



LONG READS

Claude THERMES

PLATEFORME DE SÉQUENÇAGE I2BC

INSTITUT DE BIOLOGIE INTÉGRATIVE DE LA CELLULE

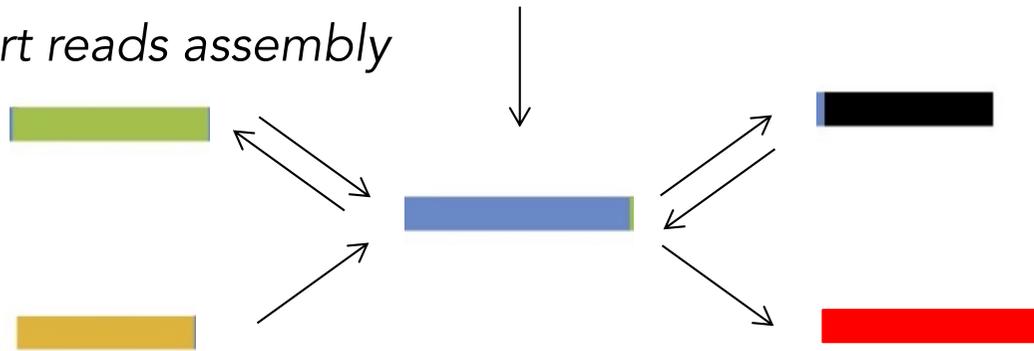
GIF-SUR-YVETTE

LONG-READS VERSUS SHORT-READS

Assembly of DNA fragments with repeated sequences



NGS short reads assembly



Several contigs → incomplete assembly, underestimation of repeats

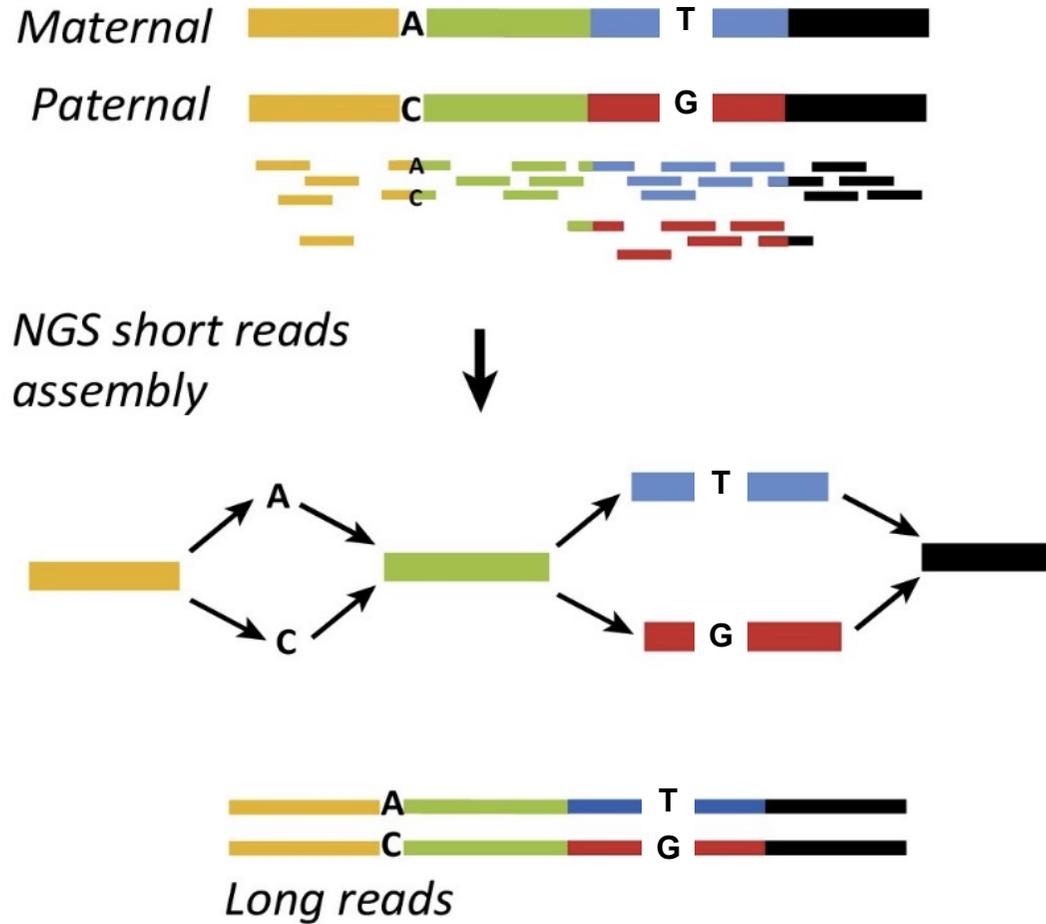
Long reads assembly



Long-reads (> 10 kb) allow assembly of large repeat-rich regions

LONG-READS VERSUS SHORT-READS

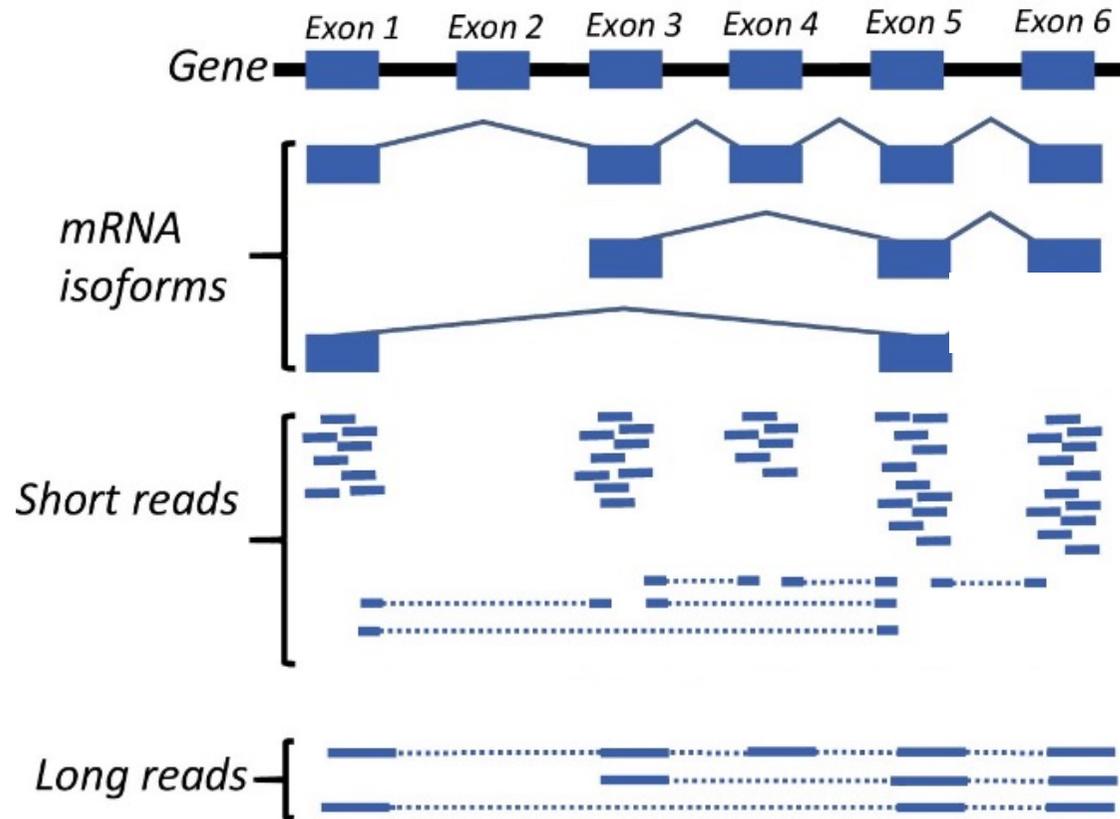
Haplotype phasing



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



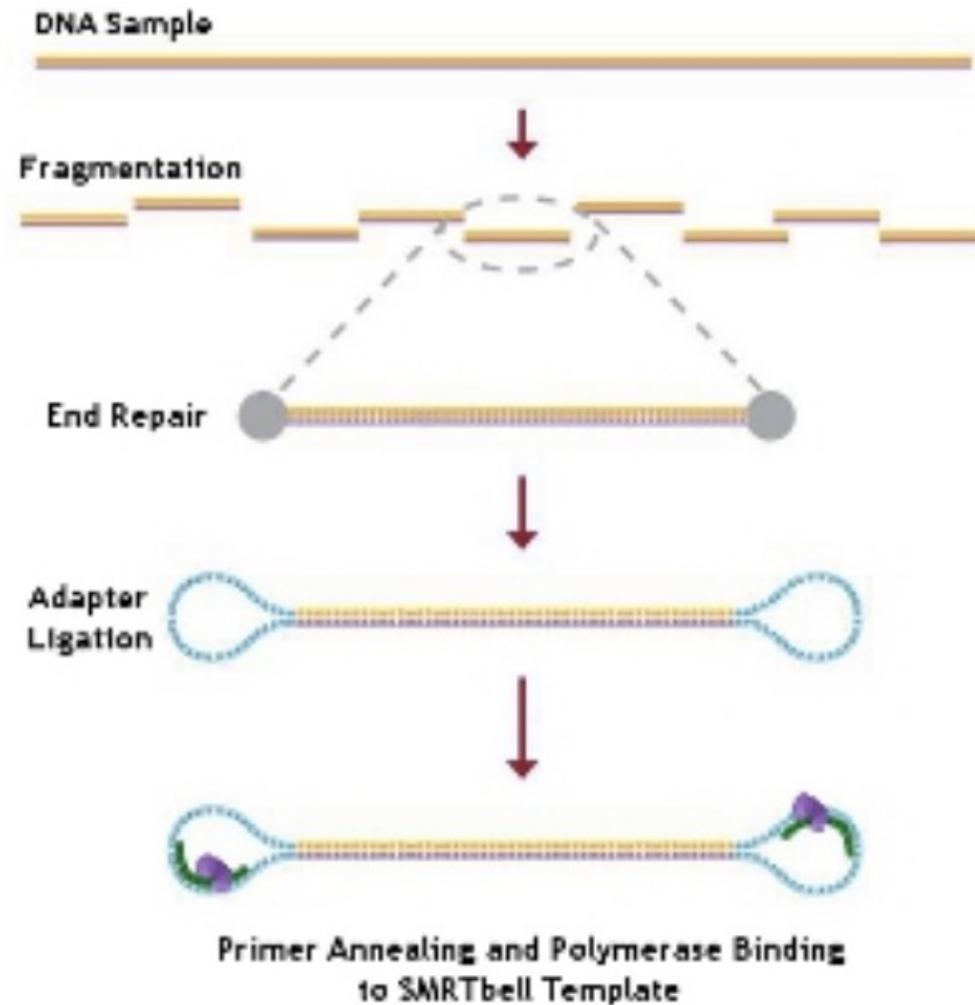
Sequel – Revio / Pacific Biosciences
Single molecules
Up to 200 kbp long



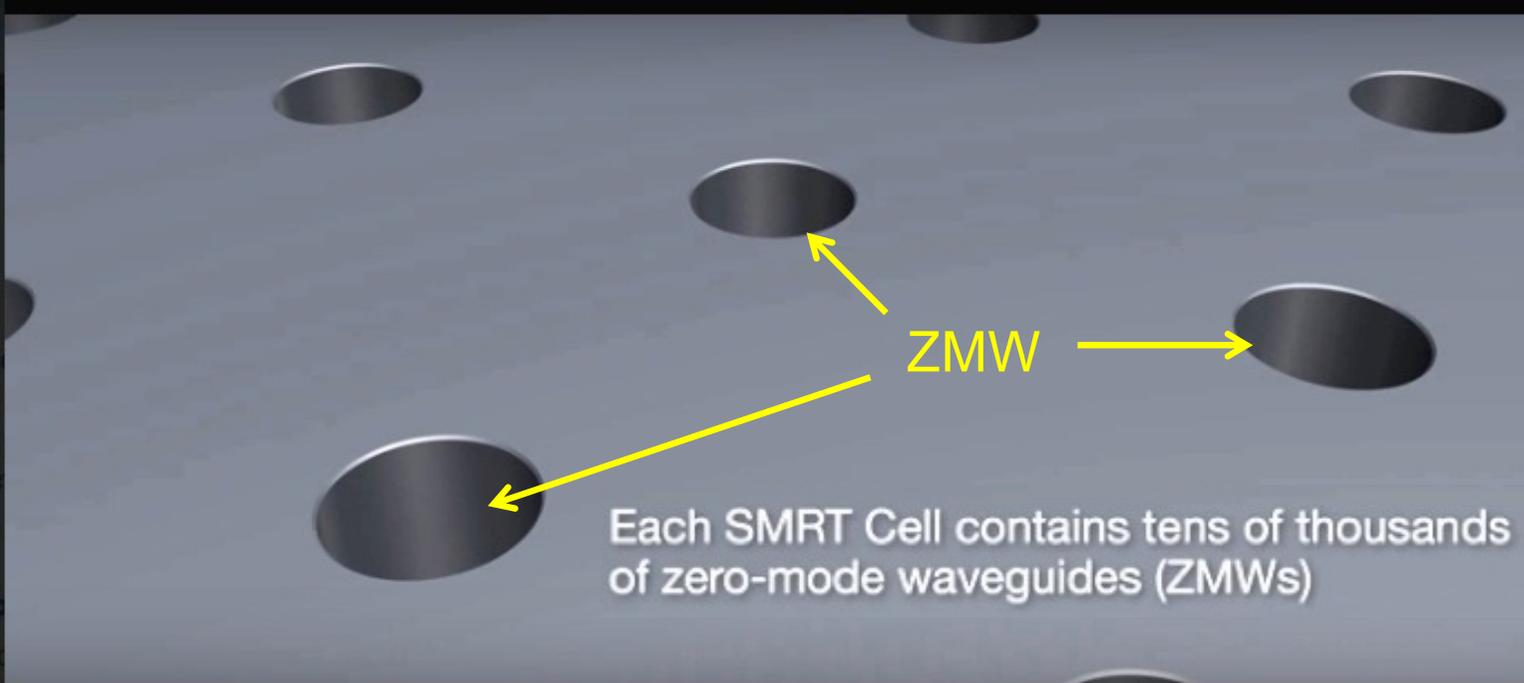
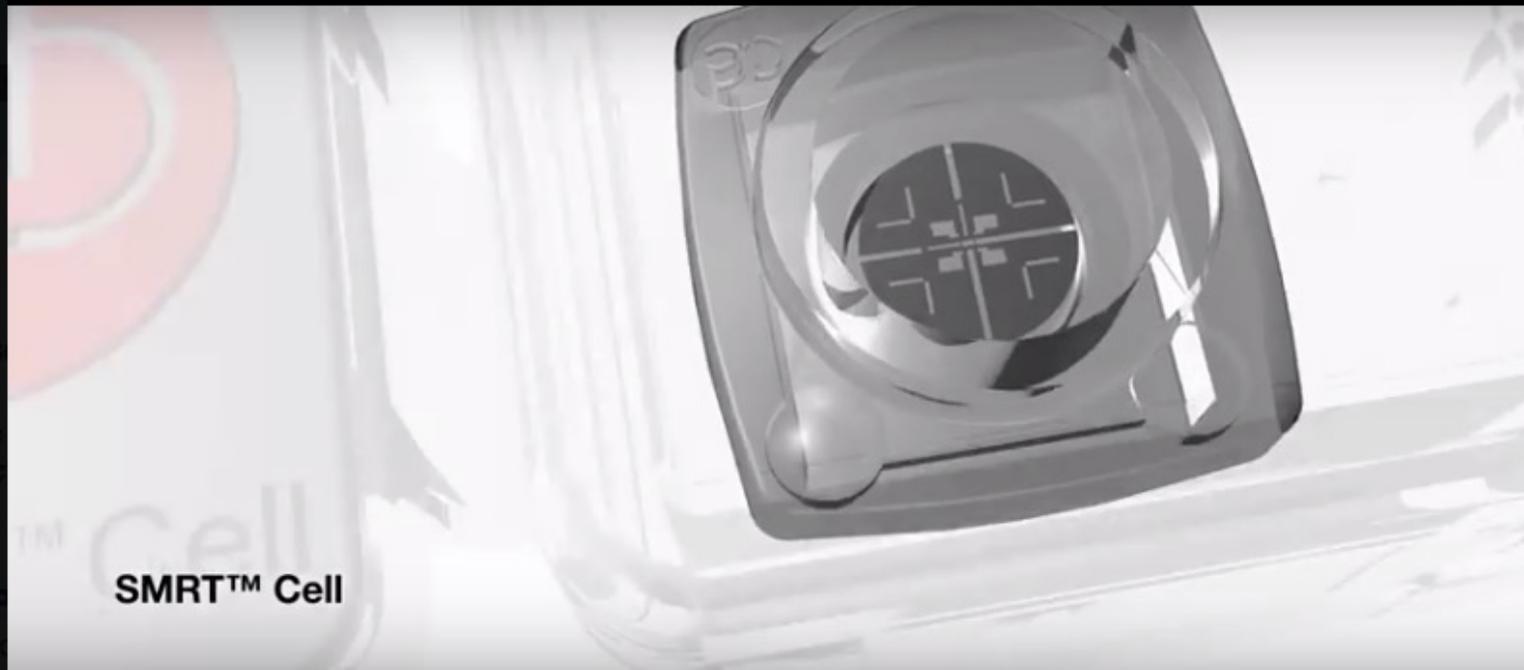
MinION – PromethION - Oxford Nanopore
Single molecules
> 1 Mbp long

— PacBio : Single Molecule Real Time (SMRT) sequencing —

PacBio DNA-seq library



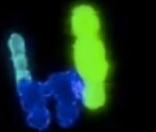
PACIFIC BIOSCIENCES



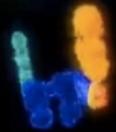
PACIFIC BIOSCIENCES

Phospholinked Nucleotides

A



C



G



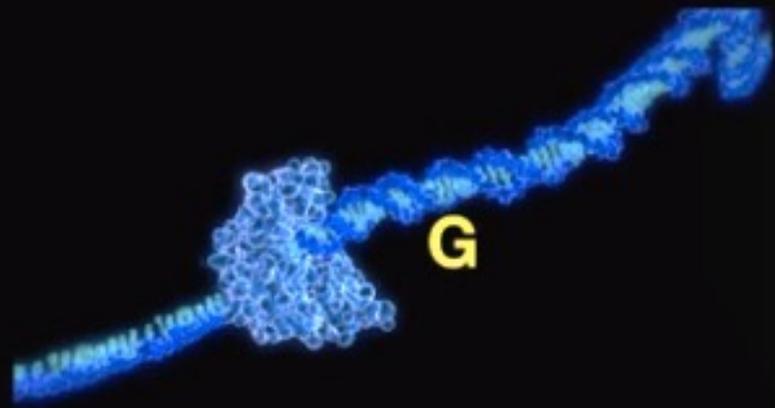
T



Phospholinked nucleotides are introduced into the ZMW chamber

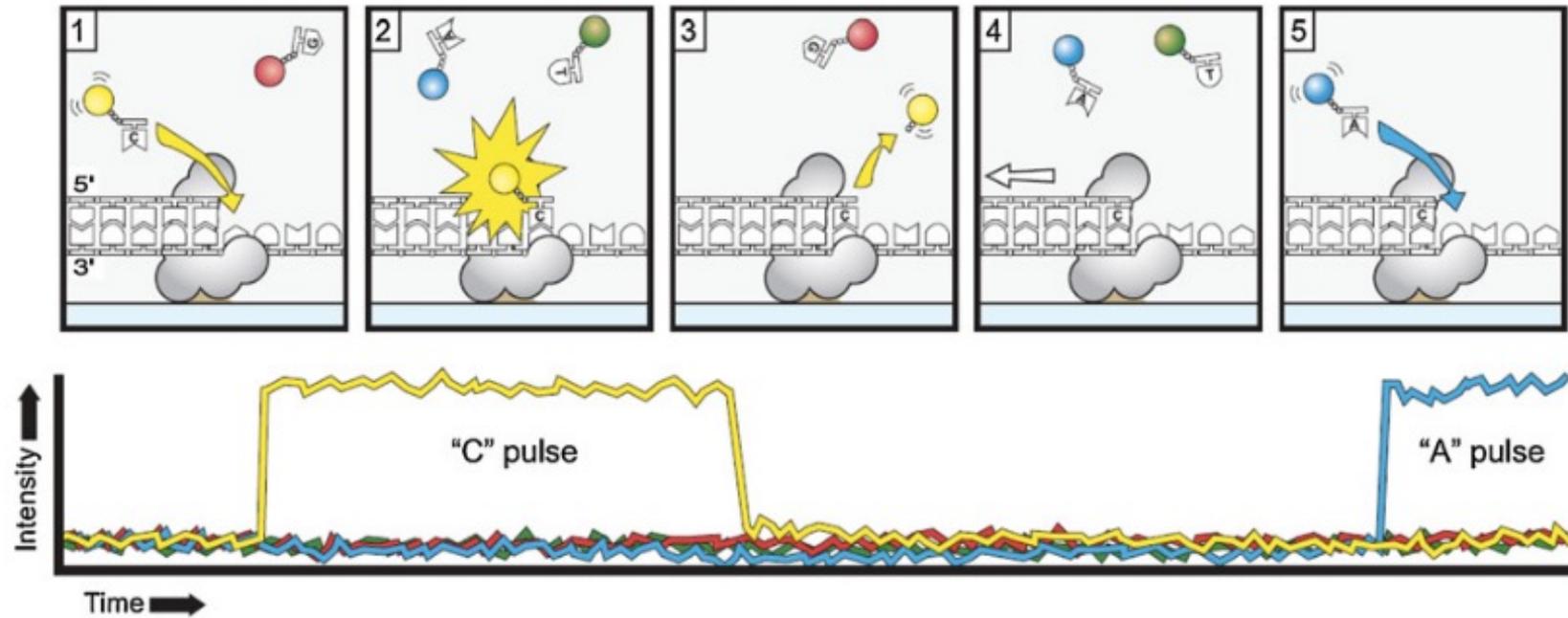


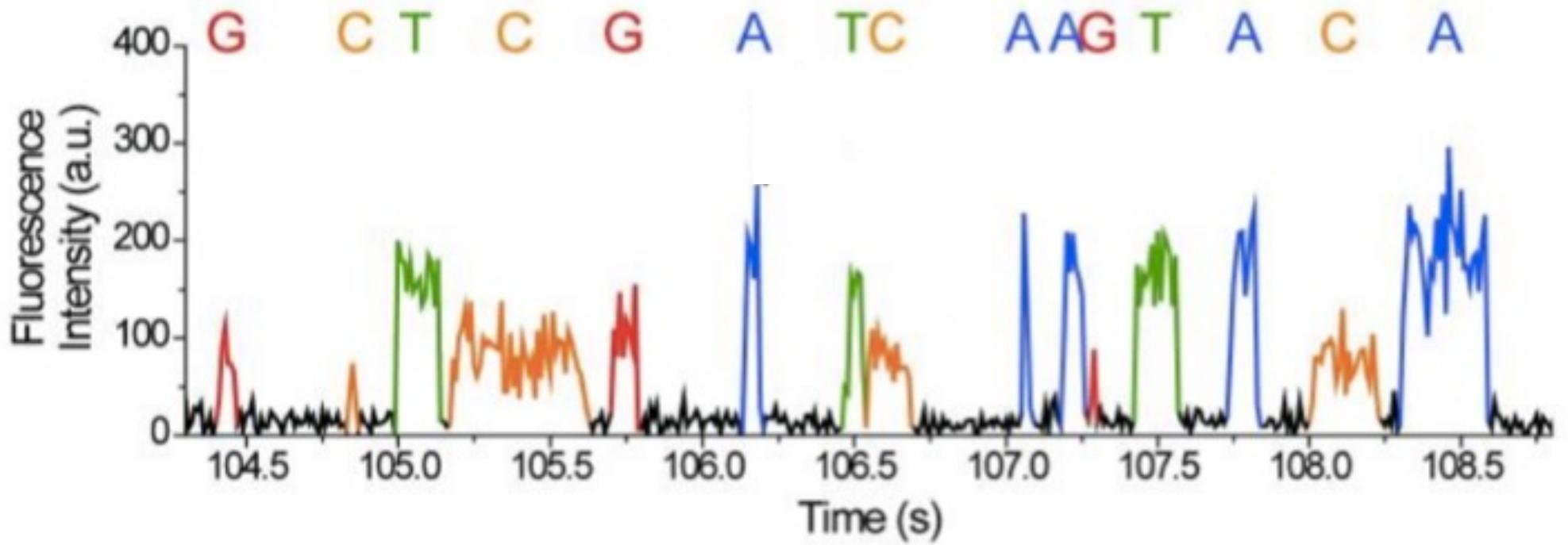
PACIFIC BIOSCIENCES



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

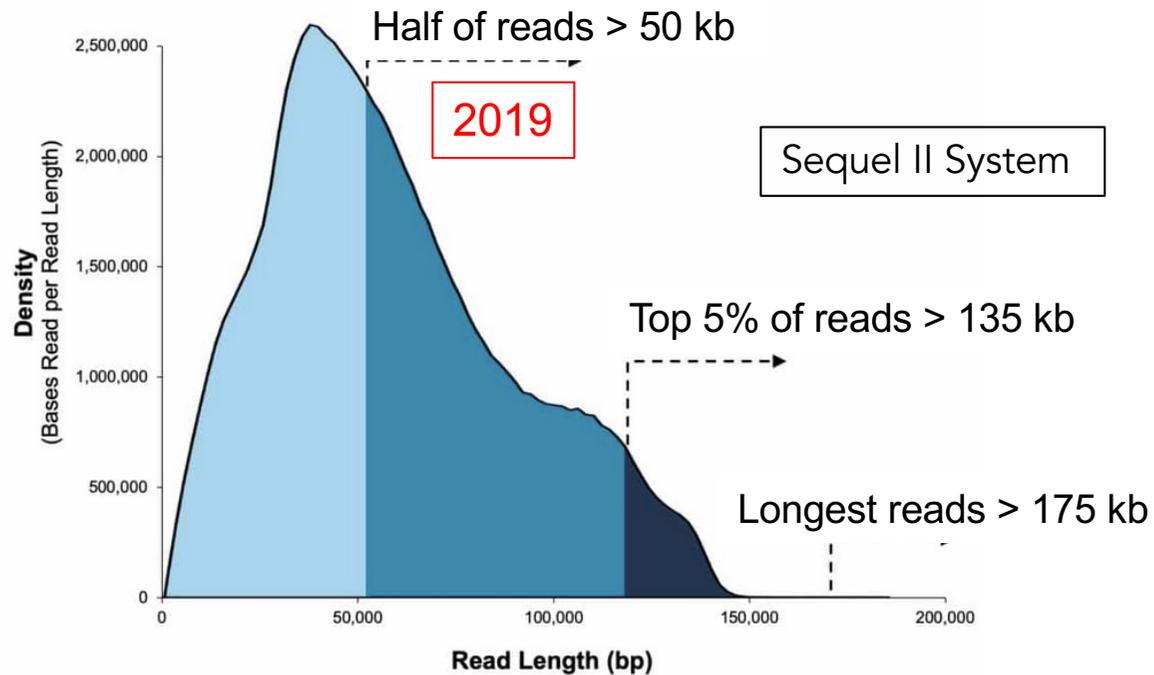
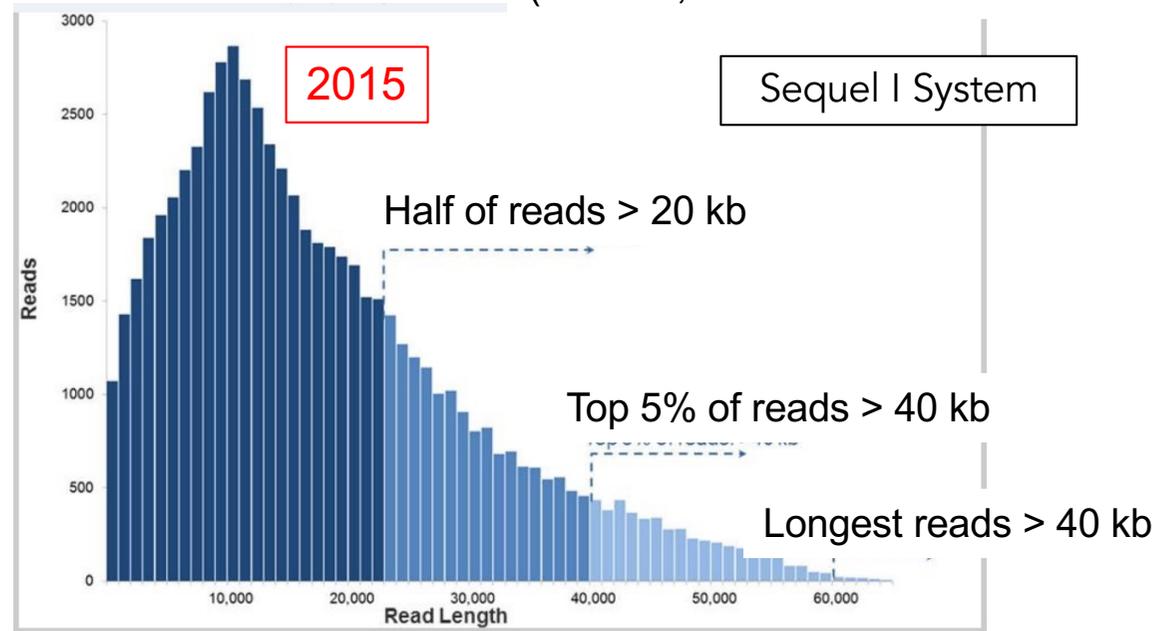




