Gladstone Scientific Retreat 2024

Unzipping the metagenome:

strain-level discovery in the gut microbiome

Byron J. Smith

Bioinformatics Fellow

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

CHAN ZUCKERBERG BIOHUB

NIH National Institutes of Health

GLADSTONE INSTITUTES

Gladstone Scientific Retreat 2024

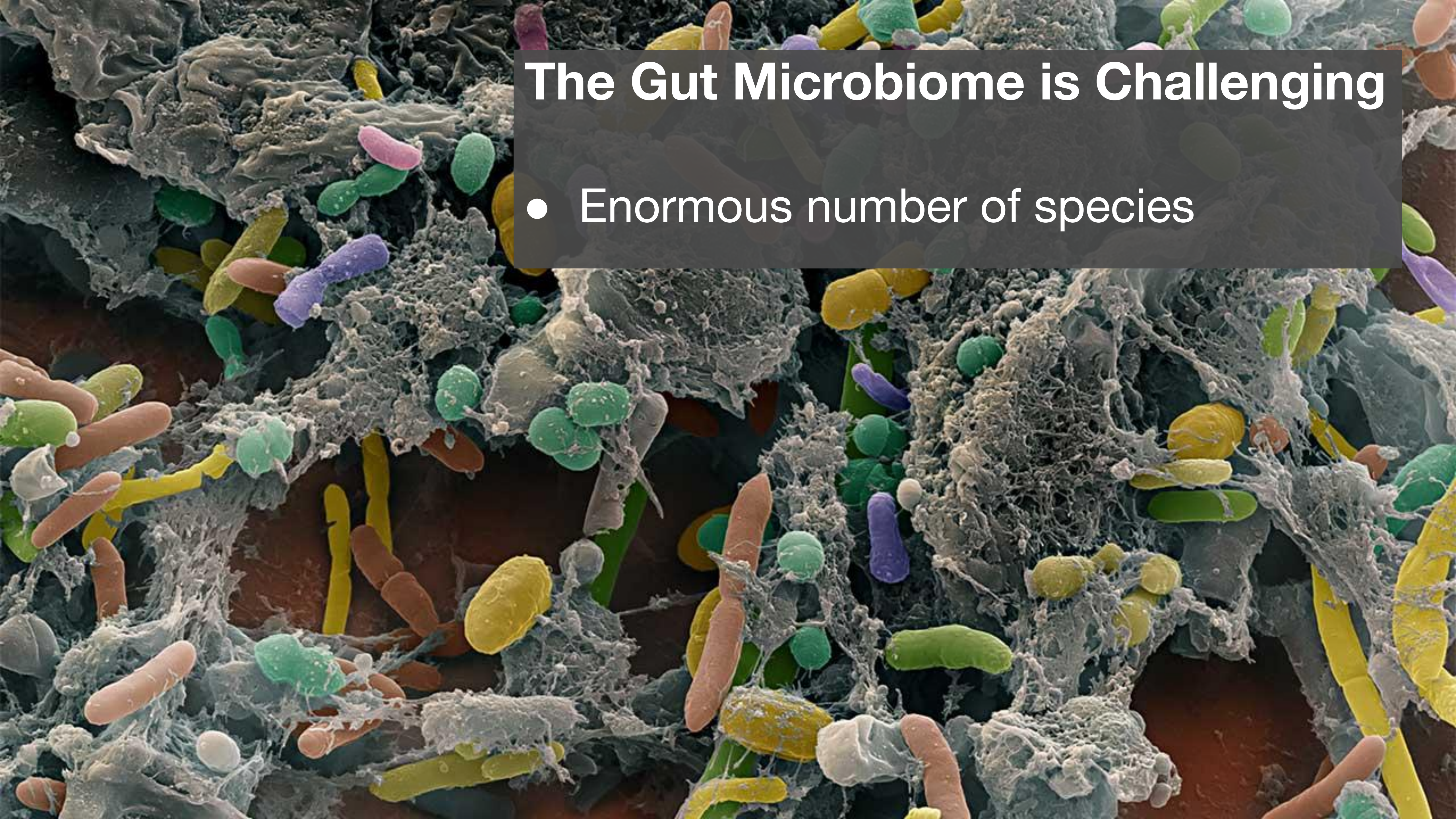
Introduction:

The gut microbiome and shotgun metagenomics

The Gut Microbiome is Challenging

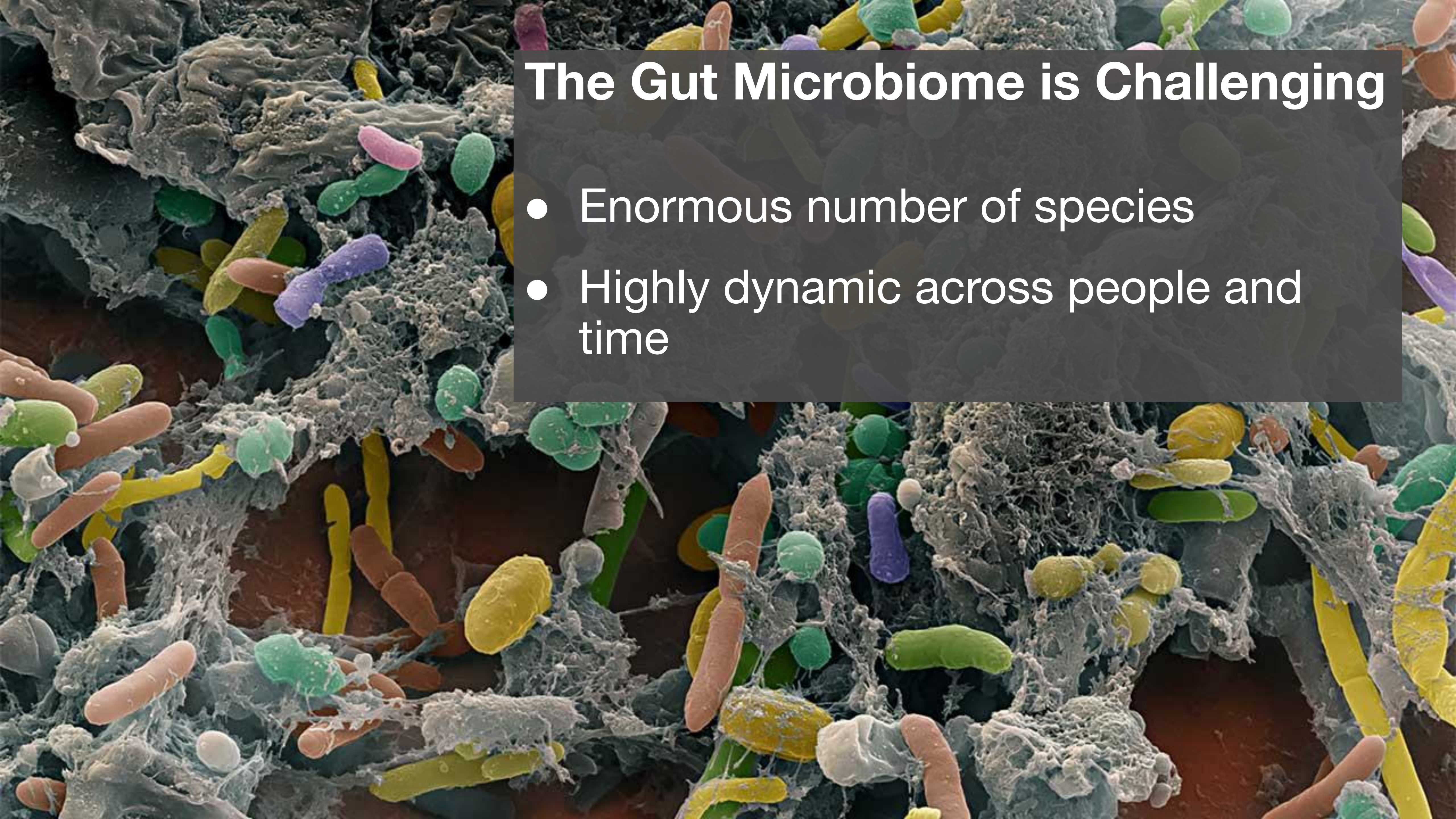
The Gut Microbiome is Challenging

- Enormous number of species



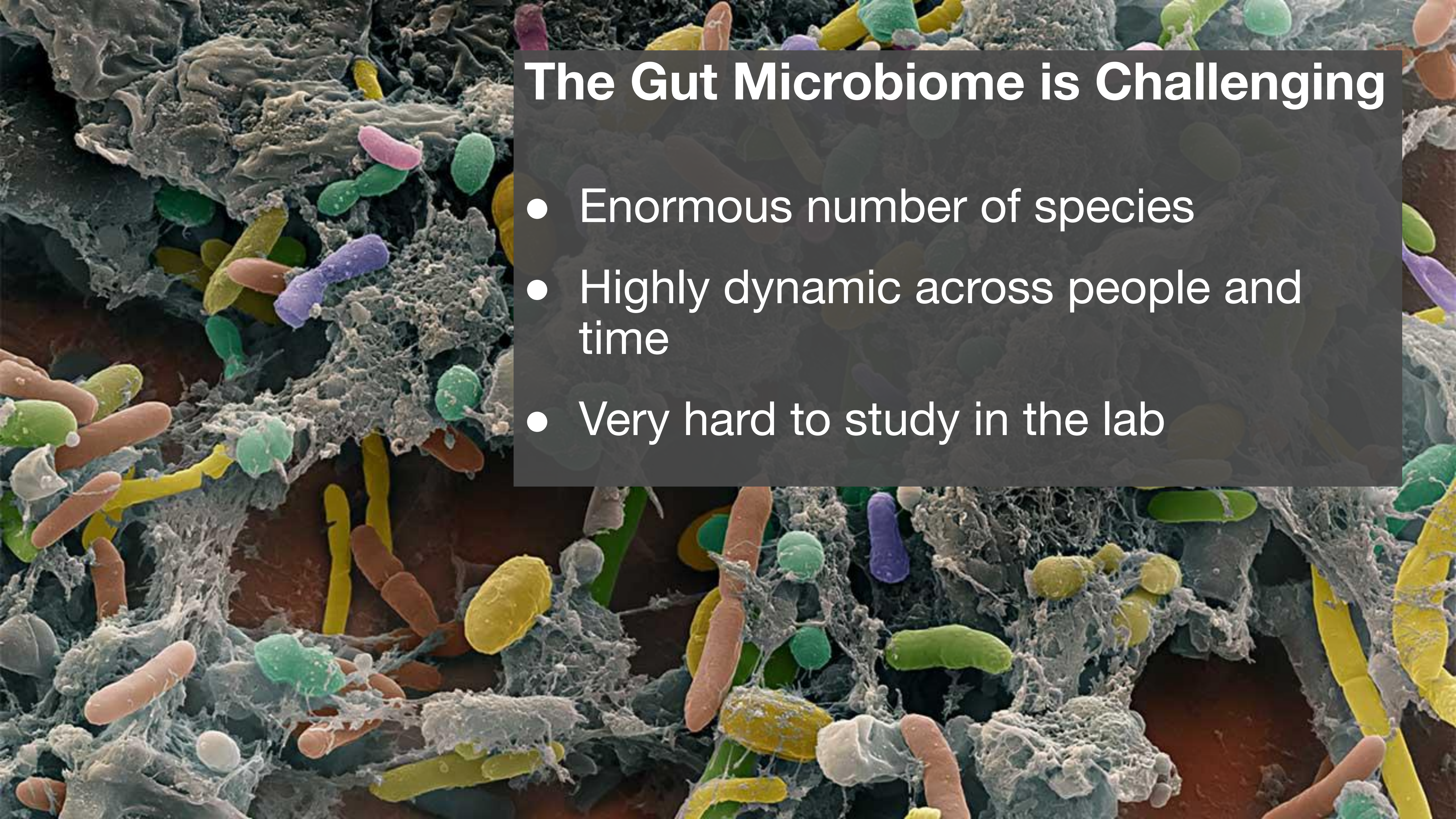
The Gut Microbiome is Challenging

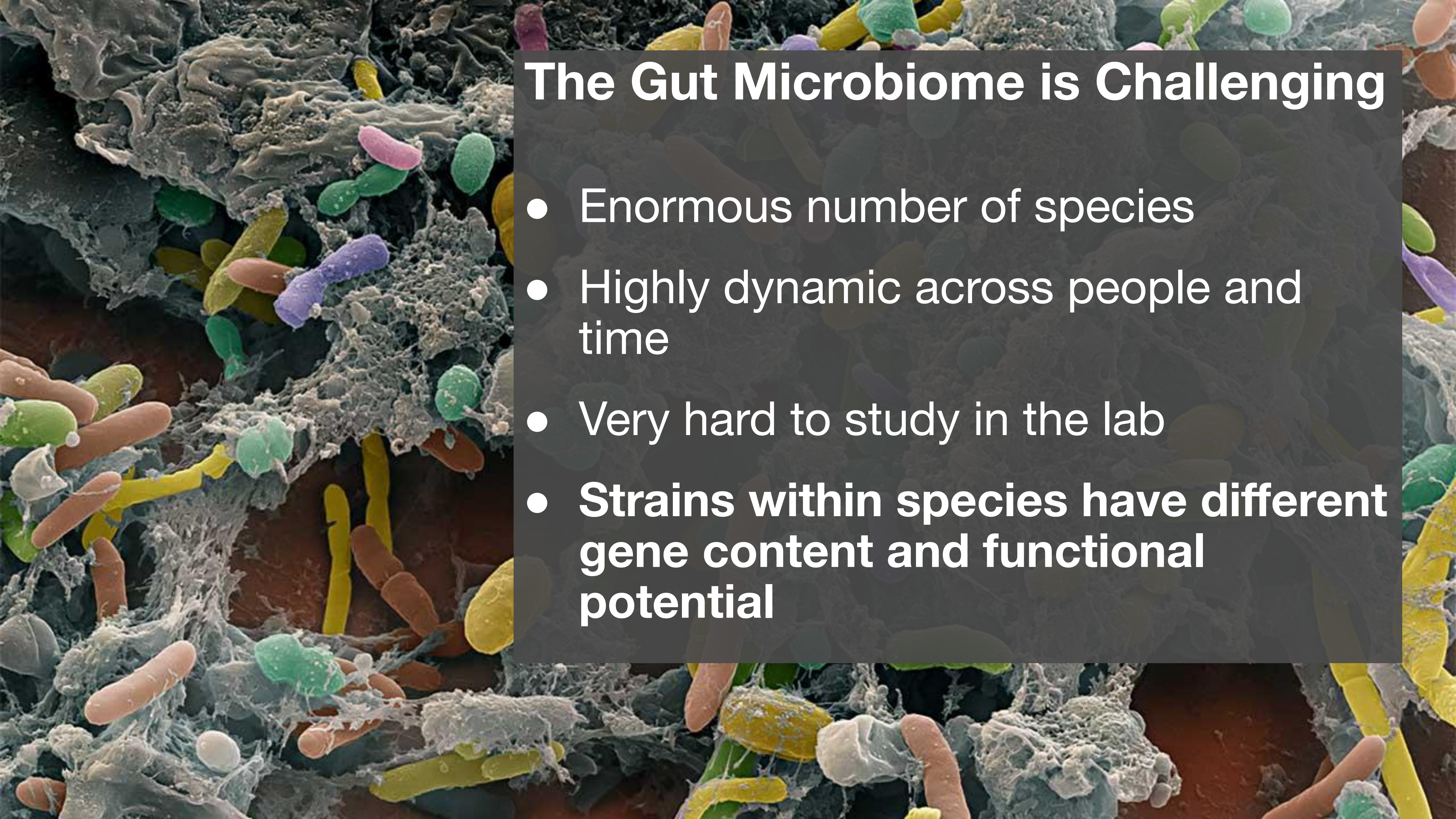
- Enormous number of species
- Highly dynamic across people and time



The Gut Microbiome is Challenging

- Enormous number of species
- Highly dynamic across people and time
- Very hard to study in the lab



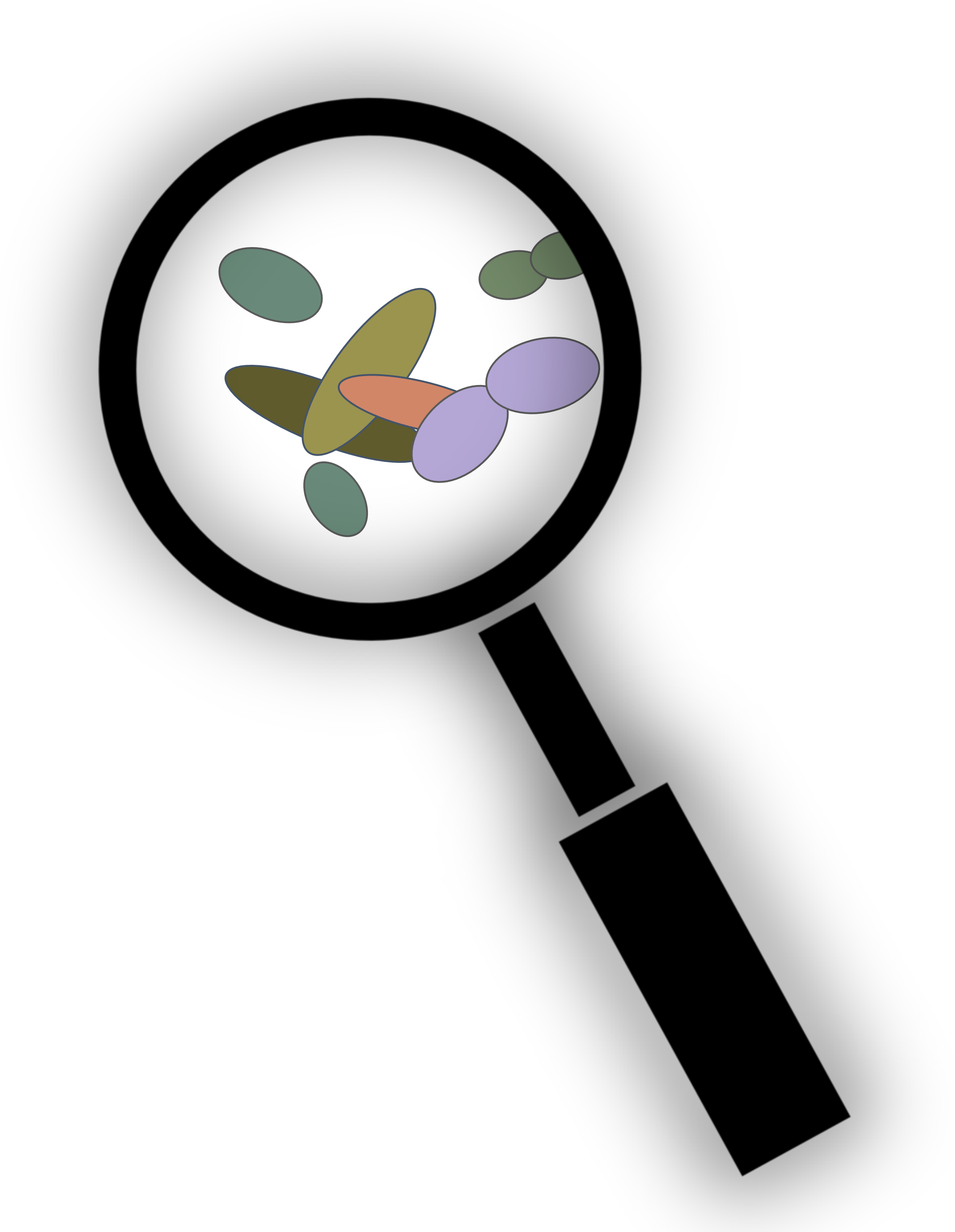


The Gut Microbiome is Challenging

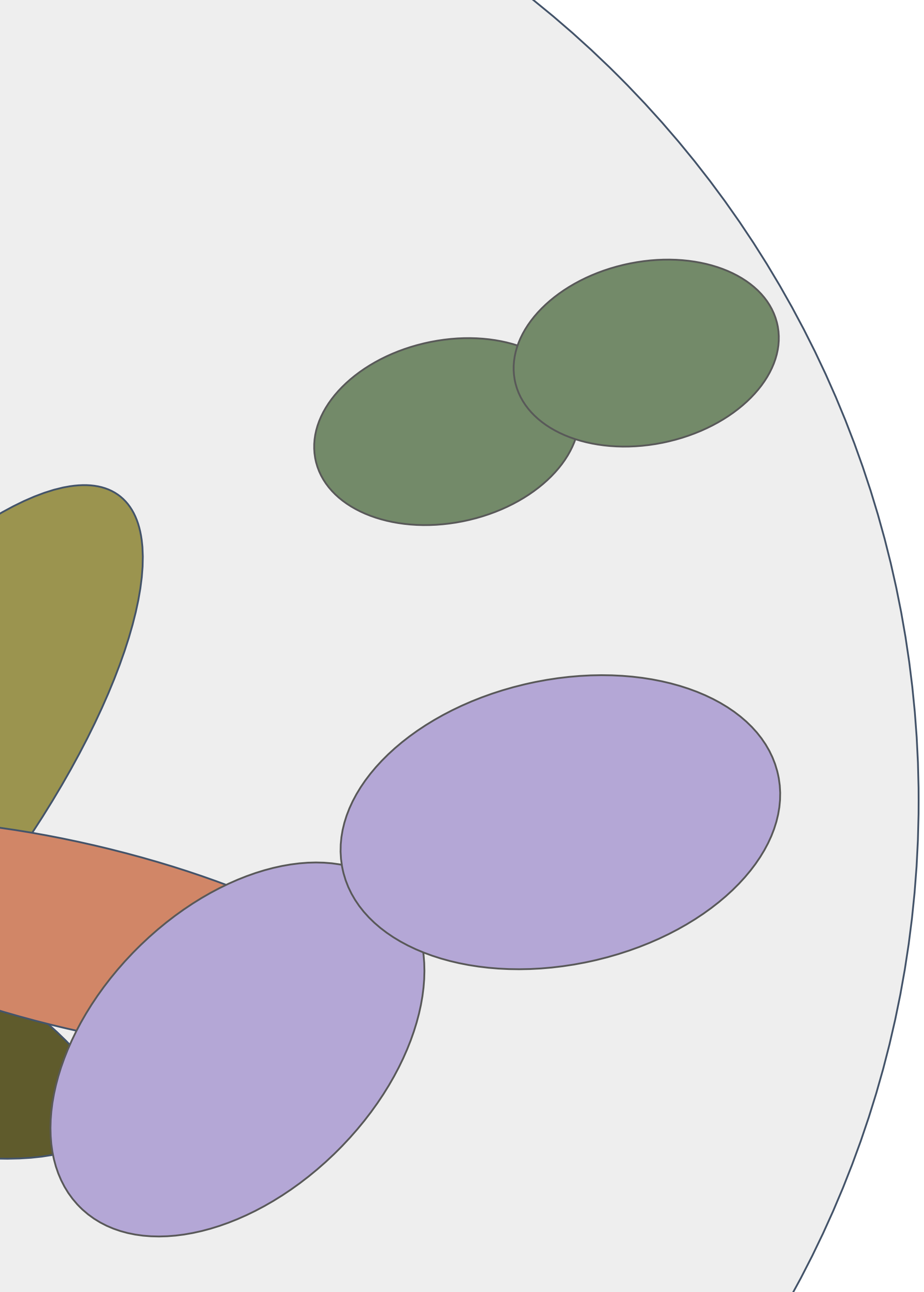
- Enormous number of species
- Highly dynamic across people and time
- Very hard to study in the lab
- **Strains within species have different gene content and functional potential**

