DNA Sequencing and Data Analysis

Prof Noam Shomron

Hadas Volkov

Lecture 10, January 6, 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	**Nanopore data analysis introduction 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, Assaf
12, 20.1	Nanopore data analysis	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

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

Know how to use Minimap2 for long read alignment

Be able to perform SV calling from long read data using Sniffles

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

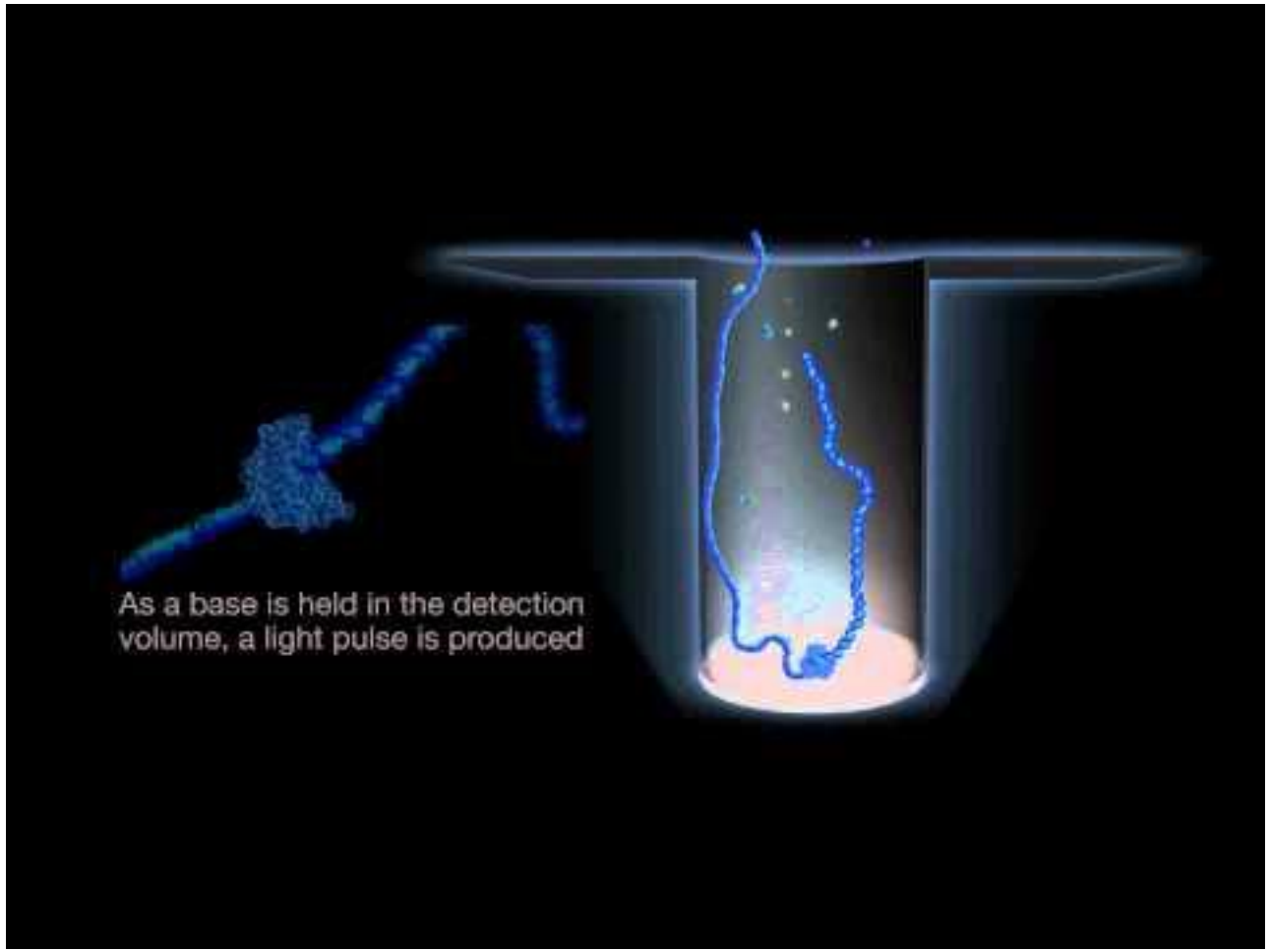
Single Molecule Real Time

No amplification step

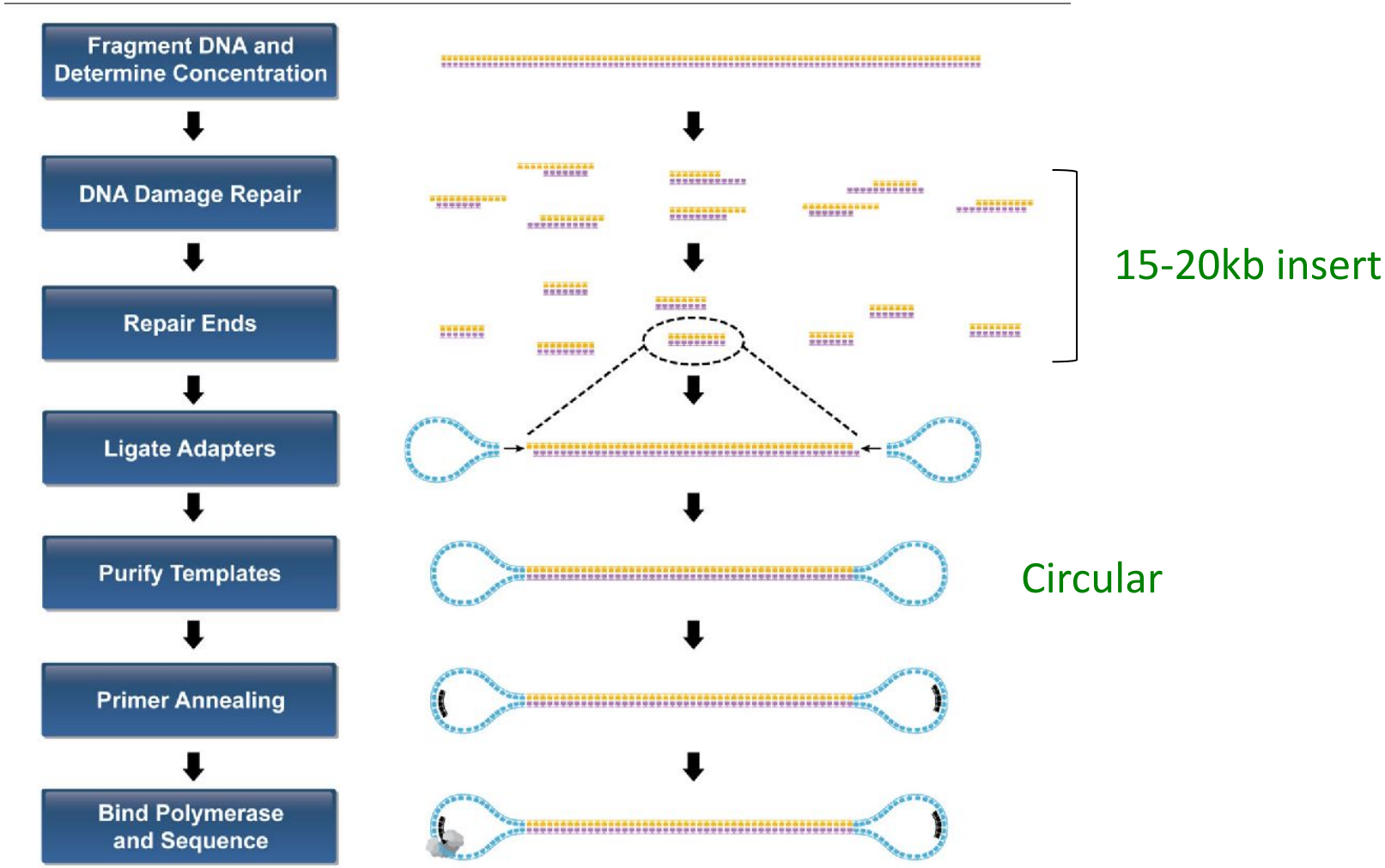
Based on the ability to analyze very small volumes

