Next Generation DNA Sequencing Technology & Data

BIOL 435/535: Bioinformatics

March 22, 2022

Illumina Sequencing Platforms

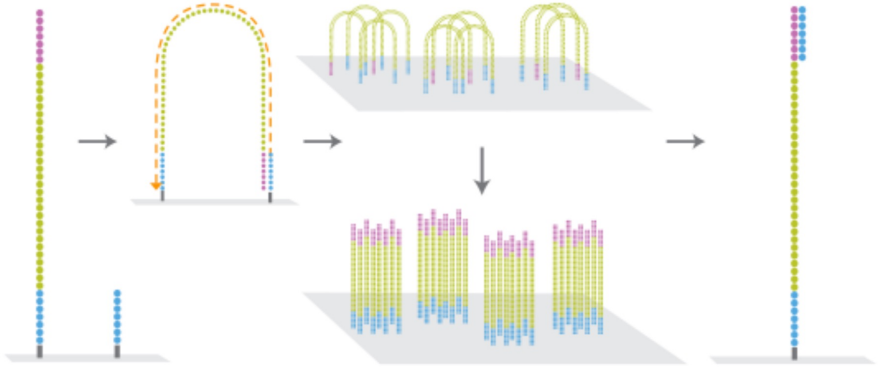
	iSeq	MiniSeq	MiSeq	NextSeq	HiSeq	NovaSeq
Max Yield	1.2 Gb	7.5 Gb	15 Gb	120 Gb	1800 Gb	6000 Gb
Max Length	150 bp	150 bp	300 bp	150 bp	150 bp	250 bp

- Illumina sequencers differ predominantly in amount of output (and cost) per run.
- The same sequencing library will generally work on all instruments.

Slide courtesy of Dan Sloan

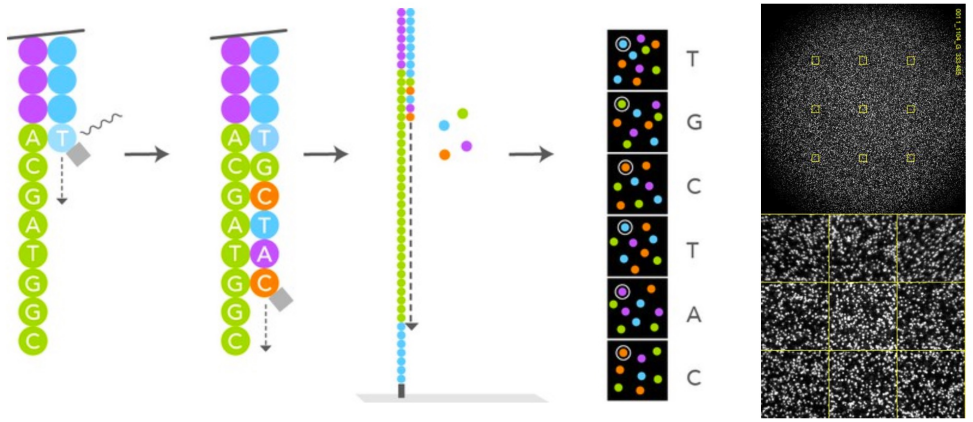
Illumina Sequencing by Synthesis

Cluster Generation



Sequencing libraries are loaded on a flow cell, where each library molecule seeds a cluster and is amplified into thousands of clonal copies by either bridge PCR or “exclusion amplification” (ExAmp).

Sequencing by Primer Extension

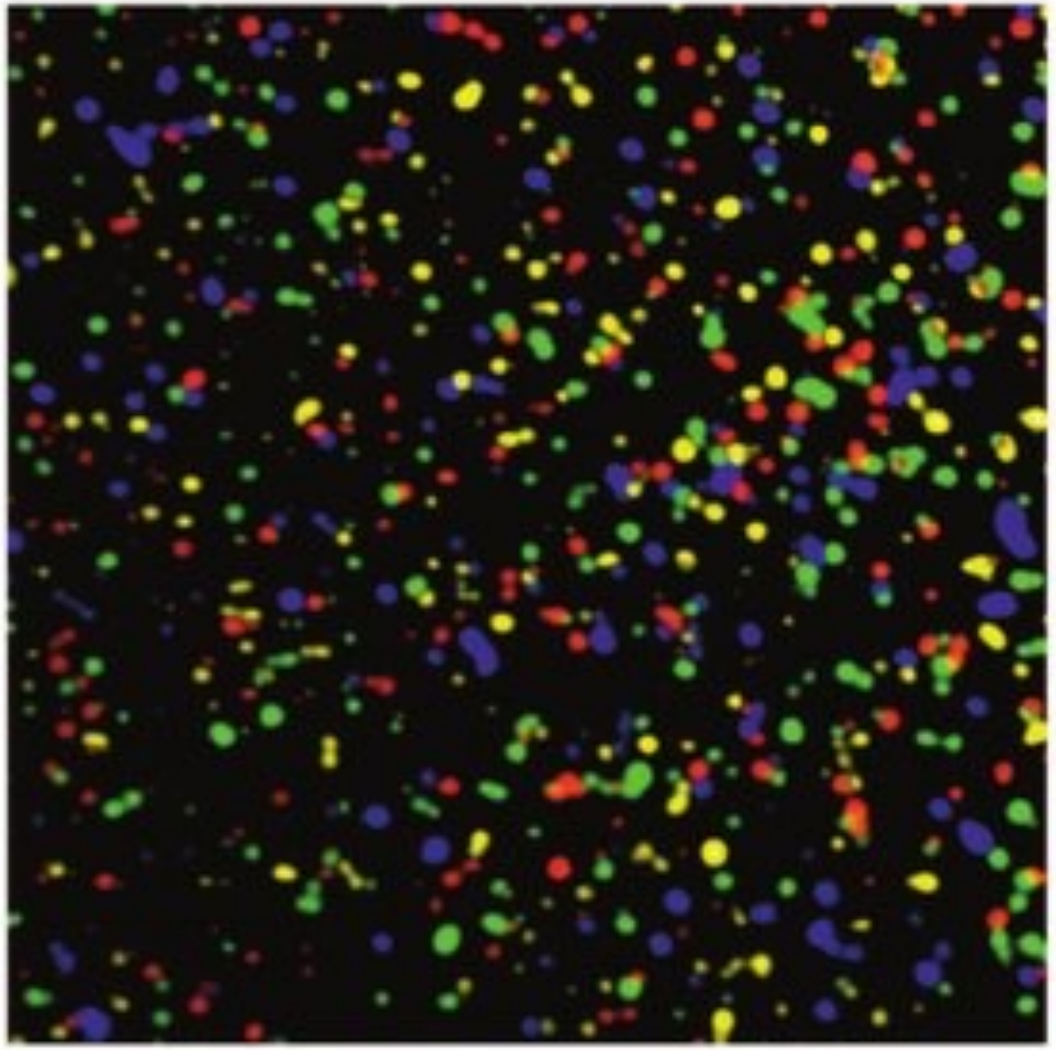


Sequencing proceeds by extending a primer one base at a time (a “cycle”) using reversible chain-terminating and fluorescently labeled nucleotides. The flow cell is imaged after each cycle before proceeding to the next base. Illumina instruments use either 4-color, 2-color, or 1-color chemistry.

https://www.youtube.com/watch?v=fCd6B5HRaZ8

Slide courtesy of Dan Sloan

Illumina Sequencing by Synthesis



Illumina adapters and library molecule structure

An Illumina library molecule

The four functions of Illumina adapters

- Library amplification (PCR primer binding)
- Flow cell binding
- Index sequences (barcodes) for multiplexing
- Sequencing primer binding

Slide courtesy of Dan Sloan

Illumina adapters

1. Library Amplification

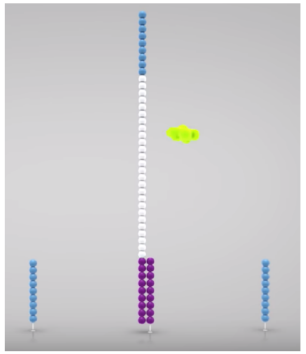
- Most library construction protocols include a PCR amplification step (in addition to the amplification that occurs on the flow cell during cluster generation).
- Adapters provide universal sequences such that all library molecules can be amplified with a common set of primers.

Slide courtesy of Dan Sloan

Illumina adapters

2. Flow Cell Binding

- The same adapter regions used for initial library amplification are also complementary to the oligos that are anchored to the surface of Illumina flow cells.
- Flow cell binding enables subsequent cluster generation.



Slide courtesy of Dan Sloan

Illumina adapters

3. Index Sequences (Barcodes) for Multiplexing

- “Multiplexing” involves pooling libraries from different biological samples to be sequenced together on the same flow cell.
- The i5 and i7 index sequences are barcodes that are shared by all molecules from the same library so that libraries can be distinguished from each other during data analysis.

Slide courtesy of Dan Sloan

Illumina adapters

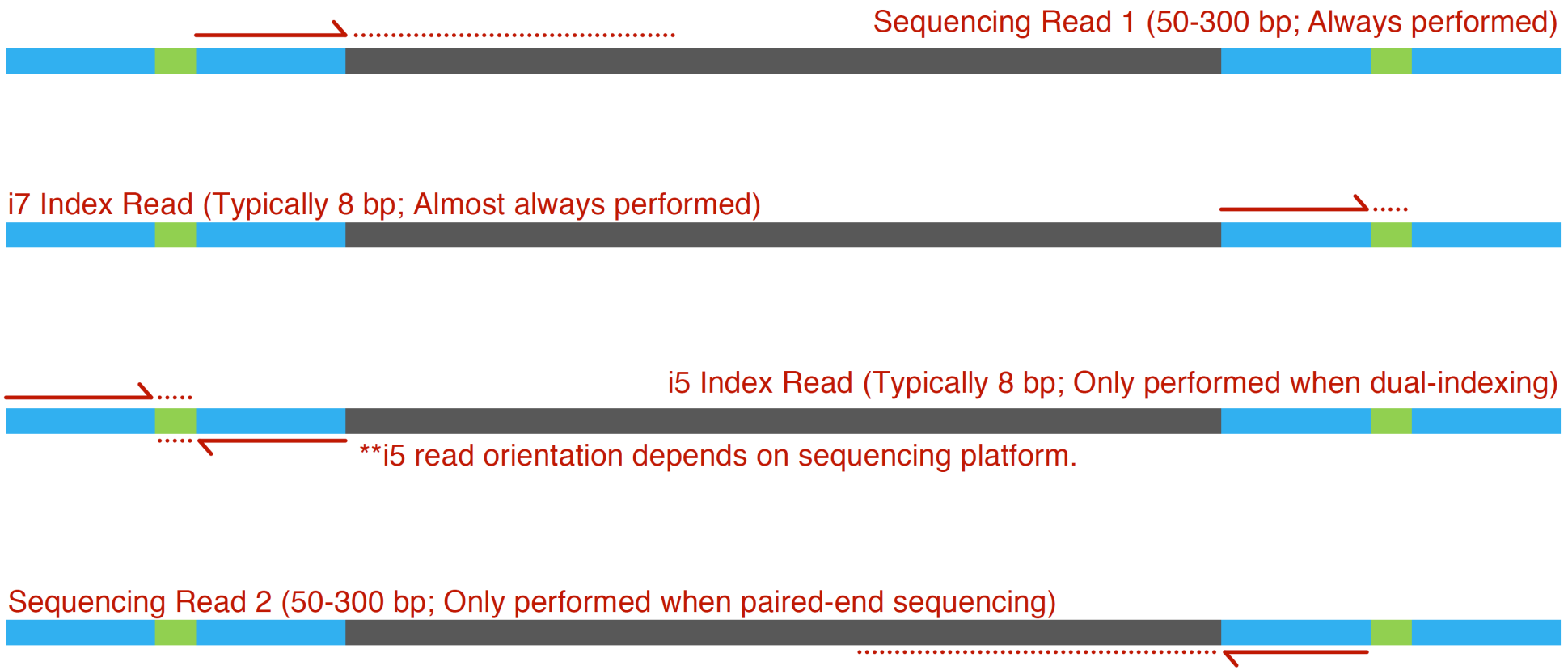
4. Sequencing Primer Binding

- The Sequencing by Synthesis (SBS) process is initiated by primers that bind to specific regions of the Illumina adapters.
- A second sequencing read is initiated from the other sides of the insert when performing paired-end sequencing.

