



# LONG-READ SEQUENCING

Claude THERMES

PLATEFORME DE SÉQUENÇAGE I2BC

INSTITUT DE BIOLOGIE INTÉGRATIVE DE LA CELLULE

GIF-SUR-YVETTE

14<sup>ème</sup> ÉCOLE DE BIOINFORMATIQUE EBAIL - 17/11/2025

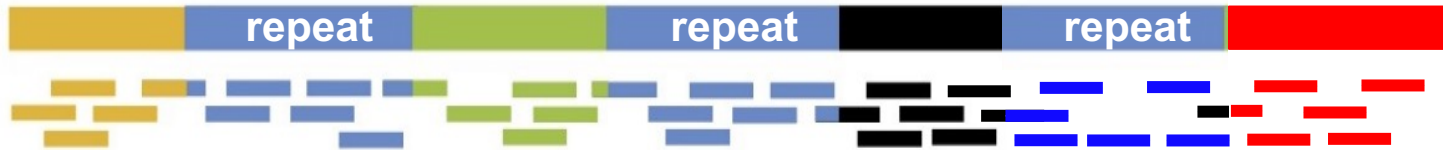
## LONG-READ SEQUENCING

Long-reads:  $\approx 1$  kb to  $\approx 100$  kb (ultra-long reads:  $> 100$  kb)

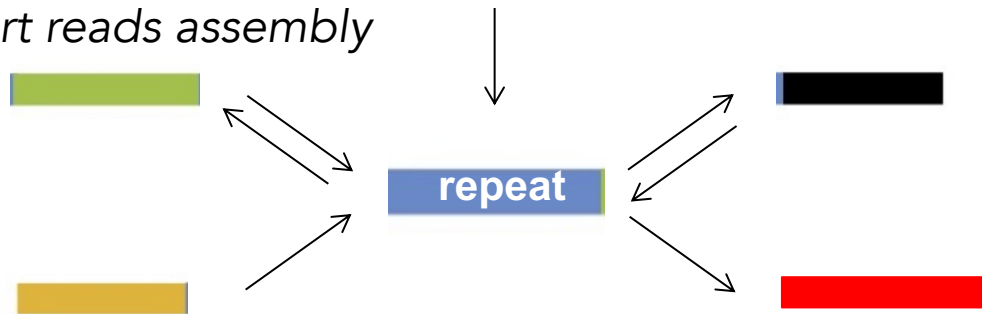
- Genome assembly
- Haplotype phasing
- Transcriptomics : splicing isoforms

# LONG-READS VERSUS SHORT-READS : GENOME ASSEMBLY

Assembly of DNA fragments with repeated sequences



*NGS short reads assembly*



Several contigs → incomplete assembly, underestimation of repeats

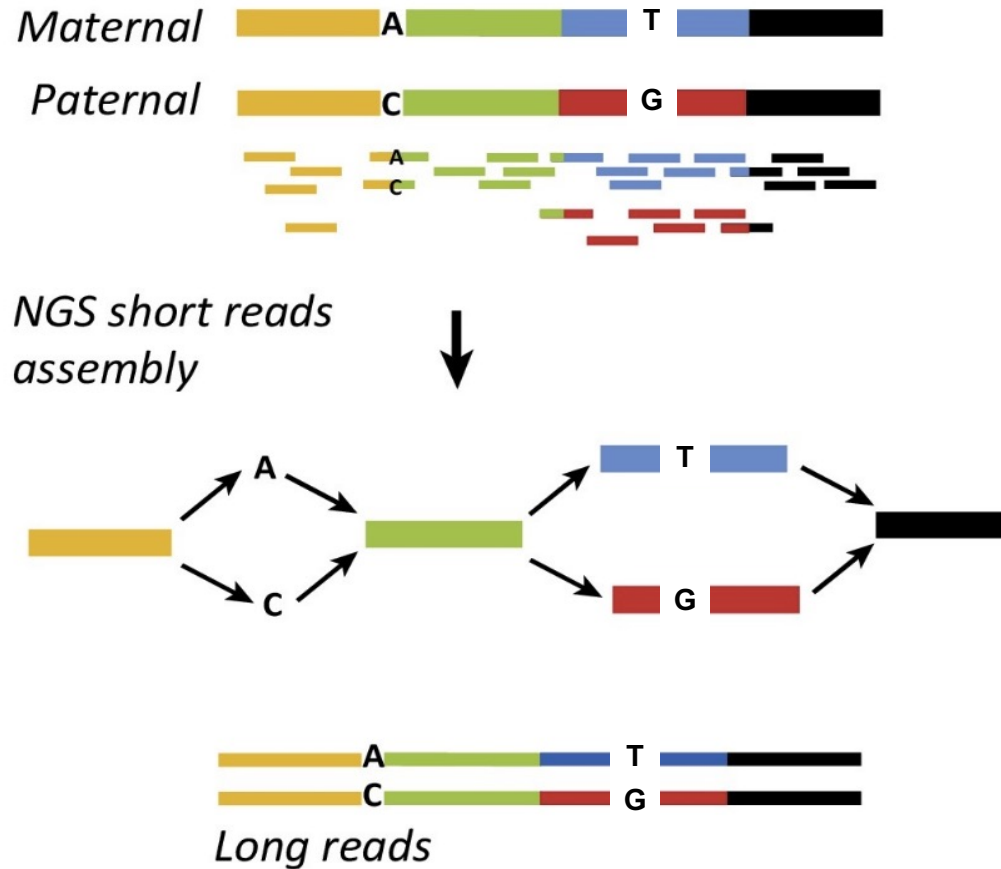
*Long reads assembly*



Long-reads (1- 200 kb) allow assembly of large repeat-rich regions  
(centromeres, telomeres...)

# LONG-READS VERSUS SHORT-READS : HAPLOTYPING

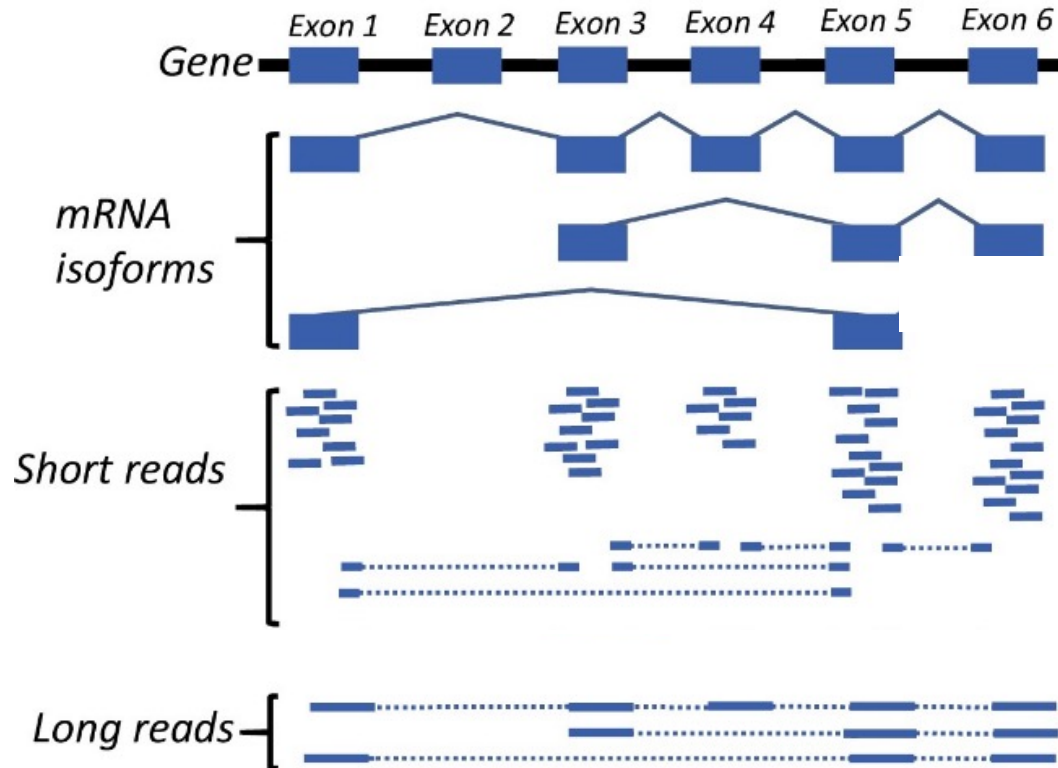
## Haplotype phasing



Long-reads allow phasing of maternal and paternal haplotypes

# LONG-READS VERSUS SHORT-READS : SPLICING ISOFORMS

## Detection of splicing isoforms



Long-reads allow identification of multiple splicing events  
along each mRNA molecule

