Lecture 19
Genome assembly

Course: Practical Bioinformatics (BIOL 4220)
Instructor: Michael Landis
Email: michael.landis@wustl.edu

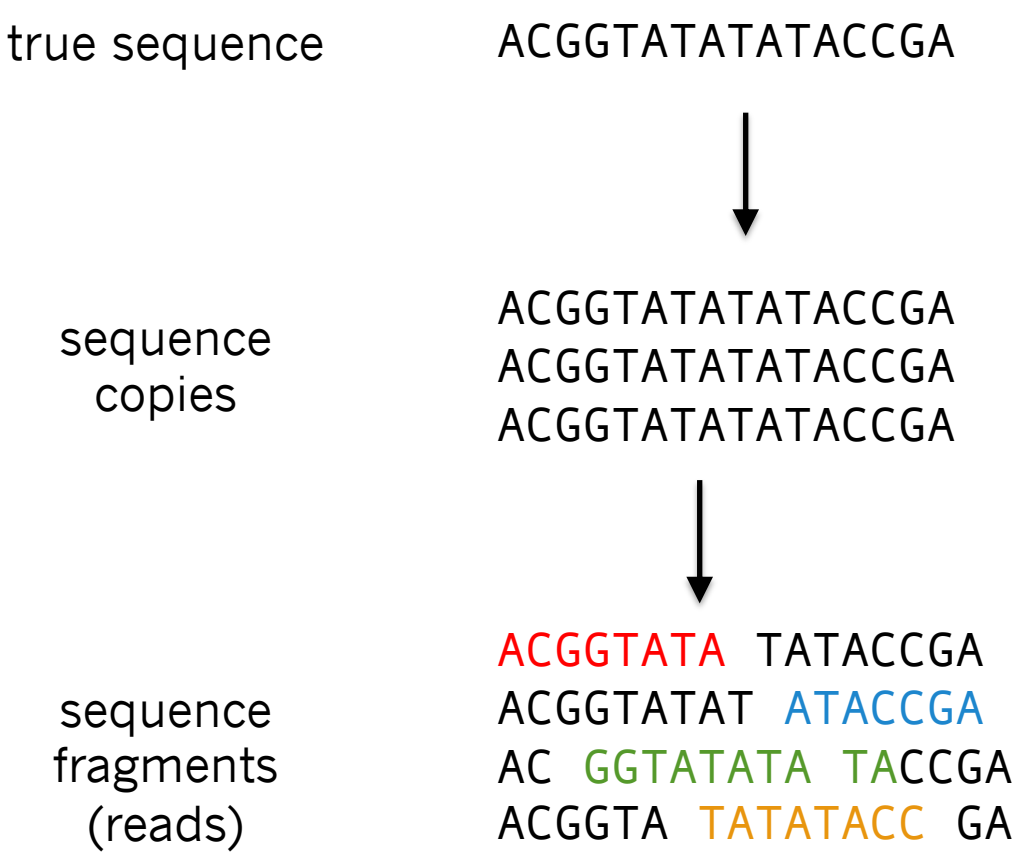
Lecture 19 outline

Last time: jupyter, matplotlib

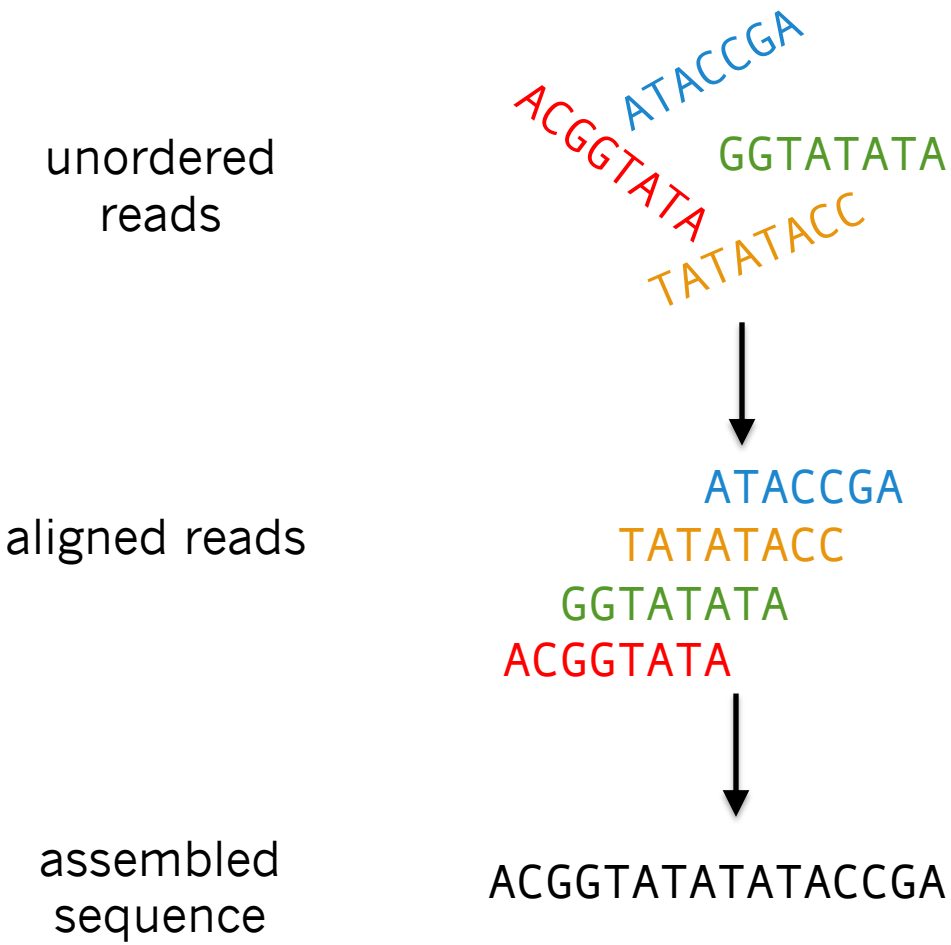
This time: genome assembly

- genome sequences
- genome sizes
- genome assembly

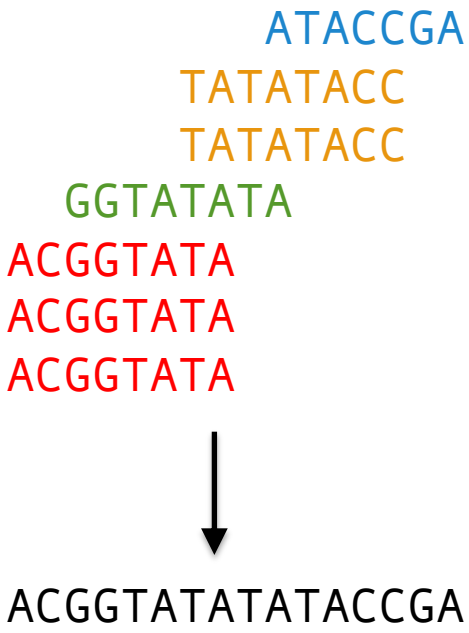
Sequencing


true sequence
ACGGTATATATACCGA
sequence copies
ACGGTATATATACCGA
ACGGTATATATACCGA
ACGGTATATATACCGA
sequence fragments (reads)