Slide courtesy of Dan Sloan

Illumina reads

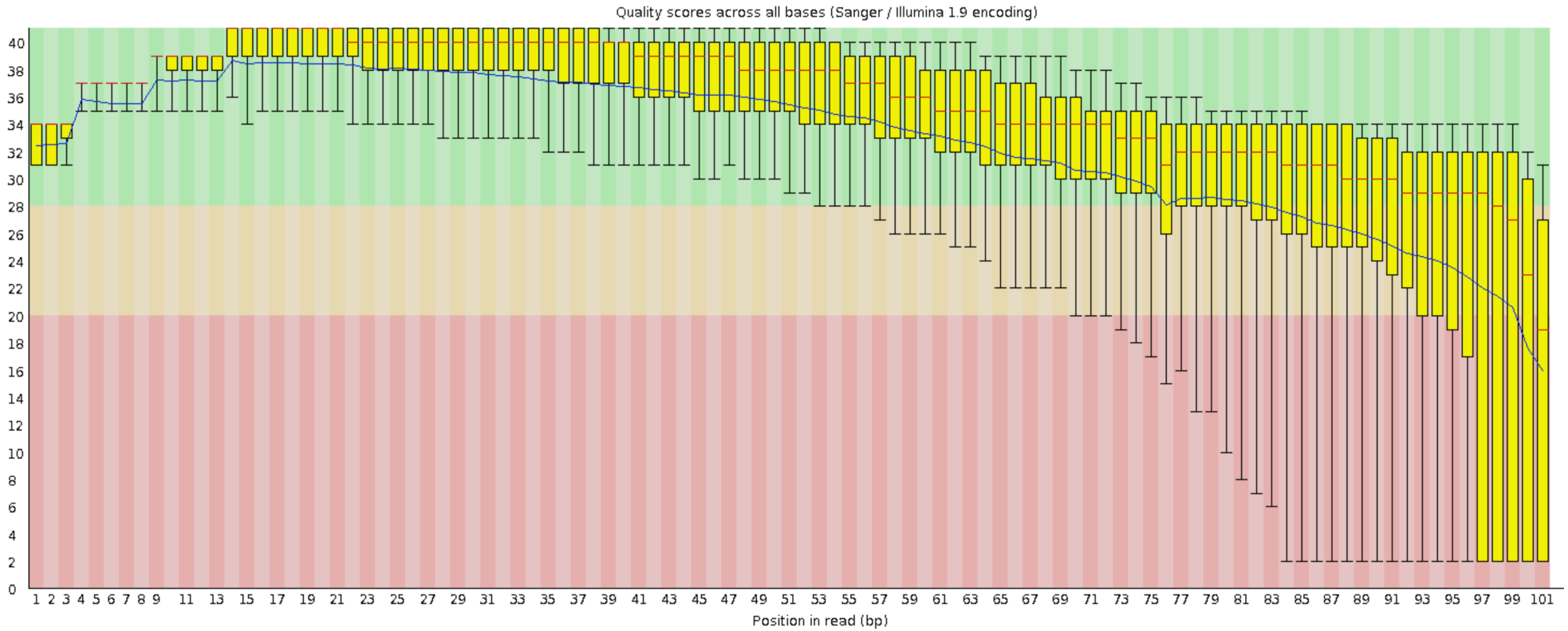
An Illumina run will actually produce up to four “reads” per molecule.



Slide courtesy of Dan Sloan

Read quality

Quality declines with increasing cycle number because amplicons within clusters get out of phase.



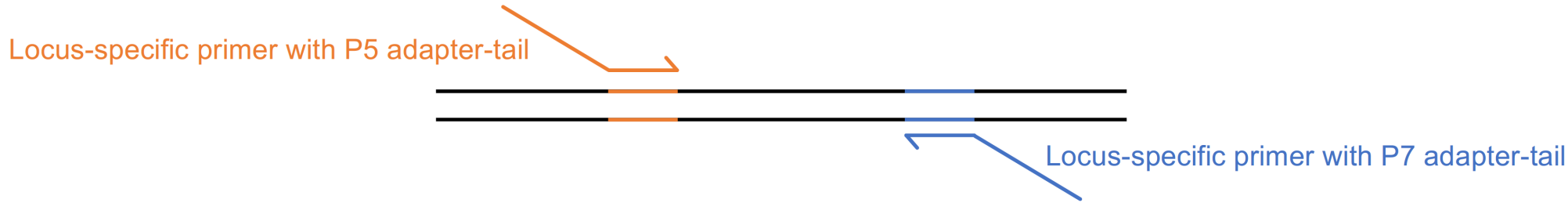
Slide courtesy of Dan Sloan

Methods to attach Illumina adapters

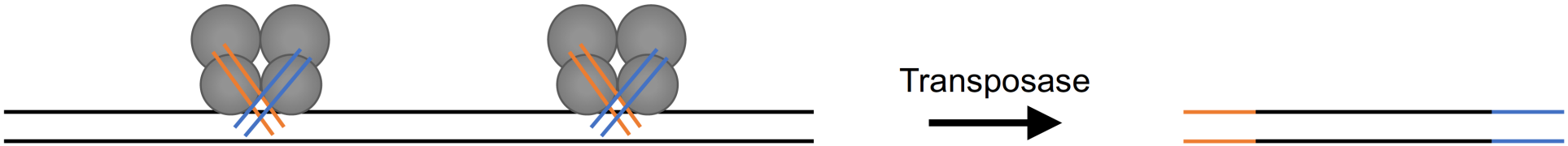
There are three primary ways to attach adapters to biological inserts.

1. Ligation

2. PCR Primers



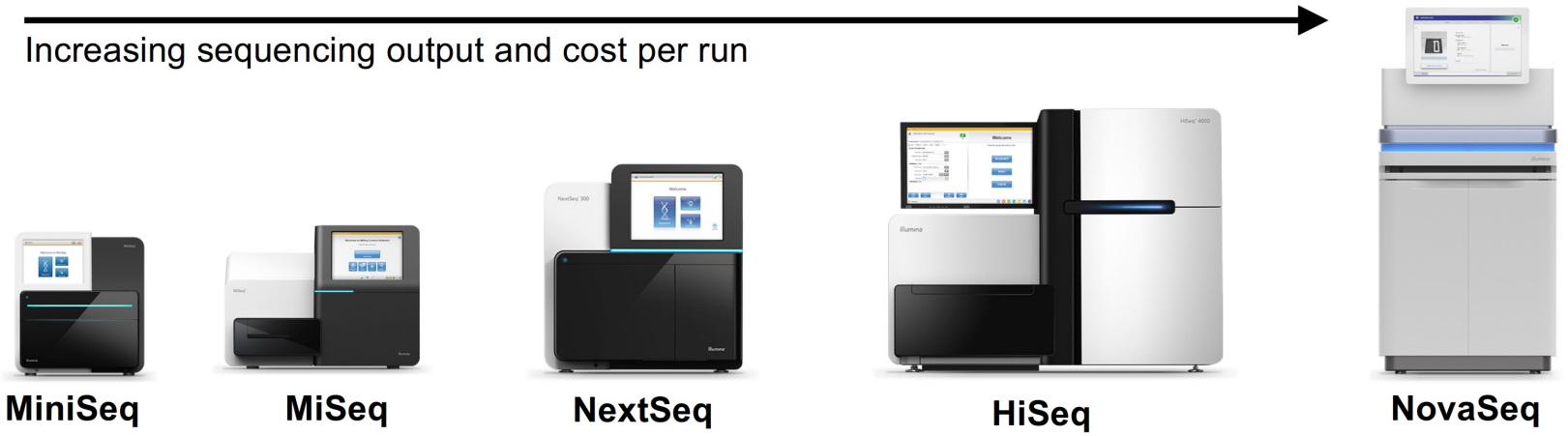
3. Tagmentation: Simultaneous fragmentation and adapter incorporation by transposase



Slide courtesy of Dan Sloan

Illumina sequencing platforms, costs, and outputs

	Clusters (millions)	Max Read-Length	Max Output (Gb)	Cost	Bacterial Genomes	Eukaryotic Transcriptomes
MiniSeq	25	150 bp	7.5	$1,500	15	1.5
MiSeq	25	300 bp	15	$1,530	30	3
NextSeq 500 (mid)	130	150 bp	40	$1,650	80	8
NextSeq 500 (high)	400	150 bp	120	$4,240	240	24
HiSeq 4000 Lane	300	150 bp	90	$1,925	180	18
NovaSeq S4 Lane	2500	150 bp	750	$6,000	1500	150



Slide courtesy of Dan Sloan

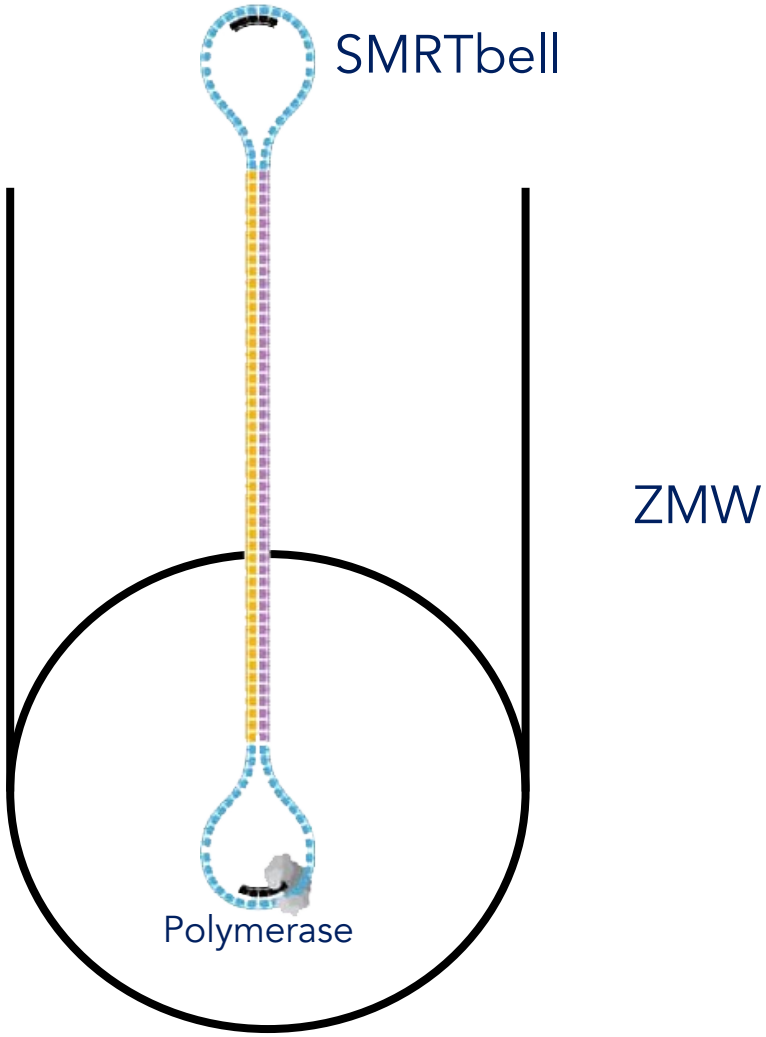
PacBio SMRTbells

Polymerase binds to SMRTbell (Single-Molecule Real-Time), performs sequencing-by-synthesis inside ZMWs

