CNRS UPMC
Station Biologique Roscoff
1872
aviesan
alliance nationale pour les sciences de la vie et de la santé

LONG-READ SEQUENCING

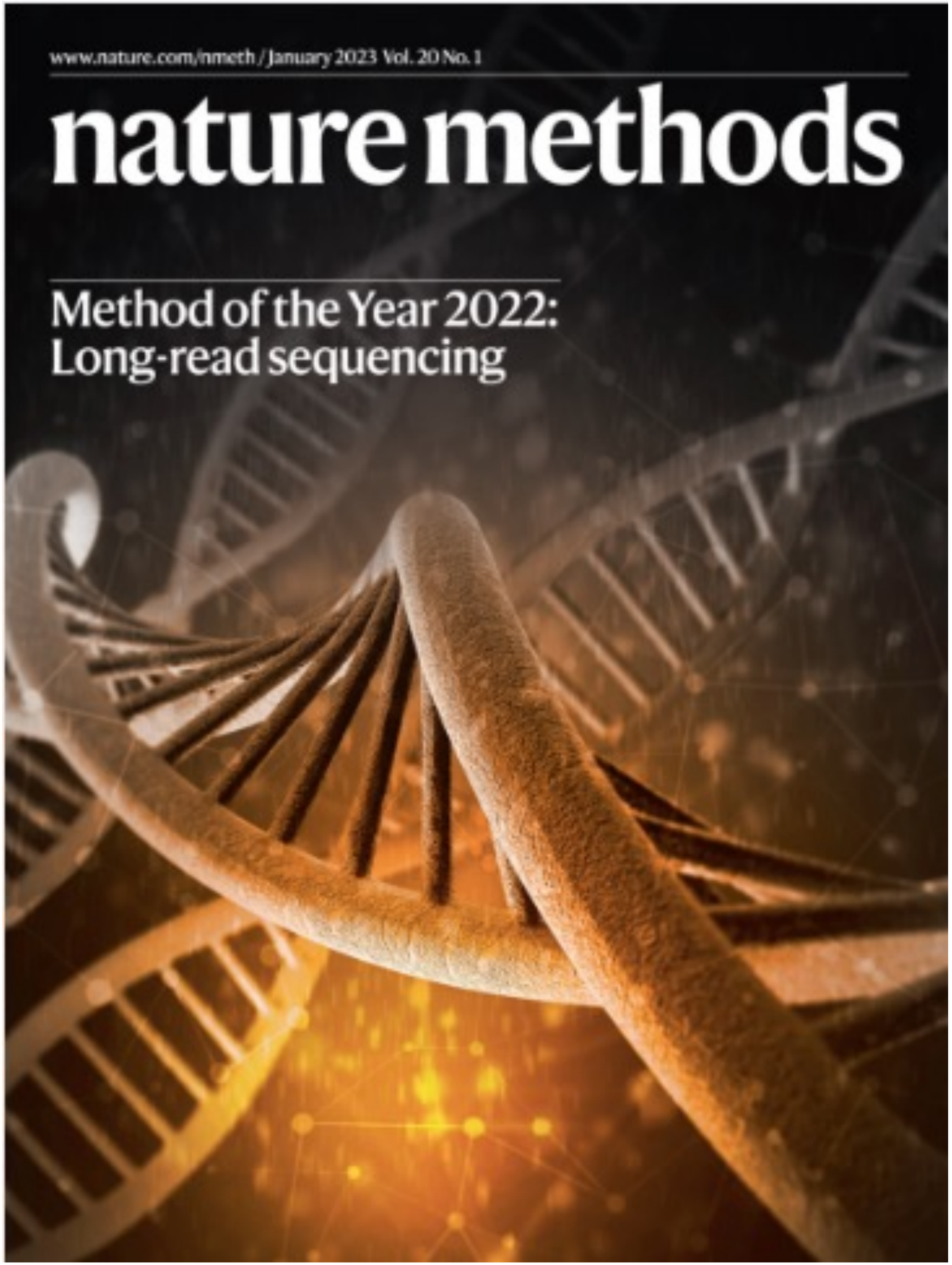
Claude THERMES

PLATEFORME DE SÉQUENÇAGE I2BC

INSTITUT DE BIOLOGIE INTÉGRATIVE DE LA CELLULE

GIF-SUR-YVETTE

12ème ÉCOLE DE BIOINFORMATIQUE EBAII - 06/11/2023

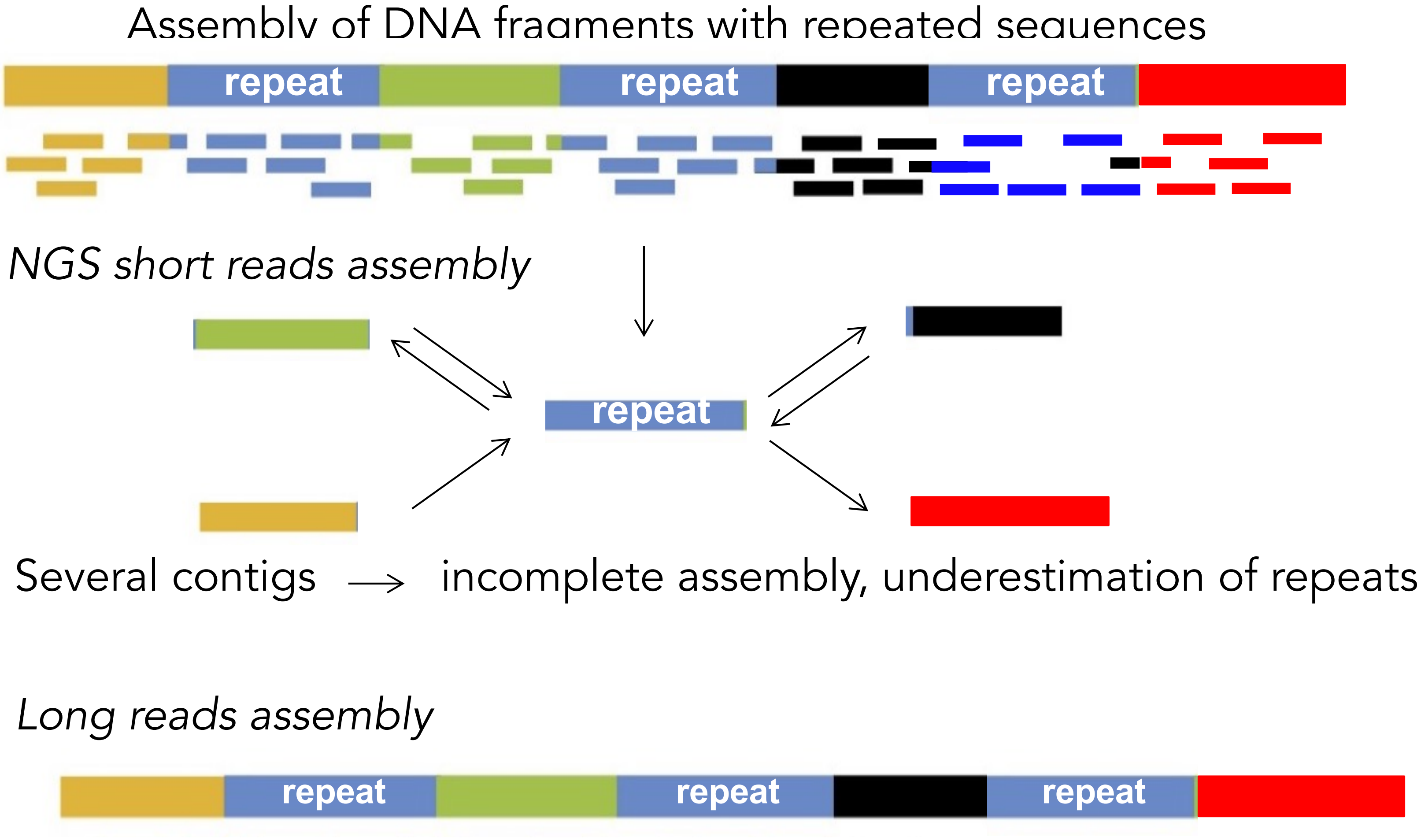


www.nature.com/nmeth / January 2023 Vol. 20 No. 1

nature methods

Method of the Year 2022:
Long-read sequencing

LONG-READS VERSUS SHORT-READS



Assembly of DNA fragments with repeated sequences

NGS short reads assembly

Several contigs → incomplete assembly, underestimation of repeats

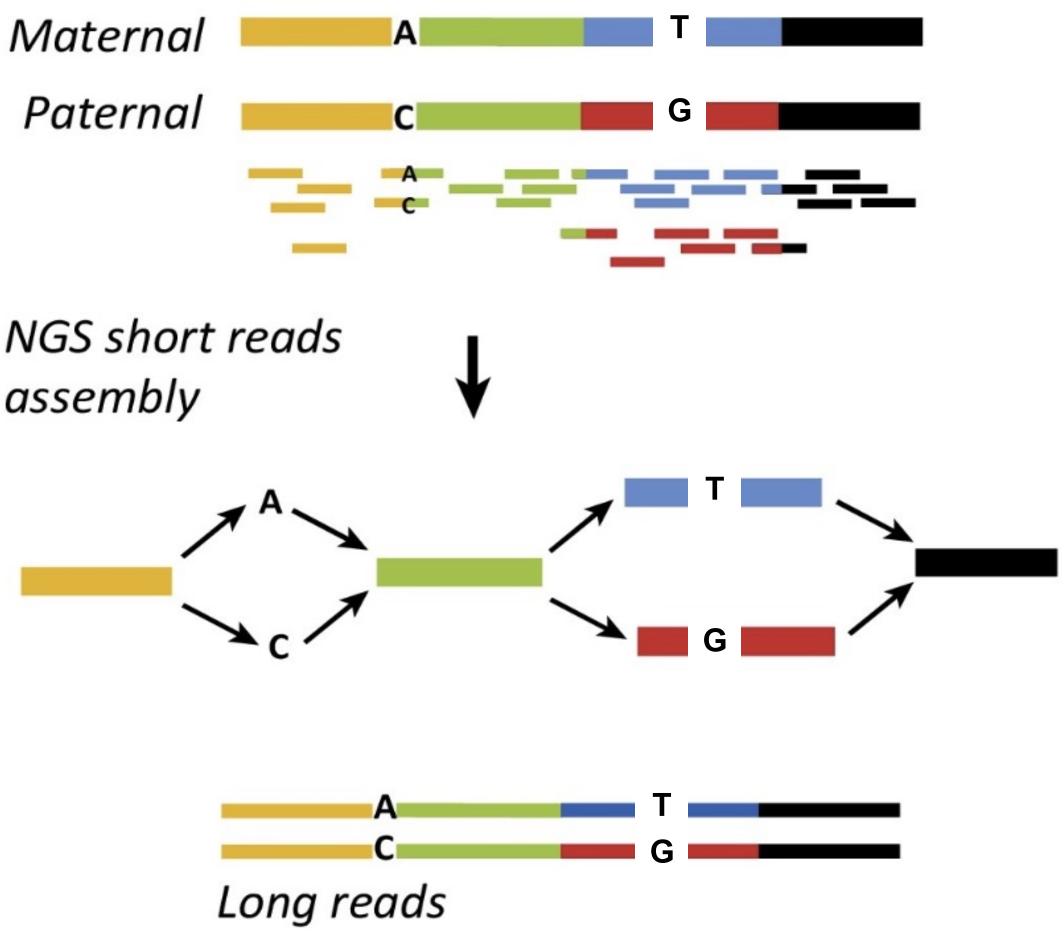
Long reads assembly

Long-reads (1- 200 kb) allow assembly of large repeat-rich regions

(centromeres, telomeres…)

LONG-READS VERSUS SHORT-READS

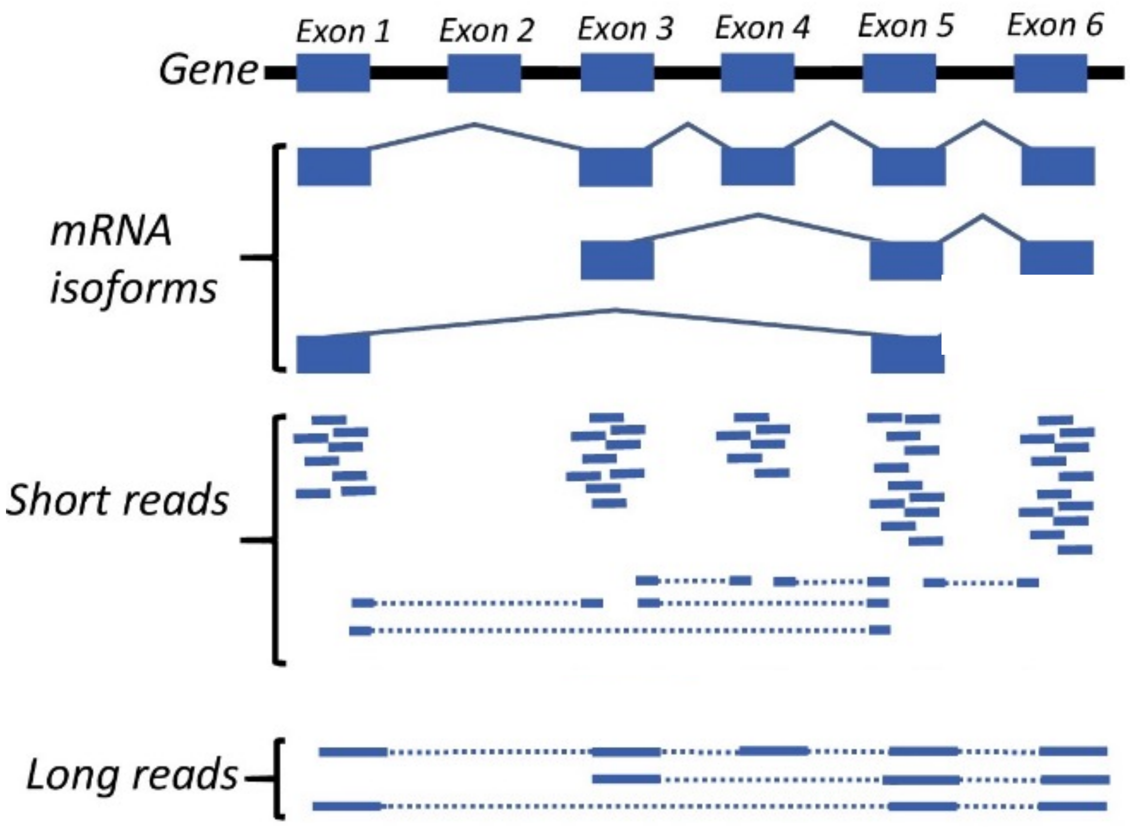
Haplotype phasing



Long-reads facilitate phasing of maternal and paternal haplotypes

LONG-READS VERSUS SHORT-READS

Detection of splicing isoforms



Long-reads allow identification of multiple splicing events along mRNAs

The 3rd generation winning technologies

Pacific Biosciences

Sequel – Revio

Single molecules
Up to 200 kbp long

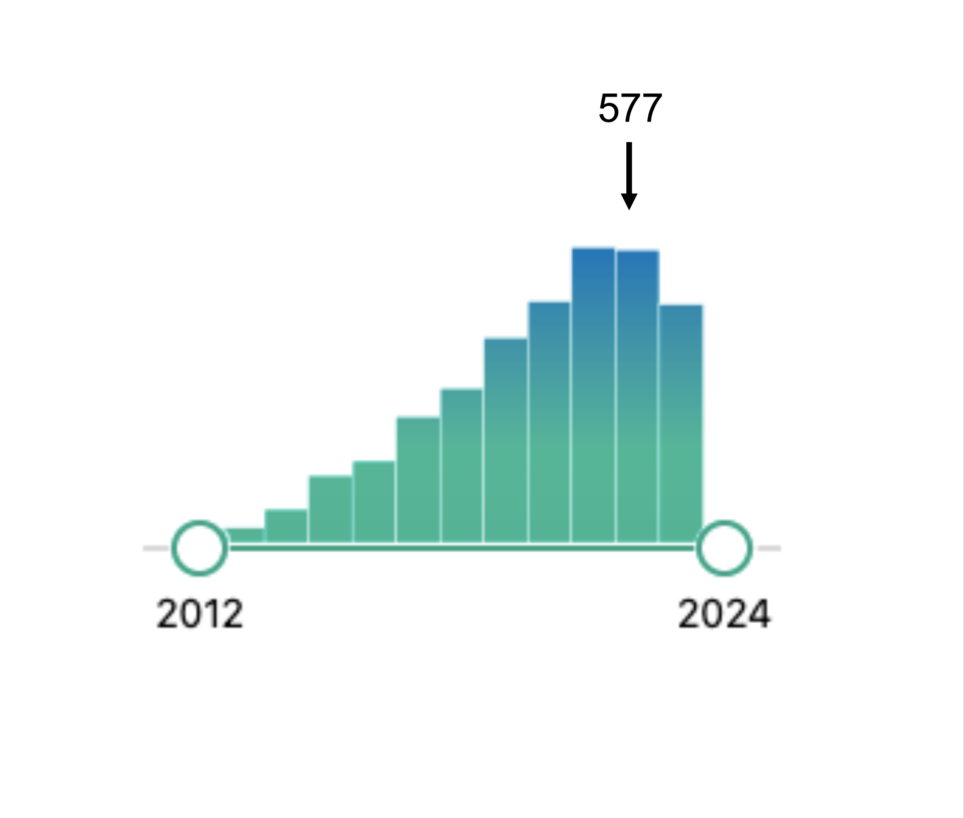
Oxford Nanopore

MinION – PromethION

Single molecules
Up to 1 Mbp long

The 3rd generation winning technologies

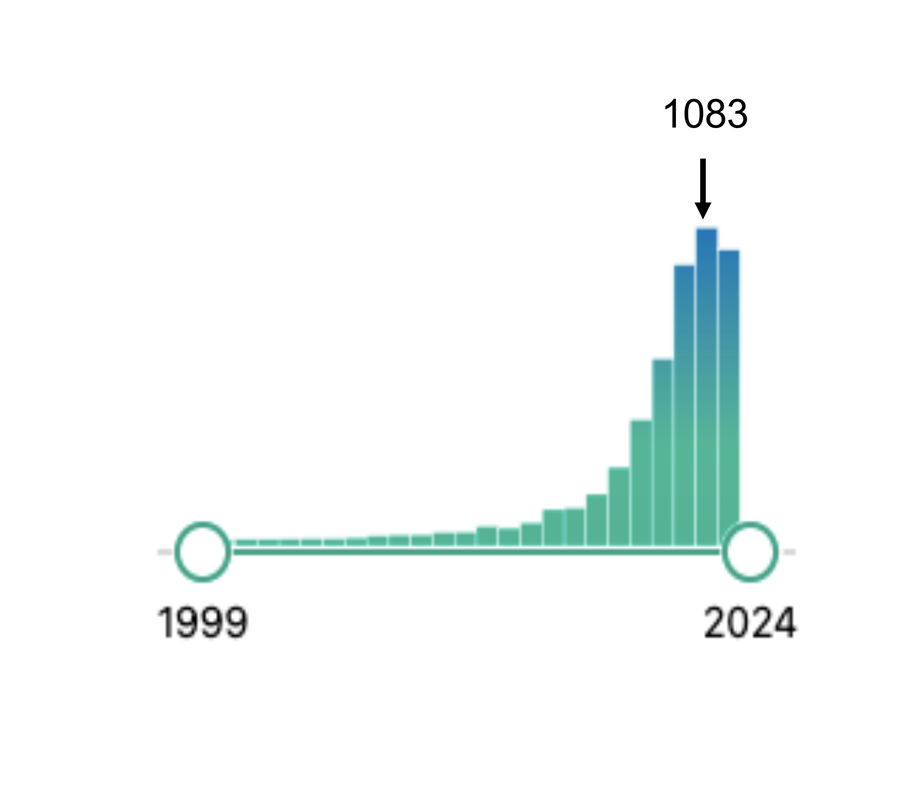
Pacific Biosciences



Sequel – Revio

Single molecules
Up to 200 kbp long

Oxford Nanopore

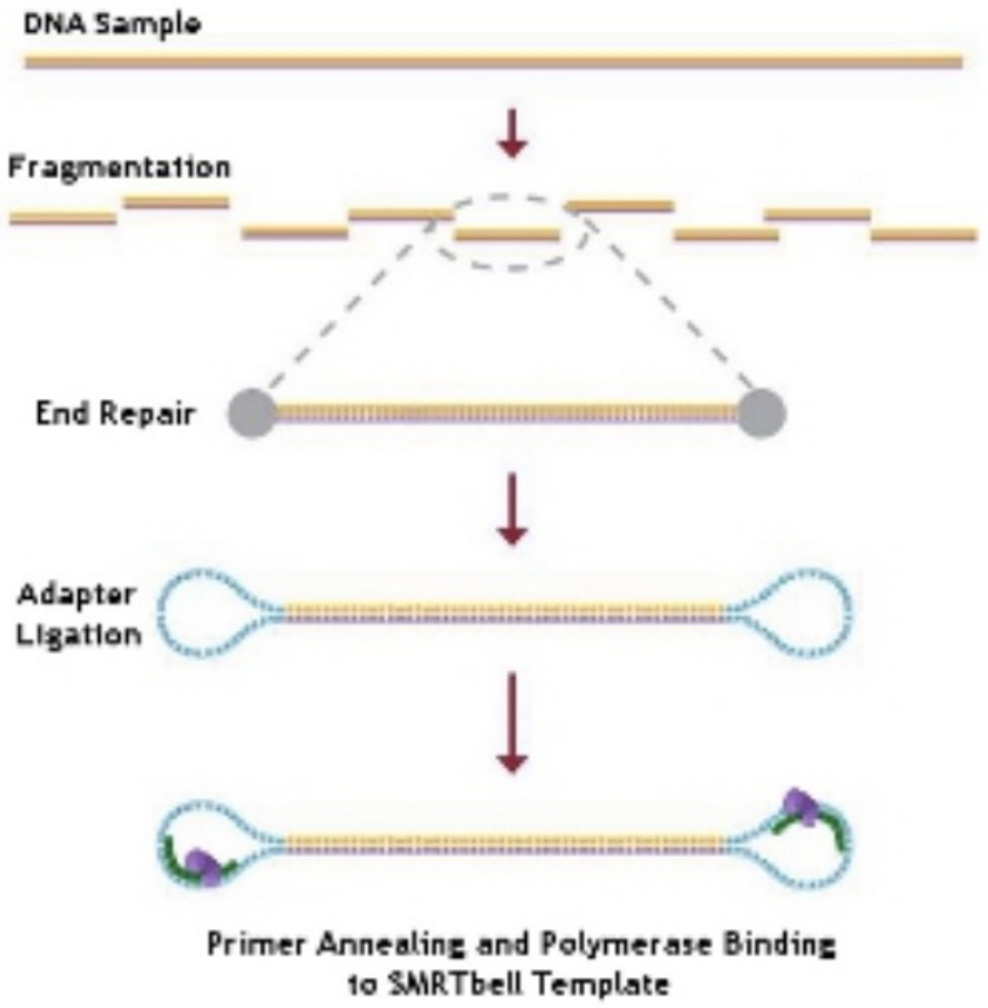


MinION – PromethION

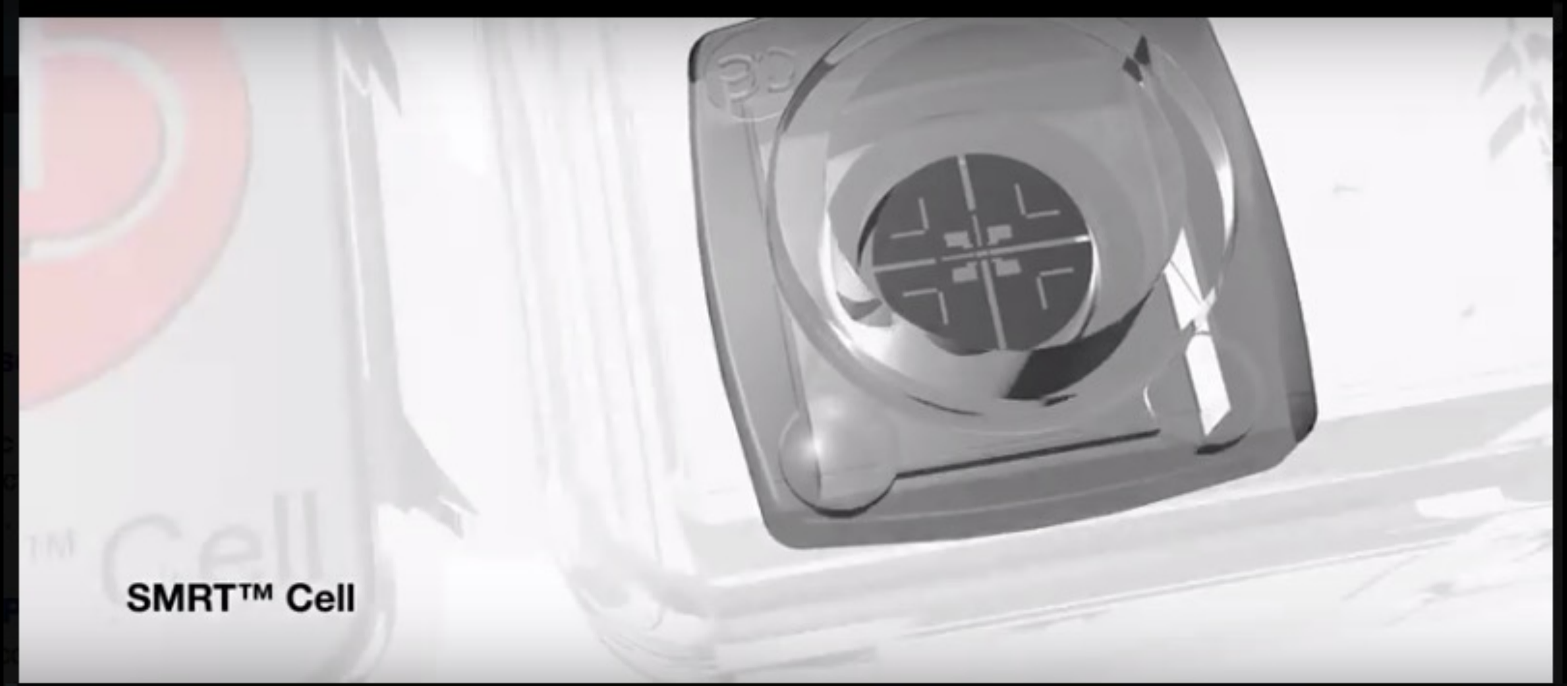
Single molecules
Up to 1 Mbp long

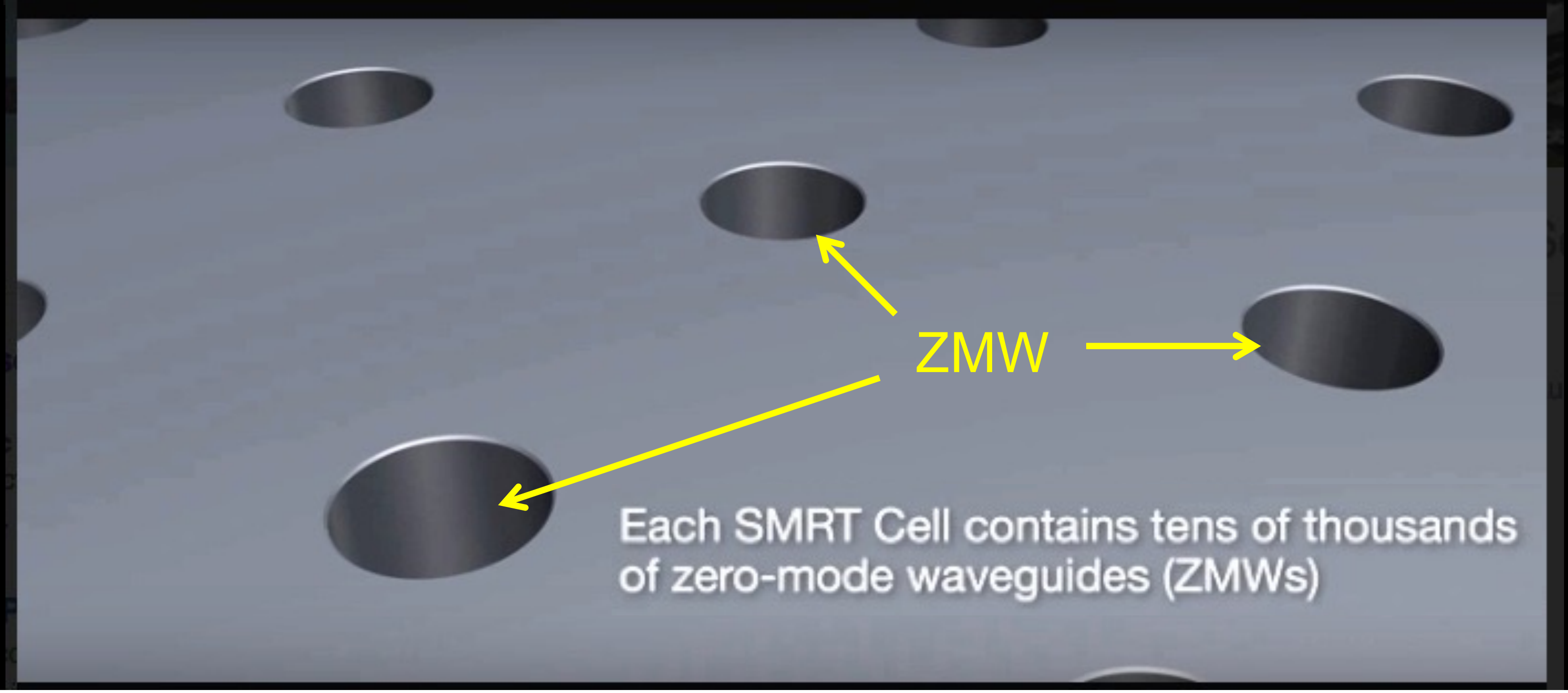
PacBio : Single Molecule Real Time (SMRT) sequencing

PacBio DNA-seq library


DNA Sample
Fragmentation
End Repair
Adapter Ligation
Primer Annealing and Polymerase Binding to SMRTbell Template

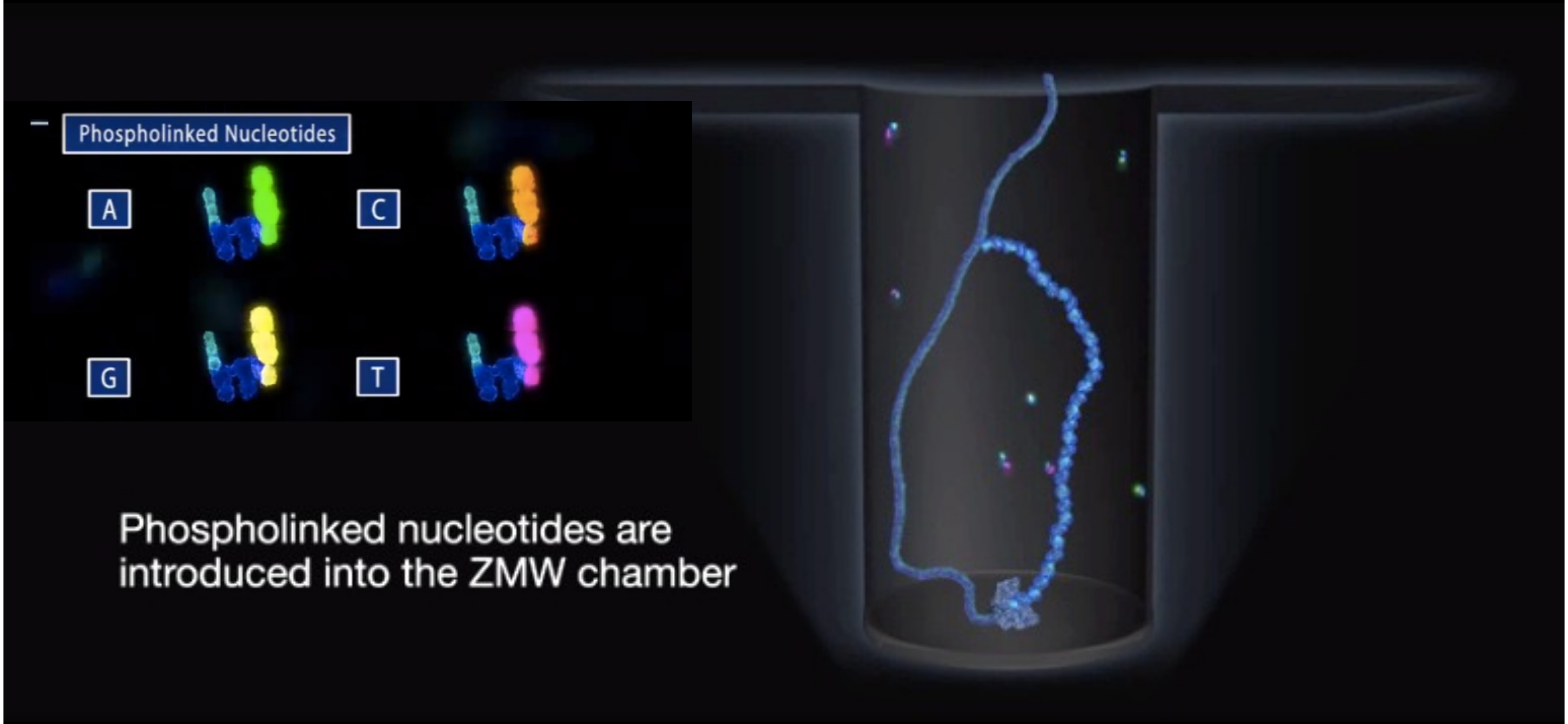
PACIFIC BIOSCIENCES





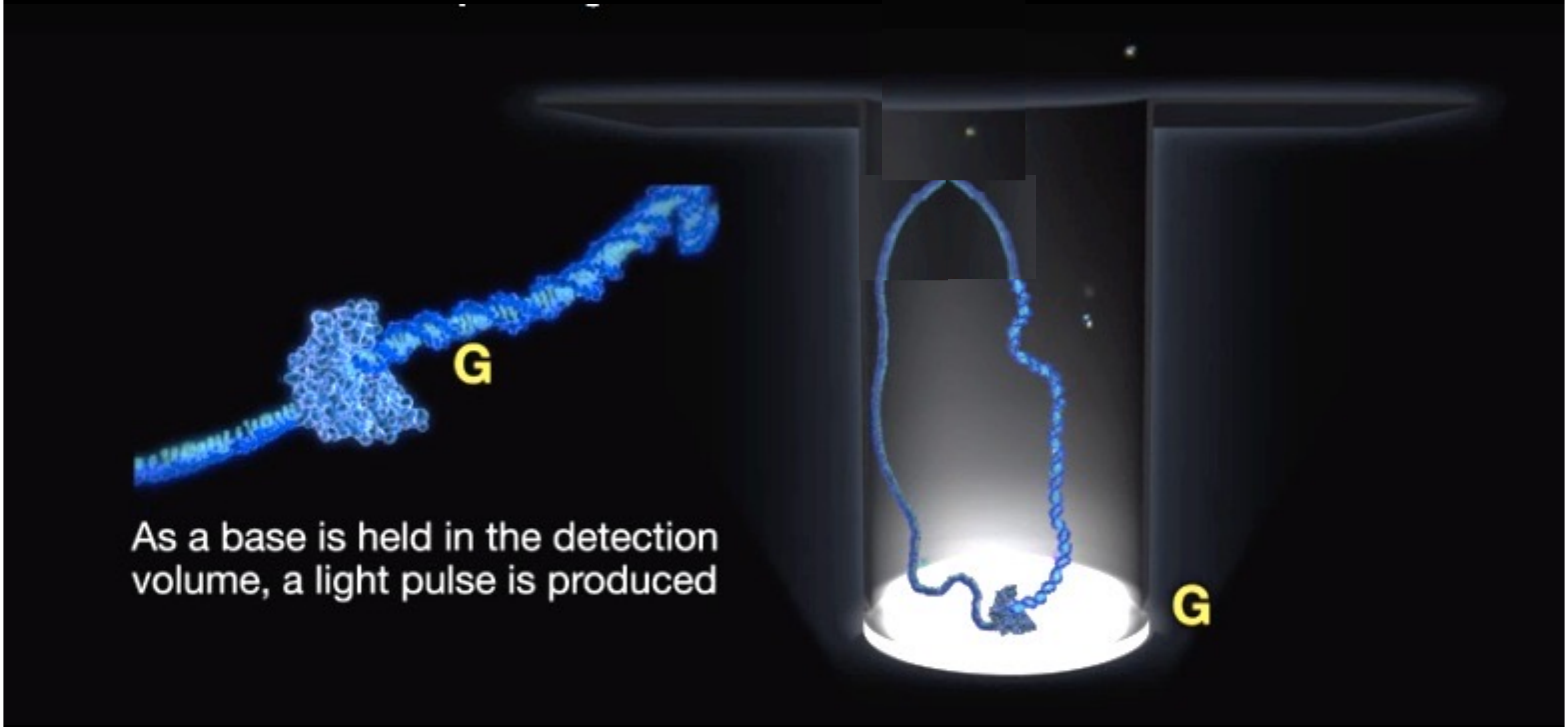
Each SMRT Cell contains tens of thousands of zero-mode waveguides (ZMWs)

