Genome Assembly

BIS180L

Julin Maloof and Matt Davis

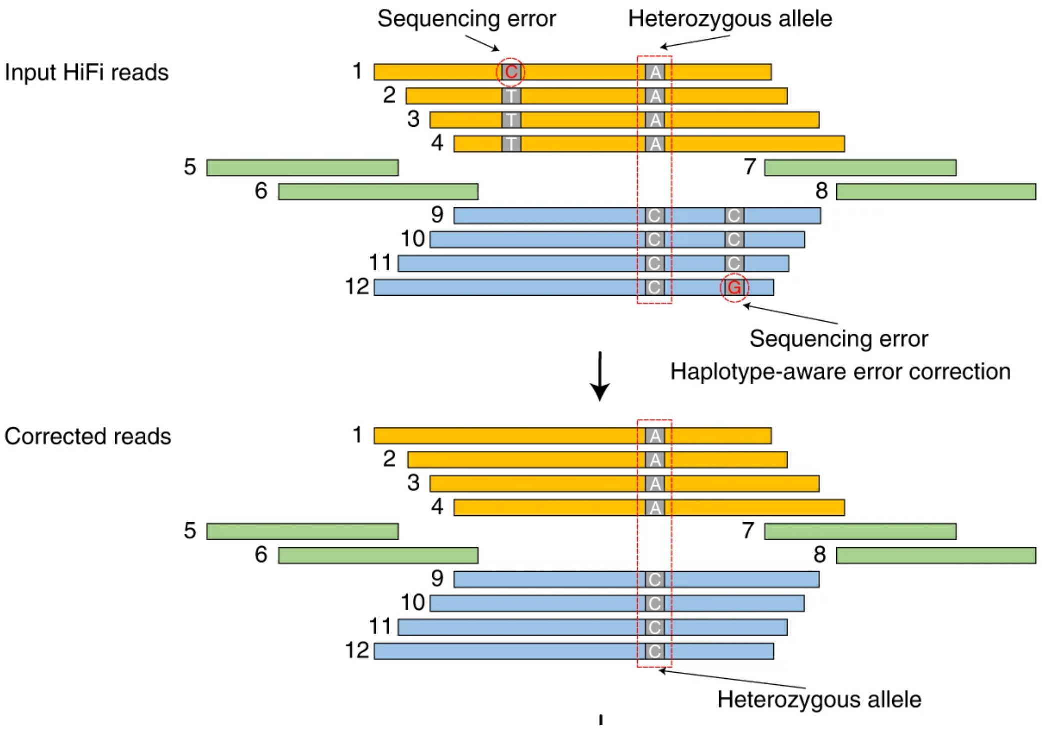
DOWNSIZE INSTANCES

Genome Assembly Overview

- Chromosomes are often 10s to 100s of megabases in length
- Even the longest read technologies only give reads 10-100 kilobases long
- Genetic, genomic, and evolutionary studies are aided (of need) longer contigs / scaffolds
- The ultimate goal is telomere-to-telomere (T-to-T) assemblies

