Lesson 11

Third Generation Sequencing

Final Assignment

By the end of next week (6.2) let me know your group by e-mail!

Group = 1/2/3 students

Final project submission via Moodle - 02.03.25
Do you need more time?

By the end of this lesson you will...

- Understand the stages involved in DE analysis
- Be able to generate and explore RNA-seq read count tables
- Be familiar with the statistical models behind differential gene expression analysis
- Know how to perform differential gene expression analysis using the pyDESeq2 Python package

By the end of this lesson you will...

Be familiar with the main 3rd generation sequencing technologies:

- PacBio SMRT sequencing
- ONT sequencing
- 10X linked reads

Understand various applications of long and linked reads

- RNA-seq
- De novo assembly
- Structural variant calling

What is 3rd Gen Sequencing

Sequencing technologies other than Illumina sequencing

Focus on producing **long-distance** information

- **Long reads**
- **Linked reads**

Developed or matured in the last decade

Actively being developed

Main technologies:

- Pacific Biosciences SMRT sequencing - **PacBio**
- Oxford Nanopore Technology - **ONT**
- 10X Genomics Chromium - **10X**

PacBio SMRT Sequencing

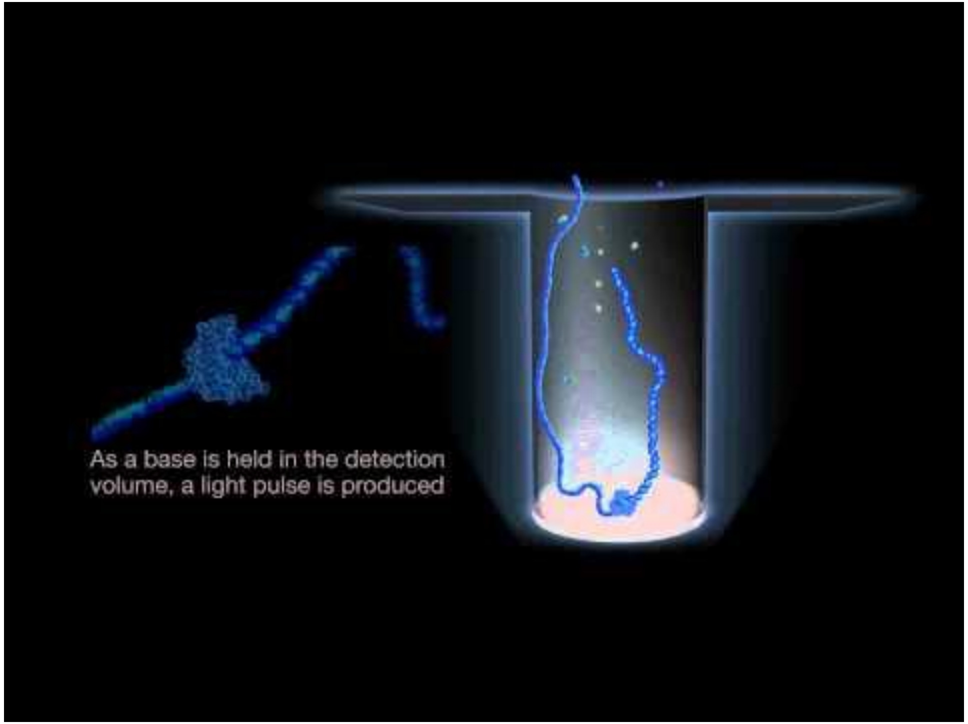
Single Molecule Real Time

No amplification step

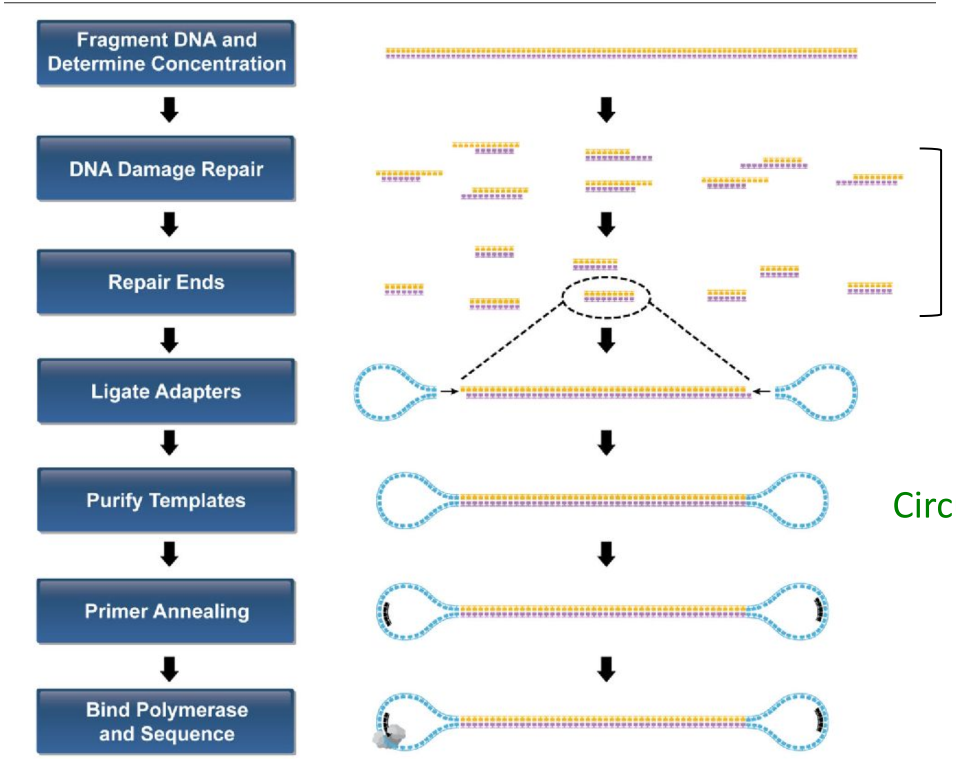
Based on the ability to analyze very small volumes