ACGGTATA TATACCGA
ACGGTATAT ATACCGA
AC GGTATATA TACCGA
ACGGTA TATATACC GA

Assembly


unordered reads
ATACCGA
ACGGTATA
GGTATATA
TATATACC
aligned reads
ATACCGA
TATATACC
GGTATATA
ACGGTATA
assembled sequence
ACGGTATATATACCGA

Assembly



Would be easy if we knew how reads were aligned

We would retrieve the original genome sequence with no effort

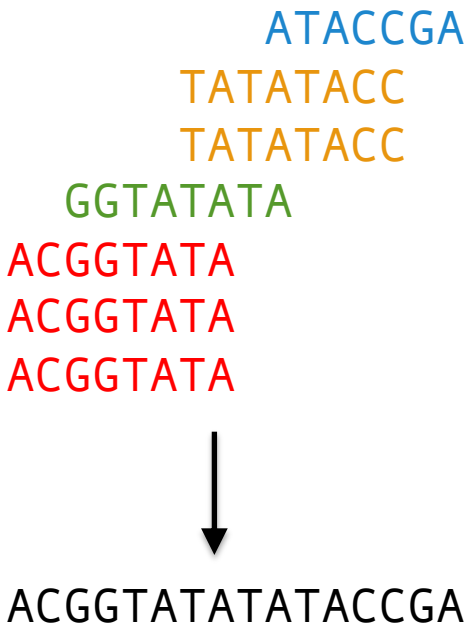
Instead, we have an unordered and unaligned bag of reads

True genome

Sequenced reads

Assembled genome

How do we assemble reads?



