Genomics Part 1: The basics of sequencing technology

Bridging the Bench-Machine Learning Gap

Dr. Emily A. Beck

Dr. Jake Searcy

At the end I’m going to ask you to brainstorm questions you may have about the types of data you can generate. Keep this in mind as we move through…

Learning Objectives

Understand how sequencing technologies have developed over time

Learn the fundamental biology behind Sanger/Illumina/PacBio/Nanopore technology

Learn about the different types of sequencing reads you can generate

Brainstorm a list of questions to ask about the options available

Doug Turnbull from The Genomics and Cell Characterization Core Facility (**GC3F**) will be in on Thursday to talk about the machines available, today we will focus on the biology behind the sequencing

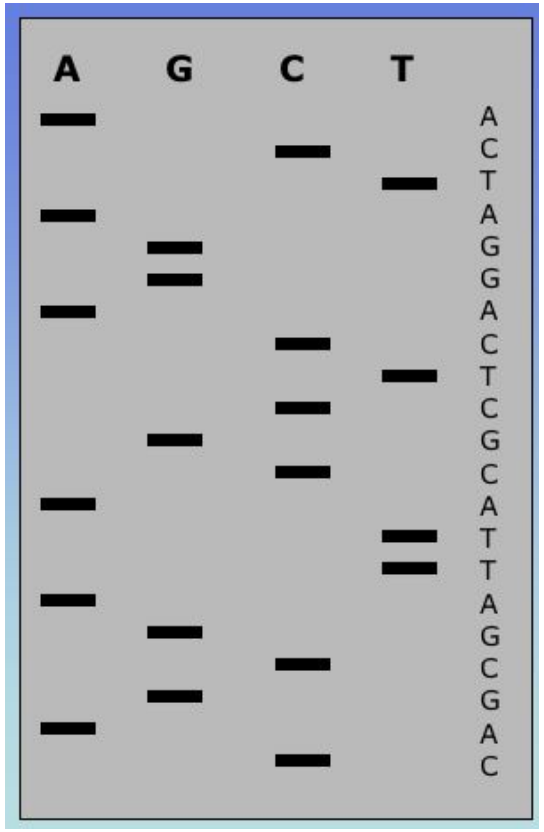
Sanger Sequencing: running gels to automation with fluorometry

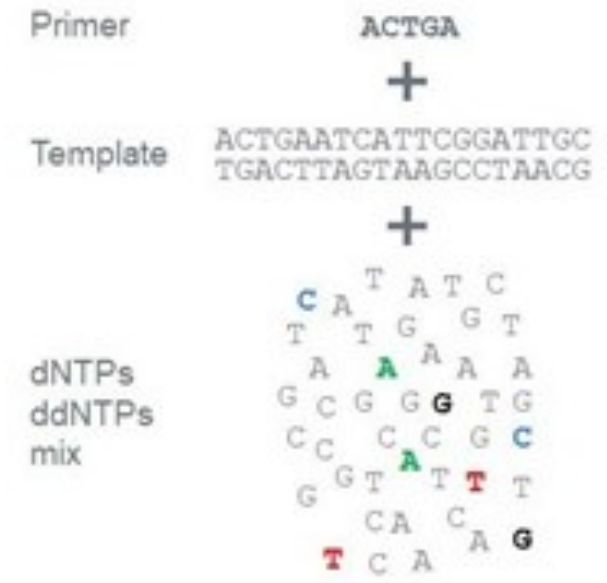
"First Generation"

Building on the principles of Polymerase Chain Reaction (PCR)

Template + primer + dNTPs (nucleotides) + polymerase +buffer

Sequencing reactions add ddNTPs (chain terminating nucleotides)





Nucleotides are added like in a normal PCR, and **fluorescently labelled ddNTPs** stop the reaction statistically after every nucleotide

Sanger Sequencing: Historical Development of Automated DNA Sequencing (2012)

“Next Generation”: Short Read sequencing

“Second Generation”, “Massively parallel”, or “Deep sequencing”

- Sometimes these alternative names are used which can be misleading
- You can use short-read sequencing to sequence to a shallow depth. (We will talk more about coverage and depth later)

I can’t keep track of who owns who, but let’s focus on Short Read sequencing biology on an Illumina platform

“Next Generation”: Short Read sequencing

Bench Top Sequencers



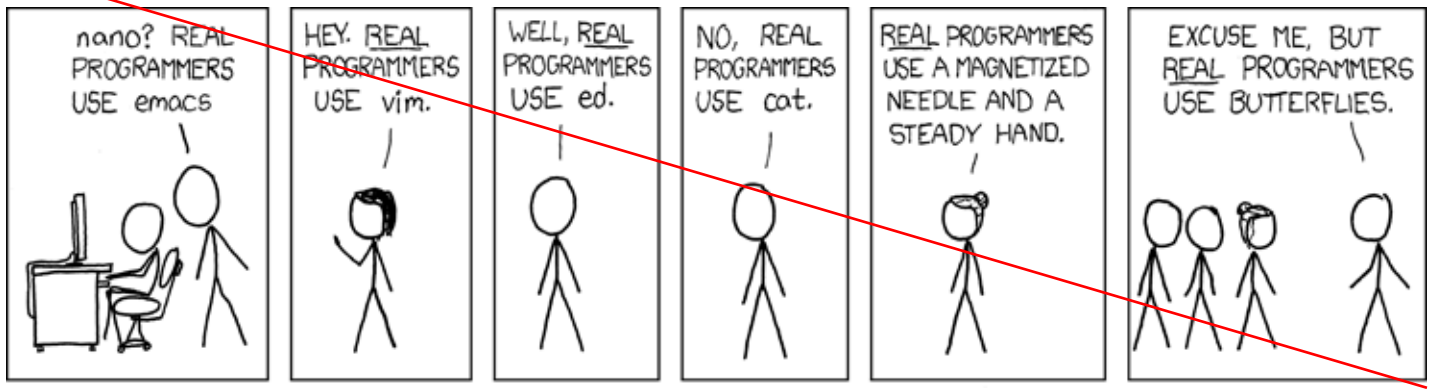
Popular Applications & Methods	MiSeq Series Key Application	NextSeq 550 Series Key Application	NextSeq 1000 & 2000 Key Application
Large Whole-Genome Sequencing (human, plant, animal)			
Small Whole-Genome Sequencing (microbe, virus)	●	●	●
Exome & Large Panel Sequencing (enrichment-based)		●	●
Targeted Gene Sequencing (amplicon-based, gene panel)	●	●	●
Single-Cell Profiling (scRNA-Seq, scDNA-Seq, oligo tagging assays)		●	●
Transcriptome Sequencing (total RNA-Seq, mRNA-Seq, gene expression profiling)		●	●
Targeted Gene Expression Profiling	●	●	●
miRNA & Small RNA Analysis	●	●	●
DNA-Protein Interaction Analysis (ChIP-Seq)	●	●	●
Methylation Sequencing		●	●
16S Metagenomic Sequencing	●	●	●
Metagenomic Profiling (shotgun metagenomics, metatranscriptomics)		●	●
Cell-Free Sequencing & Liquid Biopsy Analysis		●	●

