Long-read sequence analysis

Sequencing technologies

What is a long read?

- Short read: 50-300 bp, often paired-end
- Long read: > 1kb, up to 20 Mb

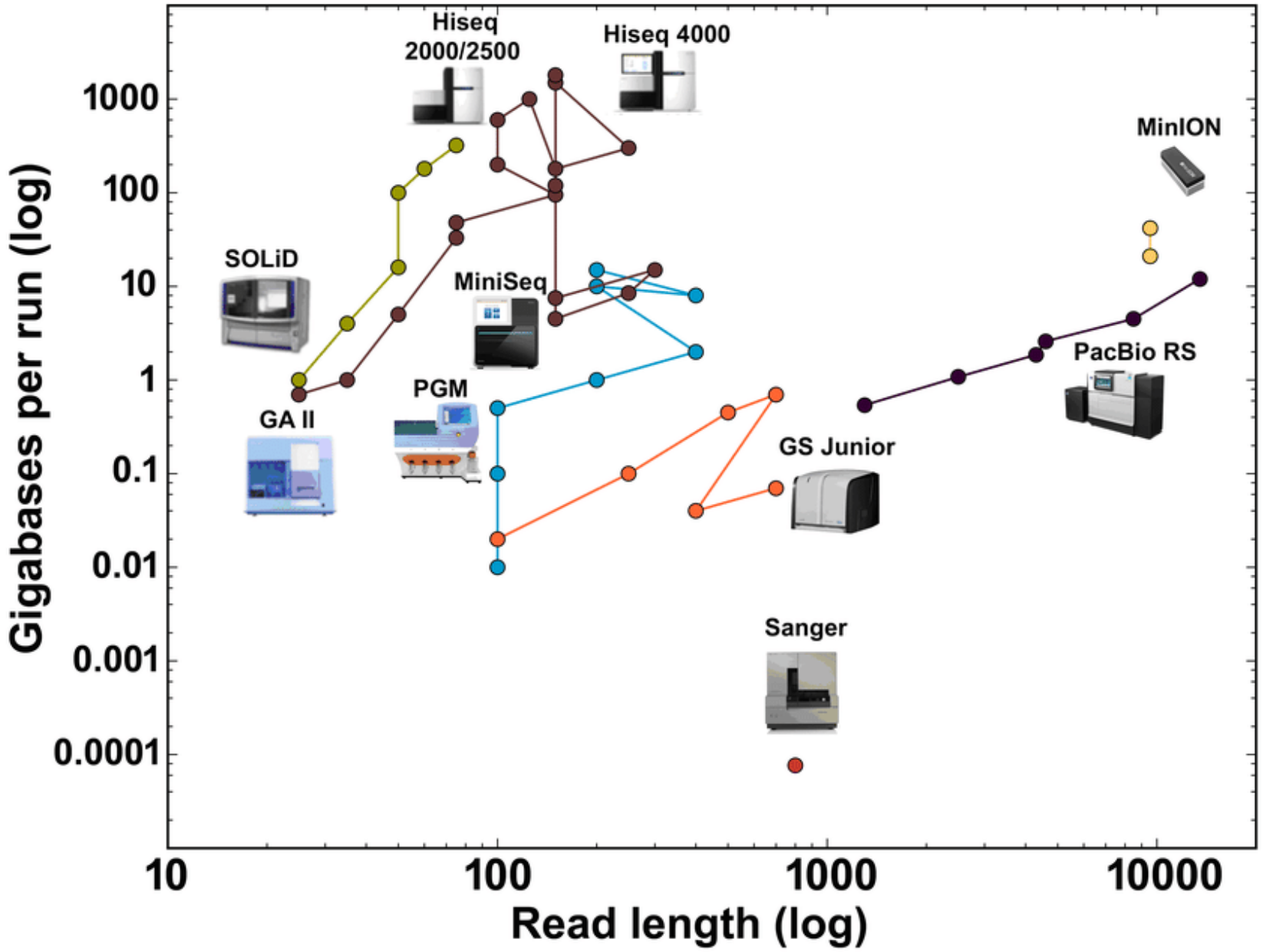


Image from: G. Silva (2016)