PACIFIC BIOSCIENCES



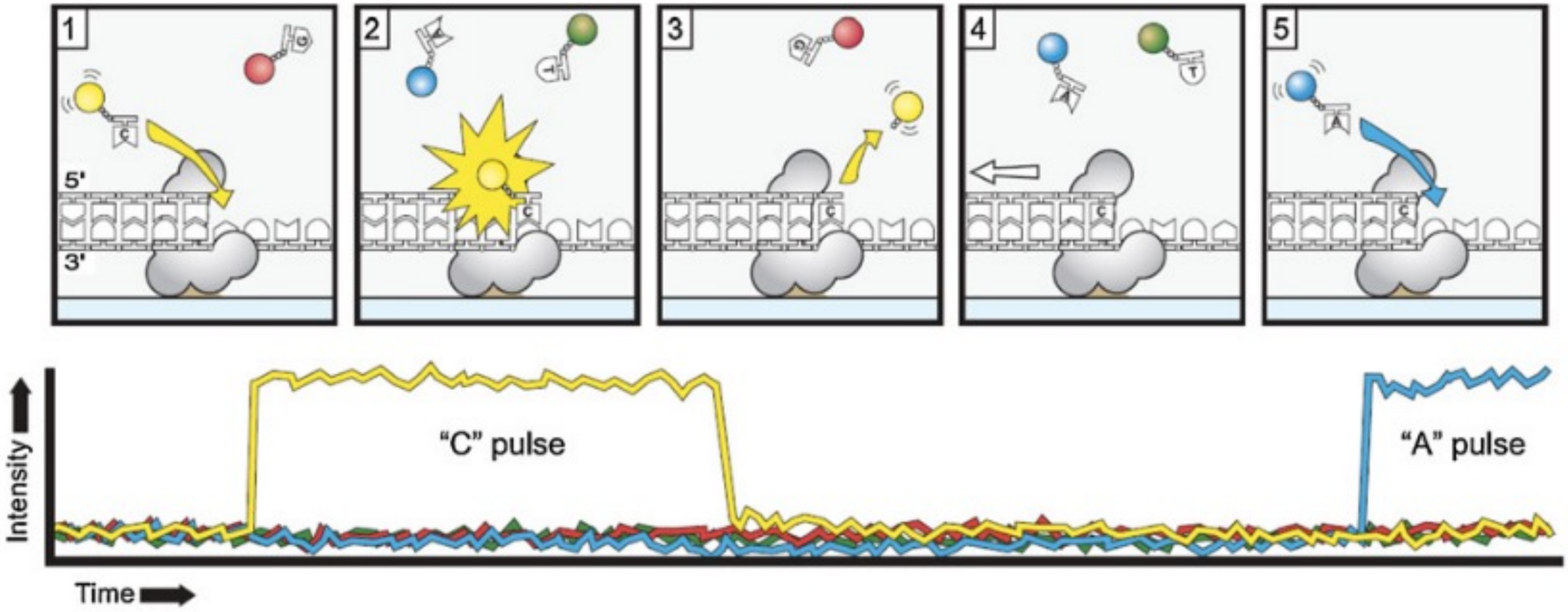
Phospholinked nucleotides are introduced into the ZMW chamber

PACIFIC BIOSCIENCES



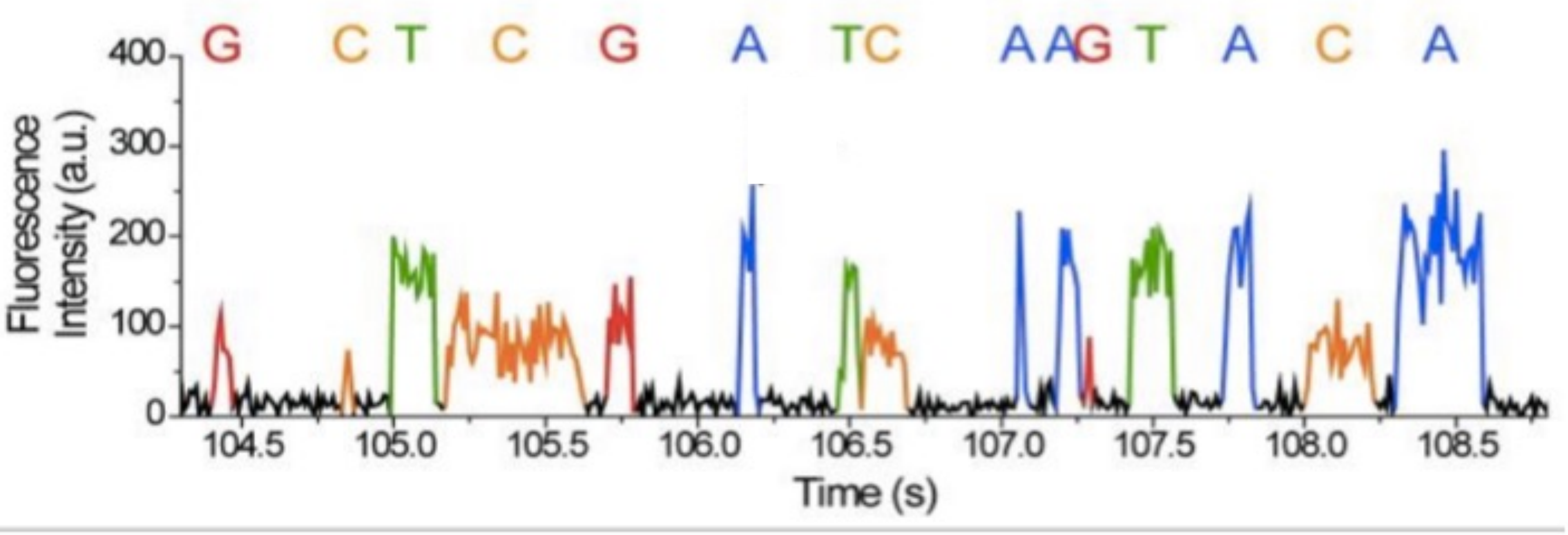
As a base is held in the detection volume, a light pulse is produced

PACIFIC BIOSCIENCES



Eid, J., et al. Science (2009)

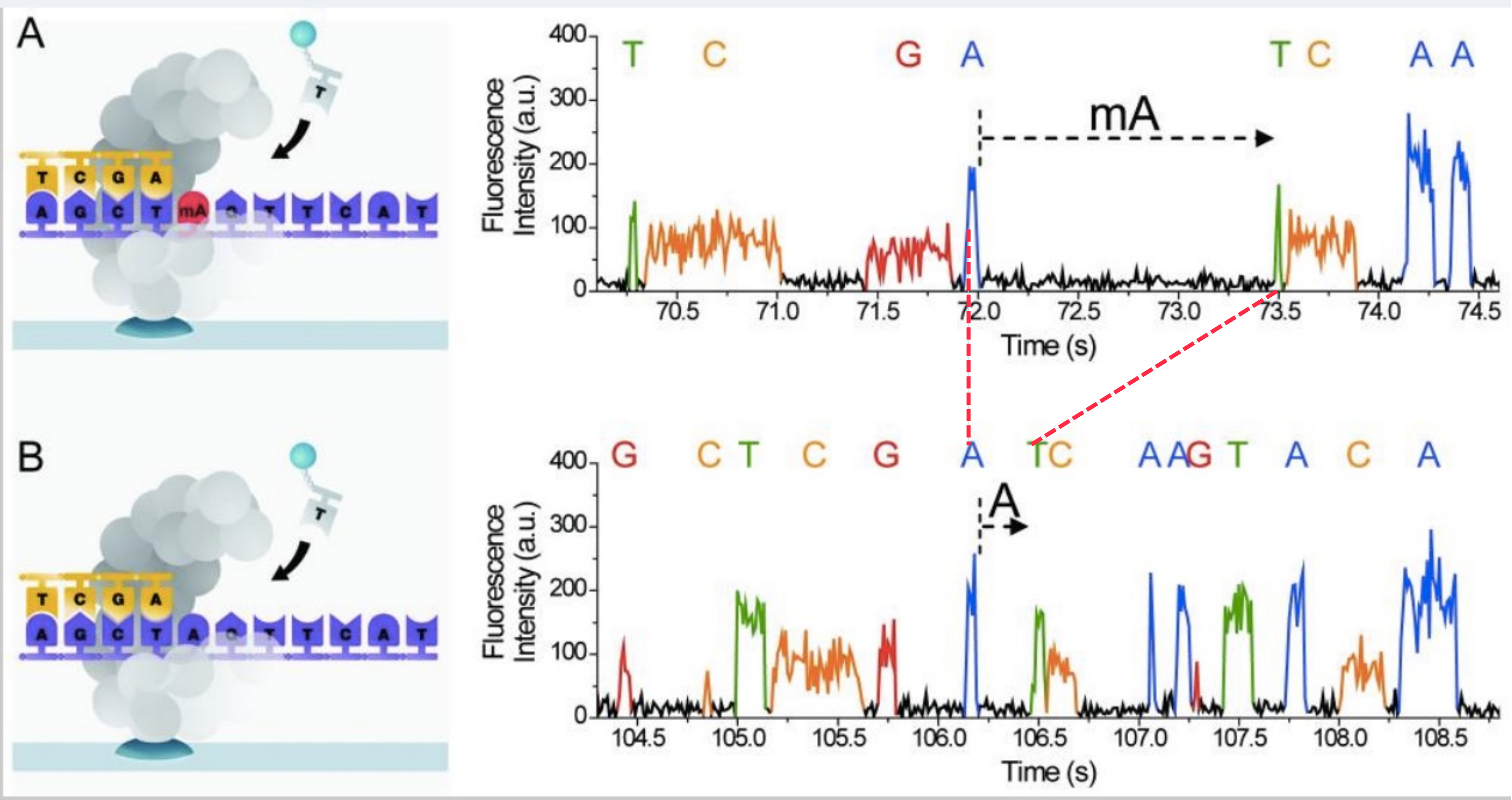
PACIFIC BIOSCIENCES



High error rate : 10% - 15%

Fusberg et al. *Nature Methods* (2010)

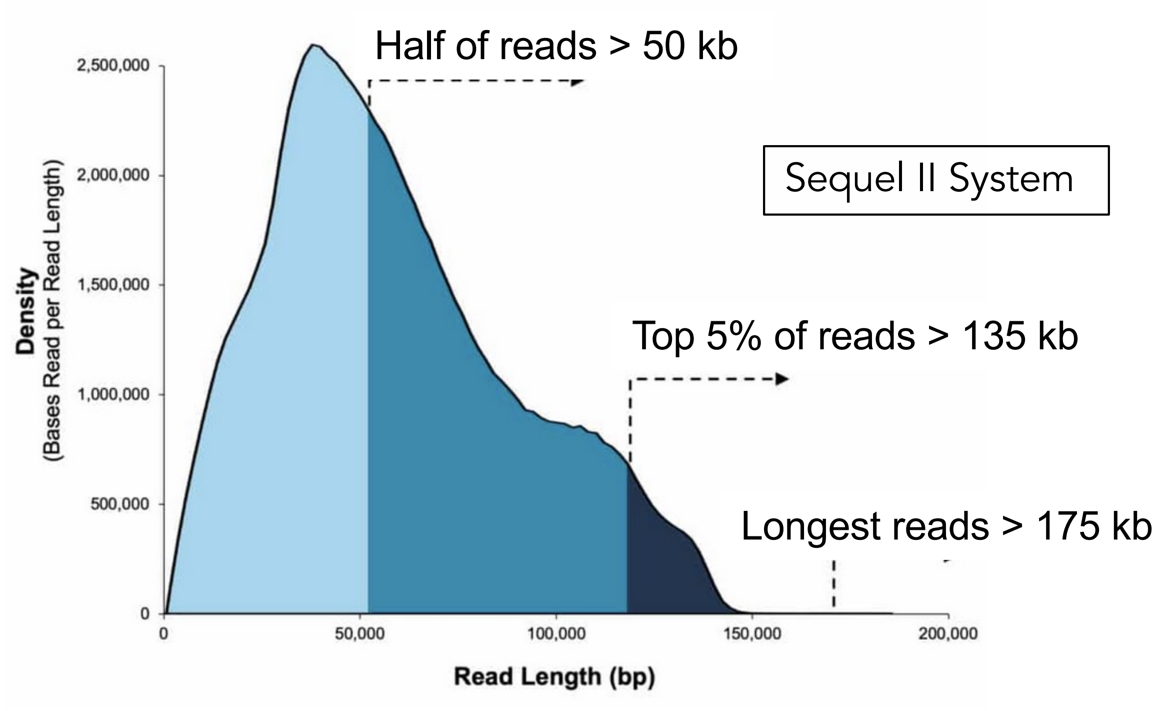
DETECTION OF MODIFIED DNA BASES



Signal modification depends on the neighbor nucleotides (sequence context)

Fusberg et al. *Nature Methods* (2010)

LENGTH OF PACBIO READS



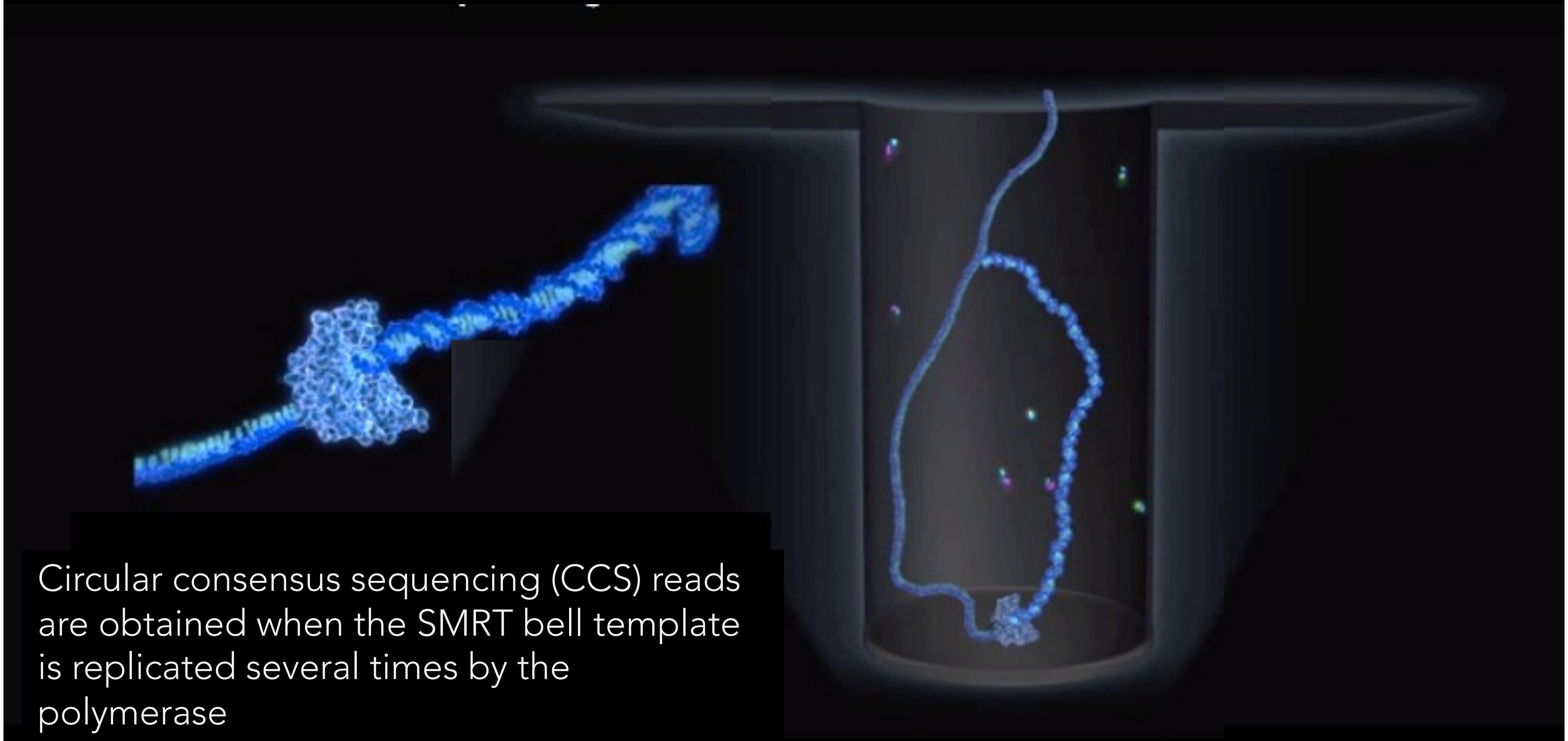
Half of reads > 50 kb

Sequel II System

Top 5% of reads > 135 kb

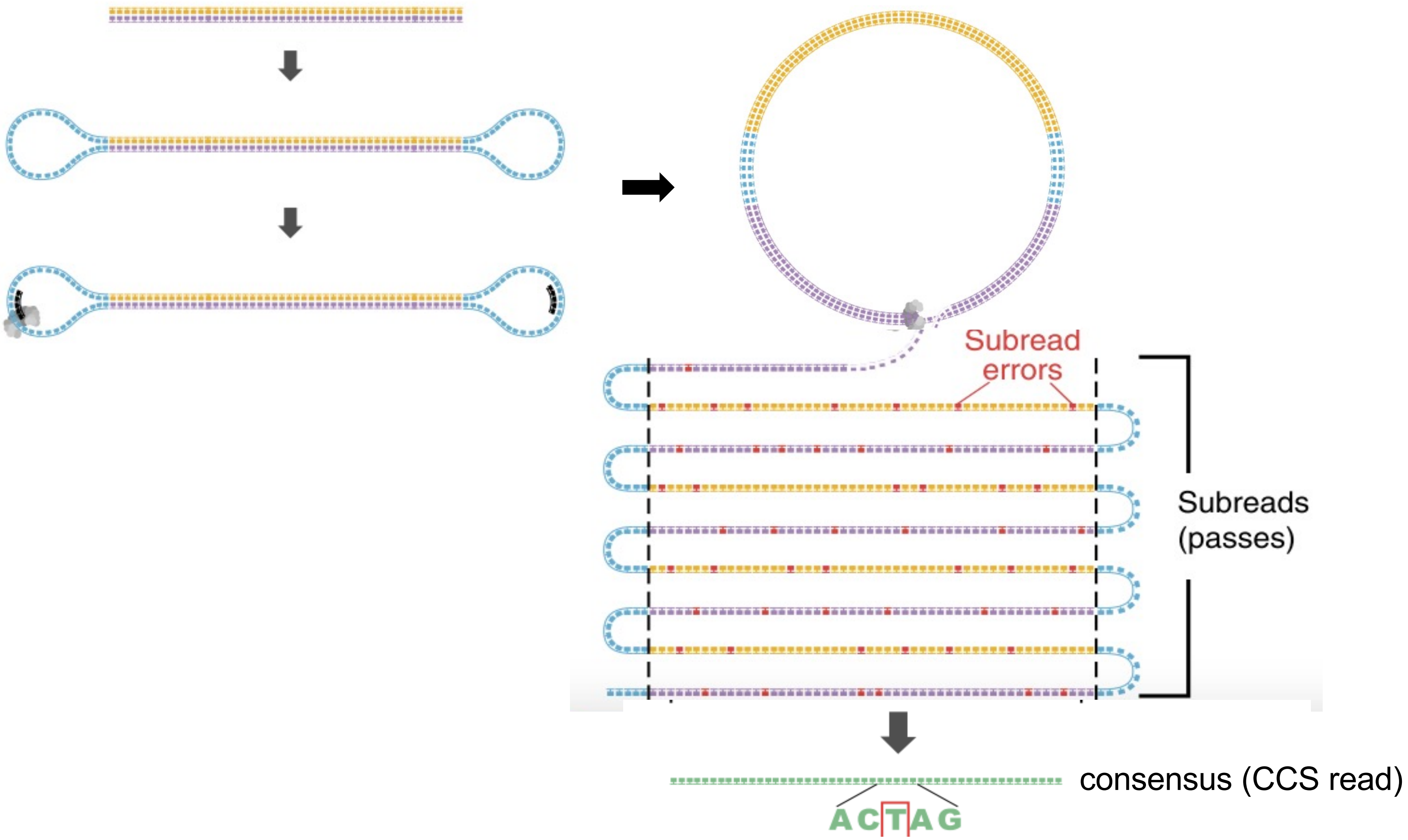
Longest reads > 175 kb

PACIFIC BIOSCIENCES



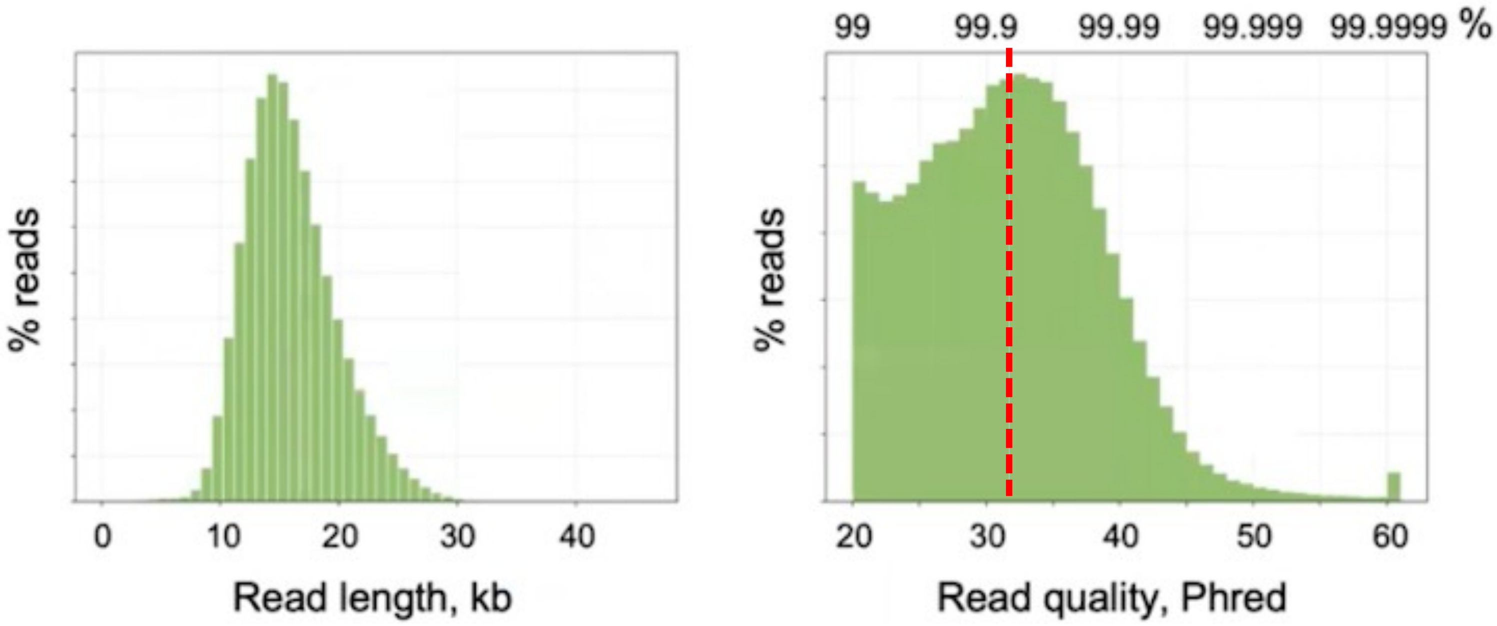
Circular consensus sequencing (CCS) reads are obtained when the SMRT bell template is replicated several times by the polymerase

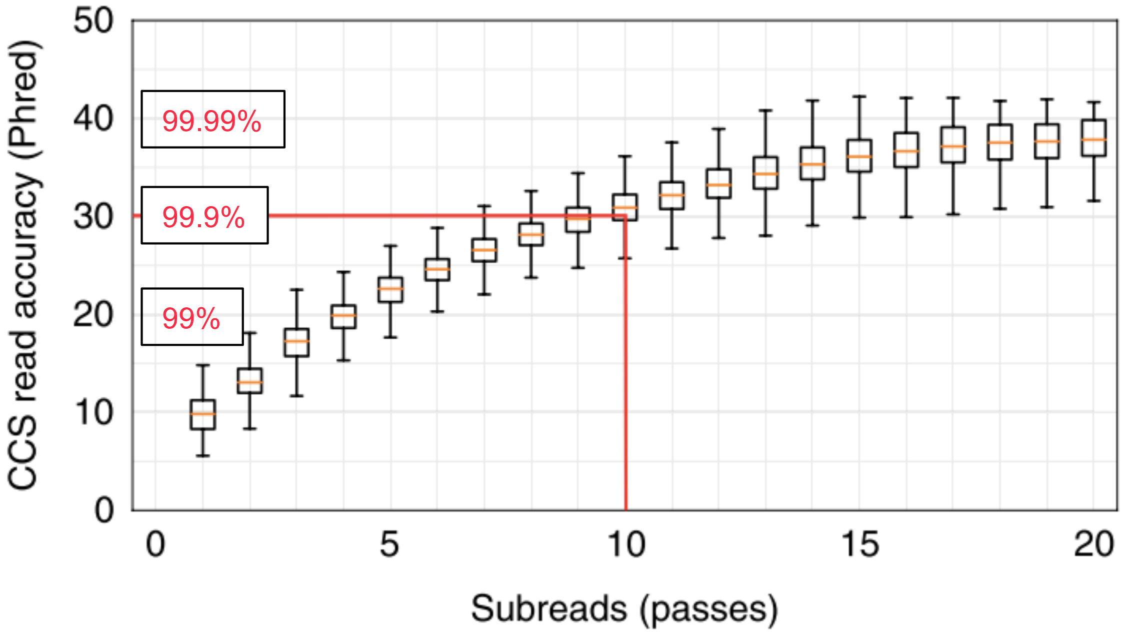
CIRCULAR CONSENSUS SEQUENCES (CCS) : HIFI READS



Randomly positioned errors ➡ they can be corrected

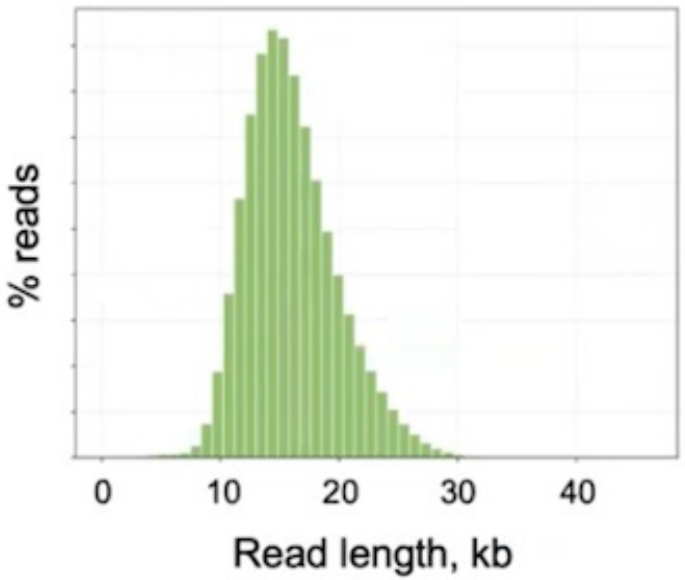
CIRCULAR CONSENSUS SEQUENCES (CCS) : HIFI READS

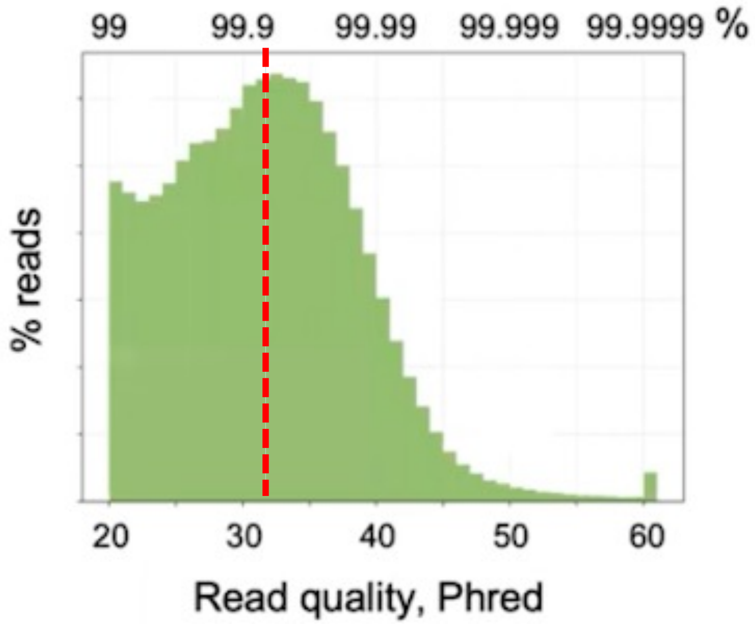




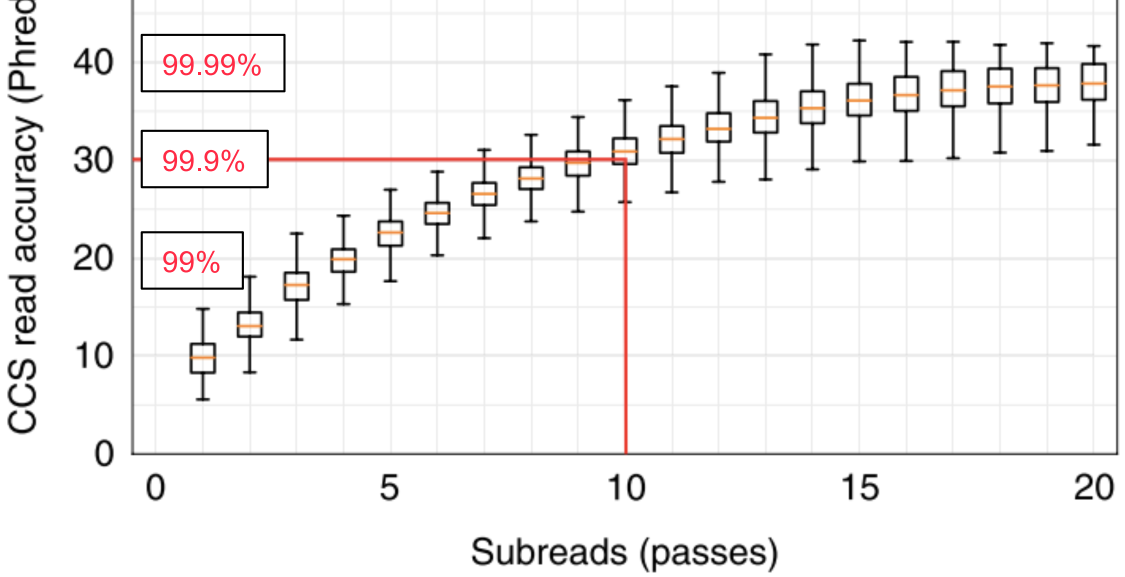
er et al. *Nat. Biotechnol.* (2019)

CIRCULAR CONSENSUS SEQUENCES (CCS) : HIFI READS





Mean accuracy > 99.9 % (Q > 30)

