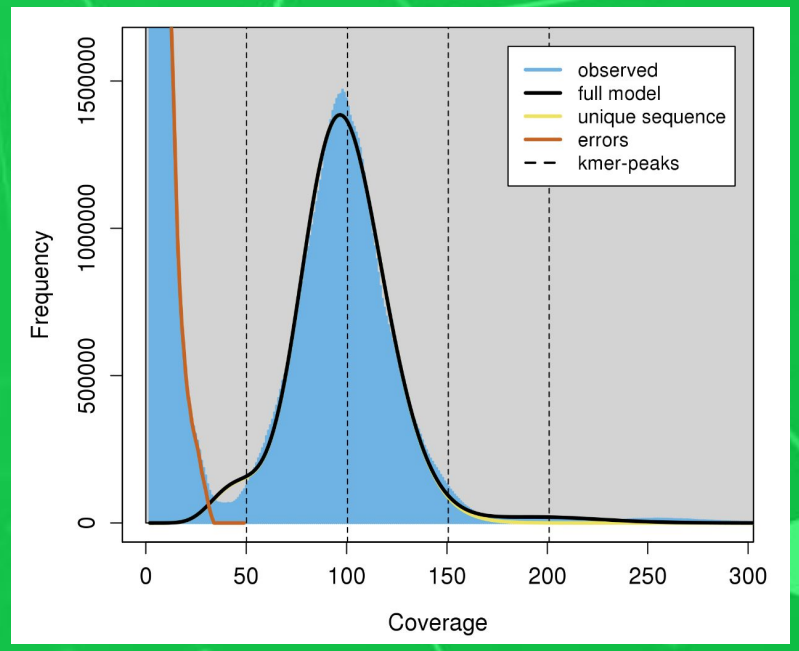
Centre for Genomic Research

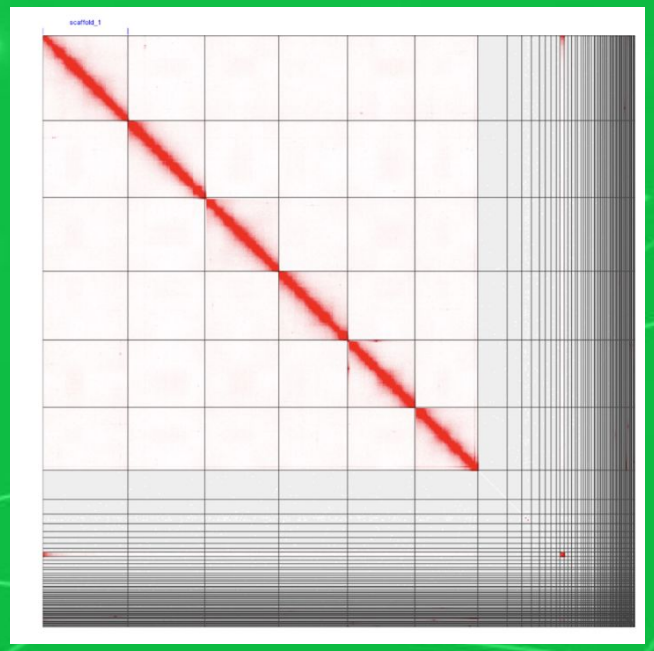
UKRI Natural Environment Research Council

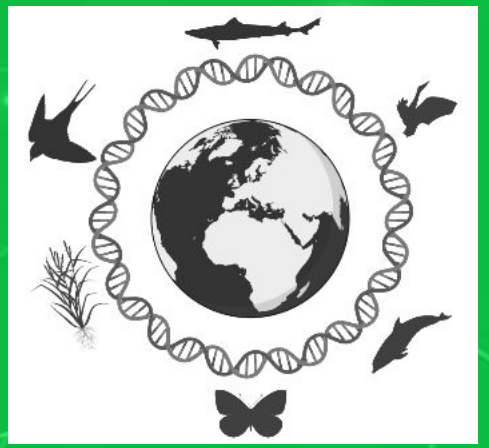
NEOF NERC ENVIRONMENTAL OMICS FACILITY

wellcome sanger institute

# Eukaryote Genome Assembly





UNIVERSITY OF LIVERPOOL

The University Of Sheffield.

UNIVERSITY OF LIVERPOOL

Natural Environment Research Council

Centre for Genomic Research

NEOF

NERC ENVIRONMENTAL OMICS FACILITY

The University Of Sheffield.

wellcome sanger institute

Websites

NEOF: https://neof.org.uk/

NERC: https://nerc.ukri.org/

CGR: https://www.liverpool.ac.uk/genomic-research/

ToL:https://www.sanger.ac.uk/programme/tree-of-life/

Twitter

NEOF: @NERC_EOF

NERC: @NERCscience

CGR: @CGR_UoL

# Upcoming workshops

https://neof.org.uk/training/

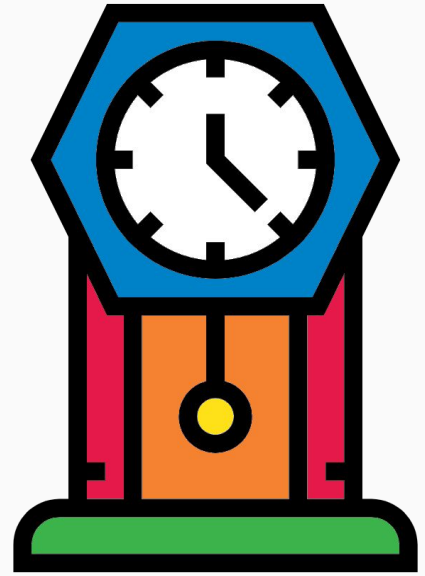
- ONLINE:
  - Community analysis in R
    - 9th & 11th May 2023
  - RNA-seq gene expression and pathway analysis
    - 6th & 8th June 2023
- IN PERSON (Sheffield)
  - Population Genomics - July 2023 TBA
  - Metabarcoding for diet analysis and environmental DNA - Sept 2023 TBA
- More!

# Format & Schedule

This intro
Bookdown
Theory
Practice
Exercises
Optional materials

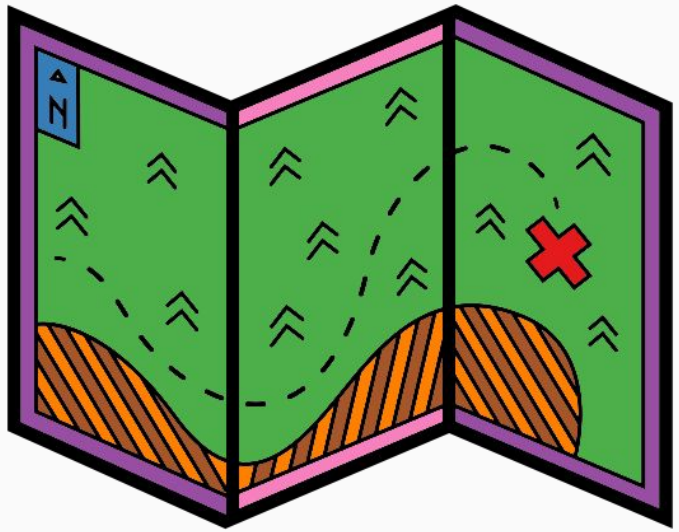
Work at your own pace
We are here to help
Time with breaks in between

