Guest lecturer: Jillian Cieslik

Lecture 19
Genome assembly

Spigelia marilandica
(c) Matilda Adams/
Missouri Botanical Garden

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

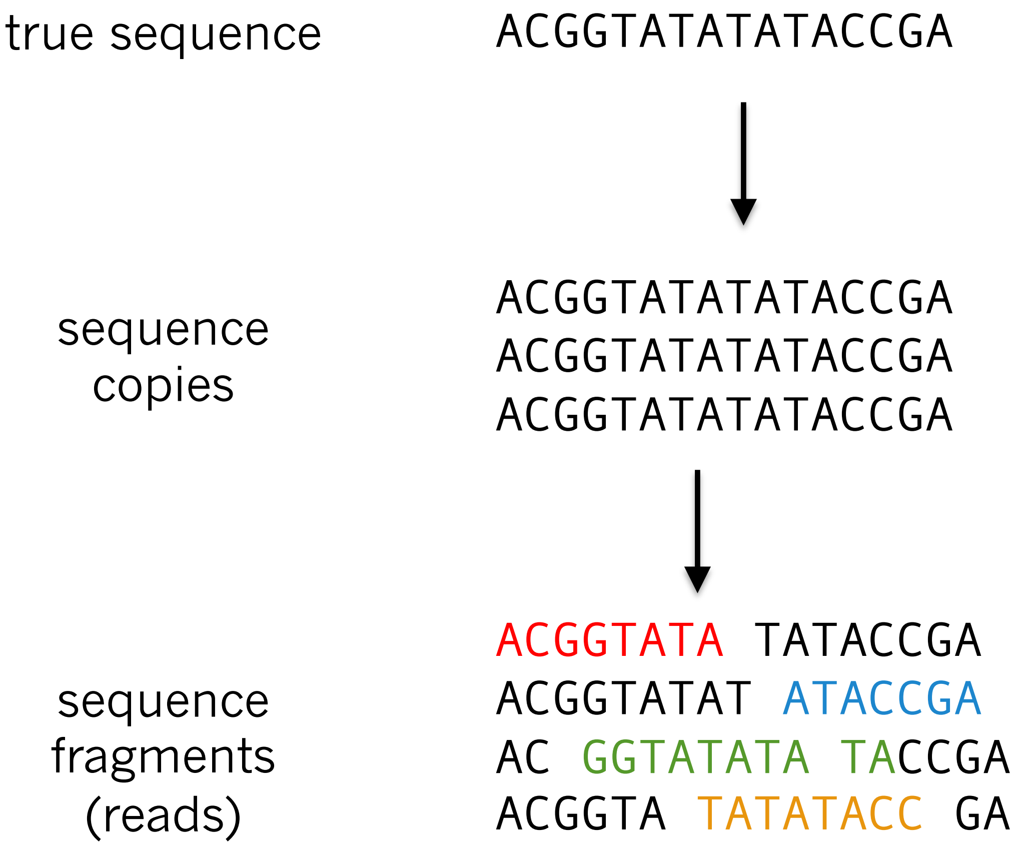
Lecture 19 outline

Last time: jupyter, matplotlib

This time: genome assembly

- genome sequences
- genome sizes
- genome assembly

Sequencing



true sequence

```
ACGGTATATATACCGA
```

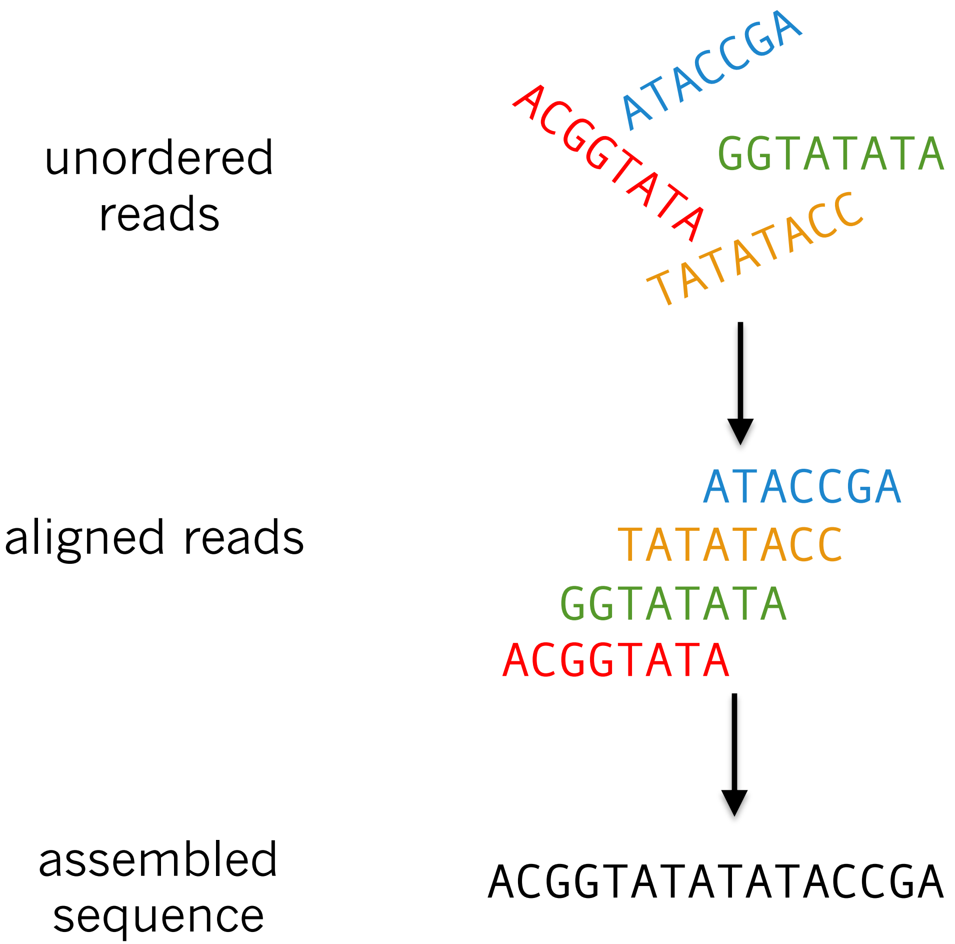
sequence copies

```
ACGGTATATATACCGA
ACGGTATATATACCGA
ACGGTATATATACCGA
```

sequence fragments (reads)

```
ACGGTATA TATACCGA
ACGGTATAT ATACCGA
AC GGTATATA TACCGA
ACGGTA TATATACC GA
```