The Gut Microbiome is Challenging

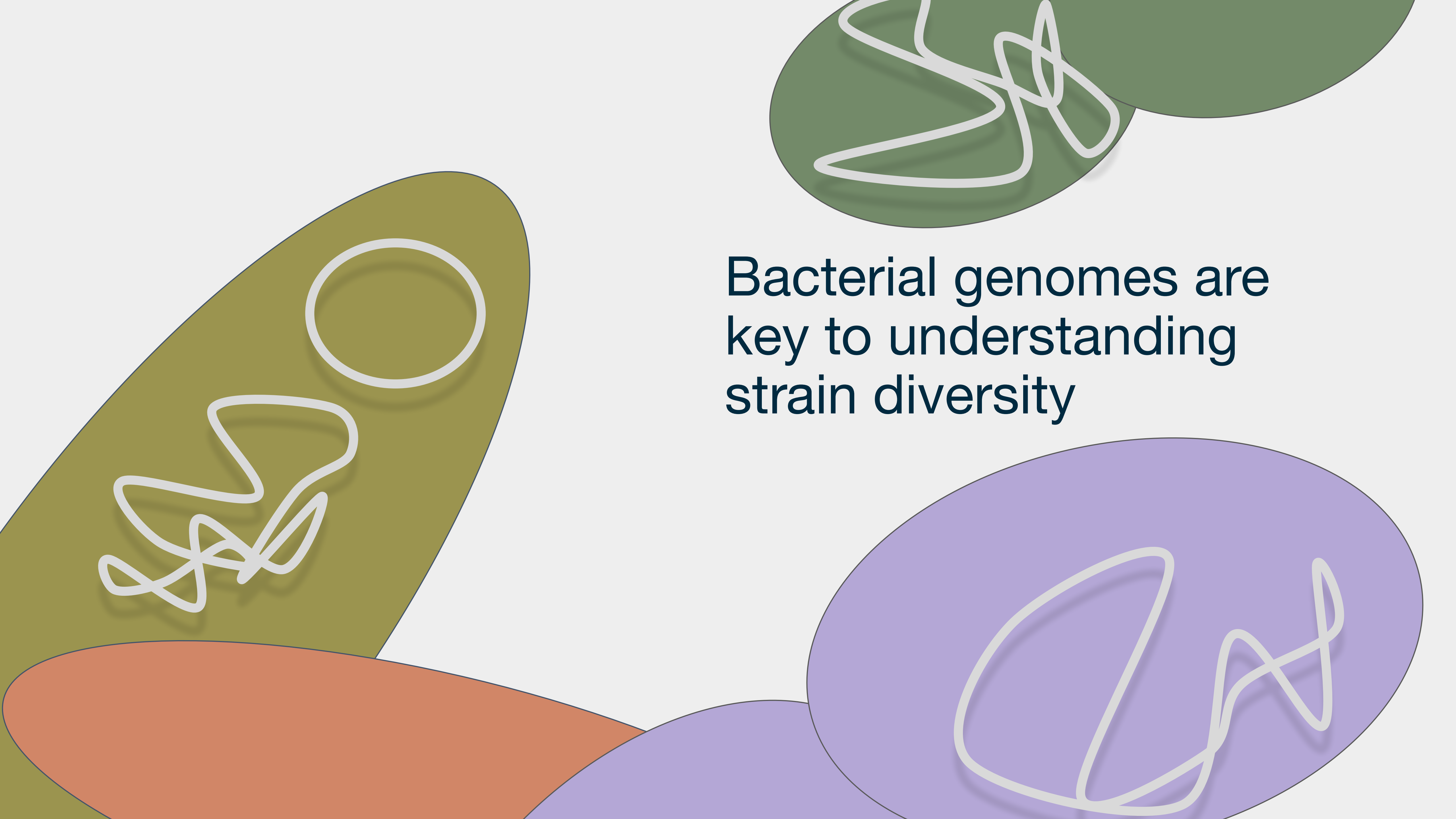
- Enormous number of species
- Highly dynamic across people and time
- Very hard to study in the lab
- **Strains within species have different gene content and functional potential**



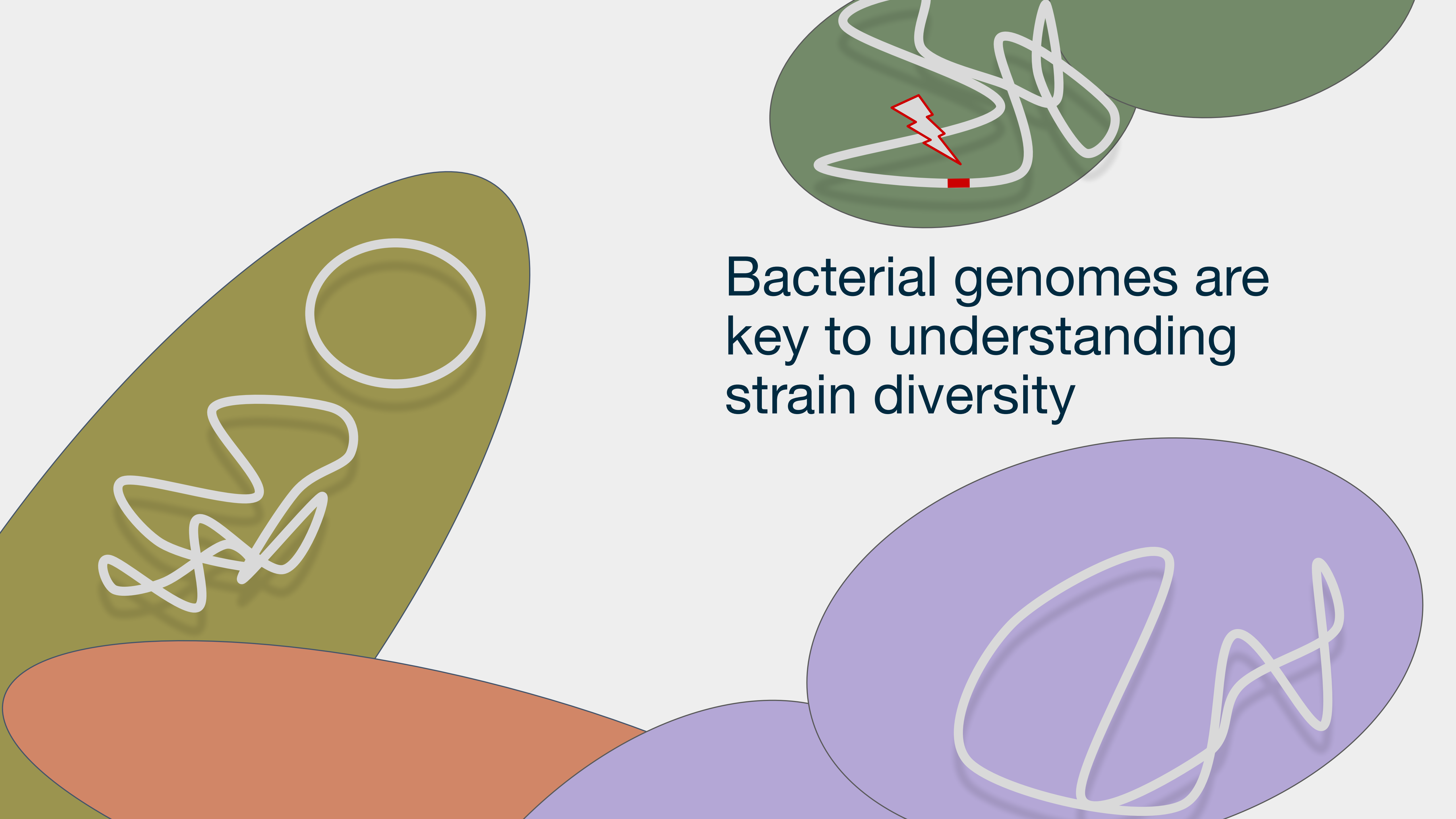
Bacterial genomes are key to understanding strain diversity



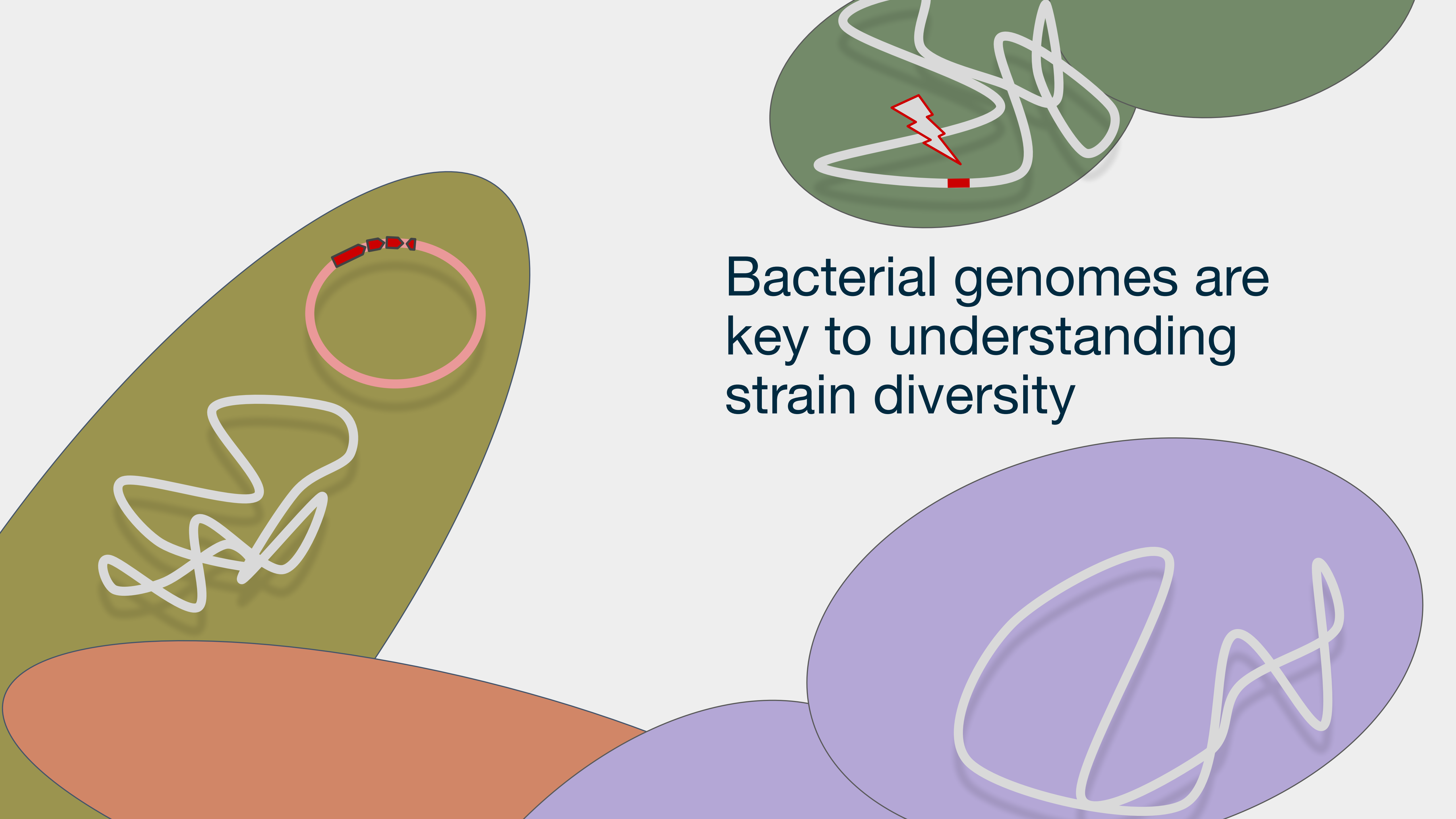
Bacterial genomes are key to understanding strain diversity



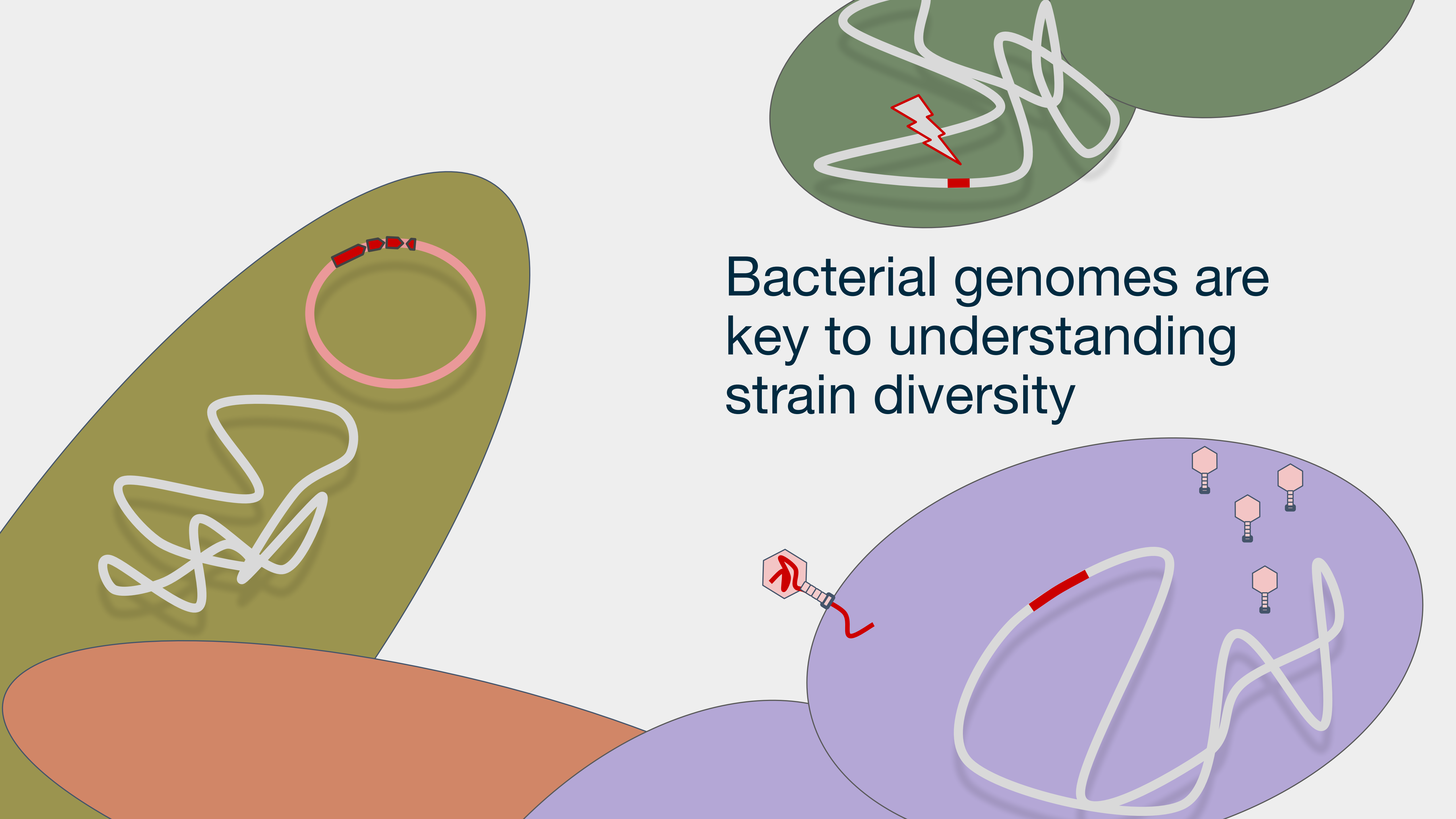
Bacterial genomes are key to understanding strain diversity



Bacterial genomes are key to understanding strain diversity

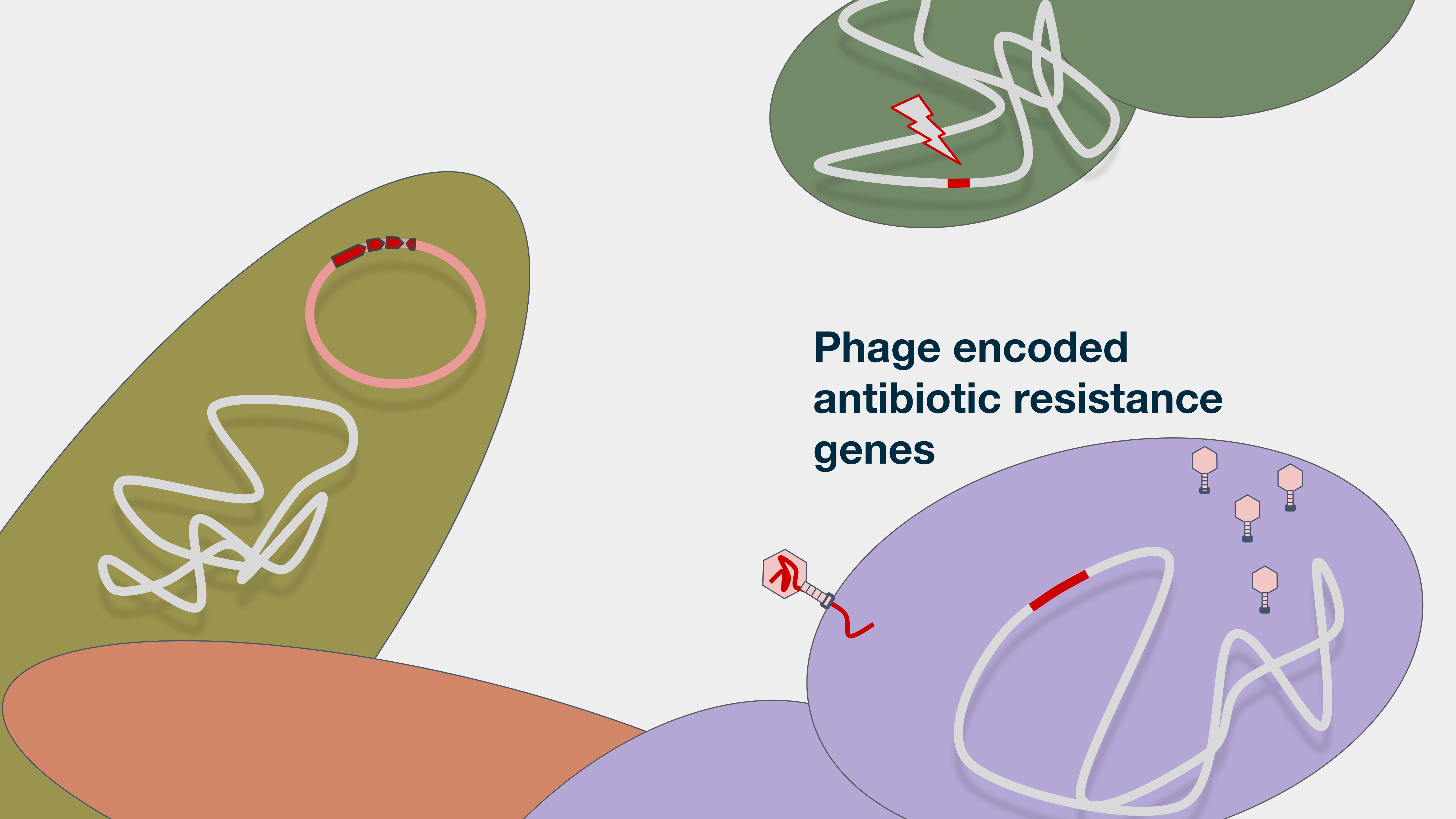


Bacterial genomes are key to understanding strain diversity

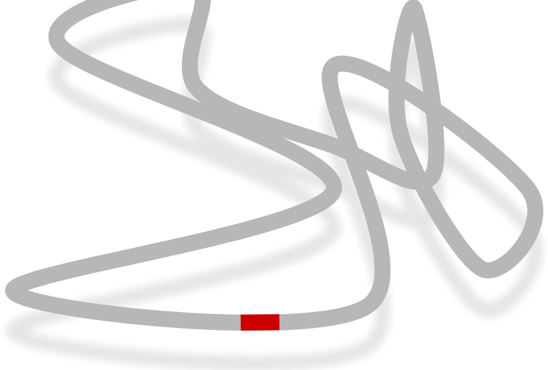


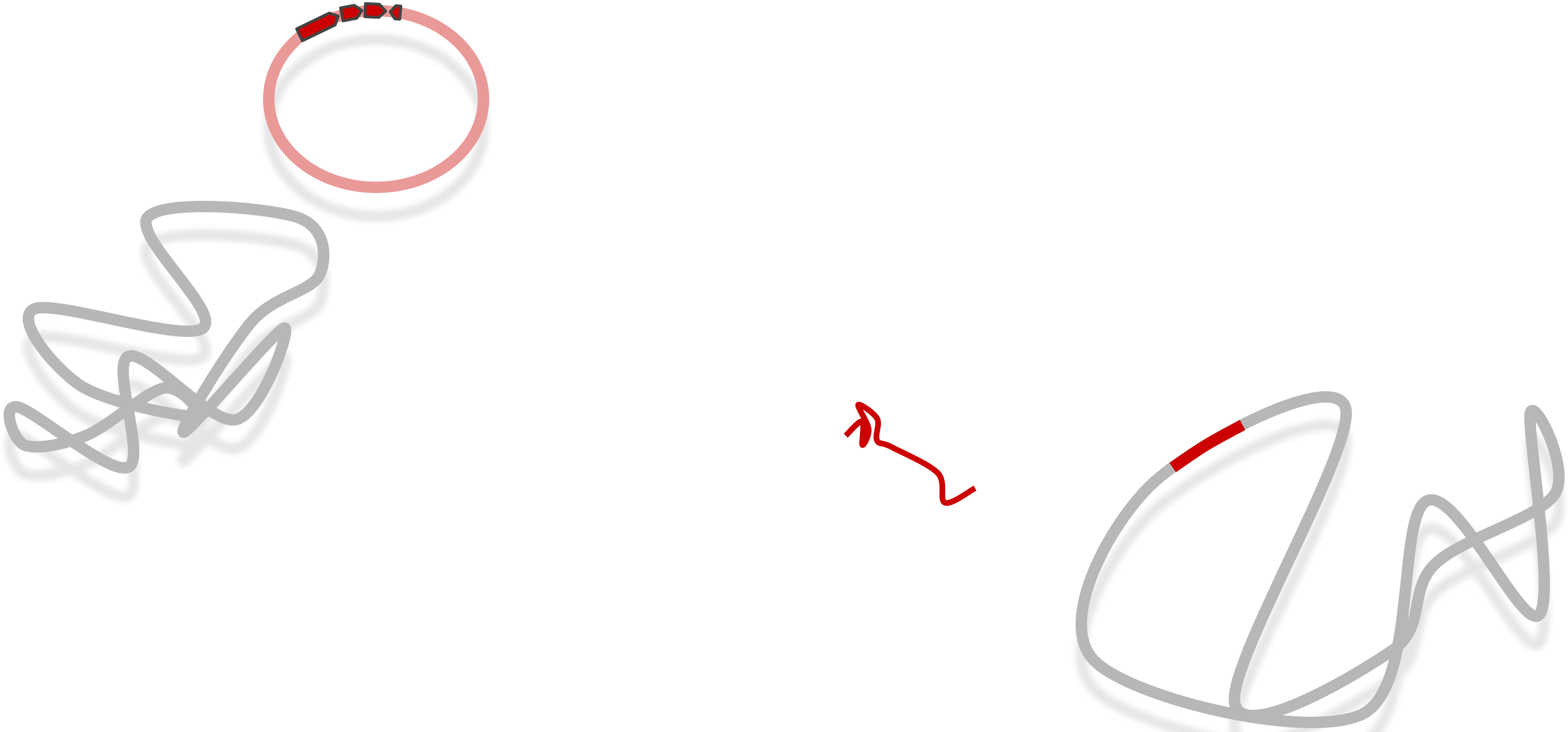
Bacterial genomes are key to understanding strain diversity

