LONG READS

Claude THERMES

PLATEFORME DE SÉQUENÇAGE I2BC

INSTITUT DE BIOLOGIE INTÉGRATIVE DE LA CELLULE

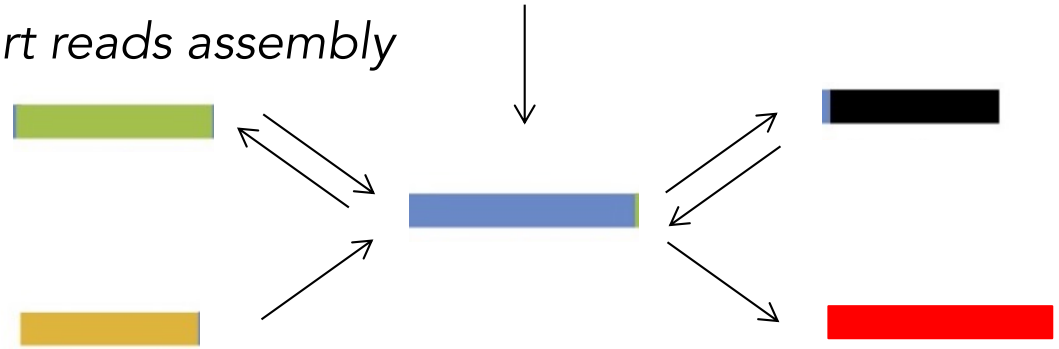
GIF-SUR-YVETTE

11ème ÉCOLE DE BIOINFORMATIQUE EBAII - 14/11/2022

LONG-READS VERSUS SHORT-READS

Assembly of DNA fragments with repeated sequences

NGS short reads assembly

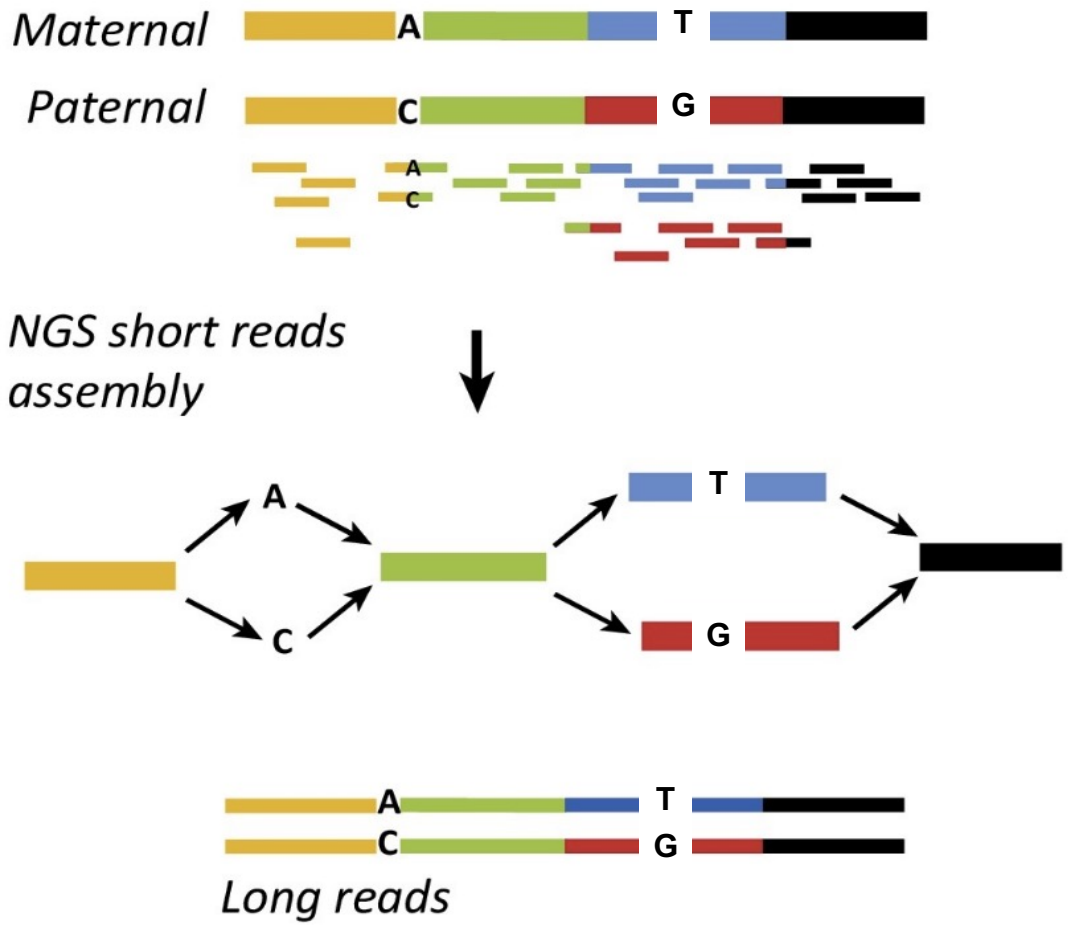


Several contigs → incomplete assembly, underestimation of repeats

Long reads assembly

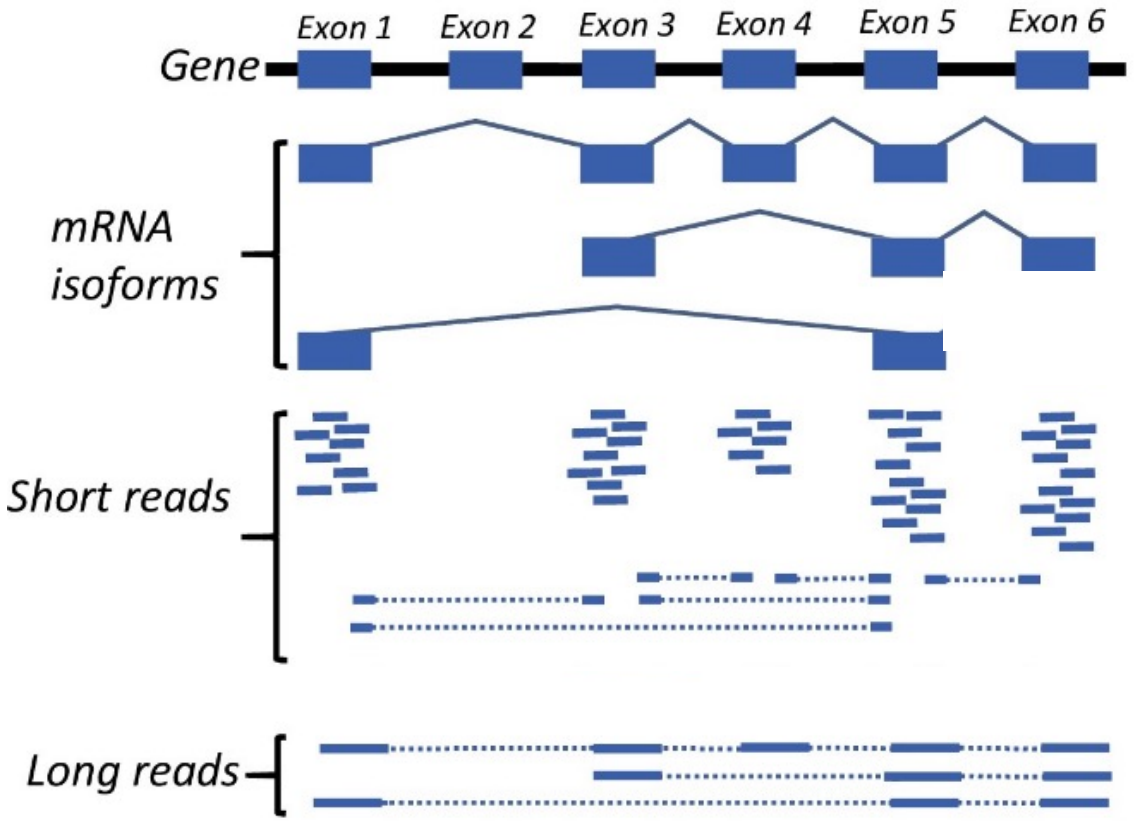
Long-reads (> 10 kb) allow assembly of large repeat-rich regions

Haplotype phasing



Long-reads facilitate phasing of maternal and paternal haplotypes

Detection of splicing isoforms



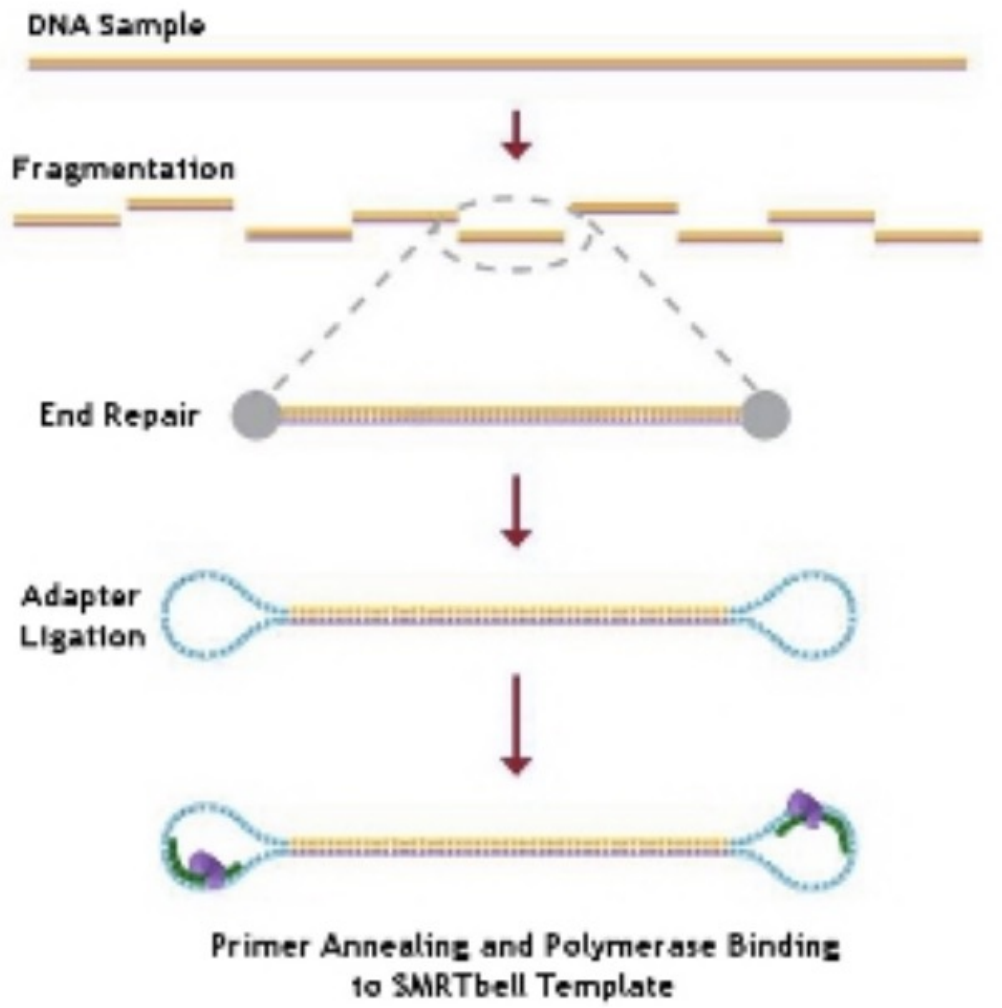
Long-reads allow identification of multiple splicing events along mRNAs

The 3rd generation winning technologies

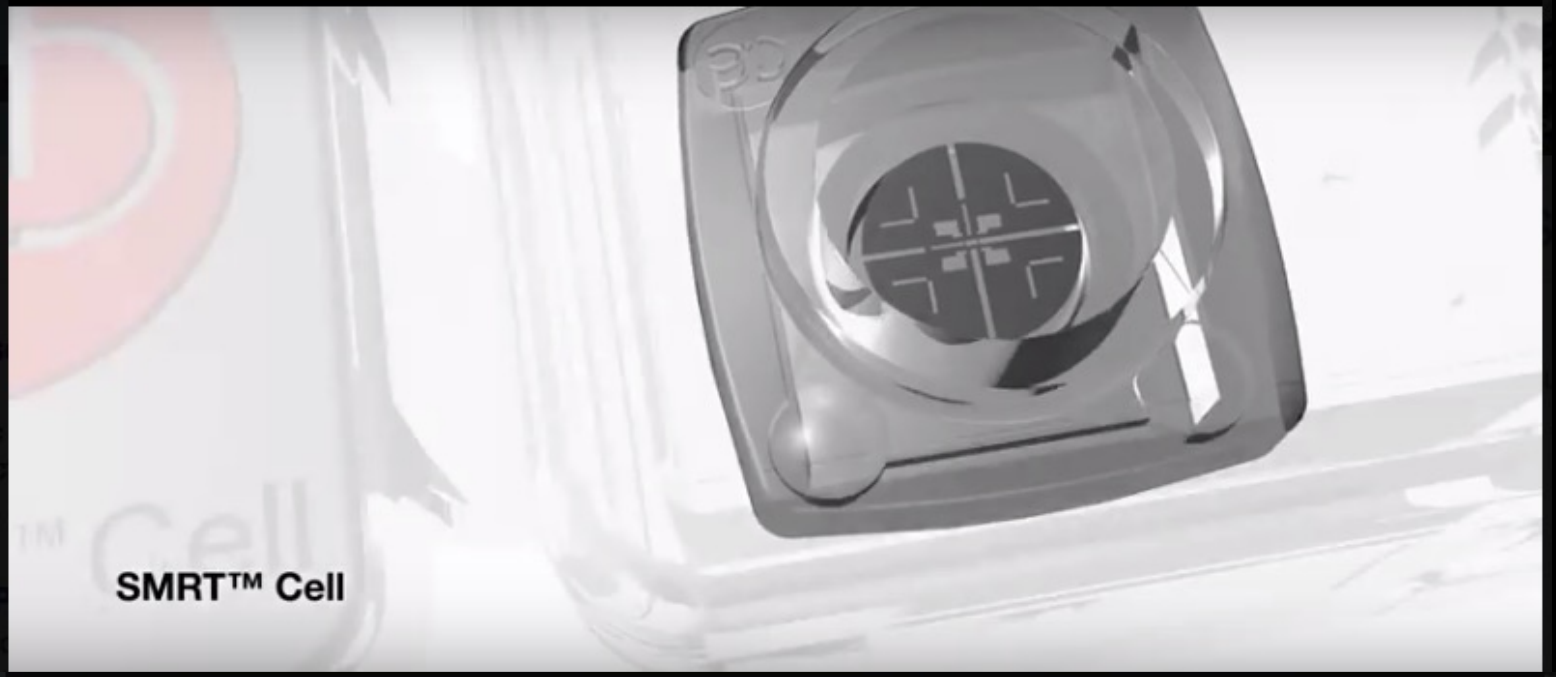
Sequel – Revio / Pacific Biosciences
Single molecules
Up to 200 kbp long

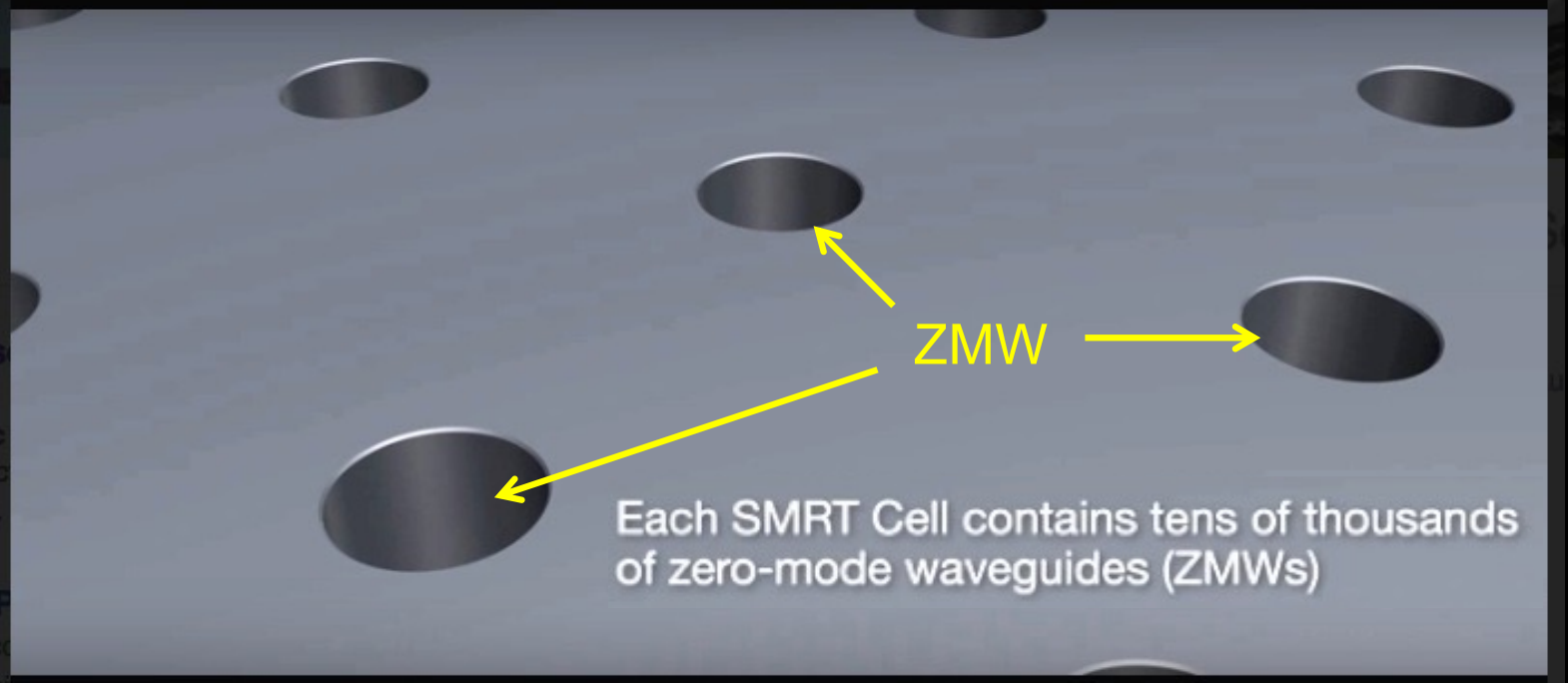
MinION – PromethION - Oxford Nanopore
Single molecules
> 1 Mbp long

