

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





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**

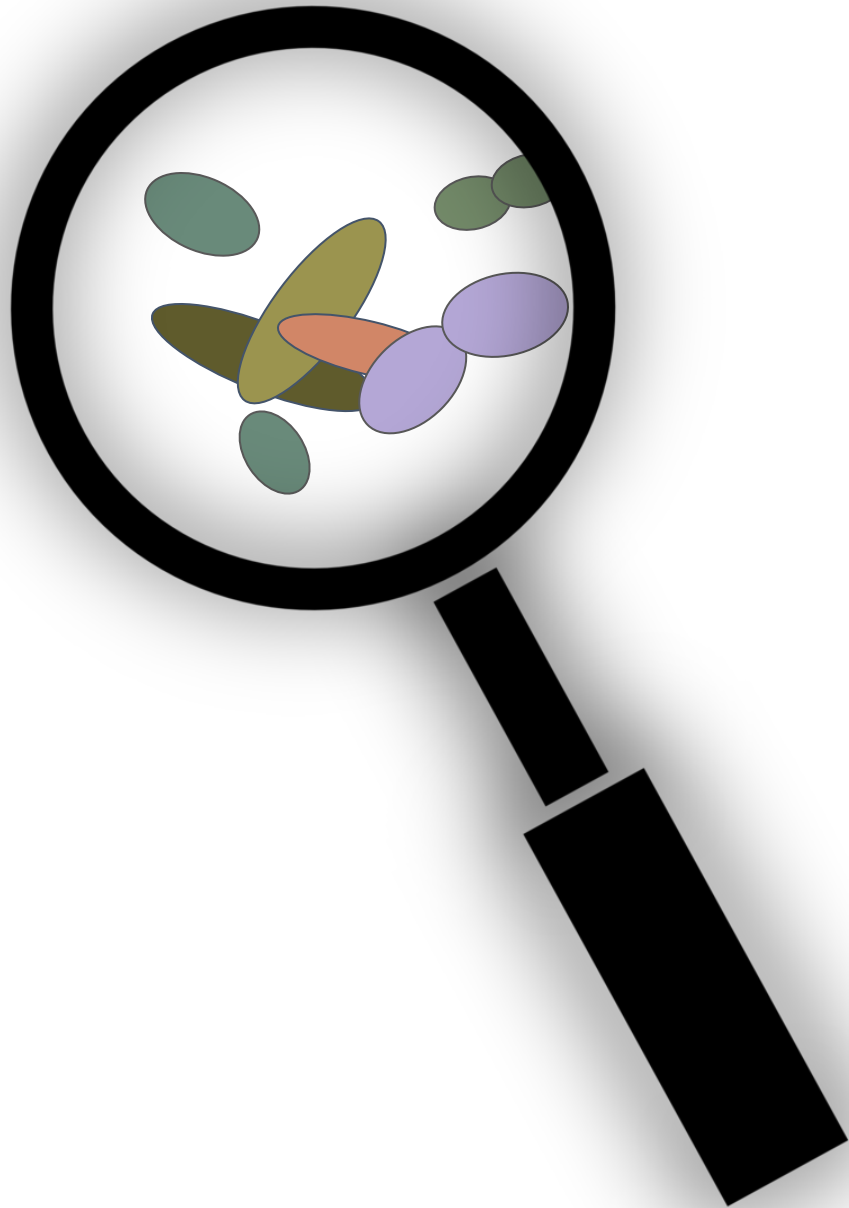


# The Gut Microbiome is Challenging



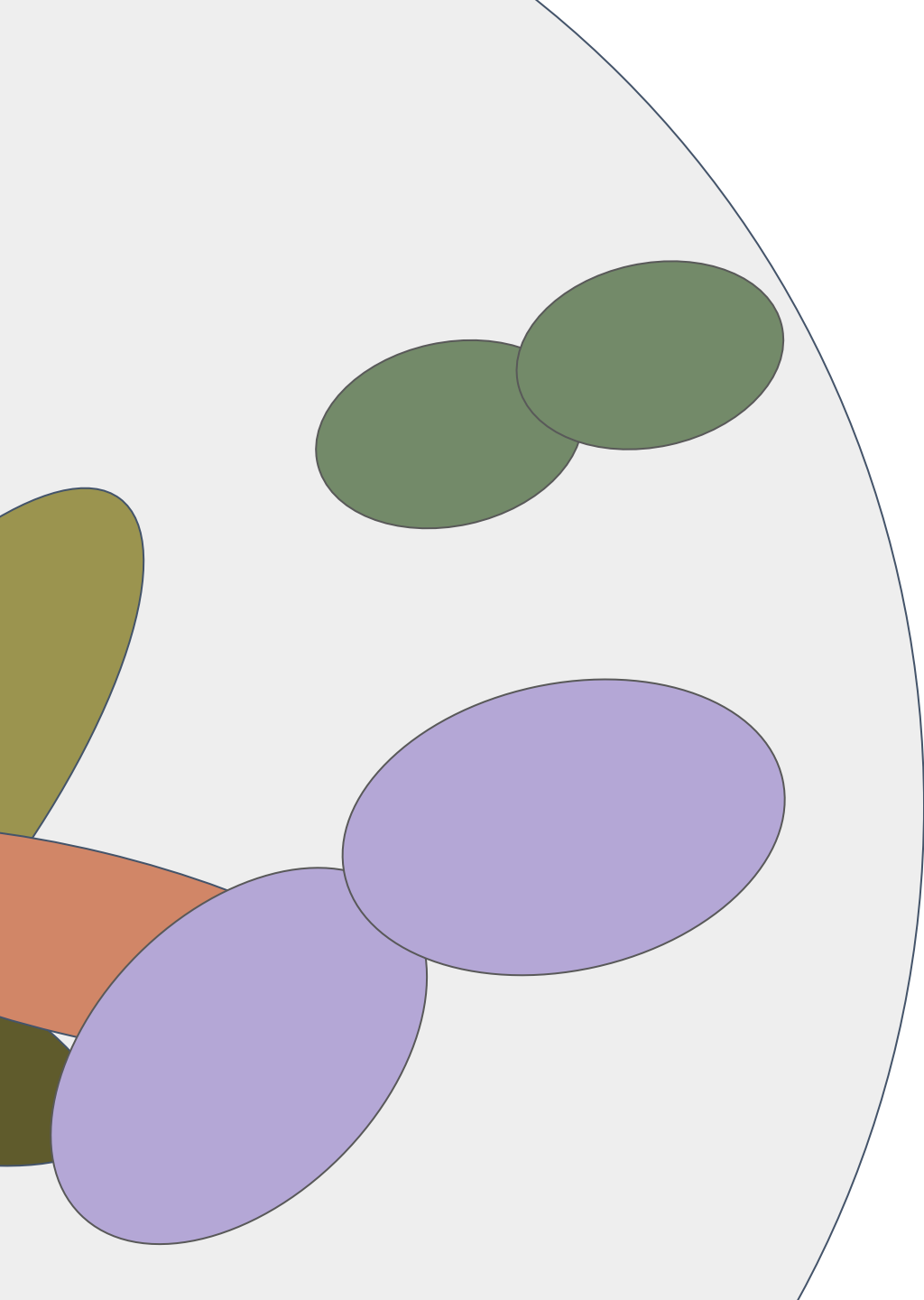
- Enormous number of species
- Highly dynamic across people and time
- Very hard to study in the lab
- **Strains within species have different gene content and functional potential**





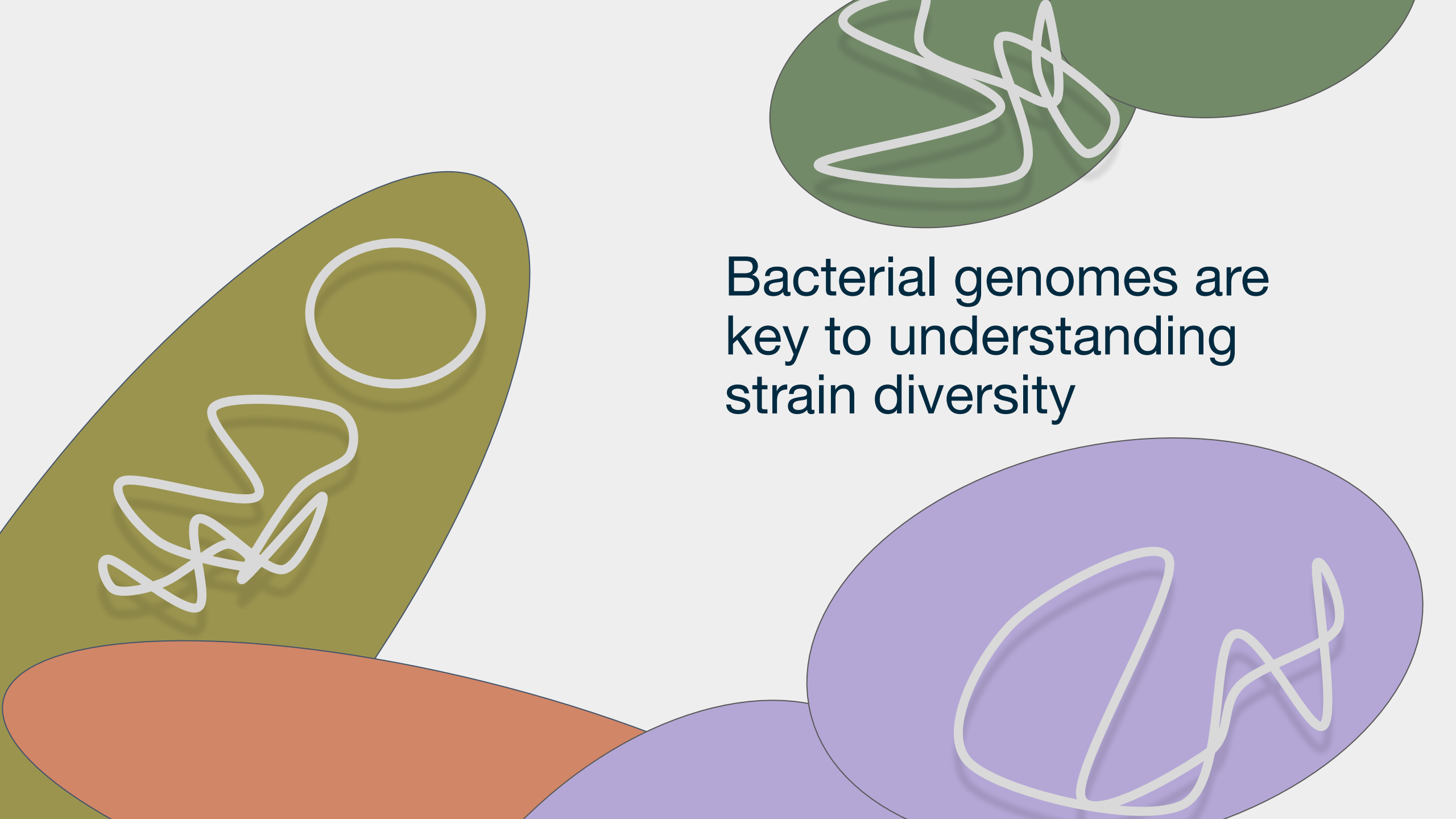
Bacterial genomes are  
key to understanding  
strain diversity





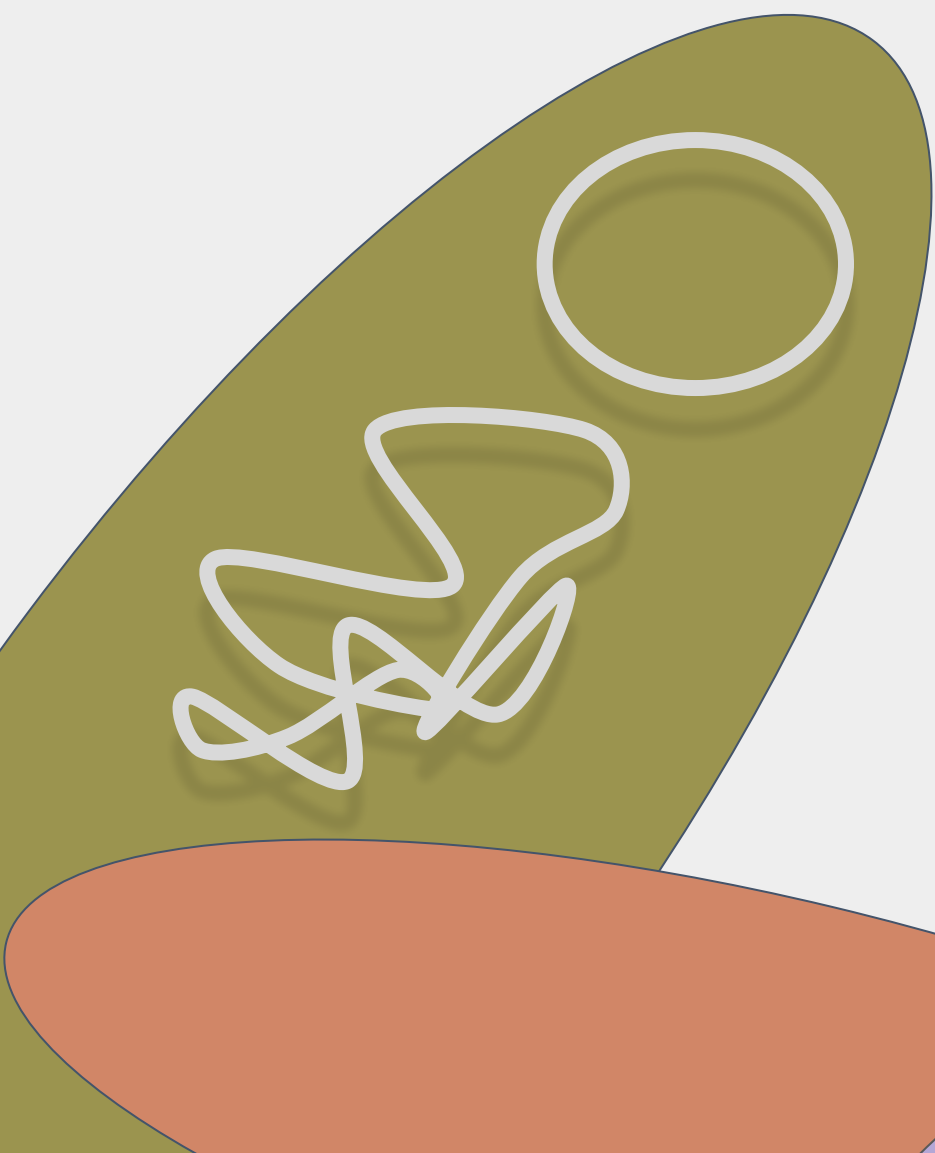
Bacterial genomes are  
key to understanding  
strain diversity



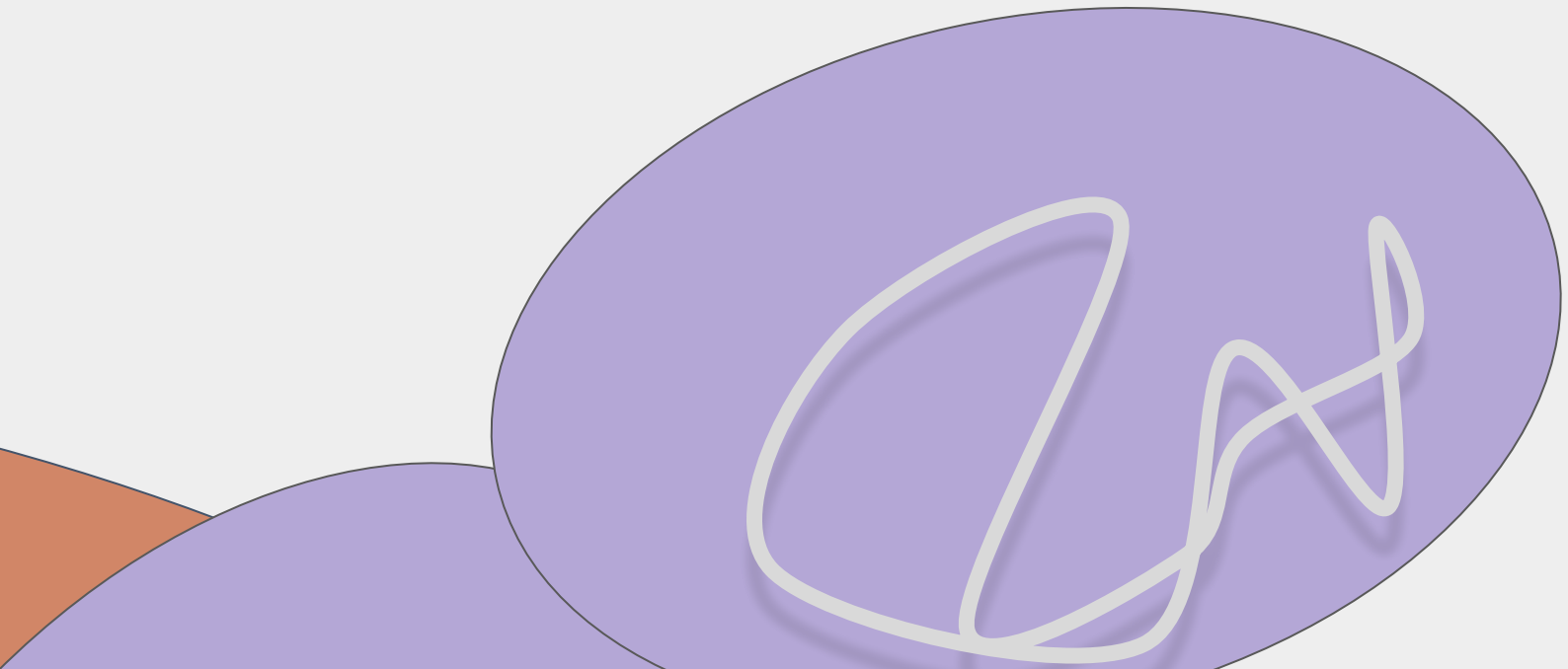
The image features a white background with several overlapping, semi-transparent colored shapes: a dark green oval at the top right, a light green oval on the left, a purple oval at the bottom right, and an orange oval at the bottom left. Inside these shapes are white line drawings representing bacterial genomes. The dark green oval contains a complex, tangled circular structure. The light green oval contains a circular structure with a smaller circle inside it and a tangled structure below. The purple oval contains a large circular structure and a tangled structure to its right. The text "Bacterial genomes are key to understanding strain diversity" is centered in the white space between the shapes.