Phage encoded antibiotic resistance genes

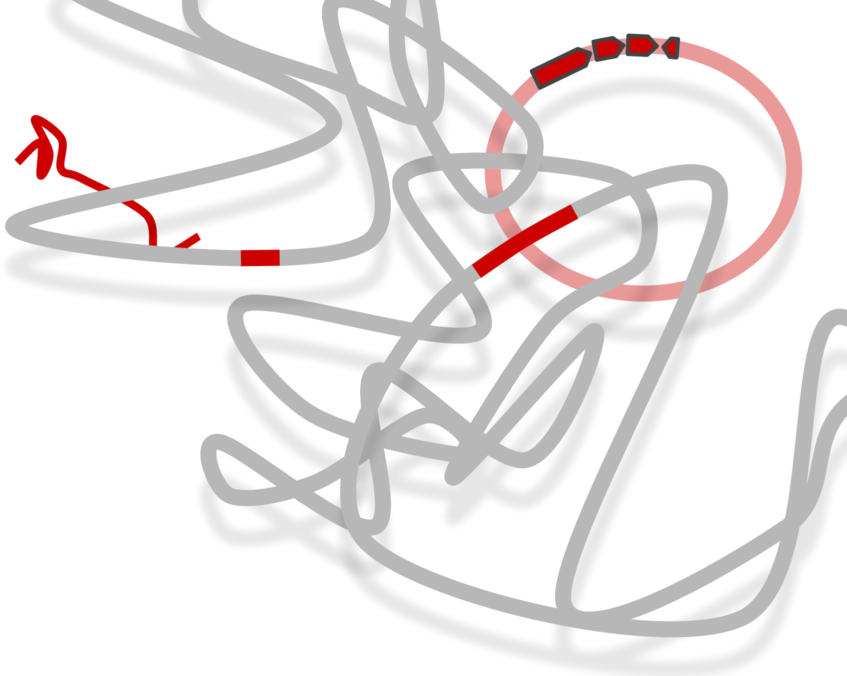


Metagenomic sequencing surveys all genomes



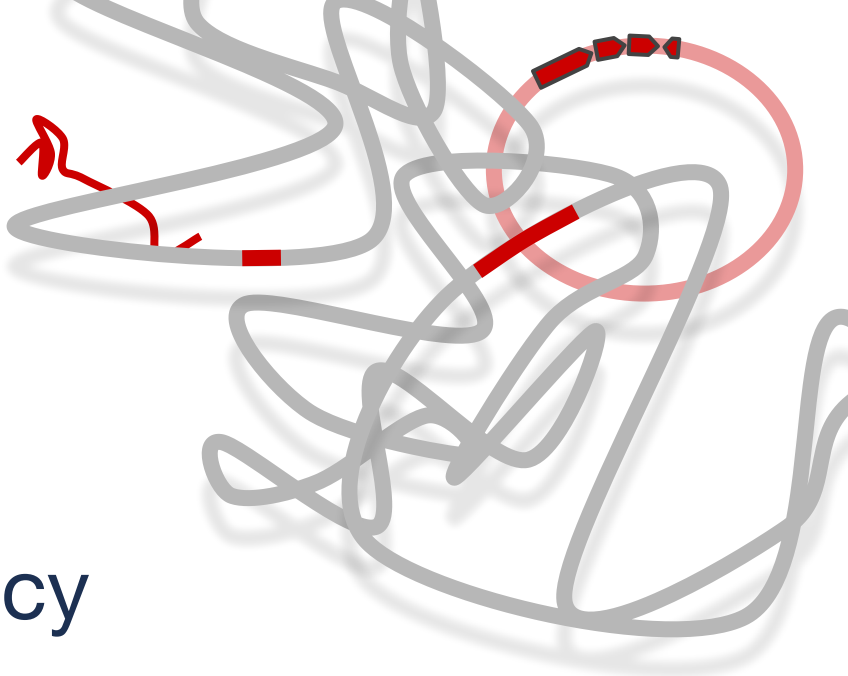


Short-read, shotgun metagenomes enable modern microbiome science



Requirements:

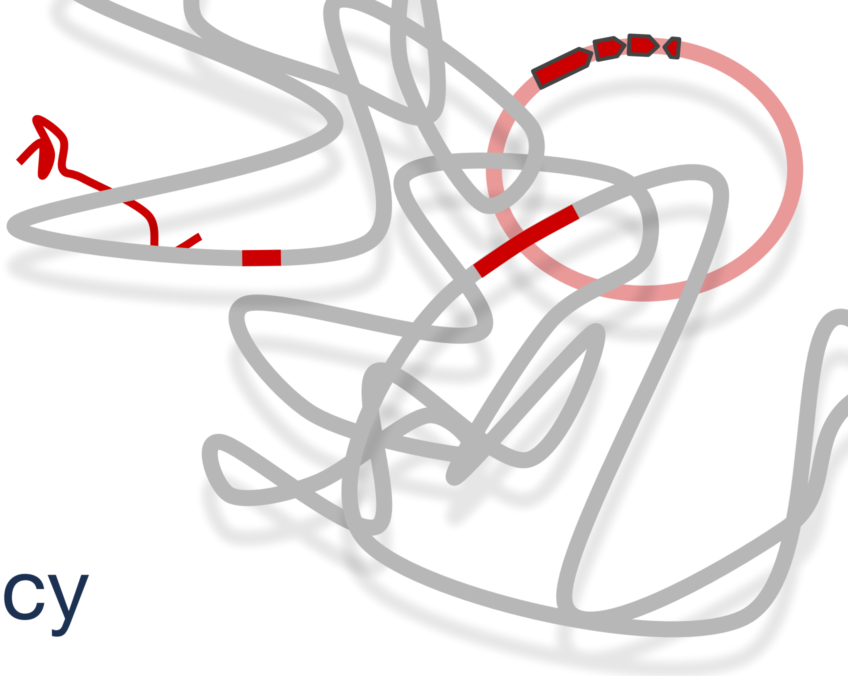
Short-read, shotgun metagenomes enable modern microbiome science



Requirements:

- strain-resolved genome sequences
  - ➢ high accuracy

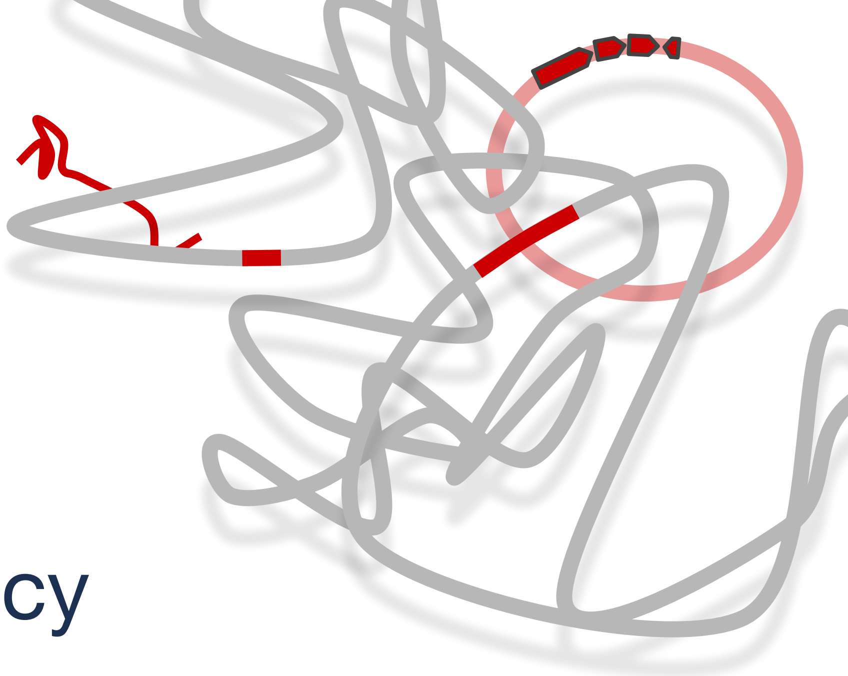
Short-read, shotgun metagenomes enable modern microbiome science



Requirements:

- strain-resolved genome sequences ➢ high accuracy
- capture low-abundance organisms ➢ very deep sequencing

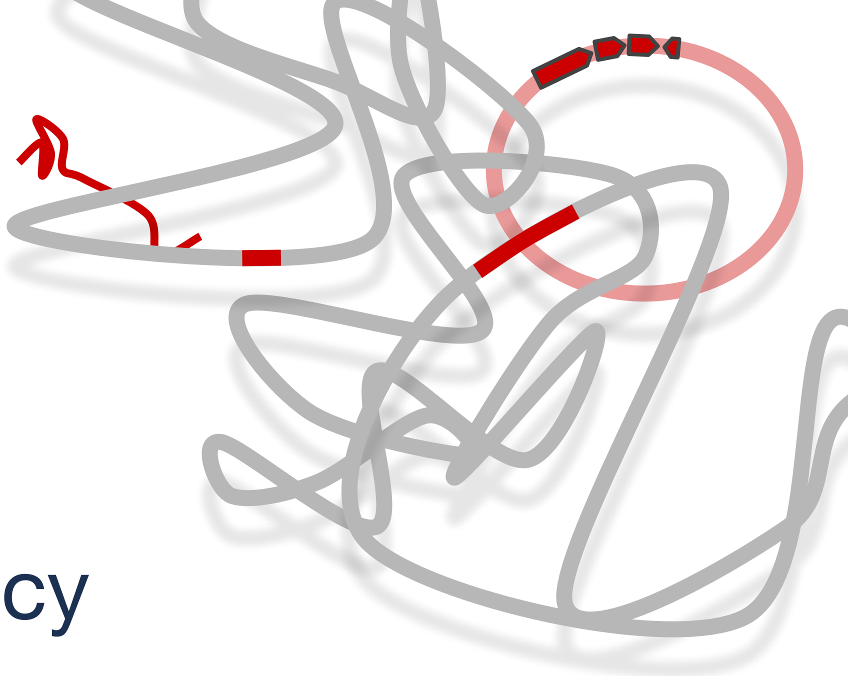
Short-read, shotgun metagenomes enable modern microbiome science



Requirements:

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

Short-read, shotgun metagenomes enable modern microbiome science



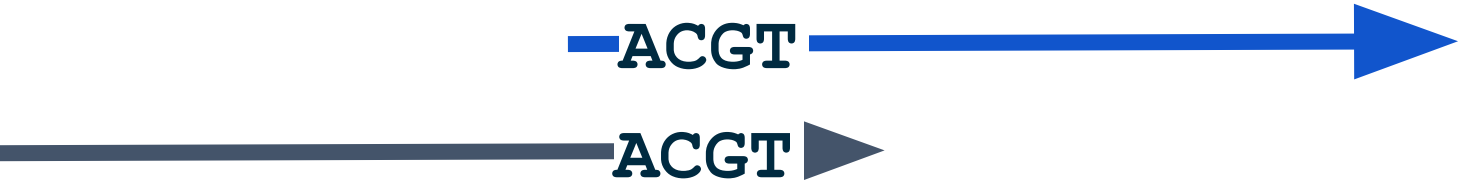
Requirements:

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

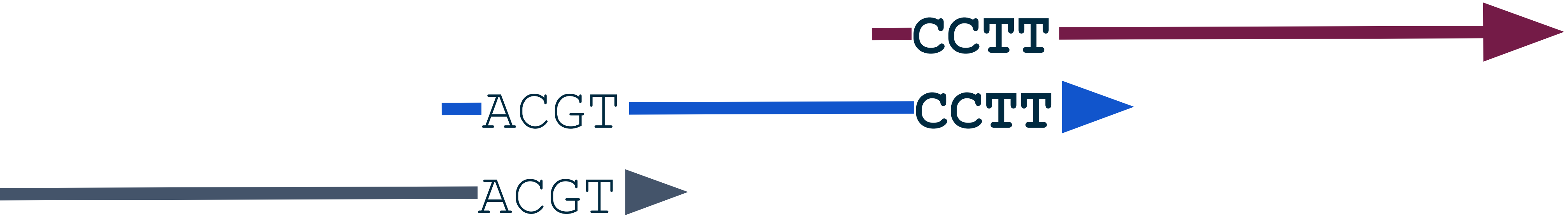
Turning short reads into long sequences

ACGT

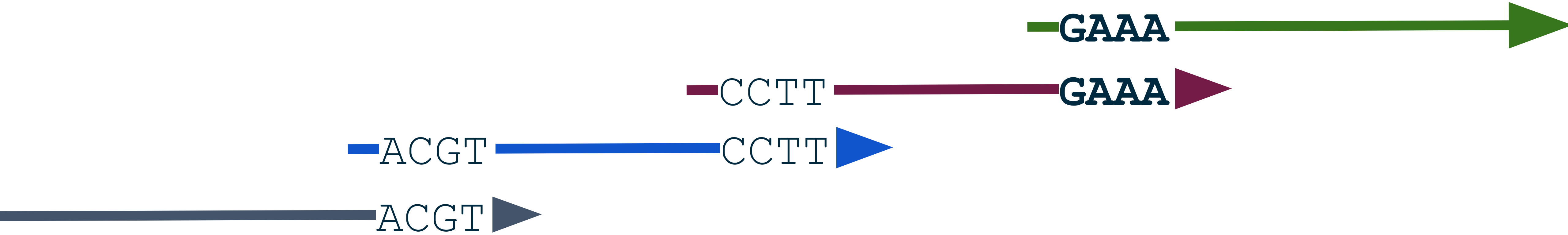
Turning short reads into long sequences



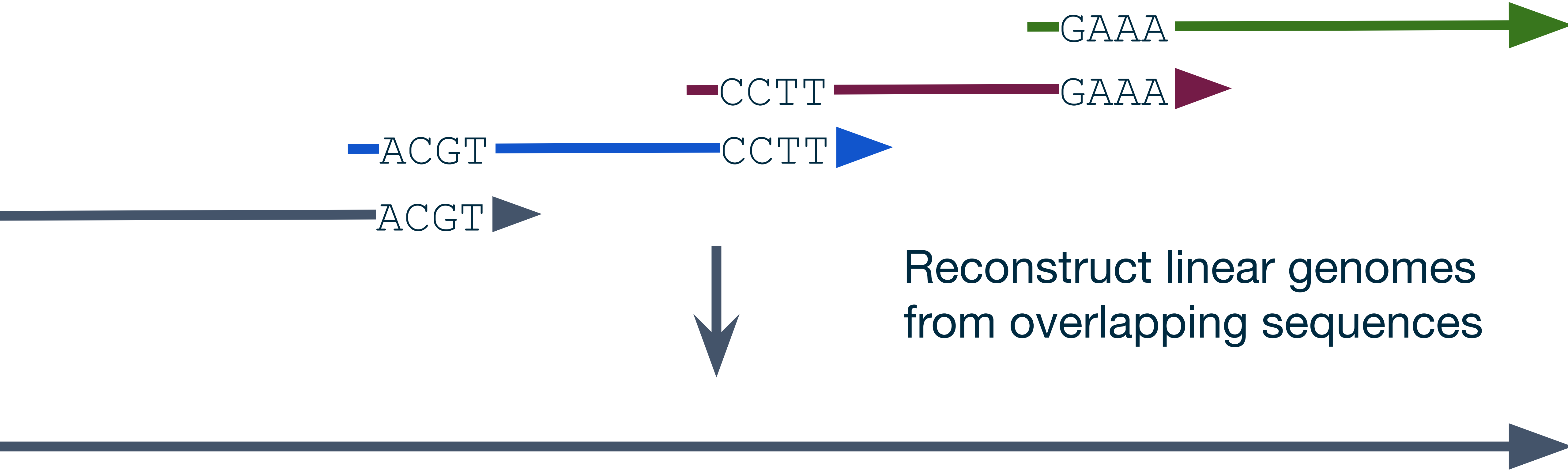
Turning short reads into long sequences



Turning short reads into long sequences



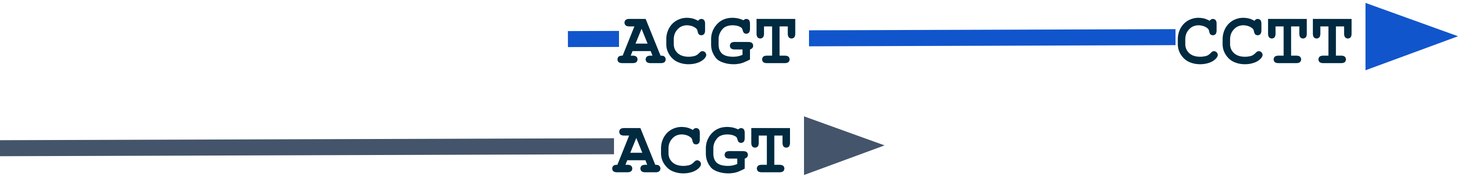
Turning short reads into long sequences



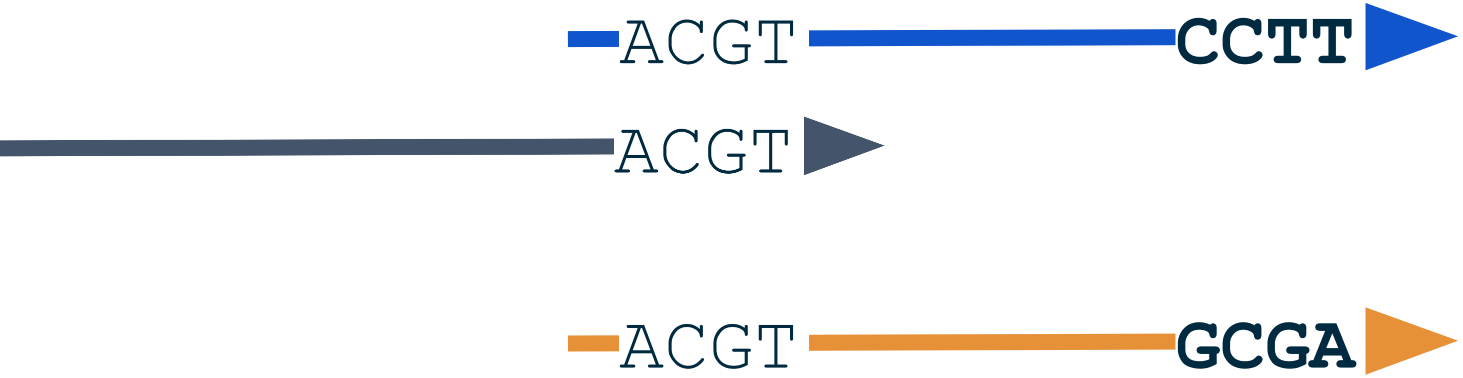
Reconstruct linear genomes from overlapping sequences

...GGTAGAGCGTGGACGTAGGGTTAACCTTAGAAAGCTAGAAAACCGCGCGCCCT...

