Read mapping

Instructor: Emma Bell

(she/her)

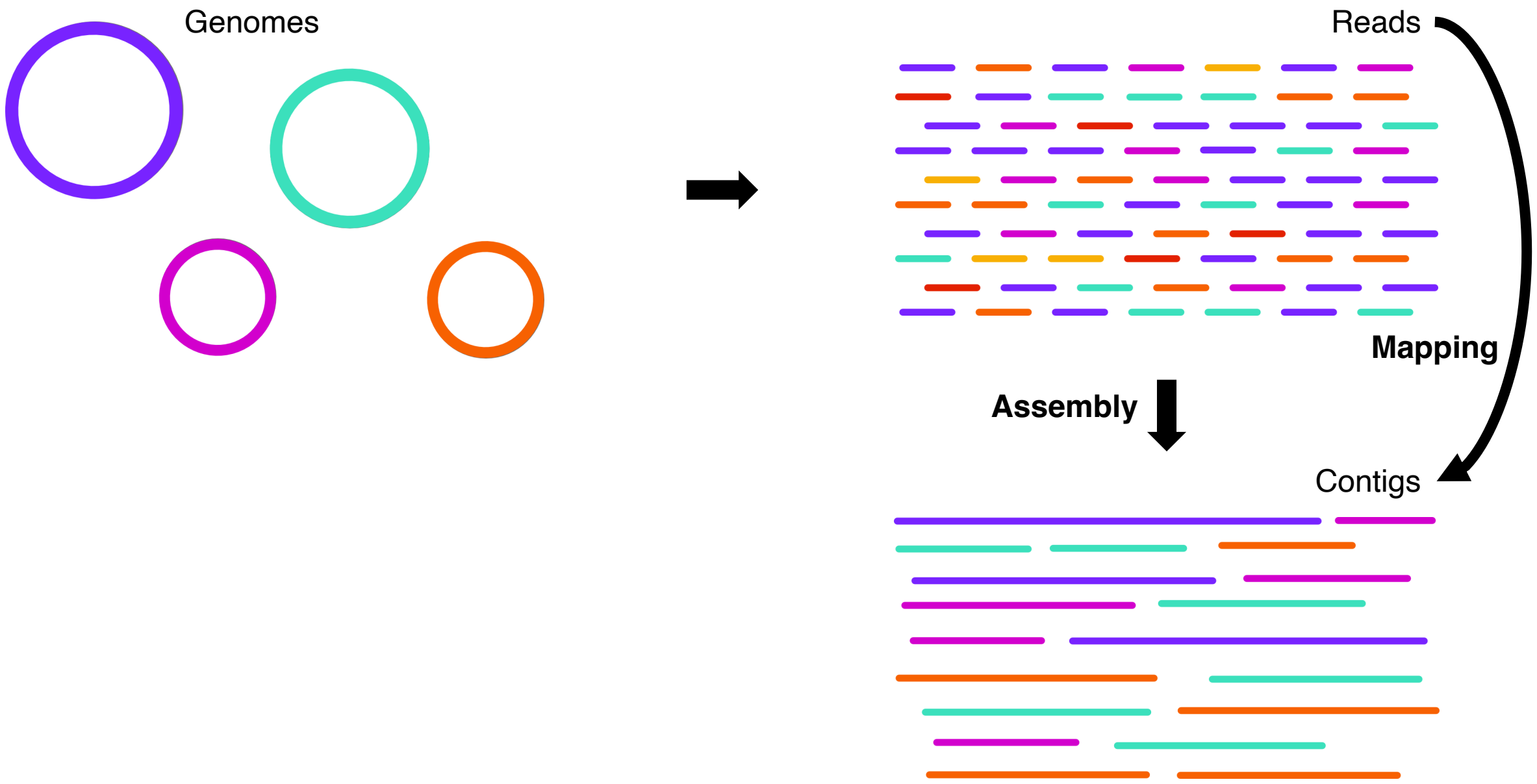
December 2023

Hands-on bioinformatics for microbial meta-omics (ENV 621)

Learning outcomes

- Lecture
  - Define mapping, coverage, and breadth
  - Understand different uses of read mapping
  - Understand important caveats of coverage

- Practical
  - Use read mapping to generate differential coverage
  - Evaluate total number of reads mapped to an assembly


Genomes
Reads
Mapping
Assembly
Contigs

Mapping

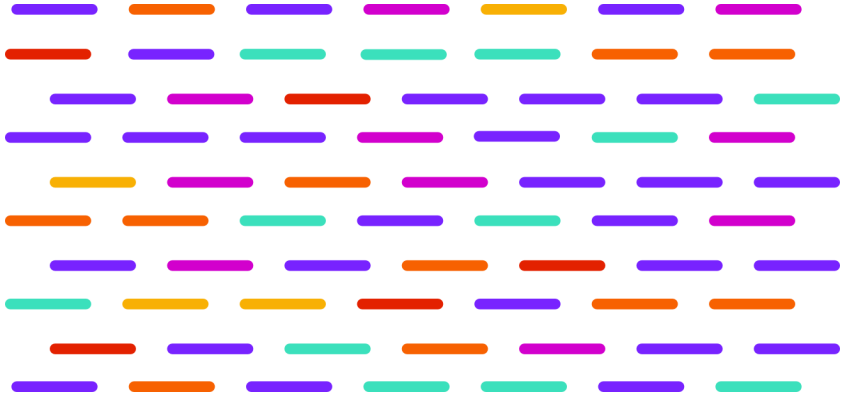
- Mapping = read recruitment

- Input
  - Reads (FASTQ or FASTA)
    - Metagenomic short reads
  - Reference sequences (FASTA)
    - **Contigs**, metagenome-assembled genomes, isolate genomes…

- Output
  - An alignment file (SAM or BAM)

