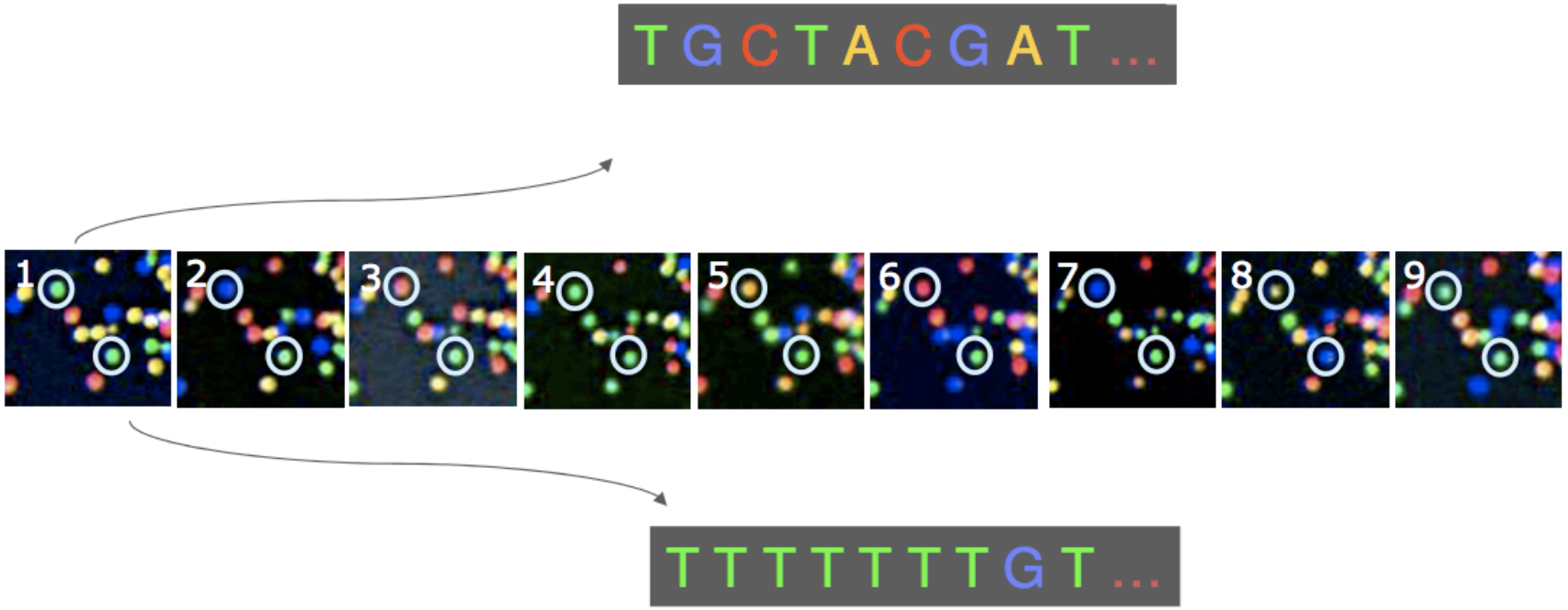
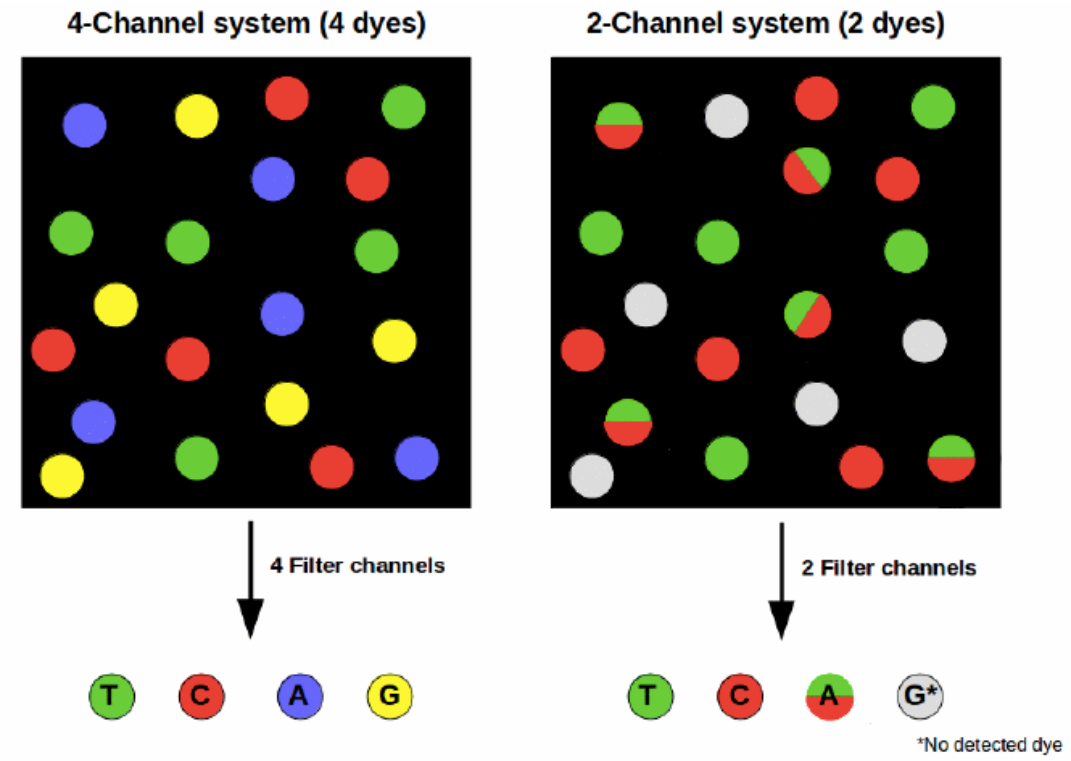
Base calling from raw data



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

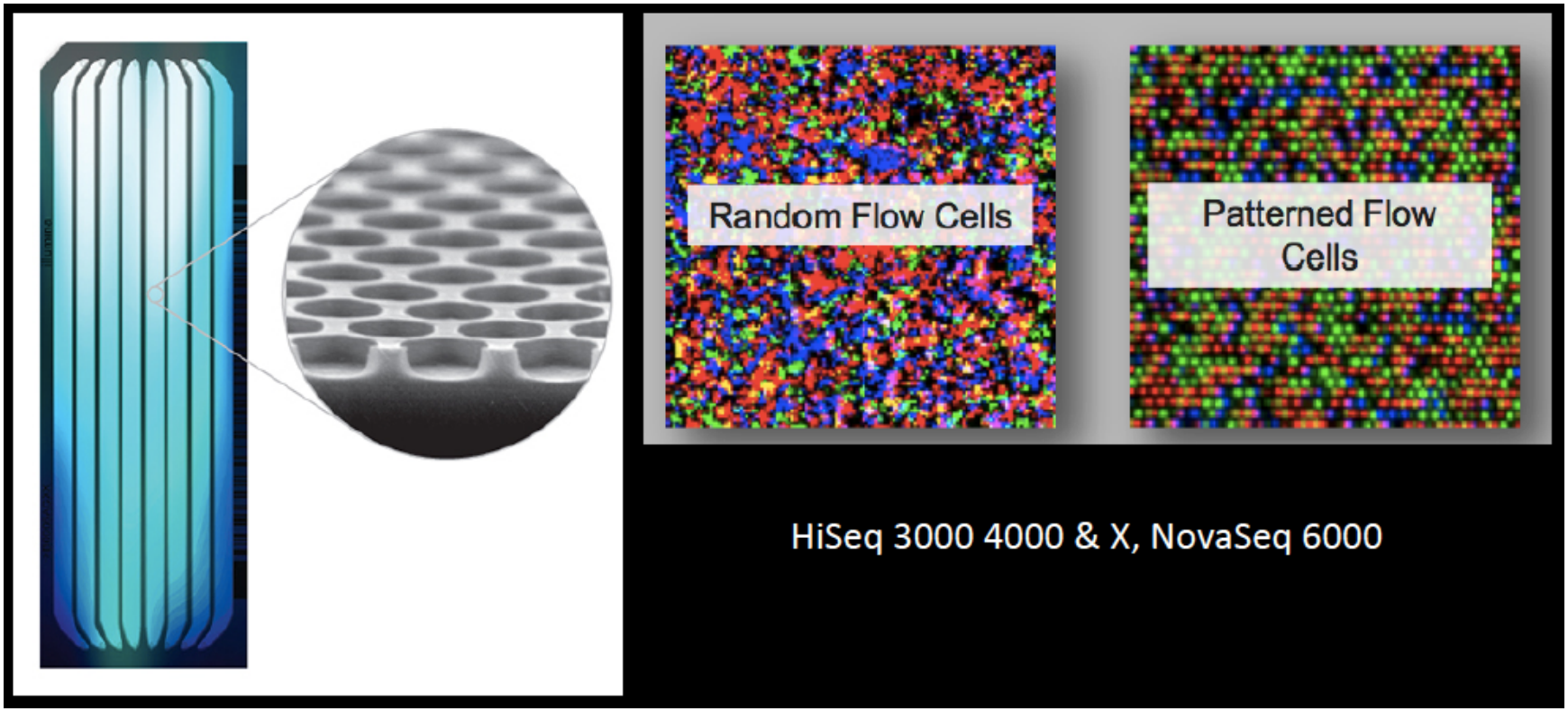
4 channel vs 2 channel detection

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



Patterned flow cells

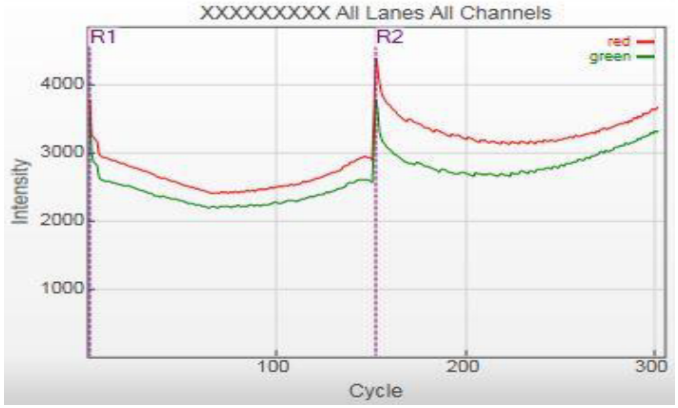
- Improves regularity of densities and qualities
- Reduces analysis time



Sequencing qualities

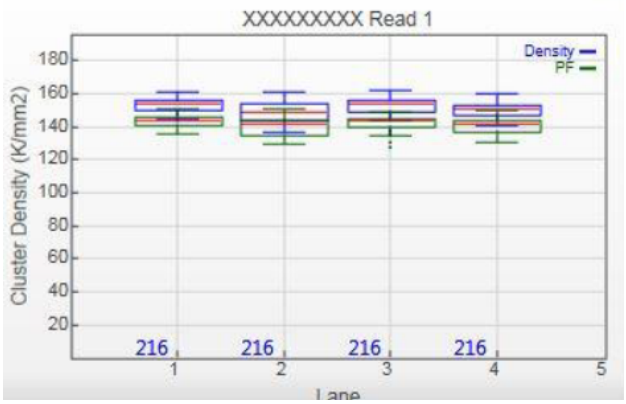
Intensities

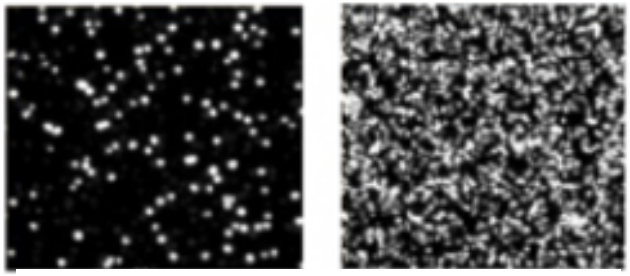
Intensity values of the four incorporated bases are measured to determine if a clustering failure has occurred



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

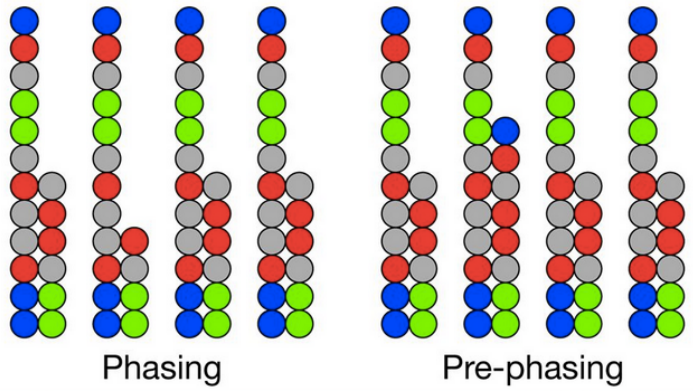
Percentage of reads passing filter (% PF)

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%.

Sequencing qualities

Phasing/Prephasing

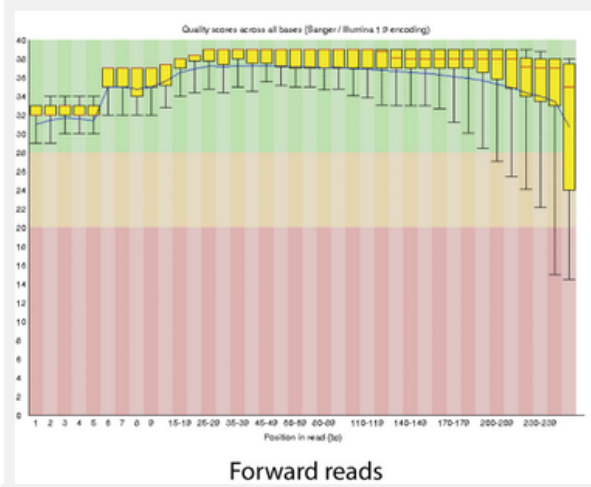
Percentages of bases that fell behind or jumped ahead the current cycle within a read. It increases with cycle number, hampering correct base identification for long reads. Optimal values are below 0.5 or 0.2% depending on platform. Issues may arise when read length is above specifications.