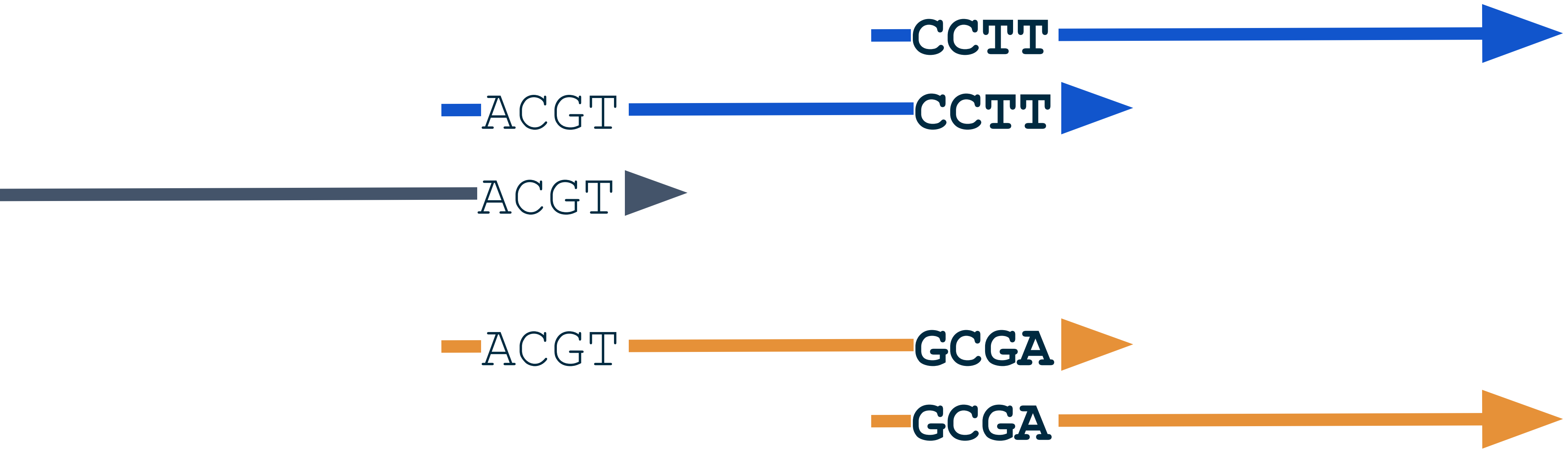
Problem:



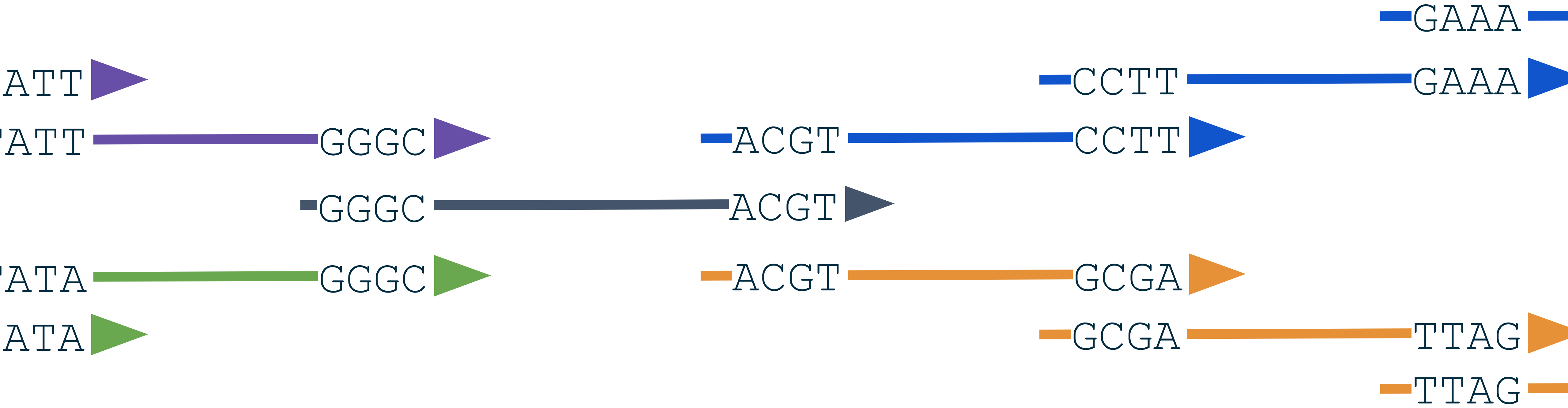
Problem: Closely related strains make read-chaining ambiguous



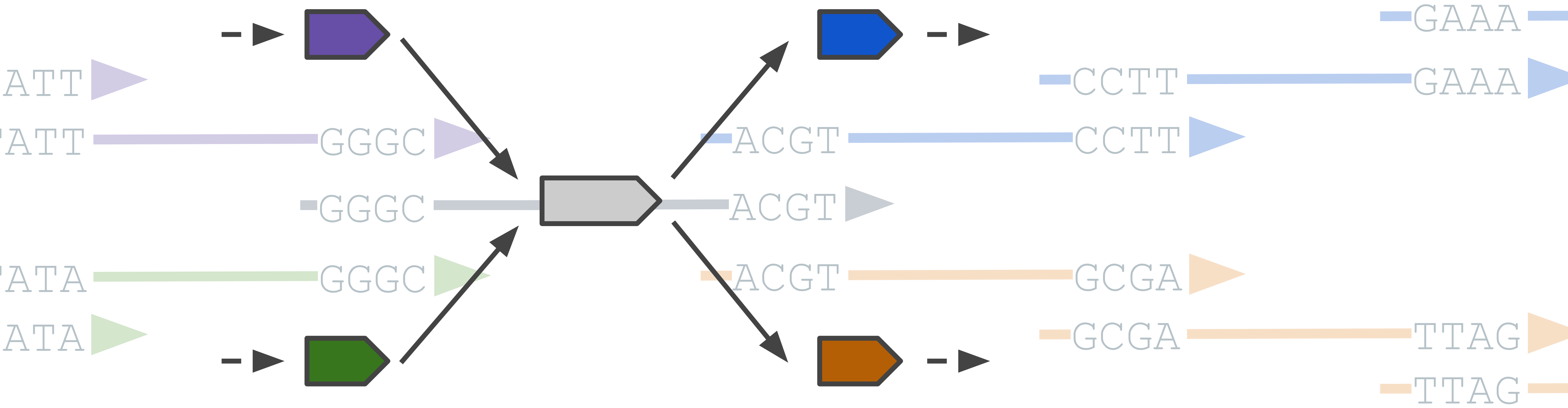
Problem: Closely related strains make read-chaining ambiguous



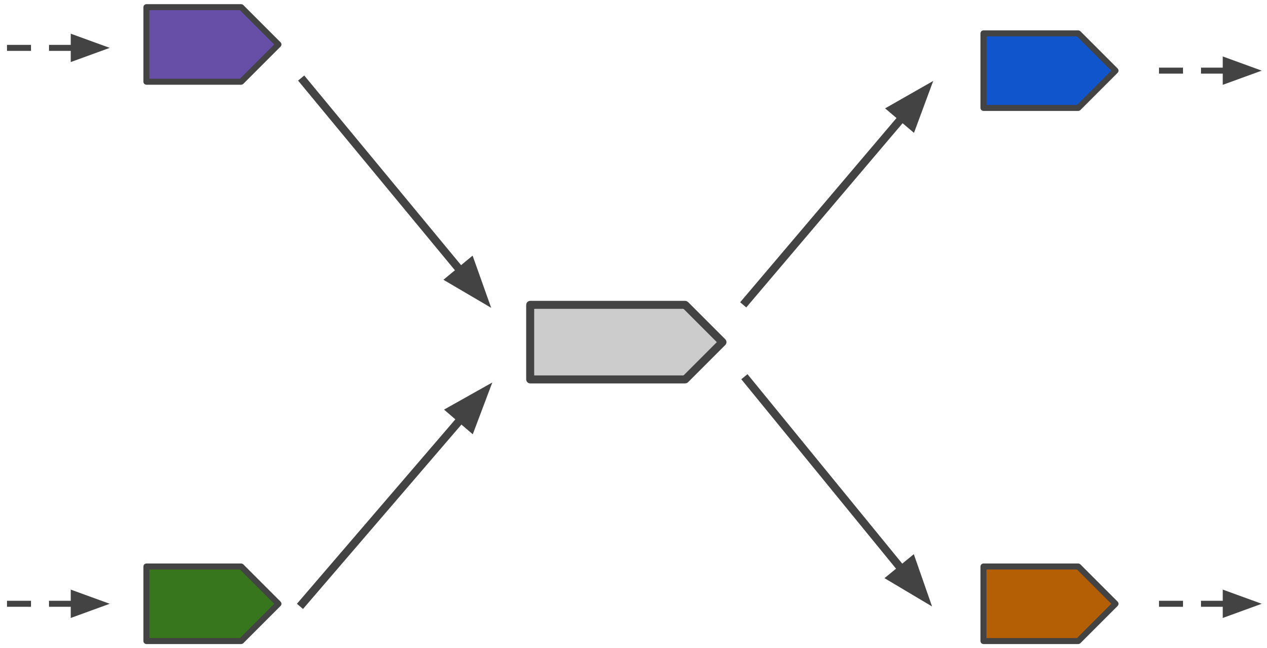
Problem: Closely related strains make read-chaining ambiguous



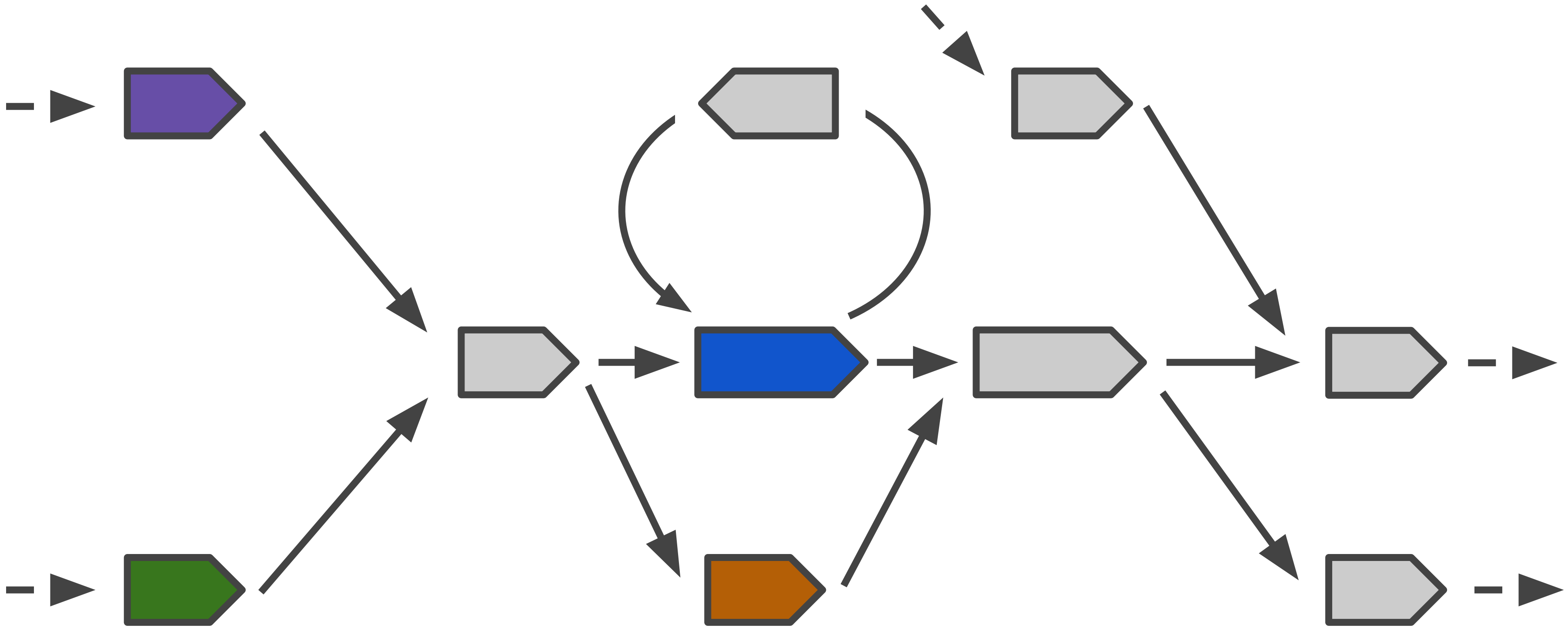
Can be represented as a graph of sequences linked by their overlaps



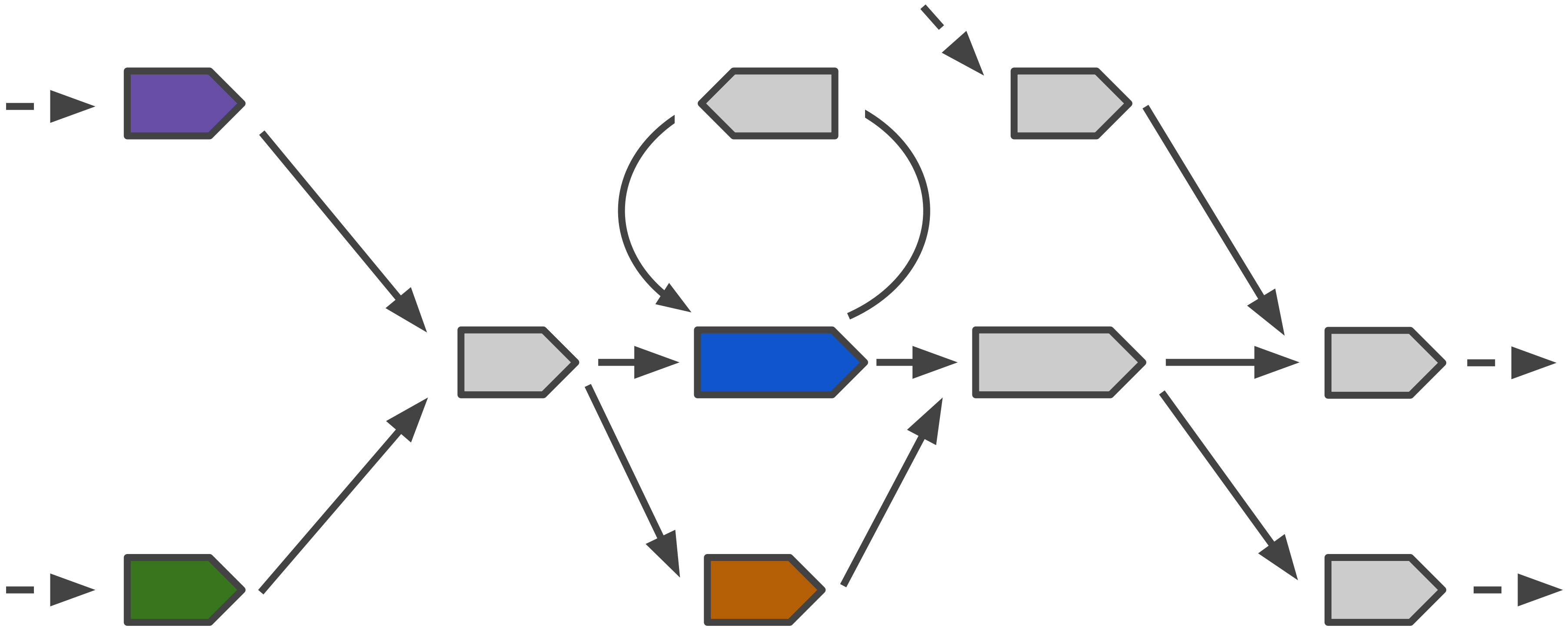
Can be represented as a graph of sequences linked by their overlaps



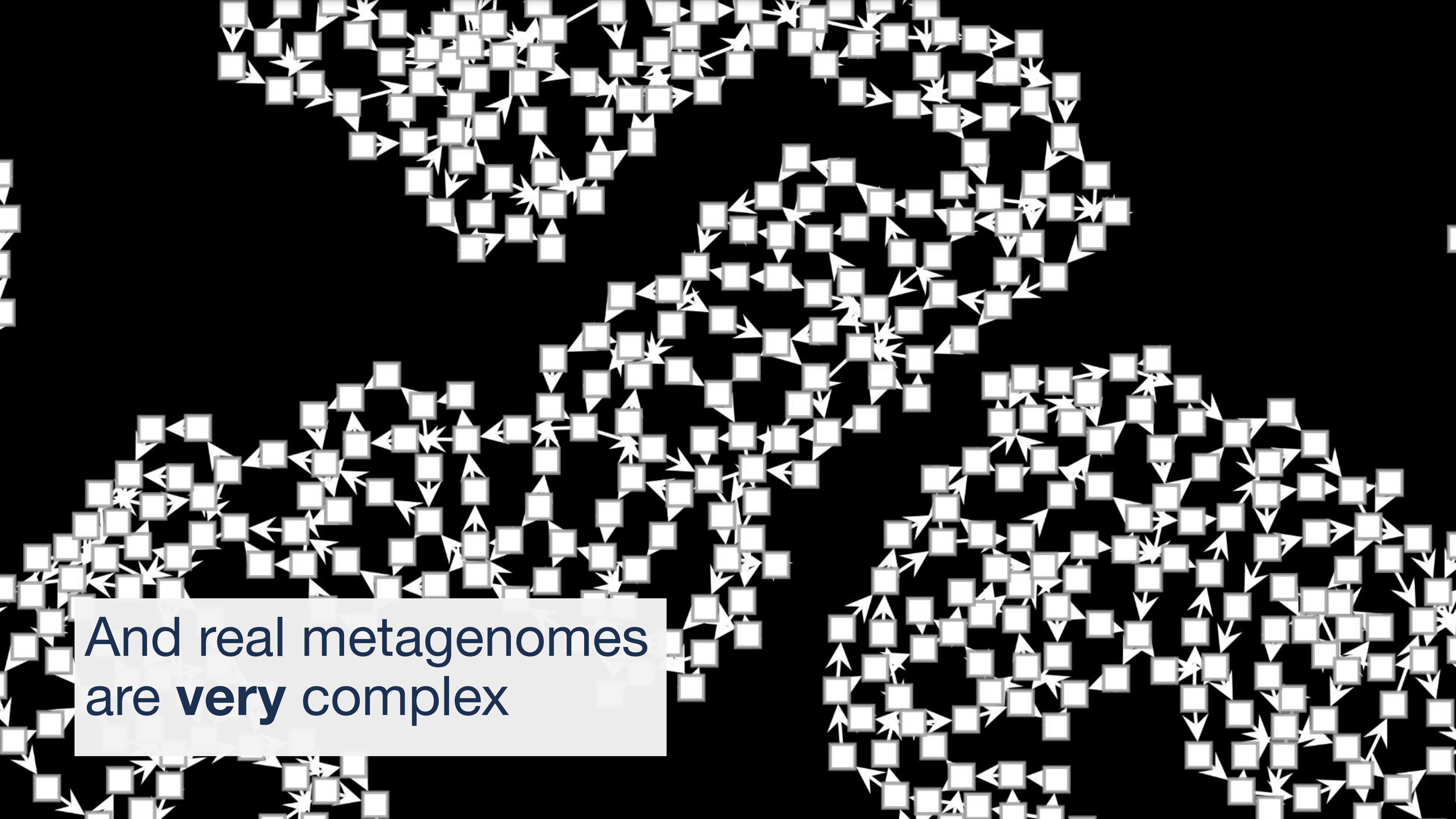
Can be represented as a graph of sequences linked by their overlaps



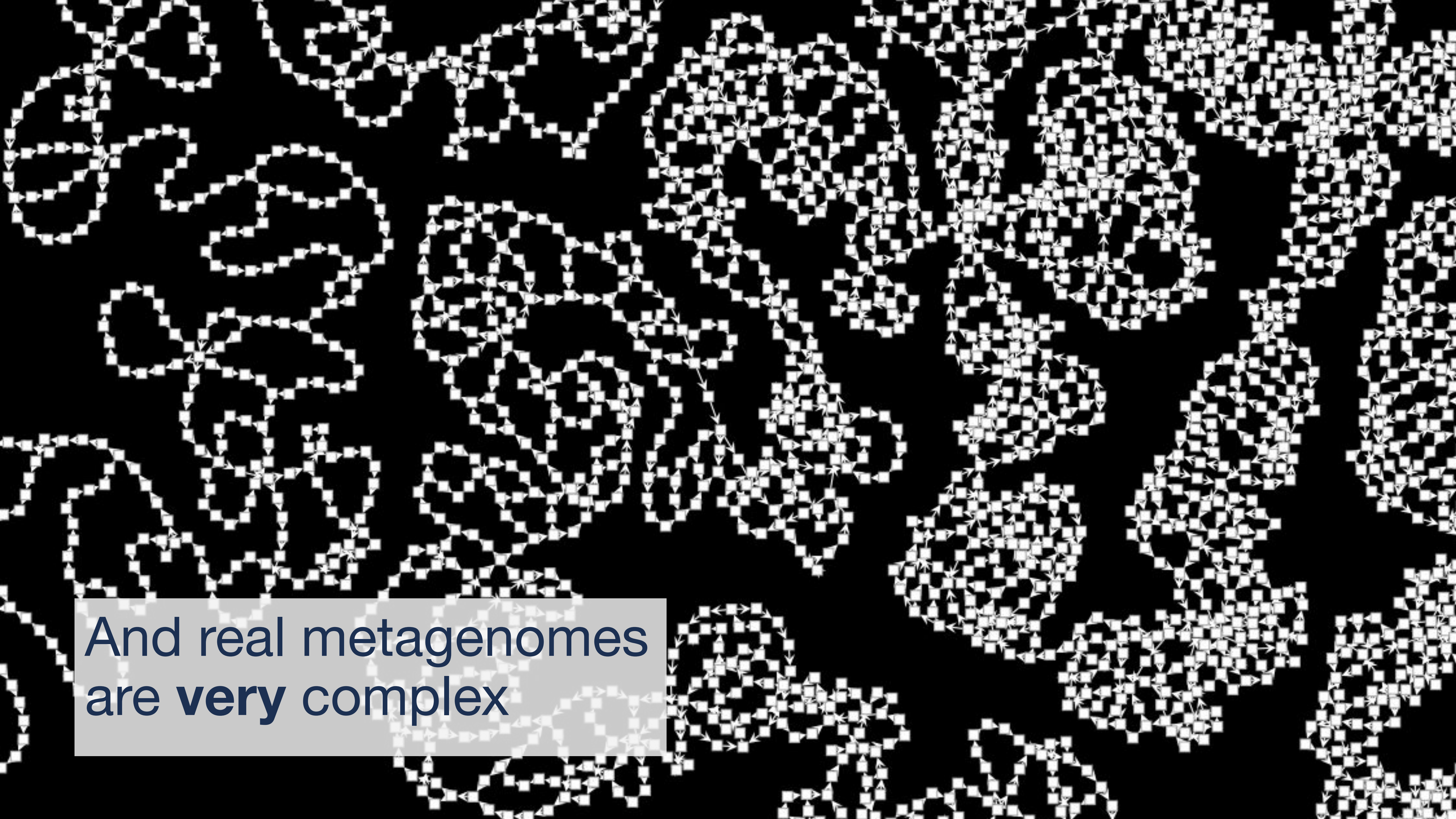
Can be represented as a graph of sequences linked by their overlaps