Bacterial genomes are key to understanding strain diversity

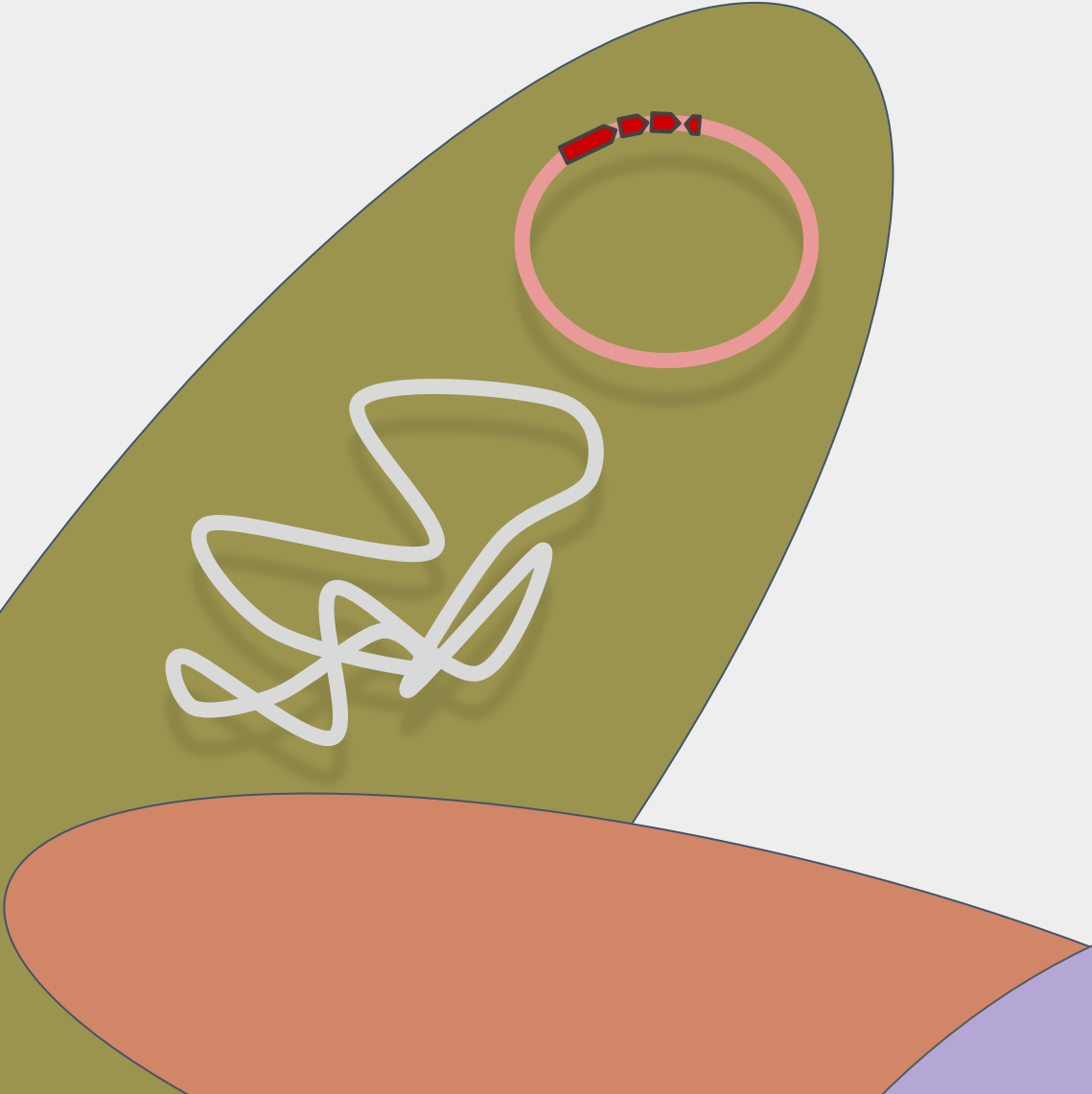




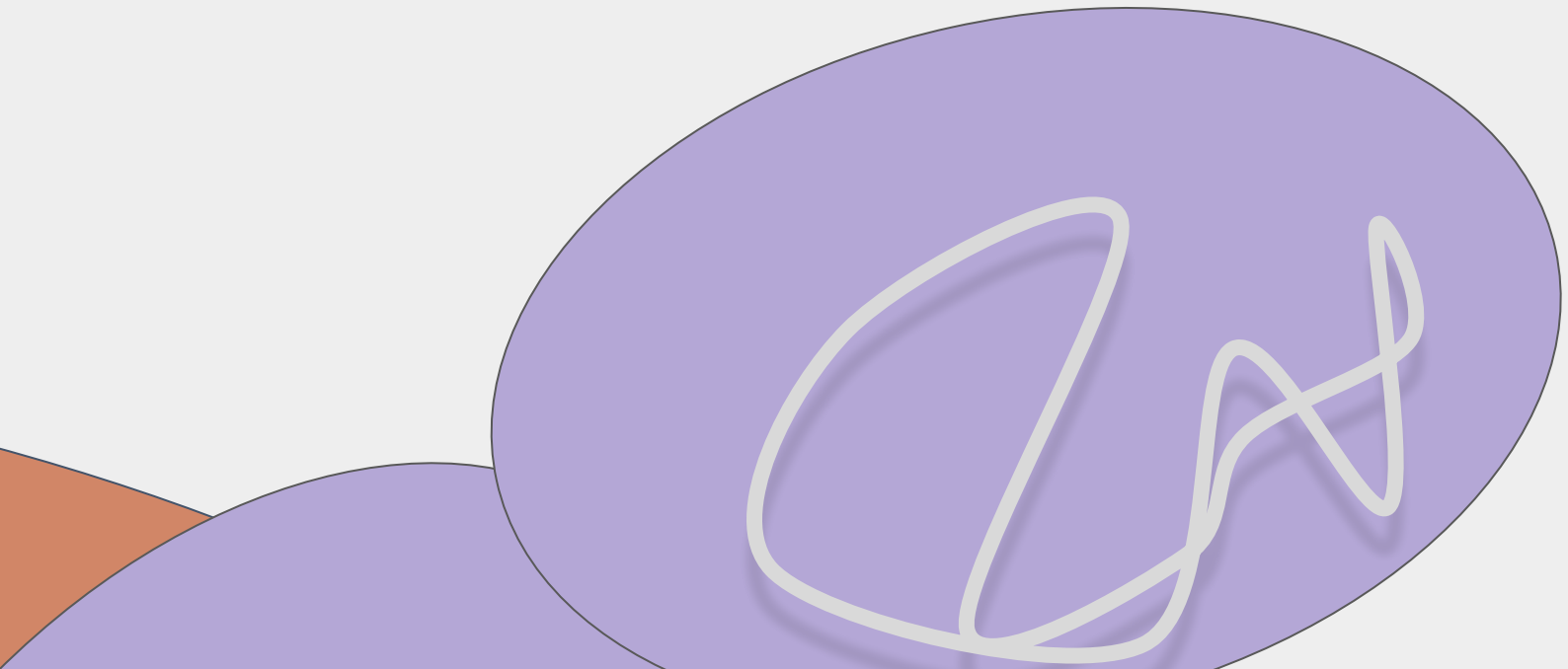
Bacterial genomes are key to understanding strain diversity



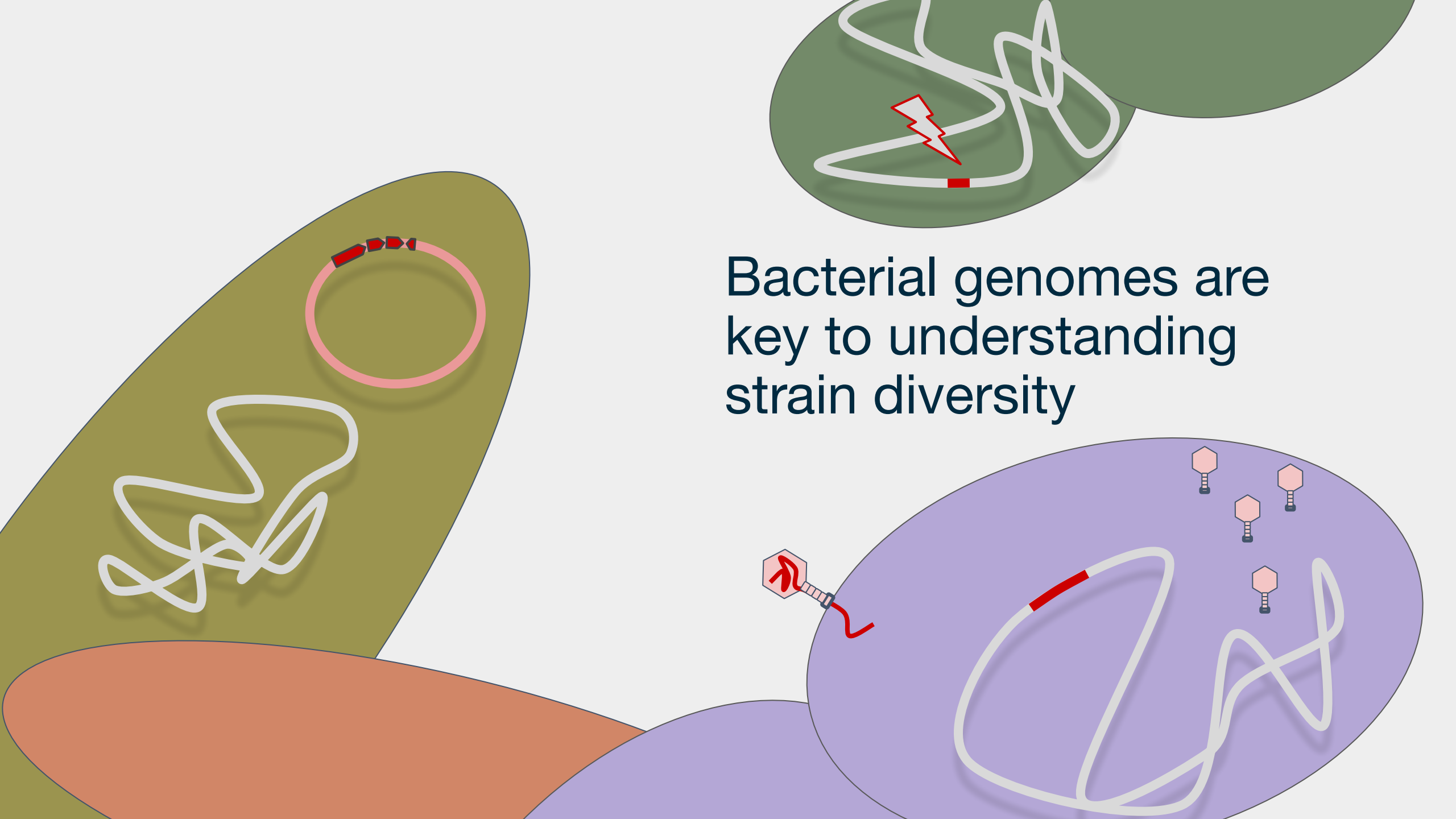




Bacterial genomes are key to understanding strain diversity





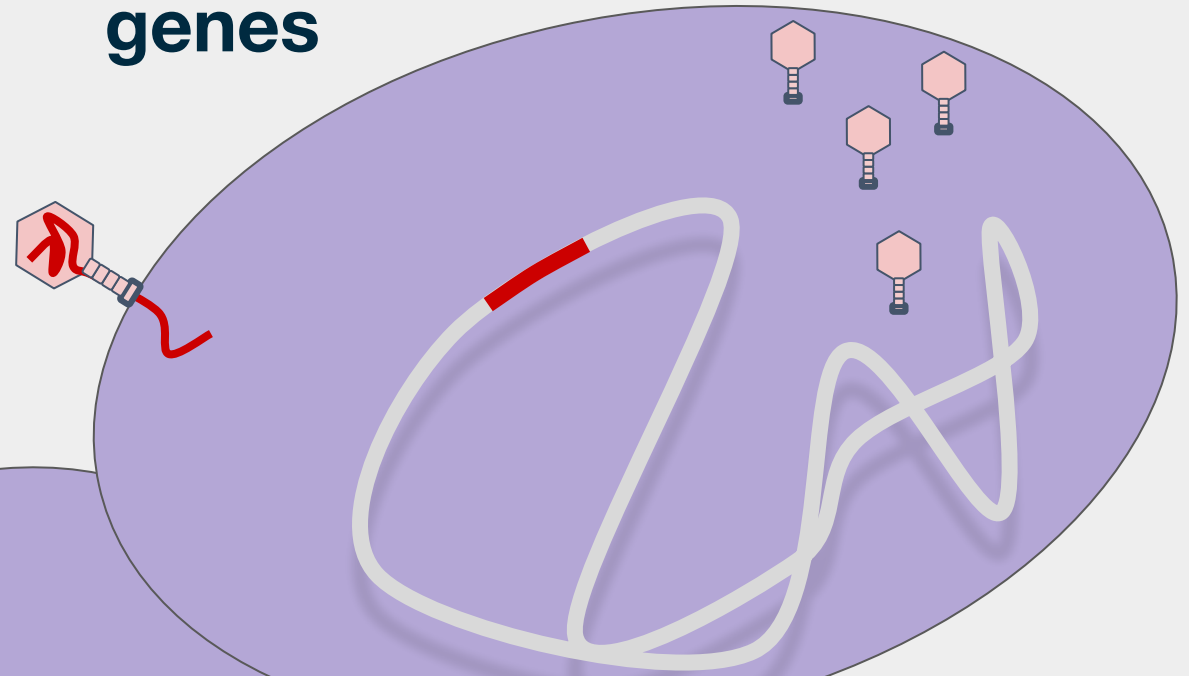
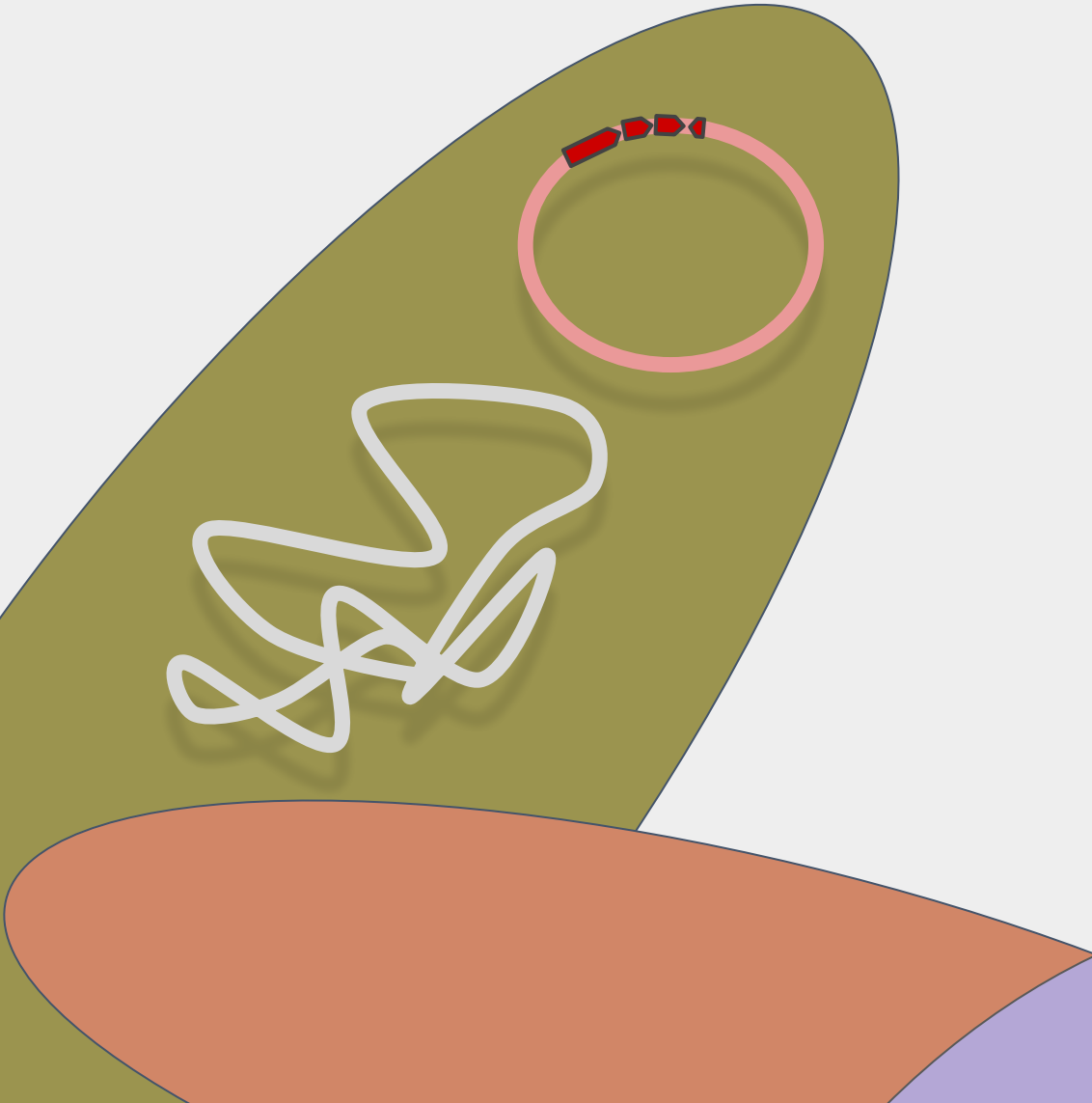


