DNA Sequencing and Data Analysis

Prof Noam Shomron

Hadas Volkov

Lecture 12, January 20, 2023

DNA Sequencing and Data Analysis

Friday 8:45 AM to 11:15 AM

Arazi-Ofer Building, C.L03

nshomron@gmail.com

hadas.volkov@post.runi.ac.il

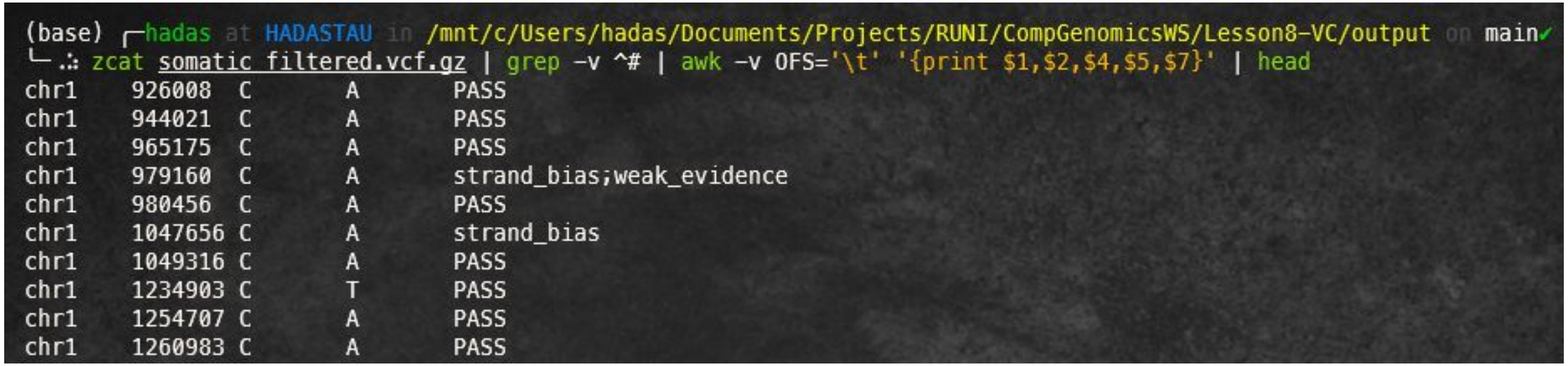
DNA Sequencing and Data Analysis

Introduction to Third Generation Sequencing

Class	Title	Content/assignments	Activity, location
1, 4.11	Introduction to Cells and DNA	Basic knowledge of biology	In the lecture hall, Noam
2, 11.11	DNA Sequencing past and present	Basic knowledge of molecular DNA	In the lecture hall, Noam
3, 18.11	Genomics technologies	DNA, RNA, technologies	In the lecture hall, Noam
4, 25.11	Introduction to Bioinformatics challenges in reading DNA	Focus on three methods: WES/WGS, RNA-seq, cell-free DNA	In the lecture hall, Noam
5, 2.12	Modern DNA Sequencing, 2nd wave File Formats, tools.	Analysis approaches for WES/WGS, RNA-seq, cell-free DNA	In the lecture hall, Hadas and Noam
6, 9.12	De novo Shotgun Assembly	The algorithms and methods behind the assembly problem	In computer class, Hadas and Noam
7, 16.12	Sequence Mapping and Alignment	The algorithms behind mapping and alignment, fast and heuristics	In computer class, Hadas and Noam
8, 23.12	Variant Calling and Somatic Variant Analysis	The bioinformatics behind discovery of novel mutations in cancer	In computer class, Hadas and Noam
9, 30.12	RNA-Seq	The bioinformatics behind RNA-Seq and Differential Gene Expression	In computer class, Hadas and Noam
10, 6.1	Practice molecular biology techniques	Pipetting, transferring small amounts of fluids, running a dry Nanopore experiment	In biology class, Meitar and Noam
11, 13.1	Nanopore DNA sequencing	Nanopore DNA sequencing, experimental run	In biology class, Meitar, Hadas
12, 20.1	**Nanopore data analysis introduction**	**Nanopore DNA analysis, experimental run**	**In computer class, Hadas and Noam**
13, 27.1	Nanopore data analysis and presentations	Groups present their results	In the lecture hall, Hadas and Noam

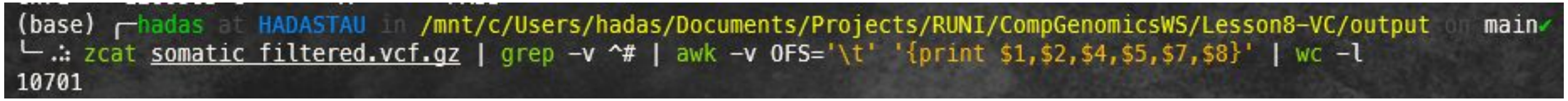
HW 8 - Variant Calling

```
gatk FilterMutectCalls -R genome.fa -V somatic.vcf.gz -O somatic_filtered.vcf.gz
```



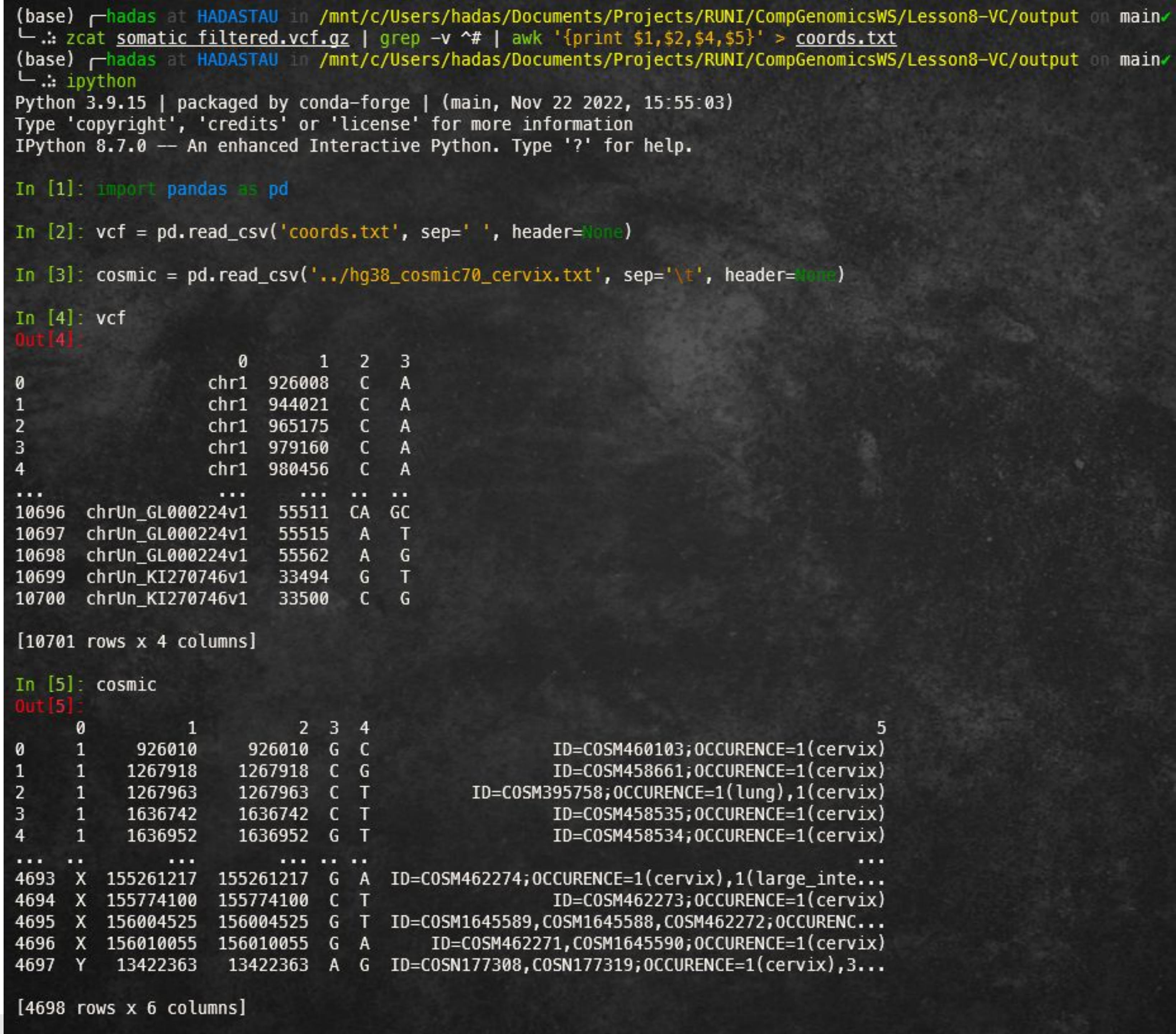
```
(base) ┌─hadas at HADASTAU in /mnt/c/Users/hadas/Documents/Projects/RUNI/CompGenomicsWS/Lesson8-VC/output on main✓
└─.∴ zcat somatic_filtered.vcf.gz | grep -v ^# | awk -v OFS='\t' '{print $1,$2,$4,$5,$7}' | head
chr1    926008  C       A       PASS
chr1    944021  C       A       PASS
chr1    965175  C       A       PASS
chr1    979160  C       A       strand_bias;weak_evidence
chr1    980456  C       A       PASS
chr1    1047656 C       A       strand_bias
chr1    1049316 C       A       PASS
chr1    1234903 C       T       PASS
chr1    1254707 C       A       PASS
chr1    1260983 C       A       PASS
```

Q8: How many filtered candidates are present in the final VCF?



```
(base) ┌─hadas at HADASTAU in /mnt/c/Users/hadas/Documents/Projects/RUNI/CompGenomicsWS/Lesson8-VC/output on main✓
└─.∴ zcat somatic_filtered.vcf.gz | grep -v ^# | awk -v OFS='\t' '{print $1,$2,$4,$5,$7,$8}' | wc -l
10701
```

Q9: Are all mutations found in the database are also present in our current case? Give one assumption on why this is so?