Assembly



unordered
reads

ATACCGA
ACGGTATA
GGTATATA
TATATACC

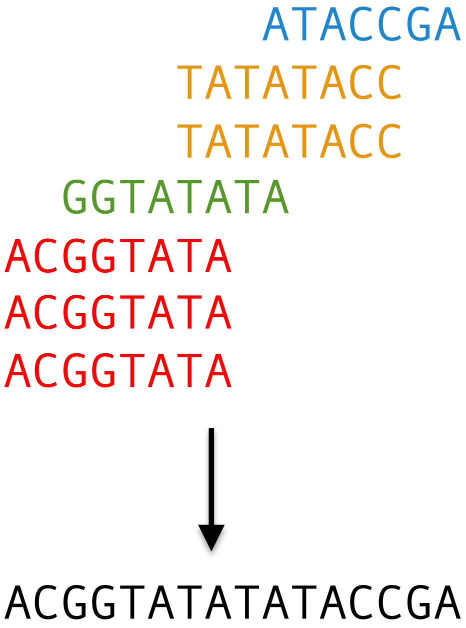
aligned reads

```
       ATACCGA
    TATATACC
  GGTATATA
ACGGTATA
```

assembled
sequence

ACGGTATATATACCGA

Assembly



```
         ATACCGA
      TATATACC
      TATATACC
  GGTATATA
ACGGTATA
ACGGTATA
ACGGTATA

ACGGTATATATACCGA
```

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

"All the News That's Fit to Print"

The New York Times

Late Edition

VOL. CXLIX . . . No. 51,432 NEW YORK, TUESDAY, JUNE 27, 2000 75 CENTS

Genetic Code of Human Life Is Cracked by Scientists

JUSTICES REAFFIRM MIRANDA RULE, 7-2; A PART OF 'CULTURE'

By LINDA GREENHOUSE

The Book of Life

A SHARED SUCCESS

2 Rivals' Announcement Marks New Medical Era, Risks and All

By NICHOLAS WADE

Science Times

A Pearl and a Hodgepodge: Human DNA

Sequenced reads

Assembled genome

"All the News That's Fit to Print"

The New York Times

Late Edition

VOL. CXLIX . . . No. 51,432 NEW YORK, TUESDAY, JUNE 27, 2000 75 CENTS

Genetic Code of Human Life Is Cracked by Scientists

JUSTICES REAFFIRM MIRANDA RULE, 7-2; A PART OF 'CULTURE'

By LINDA GREENHOUSE

The Book of Life

A SHARED SUCCESS

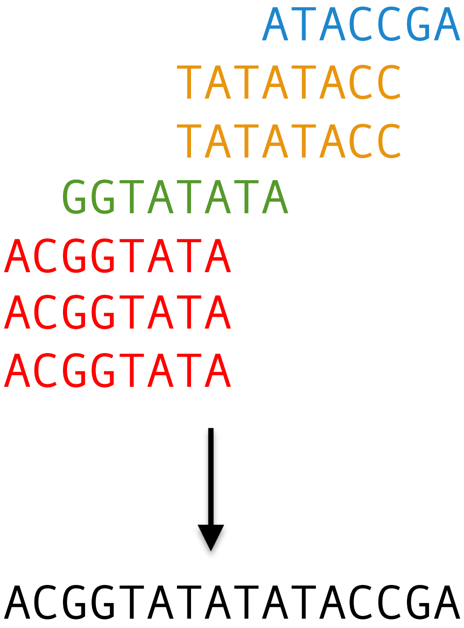
2 Rivals' Announcement Marks New Medical Era, Risks and All

By NICHOLAS WADE

Science Times

A Pearl and a Hodgepodge: Human DNA

How do we assemble reads?



```
         ATACCGA
      TATATACC
      TATATACC
  GGTATATA
ACGGTATA
ACGGTATA
ACGGTATA

ACGGTATATATACCGA
```

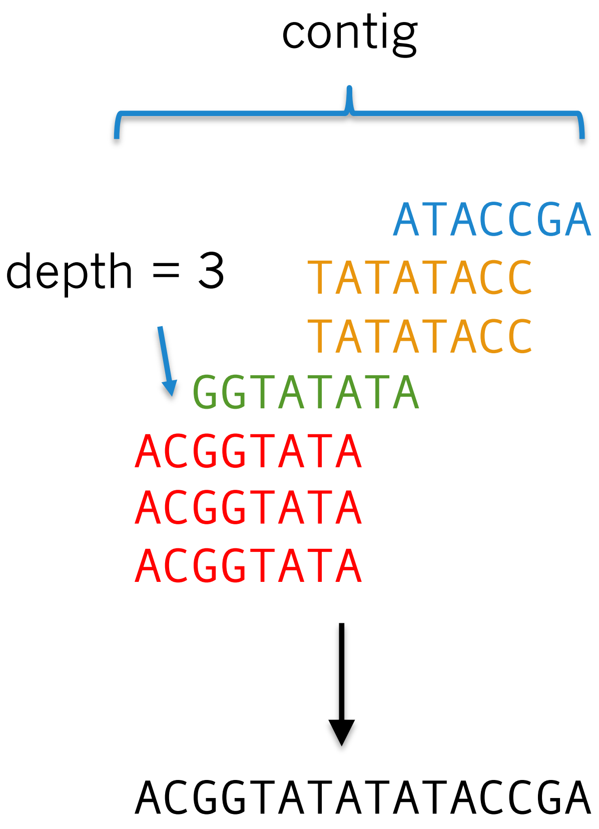
Can we do global pairwise alignments for each pair of reads?

