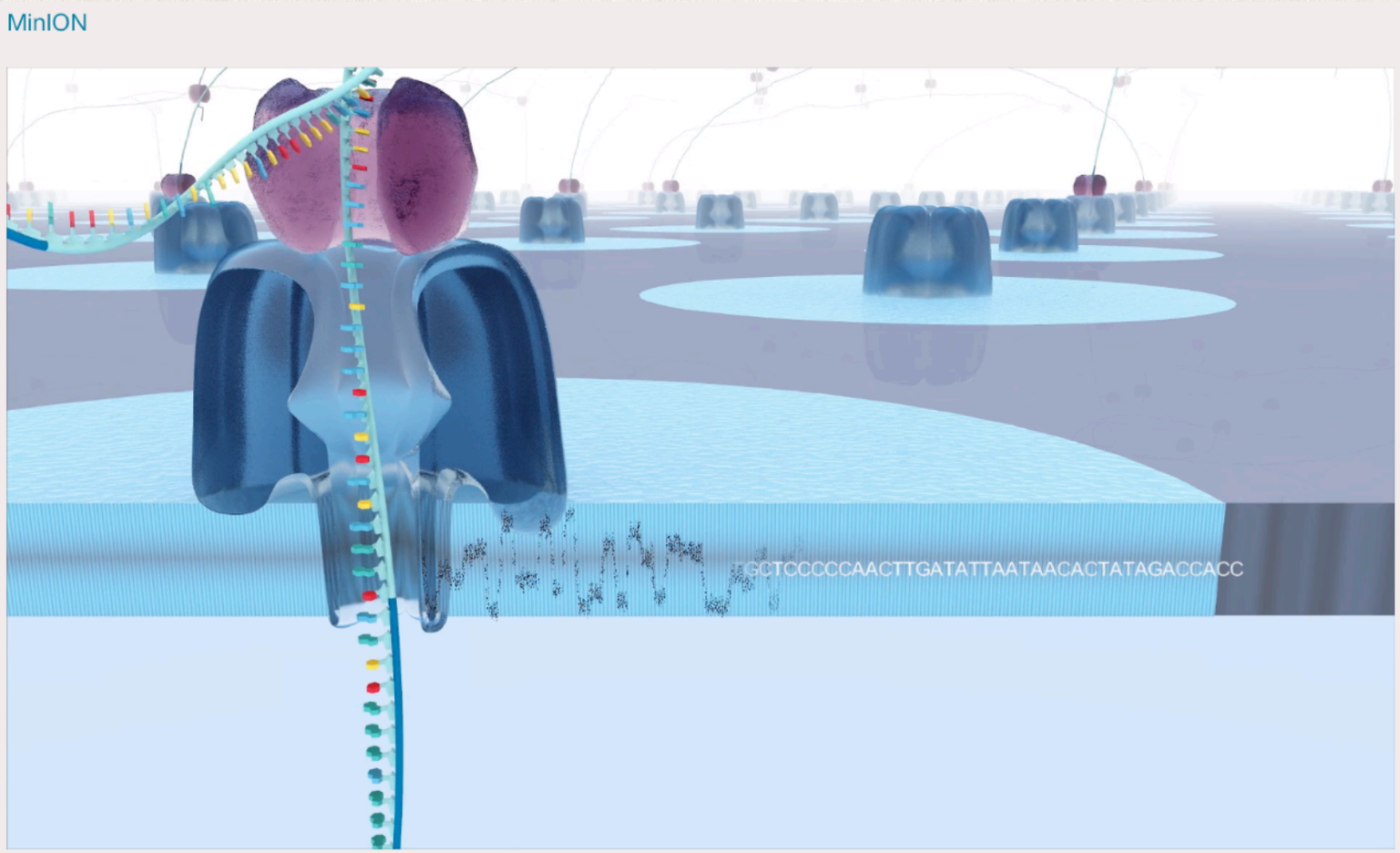


Genome Assembly and Comparative Transcriptomics

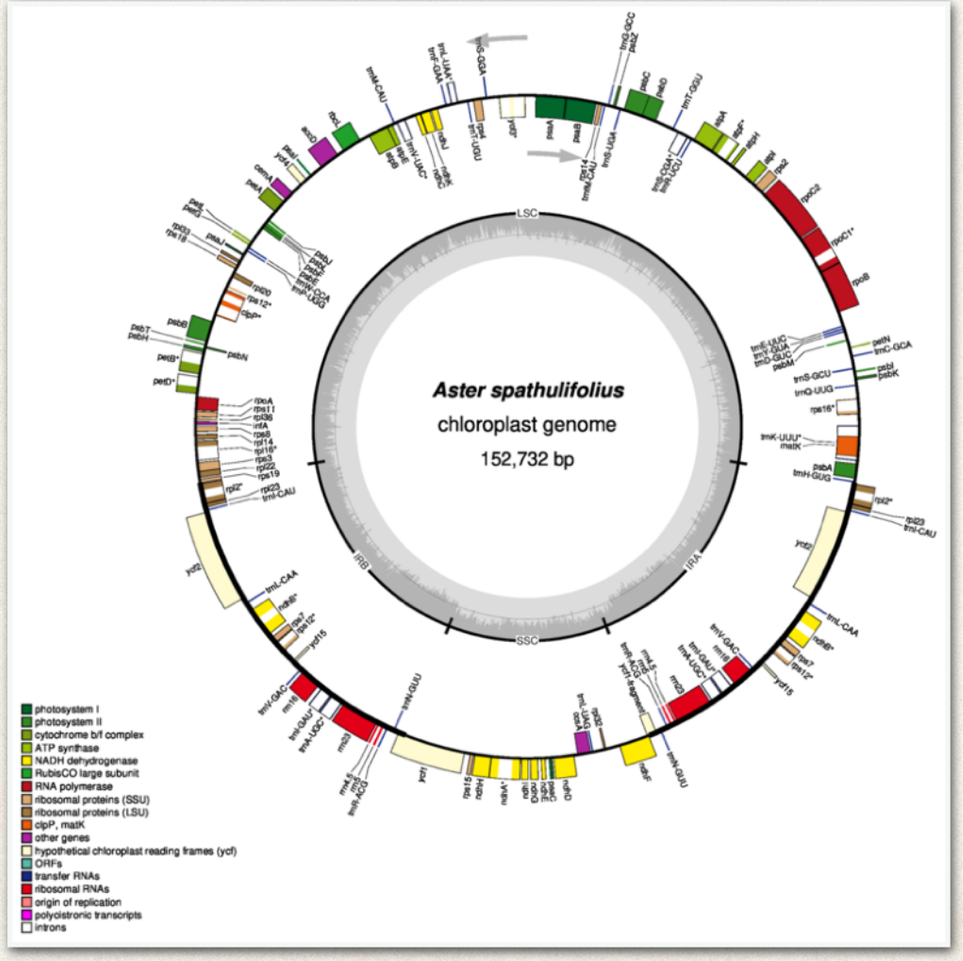
Florencia Díaz Viraqué

Institut Pasteur Montevideo

6 Julio 2021

Genome assembly

- A genome is the complete genetic information of an organism or a cell.

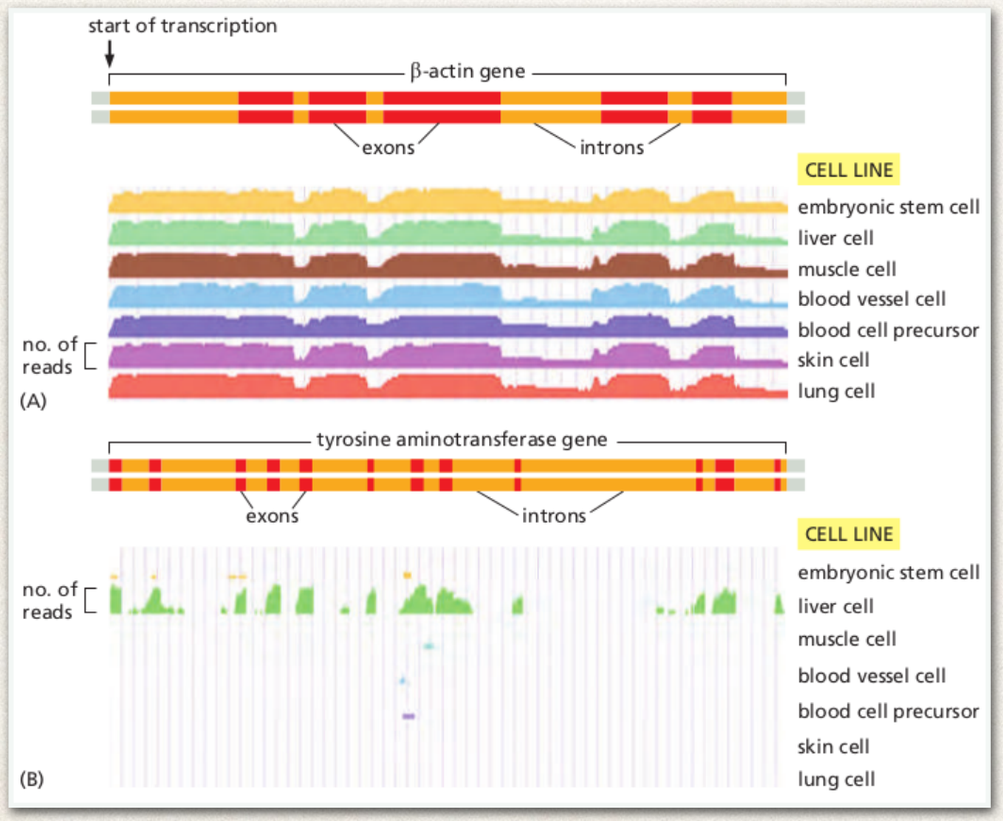


Tyagi et al., 2020. *Plants*.

- Single or double stranded nucleic acids store this information in a linear or in a circular sequence.

Comparative transcriptomics

- A transcriptome is the full range of RNAs, or mRNA, molecules expressed by an organism, tissue or cell.



Molecular Biology of the Cell 6th edition - Alberts

- The transcriptome actively changes, varies depending on many factors, including stage of development and environmental conditions.

- To precisely determine these sequences (genome and transcriptome), sequencing technologies have been developed.

Sequencers

- Sequencers generate sequences, known as **reads**, comprised only in defined ranges of lengths, usually far shorter than the size of the genomes investigated.

Illumina

Oxford Nanopore

PacBio

Genome or transcriptome assembly

- The complete genome (or transcript) sequence has to be deduced from the overlaps of these shorter fragments, a process defined as *de novo* **genome/transcriptome assembly**

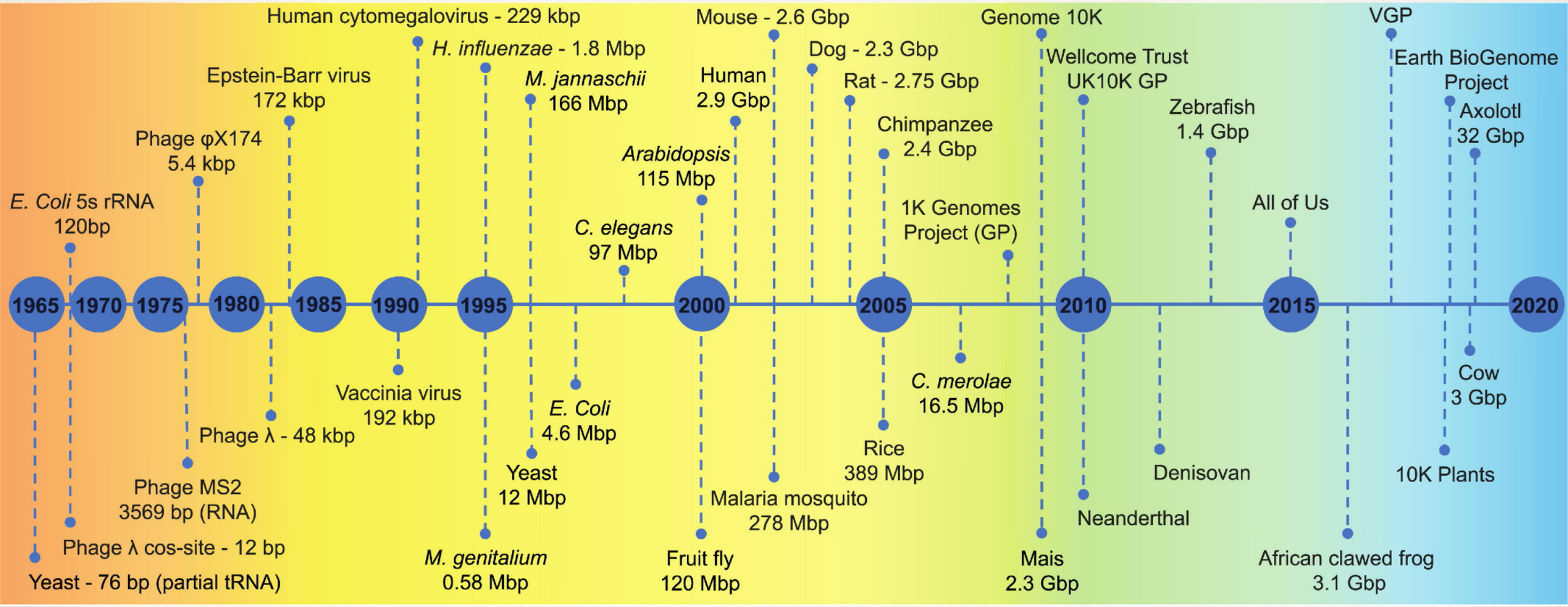
PacBio

Oxford Nanopore

Illumina

Timeline illustrating the major genome assembly achievements

pb = base pair / kbp = 1,000 pb / Mbp = 1,000,000 pb / Gbp = 1,000,000,000 pb

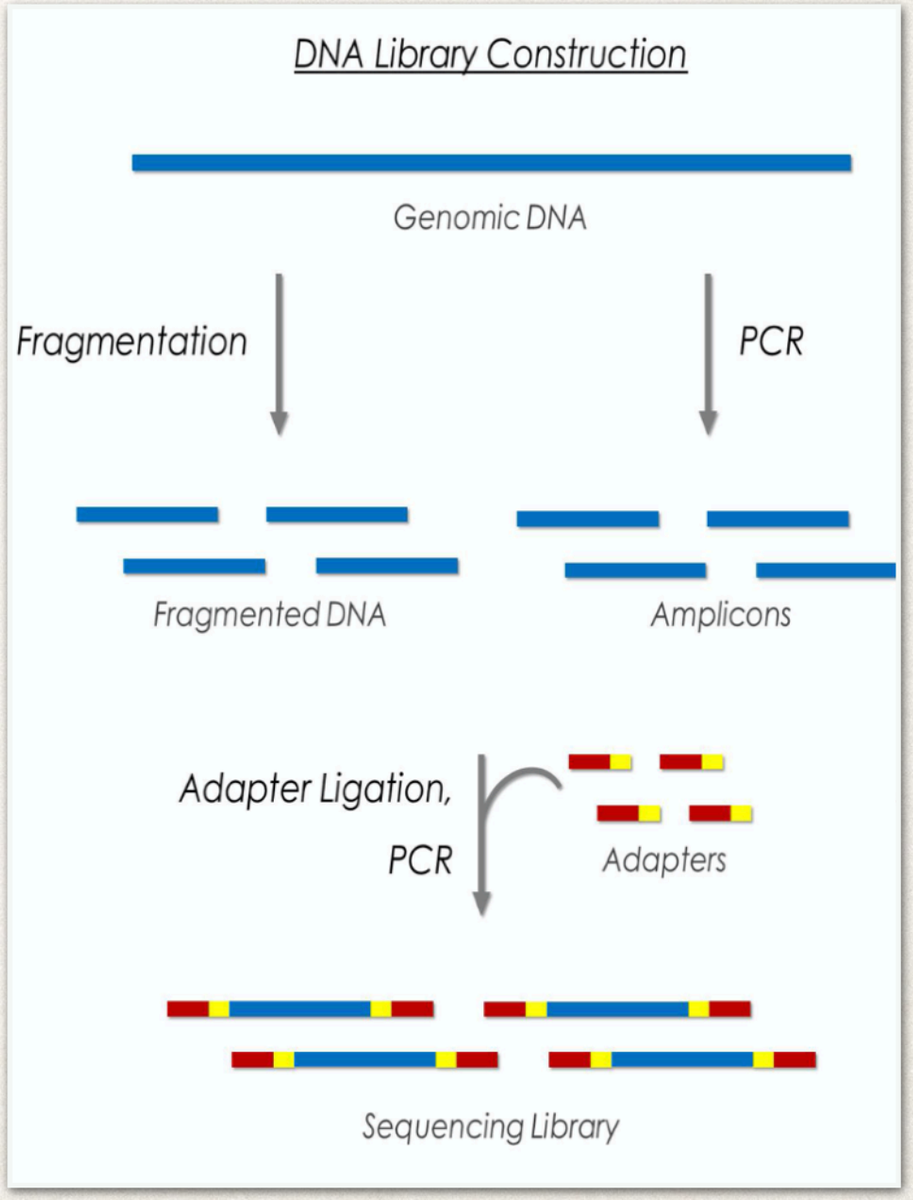


Giani et al., 2020. *Computational and Structural Biotechnology Journal*.

Light red: early sequencing methods, Yellow: Sanger-based shotgun sequencing, Green: **SGS**, Light blue: **TGS**.

Second-Generation Sequencing (SGS)

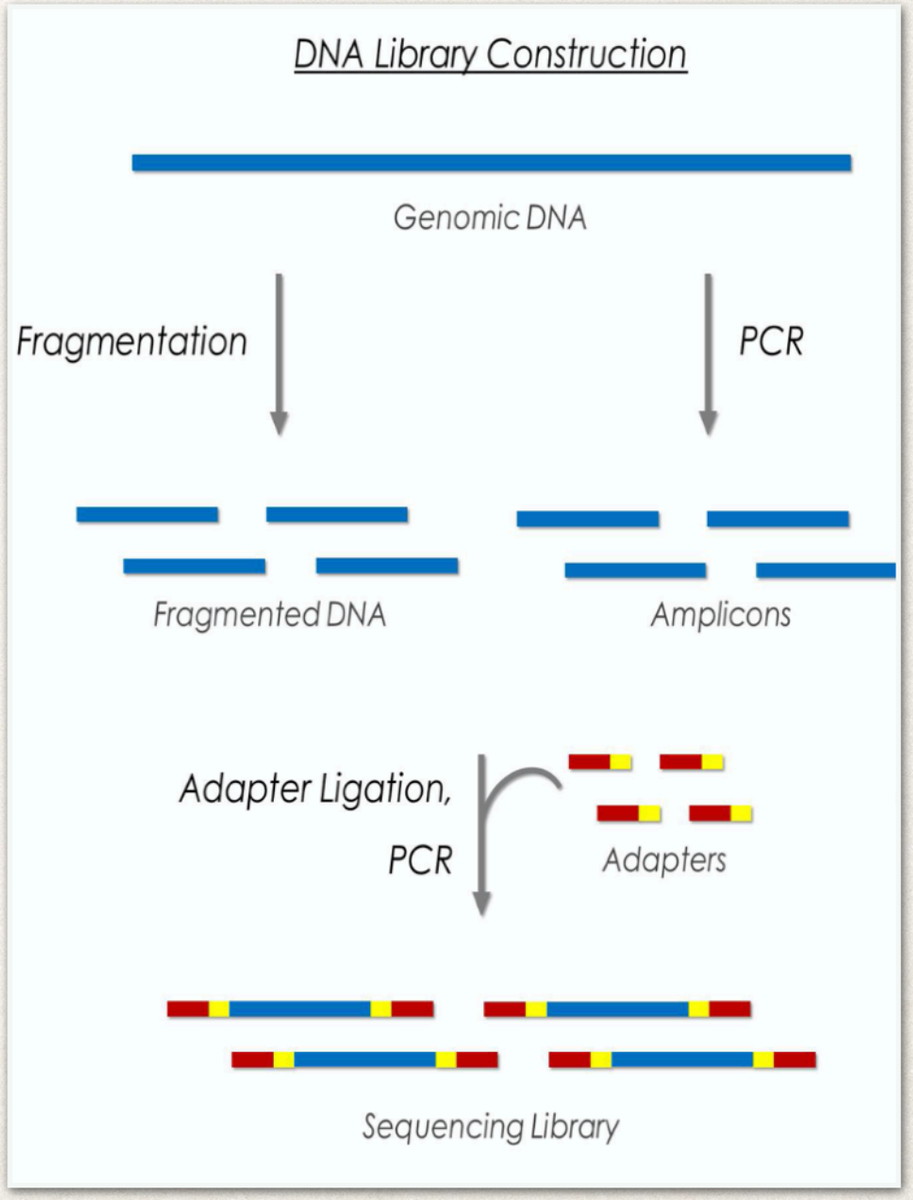
- SGS platforms have been developed to address larger genomes, in a process called Whole Genome Sequencing (WGS).



Modified from https://www.biocompare.com/Molecular-Biology/9187-Next-Generation-Sequencing/

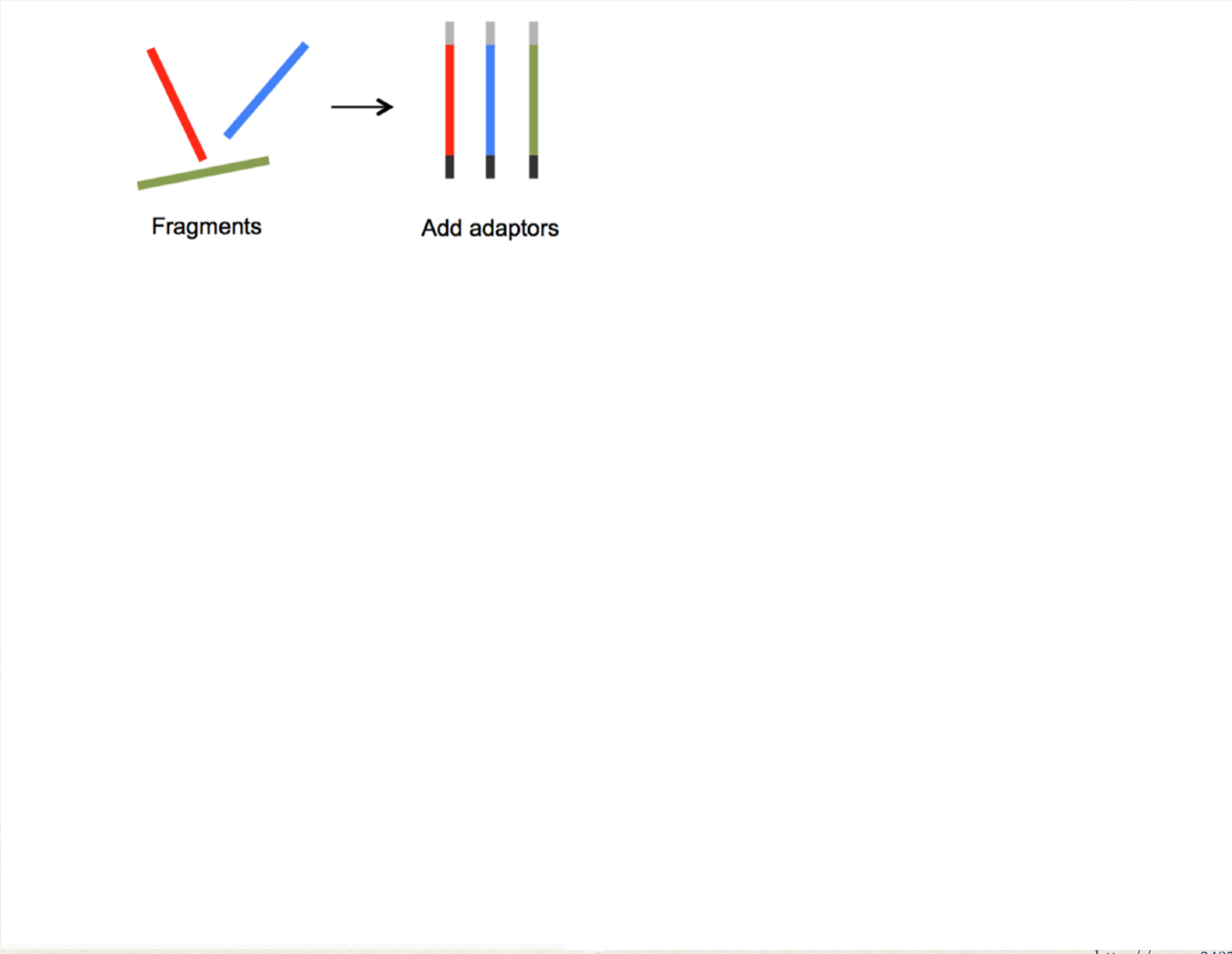
Second-Generation Sequencing (SGS)

- <u>Classical lib prep protocol:</u>
- DNA fragmentation
- Fragment size selection
- Adapters are ligated to the ends of each fragment
- Step of DNA amplification
- Library is loaded on a flow cell and sequenced in Massive Parallel Sequencing reactions



Modified from https://www.biocompare.com/Molecular-Biology/9187-Next-Generation-Sequencing/

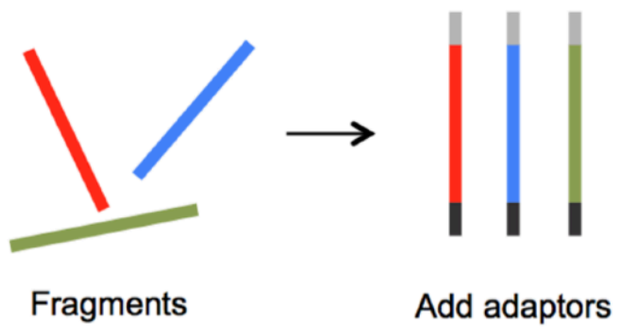
SGS: Illumina sequencing



http://www.3402bioinformaticsgroup.com/service/

SGS: Illumina seque

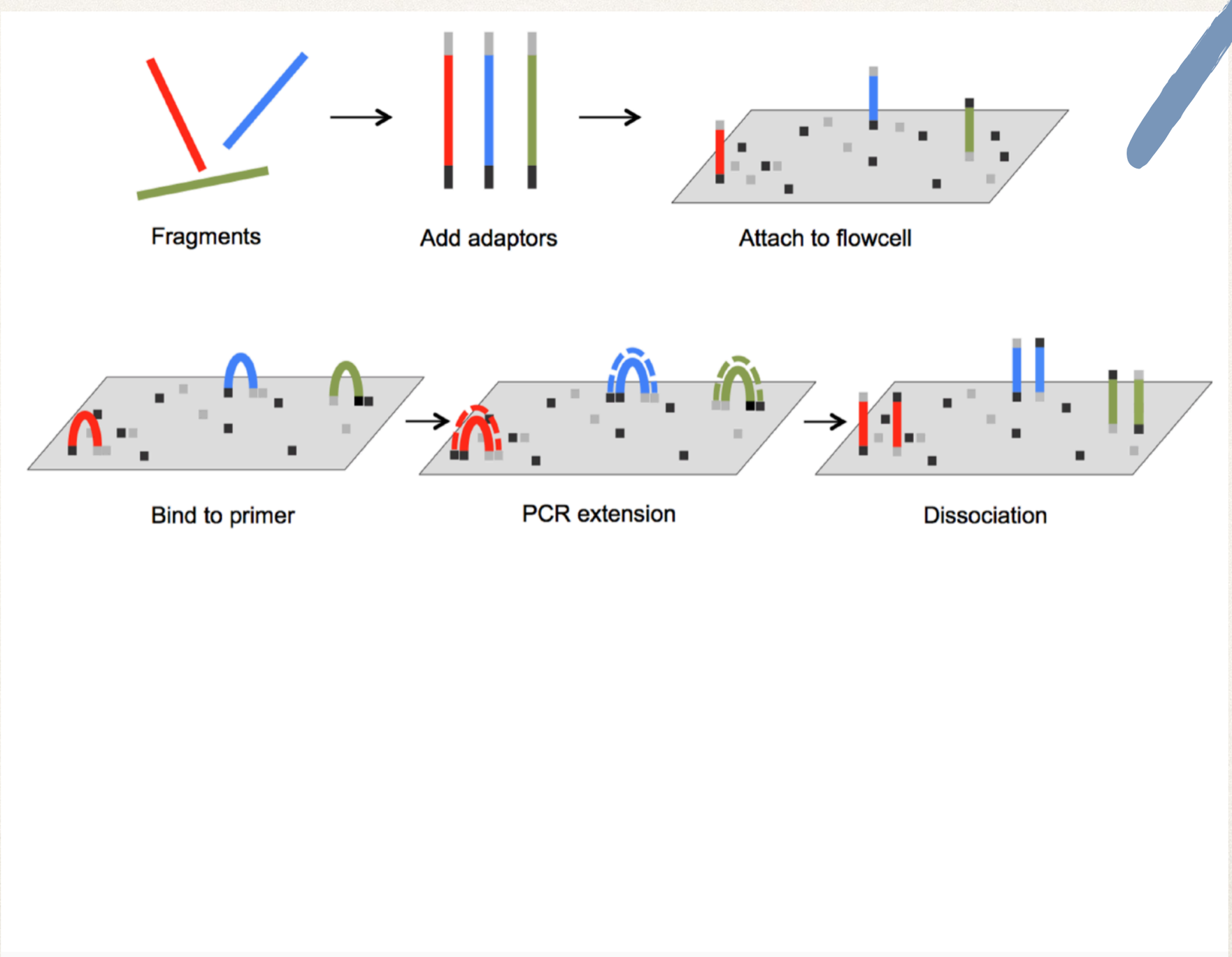
Genomic DNA is randomly fragment and adaptors are annealed to the ends of sequence fragments



http://www.3402bioinformaticsgroup.com/service/

SGS: Illumina sequer

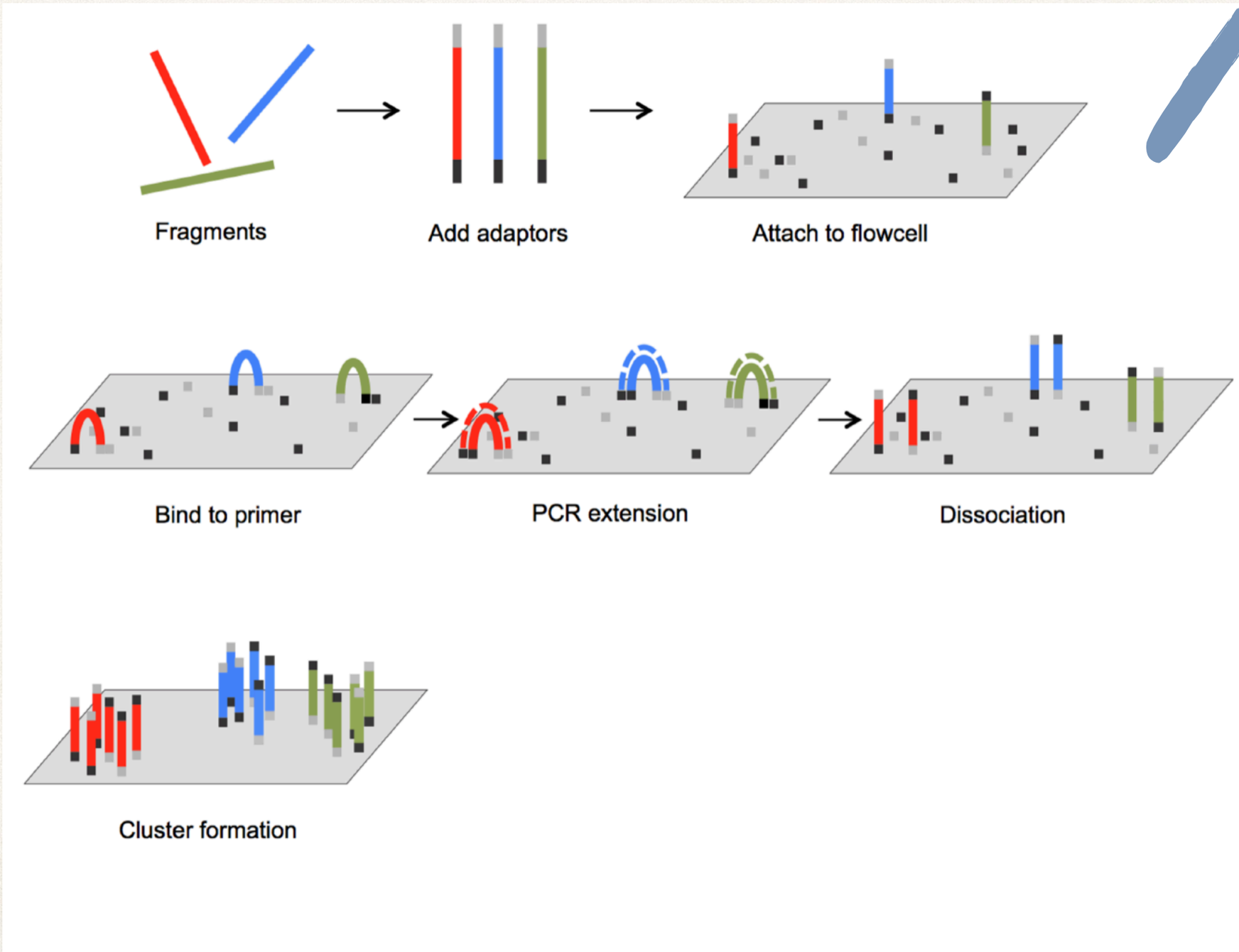
Fragments bind to the flow cell and bridge PCR reactions amplify each bound fragment to produce clusters of fragments



