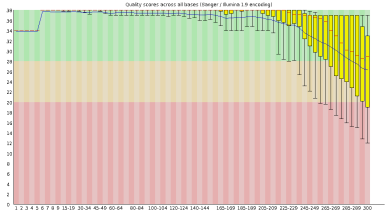
NGS - quality control, alignment, visualisation

Sequencing technologies

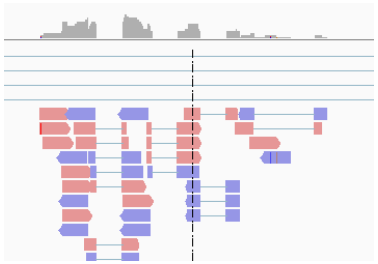
Major applications

- Transcriptome characterization
  - e.g. RNA-seq
- Epigenome characterization:
  - e.g. ATAC-seq
- DNA-protein interactions:
  - e.g. ChIP-seq
- Whole genome (assembly)
- Variant detection
- Metagenome characterization
- Any others?

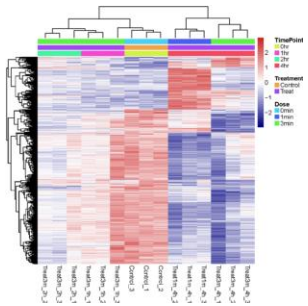
Sequencing



Quality control



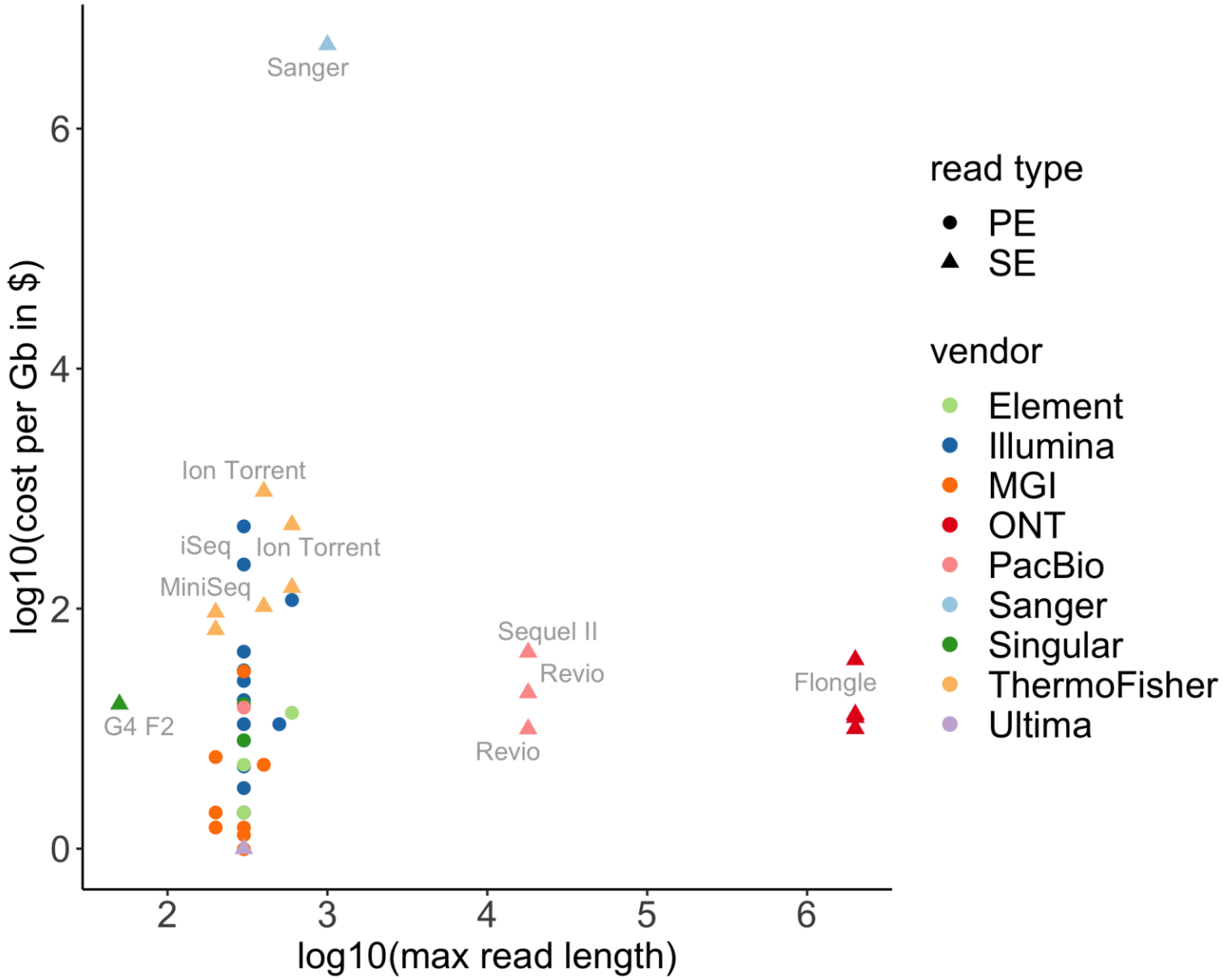
Alignment



Down-stream analysis → Visualisation

How to choose your sequencing method?