Genome Assembly With HiFiAsm

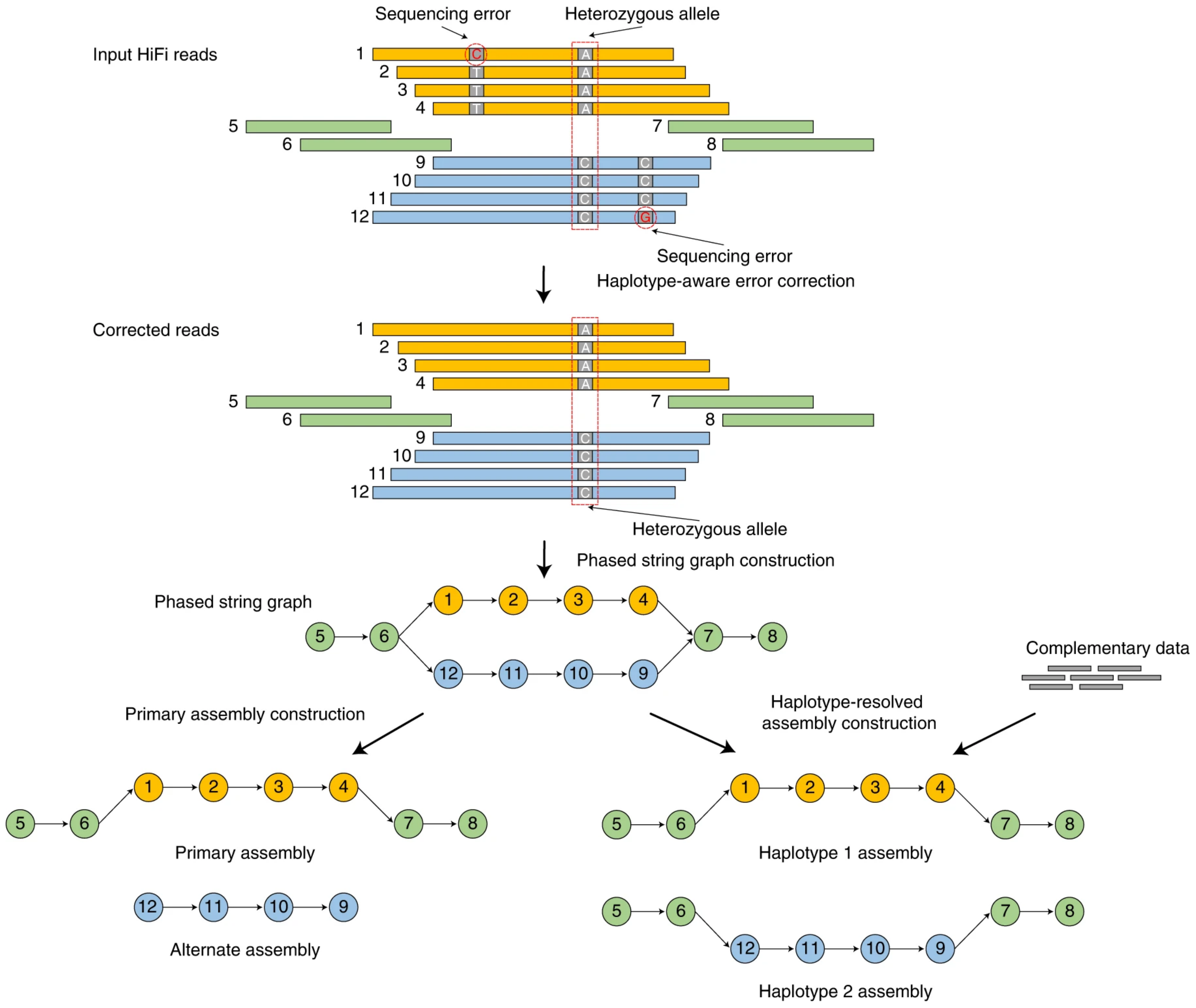
- Align Reads
- Correct Errors



Cheng et al (2021), Figure 1

