Microbiome Center Bootcamp: Sequencing Technology

Dr. Darrell Cockburn Assistant Professor of Food Science

Why does sequencing technology matter?

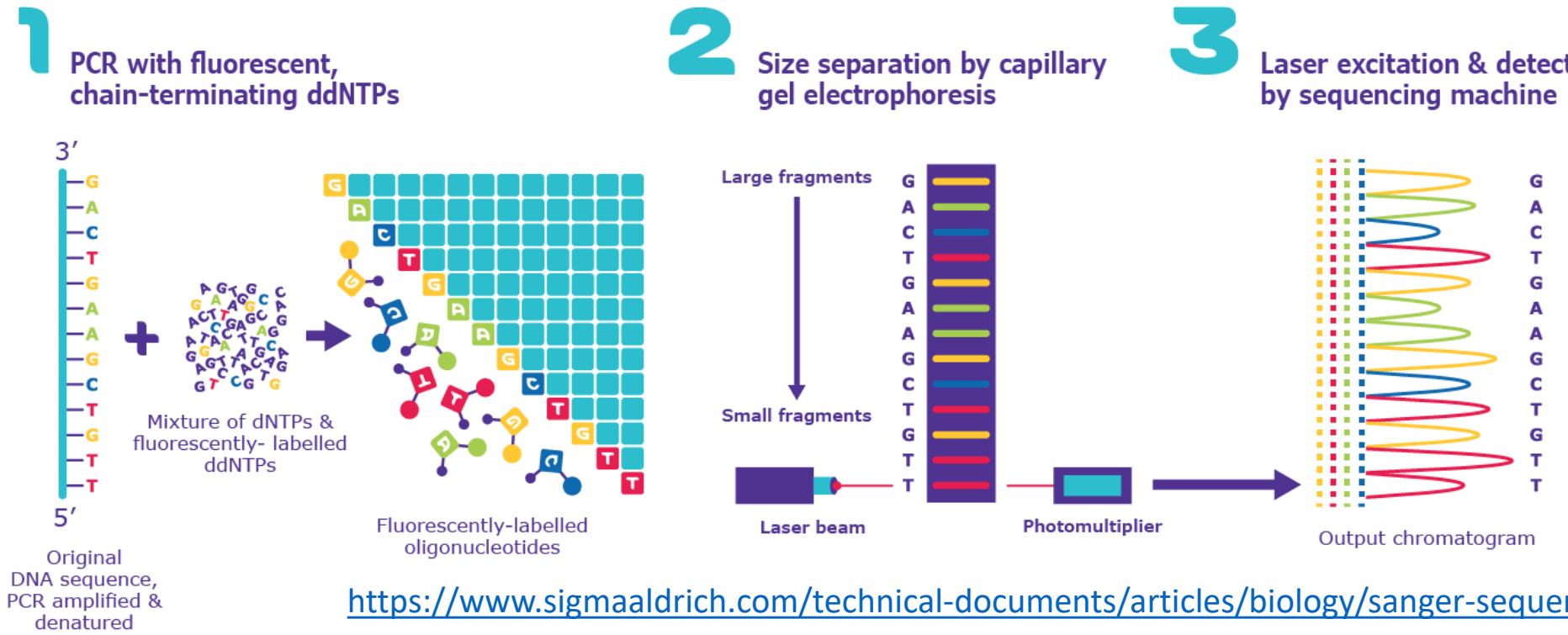
- The various sequencing technologies out there generally all have the same goal of finding the sequence of the DNA (or RNA) in your sample, so why does the technology used matter?
- It comes down to a series of tradeoffs that the various sequencing platforms offer. There are some differences in technical capabilities, but it mostly comes down to three factors:
  - Cost/Number of Reads/total output
  - Error Rate
  - Read Length

Long vs. short read technologies

- Why does read length matter? – Brainstorm
- Shorter reads mean more depth – more likely to find rarer things
- Shorter reads can be (but are not always) associated with greater total throughput, so can give better total coverage
- Longer reads make it easier to join things together into even bigger pieces such as for full genes, operons, chromosomes or genomes
- Longer reads are much better for detecting structural variants, i.e. insertions, deletions, duplications, rearrangements