- 10:00-11:15
- 11:30-12:30
- 13:30-14:45
- 15:00-16:00

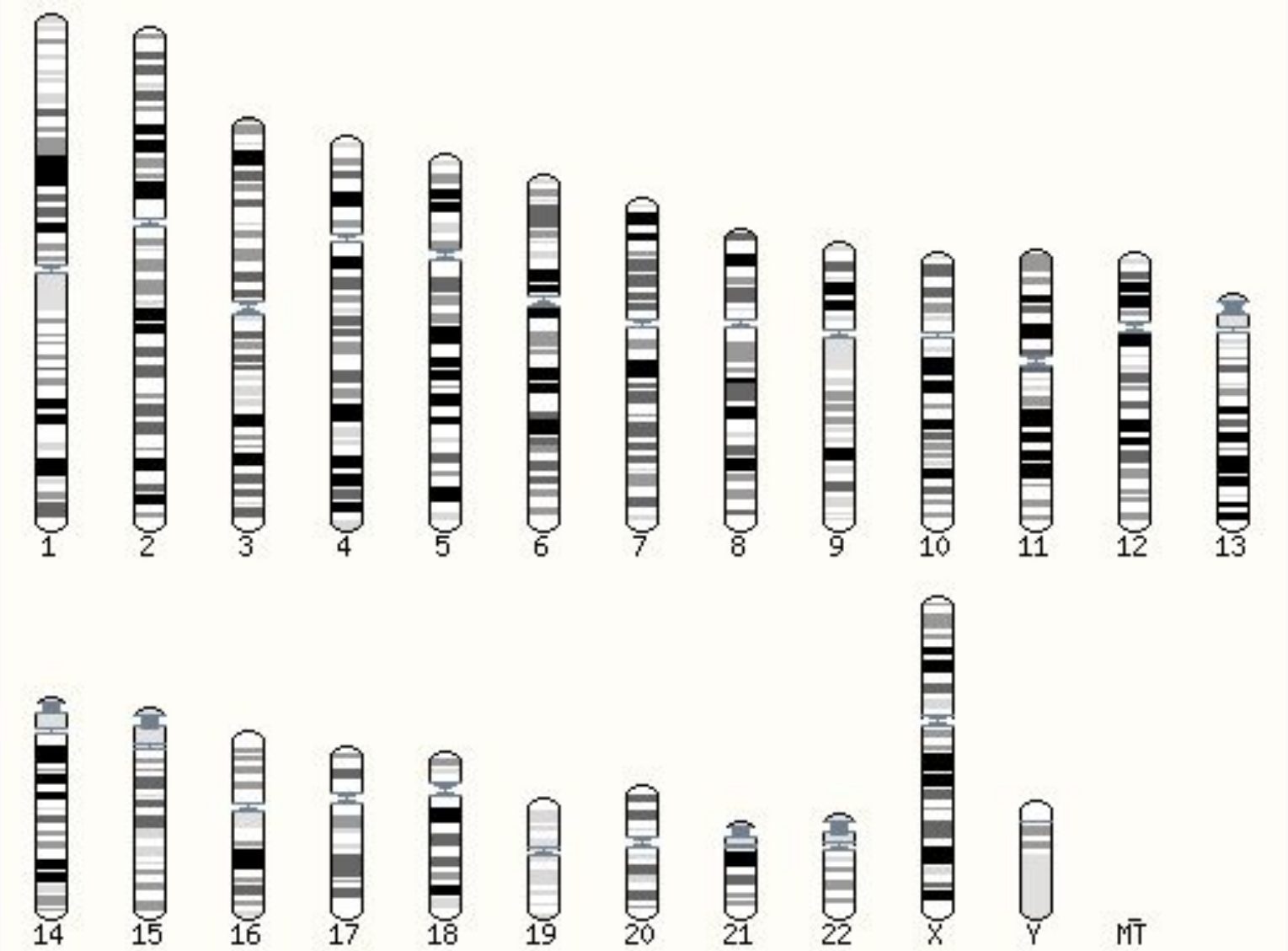


# Outline for Today

- Sequencing technologies
- Pre-assembly
- Assembly
- Assessment

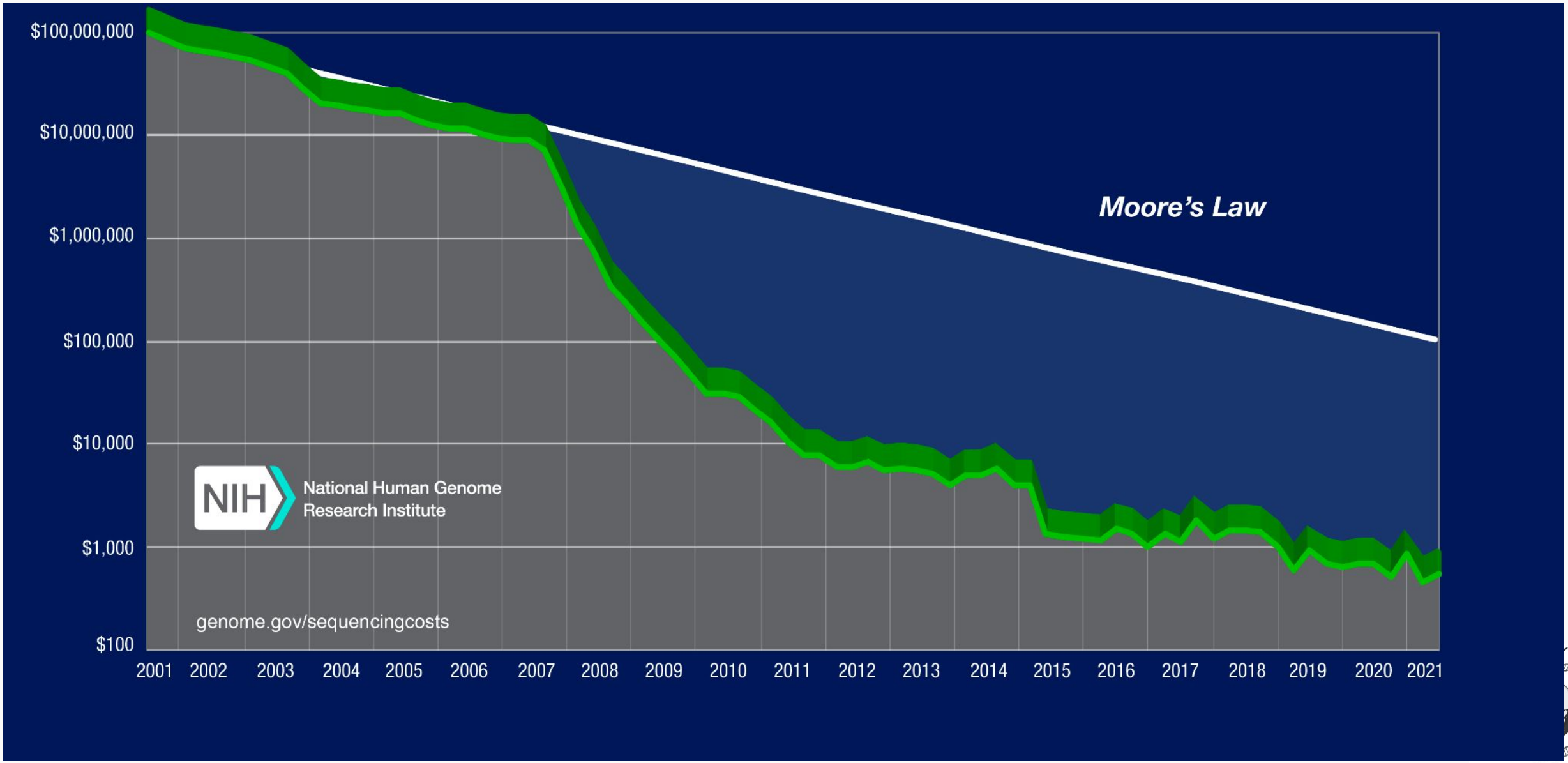


# Human genome sequence




- 3.2 Gb genome
- Published 14 Apr 2003
- Sanger sequencing
- Took 13 years
- Cost $3 billion

# Cost per human genome



# Next generation sequencing

Much higher degree of parallelism than Sanger sequencing

Much lower costs

Different platforms differ in terms:

- read lengths
- bp output
- costs of run
- costs of library preparation
- error rates

illumina

PACBIO®

Oxford NANOPORE Technologies

# illumina

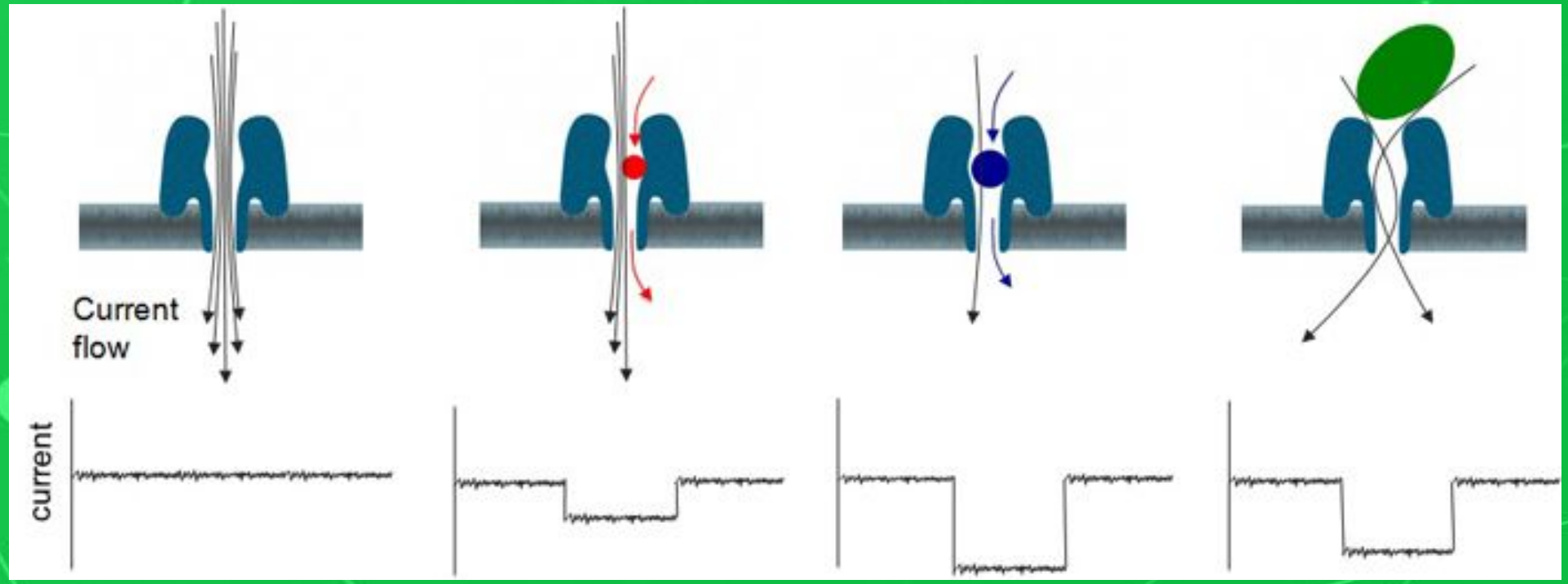
- High output; cheap (£/Mb); low error rate; short reads
- Many tools
- Low Insertion/deletion errors
- de novo genome or transcriptome sequencing, re-sequencing, GBS/capture sequencing


PACBIO®

Oxford NANOPORE Technologies

- 10+ kb reads
- de novo genomes, scaffolding
- Lower quality than Illumina, although has improved
- Can span over long repeats
- De novo assembly, detection of structural variants
- Transcriptomics: full-length transcripts/isoforms
- Epigenetics

