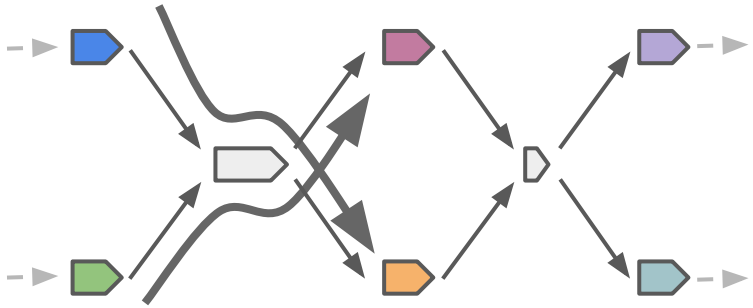
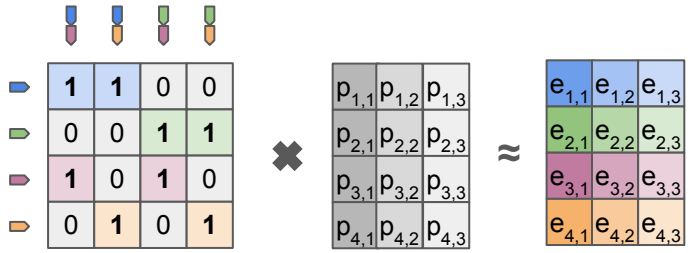
Unzipping the metagenome:

strain-level discovery in the gut microbiome



Byron J. Smith

Bhatt Lab Computational Subgroup
2024-09-10



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

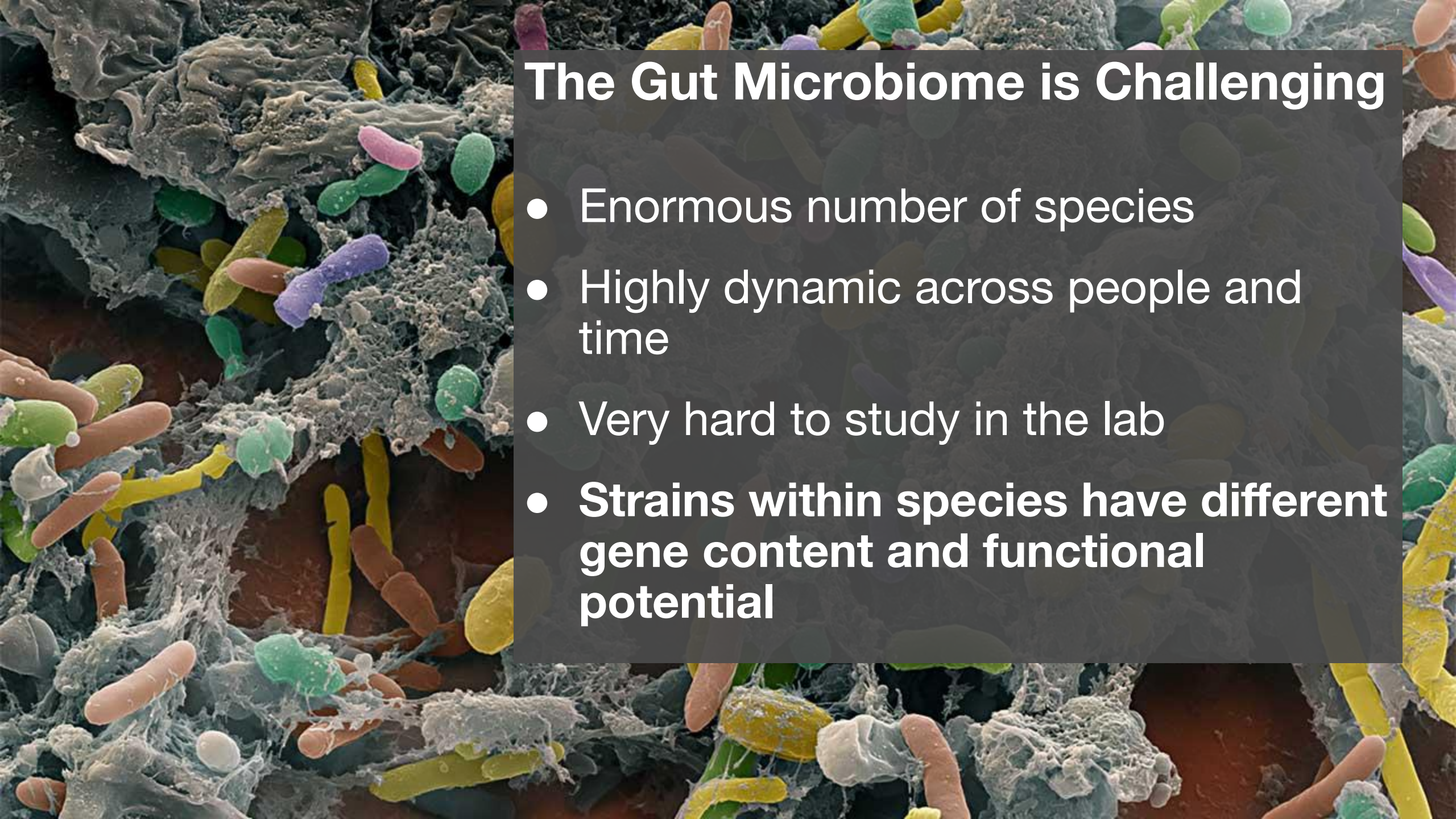
CHAN ZUCKERBERG BIOHUB

NIH National Institutes of Health

GLADSTONE INSTITUTES

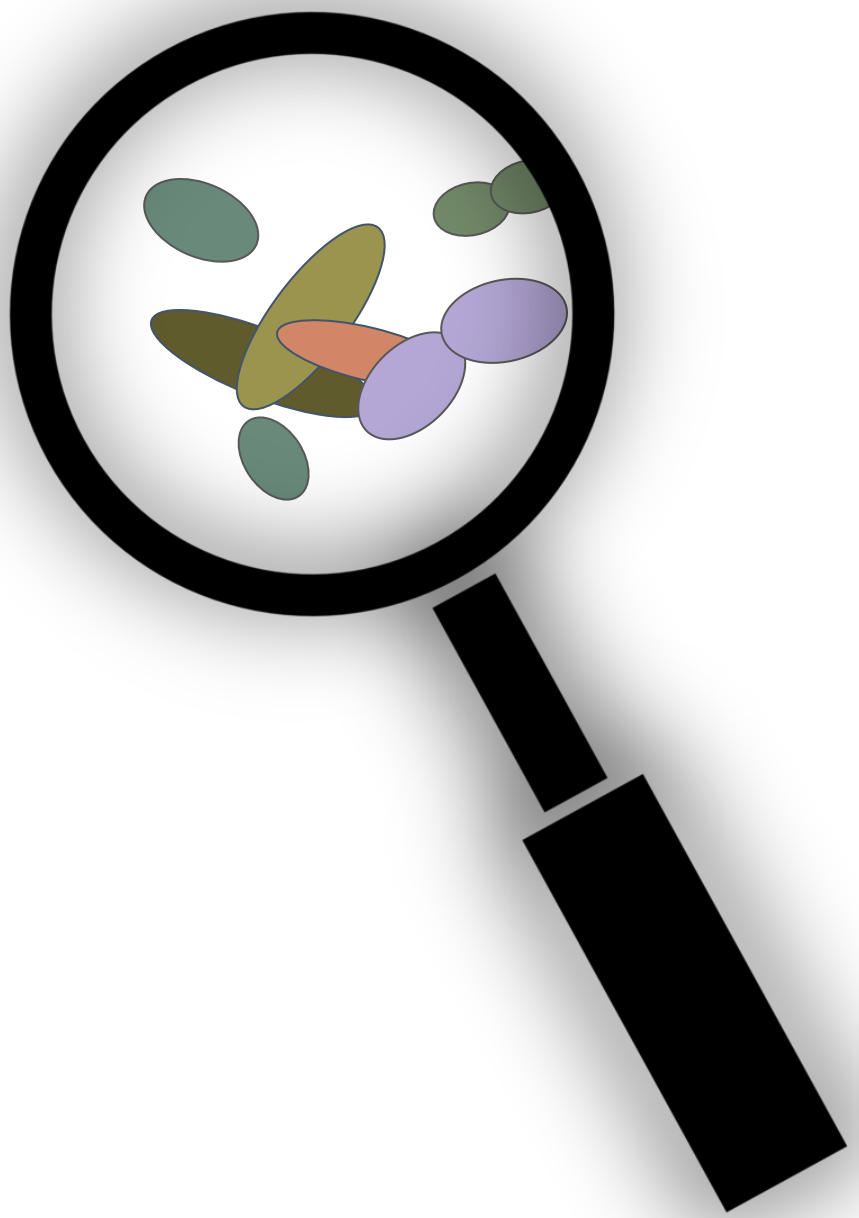
Introduction:

The gut microbiome and shotgun metagenomics

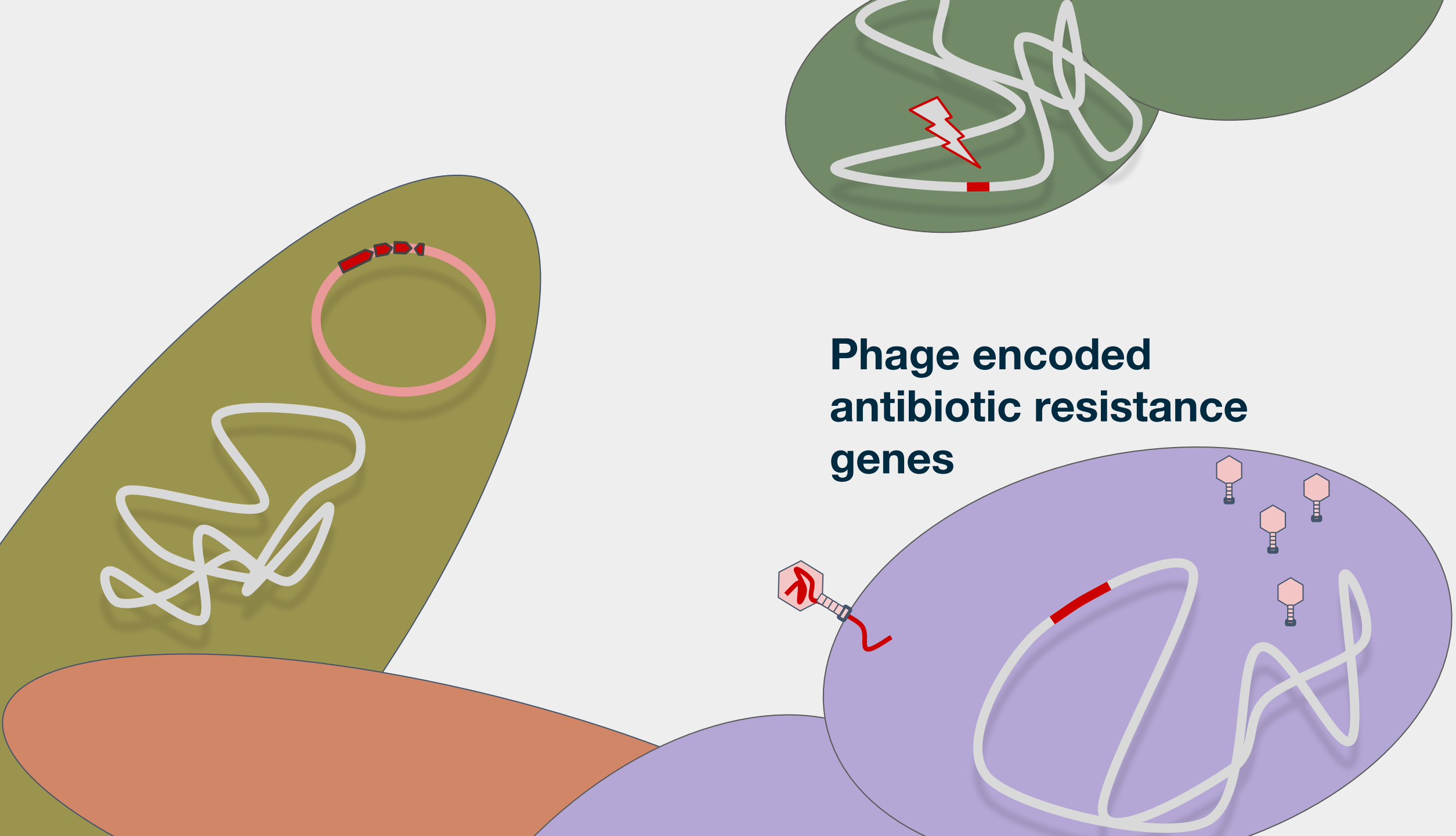


The Gut Microbiome is Challenging

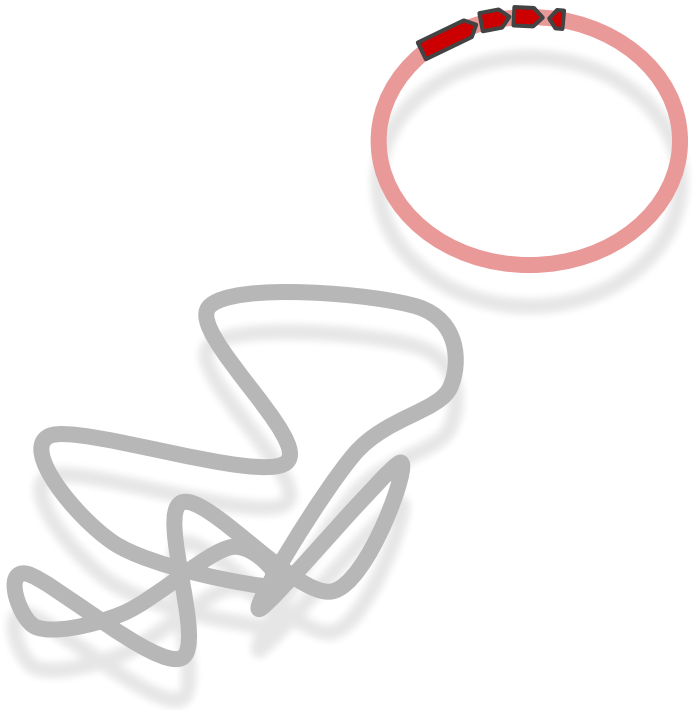
- Enormous number of species
- Highly dynamic across people and time
- Very hard to study in the lab
- **Strains within species have different gene content and functional potential**



Bacterial genomes are key to understanding strain diversity


Phage encoded
antibiotic resistance
genes

Metagenomic sequencing surveys all genomes

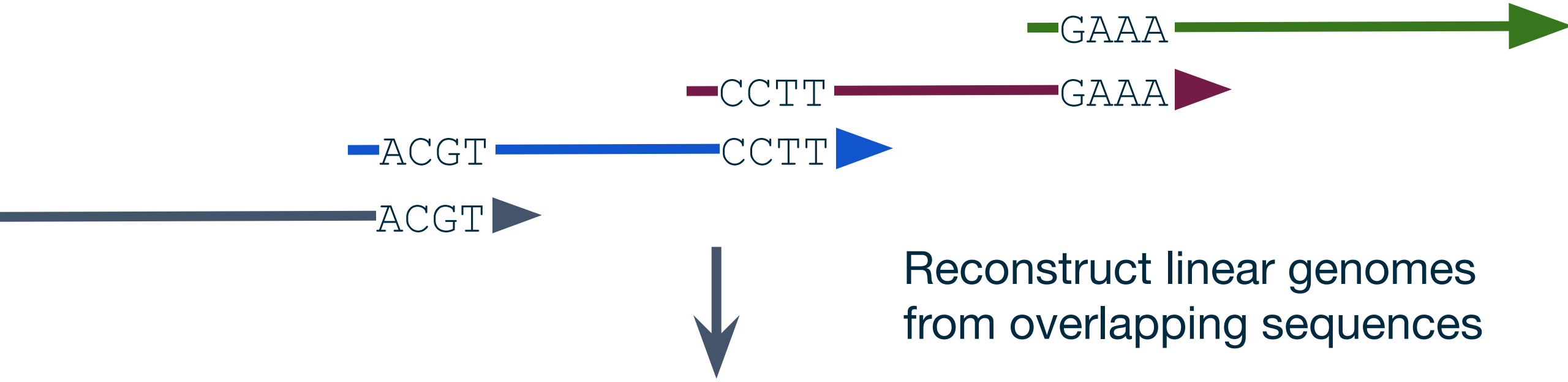


Short-read, shotgun metagenomes enable modern microbiome science

Requirements:

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

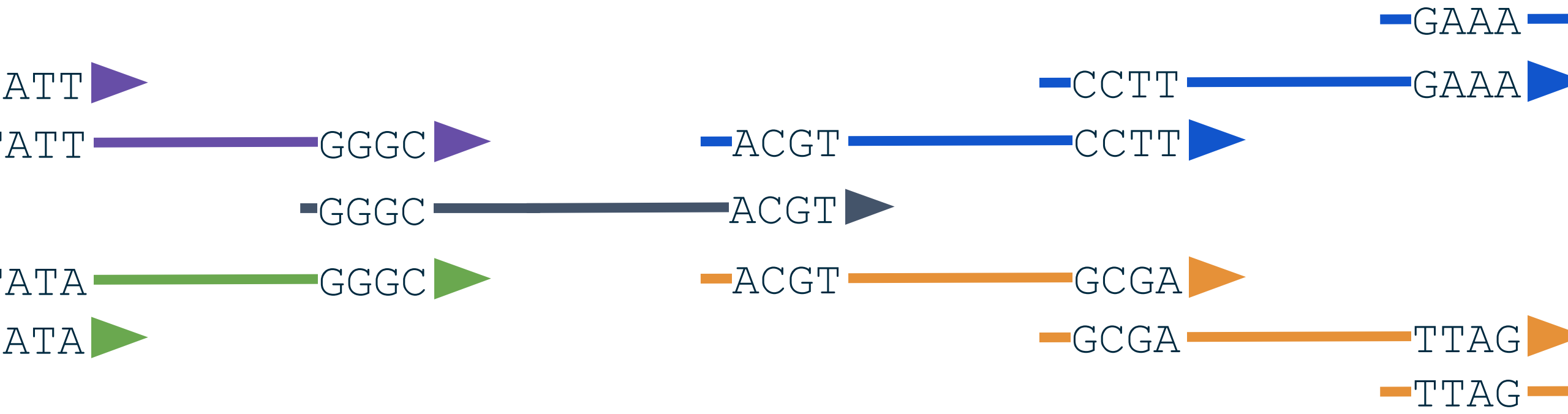
Turning short reads into long sequences



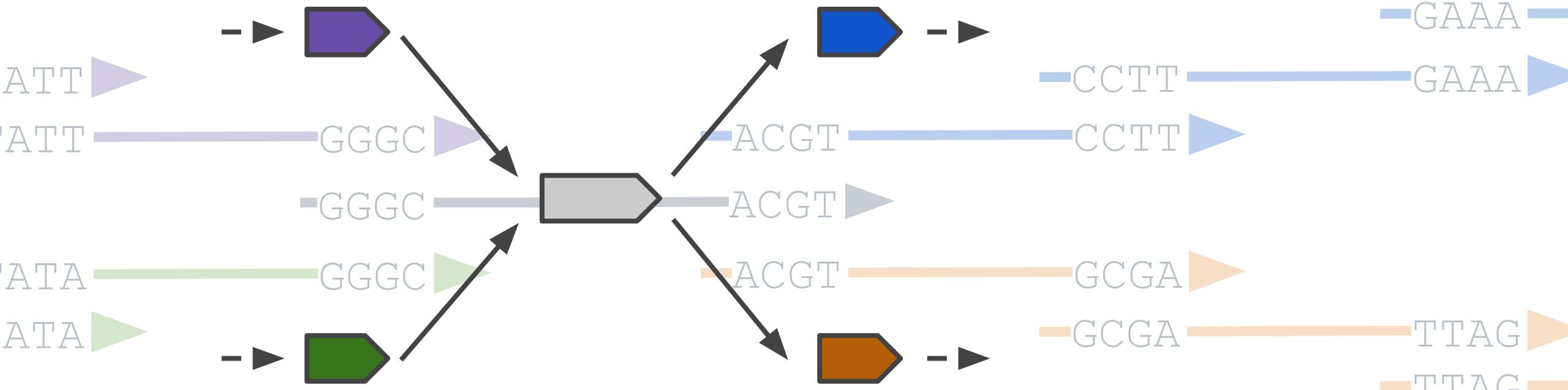
Reconstruct linear genomes from overlapping sequences

...GGTAGAGCGTGGGACGTAGGGTTAACCTTAGAAAGCTAGAAAACCGCGCGCCCT...

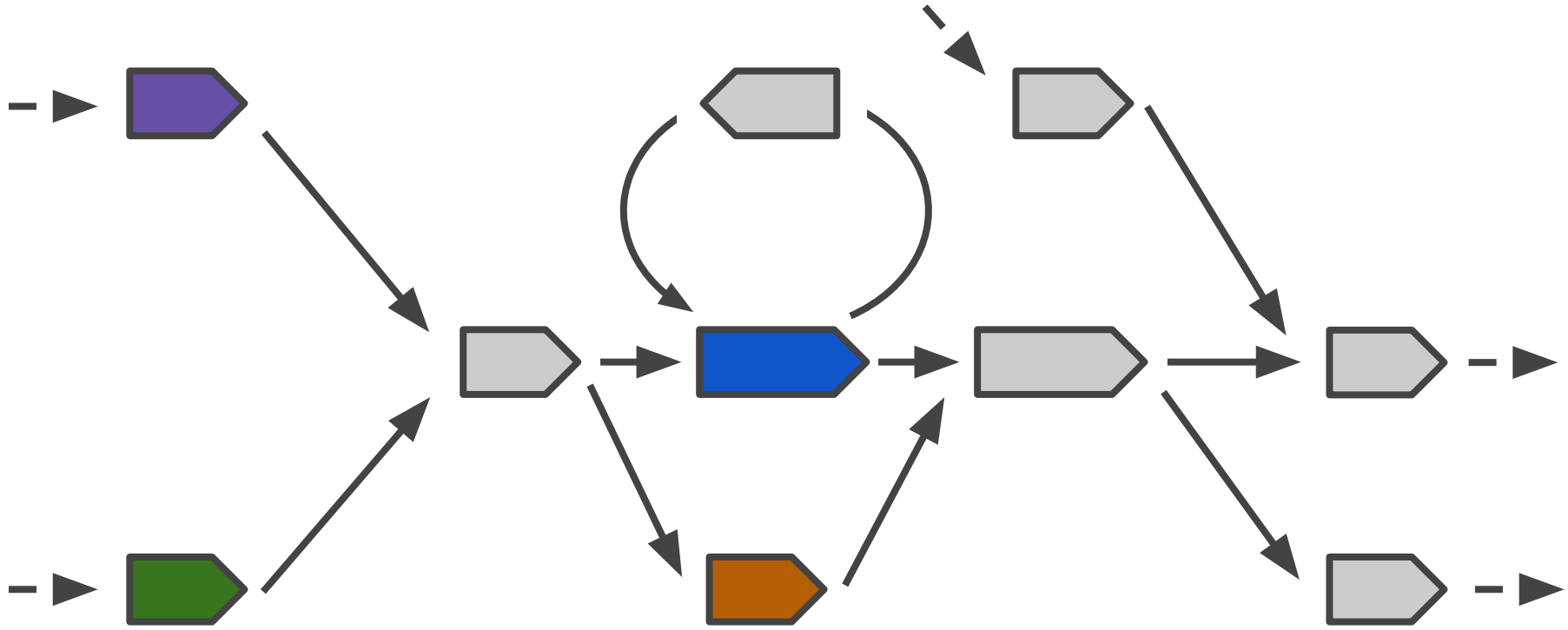
Problem: Closely related strains make read-chaining ambiguous



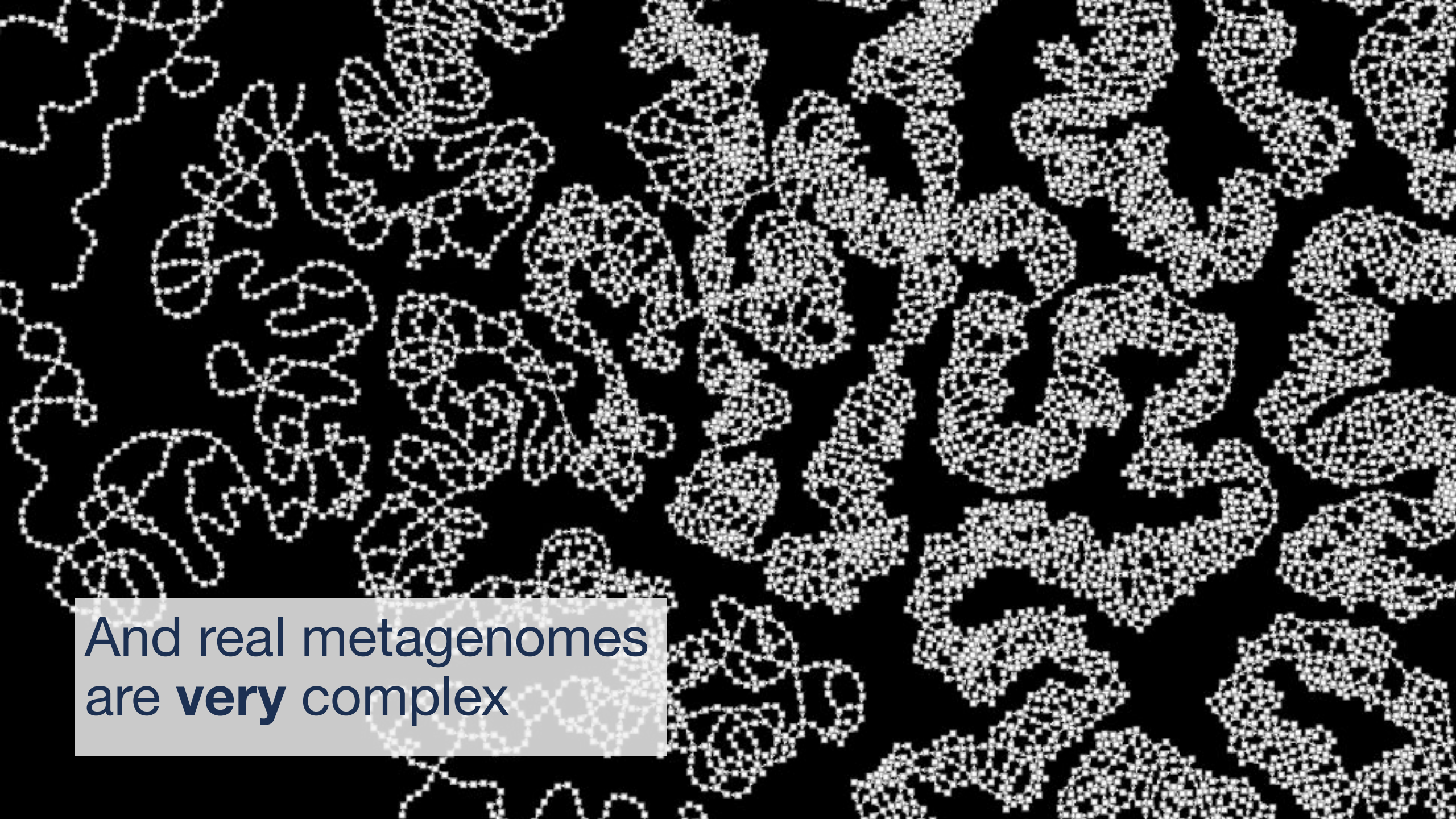
Can be represented as a graph of sequences linked by their overlaps



