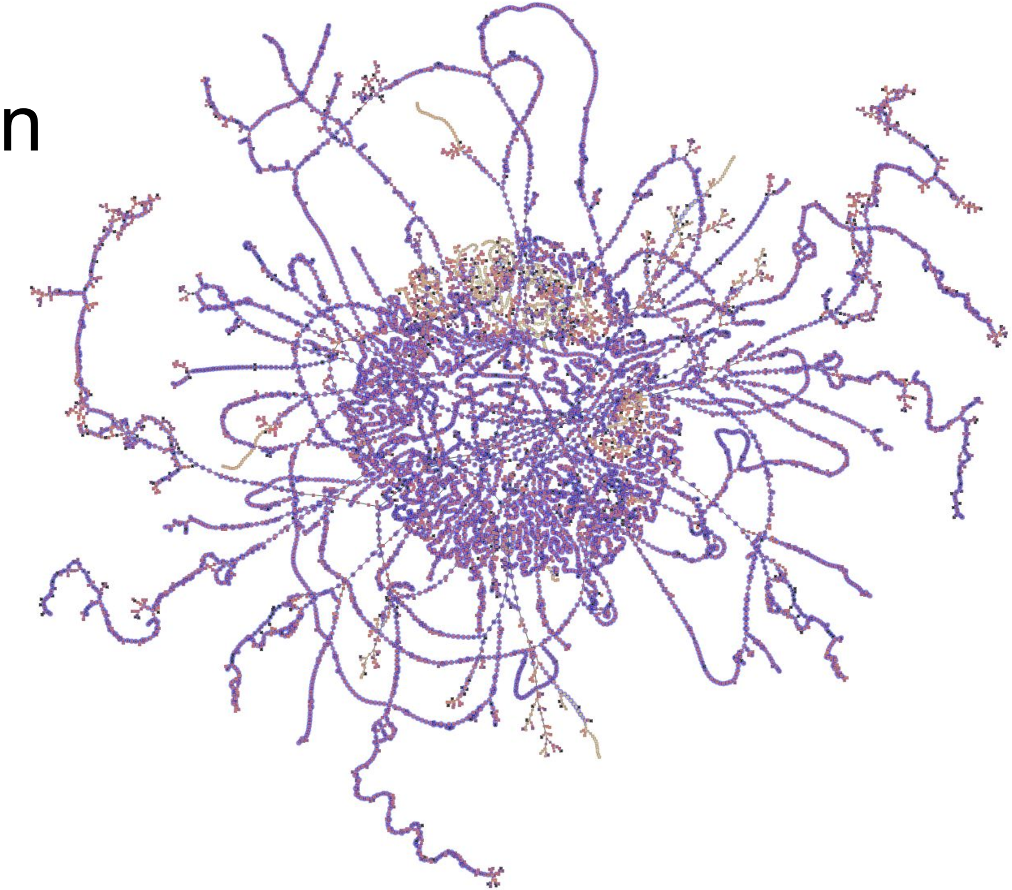
Disentangling depth on the De Bruijn graph:

Combining quantification and *de novo* assembly

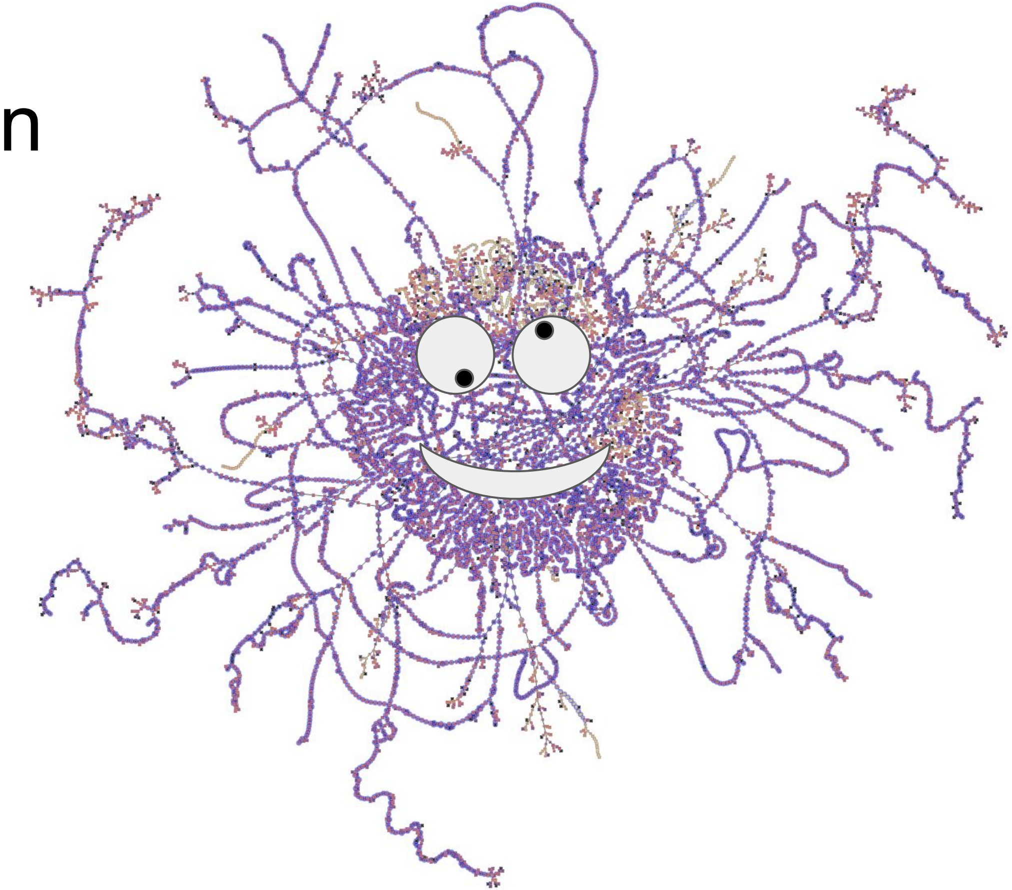
Byron J. Smith
IGGSy
2024-07-04



Disentangling depth on the De Bruijn graph:

Combining quantification and *de novo* assembly

Byron J. Smith
IGGSy
2024-07-04



Acknowledgments / Find Me

Acknowledgments / Find Me

Pollard Lab

- *Katie Pollard*
- Veronika Dubinkina
- et al.

Engelhardt Lab

- *Archit Verma*
- Dylan Cable

GLADSTONE
INSTITUTES

Twitter: @ByronJSmith

Email: byron.smith
@gladstone.ucsf.edu

Outline of the presentation

- **Motivation:**

  Profiling the microbiome with metagenomics

- **Method:**

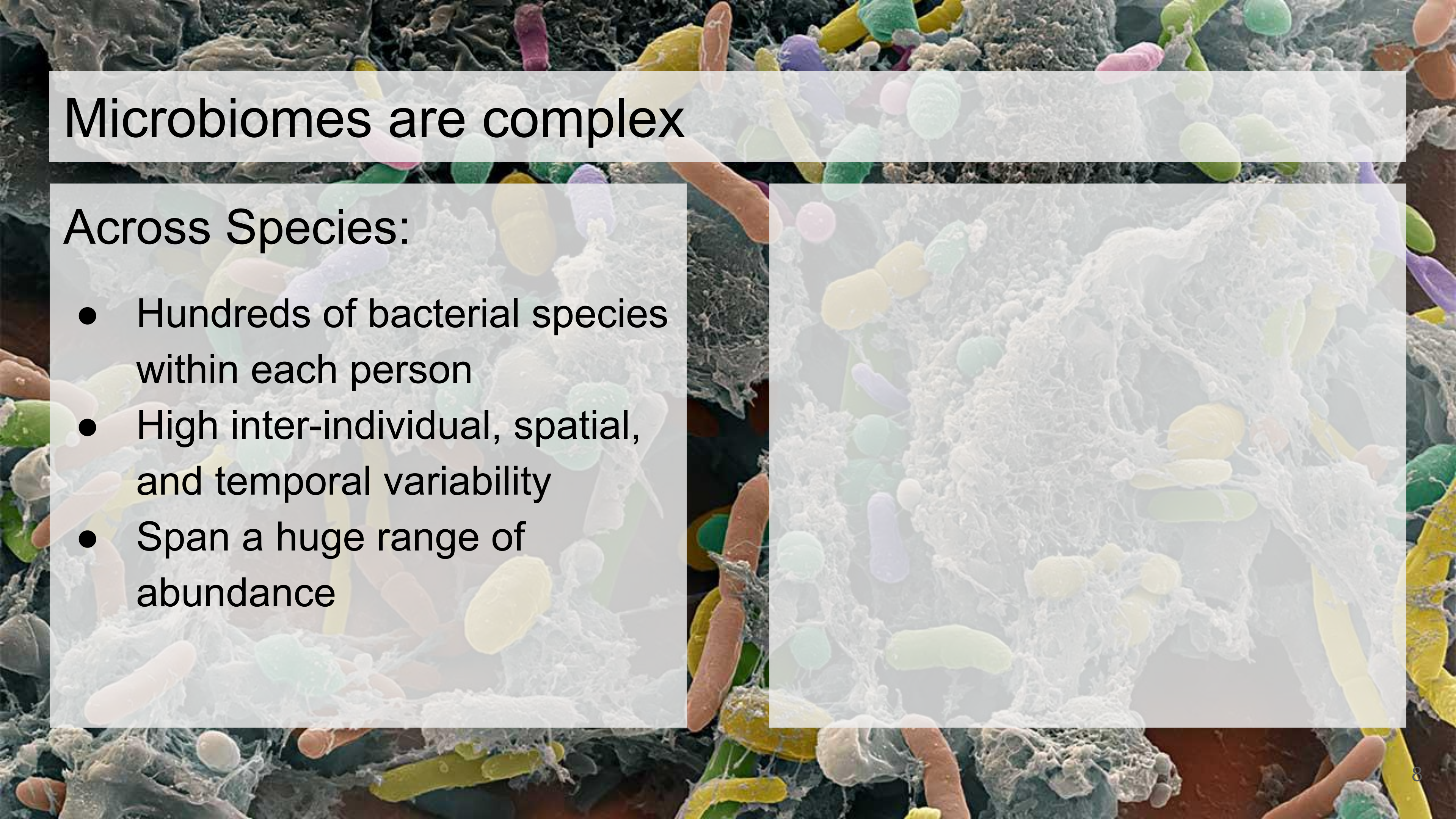
  Assembly graph deconvolution with StrainZip

- **Demonstration:**

  Application to a complex, ground-truthed dataset

Motivation

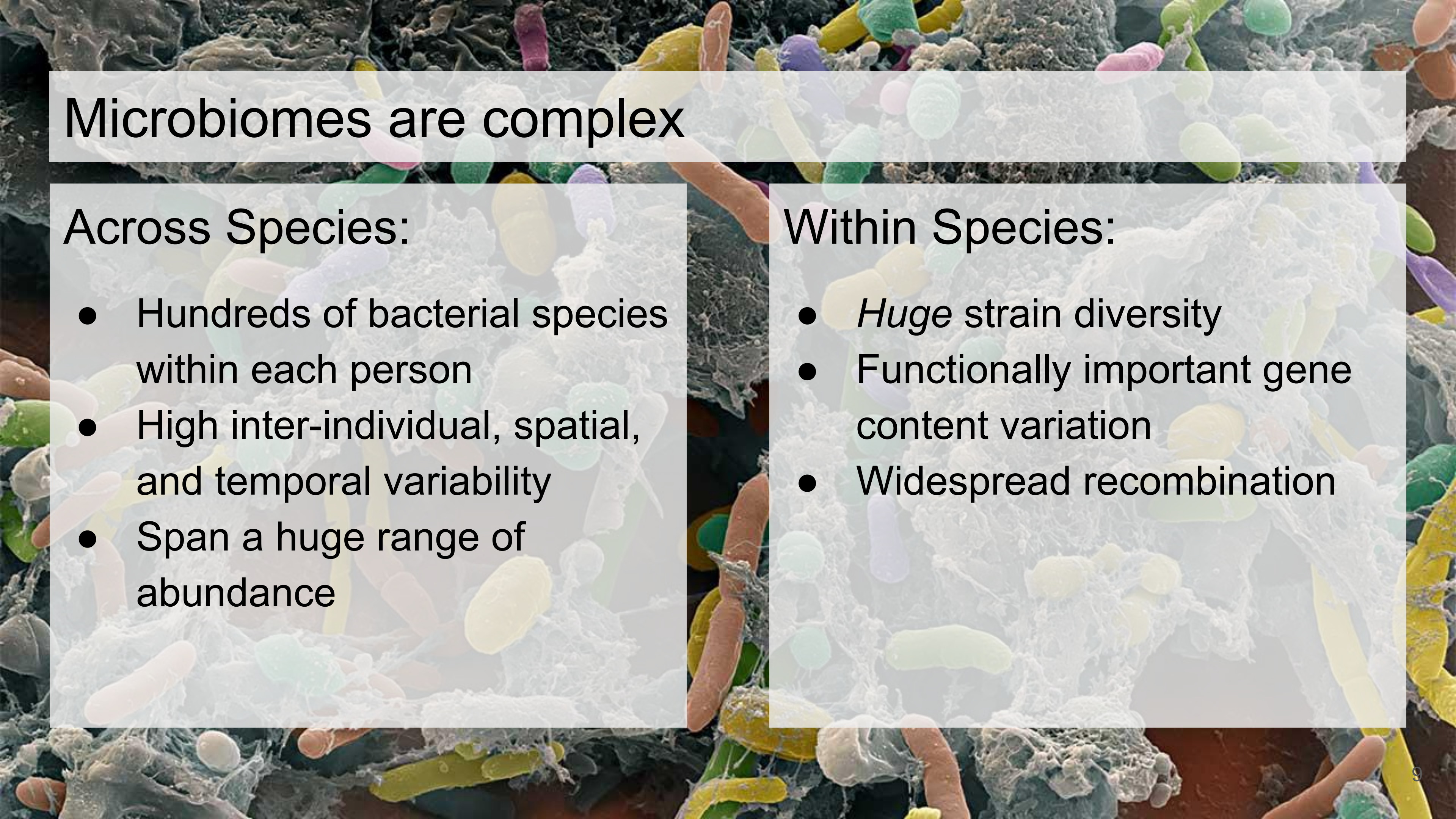
Microbiomes are complex



Microbiomes are complex

Across Species:

- Hundreds of bacterial species within each person
- High inter-individual, spatial, and temporal variability
- Span a huge range of abundance



Microbiomes are complex

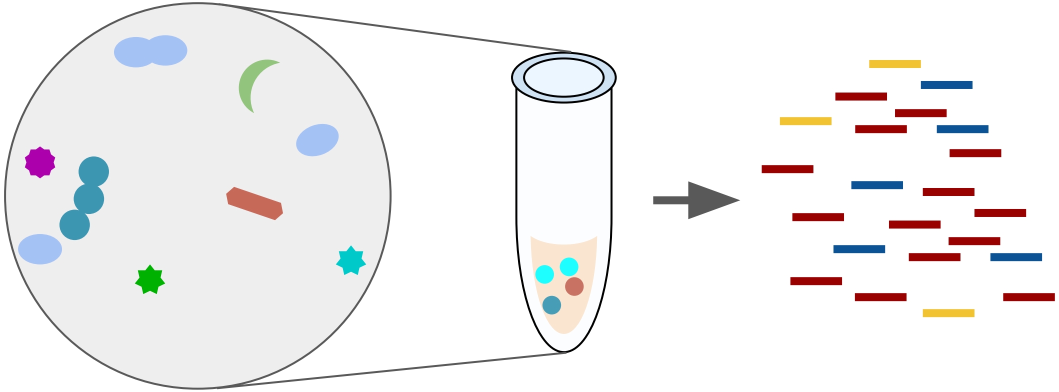
Across Species:

- Hundreds of bacterial species within each person
- High inter-individual, spatial, and temporal variability
- Span a huge range of abundance

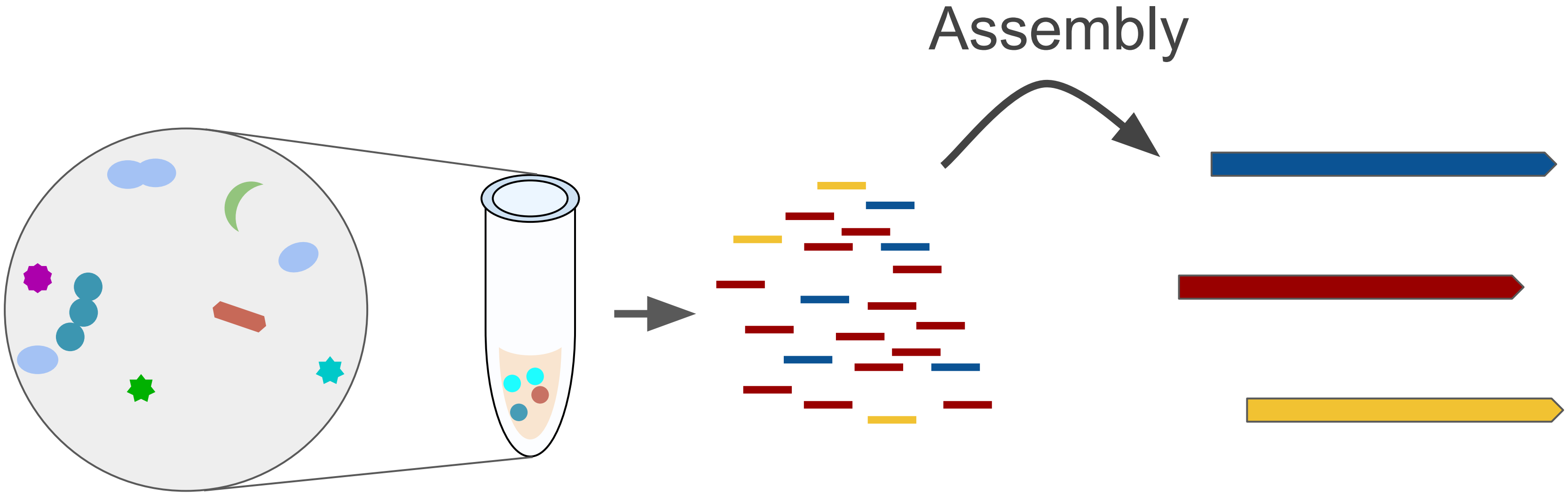
Within Species:

- *Huge* strain diversity
- Functionally important gene content variation
- Widespread recombination

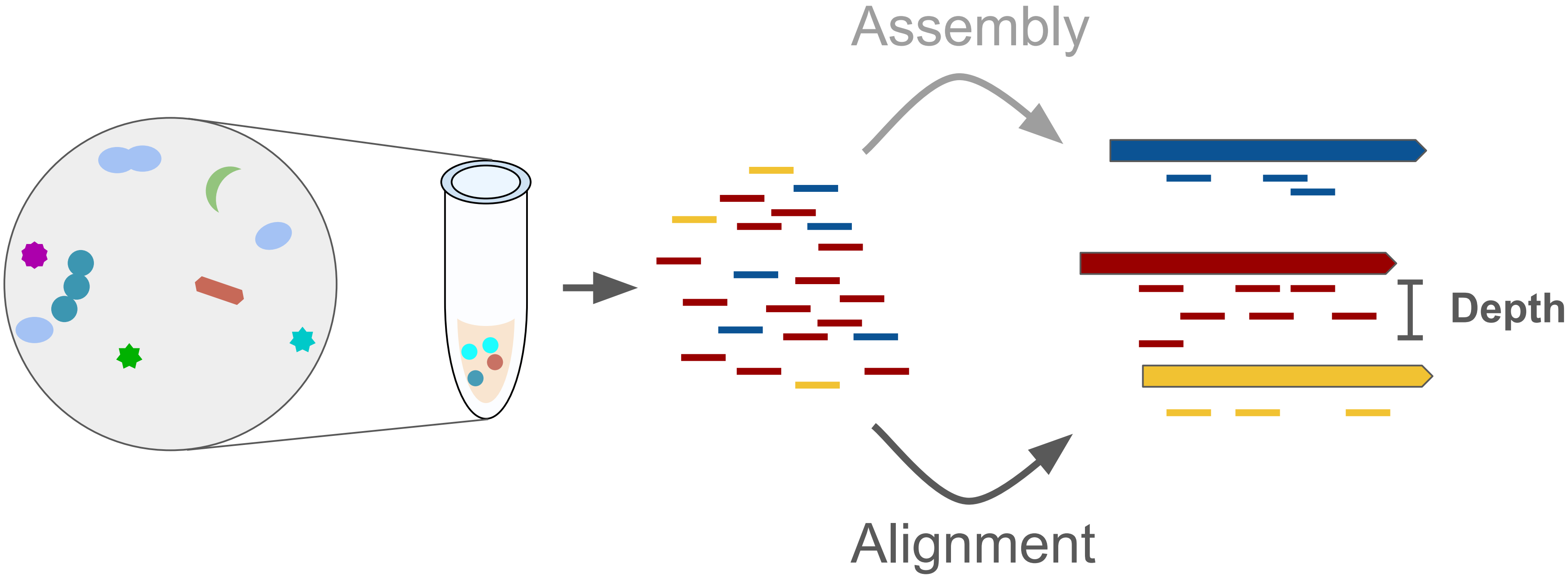
Metagenomics enables modern microbiome science



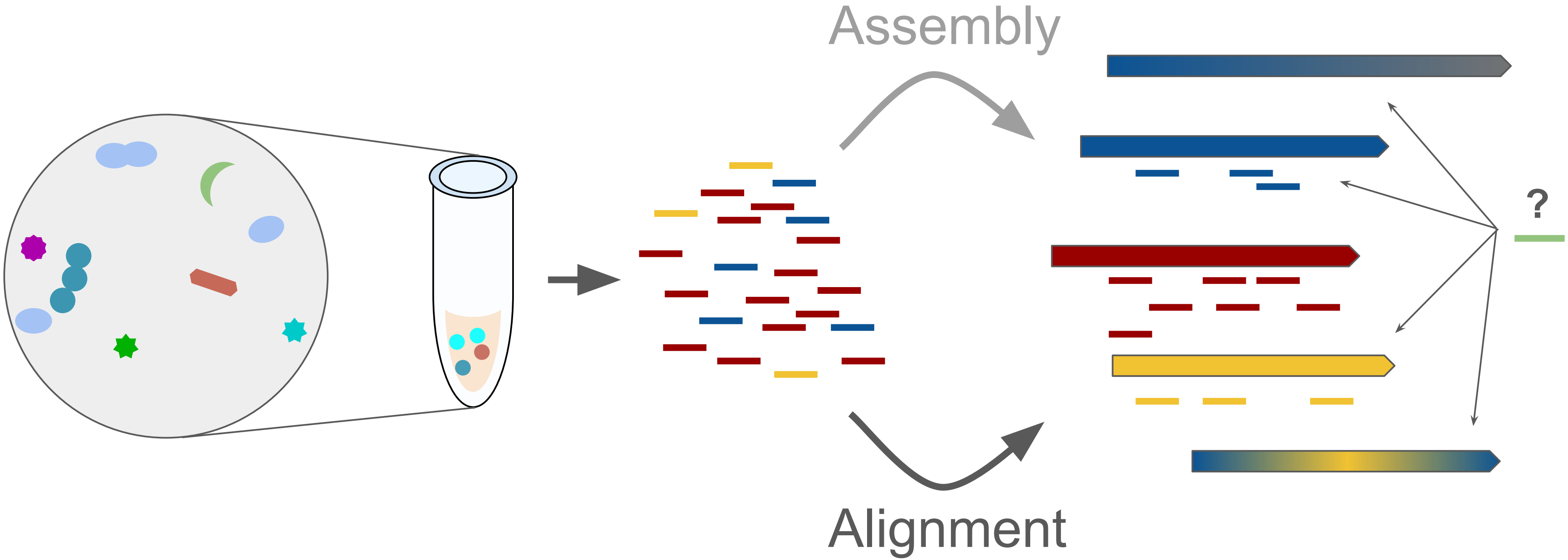
Metagenomics enables modern microbiome science

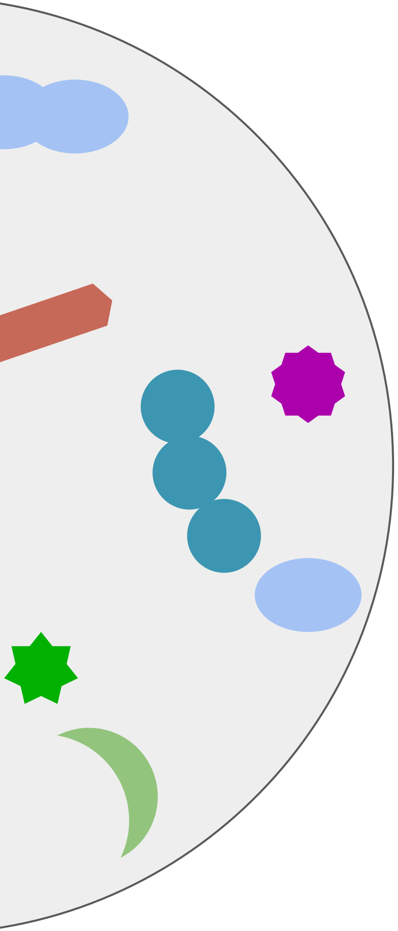


Assembly and depth quantification are complementary



Closely related sequences are a major challenge for alignment





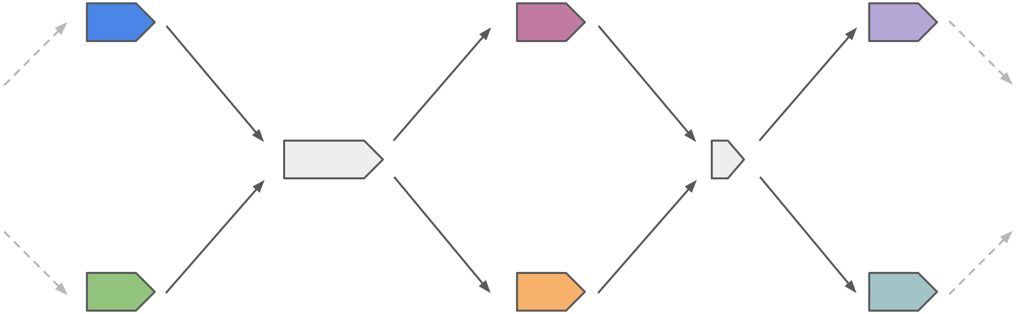
Goal:

Reference-free depth estimation among related sequences in metagenomes

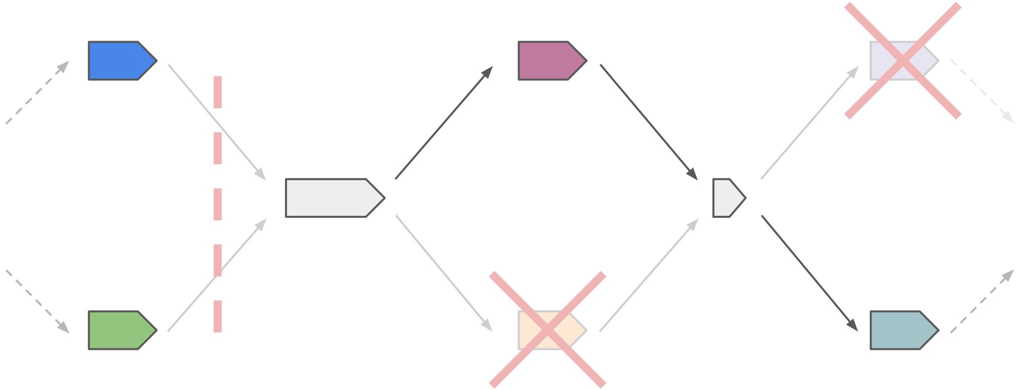
Method

Closely related sequences are a major challenge for both metagenomic assembly and alignment

Complex graph structure

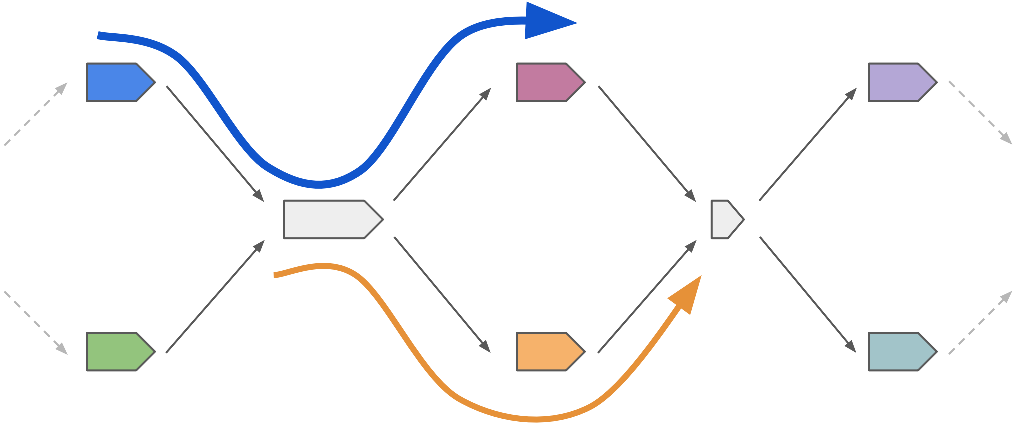


Closely related sequences are a major challenge for alignment



Complex graph structure leads to low-quality assembly

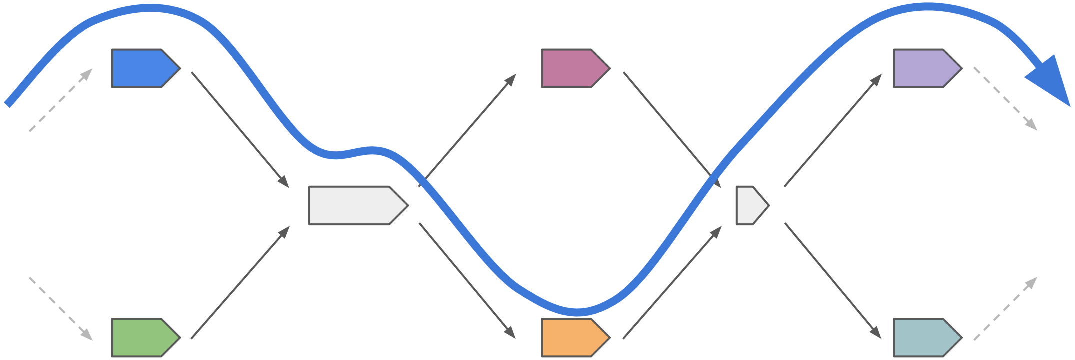
Closely related sequences are a major challenge for alignment



Complex graph structure leads to low-quality assembly

Graph-pangenome approaches account for this variability better than linear references

Closely related sequences are a major challenge for alignment

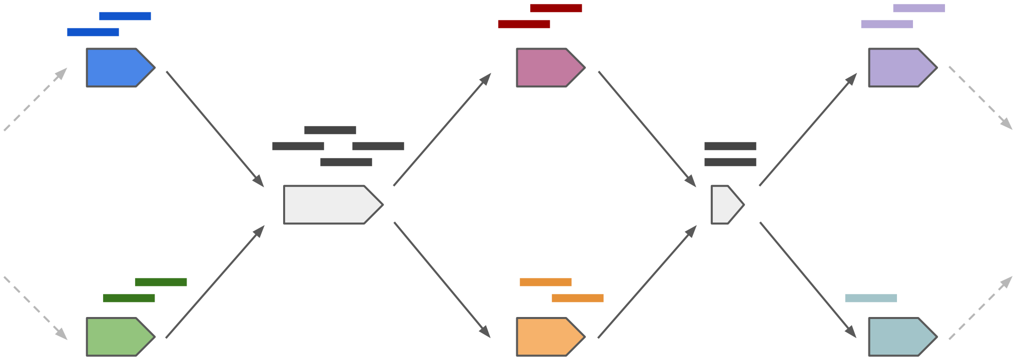


Complex graph structure leads to low-quality assembly

Graph-pangenome approaches account for this variability better than linear references

But long reads too expensive for profiling multiple samples

Closely related sequences are a major challenge for alignment



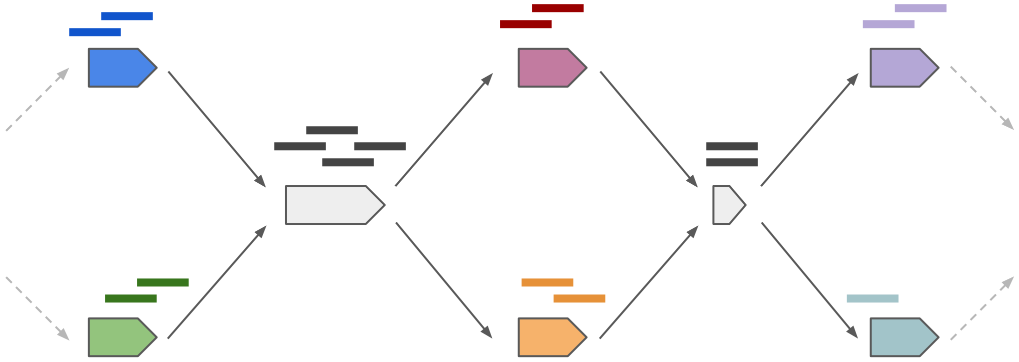
Complex graph structure leads to low-quality assembly

Graph-pangenome approaches account for this variability better than linear references

But long reads too expensive for profiling multiple samples

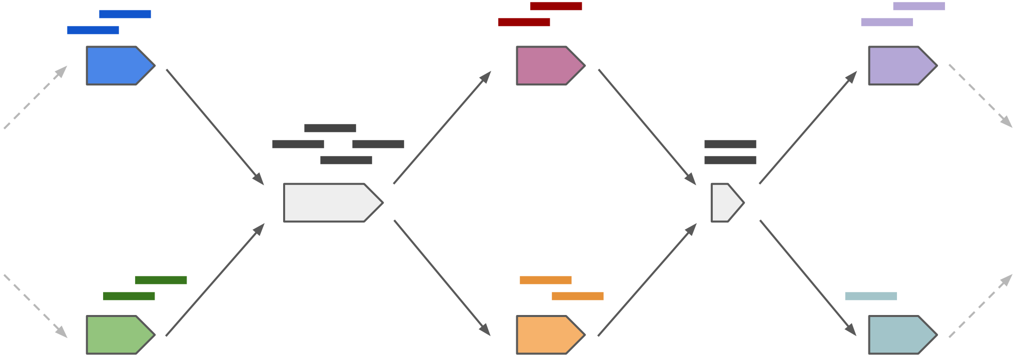
And short-reads are inherently ambiguous

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



$$\sum_{p} x_{pk}\beta_p \approx Y_k$$

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



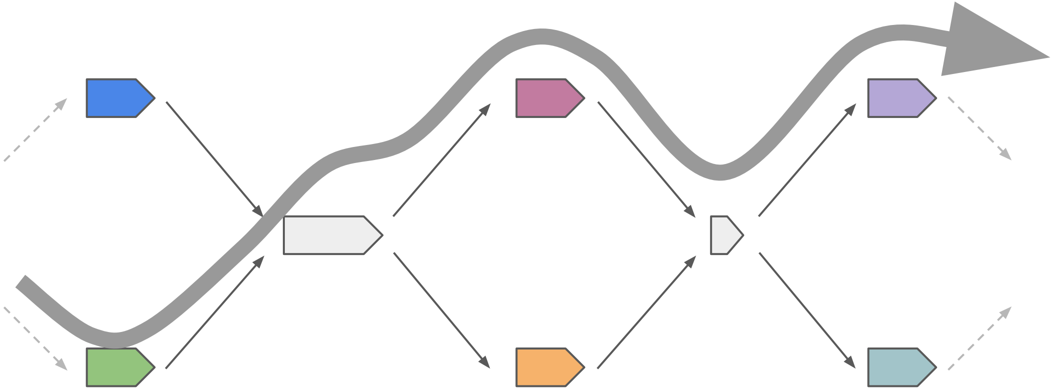
Path depths (unknown)