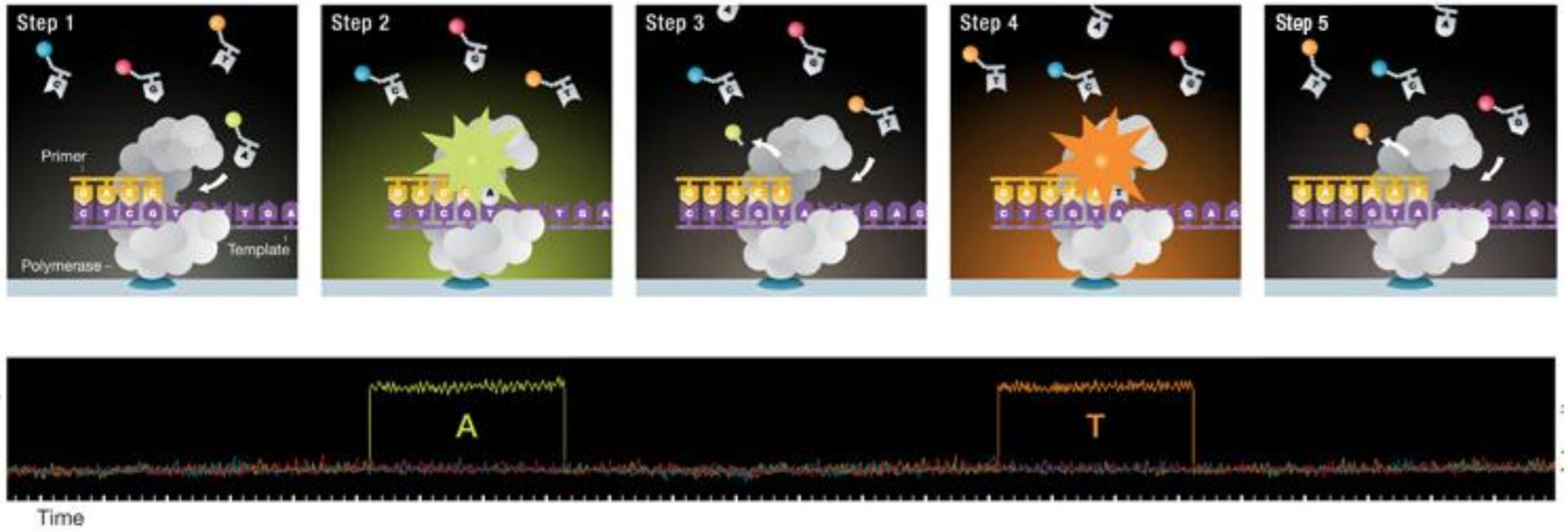
# Single molecule sequencing - nanopore




- No PCR.
- Large (unlimited?) maximum read length.
- 500 – 8000 nanopores, 500bp per minute.

https://nanoporetech.com/resource-centre/how-nanopore-sequencing-works-animation

# PROCESSIVE SYNTHESIS WITH PHOSPHOLINKED NUCLEOTIDES




**Step 1**: Phospholinked nucleotides are introduced into the zero-mode waveguide (ZMW)

**Step 2**: The nucleotide is held in the detection volume for tens of milliseconds, fluoresces when excited by light. The captured light is converted into a base call with associated quality metrics

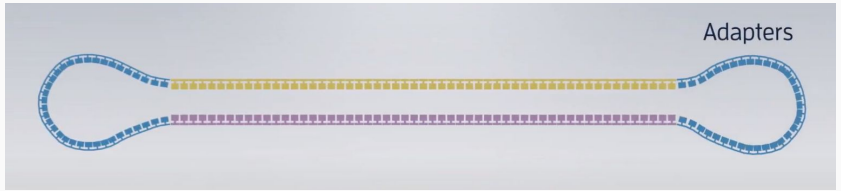
**Step 3**: The polymerase incorporates the nucleotide, releasing the attached dye molecule

Pulses of light are converted in real time into base calls, kinetic measurements are recorded, and base quality values (QVs) are generated

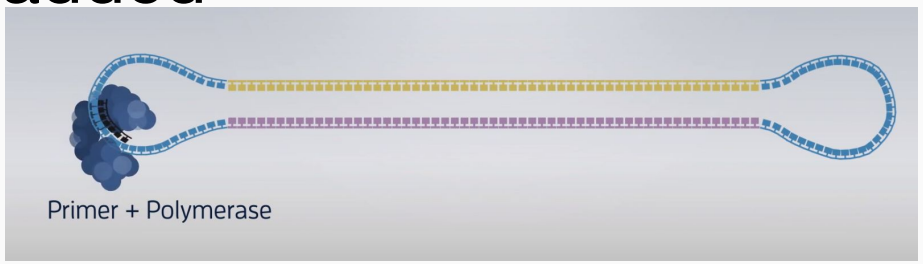
# SMRT (Single Molecule, Real-Time) Sequencin

1: Isolated DNA/RNA

2: SMRTbell Library
(Adapters ligated to dsDNA)




3. Primer and Polymerase added



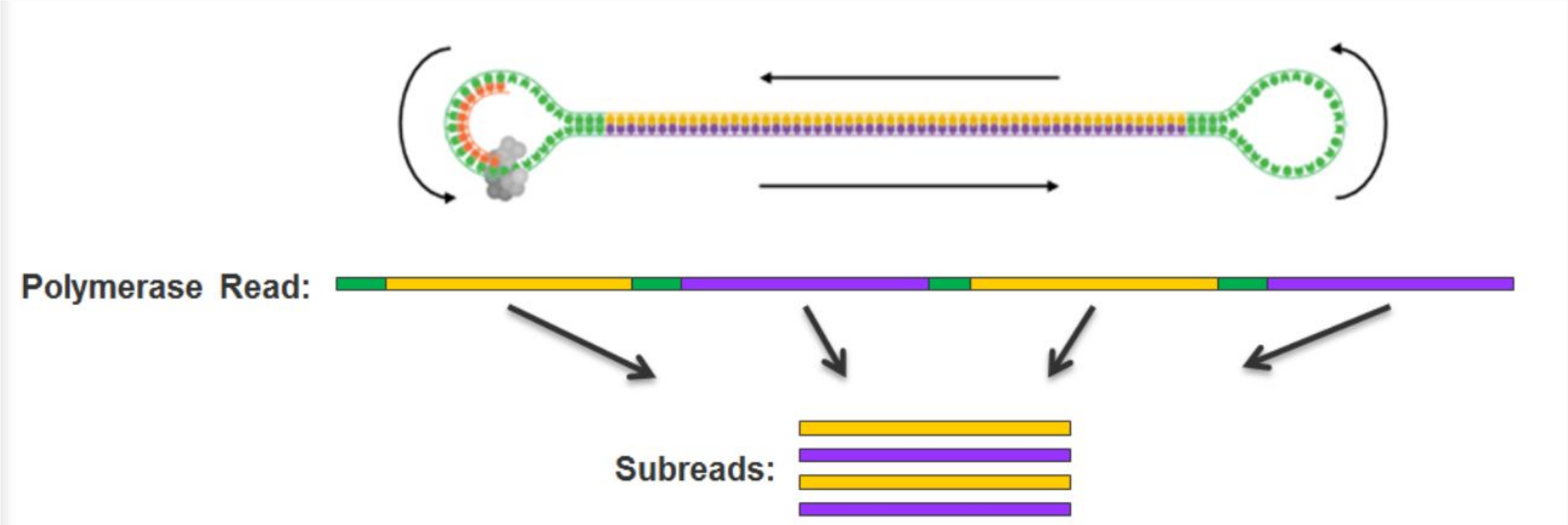

4: One of millions of wells in Flow Cell.

Wells = Zero-Mode Waveguides (ZMWs)

