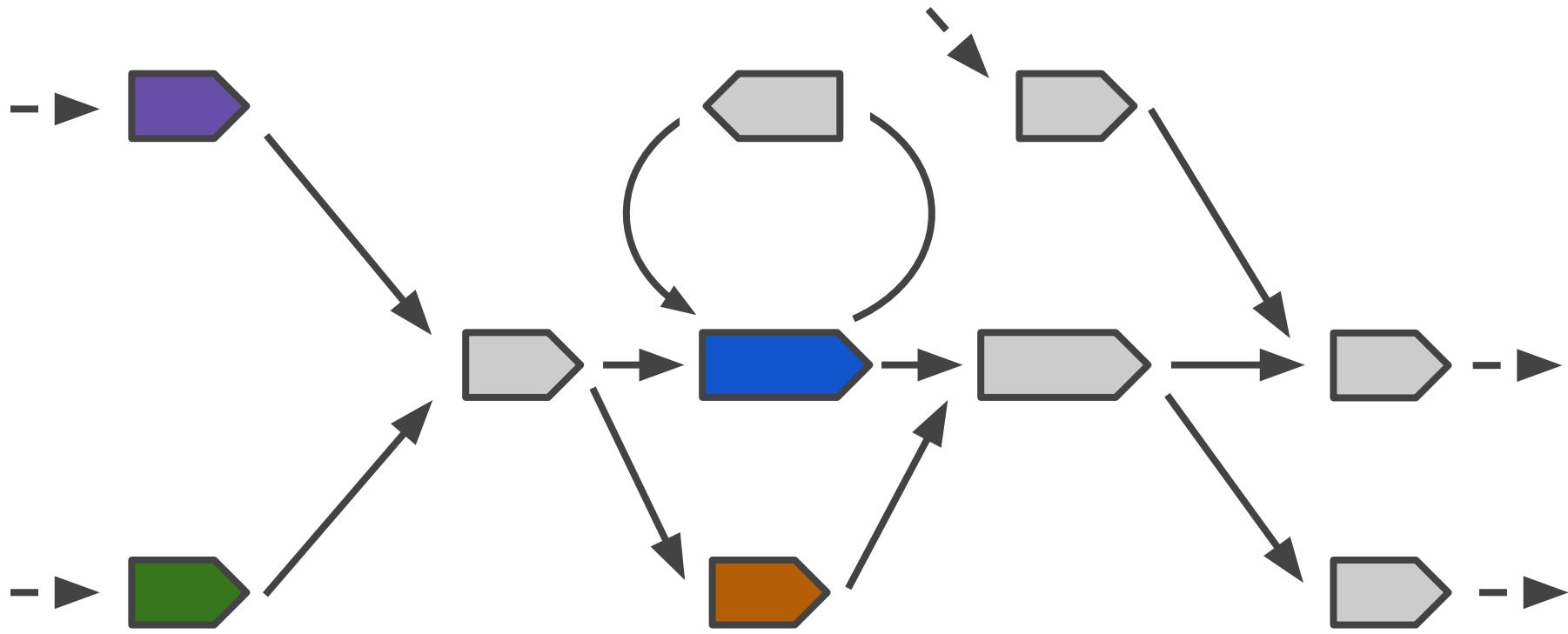
Untangling the hairball



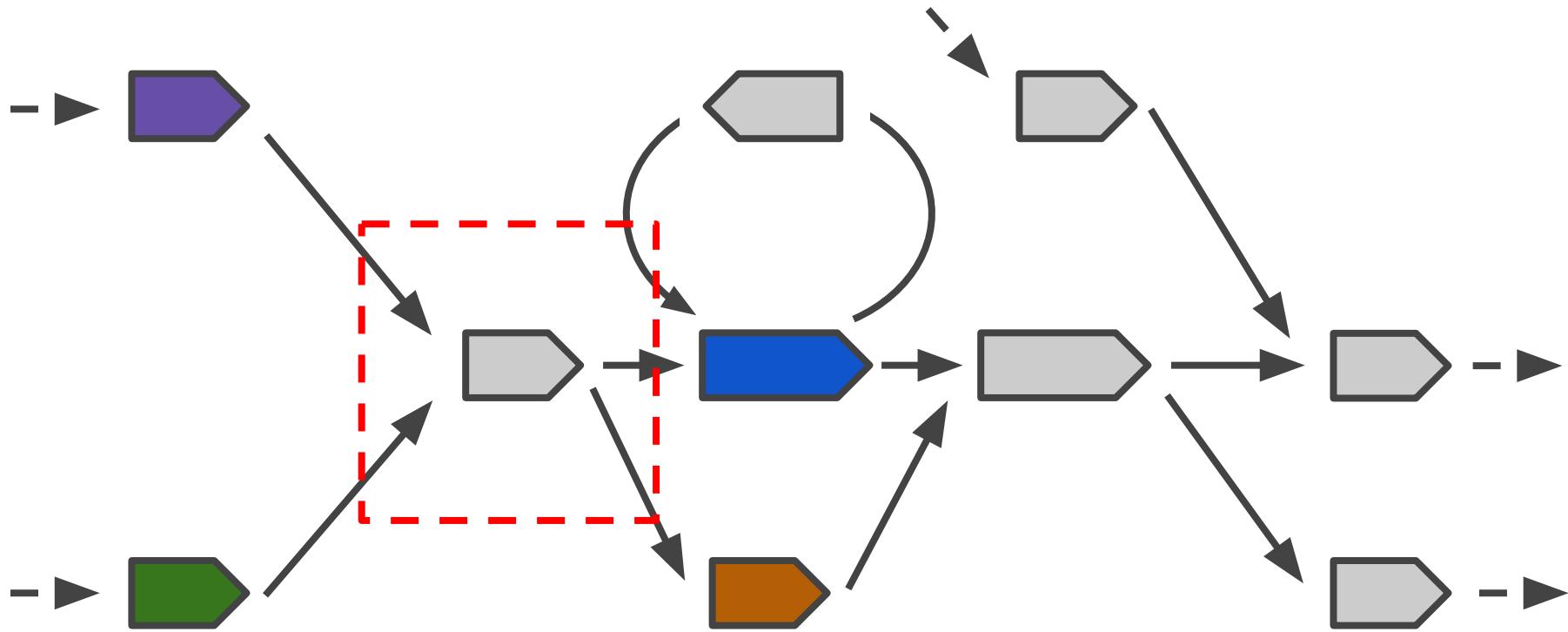
StrainZip:

Untangling the metagenome graph

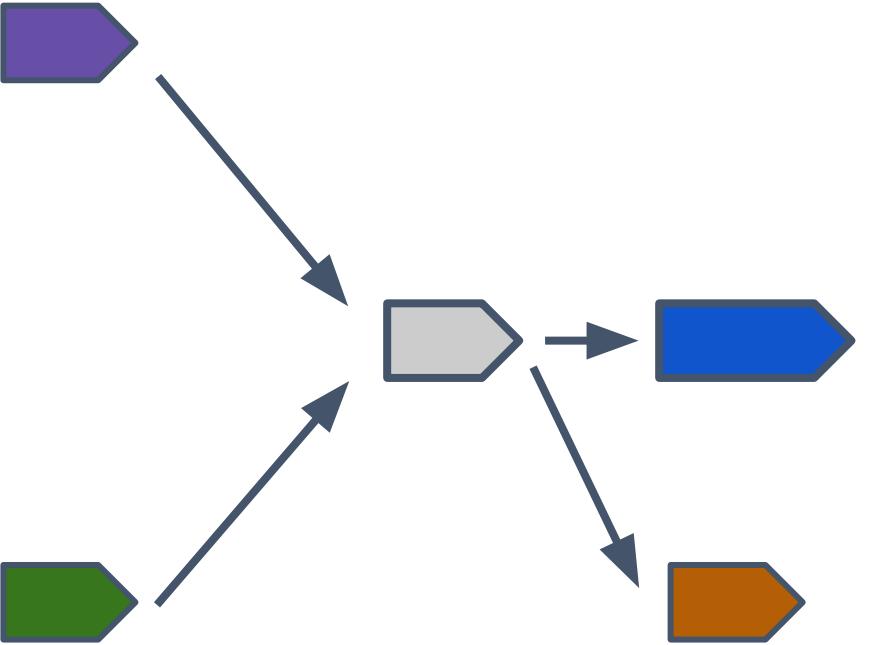
How can we recover long, accurate genome sequences from short reads?



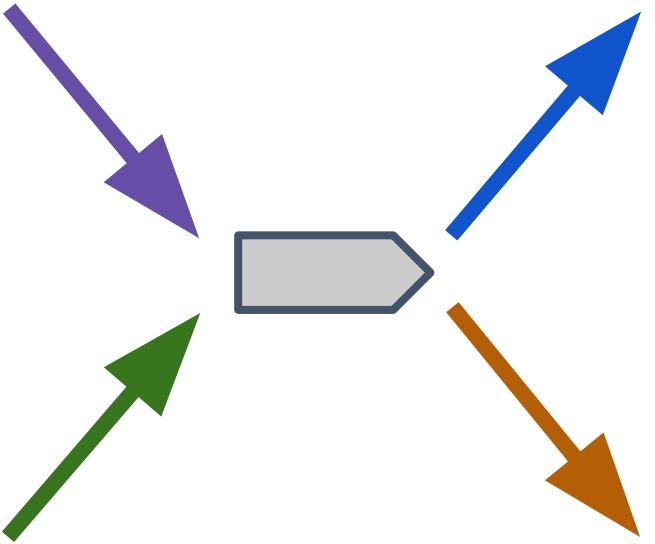
How can we recover long, accurate genome sequences from short reads?



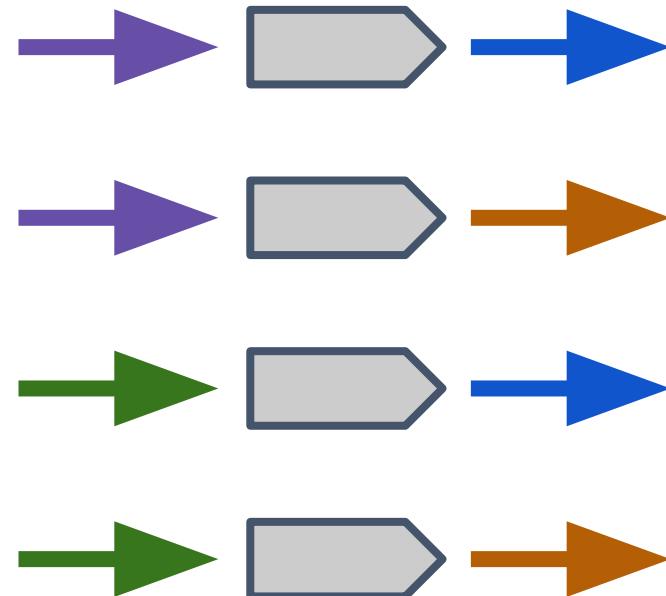
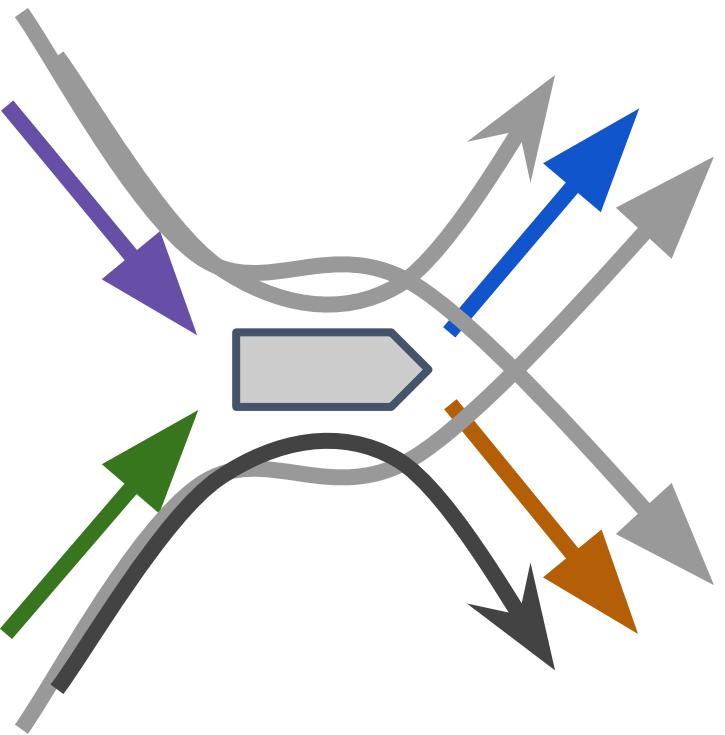
Focus on just one junction at a time



Focus on just one junction at a time

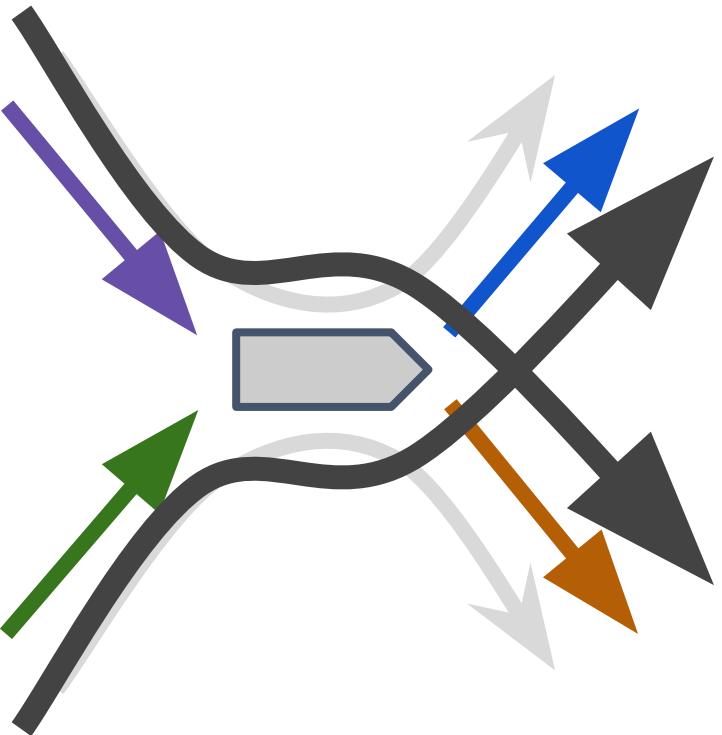


Focus on just one junction at a time



Focus on just one junction at a time

Select local paths



$$\begin{matrix} & \downarrow & \downarrow & \downarrow & \downarrow \\ \rightarrow & 1 & 1 & 0 & 0 \\ \textcolor{green}{\downarrow} & 0 & 0 & 1 & 1 \\ \textcolor{purple}{\downarrow} & 1 & 0 & 1 & 0 \\ \textcolor{orange}{\downarrow} & 0 & 1 & 0 & 1 \end{matrix} \times \begin{matrix} p_{1,1} & p_{1,2} & p_{1,3} \\ p_{2,1} & p_{2,2} & p_{2,3} \\ p_{3,1} & p_{3,2} & p_{3,3} \\ p_{4,1} & p_{4,2} & p_{4,3} \end{matrix} \approx \begin{matrix} e_{1,1} & e_{1,2} & e_{1,3} \\ e_{2,1} & e_{2,2} & e_{2,3} \\ e_{3,1} & e_{3,2} & e_{3,3} \\ e_{4,1} & e_{4,2} & e_{4,3} \end{matrix}$$