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


contig

depth = 3

```
         ATACCGA
      TATATACC
      TATATACC
  GGTATATA
ACGGTATA
ACGGTATA
ACGGTATA

ACGGTATATATACCGA
```

We want high-coverage contigs

depth = # reads mapped for one site

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

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

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

Short read dataset sizes

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

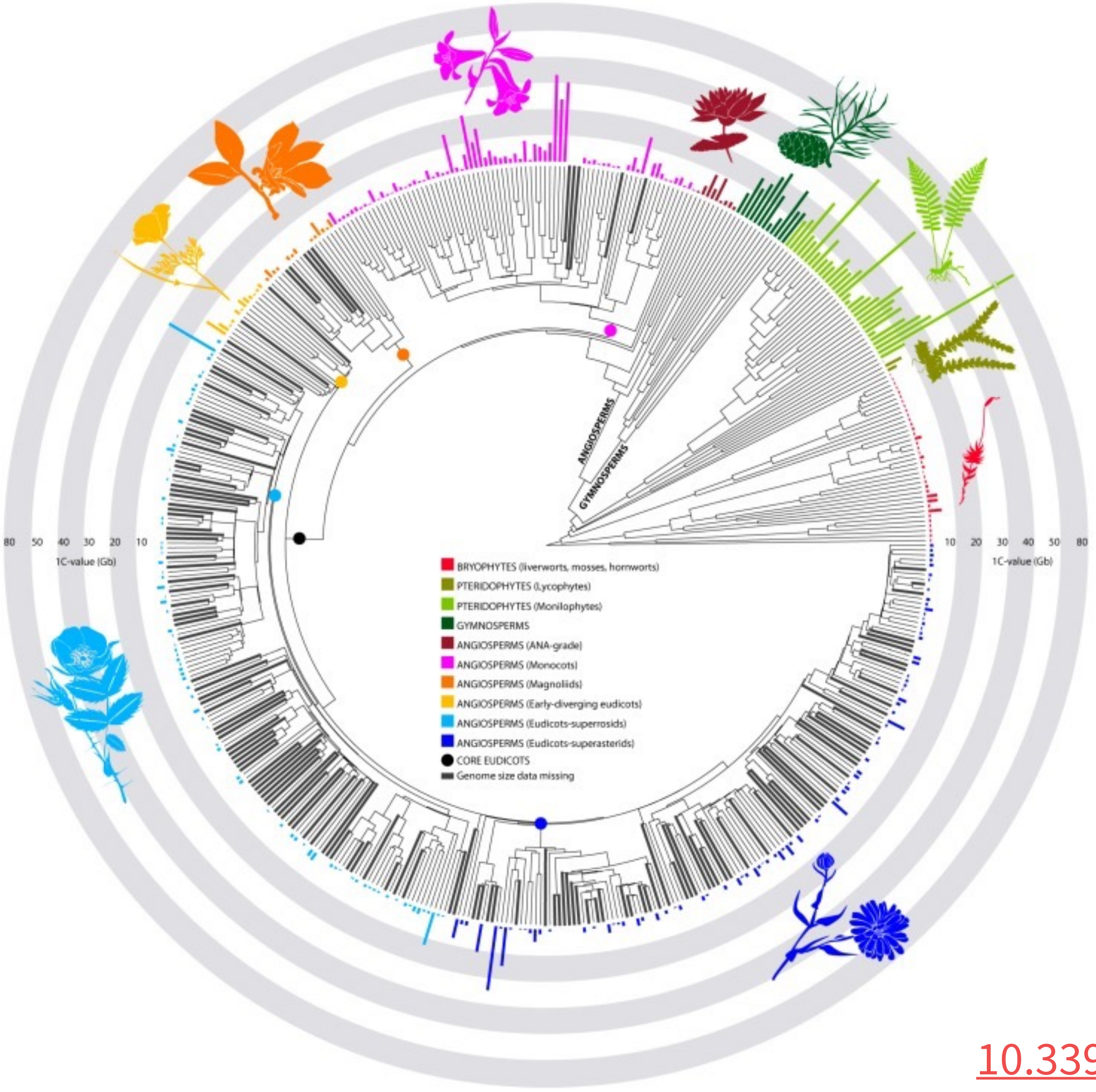
Species	**#bp**	**#reads**
SARS-CoV-2	$2x10^4$	$4x10^3$
E. coli	$4.5x10^6$	$9x10^5$
Human	$3.2x10^9$	$6.4x10^8$
Fern	$1.6x10^{11}$	$3.2x10^{10}$

Genome sizes

What generates variation in genome size?

