Long-Read Sequencing

HGM2026
April 22, 2026

Professor Qasim Ayub
Director
Monash University Malaysia Genomics Platform
Deputy Head of School (Research)
School of Science

qasim.ayub@monash.edu

MONASH University | MALAYSIA

Genomics Platform

bsi ISO 9001 Quality Management Systems CERTIFIED

Outline

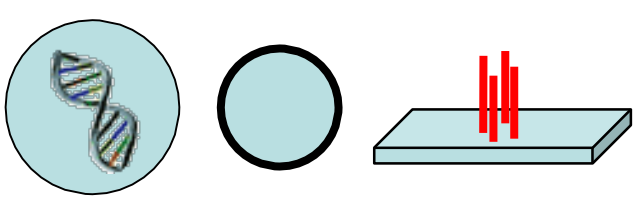
- Massively parallel high-throughput sequencing technologies.
- Short- and Long-read sequencing technologies.
- PacBio sequencing.
- Oxford Nanopore Sequencing Technologies.

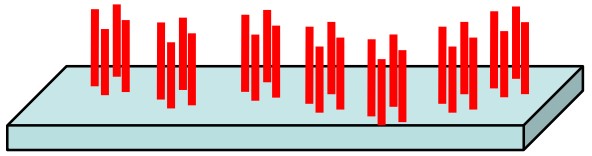
Learning Outcomes

- Understand differences between long and short-read technologies.
- Differentiate between PacBio and Oxford Nanopore Technologies.
- Applications of long-read sequencing in human genetics.

Massively Parallel Sequencing

- From 2004 – present.
- Also known as next generation sequencing (NGS).
- Utilize technologies that sequence multiple DNA fragments obtained from an individual genome in parallel – i.e. conduct massively parallel high-throughput sequencing.
- Not limited to few reactions per run.





1 feature.
1 template.

1 chip, thousands or millions of features.
Output Mb (one million) –Tb (one trillion)

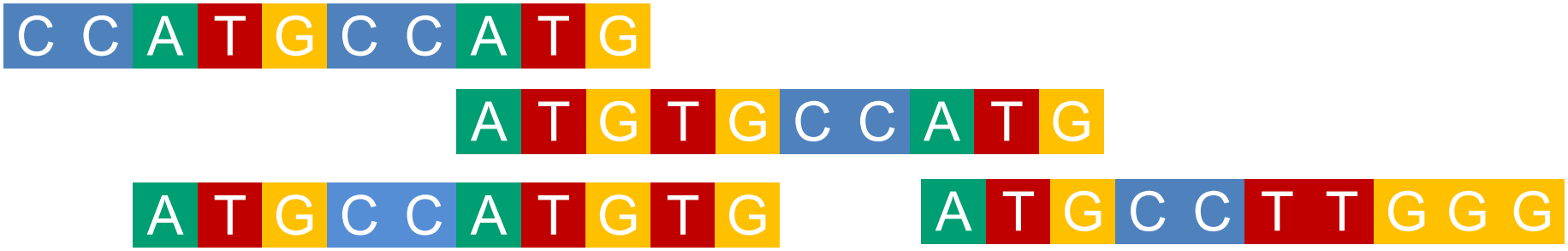
Massively Parallel Sequencing

Short Read Technologies
Illumina
MGI
ThermoFisher Scientific
GeneMind
Element Biosciences
Ultima Genomics
Salus BioMed
PacBio Onso System
Long Read Technologies
PacBio
Oxford Nanopore Technologies
MGI CycloneSeq™

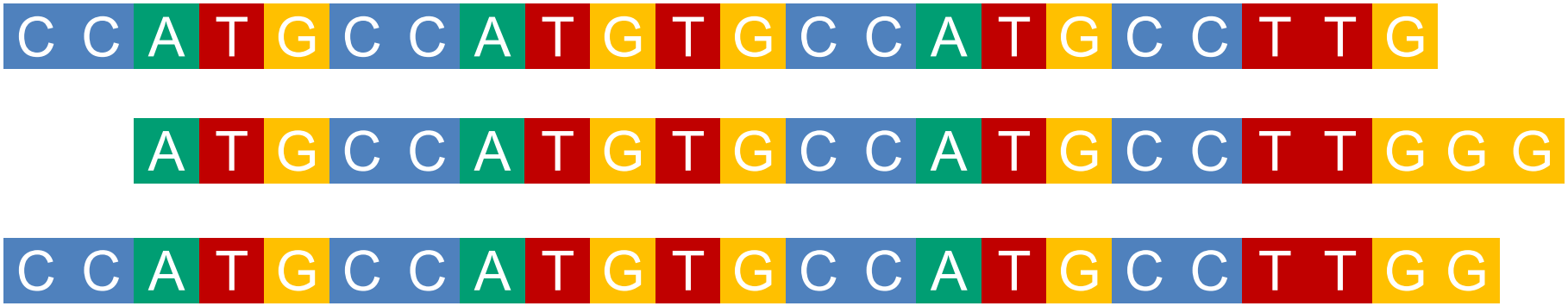
- Instrumentation and flow-cells.
- Library preparation step.
- Sequencing chemistry.
- Detection systems.
- Throughput.
- Sequence quality.

Short vs Long Read Sequencing

Short-read 36 – 1,000 bp (1 kb)



Long-read > 10 kb > 4 Mb (10^6 bases)



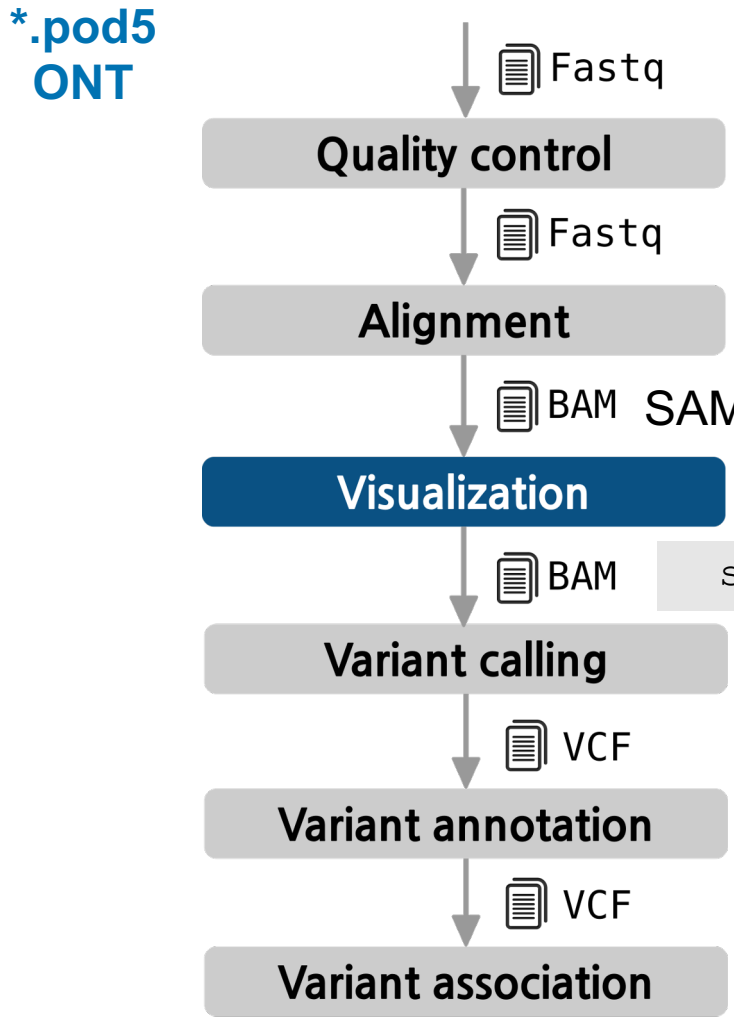
Sequencing Steps

- Extract and/or fragment DNA.

  Shearing.
  Nebulization.
  Sonication.
  Enzymatic digestion.
  Transposon mediated fragmentation.

- Prepare DNA fragment library.
- Sequence fragments 36 – > 4 Mb (10^6 bases)
- Assemble fragments:
  - Map fragments to reference sequence.
  - *De-novo* assembly.
- Call DNA variants.

