INTRODUCTION AU SÉQUENÇAGE À HAUT DÉBIT POUR LA GÉNOMIQUE

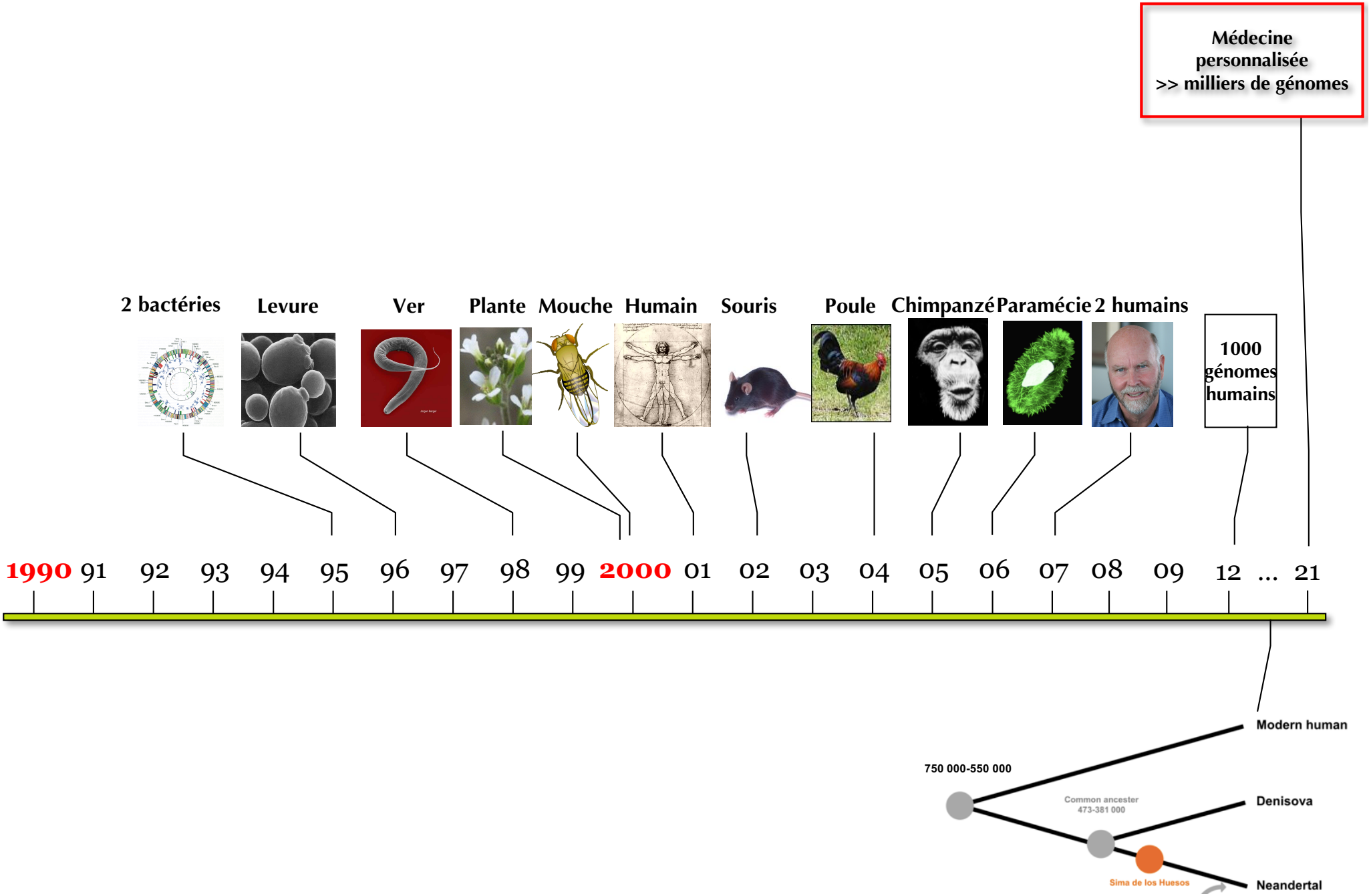
Claude THERMES

INSTITUT DE BILOGIE INTÉGRATIVE DE LA CELLULE – I2BC

GIF-SUR-YVETTE

Diplôme Universitaire en Bioinformatique intégrative - DU-Bii - 08/03/2021

Premiers génomes entièrement séquencés

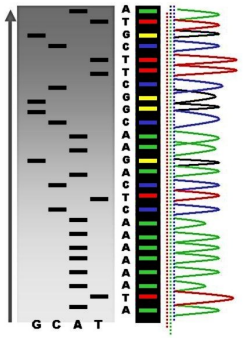


1rst generation : Sanger sequencing

- Has been the major method up to 2005

Limitations

- Extremely high cost
- Long experimental set up times
- High DNA concentrations needed



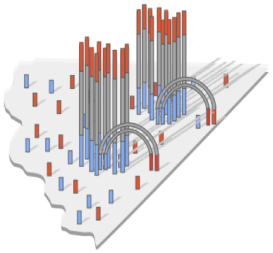
2d generation

- Single DNA molecules replicated in clusters
- Very high troughput

Limitations

- Maximum read length ≤ 300bp

Illumina



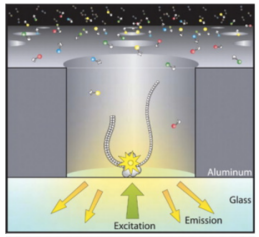
3rd generation

- Single molecules sequencing
- Very long reads

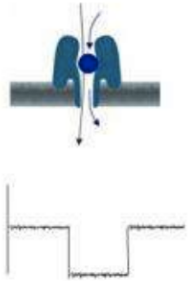
Limitations

- High error rates

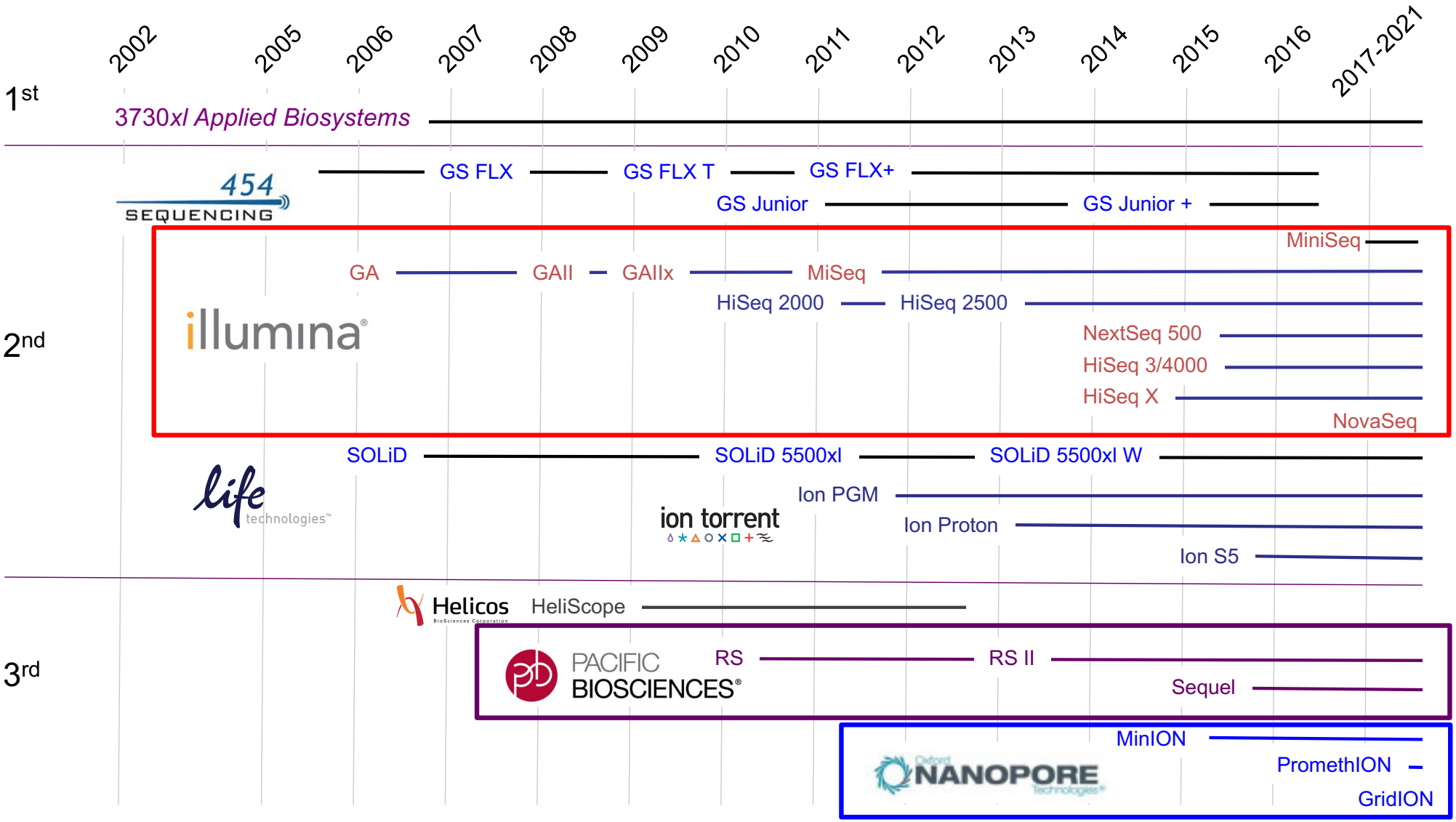
PacBio



Oxford Nanopore



Sequencing technologies



Illumina : the winning technology

MiniSeq
25 million reads

MiSeq
25 millions reads, 2 x 300 bp

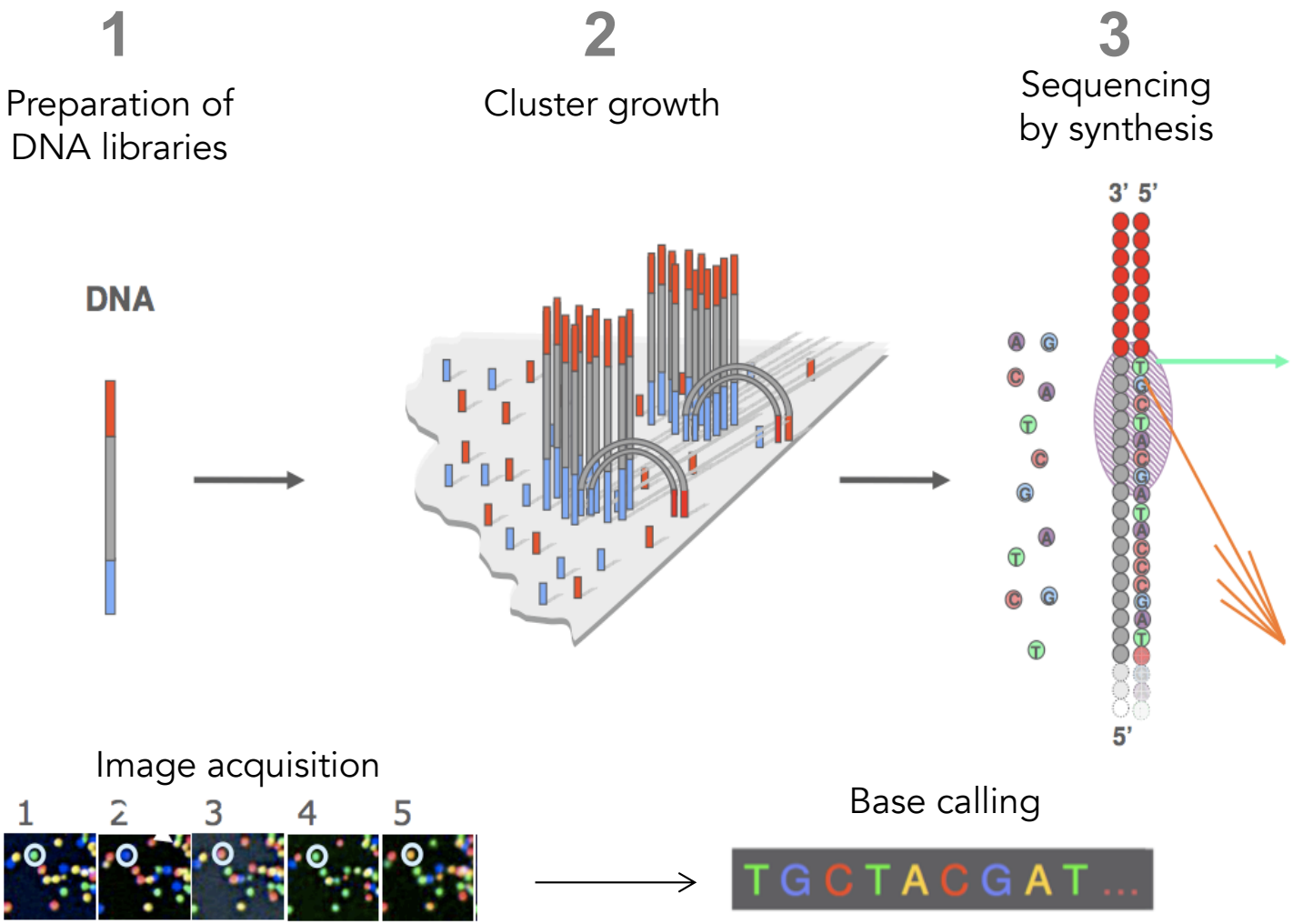
NextSeq
400 million reads

HisSeq 4000
5 billion reads

HisSeq X
6 billion reads

NovaSeq 6000
20 billion reads

General scheme of Illumina sequencing



1 - Preparation of DNA-seq Libraries

Illumina TruSeq technology

Genomic DNA

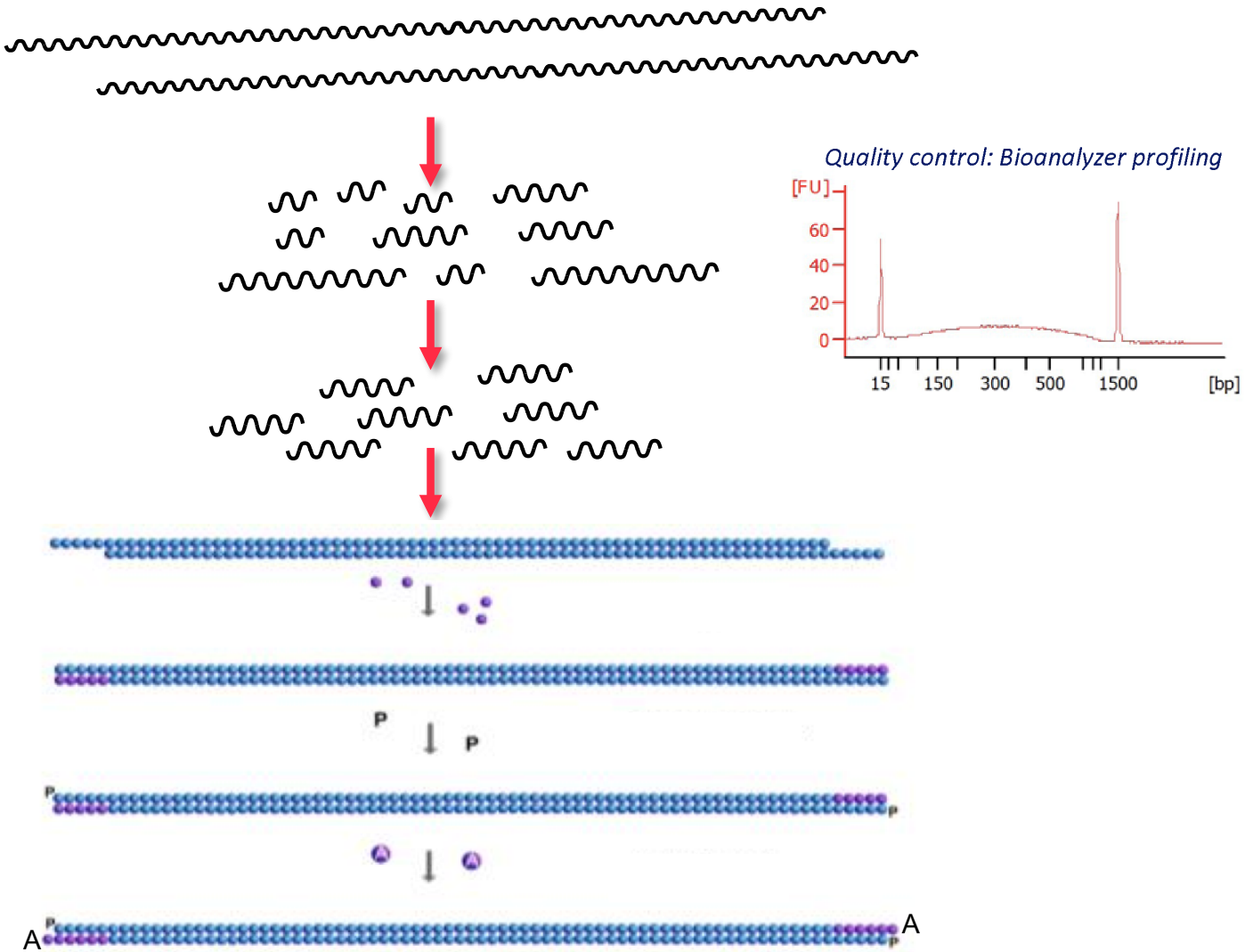
Sonication

Size selection

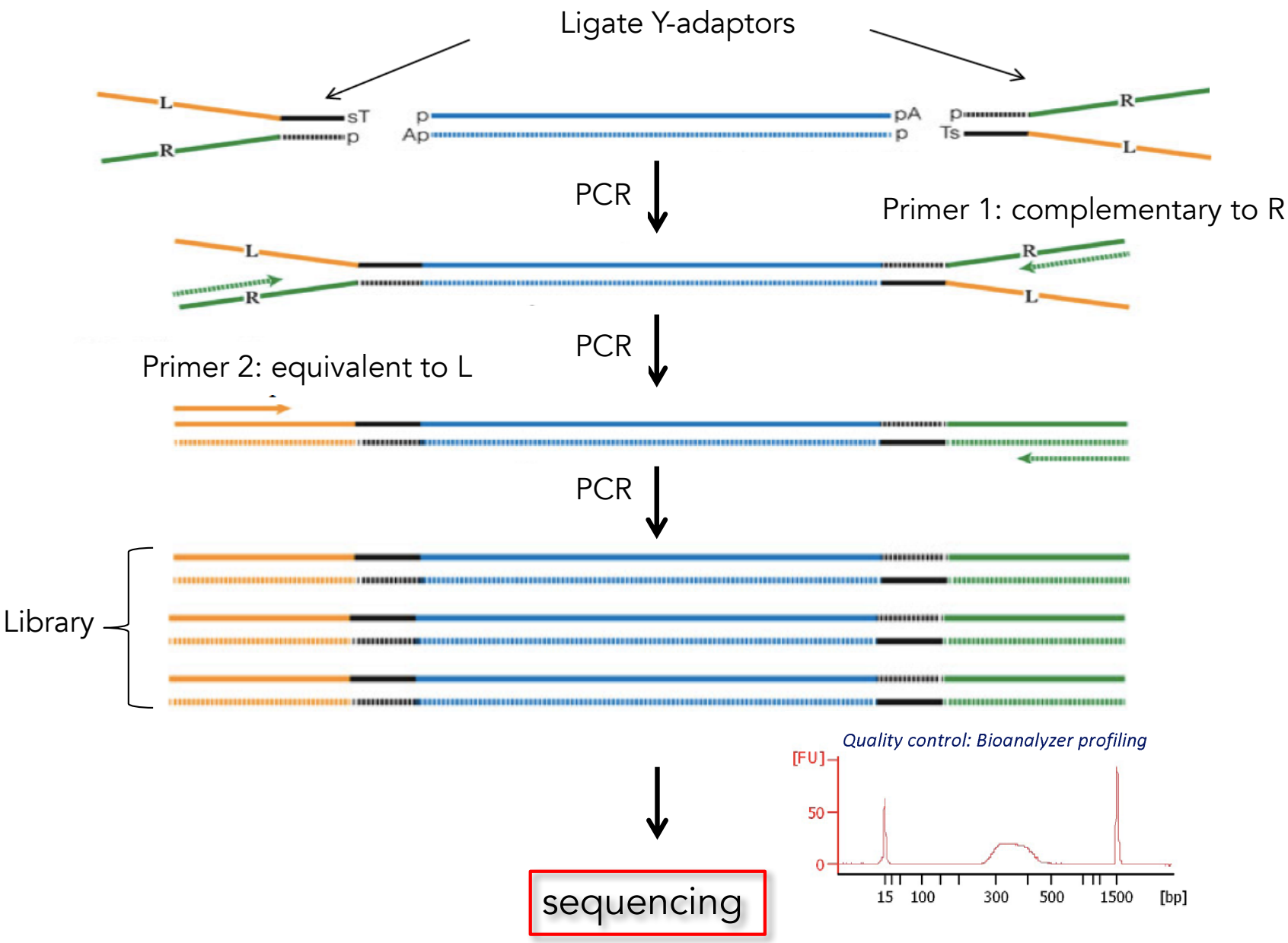
End repair

Phosphorylation

A - overhang



1 - Preparation of DNA-seq Libraries



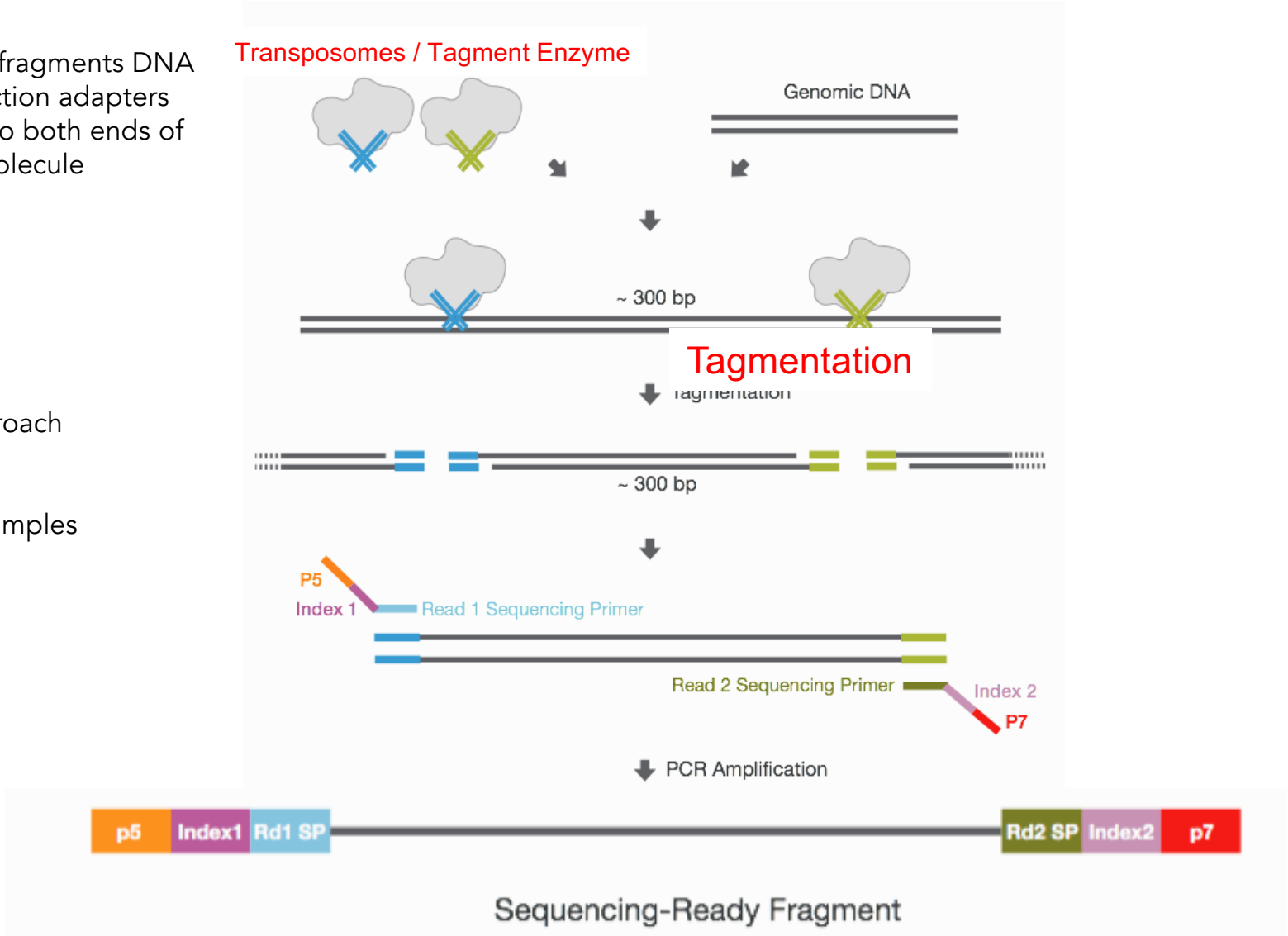
1 - Preparation of DNA-seq Libraries

Nextera "tagmentation"

Tagment Enzyme fragments DNA and attaches junction adapters (blue and green) to both ends of the tagmented molecule

Dual barcode approach

up to 96 indexed samples



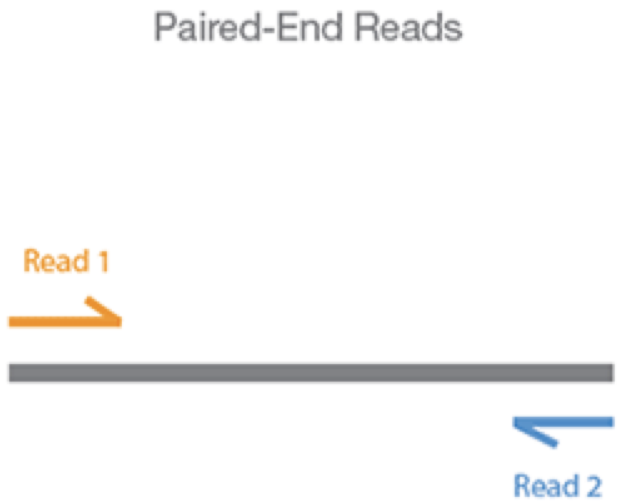
requires small quantities 1ng (bacteria) to 50 ng (human)

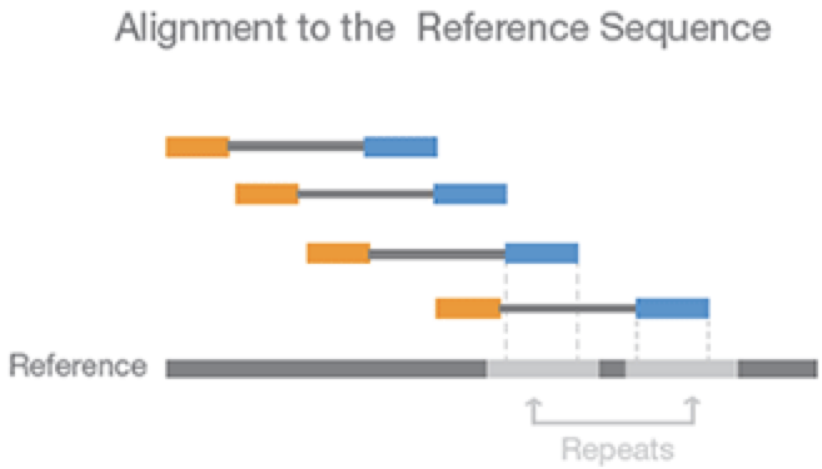
1 - Preparation of DNA-seq Libraries

SINGLE READ and PAIRED-END SEQUENCING

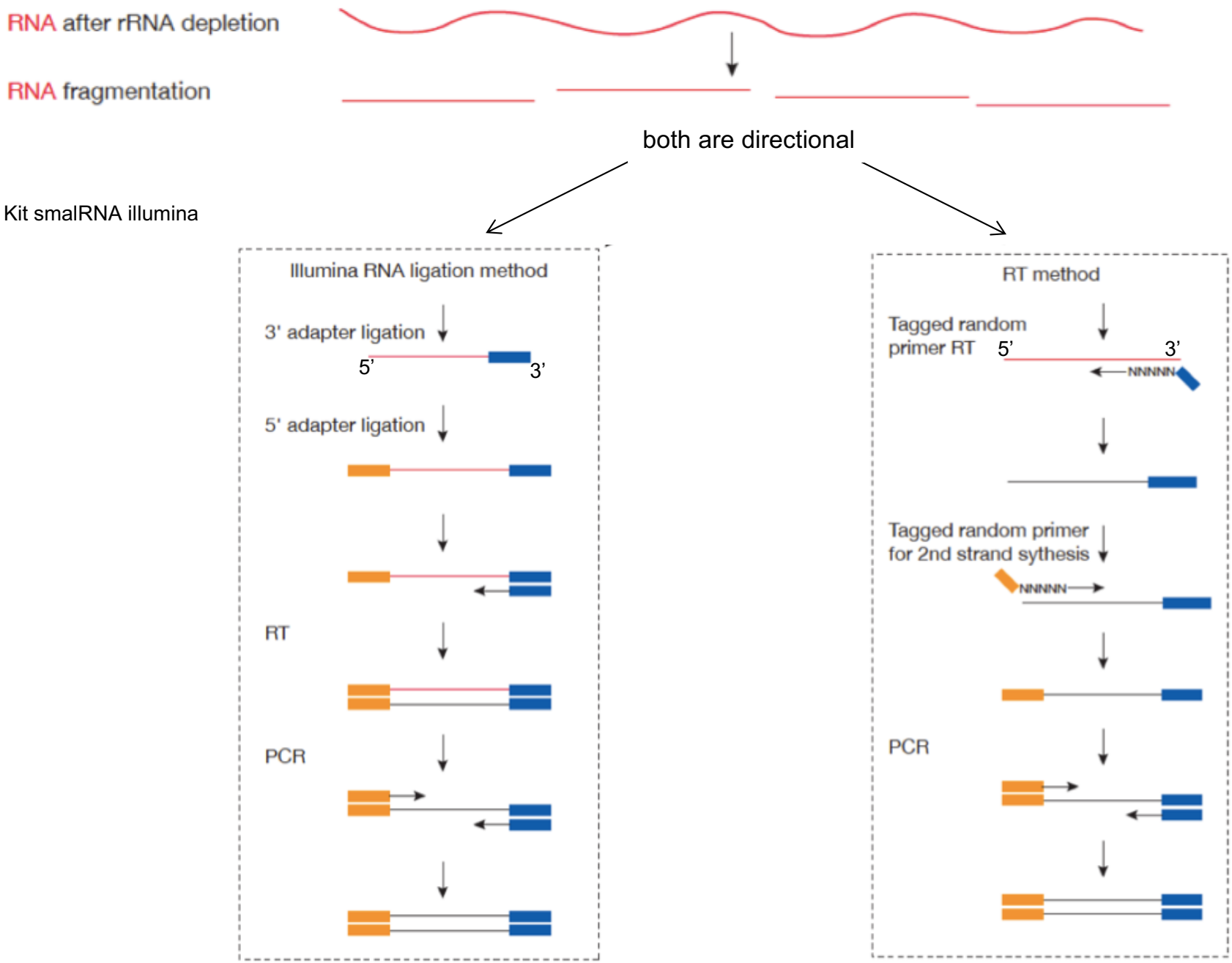
- **Single end**: Sequence one physical end of DNA fragment

- **Paired End:** Sequence both physical ends of DNA fragment
  - End distance: < 800nt





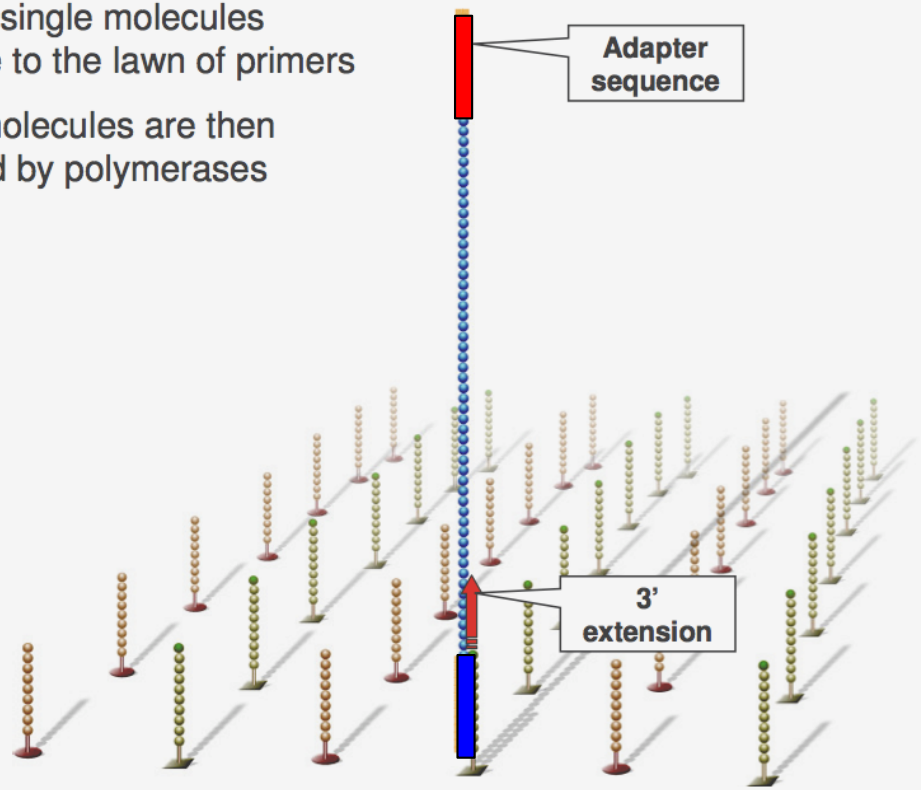
1 - Preparation of RNA-seq Libraries



2 – Cluster growth

Cluster Generation:
Hybridize Fragment & Extend

- > 100 M single molecules hybridize to the lawn of primers
- Bound molecules are then extended by polymerases



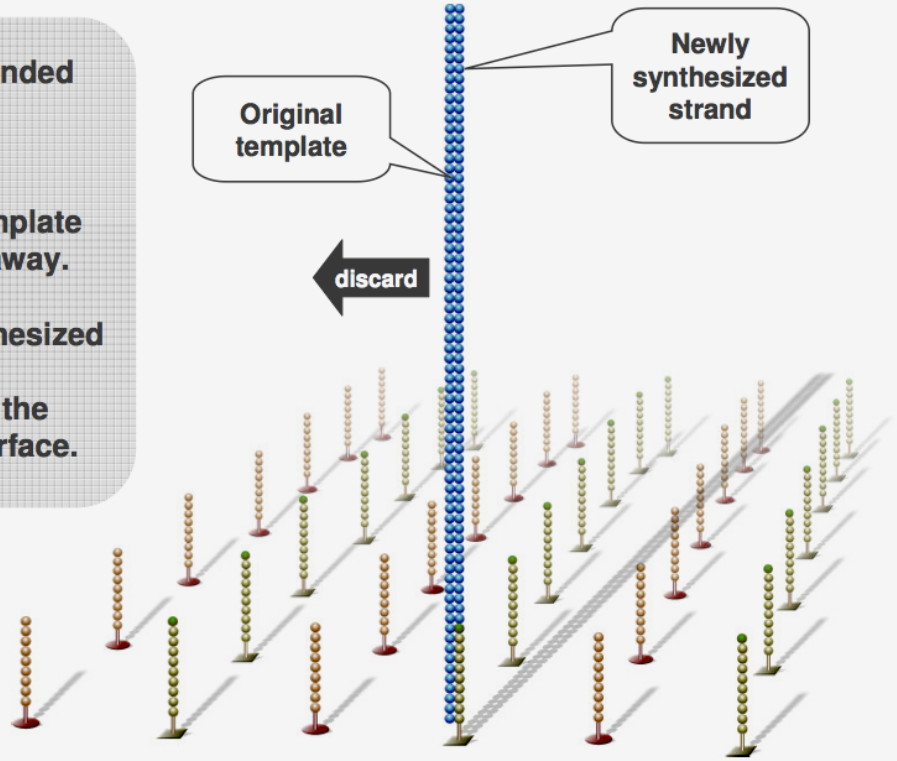
2 – Cluster growth

Cluster generation:
Denature double-stranded DNA

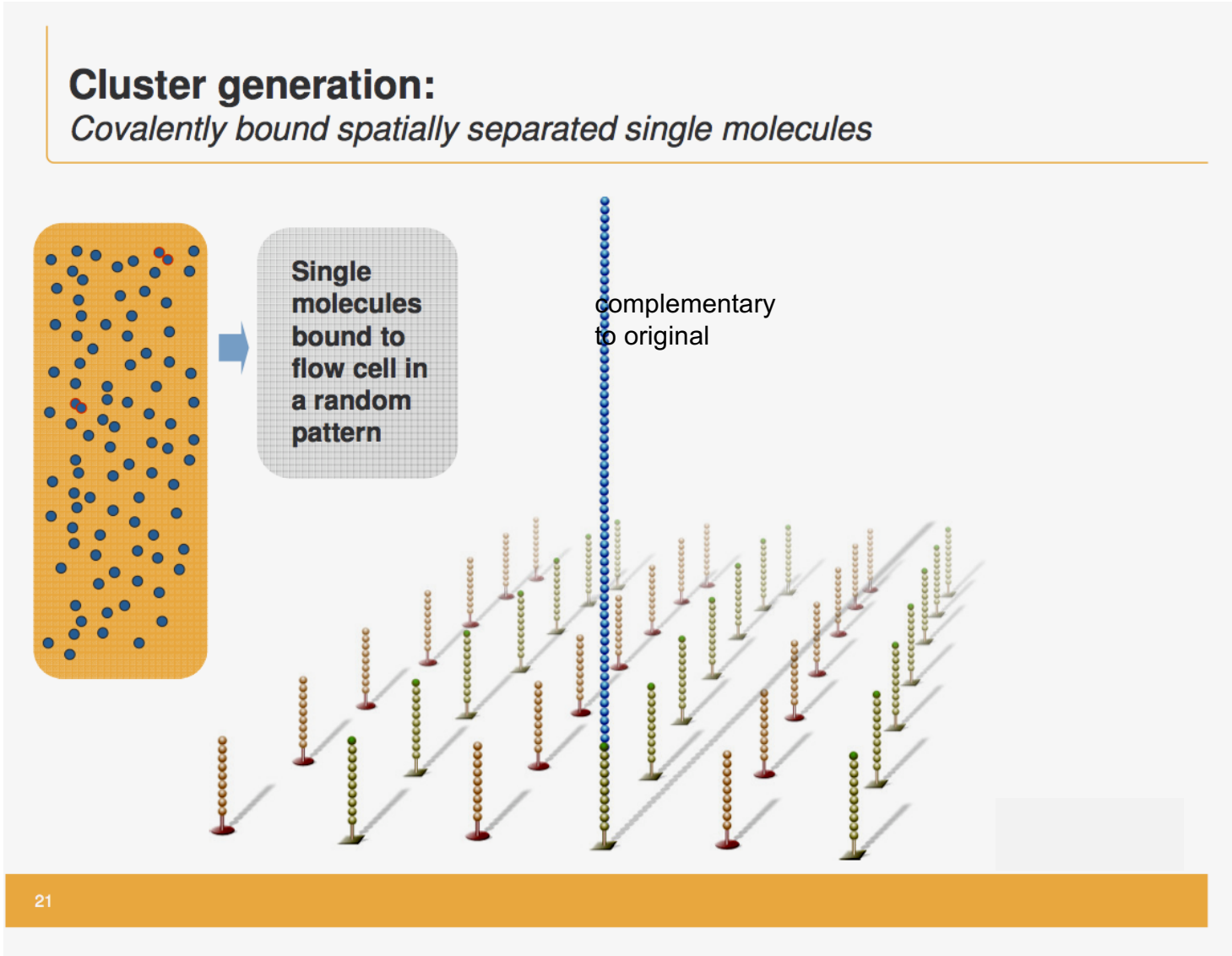
Double-stranded molecule is denatured.