$X \quad \beta \quad Y$

Sparse linear
regression across
multiple samples

Focus on just one junction at a time

Select local paths

Unzip

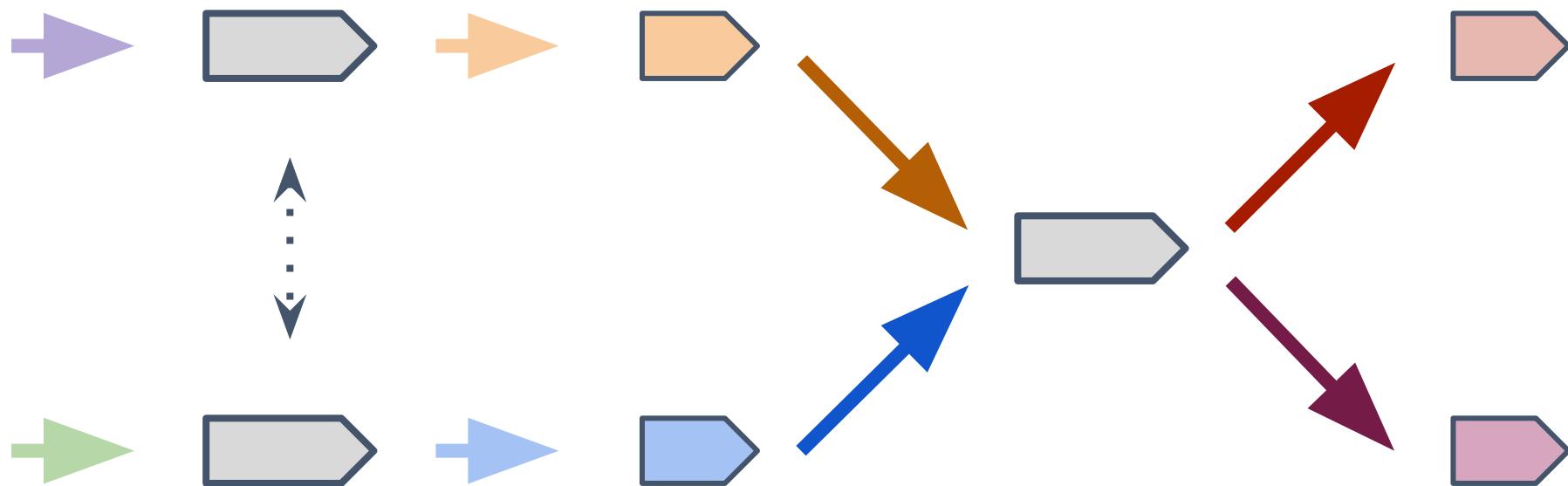


Focus on just one junction at a time

Select local paths

Unzip

Repeat



Focus on just one junction at a time

Select local paths

Unzip

Repeat



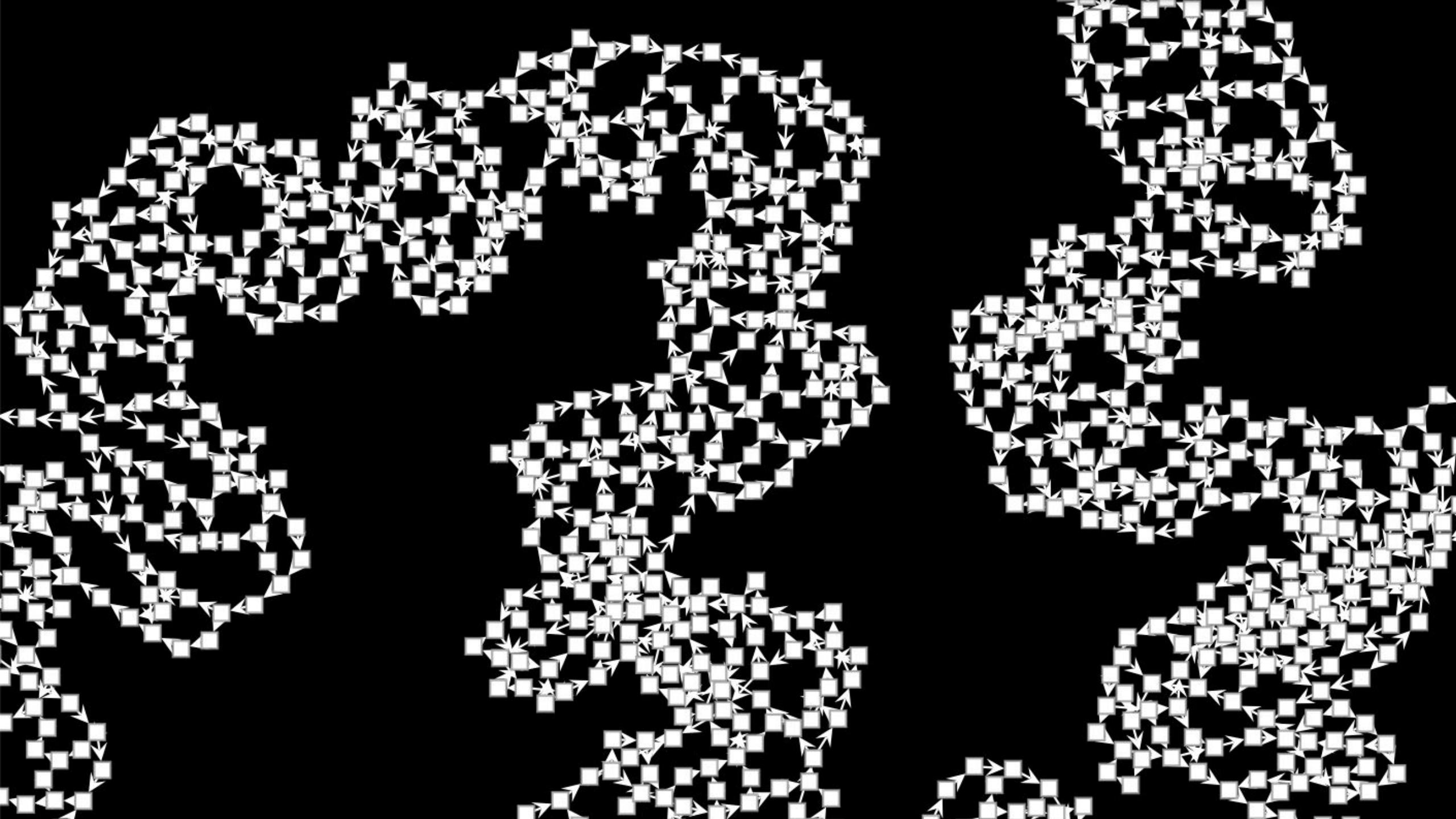
Focus on just one junction at a time

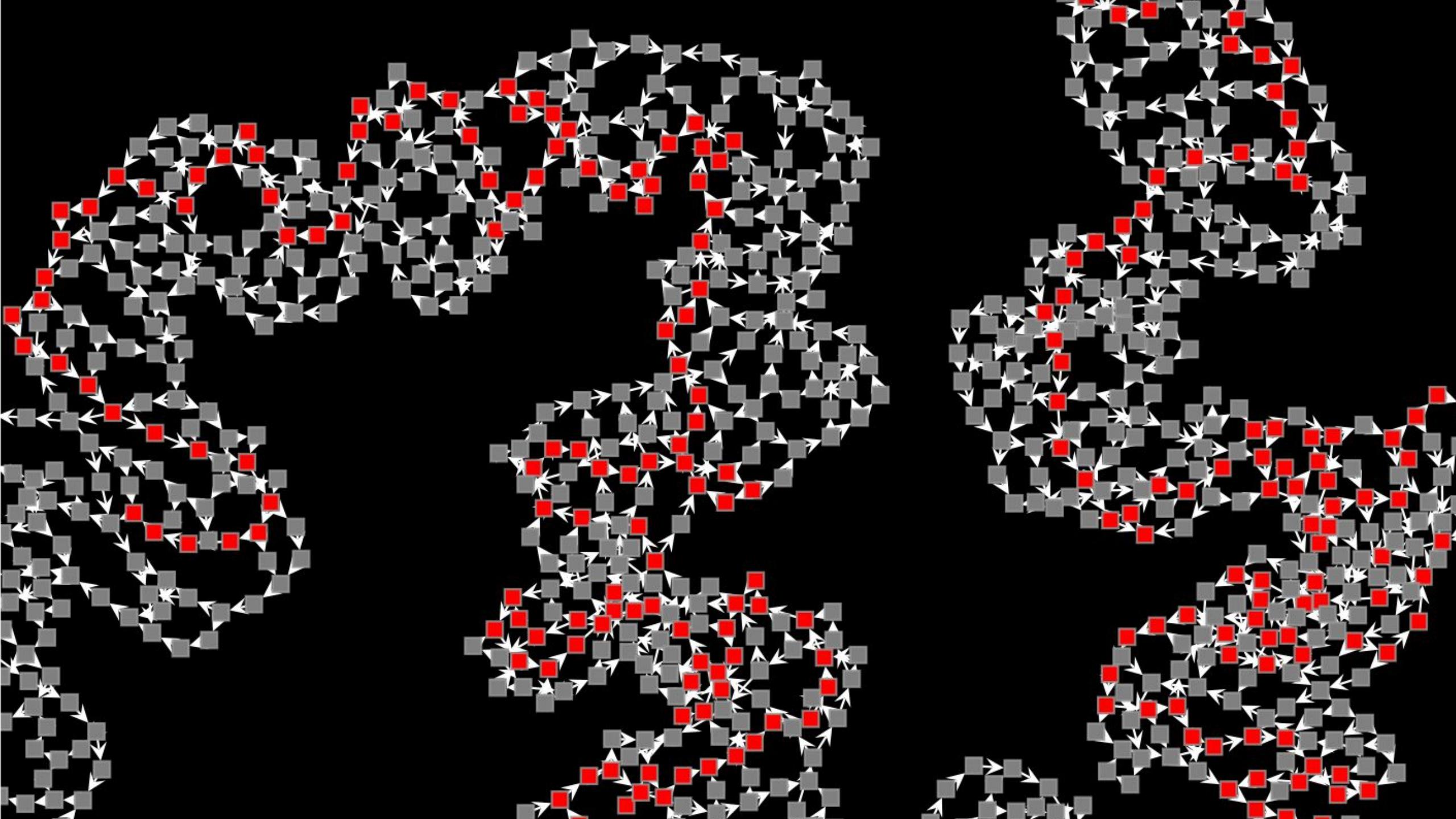
Select local paths

Unzip

Repeat

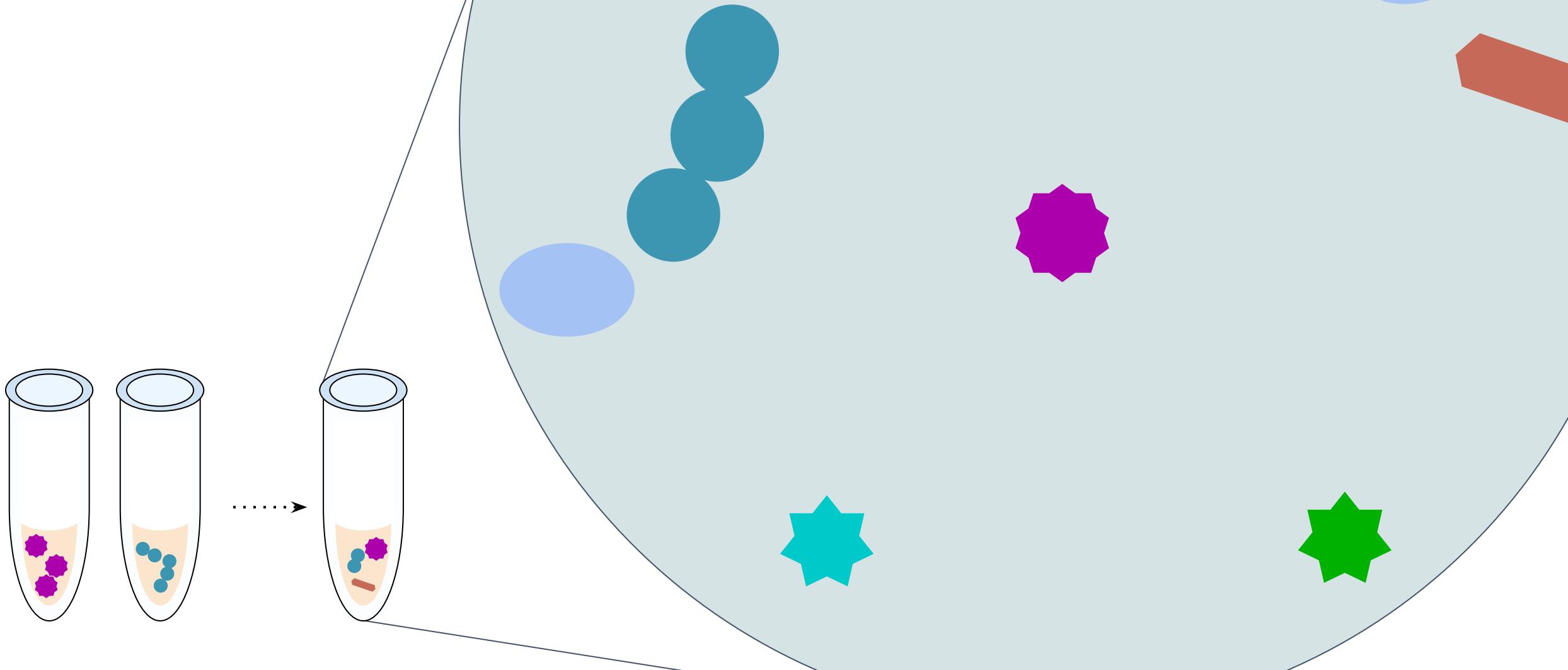






Strain-resolved discovery

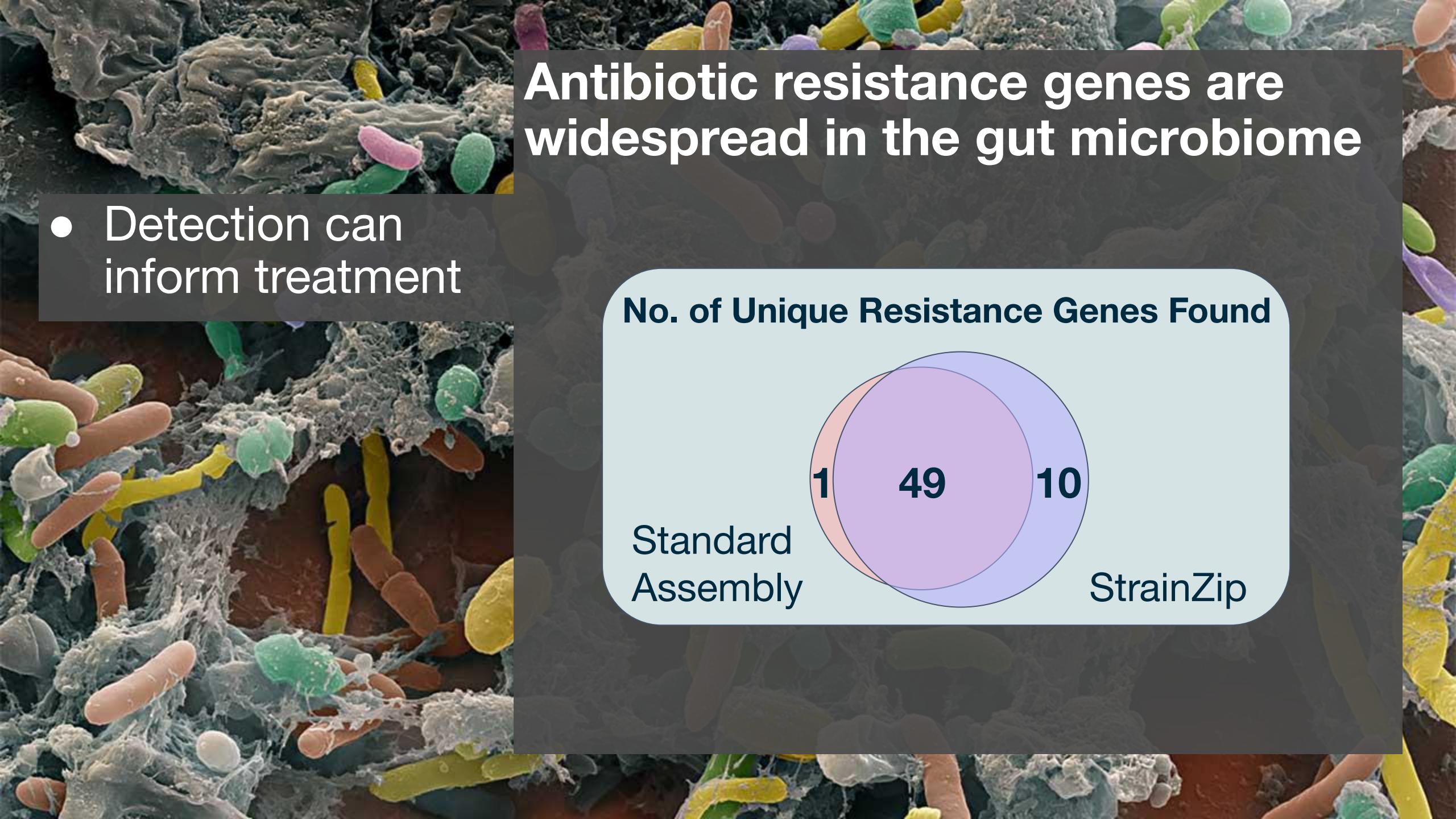
Performance benchmarked on a complex, synthetic community





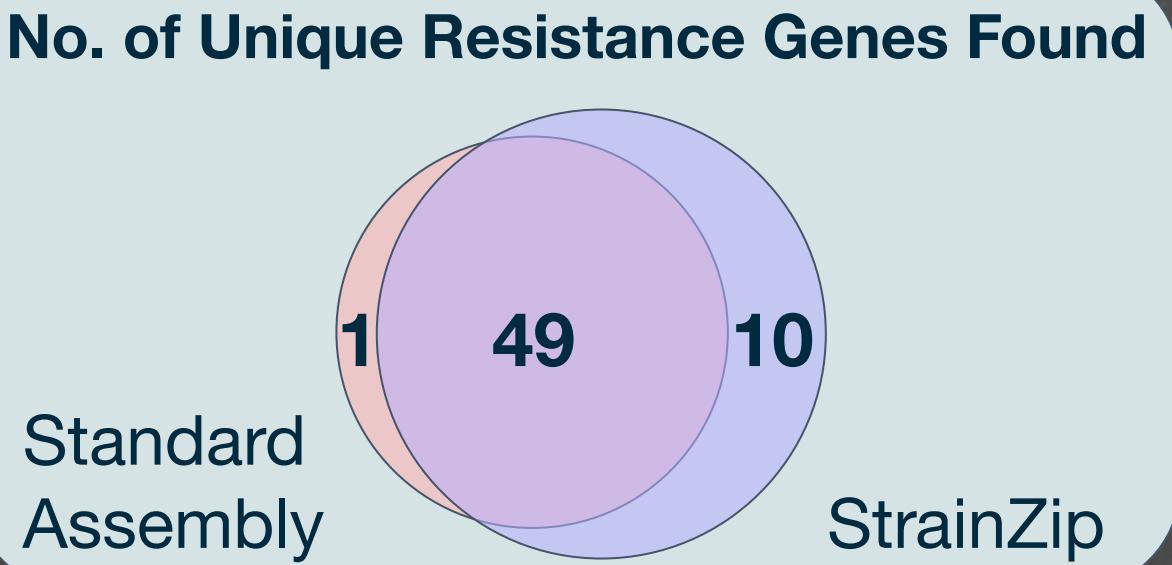
Antibiotic resistance genes are widespread in the gut microbiome

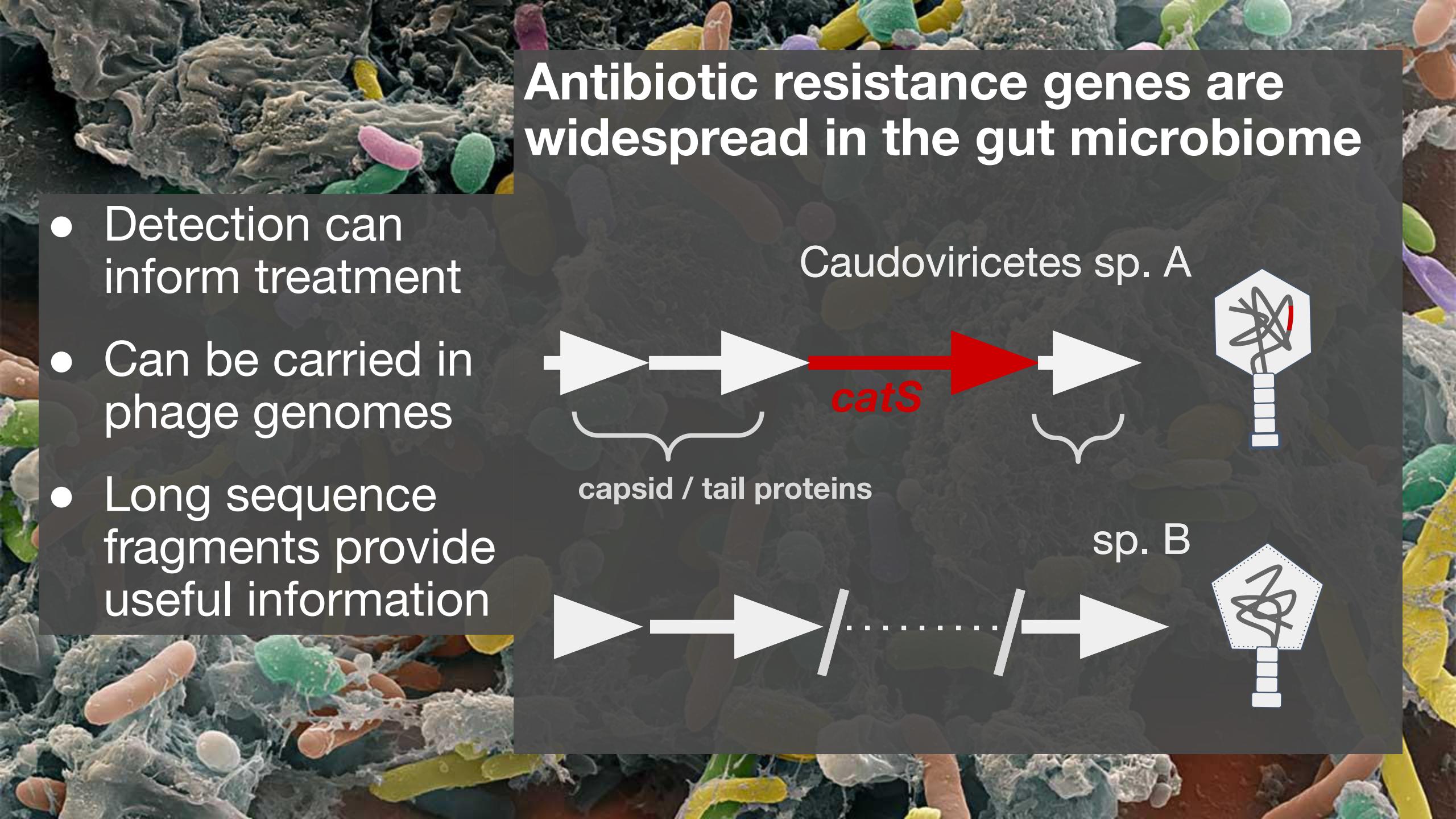
- Detection can inform treatment

A scanning electron micrograph (SEM) showing a diverse community of gut microbiota. Various colored bacteria, including rod-shaped and spherical ones, are visible against a dark, textured background.

Antibiotic resistance genes are widespread in the gut microbiome

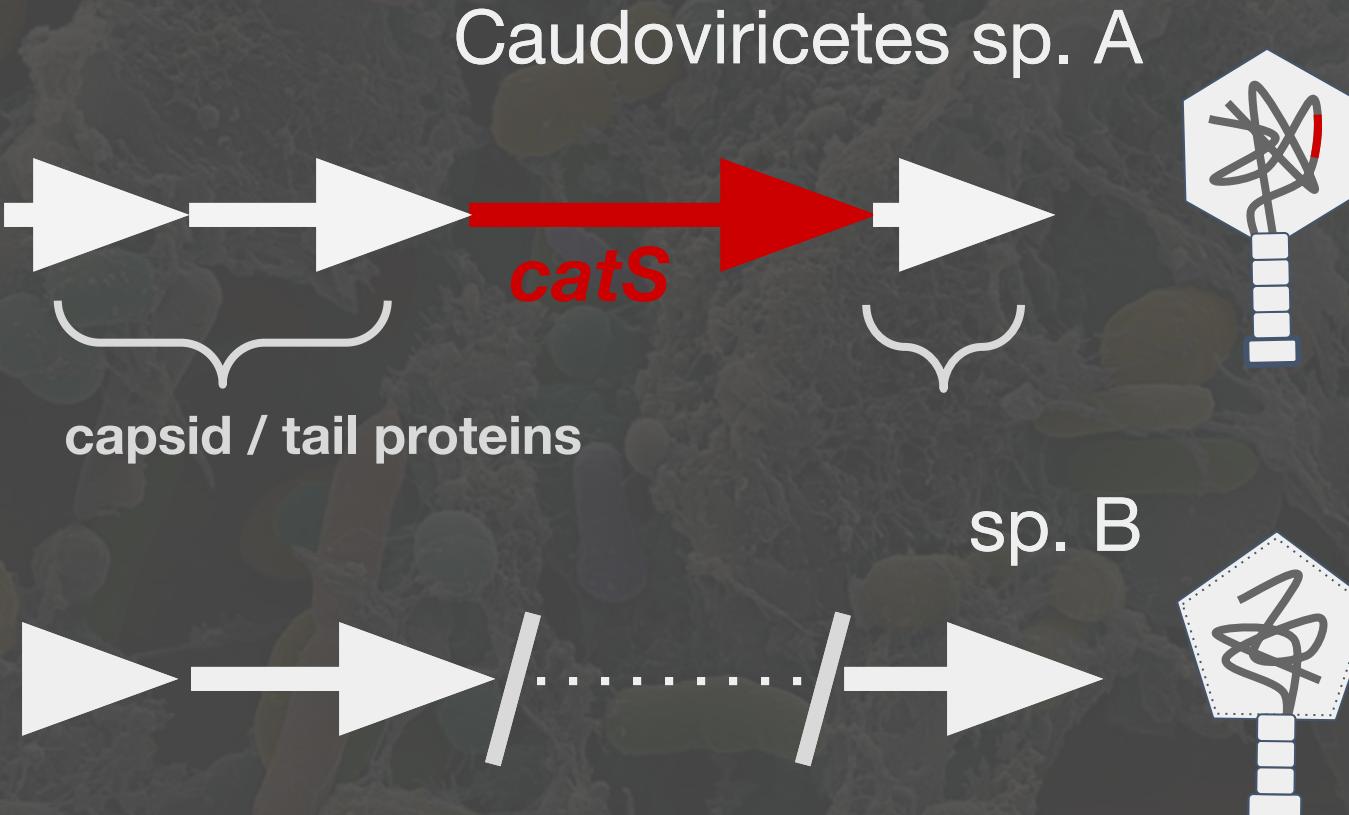
- Detection can inform treatment





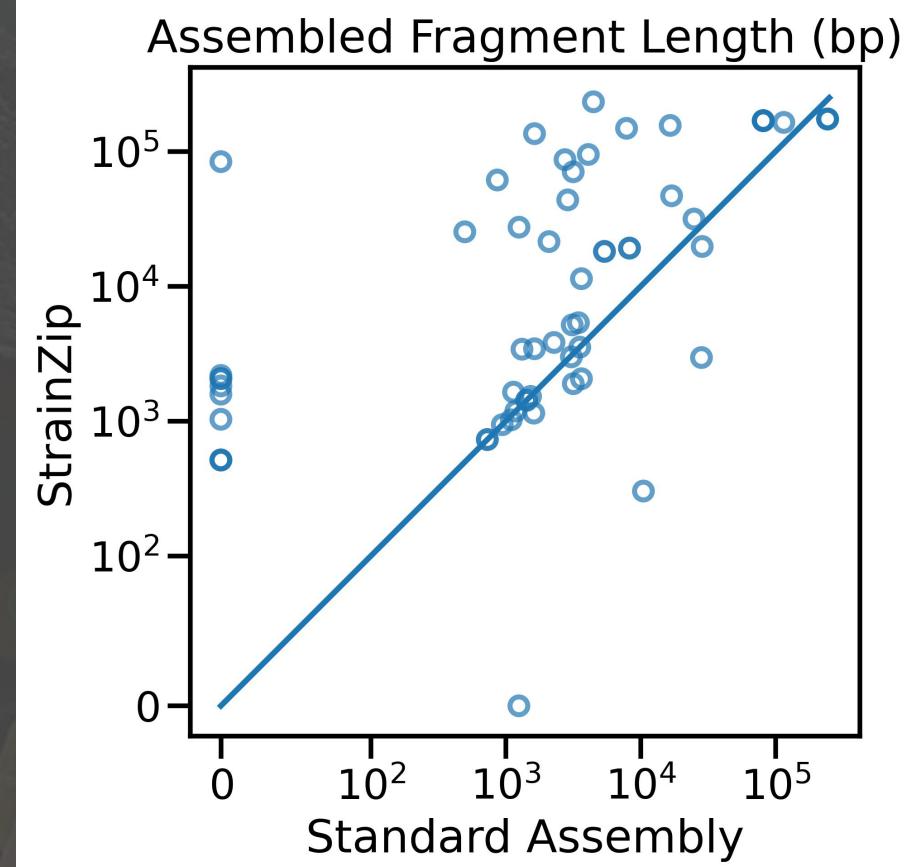
Antibiotic resistance genes are widespread in the gut microbiome

