

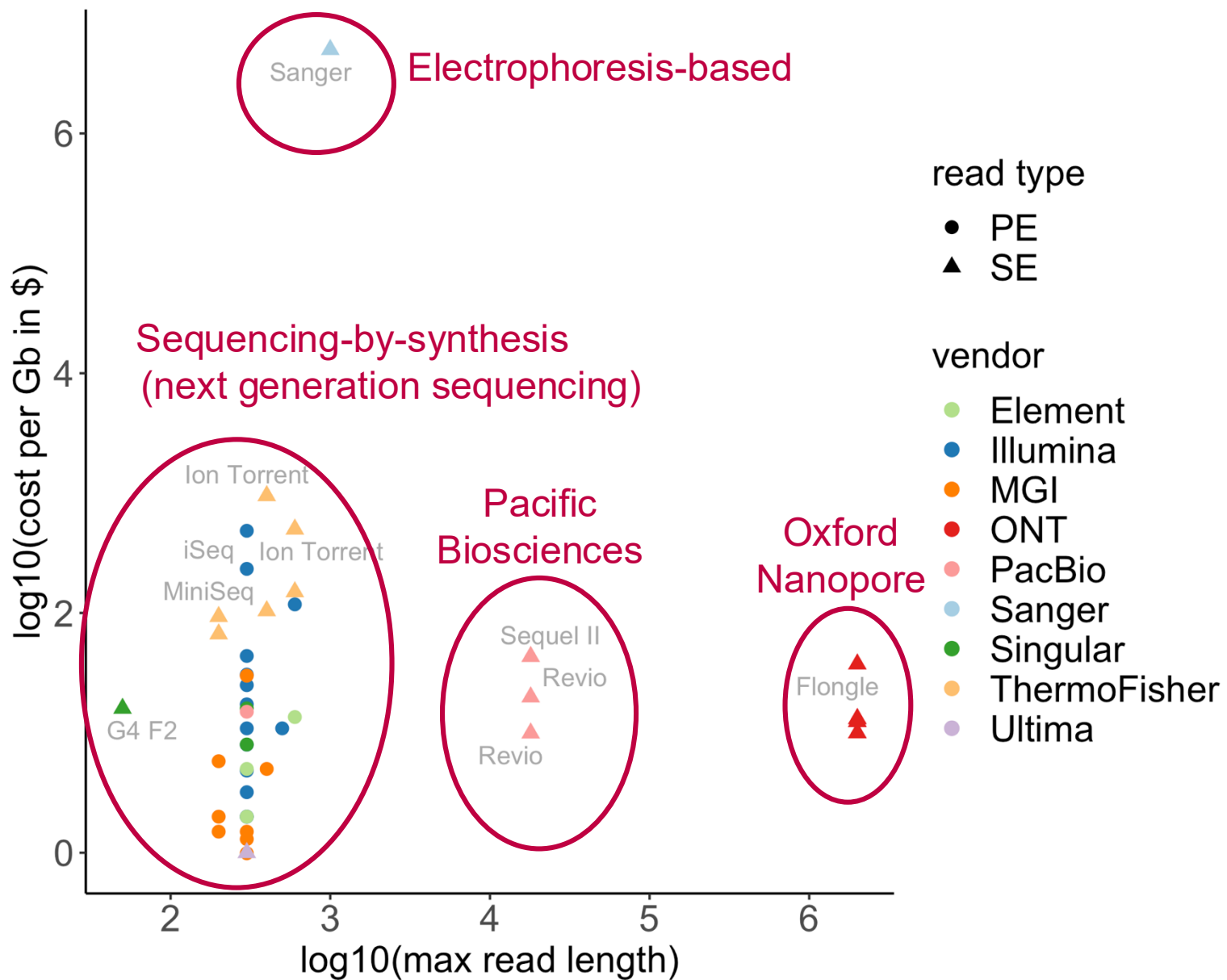
Long-read sequence analysis

Sequencing technologies

Question 4

What is a long read?