http://www.3402bioinformaticsgroup.com/service/

SGS: Illumina sequer

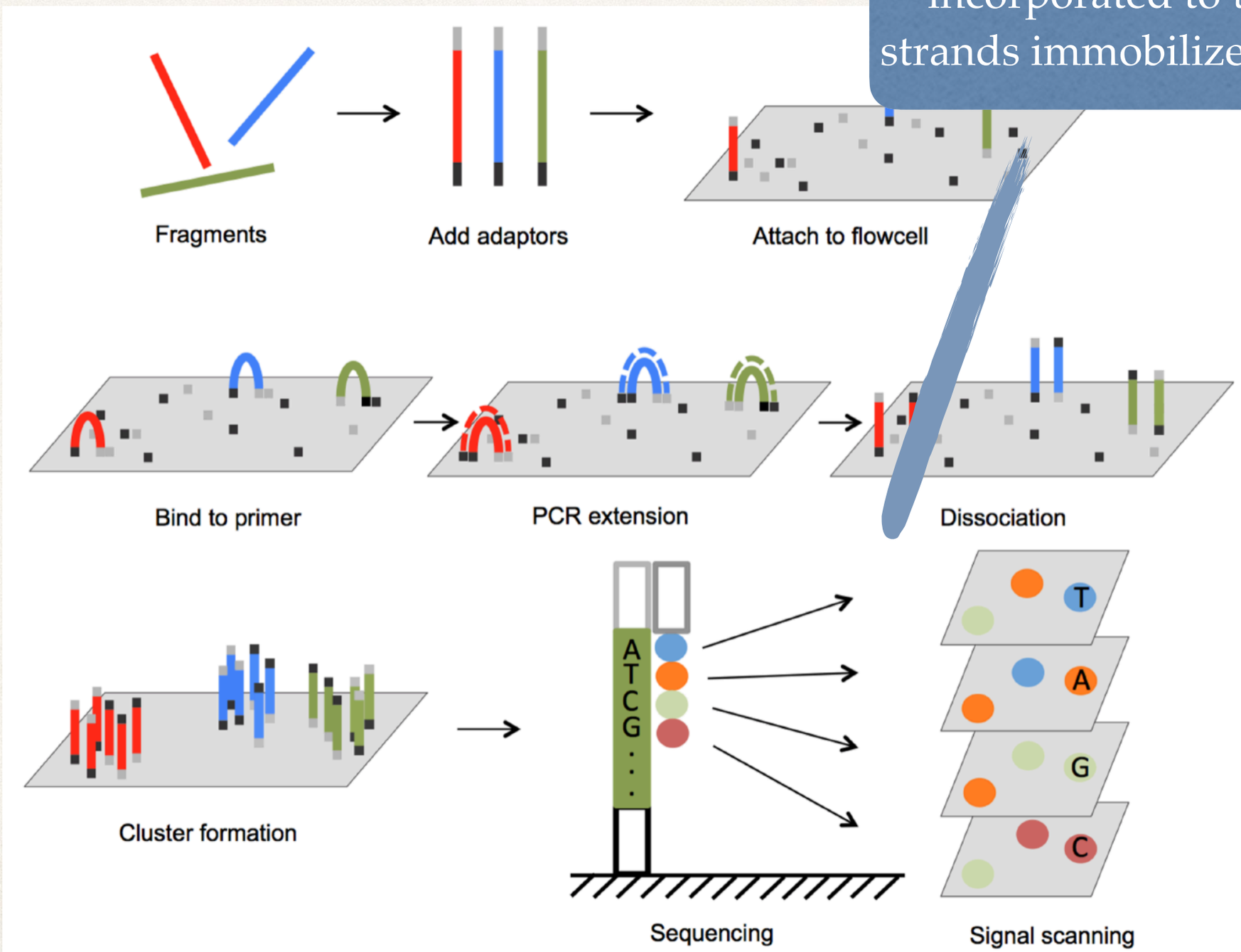
Fragments bind to the flow cell and bridge PCR reactions amplify each bound fragment to produce clusters of fragments



http://www.3402bioinformaticsgroup.com/service/

SGS: Illumina sequen

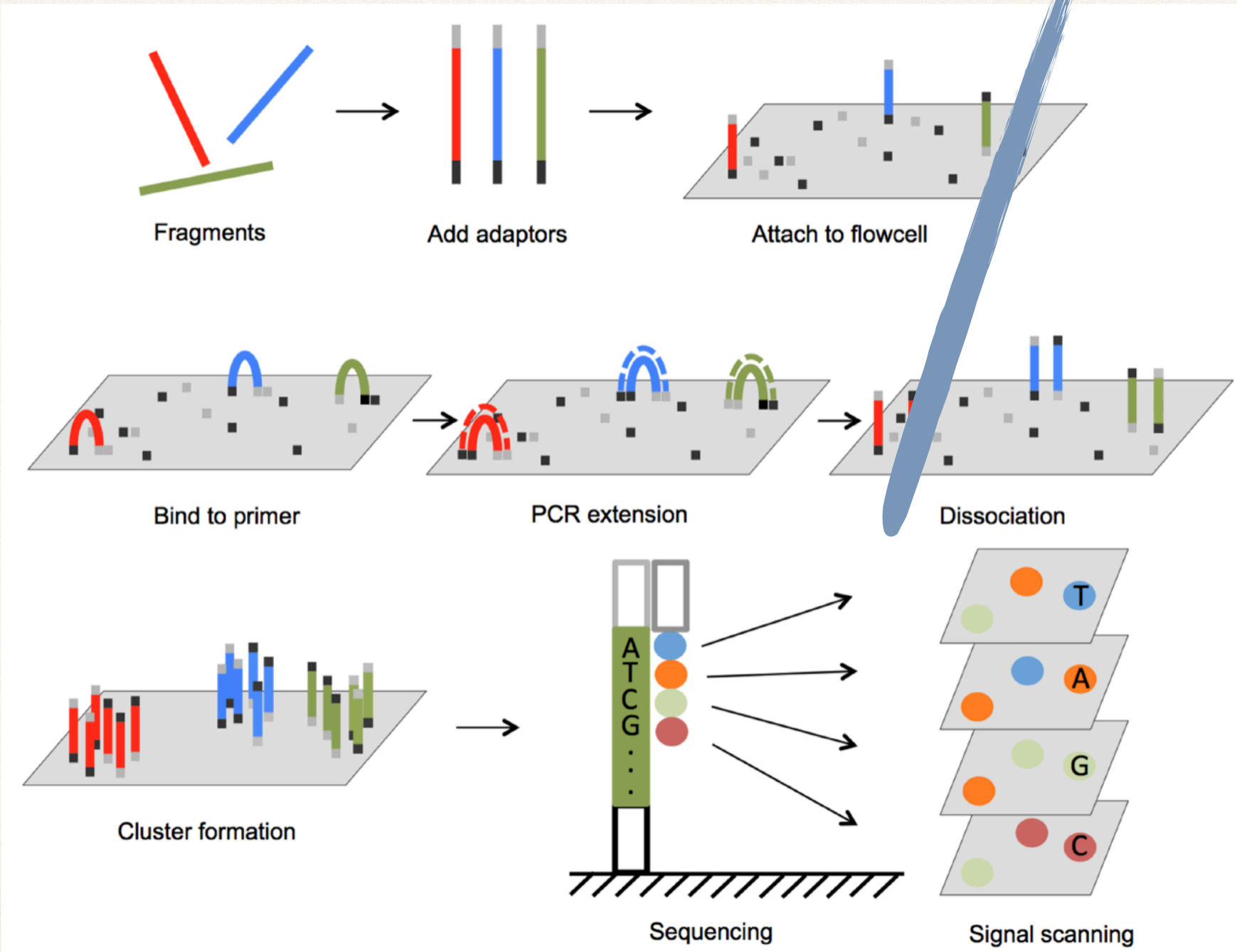
Illumina use sequencing by synthesis technology in which one fluorophore attached nucleotide is added each cycle, polymerase-mediated incorporated to the growing strands immobilized on a surface



http://www.3402bioinformaticsgroup.com/service/

SGS: Illumina sequen

Laser excites the fluorophores in all the fragments that are being sequenced and an optic scanner collects the signals from each fragment cluster.

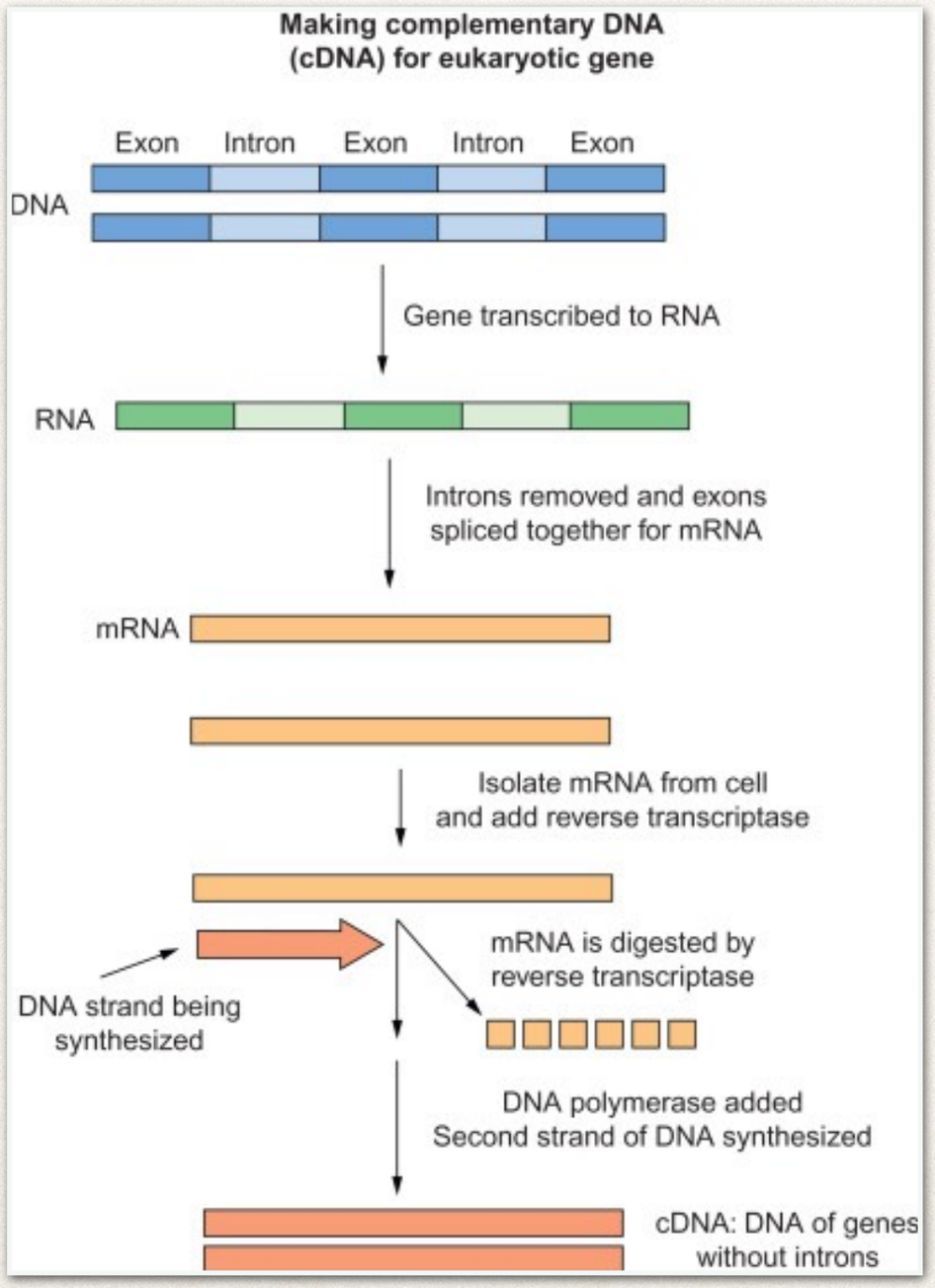


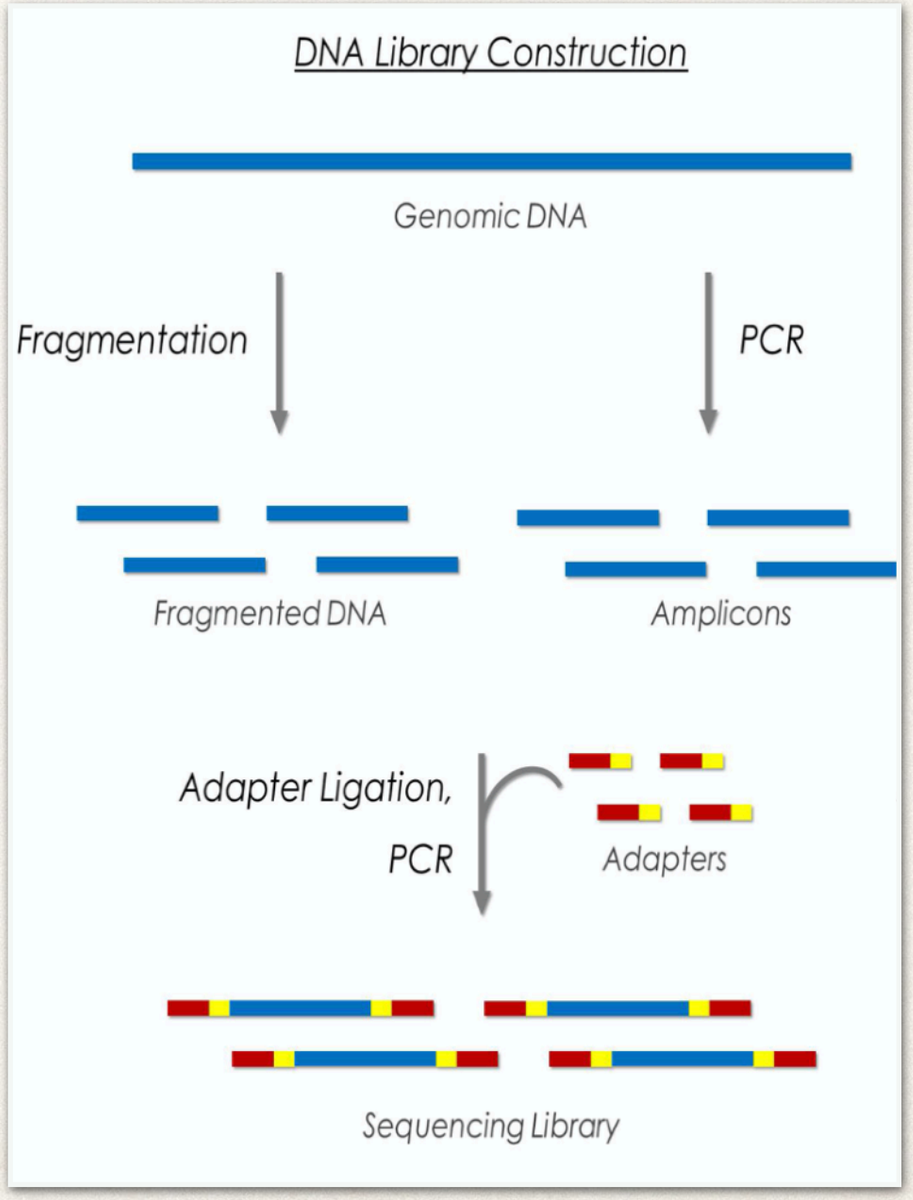
http://www.3402bioinformaticsgroup.com/service/

SGS: Illumina sequencing

https://www.youtube.com/watch?v=fCd6B5HRaZ8&t=54s

SGS: Illumina sequencing





Modified from https://www.biocompare.com/Molecular-Biology/9187-Next-Generation-Sequencing/

https://www.sciencedirect.com/topics/pharmacology-toxicology-and-pharmaceutical-science/complementary-dna

SGS features

- Allowed cost-effective and rapid sequencing of many genomes
- Very accurate
- High throughput
- Polymerase-mediated
- Requiers PCR amplification
- Short reads

- ⊗ Determination of complex genomic regions
- ⊗ Isoform detection
- ⊗ Methylation detection

Third-Generation Sequencing (TGS)

- Technologies capable of **sequencing single molecules** in real time without amplification

- These technologies allow to produce **reads far longer** than SGS, each spanning several to hundreds kbps.

PacBio

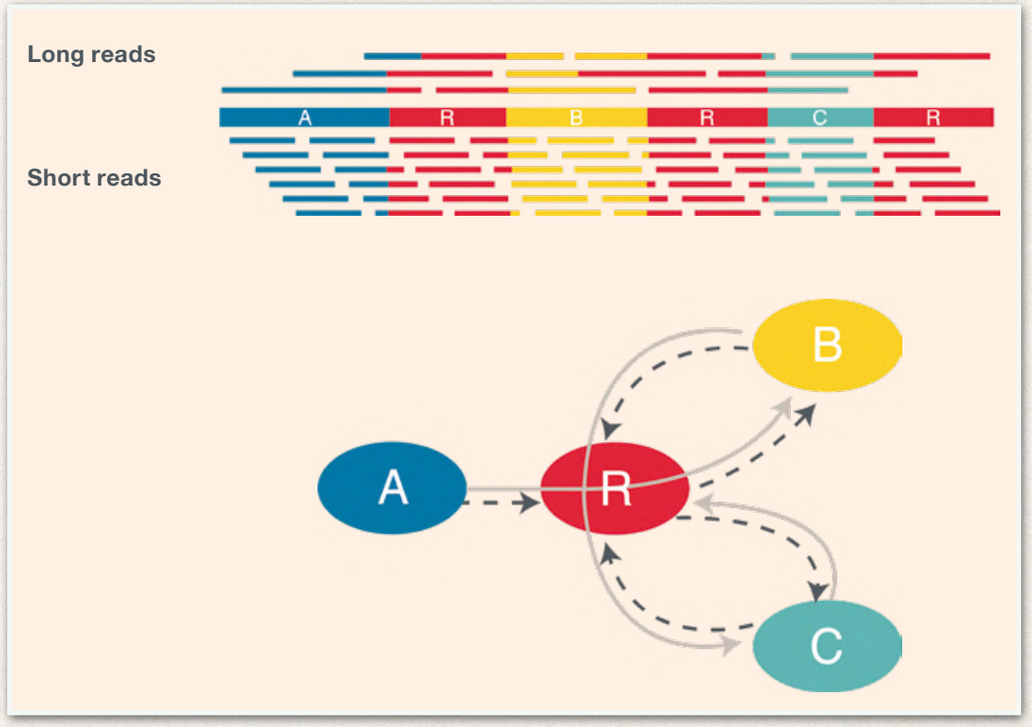
Oxford Nanopore

Third-Generation Sequencing (TGS)

1 - Ease of assembly

- Long read sequencing technology offers simplified and less ambiguous genome assembly

- Whole-genome assembly — solving the puzzle

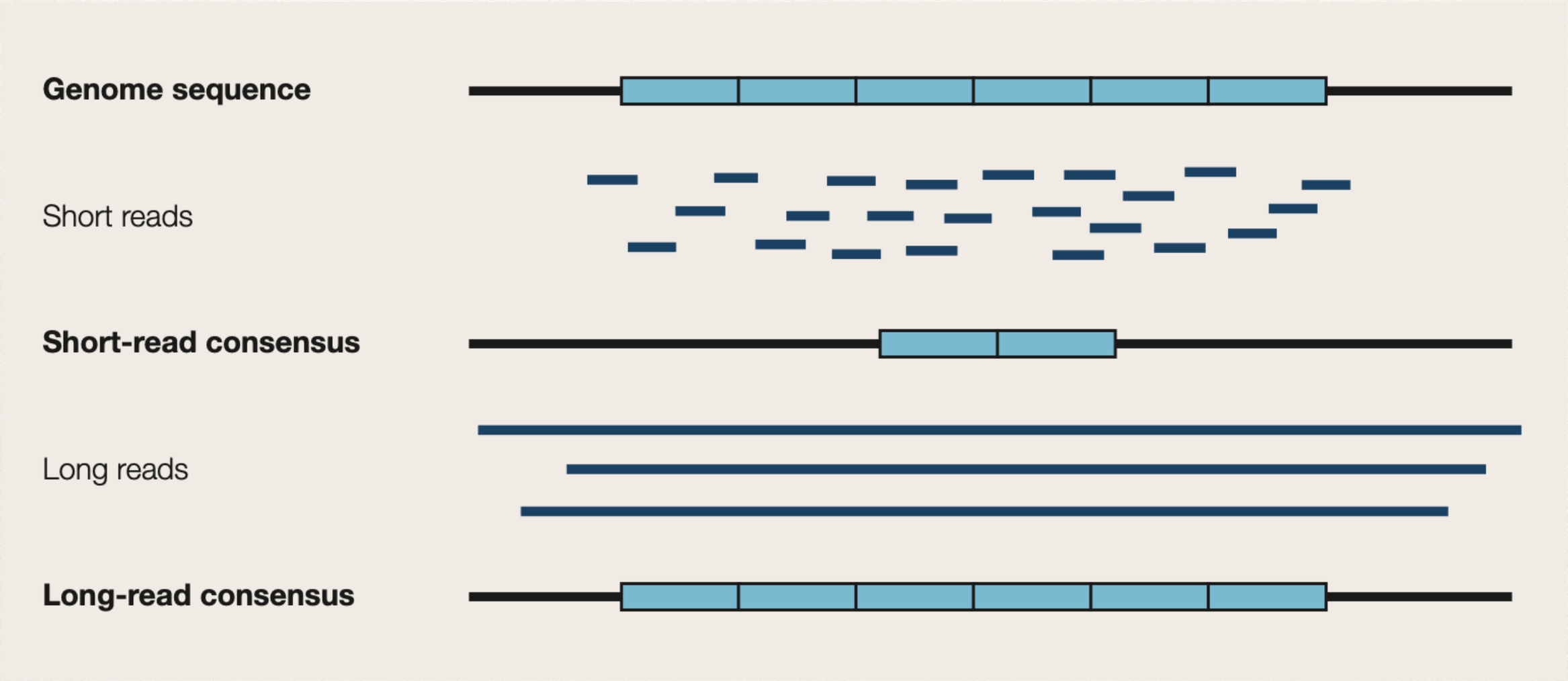


Long reads (solid arrows) have greater overlap with other reads than is provided by short reads (dashed arrows), allowing more accurate assemblies, especially in repeat regions (R).

https://nanoporetech.com/sites/default/files/s3/white-papers/genome-assembly-white-paper-oct19.pdf

Third-Generation Sequencing (TGS)

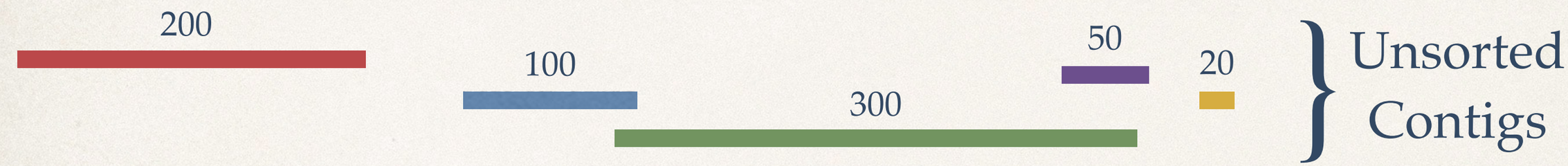
2 - Facility to span repetitive genomic regions & large structural variation

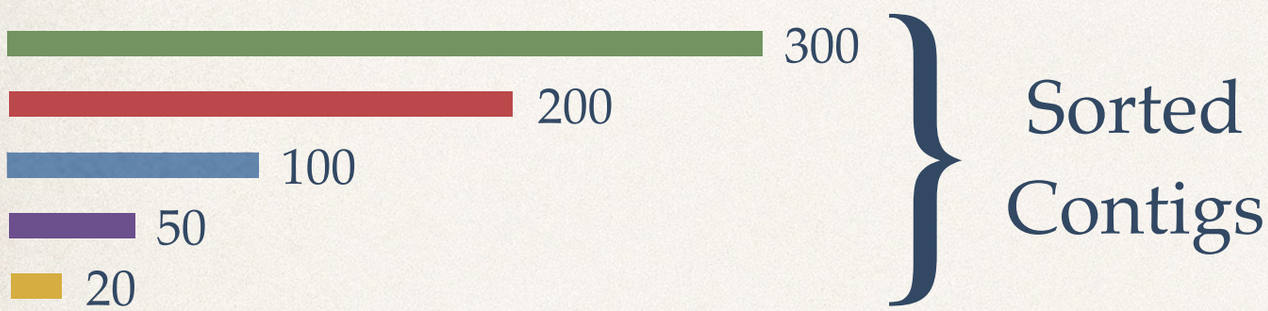


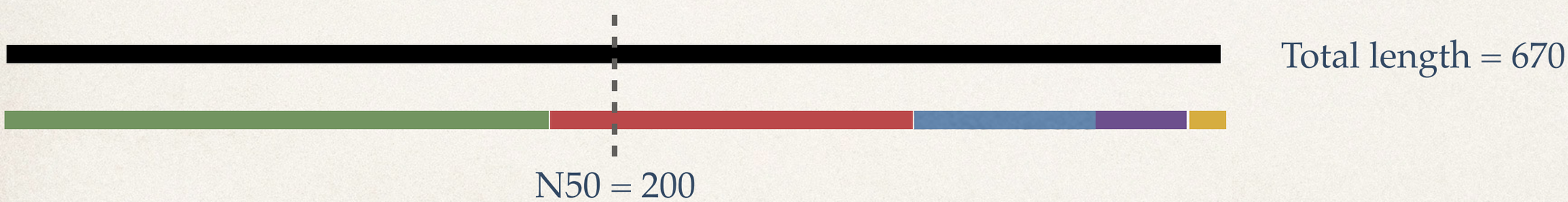
https://nanoporetech.com/sites/default/files/s3/white-papers/genome-assembly-white-paper-oct19.pdf

Third-Generation Sequencing (TGS)

3 - Assembly quality - N50



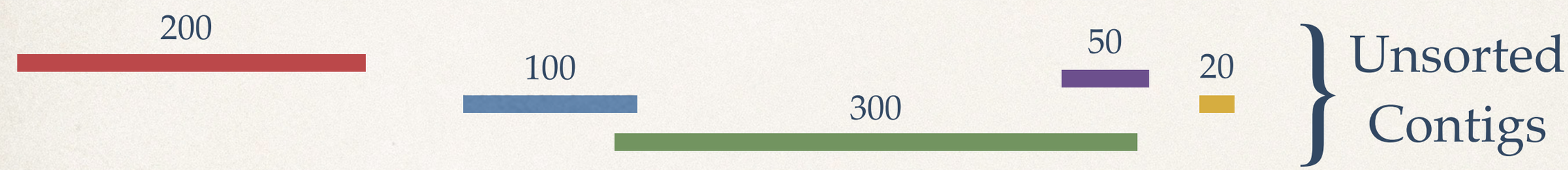


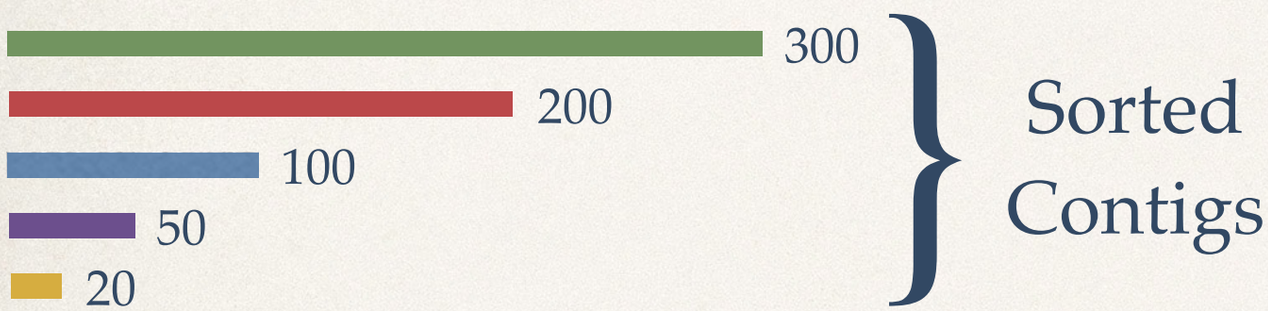


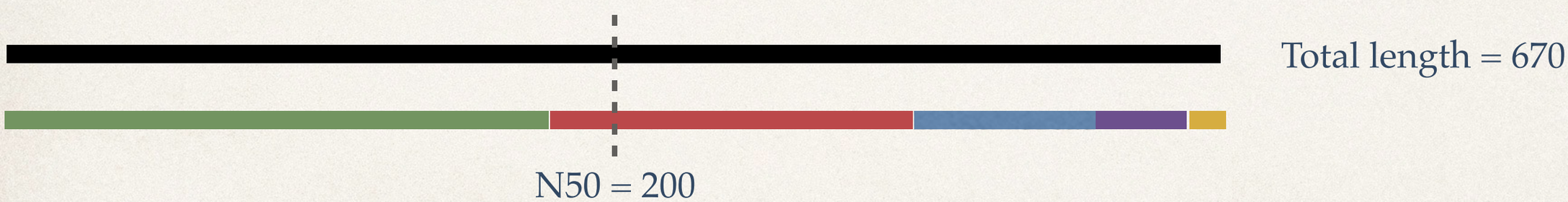
Given a set of contigs, the N50 is defined as the sequence length of the shortest contig at 50% of the total genome length.

Third-Generation Sequencing (TGS)

3 - Assembly quality - N50





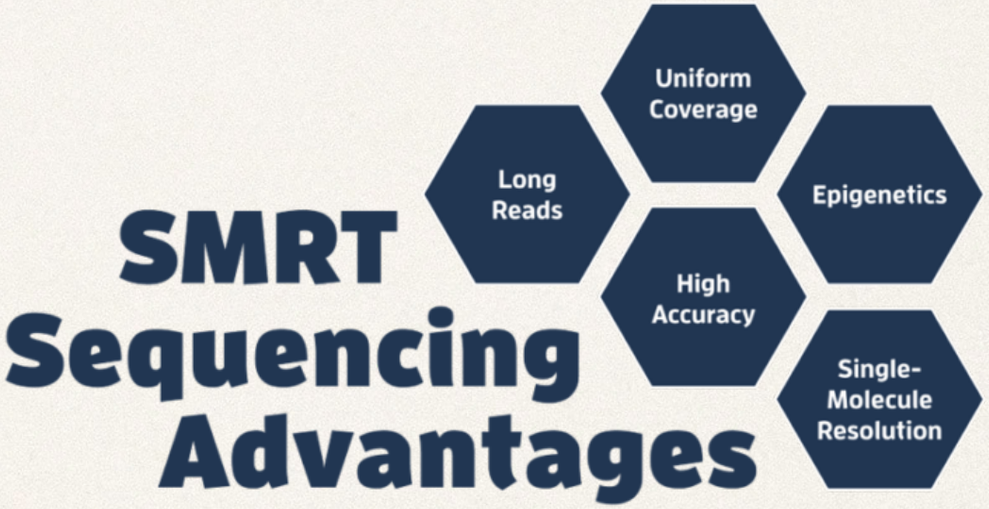


The higher the contig N50 value, the more contiguous the assembly.

PacBio Sequencing

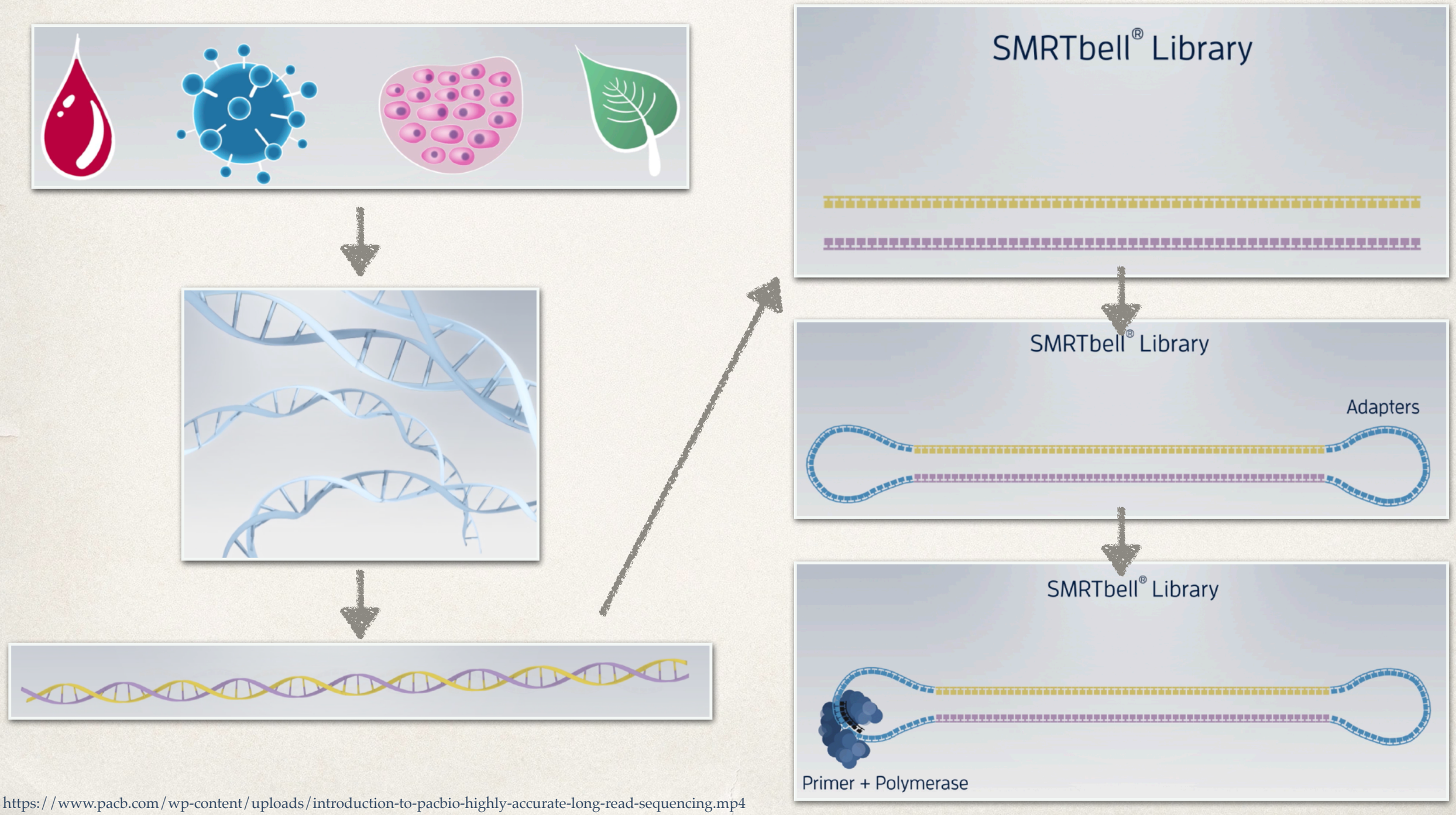
- Single Molecule, Real-Time (SMRT) Sequencing is the core technology powering these **long-read sequencing** platforms.

