Microbiome Center Bootcamp: Sequencing Technology

Dr. Darrell Cockburn Associate Professor of Food Science

Why does sequencing technology matter?

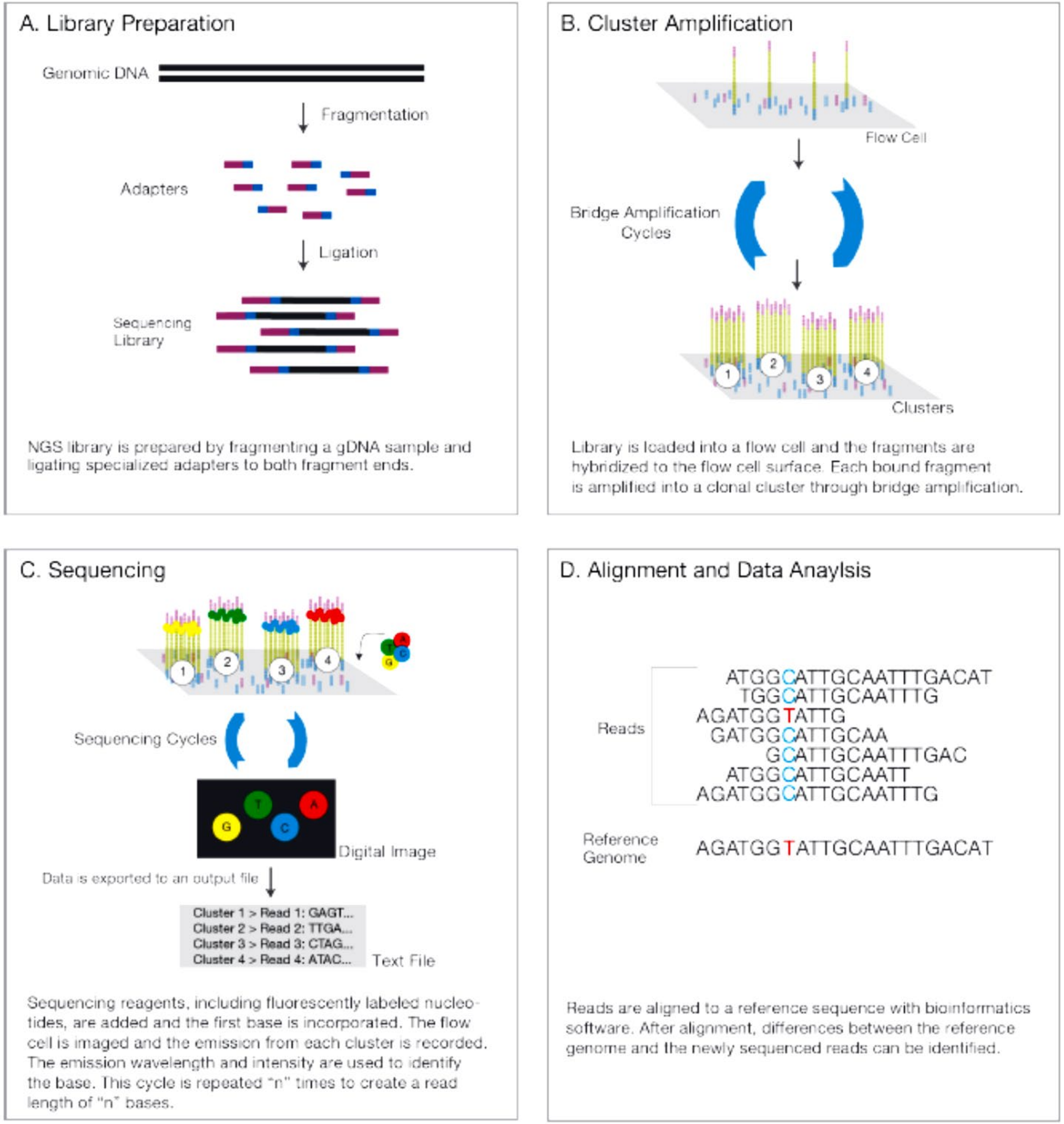
- The various sequencing technologies out there generally all have the same goal of finding the sequence of the DNA (or RNA) in your sample, so why does the technology used matter?
- It comes down to a series of tradeoffs that the various sequencing platforms/kits offer. There are some differences in 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?
- 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

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
- Known for excellent error rates



```
Cluster 1 > Read 1: GAGT...
Cluster 2 > Read 2: TTGA...
Cluster 3 > Read 3: CTAG...
Cluster 4 > Read 4: ATAC...
```

NGS library is prepared by fragmenting a gDNA sample and ligating specialized adapters to both fragment ends.