- Detection can inform treatment
- Can be carried in phage genomes
- Long sequence fragments provide useful information

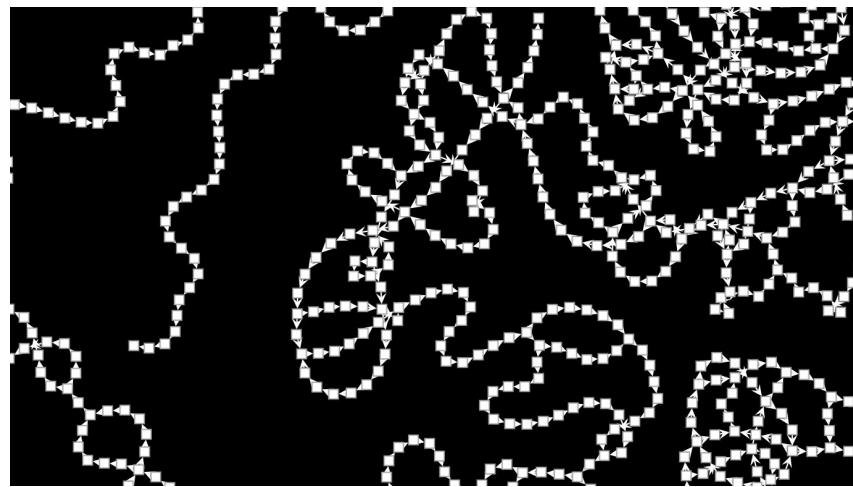


Antibiotic resistance genes are widespread in the gut microbiome

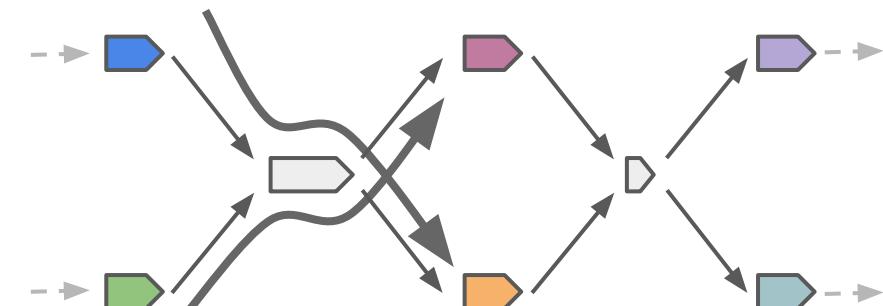
- Detection can inform treatment
- Can be carried in phage genomes
- Long sequence fragments provide useful information



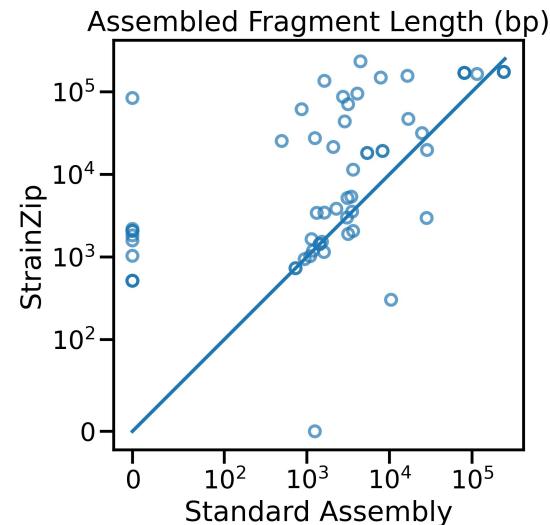
Complex Metagenome Graphs



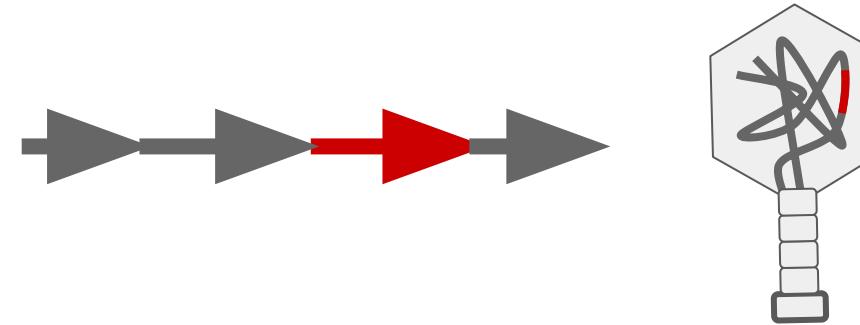
StrainZip Iteratively Unzips Junctions



Strain-Resolved Metagenomics

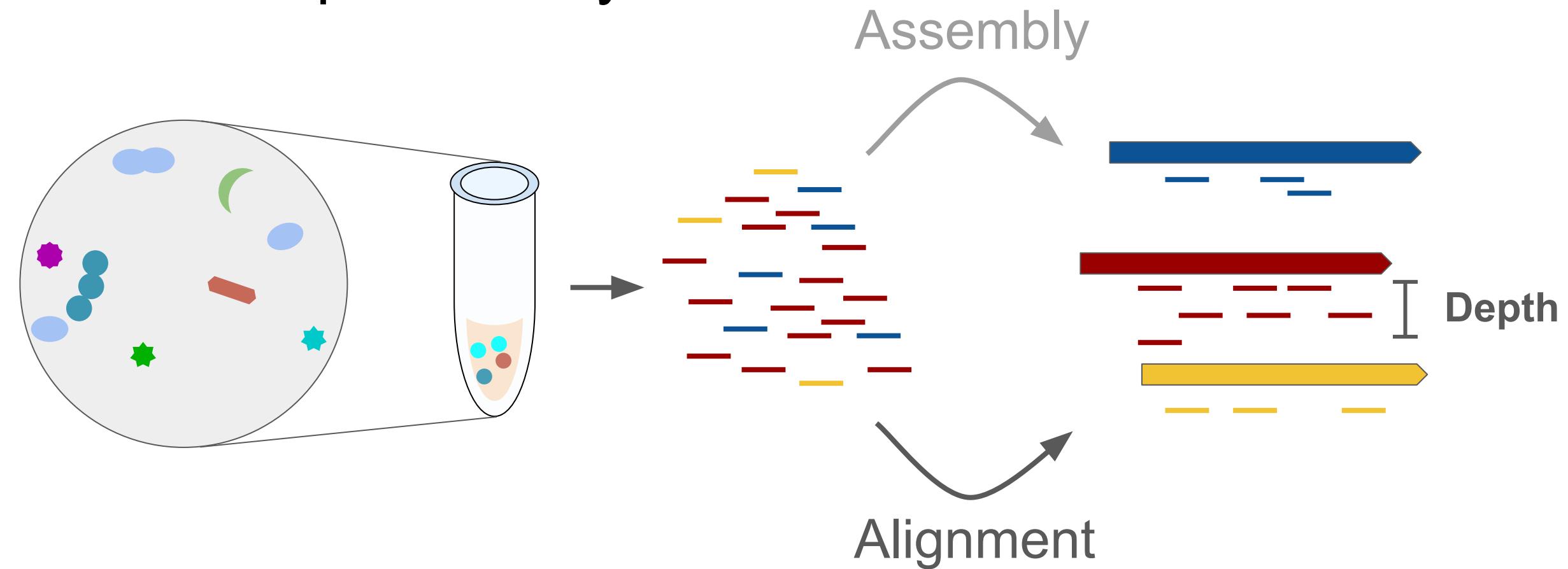


Antibiotic Resistance Potential of Phage

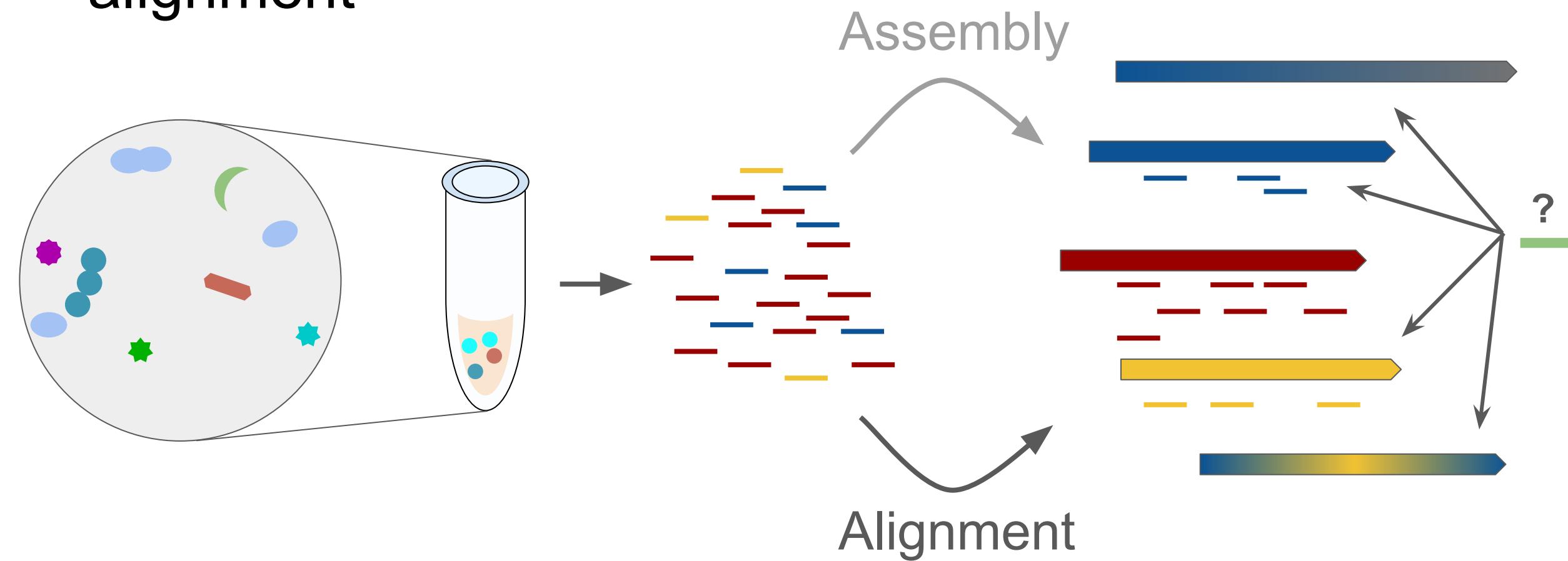


Rewind: I also care about
depth quantification

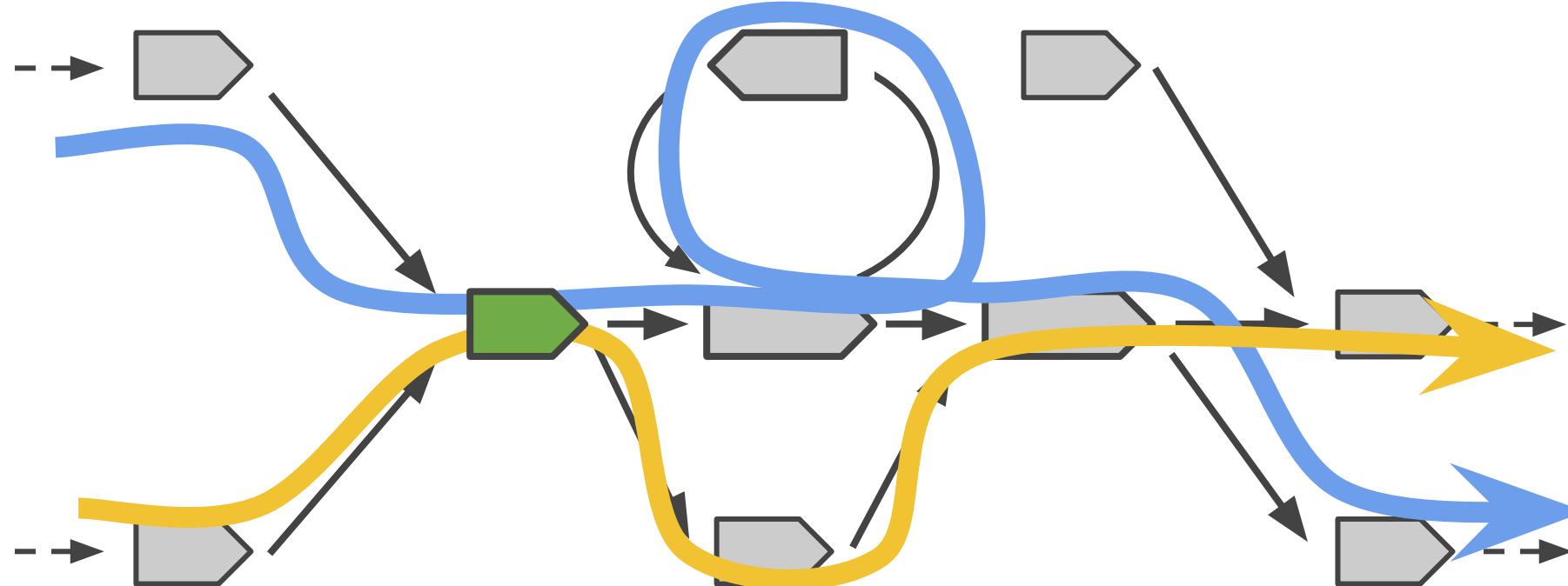
Assembly and depth quantification are complementary



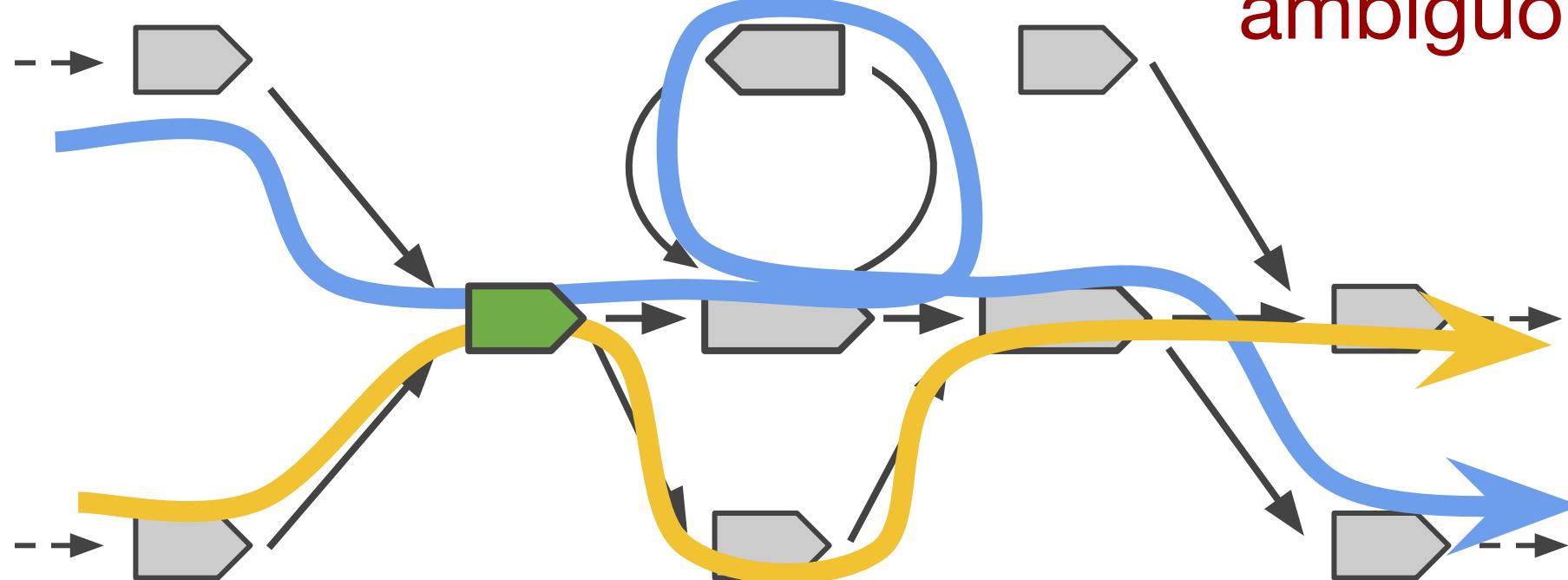
Closely related sequences
are a major challenge for
alignment



Shared sequences
mean reads map
ambiguously



Shared sequences
mean reads map
~~ambiguously~~
**kmers are
ambiguous**



Quick intro to de Bruijn graphs

Read #1

...CGTA CCTGGATTAC...

Assembly

...CGTA CCTGGATTAC**TTAA**...

Read #2

CCTGGATTAC**TTAA**...

De Bruijn graphs

Motivation: **Assembly** - stitching together longer sequences using overlapping portions

Fragment reads into k-mers

Read #1

...CGTA CCTGGATTAC

CGTA

GTAC

TACC

ACCT

CCTG

CTGG

TGGA

GGAT

GATT

ATTA

TTAC

Read #2

CCTGGATTACTTAA...

CCTG

CTGG

TGGA

GGAT

GATT

ATTA

TTAC

TACT

ACTT

CTTA

TTAA

All k-mers

...

CGTA

GTAC

TACC

ACCT

CCTG (x2)

CTGG (x2)

TGGA (x2)

GGAT (x2)

GATT (x2)

ATTA (x2)

TTAC (x2)

TACT

ACTT

CTTA

TTAA

...

Collect unique k-mers

CGTA GTAC TACC ACCT CCTG CTGG TGGA GGAT GATT ATTA TTAC TACT ACTT CTTA TTAA

Identify k-mer pairs where (k-1) suffix on one is same as other's prefix

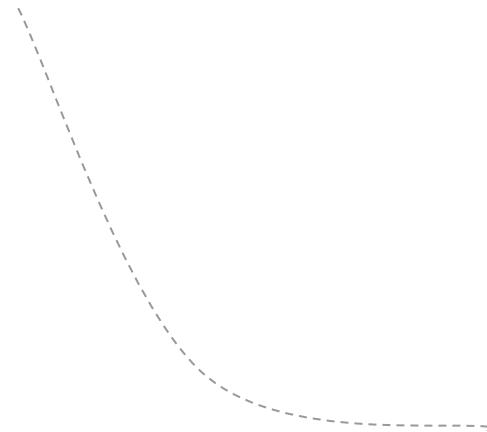
CGTA GTAC TACC ACCT CCTG CTGG TGGA GGAT GATT ATTA TTAC TACT ACTT CTTA TTAA

Draw edge

CGTA → GTAC TACC ACCT CCTG CTGG TGGA GGAT GATT ATTA TTAC TACT ACTT CTTA TTAA

Linear paths (unitigs) are assembled sequence

CGTA → GTAC → TACC → ACCT → CCTG → CTGG → TGGA → GGAT → GATT → ATTA → TTAC → TACT → ACTT → CTTA → TTAA



Unitig:

...CGTACCTGGATTAC**TTAA**...