Mapping

Reads



Mapping

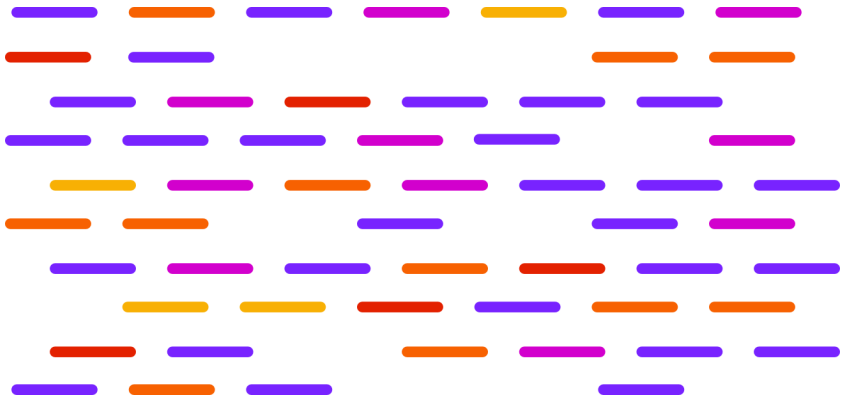
Contig #1

Contig #2

Contig #3

Mapping

Reads



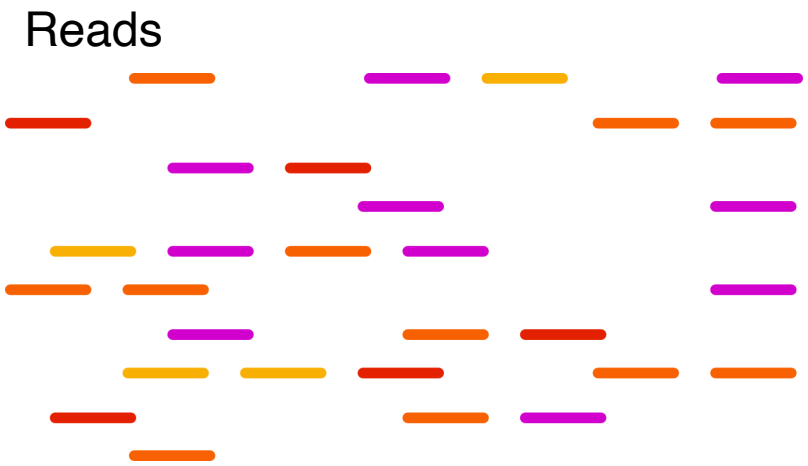
Mapping

Contig #1

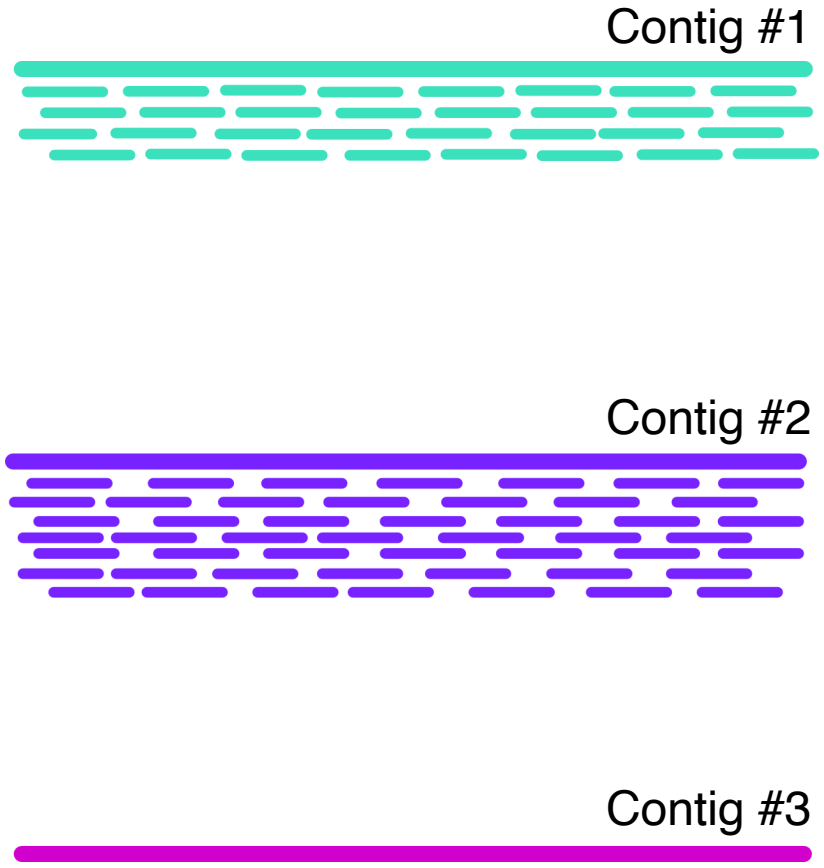
Contig #2

Contig #3

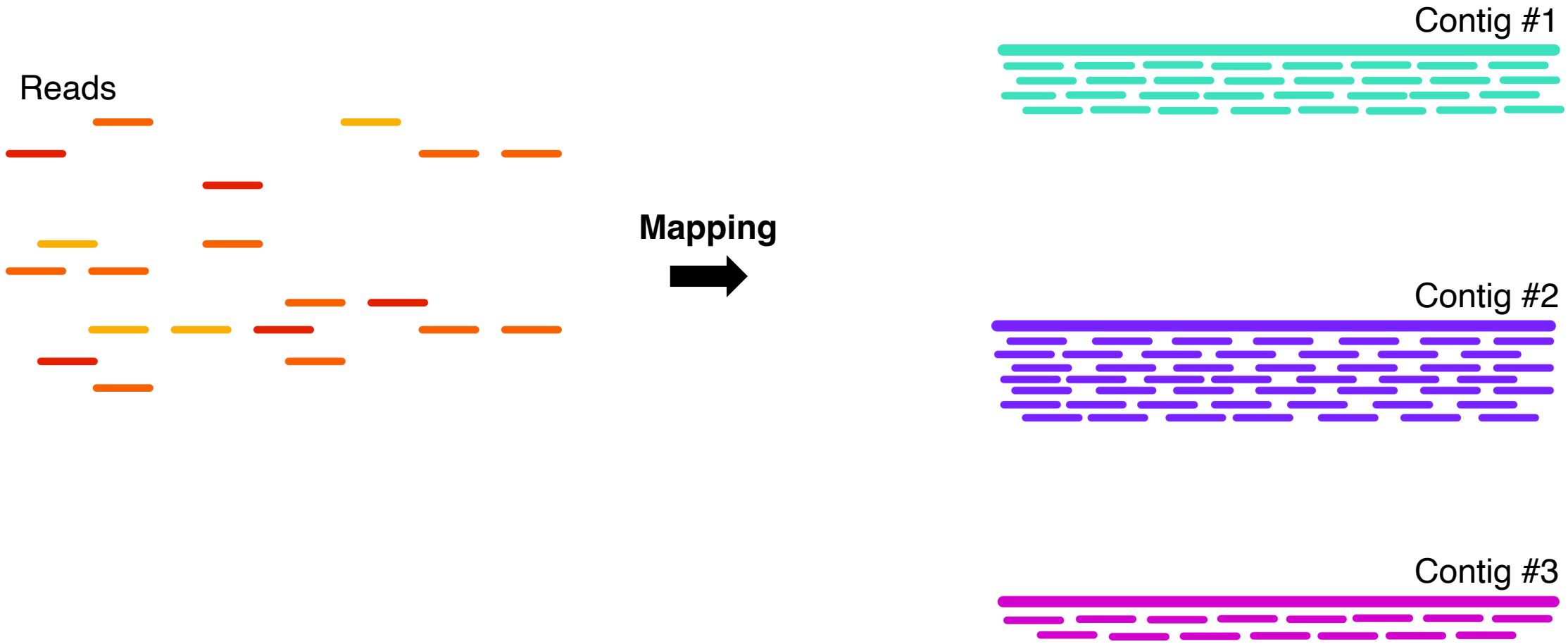
Mapping



Mapping

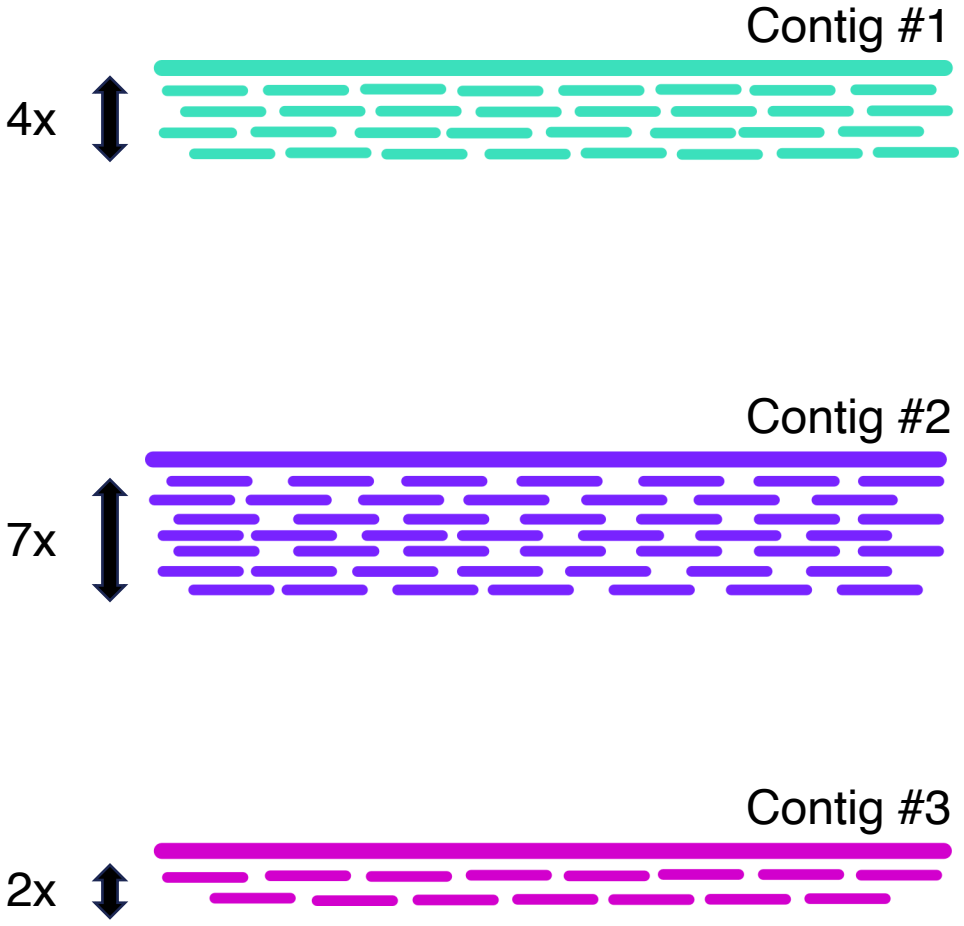


Mapping

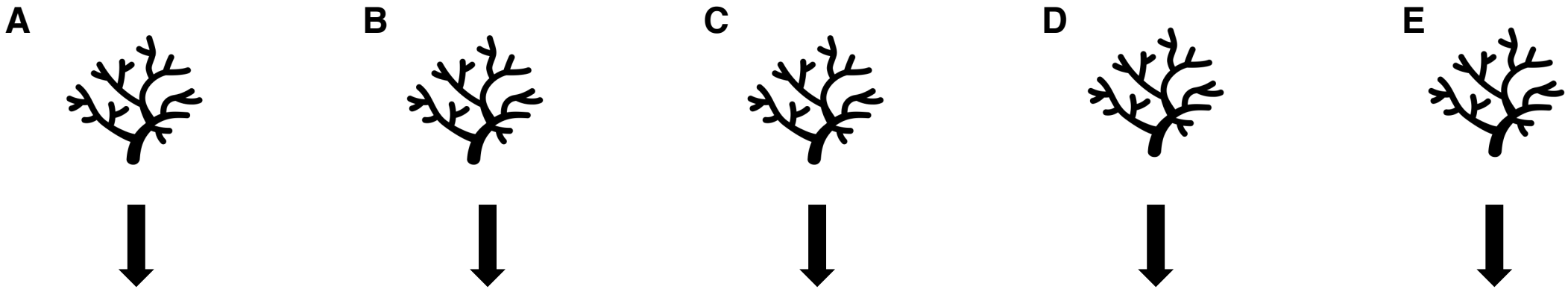


Coverage

- Calculated by mapping reads back to contigs



Cross-mapping



A reads > A contigs
B reads > A contigs
C reads > A contigs
D reads > A contigs
E reads > A contigs

5x bam files

A reads > B contigs
B reads > B contigs
C reads > B contigs
D reads > B contigs
E reads > B contigs