Sequencing by synthesis

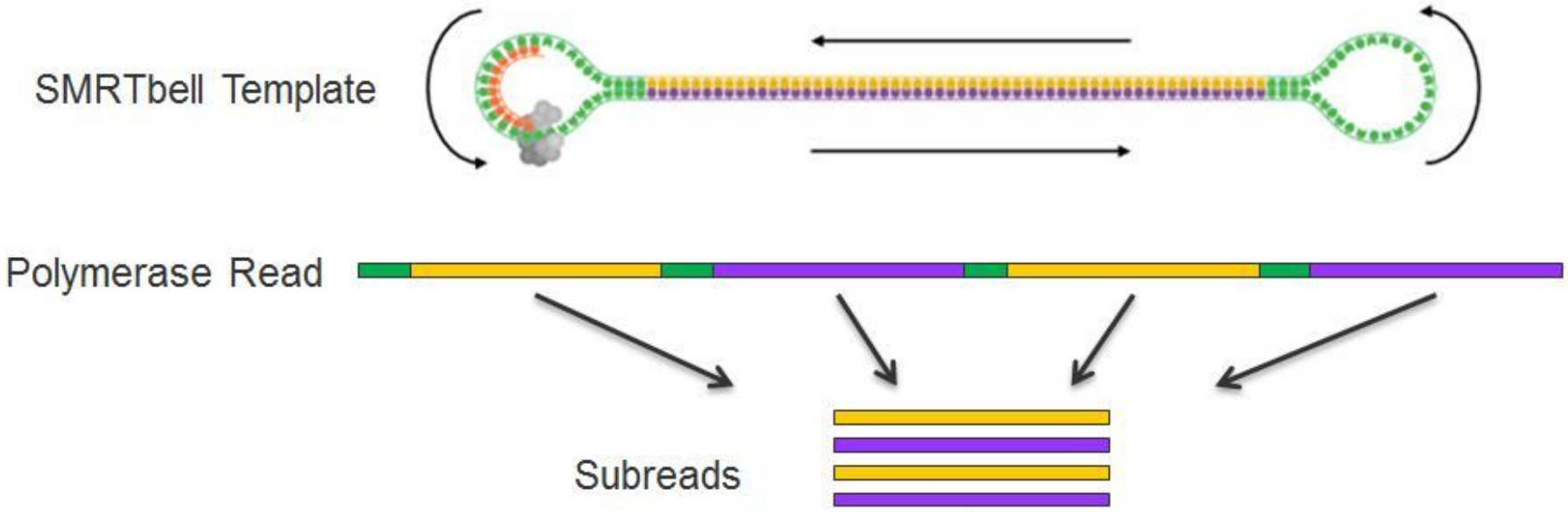
Sequel II



PacBio Library Prep



PacBio Sequencing



Properties of PacBio Sequencing

Read length

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

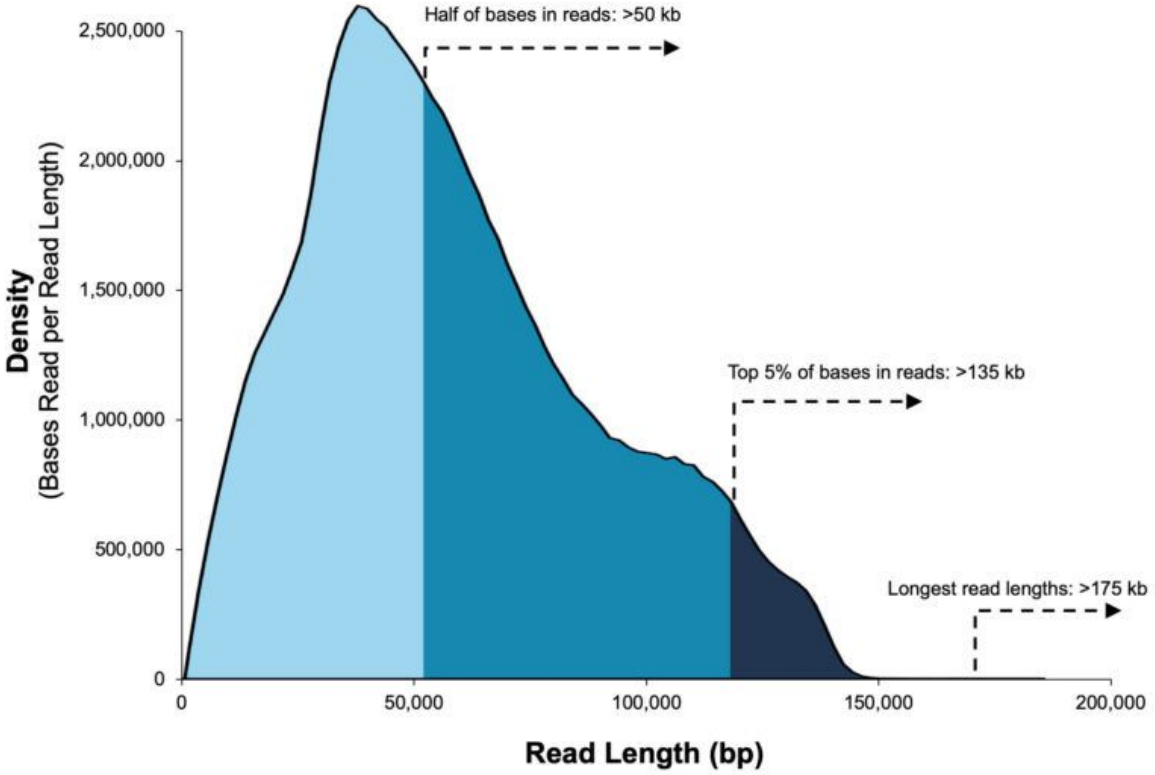
No paired-end option

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

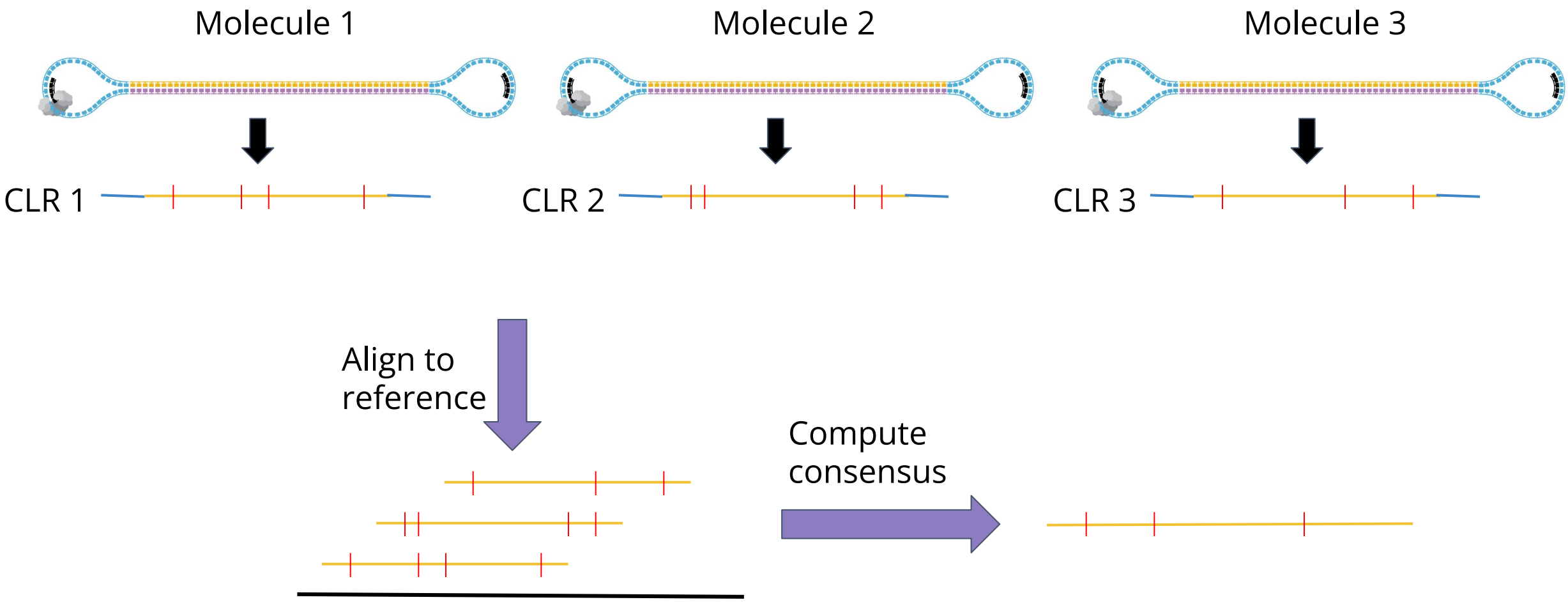
Polymerase read length = sub-read length

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**

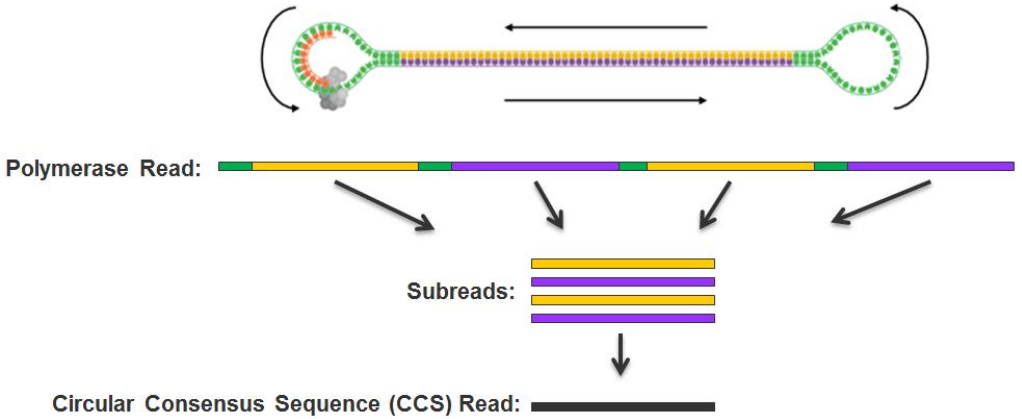
Polymerase read length > sub-read length

Align CCSs to one another and correct errors

Find the consensus of a single molecule

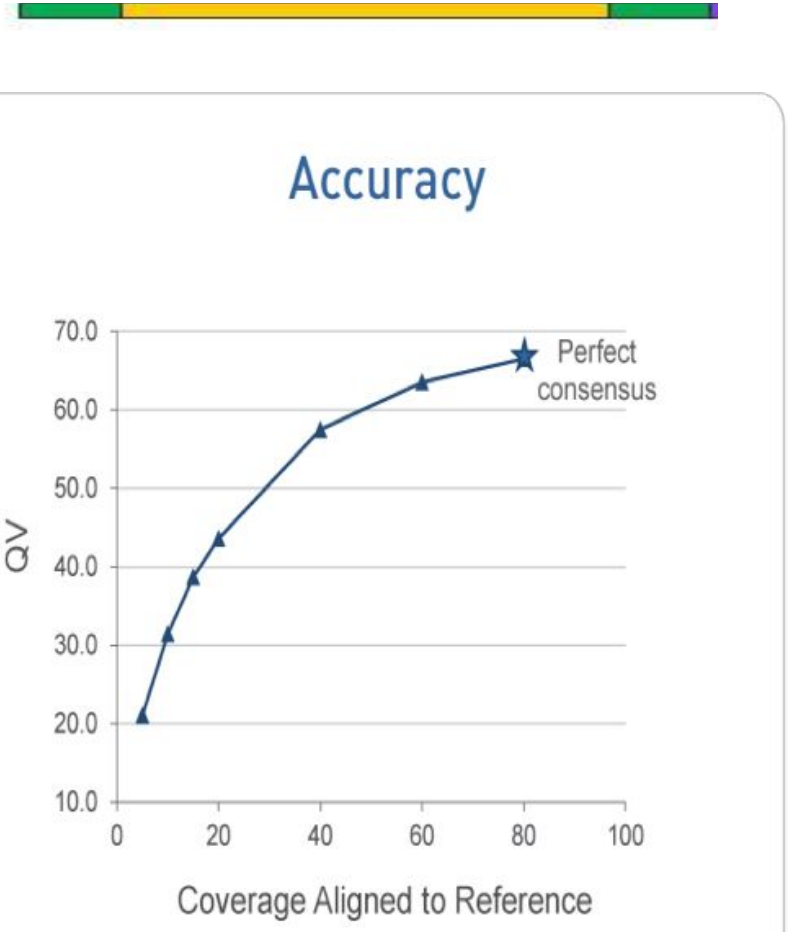
Accuracy >99%

Shorter reads (<20kb)