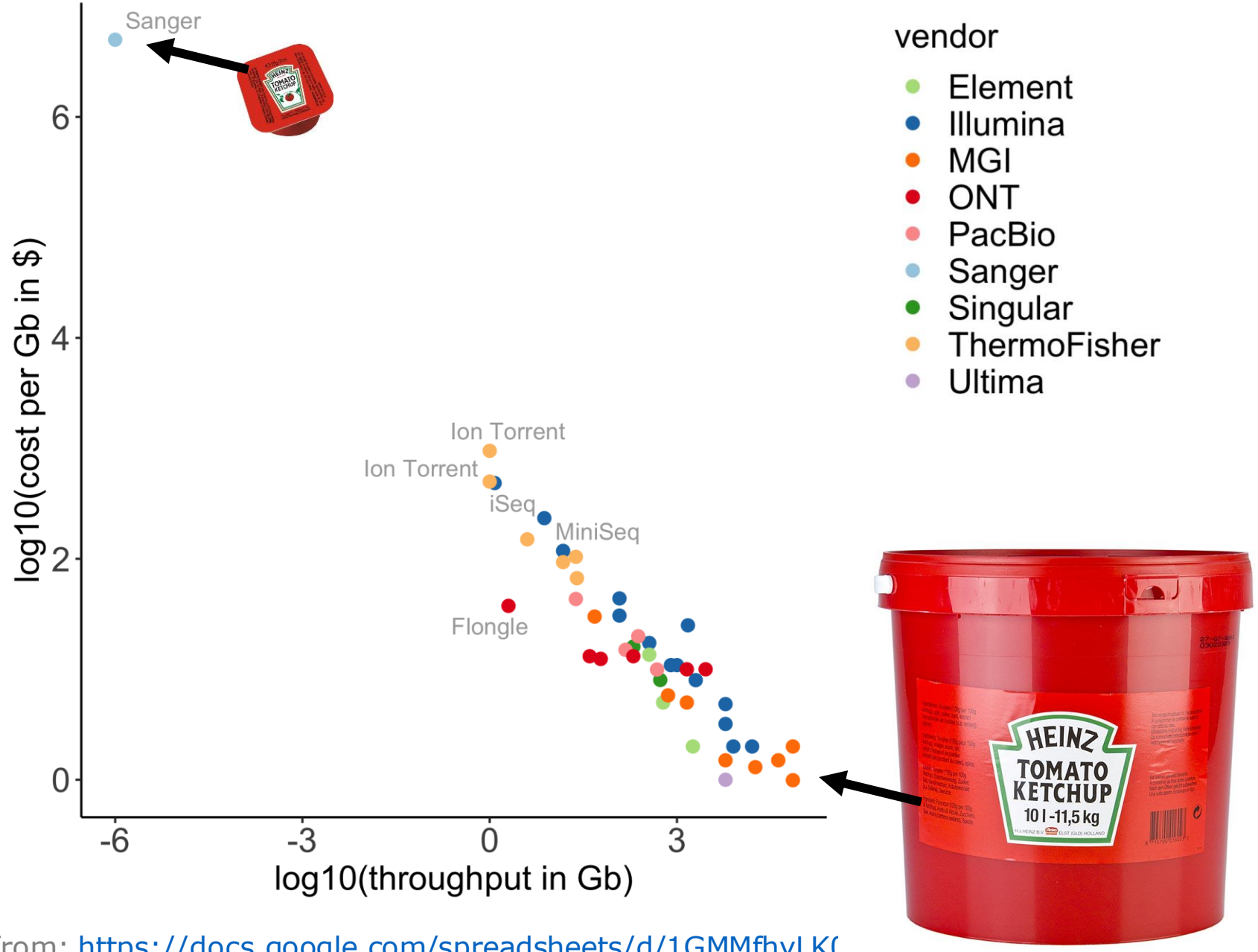
- Read length
- Accuracy
- Availability
- Costs
- Throughput



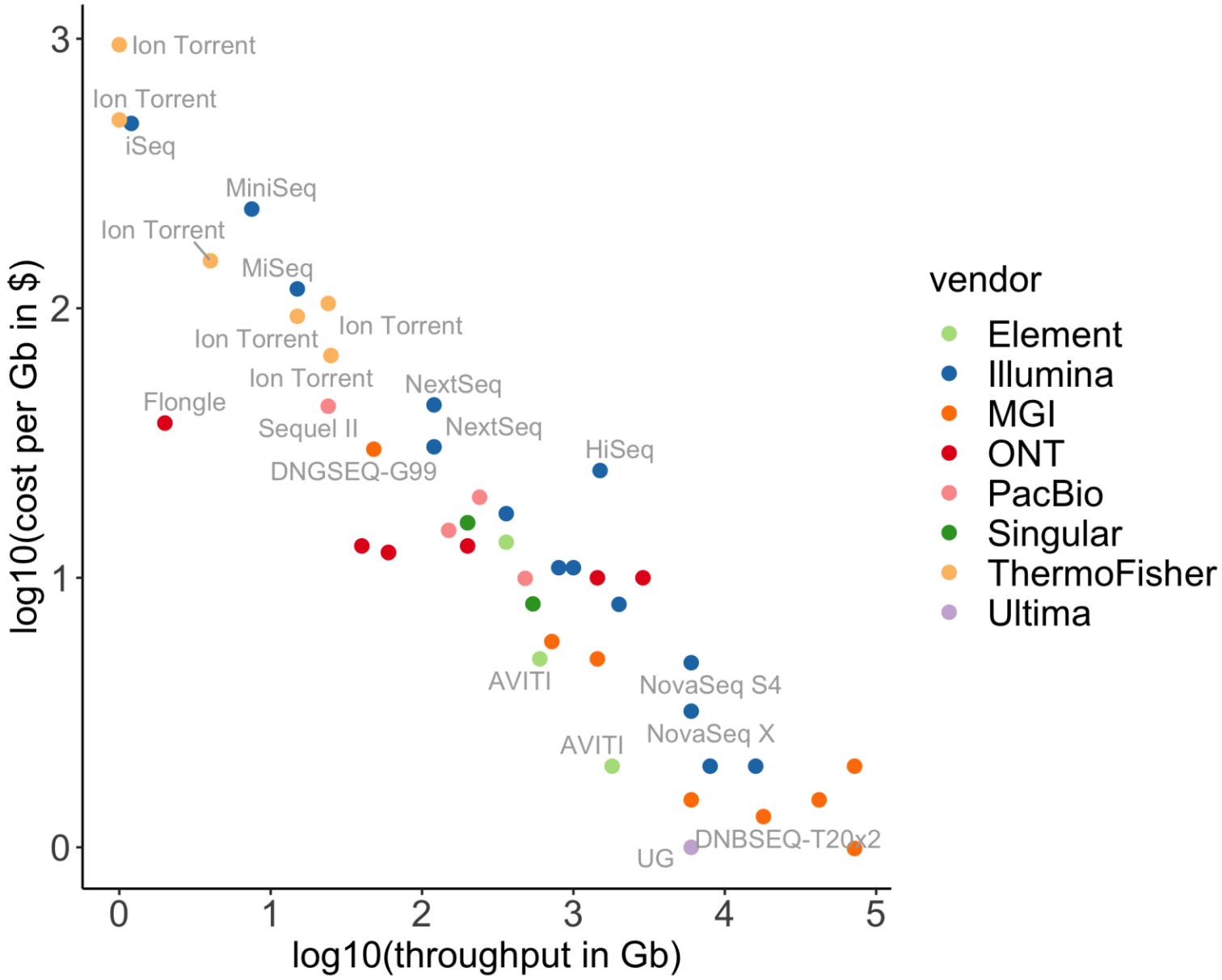
drawn from: https://docs.google.com/spreadsheets/d/1GMMfhyLK0-q8XkIo3YxlWaZA5vVMuhU1kg41g4xLkXc/ Albert Vilella