Original template is washed away.

Newly synthesized covalently attached to the flow cell surface.



2 – Cluster growth

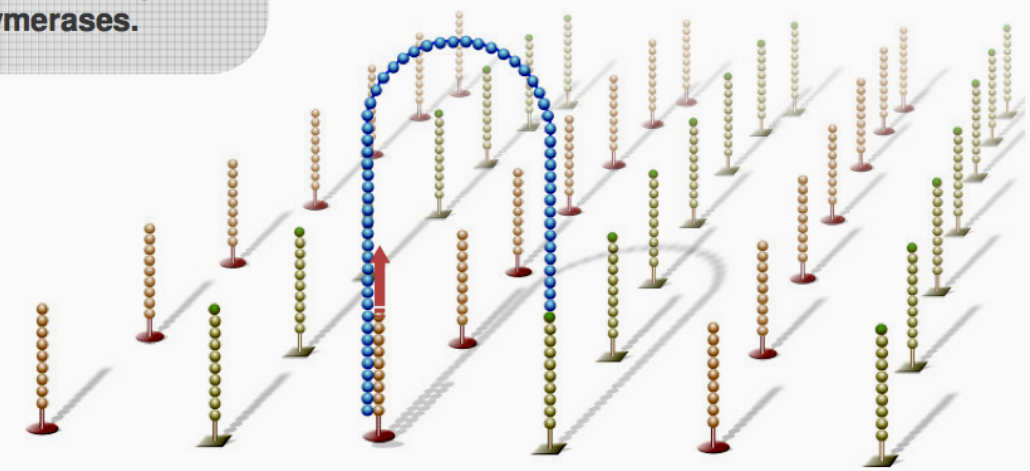


2 – Cluster growth

Cluster generation:
Bridge amplification

Single-strand flips over to hybridize to adjacent primers to form a bridge.

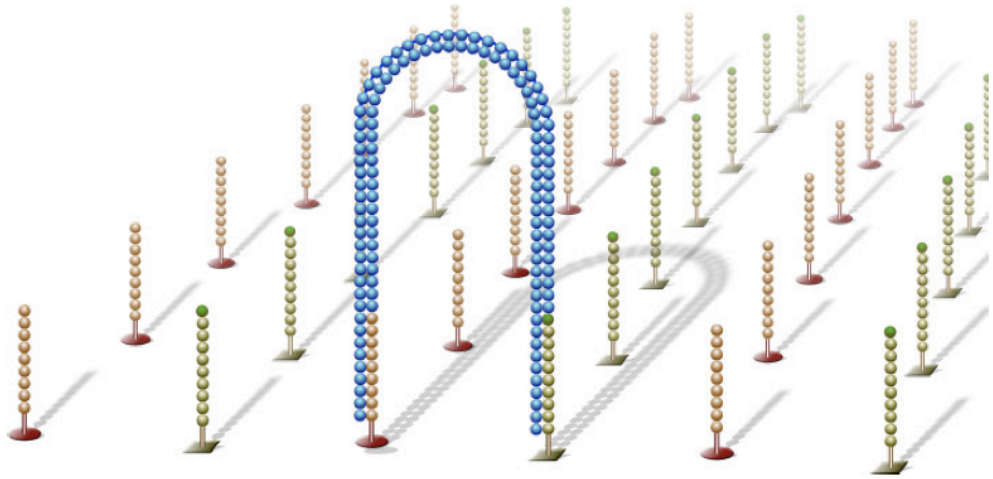
Hybridized primer is extended by polymerases.



2 – Cluster growth

Cluster generation:
Bridge amplification

→ double-stranded bridge is formed.

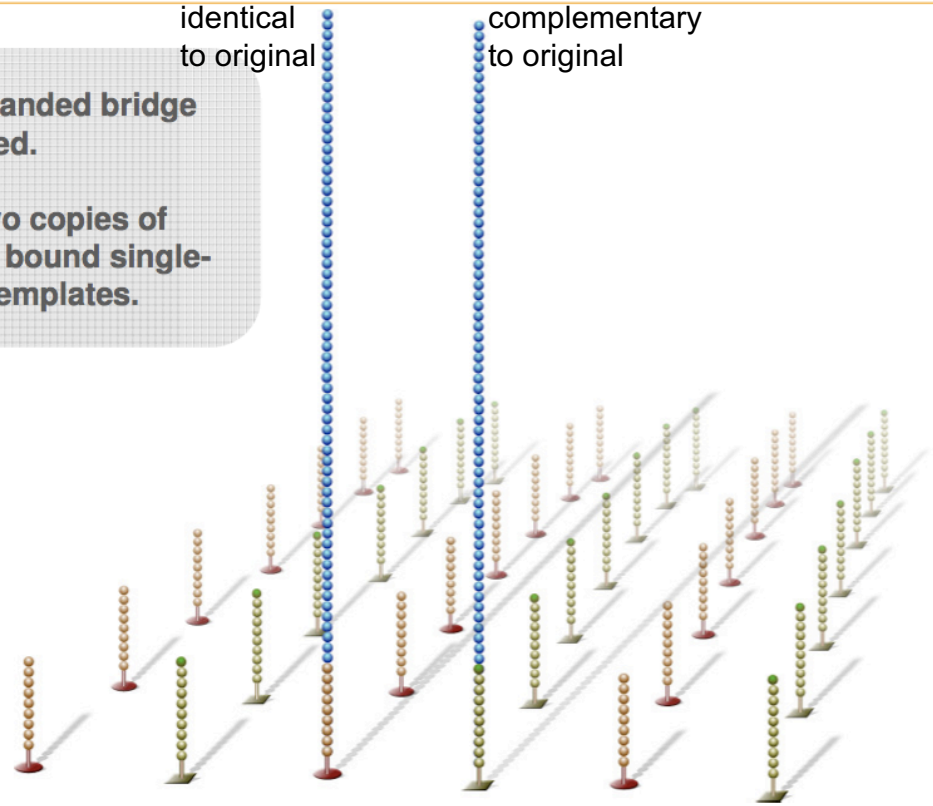


2 – Cluster growth

Cluster generation:
Bridge amplification

Double-stranded bridge is denatured.

Result: Two copies of covalently bound single-stranded templates.

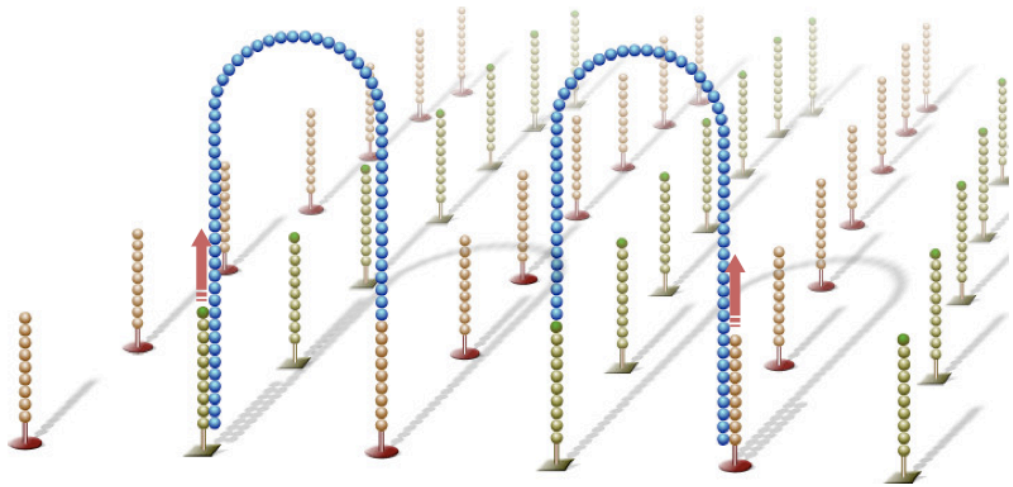


2 – Cluster growth

Cluster generation:
Bridge amplification

Single-strands flip over to hybridize to adjacent primers to form bridges.

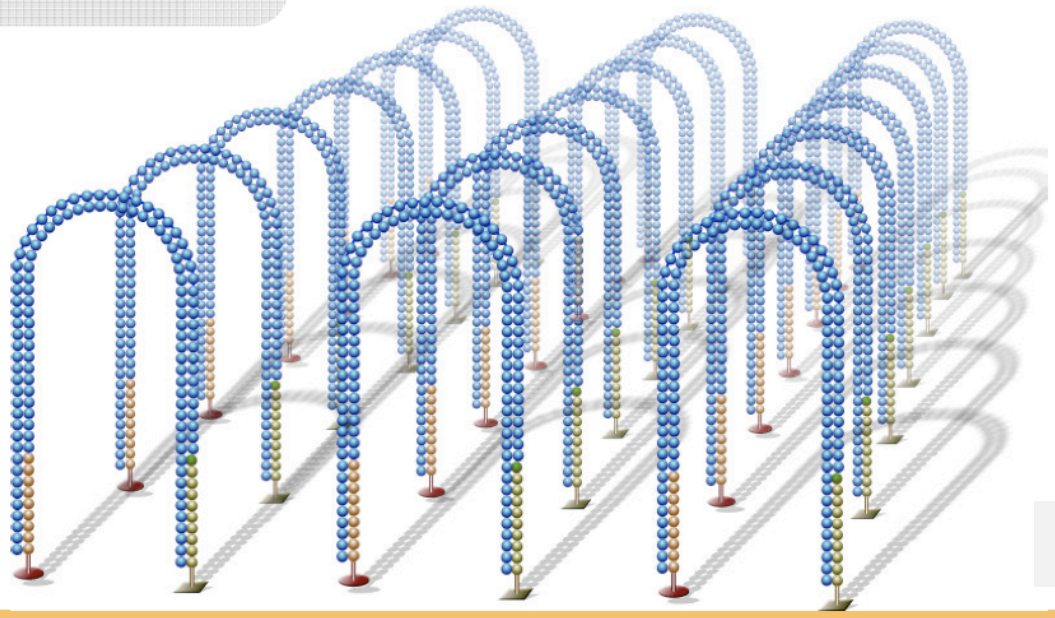
Hybridized primer is extended by polymerase.



2 – Cluster growth

Cluster generation:
Bridge amplification

Bridge amplification cycle repeated till multiple bridges are formed

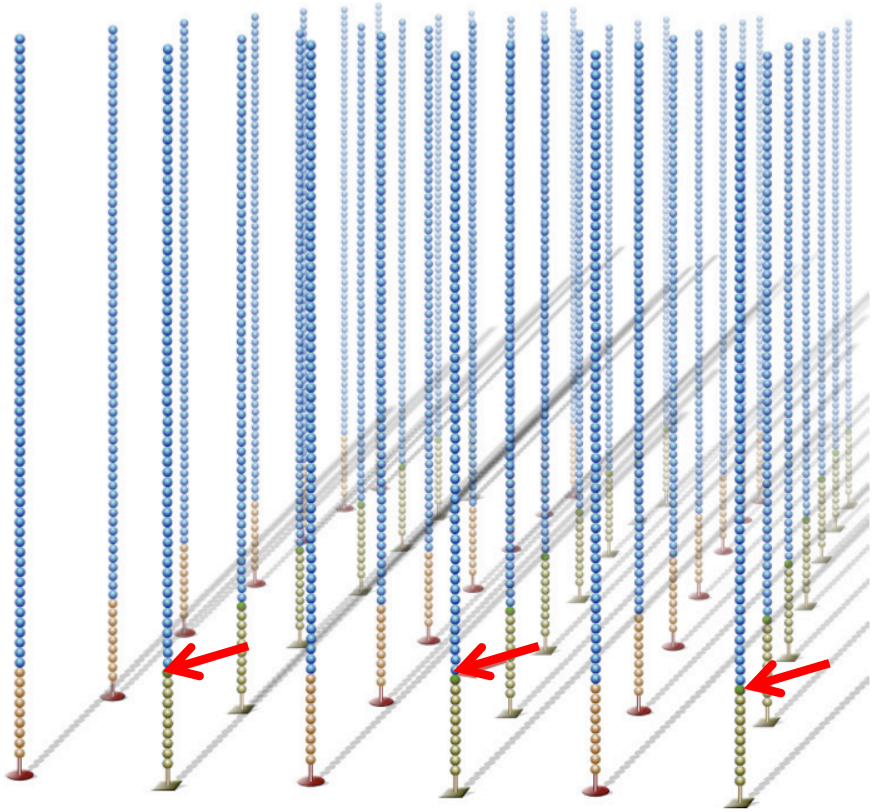


2 – Cluster growth

Cluster generation

dsDNA bridges denatured.

Reverse strands cleaved and washed away.

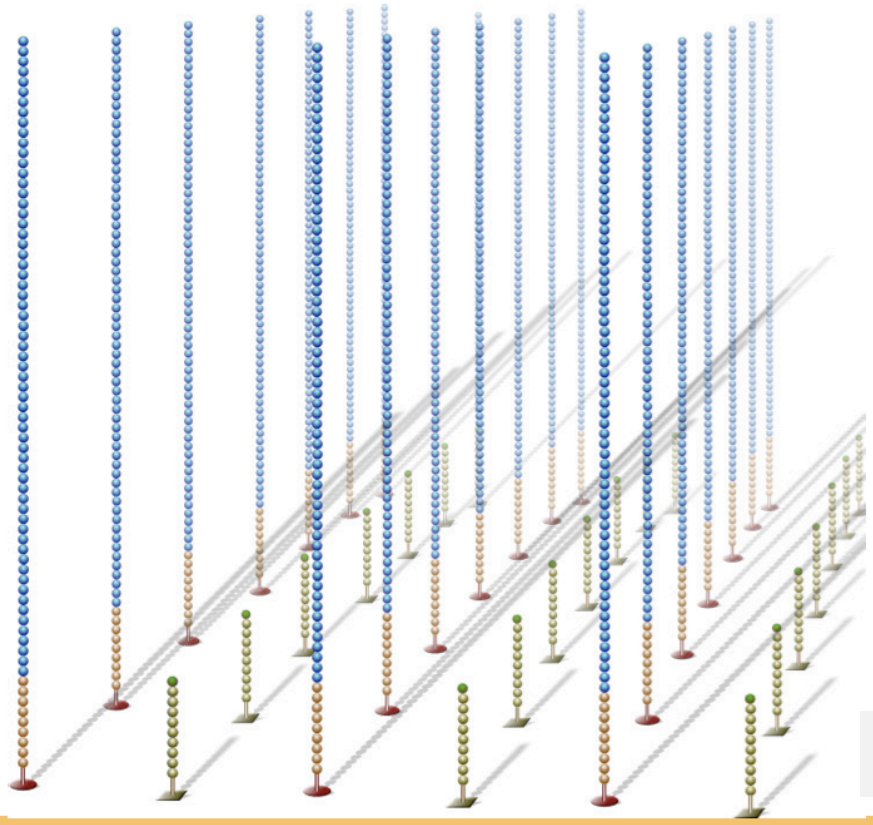


Cleavage of a chemically modified nucleotide

2 – Cluster growth

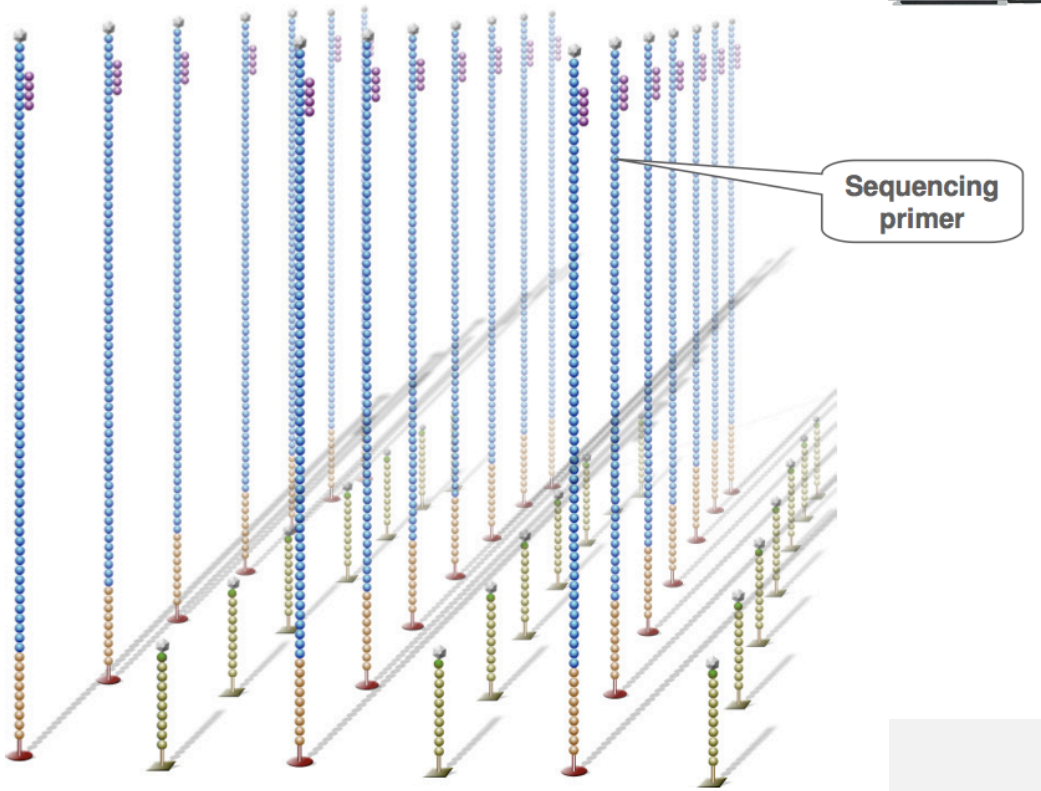
Cluster generation

... leaving a cluster with forward strands only.

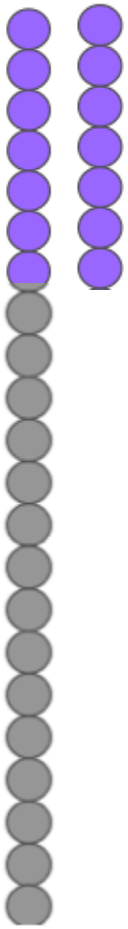


3 – Sequencing

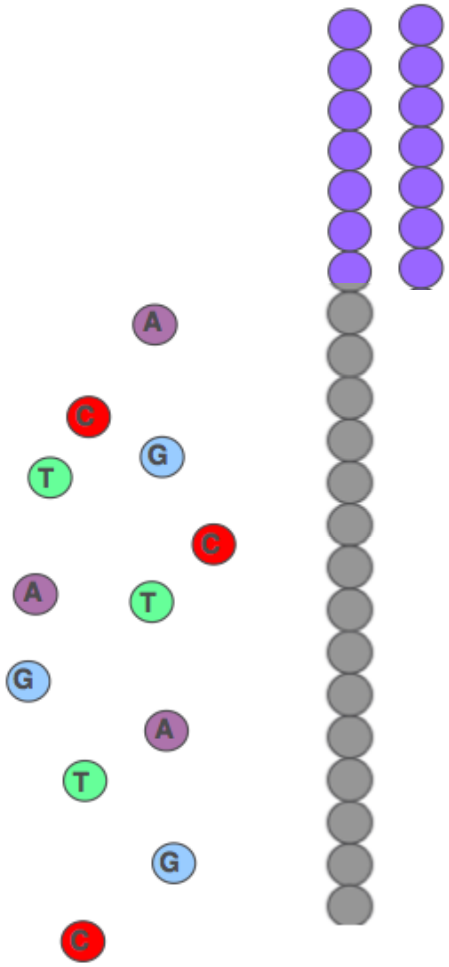
Sequencing primer is hybridized to adapter sequence.



3 - Sequencing By Synthesis (SBS)

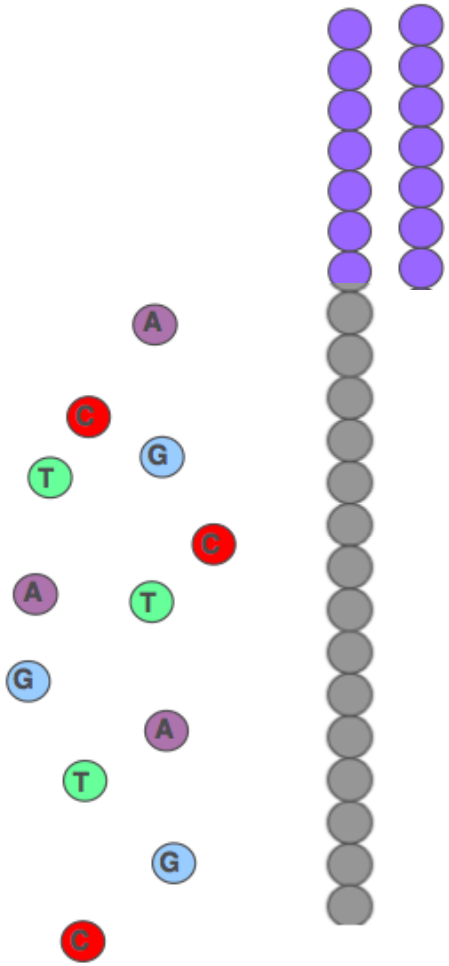


3 - Sequencing By Synthesis (SBS)



fluorescent-labeled terminator bound to each dNTP

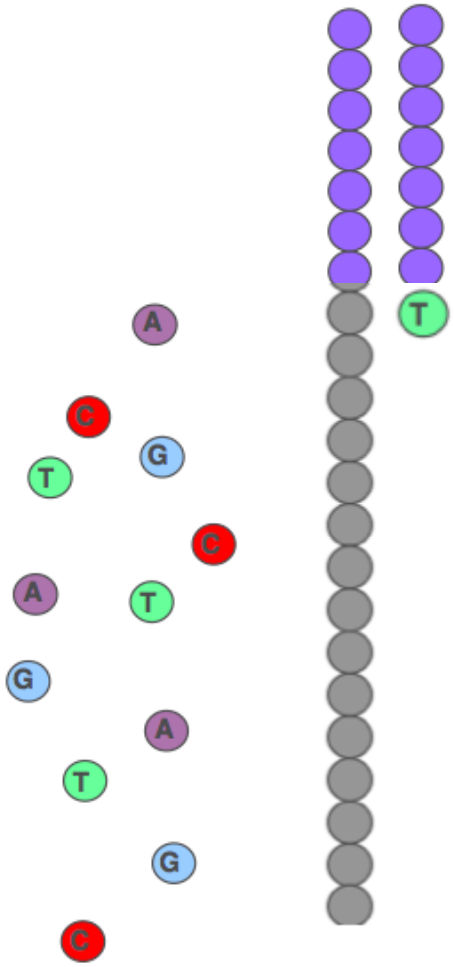
3 - Sequencing By Synthesis (SBS)



fluorescent-labeled terminator bound to each dNTP

Cycle 1 : add sequencing reagents

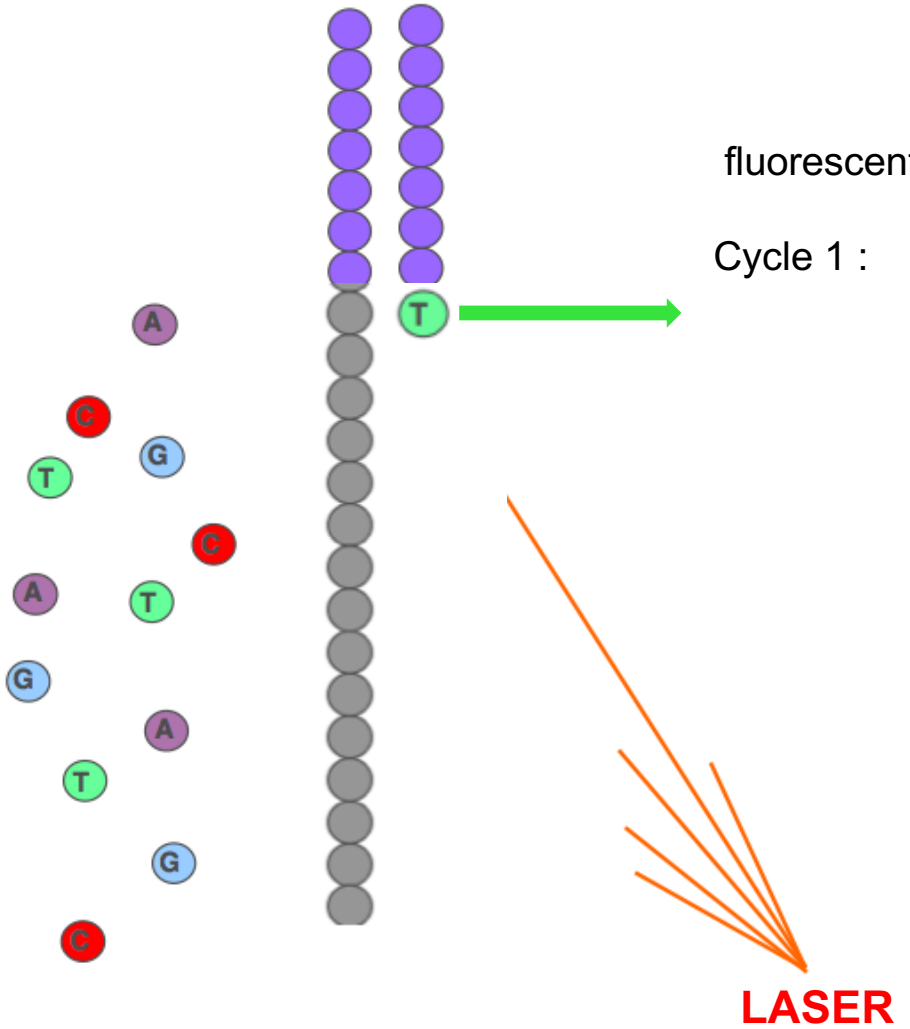
3 - Sequencing By Synthesis (SBS)



fluorescent-labeled terminator bound to each dNTP

Cycle 1 : add sequencing reagents
first base incorporated

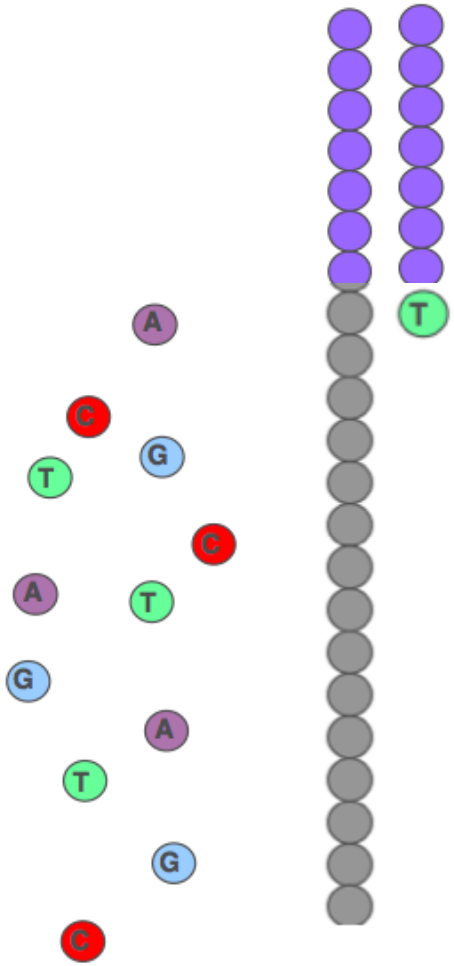
3 - Sequencing By Synthesis (SBS)



fluorescent-labeled terminator bound to each dNTP

Cycle 1 : add sequencing reagents
first base incorporated
detect signal / scan

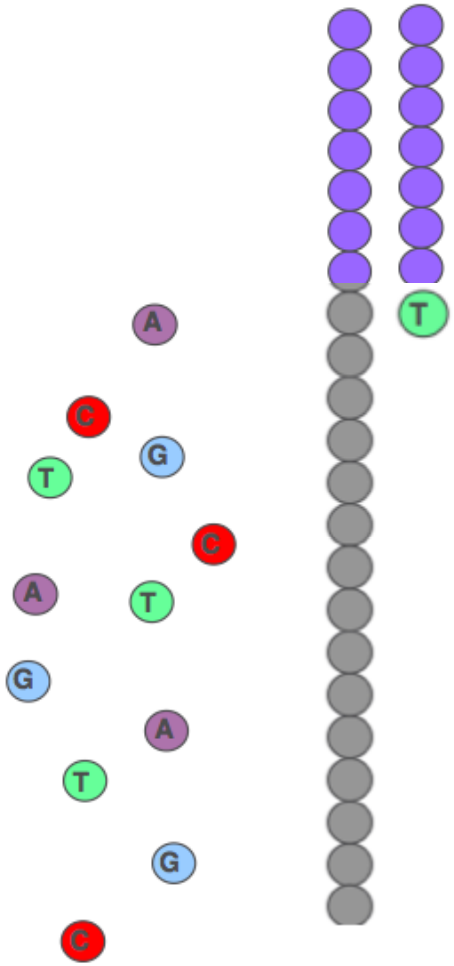
3 - Sequencing By Synthesis (SBS)



fluorescent-labeled terminator bound to each dNTP

Cycle 1 : add sequencing reagents
first base incorporated
detect signal / scan
cleave terminator and dye

3 - Sequencing By Synthesis (SBS)

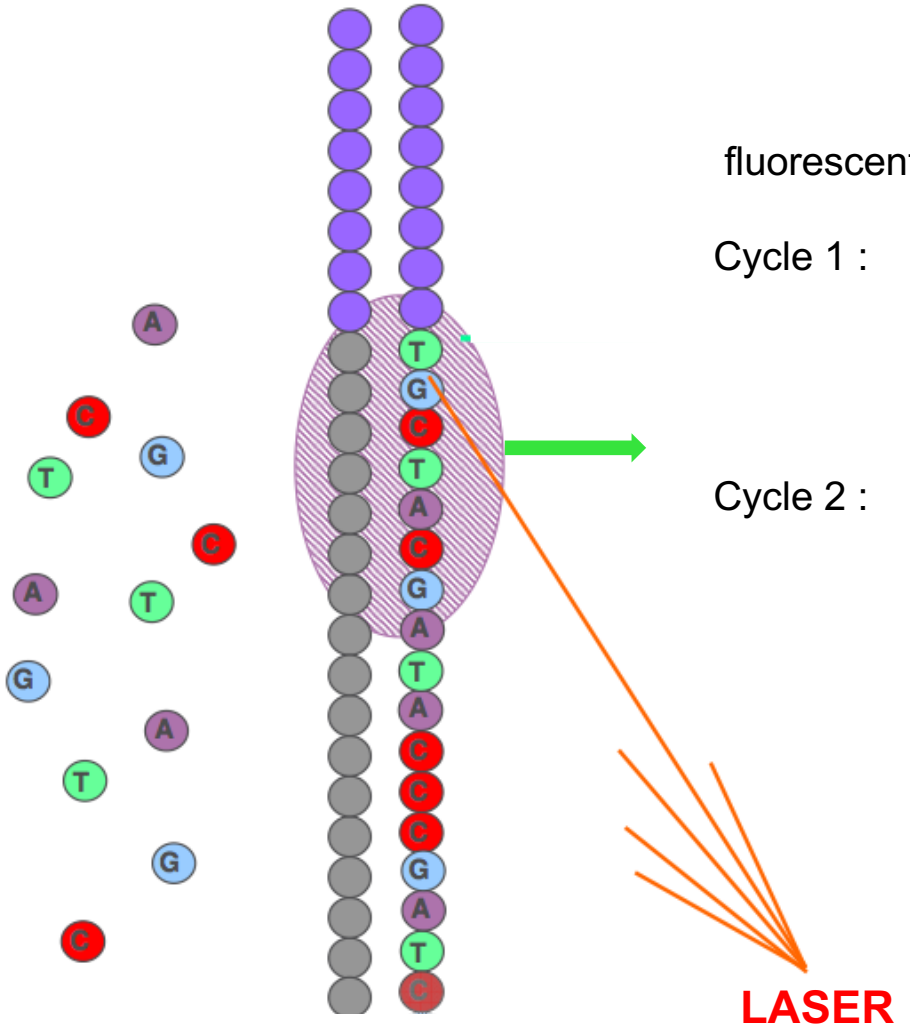


fluorescent-labeled terminator bound to each dNTP

Cycle 1 : add sequencing reagents
first base incorporated
detect signal / scan
cleave terminator and dye

Cycle 2 : add sequencing reagents and repeat

3 - Sequencing By Synthesis (SBS)

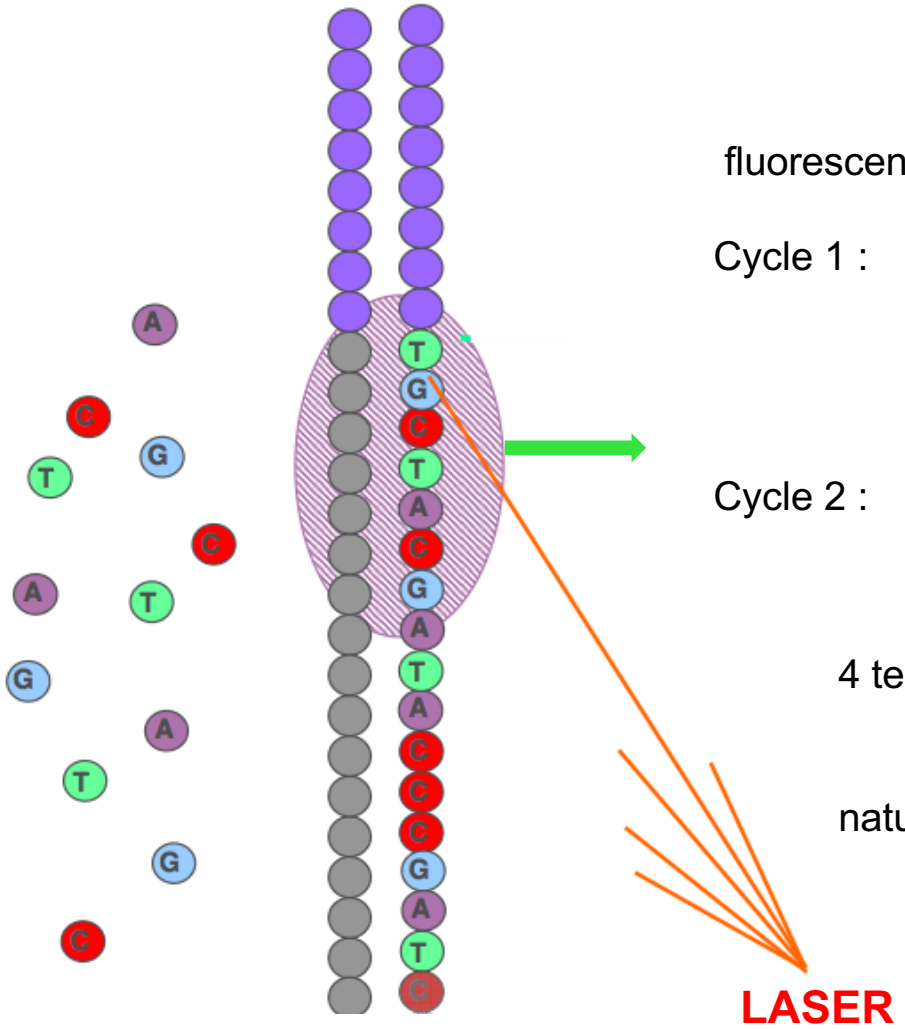


fluorescent-labeled terminator bound to each dNTP

Cycle 1 : add sequencing reagents
first base incorporated
detect signal / scan
cleave terminator and dye

Cycle 2 : add sequencing reagents and repeat

3 - Sequencing By Synthesis (SBS)



fluorescent-labeled terminator bound to each dNTP

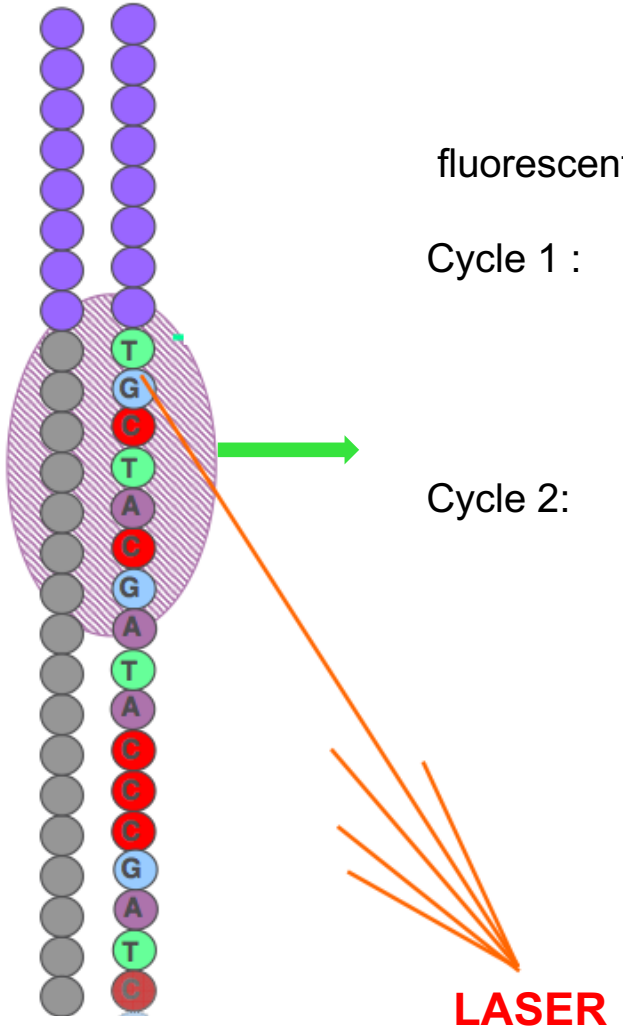
Cycle 1 : add sequencing reagents
first base incorporated
detect signal / scan
cleave terminator and dye