Sequel II System



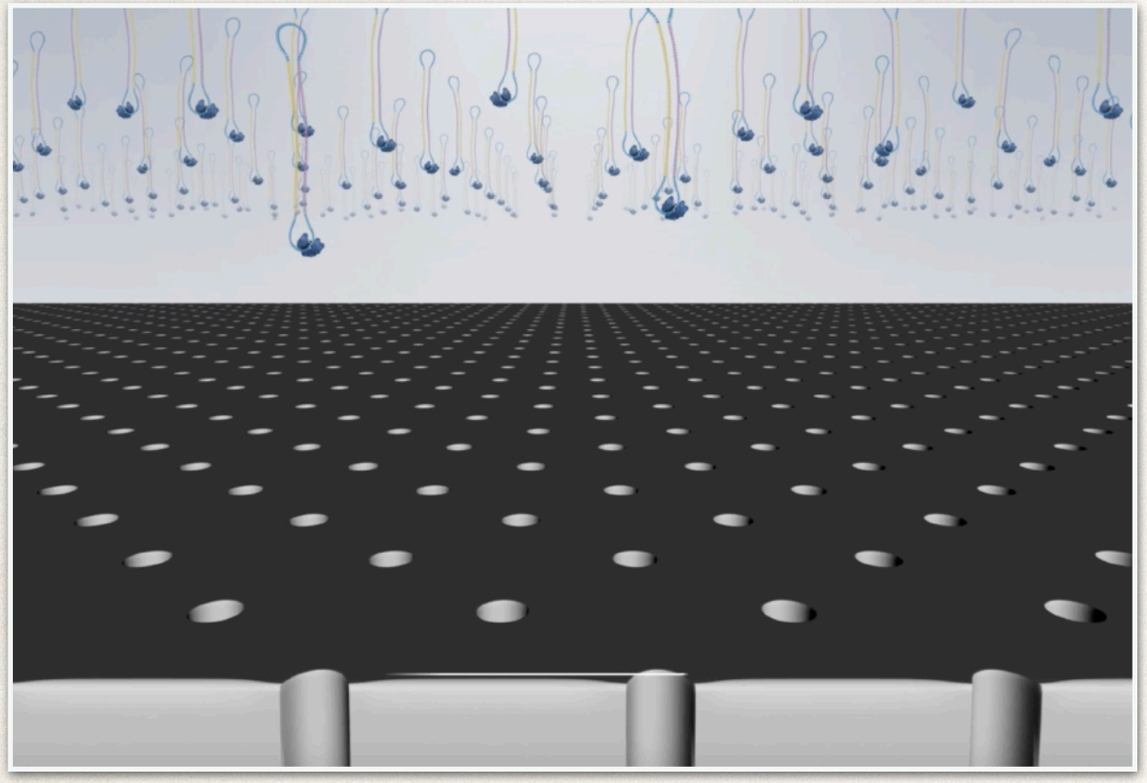
- SMRT uses the natural process of DNA replication

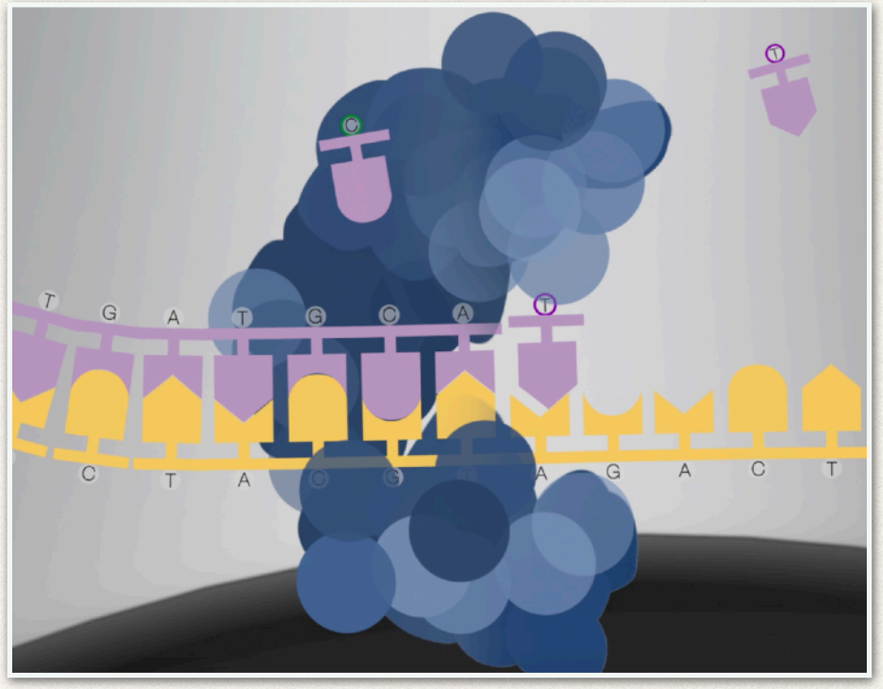
PacBio SMRT Sequencing

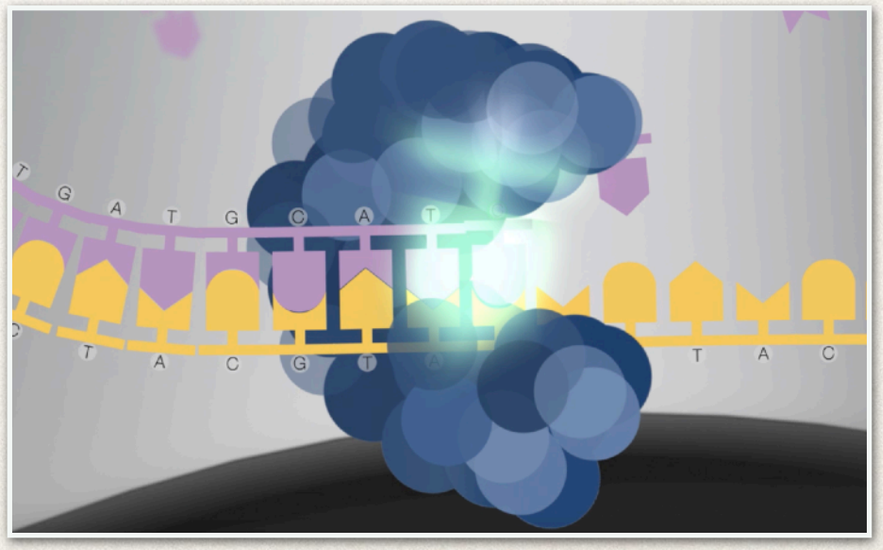


https://www.pacb.com/wp-content/uploads/introduction-to-pacbio-highly-accurate-long-read-sequencing.mp4

PacBio SMRT Sequencing

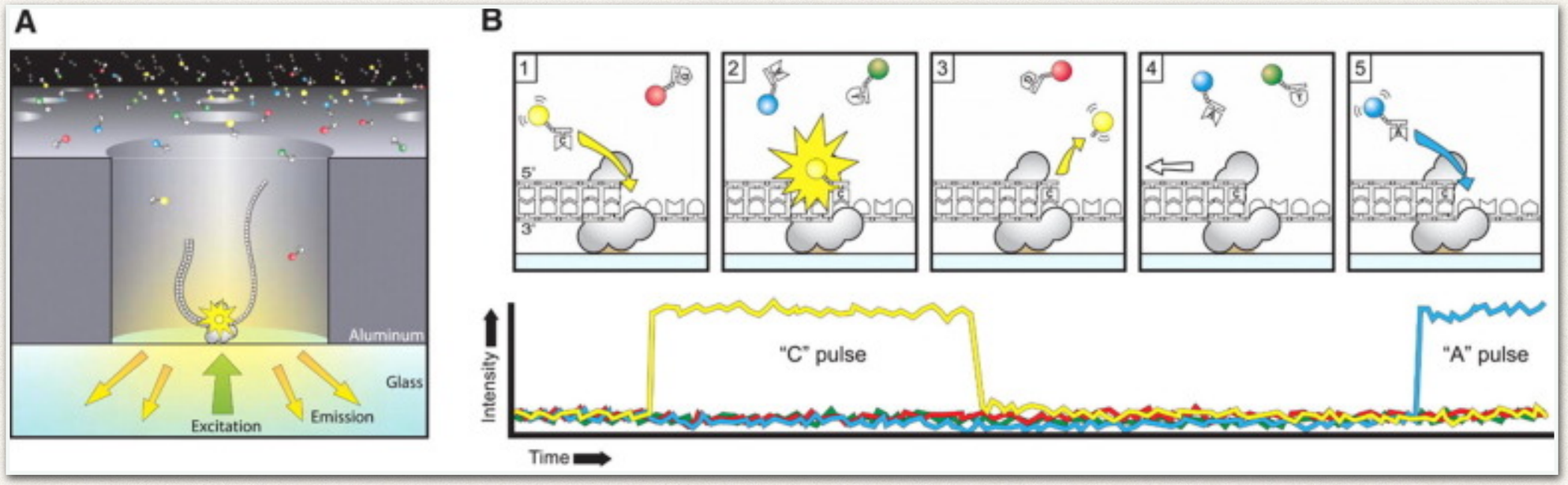






https://www.pacb.com/wp-content/uploads/introduction-to-pacbio-highly-accurate-long-read-sequencing.mp4

PacBio SMRT Sequencing

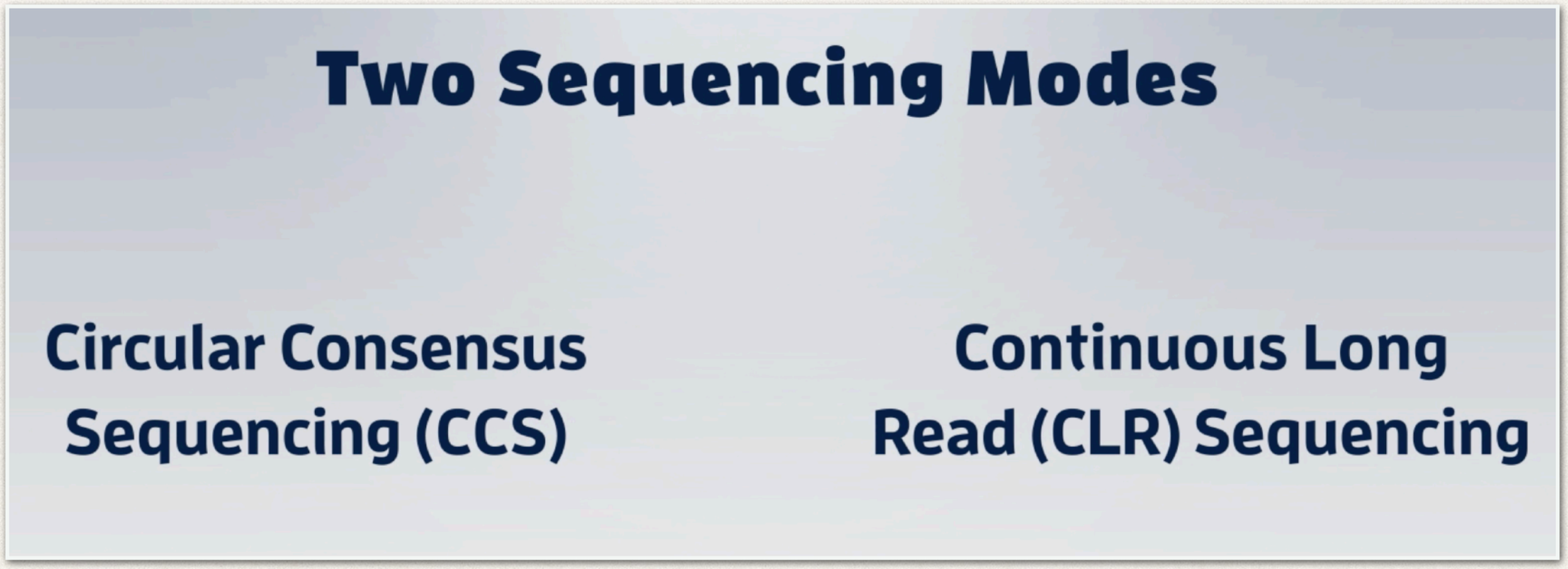


- A SMRTbell diffuses into a well, and the adaptor binds to a polymerase immobilized at the bottom. Each SMRT cell contains 150,000 well.
- Each of the four nucleotides is labeled with a different fluorescent dye so that they have distinct emission spectrums.

PacBio SMRT Sequencing

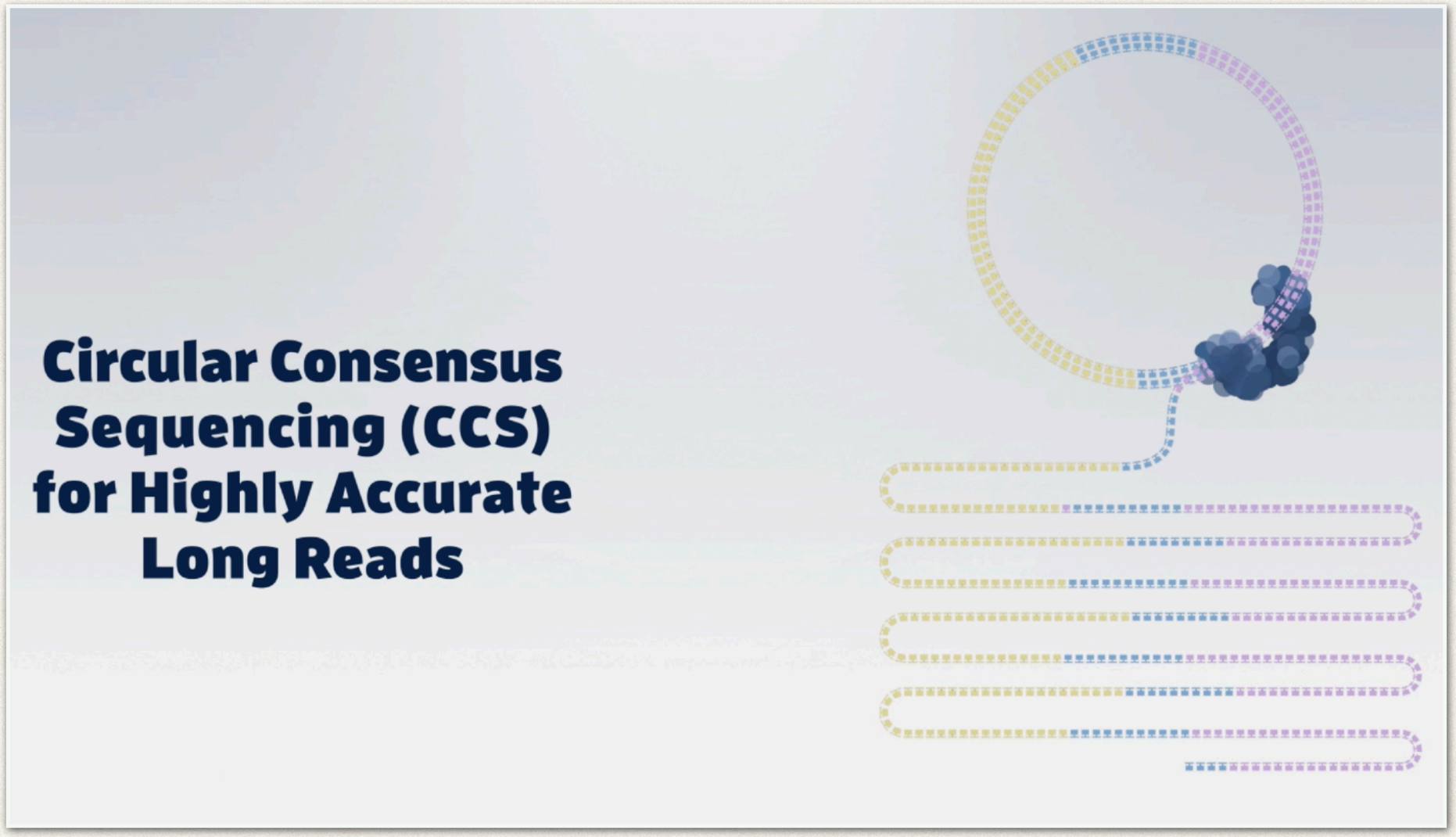
https://www.youtube.com/watch?v=_lD8JyAbwEo&t=10s

PacBio SMRT Sequencing



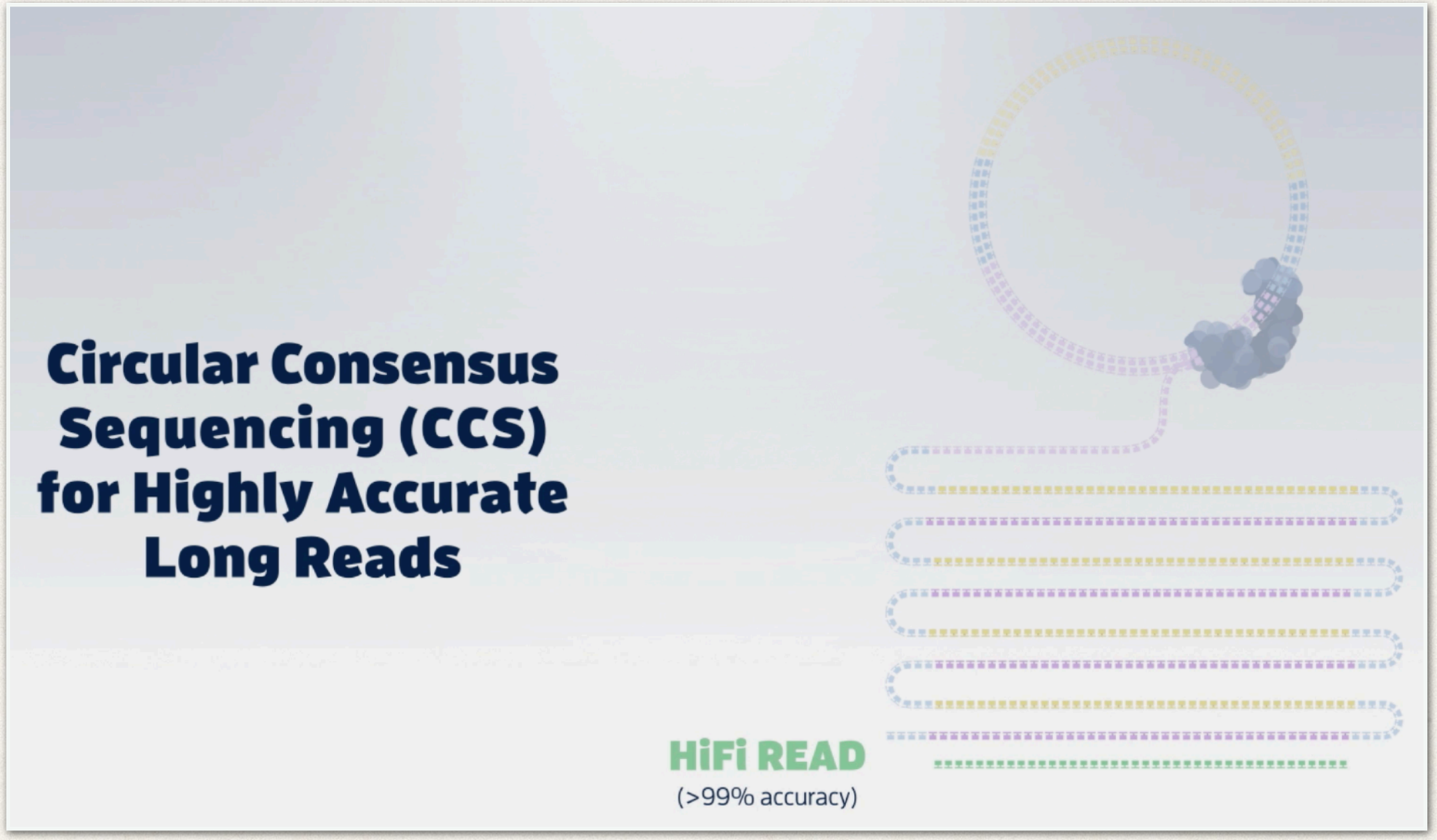
https://www.pacb.com/wp-content/uploads/introduction-to-pacbio-highly-accurate-long-read-sequencing.mp4

PacBio SMRT Sequencing



https://www.pacb.com/wp-content/uploads/introduction-to-pacbio-highly-accurate-long-read-sequencing.mp4

PacBio SMRT Sequencing

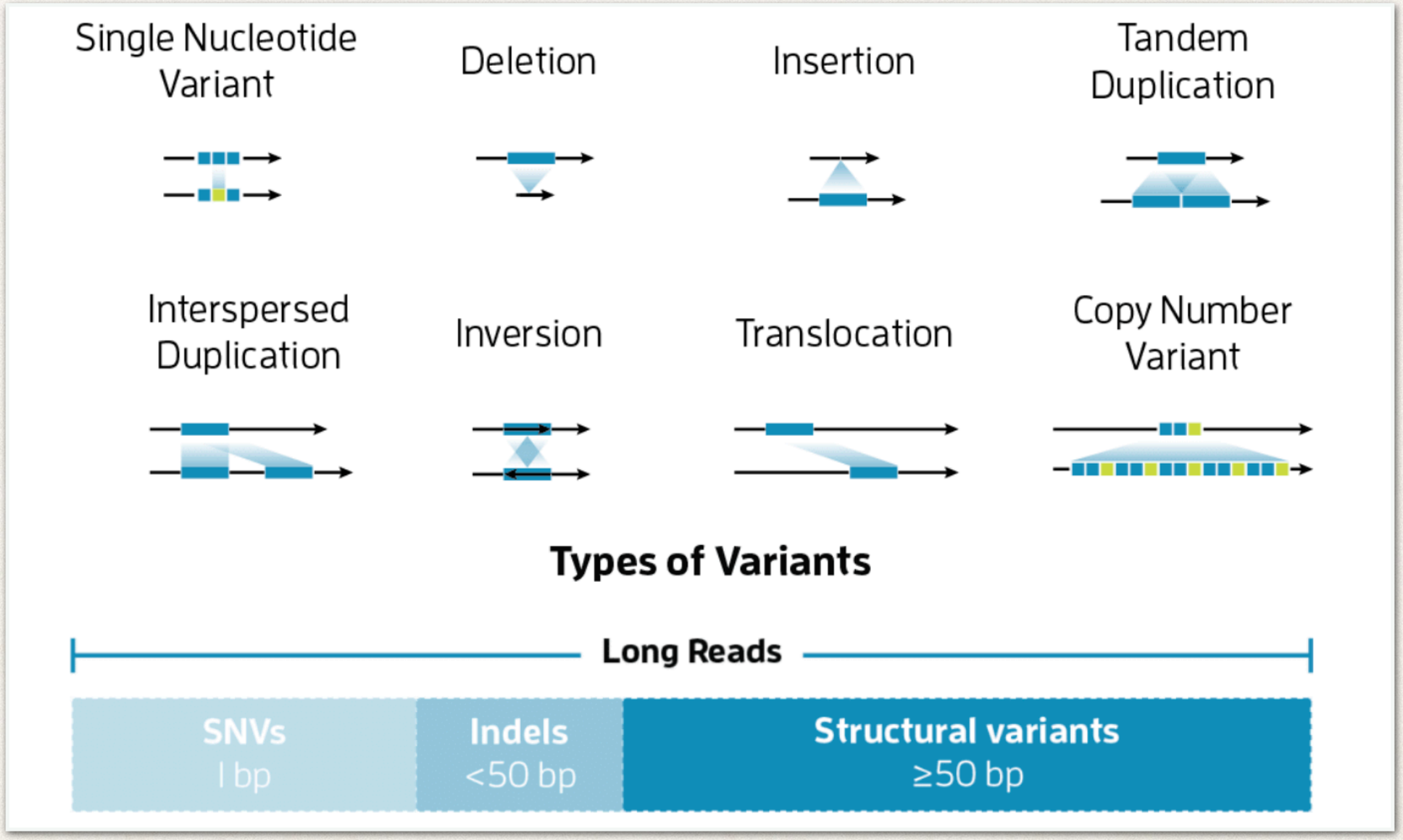


https://www.pacb.com/wp-content/uploads/introduction-to-pacbio-highly-accurate-long-read-sequencing.mp4

SMRT Sequencing Applications

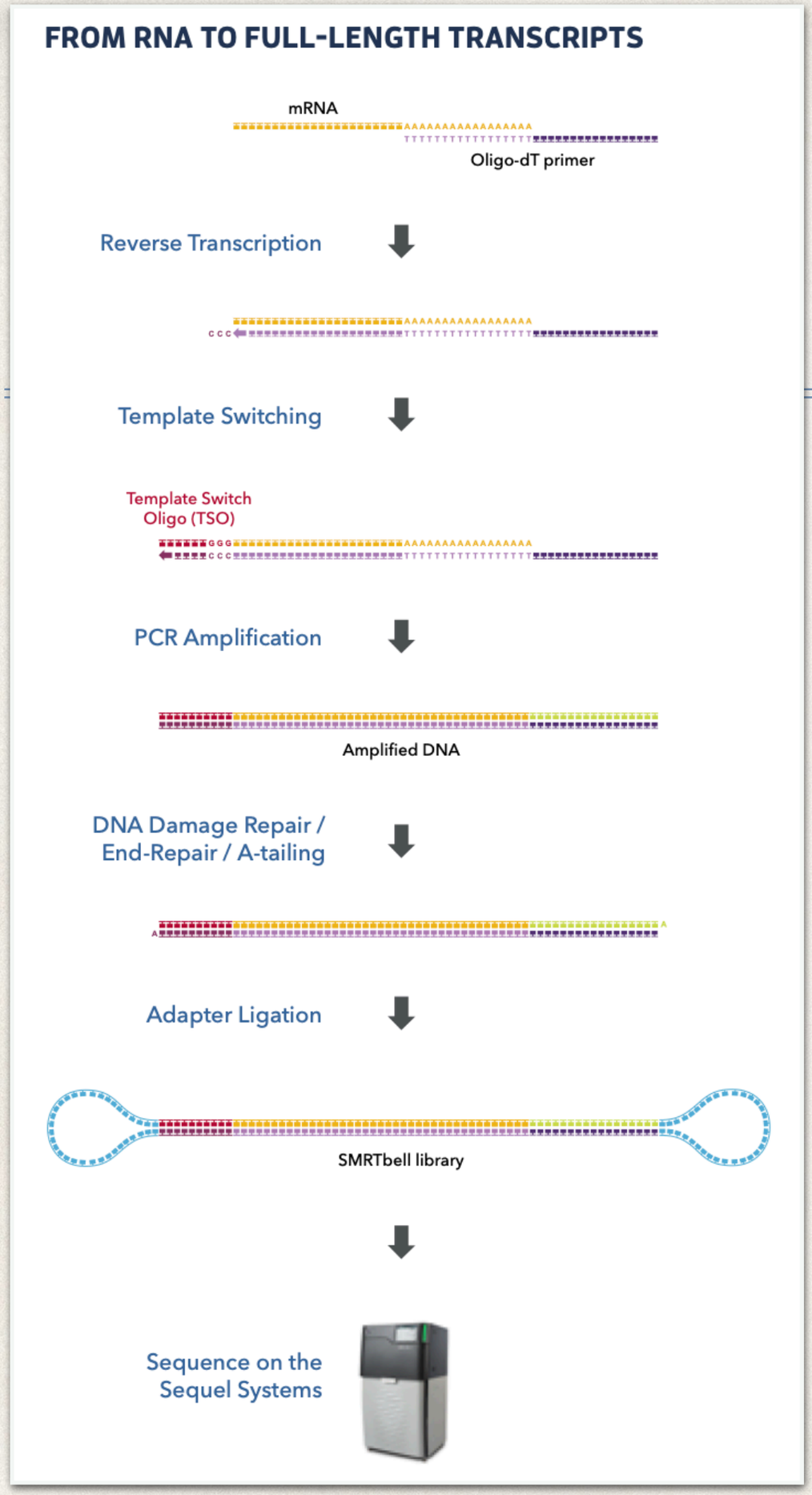
https://www.pacb.com/wp-content/uploads/introduction-to-pacbio-highly-accurate-long-read-sequencing.mp4

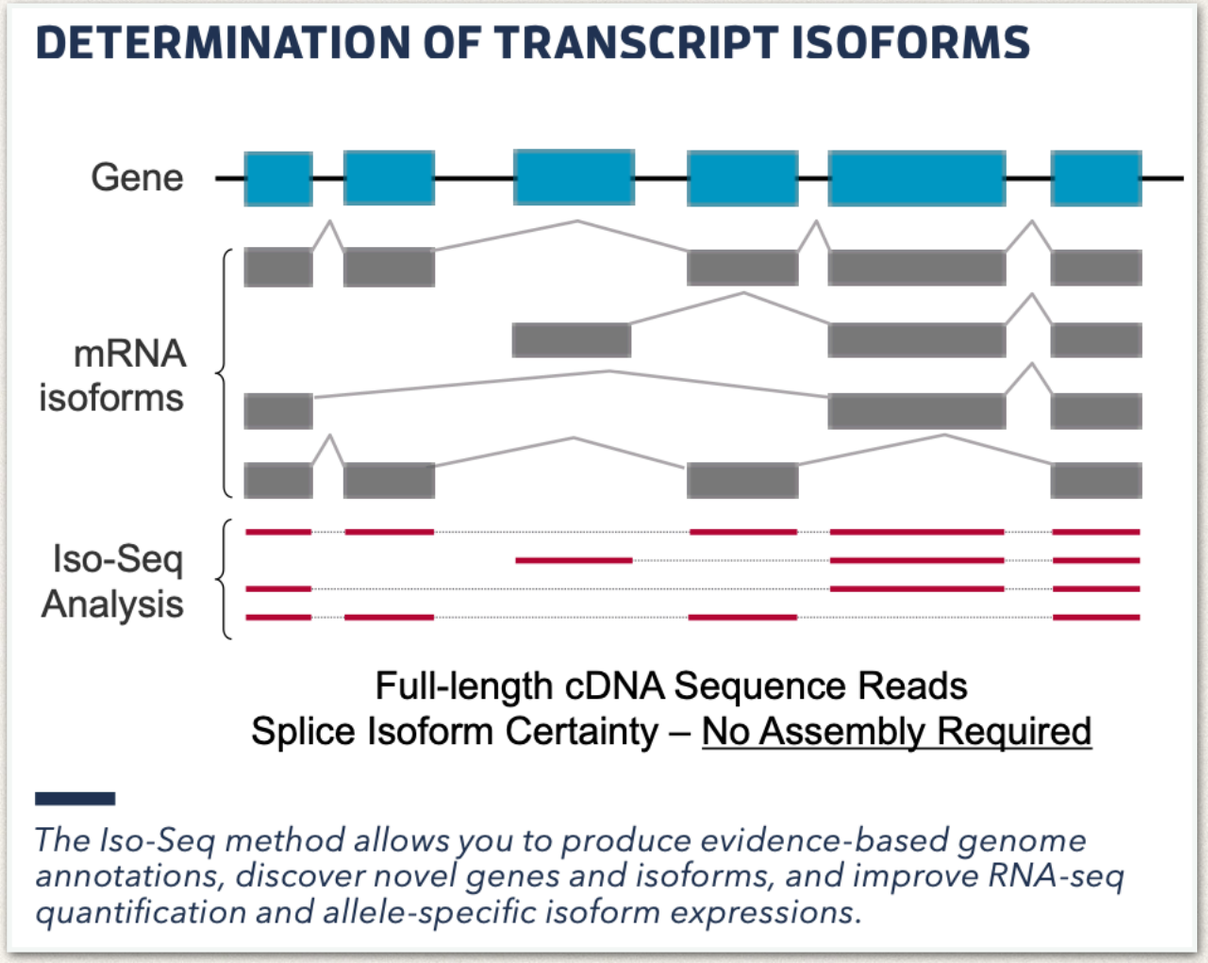
Variant detection



- PacBio long-read sequencing enables comprehensive detection of all variant types.