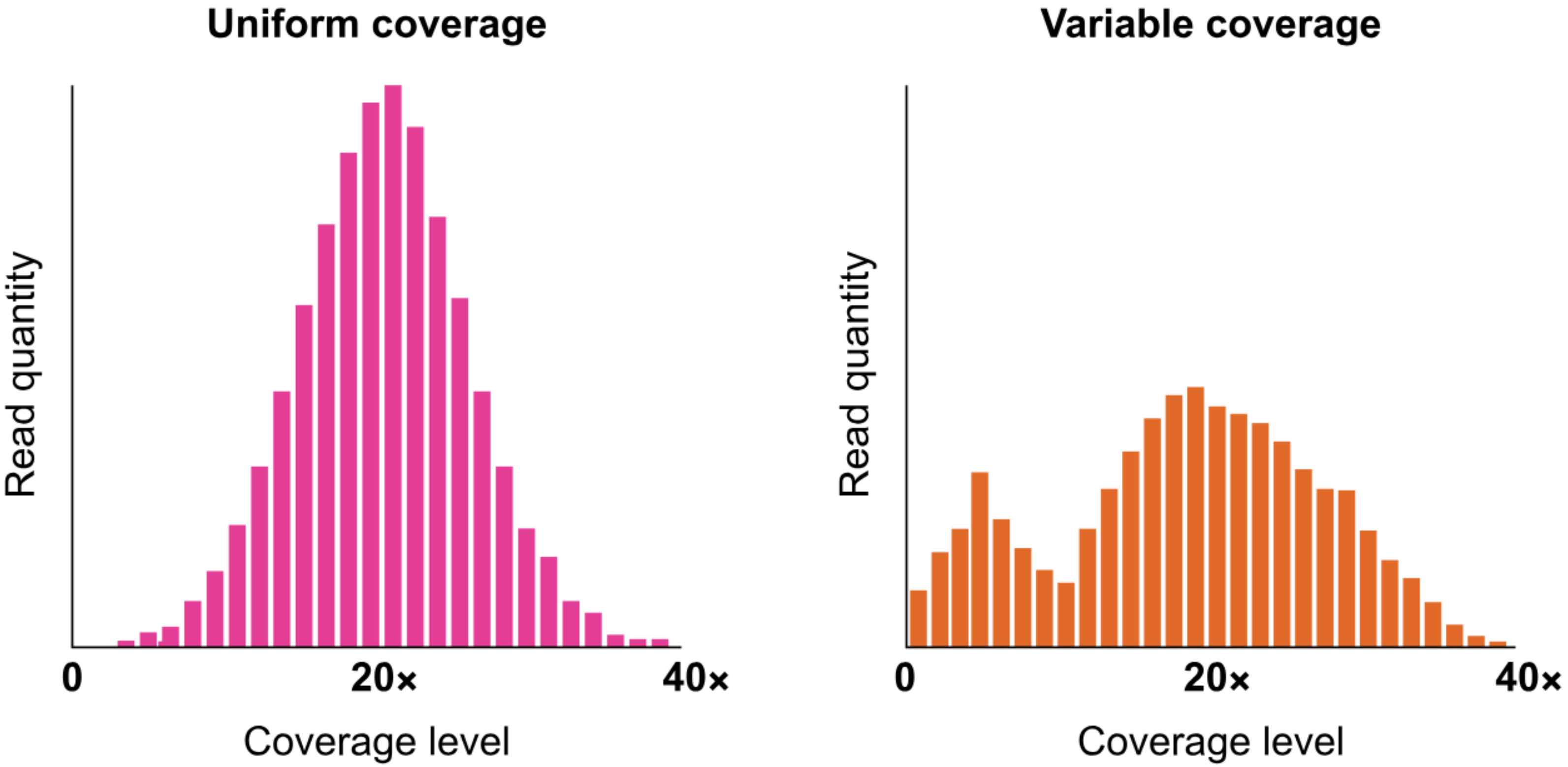
- Intron length
- Repetitive regions
- Transposable elements
- Whole genome duplication/polyploidy
- Number of genes (not always...)

Genome sizes



10.3390/genes9020088

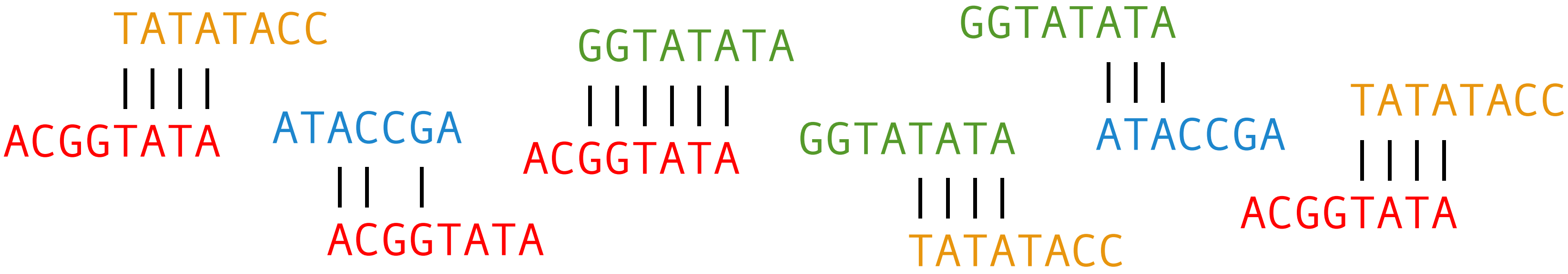
Coverage distributions



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

Assembly problem

Naive assembly would require N^2 pairwise alignments.