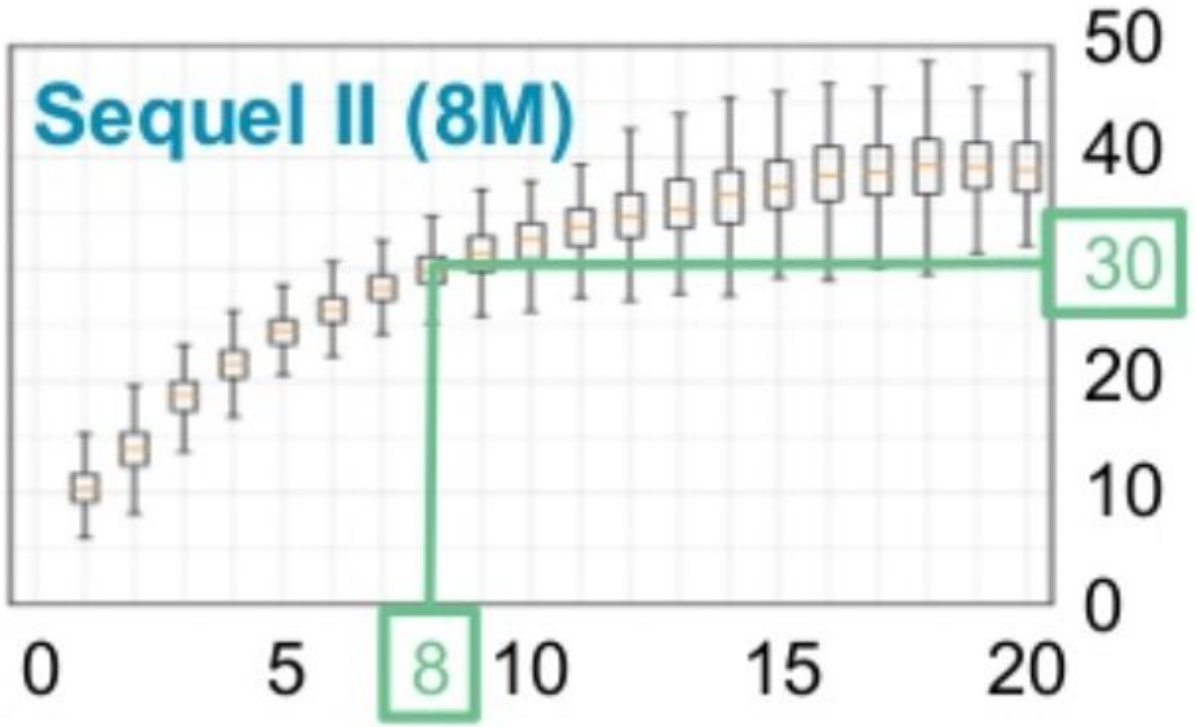


Accuracy CLR consensus Vs. CCS

CLR consensus



CCS



Oxford Nanopore Sequencing (ONT)

Single molecule

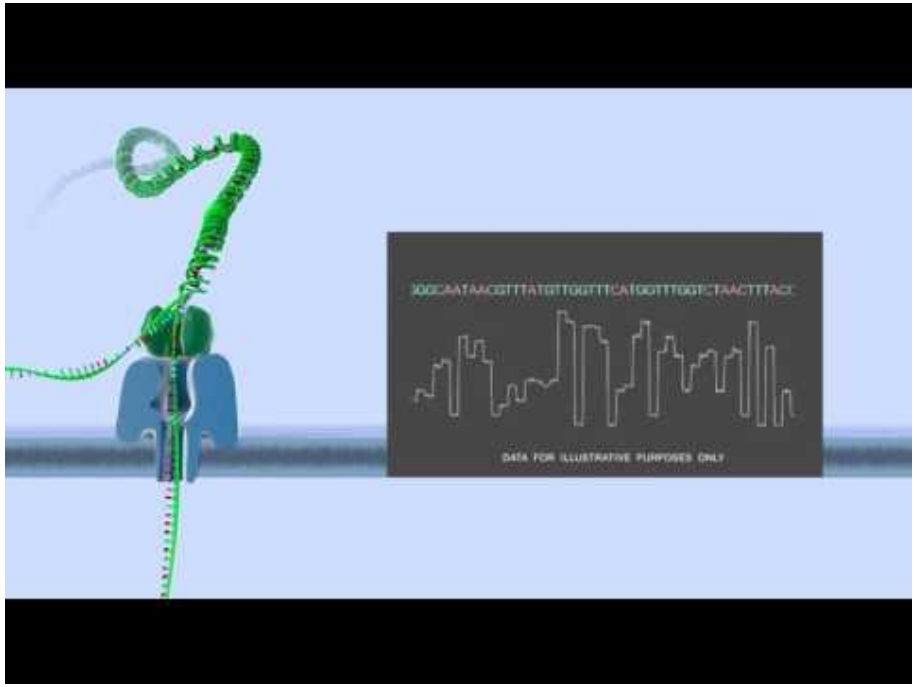
Real time

Not SBS

Palm-sized machine

MinION Mkl: portable, real time biological analyses

MinION



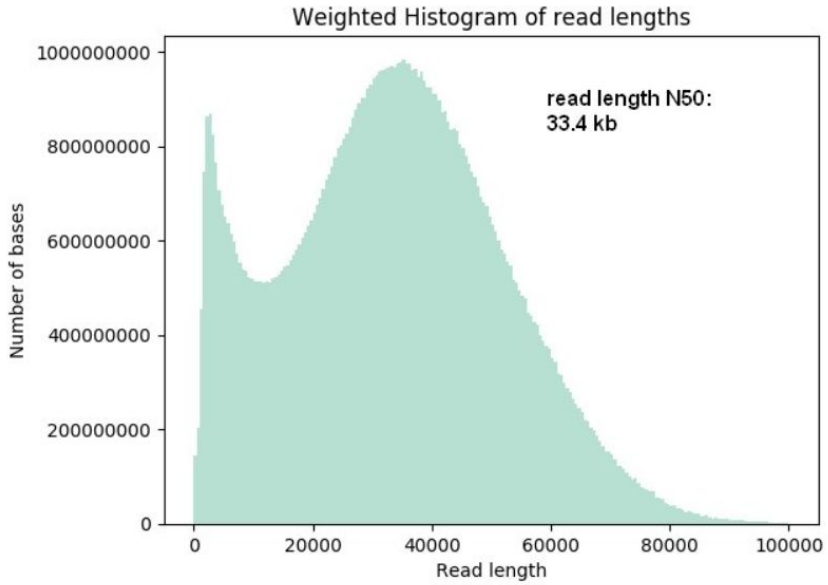
Properties of ONT Sequencing

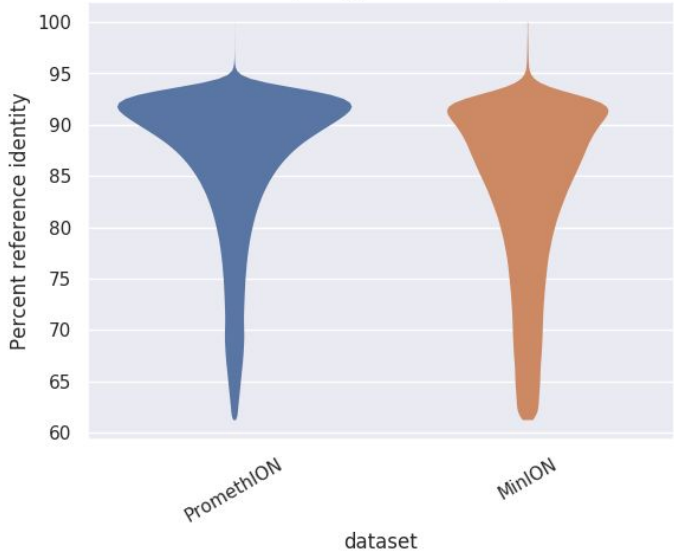
Read length - theoretically unlimited

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		
Throughput	7 Gb/h	2.5 Gb/h		

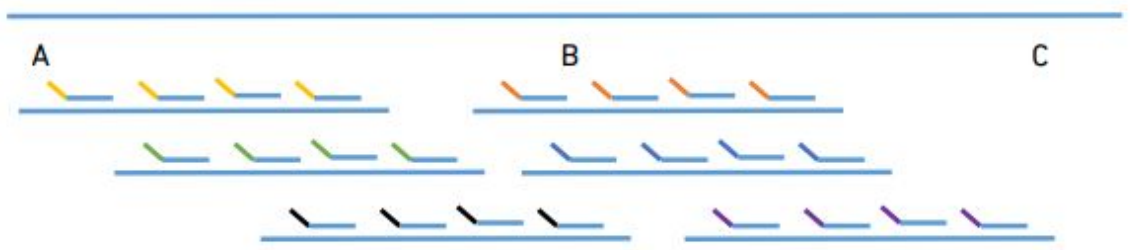
Amarasinghe, Shanika L., et al. "Opportunities and challenges in long-read sequencing data analysis." *Genome biology* 21.1 (2020): 1-16.