RNA sequencing

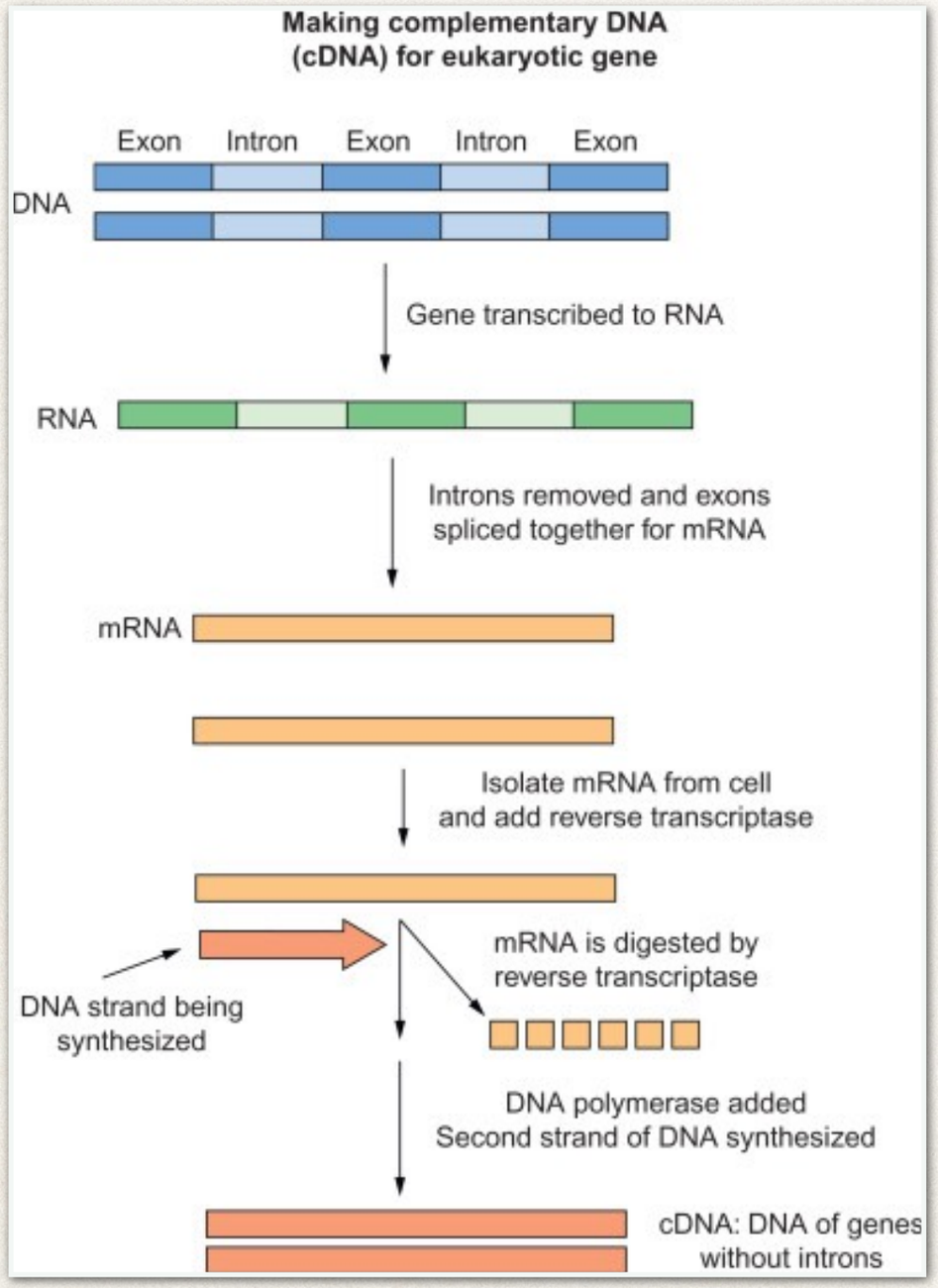


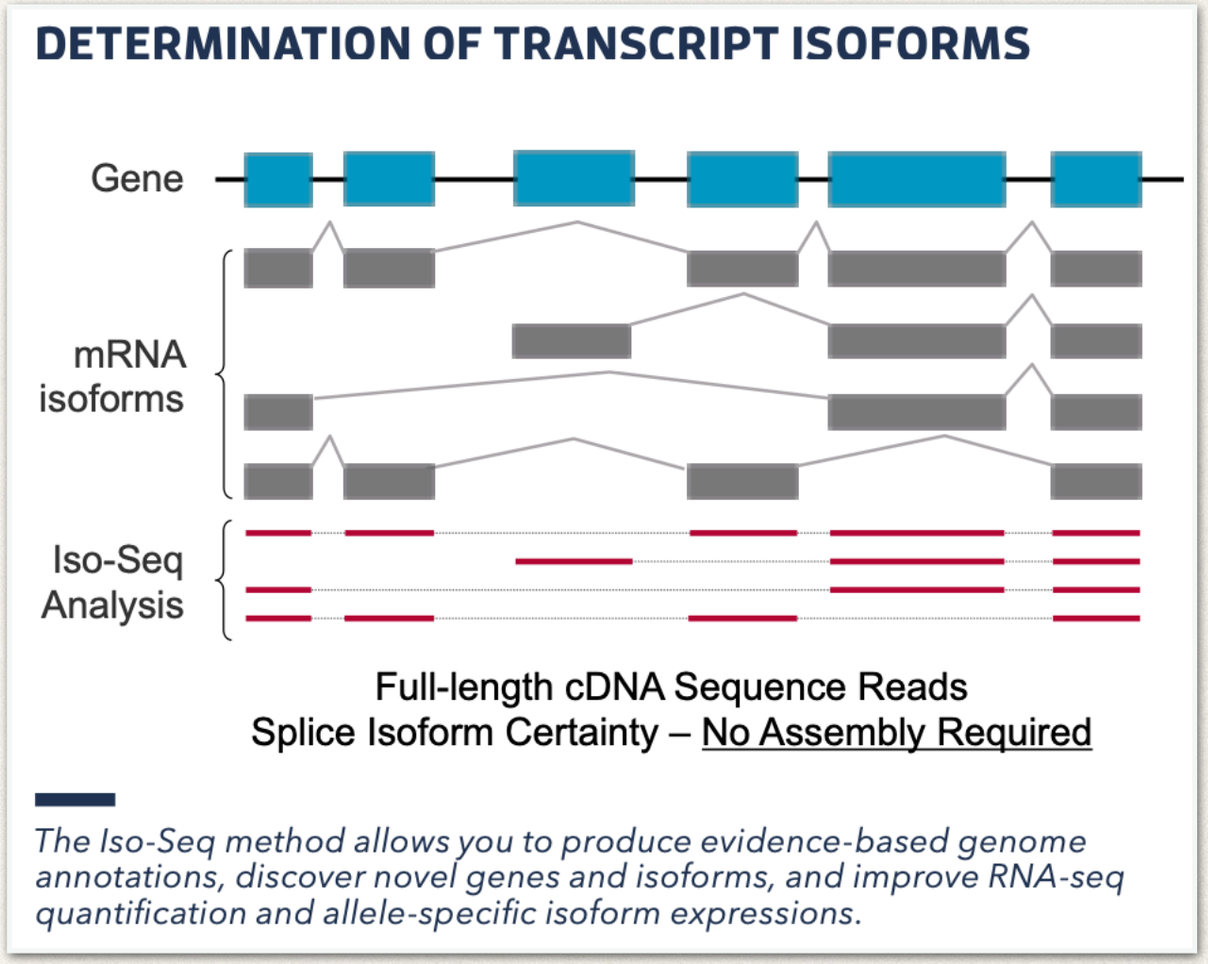


- The Iso-Seq® method allows users to generate full-length cDNA sequences up to 10 kb in length — with no assembly required — to confidently characterize full-length transcript isoforms.

https://www.pacb.com/wp-content/uploads/Application-Brief-RNA-sequencing-Best-Practices.pdf

RNA sequencing

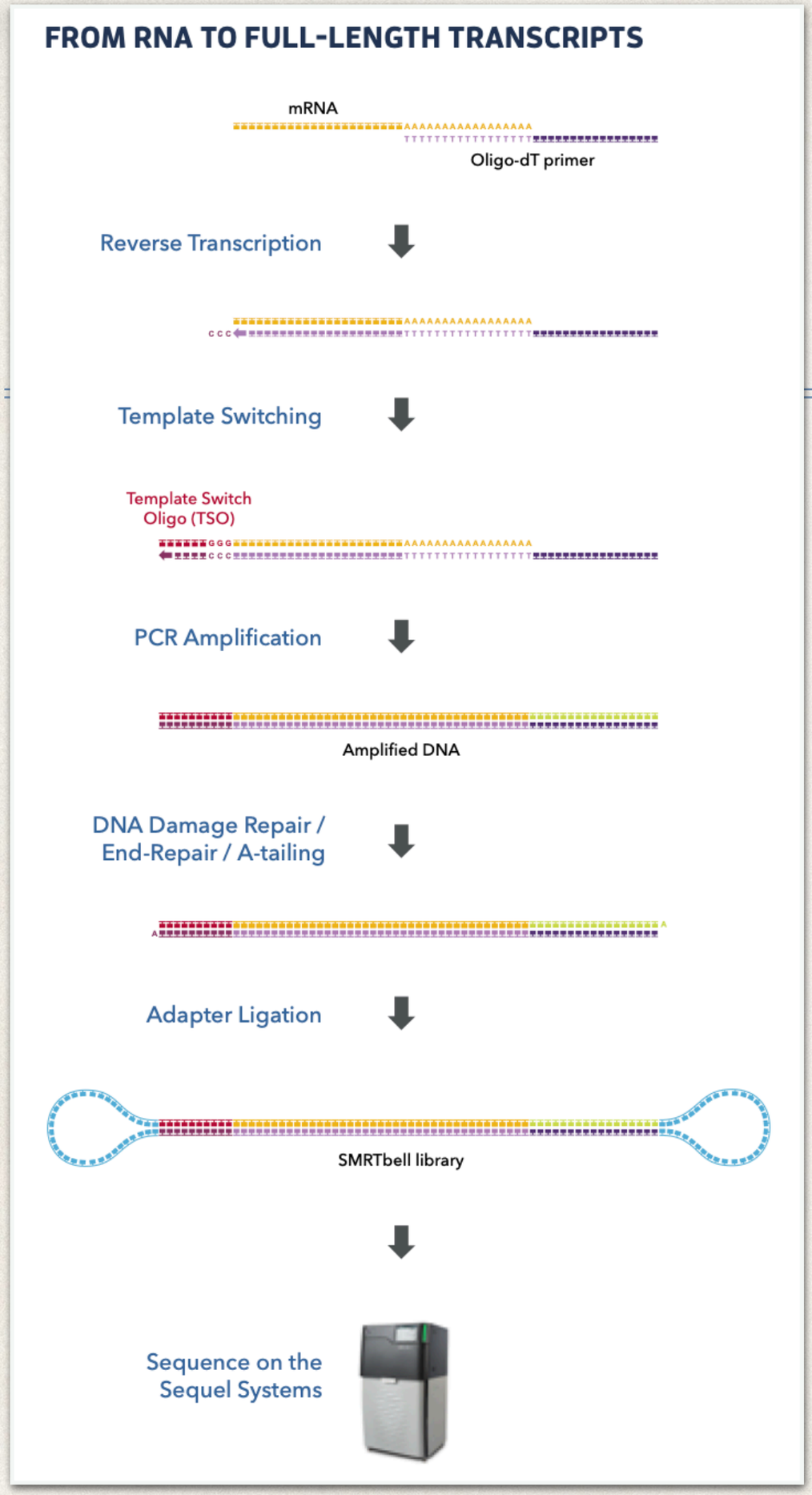


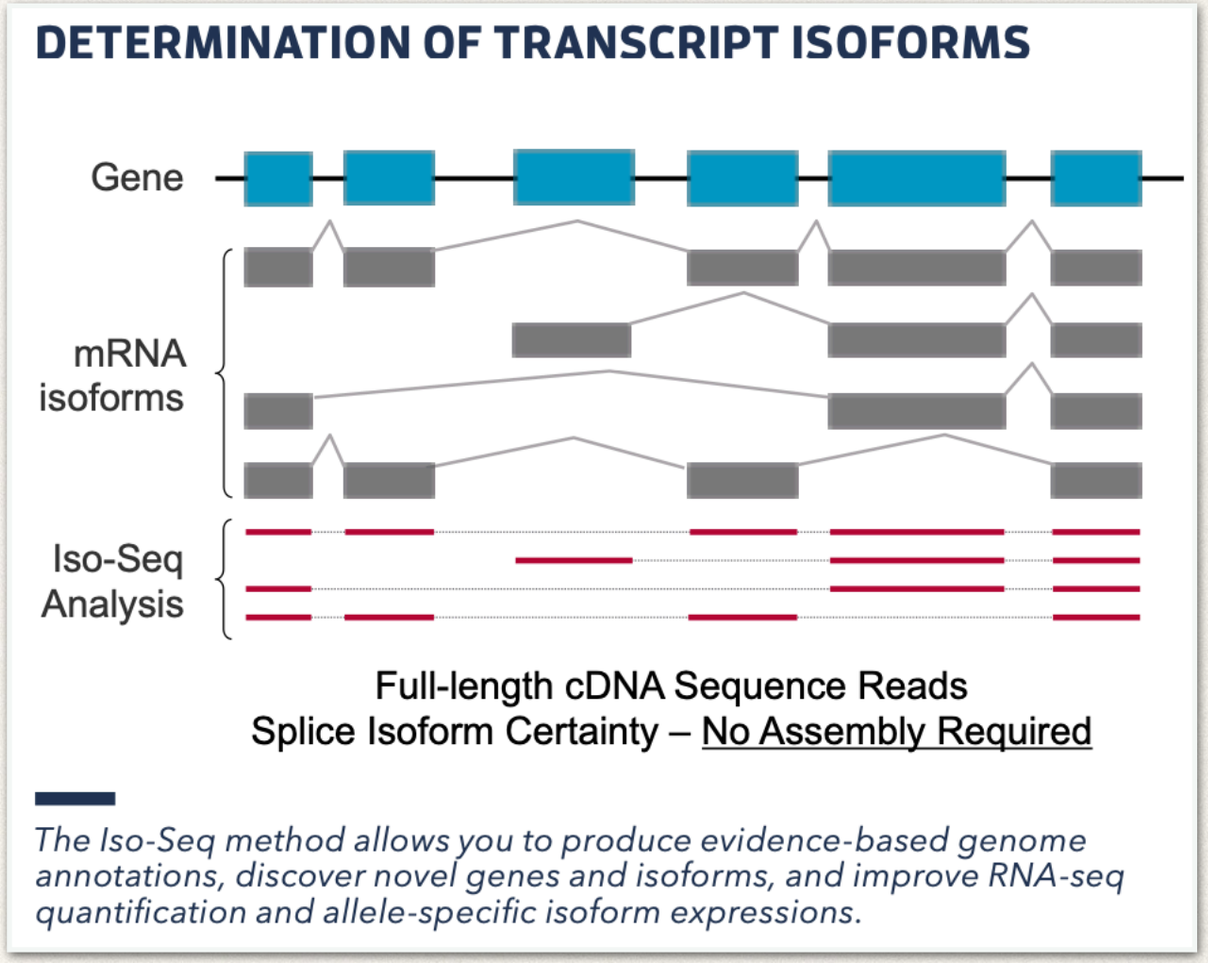


- The Iso-Seq® method allows users to generate full-length cDNA sequences up to 10 kb in length — with no assembly required — to confidently characterize full-length transcript isoforms.

https://www.sciencedirect.com/topics/pharmacology-toxicology-and-pharmaceutical-science/complementary-dna

RNA sequencing



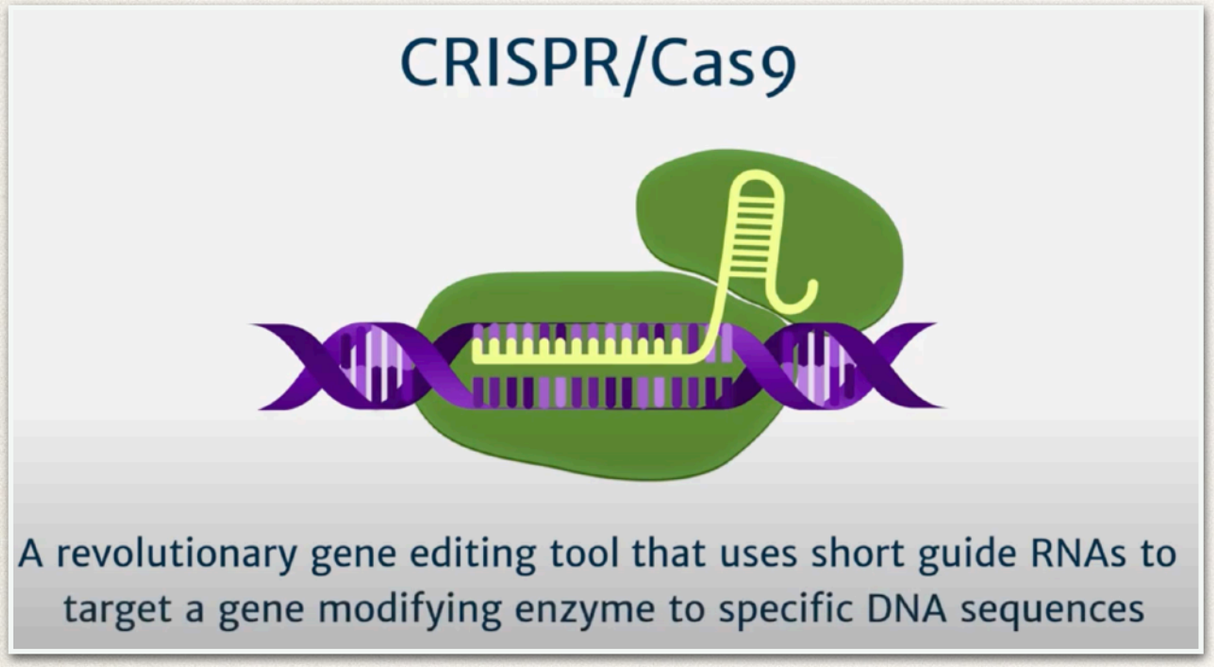


- The Iso-Seq® method allows users to generate full-length cDNA sequences up to 10 kb in length — with no assembly required — to confidently characterize full-length transcript isoforms.

https://www.pacb.com/wp-content/uploads/Application-Brief-RNA-sequencing-Best-Practices.pdf

Targeted Sequencing

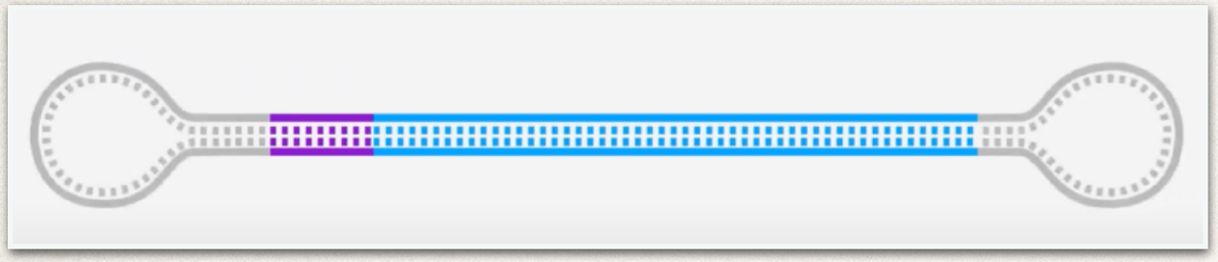
No amplification targeted sequencing

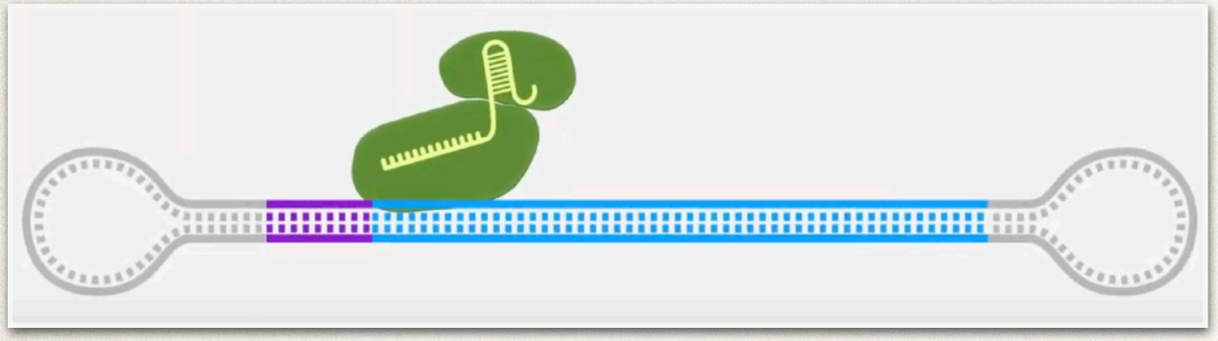


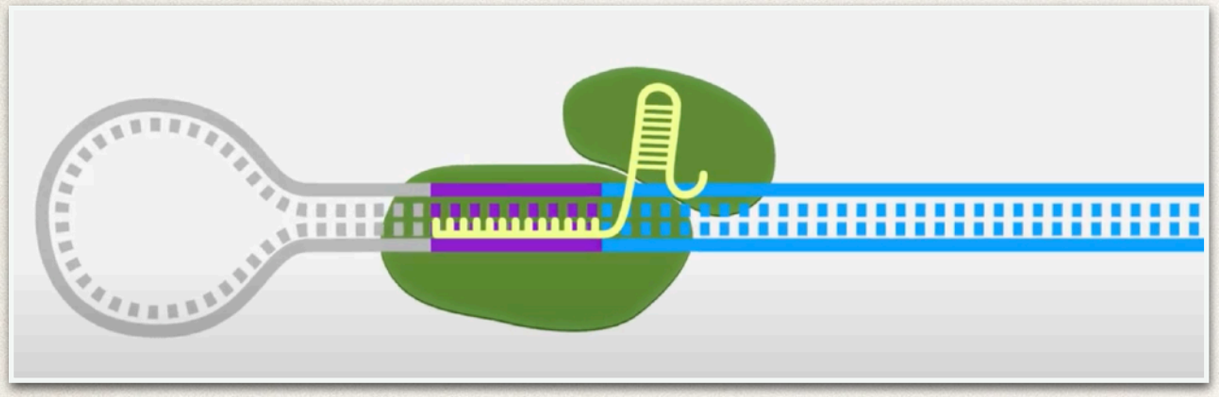
https://www.pacb.com/applications/targeted-sequencing/no-amp-targeted-sequencing/

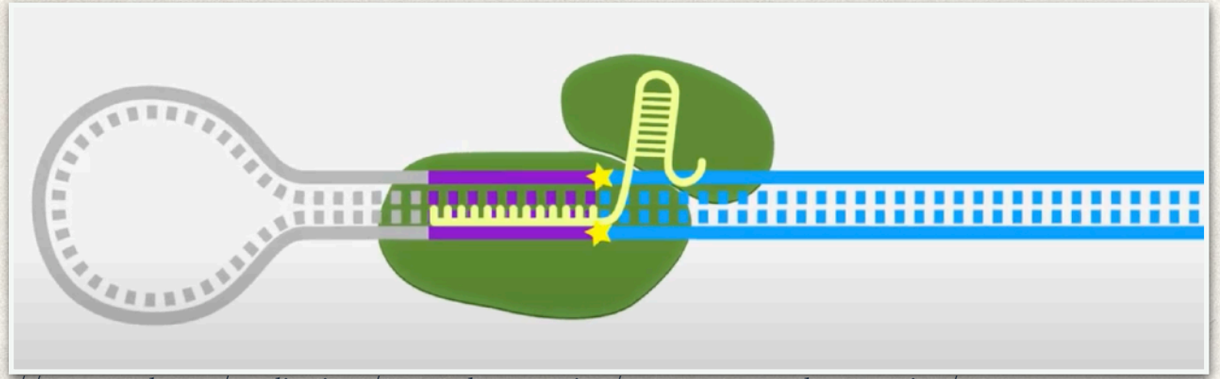
Targeted Sequencing

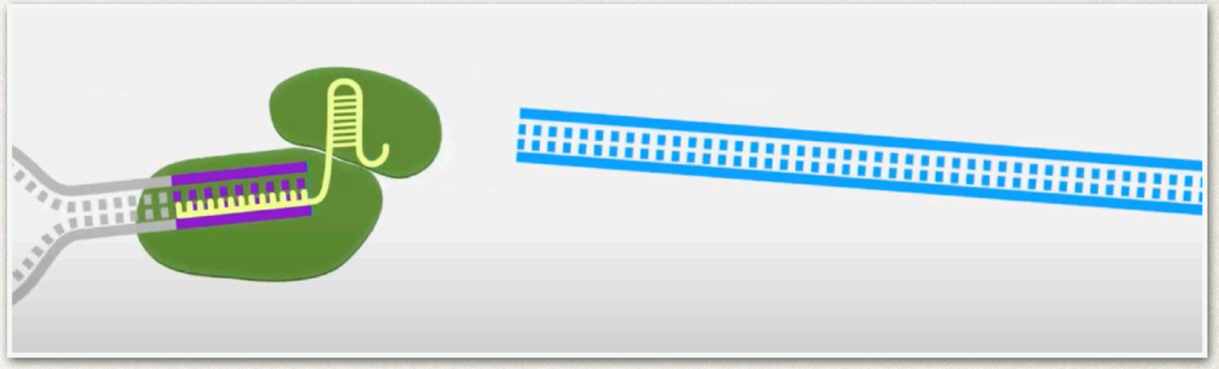
No amplification targeted sequencing

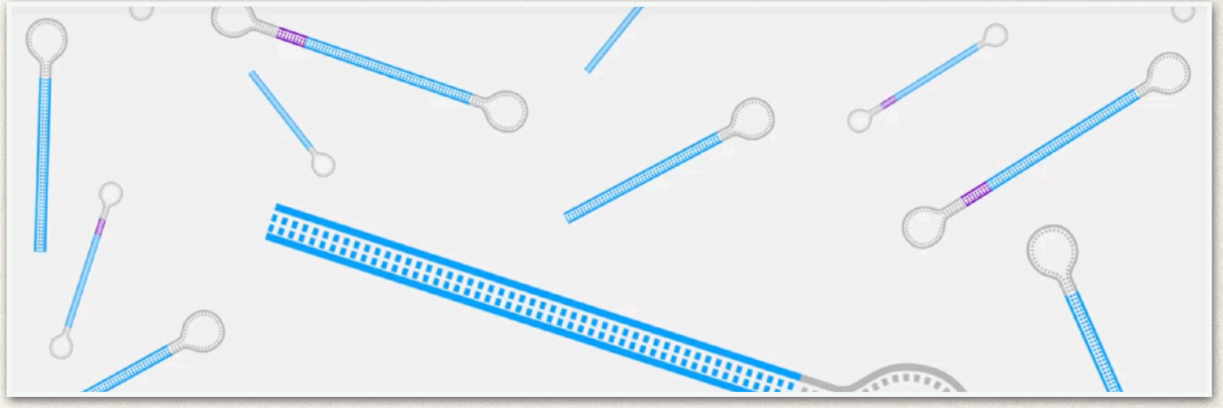


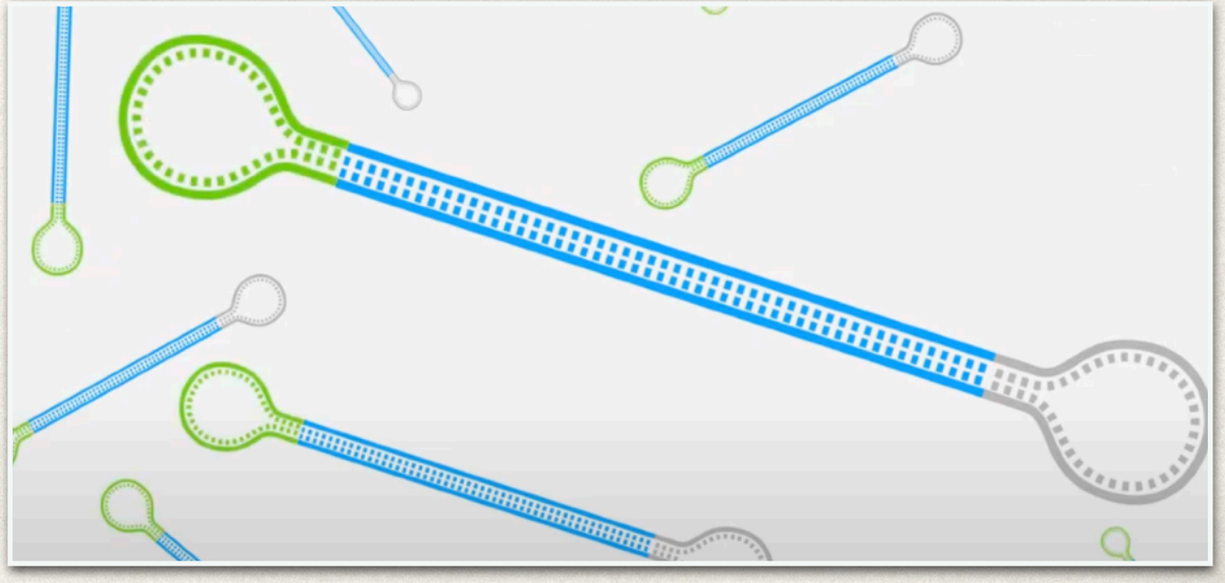








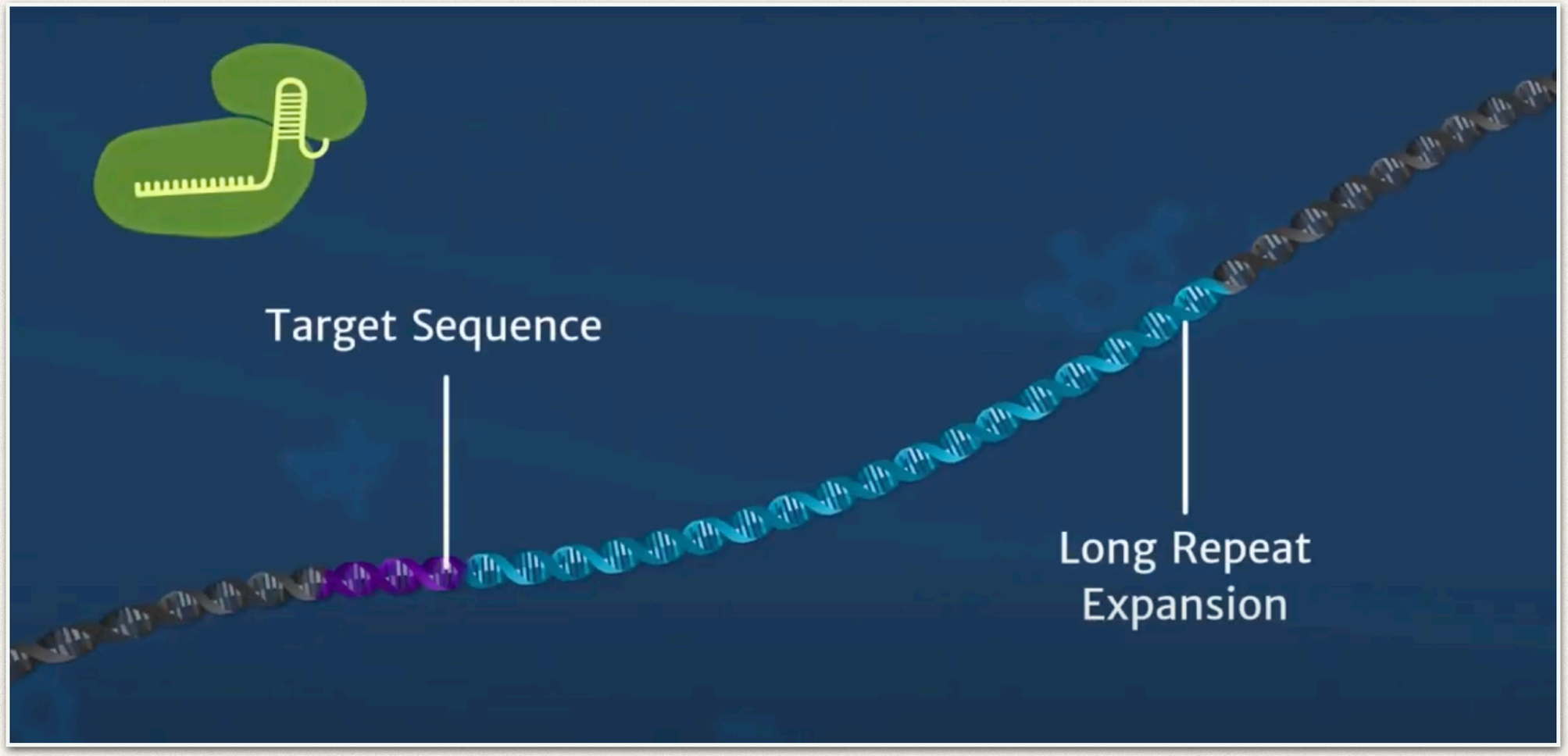


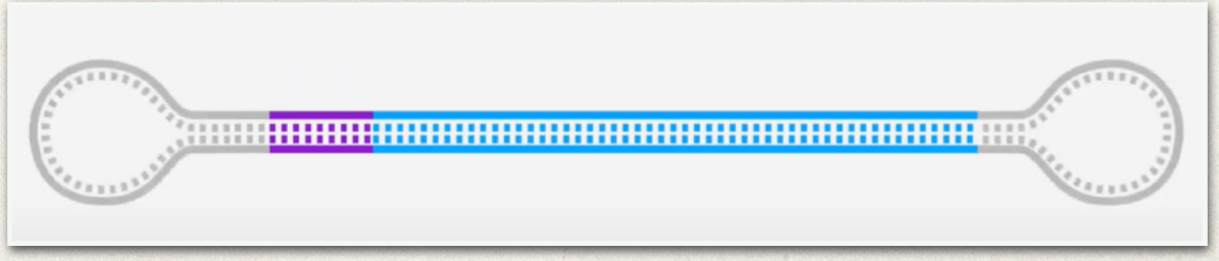


https://www.pacb.com/applications/targeted-sequencing/no-amp-targeted-sequencing/