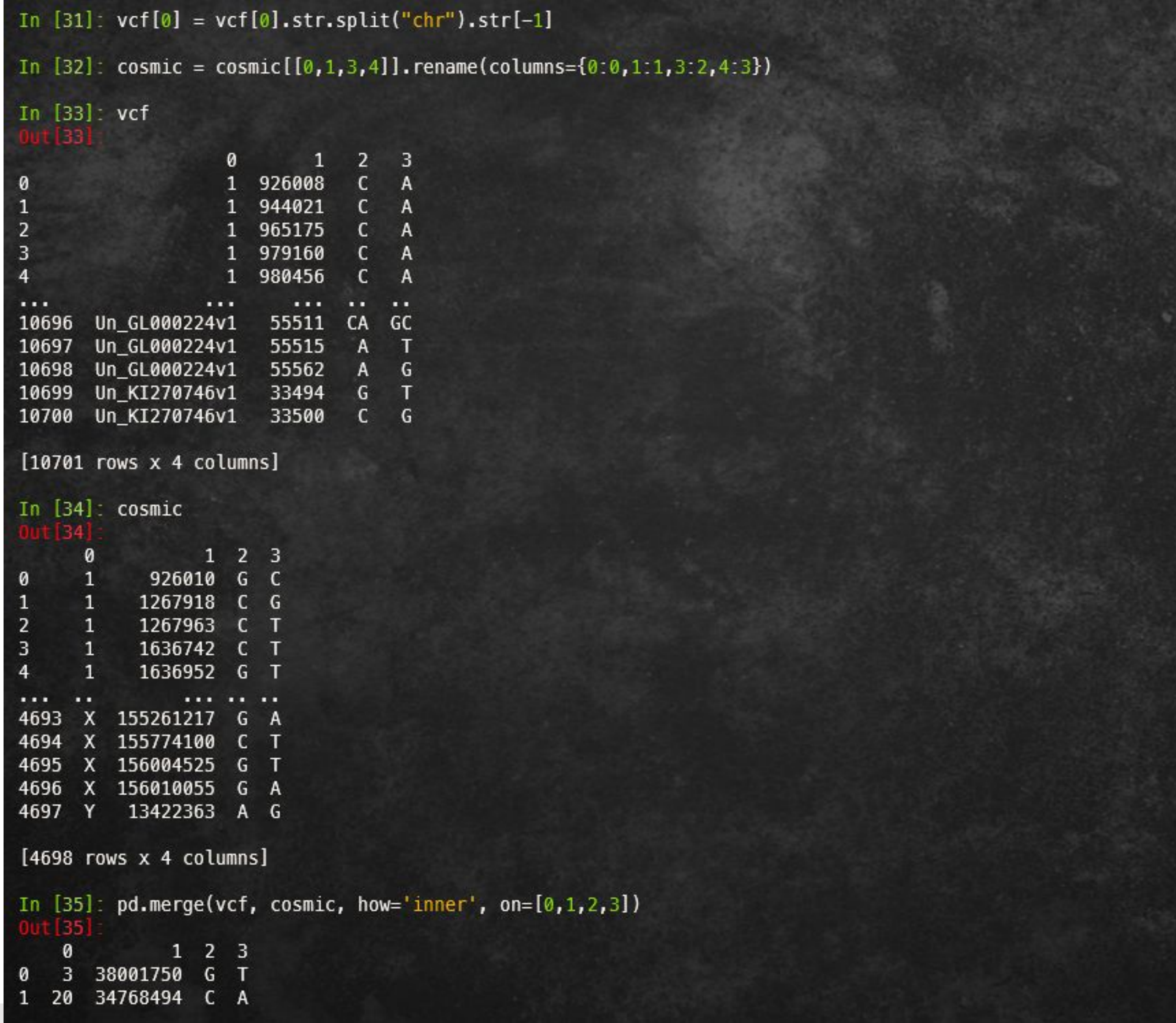
```
(base) ┌─hadas at HADASTAU in /mnt/c/Users/hadas/Documents/Projects/RUNI/CompGenomicsWS/Lesson8-VC/output on main✓
└─∴ zcat somatic_filtered.vcf.gz | grep -v ^# | awk '{print $1,$2,$4,$5}' > coords.txt
(base) ┌─hadas at HADASTAU in /mnt/c/Users/hadas/Documents/Projects/RUNI/CompGenomicsWS/Lesson8-VC/output on main✓
└─∴ ipython
Python 3.9.15 | packaged by conda-forge | (main, Nov 22 2022, 15:55:03)
Type 'copyright', 'credits' or 'license' for more information
IPython 8.7.0 -- An enhanced Interactive Python. Type '?' for help.

In [1]: import pandas as pd

In [2]: vcf = pd.read_csv('coords.txt', sep=' ', header=None)

In [3]: cosmic = pd.read_csv('../hg38_cosmic70_cervix.txt', sep='\t', header=None)

In [4]: vcf
Out[4]:
                     0       1   2   3
0                 chr1  926008   C   A
1                 chr1  944021   C   A
2                 chr1  965175   C   A
3                 chr1  979160   C   A
4                 chr1  980456   C   A
...                ...     ...  ..  ..
10696  chrUn_GL000224v1   55511  CA  GC
10697  chrUn_GL000224v1   55515   A   T
10698  chrUn_GL000224v1   55562   A   G
10699  chrUn_KI270746v1   33494   G   T
10700  chrUn_KI270746v1   33500   C   G

[10701 rows x 4 columns]

In [5]: cosmic
Out[5]:
      0          1          2  3  4                                                  5
0     1     926010     926010  G  C                   ID=COSM460103;OCCURENCE=1(cervix)
1     1    1267918    1267918  C  G                   ID=COSM458661;OCCURENCE=1(cervix)
2     1    1267963    1267963  C  T           ID=COSM395758;OCCURENCE=1(lung),1(cervix)
3     1    1636742    1636742  C  T                   ID=COSM458535;OCCURENCE=1(cervix)
4     1    1636952    1636952  G  T                   ID=COSM458534;OCCURENCE=1(cervix)
...  ..        ...        ... .. ..                                                ...
4693  X  155261217  155261217  G  A  ID=COSM462274;OCCURENCE=1(cervix),1(large_inte...
4694  X  155774100  155774100  C  T                   ID=COSM462273;OCCURENCE=1(cervix)
4695  X  156004525  156004525  G  T  ID=COSM1645589,COSM1645588,COSM462272;OCCURENC...
4696  X  156010055  156010055  G  A      ID=COSM462271,COSM1645590;OCCURENCE=1(cervix)
4697  Y   13422363   13422363  A  G  ID=COSN177308,COSN177319;OCCURENCE=1(cervix),3...

[4698 rows x 6 columns]
```

Q9: Are all mutations found in the database are also present in our current case? Give one assumption on why this is so?

Exome

Low frequency mutations

Variant caller



```
In [31]: vcf[0] = vcf[0].str.split("chr").str[-1]

In [32]: cosmic = cosmic[[0,1,3,4]].rename(columns={0:0,1:1,3:2,4:3})

In [33]: vcf
Out[33]:
                   0       1   2   3
0                  1  926008   C   A
1                  1  944021   C   A
2                  1  965175   C   A
3                  1  979160   C   A
4                  1  980456   C   A
...              ...     ...  ..  ..
10696  Un_GL000224v1   55511  CA  GC
10697  Un_GL000224v1   55515   A   T
10698  Un_GL000224v1   55562   A   G
10699  Un_KI270746v1   33494   G   T
10700  Un_KI270746v1   33500   C   G

[10701 rows x 4 columns]

In [34]: cosmic
Out[34]:
      0          1  2  3
0     1     926010  G  C
1     1    1267918  C  G
2     1    1267963  C  T
3     1    1636742  C  T
4     1    1636952  G  T
...  ..        ... .. ..
4693  X  155261217  G  A
4694  X  155774100  C  T
4695  X  156004525  G  T
4696  X  156010055  G  A
4697  Y   13422363  A  G

[4698 rows x 4 columns]

In [35]: pd.merge(vcf, cosmic, how='inner', on=[0,1,2,3])
Out[35]:
    0         1  2  3
0   3  38001750  G  T
1  20  34768494  C  A
```