Sanger sequencing

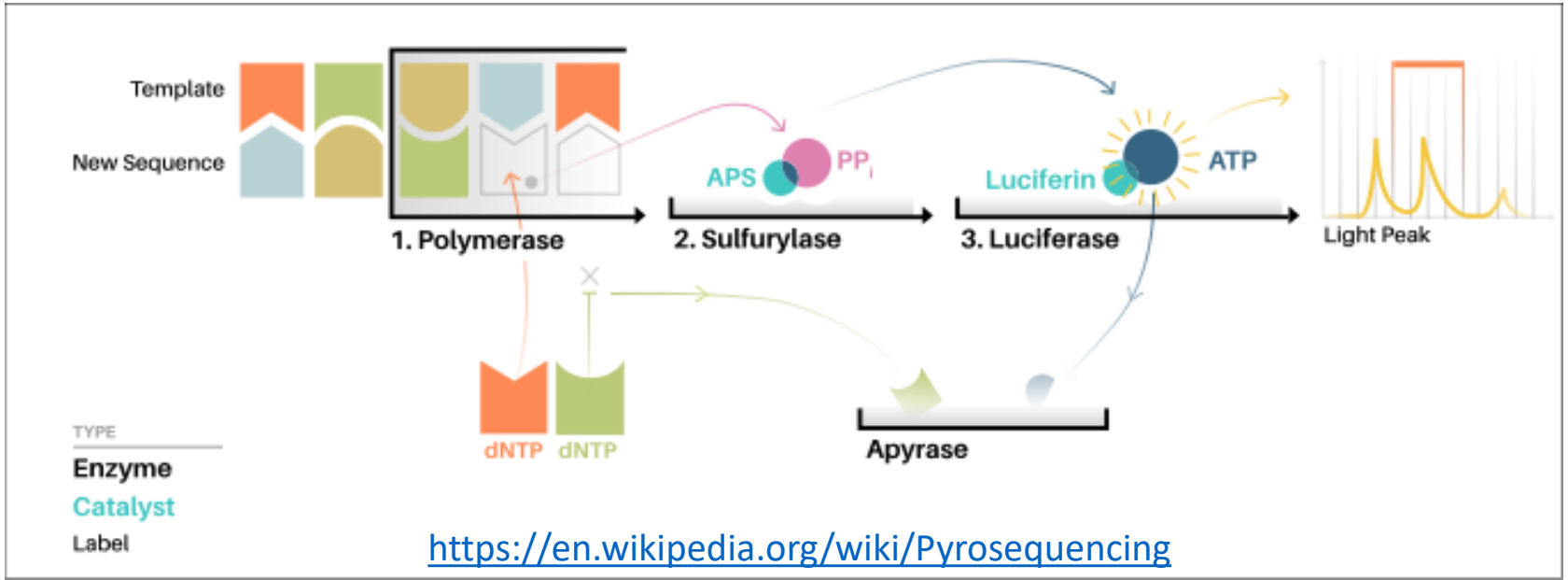
- The original revolutionary sequencing technology that is still in extensive use today for high accuracy fragment sequencing
- Limited to about 1000 bp and must be a pure sample (no mixed cultures)



https://www.sigmaaldrich.com/technical-documents/articles/biology/sanger-sequencing.html

Pyrosequencing

- The first of the next generation sequencing technologies
- Uses sequencing by synthesis (monitors action of DNA polymerase) and detection by chemiluminescence
- Sequences 300-500 bp per run
- Popularized by Roche with the 454 system, now discontinued (2013)