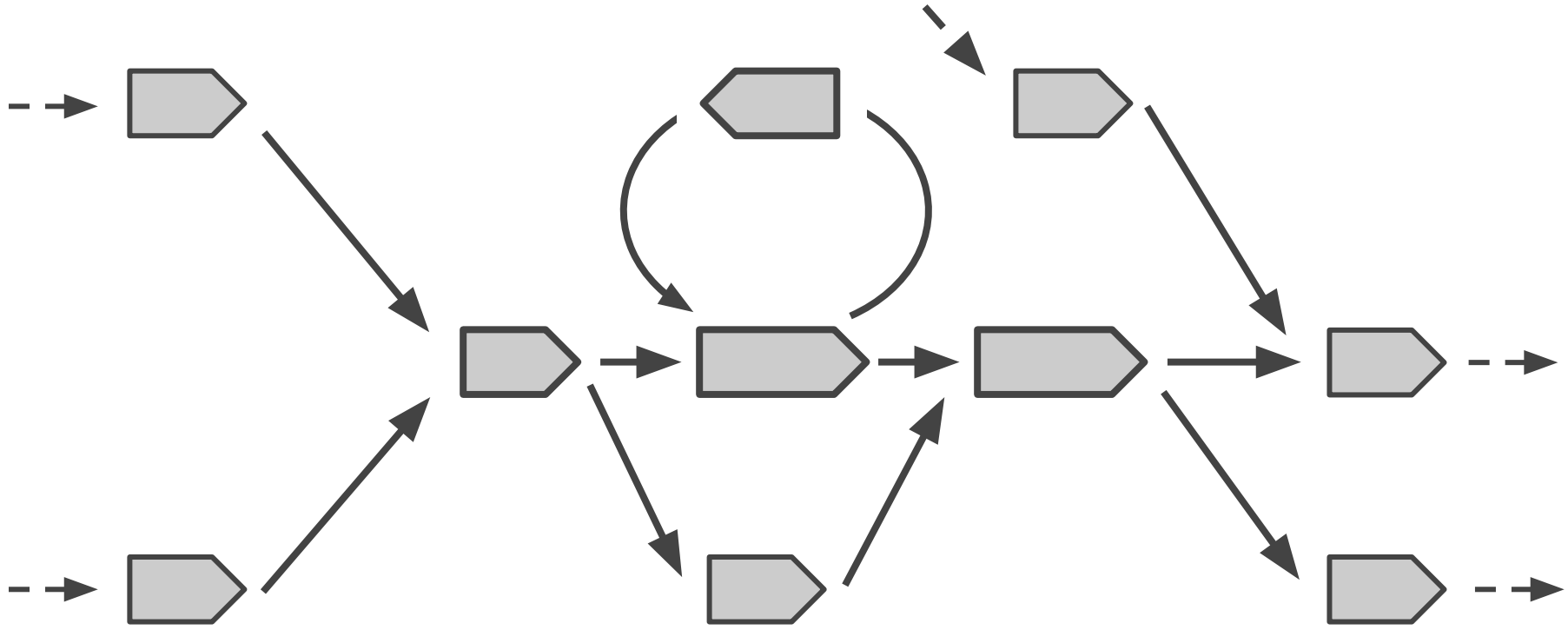
Can be represented as a graph of sequences linked by their overlaps



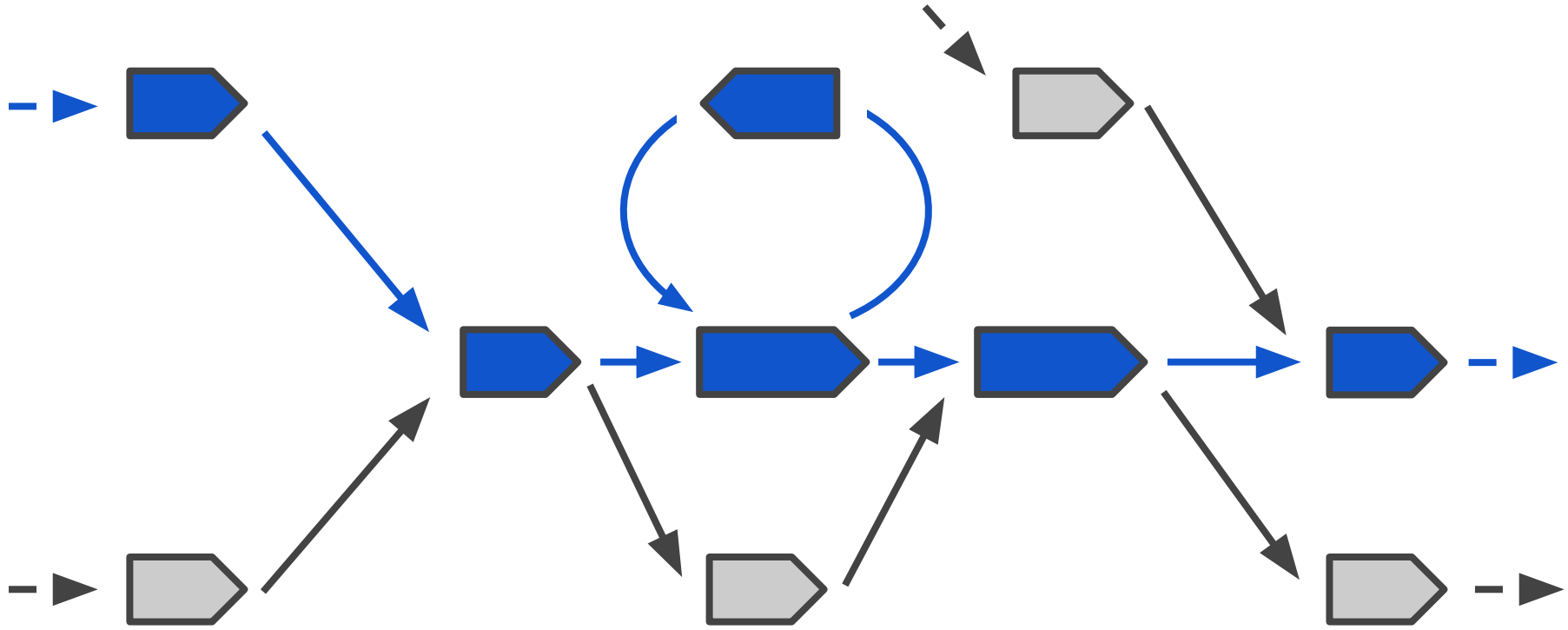
(This problem also comes up for mRNA alternative splicing)


And real metagenomes
are **very** complex

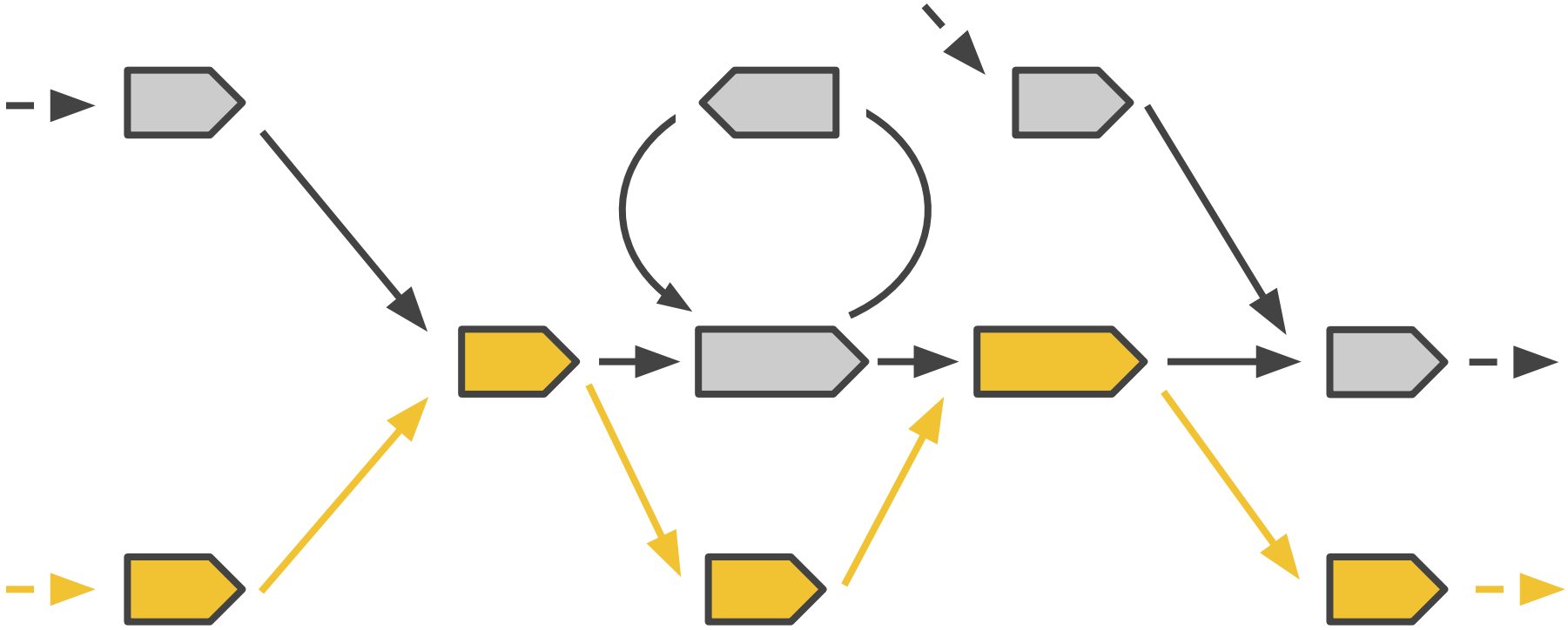
Real genomic sequences are paths on the graph



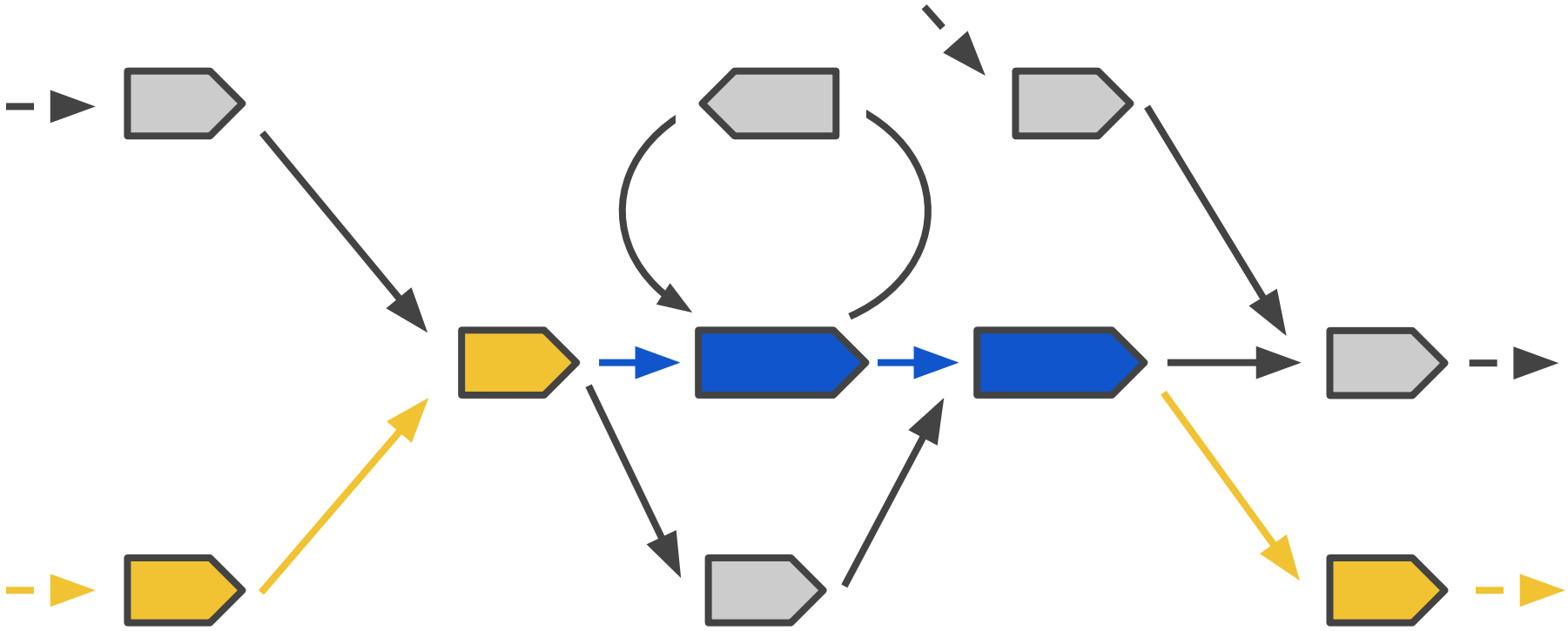
Real genomic sequences are paths on the graph



Real genomic sequences are paths on the graph



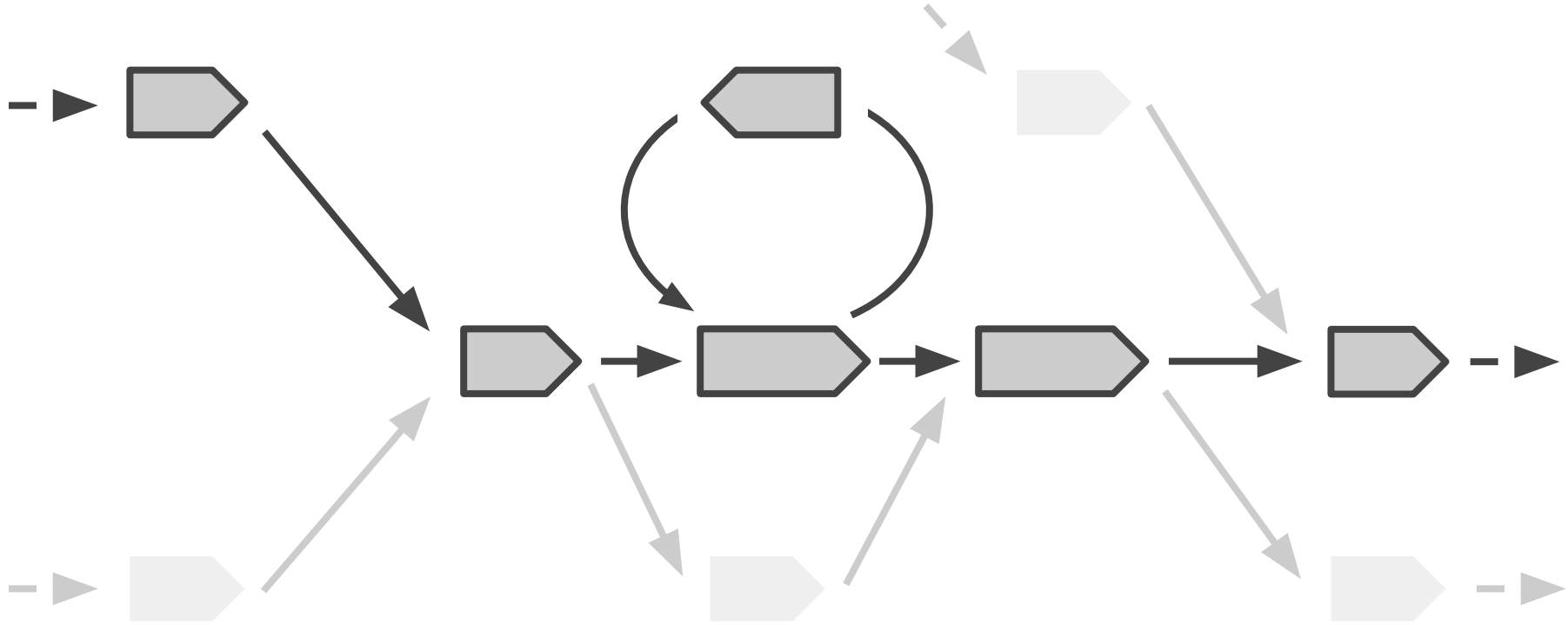
Lots of incorrect paths also exist…
How do we avoid these?



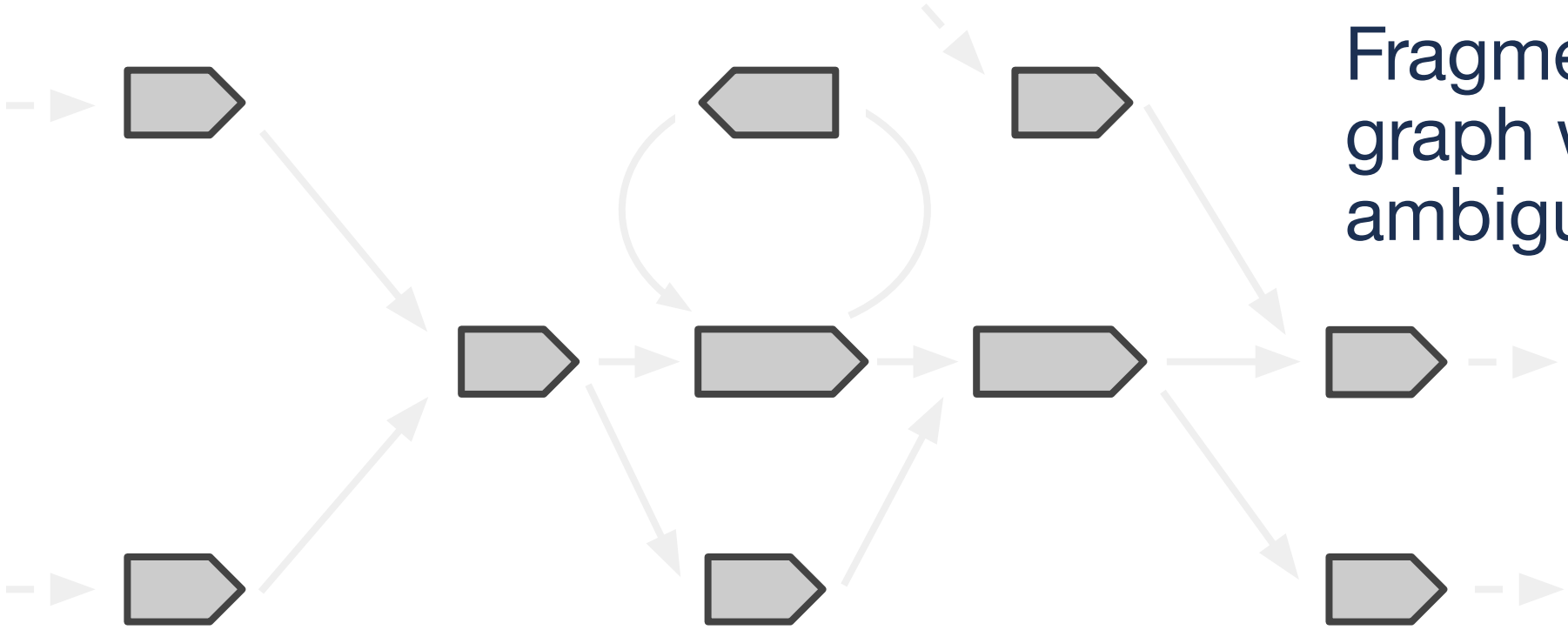
Lots of incorrect paths also exist…
How do we avoid these?

Standard Tools:

Filter out low-abundance sequences



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

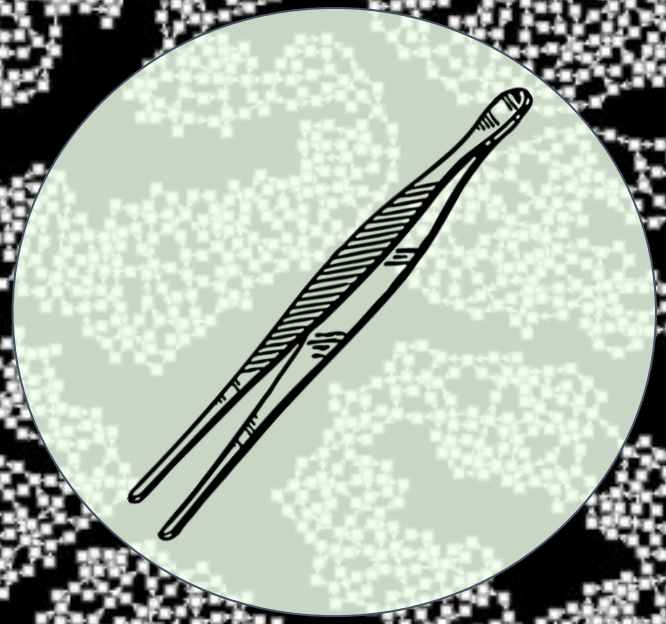


Standard Tools:

Filter out low-abundance sequences

Fragment the graph when it's ambiguous

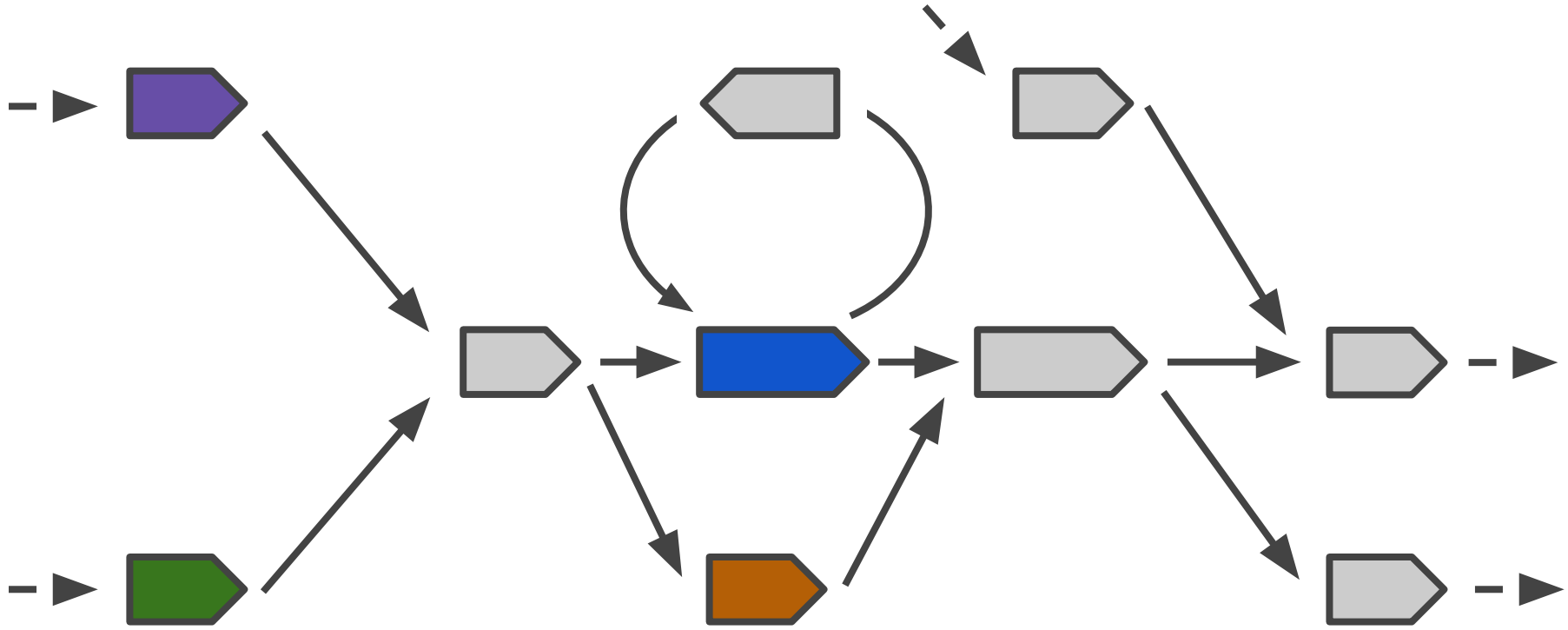
Untangling the hairball



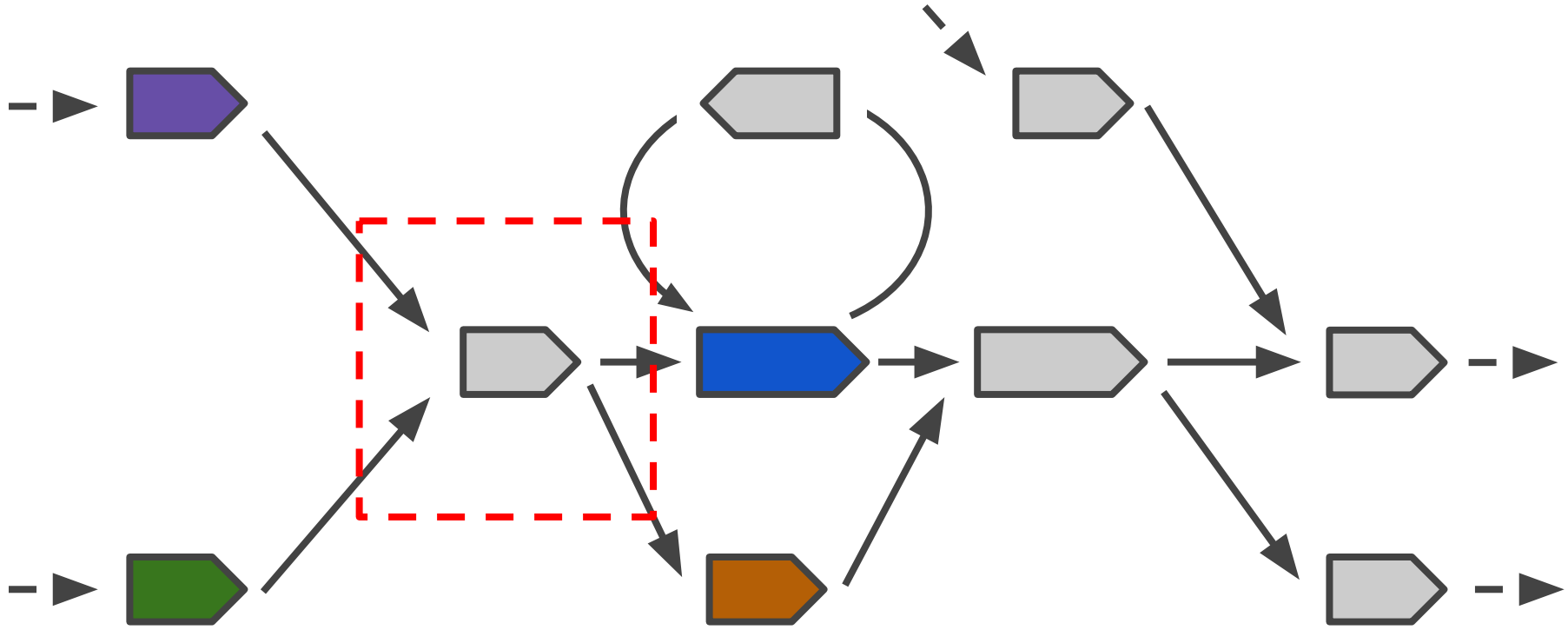
StrainZip:

Untangling the metagenome graph

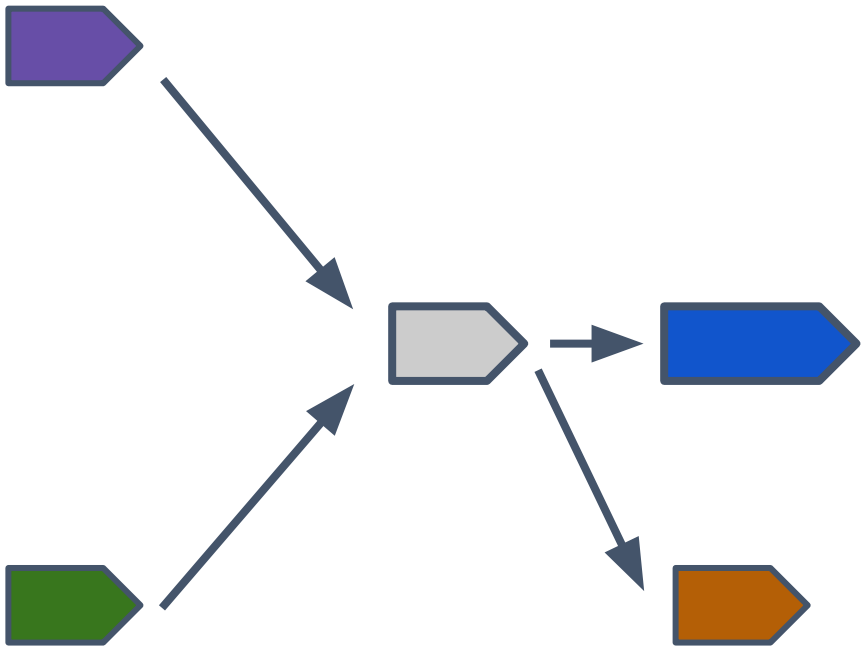
How can we recover long, accurate genome sequences from short reads?