Lesson Goals

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

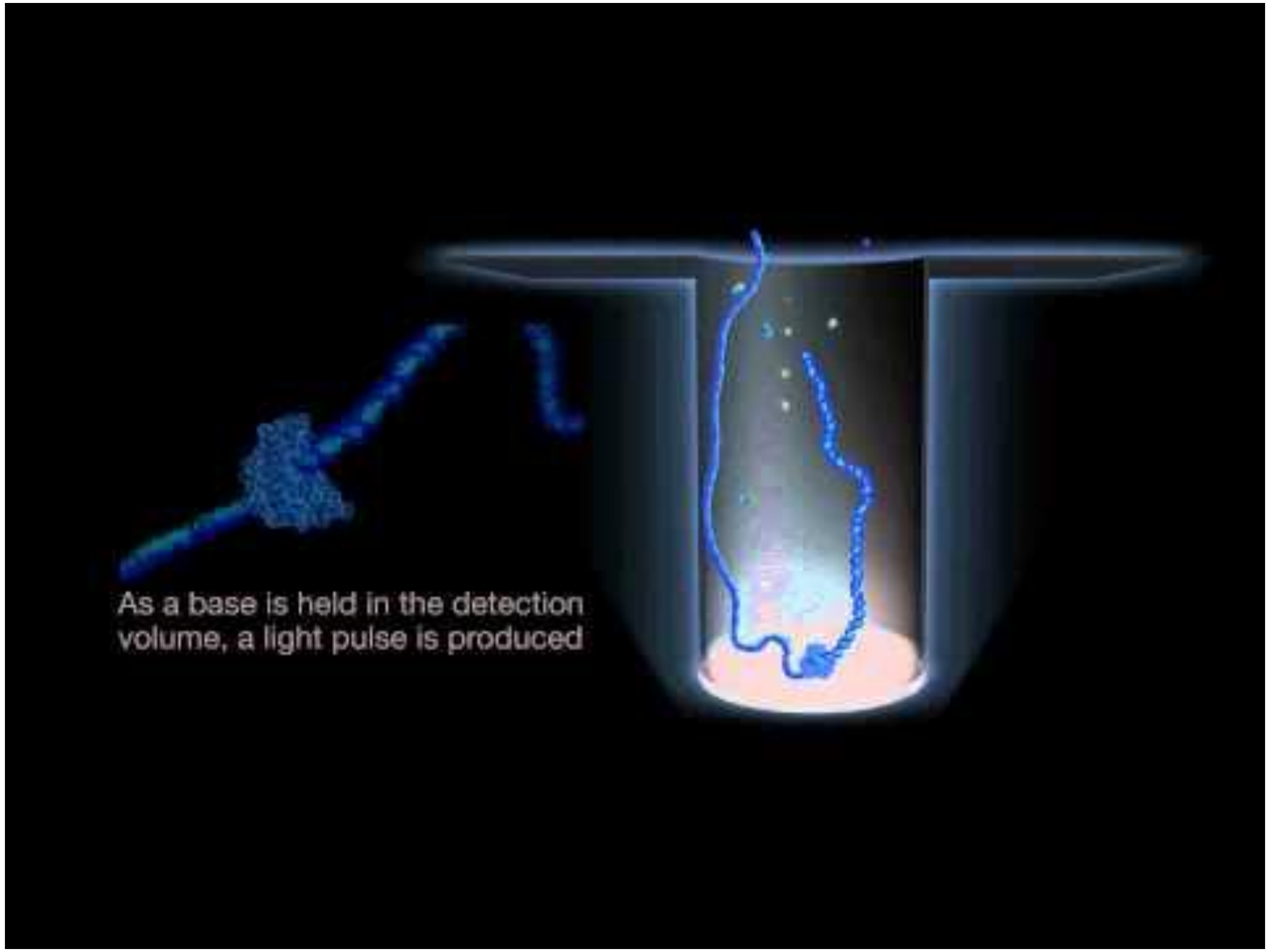
Single Molecule Real Time

No amplification step

Based on the ability to analyze very small volumes

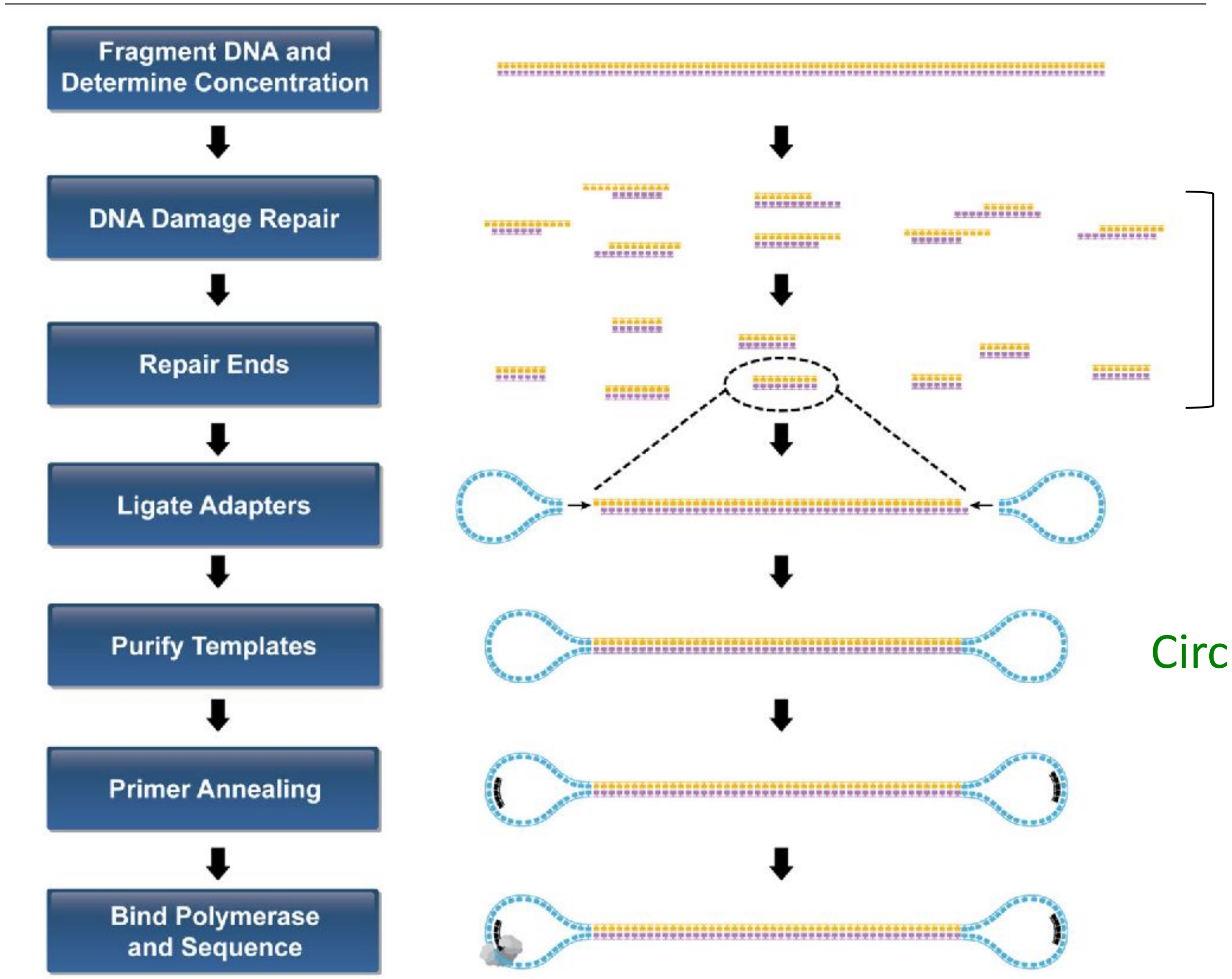
Sequencing by synthesis

Sequel II



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

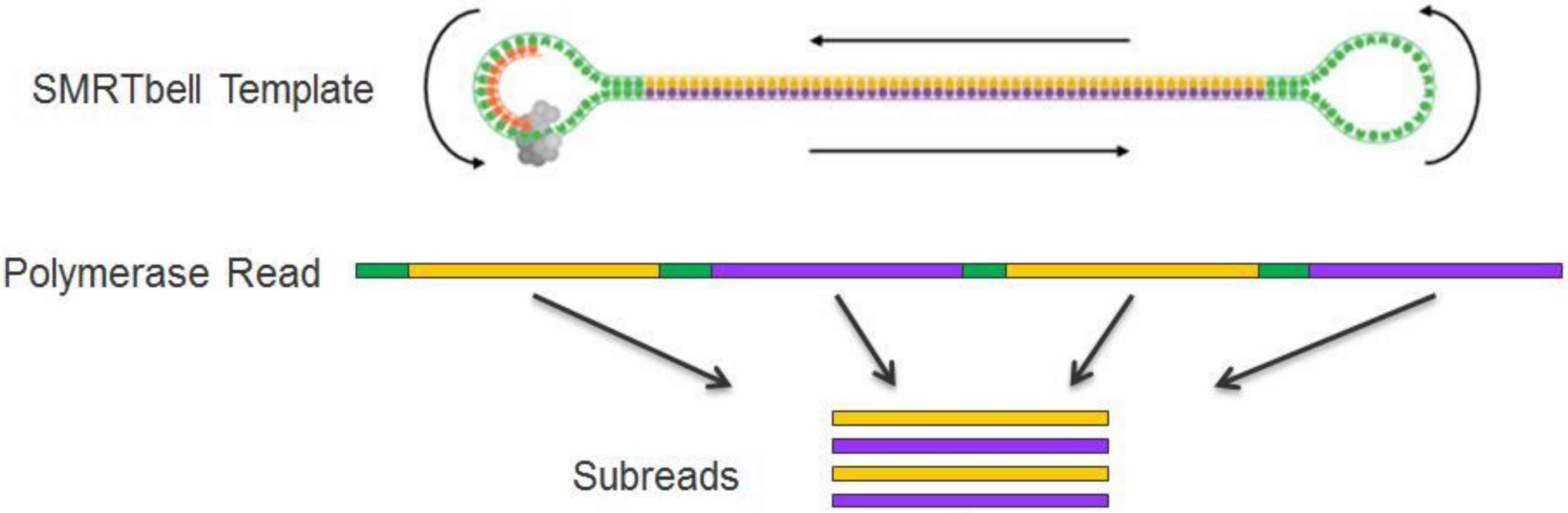
PacBio Library Prep



15-20kb insert

Circular

PacBio Sequencing



Properties of PacBio Sequencing

Read length

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

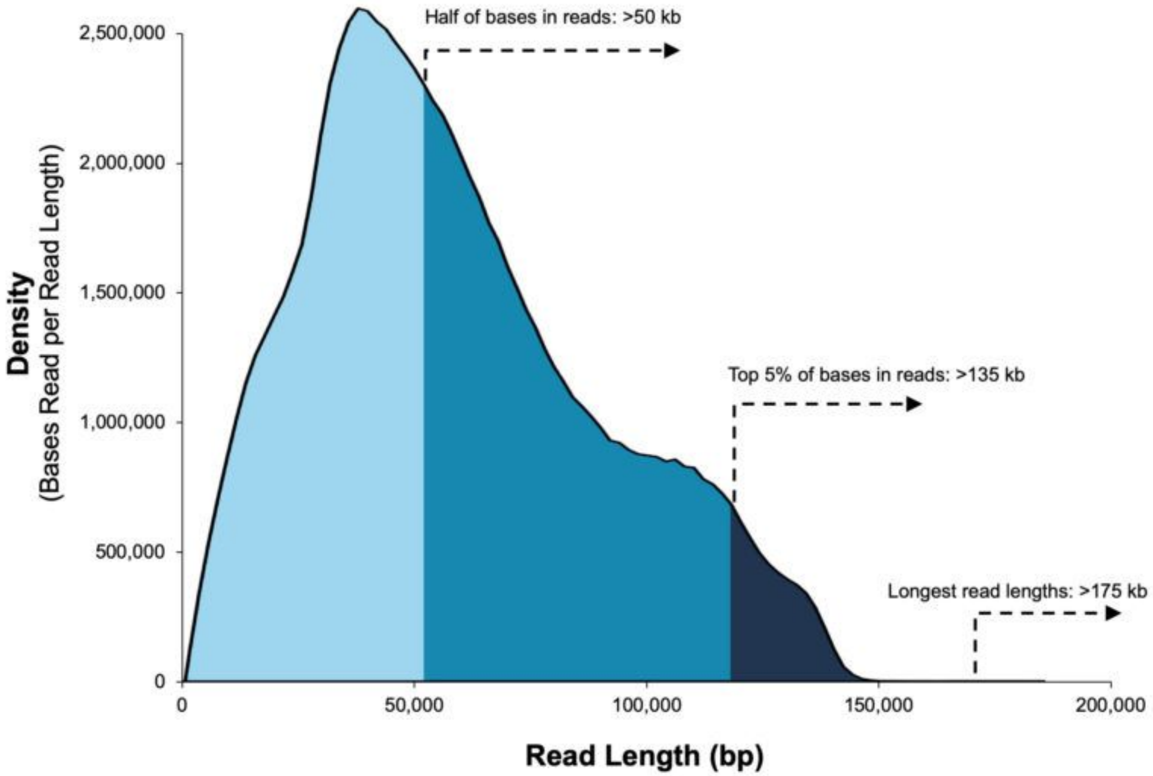
No paired-end option

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

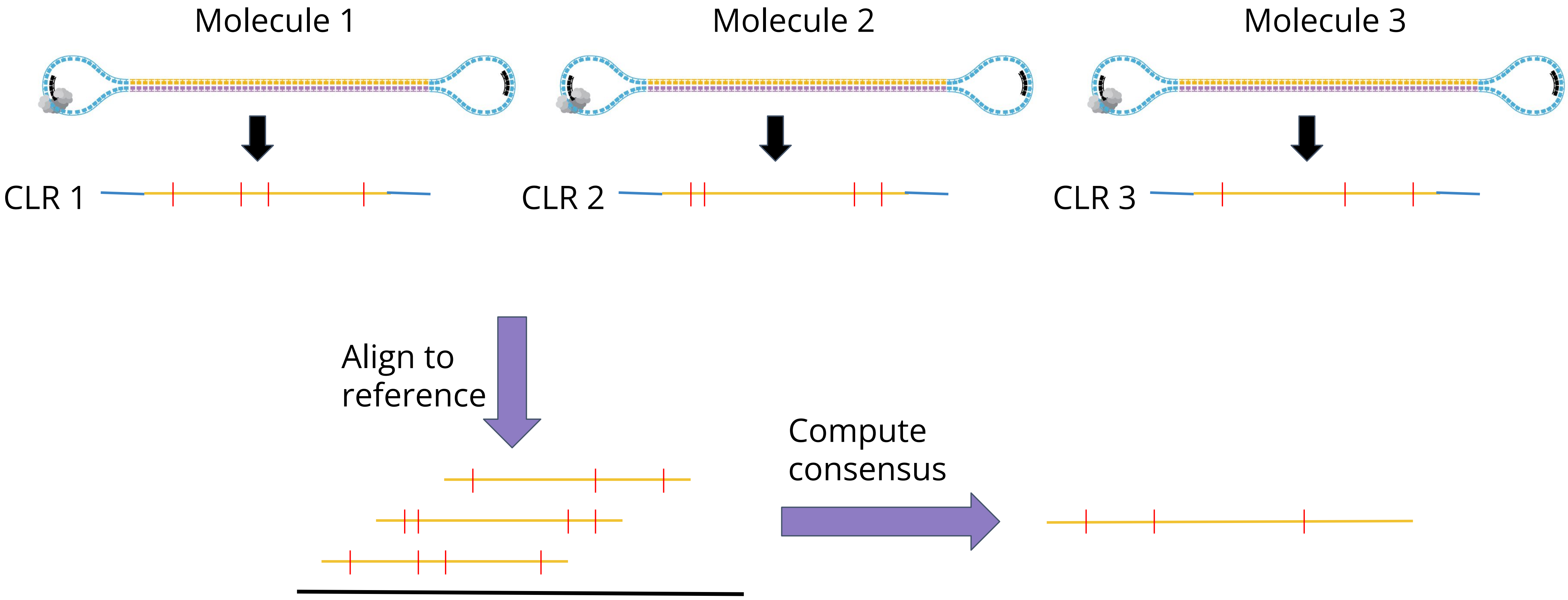