Quiz Question 4


HEINZ
TOMATO
KETCHUP
10 l -11,5 kg
H.J.HEINZ B.V. ELST (GLD) HOLLAND



drawn from: https://docs.google.com/spreadsheets/d/1GMMfhyLK(_q8XkIo3YxlWaZA5vVMuhU1kg41g4xLkXc/ Albert Vilella

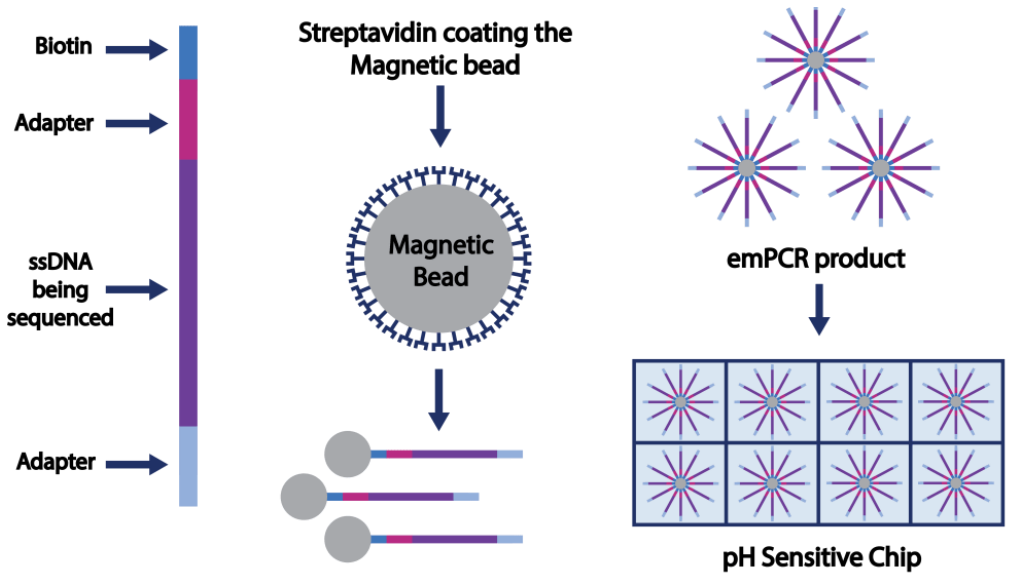


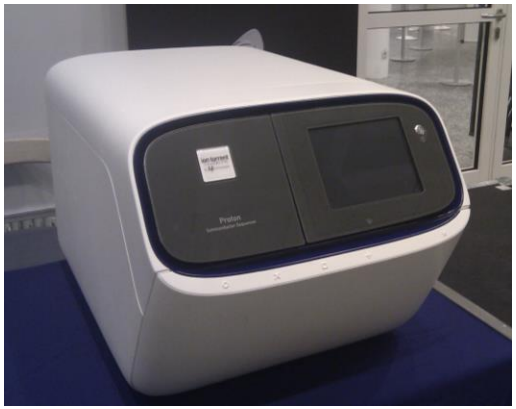
drawn from: https://docs.google.com/spreadsheets/d/1GMMfhyLK0-q8XkIo3YxlWaZA5vVMuhU1kg41g4xLkXc/ Albert Vilella

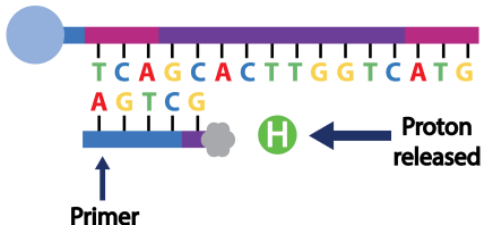
This course

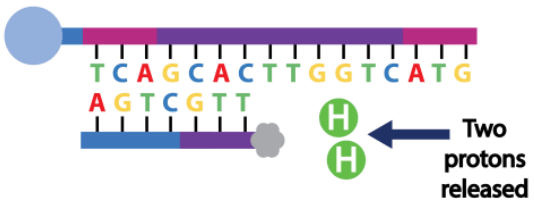
- 2nd generation (sequencing by synthesis):
  - Ion Torrent
  - Illumina
- 3rd generation:
  - Pacific Biosciences
  - Oxford Nanopore Technology