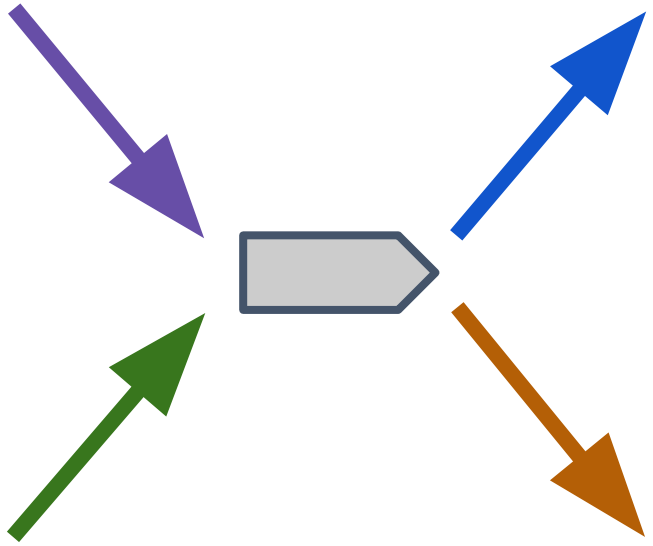
How can we recover long, accurate genome sequences from short reads?



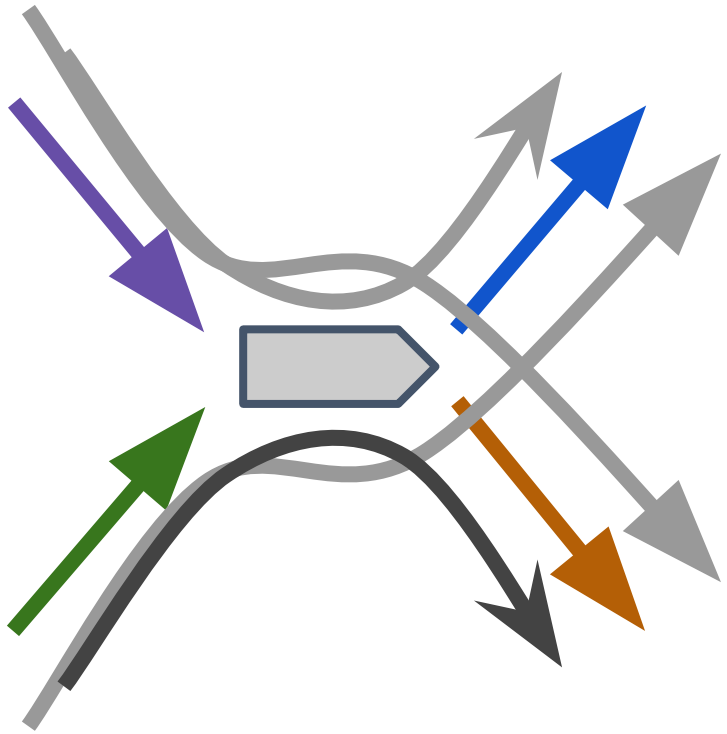
Focus on just one junction at a time

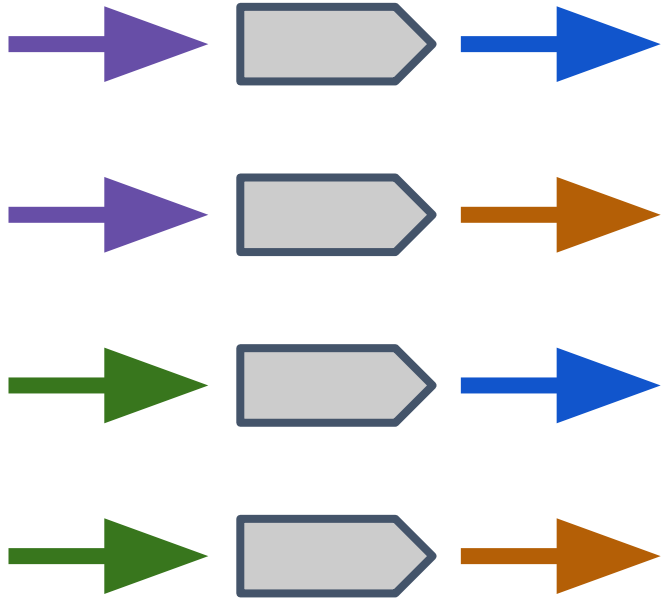


Focus on just one junction at a time

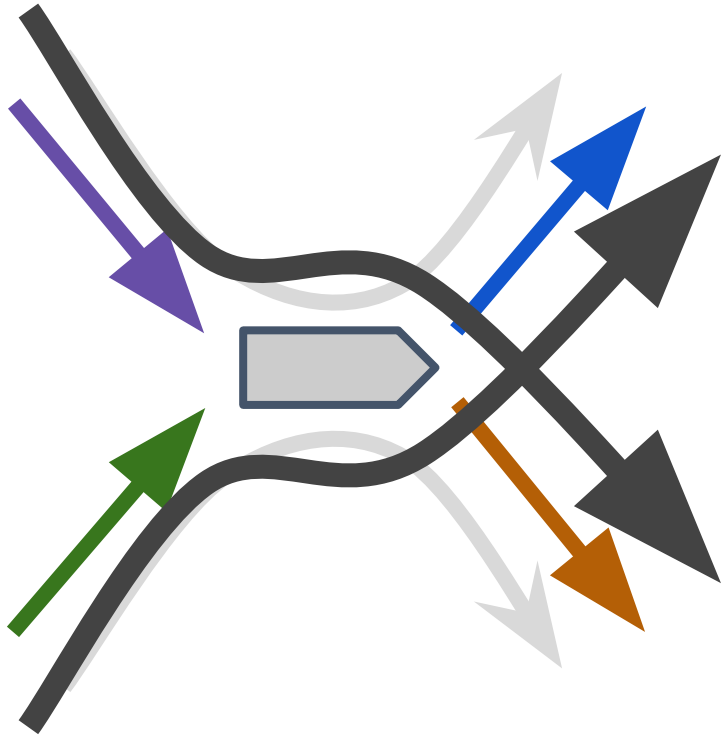


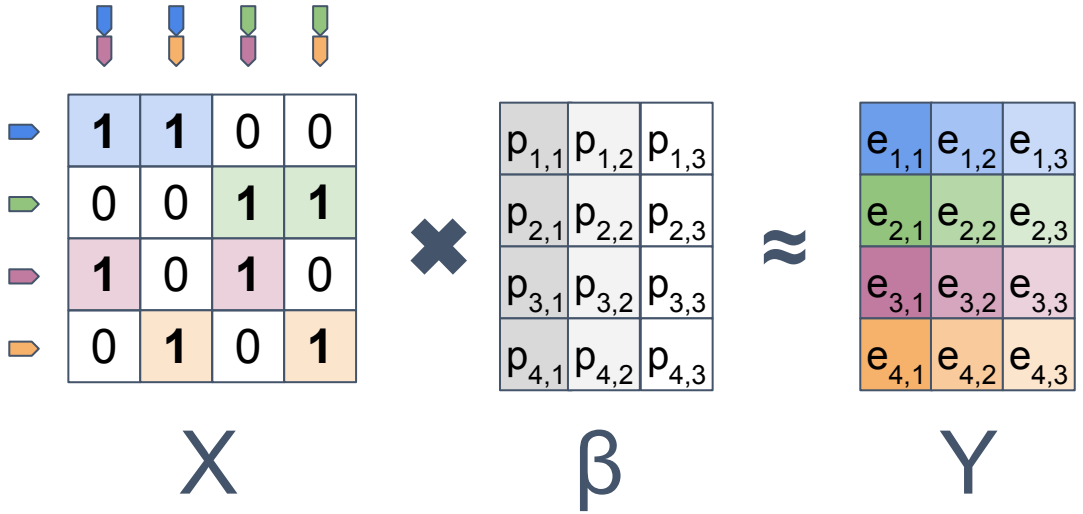
Focus on just one junction at a time





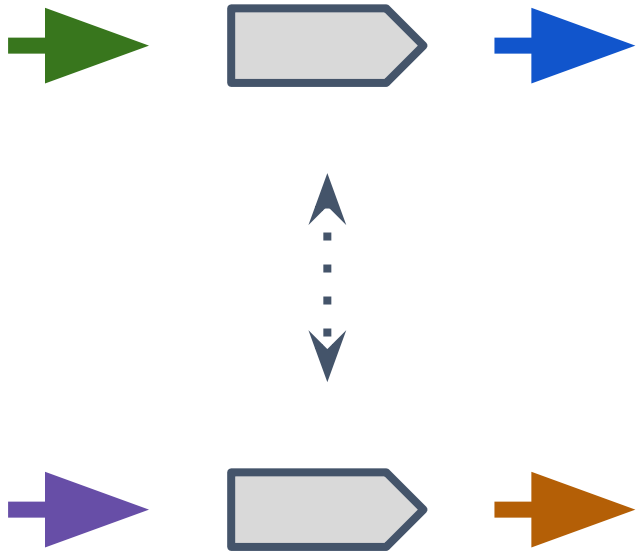
Focus on just one junction at a time
Select local paths



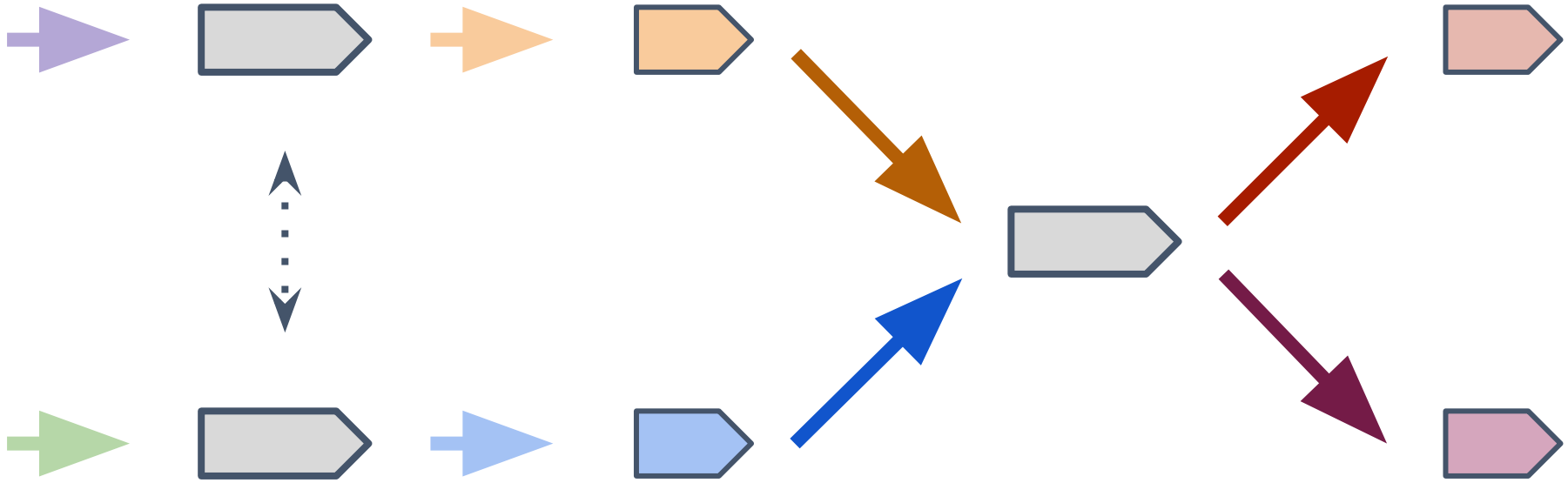


Sparse linear regression across multiple samples

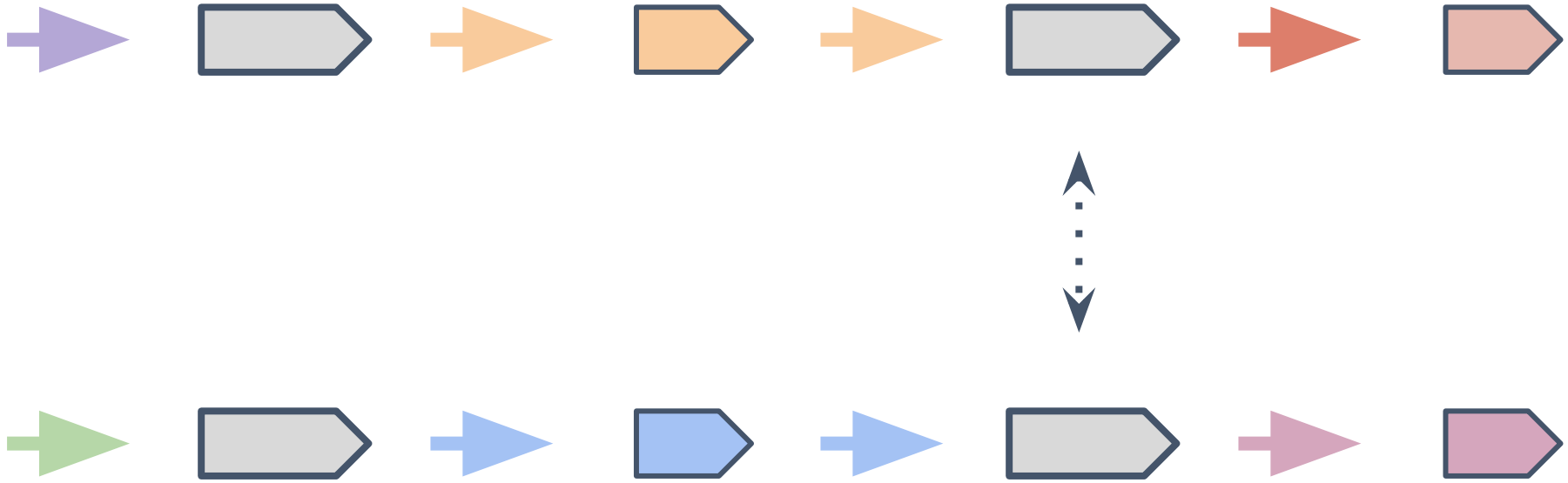
Focus on just one junction at a time
Select local paths
Unzip



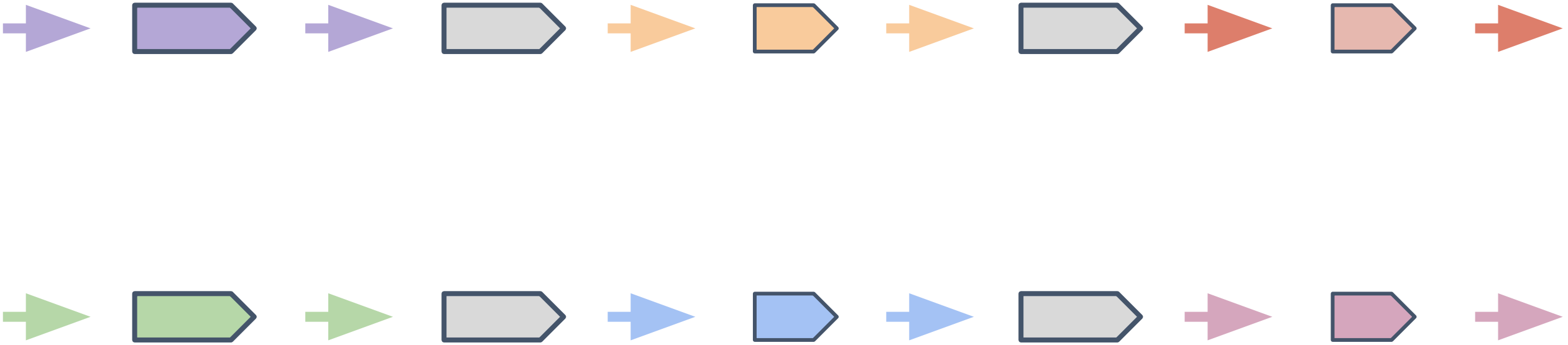
Focus on just one junction at a time
Select local paths
Unzip
Repeat

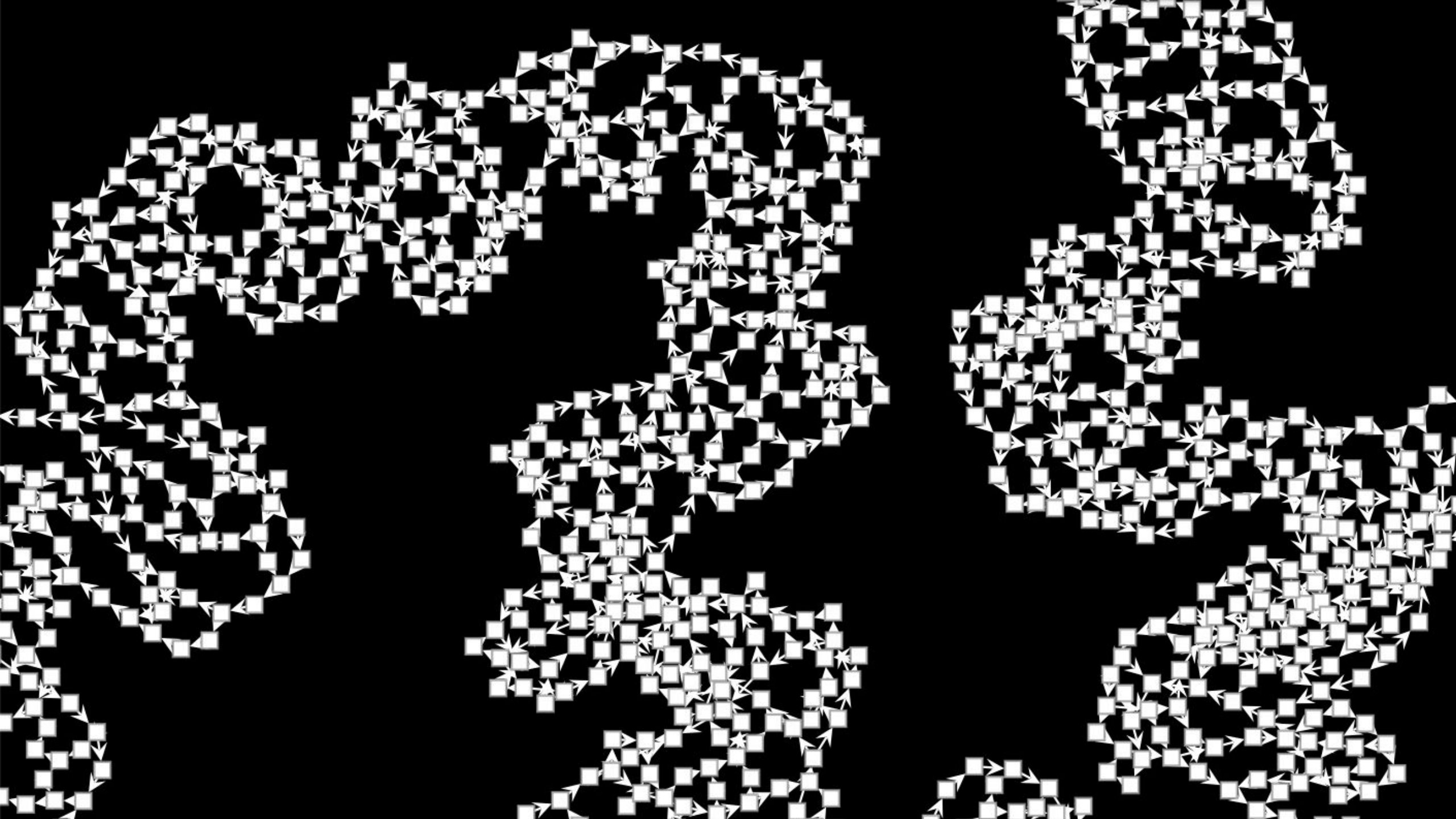


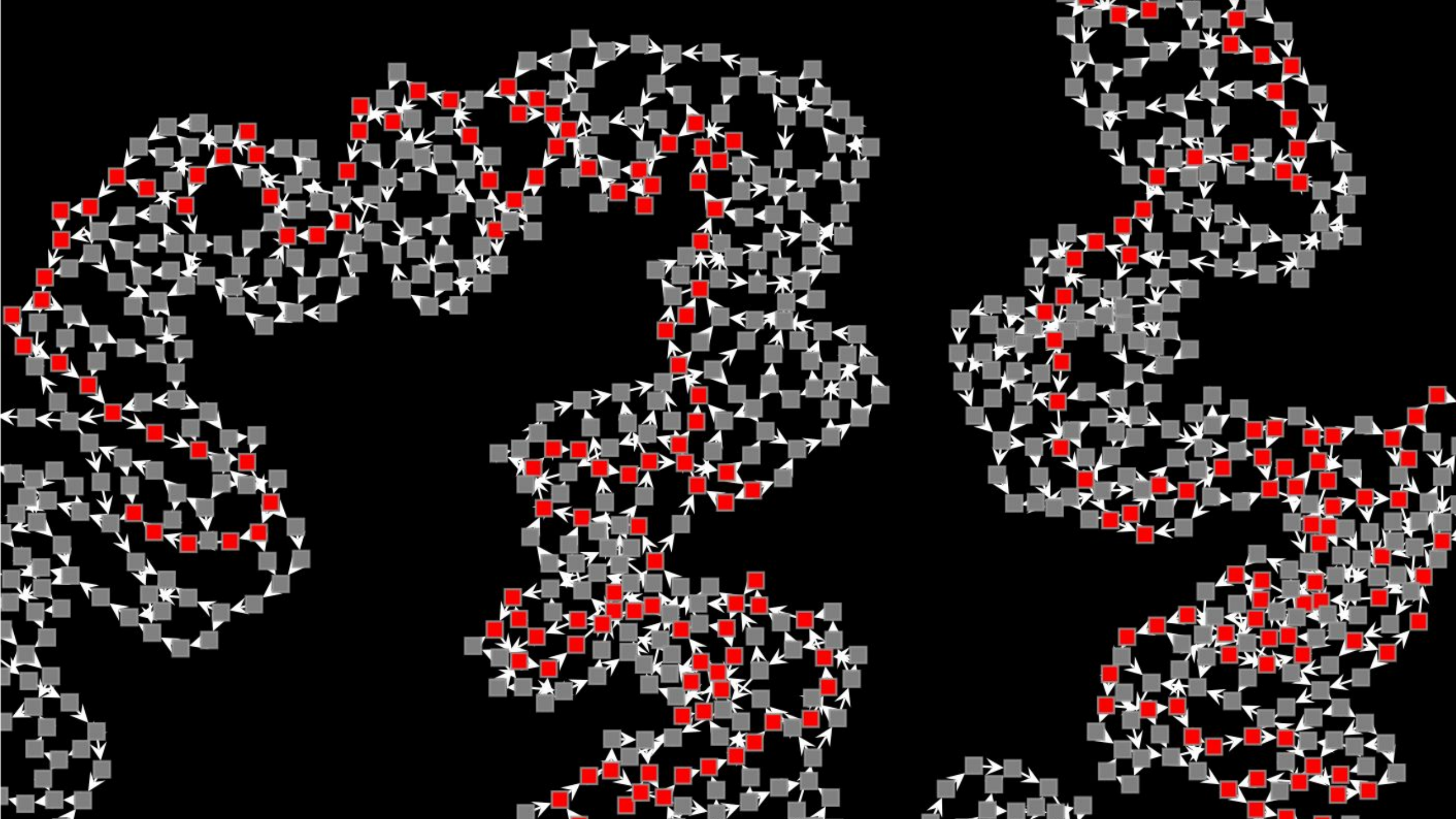
Focus on just one junction at a time
Select local paths
Unzip
Repeat