Difference Machines use tweaks to the same foundational biological principles optimizing them for different types of projects

“Next Generation”: Short Read sequencing

Production-Scale Sequencers: NextSeq and NovaSeq
*HiSeq has been discontinued

NovaSeq 6000

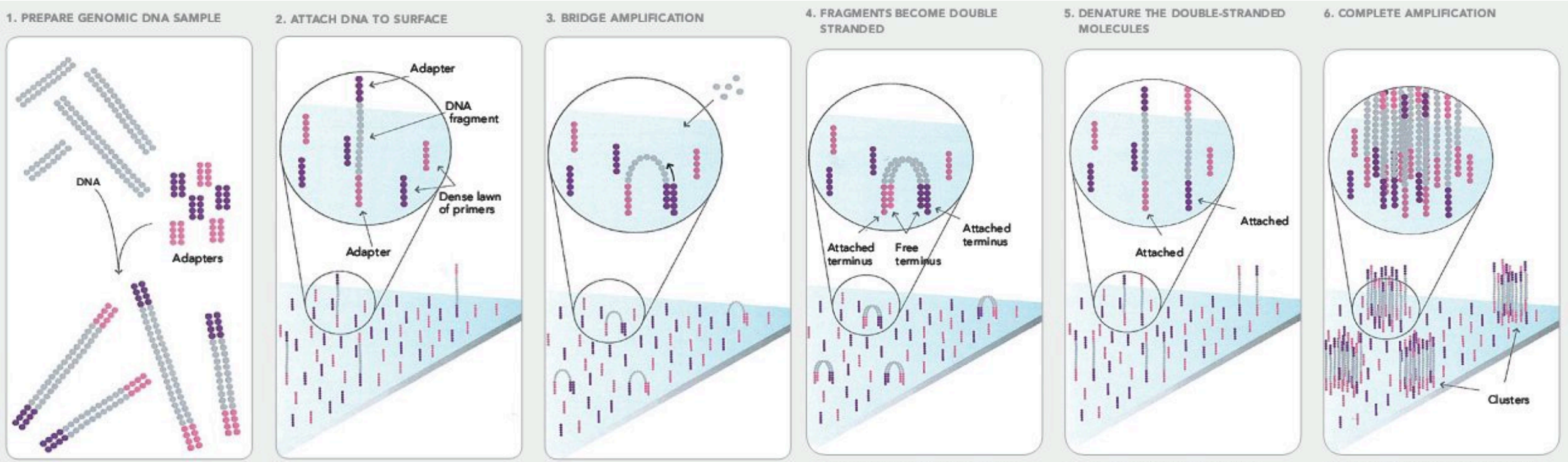


Popular Applications & Methods	Key Application
Large Whole-Genome Sequencing (human, plant, animal)	●
Small Whole-Genome Sequencing (microbe, virus)	●
Exome & Large Panel Sequencing (enrichment-based)	●
Targeted Gene Sequencing (amplicon-based, gene panel)	●
Single-Cell Profiling (scRNA-Seq, scDNA-Seq, oligo tagging assays)	●
Transcriptome Sequencing (total RNA-Seq, mRNA-Seq, gene expression profiling)	●
Chromatin Analysis (ATAC-Seq, ChIP-Seq)	●
Methylation Sequencing	●
Metagenomic Profiling (shotgun metagenomics, metatranscriptomics)	●
Cell-Free Sequencing & Liquid Biopsy Analysis	●

Proposing to use an inappropriate or discontinued sequencing platform is a major red flag in grant and fellowship proposals!

If you have questions consult with GC3F or someone who analyzes the type of data you are trying to generate.

Foundational Biology of Short-Read Illumina Sequencing



Foundational Biology of Short-Read Illumina Sequencing

<u>Step 1</u>: Obtain Template Material from your sample

how you do this matters!

- You may need high molecular weight DNA
- You may need to account for kit contaminations (16S-seq)

Consult! Consult! Consult!

Images @HenriksLab

Foundational Biology of Short-Read Illumina Sequencing

<u>Step 2:</u> Library Preparation

2a. Shear template to desired size (fragmentation)

-Lots of ways to shear: sonication, acoustic shearing, hydrodynamic shearing, enzymatic shearing, and more.

-May need to check on the quality (QC) of your library at this stage before spending more money (**Fragment Analysis**)

Images @HenriksLab

Foundational Biology of Short-Read Illumina Sequencing

Step 2: Library Preparation

2b. Ligation of adaptors (two different types)