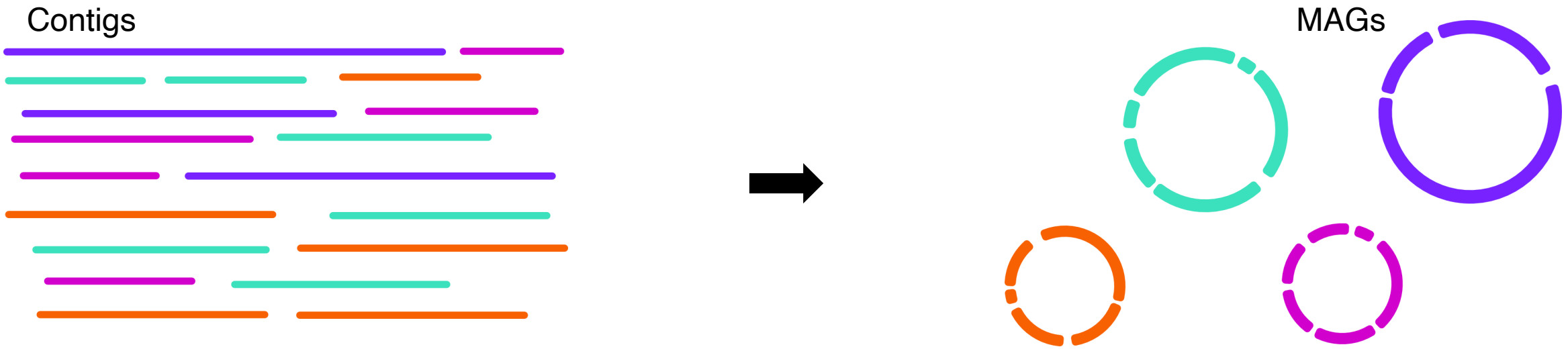
5x bam files

5 samples = 25 read alignments

- BBMap
- Bowtie2
- **Strobealign**

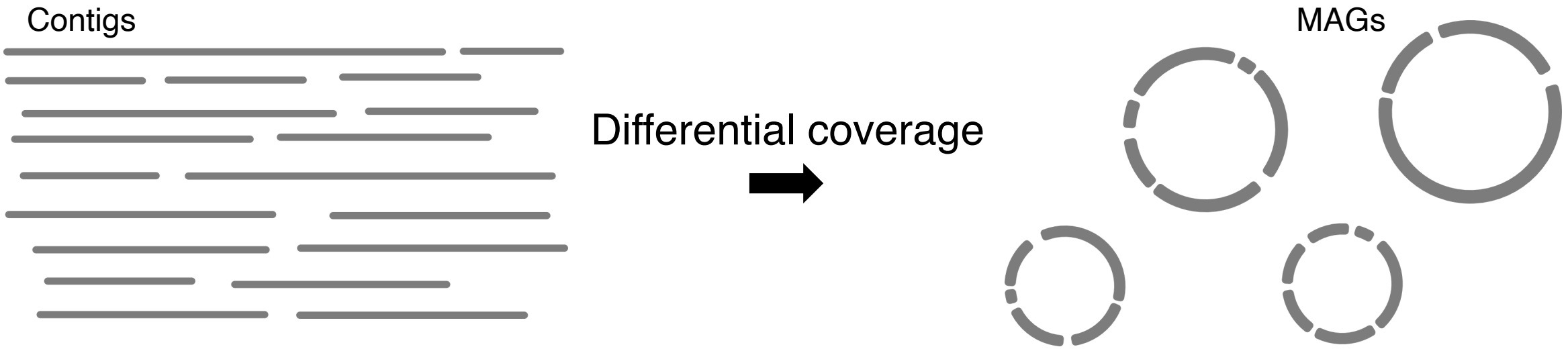
Why cross-map?

- To generate differential coverage for binning i.e., organising our contigs into **m**etagenome-**a**ssembled **g**enomes (MAGs)



Why cross-map?

- To generate differential coverage for binning i.e., organising our contigs into **m**etagenome-**a**ssembled **g**enomes (MAGs)