Focus on just one junction at a time
Select local paths
Unzip
Repeat

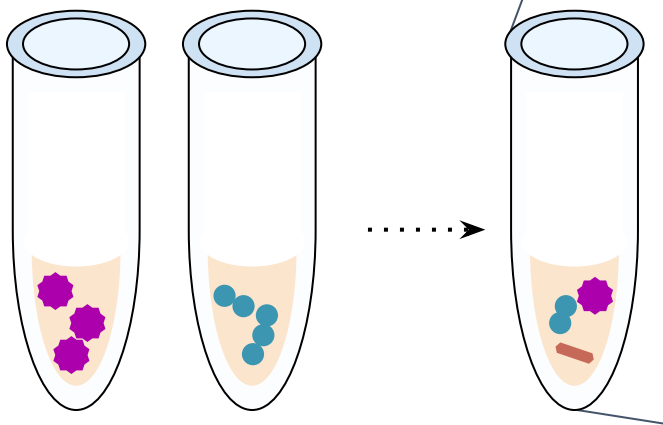






Strain-resolved discovery

Performance benchmarked on a complex, synthetic community

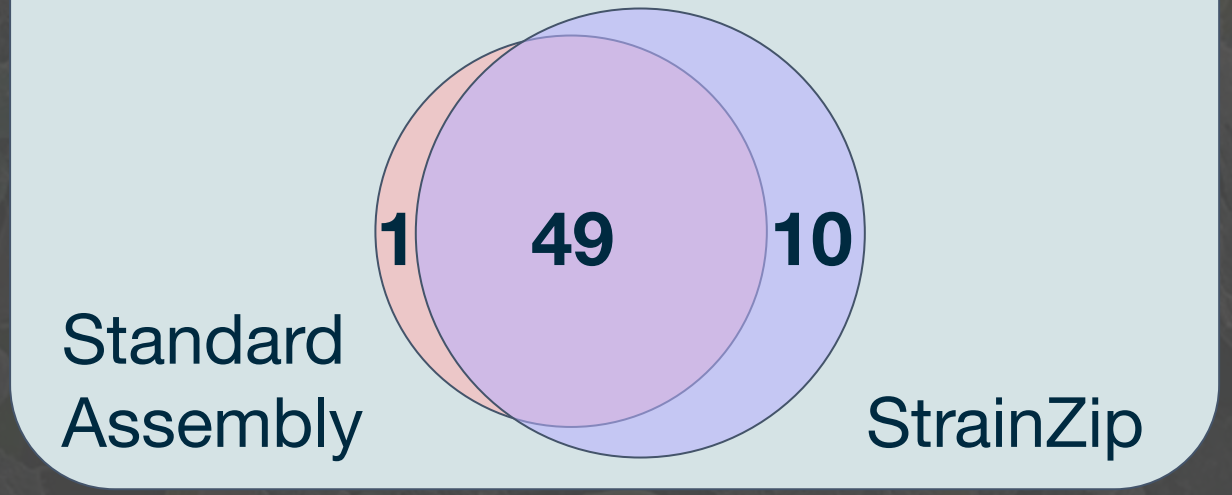


Antibiotic resistance genes are widespread in the gut microbiome

- Detection can inform treatment

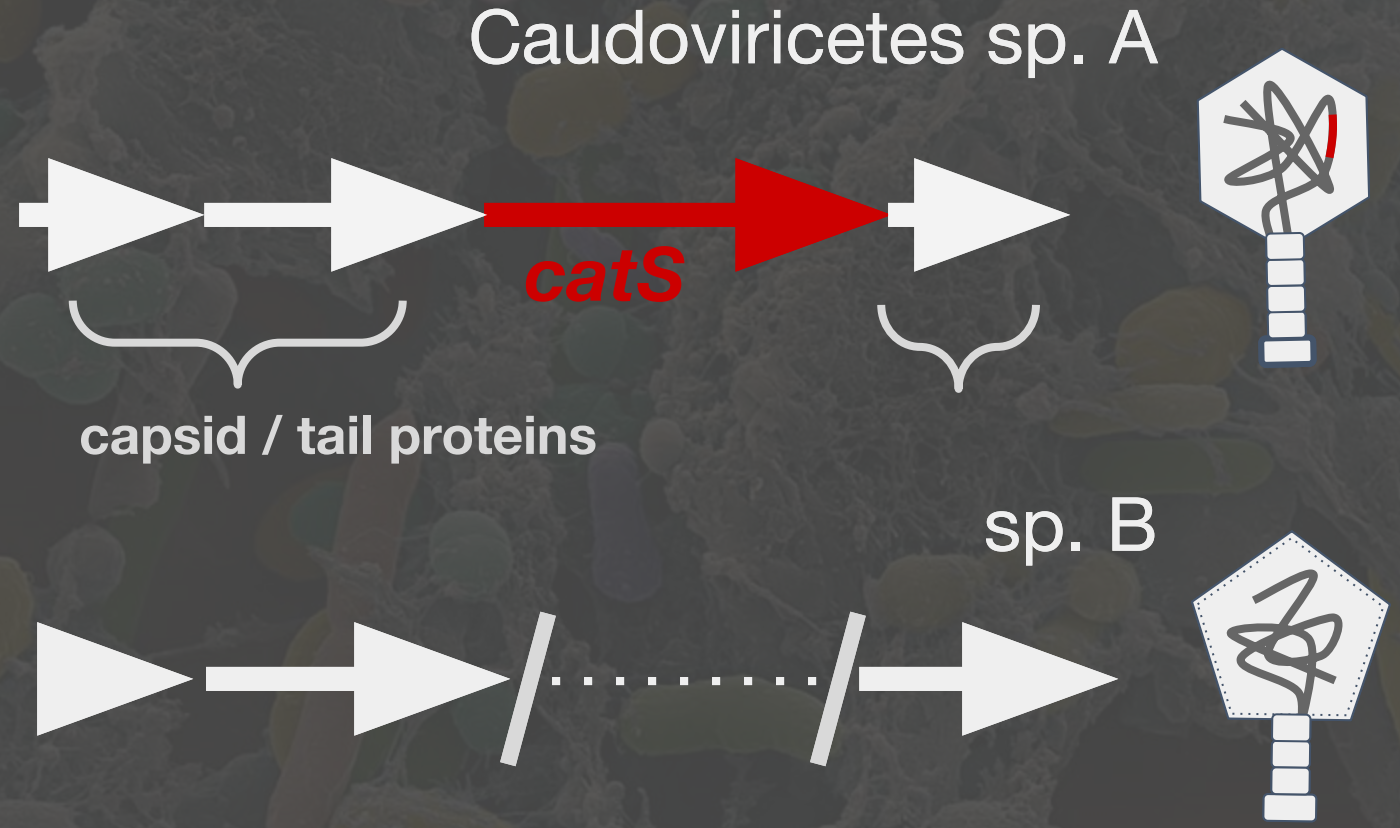
Antibiotic resistance genes are widespread in the gut microbiome

- Detection can inform treatment



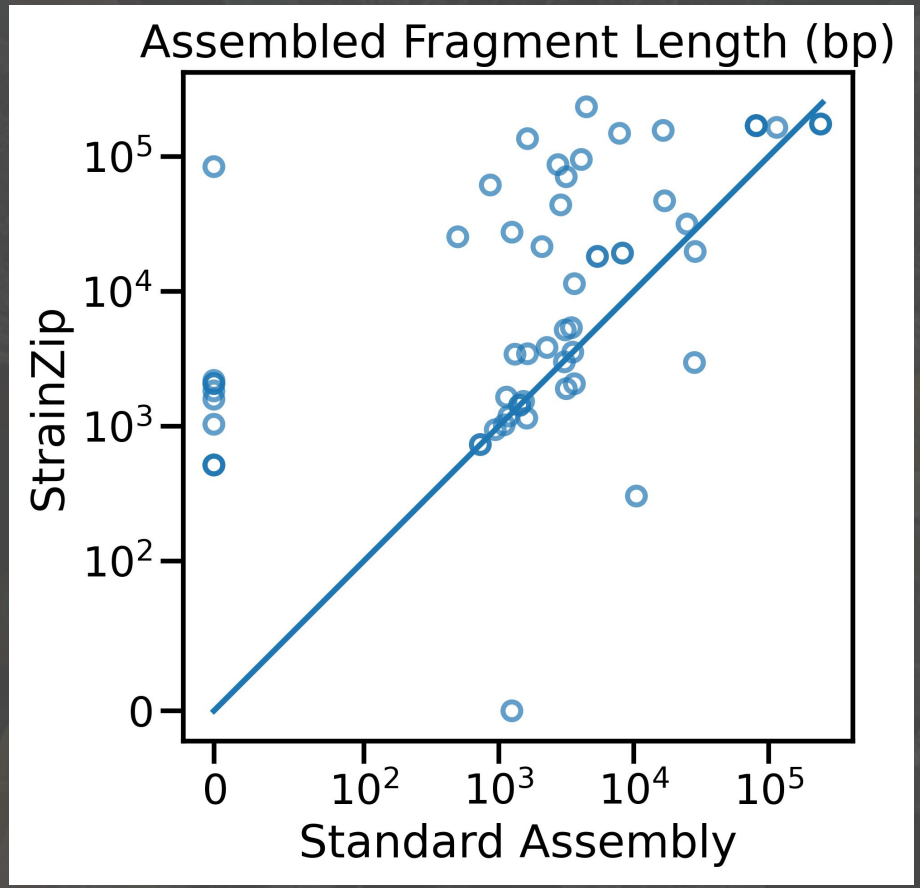
Antibiotic resistance genes are widespread in the gut microbiome

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

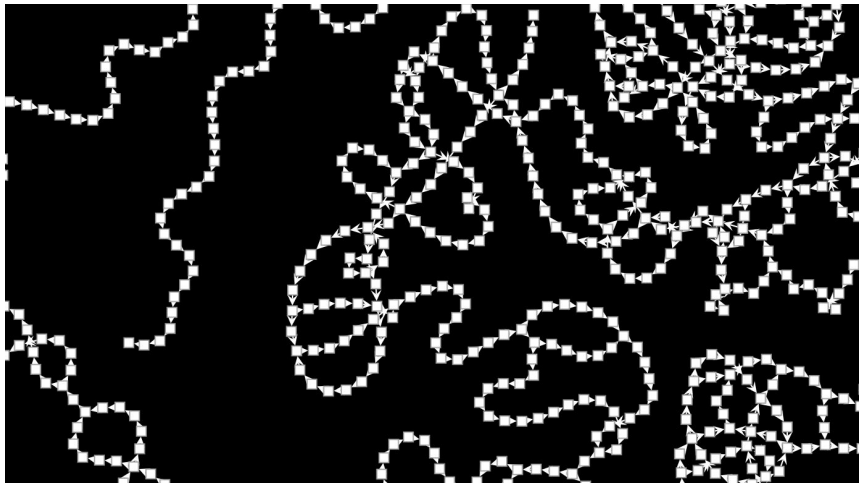


Antibiotic resistance genes are widespread in the gut microbiome

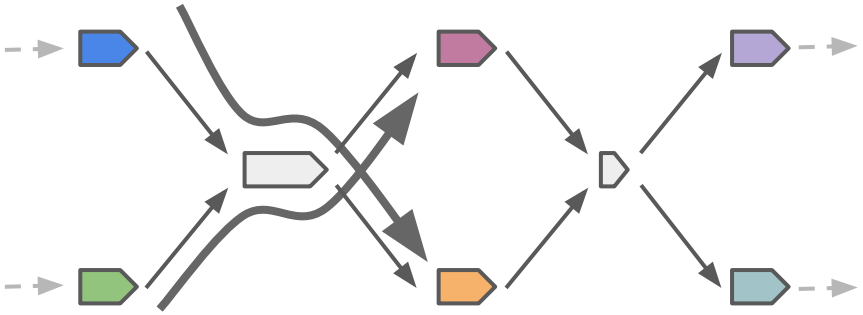
- Detection can inform treatment
- Can be carried in phage genomes
- Long sequence fragments provide useful information



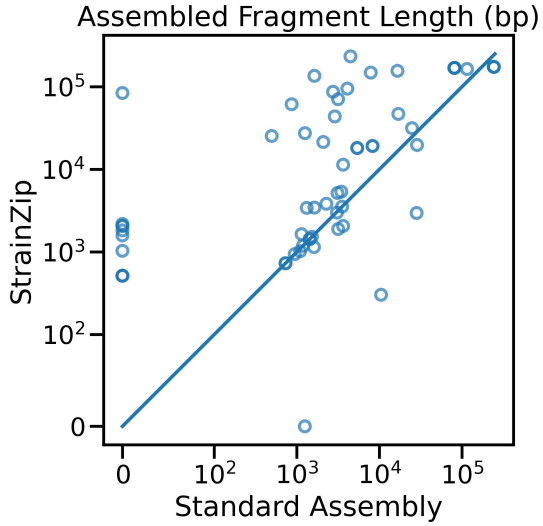
Complex Metagenome Graphs



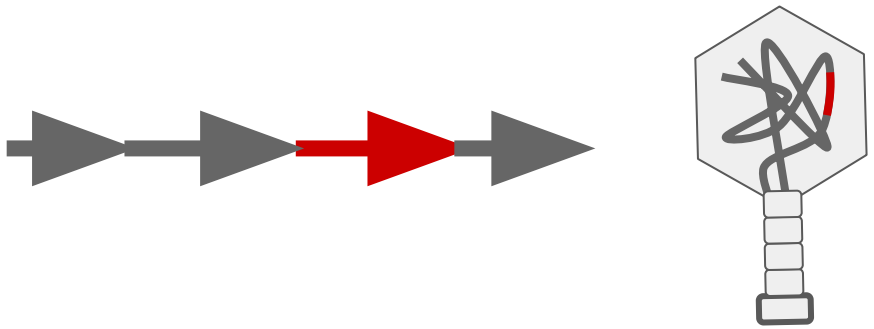
StrainZip Iteratively Unzips Junctions



Strain-Resolved Metagenomics

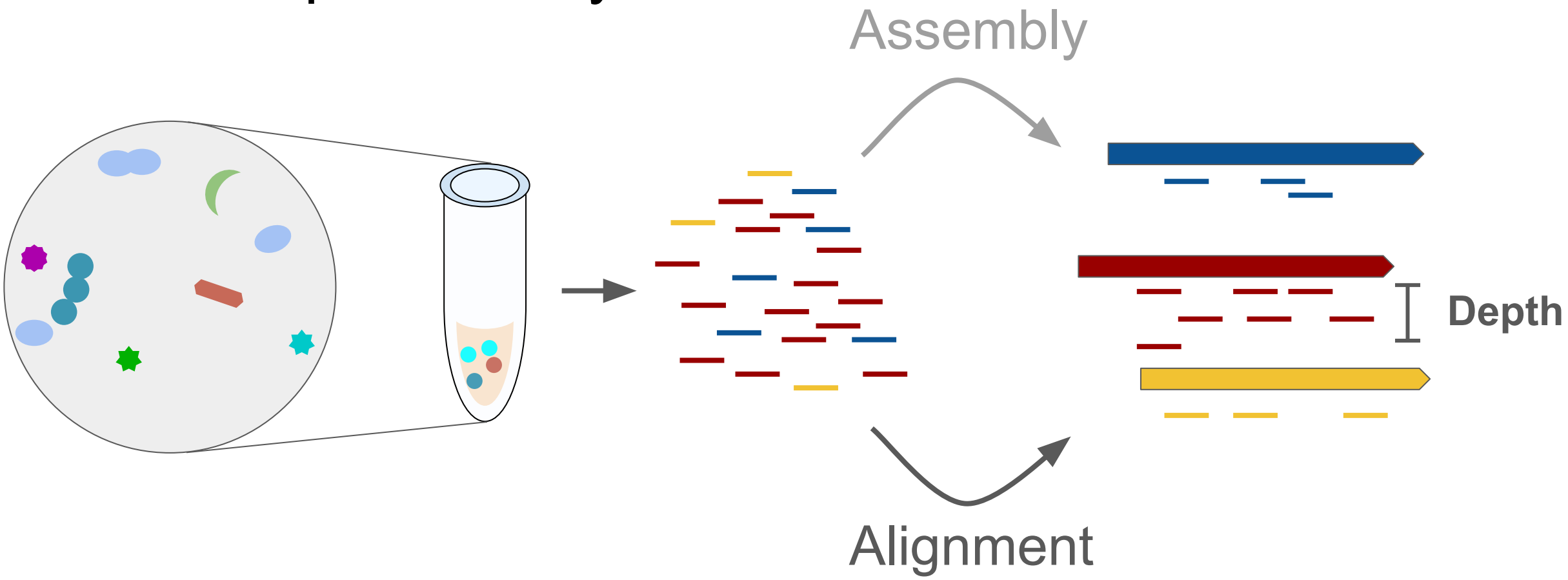


Antibiotic Resistance Potential of Phage

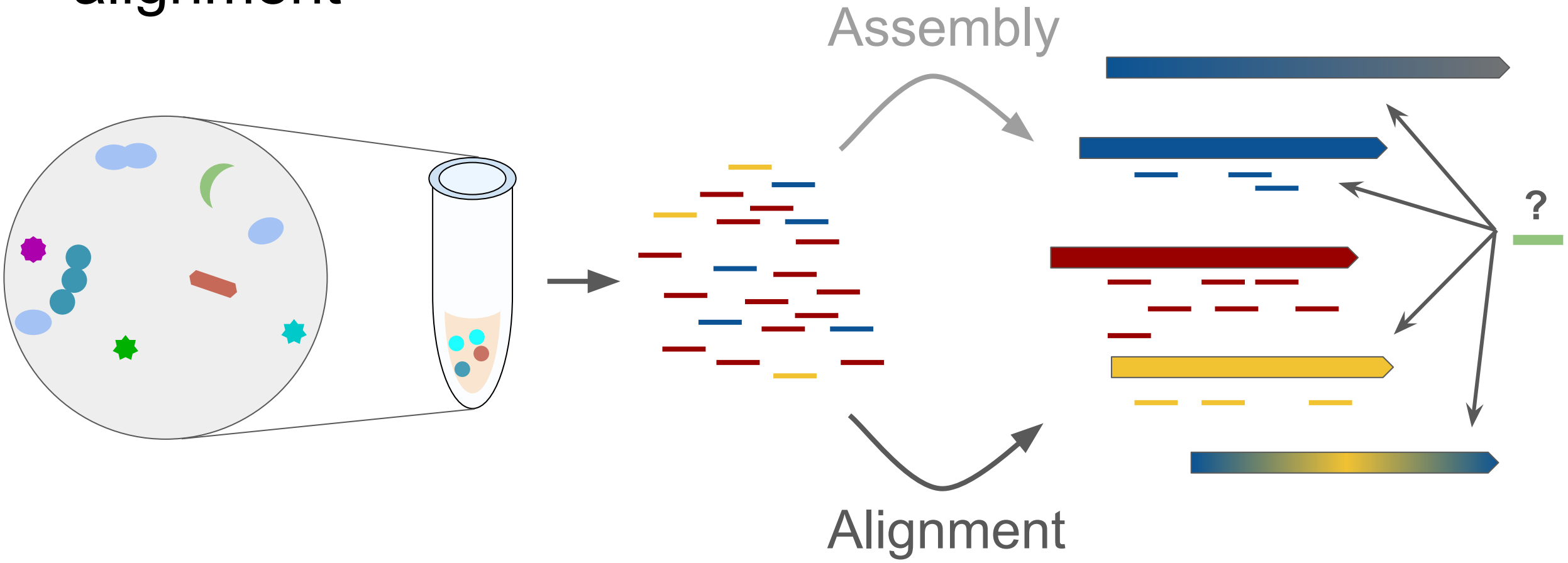


Rewind: I also care about depth quantification

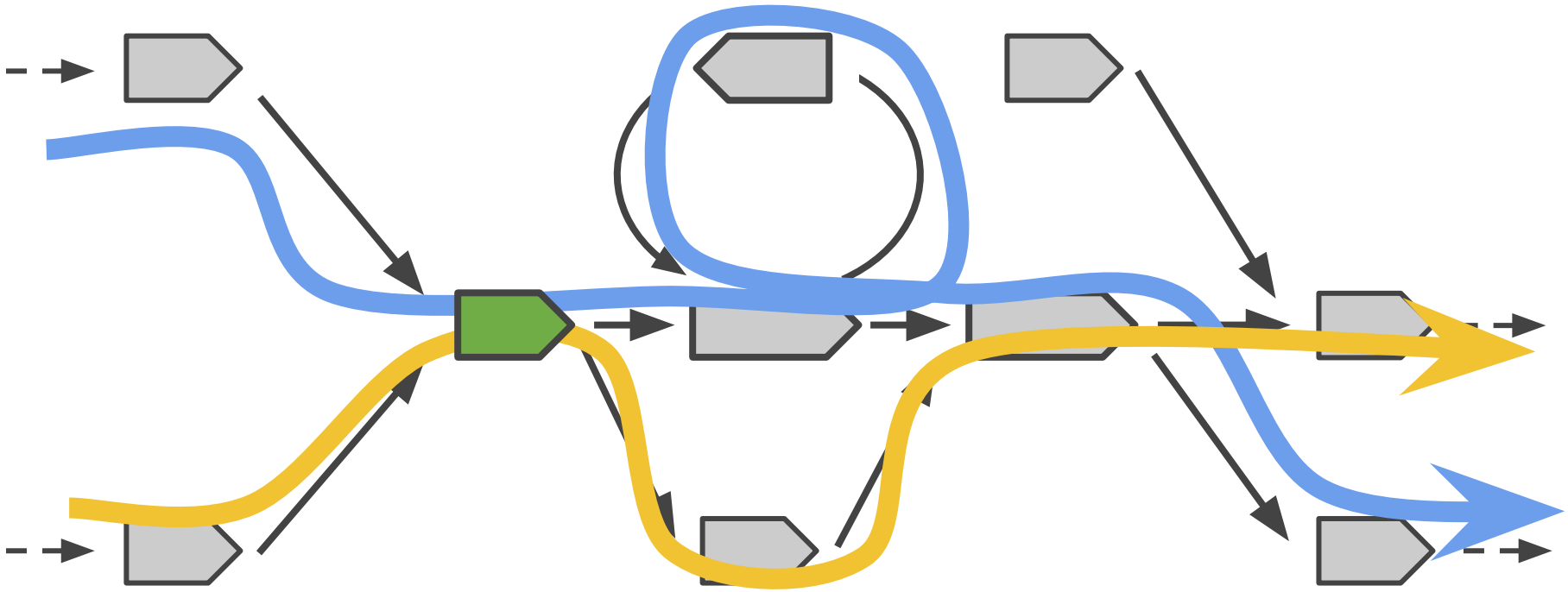
Assembly and depth quantification are complementary



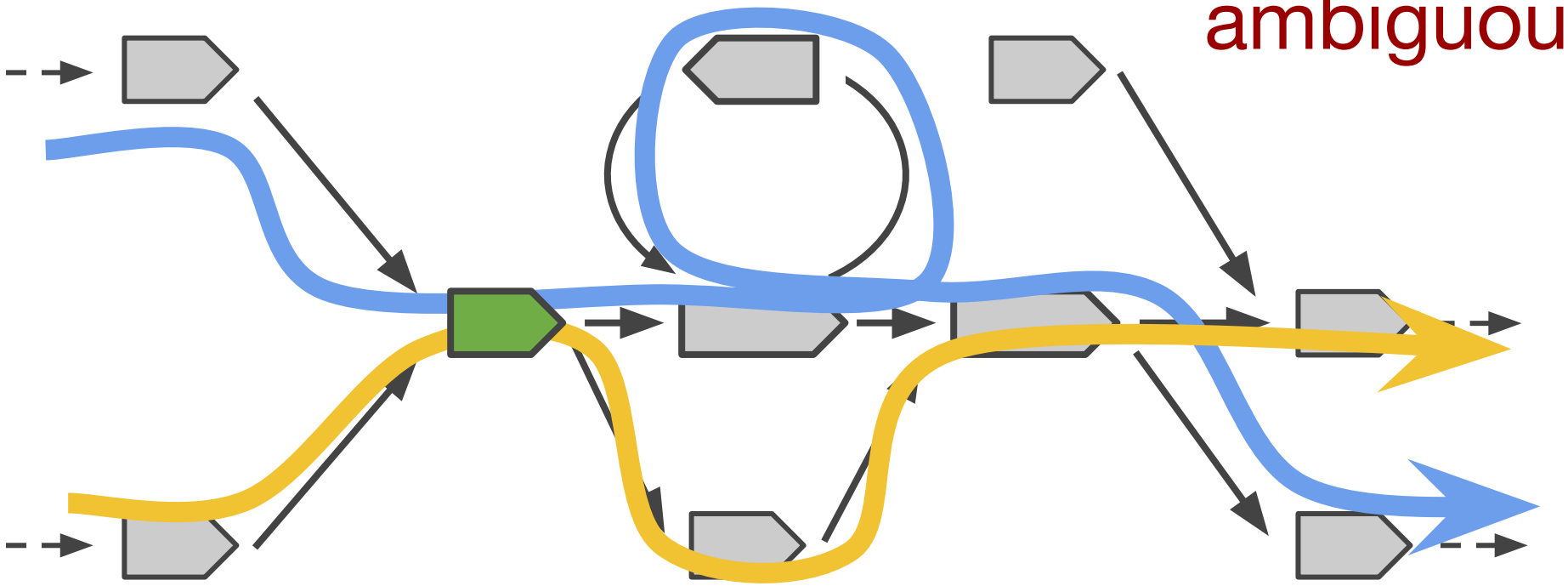
Closely related sequences are a major challenge for alignment



Shared sequences mean reads map ambiguously



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



Quick intro to de Bruijn graphs

Read #1

...CGTACCTGGATTAC...

Assembly

...CGTACCTGGATTACTTAA...

Read #2

CCTGGATTACTTAA...