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



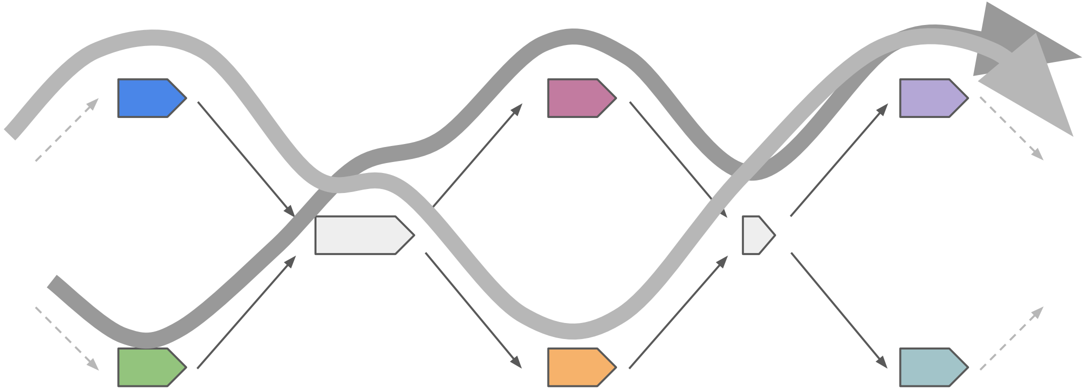
Path depths (unknown)

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



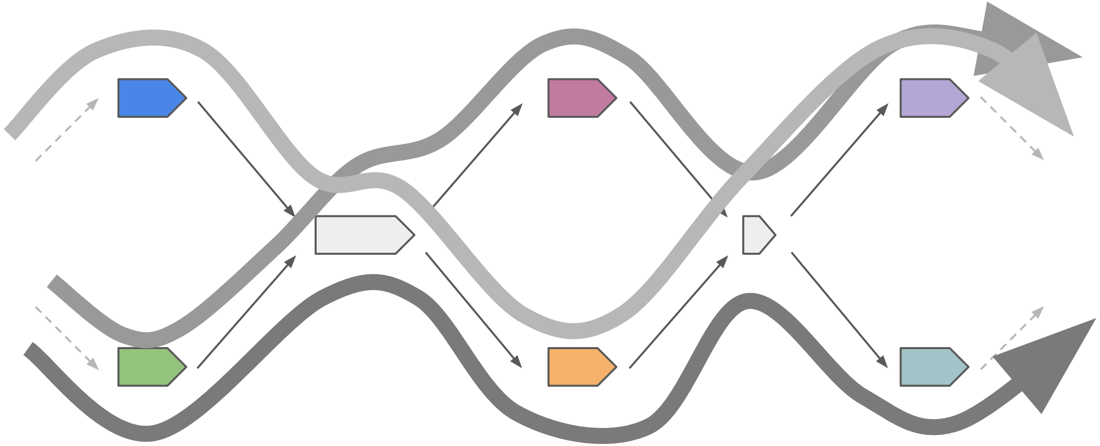
Path depths (unknown)

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



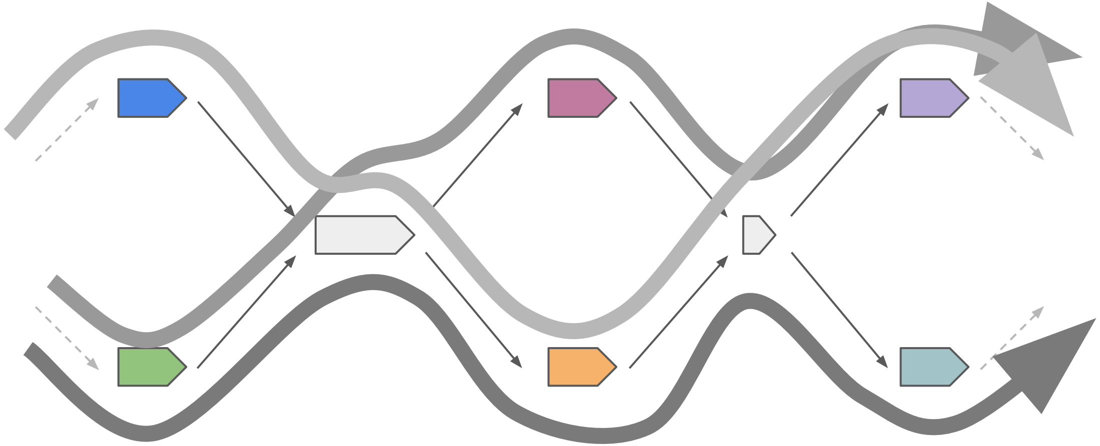
Path depths (unknown)

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



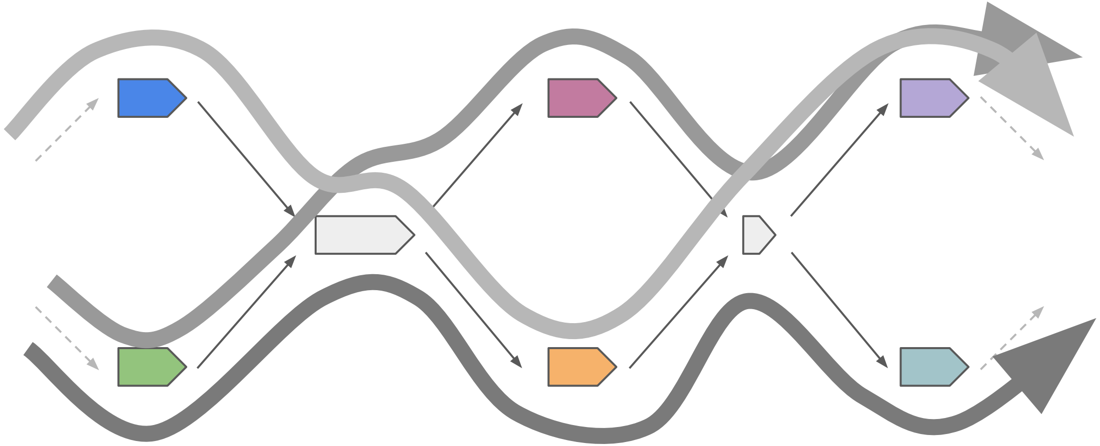
Path depths (unknown)

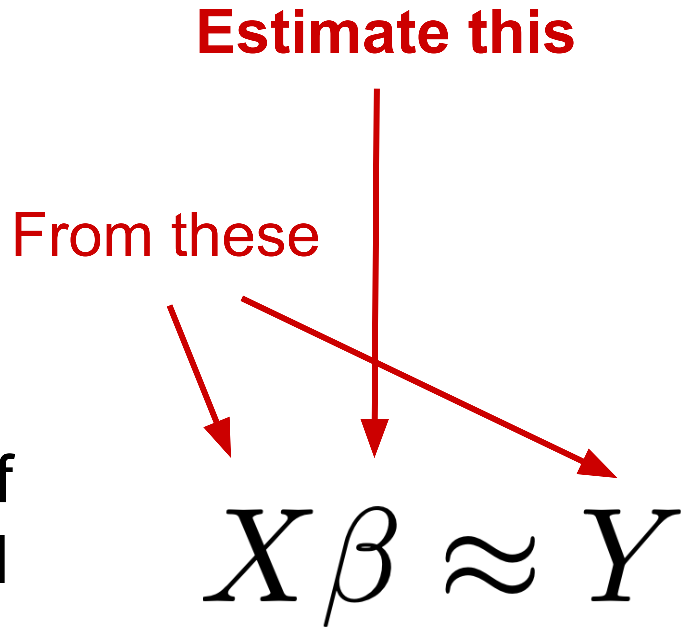
Path encoding

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



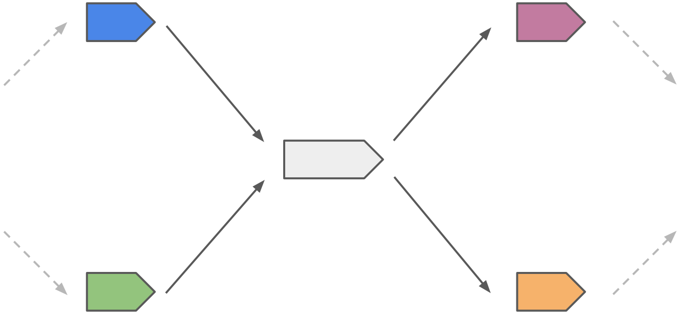

Estimate this

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

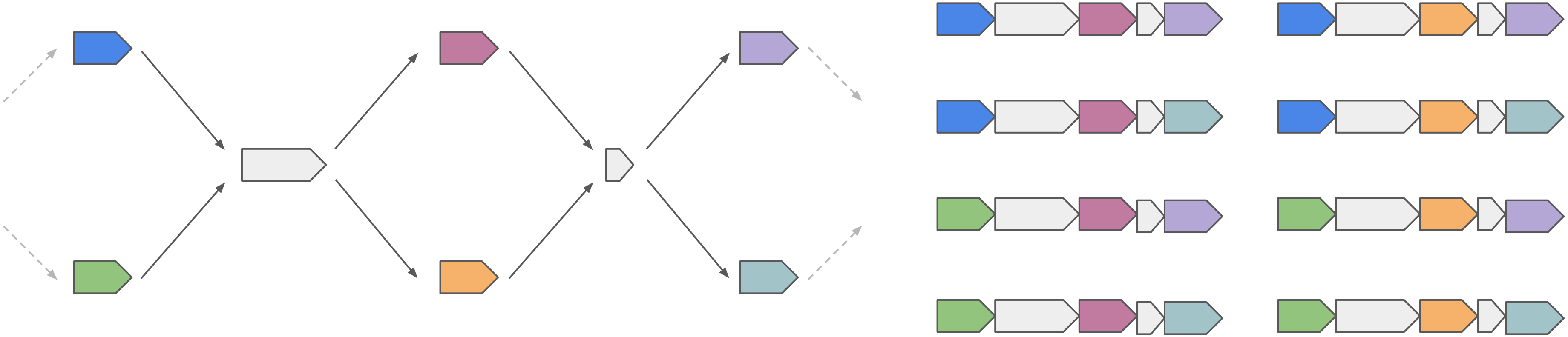


$$X\beta \approx Y$$

We can enumerate all possible paths on our assembly graph

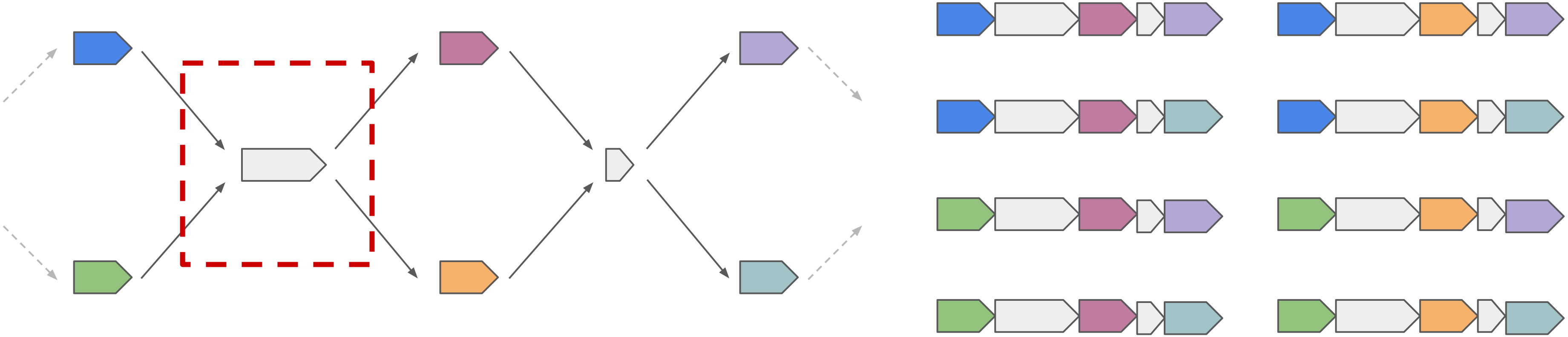
$$\boxed{X}\beta \approx Y$$

We can enumerate all possible paths on our assembly graph

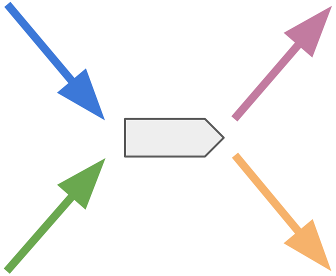


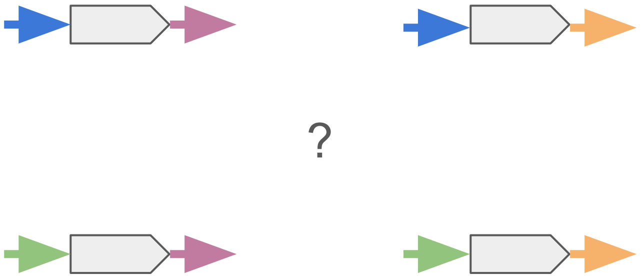
…but this grows exponentially with graph complexity

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

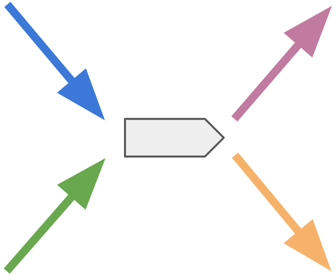


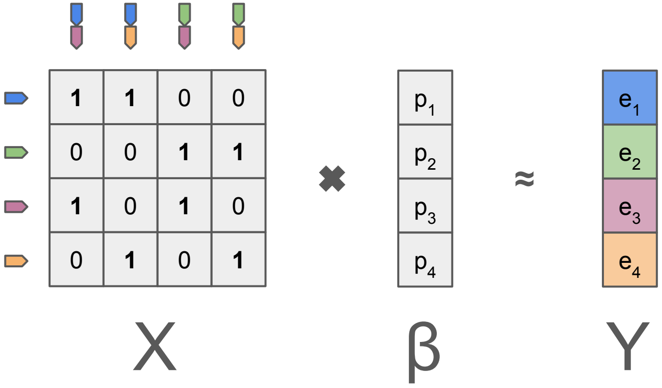
Divide and conquer: a single "junction" is the minimum unit of deconvolution





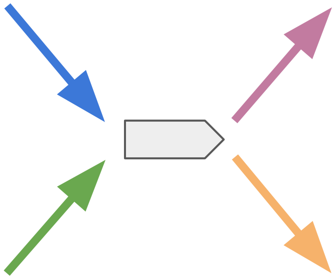
Linear model of path depths

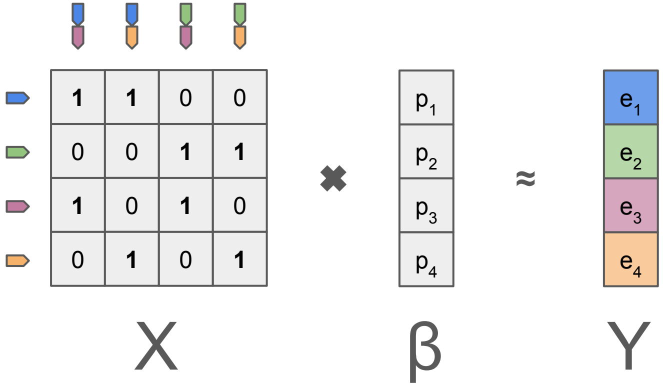




$$\underset{X}{\begin{bmatrix}1&1&0&0\\0&0&1&1\\1&0&1&0\\0&1&0&1\end{bmatrix}} \times \underset{\beta}{\begin{bmatrix}p_1\\p_2\\p_3\\p_4\end{bmatrix}} \approx \underset{Y}{\begin{bmatrix}e_1\\e_2\\e_3\\e_4\end{bmatrix}}$$

But not all paths exist: picking active paths is model selection

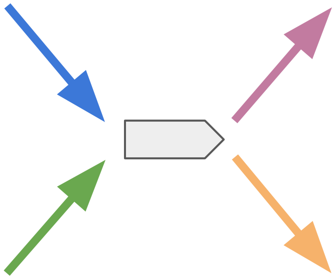


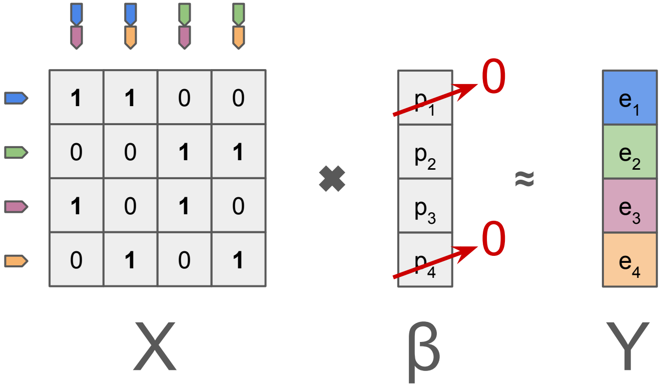


$$\underbrace{\begin{pmatrix}1&1&0&0\\0&0&1&1\\1&0&1&0\\0&1&0&1\end{pmatrix}}_{X} \times \underbrace{\begin{pmatrix}p_1\\p_2\\p_3\\p_4\end{pmatrix}}_{\beta} \approx \underbrace{\begin{pmatrix}e_1\\e_2\\e_3\\e_4\end{pmatrix}}_{Y}$$

$$\hat{\beta} = \operatorname{argmin} L(X\beta \mid Y)$$

But not all paths exist: picking active paths is model selection

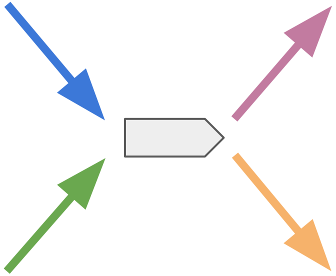


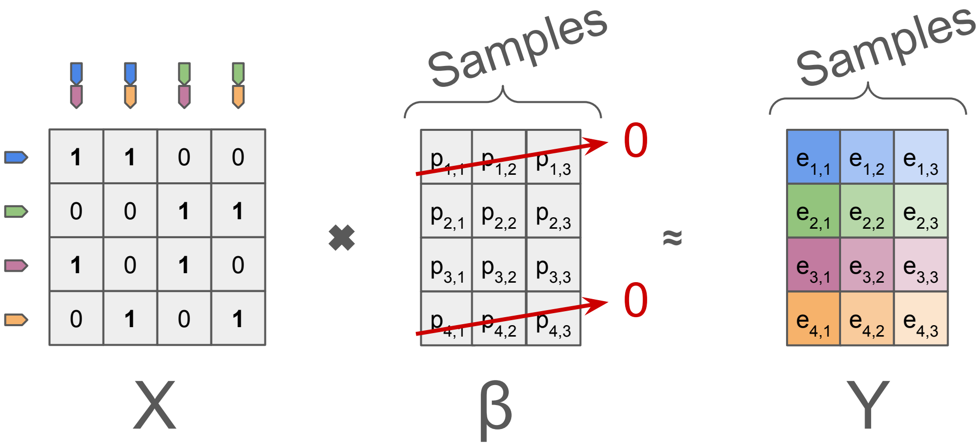


$$\underset{X}{\begin{pmatrix}1&1&0&0\\0&0&1&1\\1&0&1&0\\0&1&0&1\end{pmatrix}} \times \underset{\beta}{\begin{pmatrix}\cancel{p_1} \rightarrow 0\\ p_2\\ p_3\\ \cancel{p_4} \rightarrow 0\end{pmatrix}} \approx \underset{Y}{\begin{pmatrix}e_1\\ e_2\\ e_3\\ e_4\end{pmatrix}}$$