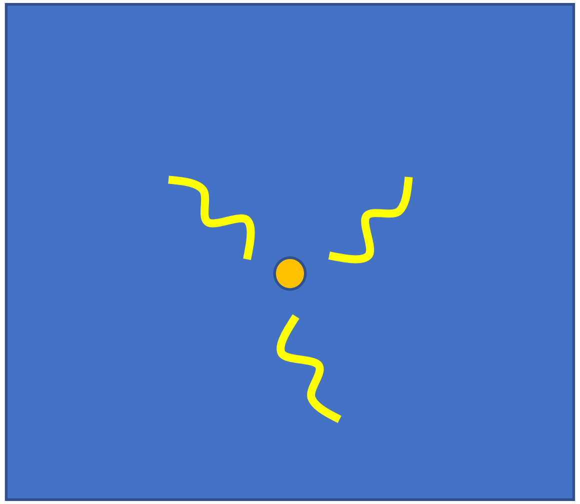
Library is loaded into a flow cell and the fragments are hybridized to the flow cell surface. Each bound fragment is amplified into a clonal cluster through bridge amplification.

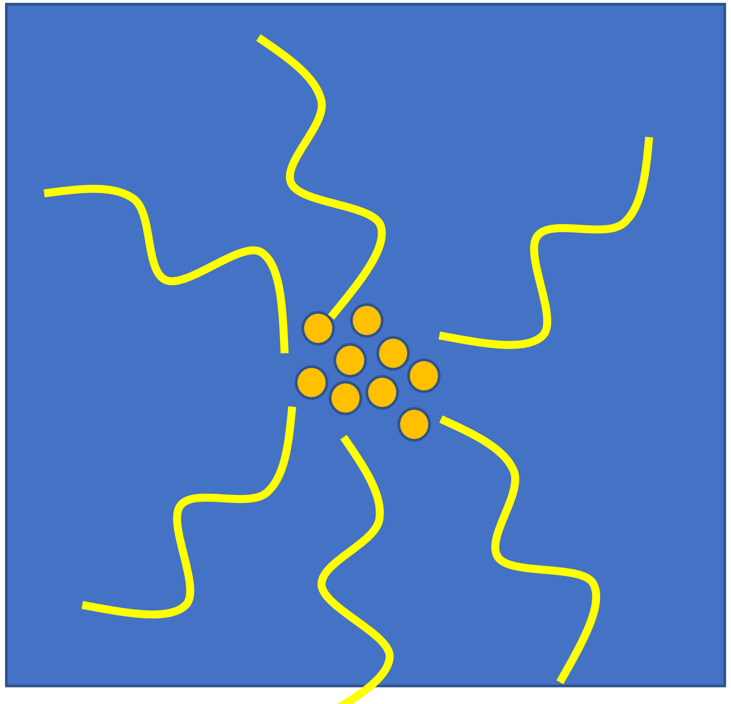
Sequencing reagents, including fluorescently labeled nucleotides, are added and the first base is incorporated. The flow cell is imaged and the emission from each cluster is recorded. The emission wavelength and intensity are used to identify the base. This cycle is repeated "n" times to create a read length of "n" bases.

Reads are aligned to a reference sequence with bioinformatics software. After alignment, differences between the reference genome and the newly sequenced reads can be identified.

Question 1: Why use clusters?