10X Genomics

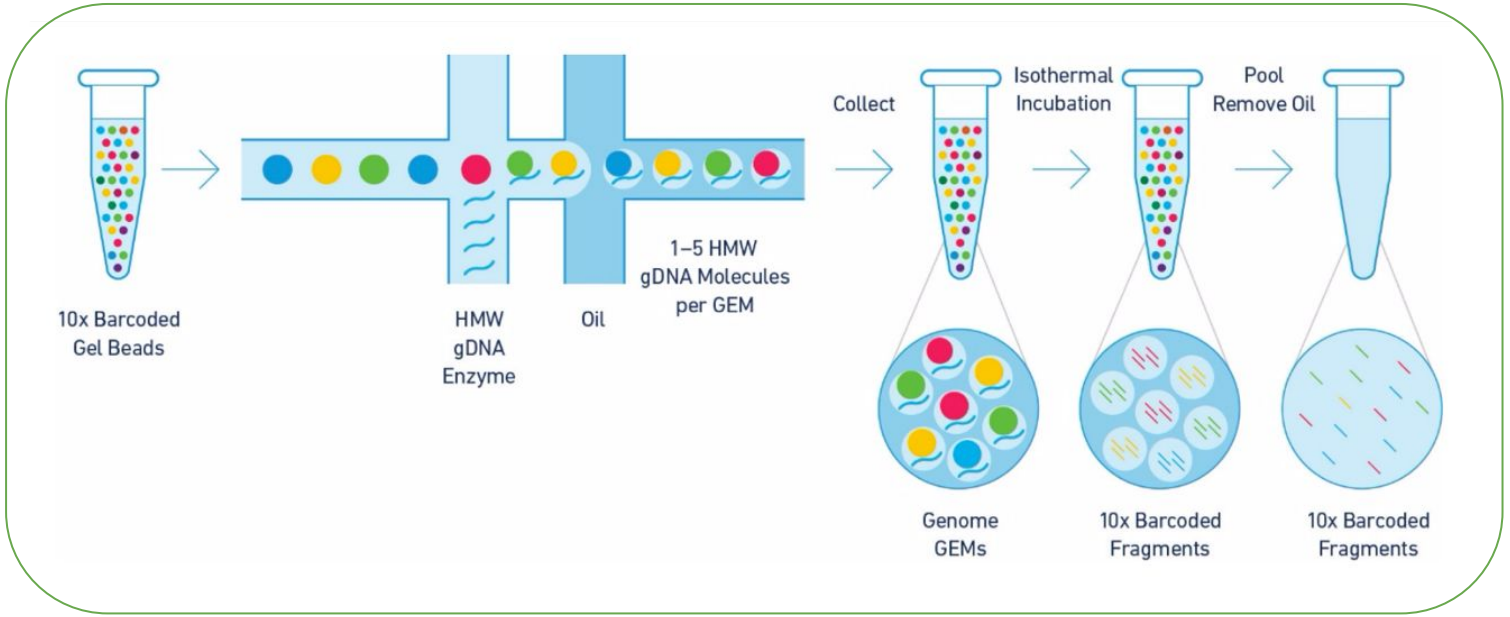
Not a long read technology

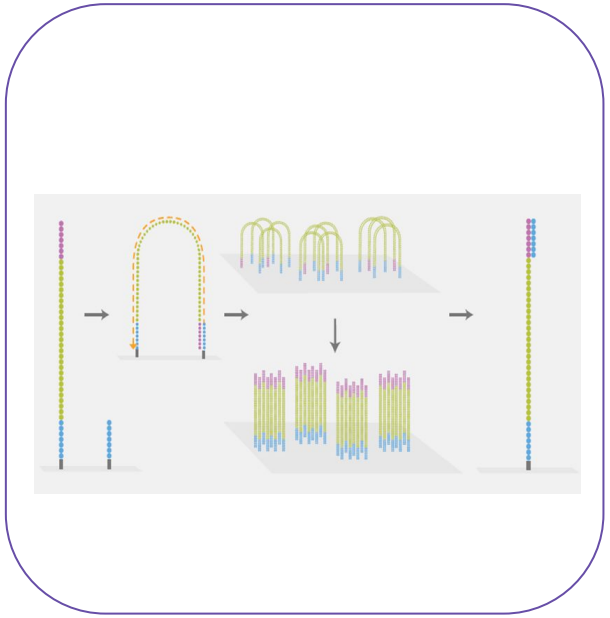
But provides long-range information through **linked reads**

Short reads originating from the same long molecule



Based on standard short read Illumina technology


10x Barcoded
Gel Beads
HMW
gDNA
Enzyme
Oil
1–5 HMW
gDNA Molecules
per GEM
Collect
Isothermal
Incubation
Pool
Remove Oil
Genome
GEMs
10x Barcoded
Fragments
10x Barcoded
Fragments




R1
Illumina
adaptor
10X
barcode
gDNA

Linked Reads

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

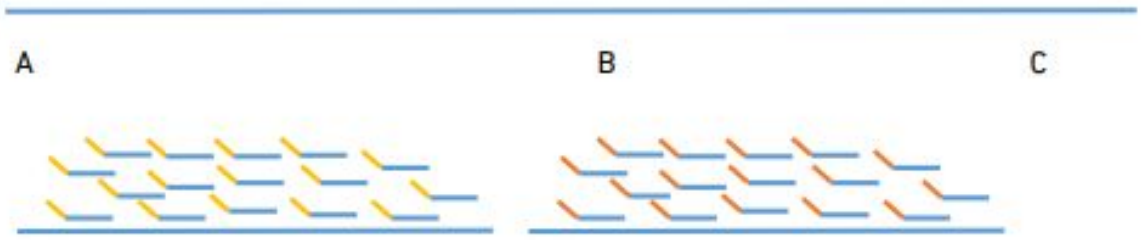
gDNA fragment size is usually 50-60kb

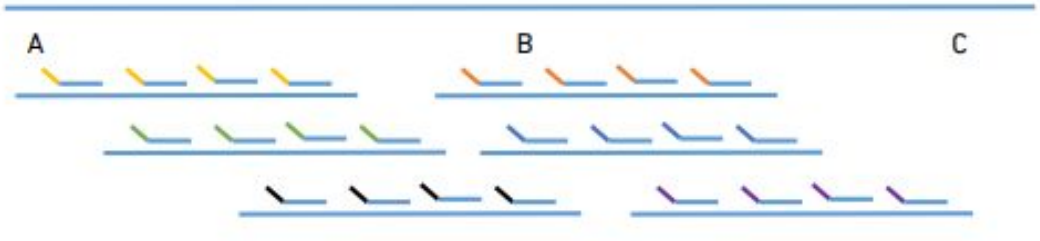
If ~x3 depth is used - we can produce “synthetic long reads”

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

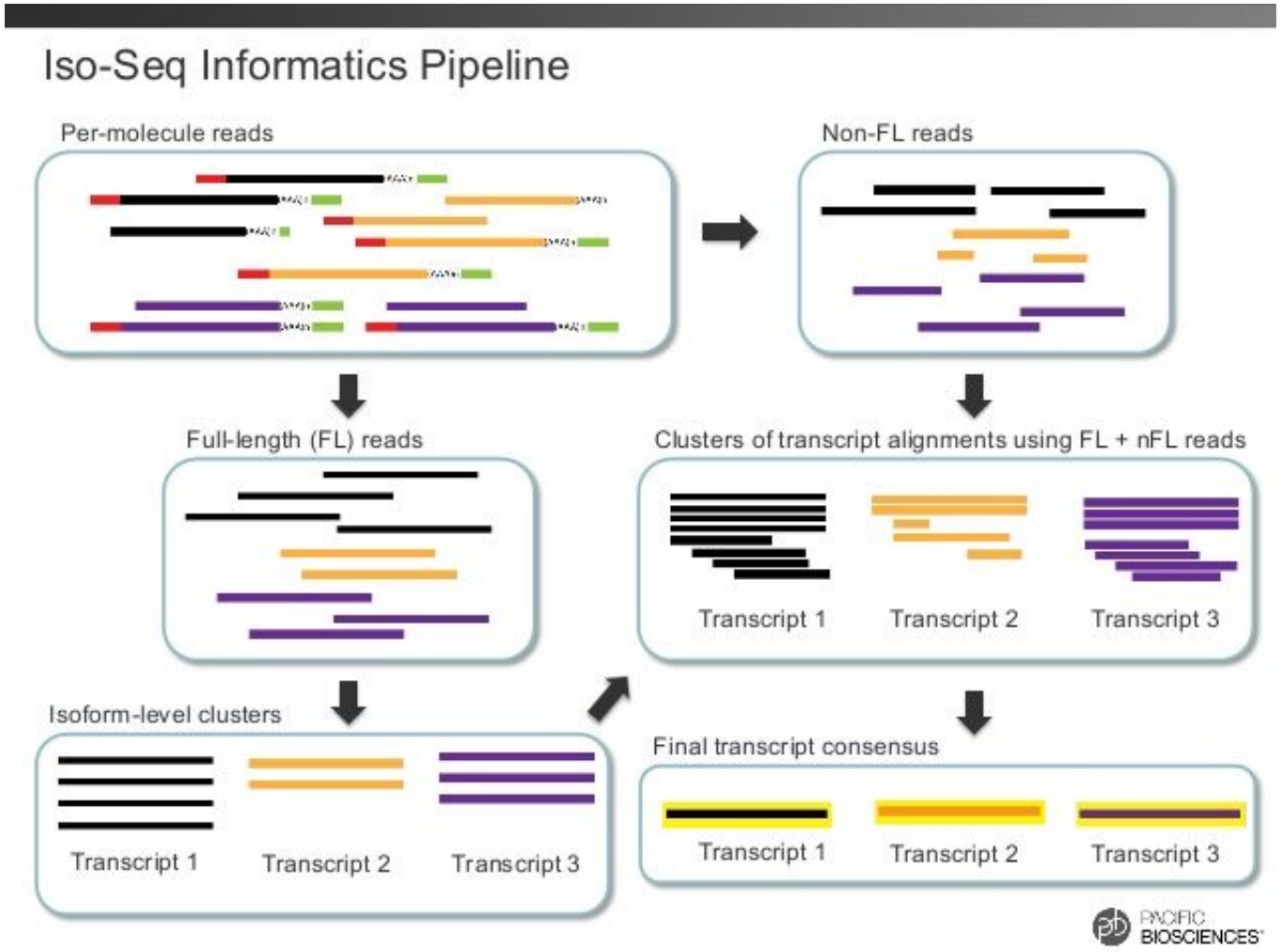
RNA-Seq and Long Reads

Read length is usually larger than mRNA size

Full-length transcripts

No transcript assembly is needed

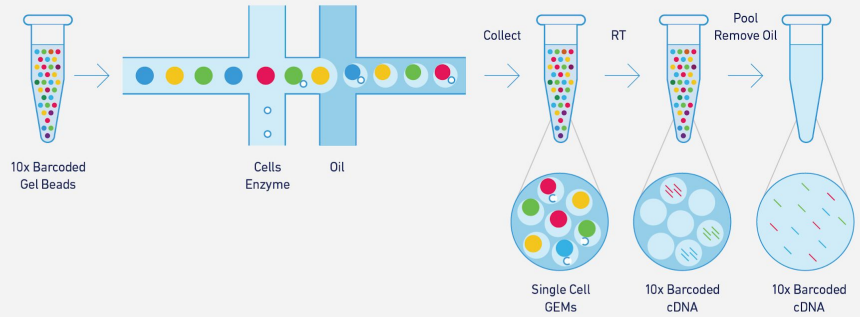
Easier to detect and quantify isoforms



10X for Single Cell RNA-Seq

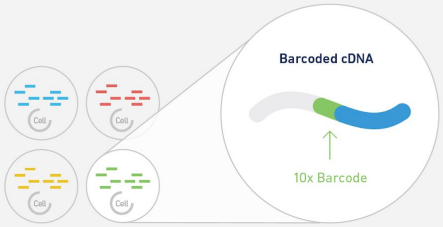
GemCode™ Technology for Single Cell Partitioning

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.