Cycle 2 : add sequencing reagents and repeat

4 terminator-bound dNTPs present during each cycle

natural competition minimizes incorporation bias

3 - Sequencing By Synthesis (SBS)



fluorescent-labeled terminator bound to each dNTP

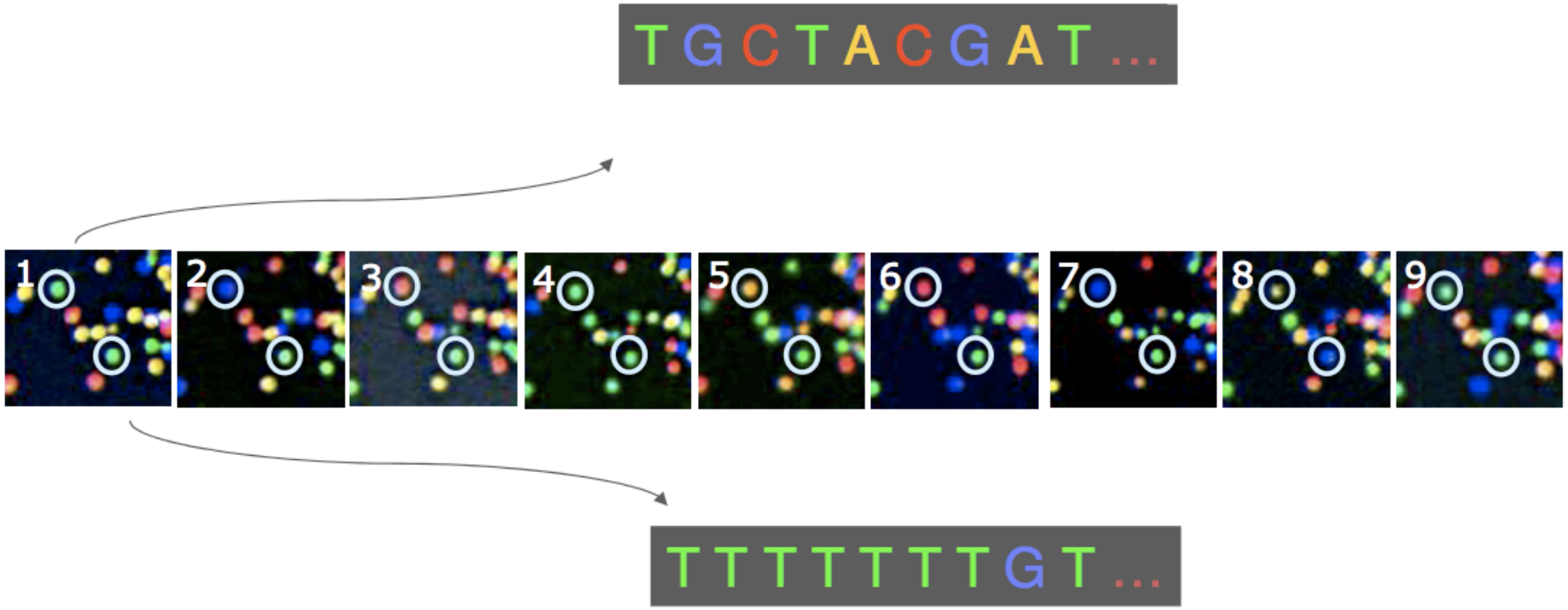
Cycle 1 : add sequencing reagents
first base incorporated
detect signal / scan
cleave terminator and dye

Cycle 2: add sequencing reagents and repeat

$(C_{eff})^{RL} = 0.5$

Cycle Efficiency (%)	Read-length (bases)
90.0	7
95.0	14
98.0	35
99.0	69
99.5	149
99.9	693

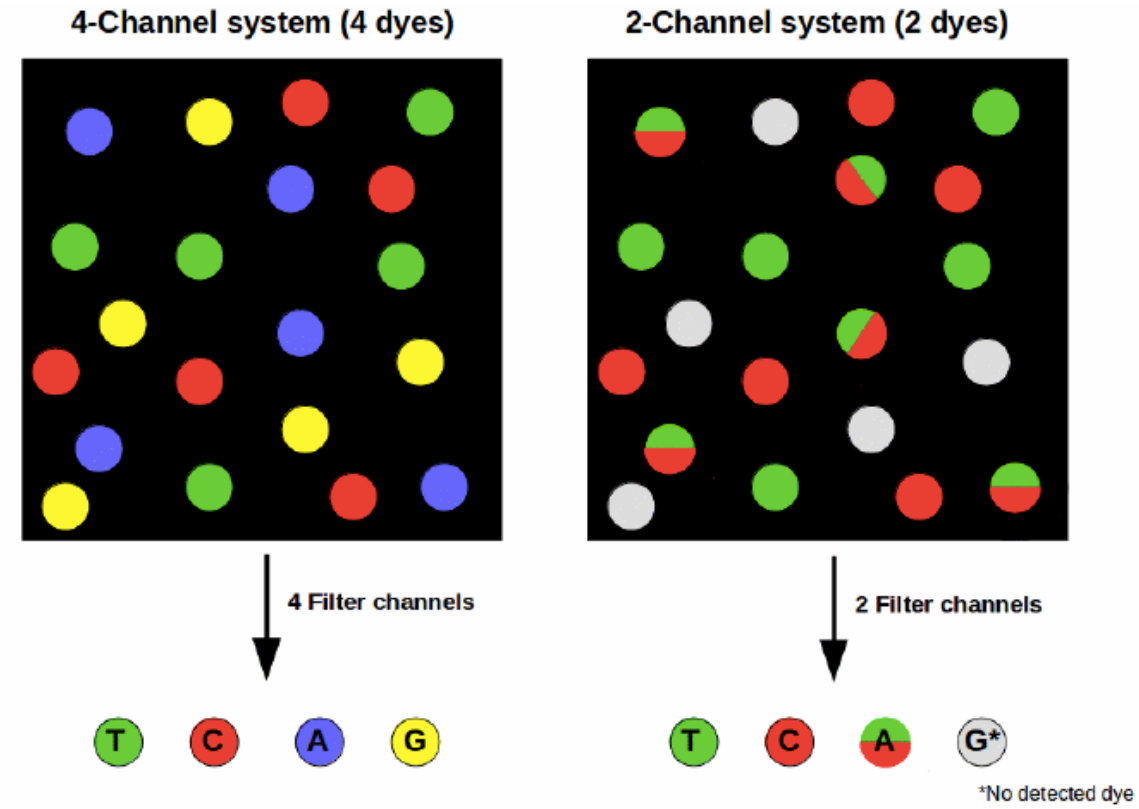
Base calling from raw data



The identity of each base of a cluster is read off from sequential images.

4 channel vs 2 channel detection

2 channel detection can cause sequencing errors due to phasing issues
and can cause polyG tracts at the end of fragments

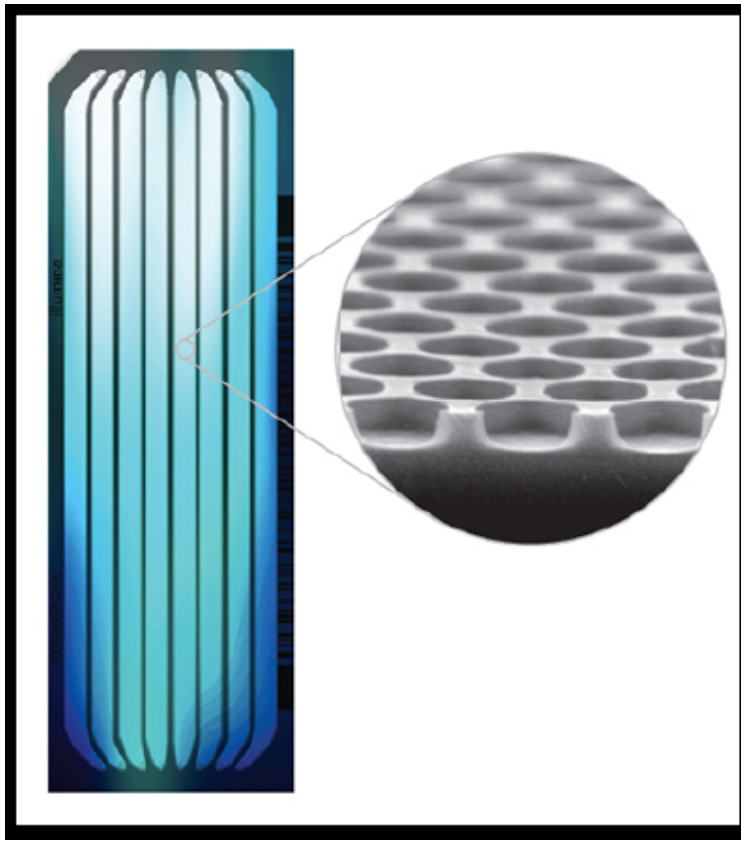


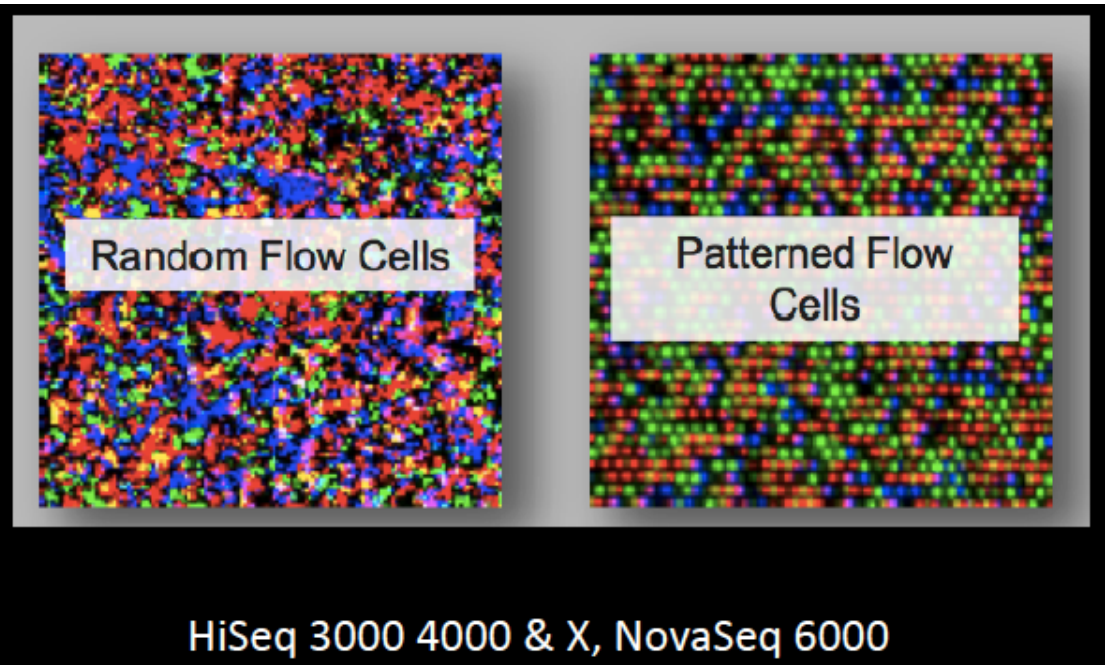
MiSeq, HiSeq 4000 and X

MiniSeq, NextSeq, NovaSeq

Patterned flow cells

- Improves regularity of densities and qualities
- Reduces analysis time



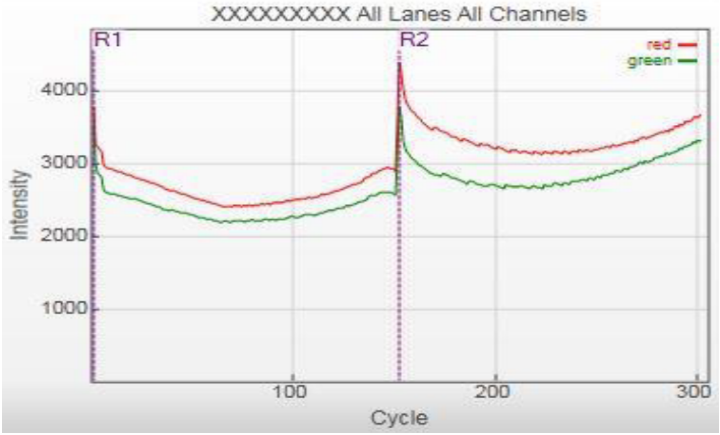


HiSeq 3000 4000 & X, NovaSeq 6000

Sequencing qualities

Intensities

Intensity values of the four incorporated bases are measured to determine if a clustering failure has occurred

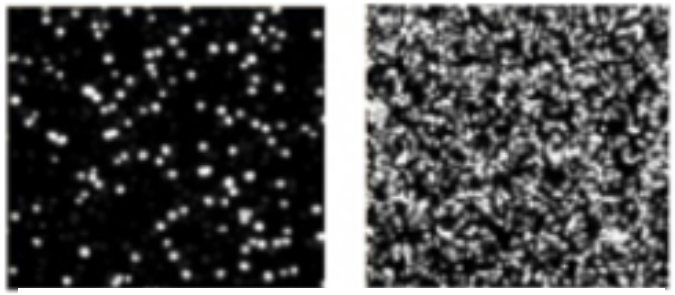


Percentage of reads passing filter (% PF)

Reflects the purity of the signal from each cluster. Ratio of the brightest intensity divided by the sum of the brightest and second brightest intensities for each cycle. Optimal values for MiSeq and HiSeq 2500 from 80-95%.

Cluster densities

Density of clusters for each tile in thousands/mm^2. Overloading of the library may cause merging of clusters, lowering of the % PF and reduction of the quality score (Q30)

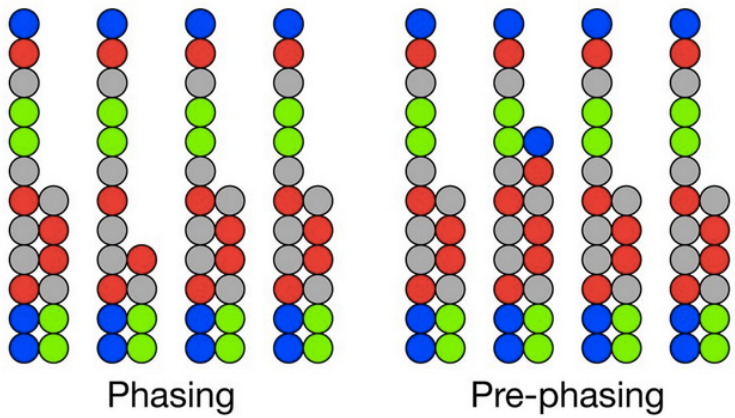


Underclustered ⟶ Overclustered

Sequencing qualities

Phasing/Prephasing

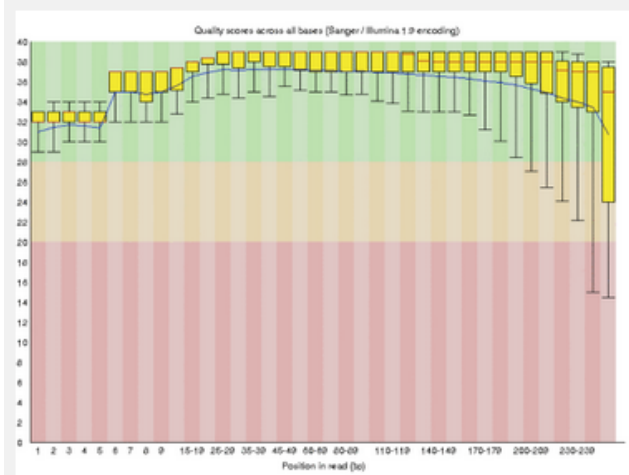
Percentages of bases that fell behind or jumped ahead the current cycle within a read. It increases with cycle number, hampering correct base identification for long reads. Optimal values are below 0.5 or 0.2% depending on platform. Issues may arise when read length is above specifications.

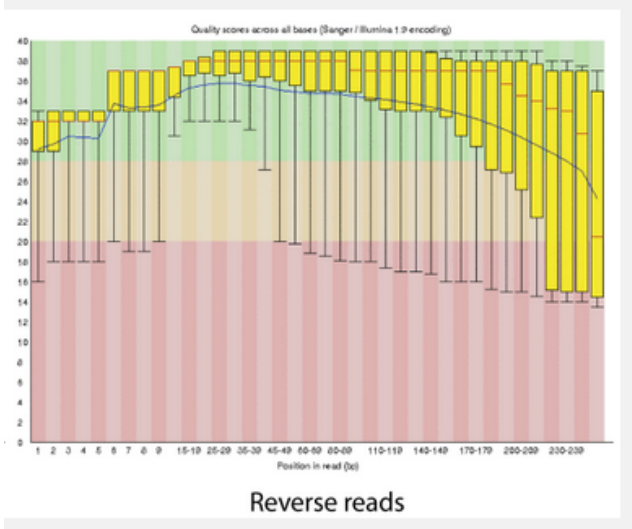


Q-score

Base calling error probabilities P : $Q = -10 \log_{10} P$

% Q-score > 30 : percentage of bases that have a a probability of incorrect base calling of 1/1000. Low Q scores can increase false-positive variant calls, which can result in inaccurate conclusions.

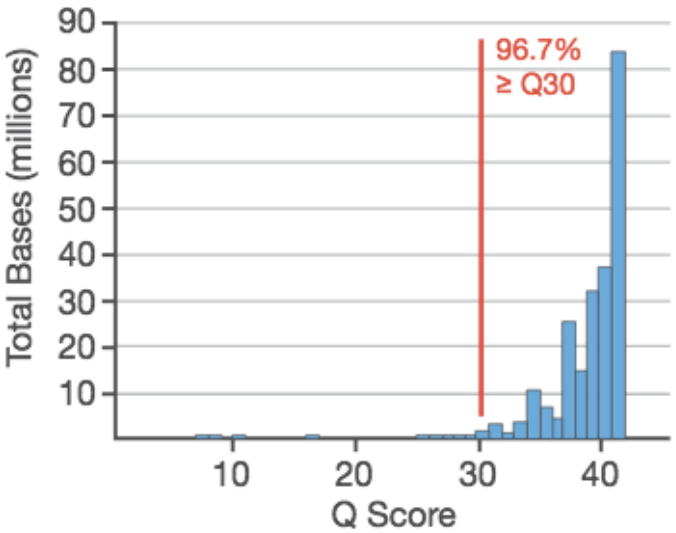




Sequencing qualities

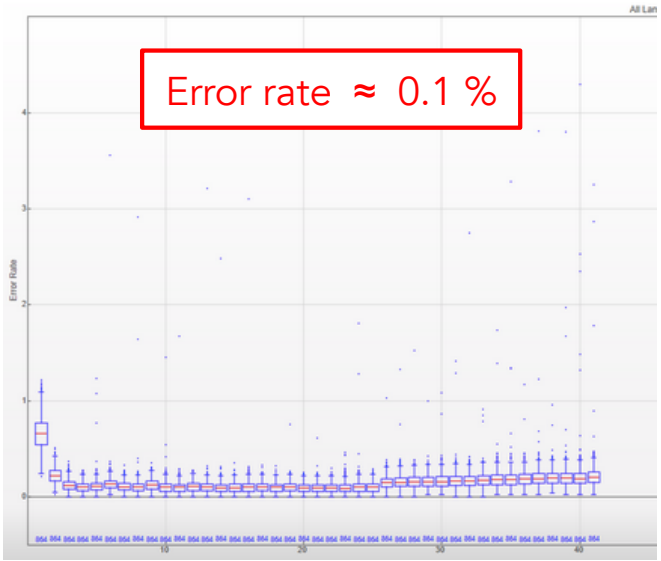
Percentage of alignment

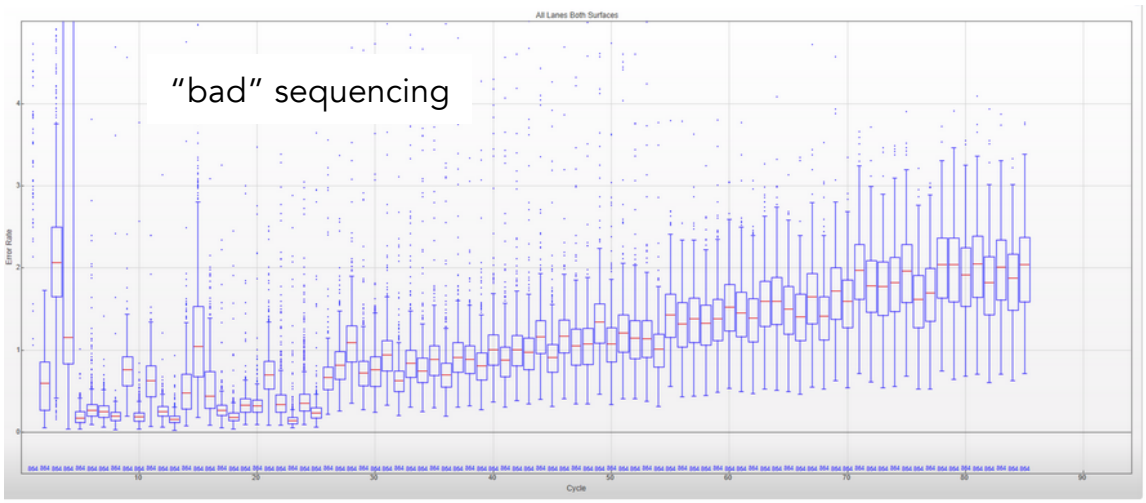
% align: percentage of the reads aligned to PhiX control genome. It is important to estimate the error rate. Indication on the success of the clustering process.



Error rate

Calculated for each read based on the aligned % of PhiX control sample. Important to determine if the sequencing has proceeded as expected.





SINGLE CELL TECHNOLOGIES

Single cell transcriptomics allows to study transcriptome heterogeneity,

to investigate differences in transcript expression and gene regulation *in individual cells* :

- ❖ Differences in trancript abundance
- ❖ Alternative splicing and differential expression of isoforms

Most widely used device to study single-cell transcriptomics : *Chromium controller (10x Genomics)*

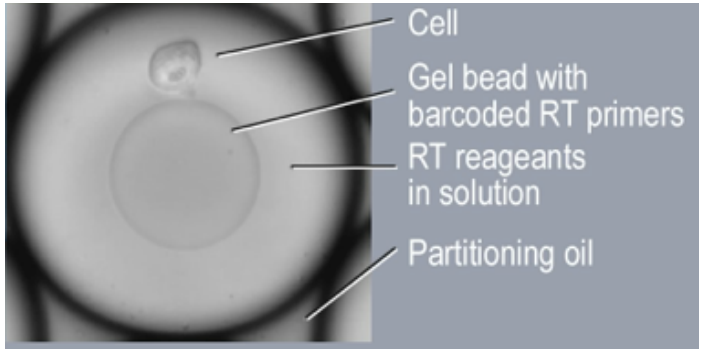
Several applications :

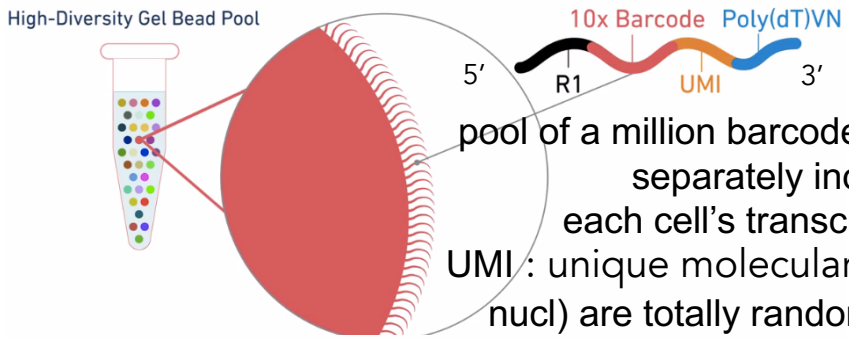
Single cell Gene expression
Measures gene activity on a cell-by-cell basis, characterize cell populations, cell types, …

Linked read genomics
Performs diploid de novo assembly, phase haplotypes, genetic variations

Single cell ATAC
Measures epigenetics by detecting open chromatin regions

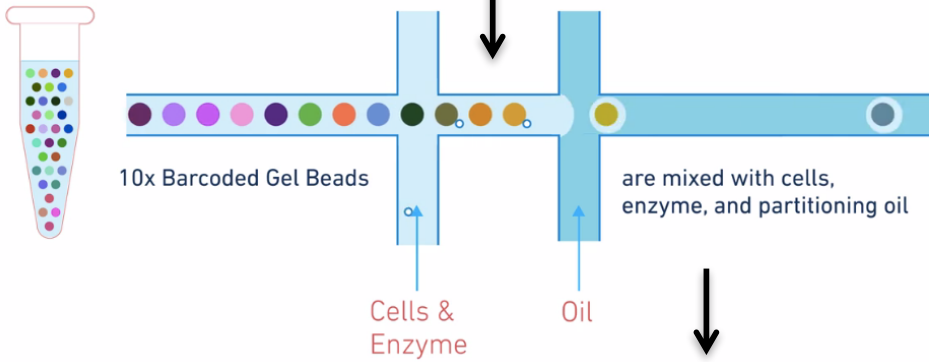
CHROMIUM SINGLE-CELL RNA SEQUENCING

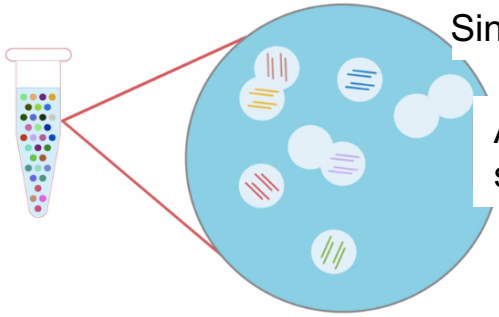




pool of a million barcodes (16 nucl) to separately index each cell's transcriptome

UMI : unique molecular identifier (10 nucl) are totally random seqences

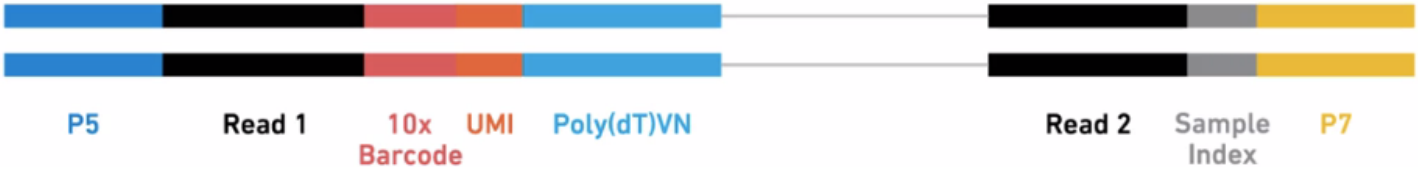




Single cells undergo RT

All cDNAs from one cell share the same barcode

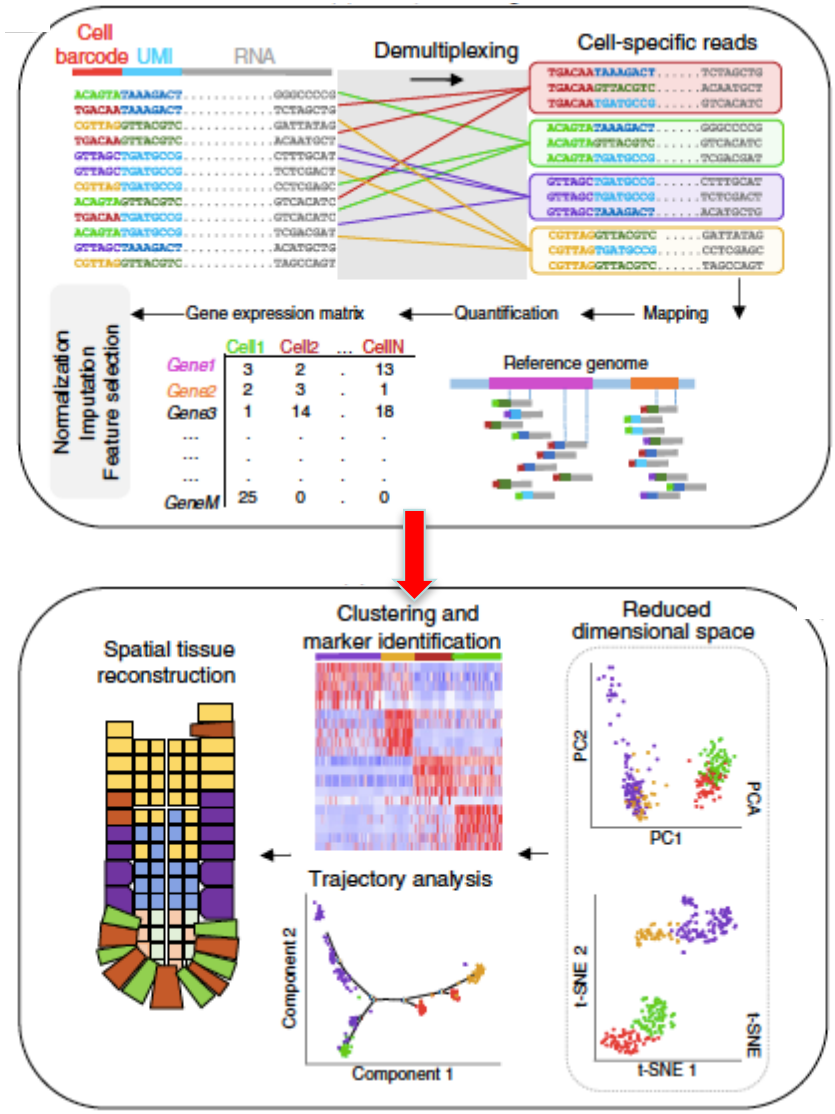
barcoded sequencing library



DATA ANALYSES

CELL RANGER :

Chromium data analysis pipelines



Recent advances : Illumina sequencing with Smart-seq3

Single-cell RNA counting at allele- and isoform-resolution using Smart-seq3
Hagemann-Jensenn et al. *Nature Biotechnology*, 2020

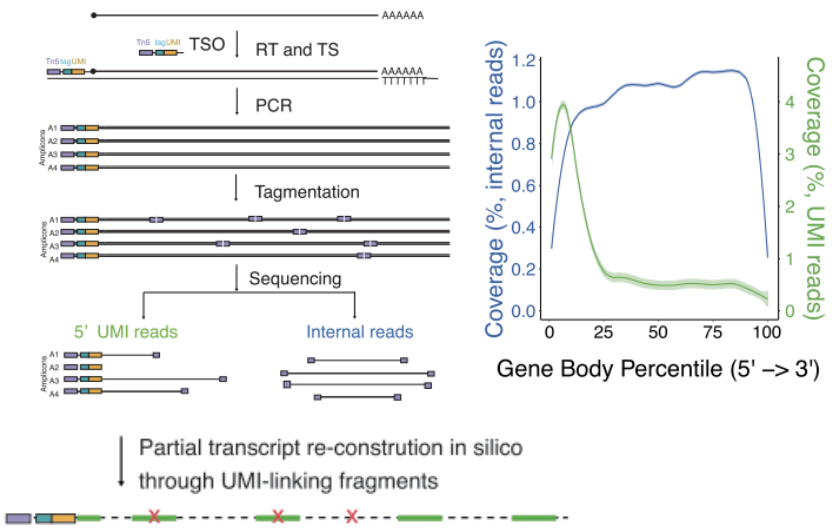
- Single cell methods capture only a small fraction of cellular mRNA (~5–10%)
- They frequently lack splice isoform information

Smart-seq3 :

- full-length transcriptome coverage
- 5′ UMI RNA -> in silico reconstruction of thousands of RNA molecules/cell
- 60% assignments to allelic origin
- 30–50% assignments to specific isoforms

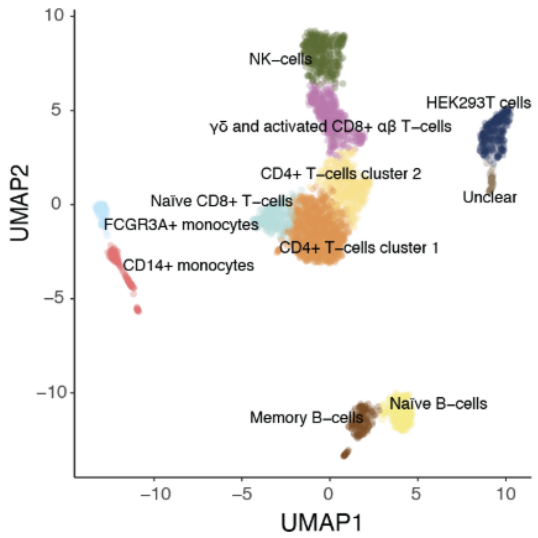
Smart-seq3 greatly increased sensitivity compared to Smart-seq2 detecting thousands more transcripts/cell

TS0 : Template-Switching Oligo
UMI : unique molecular identifier

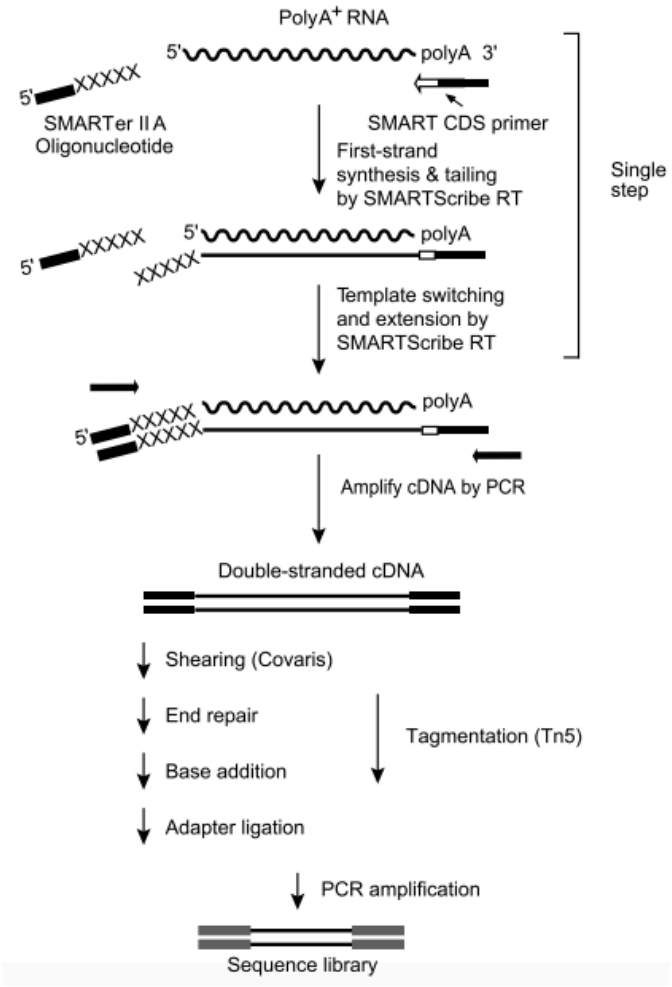


Count RNA molecules at allelic and transcript isoform resolution

Dimensionality reduction of 3,890 human cells colored by annotated cell type



The Smart-Seq protocol

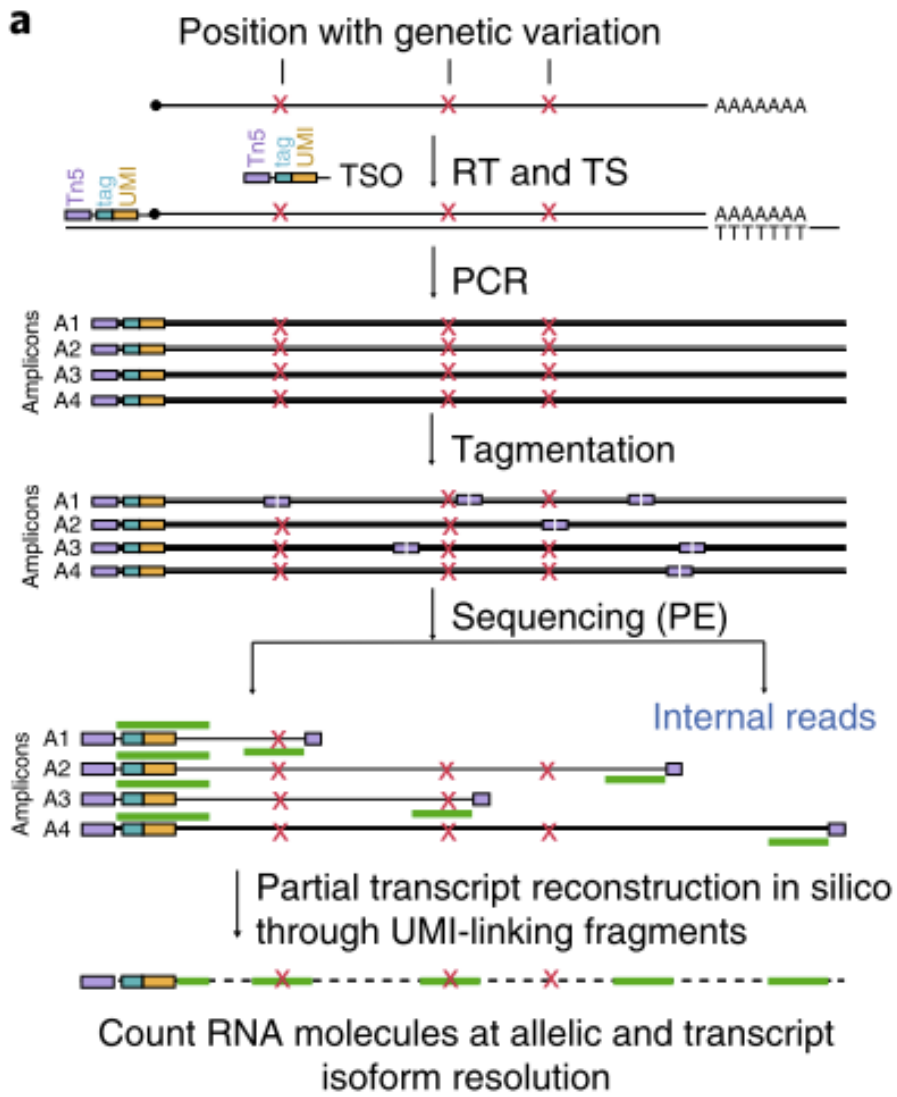


Overview of single-cell RNA sequencing in Smart-seq3

Template-Switching Oligo

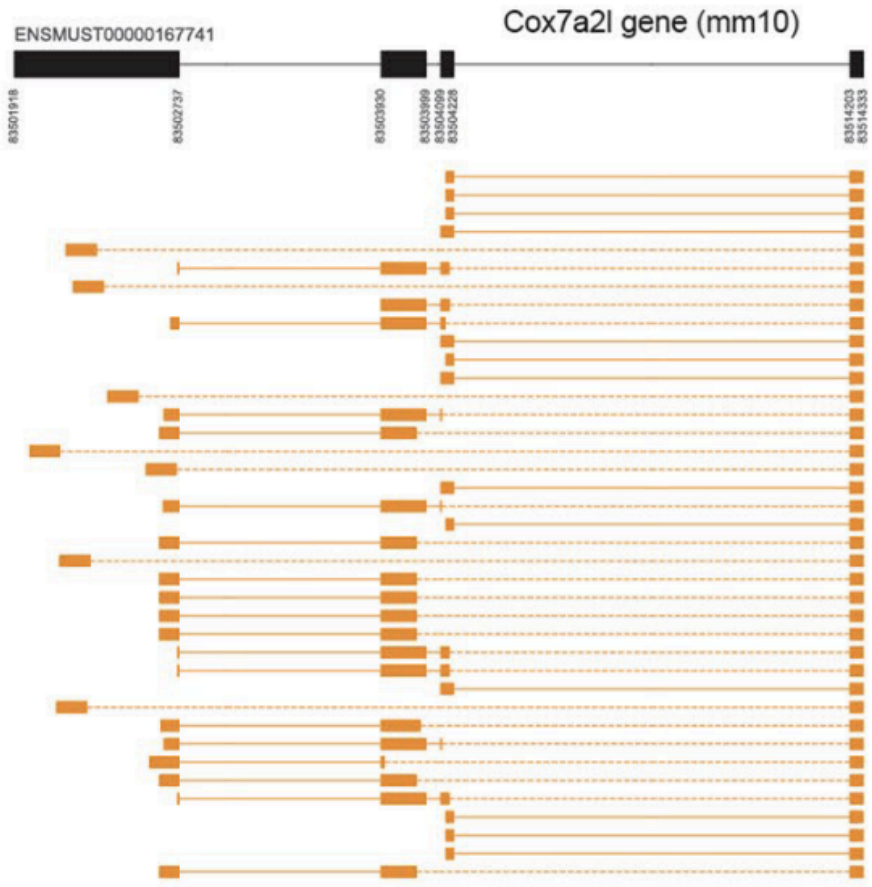
☐TSO consists of :