Polymerase read length ~= sub-read length

Align CLRs to a reference genome and correct errors

Find the consensus of multiple molecules

Accuracy increases with sequencing depth

CLR Error Correction



Dealing With High Error Rates

Option 2:

CCS - circular consensus read

Also called **HiFi reads**

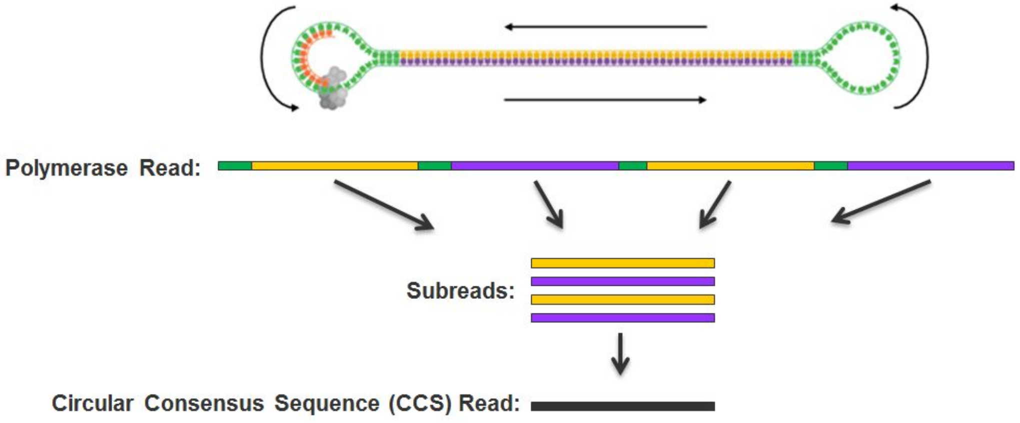
Polymerase read length > sub-read length

Align CCSs to one another and correct errors

Find the consensus of a single molecule

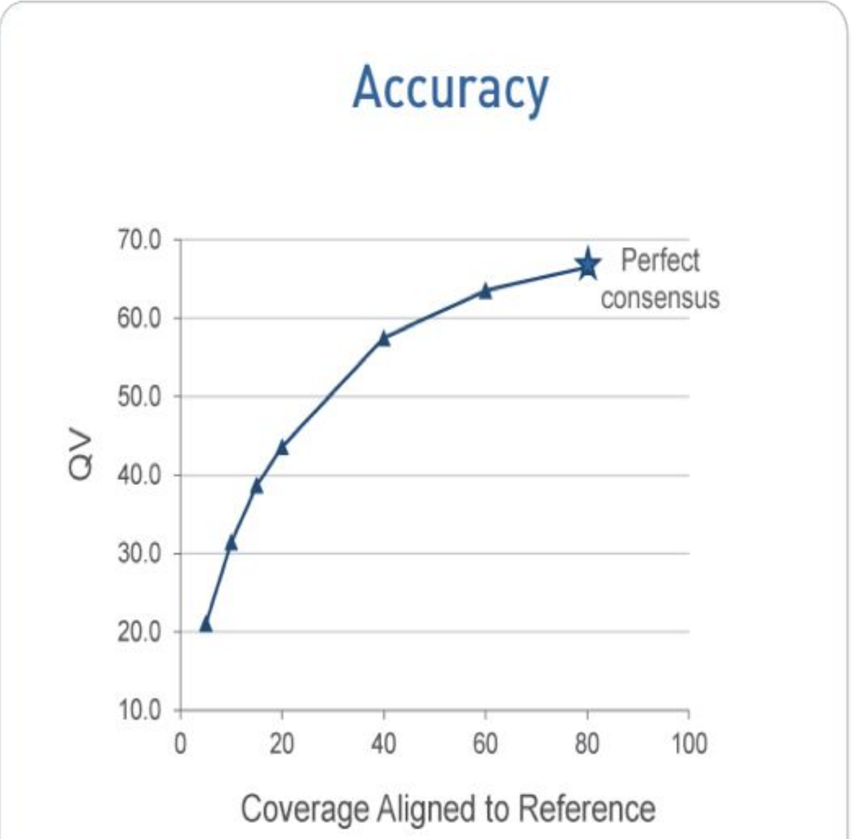
Accuracy >99%

Shorter reads (<20kb)

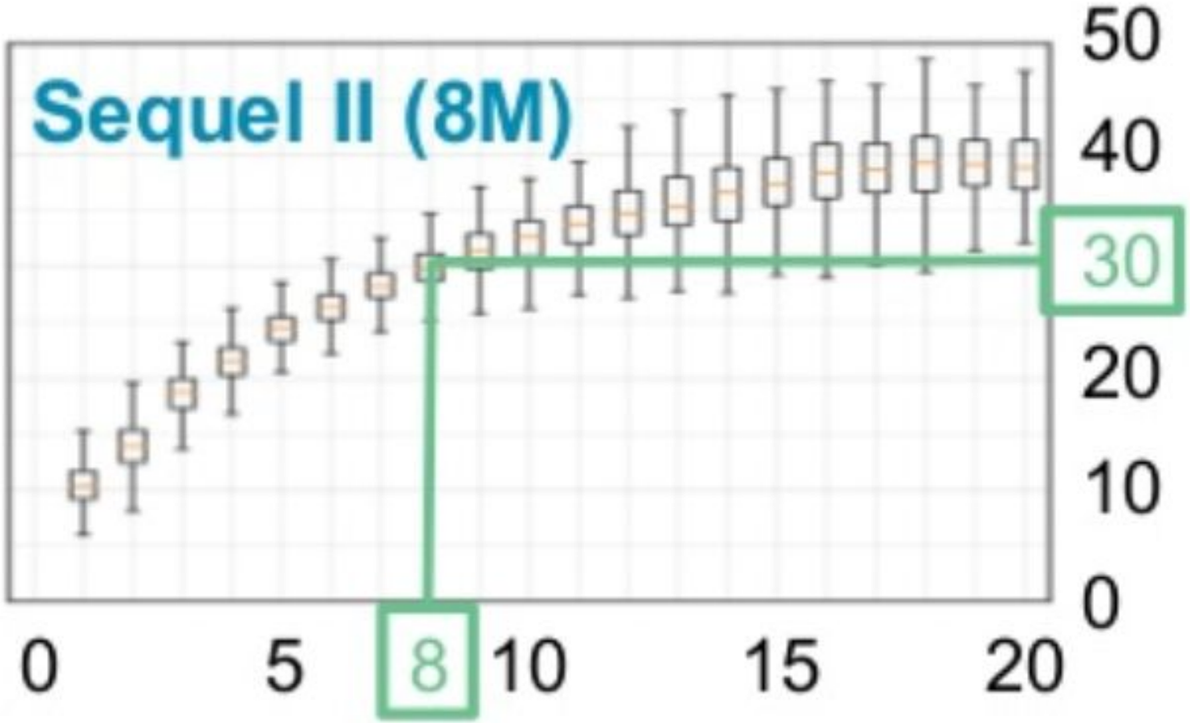


Accuracy CLR consensus Vs. CCS

CLR consensus



CCS



Passes

Oxford Nanopore Sequencing (ONT)