Q-score

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

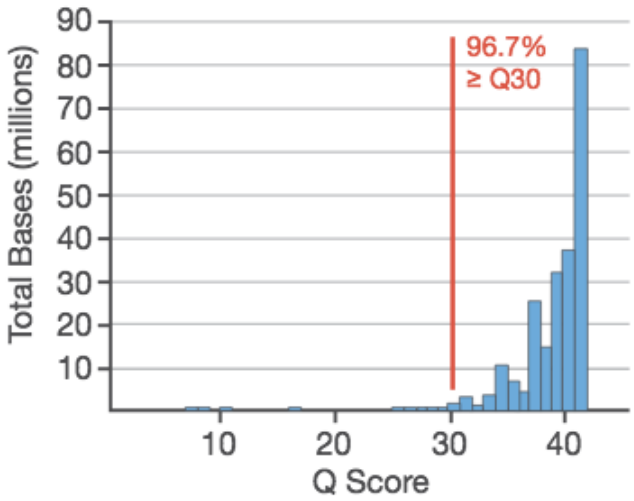
% Q-score > 30 : percentage of bases that have a a probability of incorrect base calling of 1/1000. Low Q scores can increase false-positive variant calls, which can result in inaccurate conclusions.



Sequencing qualities

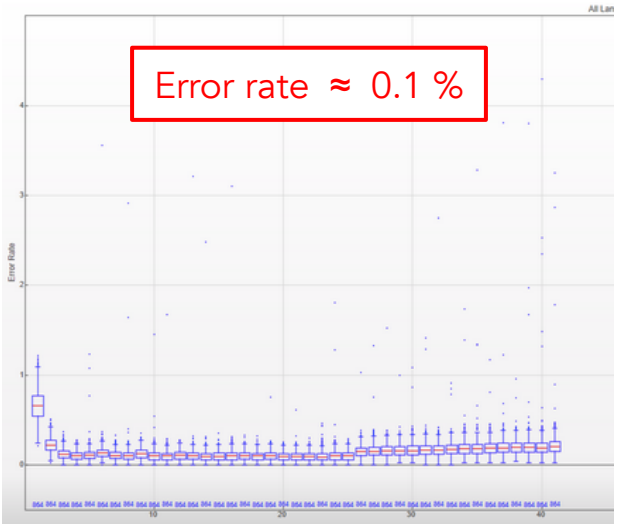
Percentage of alignment

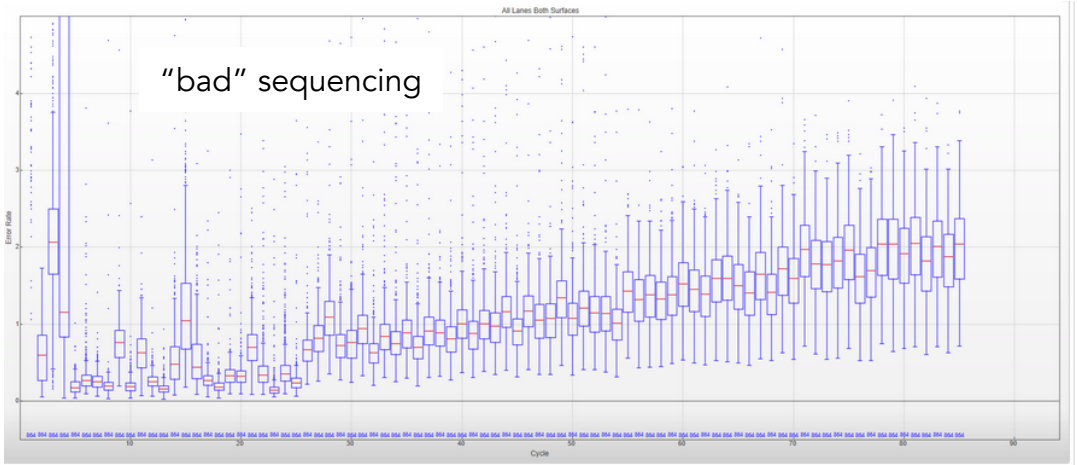
% align: percentage of the reads aligned to PhiX control genome. It is important to estimate the error rate. Indication on the success of the clustering process.



Error rate

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






Many applications
PARS-SEQ
SHAPE-SEQ
RRBS-SEQ
RIBO-SEQ
CLASH-SEQ
TIF-SEQ/PEAT
ICE
OS-SEQ
IN-SEQ
TAB-SEQ
CHIP-SEQ
5-C
SMMIP
OXBS-SEQ
RIP-SEQ
4-C
TC-SEQ
NET-SEQ
UMI
CAP-SEQ
DNASE-SEQ
FAIRE-SEQ
DUPLEX-SEQ
PAR-CLIP-SEQ
BS-SEQ
MEDIP-SEQ
GRO-SEQ
MERIP-SEQ
DIGITAL
CHIA-PET
ATAC-SEQ
CIP-TAP
ICLIP
HITS-CLIP
MBDCAP-SEQ
PARE-SEQ/GMUT
HI-C/3-C
TRAP-SEQ
RC-SEQ
FRAG-SEQ

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 :

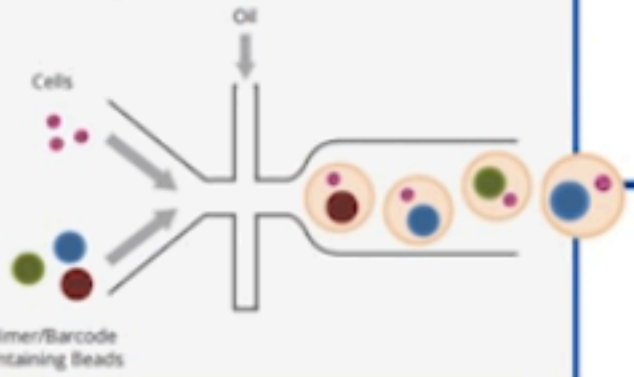
Single cell Gene expression
Measures gene activity on a cell-by-cell basis, characterize cell populations, cell types, …

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

Single cell ATAC
Measures epigenetics by detecting open chromatin regions

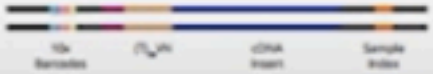
Project Workflow

1. Cells + barcoded beads isolated in oil droplet with 10X Genomics Chromium



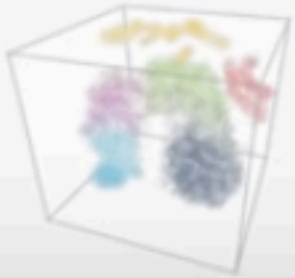
2. Reverse transcription incorporates cell and transcript-specific barcodes

3. Library generation

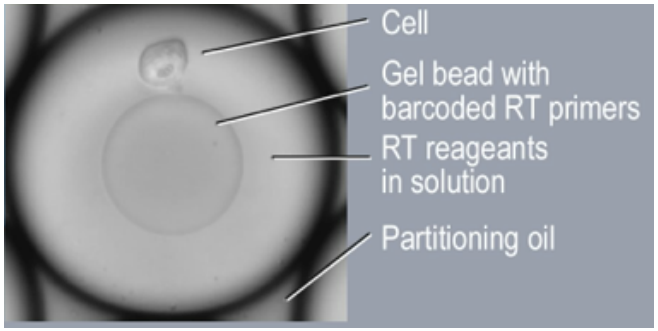


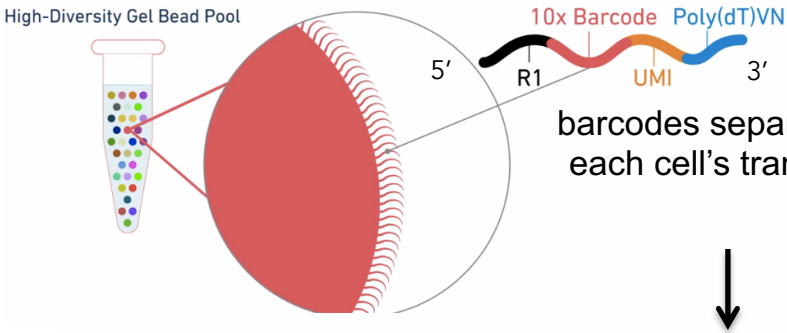
4. High-throughput sequencing with Illumina HiSeq