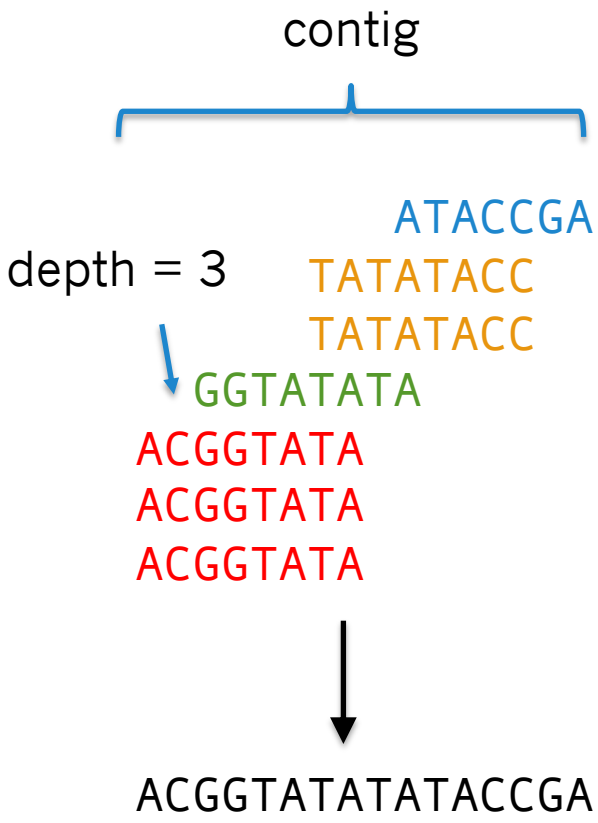
Can we do global pairwise alignments for each pair of reads?

Let's make a block of contiguously mapped reads (contig)