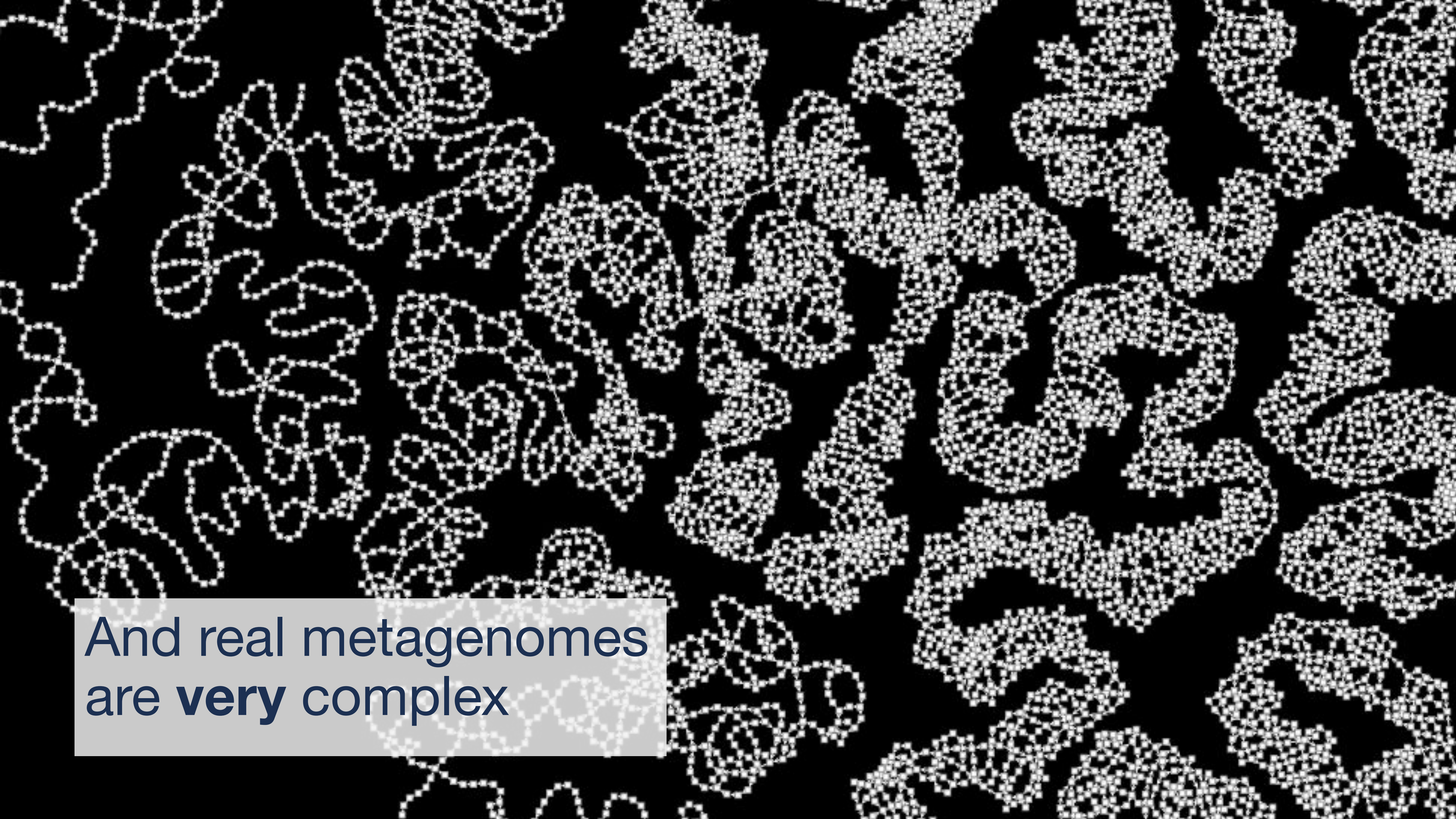
(This problem also comes up for mRNA alternative splicing)



And real metagenomes are **very** complex

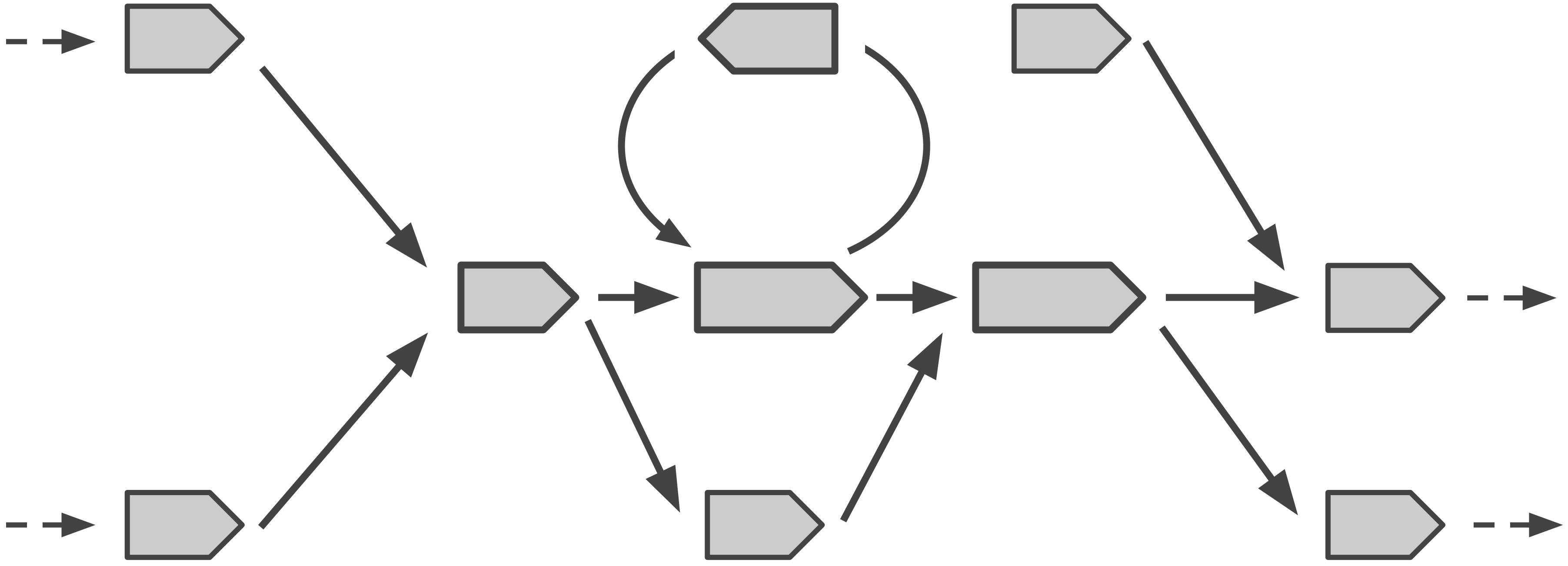


And real metagenomes are **very** complex

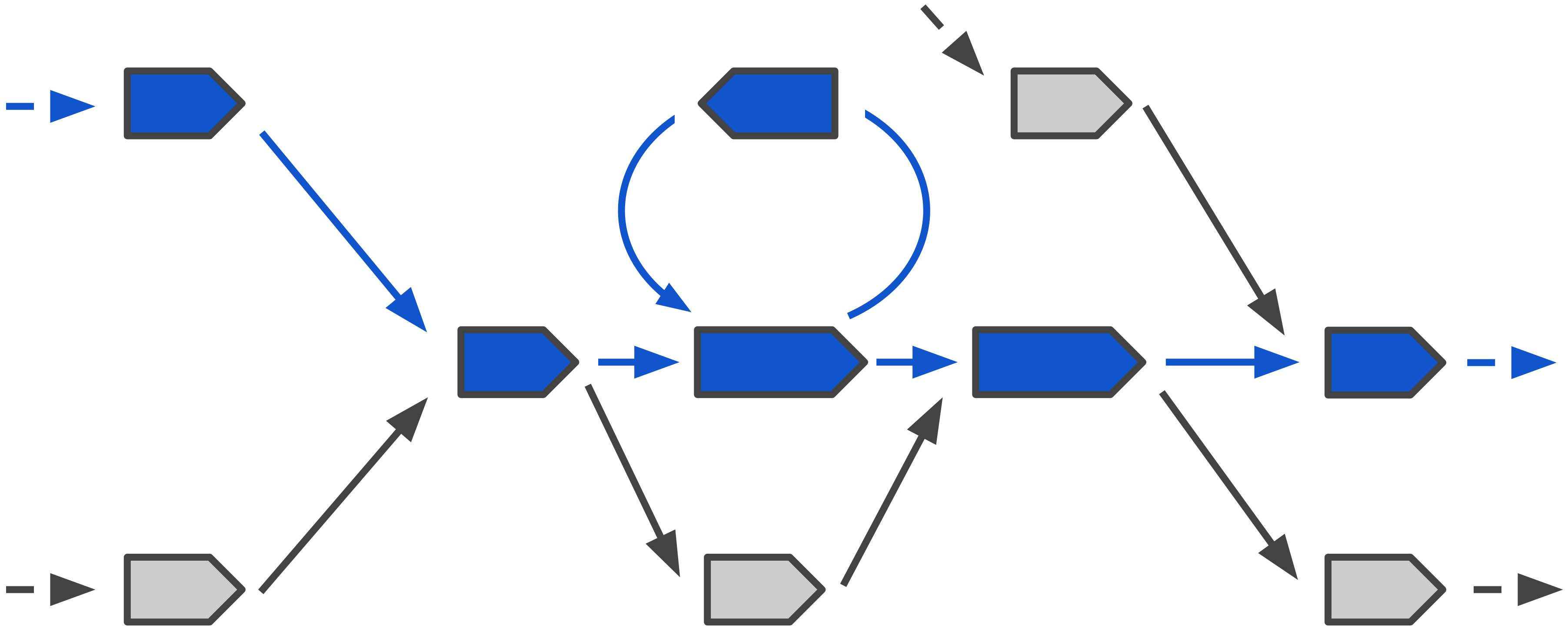


And real metagenomes are **very** complex

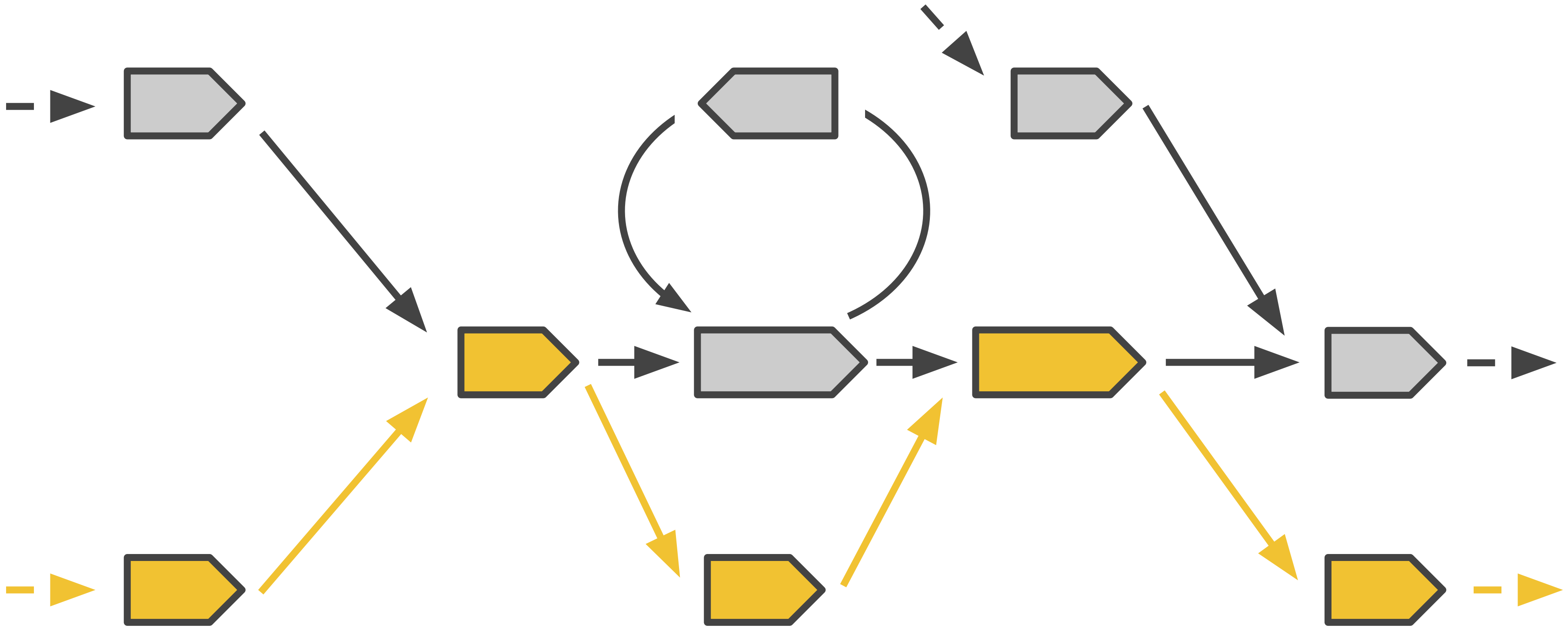
Real genomic sequences are paths on the graph



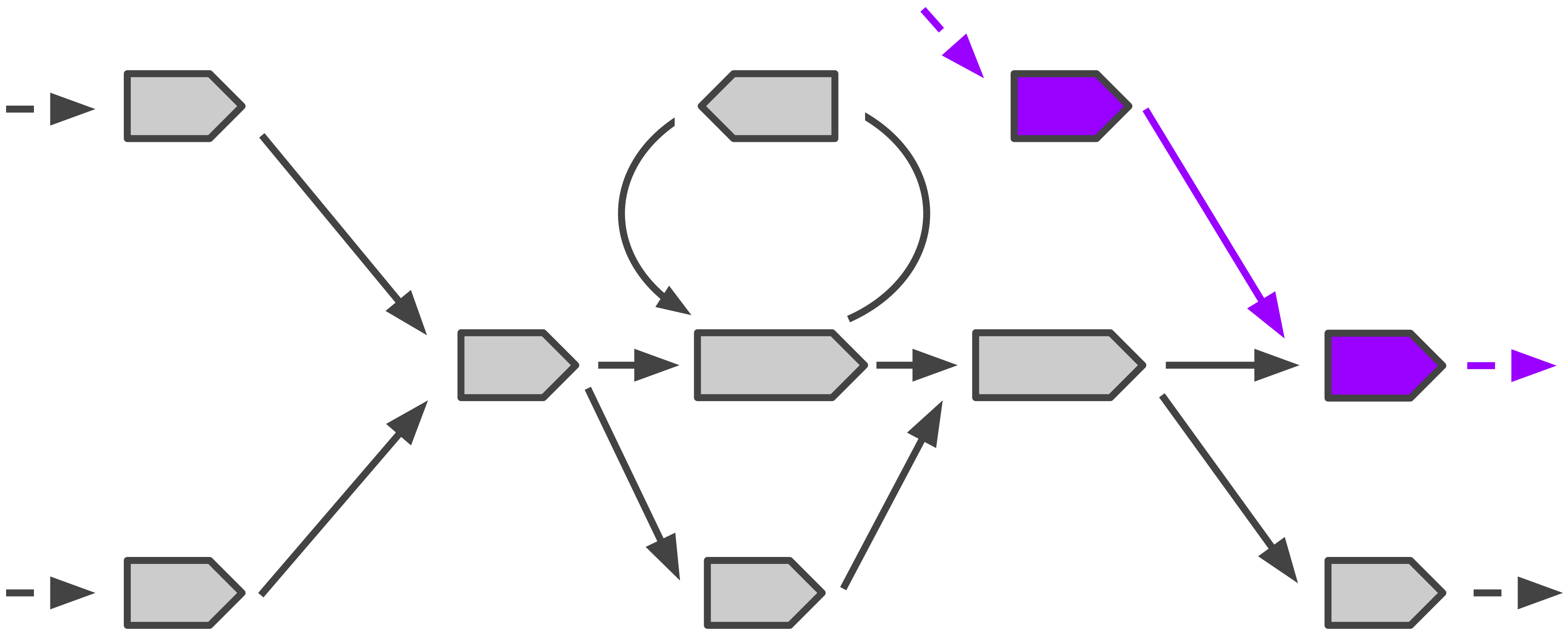
Real genomic sequences are paths on the graph



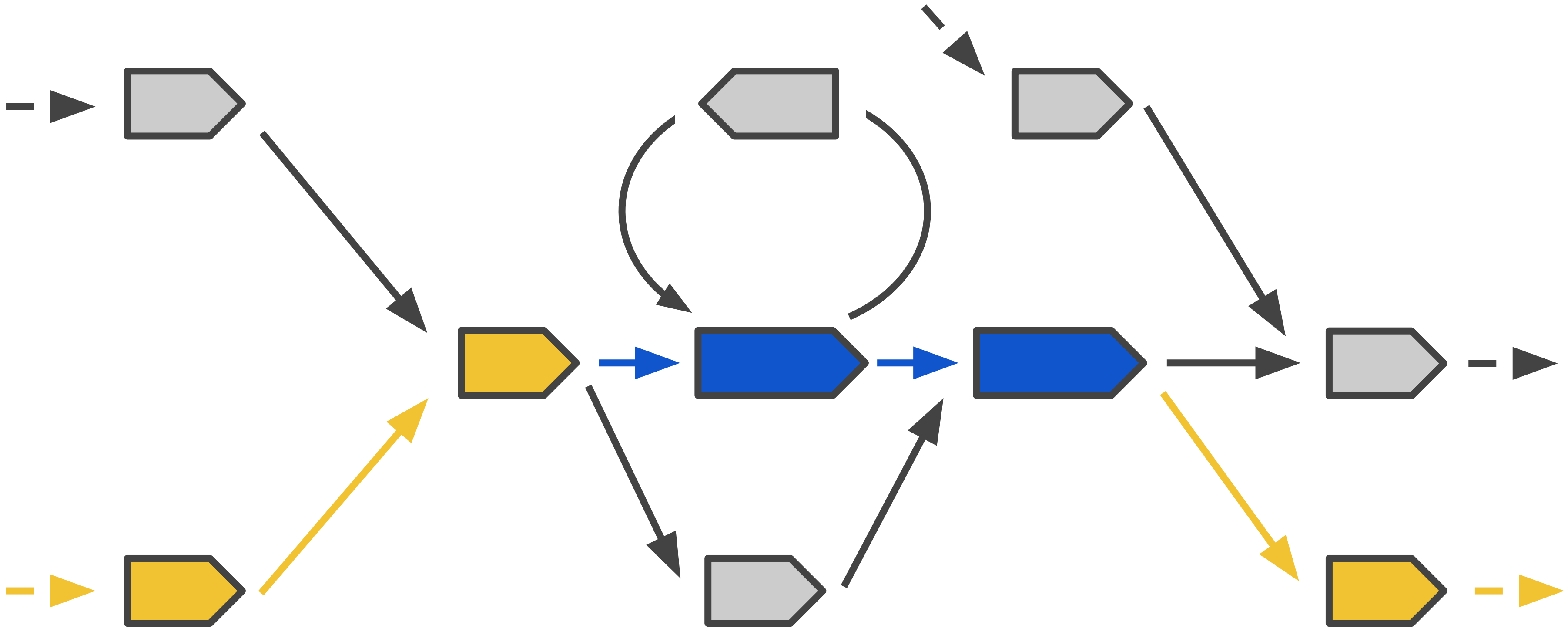
Real genomic sequences are paths on the graph



Real genomic sequences are paths on the graph

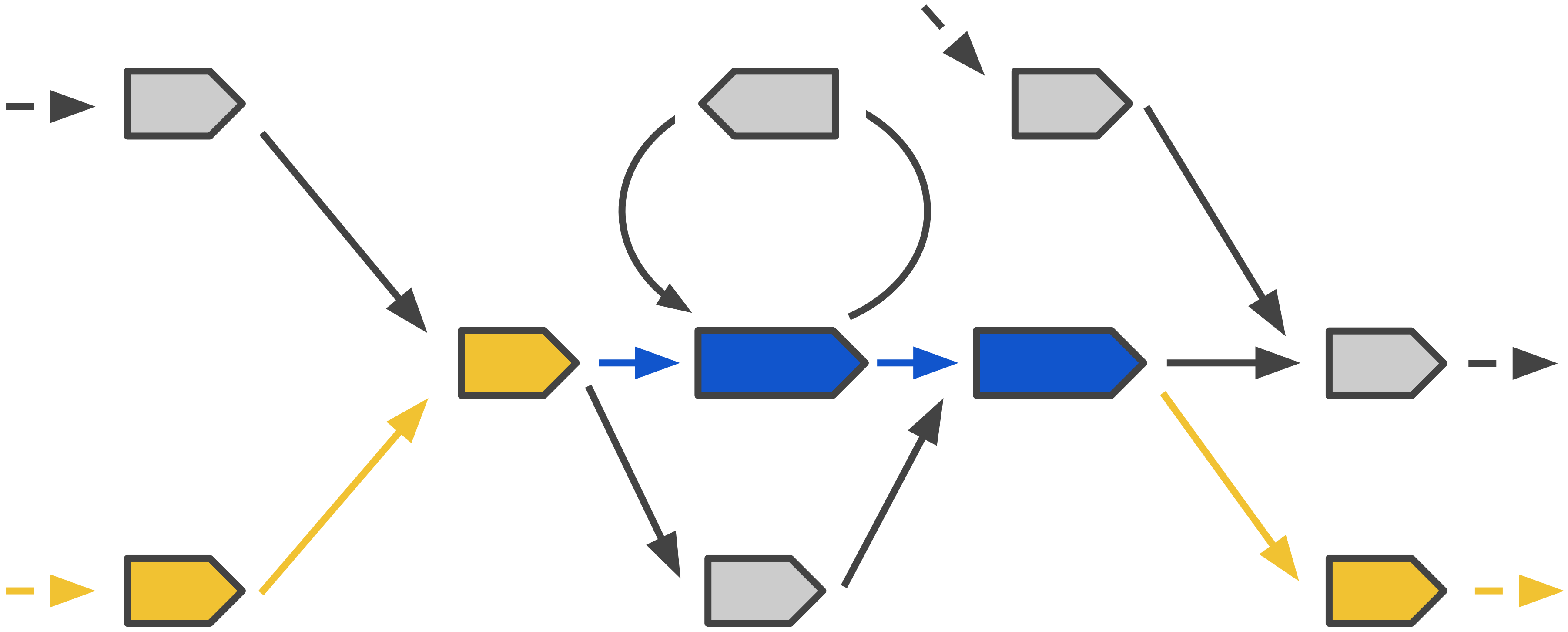


Lots of incorrect paths also exist…



Lots of incorrect paths also exist…

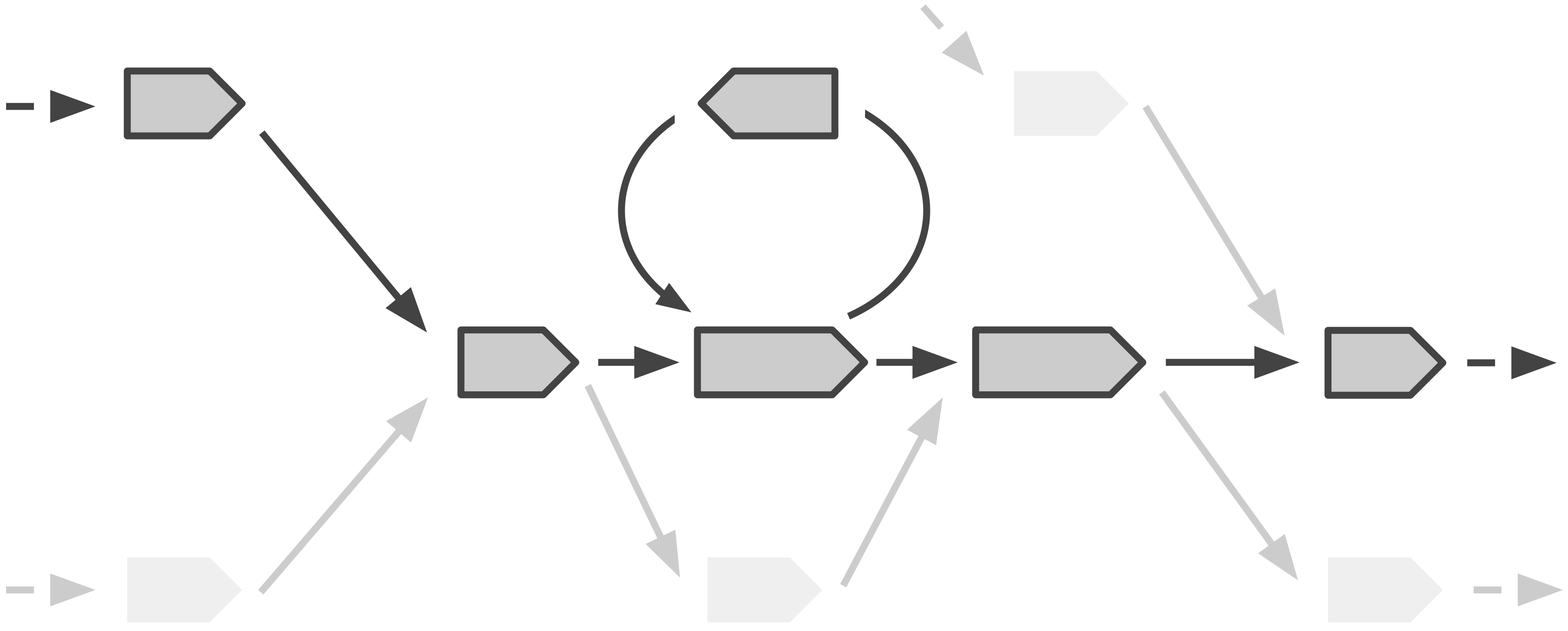
How do we avoid these?