Bacterial genomes are key to understanding strain diversity



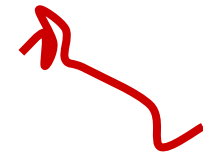
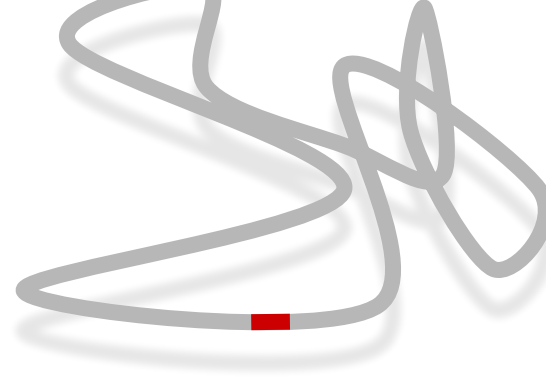
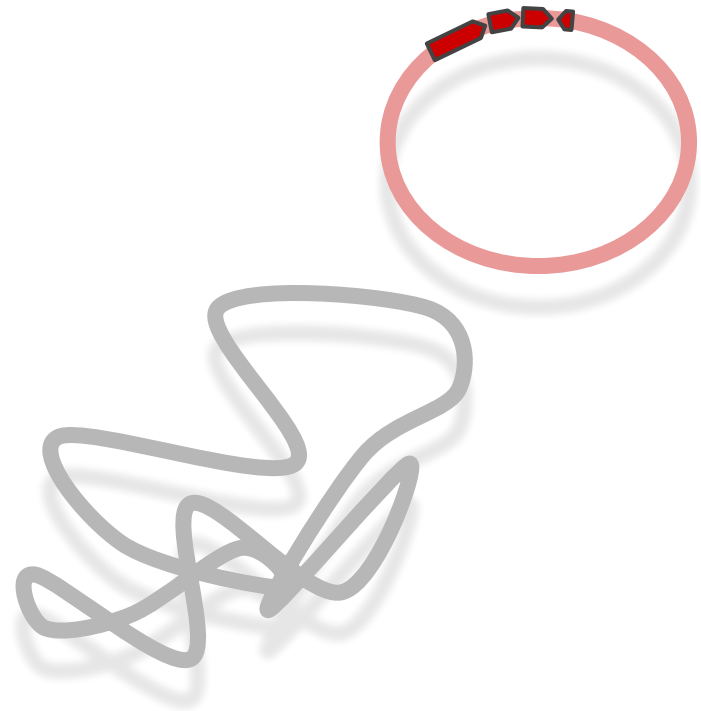


**Phage encoded  
antibiotic resistance  
genes**





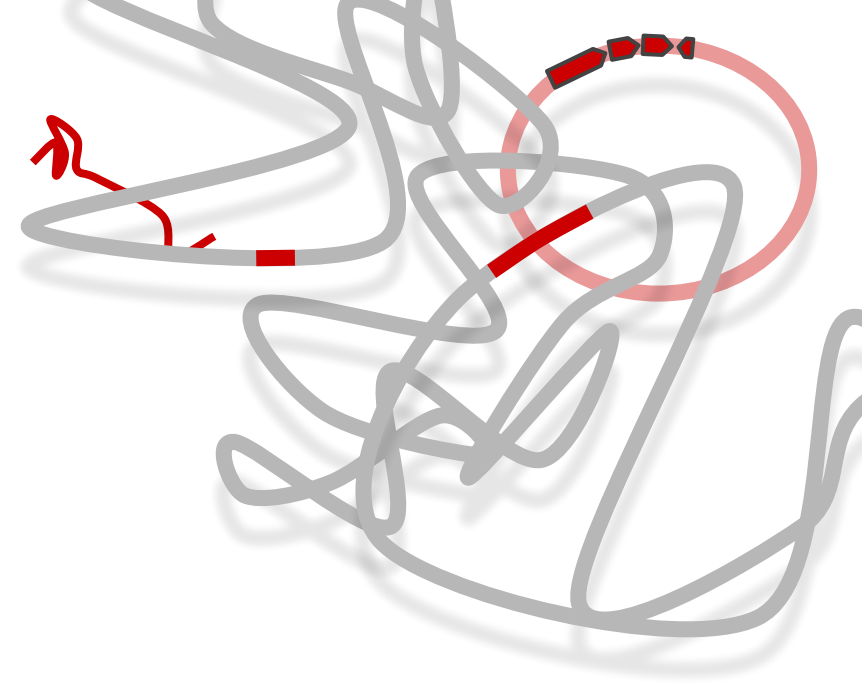
# Metagenomic sequencing surveys all genomes





# Short-read, shotgun metagenomes enable modern microbiome science

Requirements:





# Short-read, shotgun metagenomes enable modern microbiome science

## Requirements:

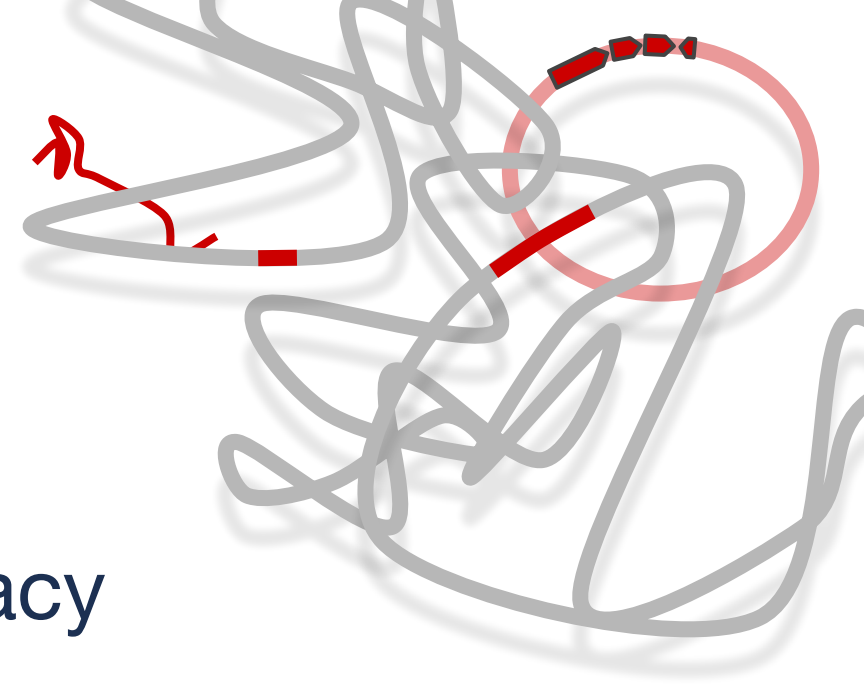
- strain-resolved genome sequences
- high accuracy



# Short-read, shotgun metagenomes enable modern microbiome science

## Requirements:

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





# Turning short reads into long sequences

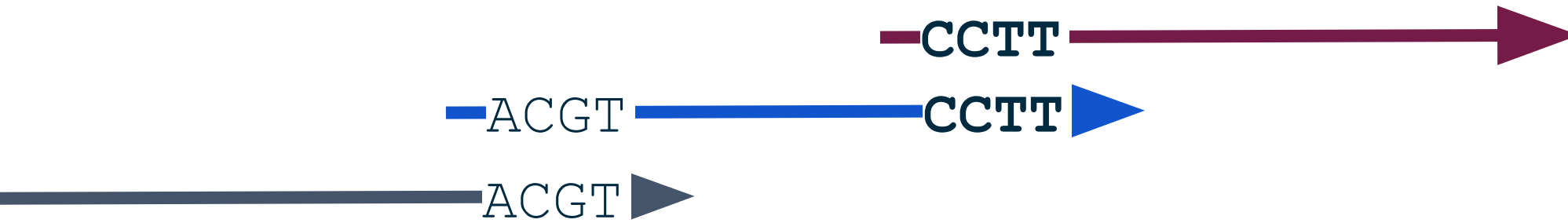


# Turning short reads into long sequences

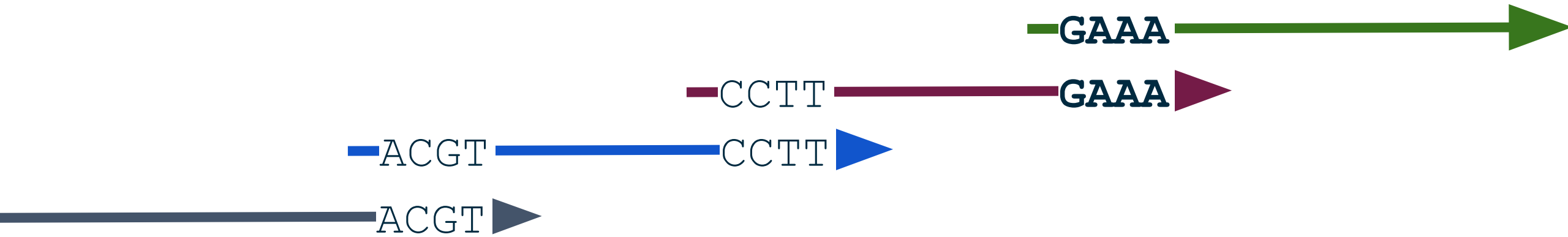




# Turning short reads into long sequences

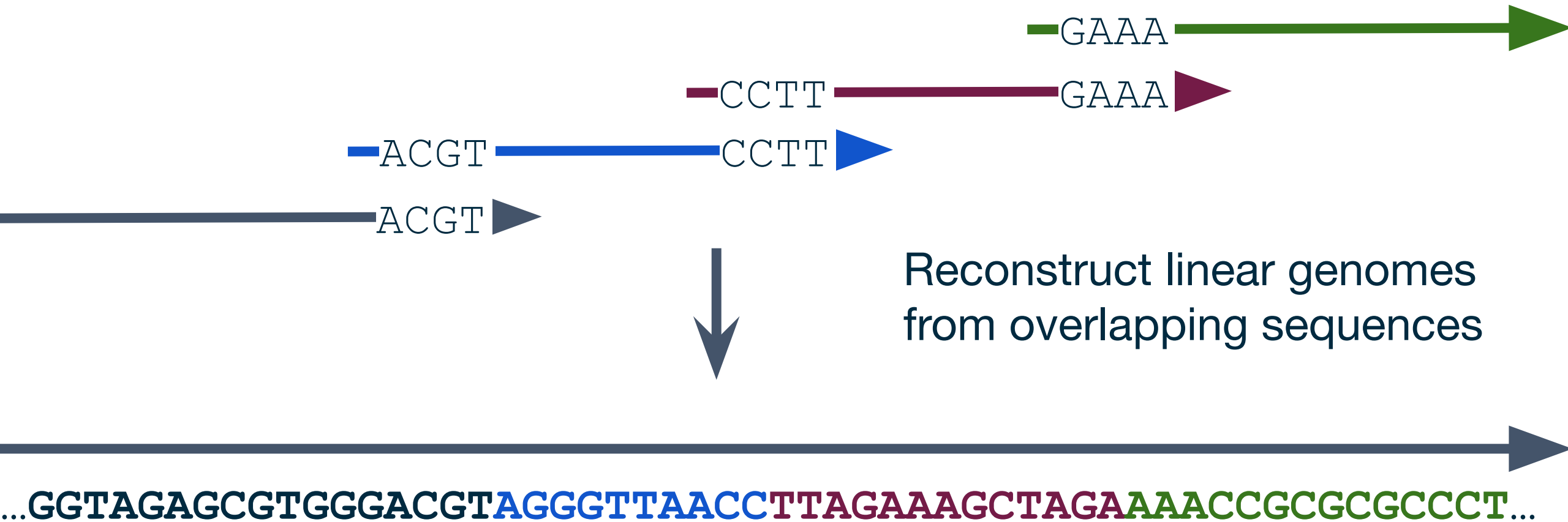


# Turning short reads into long sequences





# Turning short reads into long sequences



# Problem:



# Problem: Closely related strains make read-chaining ambiguous

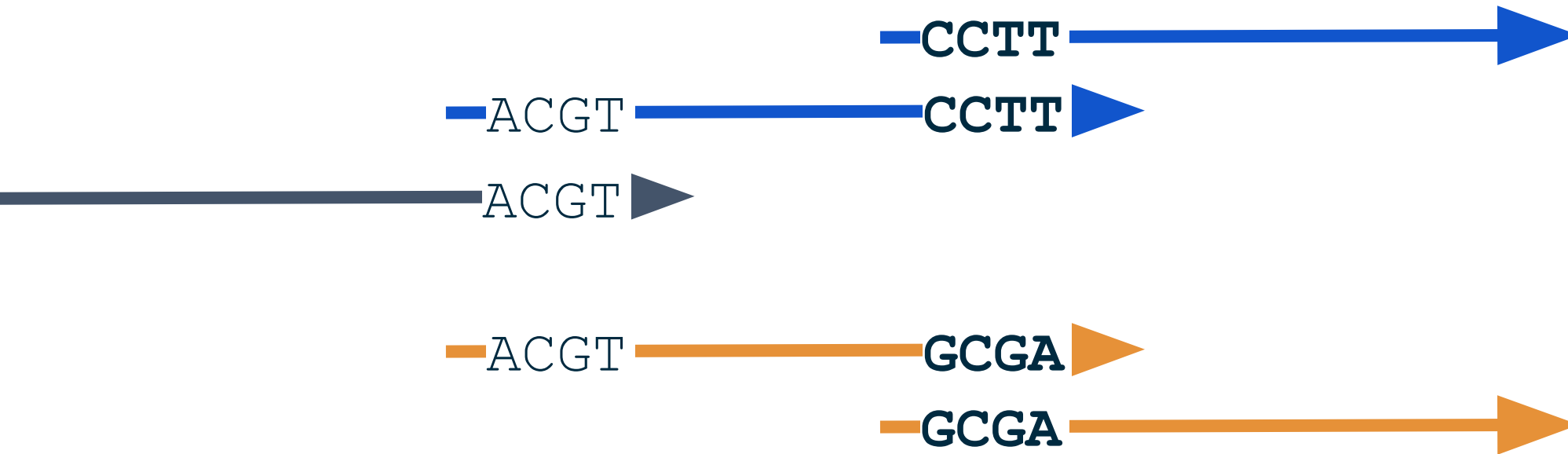
—ACGT ————— CCTT ▶

—ACGT ▶

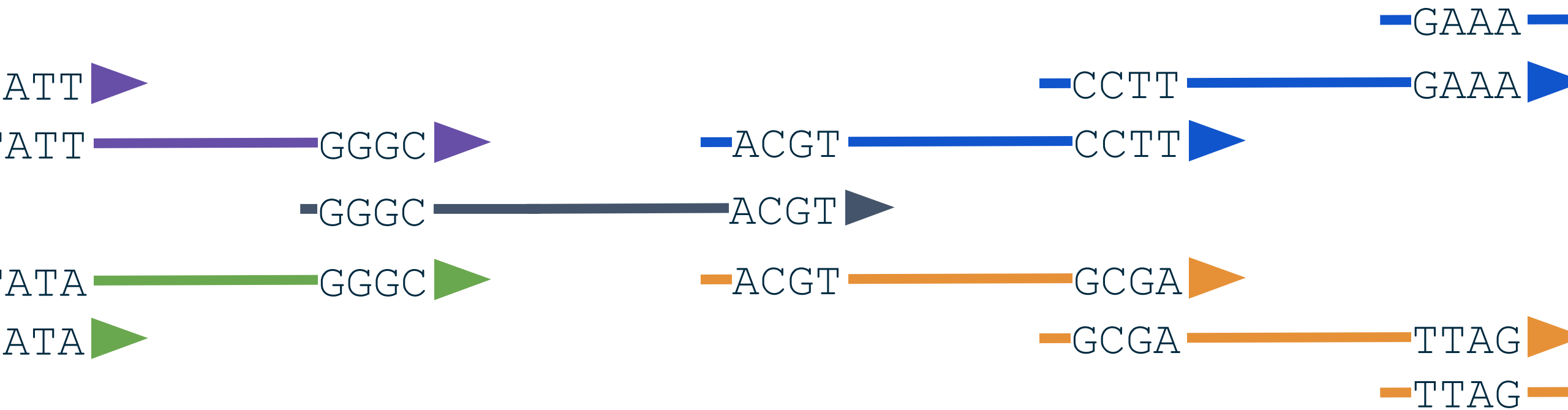
—ACGT ————— GCGA ▶



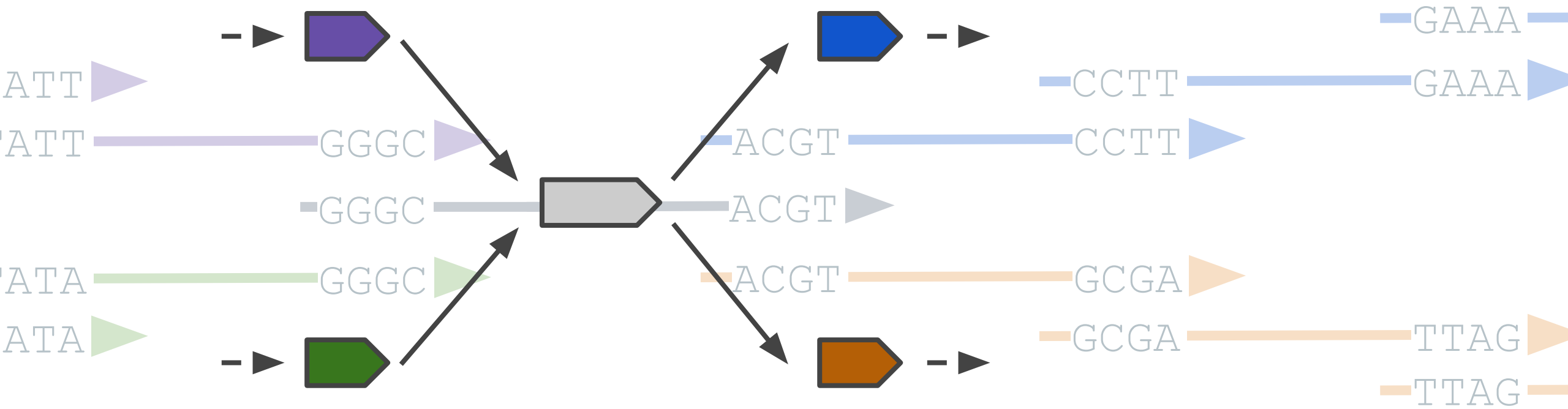
# Problem: Closely related strains make read-chaining ambiguous