$$\hat{\beta} = \operatorname{argmin} L(X\beta \,|\, Y)$$

Last trick: To increase our power to pick paths, combine multiple samples

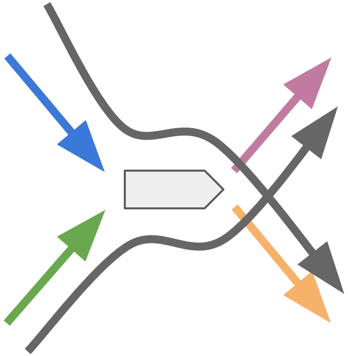


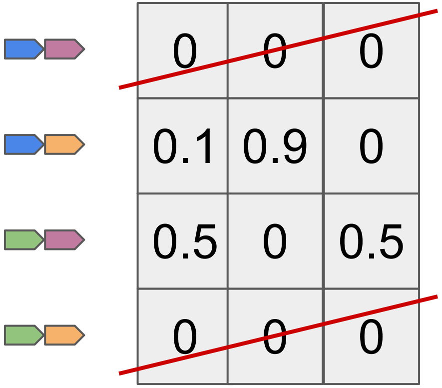


$$\begin{pmatrix}1&1&0&0\\0&0&1&1\\1&0&1&0\\0&1&0&1\end{pmatrix} \times \begin{pmatrix}p_{1,1}&p_{1,2}&p_{1,3}\\p_{2,1}&p_{2,2}&p_{2,3}\\p_{3,1}&p_{3,2}&p_{3,3}\\p_{4,1}&p_{4,2}&p_{4,3}\end{pmatrix} \approx \begin{pmatrix}e_{1,1}&e_{1,2}&e_{1,3}\\e_{2,1}&e_{2,2}&e_{2,3}\\e_{3,1}&e_{3,2}&e_{3,3}\\e_{4,1}&e_{4,2}&e_{4,3}\end{pmatrix}$$

$$\hat{\beta} = \operatorname{argmin} L(X\beta \,|\, Y)$$

Drop paths with no depth in any sample



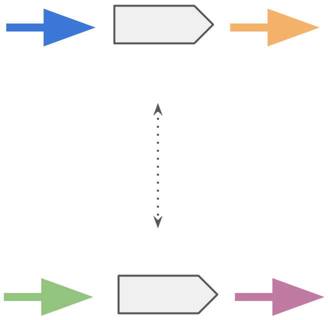


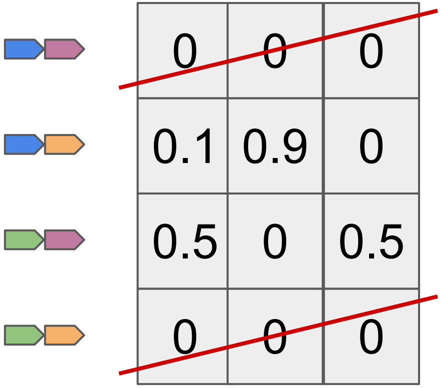
🟦🟪	~~0~~	~~0~~	~~0~~
🟦🟧	0.1	0.9	0
🟩🟪	0.5	0	0.5
🟩🟧	~~0~~	~~0~~	~~0~~