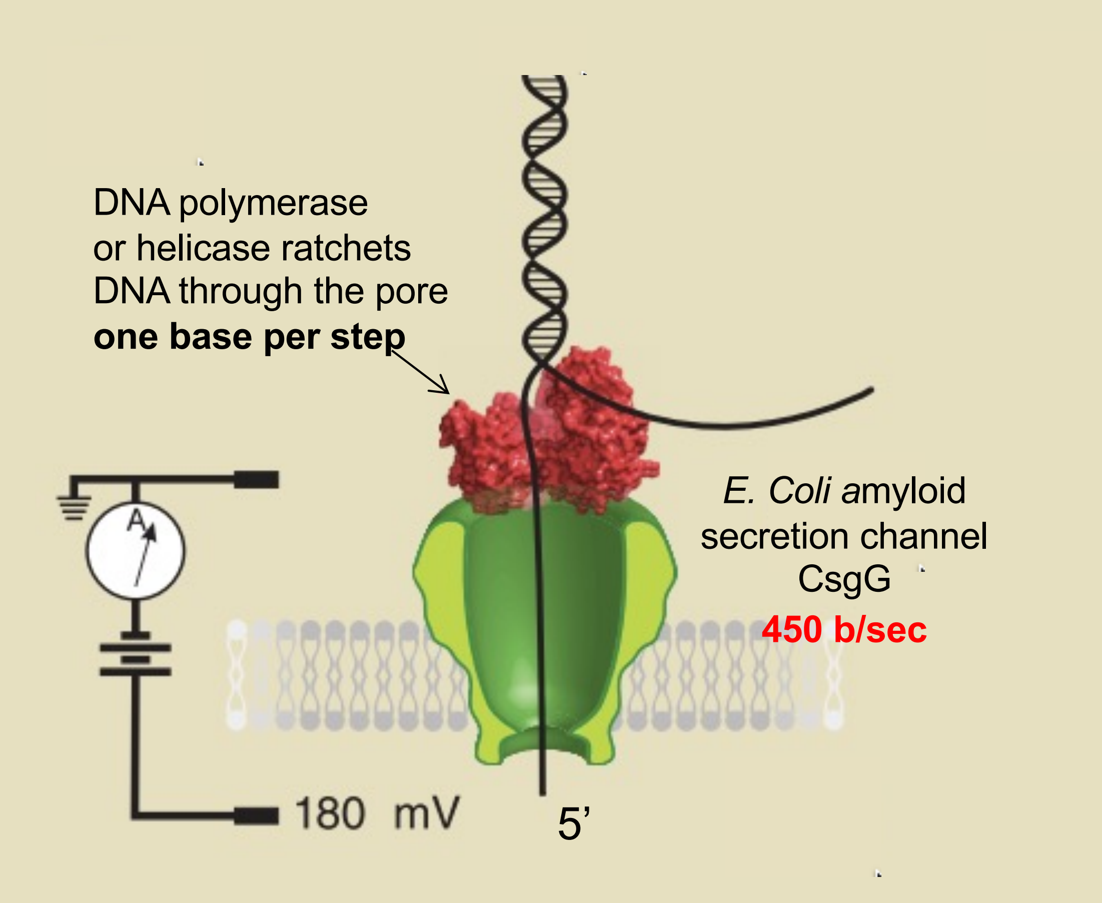


er et al. *Nat. Biotechnol.* (2019)

Next Generation Sequencing

Oxford NANOPORE Technologies®

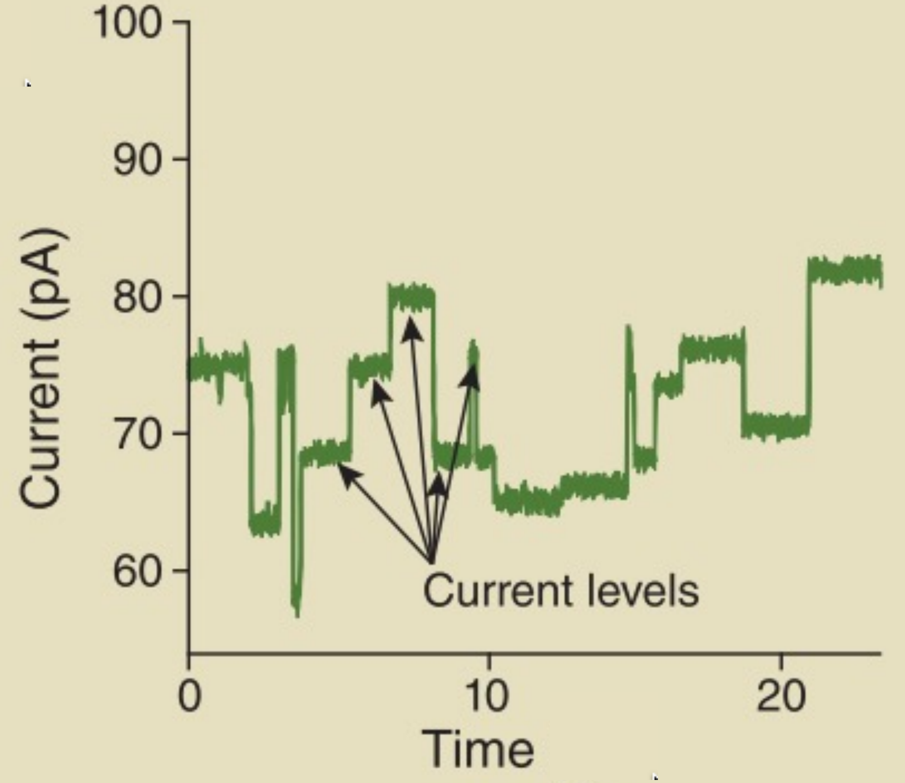
BASIC CONCEPTS



DNA polymerase or helicase ratchets DNA through the pore **one base per step**

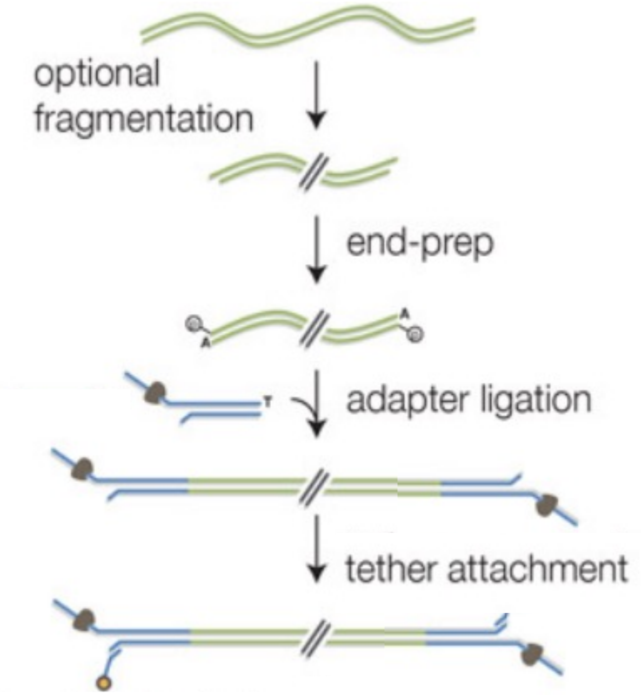
*E. Coli a*myloid secretion channel CsgG

450 b/sec

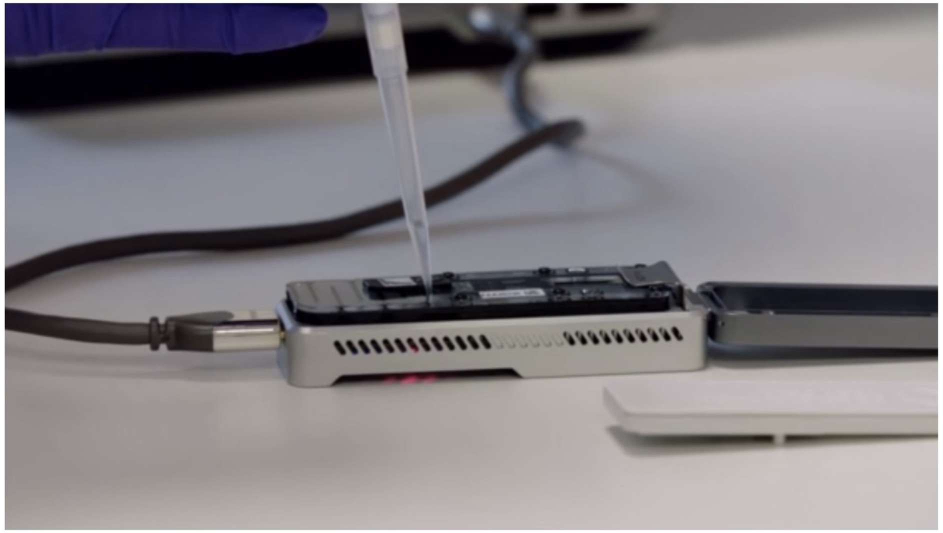


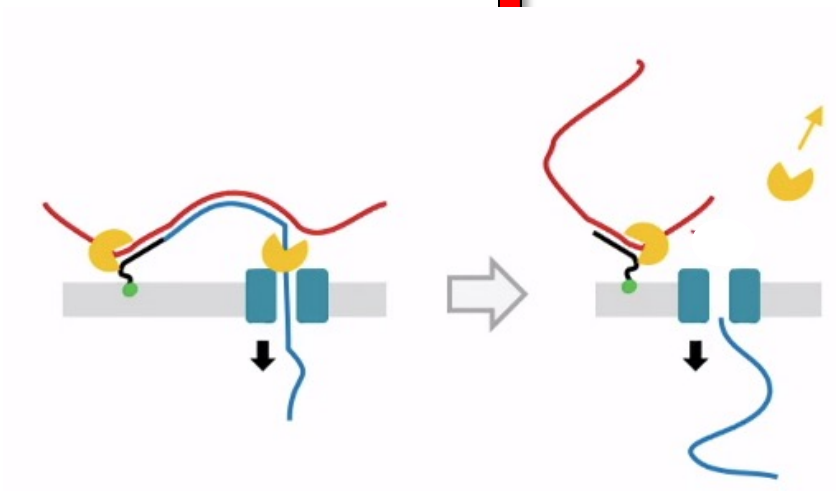
SEQUENCING PROCESS

Library preparation

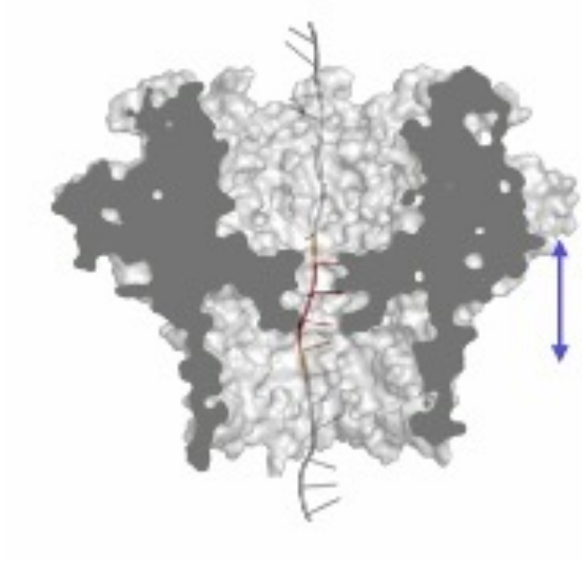


SEQUENCING



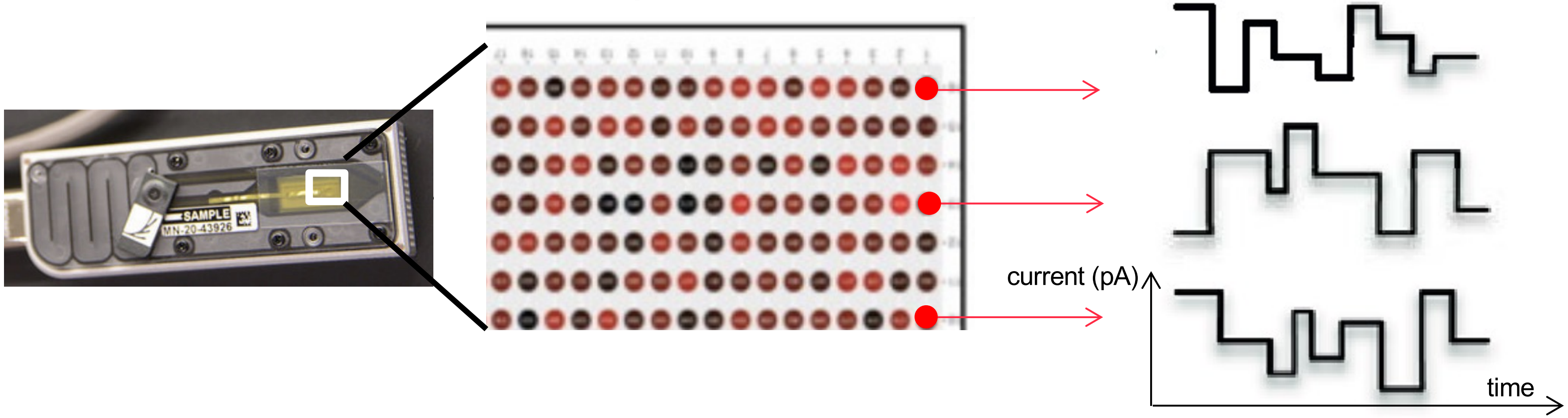


SEQUENCING PROCESS : MinION FLOW CELL



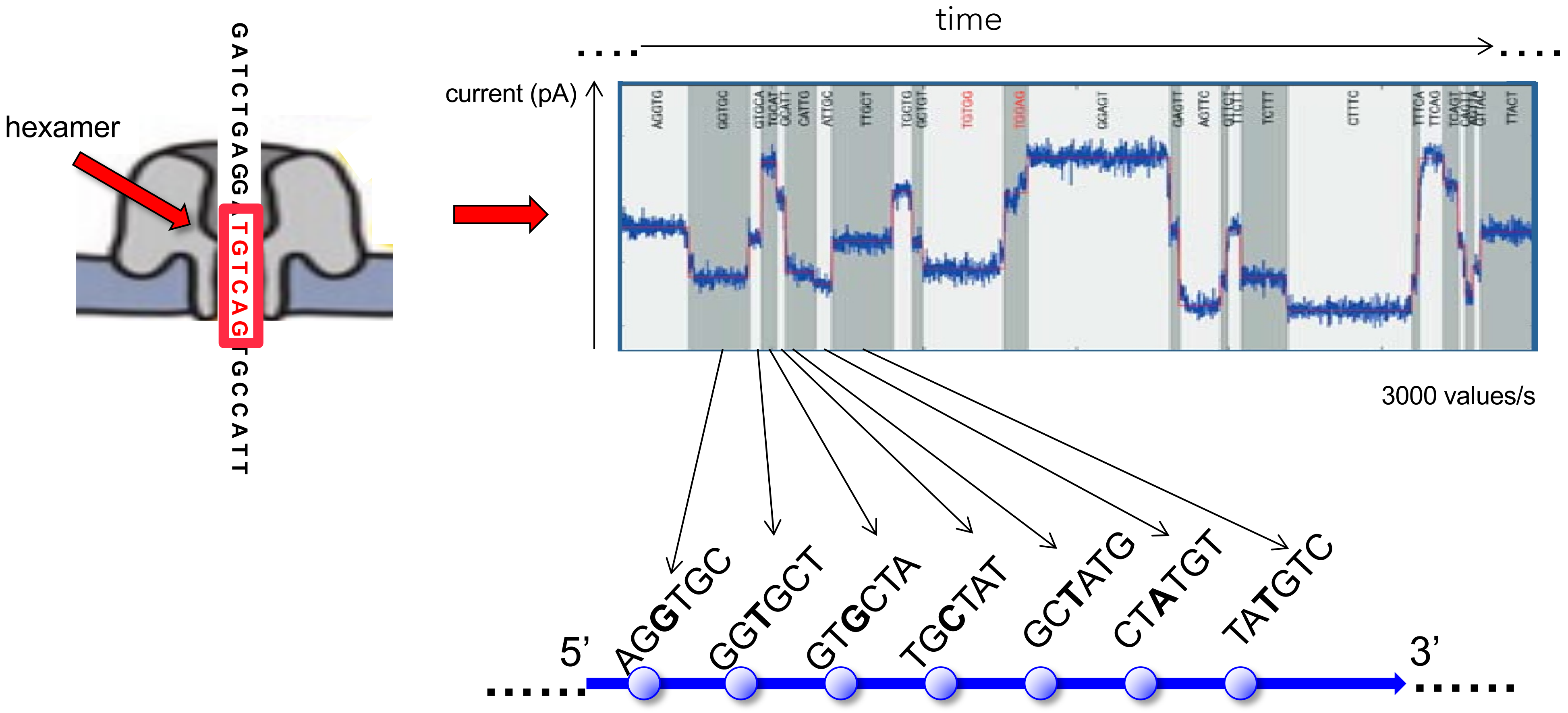
5-6 bases dominate the current signal

MinION : 1 flow cell → 512 pores



PromethION : 1 flow cell → 3000 pores (48 flow cells)

BASE CALLING

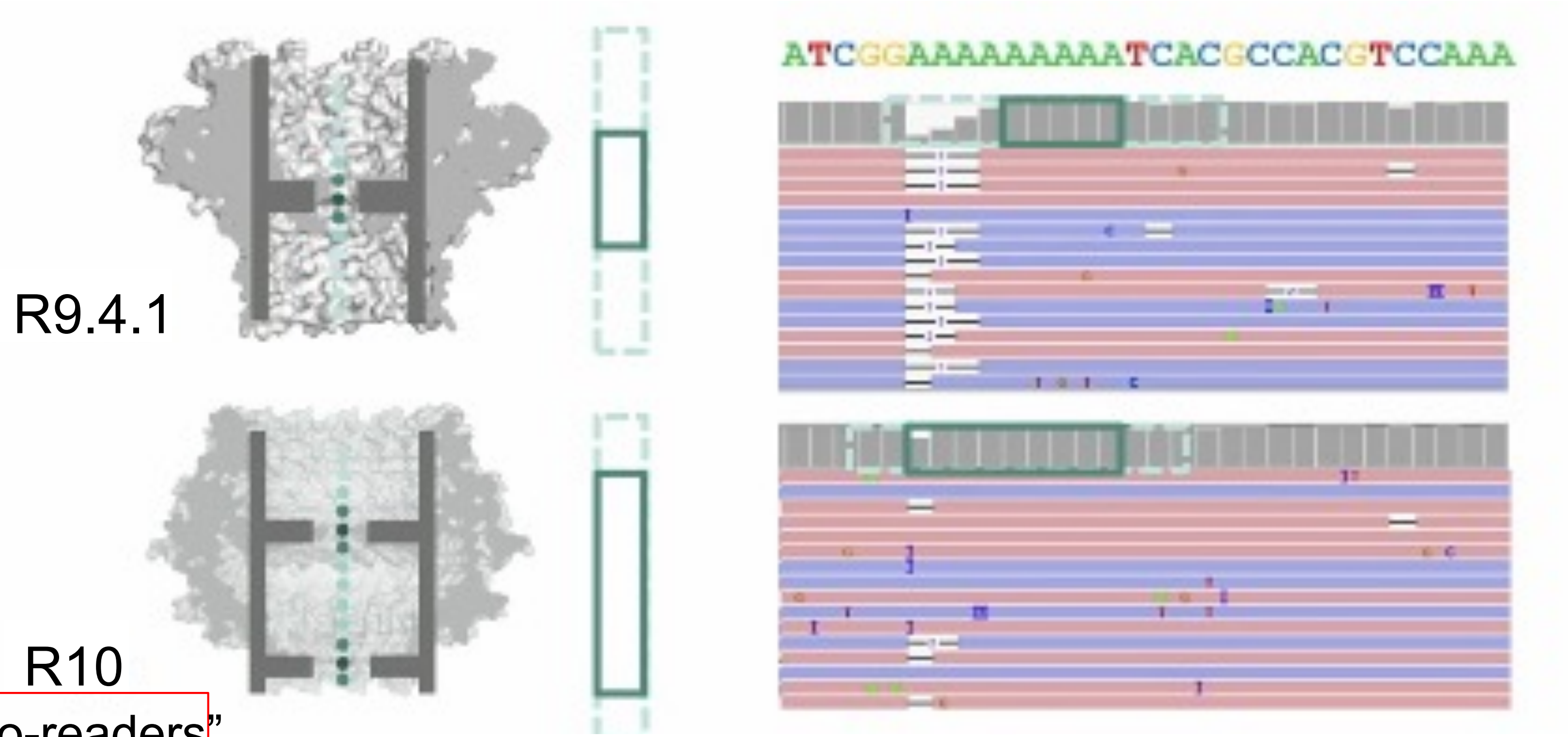


Basecalling : finding the optimal path of successive 6-mers

….. AGGTGCTATGTCT….

”TWO READERS” NANOPORE

“One-reader” pore has difficulty to read homopolymers



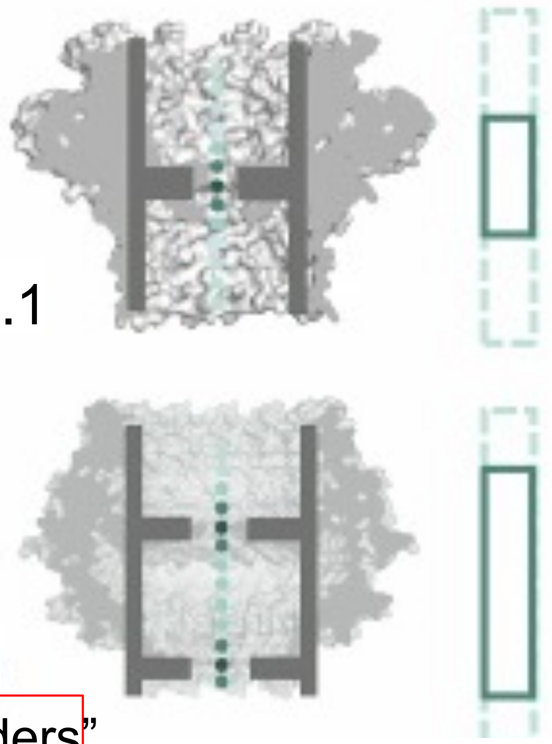
“two-readers”

Sereika et al. *Nature Methods,* 2022

Long homopolymers are better “seen” by the pore and can be decoded with higher accuracy

”TWO READERS” NANOPORE

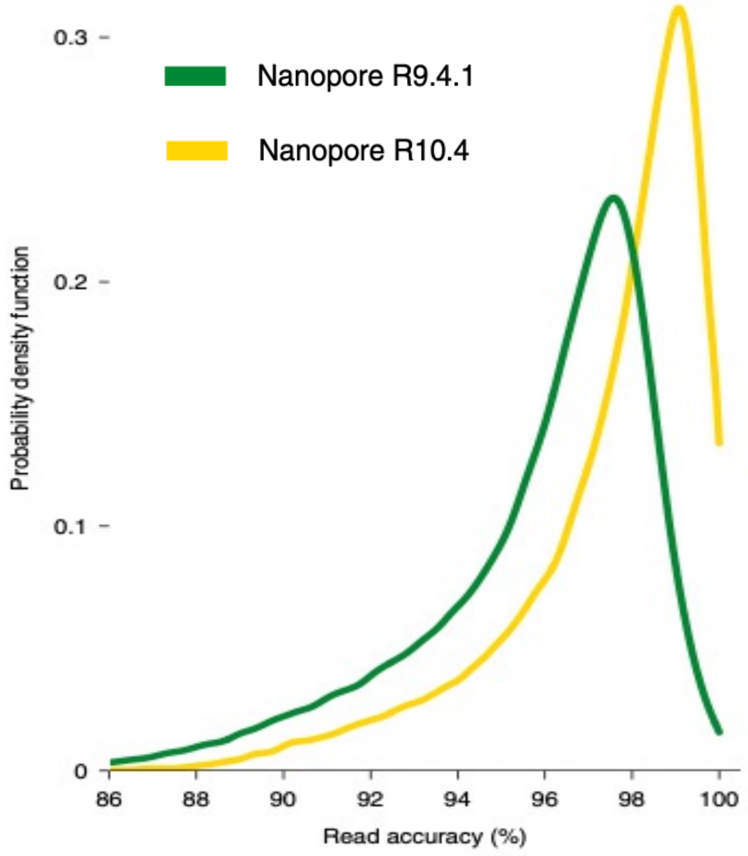
“One-reader” pore has difficulty to read homopolymers



R9.4.1

R10

“two-readers”

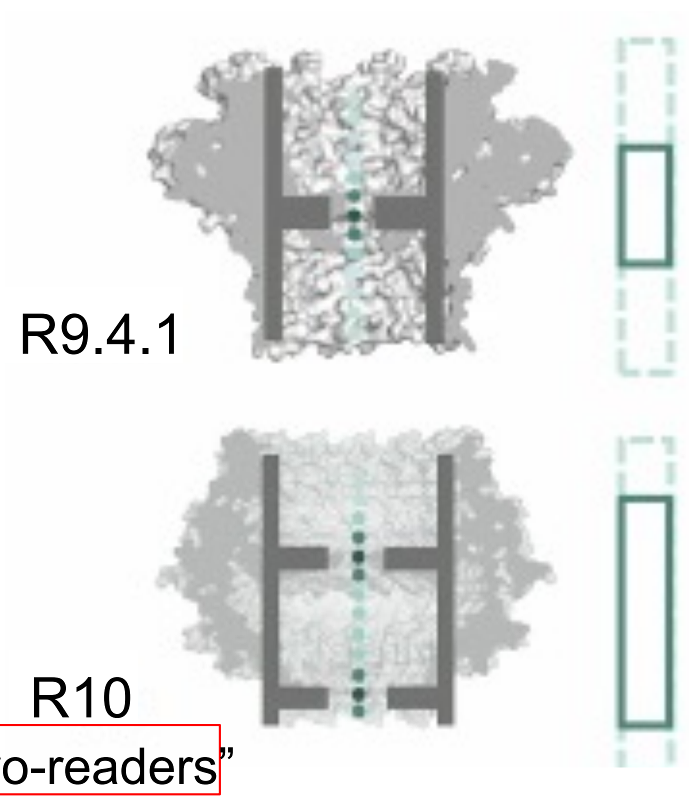


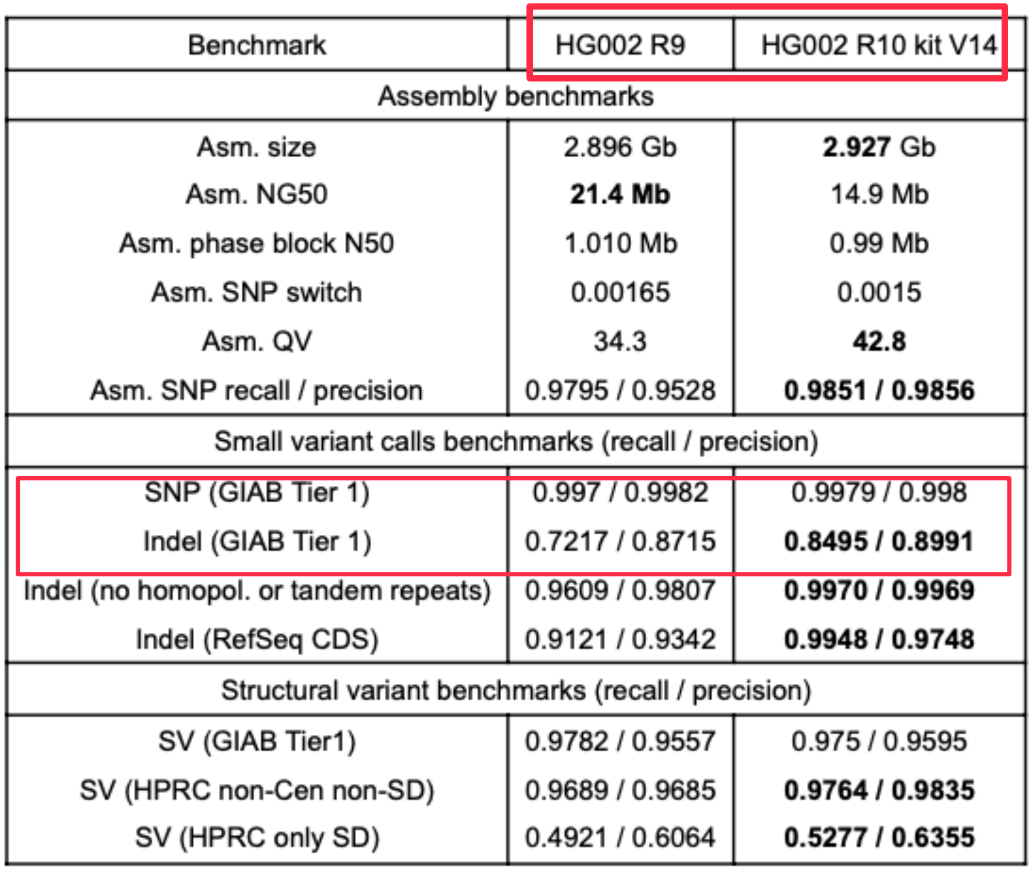
Sereika et al. *Nature Methods,* 2022

Mean accuracy (R10) > 99% ⟶ Q20+

”TWO READERS” NANOPORE

“One-reader” pore has difficulty to read homopolymers



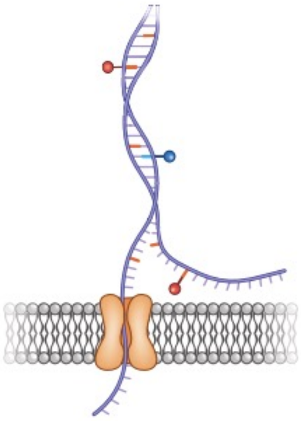


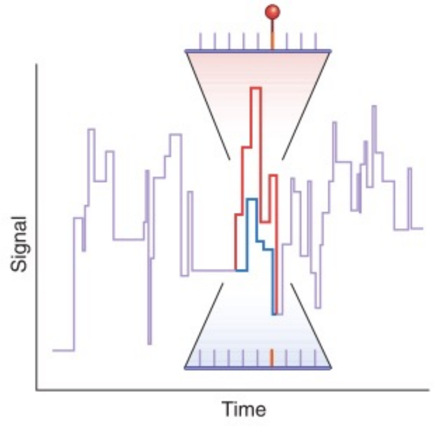
Benchmark	HG002 R9	HG002 R10 kit V14
Assembly benchmarks		
Asm. size	2.896 Gb	**2.927** Gb
Asm. NG50	**21.4 Mb**	14.9 Mb
Asm. phase block N50	1.010 Mb	0.99 Mb
Asm. SNP switch	0.00165	0.0015
Asm. QV	34.3	**42.8**
Asm. SNP recall / precision	0.9795 / 0.9528	**0.9851 / 0.9856**
Small variant calls benchmarks (recall / precision)		
SNP (GIAB Tier 1)	0.997 / 0.9982	0.9979 / 0.998
Indel (GIAB Tier 1)	0.7217 / 0.8715	**0.8495 / 0.8991**
Indel (no homopol. or tandem repeats)	0.9609 / 0.9807	**0.9970 / 0.9969**
Indel (RefSeq CDS)	0.9121 / 0.9342	**0.9948 / 0.9748**
Structural variant benchmarks (recall / precision)		
SV (GIAB Tier1)	0.9782 / 0.9557	0.975 / 0.9595
SV (HPRC non-Cen non-SD)	0.9689 / 0.9685	**0.9764 / 0.9835**
SV (HPRC only SD)	0.4921 / 0.6064	**0.5277 / 0.6355**

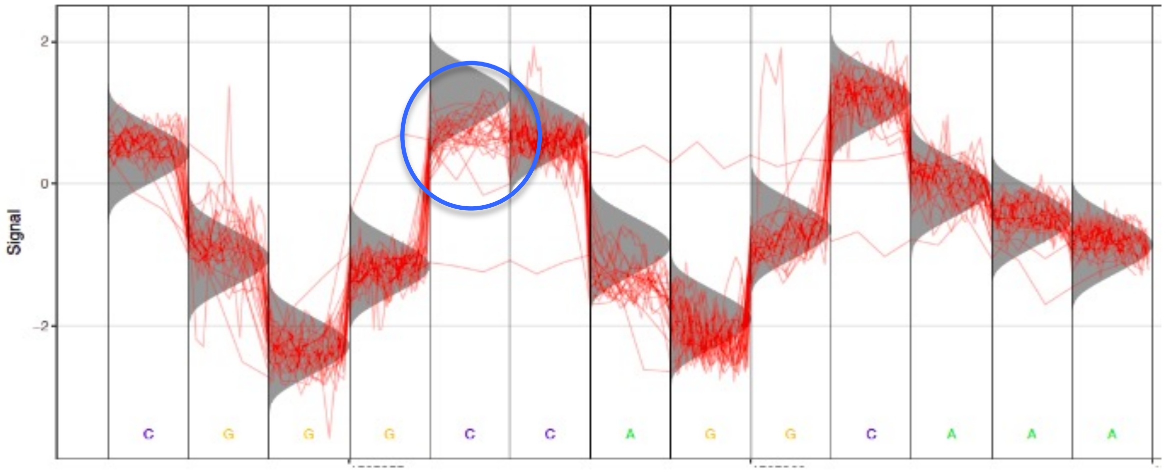
Kolmogorov et al. *bioRxiv.* (2023)