Fluorophore emits light at nucleotide incorporation

Movie for each ZMW is parsed to produce read calls

- 16hr, 20hr, 30hr movies

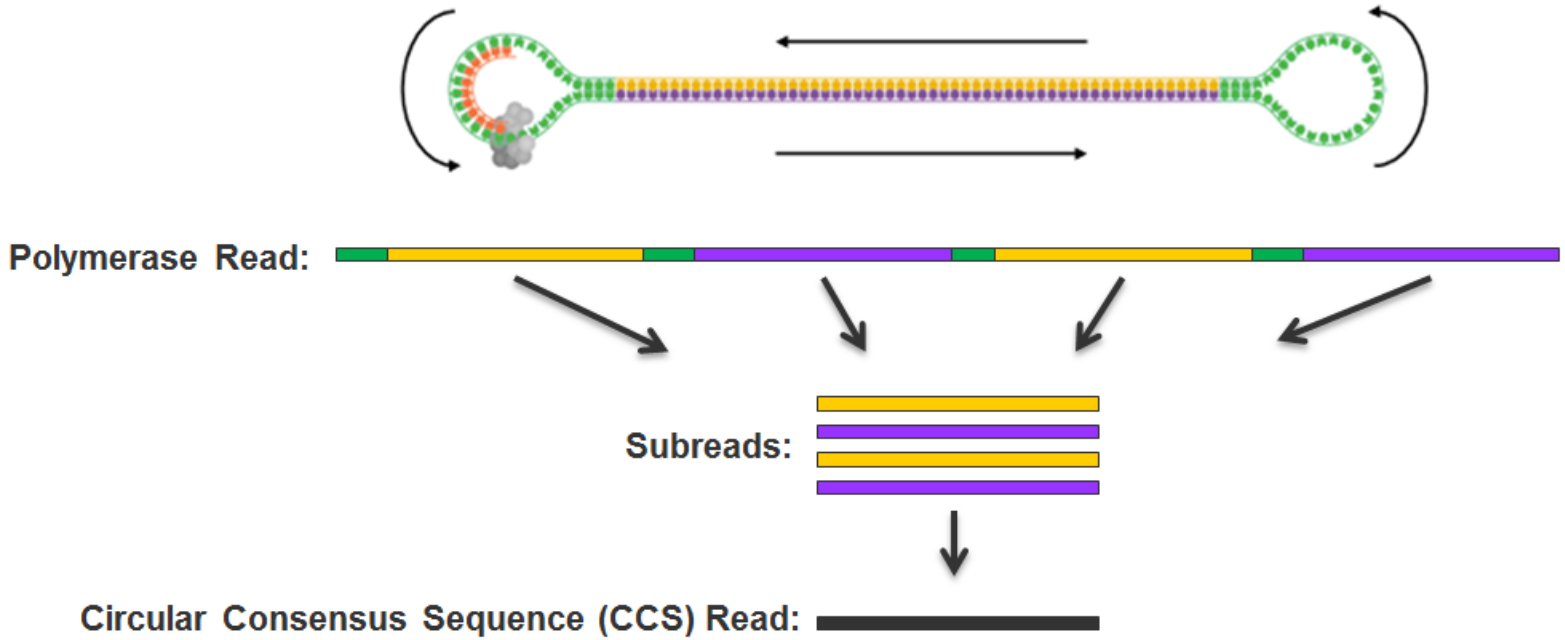


PacBio SMRTbells

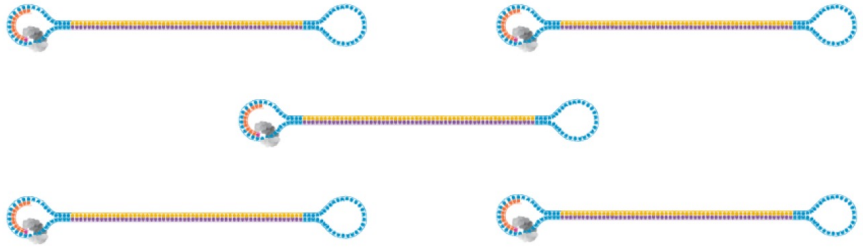
Circular Consensus Sequencing Reads (CCS Reads) are produced when polymerase goes around SMRTbell ≥3 times

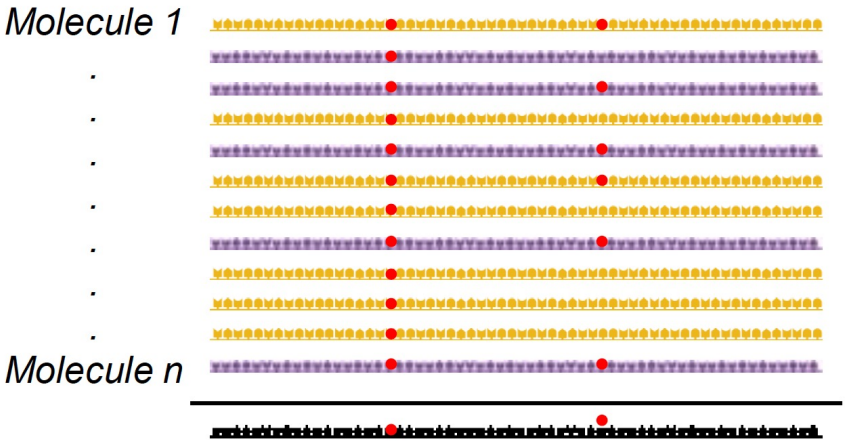
Can provide confidence for allele calling from single molecule, as a CCS read

Large inserts (≥50 kbp) are unlikely to form CCS reads

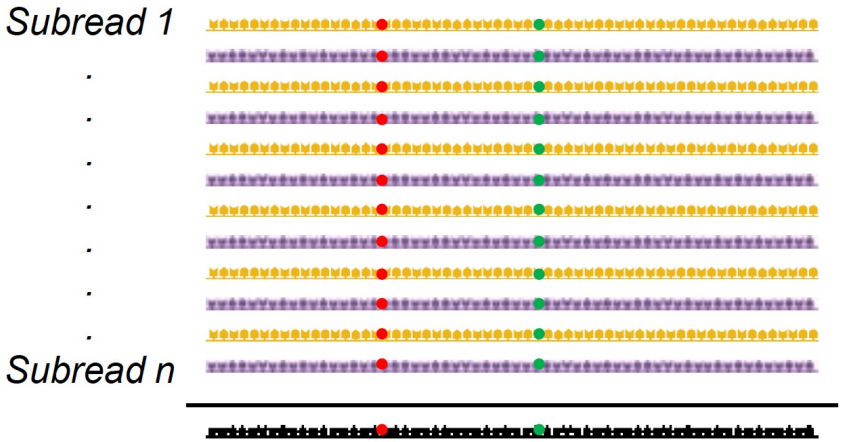


PacBio sequencing strategies





Long inserts, few CCS reads
De novo assembly

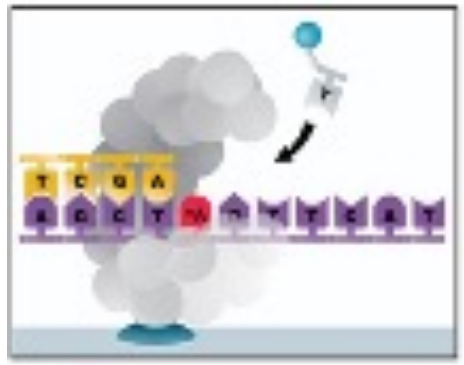


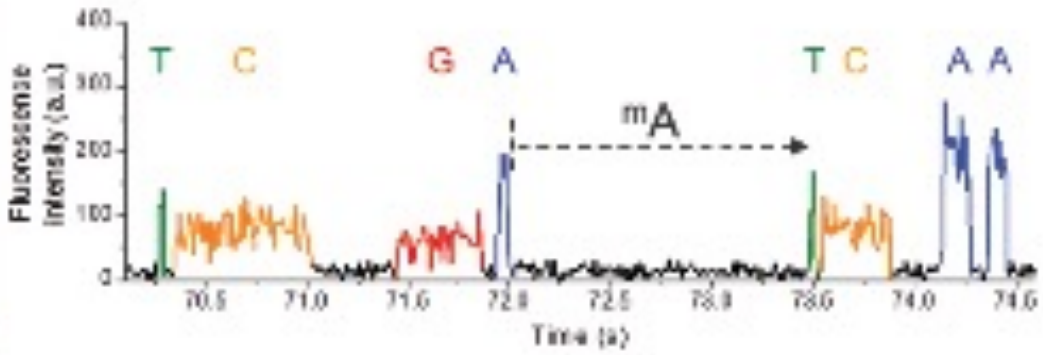
Short inserts, many CCS reads
Isoform Sequencing (Iso-Seq)

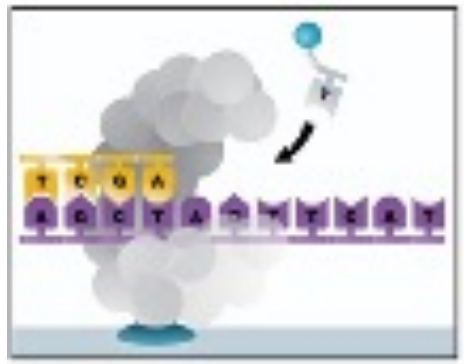
Detecting Base Modifications/Damage with PacBio SMRT bells

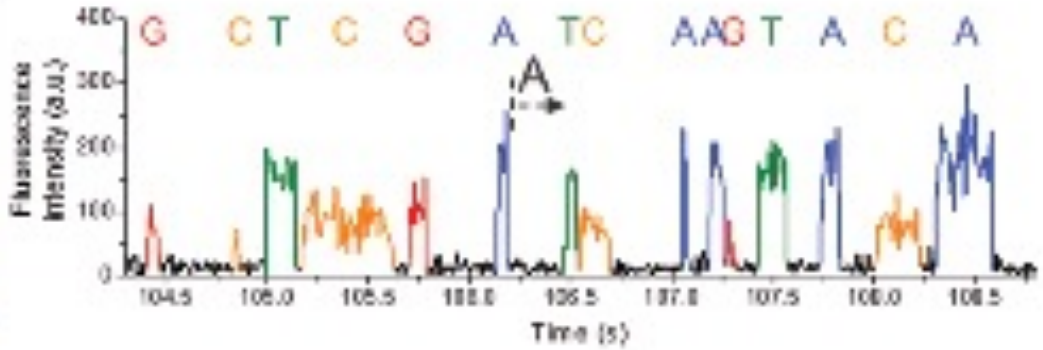
Base modifications impede polymerase processivity in a predictable manner

Can be measured with Inter-pulse Distance (IPD)