PacBio DNA-seq library

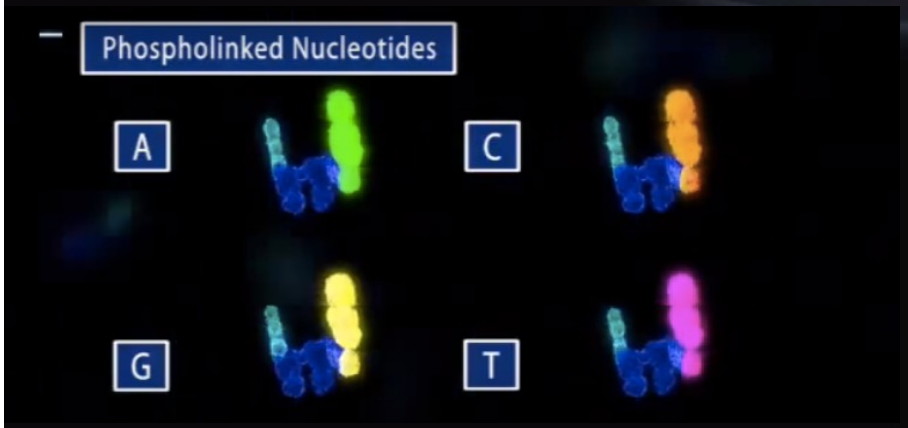


PACIFIC BIOSCIENCES

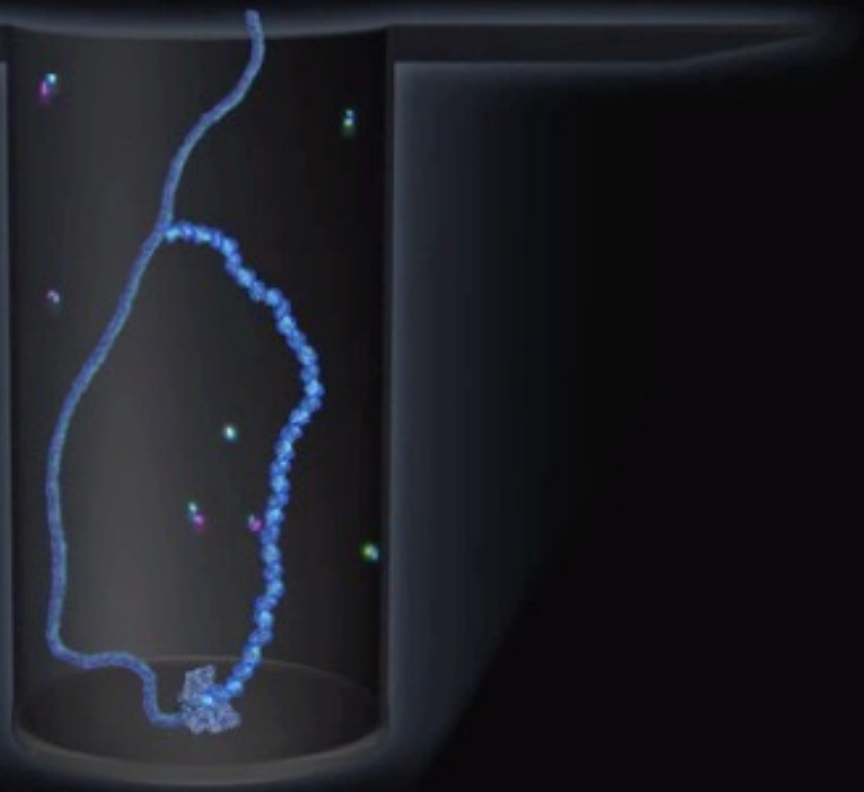




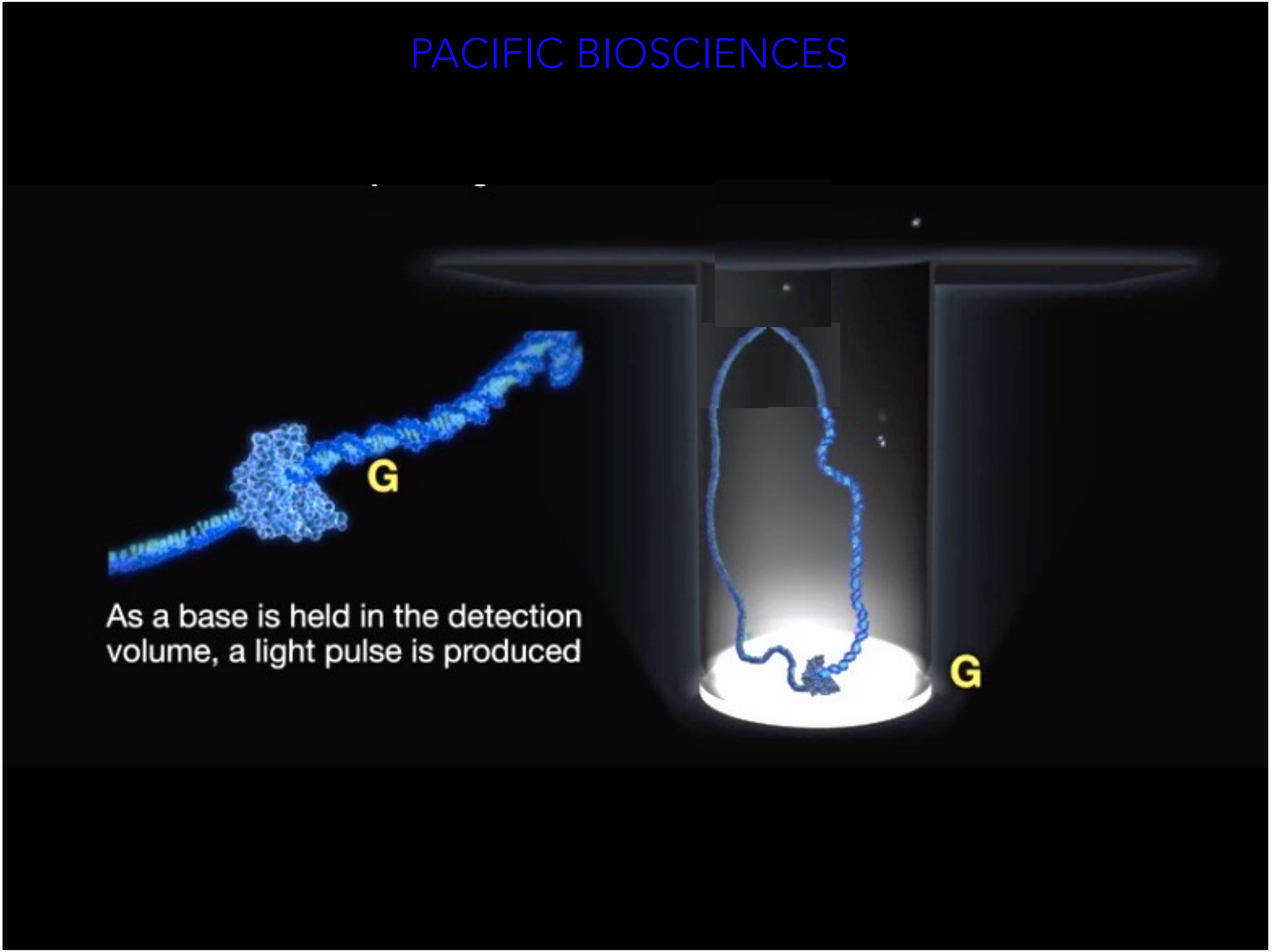
PACIFIC BIOSCIENCES

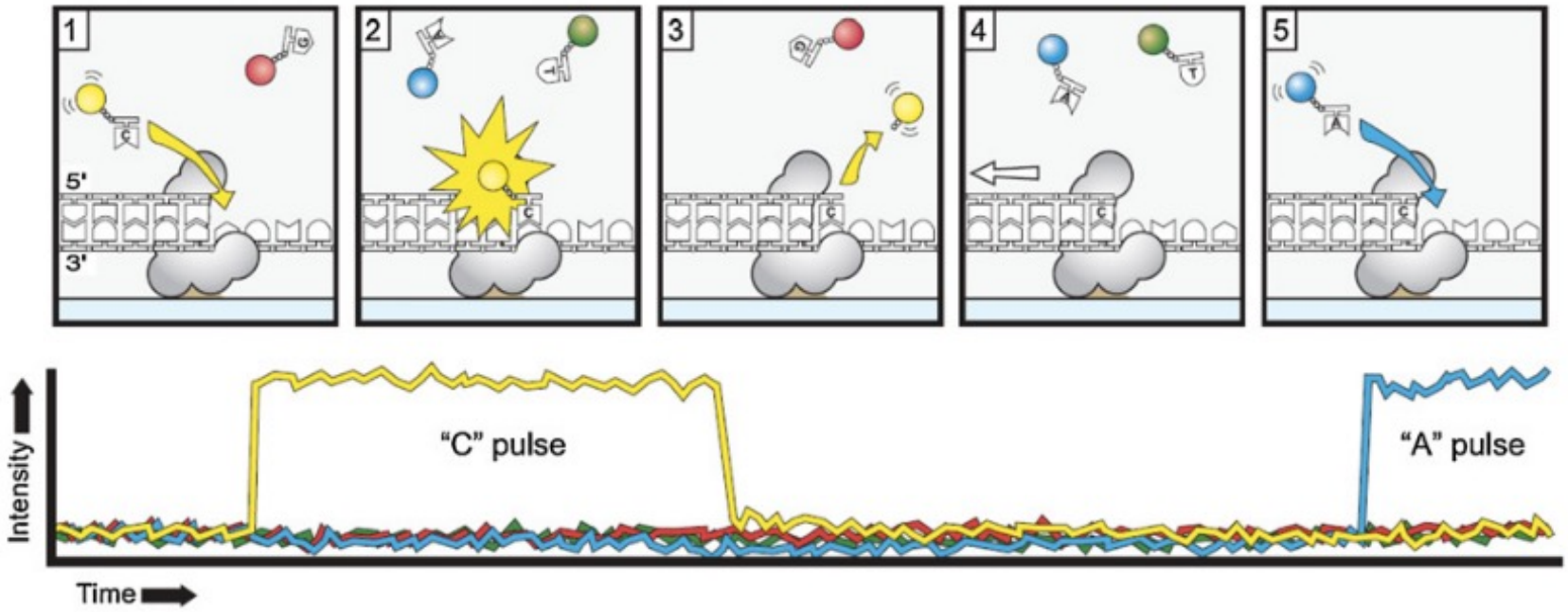


Phospholinked nucleotides are introduced into the ZMW chamber

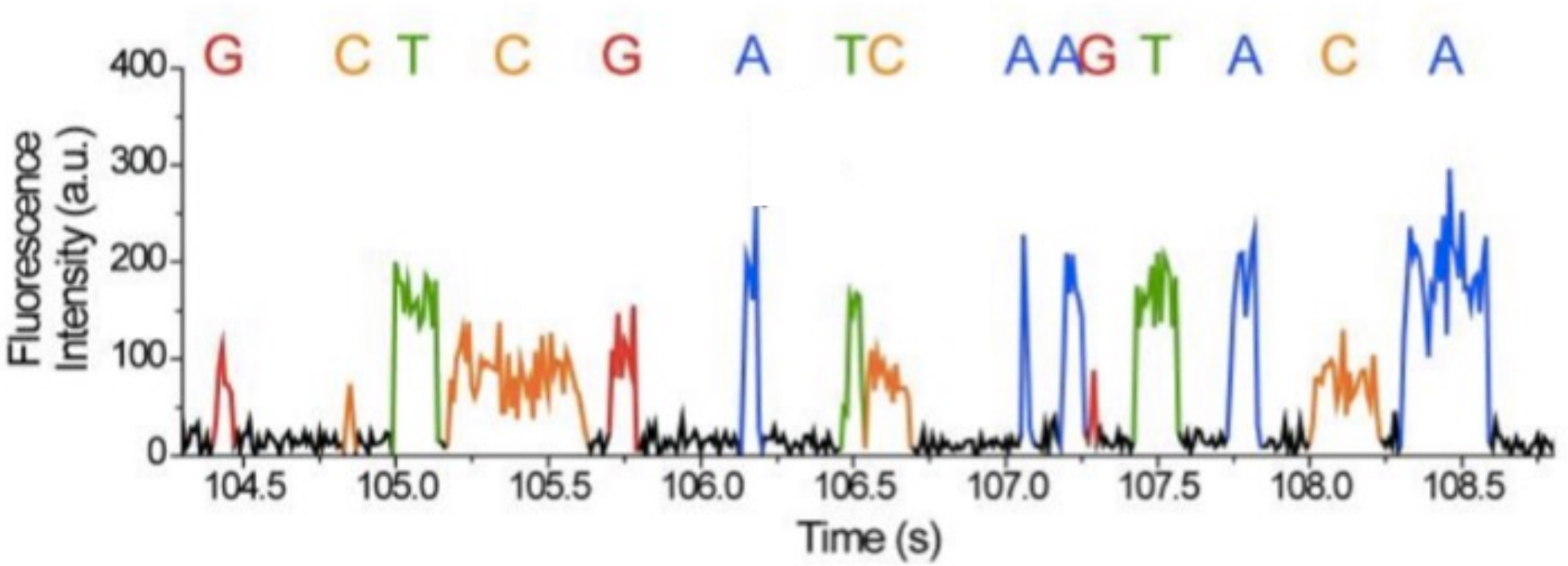


PACIFIC BIOSCIENCES

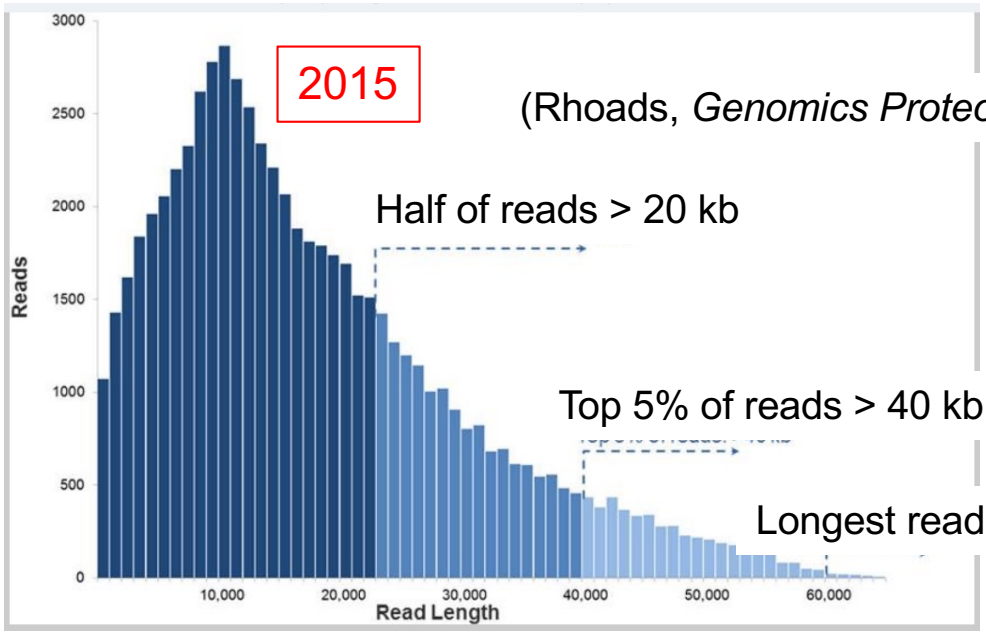




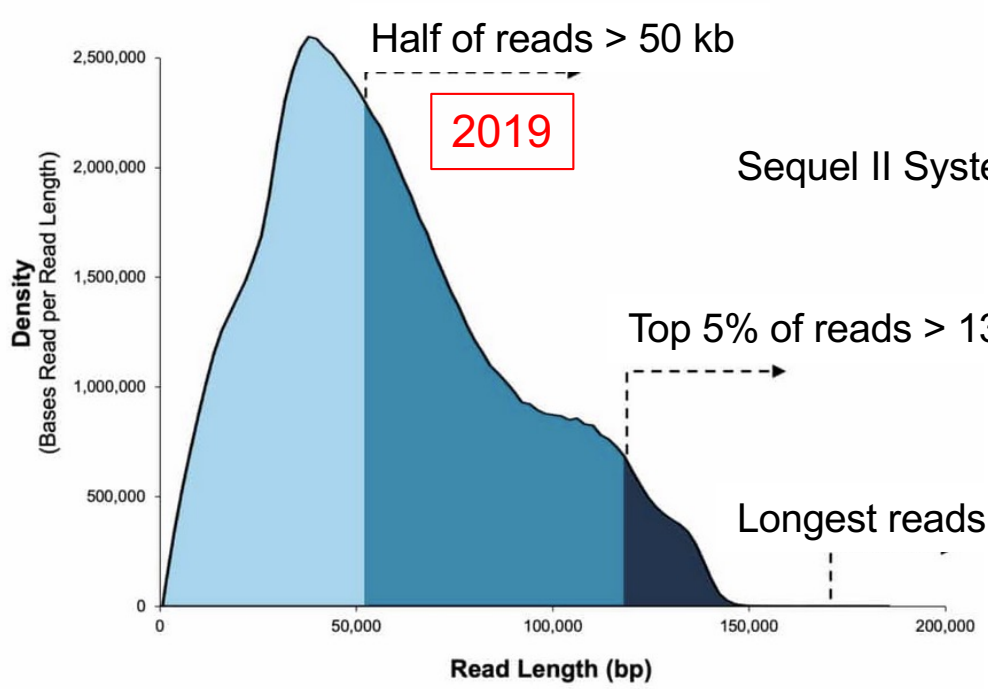
Eid, J., et al. Science (2009)


