

Long read sequencing

Next-Generation-Sequencing Analysis

08-01-2023

Long read sequencing

Frederikke Byron Pedersen MscEng

Clinical Scientist

Department of Clinical Immunology

*The University Hospital of Copenhagen,
Rigshospitalet*

Frederikke.Byron.Pedersen@regionh.dk

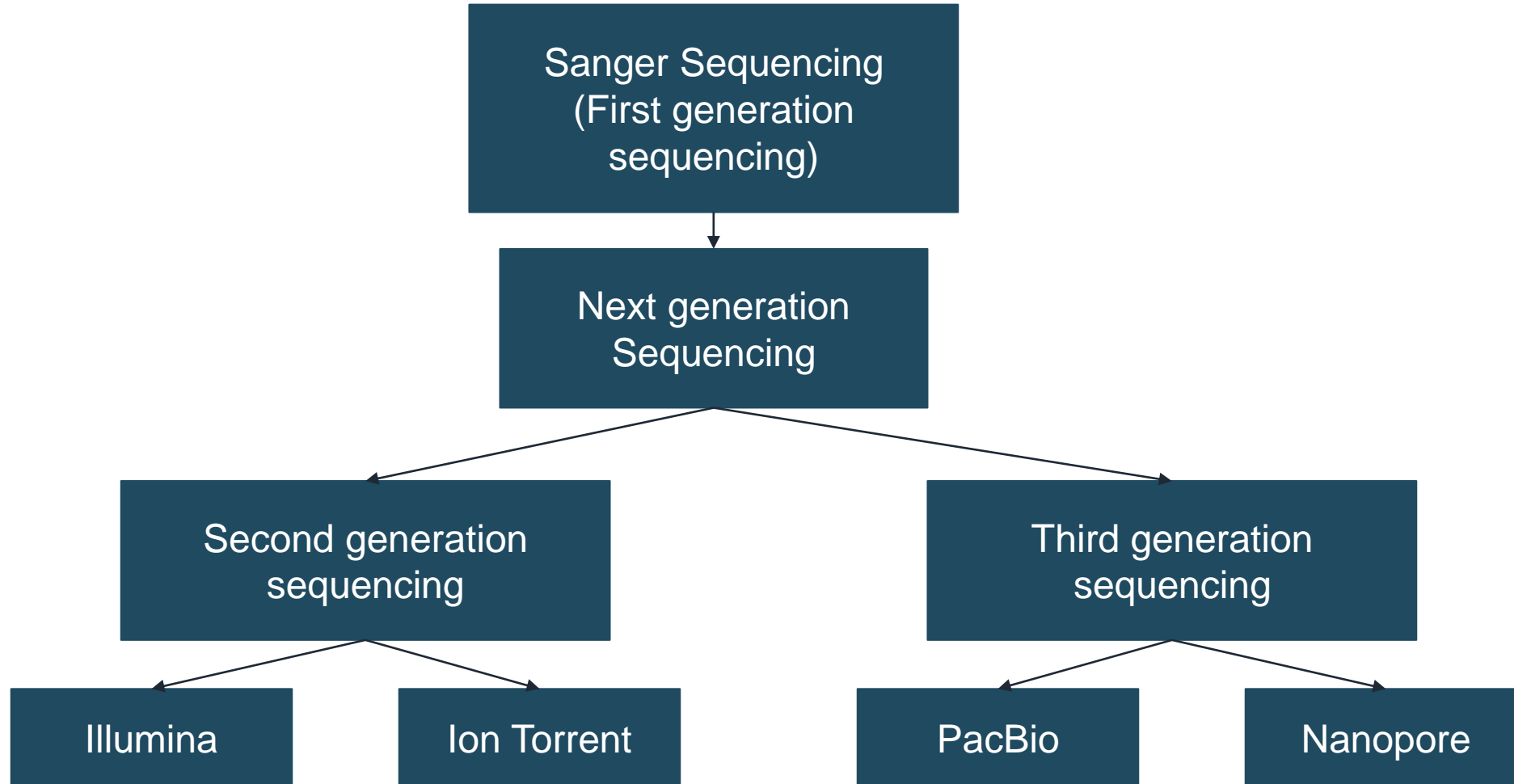
About the presenter

- Part time PhD student (Public Health and Epidemiology, KU), part time Clinical Scientist at Rigshospitalet
- Worked fulltime at the Department of Clinical Immunology at Rigshospitalet from 2021-2023
- Worked at Nanopore Technologies from 2019 to 2021
- Graduated from DTU in 2019, master in Biotechnology
- Worked with longread sequencing since 2018, and short reads since 2021

How to get in touch?

Frederikke.Byron.Pedersen@regionh.dk

Sequencing Techniques: An overview



Quiz

5 multiple choice questions

Note your answers on a piece of paper or laptop

Question 1:

What is the longest read that can be obtained using Nanopore Sequencing? Note, not average read length

1. 25kb
2. 100kb
3. 10kb
4. No limit

Question 2:

What is the longest read that can be obtained using PacBio Sequencing? Note, not average read length

1. 25kb
2. 100kb
3. 10kb
4. No limit

Question 3: What is phasing?

1. To introduce something in stages over a particular period of time
2. Phasing is the rhythmic equivalent of cycling through the phase of two waveforms as in phasing
3. The process of statistical estimation of haplotypes from genotype data
4. All of the above

Question 4:

What is one of the main advantages of Long Read Sequencing?

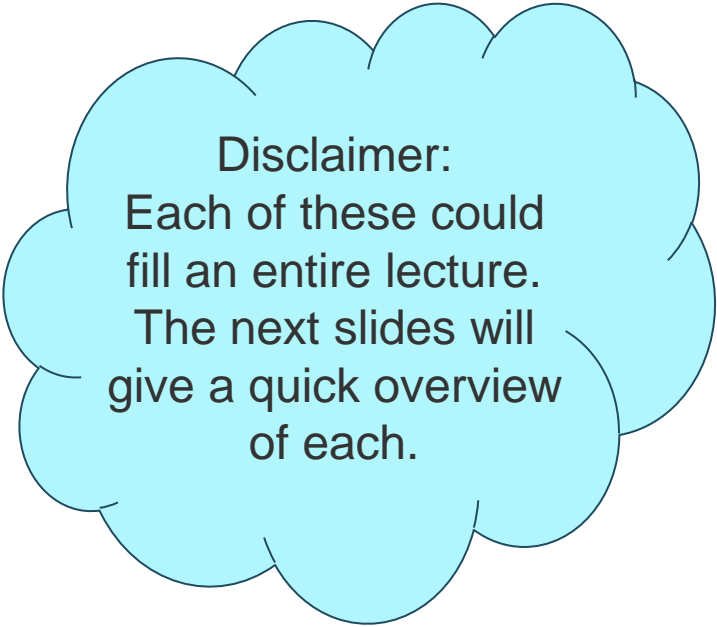
1. Haplotyping and SVs are easier resolved
2. Cheaper and Faster
3. Higher quality/Q-Score
4. All of the above

Question 5: Which sequencing platform is the best?

1. PacBio
2. Nanopore
3. Illumina
4. It depends on what you want to do

Why Long Read Sequencing?

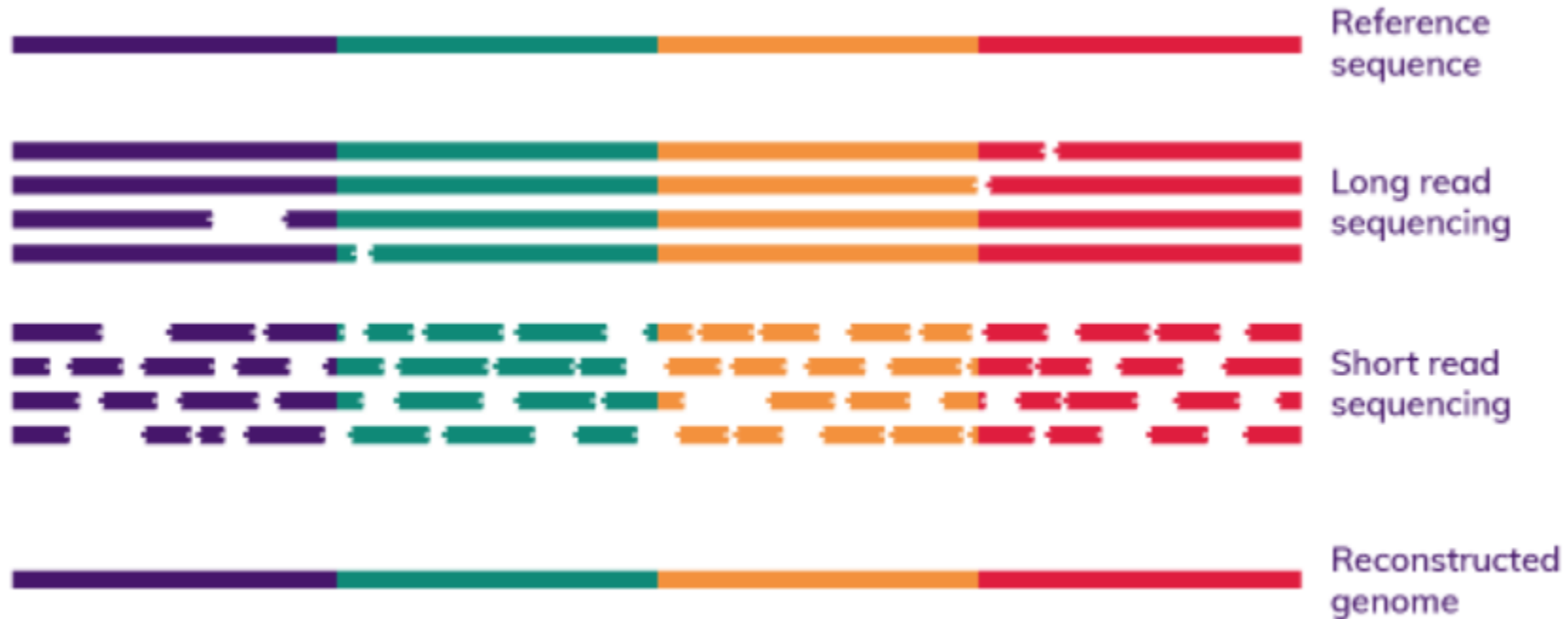
- Assembly
- Phasing/Haplotyping
- Plasmids
- Methylation
- Structural Variants (SVs)
- Large Repetitive Regions
- Plant genomes
- More?



Disclaimer:
Each of these could
fill an entire lecture.
The next slides will
give a quick overview
of each.