https://en.wikipedia.org/wiki/Pyrosequencing

Other Technologies that I won't go into detail about

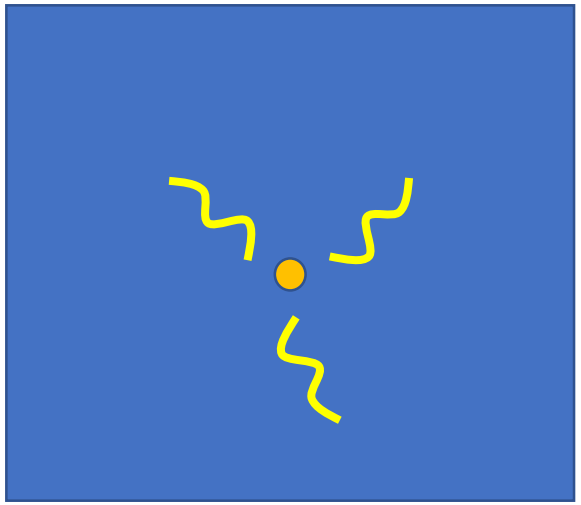
- DNA Microarrays
  - Not sequencing but can be used for genotyping or gene expression analysis
  - Will not detect novel genes/organisms, best suited for well studied organisms
- Ion Torrent
  - Detects release of hydrogen ions (pH changes) as DNA is synthesized
  - A short-read technology that hasn't caught on to the same extent as Illumina
- HiC and variants
  - Joins pieces of DNA that are 3-dimensionally close to one another, but not necessarily 1-dimensionally close
- 10X Genomics
  - Single cell sequencing technologies using barcoded nano-droplets

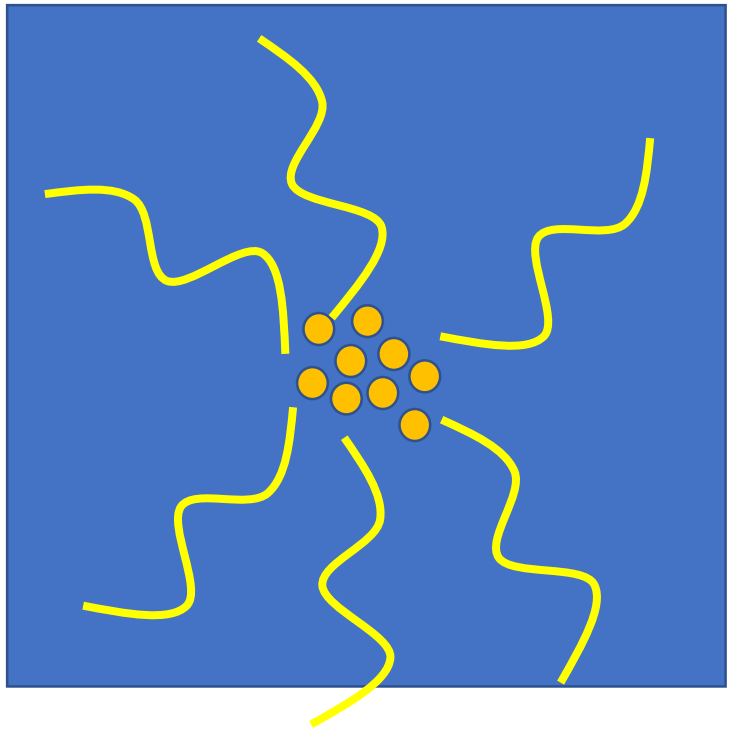
Illumina Technology

- Illumina technology is the most popular in use today
- It is a short-read technology, technically limited to about 500 bp, but often less depending on how the sequencing is performed
- https://www.illumina.com/science/technology/next-generation-sequencing/beginners.html

Question 1: Why use clusters?

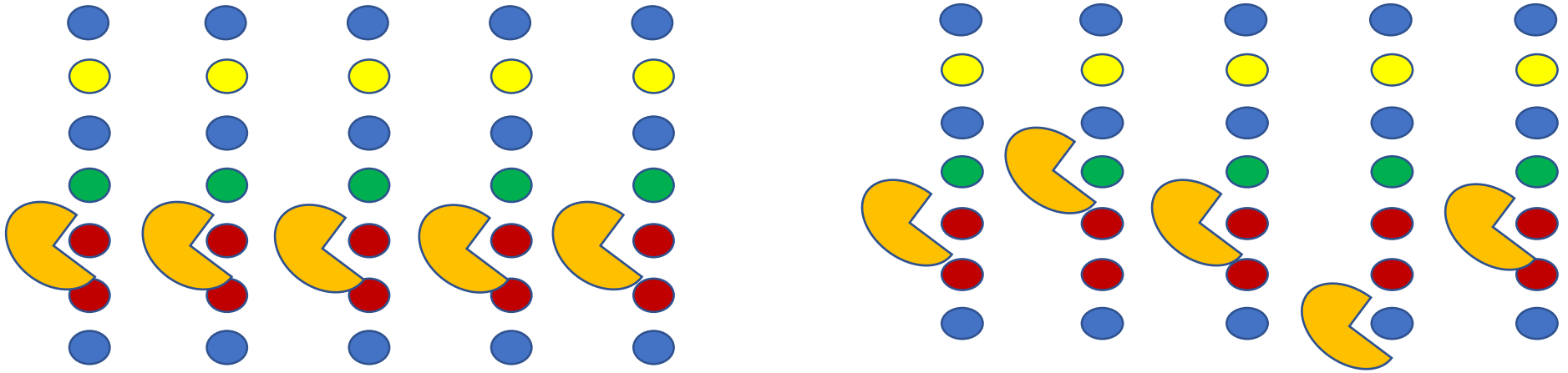
- In Illumina technology each bound DNA molecule is replicated about 1000x to make clonal clusters of each molecule before sequencing takes place. Why is this necessary?