Illumina sequencing

- Sequencing-by-synthesis: 2nd generation sequencing
- Massive throughput: up to $500x10^9$ bases/run
- Most used platform today

illumina®

Illumina sequencing

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

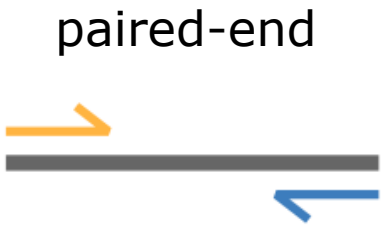
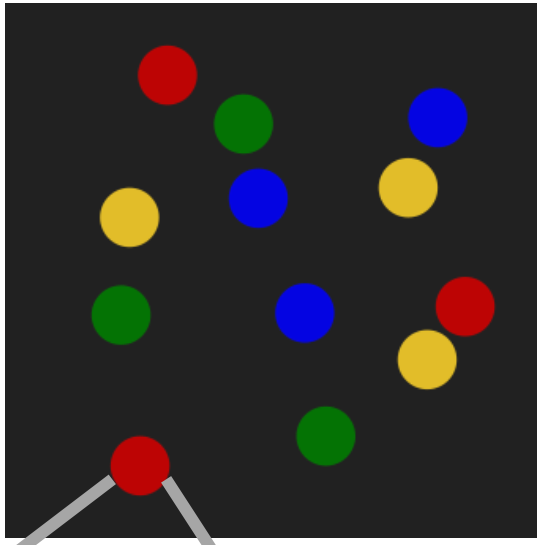
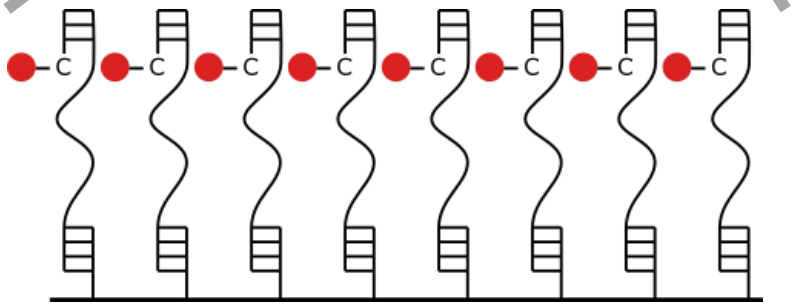


Image from: Illumina (2020)