- Short read: 50-300 bp
- Long read: > 1kb, up to 20 Mb:
 - single molecule sequencing or
 - 3rd generation sequencing



Sequencing-by-synthesis

- 2nd generation sequencing
- Massive throughput: up to 500×10^9 bases/run
- Illumina still most used platform today



S I N G U L A R
GENOMICS

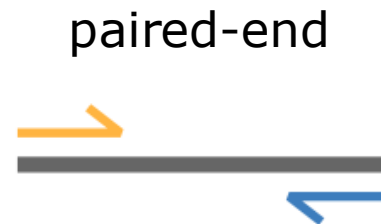
Sequencing-by-synthesis

- 50 – 300 bp
- Paired-end (or single-end)

Question 5

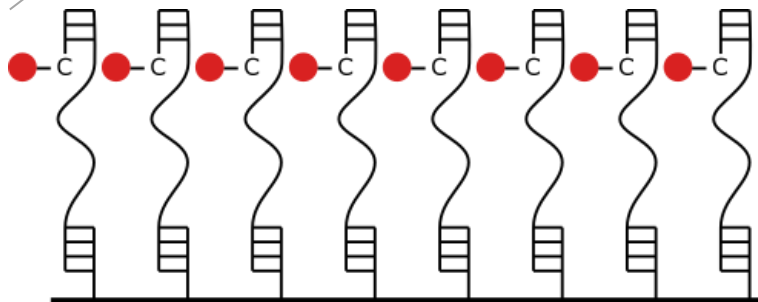
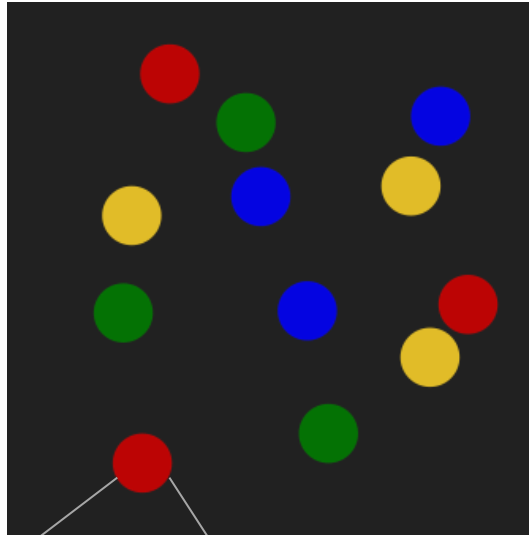
Sequencing-by-synthesis

- 50 – 300 bp
- Paired-end (or single-end)

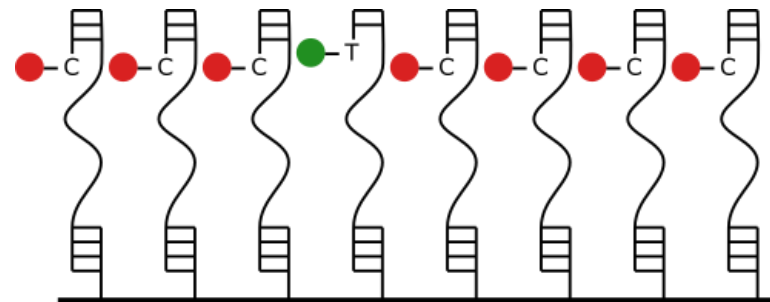


SBS - limitations

- Maximum read length: 300 bp
- How to reconstruct:
 - Repeats?
 - Isoforms?
 - Structural variation?
 - Haplotypes?
 - Genomes?
- Why not longer read lengths with Illumina?