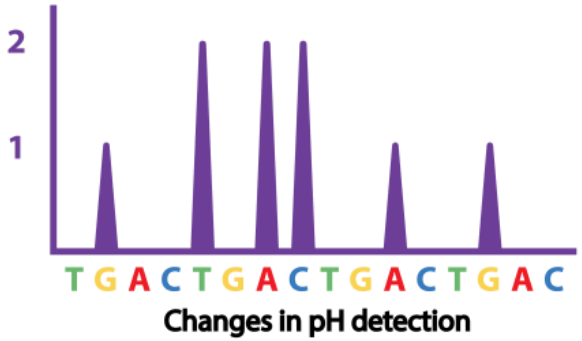
Ion Torrent sequencing





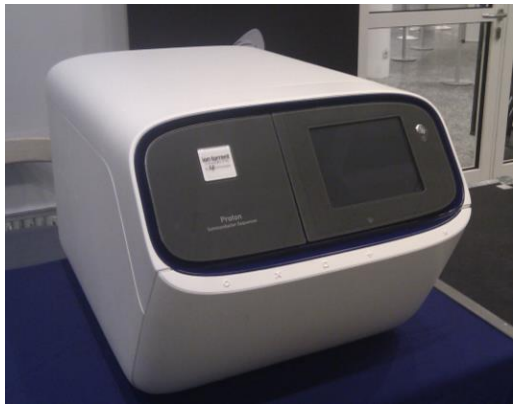






Ion Torrent sequencing

- Up to ± 400 bp read length
- Scalable (but Illumina has machines with similar thoughput nowadays)
- Homopolymers (e.g. TTTTT) are a challenge (impossible) to sequence

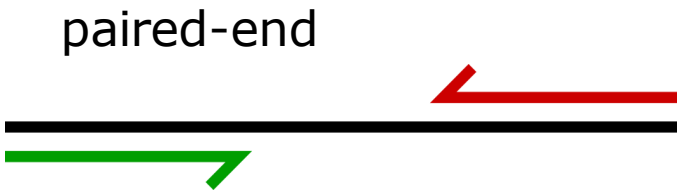


Illumina sequencing

- Massive throughput: up to $16x10^{12}$ bases/run (NovaSeq X) = ~9,000 whole exomes
- Most used platform today