Sequencing by synthesis

Sequel II


As a base is held in the detection volume, a light pulse is produced

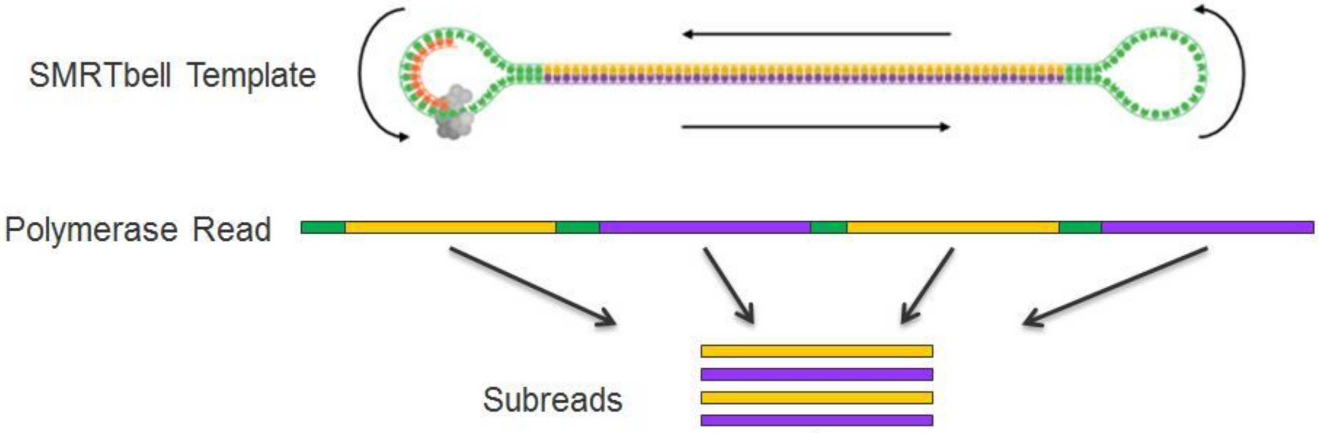
PacBio Library Prep



15-20kb insert

Circular

PacBio Sequencing



Properties of PacBio Sequencing

Read length

- Non-uniform
- Depends on selected insert size
- Usually 10-100kb

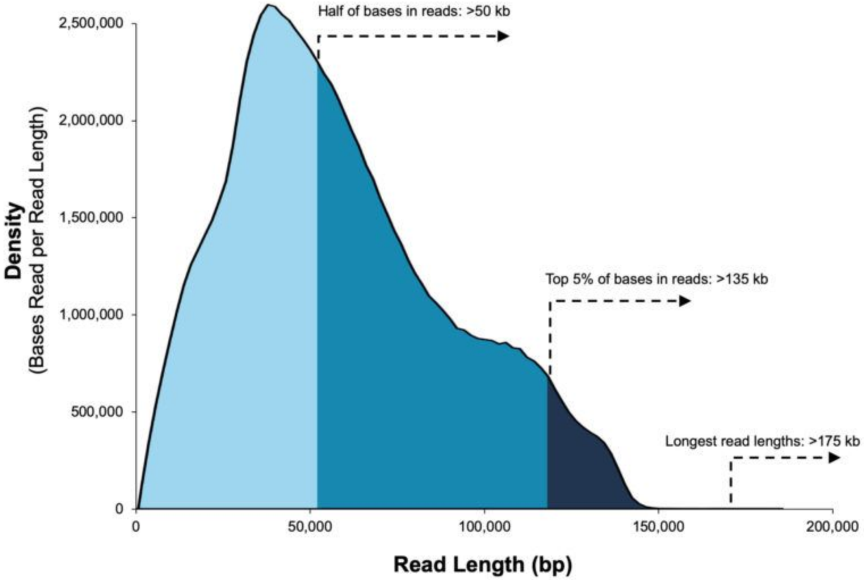
No paired-end option

One run can produce 4-5M reads - ~40Gb

Runs take several hours

Mostly uniform coverage - no GC-content bias

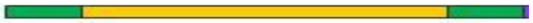
Raw reads error rate - **~10%**



Dealing With High Error Rates

Working with 10% error rate is impractical

Option 1:



Polymerase Read

CLR - continuous long read

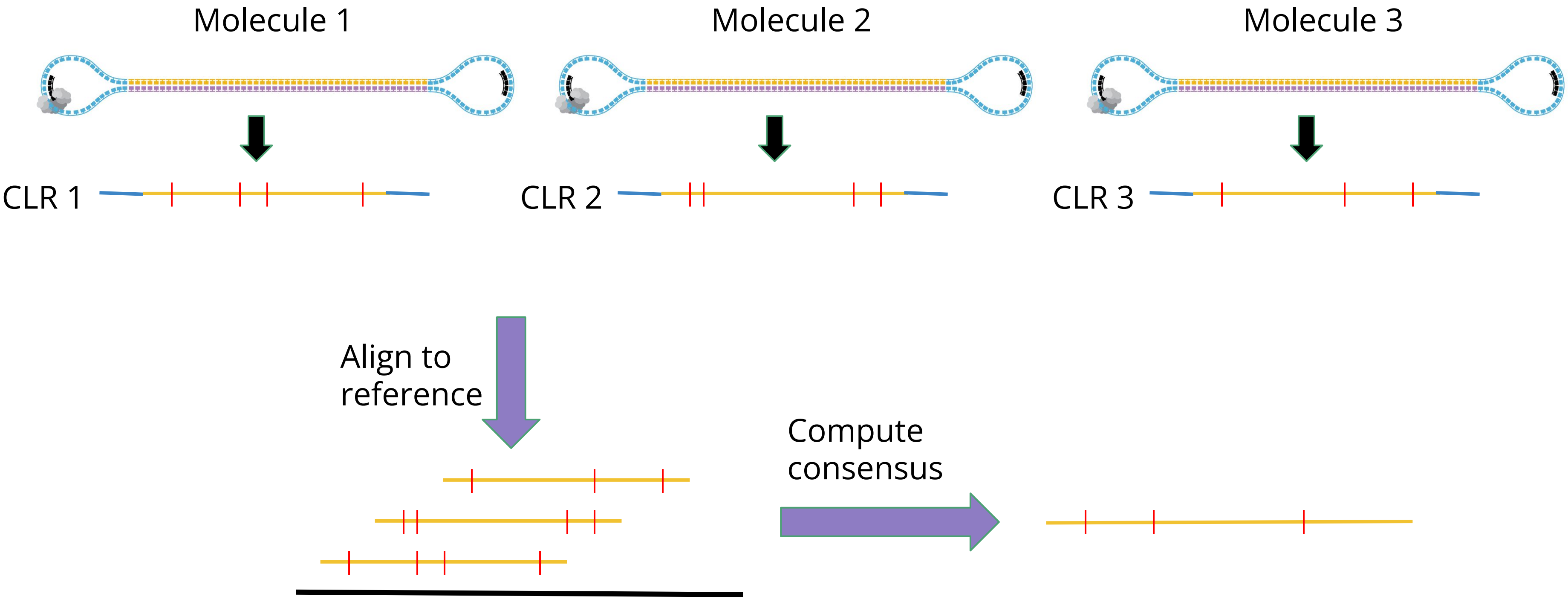
Polymerase read length ~= sub-read length