$$\hat{\beta}$$

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

Resolve ambiguity, longer linear sequences



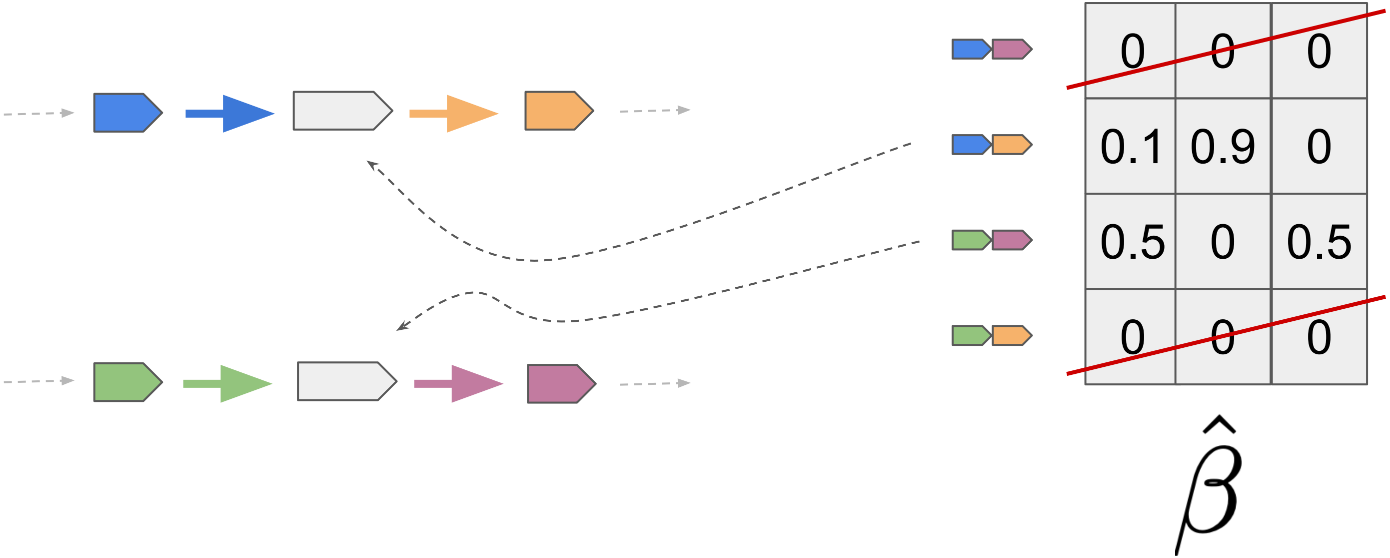


🟦🟪	~~0~~	~~0~~	~~0~~
🟦🟧	0.1	0.9	0
🟩🟪	0.5	0	0.5
🟩🟧	~~0~~	~~0~~	~~0~~

$\hat{\beta}$

Can "unzip" this unitig into two paths

Resolve ambiguity, longer linear sequences

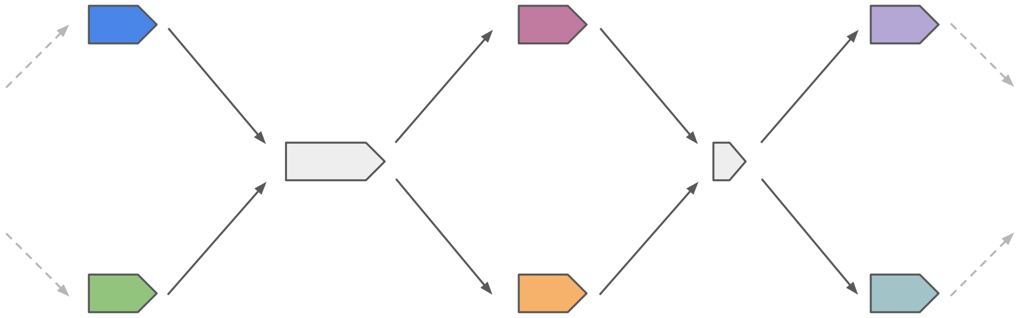


🟦🟪	~~0~~	~~0~~	~~0~~
🟦🟧	0.1	0.9	0
🟩🟪	0.5	0	0.5
🟩🟧	~~0~~	~~0~~	~~0~~

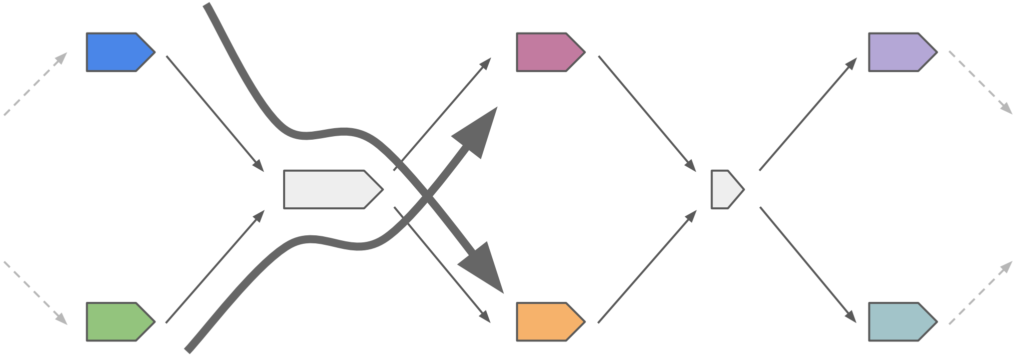
$\hat{\beta}$

Newly split unitigs already have depths estimated across samples

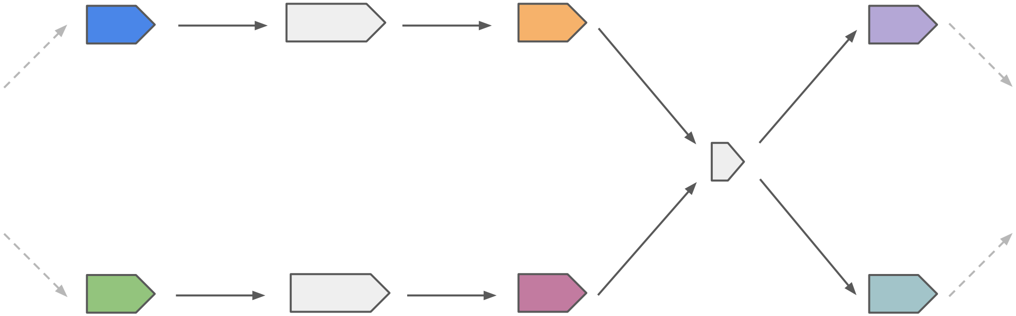
Iteratively unzipping local junctions



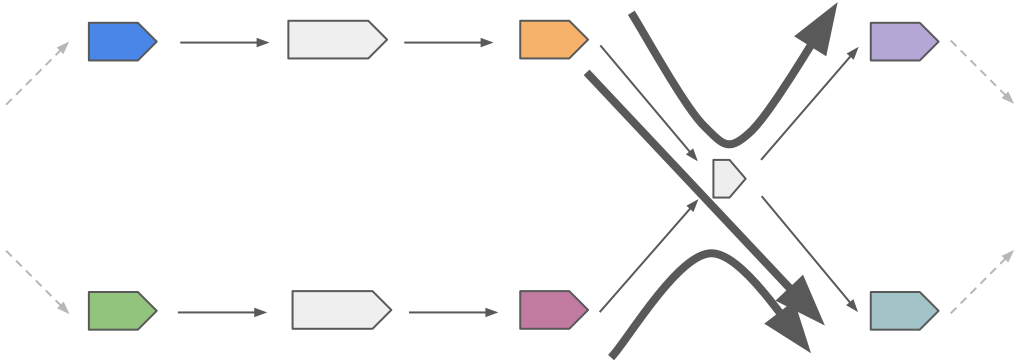
Iteratively unzipping local junctions



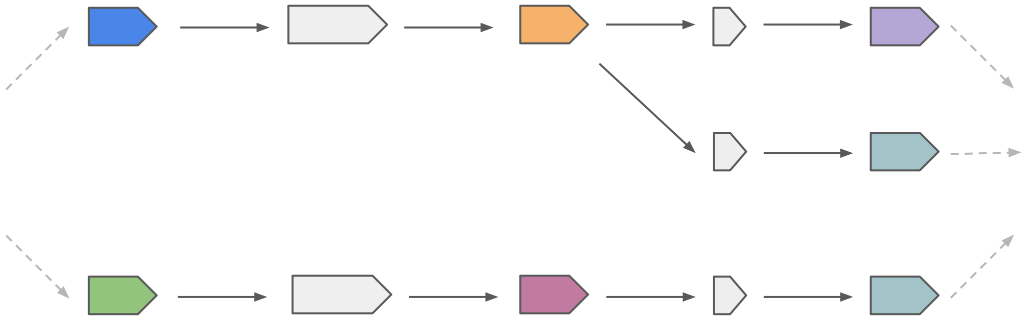
Iteratively unzipping local junctions



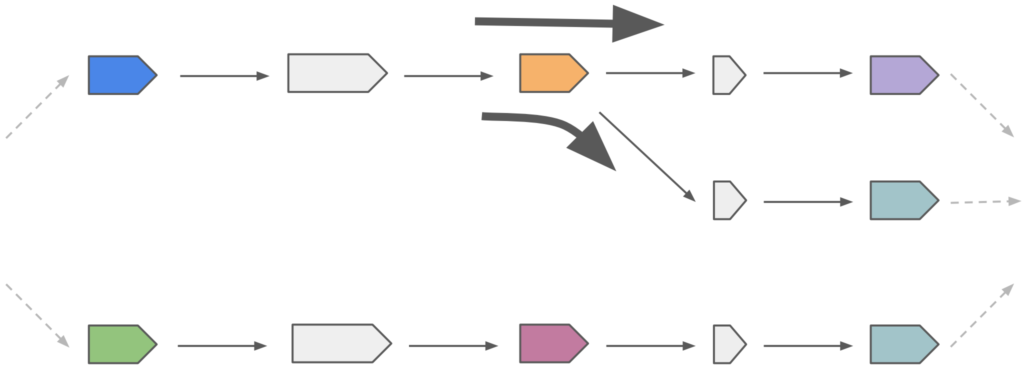
Iteratively unzipping local junctions



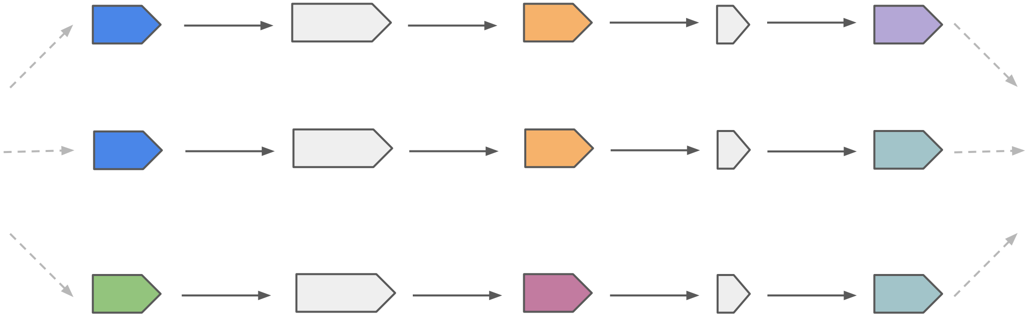
Iteratively unzipping local junctions



Iteratively unzipping local junctions



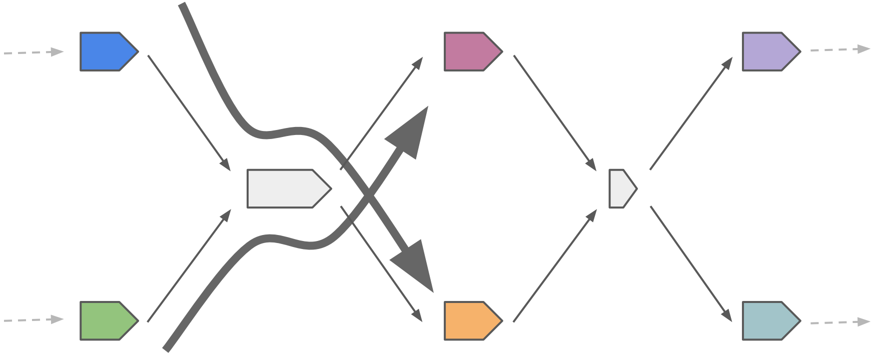
Iteratively unzipping local junctions

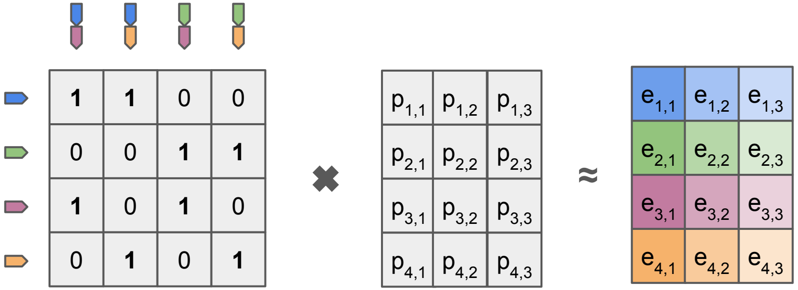


StrainZip

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

https://github.com/bsmith89/StrainZip

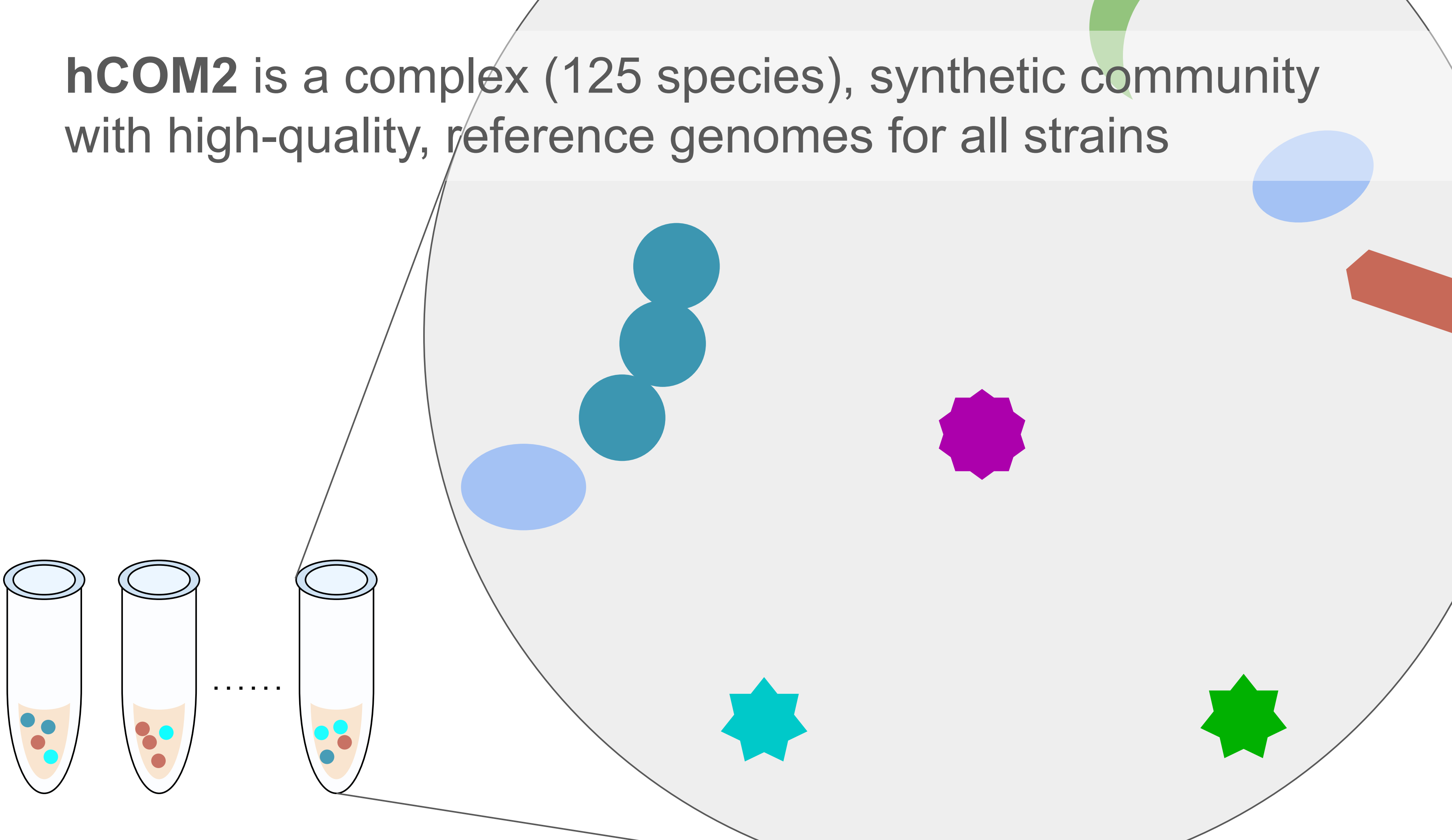




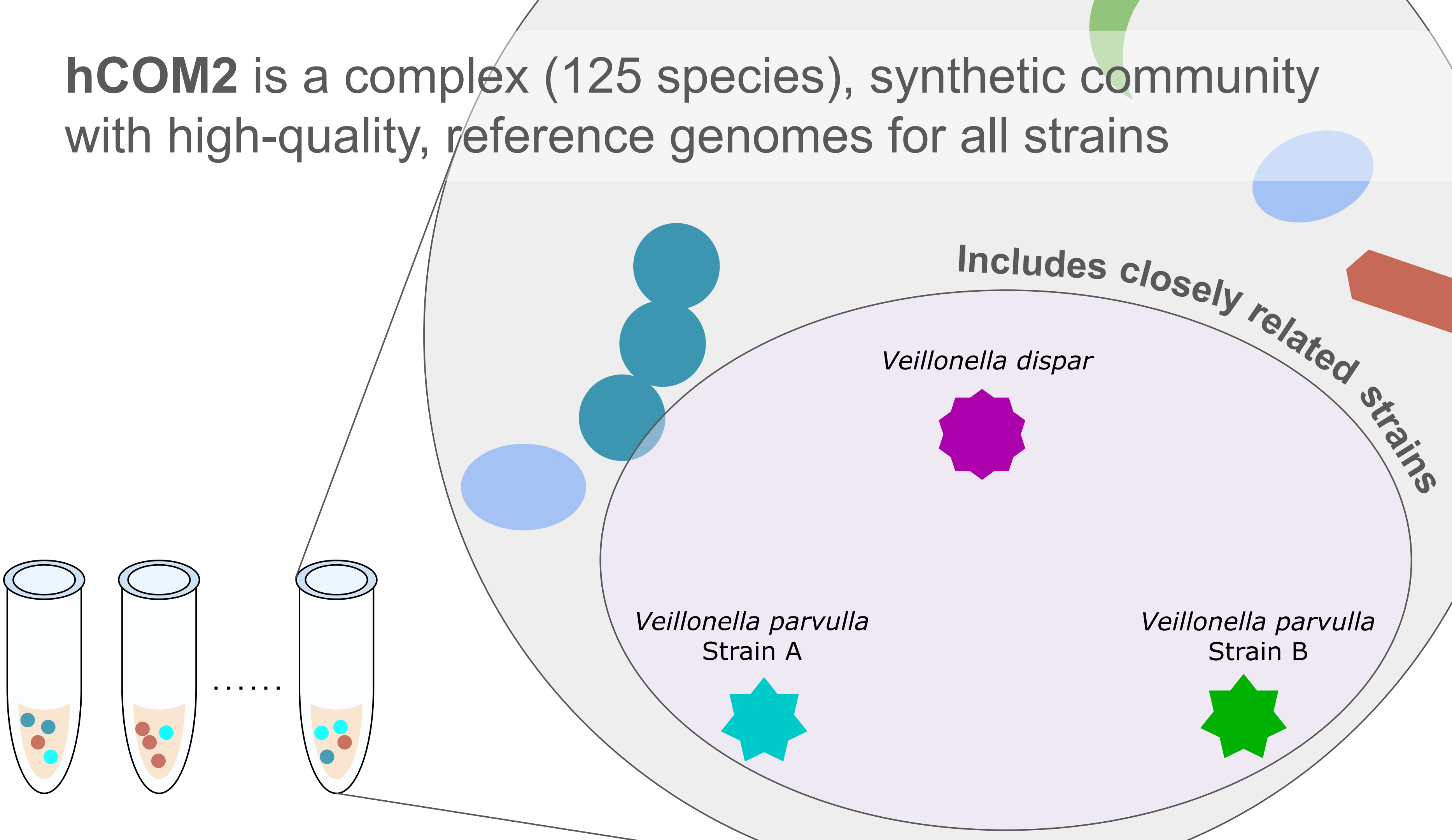
$$\begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} p_{1,1} & p_{1,2} & p_{1,3} \\ p_{2,1} & p_{2,2} & p_{2,3} \\ p_{3,1} & p_{3,2} & p_{3,3} \\ p_{4,1} & p_{4,2} & p_{4,3} \end{bmatrix} \approx \begin{bmatrix} e_{1,1} & e_{1,2} & e_{1,3} \\ e_{2,1} & e_{2,2} & e_{2,3} \\ e_{3,1} & e_{3,2} & e_{3,3} \\ e_{4,1} & e_{4,2} & e_{4,3} \end{bmatrix}$$

Demonstration

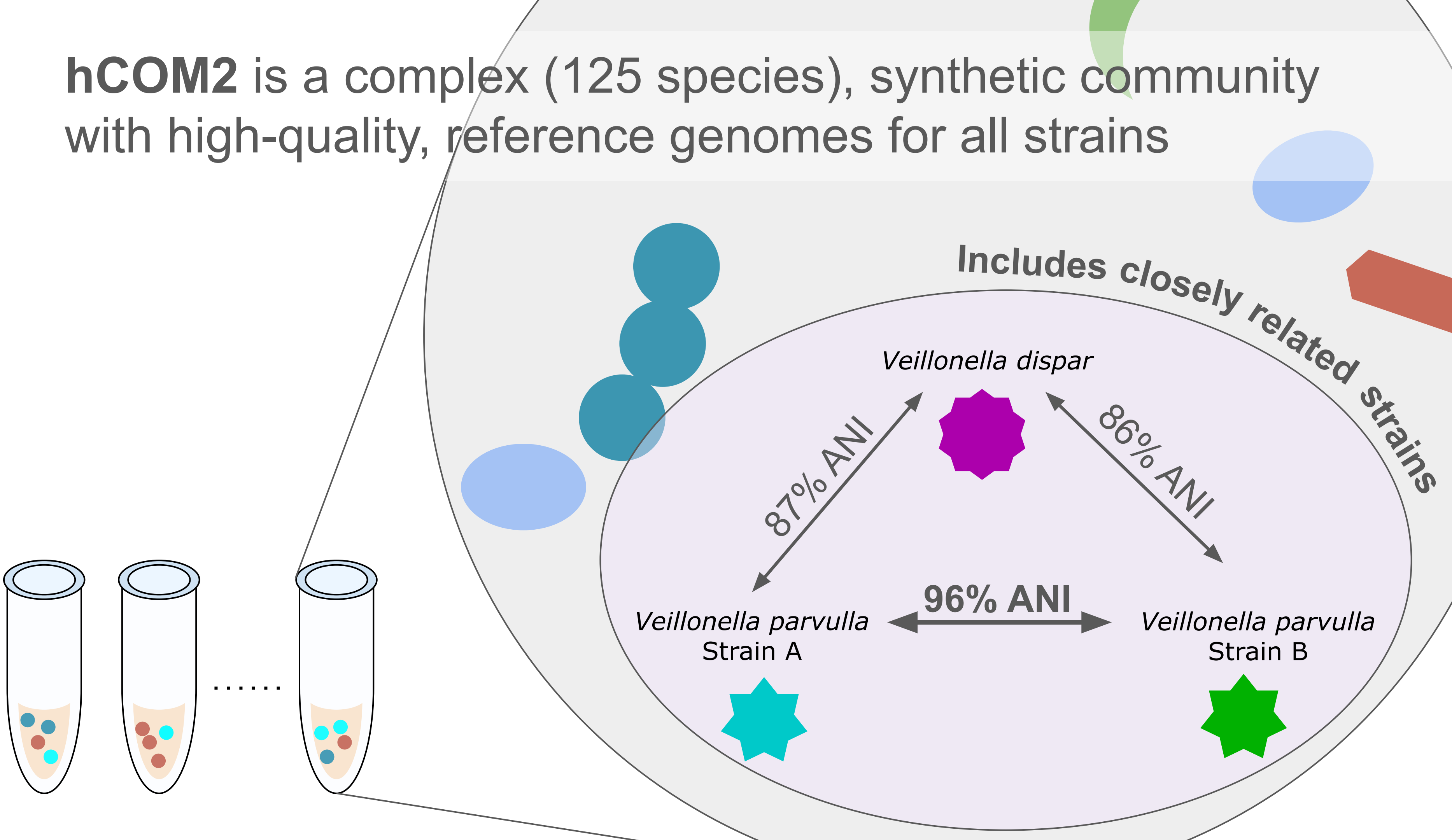
hCOM2 is a complex (125 species), synthetic community with high-quality, reference genomes for all strains

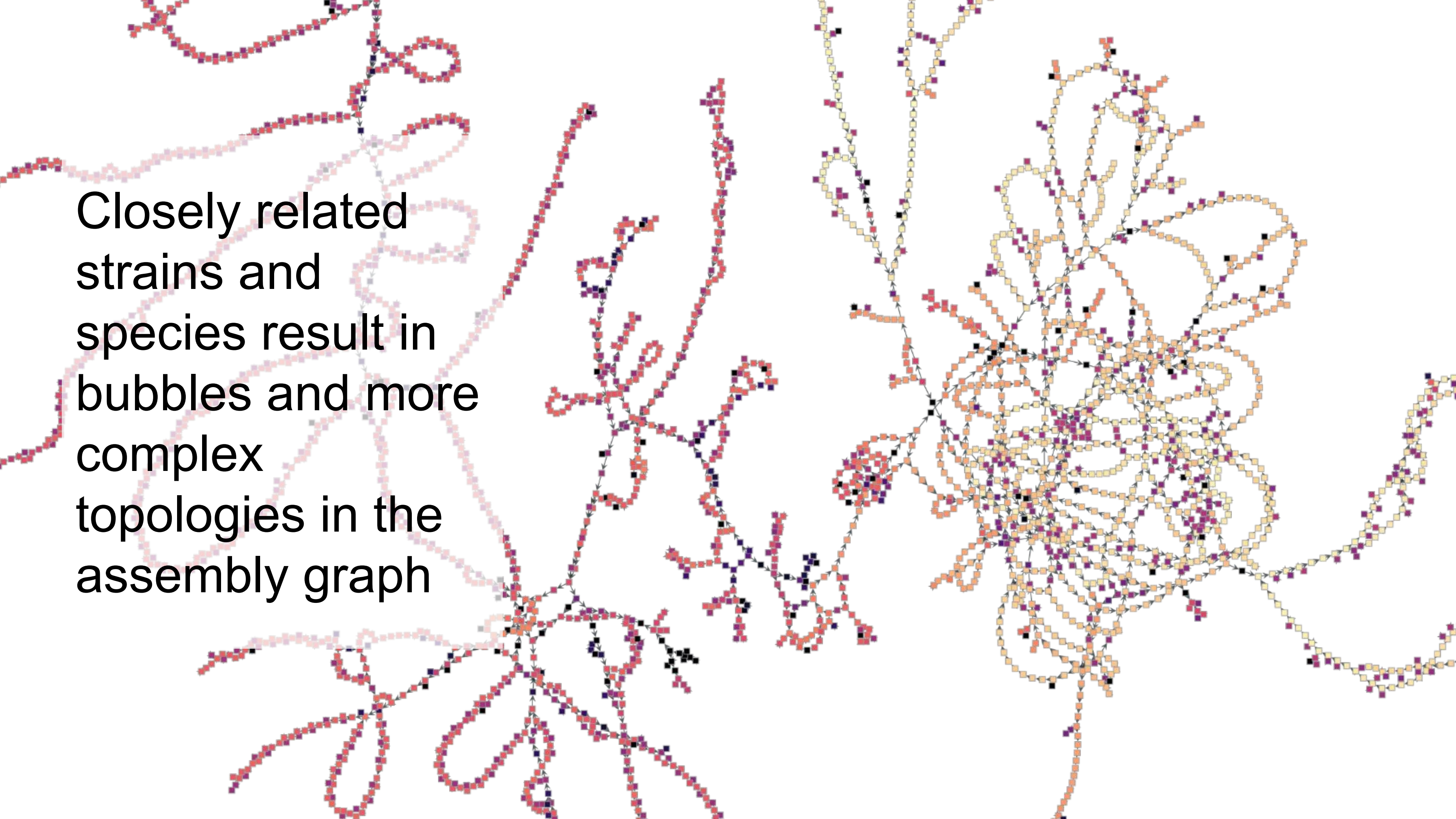


hCOM2 is a complex (125 species), synthetic community with high-quality, reference genomes for all strains

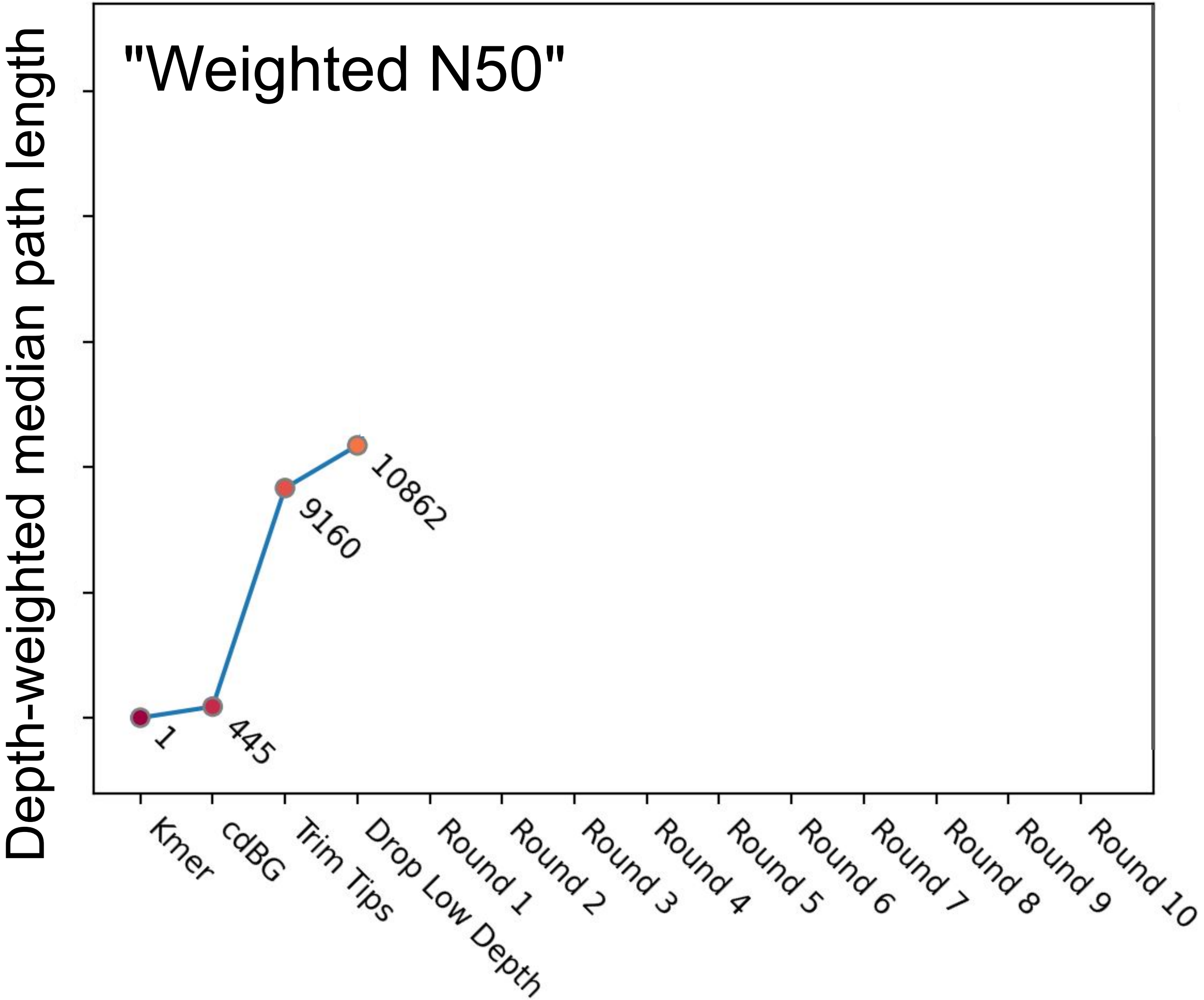


hCOM2 is a complex (125 species), synthetic community with high-quality, reference genomes for all strains