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

Length of PacBio reads

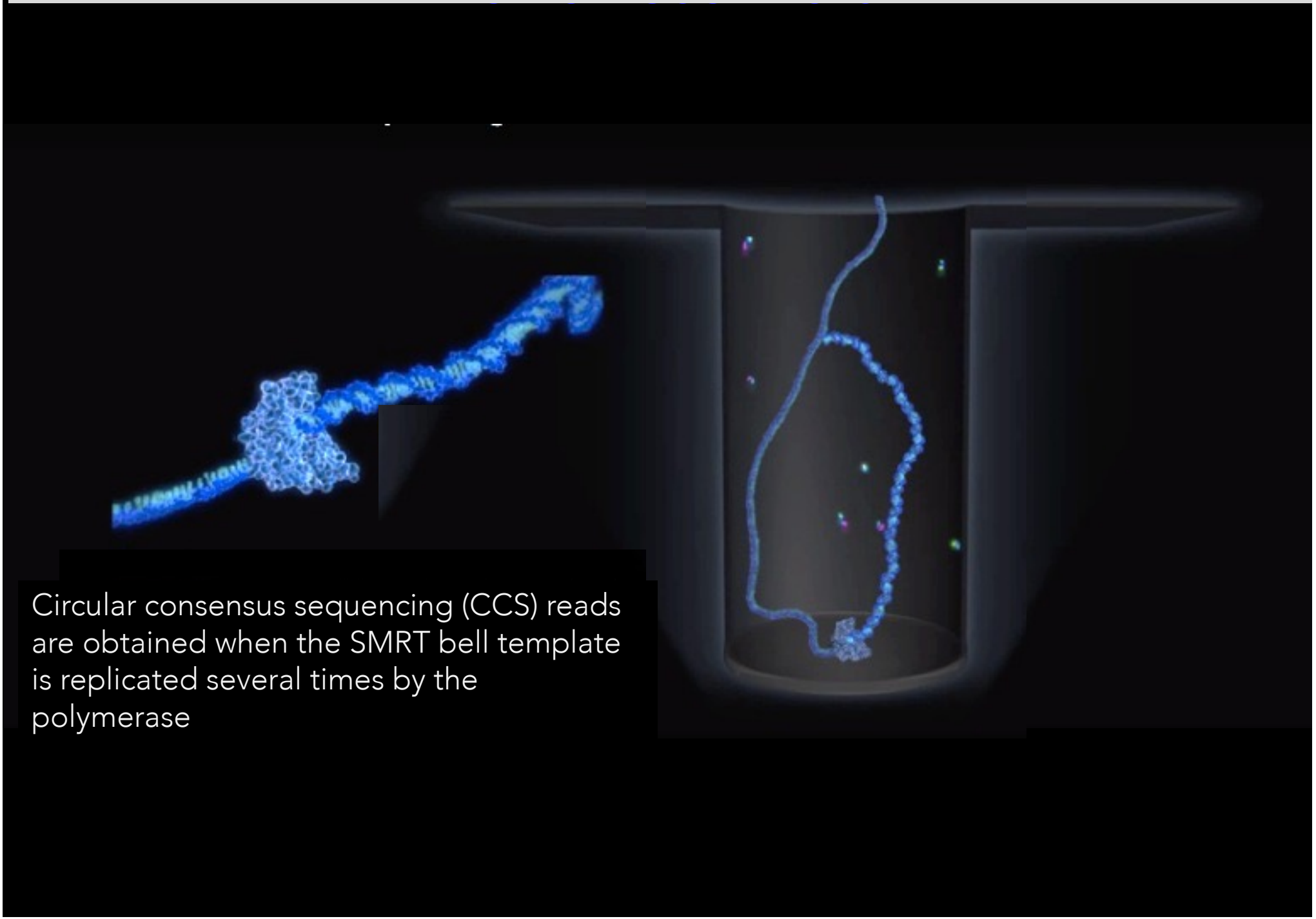


(Rhoads, *Genomics Proteomics Bioinformatics*, 2015)



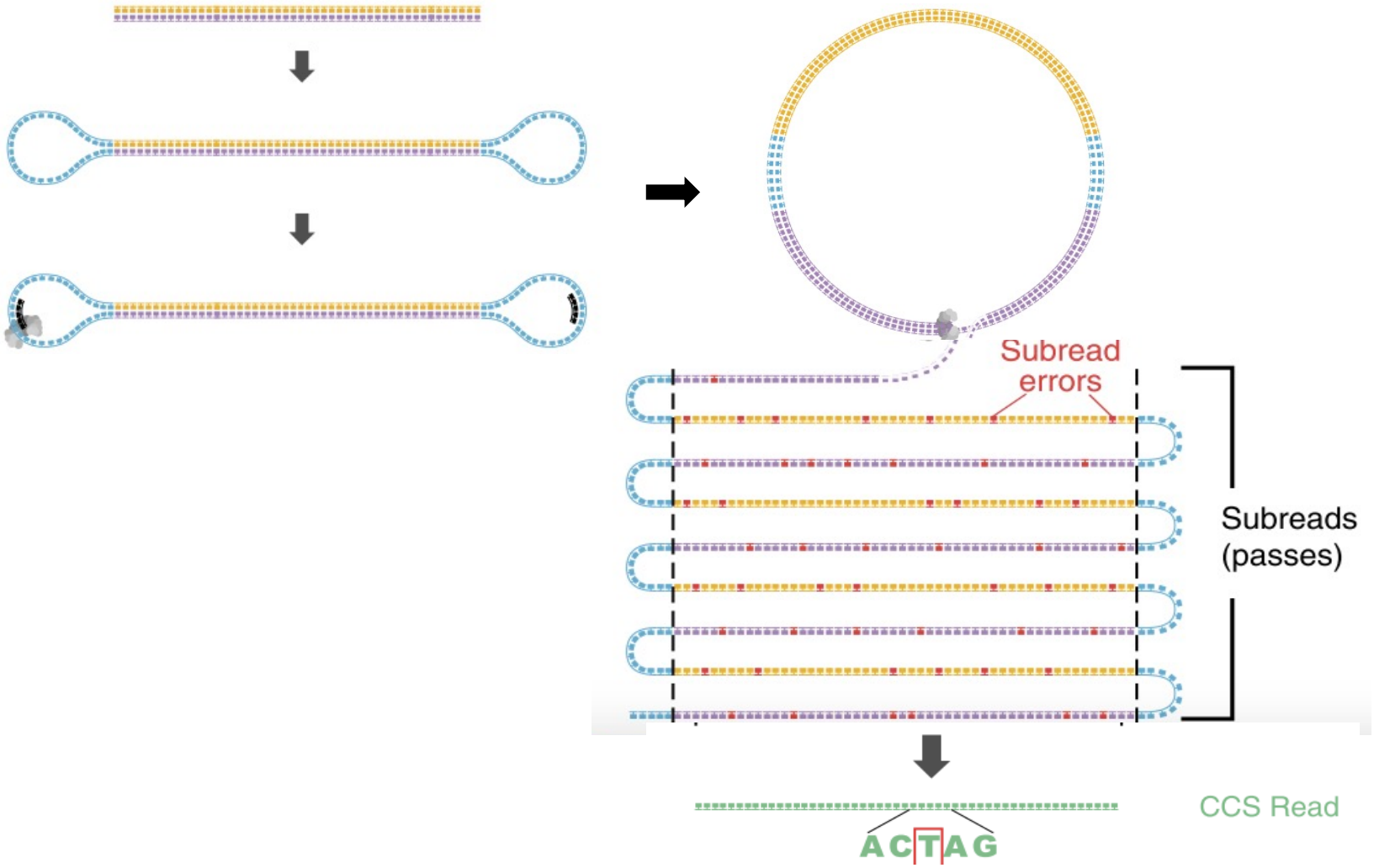
Sequel II System ; 2.0 Chemistry

Improvement with new chemistry : Circular Consensus Sequence (CCS)

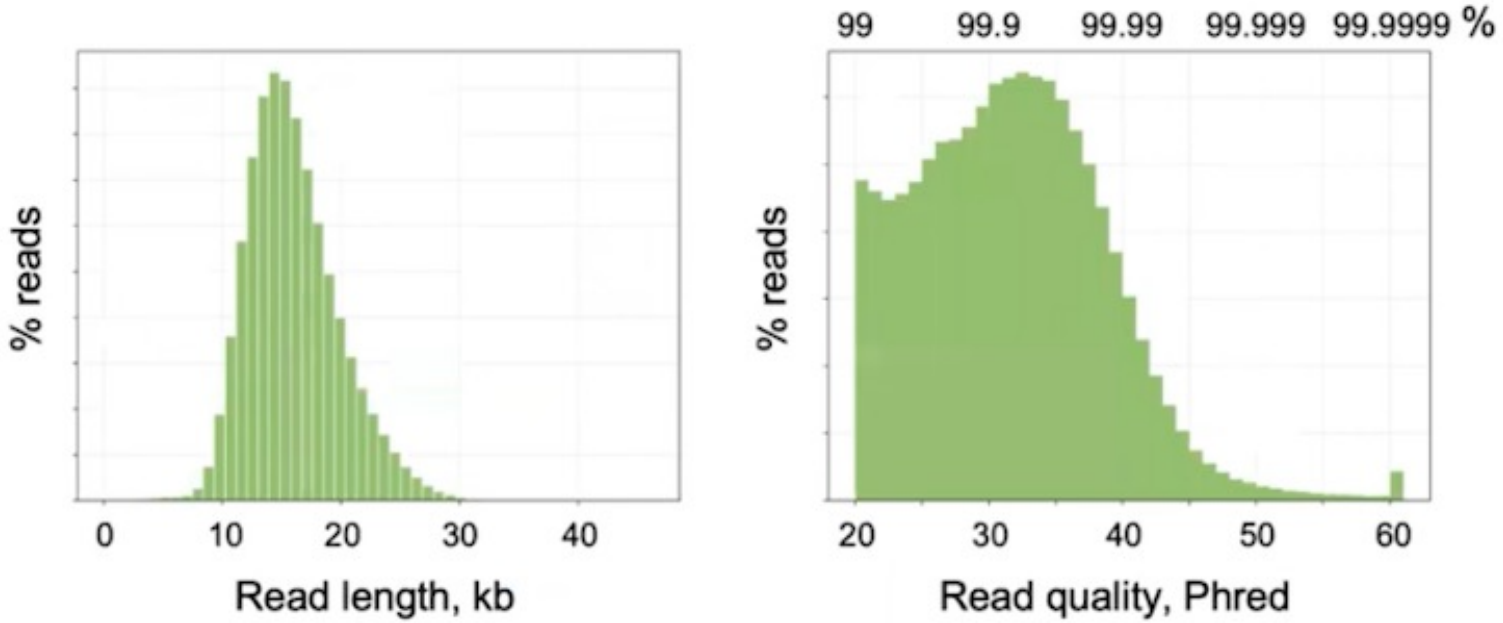


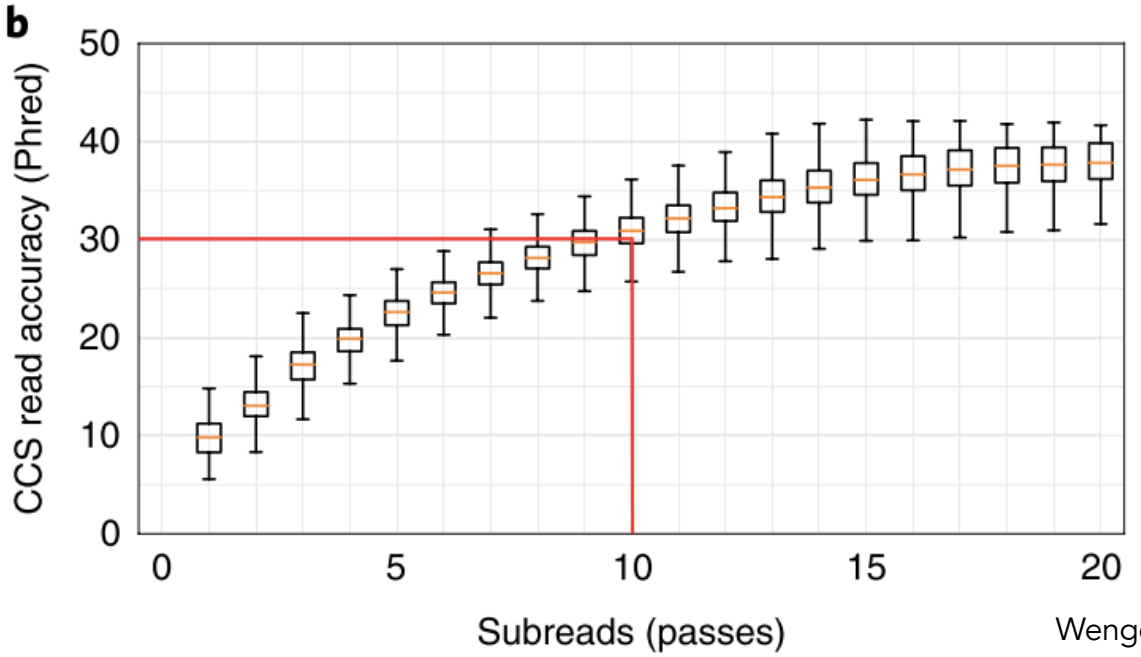
Circular consensus sequencing (CCS) reads are obtained when the SMRT bell template is replicated several times by the polymerase

Circular Consensus Sequences (CCS): HIFI READS



GENOME ASSEMBLY WITH CCS





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

GENOME ASSEMBLY WITH CCS

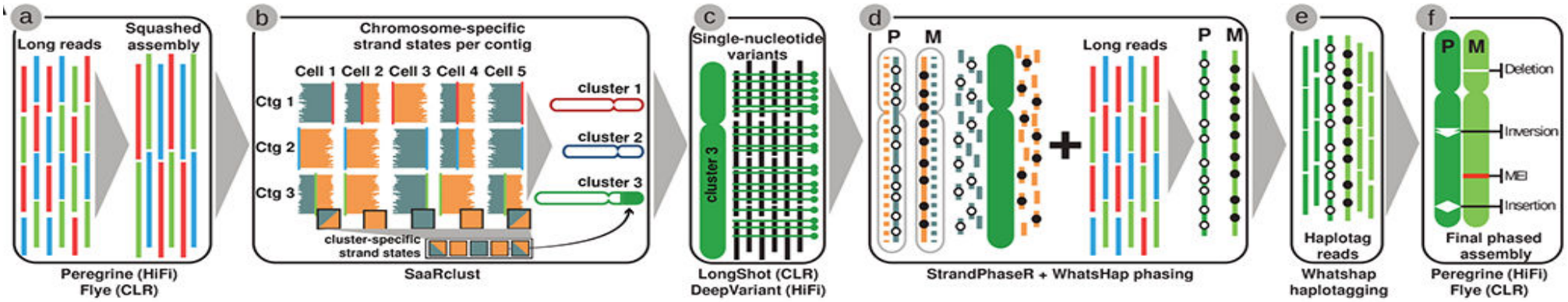
1 - Haplotype-resolved diverse human genomes and integrated analysis of structural variation
Ebert et al. *Science* 2021

New methodology that combines :

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

Methodology

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

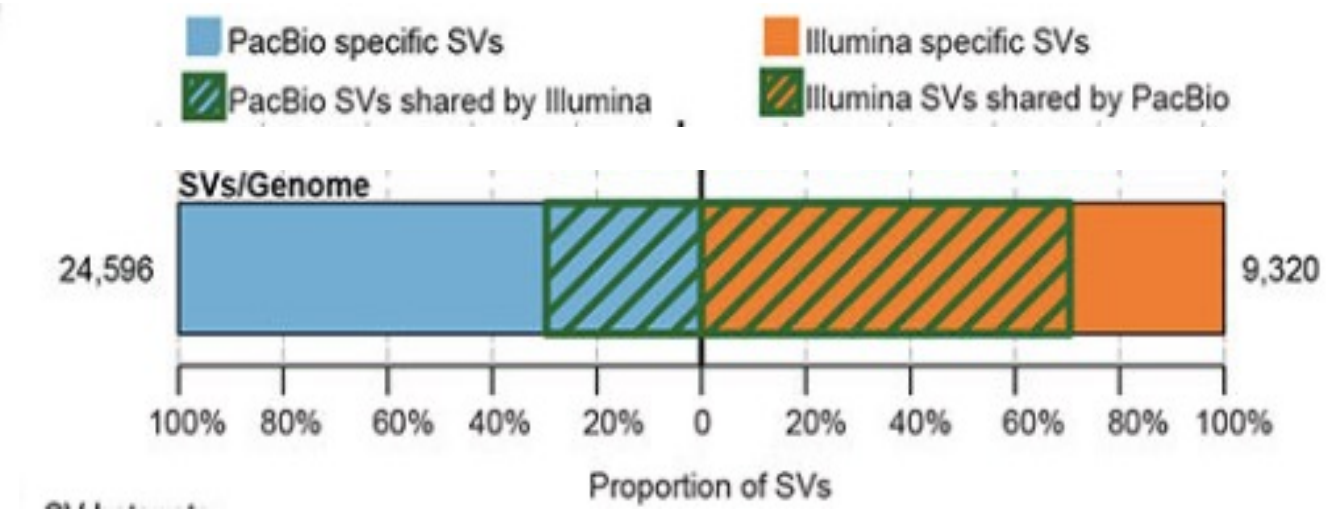


GENOME ASSEMBLY WITH CCS

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



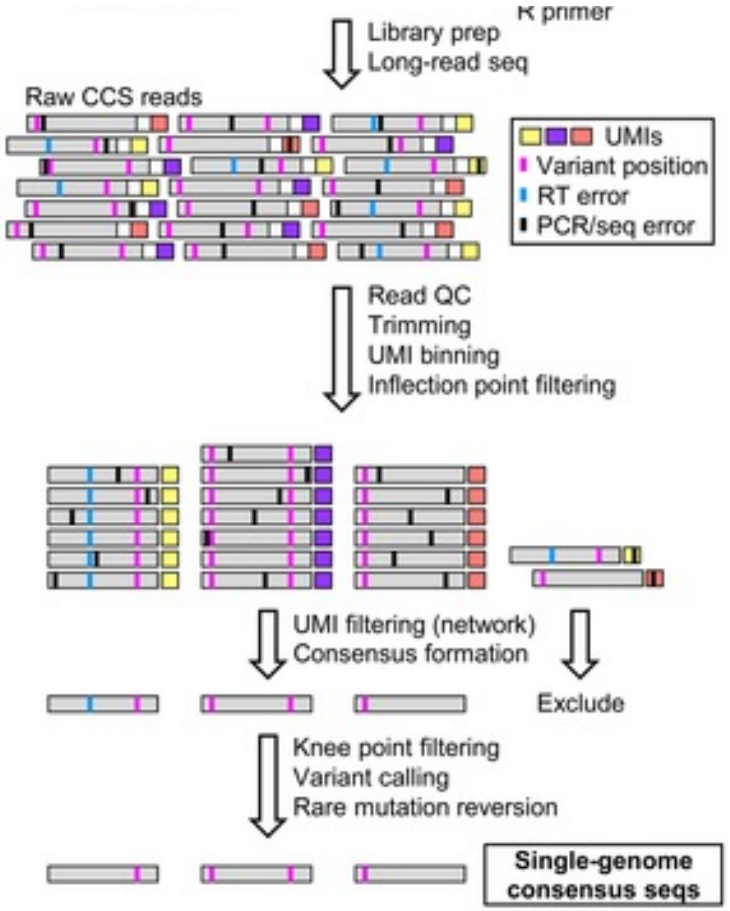
- Two important types of regions not fully resolved :
  - Gene-rich regions in segmental duplications
  - Larger repeat-rich regions such as centromeres
- " Recent advances coupling HiFi and ultra-long-read Nanopore may solve these more complex regions "