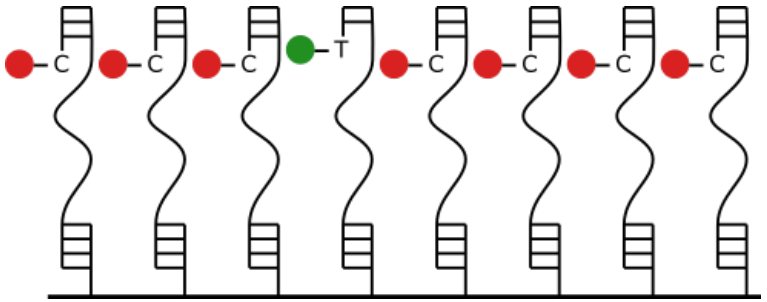
Illumina - limitations

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





in phase



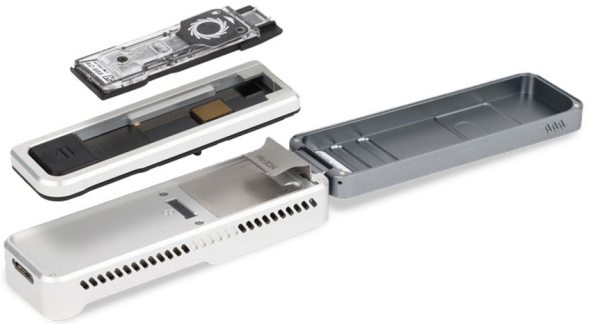
out of phase

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

Oxford Nanopore technology

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



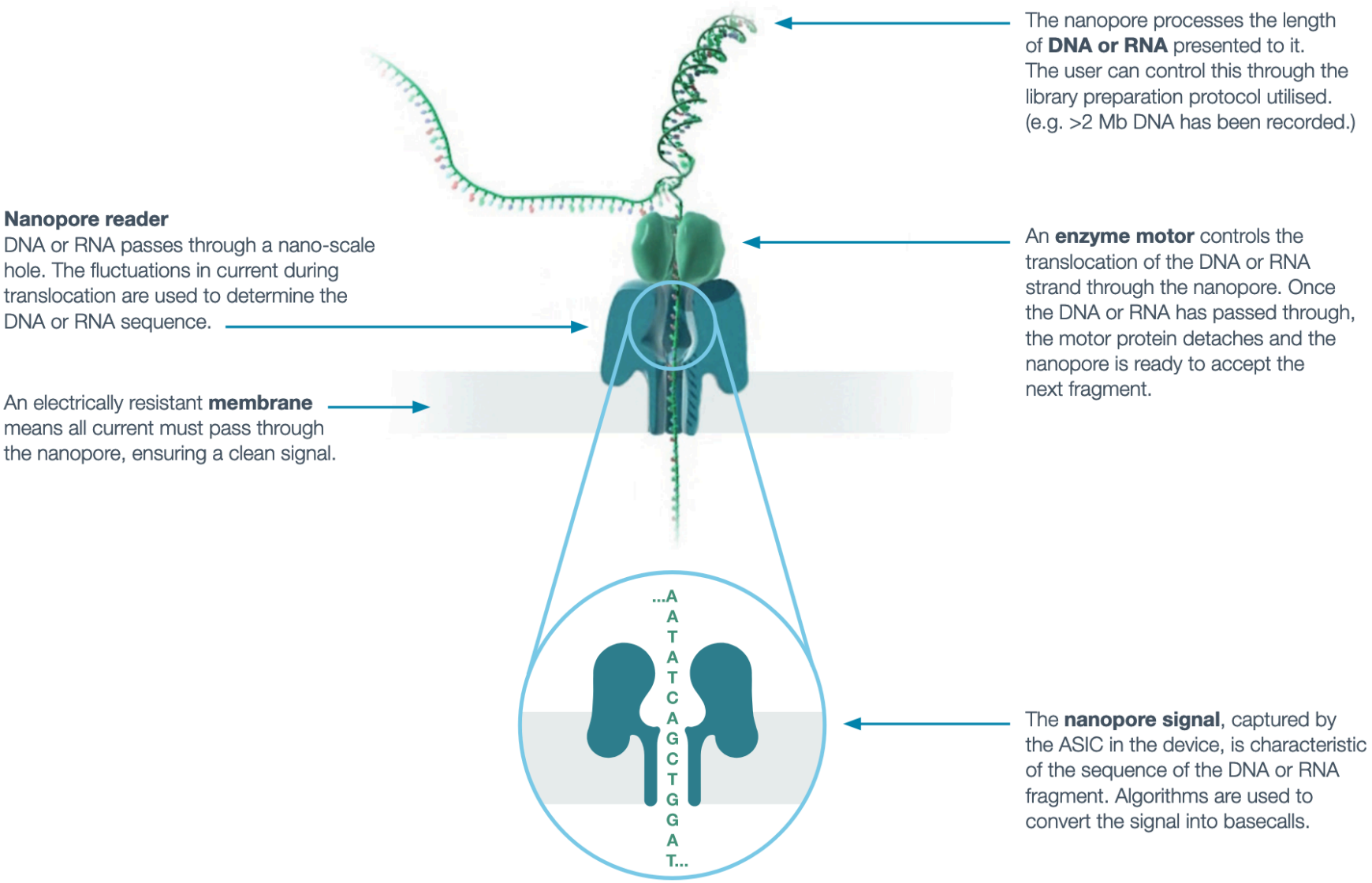
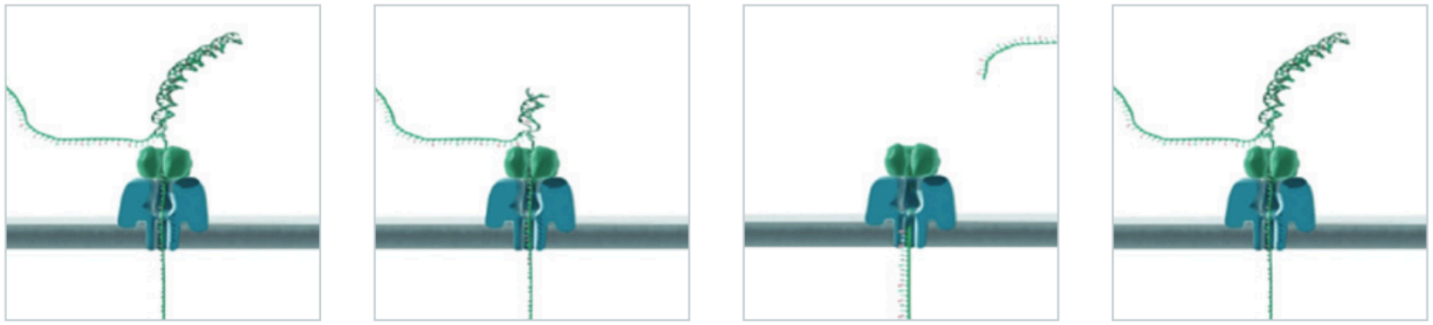