Illumina sequencing

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

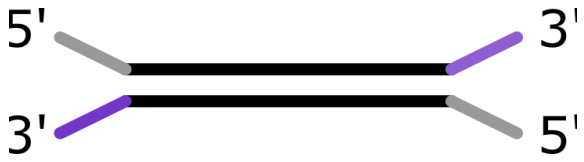


Illumina libray prep

shear + size select DNA

↓

Ligate adapters



↓

Barcode + p5/p7 sites

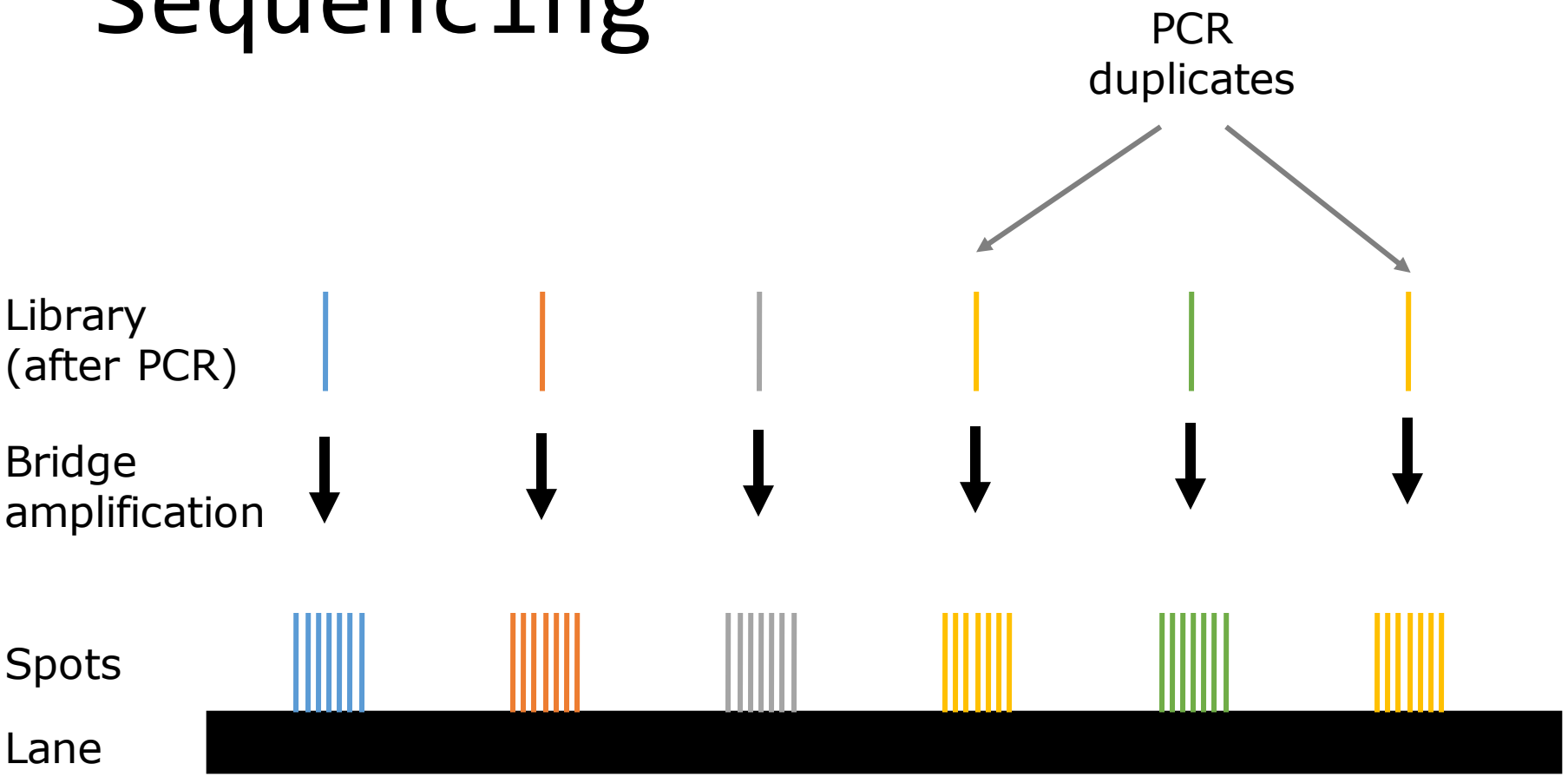