GENOME ASSEMBLY WITH CCS

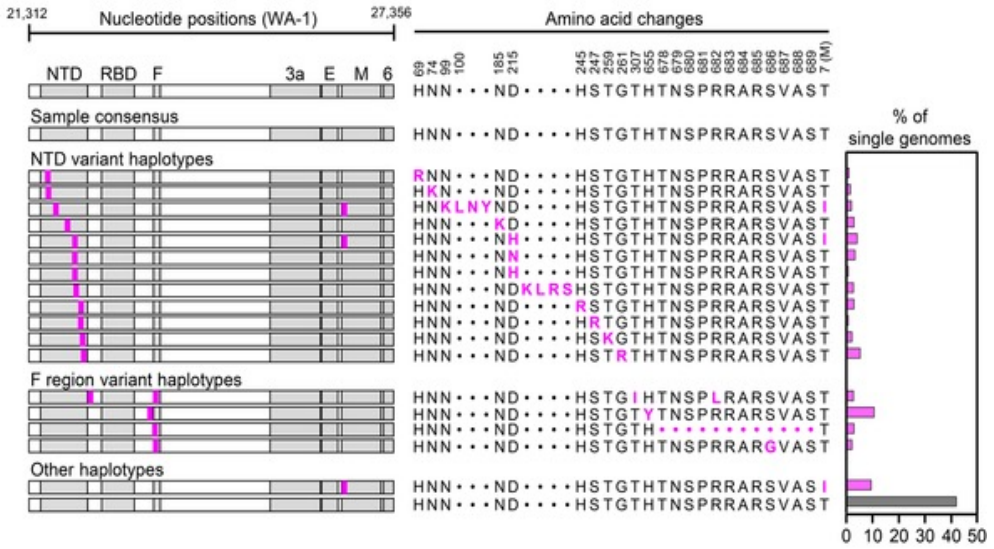
2 - High-throughput, single-copy sequencing reveals SARS-CoV-2 spike variants coincident with mounting humoral immunity during acute COVID-19, Ko S.H. et al. *PLOS Pathogens* 2021

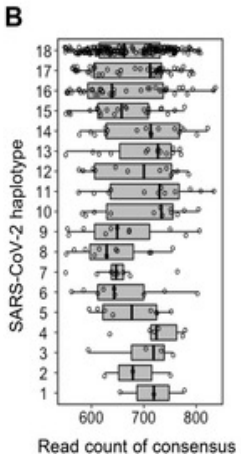
Study of intra-individual evolution of SARS-CoV-2 : for each sample -> multiple sequences representing virus diversity

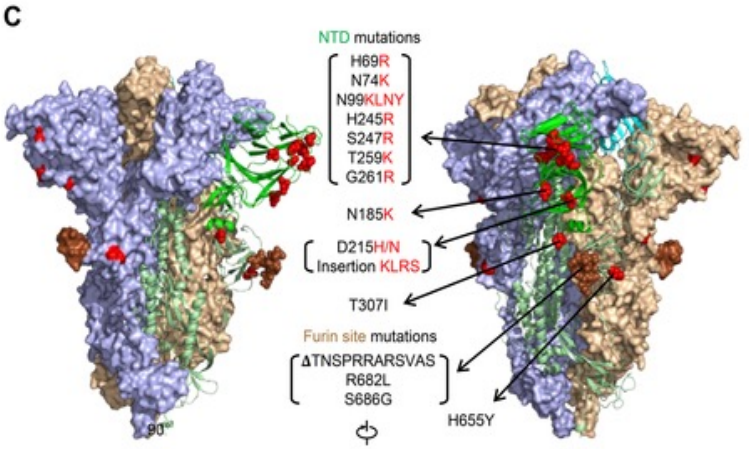
Analysis of CCS reads



Each sequence corresponds to a single viral genome

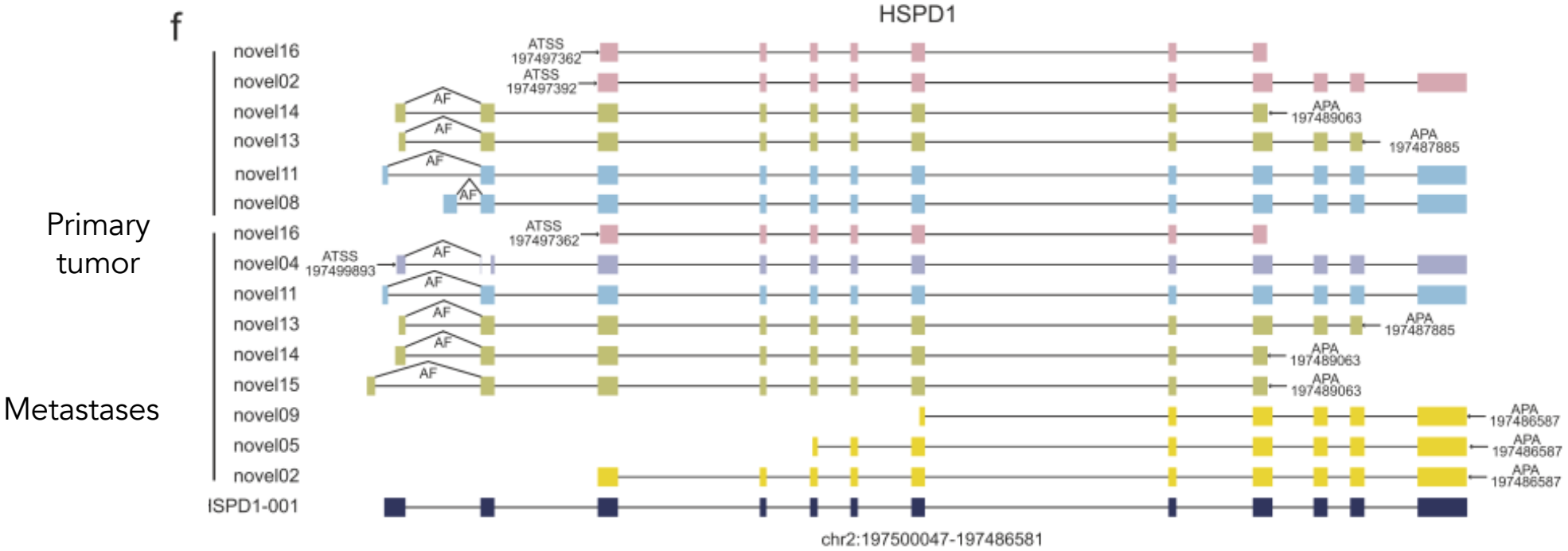






PacBio cDNA SEQUENCING WITH CCS

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

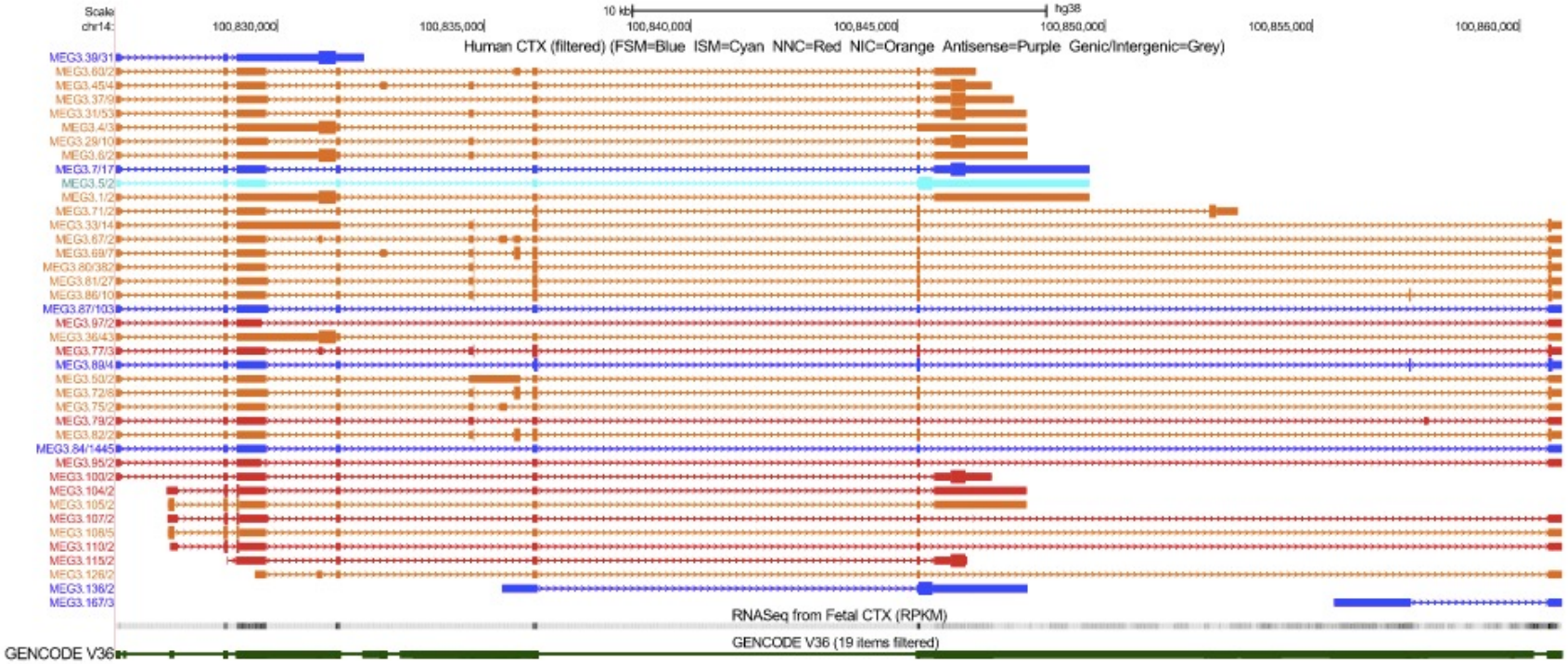


Long-read full-length transcriptome analysis improves molecular diagnostic

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

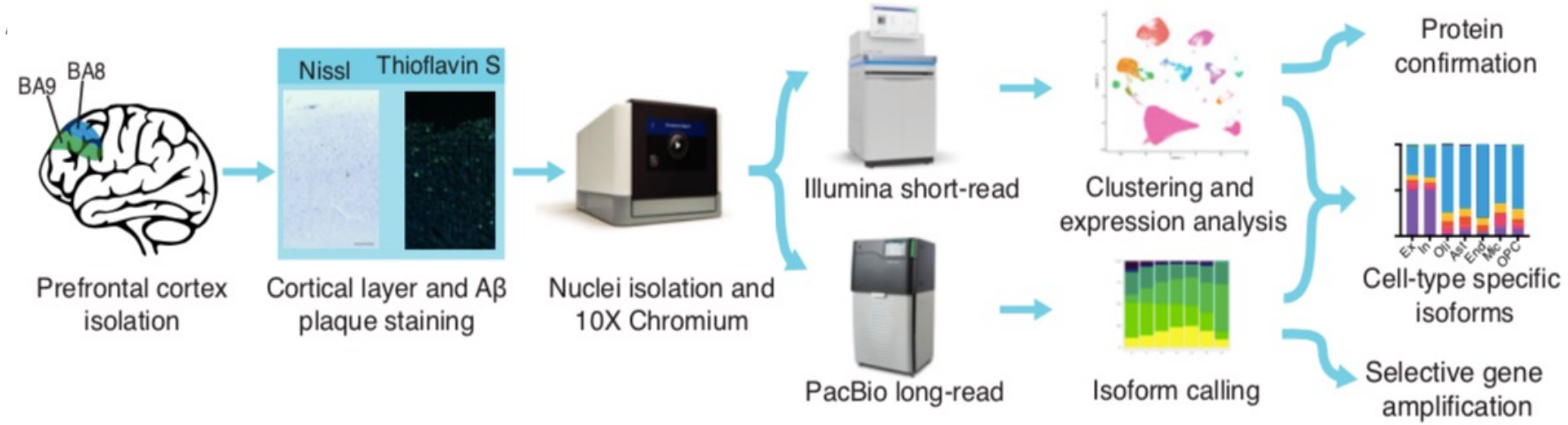
Transcripts mapping to disease-associated genes in human

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

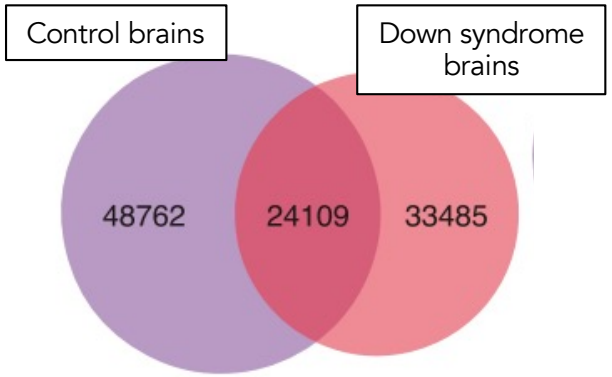
SINGLE CELL PacBio cDNA SEQUENCING

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

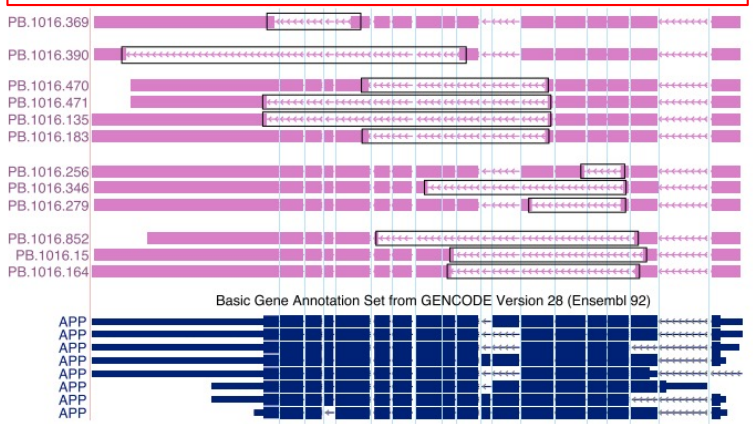
Down syndrome (trisomy 21) :
- single-nucleus long read RNA sequencing
- >170,000 cells from 29 aging DS and control brains



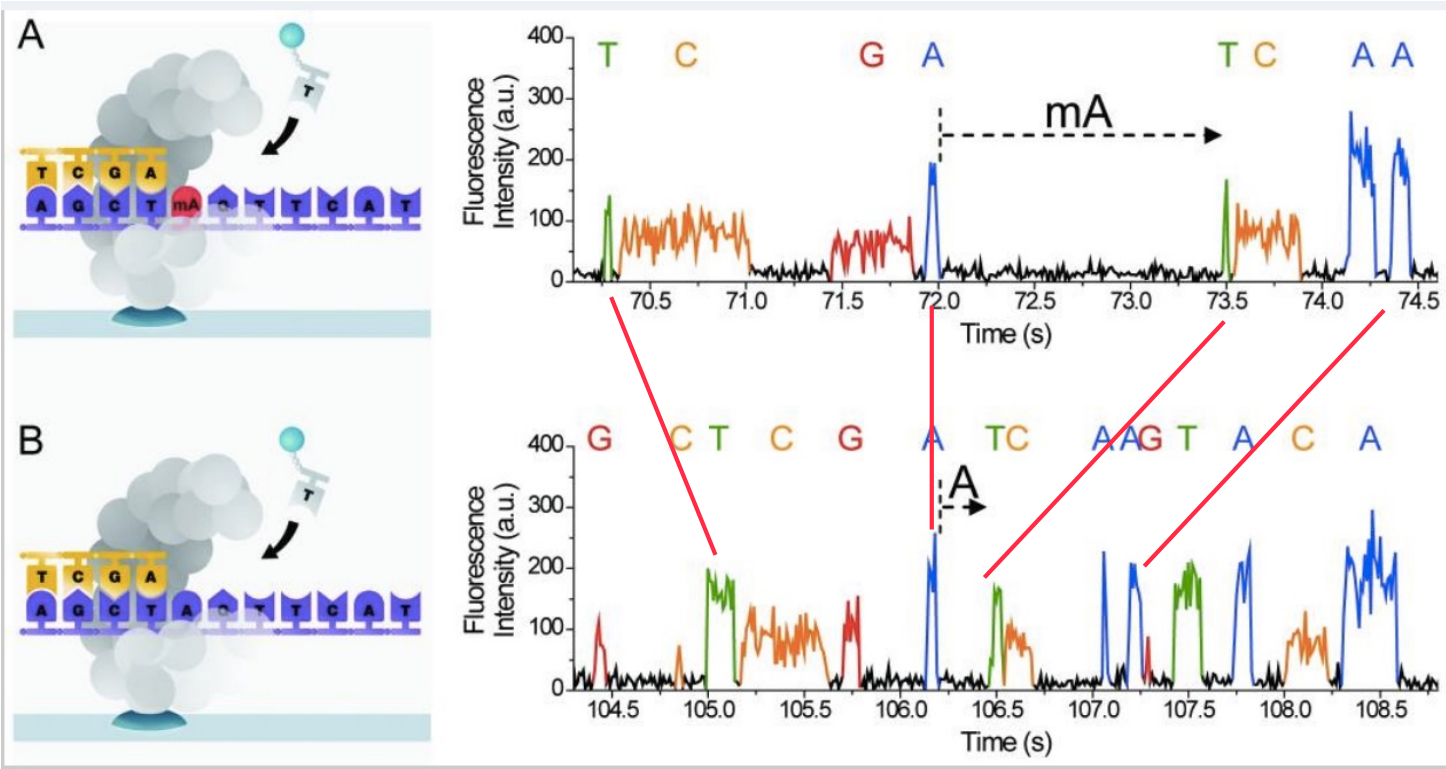
New splicing isoforms :
- new splice sites
- novel exon junctions
- entirely new exons
- intron retention



Amyloid precursor protein (Alzheimer's disease gene)



DETECTION OF MODIFIED DNA BASES



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

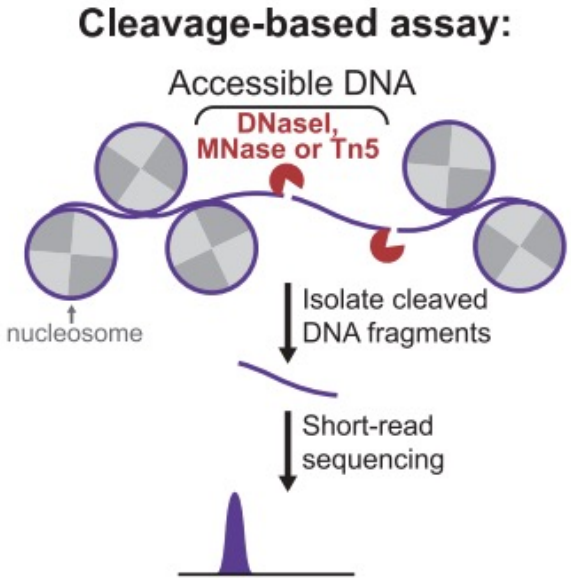
Detection of 5mA with strong influence of sequence contexts : requires high coverage

Feng et al. *PLOS Comput Biol* 2013

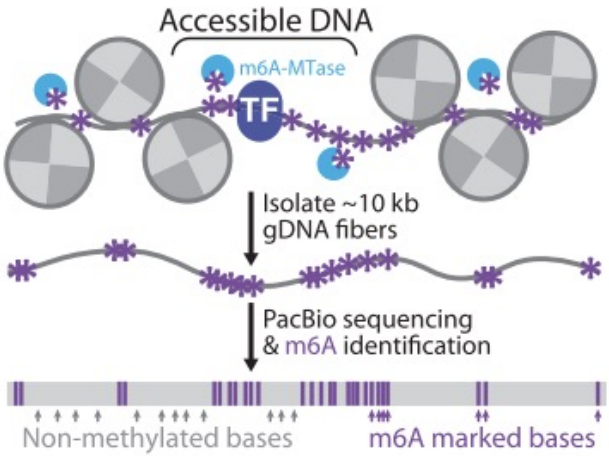
DETECTION OF DNA m6A WITH CCS

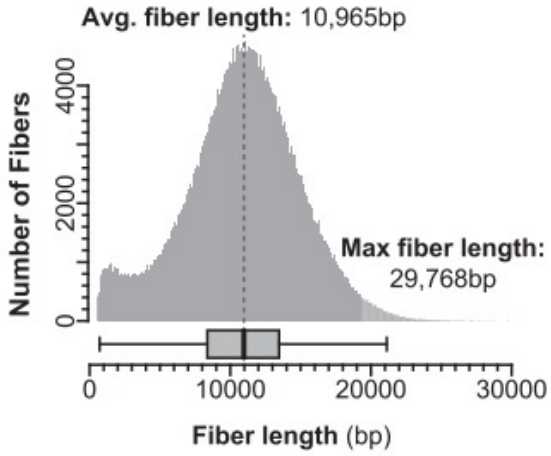
Single-molecule regulatory architectures captured by chromatin fiber sequencing
Stergachis et al. *Science* (2020)

DnaseI-seq.



