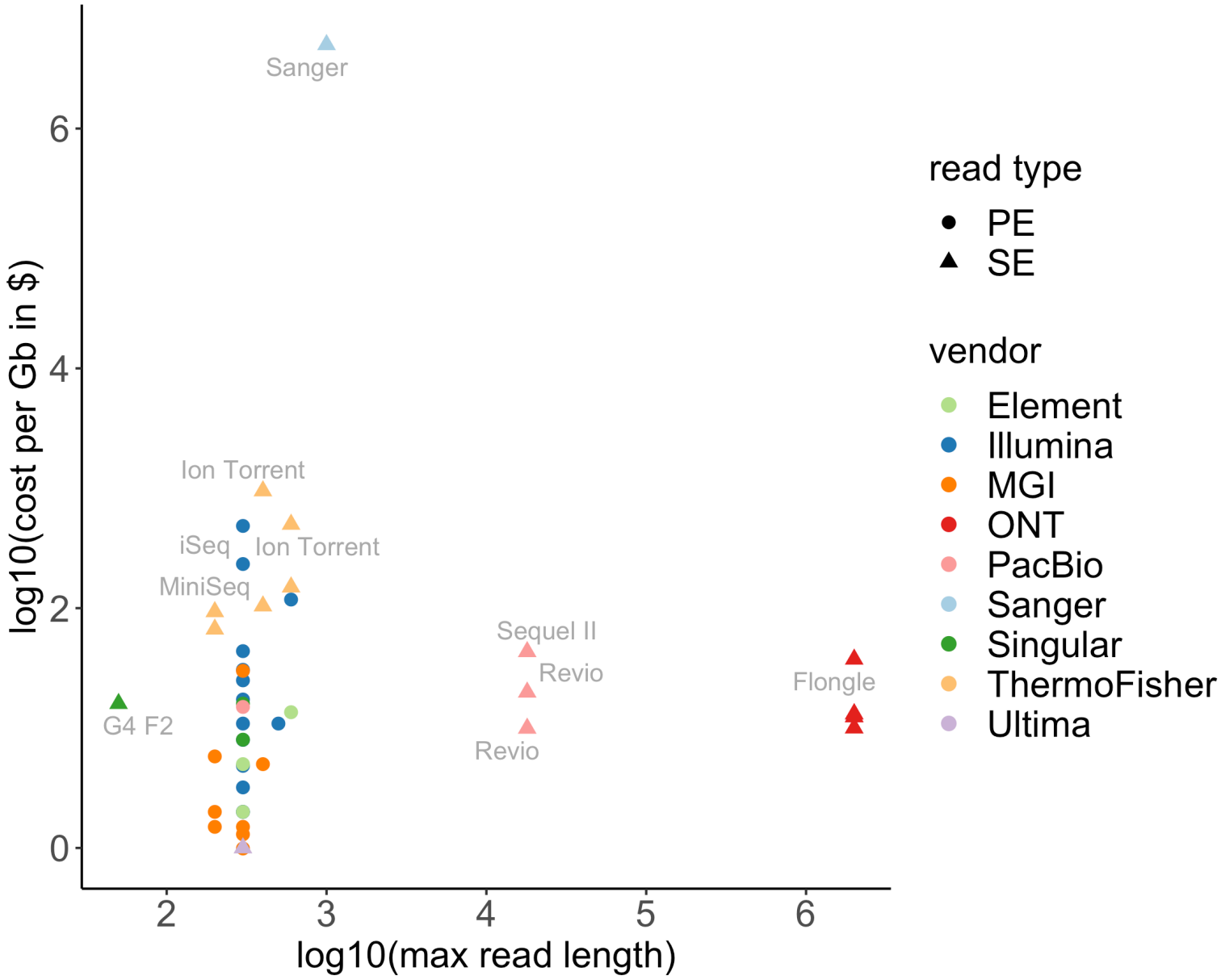
Long-read sequence analysis

Sequencing technologies

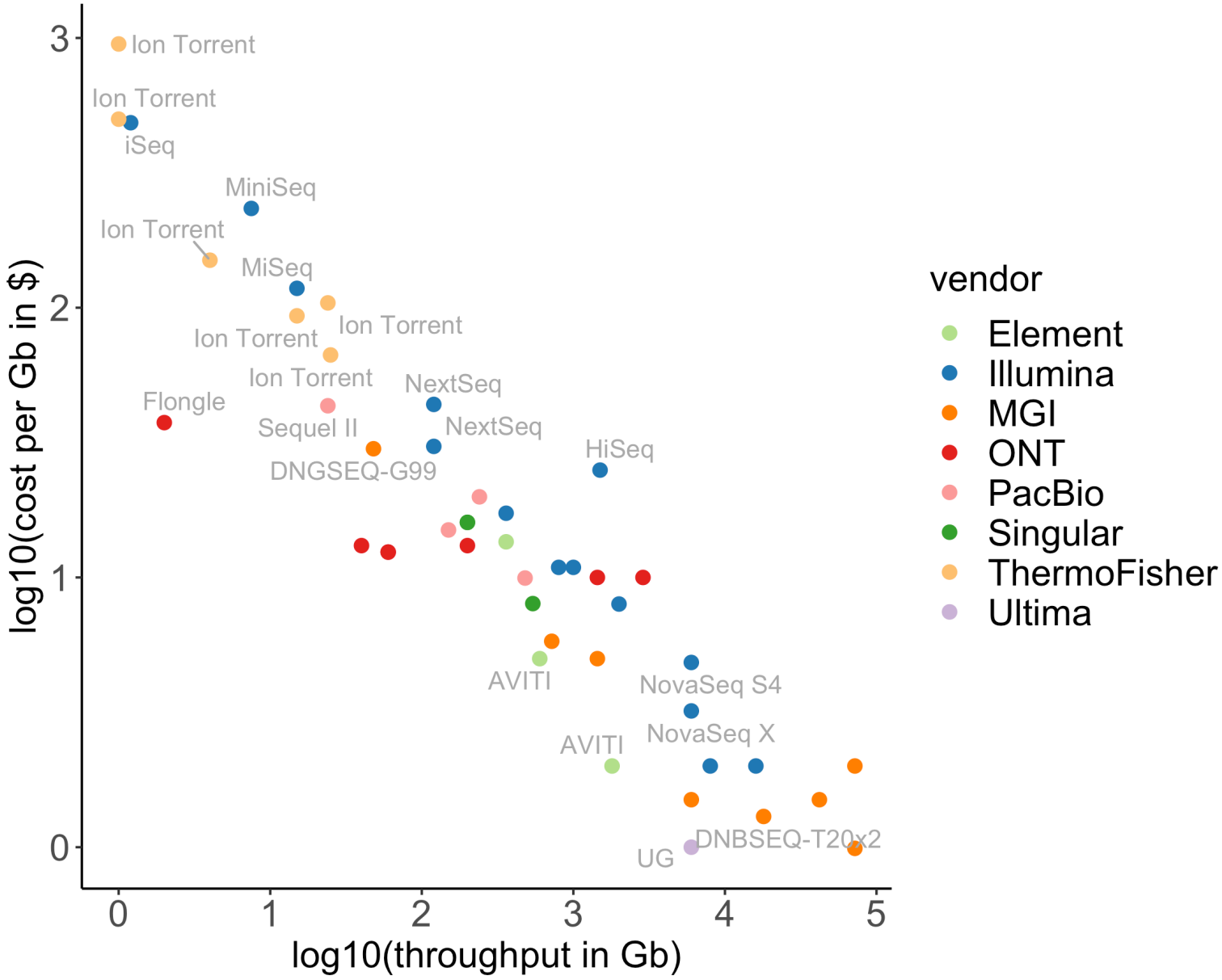
Question 4

What is a long read?

- Short read: 50-300 bp, often paired-end (Illumina sequencing)
- Long read: > 1kb, up to 20 Mb:
  - single molecule sequencing or
  - 3rd generation sequencing



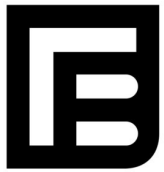
drawn from: https://docs.google.com/spreadsheets/d/1GMMfhyLK0-q8XkIo3YxlWaZA5vVMuhU1kg41g4xLkXc/ Albert Vilella



drawn from: https://docs.google.com/spreadsheets/d/1GMMfhyLK0-q8XkIo3YxlWaZA5vVMuhU1kg41g4xLkXc/ Albert Vilella

Sequencing-by-synthesis

- 2nd generation sequencing
- Massive throughput: up to $500x10^9$ bases/run
- Illumina still most used platform today