Single molecule

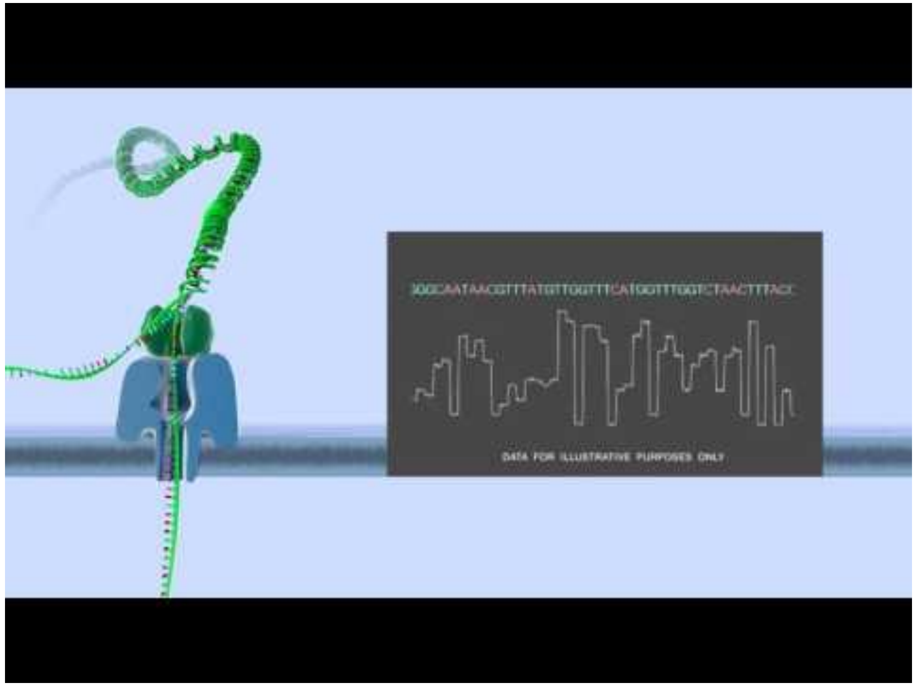
Real time

Not SBS

Palm-sized machine

MinION MkI: portable, real time biological analyses

MinION



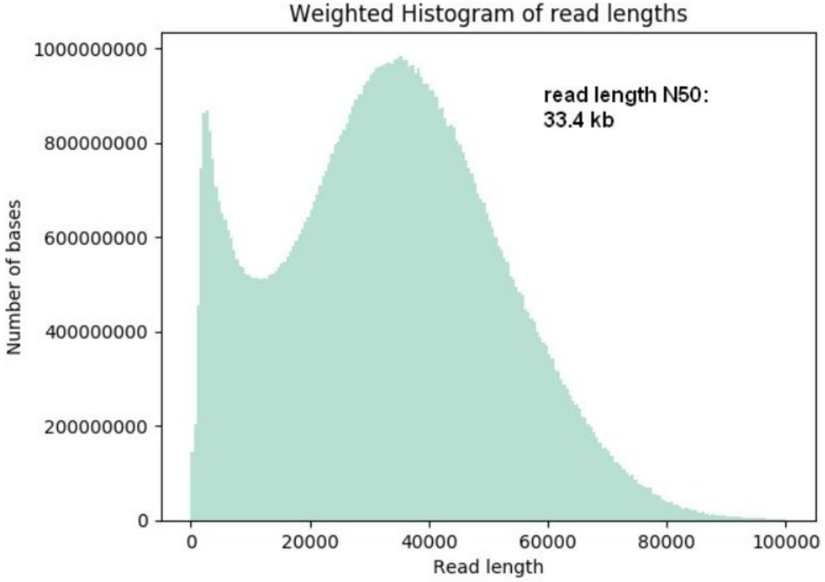
Properties of ONT Sequencing

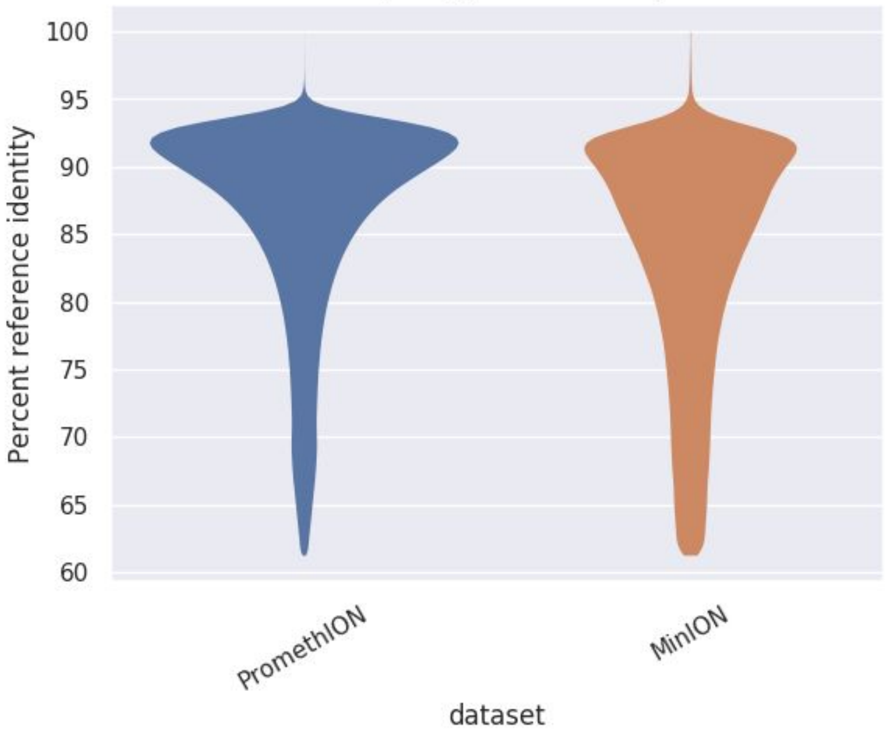
Read length - theoretically unlimited

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*

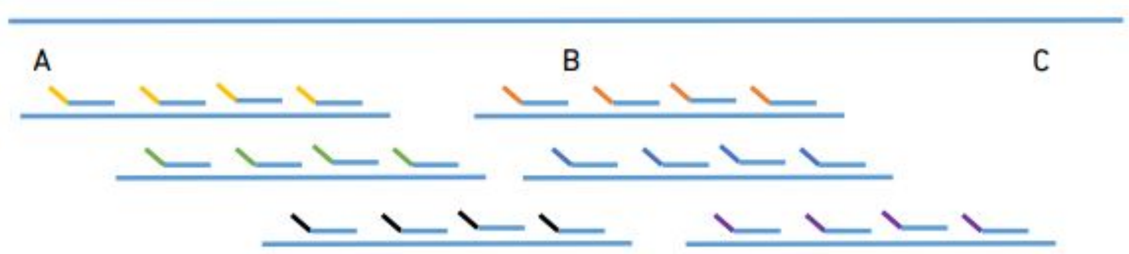
Amarasinghe, Shanika L., et al. "Opportunities and challenges in long-read sequencing data analysis." *Genome biology* 21.1 (2020): 1-16.

10X Genomics

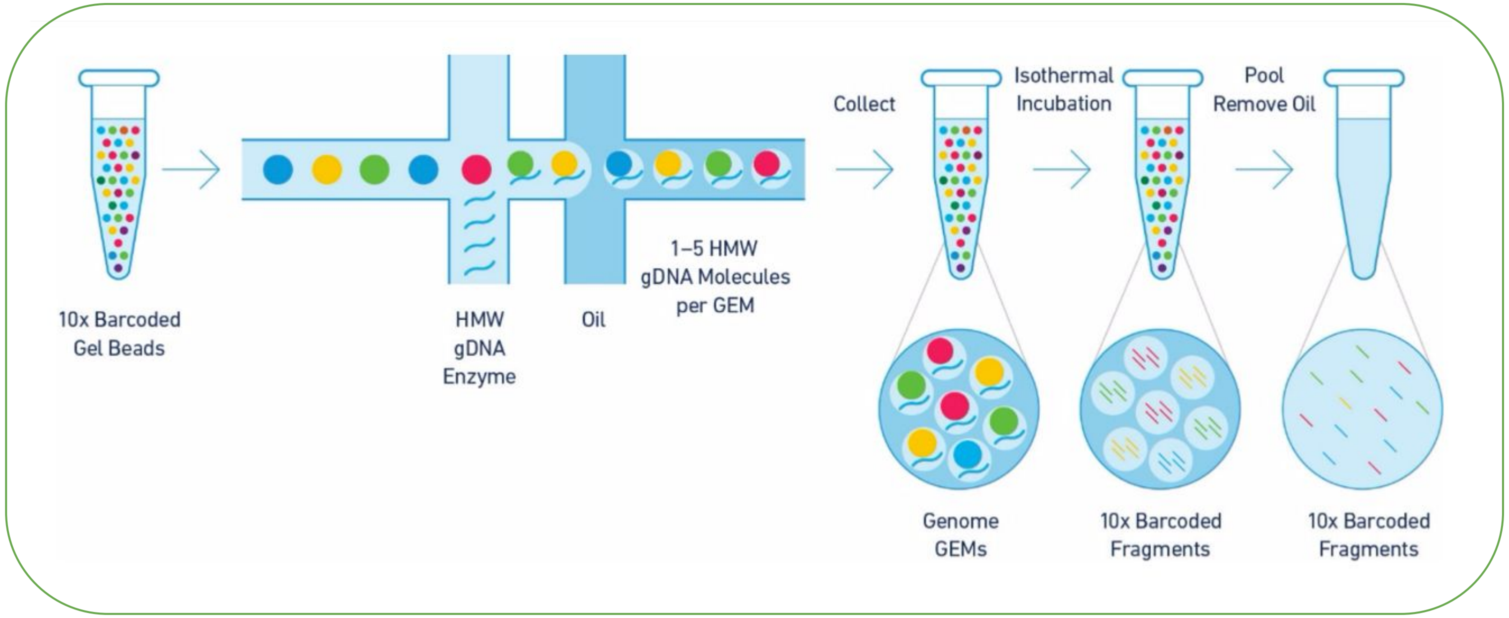
Not a long read technology

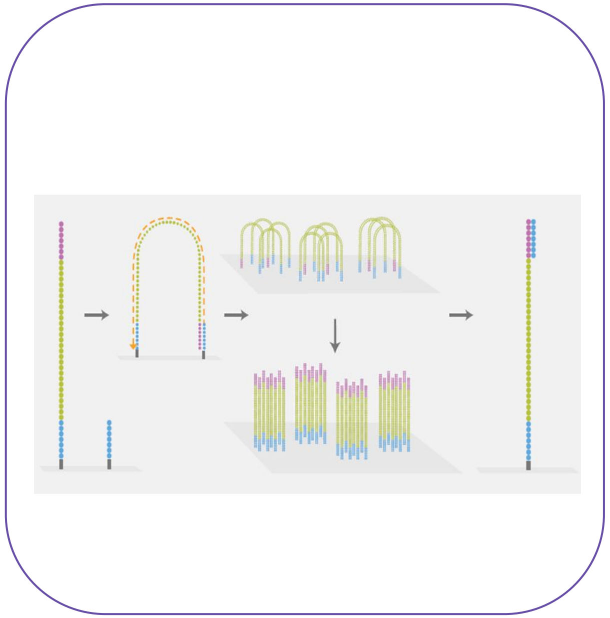
But provides long-range information through **linked reads**