Targeted Sequencing

No amplification targeted sequencing

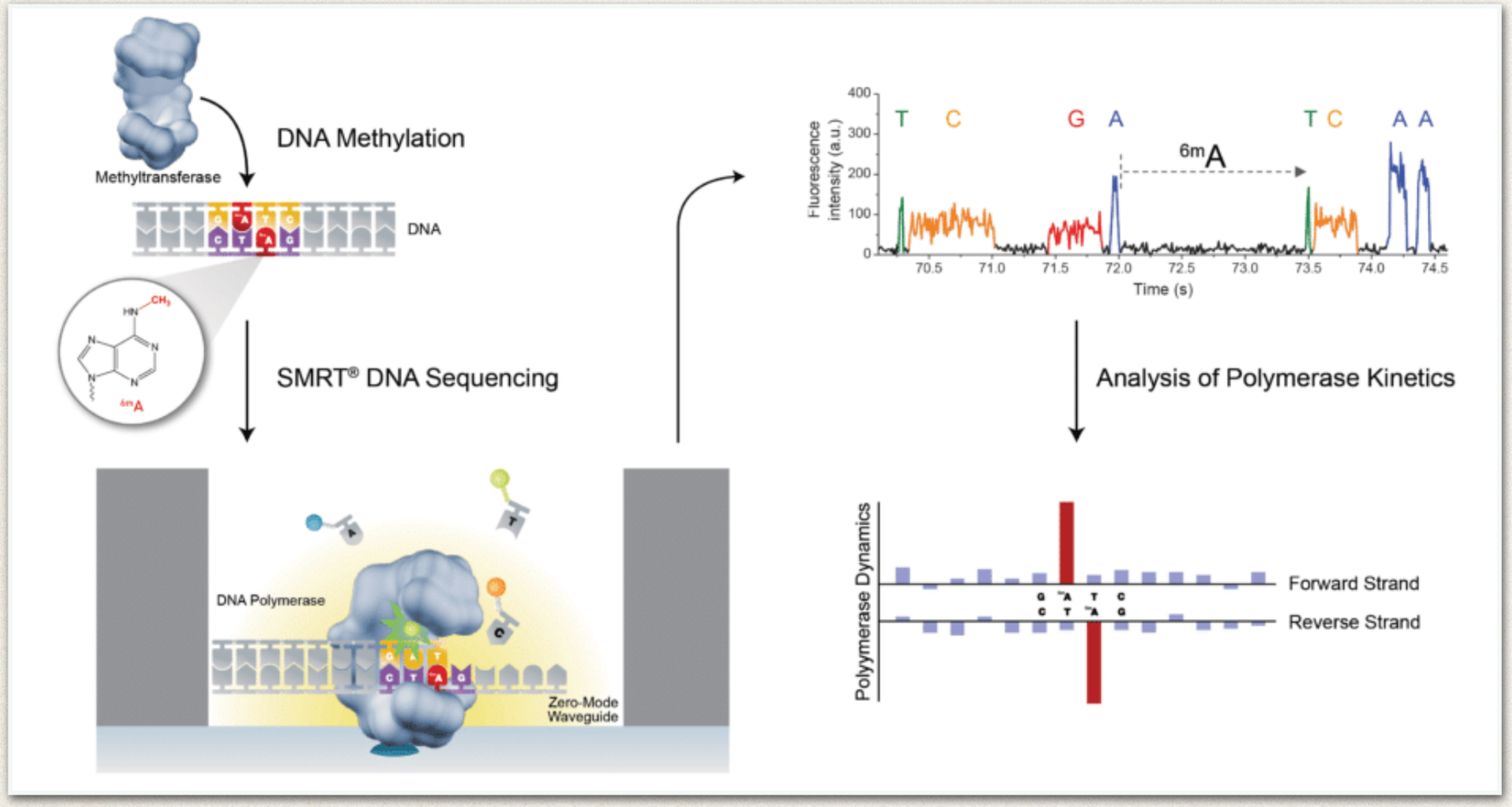


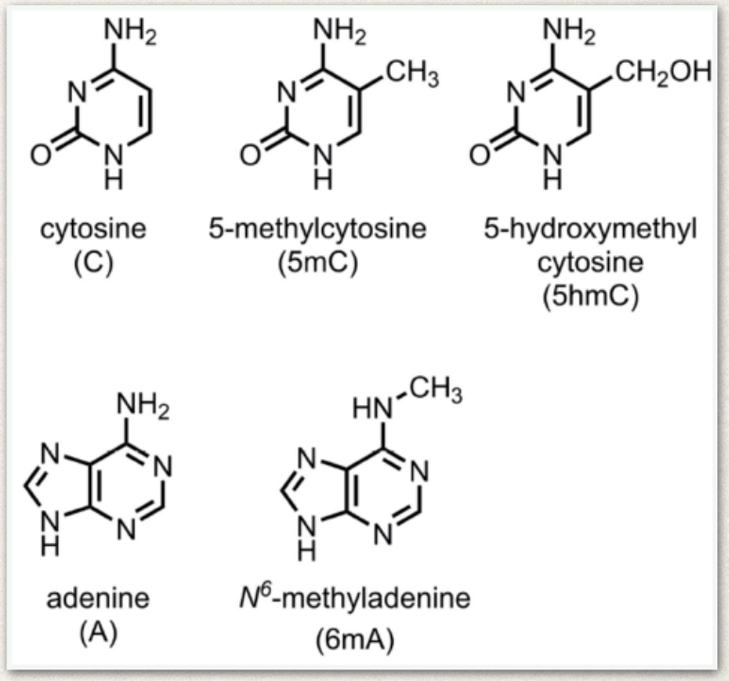


https://www.pacb.com/applications/targeted-sequencing/no-amp-targeted-sequencing/

Epigenetics

- SMRT Sequencing directly detects epigenetic modifications by measuring kinetic variation during base incorporation.





By capturing these modifications simultaneously with sequence data, this method eliminates the need for special sample preparation and additional sequencing.

https://www.pacb.com/applications/epigenetics/eukaryotic/
Shi et al., 2017 *Frontiers in Genetics* DOI: 10.3389/fgene.2017.00100

PacBio features

- PacBio improved assembly and determination of complex genomic regions, gene isoform detection, and methylation detection
- Single-molecule sequencing technology
- Generate real-time long read data (~10 Kb)
- Low throughput . There are 150,000 well on a single SMRT cell, each of which can produce one subread or CCS read. Typically, only 35,000–70,000 of the 150,000 wells on a SMRT cell produce successful reads
- High error rate (around 11%–15%)
- Based on fluorescence detection and Sequencing by Synthesis. It worked similar to Illumina sequencers, but without any bridge amplification, thereby avoiding DNA amplification-associated biases.

PacBio features

- PacBio sequencing overcome many of the obstacles faced by SGS via providing longer read lengths, kinetic variation information, and shorter run times

- Yet the technology still has to improve other aspects:

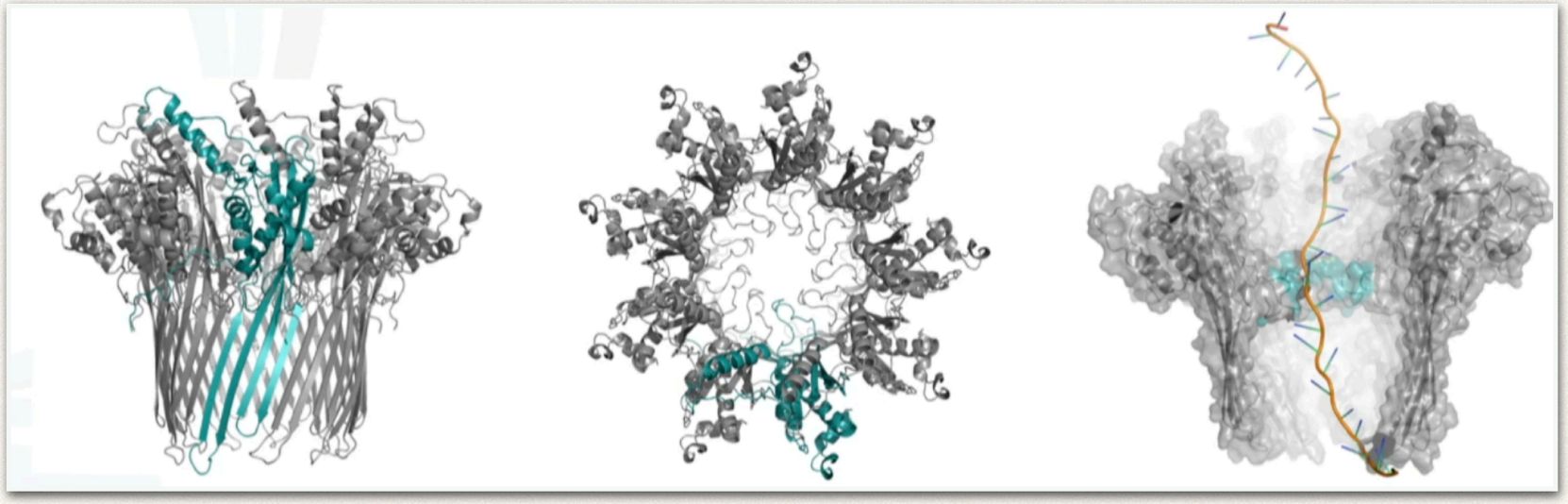
    1. the high error rate of raw single-pass data
    2. better detection of epigenetic modification
    3. read length
    4. sufficient read depth
    5. increase the rate of successfully loading a single polymerase in each well

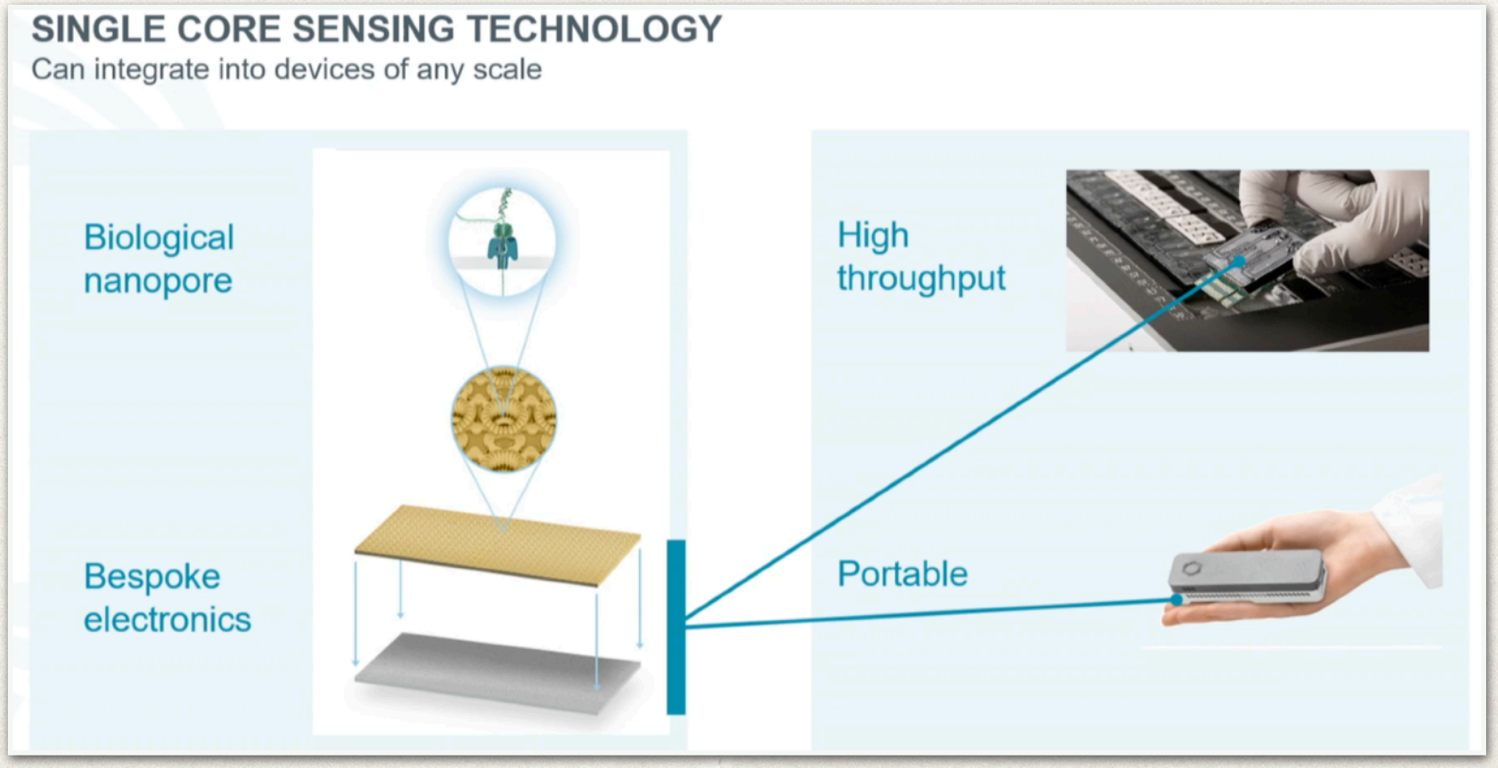
Oxford Nanopore Technologies

- Nanopore sequencing enables direct, real-time analysis of long DNA or RNA fragments.

https://nanoporetech.com

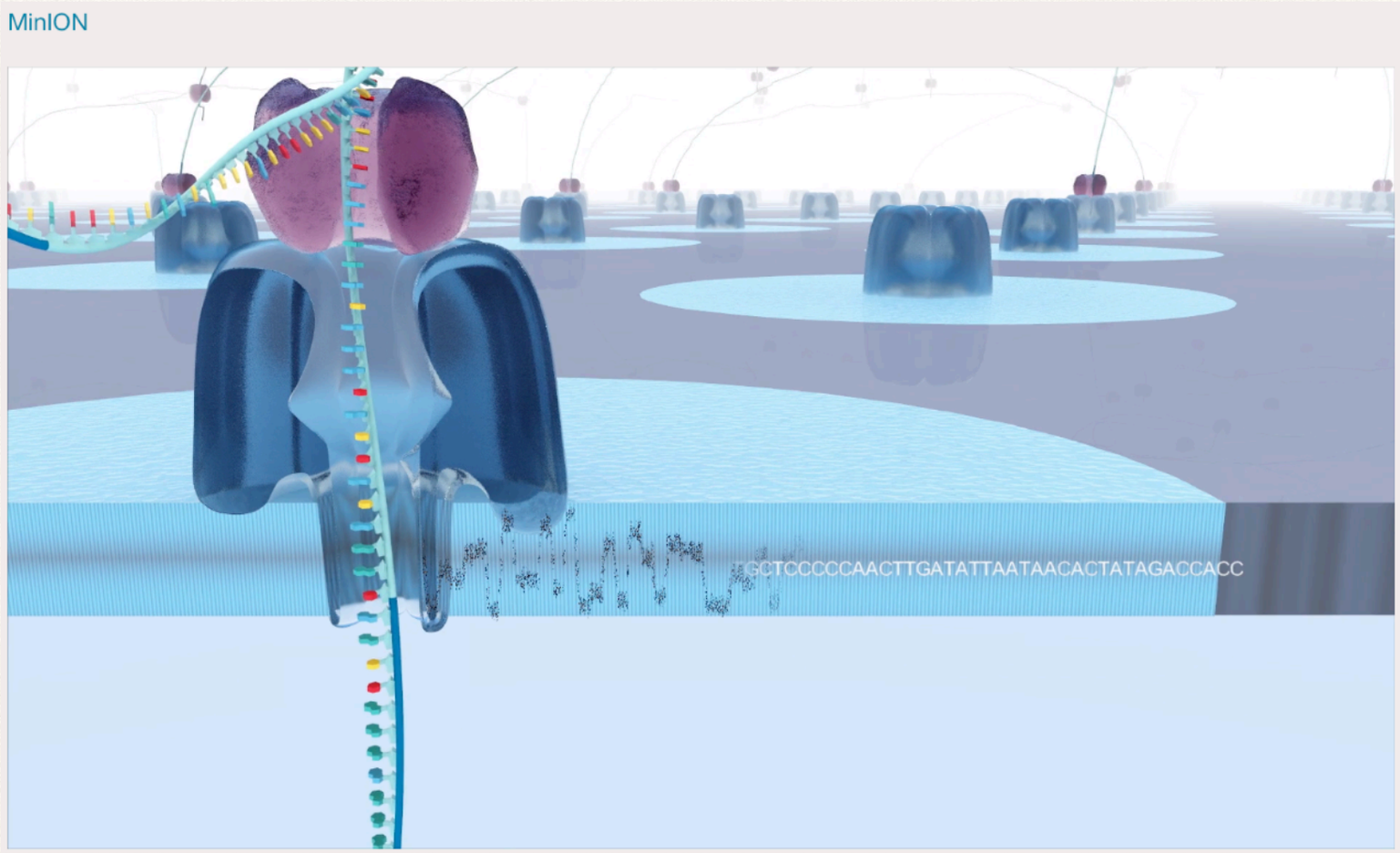
Oxford Nanopore Technologies





https://nanoporetech.com

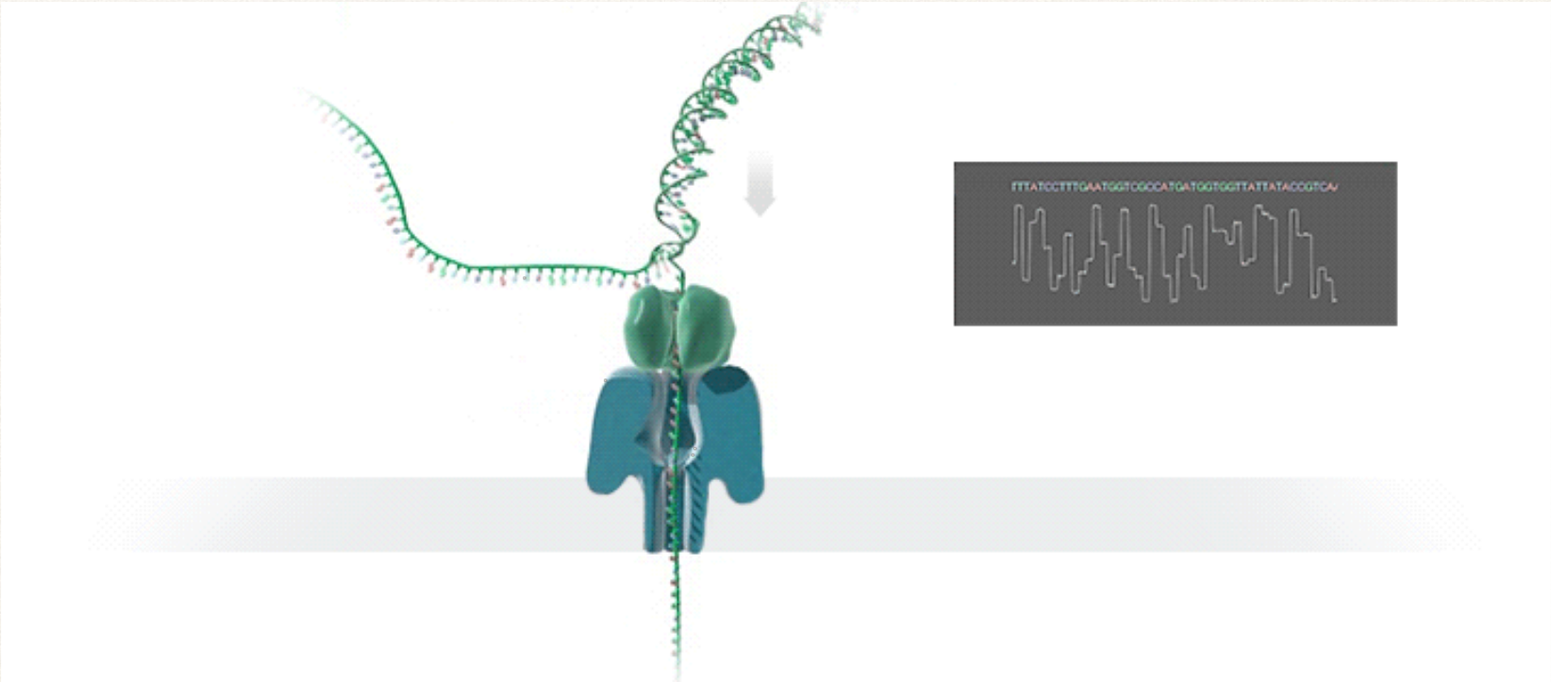
Oxford Nanopore Technologies



- Works monitoring changes to an electrical current as nucleic acids are passed through a protein nanopore.
- The resulting signal is decoded to provide the specific DNA or RNA sequence.

https://nanoporetech.com

Nanopore sequencing

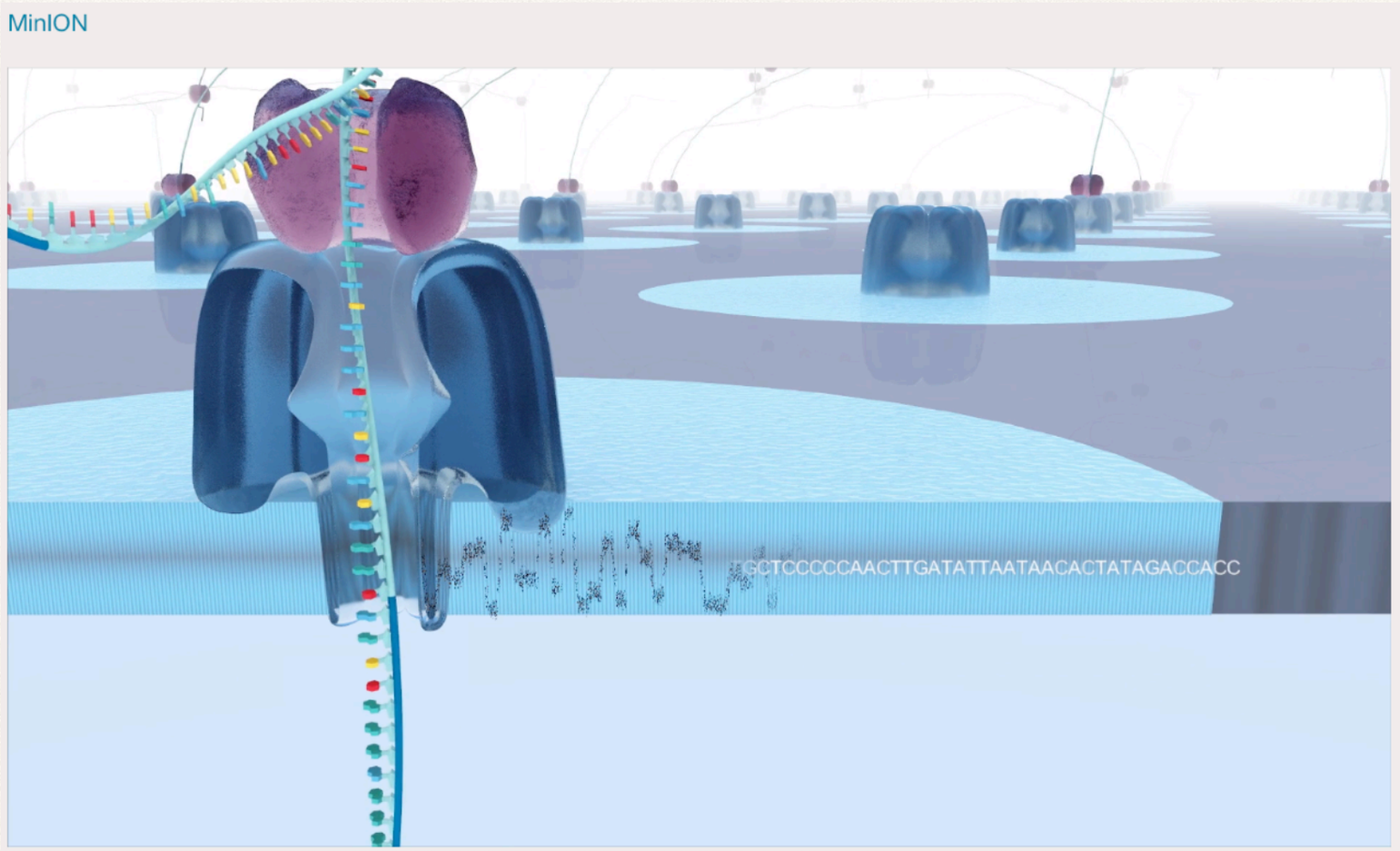


A strand of DNA is passed through a nanopore. The current is changed as the bases G, A, T and C pass through the pore in different combinations.

- The technology works by passing an ionic current through the nanopores. As molecules such DNA or RNA move through the nanopore they produce a characteristic disruption on the current. This signal can be analyze in real time to determined the sequence of bases

https://nanoporetech.com

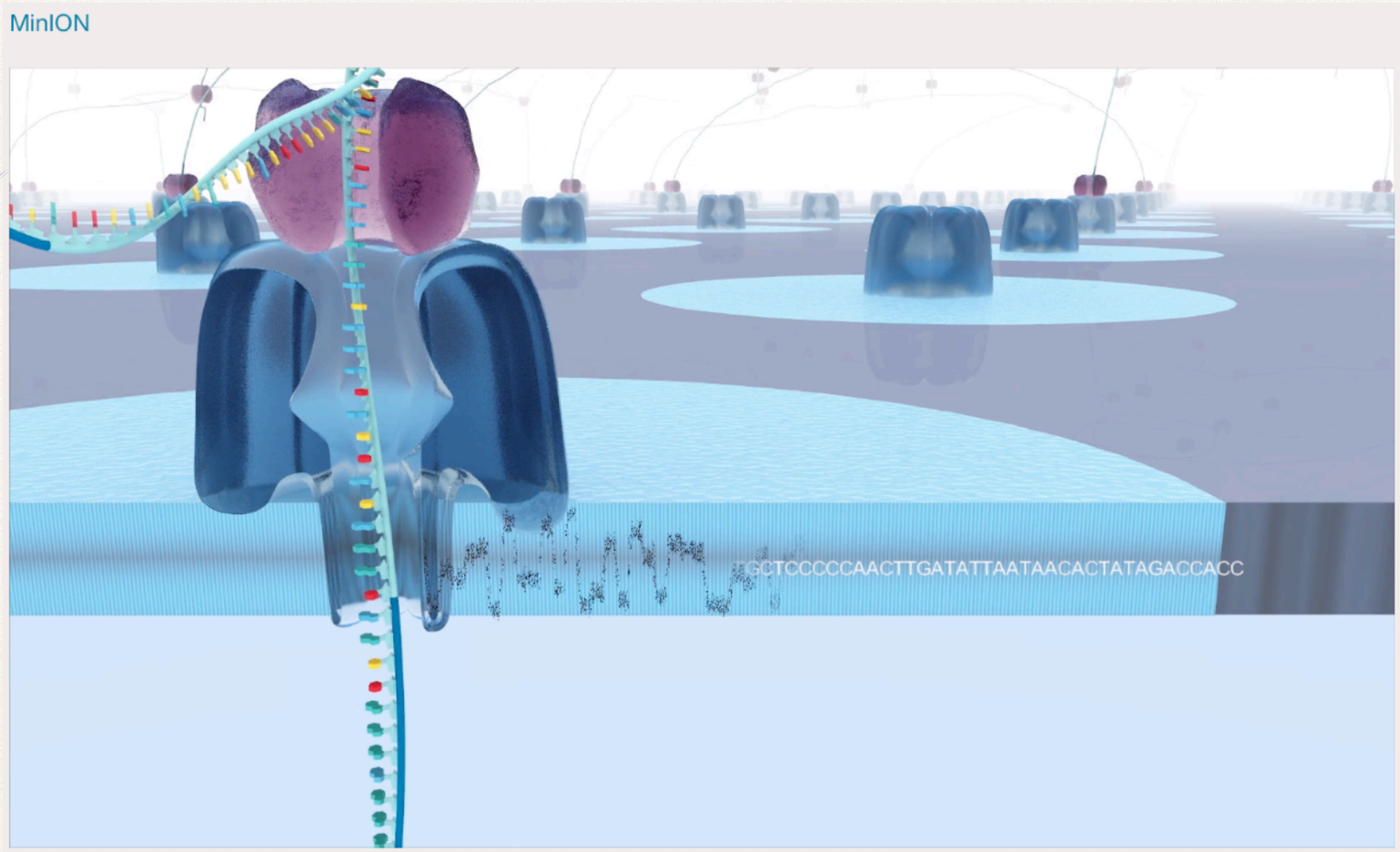
Oxford Nanopore Technologies



- The nanopore processes the length of DNA or RNA present to it.
- Read length = DNA/RNA length
- Not polymerase dependent

https://nanoporetech.com

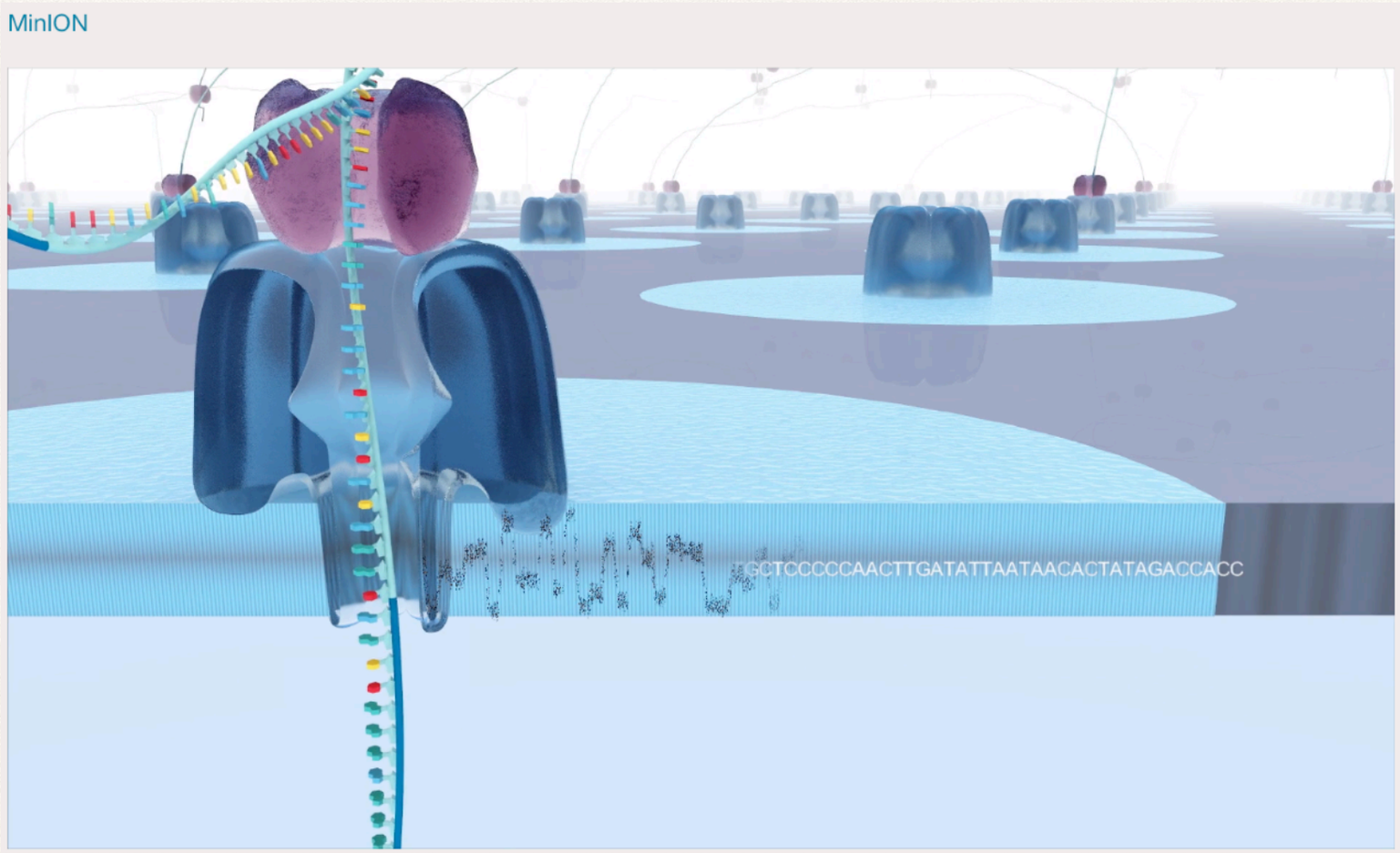
Oxford Nanopore Technologies



- An enzyme motor controls the translocation of the DNA/RNA through the nanopore.
- Once the molecule has passed through, the motor protein detaches and the nanopore is ready to accept the next fragment