De Bruijn graphs

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

Fragment reads into k-mers

Read #1

…CGTACCTGGATTAC

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**

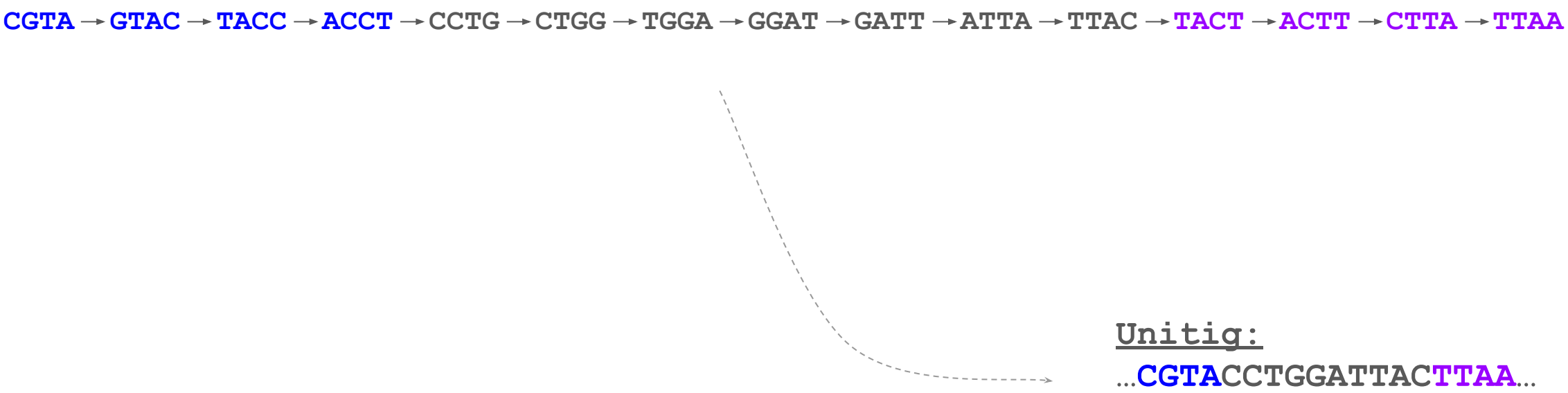
Identify k-mer pairs where (k-1) suffix on one is same as other's prefix

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



Mutations / errors introduce new k-mers

Read #1

…CGTACCTGGATTAC

CGTA

GTAC

TACC

ACCT

CCTG

CTGG

TGGA

GGAT

GATT

ATTA

TTAC

Read #2

CCTGCATTACTTAA…

CCTG

CTGC

TGCA

GCAT

CATT

ATTA

TTAC

TACT

ACTT

CTTA

TTAA

Diversity / Errors

…CGTACCTGGATTACTTAA…

…CGTACCTGCATTACTTAA…

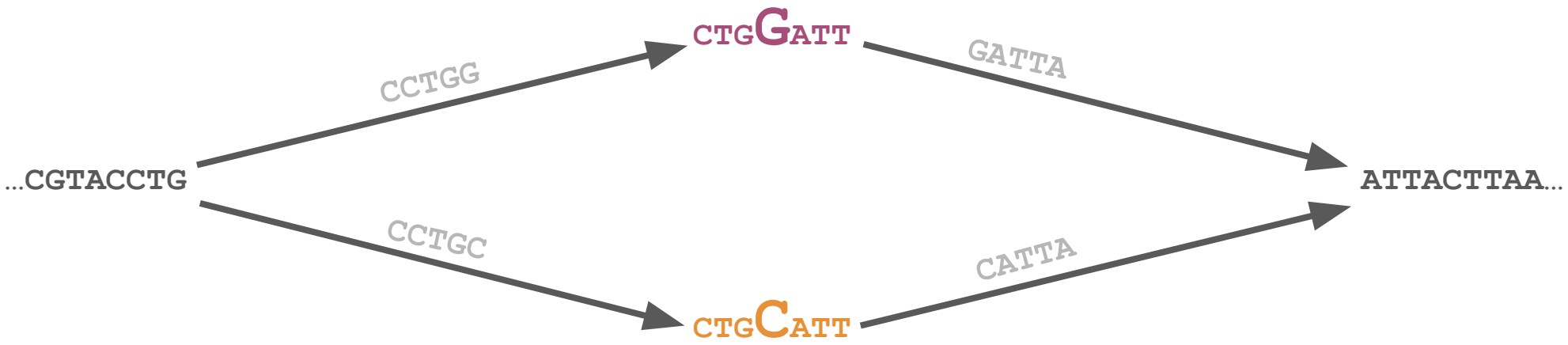
Same edge-drawing process

Same edge-drawing process

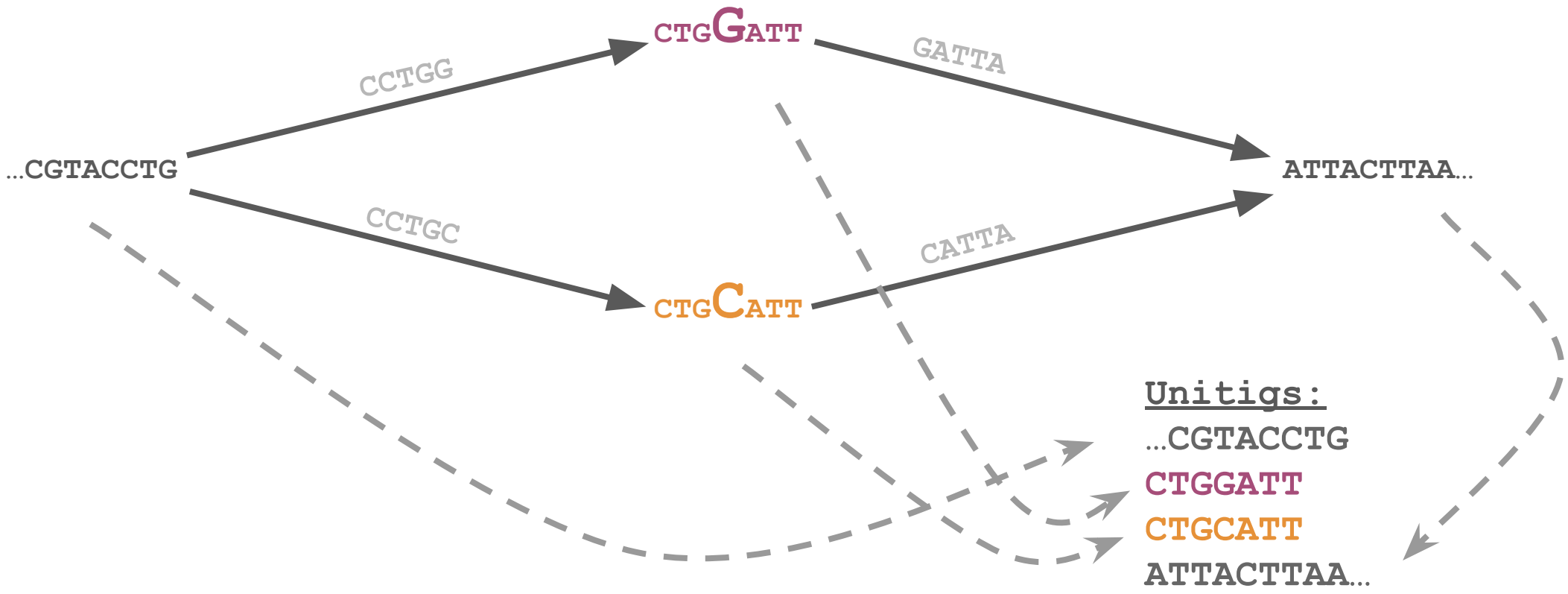
But now some k-mers have multiple edges

This introduces a "bubble"

The two sides of the bubble reflect the observed diversity



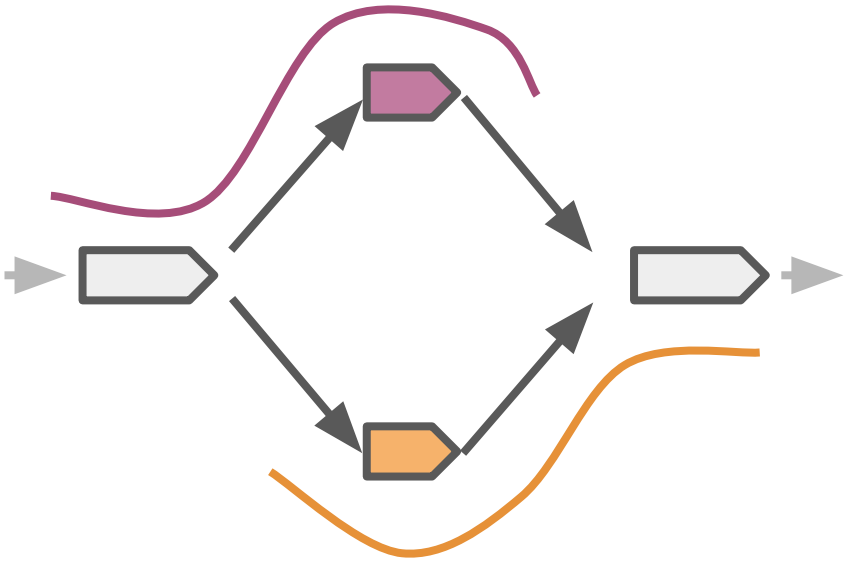
Again we extract unitigs, but now they're shorter, fragmented



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

Read #1
…CGTACTGGATTAC

Read #2
CCTGCATTACTTAA…



Alternatively: Exact k-mer counting

Unitig #1	Unitig #2	Unitig #3	Unitig #4
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

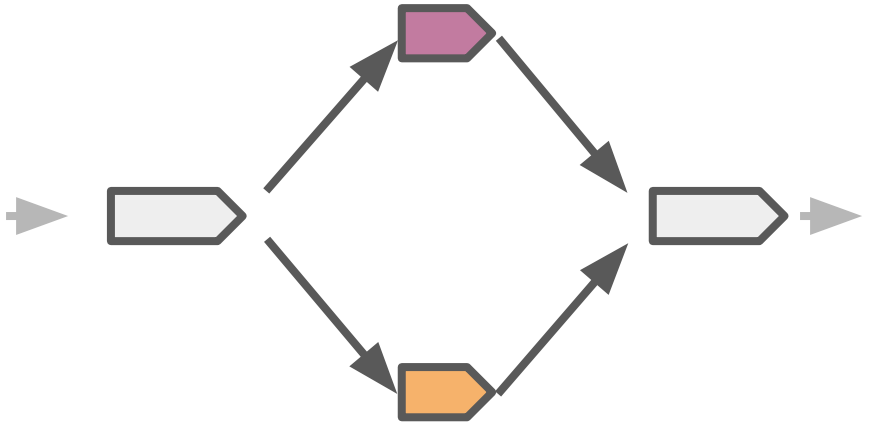
Every k-mer in the sample is in the dBG, by construction

Alternatively: Exact k-mer counting

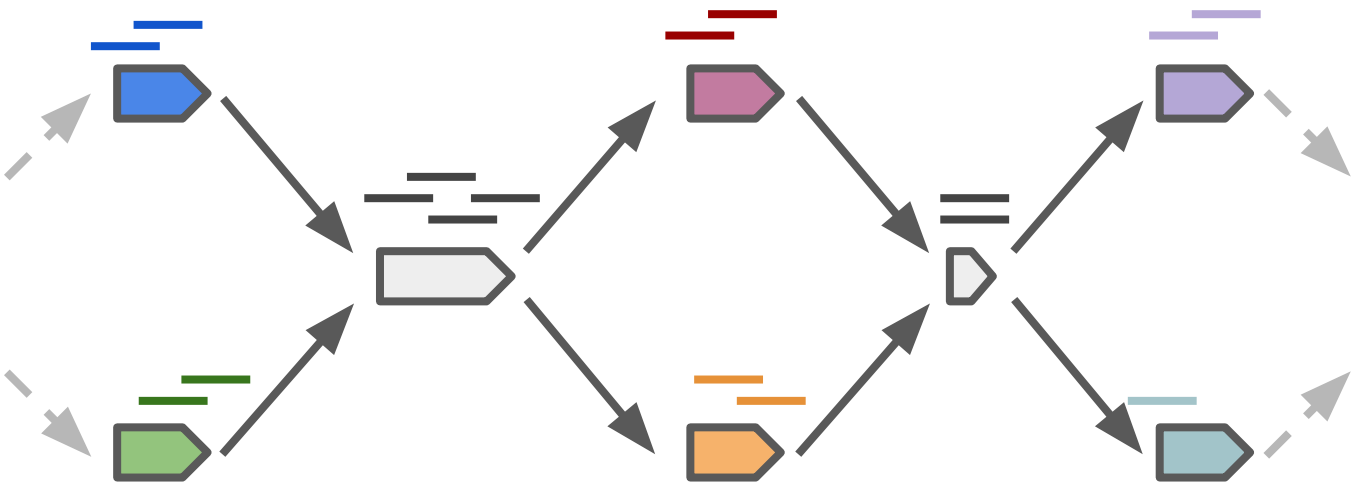
Much faster than read alignment

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



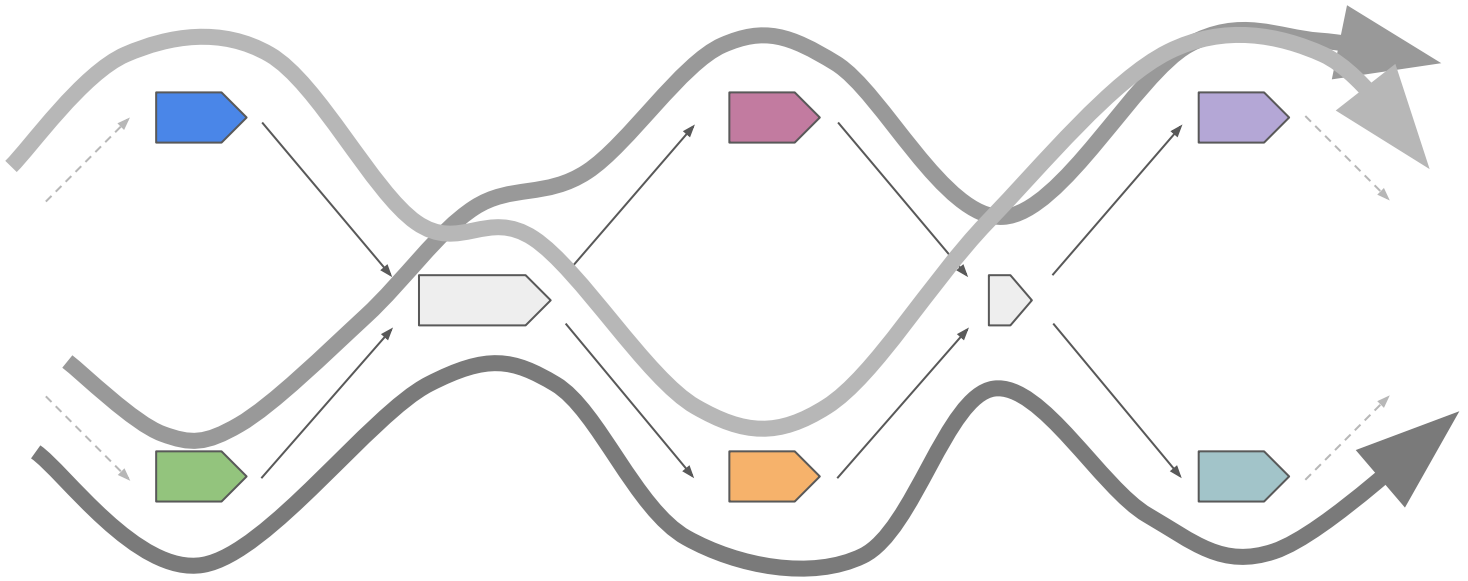
Path depths
(unknown)

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



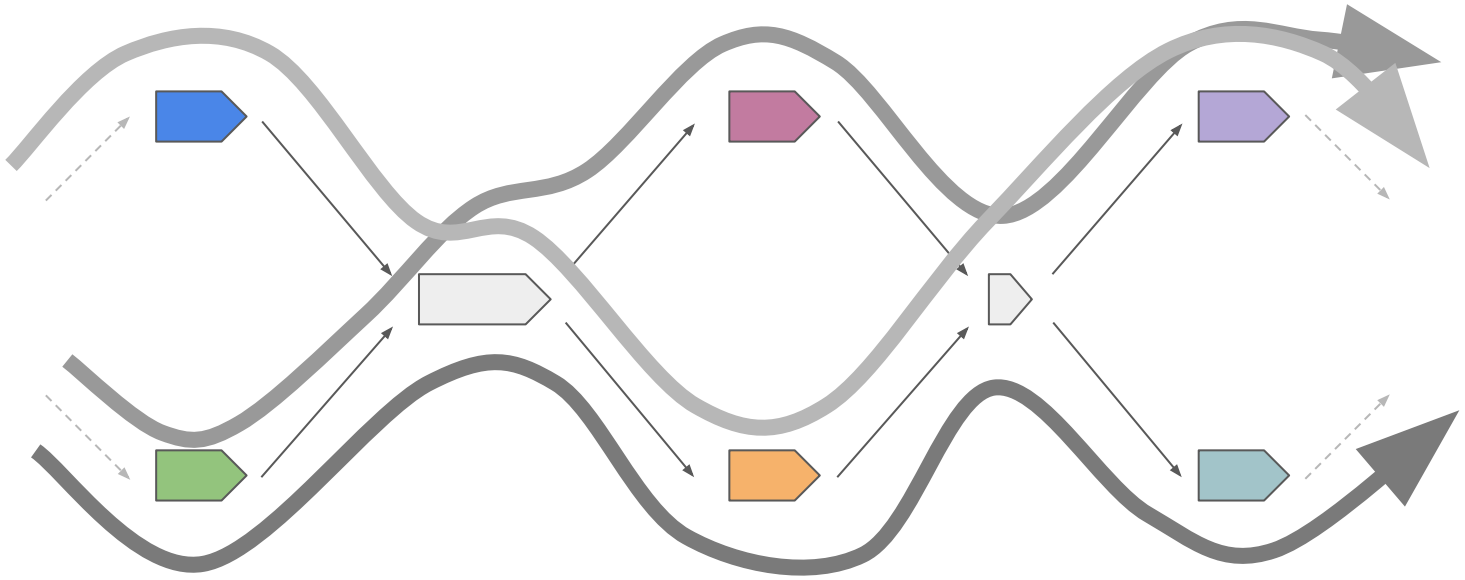
Path depths
(unknown)

Path
encoding

Observed
depths

$$X\beta \approx Y$$

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



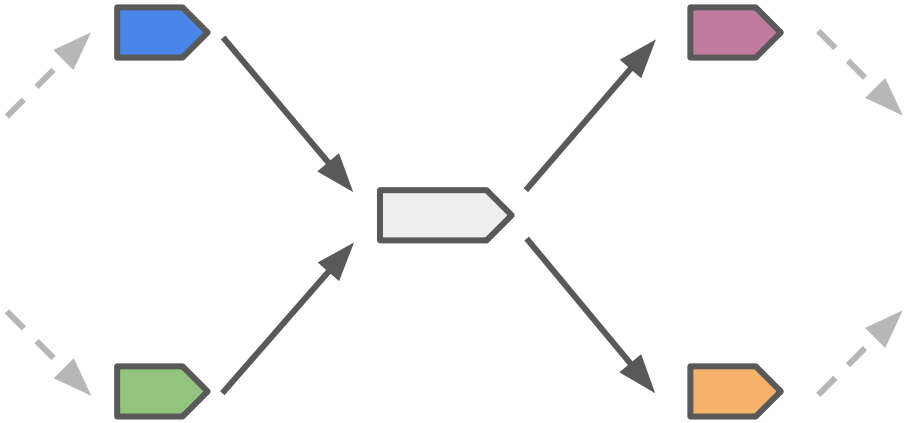
Estimate this

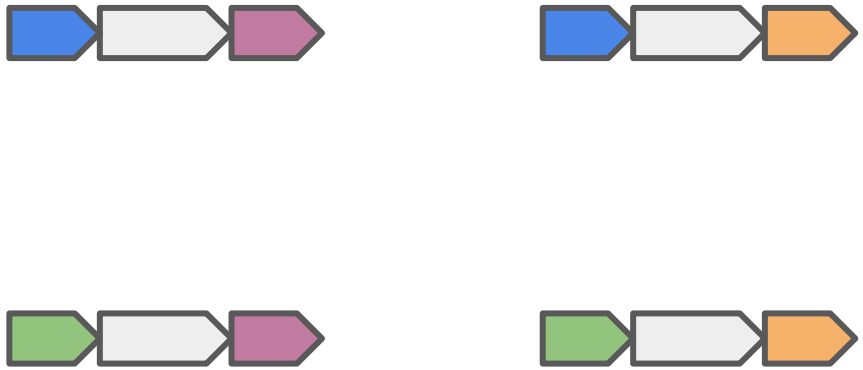
From these

$$X\beta \approx Y$$

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

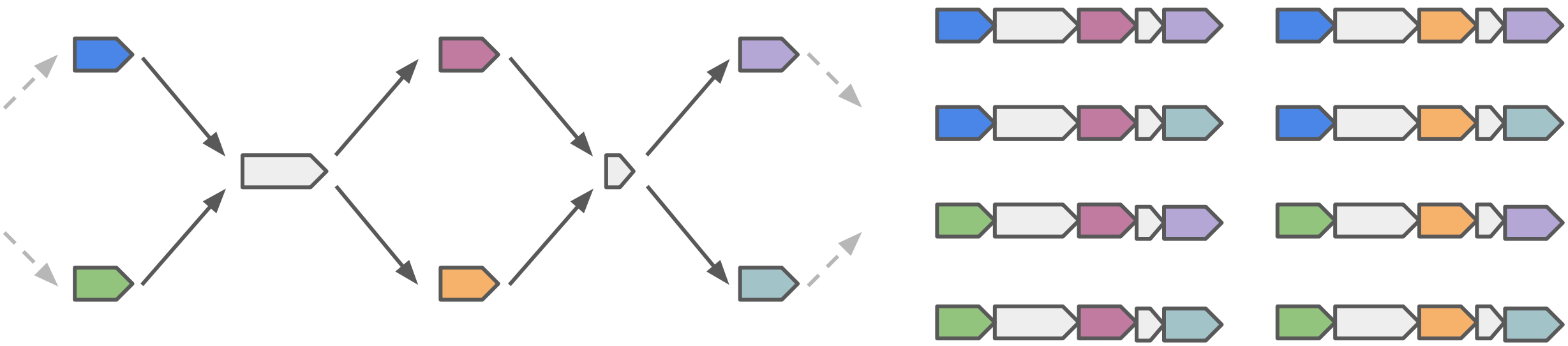
We can enumerate all possible paths on our assembly graph





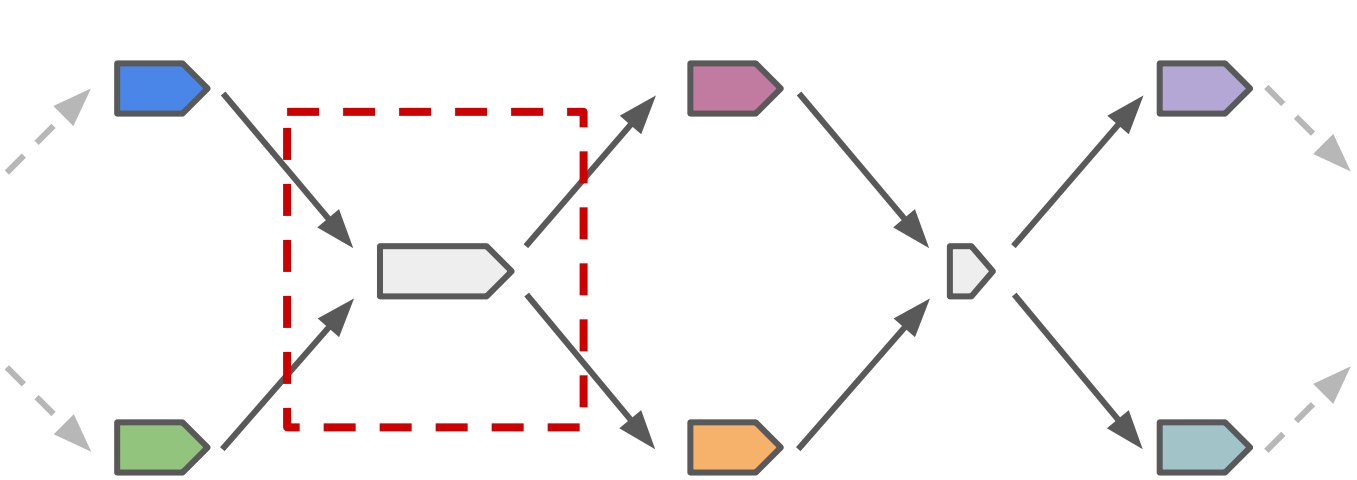
$$X\beta \approx Y$$

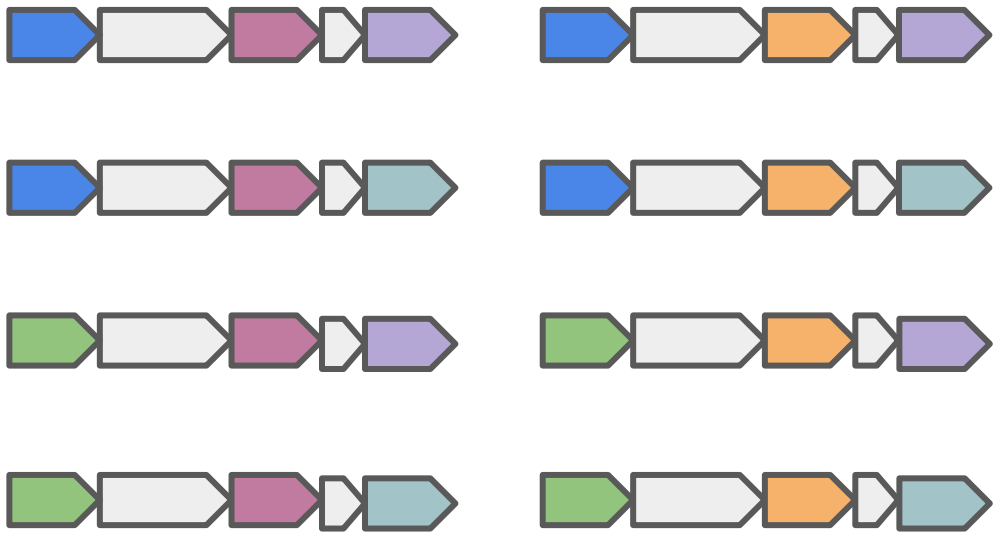
We can enumerate all possible paths on our assembly graph



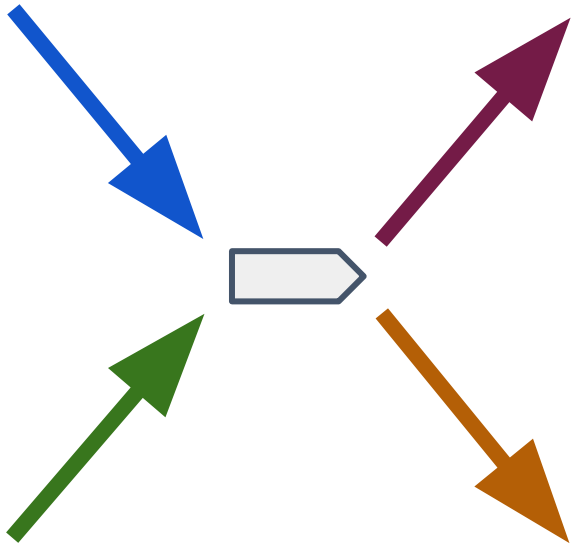
…but this grows exponentially with graph complexity