A Typical Sequence Analysis Pipeline

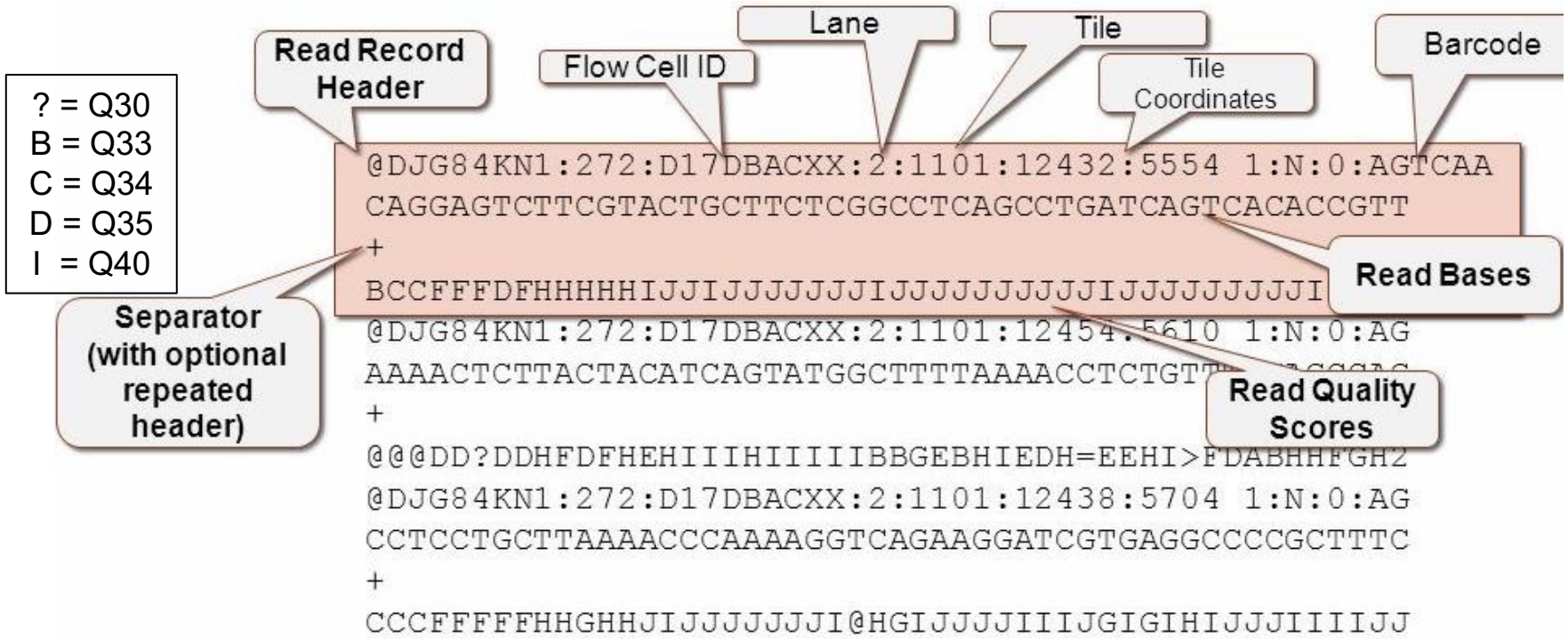


*.bam and *.pbi
PacBio

```
samtools index example.bam
```

- Index files must reside in the same directory as the BAM or SAM file.
- CRAM are compressed versions of alignment files like BAM.

The FASTQ File Format



Probability of Incorrect Base Calls

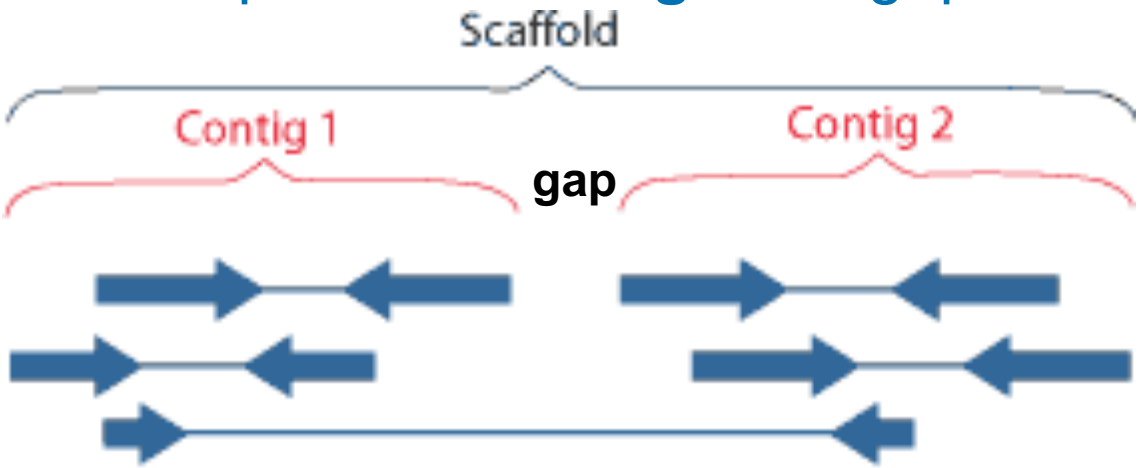
- Assess or measure accuracy of base calling.
- Sequence quality Q is reported on a log scale.
- Defined as a property related to the base calling error probabilities (P): $Q = -10 \log_{10}(P)$

Phred Quality Score (ASCII QS)	Probability of Incorrect Base Call	Call Accuracy (%)
Q10 (+)	1 in 10 bases	90
Q20 (5)	1 in 100 bases	99
Q30 (?)	1 in 1,000 bases	99.9
Q40 (I)	1 in 10,000 bases	99.99
Q50 (S)	1 in 100,000 bases	99.999

- **Q30 means that virtually all bases in a read are called correctly.**

Contigs and Scaffolds

- A **contig** is a contiguous length of unambiguously assembled genomic **sequences** in which the order of bases is known to a high confidence level.
- A **scaffold** is a portion of the genome **sequence** reconstructed from end-**sequenced** whole-genome shotgun clones. **Scaffolds** are composed of **contigs** and gaps.



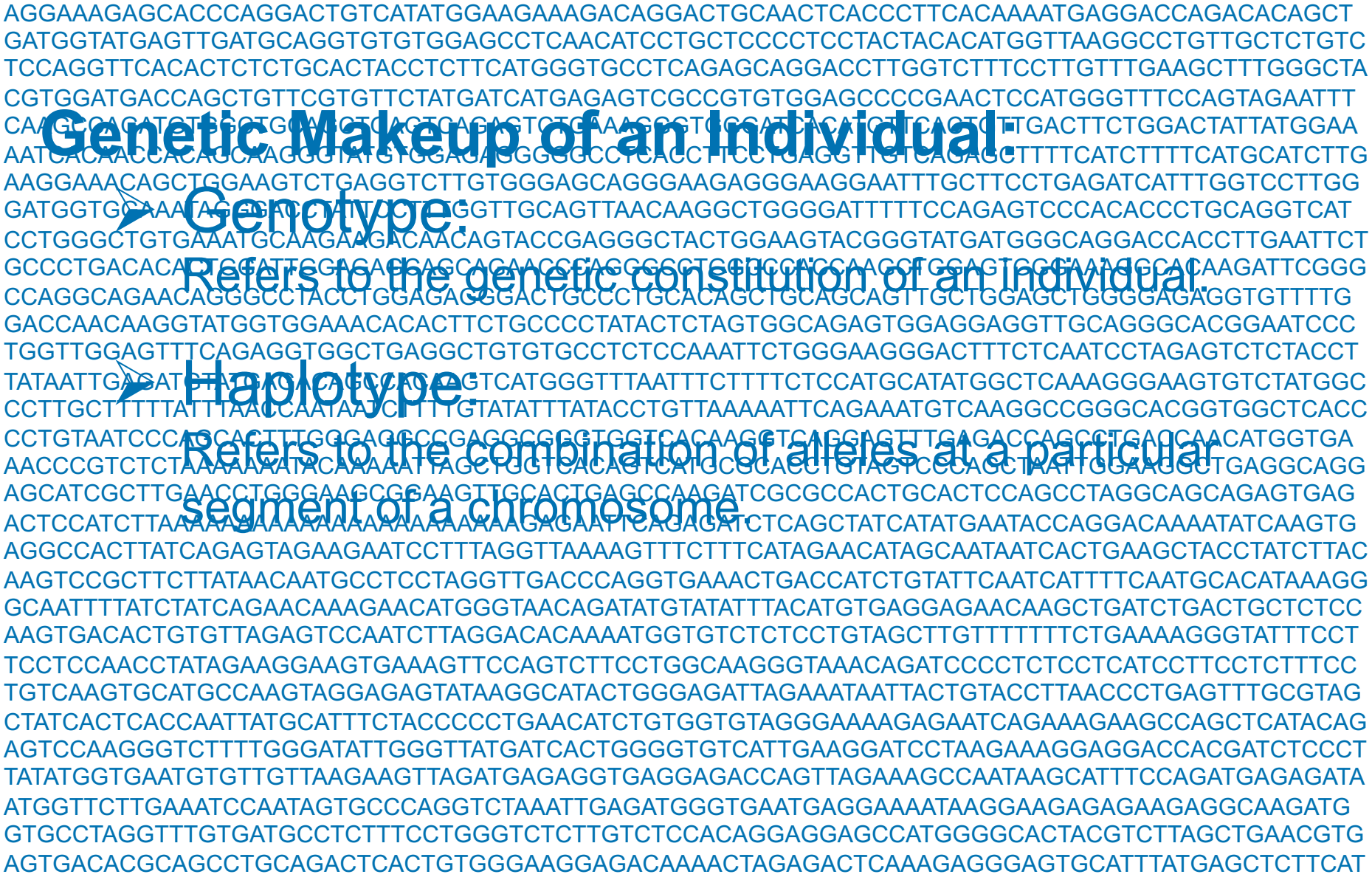
The order and gaps between the contigs are unknown.

Fragment
Read (known sequence)
Roughly known length but not known sequence

N50 Values

- **N50** is a measure to describe the quality of assembled genomes that are fragmented in contigs of different lengths.
- **N50** is defined as the minimum contig length needed to cover 50% of the genome.
- It means, half of the genome sequence is in contigs larger than, or equal to, the **N50** contig size.
- Higher **N50** contig size is usually always better.
- More repetitive genomes, and lower-quality or shorter reads will reduce the **N50.**

What we Ascertain?

