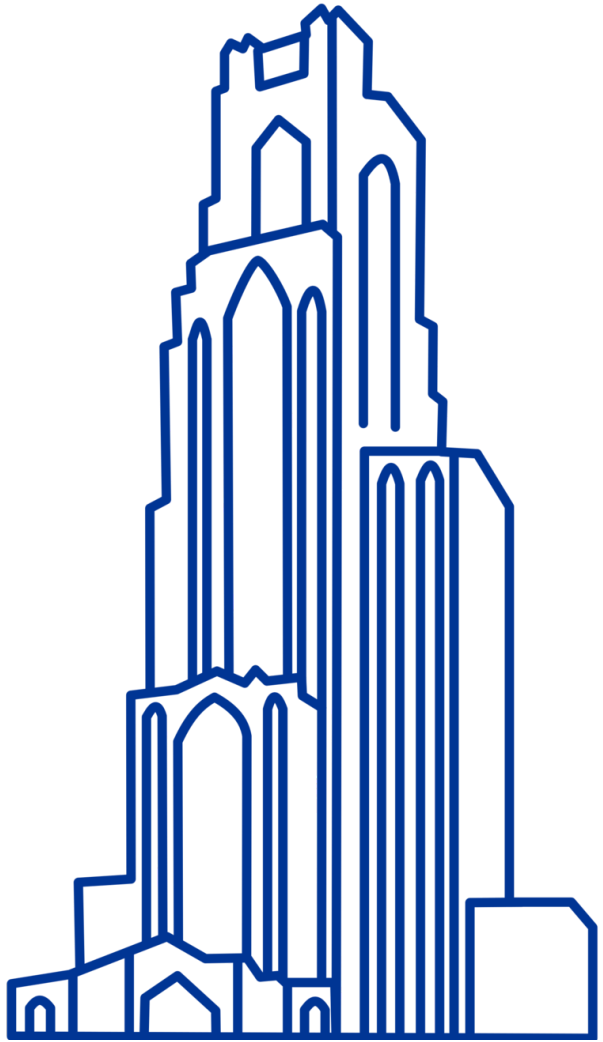


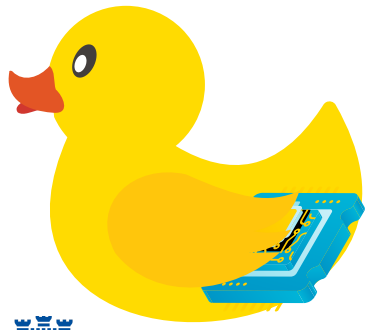
Computational Biology
(BIOSC 1540)

Lecture 03A

Genome assembly

Methodology

Jan 21, 2025



University of Pittsburgh

Announcements

Assignments

- Assignment P01B will be released tomorrow (Jan 17) at 11:59 pm

Quizzes

- Quiz 01 is next week (Jan 28) and will cover lectures 02A to 03B

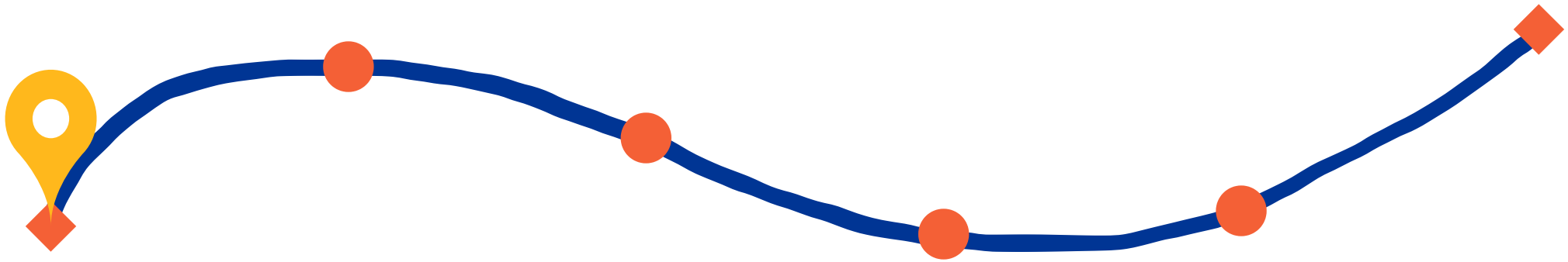
CBytes

- CByte 01 is live and will expire on Feb 1
- CByte 02 will be released Friday (Jan 24) and expire on Feb 7

Next reward: Checkpoint Submission Feedback

ATP until the next reward: 1,903

After today, you should have a better understanding of

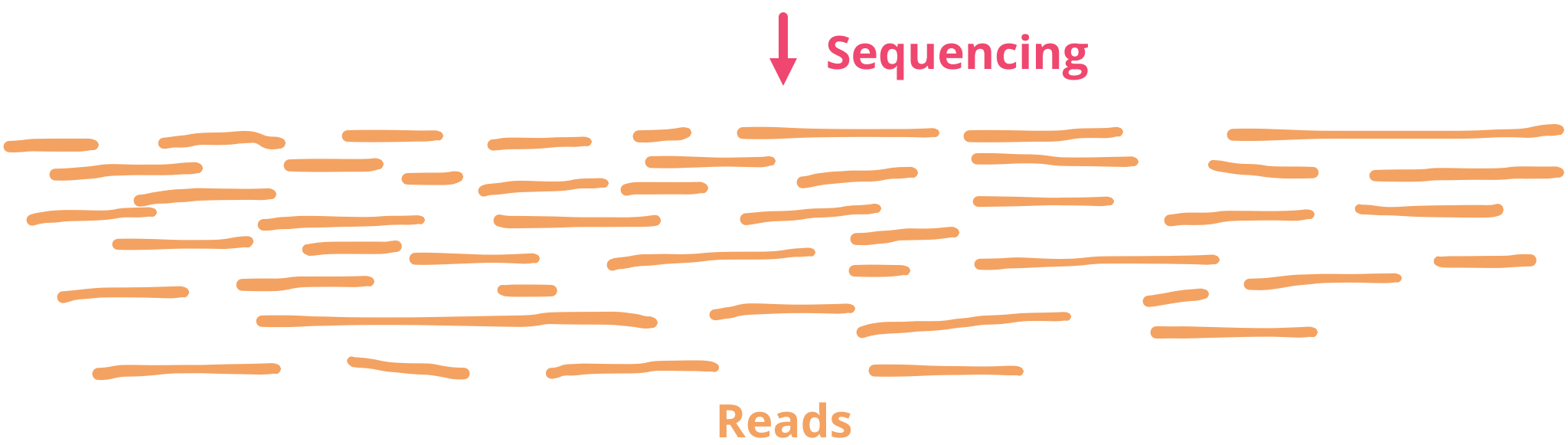


Where genome assembly fits

in the genomics pipeline

High-throughput sequencing produces short DNA fragments called reads

Modern sequencing technologies generate millions to billions of short reads (i.e., DNA fragments) from a DNA Sample



Reads are typically 100–300 base pairs long for short-read technologies and up to tens of kilobases for long-read technologies

Reads are assembled into contigs by identifying overlaps between them

Reads overlap where they represent the same genomic region

Reads

Assembly

Assembled DNA sequence (i.e., contig)

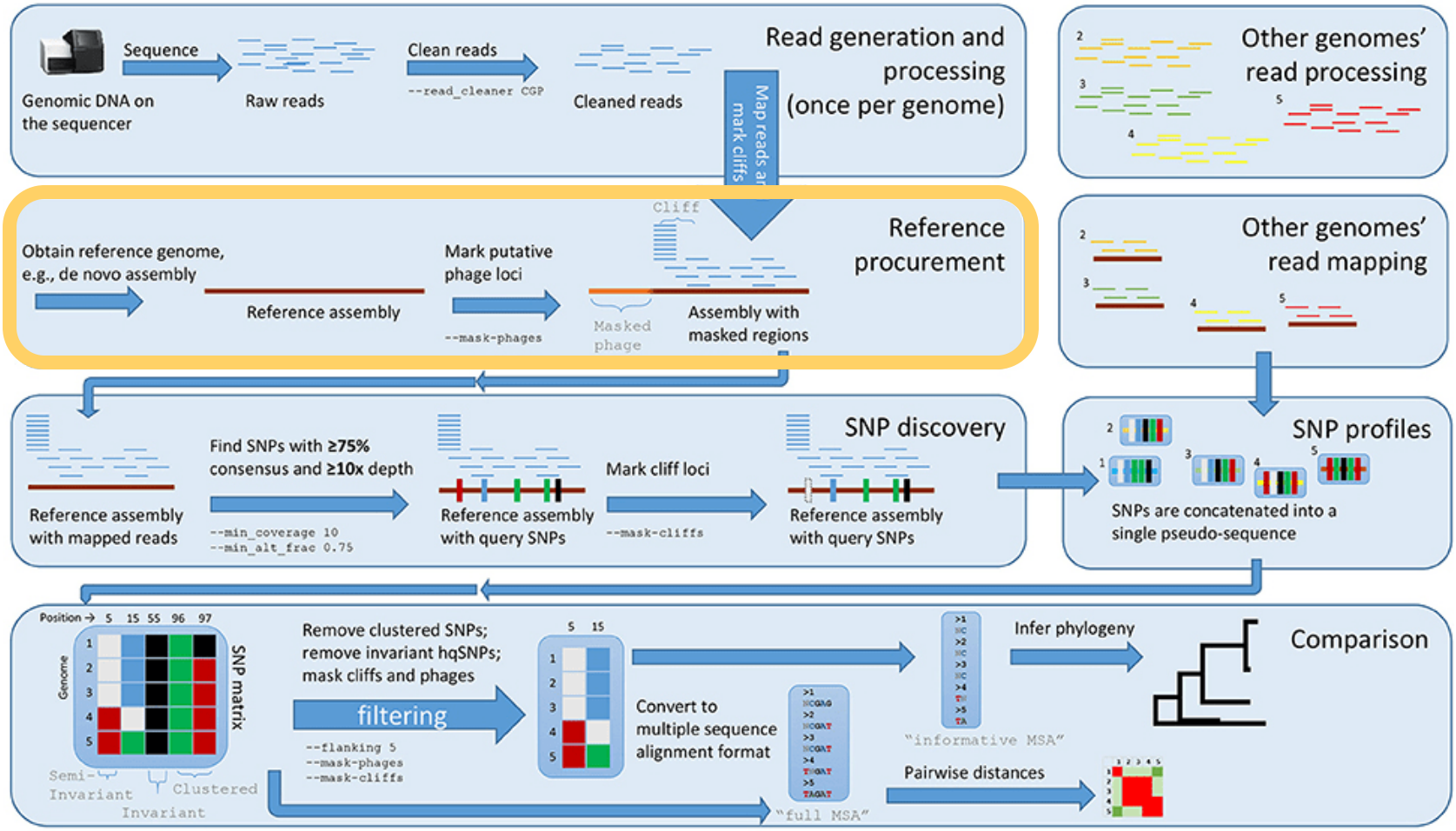
Overlap information is used to merge reads into contiguous DNA sequences called contigs