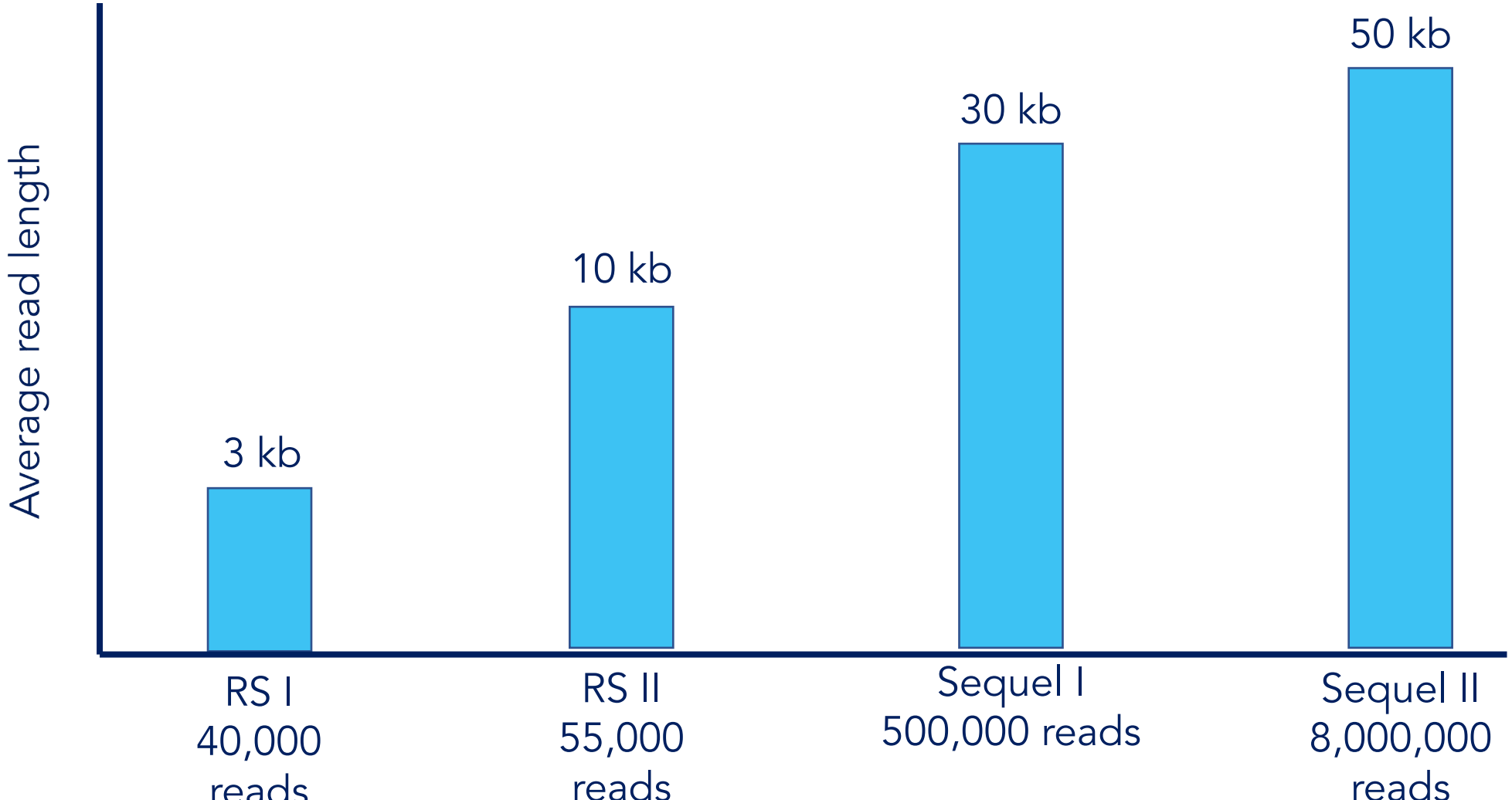








PacBio read length is increasing

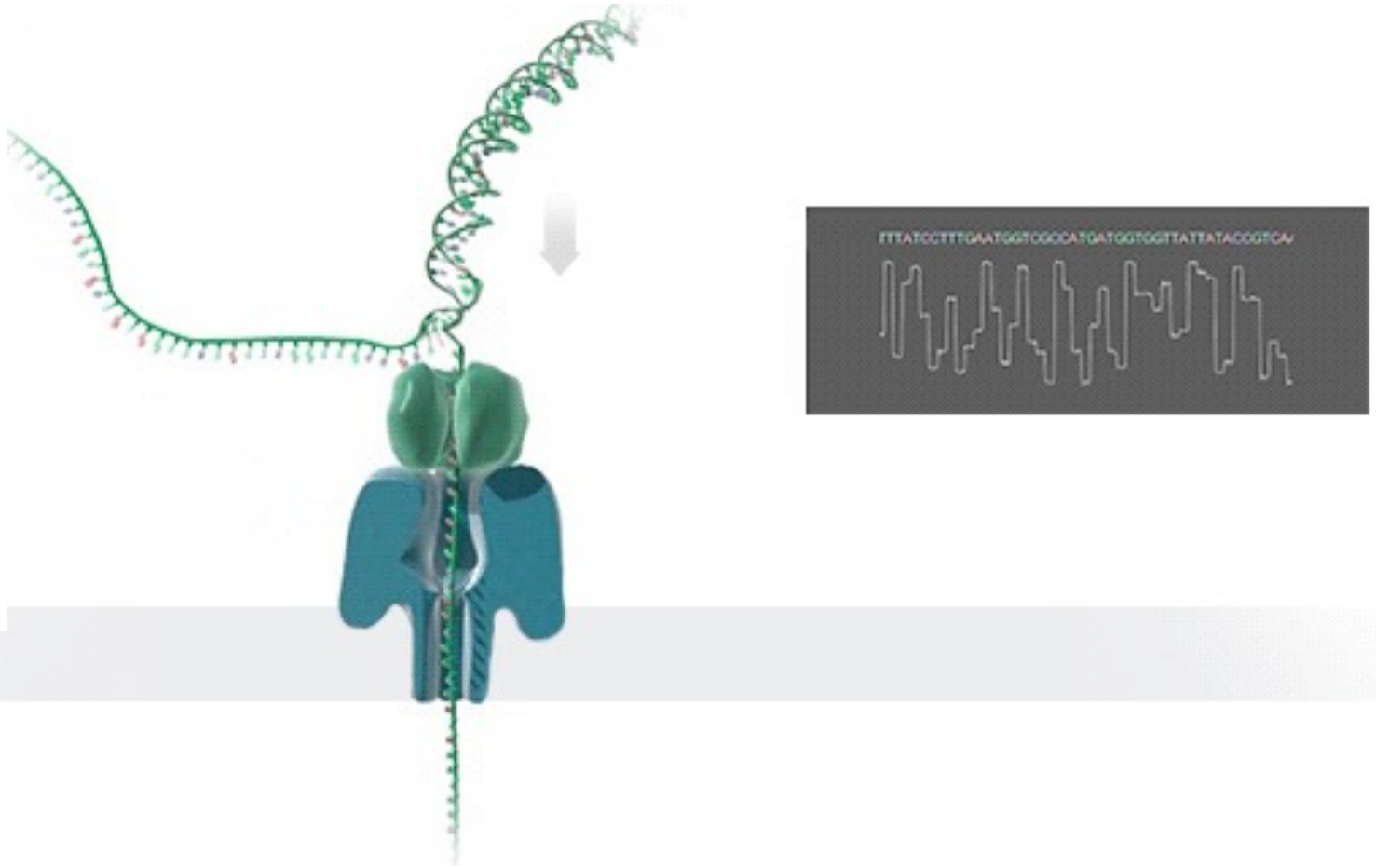


Oxford Nanopore

E. coli channel protein embedded in membrane nanopore

Double-stranded DNA is unwound and fed through a channel

Change in voltage across membrane measured by flow of ions through channel



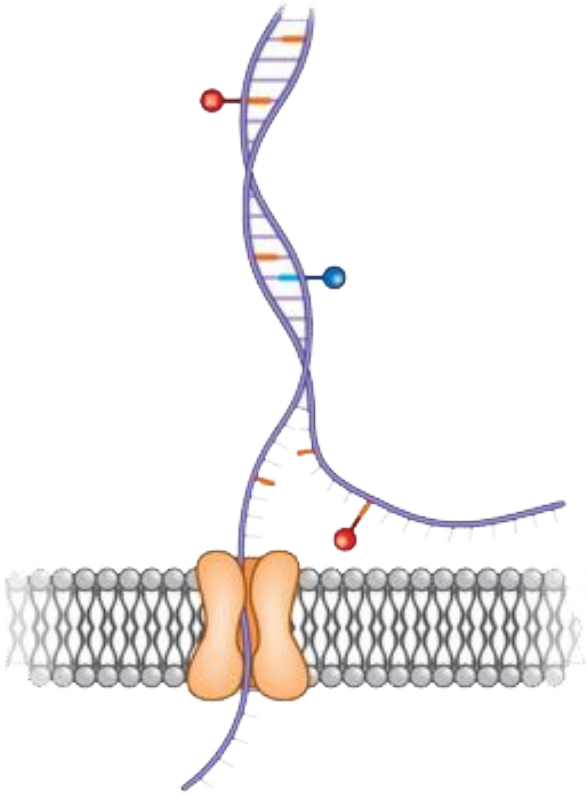
Oxford Nanopore

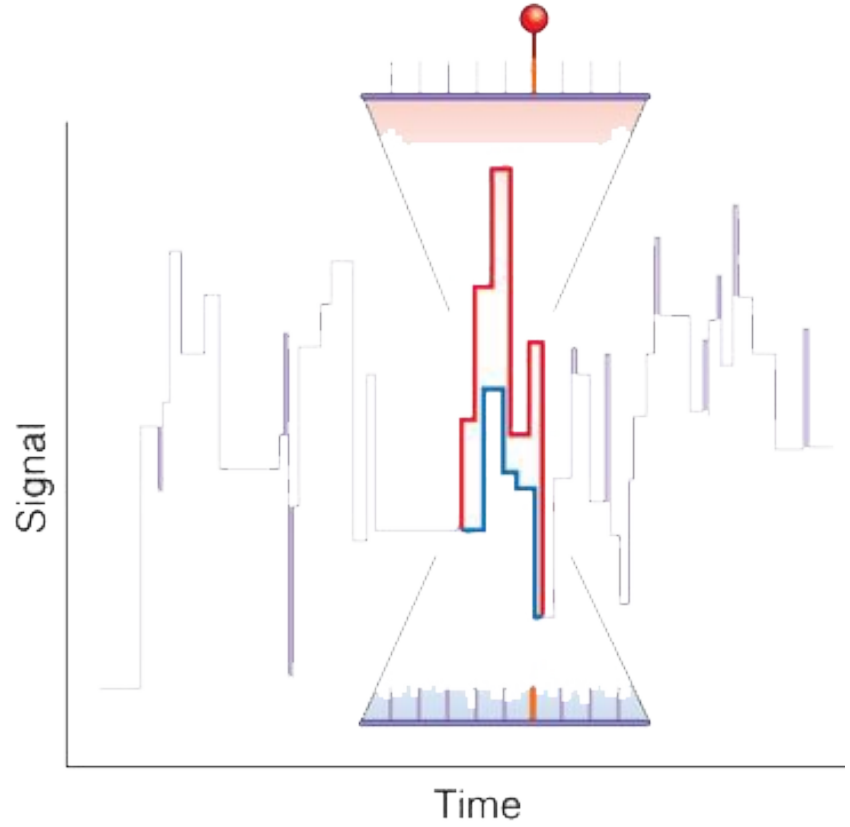
E. coli channel protein embedded in membrane nanopore