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

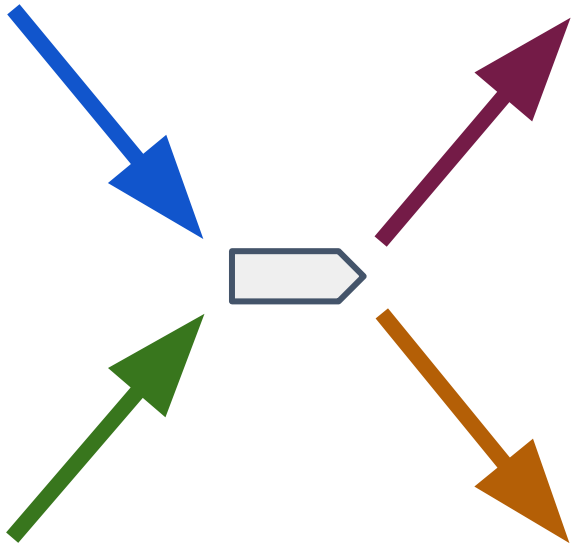


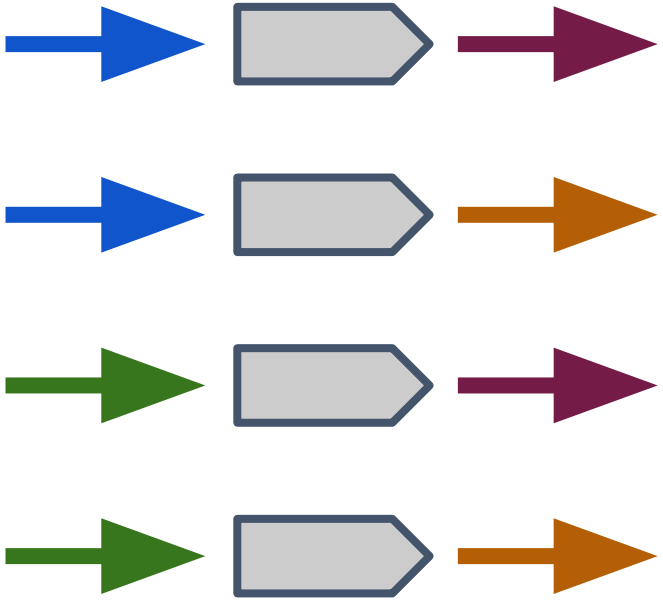


KEY IDEA: A single "junction" is the minimum unit of deconvolution

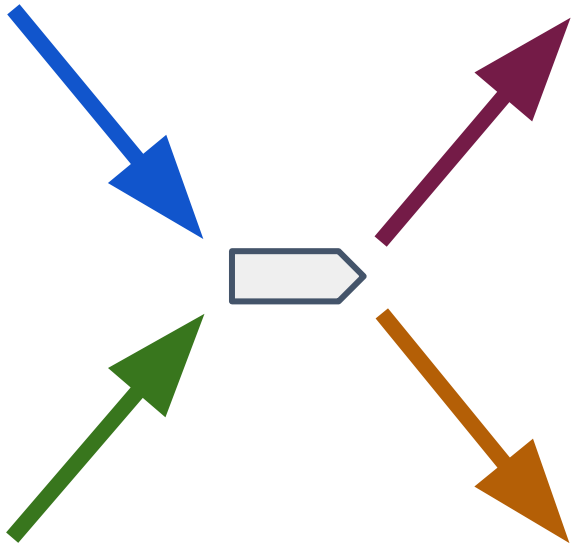


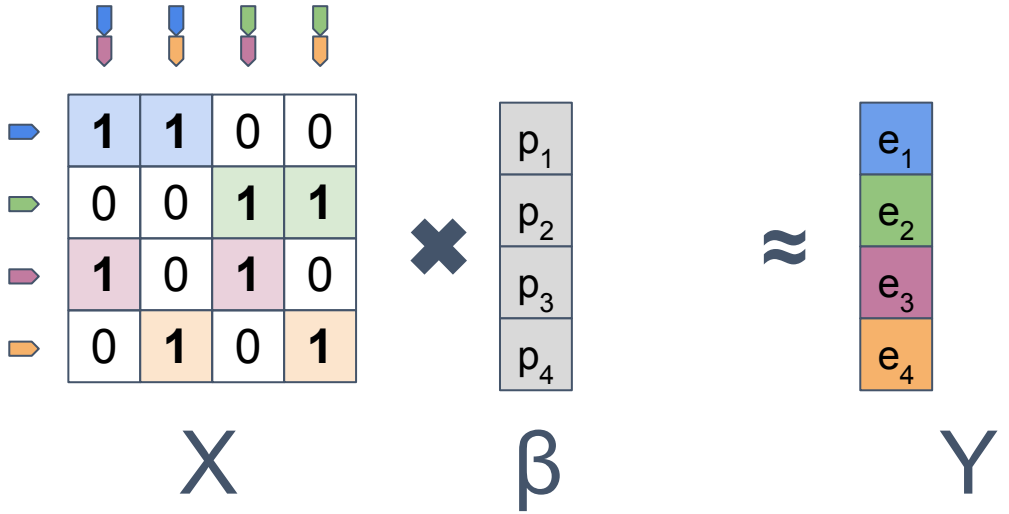
KEY IDEA: A single "junction" is the minimum unit of deconvolution





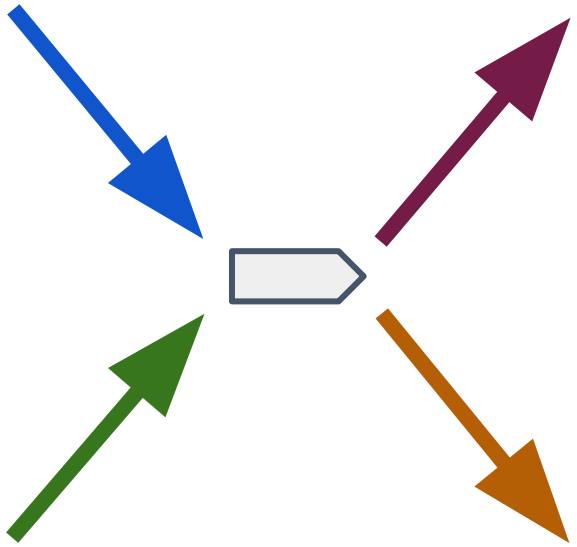
Focus on just one junction at a time
Quantify local paths

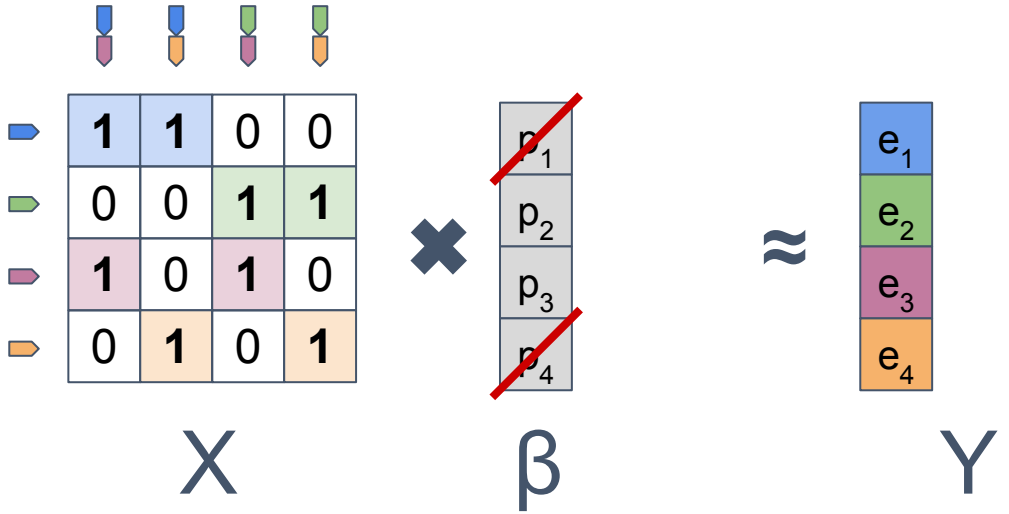




Linear regression

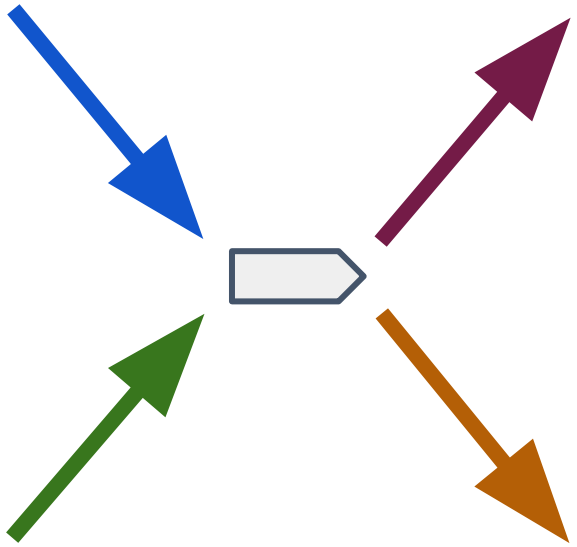
Focus on just one junction at a time
Select (and quantify) local paths

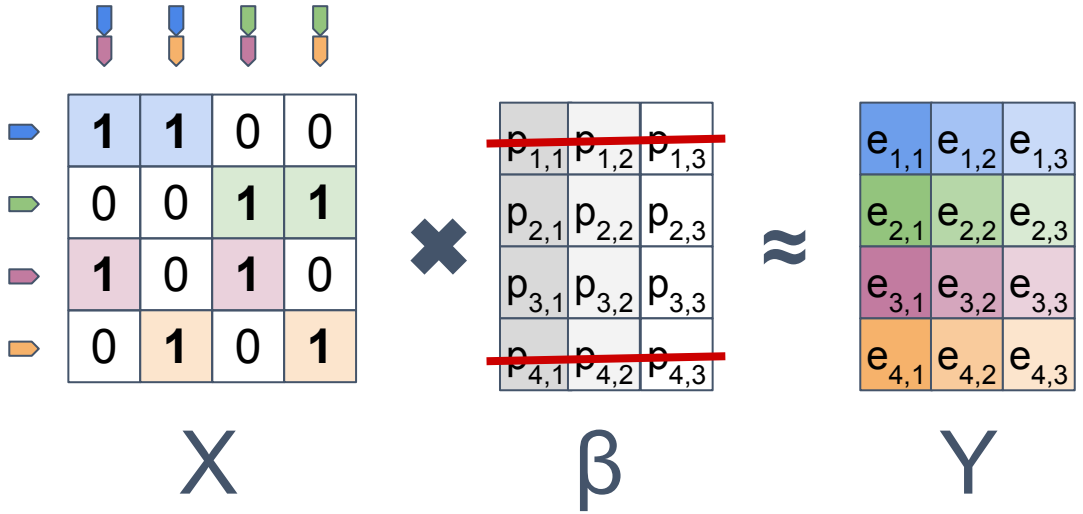




Linear regression
Model selection

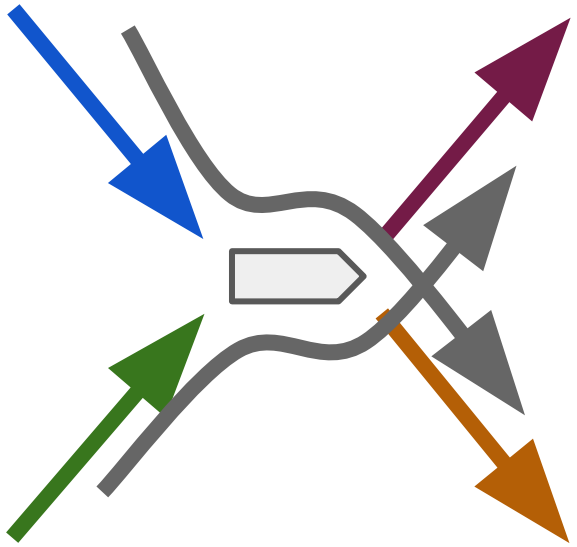
Focus on just one junction at a time
Select (and quantify) local paths





Linear regression
Model selection
Across multiple samples

Drop paths with no depth in any sample

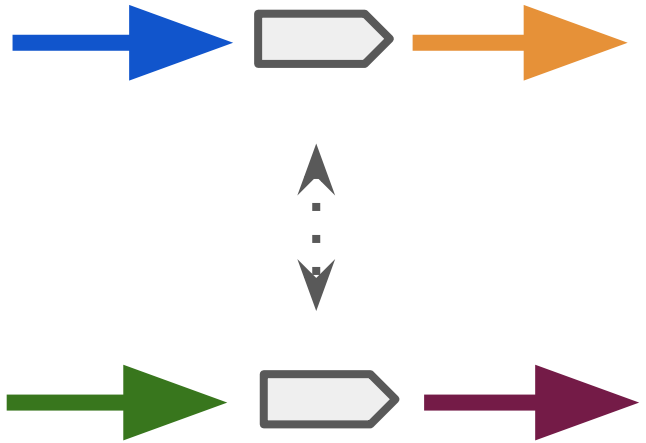


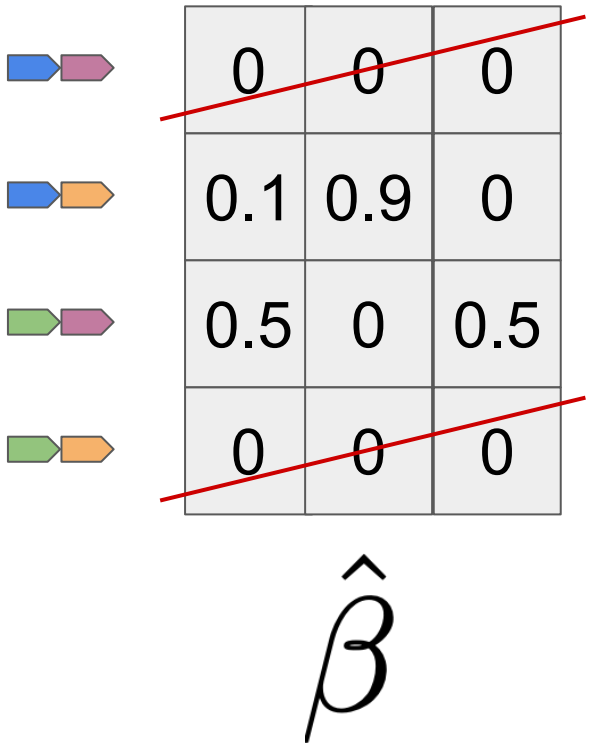
0	0	0
0.1	0.9	0
0.5	0	0.5
0	0	0

$\hat{\beta}$

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

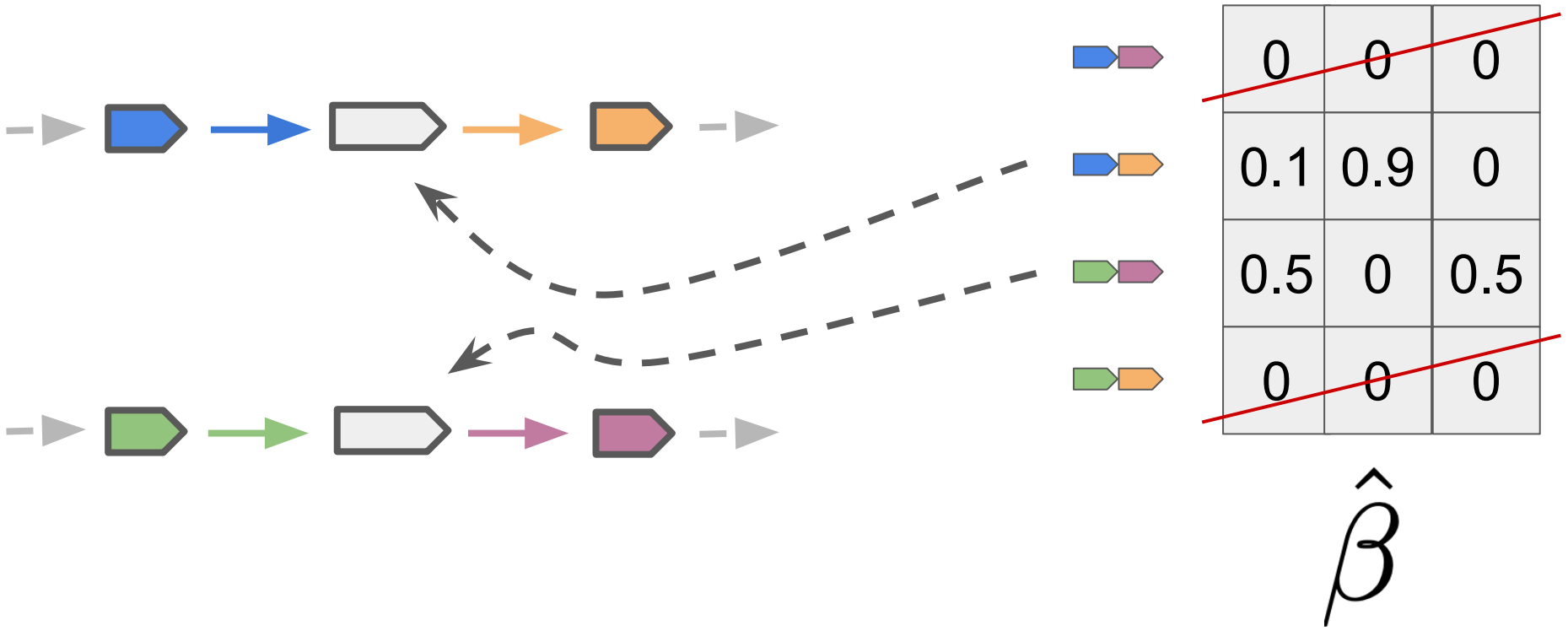
Resolve ambiguity, longer linear sequences