Assuming perfect sequencing and assembly, the resulting contig will match our original DNA sample

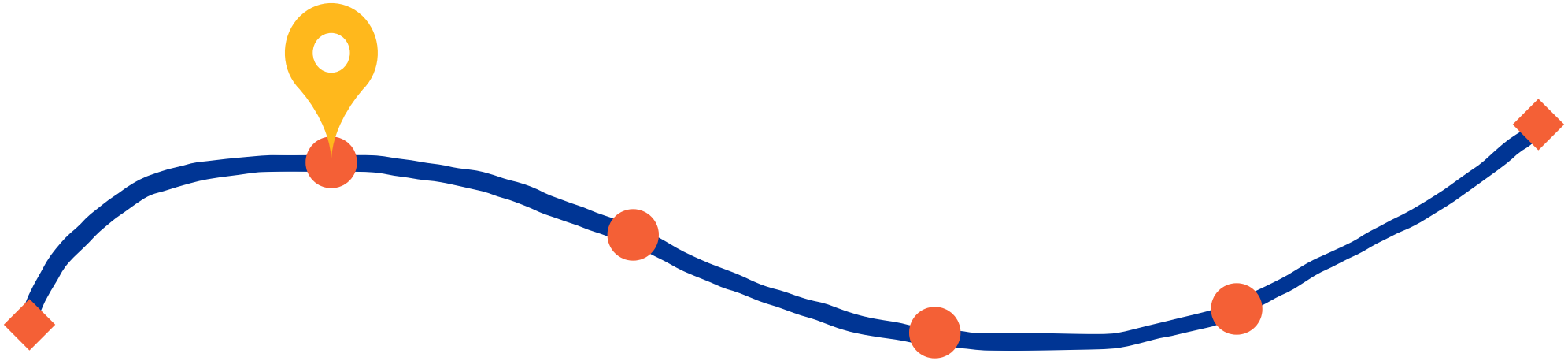
Genome assembly bridges sequencing data and biological insights

Assembled genomes are essential for identifying genes and understanding regulatory elements

Provides a foundation for downstream analyses, including functional and structural genomics



After today, you should have a better understanding of



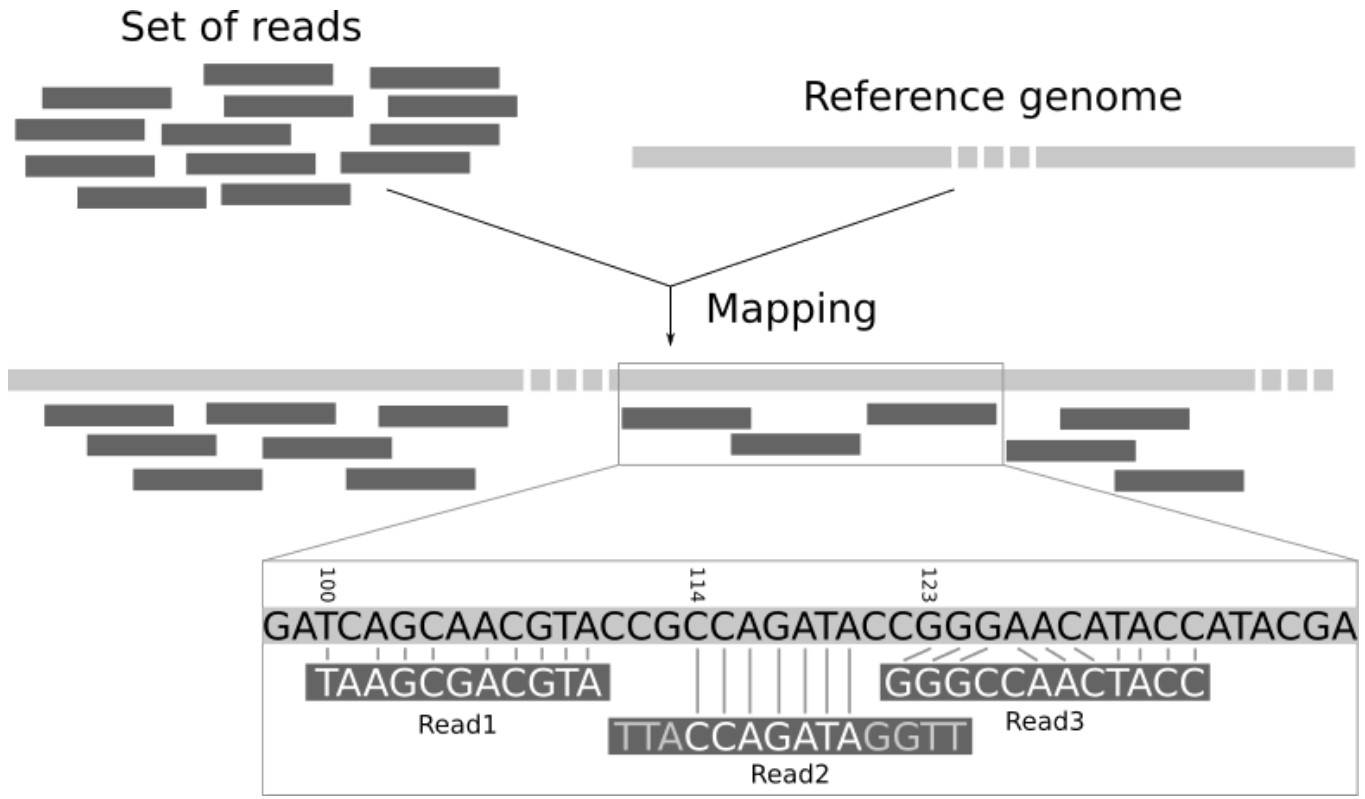
Types of genome assembly

Reference-based

Reference-based assembly aligns reads to a known genome

RefSeq provides high-quality reference genomes, transcriptomes, and proteins

Sequencing reads are matched to a reference genome to determine their correct positions



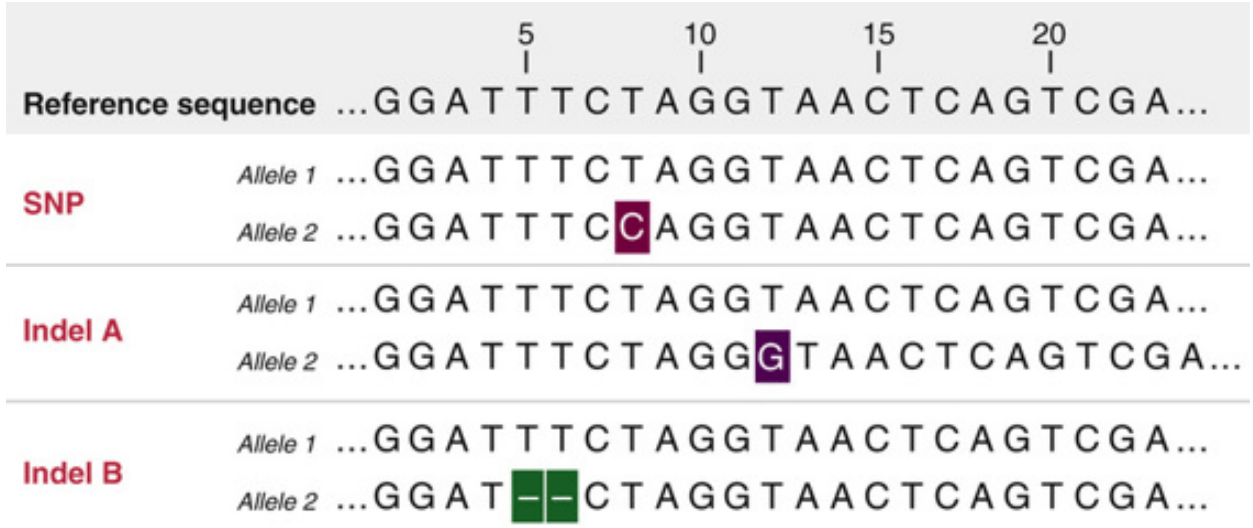
Alignment relies on identifying overlaps and shared sequences between the reads and the reference