Genome Assembly With HiFiAsm

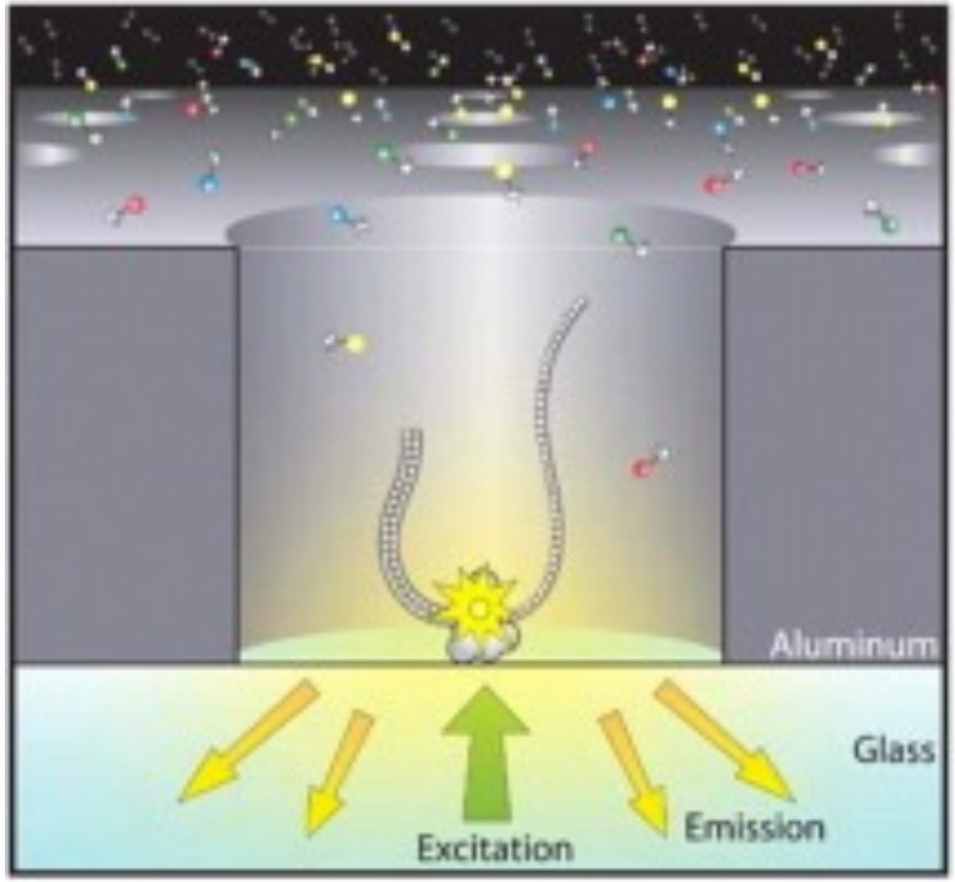
- Align Reads
- Correct Errors
- Create phased graphs
- Separate into haplotypes