- a partial Tn5 motif
- a 11-bp tag sequence
- a 8-bp UMI sequence and 3 riboguanosines



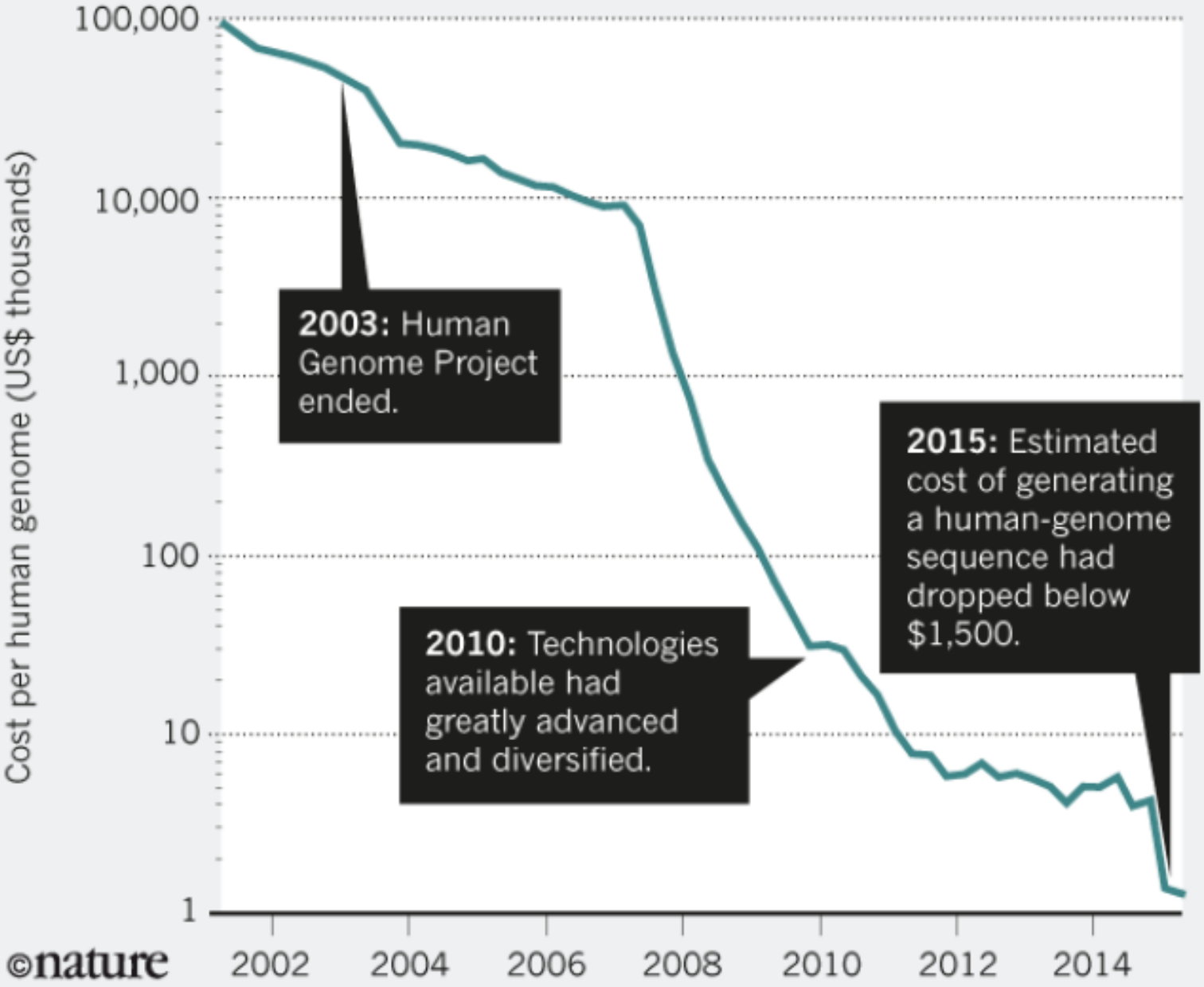
Example of read pairs from a single transcribed molecule

Each row shows a unique read pair



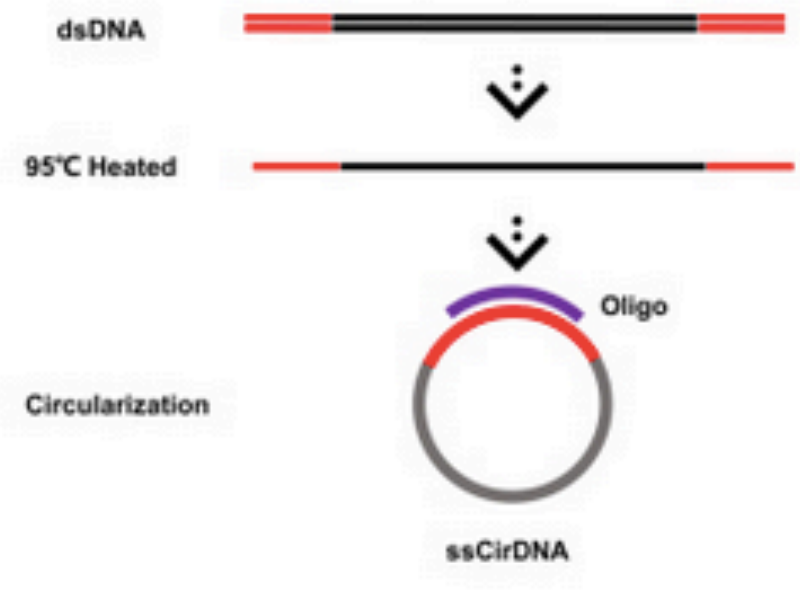
BETTER, CHEAPER, FASTER

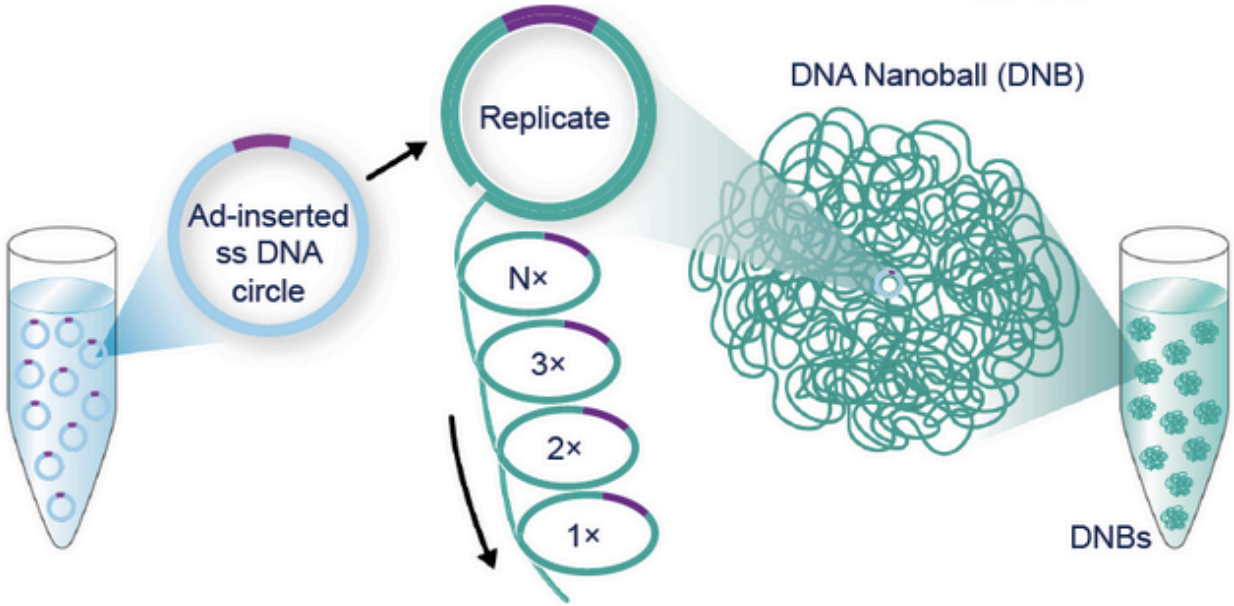
The cost of DNA sequencing has dropped dramatically over the past decade, enabling many more applications.

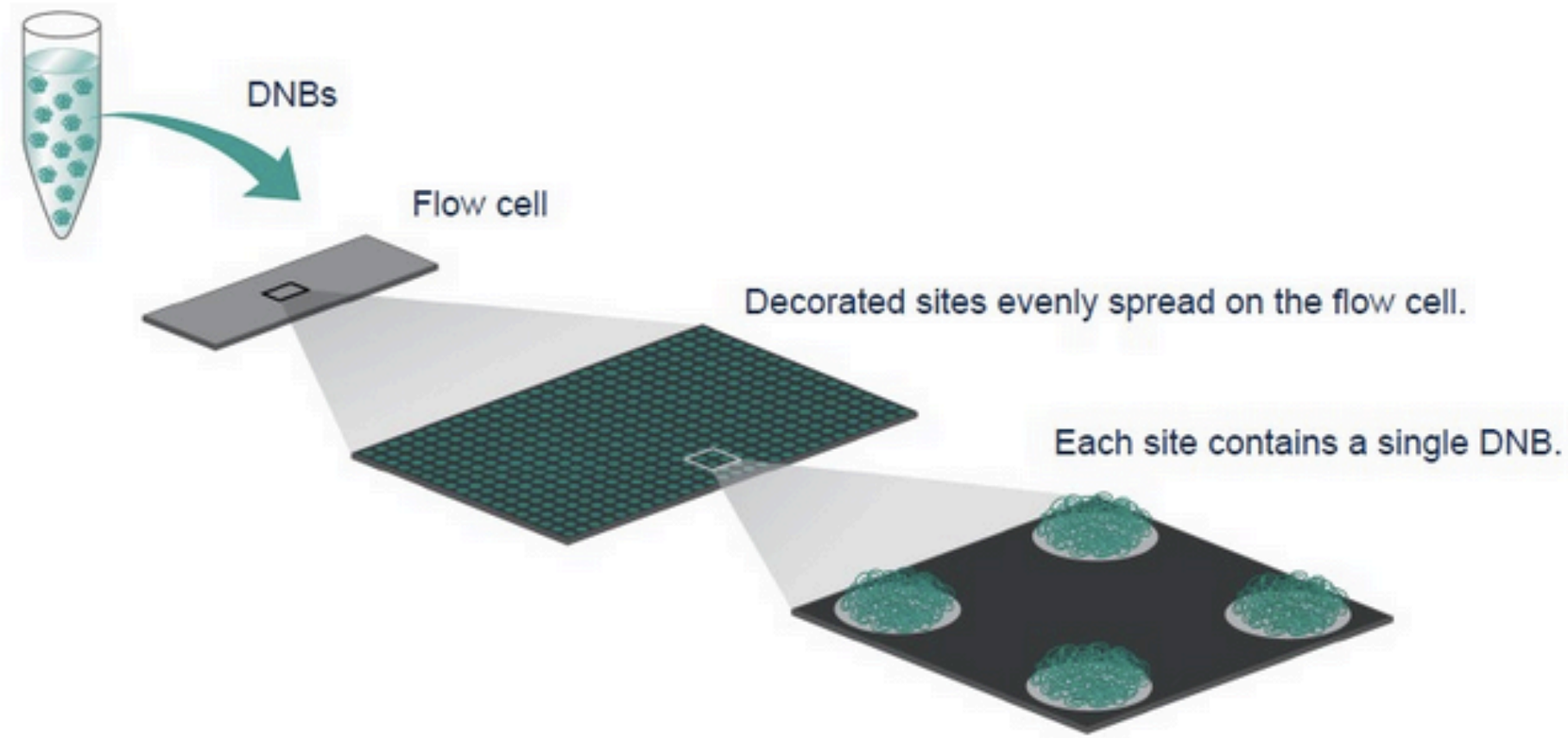


MGI 华大智造

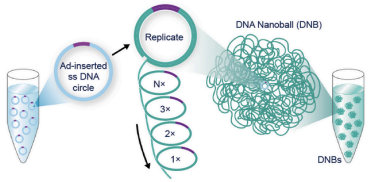
acquisition of U.S. company
Complete Genomics in 2013

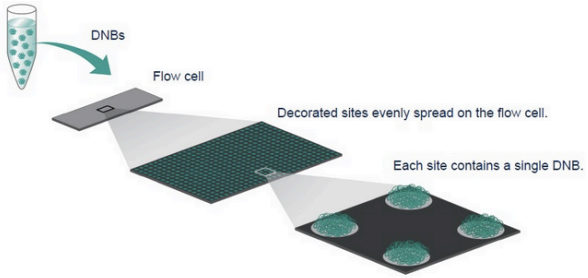





DNBs
Flow cell
Decorated sites evenly spread on the flow cell.
Each site contains a single DNB.

NEW SEQUENCING TECHNOLOGY





Tests préliminaires (Genoscope, CNRGH)

- Qualité : Q30 moyen élevé (même en 2x200)
- Taux d'erreur MGI < taux d'erreur NovaSeq
- % reads mappés MGI > % reads mappés Illumina

Mais :

- Runs plus longs (+ lavage 6h)
- Preparation of nanoballs « délicate »
- Régions riches A/T et G/C moins couvertes

	G400	NovaSeq S1	NovaSeq S4
Run time	66h	24h	48h
Output	500 Gb	500 Gb	1600-2000 Gb
Cost/Gb (€)	3.8 - 4.5	9.9 - 12.4	6.6 - 8.3

PART 2

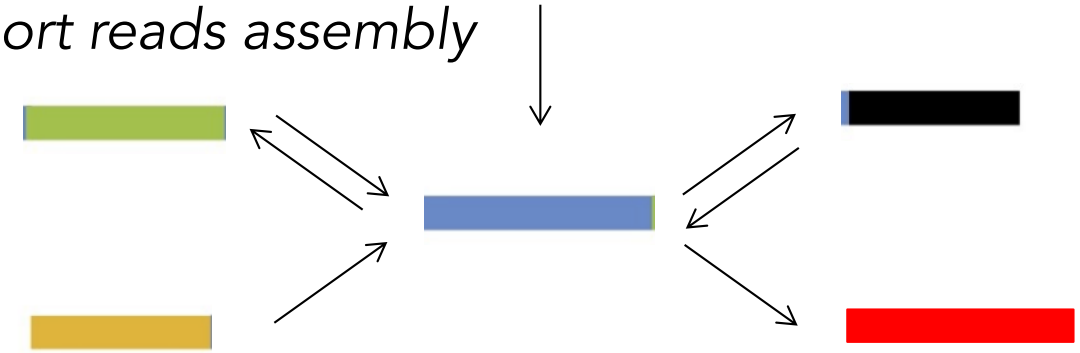
3rd GENERATION SEQUENCING

LONG READS

LONG-READS VERSUS SHORT-READS

Assembly of DNA fragments with repeated sequences

NGS short reads assembly

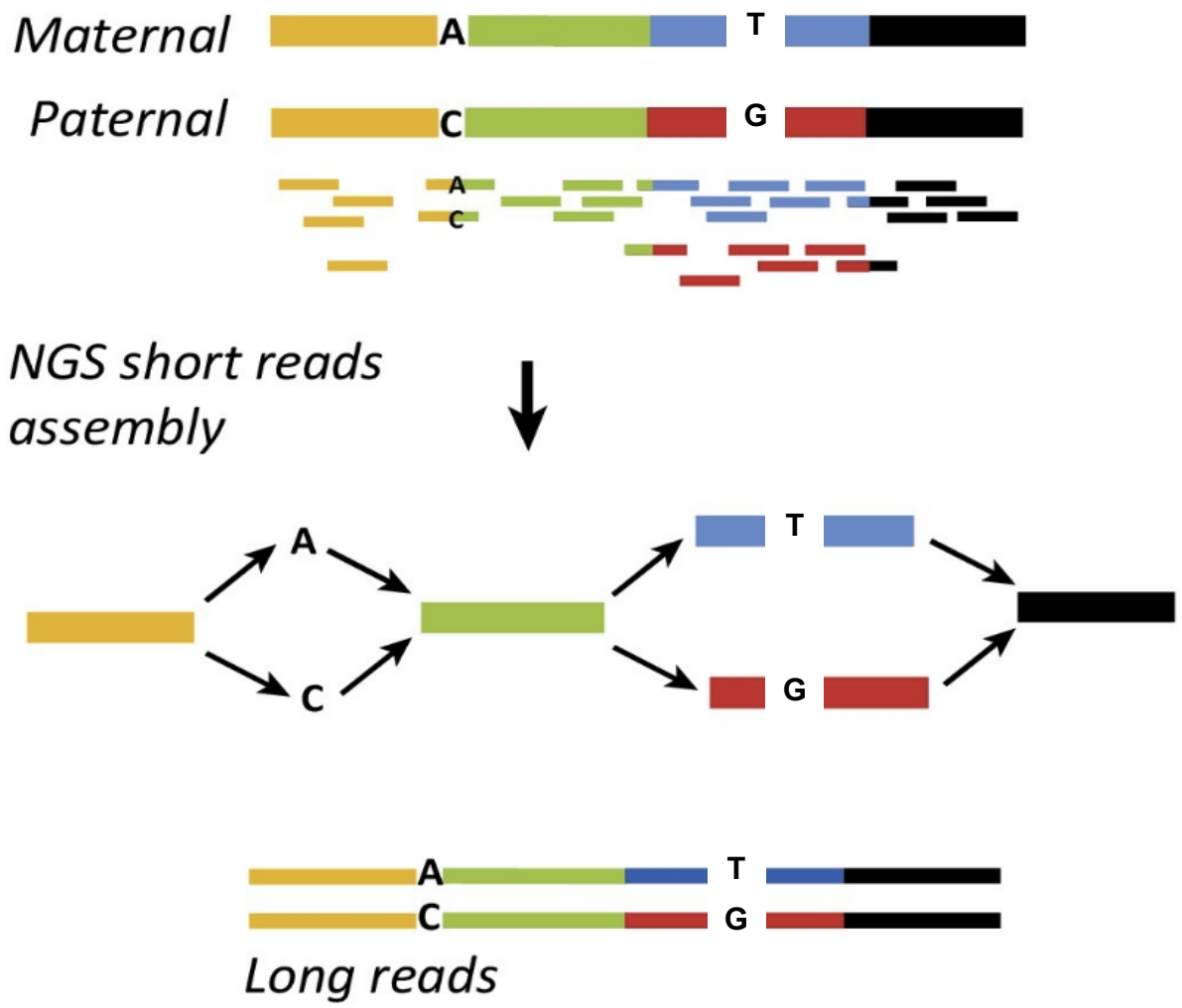


Several contigs $\longrightarrow$ incomplete assembly, underestimation of repeats

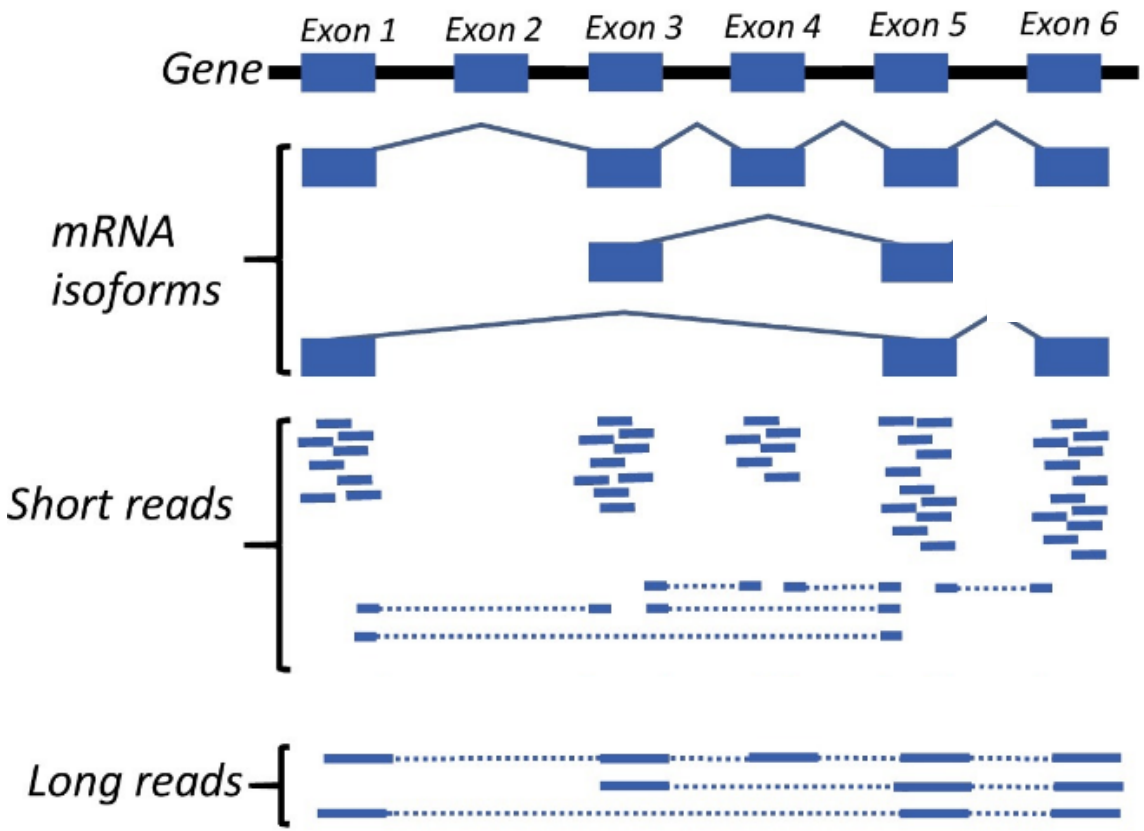
Long reads assembly

LONG-READS VERSUS SHORT-READS

Haplotype phasing



Detection of splicing isoforms



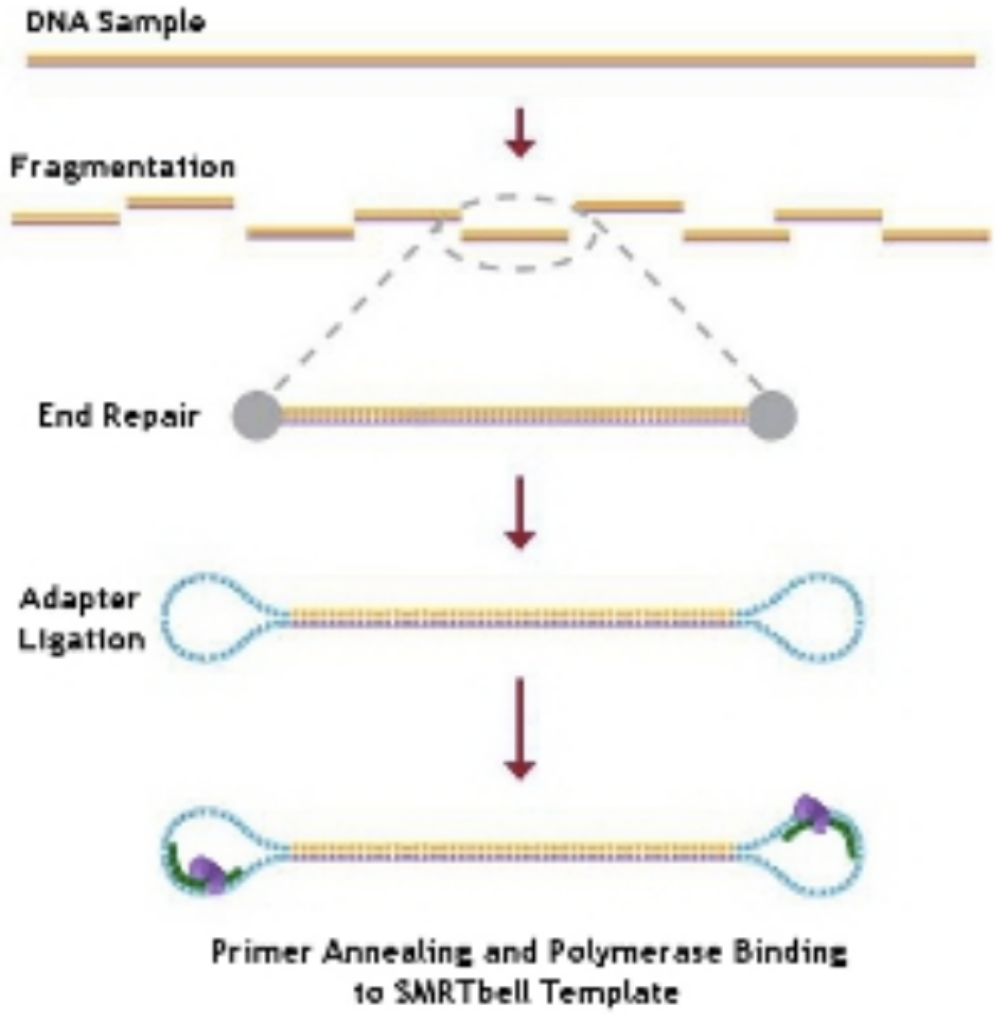
The 3rd generation winning technologies

Sequel - Pacific Biosciences
Single molecules
Up to 80,000 bp long
Error rate ≈ 10-15 % - CCS: <1%
Compensated by coverage

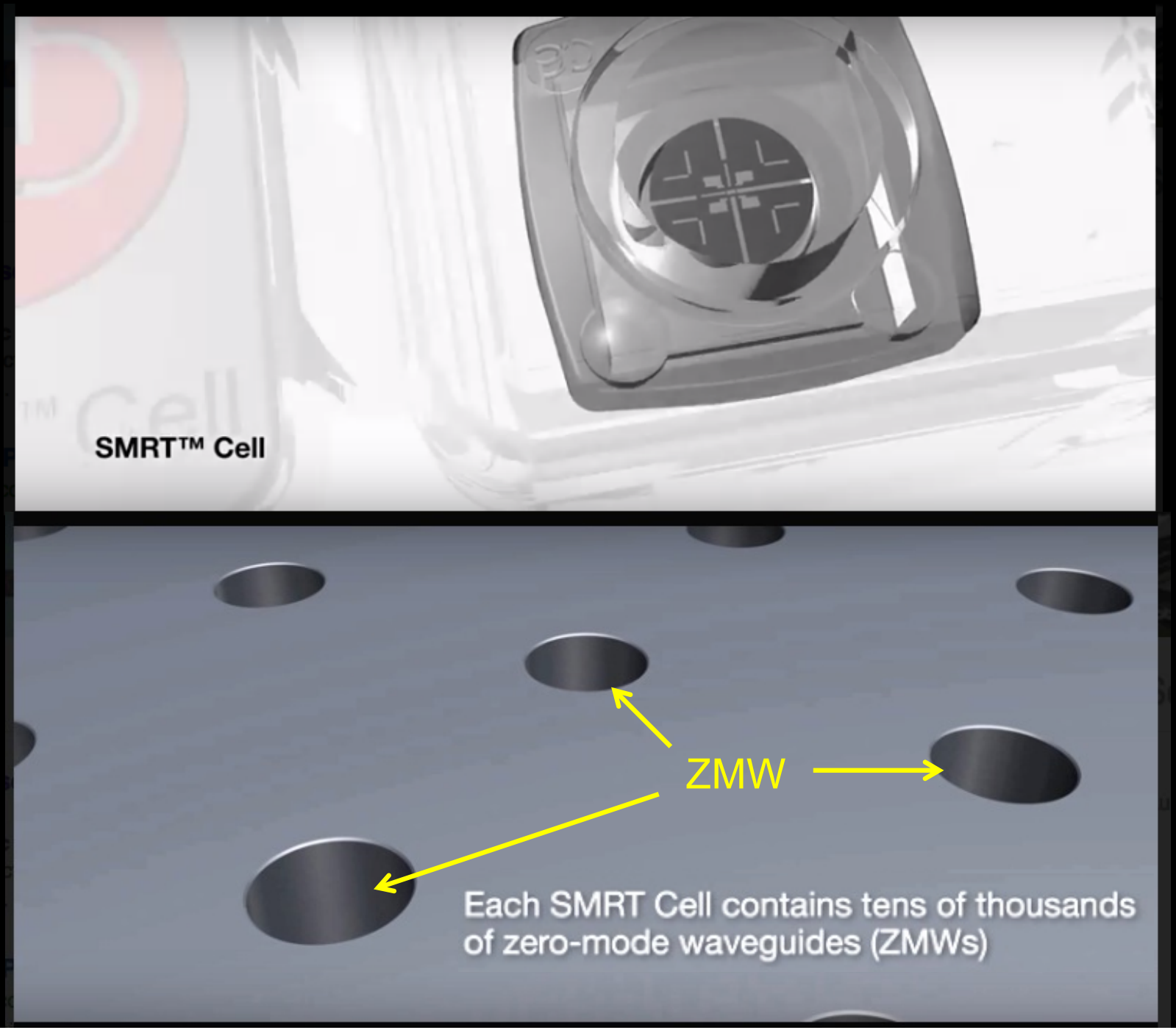
MinION - Oxford Nanopore
Single molecules
> 200 000 bp long
Error rate ≈ 10-15 %
Compensated by coverage

PacBio : Single Molecule Real Time (SMRT) sequencing

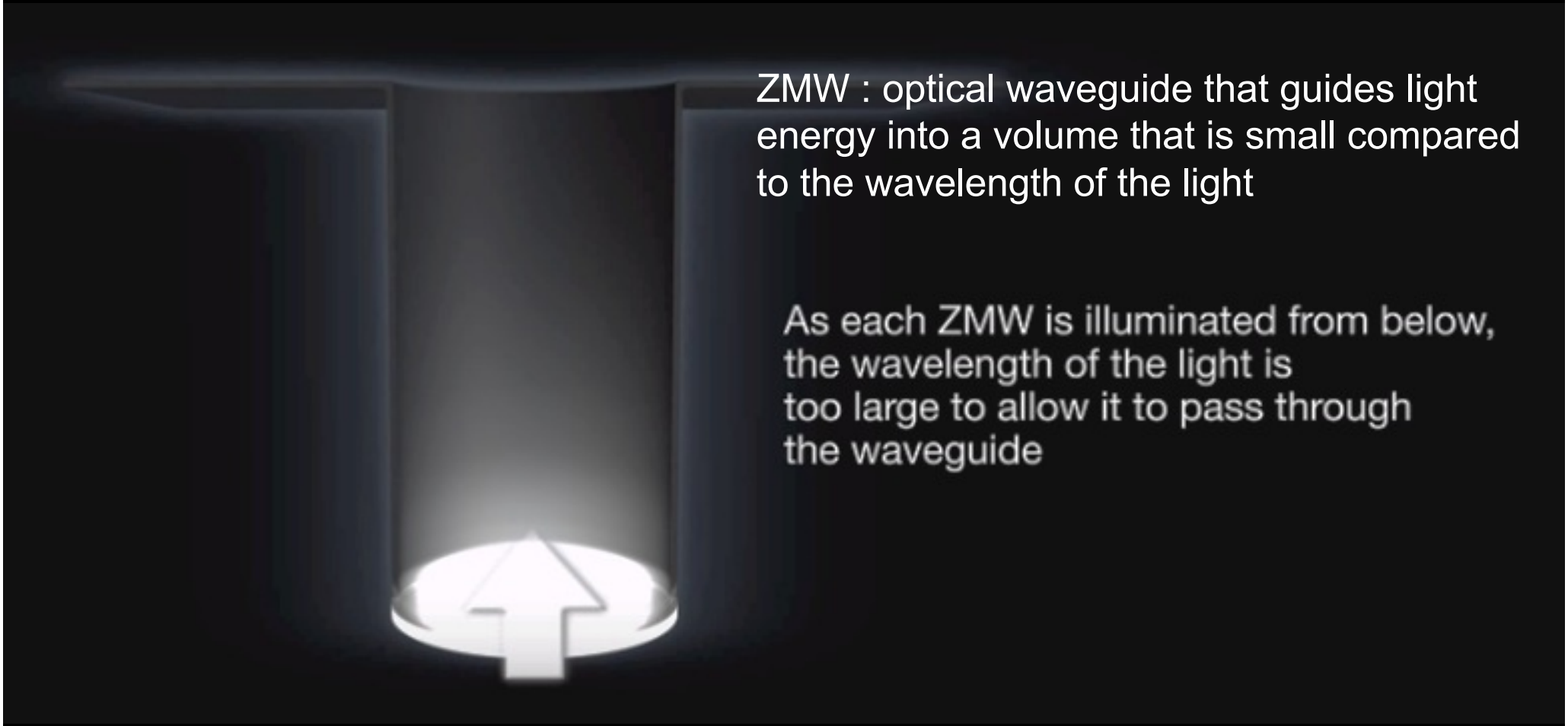
PacBio DNA-seq library



PACIFIC BIOSCIENCES

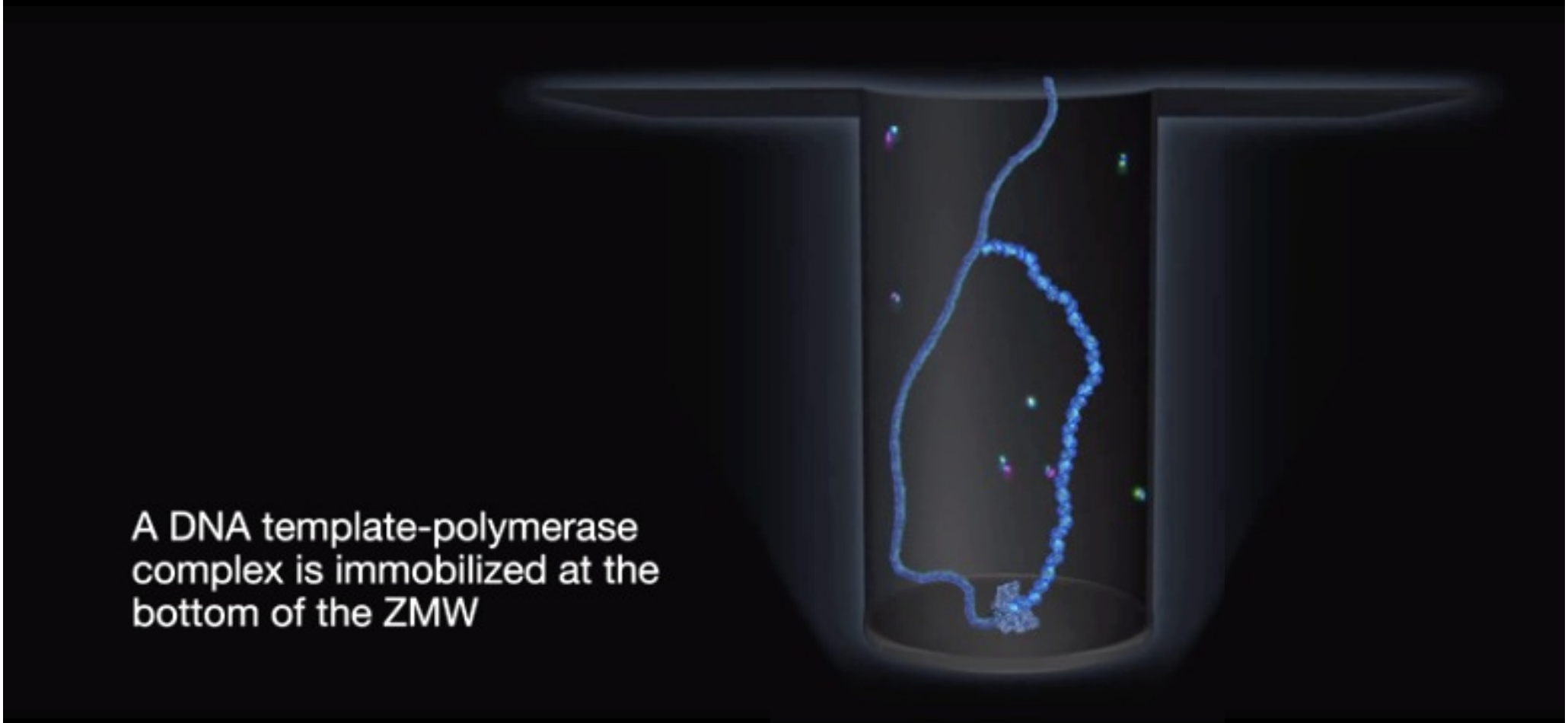


PACIFIC BIOSCIENCES



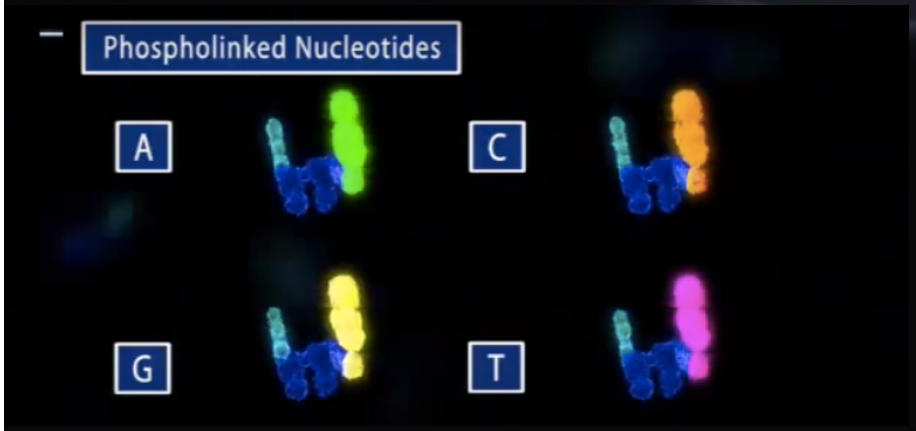
ZMW : optical waveguide that guides light energy into a volume that is small compared to the wavelength of the light