Mapping reveals variations like SNPs and small insertions or deletions

Variations occur when a read differs from the reference genome

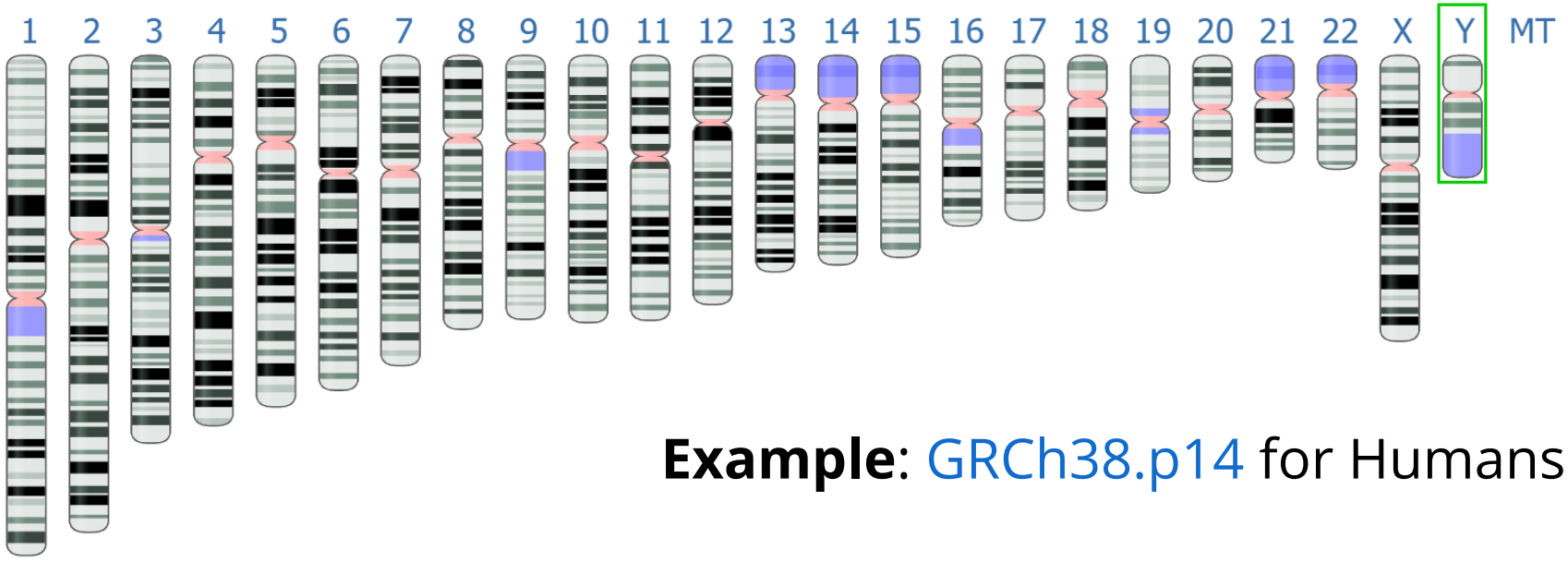
Single-nucleotide polymorphisms (SNPs) are single base changes between the read and the reference

Indels are small insertions or deletions that alter the alignment pattern



Reference-based assembly is ideal for organisms with well-annotated genomes

- Works effectively when a complete, accurate reference genome is available.
- It is commonly used for model organisms like **humans, mice,** or **fruit flies** with high-quality reference genomes, not recommended for novel organisism.

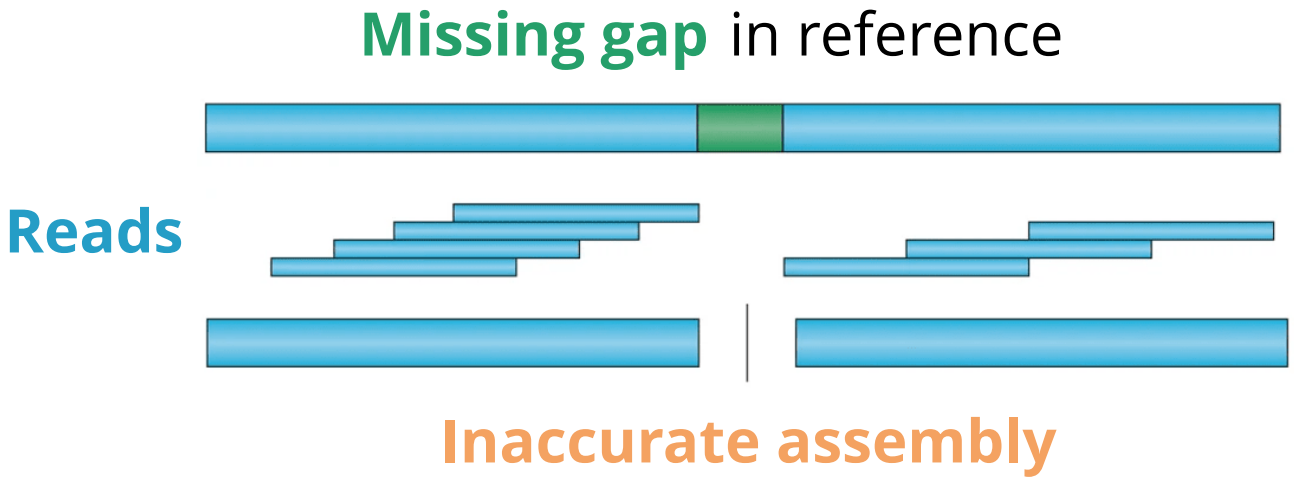


Example: GRCh38.p14 for Humans

Reduces time and cost for studies focused on variant detection or evolutionary comparisons.

Regions absent in the reference genome result in gaps in the assembly

Gaps can occur due to incomplete reference sequences or highly divergent regions in the sample genome.

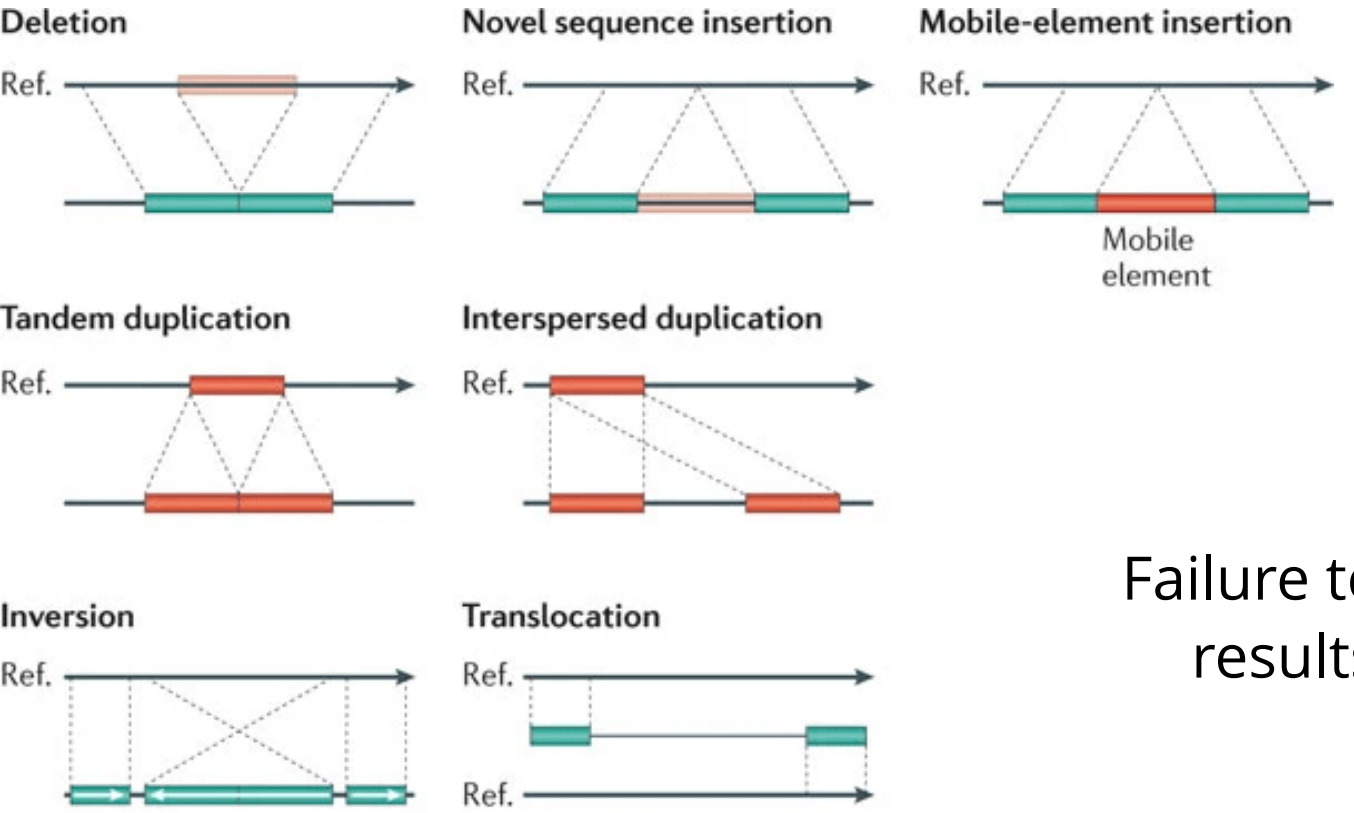


These gaps can affect downstream analysis, especially for novel genes or functional elements.