Double-stranded DNA is unwound and fed through a channel

Change in voltage across membrane measured by flow of ions through channel

The extent to which ssDNA blocks the flow of ions is the output signal

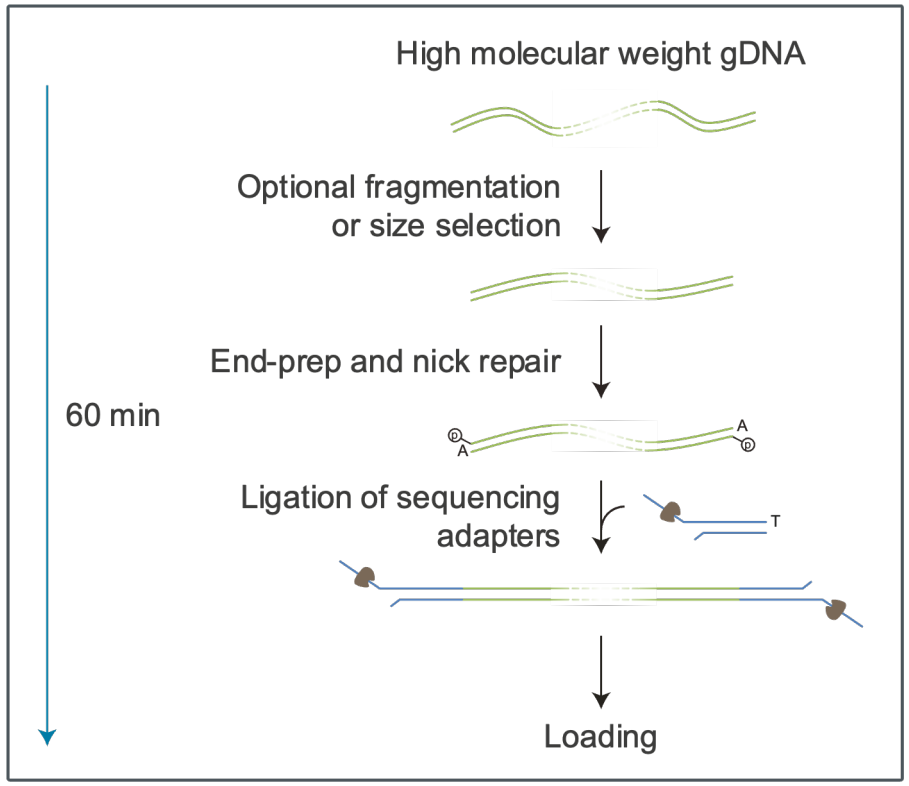




Oxford Nanopore Library Preps

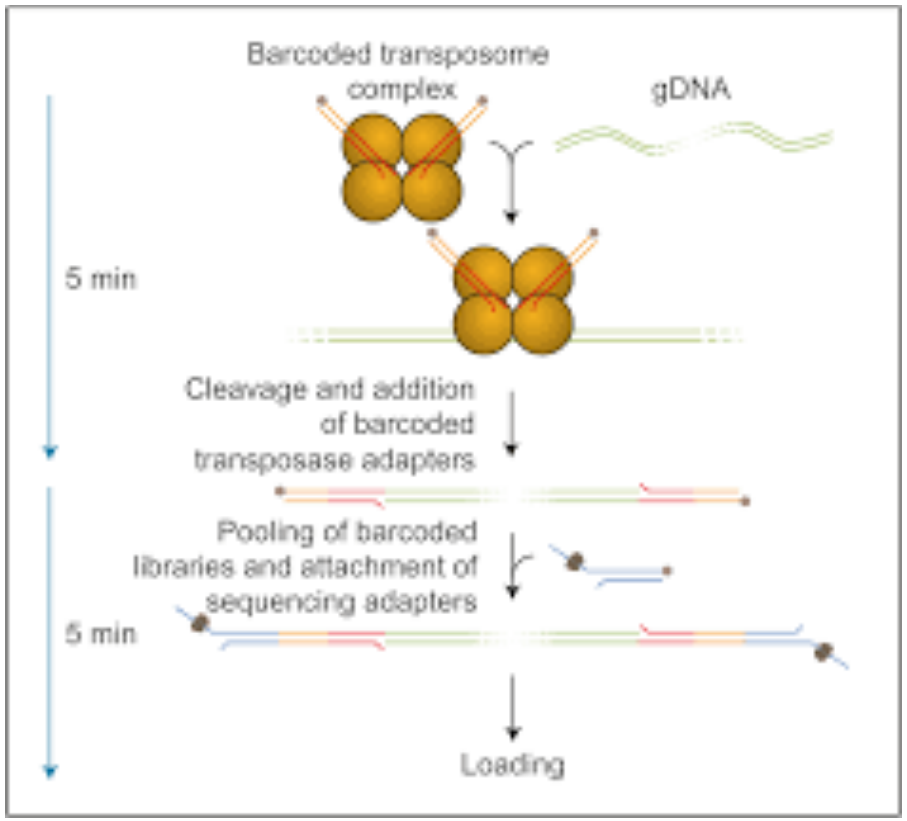
Ligation Prep

(longer reads, more prep time)



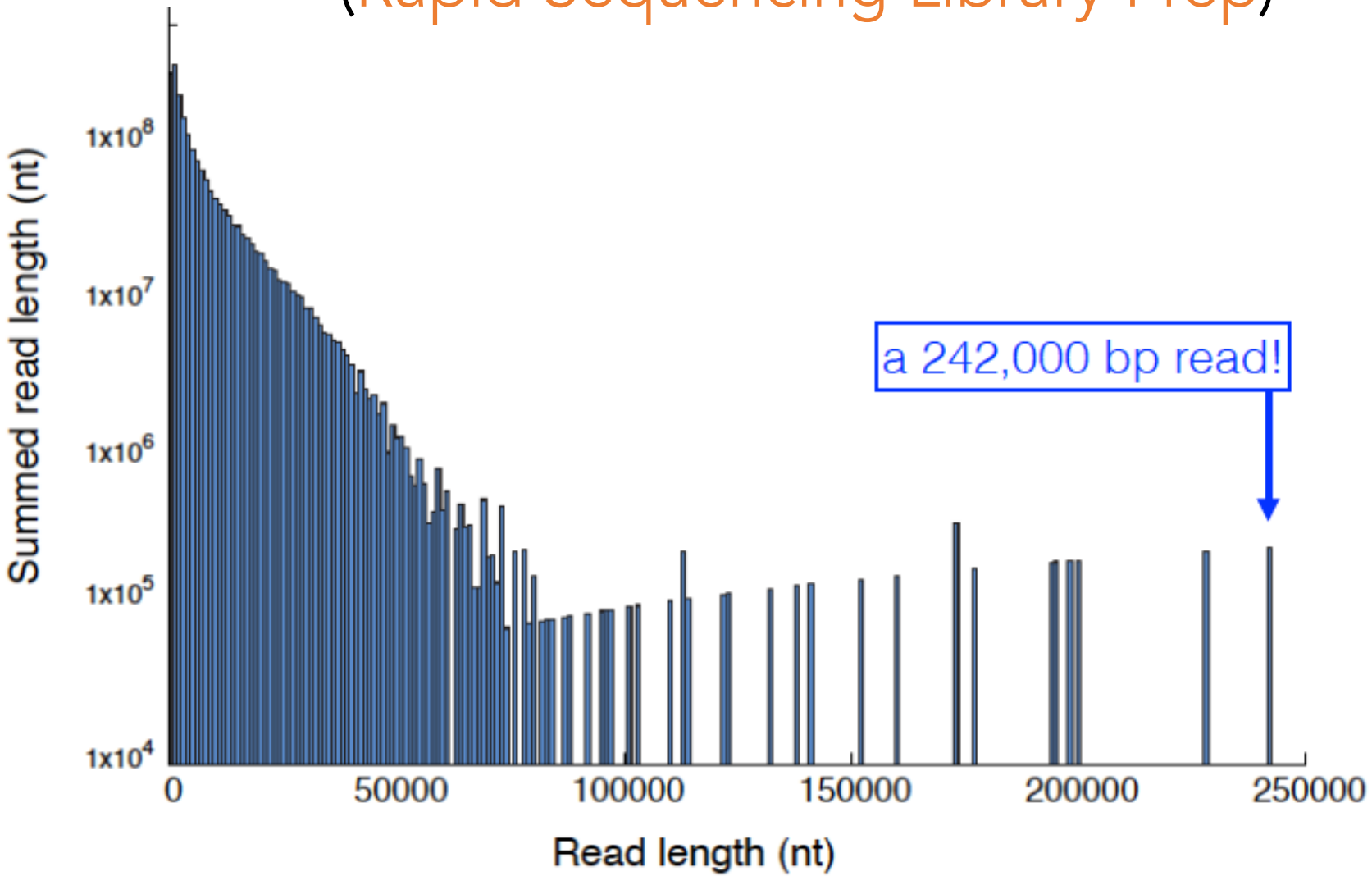
Rapid/Field Prep

shorter reads, less prep time



Nanopore read length distribution

(Rapid Sequencing Library Prep)



Oxford Nanopore Sequencing Platforms

SmidgION

-Will fill up your phone in seconds

Flongle

-Low-throughput (126 channels)
-Cheap
-Long queue

MinION

-Mid-throughput
30 Gb per flow cell
7-12 million reads
~$1000 starter kit
~$900 per flow cell after