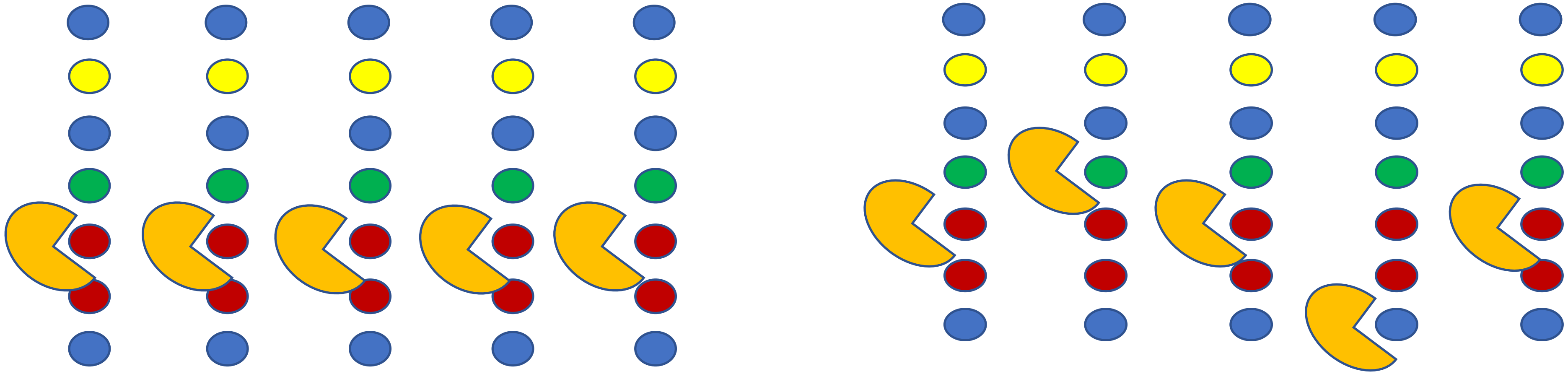
- In Illumina technology each bound DNA molecule is replicated about 1000x to make clonal clusters of each molecule before sequencing takes place. Why is this necessary?
- Improves sensitivity and allows for that really good error rate as you are getting the consensus for the same molecule being sequenced many times in the cluster.





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 off dramatically? Why is this? Over time the DNA polymerases on the various molecules in the cluster get out of sync and error rate will shoot up.



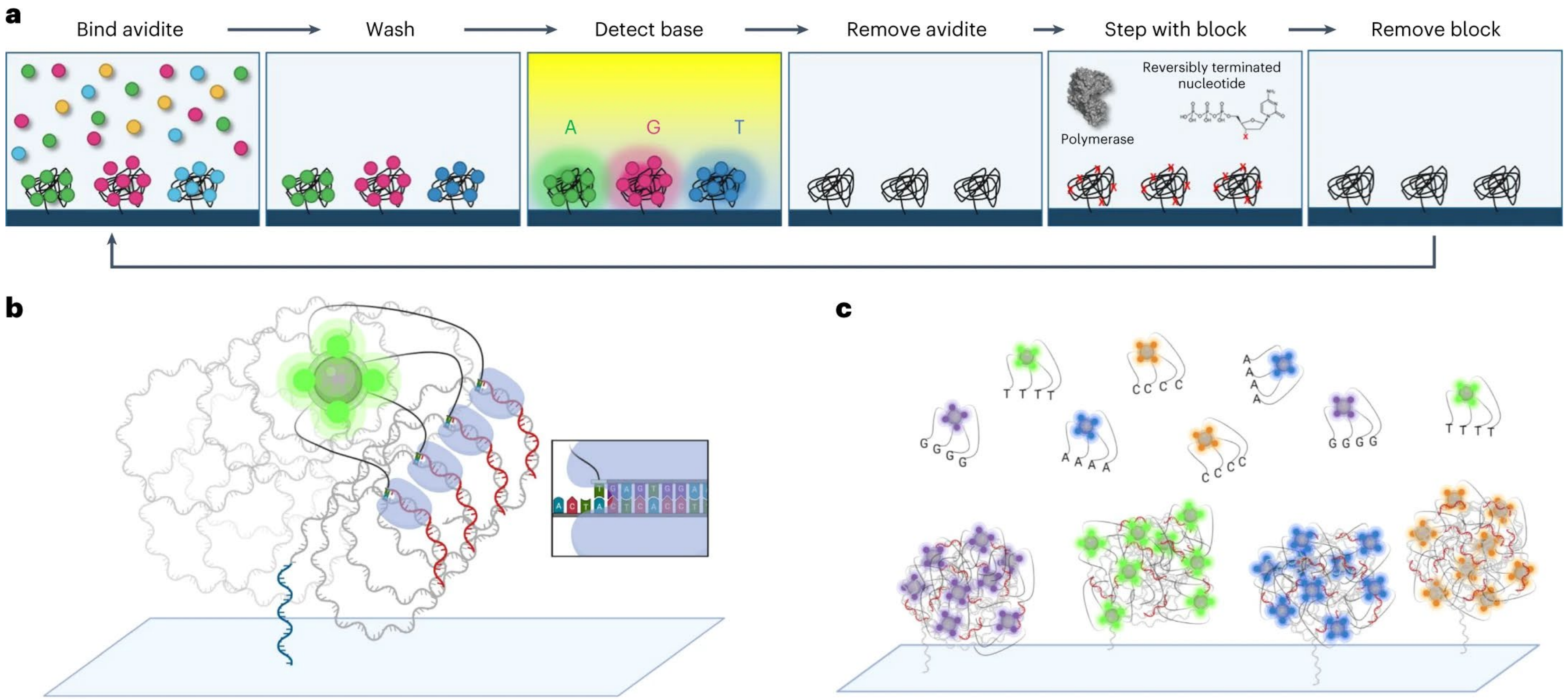
What’s with all the Seqs?