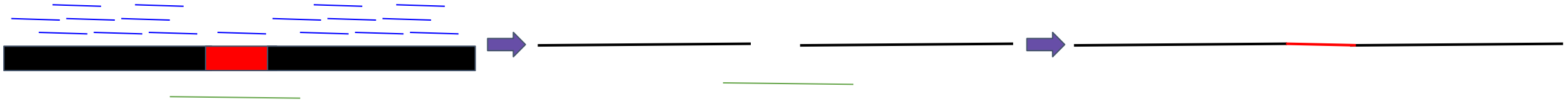
Long and Linked Reads in Genome Assembly

Many modern assemblers can work with 3rd generation reads

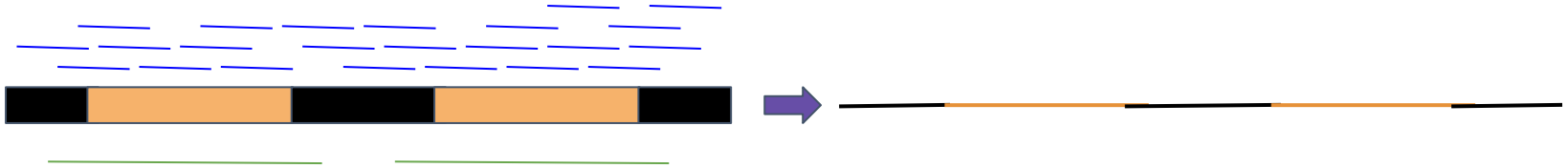
- Falcon - PacBio reads
- Canu, SPAdes - PacBio and ONT reads
- Supernova - 10X reads

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

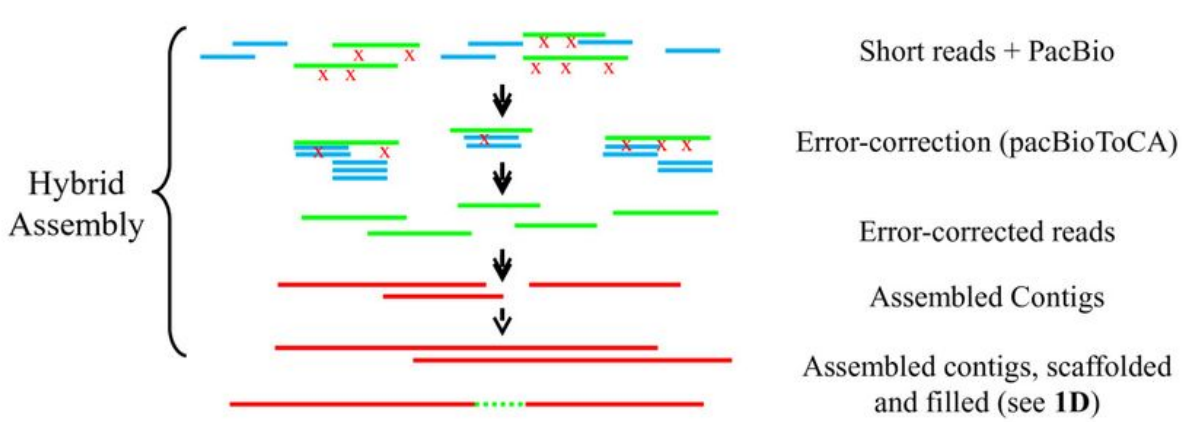
Long/linked reads can help link contigs by bridging over difficult regions

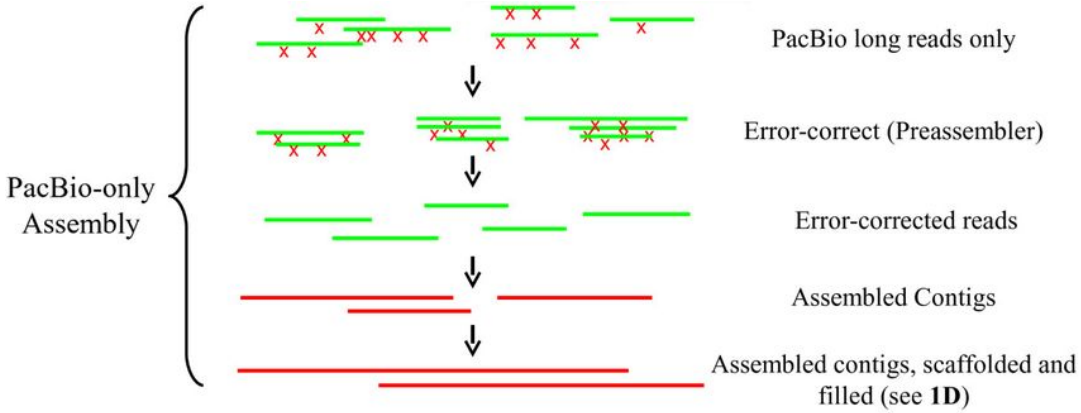


Long reads can help solve long repeats



Different Assembly Strategies





Haplotype Phasing

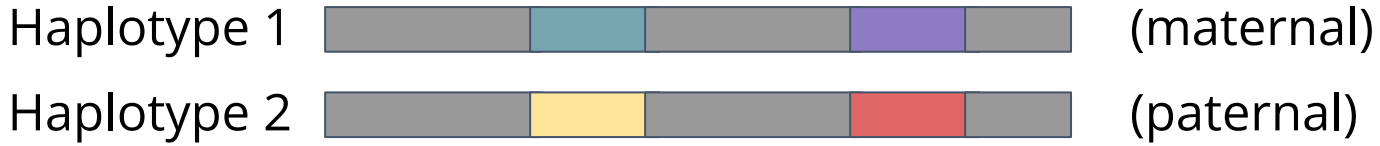
Many interesting eukaryote genomes are diploid or polyploid