Align CLRs to a reference genome and correct errors

Find the consensus of multiple molecules

Accuracy increases with sequencing depth

CLR Error Correction



Molecule 1

Molecule 2

Molecule 3

CLR 1

CLR 2

CLR 3

Align to reference

Compute consensus

Dealing With High Error Rates

Option 2:

CCS - circular consensus read

Also called **HiFi reads**

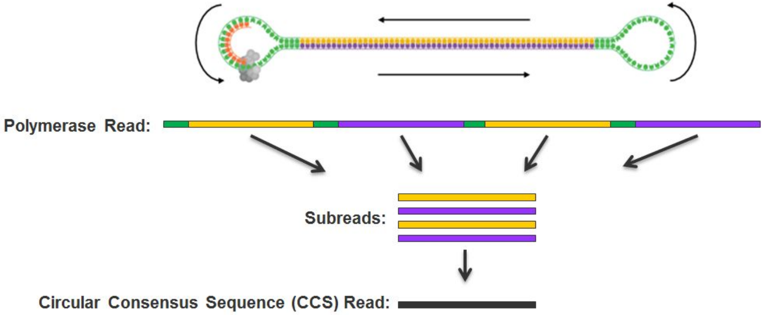
Polymerase read length > sub-read length

Align CCSs to one another and correct errors

Find the consensus of a single molecule

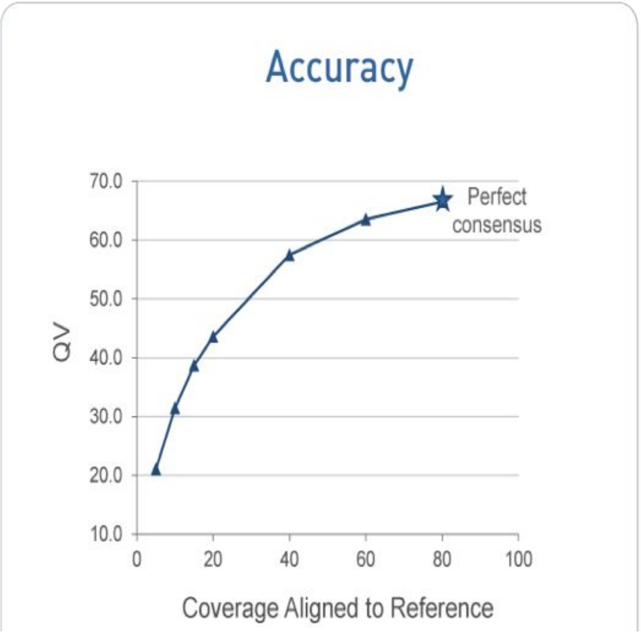
Accuracy >99%

Shorter reads (<20kb)

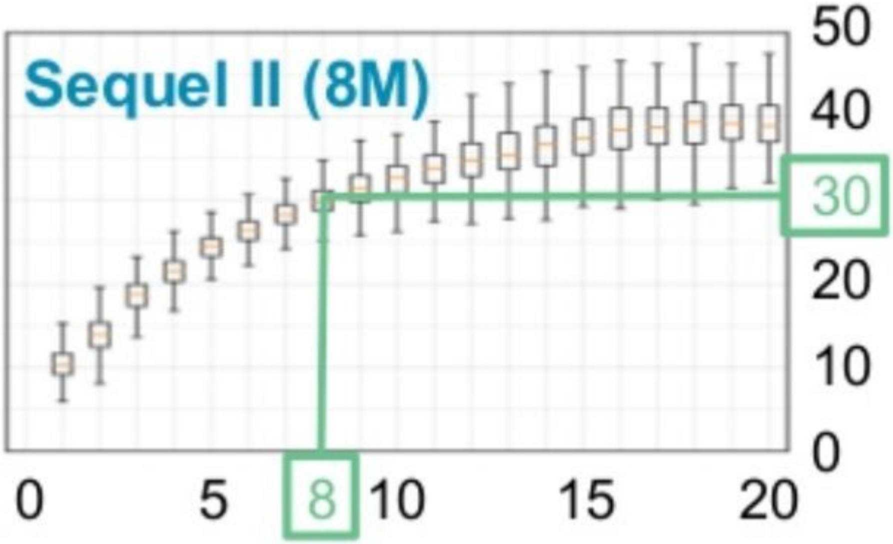


Accuracy CLR consensus Vs. CCS

CLR consensus



CCS



Passes

Oxford Nanopore Sequencing (ONT)