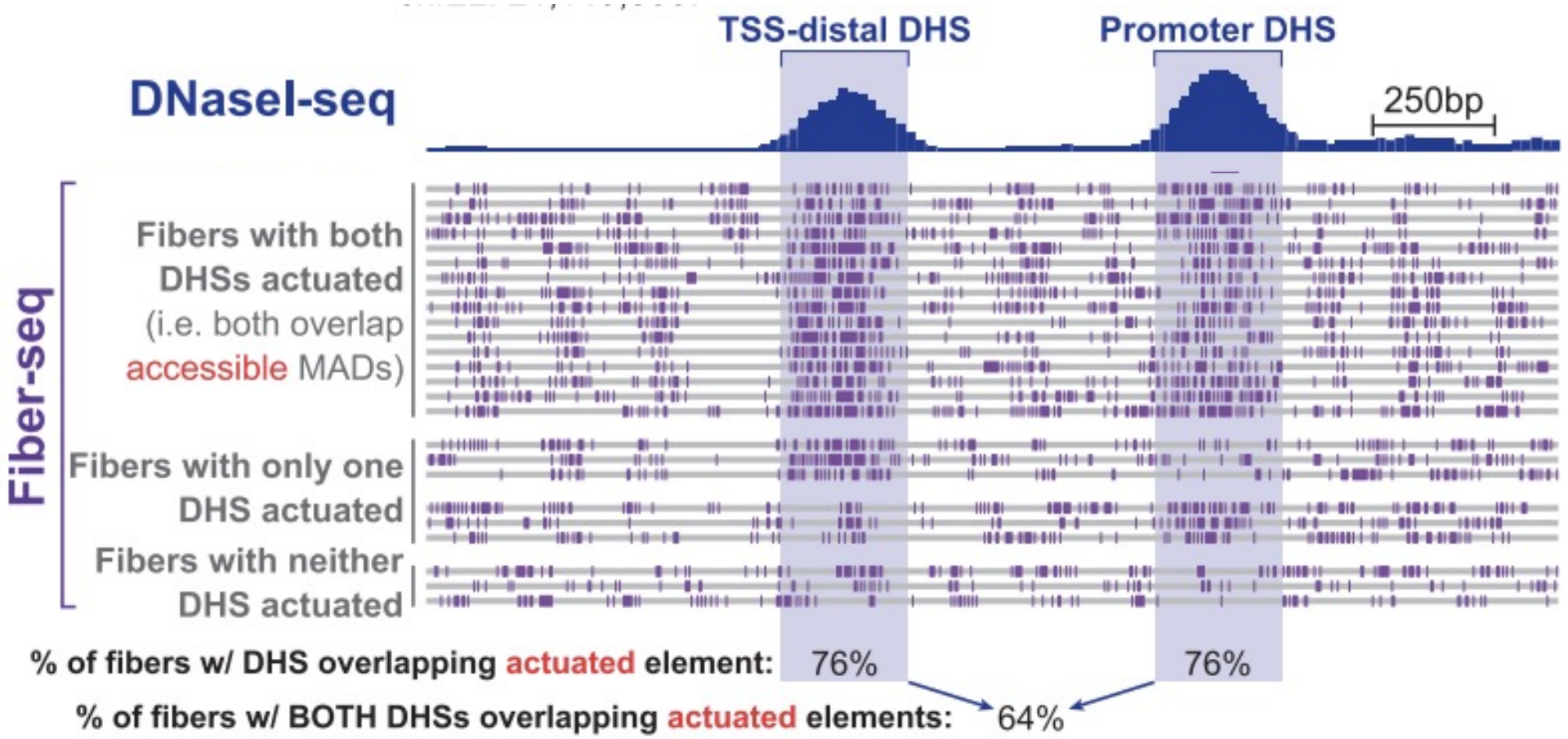
Fiber-seq.





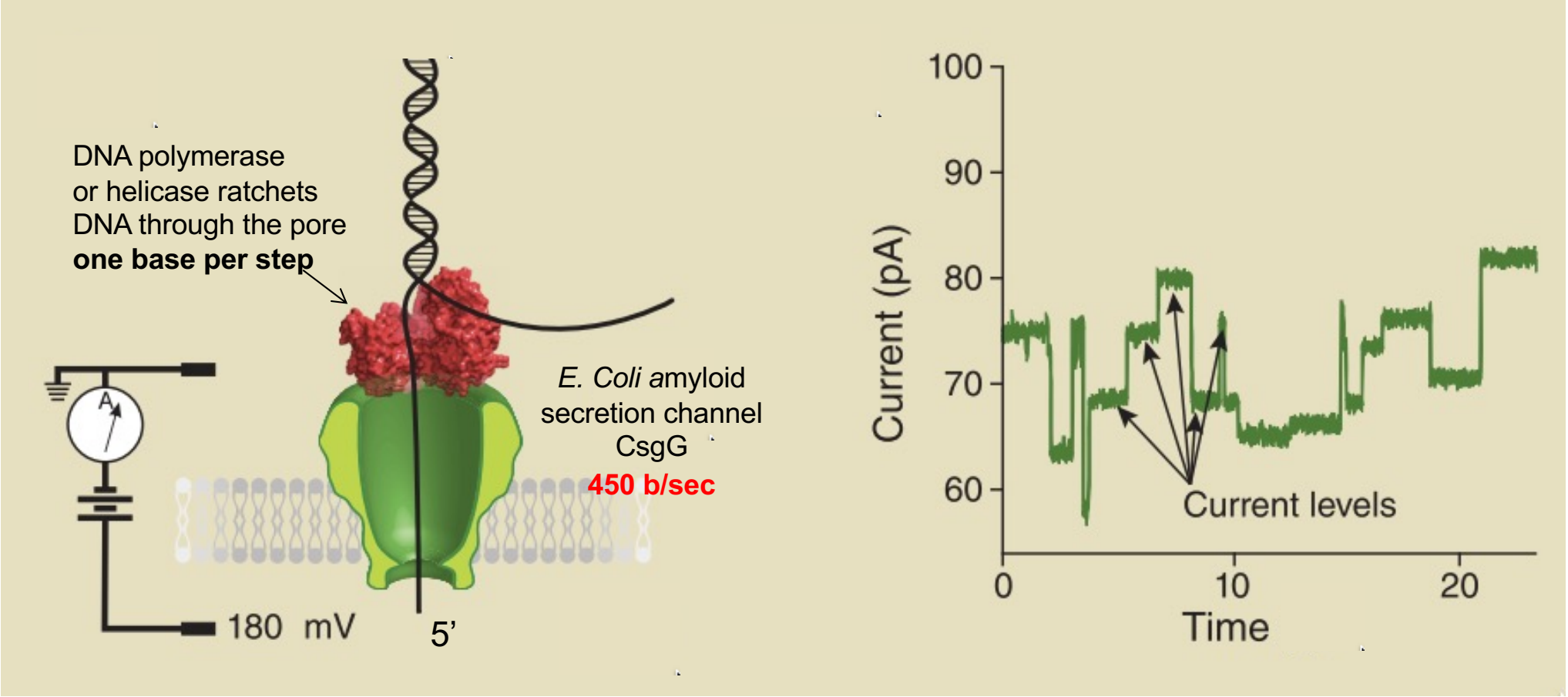
DETECTION OF DNA m6A WITH CCS

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



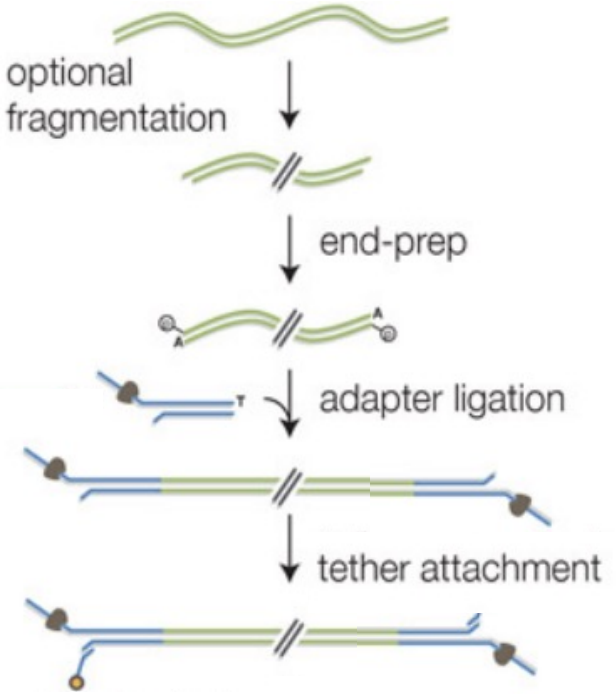
Next Generation Sequencing

Oxford NANOPORE Technologies®

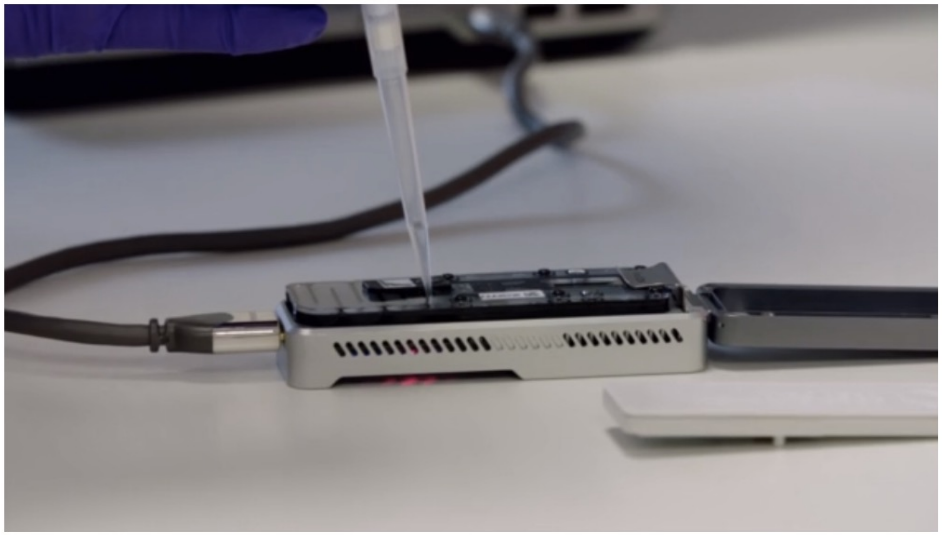

DNA polymerase
or helicase ratchets
DNA through the pore
one base per step
E. Coli amyloid
secretion channel
CsgG
450 b/sec
A
180 mV
5'
Current (pA)
100
90
80
70
60
Current levels
0
10
20
Time

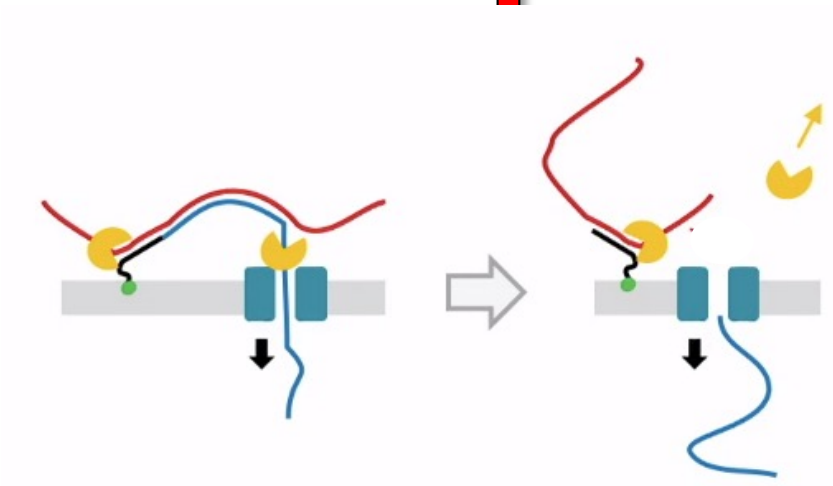
SEQUENCING PROCESS

Library preparation

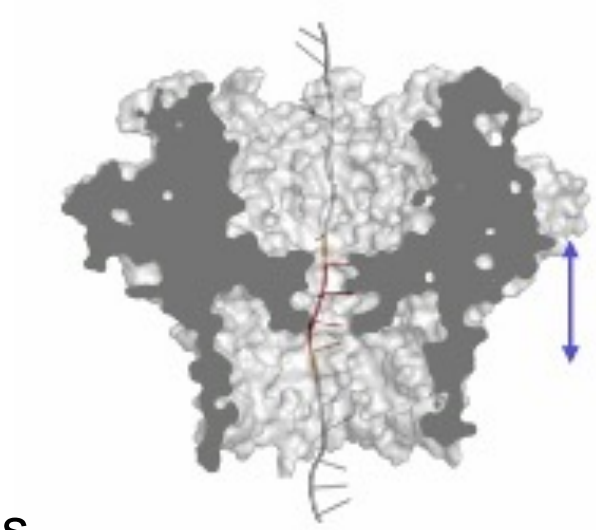


SEQUENCING



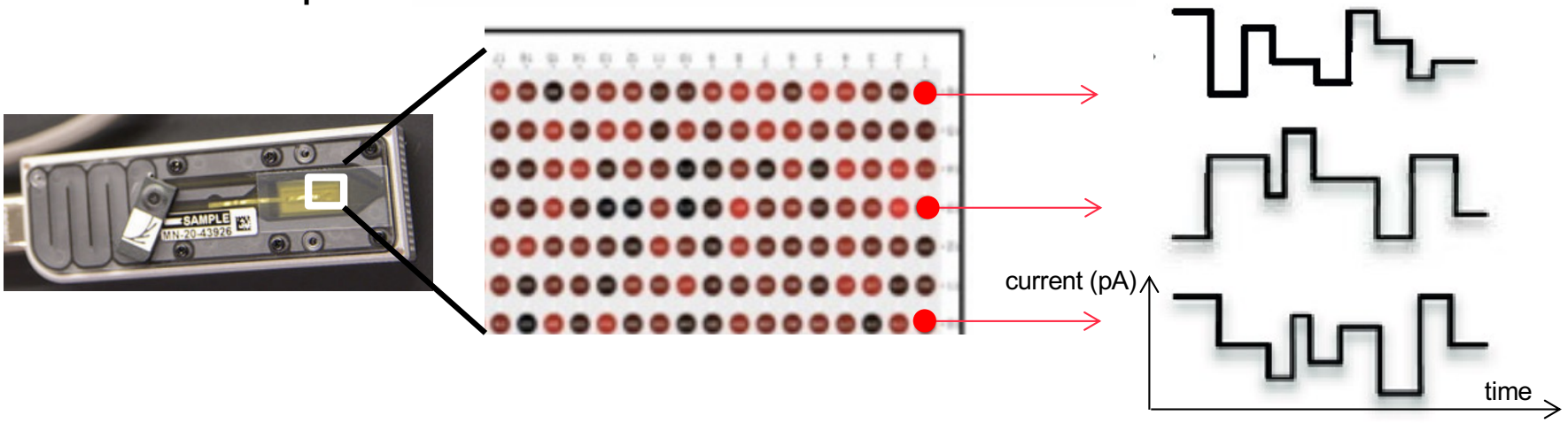


SEQUENCING PROCESS : MinION FLOW CELL



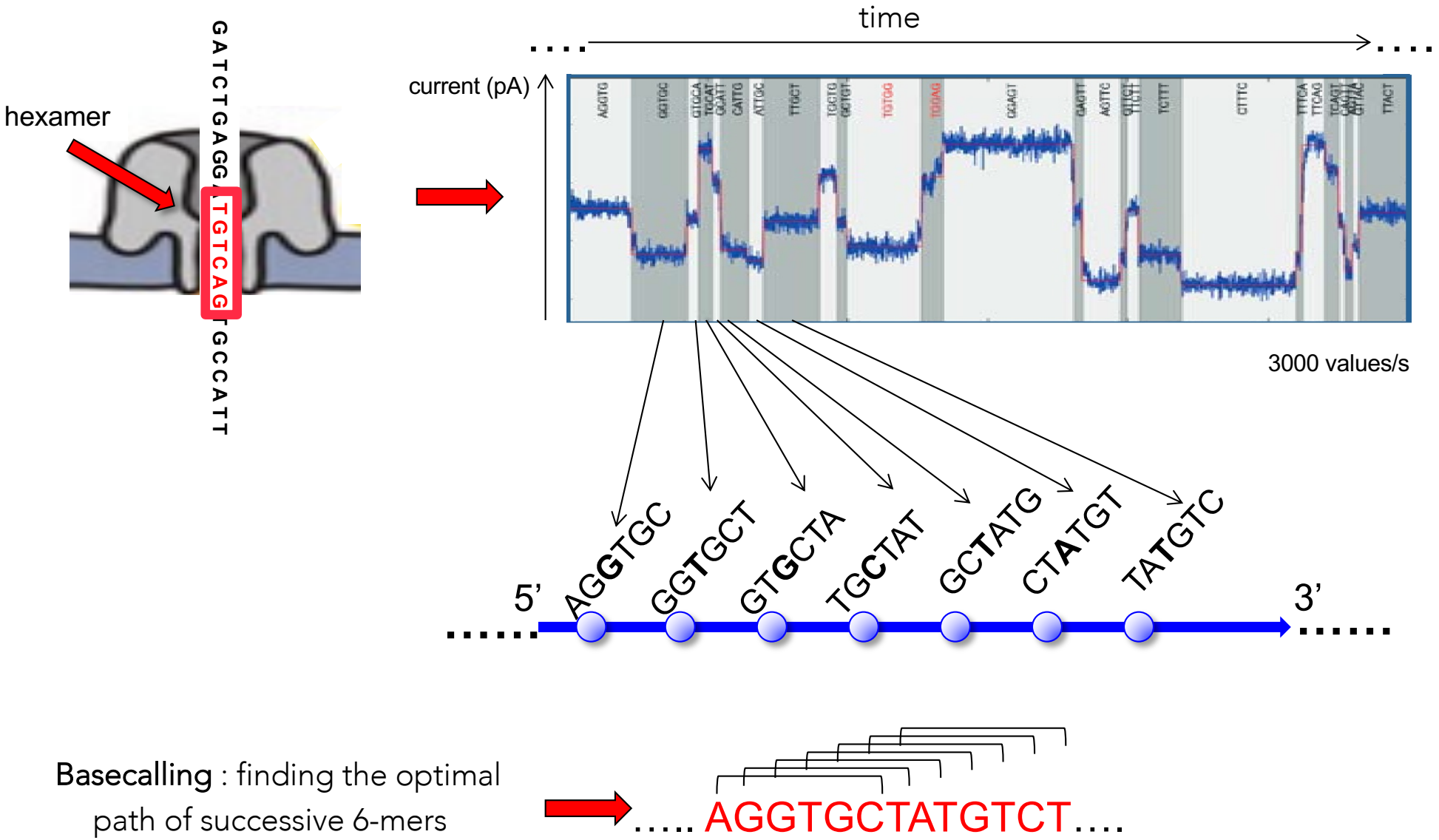
5-6 bases
dominate the
current signal

MinION : 512 pores



PromethION : 144000 pores (48 x 3000)