https://nanoporetech.com

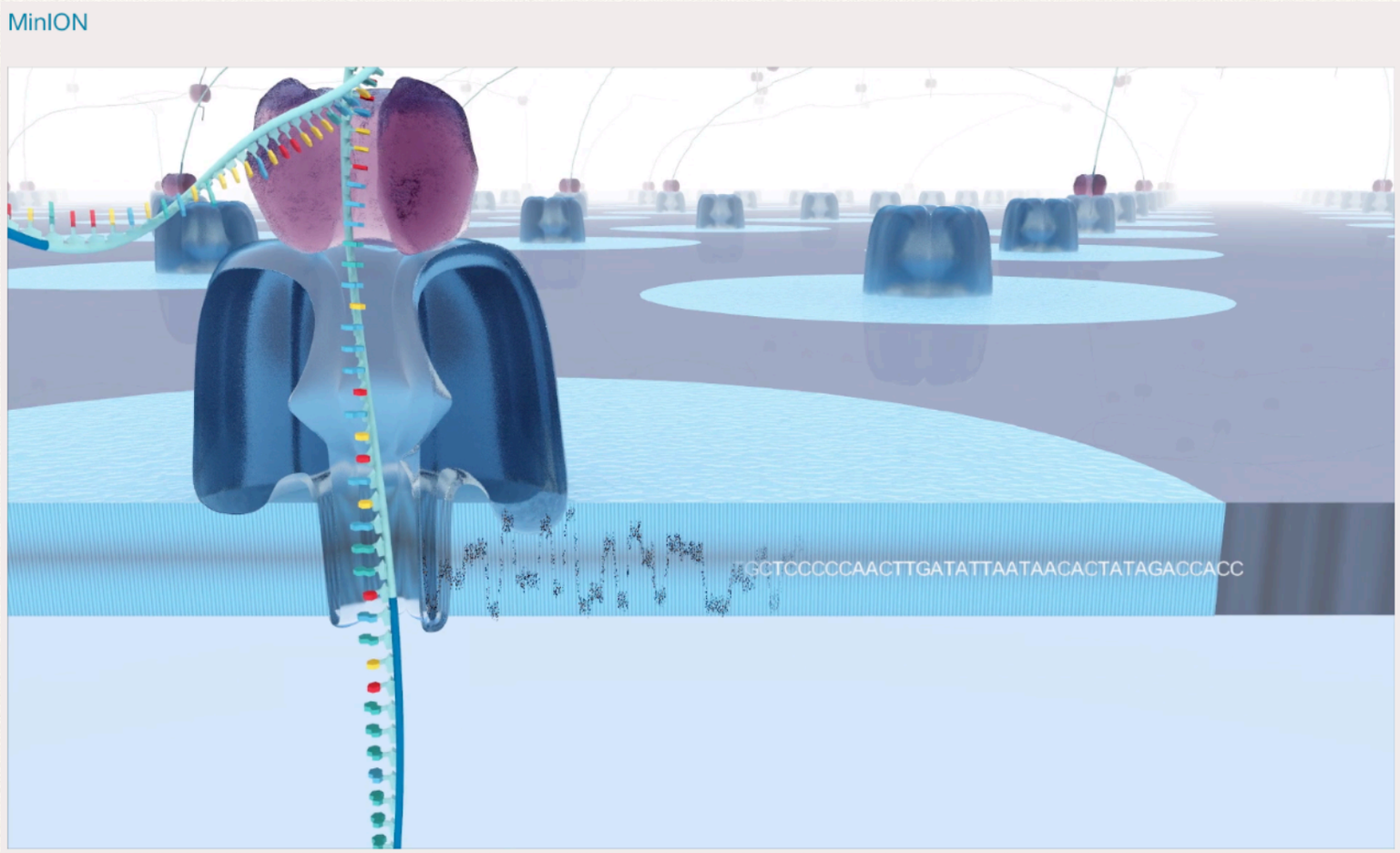
Oxford Nanopore Technologies



- An electrically resistant membrane means all current must pass through the nanopore, ensuring a clean signal

https://nanoporetech.com

Oxford Nanopore Technologies



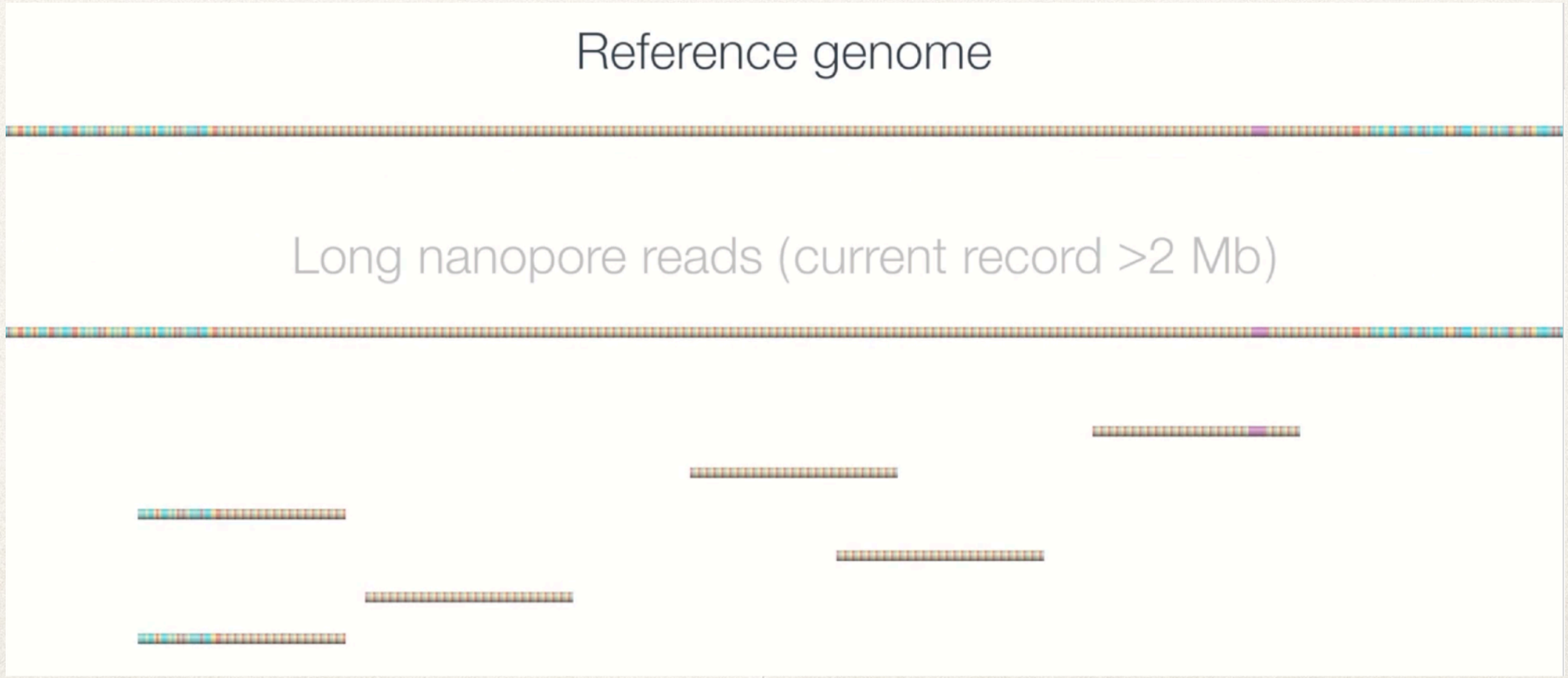
- Library preparation results in the addition of a sequencing adapter and motor protein at each end if the fragment

https://nanoporetech.com

Nanopore sequencing

- Nanopore analyze the entire fragment of DNA/RNA, so the read length is directly related to the length of the DNA/RNA in the sample (>2Mb)

- Long read provide more unambiguous approach to mapping reads enabling too much simple assemblies

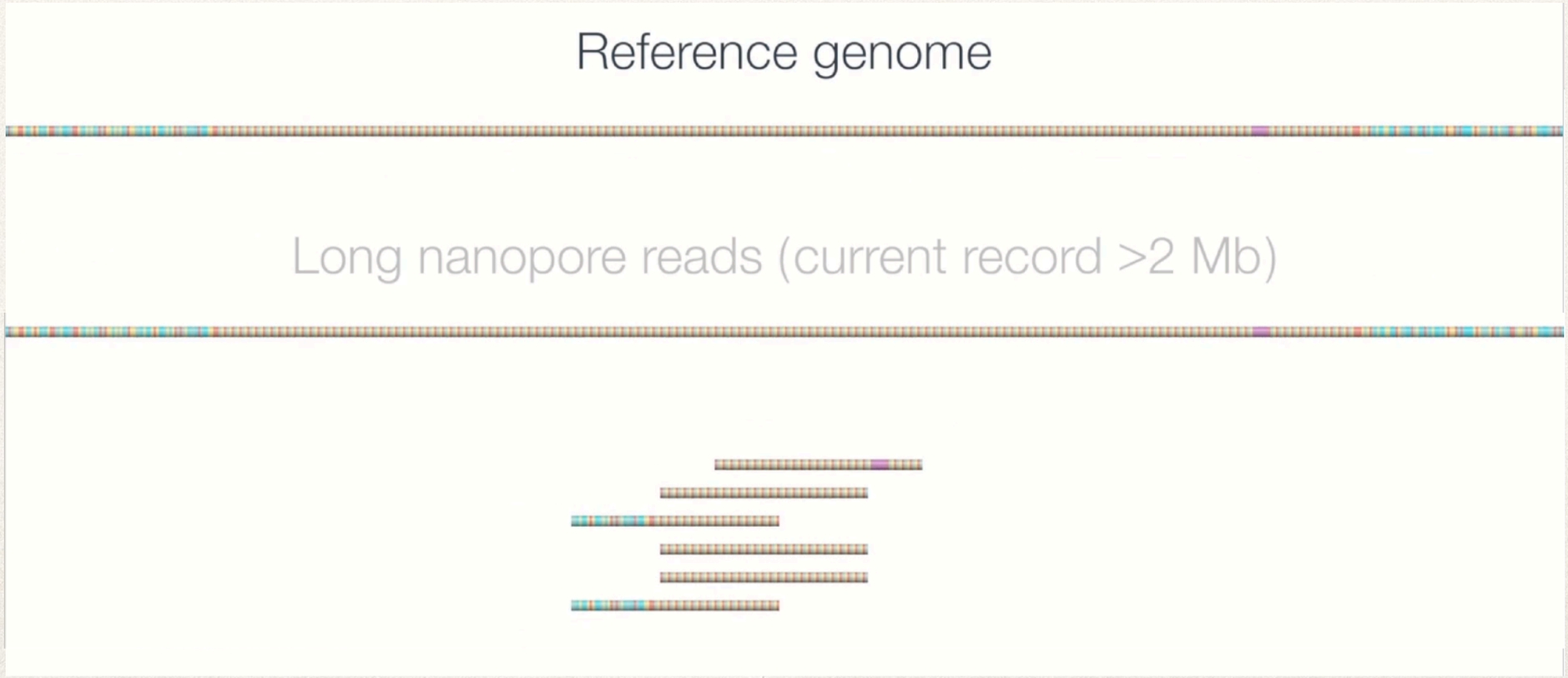


https://nanoporetech.com

Nanopore sequencing

- Nanopore analyze the entire fragment of DNA/RNA, so the read length is directly related to the length of the DNA/RNA in the sample (>2Mb)

- Long read provide more unambiguous approach to mapping reads enabling too much simple assemblies



https://nanoporetech.com

Nanopore sequencing

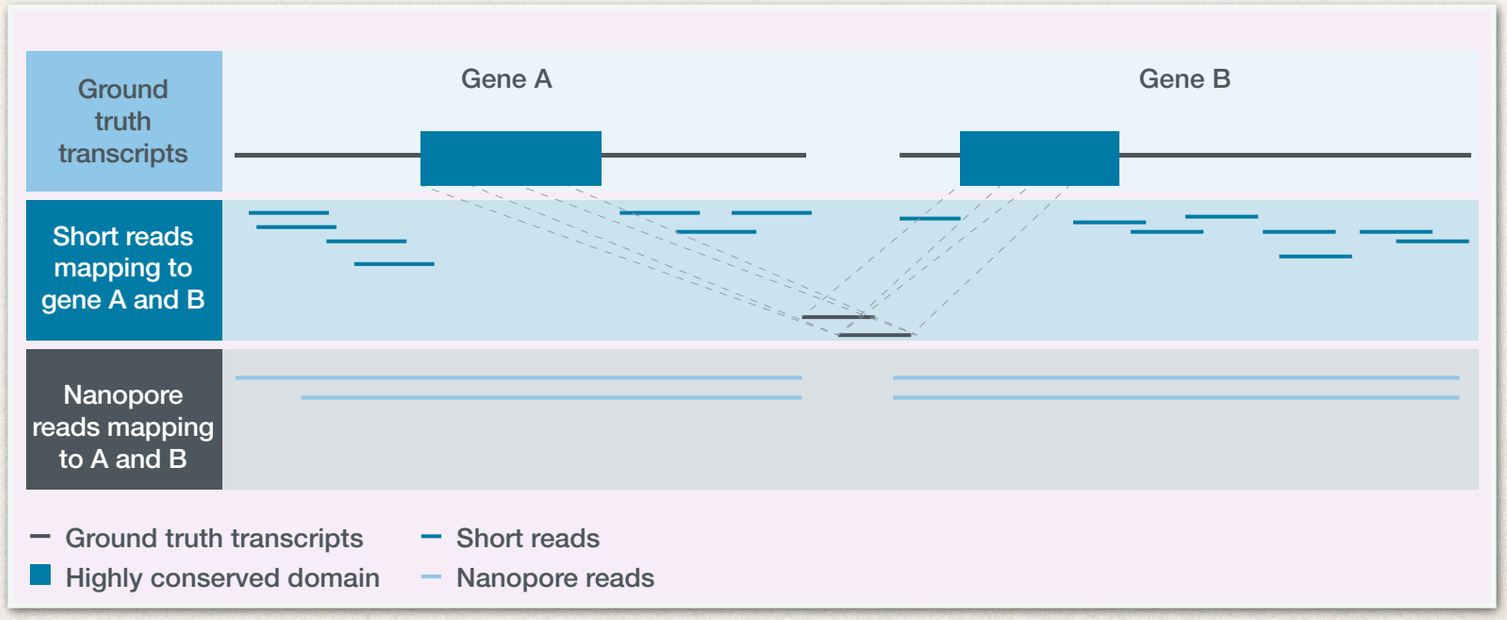
- Nanopore analyze the entire fragment of DNA/RNA, so the read length is directly related to the length of the DNA/RNA in the sample (>2Mb)

- Long read provide more unambiguous approach to mapping reads enabling too much simple assemblies

https://nanoporetech.com

RNA sequencing

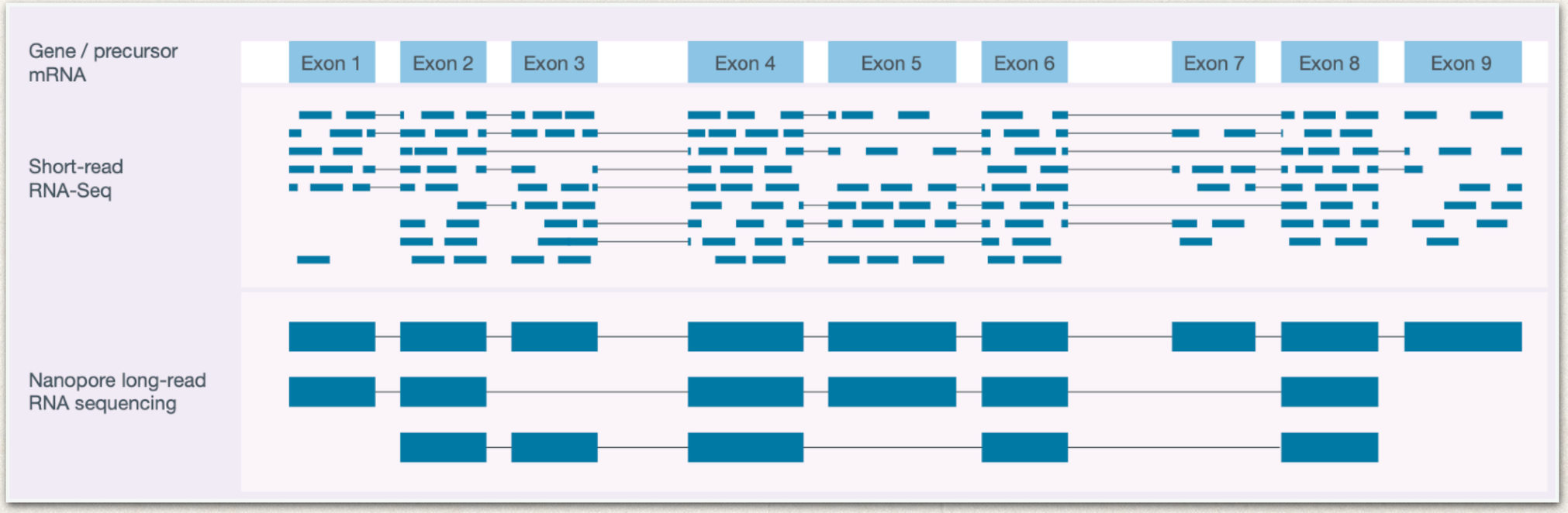
- Characterize and quantify full-length RNA transcripts, splice variants, and fusions
- Accurately analyze differential gene expression and transcript usage.
- Sequence native RNA directly, without amplification or reverse transcription, and identify **base modifications**.



https://nanoporetech.com

RNA sequencing

- Characterize and quantify full-length RNA transcripts, splice variants, and fusions
- Accurately analyze differential gene expression and transcript usage.
- Sequence native RNA directly, without amplification or reverse transcription, and identify **base modifications**.



https://nanoporetech.com

RNA sequencing

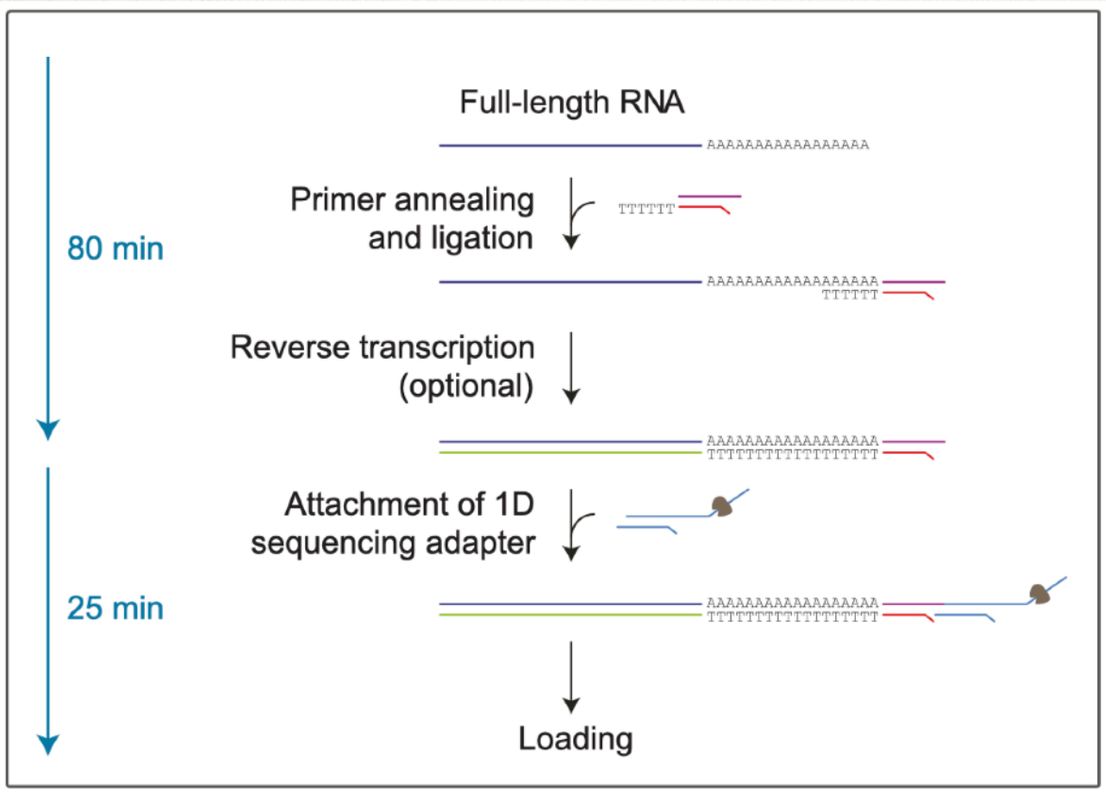
	Direct RNA Sequencing Kit	cDNA-PCR Sequencing Kit	Direct cDNA Sequencing Kit
Preparation time	105 min	165 min	275 min
Input requirement	500 ng RNA (poly-A+)	1 ng RNA (poly-A+)	100 ng RNA (poly-A+)
Reverse transcription required	Optional	Yes	Yes
PCR required	No	Yes	No
Read length	Equal to RNA length	Enriched for full-length cDNA	Enriched for full-length cDNA
Typical throughput	●○○	●●●	●●○
Multiplexing options	In development	Yes	Yes
Overview	Sequencing RNA molecules directly; identify base modifications and poly-A tail length	Optimised for throughput	No PCR bias

https://nanoporetech.com

RNA sequencing

Direct RNA Sequencing Kit

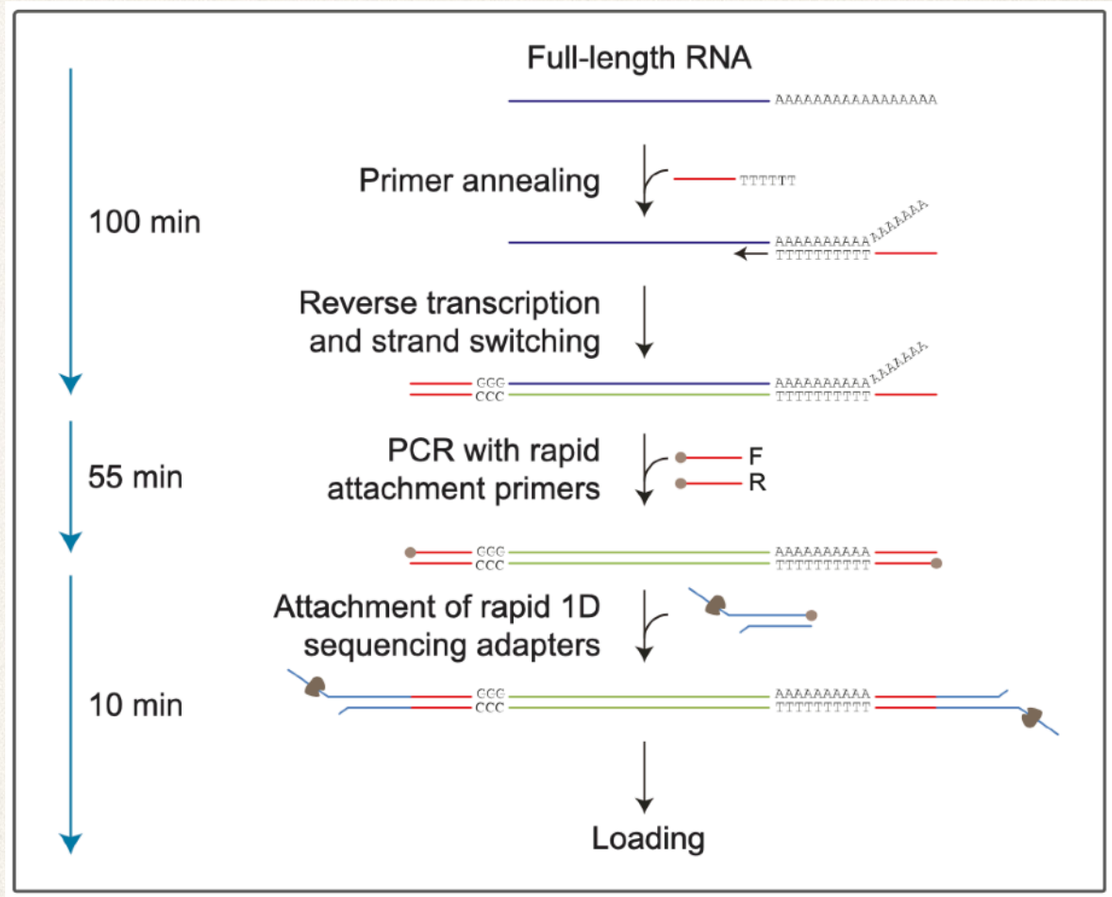
Sequencing RNA directly



- RT optional
- Sequencing adapter attached
- Read length reflects RNA length
- Epigenetic modifications

PCR-cDNA Sequencing Kit

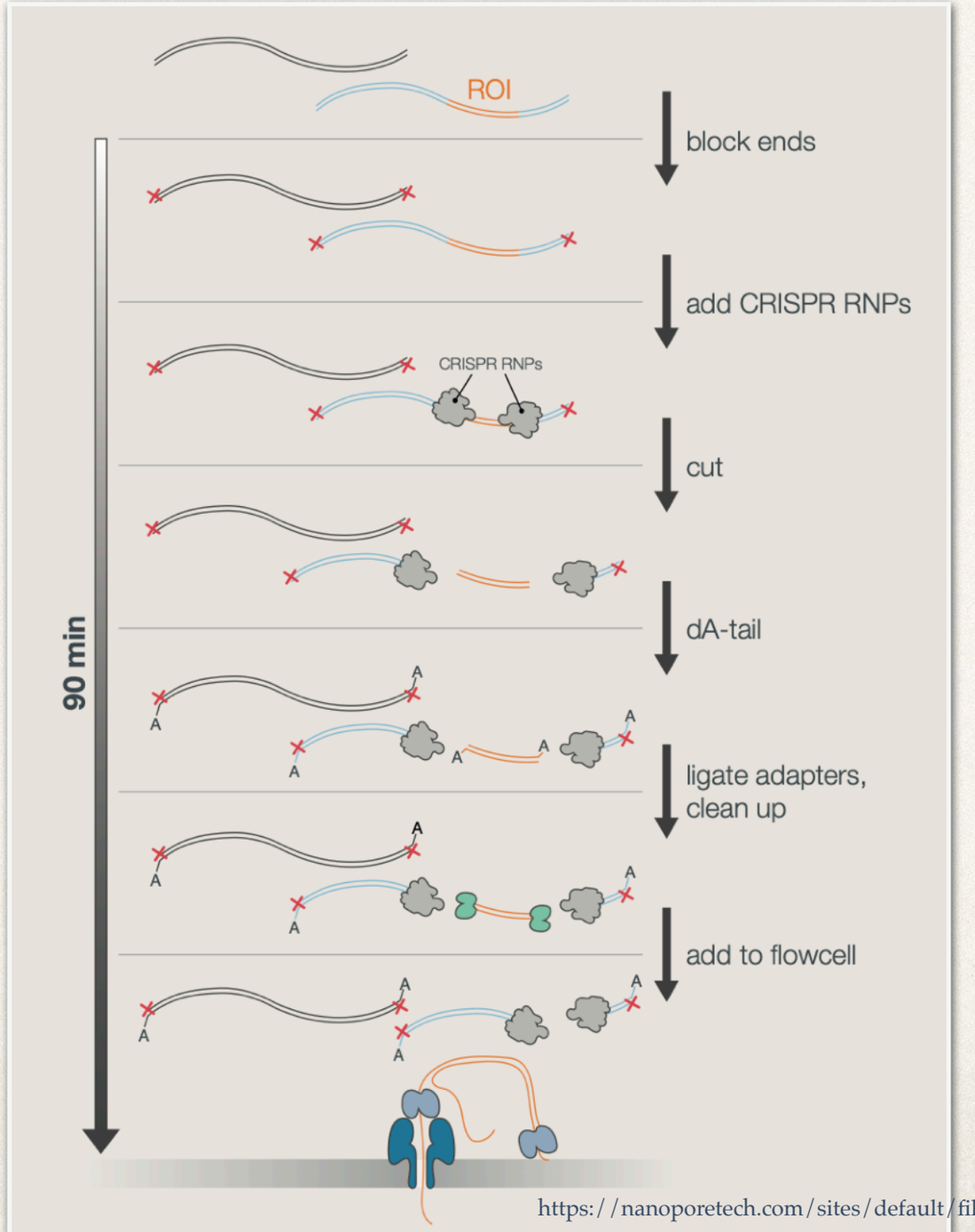
Full-length transcript analysis with high throughput



- RT and PCR amp
- Sequencing adapter attached
- Read length reflects RNA length

https://nanoporetech.com

Targeted sequencing



https://nanoporetech.com/sites/default/files/s3/white-papers/structural-variation-white-paper-oct19.pdf

Benefits of Nanopore technology

Ultra-long reads — longest 2.3 Mb[1]

- Easier genome assembly
- Resolve structural variants, repeats, and phasing
- Full-length transcripts

Direct sequencing

- Sequence native DNA or RNA, not a copy
- Eliminate amplification bias
- Identify base modifications

Streamlined library prep

- Rapid 10-minute (DNA) library prep
- Automated, portable prep — VolTRAX™
- High DNA and RNA yields from low input amounts
- Maximise throughput with barcoding

Real-time analysis

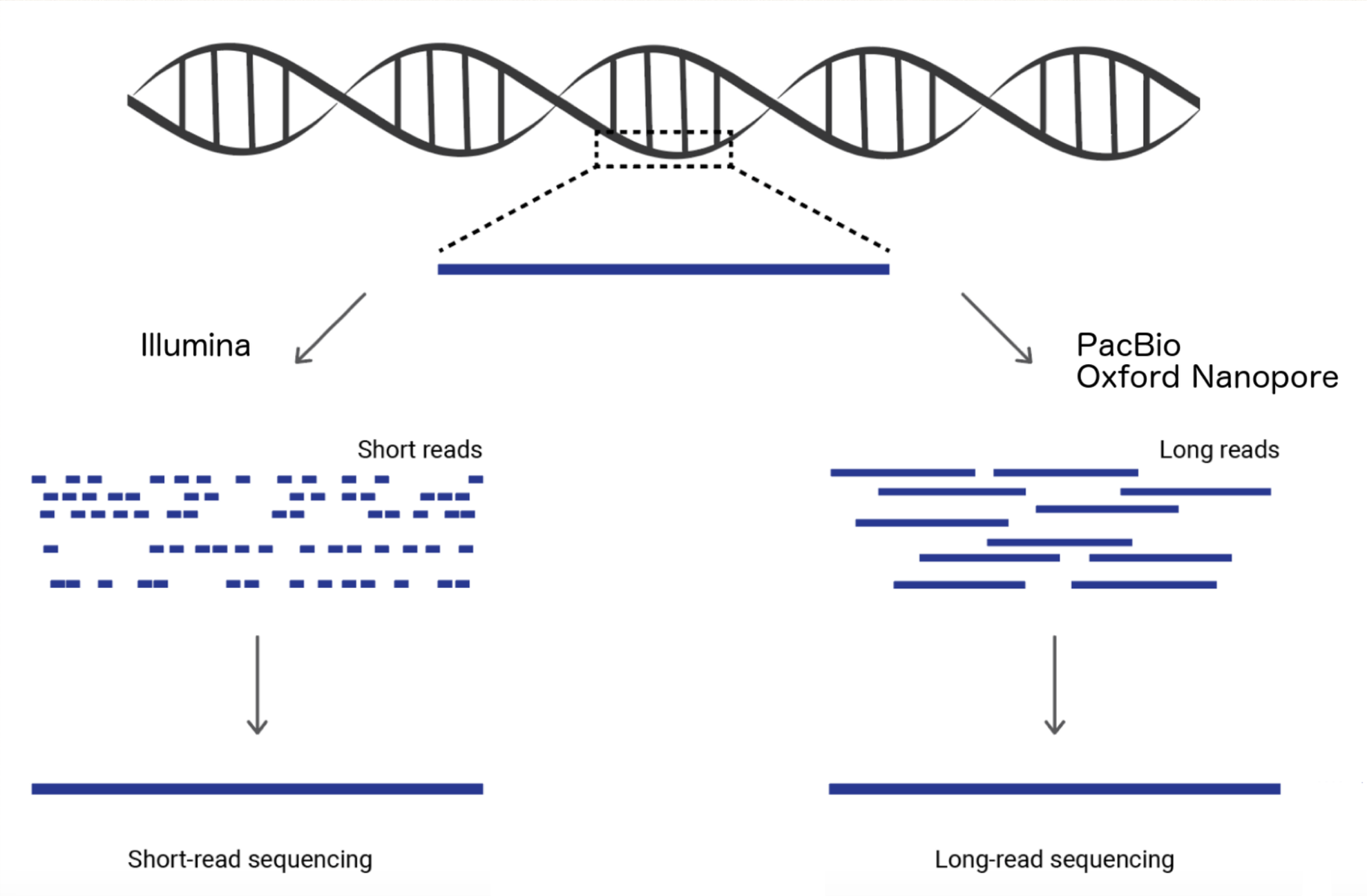
- Immediate access to actionable results
- Rapid species identification
- Early sample insights and QC
- Enough data? Stop, wash, store, or run another sample

Scalable — portable to ultra-high throughput

- One technology across all devices — scale to your needs
- Sequence at sample source with Flongle™ and MinION™
- Compact, high-throughput benchtop sequencing with GridION™ and PromethION™

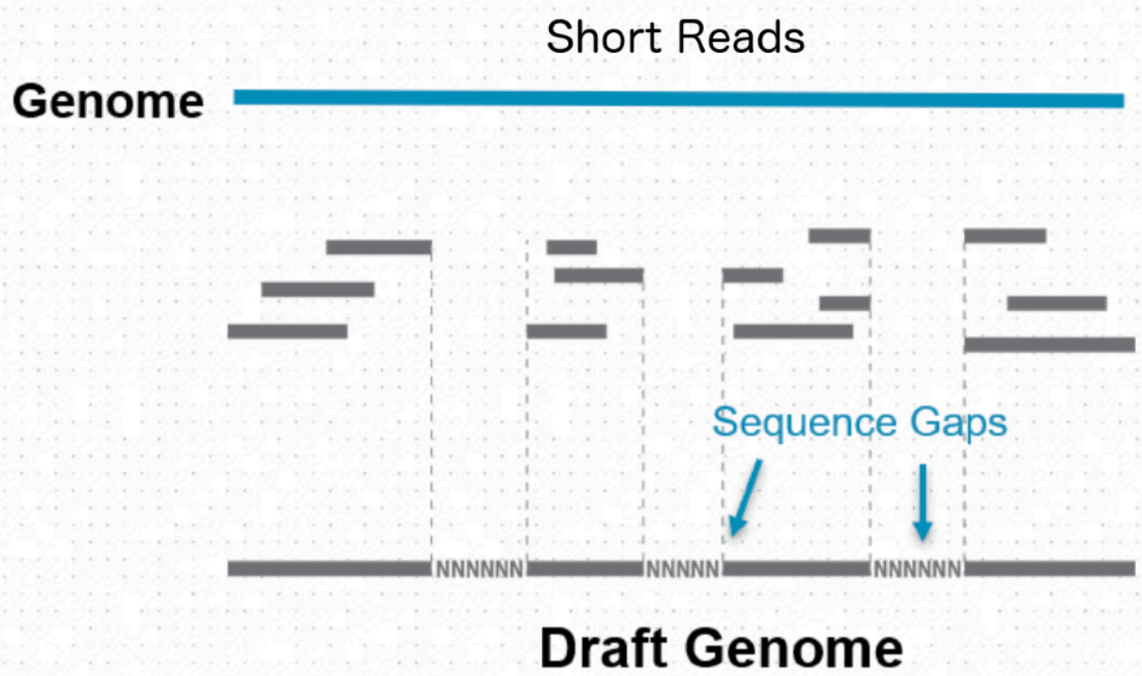
- Direct DNA/RNA sequencing
- REAL Real-time
- No Capital cost required
- Ultra-long reads - up to 2 Mb
- Scalable to portable or desktop
- Simple & rapid, or automated, library prep
- High yields for large genomes
- Comparison table