One molecule of DNA / ZMW

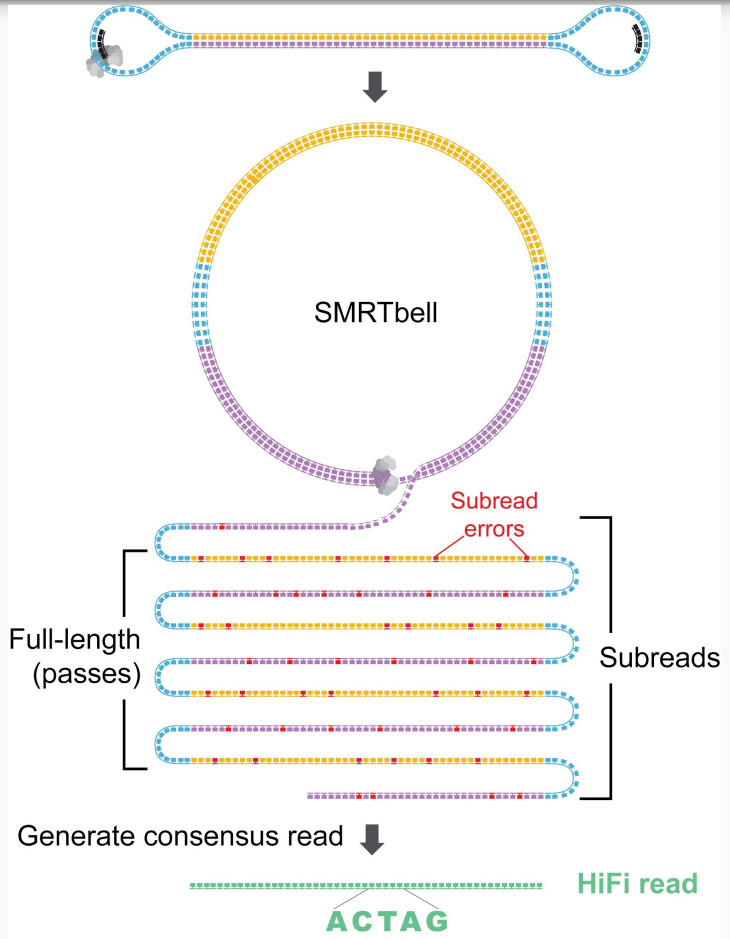
# PacBio Polymerase reads and Subreads

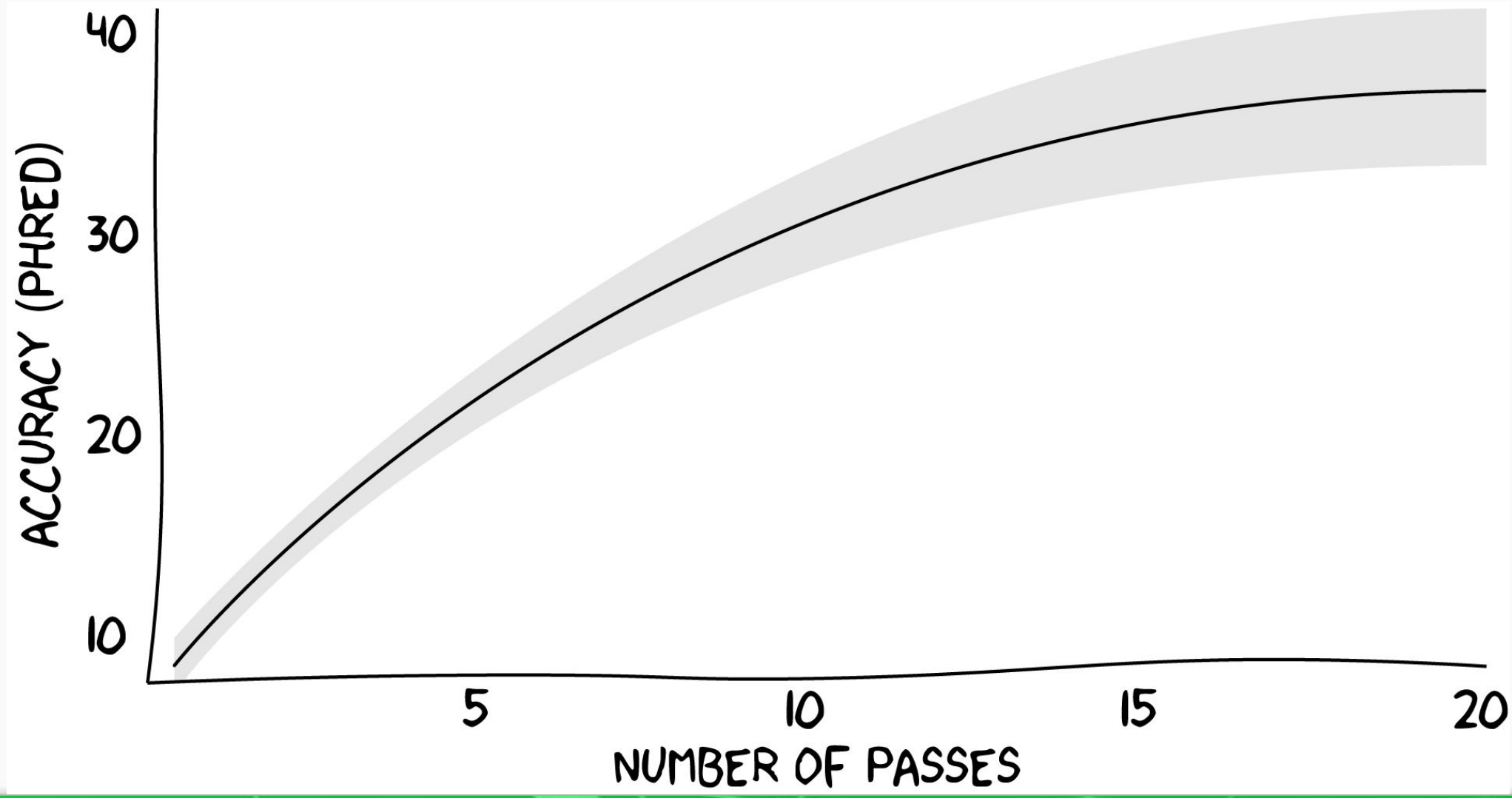
Subreads are commonly used for many applications e.g de novo assembly, resequencing, base modification analysis.



# Circular Consensus Sequencing (CCS)

- Multiple subreads combined into one HiFi read.
- Statistical model
- HiFi read: Highly accurate consensus sequence.
- https://ccs.how/