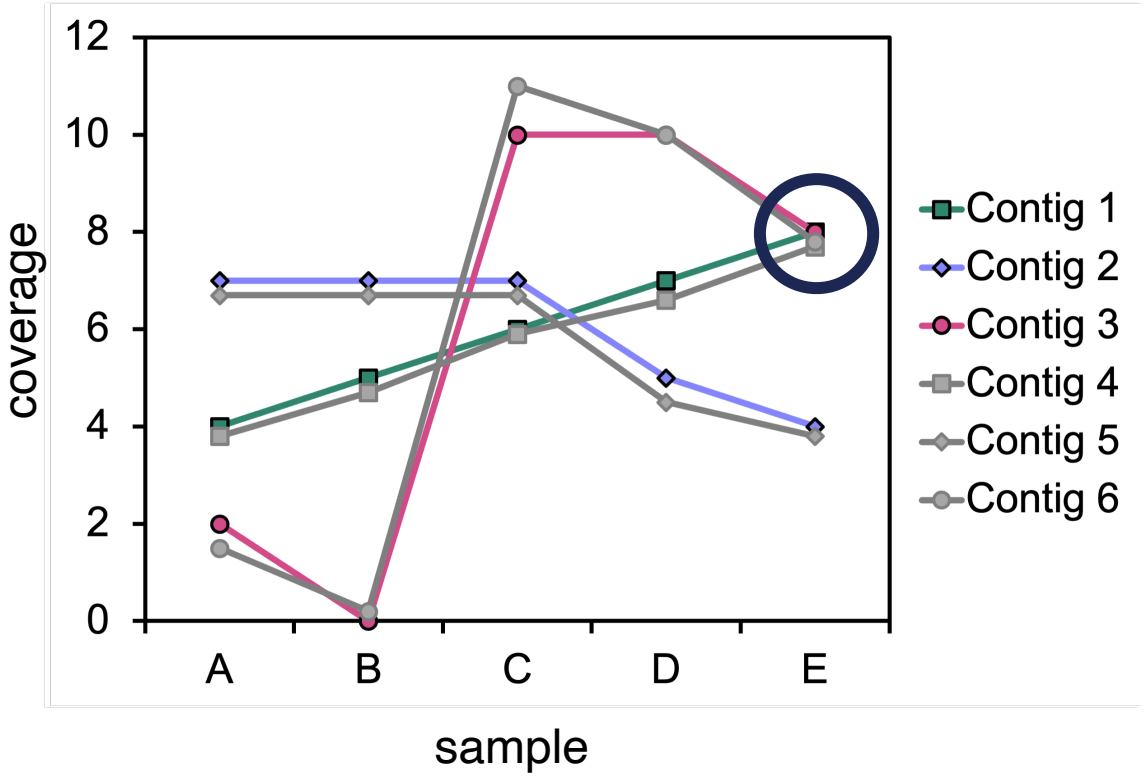


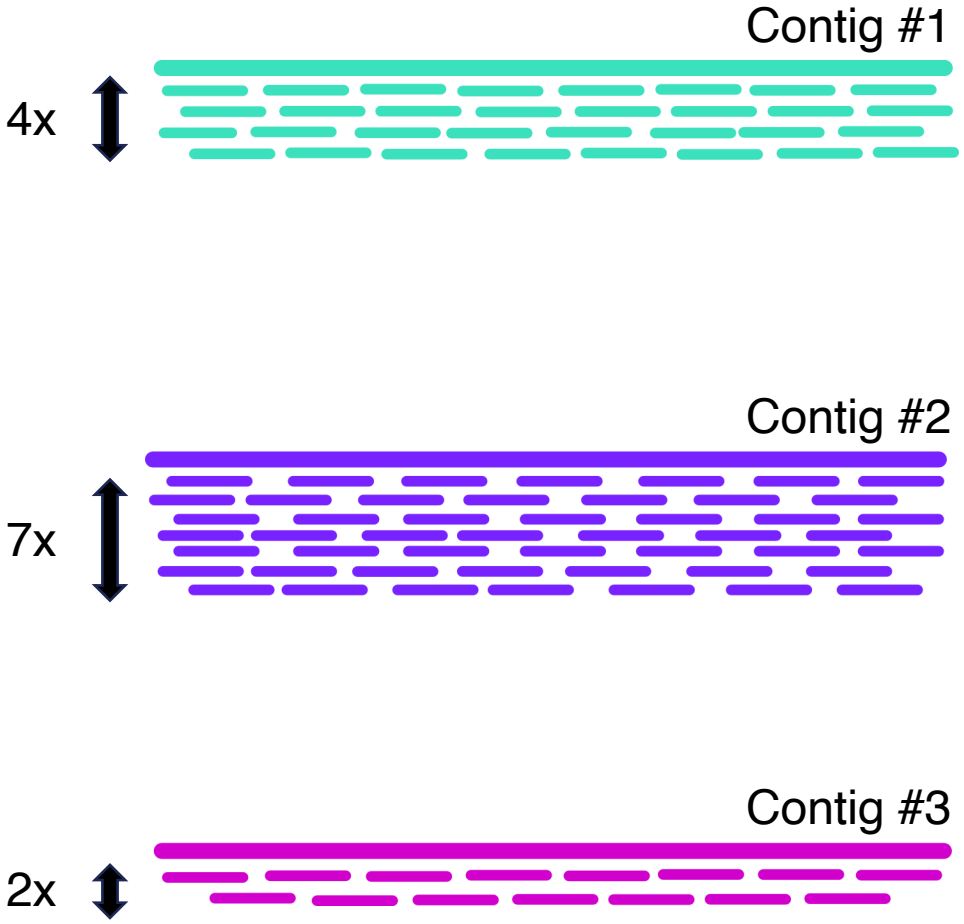
Differential coverage

- Calculated by mapping reads back to contigs over multiple samples

Differential coverage

- Calculated by mapping reads back to contigs over multiple samples





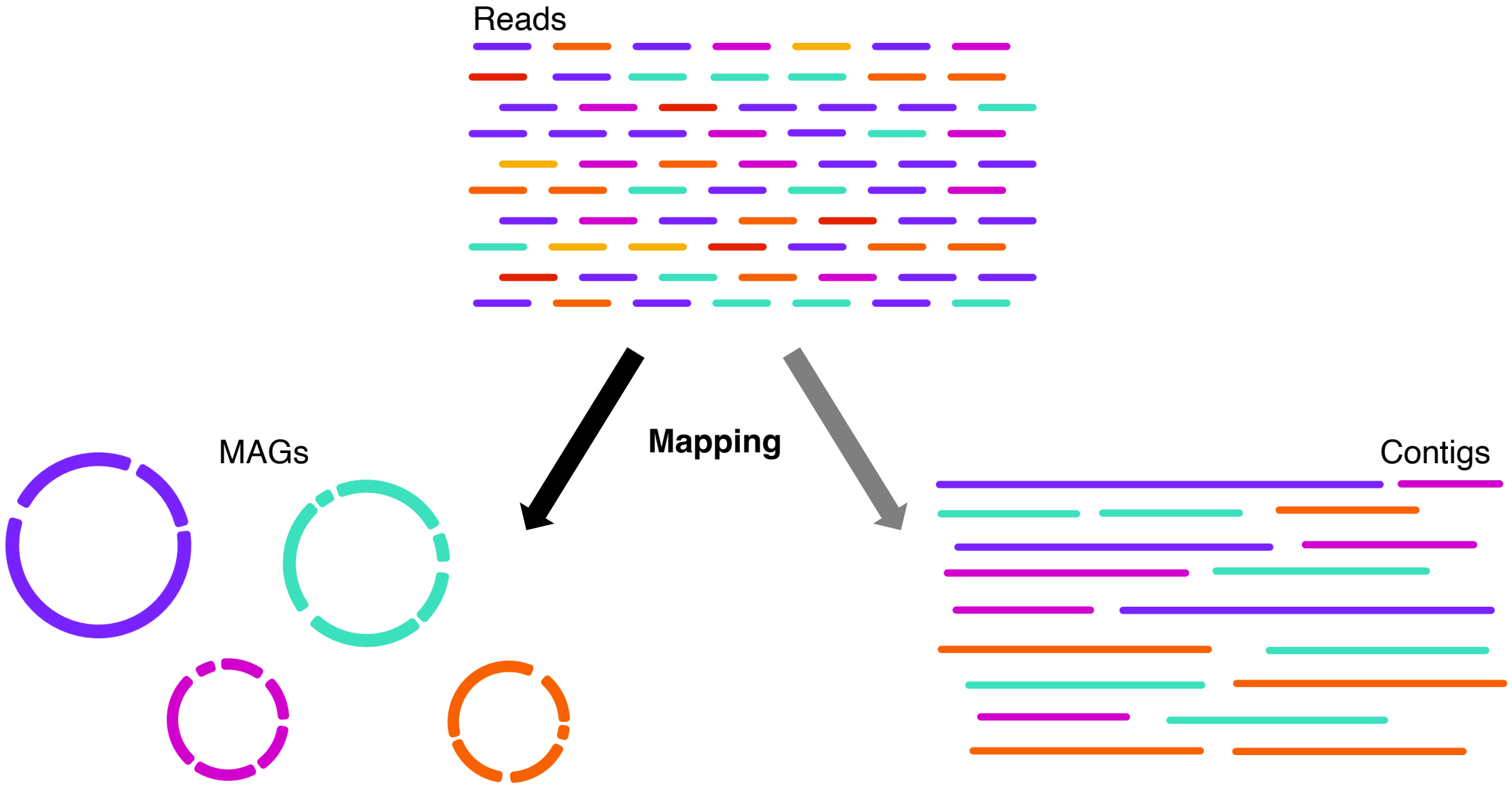
Mapping

- Mapping = read recruitment

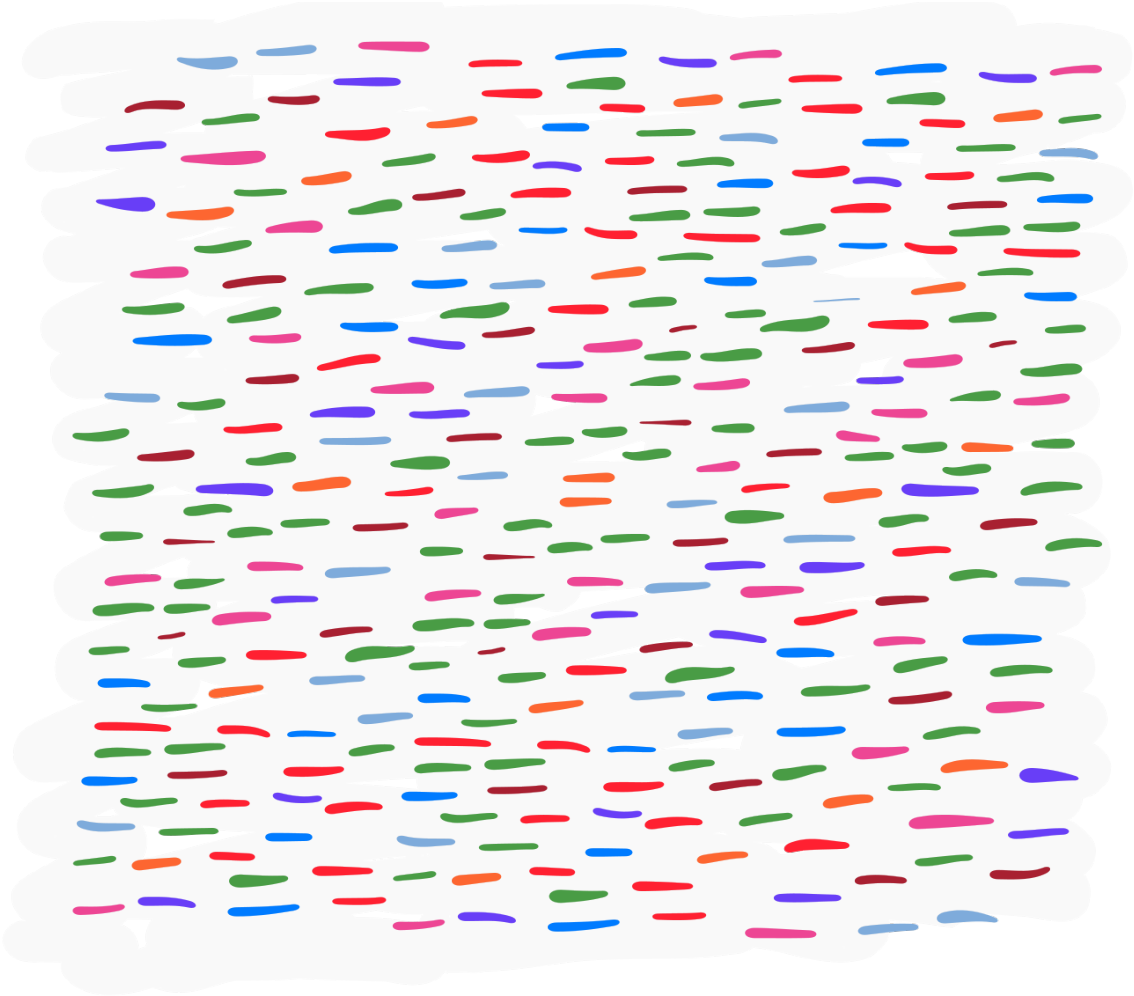
- Input
  - Reads (FASTQ or FASTA)
    - Metagenomic short reads
  - Reference sequences (FASTA)
    - Contigs, **metagenome-assembled genomes, isolate genomes…**

- Output
  - An alignment file (SAM or BAM)