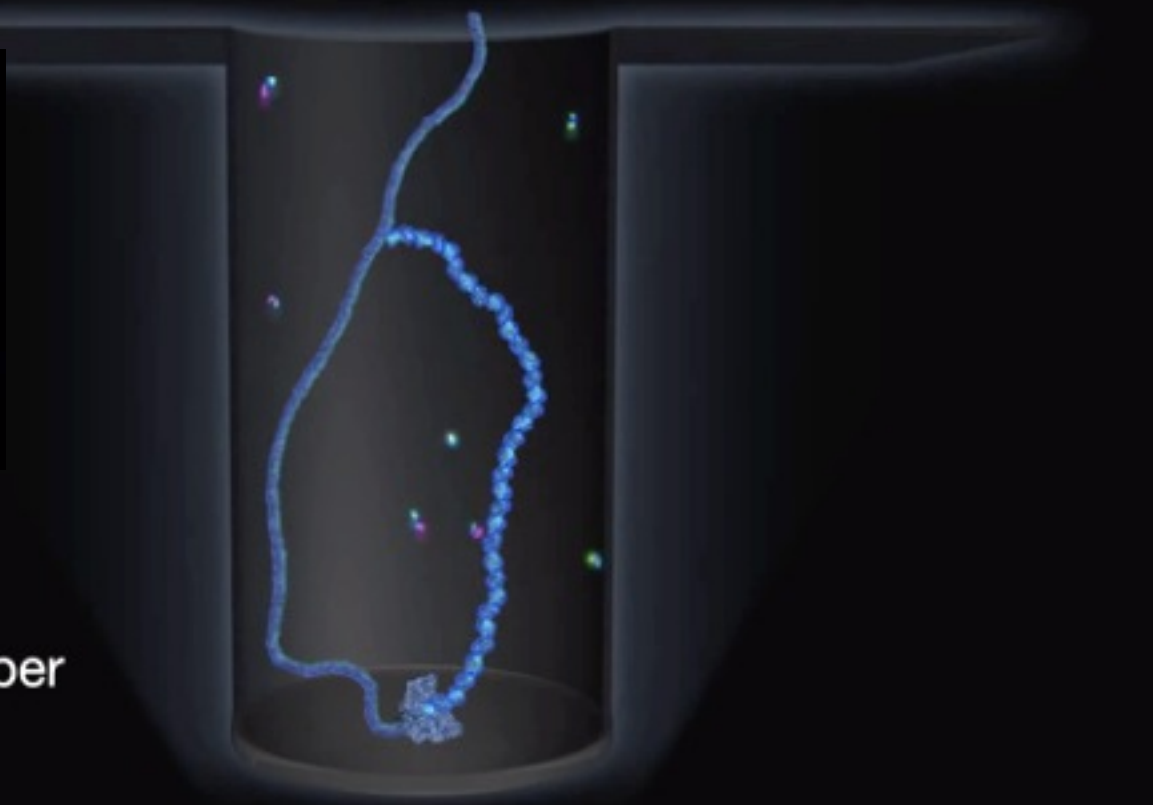
PACIFIC BIOSCIENCES



A DNA template-polymerase complex is immobilized at the bottom of the ZMW

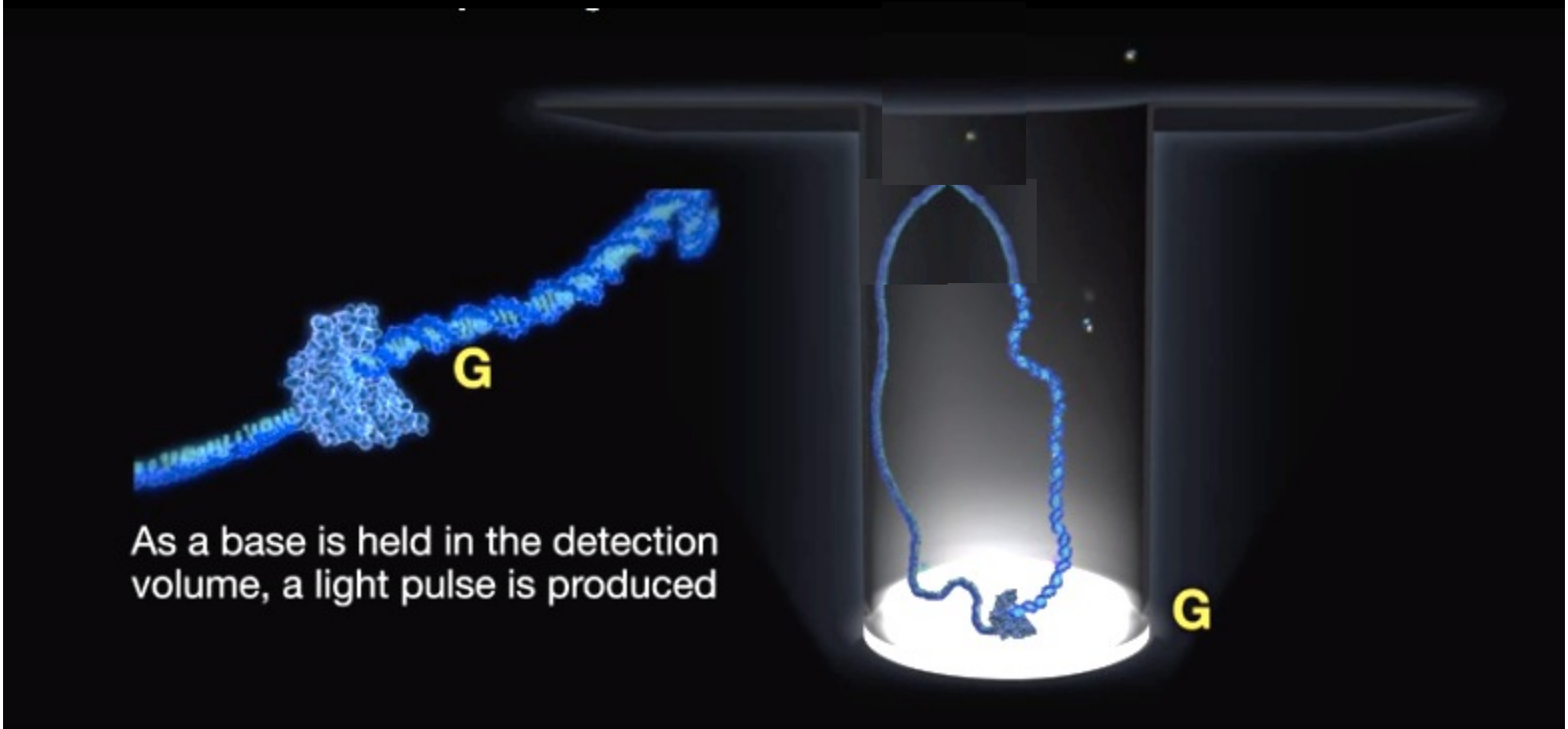
PACIFIC BIOSCIENCES



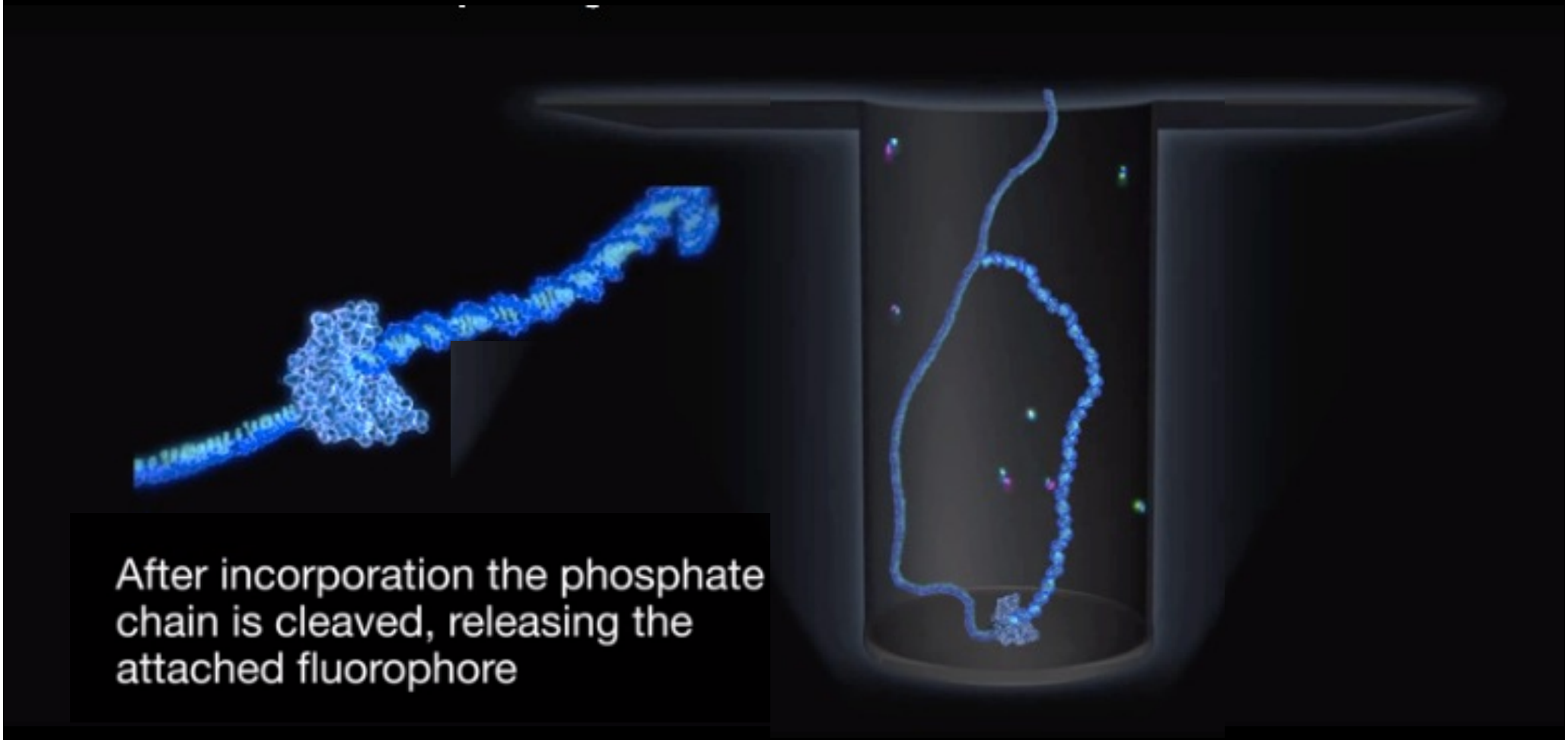


Phospholinked nucleotides are introduced into the ZMW chamber

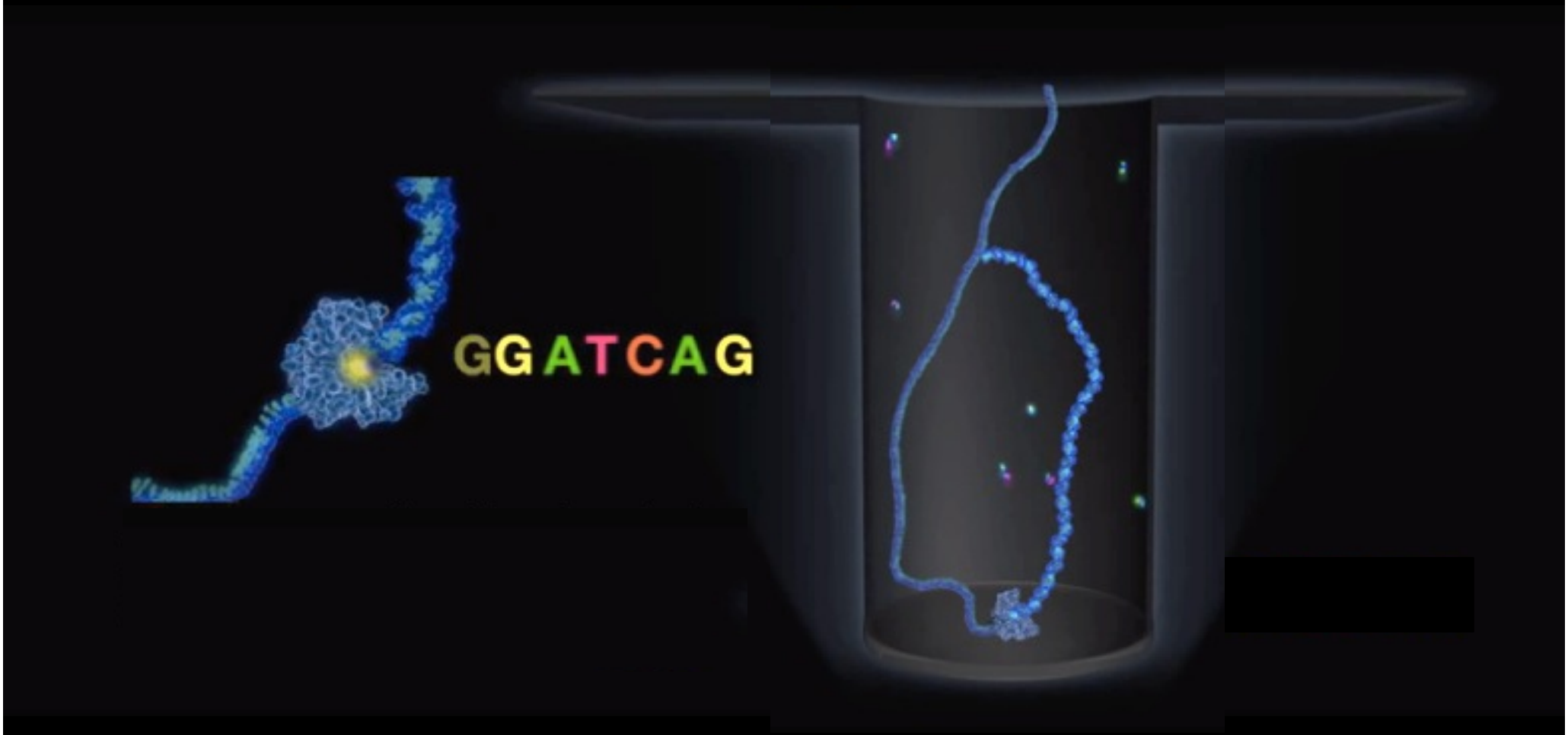
PACIFIC BIOSCIENCES

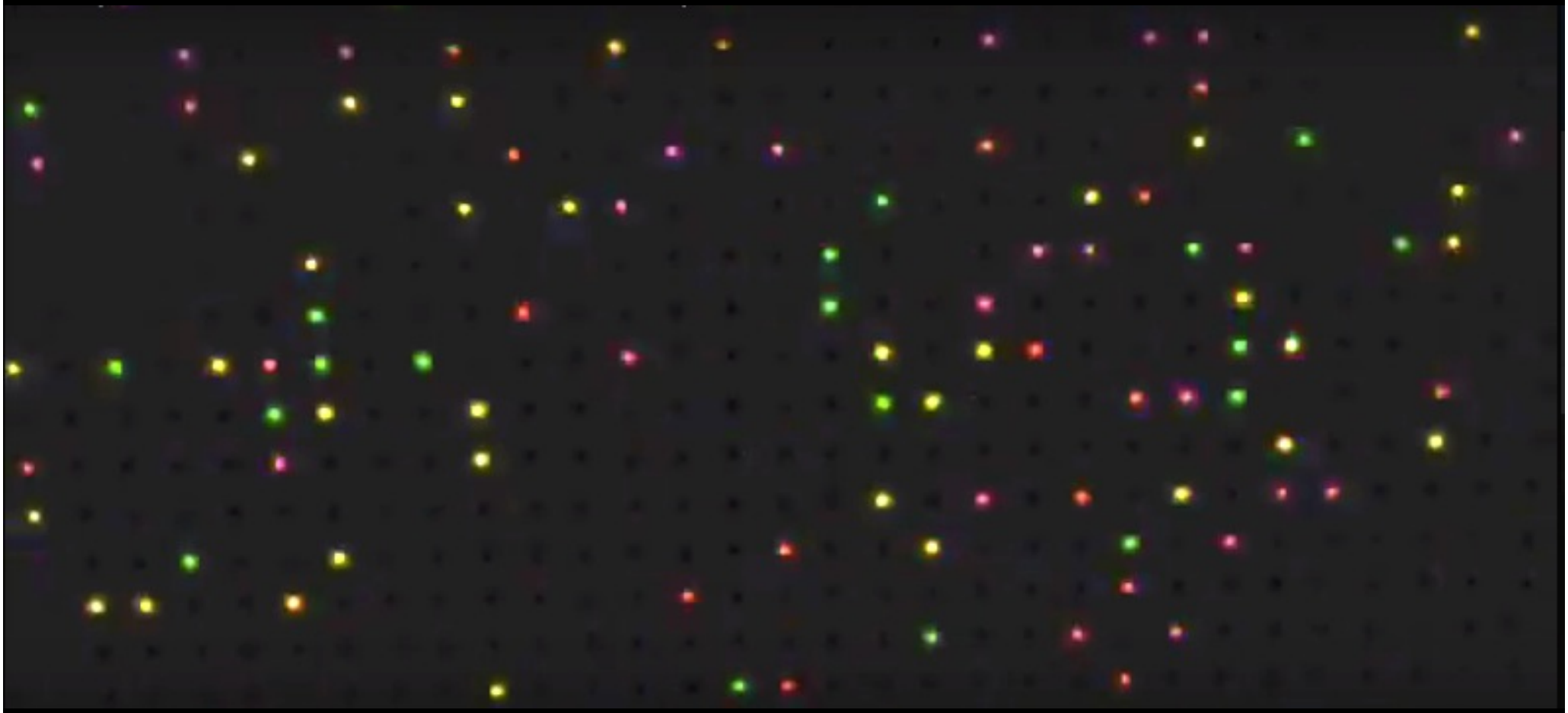


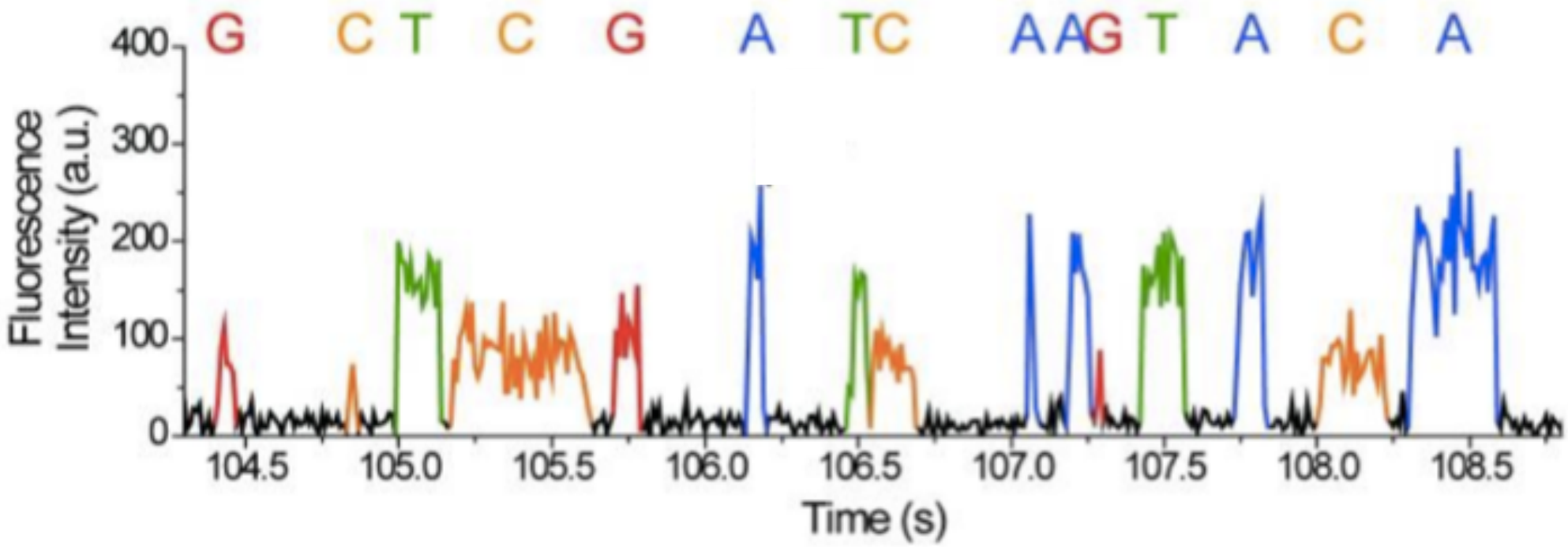
PACIFIC BIOSCIENCES



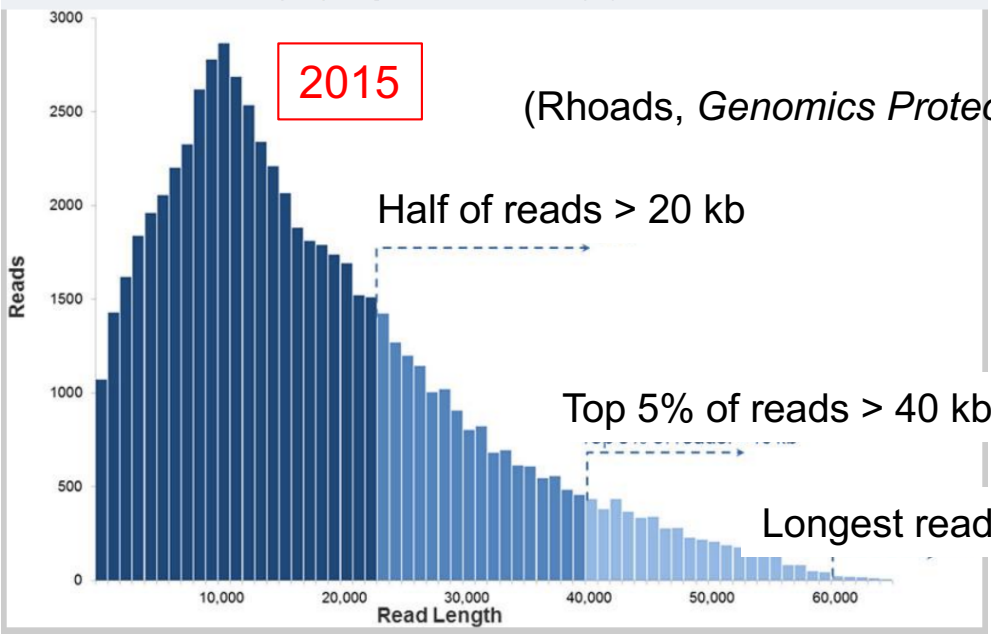
PACIFIC BIOSCIENCES






G C T C G A TC AAG T A C A
400
300
200
100
0
Fluorescence Intensity (a.u.)
104.5
105.0
105.5
106.0
106.5
107.0
107.5
108.0
108.5
Time (s)

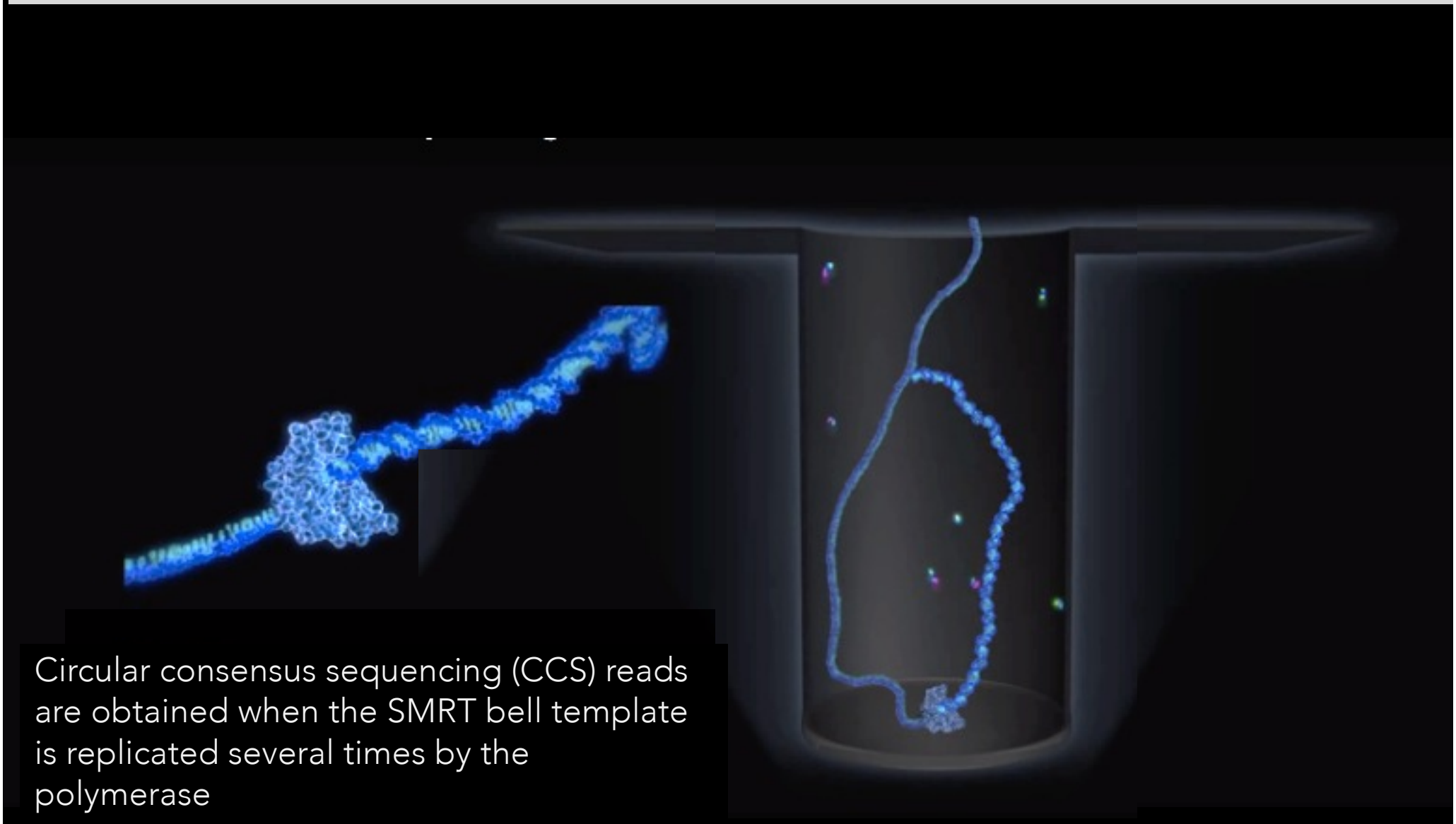
Length of PacBio reads



(Rhoads, *Genomics Proteomics Bioinformatics*, 2015)

Data from 35 kb size-selected *E. coli* library
Sequel II System ; 2.0 Chemistry

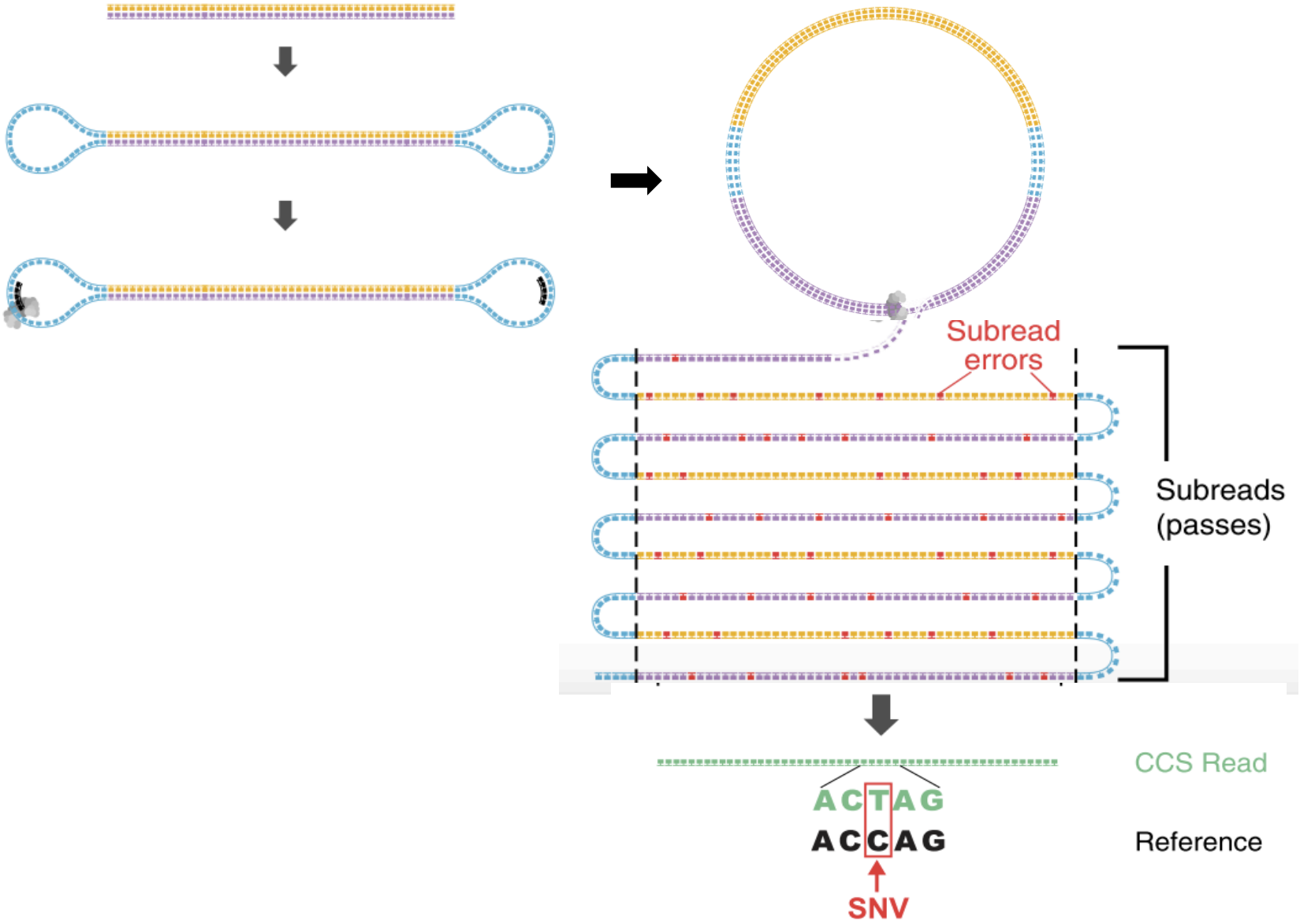
RECENT IMPROVEMENT WITH NEW CHEMISTRY



Circular consensus sequencing (CCS) reads are obtained when the SMRT bell template is replicated several times by the polymerase

This allows a highly accurate sequencing by correction of random errors

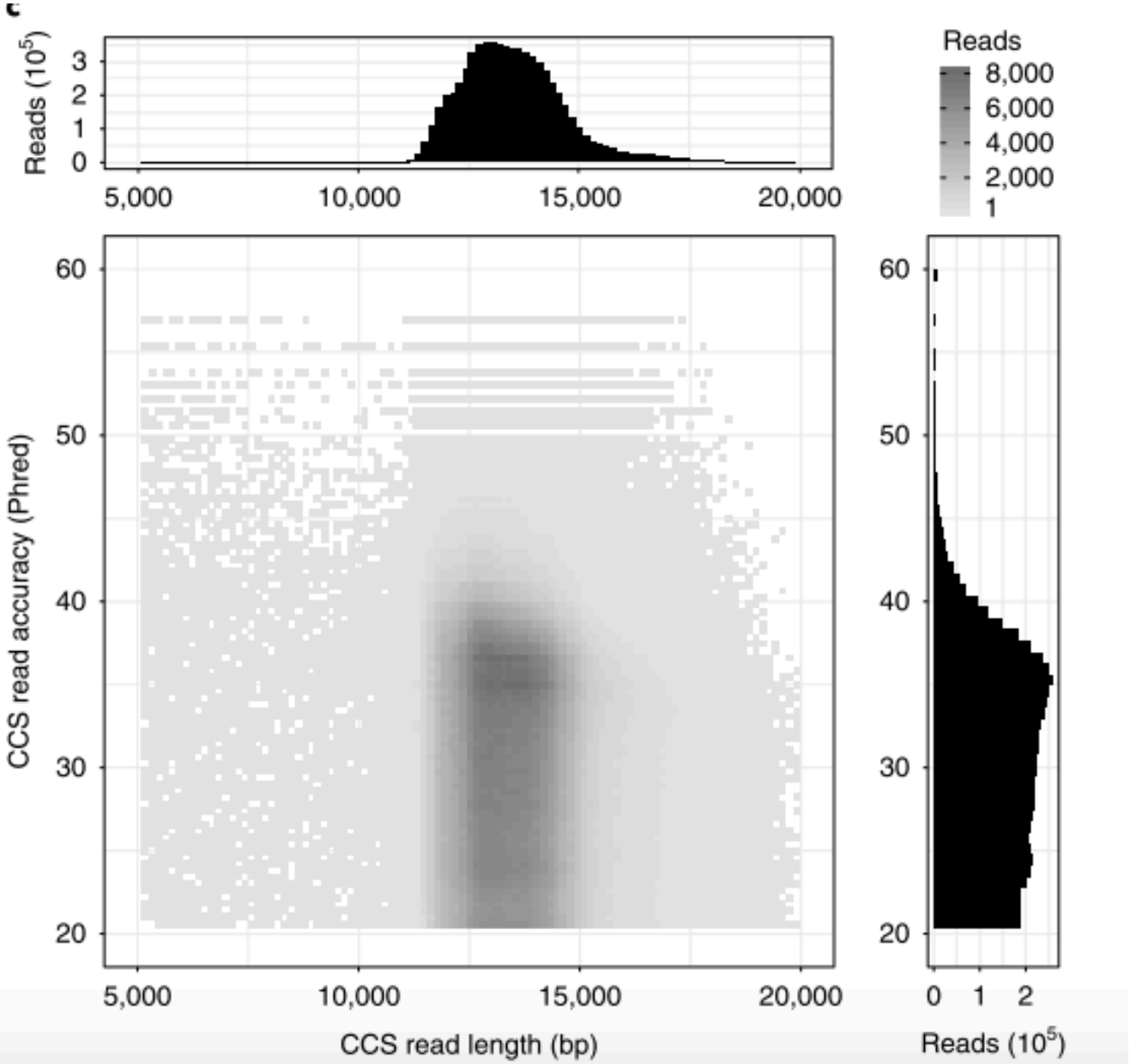
Circular Consensus Sequences (CCS): HIFI READS



RECENT IMPROVEMENT: GENOME ASSEMBLY WITH CCS

Circular consensus assembly of a human genome

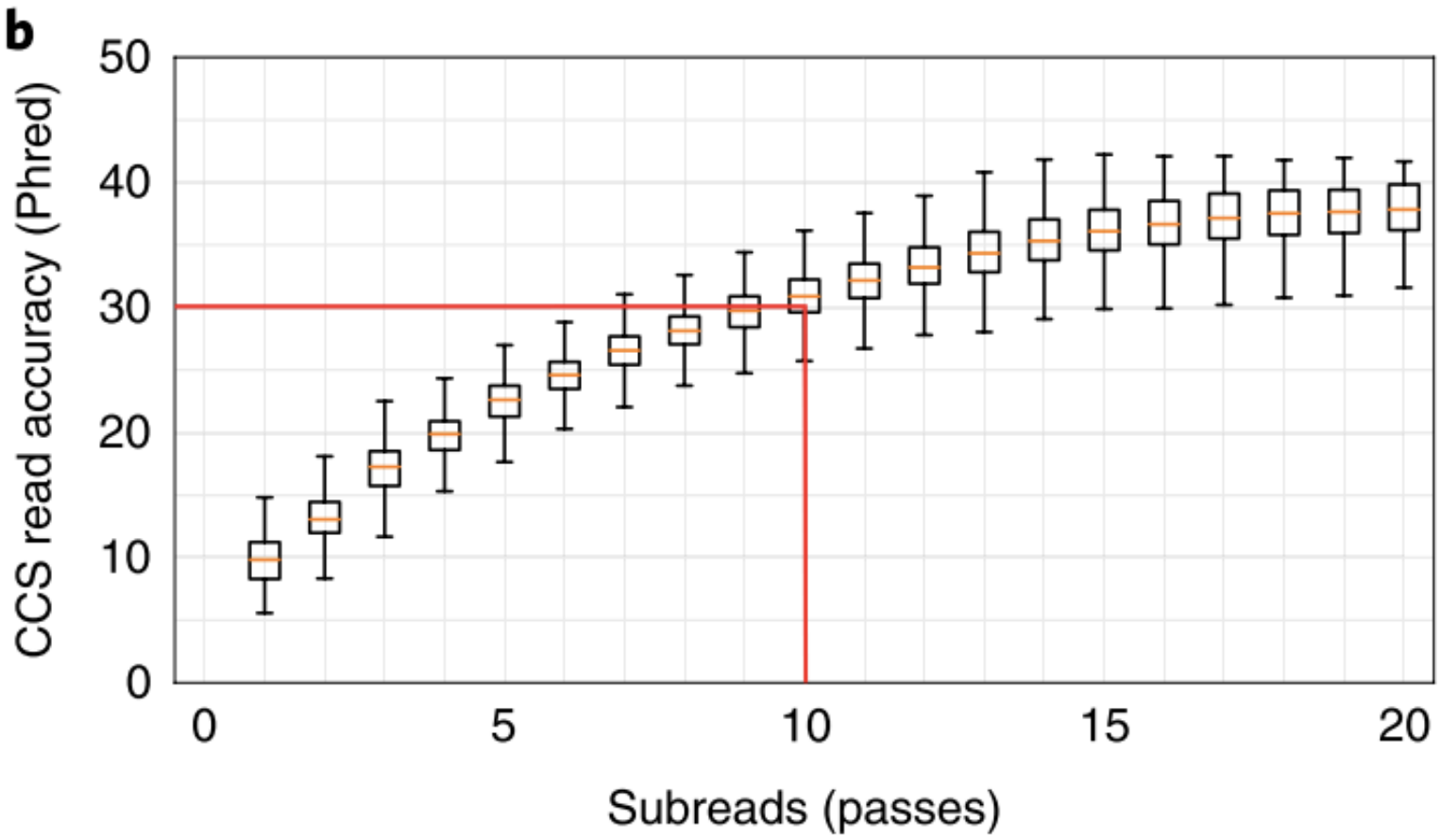
Wenger et al. *Nat. Biotechnol.* oct. 2019