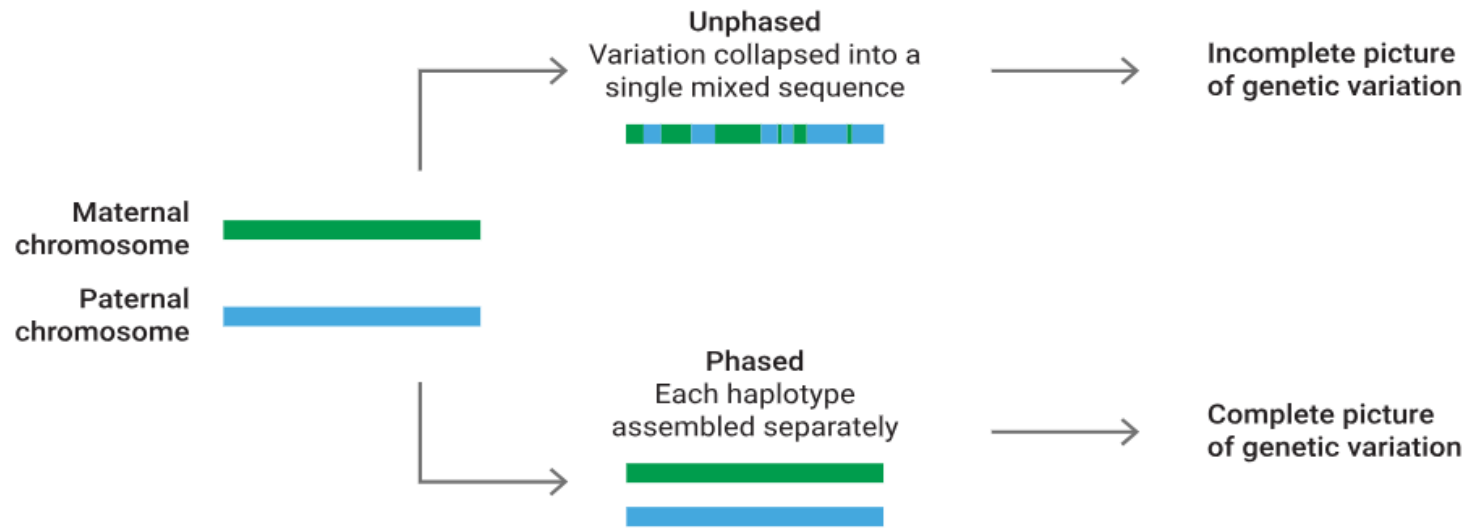
Another thing that is much easier with Long Reads is De novo assembly

Assembly with reference genome



<https://www.phgfoundation.org/briefing/clinical-long-read-sequencing>

Phasing



Diploid (2N)



2 × 3 Gb genome
= 6 Gb of DNA content



Hexaploid (6N)

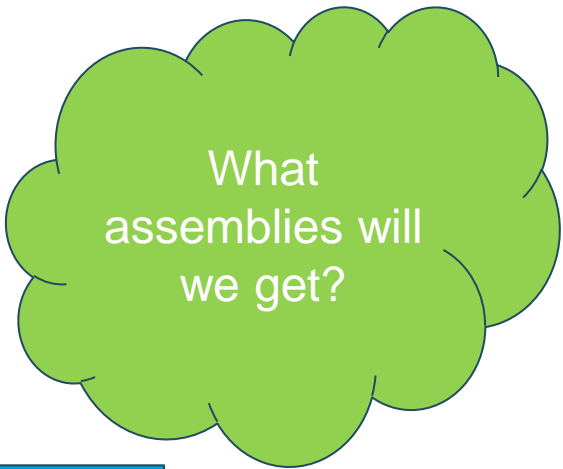


6 × 9 Gb genome
= 54 Gb of DNA content

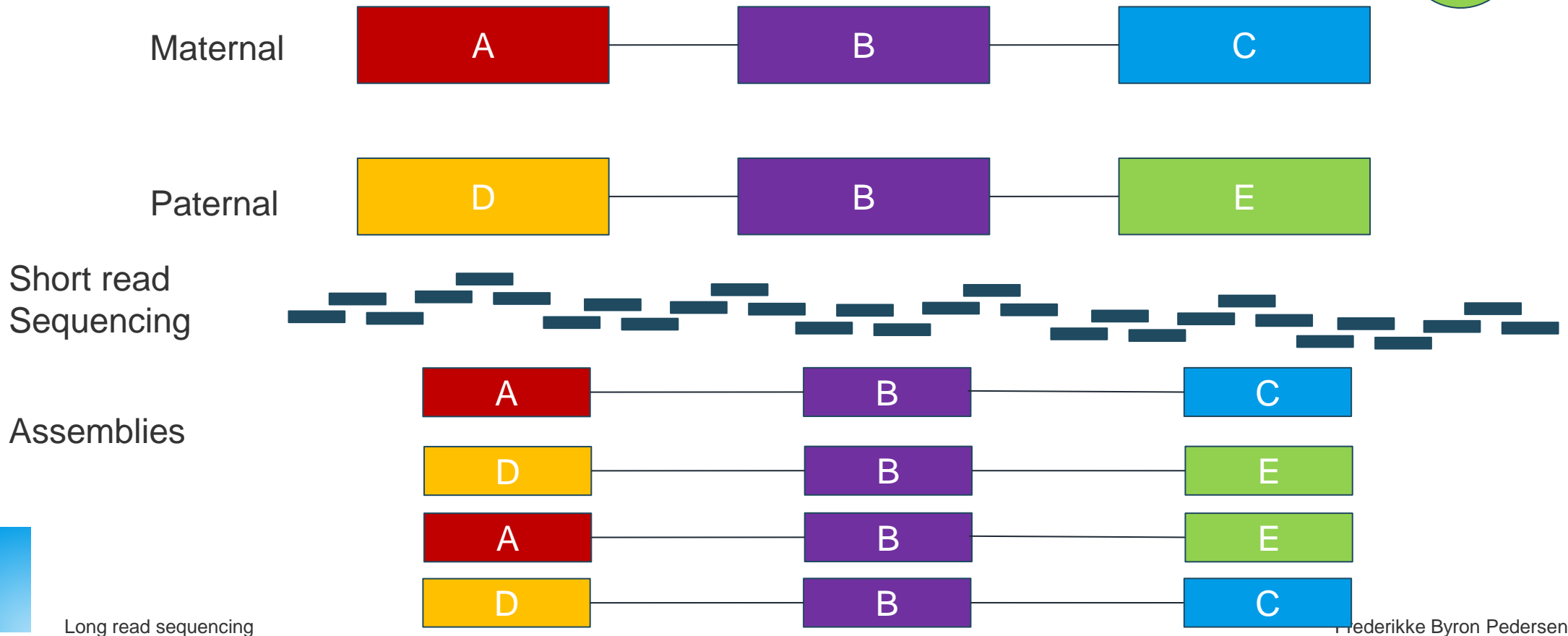
The ploidy – or number of copies of each chromosome in a genome – affects not only the size but also the complexity of the genome.

Phasing involves separating maternally and paternally inherited copies of each chromosome into haplotypes to get a complete picture of genetic variation.

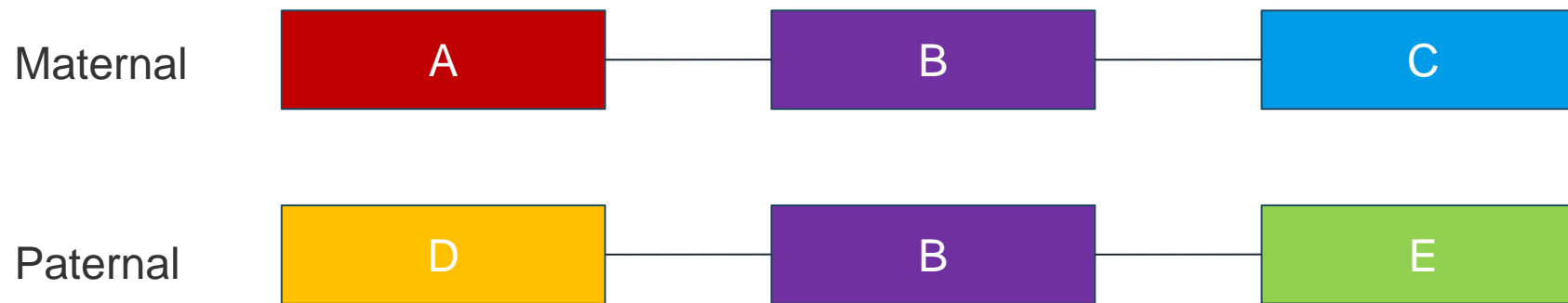
<https://www.pacb.com/blog/ploidy-haplotypes-and-phasing/>



Assembly/Haplotyping, diploid (Human)



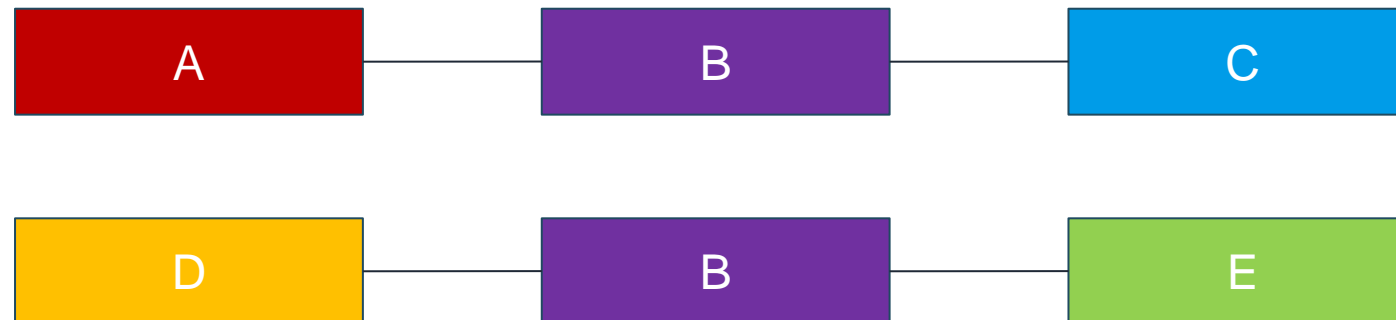
Assembly/Haplotyping, diploid (Human)



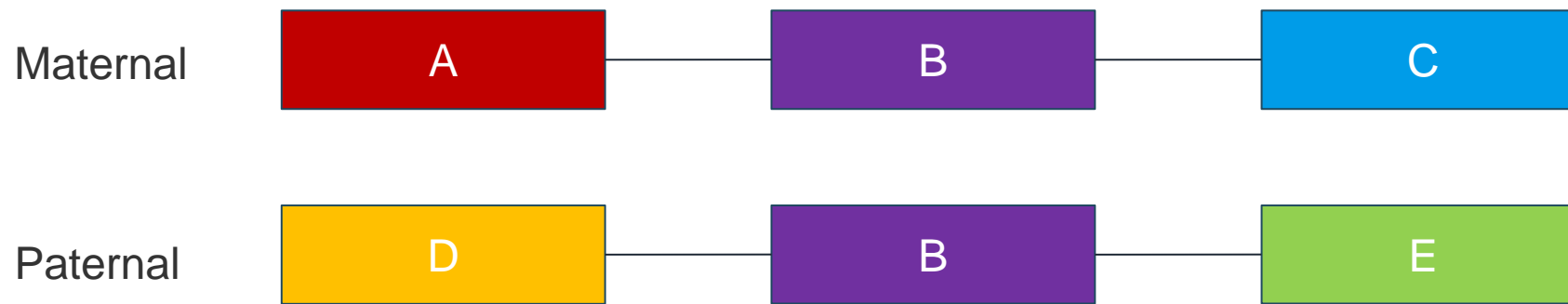
Long read
Sequencing



Assembly



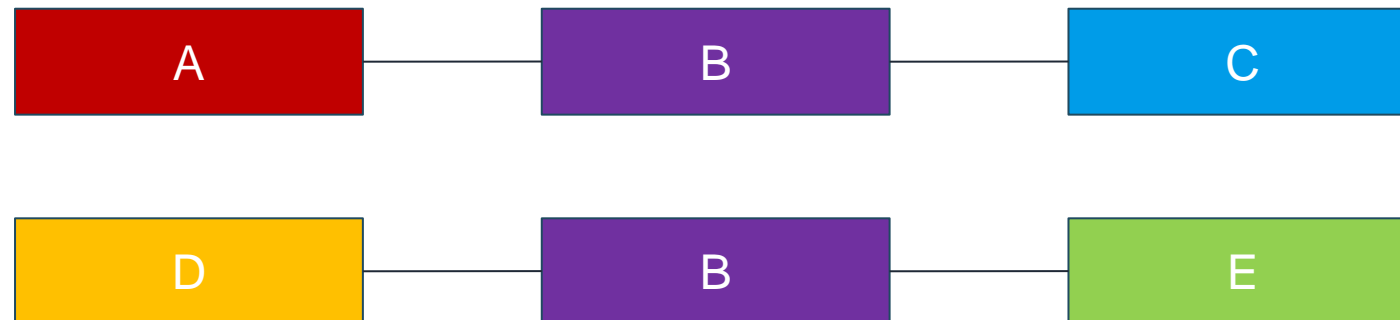
Assembly/Haplotyping, diploid (Human)