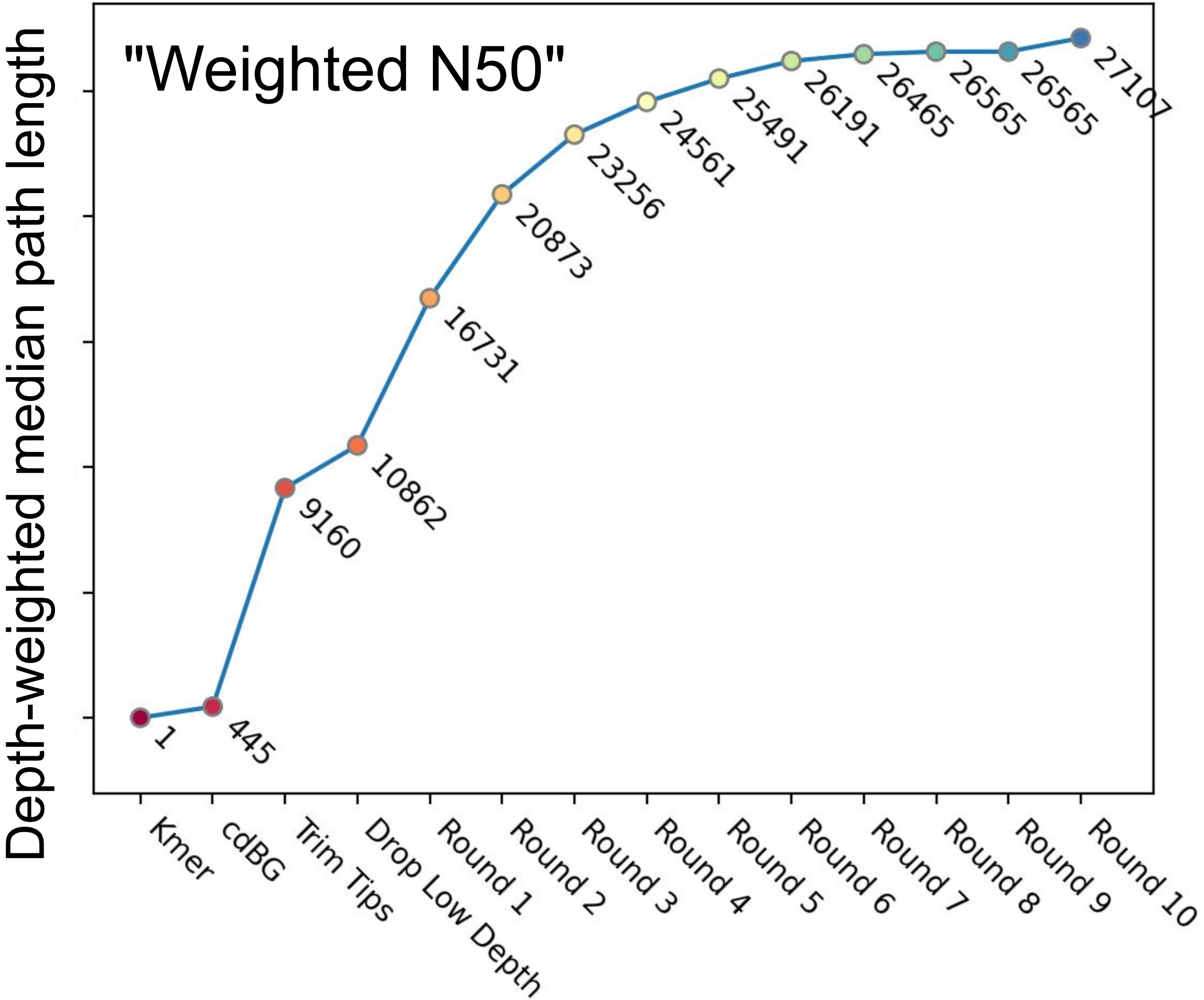




Closely related strains and species result in bubbles and more complex topologies in the assembly graph



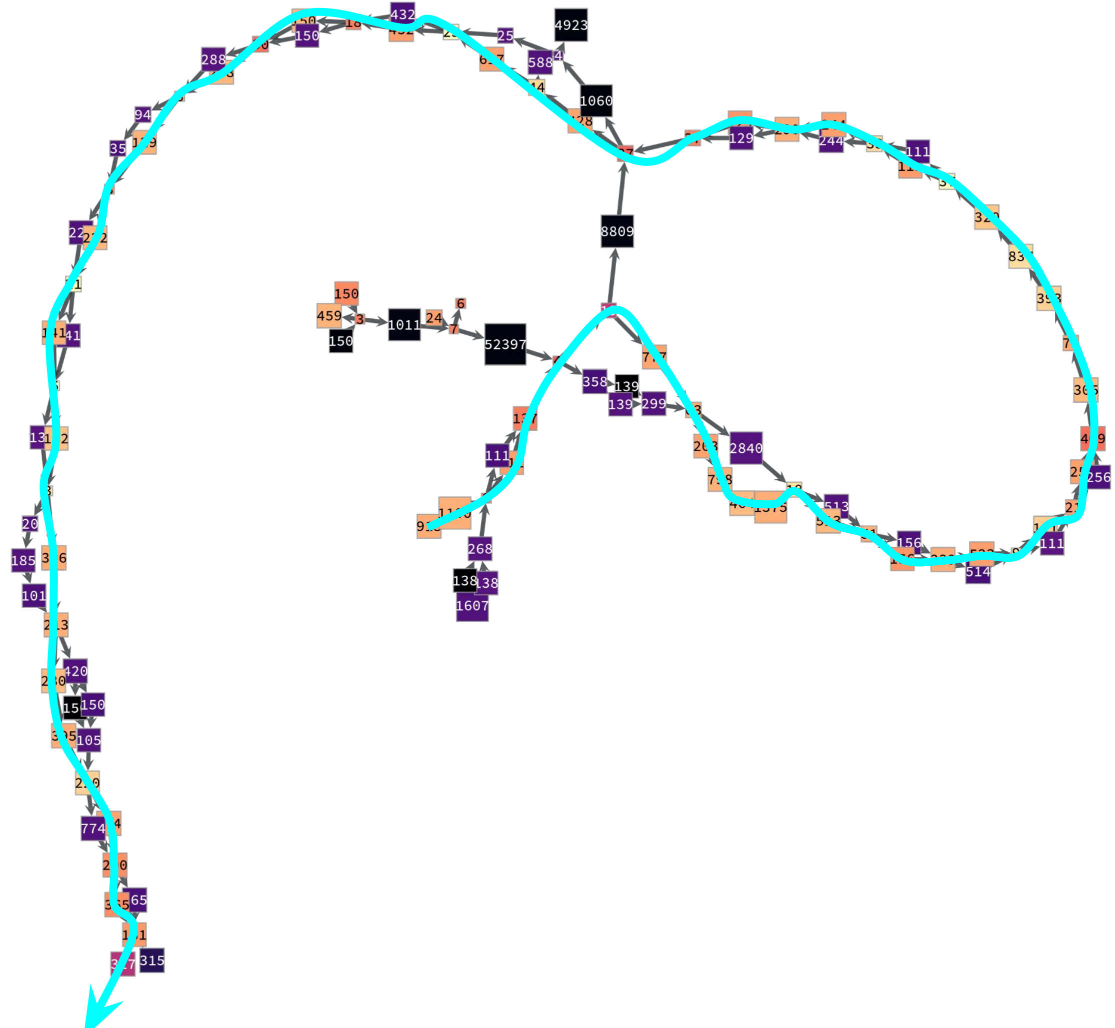
Complex assembly graph results in short path lengths



Path lengths increase over successive rounds of deconvolution

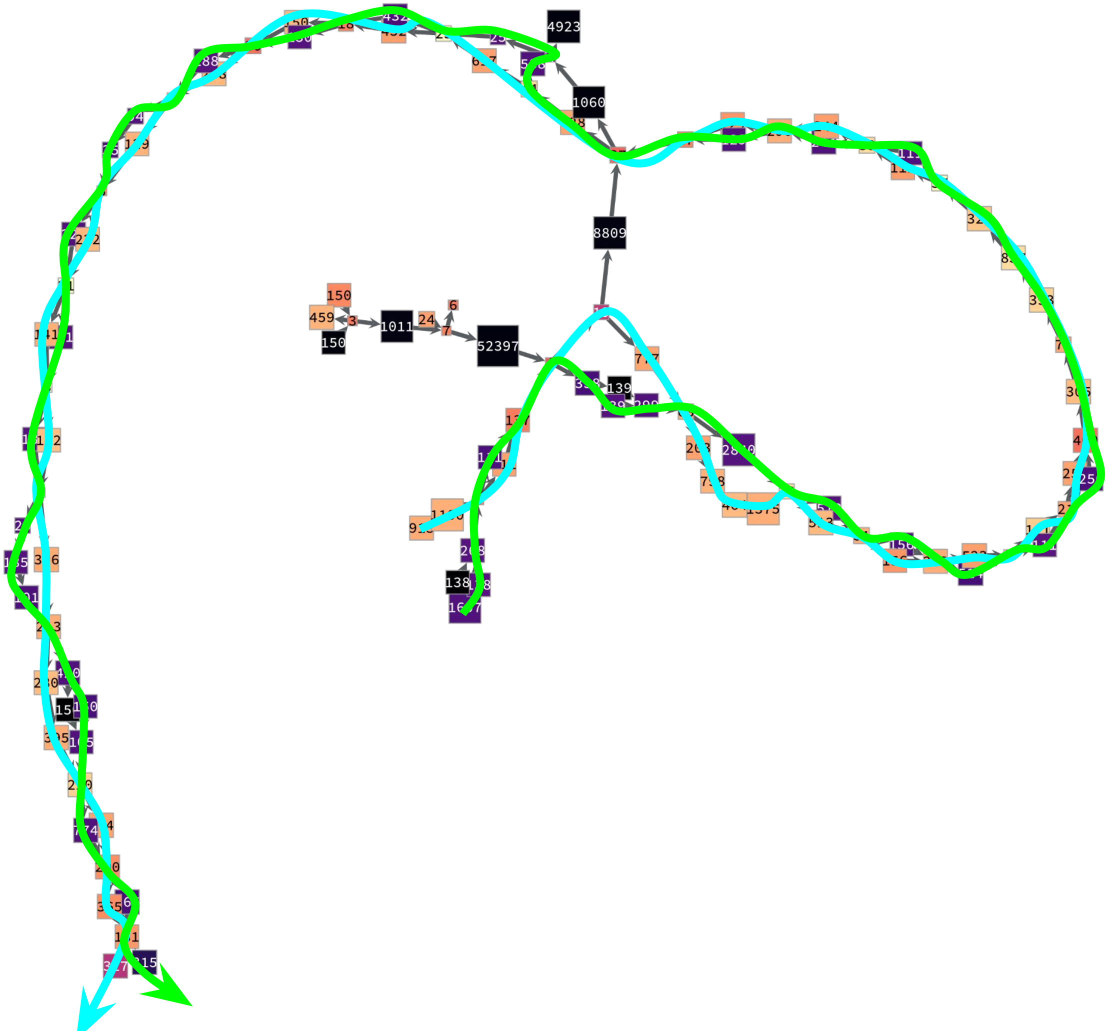
Closely related strains are interspersed in the assembly graph

Deconvolution recovers longer, strain-specific sequences



Deconvolution recovers longer, strain-specific sequences

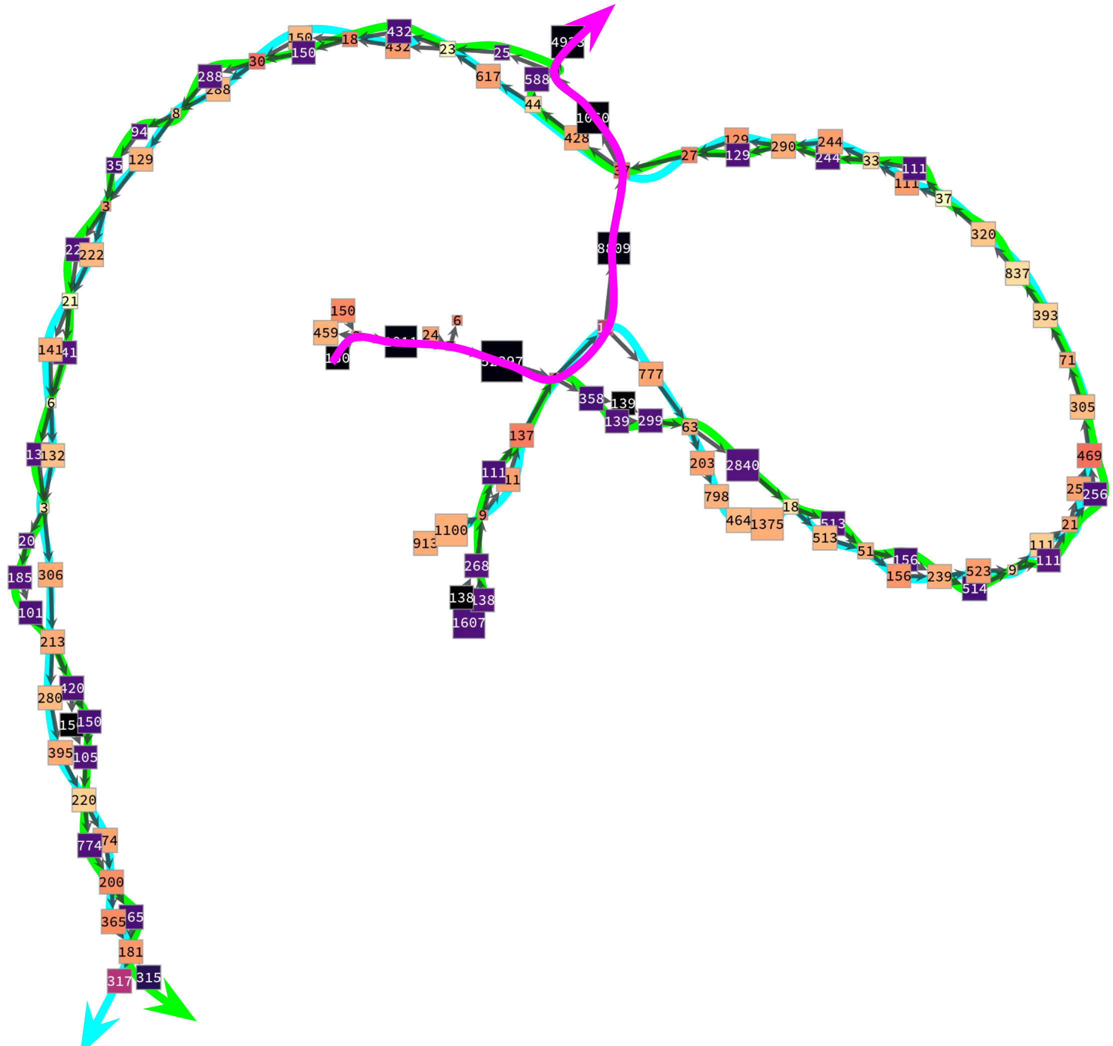
…including lower-abundance strains



Deconvolution recovers longer, strain-specific sequences

…including lower-abundance strains

…and species

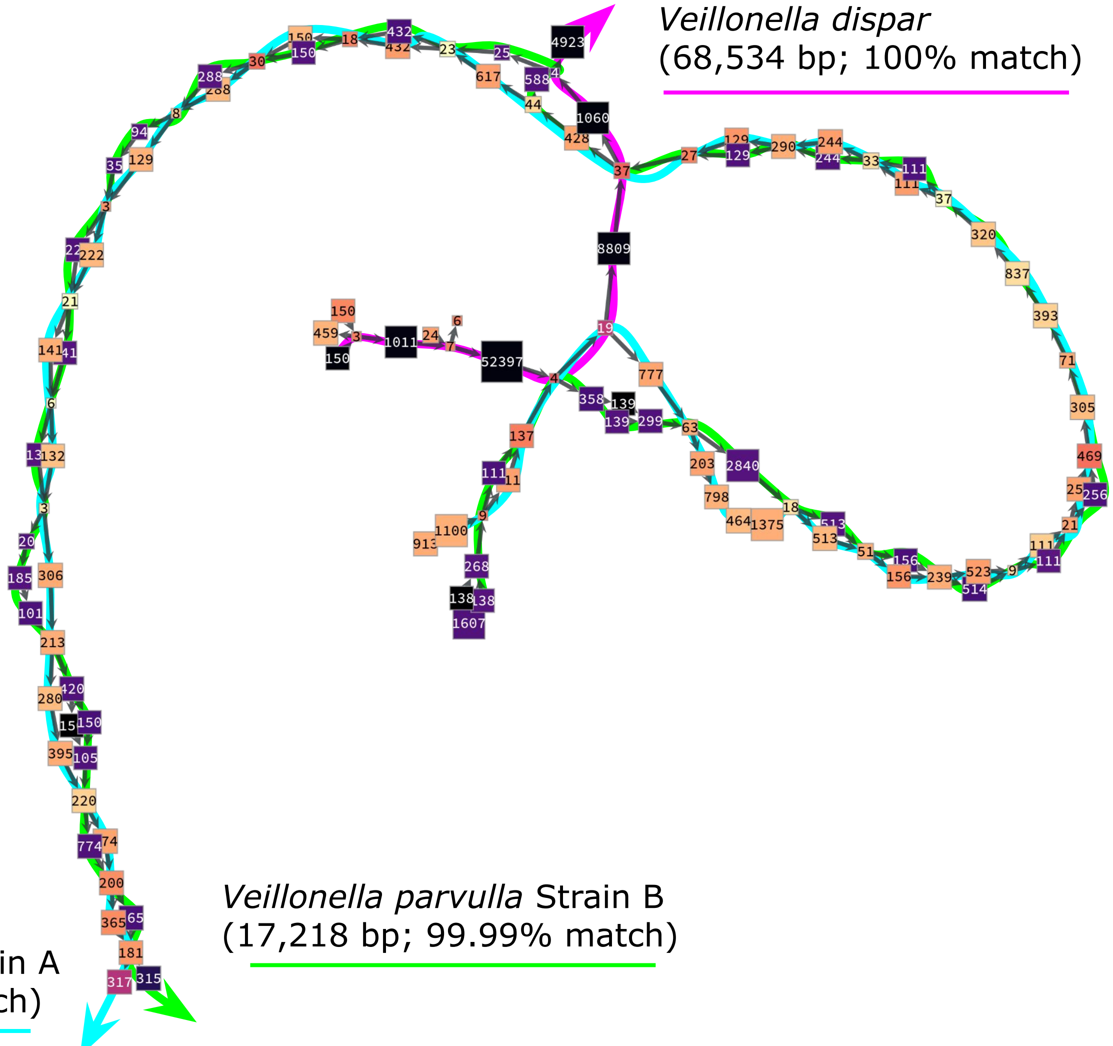


Deconvolution recovers longer, strain-specific sequences

…including lower-abundance strains

…and species

…accurately

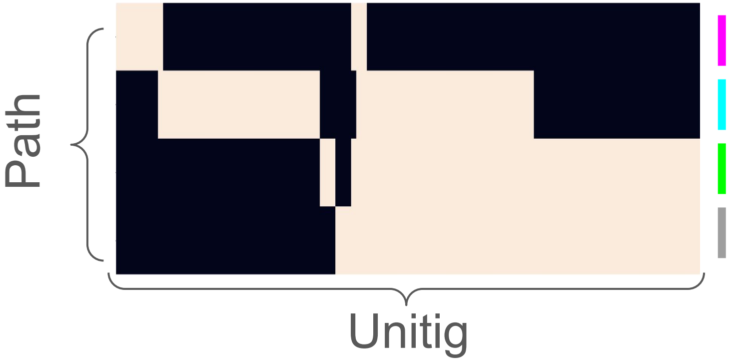


Veillonella dispar
(68,534 bp; 100% match)