TATATACC
||||
ACGGTATA

ATACCGA
|| |
ACGGTATA

GGTATATA
||||||
ACGGTATA

GGTATATA
||||
TATATACC

GGTATATA
|||
ATACCGA

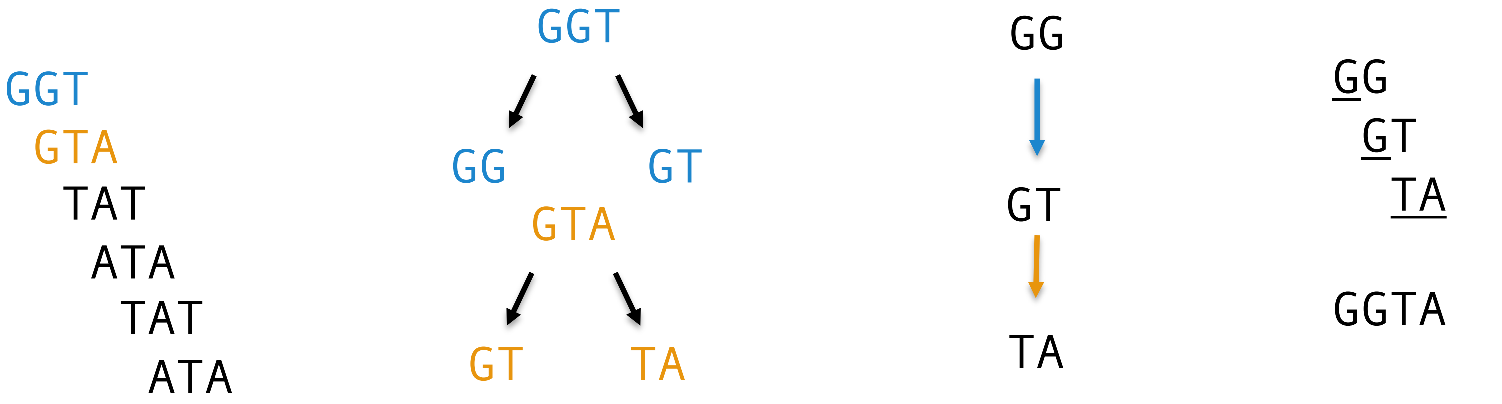
TATATACC
||||
ACGGTATA

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



kmers

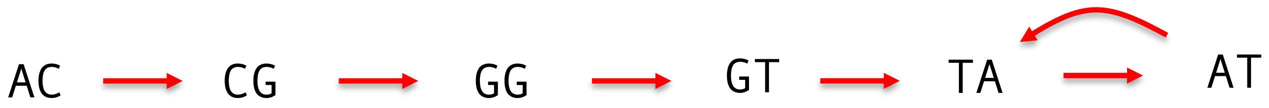
(k-1)mers

de Bruijn graph

sequence

Graph construction

One read



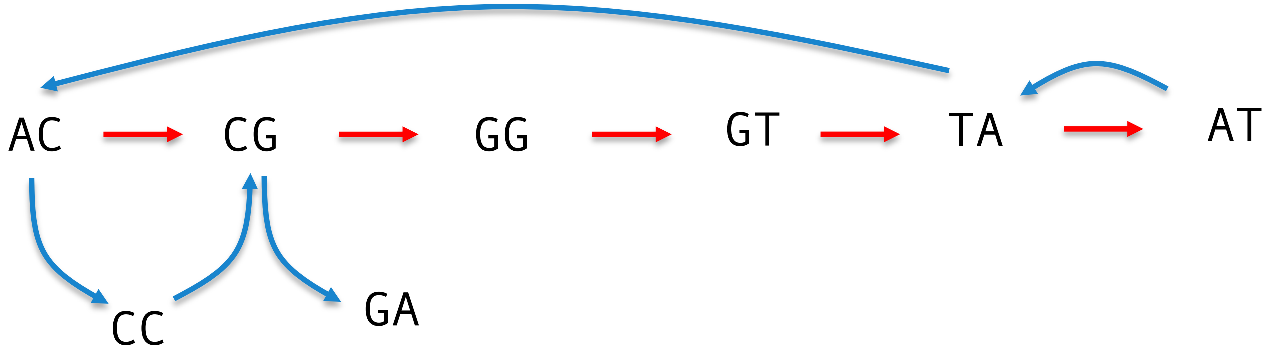
AC CG GG GT TA AT

```
ACGGTATA

ACG
 CGG
  GGT
   GTA
    TAT
     ATA
```

Graph construction

Two reads



Read 1

ACGGTATA

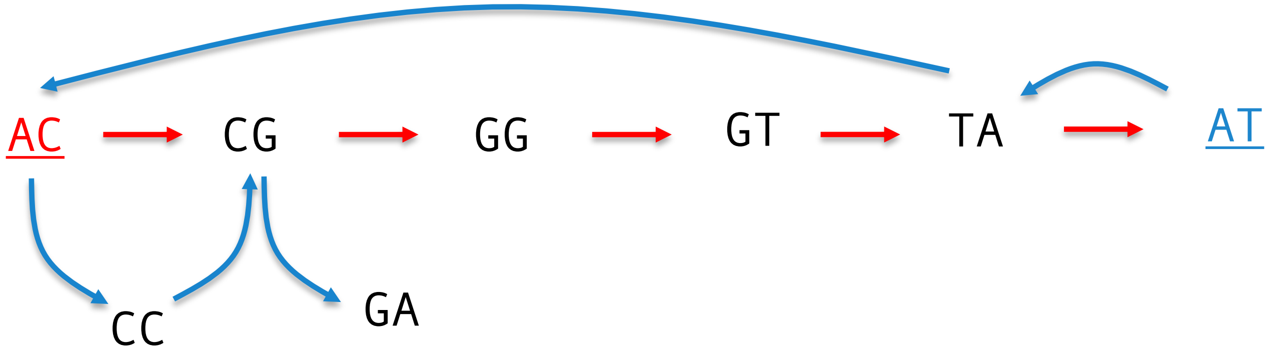
ACG
CGG
GGT
GTA
TAT
ATA

Read 2

ATACCGA

~~ATA~~
TAC
ACC
CCG
CGA

Graph traversal



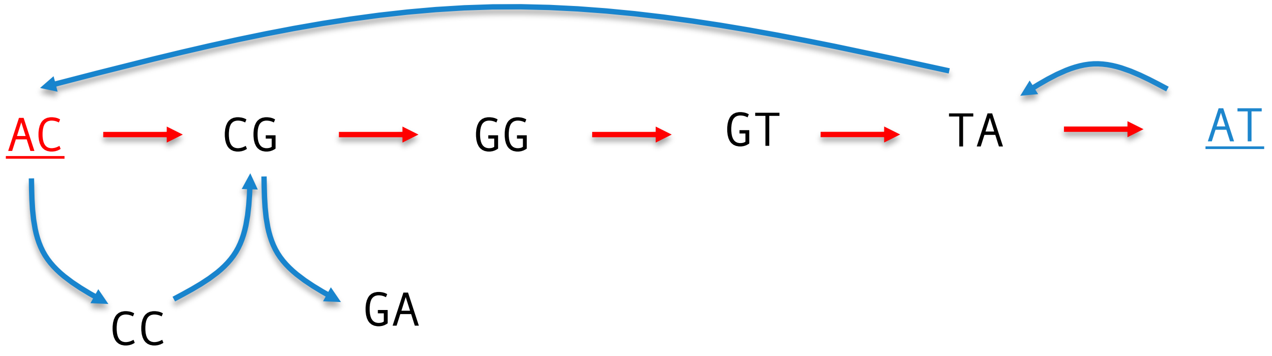
Eulerian path: visit all nodes using each edge once

Starting at AC

ACGGTATACCGA

Is this an example of a Eulerian path?