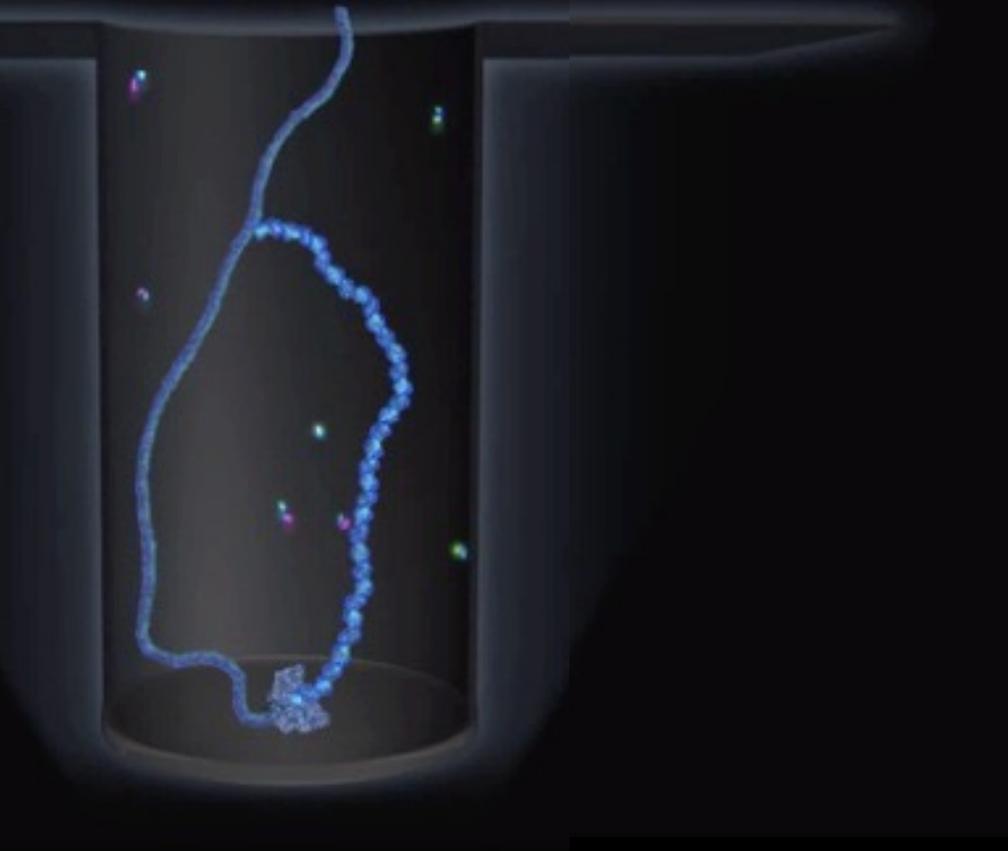
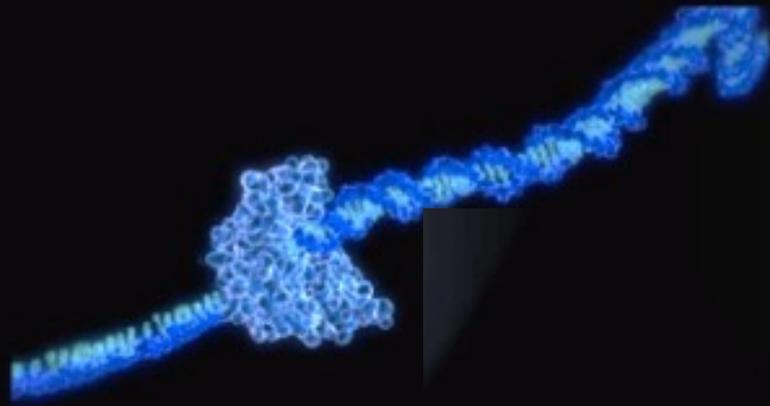


Length of PacBio reads

(Rhoads, *Genomics Proteomics Bioinformatics*, 2015)

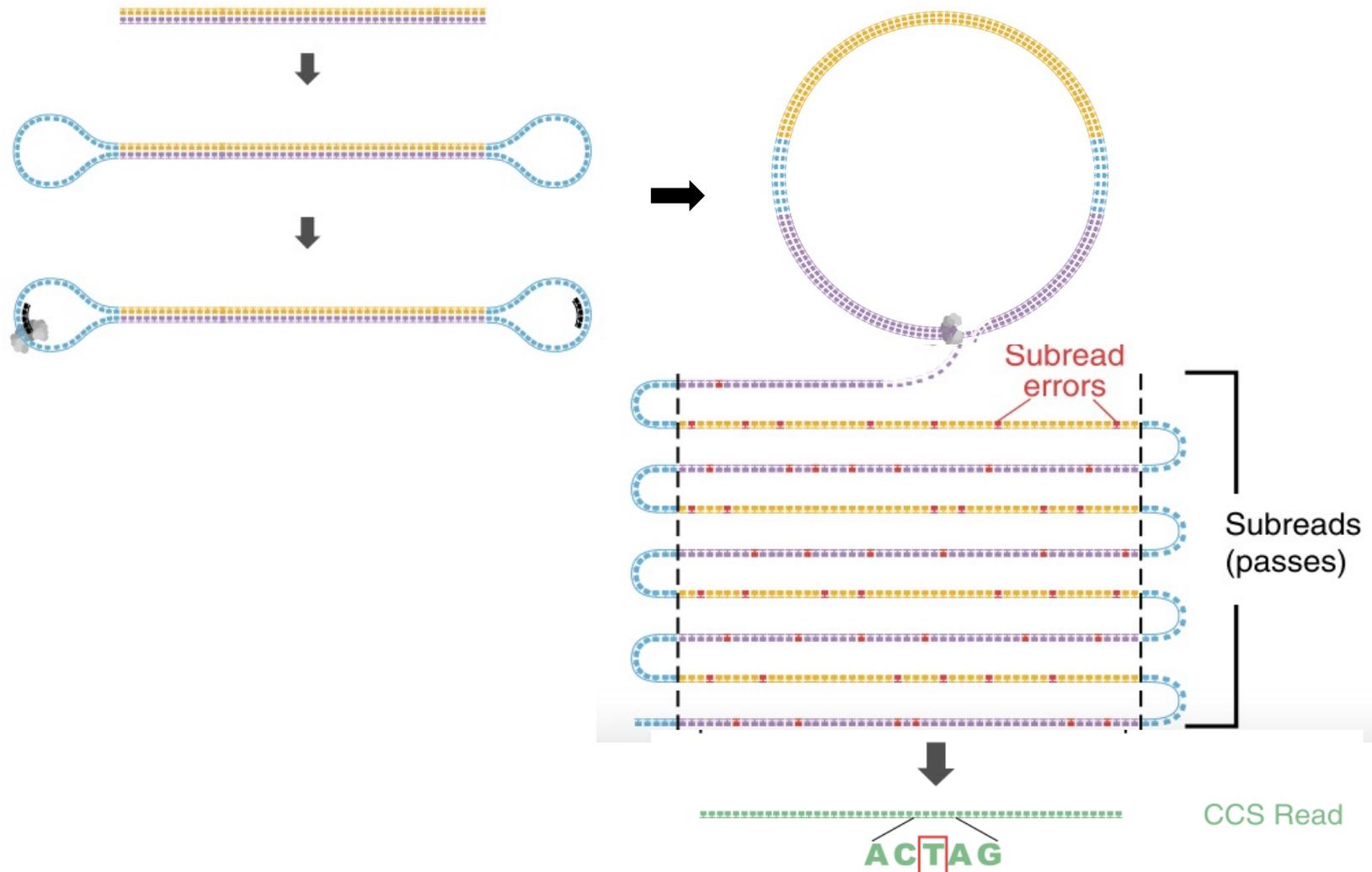


Improvement with new chemistry : Circular Consensus Sequence (CCS)

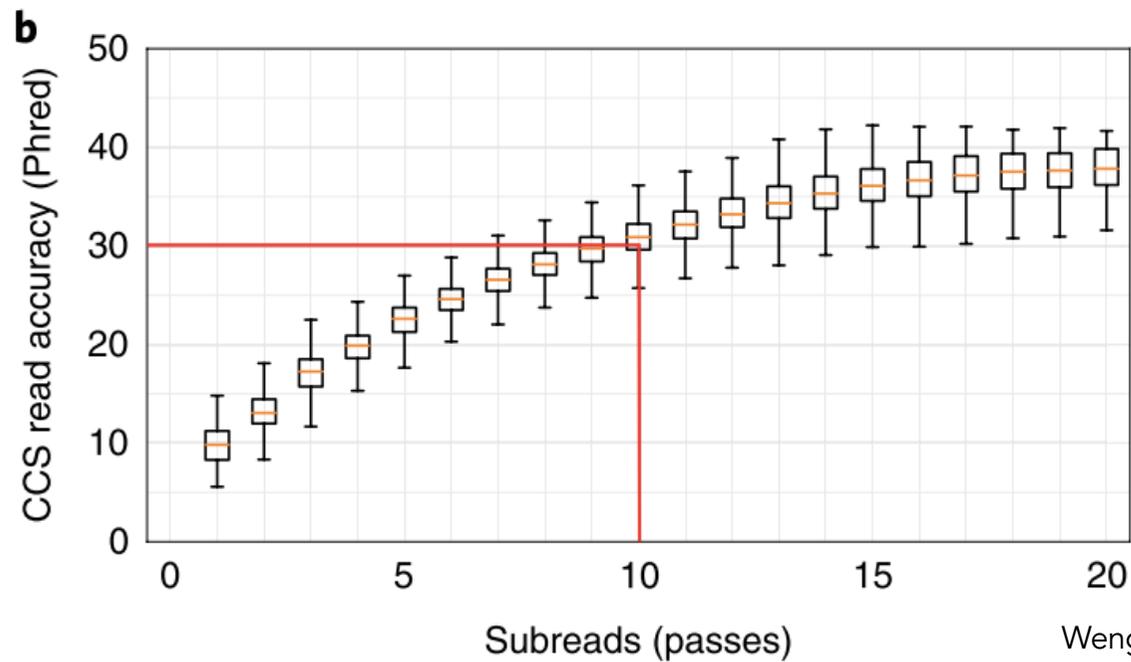
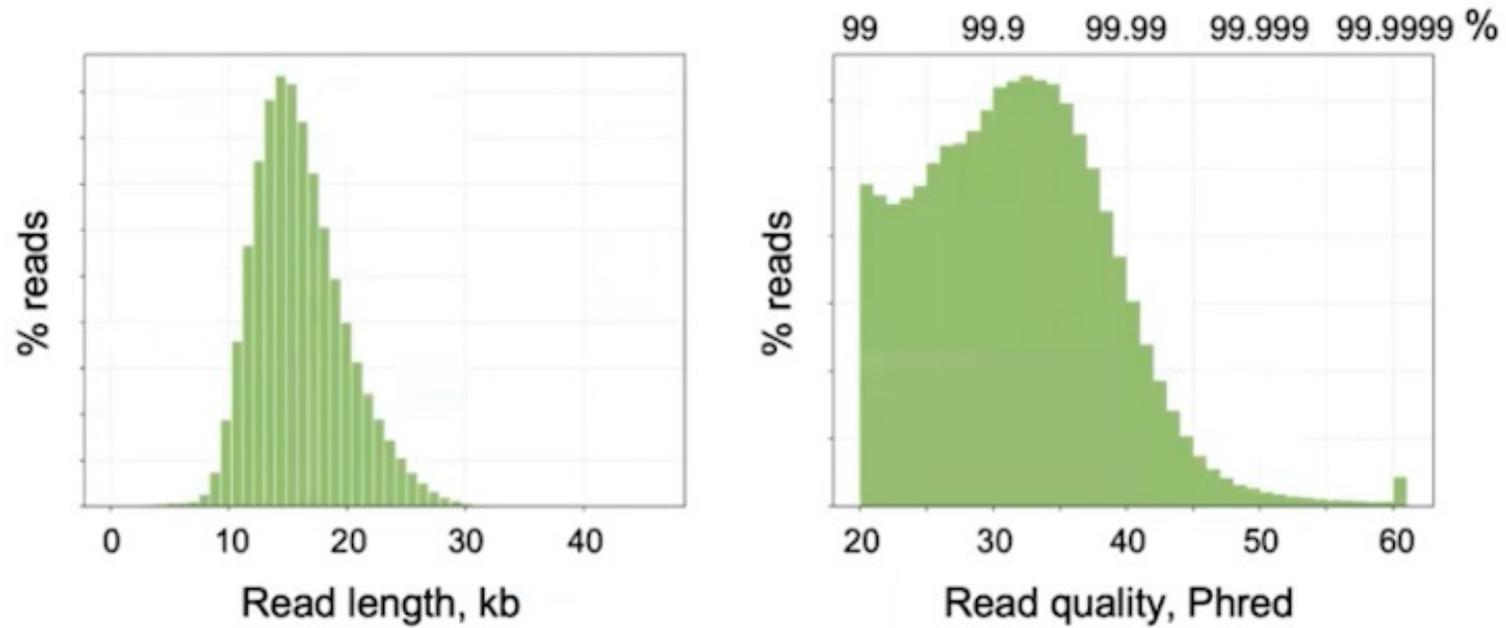


Circular consensus sequencing (CCS) reads are obtained when the SMRT bell template is replicated several times by the polymerase

— Circular Consensus Sequences (CCS): HIFI READS



— GENOME ASSEMBLY WITH CCS



GENOME ASSEMBLY WITH CCS

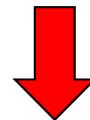
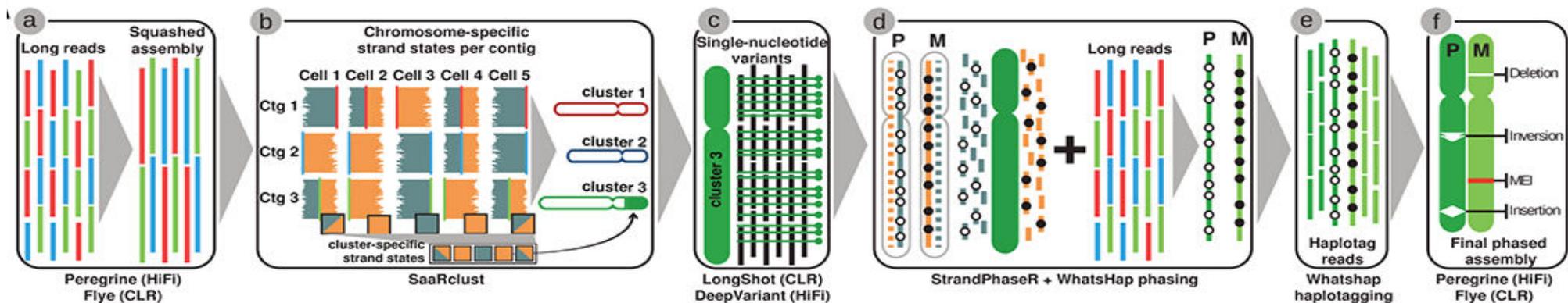
1 - Haplotype-resolved diverse human genomes and integrated analysis of structural variation
Ebert et al. *Science* 2021

New methodology that combines :

- Long-reads PacBio : CLR (continuous long reads) and CCS reads (20X) generated with **Sequel II System**
- Strand-seq Illumina

Methodology

- generation of a non-haplotype-resolved clustered assembly
- clustering of assembled contigs into "chromosome" clusters based on Strand-seq Illumina
- calling of single-nucleotide variants (SNVs) relative to the clustered assembly
- chromosome-wide phasing
- tagging of input long reads by haplotype
- phased genome assembly based on haplo-tagged long reads

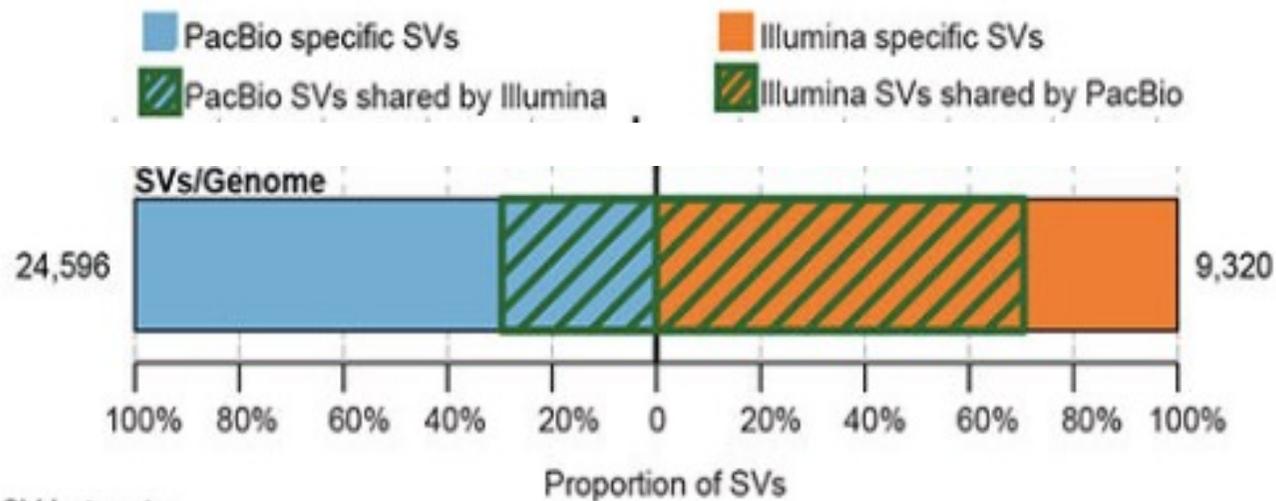


GENOME ASSEMBLY WITH CCS

1 - Haplotype-resolved diverse human genomes and integrated analysis of structural variation
Ebert et al. *Science* 2021

64 ASSEMBLED HAPLOTYPES FROM 32 DIVERSE HUMAN GENOMES

Comparison with GRCh38 -> 107,590 structural variants of which 68% not discovered by short-reads



- Two important types of regions **not fully resolved** :
 - Gene-rich regions in segmental duplications
 - Larger repeat-rich regions such as centromeres
- “ Recent advances coupling HiFi and ultra-long-read Nanopore may solve these more complex regions ”

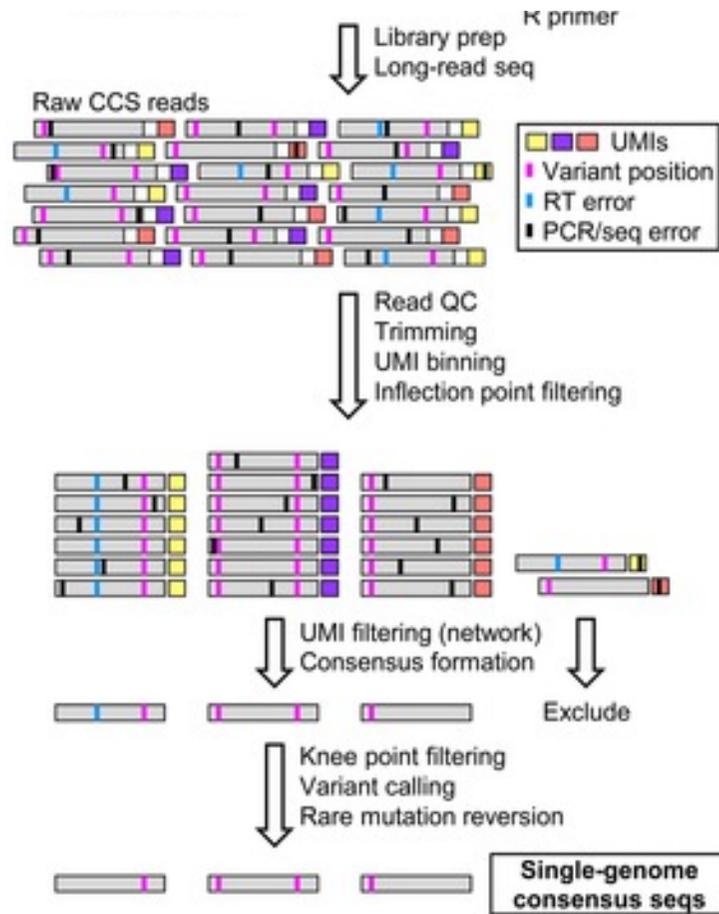
GENOME ASSEMBLY WITH CCS

2 - High-throughput, single-copy sequencing reveals SARS-CoV-2 spike variants coincident with mounting humoral immunity during acute COVID-19, Ko S.H. et al. *PLOS Pathogens* 2021

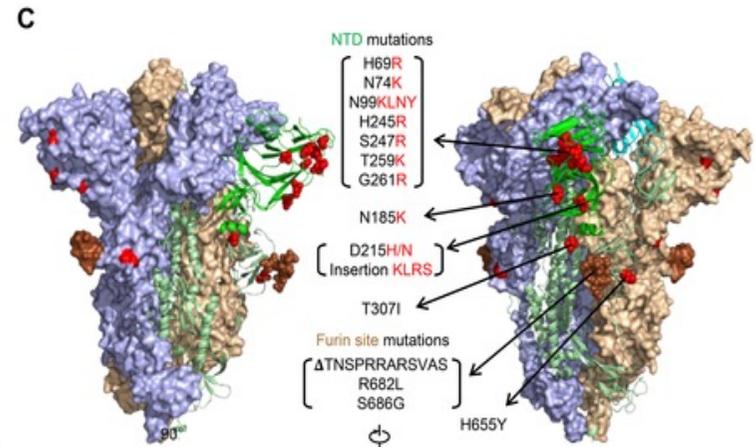
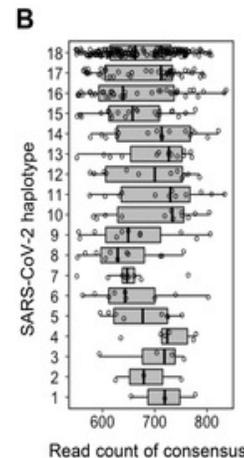
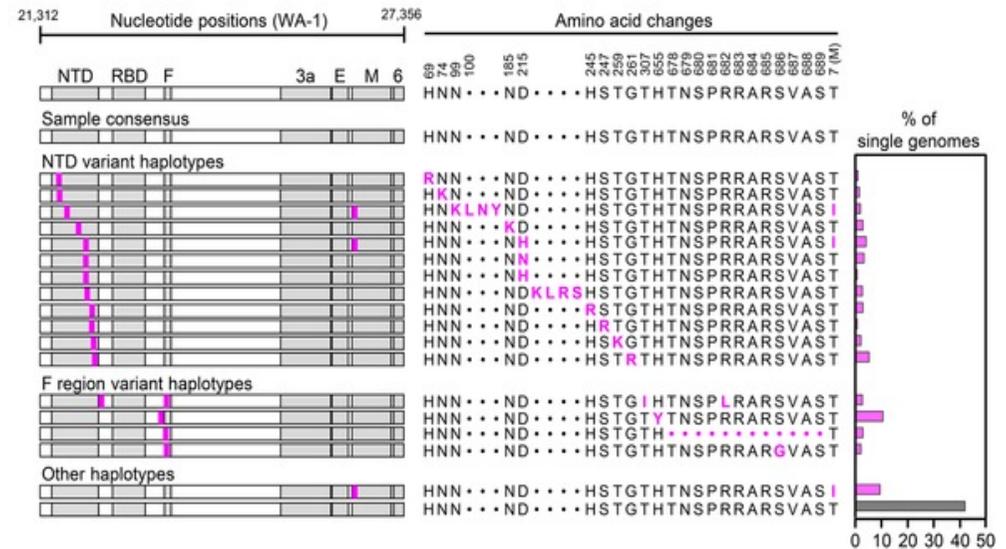
Single-genome amplification and sequencing (SGS) approach

Study of intra-individual evolution of SARS-CoV-2 : for each sample -> multiple sequences representing virus diversity

Analysis of CCS reads

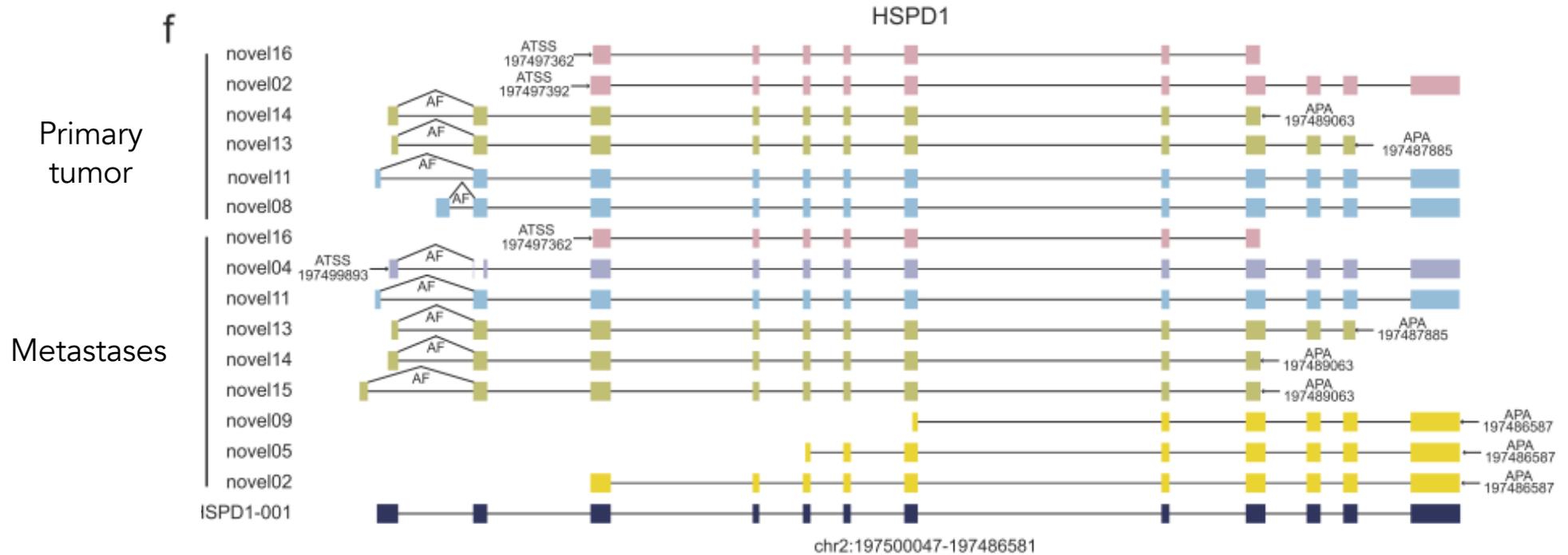


Each sequence corresponds to a single viral genome



PacBio cDNA SEQUENCING WITH CCS

1 - Hybrid full-length transcriptome in metastatic ovarian cancer
Jing et al. *Oncogene* 2019