Reads can align to any contig

Read mapped to contig with best score

Basic unit of assembly



We want high-coverage contigs

depth = # reads mapped for one site

avg. coverage = $\frac{\text{\# mapped sites}}{\text{contig size}}$

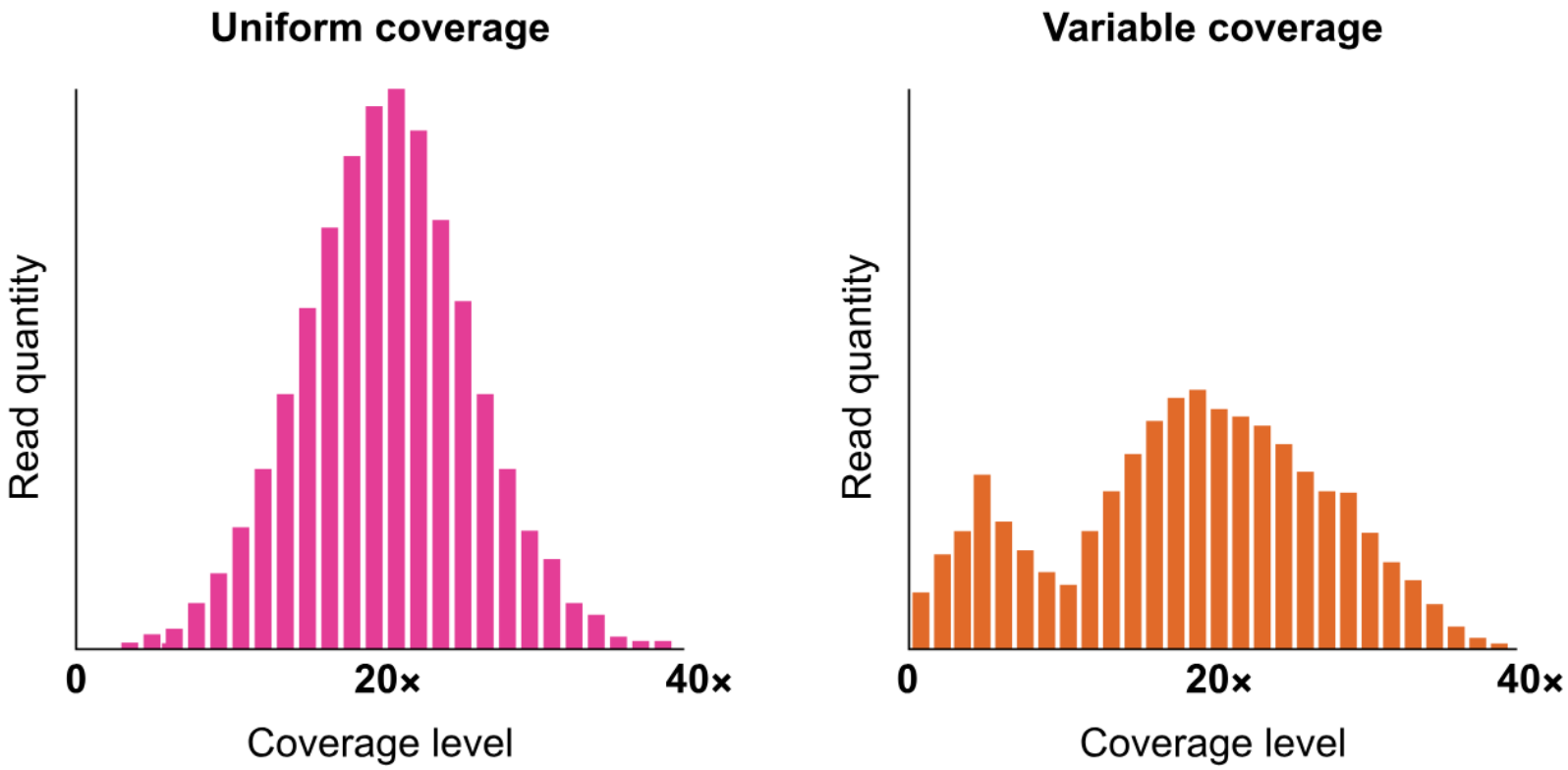
est. coverage = $\frac{\text{\# reads * read length}}{\text{genome size}}$

avg. coverage = $\frac{(8 + 8 + 8 + 8 + 8 + 8 + 7)}{16}$