Mean accuracy (R10) > 99% → Q20+

DETECTION OF MODIFIED DNA BASES





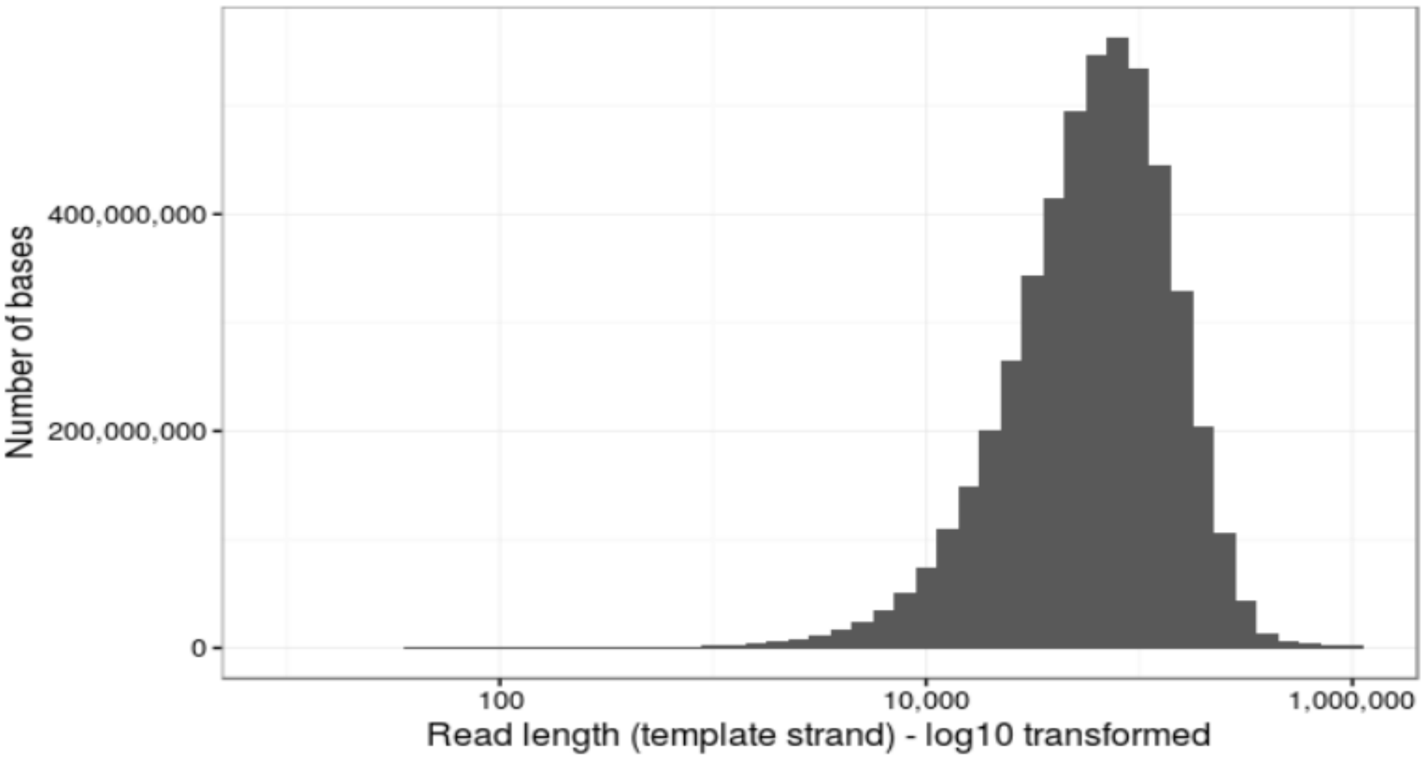


— Electric signal

▶ Canonical base distribution

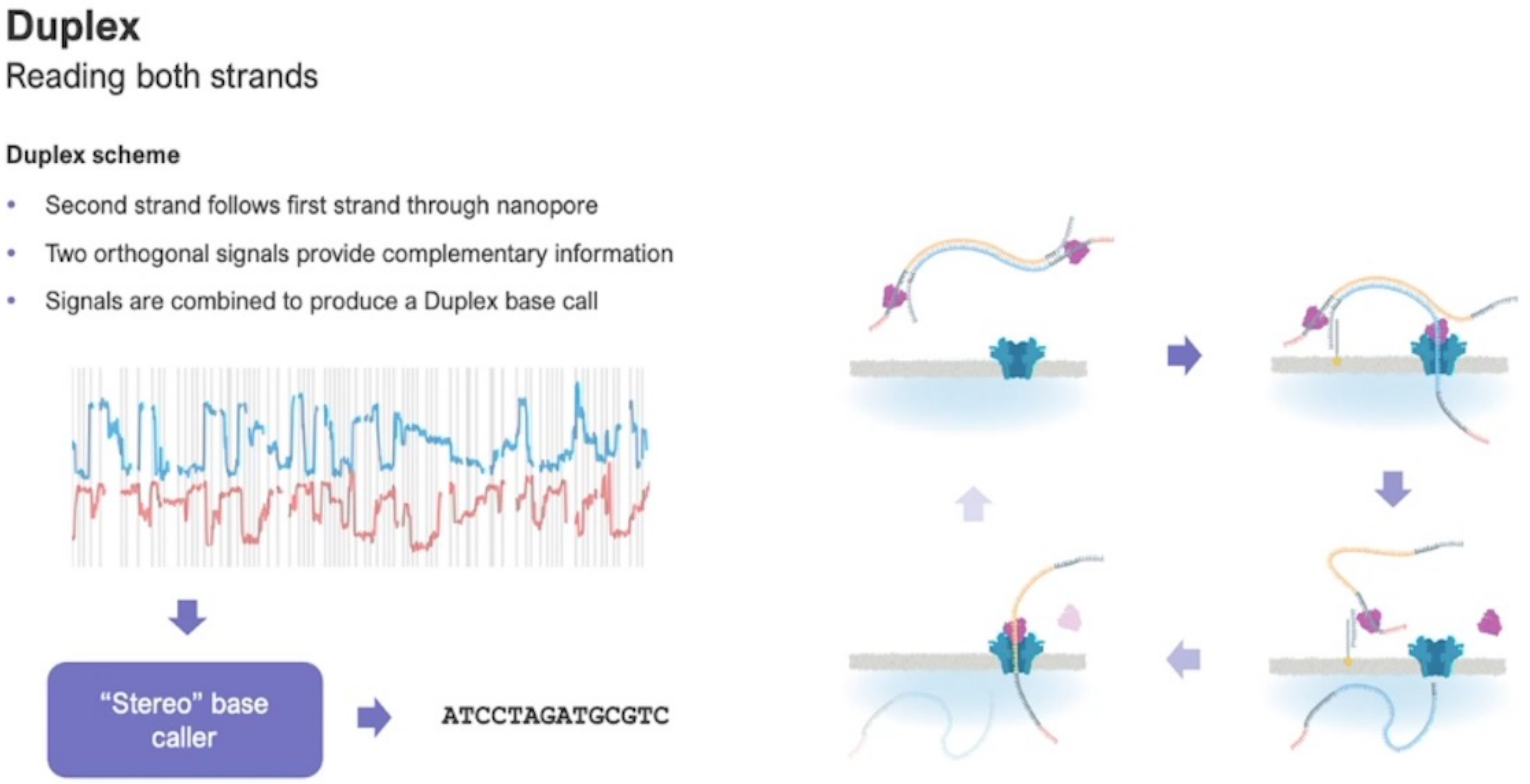
LENGTH OF NANOPORE READS

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



Size of the longest read > 1 Mb

DUPLEX SEQUENCING



Duplex

Reading both strands

Duplex scheme

- Second strand follows first strand through nanopore
- Two orthogonal signals provide complementary information
- Signals are combined to produce a Duplex base call

“Stereo” base caller

ATCCTAGATGCGTC

C. Brown London Calling 2023

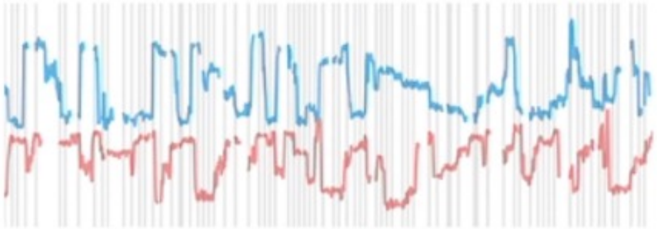
DUPLEX SEQUENCING

Duplex

Reading both strands

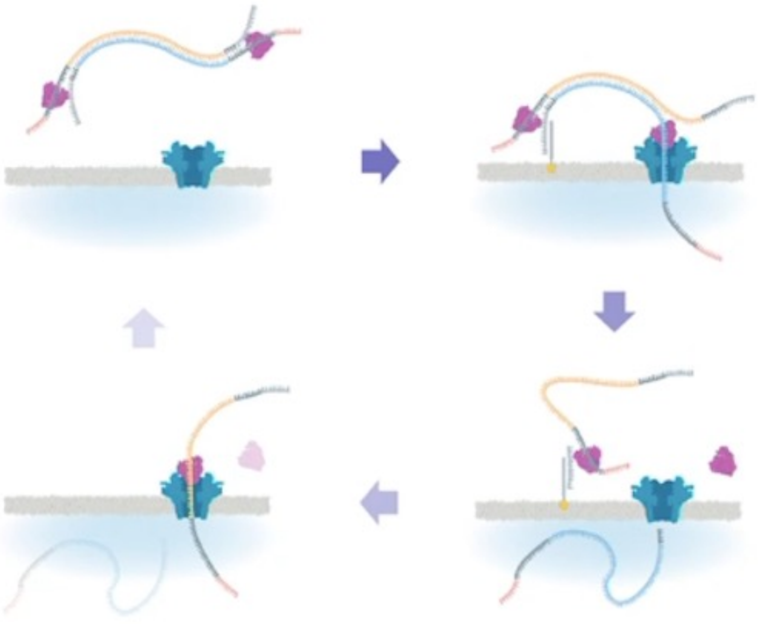
Duplex scheme

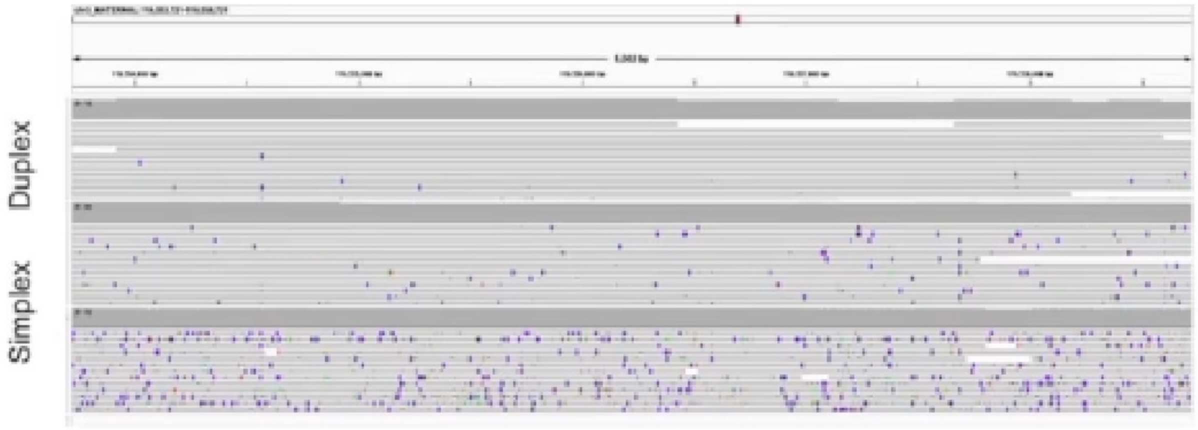
- Second strand follows first strand through nanopore
- Two orthogonal signals provide complementary information
- Signals are combined to produce a Duplex base call



"Stereo" base caller

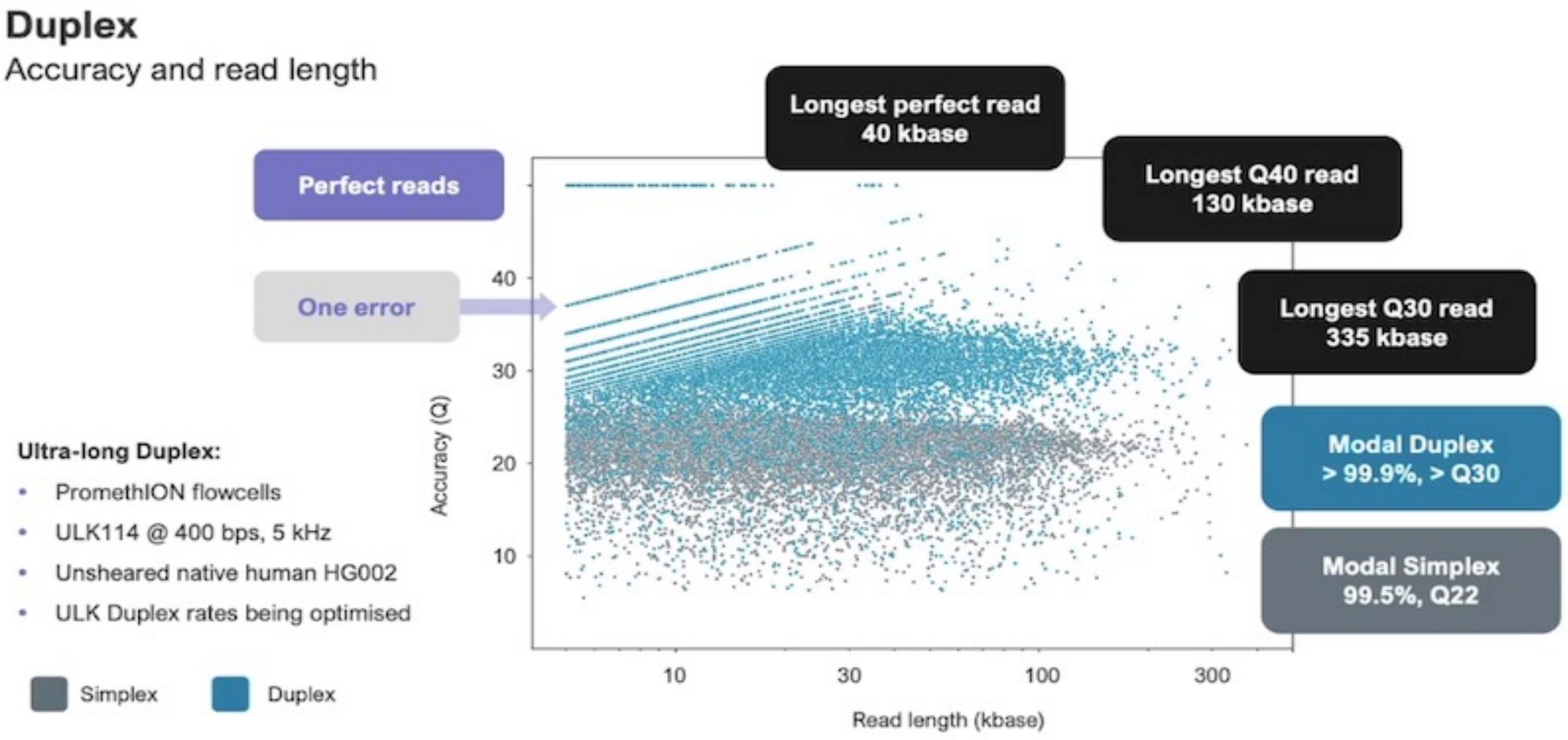
ATCCTAGATGCGTC





C. Brown London Calling 2023

DUPLEX SEQUENCING



Duplex

Accuracy and read length

Perfect reads

One error

Longest perfect read 40 kbase

Longest Q40 read 130 kbase

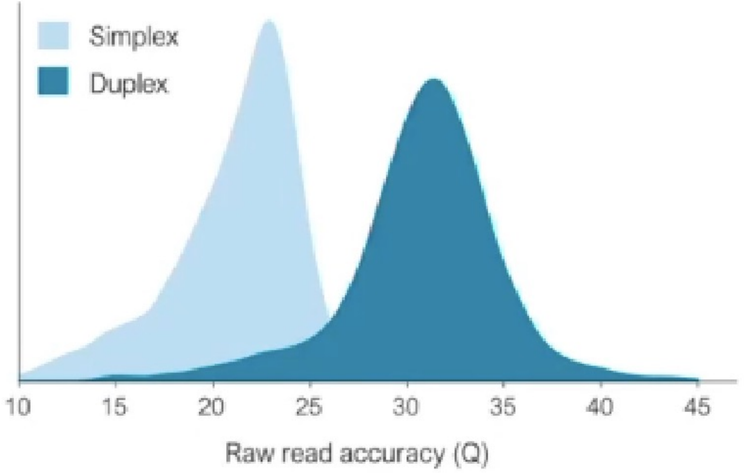
Longest Q30 read 335 kbase

Modal Duplex > 99.9%, > Q30

Modal Simplex 99.5%, Q22

Ultra-long Duplex:

- PromethION flowcells
- ULK114 @ 400 bps, 5 kHz
- Unsheared native human HG002
- ULK Duplex rates being optimised



Duplex mean accuracy > Q30

Duplex outputs

- Outputs rates of Duplex greatly increased recently
- Now achieving **> 50 Gb Duplex** from a single PromethION flowcell

C. Brown London Calling 2023

SMALL GENOMES ASSEMBLY :

NANOPORE VS PACBIO

SMALL GENOMES ASSEMBLY : NANOPORE VS PACBIO

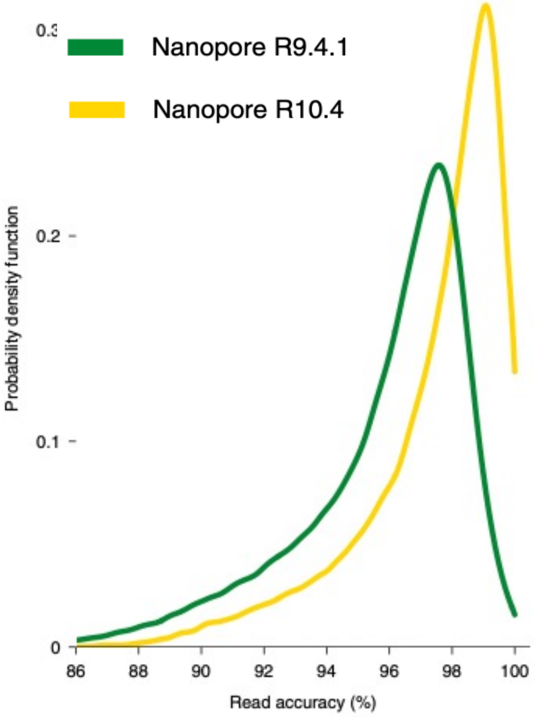
Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing
Sereika et al. *Nature Methods* 2022

- Samples :
  - Seven bacteria
  - *Saccharomyces cerevisiae*
  - Metagenome : anaerobic digester

- Sequenced with :
  - Illumina MiSeq (2 × 300 bp)
  - PacBio Sequel II HiFi
  - Oxford Nanopore R9.4.1 (MinION) and R10.4 (PromethION)

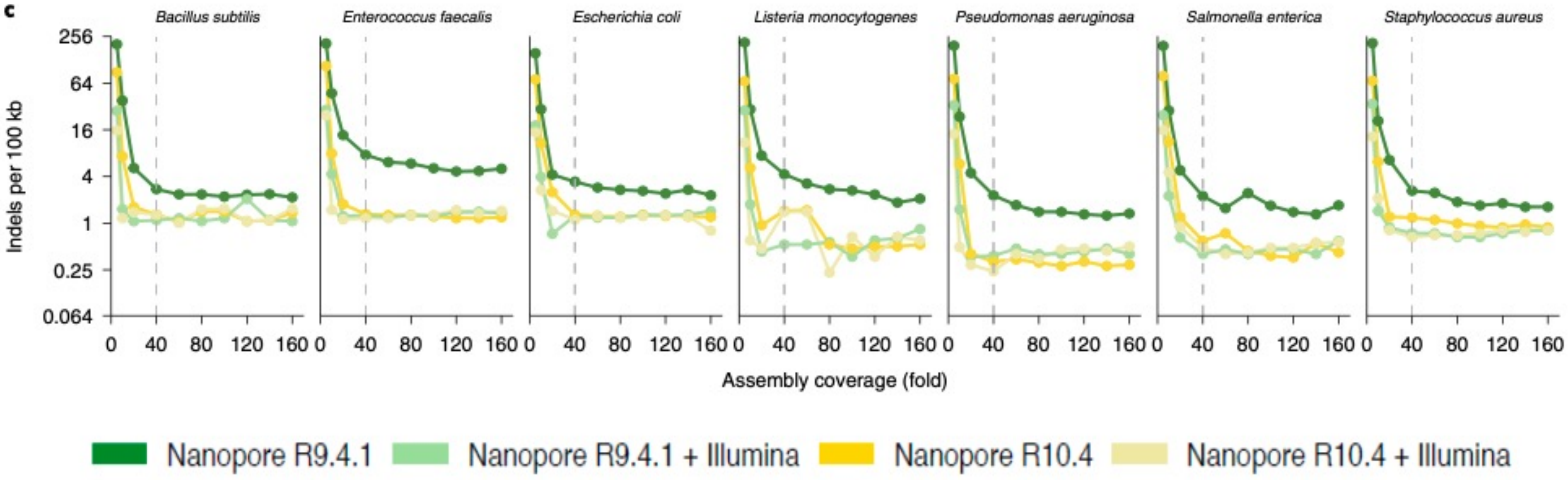
- Read processing
  - reads assembled with Flye

Read accuracies



SMALL GENOMES ASSEMBLY : NANOPORE VS PACBIO

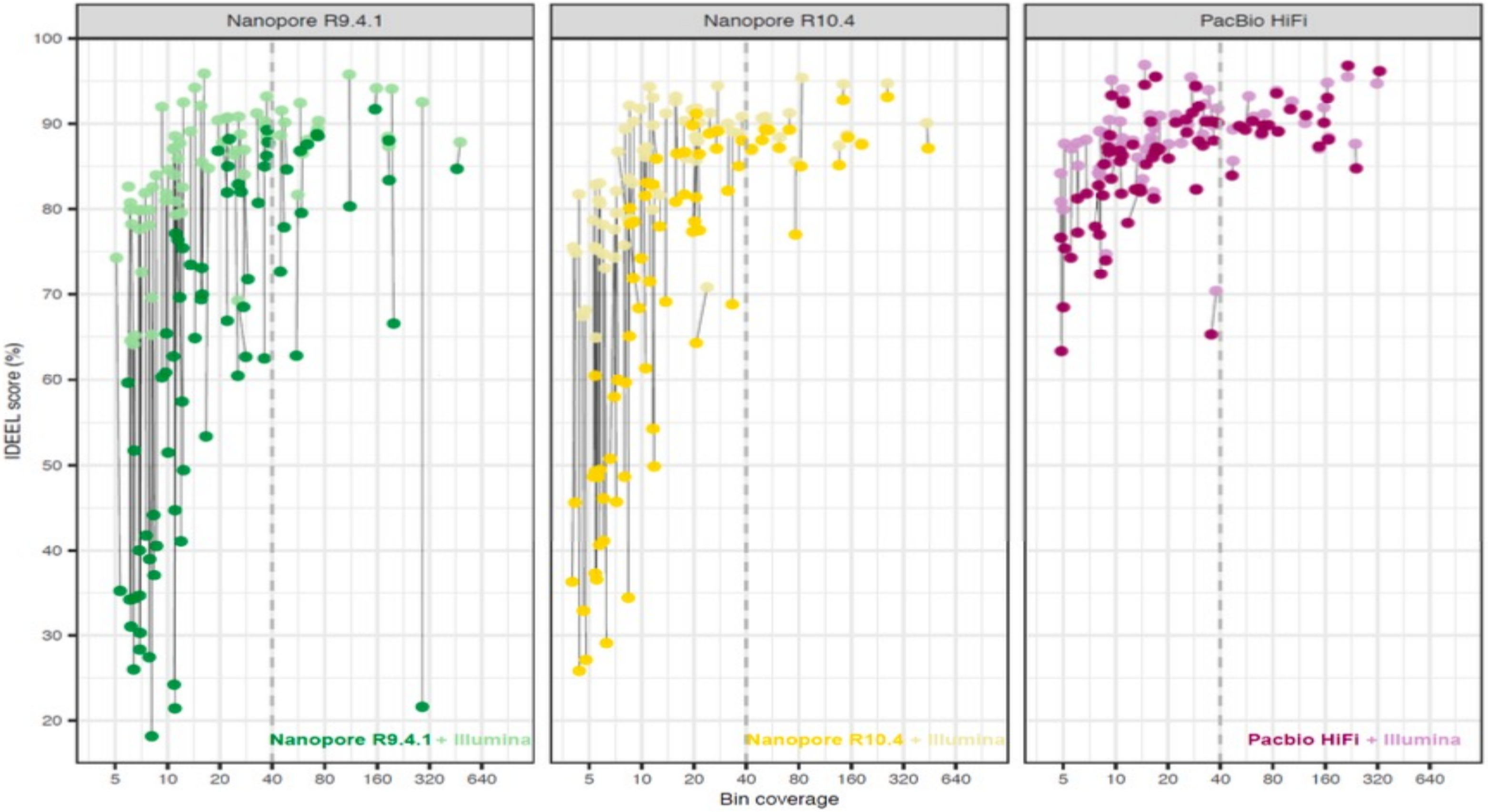
Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing
Sereika et al. *Nature Methods* 2022



SMALL GENOMES ASSEMBLY : NANOPORE VS PACBIO

Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing
Sereika et al. *Nature Methods* 2022

Metagenome-assembled genome (MAG) from the anaerobic digester sample



IDEEL score : proportion of predicted proteins that are ≥95% the length of their best-matching known protein in a database

SMALL GENOMES ASSEMBLY : NANOPORE VS PACBIO

Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing
Sereika et al. *Nature Methods* 2022

Conclusions

- HiFi reads : very low error rate, best genome assembly
- Nanopore reads : the improvement in assembly accuracy from R9.4.1 to R10.4 is largely due to an improved ability to call homopolymers
- No significant improvement for R10.4 by the addition of Illumina polishing
- -> Near-finished microbial reference genomes can be obtained from R10.4 data alone at a coverage of approximately 40-fold
- ONT more cost-effective than PacBio

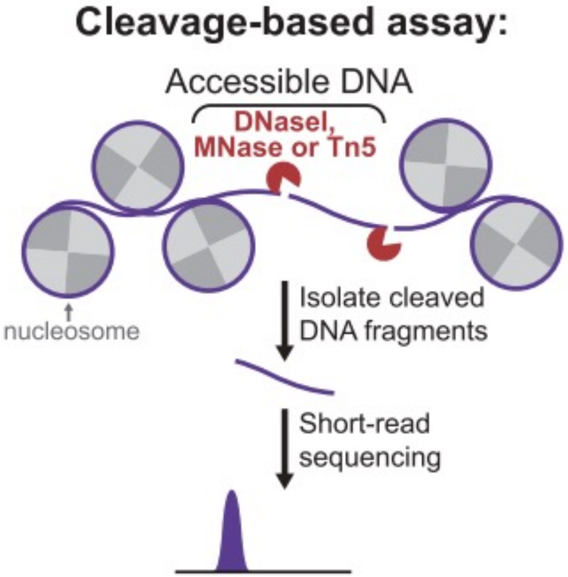
DNA MODIFICATIONS

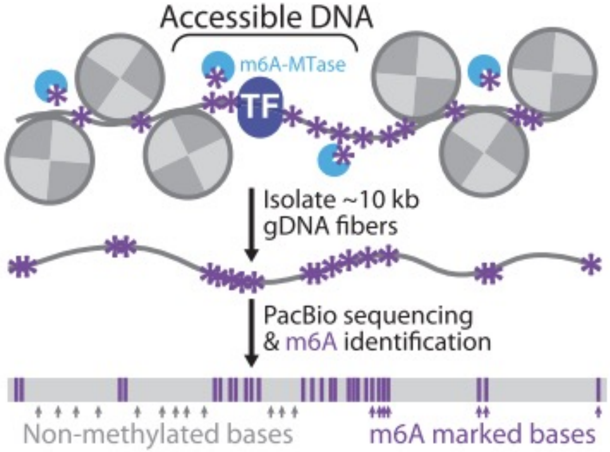
DETECTION OF DNA m6A WITH CCS

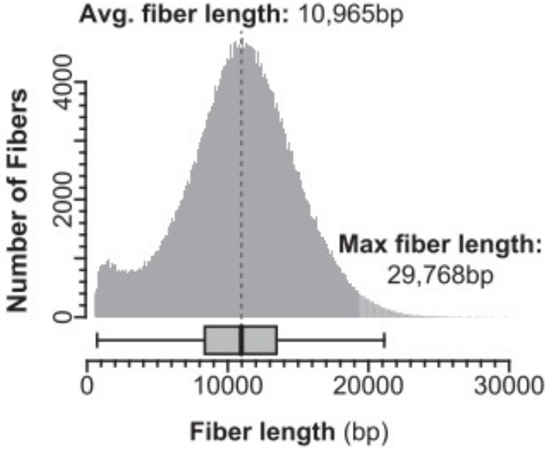
Single-molecule regulatory architectures captured by chromatin fiber sequencing
Stergachis et al. *Science* (2020)

DnaseI-seq.

Fiber-seq.

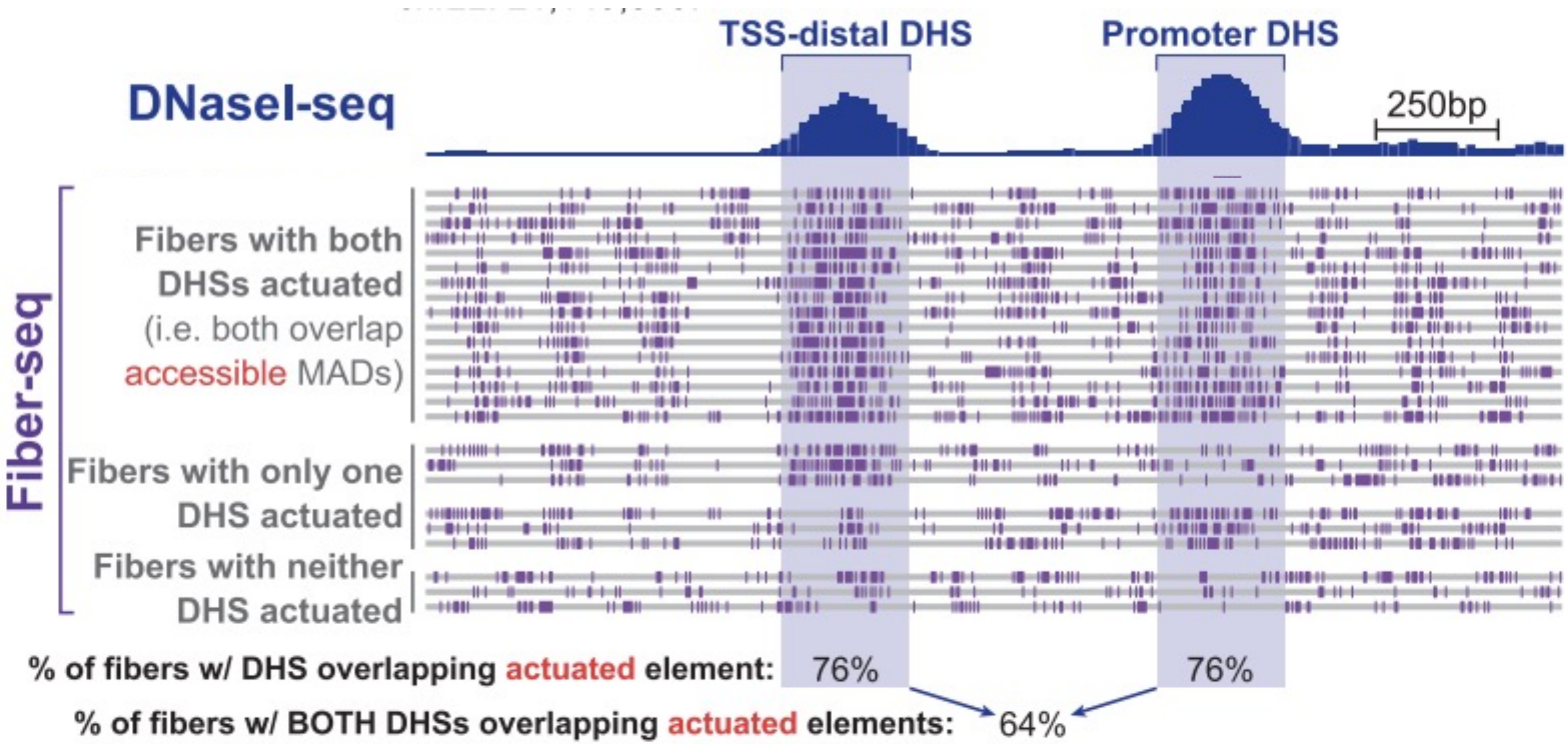



Accessible DNA
m6A-MTase
TF
Isolate ~10 kb gDNA fibers
PacBio sequencing & m6A identification
Non-methylated bases
m6A marked bases



DETECTION OF DNA m6A WITH CCS

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



DETECTION OF DNA 5mCpG WITH NANOPORE

Robust methylation-based classification of brain tumours using nanopore sequencing
Kuschel et al. *Neuropathol Appl Neurobiol.* 2023

DNA methylation profiling (5mC) of human brain tumours ➡ profound impact on clinical neuro-oncology

Hybridisation microarrays :

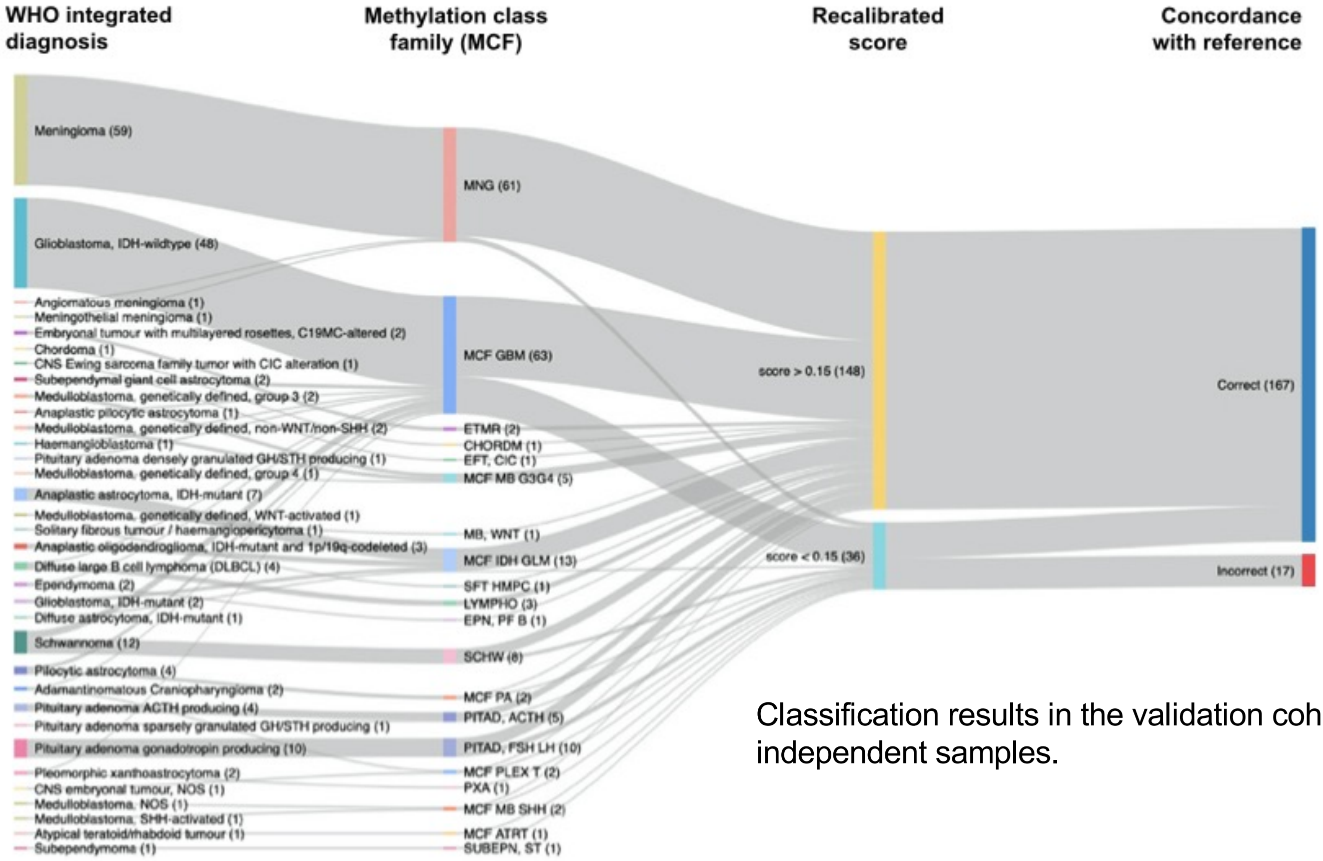
- time consuming
- costly

Nanopore genome sequencing (R9.4.1 flow cell) :

- 382 tissue samples
- 46 brain tumour (sub)types
- Bootstrap sampling in a cohort of 55 cases :
  - classification by ad hoc random forests
  - sensitivity 80.4%

DETECTION OF DNA 5mCpG WITH NANOPORE

Robust methylation-based classification of brain tumours using nanopore sequencing
Kuschel et al. *Neuropathol Appl Neurobiol.* 2023



Classification results in the validation cohort of N = 184 independent samples.