Single molecule

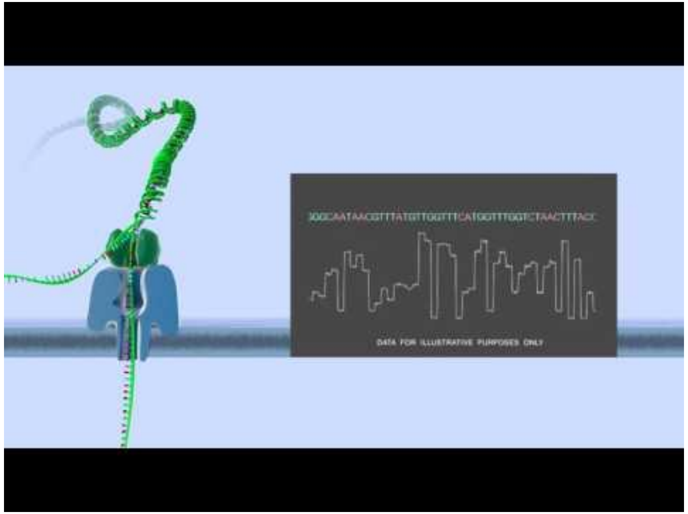
Real time

Not SBS

Palm-size

MinION MkI: portable, real time biological analyses

MinION



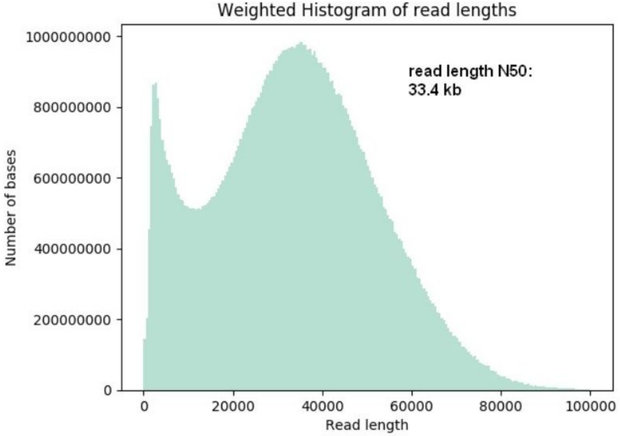
Properties of ONT Sequencing

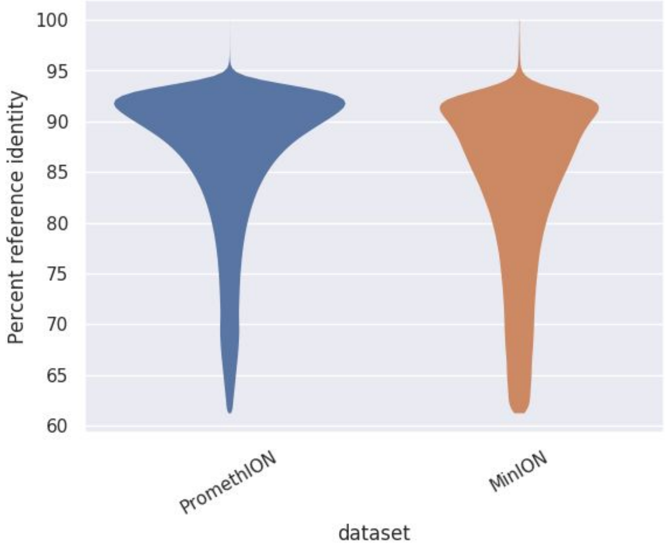
Read length - theoretically unlimited

In practice depends on DNA fragmentation - can produce reads > 2Mb

Yield - depends on machine model - 50Gb to 10Tb

Accuracy - ~10% error





Comparing Technologies

	Illumina	**PacBio CLR**	**PacBio CCS**	**ONT**
Read length	150-250 bp	50 kb	30 kb	10-30 kb
Overall error rate	0.1 %	10-15 %	<1 %	<5 %
Mismatch	~ 100 %	37 %	4 %	41 %
InDel	~ 0 %	63 %	96 %	59 %
Cost	$29/Gb	$85/Gb		$30/Gb*
Throughput	7 Gb/h	2.5 Gb/h		0.5 Gb/h*

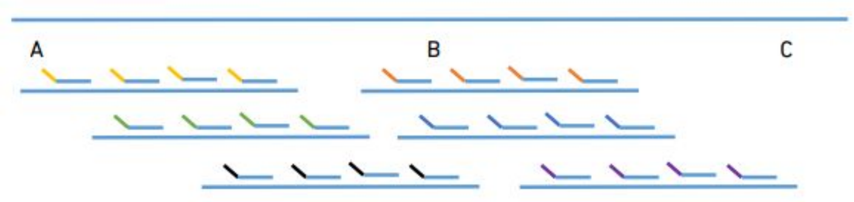
Amarasinghe, Shanika L., et al. "Opportunities and challenges in long-read sequencing data analysis." *Genome biology* 21.1 (2020): 1-16.

10X Genomics

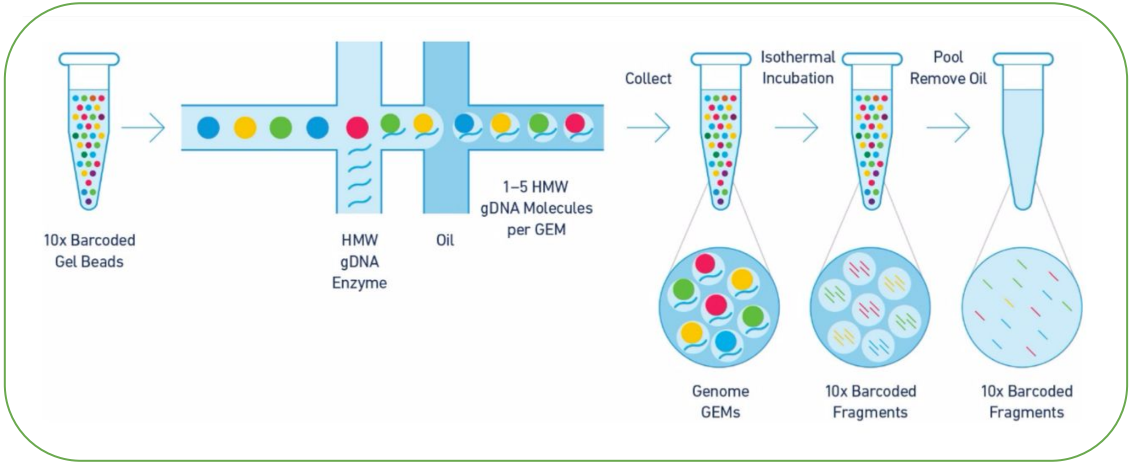
Not a long read technology

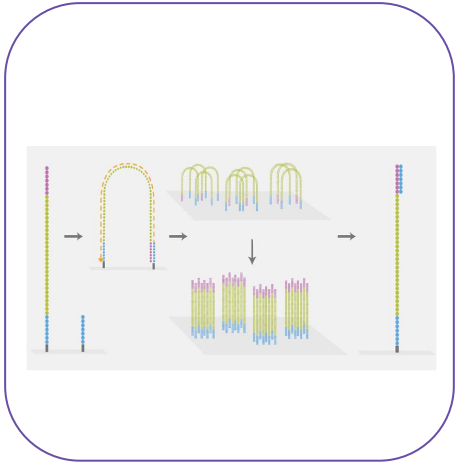
But provides long-range information through **linked reads**