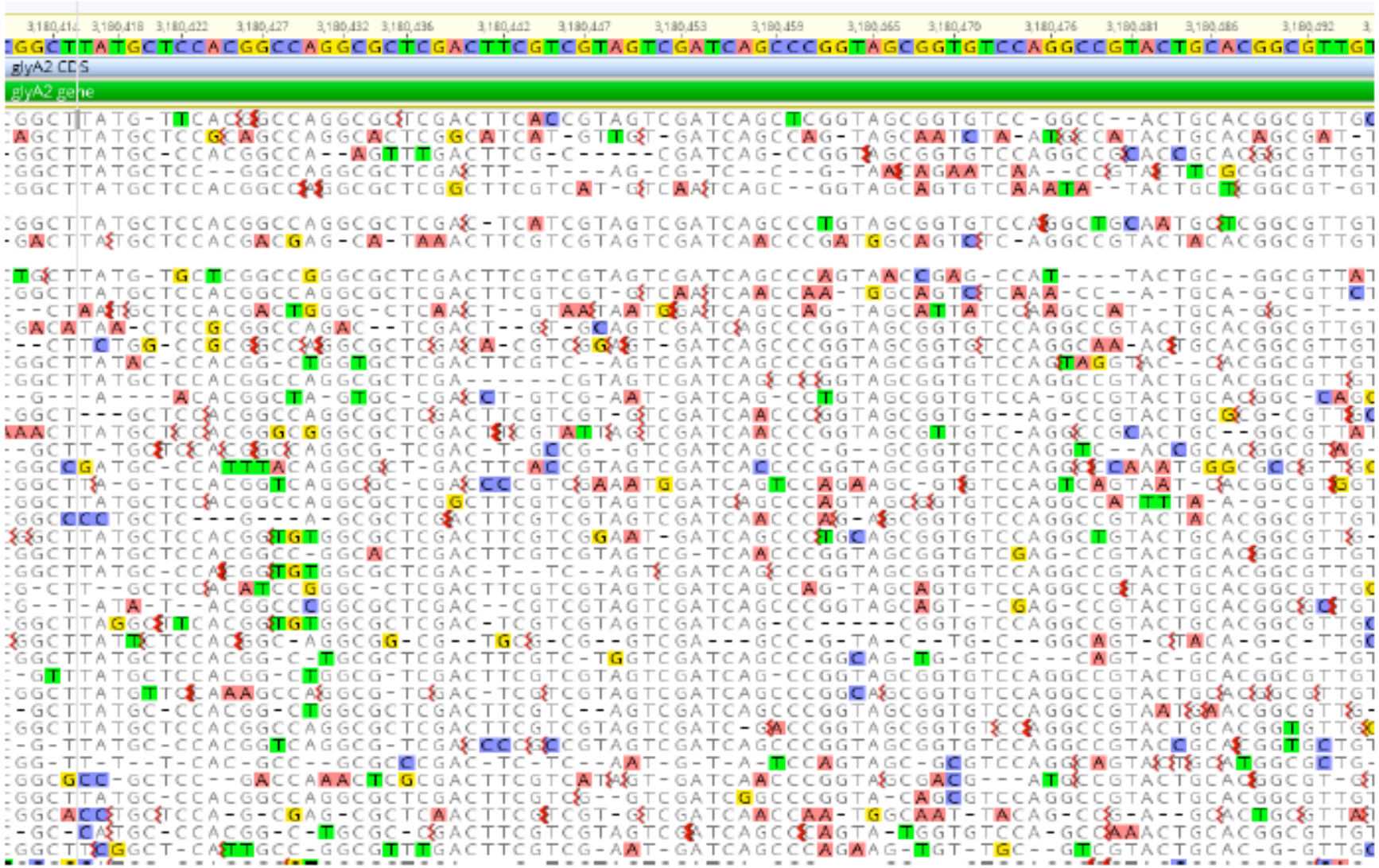
GridION

-Mid/high-throughput
5 x Flow Cells

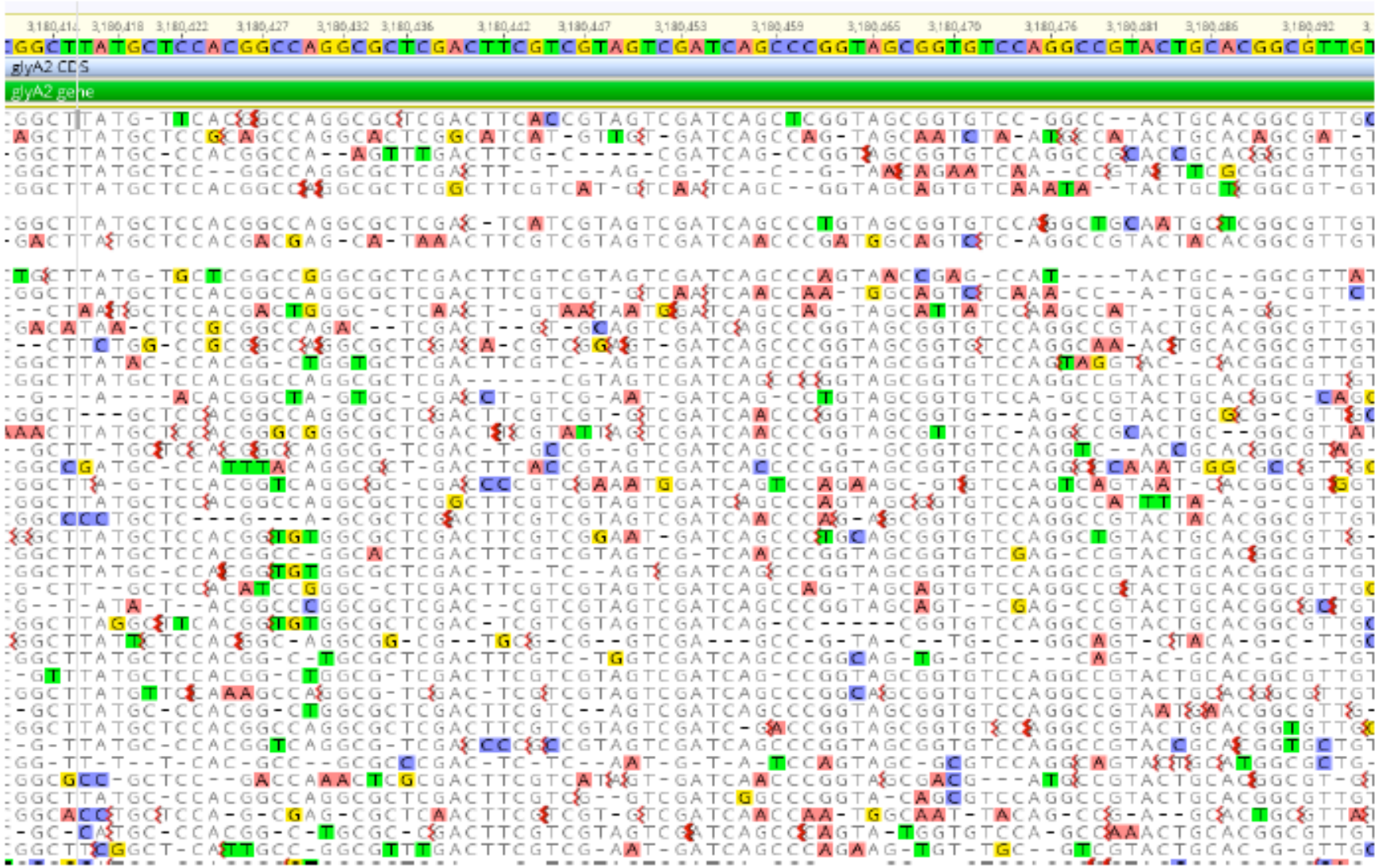
PromethION

-High-throughput
24/48 x Flow Cells

Long reads have high error rates



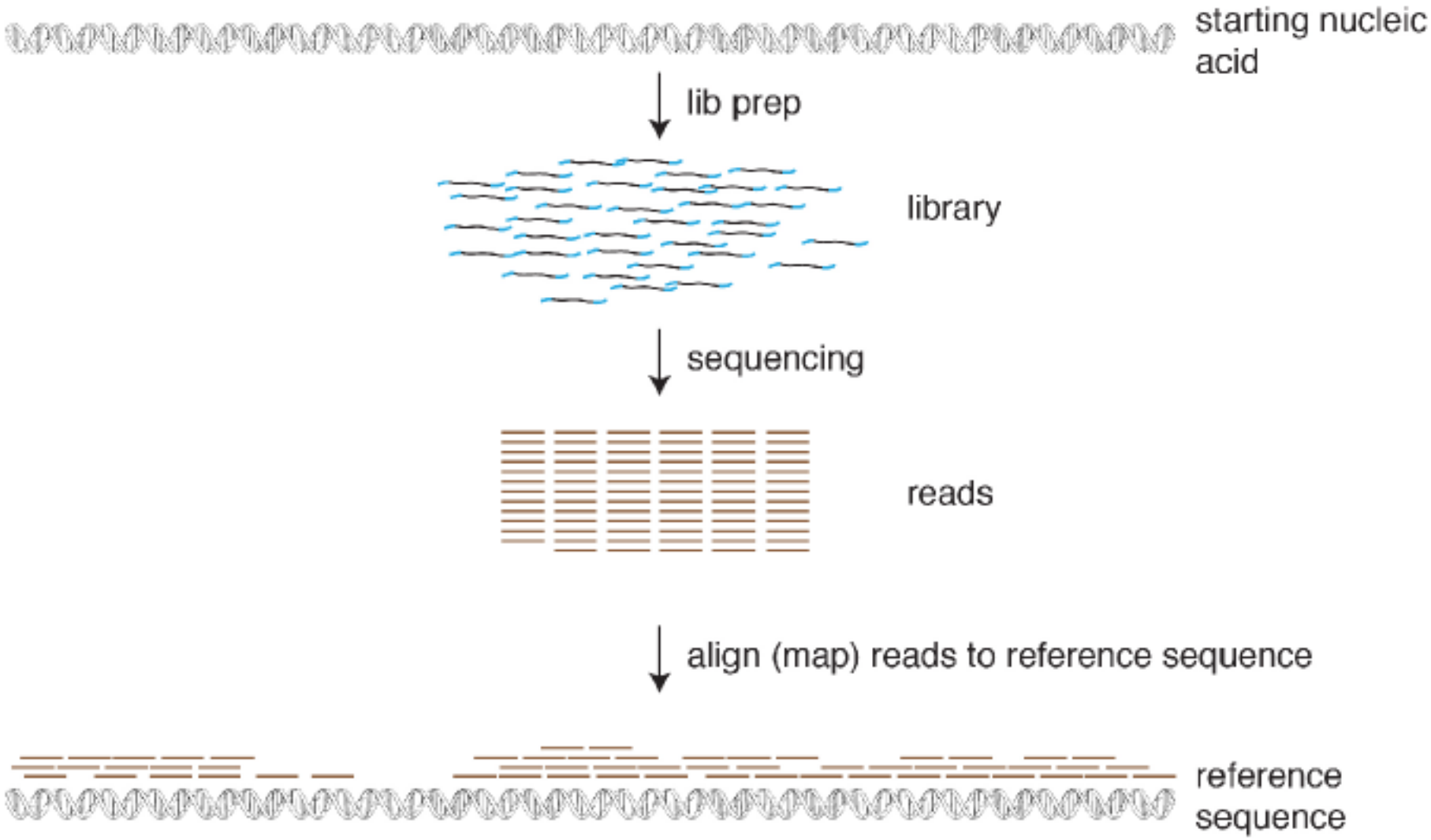
But you can use the consensus sequence



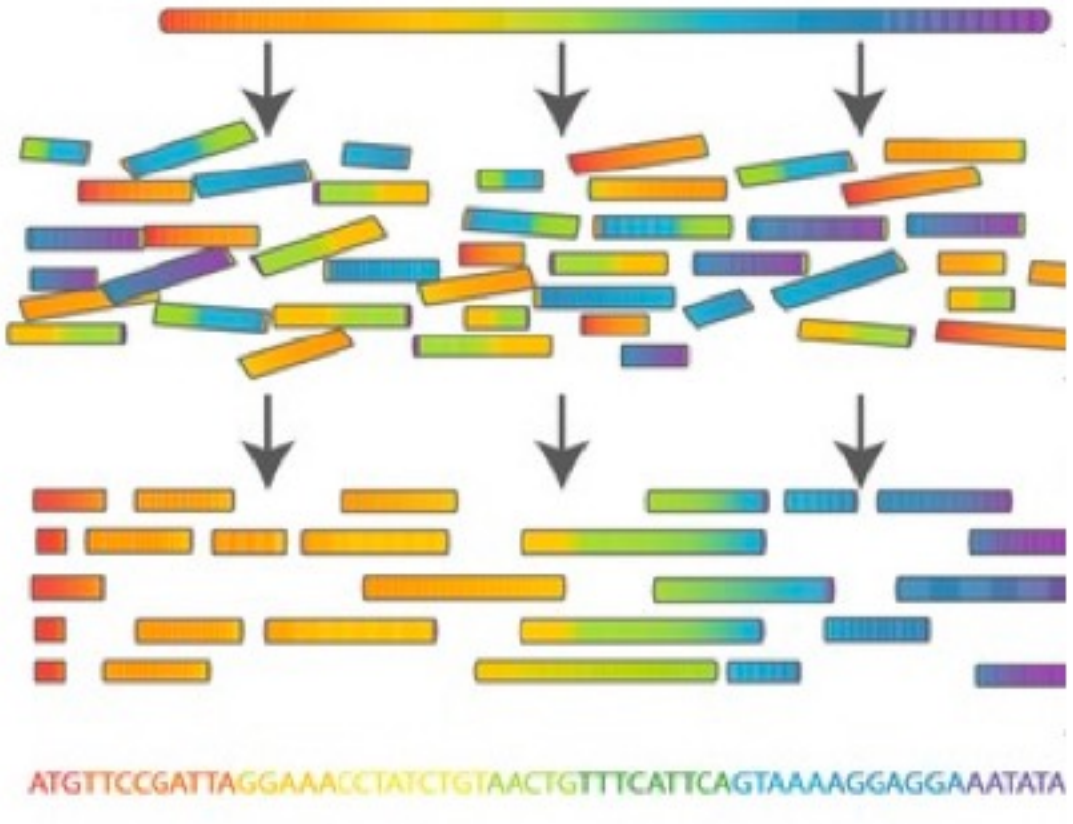
What is the consensus of these molecules? ("majority rules" at each site)

```
C G T C A C T C A G C A
C G T C A C T C A A A A
C A T C A C T C A G C A
C A C C A C T C A G C A
C A T C A C T T T G C A
C G T T A C T C A G C A
C A T C G C A C A G C A
C G T C A T A C A G C A
C G T C A T T C A G C A
C G T C A C T C A G C A
```