Mutations / errors introduce new k-mers

Read #1

...CGTA C CTGG ATTAC

CGTA

GTAC

TACC

ACCT

CCTG

CTGG

TGGA

GGAT

GATT

ATTA

TTAC

Read #2

CCTG CATTAC TAA...

CCTG

CTGC

TGCA

GCAT

CATT

ATTA

TTAC

TACT

ACTT

CTTA

TTAA

Diversity / Errors

...CGTACCTG GATTACTTAA...

...CGTACCTG CATTACTTAA...

Same edge-drawing process

CTGG TGGA GGAT GATT

CGTA → GTAC TACC ACCT CCTG

ATTA TTAC TACT ACTT CTTA TTAA

CTGC TGCA GCAT CATT

Same edge-drawing process



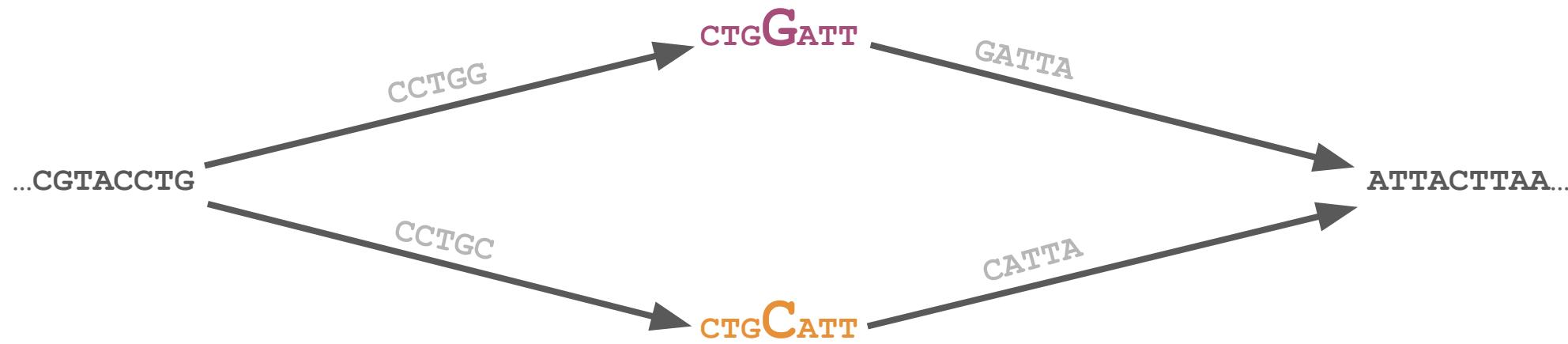
But now some k-mers have multiple edges



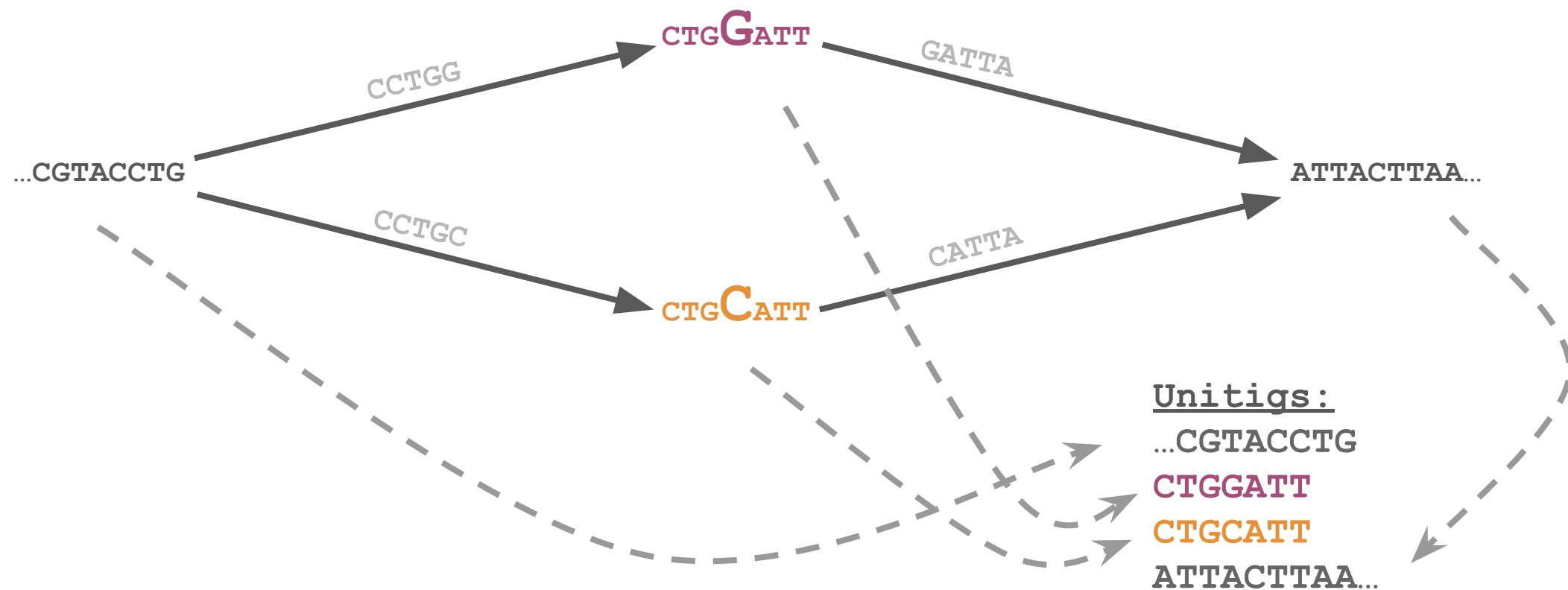
This introduces a "bubble"



The two sides of the bubble reflect the observed diversity



Again we extract unitigs, but now they're shorter, fragmented



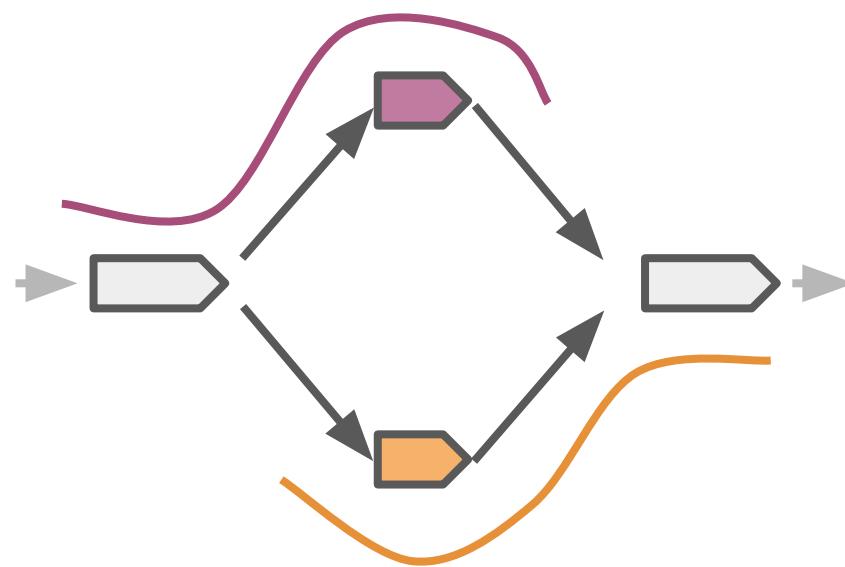
Sequences are walks along the graph;
can align reads without worrying about fragmentation

Read #1

...CGTACTGGATTAC

Read #2

CCTGCATTACTTAA...



Alternatively: Exact k-mer counting

<u>Unitig #1</u>	<u>Unitig #2</u>	<u>Unitig #3</u>	<u>Unitig #4</u>
CGTA	CTGG	CTGC	ATTA (x2)
GTAC	TGGA	TGCA	TTAC (x2)
TACC	GGAT	GCAT	TACT
ACCT	GATT	CATT	ACTT
CCTG (x2)			CTTA
			TTAA



Alternatively: Exact k-mer counting

Much faster than read alignment

Every k-mer in the sample is in the dBG, by construction

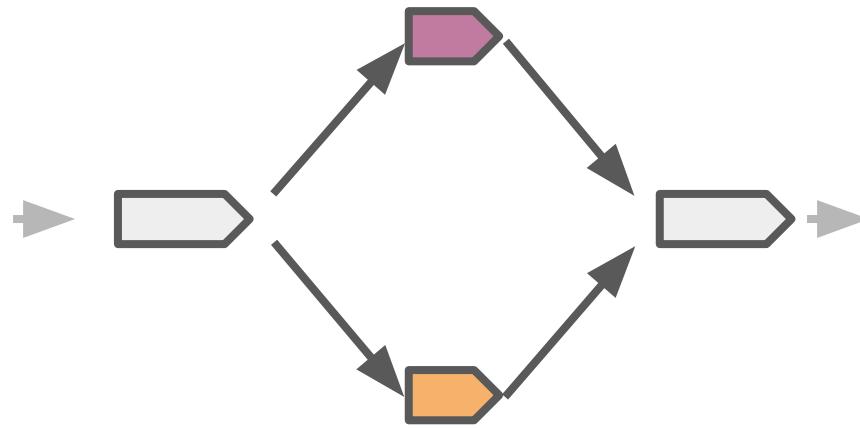


Alternatively: Exact k-mer counting

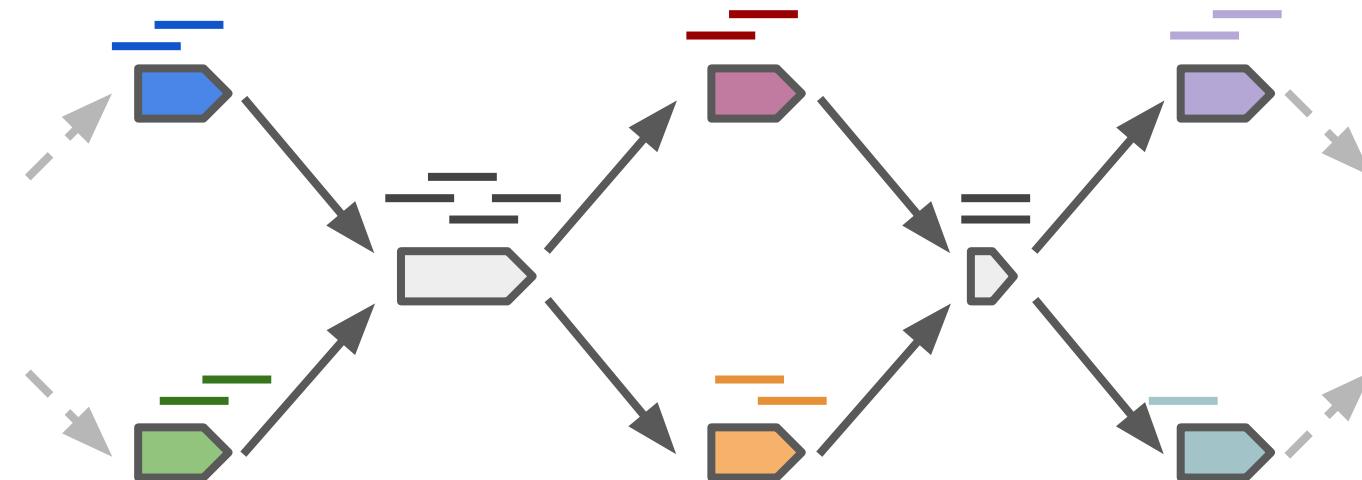
Much faster than read alignment

Every k-mer in the sample is in the dBG, by construction

No ambiguity about what is being quantified: it's unitigs



KEY IDEA: The expected depth of a k-mer is the sum of the paths that include that k-mer



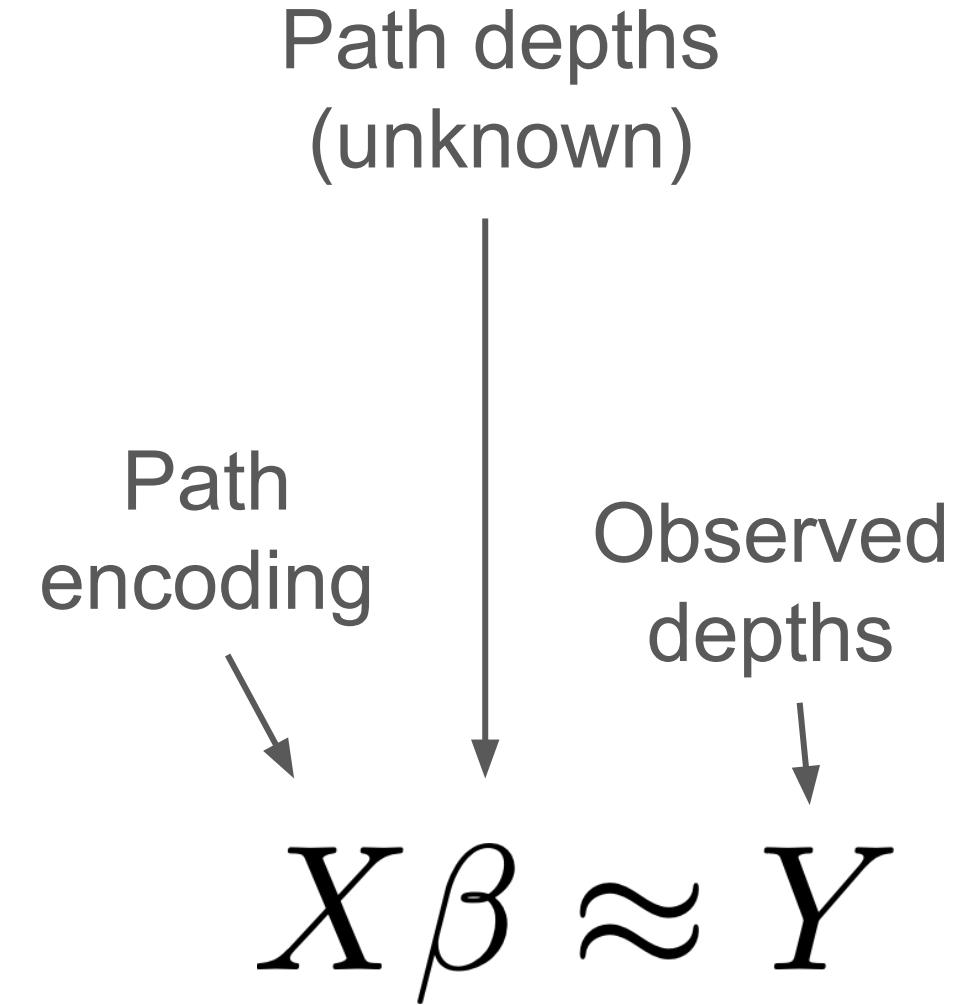
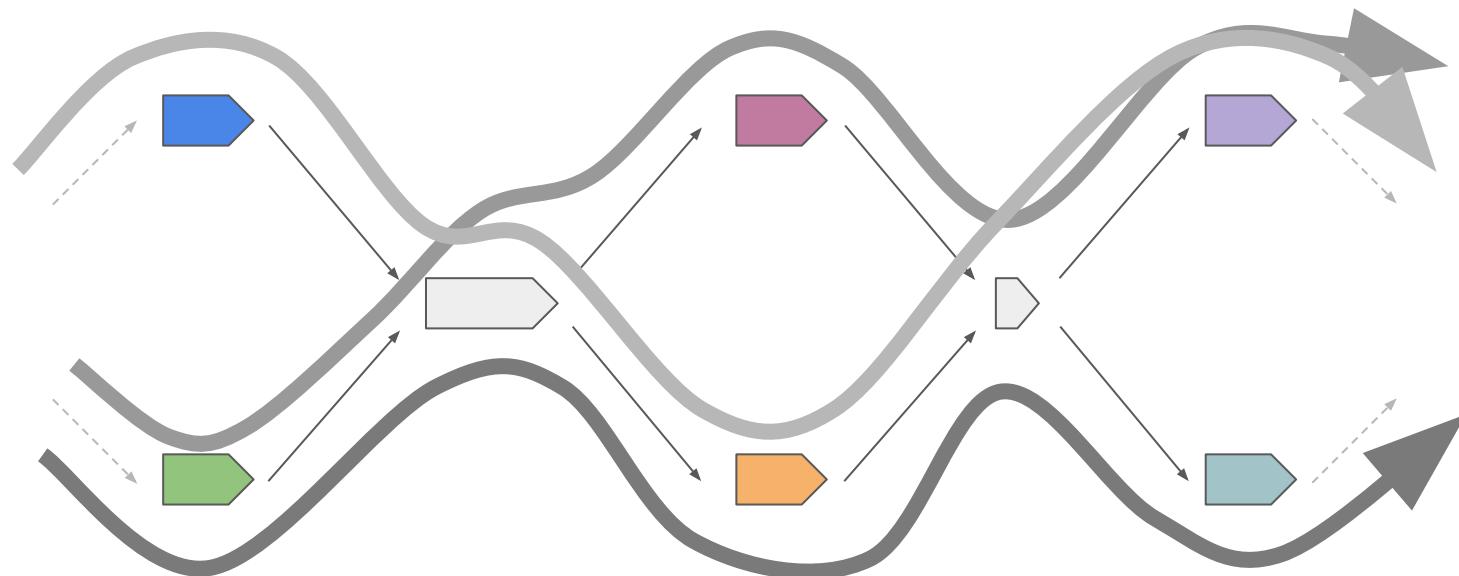
Path depths
(unknown)

Indicator:
k-mer in path

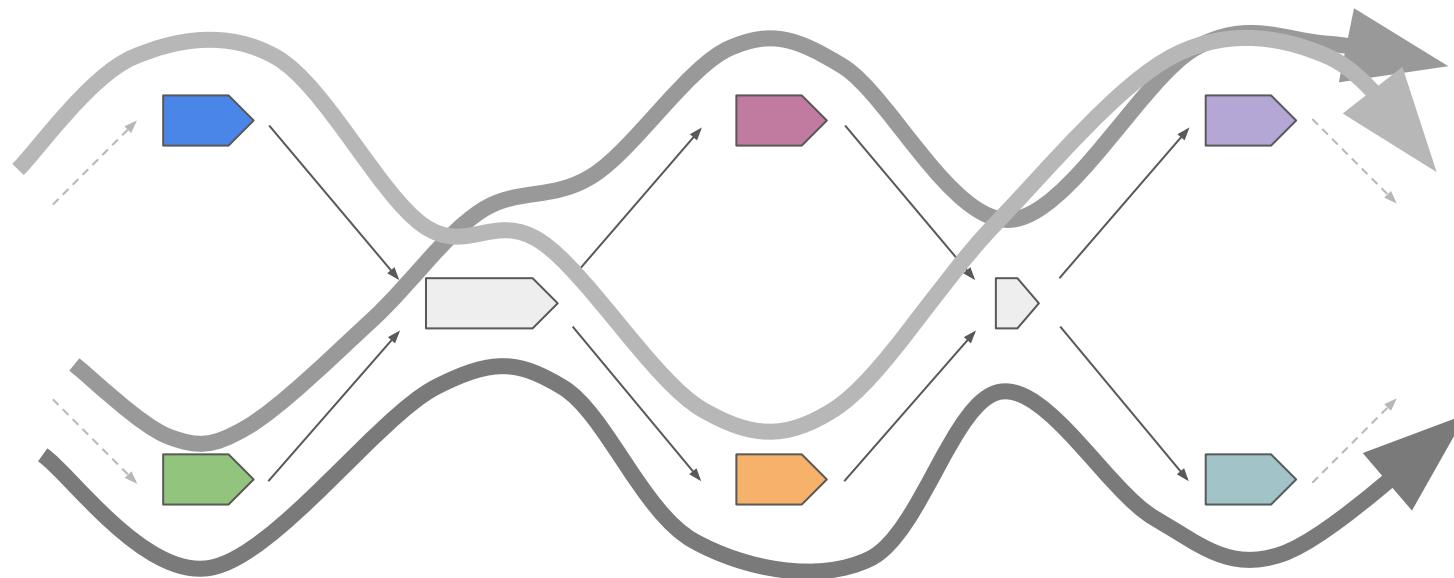
Observed
depths

$$\sum_p x_{pk} \beta_p \approx Y_k$$

KEY IDEA: The expected depth of a k-mer
is the sum of the paths that include that k-mer



KEY IDEA: The expected depth of a k-mer is the sum of the paths that include that k-mer