- iSeq100 – 4 million max reads per run (2x150bp)
- MiniSeq – 25 million max reads per run (2x150bp)
- *MiSeq – 25 million max reads per run (up to 2x300bp)
- *NextSeq – 1.2 billion max reads per run (up to 2x300bp)
- **NovaSeq – 20 billion max reads per run (up to 2x250bp)
- NovaSeqX – 52 billion max reads per run (2x150bp)

- * Present at Huck Genomics Core
- ** Available at Hershey Sequencing Core

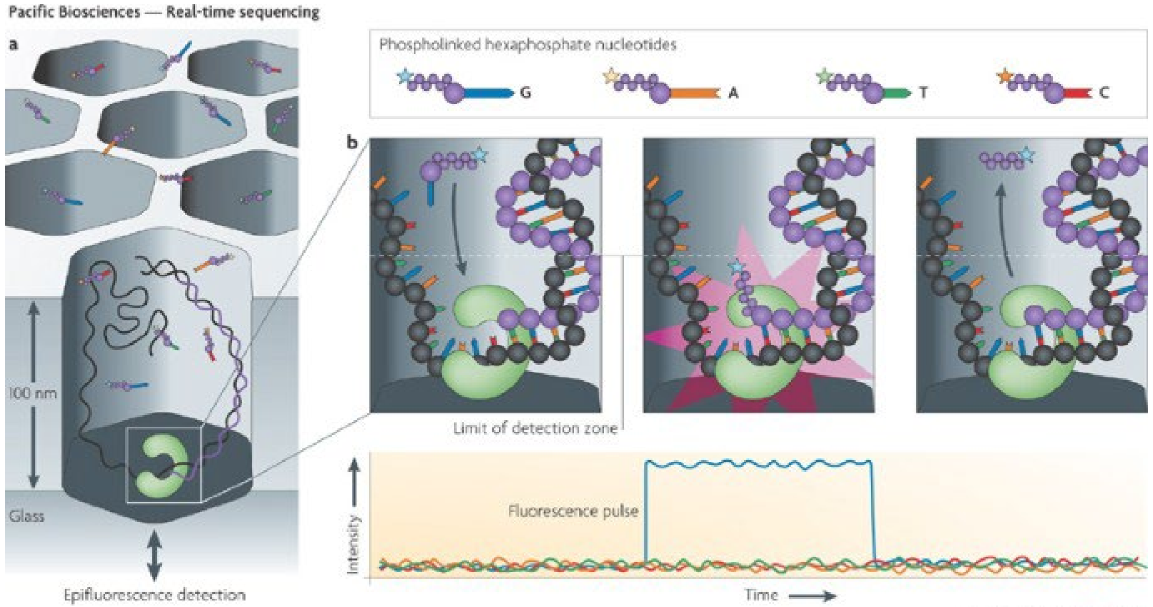
Element Biosciences Aviti

- An emerging variation on the sequencing by synthesis paradigm, this process separates base detection and synthesis steps.
- Also adds multiple interactions between flurophore and polymerase for avidity effects