# 3rd generation winning technologies

## Pacific Biosciences



Vega

Revio

Single molecules  
Up to 200 kbp long

## Oxford Nanopore

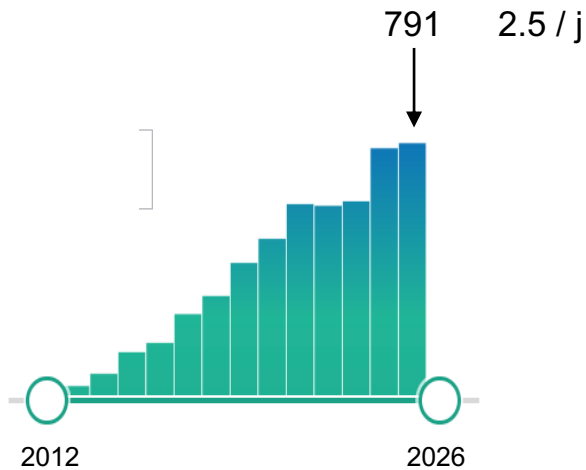


MinION – GridION – PromethION

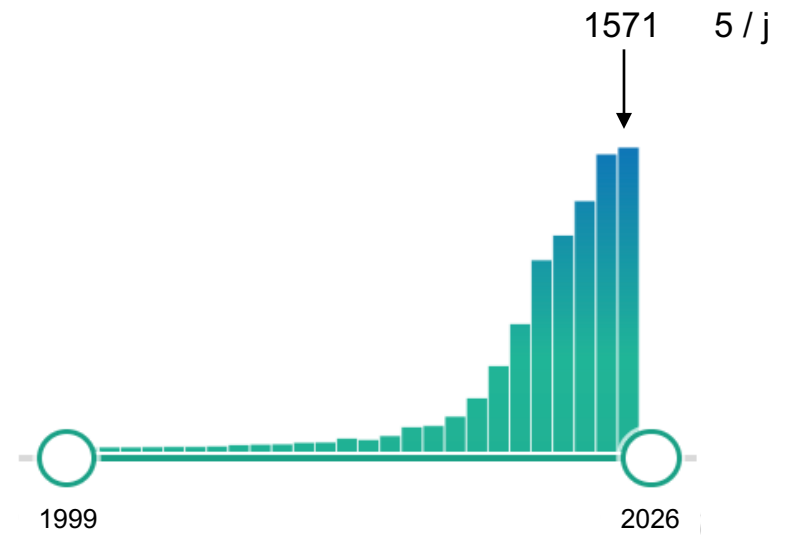
Single molecules  
Up to 1 Mbp long

# 3rd generation technologies

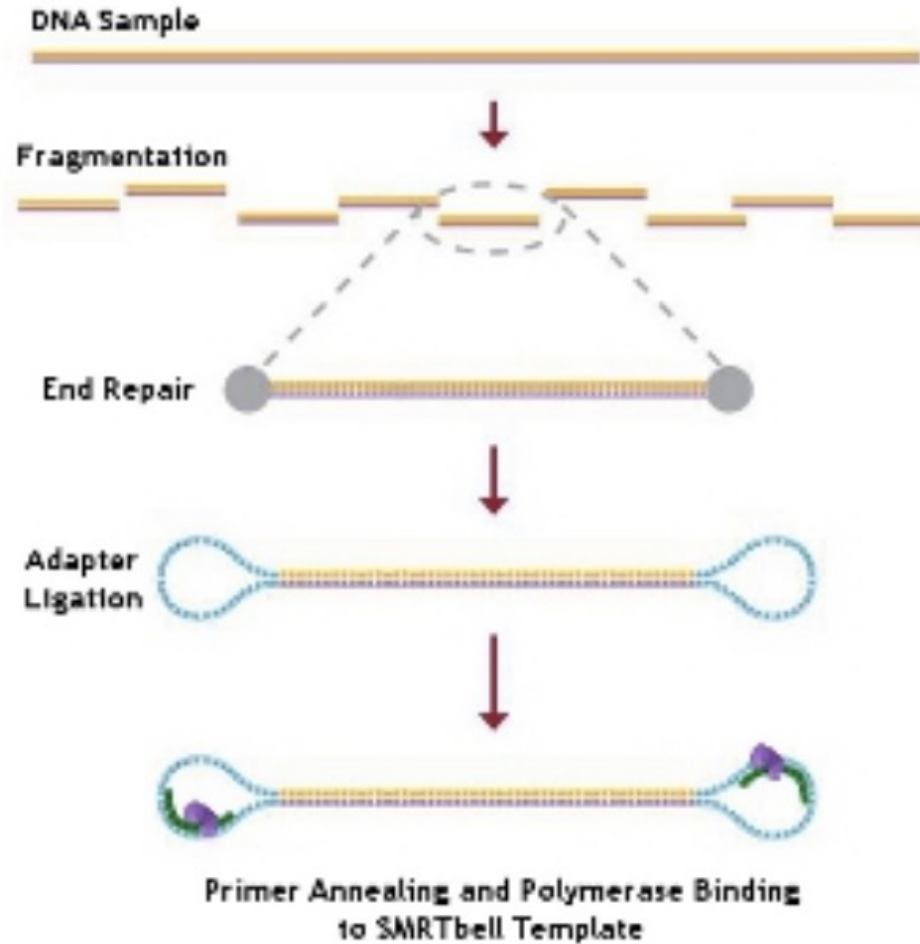
## Pacific Biosciences



## Oxford Nanopore



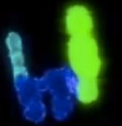
## PacBio DNA-seq library



# PACIFIC BIOSCIENCES

## Phospholinked Nucleotides

A



C