Long reads can map reads uniquely in a reference



Using long-read overlaps to perform de novo assembly



not yet sequenced genome

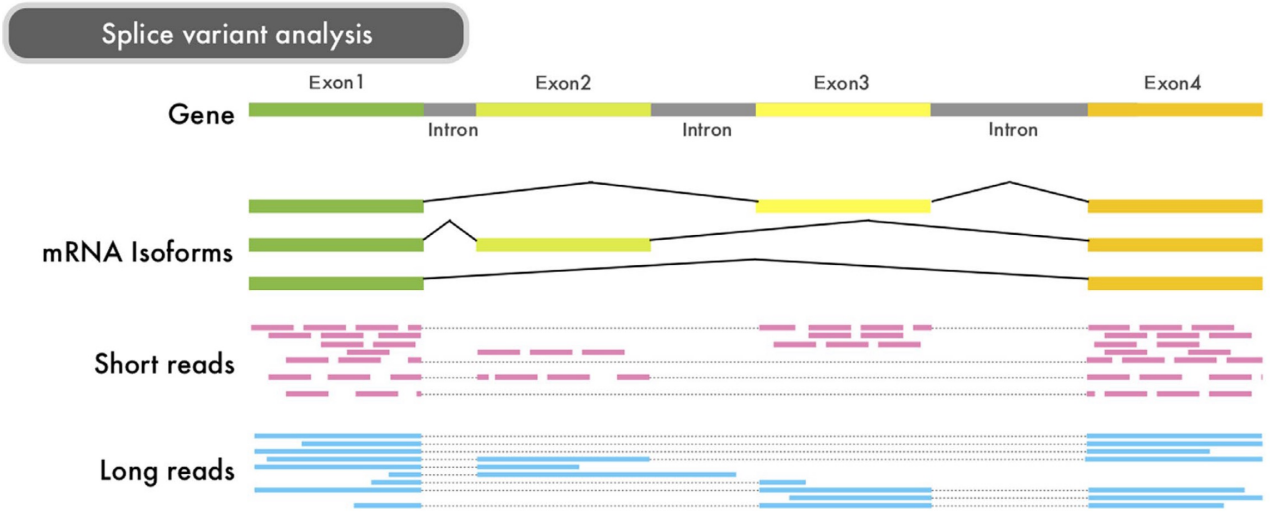
reads

contigs

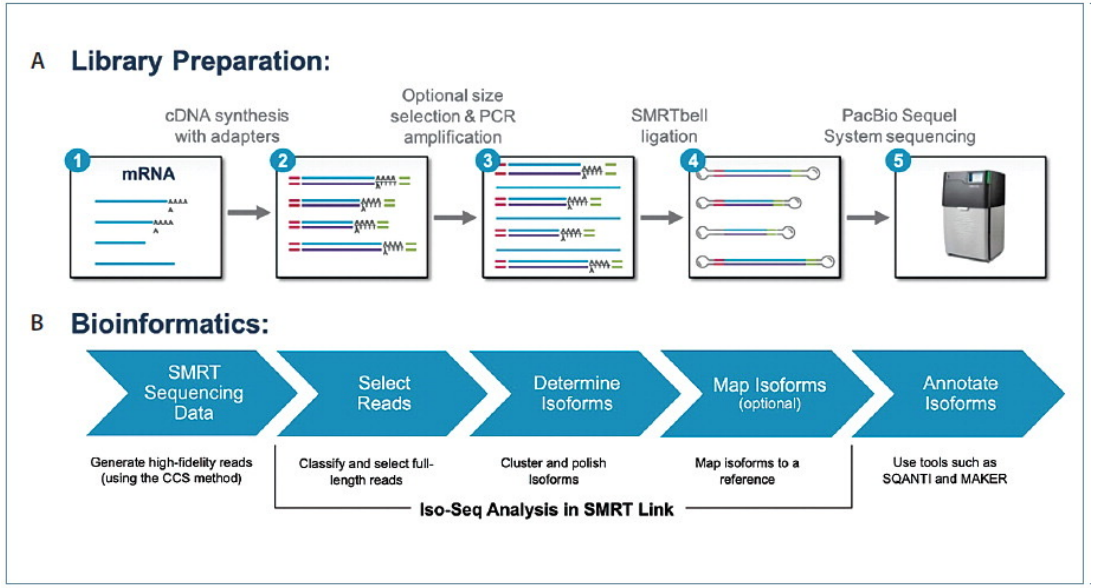
reconstructed (assembled) genome

Image: Commins, J., Toft, C., Fares, M. A (2009)

Isoform profiling with long reads removes the assembly step



Nanopore RNA sequencing



PacBio IsoSeq

Use long reads

- When linkage is more important than nucleotide identity

- Identify structural variants
- Resolve complex DNA structures
- Sequence though repeats
- Identify distinct splice variants
- Assembling a reference genome

Don’t use long reads

- When nucleotide identity is more important than linkage

- Identify low-frequency SNVs
- Quantify gene expression
- Re-sequencing in populations (for now)

Which technology would you use?

- Quantifying gene expression among different isoforms in a non-model species
- Linkage analysis between SNPs that are on average 10kb apart
- Assemble a plant mitochondrial genome

FASTQ Format

```
[joel.sharbrough_nmt@discovery-l2 data]$ head -16 *.ERR2000666.fastq
@ERR2000666.1 1 length=60
CCCTTCGCTGGGCAGCGCAGTACTGGCTGCCGATGGTGCAGGTCACGCCGCGCGGGATGT
+ERR2000666.1 1 length=60
I]]]]]]]]]]GIBHI]ICI]]]]]HIDHIIICDIIDIIDI]]ICIHIHFIFGEFD?@C>
@ERR2000666.2 2 length=152
NCGGCGTGCGTNNNNNNNNNGCGANNNNNNNNNNNNNNNNTNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
+ERR2000666.2 2 length=152
!DHGFH?FFG@!!!!!!!!!IIIC!!!!!!!!!!!!!!!!8!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
@ERR2000666.3 3 length=152
CNCCGGTGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
+ERR2000666.3 3 length=152
3!D=?415!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
@ERR2000666.4 4 length=43
TCGTGGTGCTGGATCATCTTGGTGAGGAAGGCCTGGTCGAACG
+ERR2000666.4 4 length=43
]]]]]]S]]]]]]BI]]]]]]]]]]]>]B]]]A?]FA?@:>]W]
```

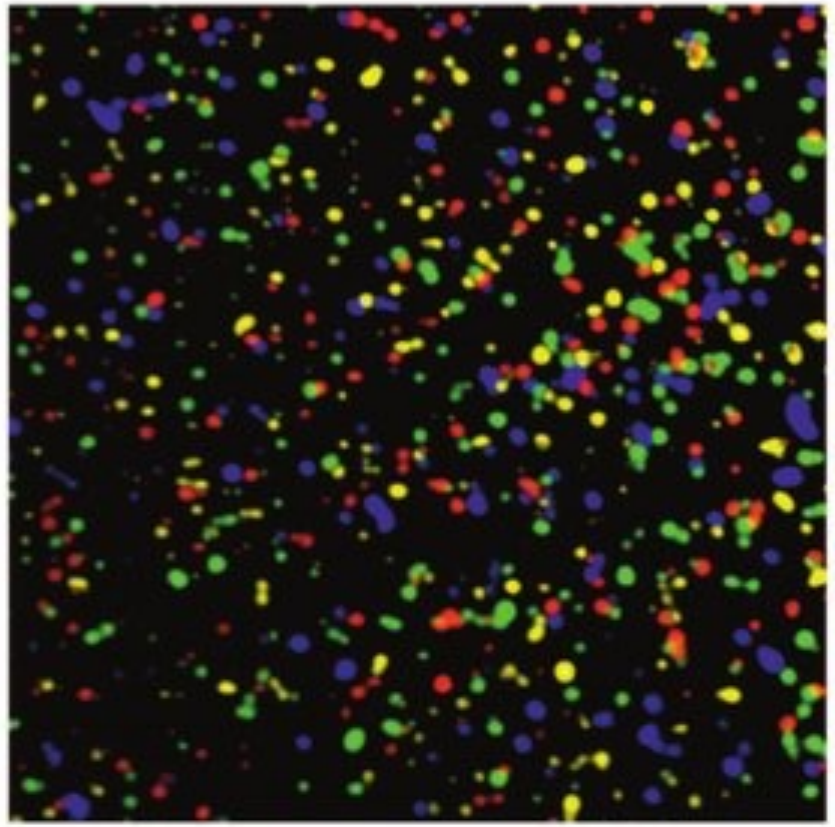
SAM/BAM format

SAM = Sequence Alignment Map

BAM = Binary Alignment Map

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,254}`	Query template NAME
2	FLAG	Int	$[0, 2^{16} - 1]$	bitwise FLAG
3	RNAME	String	`\*|[:rname:^*=][:rname:]*`	Reference sequence NAME[11]
4	POS	Int	$[0, 2^{31} - 1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0, 2^{8} - 1]$	MAPping Quality
6	CIGAR	String	`\*|([0-9]+[MIDNSHPX=])+`	CIGAR string
7	RNEXT	String	`\*|=|[:rname:^*=][:rname:]*`	Reference name of the mate/next read
8	PNEXT	Int	$[0, 2^{31} - 1]$	Position of the mate/next read
9	TLEN	Int	$[-2^{31} + 1, 2^{31} - 1]$	observed Template LENgth
10	SEQ	String	`\*|[A-Za-z=.]+`	segment SEQuence
11	QUAL	String	`[!-~]+`	ASCII of Phred-scaled base QUALity+33

Read files (Fastq or BAM) come with quality scores



How out-of-phase is each cluster on the image?

- Use Phi-X control library in every Illumina run as the calibration point for quality
- Quality encoded in a Phred score

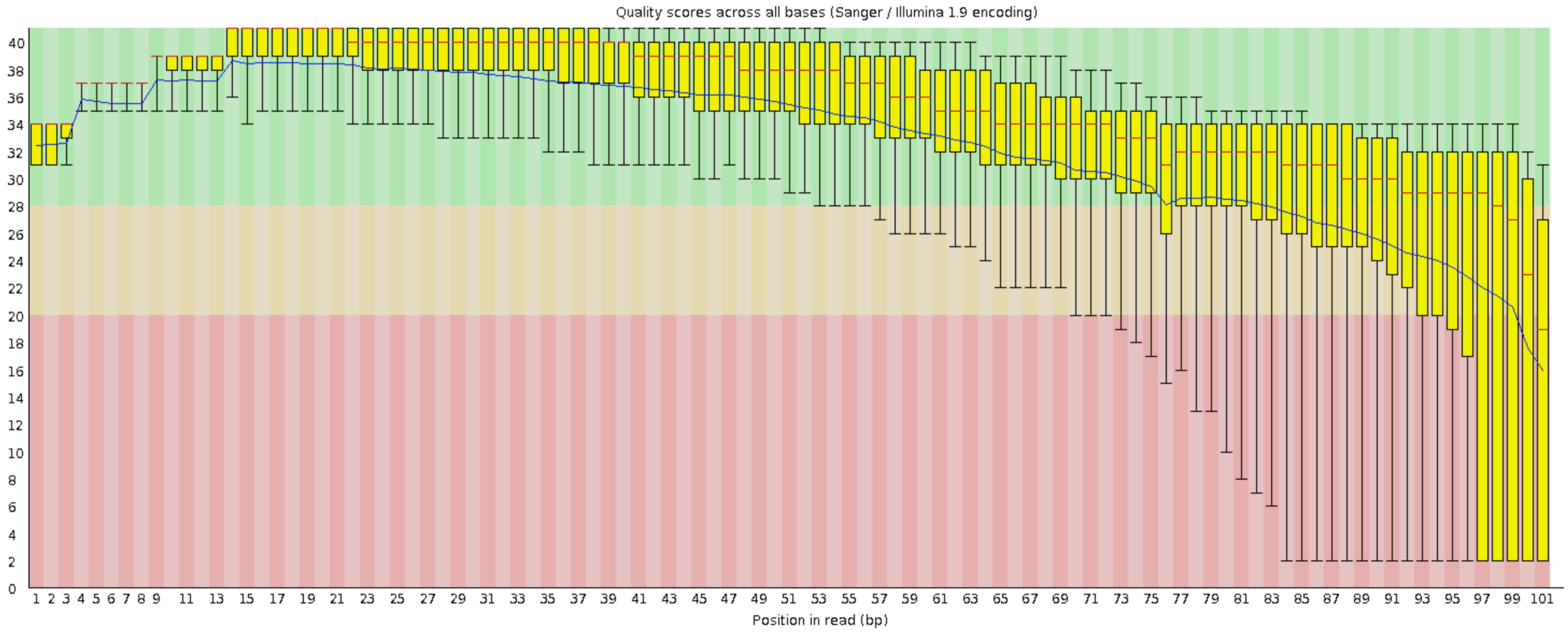
$$Q = -10 \log E$$

Read files (Fastq or BAM) come with quality scores

Phred Quality Score	Error	Accuracy (1 - Error)
10	1/10 = 10%	90%
20	1/100 = 1%	99%
30	1/1000 = 0.1%	99.9%
40	1/10000 = 0.01%	99.99%
50	1/100000 = 0.001%	99.999%
60	1/1000000 = 0.0001%	99.9999%

Visualizing your data with fastqc

Quality declines with increasing cycle number because amplicons within clusters get out of phase.