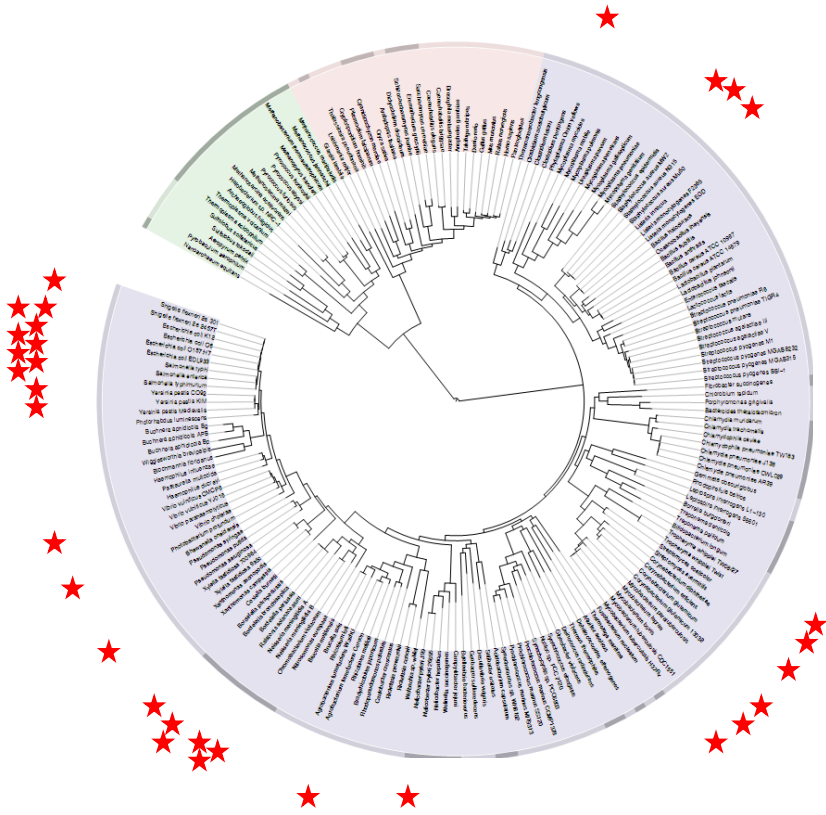


Genetic Makeup of an Individual:

- **Genotype:**
  Refers to the genetic constitution of an individual.
- **Haplotype:**
  Refers to the combination of alleles at a particular segment of a chromosome.

What Matters Most?

- Read length.
- Throughput.
- Data quality.



Pacific Biosciences (PacBio)

PacBio

- Single Molecule Real Time (SMRT) DNA Sequencing
- Current technology of choice for *de-novo* sequencing projects.

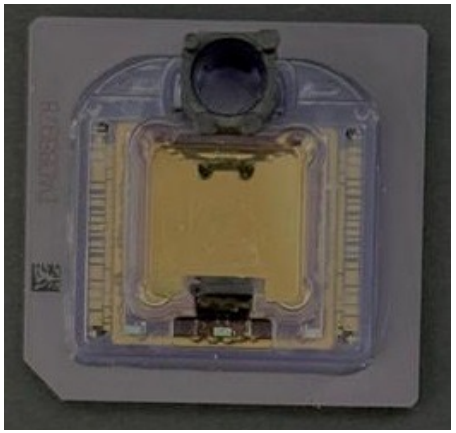
RSII : 1800 lbs. and ~11 feet long !

https://youtu.be/WMZmG00uhwU

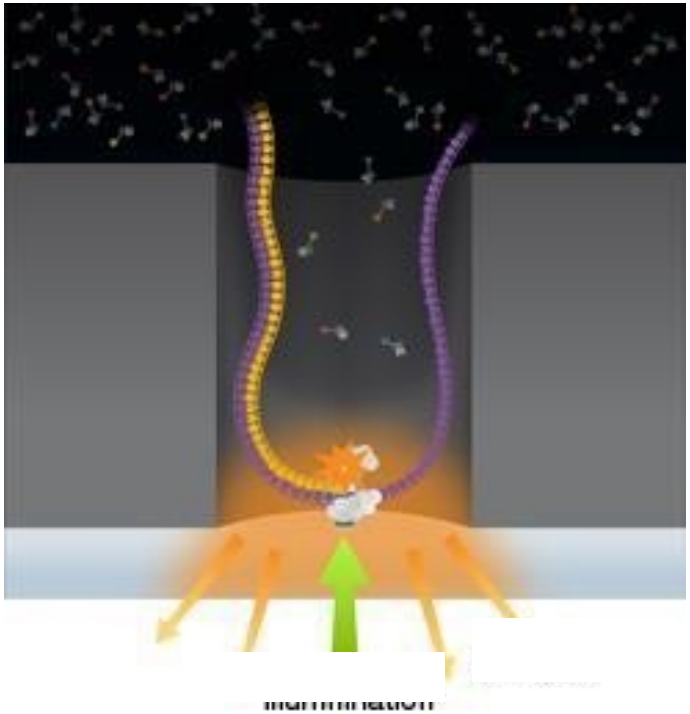
Sequel IIe

PacBio Flow Cell

- Single polymerase molecule bound to SMRT bell loaded in a 20 nm chamber, termed zero-mode waveguide (ZMW).
- Records incorporation in real time.
- ~2 bases per second.



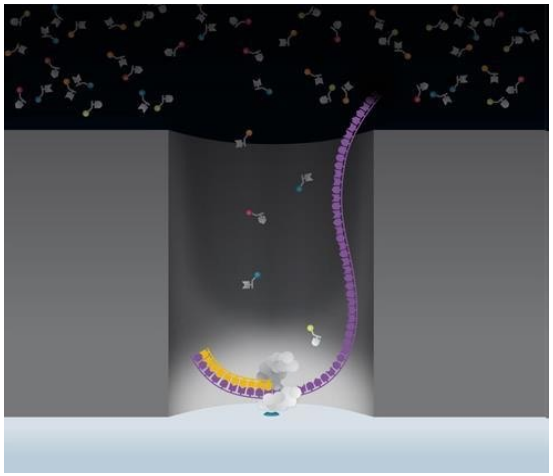
PacBio Sequel IIe Flow Cell
8M ZMW



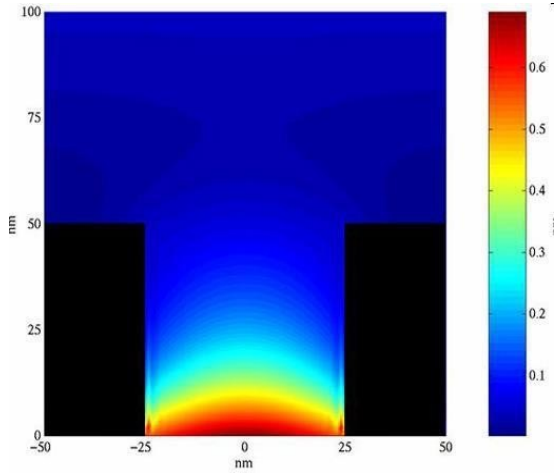
PacBio ZMWs

- Laser light illuminates and penetrates through the lower 20 - 30 nm of each ZMWs.
- Detection volume is 20 zeptoliter (10^{-21} liters).