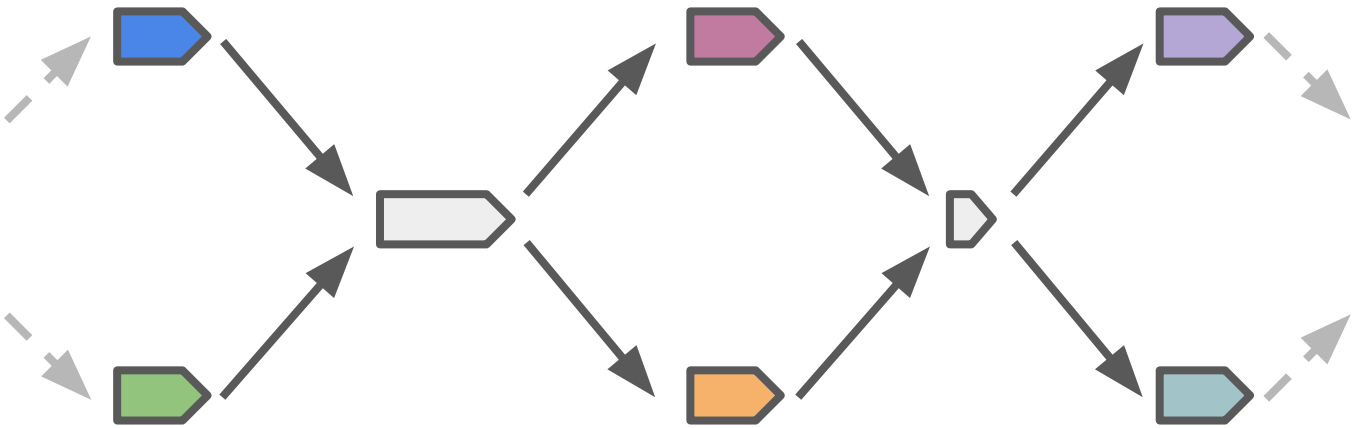
Can "unzip" this unitig into two paths

Resolve ambiguity, longer linear sequences

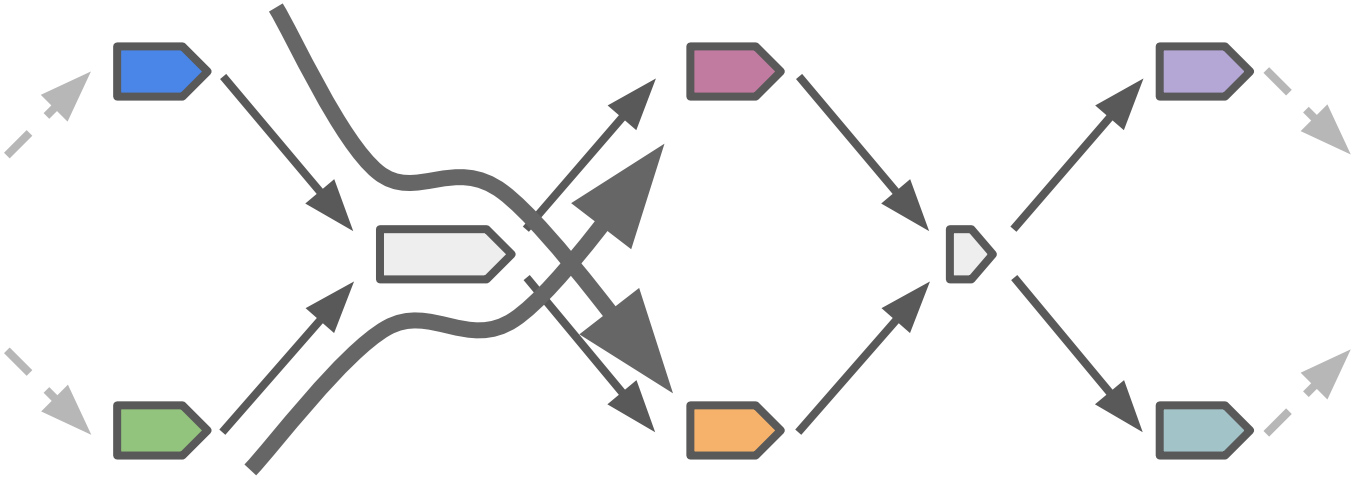


Newly split unitigs already have depths estimated across samples

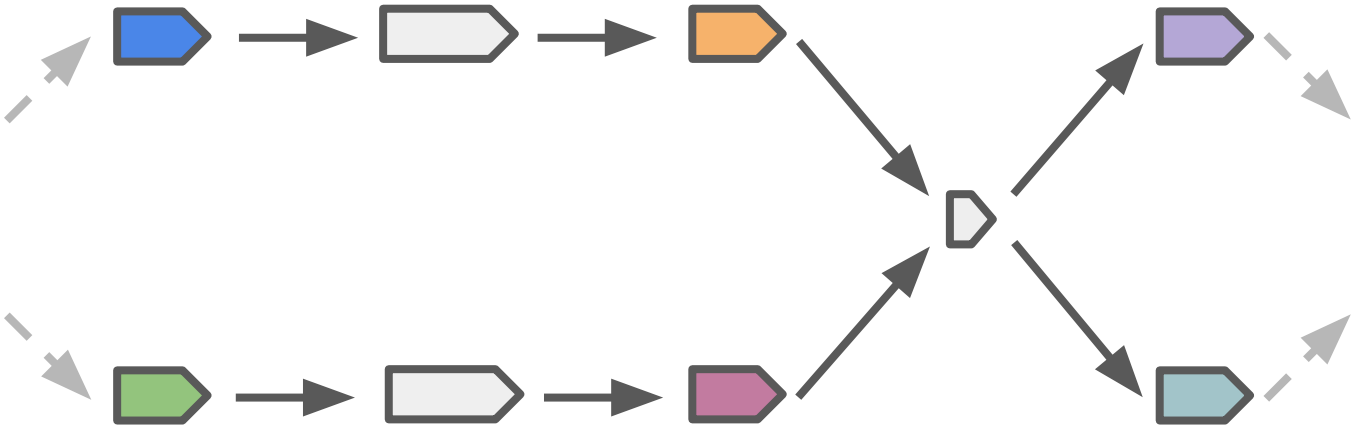
Iteratively unzipping local junctions



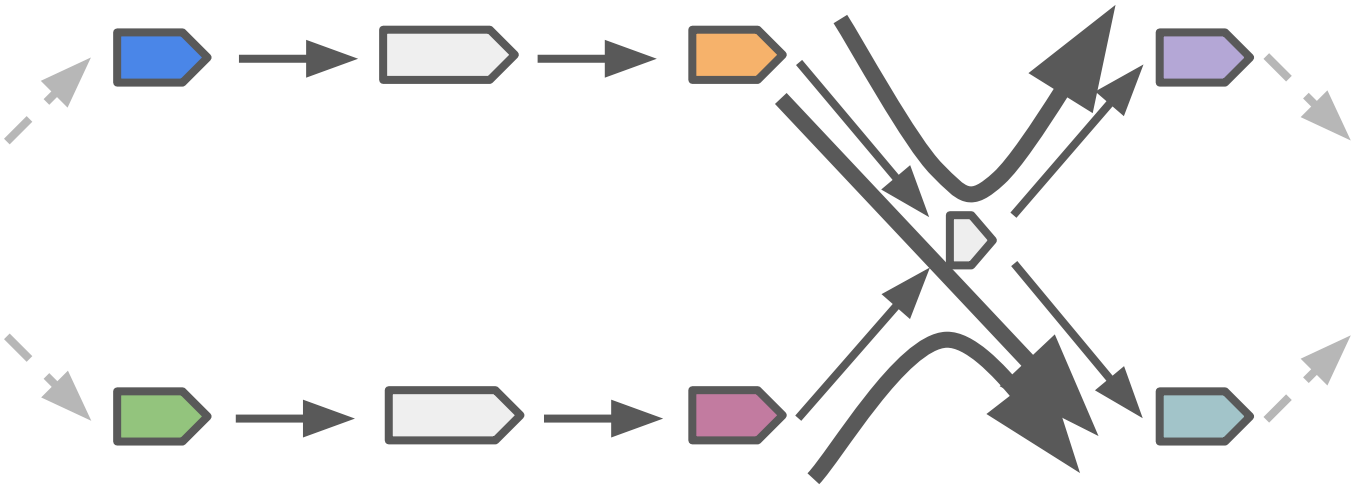
Iteratively unzipping local junctions



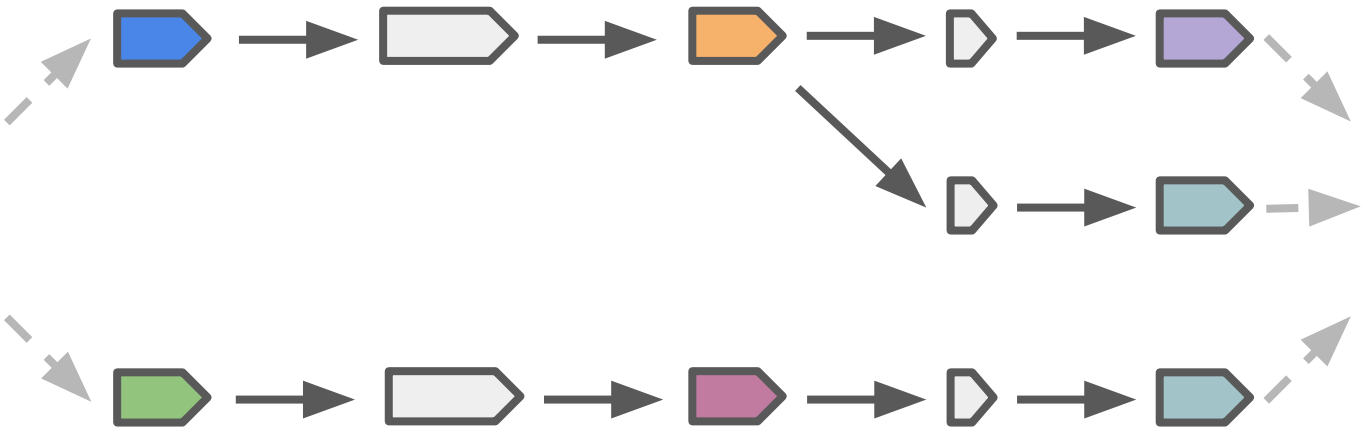
Iteratively unzipping local junctions



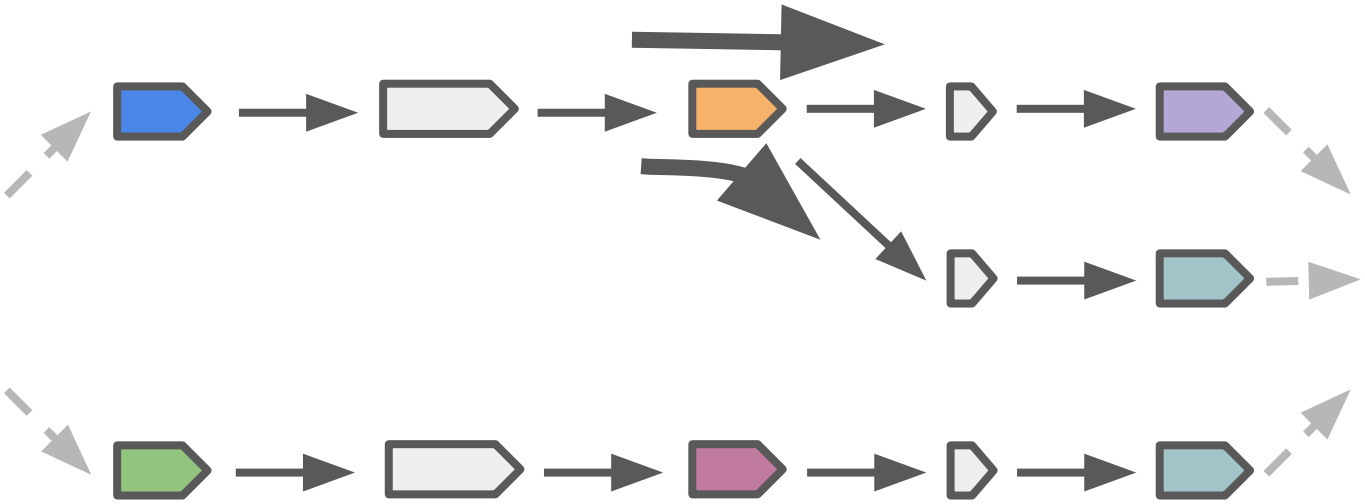
Iteratively unzipping local junctions



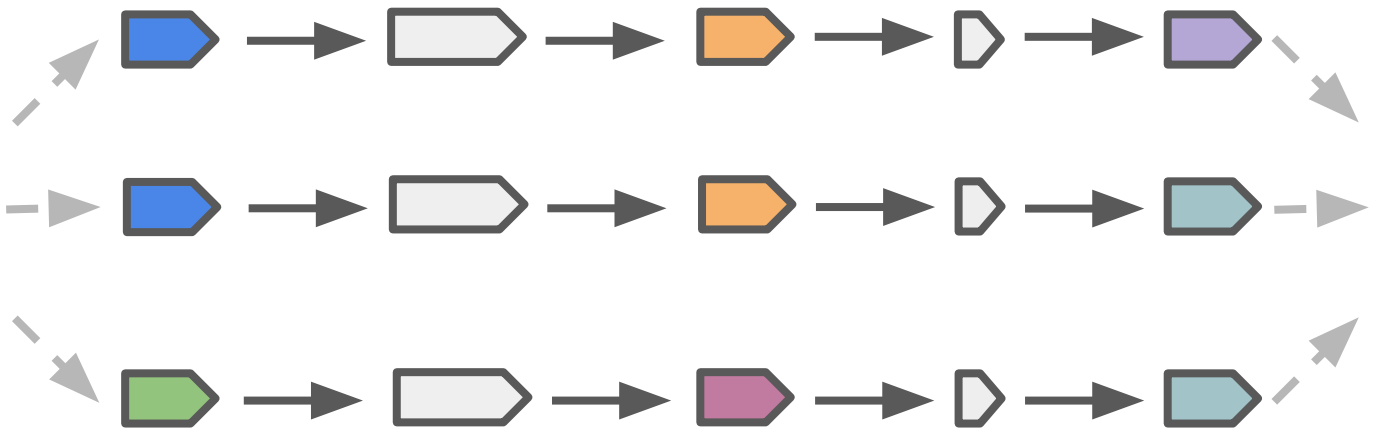
Iteratively unzipping local junctions



Iteratively unzipping local junctions



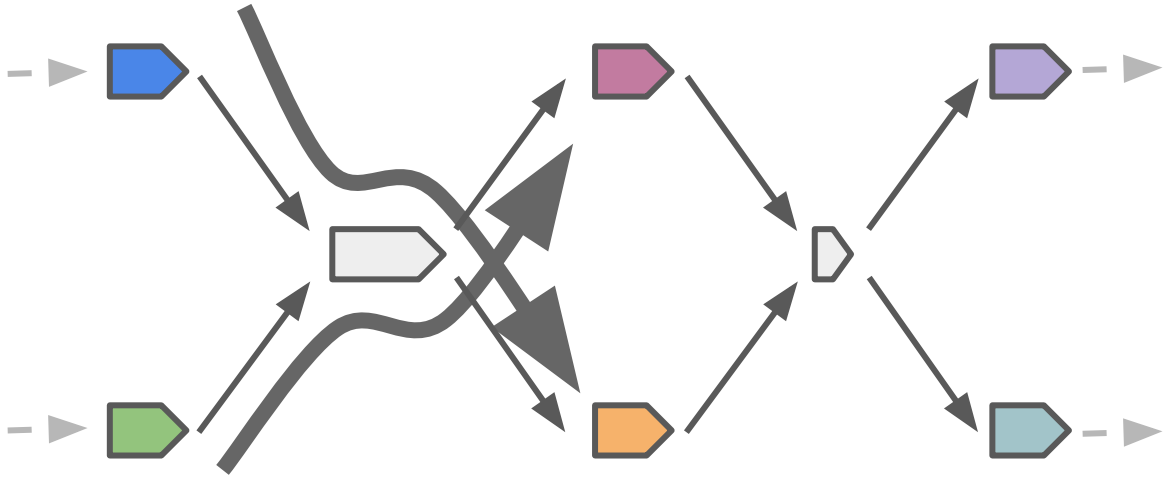
Iteratively unzipping local junctions

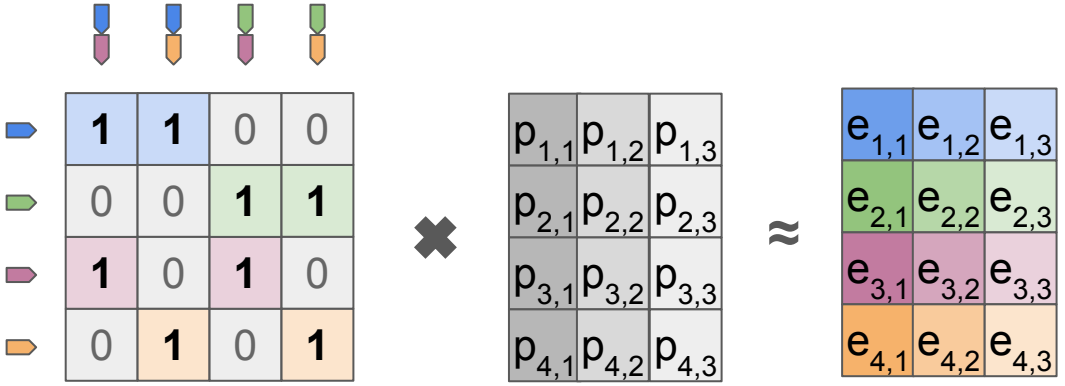


StrainZip

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

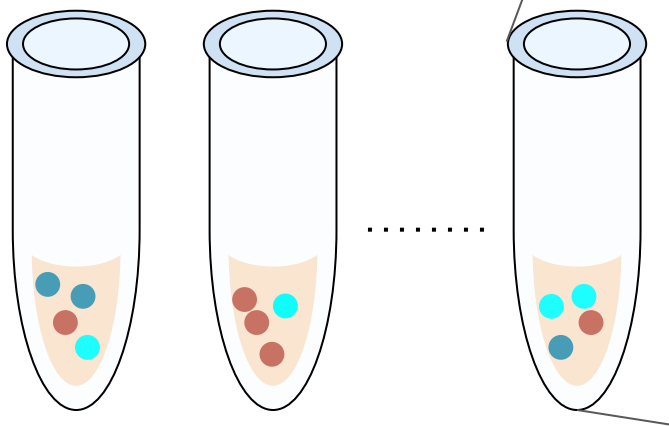
https://github.com/bsmith89/StrainZip

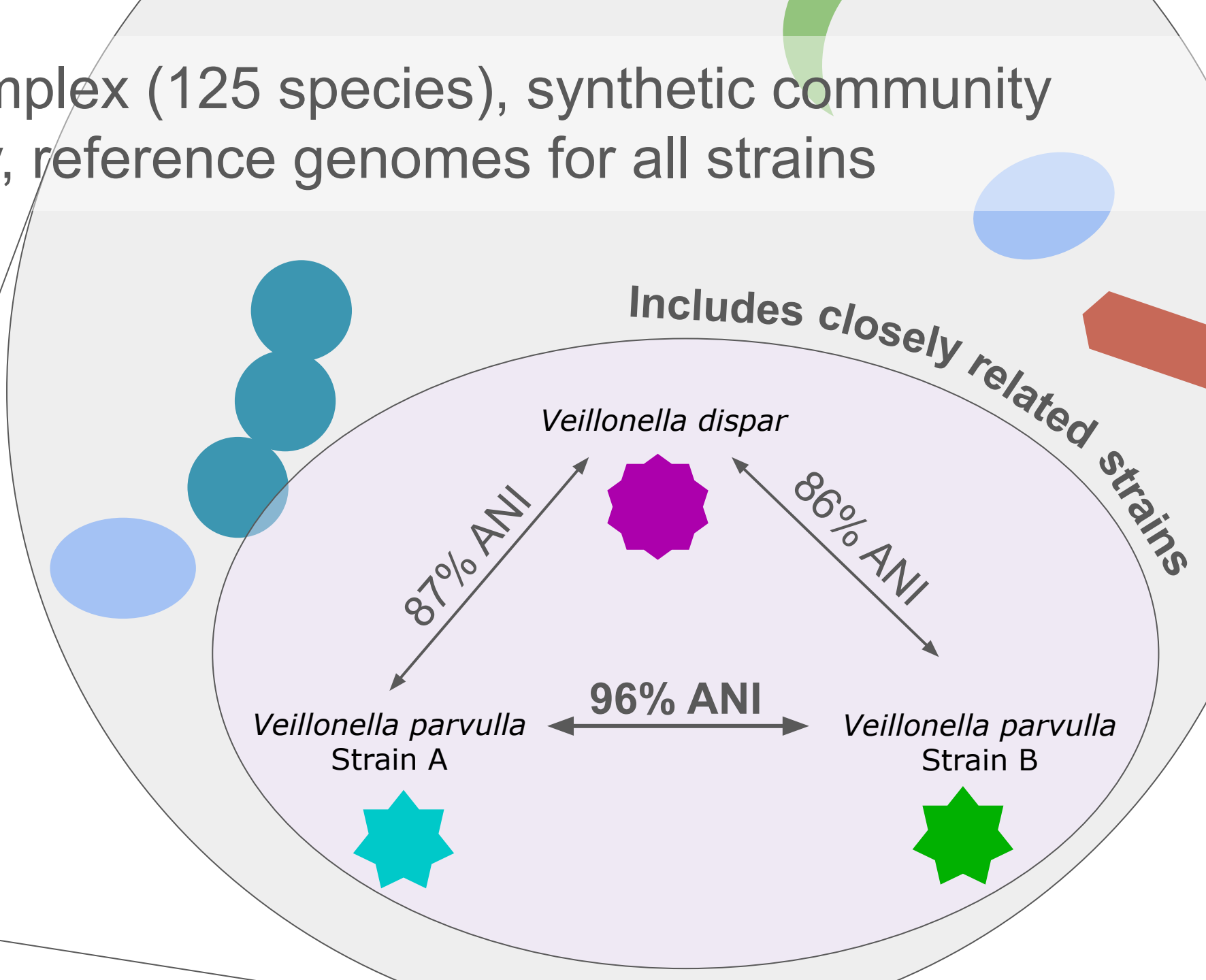


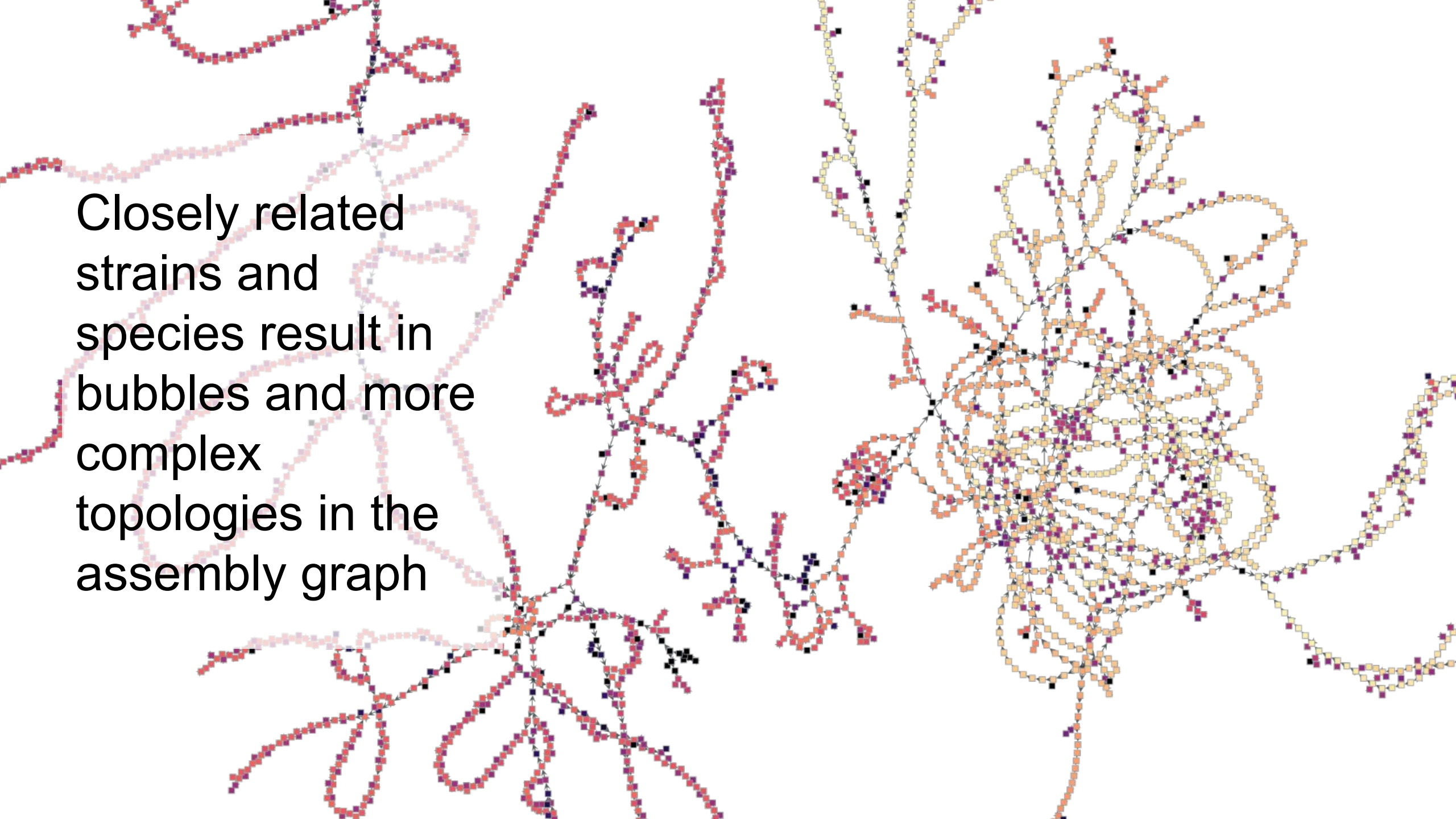


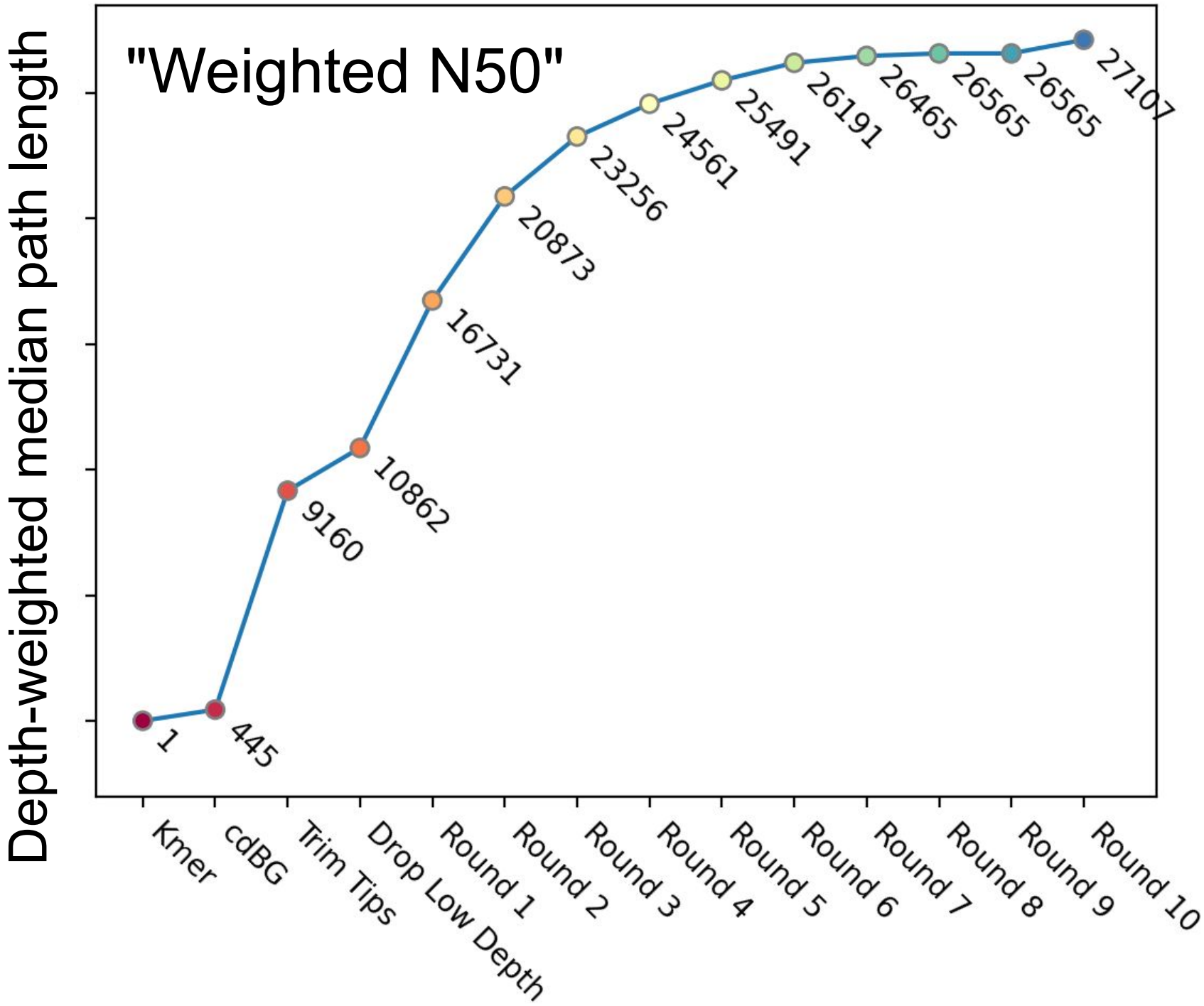
Benchmarking

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



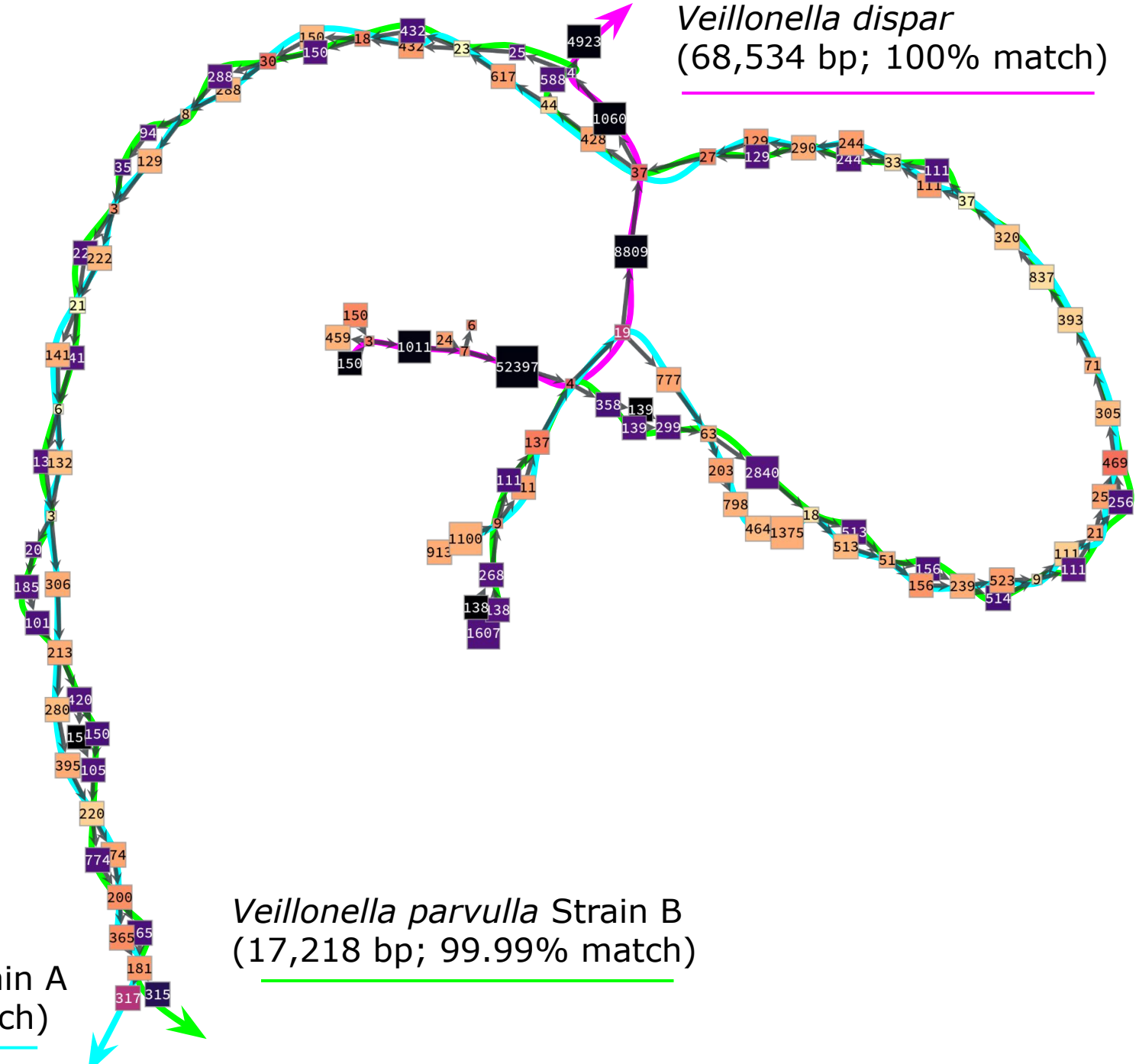
Path lengths increase over successive rounds of deconvolution

Deconvolution recovers longer, strain-specific sequences

…including lower-abundance strains

…and species

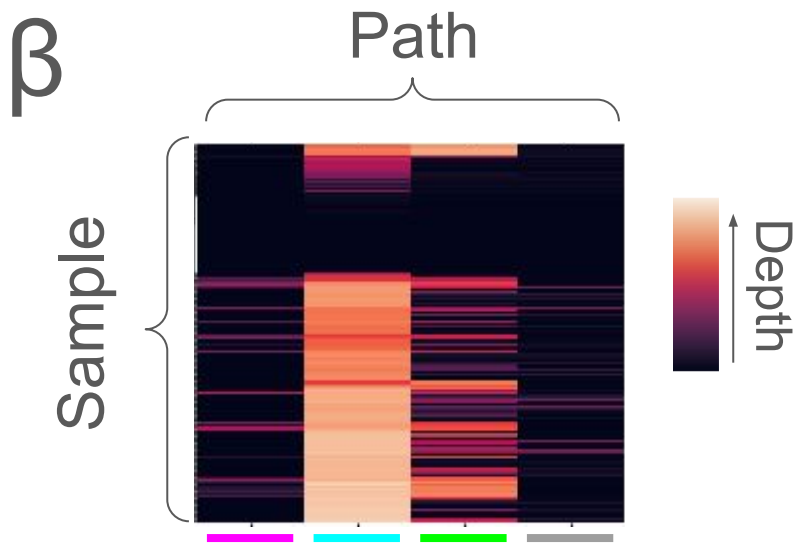
…accurately

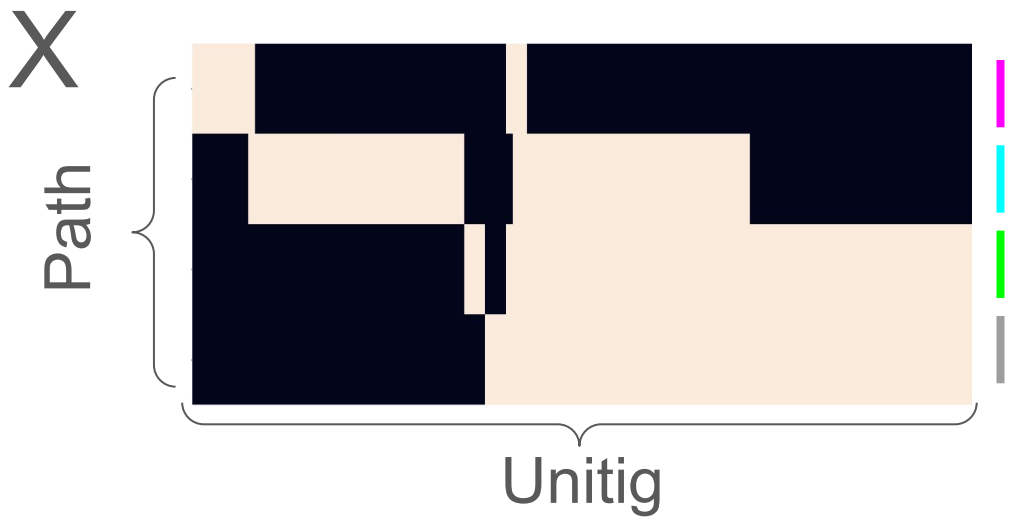


Veillonella dispar
(68,534 bp; 100% match)