RECENT IMPROVEMENT: GENOME ASSEMBLY WITH CCS

Circular consensus assembly of a human genome

Wenger et al. *Nat. Biotechnol.* oct. 2019



Genome assembly with CCS

Circular consensus assembly of a human genome
Wenger et al. *Nat. Biotechnol.* oct. 2019

CCS reads alone : high quality contiguous genome : concordance of 99.997%

Assembler	Total size (Gb)	Contigs	N50 (Mb)	Ensembl genes (%)
Canu	3.42	18,006	22.78	93.2
FALCON	2.91	2,541	28.95	97.6
wtdbg2	2.79	1,554	15.43	96.1

Canu assembly

- genome size > expected haploid genome because it resolves some heterozygous alleles into separate contigs

Majority of CCS read discordances

- 3.4% mismatches → 1 mismatch every 13,048 bp
- 4.6% indels in non homopolymers. → 1 non-homopolymer indel every 9,669 bp
- 92.0% indels in homopolymers → 1 homopolymer indel every 477 bp

Comparison with NovaSeq

- CCS mismatch rate is 17× lower than reads from NovaSeq
- CCS indel rate is 181× higher than reads from NovaSeq

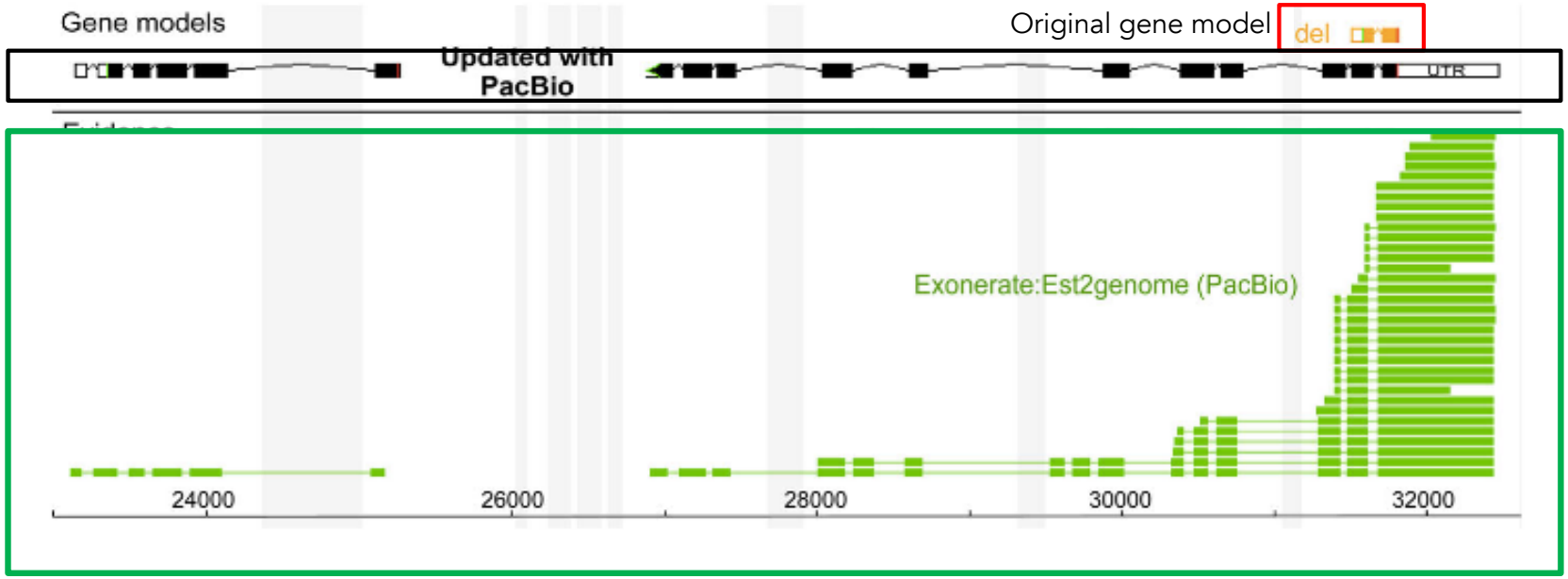
SEQUENCING cDNA USING CIRCULAR CONSENSUS SEQUENCES

Genome annotation of the parasitic hookworm *Ancylostoma ceylanicum* using single molecule mRNA sequencing

Magrini et al. *BMC Genomics*, 2018

RNA → cDNA —*PacBio* sequencing→ 193 000 CCS

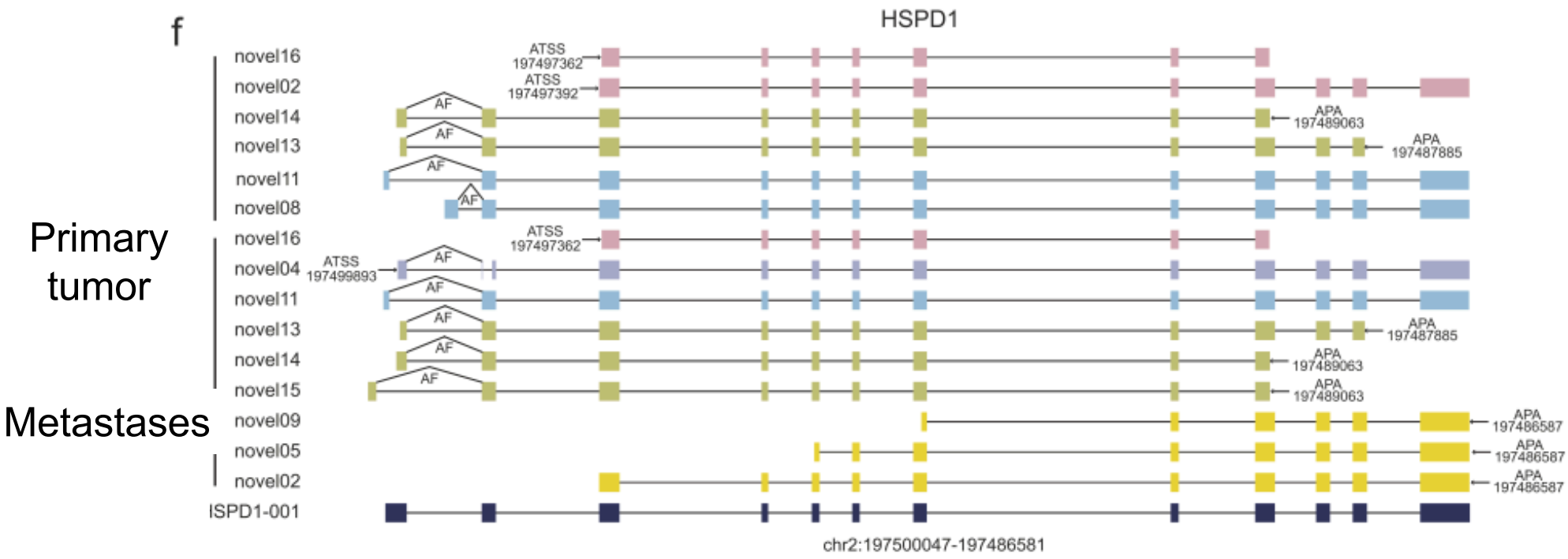
- Increased the total genomic exon length by 1.9 Mb (12.4%)
- 1609 (9.2%) new genes



PacBio cDNA SEQUENCING

Hybrid full-length transcriptome in metastatic ovarian cancer

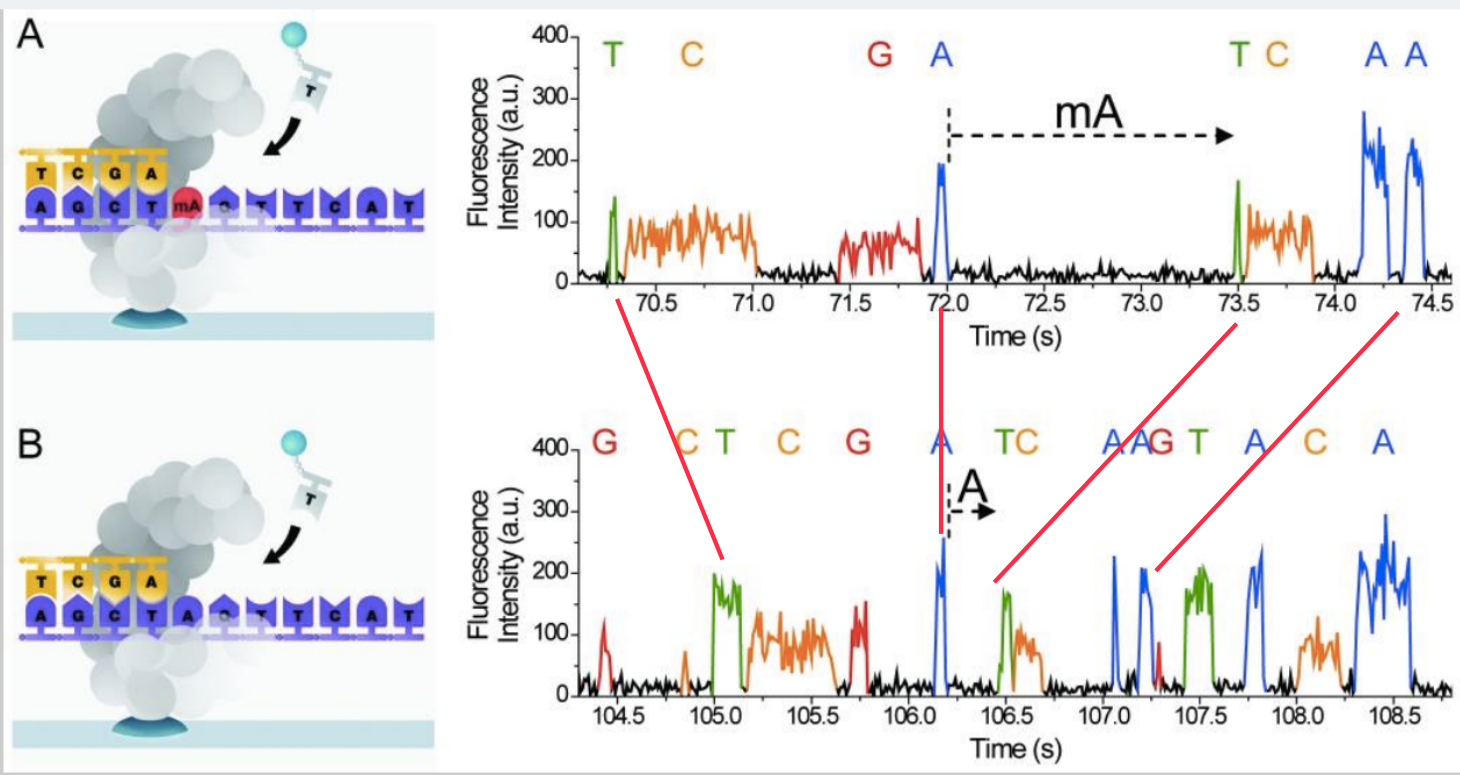
Jing et al. *Oncogene* 2019



Long-read full-length transcriptome analysis

- improves molecular diagnostic
- reveals novel therapeutic vulnerabilities

DETECTION OF MODIFIED BASES



from Fusberg et al. *Nature Methods* (2010)

Detection of 5mA with strong influence of sequence contexts: requires high coverage

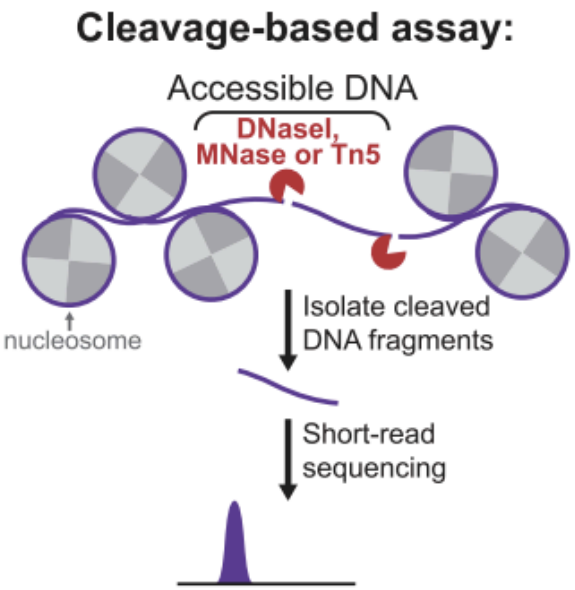
Feng et al. *PLOS Comput Biol* (2013)

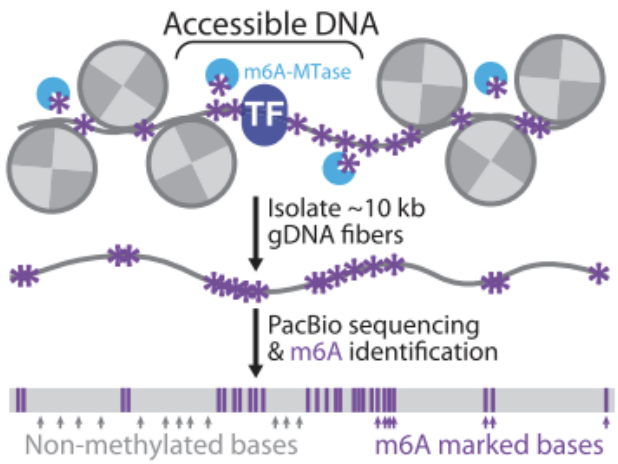
Detection of m6A with CCS

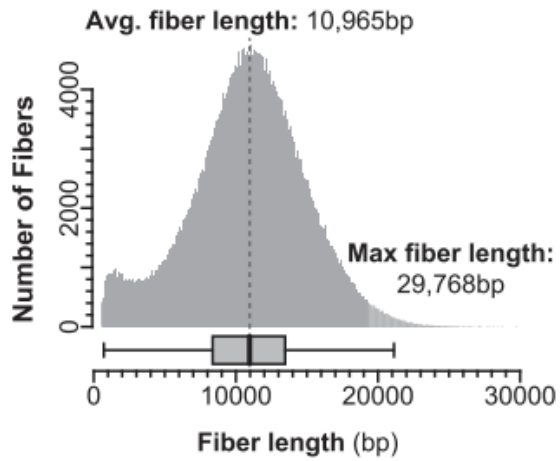
Single-molecule regulatory architectures captured by chromatin fiber sequencing
Stergachis et al. *Science* (2020)

DnaseI-seq.

Fiber-seq.



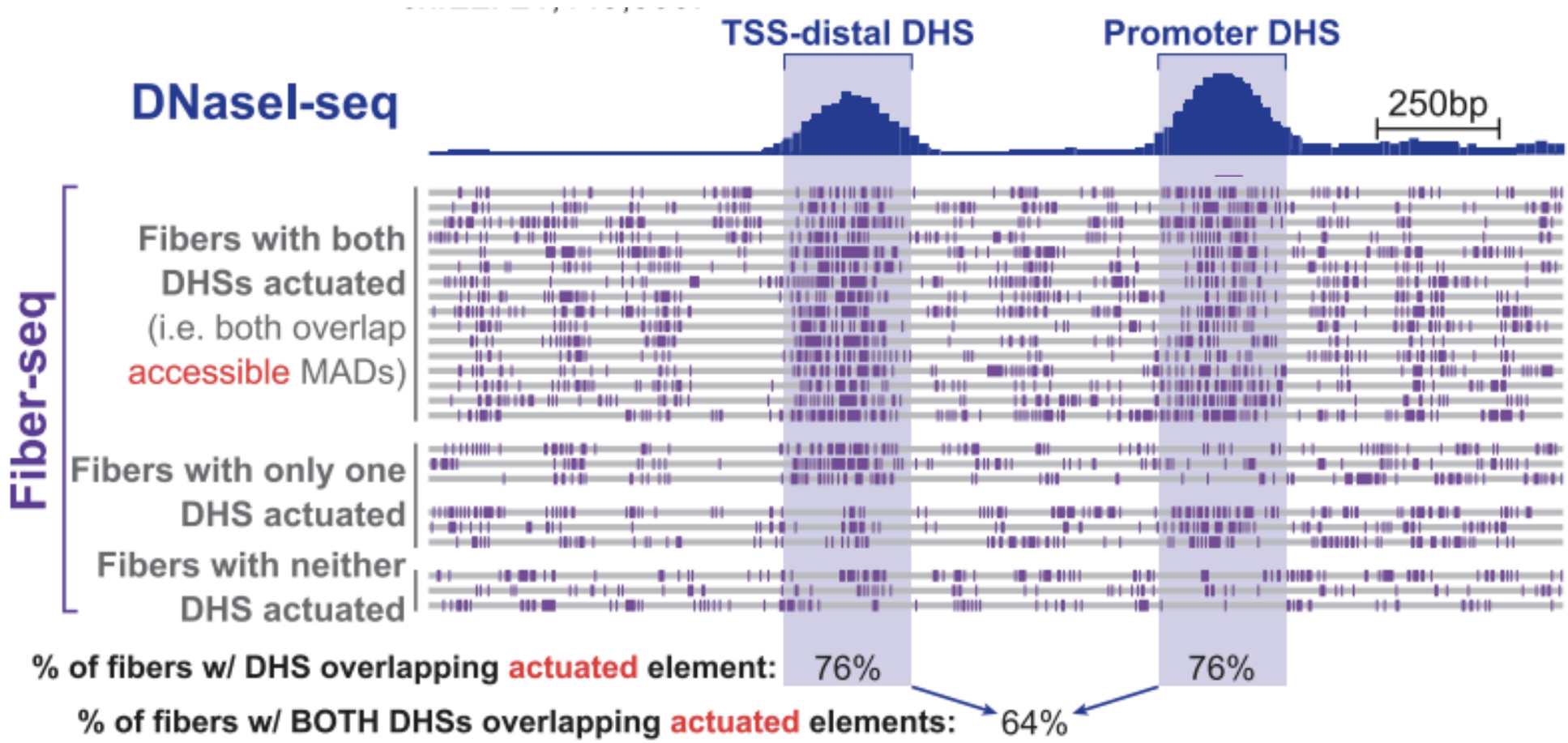




Detection of m6A with CCS

Single-molecule regulatory architectures captured by chromatin fiber sequencing
Stergachis et al. *Science* (2020)

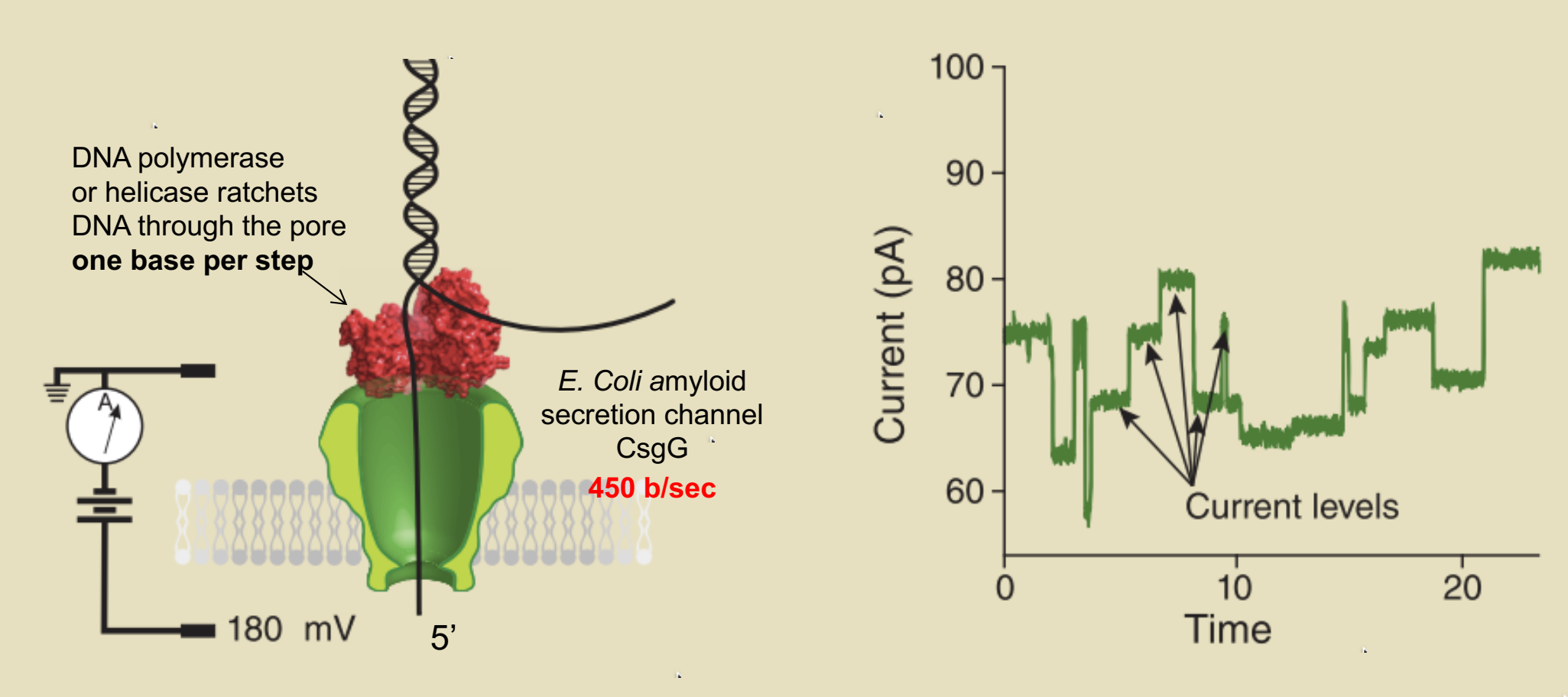
DHS : DNaseI Hypersensitive Site



Next Generation Sequencing

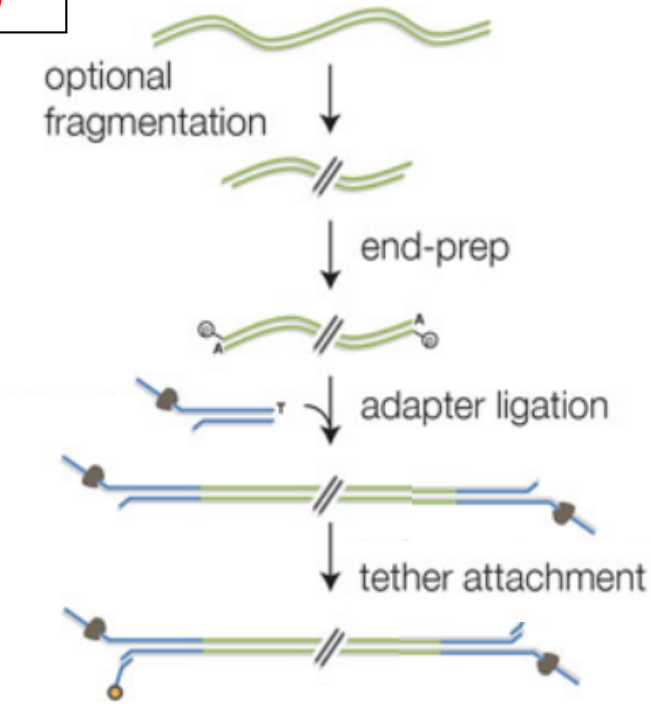
Oxford NANOPORE Technologies®

BASIC CONCEPTS

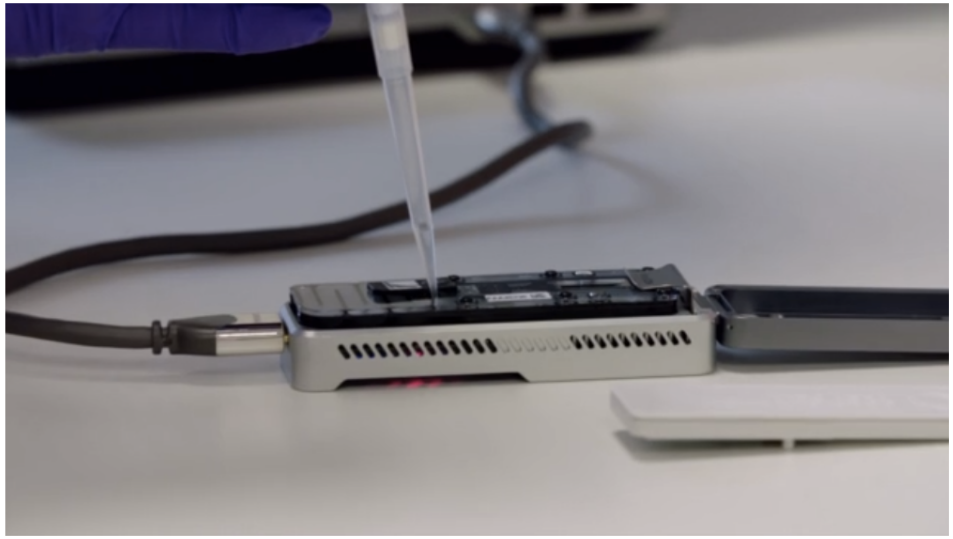


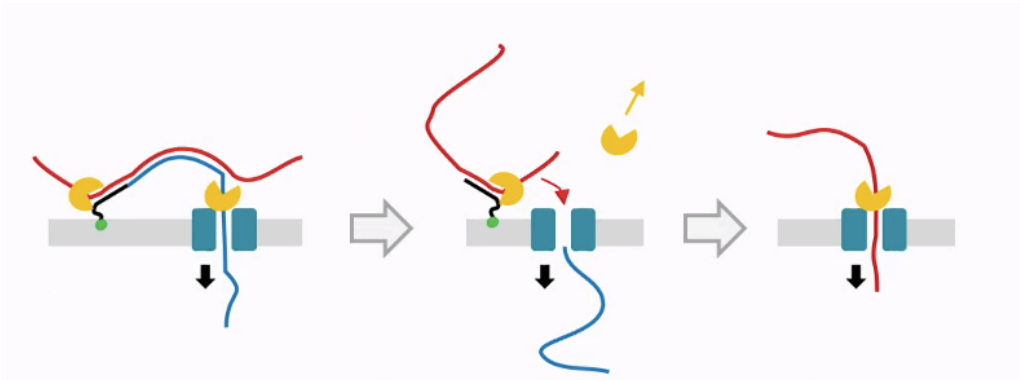
SEQUENCING PROCESS

1D^2 Library
(2017)

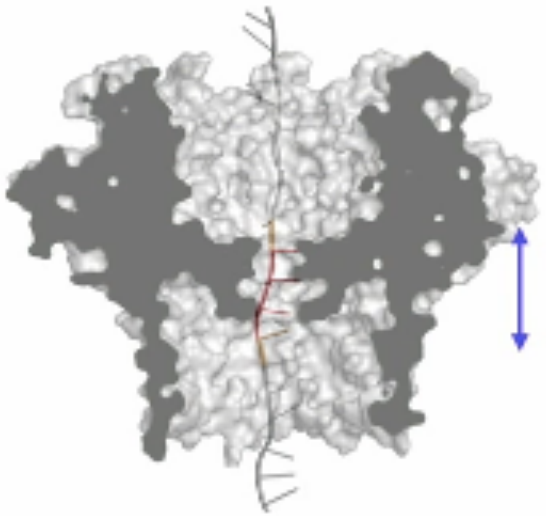


SEQUENCING



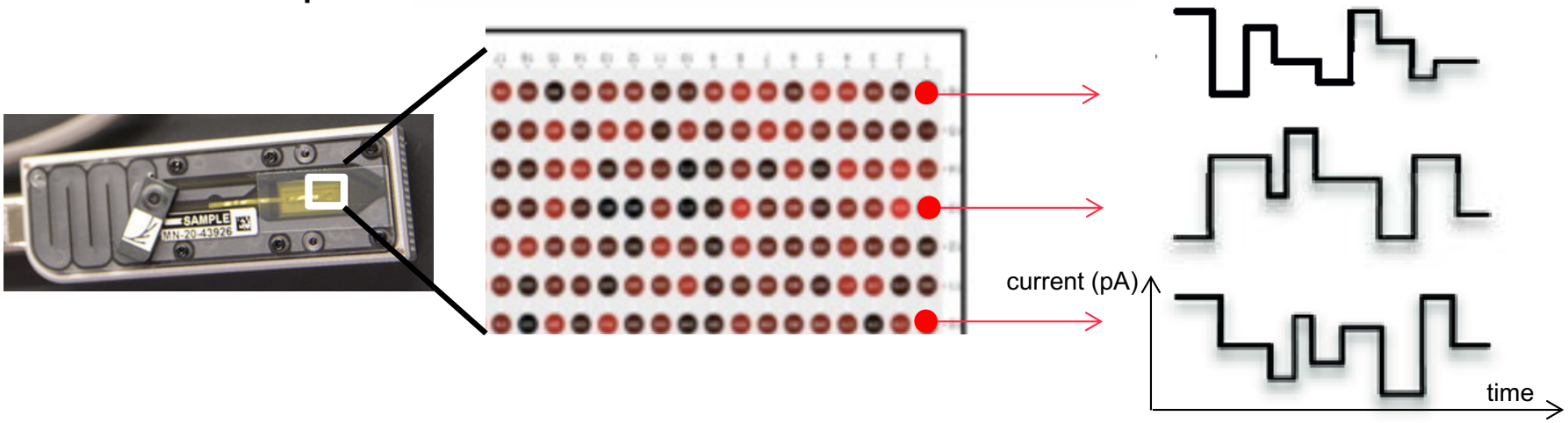


SEQUENCING PROCESS : MinION FLOW CELL



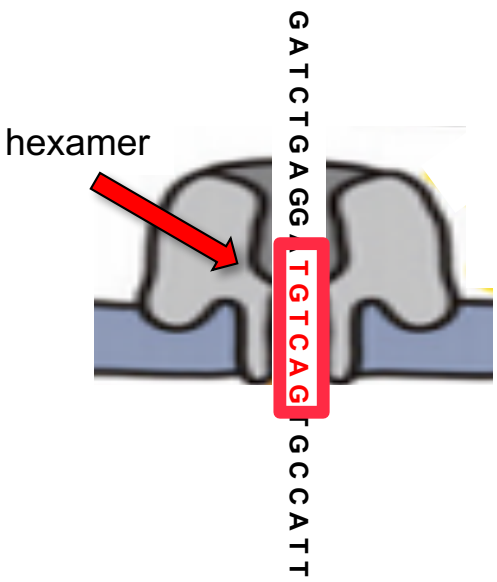
5-6 bases dominate the current signal

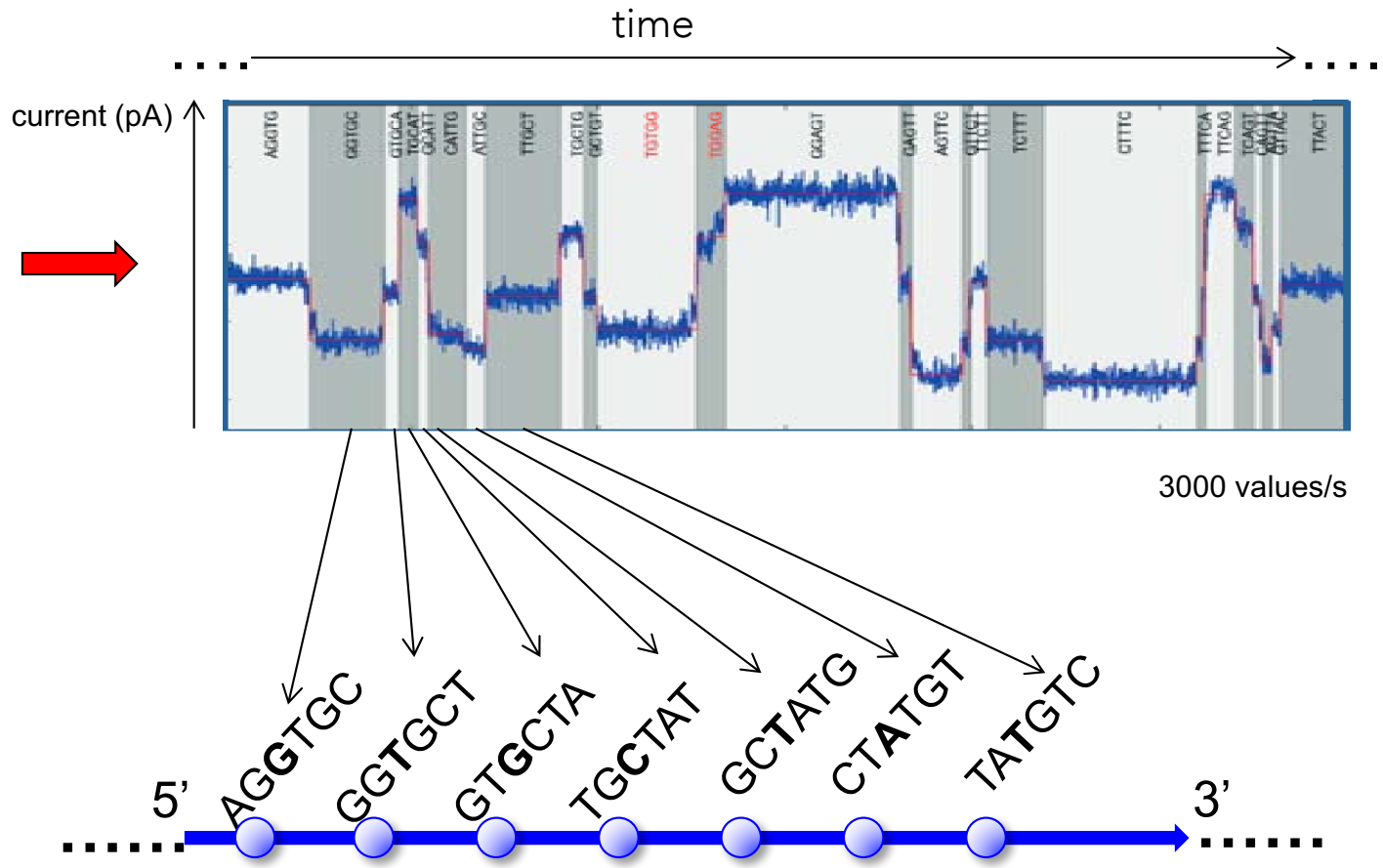
MinION : 512 pores



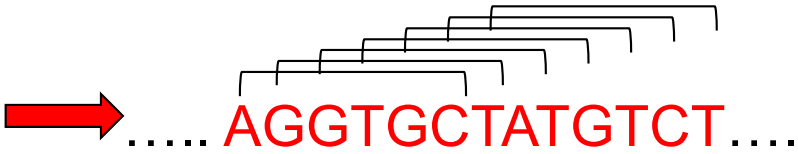
PromethION : 144000 pores (48 x 3000)