Reads corresponding to regions missing in the reference cannot be mapped, leaving unassembled gaps.

Structural variations can be overlooked or incorrectly assembled

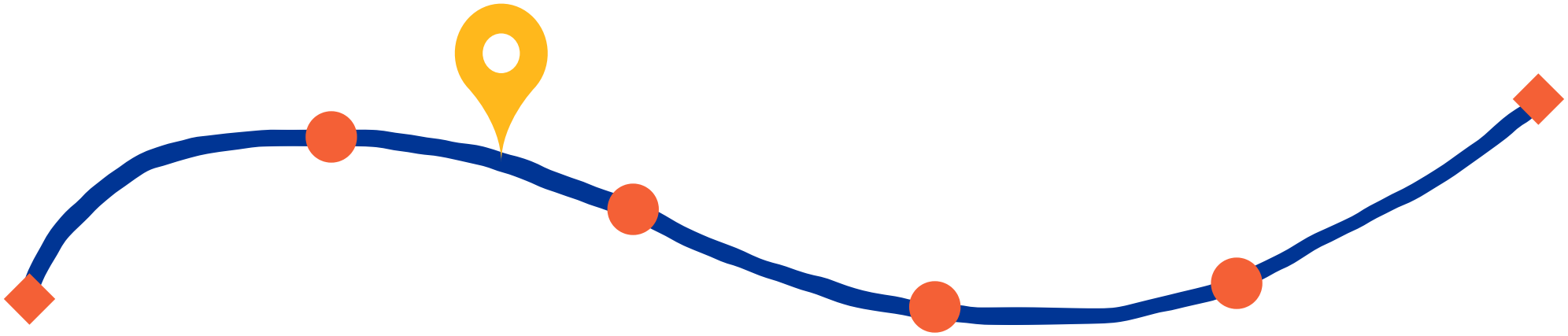
Variations like insertions, deletions, inversions, or translocations may not align correctly to the reference.



Assemblers may interpret these variations as mismatches or sequencing errors.

Failure to account for structural variations can skew results and mask important genomic differences.

After today, you should have a better understanding of

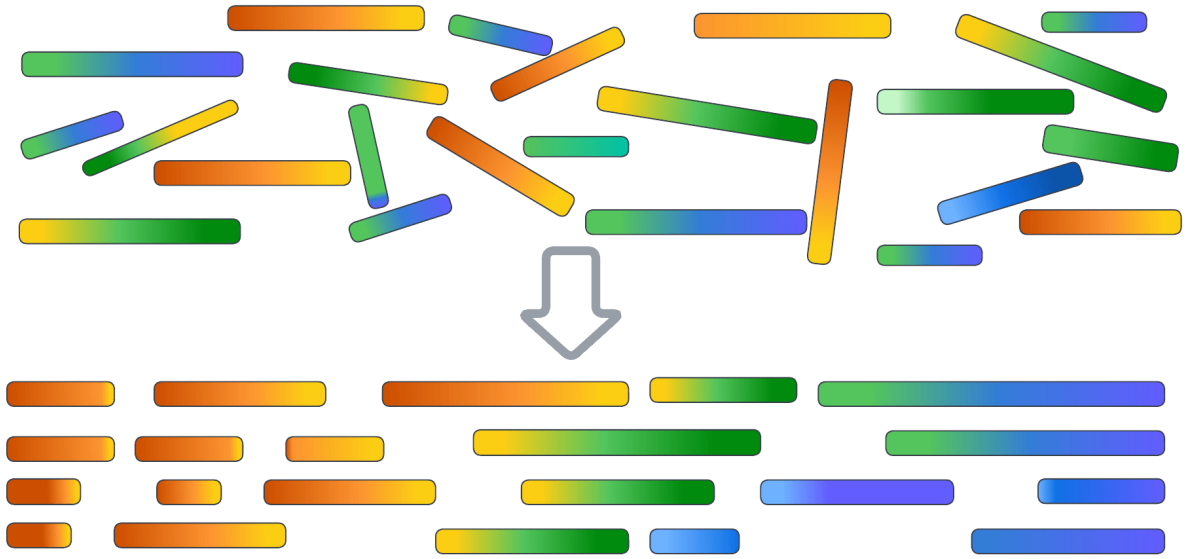


Types of genome assembly

De novo

De novo assembly reconstructs genomes without a reference

Instead of mapping to a reference, reads are assembled by finding overlaps between reads and merging them



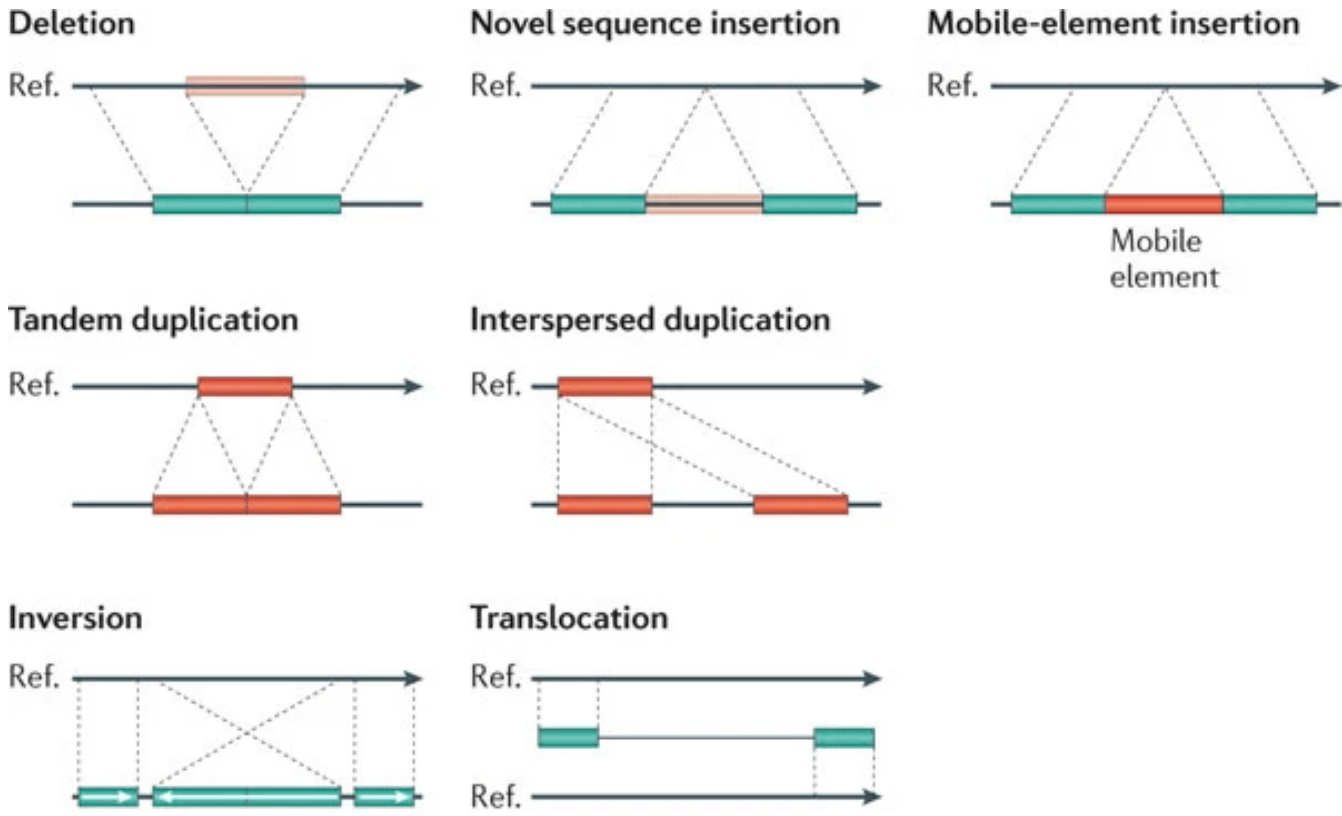
It does not rely on pre-existing data, allowing for unbiased genome reconstruction. Essential for novel organisms or those with no reference genome.

De novo assembly captures the full genome, including novel regions

Unbiased assembly enables the discovery of unique and divergent sequences.

Resolves structural variations that reference-based methods might miss.