Short read dataset sizes

How many 150 bp length reads needed for 30x coverage?

Species	**#bp**	**#reads**
SARS-CoV-2	$2x10^4$	$4x10^3$
E. coli	$4.5x10^6$	$9x10^5$
Human	$3.2x10^9$	$6.4x10^8$
Fern	$1.6x10^{11}$	$3.2x10^{10}$



https://www.pacb.com/blog/sequencing-101-sequencing-coverage/

Assembly problem

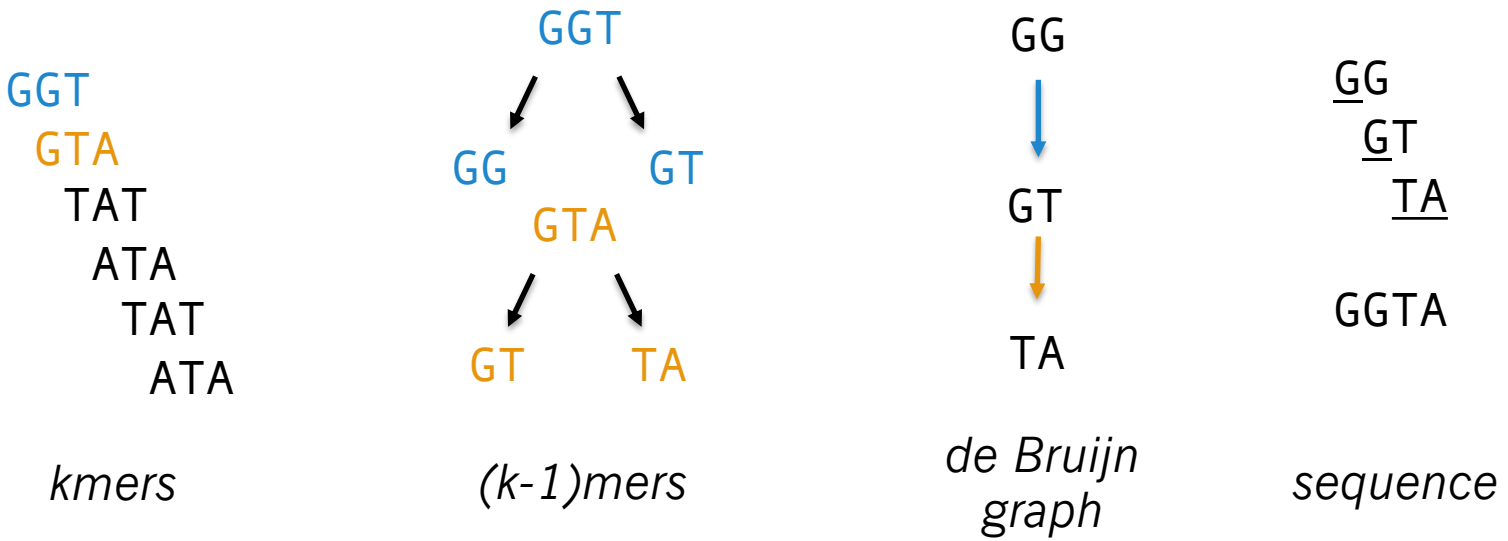
Naive assembly would require N^2 pairwise alignments.