CONCLUSION

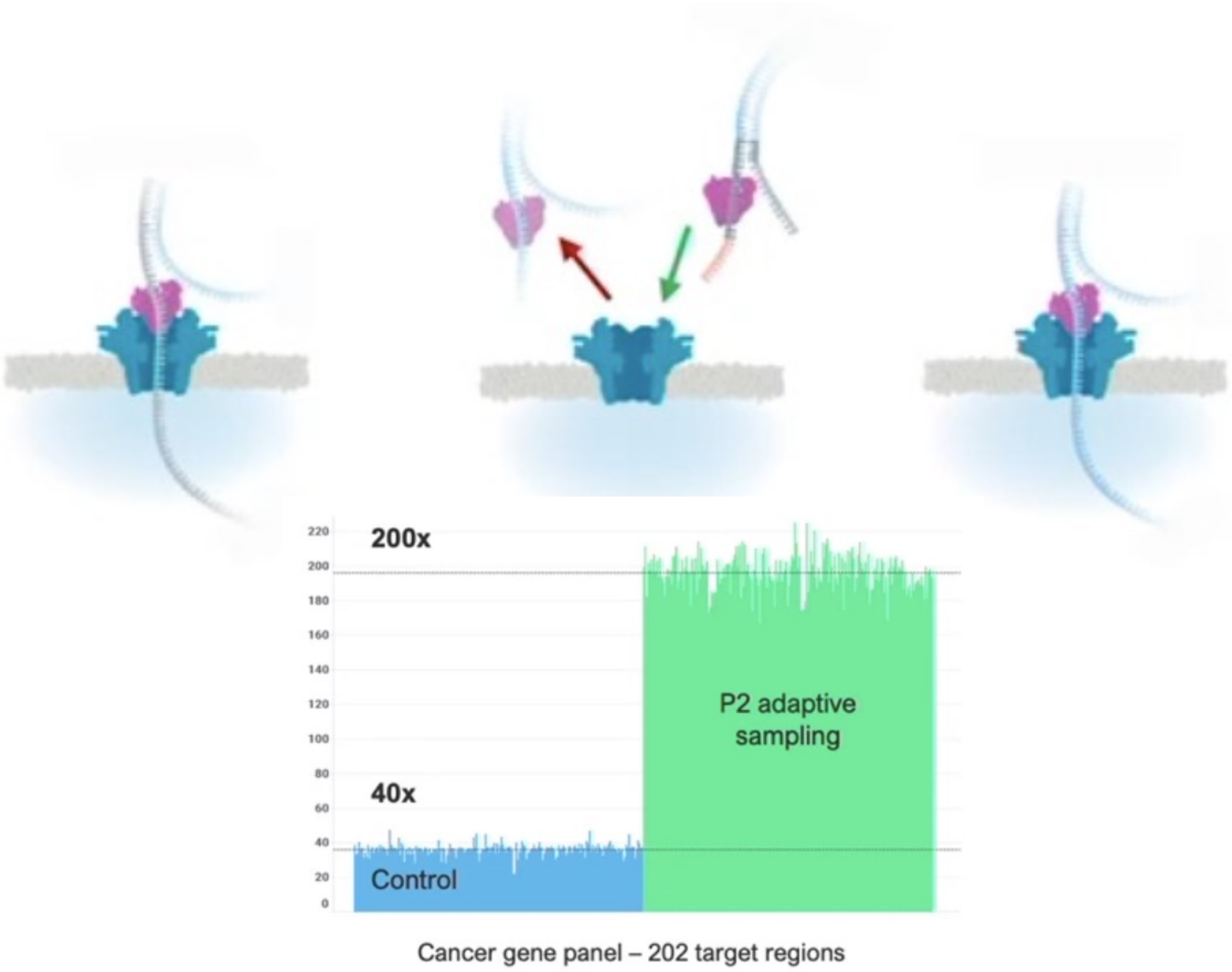
Nanopore sequencing ➡ DNA methylation-based classification in brain tumour diagnostics :

- rapid and cost-effective
- shorten the time to diagnosis
- augment neuropathological decision making
- improve diagnostic precision

TARGETED SEQUENCING

NANOPORE ADAPTIVE SAMPLING

- Specification of target regions
- Real time basecalling
- Mapping of ~ 500 first bases
- Before the molecule is fully sequenced : If it differs from target -> reversion of polarity and ejection



NANOPORE ADAPTIVE SAMPLING

Adaptive nanopore sequencing to determine pathogenicity of *BRCA1* exonic duplication
Filser et al. *J. Med. Genet.* Jun. 2023

Patient with a breast tumor : Initial molecular analysis

- germline DNA extracted from blood cells -> sequenced with Illumina
- NGS panel (HBOC) -> duplication encompassing *BRCA1* exons 18–20
- But :
  - NGS data could not demonstrate that reading frame of BRCA1 transcript was altered
  - ie, that the event was a tandem duplication
  - ➡ further cDNA analysis required to confirm pathogenicity
  - but RNA is not routinely available
  - and the technique is very time-consuming (~2 months for analysis)

⬇

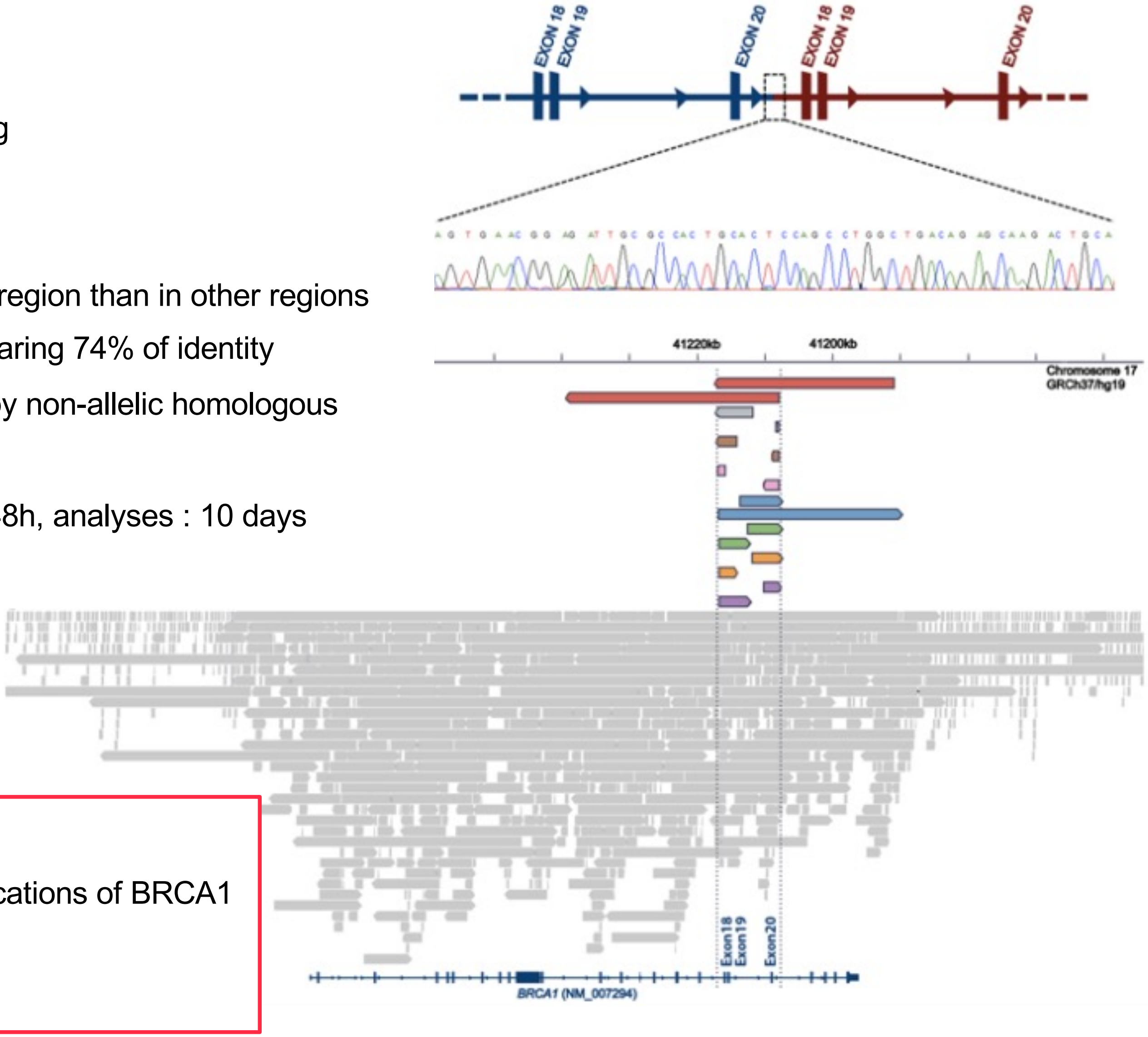
Decision of Nanopore sequencing with adaptive sampling

NANOPORE ADAPTIVE SAMPLING

Adaptive nanopore sequencing to determine pathogenicity of *BRCA1* exonic duplication
Filser et al. *J. Med. Genet.* Jun. 2023

Nanopore sequencing with adaptive sampling

- Depth of coverage: 24x
- 10 times higher in the targeted genomic region than in other regions
- SV breakpoints located in two Alu RE sharing 74% of identity
- -> supports that this SV was mediated by non-allelic homologous recombination
- Fast (library preparation - sequencing : 48h, analyses : 10 days

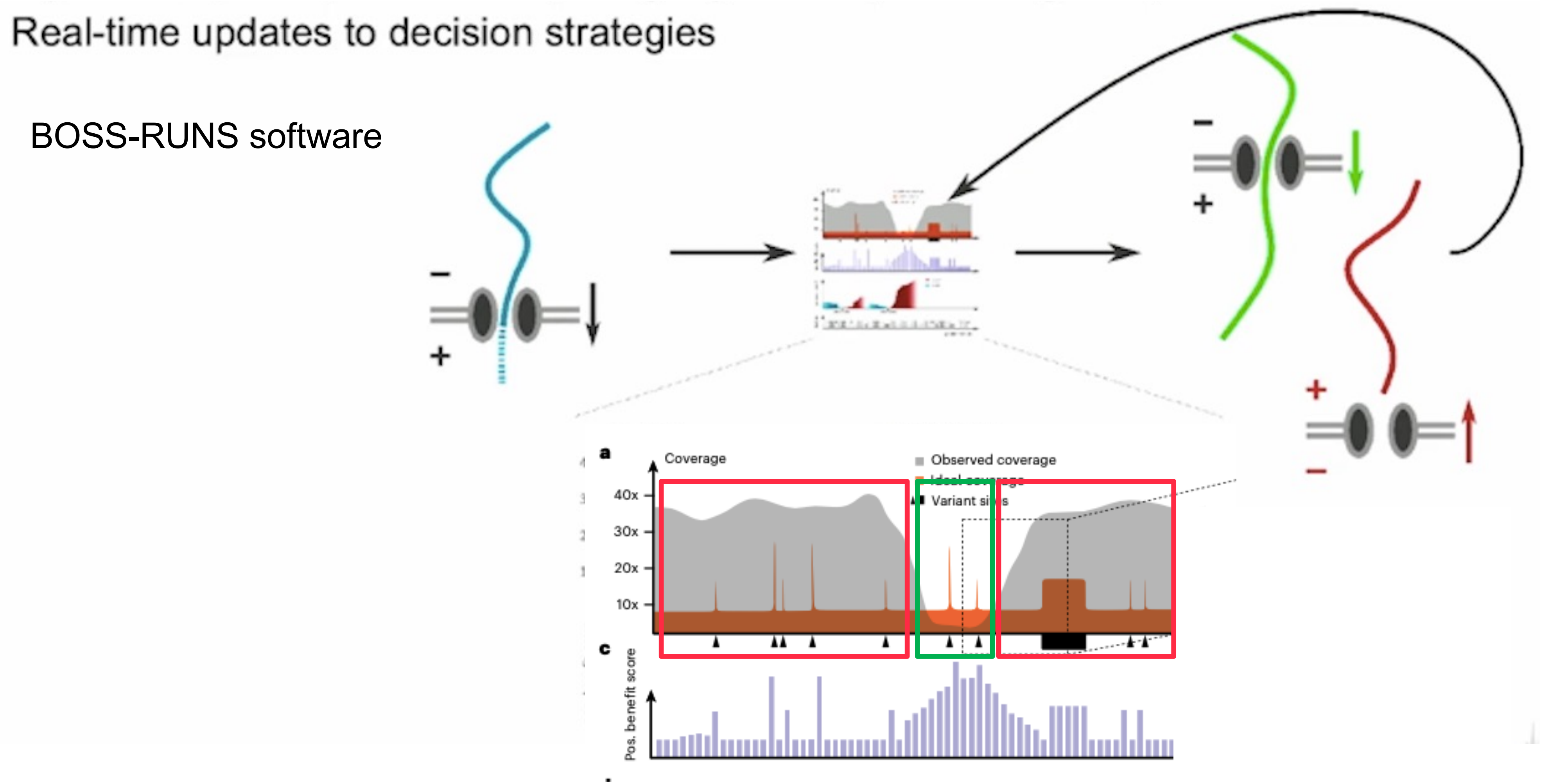


Conclusions

- Accurate resolution of an intragenic duplications of BRCA1
- Classification as a pathogenic variant
- Ultimately guiding the clinician’s decision

NANOPORE ADAPTIVE SAMPLING

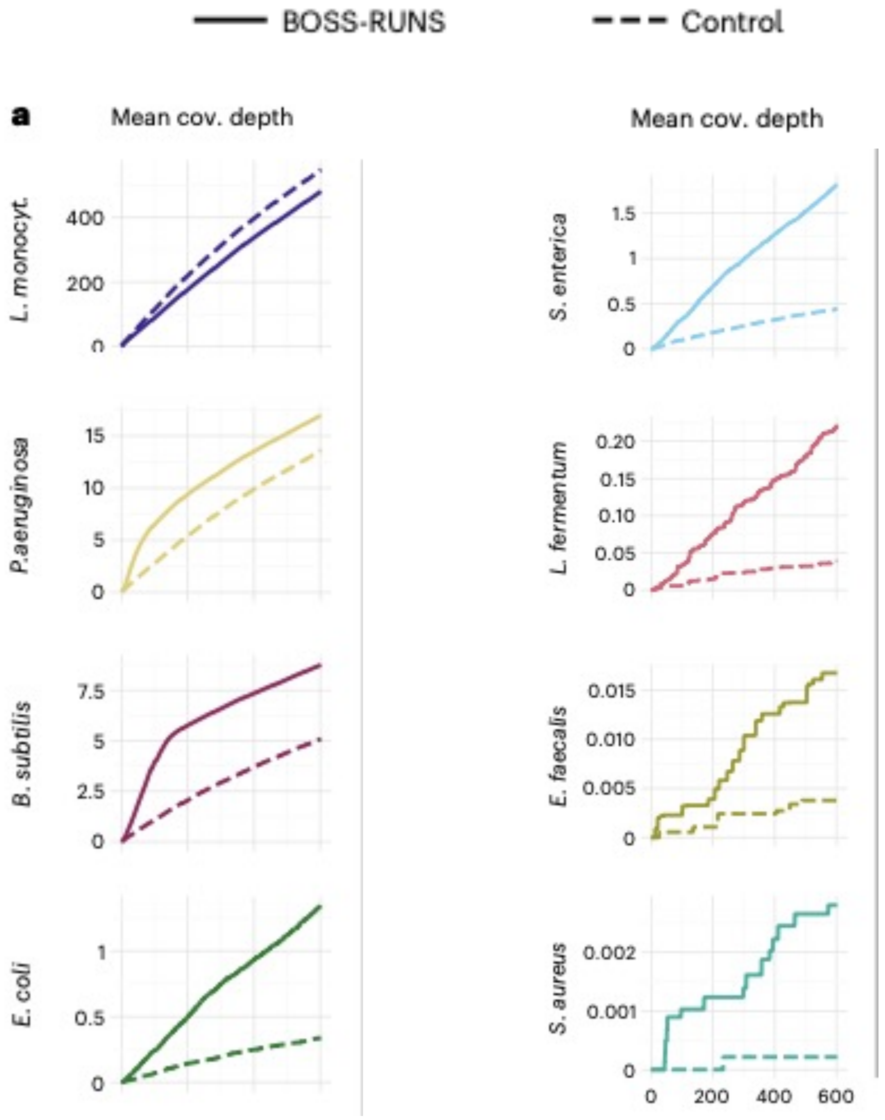
Dynamic, adaptive sampling during nanopore sequencing using Bayesian experimental design
Weilguny et al. *Nature Biotechnology* Jan. 2023



BOSS-RUNS software

NANOPORE ADAPTIVE SAMPLING

Dynamic, adaptive sampling during nanopore sequencing using Bayesian experimental design
Weilguny et al. *Nature Biotechnology* Jan. 2023



NANOPORE ADAPTIVE SAMPLING

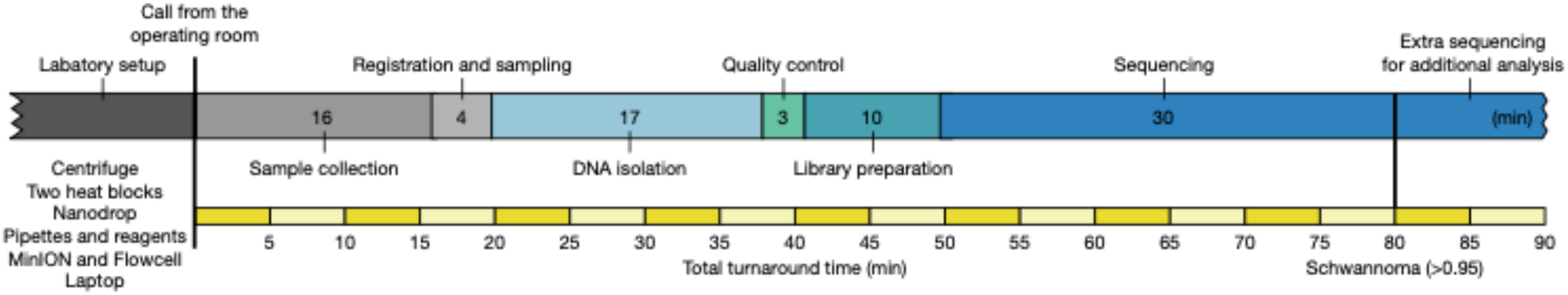
Ultra-fast deep-learned CNS tumour classification during surgery
Vermeulen et al. *Nature Oct.* 2023

Using nanopore adaptive sampling to obtain a methylation profile (5mCpG sites) during surgery :

- Development of Sturgeon software
- patient-agnostic transfer-learned neural network
- enables molecular subclassification of central nervous system tumours based on such profiles

Sturgeon delivered :

- Diagnosis within 40 minutes after starting sequencing
- Diagnostic turnaround time of less than 90 min
- Accurate diagnosis in 45 out of 50 retrospectively sequenced samples
- Applicability in real time during 25 surgeries
- Of these, 18 (72%) diagnoses were correct



LARGE GENOME ASSEMBLY

VERY BRIEF SUMMARY OF HUMAN GENOME ASSEMBLY

- 2001: Celera Genomics and International Human Genome Sequencing Consortium :
  - initial drafts of the human genome
- But many complex regions were left unfinished or incorrectly assembled for over 20 years :
  - they represent 8% of the genome

↓

T2T : telomere to telomere assembly: largest addition of new content to human genome in the past 20 years

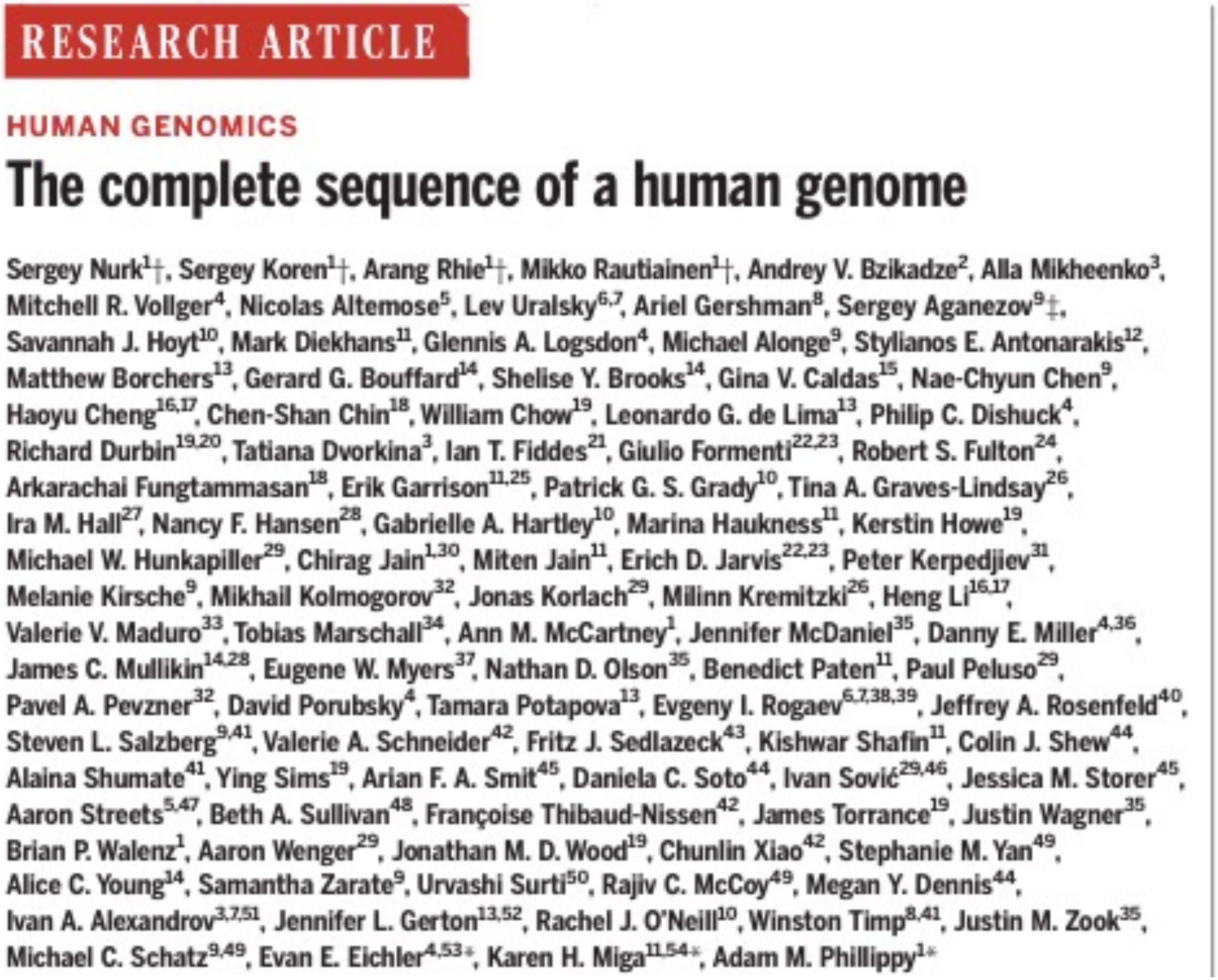
1 - The complete sequence of a human genome
Nurk et al. *Science* 2022

2 - Chasing perfection: validation and polishing strategies for telomere-to-telomere genome assemblies
Mc Cartney et al. *Nature Methods* 2022

3 - The complete sequence of a human Y chromosome
Rhie et al. *Nature* Sept. 2023

LARGE GENOME ASSEMBLY

The complete sequence of a human genome
Nurk et al. *Science* 2022



RESEARCH ARTICLE

HUMAN GENOMICS

The complete sequence of a human genome

Sergey Nurk[1]†, Sergey Koren[1]†, Arang Rhie[1]†, Mikko Rautiainen[1]†, Andrey V. Bzikadze[2], Alla Mikheenko[3], Mitchell R. Vollger[4], Nicolas Altemose[5], Lev Uralsky[6,7], Ariel Gershman[8], Sergey Aganezov[9]‡, Savannah J. Hoyt[10], Mark Diekhans[11], Glennis A. Logsdon[4], Michael Alonge[9], Stylianos E. Antonarakis[12], Matthew Borchers[13], Gerard G. Bouffard[14], Shelise Y. Brooks[14], Gina V. Caldas[15], Nae-Chyun Chen[9], Haoyu Cheng[16,17], Chen-Shan Chin[18], William Chow[19], Leonardo G. de Lima[13], Philip C. Dishuck[4], Richard Durbin[19,20], Tatiana Dvorkina[3], Ian T. Fiddes[21], Giulio Formenti[22,23], Robert S. Fulton[24], Arkarachai Fungtammasan[18], Erik Garrison[11,25], Patrick G. S. Grady[10], Tina A. Graves-Lindsay[26], Ira M. Hall[27], Nancy F. Hansen[28], Gabrielle A. Hartley[10], Marina Haukness[11], Kerstin Howe[19], Michael W. Hunkapiller[29], Chirag Jain[1,30], Miten Jain[11], Erich D. Jarvis[22,23], Peter Kerpedjiev[31], Melanie Kirsche[9], Mikhail Kolmogorov[32], Jonas Korlach[29], Milinn Kremitzki[26], Heng Li[16,17], Valerie V. Maduro[33], Tobias Marschall[34], Ann M. McCartney[1], Jennifer McDaniel[35], Danny E. Miller[4,36], James C. Mullikin[14,28], Eugene W. Myers[37], Nathan D. Olson[35], Benedict Paten[11], Paul Peluso[29], Pavel A. Pevzner[32], David Porubsky[4], Tamara Potapova[13], Evgeny I. Rogaev[6,7,38,39], Jeffrey A. Rosenfeld[40], Steven L. Salzberg[9,41], Valerie A. Schneider[42], Fritz J. Sedlazeck[43], Kishwar Shafin[11], Colin J. Shew[44], Alaina Shumate[41], Ying Sims[19], Arian F. A. Smit[45], Daniela C. Soto[44], Ivan Sović[29,46], Jessica M. Storer[45], Aaron Streets[5,47], Beth A. Sullivan[48], Françoise Thibaud-Nissen[42], James Torrance[19], Justin Wagner[35], Brian P. Walenz[1], Aaron Wenger[29], Jonathan M. D. Wood[19], Chunlin Xiao[42], Stephanie M. Yan[49], Alice C. Young[14], Samantha Zarate[9], Urvashi Surti[50], Rajiv C. McCoy[49], Megan Y. Dennis[44], Ivan A. Alexandrov[3,7,51], Jennifer L. Gerton[13,52], Rachel J. O'Neill[10], Winston Timp[8,41], Justin M. Zook[35], Michael C. Schatz[9,49], Evan E. Eichler[4,53]*, Karen H. Miga[11,54]*, Adam M. Phillippy[1]*

The complete sequence of a human genome
Nurk et al. *Science* 2022

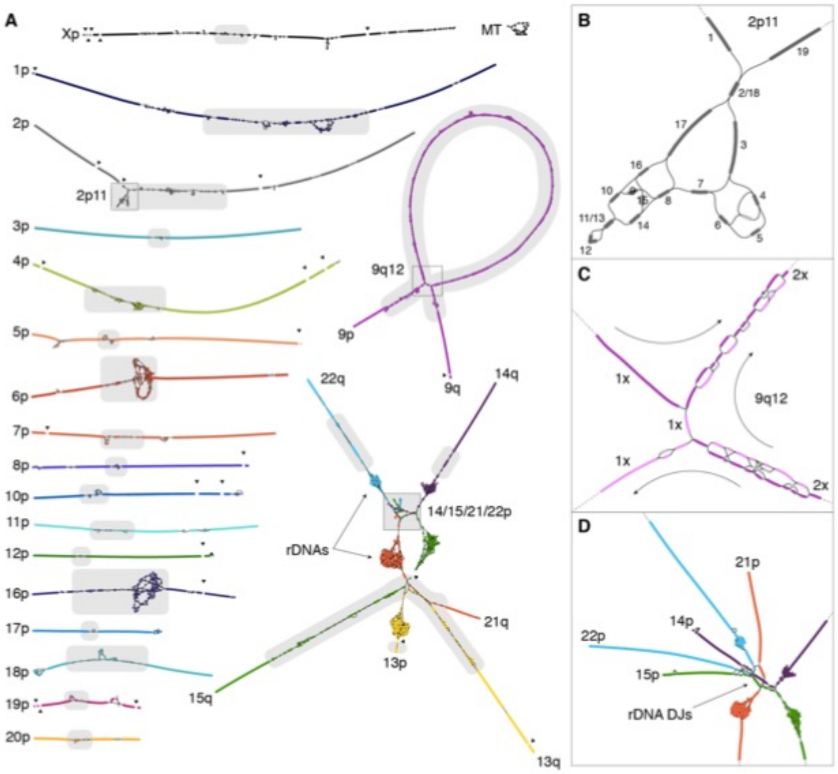
SEQUENCING

Data were obtained with a “complete hydatidiform mole” (CHM13) cell line (homozygous with a 46,XX karyotype) :

- 30× PacBio HiFi
- 120× Nanopore ultra-long read
- BioNano optical maps
- 70× Hi-C
- 100× Illumina PCR-Free sequencing

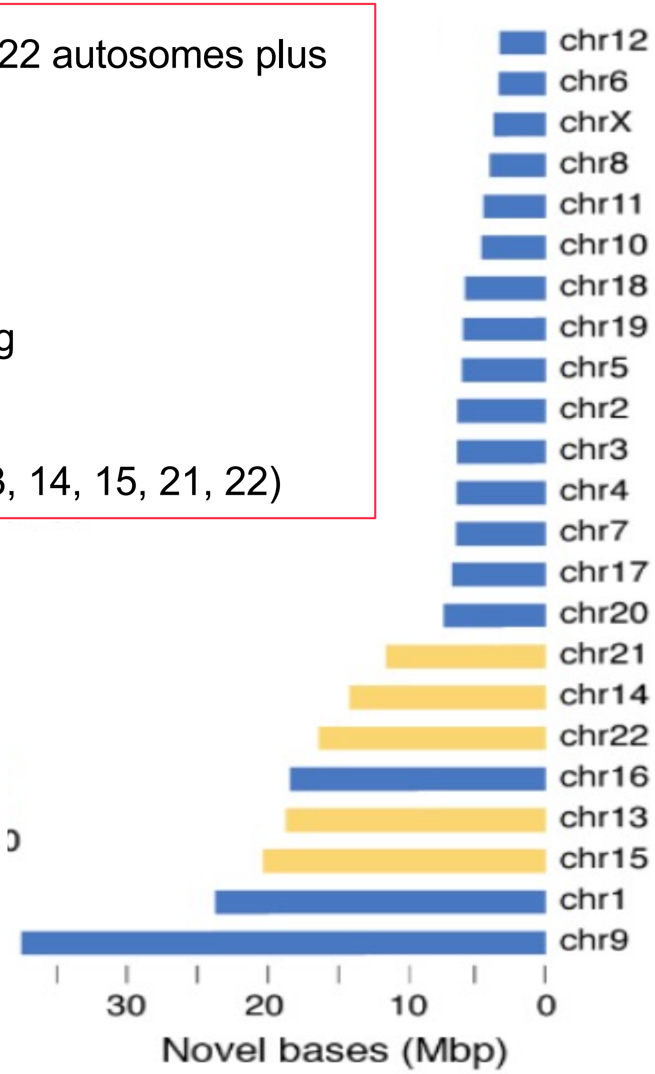
WHOLE GENOME ASSEMBLY

1. HiFi-based graph construction
2. ONT-based tangle resolution
3. Gap filling
4. Polishing



The complete sequence of a human genome
Nurk et al. *Science* 2022

- 8% of the genome completed by this T2T assembly including all 22 autosomes plus Chromosome X :
  - Corrects numerous errors
  - Introduces **200 million bp of novel sequence** containing :
    - 1956 gene predictions, 99 predicted as protein coding
    - all centromeric regions
    - entire short arms (p) of acrocentric chromosomes (13, 14, 15, 21, 22)



LARGE GENOME ASSEMBLY

Chasing perfection: validation and polishing strategies for telomere-to-telomere genome assemblies

Mc Cartney et al. *Nature Methods* 2022

Recent Telomere-to-Telomere (T2T) human genome assembly

- this assembly has evidence of small errors and structural misassemblies
- polishing strategy :
  - ✓ Make corrections in large repeats without over-correction
  - ✓ Ultimately fixing 51% of errors and improving the assembly QV to 73.9
  - ✓ show sequencing biases in PacBio HiFi and ONT reads that cause errors that can be correcte

- 1,457 corrections :
  - ✓ replacing a total of 12,234,603 bp with 10,152,653 bp
  - ✓ ultimately leading to the first complete human genome ever assembled

LARGE GENOME ASSEMBLY

The complete sequence of a human Y chromosome
Rhie et al. *Nature* Sept. 2023

- HG002 diploid genome
- PacBio HiFi reads (60 × haploid genome coverage)
- ONT ultralong reads (90 × in reads > 100 kb)

↓

- last chromosome completed from telomere to telomere
- Addition of T2T-Y with previous assembly of the CHM13 genome

↓

T2T-chm13

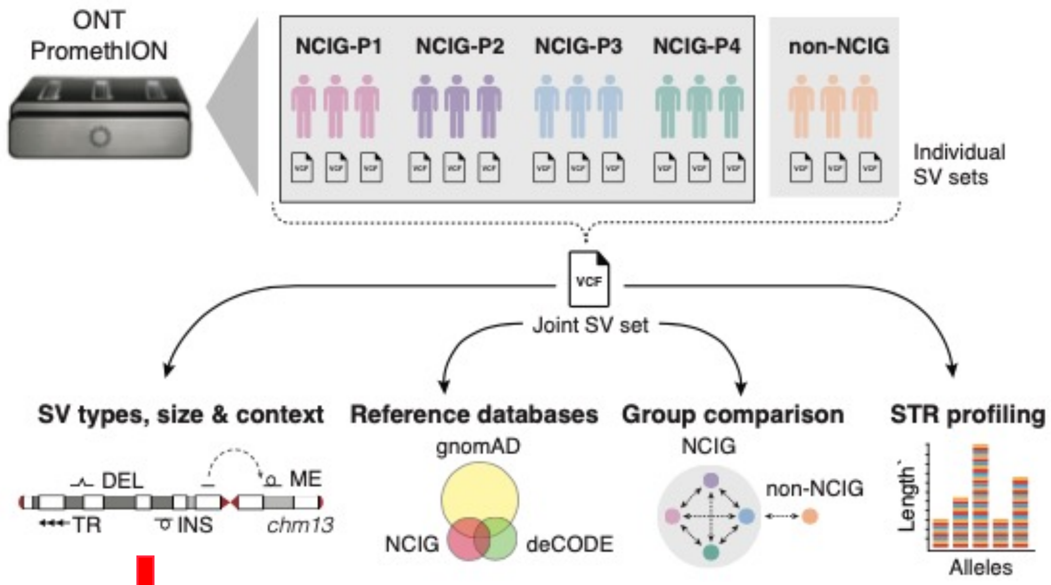
Complete and comprehensive reference sequence for all 24 human chromosomes