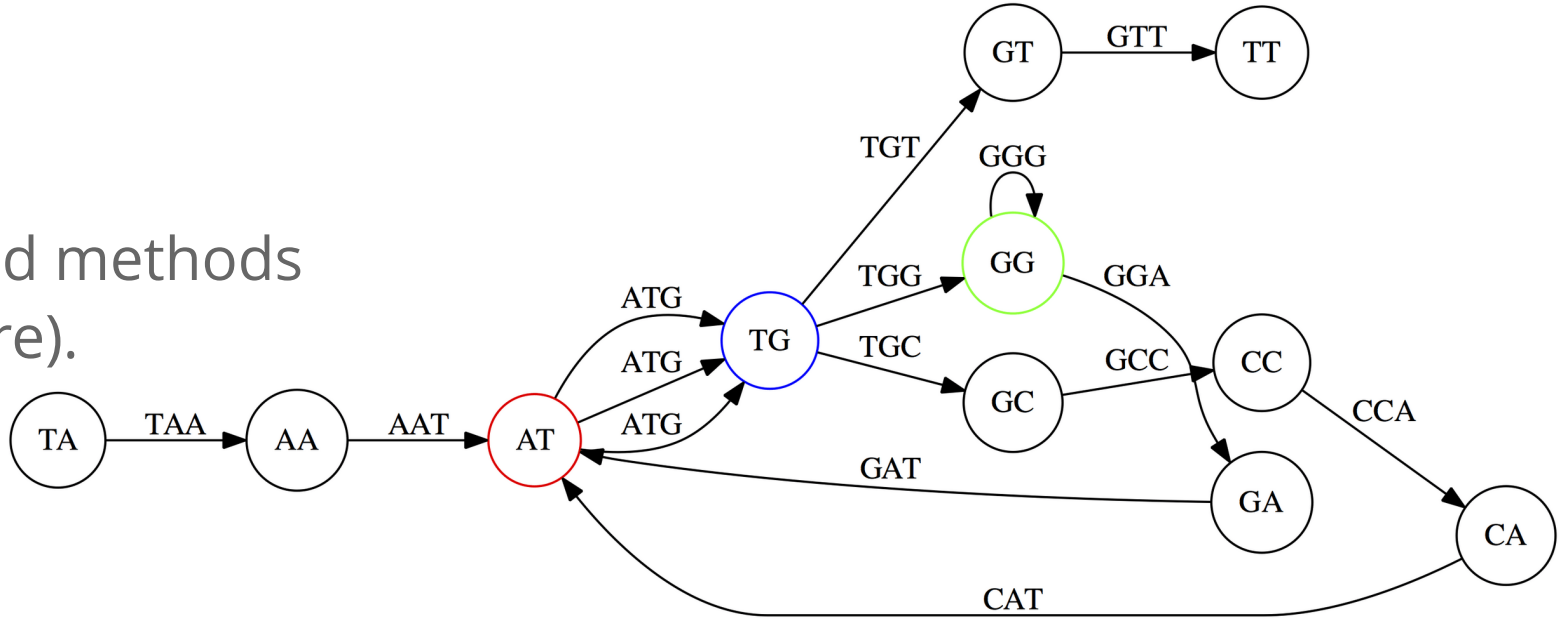
Ideal for exploring non-model organisms and highly variable regions.



De novo assembly faces computational and biological challenges

High computational requirements due to complex algorithms

Most methods use graph-based methods (more on this in the next lecture).



Struggles with repeats, sequencing errors, and low-coverage regions (more on this later).

Reference-Based vs. De Novo Genome Assembly

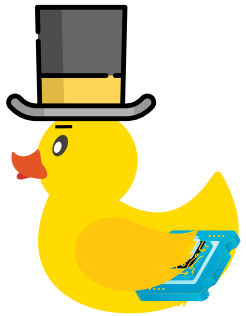
Researchers are analyzing the genome of a newly discovered bacterial strain suspected to carry antibiotic-resistance genes. They have access to a draft reference genome from a closely related strain, but it is incomplete and poorly annotated. Their main goal is identifying novel resistance genes while ensuring assembly accuracy and minimizing computational costs.
Which approach would you recommend for assembling the genome, and why?

A. Use reference-based assembly to ensure computational efficiency and focus on conserved regions.

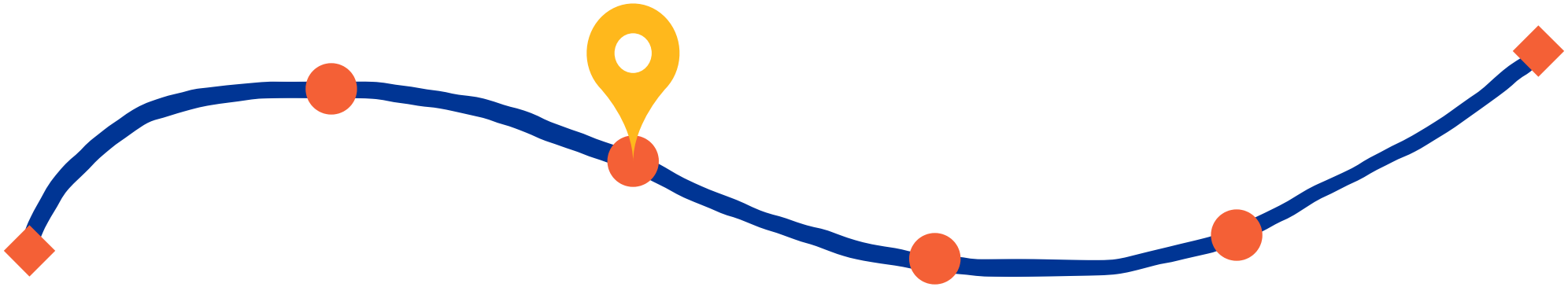
B. Use de novo assembly to avoid reference bias and discover novel resistance genes.

C. Use hybrid assembly, starting with reference-based assembly and refining with de novo assembly for poorly aligned regions.

D. BLAST reads that fail to align to the reference genome but avoid de novo assembly to reduce computational cost.



After today, you should have a better understanding of



Challenges in genome assembly

Genome assembly faces biological and technical challenges

These challenges complicate the process of accurately reconstructing a genome.

Biological factors: Repetitive sequences, structural variations, and genome size.

Technical issues: Sequencing errors, low coverage, and short read lengths.

Overcoming these challenges requires balancing biological and technical factor

Advances in sequencing technology (e.g., long-read sequencing)

Careful experimental design (e.g., choosing read length and depth)

After today, you should have a better understanding of

Challenges in genome assembly

Repetitive DNA (i.e., repeats)

Repeats are a widespread feature of many genomes

Repeats are sequences of DNA that occur multiple times in the genome

Common types of repeats

Tandem repeats: Consecutive copies of the same sequence.

AGCTGATC CGAT CGAT CGAT CGAT TTAGCCGA

Interspersed repeats: Similar sequences scattered throughout the genome

AGCTGATC CGAT CGAT TTAGCCGA CGAT CGAT

Note: Repeats are especially abundant in eukaryotic genomes, comprising up to 50% of human DNA.

Repeats create ambiguity in placing reads during assembly

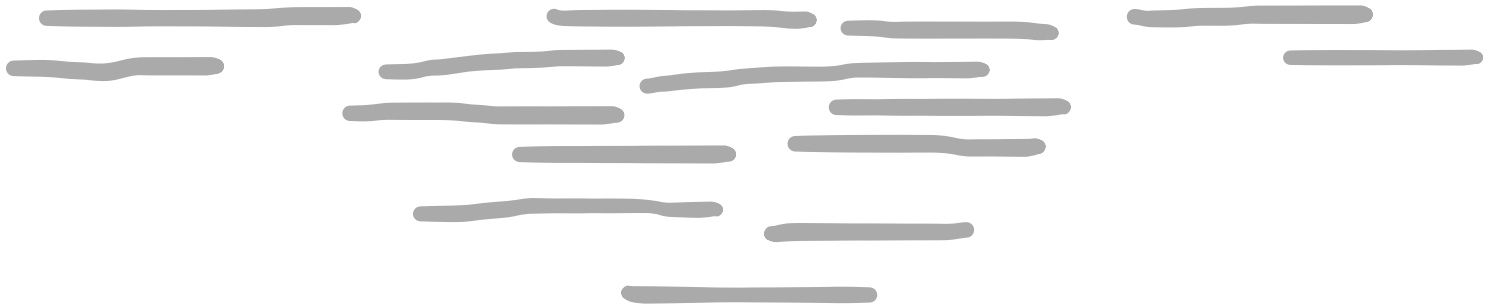
How will the assembler know the difference between these two options?
Maybe it has high coverage instead of more repeats?

Option 1

AGCTGATC CGAT CGAT CGAT CGAT CGAT CGAT CGAT CGAT CGAT CGAT TTAGCCGA

Option 2

AGCTGATC CGAT CGAT CGAT CGAT CGAT TTAGCCGA



Repeats create ambiguity in placing reads during assembly

Reads from repeats may align to multiple locations, making it unclear where they belong.

AGCTGATC CGAT CGAT CGAT CGAT CGAT TTAGCCGA CGAT CGAT CGAT CGAT CGAT

Which repeat did a read come from?

Who knows ...

Repeats can lead to fragmented assemblies or misassembled contigs

Fragmentation: Assemblers may break contigs at repetitive regions, resulting in gaps.