Short reads originating from the same long molecule

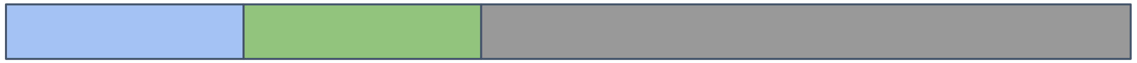


Based on standard short read Illumina technology





R1



Illumina adaptor

10X barcode

gDNA

Linked Reads

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

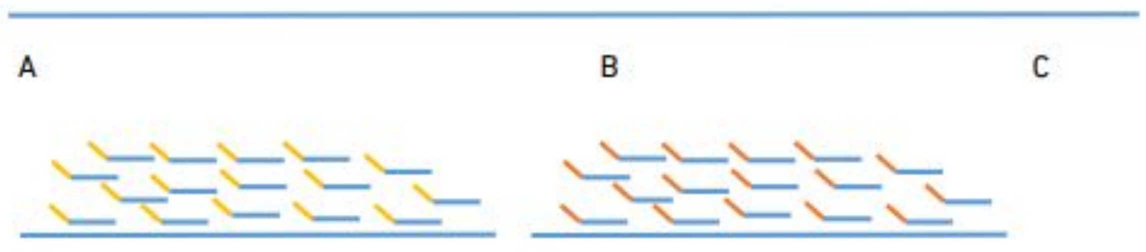
gDNA fragment size is usually 50-60kb

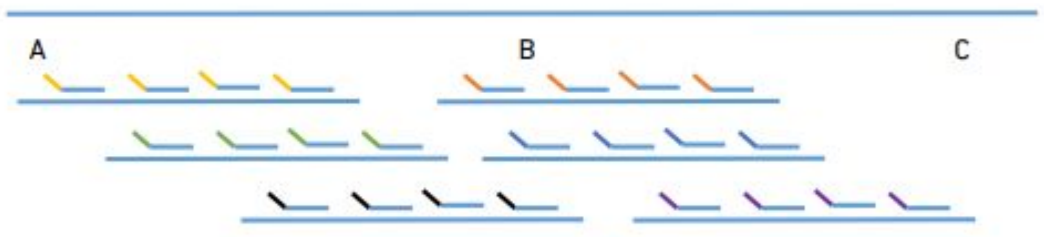
If ~x3 depth is used - we can produce “synthetic long reads”

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

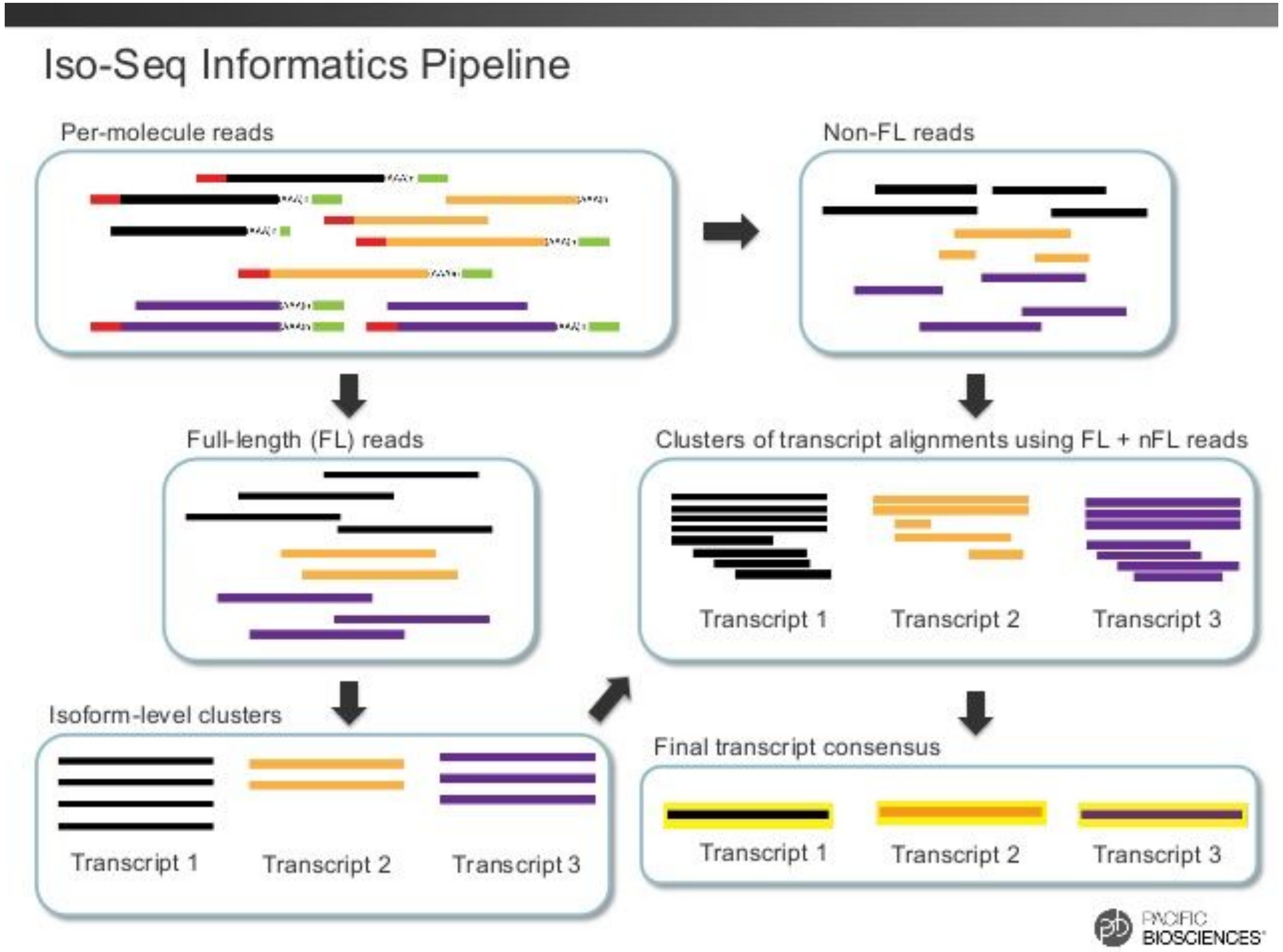
RNA-Seq and Long Reads

Read length is usually larger than mRNA size

Full-length transcripts

No transcript assembly is needed

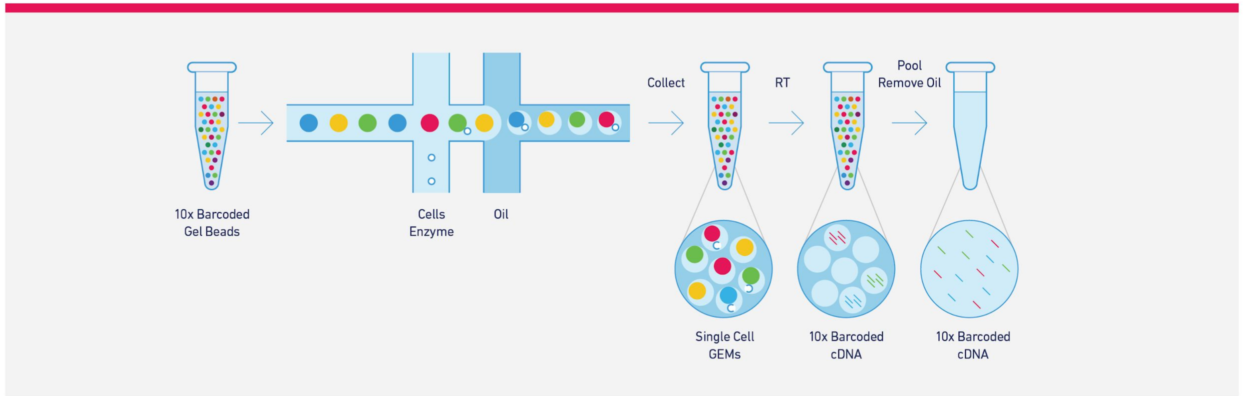
Easier to detect and quantify isoforms



10X for Single Cell RNA-Seq

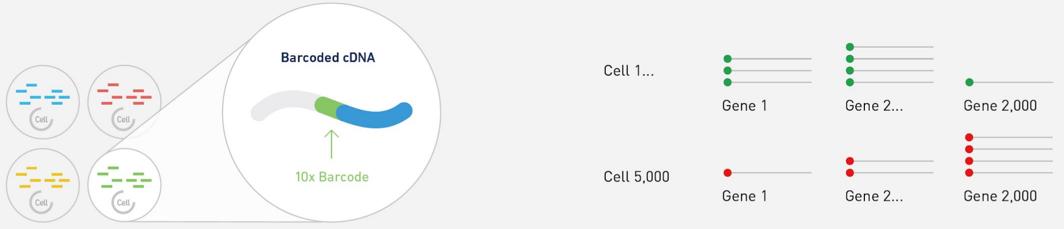
GemCode™ Technology for Single Cell Partitioning

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.