# Problem: Closely related strains make read-chaining ambiguous

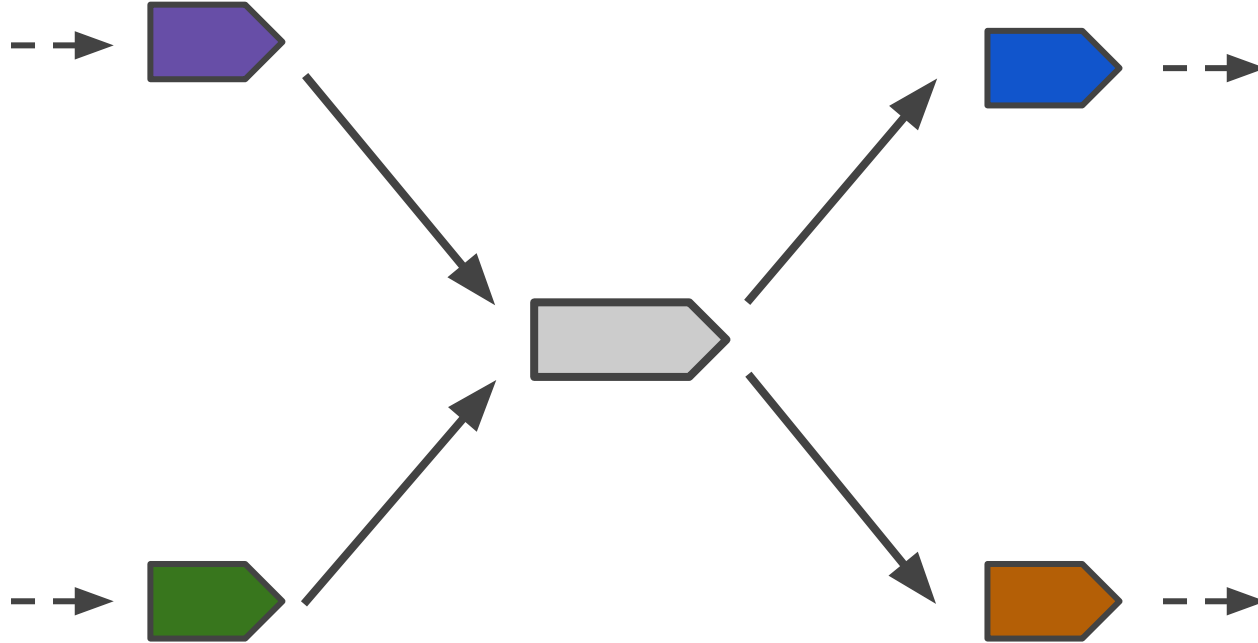


Can be represented as  
a graph of sequences  
linked by their overlaps

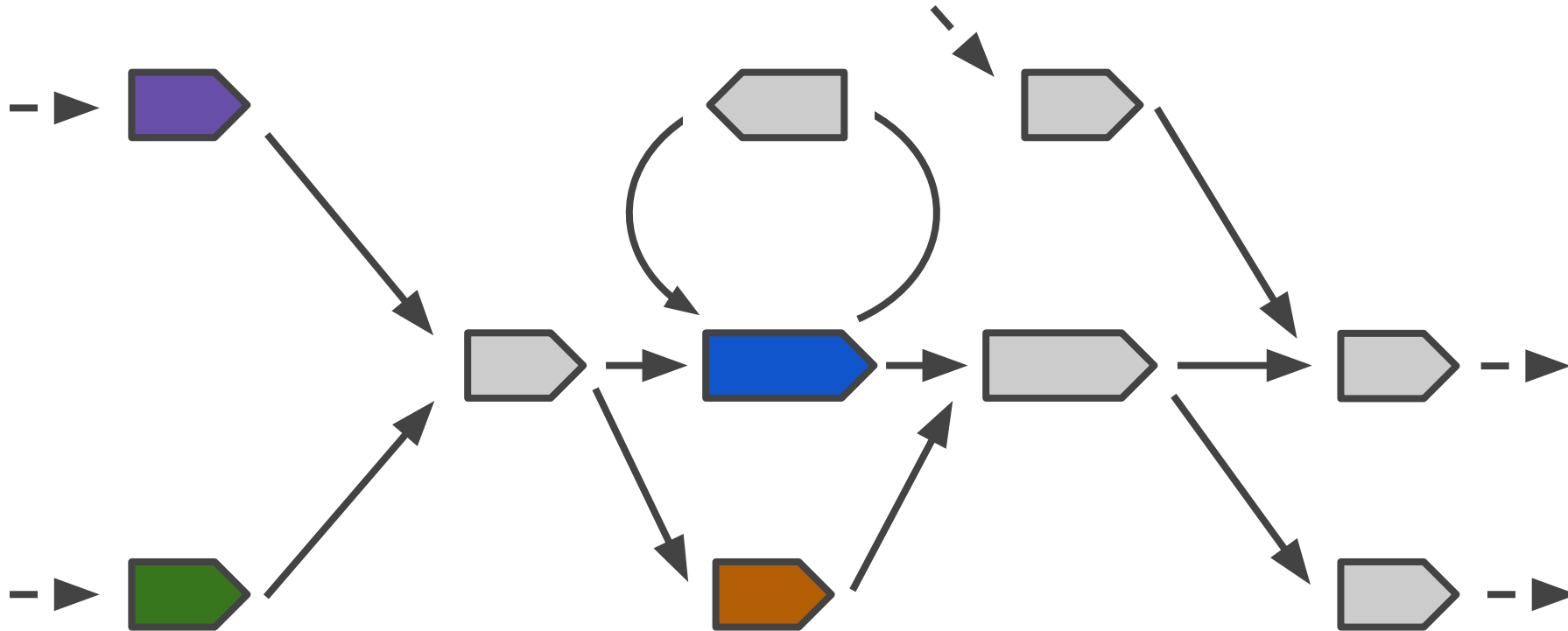




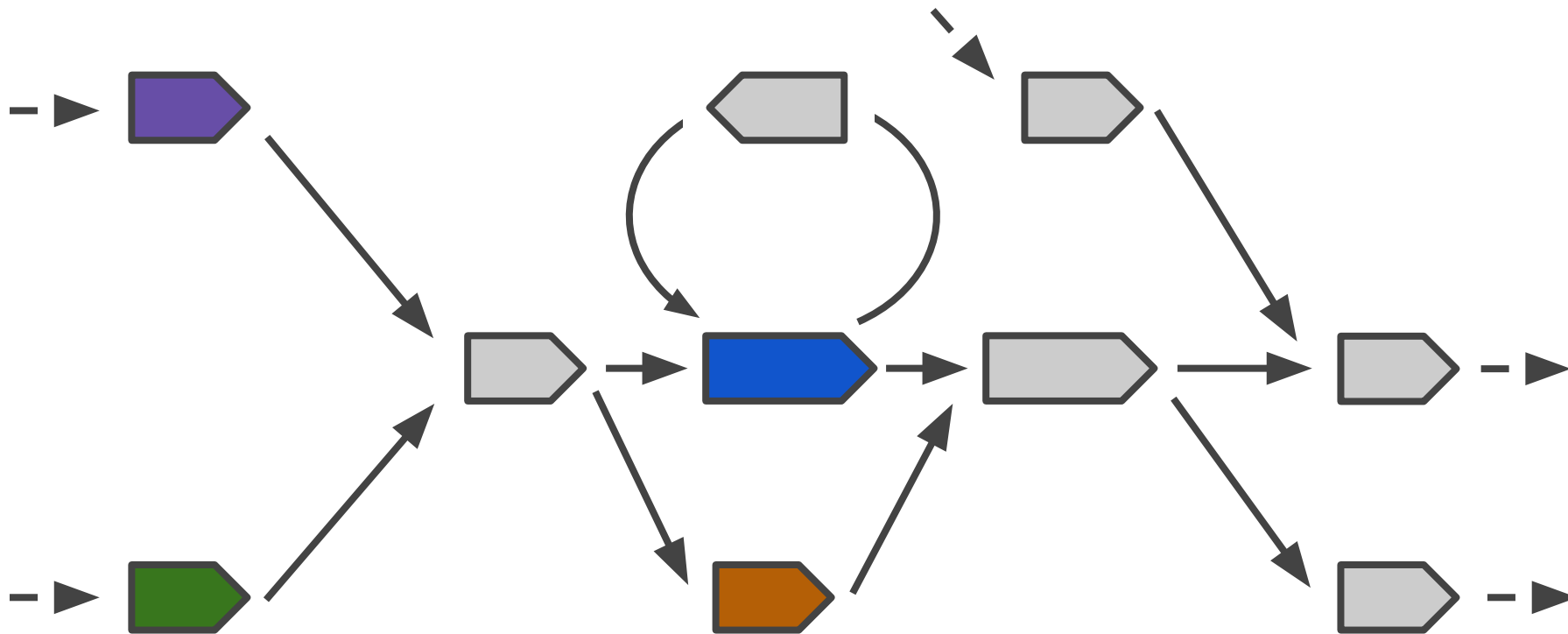
Can be represented as  
a graph of sequences  
linked by their overlaps



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)



The background of the slide is a dense, intricate network of white squares and arrows on a black background. The squares are arranged in a roughly circular pattern, with arrows pointing in various directions, creating a complex, interconnected web of relationships. The overall appearance is that of a highly complex system or network.

And real metagenomes  
are **very** complex



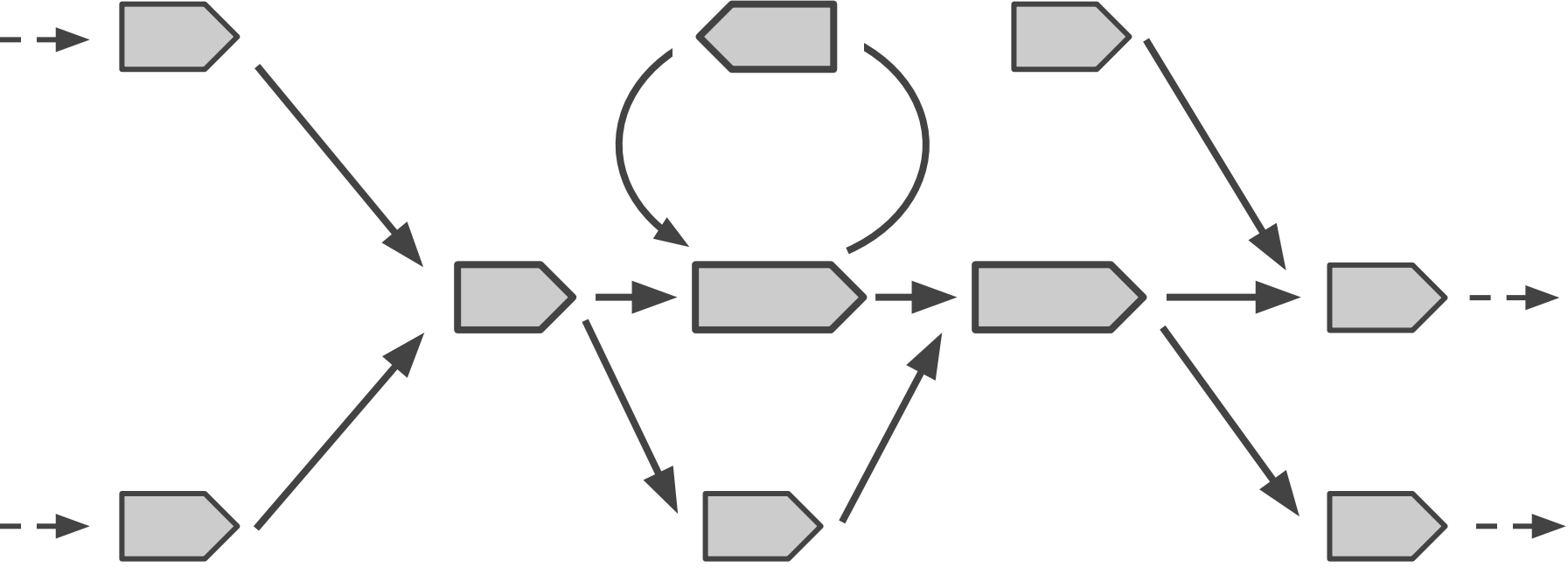
And real metagenomes  
are **very** complex





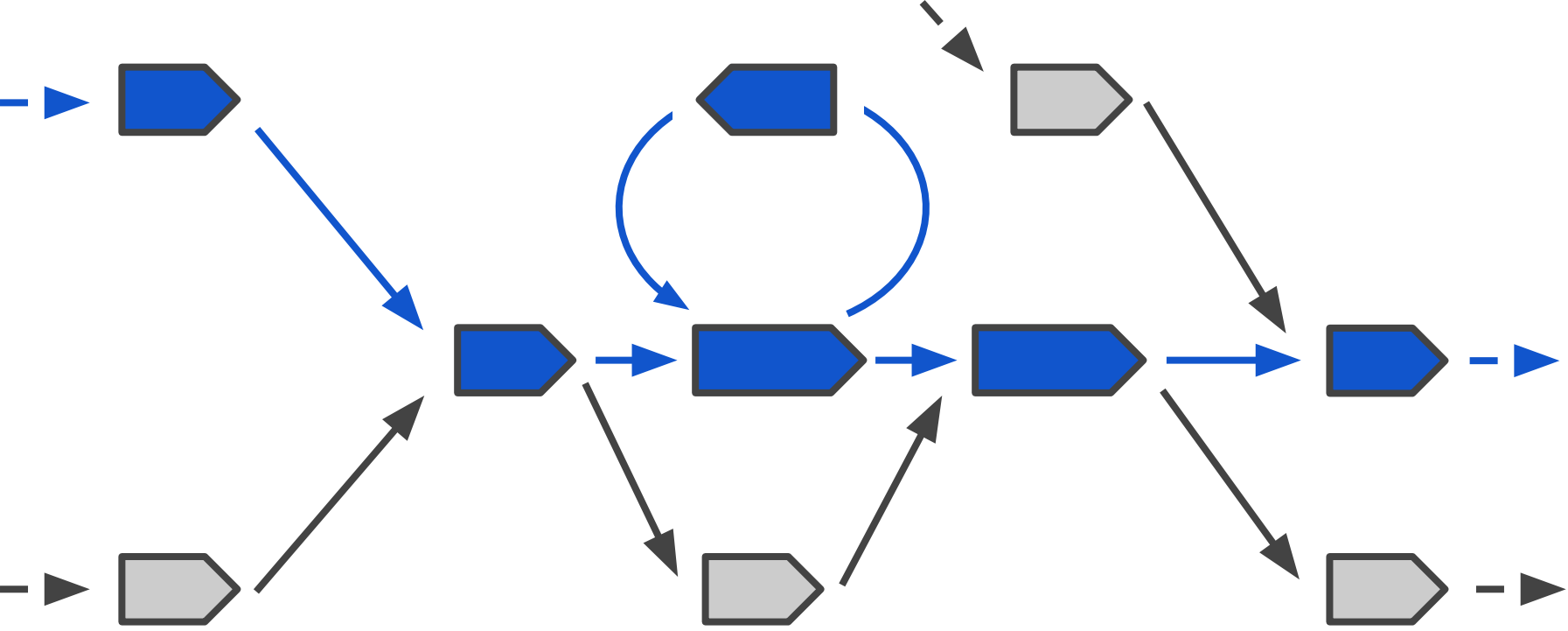
And real metagenomes  
are **very** complex