Collapsing repeats: Similar repeats may be merged into a single copy, leading to incorrect assemblies.

AGCTGATC CGAT CGAT CGAT CGAT CGAT CGAT CGAT CGAT CGAT CGAT TTAGCCGA

versus

CGAT CGAT CGAT CGAT CGAT TTAGCCGA

AGCTGATC CGAT CGAT CGAT CGAT CGAT

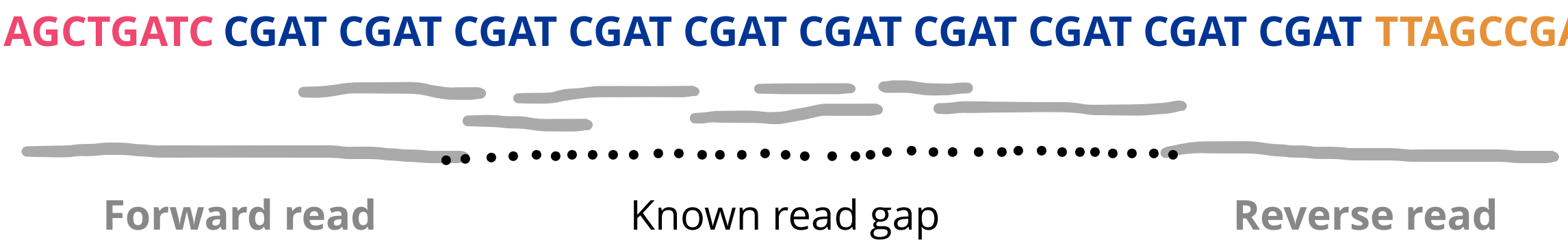
Read length affects the ability to resolve repeats during assembly

Short reads: Often shorter than repeat regions, making it difficult to span and resolve repeats.

Long reads: Can span entire repetitive regions, reducing ambiguity and improving assembly accuracy.

Paired-end reads span repetitive regions, providing distance information

Paired-end reads are sequenced from both ends of a DNA fragment, with a known distance between the reads



If I have two reads at the ends of a repeat, and I know the distance between the reads, I know the length of repeat

(This is why having paired-end reads that do not overlap is helpful.)

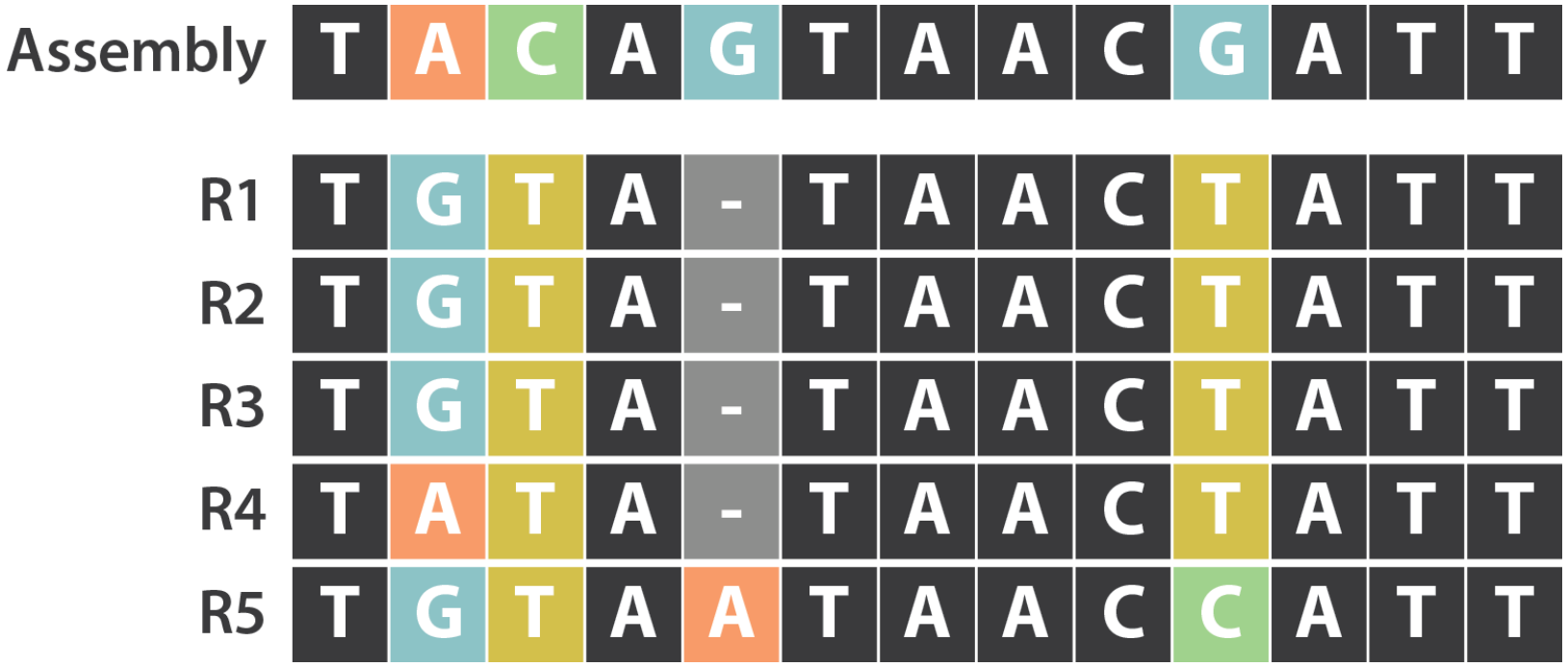
After today, you should have a better understanding of

Challenges in genome assembly

Sequence errors

Sequencing errors disrupt overlaps, complicating assembly

Sequencing errors interfere with overlaps by creating mismatches between reads

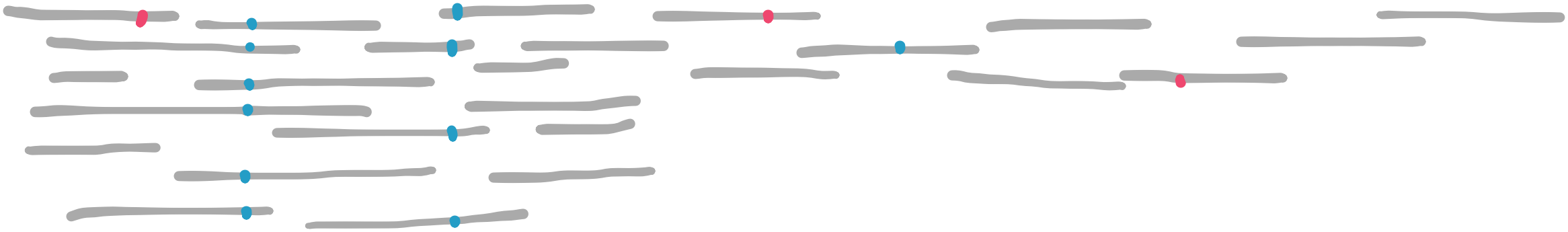


Assemblers must distinguish true overlaps from errors, which dramatically increases computational complexity.

Assemblers use error correction and redundancy to handle sequencing errors

Redundant data (high coverage) helps correct errors by identifying the most likely base (i.e., consensus)

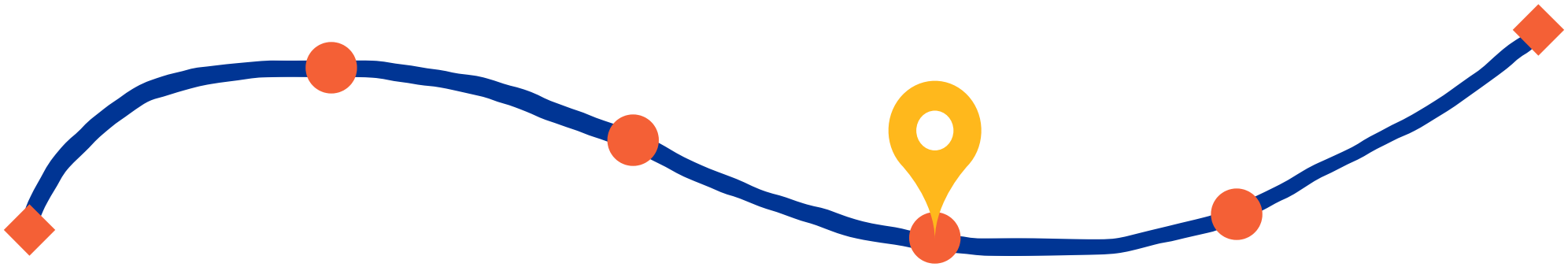
TACGATCGGATTACGCGTAGGCTAGCTTACGGACTCGATGTACGATCGGATTACGCGTAGG



Real sequencing errors can be fixed in high-coverage areas

Real SNPs can be confidently detected when all reads have the same base

After today, you should have a better understanding of



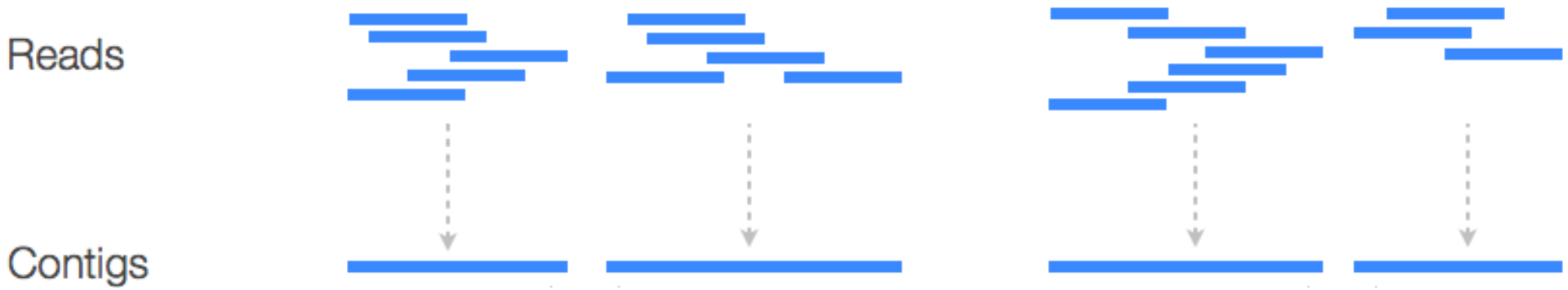
Outputs of genome assembly tools

Contigs are continuous sequences assembled from overlapping reads

Contigs are the first level of assembly, where reads are merged based on overlaps. In other words, they represent reconstructed DNA without gaps (i.e., continuous).