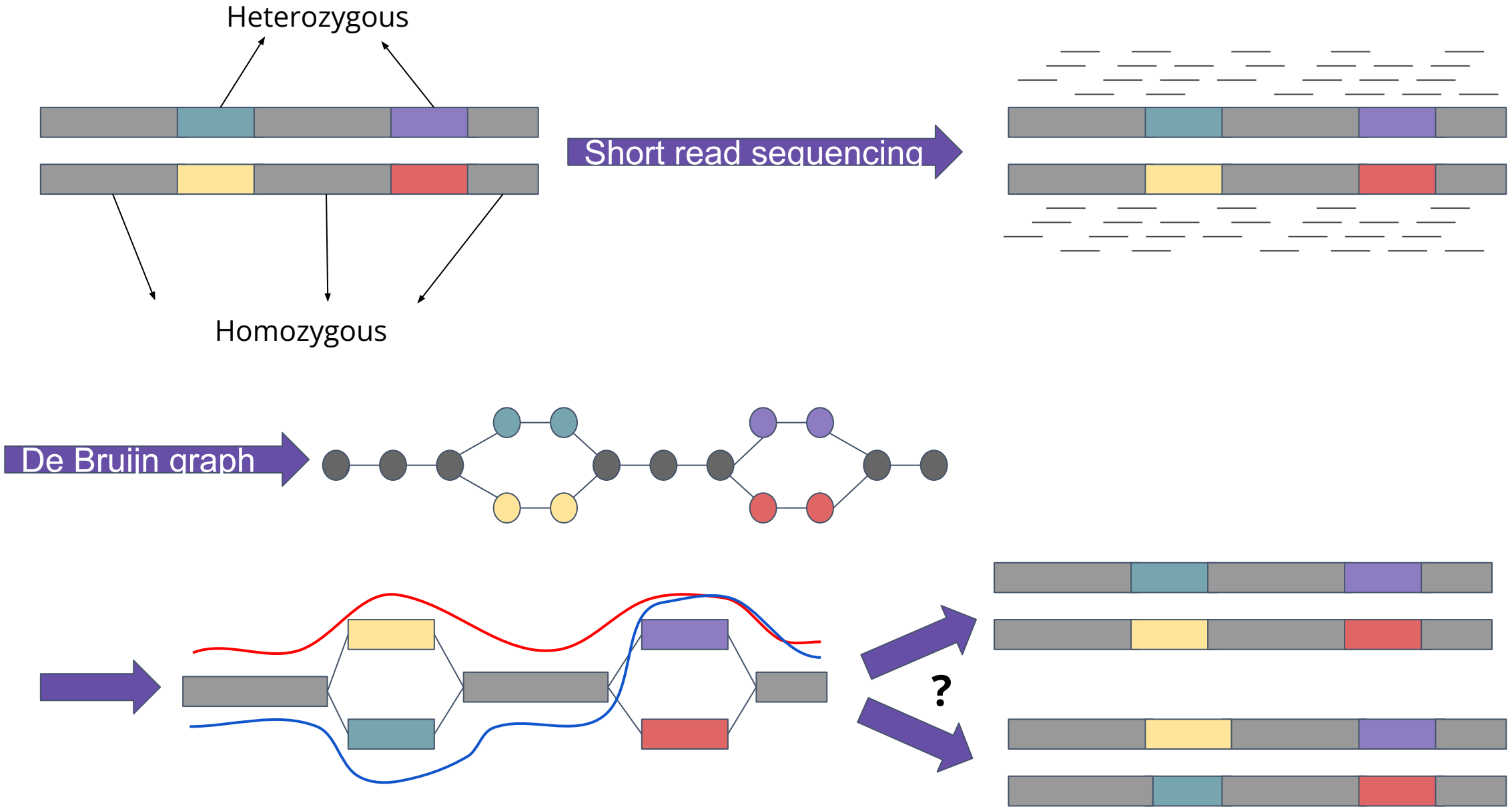
Still, most assemblies are haploid

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

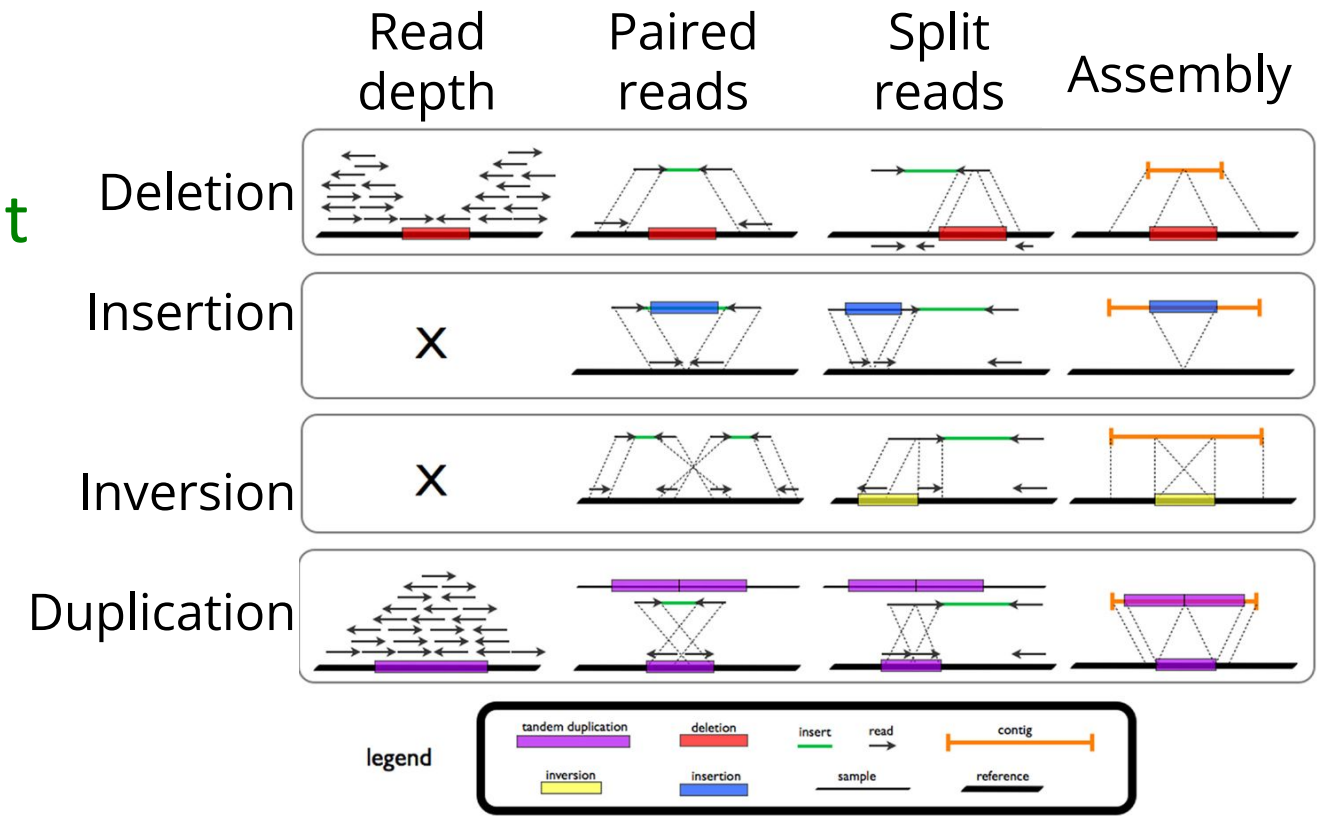
SVs are generally hard to detect with short reads

Many SVs are located in regions that are hard to sequence

SV detection is usually based on mapping reads to a reference

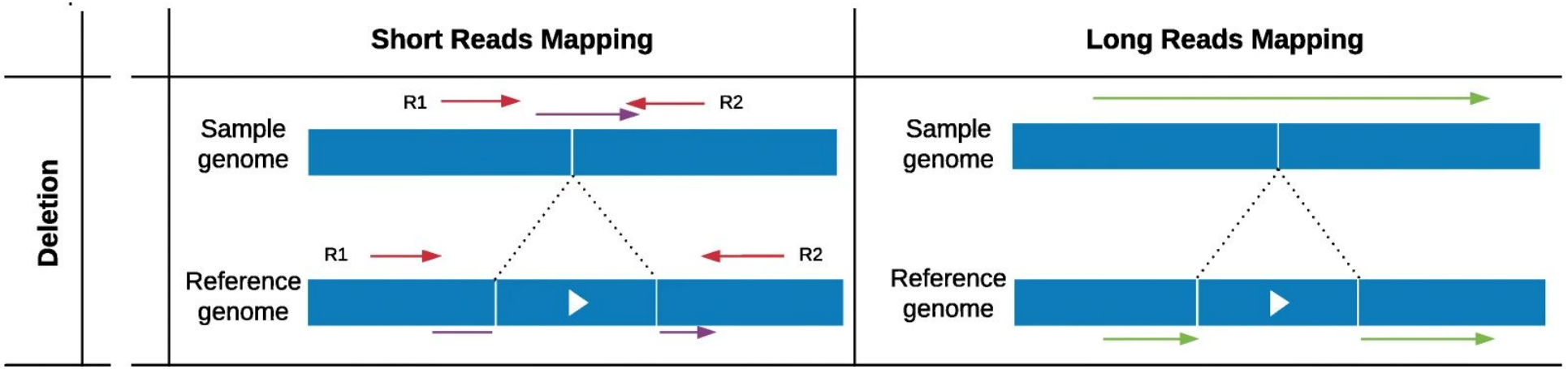
Long reads are useful because:

- 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

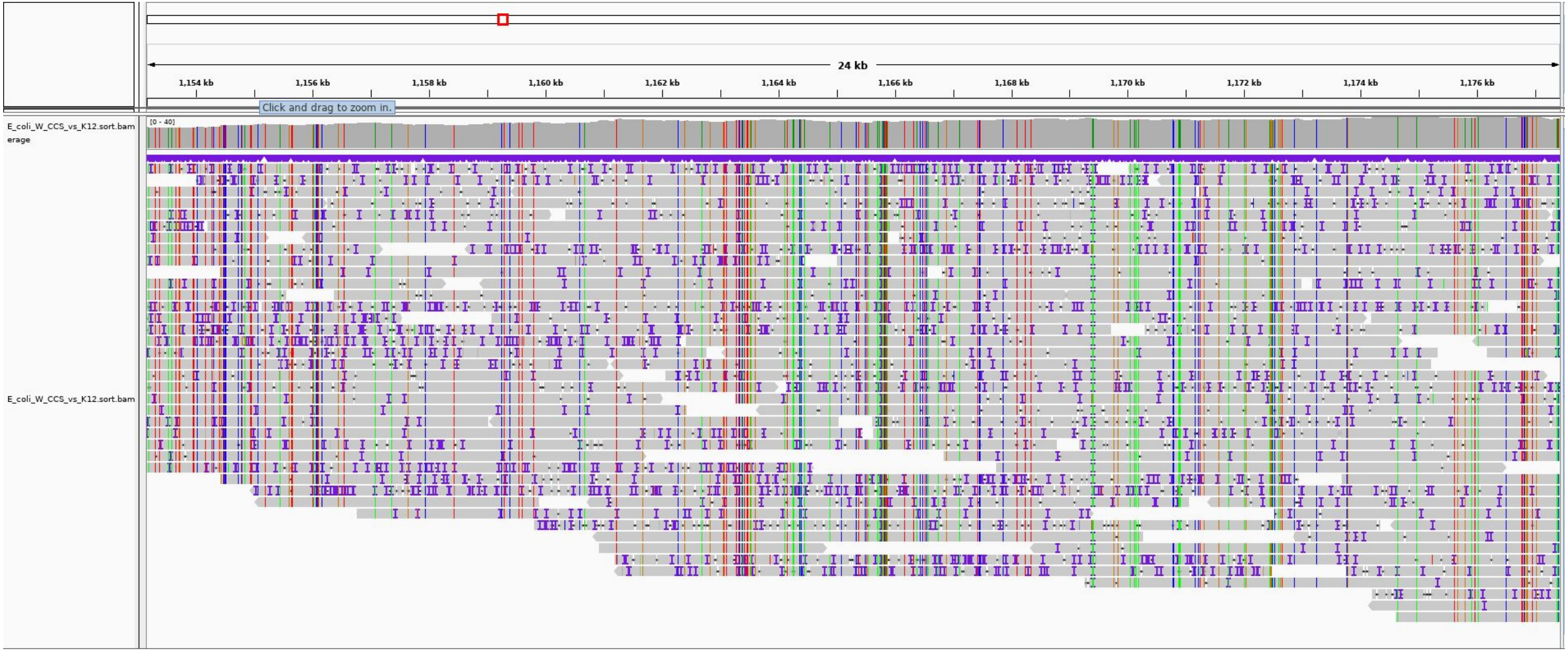
There are various mapping modes like:

- 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

Output format is SAM or PAF


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

SV Calling With Sniffles

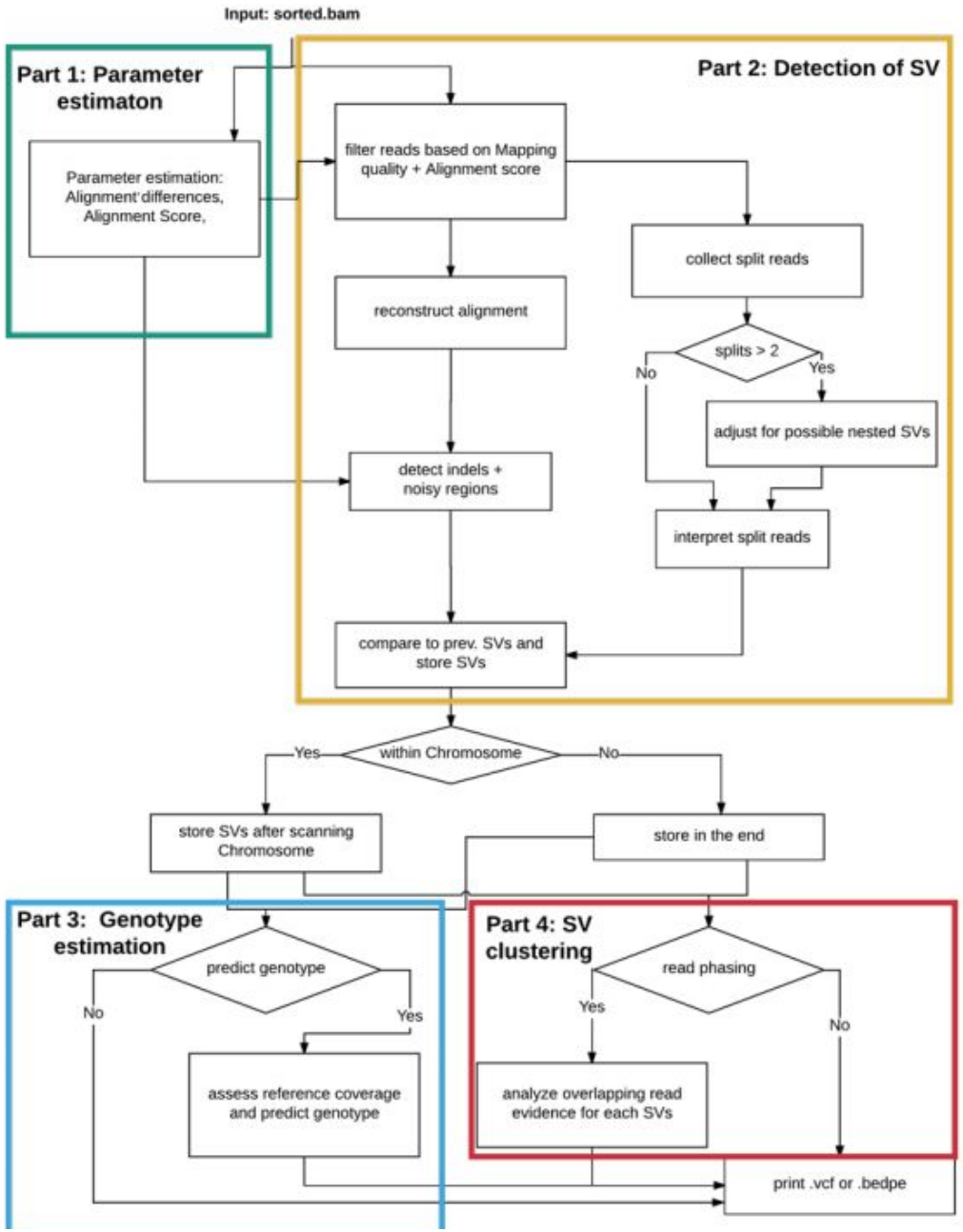
Sniffles applies various strategies to call SVs from long read data

It can detect insertions, deletions, duplications, inversions and translocations.

Input: aligned long reads in BAM format

Output: SV calls in VCF format

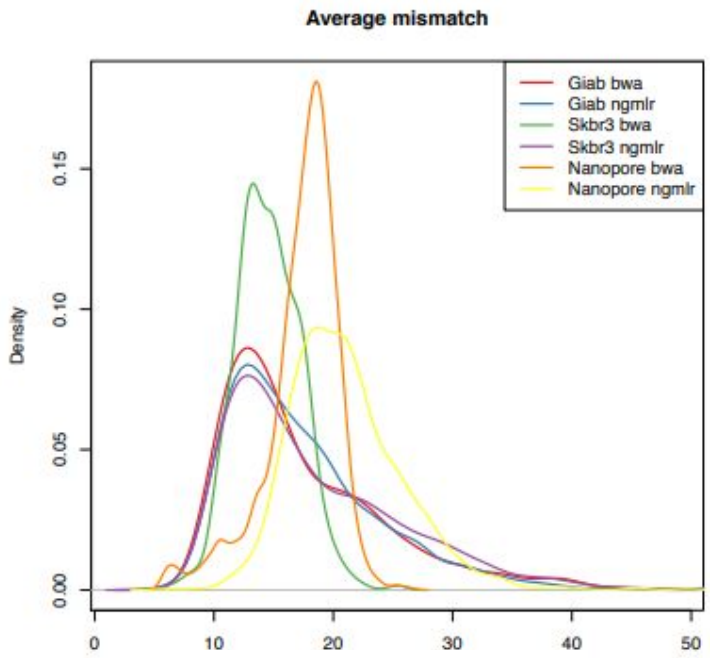
Sniffles workflow



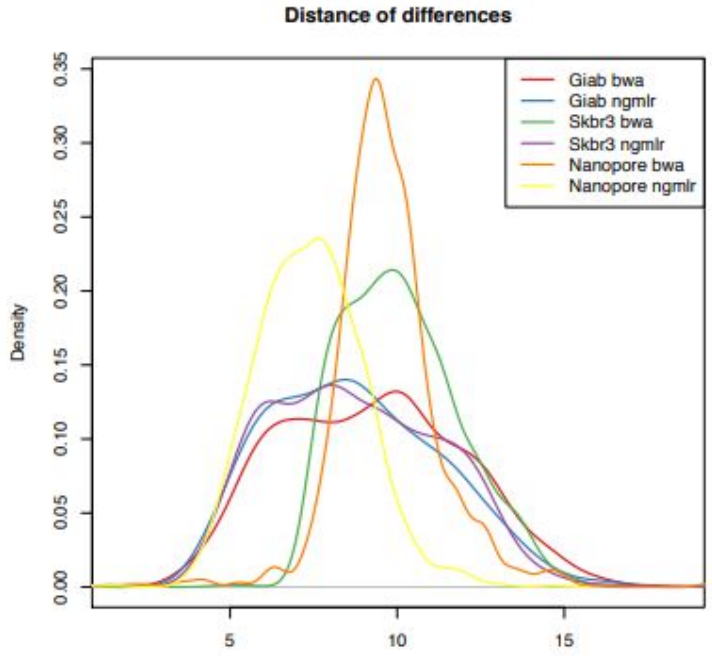
Sedlazeck, Fritz J., et al. "Accurate detection of complex structural variations using single-molecule sequencing." *Nature methods* 15.6 (2018): 461-468.

SV Calling With Sniffles

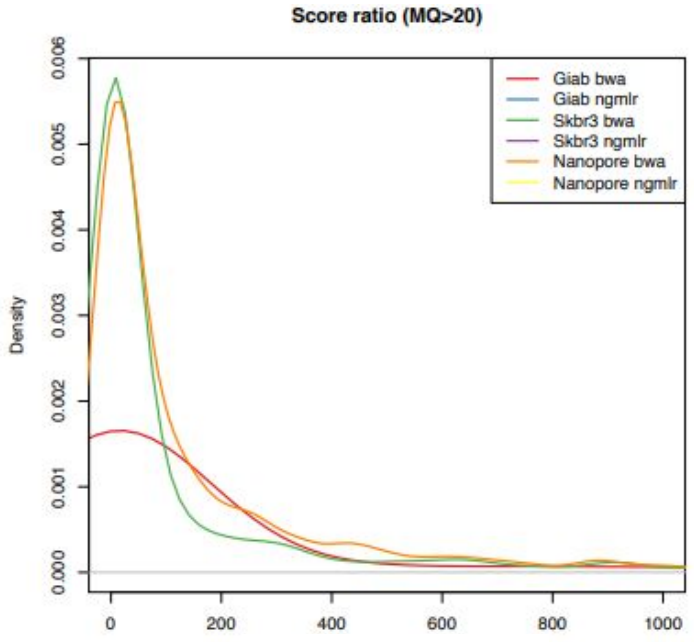
Step 1 - estimate data parameters (from sub-sample)



Differences / 100 bp



Distance (bp)



MAPQ ratio

Sedlazeck, Fritz J., et al. "Accurate detection of complex structural variations using single-molecule sequencing." *Nature methods* 15.6 (2018): 461-468.