The landscape of genomic structural variation in Indigenous Australians
Reis et al. *bioRxiv,* Oct. 2023

Genome sequencing
4 Aboriginal communities in Australia :

Northern
Australia
1
2
3
4

	1	2	3	4
	NCIGP1	NCIGP2	NCIGP3	NCIGP4
	Tiwi Islands	Galiwin'ku	Titjikala	Yarrabah
Population:	~2,348	~2,000	~201	~4,000
Participants:	41	33	9	39



Samples : 121 Australian Indigenous + 18 non-indigenous

- Sequenced on Promethion flow cells R10.4.1
- ~30-fold genome coverage ; ~9.2 kb read-length

- T2T Consortium -> T2T-chm13 chosen as reference genome for mapping and structure variant detection
- By comparison to hg38 : → T2T-chm13 affords additional ~125 Mbases accessible to analysis

- abundance of large indels (n=136,797) structural variants (n=159,912)
- 73% not previously annotated
- large fraction (30%) exclusive to Indigenous Australians

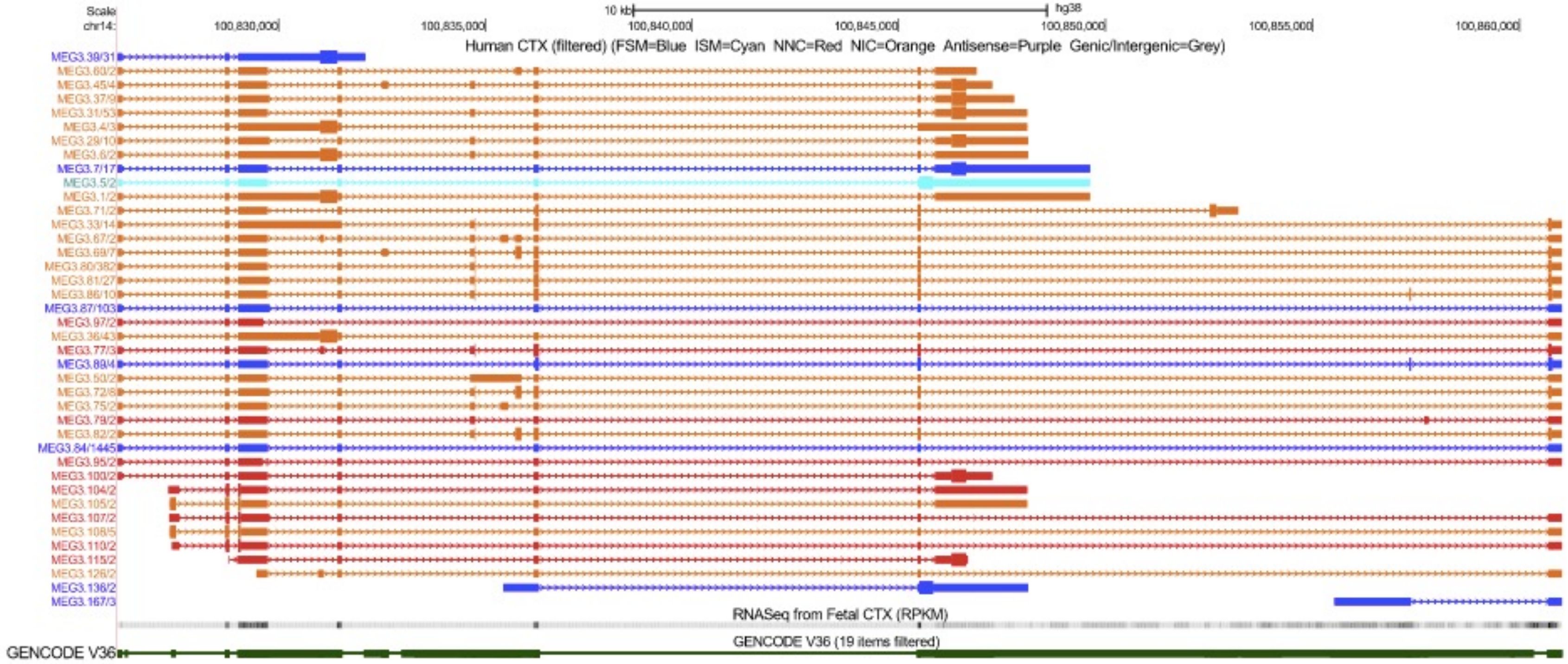
- Large diversity of genomic structural variation within Aboriginal communities

LONG READ cDNA SEQUENCING

PacBio cDNA SEQUENCING

Full-length transcript sequencing of human and mouse cerebral cortex identifies widespread isoform diversity and alternative splicing
Leung et al. *Cell Report* 2021

Transcripts annotated to MEG3 gene in the human cortex
(blue = FSM; cyan = ISM; red = NIC; orange = NNC



- 11,913 novel transcripts associated with 5,327 genes mean size = 2.84 kb, mean number of exons =11.1
- "novel in catalog" (NIC: n=8,721) contain a combination of known donor and acceptor splice sites
- "novel not in catalog" (NNC: n=3021) with at least one novel donor or acceptor site
- Novel transcripts are generally less abundant than annotated and presumably harder to detect using standard RNA-seq
- They are longer with more exons
- Our data confirm the importance of alternative splicing in the cortex, dramatically increasing transcriptional diversity and representing an important mechanism underpinning gene regulation in the brain

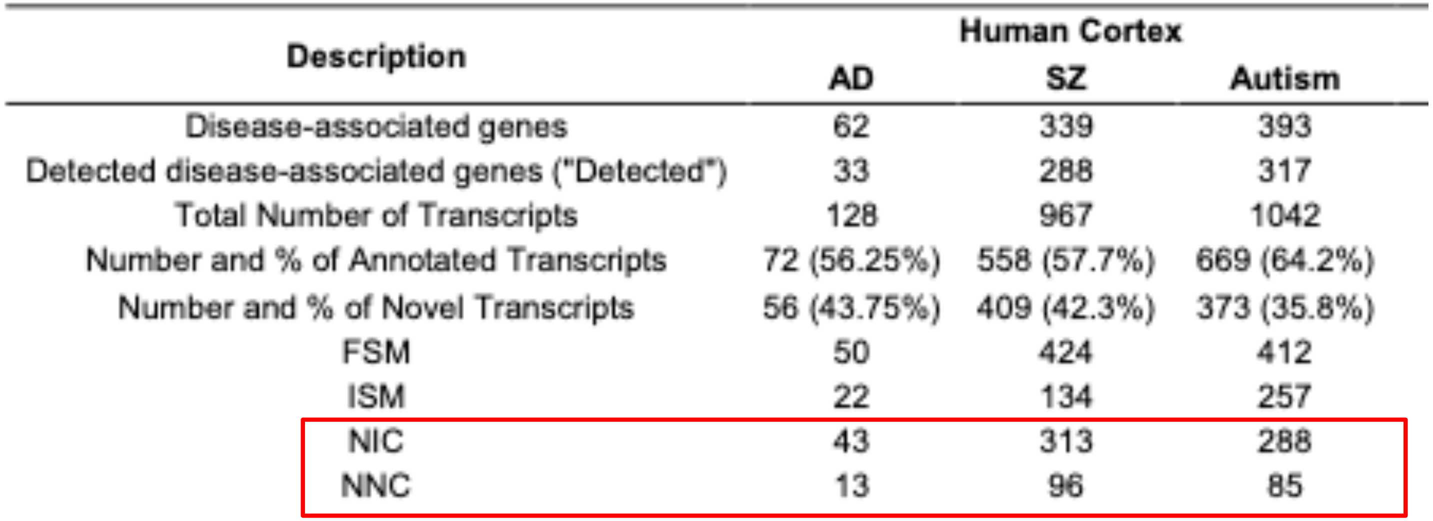
PacBio cDNA SEQUENCING

Full-length transcript sequencing of human and mouse cerebral cortex identifies widespread isoform diversity and alternative splicing
Leung et al. *Cell Report* 2021

Increasing interest in the role of AS (alternative splicing) in human disease :

- correction of AS deficits has therapeutic benefit in several disorders including spinal muscular atrophy.
- AS impacts neurodevelopment and key neural functions
- AS is a common feature of many neuropsychiatric and neurodegenerative diseases with recent studies highlighting splicing differences associated with autism

Transcripts mapping to disease-associated genes in human



Description	Human Cortex		
	AD	SZ	Autism
Disease-associated genes	62	339	393
Detected disease-associated genes ("Detected")	33	288	317
Total Number of Transcripts	128	967	1042
Number and % of Annotated Transcripts	72 (56.25%)	558 (57.7%)	669 (64.2%)
Number and % of Novel Transcripts	56 (43.75%)	409 (42.3%)	373 (35.8%)
FSM	50	424	412
ISM	22	134	257
NIC	43	313	288
NNC	13	96	85

SINGLE CELL SEQUENCING

Single-cell transcriptome :

- 10 000 to 50 000 reads / single-cell

PacBio system Sequel II :

- ~ 8 million Hi-Fi reads -> hundreds of single-cell transcriptomes

PromethION :

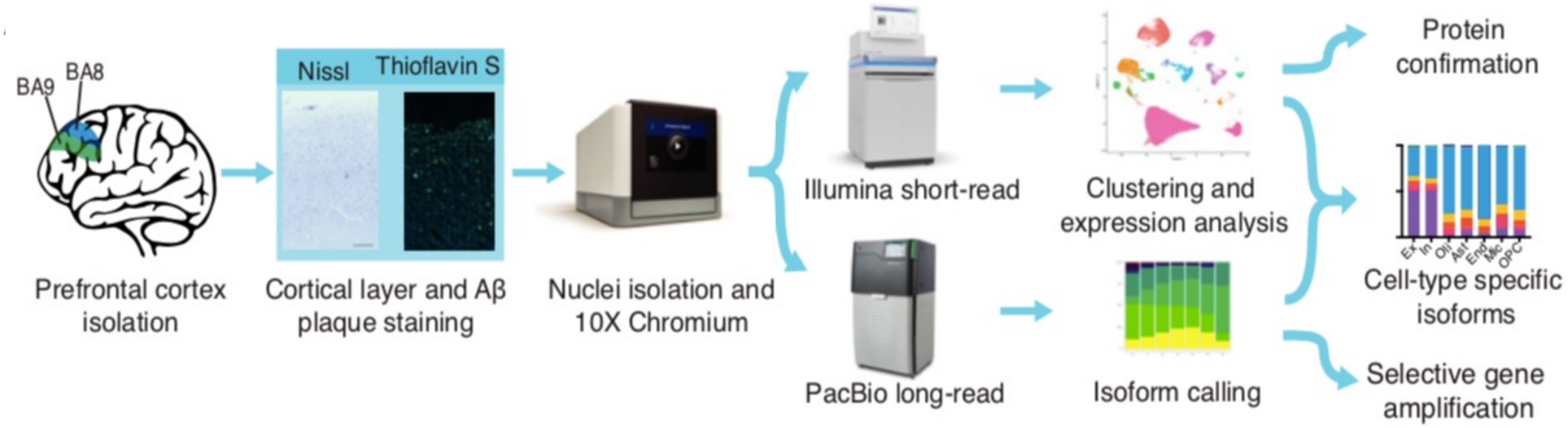
- ~ 100 million reads / flow cell -> thousands of single-cell transcriptomes

SINGLE CELL PacBio SEQUENCING

Altered cell and RNA isoform diversity in aging Down syndrome brains
Palmer et al. *PNAS* 2021

Down syndrome (trisomy 21) :

- single-nucleus long read RNA sequencing
- >170,000 cells from 29 aging DS and control brains



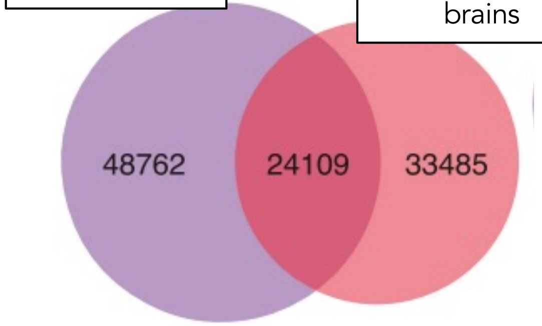
BA9
BA8
Prefrontal cortex isolation
Nissl
Thioflavin S
Cortical layer and Aβ plaque staining
Nuclei isolation and 10X Chromium
Illumina short-read
PacBio long-read
Clustering and expression analysis
Isoform calling
Protein confirmation
Ex
In
Oli
Ast
End
Mic
OPC
Cell-type specific isoforms
Selective gene amplification

New splicing isoforms :

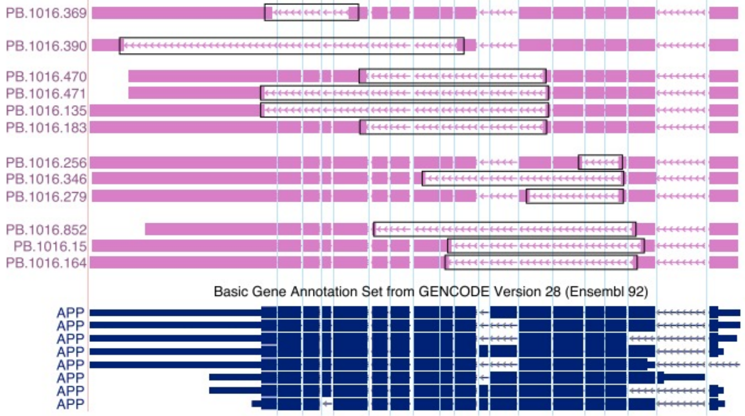
- new splice sites
- novel exon junctions
- entirely new exons
- intron retention

Control brains

Down syndrome brains

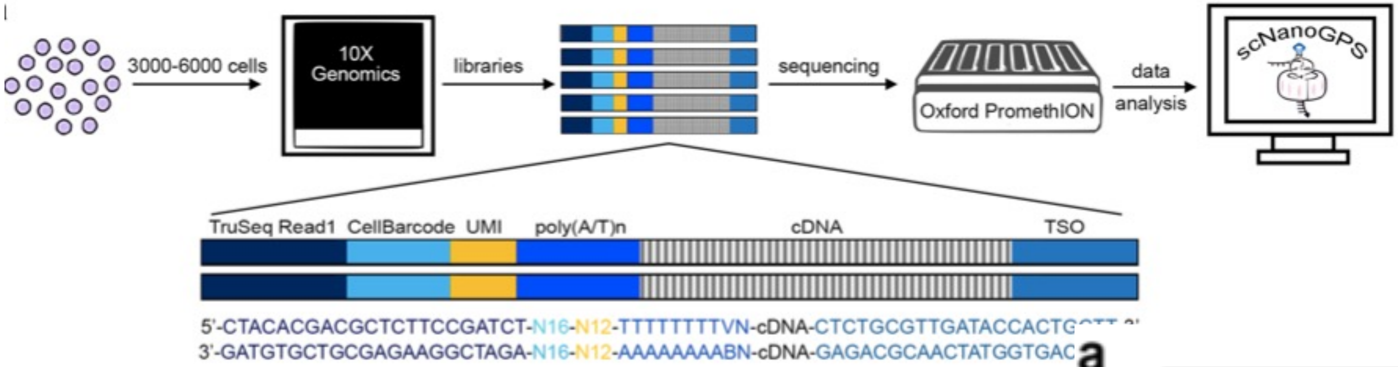


Amyloid precursor protein (Alzheimer's disease gene)



SINGLE CELL NANOPORE SEQUENCING

High throughput single cell long-read sequencing analyses of same-cell genotypes and phenotypes in human tumors
Schiau et al. *Nature Communications* July 2023

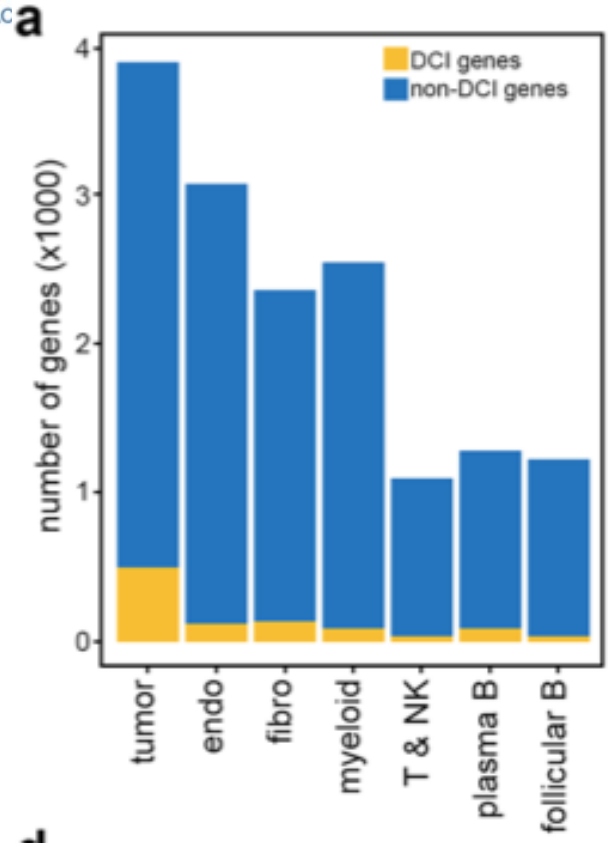

3000-6000 cells
10X Genomics
libraries
sequencing
Oxford PromethION
data analysis
scNanoGPS
TruSeq Read1
CellBarcode
UMI
poly(A/T)n
cDNA
TSO
5'-CTACACGACGCTCTTCCGATCT-N16-N12-TTTTTTTTVN-cDNA-CTCTGCGTTGATACCACTG
3'-GATGTGCTGCGAGAAGGCTAGA-N16-N12-AAAAAAAABN-cDNA-GAGACGCAACTATGGTGAC

Identification of cell-type-specific:

- isoforms (LIQA software)
- mutations
- gene expression

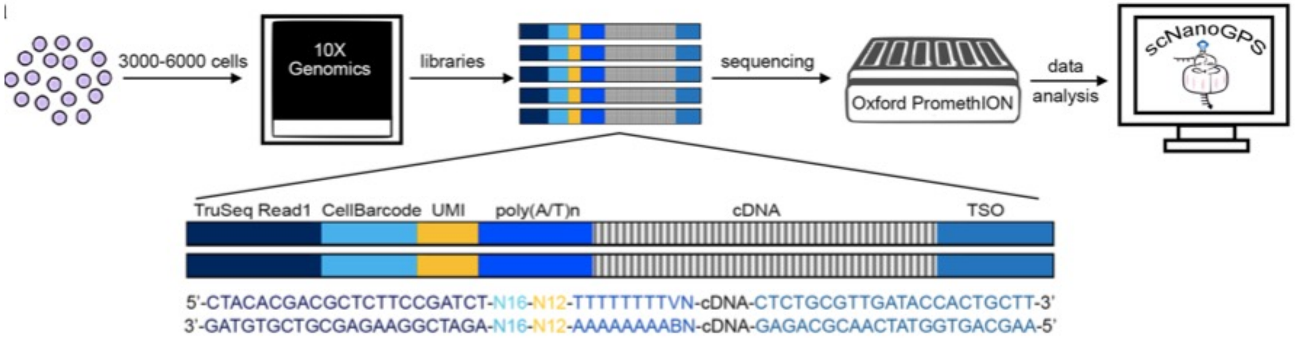
→ synchronous cell-lineage (genotype) and cell-fate (phenotype)

- 2-4 times more genes with different combination of isoforms in tumor cells (chimio-resistance pathway) compared to immune and stromal cell types



SINGLE CELL NANOPORE SEQUENCING

High throughput single cell long-read sequencing analyses of same-cell genotypes and phenotypes in human tumors
Schiau et al. *Nature Communications* July 2023



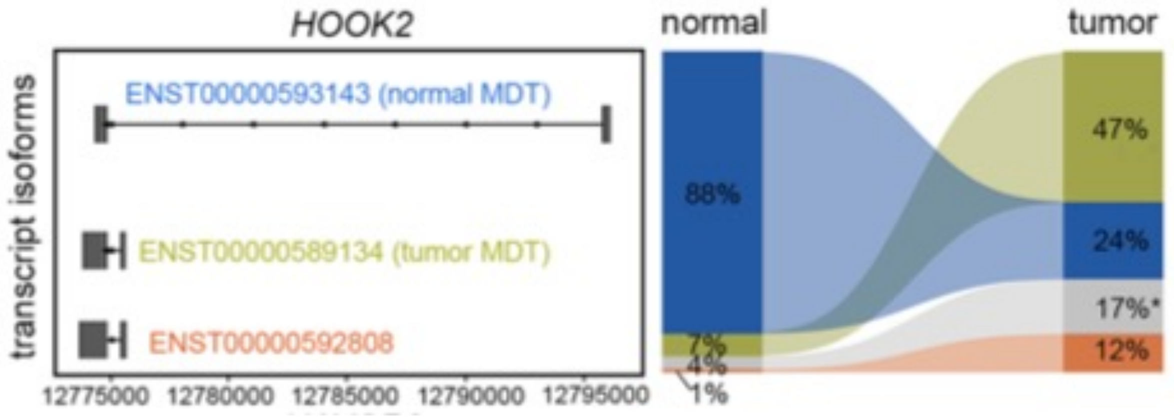
3000-6000 cells
10X Genomics
libraries
sequencing
Oxford PromethION
data analysis
scNanoGPS
TruSeq Read1
CellBarcode
UMI
poly(A/T)n
cDNA
TSO
5'-CTACACGACGCTCTTCCGATCT-N16-N12-TTTTTTTTVN-cDNA-CTCTGCGTTGATACCACTGCTT-3'
3'-GATGTGCTGCGAGAAGGCTAGA-N16-N12-AAAAAAAABN-cDNA-GAGACGCAACTATGGTGACGAA-5'

Identification of cell-type-specific:

- isoforms (LIQA software)
- mutations
- gene expression

→ synchronous cell-lineage (genotype) and cell-fate (phenotype)

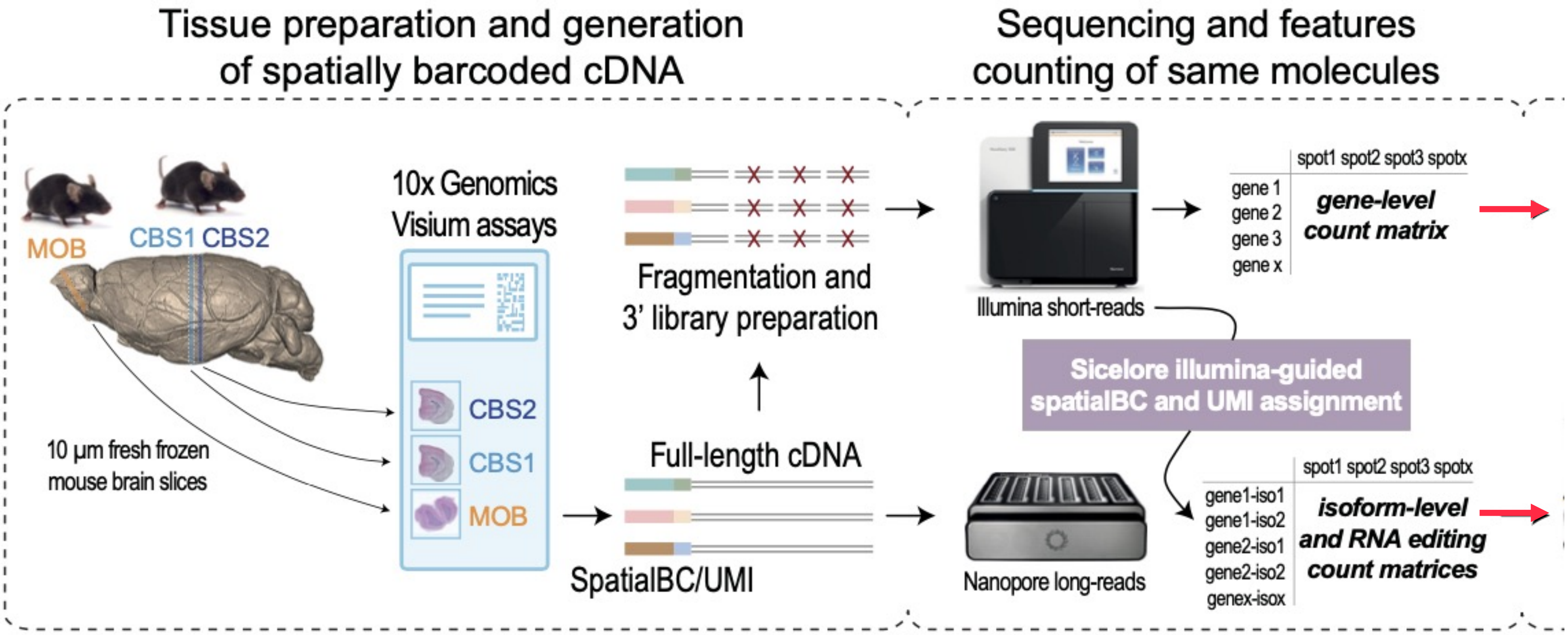
- 2-4 times more genes with different combination of isoforms in tumor cells (drug-resistance pathway) compared to immune and stromal cell types



SPATIAL TRANSCRIPTOMICS

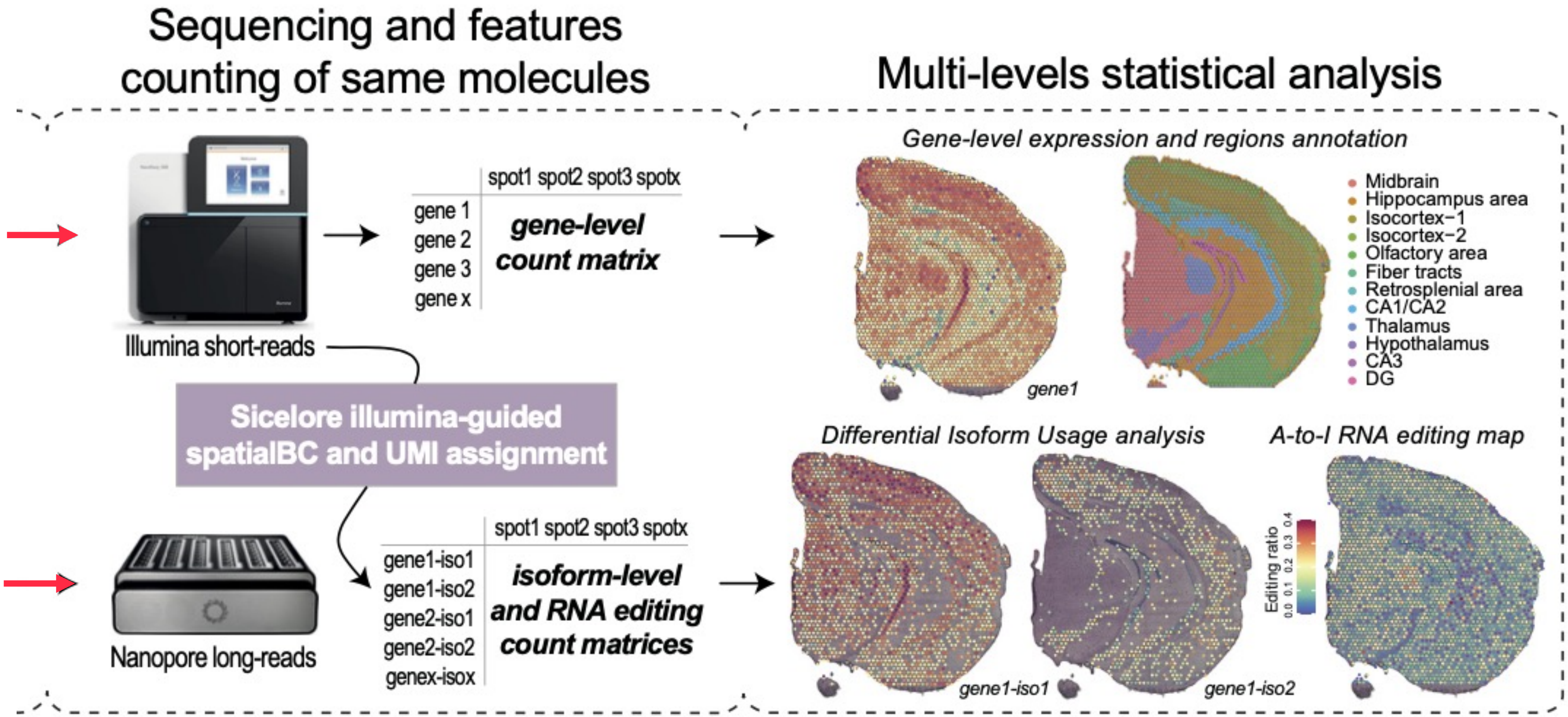
NANOPORE SPATIAL ISOFORM TRANSCRIPTOMICS

The spatial landscape of gene expression isoforms in tissue sections
Lebrigand et al. *NAR* March 2023


Tissue preparation and generation of spatially barcoded cDNA
Sequencing and features counting of same molecules
MOB
CBS1 CBS2
10 µm fresh frozen mouse brain slices
10x Genomics Visium assays
CBS2
CBS1
MOB
Fragmentation and 3' library preparation
Full-length cDNA
SpatialBC/UMI
Illumina short-reads
Nanopore long-reads
Sicelore illumina-guided spatialBC and UMI assignment
spot1 spot2 spot3 spotx
gene 1
gene 2
gene 3
gene x
gene-level count matrix
spot1 spot2 spot3 spotx
gene1-iso1
gene1-iso2
gene2-iso1
gene2-iso2
genex-isox
isoform-level and RNA editing count matrices

NANOPORE SPATIAL ISOFORM TRANSCRIPTOMICS

The spatial landscape of gene expression isoforms in tissue sections
Lebrigand et al. *NAR* March 2023



Spatial isoform transcriptomics (SiT) combines :

- short-read sequencing of cDNA -> spatial gene expression
- long-read sequencing -> spatial full-length isoforms and sequence data

NANOPORE SPATIAL ISOFORM TRANSCRIPTOMICS

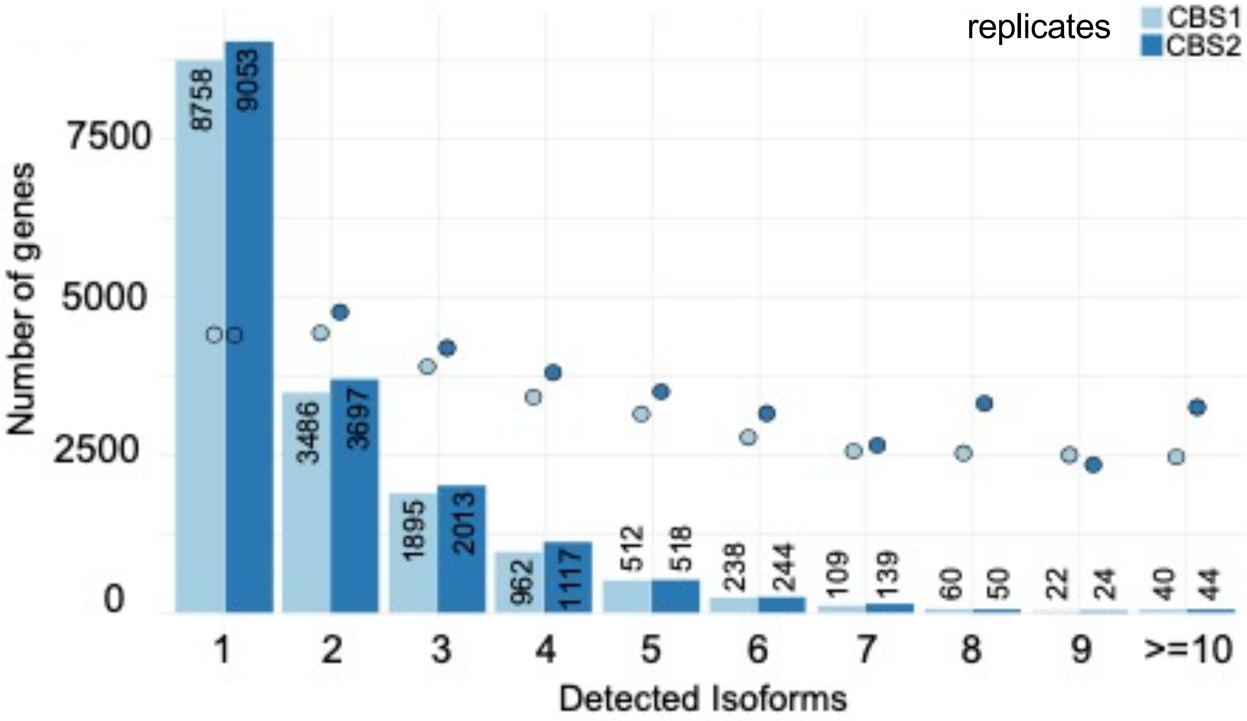
The spatial landscape of gene expression isoforms in tissue sections
Lebrigand et al. *NAR* March 2023

Coronal brain sections :

10 million UMIs assigned to a precise isoform

➡ **33097 isoforms** encoded by **16899 genes**

126 genes present regional isoform switching



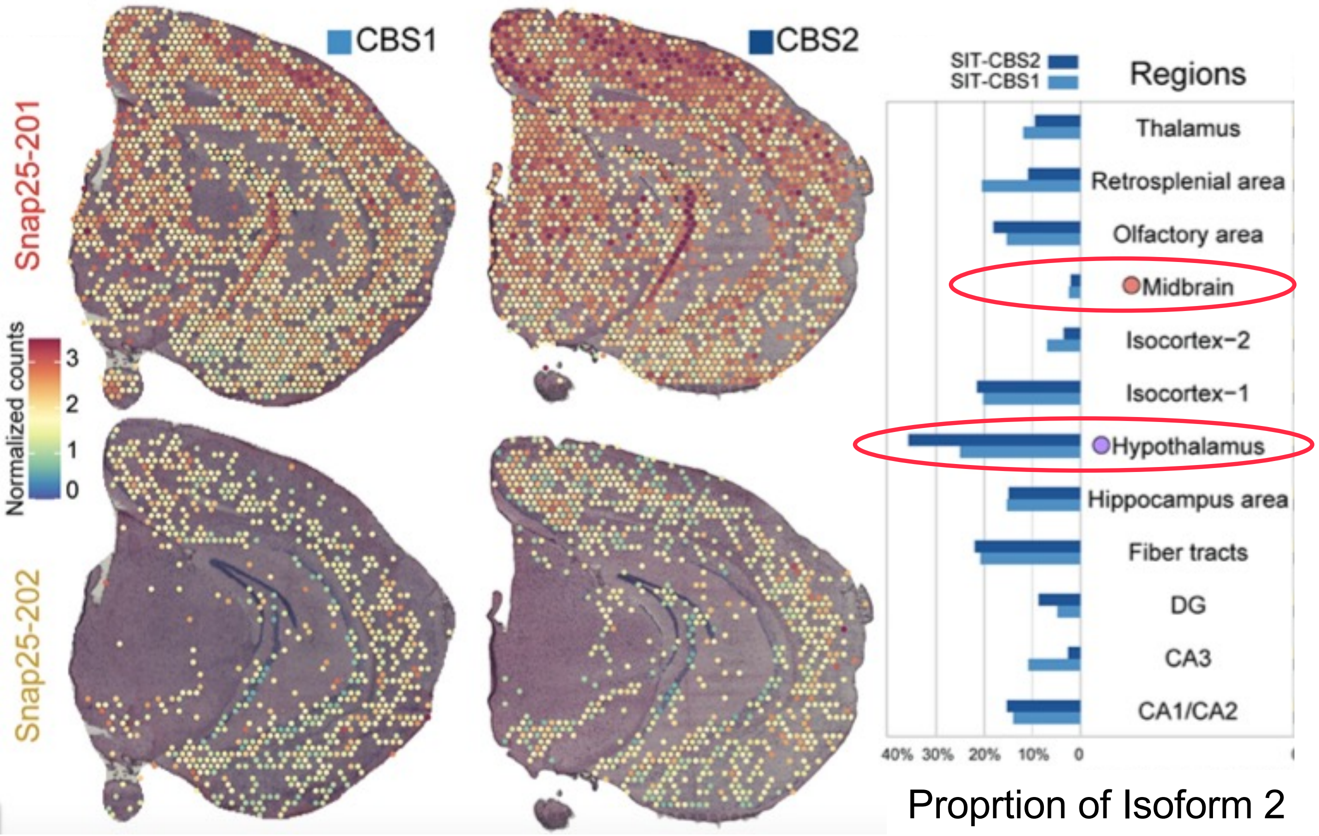
NANOPORE SPATIAL ISOFORM TRANSCRIPTOMICS

The spatial landscape of gene expression isoforms in tissue sections
Lebrigand et al. *NAR* March 2023

Regional isoform switching gene *Snap25* : codes 2 isoforms (role in synaptic plasticity)

Isoform 1

Isoform 2



Proprtion of Isoform 2

NANOPORE SPATIAL ISOFORM TRANSCRIPTOMICS

The spatial landscape of gene expression isoforms in tissue sections
Lebrigand et al. *NAR* March 2023

Conclusion

1 Nanopore flowcell (100 millions reads) is sufficient to :

- explore the spatial landscape of mRNA isoform expression in a typical Visium experiment
- resolve spatially the expression of pathological isoforms (e.g. fusion transcripts) and cancer mutations
- better characterize the heterogeneity of tumor biopsies

DIRECT RNA SEQUENCING

DETECTION OF MODIFICATIONS

MODIFIED RNA

RNA modifications (> 150) play important roles in regulating RNA fate :

- RNA folding and structure
- base pairing
- recruitment of RNA-binding proteins
- can be dynamic and reversible

In mRNAs (translation, stability, splicing..)