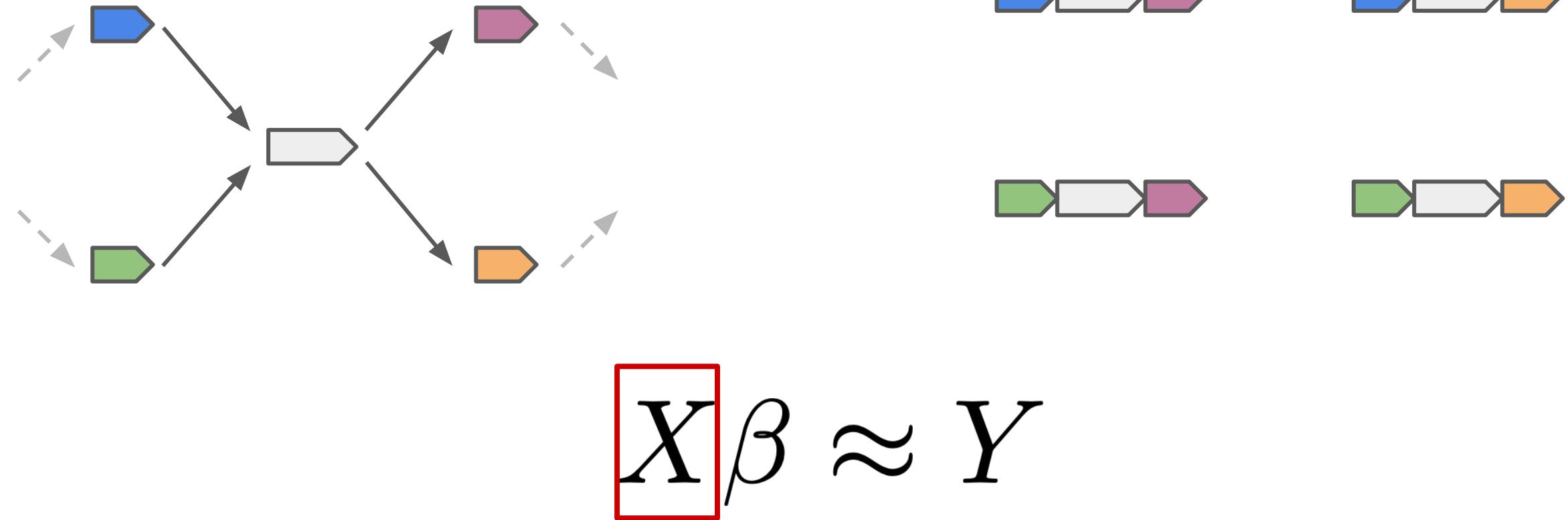


Deconvolution: Inferring the depth of these latent paths based on observed k-mer depths

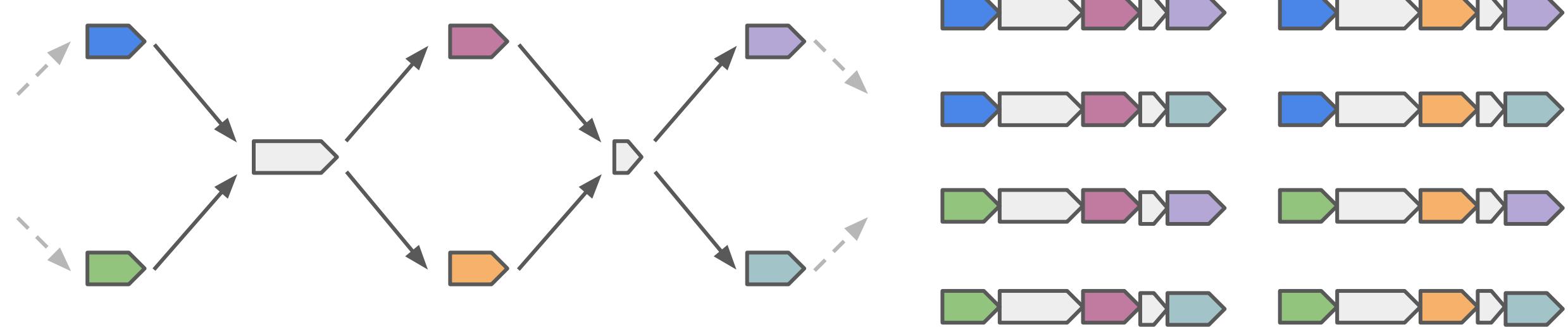
Estimate this
From these

$$X\beta \approx Y$$

We can enumerate all possible paths on our assembly graph

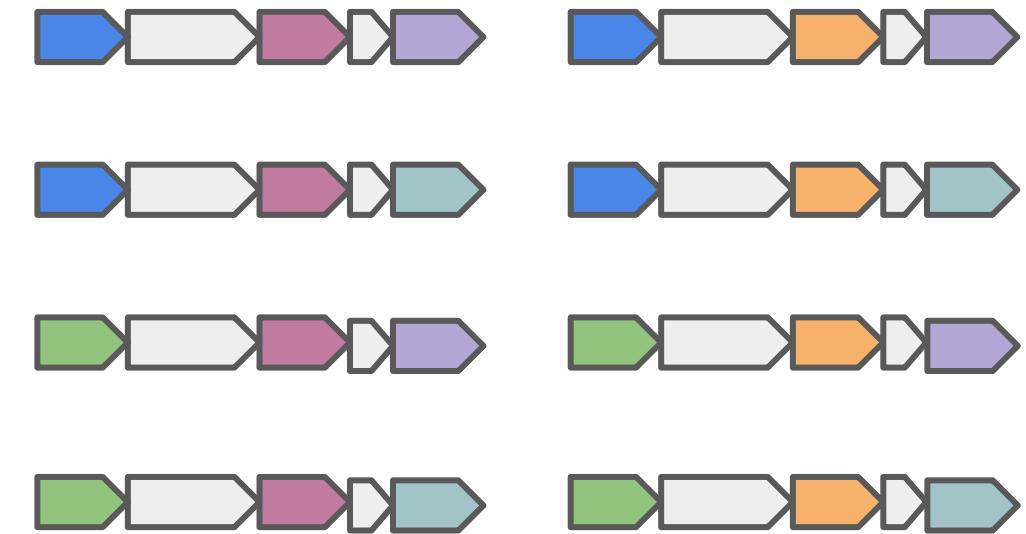
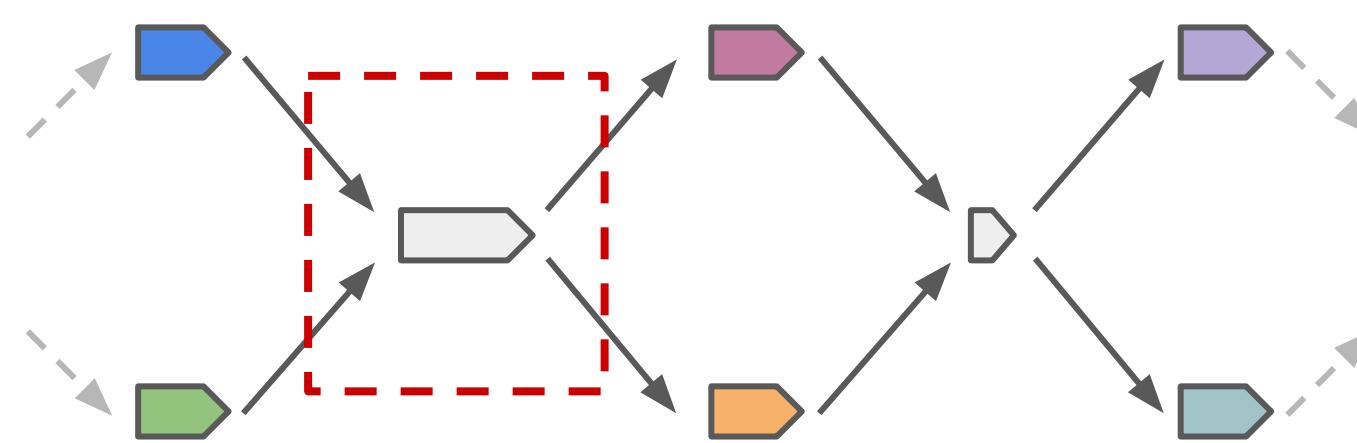


We can enumerate all possible paths on our assembly graph

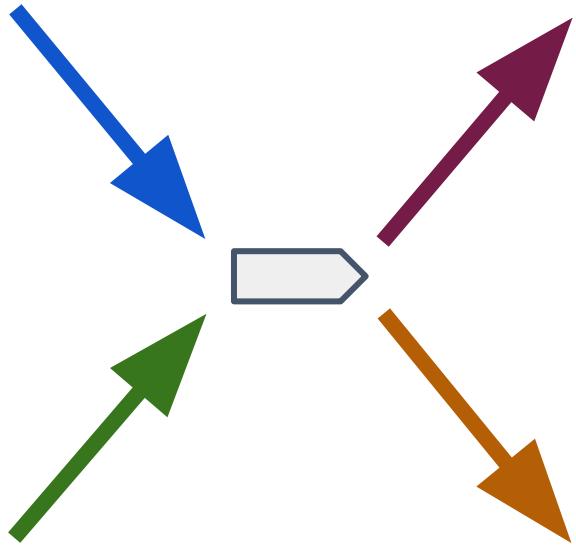


...but this grows exponentially with graph complexity

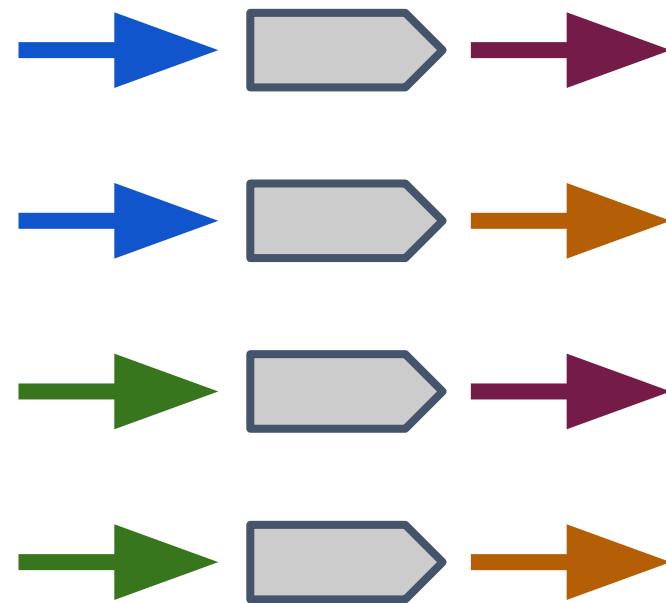
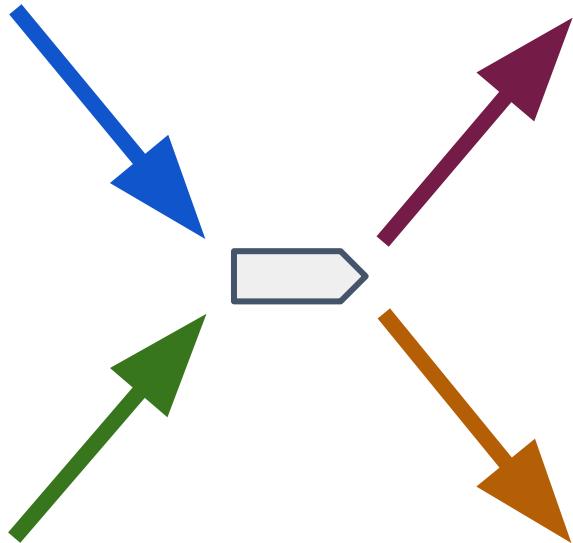
KEY IDEA: A single "junction" is the minimum unit of deconvolution



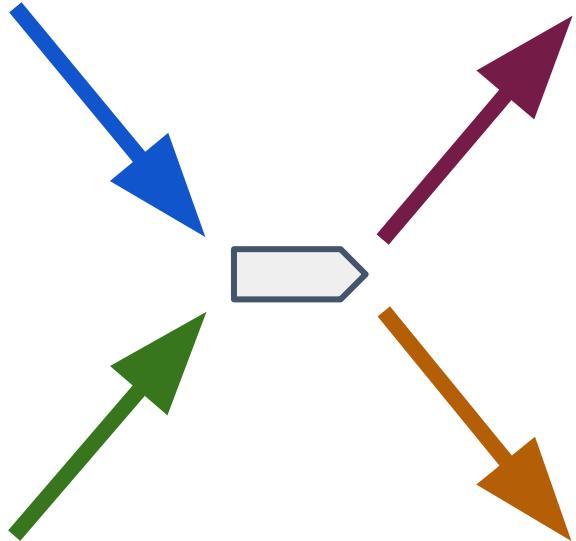
KEY IDEA: A single "junction" is the minimum unit of deconvolution



KEY IDEA: A single "junction" is the minimum unit of deconvolution



Focus on just one junction at a time
Quantify local paths

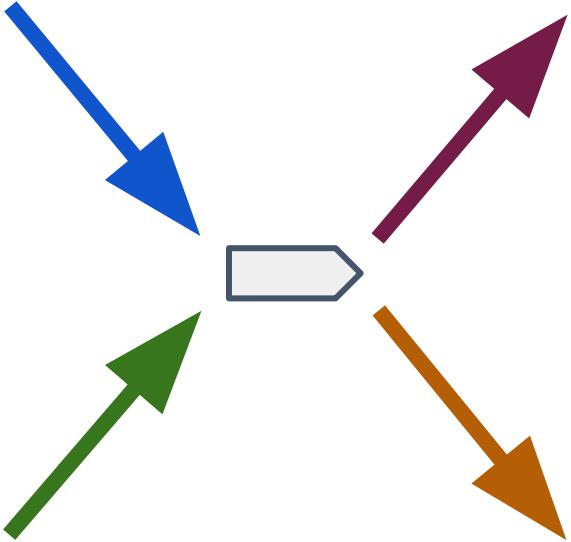


$$\begin{matrix} & \downarrow & \downarrow & \downarrow & \downarrow \\ \rightarrow & 1 & 1 & 0 & 0 \\ \text{green} & 0 & 0 & 1 & 1 \\ \text{purple} & 1 & 0 & 1 & 0 \\ \rightarrow & 0 & 1 & 0 & 1 \end{matrix} \times \begin{matrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{matrix} \approx \begin{matrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{matrix}$$

The diagram illustrates a linear regression model. On the left, a 4x4 matrix X is shown with colored arrows indicating specific entries: the first row has blue arrows pointing to the first two columns; the second row has green arrows pointing to the third and fourth columns; the third row has purple arrows pointing to the second, third, and fourth columns; and the fourth row has orange arrows pointing to the second, third, and fourth columns. To the right of X is a multiplication sign followed by a beta symbol (β). Below the beta symbol is a vector of parameters p_1, p_2, p_3, p_4 . To the right of the approximation symbol (\approx) is a vector of error terms e_1, e_2, e_3, e_4 .

Linear regression

Focus on just one junction at a time
Select (and quantify) local paths



$$\begin{matrix} & \downarrow & \downarrow & \downarrow & \downarrow \\ \rightarrow & \begin{matrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{matrix} & X & \times & \begin{matrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{matrix} & \beta & \approx & \begin{matrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{matrix} & Y \end{matrix}$$

The diagram illustrates a linear regression model selection process. It shows a matrix X with four columns and four rows, representing local paths. Above X are four downward arrows of different colors (blue, green, maroon, orange). To the right of X is a multiplication sign (\times). To the right of \times is a parameter vector β , with a red diagonal slash through it. To the right of β is an approximation symbol (\approx). To the right of \approx is a vector Y . The matrix X has values: Row 1: 1, 1, 0, 0. Row 2: 0, 0, 1, 1. Row 3: 1, 0, 1, 0. Row 4: 0, 1, 0, 1.

Linear regression
Model selection

Focus on just one junction at a time
Select (and quantify) local paths

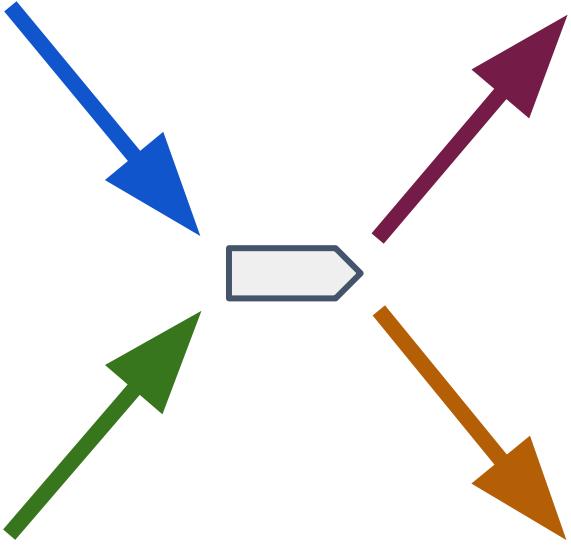
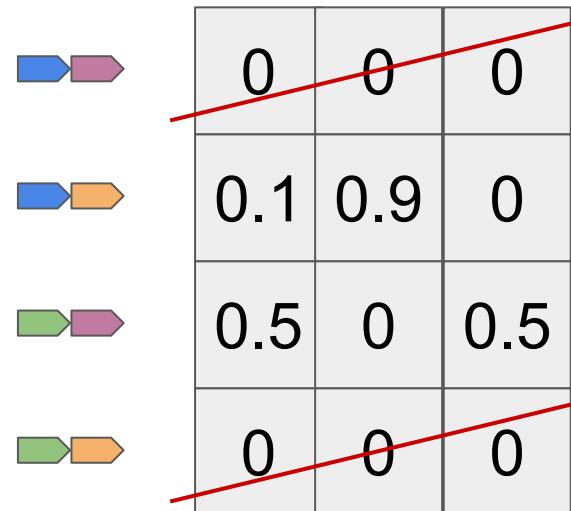
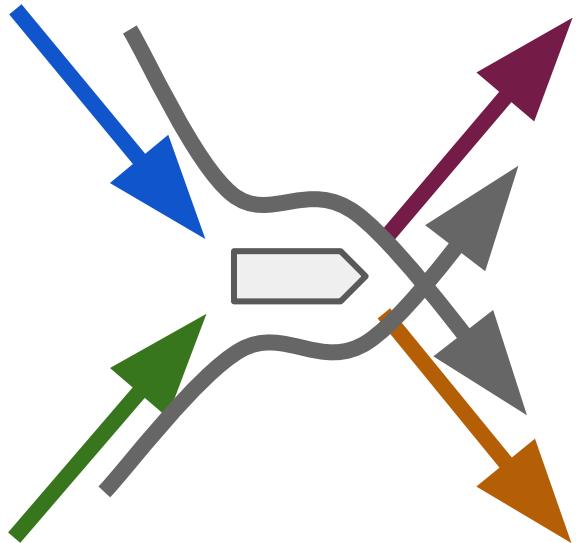


Diagram illustrating the matrix multiplication of two 4x3 matrices X and β , resulting in a 4x3 matrix Y . The first row of X has all zeros except for the second column. The second row of X has all zeros except for the third column. The third row of X has all zeros except for the first column. The fourth row of X has all zeros except for the fourth column. The first column of β has all zeros except for the second row. The second column of β has all zeros except for the third row. The third column of β has all zeros except for the first row. The result matrix Y has all zeros except for the second column, which contains the elements of the second row of β .