5. Interactive, easy to interpret report

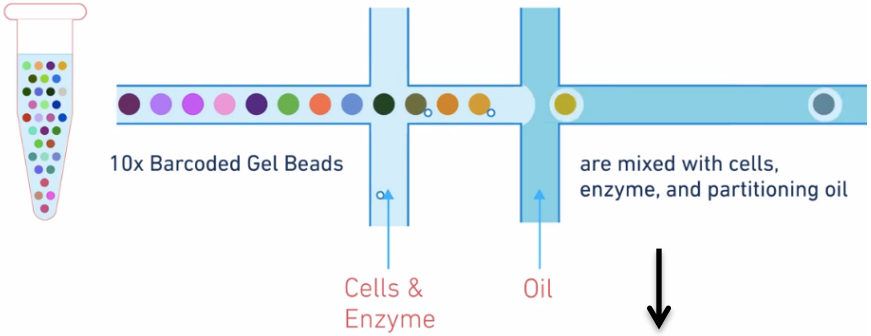


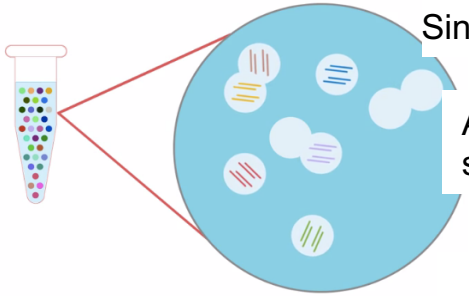
CHROMIUM SINGLE-CELL RNA SEQUENCING





barcodes separately index each cell's transcriptome

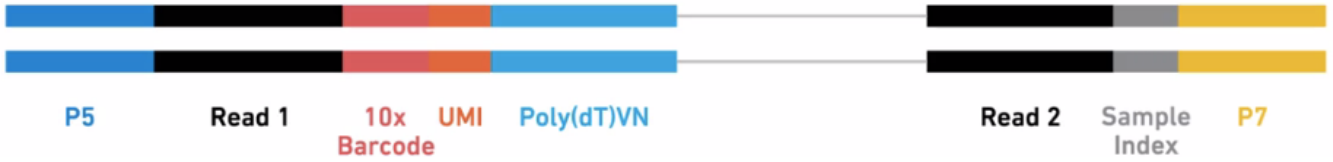




Single cells undergo RT

All cDNAs from one cell share the same barcode

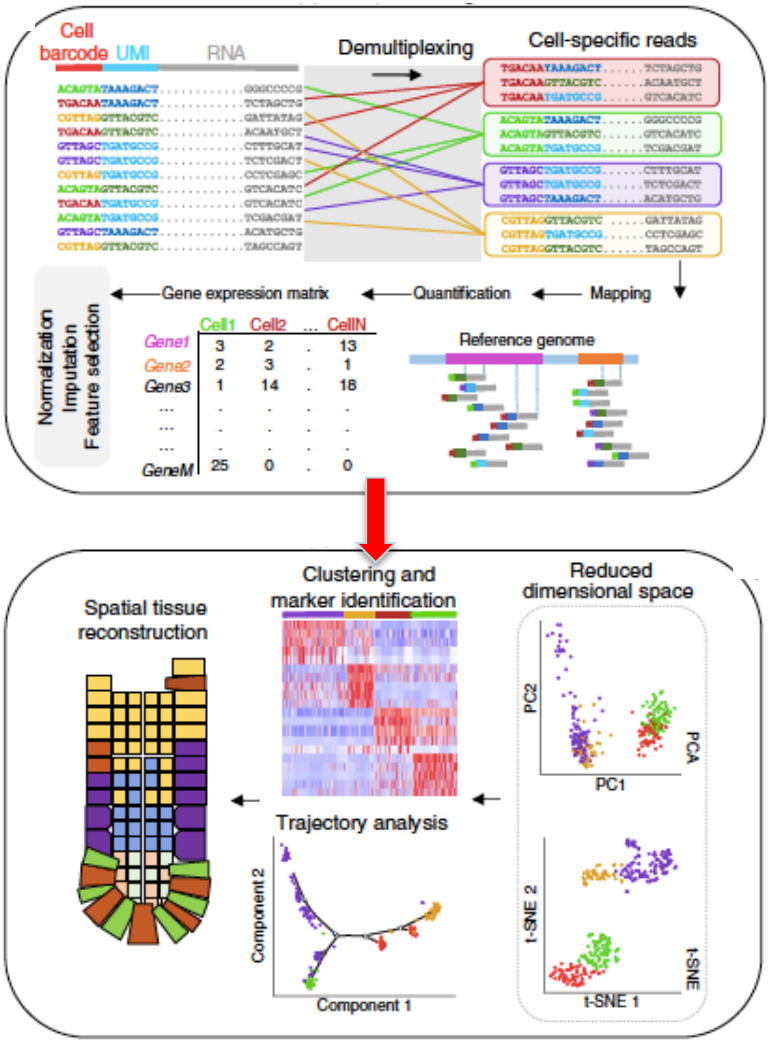
barcoded sequencing library



DATA ANALYSES

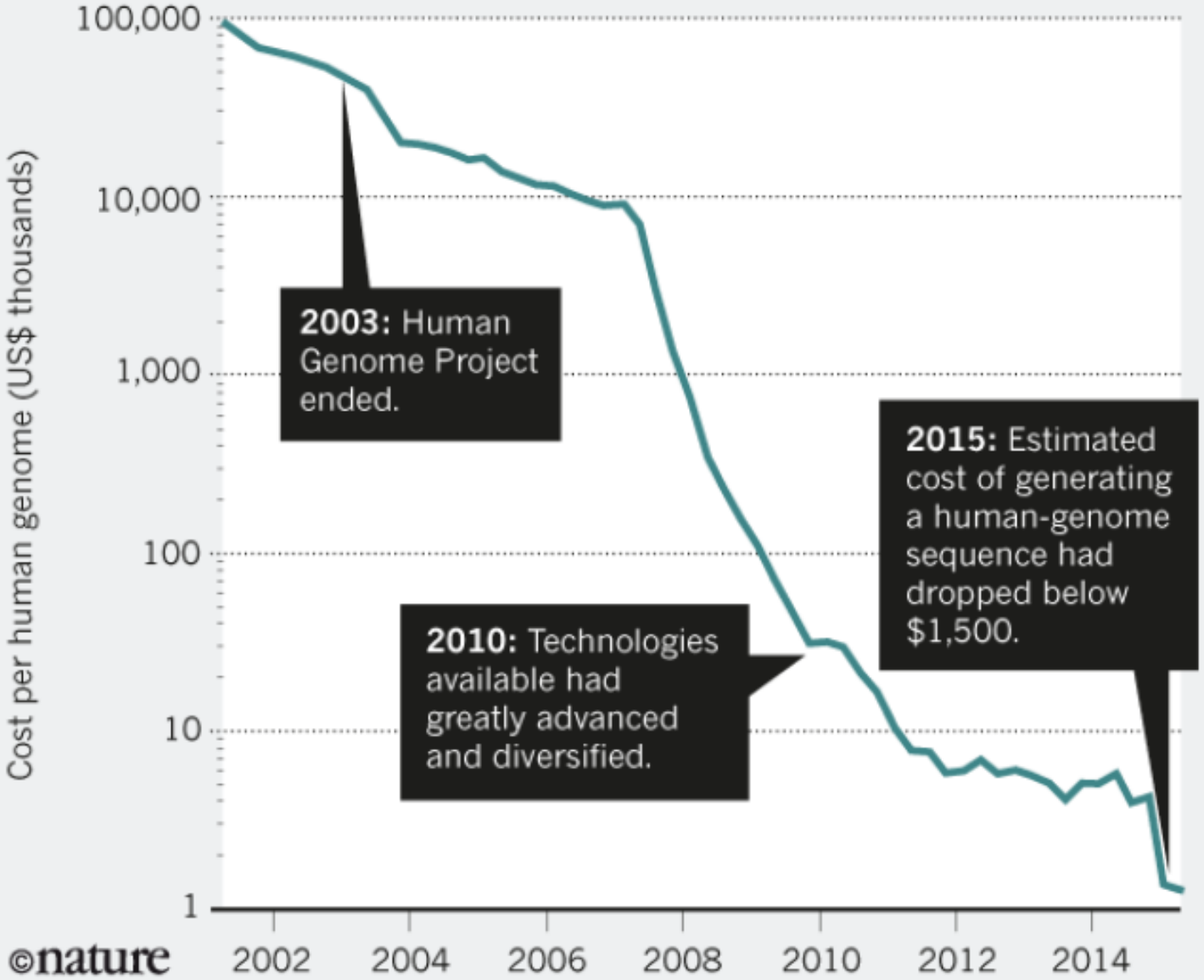
CELL RANGER :