Individual ZMW

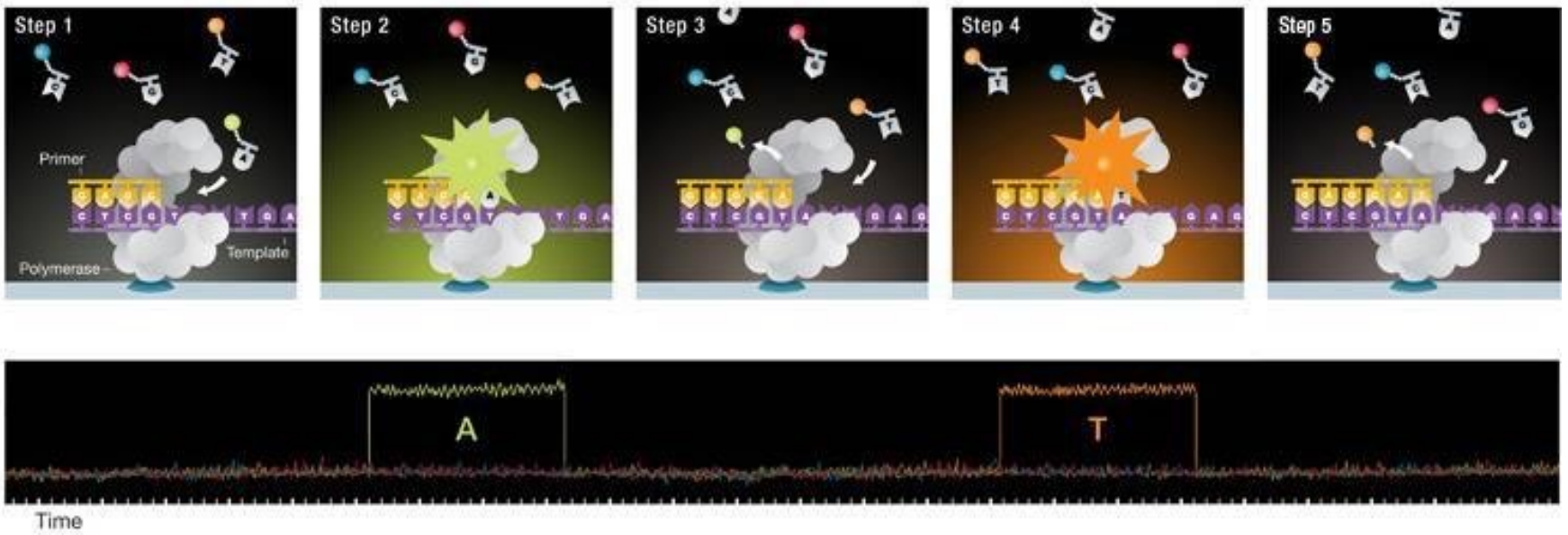


ZMW with polymerase and nucleotides

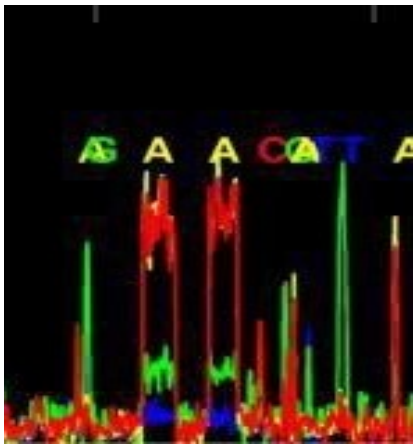


Laser light illuminates the ZMW

PacBio Sequencing Cycles



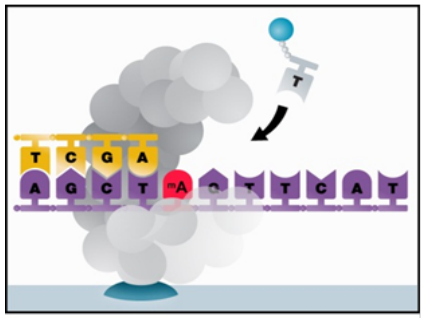
PacBio Sequencing

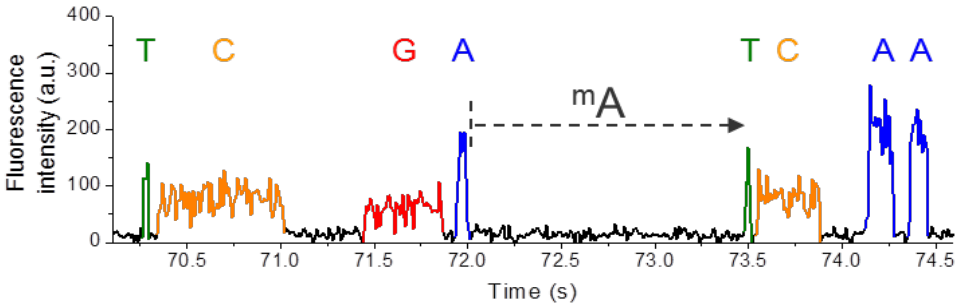


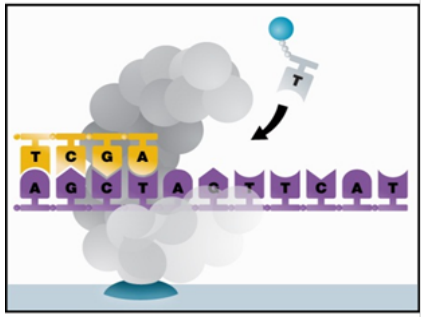
- Labelled dNTPs.
- Some bases added very quickly and missed.
- Some wrong bases flirt with active site and go away.
- Captures base modification via kinetic information.

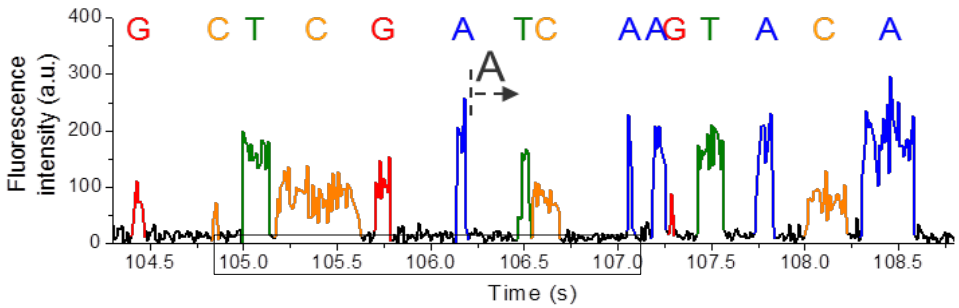
Detection of DNA Base Modifications

Example: N^6-methyladenine









- SMRT Sequencing uses kinetic information from each nucleotide addition to call bases.
- This same information can be used to distinguish modified and native bases by comparing results of SMRT sequencing to an *in silico* kinetic reference for incorporation dynamics without modifications.

PacBio Long-Read Sequencers

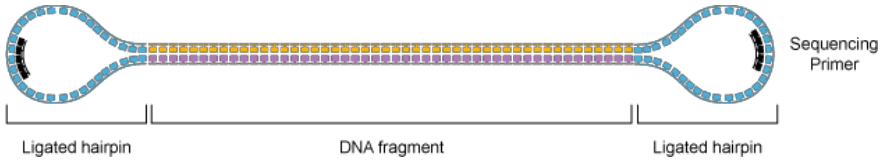
Platform (Run Time)	Flow Cell	Maximum Output (Gbp/run)	Cost/ Human Genome (US$)
Sequel (1 day)	SMRT 8M	0.144 - 1.2 Gb	4,000
Sequel II and Iie (1 – 2 days)	SMRT 8M	24 Gb	4,000
Revio (1 day)	SMRT 25M	120 - 480 Gb	1,100
Vega	SMRT 25M	60 Gb	1,100
Capillary sequencing	700-1000 bp	0.6 Gb	3,000,000,000

- Revio
- 2023 release.
- 25 million ZMWs/SMRT.
- 120 Gb/SMRT cell (4 cells).
- US$800K instrument.

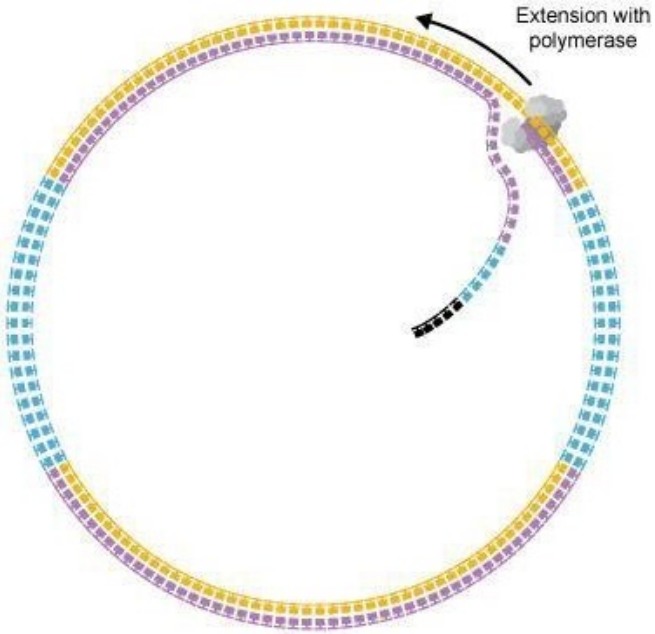
- Vega
- 2025 release.
- 25 million ZMWs/SMRT.
- 30 - 60 Gb/run.
- US$169K instrument.

PacBio Template Preparation

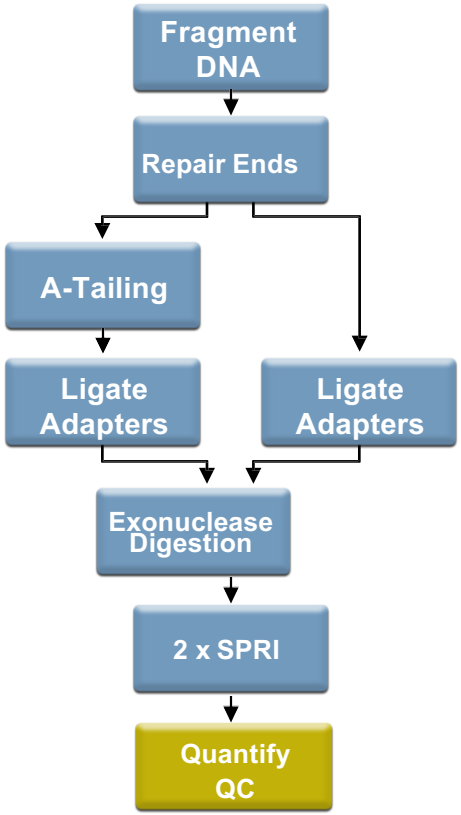
- The aim of library preparation is to obtain nucleic acid fragments with adapters attached on both ends.