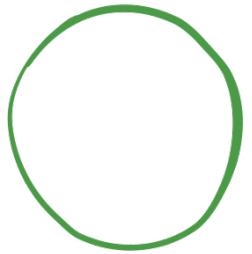
Mapping to a set of genomes



Mapping to a genome

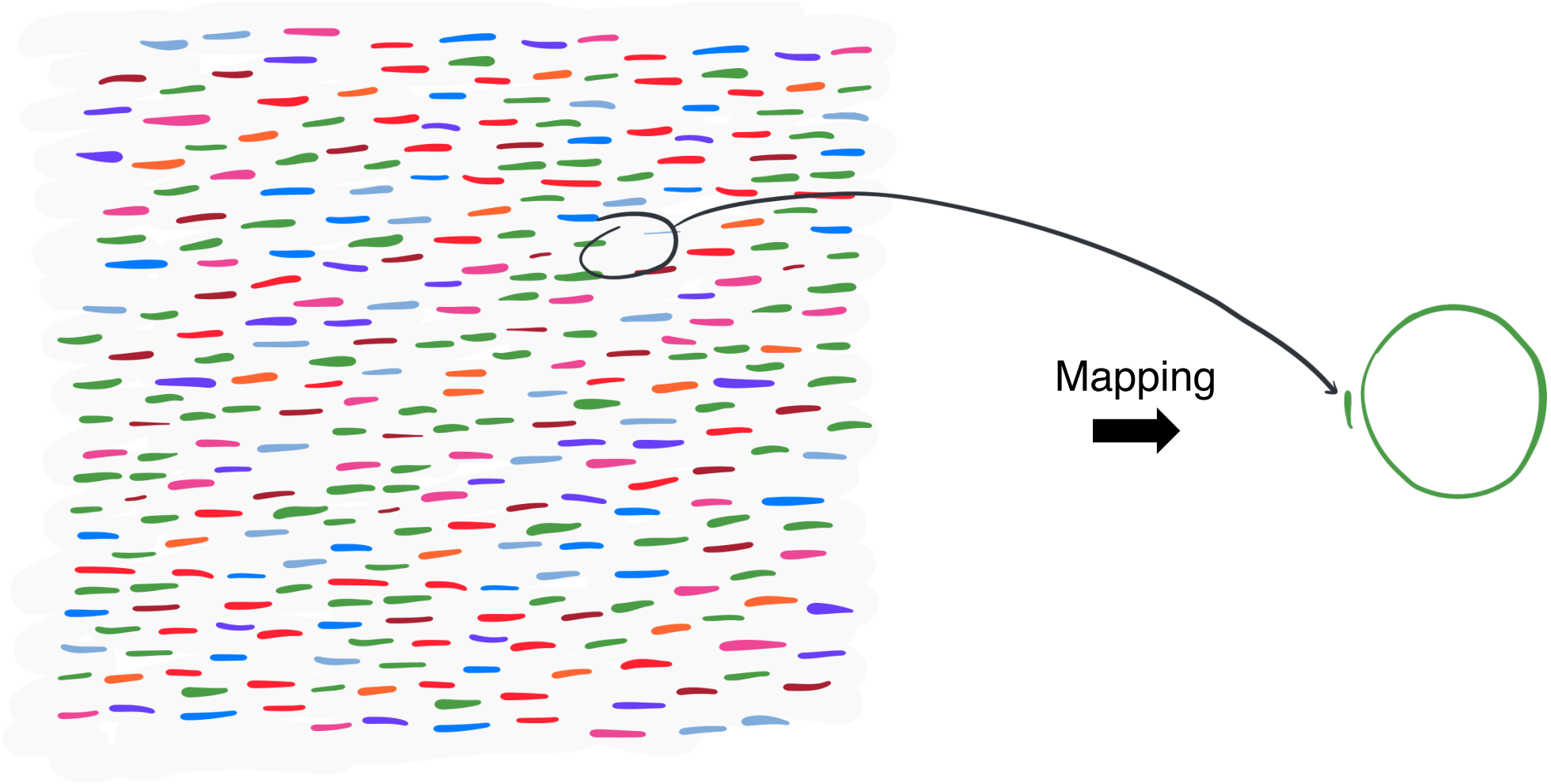


Reads



Genome

Mapping to a genome

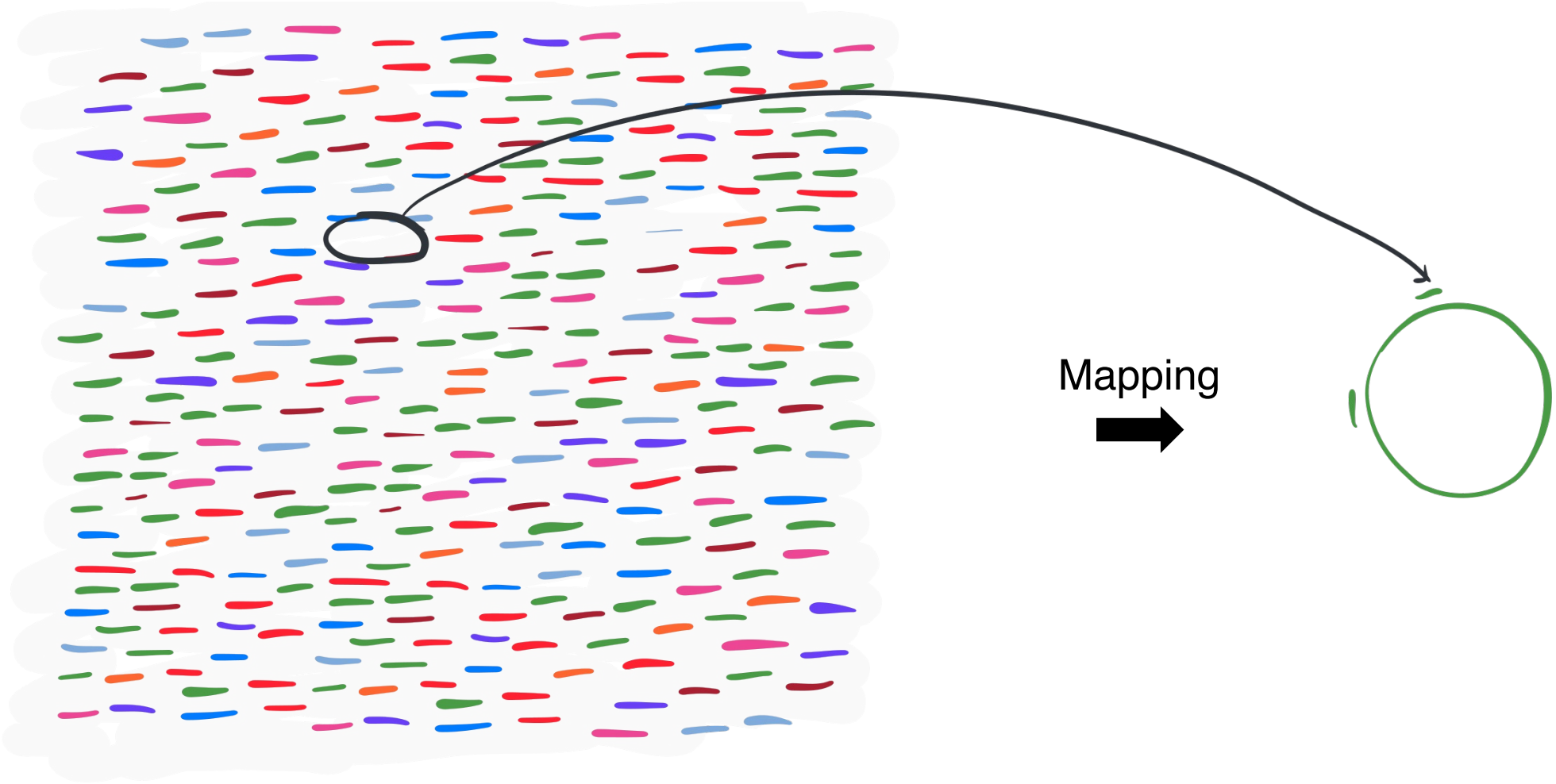


Reads

Genome

https://merenlab.org

Mapping to a genome

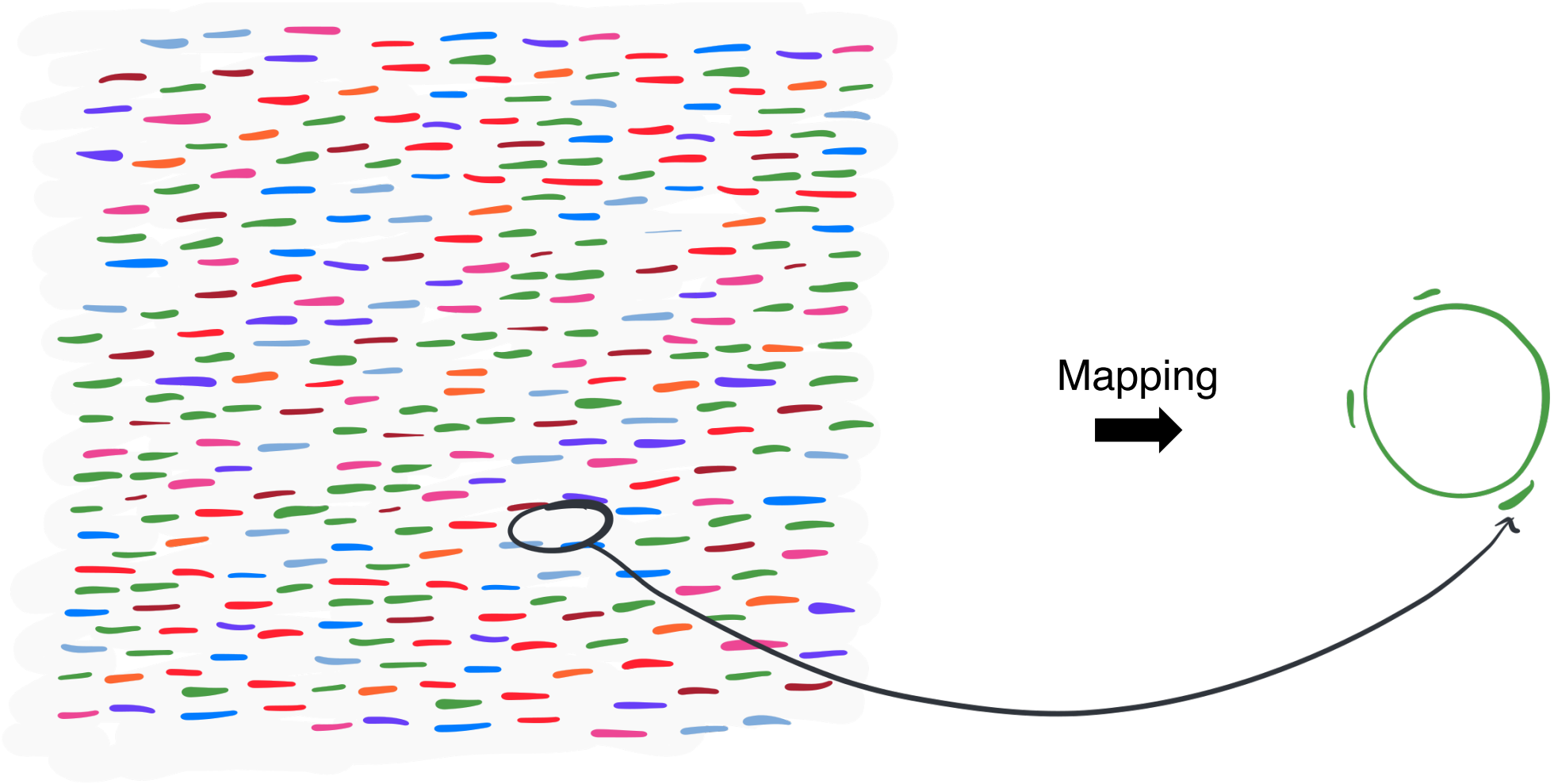


Reads

Genome

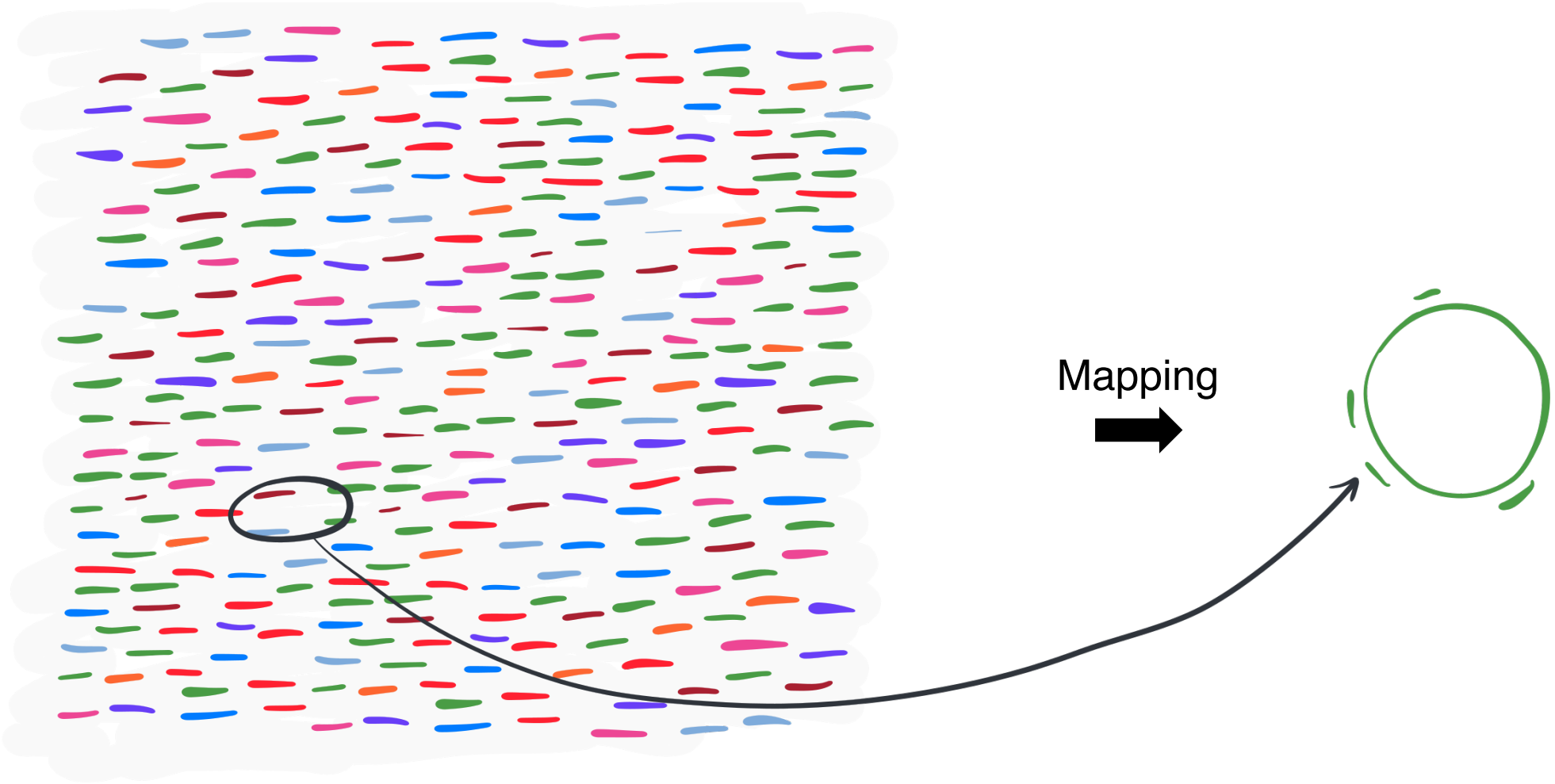
https://merenlab.org

Mapping to a genome



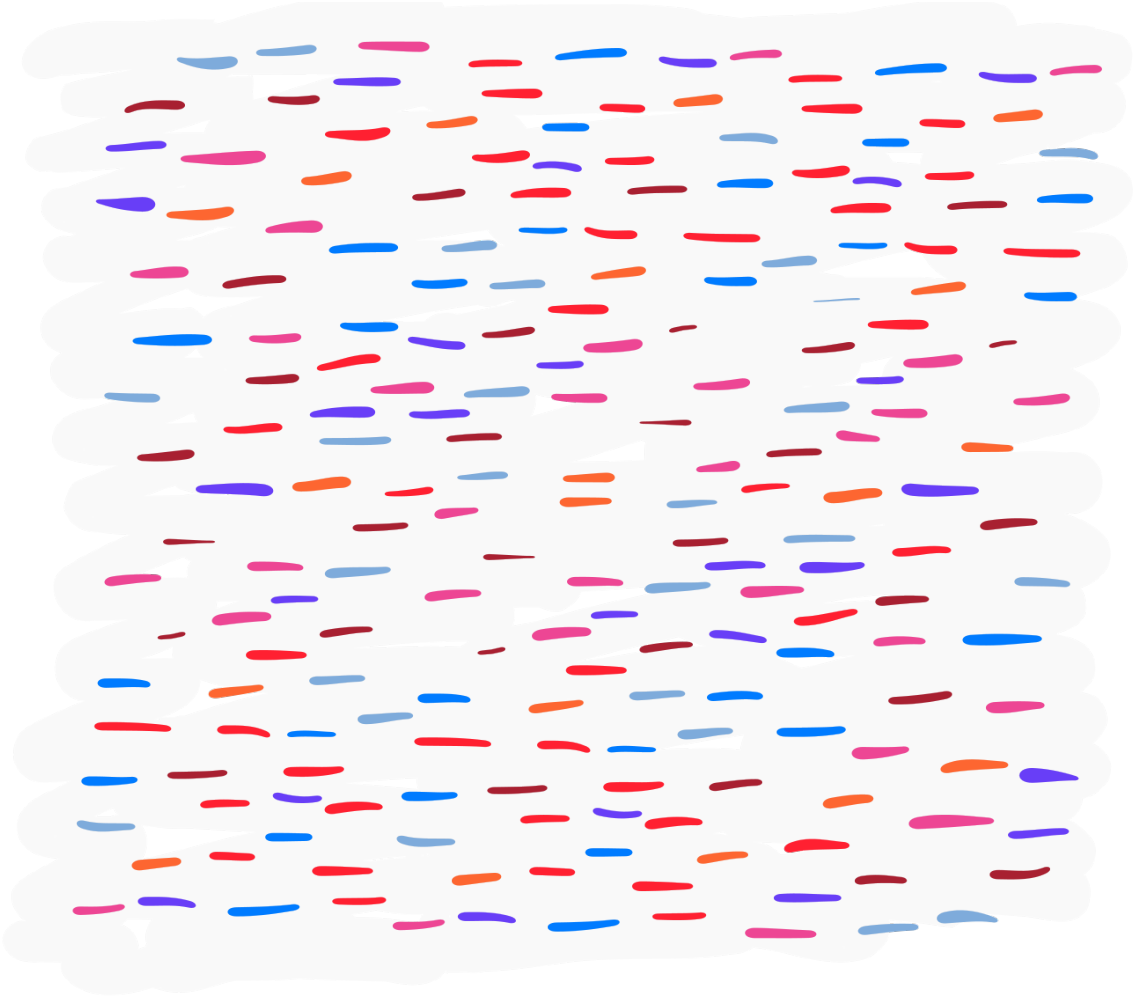
https://merenlab.org