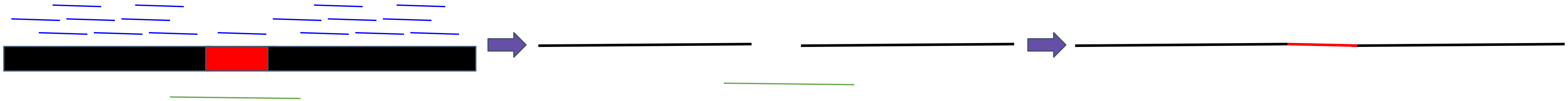
Long and Linked Reads in Genome Assembly

Many modern assemblers can work with 3rd generation reads

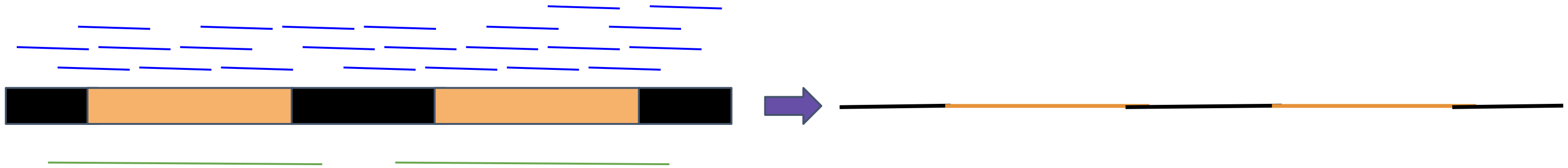
- Falcon - PacBio reads
- Canu, SPAdes - PacBio and ONT reads
- Supernova - 10X reads

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

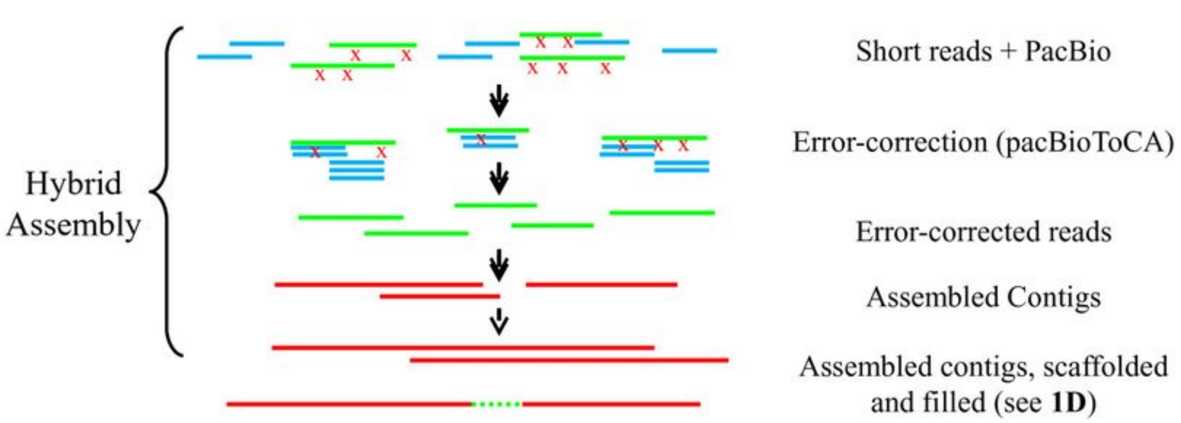
Long/linked reads can help link contigs by bridging over difficult regions



Long reads can help solve long repeats



Different Assembly Strategies



Hybrid Assembly

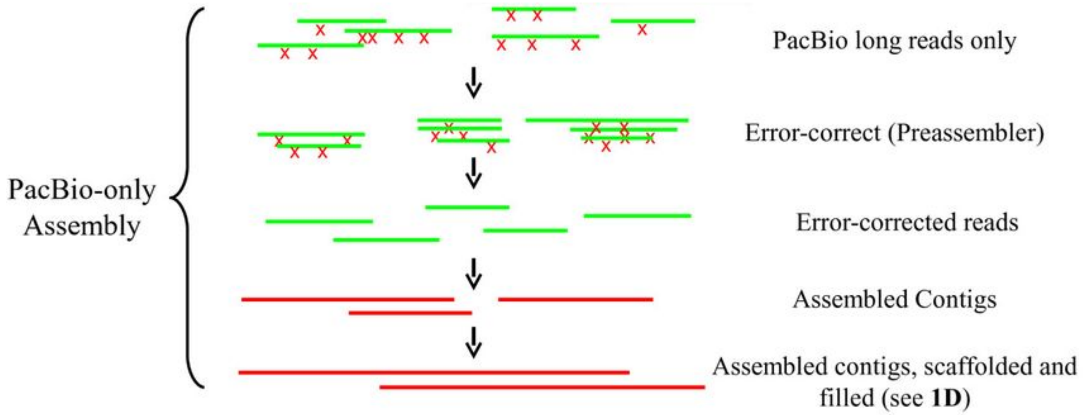
Short reads + PacBio

Error-correction (pacBioToCA)

Error-corrected reads

Assembled Contigs

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



PacBio-only Assembly

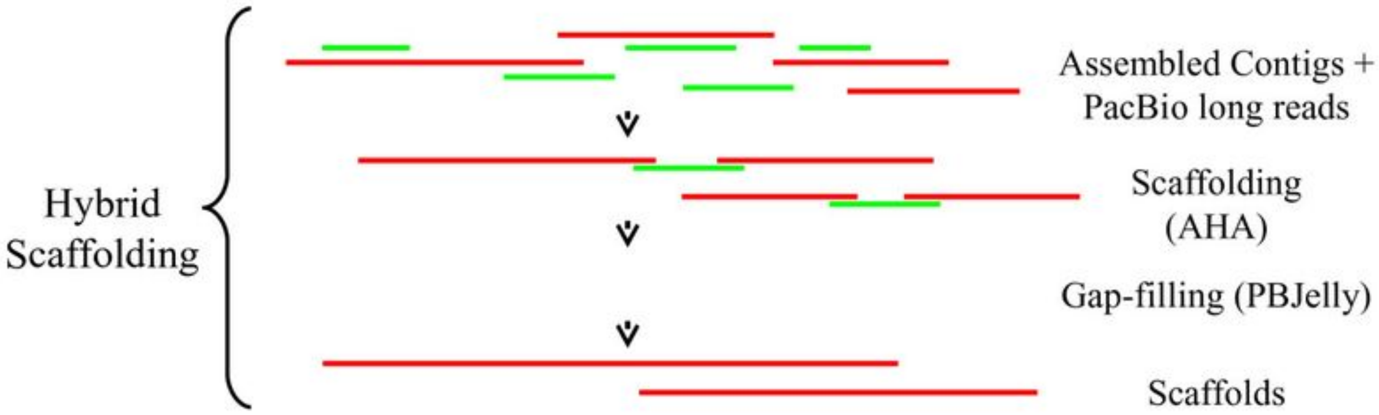
PacBio long reads only

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

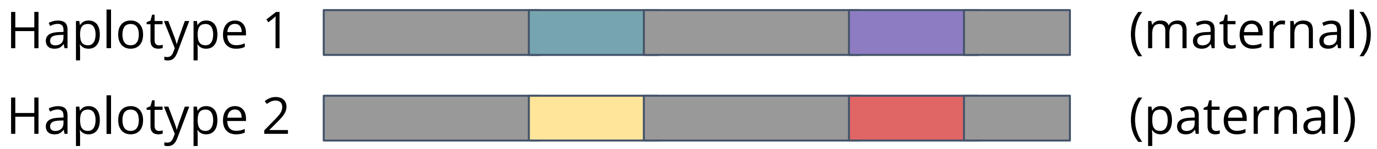
Many interesting eukaryote genomes are diploid or polyploid

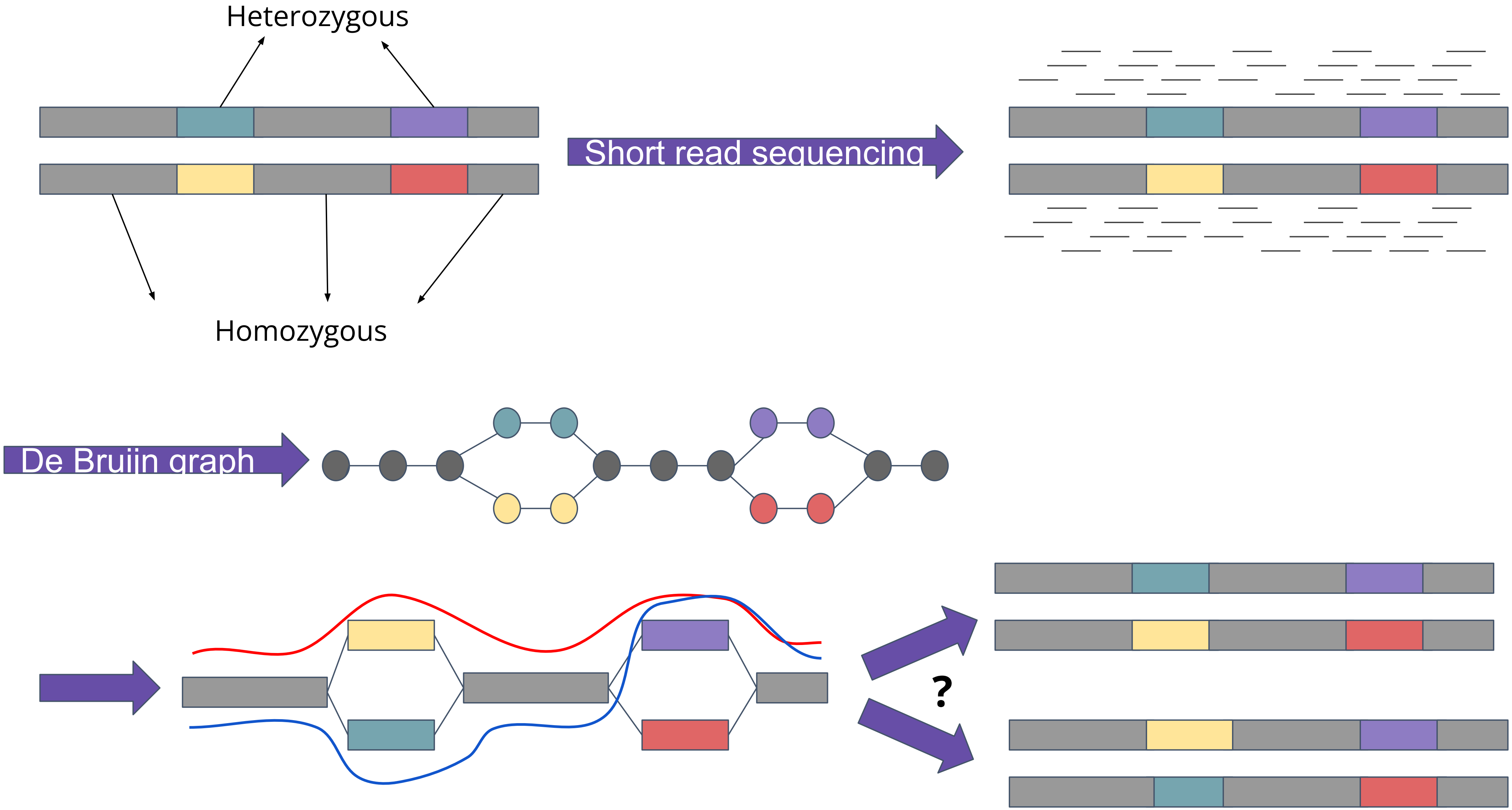
Still, most assemblies are haploid

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**





Heterozygous

Homozygous

Short read sequencing

De Bruijn graph

?