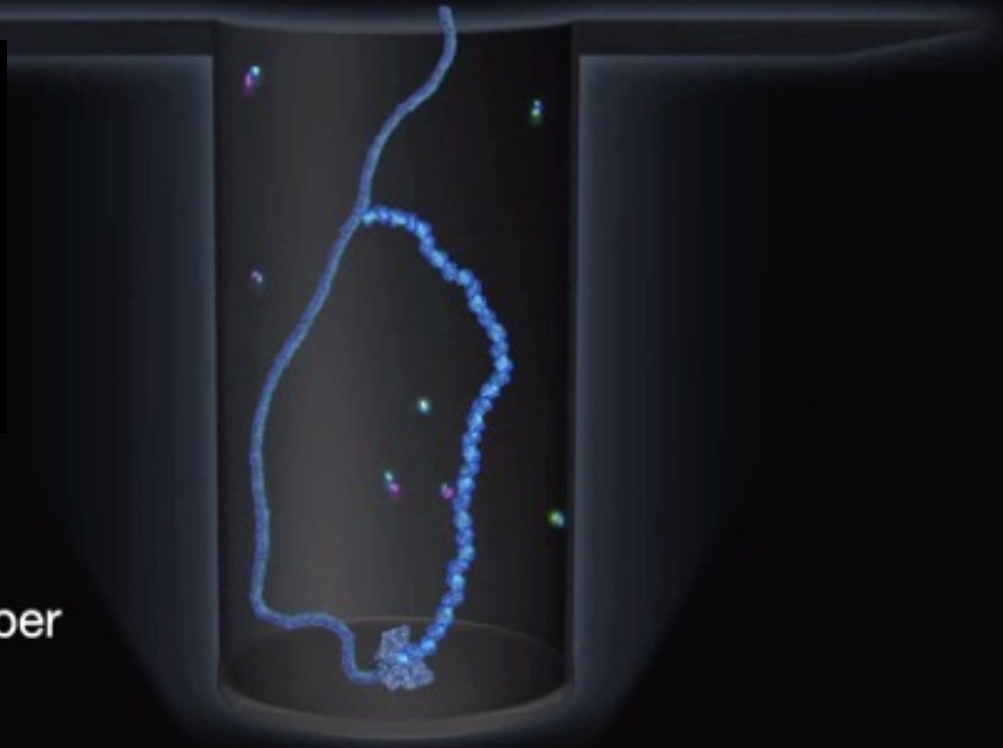
G



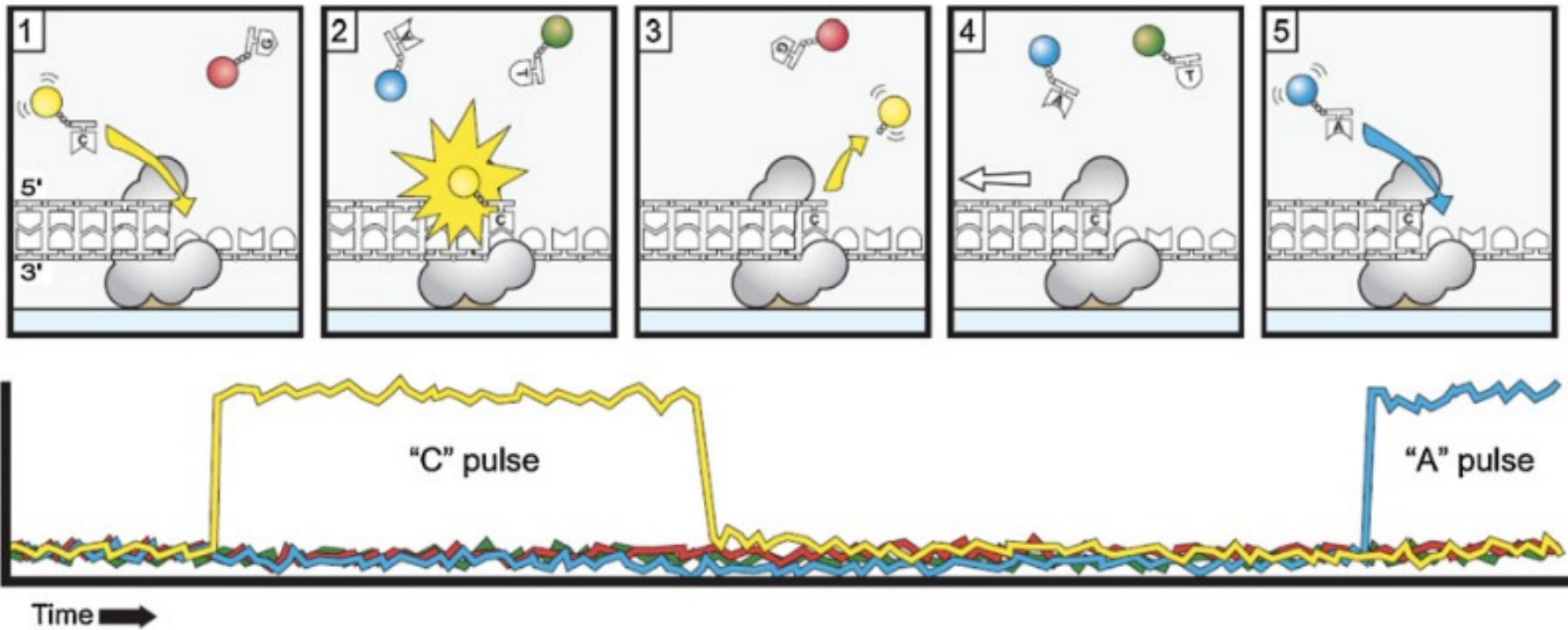
T



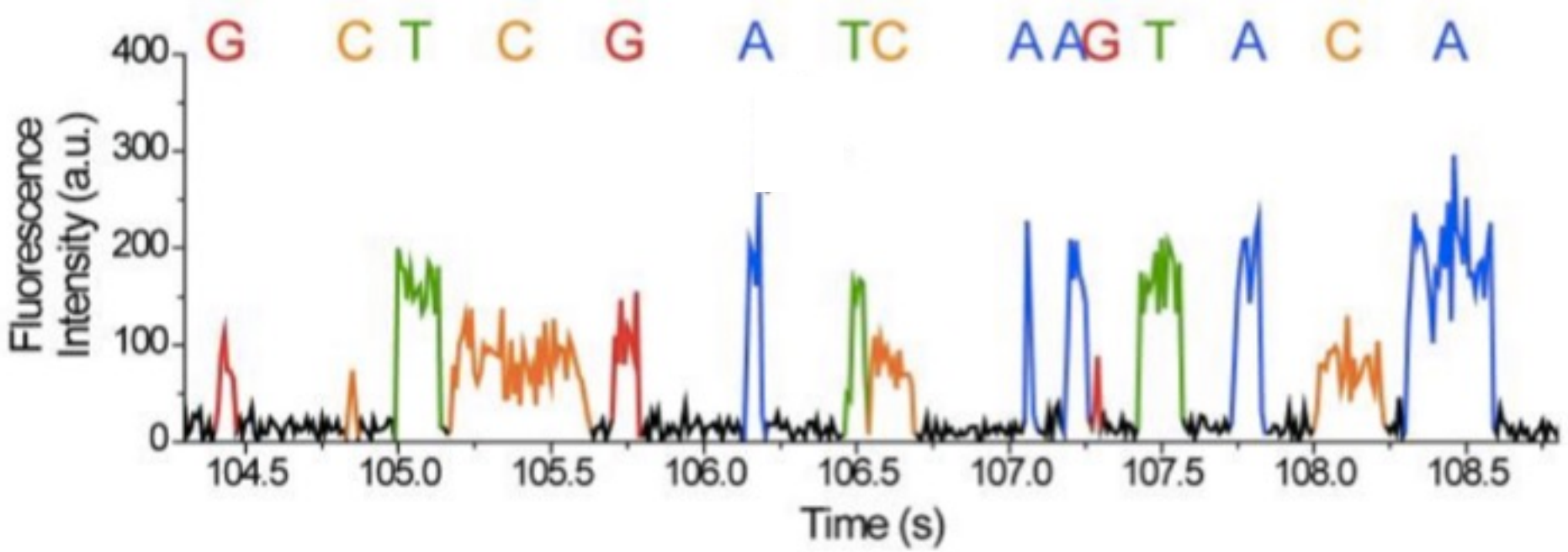
Phospholinked nucleotides are introduced into the ZMW chamber



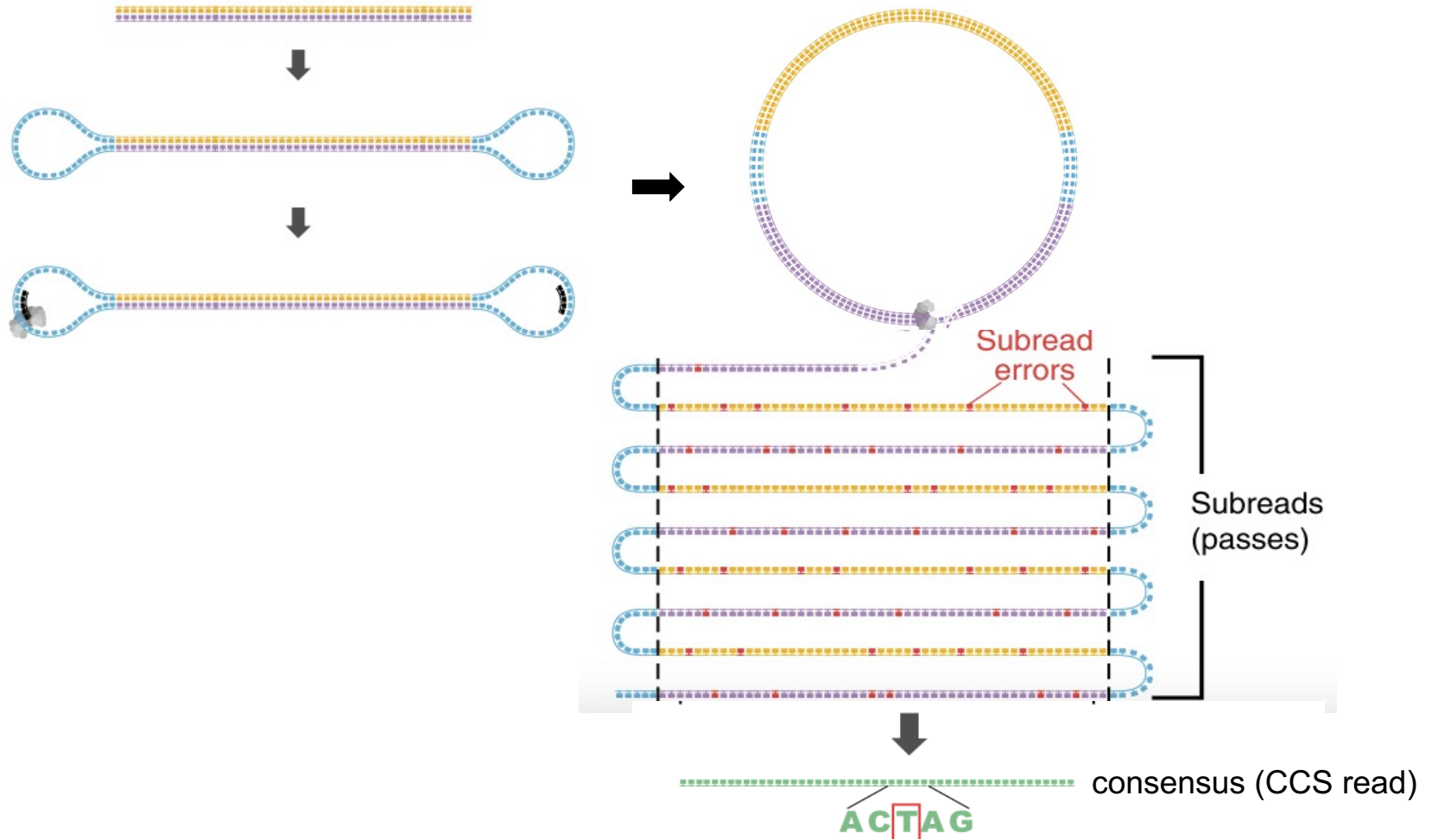
# PACIFIC BIOSCIENCES



# PACIFIC BIOSCIENCES

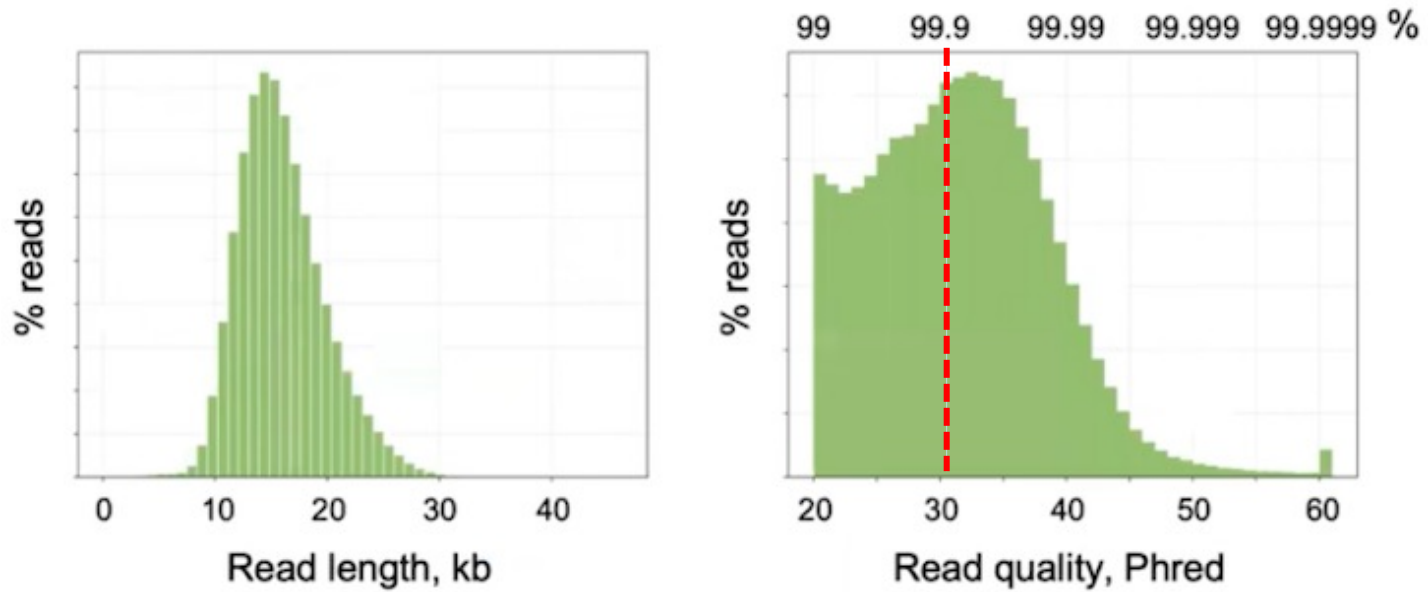


# CIRCULAR CONSENSUS SEQUENCES (CCS) : HIFI READS

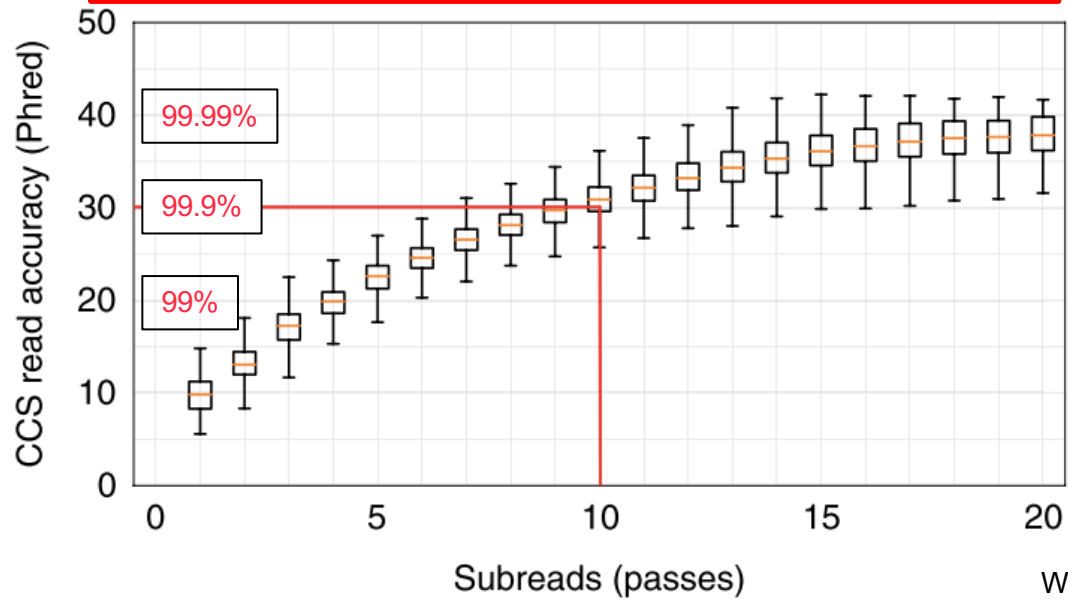


Randomly positioned errors → they can be corrected

# CIRCULAR CONSENSUS SEQUENCES (CCS) : HIFI READS



Mean accuracy > 99.9 % ↔ >Q30 (< 1 err/kb)