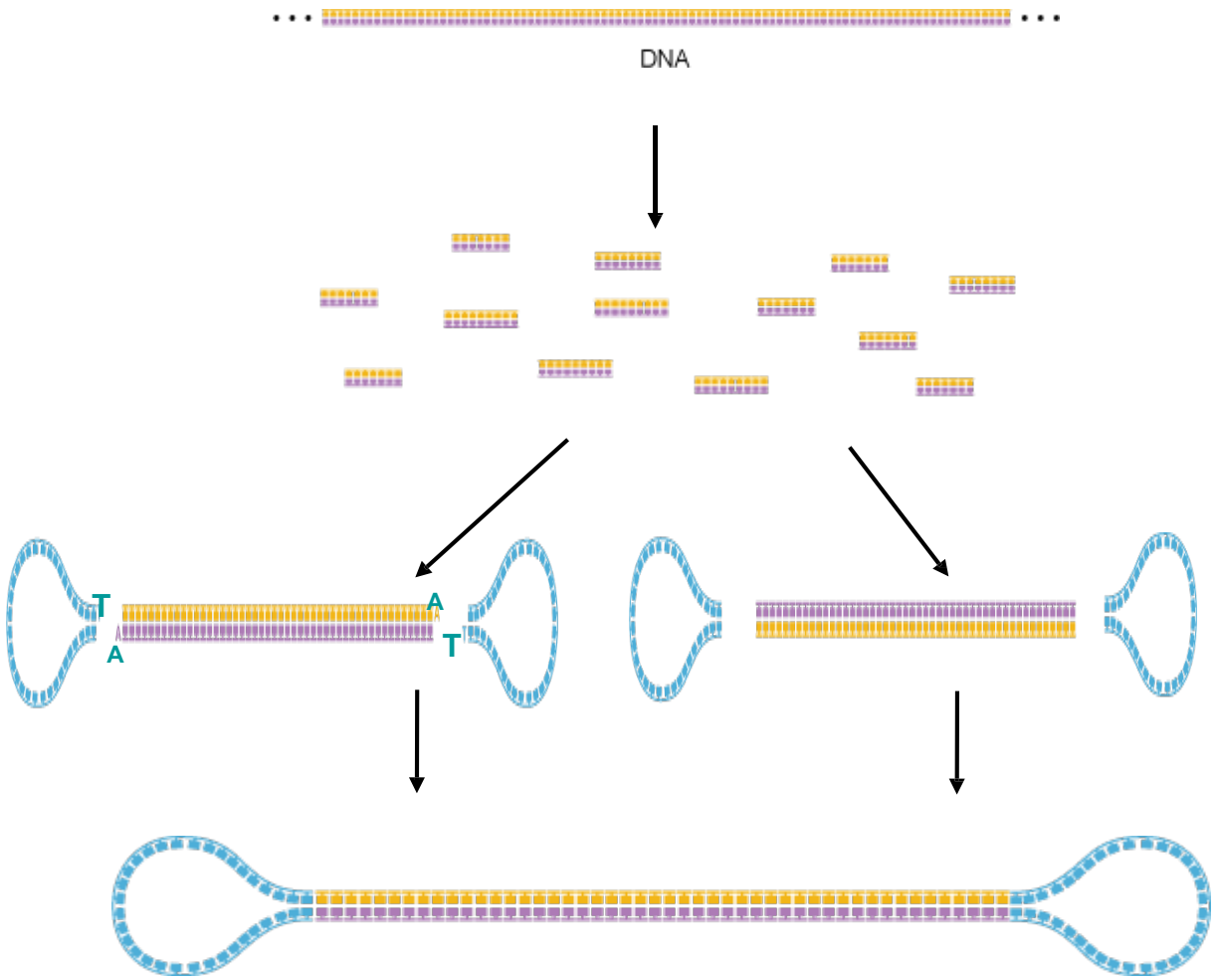


1. **Anneal primer.**
2. **Bind polymerase.**
3. **Sequence.**

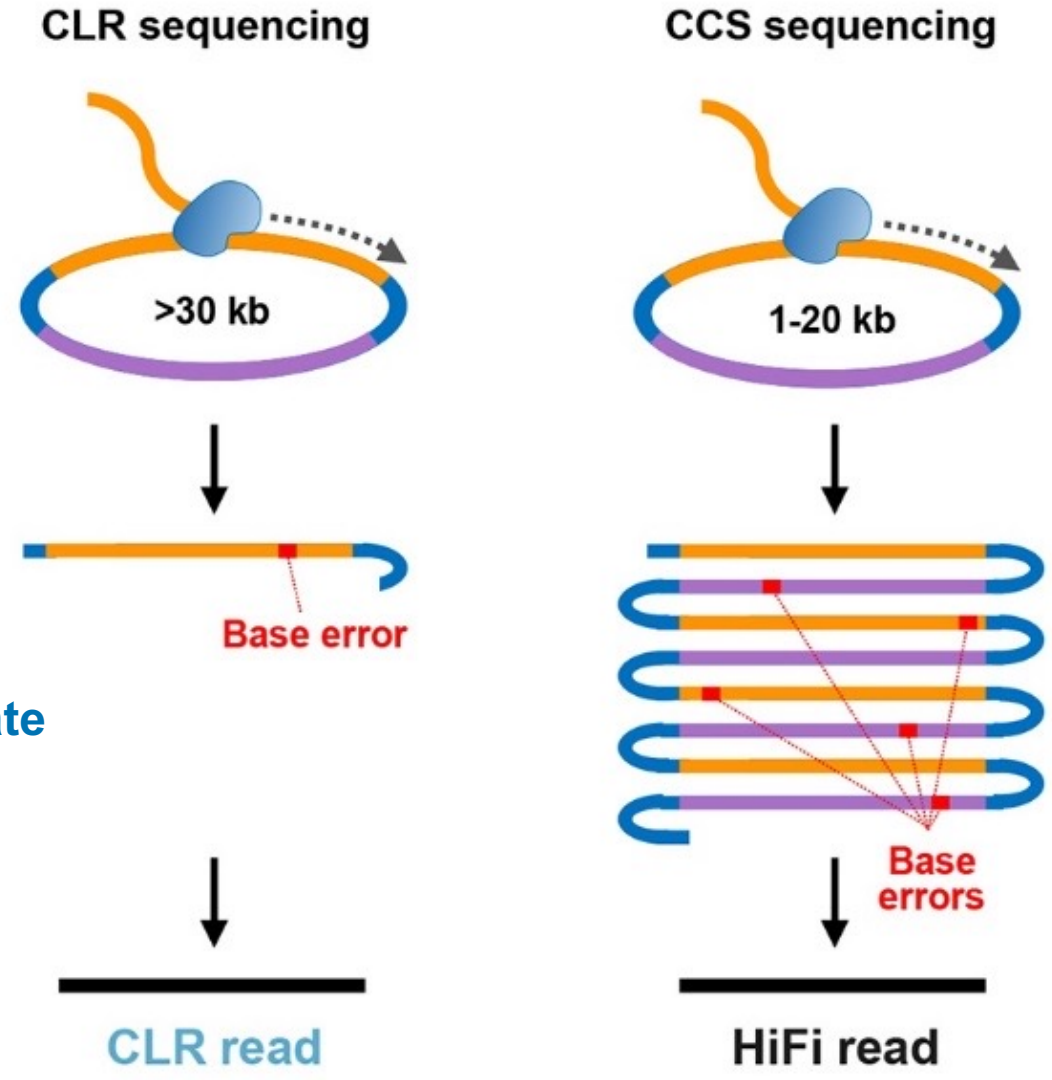


PacBio Library Prep





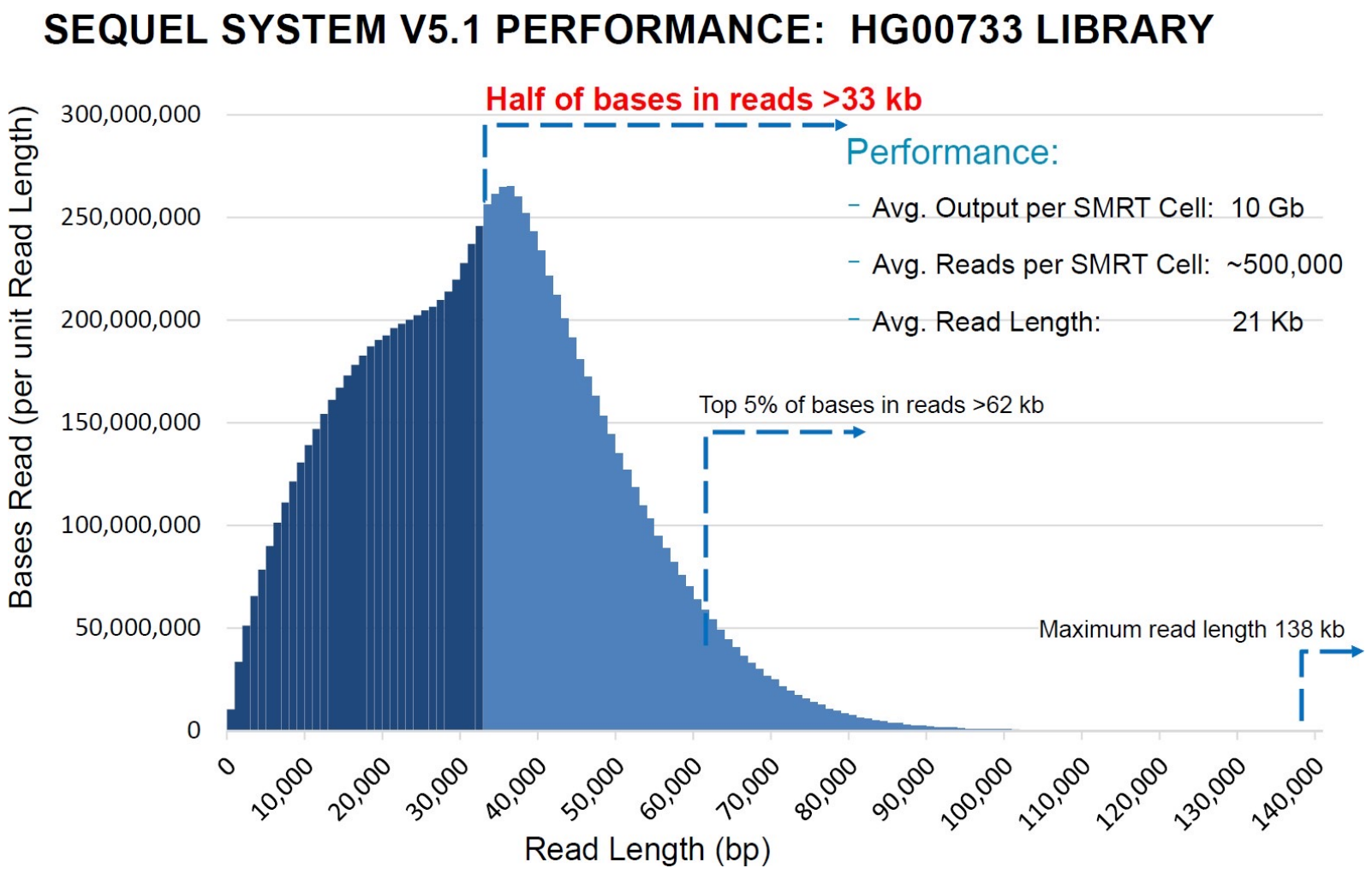
CLR and Hi-Fi Reads



8 - 15% error rate

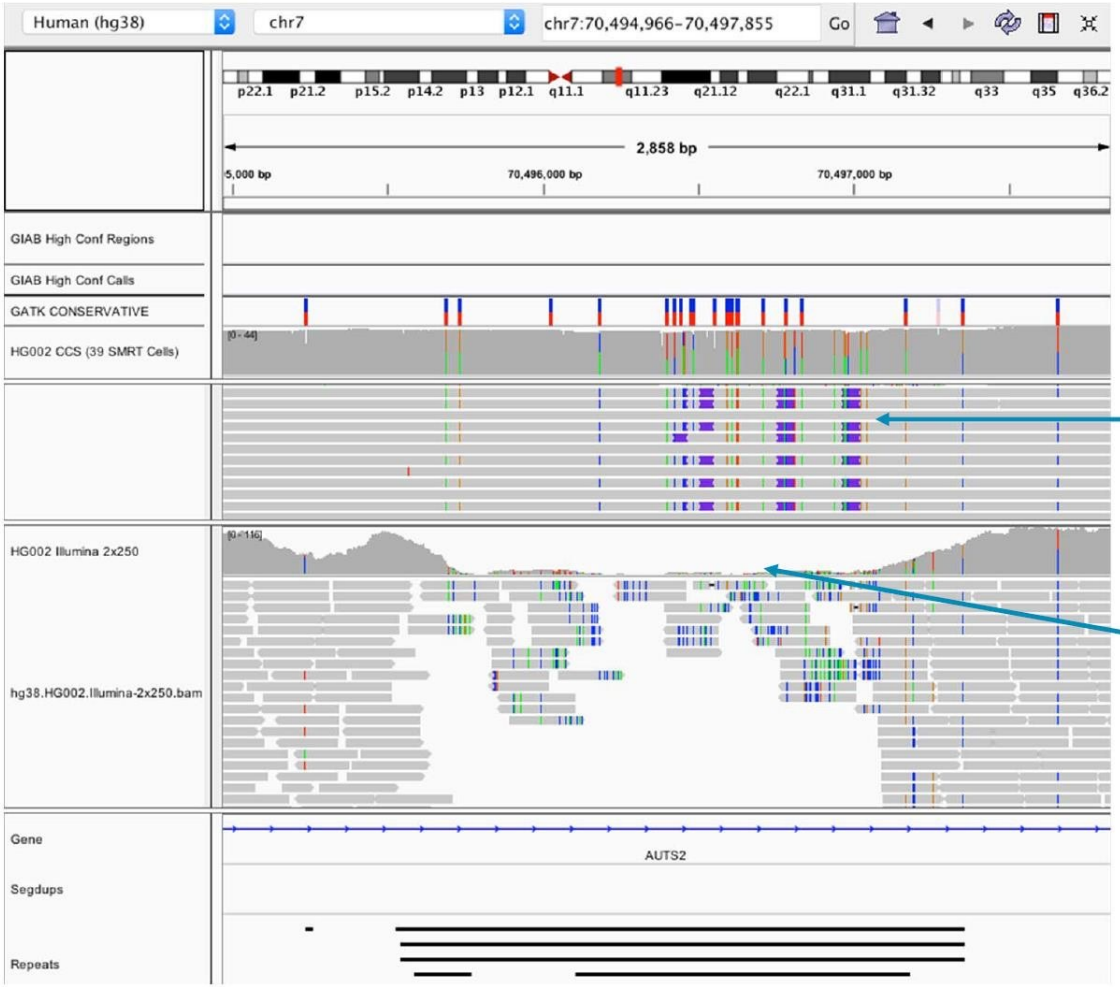
<1% error rate

PacBio HiFi Sequencing



PacBio Detecting Structural Variation

INTRONIC INSERTION IN *AUTS2* (AUTISM)



PacBio reads show insertions

Illumina reads do not map

Tandem repeat

Oxford Nanopore Technologies (ONT)

ONT

- Another long read sequencing platform.
- Sequences single molecules of DNA or RNA as they travel through a nanopore.