(1) Universal region that matches sequencing primers

(2) Hybridization sequence to allow fragments to adhere to a chip

(3) Unique barcodes to identify which fragments came from which individuals

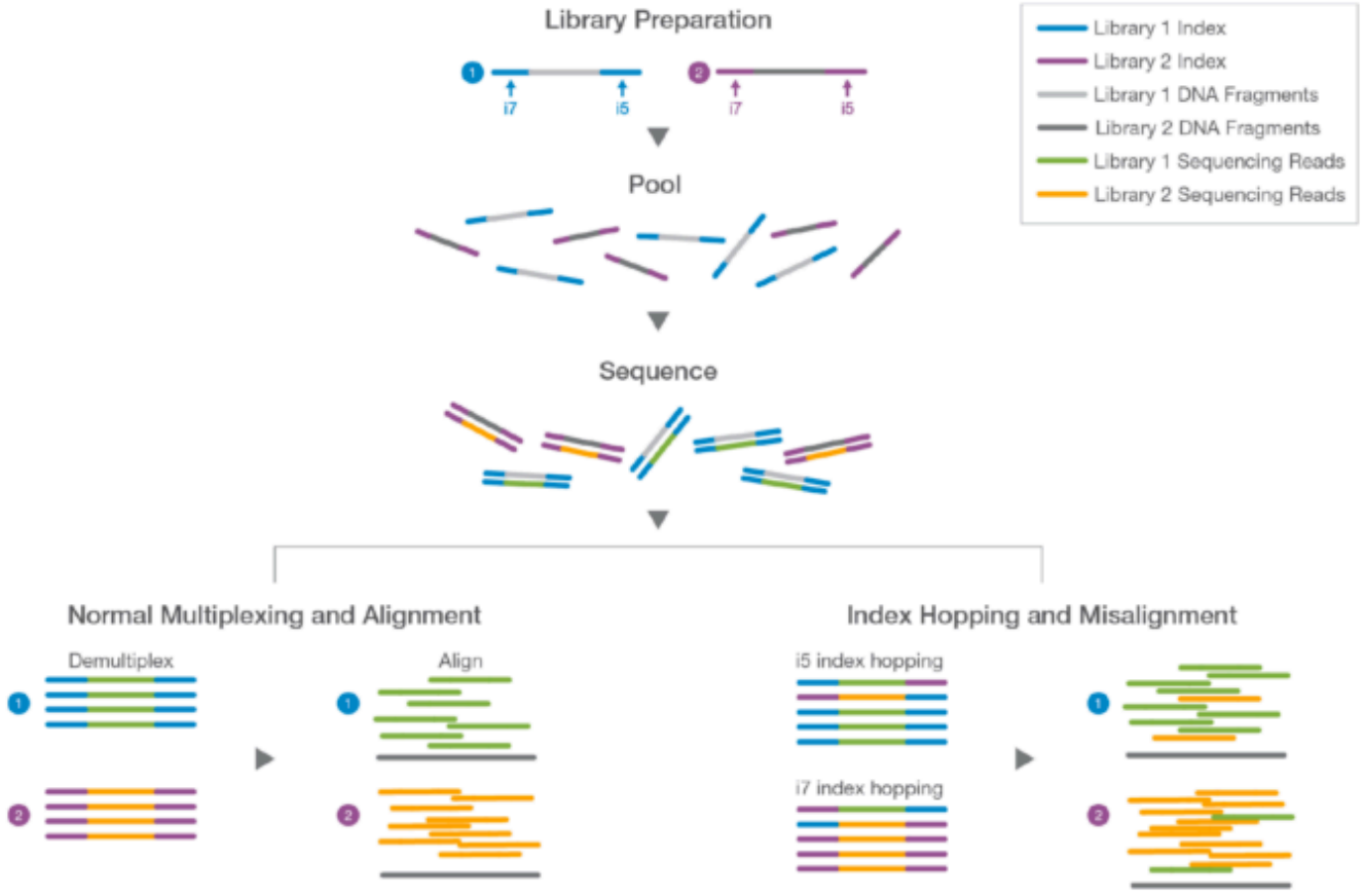
To reduce costs you may pool projects together into a single sequencing run.

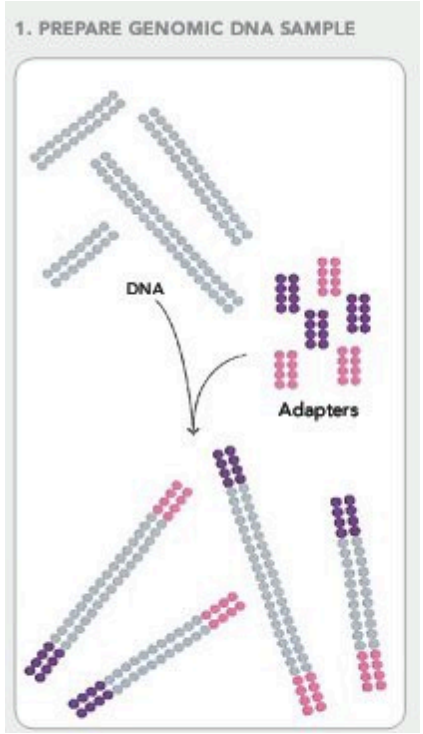
You need to make sure all projects are compatible and able to be pooled.

Images @HenriksLab

Foundational Biology of Short-Read Illumina Sequencing

One concern about library prep has been to reduce probability of index hopping





(The Sequencing Center)

Foundational Biology of Short-Read Illumina Sequencing

Step 2: Library Preparation

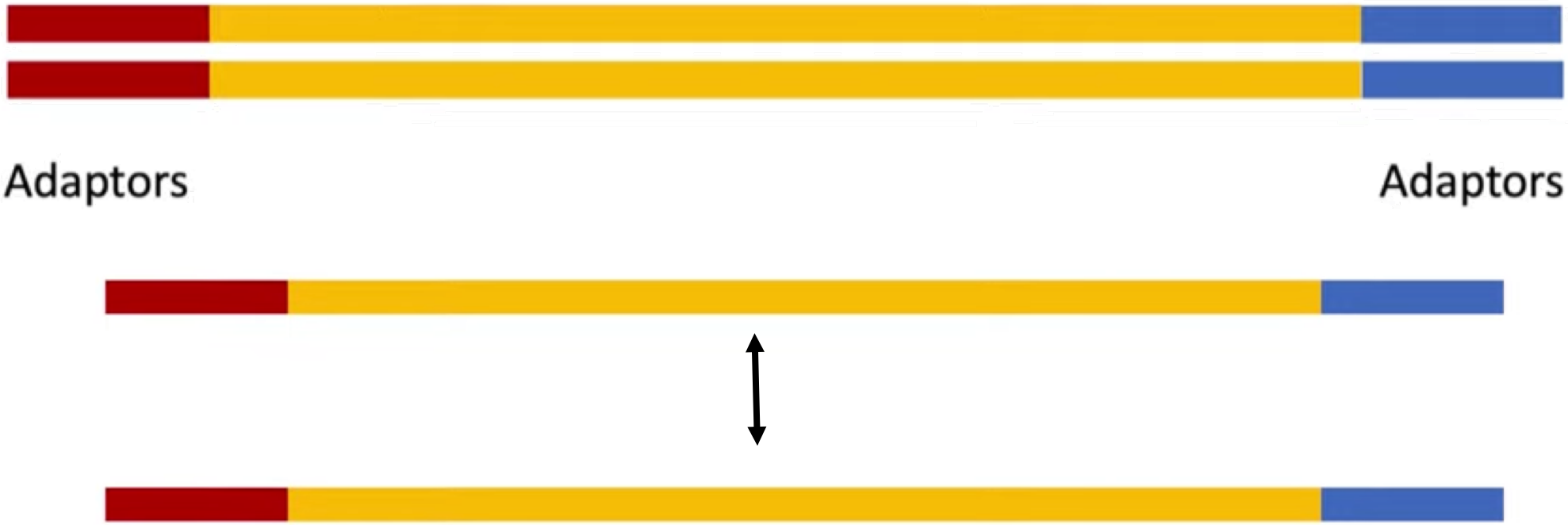
2c. Amplification of the library
Need enough material to sequence