Mapping to a genome



https://merenlab.org

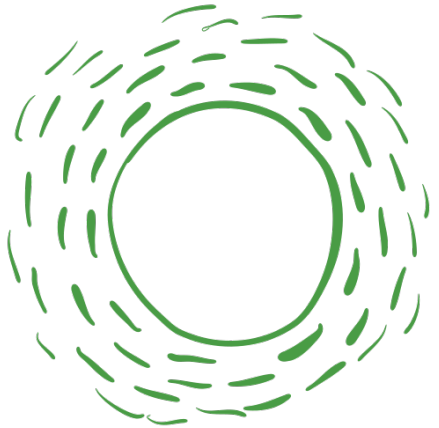
Mapping to a genome



Reads

Is the genome of interest in your environment of interest…?

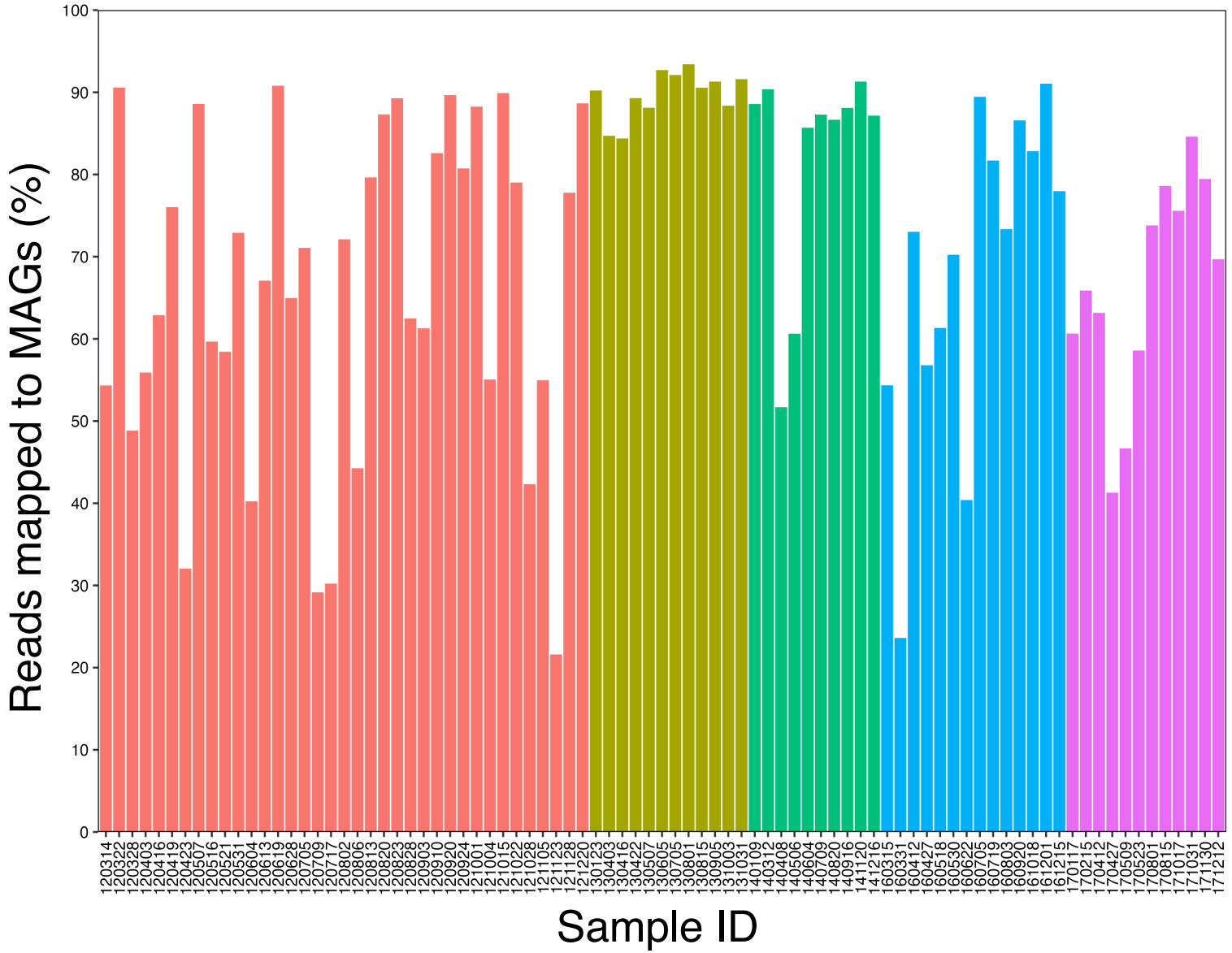
Mapping



Genome

https://merenlab.org

Mapping to a set of genomes



How representative is my genome set…?

Data: x 81

x 568

Soil ~20-40%
Human microbiome 60-80%
Simple, well defined communities 90%+

Coverage and breadth

Coverage: The average depth of coverage.