↓

PCR: 8-16 cycles

↓

Sequencing

Sequencing



Each spot represents one read pair

Sequencing by synthesis

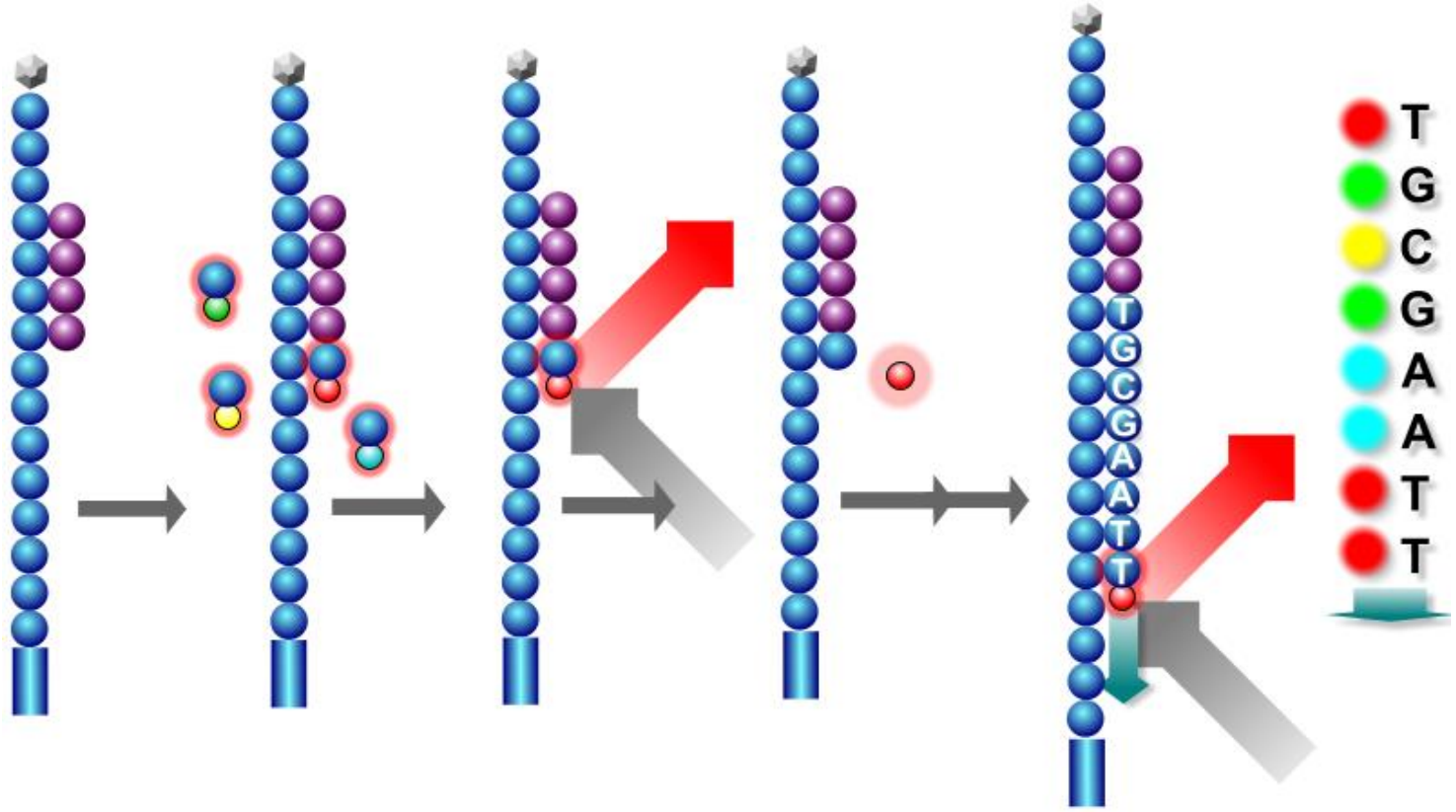
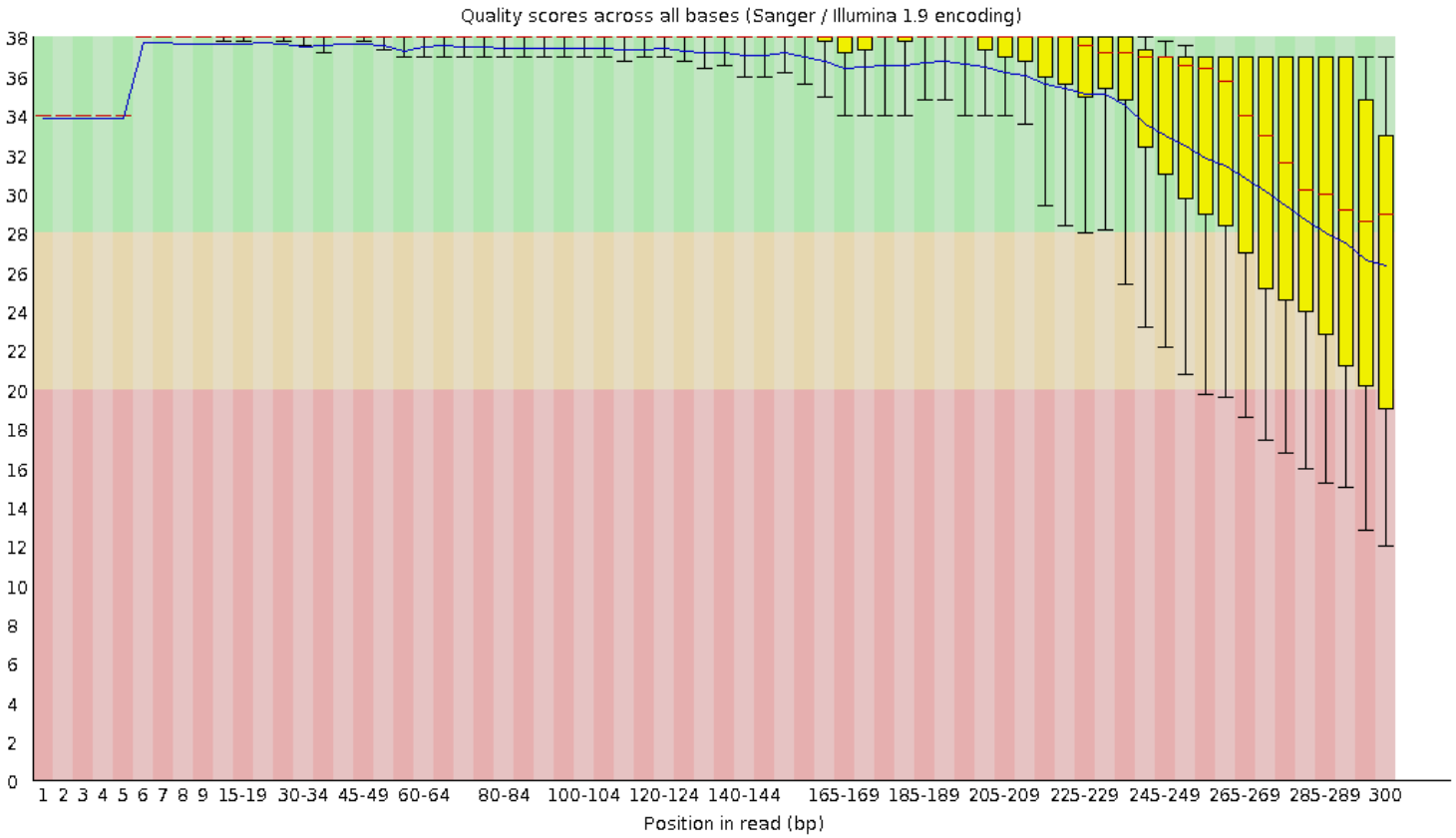
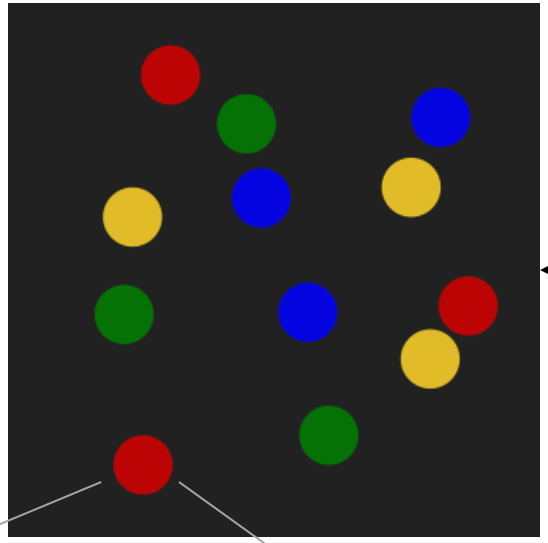