2d. Multiplex (Pool libraries together into a single run)

Many QC steps along the way but we are now finally ready to actually start sequencing

Uniquely barcoded fragments are pooled any ready to interact with our chip

Step 1: Denature

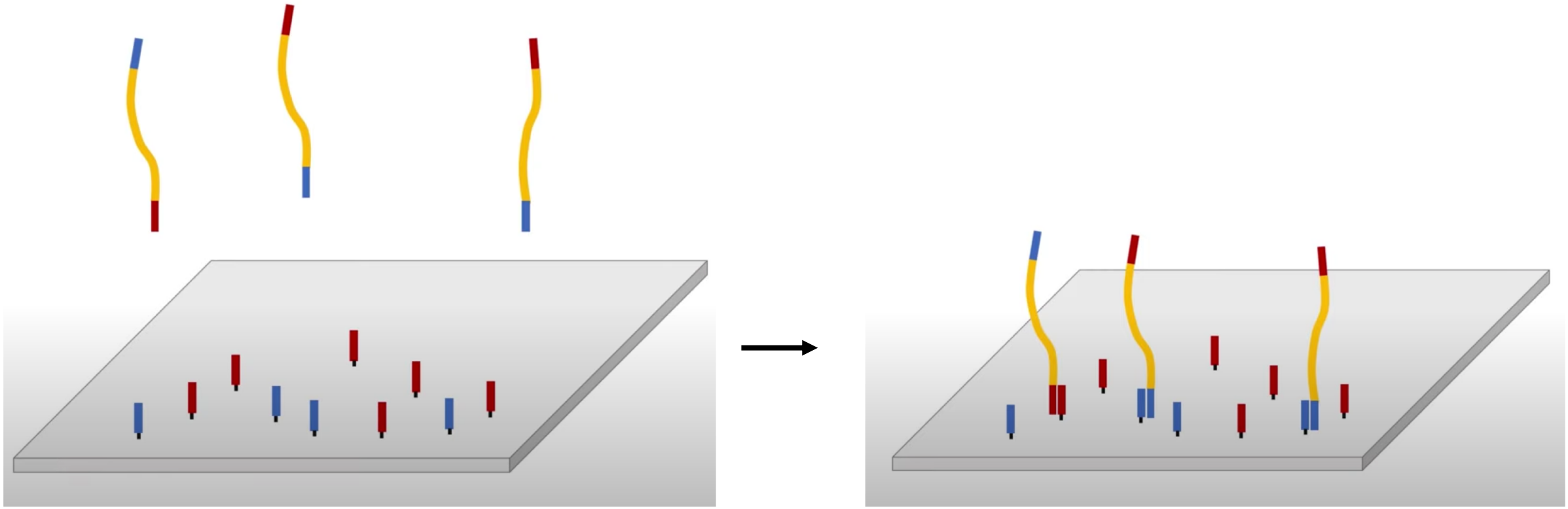


Images @HenriksLab

Uniquely barcoded fragments are pooled any ready to interact with our chip

Step 2: Bind to the Chip

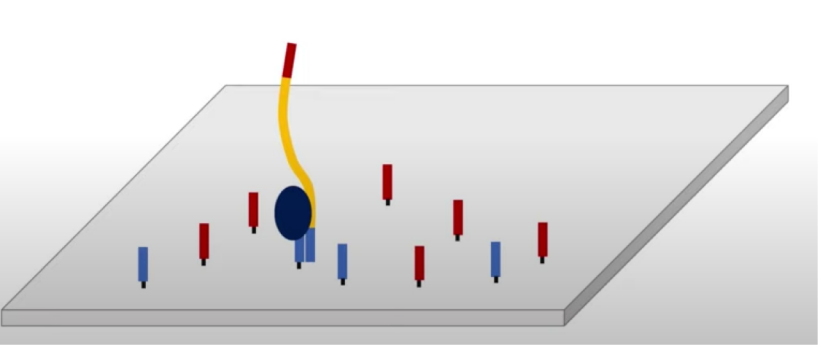
REMEMBER: Our adapters have hybridization seqs designed to bind the chip