Chromium data analysis pipelines



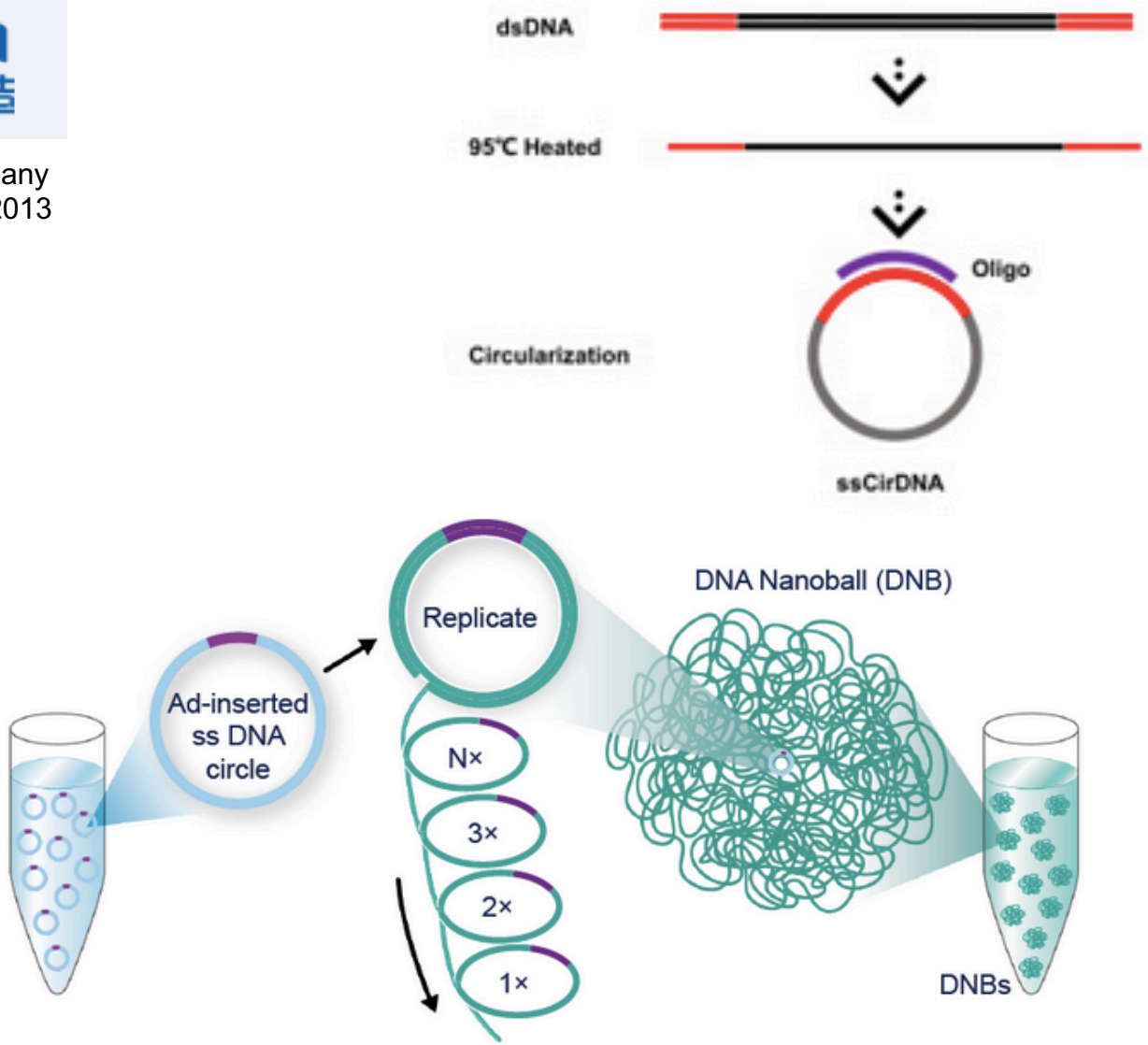
BETTER, CHEAPER, FASTER

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

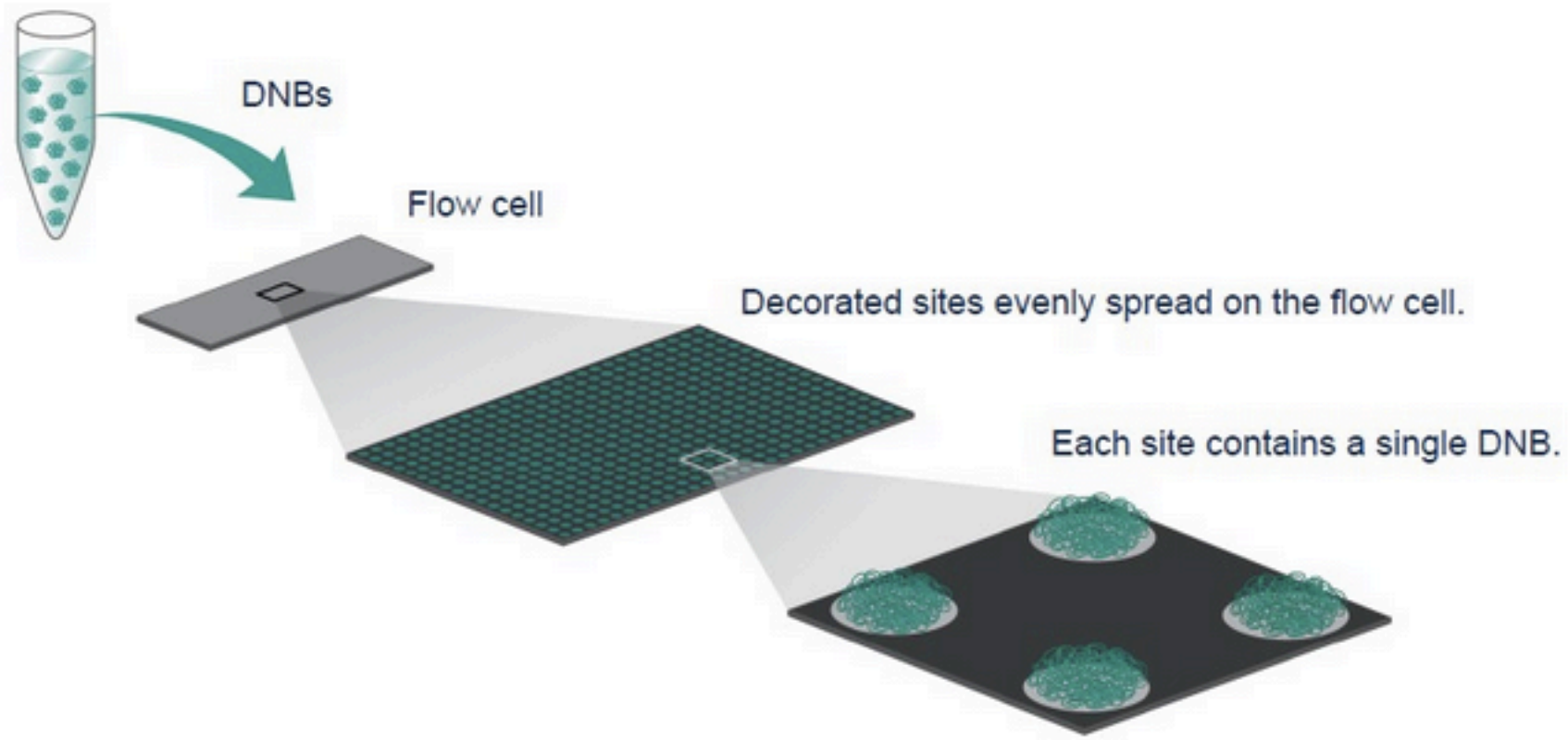


MGI 华大智造

acquisition of U.S. company
Complete Genomics in 2013



MGI 华大智造


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

PART 2

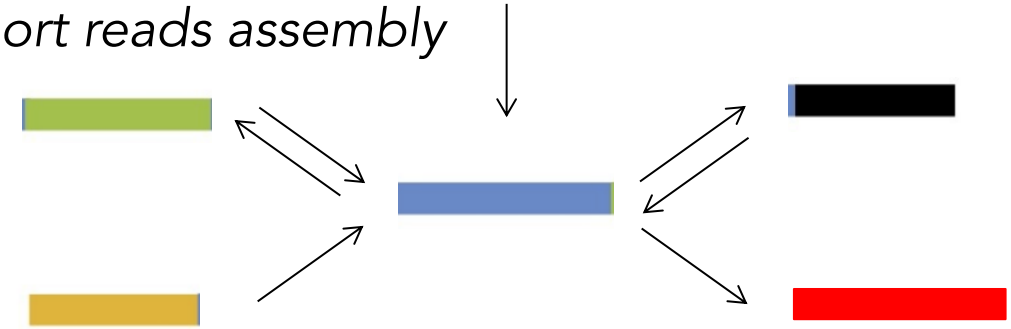
3rd GENERATION SEQUENCING

LONG READS

LONG-READS VERSUS SHORT-READS

Assembly of DNA fragments with repeated sequences

NGS short reads assembly

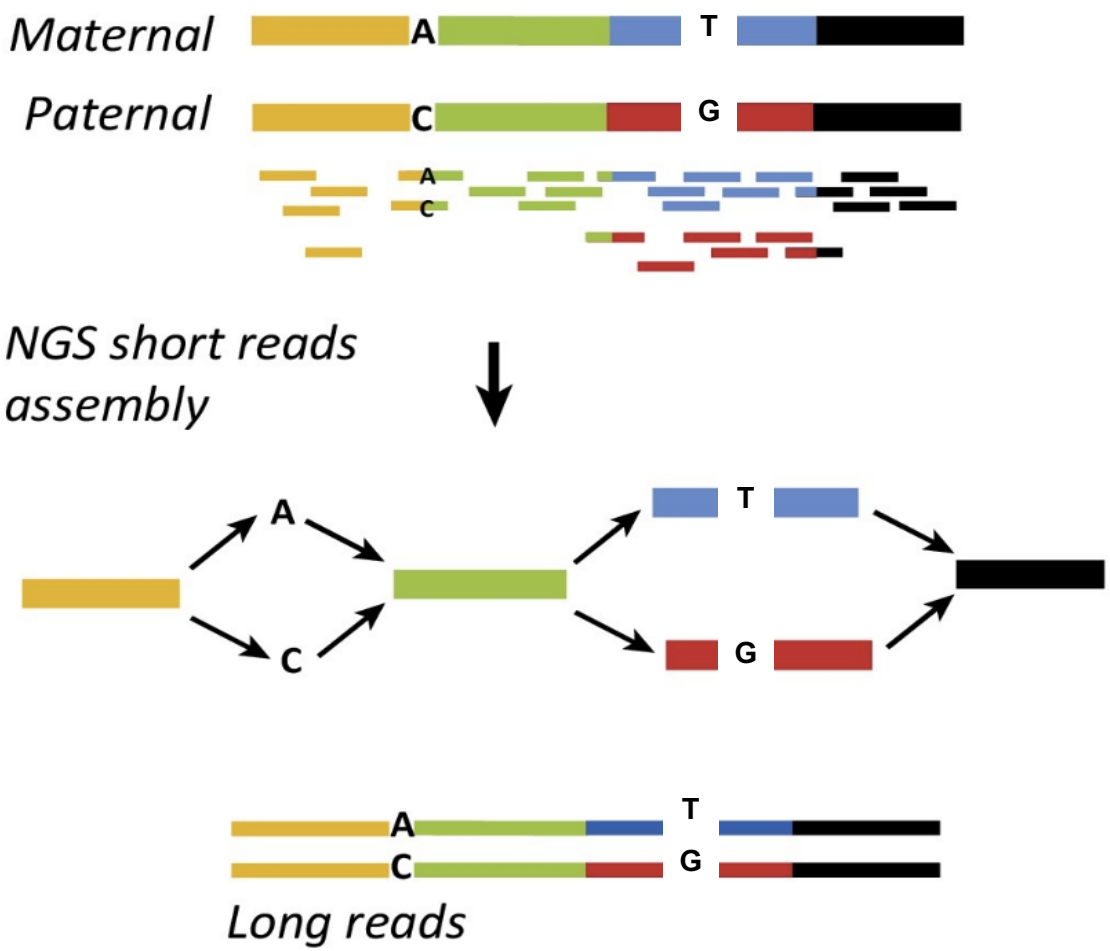