Real genomic sequences are paths on the graph

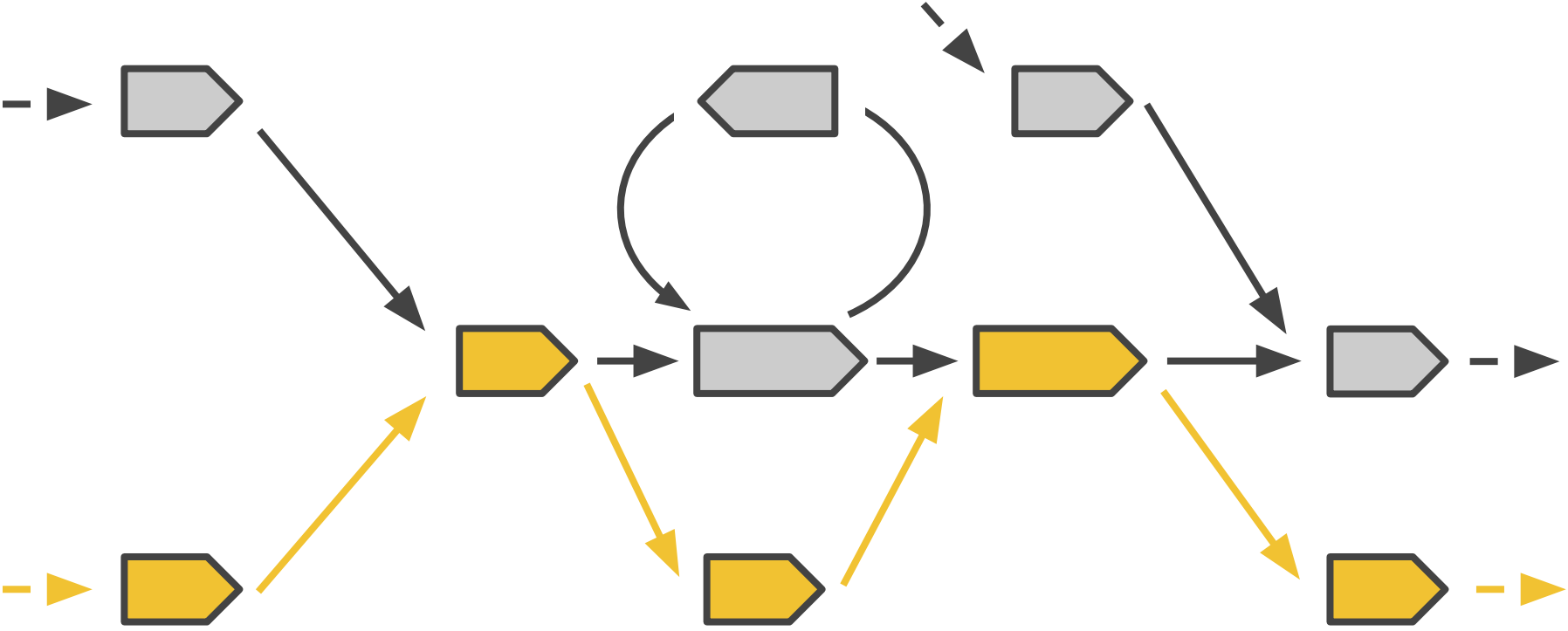




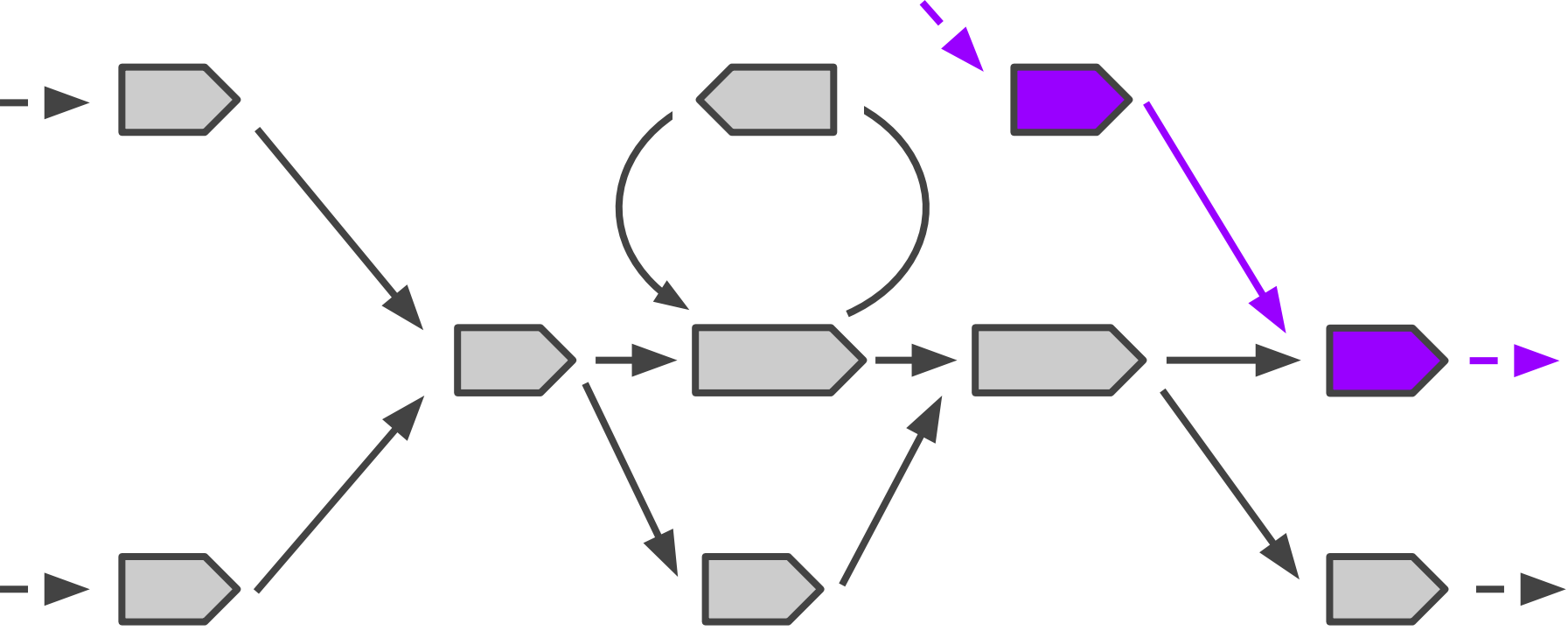
Real genomic sequences are paths on the graph



Real genomic sequences are paths on the graph

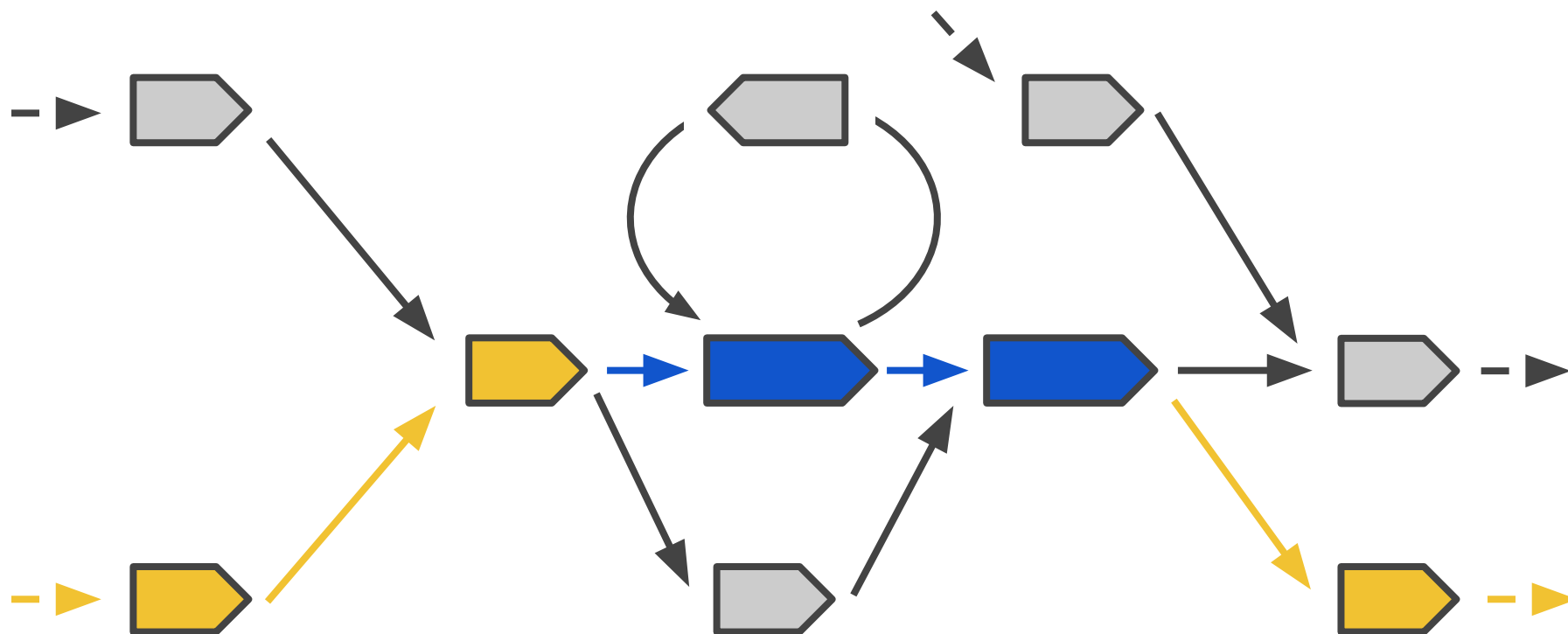


Real genomic sequences are paths on the graph

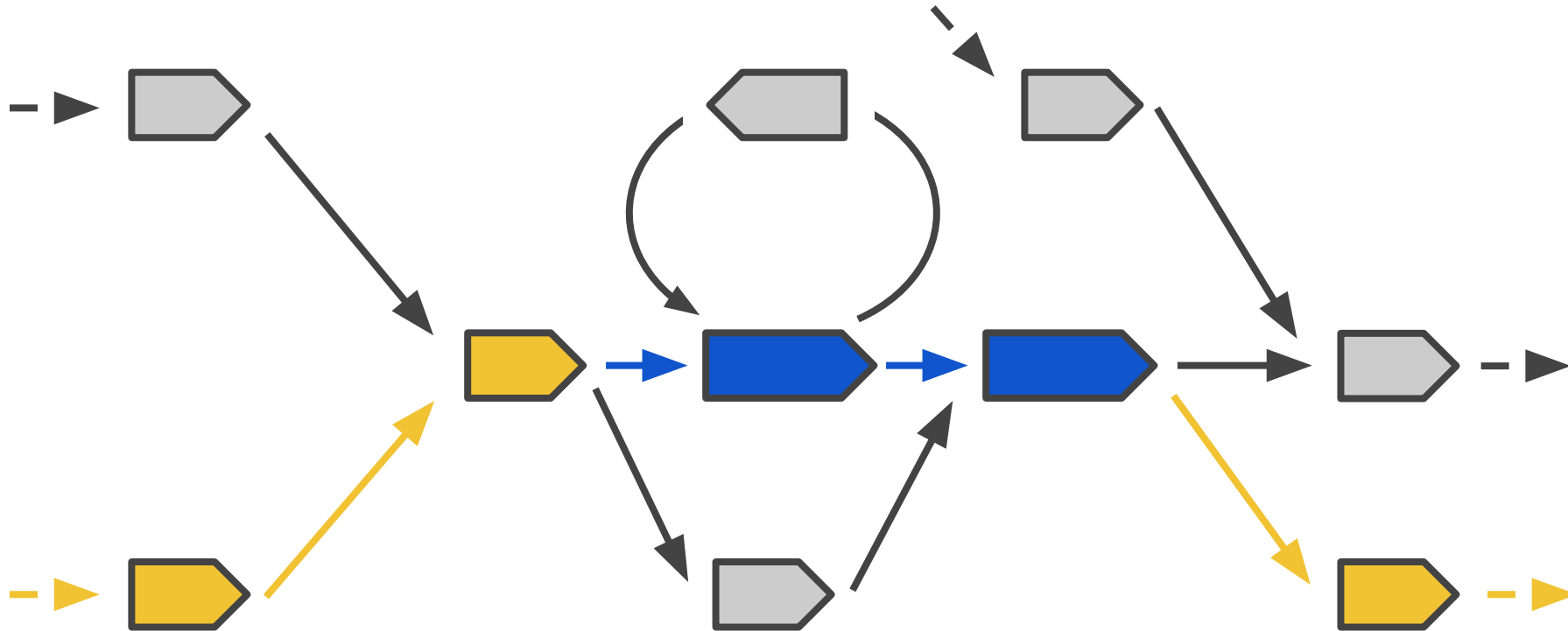




Lots of incorrect paths also exist...