Image: Abizar Lakdawalla CC BY-SA 3.0

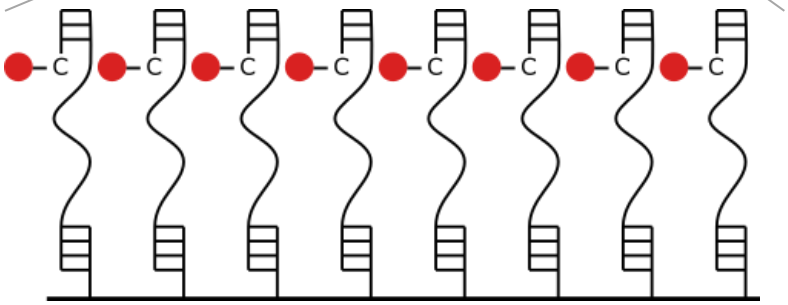
Illumina - limitations

Sequence quality declines towards the end

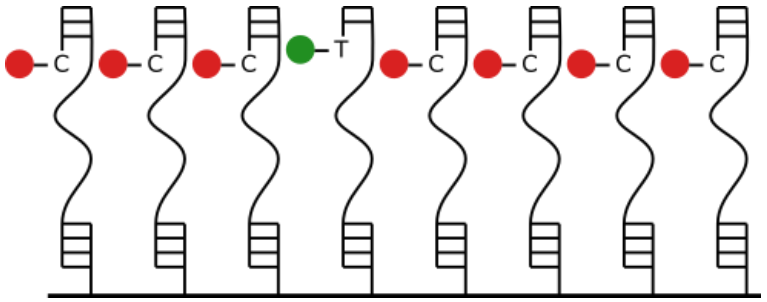




View from above

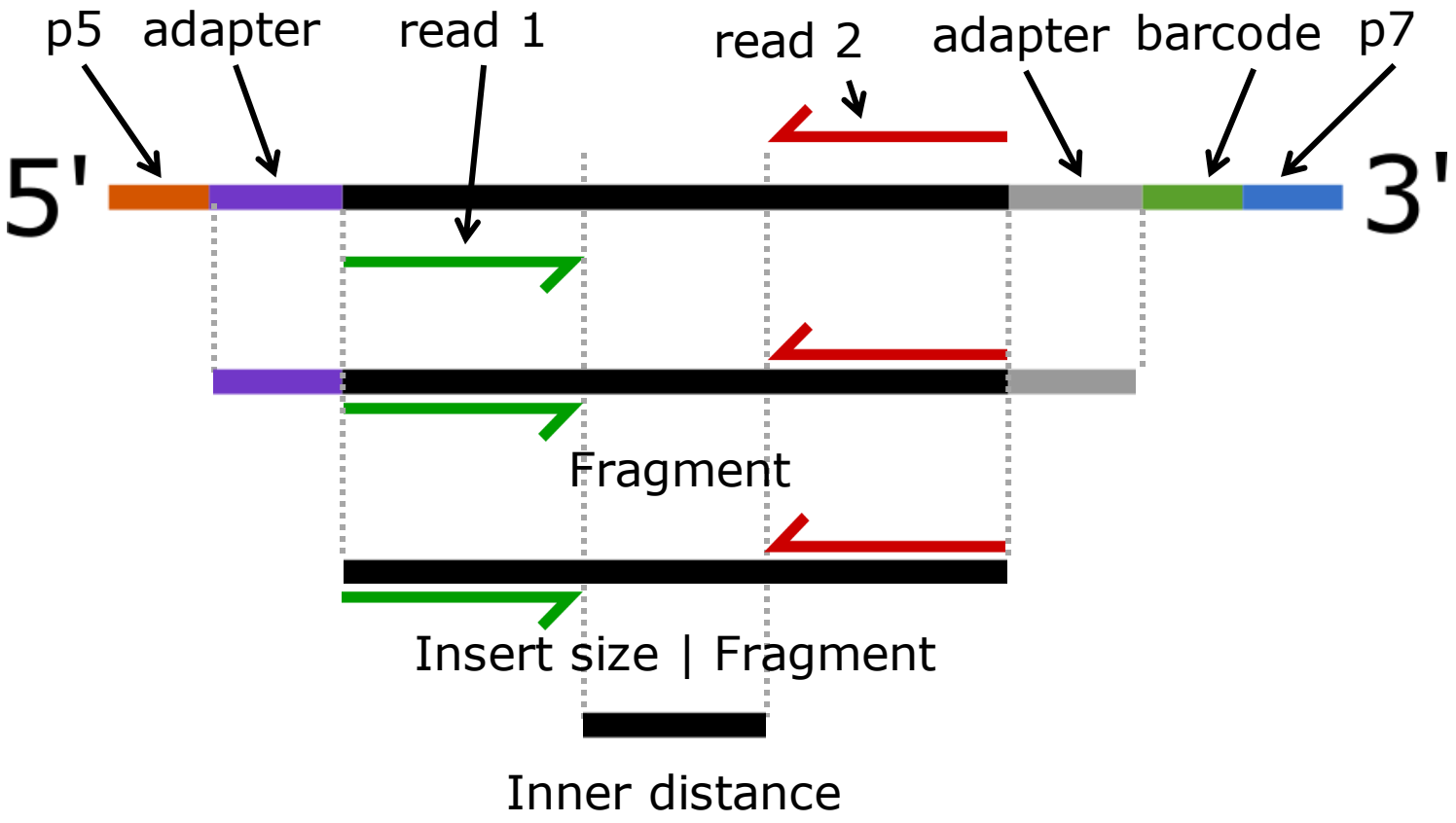


in phase



out of phase

Some definitions



Some more definitions..

- **Library:** fragments from one (c)DNA sample that share a barcode
- **Sequencing run:** complete cycle of generating reads on a machine
- **Flow cell:** physical platform where sequencing reactions take place. Used once in a sequencing run.
- **Lane:** compartment within the flow cell. An Illumina flow cell often has multiple lanes (2 or 4)

Quiz Question 5

Illumina - limitations

- Maximum read length: 300 bp
- Read length is limited by out-of-phase signal
- How to reconstruct:
  - Repeats?
  - Isoforms?
  - Structural variation?
  - Haplotypes?
  - Genomes?

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

Oxford Nanopore technology

- Based on changes in electrical current
- Well-known for its scalability and portability
- ~95-97% accuracy

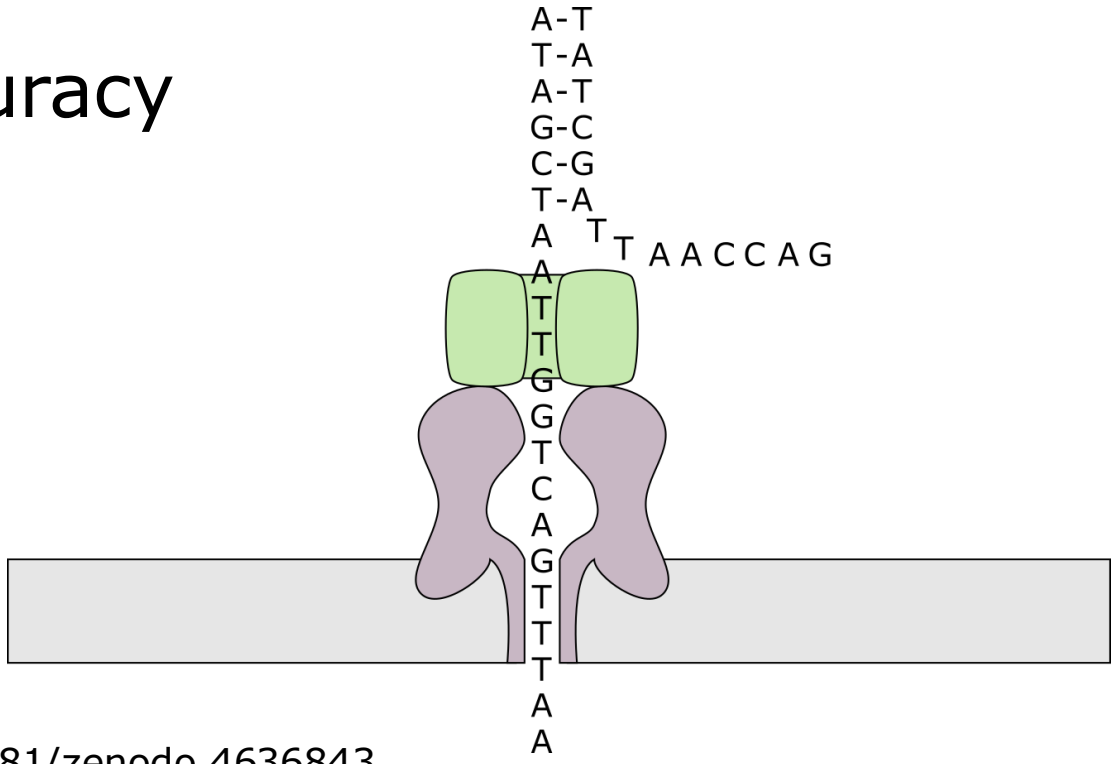
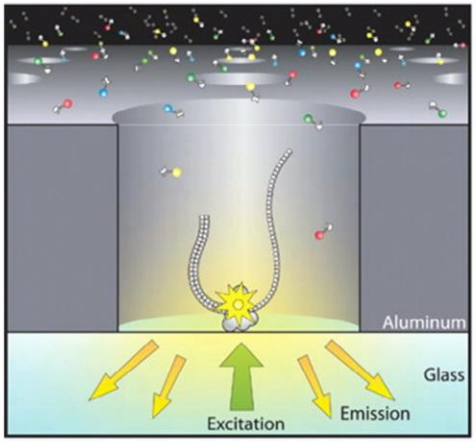
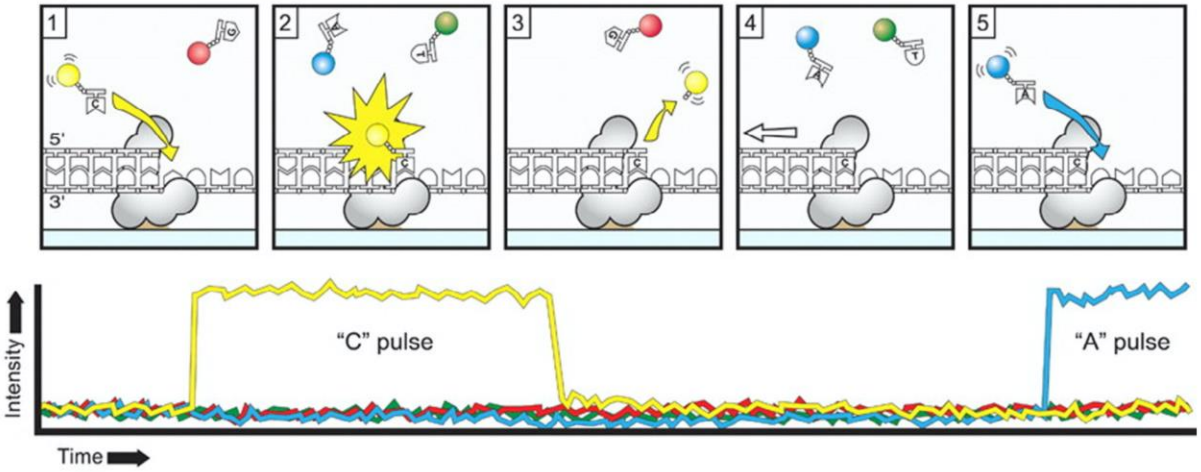