Not possible for short read datasets!

e.g. 10^{18} alignments for 10^9 reads

de Bruijn graph

- Choose kmer length (often 40 < k < 100)
- Make left and right (k-1)mers for each kmer
- Add node for (k-1)mer if it doesn't exist
- Add edge from left (k-1)mer to right (k-1)mer

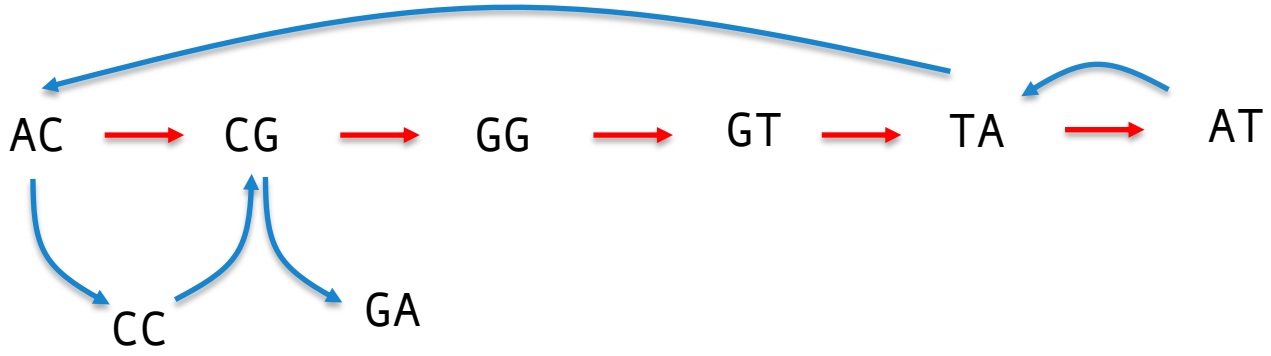


Graph construction

ACGGTATA

ACG
CGG
GGT
GTA
TAT
ATA

Graph construction



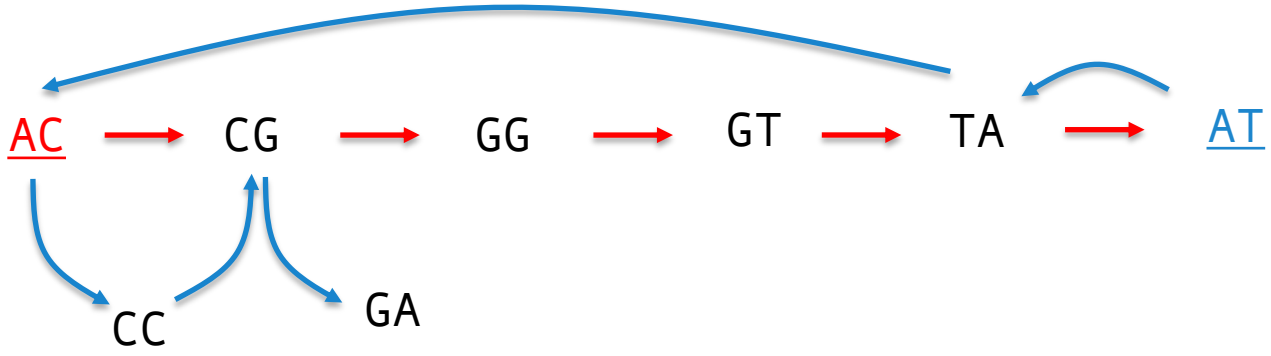
ACGGTATA

ACG
CGG
GGT
GTA
TAT
ATA

ATACCGA

ATA
TAC
ACC
CCG
CGA

Graph traversal



Eulerian path: visit all nodes using each edge once

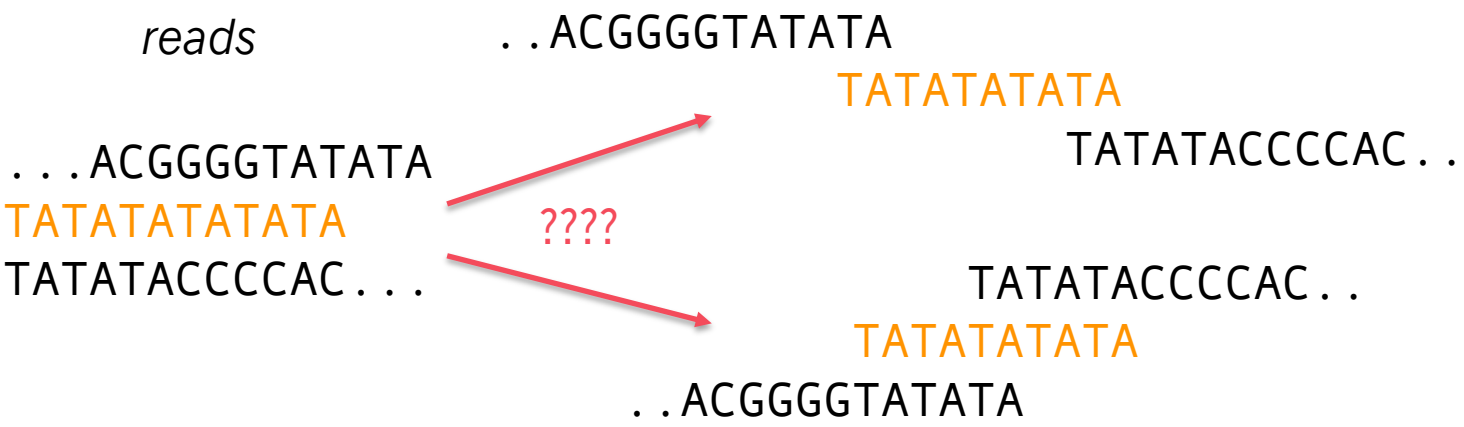
Starting at AC

ACGGTATACCGA

Starting at AT

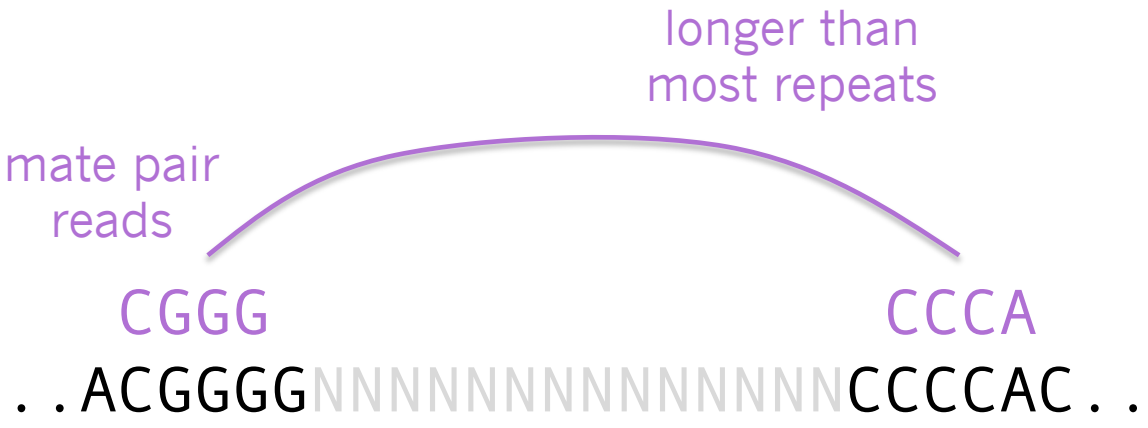