in phase



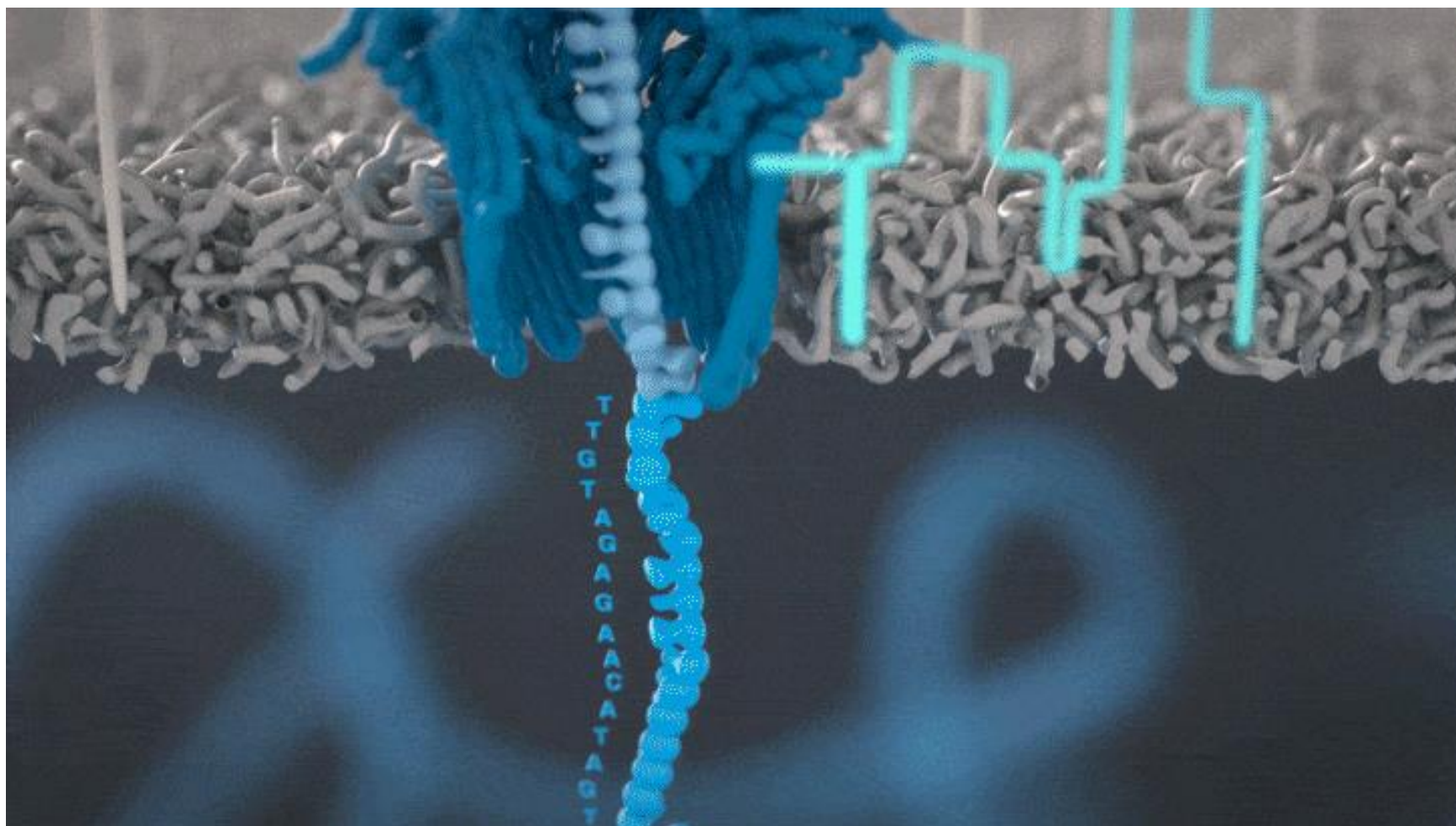
out of phase

Long reads (3rd generation)

- Crux: maximizing signal from a single-molecule base read-out
- Single molecule, so no out-of-phase signal
- Two frequently used platforms:
 - PacBio SMRT sequencing
 - Oxford Nanopore Technology