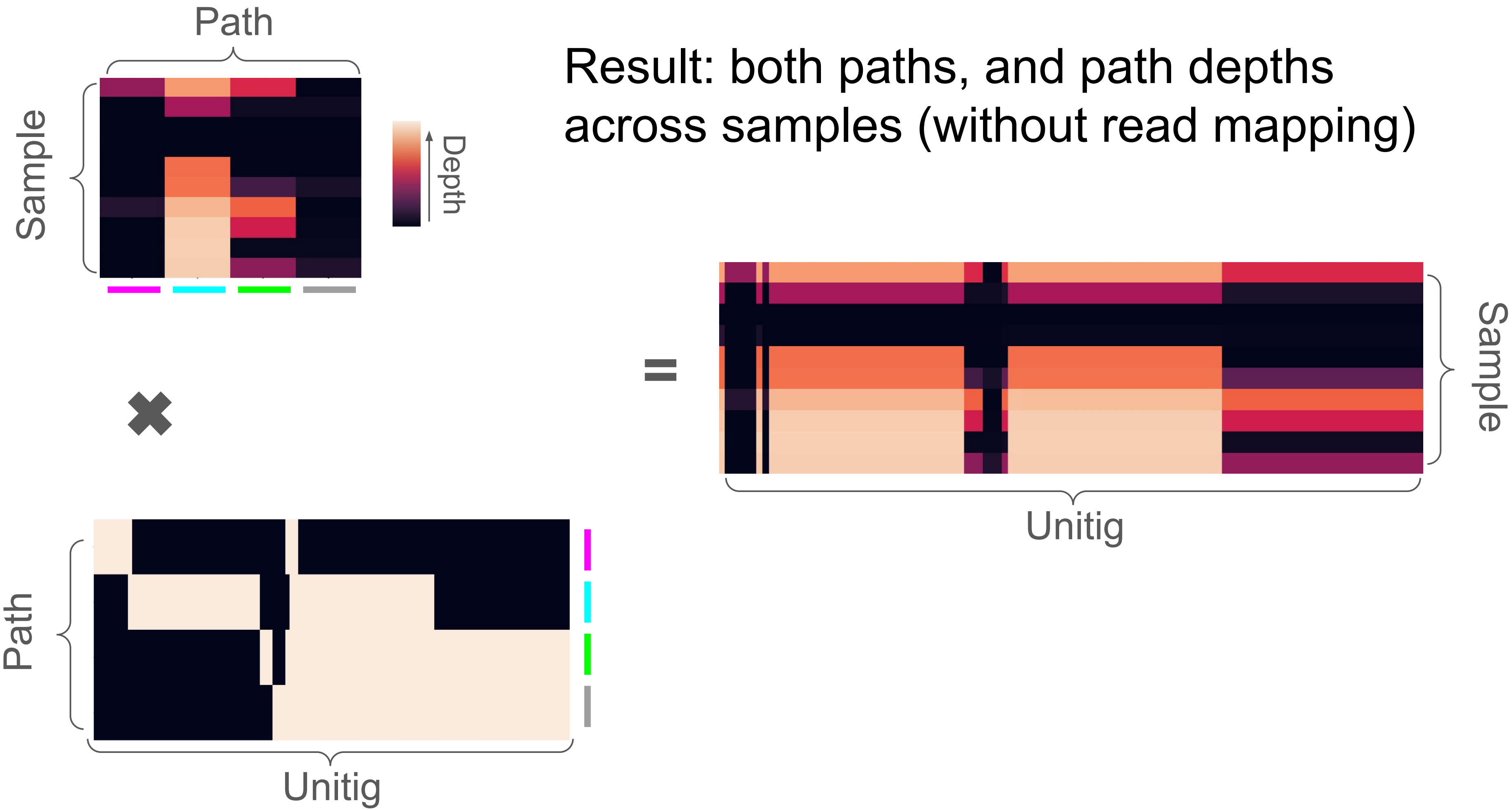
Veillonella parvulla Strain B
(17,218 bp; 99.99% match)

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

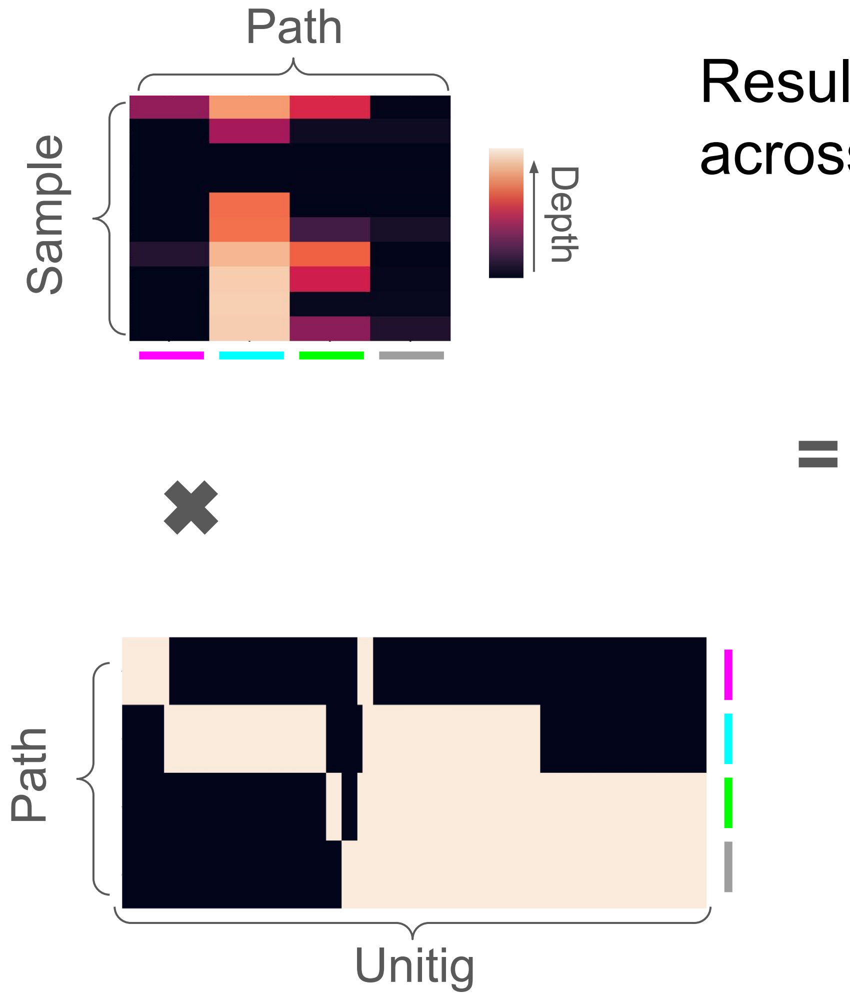
Result: both paths, and path depths across samples (without read mapping)

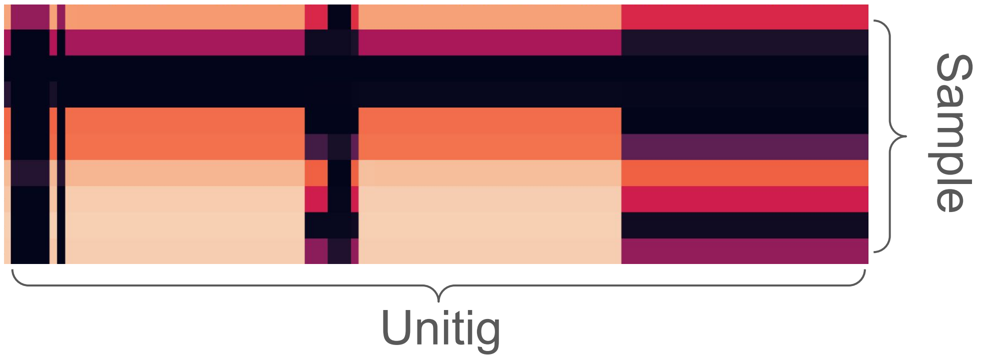


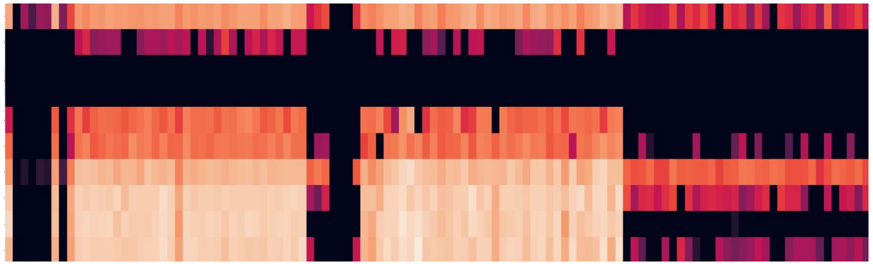
Result: both paths, and path depths across samples (without read mapping)



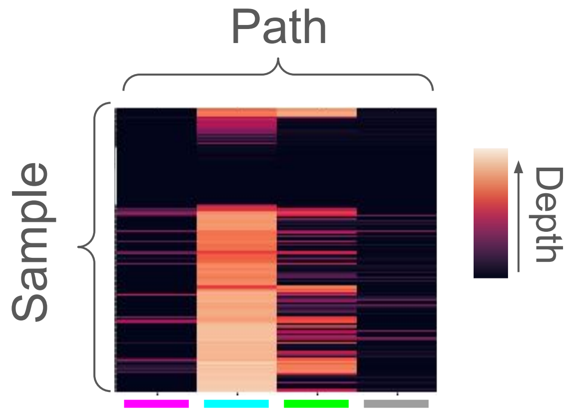
Result: both paths, and path depths across samples (without read mapping)



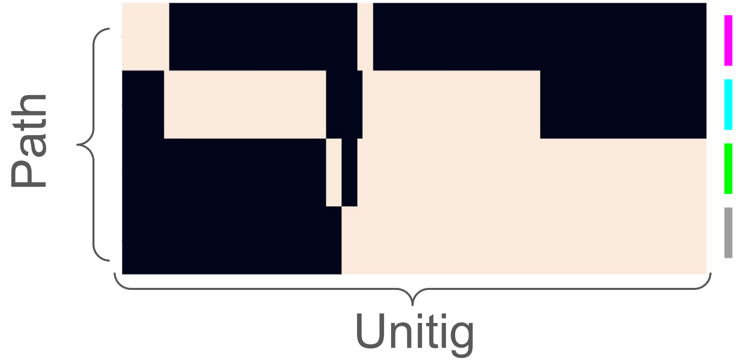




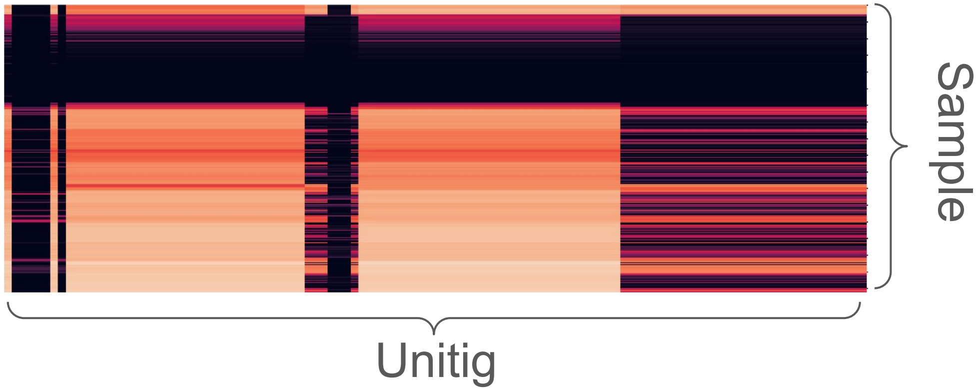
Result: both paths, and path depths across samples (without read mapping)