Several contigs ⟶ incomplete assembly, underestimation of repeats

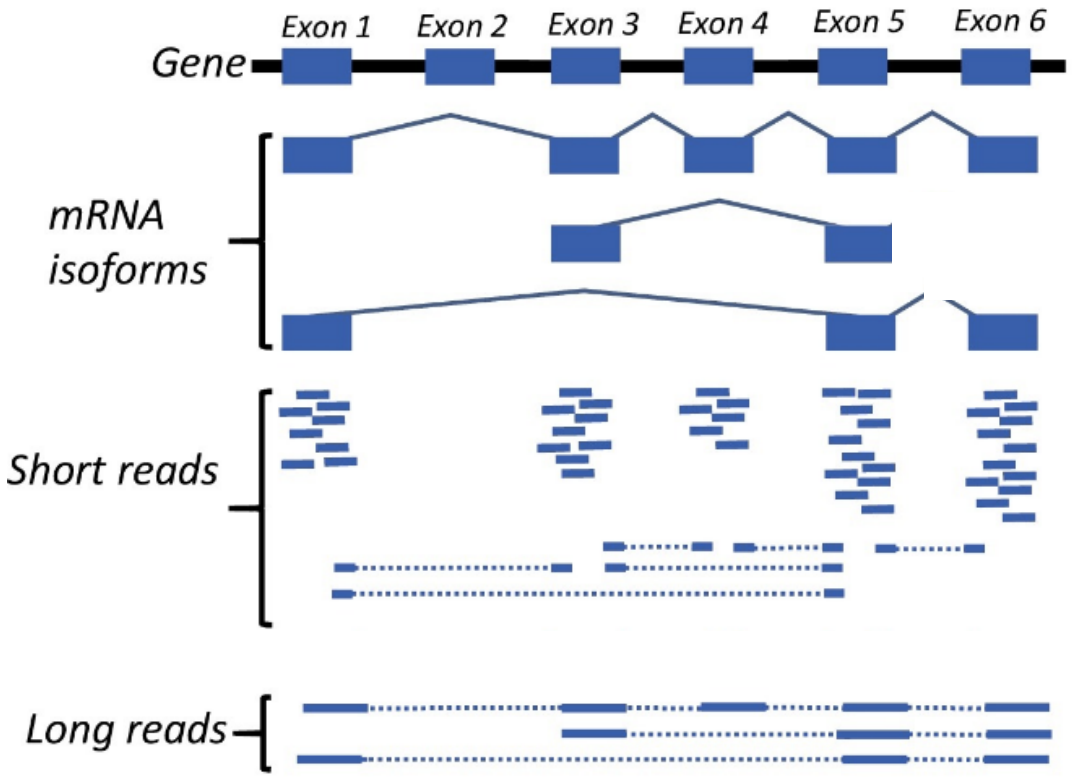
Long reads assembly

LONG-READS VERSUS SHORT-READS

Haplotype phasing



Detection of splicing isoforms



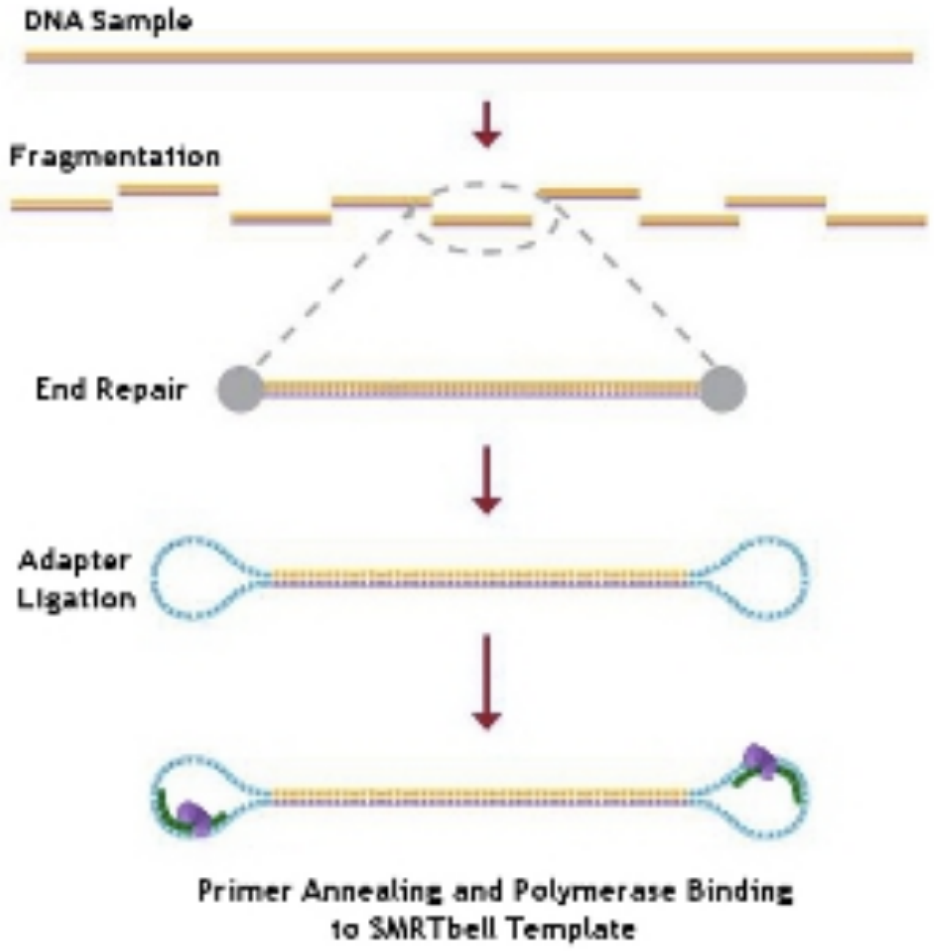
The 3rd generation winning technologies

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

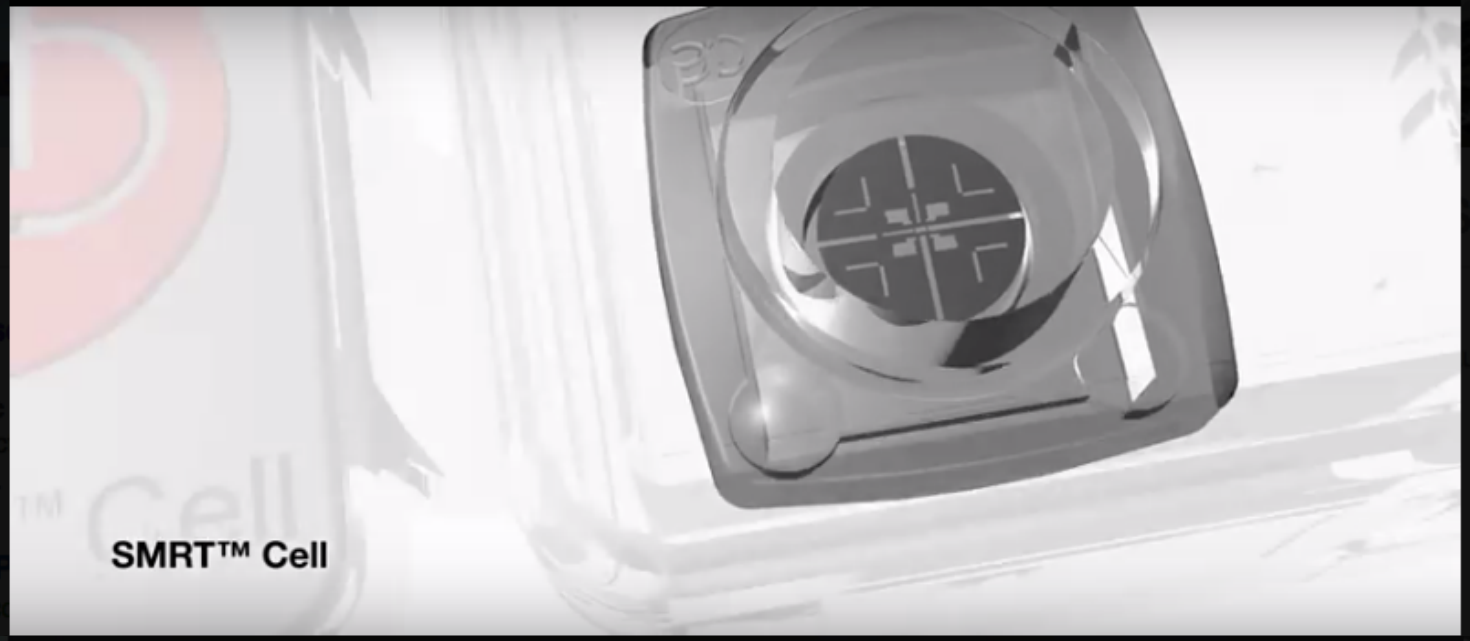
MinION - Oxford Nanopore
Single molecules
> 200 000 bp long
Error rate ≈ 10-15 %
Compensated by coverage

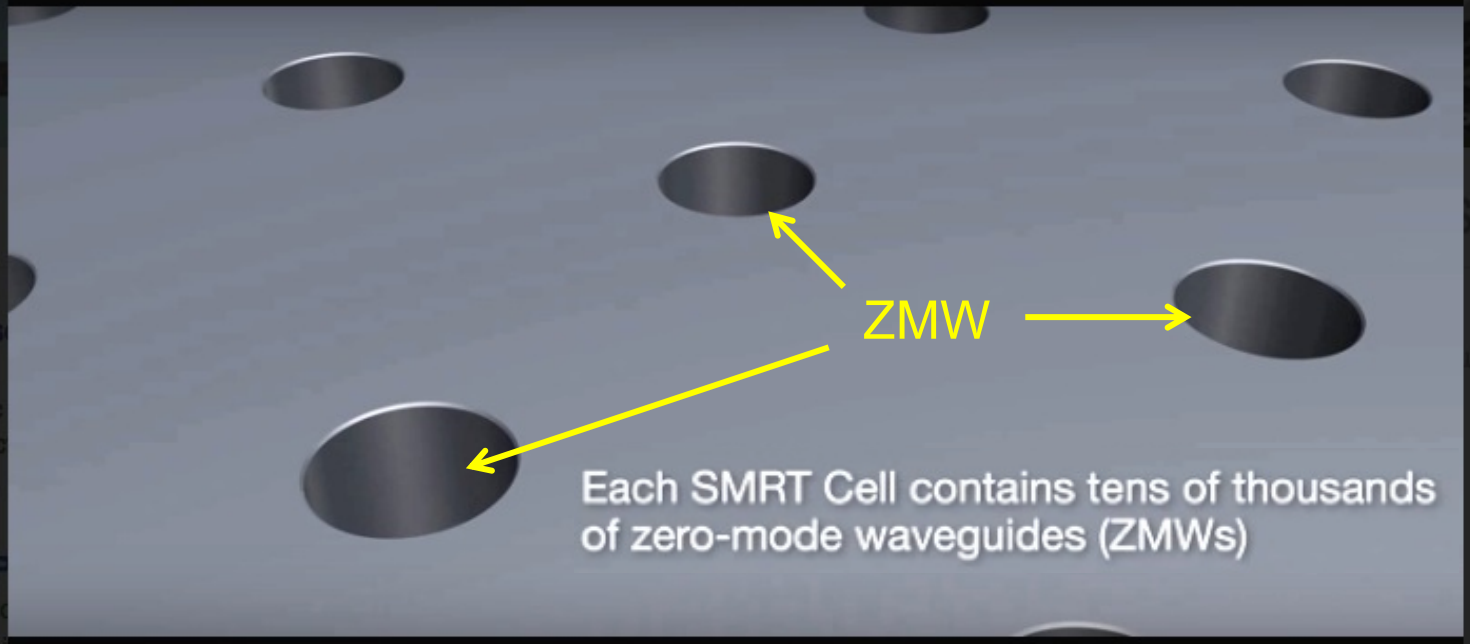
PacBio : Single Molecule Real Time (SMRT) sequencing