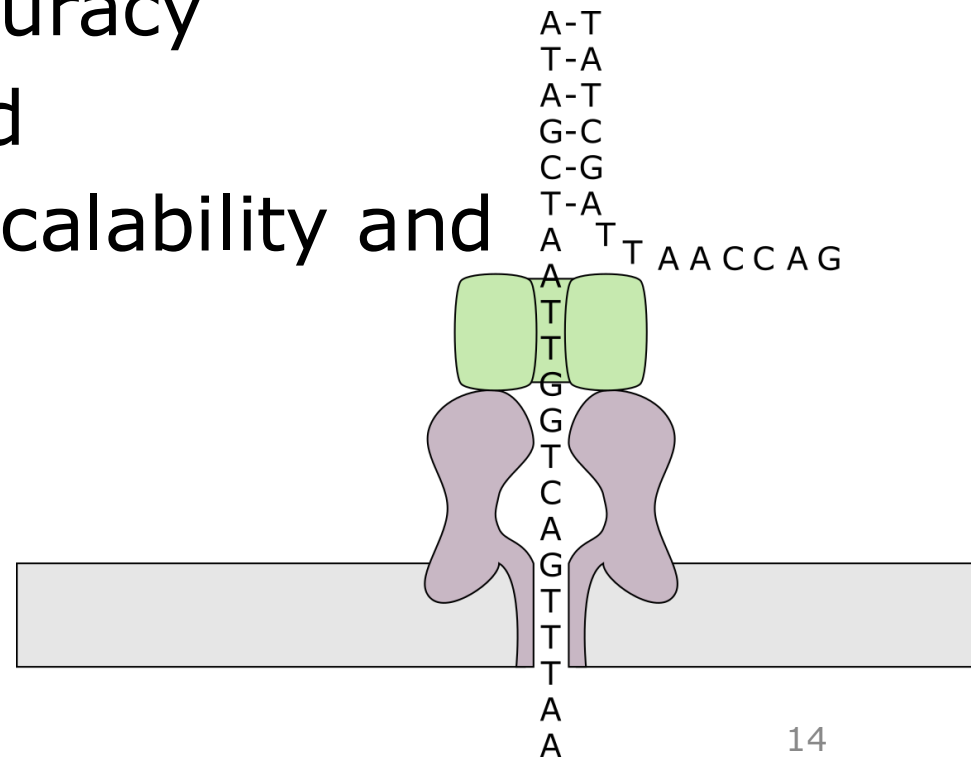


Question 6



Oxford Nanopore technology

- Based on changes in electrical current
- 4 bp (or more) read at a time
- Up to ~95-99% accuracy
- Errors can be biased
- Well-known for its scalability and portability