Element Biosciences

illumina®

SINGULAR GENOMICS

Sequencing-by-synthesis

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

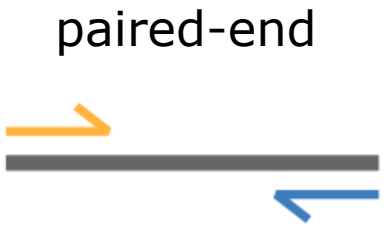
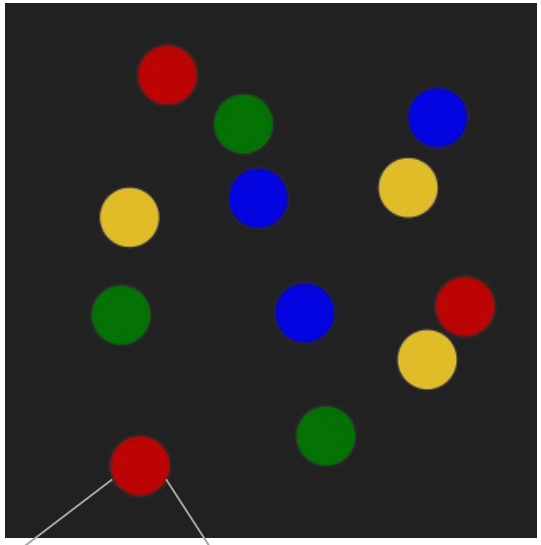


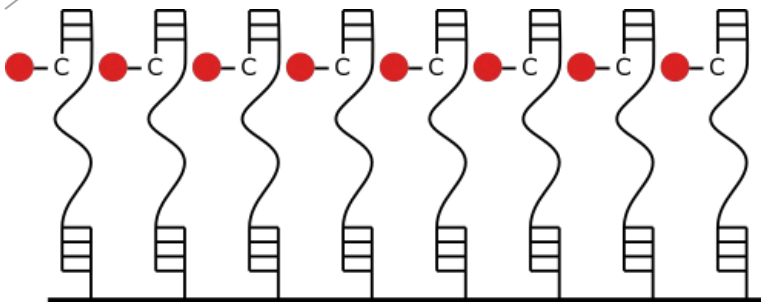
Image from: Illumina (2020)