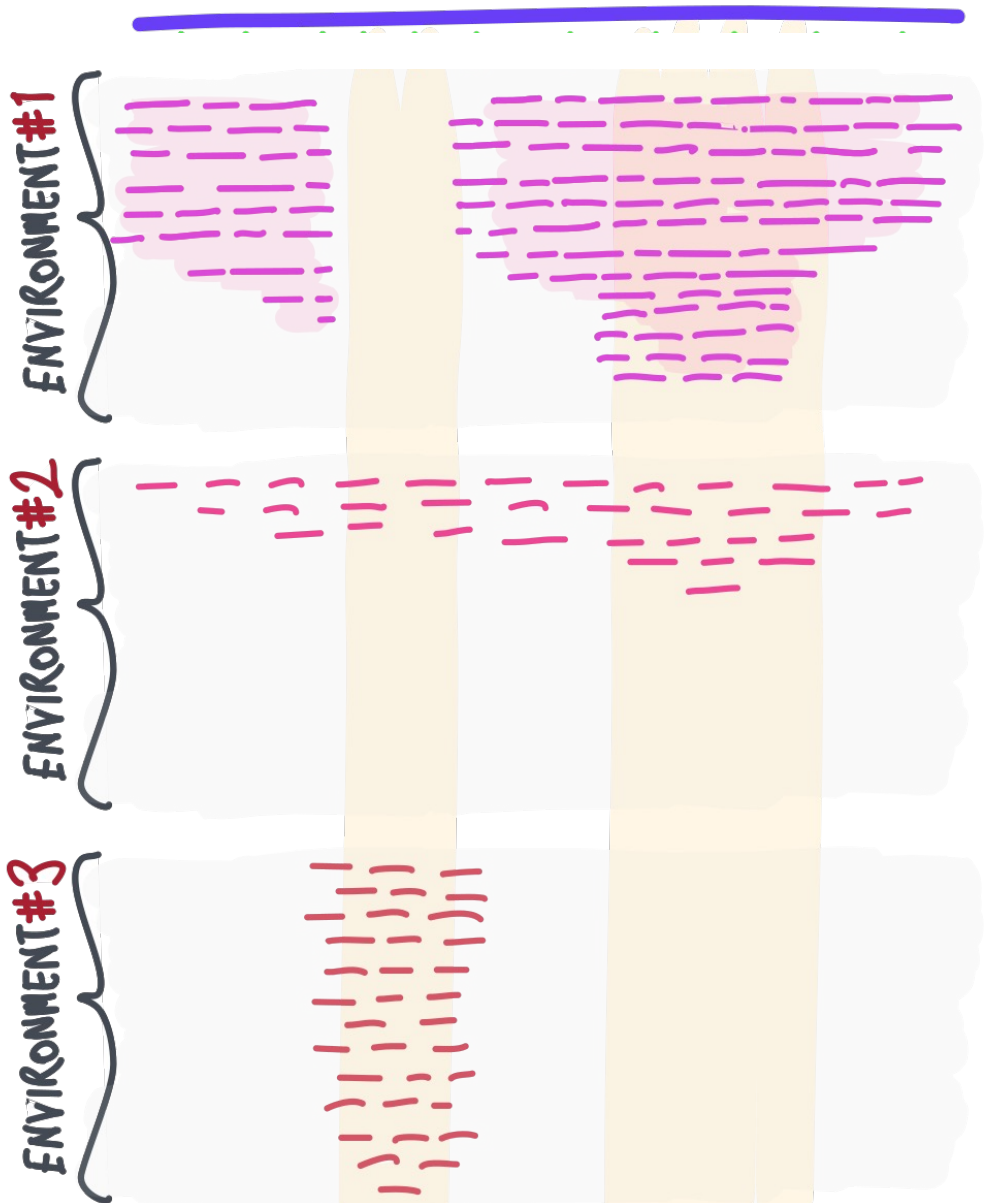
Coverage and breadth

Coverage: 10x

Coverage: The average depth of coverage.

https://merenlab.org

Coverage and breadth



Coverage: 10x

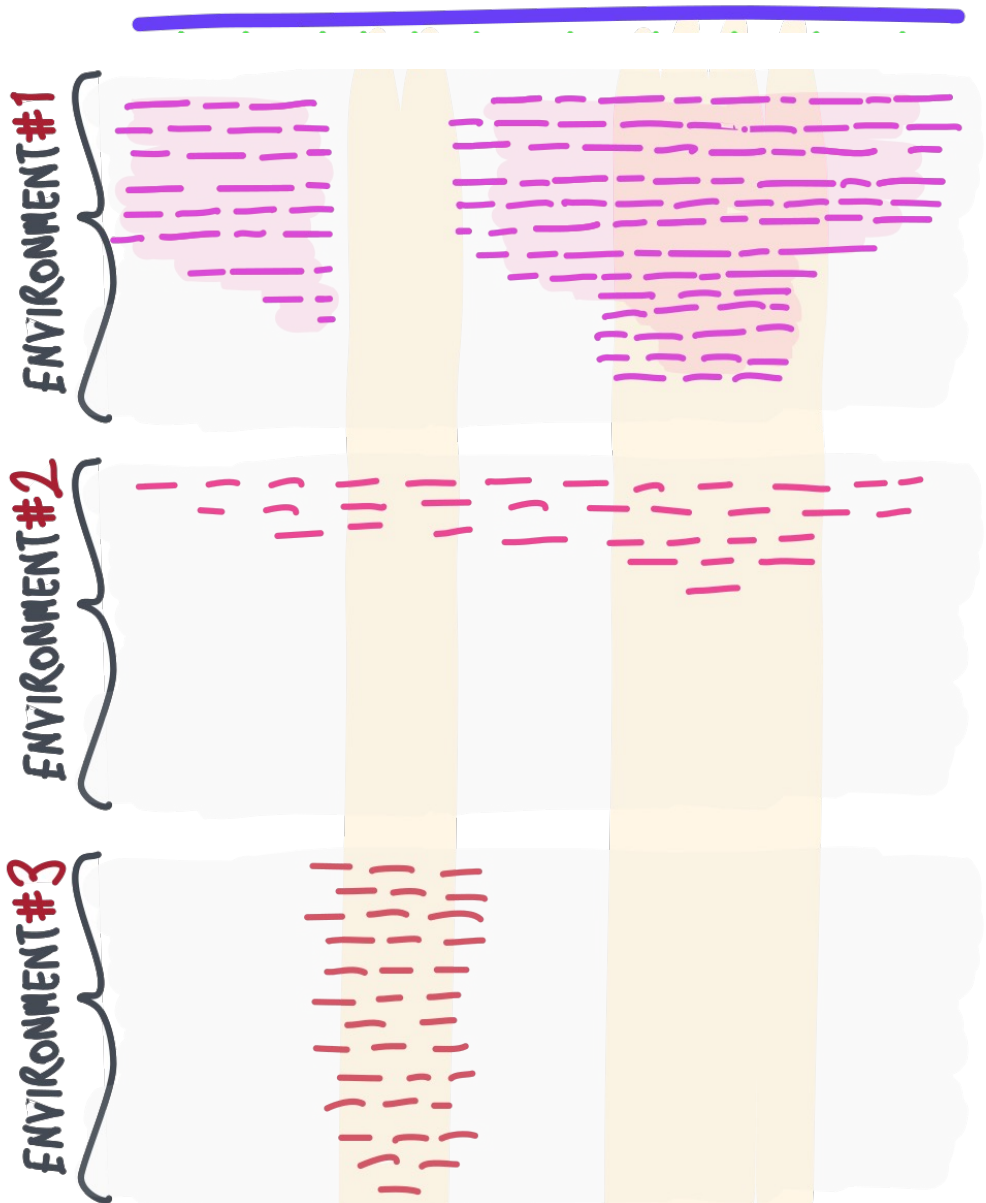
Coverage: 3x

Coverage: 3x

Coverage: The average depth of coverage.

https://merenlab.org

Coverage and breadth



Coverage: 10x

Coverage: 3x

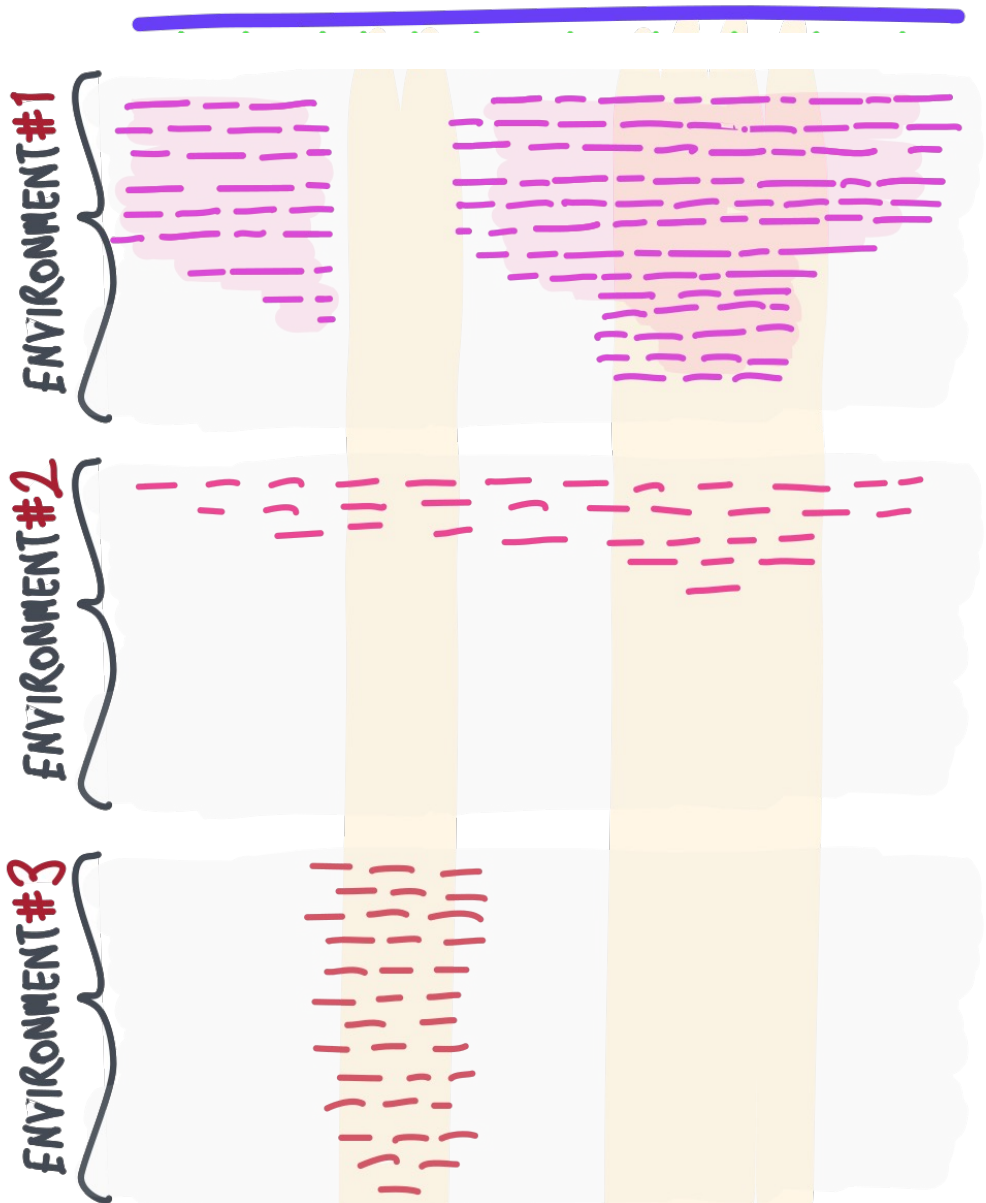
Coverage: 3x

Coverage: The average depth of coverage.