BASECALLING

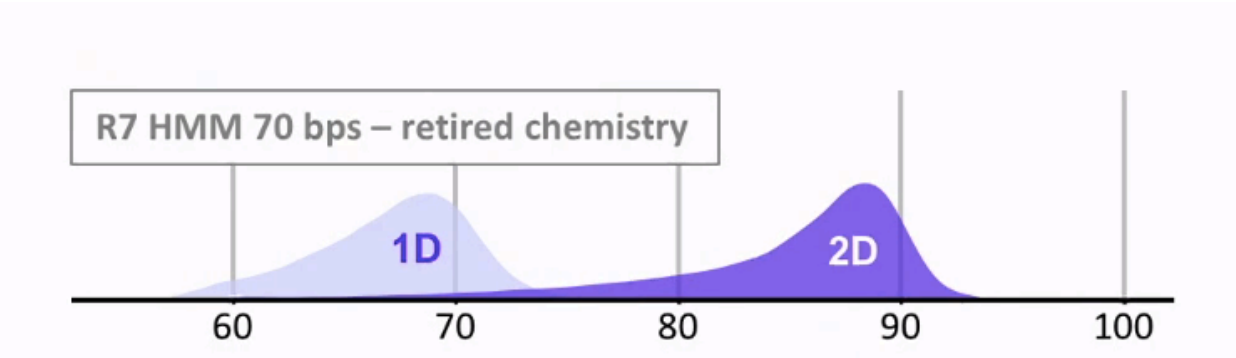


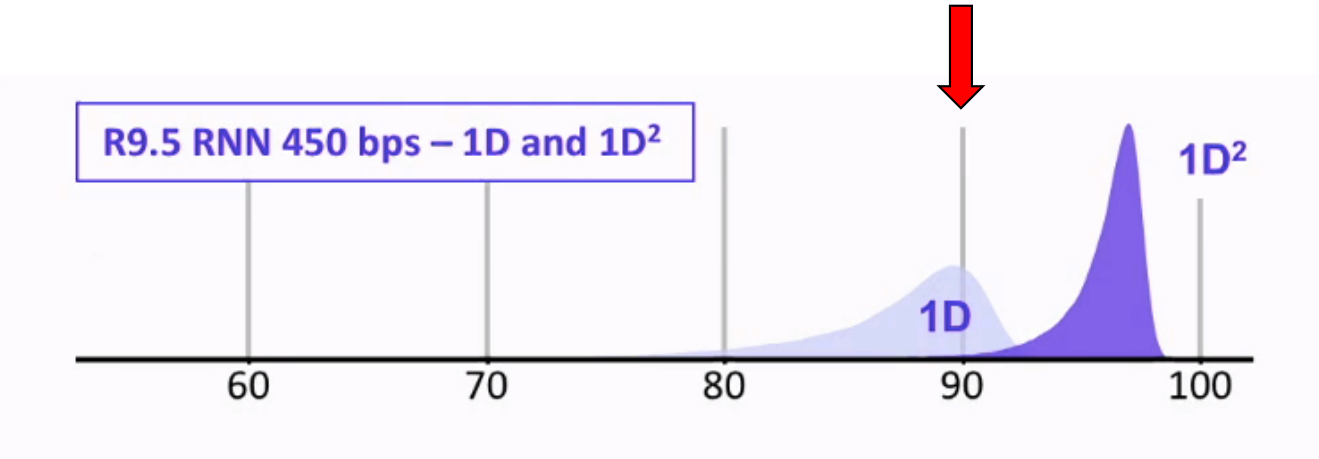


Basecalling : finding the optimal path of successive 6-mers



QUALITY

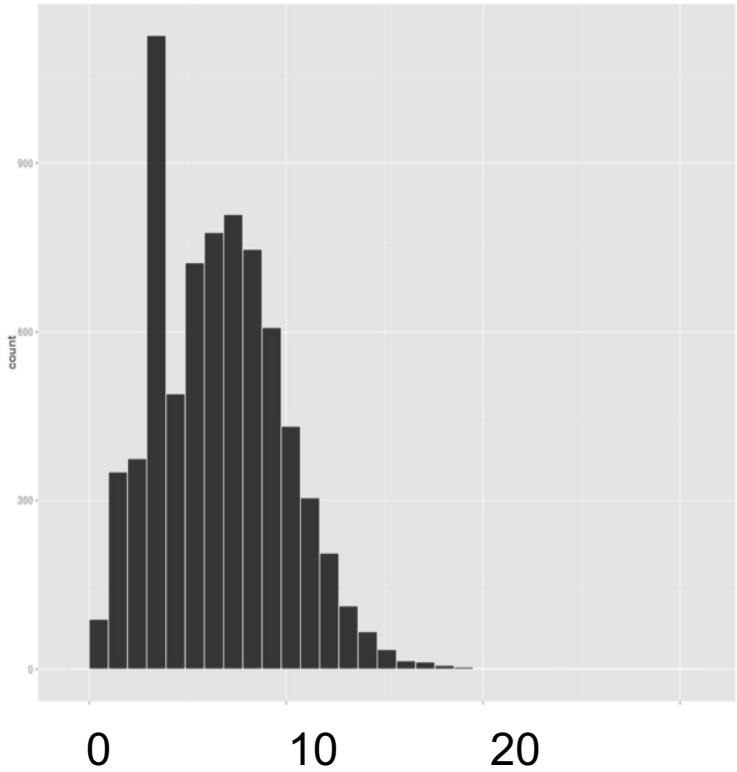




Homopolymers difficult to sequence

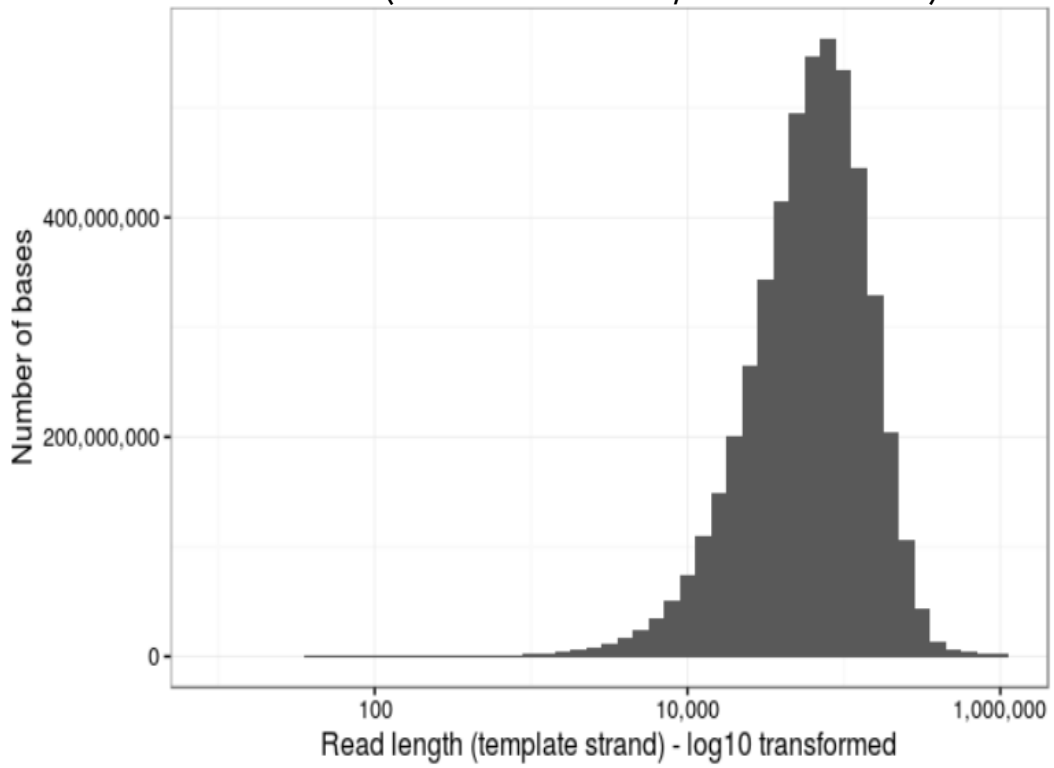
SIZE OF SEQUENCED DNA FRAGMENTS

Typical profile of fragment size



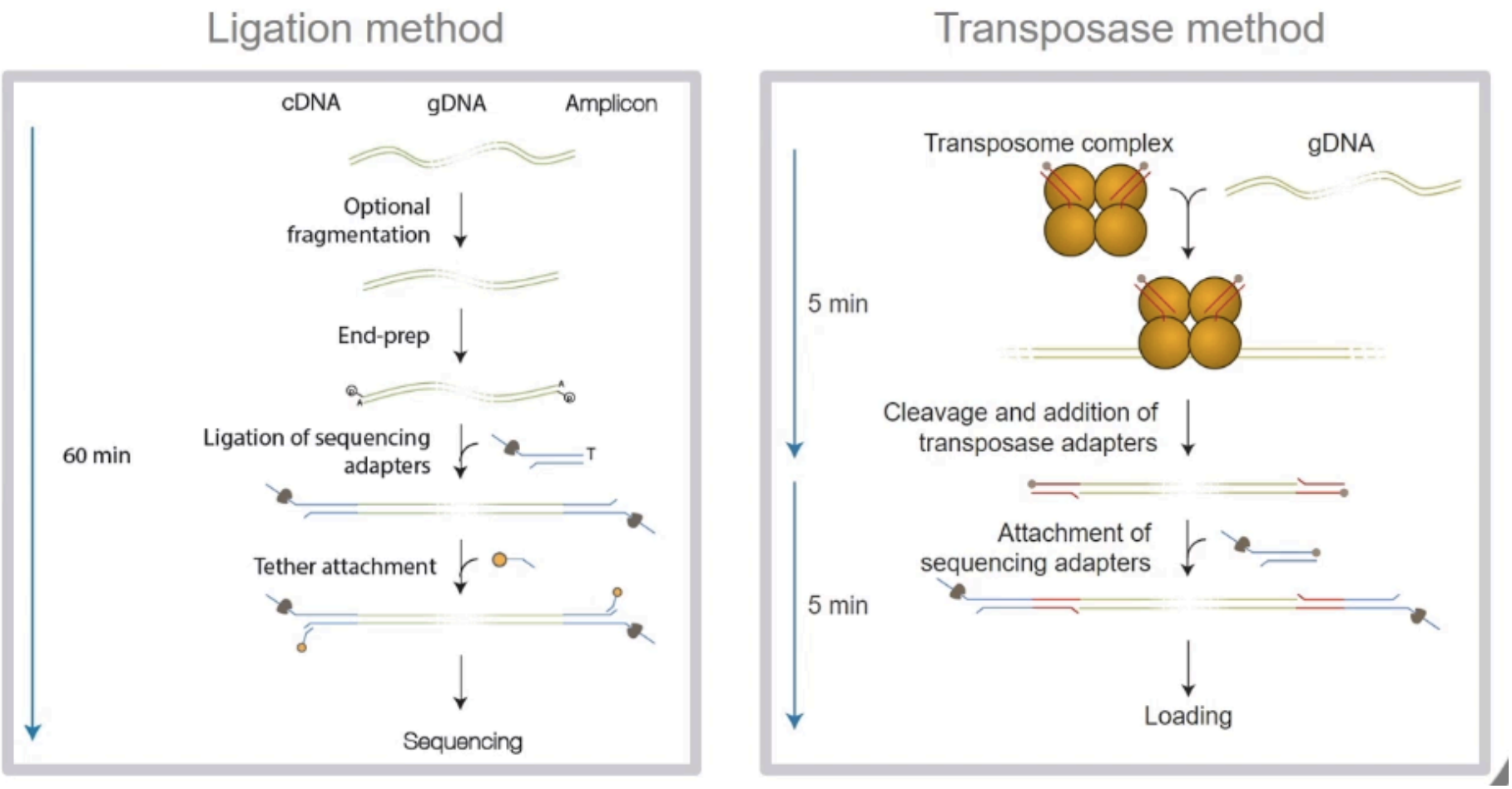
(Risse et al. *GigaScience*, 2015)

"Ultra long" reads
(lab.loman.net, March 2017)

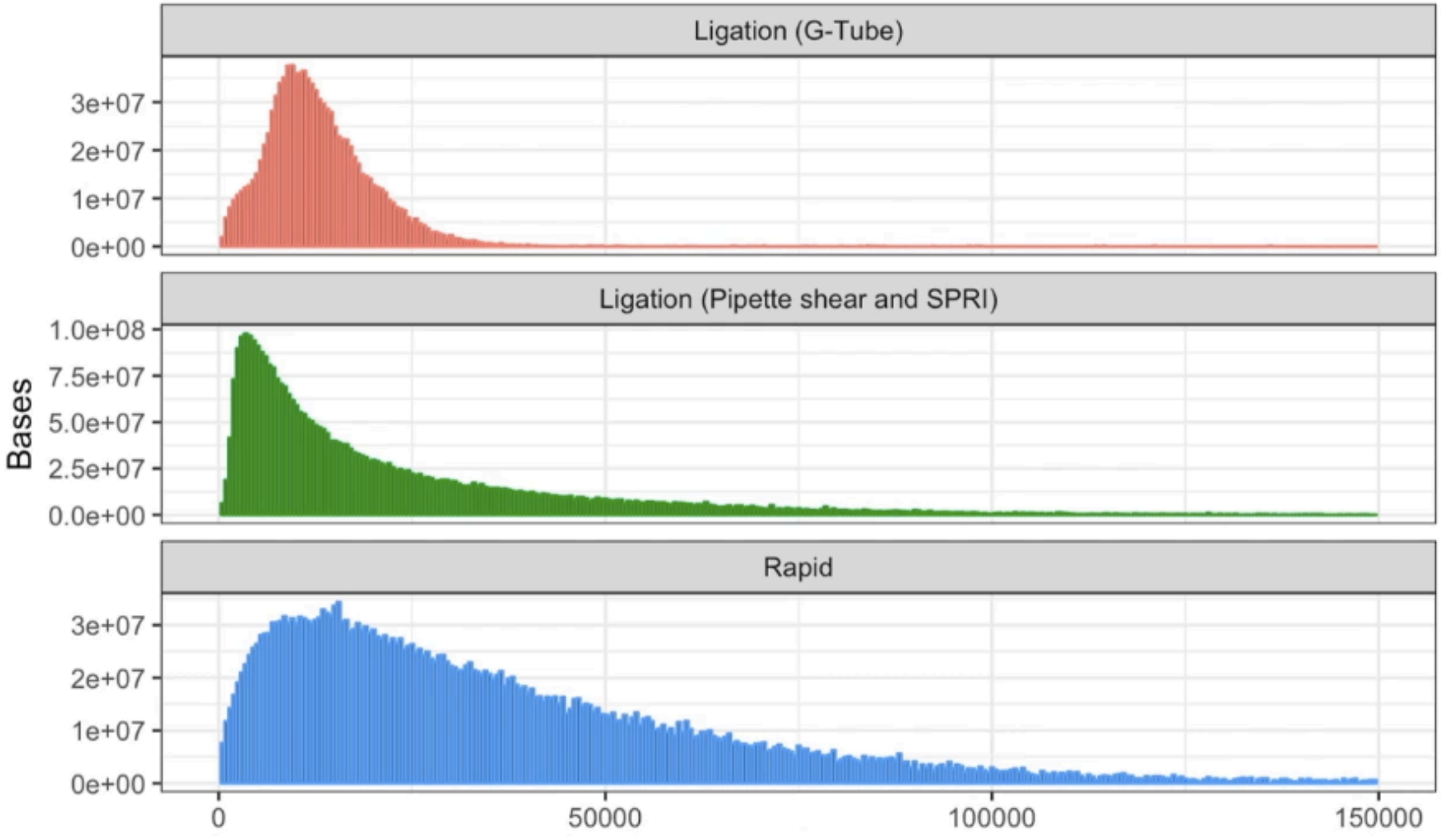


Size of the longest read : 778 kb
1 contig of the 4.6Mb chromosome of *E. coli*
obtained with just the 7 longest reads

SIZE OF SEQUENCED DNA FRAGMENTS



SIZE OF SEQUENCED DNA FRAGMENTS



Josh Quick, Nick Loman

see John Tyson's video (ONT website)

HYBRID GENOME ASSEMBLY : NANOPORE AND ILLUMINA DATA

Acinetobacter baylyi (data from Oxford Nanopore)

Assemblies	Illumina only	Illumina + MinION
Input Coverage	50X	13X
# contigs	20	1
Assembly size (Mb)	3.59	3.62
N90 size (Kb)	326	3 621
NA75 size (Kb)	194	1 002
Genome fraction (%)	99.73	99.997
# misassemblies	4	2
# local misassemblies	3	4
# mismatches per 100 Kb	6.49	3.11
# indels per 100 Kb	0.33	0.14

NANOPORE GENOME SEQUENCING

ARTICLES

nature biotechnology

OPEN

Jan. 2018

Nanopore sequencing and assembly of a human genome with ultra-long reads

Miten Jain[1,13], Sergey Koren[2,13], Karen H Miga[1,13], Josh Quick[3,13], Arthur C Rand[1,13], Thomas A Sasani[4,5,13], John R Tyson[6,13], Andrew D Beggs[7], Alexander T Dilthey[2], Ian T Fiddes[1], Sunir Malla[8], Hannah Marriott[8], Tom Nieto[7], Justin O'Grady[9], Hugh E Olsen[1], Brent S Pedersen[4,5], Arang Rhie[2], Hollian Richardson[9], Aaron R Quinlan[4,5,10], Terrance P Snutch[6], Louise Tee[7], Benedict Paten[1], Adam M Phillippy[2], Jared T Simpson[11,12], Nicholas J Loman[3] & Matthew Loose[8]

Using nanopore reads alone assembly of a human genome :

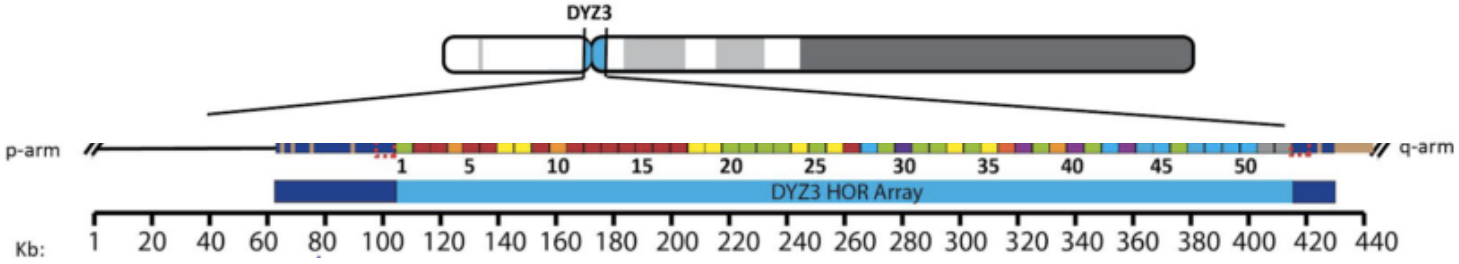
- NG50 contig size of ~6.4 Mb
- covers >85% of the reference
- 99.88% accuracy
- MHC locus on a single contig, phased over its full length
- closure of 12 large (>50 kb) gaps in the reference human genome

ASSEMBLY OF A HUMAN Y CENTROMERE

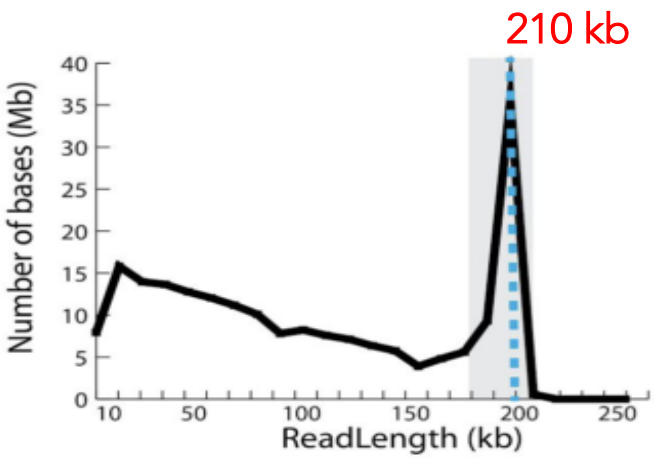
(Jain et al., *bioRxiv*, 2017)

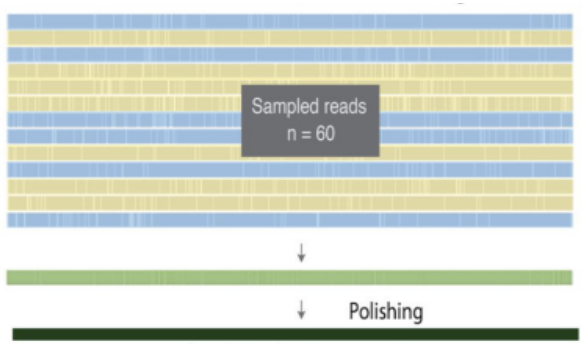
300 kb array of 5.8 kb sequence repeated in an uninterrupted head-to-tail orientation

To date, no technology has been capable of sequencing centromeres due to requirement for extremely high-quality long reads



9 BACs
100kb to 210kb





Final high quality consensus BAC sequence

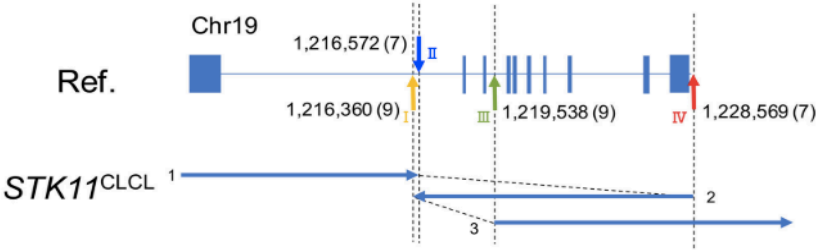
FIRST COMPLETE SEQUENCE OF A HUMAN CENTROMERE

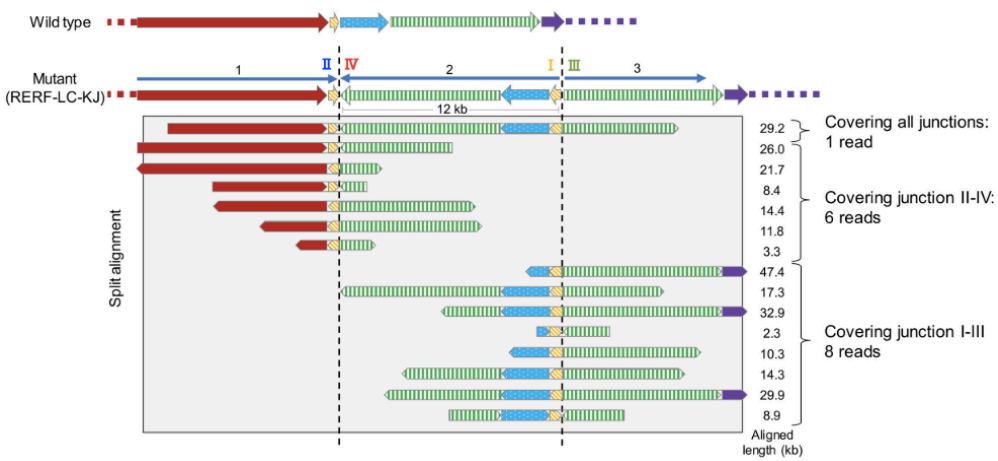
GENOME SEQUENCING

Long-read sequencing for non-small-cell lung cancer genomes
Sakamoto et al. *Genome Research*, Sept. 2020

Nanopore sequencing of cancer cell lines (Promethion)

- Maximum length : 0.99 Mb
- N50 : 32 kb
- Average mapped reads : 14 kb





Biological relevance of SV further revealed by :

- epigenome,
- transcriptome,
- protein analyses

Sequencing of clinical tumor samples

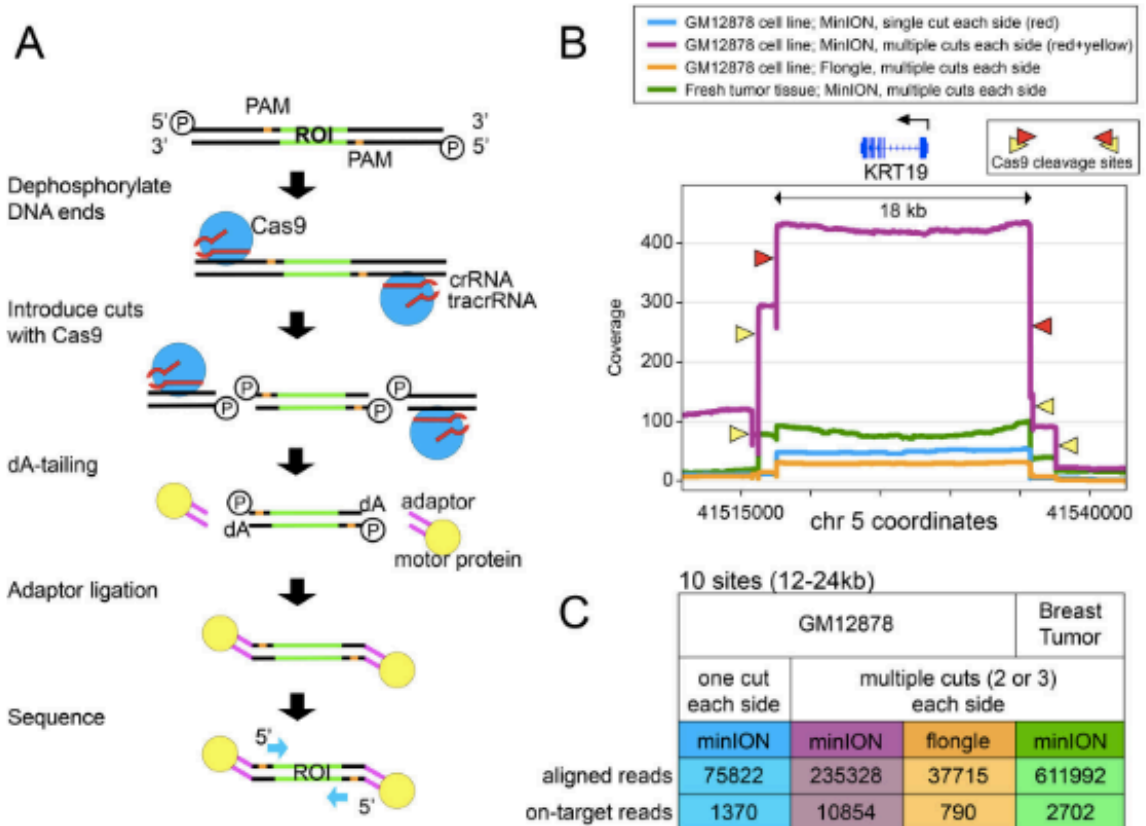
- Structural aberrations also found in clinical lung adenocarcinoma specimens

Structural variants : comparison with PacBio sequencing

"These results indicated that neither the PacBio nor the PromethION platform is currently perfect; therefore, they should be used to complement each other."

GENOME SEQUENCING : TARGETED SEQUENCING

Targeted nanopore sequencing with Cas9-guided adaptor ligation
Gilpatrick et al. *Nature Biotechnology* April 2020



C

10 sites (12-24kb)

	GM12878			Breast Tumor
	one cut each side	multiple cuts (2 or 3) each side		
	minION	minION	flongle	minION
aligned reads	75822	235328	37715	611992
on-target reads	1370	10854	790	2702
on-target percent	1.81%	4.61%	2.09%	0.44%
median coverage	80	680	34	93

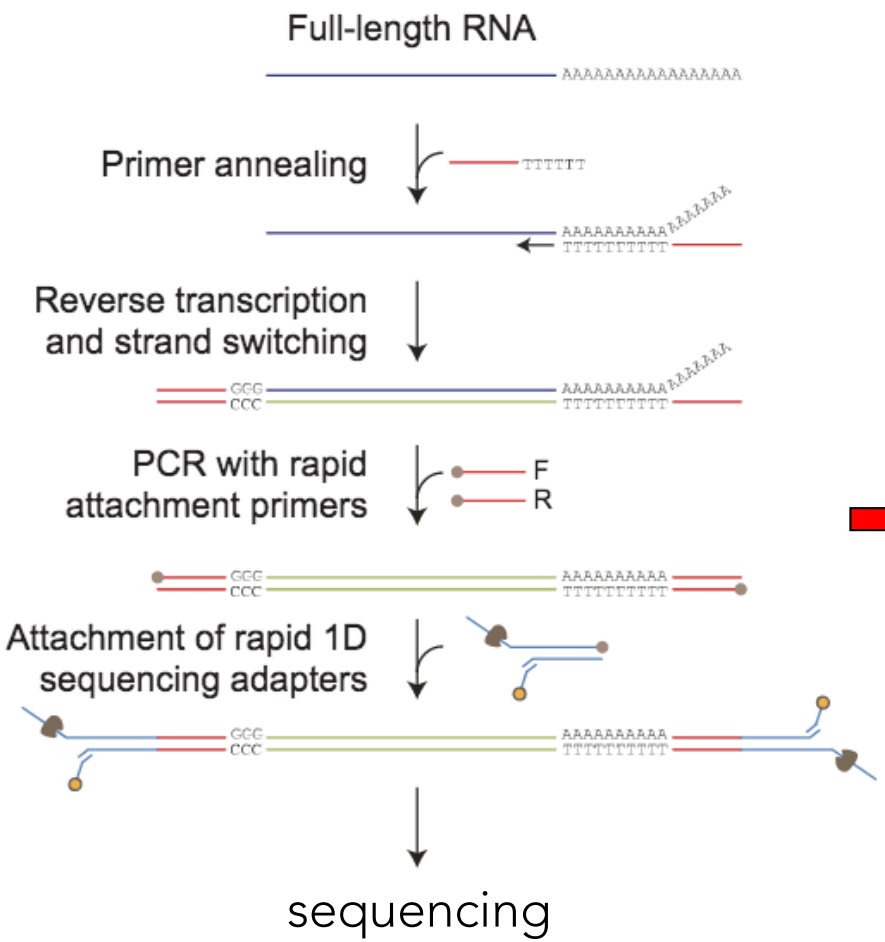
nCATS = nanopore Cas9-targeted sequencing : enrichment strategy using targeted cleavage of DNA to ligate adapters for nanopore

nCATS can simultaneously assess :

- haplotype-resolved single-nucleotide variants (SNVs)
- structural variations (SVs)
- CpG methylation…
- Best median sequencing coverage : 680 X
- nCATS uses only ~3 µg of genomic DNA + can target a large number of loci in a single reaction.

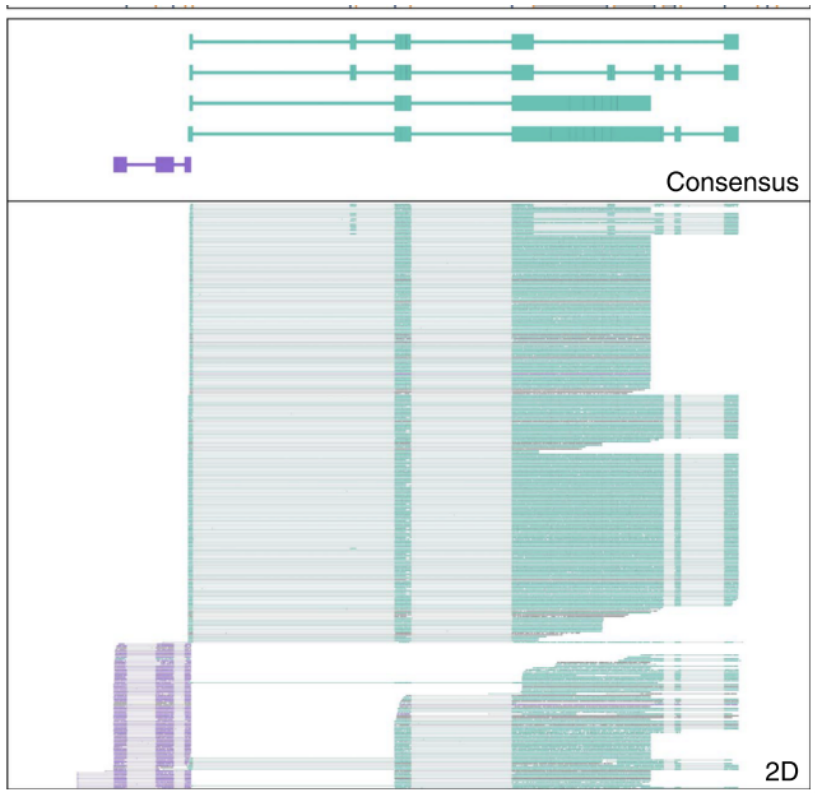
cDNA SEQUENCING

Library preparation



Detection of splice variants in surface receptor of B cells

(Byrne et al. *Nat. Comm.* 2017)

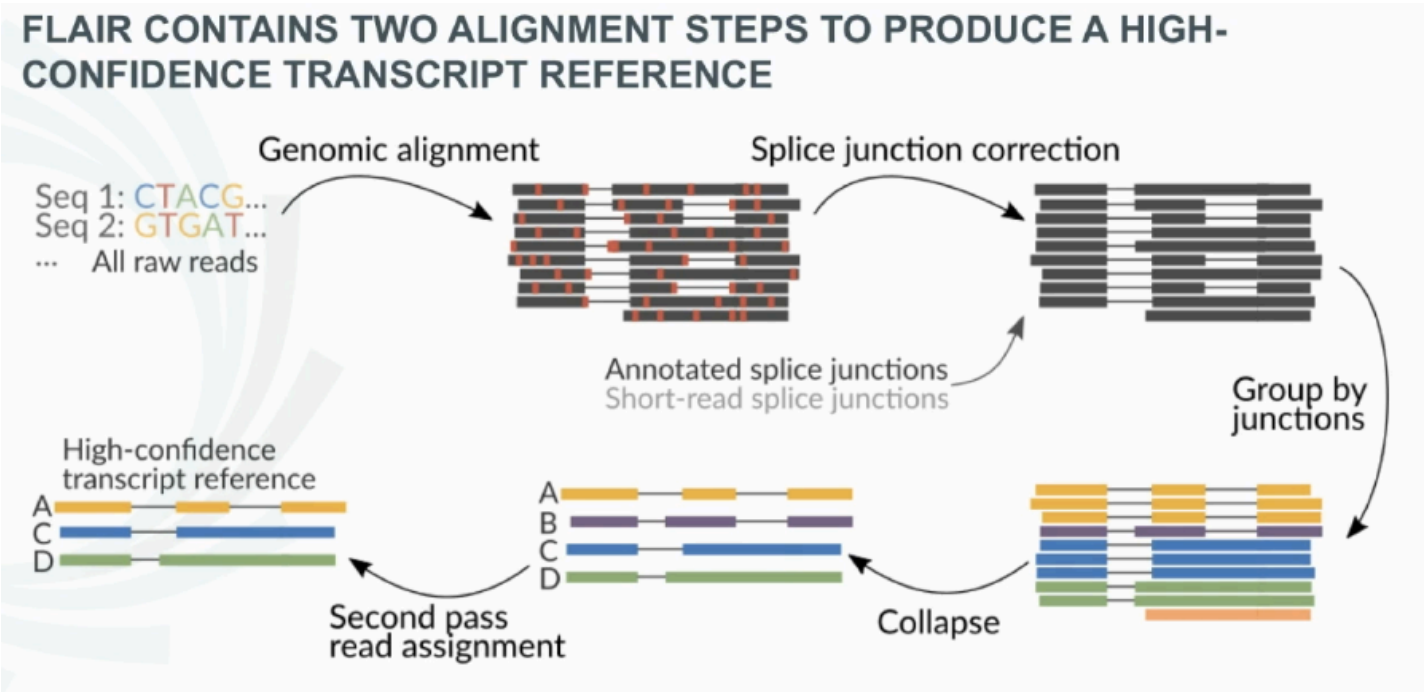


- Splice alignment uneasy due to high (5-10%) error rate
- Reads are frequently truncated from 5' end

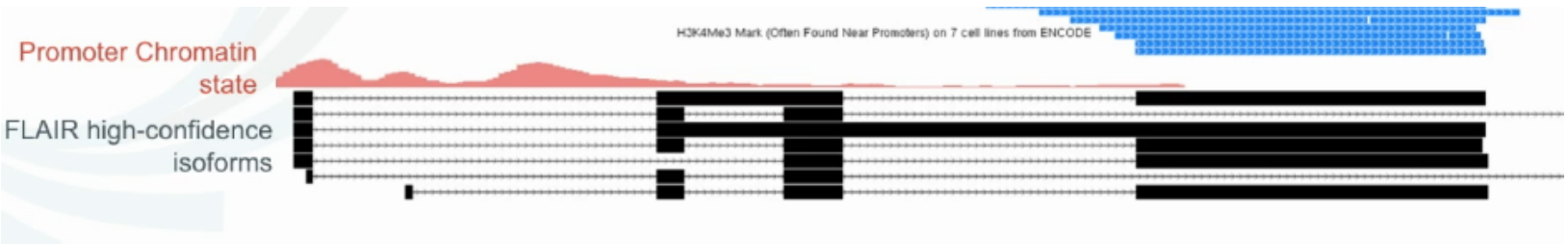
CHALLENGES OF NANOPORE TRANSCRIPTOME ANALYSIS

FLAIR : a pipeline for splicing isoform determination

Tang et al. *bioRxiv* 2018



FLAIR incorporates promoter chromatin states to distinguish 5' truncations from true novel start sites



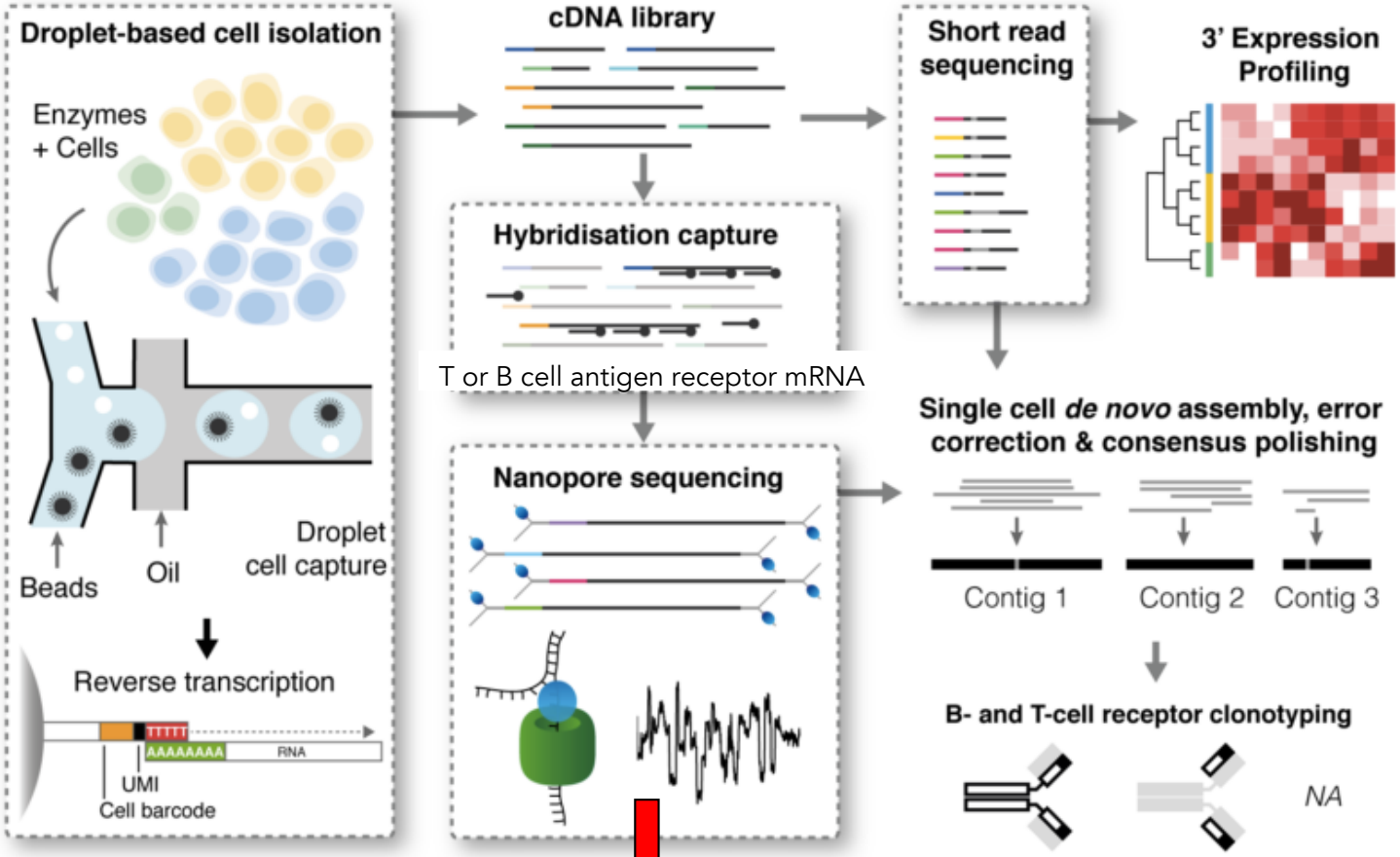
cDNA from chronic lymphocytic leukemia (CLL)

NANOPORE and SINGLE CELL cDNA SEQUENCING

High-throughput targeted long-read single cell sequencing reveals the clonal and transcriptional landscape of lymphocytes
Singh et al., *bioRxiv*, 2018

RAGE-seq (Repertoire And Gene Expression sequencing) : combines targeted long-read sequencing with short-read transcriptome of barcoded single cell libraries

10x Chromium Single Cell 3' (V2 chemistry)