Question 5

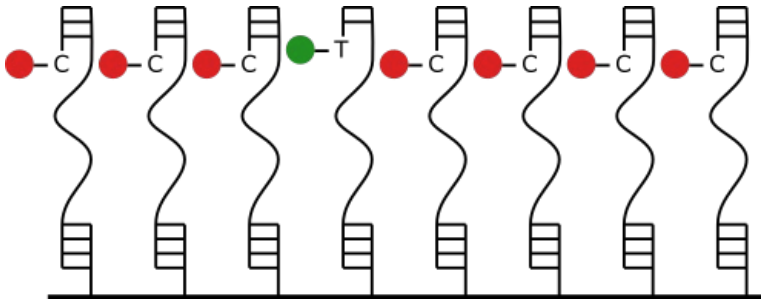
SBS - limitations

- 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

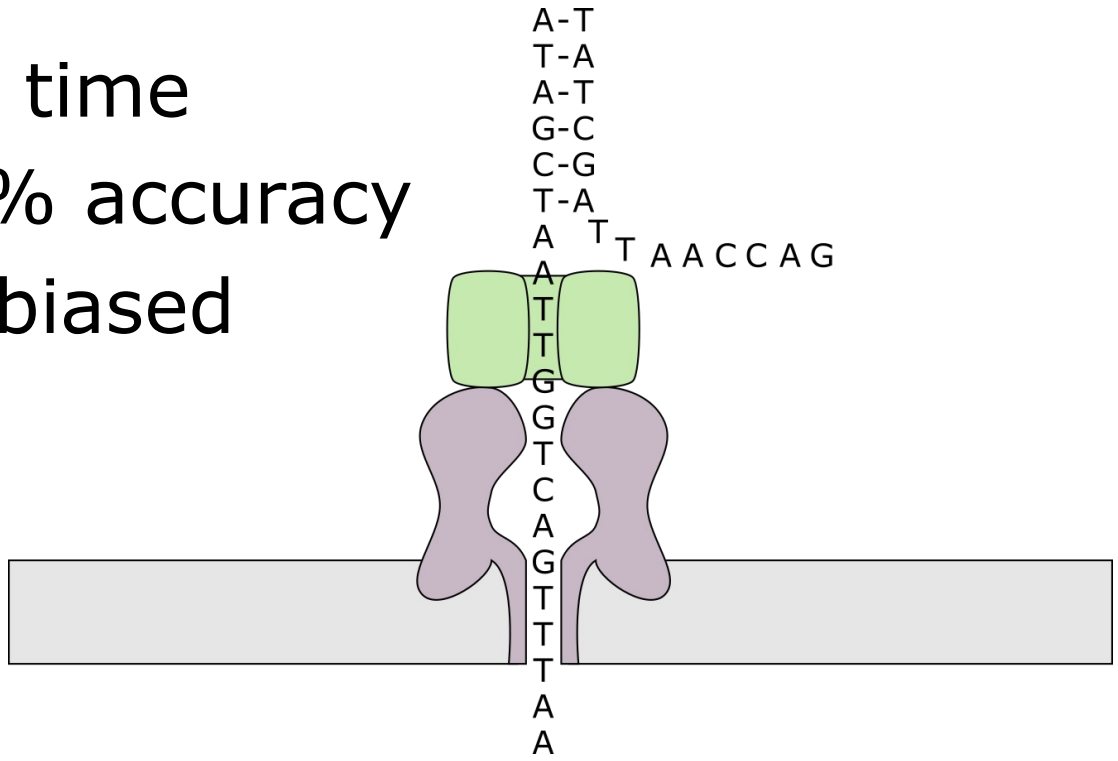
PacBio

Oxford NANOPORE Technologies

Question 6

Oxford Nanopore technology

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



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

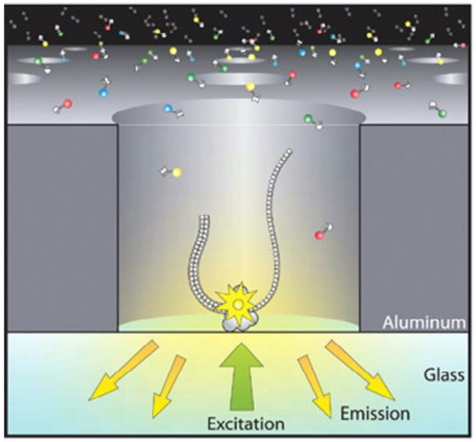
PromethION

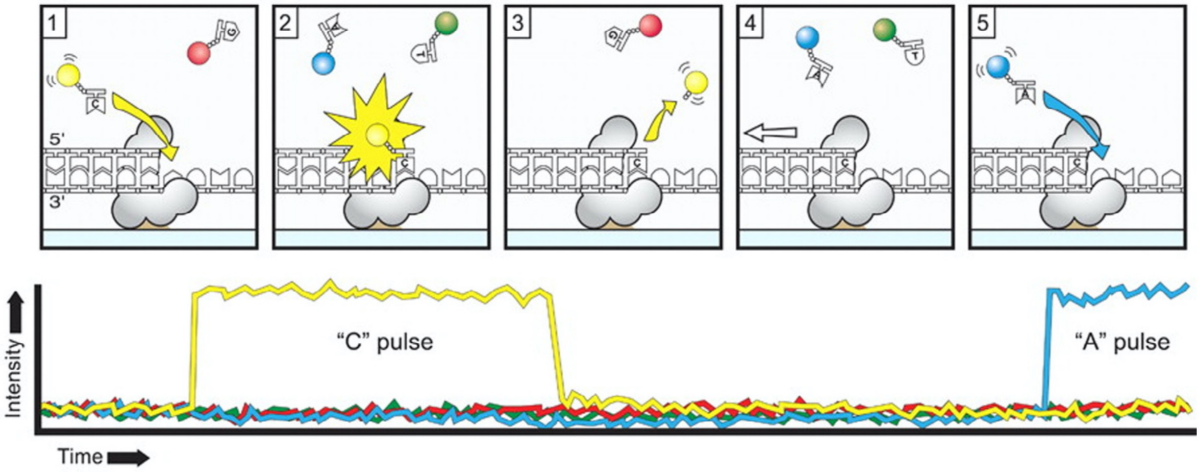
https://nanoporetech.com/products#comparison

ONT library prep

- Standard kit:
  - >1 μg HMW DNA
  - Shearing + size selection is optional
  - Multiplexing requires PCR step