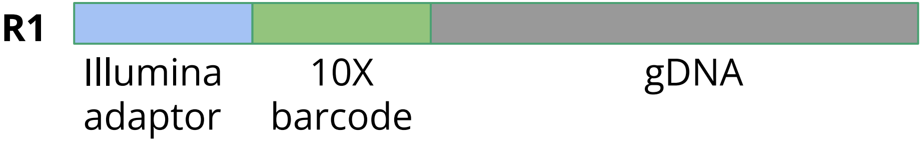
Short reads originating from the same long molecule



Based on standard short read Illumina technology







R1

Illumina adaptor

10X barcode

gDNA

Linked Reads

Reads with the same barcode likely come from the same gDNA fragment

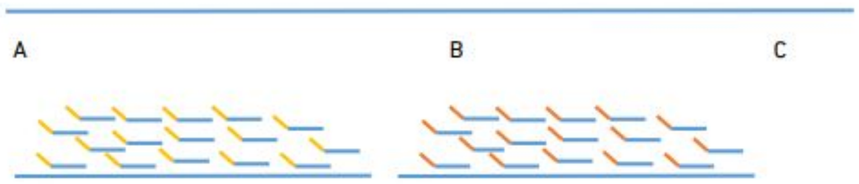
gDNA fragment size is usually 50-60kb

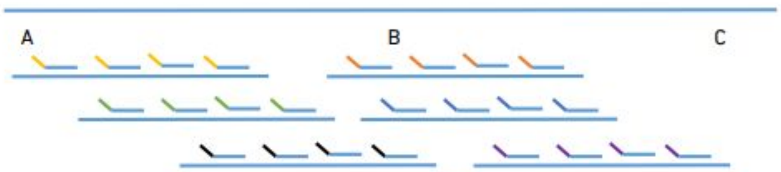
If ~x3 depth is used - we can produce “synthetic long reads”

Usually each molecule is sequenced at ~x0.2

We can still get useful long-range information

Non-trivial computational analysis is needed





Applications of 3rd Gen Sequencing

Transcriptomics

Genome assembly

Structural variation detection

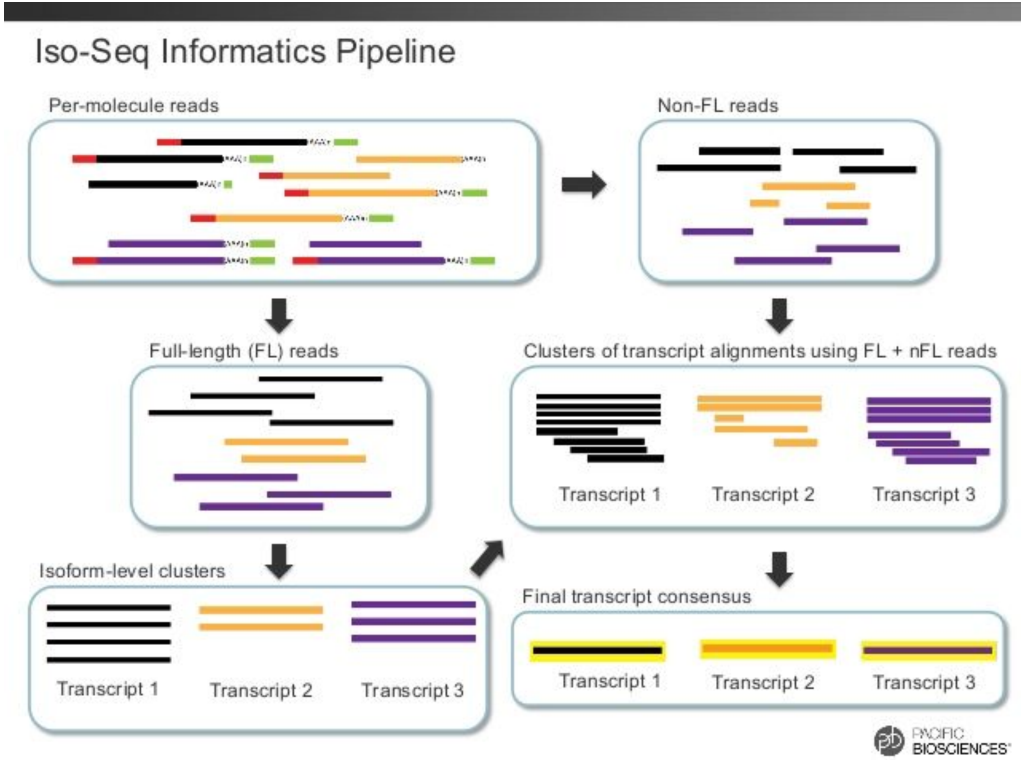
RNA-Seq and Long Reads

Read length is usually larger than mRNA size

Full-length transcripts

No transcript assembly is needed

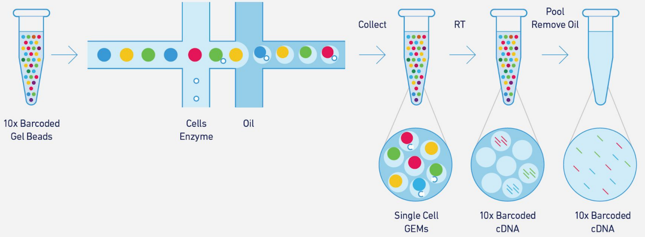
Easier to detect and quantify isoforms



10X for Single Cell RNA-Seq

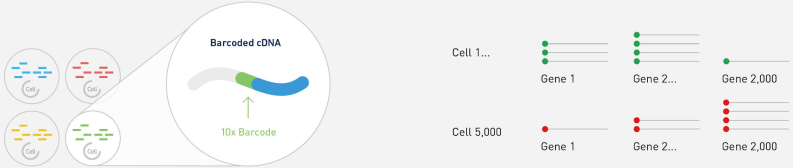
GemCode™ Technology for Single Cell Partitioning

Utilize an efficient droplet-based system to encapsulate up to 100-80,000+ cells in a single 10-minute run.