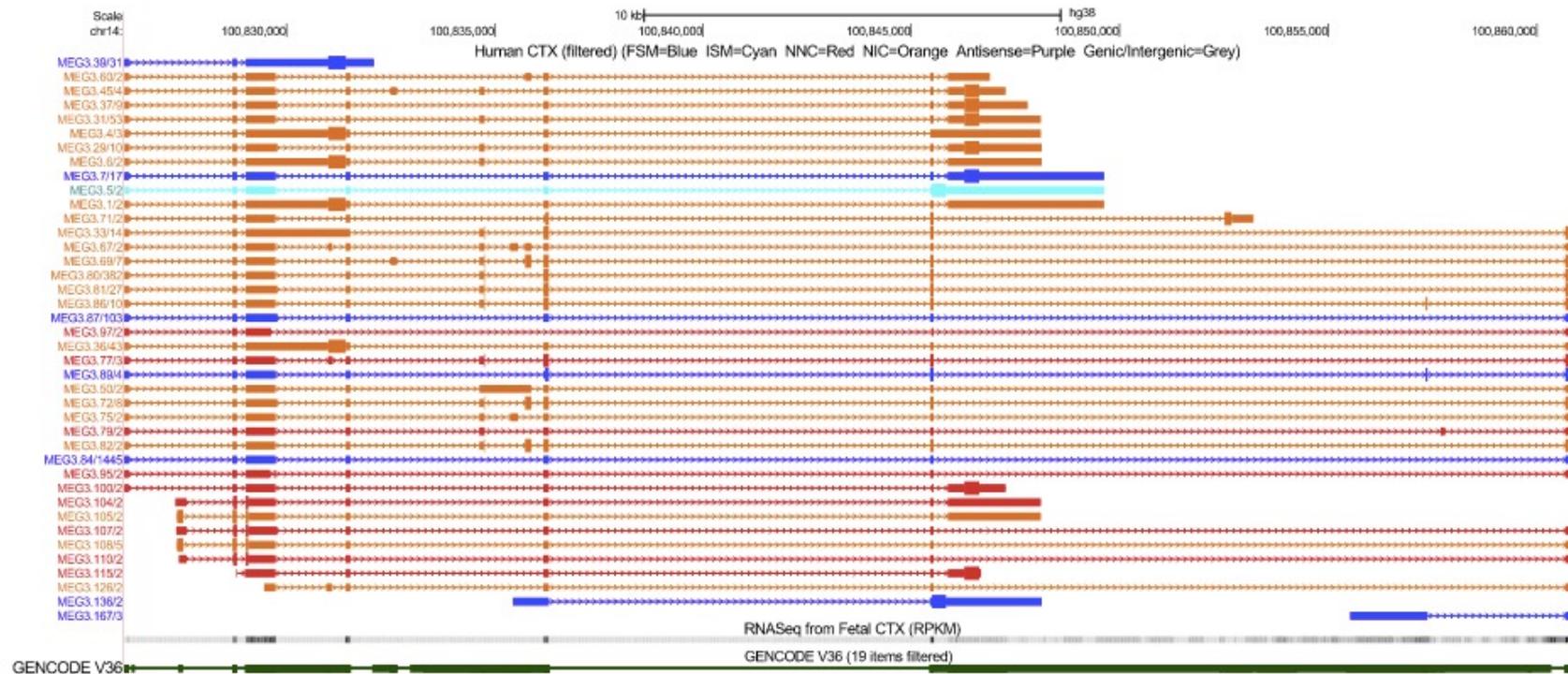


Long-read full-length transcriptome analysis improves molecular diagnostic

PacBio cDNA SEQUENCING WITH CCS

2 - 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 WITH CCS

2 - 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

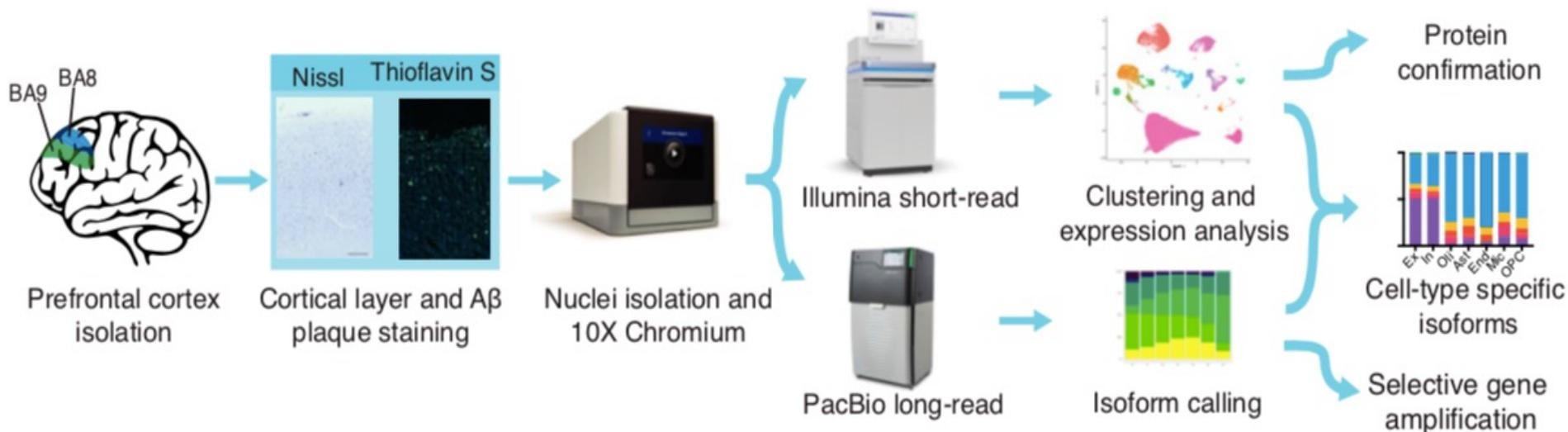
Description	Human Cortex		
	AD	SZ	Autism
Disease-associated genes	62	339	393
Detected disease-associated genes ("Detected")	33	288	317
Total Number of Transcripts	128	967	1042
Number and % of Annotated Transcripts	72 (56.25%)	558 (57.7%)	669 (64.2%)
Number and % of Novel Transcripts	56 (43.75%)	409 (42.3%)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

SINGLE CELL PacBio cDNA 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

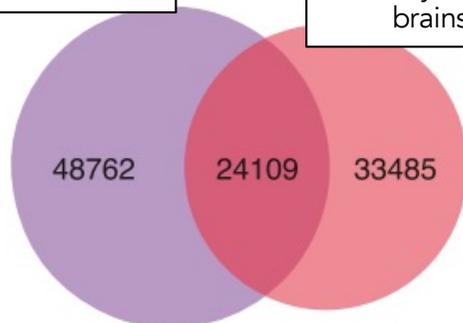


New splicing isoforms :

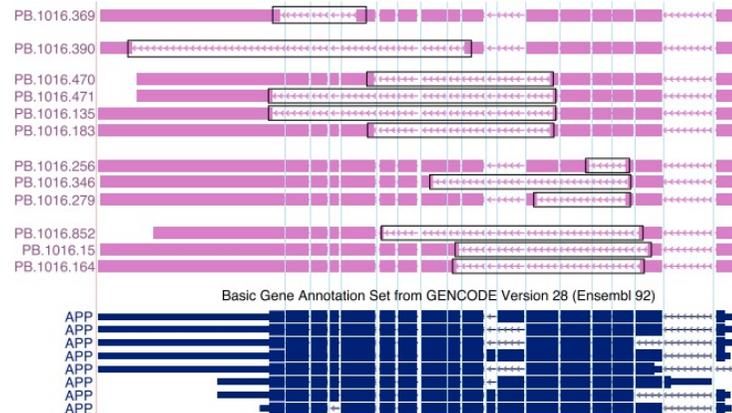
- new splice sites
- novel exon junctions
- entirely new exons
- intron retention

Control brains

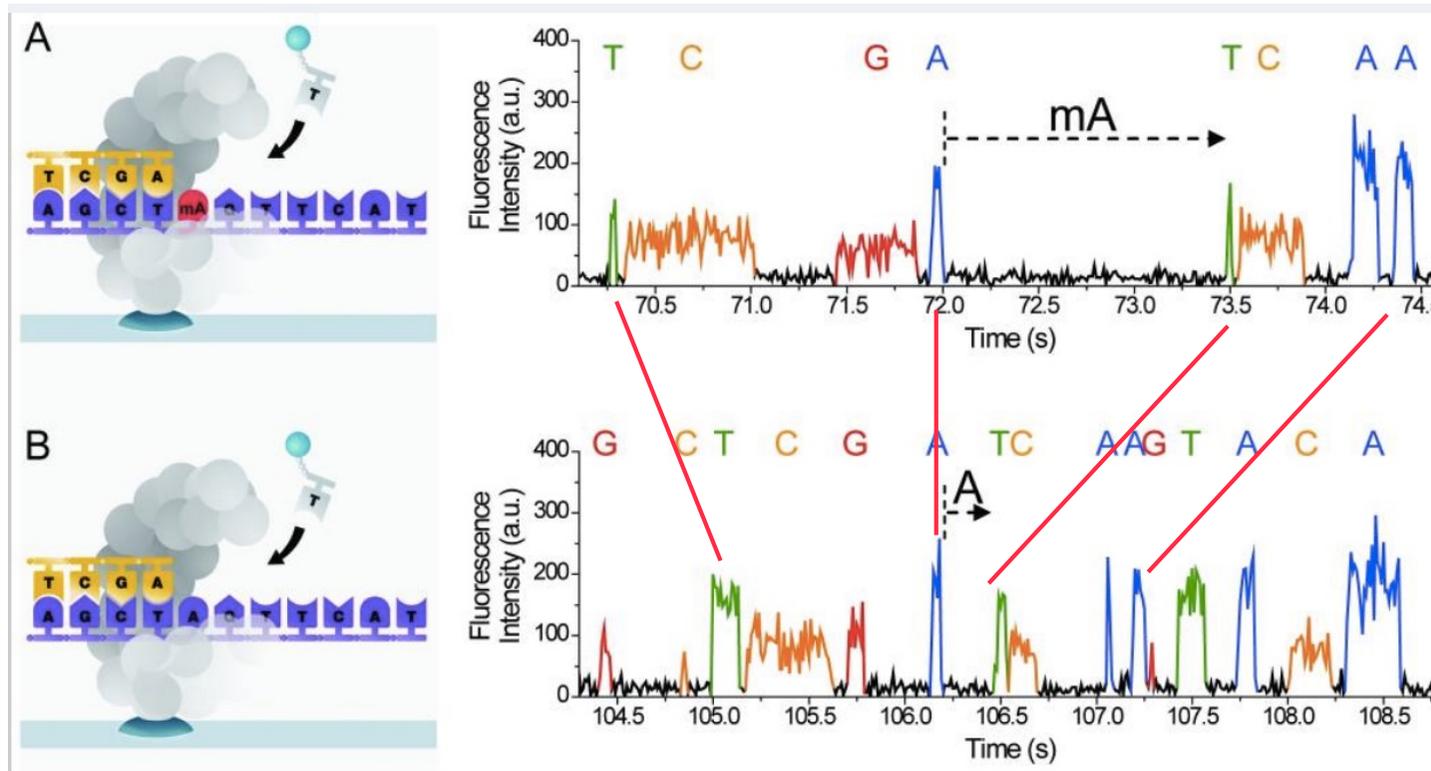
Down syndrome brains



Amyloid precursor protein (Alzheimer's disease gene)



DETECTION OF MODIFIED DNA BASES



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

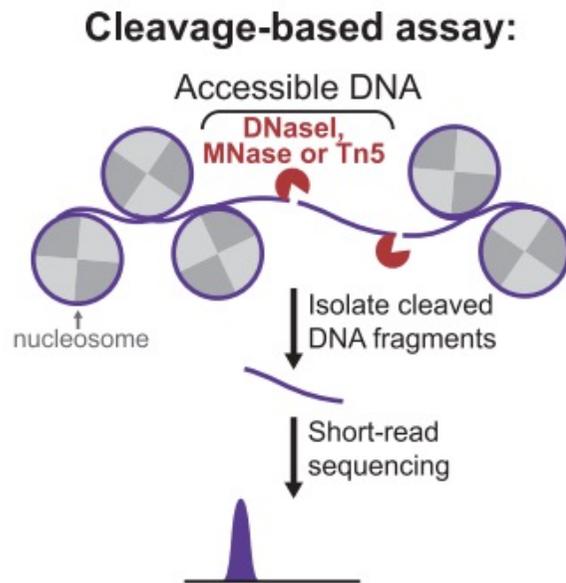
Detection of 5mA with strong influence of sequence contexts : requires high coverage

Feng et al. *PLOS Comput Biol* 2013

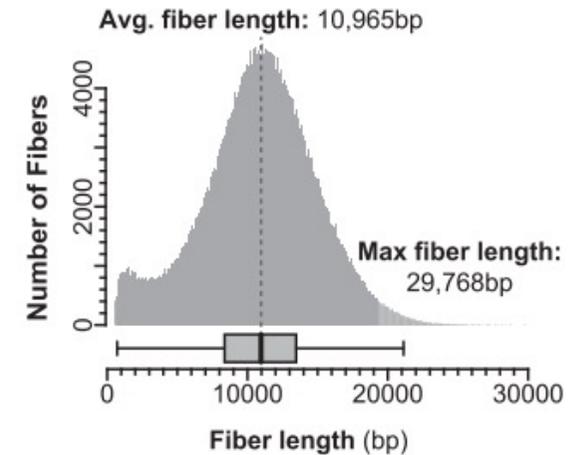
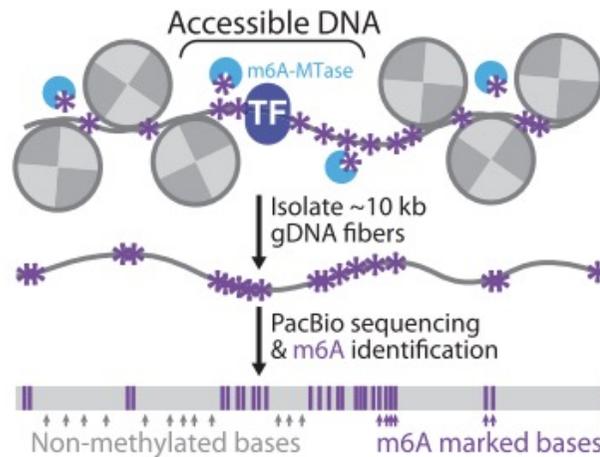
DETECTION OF DNA m6A WITH CCS

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

DnaseI-seq.

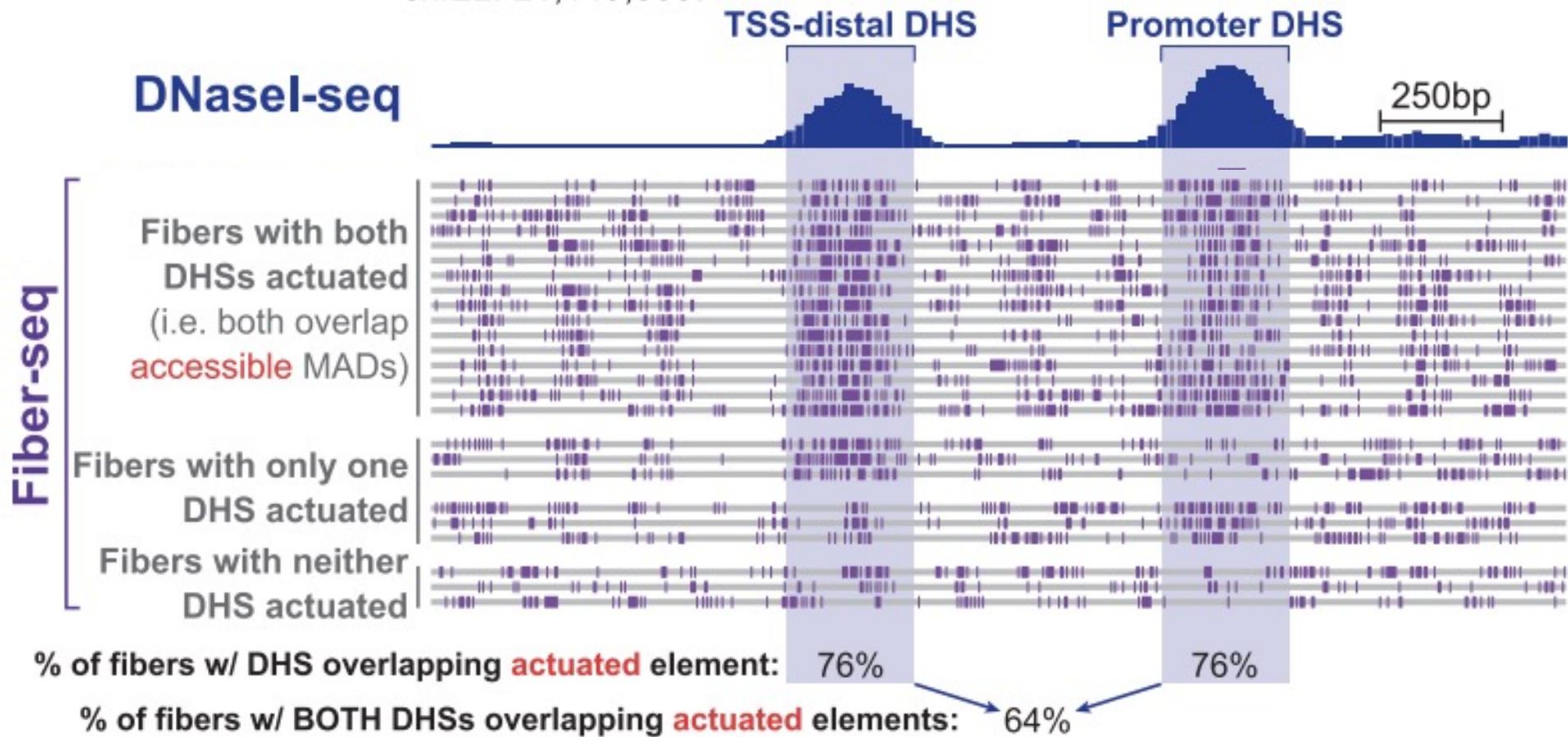


Fiber-seq.



DETECTION OF DNA m6A WITH CCS

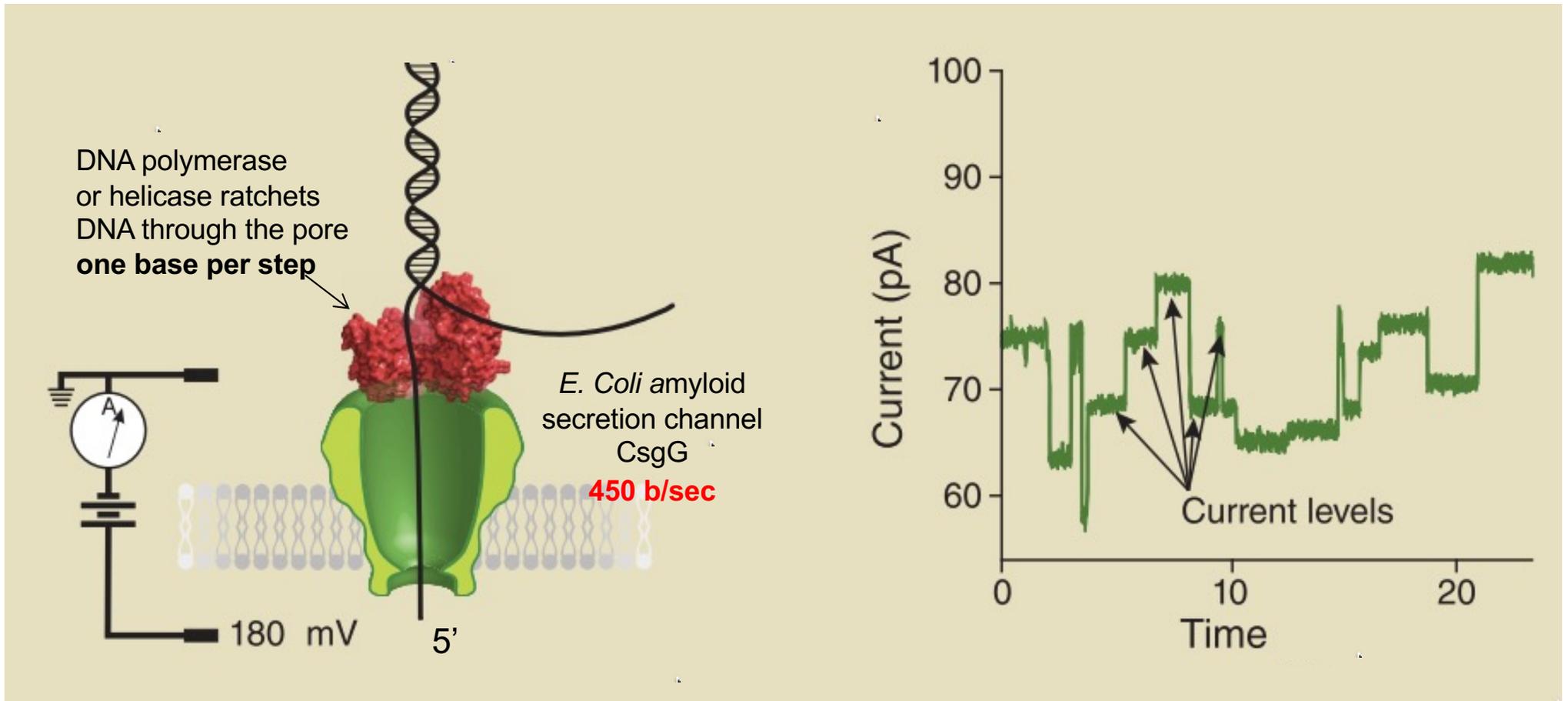
Single-molecule regulatory architectures captured by chromatin fiber sequencing
Stergachis et al. *Science* (2020)



Next Generation Sequencing



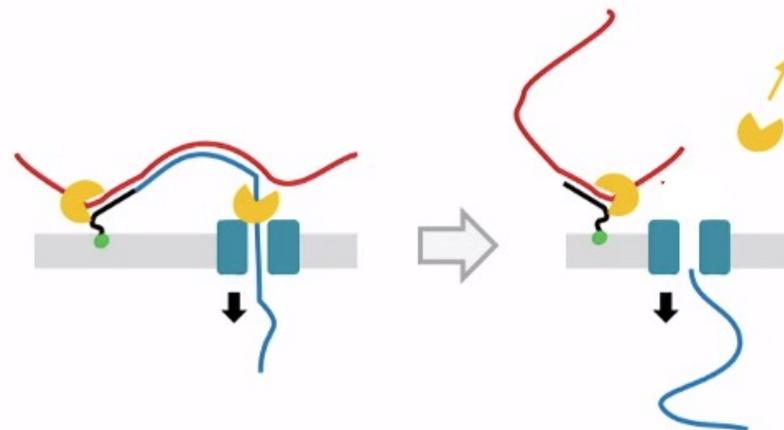
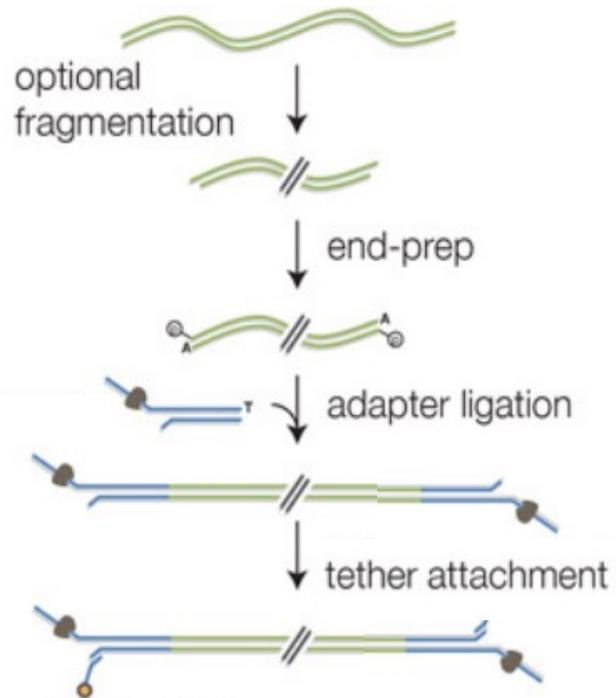
BASIC CONCEPTS



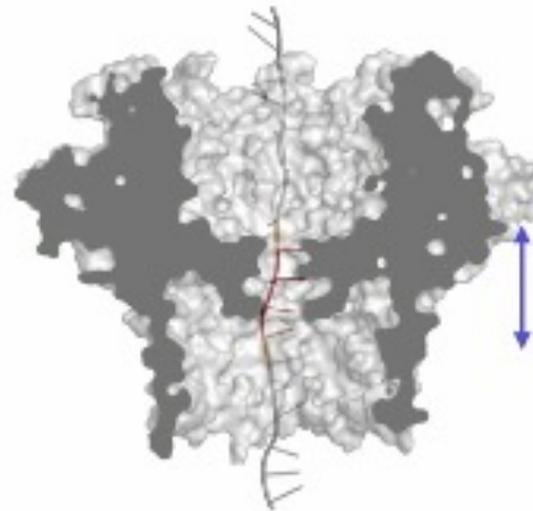
SEQUENCING PROCESS

SEQUENCING

Library preparation

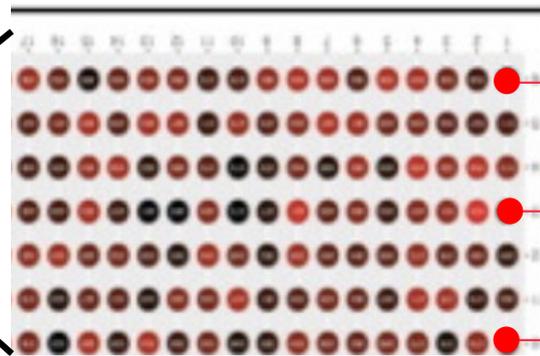


SEQUENCING PROCESS : MinION FLOW CELL

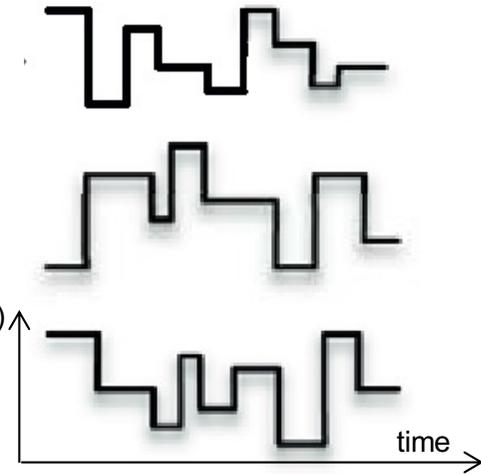


5-6 bases
dominate the
current signal

MinION : 512 pores

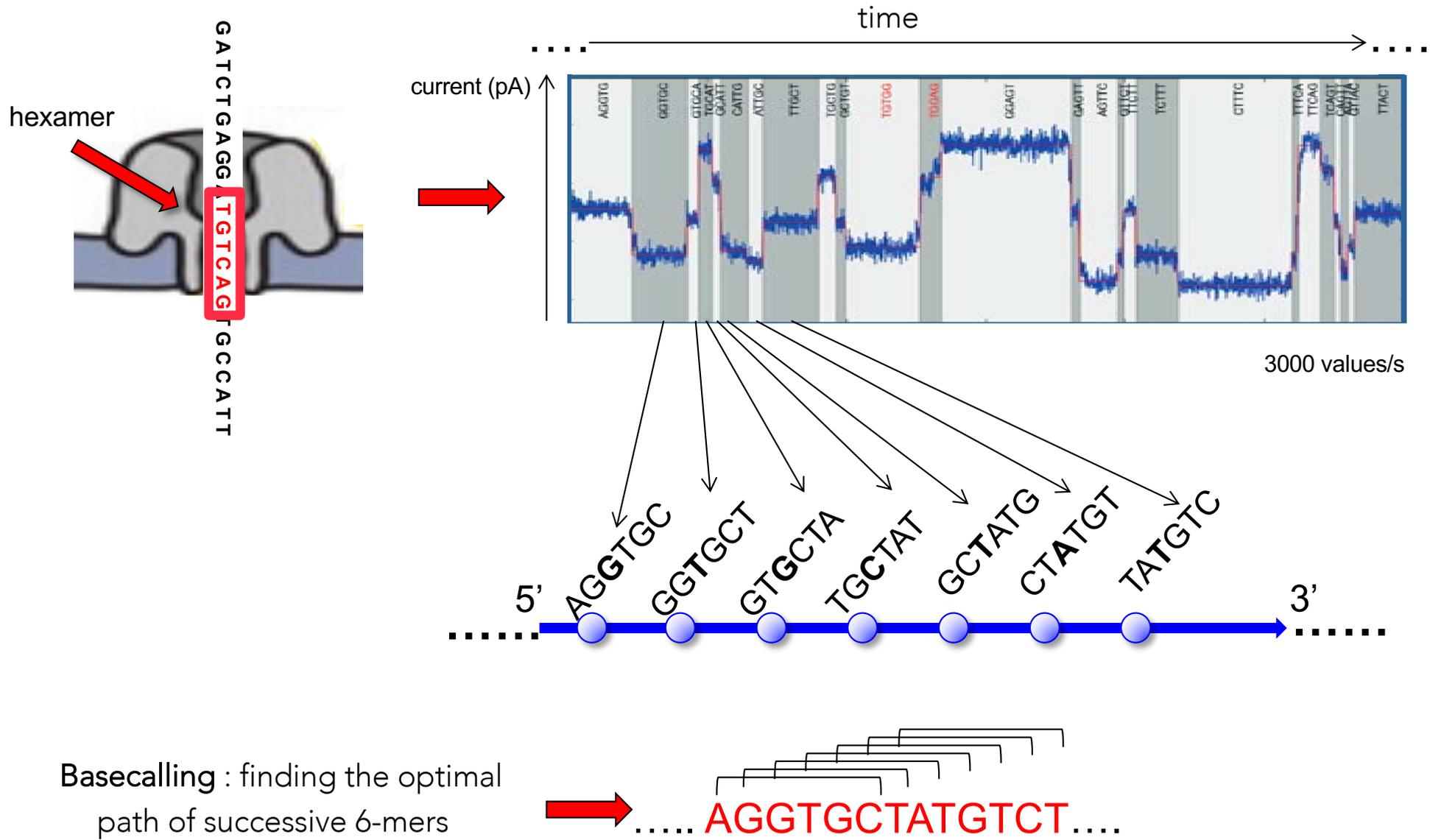


current (pA)