ONT scalability

1 small
flow cell:
1 x 2.8 Gb



Flongle

1 medium
flow cell:
1 x 50 Gb



MinION

5 medium
flow cells:
5 x 50 Gb



GridION

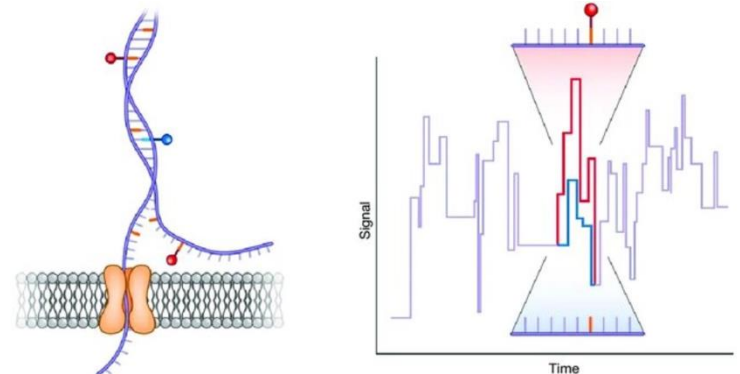
24-48 big
flow cells:
48 x 290 Gb



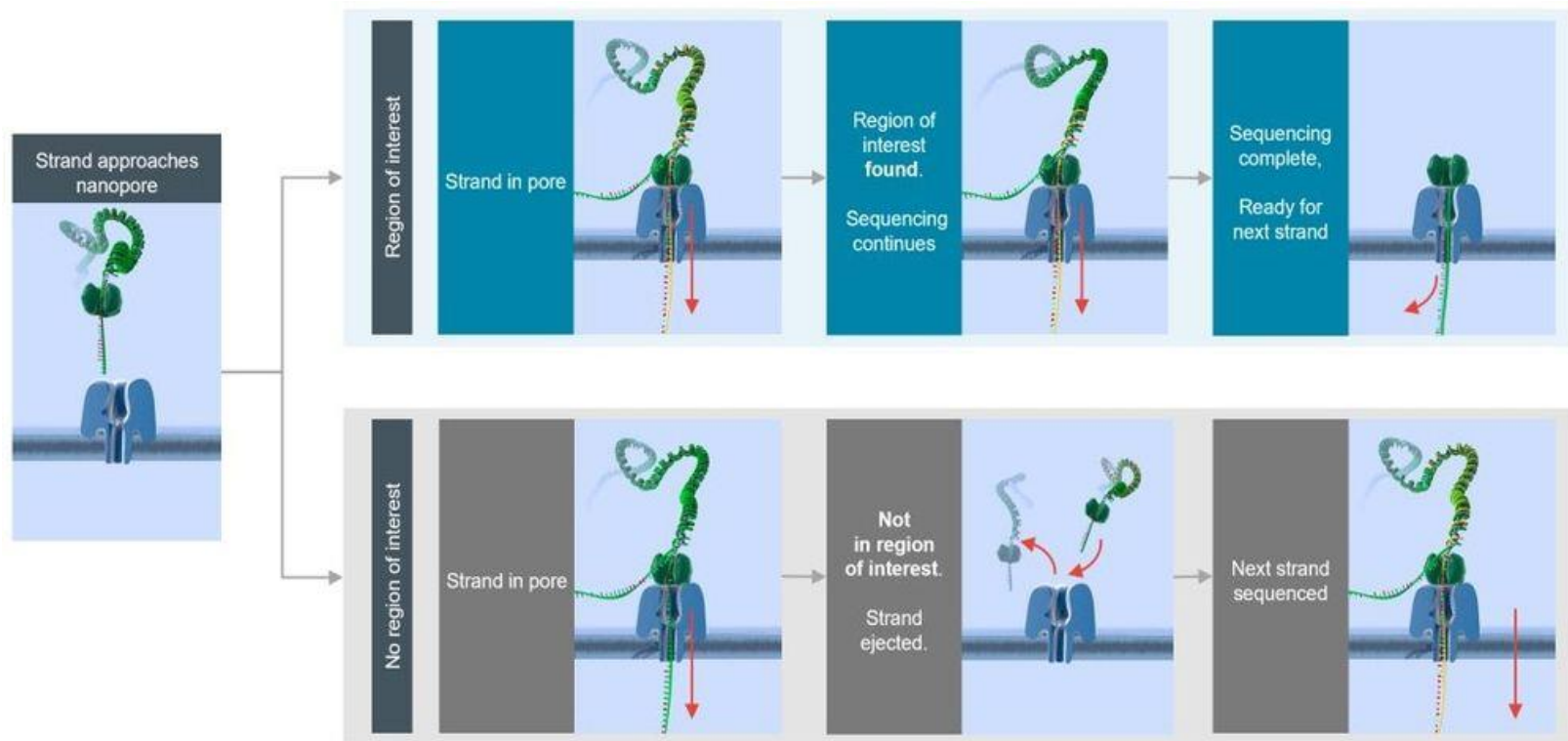
PromethION

ONT sequencing

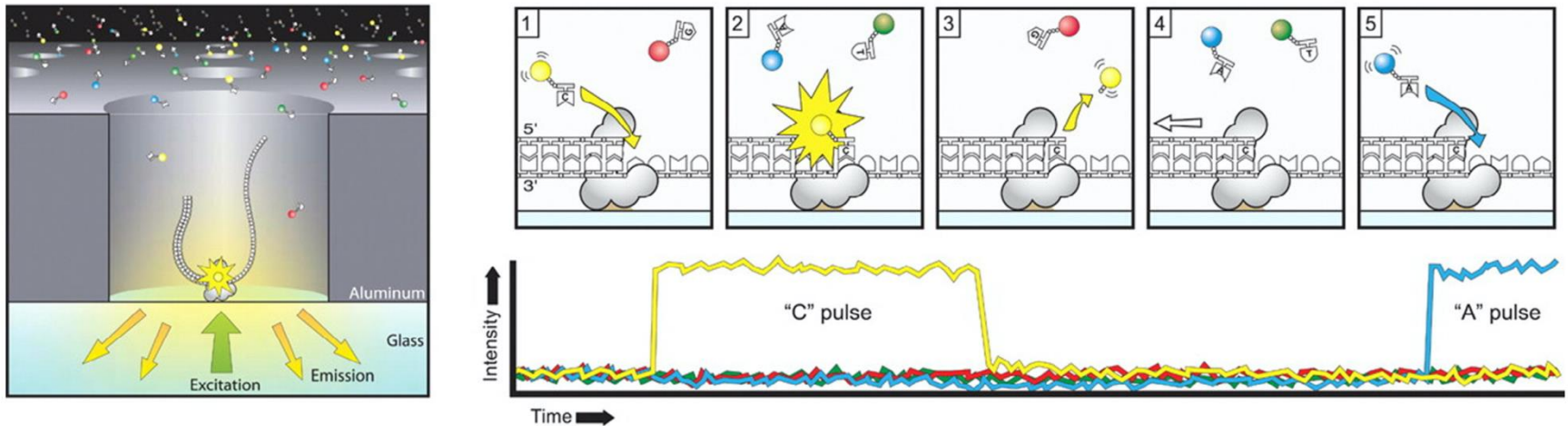
- Standard kit (library prep):
 - $>1 \mu\text{g}$ HMW DNA
 - Shearing + size selection is optional
 - Library prep with and without PCR
- Full-length native RNA or cDNA
- Enables simultaneous detection of epigenetic modifications