Single Cell Digital Gene Expression

Enable digital quantification of transcripts in every cell, for single cell digital gene expression analysis.



mRNA is transcribed by a reverse transcriptase that creates barcoded cDNA.

A barcode identifies transcripts originating from a single cell, which are then counted.

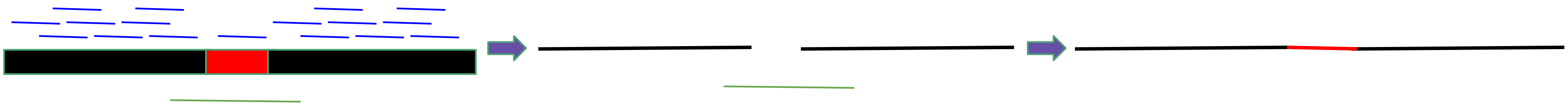
Long and Linked Reads in Genome Assembly

Many modern assemblers can work with 3rd generation reads

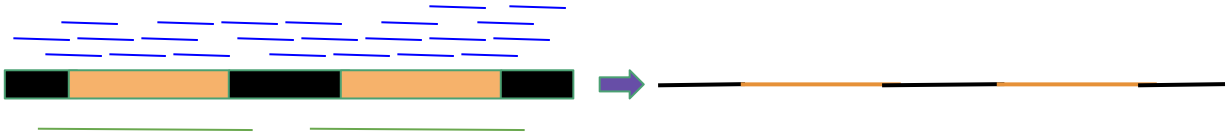
- Falcon - PacBio reads
- Canu, SPAdes - PacBio and ONT reads
- Supernova - 10X reads

Most assemblers take a “hybrid” approach - long + short reads

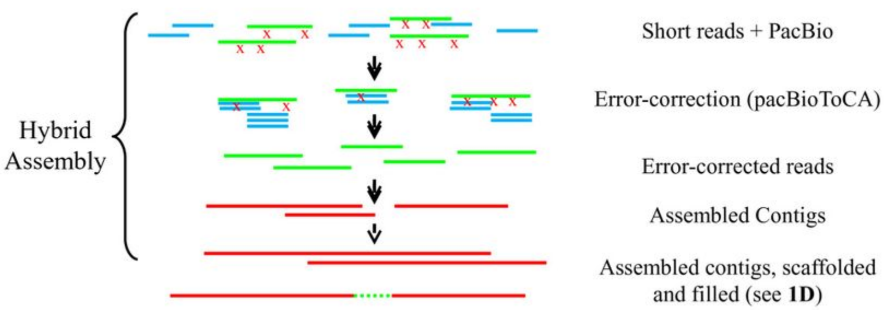
Long/linked reads can help link contigs by bridging over difficult regions



Long reads can help solve long repeats



Different Assembly Strategies



Hybrid Assembly

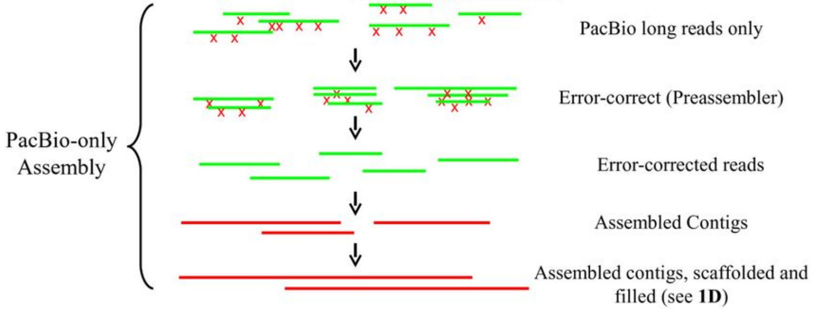
Short reads + PacBio

Error-correction (pacBioToCA)

Error-corrected reads

Assembled Contigs

Assembled contigs, scaffolded and filled (see **1D**)



PacBio-only Assembly

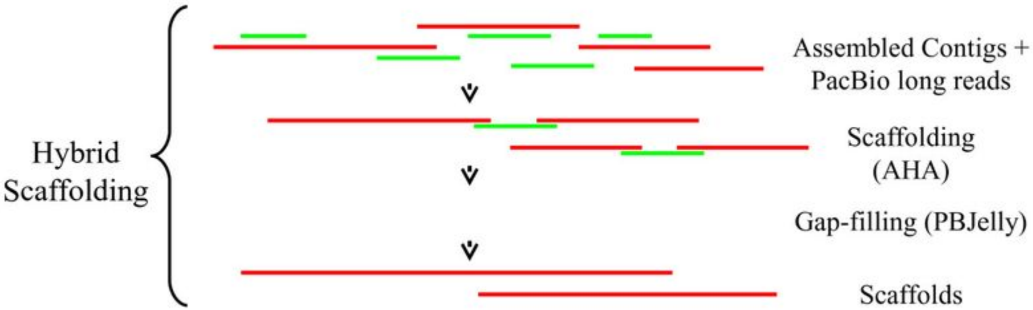
PacBio long reads only

Error-correct (Preassembler)

Error-corrected reads

Assembled Contigs

Assembled contigs, scaffolded and filled (see **1D**)



Hybrid Scaffolding

Assembled Contigs + PacBio long reads

Scaffolding (AHA)

Gap-filling (PBJelly)

Scaffolds

Haplotype Phasing

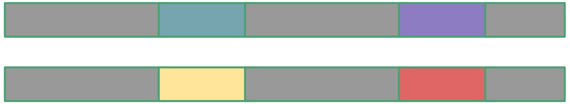
Many interesting eukaryote genomes are diploid or polyploid