Question 2: Why are the reads short?

- The Illumina technology is limited to about 150 or 250 bp read lengths and while you could make the instrument run longer, quality of reads would drop of dramatically? Why is this? Hint: there are a couple of sources of error, but the fundamental limitation to read length is related to the clustering.



What's with all the Seqs (MiSeq, HiSeq, NextSeq)?

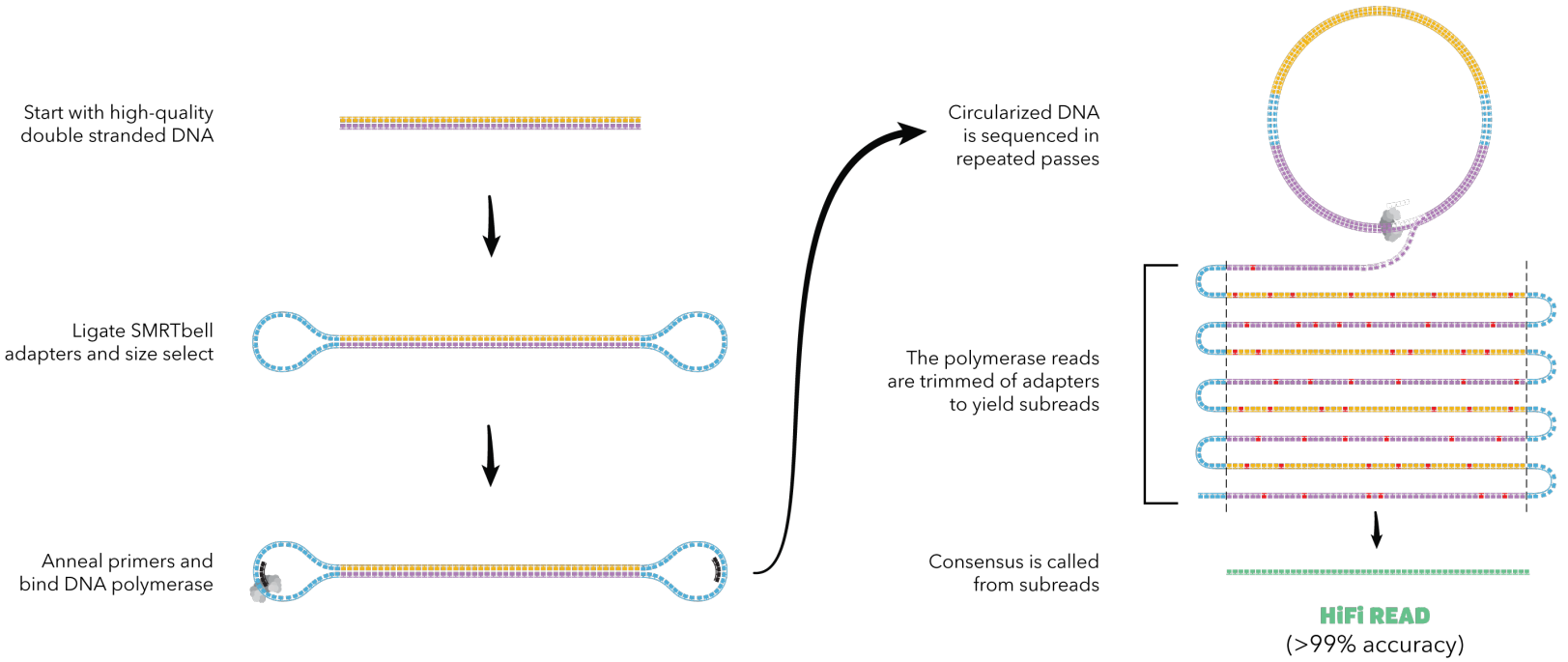
- iSeq100 – Up to 4 million reads per run (2x150bp)
- MiniSeq – Up to 25 million reads per run (2x150bp)
- MiSeq – Up to 25 million reads per run (2x300bp)
- NextSeq – Up to 1 billion reads per run (2x150bp)
- NovaSeq – Up to 20 billion reads per run (up to 2x250bp – reduces #)
- (Note: NovaSeq replaces HiSeq which is available in our sequencing core and limited to 3 billion reads per run. There is a NovaSeq at Hershey)