Structural Variant Detection

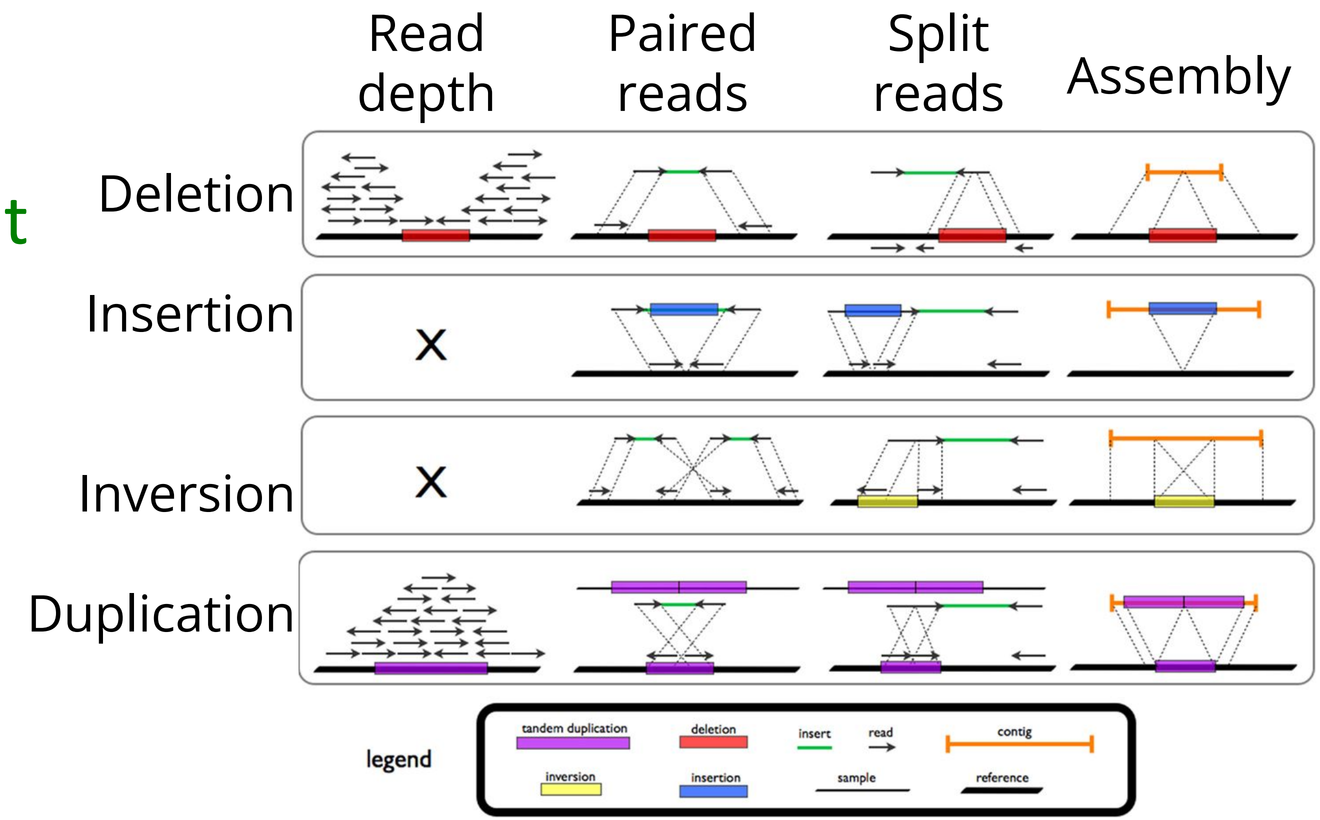
SVs are generally hard to detect with short reads

Many SVs are located in regions that are hard to sequence

SV detection is usually based on mapping reads to a reference

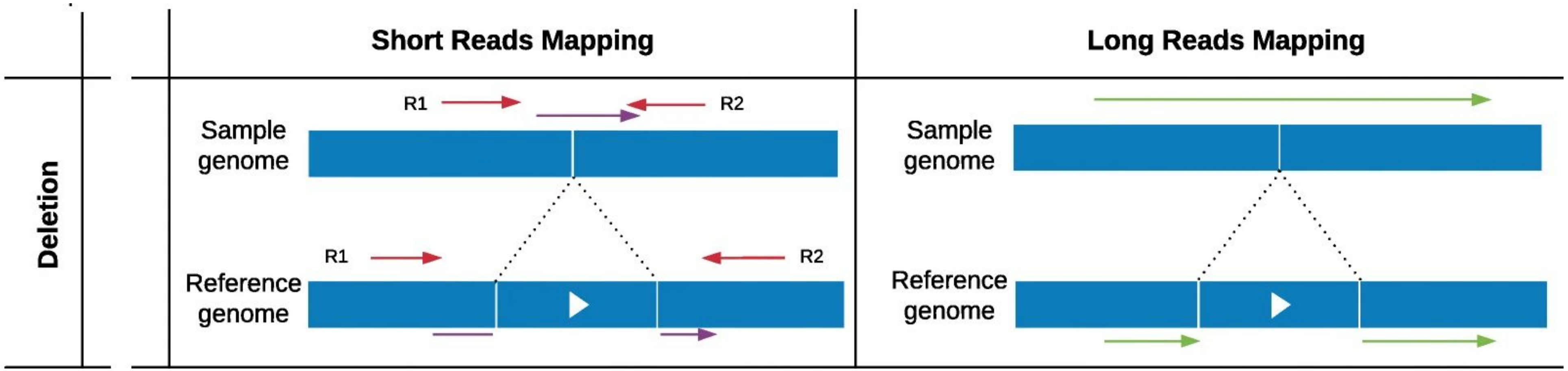
Long reads are useful because:

- 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

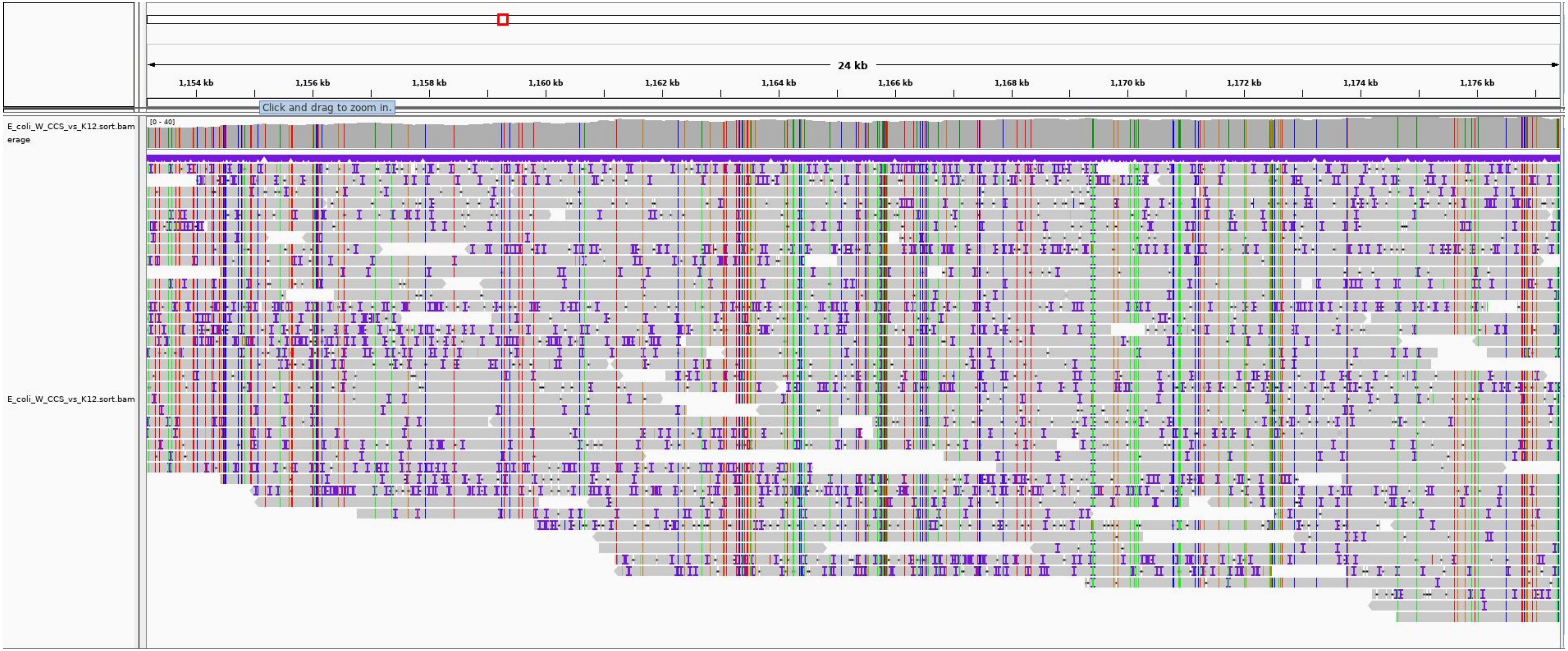
Minimap2 is a generic sequence mapping software

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!