Still, most assemblies are haploid

Heterozygosity is “squished” into consensus sequences

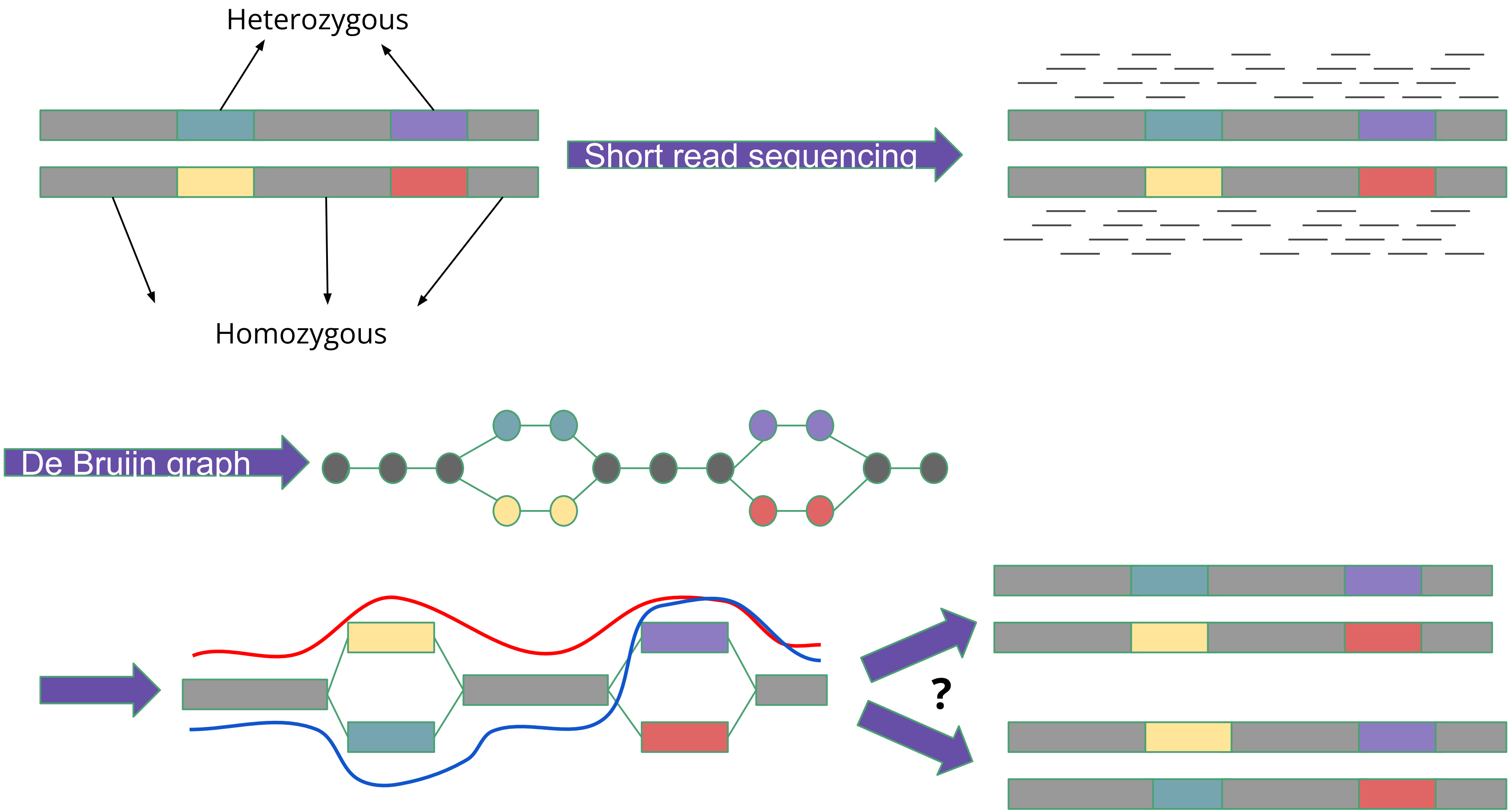
A **haplotype** is a group of alleles arising from the same molecule

Splitting an assembly into haplotypes is called **phasing**



Haplotype 1 (maternal)

Haplotype 2 (paternal)



Heterozygous

Homozygous

Short read sequencing

De Bruijn graph

?

Structural Variant Detection

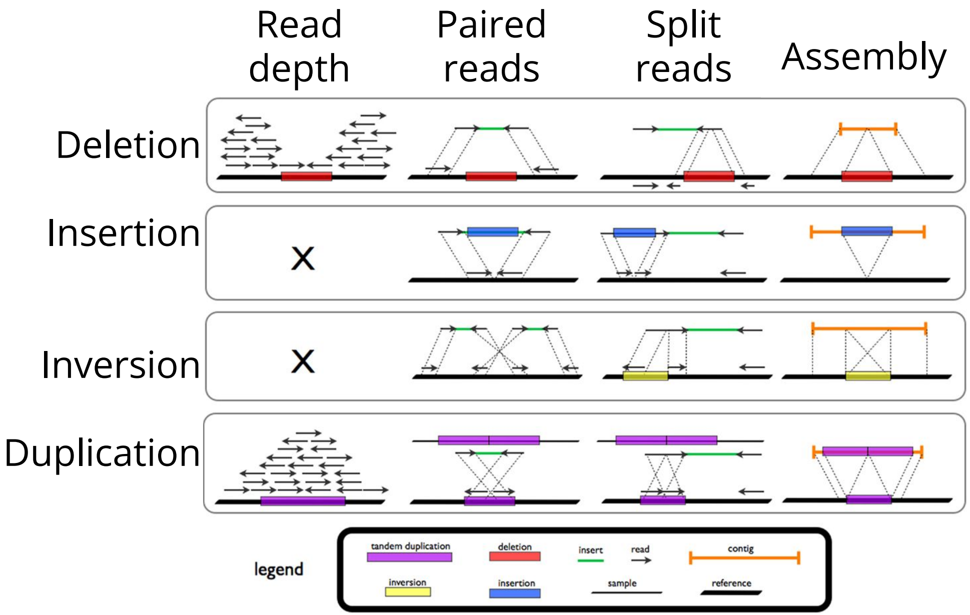
SVs are generally hard to detect with short reads