- *6mA* most abundant and better characterized
- *pseudo U*
- *2'O-methyl*
- ….

Also found in ncRNAs

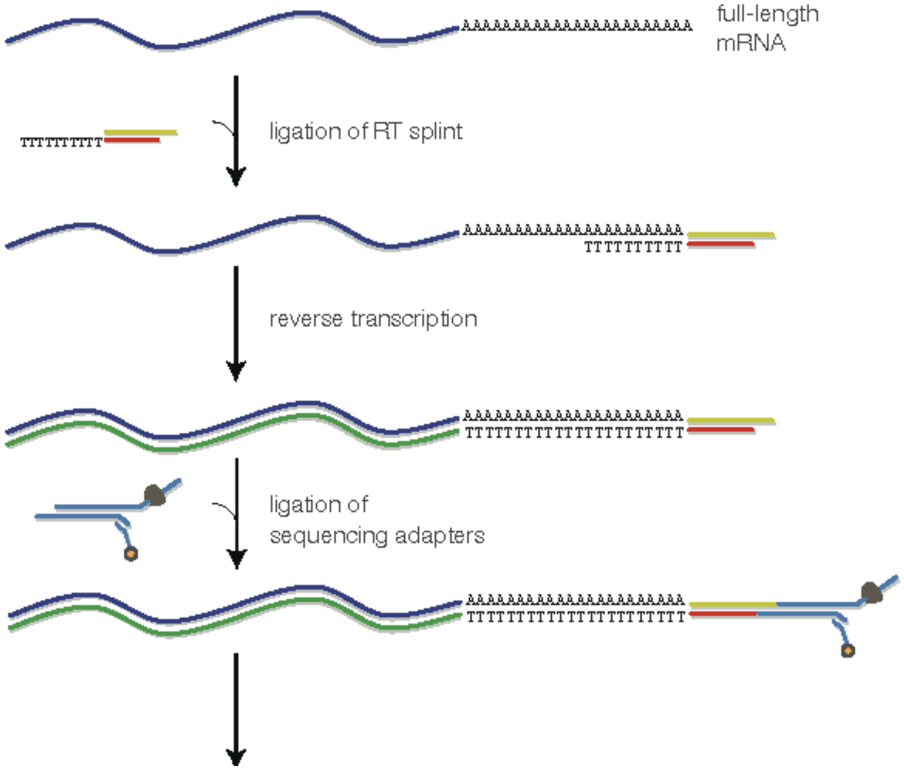
- microRNAs (miRNAs)
- long non-coding RNAs (lncRNAs)
- circular RNAs (circRNAs)

Viral RNAs contain high levels of modifications (modulate virus cycle)

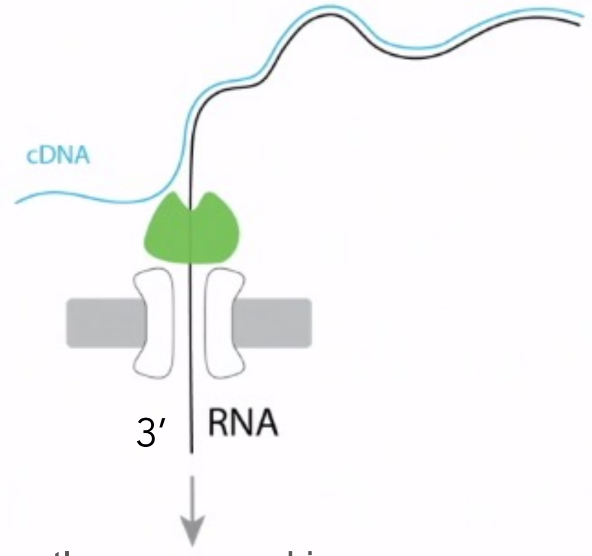
- HIV RNA rich in *6mA*

DIRECT RNA SEQUENCING

Library preparation



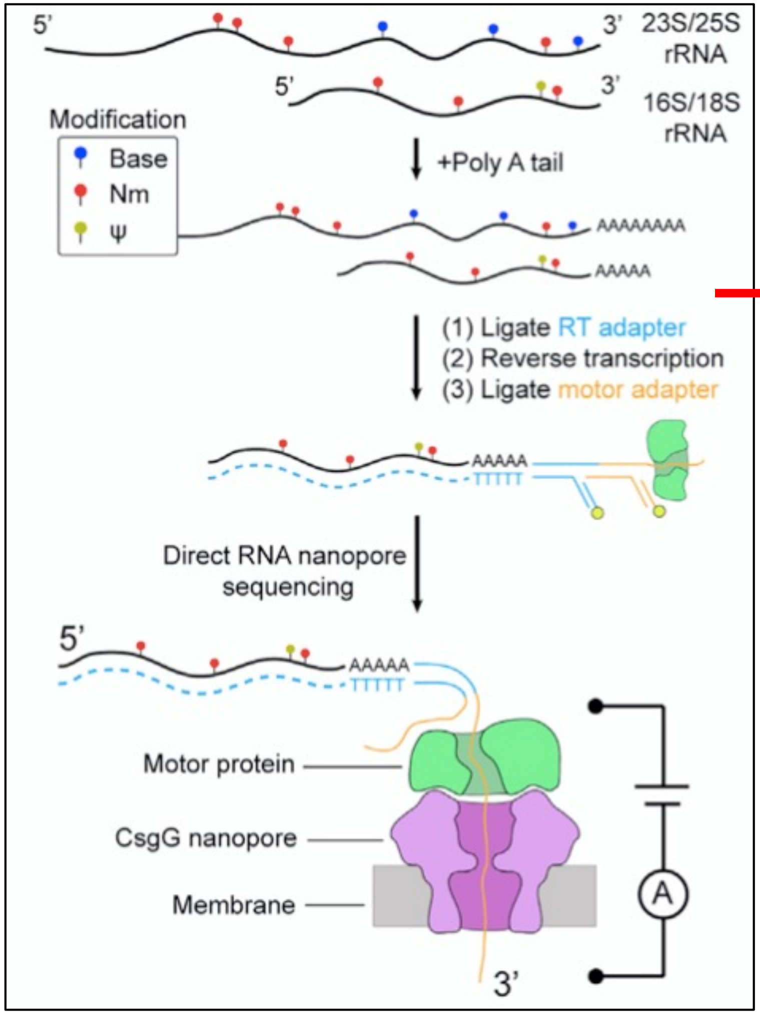
sequencing

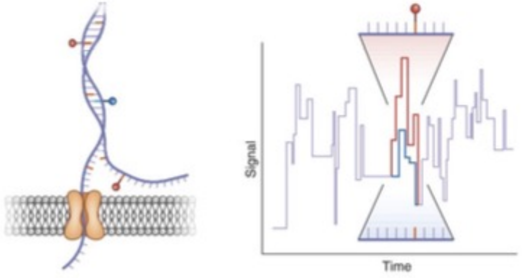


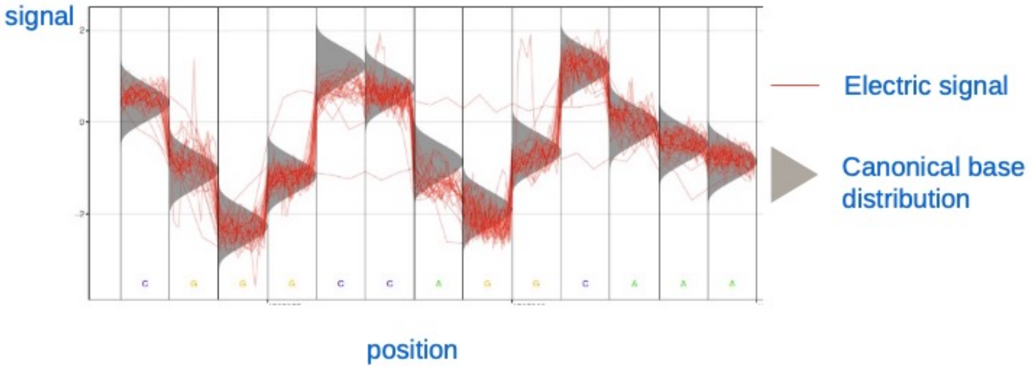
RNA directly sequenced in nanopore

- No PCR bias
- Quantitative

DIRECT RNA SEQUENCING : DETECTION OF MODIFICATIONS





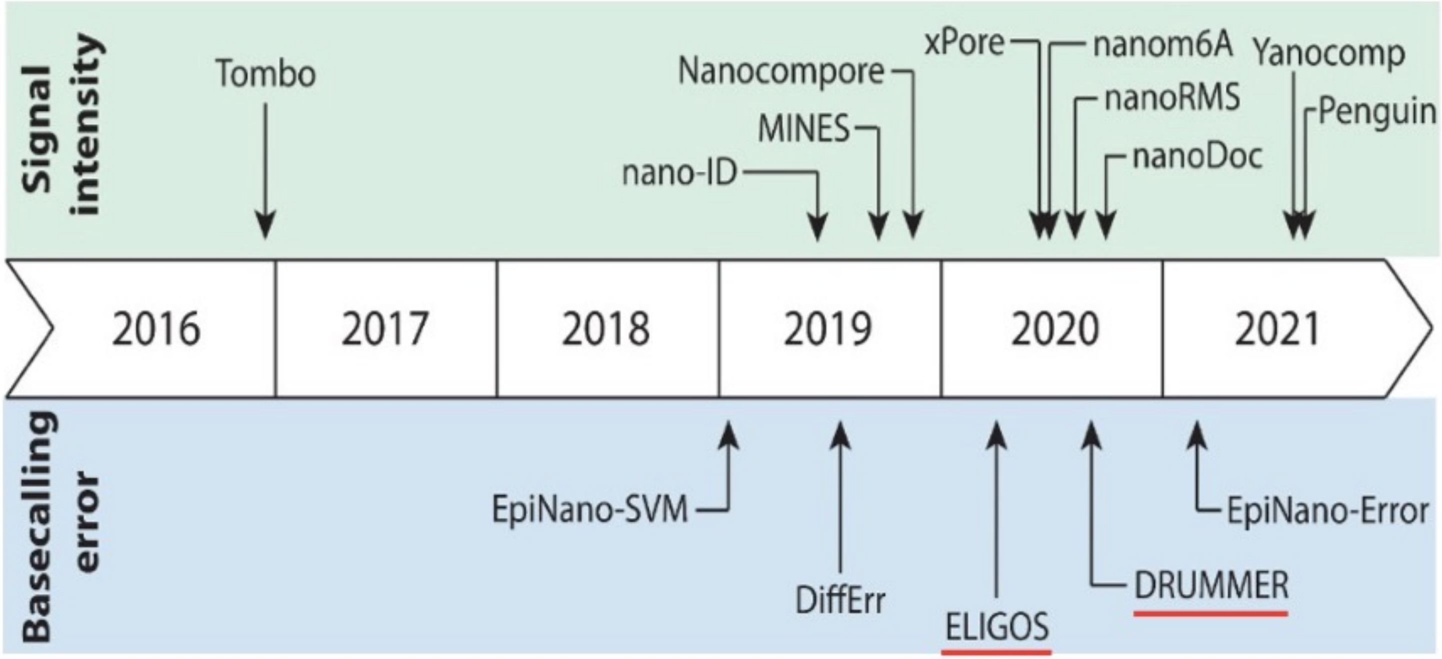


Base-calling

- analysis of **FAST5/POD5** features (signal intensity)
- analysis of **base-called** 'error' features

DIRECT RNA SEQUENCING : DETECTION OF MODIFICATIONS

DIFFERENT TOOLS



Furlan et al, RNA Biology, 2021.

DIRECT RNA SEQUENCING : DETECTION OF MODIFIED RNA

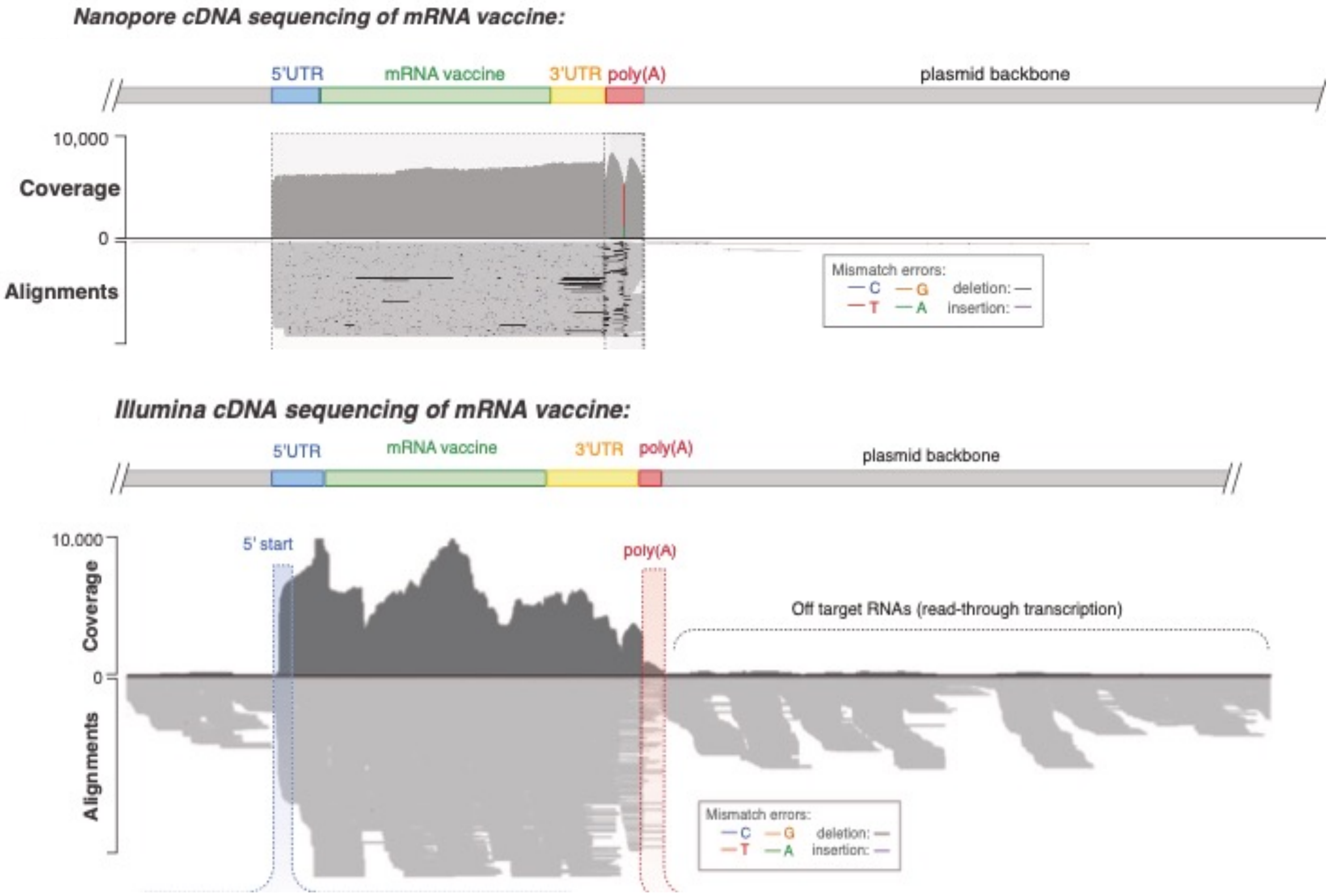
mRNA vaccine quality analysis using RNA sequencing
Gunter et al. *Nat. Comm. Sept.* 2023

mRNA vaccines must be rigorously analyzed :

- to measure their integrity
- detect contaminants that reduce their effectiveness and induce side-effects
- Currently, mRNA vaccines and therapies are analysed using time-consuming and costly methods
- This work describes a how to analyse mRNA vaccines using long-read nanopore sequencing.

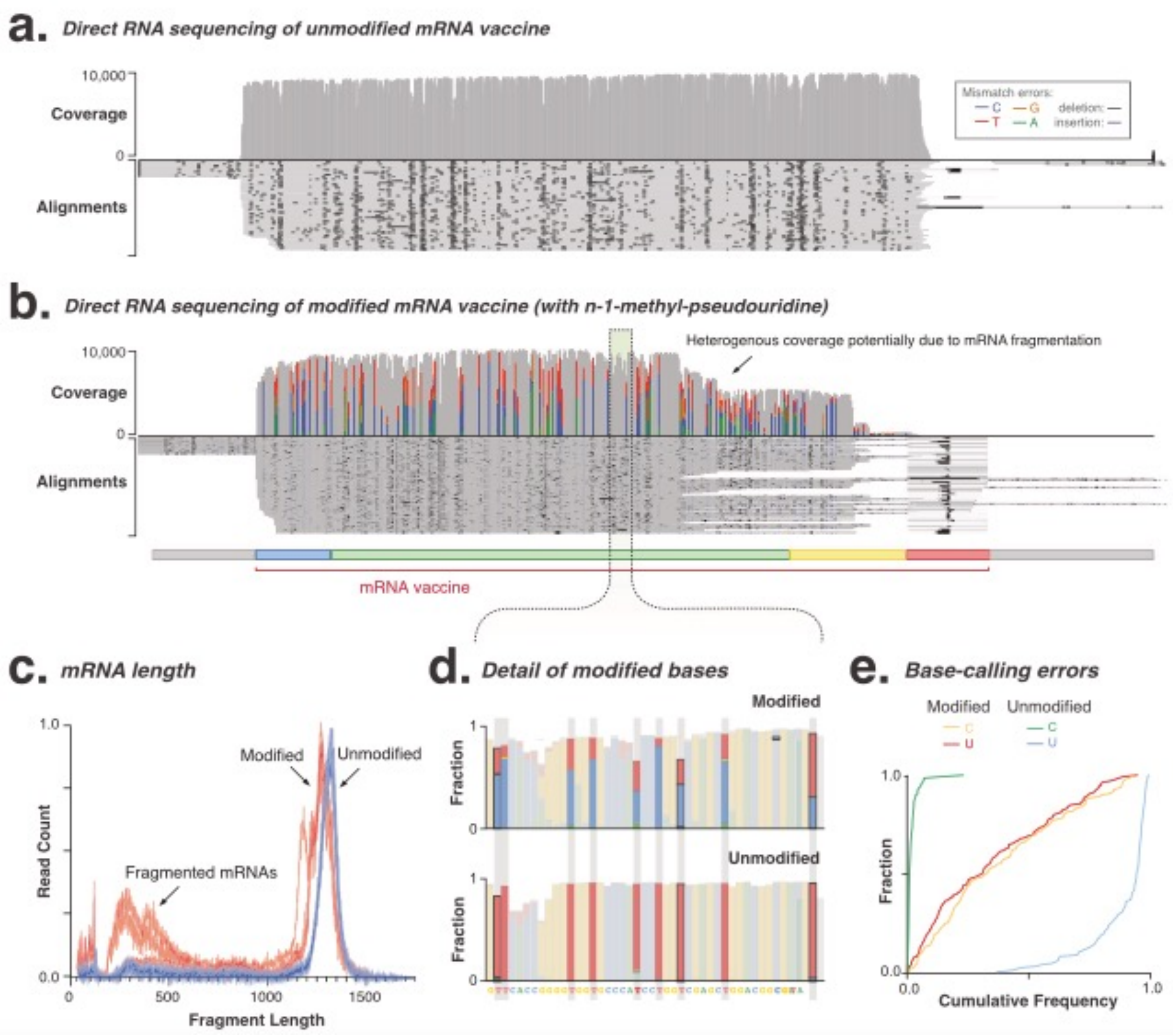
DIRECT RNA SEQUENCING : DETECTION OF MODIFIED RNA

mRNA vaccine quality analysis using RNA sequencing
Gunter et al. *Nat. Comm. Sept.* 2023



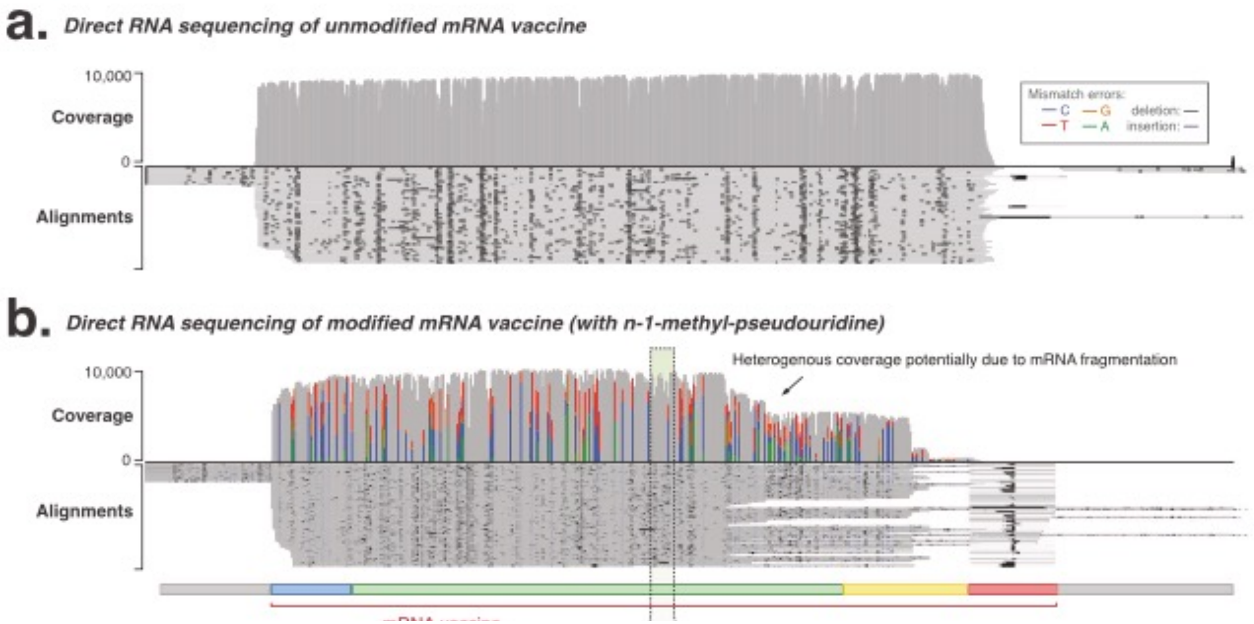
DIRECT RNA SEQUENCING : DETECTION OF MODIFIED RNA

mRNA vaccine quality analysis using RNA sequencing
Gunter et al. *Nat. Comm. Sept.* 2023



DIRECT RNA SEQUENCING : DETECTION OF MODIFIED RNA

mRNA vaccine quality analysis using RNA sequencing
Gunter et al. *Nat. Comm. Sept.* 2023



a. *Direct RNA sequencing of unmodified mRNA vaccine*

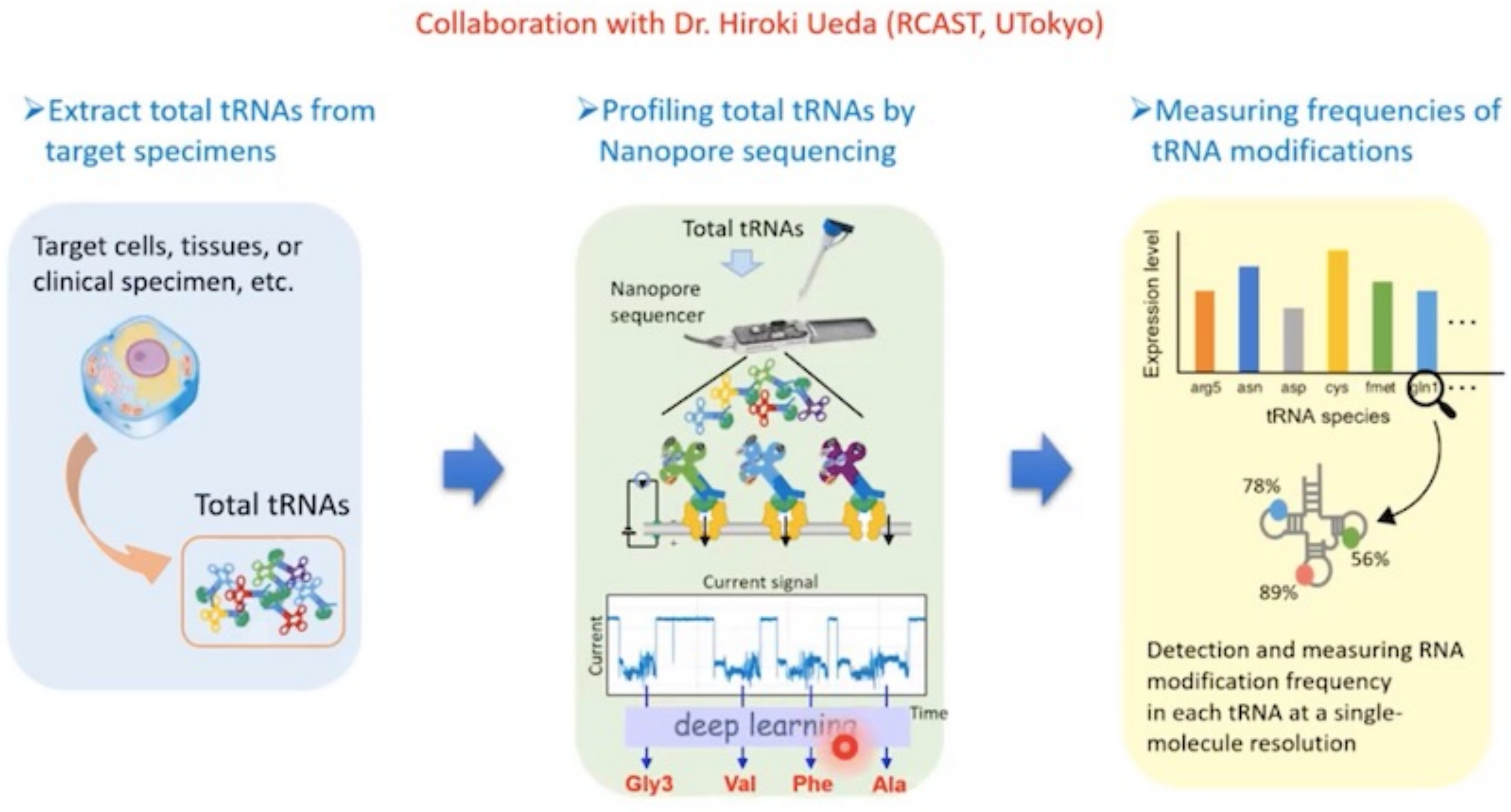
b. *Direct RNA sequencing of modified mRNA vaccine (with n-1-methyl-pseudouridine)*

Compared to other industry-standard techniques, VAX-seq can comprehensively measure key mRNA vaccine quality attributes, including sequence, length, integrity, and purity.

Direct RNA sequencing can analyse mRNA chemistry, including the detection of nucleoside modifications.

DIRECT RNA SEQUENCING : DETECTION OF MODIFIED RNA

tRNA profiling using Nanopore sequencer
Tsutomu Suzuki, London Calling Nanopore meeting 2023



Collaboration with Dr. Hiroki Ueda (RCAST, UTokyo)

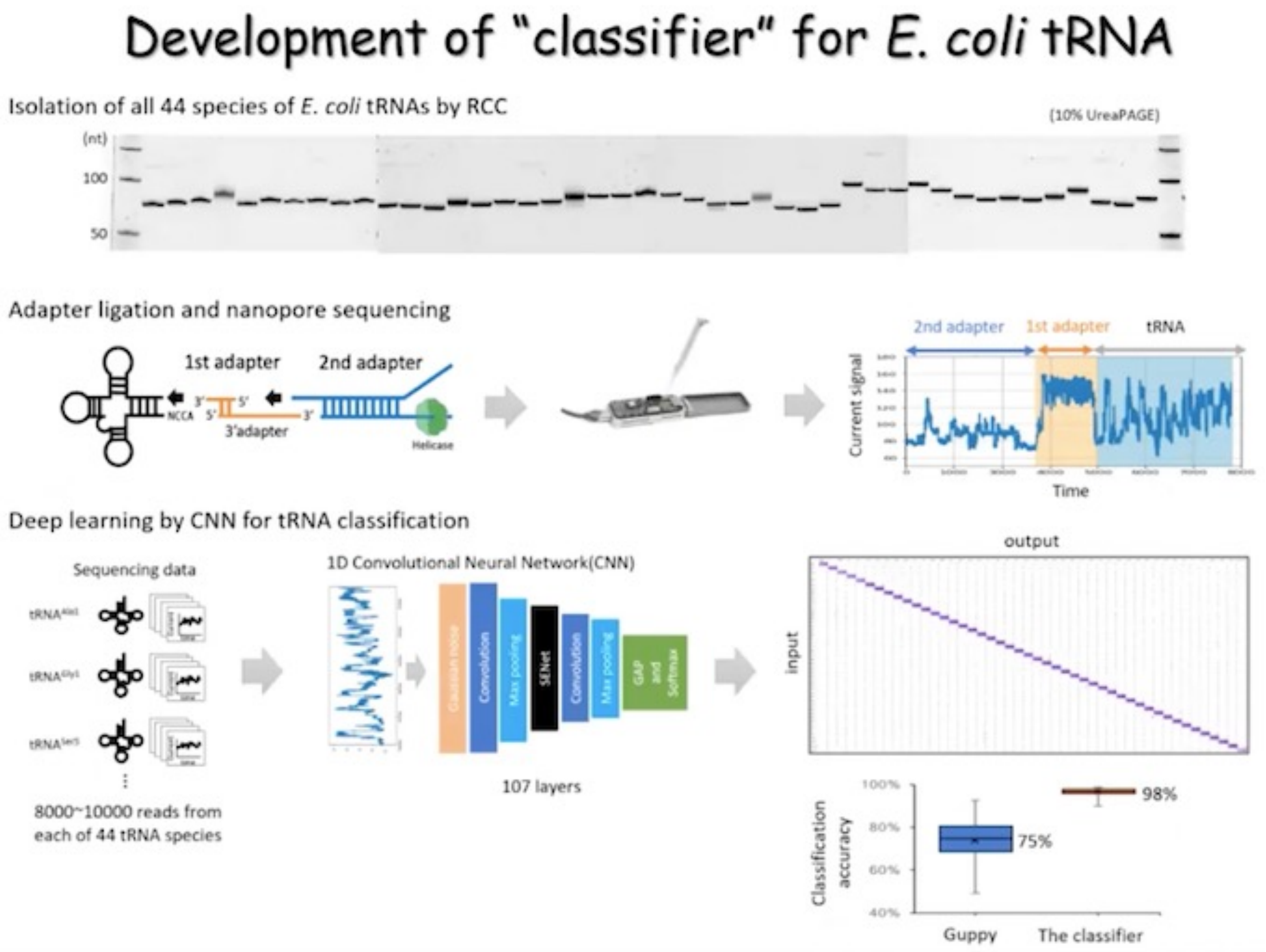
- Extract total tRNAs from target specimens
- Profiling total tRNAs by Nanopore sequencing
- Measuring frequencies of tRNA modifications

Target cells, tissues, or clinical specimen, etc.