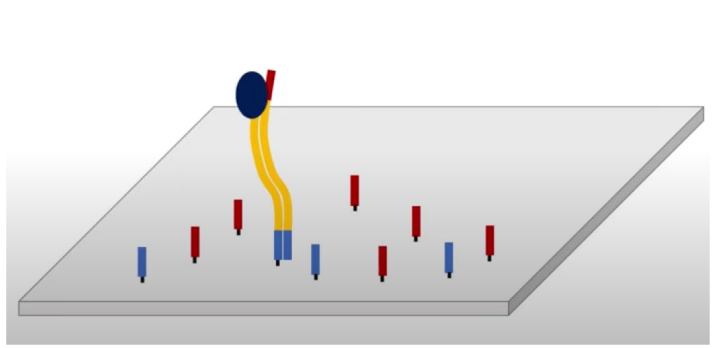


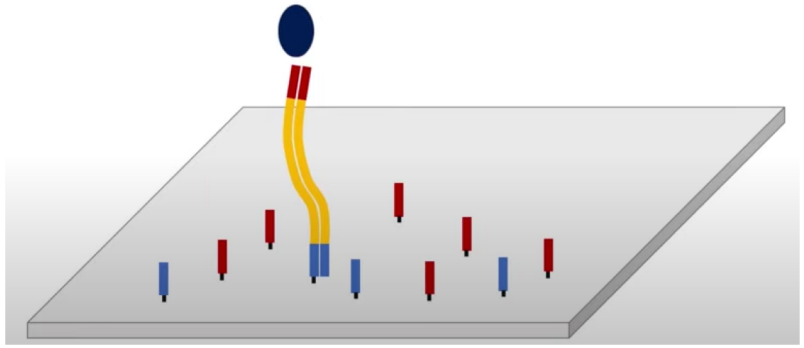
Images @HenriksLab

Uniquely barcoded fragments are pooled any ready to interact with our chip

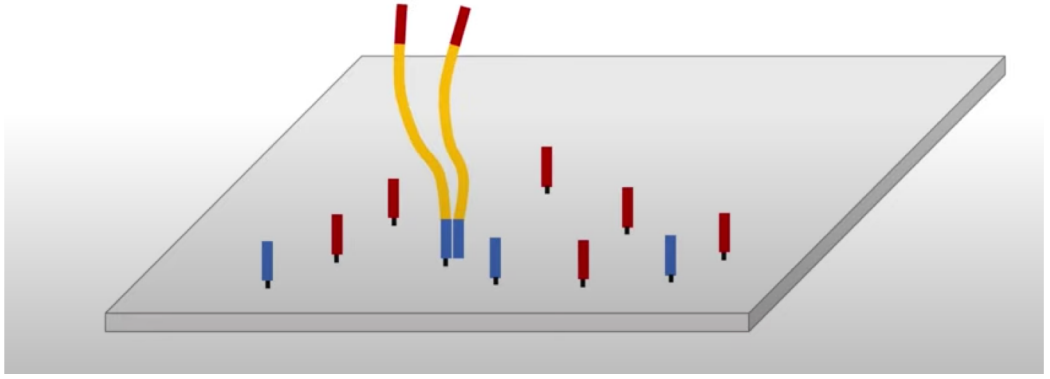
Step 3: PCR amplification of the bound fragment

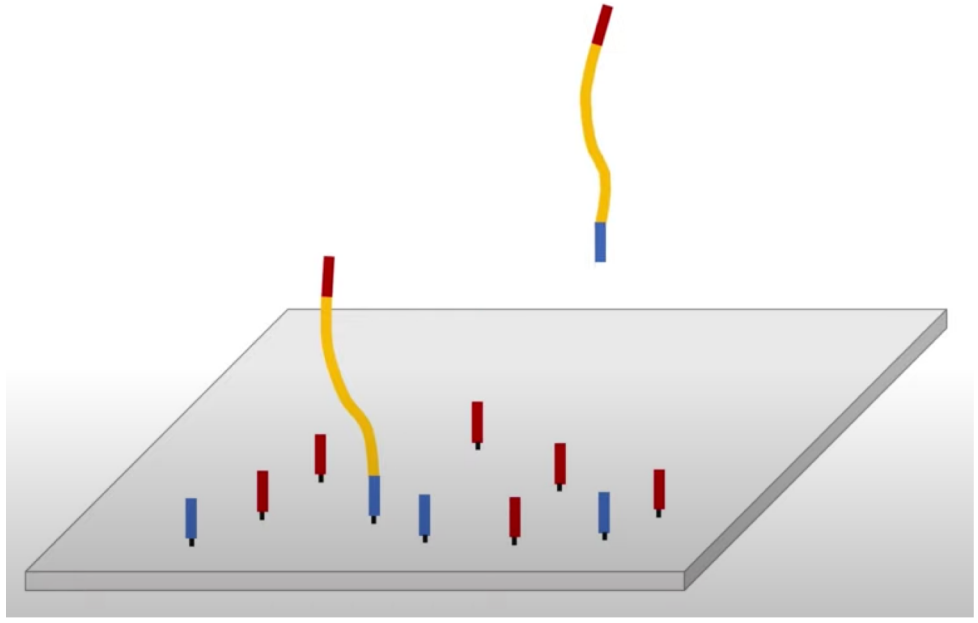






Step 4: Denature and wash



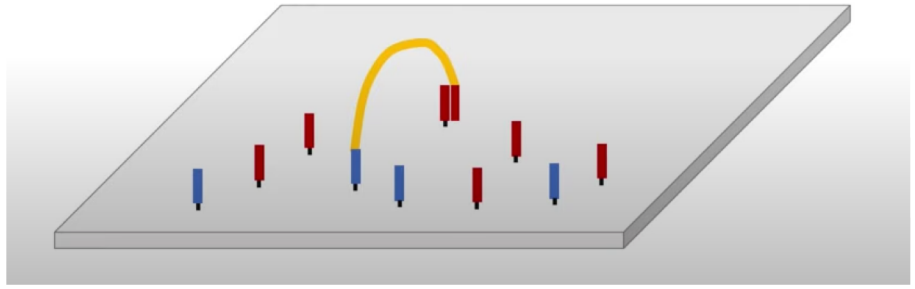


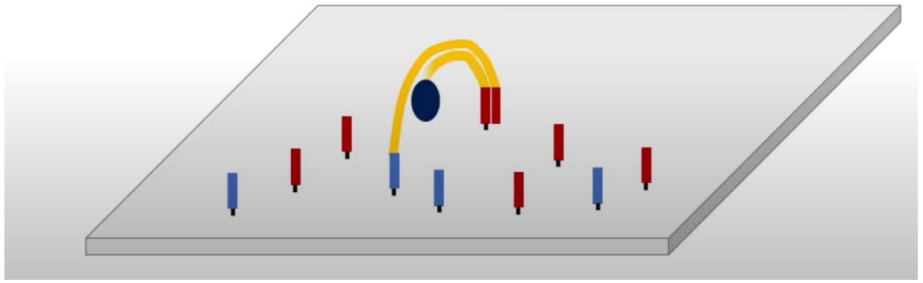
Images @HenriksLab

Uniquely barcoded fragments are pooled any ready to interact with our chip

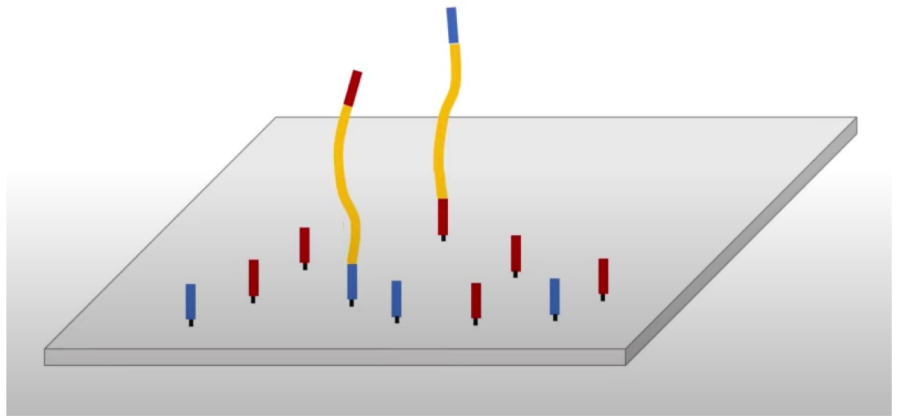
Step 5: Bridge Building and Bridge Amplification