Ultra
Long read
Sequencing

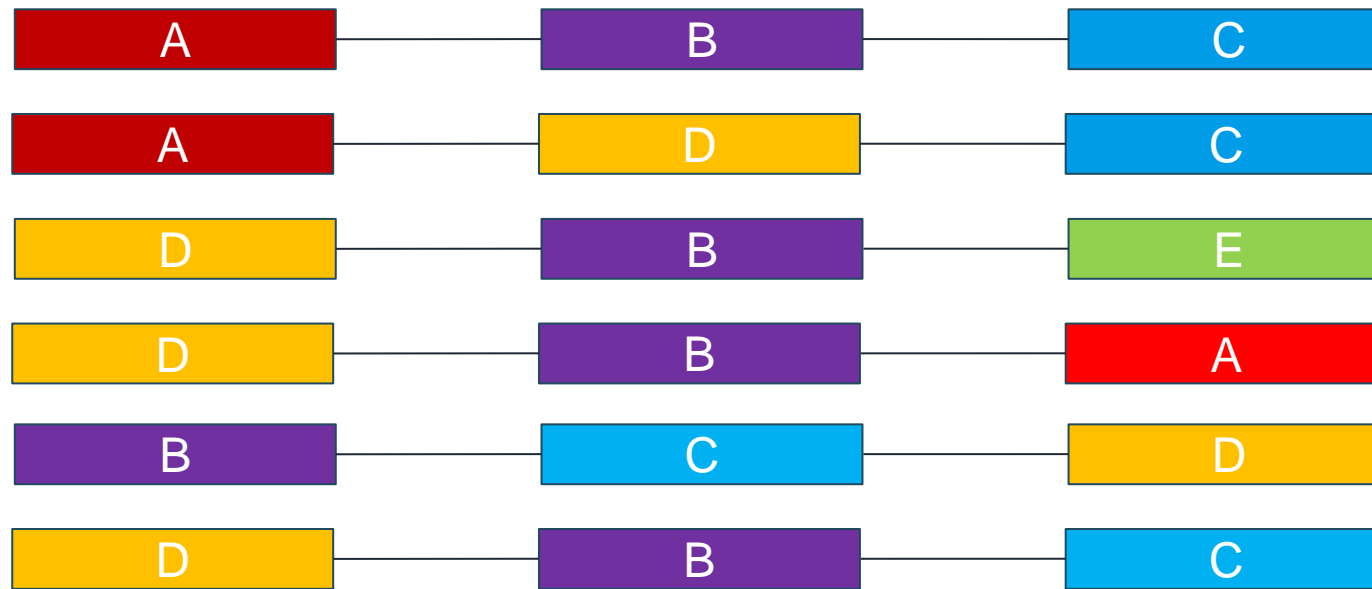


Assembly



Would short reads ever be able to properly phase it?

Assembly/Haplotyping, hexaploid (Plant)



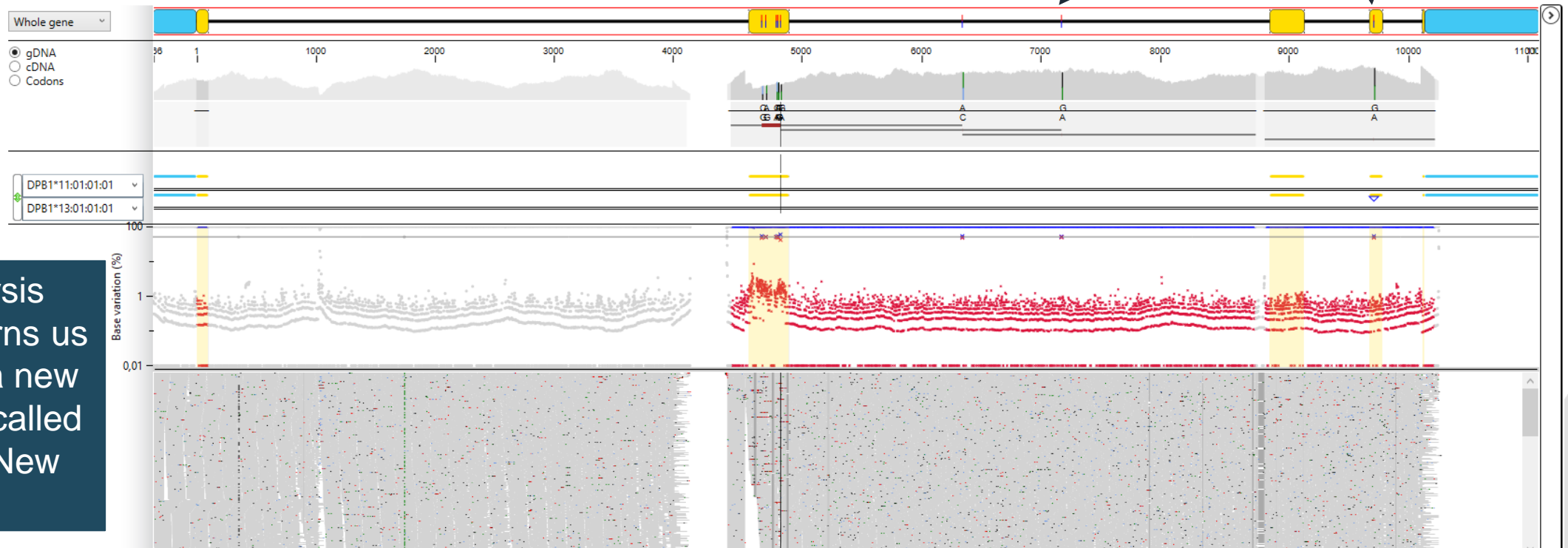
Short read
Sequencing



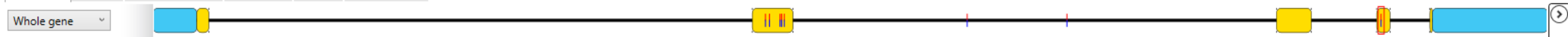
Long read sequencing

Phasing – A real life example, using Illumina sequencing

Over 2kb to nearest heterozygote position



Our analysis program warns us that this is a new DPB1 type called DPB1*13:New



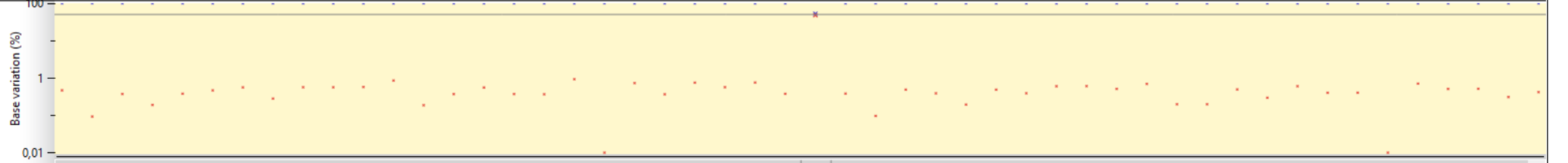
gDNA
cDNA
Codons

9695 9700 9705 9710 9715 9720 9725 9730 9735 9740

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

DPB1*11:01:01:01
DPB1*13:01:01:01

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



Read alignment tracks showing individual sequencing reads with mismatches highlighted.