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



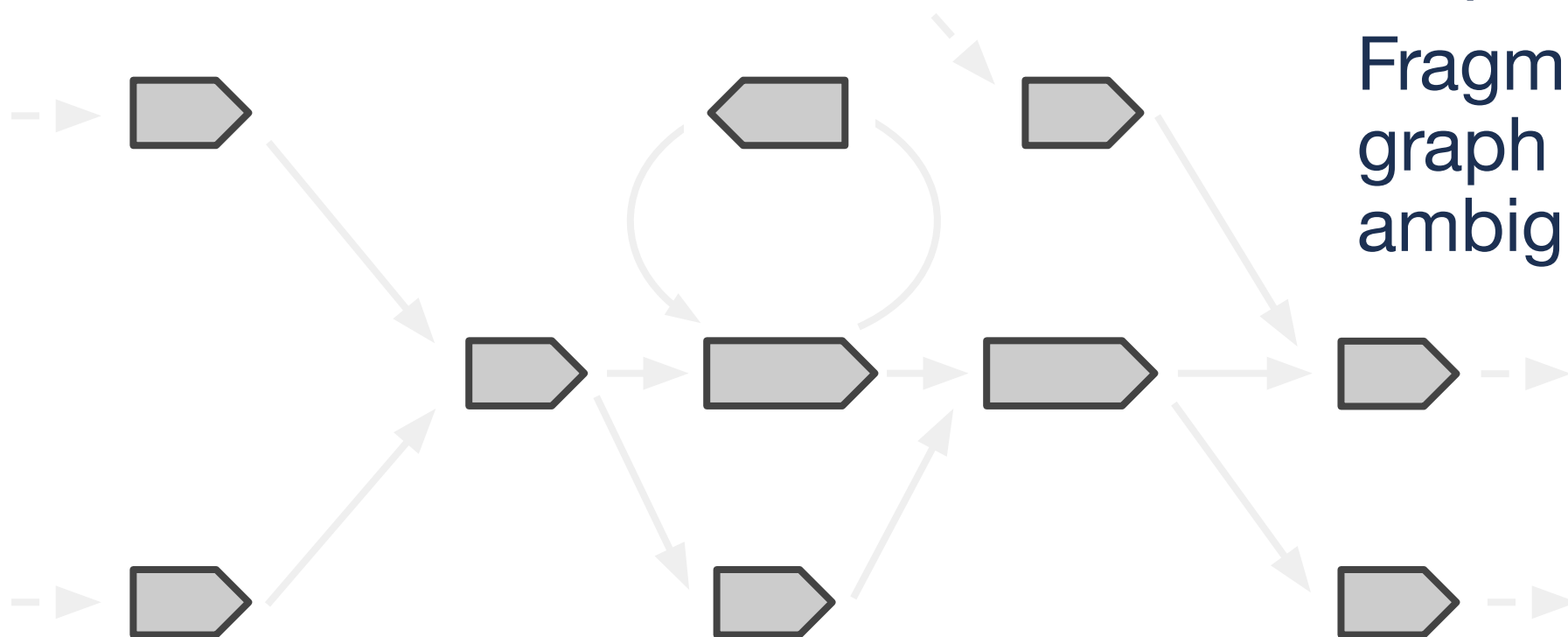


Lots of incorrect paths also exist...  
*How do we avoid these?*

**Standard Tools:**

Filter out  
low-abundance  
sequences

Fragment the  
graph when it's  
ambiguous



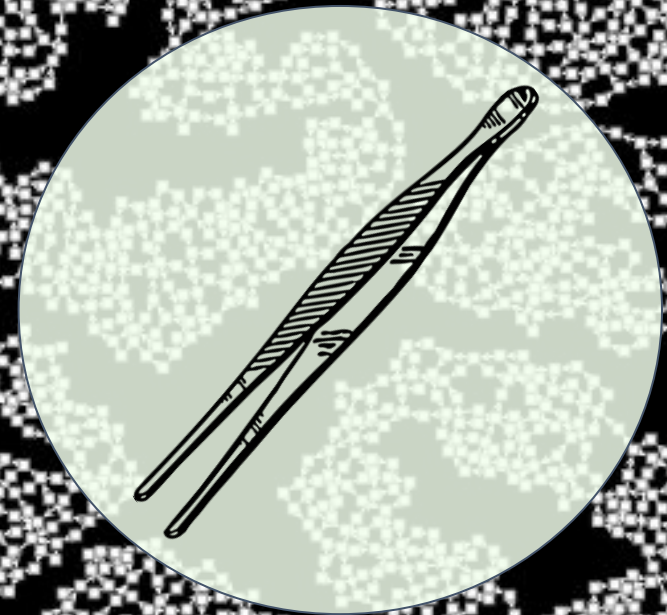


# Untangling the hairball





# Untangling the hairball



Gladstone Scientific Retreat 2024

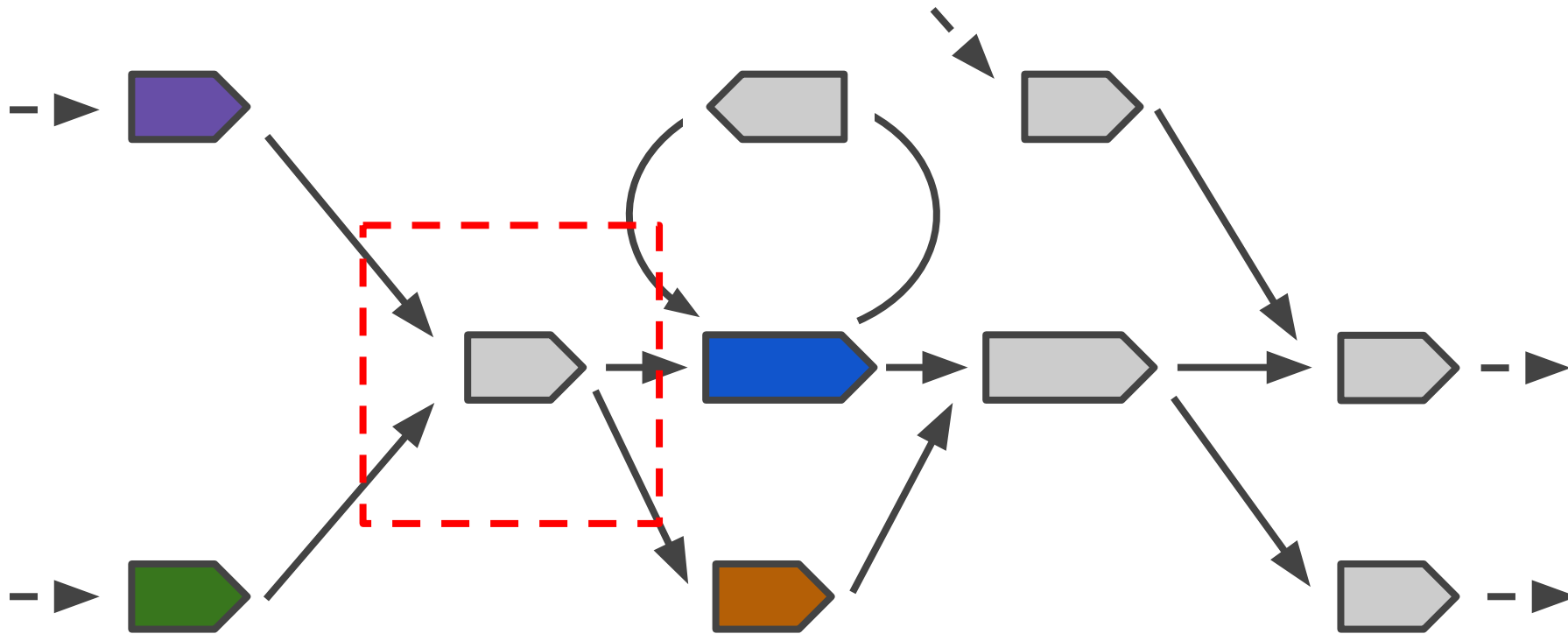
**StrainZip:**

Untangling the metagenome graph

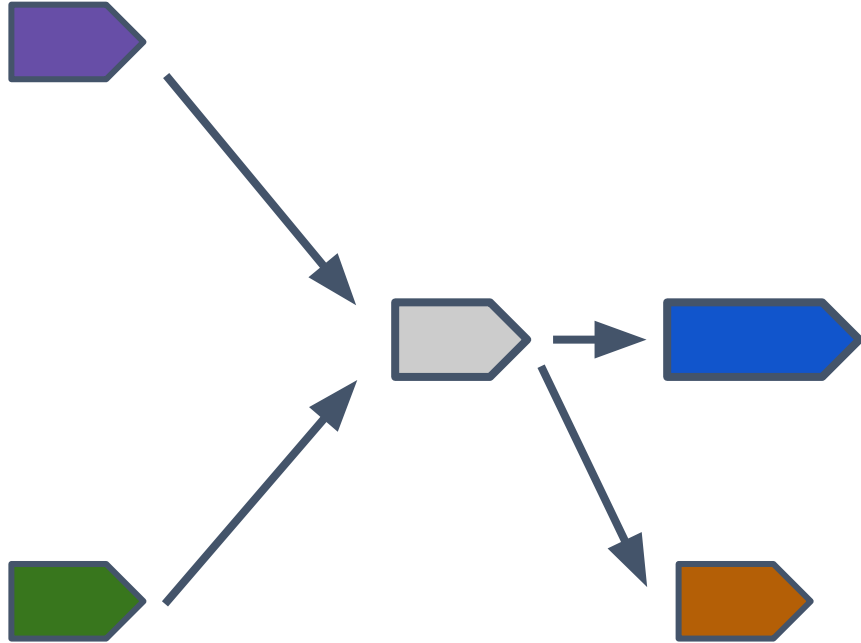




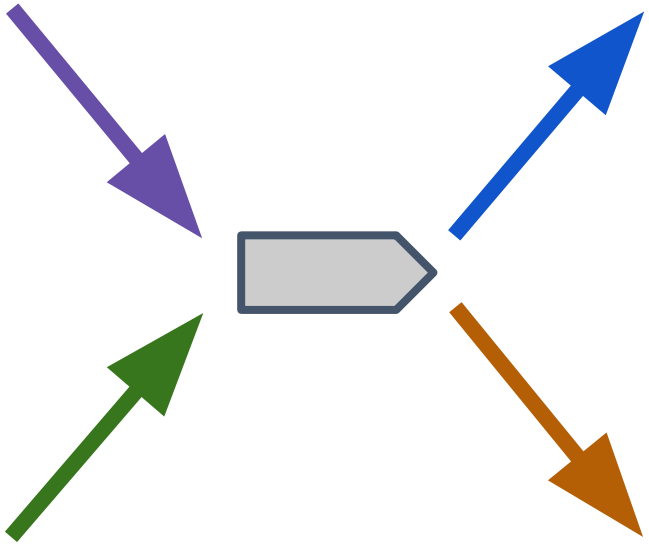
How can we recover long, accurate genome sequences from short reads?



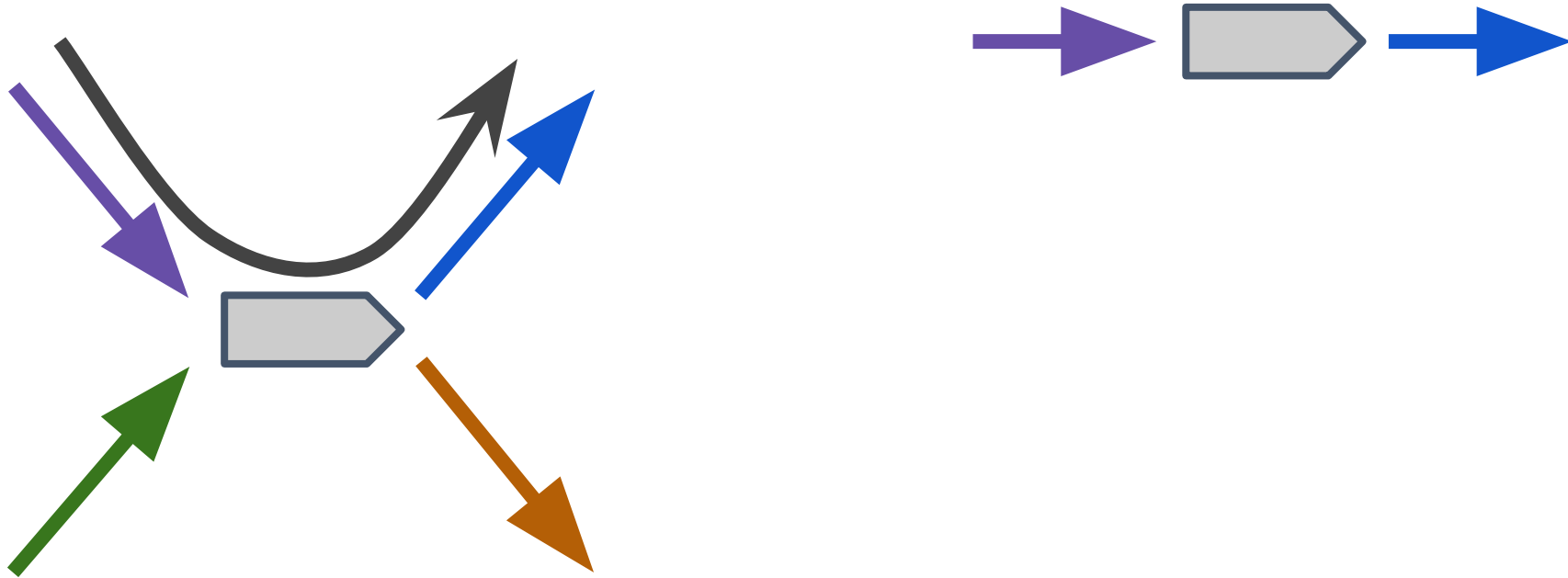
Focus on just one junction at a time



Focus on just one junction at a time

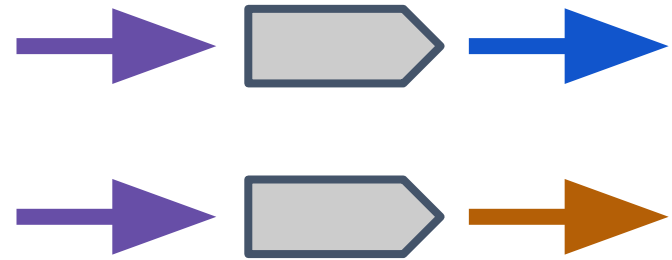
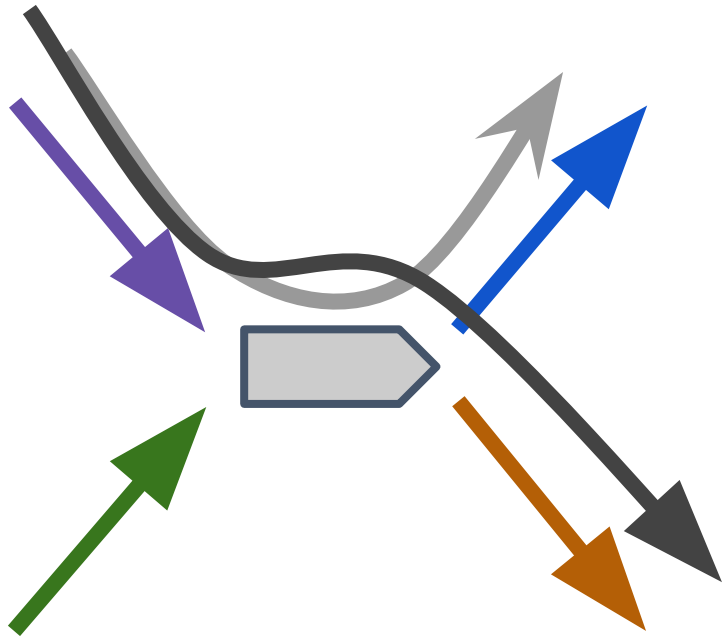


Focus on just one junction at a time

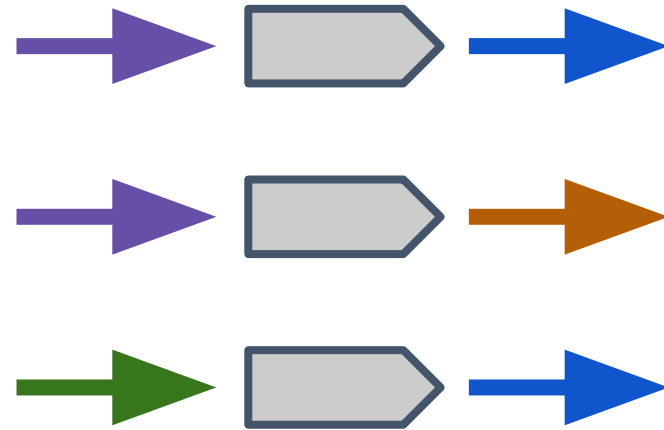
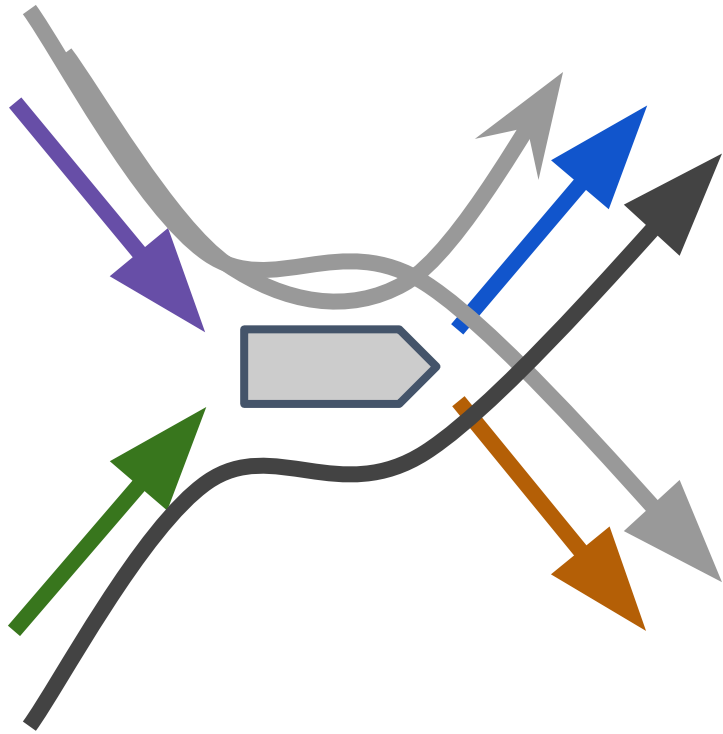




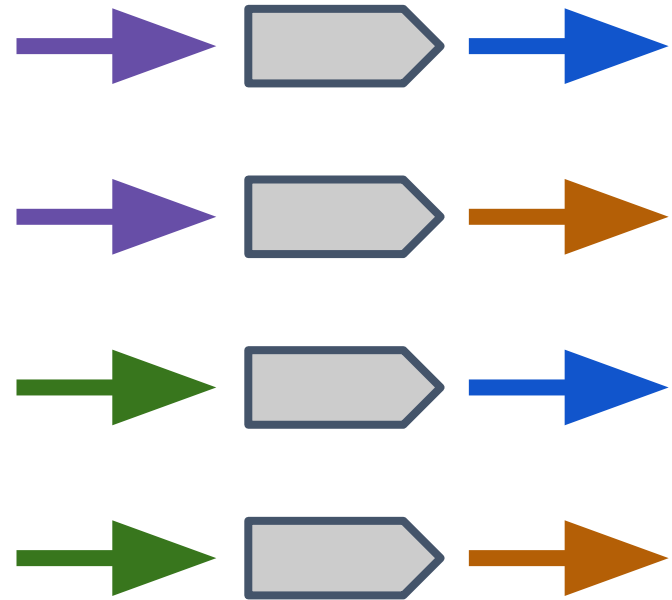
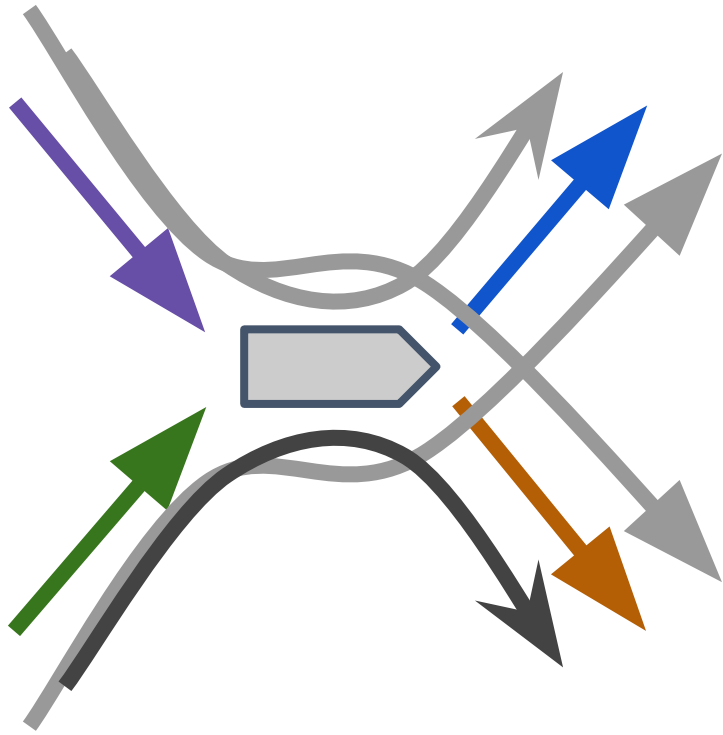
Focus on just one junction at a time