SGS vs TGS

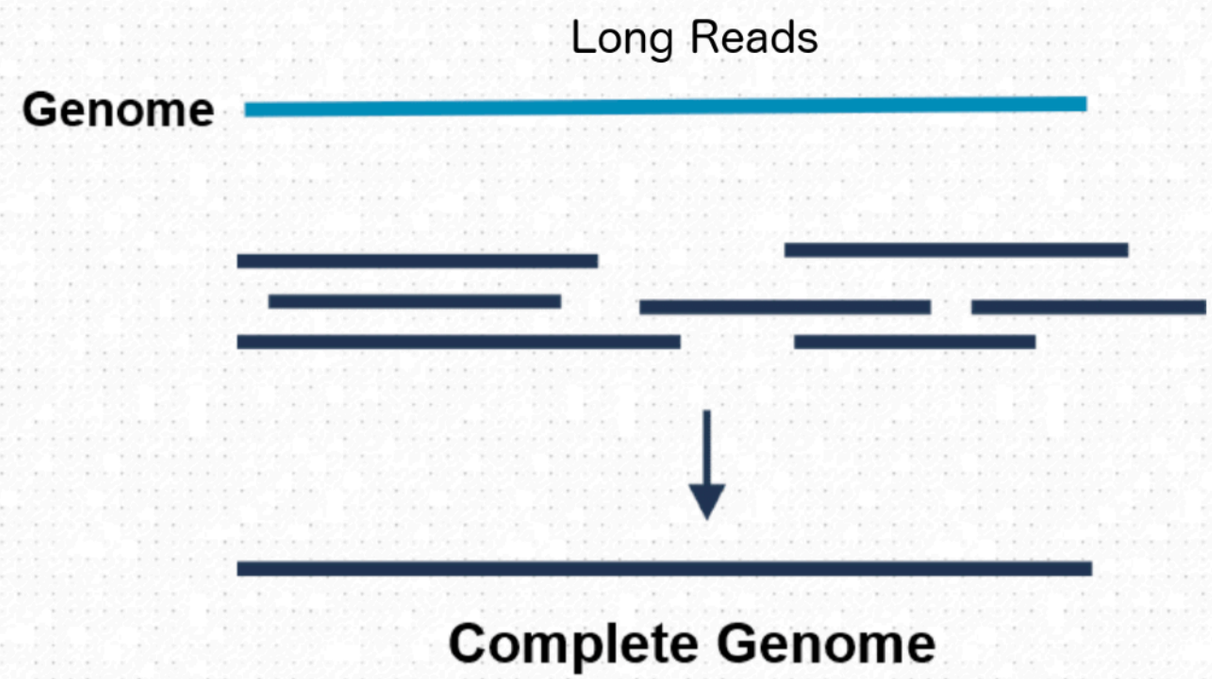


Modified from https://www.macrogen.com.au/system-3/#tab-id-3

SGS vs TGS



Missing sequencing leads to missed genes and limits biological interpretation



A comprehensive structural, functional and organizational picture of the genome

Modificado de https://www.pacb.com/blog/the-evolution-of-dna-sequencing-tools/attachment/short-reads-and-hifi-reads-genome-assembly-comparison/

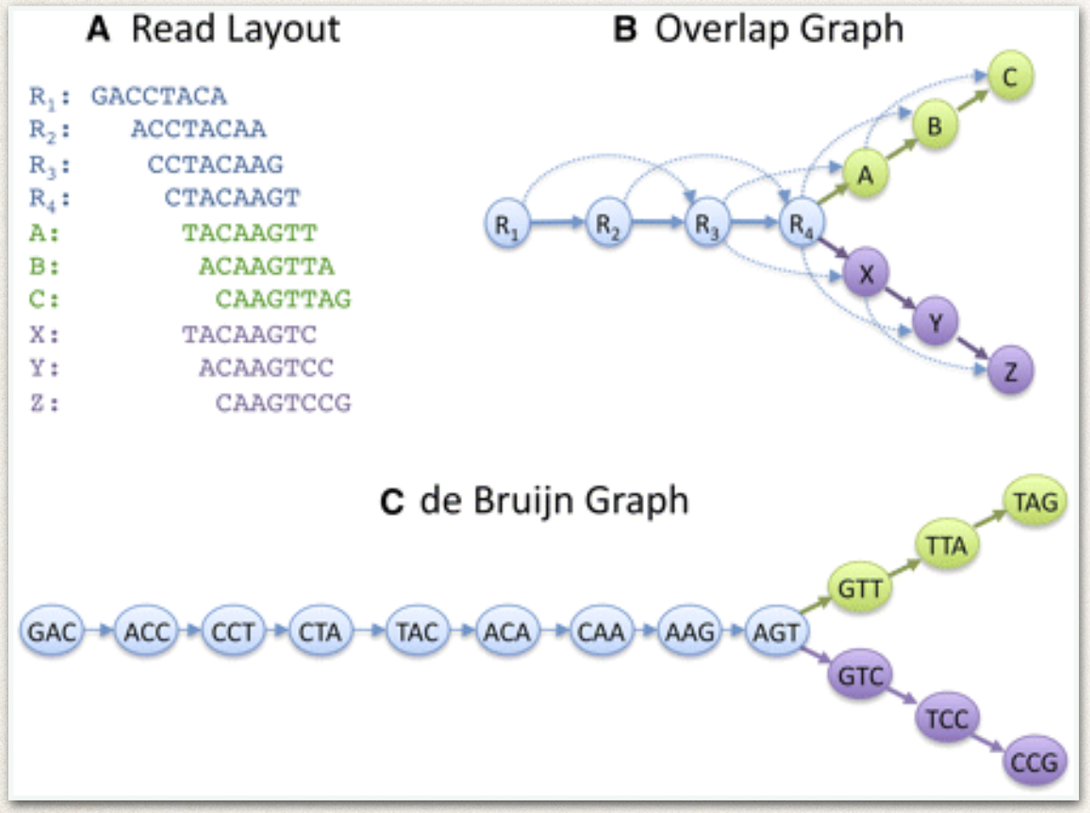
SGS vs TGS

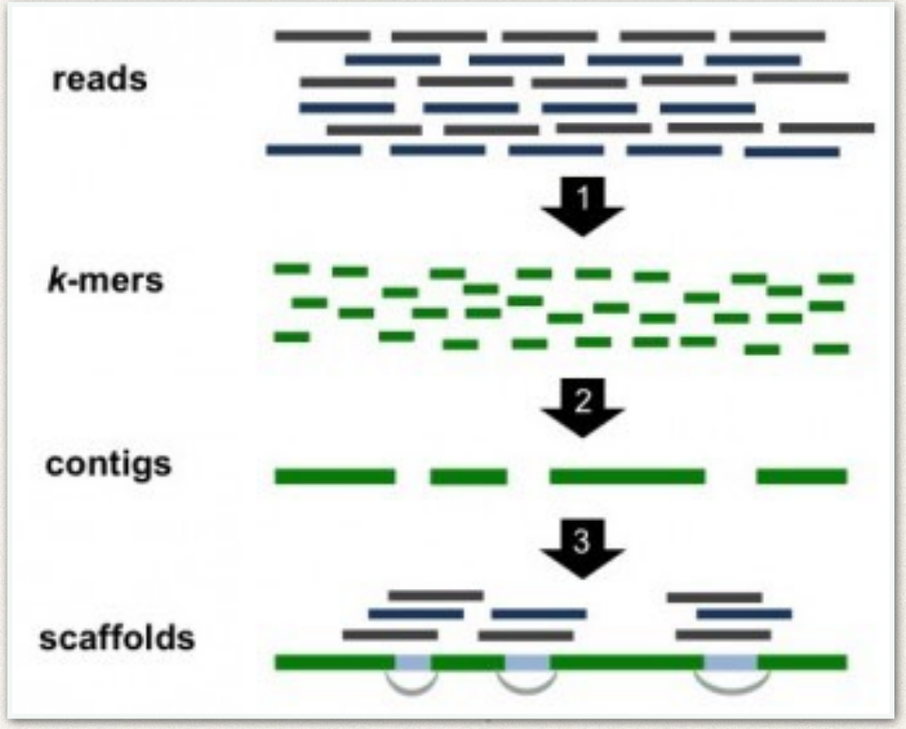
	SGS	TGS
Length	Short read	Long read
PCR amplification	Required	Optional
Throughput	High	Lower
Error rate	Low	Higher
Velocity	Low	Real-time

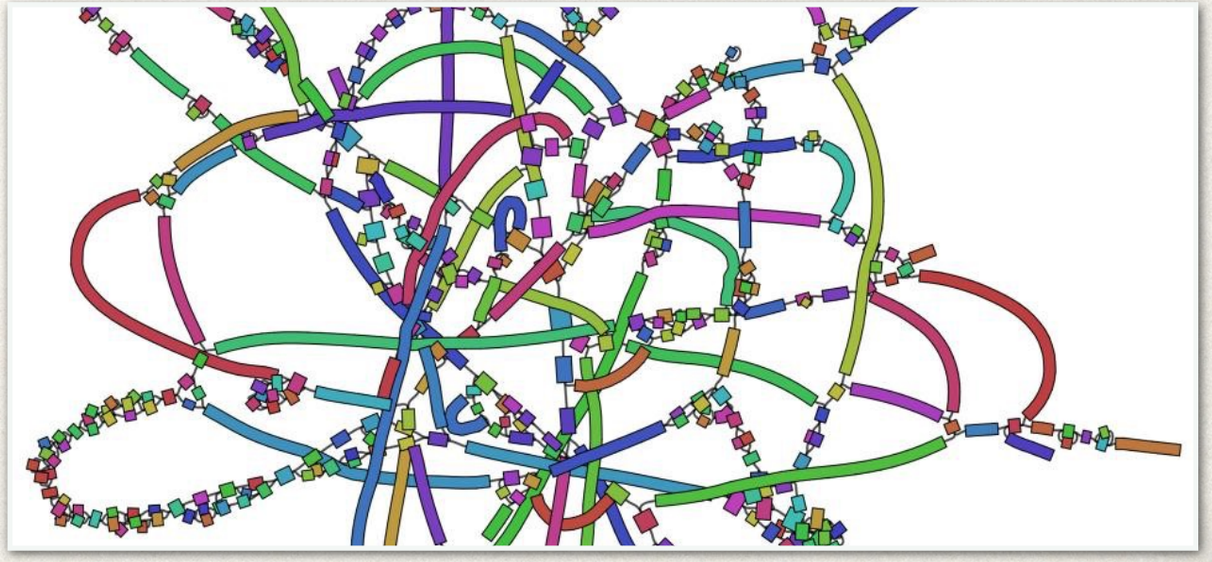
↓

Hybrid assemblies

Assembly algorithms

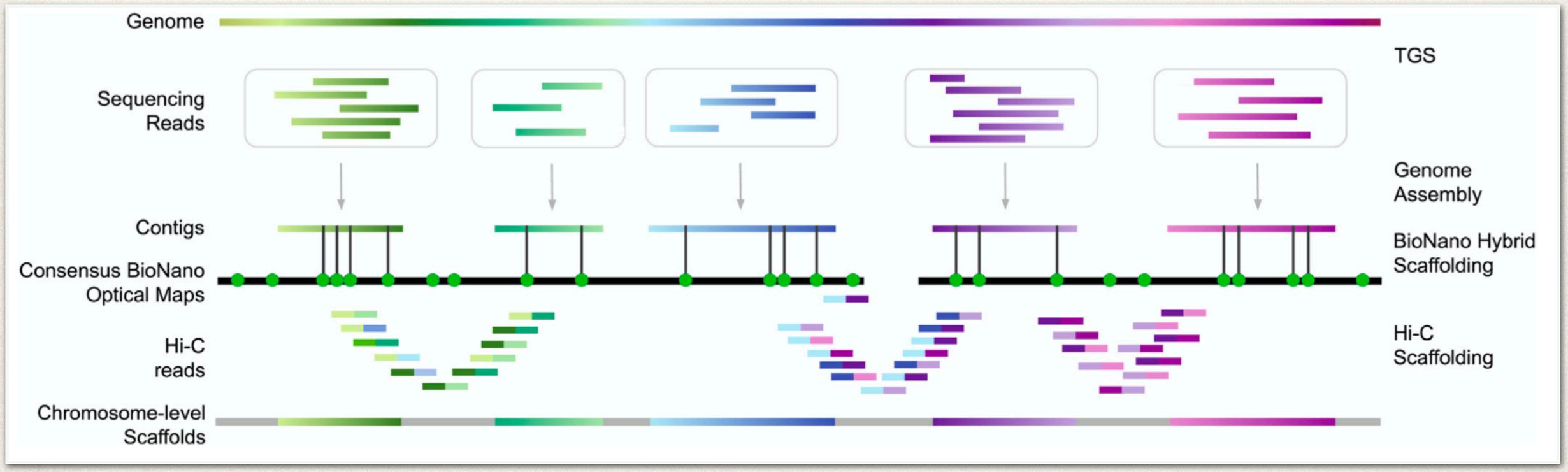






2010-Schatz
https://www.bio21.unimelb.edu.au/humpty-dumpty-program-bandage-helps-piece-dna-sequences-back-together-again-and-wins-2015-iawards

Chromosome-level scaffolding of *de novo* genome assemblies.



Giani et al., 2020. *Computational and Structural Biotechnology Journal*.

Trypanosoma cruzi

Trypanosoma cruzi is a protozoan parasite belonging to the order *Kinetoplastida* that causes Chagas disease, a neglected parasitic disease that affects 6–7 million people worldwide and is transmitted to humans and animals mainly by Triatomine insect vectors

Kinetoplastids genomes

The Genome of the Kinetoplastid Parasite, *Leishmania major*

Alasdair C.
Matthew Be
Zina Apostolo
Gabriella Bia
Laura Ciarloni,[8]
Javier De C
Alberto Carlo
David Harri
Andrew Knight
David Masuy,[15] Ke
Siri Nelson,[2] Hali
Bénédicte Purne
Johan Robbe
Jacquie Schein,[17]
Rob Squares,[1] Ste
Holger W

The Genome of the African Trypanosome *Trypanosoma brucei*

Matthew Berriman,[1*] Elodie Ghedin,[2,3] Christiane Hertz-Fowler,[1] Gaëlle Blandin,[2] Hubert Renauld,[1]
Daniella C. Bartholomeu,[2]
Linda Hannick,[2] Marti
U. Cecilia M. Alsmark,[6] Claire
Mark Carrington,[8] Inna Cherev
Ann Cronin,[1] Rob M. Davie
Audrey Fraser,[1] Ian Goodhead,
Al Ivens,[1] Kay Jagels,[1] Dav
Natasha Larke,[1] Scott La
Annette MacLeod,[4] Paul J. M
Halina Norbertczak,[1] Dou
Ester Rabbinowitsch,[1] Marie-A
Sarah Sharp,[1] Mark Simm
Adrian R. Tivey,[1] Susan Van A
Sally Whitehead,[1] John W
Elisabetta Ullu,[15] J. David Ba
Neil Ha

The Genome Sequence of *Trypanosoma cruzi*, Etiologic Agent of Chagas Disease

Najib M. El-Sayed,[1,2*†] Peter J. Myler,[3,4,5*†] Daniella C. Bartholomeu,[1] Daniel Nilsson,[6] Gautam Aggarwal,[3]
Anh-Nhi Tran,[6] Elodie Ghedin,[1,2] Elizabeth A. Worthey,[3] Arthur L. Delcher,[1] Gaëlle Blandin,[1] Scott J. Westenberger,[1,7]
Elisabet Caler,[1] Gustavo C. Cerqueira,[1,8] Carole Branche,[6] Brian Haas,[1] Atashi Anupama,[3] Erik Arner,[6] Lena Åslund,[9]
Philip Attipoe,[3] Esteban Bontempi,[6,10] Frédéric Bringaud,[11] Peter Burton,[12] Eithon Cadag,[3] David A. Campbell,[7]
Mark Carrington,[13] Jonathan Crabtree,[1] Hamid Darban,[6] Jose Franco da Silveira,[14] Pieter de Jong,[15]
Kimberly Edwards,[6] Paul T. Englund,[16] Gholam Fazelina,[3] Tamara Feldblyum,[1] Marcela Ferella,[6]
Alberto Carlos Frasch,[17] Keith Gull,[18] David Horn,[19] Lihua Hou,[1] Yiting Huang,[3] Ellen Kindlund,[6] Michele Klingbeil,[20]
Sindy Kluge,[6] Hean Koo,[1] Daniela Lacerda,[1,21] Mariano J. Levin,[22] Hernan Lorenzi,[22] Tin Louie,[3]
Carlos Renato Machado,[8] Richard McCulloch,[12] Alan McKenna,[6] Yumi Mizuno,[6] Jeremy C. Mottram,[12]
Siri Nelson,[3] Stephen Ochaya,[6] Kazutoyo Osoegawa,[15] Grace Pai,[1] Marilyn Parsons,[3,4] Martin Pentony,[3]
Ulf Pettersson,[9] Mihai Pop,[1] Jose Luis Ramirez,[23] Joel Rinta,[3] Laura Robertson,[3] Steven L. Salzberg,[1]
Daniel O. Sanchez,[17] Amber Seyler,[3] Reuben Sharma,[13] Jyoti Shetty,[1] Anjana J. Simpson,[1] Ellen Sisk,[3]
Martti T. Tammi,[6,24] Rick Tarleton,[25] Santuza Teixeira,[8] Susan Van Aken,[1] Christy Vogt,[3]
Pauline N. Ward,[12] Bill Wickstead,[18] Jennifer Wortman,[1] Owen White,[1] Claire M. Fraser,[1]
Kenneth D. Stuart,[3,4] Björn Andersson[6†]

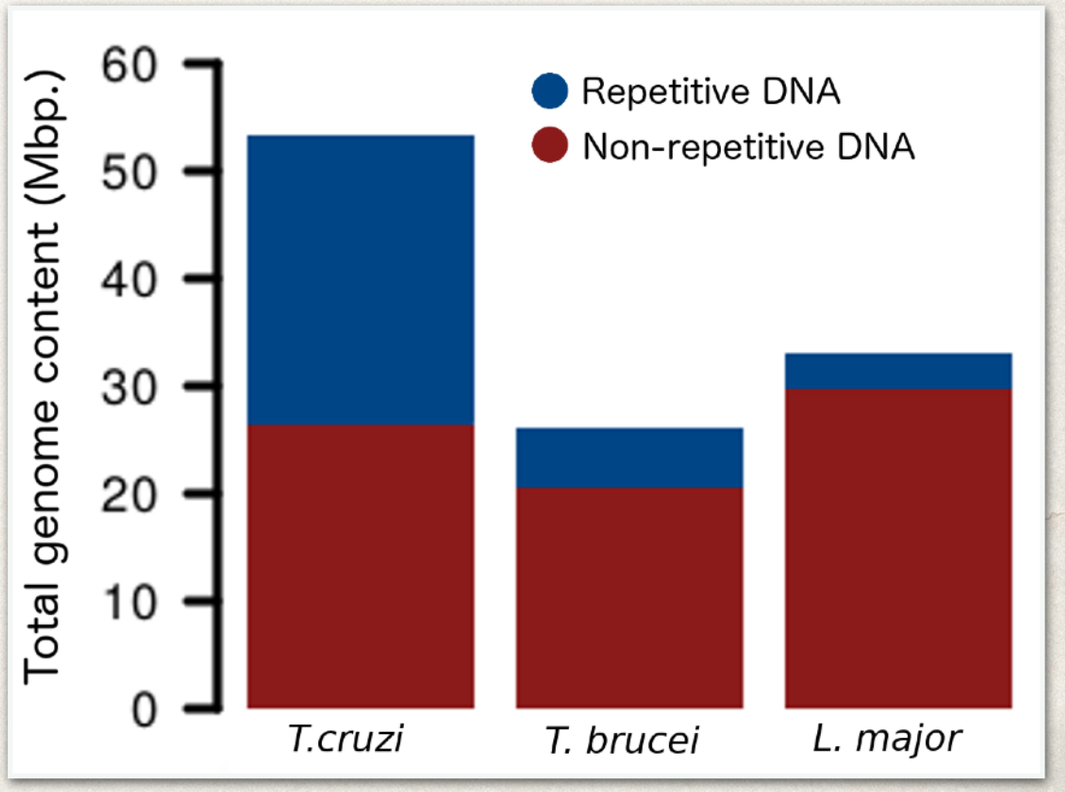
Trypanosoma cruzi genome

- Despite trypanosomatids genomes are quite small, their assembly and annotation has been challenging due to the abundance of repetitive sequences.

- ~ 50 % of the genome is composed by repetitive sequences

- Long read sequencing technology

  (PacBio & Oxford Nanopore)



Pita, Díaz-Viraqué, Iraola, Robello. *Genome Biology and Evolution*. 2019

Could ONT and PacBio improve *T. cruzi* genome assembly?

- Whole-genome assembly using Illumina, PacBio and ONT

MICROBIAL GENOMICS

RESEARCH ARTICLE
Berná *et al.*, *Microbial Genomics* 2018;4
DOI 10.1099/mgen.0.000177

MICROBIOLOGY SOCIETY
OPEN DATA OPEN MICROBIOLOGY

Expanding an expanded genome: long-read sequencing of *Trypanosoma cruzi*

Luisa Berná,[1] Matias Rodriguez,[2] María
Fernando Alvarez-Valin[2,*] and Carlos F

GBE

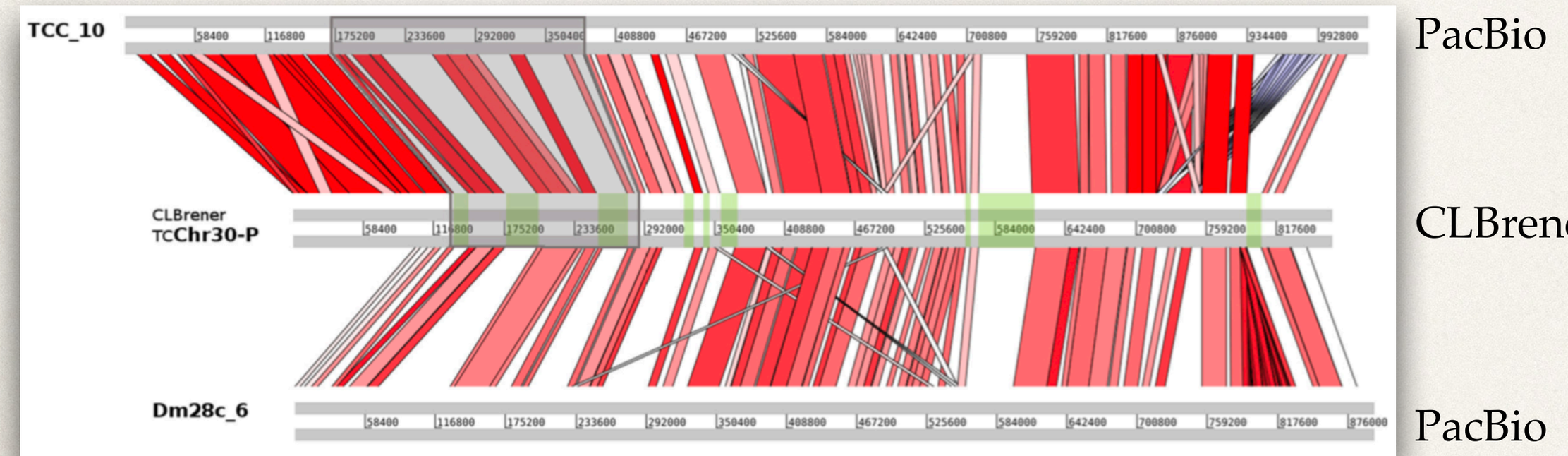
Nanopore Sequencing Significantly Improves Genome Assembly of the Protozoan Parasite *Trypanosoma cruzi*

Florencia Díaz-Viraqué[1,*], Sebastián Pita[1,2], Gonzalo Greif[1], Rita de Cássia Moreira de Souza[3], Gregorio Iraola[4,5,*], and Carlos Robello[1,6,*]

T. cruzi genome assembly with PacBio

Improvements:

1: Determine sequences filled by Ns in others alignments (in green)



PacBio

CLBrener

PacBio

Comparison between chr30P from CL Brener with TCC and Dm28c contigs reported using long reads. These Ns regions were fully resolved in Dm28c and TCC assemblies.

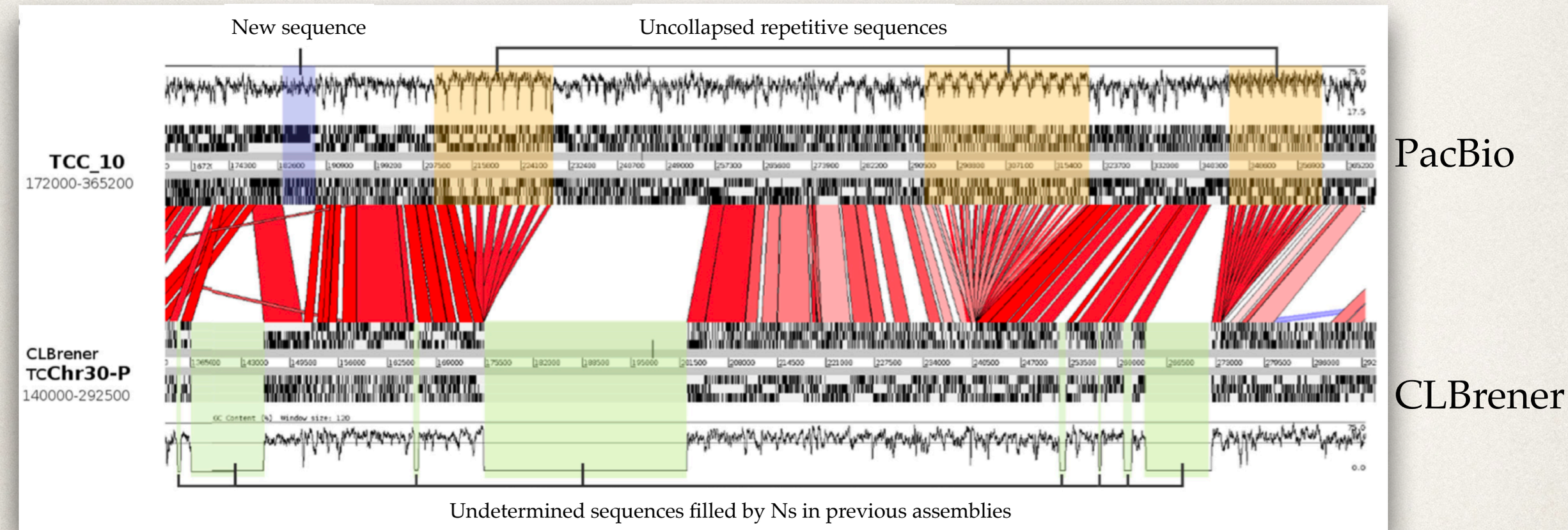
Berná et al. *Microbial Genomics*. 2018

T. cruzi genome assembly with PacBio

Improvements:

2: Determined the full sequence of large clusters of repetitive sequences

Uncollapsed repetitive sequences (showed in orange)



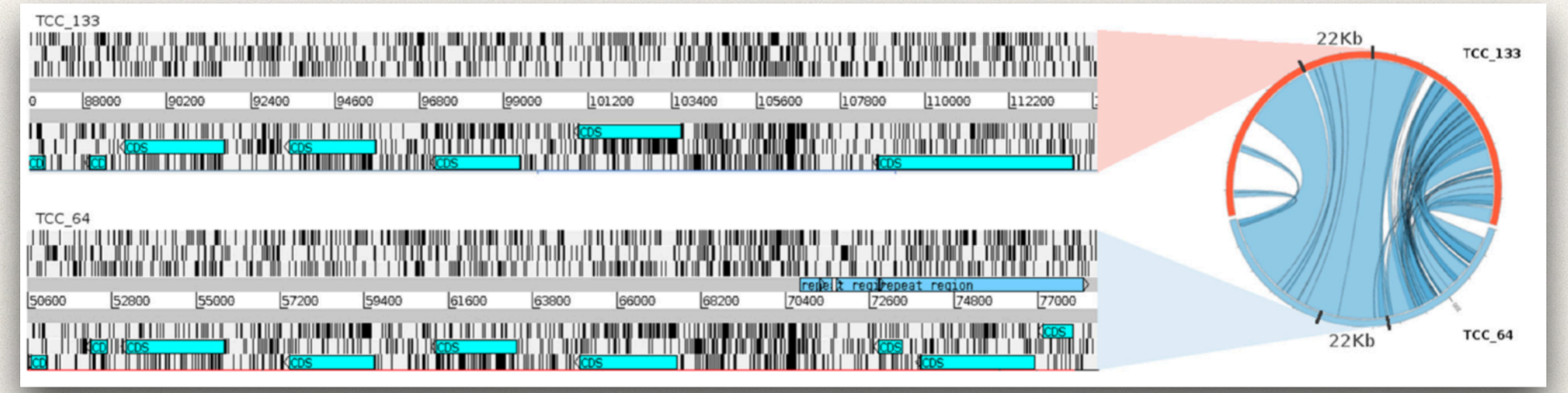
Repetitive sequences clusters, which were collapsed in the previous assembly, now are disaggregated into the actual copy number

Berná et al. *Microbial Genomics*. 2018

T. cruzi genome assembly with PacBio

Improvements:

3: Separate assembly the parental haplotypes and detect recombination



For TCC (the hybrid strain) we were able to assemble separately the parental haplotypes.

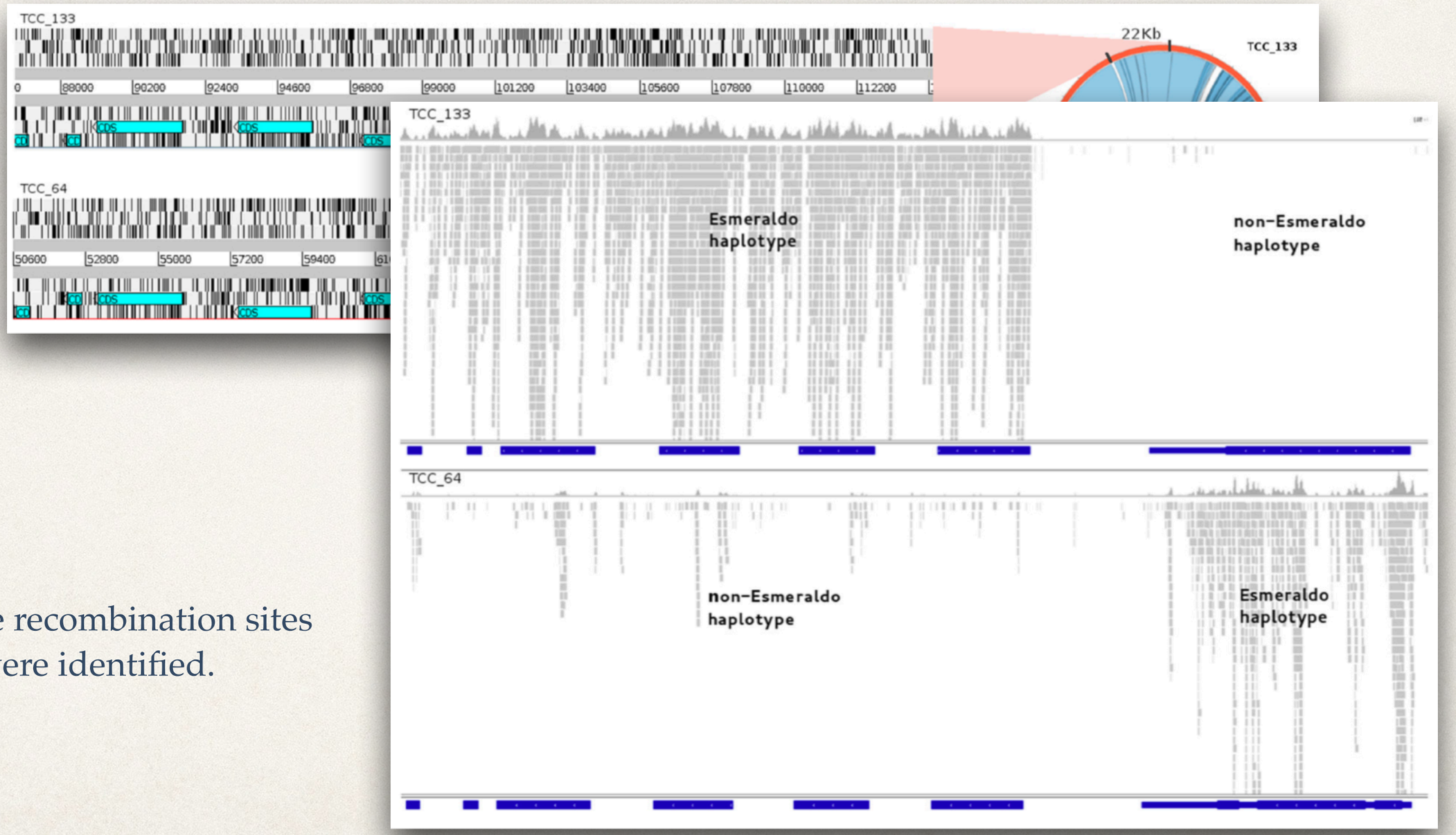
To test if homologous recombination events could be detected, the Illumina reads from Esmeraldo (one of the parental) were mapped to the genome.