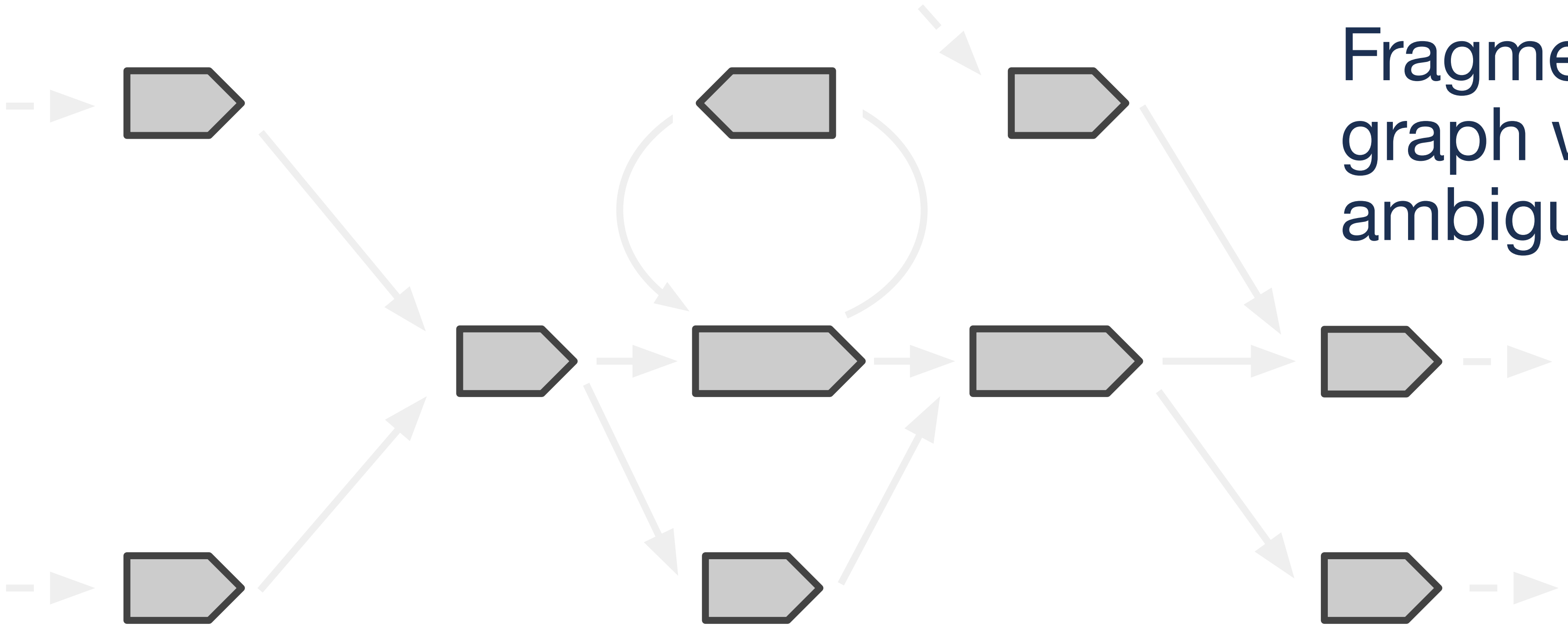
Lots of incorrect paths also exist…
How do we avoid these?

Standard Tools:

Filter out low-abundance sequences



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



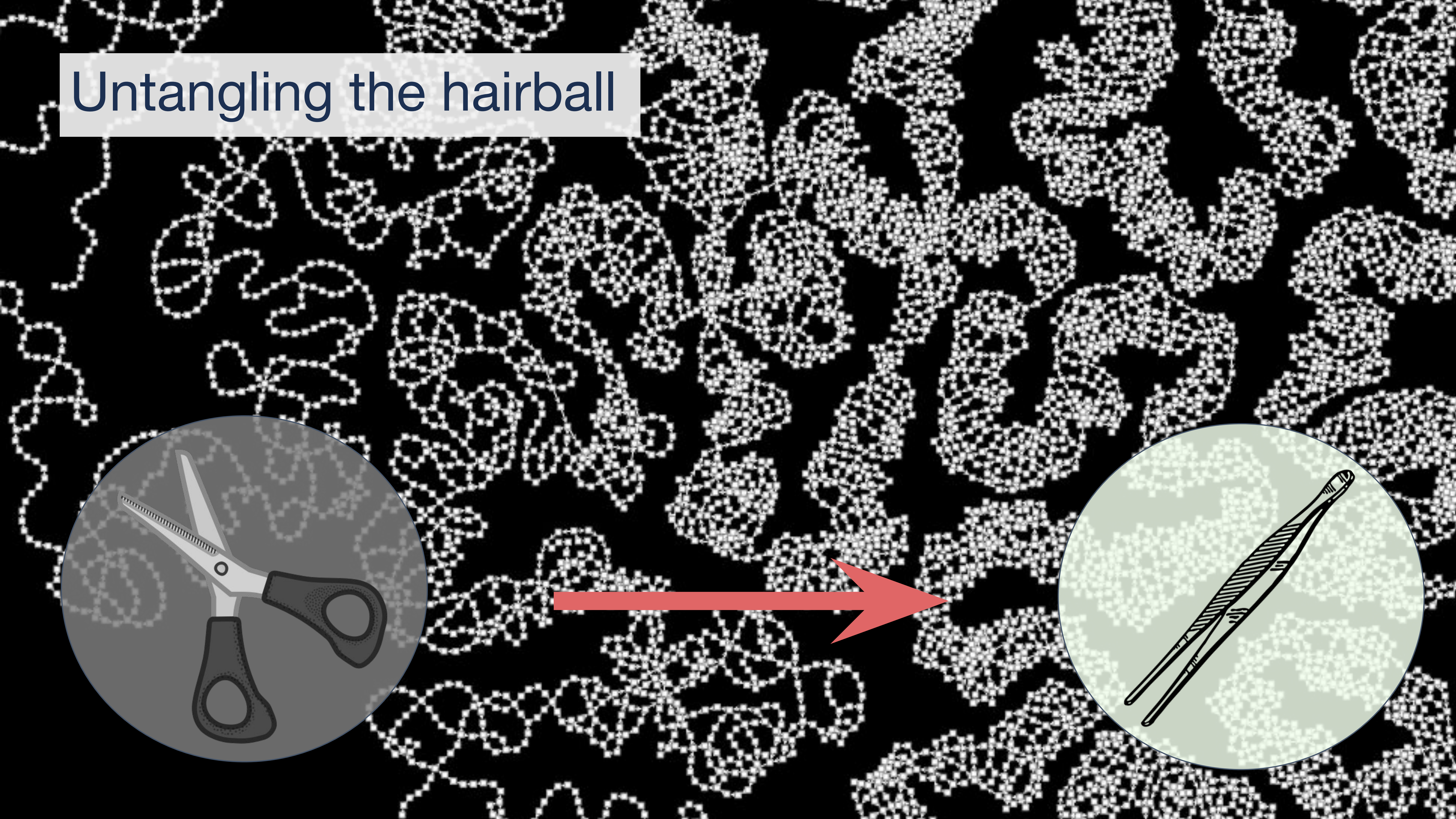
Standard Tools:

Filter out low-abundance sequences

Fragment the graph when it's ambiguous

Untangling the hairball

Untangling the hairball

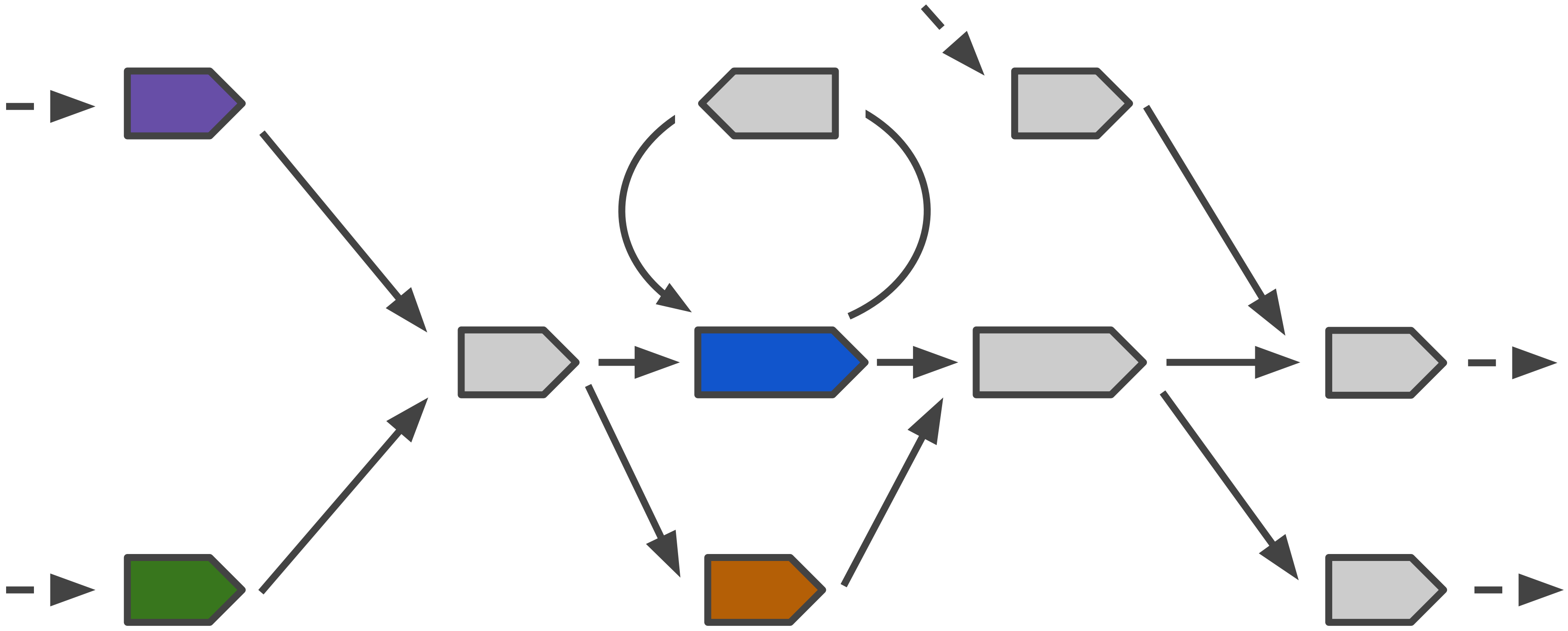


Gladstone Scientific Retreat 2024

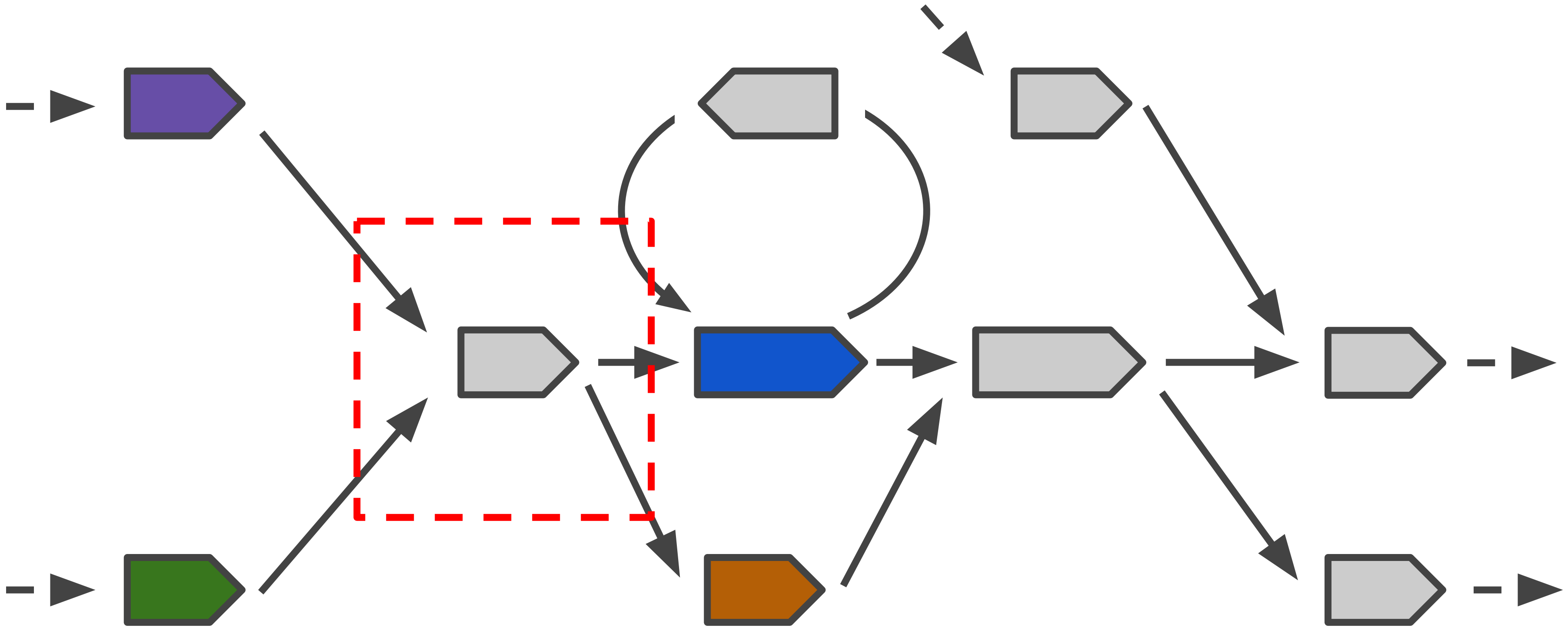
StrainZip:

Untangling the metagenome graph

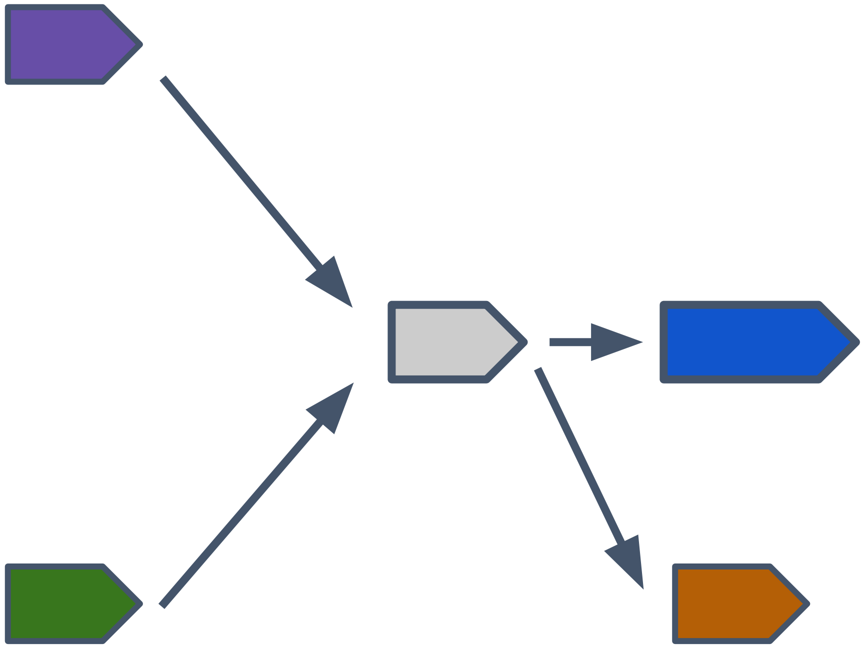
How can we recover long, accurate genome sequences from short reads?



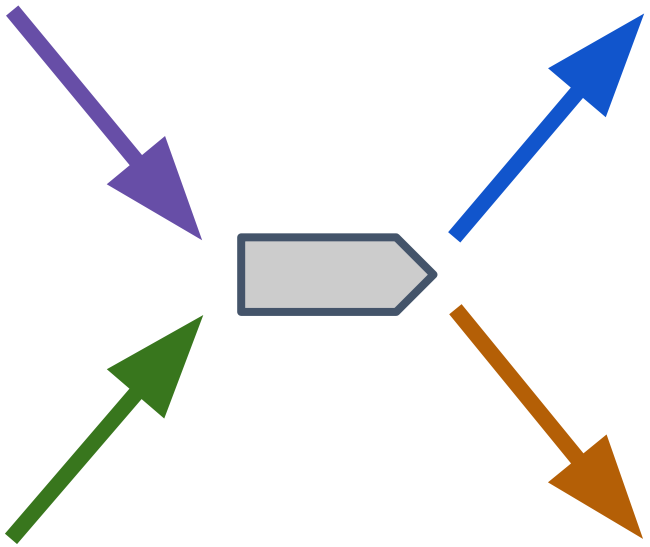
How can we recover long, accurate genome sequences from short reads?