PacBio DNA-seq library

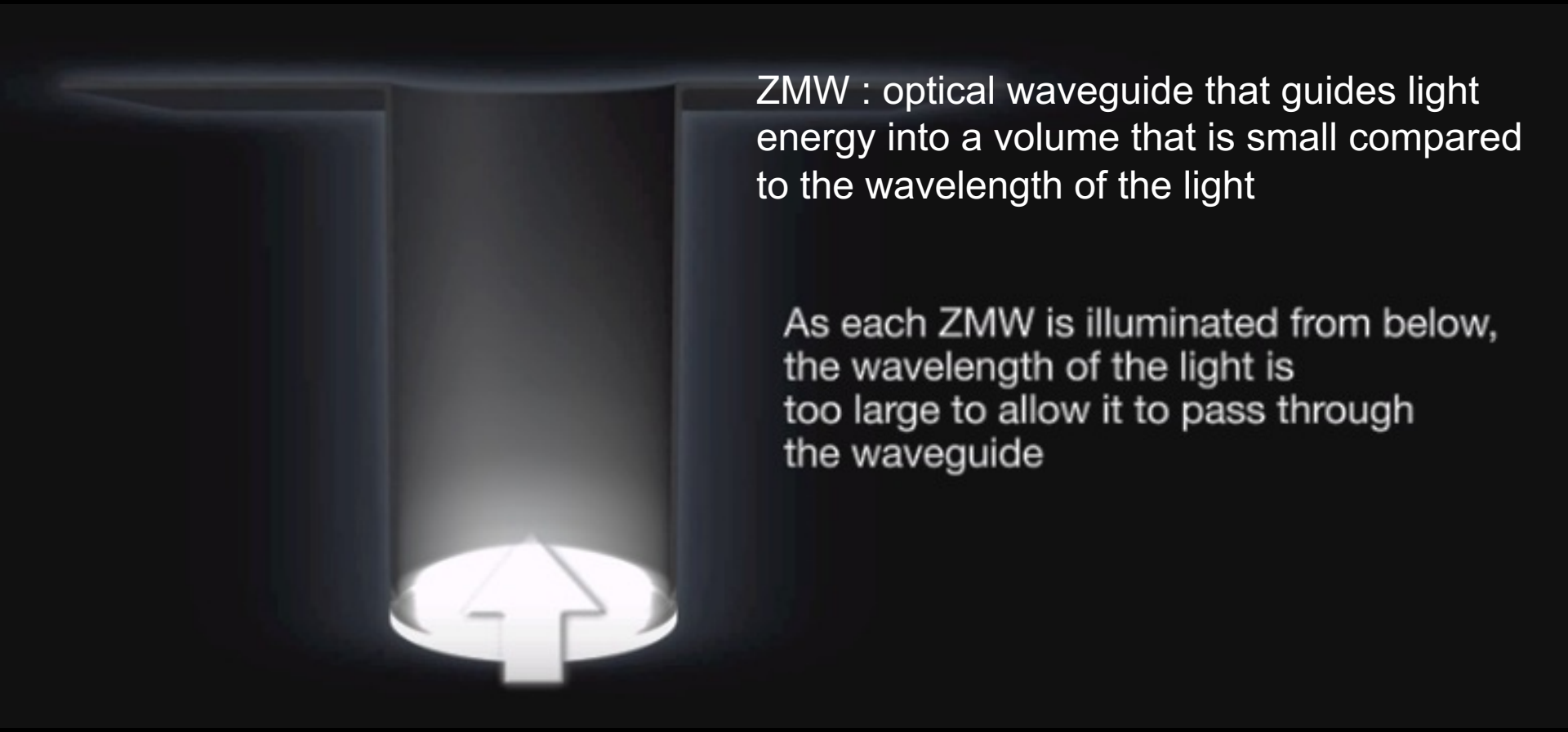


PACIFIC BIOSCIENCES

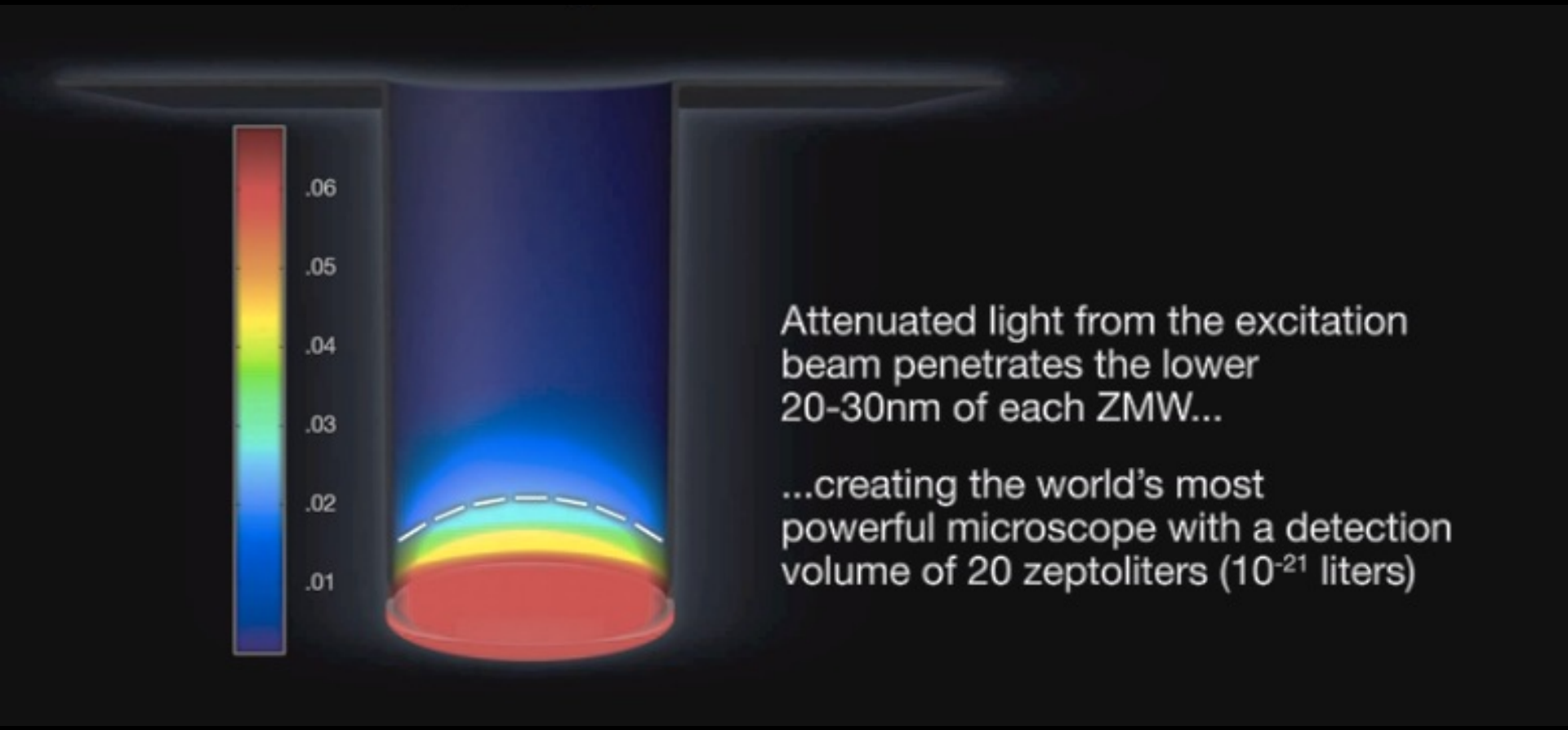




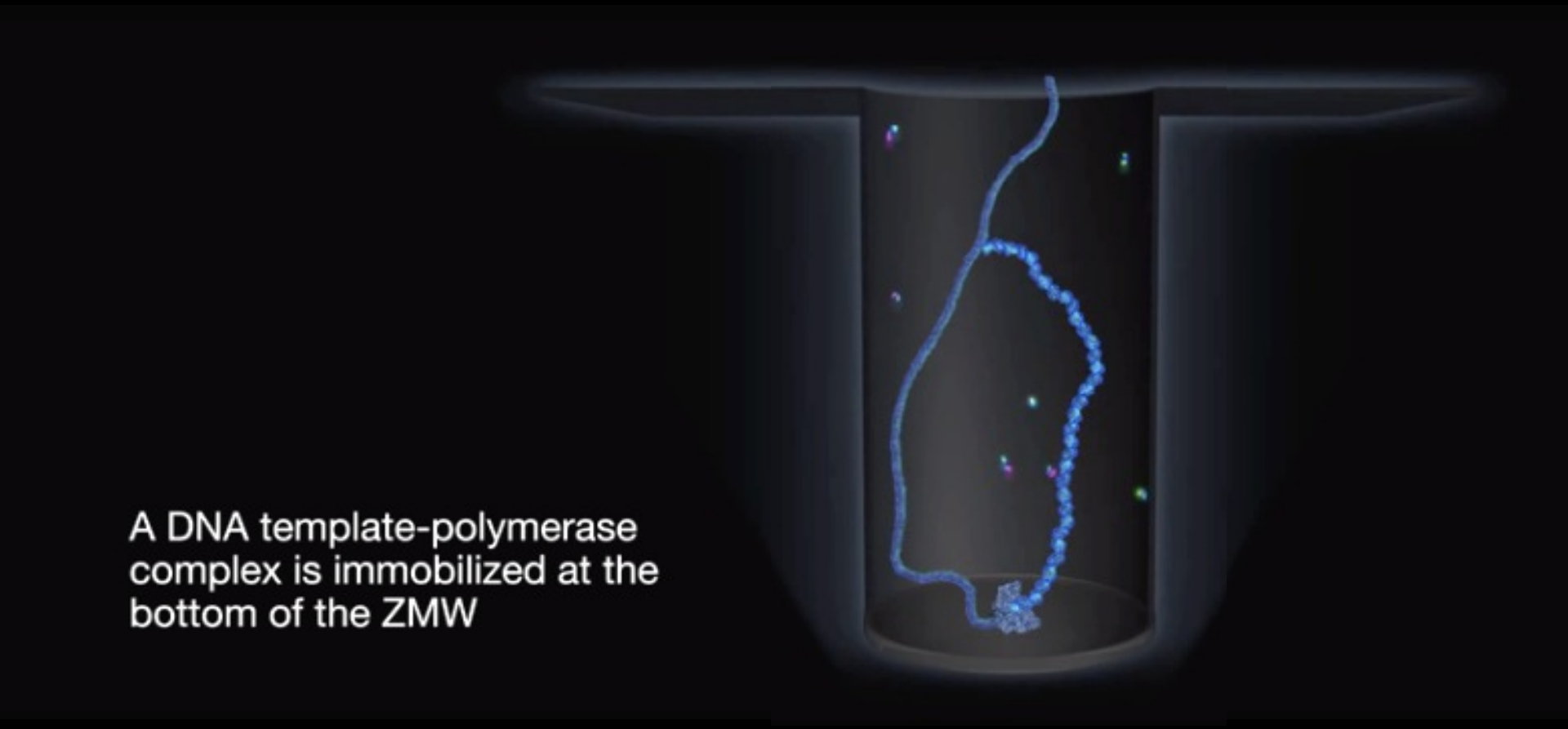
PACIFIC BIOSCIENCES



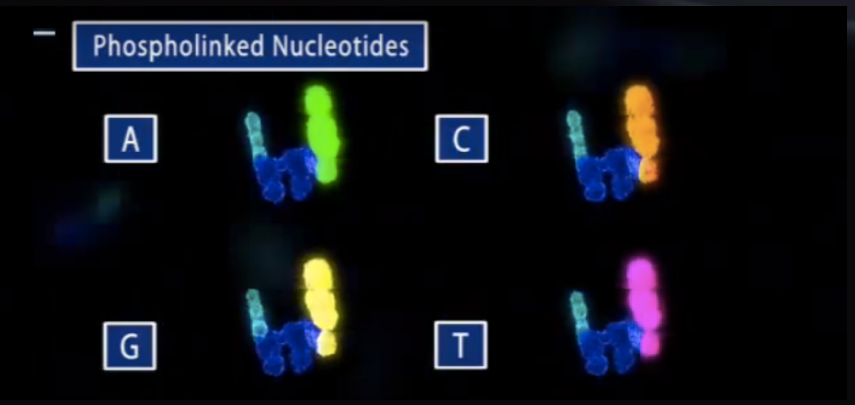
PACIFIC BIOSCIENCES



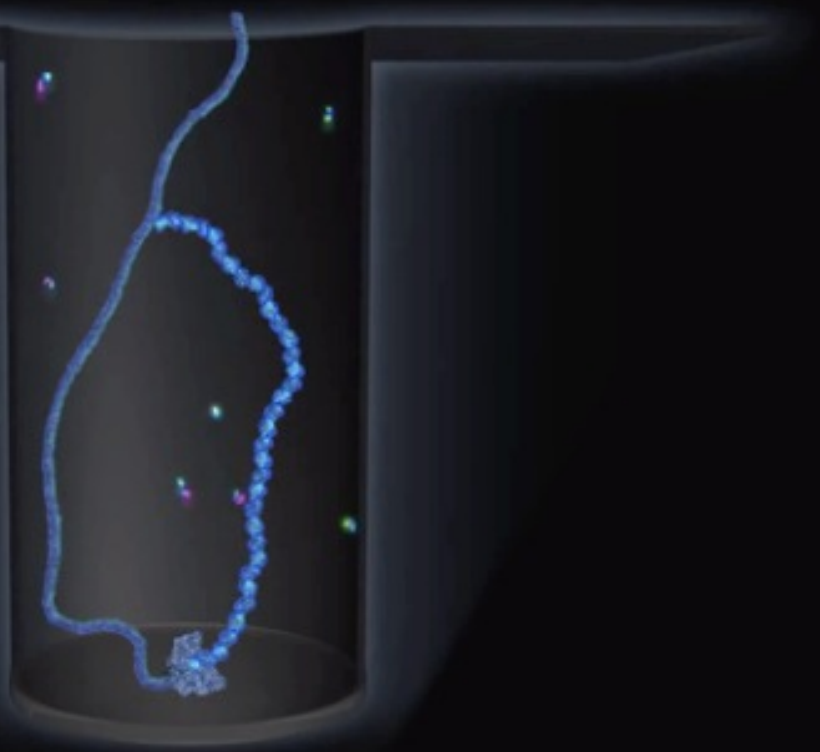
PACIFIC BIOSCIENCES



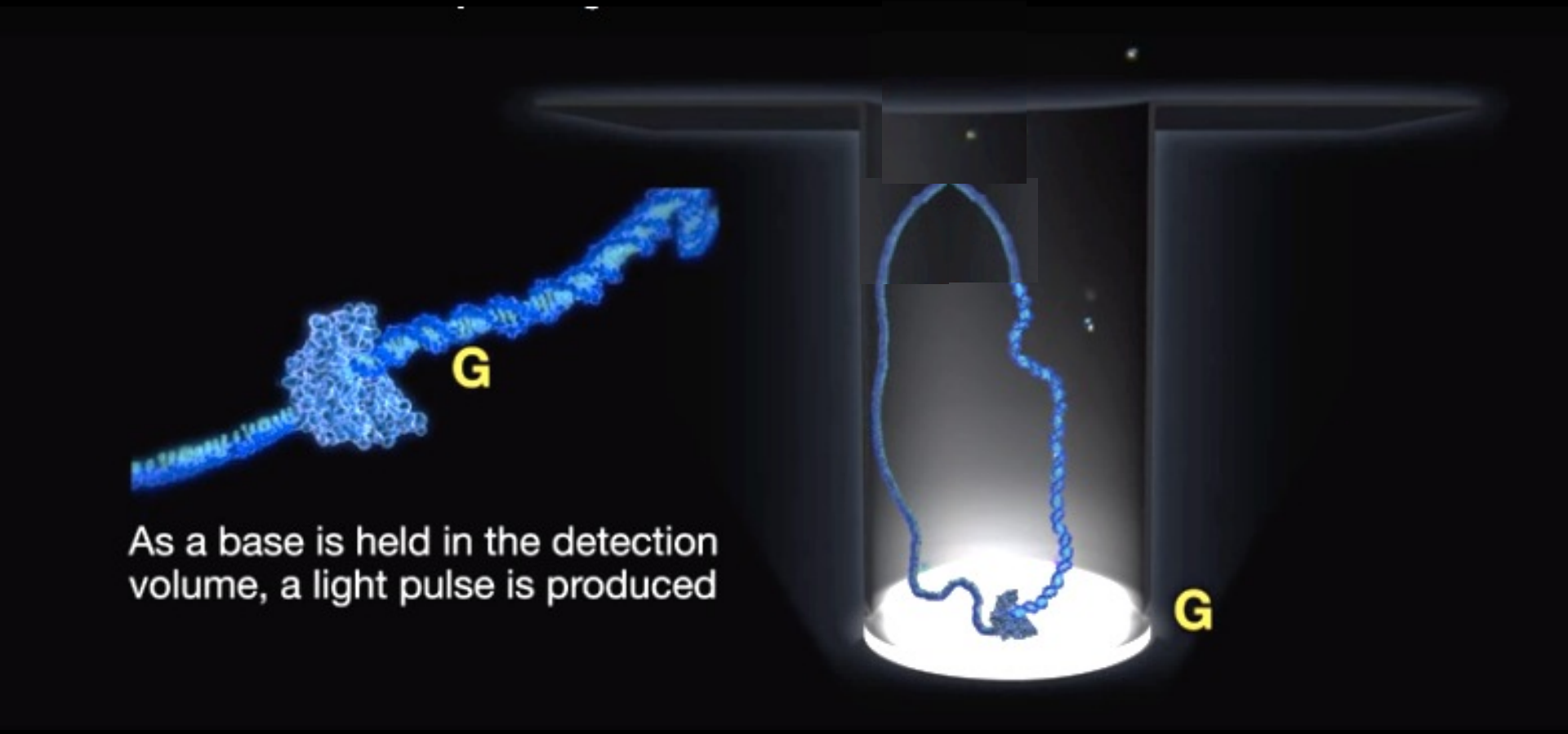
PACIFIC BIOSCIENCES



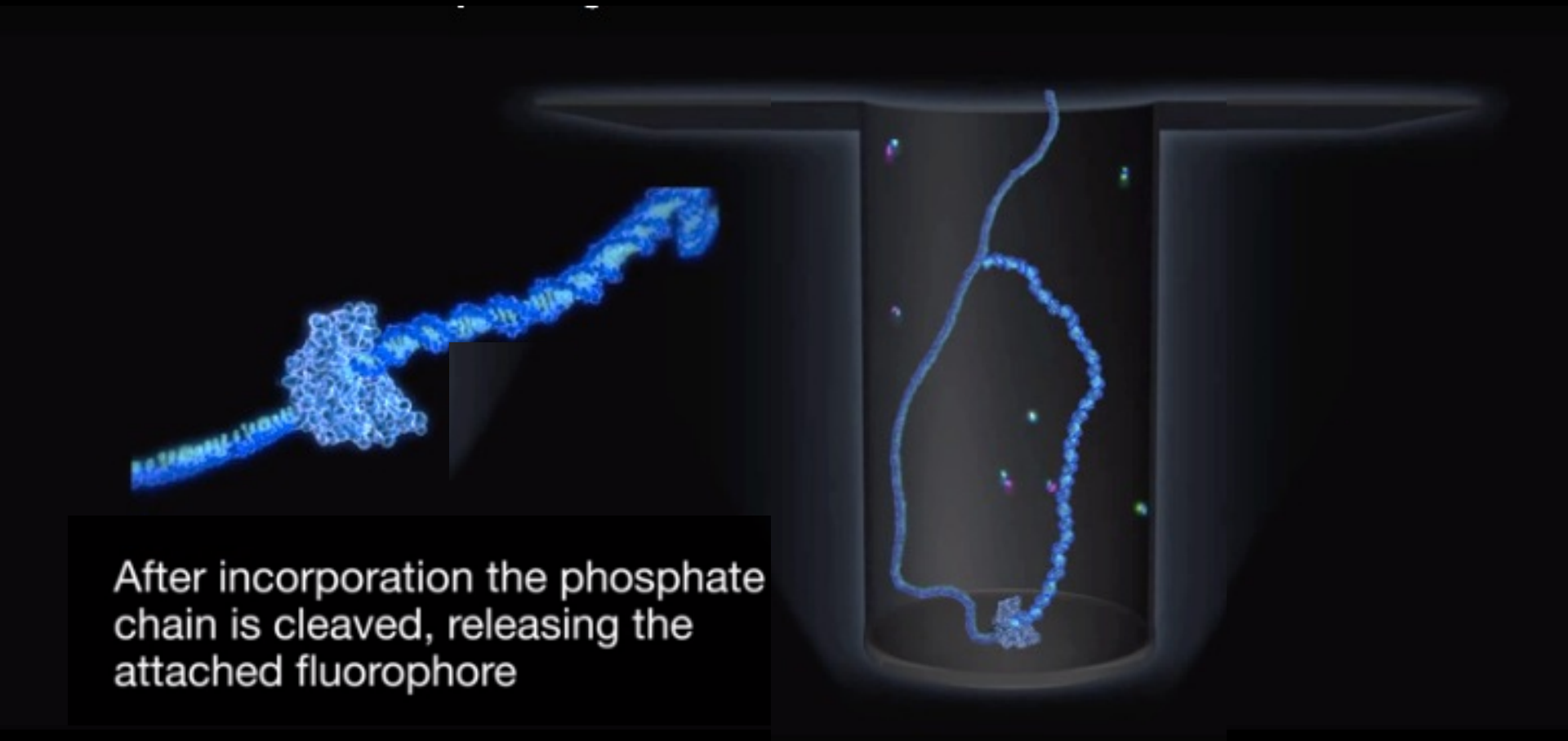
Phospholinked nucleotides are introduced into the ZMW chamber