Berná et al. *Microbial Genomics*. 2018

T. cruzi genome assembly with PacBio

Improvements:

3: Separate assembly the parental haplotypes and detect recombination



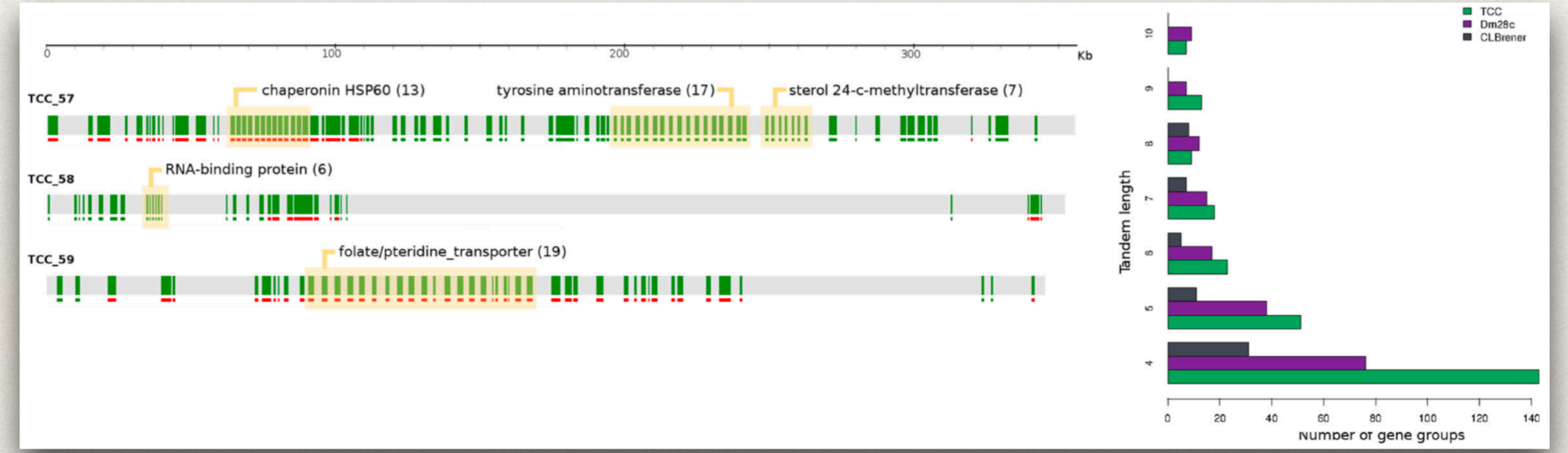
Putative recombination sites were identified.

Berná et al. *Microbial Genomics*. 2018

T. cruzi genome annotation
Multigene families

- The genes from a family were found in tandem and dispersed.

- The resolution of previously collapsed repetitive regions allowed us to visualize and measure the extent of the tandem arrays of genes



There are 5 times more gene tandems of 4 genes in TCC than those identified previously.

Berná et al. *Microbial Genomics*. 2018

Berenice

Berenice is the first *T. cruzi* strain isolated from a patient, so it has epidemiological and historical relevance.

Strategy

Two genomes assemblies
(MaSuRCA)

Using short reads
from Illumina

Illumina assembly

Combining Illumina
short reads with
Nanopore long reads

Hybrid assembly

Summary of genome assemblies

	Hybrid genome assembly	Illumina genome assembly
Assembly features		
Number of contigs	923	46,821
Largest contig	926,516	26,836
Size (bp)	40,801,262	25,004,252
GC (%)	51.20	48.67
N50	156,193	659
N75	40,889	333
Number of contigs (>= 50,000 pb)	160	0

Díaz-Viraqué et al. 2019 *Genome Biology and Evolution*

Summary of genomes assemblies

51-fold decrease in contig number

	Hybrid genome assembly	Illumina genome assembly
Assembly features		
Number of contigs	923	46,821
Largest contig	926,516	26,836
Size (bp)	40,801,262	25,004,252
GC (%)	51.20	48.67
N50	156,193	659
N75	40,889	333
Number of contigs (>= 50,000 pb)	160	0

Díaz-Viraqué et al. 2019 *Genome Biology and Evolution*

Summary of gen lies

~1 Mb vs ~26 Kb

	Hybrid genome assembly	Illumina genome assembly
Assembly features		
Number of contigs	923	46,821
Largest contig	926,516	26,836
Size (bp)	40,801,262	25,004,252
GC (%)	51.20	48.67
N50	156,193	659
N75	40,889	333
Number of contigs (>= 50,000 pb)	160	0

Díaz-Viraqué et al. 2019 *Genome Biology and Evolution*

Summary of gen lies

Nanopore increase
~16 Mb in assembly size

	Hybrid genome assembly	Illumina genome assembly
Assembly features		
Number of contigs	923	46,821
Largest contig	926,516	26,836
Size (bp)	40,801,262	25,004,252
GC (%)	51.20	48.67
N50	156,193	659
N75	40,889	333
Number of contigs (>= 50,000 pb)	160	0

Díaz-Viraqué et al. 2019 *Genome Biology and Evolution*

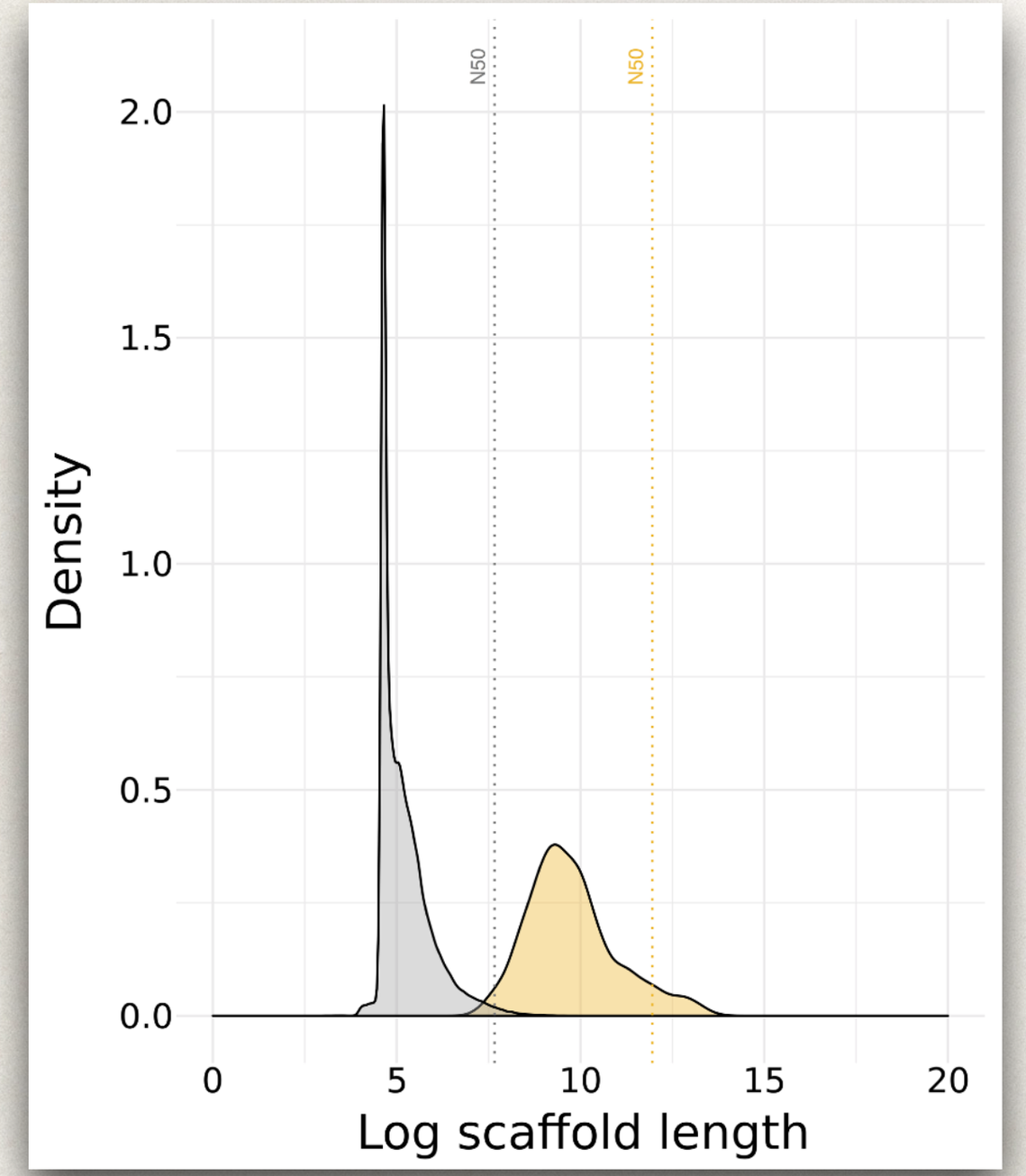
Summary of gen lies

N50: indicator of assembly quality
The higher, the better

	Hybrid genome assembly	Illumina genome assembly
Assembly features		
Number of contigs	923	46,821
Largest contig	926,516	26,836
Size (bp)	40,801,262	25,004,252
GC (%)	51.20	43.67
N50	156,193	659
N75	40,889	333
Number of contigs (>= 50,000 pb)	160	0

Díaz-Viraqué et al. 2019 *Genome Biology and Evolution*

Nanopore sequencing improves *T. cruzi* assembly contiguity and size



This improvement is evident looking at:

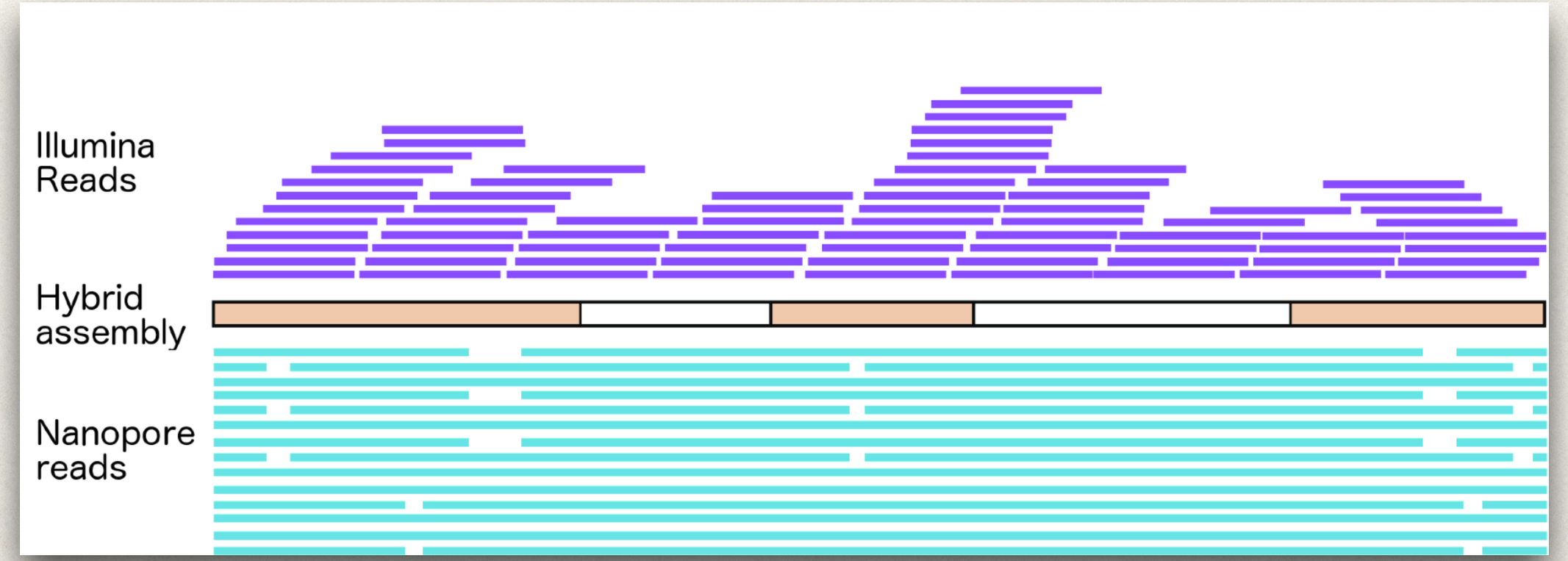
- # contig: ~47,000 in the Illumina assembly
  ~900 in the hybrid assembly
- Length distribution

Nanopore increase the assembly size in 16 Mb
(1 Mb = 1 millón pb)

Díaz-Viraqué et al. 2019 *Genome Biology and Evolution*

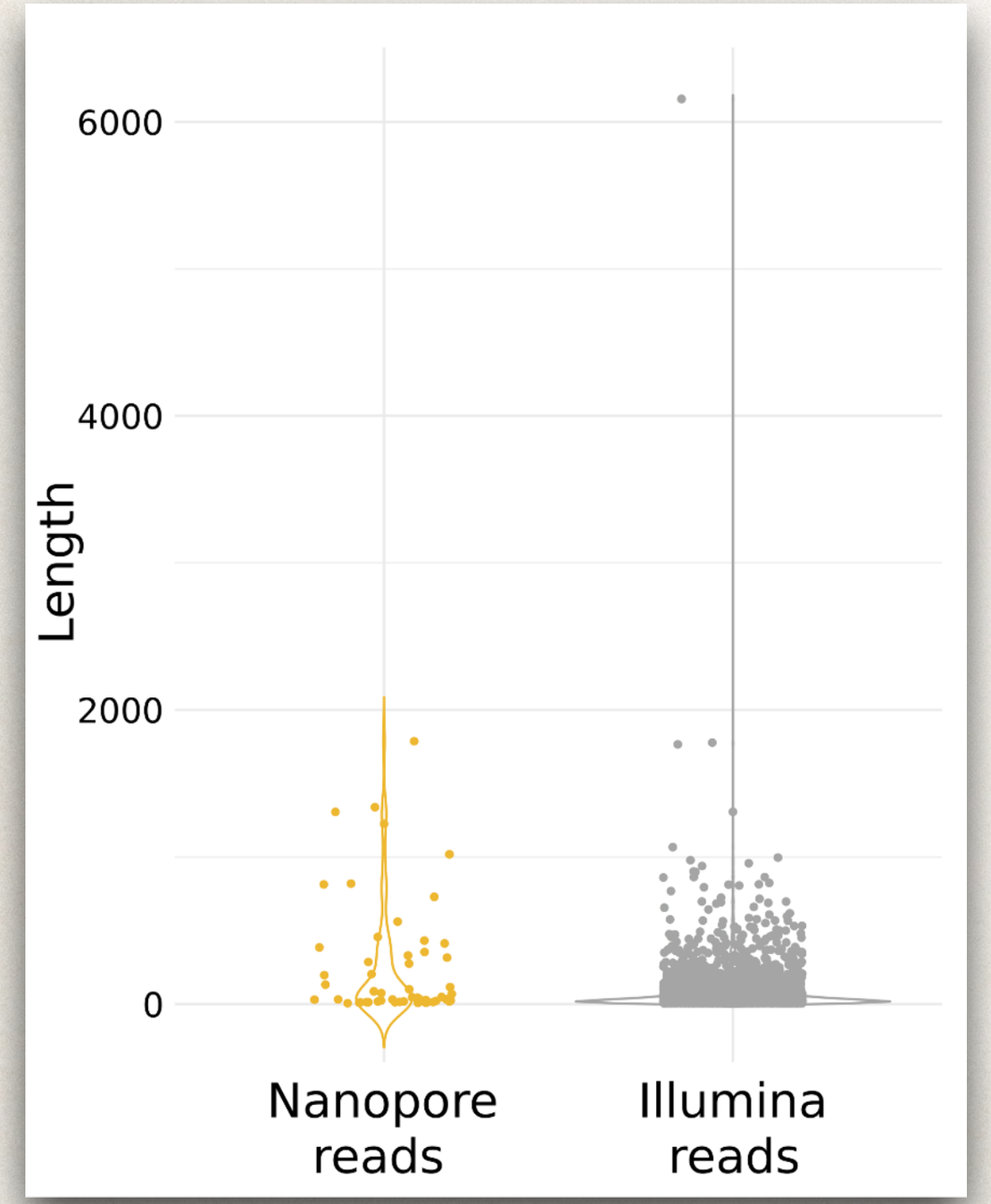
Nanopore reads close Illumina gaps

To evaluate the contribution of Illumina and Nanopore data to close gaps, we separately aligned both types of reads to the hybrid assembly.



Modified from Jaworski et al. 2017 *PLOS Pathogens*

Coverage zero regions



Coverage zero regions: no read alignment in at least 6 consecutive positions

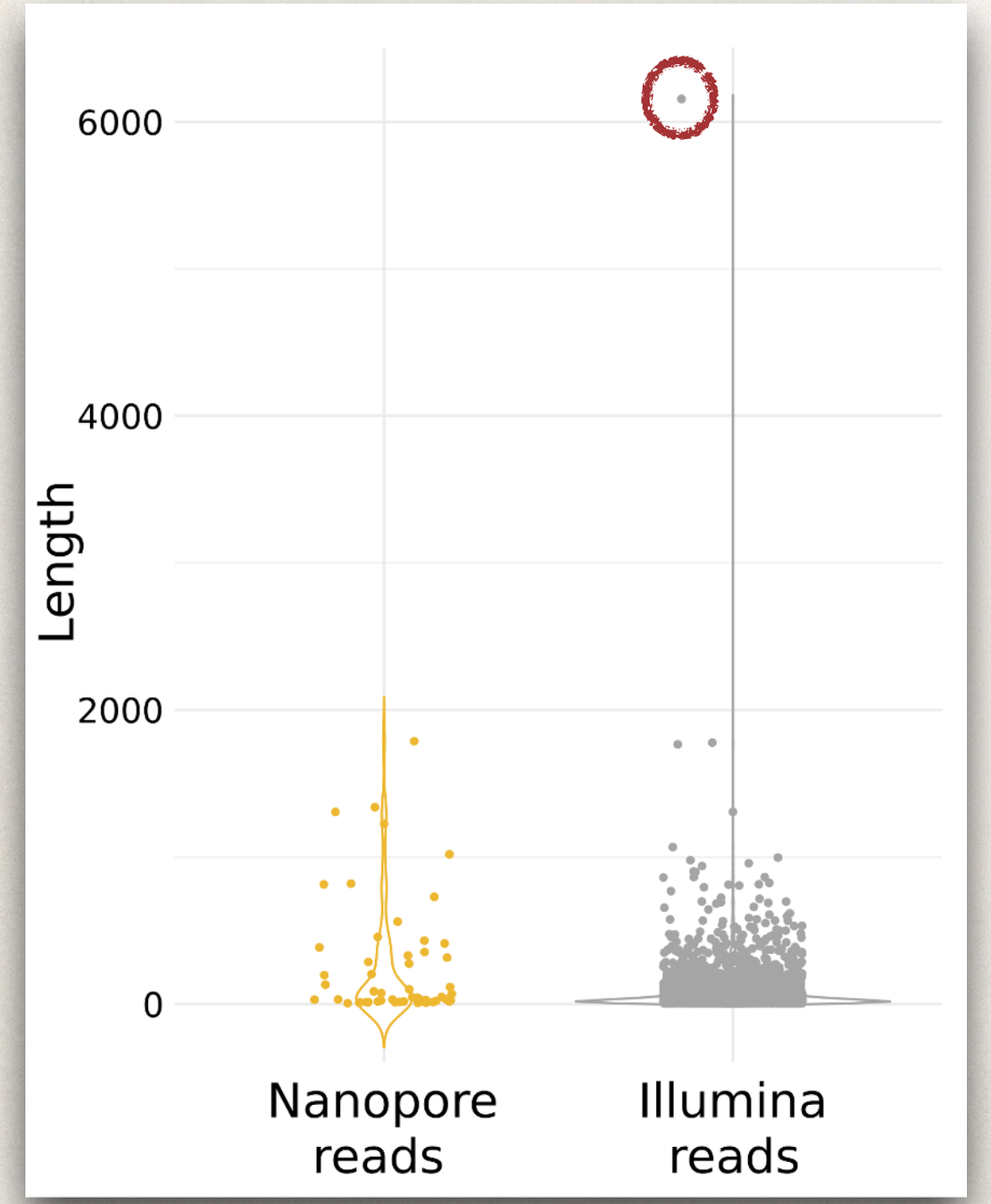
Each point represent a coverage cero region

The number of coverage zero region was higher when Illumina reads were alined (n=3624) compare to Nanopore reads (n=54)

The longest coverage zero region length was ~6,000 bp with Illumina reads while it decreased to ~2,000 bp with Nanopore reads.

Díaz-Viraqué et al. 2019 *Genome Biology and Evolution*

Coverage zero regions



Coverage zero regions: no read alignment in at least 6 consecutive positions

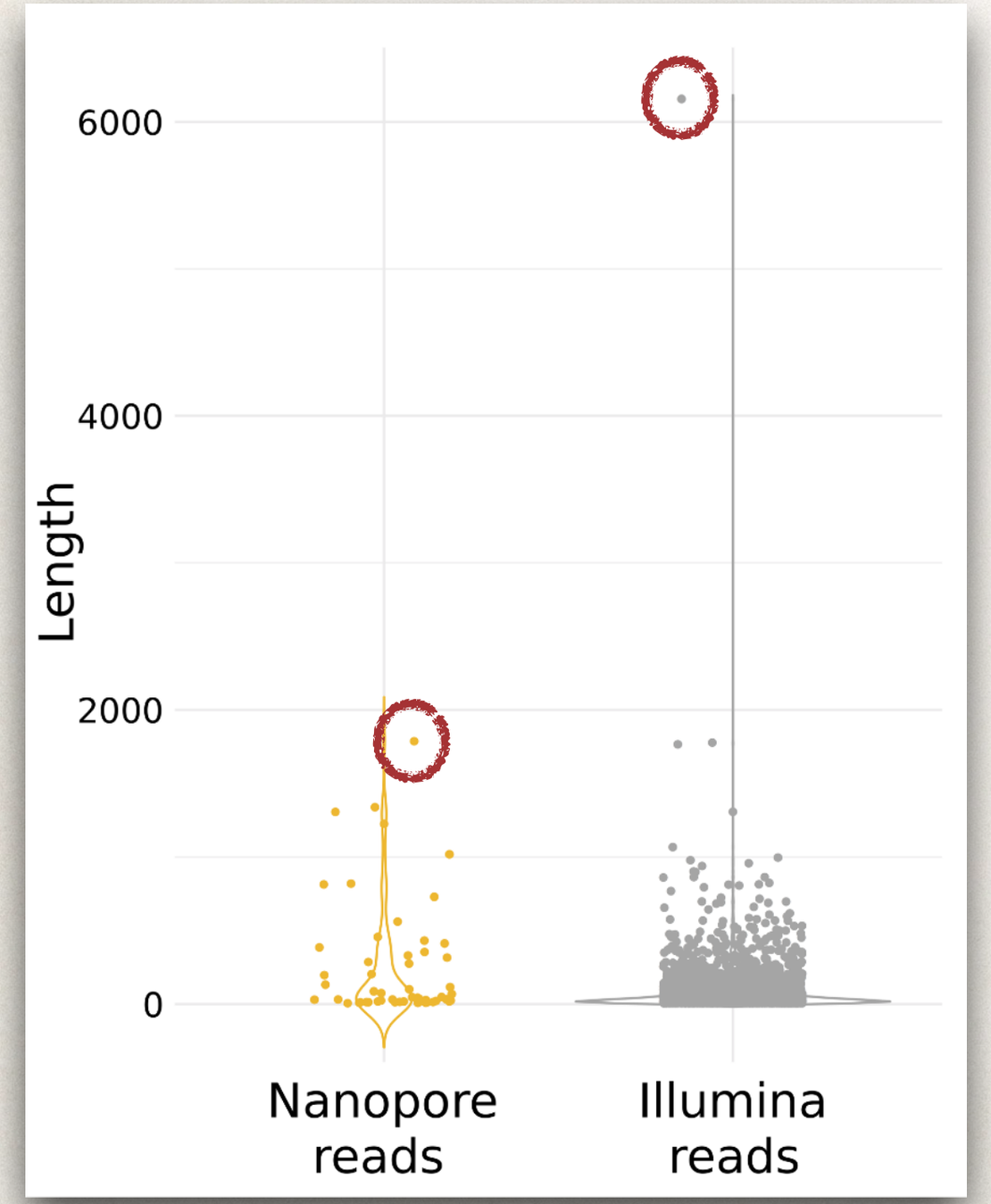
Each point represent a coverage cero region

The number of coverage zero region was higher when Illumina reads were alined (n=3624) compare to Nanopore reads (n=54)

The longest coverage zero region length was ~6,000 bp with Illumina reads while it decreased to ~2,000 bp with Nanopore reads.

Díaz-Viraqué et al. 2019 *Genome Biology and Evolution*

Coverage zero regions



Coverage zero regions: no read alignment in at least 6 consecutive positions

Each point represent a coverage cero region

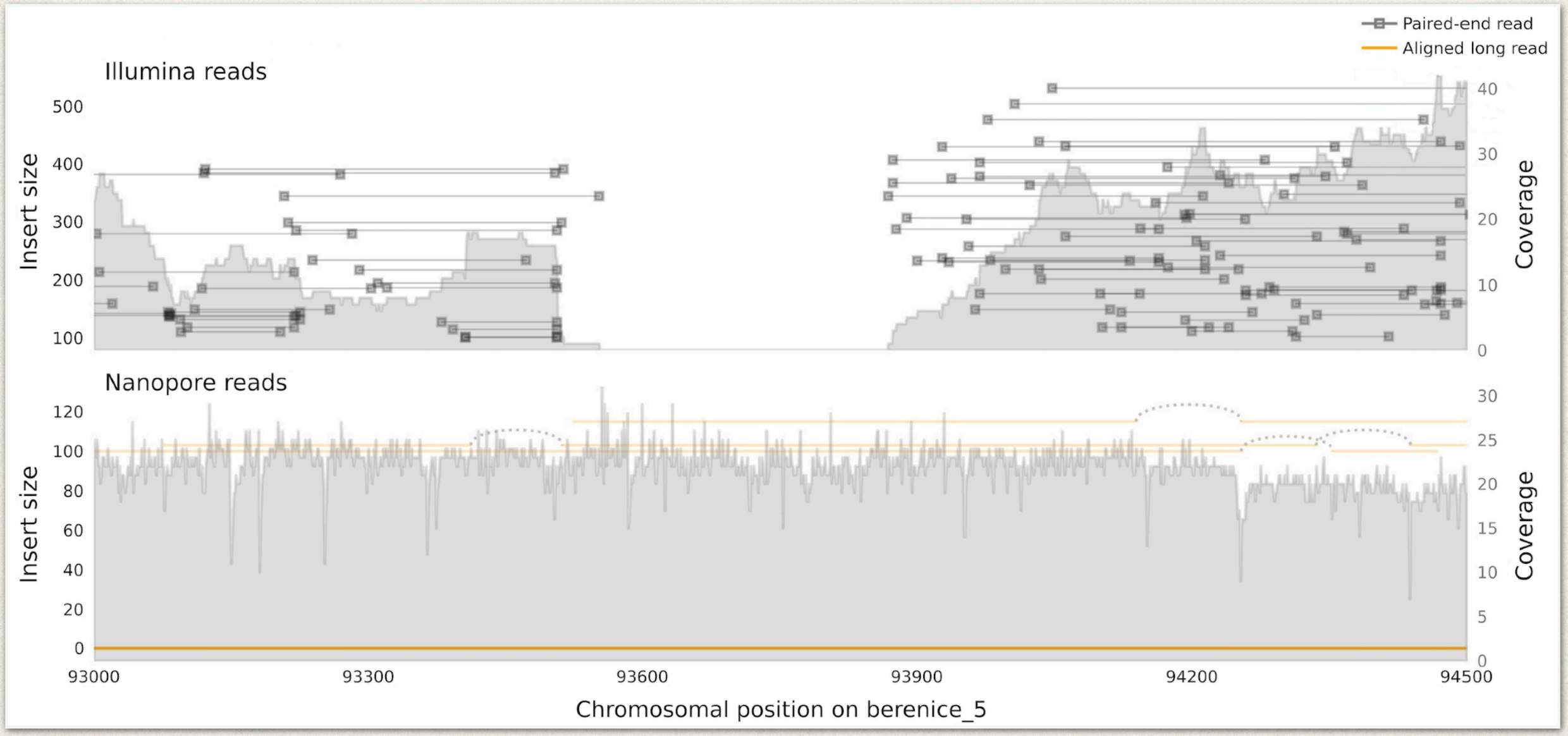
The number of coverage zero region was higher when Illumina reads were alined (n=3624) compare to Nanopore reads (n=54)

The longest coverage zero region length was ~6,000 bp with Illumina reads while it decreased to ~2,000 bp with Nanopore reads.

Díaz-Viraqué et al. 2019 *Genome Biology and Evolution*

Example of a coverage zero region

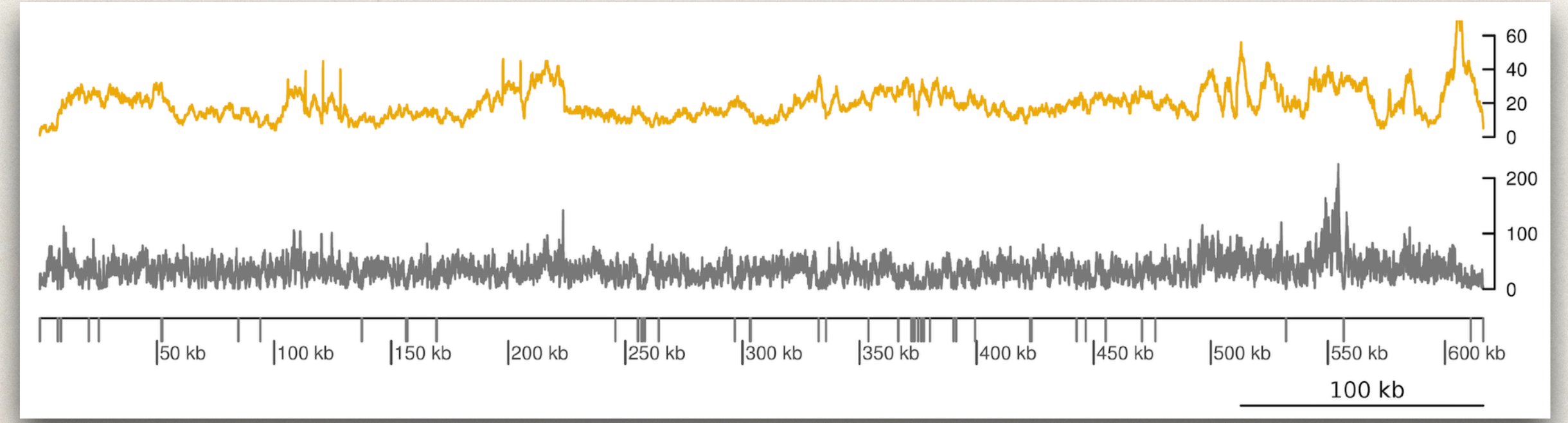
Nanopore & Illumina reads mapped to hybrid assembly



Nanopore reads uninterruptedly cover this genomic segment with a depth of ~20x while Illumina reads fail to resolve a region where coverage falls to zero, causing the break of contiguity in the assembly.

Díaz-Viraqué et al. 2019 *Genome Biology and Evolution*

Nanopore reads close Illumina gaps



There are a lot of zero regions in this contig when Illumina reads were mapped to hybrid assembly

Díaz-Viraqué et al. 2019 *Genome Biology and Evolution*

Nanopore reads improve completeness of repeated regions

We also evaluated the resolution of repeated regions analyzing the annotation of retroelements in both assemblies

Given that retroelements are very long (>5 Kb), if the genome is fragmented I expect that it would be difficult to find them complete.

Transposable elements	Hybrid assembly	Illumina assembly
CZAR	27	0
L1Tc	38	0
VIPER	50	0
NARTc	54	4
SIRE	80	0

Repeated regions are completely fragmented in the Illumina assembly

Díaz-Viraqué et al. 2019 *Genome Biology and Evolution*

Comparison with other “long read genomes”

	Sequencing method	Size (Mb)	GC (%)	Number of contigs	N50	Coverage	BUSCO eukaryota
Dm28c (TcI)	PacBio	53.16	51.6	599	317,638	76 X	202
Berenice (TcII)	Illumina Nanopore	40.80	51.2	923	156,193	41 X Illumina 28 X Nanopore	206
Bug2148 (TcV)	PacBio	55.22	51.63	934	196,760	68 X	208
TCC (TcVI)	PacBio	86.77	51.7	1142	265,169	60 X	210

Berenice becomes the best-quality assembly available for a member of TcII contributing to expand the known genetic diversity of *T. cruzi* and facilitating more accurate evolutionary inferences.

Díaz-Viraqué et al. 2019 *Genome Biology and Evolution*

References

- Giani AM, Gallo GR, Gianfranceschi L, Formenti G (2020) Long walk to genomics: History and current approaches to genome sequencing and assembly. Computational and Structural Biotechnology Journal, 18, 9-19.
- Rhoads A, Au KF. (2015). PacBio sequencing and its applications. Genomics, proteomics & bioinformatics, 13(5), 278-289.
- Pita S, Díaz-Viraqué F, Iraola G, Robello C. (2019). The Tritryps comparative repeatome: insights on repetitive element evolution in Trypanosomatid pathogens. Genome biology and evolution, 11(2), 546-551.
- Díaz-Viraqué F, Pita S, Greif G, De Souza RDCM, Iraola G, Robello C. (2019). Nanopore sequencing significantly improves genome assembly of the protozoan parasite Trypanosoma cruzi. Genome biology and evolution, 11(7), 1952-1957.
- Berná L, Rodríguez M, Chiribao ML, Parodi-Talice A, Pita S, Rijo G, ... & Robello, C. (2018). Expanding an expanded genome: long-read sequencing of Trypanosoma cruzi. Microbial Genomics, 4(5).
- PabBio & Oxford Nanopore web pages

Obrigada!

Questions??