Breadth: The percentage of bases in the scaffold that are covered by at least a single read.

https://merenlab.org

Coverage and breadth



Coverage: 10x
Breadth: 0.8

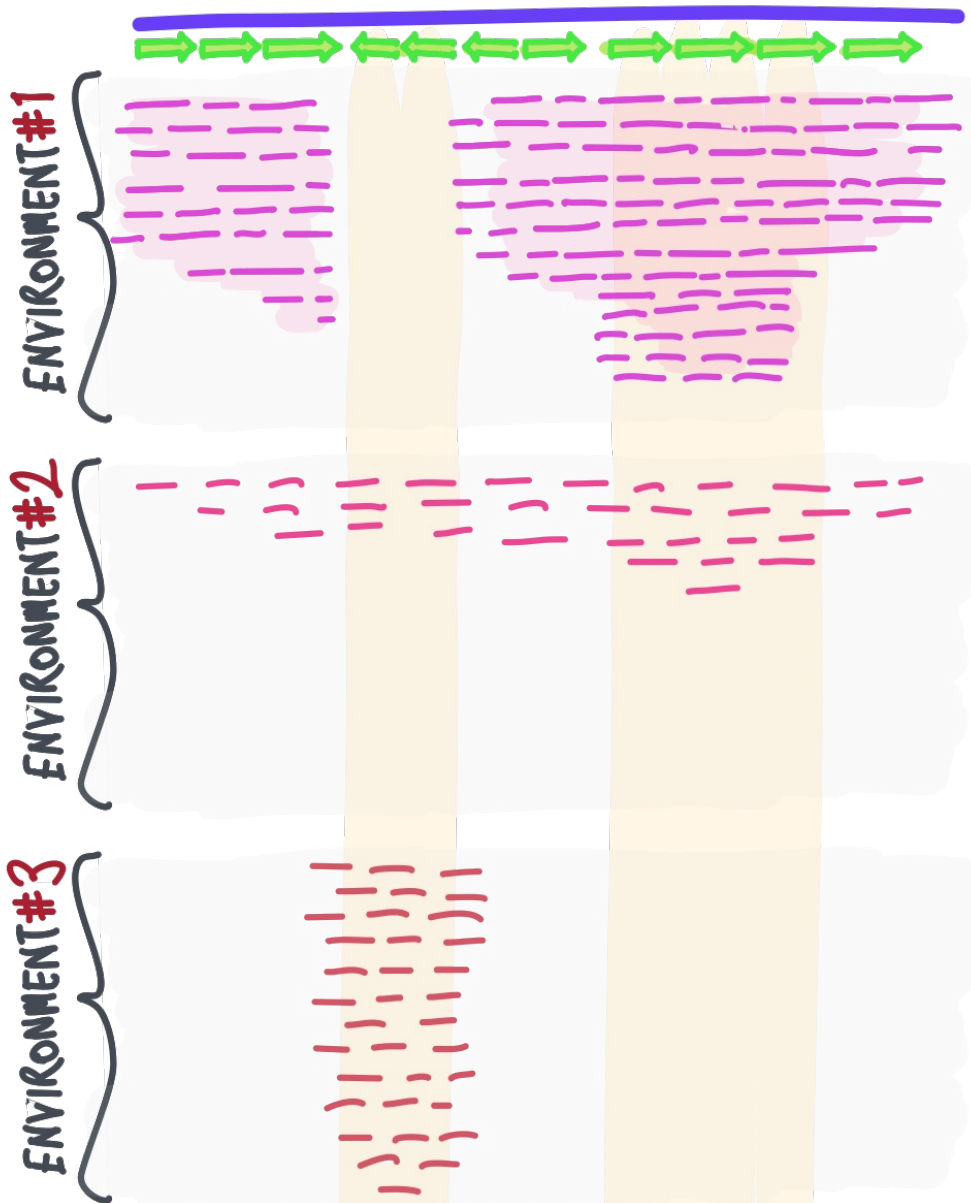
Coverage: 3x
Breadth: 1.0

Coverage: 3x
Breadth: 0.2

Coverage: The average depth of coverage.

Breadth: The percentage of bases in the scaffold that are covered by at least a single read.

Coverage and breadth



Coverage: 10x
Breadth: 0.8

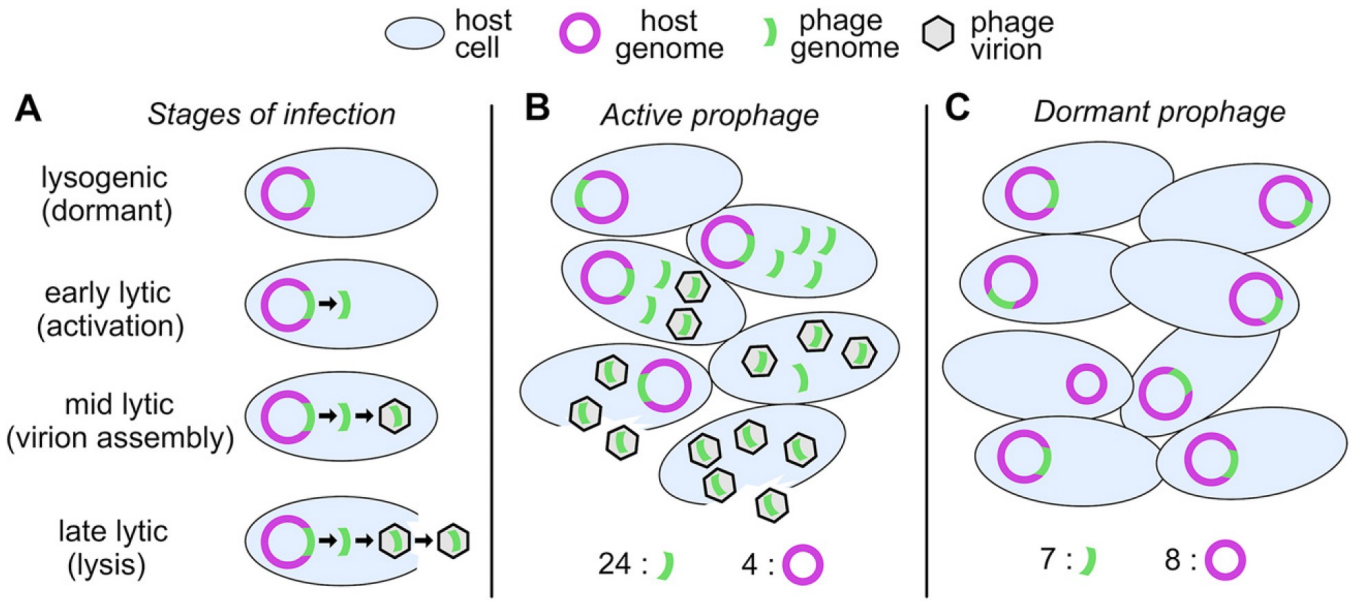
Coverage: 3x
Breadth: 1.0

Coverage: 3x
Breadth: 0.2

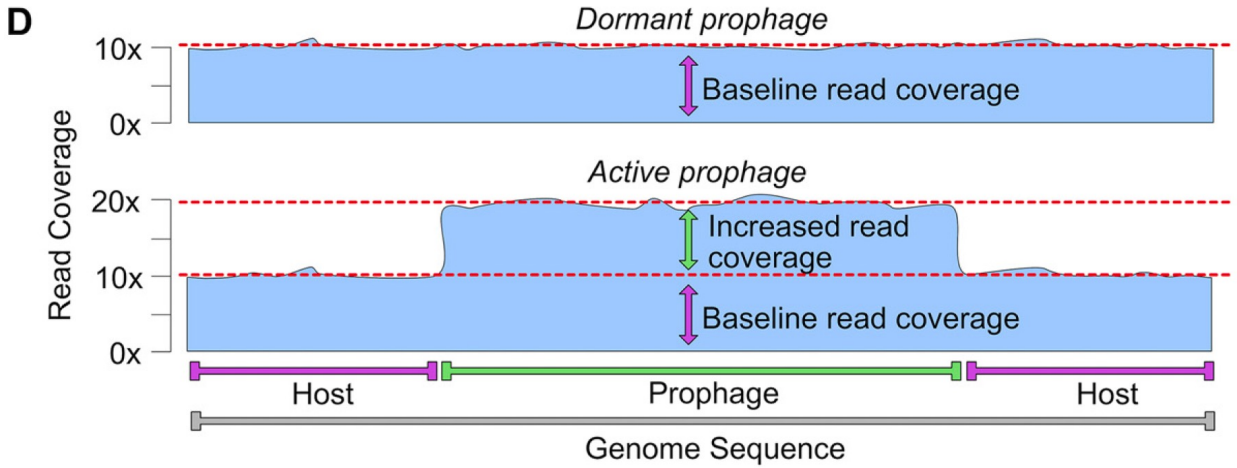
Coverage: The average depth of coverage.

Breadth: The percentage of bases in the scaffold that are covered by at least a single read.

Active prophages in metagenomes



- Prophages persist as dormant components of host cells (lysogenic stage) before activating and lysing the host (lytic stage).



PropagAtE
Prophage Activity Estimator

Kieft and Anantharaman, *mSystems* 2022

Learning outcomes

- Lecture
  - Define mapping, coverage, and breadth
  - Understand different uses of read mapping
  - Understand important caveats of coverage

- Practical
  - Use read mapping to generate differential coverage
  - Evaluate total number of reads mapped to an assembly

Popular read mapping tools

- BBMap https://jgi.doe.gov/data-and-tools/software-tools/
- Bowtie2 https://jgi.doe.gov/data-and-tools/software-tools/bbtools/
- **Strobealign** https://github.com/ksahlin/strobealign

Visualising read coverage

- Integrative genomics viewer (IGV) https://igv.org

- Anvi'o: An open-source, community-driven **an**alysis and **vi**sualization platform for microbial **'o**mics https://merenlab.org/tutorials/read-recruitment/