What do contigs indicate?

- Longer contigs suggest better assembly quality.
- Fragmented contigs indicate challenges such as repeats or low coverage.



Contig FASTA files store the reconstructed DNA sequences

What is a contig FASTA file?

- Contains the sequence of each contig in FASTA format.
- Used for downstream analysis like annotation and comparison.

```
>NODE_1_length_251580_cov_96.965763
GCCTTTTTCATATTCTTGAAACATATATAGCAGTACATCTATGTCTACTTTAGGTTTTAT
TGACATAAATAAAGCTCCCTTCAAAGTTTTCATTTTTTCAATGTCTACTTTGAAGGGAGC
ATTTCACTGAACTTTGTTCAGGCTCTTTTTAAATGTATATCAGGCATGGCGGCGACTTGA
TAGTGAAAGTCCATATATGCTTTGTAGTCAAAACTGCTAGCGGATATTGTTATCTTAACA
...
```

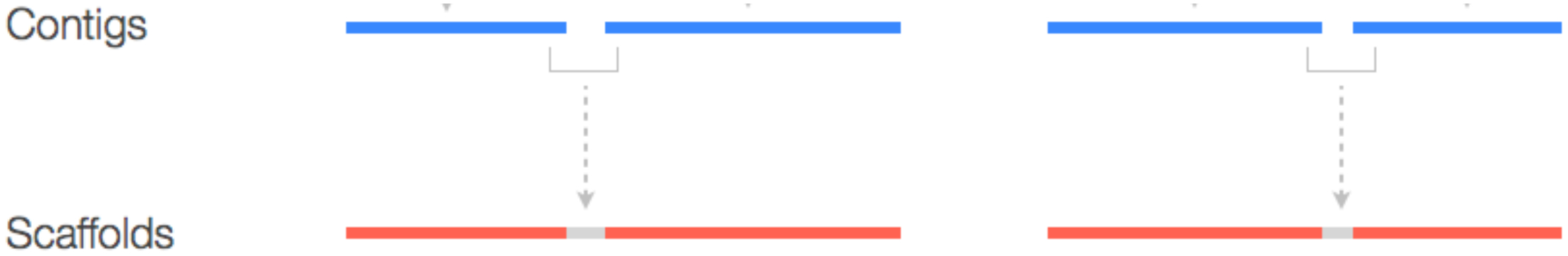
Header format: **NODE_1** is the number of the contig

length_251580 is the sequence length

cov_96.965763 is the k-mer coverage of the largest k used in assembly (will be discussed on Thursday)

Scaffolds use paired-end reads to bridge gaps between contigs

Scaffolds are higher-order assemblies formed by ordering and orienting contigs



Paired-end reads provide distance and orientation information to connect contigs.

What do scaffolds indicate?

- Larger scaffolds suggest fewer gaps and better assembly resolution.
- Remaining gaps in scaffolds are represented as "N" regions.

Scaffold FASTA files combine contigs into longer sequences

What is a **scaffold FASTA file** includes contigs linked by paired-end reads with "N"s as the base

Contigs

```
>NODE_1_length_335019_cov_108.862920
TTATATTGGCAGTAGTTGACTGAACGAAAATGCGCTTGTAACAAGCTTTTTTCAATTCTA
GTCAACCTTGCCGGGGTGGGACGACGAAATAAATTTTGCGAAAATATCATTTCTGTCCCA
CTCCCTAATTTAAACATTTTAAAATATACCAATTACTTTCATCCAAAGTGATCCTAAACC
AATCCAGATAATAAAGTAGACGAAACCTAATATTAAGTTCATTGTCCACCAACGTTTTTG
...

>NODE_5_length_181792_cov_108.741524
TGGCTCTTATGCAGTTGGAGCGAAGATCCAACTGTAAACCATAGTGTACTTATTATTTAT
AATAATTAGTGGCTCTTATGCAGTTGGAGCGAAGATCCAACTGTAAACCATAGTGTACTT
ATTATTTATAATAATTAGTGGCTCTTATGCAGTTGGAGCGAAGATCCAACTGTAAACCAT
AGTGTACTTATTATTTGTAATAATATTGTAGAGTCTGAGACATAAATCAATGTTCAATGC
...
```

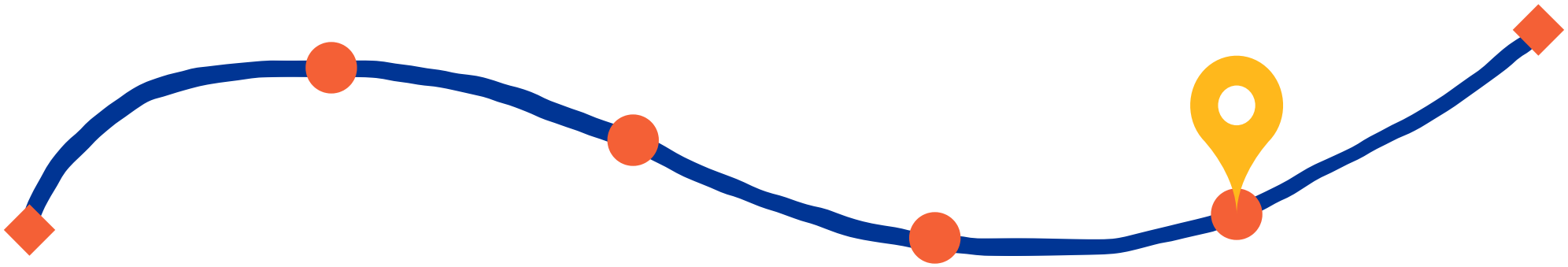
Scaffold

```
>NODE_1_length_335019_cov_108.862920
TTATATTGGCAGTAGTTGACTGAACGAAAATGCGCTTGTAACAAGCTTTTTTCAATTCTA
GTCAACCTTGCCGGGGTGGGACGACGAAATAAATTTTGCGAAAATATCATTTCTGTCCCA
CTCCCTAATTTAAACATTTTAAAATATACCAATTACTTTCATCCAAAGTGATCCTAAACC
AATCCAGATAATAAAGTAGACGAAACCTAATATTAAGTTCATTGTCCACCAACGTTTTTG
...
CATTTAAAATTTCTTGTGACATAGCATTCACCTCCTTTTAGAGCCACTTATTATTTATAA
TAATTAGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNTGGCTCTTATGCA
GTTGGAGCGAAGATCCAACTGTAAACCATAGTGTACTTATTATTTATAATAATTAGTGGC
...
TTACTTTGAAATACTTTAAAAAAATAAGACACTTTCGTA
>NODE_2_length_262462_cov_97.035104
```

Provides a higher-level view of genome assembly, bridging contigs to form scaffolds.

We almost always use this file for downstream processes.

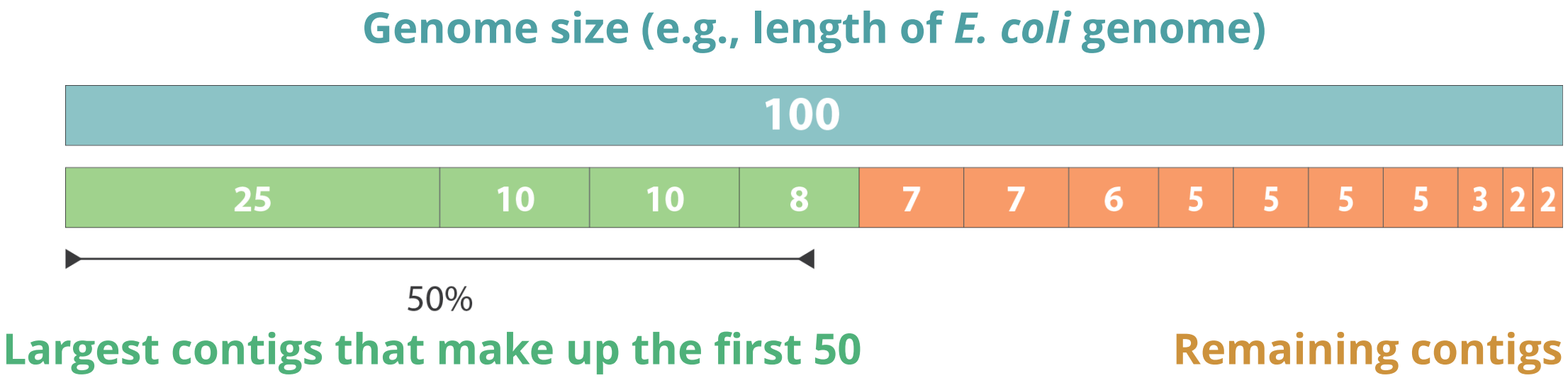
After today, you should have a better understanding of



Assessing assembly quality

N50 is the length of the shortest contig that covers 50% of the assembly

- Sort contigs by length in descending order.
- Add lengths sequentially until 50% of the total assembly length is covered.
- The length of the last contig added is the N50.