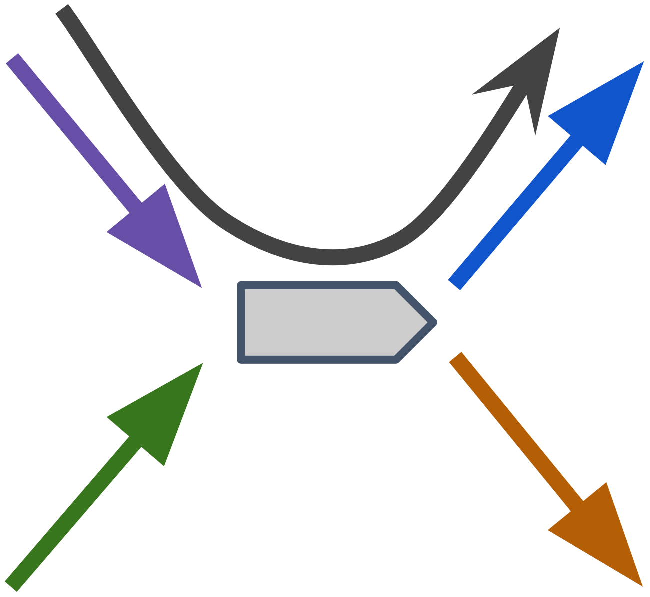
Focus on just one junction at a time



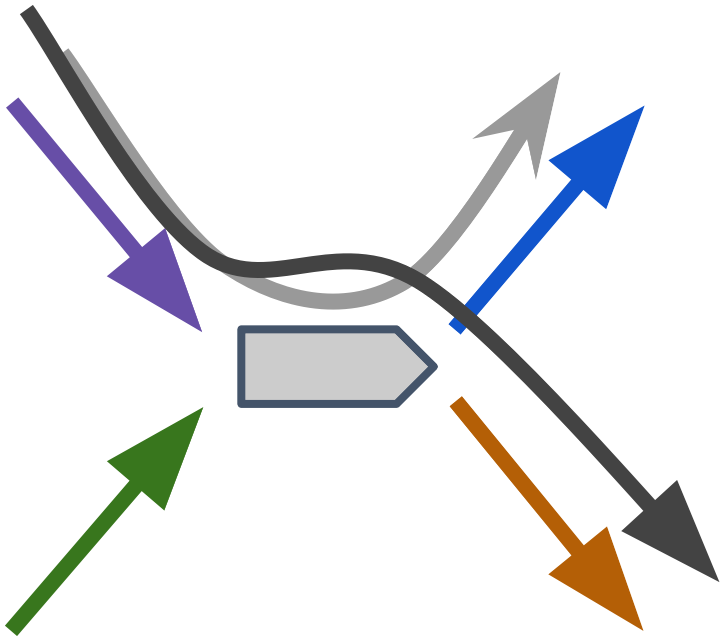
Focus on just one junction at a time

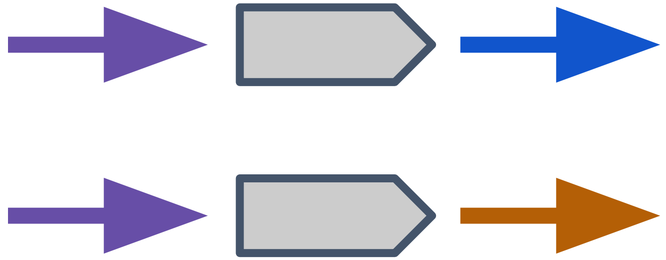


Focus on just one junction at a time

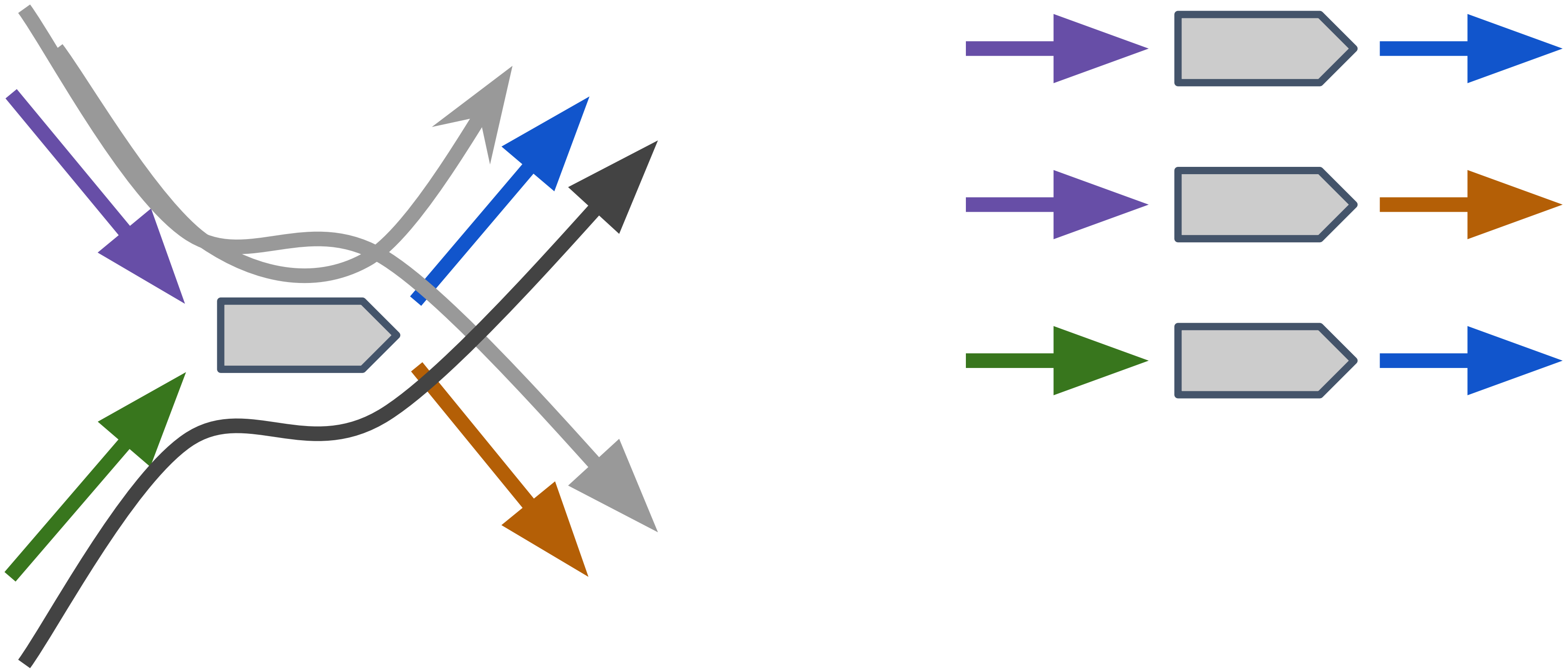


Focus on just one junction at a time

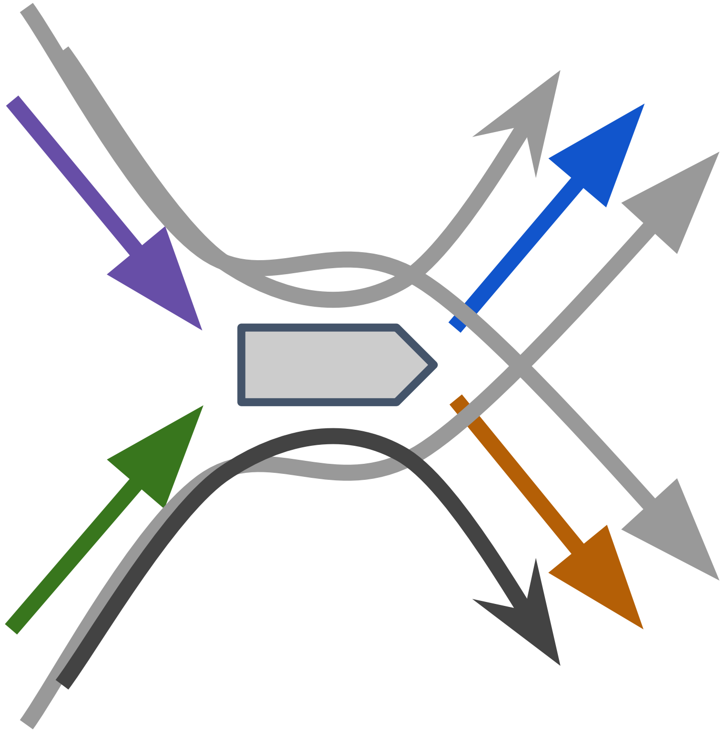


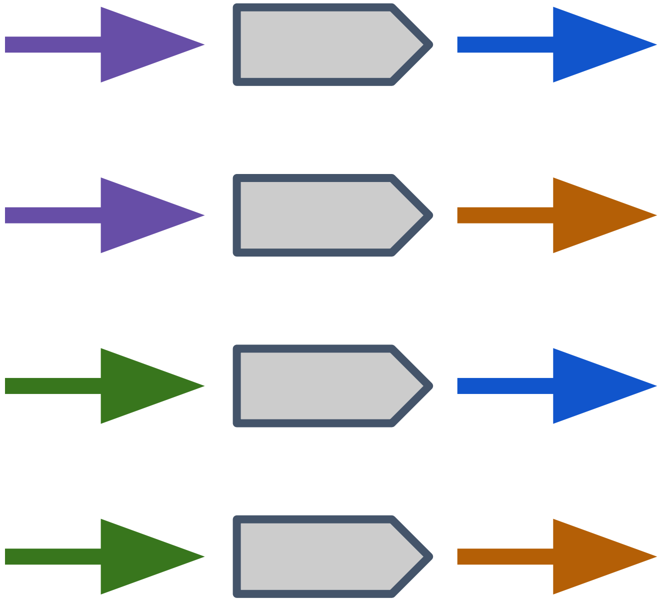


Focus on just one junction at a time

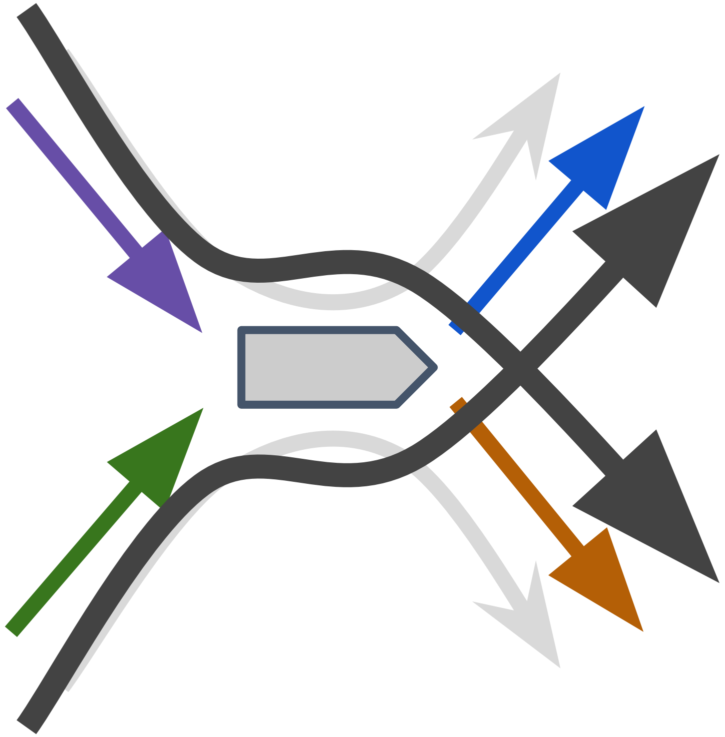


Focus on just one junction at a time





Focus on just one junction at a time
Select local paths



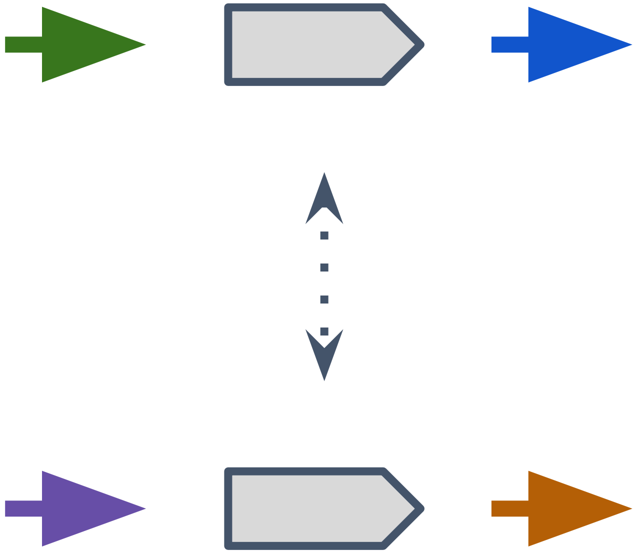
$$\underset{X}{\begin{array}{|c|c|c|c|}\hline \mathbf{1} & \mathbf{1} & 0 & 0 \\ \hline 0 & 0 & \mathbf{1} & \mathbf{1} \\ \hline \mathbf{1} & 0 & \mathbf{1} & 0 \\ \hline 0 & \mathbf{1} & 0 & \mathbf{1} \\ \hline \end{array}} \times \underset{\beta}{\begin{array}{|c|c|c|}\hline p_{1,1} & p_{1,2} & p_{1,3} \\ \hline p_{2,1} & p_{2,2} & p_{2,3} \\ \hline p_{3,1} & p_{3,2} & p_{3,3} \\ \hline p_{4,1} & p_{4,2} & p_{4,3} \\ \hline \end{array}} \approx \underset{Y}{\begin{array}{|c|c|c|}\hline e_{1,1} & e_{1,2} & e_{1,3} \\ \hline e_{2,1} & e_{2,2} & e_{2,3} \\ \hline e_{3,1} & e_{3,2} & e_{3,3} \\ \hline e_{4,1} & e_{4,2} & e_{4,3} \\ \hline \end{array}}$$