Image from: https://doi.org/10.5281/zenodo.4636843

PacBio sequencing





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

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

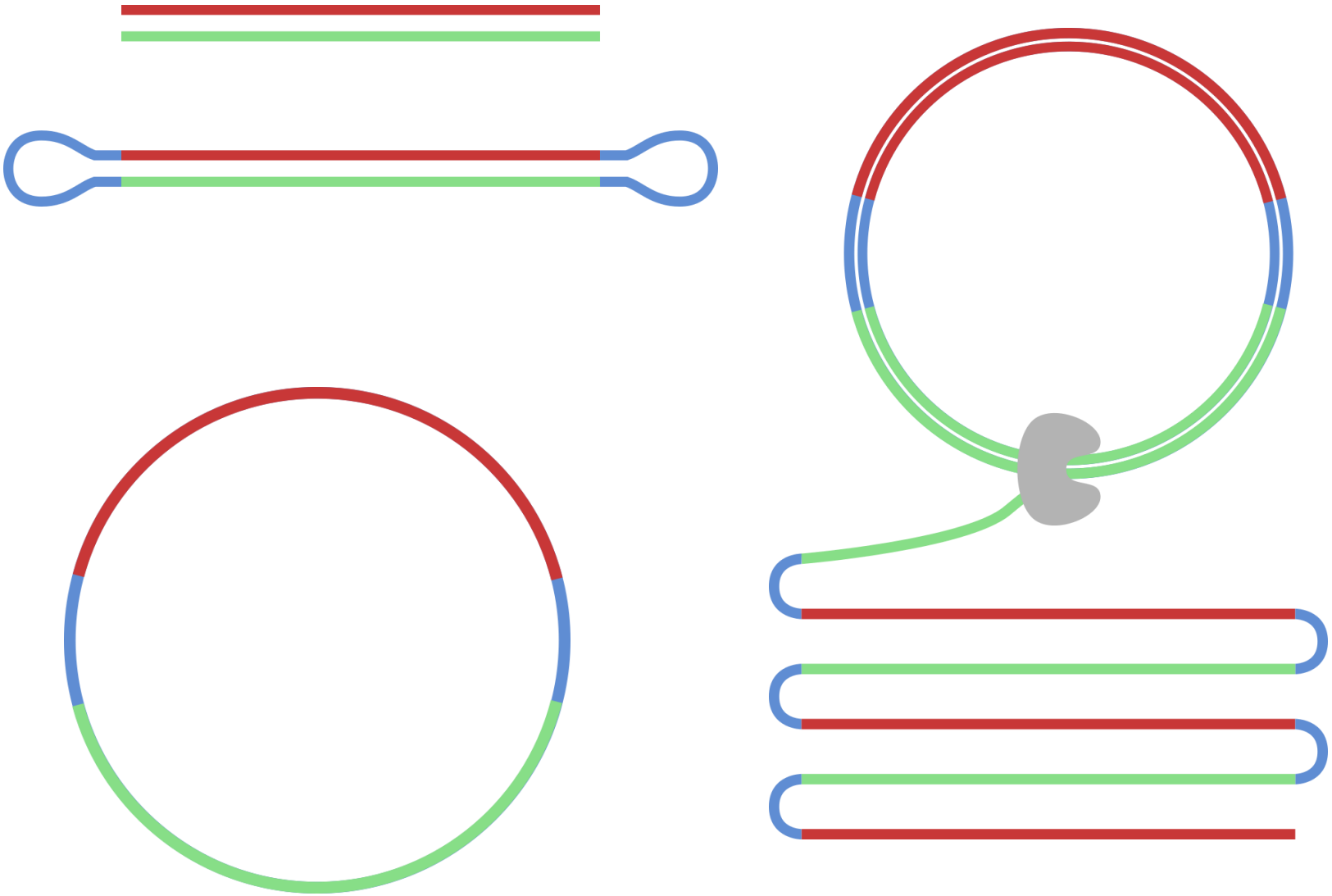


Image from: https://doi.org/10.5281/zenodo.4636860

Quiz Question 6 and 7