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

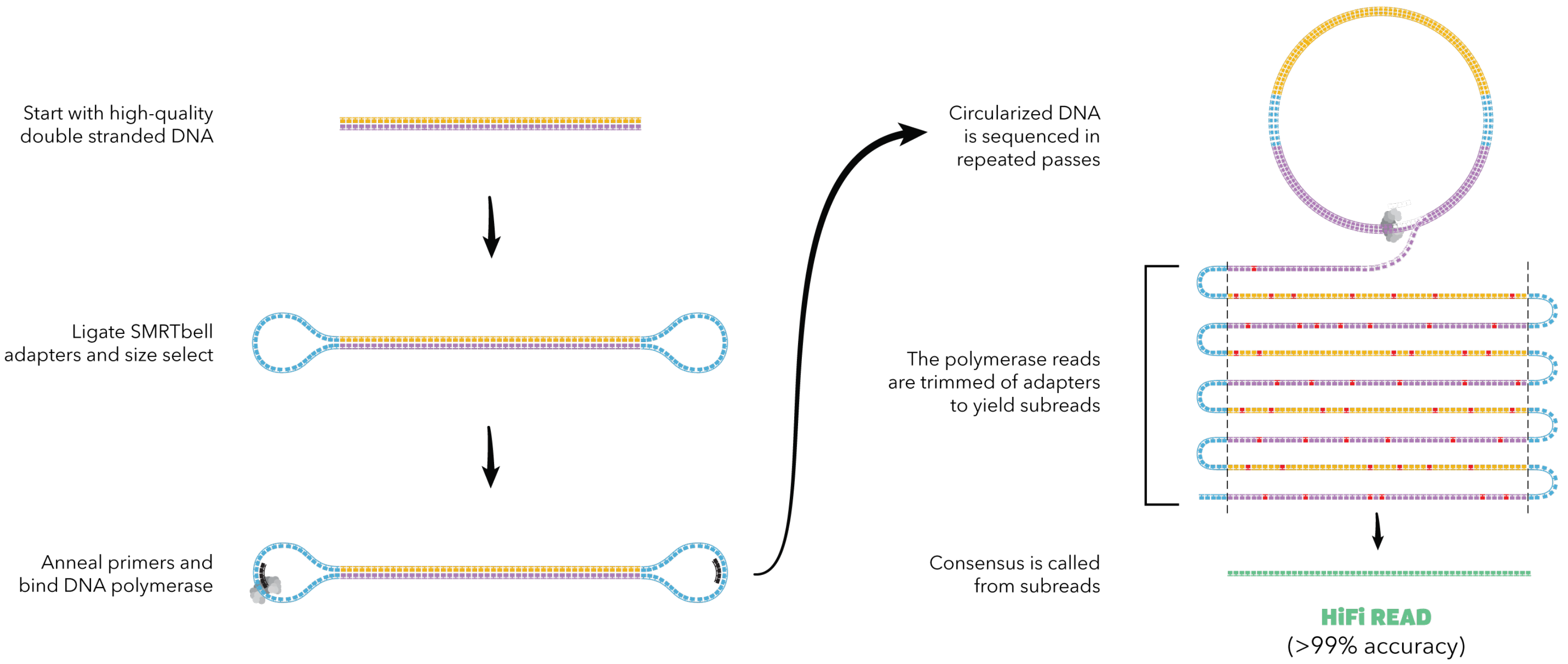

Pacific Biosciences — Real-time sequencing
a
b
Phospholinked hexaphosphate nucleotides
G
A
T
C
100 nm
Glass
Epifluorescence detection
Limit of detection zone
Fluorescence pulse
Intensity
Time