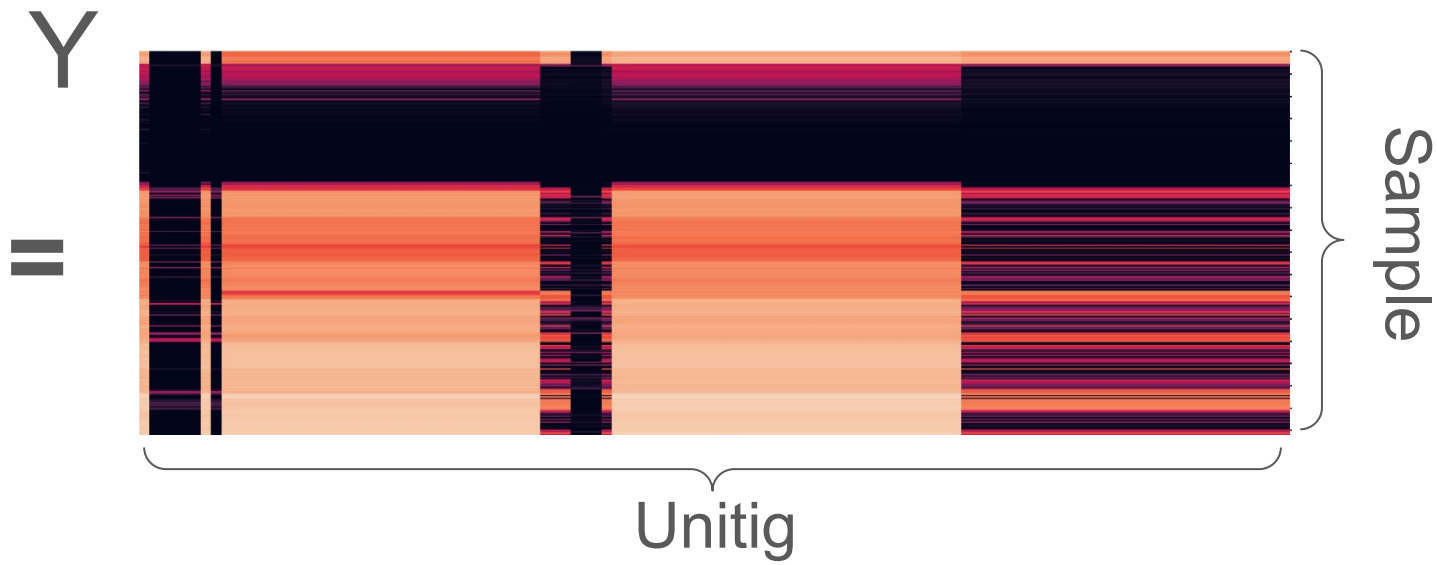
Veillonella parvulla Strain B
(17,218 bp; 99.99% match)

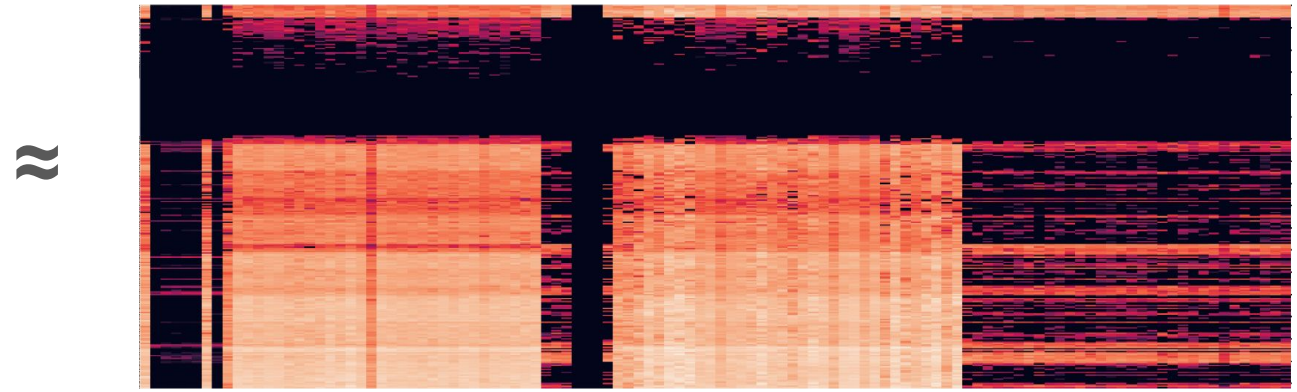
Veillonella parvulla Strain A
(17,229 bp; 100% match)



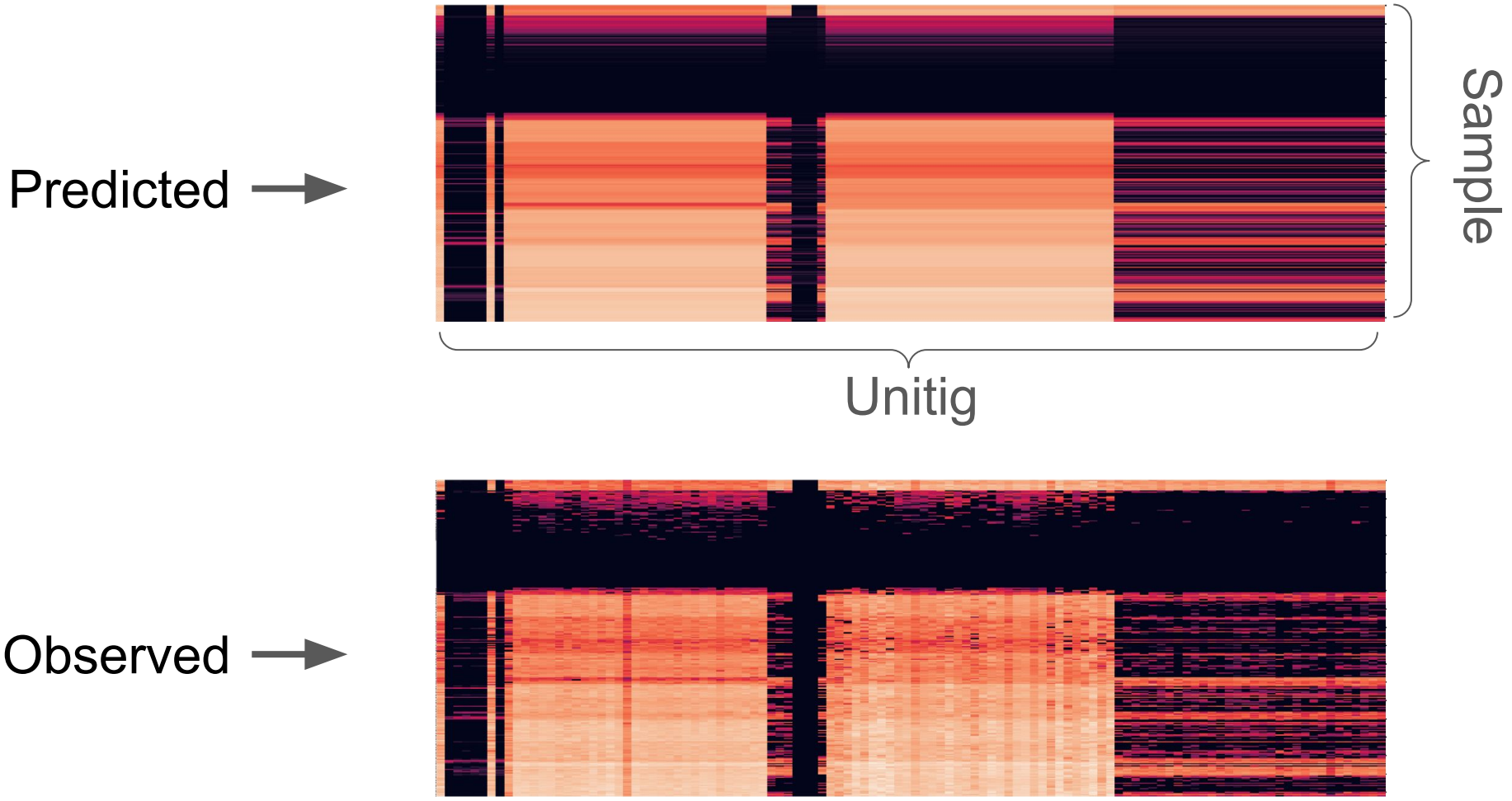


Result: both paths, and path depths across samples (without read mapping)

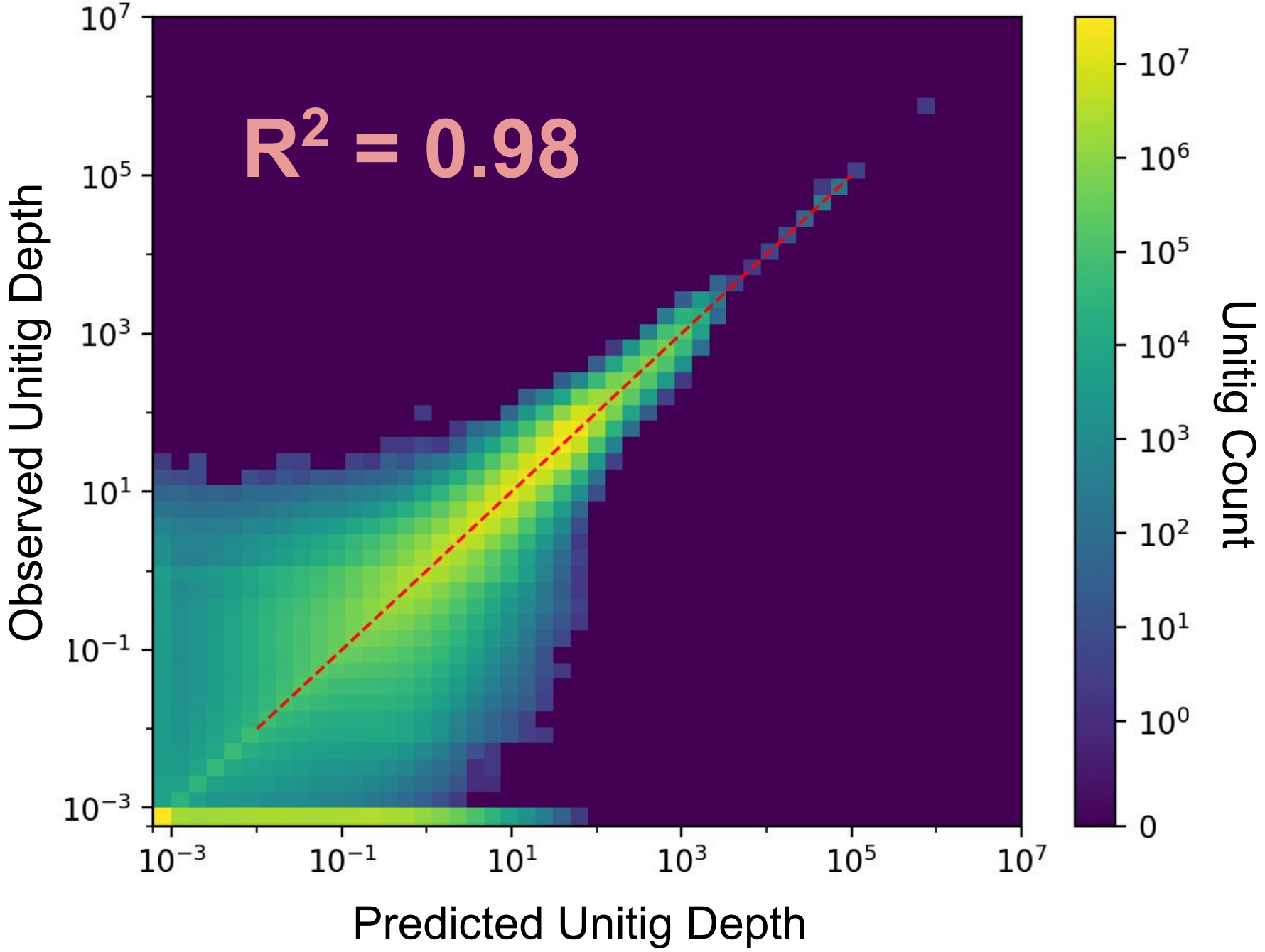




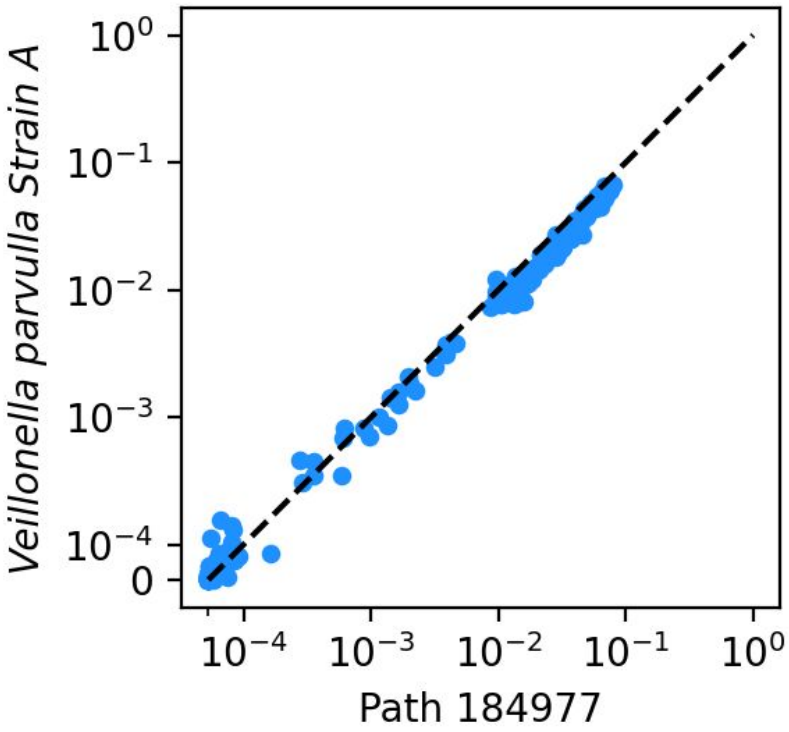
Estimated unitig depths closely match observed depths

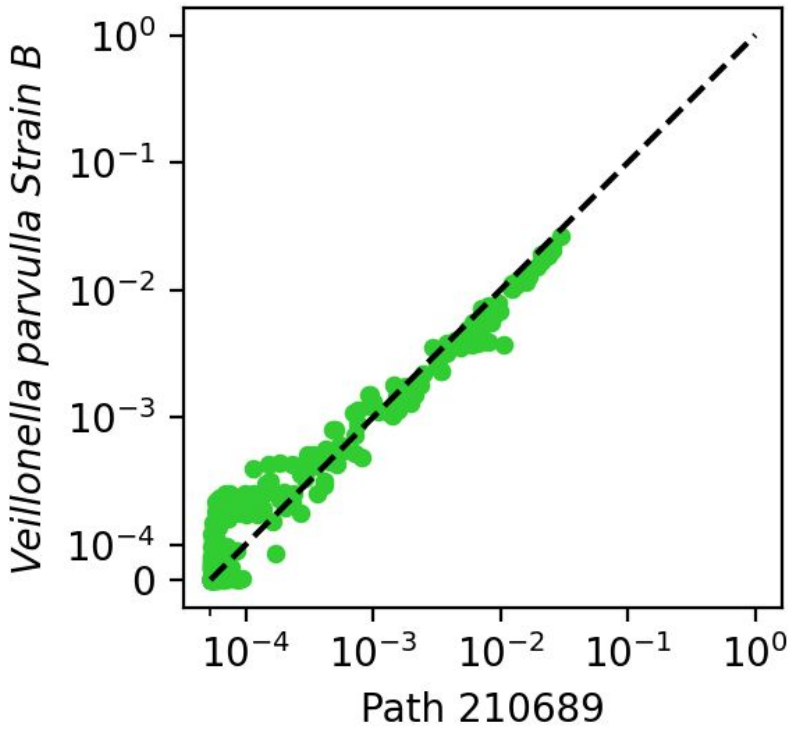


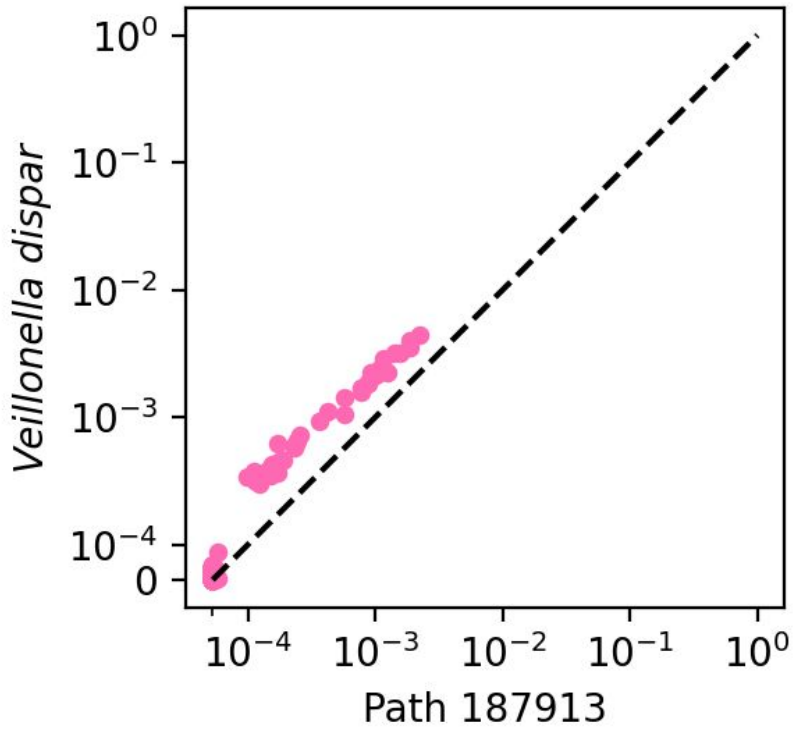
Estimated unitig depths closely match observed depths



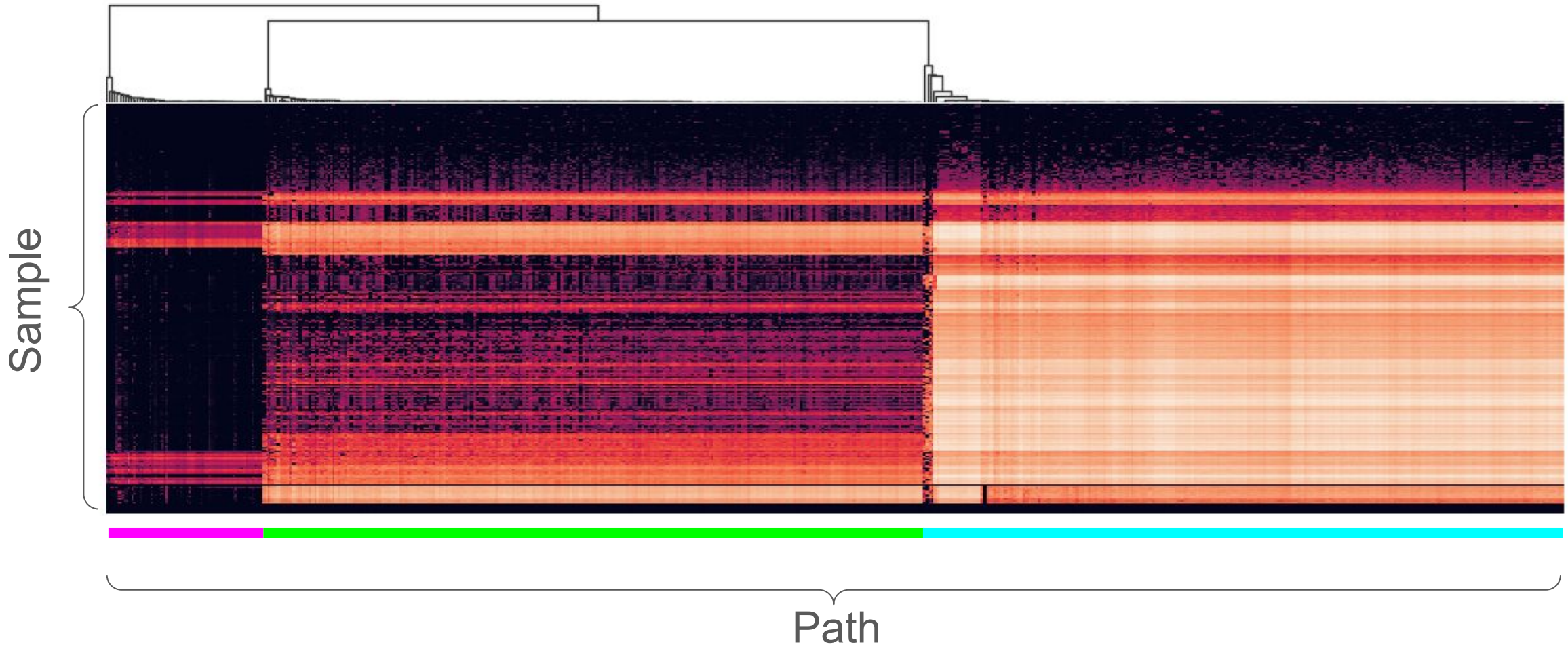
Path depths match reference-based strain depth estima



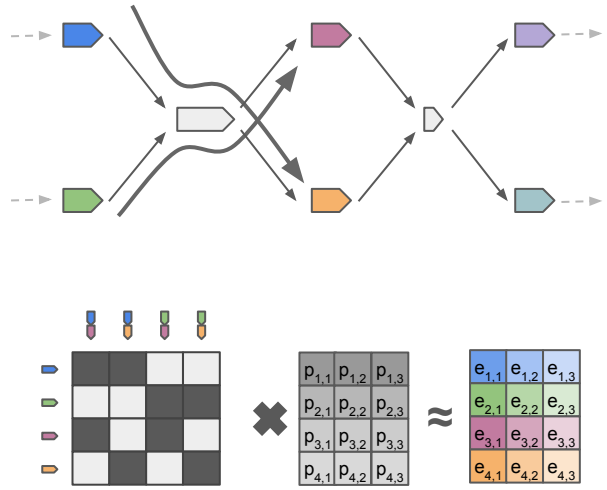




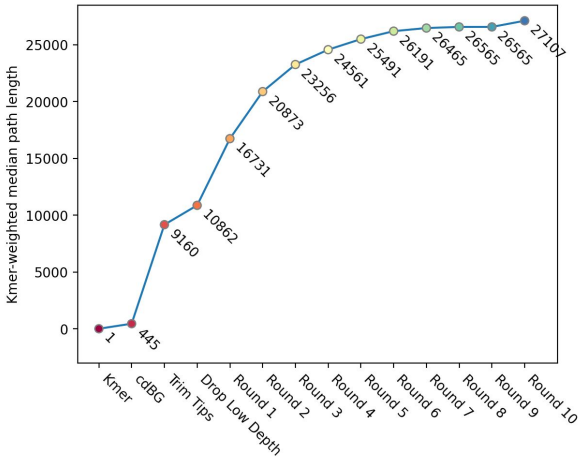
Clustering paths by depth combines multiple sequences from the same strain

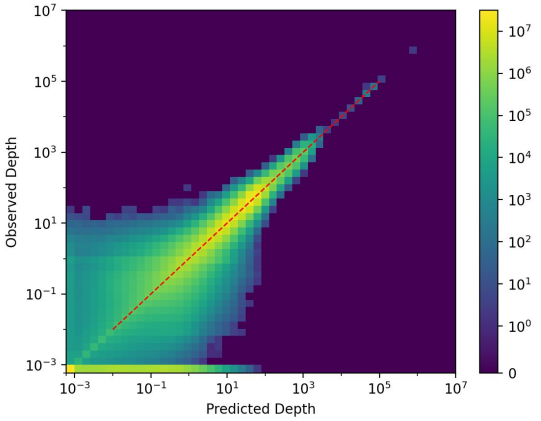


Iterative Junction Deconvolution

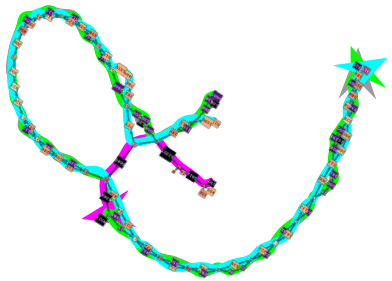


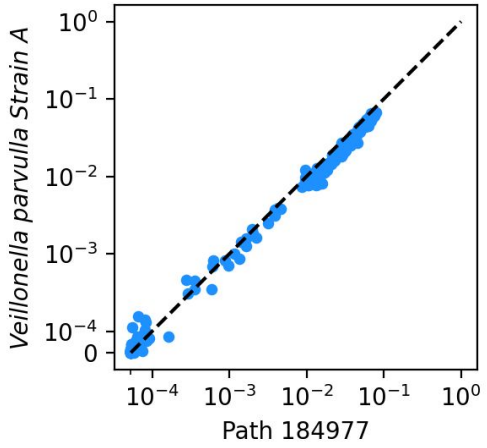
Combines Assembly, Depth Estimation





Recovers Closely Related Genomes





Enables Strain-Resolved Metagenomics