SV Calling With Sniffles

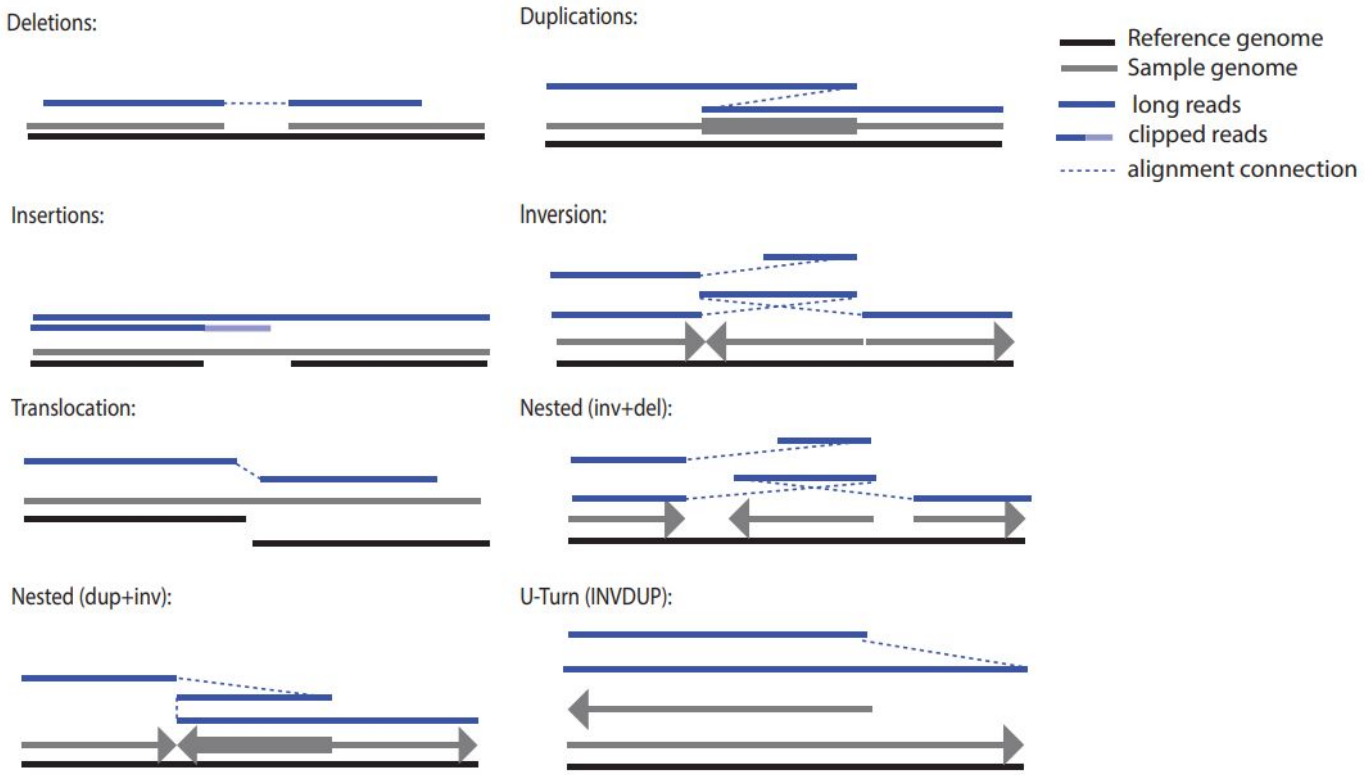
Step 2 - filter low quality mappings

- MAPQ < 20
- MAPQ ratio < 2 or smaller than minimum observed in step 1
- Too many split events

SV Calling With Sniffles

Step 3 - 4

- Scan for regions deviating from the observed parameters and detect SVs
- Merge SVs from multiple reads



Sniffles Output VCF

```
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  E_coli_W_CCS_vs_K12.sort.bam
Chromosome      0       0       N       <DUP>   .       PASS    PRECISE;SVMETHOD=Snifflesv1.0.11;CHR2=Chromosome;END=4641652;ZMW
=45;STD_quant_start=0.000000;STD_quant_stop=0.213201;Kurtosis_quant_start=7.874286;Kurtosis_quant_stop=-2.083333;SVTYPE=DUP;SUPT
YPE=SR;SVLEN=4641652;STRANDS=-+;RE=45;REF_strand=0,0;AF=1       GT:DR:DV        1/1:0:45
Chromosome      5596    1       N       GCCTTGTATTATCCCTCCAGTGCAGAGAAAATCGGCCAGTTTTCTCTGCCTGCAGTCCGCATGCCGTATCGGGCCTTGGGTTCTAACC
TGTTGCGTAGATTTATGCAGCGGACTGCCTTTCTCCCAAAGTGATAAACCGGACAGTATCATGGACCGGTTTTCCCGGTAATCCGTATTTGCAAGGTTGGTTTCACTATGGAACATGAACTTCATTAT
ATCGGTATCGACACCGCTAAAGAGAAACTGGATGTCGATGTGTTGCGTCCTGATGGTCGTCATCGCACCGAAAAAATTCGCTAACACCACTAAAGGGCACGATGAGCTGGTGAGCTGGCTGAAAGGTC
ACAAGATTGACCATGCGCATATCTGCATCGAAGCGACCGGCACCTATATGGAACCTGTCGCTGAGTGCCTTTACGATGCTGGCTACATAGTGTCAGTCATTAATCCTGCGCTGGGTAAAGCTTTCGCT
CAGAGTGAAGGACTGCGTAACAAGACTGATACCGTGGATGCGCGCATGCTGGCAGAGTTCTGTCGTCAGAAGCGCCCTGCAGCCTGGGAAGCGCCTCACCCGCTTGAACGCGCGTTGCGTGCCTGGTA
GTCCGCCACCAGGCGCTGACAGATATGCACACGCAGGAACTGAATCGCACTGAAACGGCGCGGGAAGTCCAGAGACCGAGCATTGATGCTCACCTTCTGTGGCTTGAAGCAGAGCTGAAGCGTCTTGG
AGAAGCAGATAAAAGACCTGACAGACGATGATCCGGATATGAAACACCGCAGGAAACTGCTGGAAAGCATCCCGGGTATCGGAGAGAAAACATCTGCGGTATTGCTGGCTTATATCGGTCTGAAGGAC
CGCTTCGCCCATGCCAGACAGTTCGCCGCTTTTGGCGGTCTGACACCACGGCGTTATGAATCAGGTAGCAGTGTGAGAGGGGCGAGCCGGATGAGTAAGGCCGGACATGTGTCGCTTCGCAGGGCGTT
GTATATGCCCGCAATGGTAGCCACCAGTAAGACTGAGTGGGGGACGGGCGTTCCGCGACCGGTCTGGCGGCTAATGGCAAGAAAGGAAAGGTGATTCTCGGCGCGATGATGCGCAAGCTGGCACAGGT
GGCGTATGAGTGCTGAAGTCAGGCGTGCCGTTCGATGCGTCACGGCATAATCCGTAGCGGCGTAAAAAATCGCGGAAGGGATGAAAAAAACAGCGCCTGACGGCGCTGTGTCCTGGCATGCTGCAATC
CGGGAAACCGACCAGGAAAAAACTTGCAGGCCATAACAGTA       .       PASS    PRECISE;SVMETHOD=Snifflesv1.0.11;CHR2=Chromosome;END=6876;ZMW=40
;STD_quant_start=0.000000;STD_quant_stop=0.632456;Kurtosis_quant_start=2.078125;Kurtosis_quant_stop=-1.459504;SVTYPE=INS;SUPTYPE
=AL,SR;SVLEN=1280;STRANDS=+-;RE=40;REF_strand=9,10;AF=0.677966  GT:DR:DV        0/1:19:40
Chromosome      15378   2       GGGATCACTCCCATAAGCGCTAACTTAAGGGTTGTGGTATTACGCCTGATATGATTTAACGTGCCGATGAATTACTCTCACGATAACTGGTCAGCA
ATTCTGGCCCATATTGGTAAGCCCGAAGAACTGGATACTTCGGCACGTAATGCCGGGGCTCTAACCCGCCGCCGCGAAATTCGTGATGCTGCAACTCTGCTACGTCTGGGGCTGGCTTACGGCCCCGG
GGGGATGTCATTACGTGAAGTCACTGCATGGGCTCAGCTCCATGACGTTGCAACATTATCTGACGTGGCTCTCCTGAAGCGGCTGCGGAATGCCGCCGACTGGTTTGGCATACTTGCCGCACAAACAC
TTGCTGTACGCGCCGCAGTTACGGGTTGTACAAGCGGAAAGAGATTGCGTCTTGTCGATGGAACAGCAATCAGTGCGCCCGGGGGCGGCAGCGCTGAATGGCGACTACATATGGGATATGATCCTCAT
ACCTGTCAGTTCACTGATTTTGAGCTAACCGACAGCAGAGACGCTGAACGGCTGGACCGATTTGCGCAAACGGCAGACGAGATACGCATTGCTGACCGGGGATTCGGTTCGCGTCCCGAATGTATCCG
CTCACTTGCTTTTGGAGAAGCTGATTATATCGTCCGGGTTCACTGGCGAGGATTGCGCTGGTTAACTGCAGAAGGAATGCGCTTTGACATGATGGGTTTTCTGCGCGGGCTGGATTGCGGTAAGAACG
GTGAAACCACTGTAATGATAGGCAATTCAGGTAATAAAAAAGCCGGAGCTCCCTTTCCGGCACGTCTCATTGCCGTATCACTTCCTCCCGAAAAAGCATTAATCAGTAAAACCCGACTGCTCAGCGAG
AATCGTCGAAAAGGACGAGTAGTTCAGGCGGAAACGCTGGAAGCAGCGGGCCATGTGCTATTGCTAACATCATTACCGGAAGATGAATATTCAGCAGAGCAAGTGGCTGATTGTTACCGTCTGCGATG
GCAAATTGAACTGGCTTTTAAGCGGCTCAAAAGTTTGCTGCACCTGGATGCTTTGCGTGCAAAGGAACCTGAACTCGCGAAAGCGTGGATATTTGCTAATCTACTCGCCGCATTTTTAATTGACGACA
TAATCCAGCCATCGCTGGATTTCCCCCCCAGAAGTGCCGGATCCGAAAAGAAGAACTAACTCGTTGTGGAGAATAACAAAAATGGTCATCTGGAGCTTACAGGTGGCCATTCGTGGGACAGTATCCCT
GACAGCCTACAAAACGCAATTGAAGAACGCGAGGCATCGTCTTAACGAGGCACCGAGGCGTCGCATTCTTCAGATGGTTCAACCCTTAAGTTAGCGCTTAT   N       .       PASS
PRECISE;SVMETHOD=Snifflesv1.0.11;CHR2=Chromosome;END=16727;ZMW=30;STD_quant_start=0.000000;STD_quant_stop=0.000000;Kurtosis_quan
t_start=1.301541;Kurtosis_quant_stop=3.975309;SVTYPE=DEL;SUPTYPE=AL,SR;SVLEN=-1349;STRANDS=+-;RE=30;REF_strand=3,5;AF=0.789474
GT:DR:DV        0/1:8:30
```

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

Constantly improving

Long reads have their own problems

There are strong commercial interests involved

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!