- Reads typically 5 – 50 kb but some can be as long as the template (~4 Mb). However, shorter fragments give higher yield.
- Fast moving technology that may soon be cheaper and have a higher throughput than Illumina.
- Low capital cost price structures available.

The Nanopore

PromethION Flow Cell
100 – 200 Gb

MinION Flow Cell
15 - 35 Gb

Multiple nanopore sensors arrayed in one flow cell (P2 or MinION).

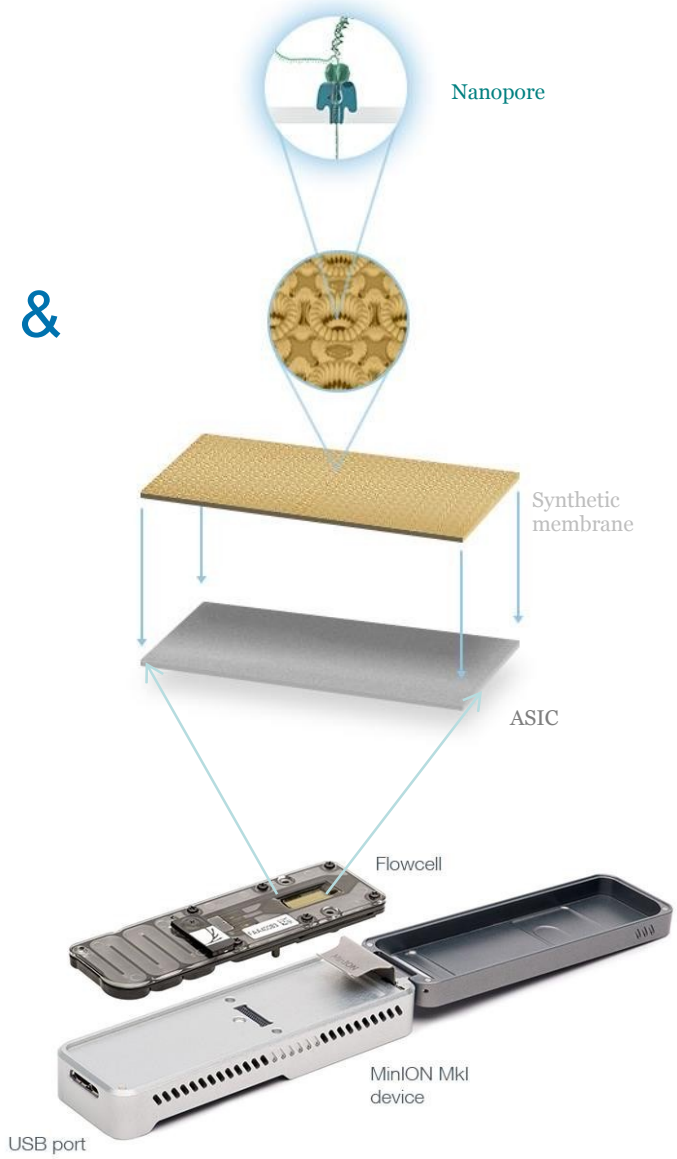
Operate independently but at the same time.

5-base words = 1024 current levels

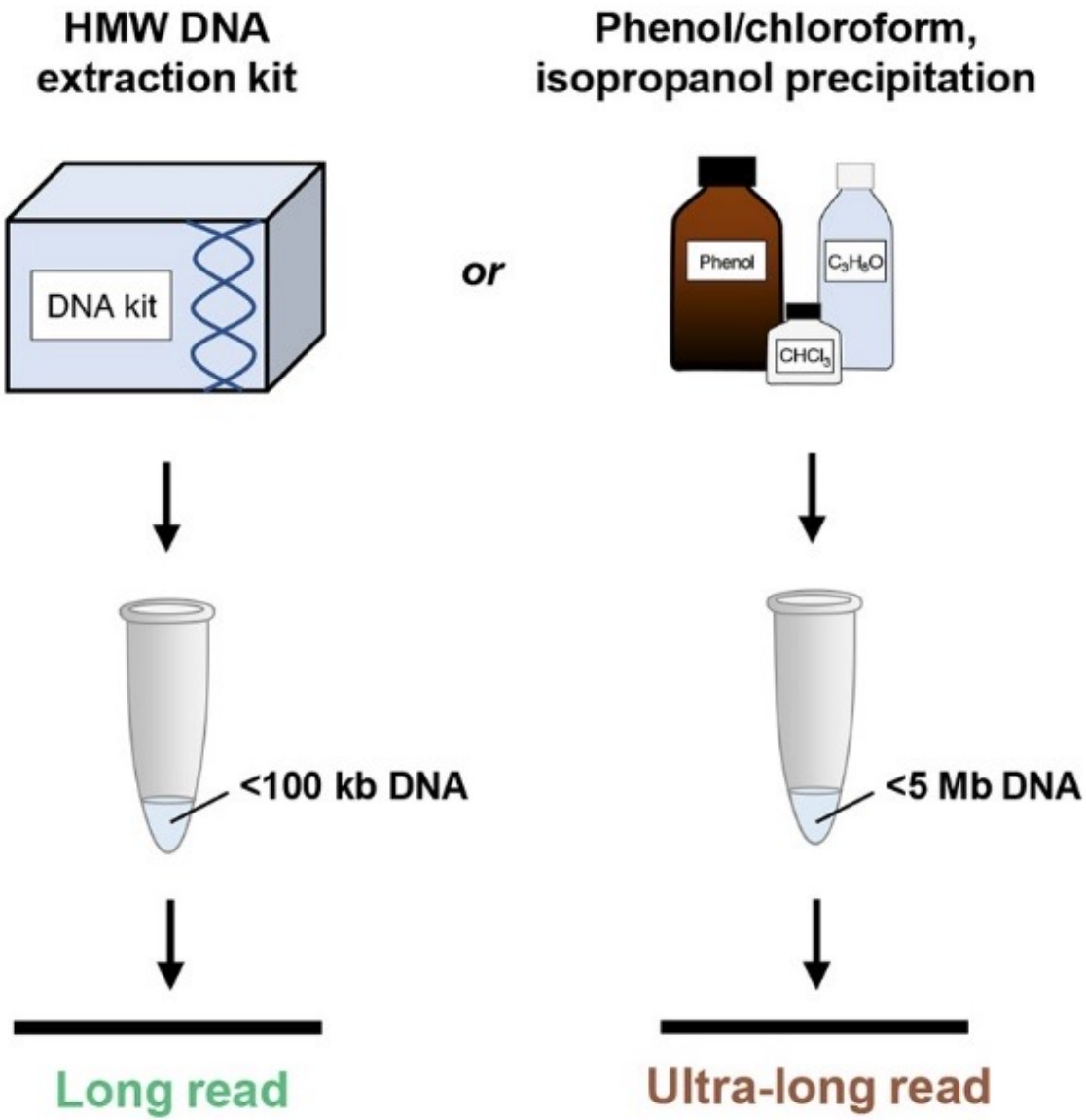
ONT Flow Cell Design

- Application-Specific Integrated Circuits (ASICs) contains 512 channels.
- Each channel is surrounded by 4 pores & records only 1 at the tie.