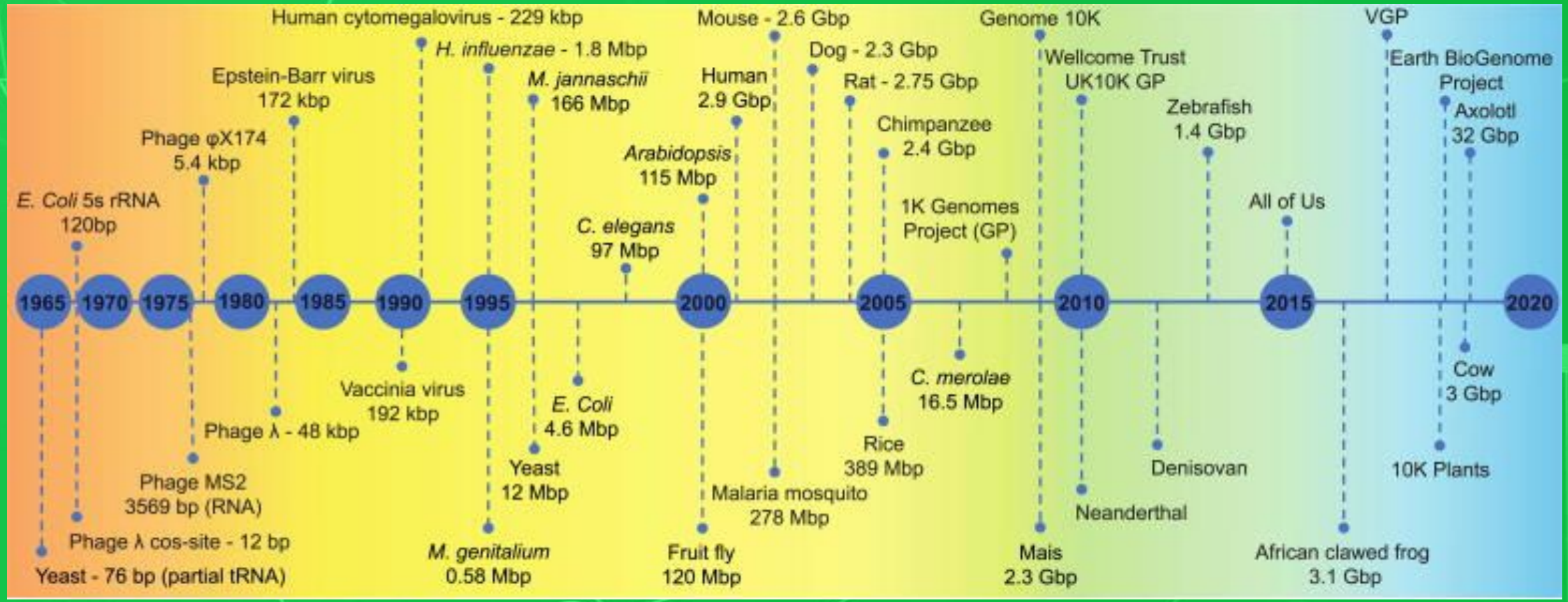





https://ccs.how/faq/accuracy-vs-passes.html

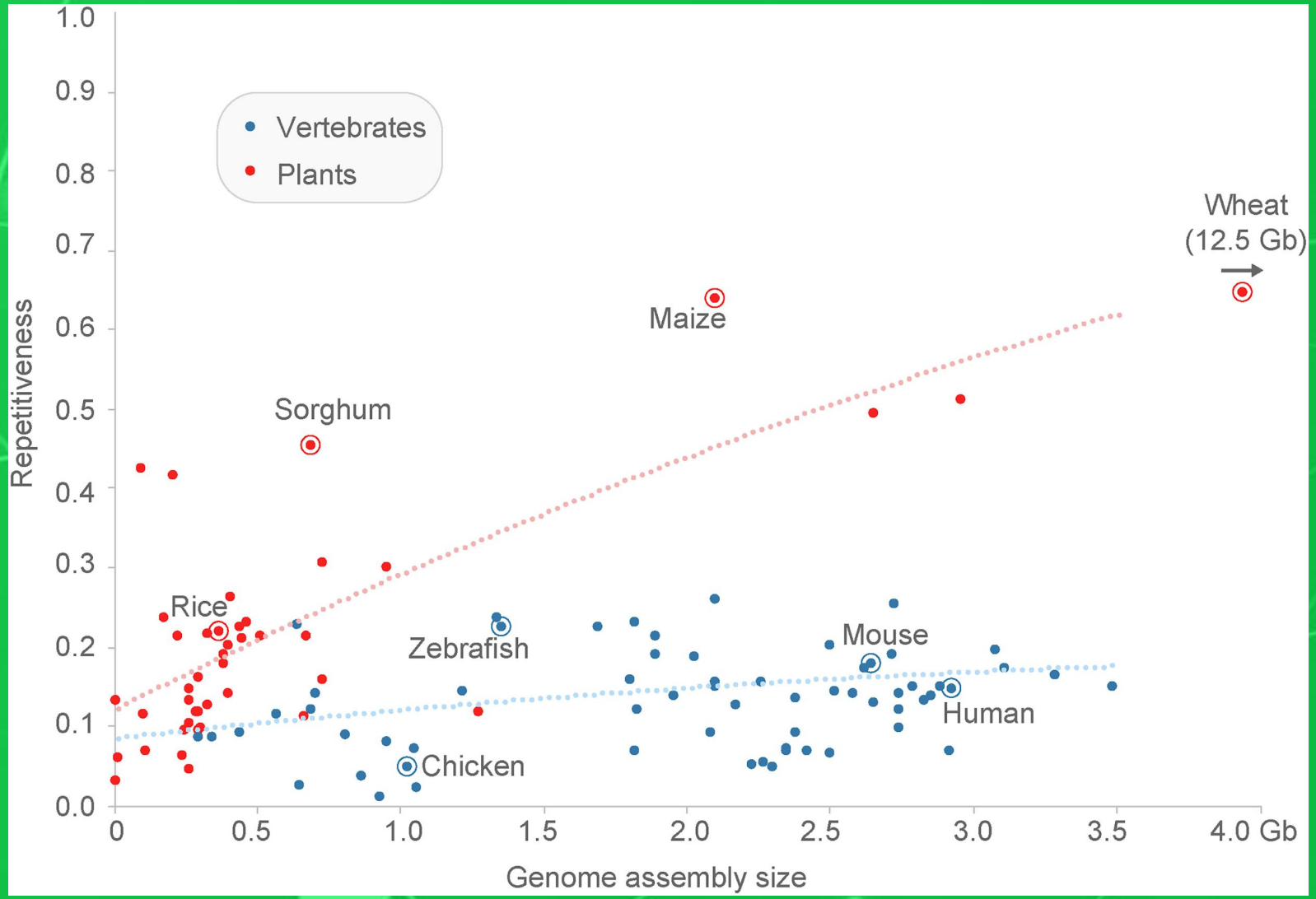
# Genome assembly: challenge

# Milestones in genome assembly




- Comput Struct Biotechnol J. 2020; 18: 9–19.

# Size and repetitiveness: plants vs animals



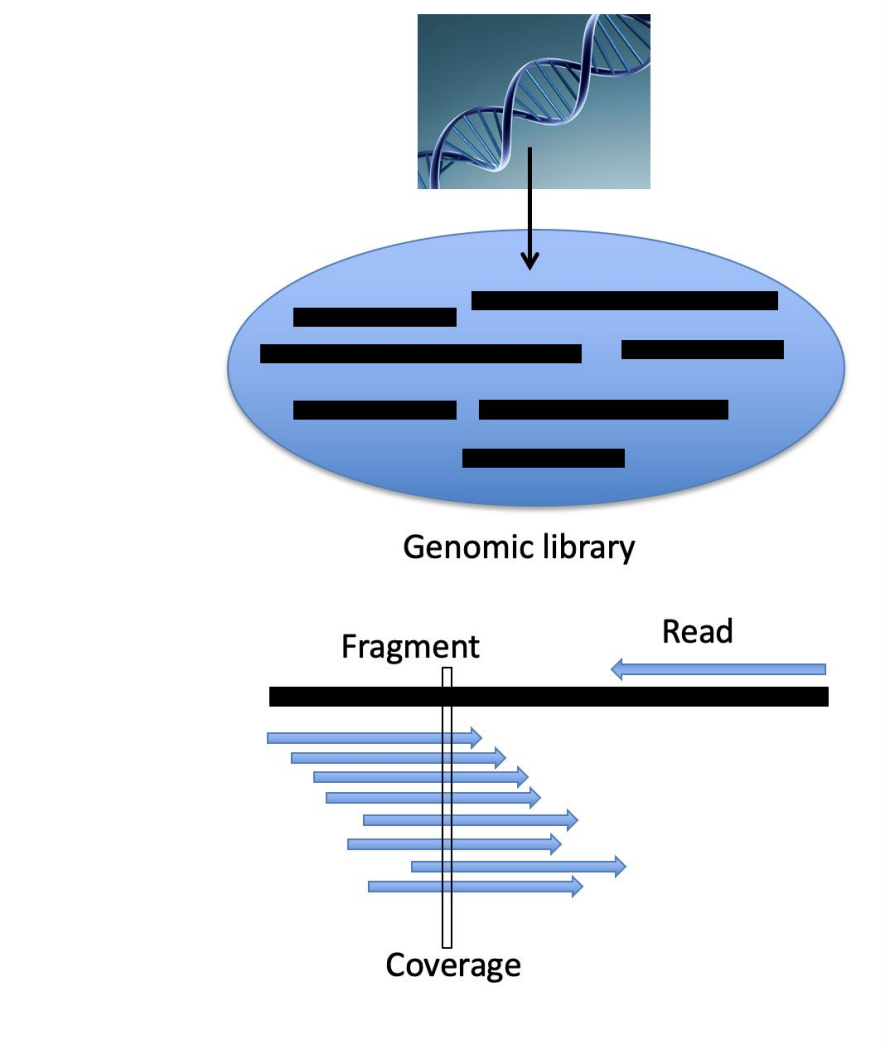

- Current Opinion in Plant BiologyVolume 36, April 2017, Pages 64-70

# Genome sequencing - Basic concepts

- **Genomic library** – collection of DNA from source organism divided into multiple fragments

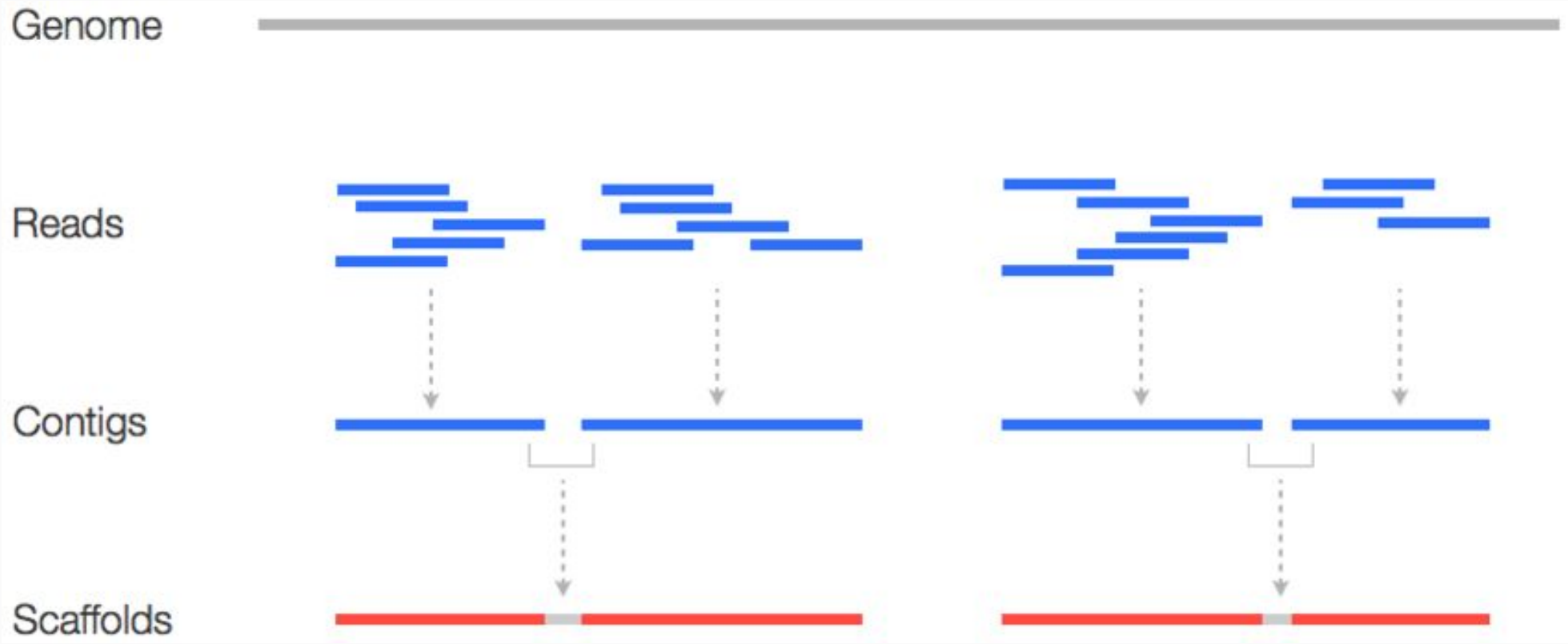
- **Read** – Substrings of genomic sequence output by sequencer

- **Sequencing coverage/depth** - average number of reads representing a given nucleotide in the reconstructed sequence