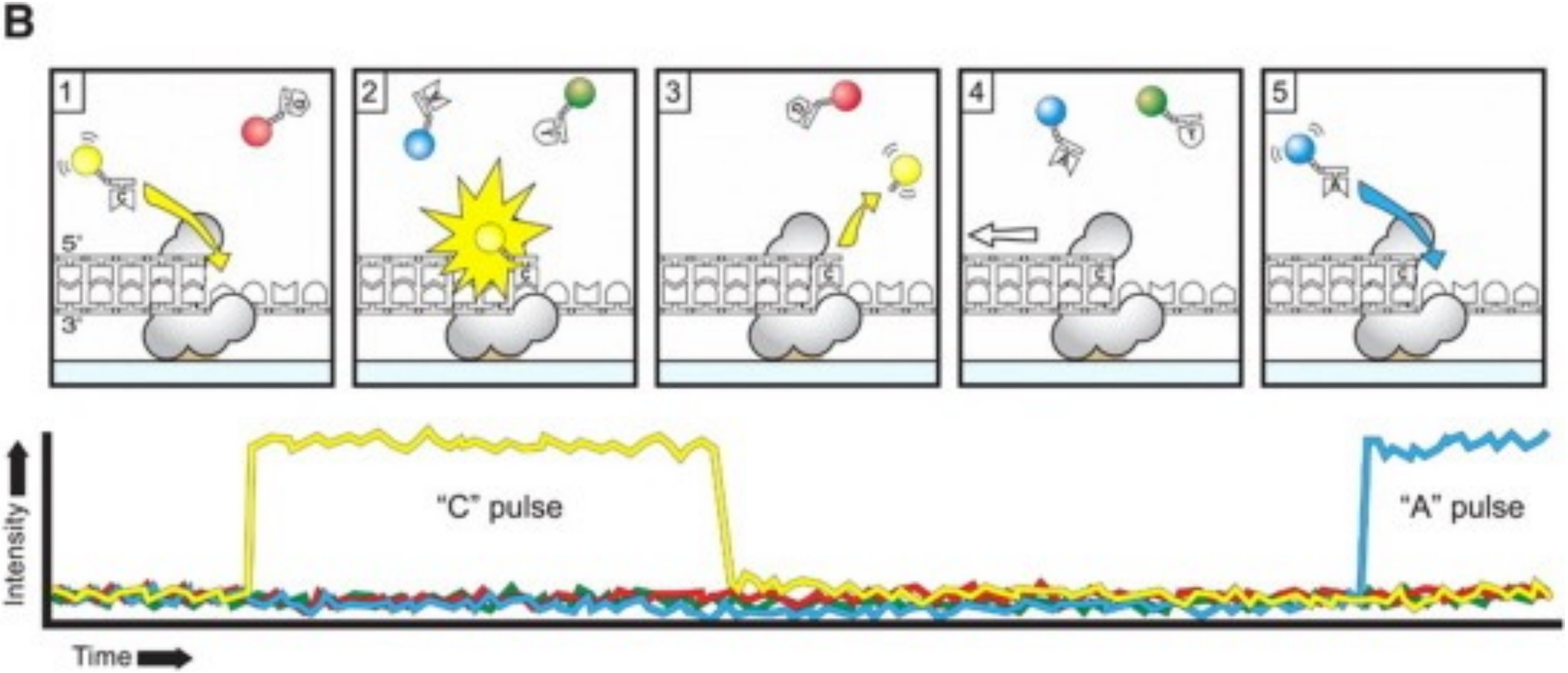
Cheng et al (2021), Figure 1

Pacific Biosciences

- Create a cell with an array of millions of tiny fluorescence detector wells
- Affix DNA polymerase at the detector
- Measure fluorescence of nucleotides as they are added
- Single molecule (no cluster needed)
- Average read length 10,000 – 15,000 bases



Rhoads and Au 2015 https://doi.org/10.1016/j.gpb.2015.08.002



Rhoads and Au 2015 https://doi.org/10.1016/j.gpb.2015.08.002

PacBio Video

- https://www.youtube.com/watch?v=_lD8JyAbwEo

Read Statistics

- Quality
- Min, average and max read length
- N50:
  - Sort reads or contigs from largest to smallest
  - Compute the running sum
  - N50 is the length of the shortest read/contig where 50% of the total sequence is contained
- Translation: 50% of the data is on reads/contigs of at least X bp in length
- What is N50 for the sequences on the right?

Length	Cumulative Sum
1543	1543
1500	3043
1323	4366
1301	5667
1276	6943
888	7831
789	8620
777	9397
743	10140
701	10841
654	11495
622	12117
300	12417
121	12538

Assembly Statistics: General

- Min, avg, max contig length
- # of contigs
- N50, N90
- How do contigs compare to expected number of chromosomes?
- How does assembly length compare to expected genome size?
- Does contigs go telomere to telomere?

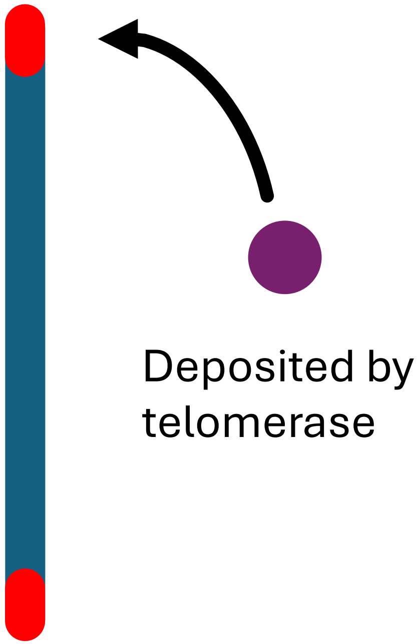
Assembly statistics: BUSCO

- Does the genome contain expected genes?
- BUSCO:
  - Benchmarking Universal Single Copy Orthologs
  - Defined set of expected orthologs
  - Different sets for different branches on the tree of life:
- Assembly evaluation:
  - What % of BUSCO genes are
    - Present in the assembly
    - Complete vs fragmented
    - Single copy vs duplicated

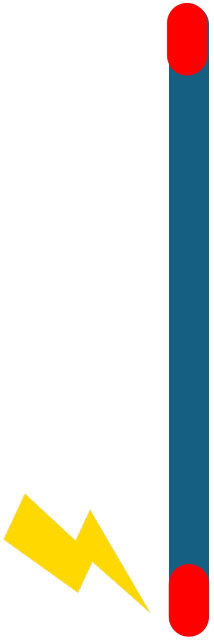
CONDA Environments

- Today we are going to use something called CONDA
- CONDA is a tool for managing packages, programs, and add other add-ons
- Originally developed for Python, useful for other languages as well
- You can create separate "environments" for different tasks or sets of programs
- This helps prevent version or package conflicts and can also keep things reproducible
- We are using a conda environment for this lab…you will see the code to activate it and install a package