image source: https://nanoporetech.com

1D



$1D^2$

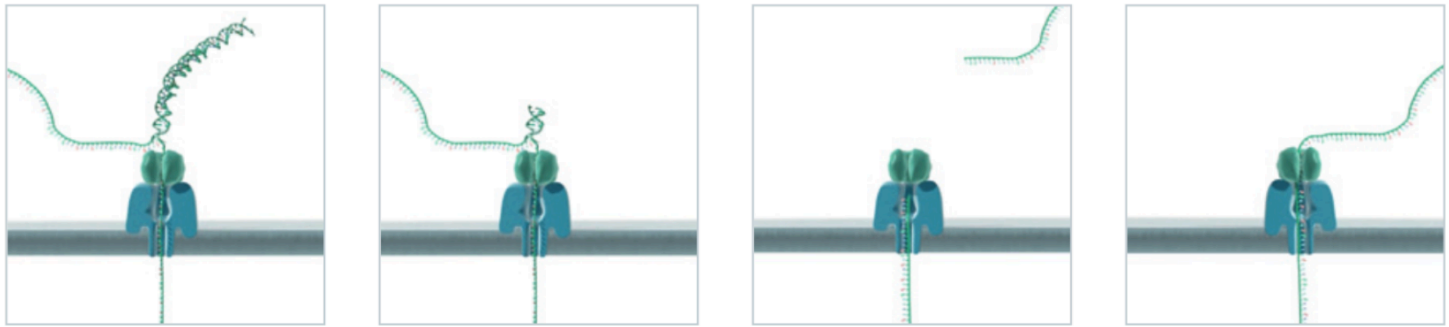
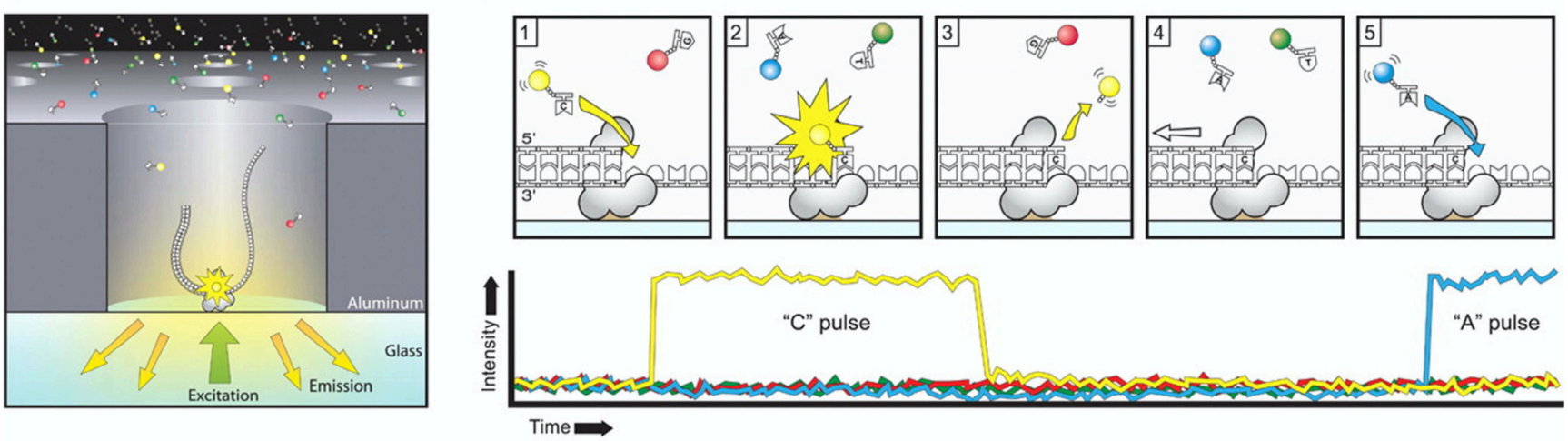


image source: https://nanoporetech.com

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. (2015). Genomics, Proteomics Bioinformatics 2015;13:278–89