gDNA Position: 9717 A: 501, C: 1, G: 551, T: 0, -: 0

Rejected reads
 Separate alleles
 Library
 Read variation
 Base variation



Whole gene

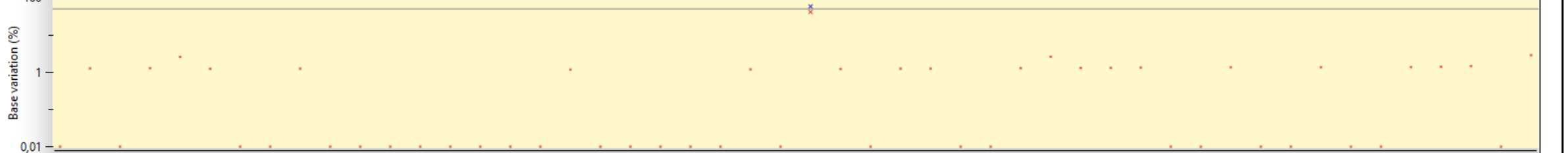
gDNA
cDNA
Codons

9695 9700 9705 9710 9715 9720 9725 9730 9735 9740

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

DPB1*11:01:01:01
DPB1*13:01:01:01

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



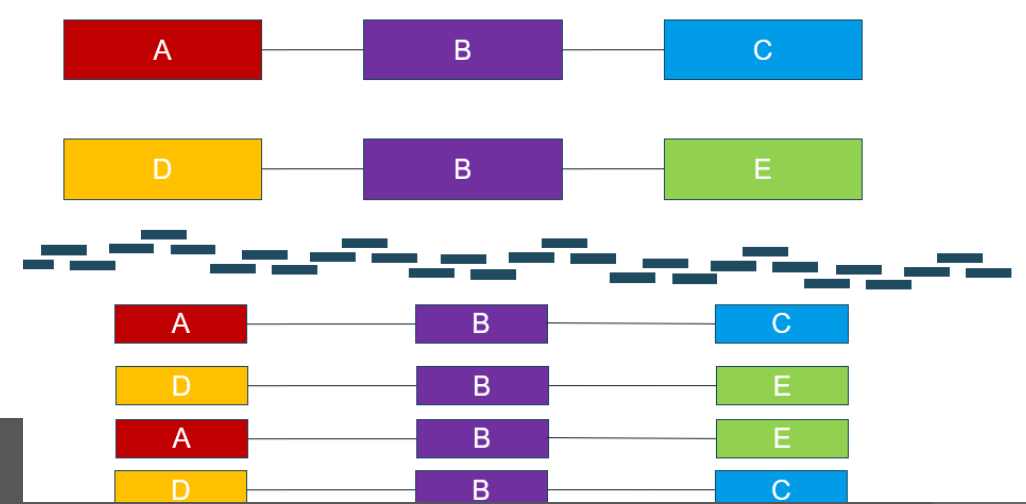
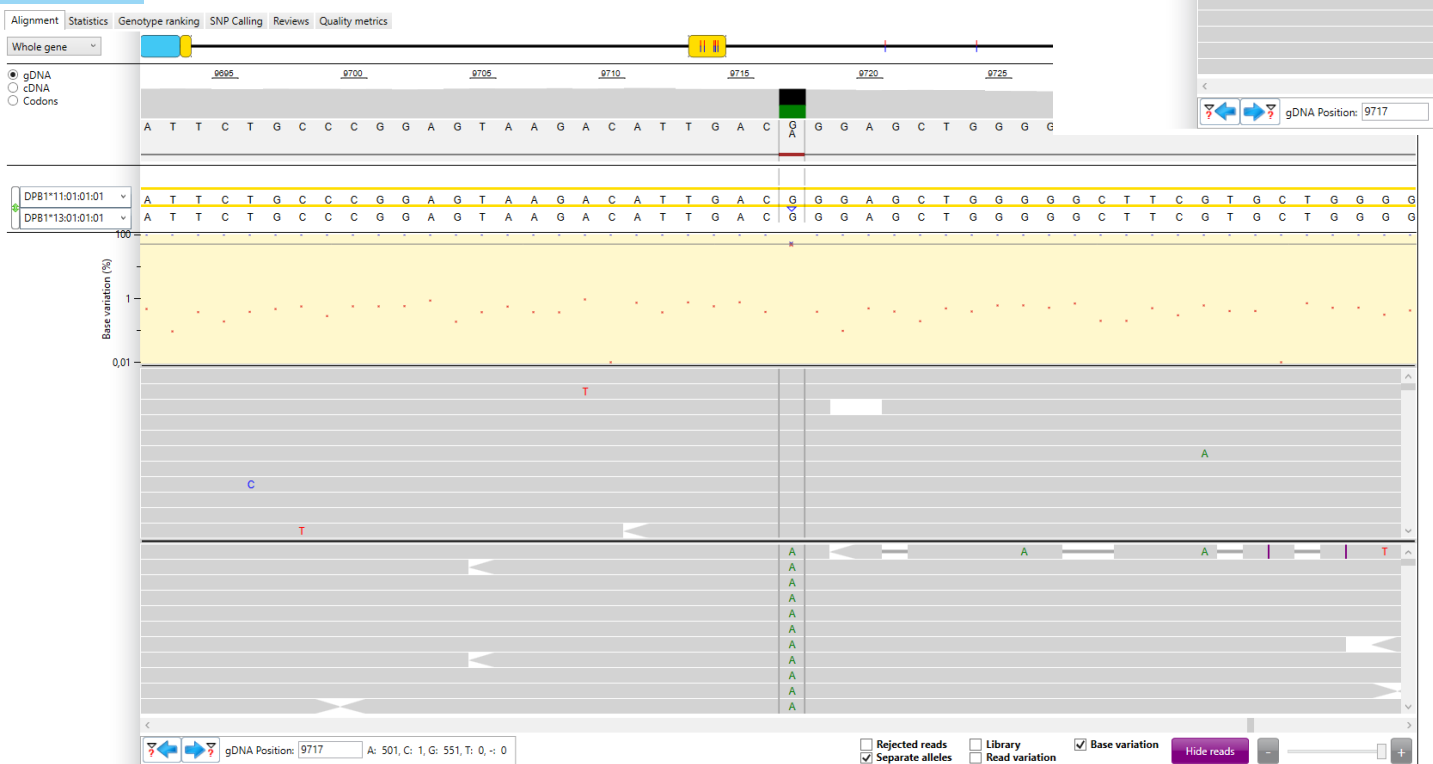
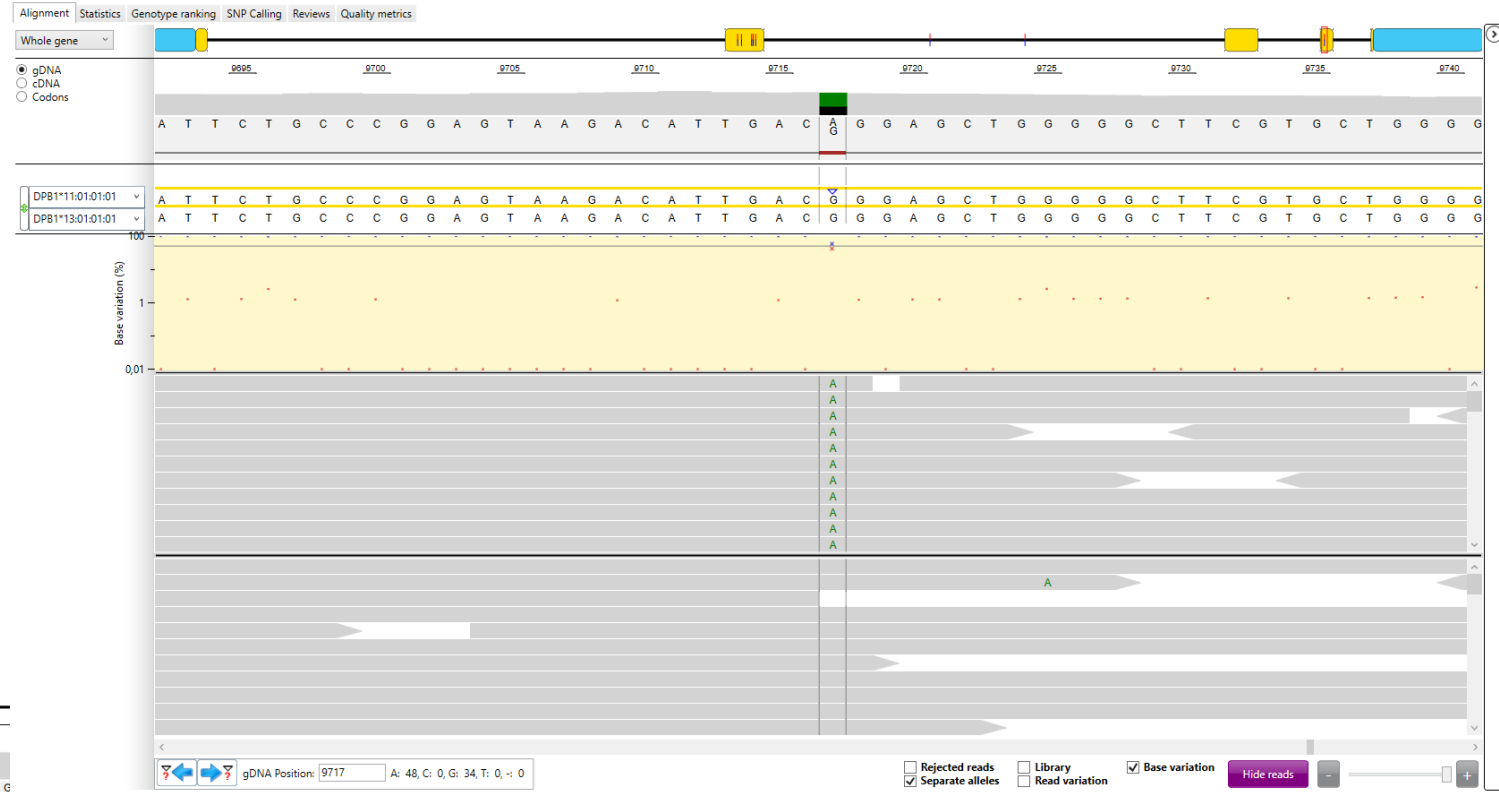
Upon re-running the sample, it now calls it a DPB1*11:New

A
A
A
A
A
A
A
A
A
A

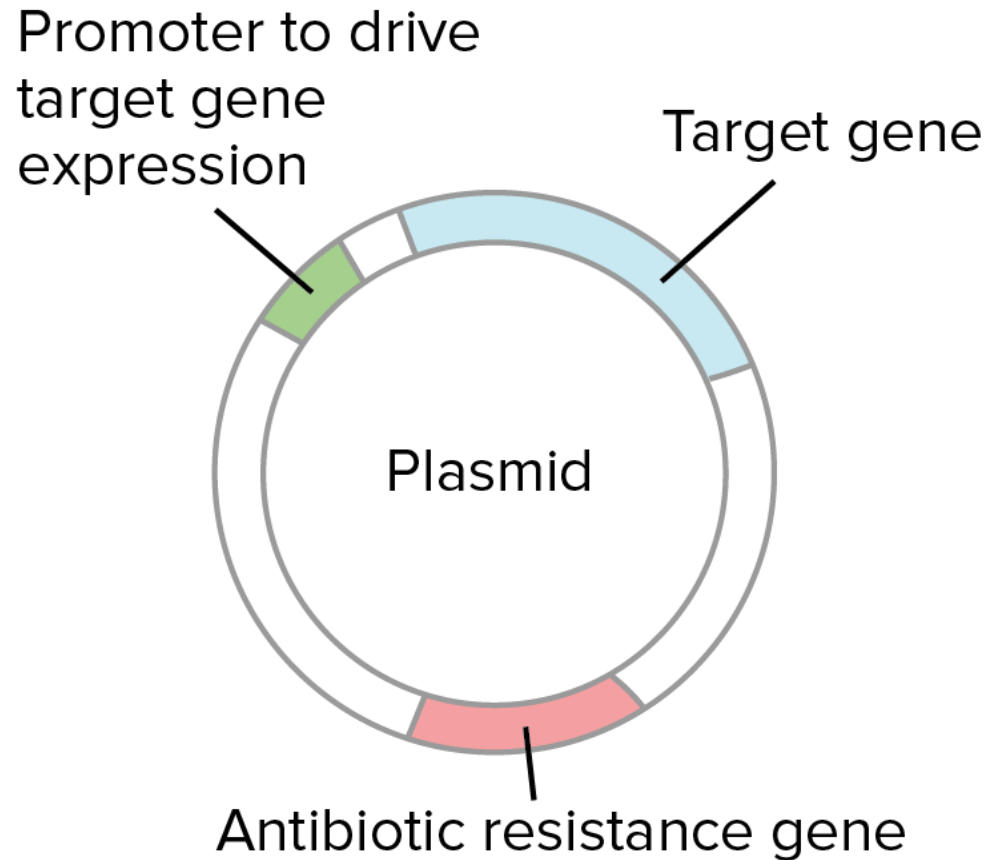
A

gDNA Position: 9717 A: 48, C: 0, G: 34, T: 0, -: 0

What happened?



Plasmids

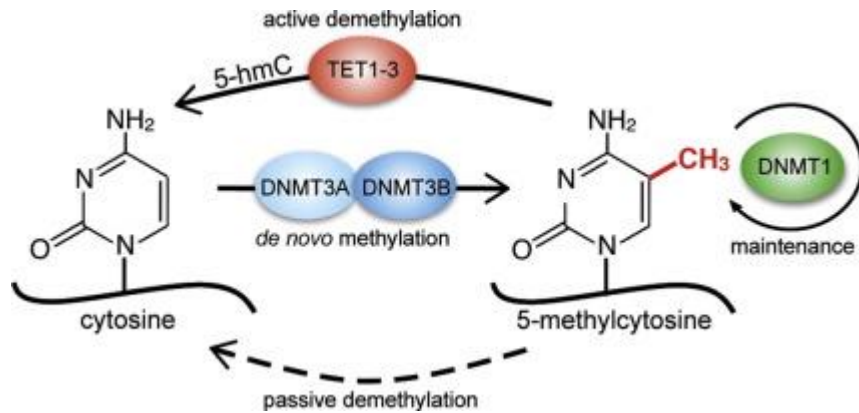


Why use long reads for Plasmids sequencing?

Plasmids can be up to several hundred kb

It is important to be able to know which plasmid an antibiotic resistance gene is on, to be able to track it.

Methylation, what is it, and is it useful information?



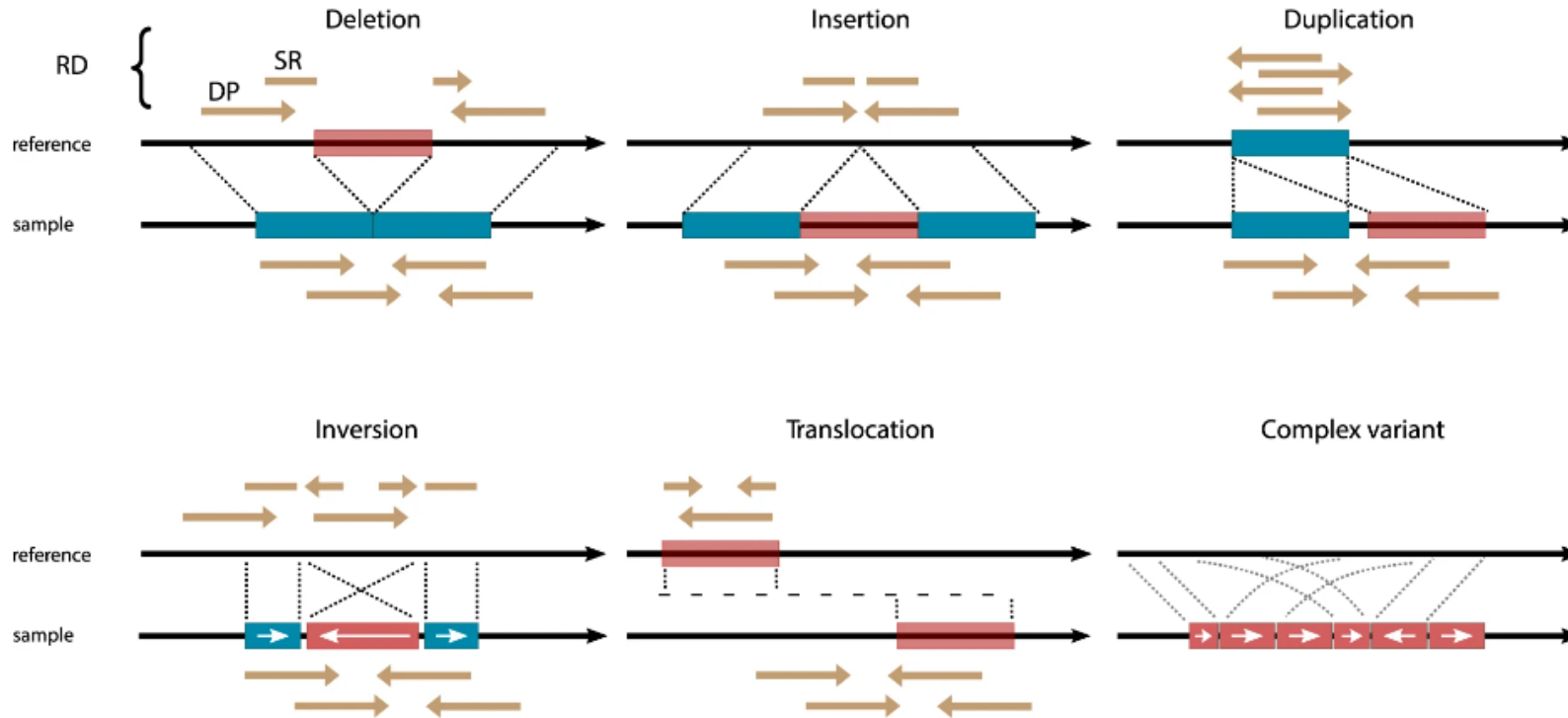
Ambrosi et al. 2017 "Dynamics and context dependent roles of DNA Methylation"

Briefly: gene expression regulation. Among other things, it influences what type of cell a naïve cell becomes, by determining which genes are expressed.