Many SVs are located in regions that are hard to sequence

SV detection is usually based on mapping reads to a reference

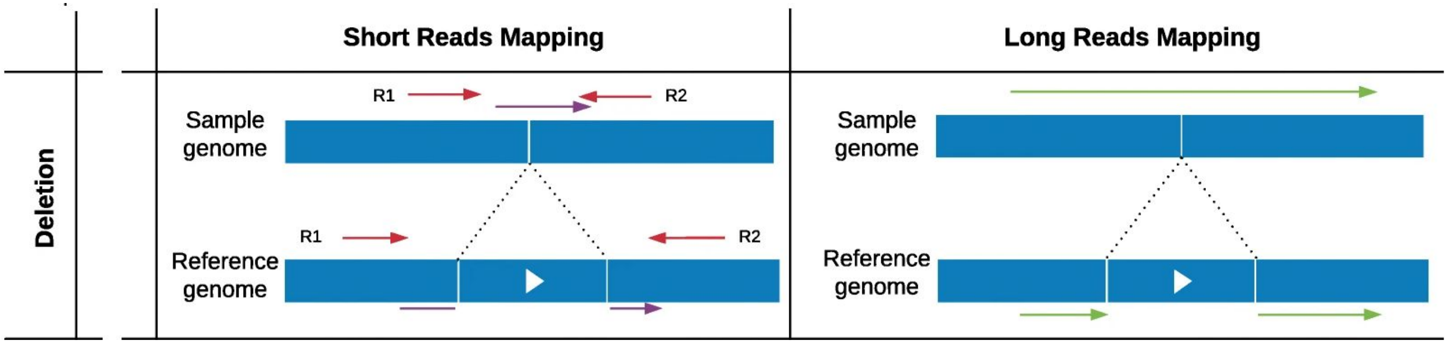
Long reads are useful because:

- They can cross long repeats
- They are not affected by GC-bias
- They can span large insertions



1. Tattini, L., D'Aurizio, R., & Magi, A. (2015). Detection of genomic structural variants from next-generation sequencing data. *Frontiers in bioengineering and biotechnology*, *3*, 92.

How Do we Detect Variants



	Sequencing	Mapping	Variant calling
SNP	short reads	BWA	GATK
SV	short reads	BWA	Manta
	long reads	Minimap2	Sniffles

Read Mapping With Minimap2

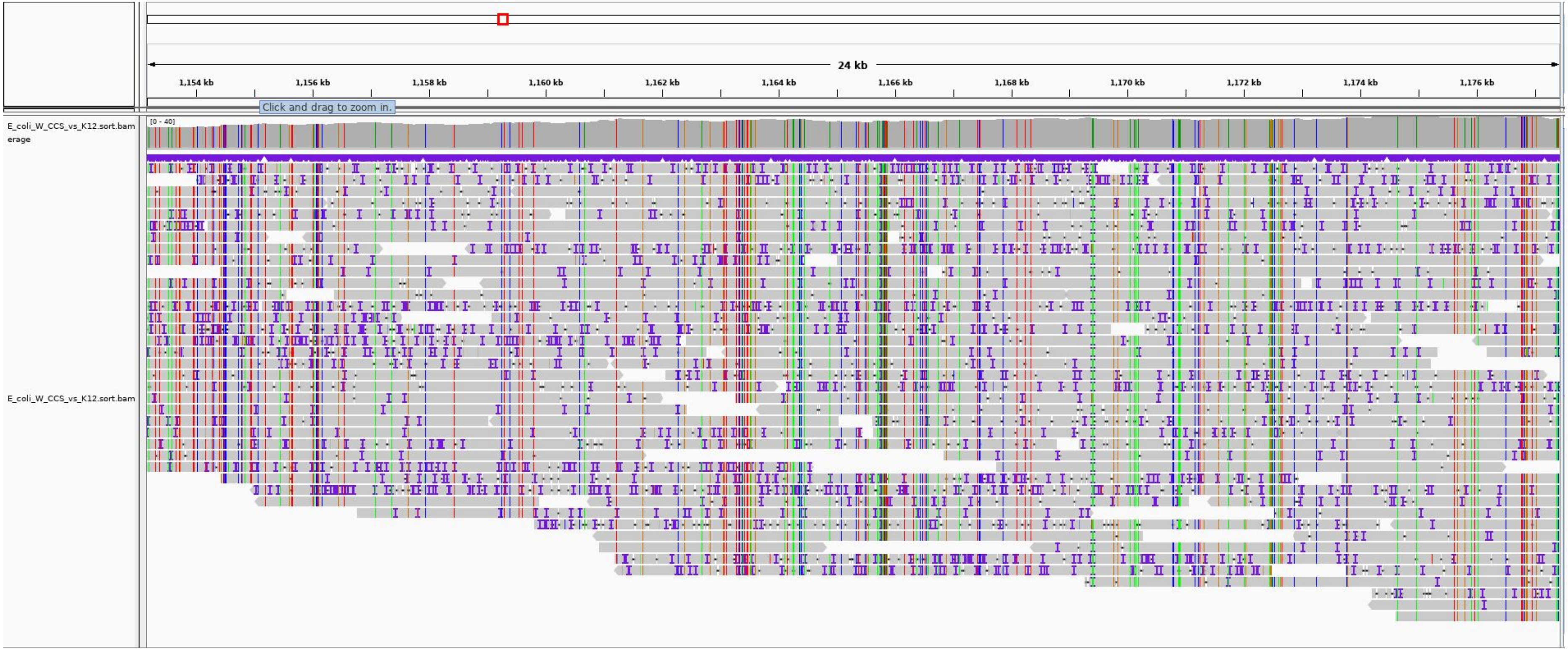
Minimap2 is a generic sequence mapping software

There are various mapping modes like:

- PacBio CLR to genome
- PacBio CCS to genome
- cDNA / PacBio Iso-Seq (transcripts) to genome
- ONT reads to genome
- PacBio reads to PacBio reads
- Short reads to genome (alternative to BWA)

Modes accounts for the specific biases of each technology

Input format is fasta/fastq


24 kb
1,154 kb
1,156 kb
1,158 kb
1,160 kb
1,162 kb
1,164 kb
1,166 kb
1,168 kb
1,170 kb
1,172 kb
1,174 kb
1,176 kb
Click and drag to zoom in.
E_coli_W_CCS_vs_K12.sort.bam
erage
[0 - 40]
E_coli_W_CCS_vs_K12.sort.bam

HYPE!

In recent years there have been **lots** of talk about long (an linked) reads

Many publications about data analysis and dedicated tools

Long reads are great! … **for some things**
Don’t trust everything you read

Always read the “small letters” (usually supplementary materials)

Vast majority of sequencing is still done with short reads

One technology can’t solve all problems in biology!