Sparse linear regression across multiple samples

Focus on just one junction at a time

Select local paths

Unzip

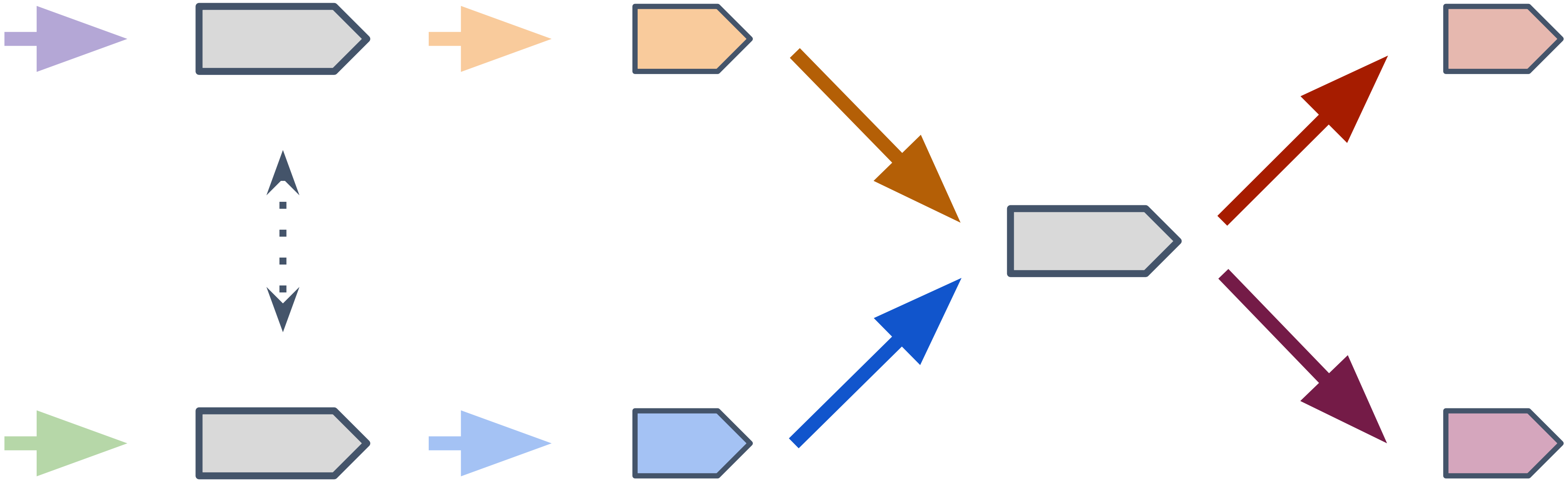


Focus on just one junction at a time

Select local paths

Unzip

Repeat

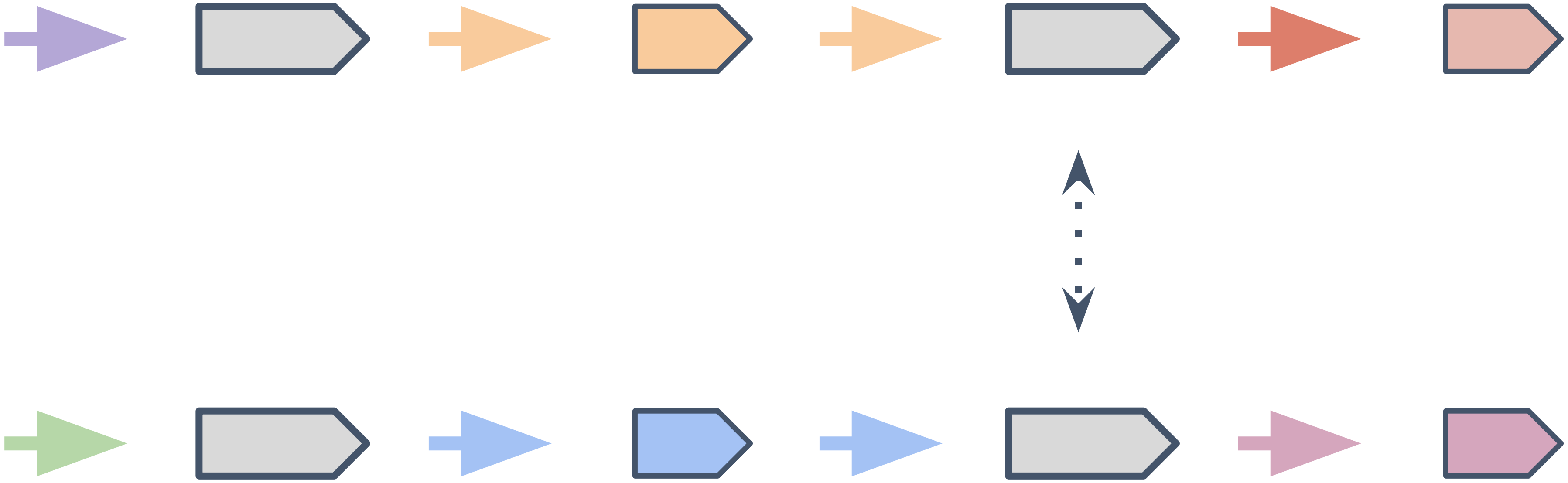


Focus on just one junction at a time

Select local paths

Unzip

Repeat

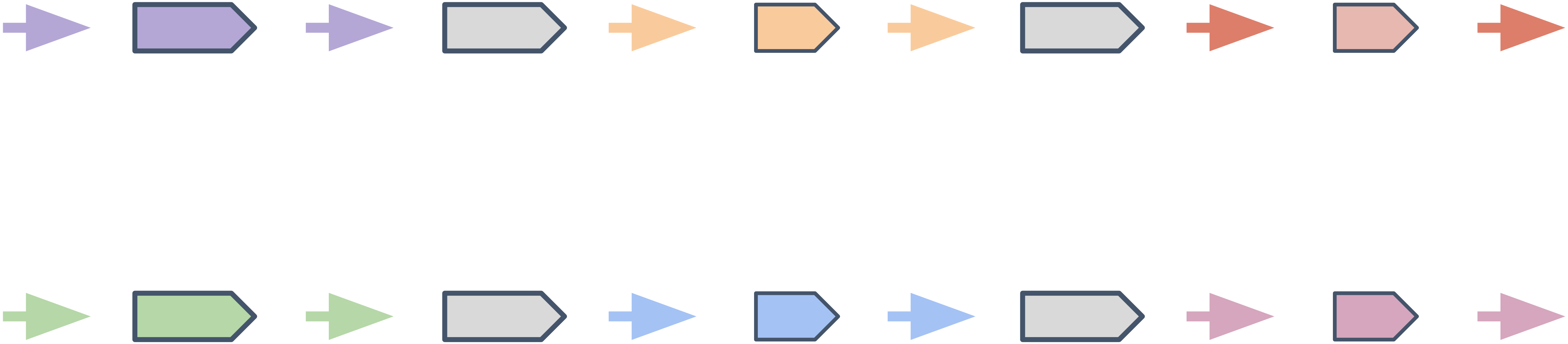


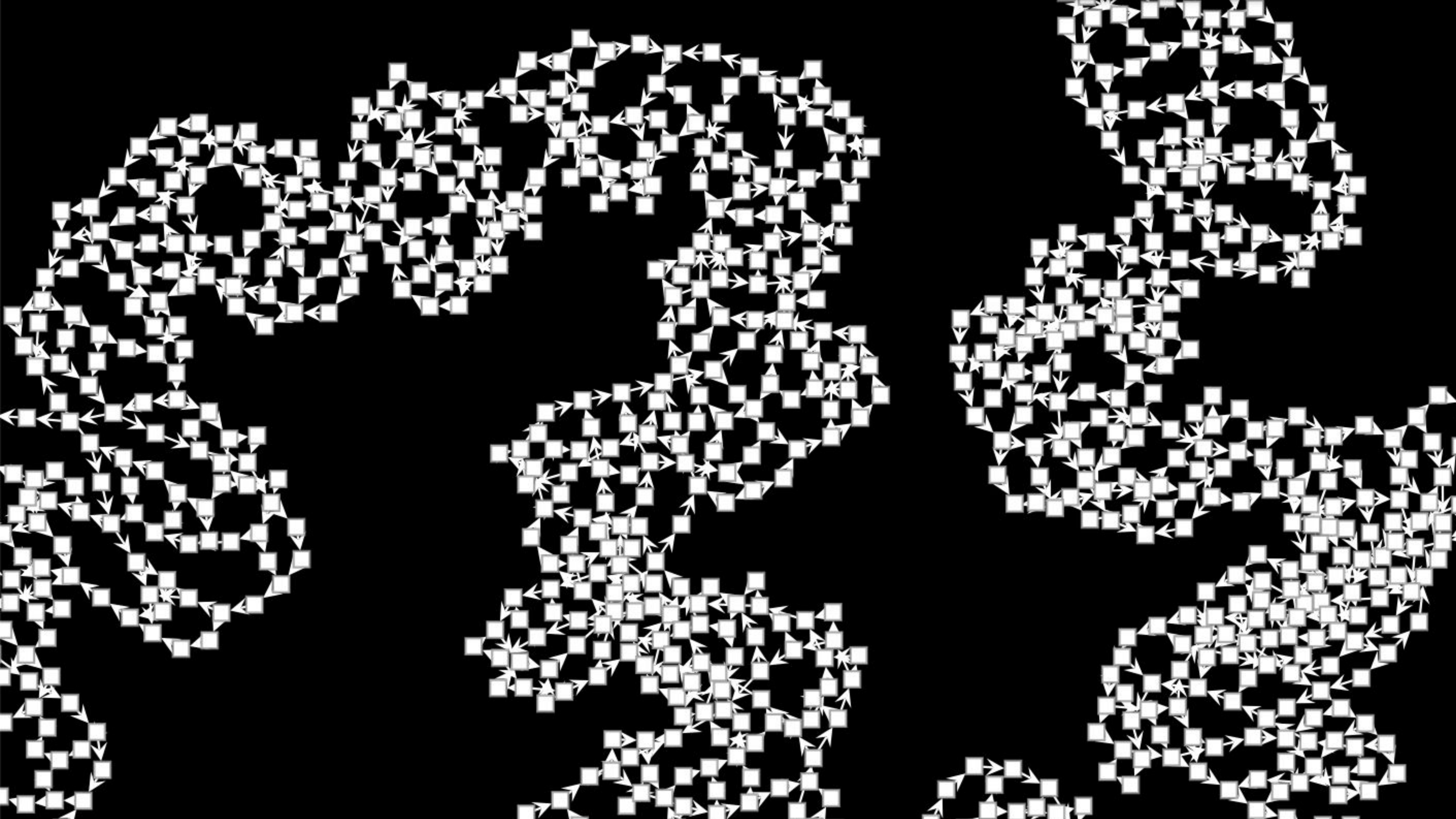
Focus on just one junction at a time

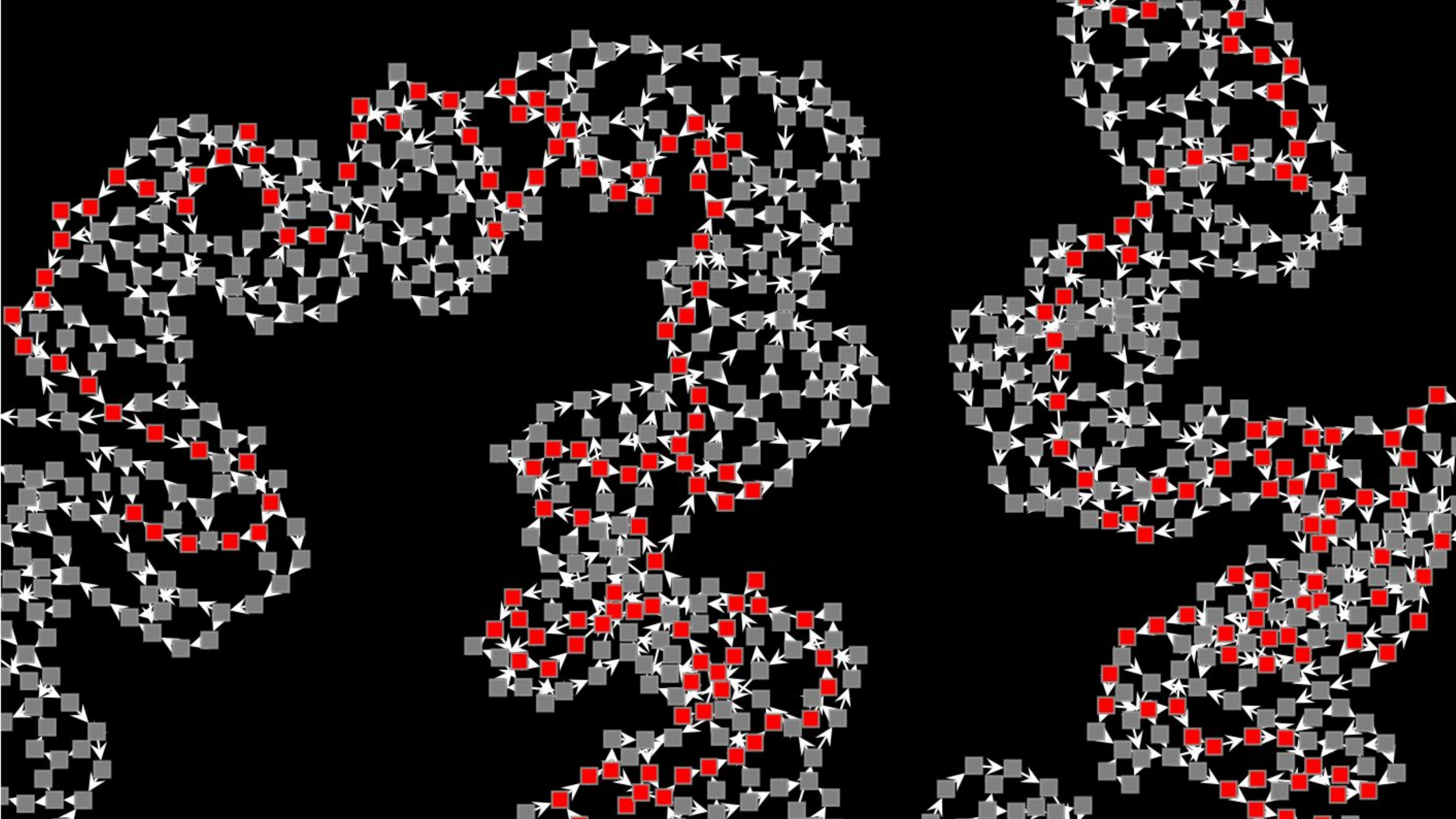
Select local paths

Unzip

Repeat



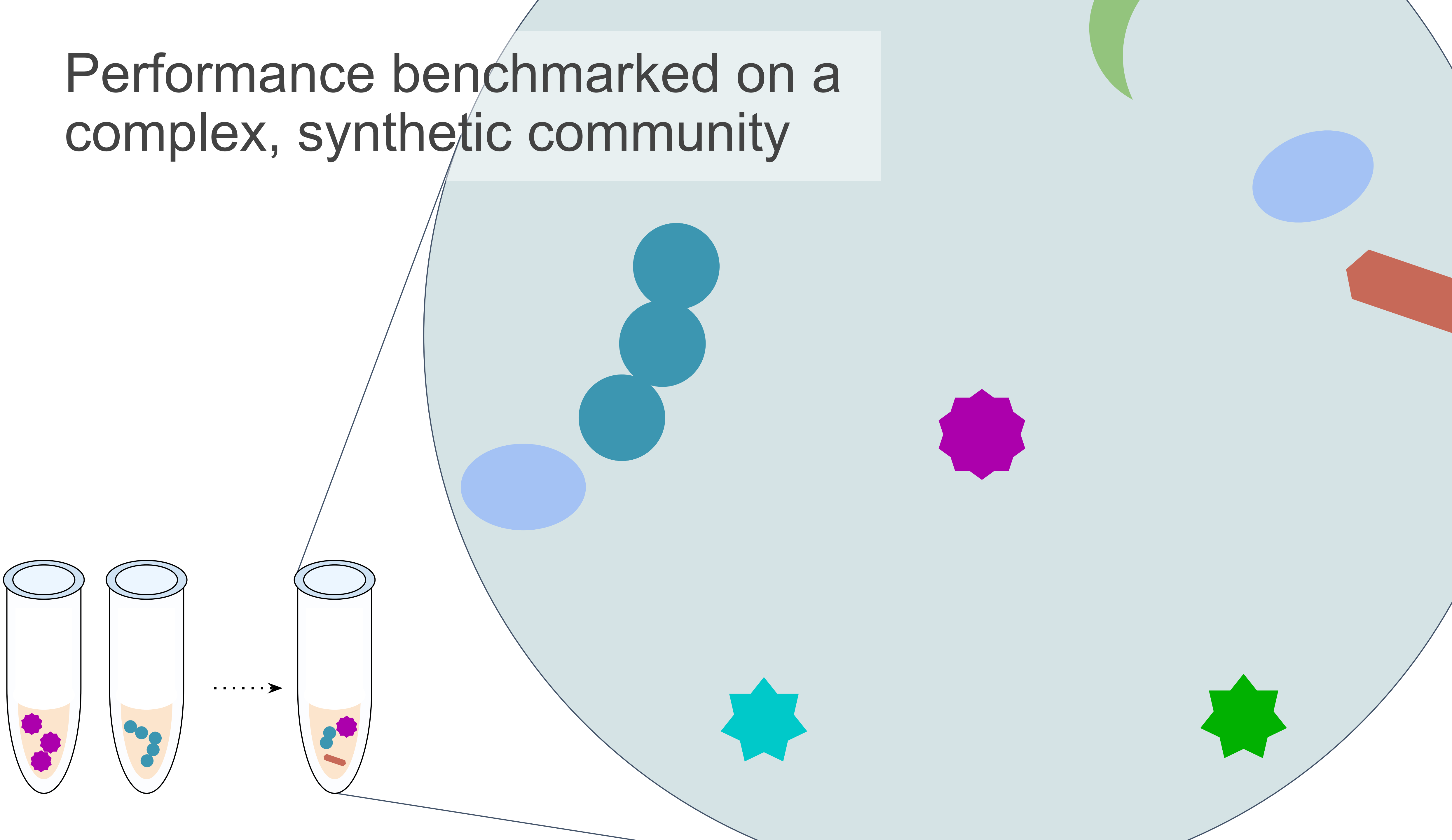




Gladstone Scientific Retreat 2024

Strain-resolved discovery

Performance benchmarked on a complex, synthetic community



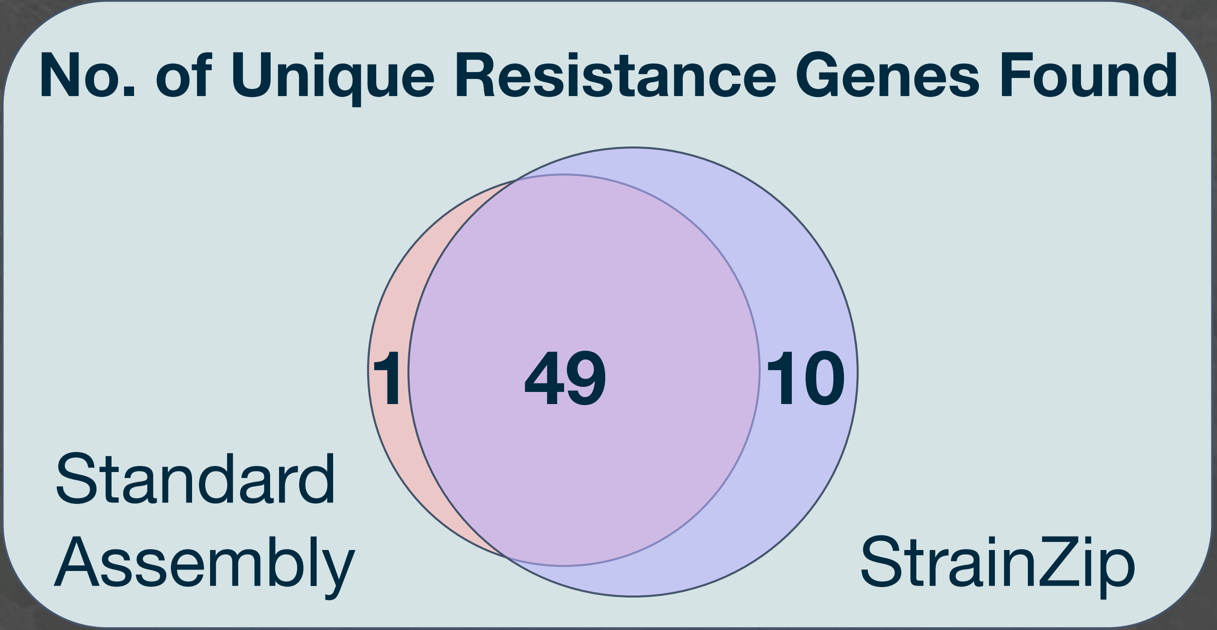
Antibiotic resistance genes are widespread in the gut microbiome

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



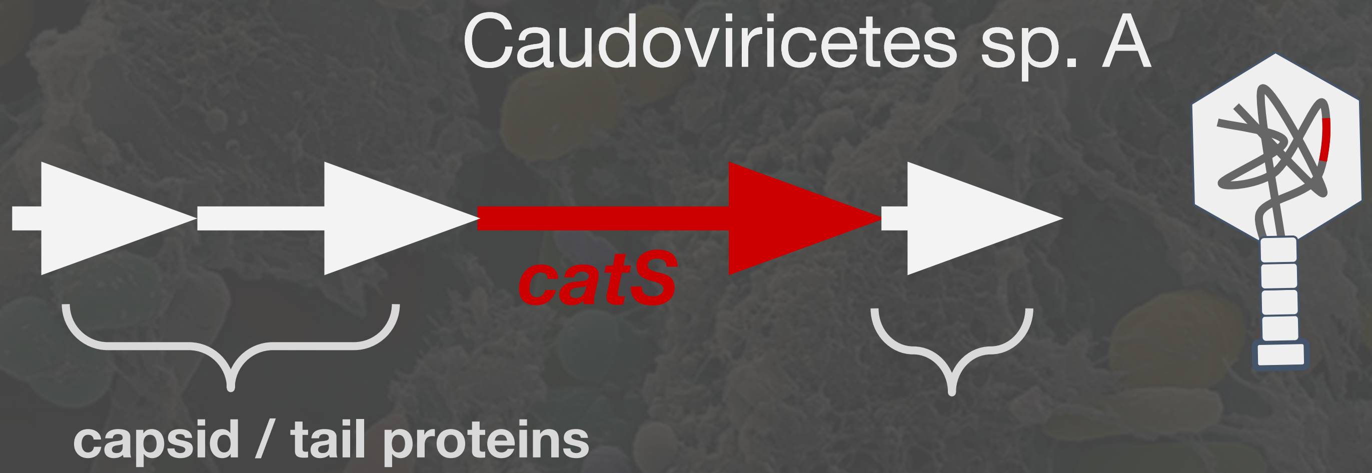
No. of Unique Resistance Genes Found

Antibiotic resistance genes are widespread in the gut microbiome

- Detection can inform treatment

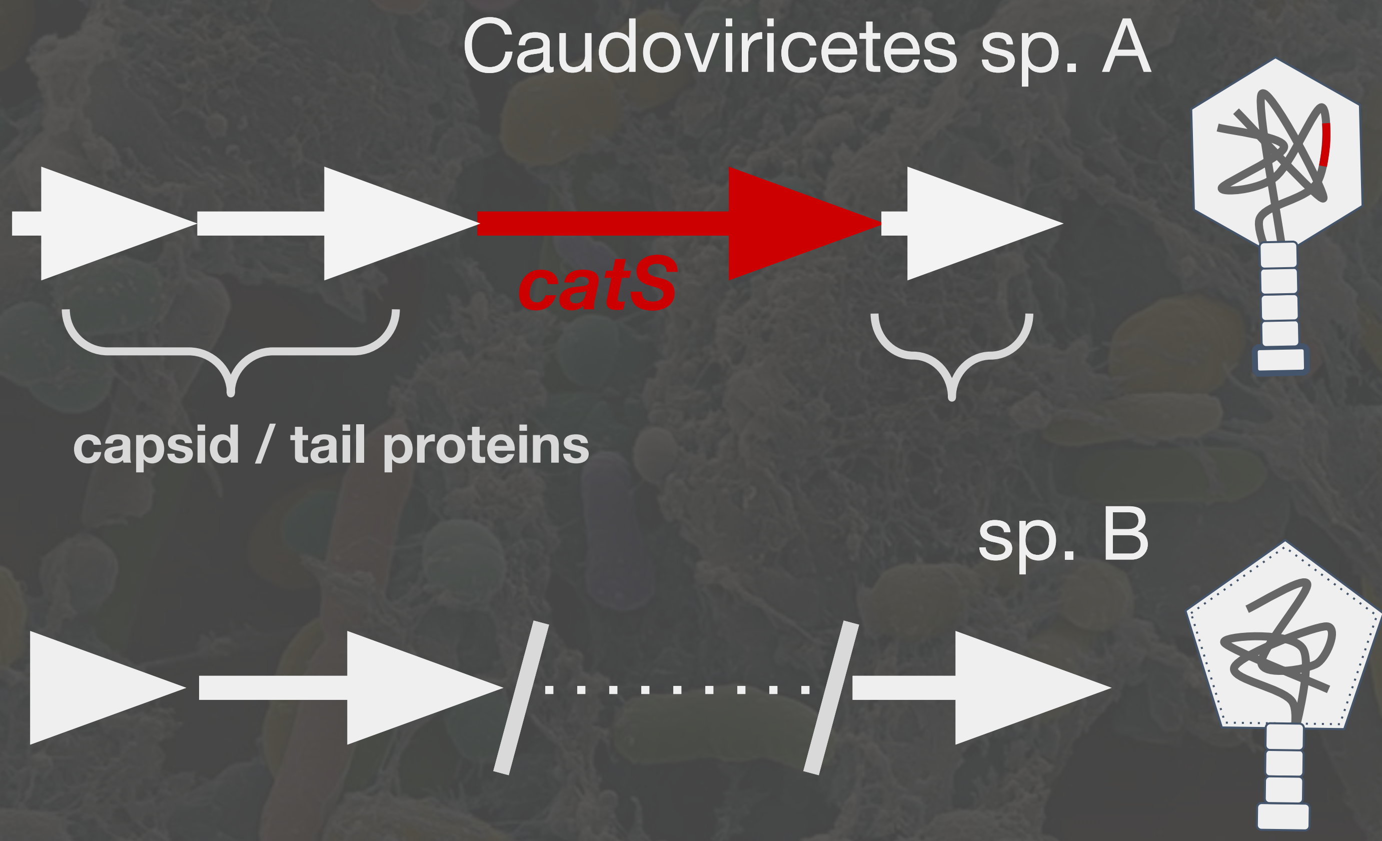
Antibiotic resistance genes are widespread in the gut microbiome

- Detection can inform treatment
- Can be carried in phage genomes



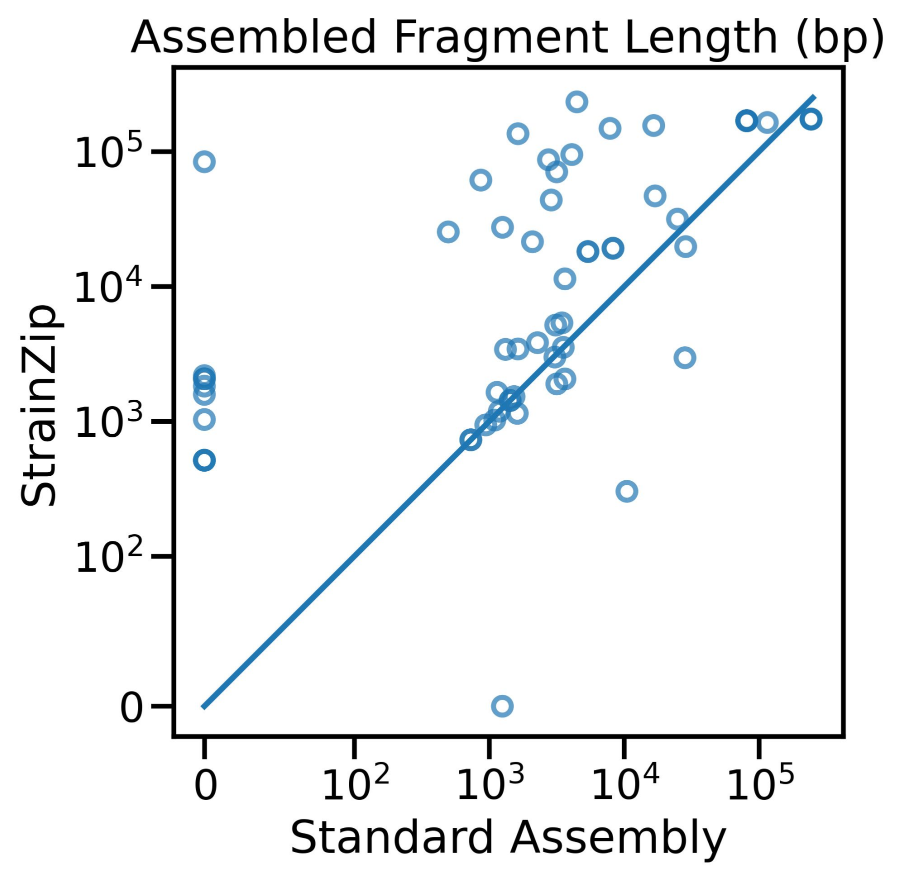
Antibiotic resistance genes are widespread in the gut microbiome

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

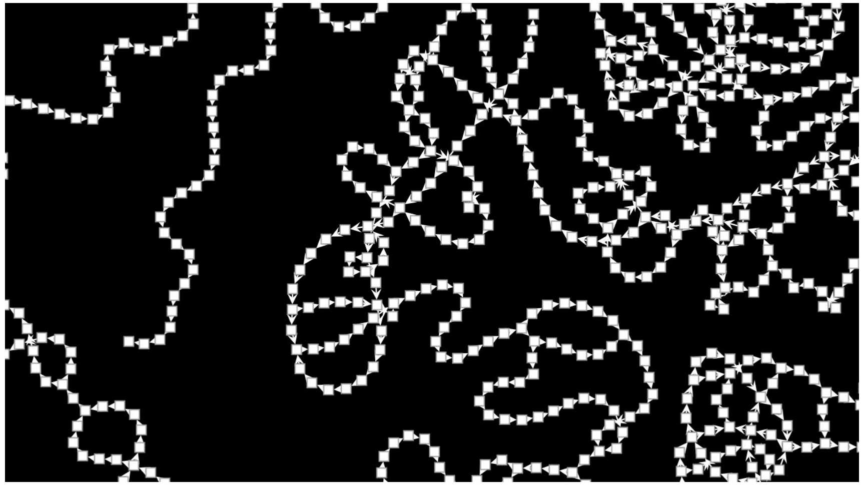


Antibiotic resistance genes are widespread in the gut microbiome

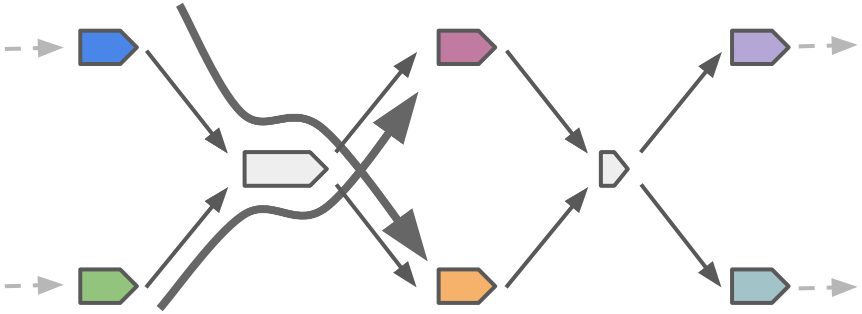
- Detection can inform treatment
- Can be carried in phage genomes
- Long sequence fragments provide useful information



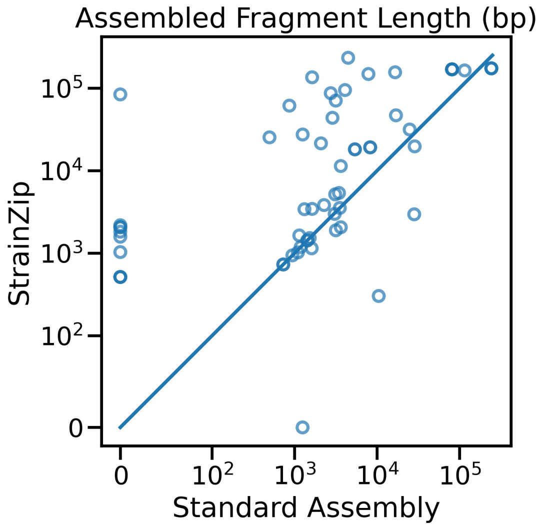
Complex Metagenome Graphs



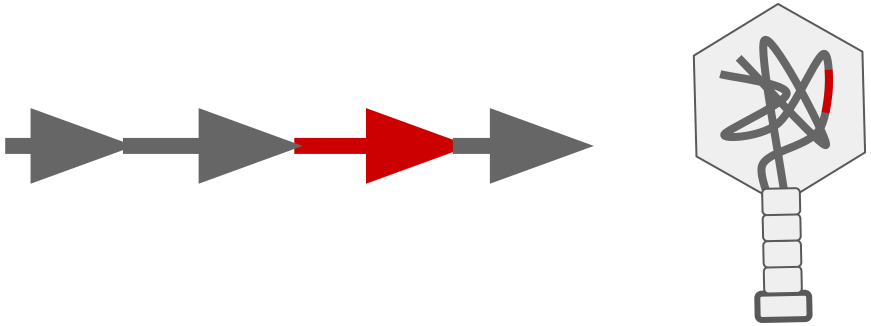
StrainZip Iteratively Unzips Junctions



Strain-Resolved Metagenomics



Antibiotic Resistance Potential of Phage



Gladstone Scientific Retreat 2024

Thank You!
Questions?