ONT adaptive sampling

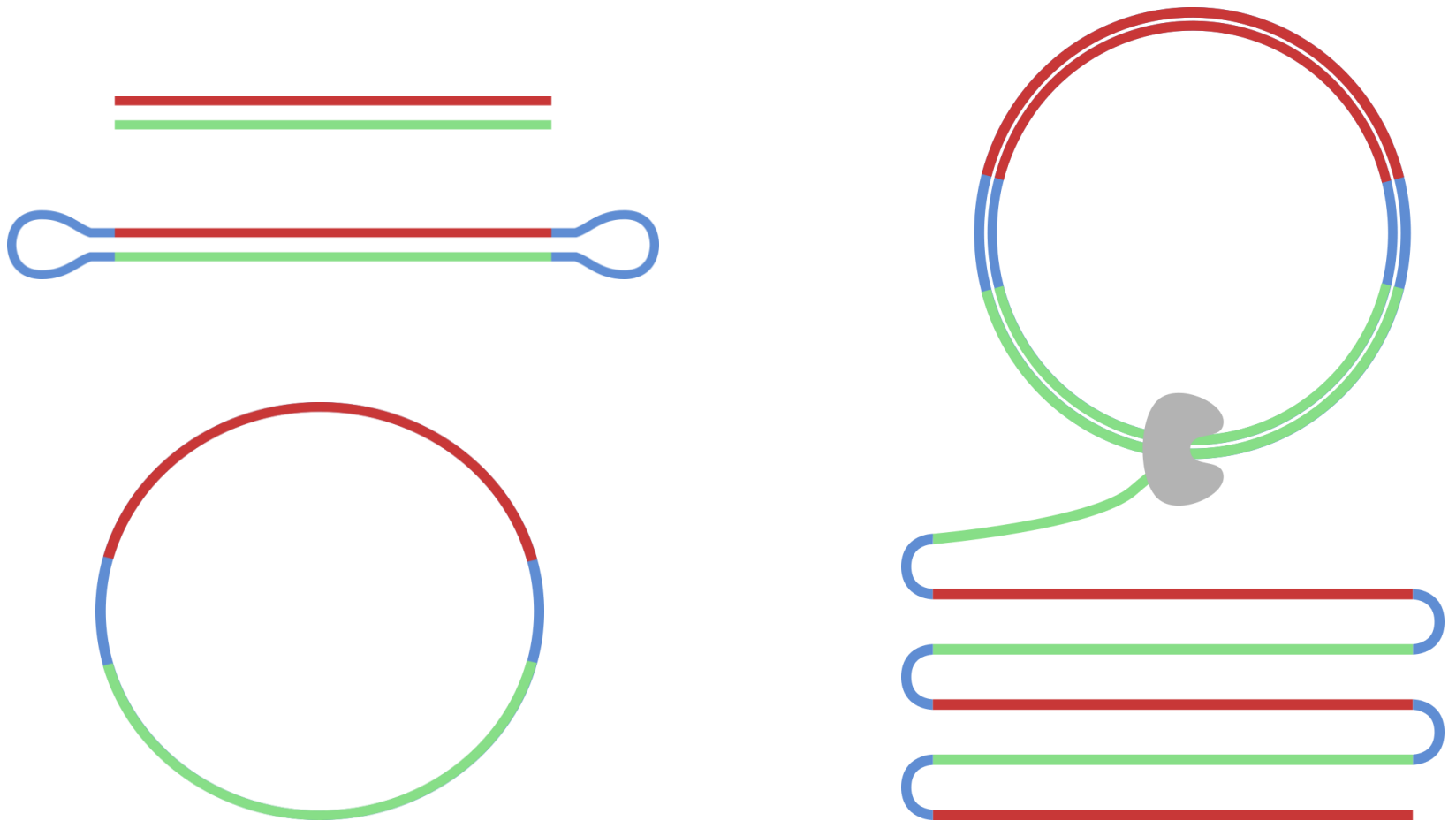


PacBio technology



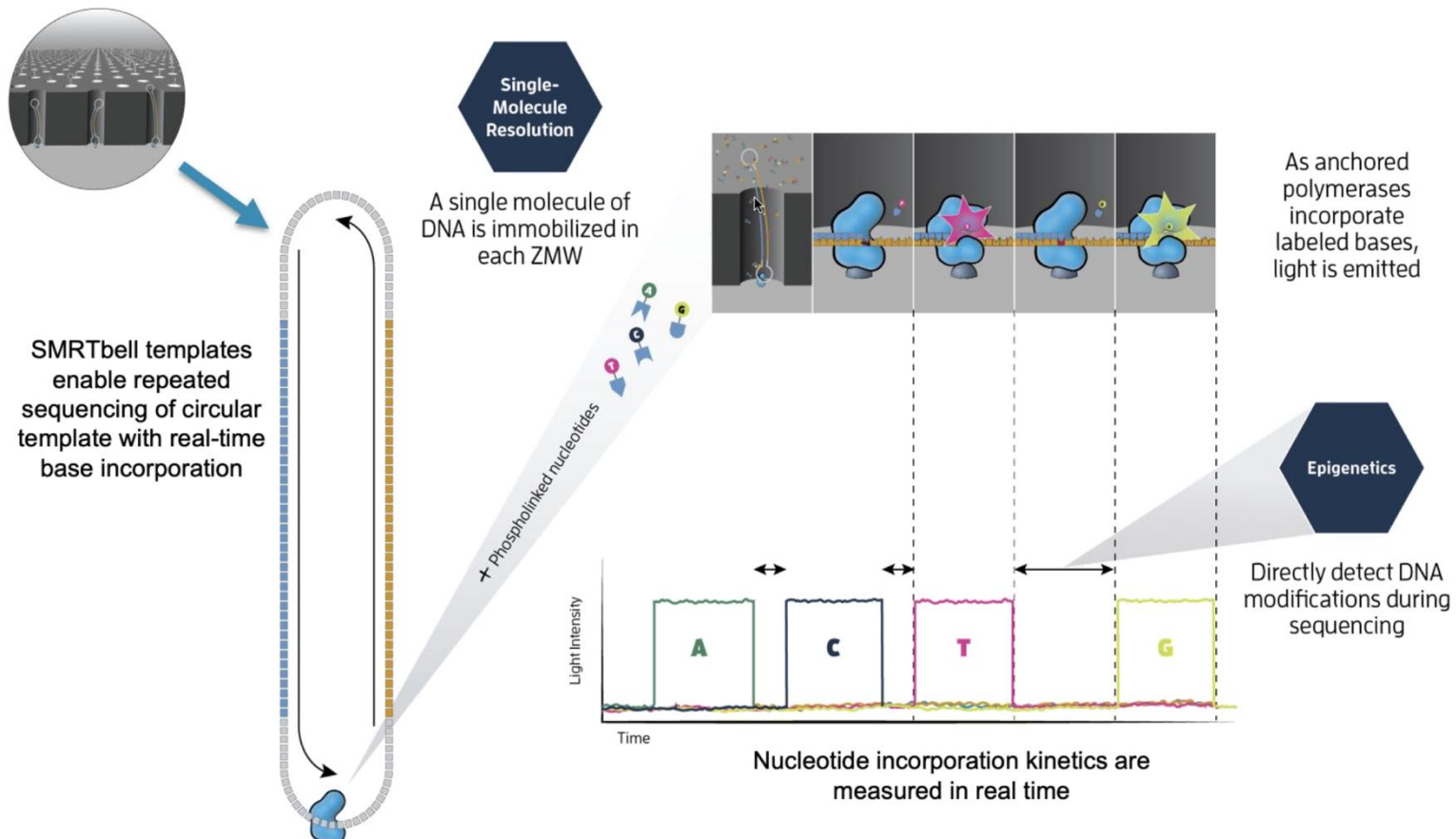
- Polymerase bound to ZMW bottom
- Circular molecules
- Single read out $\sim 90\%$ accuracy
- HiFi: single molecule sequenced multiple times
- Errors are relatively random

PacBio sequencing



Hi-Fi read

PacBio sequencing



PacBio sequencing

- Read length limitation: function of polymerase longevity and molecule length
- RNA sequencing:
 - Conversion into cDNA
 - Size selection
- Simultaneous detection of epigenetic modifications → DNA