Graph traversal



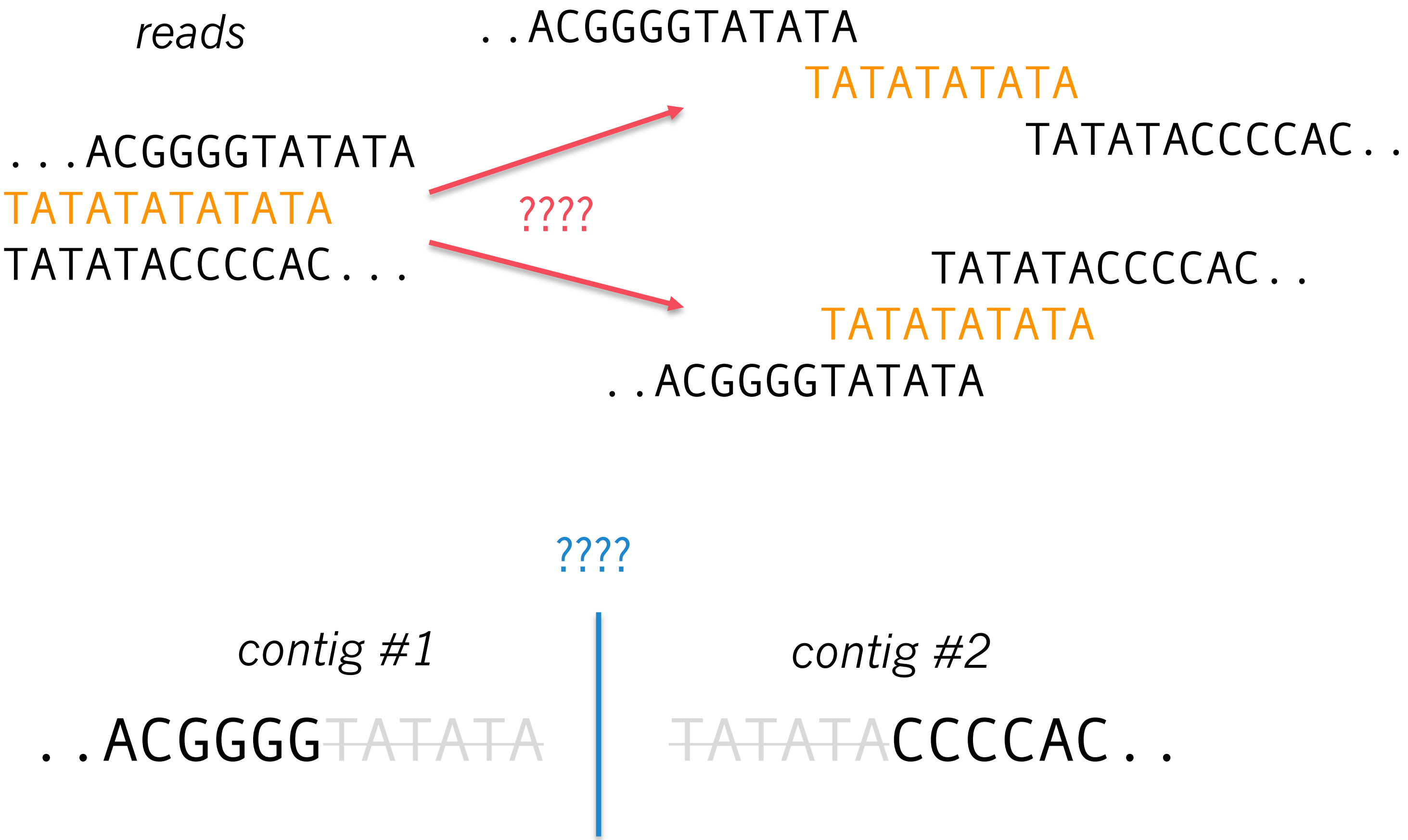
Eulerian path: visit all nodes using each edge once

Starting at AT

How about these?

ATACCGA
or
ATACCGGTAT

Repeat regions



reads

...ACGGGGTATATA
TATATATATATA
TATATACCCCAC...

????

..ACGGGGTATATA
TATATATATA
TATATACCCCAC..

TATATACCCCAC..
TATATATATA
..ACGGGGTATATA

????

contig #1

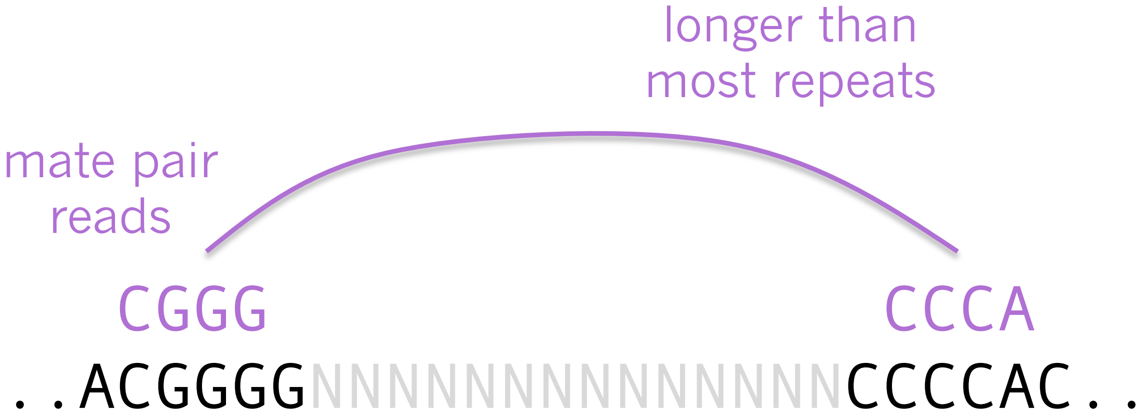
..ACGGGG~~TATATA~~

contig #2

~~TATATA~~CCCCAC..