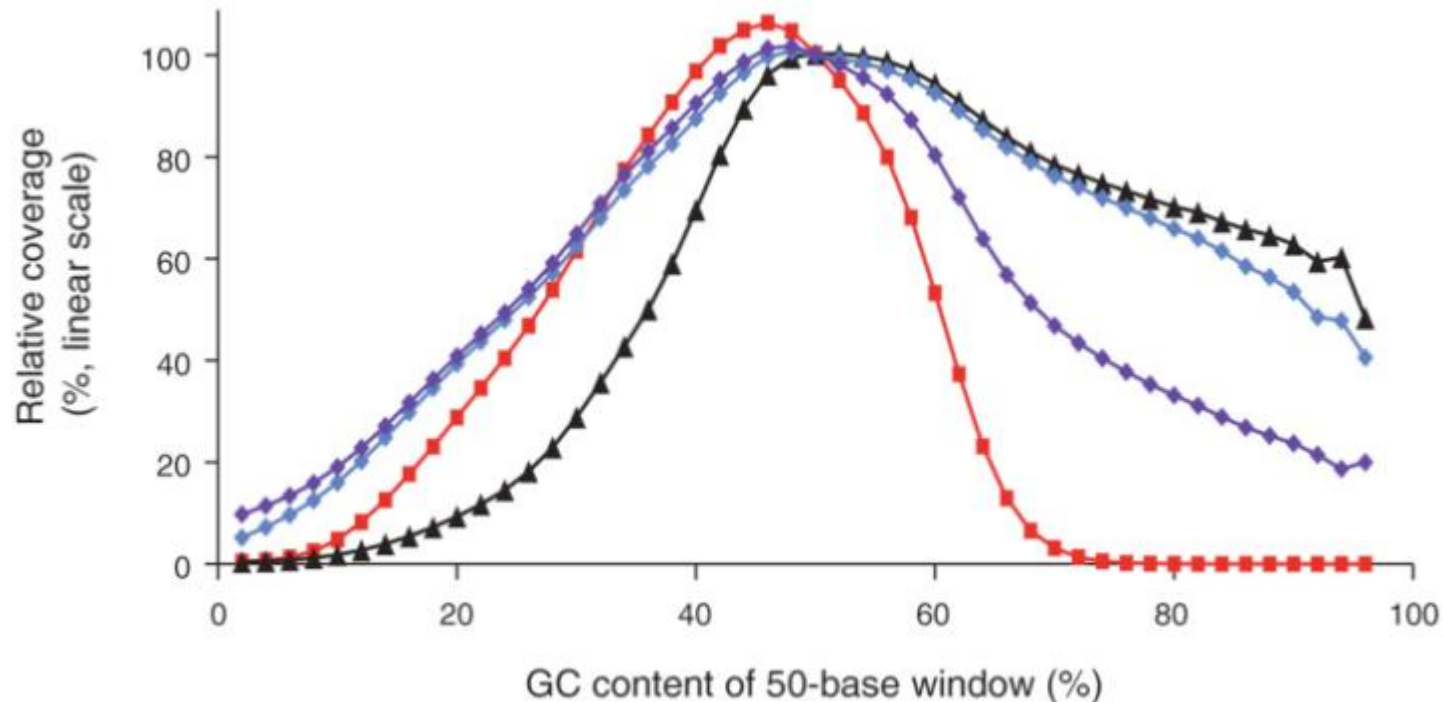
Structural Variants (SVs)

Fig. 1: Major SV types and their characteristic read-alignment patterns.

From: [Structural variant detection in cancer genomes: computational challenges and perspectives for precision oncology](https://www.nature.com/articles/s41698-021-00155-6)



Repetitive regions and their challenge



<https://www.ecseq.com/support/ngs/are-there-regions-in-the-genome-that-are-not-covered-by-dna-sequencing>

Old overview, but still illustrates the issues with short read sequencing

Long read sequencing

Repetitive sequences, why are they a problem when sequencing?:

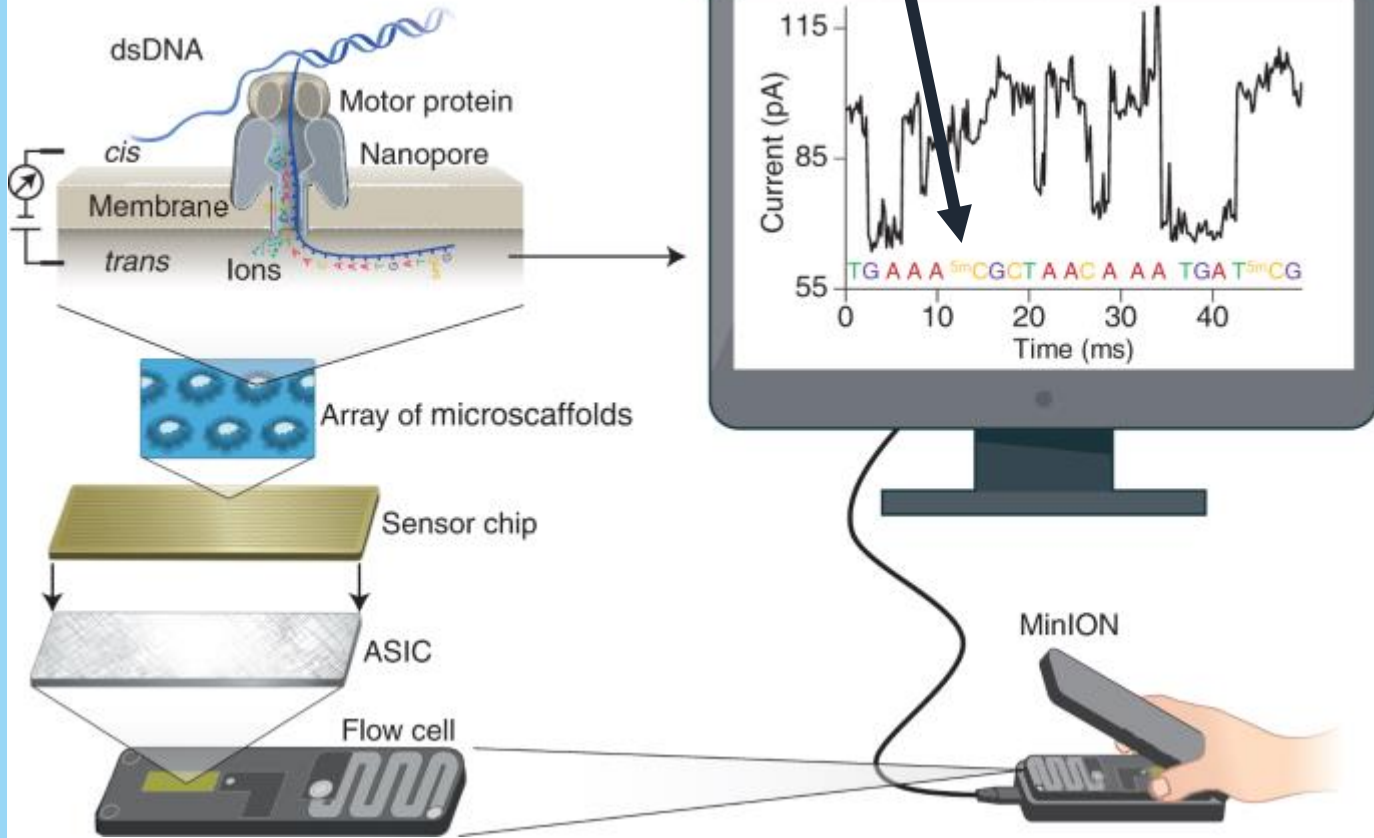
- High or low GC regions are hard to amplify
- DNA fragmentation is not random
- Hard to assemble with short reads, if the repetitive region becomes larger than 100bp. Tandem repeats can be up to 1,7Mbp in size

To overcome this, use sequencing methods that do not require PCR and fragmentation, and produces long reads. Frederikke Byron Pedersen

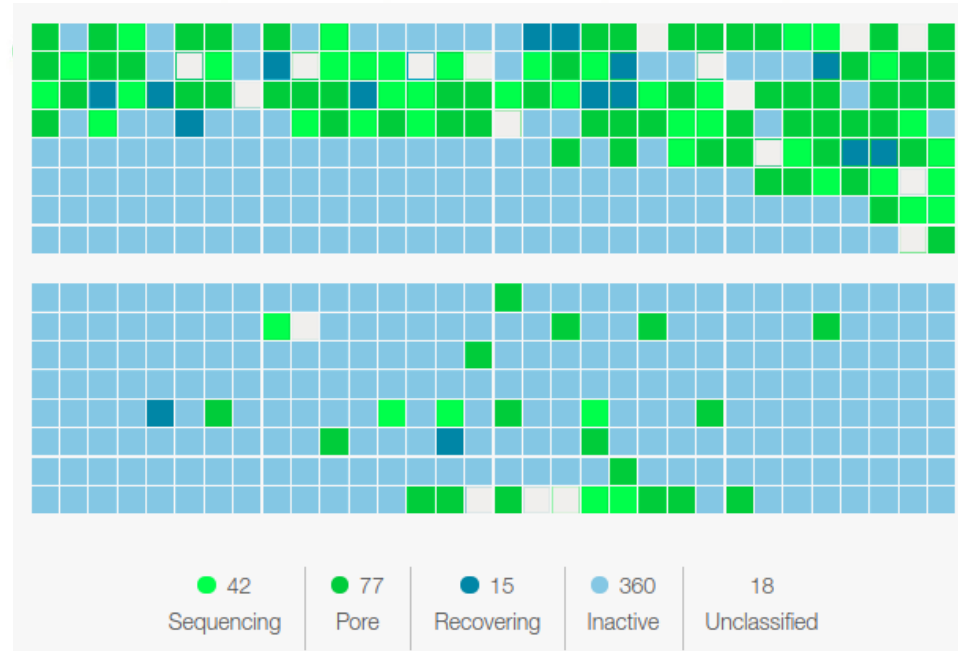
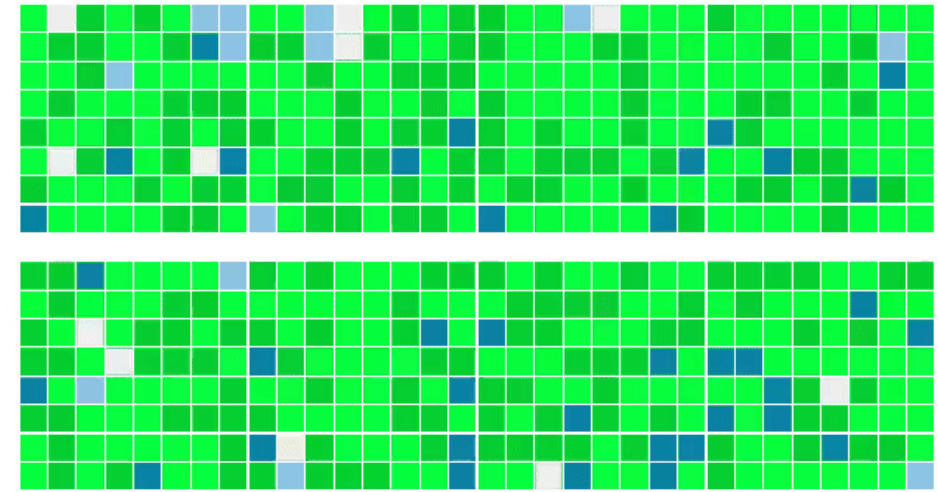
The two main Long Read Sequencing technologies