PacBio Sequel IIe



- 8M ZMW
- ~2M HiFi reads/SMRT cell

Pacbio Revio

- 25M ZMW
- ~5-6M HiFi (90-120Gb) reads/SMRT cell
- 2-4 SMRT cells/run



Pacbio Vega

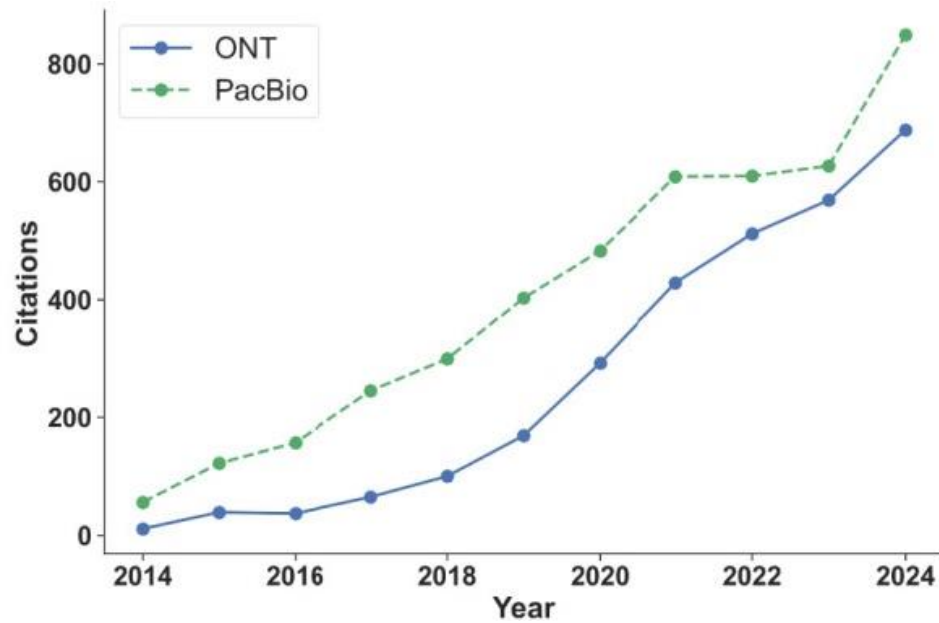
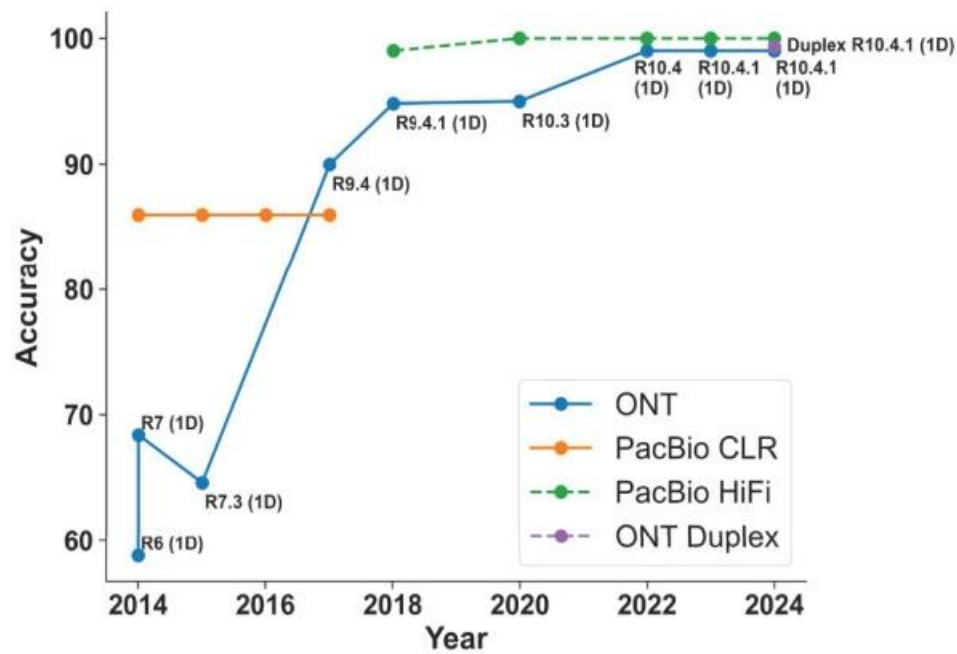
- 1 SMRT cell/run
- 60 Gb per cell



The old paradigm vs Long reads

accurate reads





Overview on technologies

	ONT	PacBio
Read accuracy	~95-99.5%	> 99% (HiFi)
Read length	up to 2 Mb	up to 30-40 kb (HiFi) typically ~15-20 kb
Type of molecule	DNA (cDNA) and RNA	DNA (cDNA)
RNA base modifications	Yes (6mA) ¹	No
DNA base modifications	Yes (5mC, 5hmC, 6mA) ²	Yes (m5C, m6A) ³
Throughput (BIF)	10 Tb ~500M reads/run ⁴	480 Gb ~25M HiFi reads/run ⁵

1. Liu, H., et al (2019). Accurate detection of m6A RNA modifications in native RNA sequences. *Nature Communications*, 10(1), 1–9

2. Liu, Q., et al (2019). Detection of DNA base modifications by deep recurrent neural network on Oxford Nanopore sequencing data. *Nature Communications*, 10(1).

3. Flusberg, B. A., et al (2010). Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nature Methods*, 7(6), 461–465

4. 48 flow cells on a PromethION

5. 4 SMRT cells on a Revio

Overview on technologies

Technology	Platform	Median read length (kb)	Median throughput per run (10^6 transcript reads)
PacBio	cDNA + IsoSeq + Sequel II	~2.1	~2.6
	cDNA + Kinnex + Sequel II	~1.7	~40
	cDNA + Kinnex + Revio	~1.7	~100
ONT	cDNA + MinION (R10.4)	~0.939	~20
	cDNA + PromethION (R9.4)	~1	~130
	dRNA + MinION (R9.4)	~0.8	~1.1
	dRNA + PromethION (R9.4)	~0.6	~20

Question 7&8