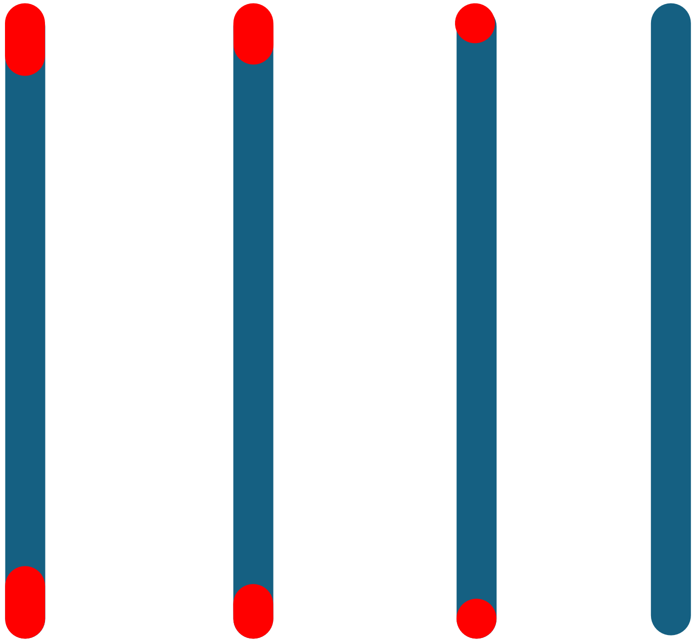
What is a telomere?



Telomeres are repetitive sequence at the ends of chromosomes



Prevent chromosomal degradation



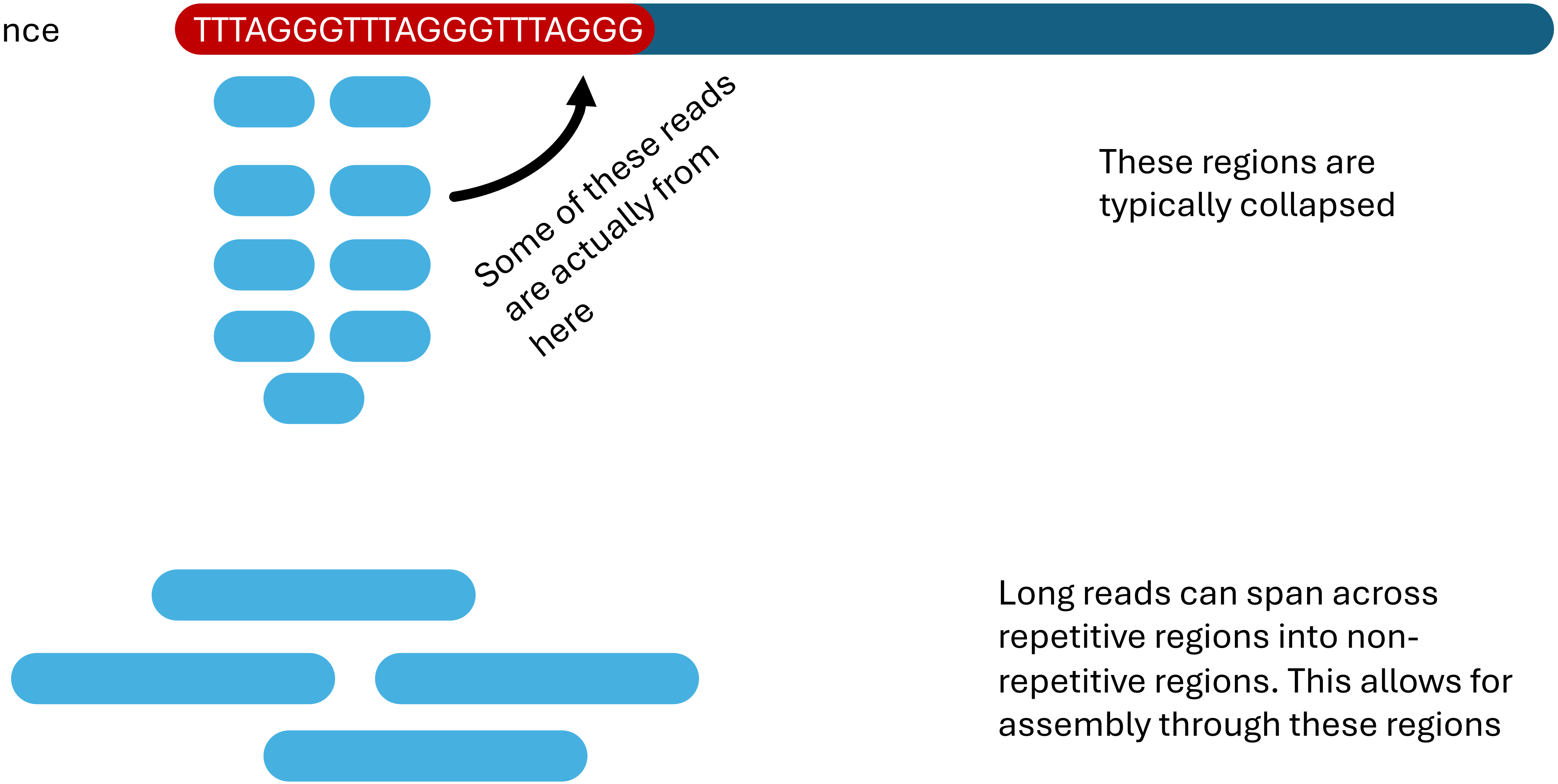
Shortening is associated with aging in somatic cells