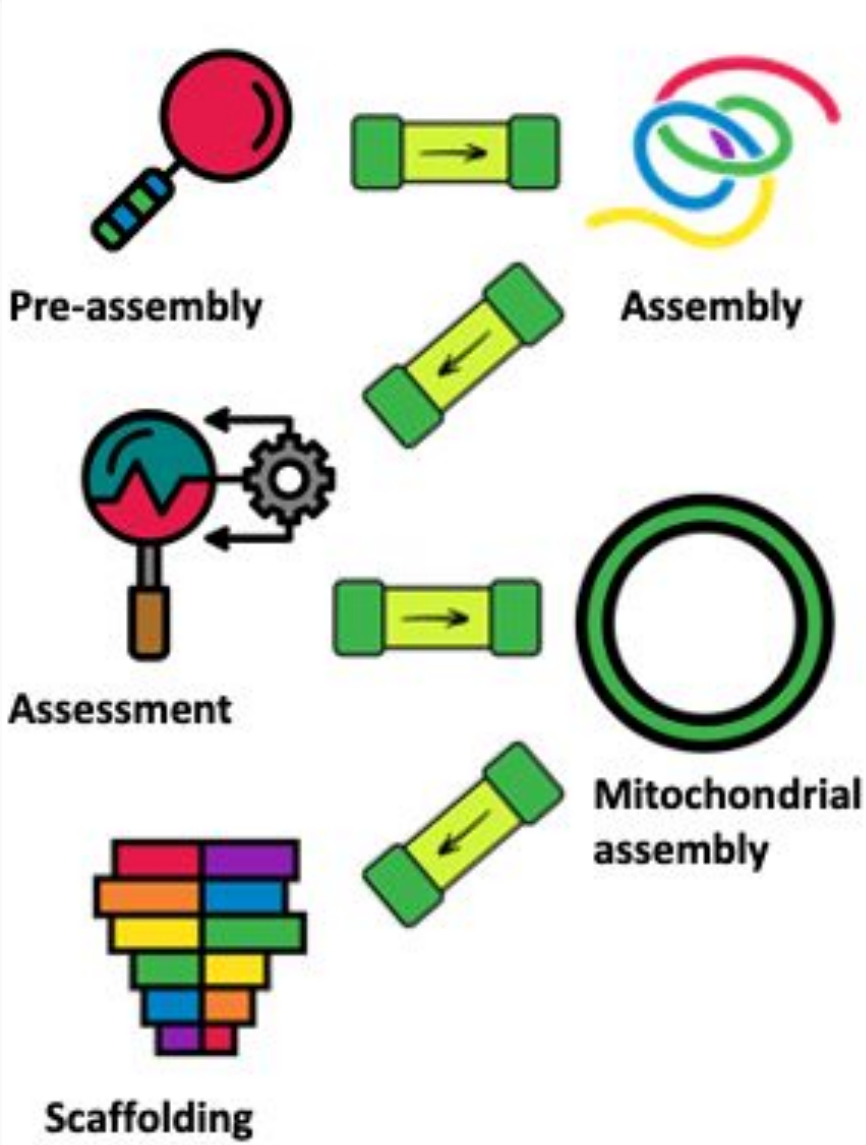


# Genome assembly

Assembling reads into contigs
Contig = contiguous sequence
Perfect contig = chromosome/plasmid/etc



# Genome assembly workflow



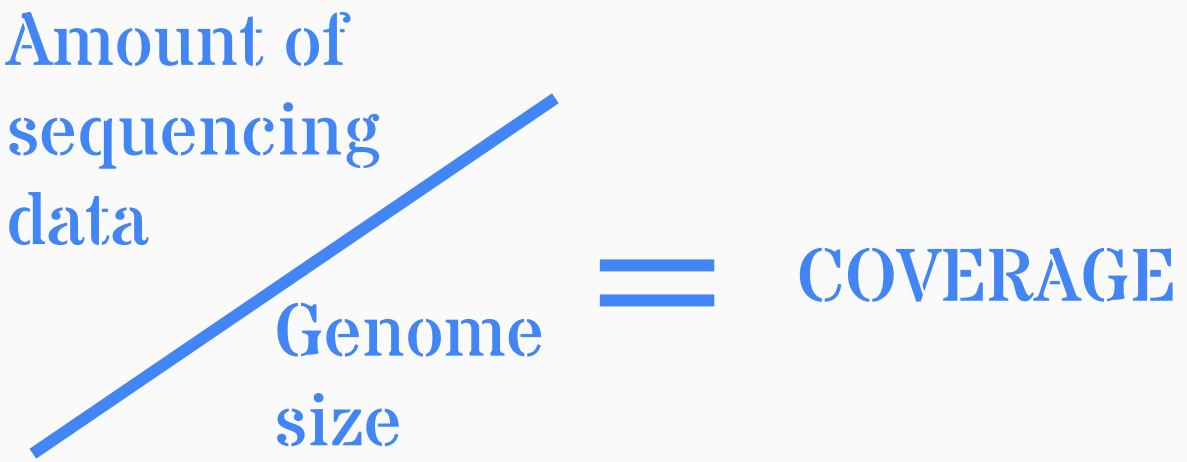
# Pre-assembly: Coverage

## Counting reads and base

- **Sequencing coverage/depth** - average number of reads representing a given nucleotide in the genome

- **SeqKit stats** – tells us the Mb of data in our sequence file

- **Genome size** – databases, NCBI.





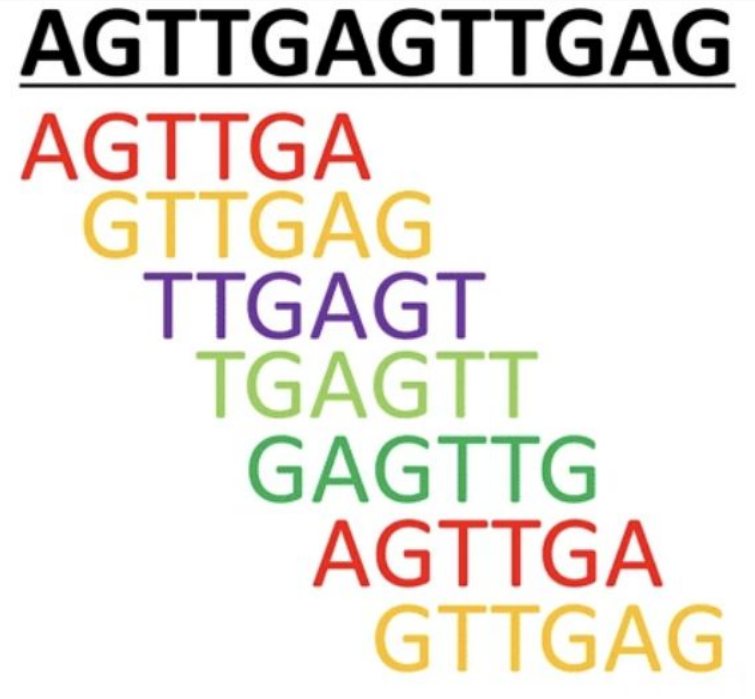
https://cvalues.science.kew.org/

https://www.genomesize.com/

# Pre-assembly: k-mers

## Counting and analysing k-mers

- **k-mer** – A sequence of **k** nucleotides in a DNA sequence
- **Coverage** and **heterozygosity** estimates
- Number & frequency of k-mers to estimate **genome size**
- **Odd values** for k
- **Length of k:** too short – not sensitive enough, too long – computational limitations