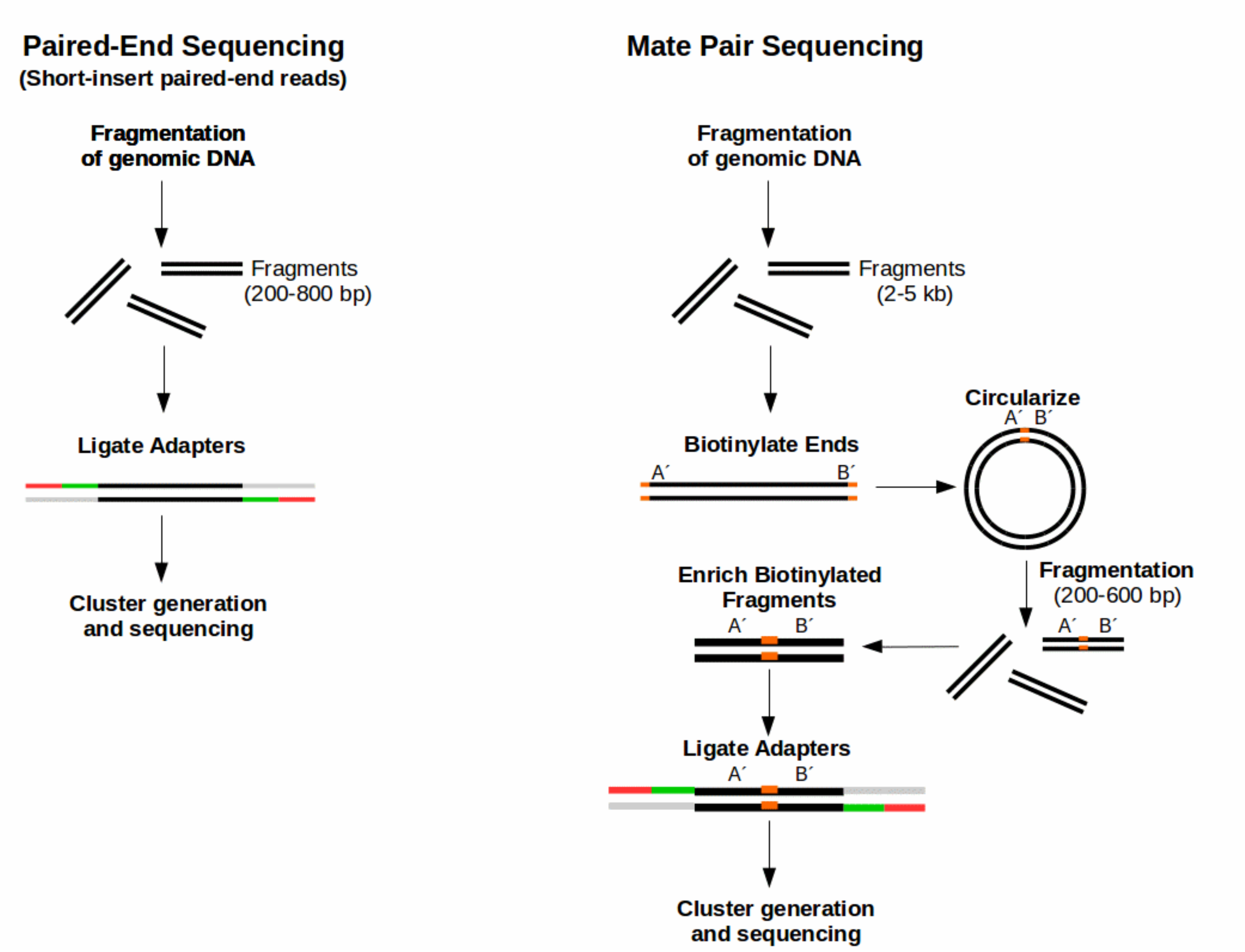
contigs form where assembly is ambiguous