PacBio Technology

- A long-read technology, PacBio utilizes a single molecule approach to generating reads – no clustering induced limitations.
- Also utilizes real-time sequencing – there is no pause between nucleotide additions. This means that kinetics of base addition can be used to detect modified bases
- Trade off in terms of number of reads and/or read length vs read quality
- https://www.pacb.com/smrt-science/smrt-sequencing/

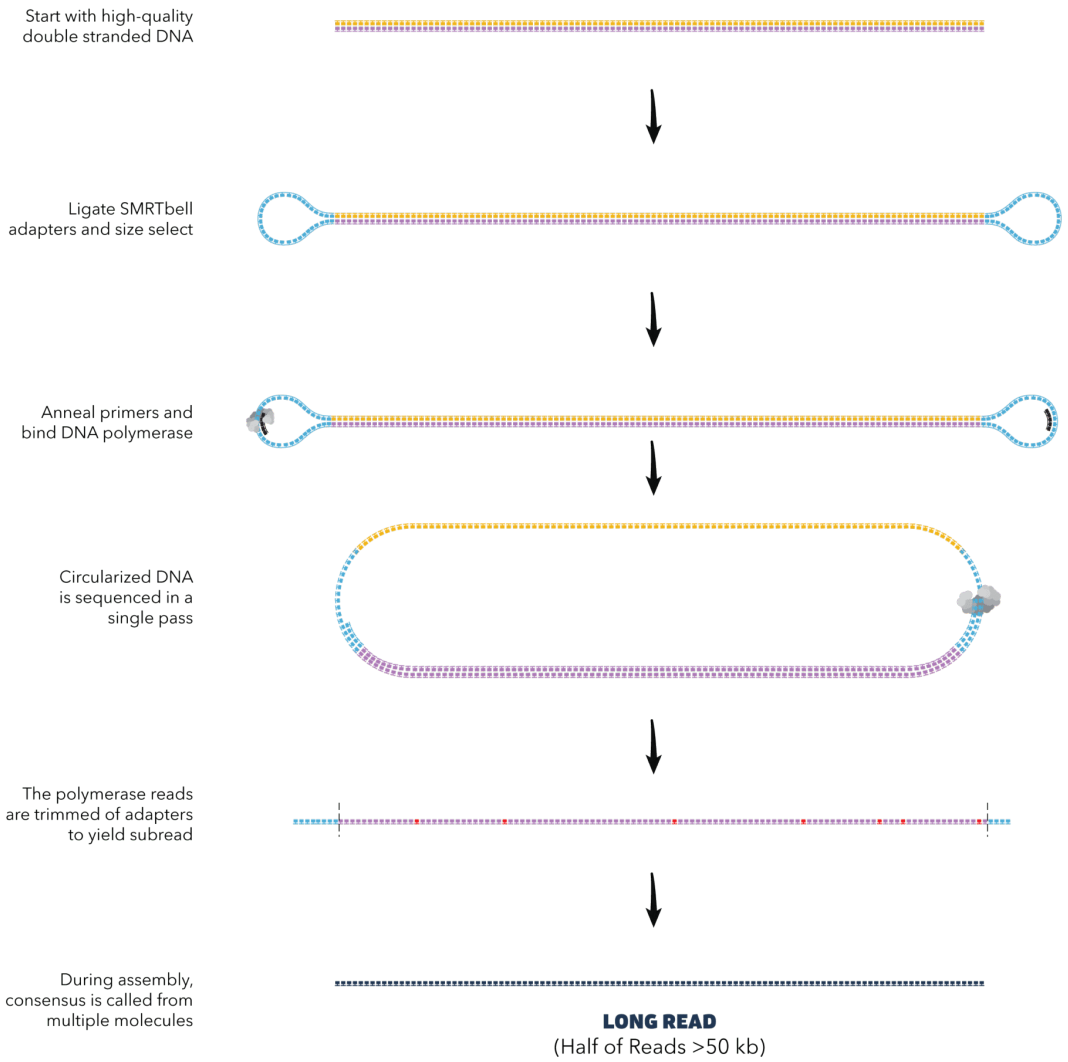
PacBio Mode 1: Circular Consensus Sequencing