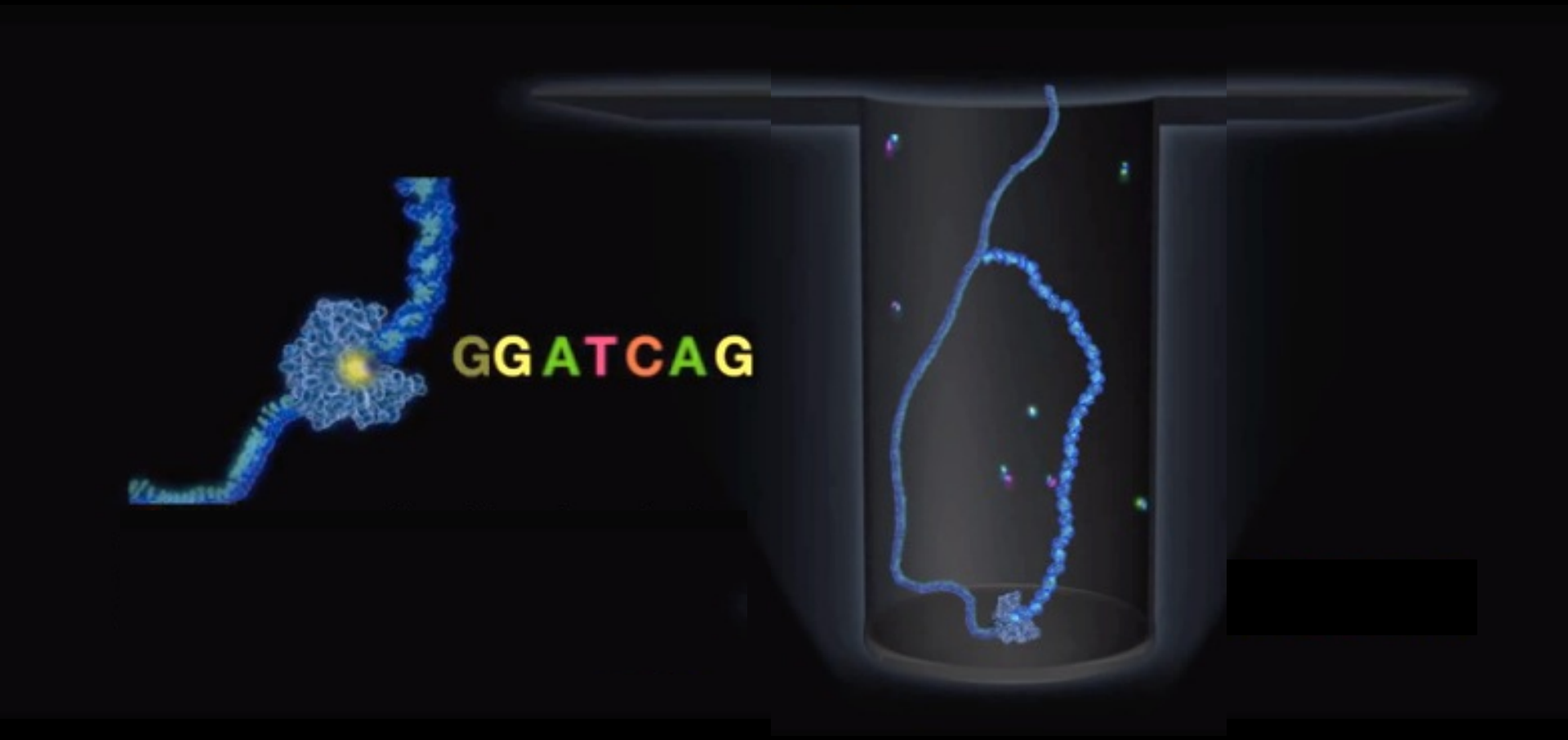
PACIFIC BIOSCIENCES

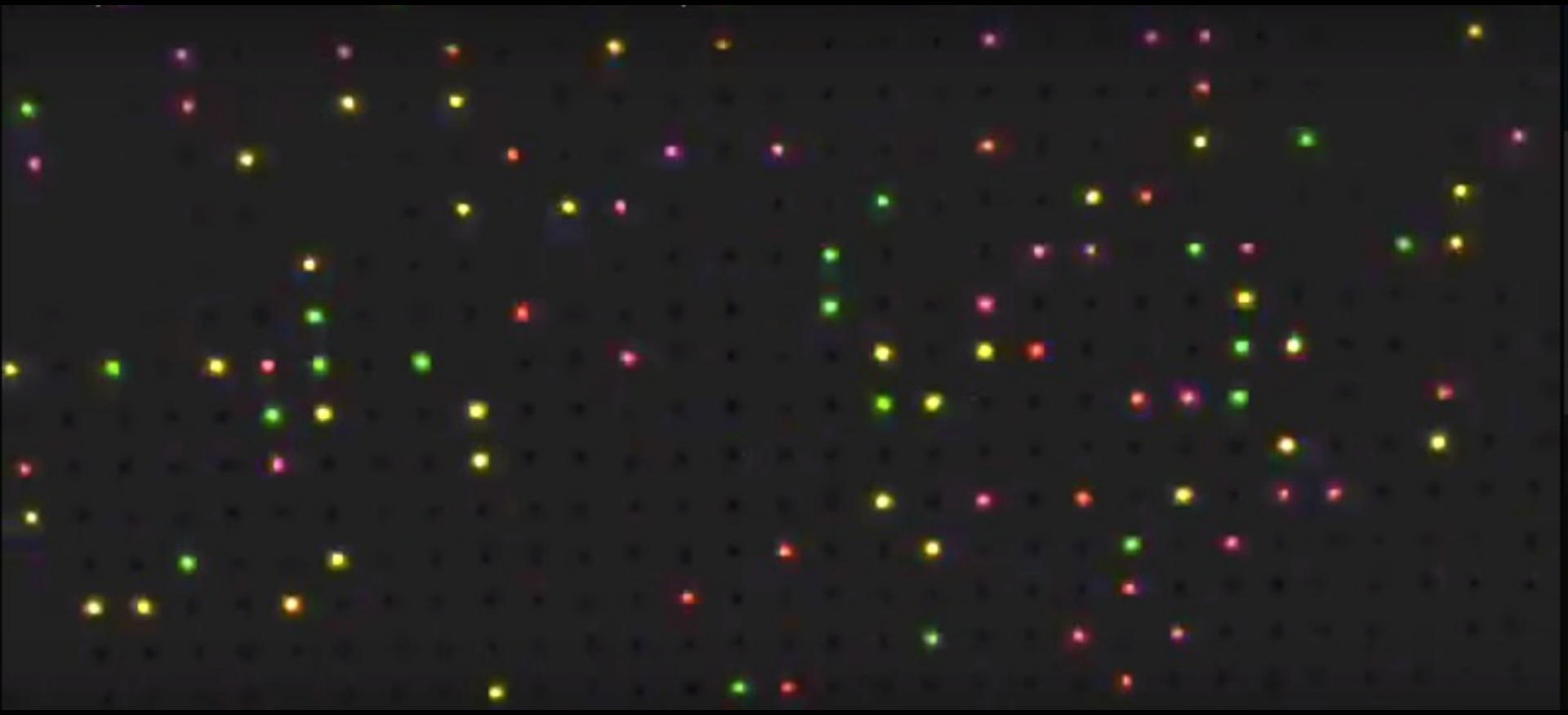


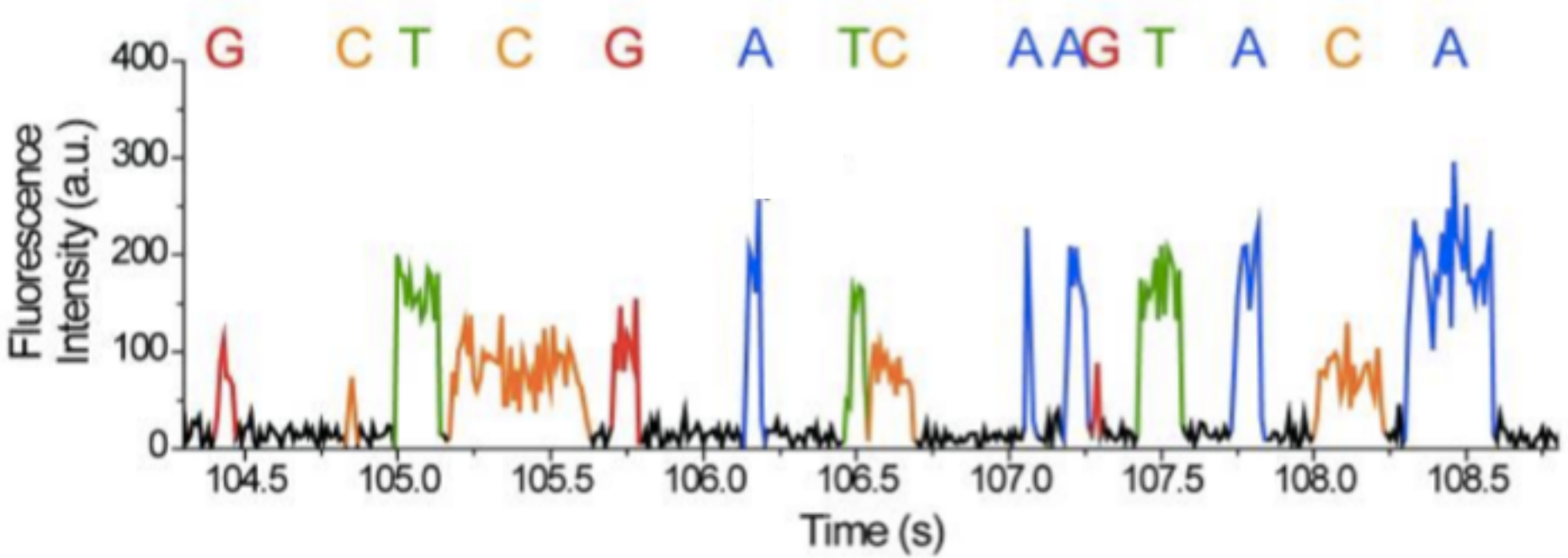
PACIFIC BIOSCIENCES



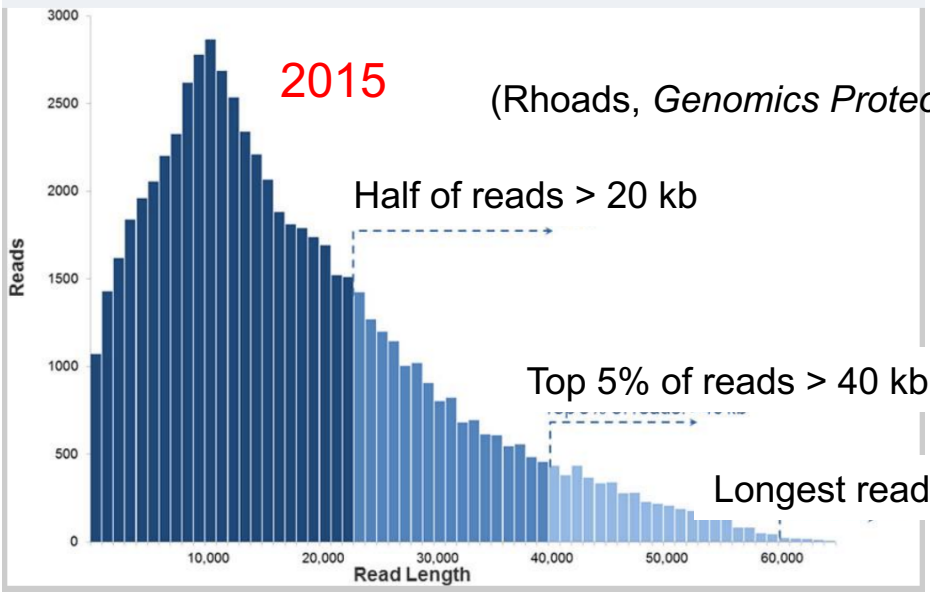
PACIFIC BIOSCIENCES





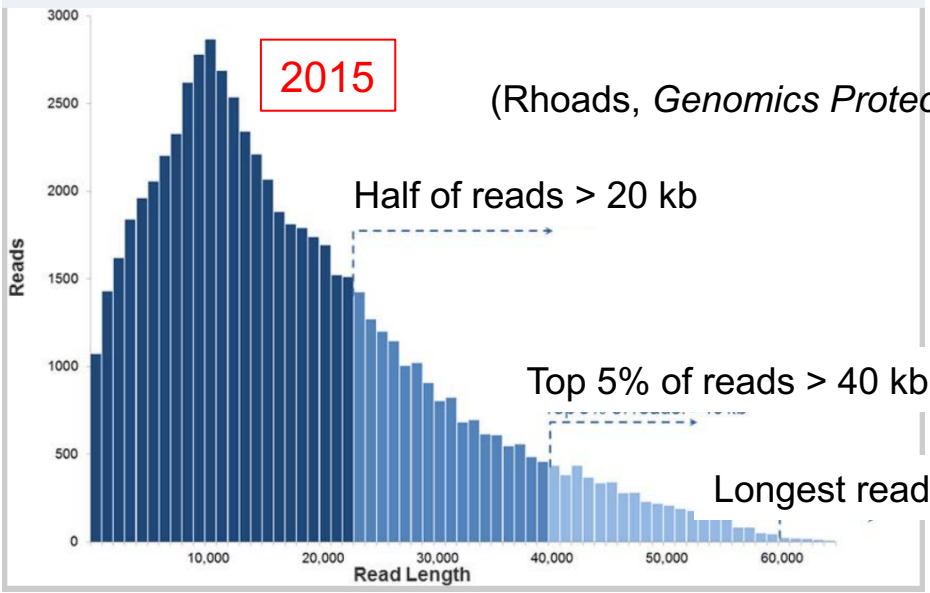

G C T C G A TC AAG T A C A
400
300
200
100
0
Fluorescence Intensity (a.u.)
104.5
105.0
105.5
106.0
106.5
107.0
107.5
108.0
108.5
Time (s)

Length of PacBio reads



(Rhoads, *Genomics Proteomics Bioinformatics*, 2015)

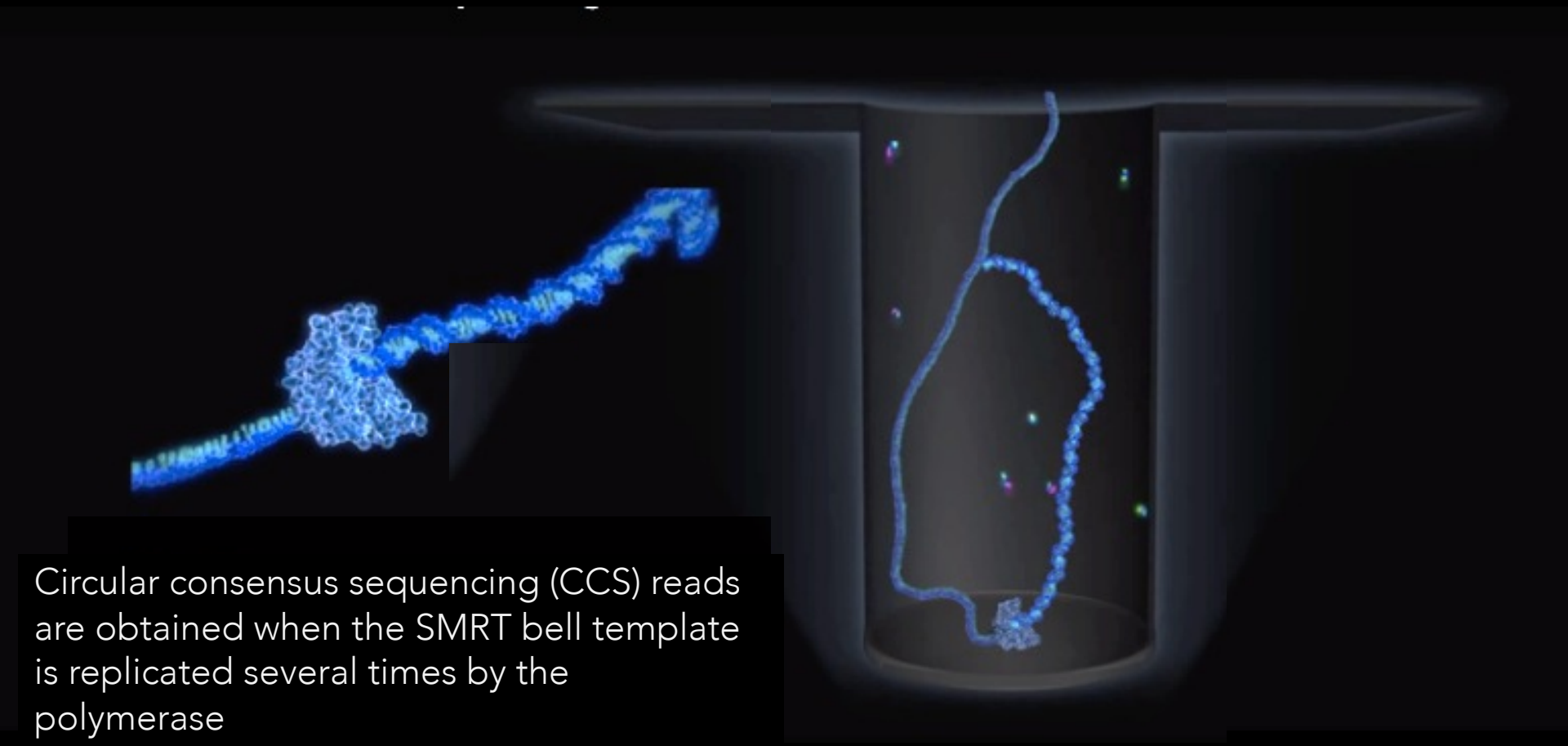
Length of PacBio reads



(Rhoads, *Genomics Proteomics Bioinformatics*, 2015)