Scaffolds from contigs



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

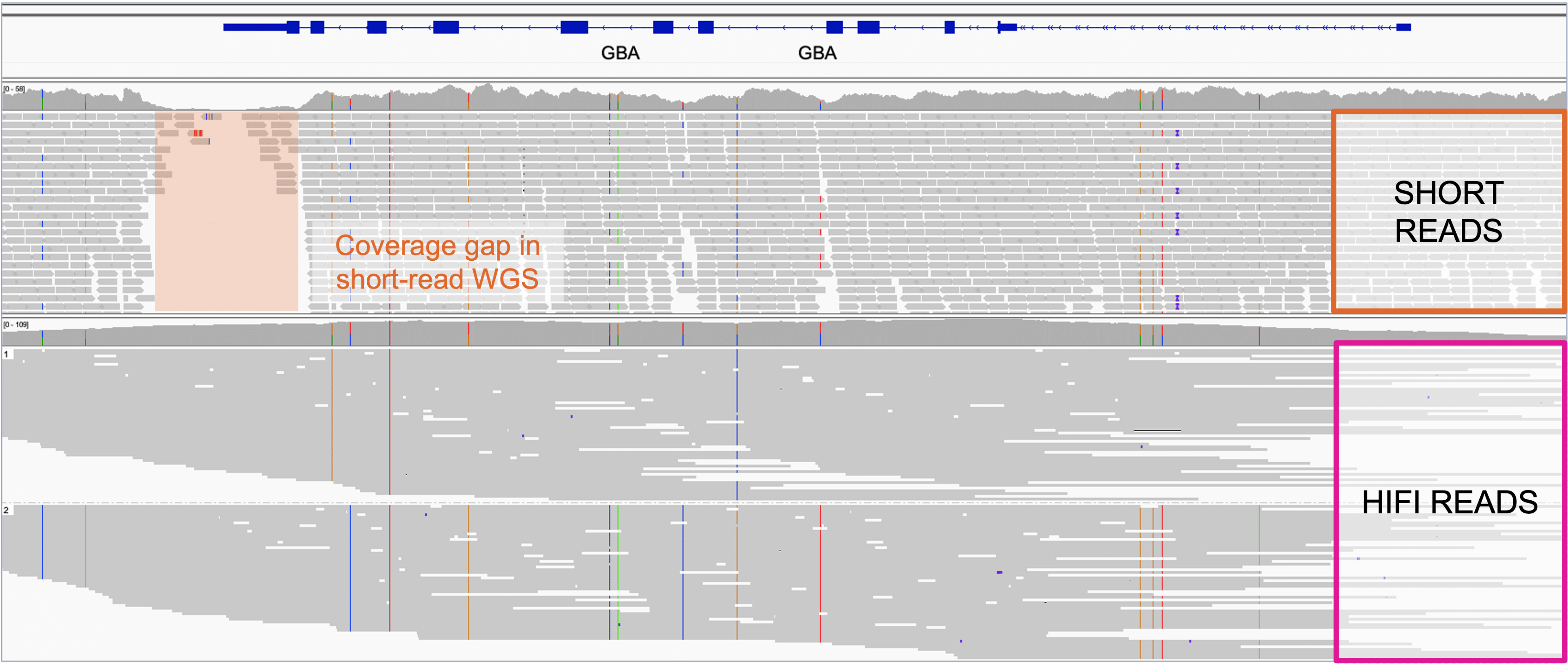
Scaffolds from contigs


Paired-End Sequencing
(Short-insert paired-end reads)
Fragmentation of genomic DNA
Fragments (200-800 bp)
Ligate Adapters
Cluster generation and sequencing
Mate Pair Sequencing
Fragmentation of genomic DNA
Fragments (2-5 kb)
Biotinylate Ends
A´
B´
Circularize
A´ B´
Fragmentation (200-600 bp)
A´ B´
Enrich Biotinylated Fragments
A´ B´
Ligate Adapters
A´ B´
Cluster generation and sequencing

https://www.ecseq.com/support/ngs/what-is-mate-pair-sequencing-useful-for

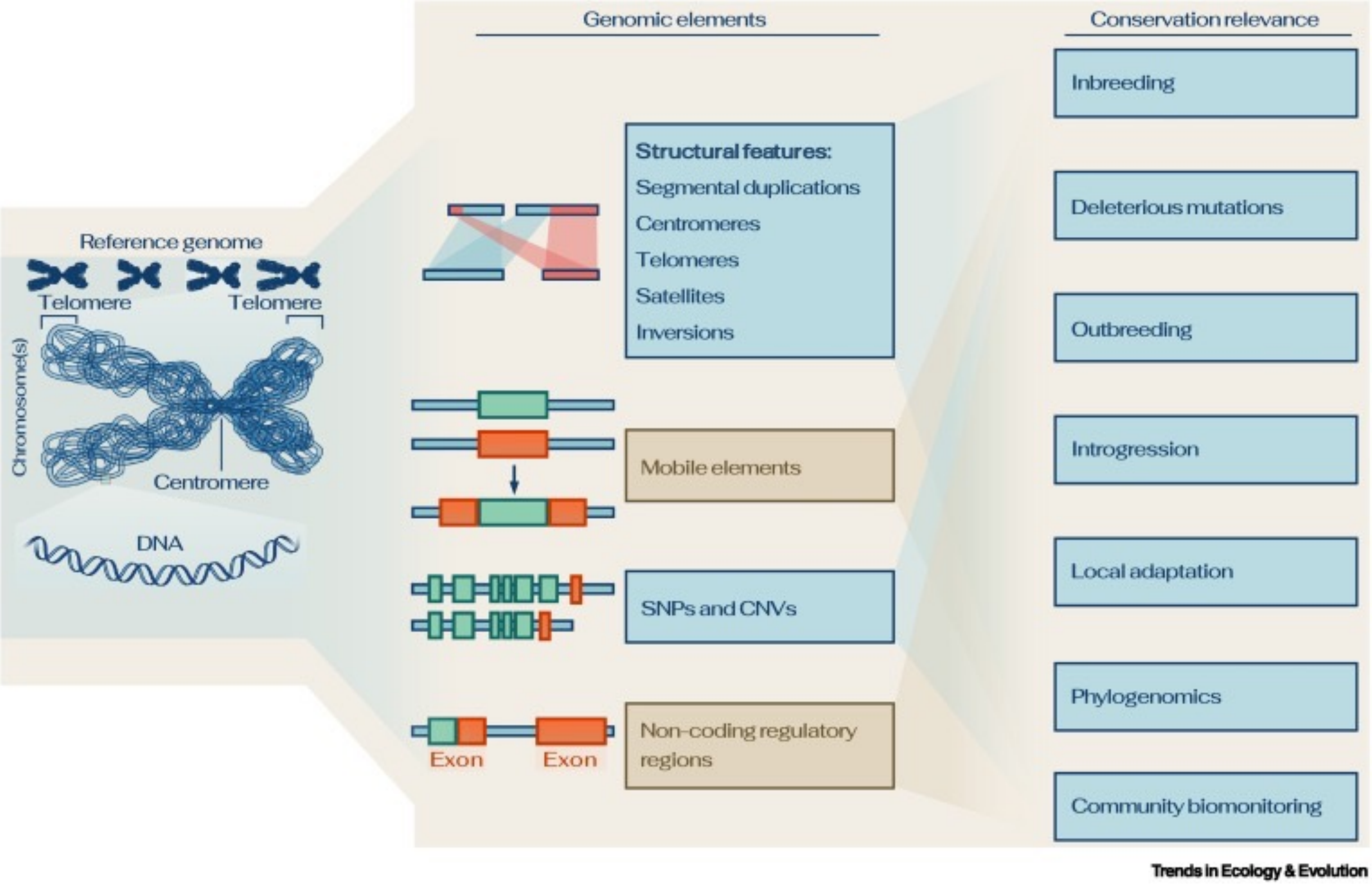
What influences number of contigs?

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



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

Reference Genomes



https://doi.org/10.1016/j.tree.2021.11.008

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