## HOW HI-FI SEQUENCING COMPARES

	HiFi sequencing	SBS sequencing	Nanopore sequencing
Read length	1–25 kb	2x150 bp	10–100 kb
Read accuracy	99.95% (Q33)	99.92% (Q31)	99.26% (Q21)
Run time	0.5 error/kb	0.8 error/kb	
Yield	90 Gb <sup>2,5</sup>	2,400–3,000 Gb	50–110 Gb
Variant calling – SNVs	✓	✓	✓
Variant calling – indels	✓	✓	✗
Variant calling – SVs	✓	✗	✓
5mC methylation	✓	✗	✓
Phasing	✓	✗	✓

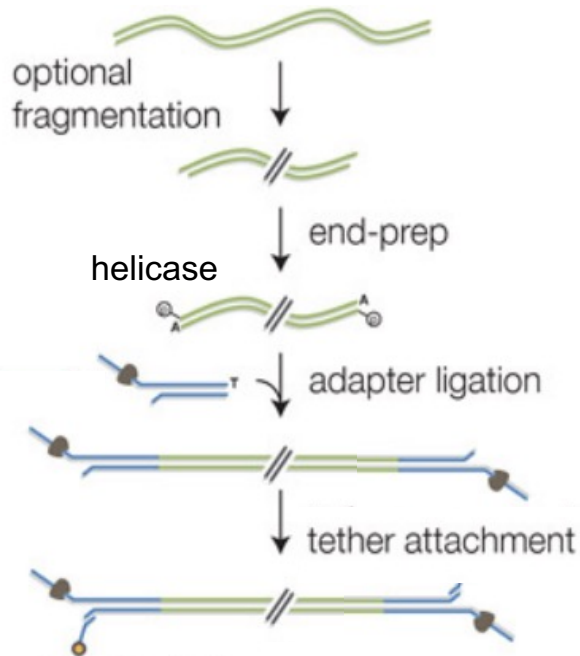
- <sup>2</sup>HiFi yield specification based on HG002/GM24385 human DNA extracted with Nanobind CBB kit and prepared with SMRTbell prep kit 3.0.
- <sup>3</sup>Run time specification is for the sequencing reaction.
- <sup>5</sup>HiFi yield is dependent on library fragment size. Yield is typically lower for shorter libraries.

# Next Generation Sequencing

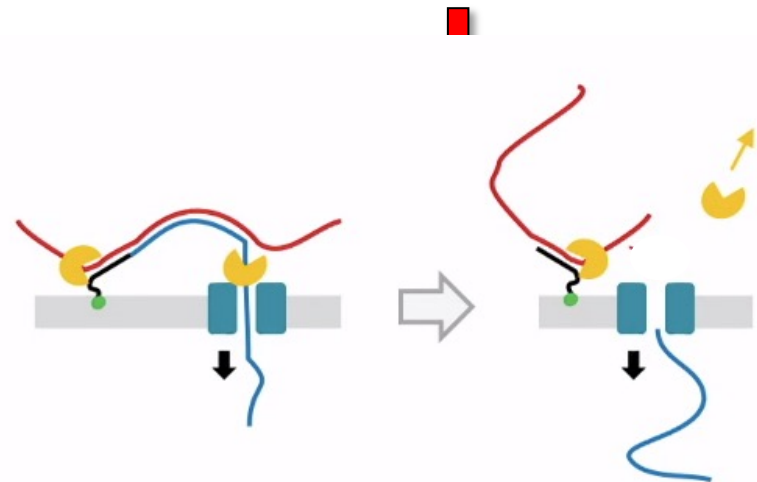


# SEQUENCING PROCESS

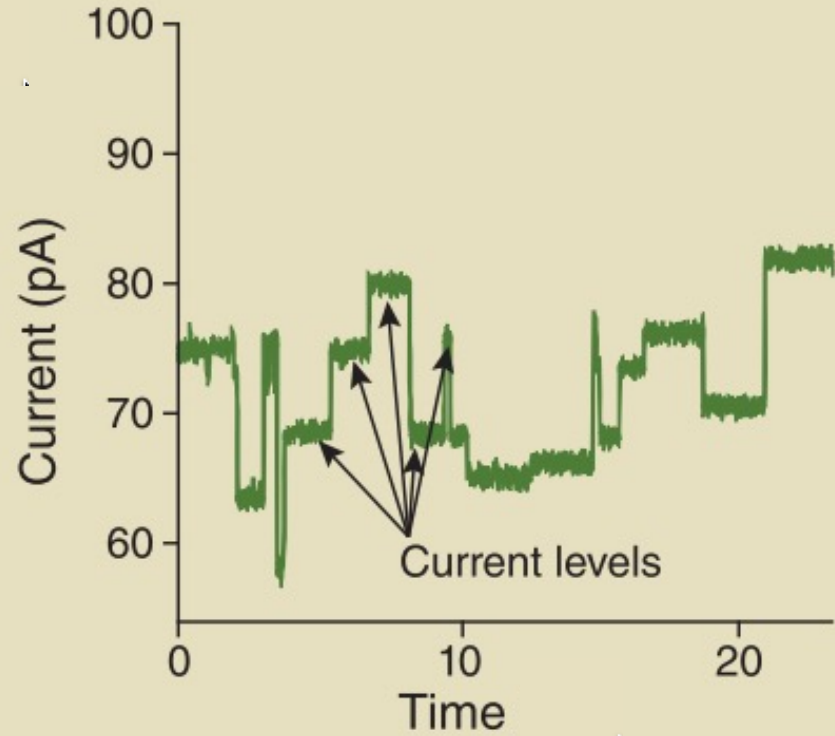
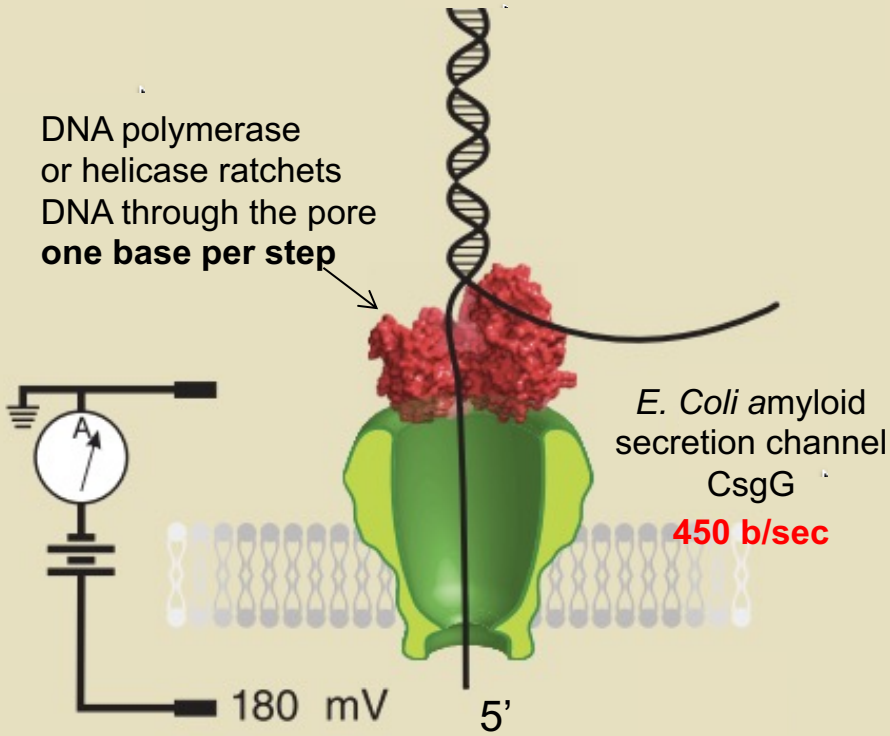
## Library preparation