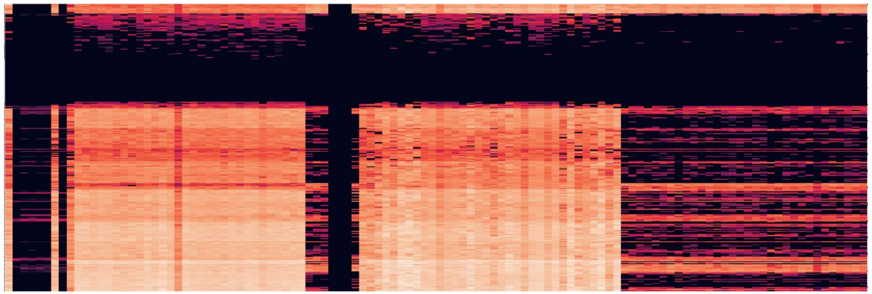


×

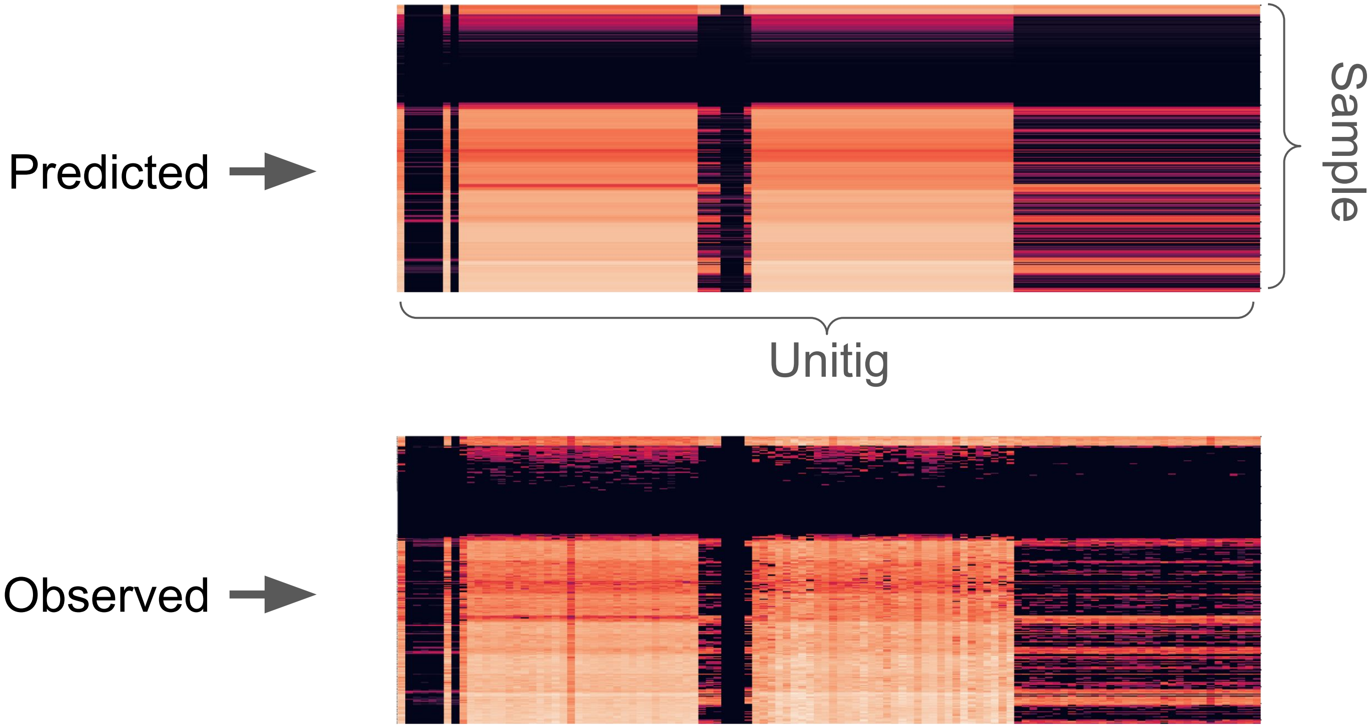


=

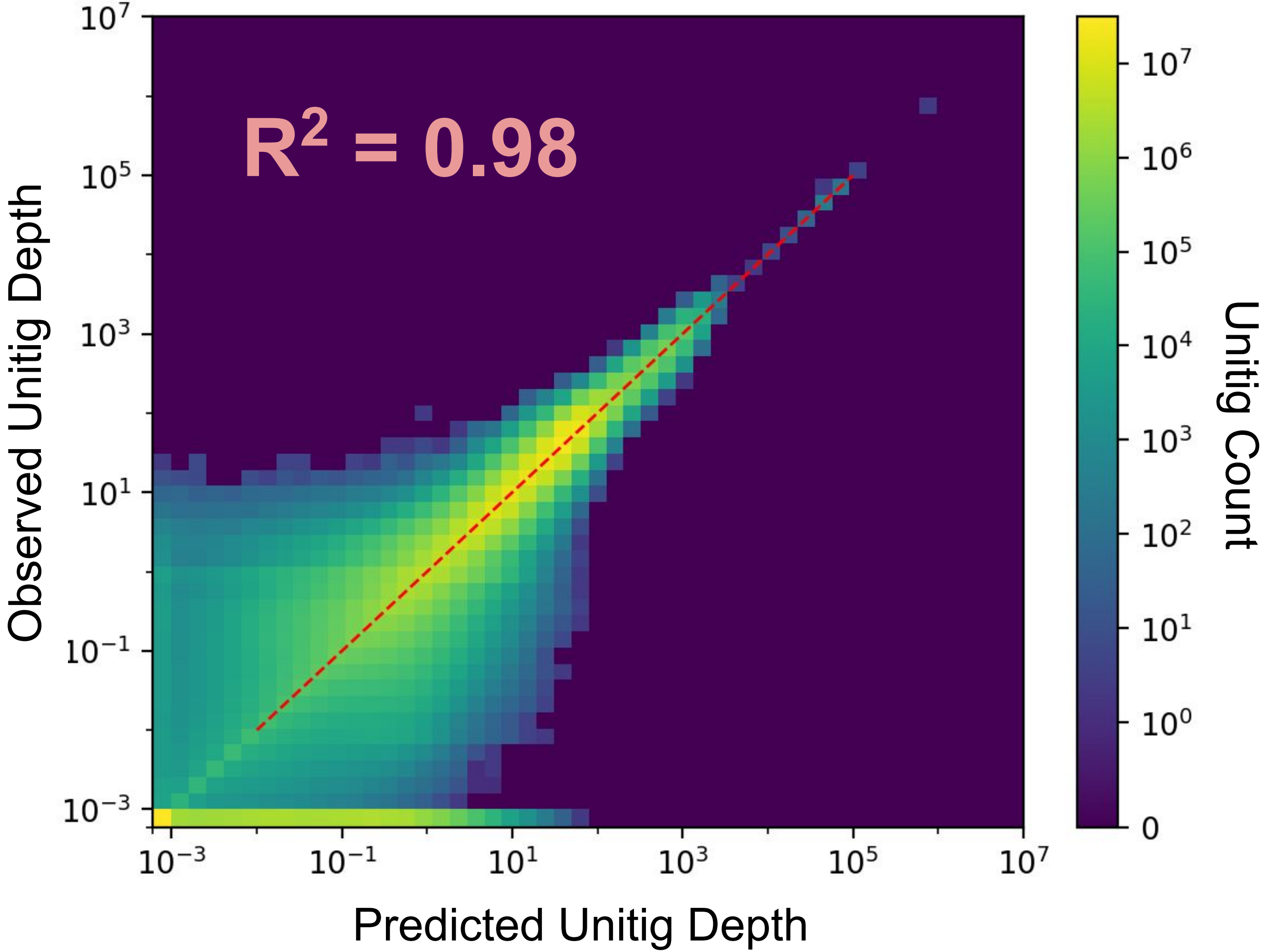




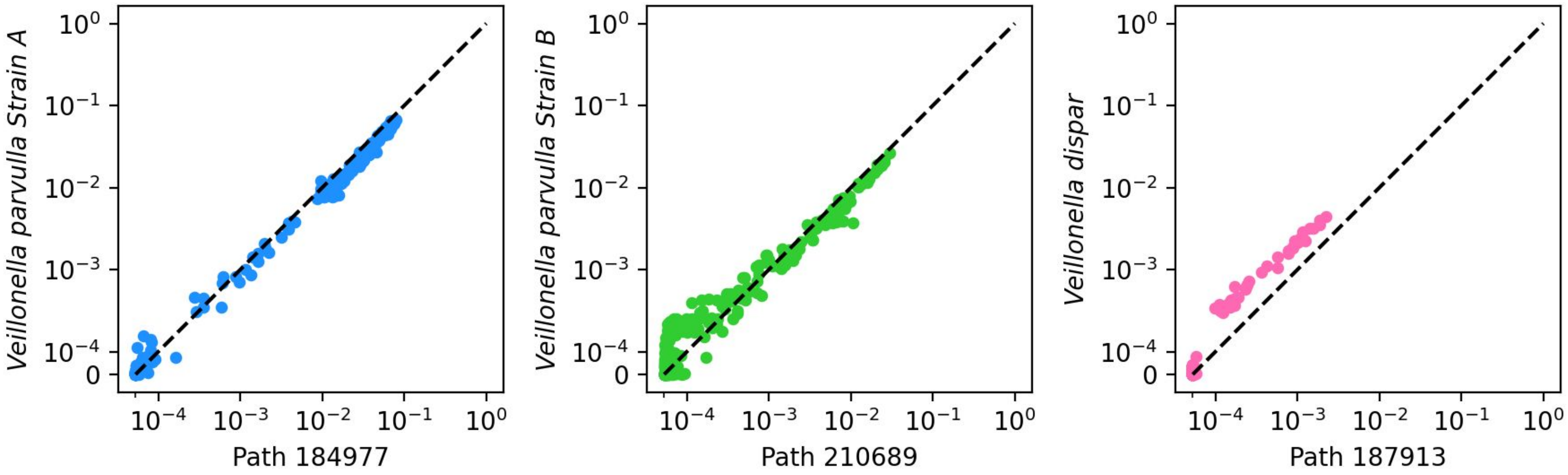
Estimated unitig depths closely match observed depths



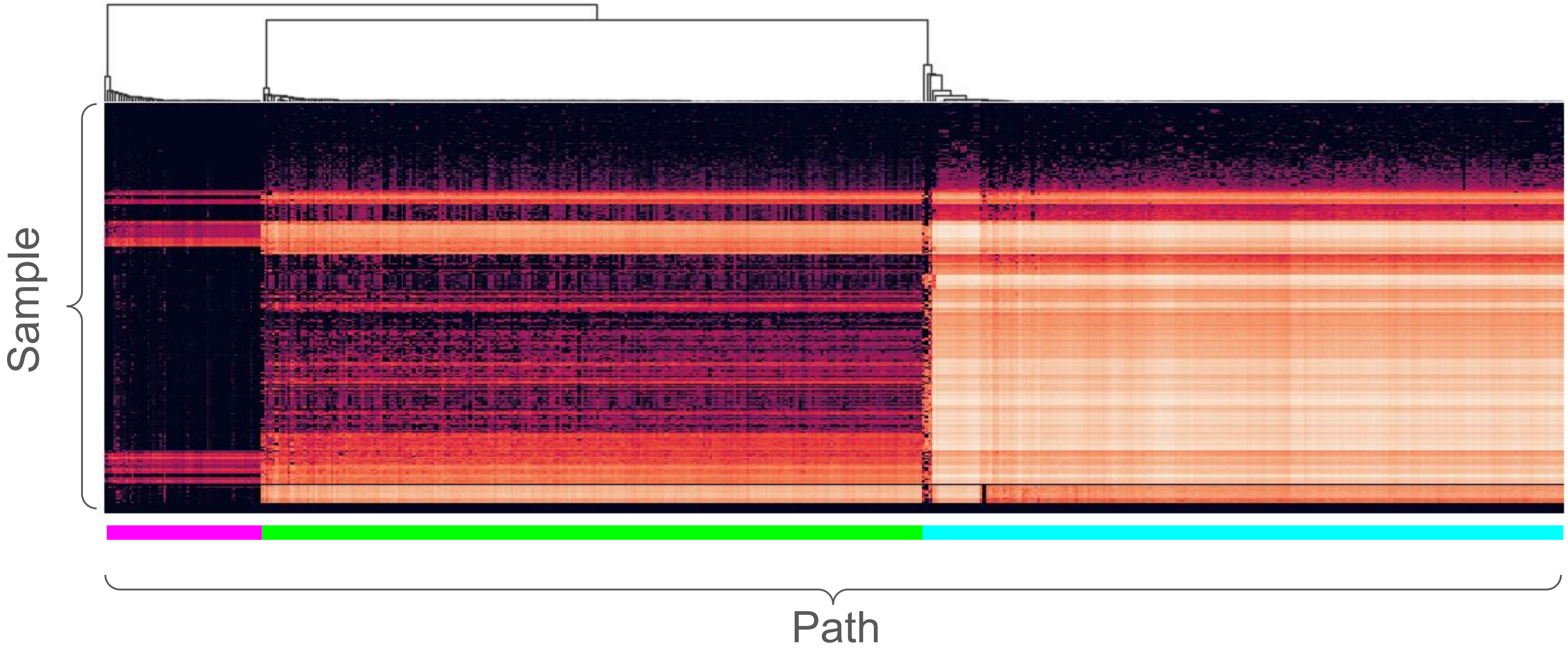
Estimated unitig depths closely match observed depths

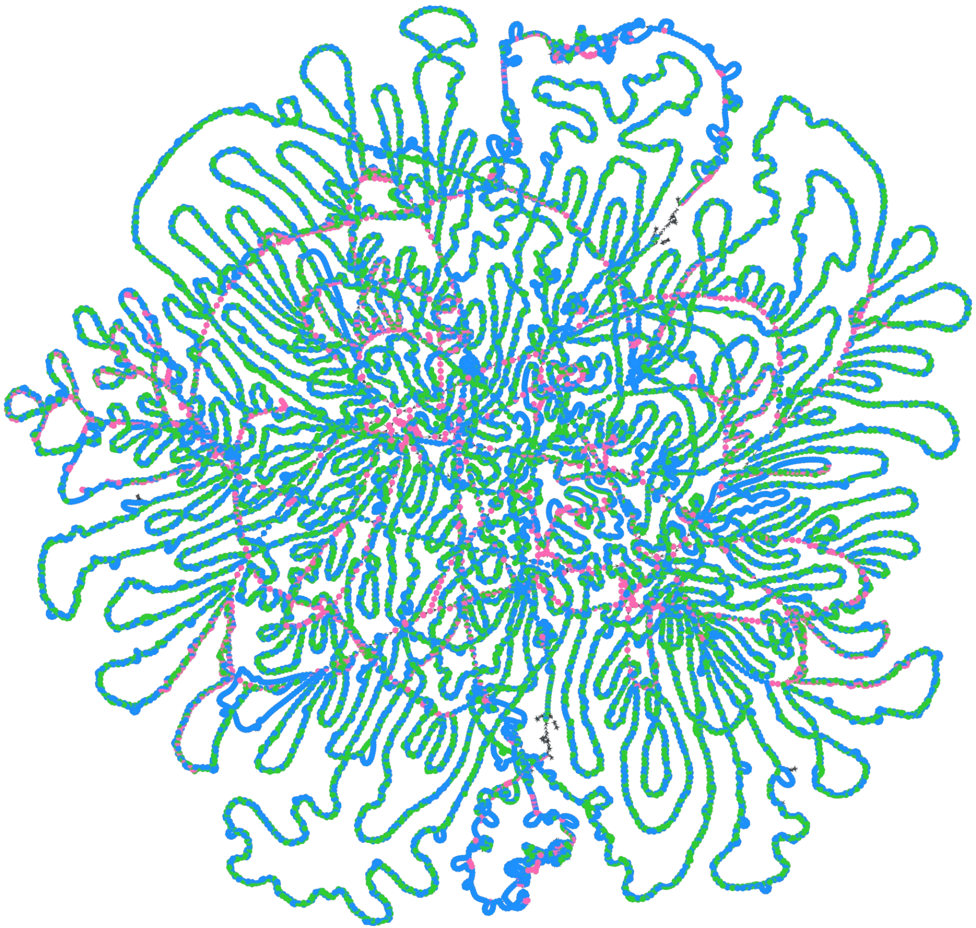


Path depths match reference-based strain depth estimates



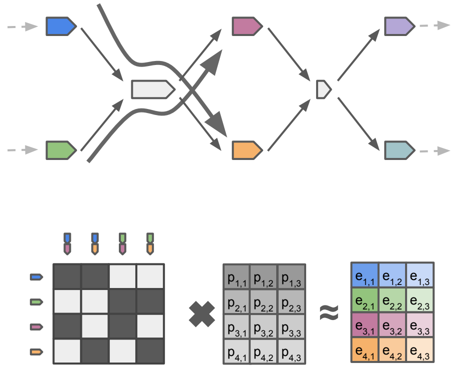
Clustering paths by depth combines multiple sequences from the same strain



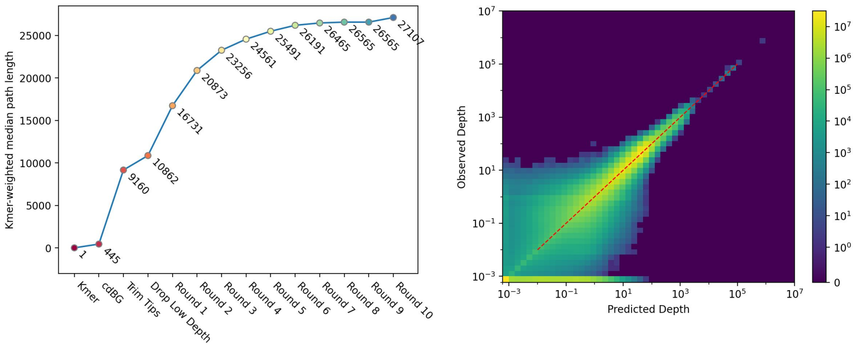


Enables strain-resolved genome assembly from metagenomes

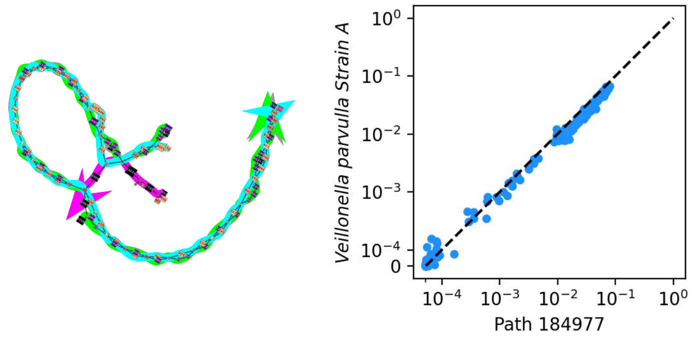
Iterative Junction Deconvolution



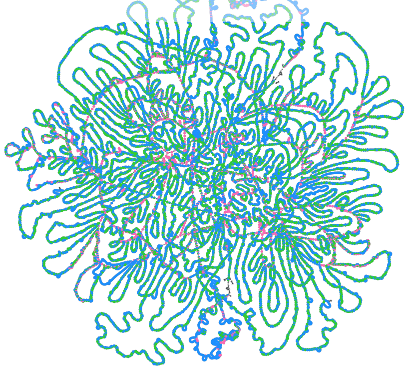
Combines Assembly, Depth Estimation



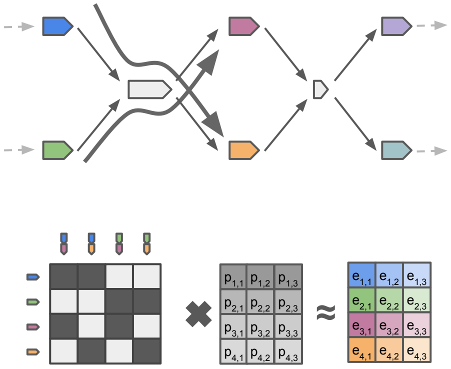
Recovers Closely Related Genomes



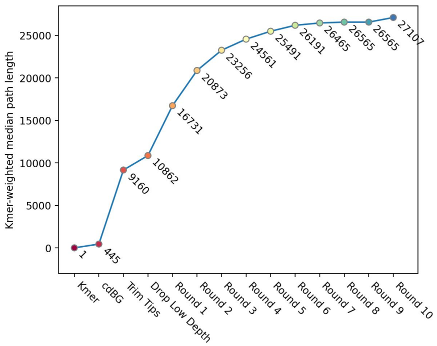
Enables Strain-Resolved Metagenomics

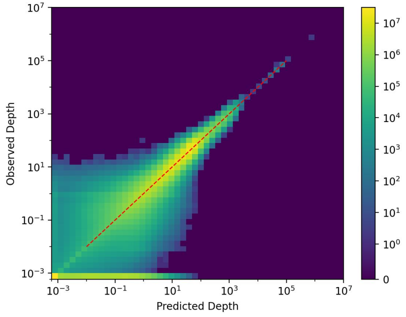


Iterative Junction Deconvolution

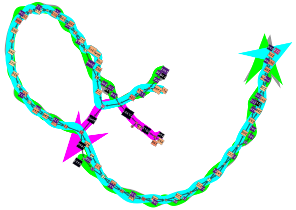


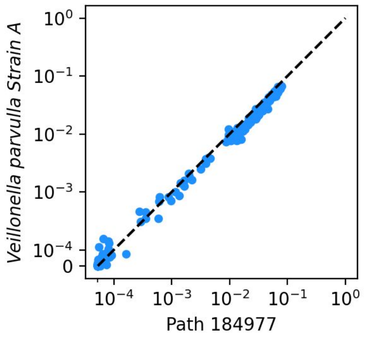
Combines Assembly, Depth Estimation



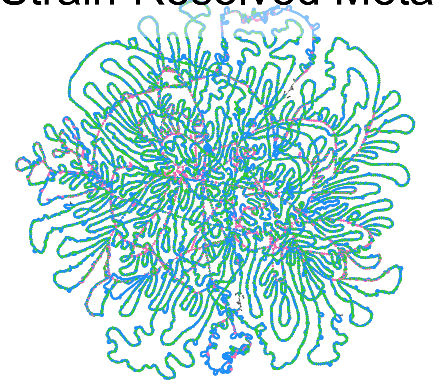


Recovers Closely Related Genomes





Enables Strain-Resolved Metagenomics



Thank You!