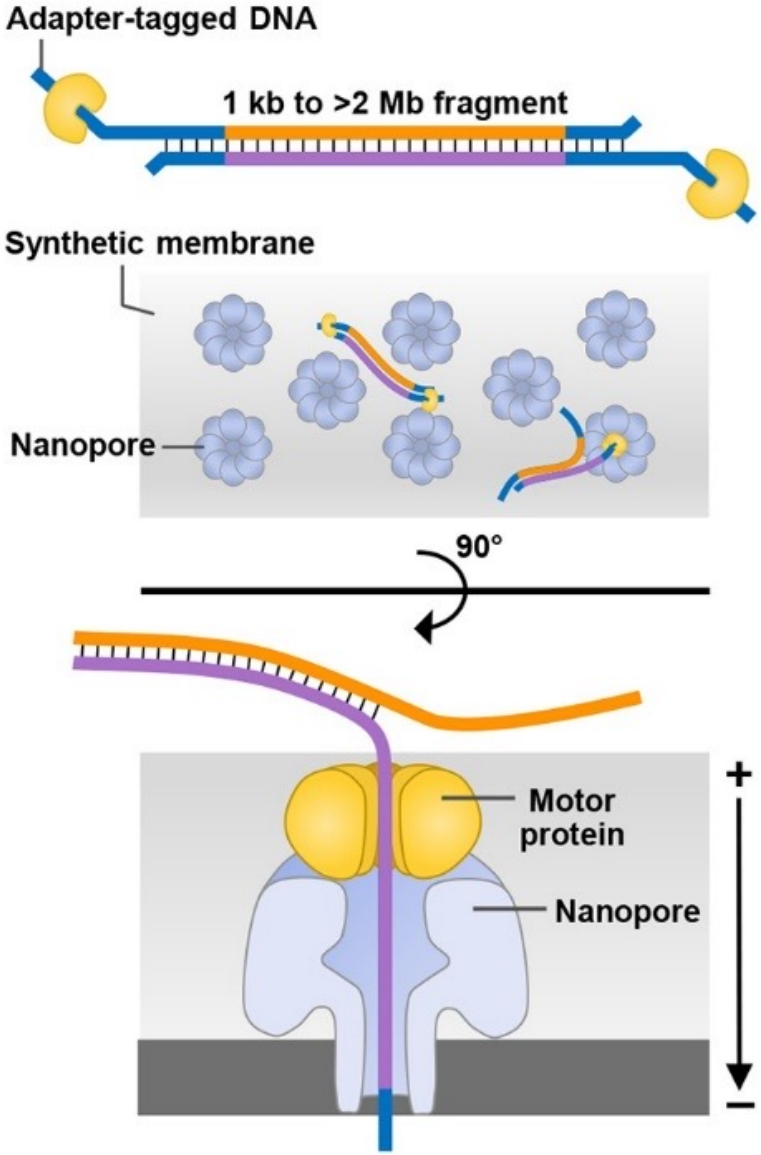
- A maximum total of 512 pores are recorded at a time.
- Scan for “fresh” active pores automatically every 24 hours or when manually restarted.