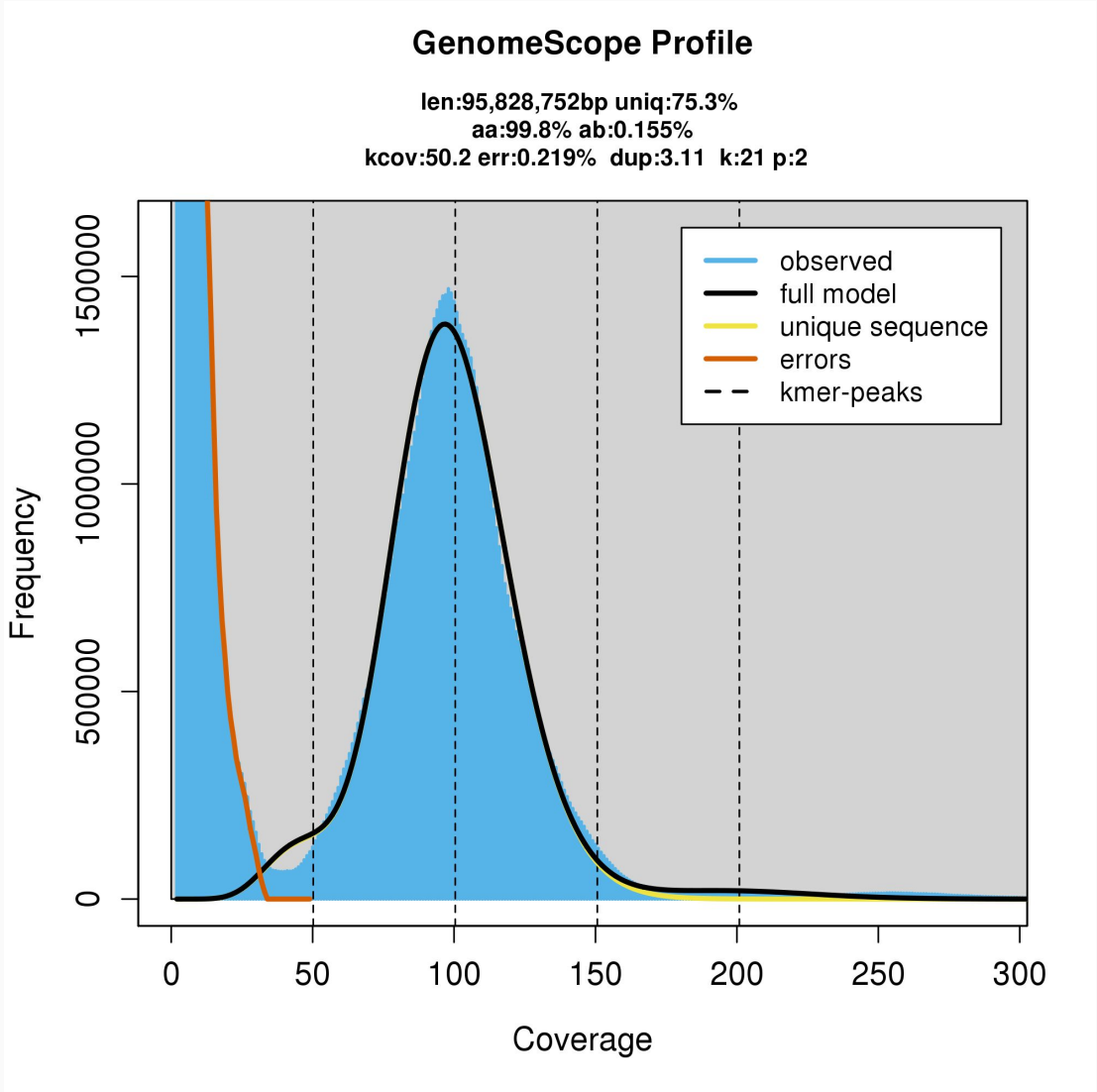


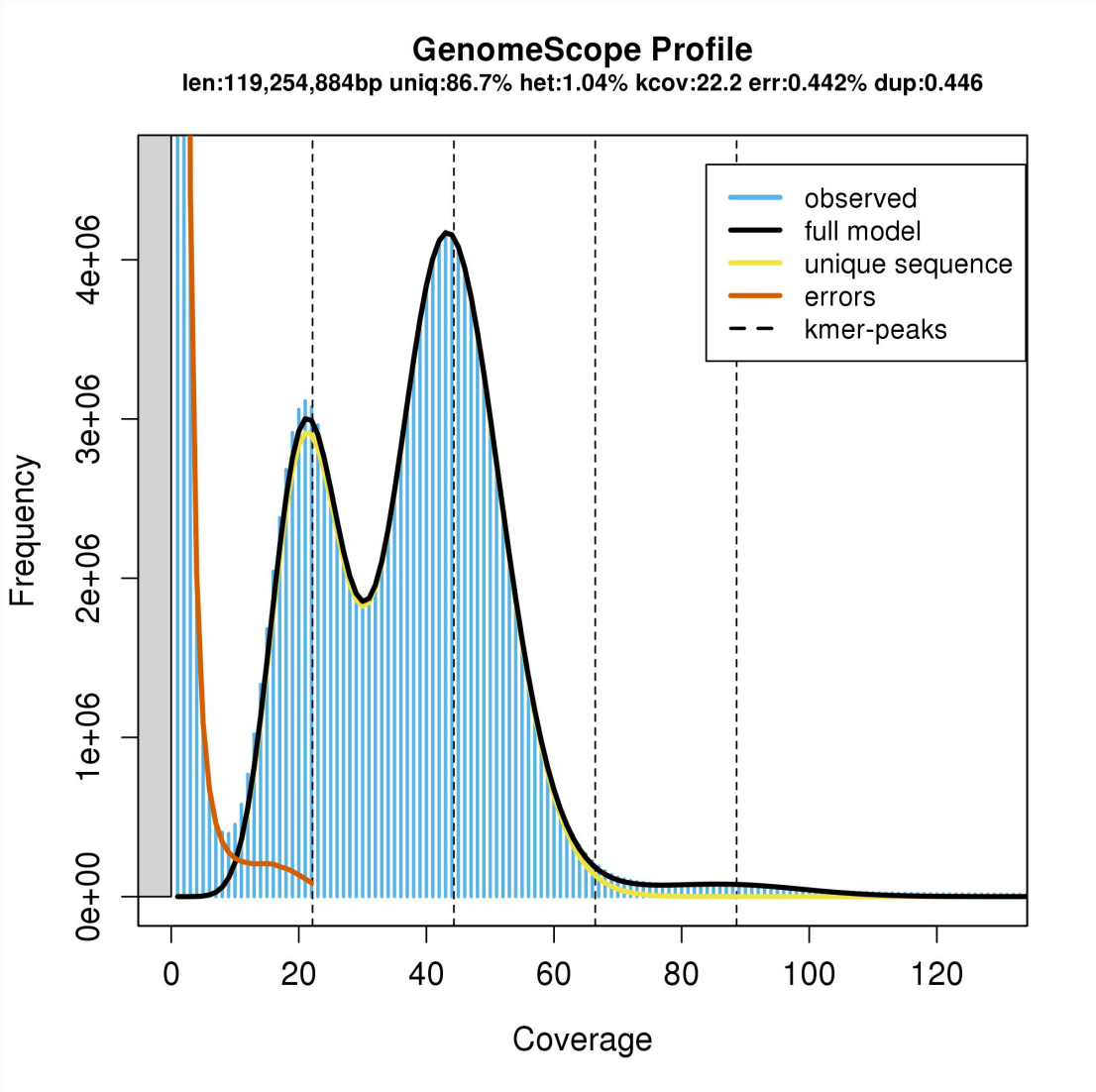

k-mer length of 6

https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-019-5467-x#Fig1

# Pre-assembly: k-mers

## Interpreting k-mer counts - GenomeScope



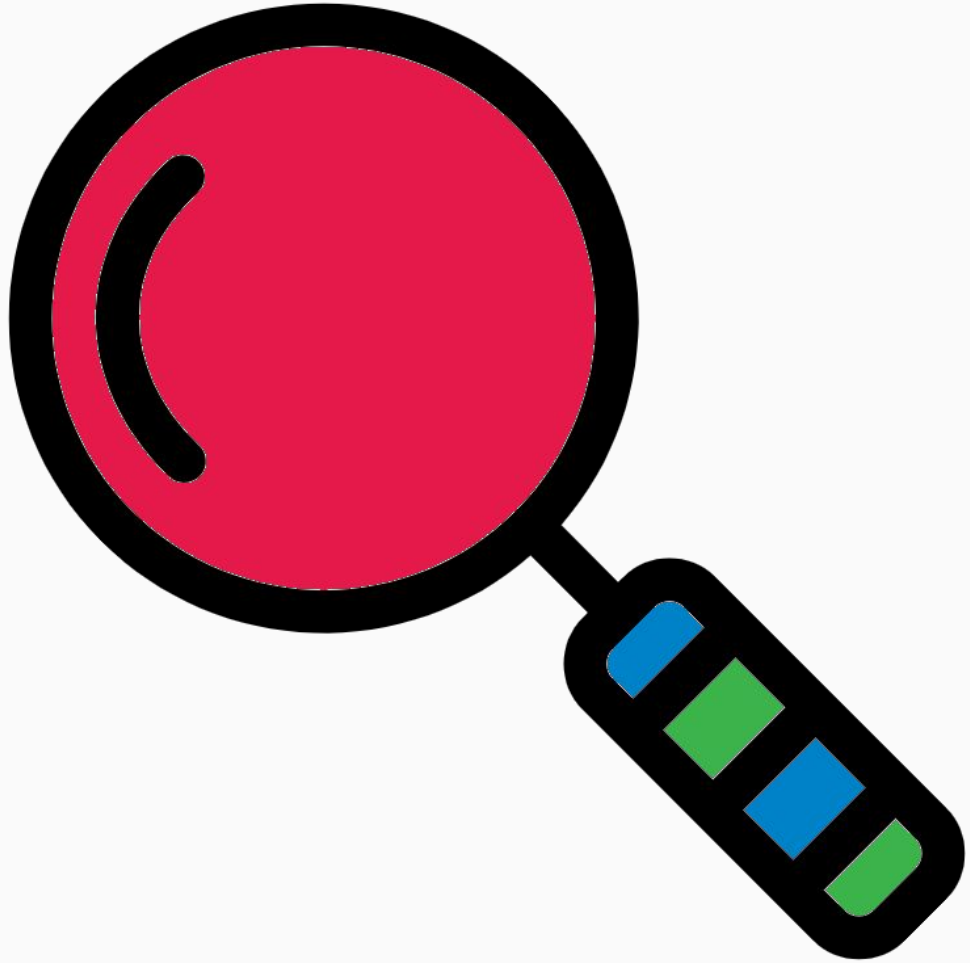
# Pre-assembly: summary

- **How much coverage do we have?**

30x recommended for PacBio HiFi, up to 100x for PacBio CLR or ONT

- **Sequence composition & quality**

k-mer frequency plots – low errors, expected genome size and level of heterozygosity for the sample/study species



# Assembly: Hifiasm

***de novo* assembler for PacBio HiFi reads**

- **fast**

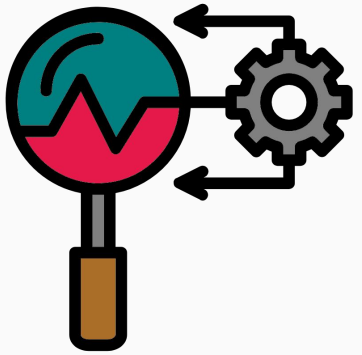
- **haplotype aware error correction –** corrects sequencing errors but retains heterozygous alleles

- **Outputs: phased assembly –** primary assembly and alternate contigs (haplotigs = contig from same haplotype)

- **If you don't have HiFi data?** Other long read assemblers are **miniasm**, **canu**, **wtdbg2**, **Falcon**, **Flye** – more coverage, slower assembly

# Genome assessment – basic stats

- **Contiguity assessment**
- #contigs



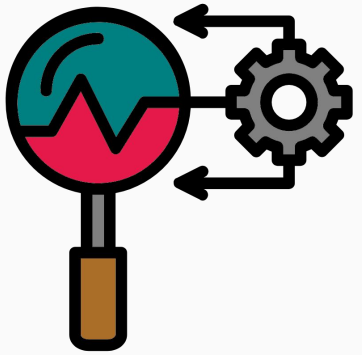
- Length/size – is it what we expected?