Nanopore sequencing

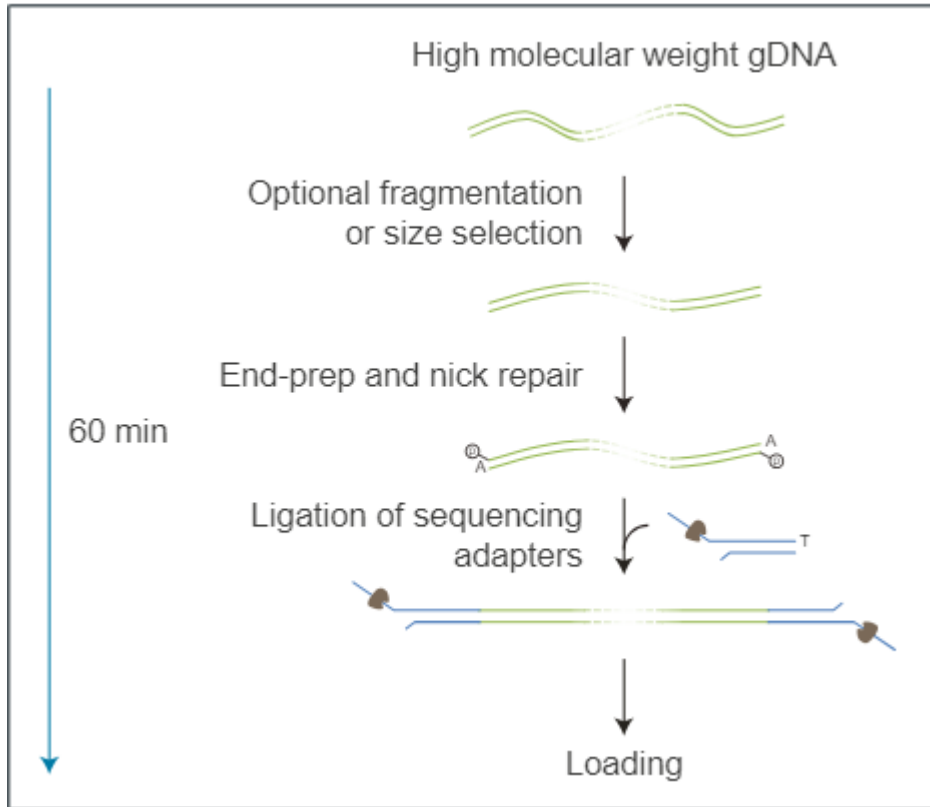


Source: Nature.com
 Nanopore sequencing technology, bioinformatics and applications



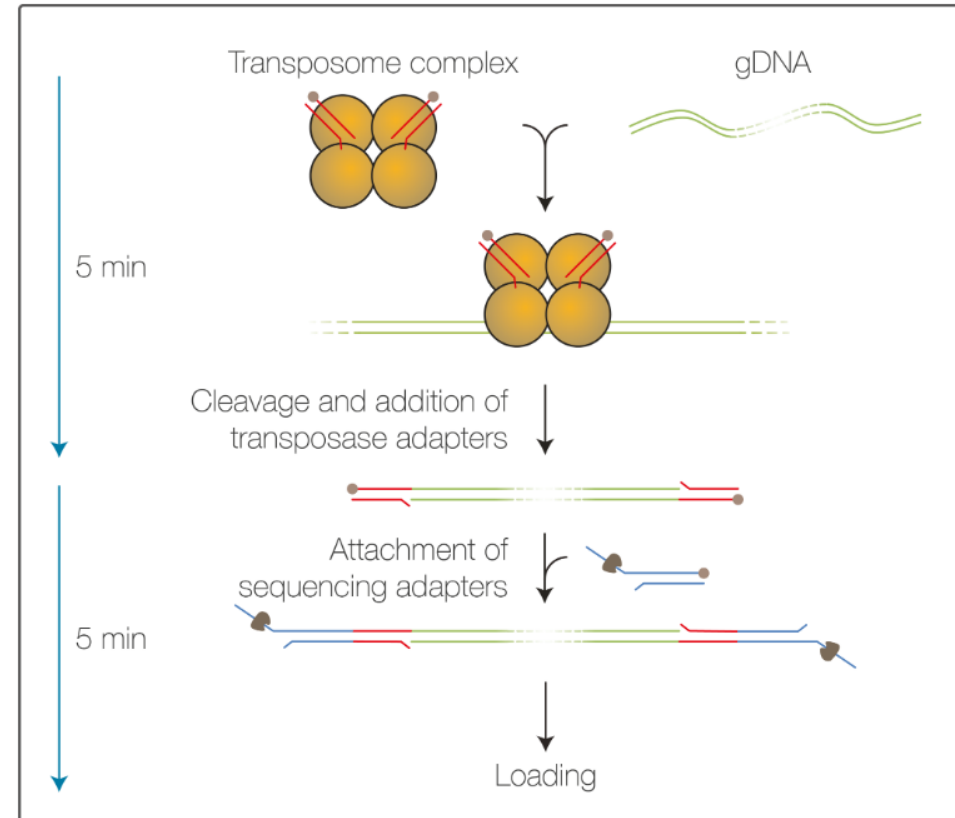
Library preparation: highlighted methods

Ligation Sequencing Kit



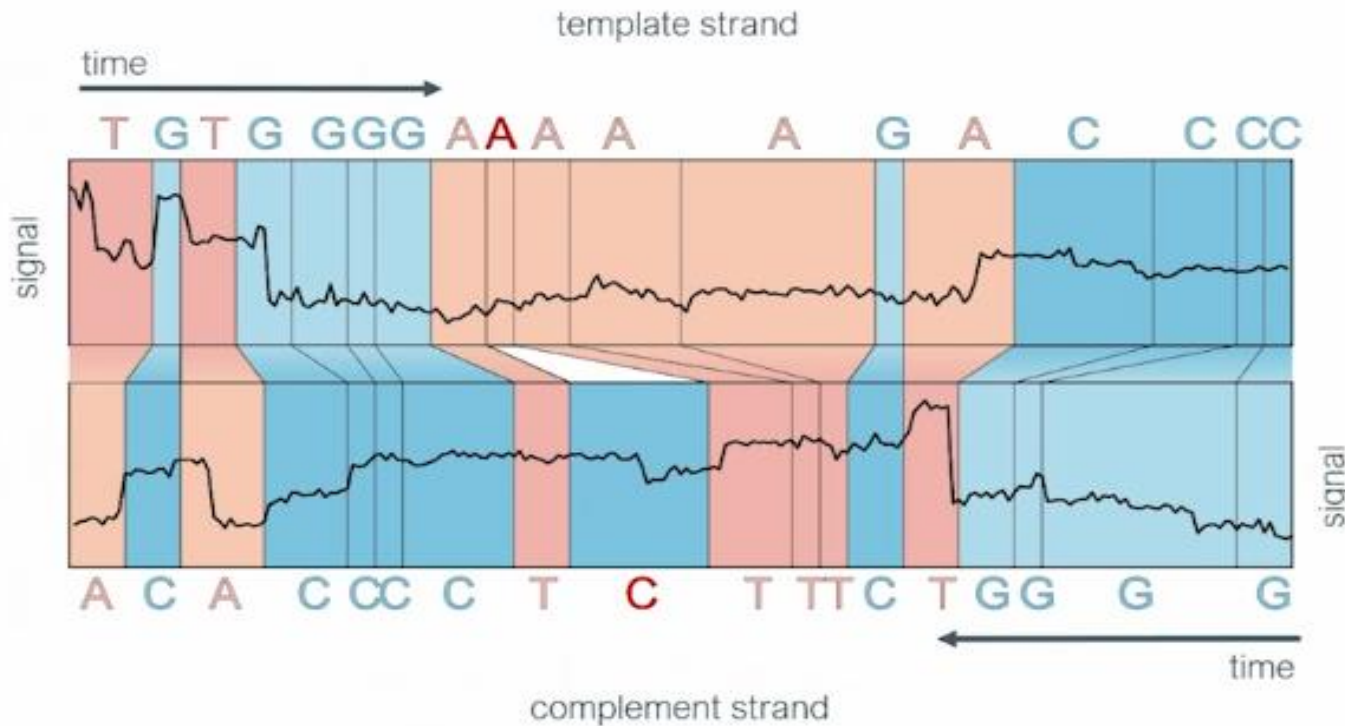
<https://nanoporetech.com/products/prepare/dna-library-preparation>

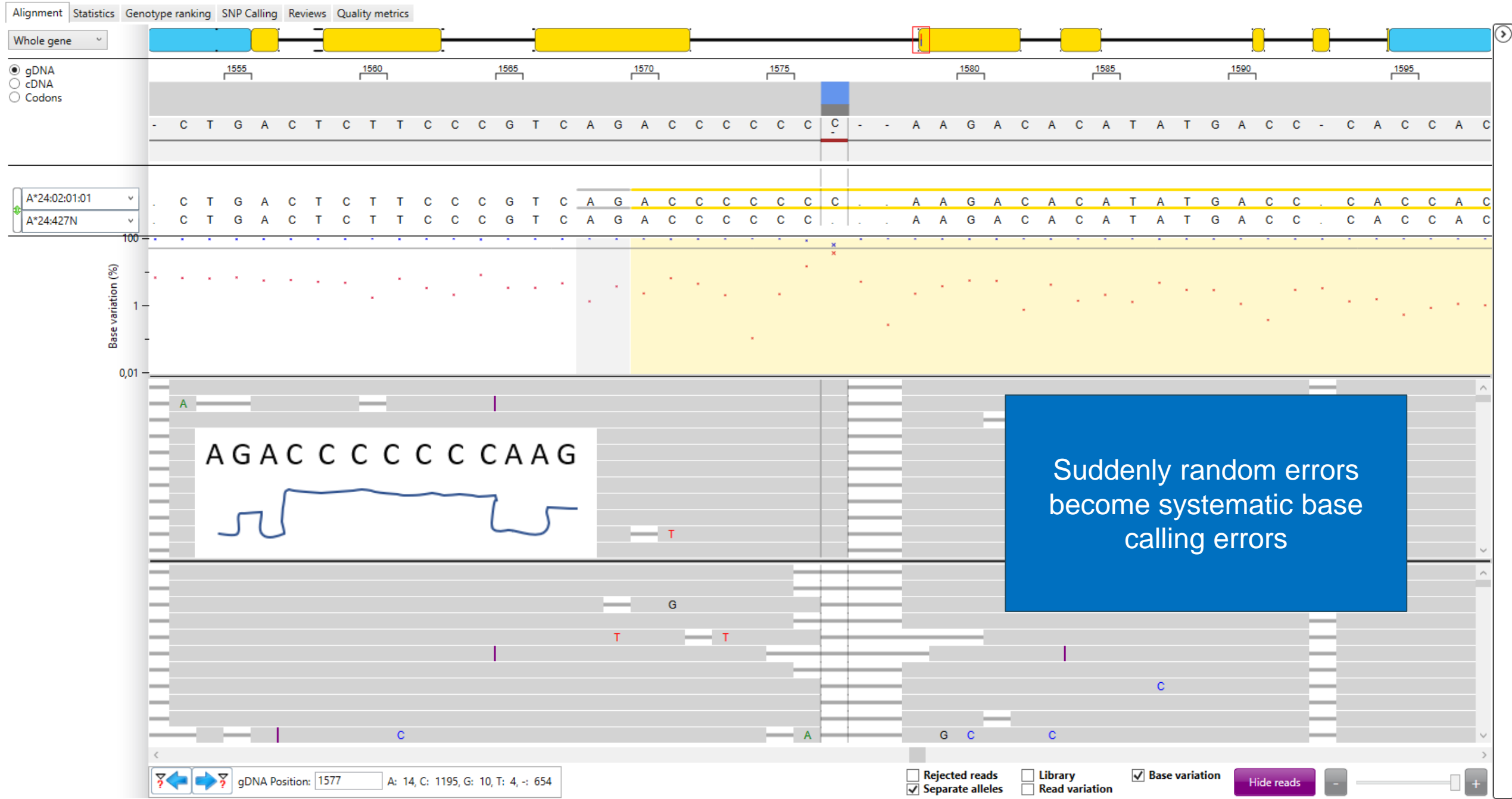
Rapid Sequencing Kit



Both have barcoding options up to 96, and there are many other library preparation methods, including for RNA and cDNA