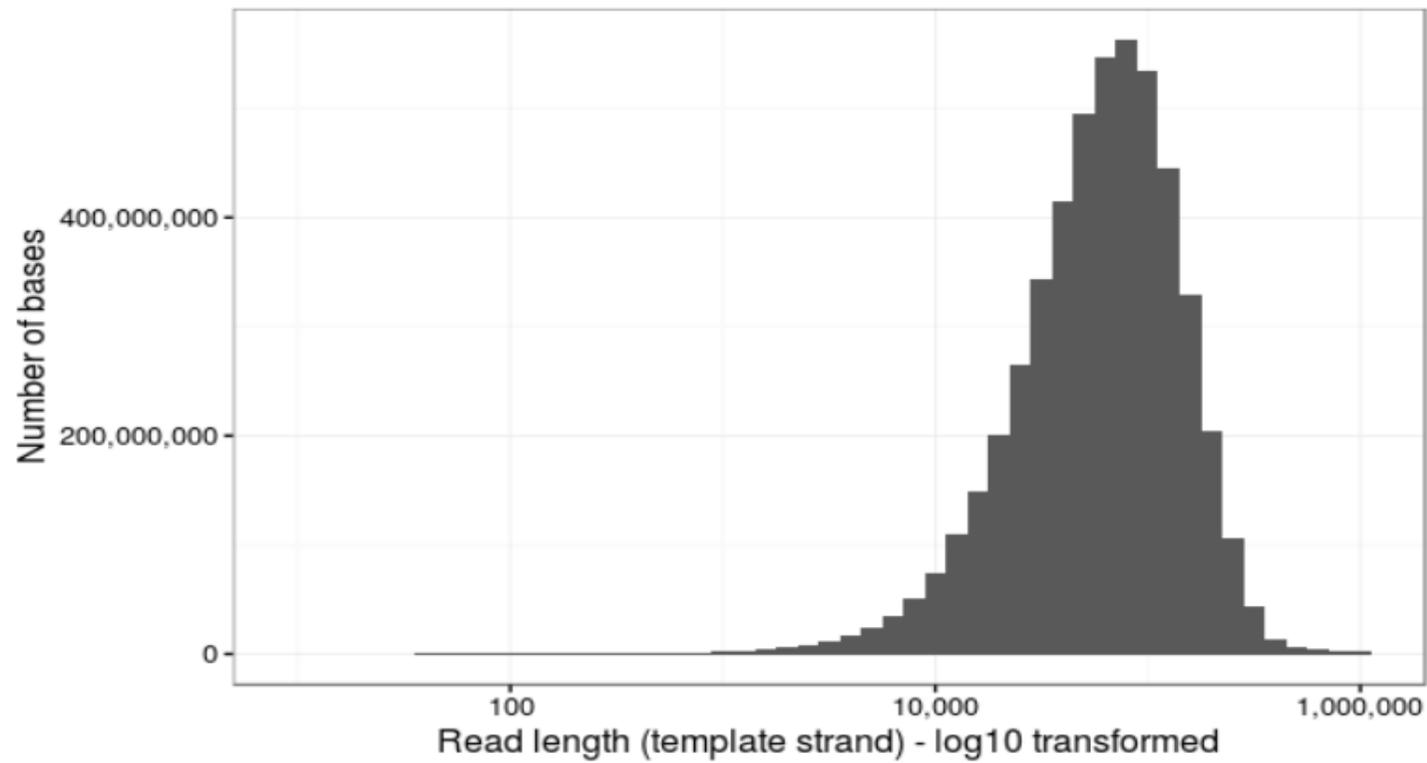
PromethION : 144000 pores (48 x 3000)

BASE CALLING



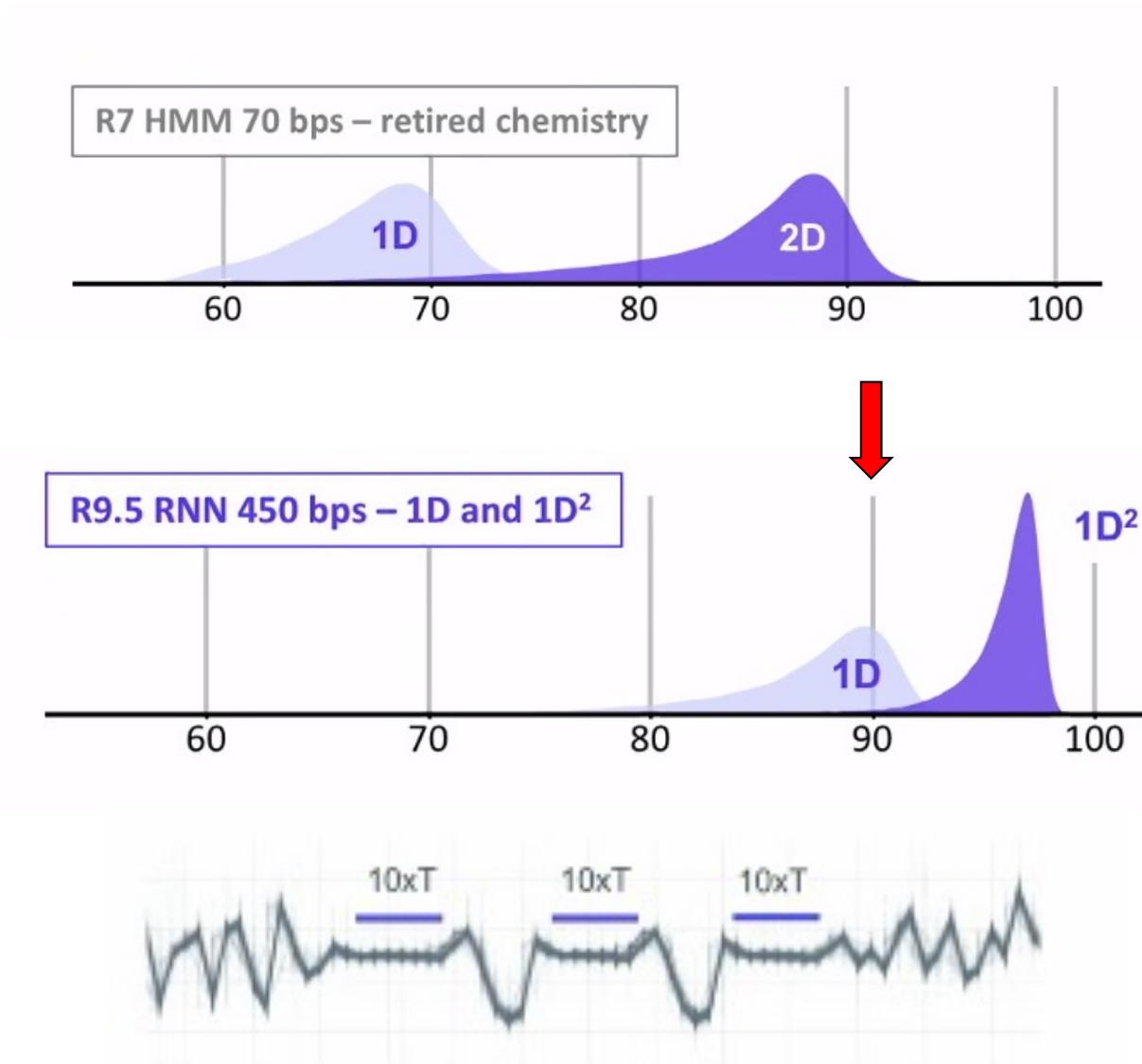
— SIZE OF SEQUENCED DNA FRAGMENTS —

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



Size of the longest read > 1 Mb

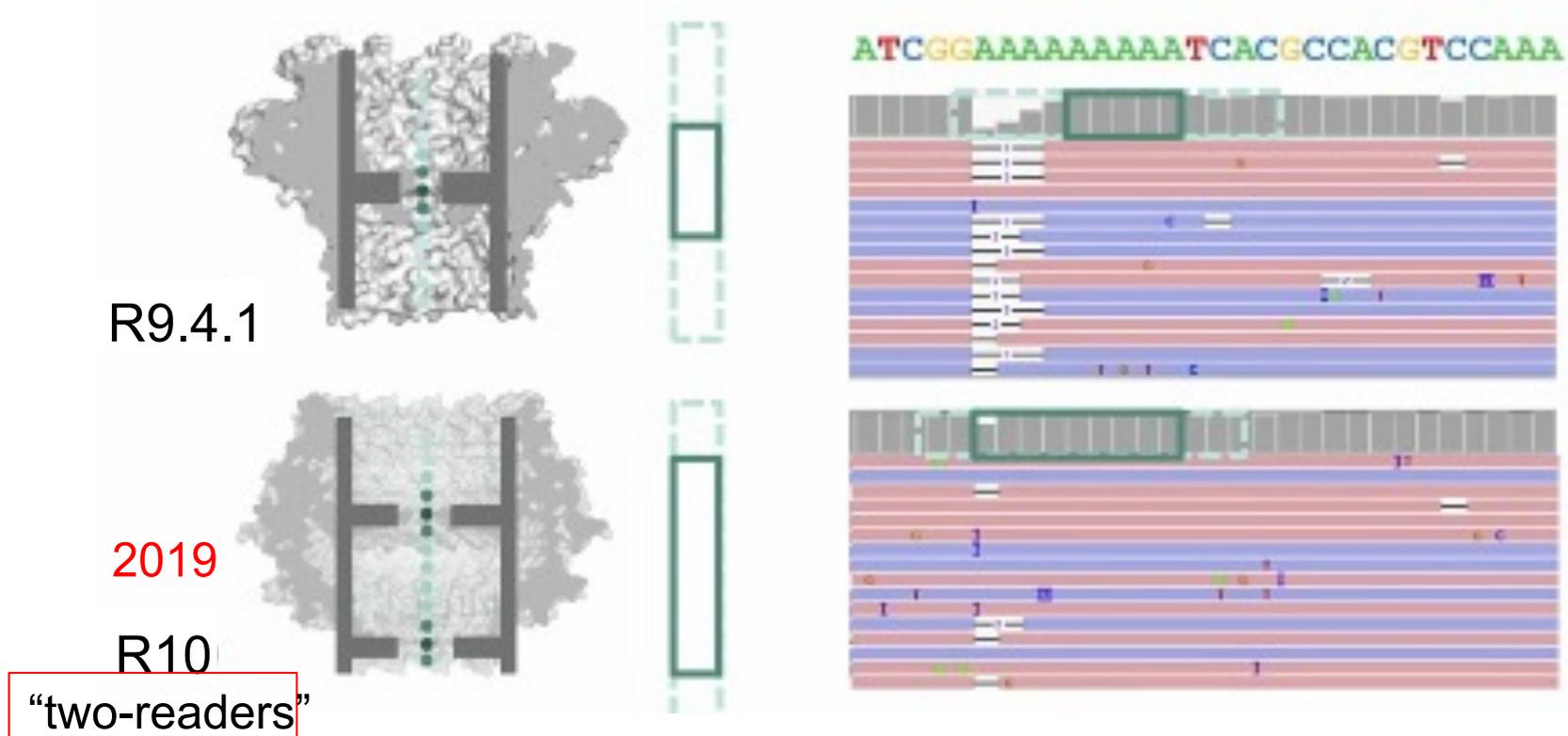
READ QUALITY



Homopolymers difficult to sequence

Recent improvements: "Two readers" nanopore

"One-reader" pore has difficulty to read homopolymers

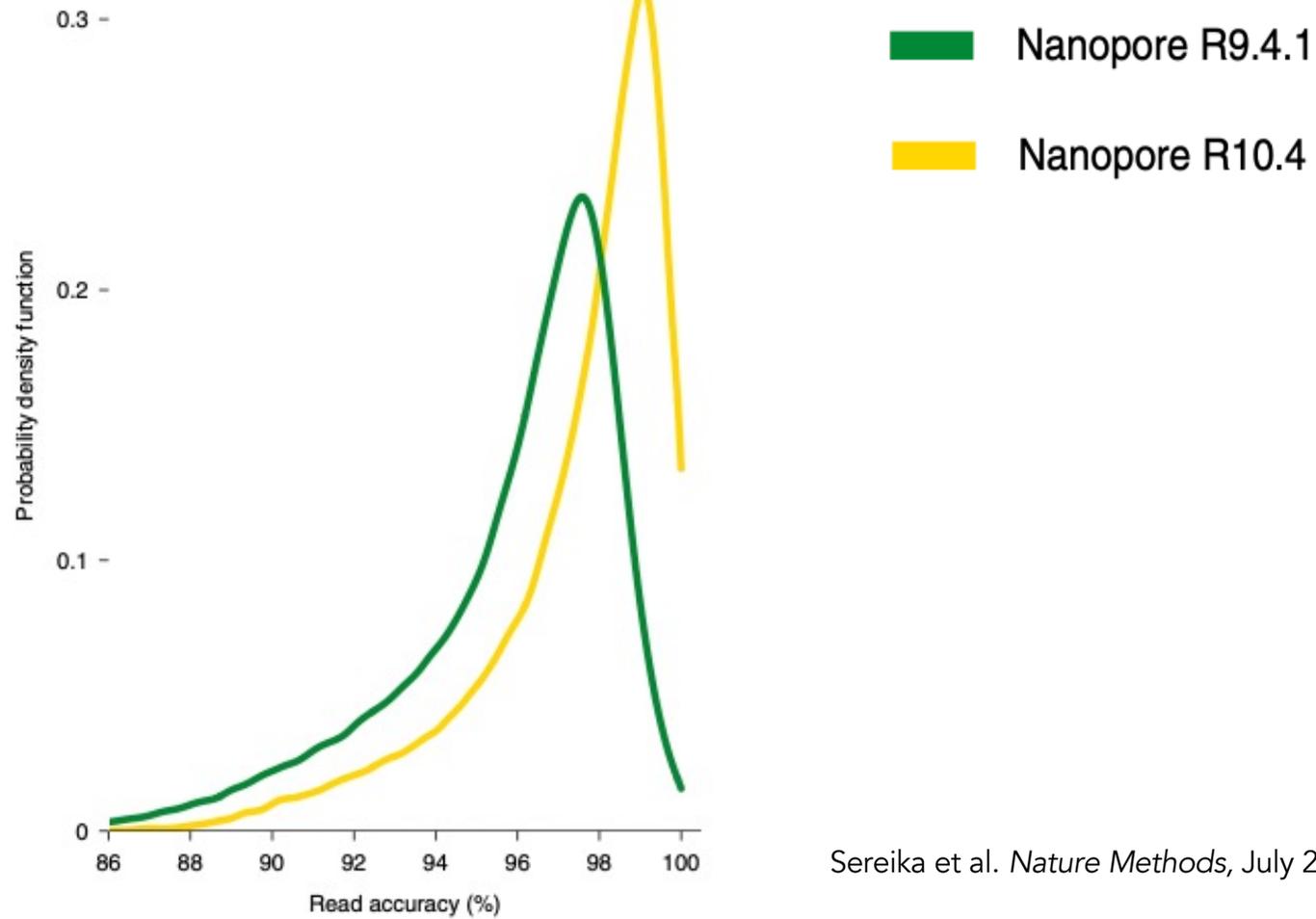


New pore accurately calls homopolymers

- A pore with a longer or multiple "readers" has more bases dominating the signal
- Longer homopolymers are "seen" by the pore and can be decoded with high accuracy

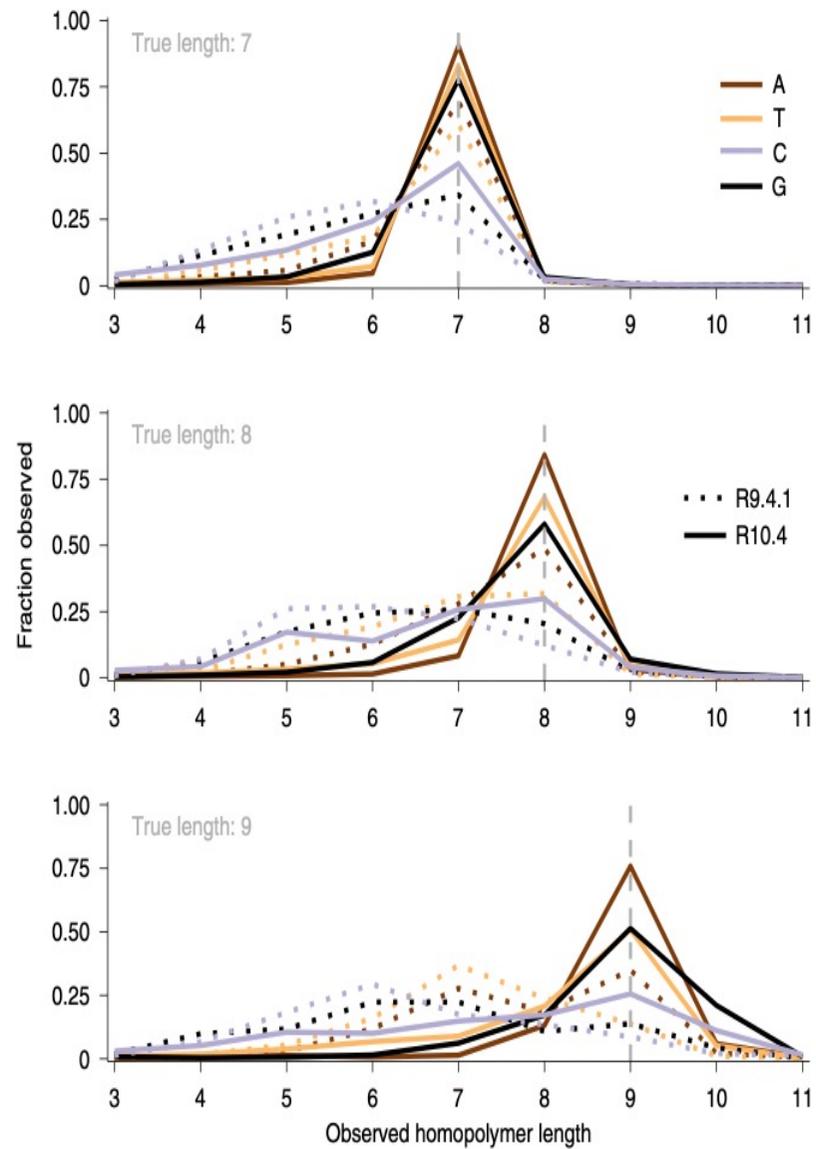
Recent improvements: "Two readers" nanopore

Read accuracies measured through read-mapping



Sereika et al. *Nature Methods*, July 2022

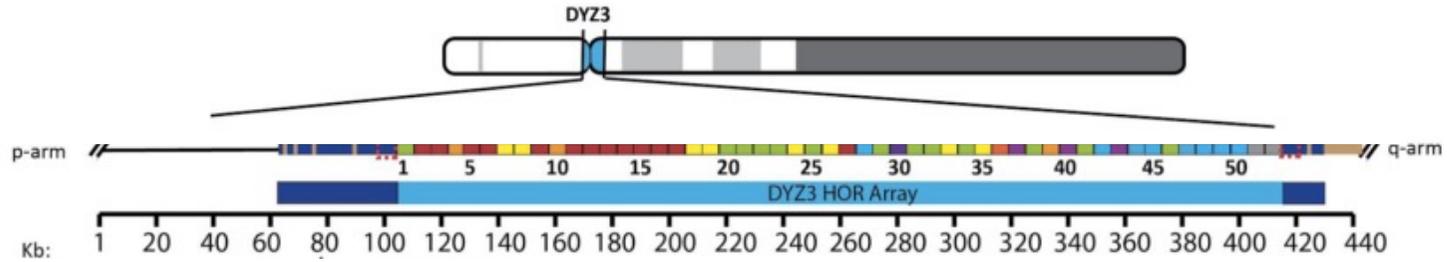
Recent improvements: "Two readers" nanopore



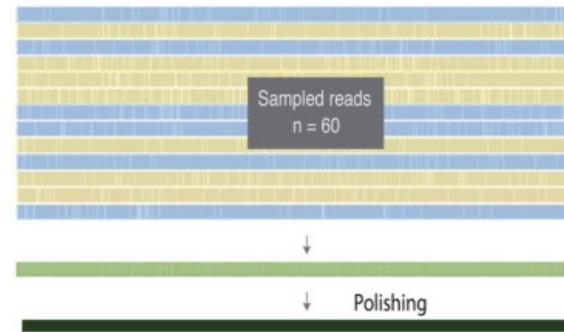
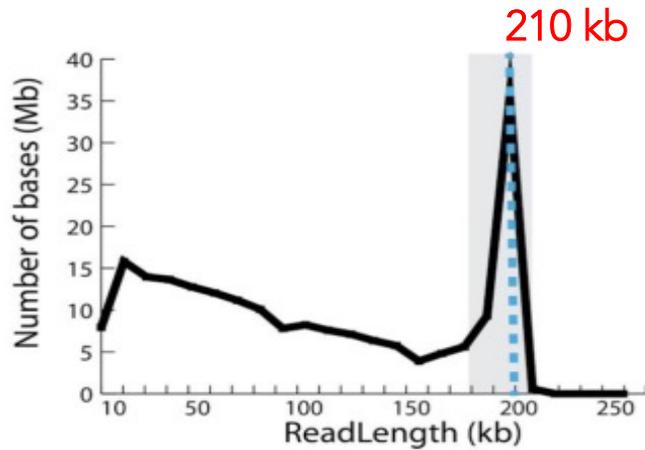
➤ R10.4 : improved ability to call homopolymers

GENOME ASSEMBLY WITH NANOPORE

1 - Linear Assembly of a Human Y Centromere using Nanopore Long Reads
Jain et al., *bioRxiv*, 2017



9 BACs
100 kb to 210 kb



Final high quality consensus BAC sequence

FIRST COMPLETE SEQUENCE OF A HUMAN CENTROMERE

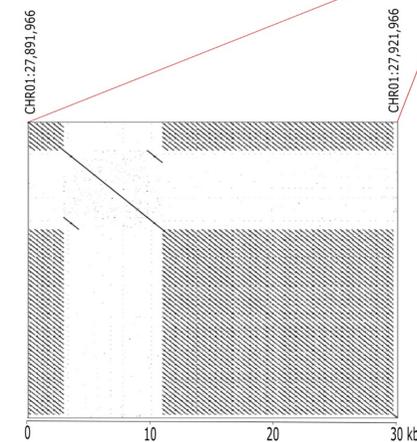
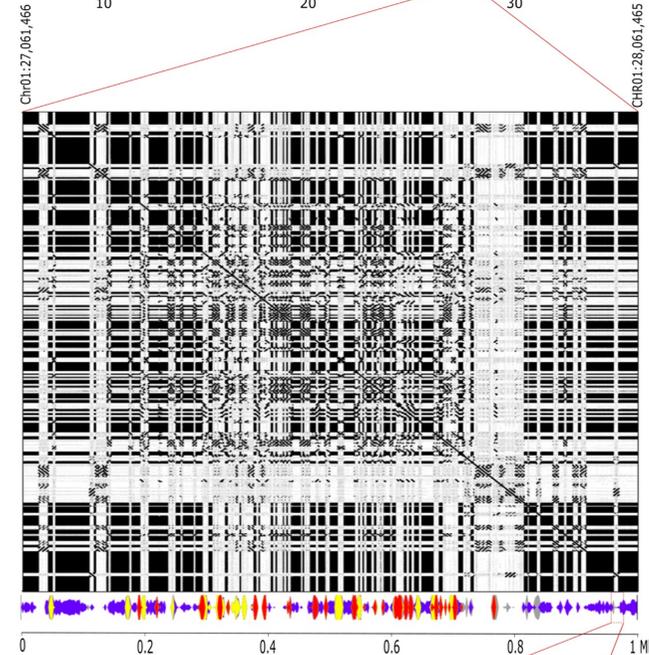
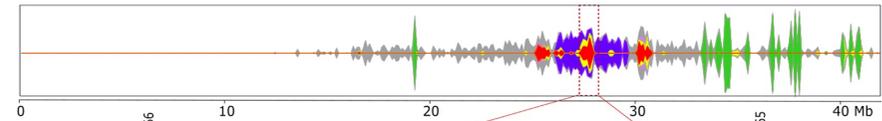
GENOME ASSEMBLY WITH NANOPORE

2 - Telomere-to-telomere gapless chromosomes of banana using nanopore sequencing
Belser et al. *Communications Biology* Sept 2021

- haploid genome :
 - ~500 Mbp,
 - 11 chromosomes:
- 3 samples of reads:
 - 177X of all reads
 - 30X of the longest reads
 - 30X of the **Filtlong** highest-score reads
- assembler: NECAT11,
- 124 contigs polished with:
 - Racon (nanopore reads)
 - Medaka (nanopore reads)
 - Hapo-G (Illumina reads) : incorporates phasing information (Aury & Istace, NAR Apr. 2021)
- Bionano:
 - validate order and orient the contigs:
 - all contigs but 1 in accordance with optical maps
- **➡ 5 chromosomes reconstructed telomere to telomere**
- reveal centromeres, clusters of paralogous genes
- Ex. : in previous versions : 130 5S rDNA genes
- New version : 7696 rDNA genes

Fine structure of repeated elements

Chromosome 01



■ Nanica ■ 45S ■ 5S ■ CRM ■ CL33 ■ CL18 ■ Maximus

GENOME ASSEMBLY WITH NANOPORE

3 - Long-read and chromosome-scale assembly of the hexaploid wheat genome
Aury et al., *bioRxiv*, Aug 2021