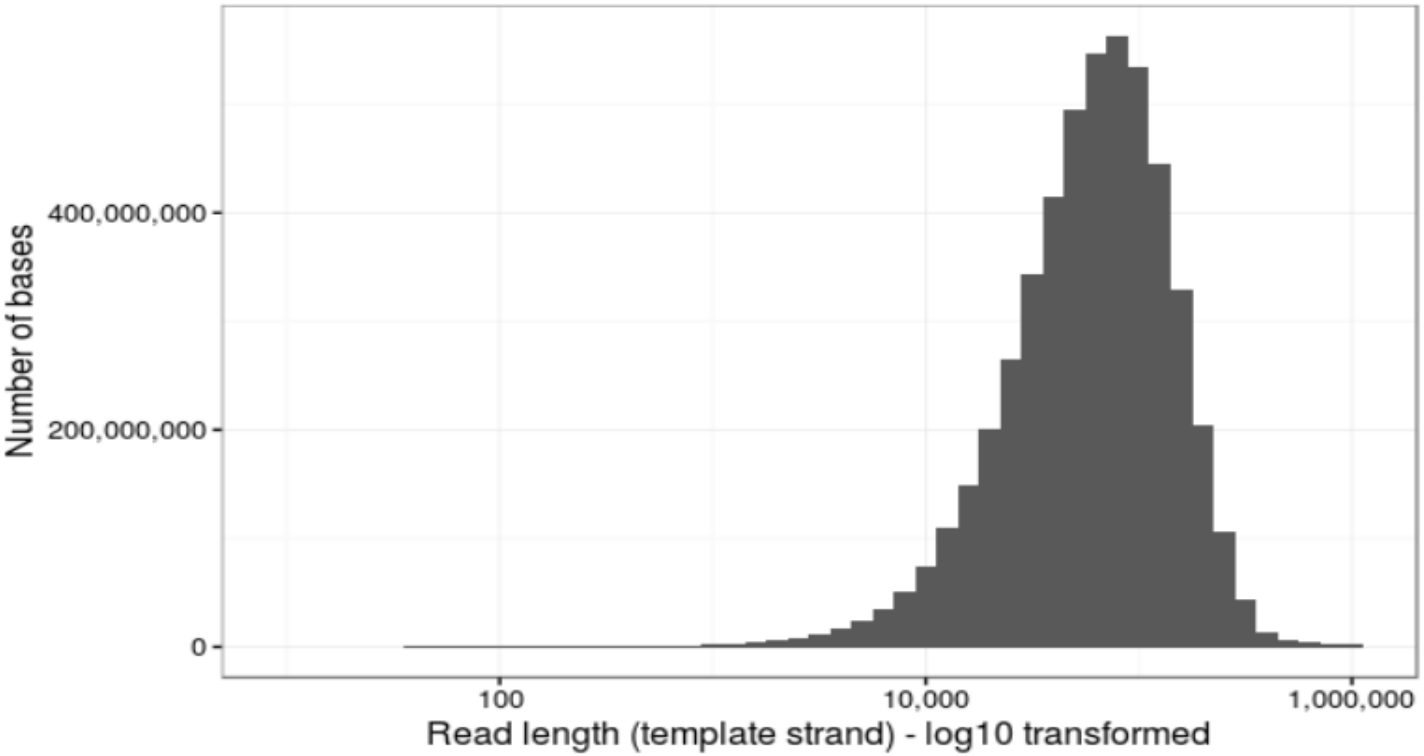
BASE CALLING



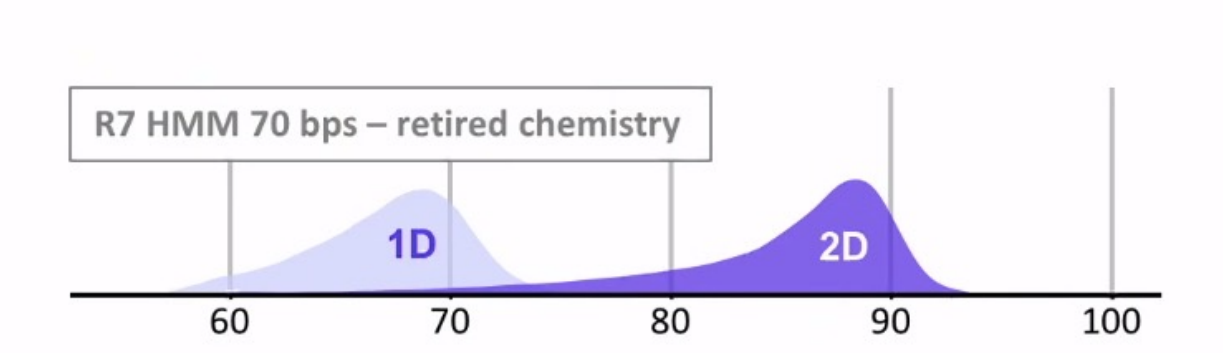
Basecalling : finding the optimal path of successive 6-mers

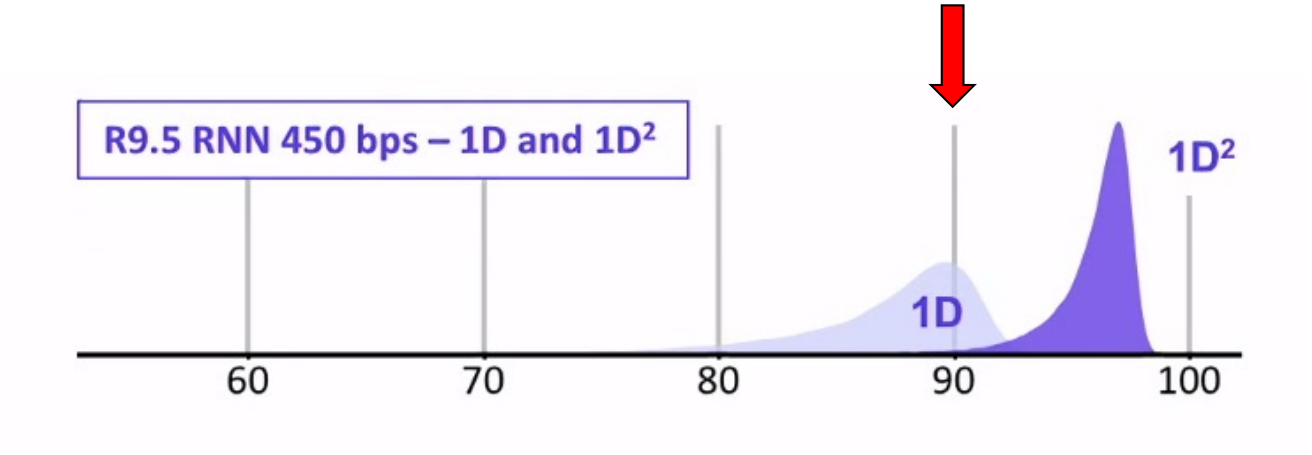
SIZE OF SEQUENCED DNA FRAGMENTS

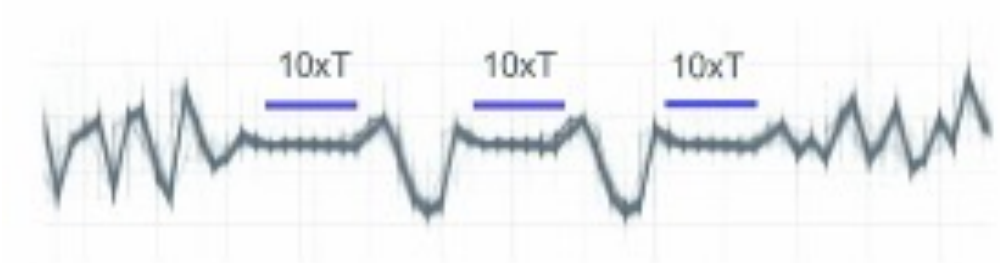
"Ultra long" reads
(lab.loman.net, March 2017)



Size of the longest read > 1 Mb



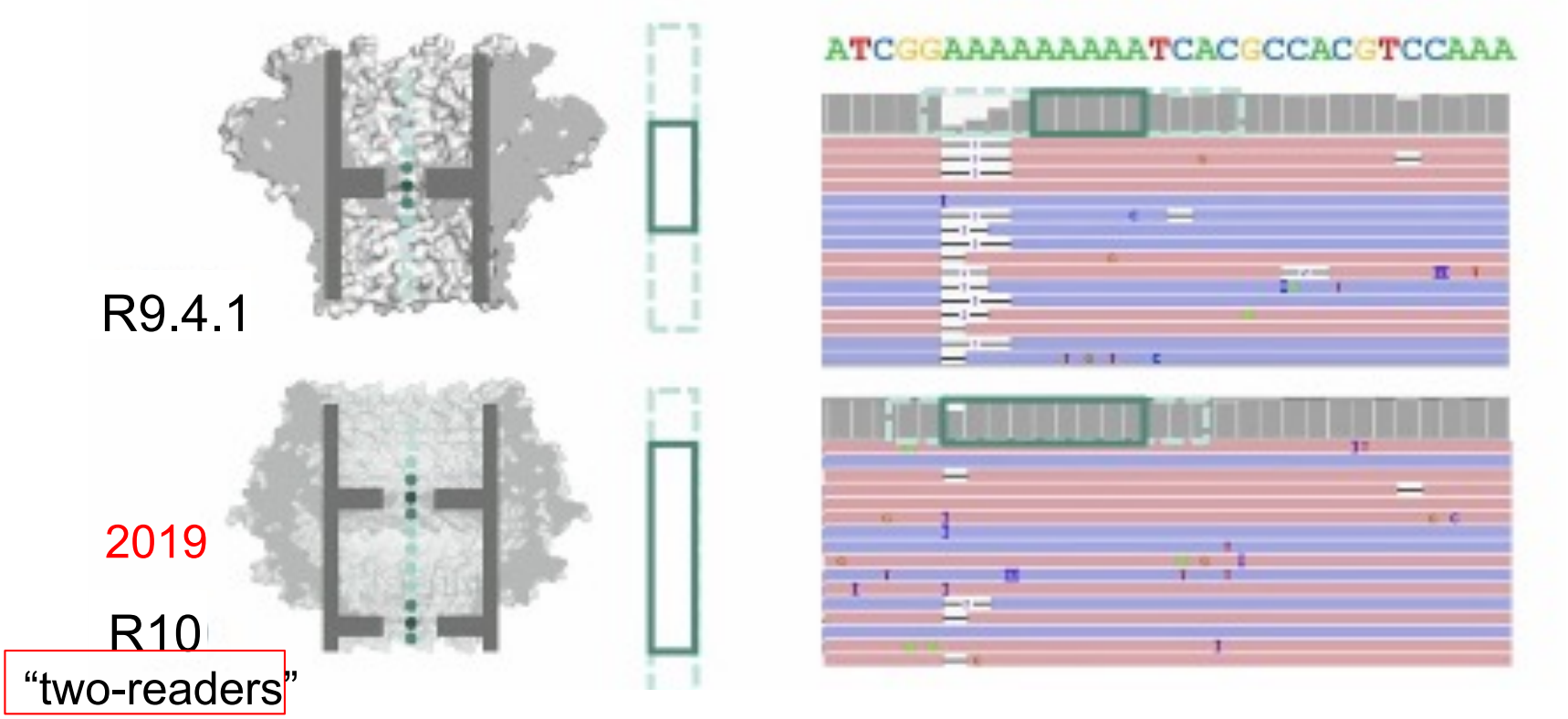




Homopolymers difficult to sequence

Recent improvements: "Two readers" nanopore

"One-reader" pore has difficulty to read homopolymers

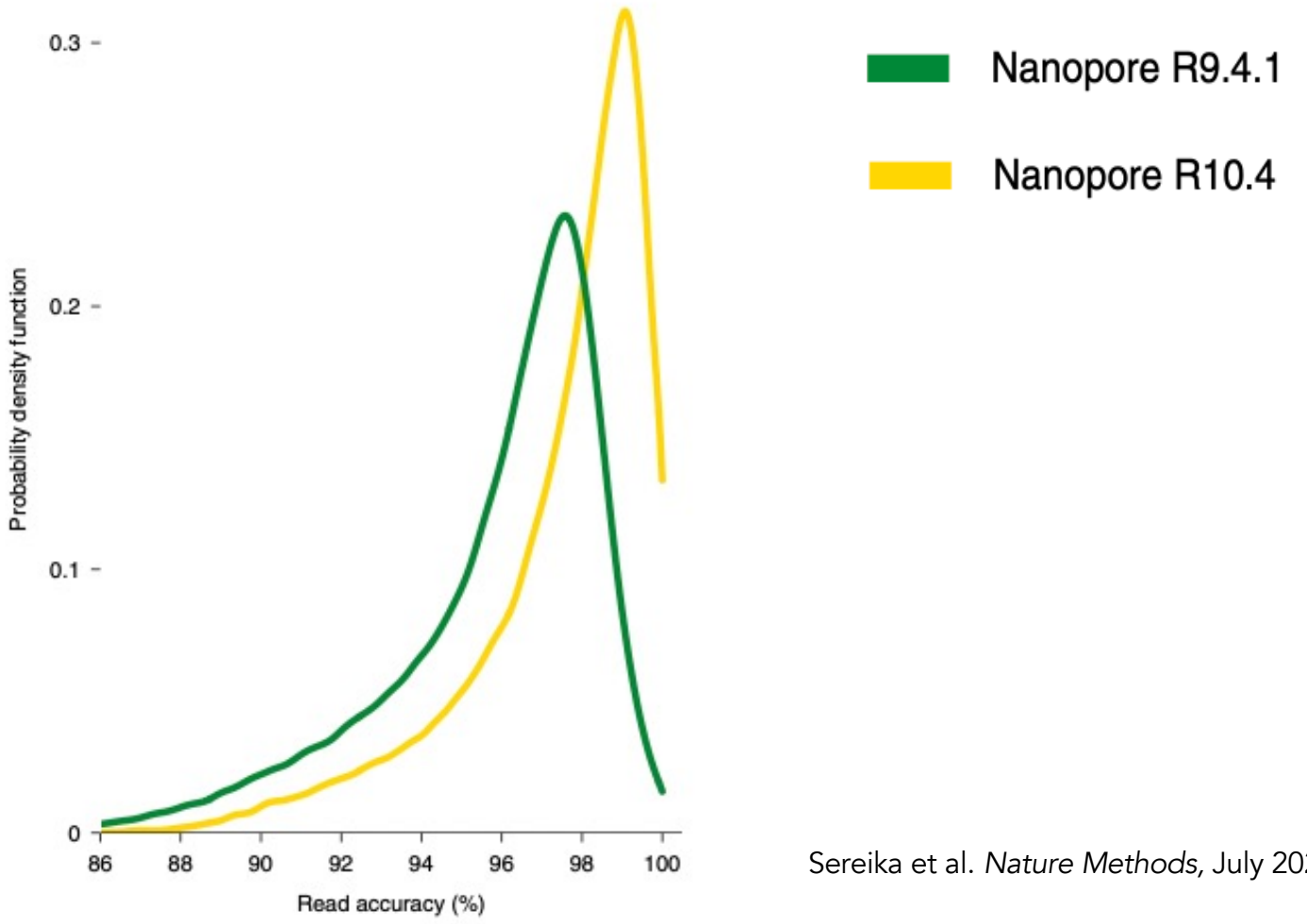


New pore accurately calls homopolymers

- 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

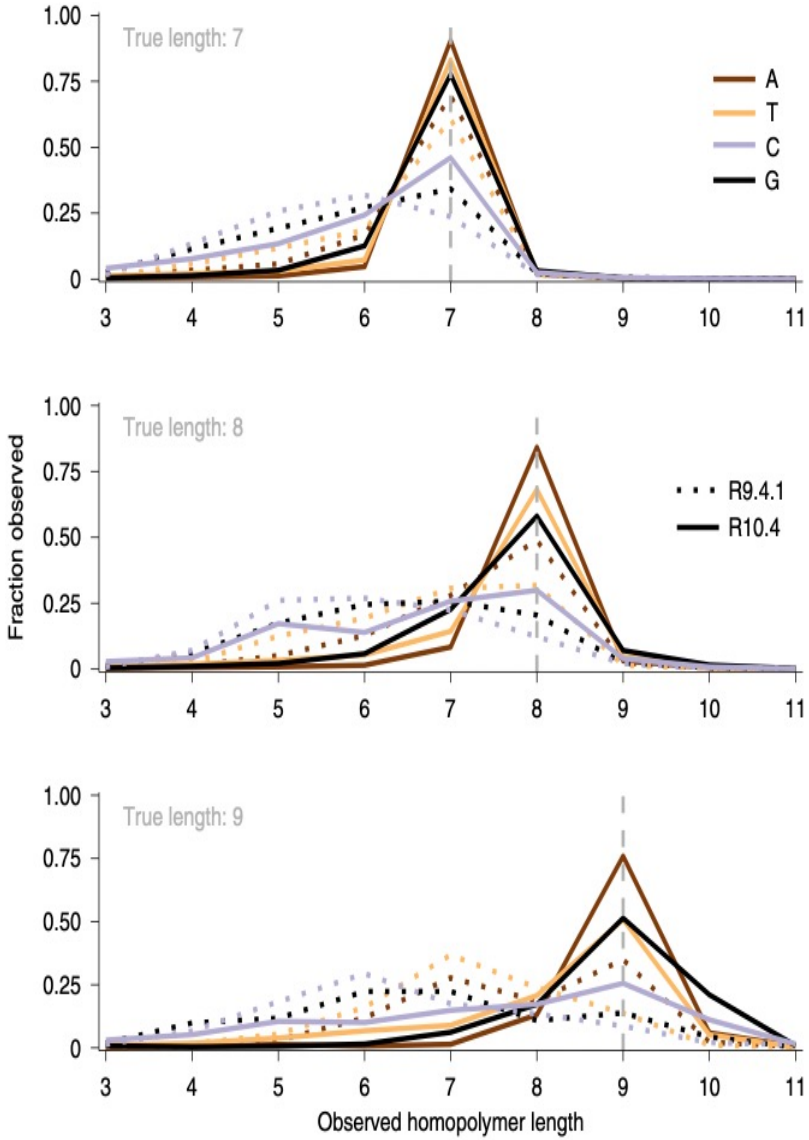
Recent improvements: "Two readers" nanopore