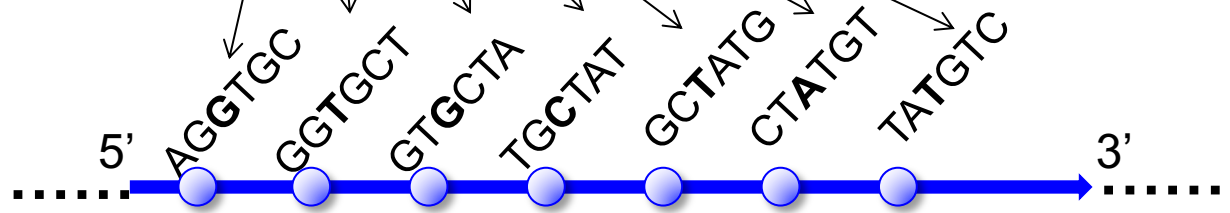
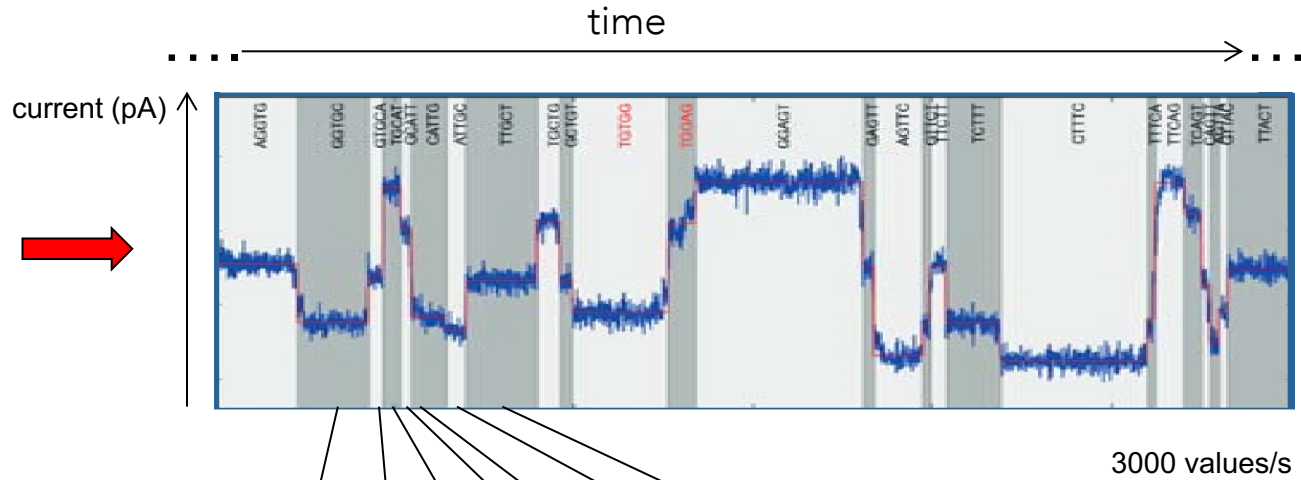
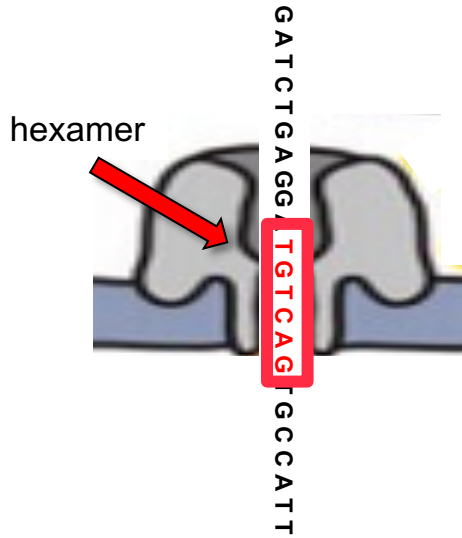
## SEQUENCING



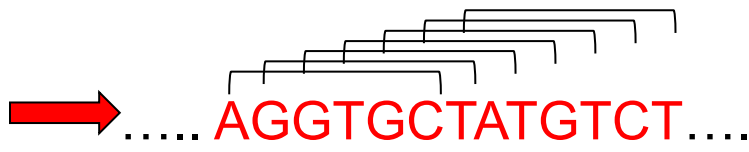
# BASIC CONCEPTS



# BASE CALLING

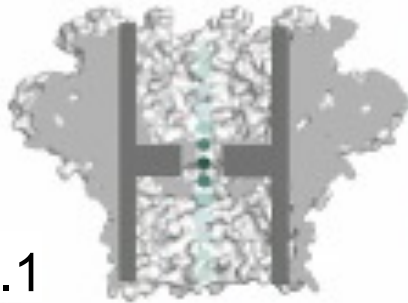


Basecalling : finding the optimal path of successive 6-mers



# "TWO READERS" NANOPORE

R9.4.1

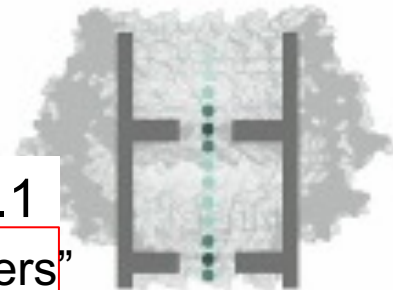


ATCGGAAAAAAAAAATCACGCCACGTCCAAA



R10.4.1

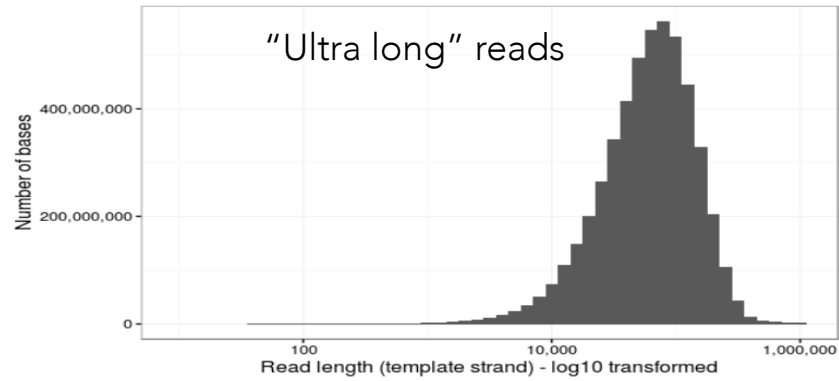
"two-readers"



Sereika et al. *Nature Methods*, 2022

Long homopolymers are better "seen" by the pore and can be decoded with higher accuracy

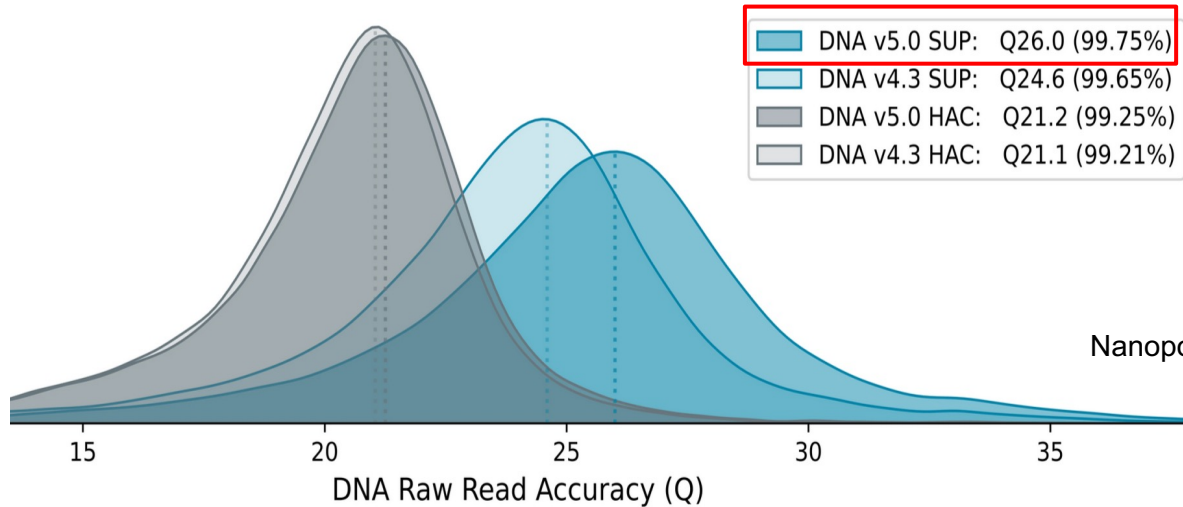
# LENGTH OF NANOPORE READS



(lab.loman.net, March 2017)

Size of longest reads > 1 Mb

# READ ACCURACY R10.4.1



Nanopore website 2024

GENOME ASSEMBLY  
SMALL GENOMES

# NANOPORE : ASSEMBLY OF BACTERIAL GENOMES

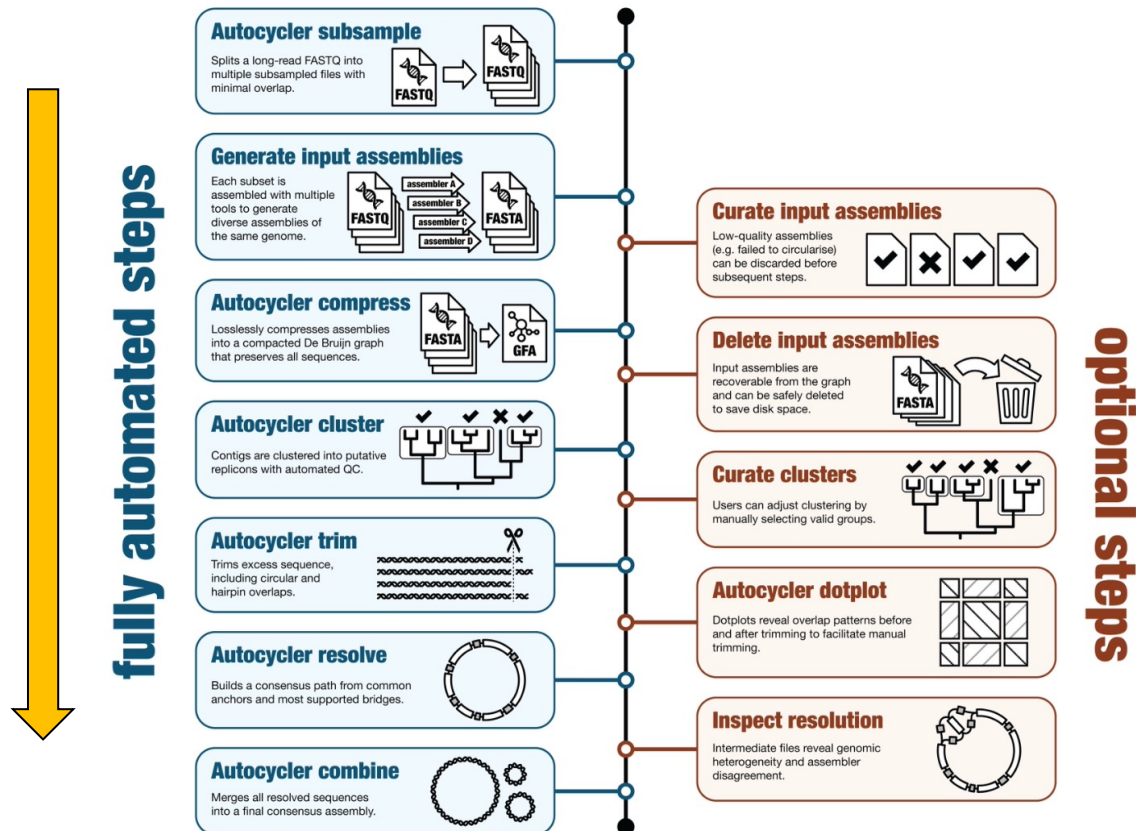
Autocycler: long-read consensus assembly for bacterial genomes  
Wick et al. *Bioinformatics* Aug 2025