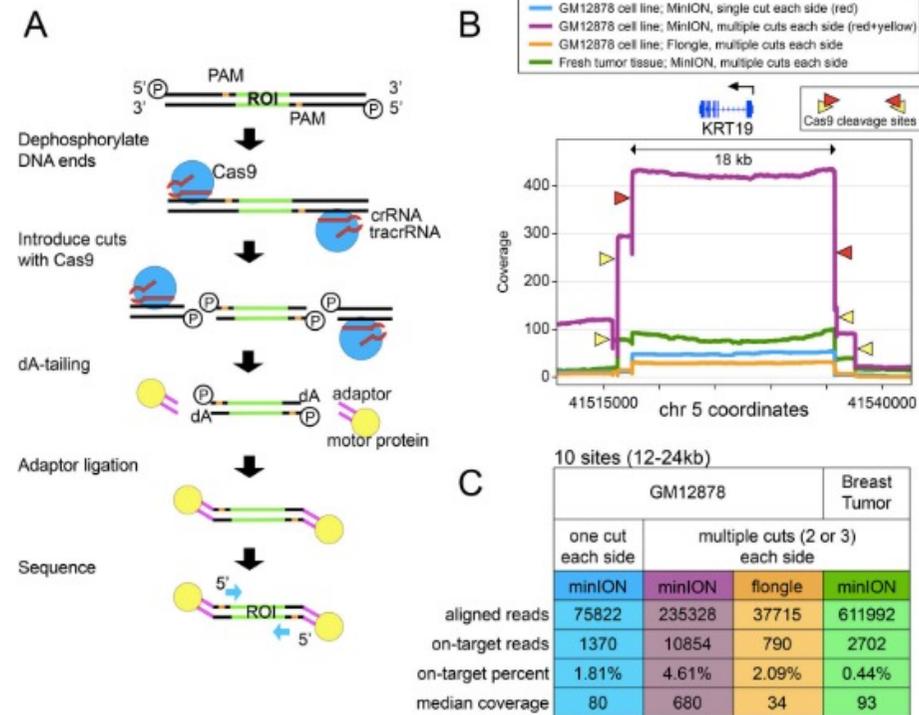
- **First hexaploid wheat genome based on ONT long-reads**
- hexaploid genome (15.5 Gb)
- sequencing began in 2005 : International Wheat Genome Sequencing Consortium (IWGSC)
- first sequence in 2018
- This work:
 - ✓ organize contigs in chromosomes using:
 - ONT
 - 20 ONT flow cells (2 MinION and 18 PromethION)
 - produced 12M reads representing 1.1 Tb
 - base calling: (i) guppy 2.0 and then guppy 3.6 (High Accuracy)
 - coverage: 63x, N50: 24.6 kb
 - 3.1M reads > 50 kb, coverage: 14x
 - Bionano Genomics (BNG) Saphyr
 - direct Label and Stain Chemistry (DLS) with the DLE-1 enzyme
 - total size: 14.9 Gb, N50: 37.5 Mb
 - Hi-C
 - 4 Hi-C libraries, Arima Genomics protocol
 - Illumina sequencing -> 537 Gb, 35x
 - We used a sample of 240 million read pairs (72 Gb, 5x) to build a Hi-C map



Most contiguous and complete chromosome-scale assembly of a bread wheat genome

— GENOME SEQUENCING : TARGETED NANOPORE SEQUENCING —

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



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

But it removes critical information such as methylation status, takes time to design and optimize

— GENOME SEQUENCING : TARGETED NANOPORE SEQUENCING —

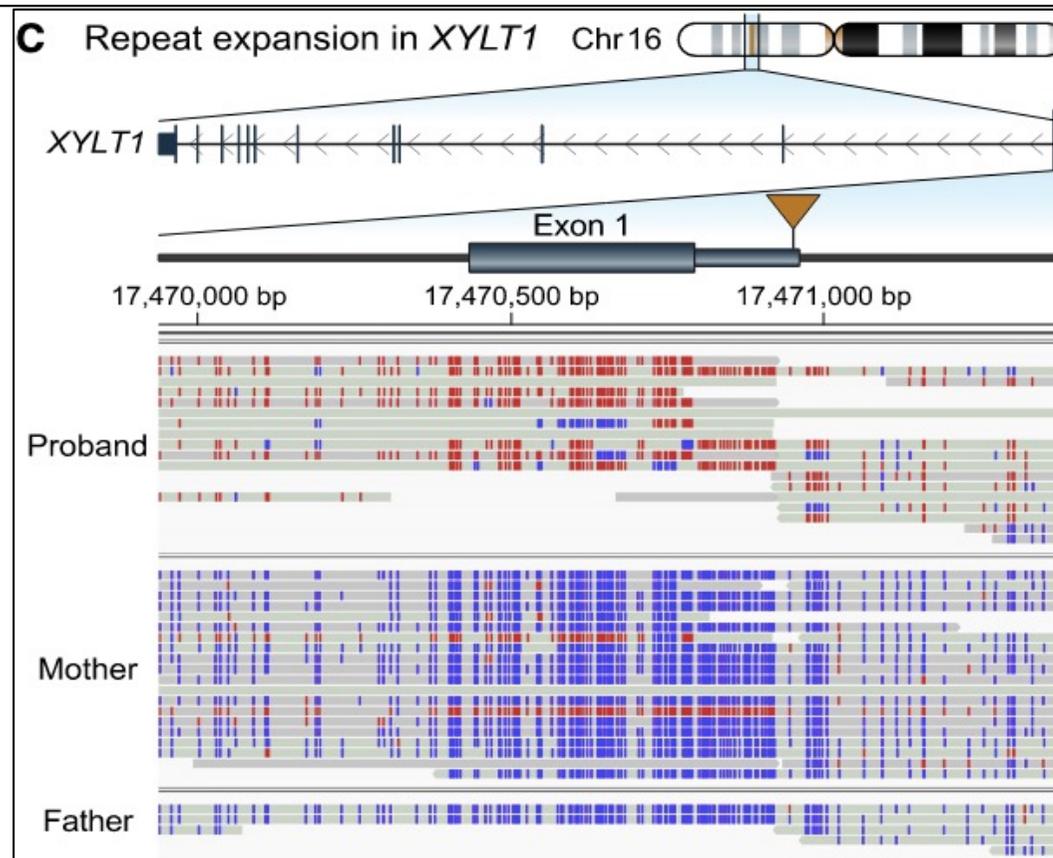
2 - Targeted long-read sequencing identifies missing disease-causing variation

Miller et al. *American Journal of Human Genetics*, 2021

- The software analyzes the signal after a DNA molecule enters a pore to determine if it lies in the region of interest
- If it does, the pore continues to sequence the molecule
- If not, the DNA molecule is ejected from the pore
- In cases with complex CNVs, large genomic regions on either side of the known aberration are targeted

Baratela-Scott syndrome mediated by methylation :

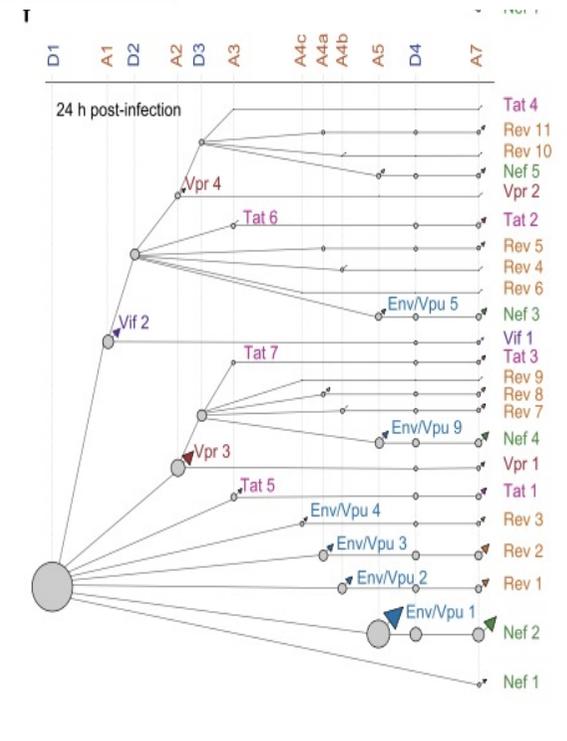
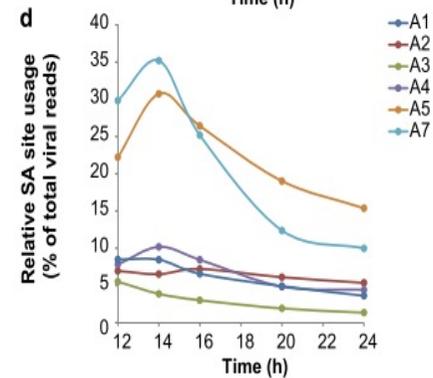
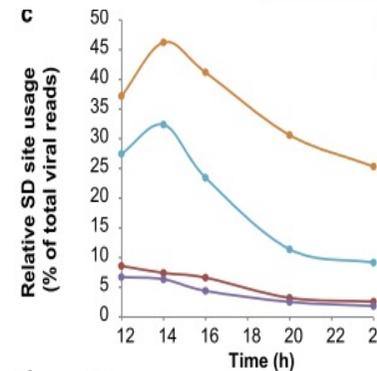
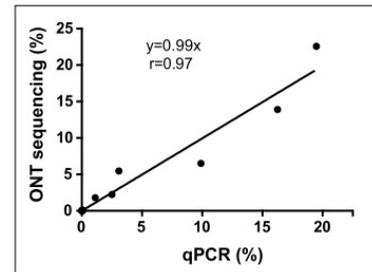
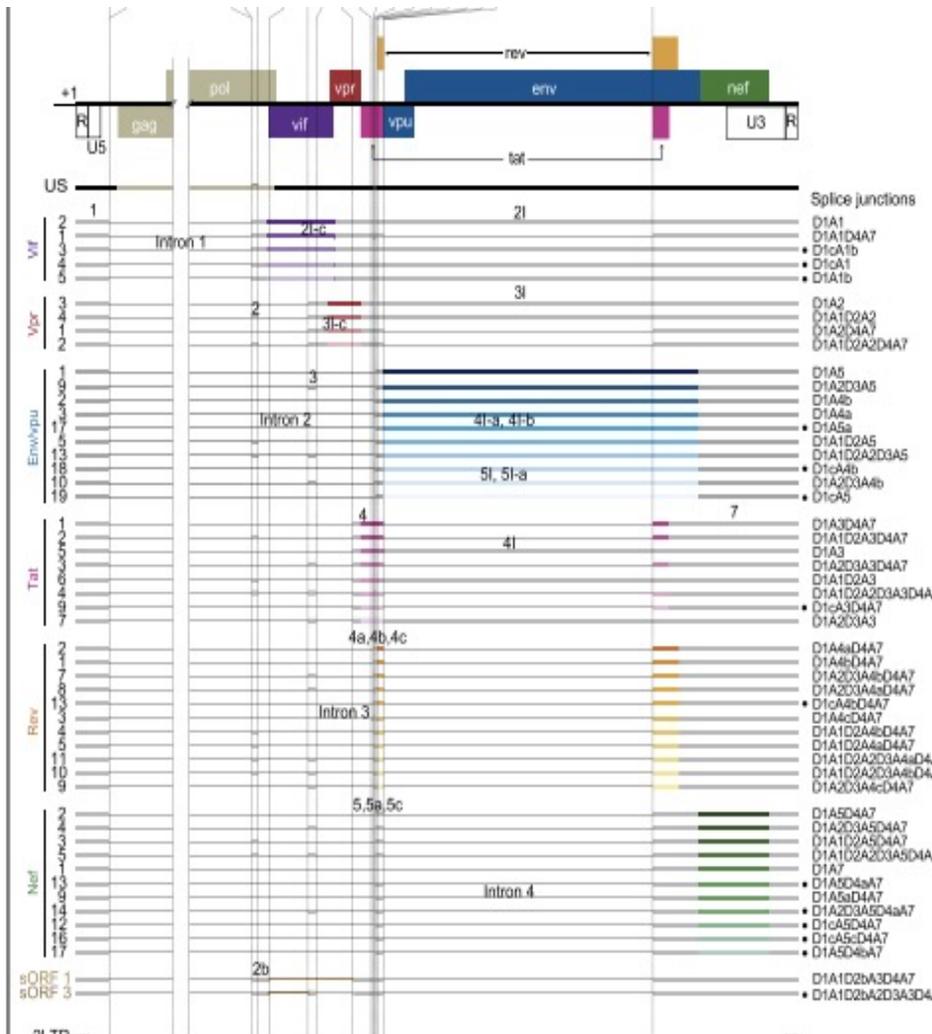
- T-LRS of native DNA molecules provides additional information not available when repeat length and methylation are assayed separately.



cDNA NANOPORE SEQUENCING

Dynamic nanopore long-read sequencing analysis of HIV-1 splicing events during the early steps of infection
 Quang et al. *Retrovirology* 2020

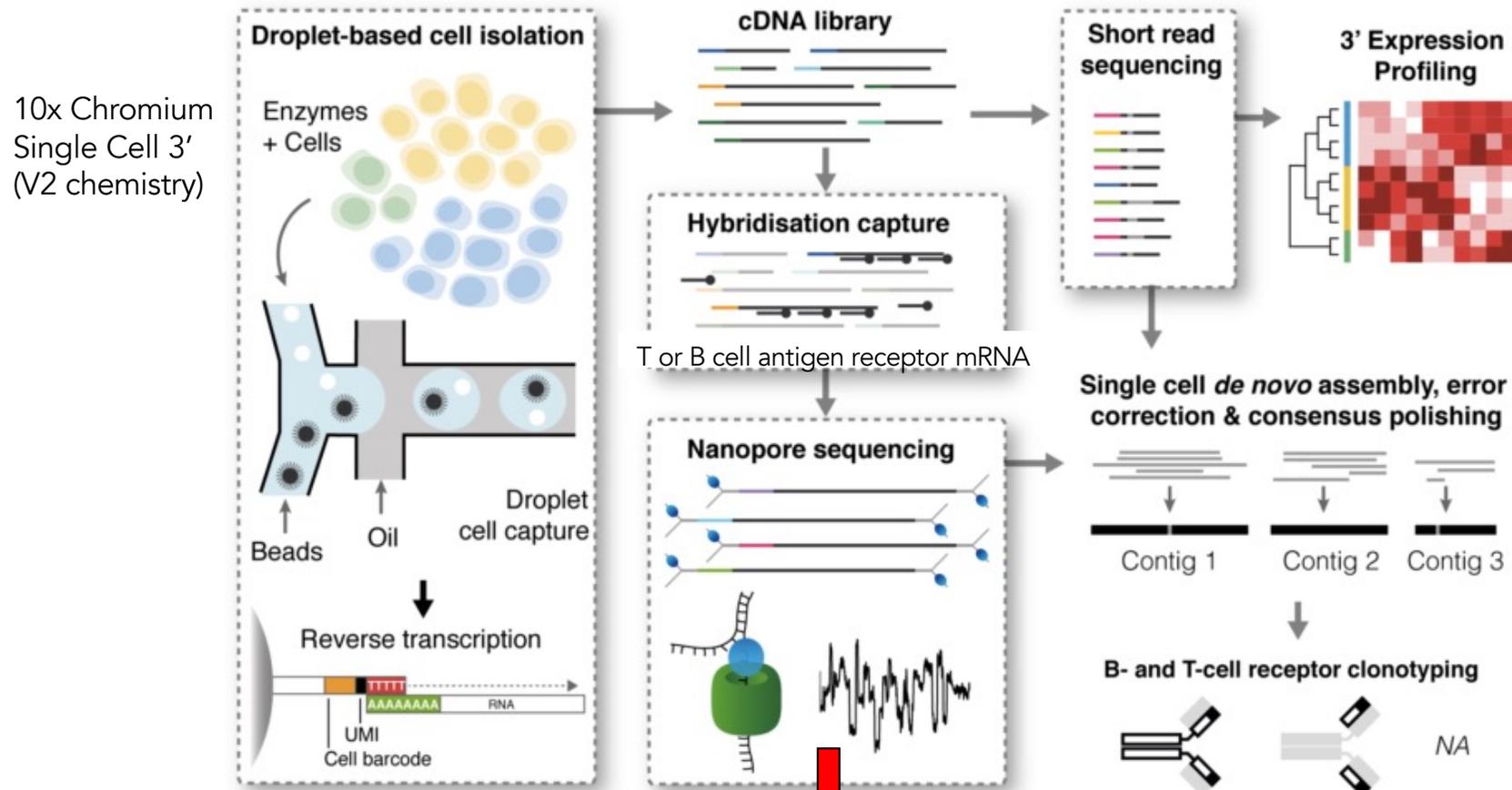
- 53 viral RNA isoforms, including 14 new ones
- Relative levels highly correlated with qPCR
- First dynamic picture of the cascade of events occurring between 12 and 24 h of viral infection
- -> importance of non-coding exons in viral RNA transcriptome regulation



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



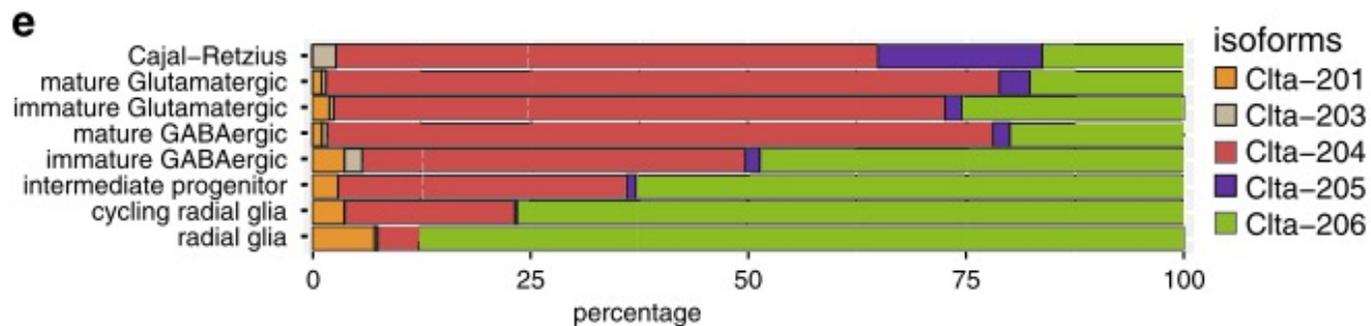
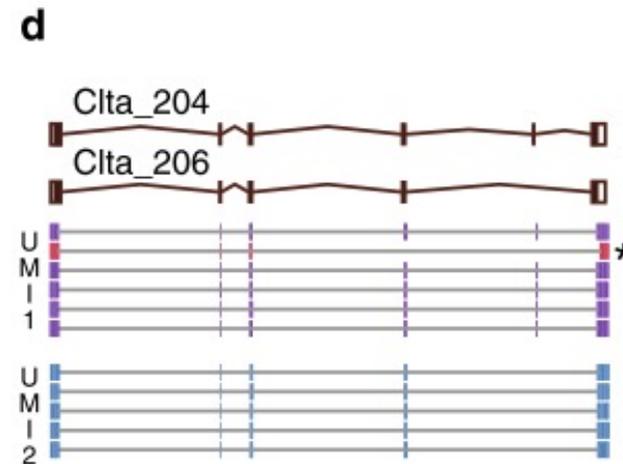
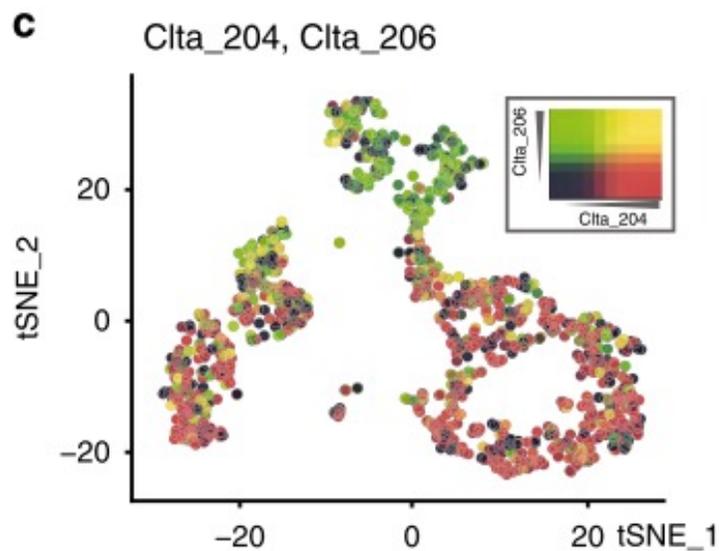
Tracking of somatic mutation, alternate splicing and clonal evolution of T and B lymphocytes
BUT
Does not correct for PCR biases

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)

- High accuracy cellBC and UMI assignment
- Analysis of splicing and sequence variation at the single-cell level

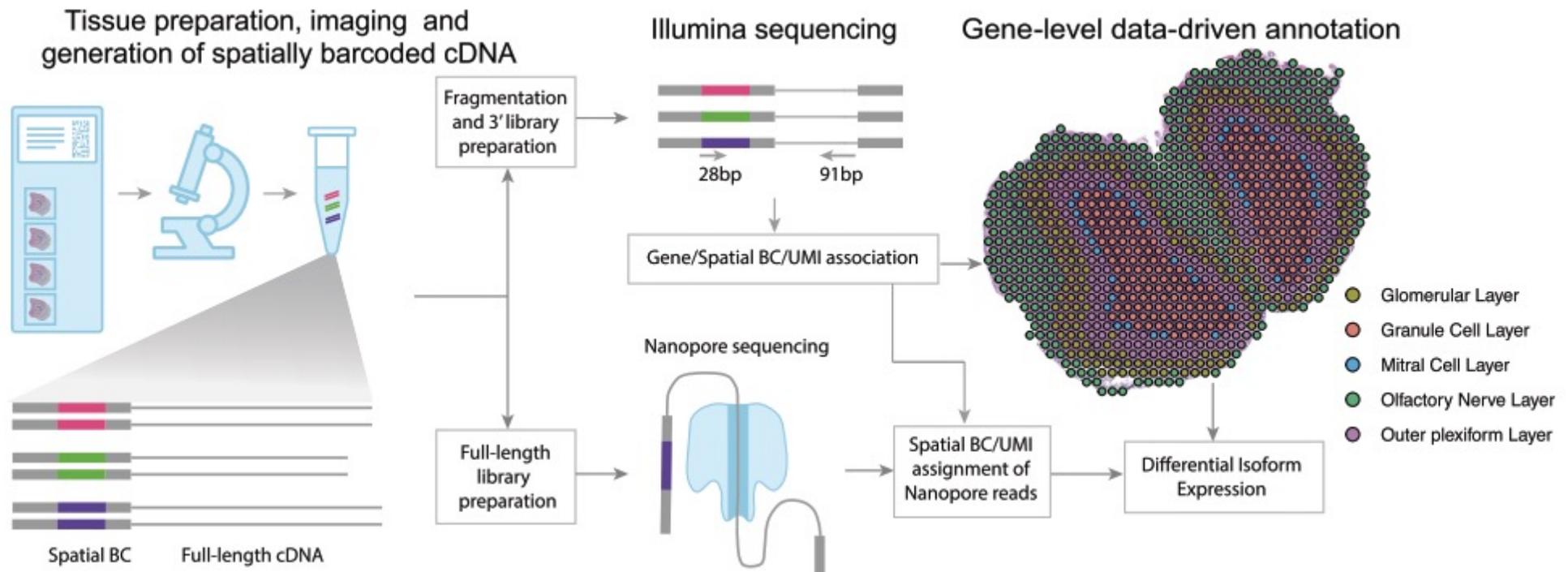


NANOPORE and 10x Genomics Visium

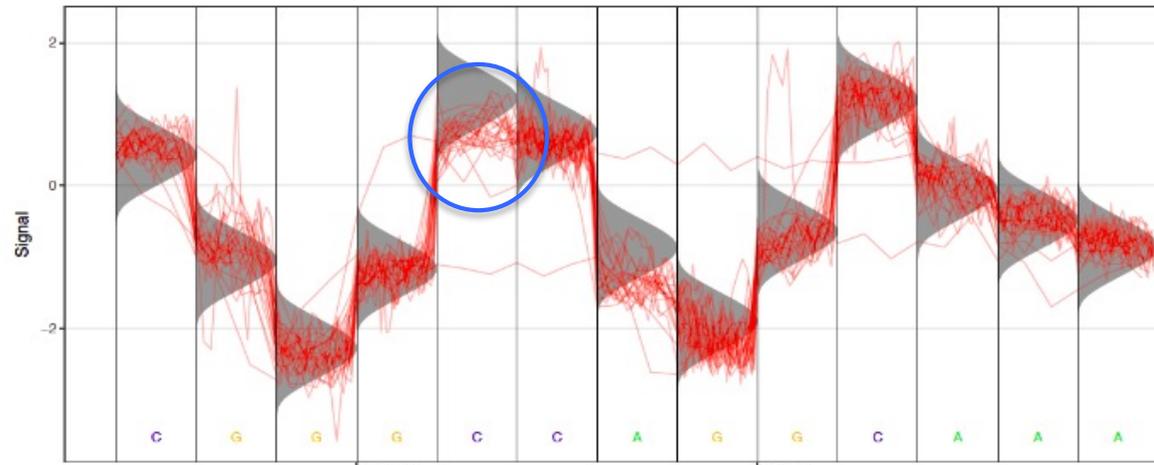
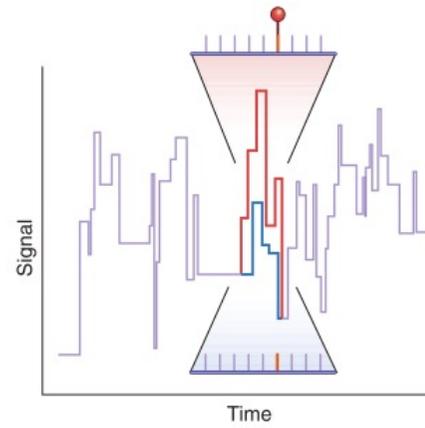
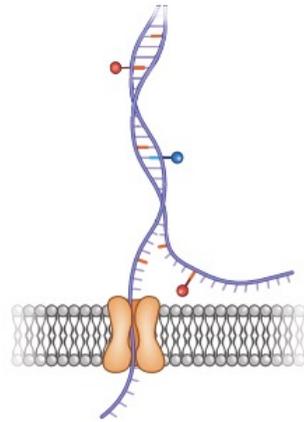
The spatial landscape of gene expression isoforms in tissue sections
Lebrigand et al., *bioRxiv*, 2020

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

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



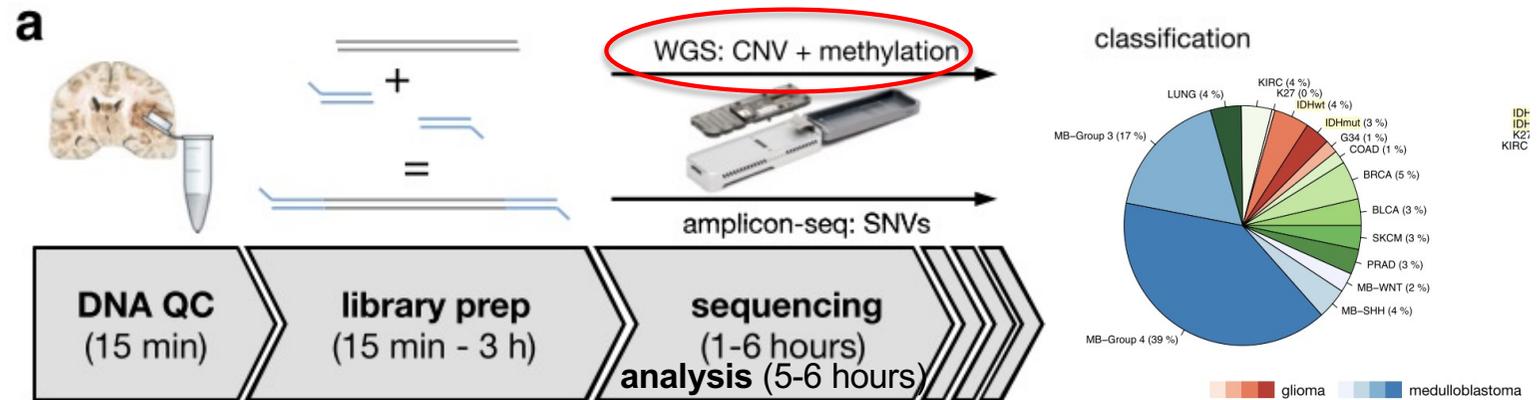
DETECTION OF MODIFIED DNA BASES



— Electric signal
▶ Canonical base distribution

— DETECTION OF MODIFIED DNA BASES : 5mCpG 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 :