Trim Adapters/Barcodes!!

Illumina read

Pacbio read

Nanopore read

Fastx Toolkit

Trimmomatic

Cut-Adapt

etc.

Trim Adapters/Barcodes!!

Illumina read

Pacbio read

Nanopore read

Fastx Toolkit

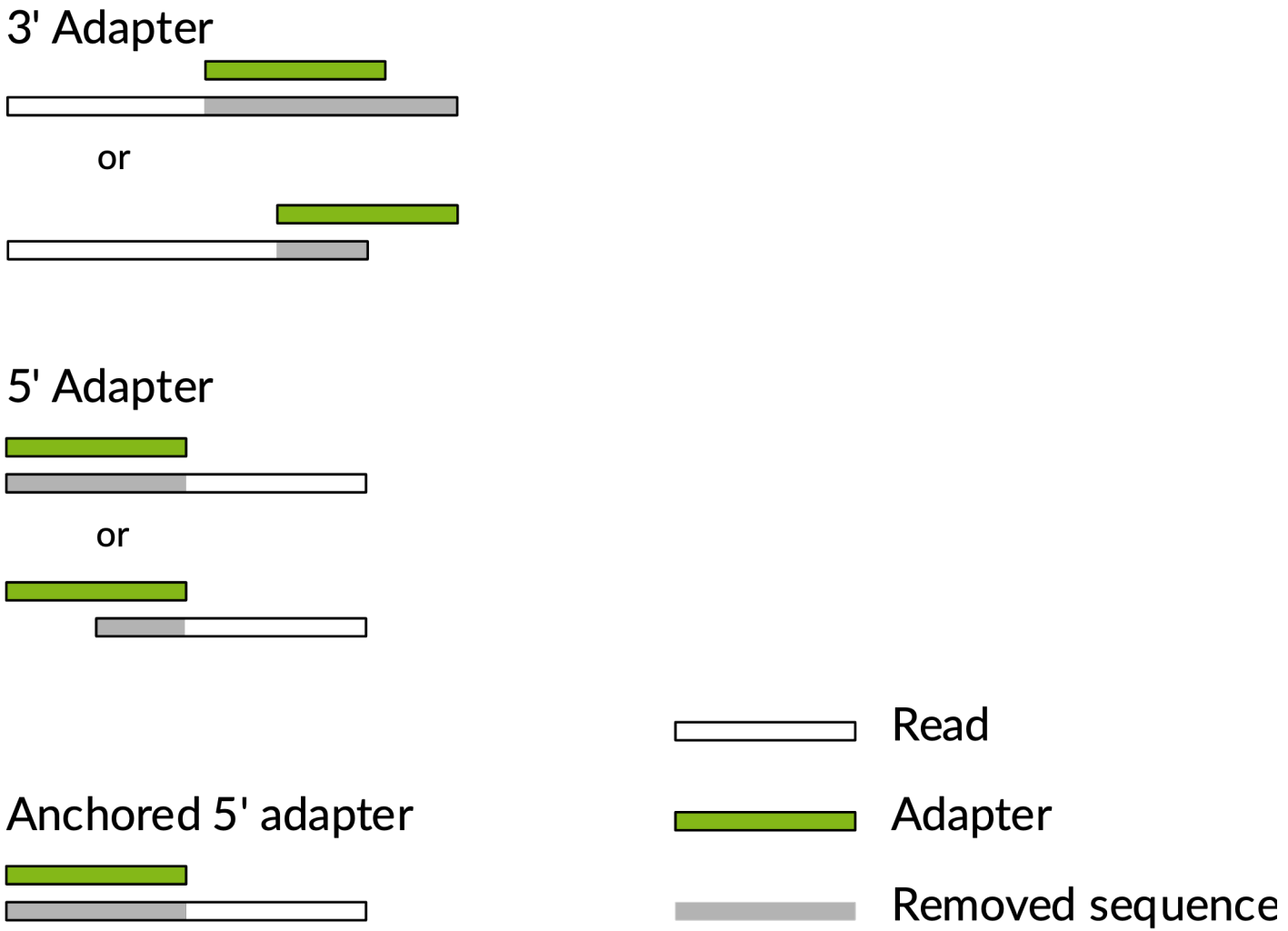
Trimmomatic

Cut-Adapt

etc.

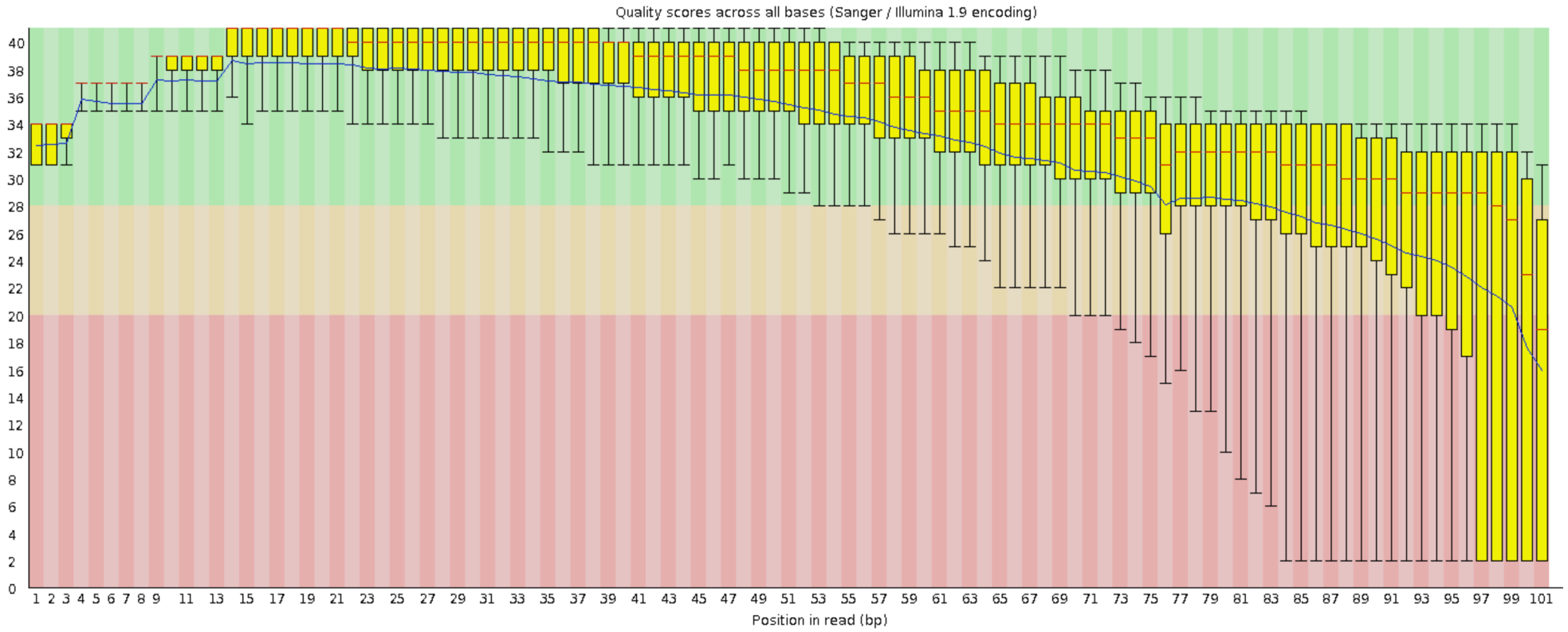
Atlantic cod genome assembly is full of adapters

Remove adapters, primers, poly A tails, with Cutadapt



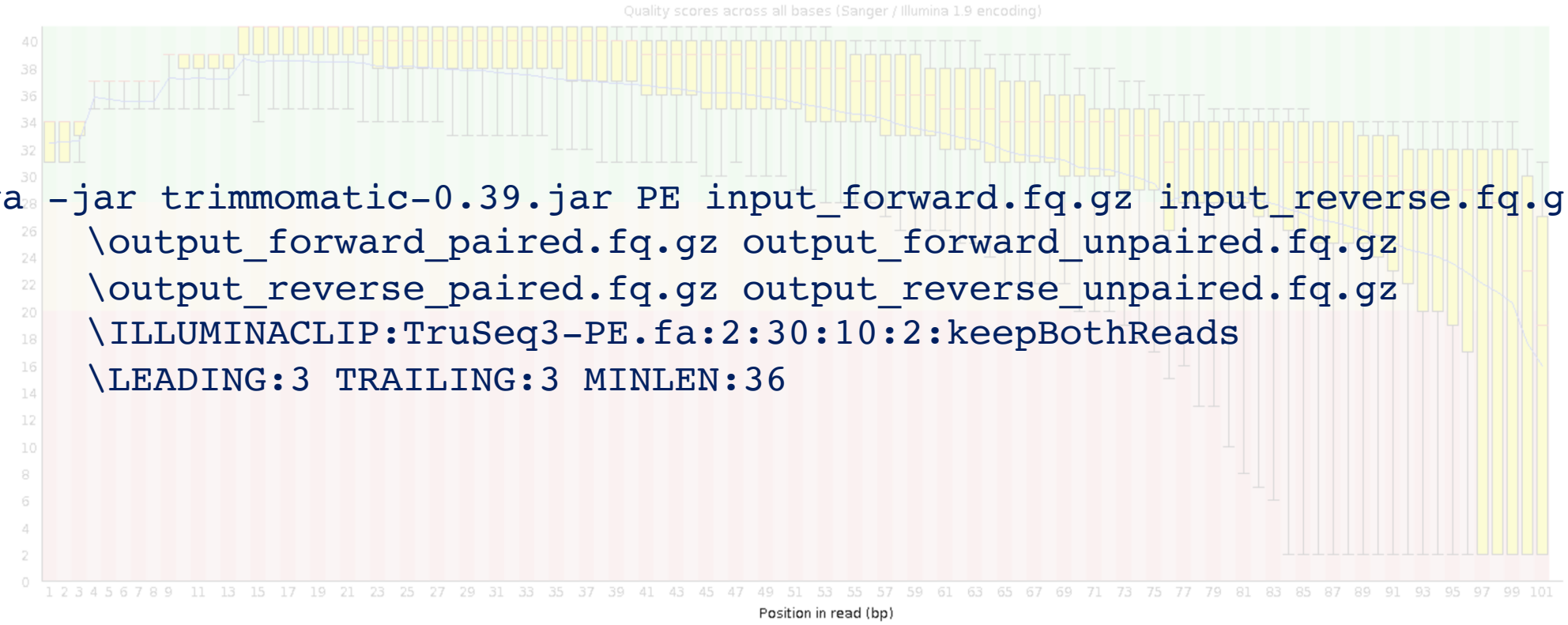
Trim low-quality bases with Trimmomatic

Quality declines with increasing cycle number because amplicons within clusters get out of phase.



Trim low-quality bases with Trimmomatic

Quality declines with increasing cycle number because amplicons within clusters get out of phase.




```
java -jar trimmomatic-0.39.jar PE input_forward.fq.gz input_reverse.fq.gz
  \output_forward_paired.fq.gz output_forward_unpaired.fq.gz
  \output_reverse_paired.fq.gz output_reverse_unpaired.fq.gz
  \ILLUMINACLIP:TruSeq3-PE.fa:2:30:10:2:keepBothReads
  \LEADING:3 TRAILING:3 MINLEN:36
```