Context :

- Individual assemblers are imperfect and often produce sequence-level and structural errors
- different tools produce different assemblies from the same input read set

Objective : Generate an accurate genome assembly by combining multiple alternative long-read assemblies of the same genome

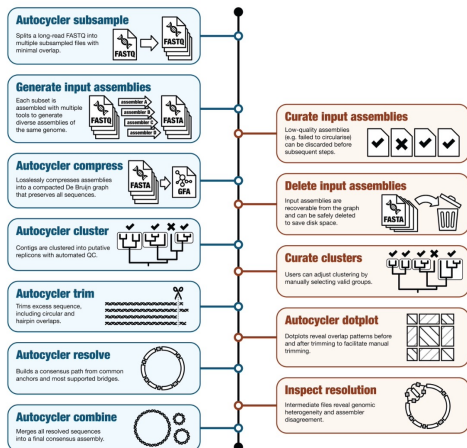
From the input assemblies, **Autocycler** resolves consensus sequences by selecting the most common variant at each locus



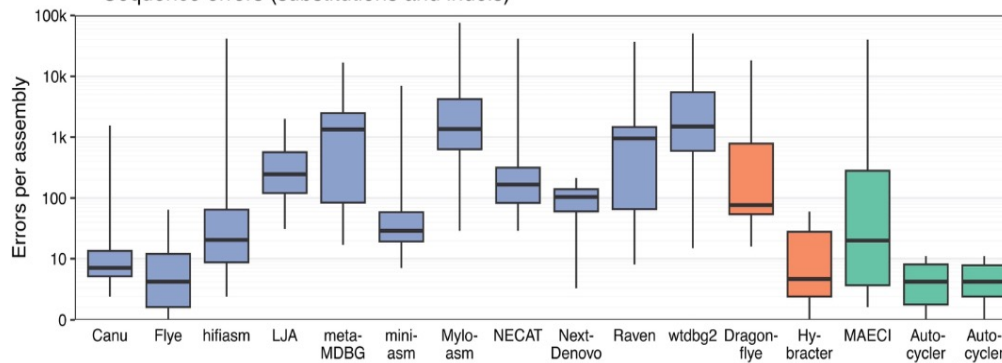
# NANOPORE : ASSEMBLY OF BACTERIAL GENOMES

Autocycler: long-read consensus assembly for bacterial genomes  
 Wick et al. *Bioinformatics* Aug 2025

fully automated steps

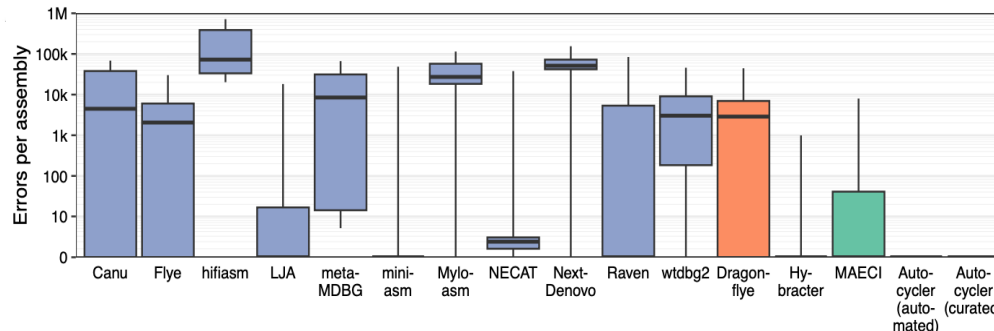


Sequence errors (substitutions and indels)



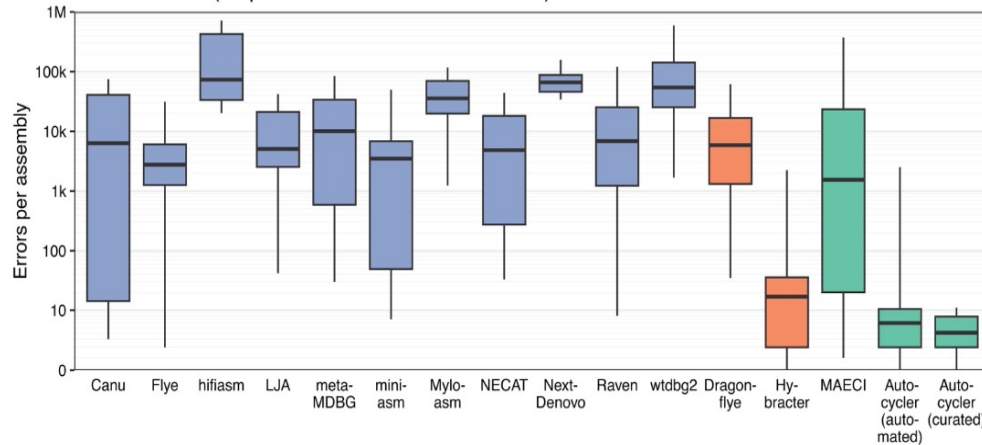
Small-scale errors

Extra bases



Structural errors

Total errors (sequence errors + structural errors)

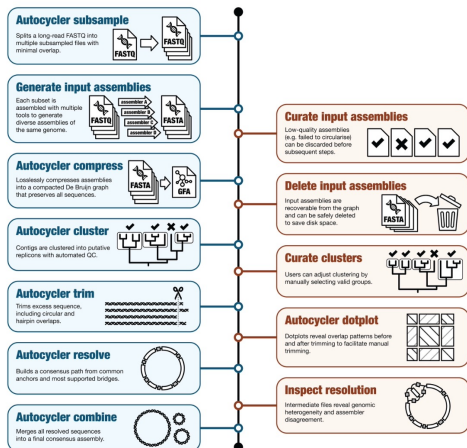


Total errors

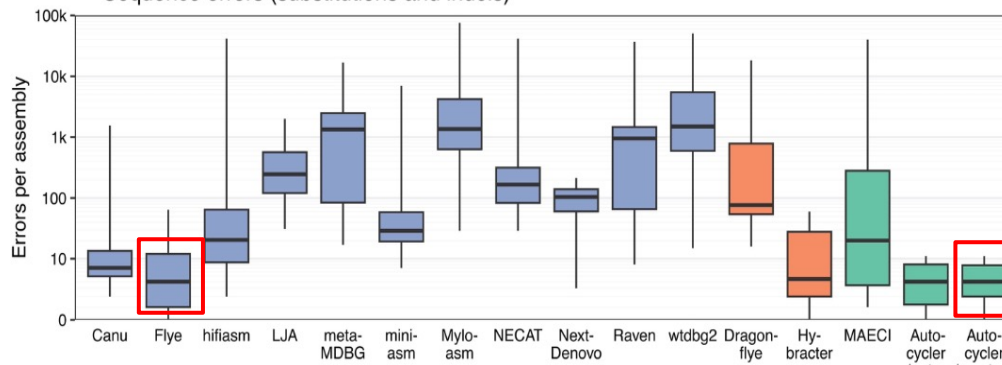
# NANOPORE : ASSEMBLY OF BACTERIAL GENOMES

Autocycler: long-read consensus assembly for bacterial genomes  
 Wick et al. *Bioinformatics* Aug 2025

fully automated steps

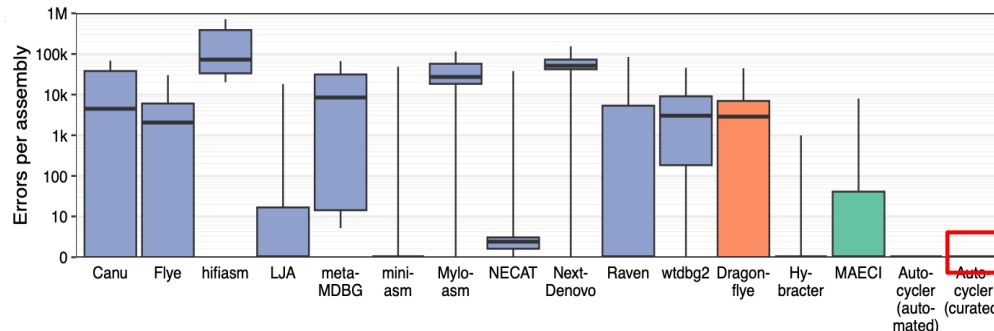


Sequence errors (substitutions and indels)



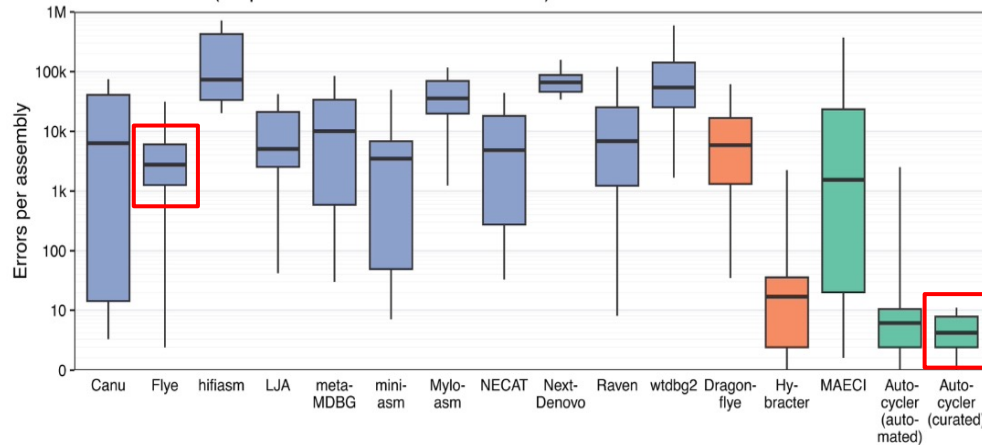
Small-scale errors

Extra bases



Structural errors

Total errors (sequence errors + structural errors)