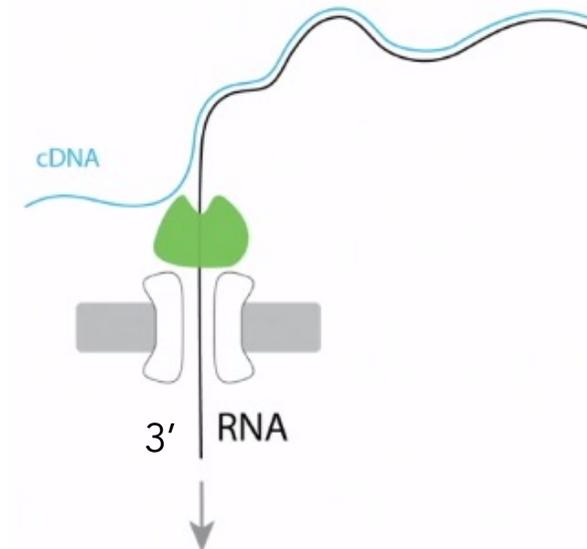
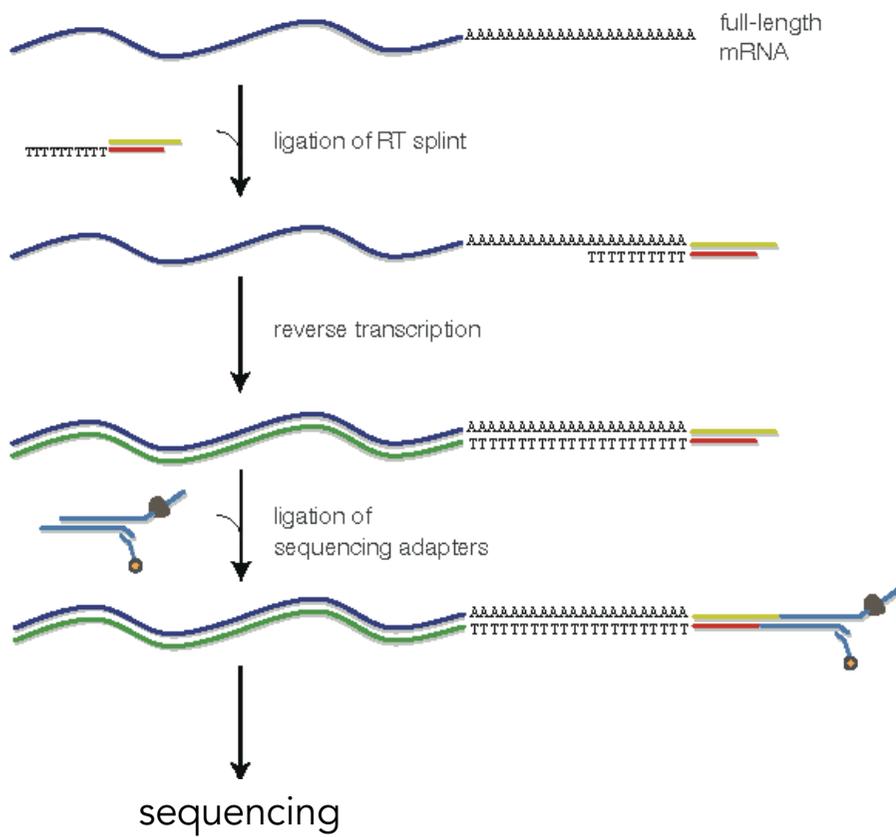
- structural variants
- point mutations
- CpG methylation profiling

Single device with negligible capital cost :

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

DIRECT RNA SEQUENCING

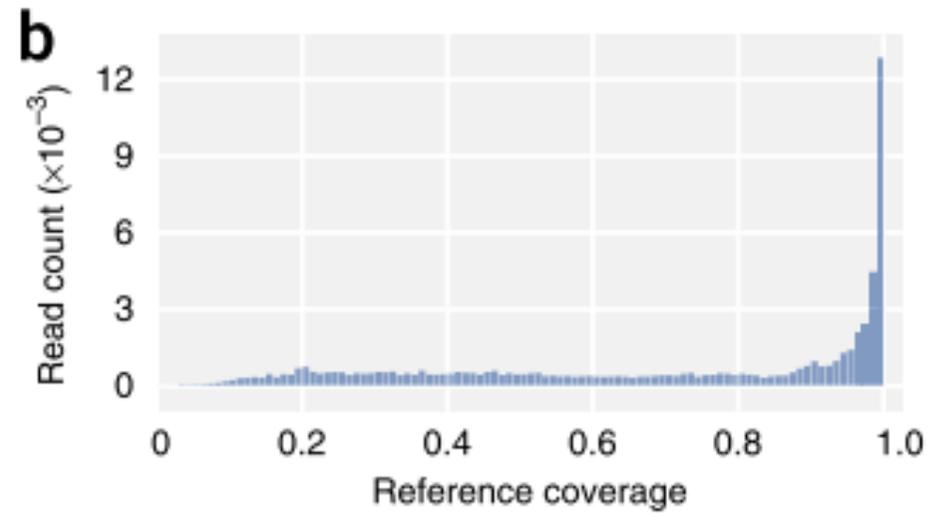
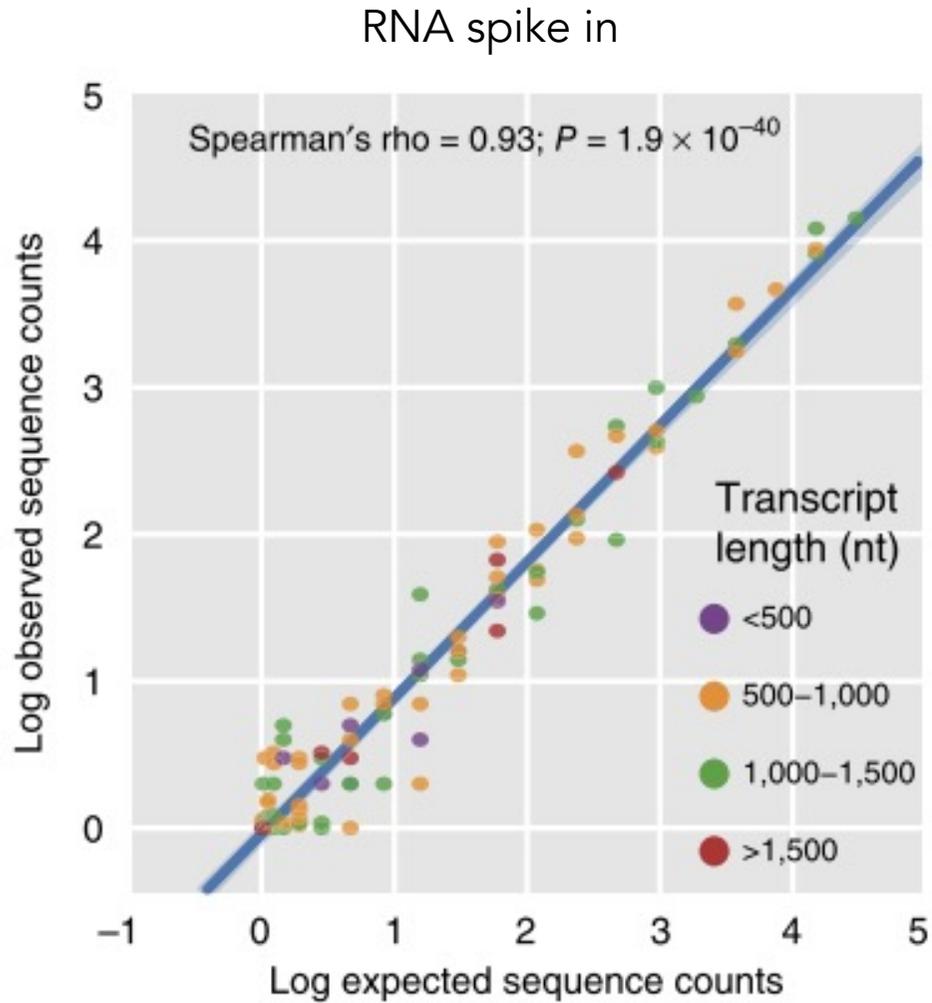
Library preparation



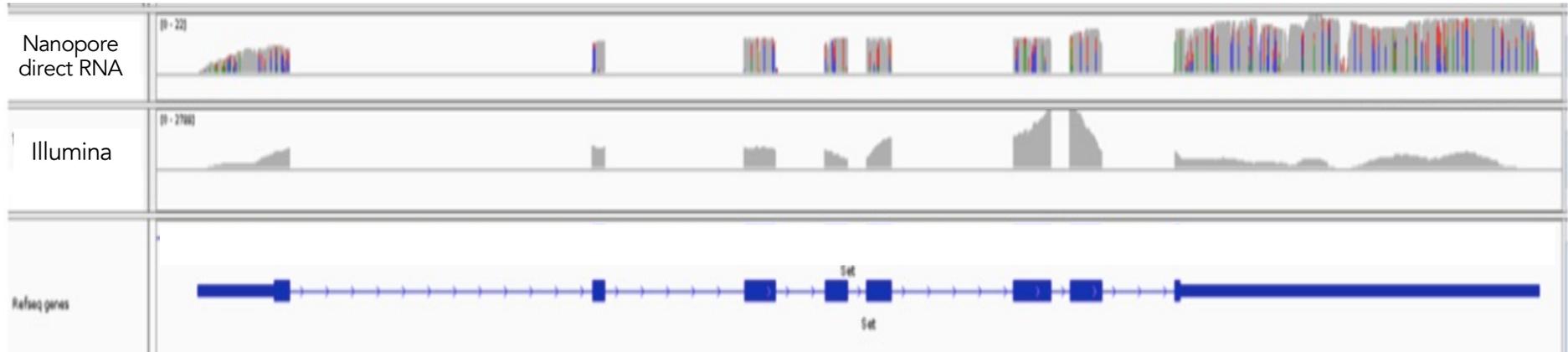
RNA directly sequenced in nanopore

- No PCR bias
- Quantitative

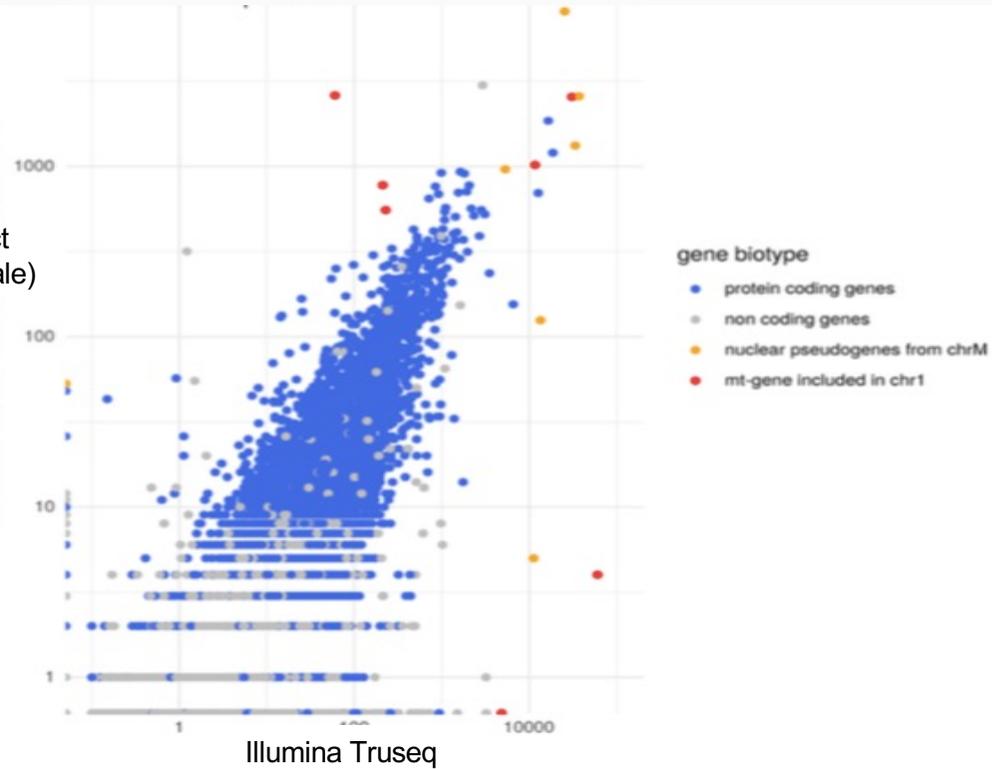
DIRECT RNA SEQUENCING : CONTROLS



DIRECT RNA SEQUENCING vs ILLUMINA

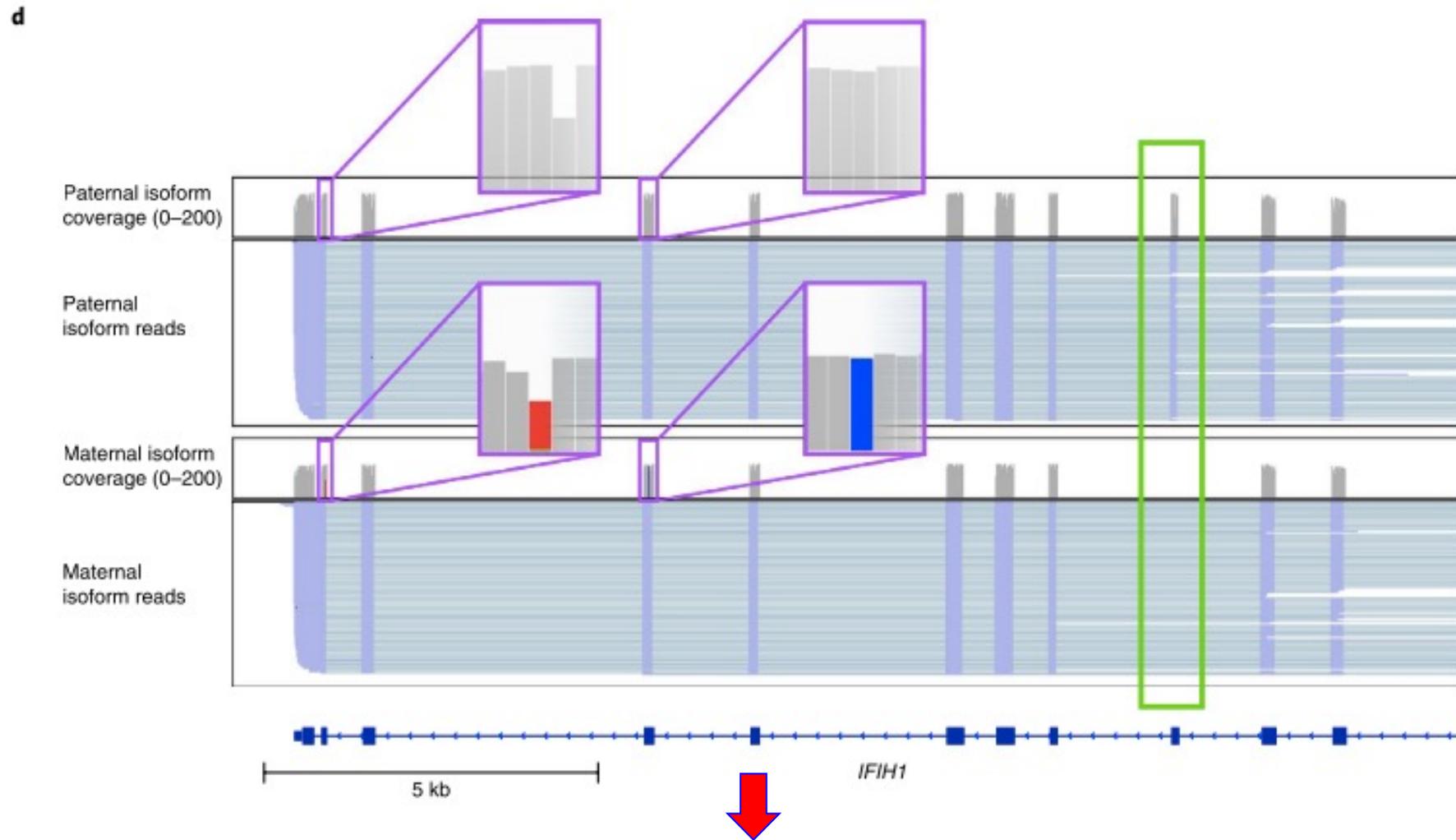


Nanopore RNA direct
(read number ; Log scale)



DIRECT RNA SEQUENCING: TRANSCRIPT HAPLOTYPE

Nanopore native RNA sequencing of a human transcriptome
Workman et al. *Nature Methods* 2019



— DIRECT RNA SEQUENCING: DETECTION OF 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)

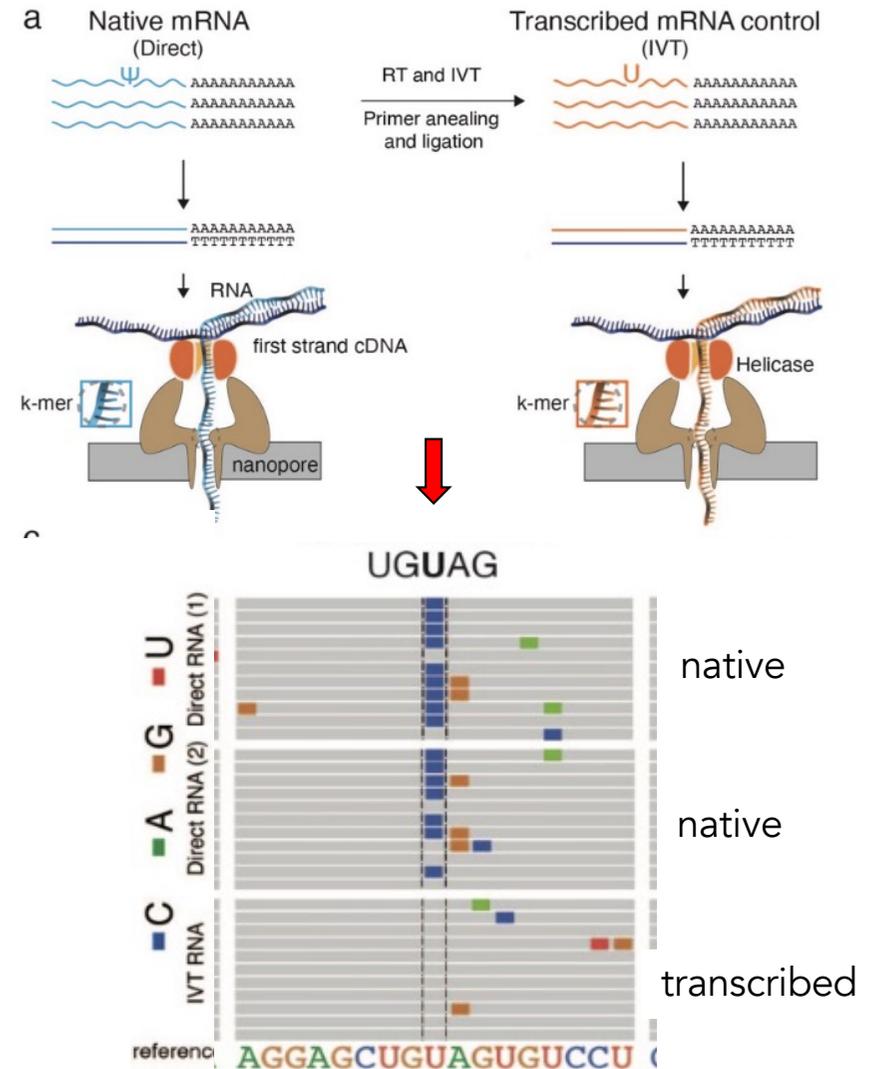
- HIV RNA rich in *6mA*

DIRECT RNA SEQUENCING : DETECTION OF pseudo-U

Detection of pseudouridine modifications and type I/II hypermodifications in human mRNAs using direct long-read sequencing.
Tavakoli et al. *bioRxiv* Nov. 2021

Detection of pseudo-U sites

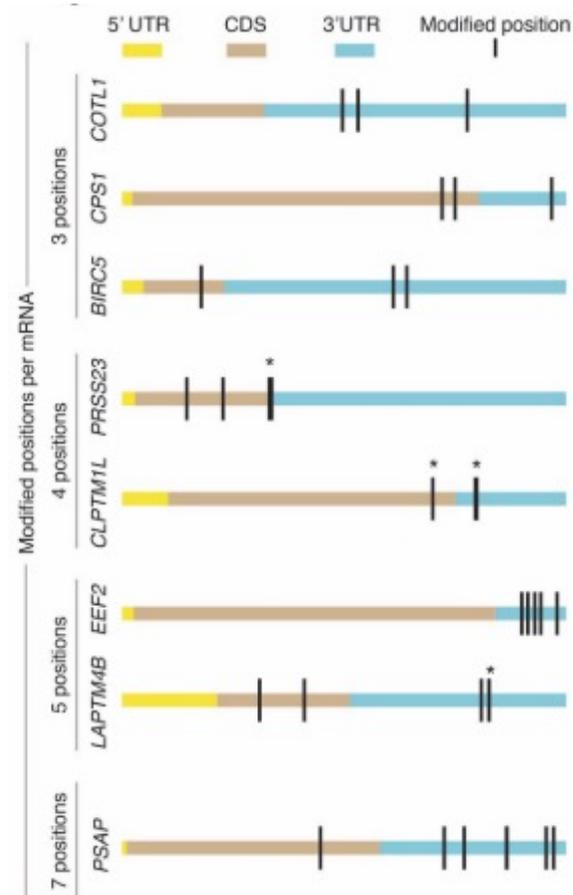
- U-to-C base-calling errors occur at pseudouridines
 - benchmarked against sites previously identified
- ↓
- Pipeline for direct identification, quantification, and detection of pseudouridine modifications and
 - Controls :
 - 1000mer synthetic RNA with single pseudouridine in center position
 - U-to-C occurs at the site of pseudouridylation
 - Discovery of human mRNAs with up to 7 unique sites of pseudouridine modification



— DIRECT RNA SEQUENCING : DETECTION OF pseudo-U

Detection of pseudouridine modifications and type I/II hypermodifications in human mRNAs using direct long-read sequencing.
Tavakoli et al. *bioRxiv* Nov. 2021

Pseudouridinylated human mRNAs :
104 at 2 positions
27 at 3 positions
4 at 4 positions
5 at 5 positions
1 at 6 positions
1 at 7 positions



Small genomes assembly :

Nanopore or PacBio ?

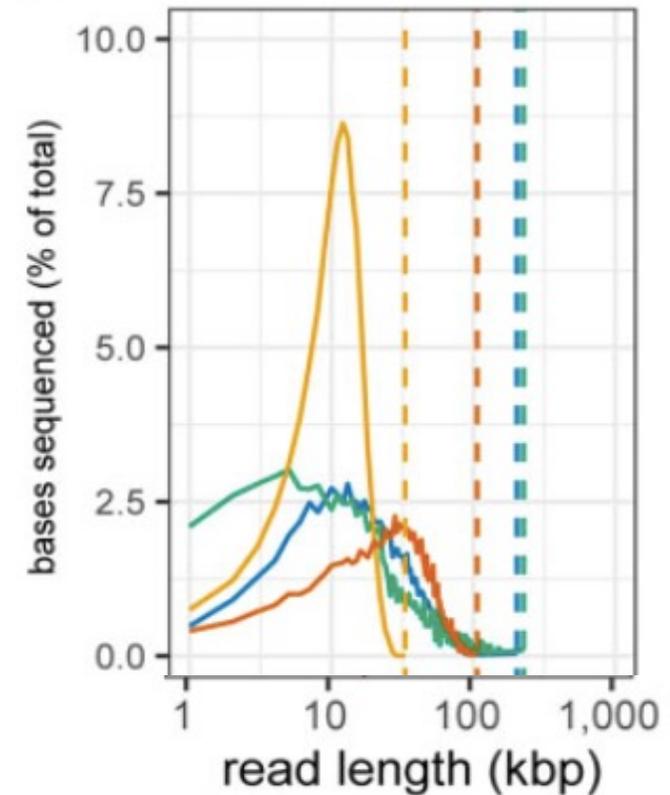
— Assembly of small genomes : Nanopore vs PacBio

1 - Comparison of long-read sequencing technologies in interrogating bacteria and fly genomes
Tvedte et al. G3, 2021

Sequence of *Escherichia coli* genome + 2 small plasmids

Protocols

- PacBio :
 - Sequel II CLR (continuous long-read sequencing)
 - Sequel II HiFi
- ONT :
 - Rapid Sequencing, R9 flow cell
 - Ligation Sequencing, R9 flow cell
 - Ligation Sequencing, R10 flow cell
- ONT + Illumina (hybrid)



— ONT RAPID R9.4.1 — PB RSII
— ONT LIG R9.4.1 — PB SQII CLR
— ONT LIG R10 unsheared — PB SQII HiFi

Assembly of small genomes : Nanopore vs PacBio

1 - Comparison of long-read sequencing technologies in interrogating bacteria and fly genomes
Tvedte et al. G3, 2021

polished with Illumina



Table 2 Summary of *E. coli* E2348/69 assemblies

Library 1	Assembler	Total contigs	Largest genome contig	Largest pMAR2 contig	Largest p5217 contig	BUSCO ^b (%)	Consensus identity ^c (%)
ONT RAPID	Canu	6	4,989,389	189,389	11,738	91.13	99.950
ONT RAPID	Canu ^a	4	4,944,380	96,603	10,423	100.00	99.997
ONT RAPID	Flye	3	4,943,164	96,555	5212	93.55	99.972
ONT RAPID	Unicycler	7	4,944,462	96,603	5218	100.00	NA
ONT LIG	Canu	4	3,093,902	141,938	NA	92.74	99.967
ONT LIG	Canu ^a	4	3,094,900	96,602	NA	100.00	99.996
ONT LIG	Flye	2	3,402,910	NA	NA	93.55	99.974
ONT LIG	Unicycler	7	4,944,462	96,603	5218	100.00	NA
PB RS II	Canu	72	265,067	28,923	NA	45.97	99.747
PB RS II	Canu ^a	67	265,619	29,066	NA	93.55	99.979
PB RS II	Flye	5	4,941,598	96,381	NA	79.84	99.898
PB RS II	Unicycler	13	4,885,846	95,943	5218	100.00	NA
PB SQ II CLR	Canu	4	4,989,961	132,660	NA	99.19	99.998
PB SQ II CLR	Canu ^a	3	5,044,086	96,604	NA	100.00	99.997
PB SQ II CLR	Flye	2	4,944,307	96,604	NA	100.00	99.997
PB SQ II CLR	Unicycler	7	4,944,462	96,603	5218	100.00	NA
PB SQ II HiFi	HiCanu	56	4,930,997	109,122	NA	100.00	99.999
PB SQ II HiFi	HiCanu ^a	10	4,931,051	96,603	NA	100.00	99.998
PB SQ II HiFi	Flye HiFi	2	4,944,462	96,603	NA	100.00	99.999
PB SQ II HiFi	Unicycler	13	4,885,847	96,603	5218	100.00	NA

— Assembly of small genomes : Nanopore vs PacBio —

1 - Comparison of long-read sequencing technologies in interrogating bacteria and fly genomes
Tvedte et al. G3, 2021

NO SINGLE TECHNOLOGY OUTPERFORMED OTHERS IN ALL METRICS :

PacBio

- HiFi
 - Highest consensus accuracy
 - Detection of only 1 plasmid (out of 2)
- CLR
 - High genome contiguity (longer reads)
 - Detection of only 1 plasmid (out of 2)

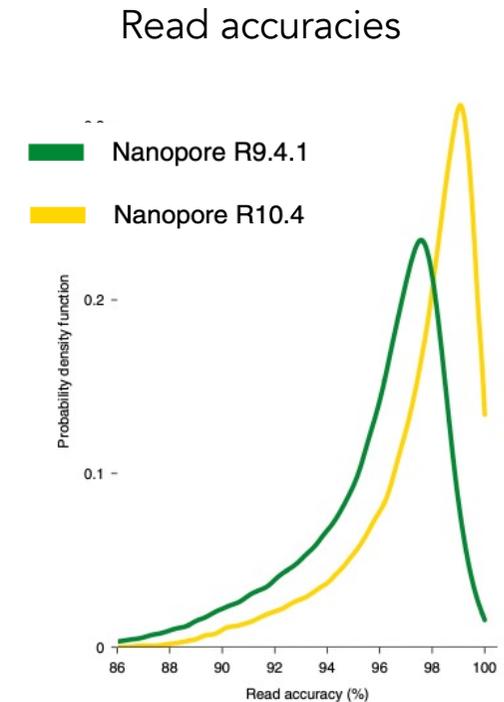
ONT

- All libraries
 - High genome contiguity (longer reads)
 - Better detection of DNA methylation motifs than PacBio (HiFi and CLR)
- ONT Rapid
 - Lowest percentage of chimeric reads (0.02%)
 - Detection of the 2 plasmids
- ONT : more cost-effective