- In CCS mode, inserts less than about 15-20 Kb get sequenced multiple times (About 10x). A consensus read can then be generated from these sub-reads to get a very high accuracy



PacBio Mode 2: Continuous Long Read

- In this mode the system is optimized to produce as long of reads as possible, but not repeatedly
- Limited in length of read by the quality of the DNA (are there intact fragments long enough) and by the maximum single run synthesis by the polymerase
- Variable lengths, but many above 50 Kb and up to 175 Kb



PacBio Instruments

- RS and RS-II – have 150,000 ZMWs and can generate 50,000-75,000 reads per run
- Sequel – have 1 million ZMWs and can generate 300K-500K reads per run
- Sequel II have 8 million ZMWs and can generate 2.5 million-4 million reads per un

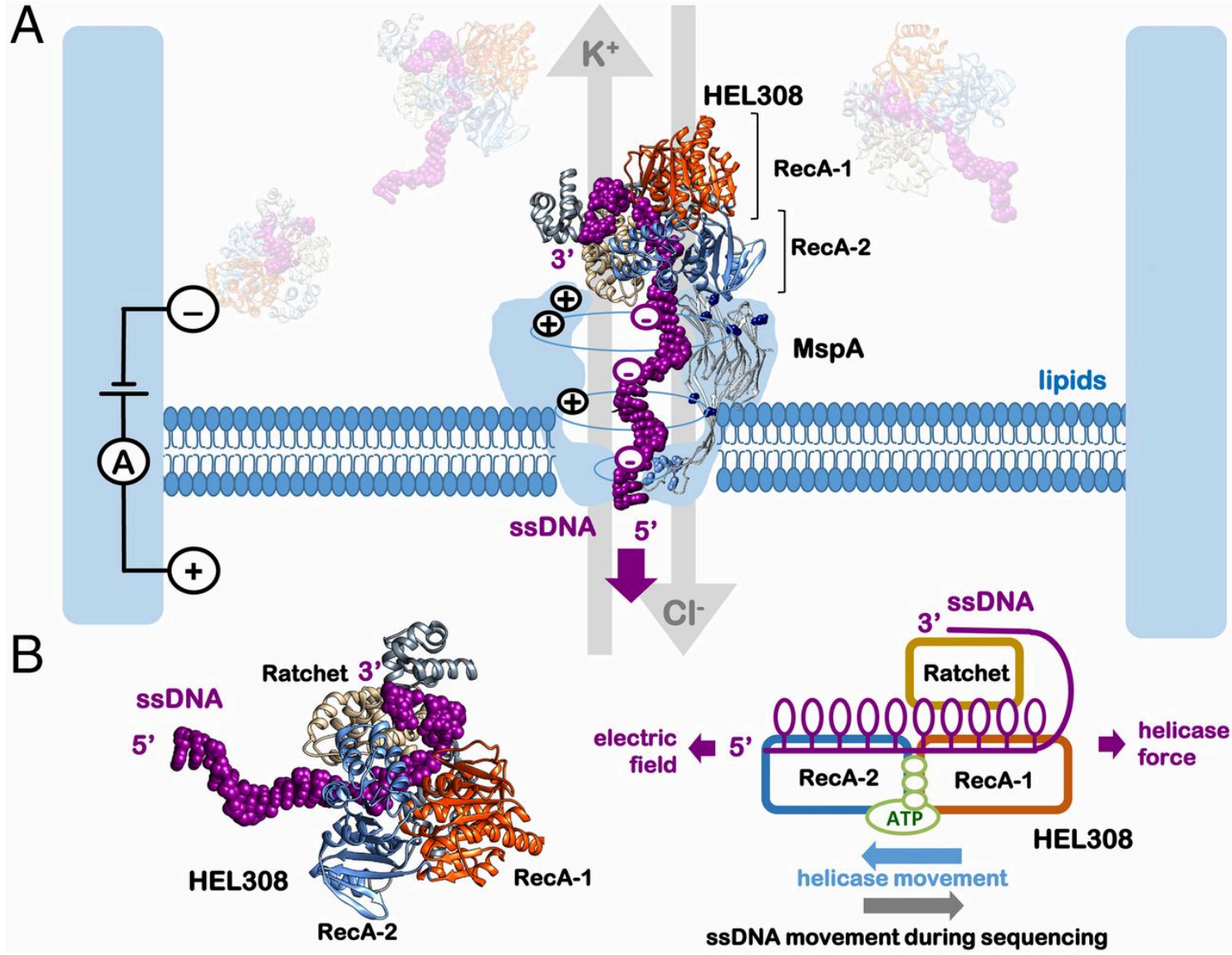
- Why do they only generate between 1/3 and ½ the number of reads relative to the number of ZMWs?