Telomeres are hard to assemble

Actual sequence

Short Reads

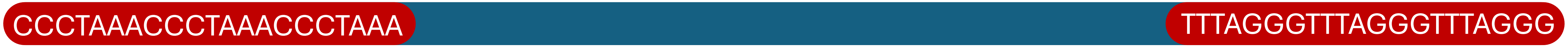
Long Reads



These regions are typically collapsed

Long reads can span across repetitive regions into non-repetitive regions. This allows for assembly through these regions

What is a telomere-to-telomere assembly?



CCCTAAACCCTAAACCCTAAA

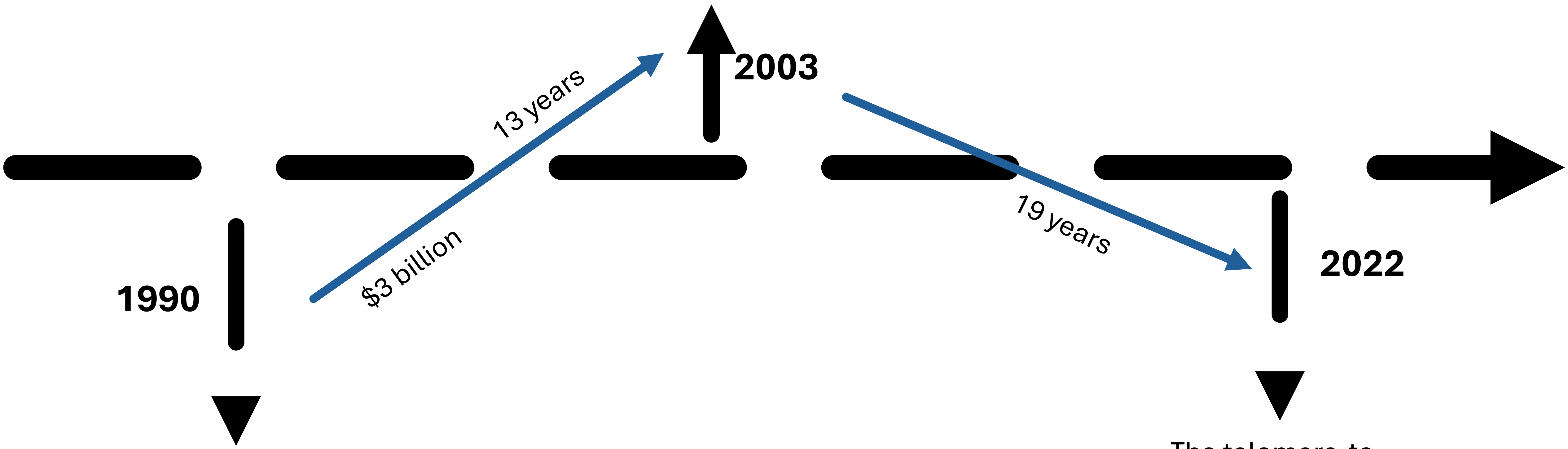
TTTAGGGTTTAGGGTTTAGGG

A genome assembly where every chromosome is sequenced from telomere-to-telomere

This is very difficult, but is now the gold standard for assemblies

Long read sequencing (like HiFi) along with better assembly software (like HiFiAsm) have enabled telomere-to-telemore assembly

The telomere-to-telomere human genome was just published in 2022



1990

The Human Genome Project officially begins

$3 billion

13 years

2003

The first human genome is published

19 years

2022

The telomere-to-telomere human genome is published

It is such a big deal that NCBI has a website for it.

Important:

Run the methylation code at the very end of today’s lab before leaving.

We need the results for Thursday