50% length = 200
N50 = 60

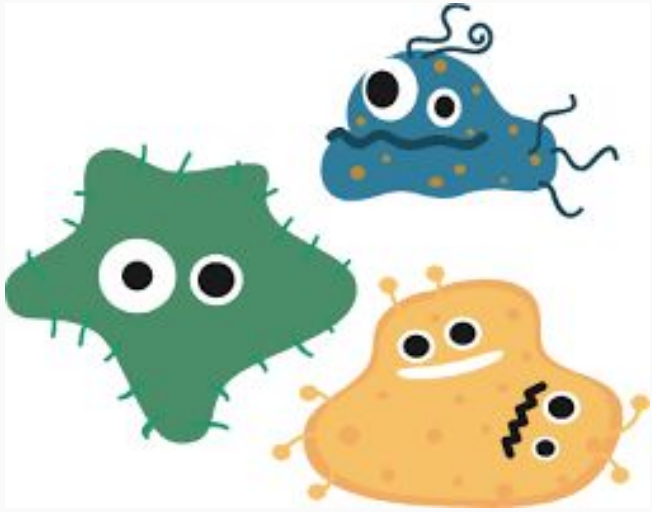
- N50 – above 1Mb is considered good for a long read assembly

# Genome assessment – contamination

## Check for contamination

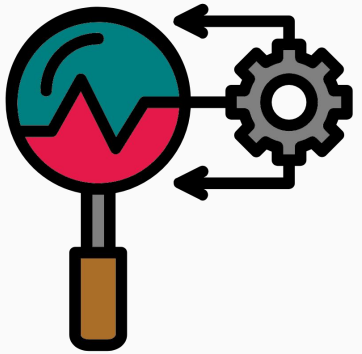


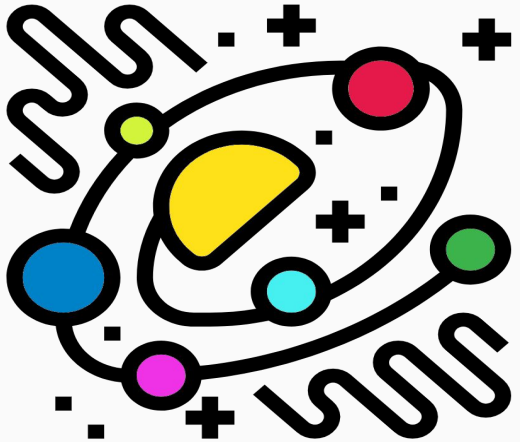
- What was the source of the DNA?
- May have contaminants / symbionts / parasites
- Check & remove from assembly
- **Tiara** - very fast tool, identifies archaea, bacteria, prokarya, eukarya, and organelle sequences.



# Genome assessment - BUSCO

- **Benchmarking Universal Single-Copy Orthologs**
- Benchmarking - Estimates completeness
- Universal - Present in all organisms within a lineage
    - Many lineage datasets to choose from
- Single-Copy - Should only be found once in the genomes
- Orthologs - Genes in different species that evolved from a common ancestral gene by speciation

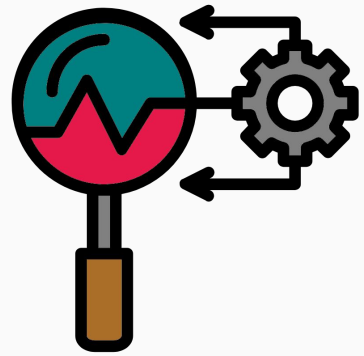




# Genome assessment – duplication

## Identifying alternate duplicated contigs

- Diploid genomes - allelic variations can mean alternate duplicated contigs are assembled.

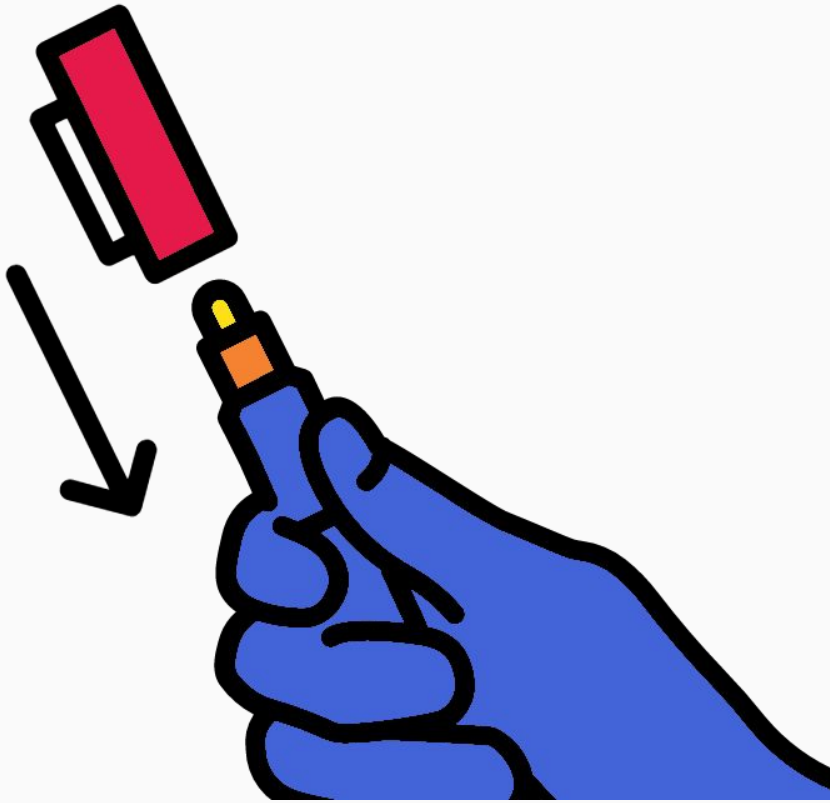


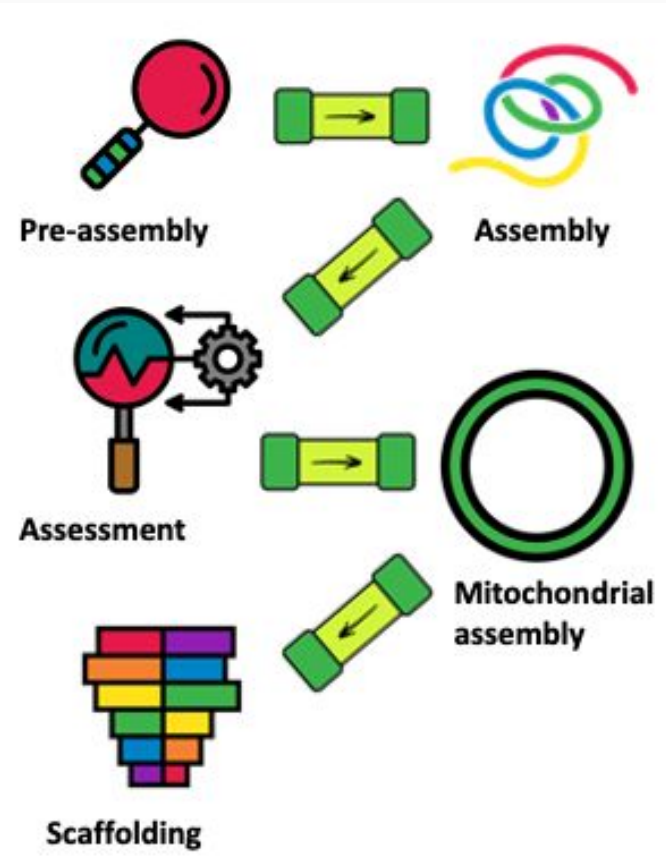
- Hifiasm – doesn't always fully identify these

- **purge_dups** - maps original reads to assembly, splits alternate haplotigs and genuine haploid genome representations.

# Recap

- Sequencing technologies
- Pre-assembly assessment
- Assembly
- Post-assembly assessment





# Reminders and Tips

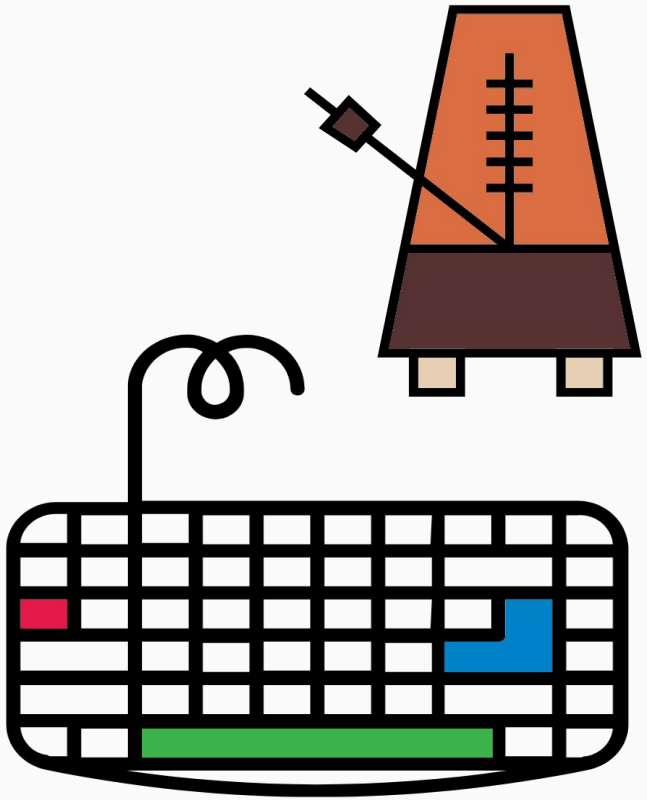
Work at your own pace

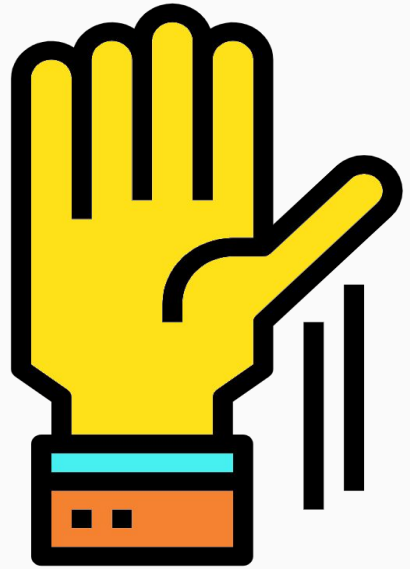
Typos

Ask questions

Breaks are important

Tab, space, and enter





# During sessions

Zoom - Ask via microphone if no question currently being asked/answered

Slack - Ask questions via the channel or ask to go into a zoom breakout room with one of us

WebVNC - We can connect to your webVNC to see and help with issues.

Breakout rooms upon request

# Thank you!

Questions?