PacBio sequencing





- Polymerase bound to ZMW bottom
- Circular molecules
- Single read out ~90% accuracy
- HiFi: single molecule sequenced multiple times

Image from: Rhoads A, Au KF. Genomics Proteomics Bioinformatics 2015;13:278–89

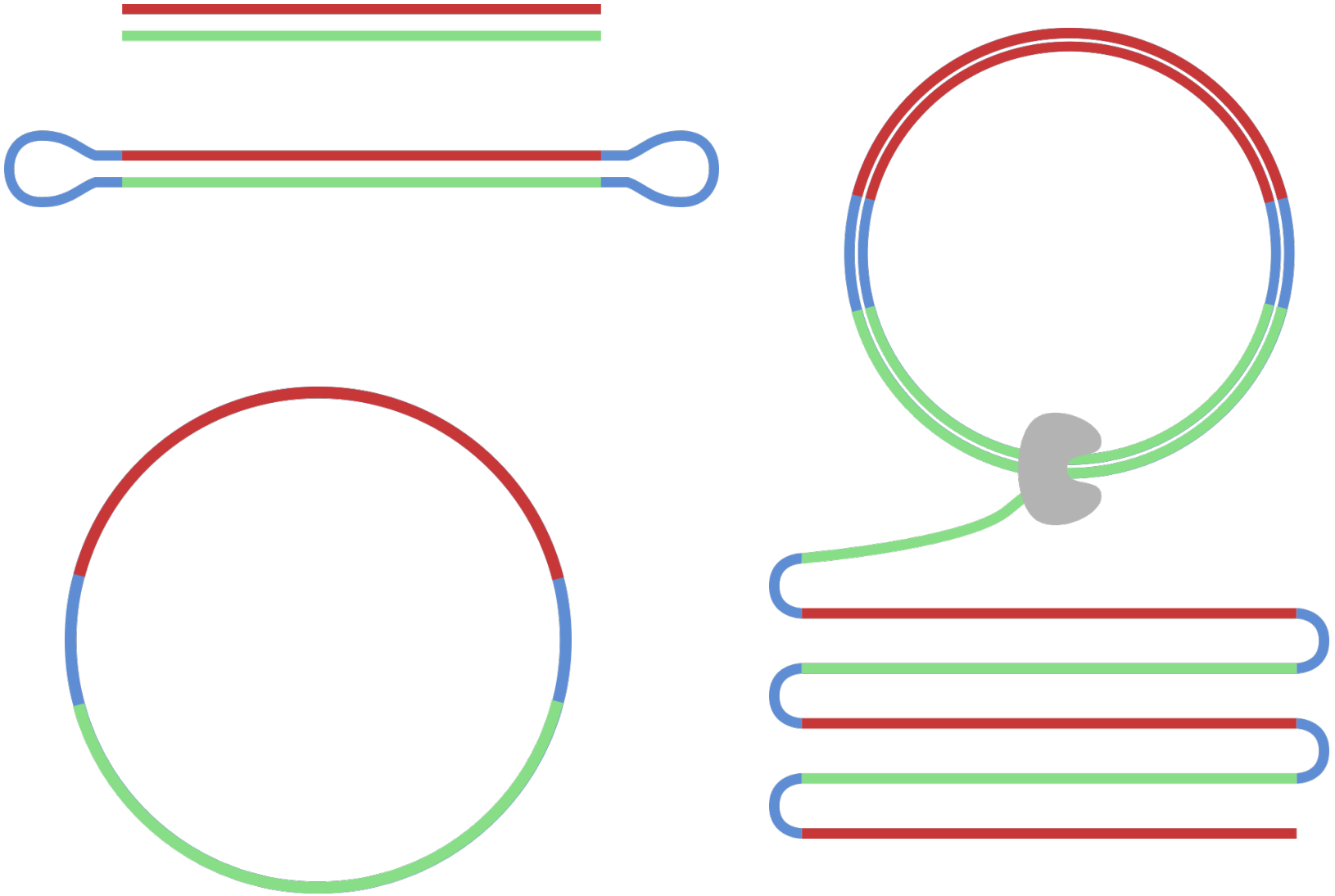
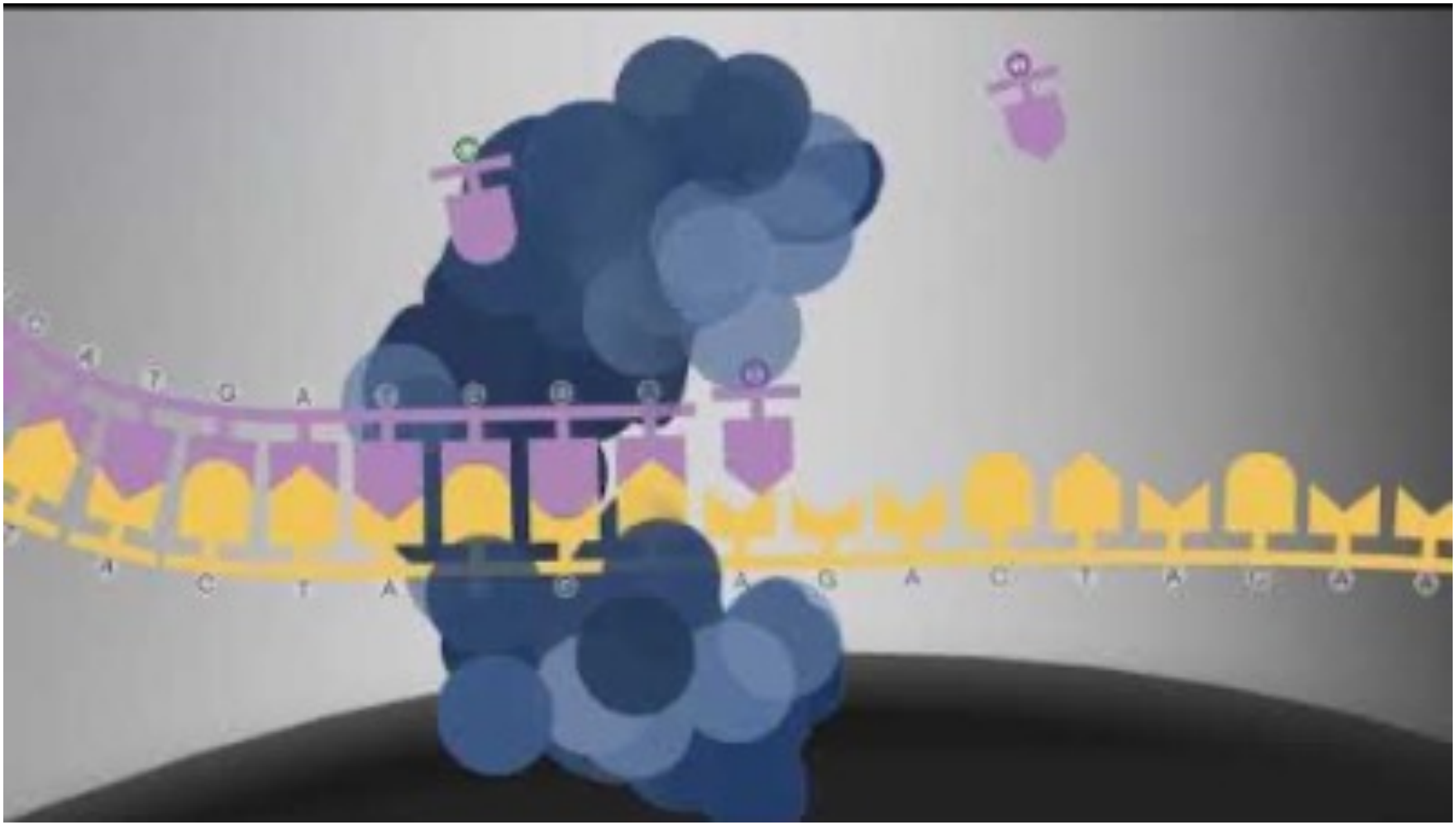


Image from: https://doi.org/10.5281/zenodo.4636860



PacBio Sequel IIe

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

Pacbio Revio

- 25M ZMW
- ~5-6M HiFi reads/SMRT cell
- 2-4 SMRT cells/run

PacBio library prep

- start with >5 μg HMW DNA
- Requires shearing + size selection
- Multiplexing requires PCR

	ONT	**PacBio**
Read accuracy	~90-95%	> 99.99% (HiFi)
Read length	up to 2 Mb	up to 30-40 kb (HiFi) typically ~15-20 kb
RNA base modifications	Yes (m6A)[1]	No
DNA base modifications	Yes (m5C, m6A)[2]	Yes (m5C, m6A, hm5C)[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

Question 7&8