Linear regression Model selection Across multiple samples

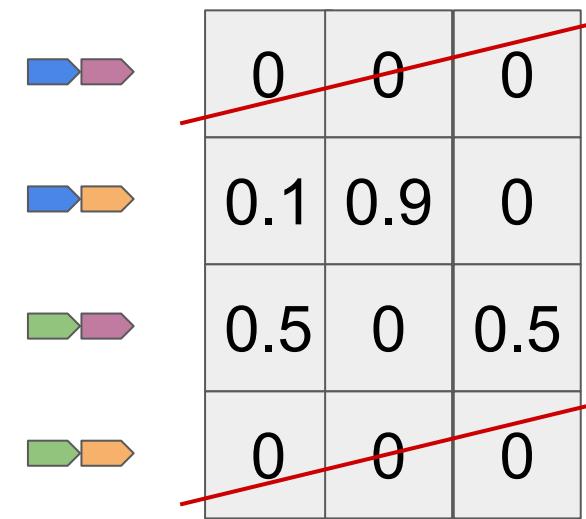
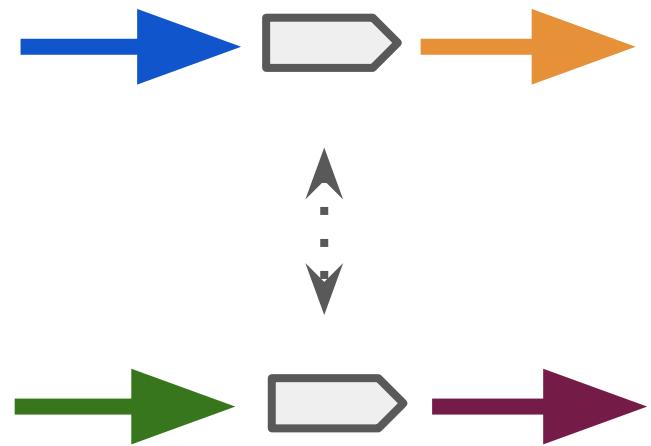
Drop paths with no depth in any sample



$$\hat{\beta}$$

Used statistical linkage to resolve ambiguity about
which of possible paths are "real"

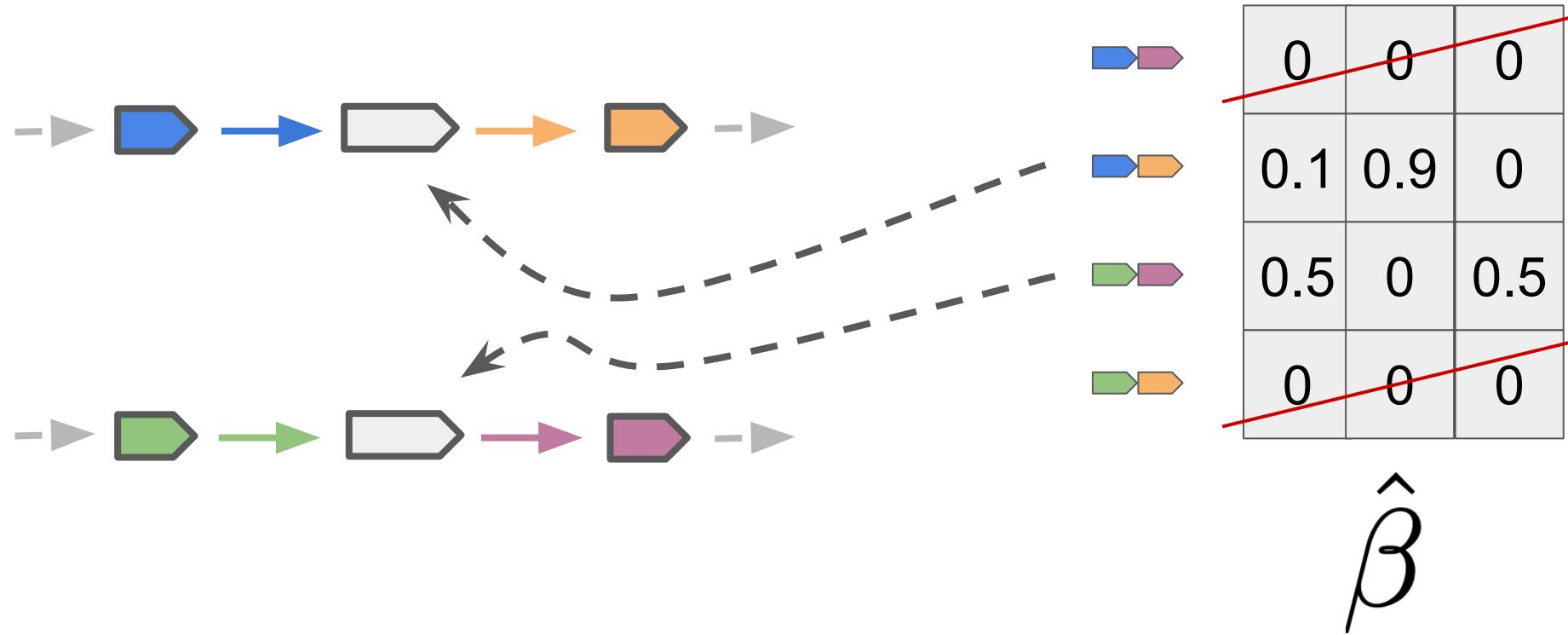
Resolve ambiguity, longer linear sequences



$\hat{\beta}$

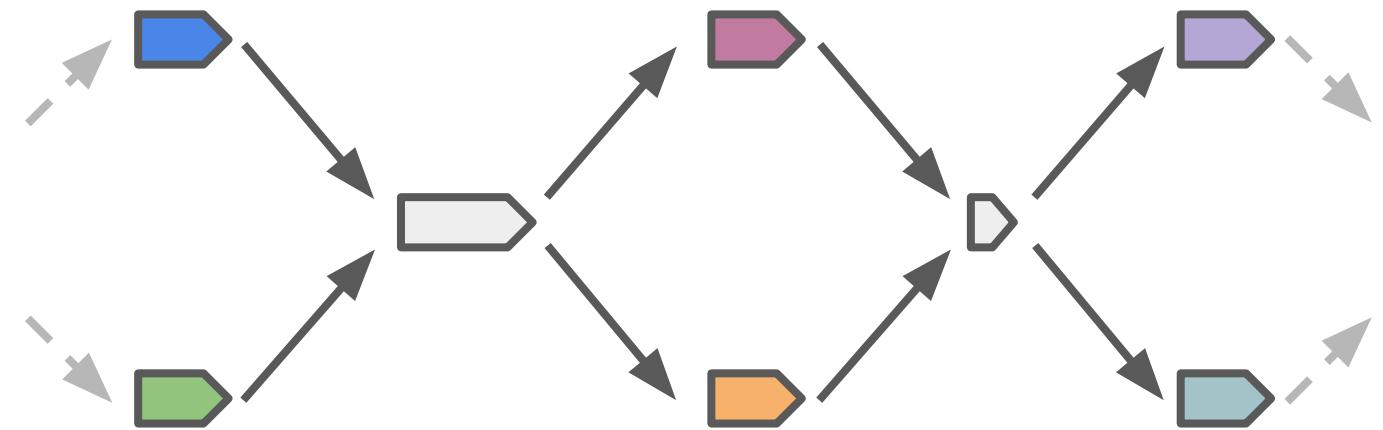
Can "unzip" this unitig into two paths

Resolve ambiguity, longer linear sequences

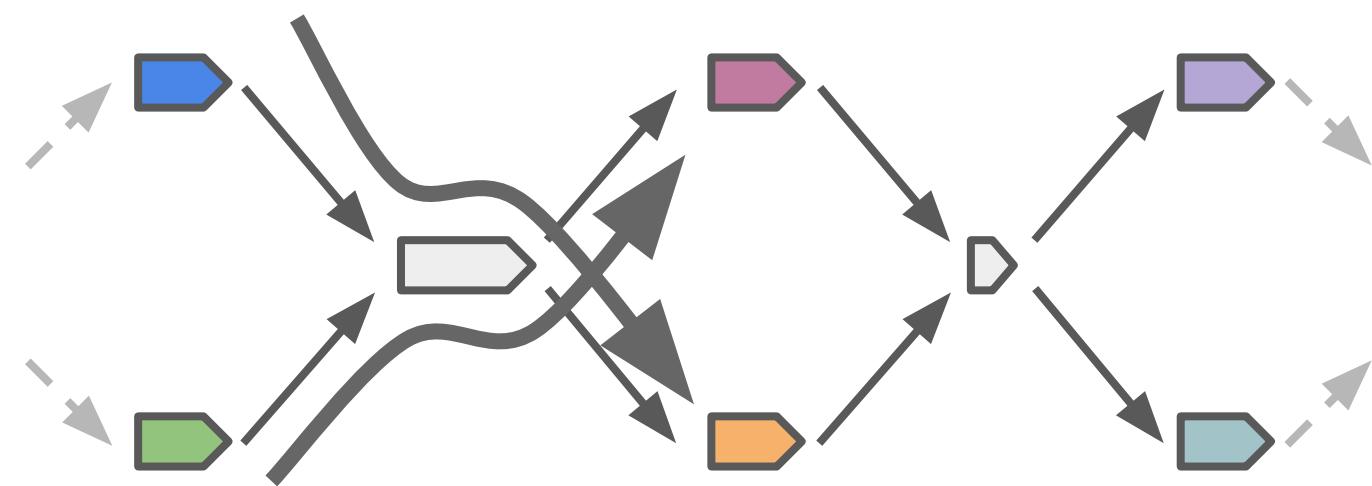


Newly split unitigs already have depths estimated across samples

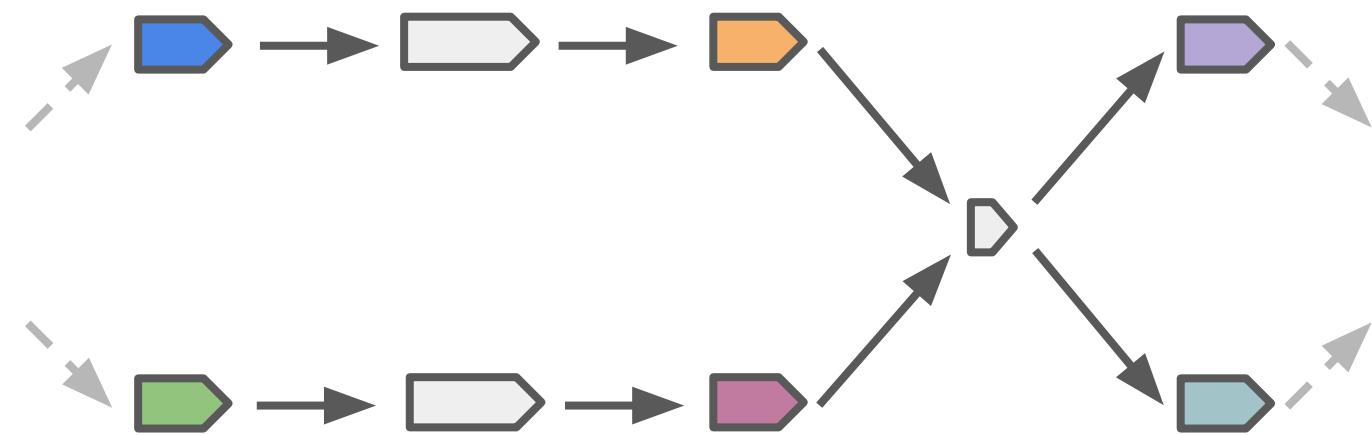
Iteratively unzipping local junctions



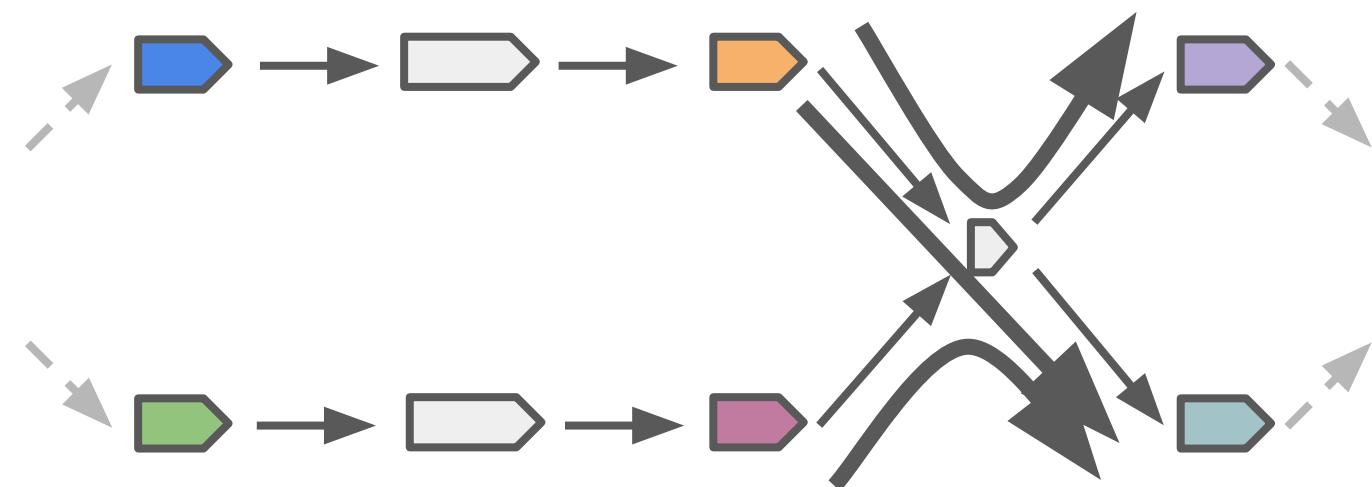
Iteratively unzipping local junctions



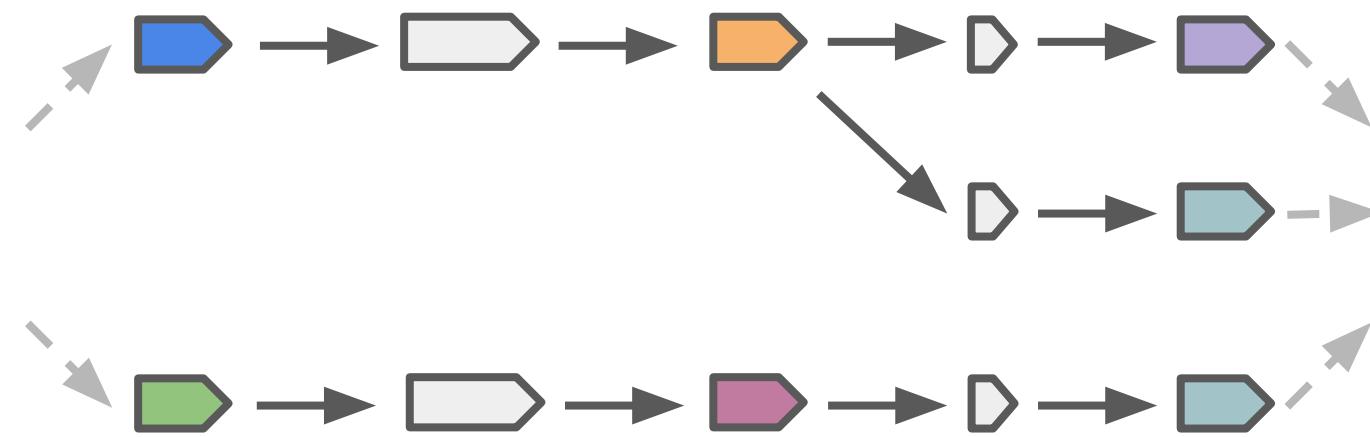
Iteratively unzipping local junctions



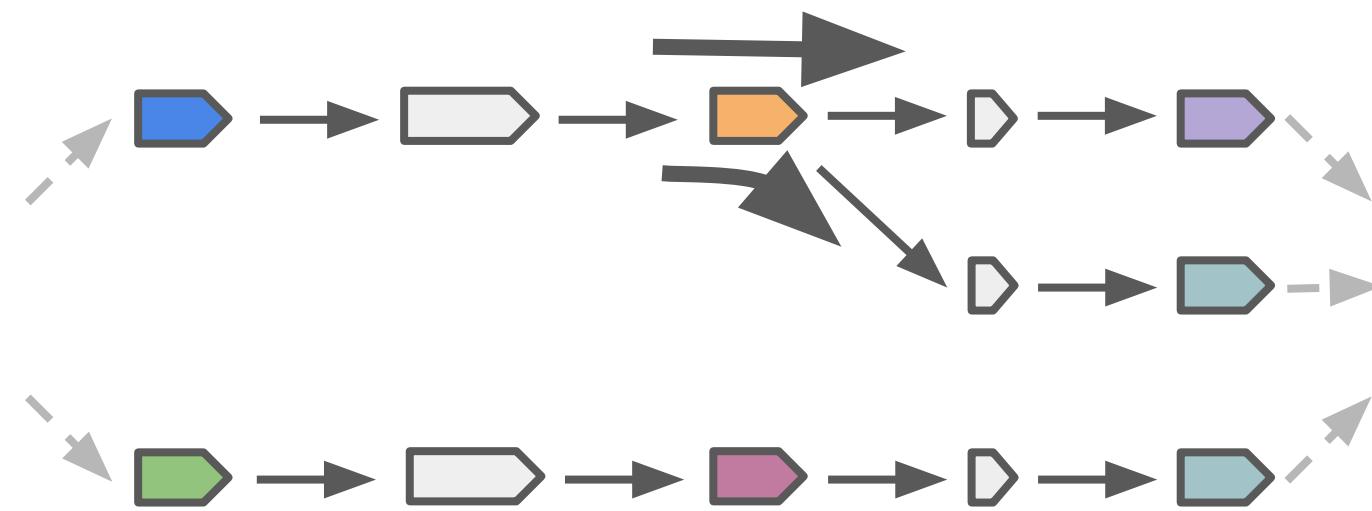
Iteratively unzipping local junctions



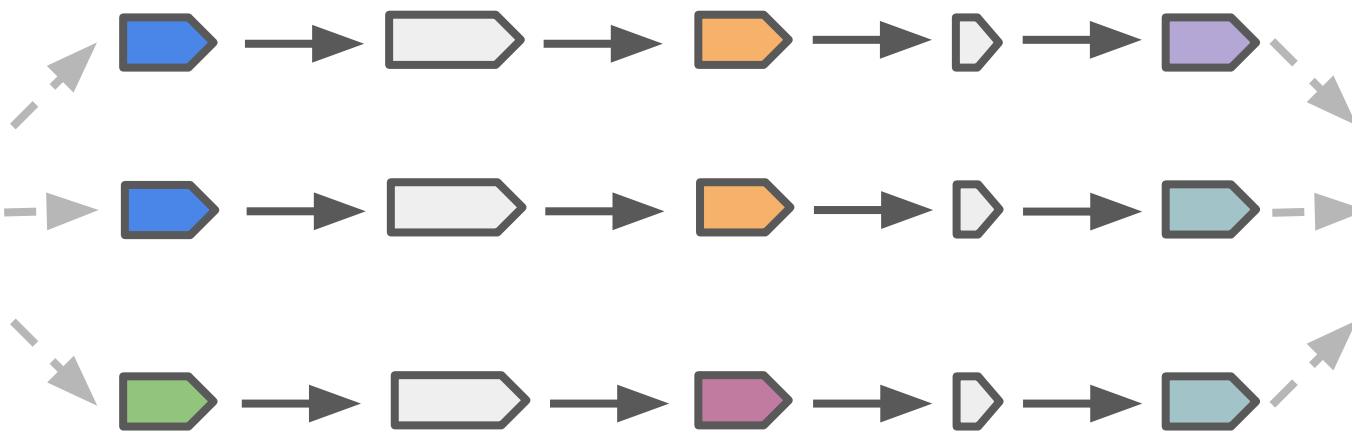
Iteratively unzipping local junctions



Iteratively unzipping local junctions



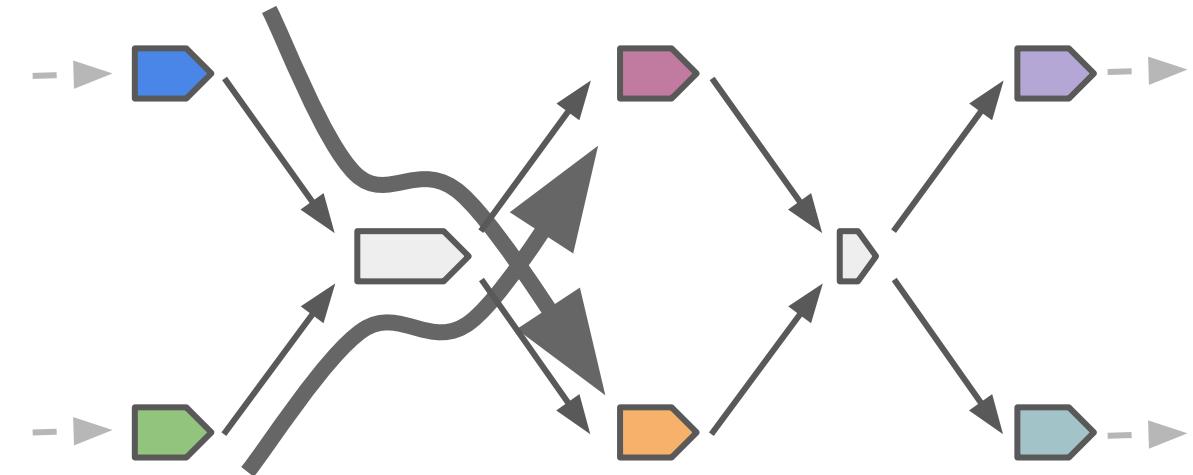
Iteratively unzipping local junctions



StrainZip

Assembly Graph Deconvolution for
Quantification of Strain-Specific
Sequences across Metagenomes