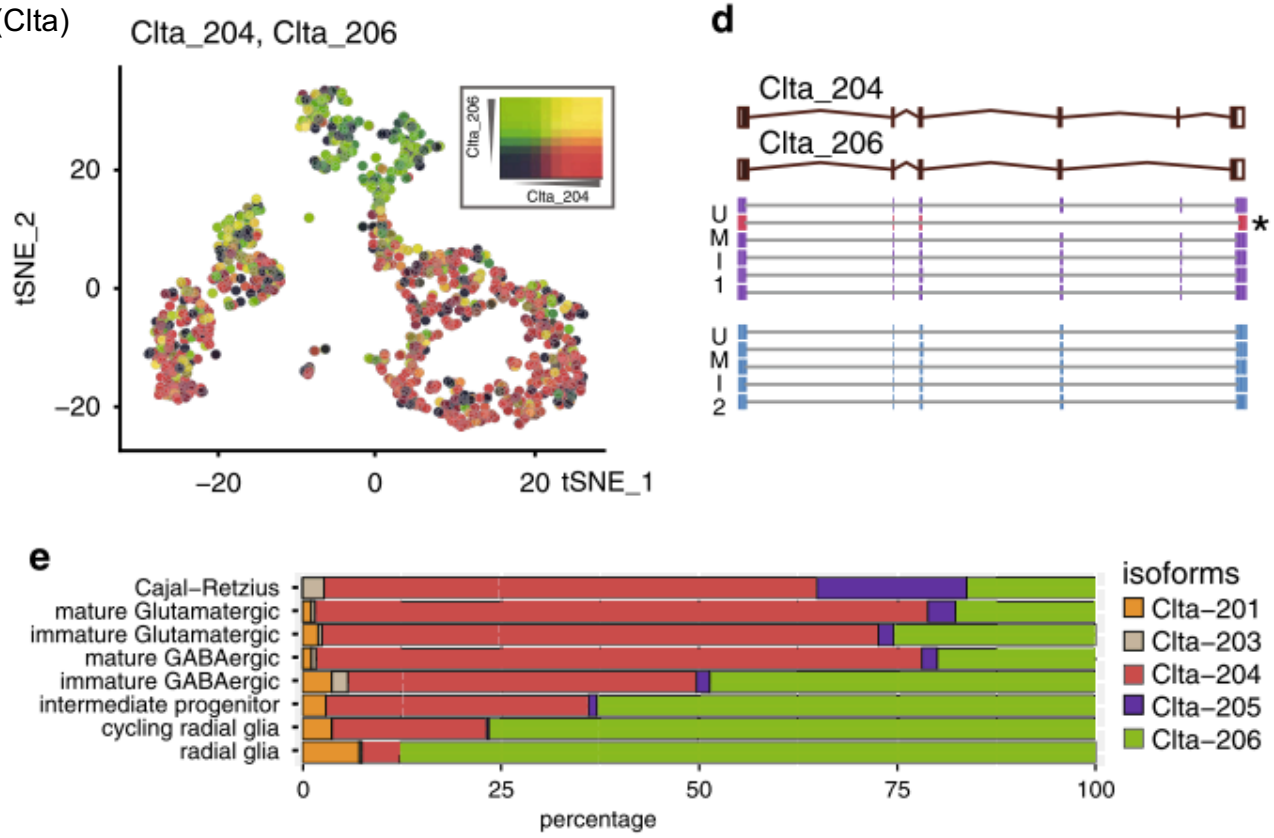
Tracking of somatic mutation, alternate splicing and clonal evolution of T and B lymphocytes

NANOPORE and SINGLE CELL cDNA SEQUENCING

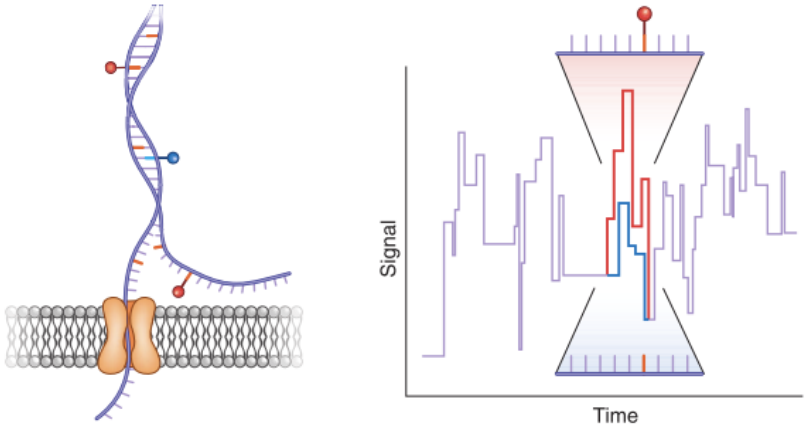
High-throughput targeted long-read single cell sequencing reveals the clonal and transcriptional landscape of lymphocytes
Lebrigand et al., *Nature Communications*, 2020

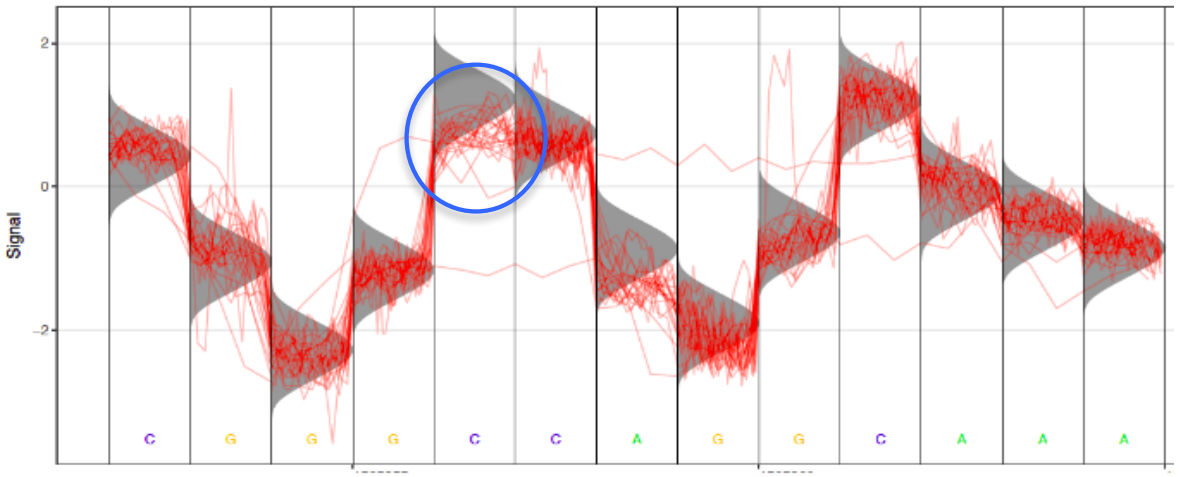
ScNaUmi-seq : Single-cell Nanopore sequencing with UMIs (10x Genomics Chromium system)
- High accuracy cell BC and UMI assignment
- Analysis of splicing and sequence variation at the single-cell level

Clathrin light chain A (Clta)



DETECTION OF MODIFIED BASES



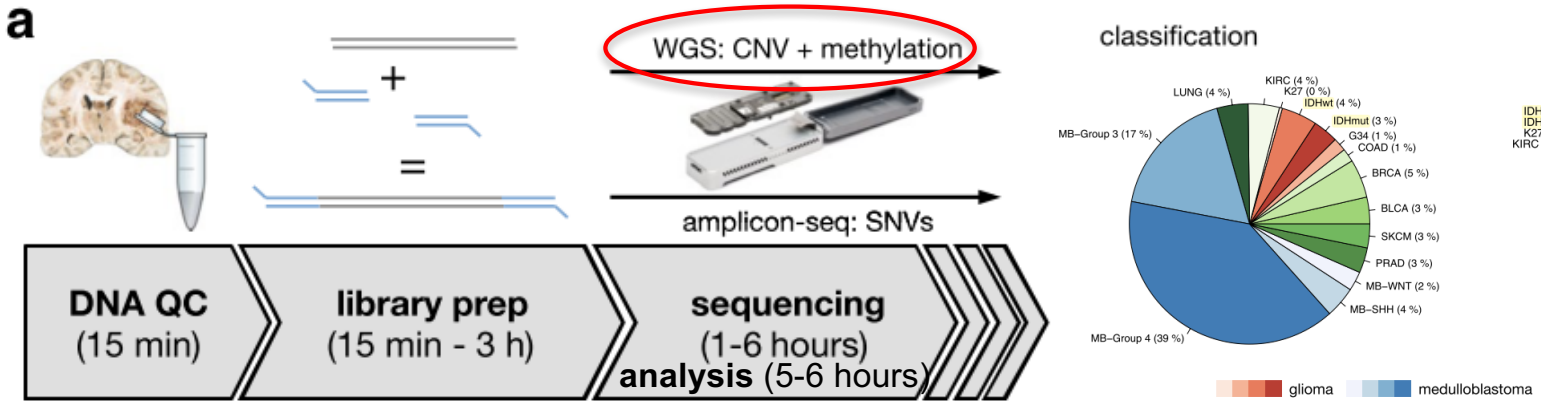


Electric signal

Canonical base distribution

DETECTION OF MODIFIED BASES IN CANCER GENOMES

Same-day genomic and epigenomic diagnosis of brain tumors (gliomas, medulloblastomas) with nanopore sequencing
Euskirchen et al., *Acta Neuropathol.* (2017)



Same-day detection of :

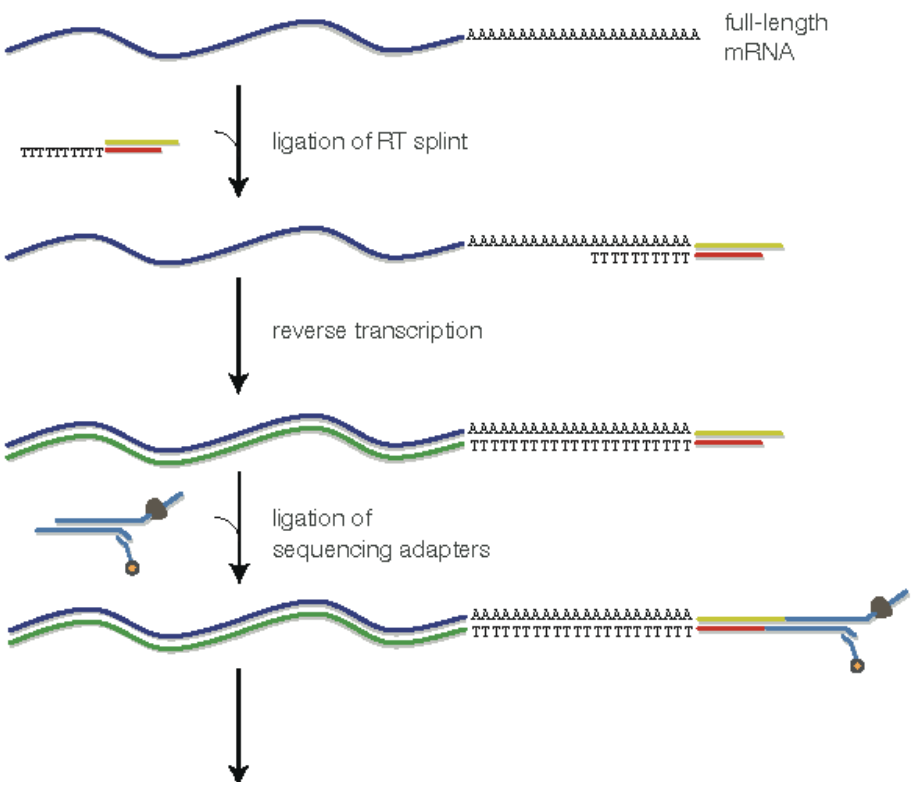
- structural variants
- point mutations
- methylation profiling

Single device with negligible capital cost :

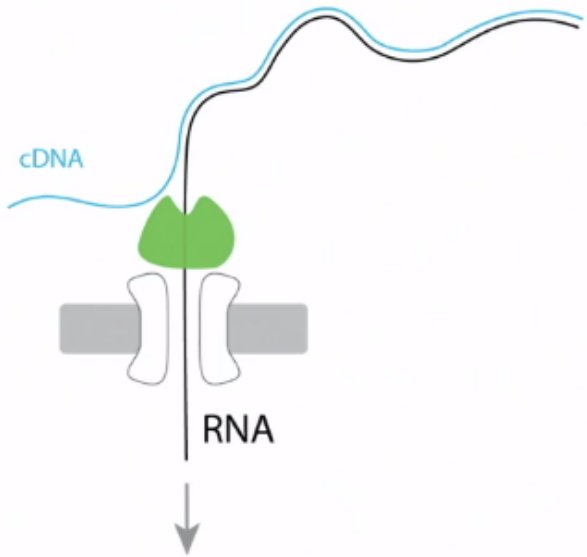
- outperforms hybridization-based and current sequencing technologies
- makes precision medicine possible for every cancer patient

DIRECT RNA SEQUENCING

Library preparation



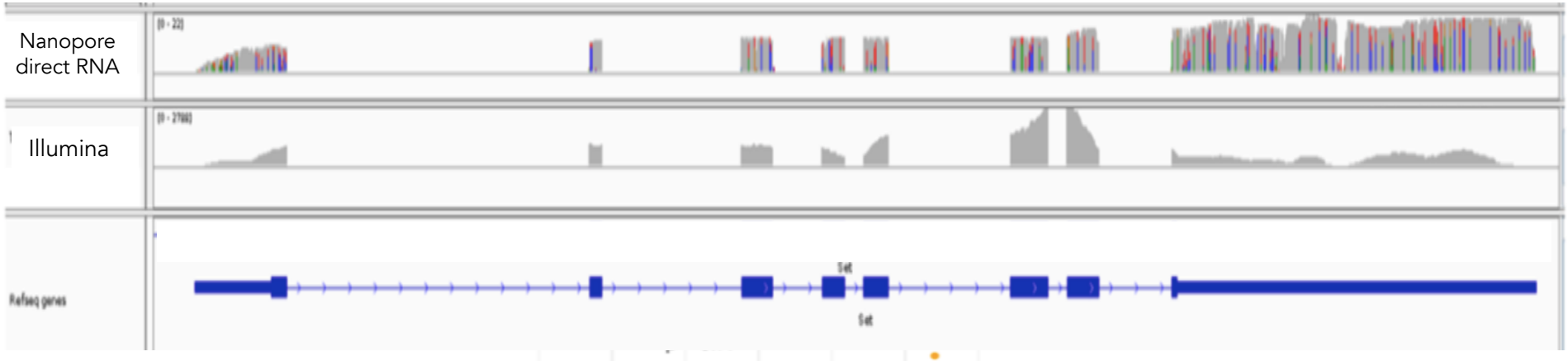
sequencing

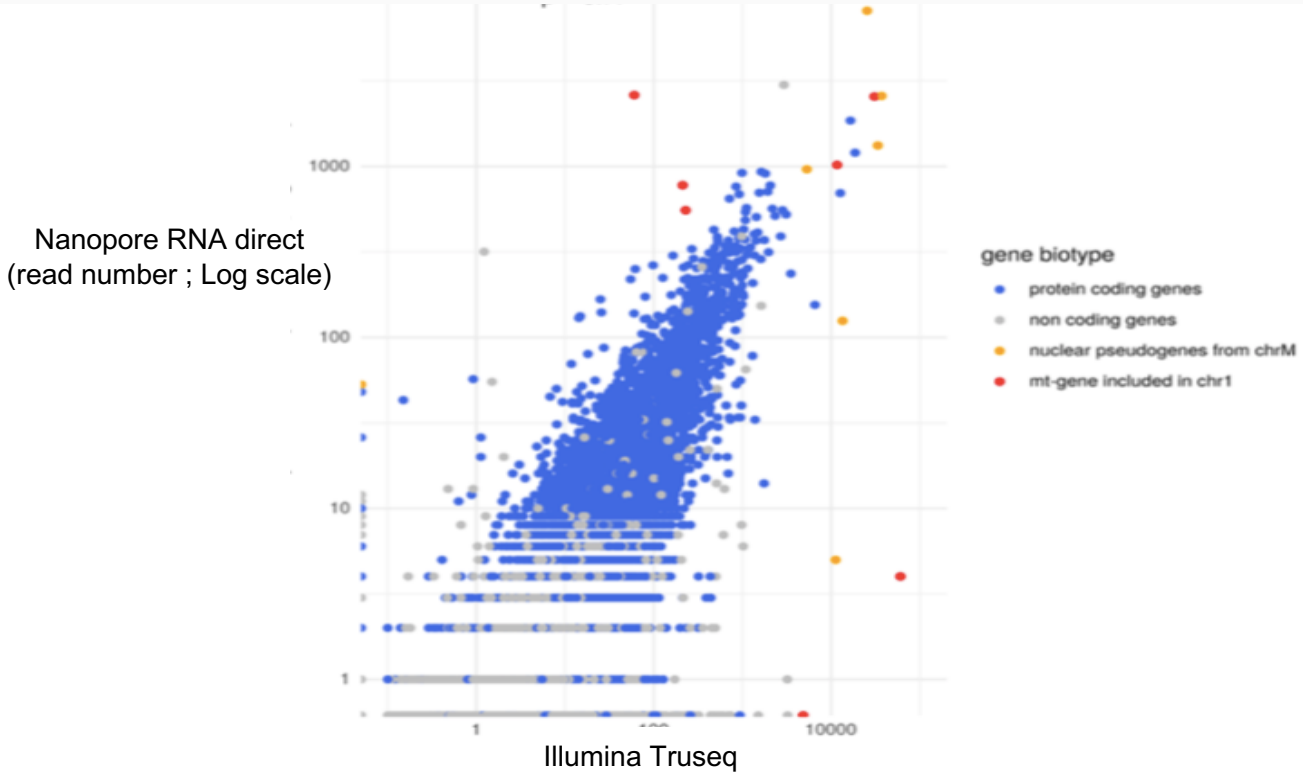


RNA directly sequenced in nanopore

- No PCR bias
- Quantitative

DIRECT RNA SEQUENCING vs ILLUMINA





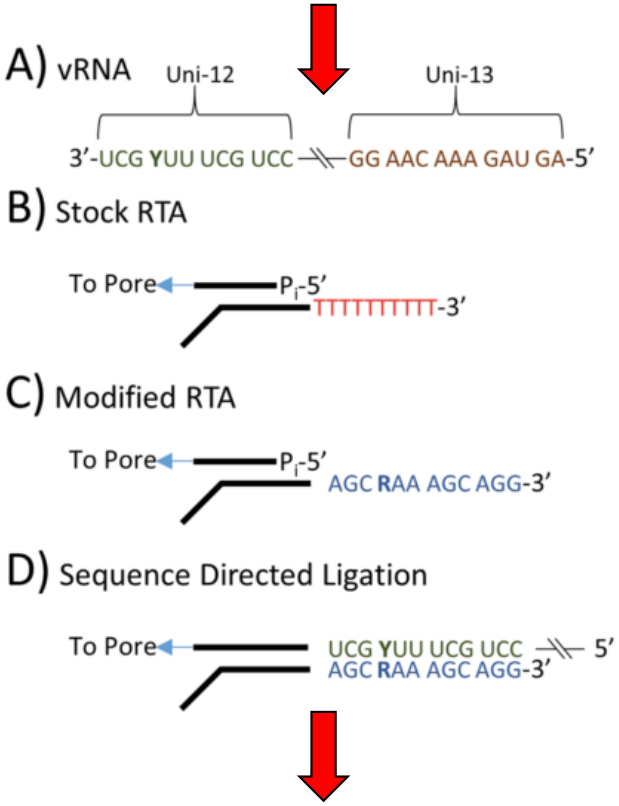
Sessegolo et al. *Sci. Reports* 2019

DIRECT RNA SEQUENCING: INFLUENZA VIRUS GENOME

Direct RNA Sequencing of the complete Influenza A Virus Genome
Keller et al. *Scientific Reports*, Sept. 2018

For the first time a complete genome of an RNA virus sequenced in its original form

Influenza A viruses are negative-sense segmented RNA viruses (8 segments)

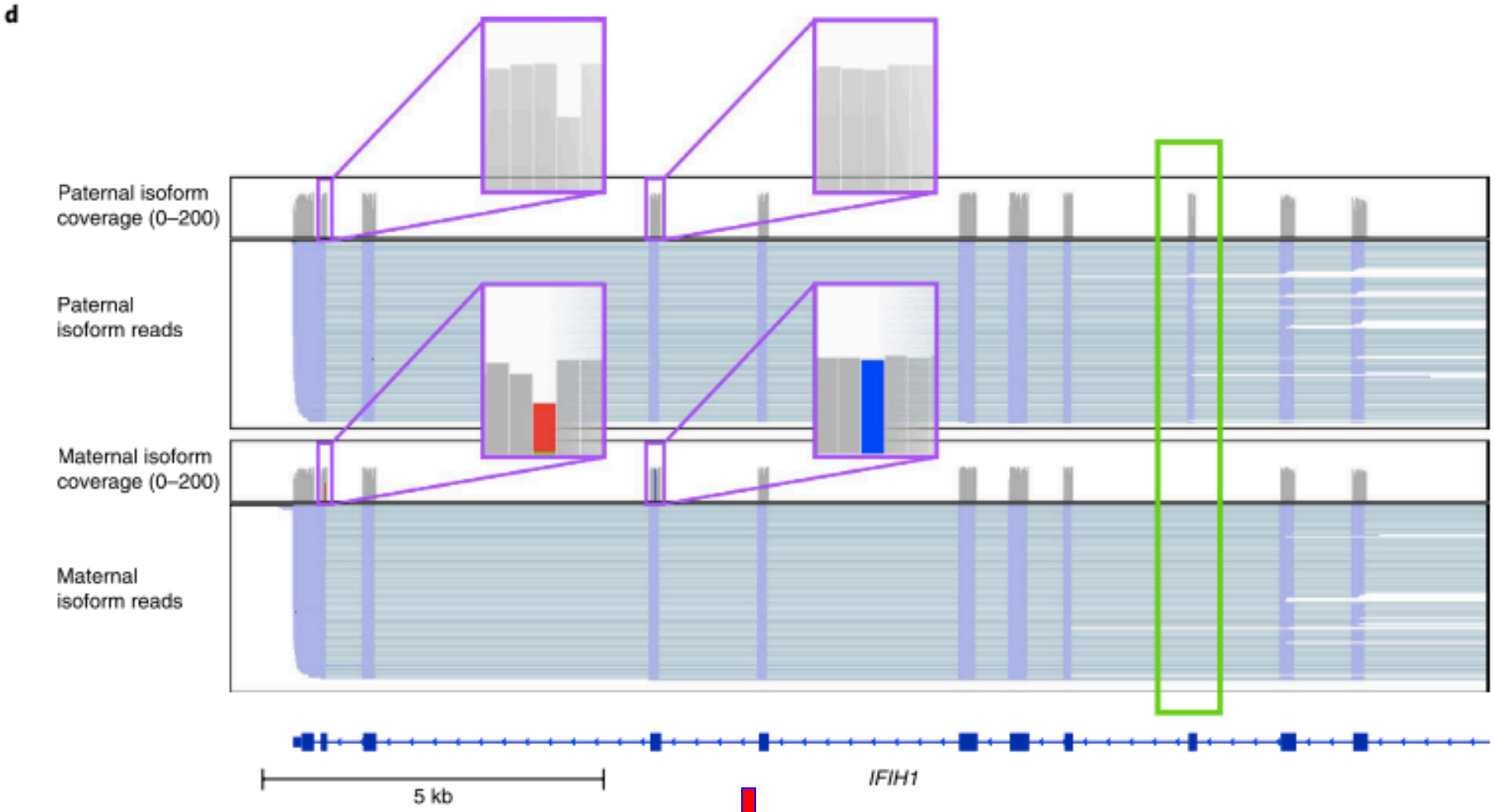


sequencing of complete genome with 100% nucleotide coverage, 99% consensus identity

Potential to identify and quantify splice variants, base modifications
not practically measurable with current methods

DIRECT RNA SEQUENCING: TRANSCRIPT HAPLOTYPE

Nanopore native RNA sequencing of a human transcriptome

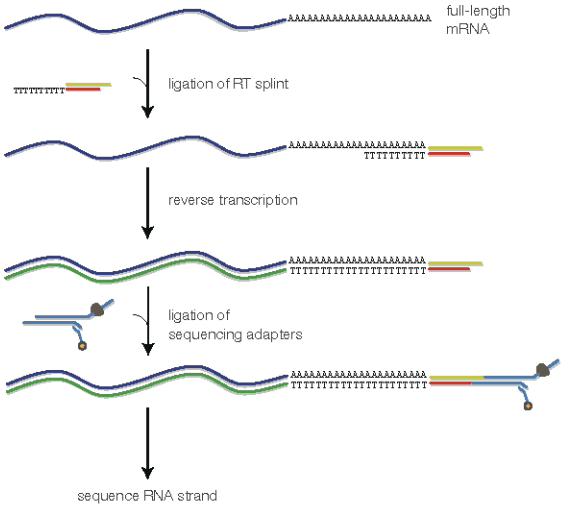


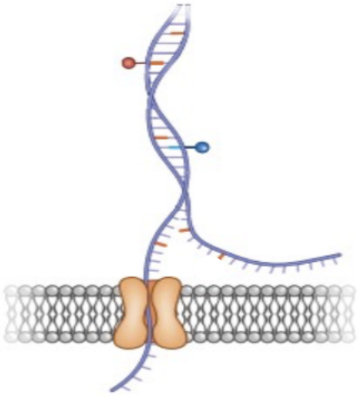
34 genes with discordant allele specificity in two isoforms

Workman et al. *Nat. Methods* (2019)

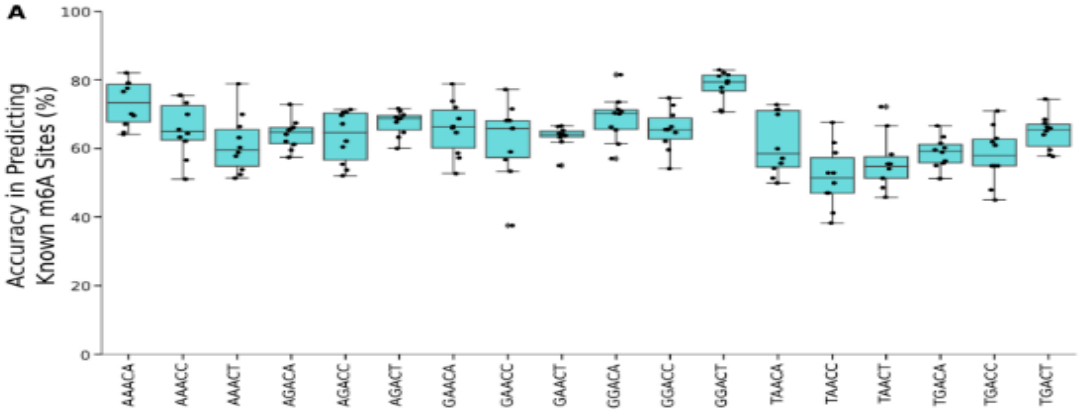
DIRECT RNA SEQUENCING: DETECTION OF m6A

Library preparation





Lorenz et al. *RNA* 2019



Detection of **m6A** with Nanopolish :

Different detection efficiency in different sites: 45% to 82%

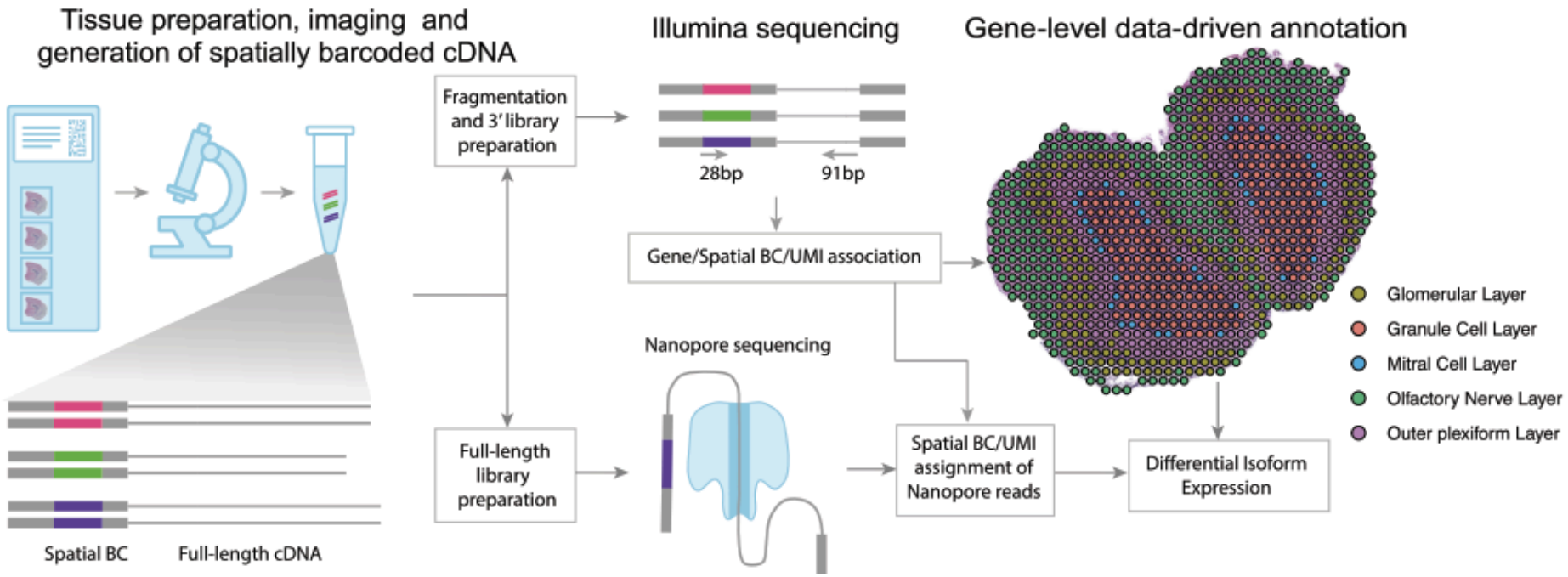
Context dependent detection efficiency

Recent advances : Nanopore and 10x Genomics Visium

The spatial landscape of gene expression isoforms in tissue sections
Lebrigand et al., *bioRxiv*, 2020

Spatial Isoform Transcriptomics (SiT) : Genome-wide approach to explore and discover in a tissue context :

- Isoform expression (bi-allelic expression)
- Sequence heterogeneity (SNP expression)



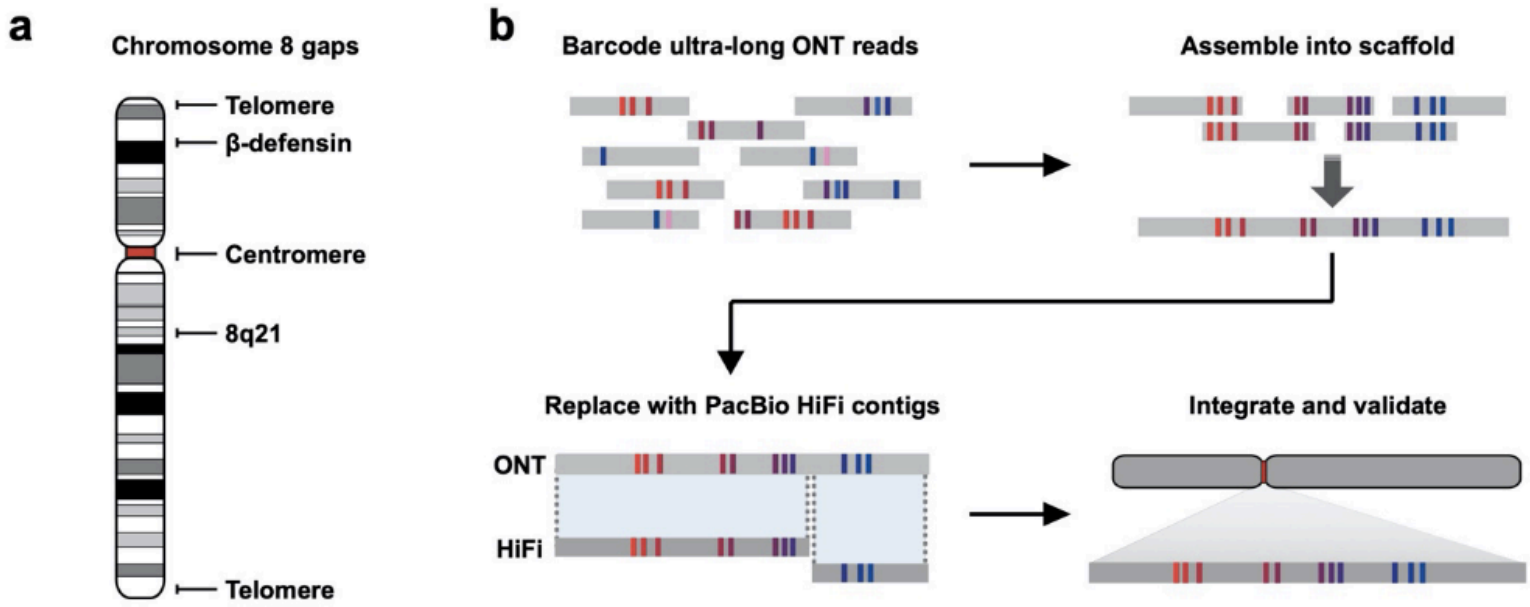
The structure, function, and evolution of a complete human chromosome 8

Logsdon et al., *bioRxiv*, Sept 2020

First complete linear assembly of a human autosomal chromosome.
It resolves the sequence of five previously long-standing gaps :

- 2.08 Mbp centromeric α-satellite array
- 644 kbp defensin copy number polymorphism
- 863 kbp variable number tandem repeat at chromosome 8q21.2 (neocentromere)
- Etc..

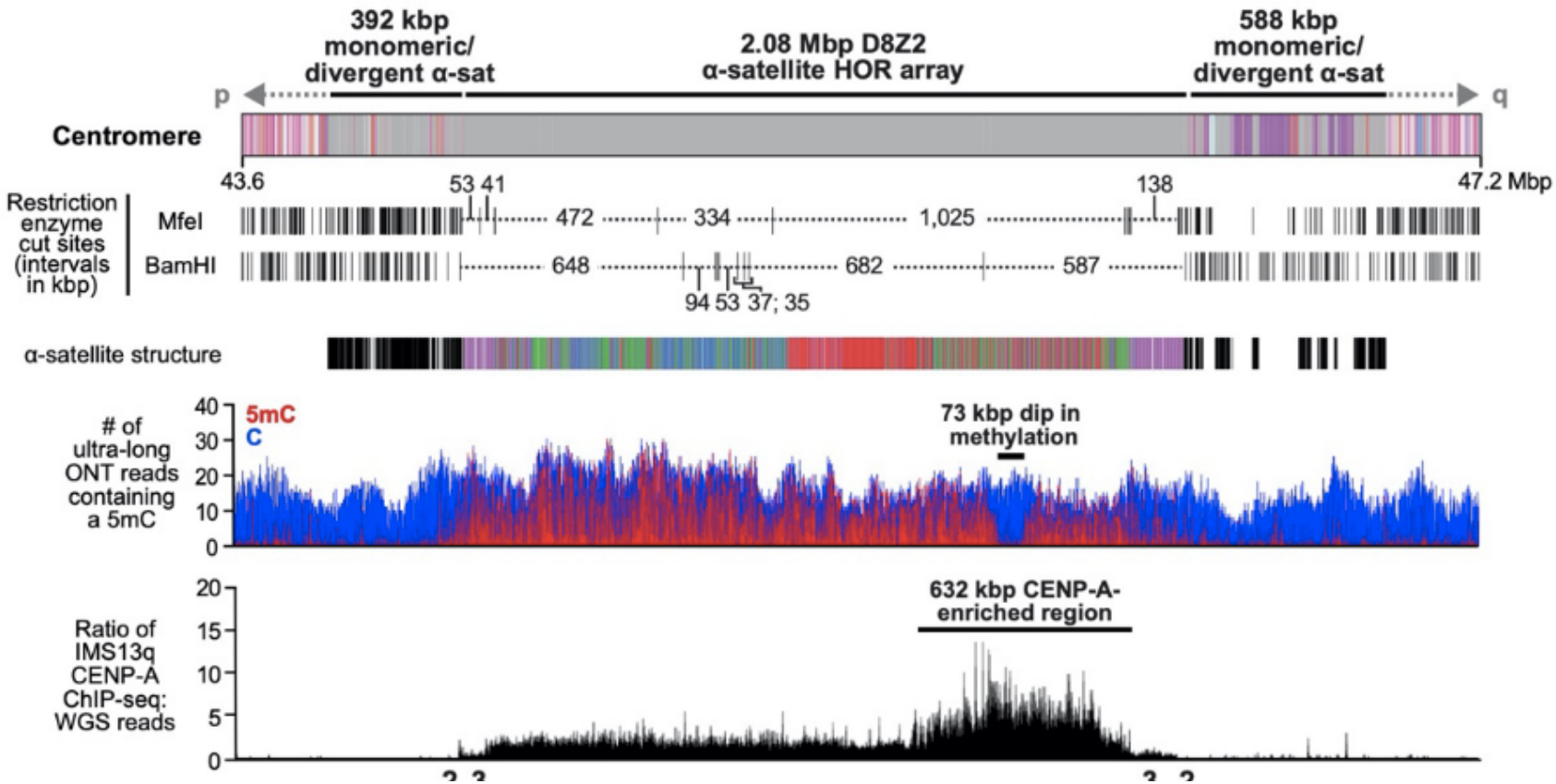
- Barcoded Ultra-long Nanopore reads assembled into a scaffold
- Regions within the scaffold with high sequence identity with PacBio HiFi contigs are replaced, thereby improving the base accuracy to >99.99%.

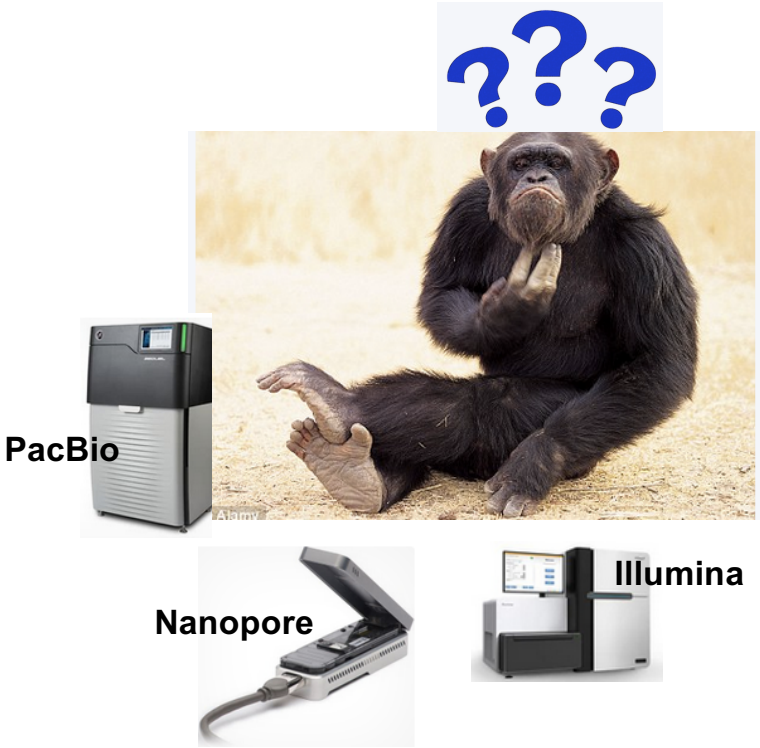


The structure, function, and evolution of a complete human chromosome 8

Logsdon et al., *bioRxiv*, Sept 2020

Epigenetic map of the chromosome 8 centromeric region




???
PacBio
Nanopore
Illumina