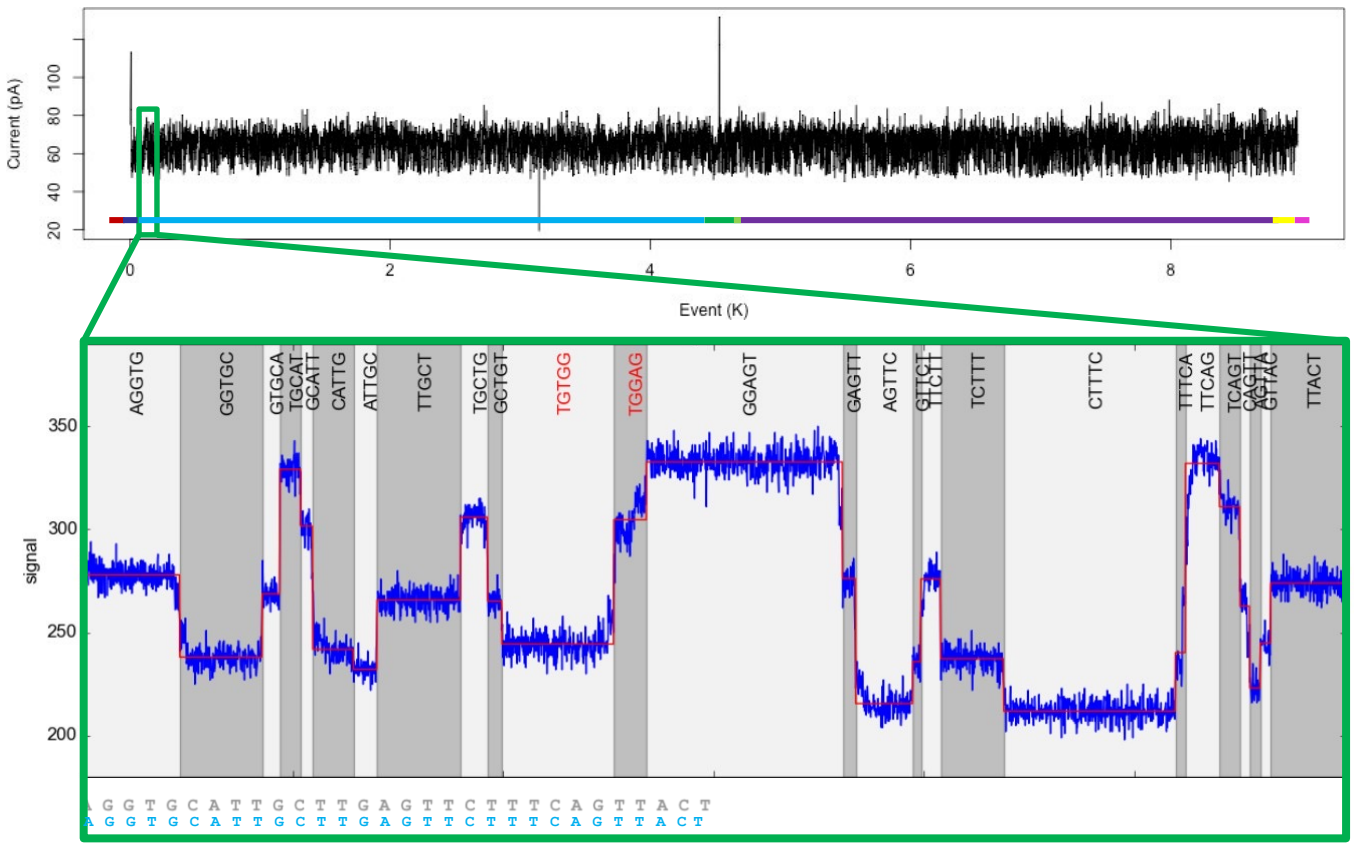
Ultra-long Reads

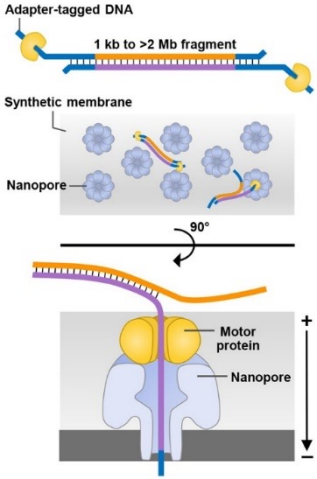


ONT Library Preparation



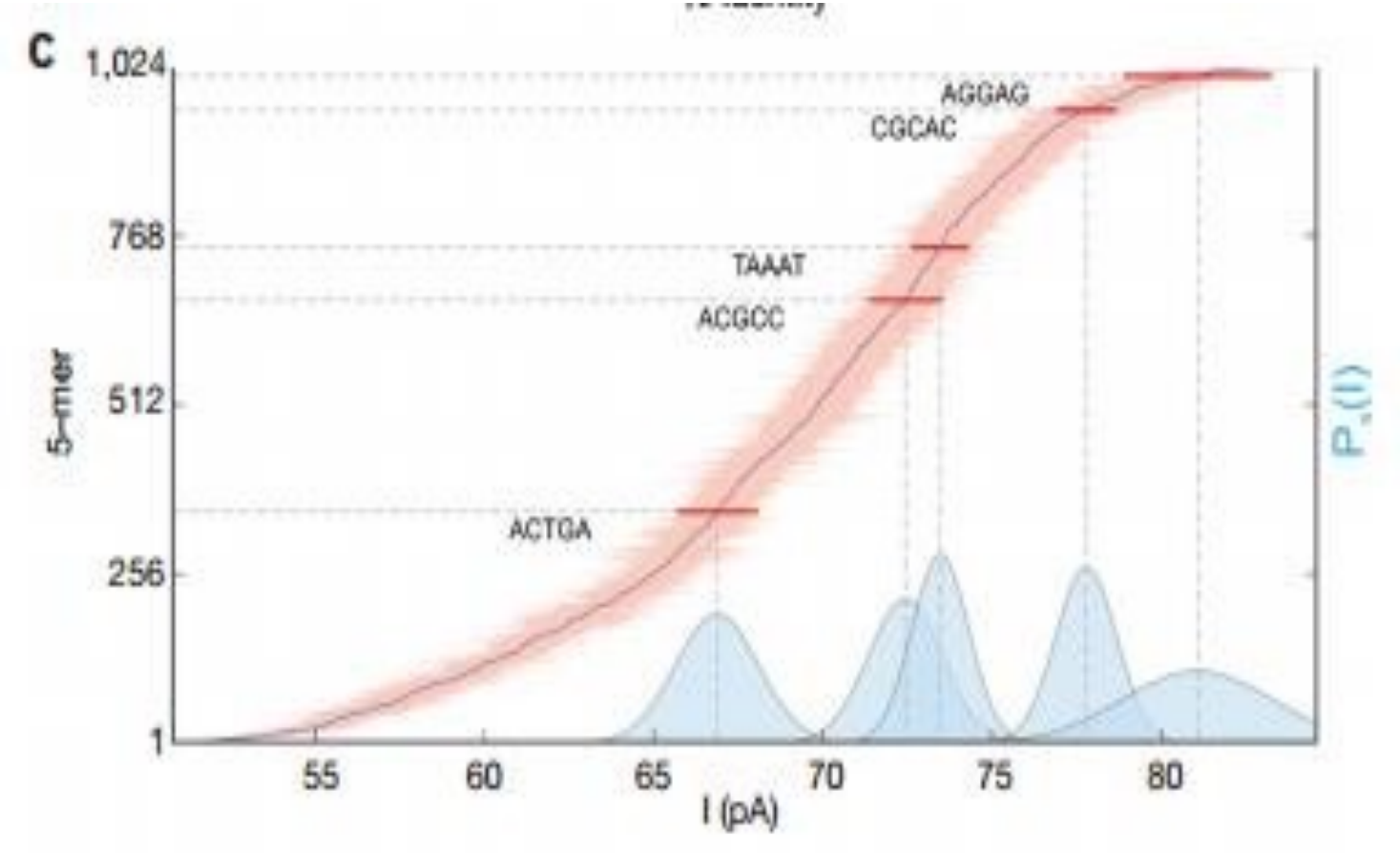
ONT: The squiggles





Slide courtesy of David Buck. WTCGH

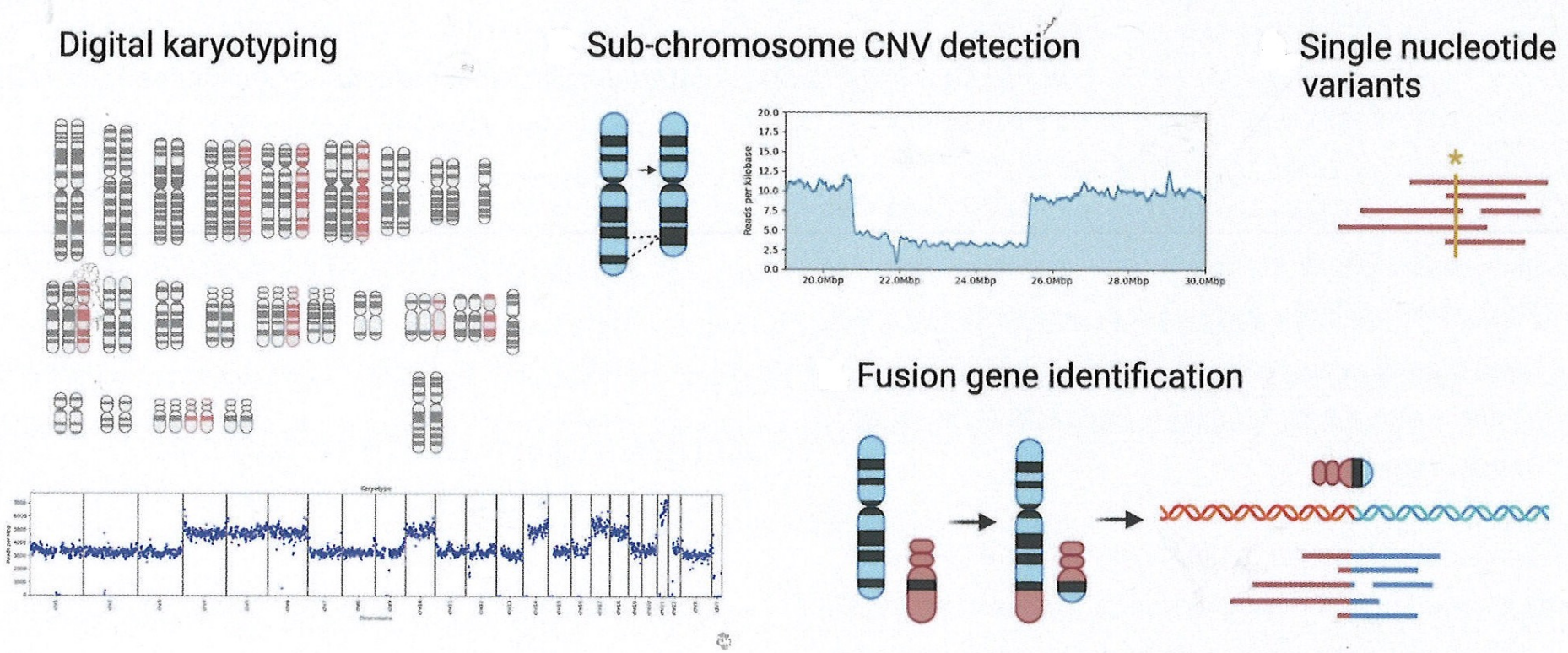
5mer Current Signals



From Szalay & Golovchenko, Nat. Biotech (2015).

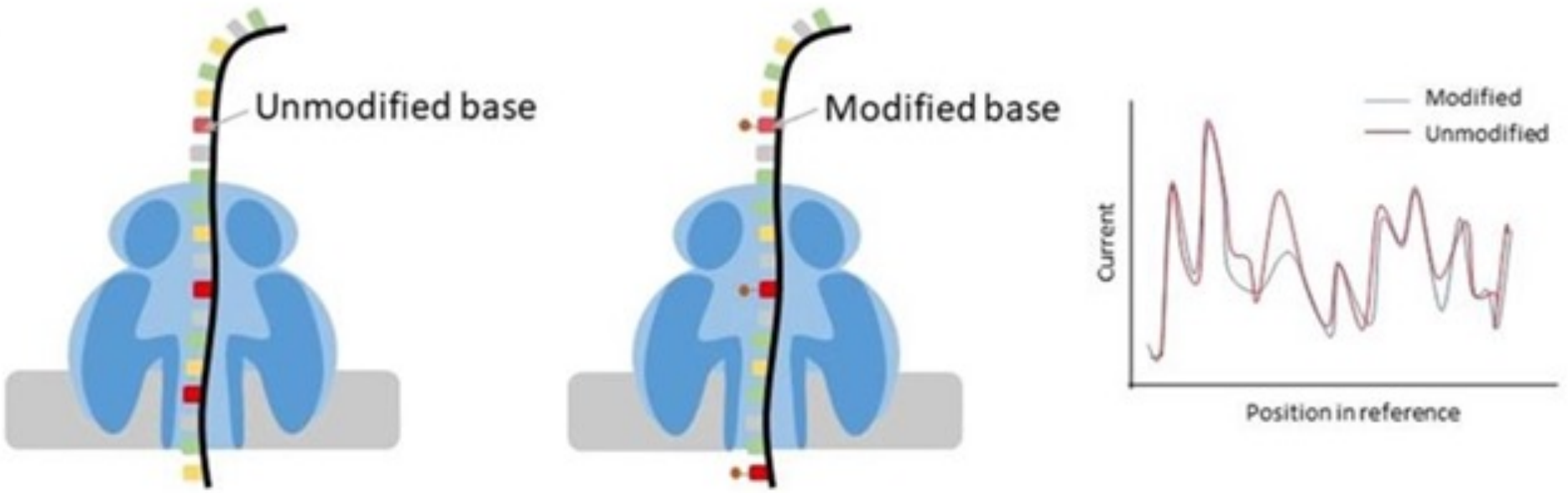
- Low error rates 1 - 5% comprising mostly of indels.

Captures All Variations



- Copy Number Variations.
- Structural Variations.
- STRs

Methylation in Real Time



From Szalay & Golovchenko, Nat. Biotech (2015).

ONT Hardware

Platform (Run Time)	Flow Cell	Maximum Output (Gbp/run)	Cost/ Human Genome (US$)
Flongle	Flongle	10 Gb	Not applicable
MinION (1 – 3 days)	Mk1D	15 - 35 Gb	Not applicable
GridION (1 - 3 days)	Mk1D	75 - 175 Gb	1,000
PromethION 2, 24, 48 (1 – 3 days)	P2	130 – 9.6 Tb	800

- Run until done.
- Selective reads.
- Mobile sequencing.

Pros and Cons of Long-Read Sequencing

- Complete chromosome sequencing from telomere to telomere.
- Ascertain all variations, including those in repetitive regions, complex structural variations and identify fusion genes.
- Allows high quality *de novo* genome assembly. Very high quality sequence in PacBio HiFi mode. Low error rates 0.1 - 2% comprising mostly of indels in ONT.
- Permit haplotyping.
- Captures full length cDNA or RNA molecules (isoform sequencing).
- Identify DNA base modifications.
- Overall more costly as compared to short-read sequencing.

Questions

Thank You For Listening

Contact Information

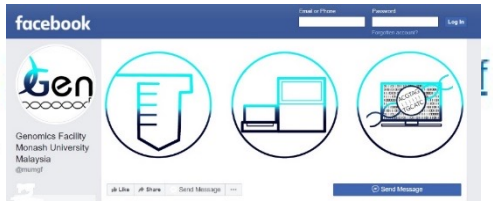
qasim.ayub@monash.edu

Central Email: mum.genomics@monash.edu

Telephone: +6 03 5514 6000 ext - Office: 61727; Lab: 61878

Find us on www.facebook.com/mumgf

Find us on: facebook



@genomicsMUMGP

Follow us on LinkedIn https://www.linkedin.com/company/mumgp