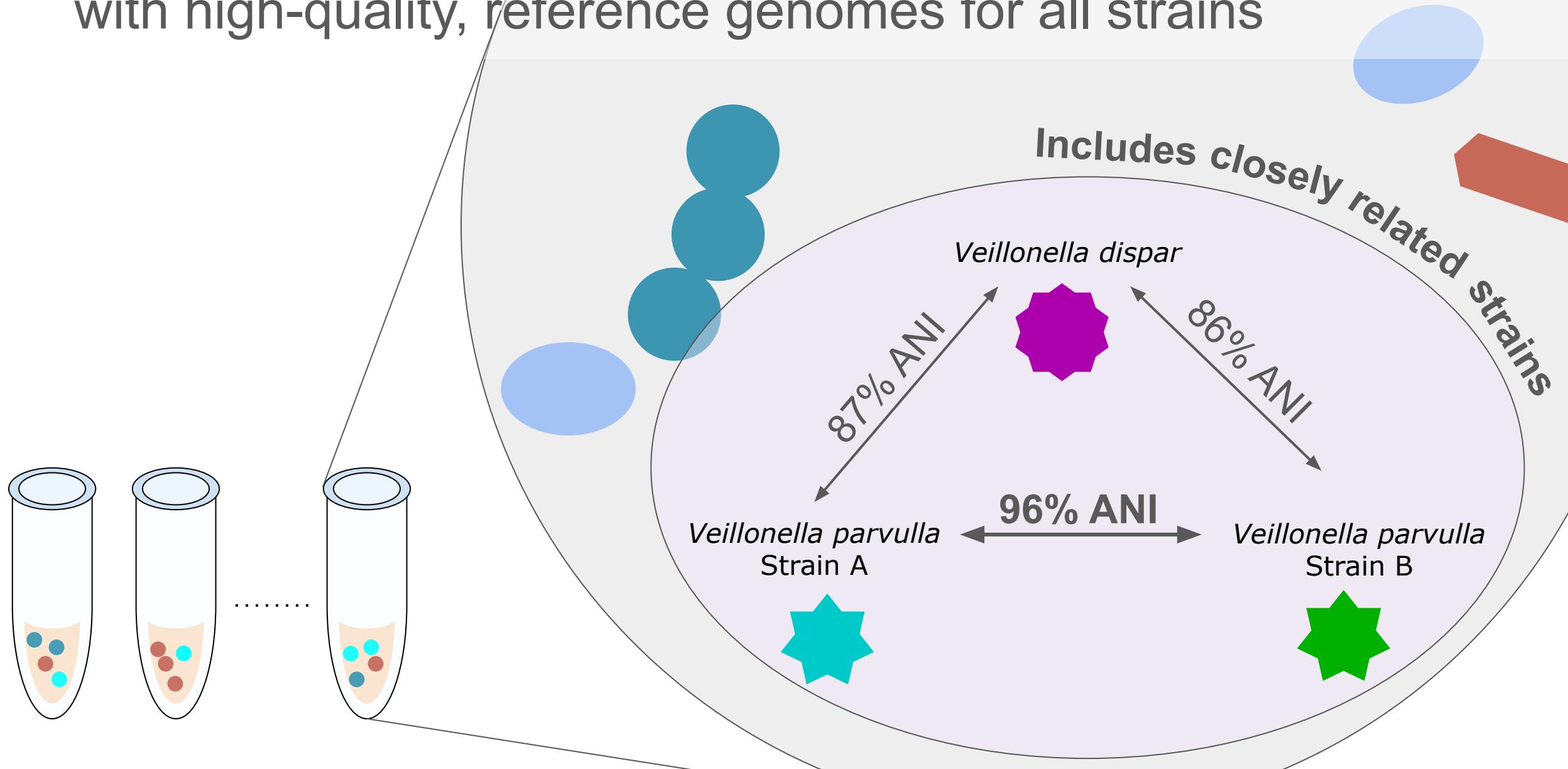
<https://github.com/bsmith89/StrainZip>

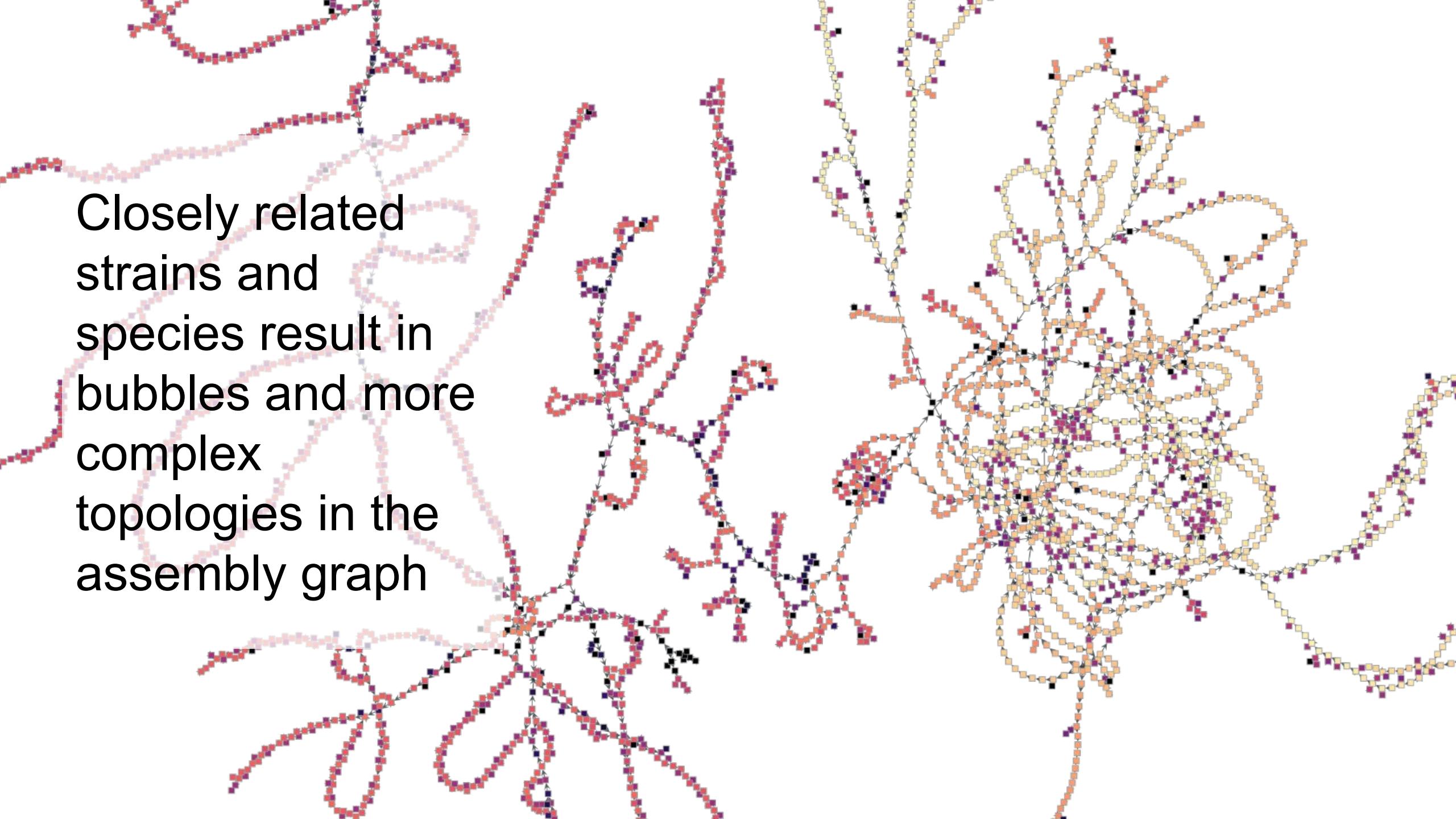


$$\begin{array}{c} \textcolor{violet}{\downarrow} \quad \textcolor{blue}{\downarrow} \quad \textcolor{green}{\downarrow} \quad \textcolor{orange}{\downarrow} \\ \textcolor{blue}{\rightarrow} \quad \begin{matrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{matrix} \\ \times \\ \begin{matrix} p_{1,1} & p_{1,2} & p_{1,3} \\ p_{2,1} & p_{2,2} & p_{2,3} \\ p_{3,1} & p_{3,2} & p_{3,3} \\ p_{4,1} & p_{4,2} & p_{4,3} \end{matrix} \\ \approx \\ \begin{matrix} e_{1,1} & e_{1,2} & e_{1,3} \\ e_{2,1} & e_{2,2} & e_{2,3} \\ e_{3,1} & e_{3,2} & e_{3,3} \\ e_{4,1} & e_{4,2} & e_{4,3} \end{matrix} \end{array}$$

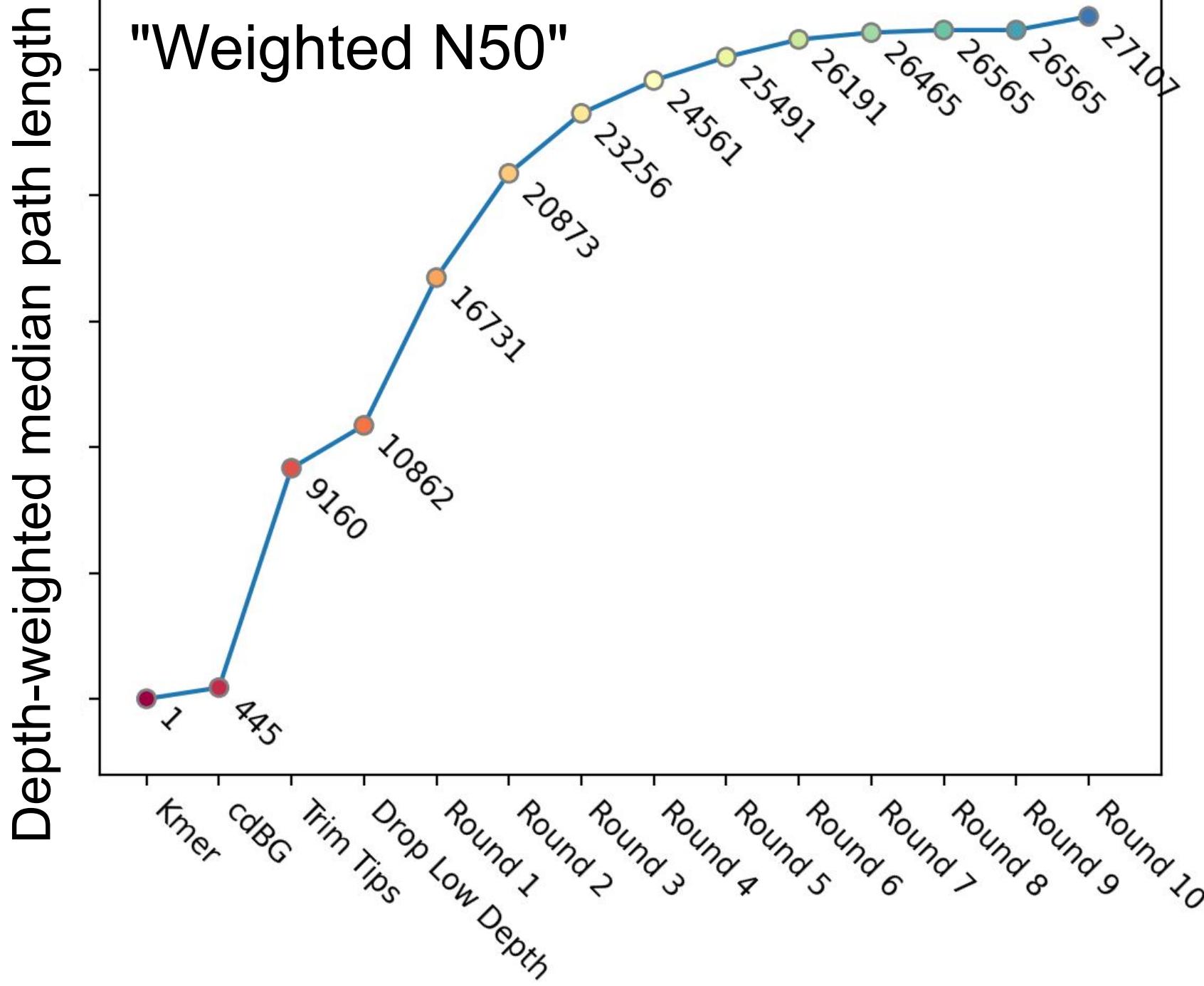
Benchmarking

hCOM2 is a complex (125 species), synthetic community with high-quality, reference genomes for all strains





Closely related strains and species result in bubbles and more complex topologies in the assembly graph

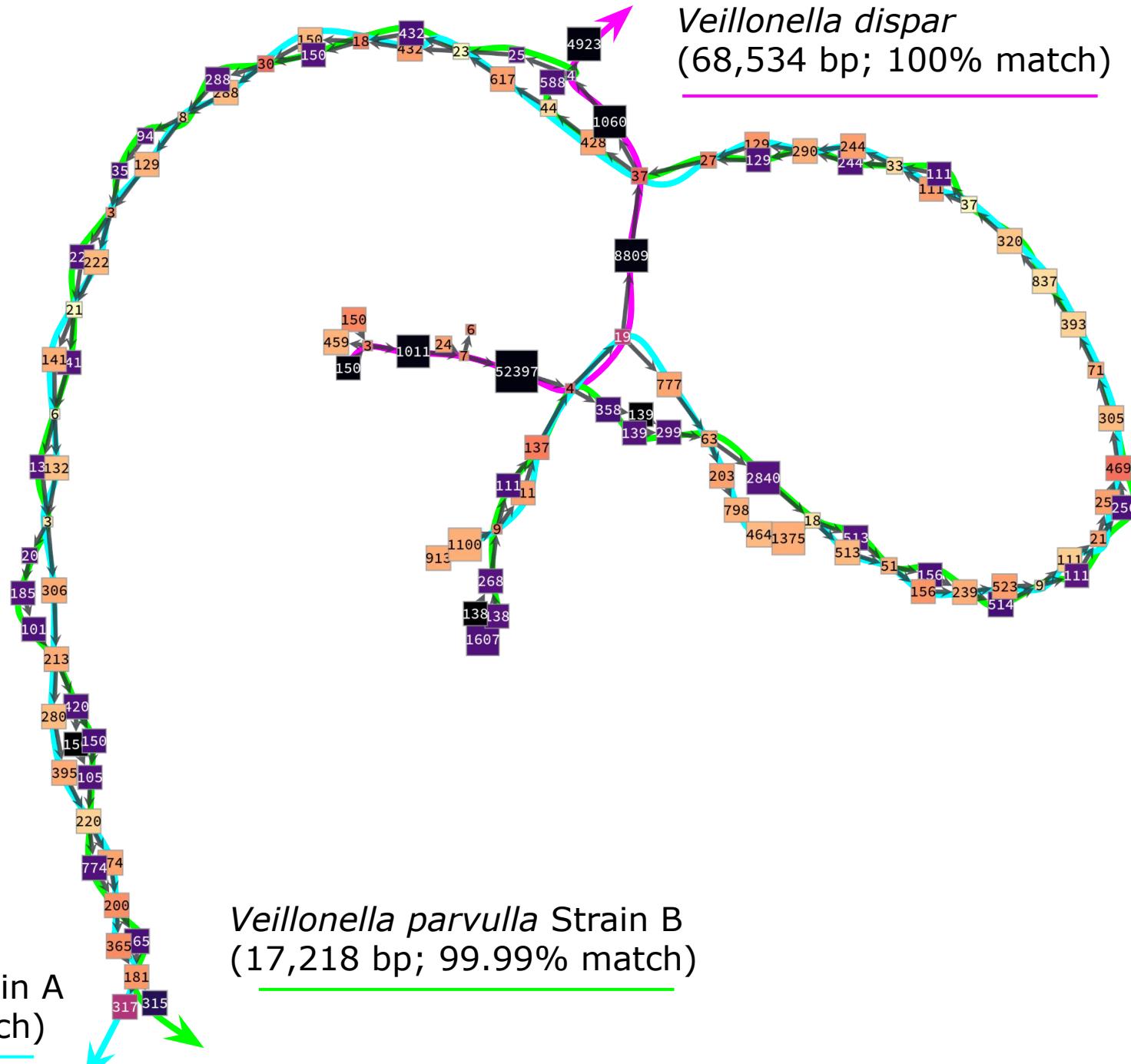


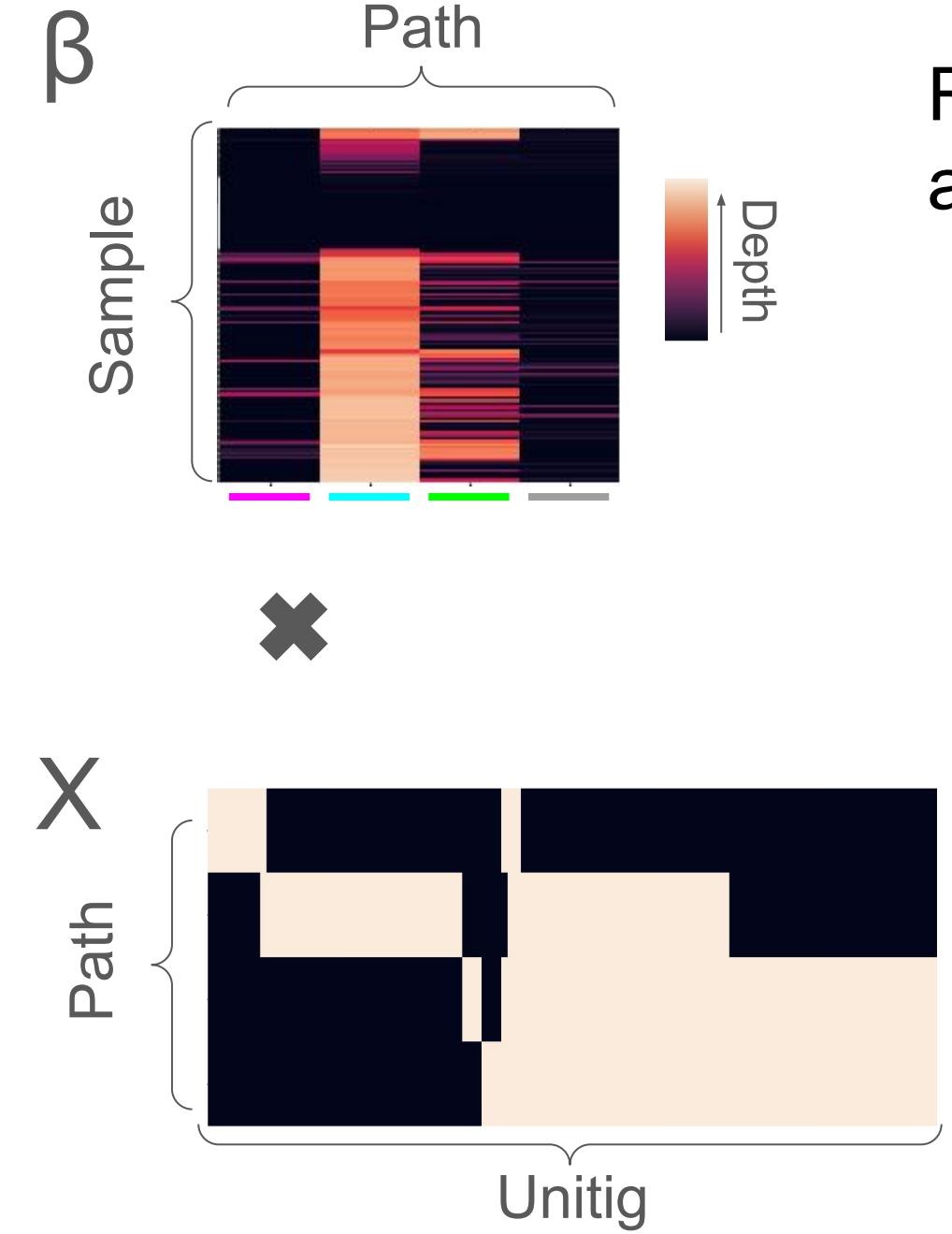
Path lengths
increase over
successive rounds
of deconvolution

Deconvolution
recovers longer,
strain-specific
sequences

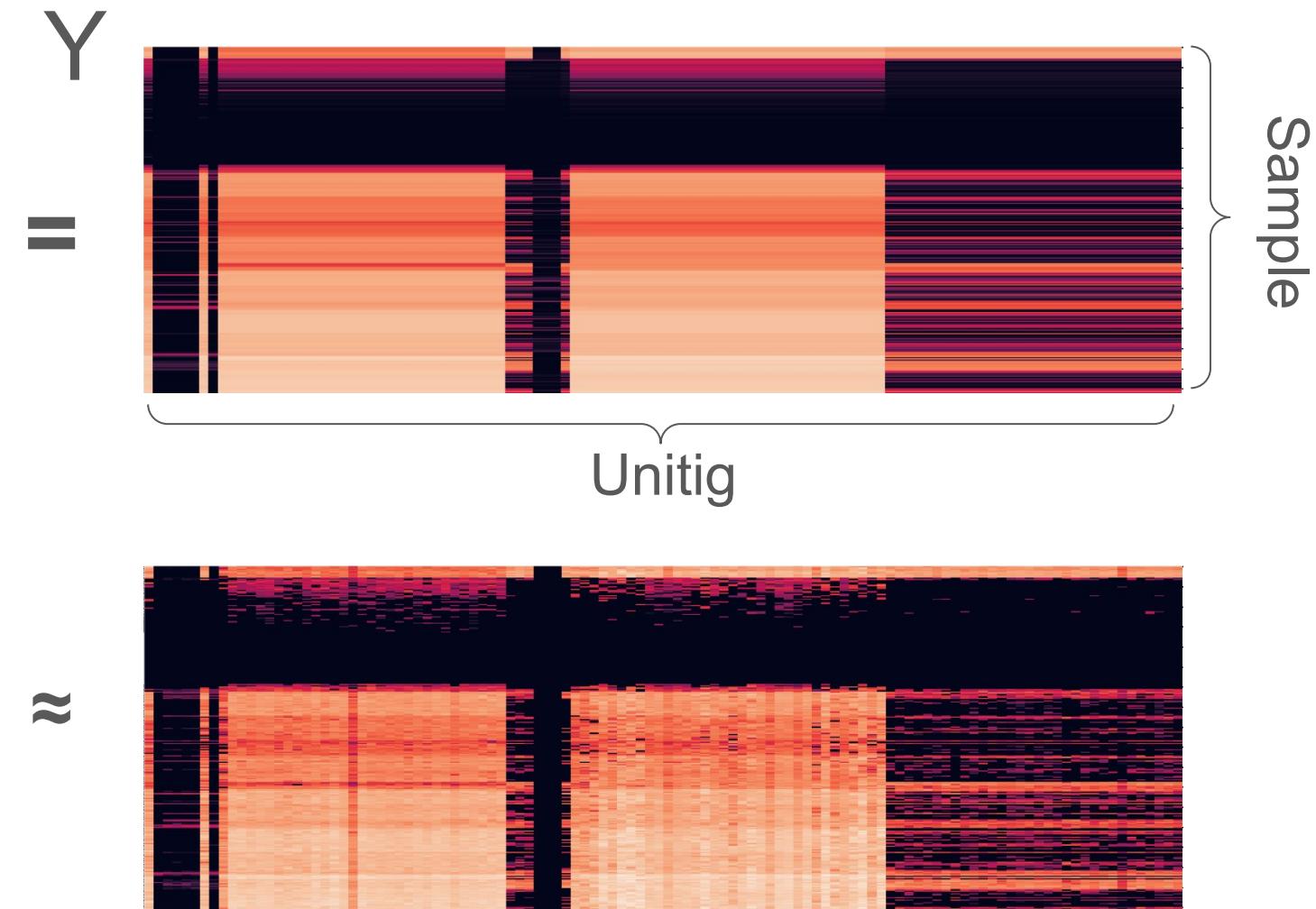
...including
lower-abundance
strains
...and species
...accurately