Detection and measuring RNA modification frequency in each tRNA at a single-molecule resolution

DIRECT RNA SEQUENCING : DETECTION OF MODIFIED RNA

tRNA profiling using Nanopore sequencer
Tsutomu Suzuki, London Calling Nanopore meeting 2023



Development of "classifier" for *E. coli* tRNA

Isolation of all 44 species of *E. coli* tRNAs by RCC

(10% UreaPAGE)

Adapter ligation and nanopore sequencing

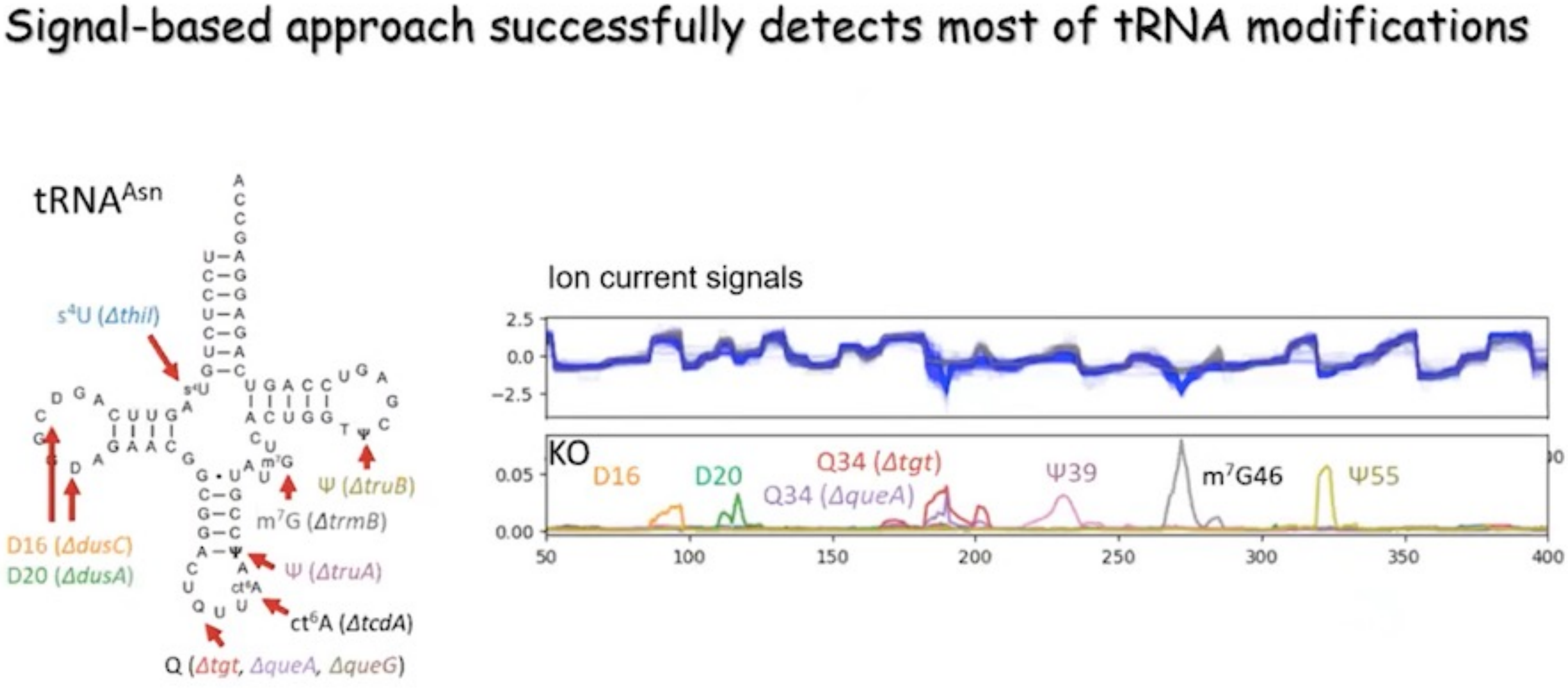
Deep learning by CNN for tRNA classification

8000~10000 reads from each of 44 tRNA species

107 layers

DIRECT RNA SEQUENCING : DETECTION OF MODIFIED RNA

tRNA profiling using Nanopore sequencer
Tsutomu Suzuki, London Calling Nanopore meeting 2023



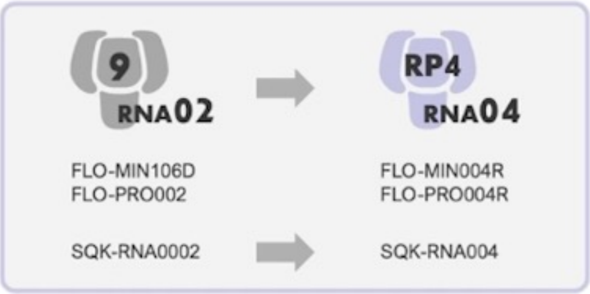
Signal-based approach successfully detects most of tRNA modifications

DIRECT RNA SEQUENCING : DETECTION OF MODIFIED RNA

Future improvements

RNA enzyme motor developed for better speed and output

- Now averaging speed of 125 bps (~2x improvement)
- Hitting outputs of 30 million reads from PromethION flowcell

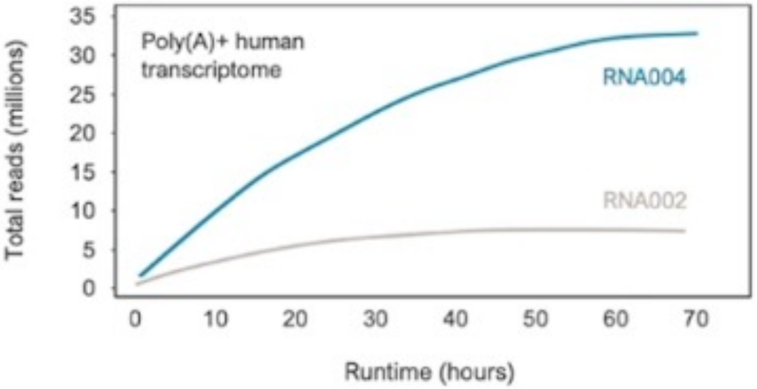


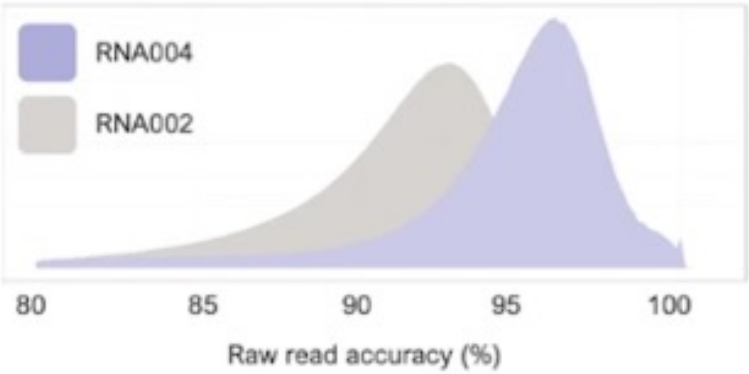
9 RNA02 → RP4 RNA04

FLO-MIN106D
FLO-PRO002

FLO-MIN004R
FLO-PRO004R

SQK-RNA0002 → SQK-RNA004





C. Brown London Calling 2023

LARGE SEQUENCING PROJECTS

VERTEBRATE GENOMES PROJECT (VGP)

Towards complete and error-free genome assemblies of all vertebrate species
Rhie et al. *Nature* 2021

International effort to generate high-quality, complete reference genomes :

- For all of the roughly 70,000 extant vertebrate species
- To enable a new era of discovery across the life sciences

Arang Rhie[1,103], Shane A. McCarthy[2,3,103], Olivier Fedrigo[4,103], Joana Damas[5],
Giulio Formenti[4,6], Sergey Koren[1], Marcela Uliano-Silva[7,8], William Chow[3],
Arkarachai Fungtammasan[9], Juwan Kim[10], Chul Lee[10], Byung June Ko[11], Mark Chaisson[12],
Gregory L. Gedman[6], Lindsey J. Cantin[6], Francoise Thibaud-Nissen[13], Leanne Haggerty[14],
Iliana Bista[2,3], Michelle Smith[3], Bettina Haase[4], Jacquelyn Mountcastle[4], Sylke Winkler[15,16],
Sadye Paez[4,6], Jason Howard[17], Sonja C. Vernes[18,19,20], Tanya M. Lama[21], Frank Grutzner[22],
Wesley C. Warren[23], Christopher N. Balakrishnan[24], Dave Burt[25], Julia M. George[26],
Matthew T. Biegler[6], David Iorns[27], Andrew Digby[28], Daryl Eason[28], Bruce Robertson[29],
Taylor Edwards[30], Mark Wilkinson[31], George Turner[32], Axel Meyer[33], Andreas F. Kautt[33,34],
Paolo Franchini[33], H. William Detrich III[35], Hannes Svardal[36,37], Maximilian Wagner[38],
Gavin J. P. Naylor[39], Martin Pippel[15,40], Milan Malinsky[3,41], Mark Mooney[42], Maria Simbirsky[9],
Brett T. Hannigan[9], Trevor Pesout[43], Marlys Houck[44], Ann Misuraca[44], Sarah B. Kingan[45],
Richard Hall[45], Zev Kronenberg[45], Ivan Sović[45,46], Christopher Dunn[45], Zemin Ning[3],
Alex Hastie[47], Joyce Lee[47], Siddarth Selvaraj[48], Richard E. Green[43,49], Nicholas H. Putnam[50],
Ivo Gut[51,52], Jay Ghurye[49,53], Erik Garrison[43], Ying Sims[3], Joanna Collins[3], Sarah Pelan[3],
James Torrance[3], Alan Tracey[3], Jonathan Wood[3], Robel E. Dagnew[12], Dengfeng Guan[2,54],
Sarah E. London[55], David F. Clayton[56], Claudio V. Mello[57], Samantha R. Friedrich[57],
Peter V. Lovell[57], Ekaterina Osipova[15,40,58], Farooq O. Al-Ajli[59,60,61], Simona Secomandi[62],
Heebal Kim[10,11,63], Constantina Theofanopoulou[6], Michael Hiller[64,65,66], Yang Zhou[67],
Robert S. Harris[68], Kateryna D. Makova[68,69,70], Paul Medvedev[69,70,71,72], Jinna Hoffman[13],
Patrick Masterson[13], Karen Clark[13], Fergal Martin[14], Kevin Howe[14], Paul Flicek[14],
Brian P. Walenz[1], Woori Kwak[63,73], Hiram Clawson[43], Mark Diekhans[43], Luis Nassar[43],
Benedict Paten[43], Robert H. S. Kraus[33,74], Andrew J. Crawford[75], M. Thomas P. Gilbert[76,77],
Guojie Zhang[78,79,80,81], Byrappa Venkatesh[82], Robert W. Murphy[83], Klaus-Peter Koepfli[84],
Beth Shapiro[85,86], Warren E. Johnson[84,87,88], Federica Di Palma[89], Tomas Marques-Bon
et[90,91,92,93], Emma C. Teeling[94], Tandy Warnow[95], Jennifer Marshall Graves[96],
Oliver A. Ryder[44,97], David Haussler[43,85], Stephen J. O'Brien[98,99], Jonas Korlach[45],
Harris A. Lewin[5,100,101], Kerstin Howe[3,104] ✉, Eugene W. Myers[15,40,102,104] ✉,
Richard Durbin[2,3,104] ✉, Adam M. Phillippy[1,104] ✉ & Erich D. Jarvis[4,6,86,104] ✉

VERTEBRATE GENOMES PROJECT (VGP)

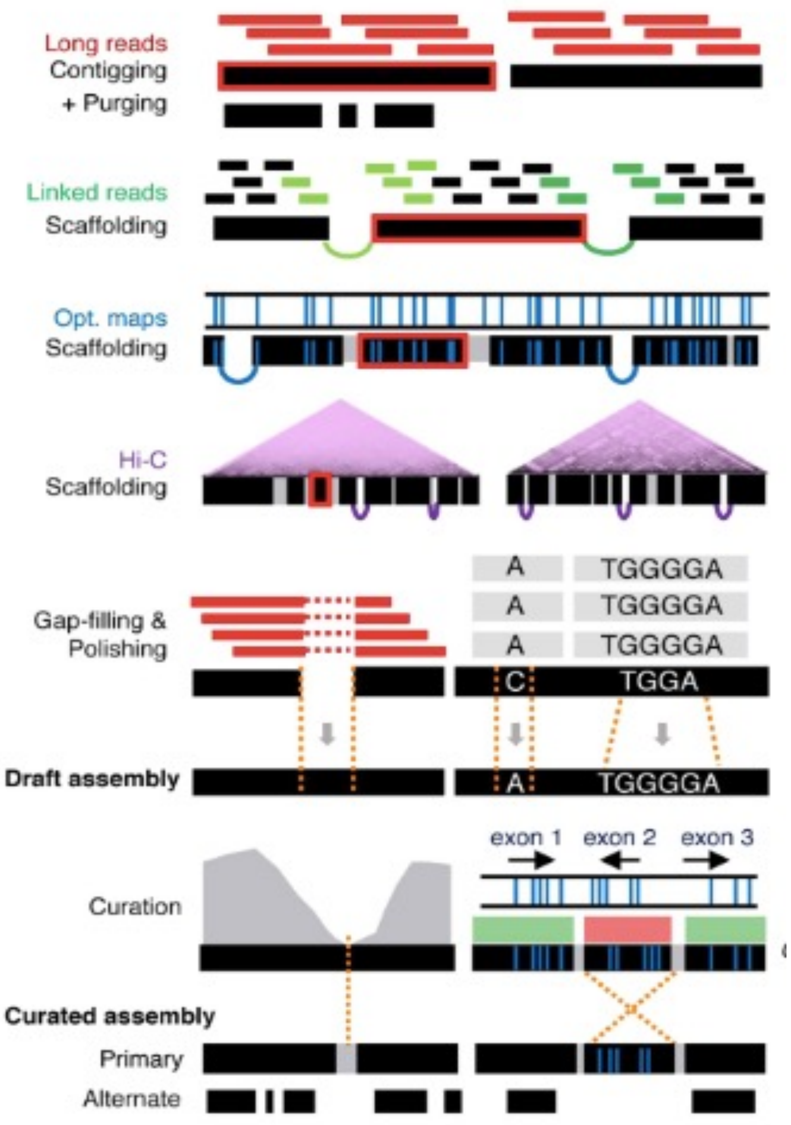
Towards complete and error-free genome assemblies of all vertebrate species
Rhie et al. *Nature* 2021

Obtain high-quality cells or tissue that would yield high-molecular-weight DNA :

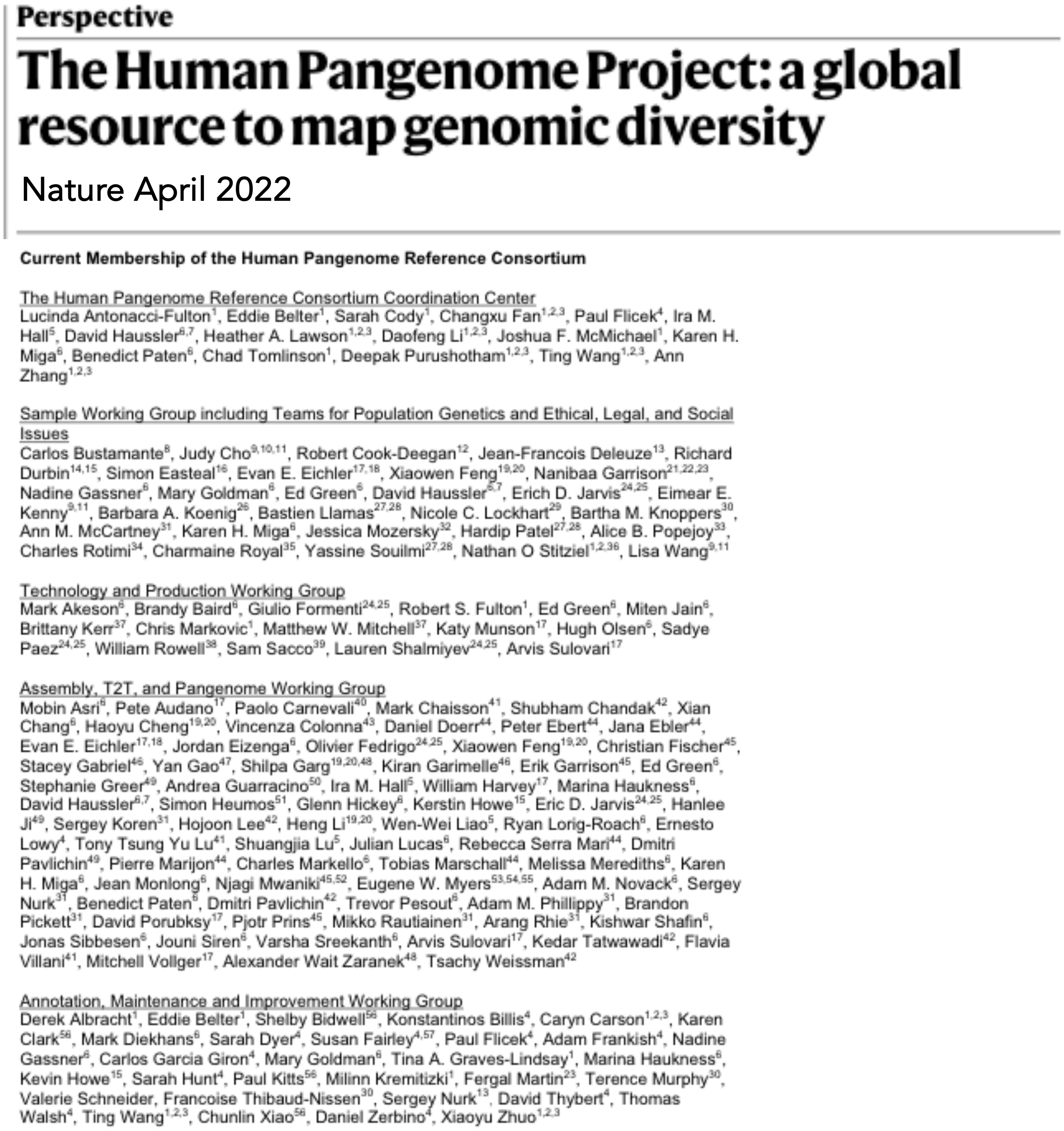
- for long-read sequencing (PacBio and ONT)
- optical mapping (Bionano)

“We will take advantage of continuing improvements in genome sequencing technology, assembly, and annotation, including advances in PacBio HiFi reads, Oxford Nanopore reads, and replacements for 10XG reads”

VGP assembly pipeline applied across multiple species



THE HUMAN PANGENOME PROJECT



Perspective

The Human Pangenome Project: a global resource to map genomic diversity

Nature April 2022

Current Membership of the Human Pangenome Reference Consortium

The Human Pangenome Reference Consortium Coordination Center

Lucinda Antonacci-Fulton[1], Eddie Belter[1], Sarah Cody[1], Changxu Fan[1,2,3], Paul Flicek[4], Ira M. Hall[5], David Haussler[6,7], Heather A. Lawson[1,2,3], Daofeng Li[1,2,3], Joshua F. McMichael[1], Karen H. Miga[6], Benedict Paten[6], Chad Tomlinson[1], Deepak Purushotham[1,2,3], Ting Wang[1,2,3], Ann Zhang[1,2,3]

Sample Working Group including Teams for Population Genetics and Ethical, Legal, and Social Issues

Carlos Bustamante[8], Judy Cho[9,10,11], Robert Cook-Deegan[12], Jean-Francois Deleuze[13], Richard Durbin[14,15], Simon Easteal[16], Evan E. Eichler[17,18], Xiaowen Feng[19,20], Nanibaa Garrison[21,22,23], Nadine Gassner[6], Mary Goldman[6], Ed Green[6], David Haussler[6,7], Erich D. Jarvis[24,25], Eimear E. Kenny[9,11], Barbara A. Koenig[26], Bastien Llamas[27,28], Nicole C. Lockhart[29], Bartha M. Knoppers[30], Ann M. McCartney[31], Karen H. Miga[6], Jessica Mozersky[32], Hardip Patel[27,28], Alice B. Popejoy[33], Charles Rotimi[34], Charmaine Royal[35], Yassine Souilmi[27,28], Nathan O Stitziel[1,2,36], Lisa Wang[9,11]

Technology and Production Working Group

Mark Akeson[6], Brandy Baird[6], Giulio Formenti[24,25], Robert S. Fulton[1], Ed Green[6], Miten Jain[6], Brittany Kerr[37], Chris Markovic[1], Matthew W. Mitchell[37], Katy Munson[17], Hugh Olsen[6], Sadye Paez[24,25], William Rowell[38], Sam Sacco[39], Lauren Shalmiyev[24,25], Arvis Sulovari[17]

Assembly, T2T, and Pangenome Working Group

Mobin Asri[6], Pete Audano[17], Paolo Carnevali[40], Mark Chaisson[41], Shubham Chandak[42], Xian Chang[6], Haoyu Cheng[19,20], Vincenza Colonna[43], Daniel Doerr[44], Peter Ebert[44], Jana Ebler[44], Evan E. Eichler[17,18], Jordan Eizenga[6], Olivier Fedrigo[24,25], Xiaowen Feng[19,20], Christian Fischer[45], Stacey Gabriel[46], Yan Gao[47], Shilpa Garg[19,20,48], Kiran Garimelle[46], Erik Garrison[45], Ed Green[6], Stephanie Greer[49], Andrea Guarracino[50], Ira M. Hall[5], William Harvey[17], Marina Haukness[6], David Haussler[6,7], Simon Heumos[51], Glenn Hickey[6], Kerstin Howe[15], Eric D. Jarvis[24,25], Hanlee Ji[49], Sergey Koren[31], Hojoon Lee[42], Heng Li[19,20], Wen-Wei Liao[5], Ryan Lorig-Roach[6], Ernesto Lowy[4], Tony Tsung Yu Lu[41], Shuangjia Lu[5], Julian Lucas[6], Rebecca Serra Mari[44], Dmitri Pavlichin[49], Pierre Marijon[44], Charles Markello[6], Tobias Marschall[44], Melissa Merediths[6], Karen H. Miga[6], Jean Monlong[6], Njagi Mwaniki[45,52], Eugene W. Myers[53,54,55], Adam M. Novack[6], Sergey Nurk[31], Benedict Paten[6], Dmitri Pavlichin[42], Trevor Pesout[6], Adam M. Phillippy[31], Brandon Pickett[31], David Porubksy[17], Pjotr Prins[45], Mikko Rautiainen[31], Arang Rhie[31], Kishwar Shafin[6], Jonas Sibbesen[6], Jouni Siren[6], Varsha Sreekanth[6], Arvis Sulovari[17], Kedar Tatwawadi[42], Flavia Villani[41], Mitchell Vollger[17], Alexander Wait Zaranek[48], Tsachy Weissman[42]

Annotation, Maintenance and Improvement Working Group

Derek Albracht[1], Eddie Belter[1], Shelby Bidwell[56], Konstantinos Billis[4], Caryn Carson[1,2,3], Karen Clark[56], Mark Diekhans[6], Sarah Dyer[4], Susan Fairley[4,57], Paul Flicek[4], Adam Frankish[4], Nadine Gassner[6], Carlos Garcia Giron[4], Mary Goldman[6], Tina A. Graves-Lindsay[1], Marina Haukness[6], Kevin Howe[15], Sarah Hunt[4], Paul Kitts[56], Milinn Kremitizki[1], Fergal Martin[23], Terence Murphy[30], Valerie Schneider, Francoise Thibaud-Nissen[30], Sergey Nurk[13], David Thybert[4], Thomas Walsh[4], Ting Wang[1,2,3], Chunlin Xiao[56], Daniel Zerbino[4], Xiaoyu Zhuo[1,2,3]

THE HUMAN PANGENOME PROJECT

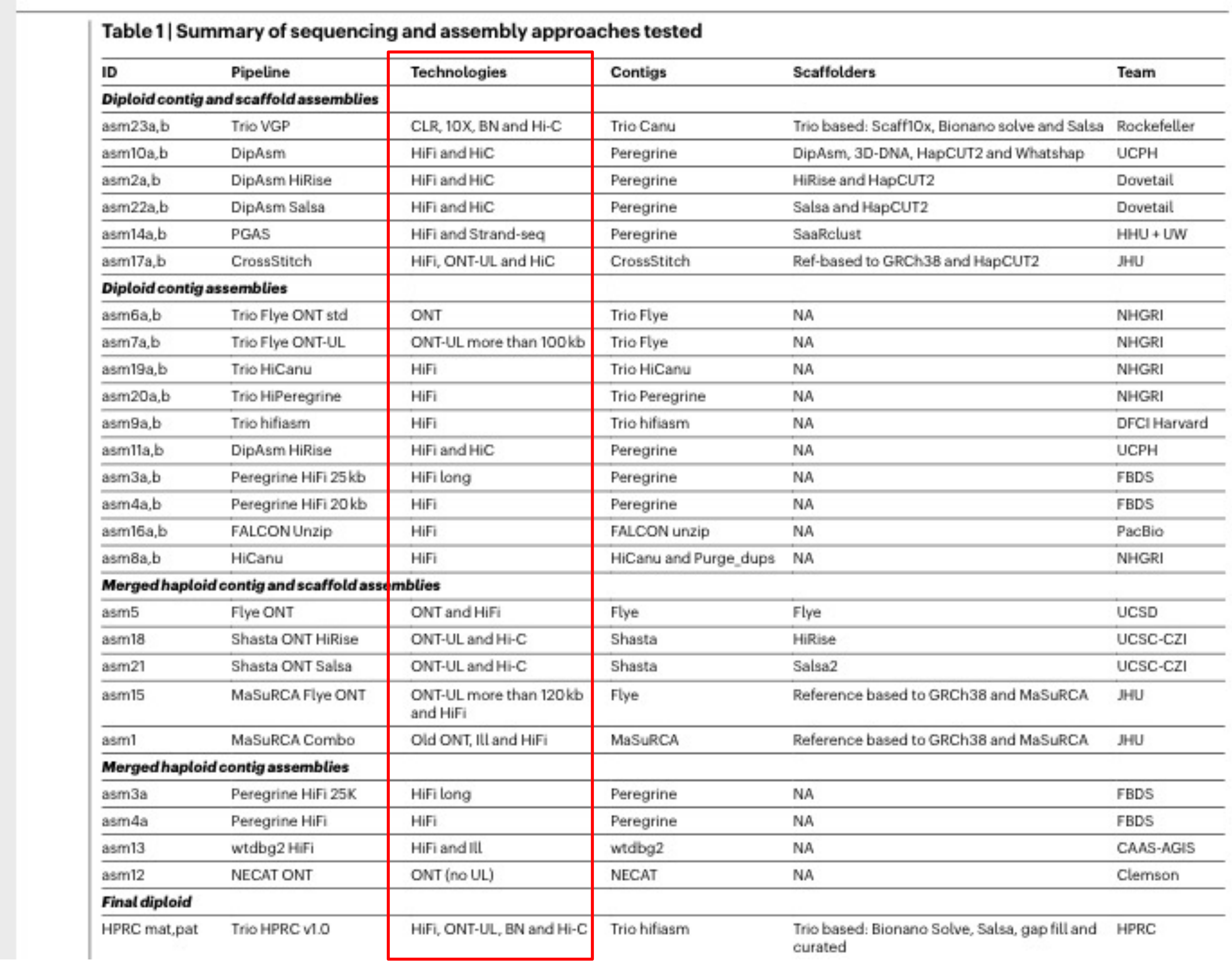


Table 1 | Summary of sequencing and assembly approaches tested

ID	Pipeline	Technologies	Contigs	Scaffolders	Team
Diploid contig and scaffold assemblies					
asm23a,b	Trio VGP	CLR, 10X, BN and Hi-C	Trio Canu	Trio based: Scaff10x, Bionano solve and Salsa	Rockefeller
asm10a,b	DipAsm	HiFi and HiC	Peregrine	DipAsm, 3D-DNA, HapCUT2 and Whatshap	UCPH
asm2a,b	DipAsm HiRise	HiFi and HiC	Peregrine	HiRise and HapCUT2	Dovetail
asm22a,b	DipAsm Salsa	HiFi and HiC	Peregrine	Salsa and HapCUT2	Dovetail
asm14a,b	PGAS	HiFi and Strand-seq	Peregrine	SaaRclust	HHU + UW
asm17a,b	CrossStitch	HiFi, ONT-UL and HiC	CrossStitch	Ref-based to GRCh38 and HapCUT2	JHU
Diploid contig assemblies					
asm6a,b	Trio Flye ONT std	ONT	Trio Flye	NA	NHGRI
asm7a,b	Trio Flye ONT-UL	ONT-UL more than 100 kb	Trio Flye	NA	NHGRI
asm19a,b	Trio HiCanu	HiFi	Trio HiCanu	NA	NHGRI
asm20a,b	Trio HiPeregrine	HiFi	Trio Peregrine	NA	NHGRI
asm9a,b	Trio hifiasm	HiFi	Trio hifiasm	NA	DFCI Harvard
asm11a,b	DipAsm HiRise	HiFi and HiC	Peregrine	NA	UCPH
asm3a,b	Peregrine HiFi 25 kb	HiFi long	Peregrine	NA	FBDS
asm4a,b	Peregrine HiFi 20 kb	HiFi	Peregrine	NA	FBDS
asm16a,b	FALCON Unzip	HiFi	FALCON unzip	NA	PacBio
asm8a,b	HiCanu	HiFi	HiCanu and Purge_dups	NA	NHGRI
Merged haploid contig and scaffold assemblies					
asm5	Flye ONT	ONT and HiFi	Flye	Flye	UCSD
asm18	Shasta ONT HiRise	ONT-UL and Hi-C	Shasta	HiRise	UCSC-CZI
asm21	Shasta ONT Salsa	ONT-UL and Hi-C	Shasta	Salsa2	UCSC-CZI
asm15	MaSuRCA Flye ONT	ONT-UL more than 120 kb and HiFi	Flye	Reference based to GRCh38 and MaSuRCA	JHU
asm1	MaSuRCA Combo	Old ONT, Ill and HiFi	MaSuRCA	Reference based to GRCh38 and MaSuRCA	JHU
Merged haploid contig assemblies					
asm3a	Peregrine HiFi 25K	HiFi long	Peregrine	NA	FBDS
asm4a	Peregrine HiFi	HiFi	Peregrine	NA	FBDS
asm13	wtdbg2 HiFi	HiFi and Ill	wtdbg2	NA	CAAS-AGIS
asm12	NECAT ONT	ONT (no UL)	NECAT	NA	Clemson
Final diploid					
HPRC mat,pat	Trio HPRC v1.0	HiFi, ONT-UL, BN and Hi-C	Trio hifiasm	Trio based: Bionano Solve, Salsa, gap fill and curated	HPRC

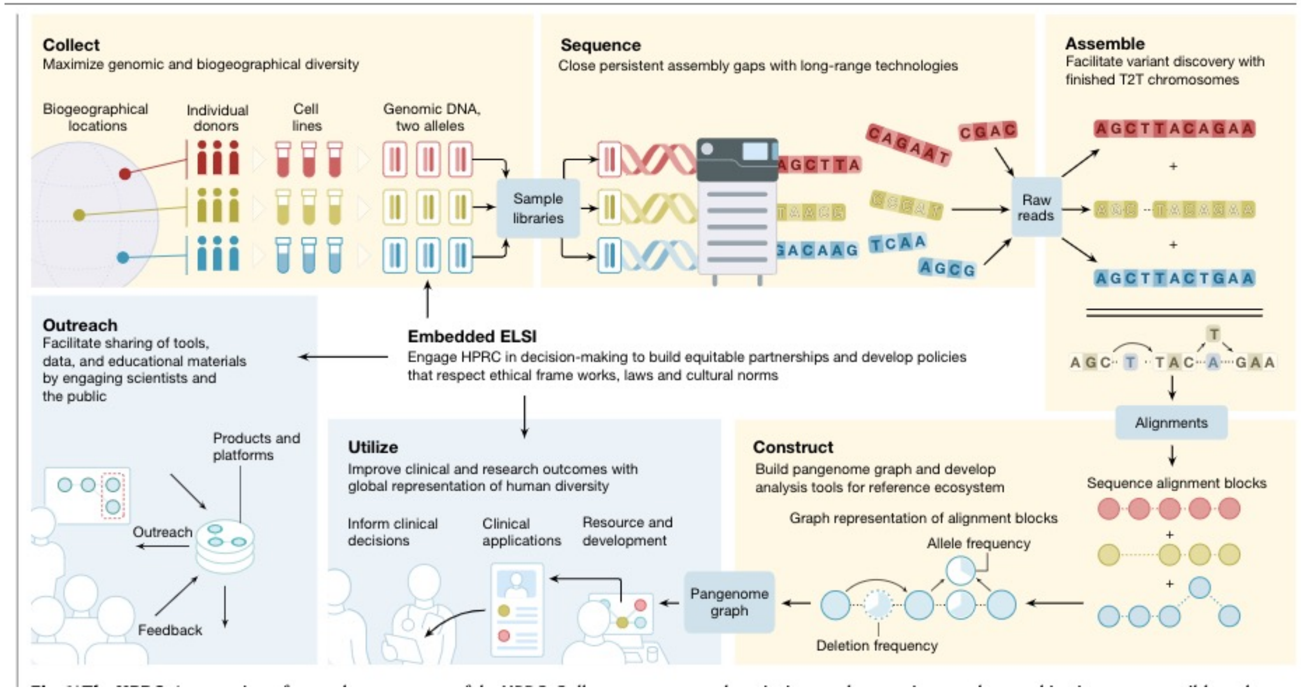
THE HUMAN PANGENOME PROJECT

Wang et al. The Human Pangenome Project: a global resource to map genomic diversity. *Nature* 2022

Goals of the Human Pangenome Project

- To generate the highest quality phased genomes possible, prioritize the use of long-read and long-range technologies for assemblies, with haplotype-aware algorithms

- As long-read sequencing costs fall and pangenome methods evolve, we predict that patient samples will probably be sequenced using long-read technology.



Collect
Maximize genomic and biogeographical diversity
Biogeographical locations
Individual donors
Cell lines
Genomic DNA, two alleles
Sample libraries
Sequence
Close persistent assembly gaps with long-range technologies
AGCTTA
CAGAAT
CGAC
TAAGG
GGGAT
GACAAG
TCAA
AGCG
Raw reads
Assemble
Facilitate variant discovery with finished T2T chromosomes
AGCTTACAGAA
+
AGC TACAGAA
+
AGCTTACTGAA
AGC T TAC A GAA
T
Alignments
Sequence alignment blocks
Outreach
Facilitate sharing of tools, data, and educational materials by engaging scientists and the public
Products and platforms
Outreach
Feedback
Embedded ELSI
Engage HPRC in decision-making to build equitable partnerships and develop policies that respect ethical frame works, laws and cultural norms
Utilize
Improve clinical and research outcomes with global representation of human diversity
Inform clinical decisions
Clinical applications
Resource and development
Construct
Build pangenome graph and develop analysis tools for reference ecosystem
Graph representation of alignment blocks
Allele frequency
Deletion frequency
Pangenome graph

Summary

PacBio

- Maximum read length : 200 kb
- CCS sequencing (HiFI reads) :
  - Very low error rate, best genome assembly
  - Sequencing of cDNAs (resolution of alternative splicing)
  - Detection of modified DNA (6mA > 5mC)
  - cDNA :
    - RNA-seq
    - Efficient for splicing isoforms detection

Nanopore

- Very light sequencing system - portability
- Very long reads : maximum length > 1 Mb
- 10.4.1 flow cells: low error rate, accurate genome assembly
- Duplex sequencing may allow higher accuracy and challenge HiFi reads
- Detection of modified DNA (5mC, 6mA)
- Direct sequencing of RNA :
  - Direct RNA sequencing :
    - RNA-seq
    - splicing isoforms detection
    - Detection of modified RNA (6mA, pseudo U, etc..)