ATACCGA
or
ATACCGGTAT

Repeat regions



contigs form where assembly is ambiguous

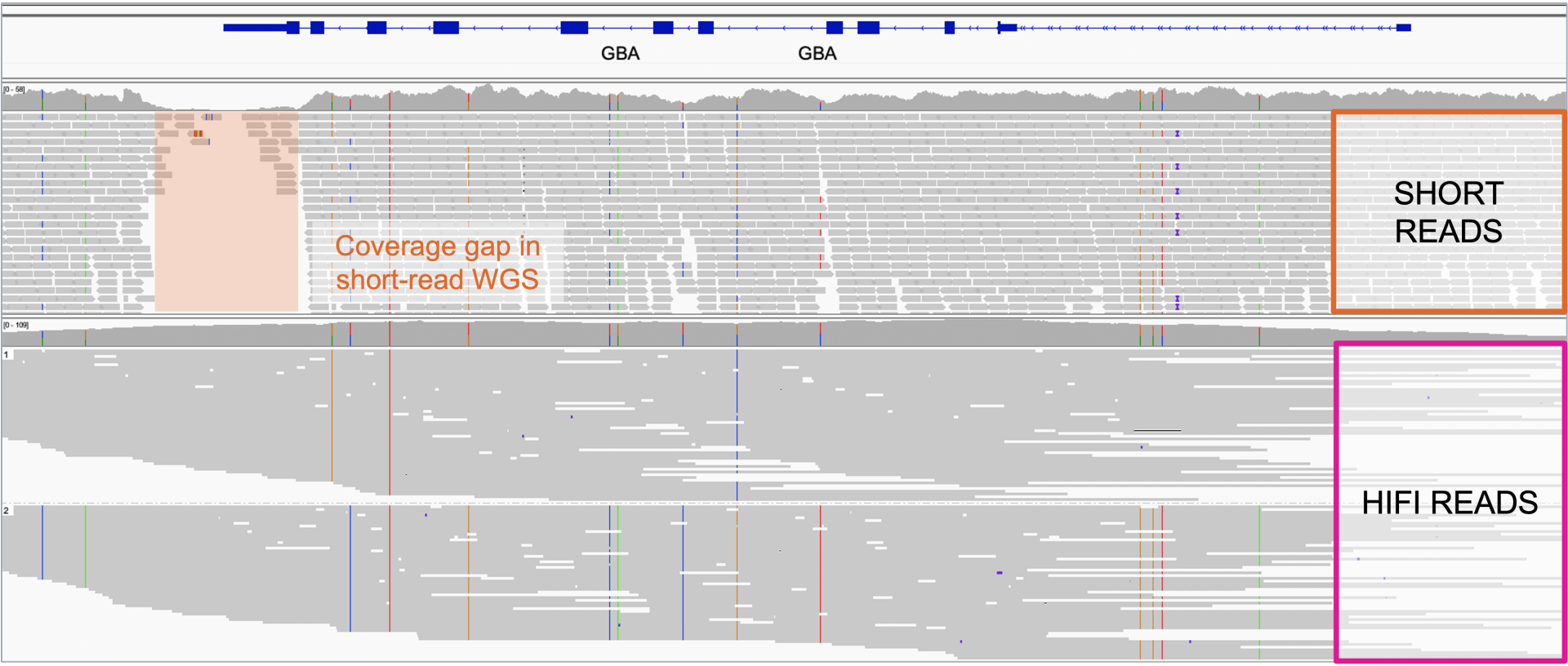
Scaffolds from contigs



Mate pair reads establish order and estimated distance between pairs of contigs

What influences number of contigs?

- genome size
- repetitiveness of genome
- number of reads
- read length



https://www.pacb.com/blog/sequencing-101-sequencing-coverage/

Short read workflow

Lab focuses
on these steps

1. Assess quality of raw reads
2. Trim raw reads based on quality
3. Assemble trimmed reads into contigs
4. Assess quality of contigs
5. Scaffold contigs into genome
6. Assess/annotate genome

Overview for Lab 19