REMEMBER: Our we have two types of adapters





Step 4: Denature and wash



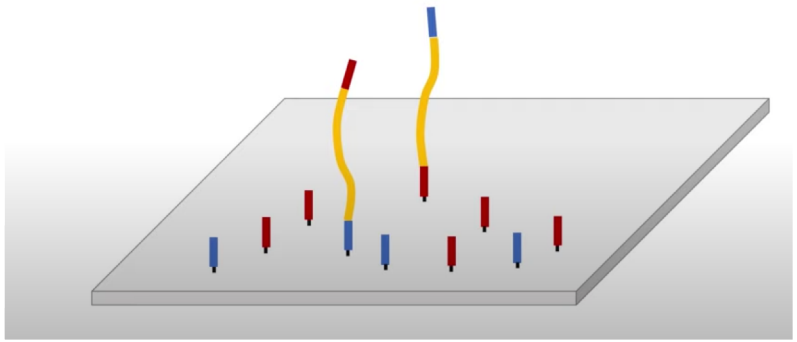
Now both strands are adhered to the chip

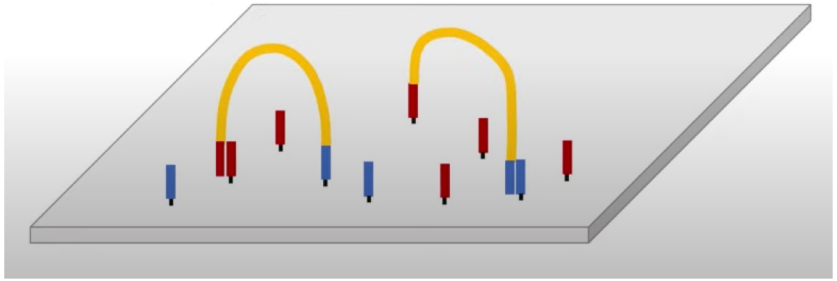
Repeat Repeat Repeat!

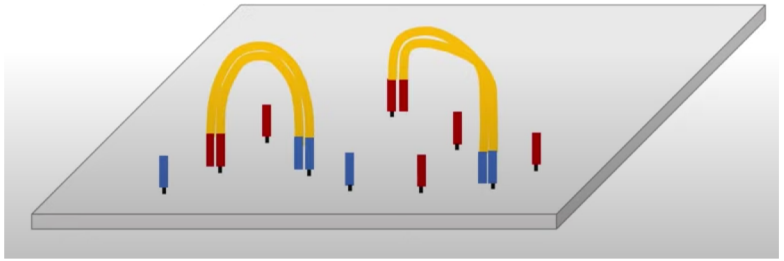
Images @HenriksLab

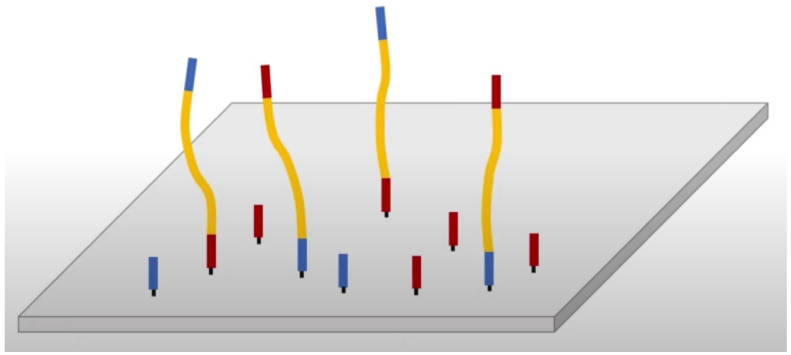
Uniquely barcoded fragments are pooled any ready to interact with our chip

Repeat Repeat Repeat!