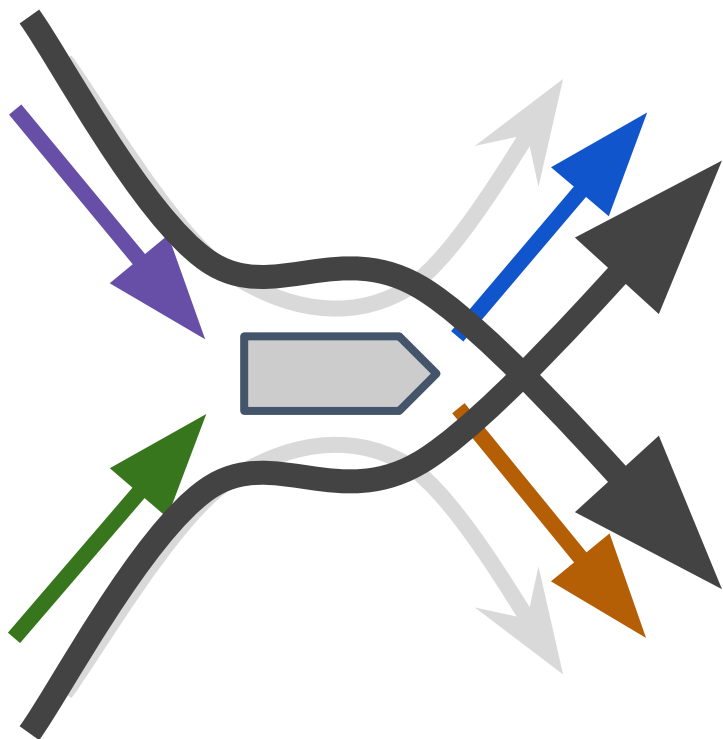
Focus on just one junction at a time



Focus on just one junction at a time



Focus on just one junction at a time  
Select local paths



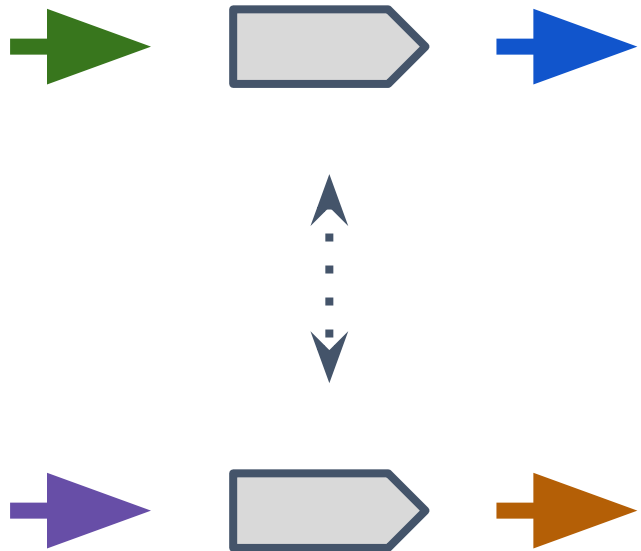
$$\begin{matrix} \begin{matrix} \downarrow & \downarrow & \downarrow & \downarrow \\ \color{blue} & \color{orange} & \color{green} & \color{green} \\ \color{blue} & \color{orange} & \color{green} & \color{green} \\ \color{purple} & & \color{purple} & \\ \color{orange} & & & \color{orange} \end{matrix}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

$X \quad \beta \quad Y$

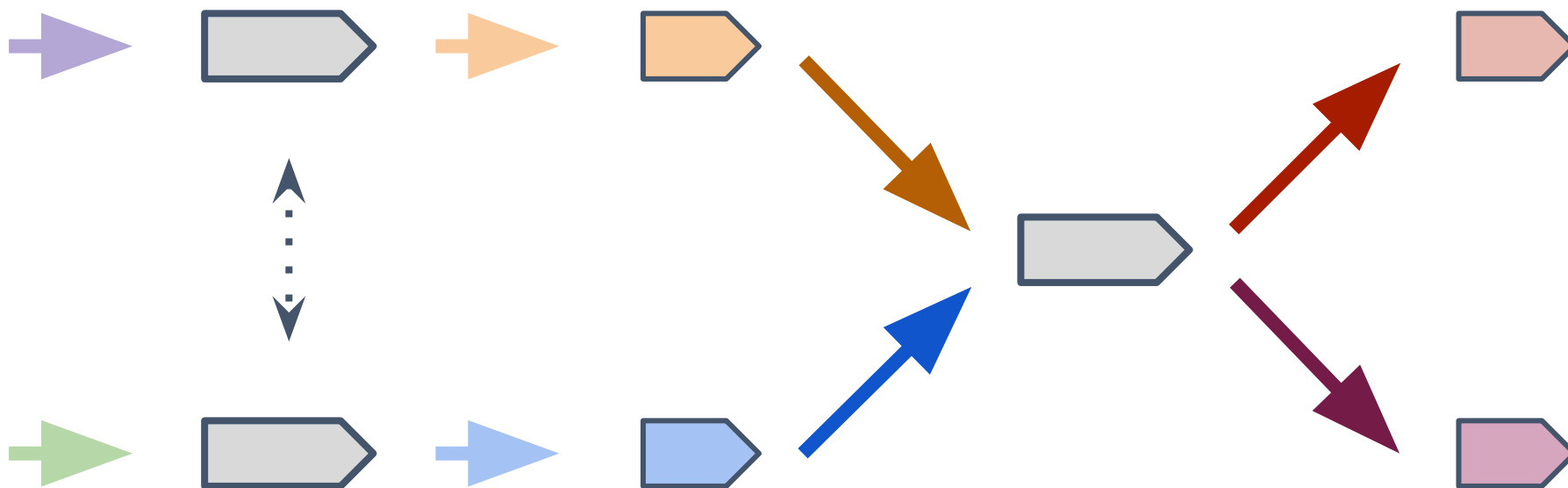
Sparse linear  
regression across  
multiple samples



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

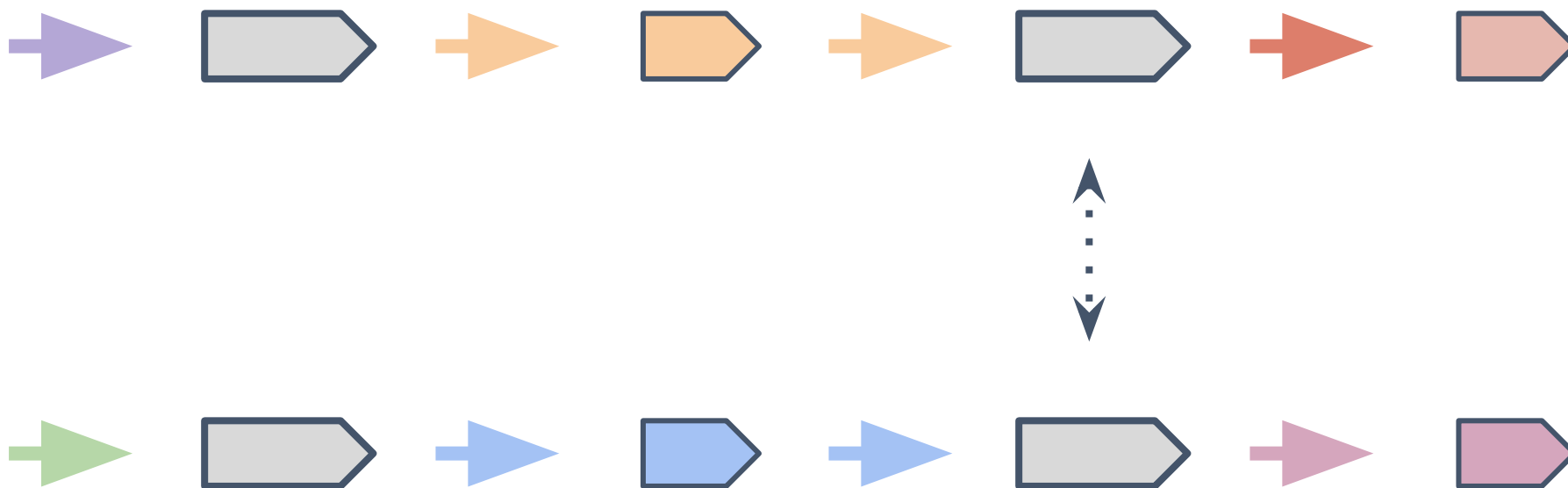


Focus on just one junction at a time

Select local paths

Unzip

Repeat



Focus on just one junction at a time

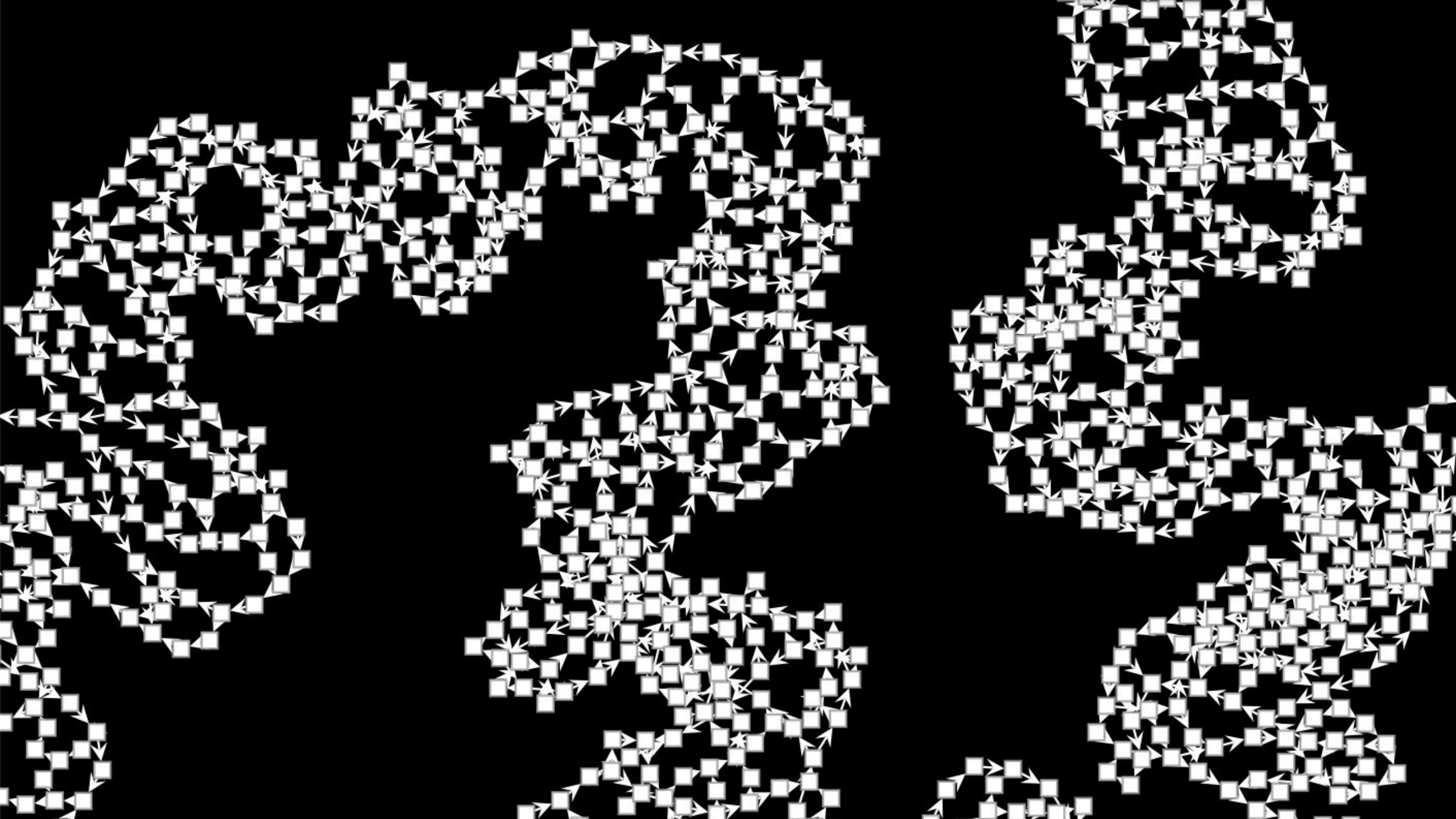
Select local paths

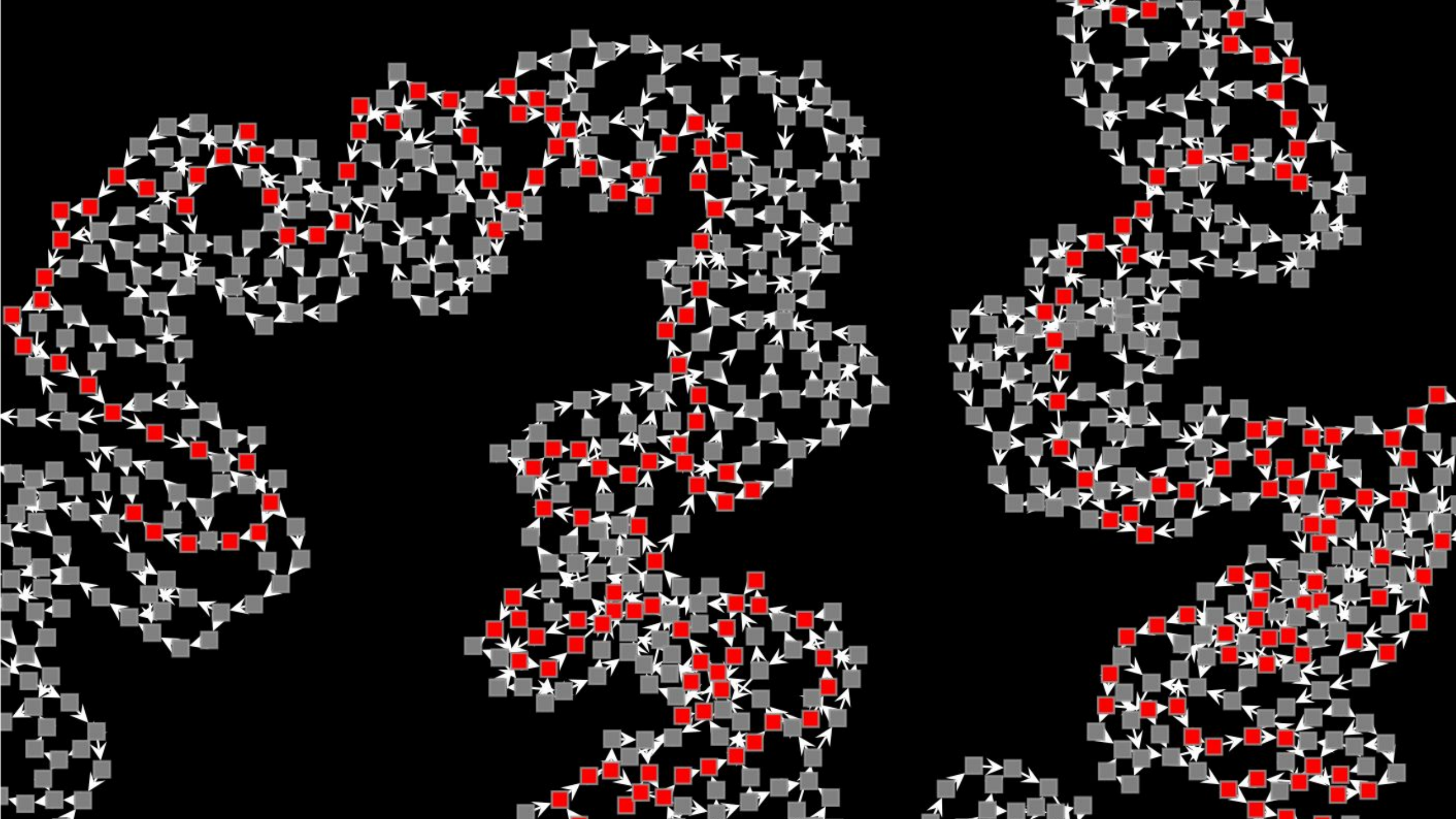
Unzip

Repeat





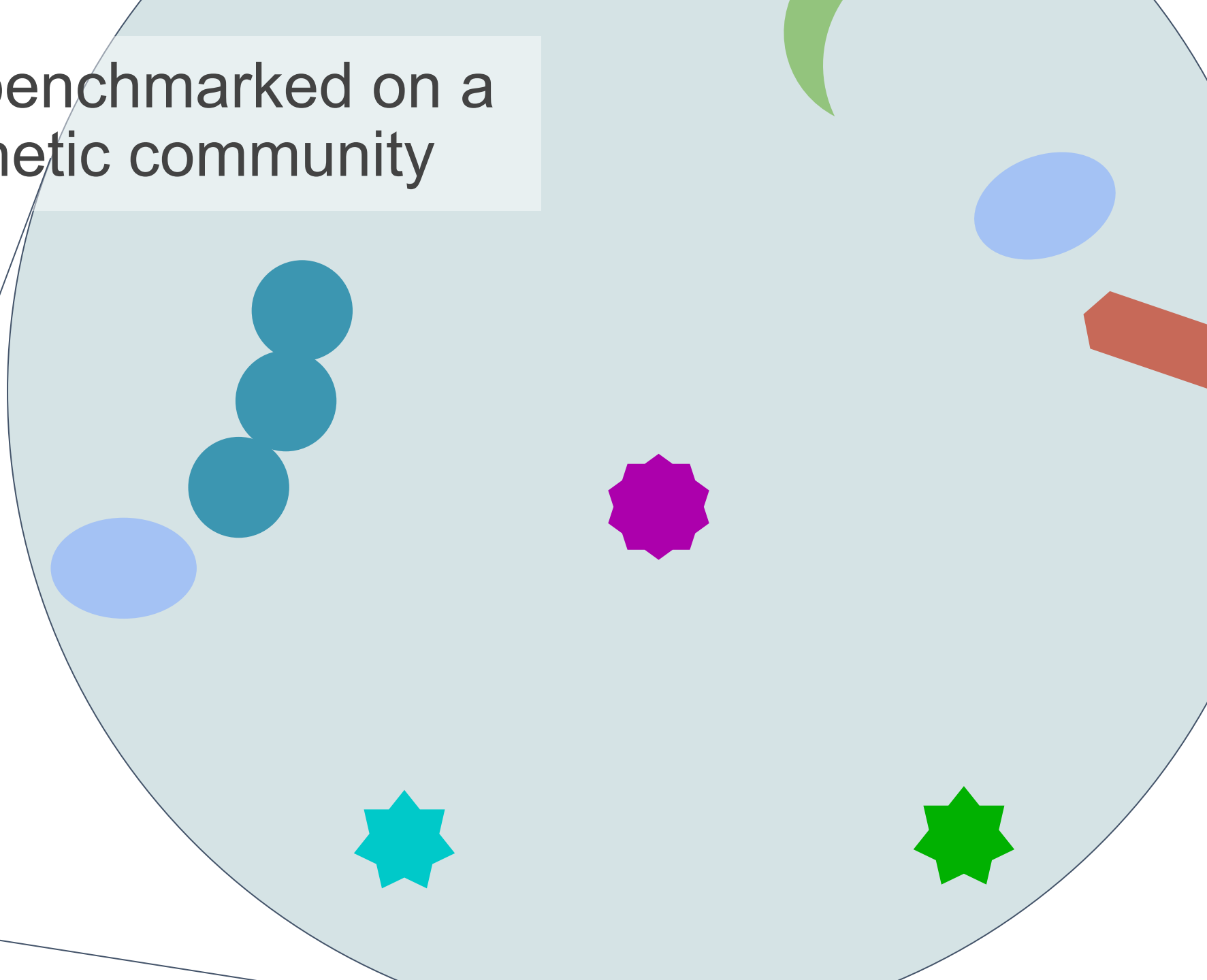
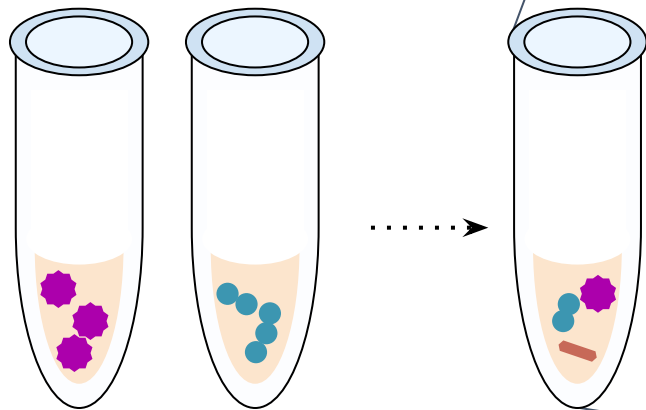




Gladstone Scientific Retreat 2024

# Strain-resolved discovery

Performance benchmarked on a complex, synthetic community







**Antibiotic resistance genes are widespread in the gut microbiome**



A scanning electron micrograph (SEM) of a gut microbiome. The image shows a dense, complex network of grey, fibrous material, likely extracellular polymeric substances (EPS) or mucus. Interspersed throughout this network are numerous individual bacterial cells of various shapes and sizes, each rendered in a different color (e.g., green, yellow, orange, pink, purple, blue). The bacteria appear to be embedded within or attached to the fibrous structure.