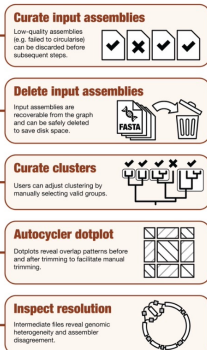
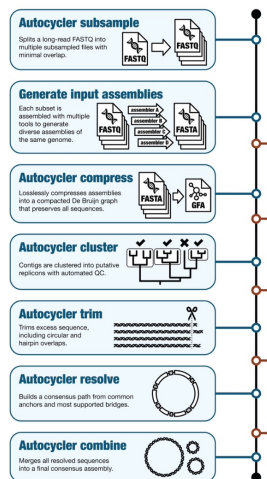
Total errors

# NANOPORE : ASSEMBLY OF BACTERIAL GENOMES

Autocycler: long-read consensus assembly for bacterial genomes

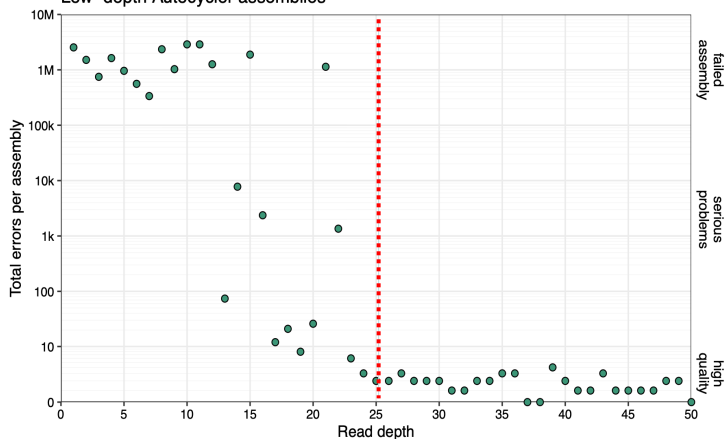
Wick et al. *Bioinformatics* Aug 2025

fully automated steps

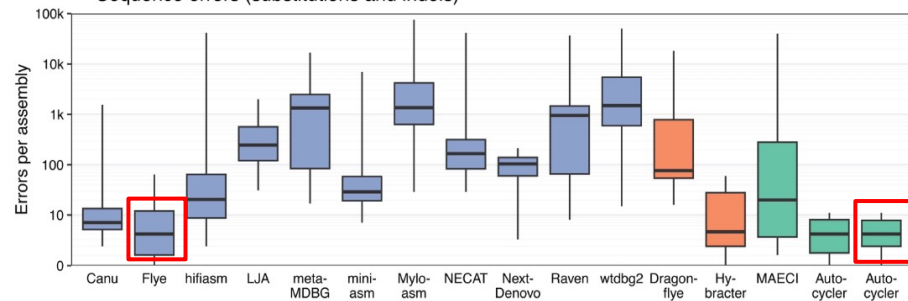


optional steps

Low-depth Autocycler assemblies

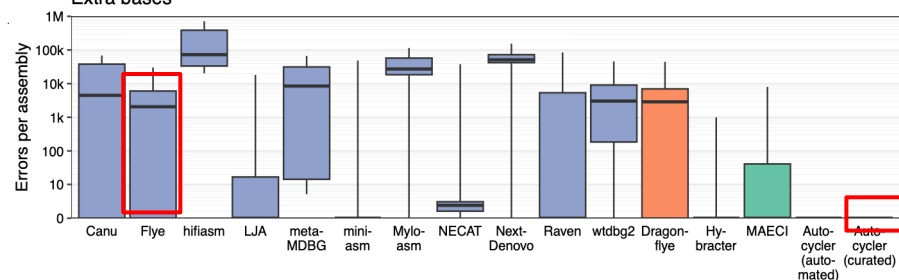


Sequence errors (substitutions and indels)



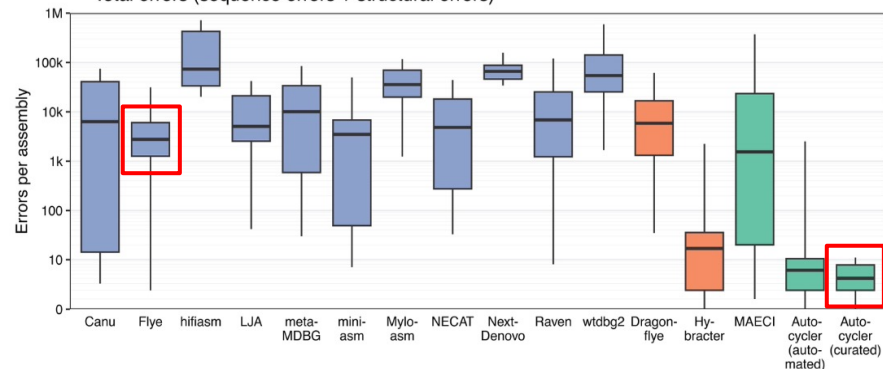
Small-scale errors

Extra bases



Structural errors

Total errors (sequence errors + structural errors)

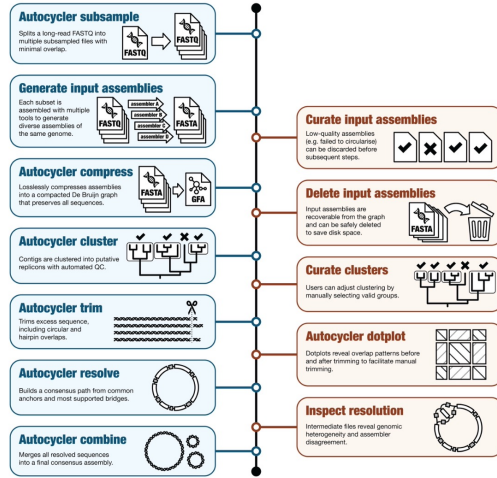


Total errors

# NANOPORE : ASSEMBLY OF BACTERIAL GENOMES

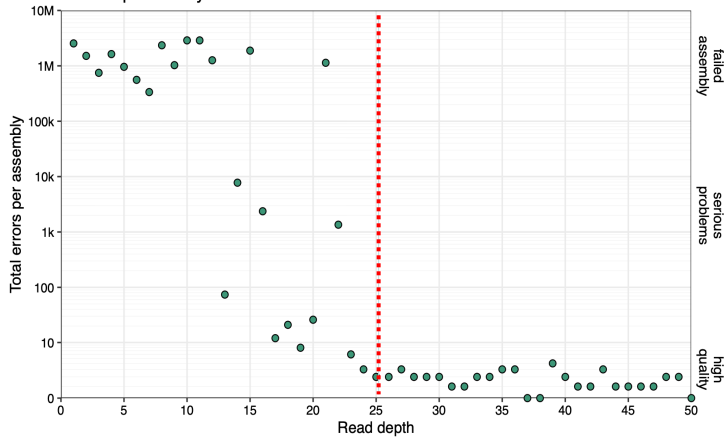
Autocycler: long-read consensus assembly for bacterial genomes  
 Wick et al. *Bioinformatics* Aug 2025

fully automated steps

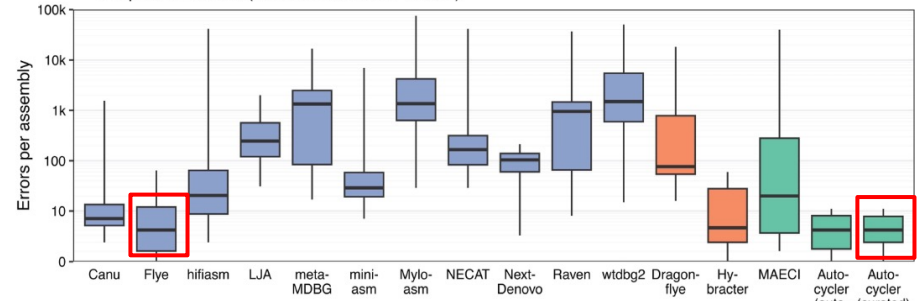


optional steps

Low-depth Autocycler assemblies

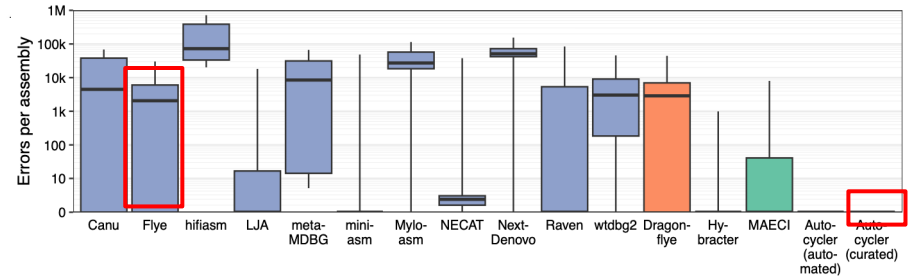


Sequence errors (substitutions and indels)



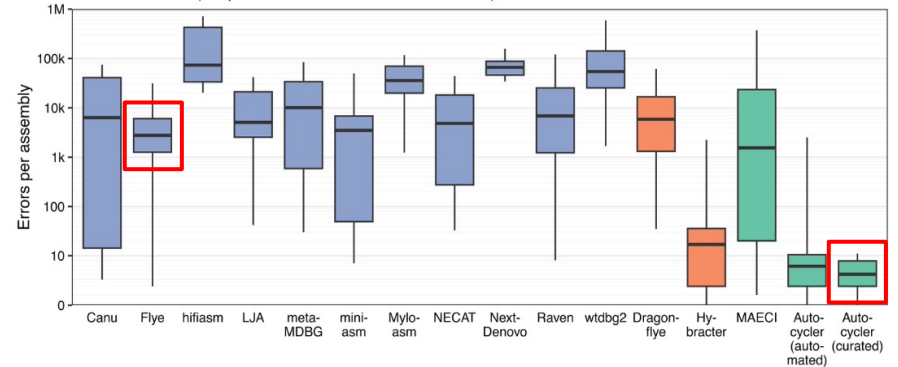
Small-scale errors

Extra bases



Structural errors

Total errors (sequence errors + structural errors)