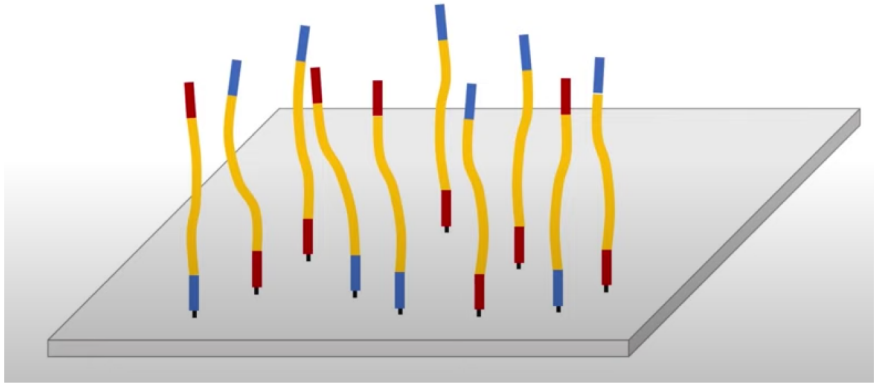








… eventually

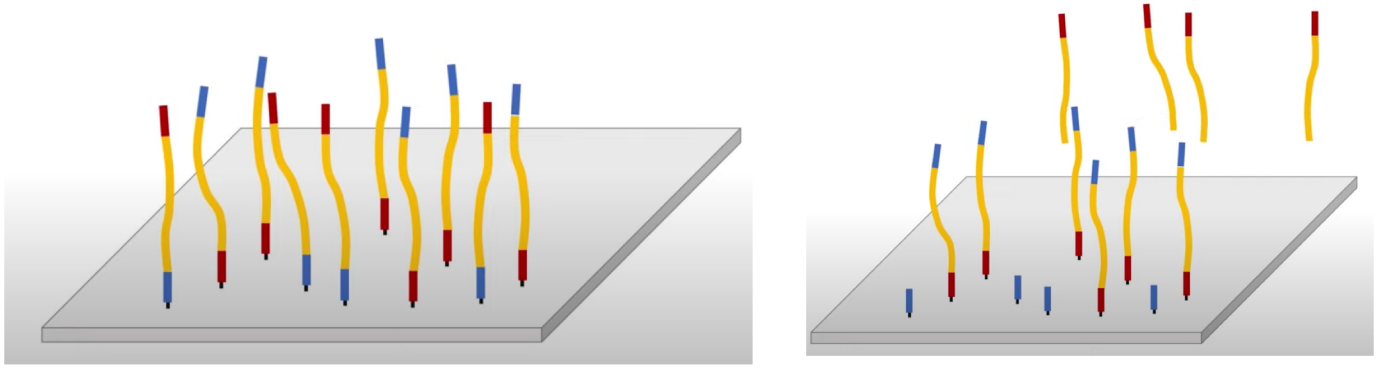


Lots of copies (forward and reverse) from a single template

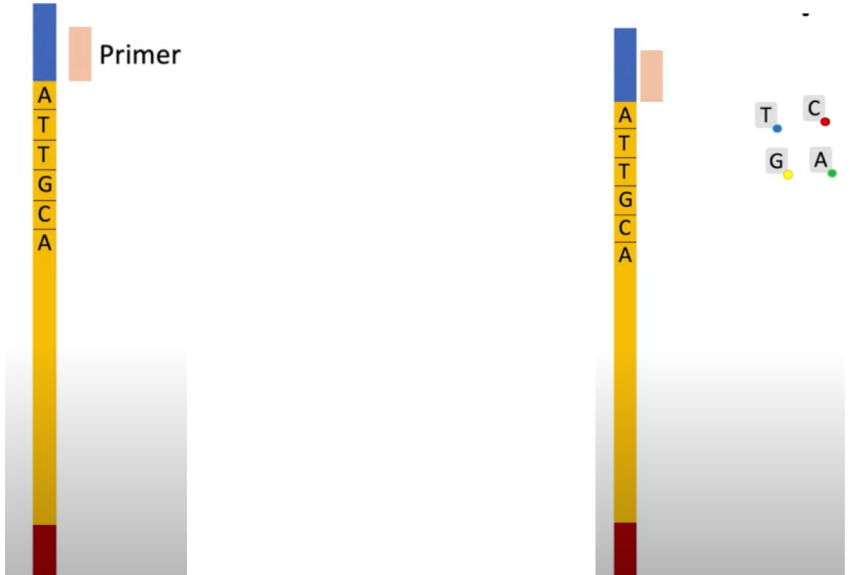
Images @HenriksLab

Uniquely barcoded fragments are pooled any ready to interact with our chip

One strand is cleaved then sequencing begins!



REMEMBER: Our adapters have primer recognition seqs



You can use Illumina sequencing to generate reads of different lengths 50-250bp depending on your needs

Images @HenriksLab

Newer technologies have made further alterations to this general protocol

Many employ **isothermal amplification chemistry** removed the need for traditional PCR temperature cycling called exclusion amplification (**ExAmp**)

Novaseq has moved away from bridge amplification entirely with a new type of Chip and **requires** the use of dual-indexed barcodes to prevent index switching.

Novaseq is also biased toward sequencing small fragments so you need a tighter range of fragment sizes after shearing

Long Read Sequencing: PacBio

Generating Long Reads can be useful to address certain biological questions, but there are problems with long read data including high error rates (later)

Let's Start with discussing how PacBio has modified what we just discussed to increase sequencing read length

Step 1: Obtain Template Material from your sample

how you do this matters!

- You may need high molecular weight DNA
- You may need to account for kit contaminations (16S-seq)

Consult! Consult! Consult!

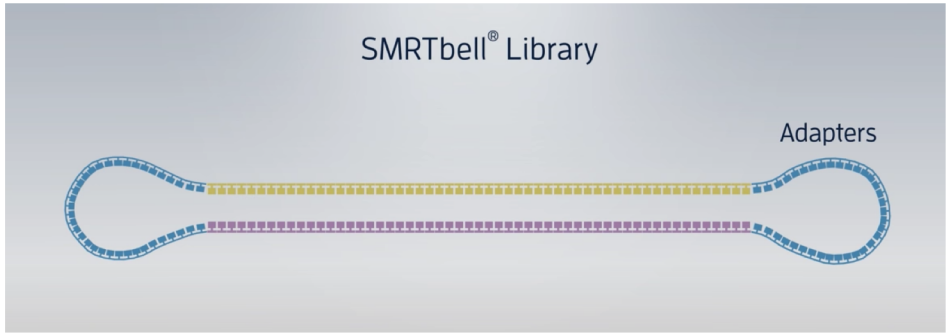
For Long Read Data you need high molecular weight DNA!!!!

Long Read Sequencing: PacBio

Generating Long Reads can be useful to address certain biological questions, but there are problems with long read data including high error rates (later)

Let's Start with discussing how PacBio has modified what we just discussed to increase sequencing read length

<u>Adapters are different for PacBio</u>



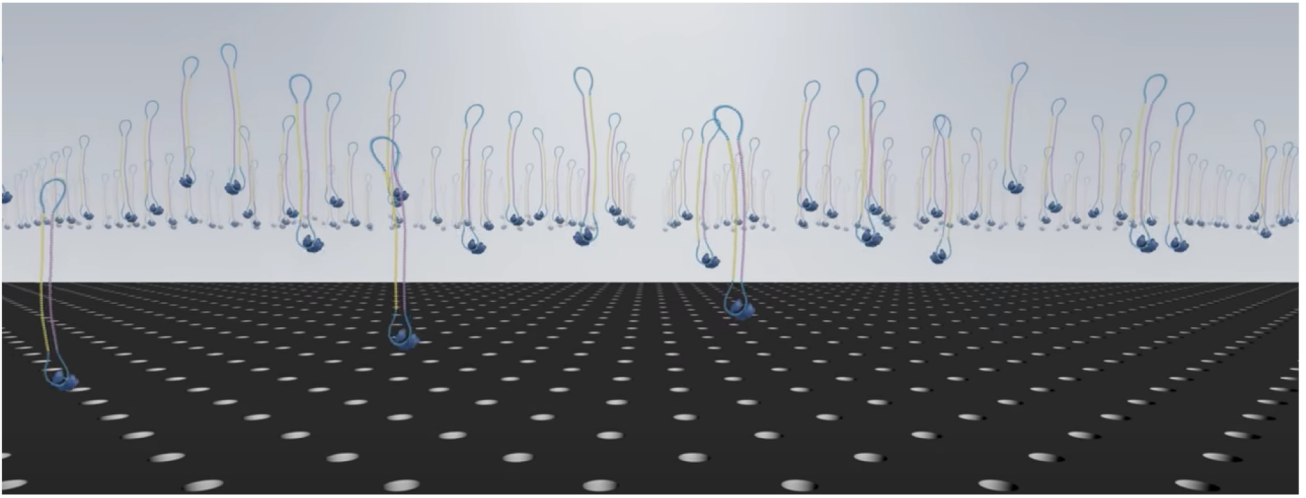
Generate circular templates containing both strands to be combined with a SMRT Cell