Veillonella parvulla Strain A
(17,229 bp; 100% match)



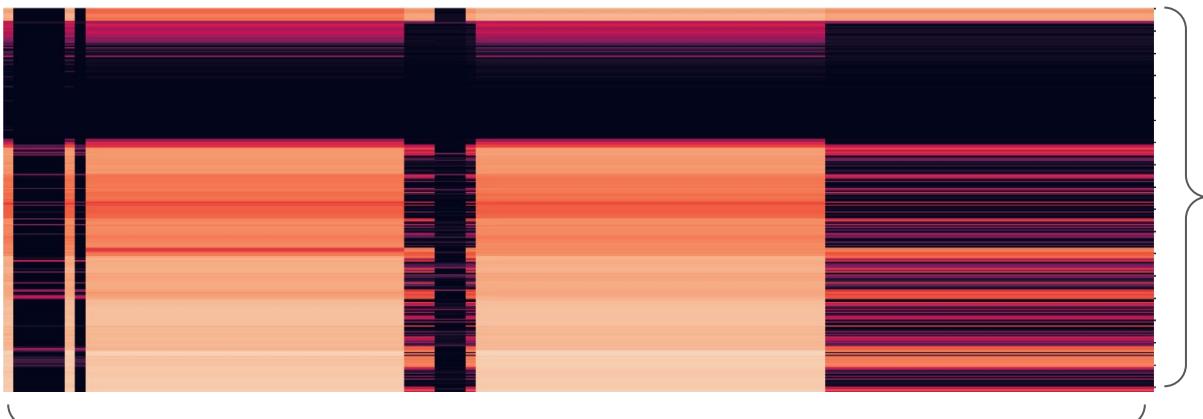


Result: both paths, and path depths
across samples (without read mapping)

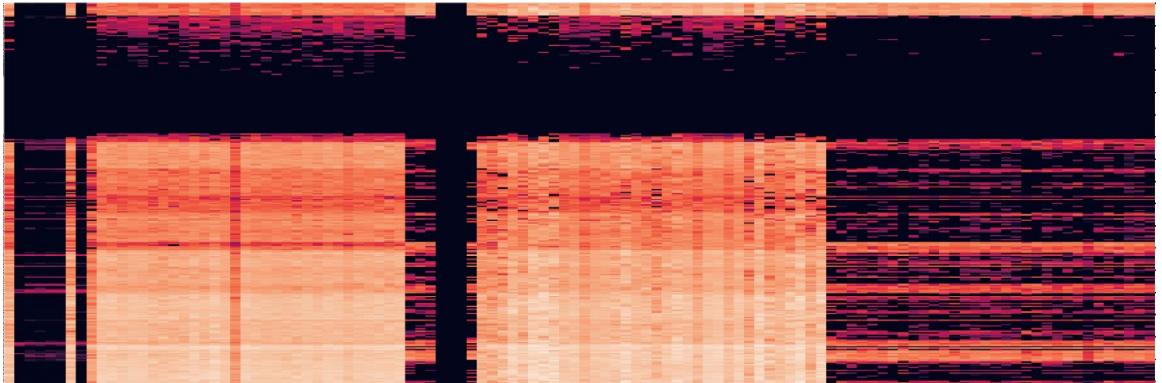


Estimated
unitig
depths
closely
match
observed
depths

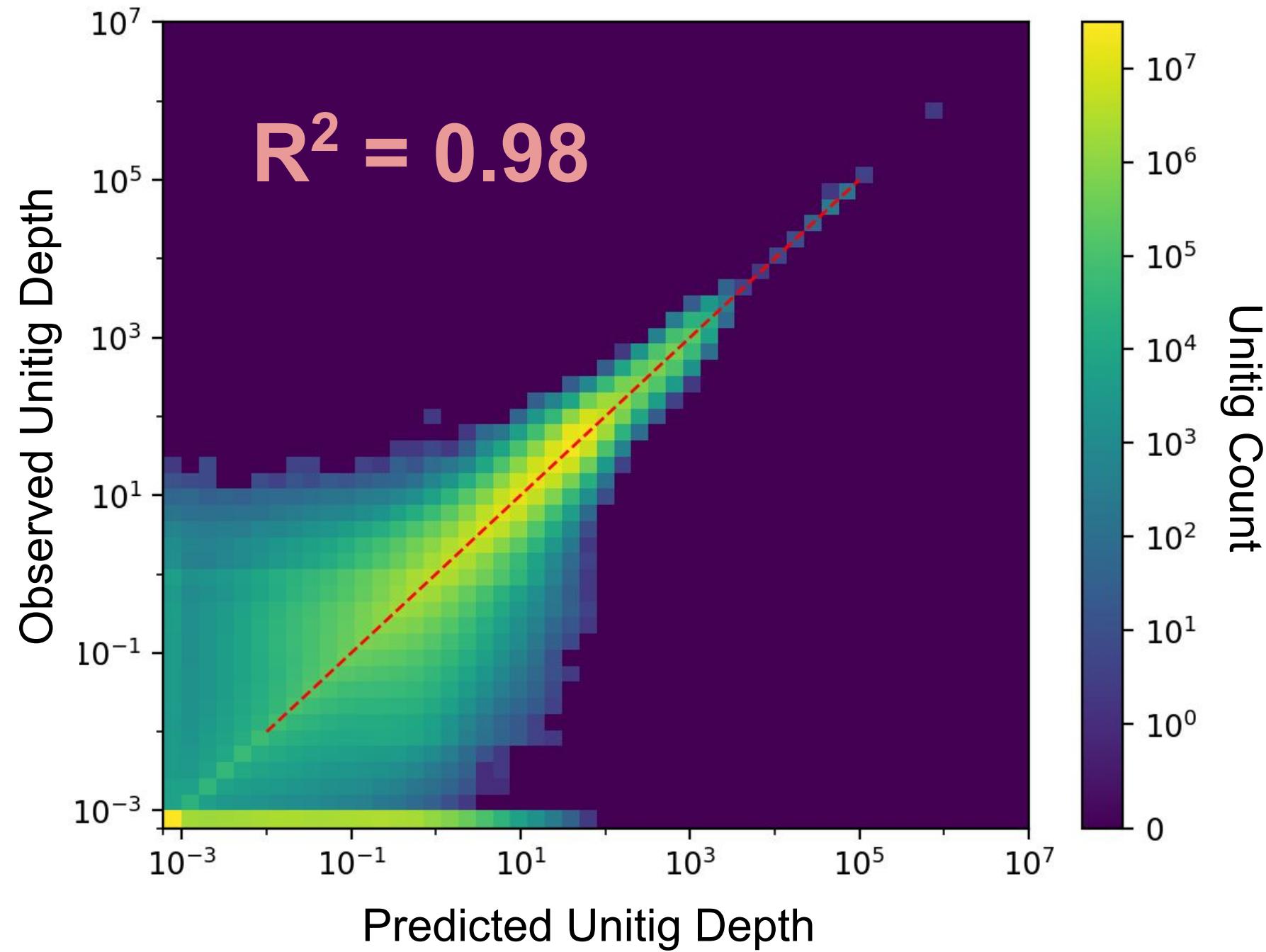
Predicted →



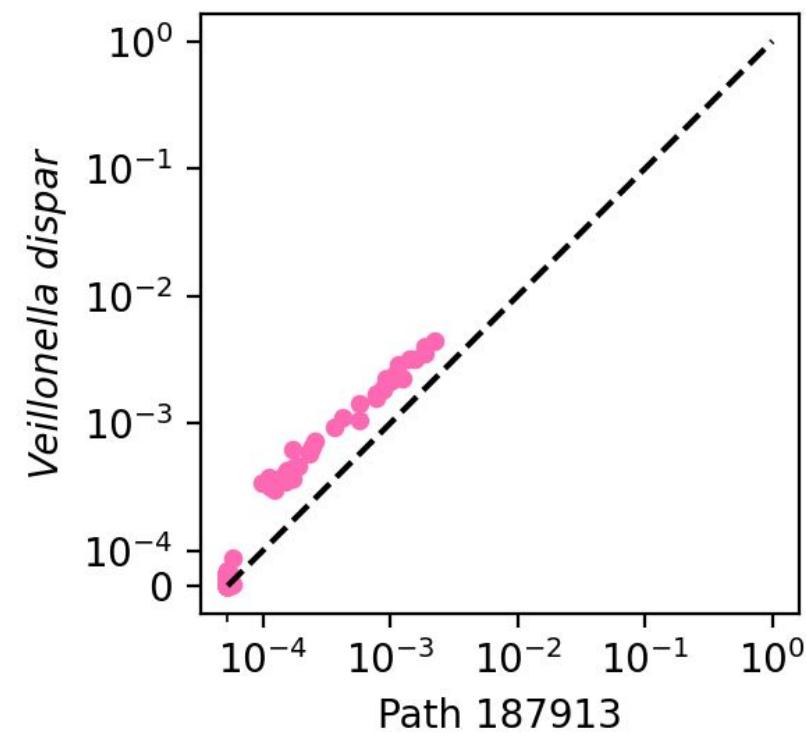
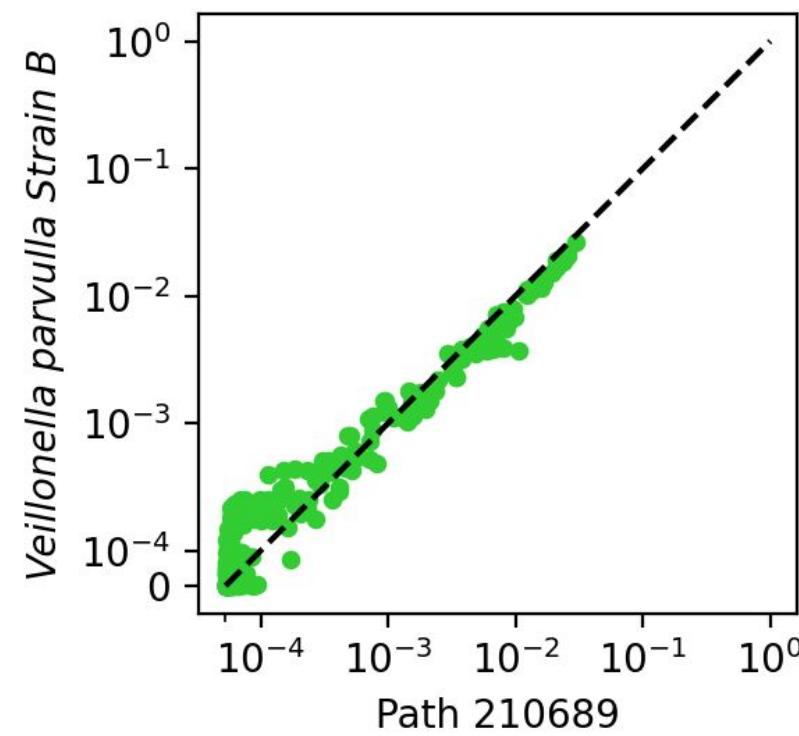
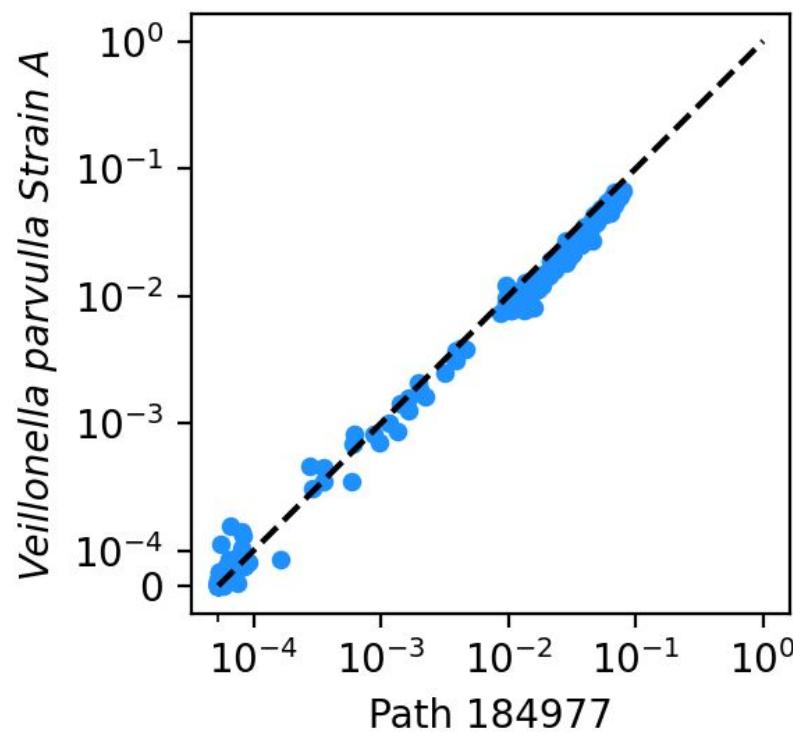
Observed →



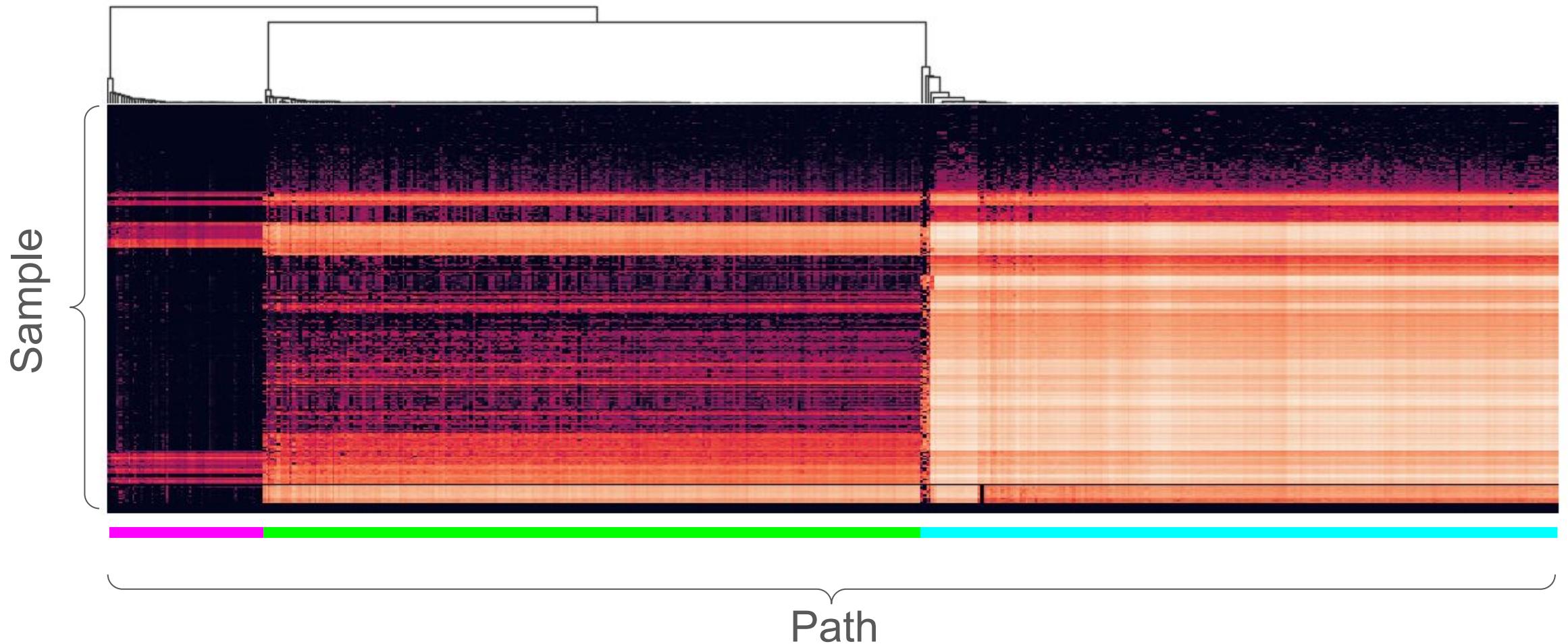
Estimated
unitig
depths
closely
match
observed
depths



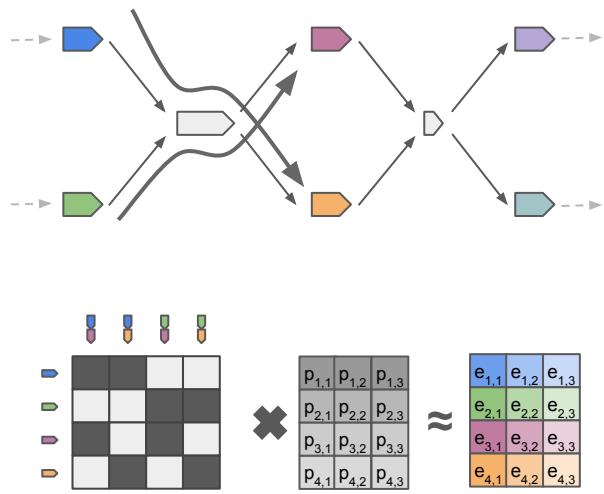
Path depths match reference-based strain depth estimation



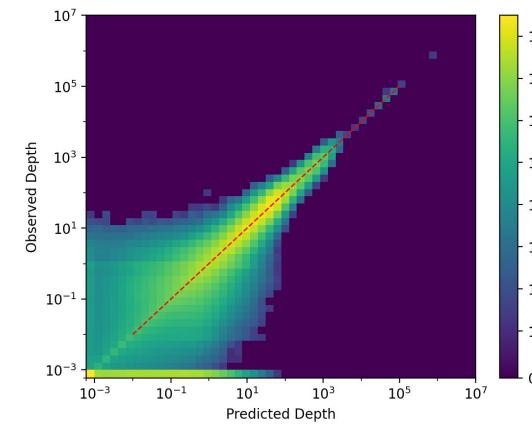
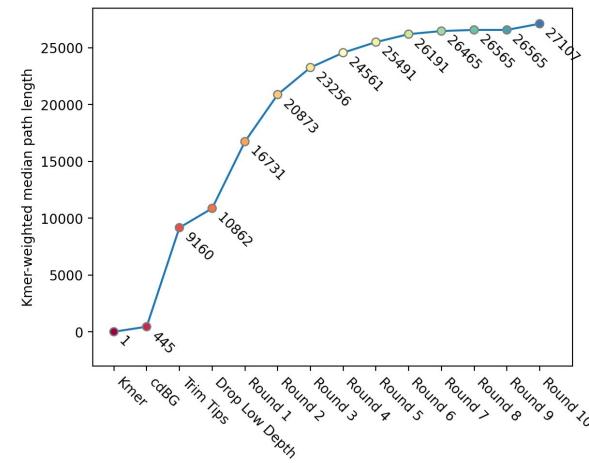
Clustering paths by depth combines multiple sequences from the same strain



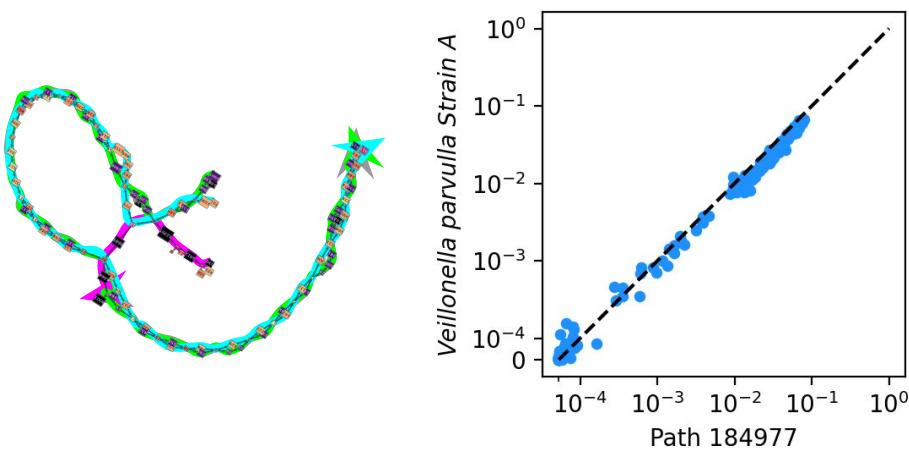
Iterative Junction Deconvolution



Combines Assembly, Depth Estimation



Recovers Closely Related Genomes



Enables Strain-Resolved Metagenomics