— Assembly of small genomes : Nanopore vs PacBio

2 - 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*, July 2022

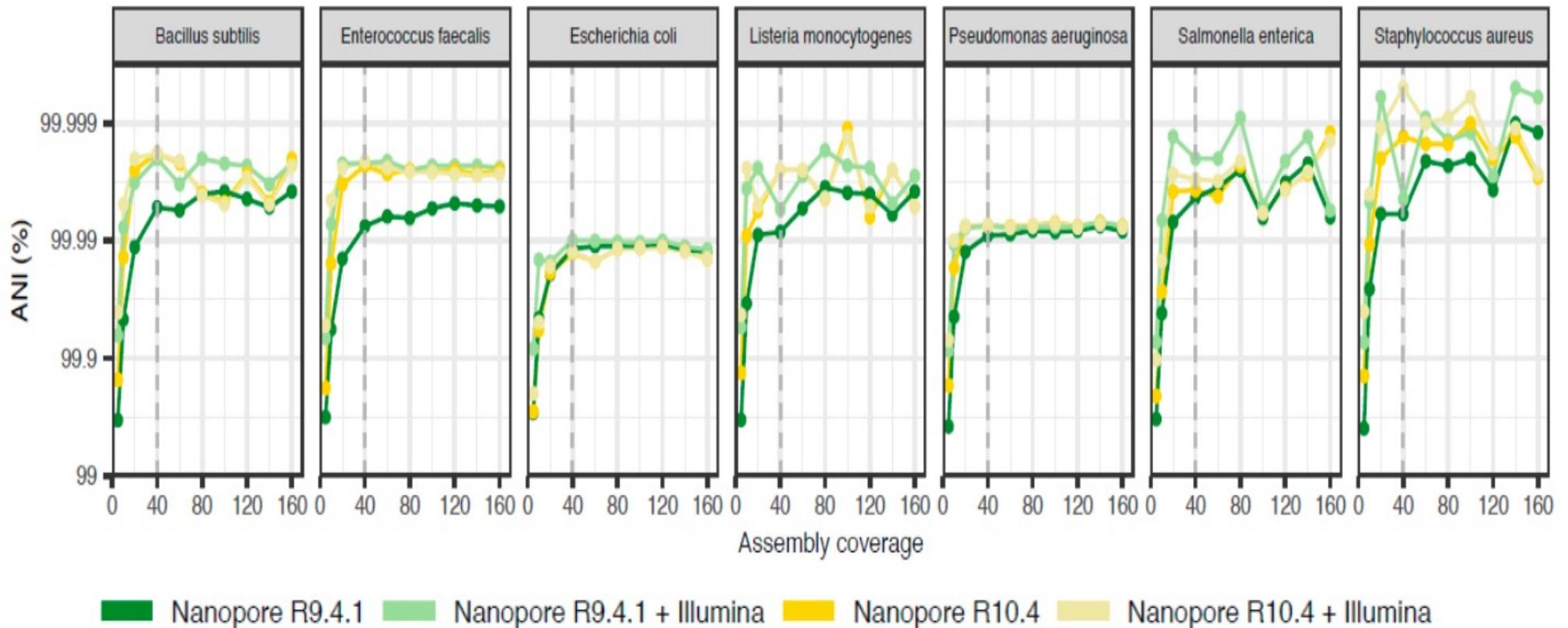
- Samples :
 - Seven bacteria
 - *Saccharomyces cerevisiae*
 - Metagenome : anaerobic digester
- Sequenced with :
 - Illumina MiSeq (2 × 300 bp)
 - PacBio Sequel II HiFi
 - Oxford Nanopore R9.4.1 (MinION) and R10.4 (PromethION)
- Read processing
 - reads assembled with Flye



— Assembly of small genomes : Nanopore vs PacBio —

2 - 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*, July 2022

Sequencing and assembly statistics for the bacterial species ($n = 7$)



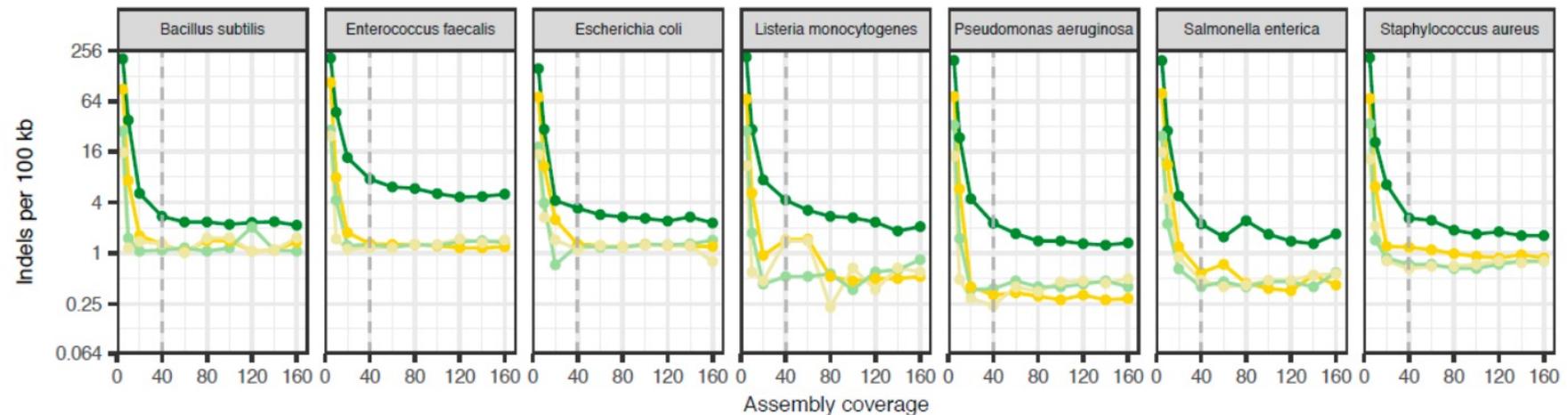
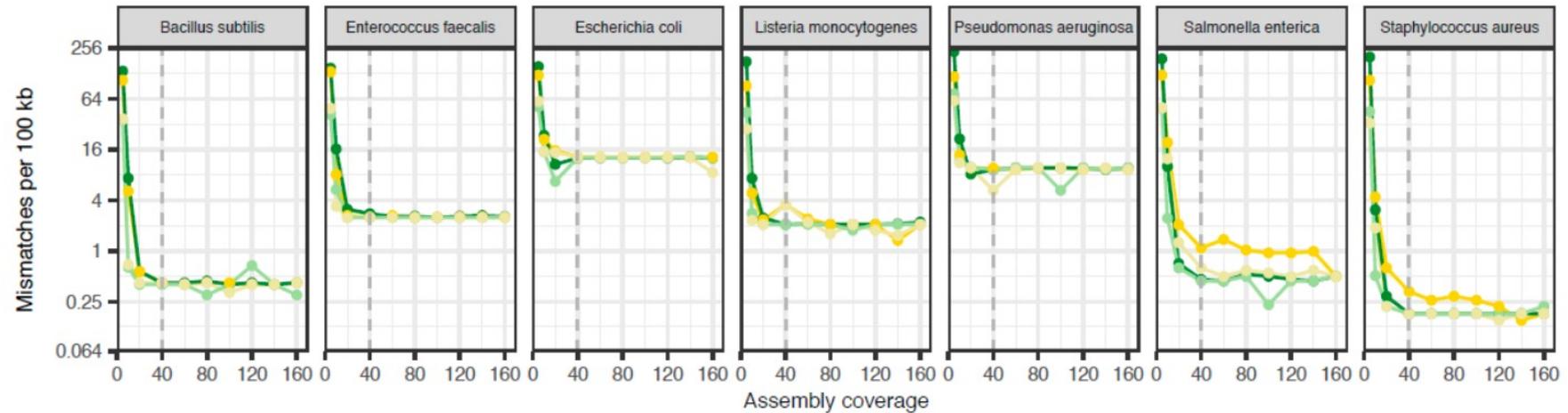
$$gANI_{G1 \rightarrow G2} = \frac{\sum_{bbh} (\text{Percent Identity} * \text{Alignment length})}{\text{lengths of BBH genes}}$$

ANI : Average Nucleotide Identity
BBH genes : genes having 70% or more identity and at least 70% coverage of the shorter gene

Assembly of small genomes : Nanopore vs PacBio

2 - 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*, July 2022

Sequencing and assembly statistics for the bacterial species ($n = 7$)

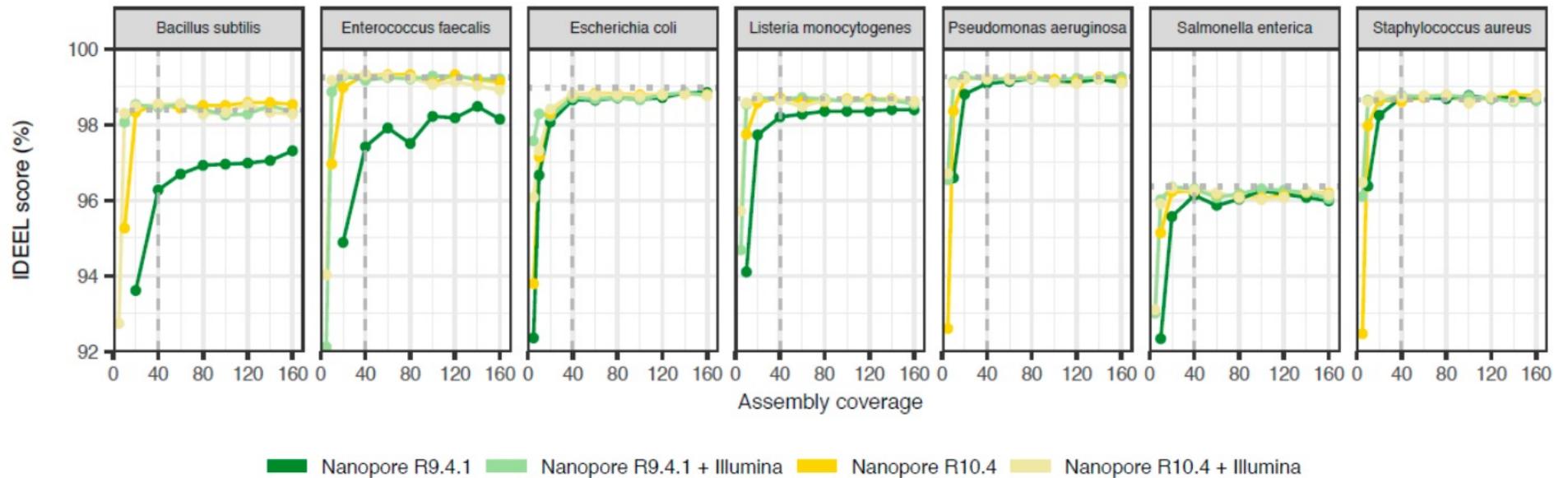


■ Nanopore R9.4.1 ■ Nanopore R9.4.1 + Illumina ■ Nanopore R10.4 ■ Nanopore R10.4 + Illumina

— Assembly of small genomes : Nanopore vs PacBio

2 - 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*, July 2022

Sequencing and assembly statistics for the bacterial species ($n = 7$)



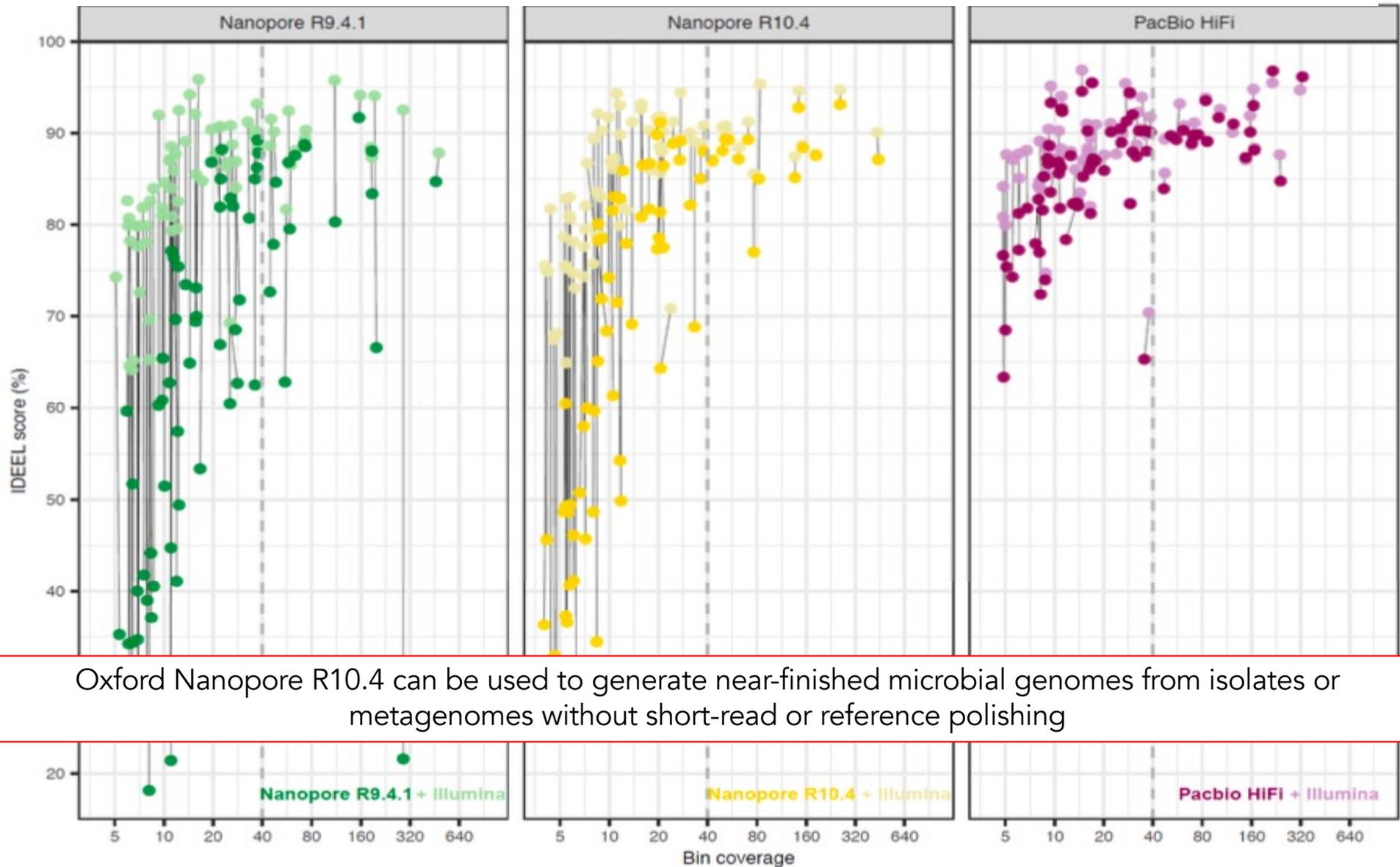
IDEEL score : proportion of predicted proteins that are $\geq 95\%$ the length of their best-matching known protein in a database

➤ No significant improvement in quality for R10.4 by addition of Illumina polishing

— Assembly of small genomes : Nanopore vs PacBio —

2 - 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*, July 2022

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



Oxford Nanopore R10.4 can be used to generate near-finished microbial genomes from isolates or metagenomes without short-read or reference polishing

— Assembly of small genomes : Nanopore vs PacBio

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

Large Genome assembly

Nanopore and/or PacBio

Large genome assembly : Nanopore + PacBio

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

Main publications

- 1 - The structure, function and evolution of a complete human chr. 8. Logsdon et al., *Nature*, May 2021
- 2 - The complete sequence of a human genome. Nurk et al., *Science* April 2022
- 3 - Chasing perfection: validation and polishing strategies for T2T genome assemblies. Cartney et al., *Nat. Methods* March 2022

Large projects using long reads

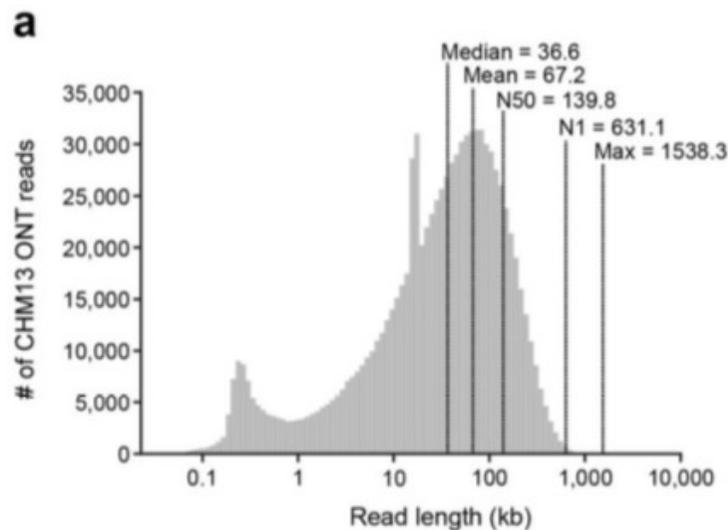
- 1 – Vertebrate Genome Project (VGP) :
Towards complete and error-free genome assemblies of all vertebrate species. Rhie et al. *Nature* 2021
- 2 – Human Pangenome Project :
 - Semi-automated assembly of high-quality diploid human reference genomes. Jarvis et al. *Nature* July 2022
 - The Human Pangenome Project: a global resource to map genomic diversity. Wang et al. *Nature* April 2022

Large genome assembly : Nanopore + PacBio

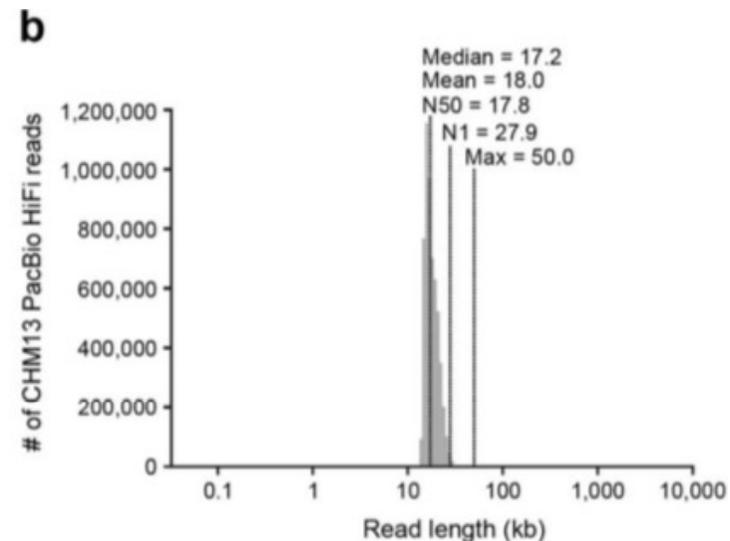
1 - The structure, function, and evolution of a complete human chromosome 8
Logsdon et al., *Nature*, May 2021

- Cell line : “complete hydatidiform mole” (CHM) derived from abnormal form of pregnancy
- Almost completely homozygous and therefore easier to assemble than heterozygous diploid genomes
- 20-fold sequence coverage of ONT ultra-long reads
- 32.4-fold coverage of PacBio HiFi

ONT ultra-long reads



PacBio HiFi reads



Large genome assembly : Nanopore + PacBio

1 - The structure, function, and evolution of a complete human chromosome 8
Logsdon et al., *Nature*, May 2021

- 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%.



- 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..

2 - The complete sequence of a human genome
Nurk et al. *Science* April 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. Bzikadze², Alla Mikheenko³, Mitchell R. Vollger⁴, Nicolas Altemose⁵, Lev Uralsky^{6,7}, Ariel Gershman⁸, Sergey Aganezov^{9†}, Savannah J. Hoyt¹⁰, Mark Diekhans¹¹, Glennis A. Logsdon⁴, Michael Alonge⁹, Stylianos E. Antonarakis¹², Matthew Borchers¹³, Gerard G. Bouffard¹⁴, Shelise Y. Brooks¹⁴, Gina V. Caldas¹⁵, Nae-Chyun Chen⁹, Haoyu Cheng^{16,17}, Chen-Shan Chin¹⁸, William Chow¹⁹, Leonardo G. de Lima¹³, Philip C. Dishuck⁴, Richard Durbin^{19,20}, Tatiana Dvorkina³, Ian T. Fiddes²¹, Giulio Formenti^{22,23}, Robert S. Fulton²⁴, Arkarachai Fungtammasan¹⁸, Erik Garrison^{11,25}, Patrick G. S. Grady¹⁰, Tina A. Graves-Lindsay²⁶, Ira M. Hall²⁷, Nancy F. Hansen²⁸, Gabrielle A. Hartley¹⁰, Marina Haukness¹¹, Kerstin Howe¹⁹, Michael W. Hunkapiller²⁹, Chirag Jain^{1,30}, Miten Jain¹¹, Erich D. Jarvis^{22,23}, Peter Kerpedjiev³¹, Melanie Kirsche⁹, Mikhail Kolmogorov³², Jonas Korlach²⁹, Milinn Kremitzki²⁶, Heng Li^{16,17}, Valerie V. Maduro³³, Tobias Marschall³⁴, Ann M. McCartney¹, Jennifer McDaniel³⁵, Danny E. Miller^{4,36}, James C. Mullikin^{14,28}, Eugene W. Myers³⁷, Nathan D. Olson³⁵, Benedict Paten¹¹, Paul Peluso²⁹, Pavel A. Pevzner³², David Porubsky⁴, Tamara Potapova¹³, Evgeny I. RogaeV^{6,7,38,39}, Jeffrey A. Rosenfeld⁴⁰, Steven L. Salzberg^{9,41}, Valerie A. Schneider⁴², Fritz J. Sedlazeck⁴³, Kishwar Shafin¹¹, Colin J. Shew⁴⁴, Alaina Shumate⁴¹, Ying Sims¹⁹, Arian F. A. Smit⁴⁵, Daniela C. Soto⁴⁴, Ivan Sovic^{29,46}, Jessica M. Storer⁴⁵, Aaron Streets^{5,47}, Beth A. Sullivan⁴⁸, Françoise Thibaud-Nissen⁴², James Torrance¹⁹, Justin Wagner³⁵, Brian P. Walenz¹, Aaron Wenger²⁹, Jonathan M. D. Wood¹⁹, Chunlin Xiao⁴², Stephanie M. Yan⁴⁹, Alice C. Young³⁴, Samantha Zarate⁹, Urvashi Surti⁵⁰, Rajiv C. McCoy⁴⁹, Megan Y. Dennis⁴⁴, Ivan A. Alexandrov^{3,7,51}, Jennifer L. Gerton^{13,52}, Rachel J. O'Neill¹⁰, Winston Timp^{8,41}, Justin M. Zook³⁵, Michael C. Schatz^{9,49}, Evan E. Eichler^{4,53*}, Karen H. Miga^{11,54*}, Adam M. Phillippy^{1*}

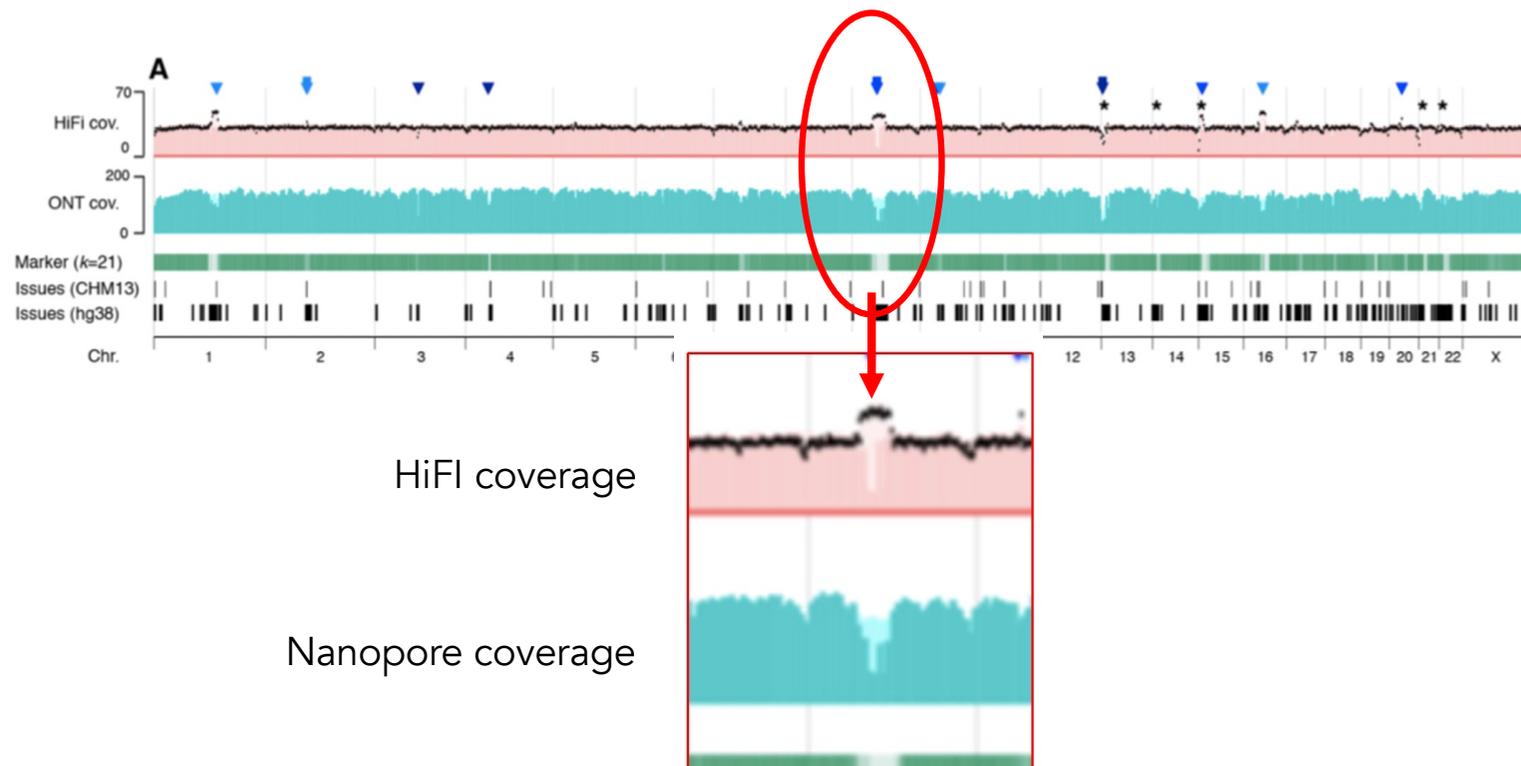
Large genome assembly : Nanopore + PacBio

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

SEQUENCING

Data were obtained with a "complete hydatidiform mole" (CHM13) cell line:

- 30× PacBio circular consensus sequencing (HiFi)
- 120× Oxford Nanopore ultra-long read sequencing (ONT)
- 100× Illumina PCR-Free sequencing
- 70× Illumina / Arima Genomics Hi-C (Hi-C)
- BioNano optical maps (11)
- Strand-seq