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

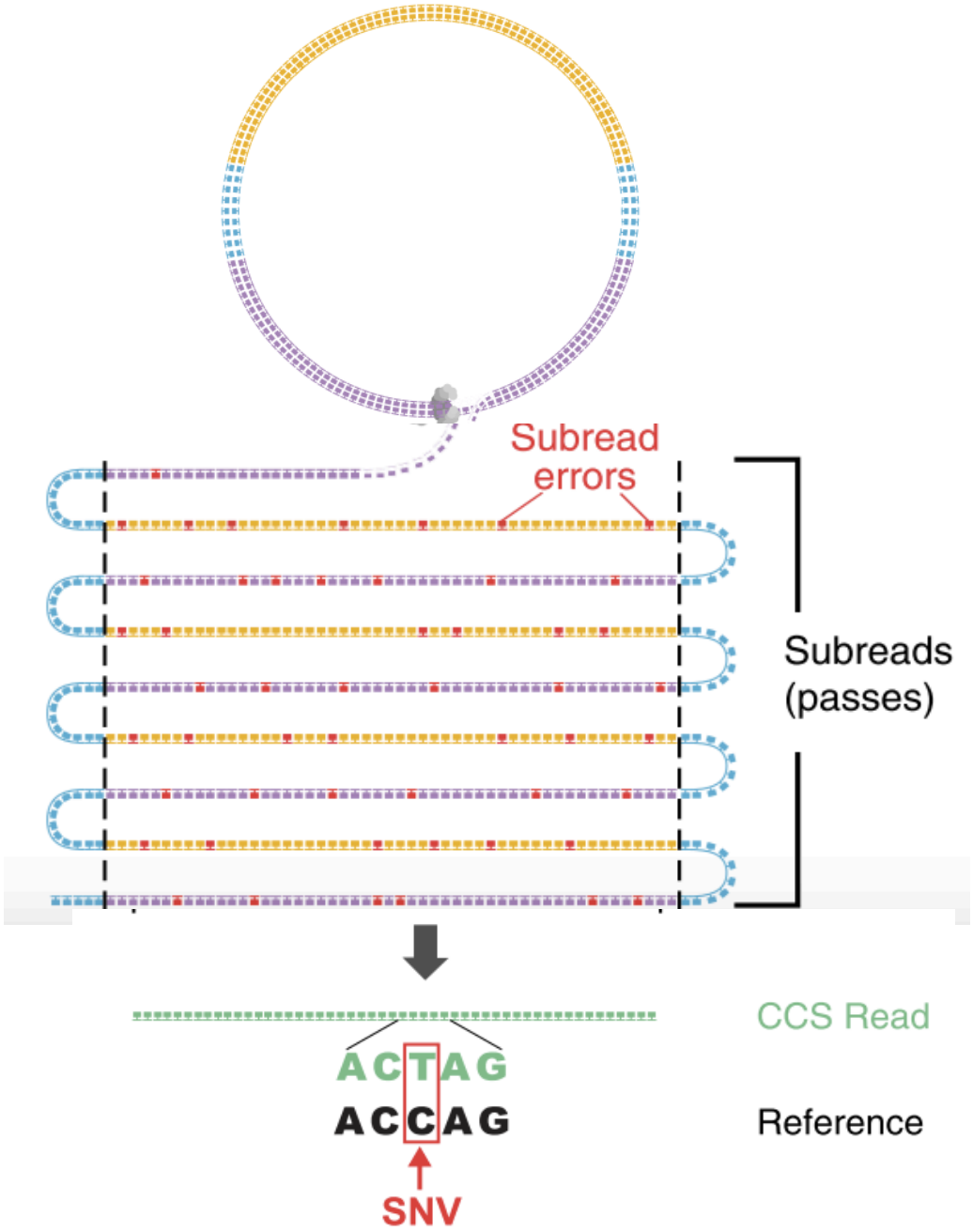
RECENT IMPROVEMENT WITH NEW CHEMISTRY



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

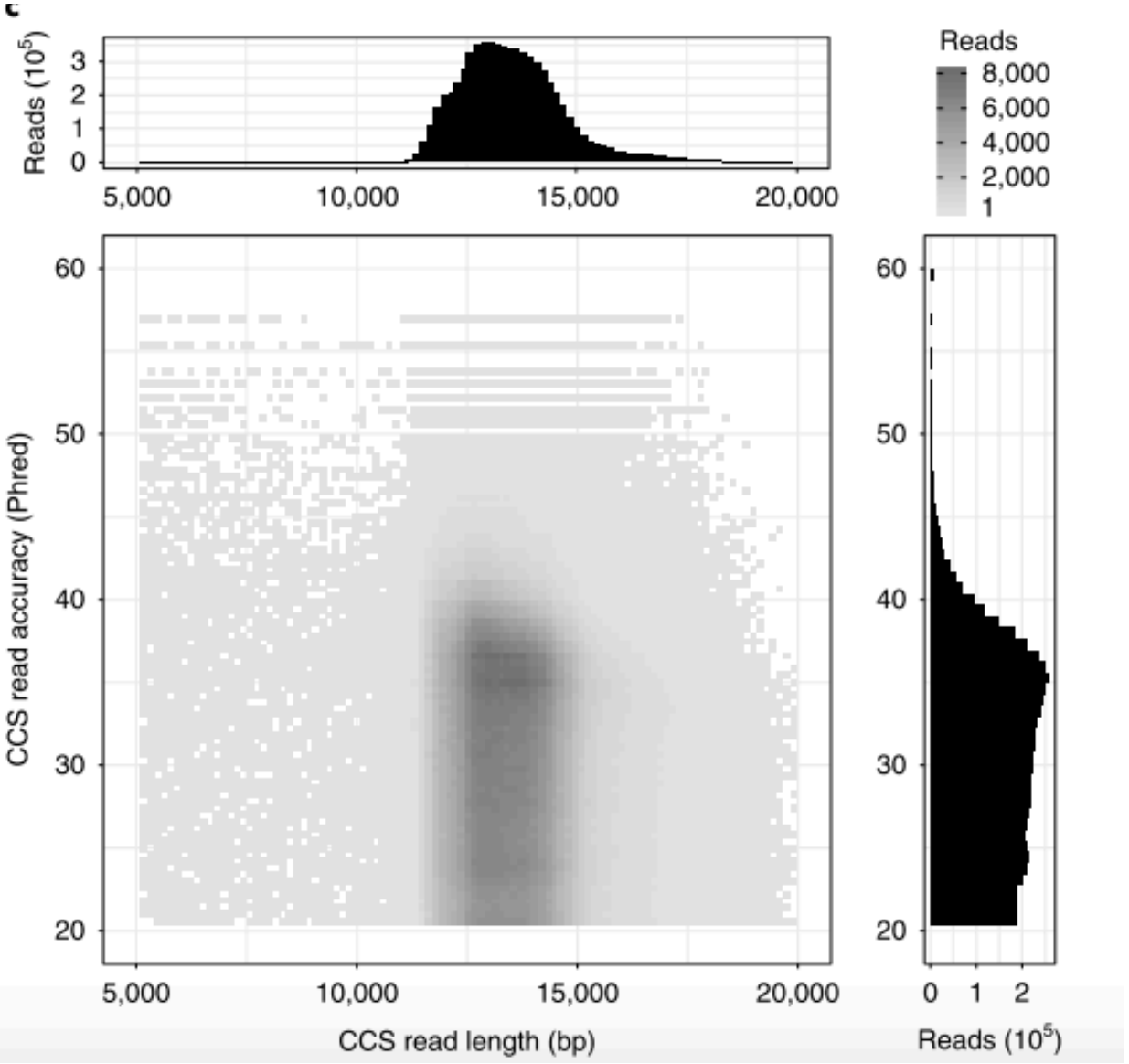
RECENT IMPROVEMENT: GENOME ASSEMBLY WITH CCS



RECENT IMPROVEMENT: GENOME ASSEMBLY WITH CCS

Circular consensus assembly of a human genome

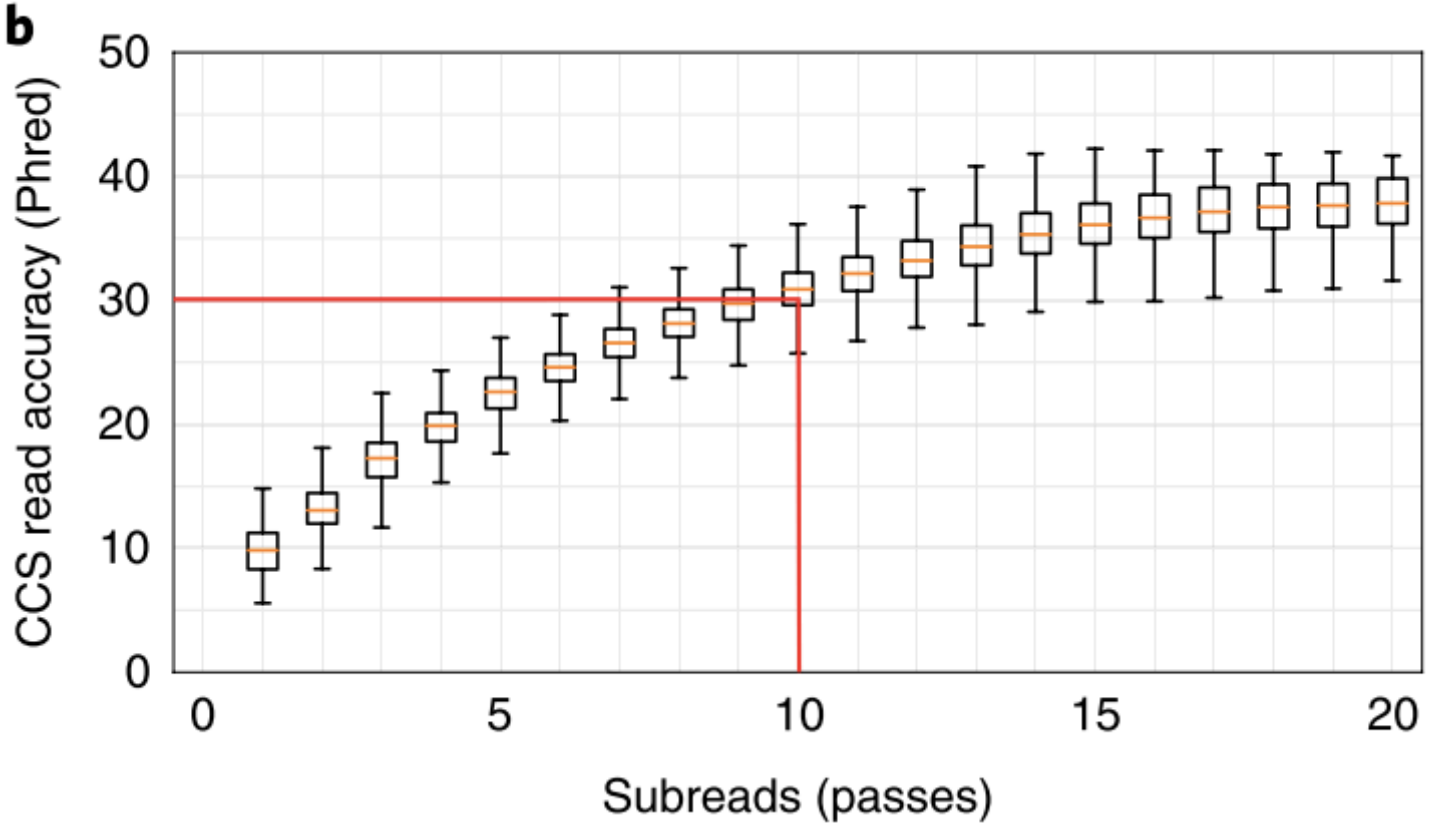
Wenger et al. *Nat. Biotechnol.* oct. 2019



RECENT IMPROVEMENT: GENOME ASSEMBLY WITH CCS

Circular consensus assembly of a human genome

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



RECENT IMPROVEMENT: GENOME ASSEMBLY WITH CCS

Circular consensus assembly of a human genome

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

Table 2 | Statistics for de novo assembly of CCS reads

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

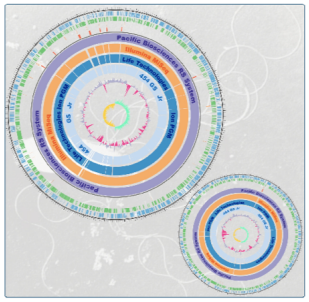
Canu assembly : genome size 3.42 Gb larger than the expected haploid human genome because it resolves some heterozygous alleles into separate contigs

CCS reads alone produced a contiguous and accurate genome with a concordance of 99.997%, substantially ***outperforming assembly with less-accurate long reads***

GENOME ASSEMBLY

Miyamoto et al. *BMC Genomics* 2014

De novo assembly of *Vibrio parahaemolyticus* (5 Mb): two circular chromosomes



Mismatches	GS Jr	Ion PGM	MiSeq	PacBio (>1 M bp)
Number of contigs	309	61	34	2
Number of mismatches	133	108	230	157
Number of indels	824	2853	184	698
Indels length	977	3018	241	794
Number of mismatches per 100 kbp	2.6	2.1	4.5	3.0
Number of indels per 100 kbp	16.3	56.2	3.6	13.5
Number of misassemblies	0	0	1	10
Number of relocations	0	0	1	10
Number of translocations	0	0	0	0
Number of inversions	0	0	0	0
Number of misassembled contigs	0	0	1	2
Genome coverage (%)	97.844	98.290	98.499	99.848
Duplication ratio	1.004	1.000	1.003	1.007

PacBio GENOME ASSEMBLY

De novo Assembly of Two Swedish Genomes Reveals Missing Segments from the Human Reference (hg38)
Ameur et al. *Genes*, 2018

- 10 Mb of the 2 genomes are absent from hg38 reference
- 1 Mb are assigned to chr. Y
- 6 Mb are shared with a Chinese personal genome
- Inclusion of these sequences in GRCh38 genome radically improves alignment and variant calling from short-read data :
- re-analysis :
  - ➢ yields > 75,000 putative novel single nucleotide variants (SNVs)
  - ➢ removes > 10,000 false positive SNV calls per individual

- It becomes possible to represent specific population groups by assembly of representative genomes from different populations.