Nature Reviews | Genetics

Doi:10.1038/nrg2626

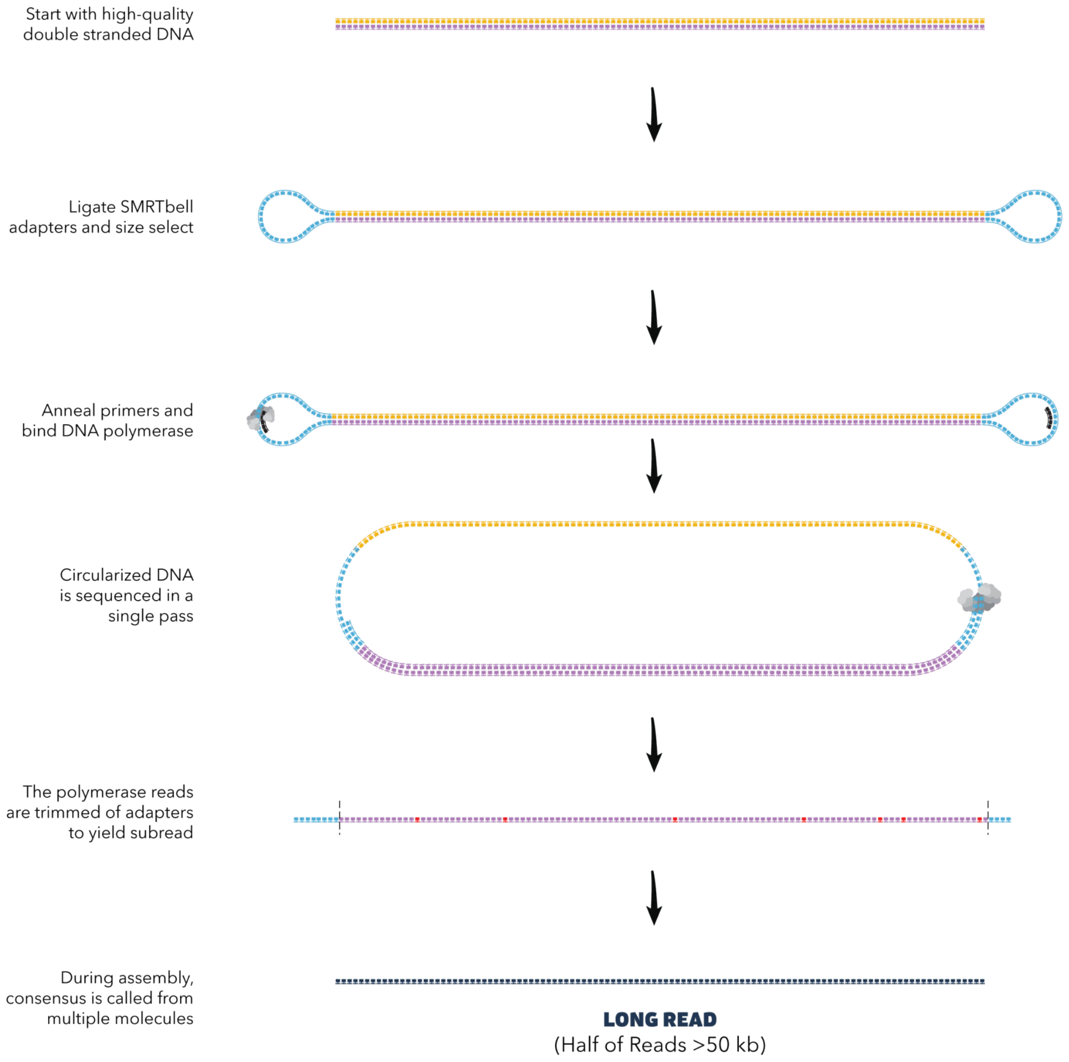
PacBio Mode 1: Circular Consensus Sequencing

- 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


Start with high-quality double stranded DNA
Ligate SMRTbell adapters and size select
Anneal primers and bind DNA polymerase
Circularized DNA is sequenced in repeated passes
The polymerase reads are trimmed of adapters to yield subreads
Consensus is called from subreads
HiFi READ
(>99% 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


Start with high-quality double stranded DNA
Ligate SMRTbell adapters and size select
Anneal primers and bind DNA polymerase
Circularized DNA is sequenced in a single pass
The polymerase reads are trimmed of adapters to yield subread
During assembly, consensus is called from multiple molecules
LONG READ
(Half of Reads >50 kb)

PacBio Instruments

- *Sequel II has 8 million ZMWs and can generate 2.5 million-4 million reads per run
- Revio has 25 million ZMWs per cell, 4 cells so up to 100 million ZMWs up to about 50 million reads
- Onso – a short read system with capabilities similar to Aviti