Read accuracies measured through read-mapping



Sereika et al. *Nature Methods*, July 2022

Recent improvements: "Two readers" nanopore

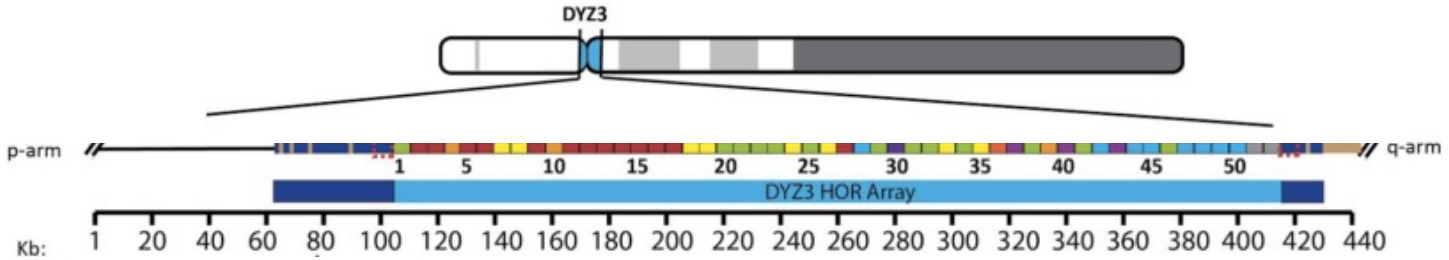


➢ R10.4 : improved ability to call homopolymers

Sereika et al. *Nature Methods*, July 2022

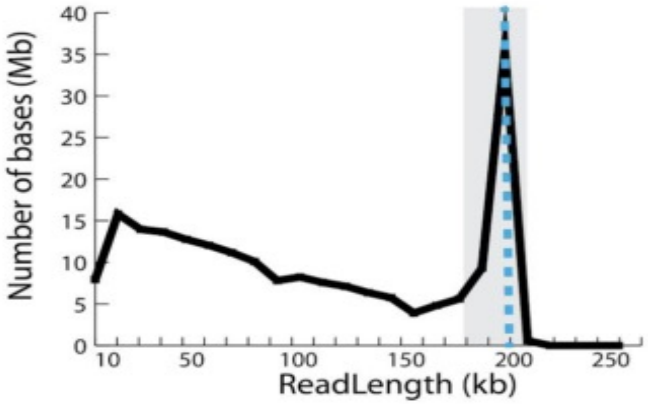
GENOME ASSEMBLY WITH NANOPORE

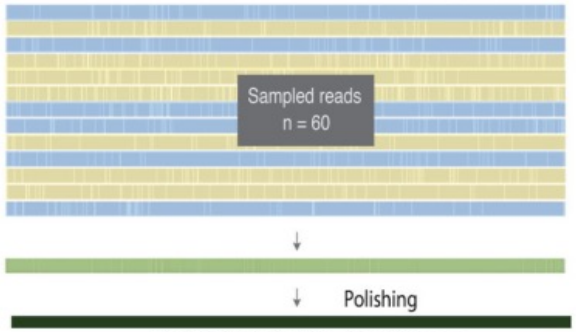
1 - Linear Assembly of a Human Y Centromere using Nanopore Long Reads
Jain et al., *bioRxiv*, 2017



9 BACs
100 kb to 210 kb

210 kb





Final high quality consensus BAC sequence

FIRST COMPLETE SEQUENCE OF A HUMAN CENTROMERE

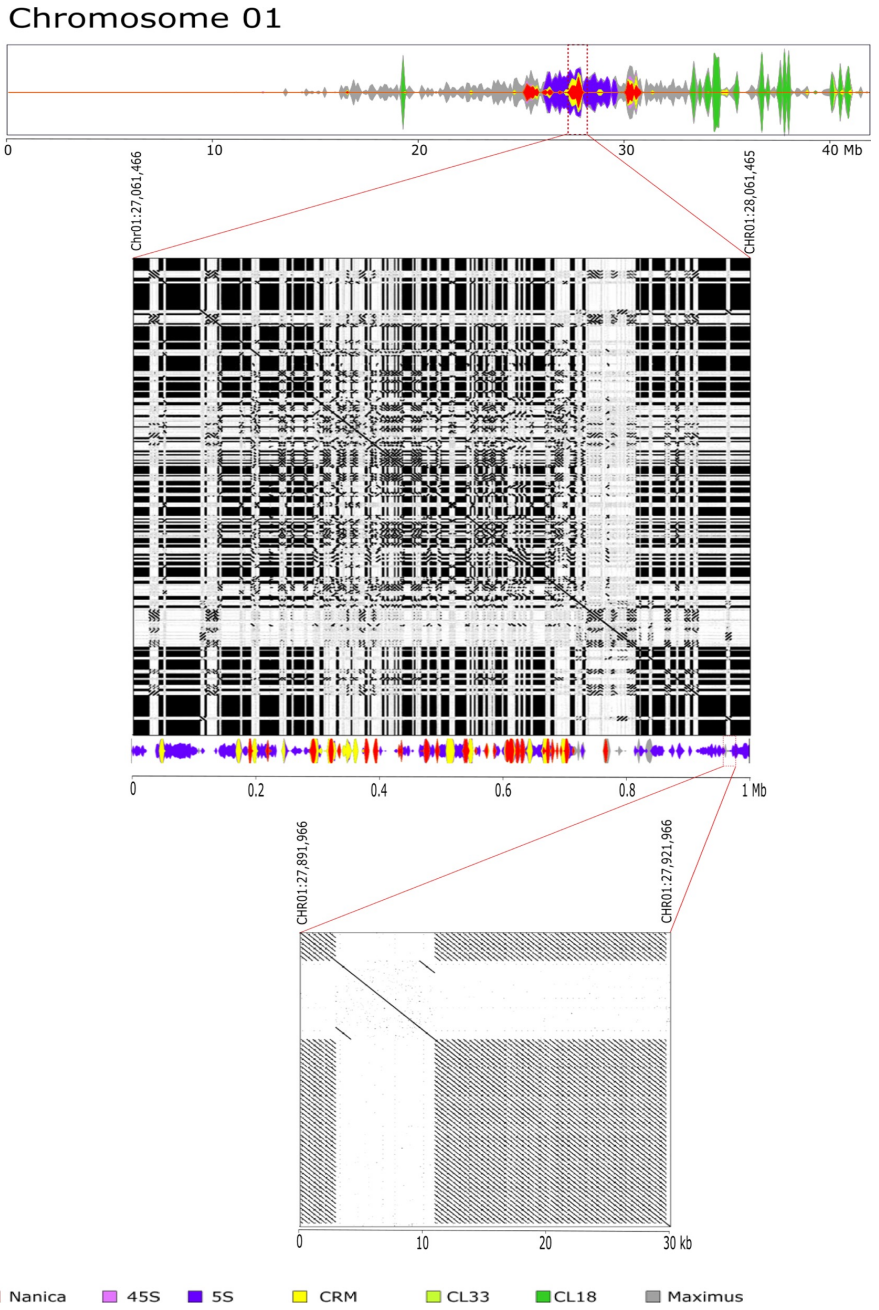
GENOME ASSEMBLY WITH NANOPORE

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



GENOME ASSEMBLY WITH NANOPORE

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

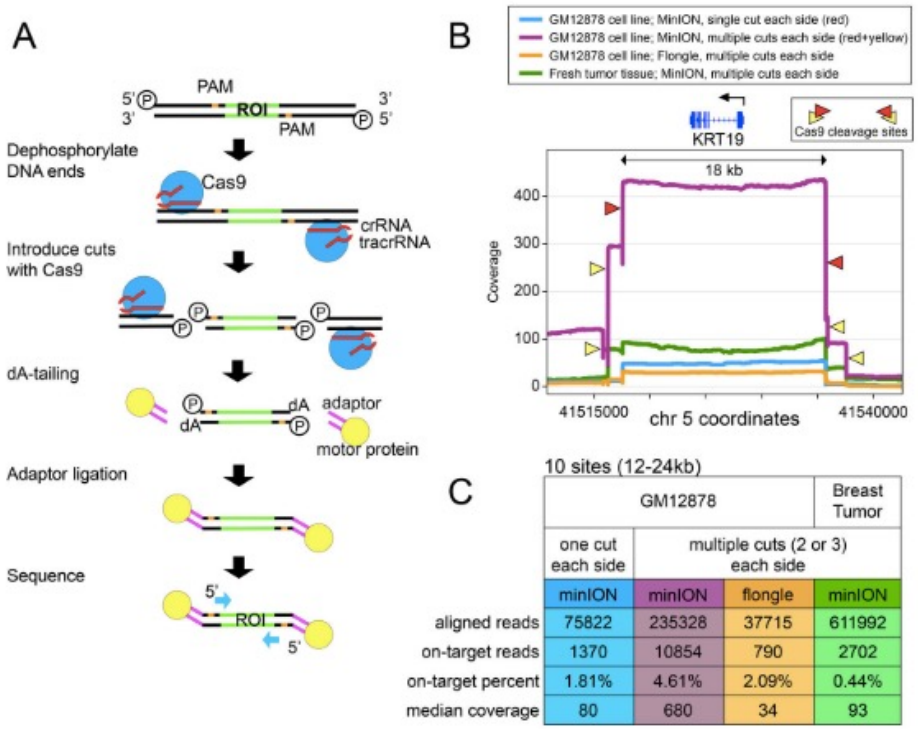
- First hexaploid wheat genome based on ONT long-reads
- hexaploid genome (15.5 Gb)
- sequencing began in 2005 : International Wheat Genome Sequencing Consortium (IWGSC)
- first sequence in 2018

- 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



C

10 sites (12-24kb)

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

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

nCATS can simultaneously assess :

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

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

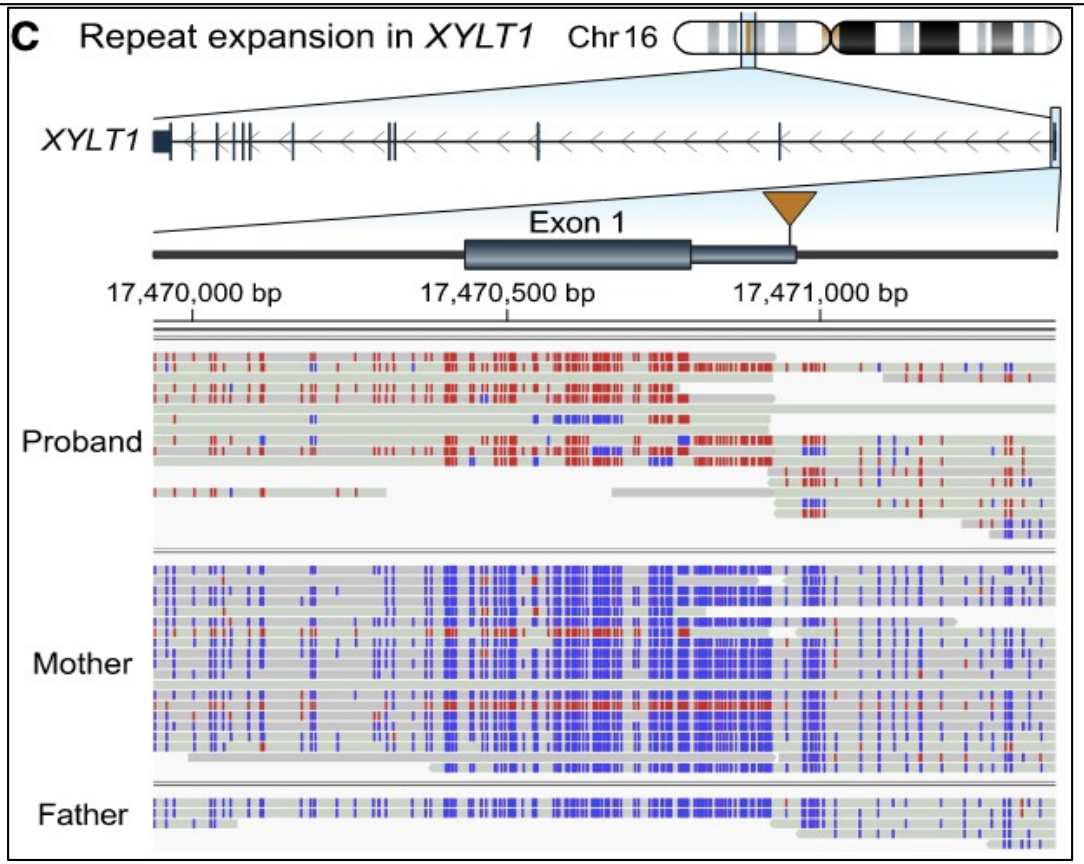
GENOME SEQUENCING : TARGETED NANOPORE SEQUENCING

2 - Targeted long-read sequencing identifiesmissing 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 :

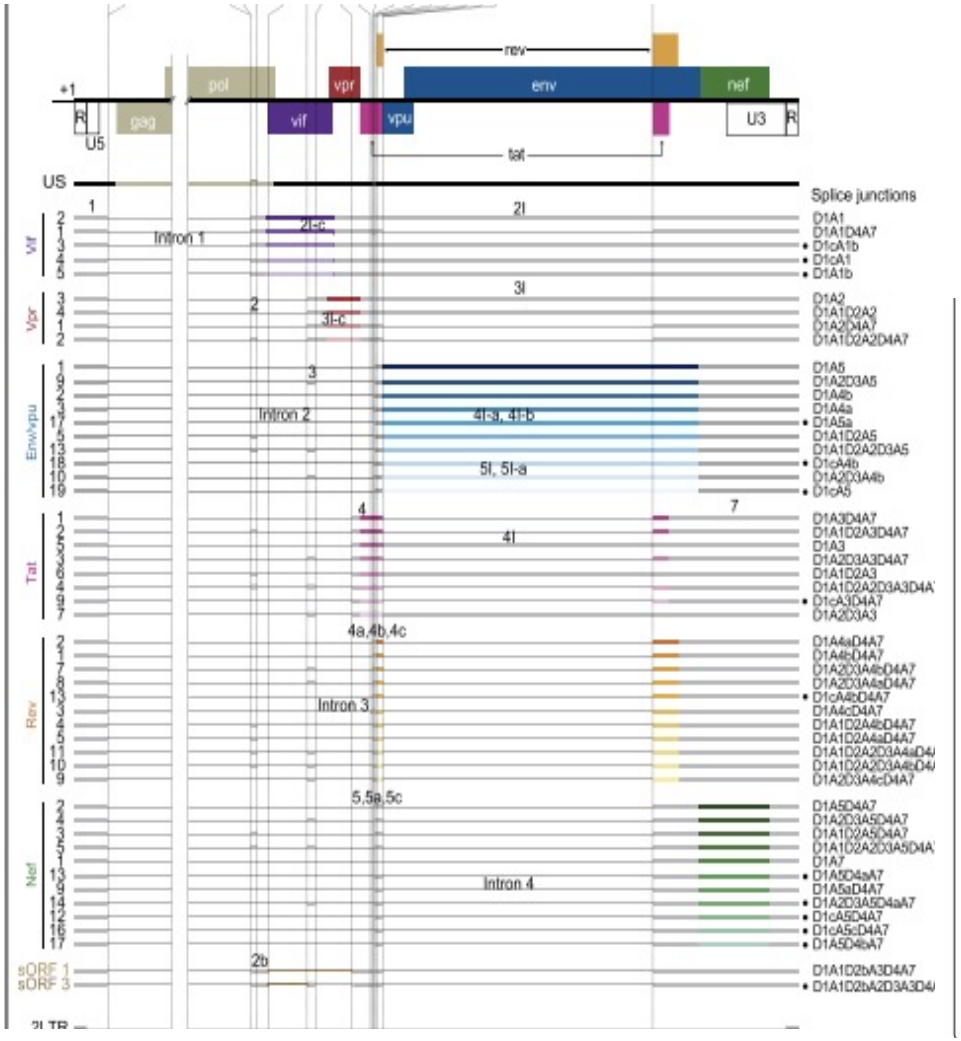
- T-LRS of native DNA molecules provides additional information not available when repeat length and methylation are assayed separately.