## Antibiotic resistance genes are widespread in the gut microbiome

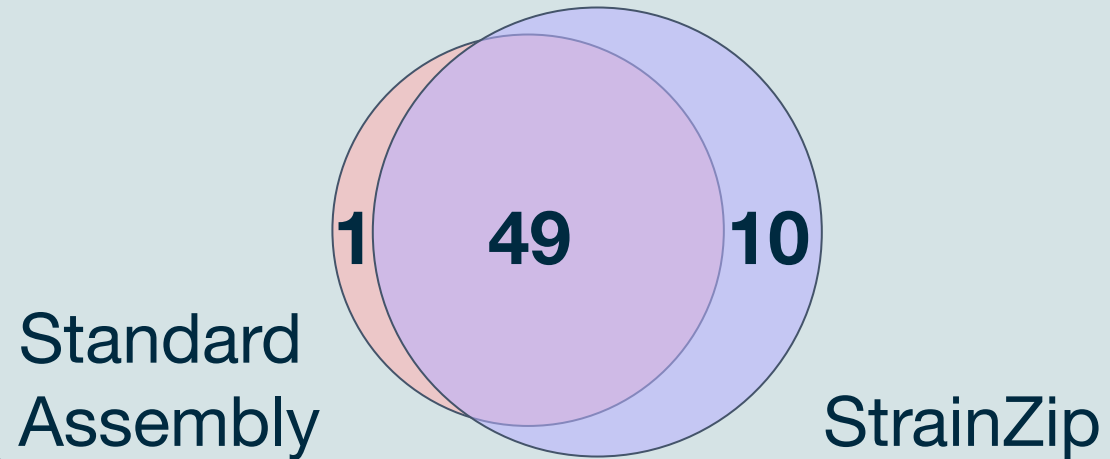
- Detection can inform treatment



# 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

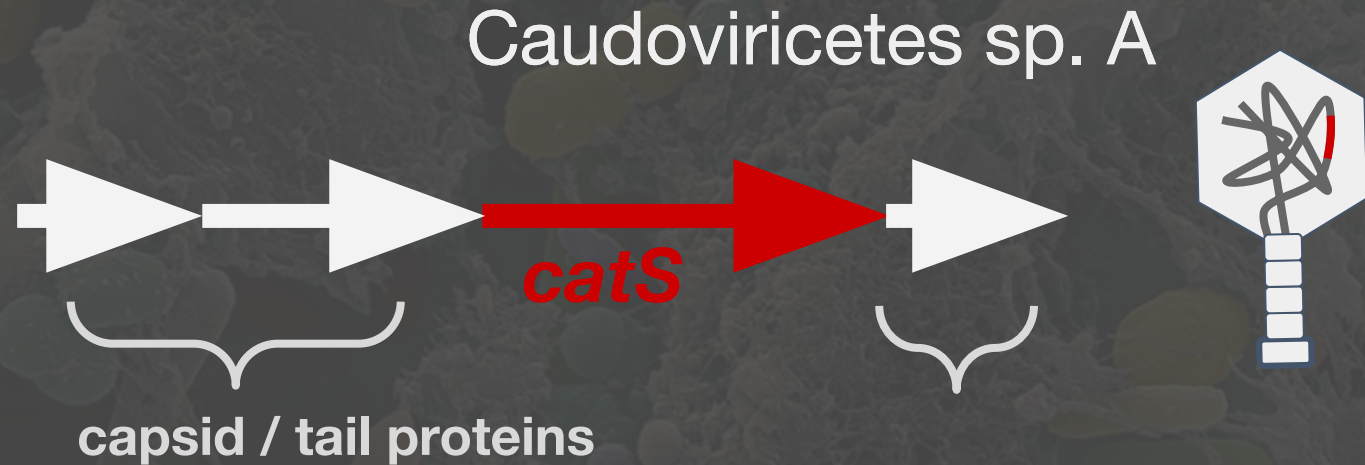
- Detection can inform treatment

*catS* 



# Antibiotic resistance genes are widespread in the gut microbiome

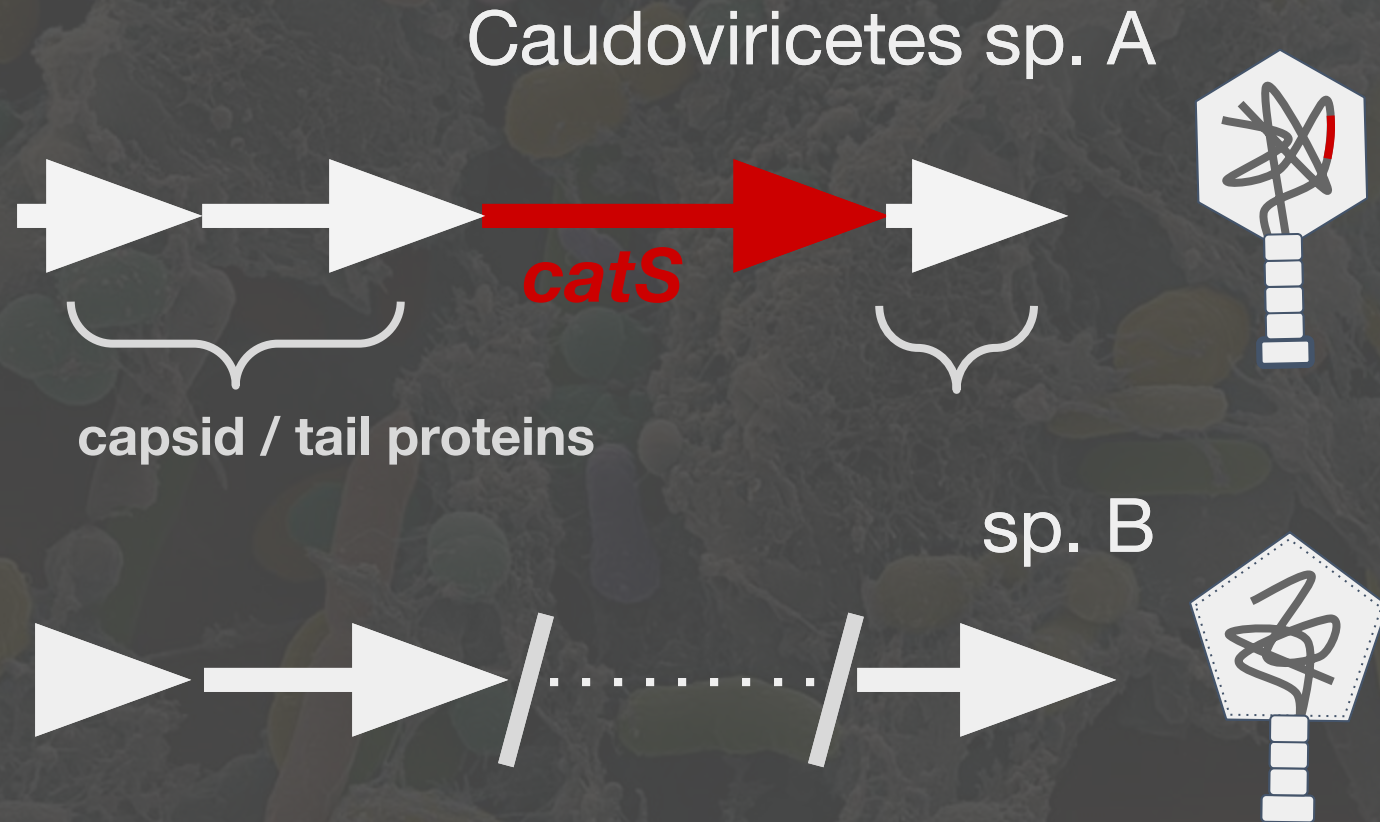
- Detection can inform treatment
- Can be carried in phage genomes





# Antibiotic resistance genes are widespread in the gut microbiome

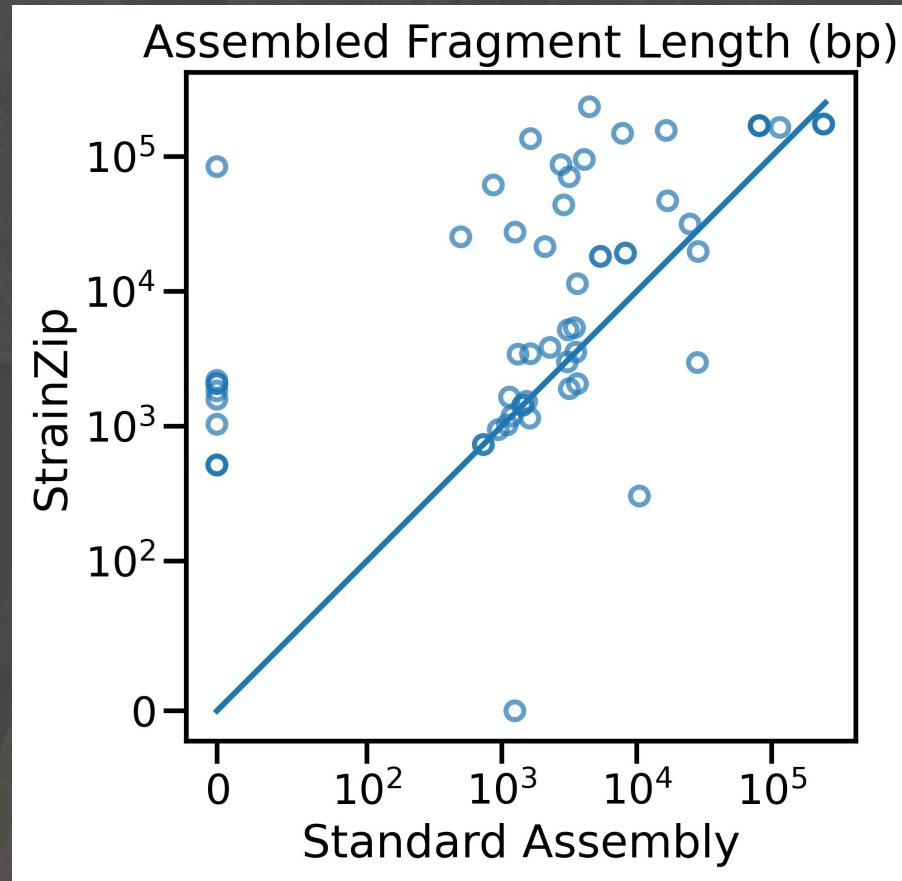
- Detection can inform treatment
- Can be carried in phage genomes
- Long sequence fragments provide useful information



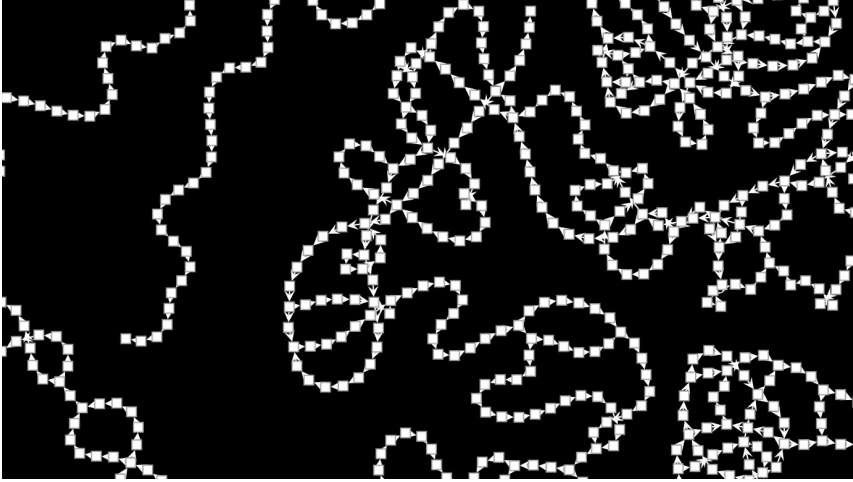


# Antibiotic resistance genes are widespread in the gut microbiome

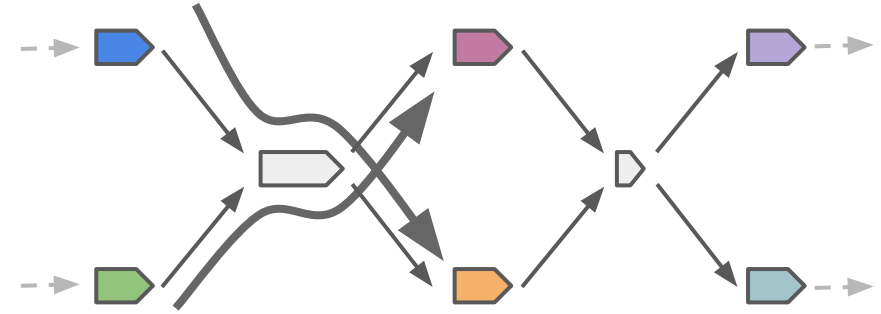
- Detection can inform treatment
- Can be carried in phage genomes
- Long sequence fragments provide useful information



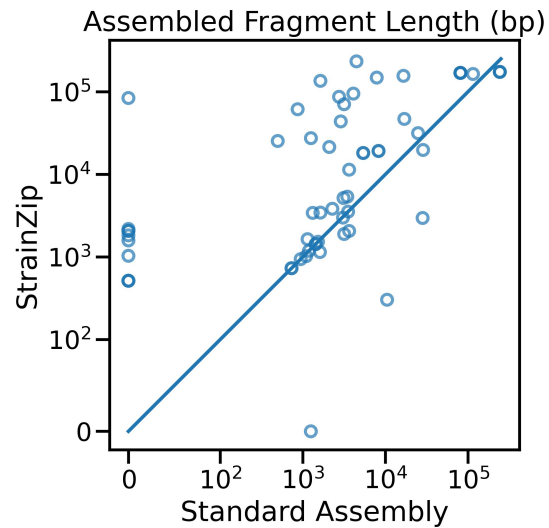
## Complex Metagenome Graphs



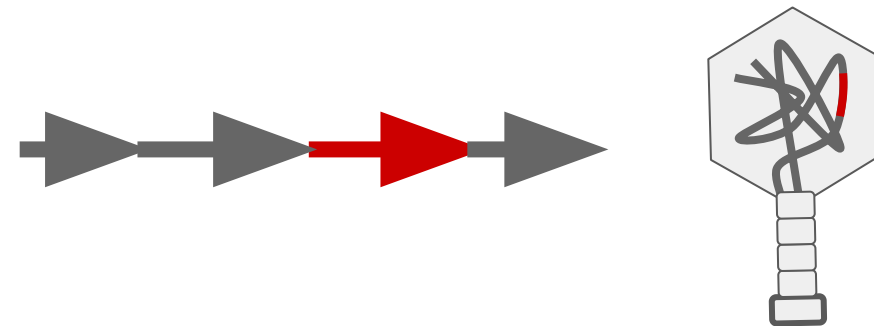
## StrainZip Iteratively Unzips Junctions



## Strain-Resolved Metagenomics



## Antibiotic Resistance Potential of Phage



Gladstone Scientific Retreat 2024

**Thank You!**  
**Questions?**