Biggest hurdle with Nanopore: Homopolymers

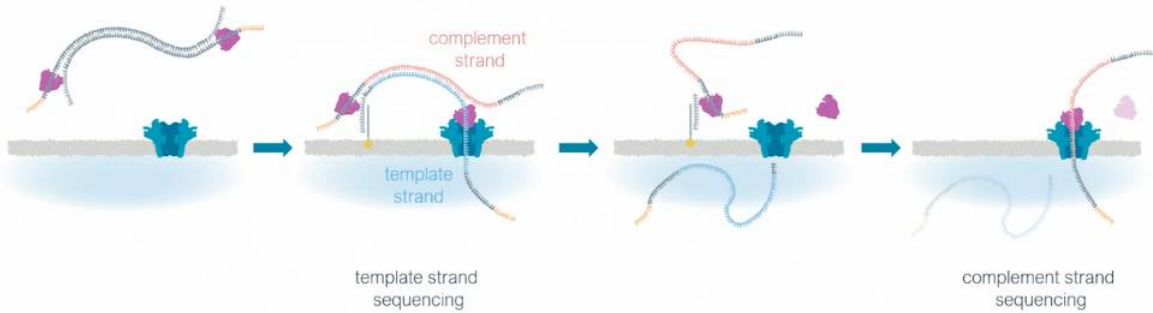




Duplex basecalling to overcome homopolymers?

What base should this have been?

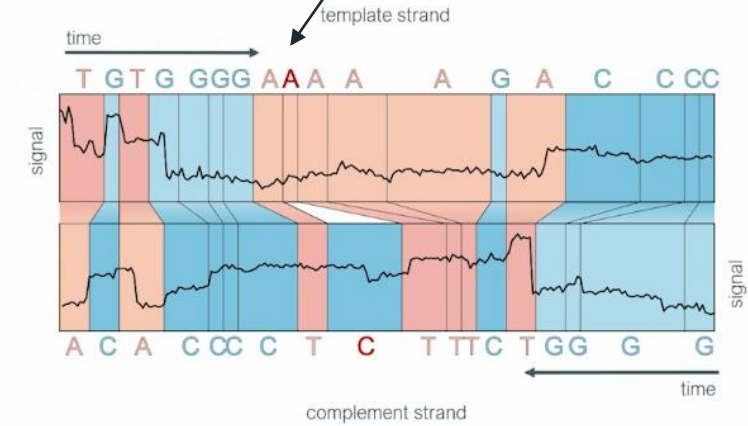
What is duplex sequencing?



Why duplex?

Two measurements are (much!) better than one

- Statistical errors (random noise) are reduced by repeated measurements
- Reverse complement sequence gives orthogonal information



49

© 2022 Oxford Nanopore Technologies plc.
 Oxford Nanopore Technologies products are not intended for use for health assessment or to diagnose, treat, mitigate, cure, or prevent any disease or condition.

• CONFIDENTIAL



54

© 2022 Oxford Nanopore Technologies plc.
 Oxford Nanopore Technologies products are not intended for use for health assessment or to diagnose, treat, mitigate, cure, or prevent any disease or condition.

• CONFIDENTIAL

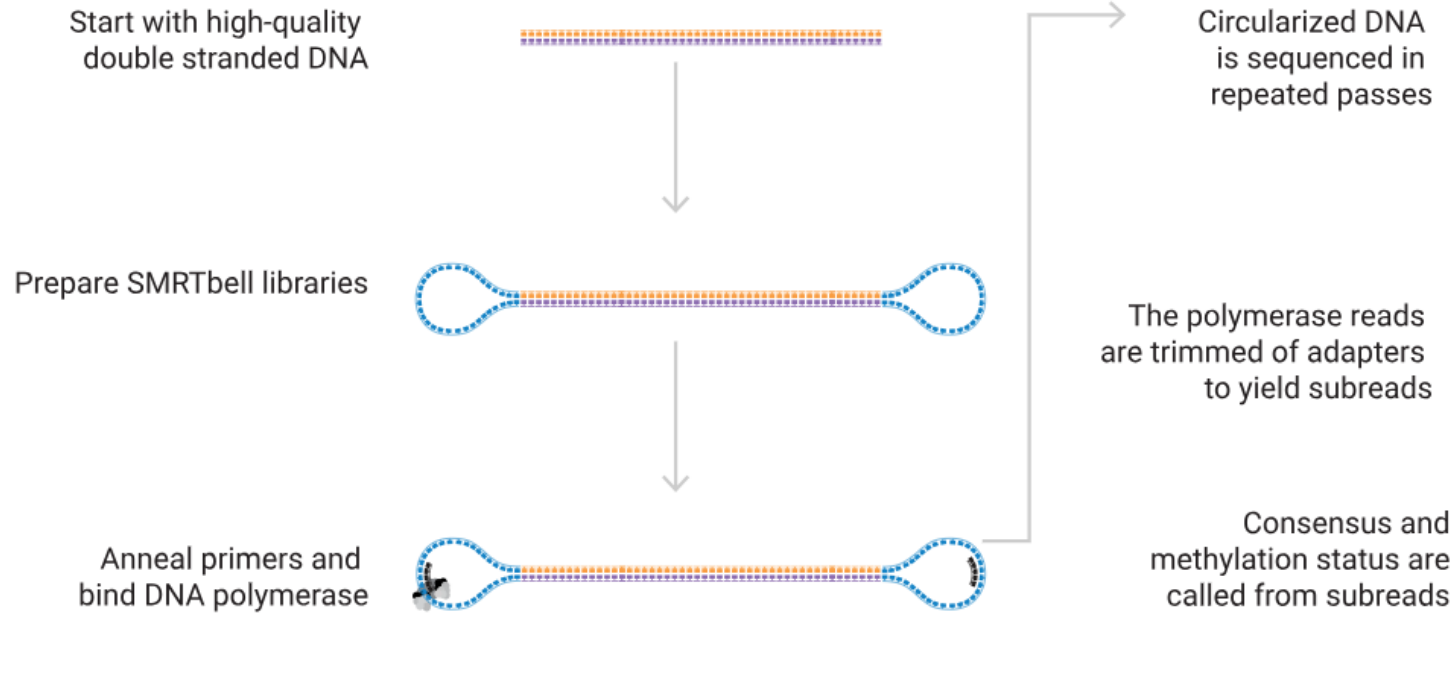


<https://nanoporetech.com/resource-centre/video/ncm22/advances-in-duplex-basecalling>

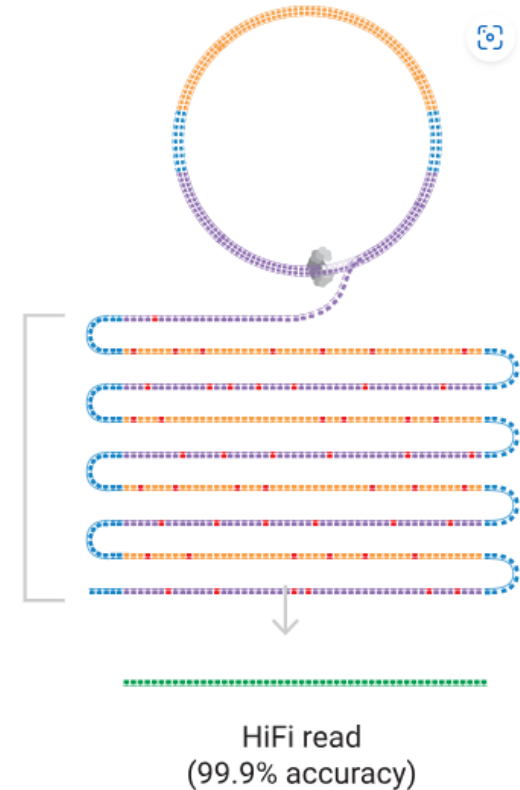
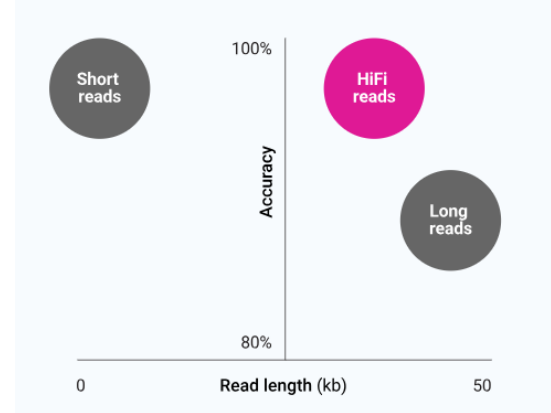
So, better Q/Phred score, but lower output. Only ~20% of the reads are duplex, and this does not work with barcoded reads (yet)

PacBio (HiFi)

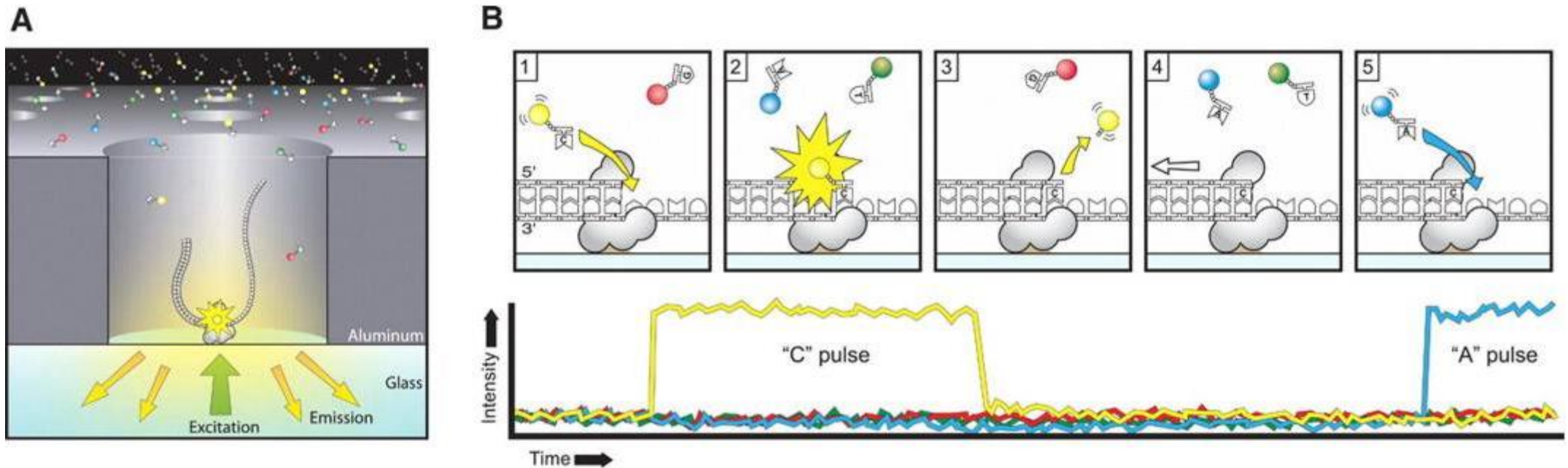
How are HiFi reads generated?