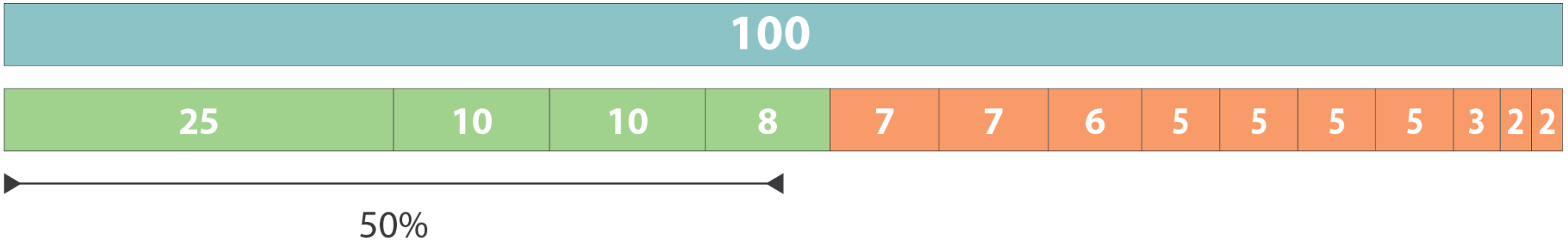


N50 = 8

Higher N50 values indicate more contiguous assemblies

L50 is the number of contigs required to cover 50% of the assembly length

Genome size (e.g., length of *E. coli* genome)



Largest contigs that make up the first 50

Remaining contigs

For L50, count the number of contigs used in the N50 calculation

L50 = 4

Lower L50 values indicate fewer, larger contigs, which is better for assembly quality

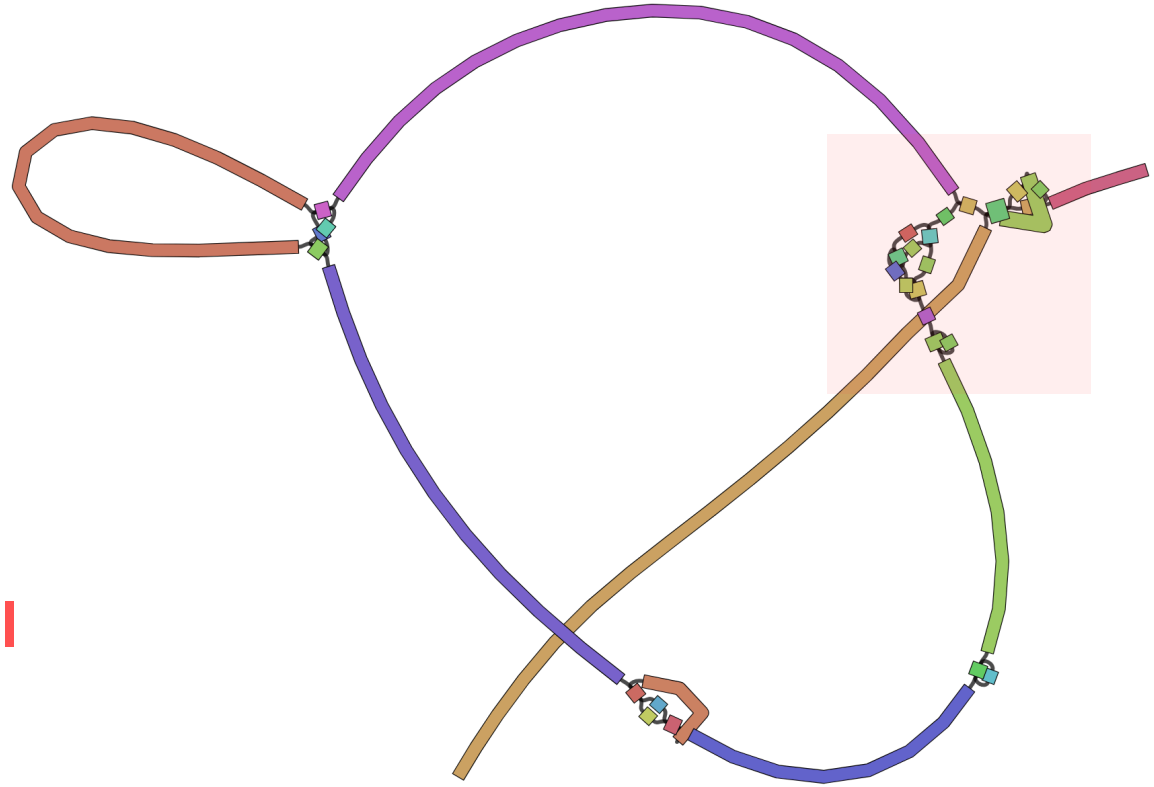
Total assembly length approximates the genome size; deviations could indicate missing data

Bandage helps interpret assembly graphs and identify unresolved regions

We can visualize contigs and how they connect with a Bandage graph

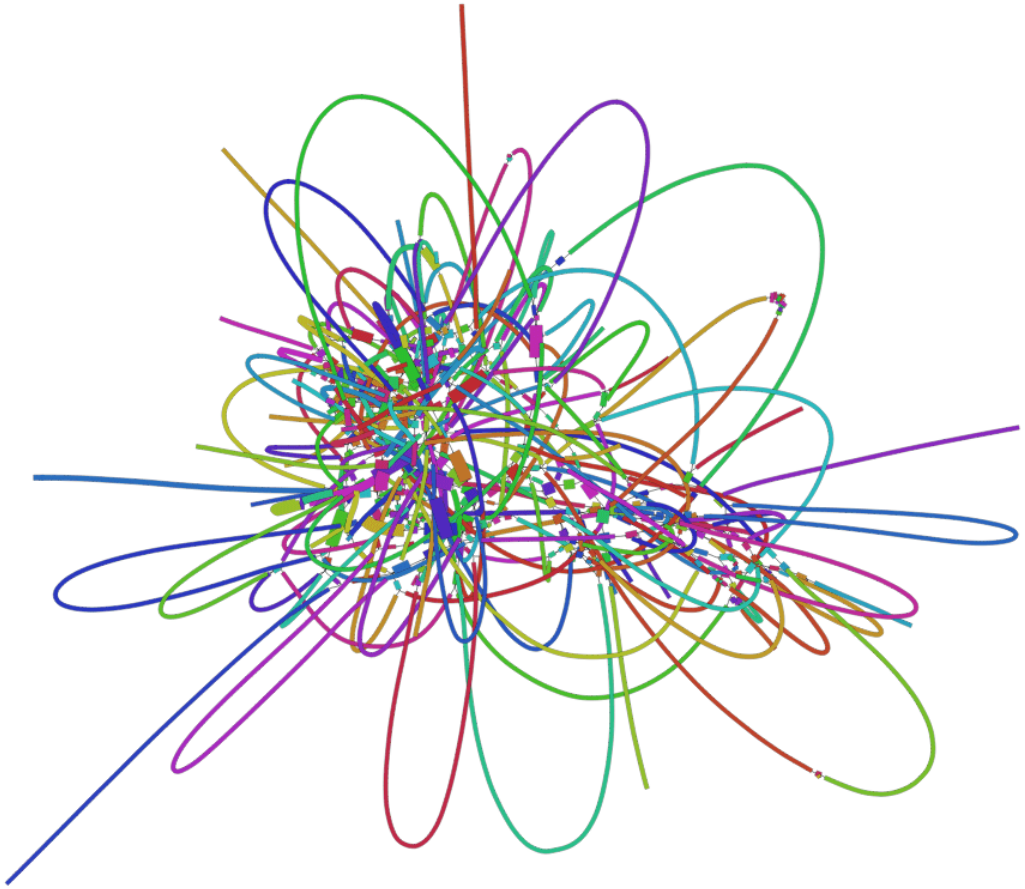
Each colored line is a contig/scaffold

Highly branched assemblies are not ideal



Bandage helps interpret assembly graphs and identify unresolved regions

Here is an example of a real, highly branched assembly.



Islands are sequences that we cannot merge into our assembly above (often sequencing errors)

Before the next class, you should

Lecture 03A:

Genome assembly - Foundations

Lecture 03B:

Genome assembly - Methodology

Today

Thursday

- Work on P01B (due Friday, Jan 24)
- Work on CByte 01
- Review Lectures 02A and 02B for quiz next week