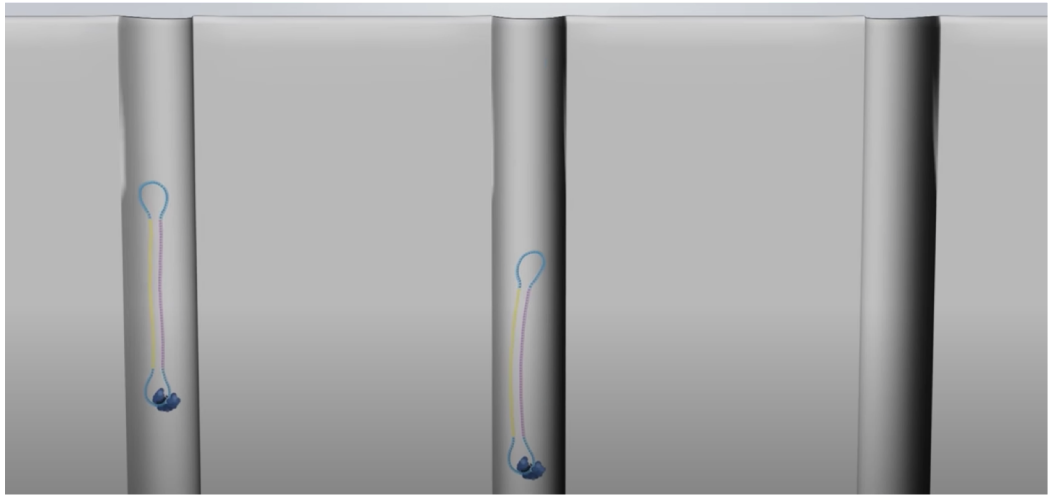
Long Read Sequencing: PacBio

Generating Long Reads can be useful to address certain biological questions, but there are problems with long read data including high error rates (later)

Let's Start with discussing how PacBio has modified what we just discussed to increase sequencing read length

<u>SMRT Cells are used in place of Illumina Chips</u>



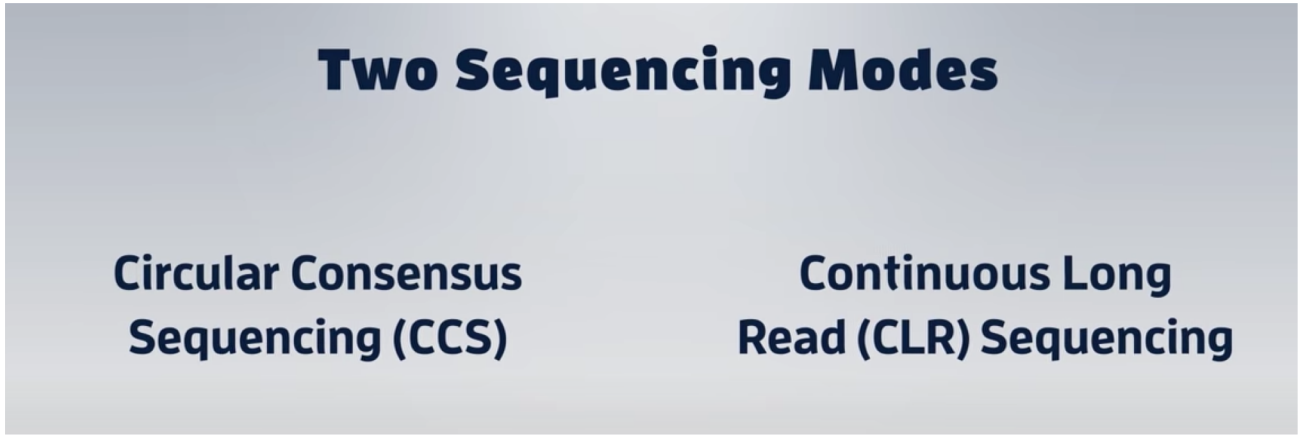


Long Read Sequencing: PacBio

Generating Long Reads can be useful to address certain biological questions, but there are problems with long read data including high error rates (later)

Let's Start with discussing how PacBio has modified what we just discussed to increase sequencing read length

<u>Two sequencing modes can be engaged</u>



shorter but more accurate (HiFi)

Longer but less accurate

Both use light emission to recognize nucleotides to get readouts in real time

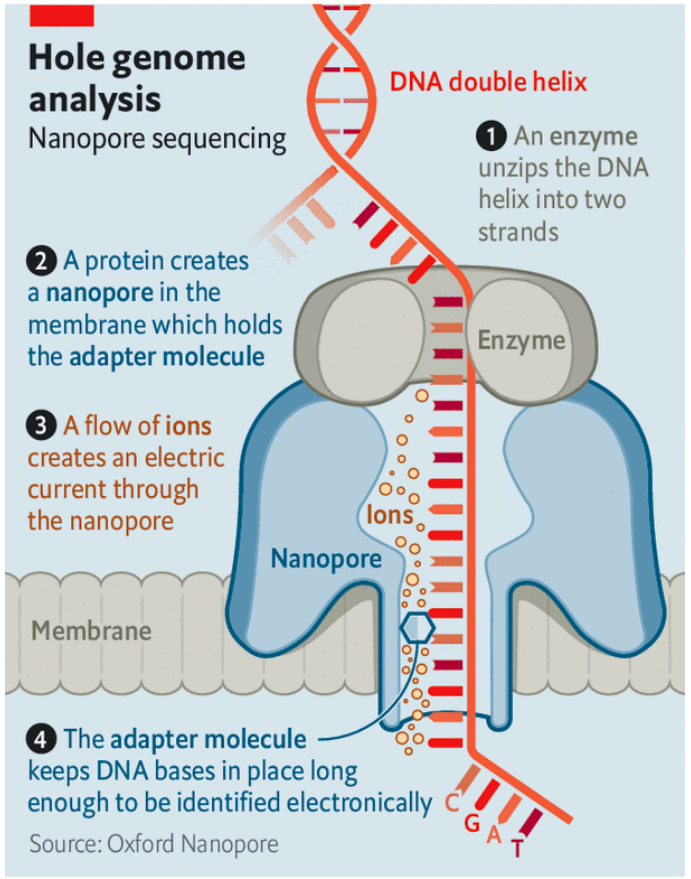
Long Read Sequencing: Nanopore

Generating Long Reads can be useful to address certain biological questions, but there are problems with long read data including high error rates (later)

NanoPore also generates long reads but using a different readout method

<u>Instead of relying on light emissions Nanopore is based on the principle that each nucleotide has different electrical properties</u>

Uses hairpin adaptors to the template (similar to PacBio)
Bases are measured by changes in current as they pass through an electrical resistant membrane in nanopore
Very fast, long reads, high error rate.



The Economist

Later we will lean how to handle large files of multiplexed reads from different individuals

Let’s take a minute as a group and write down questions you may have about the types of data you can generate!