<https://www.pacb.com/technology/hifi-sequencing/>

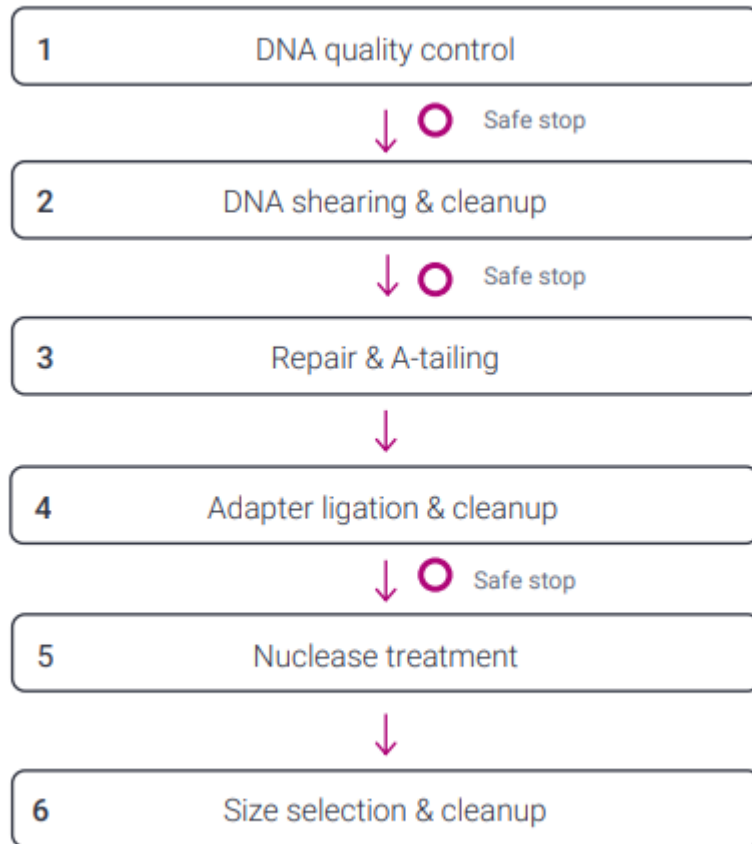


PacBio Sequencing



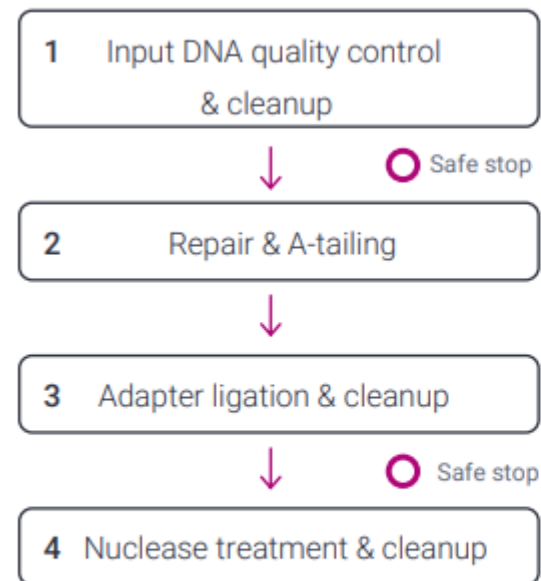
Library preparation

[SMRTbell® prep kit 3.0](#)

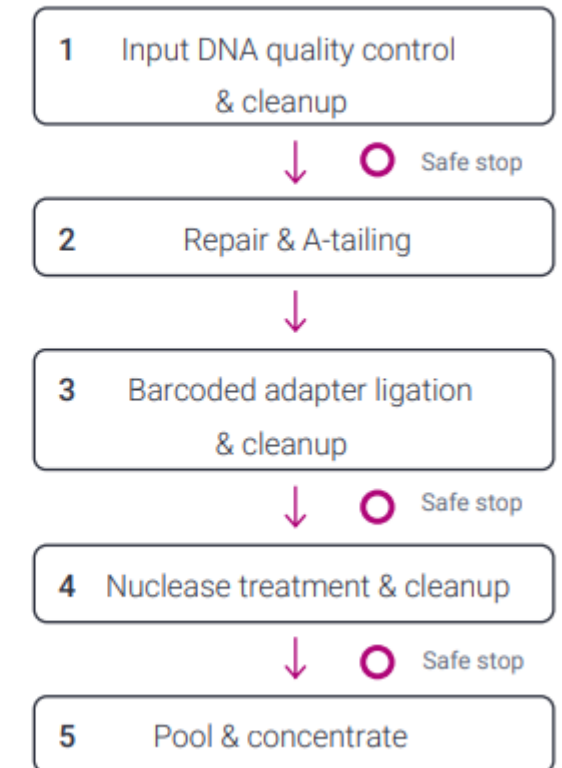


[Preparing multiplexed amplicon libraries using SMRTBell prep kit 3.0](#)

Primer-barcoded samples



Adapter-barcoded samples



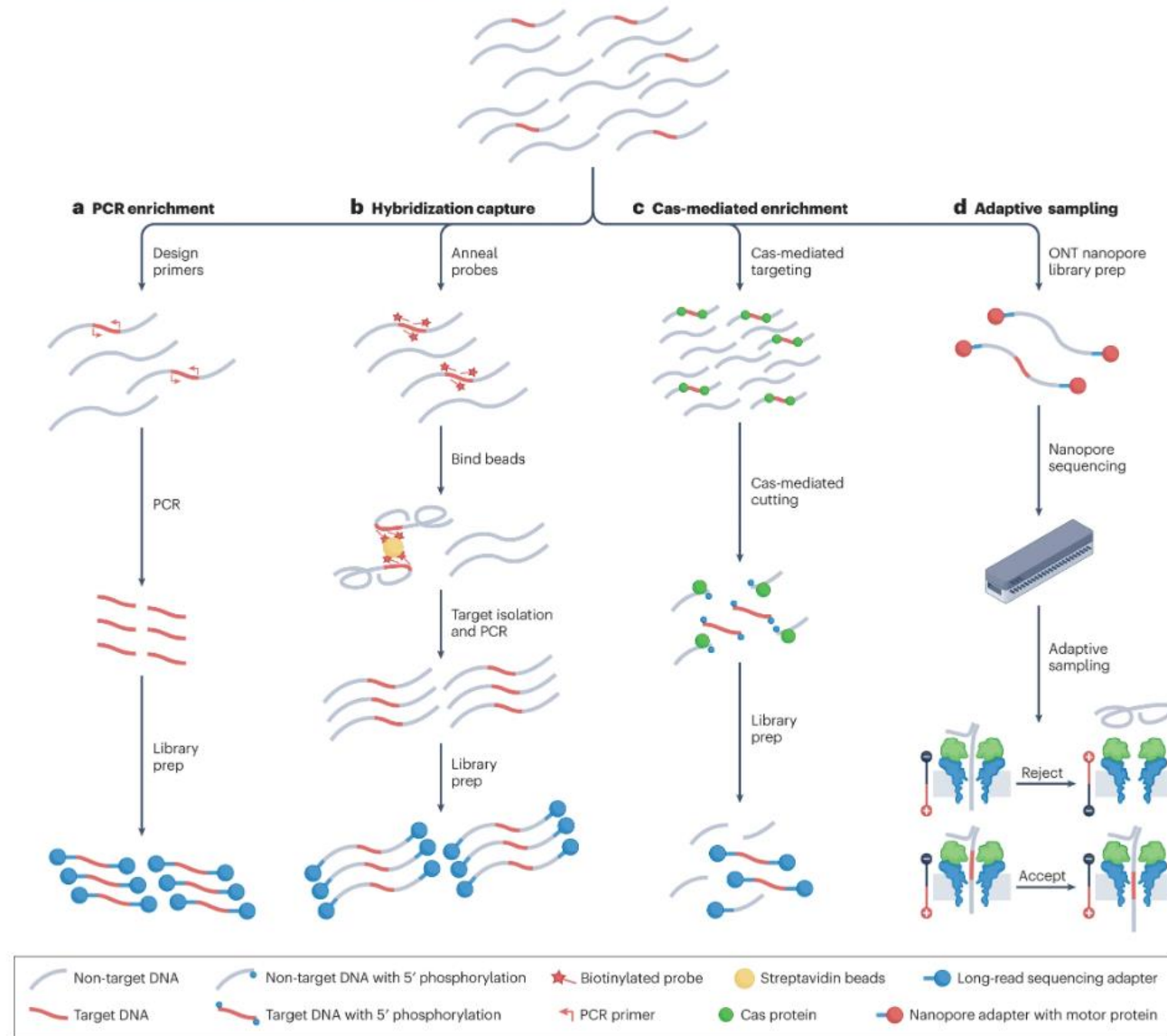
Biggest hurdles with Pacbio HiFi/Revio

- Limited read length
- Requires shearing or amplification, which may lead to low coverage of high or low GC rich areas
- Fixed sequencing time for a SMRT cell
- It's large



Fig. 1: Long-read targeted sequencing methods.

From: [Beyond assembly: the increasing flexibility of single-molecule sequencing technology](https://www.nature.com/articles/s41576-023-00600-1)



The enrichment strategies can be used on all sequencing platforms, except Adaptive Sampling, which requires the use of Nanopore's ability to flick out a read using the positive and negative charge.

My experiences with Long Read Sequencing

- Plasmids, tracking of antibiotic resistance – SSI
- Cas9 Mediated Enrichment – Oxford Nanopore Technologies (ONT)
- Leukemia panel (AML), one diagnostic tool for all types of AML – ONT
- Covid-19 test – ONT
- De Novo Assembly of viral DNA in metagenomic samples, getting a consensus/reference sequence for new variants/virus – ONT
- HLA typing – The University Hospital of Copenhagen, Rigshospitalet

Comparison of the 3 major players in sequencing

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%
50	1 in 100000	99.999%

	ILLUMINA (MiSeq)	NANOPORE (MinION)	PACBIO (Revo/SMRT)
Maximum read length	600bp	None (record >4mb)	25kb
Cost per instrument	\$128,000	\$1,000*	\$779,000
Q score	~Q40 (Min Q30)	~Q20 (Personal experience say it's Q15)	~Q30 (Min Q20)
Minimum Run Time	4h (1x36bp) 56h (2x300bp)	None	24h
Methylation?	No	Yes	Yes
Direct RNA Seq	No	Yes	No
Mobile	No	Yes	No
Reusable flow cells	No	Yes	No

*lease of the instrument

Quiz

Same questions, compare your answers to those at the start of the lecture. Have they changed?

Question 1:

What is the longest read that can be obtained using Nanopore Sequencing? Note, not average read length

1. 25kb
2. 100kb
3. 10kb
4. No limit

Question 2:

What is the longest read that can be obtained using PacBio Sequencing? Note, not average read length

1. 25kb
2. 100kb
3. 10kb
4. No limit

Question 3: What is phasing?

1. To introduce something in stages over a particular period of time
2. Phasing is the rhythmic equivalent of cycling through the phase of two waveforms as in phasing
3. The process of statistical estimation of haplotypes from genotype data
4. All of the above

Question 4:

What is one of the main advantages of Long Read Sequencing?

1. Haplotyping and SVs are easier resolved
2. Cheaper and Faster
3. Higher quality/Q-Score
4. All of the above

Question 5: Which sequencing platform is the best?

1. PacBio
2. Nanopore
3. Illumina
4. It depends on what you want to do

Questions?