Total errors

- Flye produces the fewest sequence-level errors
- Autocycler resolves structural errors

# NANOPORE : SURVEILLANCE OF BACTERIAL PATHOGENS

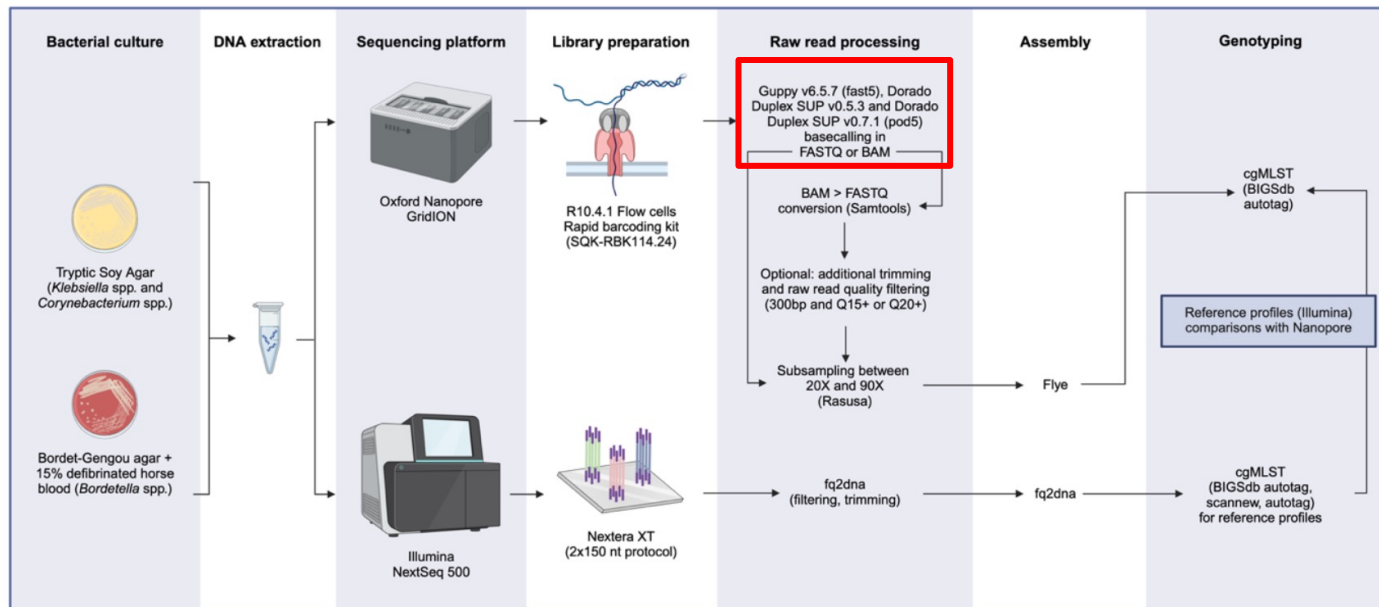
Accurate genotyping of three major respiratory bacterial pathogens with ONT R10.4.1 long-read sequencing  
Zidane et al. *Genome Res.* Aug 2025

## Context :

- Bacterial strain taxonomy so far do not accept genomes generated with ONT
- Previous ONT chemistries (e.g., wR9.4.1) -> higher error rates compared to the gold standard Illumina

## Objective : evaluate ONT R10.4.1 for genomic typing against Illumina

- ONT R10.4.1 with Rapid Barcoding Kit V14, basecall with Dorado SUP v0.9.0
- assembled with Flye (v2.9)
- Applied to 3 respiratory pathogens : *K. pneumoniae*, *C. diphtheriae*, *B. pertussis*



## NANOPORE : SURVEILLANCE OF BACTERIAL PATHOGENS

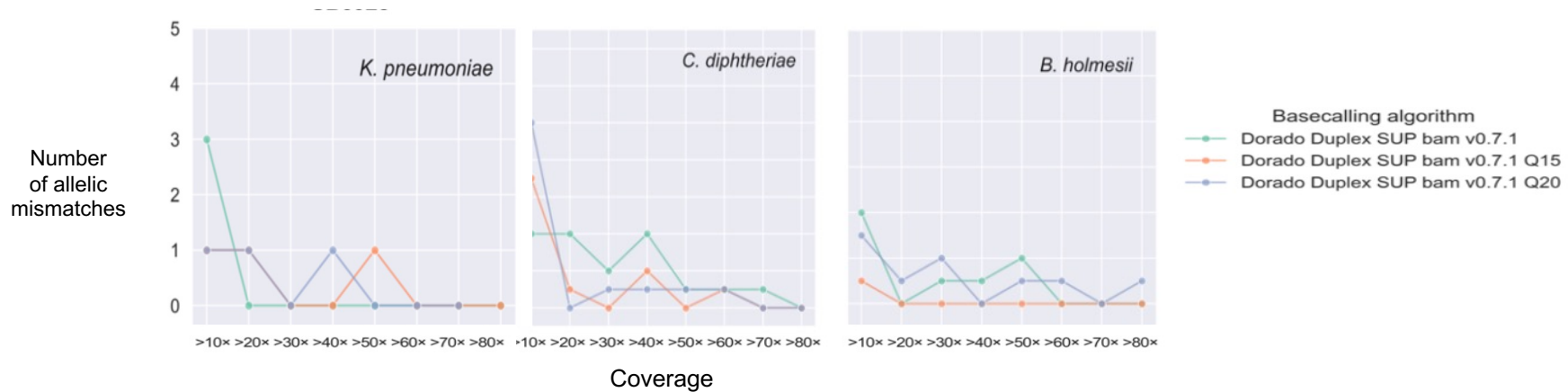
Accurate genotyping of three major respiratory bacterial pathogens with ONT R10.4.1 long-read sequencing  
Zidane et al. *Genome Res.* Aug 2025

Genomes obtained from ONT R10.4.1 :

- Basecall with Dorado SUP v0.9.0
- assembled with Flye
- minimum coverage 35 ×
- Error rates <0.5% for each cgMLST scheme



Number of allelic mismatches between assemblies generated with Illumina and with ONT



**ONT R10.4.1 suitable for genomic typing applied to outbreak and public health surveillance**

Genome assembly :  
Large genomes : T2T era

## Strategy for near-telomere-to-telomere assembly

**Table 1 | Common data types for high-quality assembly**

Data type	Technologies	Description	Roles
Accurate long reads	PacBio HiFi, ONT duplex	>10 kb in length; error rate <0.5%	Initial assembly graph construction; phasing over heterozygous variants that are less than 10 kb apart
Ultra-long reads	ONT ultra-long	>100 kb in length; error rate <10%	Resolving tangles; phasing through homozygous regions over 100 kb in length
Trio data	Short-read	Standard whole-genome shotgun sequencing of parents	Whole-genome phasing
Long-range data	Hi-C, Pore-C, Strand-seq	Information over 1 kb to over 10 Mb in length	Chromosomal phasing; chromosome-scale scaffolding

# 1. HISTORICAL PAPERS

## PacBio+Nanopore+Illumina

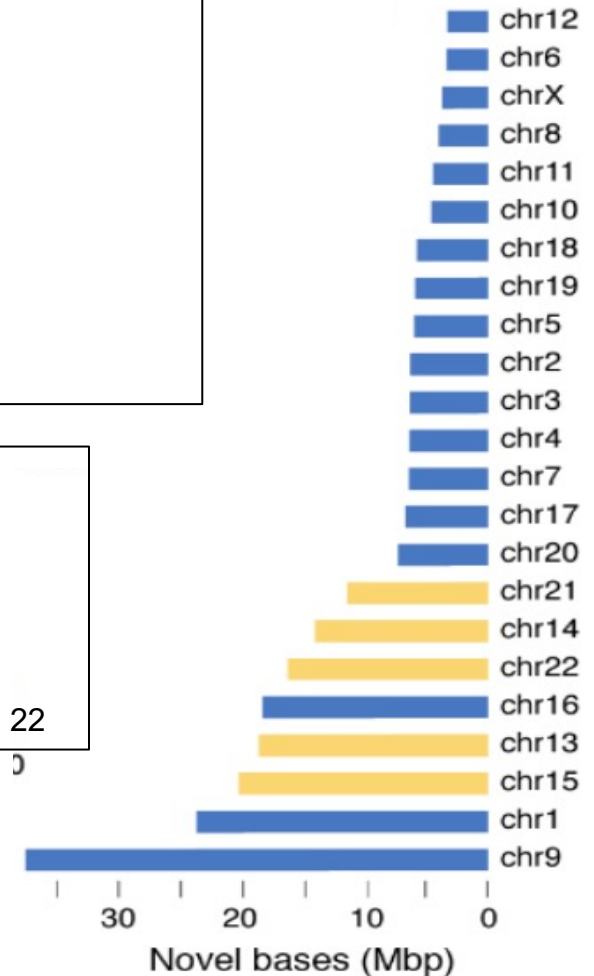
The complete sequence of a human genome  
Nurk et al. *Science* 2022

### SEQUENCING

Data were obtained with a “complete hydatidiform mole” (CHM13) cell line (haploid) :

- 30× PacBio circular consensus sequencing (HiFi)
- 120× Oxford Nanopore ultra-long read sequencing (ONT)
- 100× Illumina PCR-Free sequencing
- 70× Illumina / Arima Genomics Hi-C (Hi-C)
- BioNano optical maps (11)
- Strand-seq

- T2T assembly :including all 22 autosomes plus Chromosome X :
  - Introduces **200 million bp** of novel sequence (8% of the genome)
  - Identifies 2,226 paralogous gene copies, 115 predicted as protein coding
  - all centromeric regions
  - entire short arms (p-arms) of 5 acrocentric chromosomes : 13, 14, 15, 21, 22



## PacBio+Nanopore+Illumina

The complete sequence of a human Y chromosome  
Rhie et al. *Nature* 2023 (88 authors)

HG002 diploid genome

- Y chromosome -> last chromosome completed from telomere to telomere
- PacBio HiFi reads (60 × haploid genome coverage)
- ONT ultralong reads (90 × in reads > 100 kb)
- Strand-seq
- **combined T2T-Y with CHM13 to produce a complete reference sequence for all 24 human chromosomes**