Oxford Nanopore technology

- Another long-read (up to 2 Mb!) technology that uses single molecules
- This differs from the others we have discussed in that it is not sequencing by synthesis, rather individual bases are detected in turn as they pass through a nanopore and disrupt an ionic current
- Potentially significantly cheaper than the other technologies (depending on instrument and experiment)
- Big drawback is error rate – initially as high as 30%!, now claimed to be down to 5%, but real world results seem to be 7-10%
- https://nanoporetech.com/how-it-works

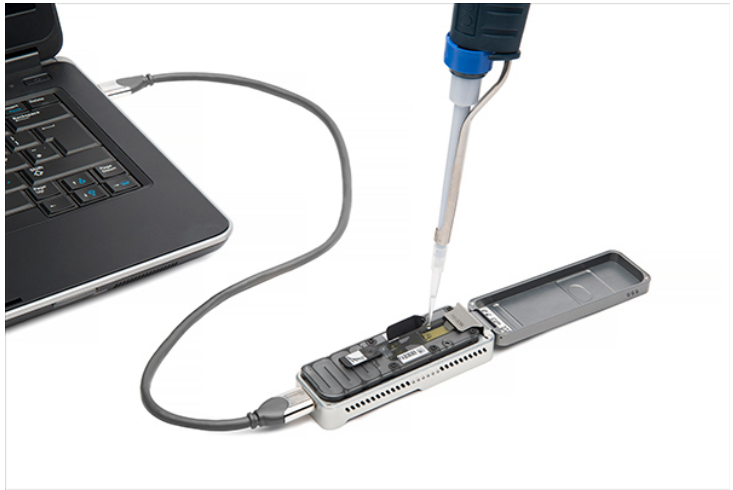
Nanopore Technology

- Each nanopore is created by a pore forming protein that DNA can be fed through
- A motor protein, DNA helicase unwinds a single strand of DNA pushes it through the pore
- Changes in ion flow between the compartments is analyzed to determine the base sequence of the DNA



The Instruments

- MinIon – Processes one flow cell with 512 nanopores, capable of up to 30 Gb of sequencing
- GridIon – Process up to 5 flow cells at a time (+ built in computing power for analysis)
- PromethIon – Process up to 48 flow cells at a time each with 3000 nanopores for up to 8 Tb of data per run!



Applications

- Real time sequencing results – can be out in the field with these instruments (especially MinIon) and be getting results within minutes for rapid diagnostic purposes
- Read Until… - possible to have the instrument selectively sequence certain sequence signatures – if it doesn’t match a certain pattern within the first few bases, DNA molecule can be expelled from pore and a new one bound
- Direct analysis of molecule without sequencing by synthesis allows direct detection of DNA/RNA modifications (including direct reading of RNA

Combining short read and long read technology – activity?

- Would there be any advantage to combining these various technologies?
- For long assemblies such as genomes, long reads can be used to provide the majority of the assembly and short reads can be used to error correct.
- Alternatively, the short reads can be used to provide the majority of the sequence and long reads can be used to help assemble them
- With metagenomes long reads can provide enhanced assemblies, while short reads can provide depth to detect rare taxa

What Technology would you use for your real or hypothetical project?