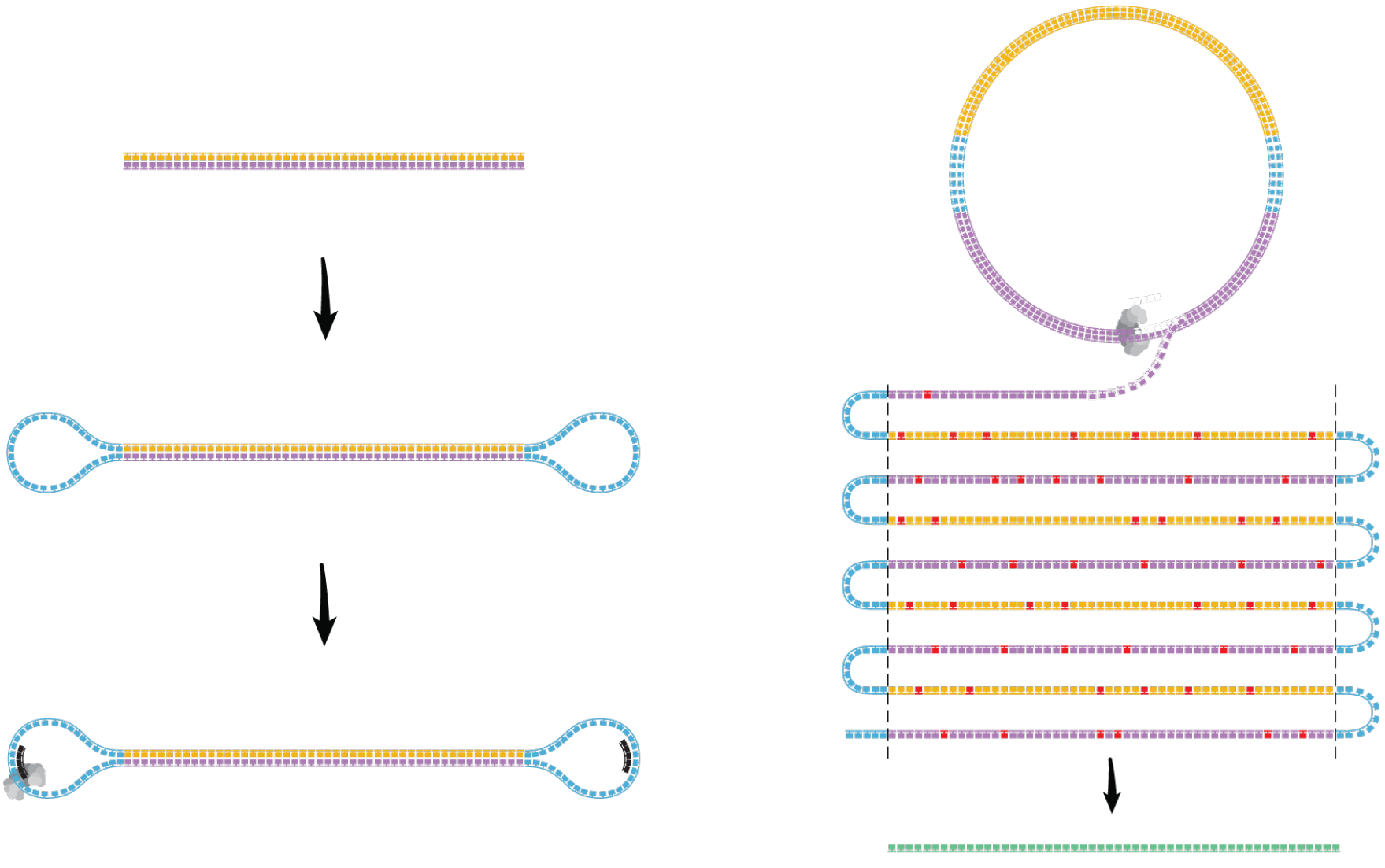


image from: https://www.pacb.com/smrt-science/smrt-sequencing/hifi-reads-for-highly-accurate-long-read-sequencing/