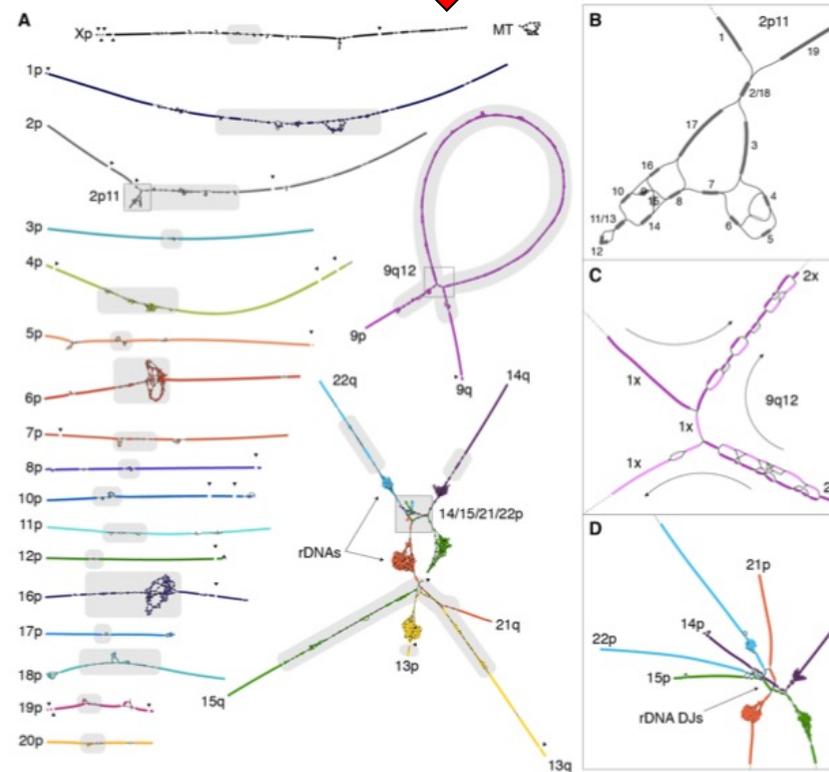


Large genome assembly : Nanopore + PacBio

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

ASSEMBLY

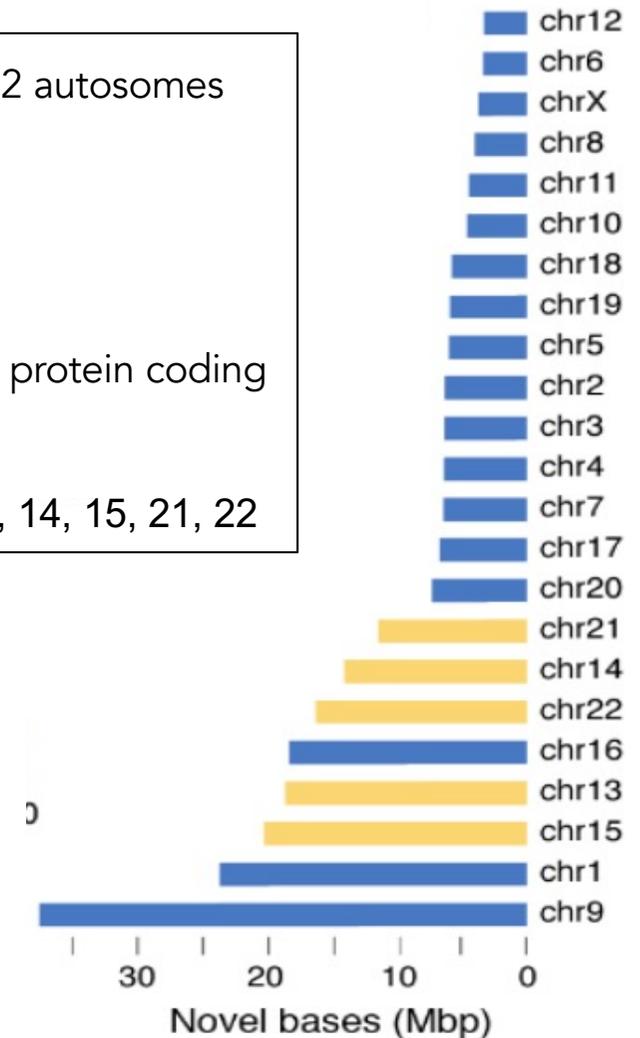
- HiFi-based string graph constructed using a purpose-built method that combines components from
 - HiCanu
 - Miniasm
 - specialized graph processing



Large genome assembly : Nanopore + PacBio

2 - The complete sequence of a human genome
Nurk et al. *Science* April 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
 - Identifies 2,226 paralogous gene copies, 115 of predicted as protein coding
 - all centromeric regions
 - entire short arms (p-arms) of 5 acrocentric chromosomes : 13, 14, 15, 21, 22



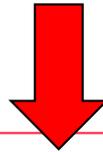
Large genome assembly : Nanopore + PacBio

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

Mc Cartney et al. *Nature Methods* March 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 corrected**



- **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 PROJECTS USING LONG READS

— 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^{1,303}, Shane A. McCarthy^{2,3,303}, Olivier Fedrigo^{4,303}, Joana Damas⁵,
Giulio Formenti^{1,6}, Sergey Koren¹, Marcela Uliano-Silva^{7,8}, William Chow⁷,
Arkarachai Fungtammasan⁹, Juwan Kim¹⁰, Chul Lee³⁰, Byung June Ko¹¹, Mark Chaisson¹²,
Gregory L. Gedman⁶, Lindsey J. Cantin⁶, Francoise Thibaud-Nissen¹³, Leanne Haggerty¹⁴,
Iliana Bista^{2,3}, Michelle Smith³, Bettina Haase⁴, Jacquelyn Mountcastle⁶, Sylke Winkler^{15,16},
Sadye Paez^{4,6}, Jason Howard¹⁷, Sonja C. Vernes^{18,19,20}, Tanya M. Lama²¹, Frank Grutzner²²,
Wesley C. Warren²³, Christopher N. Balakrishnan²⁴, Dave Burt²⁵, Julia M. George²⁶,
Matthew T. Biegler⁶, David Iorns²⁷, Andrew Digby²⁸, Daryl Eason²⁸, Bruce Robertson²⁹,
Taylor Edwards³⁰, Mark Wilkinson³¹, George Turner³², Axel Meyer³³, Andreas F. Kautt^{33,34},
Paolo Franchini³⁵, H. William Detrich III³⁵, Hannes Svoldal^{36,37}, Maximilian Wagner³⁸,
Gavin J. P. Naylor³⁹, Martin Pippel^{15,40}, Milan Malinsky^{2,41}, Mark Mooney⁴², Maria Simbirsky²,
Brett T. Hannigan⁹, Trevor Pesout⁴³, Marlys Houck⁴⁴, Ann Misuraca⁴⁴, Sarah B. Kingan⁴⁵,
Richard Hall⁴⁵, Zev Kronenberg⁴⁵, Ivan Sovic^{45,46}, Christopher Dunn⁴⁵, Zemin Ning³,
Alex Hastie⁴⁷, Joyce Lee⁴⁷, Siddarth Selvaraj⁴⁸, Richard E. Green^{43,49}, Nicholas H. Putnam⁵⁰,
Ivo Gut^{51,52}, Jay Ghurye^{40,53}, Erik Garrison⁴³, Ying Sims³, Joanna Collins³, Sarah Pelan³,
James Torrance³, Alan Tracey³, Jonathan Wood³, Robel E. Dagnew¹², Dengfeng Guan^{2,54},
Sarah E. London⁵⁵, David F. Clayton⁵⁶, Claudio V. Mello⁵⁷, Samantha R. Friedrich⁵⁷,
Peter V. Lovell⁵⁷, Ekaterina Osipova^{35,40,58}, Farooq O. Al-Ajli^{59,60,61}, Simona Secomandi⁶²,
Hee-bal Kim^{30,31,63}, Constantina Theofanopoulou⁶, Michael Hiller^{64,65,66}, Yang Zhou⁶⁷,
Robert S. Harris⁶⁸, Kateryna D. Makova^{69,69,70}, Paul Medvedev^{69,70,71,72}, Jinna Hoffman¹³,
Patrick Masterson¹³, Karen Clark¹³, Fergal Martin¹⁴, Kevin Howe¹⁴, Paul Flicek¹⁴,
Brian P. Walenz¹, Woori Kwak^{63,73}, Hiram Clawson⁴², Mark Diekhans⁴³, Luis Nassar⁴³,
Benedict Paten⁴³, Robert H. S. Kraus^{33,74}, Andrew J. Crawford⁷⁵, M. Thomas P. Gilbert^{76,77},
Guojie Zhang^{78,79,80,81}, Byrappa Venkatesh⁸², Robert W. Murphy⁸³, Klaus-Peter Koepfli⁸⁴,
Beth Shapiro^{85,86}, Warren E. Johnson^{84,87,88}, Federica Di Palma⁸⁹, Tomas Marques-Bon
et^{90,91,92,93}, Emma C. Teeling⁹⁴, Tandy Warnow⁹⁵, Jennifer Marshall Graves⁹⁶,
Oliver A. Ryder^{44,97}, David Haussler^{42,85}, Stephen J. O'Brien^{98,99}, Jonas Korlach⁴⁵,
Harris A. Lewin^{5,100,101}, Kerstin Howe^{3,104}, Eugene W. Myers^{35,40,102,104},
Richard Durbin^{2,3,304}, Adam M. Phillippy^{1,304} & Erich D. Jarvis^{4,6,86,304}

VERTEBRATE GENOMES PROJECT (VGP)

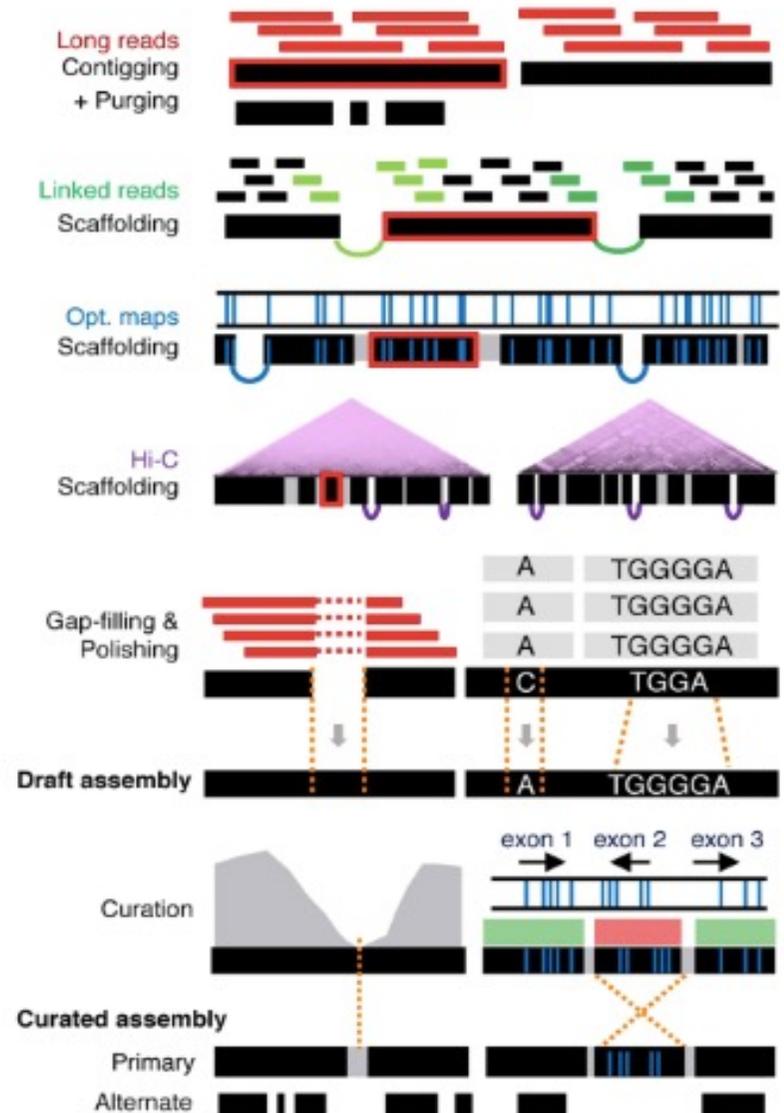
Towards complete and error-free genome assemblies of all vertebrate species
Rhie et al. *Nature* 2021

VGP assembly pipeline applied across multiple species

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

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

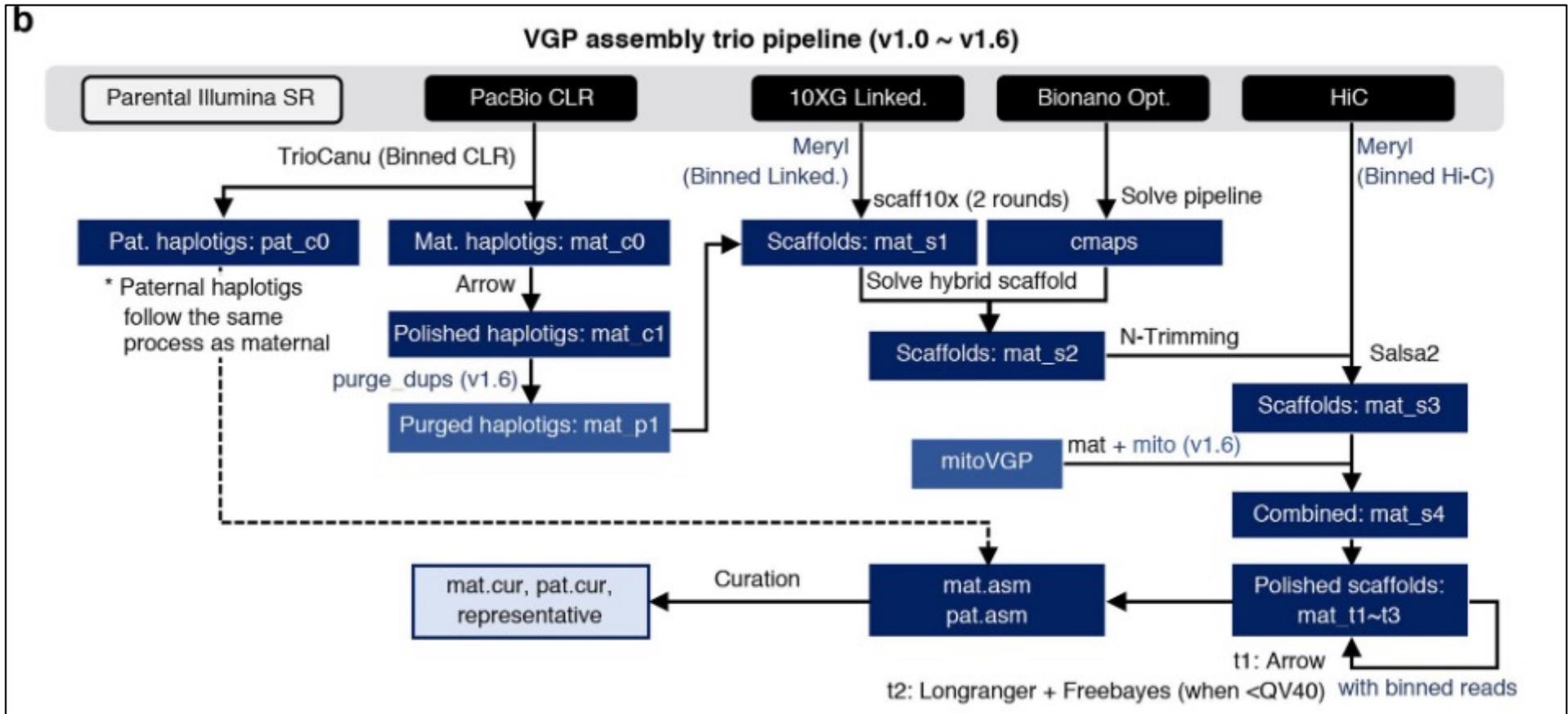
We will take advantage of continuing improvements in genome sequencing technology, assembly, and annotation, including advances in PacBio HiFi reads, Oxford Nanopore reads, and replacements for 10XG reads



— VERTEBRATE GENOMES PROJECT (VGP) —

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

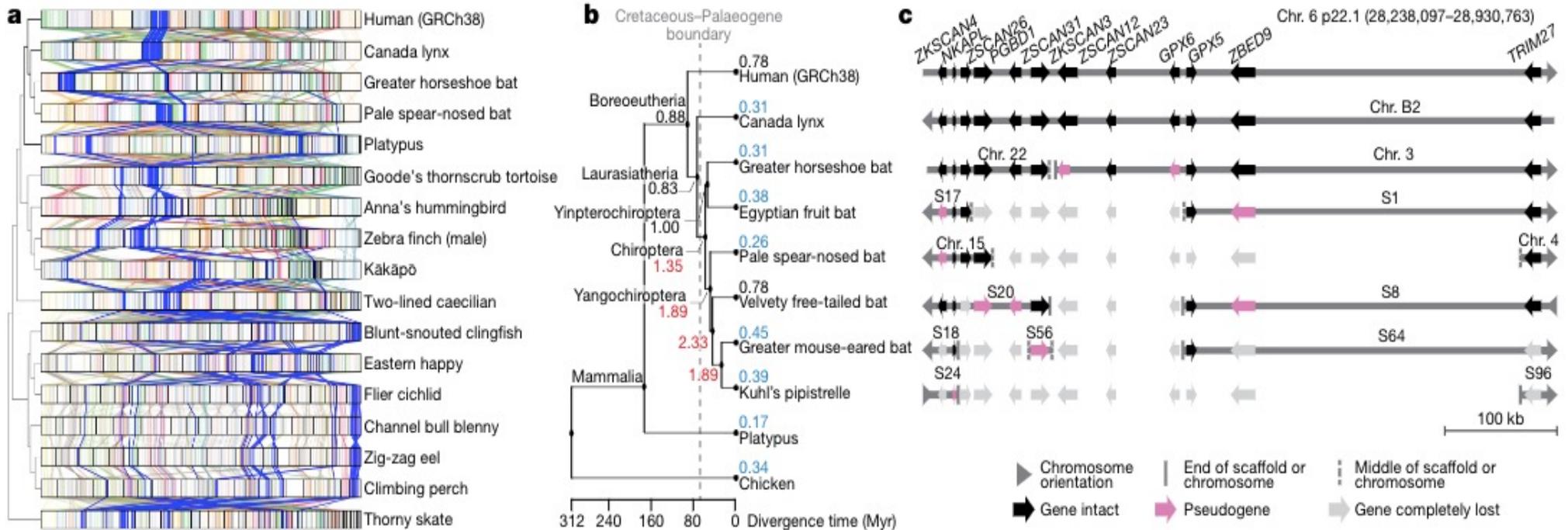
Standard VGP trio assembly pipeline when DNA is available for a child and parents²⁰.



VERTEBRATE GENOMES PROJECT (VGP)

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

One example : chromosome evolution among bats and other vertebrates



Chromosome synteny maps across the species sequenced based on BUSCO gene alignments

THE HUMAN PANGENOME PROJECT

Semi-automated assembly of high-quality diploid human reference genomes
Jarvis et al. *Nature* July 2022

- T2T-CHM13 is almost totally haploid and assembly also required a substantial amount of manual curation by dozens of people over many months -> additional developments are needed to assemble diploid genomes at high quality
- Approaches using highly accurate long reads with graph-based haplotype phasing outperformed those that did not.
- -> The goal is to determine which combination of approaches yield the most complete and accurate human diploid genome assembly with minimal manual curation.

- The high-quality Pangenome reference will :
 - represent > 99% of diversity for minor alleles of > 1% frequency in human population
 - Include at least 350 reference quality haplotype-phased human diploid genomes (700 haplotypes in total)



- 'assemblyathon' to produce the most complete and accurate genome possible in a near-automated way.



- 23 genome assemblies, generated with 23 different methods

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 120kb 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

Semi-automated assembly of high-quality diploid human reference genomes
Jarvis et al. *Nature* July 2022

- Extensive evaluation of more than 60 metrics led to an approach with the highest scores
- Key factors were :
 - the use of mother–father–offspring trio data to resolve haplotypes **during the assembly rather than before or after it**
 - amalgamating different types of sequence data and assembly tools **simultaneously, as opposed to sequentially**
- This study applied the best-performing method for producing human genome assemblies for the pangenome
- And it gave rise to the **highest-quality and most-complete diploid human genome assembled so far**

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-Fulton¹, Eddie Belter¹, Sarah Cody¹, Changxu Fan^{1,2,3}, Paul Flicek⁴, Ira M. Hall⁵, David Haussler^{6,7}, Heather A. Lawson^{1,2,3}, Daofeng Li^{1,2,3}, Joshua F. McMichael¹, Karen H. Miga⁸, Benedict Paten⁹, Chad Tomlinson¹, 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 Bustamante⁸, Judy Cho^{9,10,11}, Robert Cook-Deegan¹², Jean-Francois Deleuze¹³, Richard Durbin^{14,15}, Simon Eastaugh¹⁶, Evan E. Eichler^{17,18}, Xiaowen Feng^{19,20}, Nanibaa Garrison^{21,22,23}, Nadine Gassner⁶, Mary Goldman⁶, Ed Green⁶, David Haussler^{6,7}, Erich D. Jarvis^{24,25}, Eimear E. Kenny^{9,11}, Barbara A. Koenig²⁶, Bastien Llamas^{27,28}, Nicole C. Lockhart²⁹, Bartha M. Knoppers³⁰, Ann M. McCartney³¹, Karen H. Miga⁸, Jessica Mozerky³², Hardip Patel^{27,28}, Alice B. Popejoy³³, Charles Rotimi³⁴, Charmaine Royal³⁵, Yassine Souilmi^{27,28}, Nathan O Stitzel^{1,2,36}, Lisa Wang^{9,11}

Technology and Production Working Group

Mark Akeson⁶, Brandy Baird⁶, Giulio Formenti^{24,25}, Robert S. Fulton¹, Ed Green⁶, Miten Jain⁶, Brittany Kerr³⁷, Chris Markovic¹, Matthew W. Mitchell³⁷, Katy Munson¹⁷, Hugh Olsen⁶, Sadye Paez^{24,25}, William Rowell³⁸, Sam Sacco³⁹, Lauren Shalmiyev^{24,25}, Arvis Sulovari¹⁷

Assembly, T2T, and Pangenome Working Group

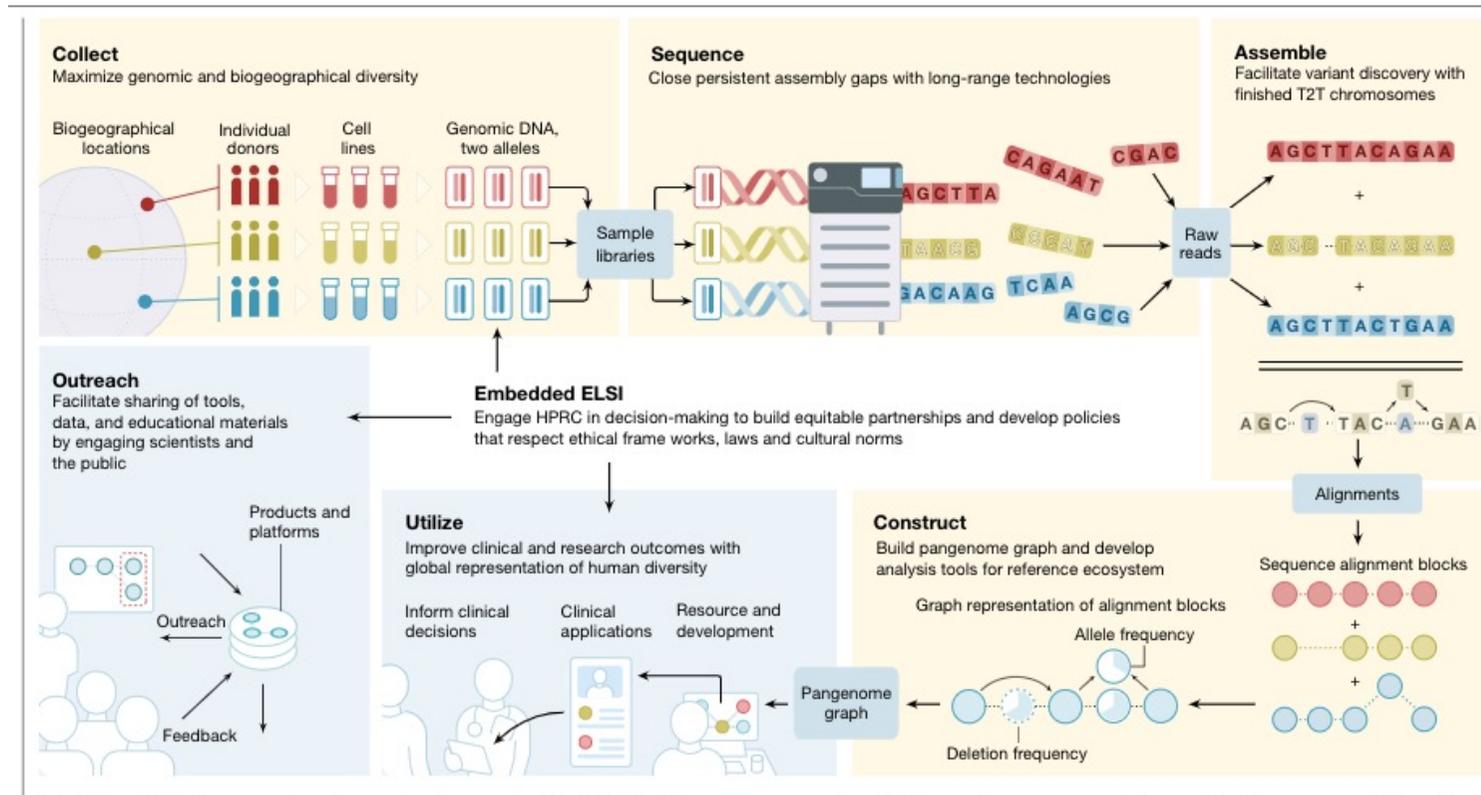
Mobin Asri⁶, Pete Audano¹⁷, Paolo Carnevali⁴⁰, Mark Chaisson⁴¹, Shubham Chandak⁴², Xian Chang⁶, Haoyu Cheng^{19,20}, Vincenza Colonna⁴³, Daniel Doerr⁴⁴, Peter Ebert⁴⁴, Jana Ebler⁴⁴, Evan E. Eichler^{17,18}, Jordan Eizenga⁶, Olivier Fedrigo^{24,25}, Xiaowen Feng^{19,20}, Christian Fischer⁴⁵, Stacey Gabriel⁴⁶, Yan Gao⁴⁷, Shilpa Garg^{19,20,48}, Kiran Garimelle⁴⁶, Erik Garrison⁴⁵, Ed Green⁶, Stephanie Greer⁴⁹, Andrea Guarracino⁵⁰, Ira M. Hall⁵, William Harvey¹⁷, Marina Haukness⁶, David Haussler^{6,7}, Simon Heumos⁵¹, Glenn Hickey⁶, Kerstin Howe¹⁵, Eric D. Jarvis^{24,25}, Hanlee Ji⁴⁹, Sergey Koren³¹, Hojoon Lee⁴², Heng Li^{19,20}, Wen-Wei Liao⁵, Ryan Lorig-Roach⁶, Ernesto Lowy⁴, Tony Tsung Yu Lu⁴¹, Shuangjia Lu⁵, Julian Lucas⁶, Rebecca Serra Mari⁴⁴, Dmitri Pavlichin⁴⁹, Pierre Marjion⁴⁴, Charles Markello⁶, Tobias Marschall⁴⁴, Melissa Merediths⁶, Karen H. Miga⁸, Jean Morlong⁶, Njagi Mwaniki^{45,52}, Eugene W. Myers^{53,54,55}, Adam M. Novack⁶, Sergey Nurk³¹, Benedict Paten⁹, Dmitri Pavlichin⁴², Trevor Pesout⁶, Adam M. Phillippy³¹, Brandon Pickett³¹, David Porubsky¹⁷, Piotr Prins⁴⁵, Mikko Rautiainen³¹, Arang Rhie³¹, Kishwar Shafiq⁶, Jonas Sibbesen⁶, Jouni Siren⁶, Varsha Sreekanth⁶, Arvis Sulovari¹⁷, Kedar Tatwawadi⁴², Flavia Villani⁴¹, Mitchell Volger¹⁷, Alexander Wait Zaranek⁴⁸, Tsachy Weissman⁴²

Annotation, Maintenance and Improvement Working Group

Derek Albracht¹, Eddie Belter¹, Shelby Bidwell⁵⁶, Konstantinos Billis⁴, Caryn Carson^{1,2,3}, Karen Clark⁵⁶, Mark Diekhans⁶, Sarah Dyer⁴, Susan Fairley^{4,57}, Paul Flicek⁴, Adam Frankish⁴, Nadine Gassner⁶, Carlos Garcia Giron⁴, Mary Goldman⁶, Tina A. Graves-Lindsay¹, Marina Haukness⁶, Kevin Howe¹⁵, Sarah Hunt⁴, Paul Kitts⁵⁶, Milinn Kremitzki¹, Fergal Martin²³, Terence Murphy³⁰, Valerie Schneider, Francoise Thibaud-Nissen³⁰, Sergey Nurk¹³, David Thybert⁴, Thomas Walsh⁴, Ting Wang^{1,2,3}, Chunlin Xiao⁵⁶, Daniel Zerbino⁴, Xiaoyu Zhuo^{1,2,3}

THE HUMAN PANGENOME PROJECT

Wang *et al.* The Human Pangenome Project: a global resource to map genomic diversity. *Nature* **604**, 437–446, 2022



Goals of the Human Pangenome Project

- Identify individuals from diverse genomic and biogeographical backgrounds
- include at least 350 reference quality haplotype-phased human diploid genomes (700 haplotypes in total)
- **to generate the highest quality phased genomes possible, prioritize the use of long-read and long-range technologies for assemblies, with haplotype-aware algorithms**
- These assemblies will pinpoint all genetic differences, both large and small, at the base-pair level.
- As long-read sequencing costs fall and pangenome methods evolve, **we predict that patient samples will probably be sequenced using long-read technology.**

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
- 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..)



Conclusion

- Whereas HiFi sequencing excels at differentiating subtly diverged repeat copies or haplotypes, ultra-long nanopore sequencing excels at spanning long, identical repeats.
- For large genomes, using these technologies simultaneously will likely improve the assembly