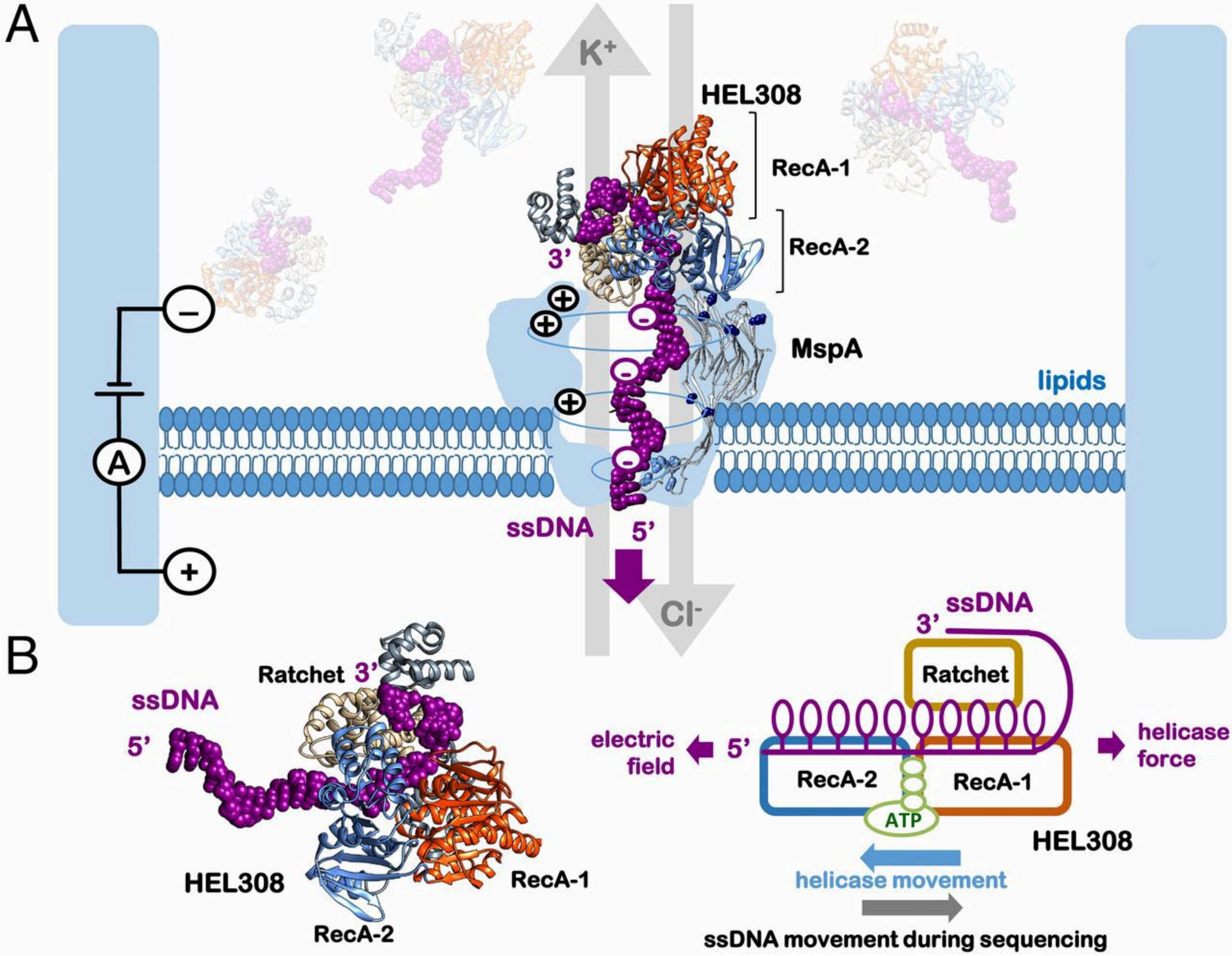
- Why do they only generate between 1/3 and ½ the number of reads relative to the number of ZMWs? Poisson distribution to get no more than one DNA molecule per ZMW
- * Available in Huck Genomics Core

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 latest flow cells and chemistries seem to be down to <2%

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 and 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 and long read technology

- 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

Questions:

- 1. What aspects of your mars sampling project would you need to consider when choosing your sequencing technology/approach

- 2. Which technology/approach would you choose?

- 3. If you had to consider costs would your answer change?