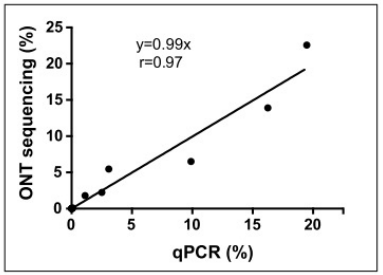


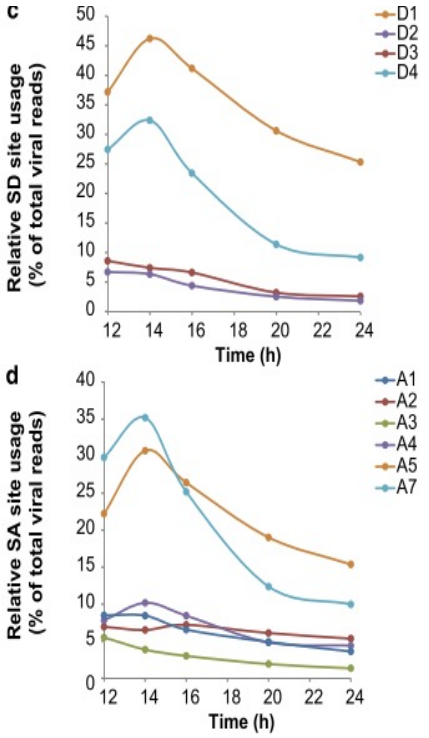
cDNA NANOPORE SEQUENCING

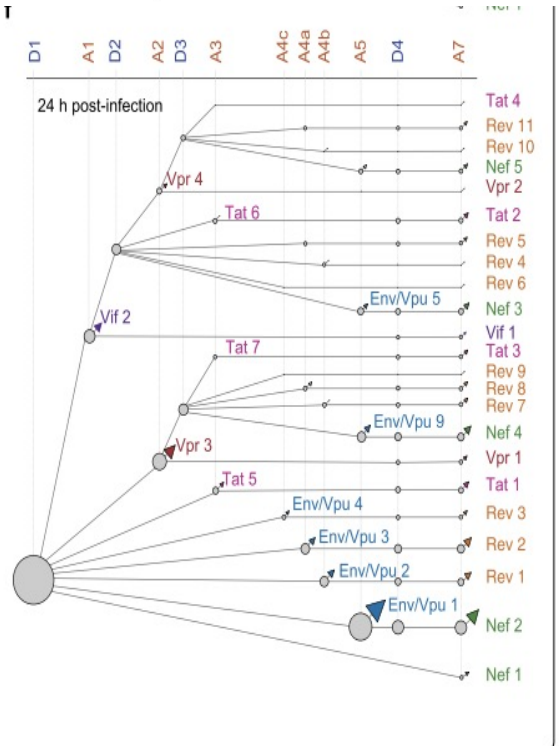
Dynamic nanopore long-read sequencing analysis of HIV-1 splicing events during the early steps of infection
Quang et al. *Retrovorology* 2020

- 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







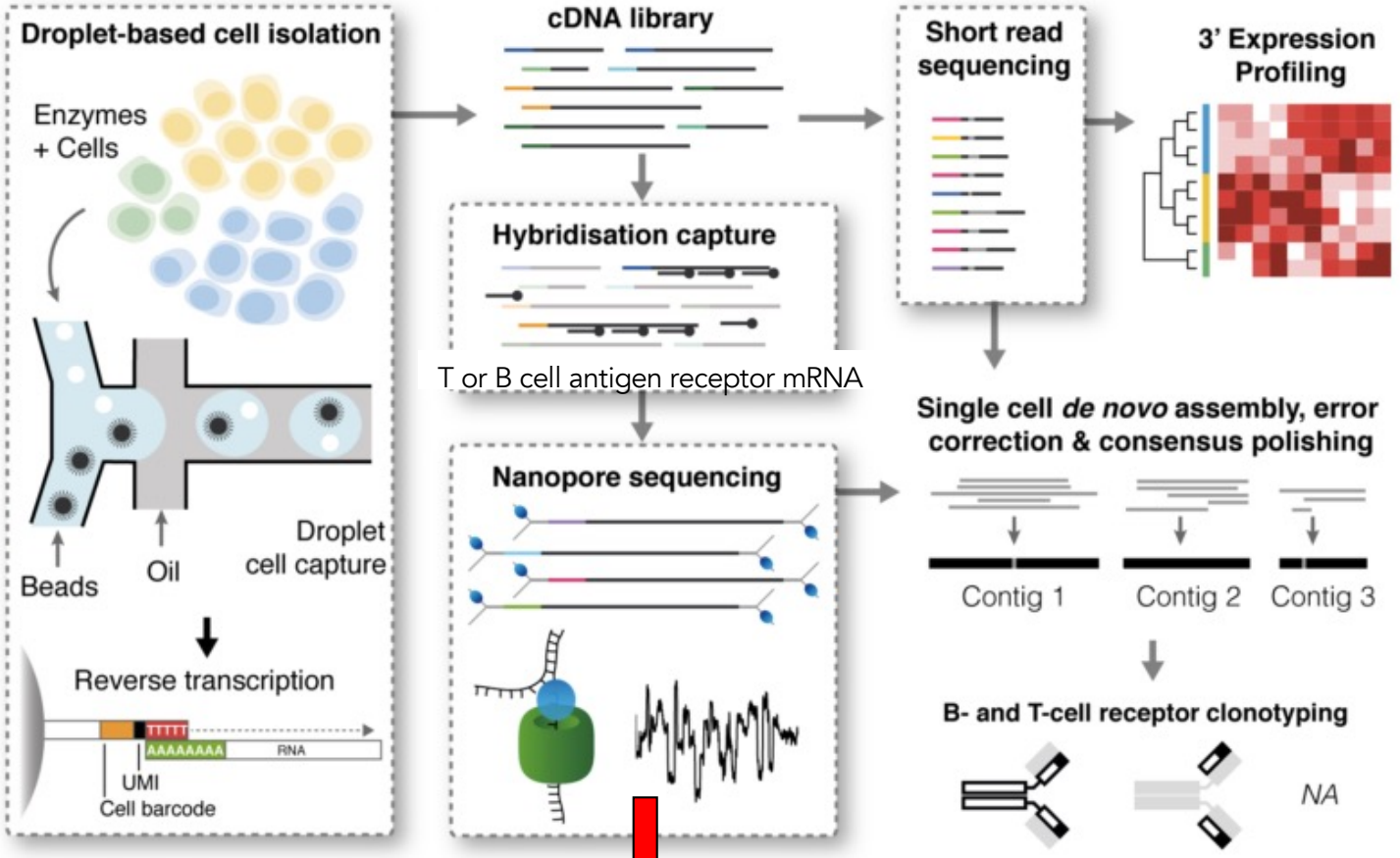


NANOPORE and SINGLE CELL cDNA SEQUENCING

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

10x Chromium Single Cell 3' (V2 chemistry)



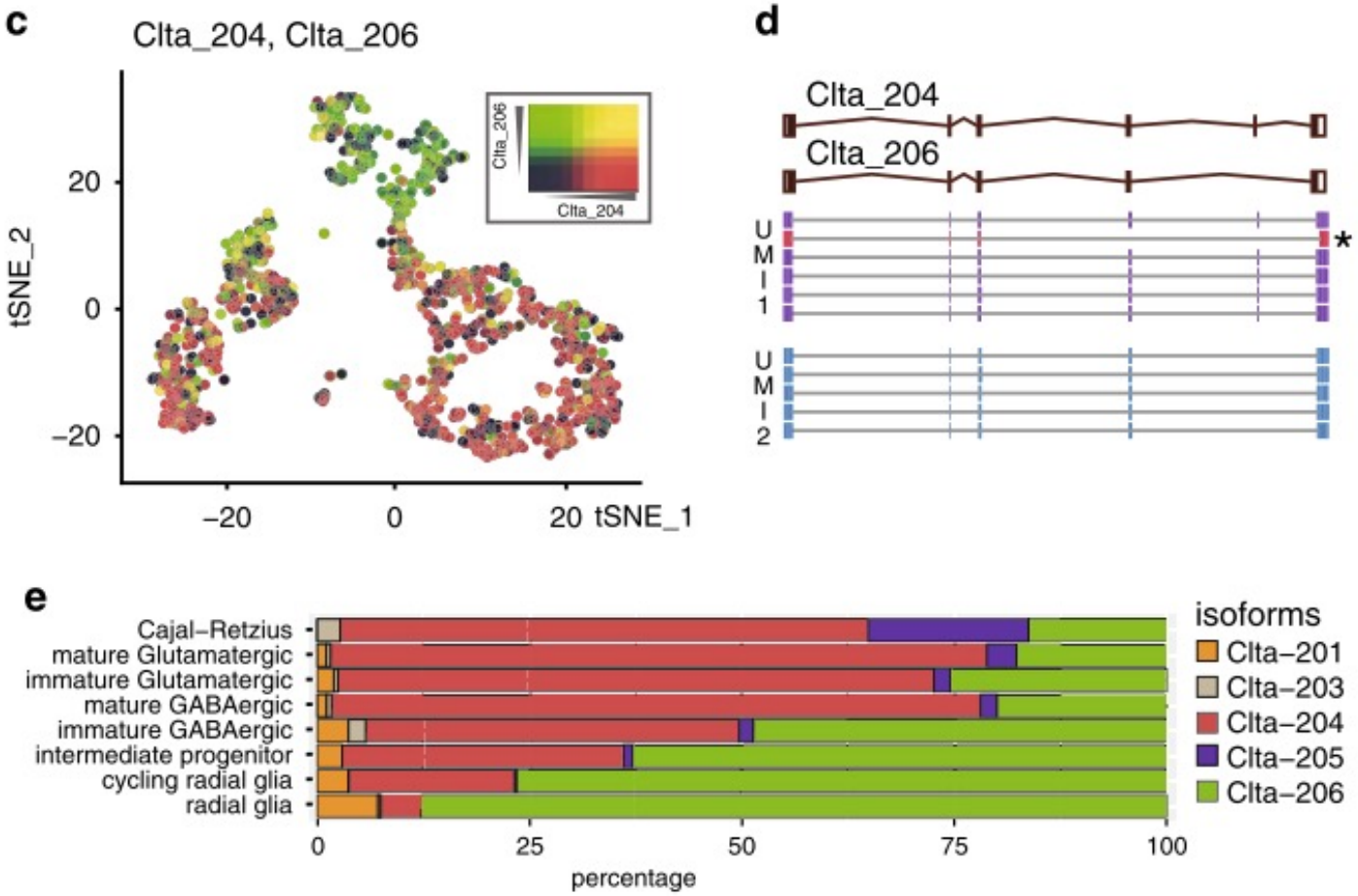
Tracking of somatic mutation, alternate splicing and clonal evolution of T and B lymphocytes
BUT
Does not correct for PCR biases

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

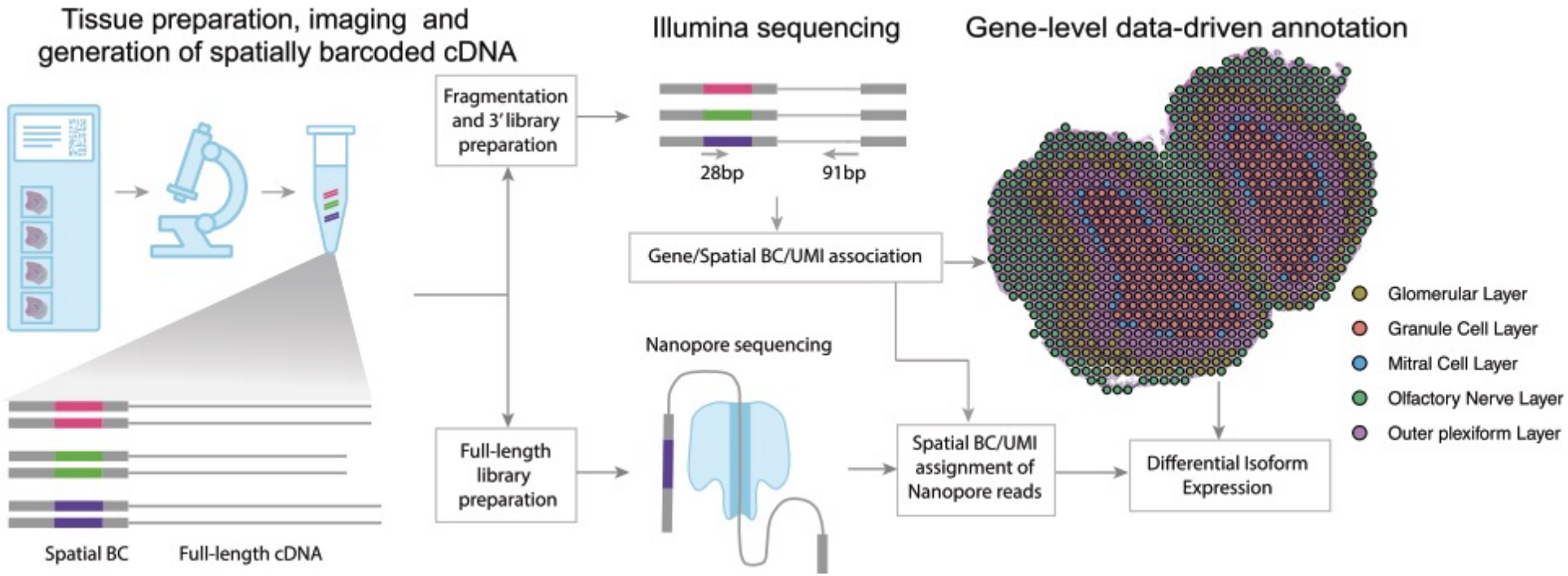


NANOPORE and 10x Genomics Visium

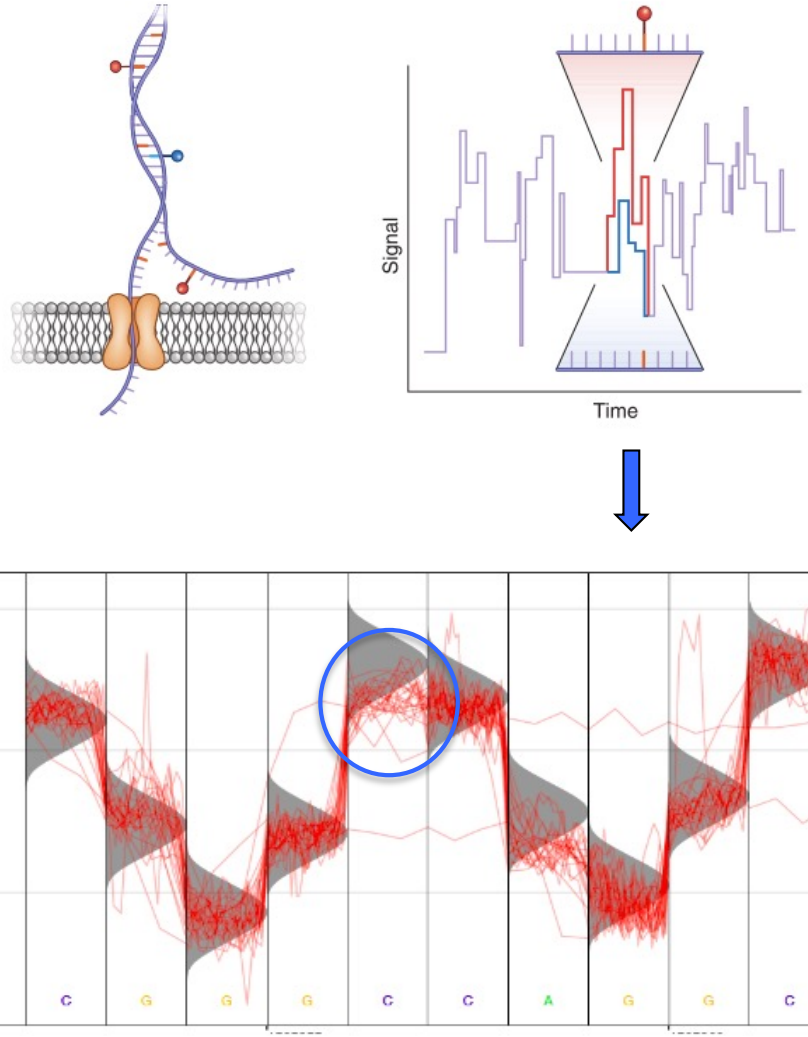
The spatial landscape of gene expression isoforms in tissue sections
Lebrigand et al., *bioRxiv*, 2020

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

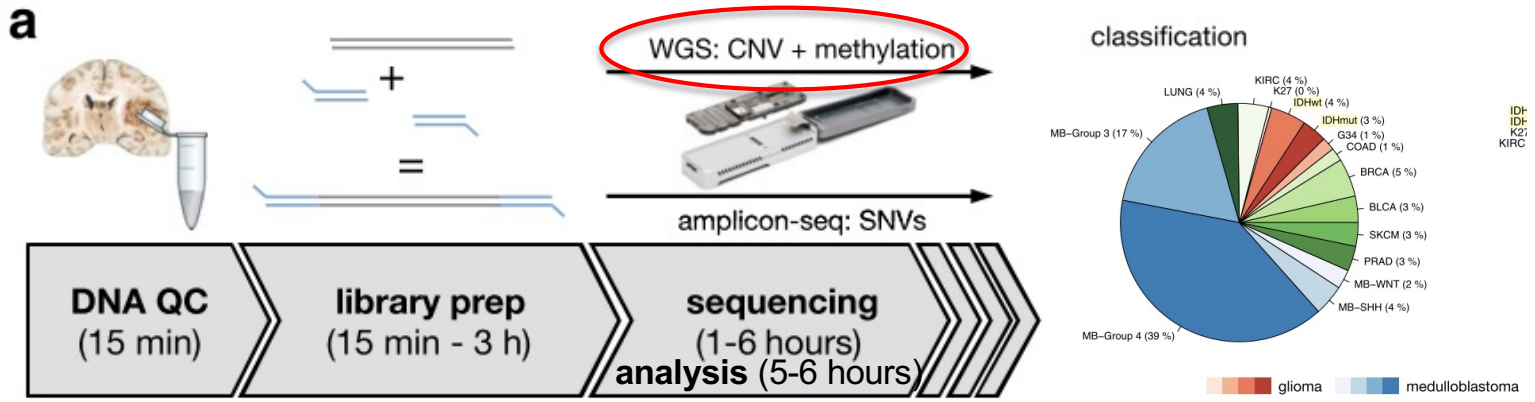


Electric signal

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)

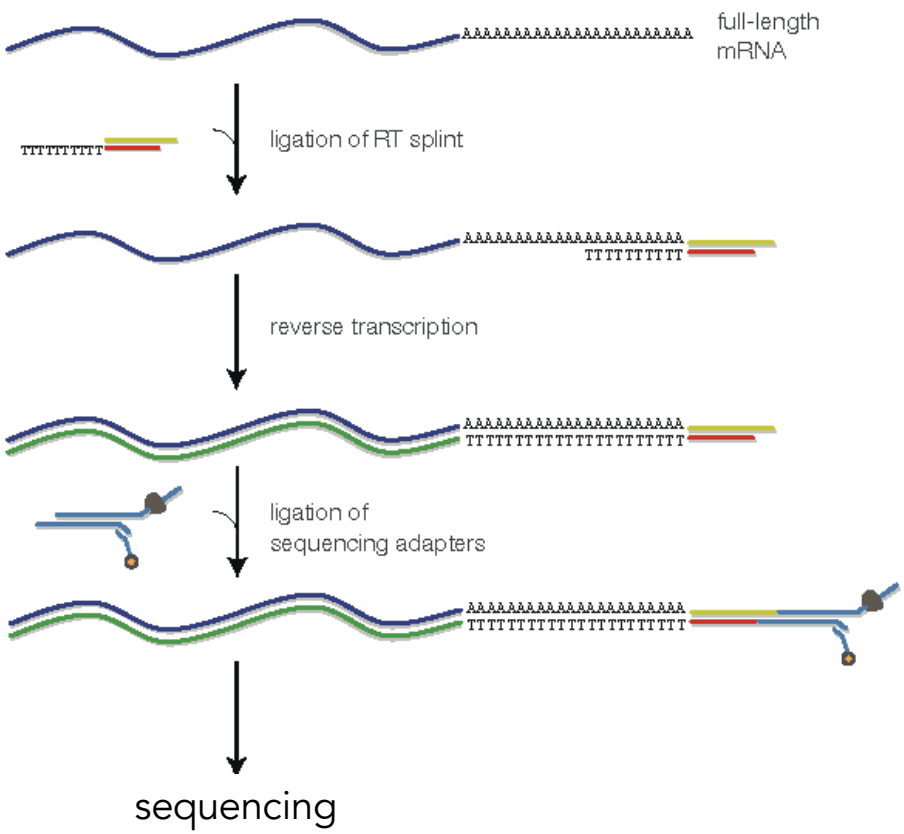


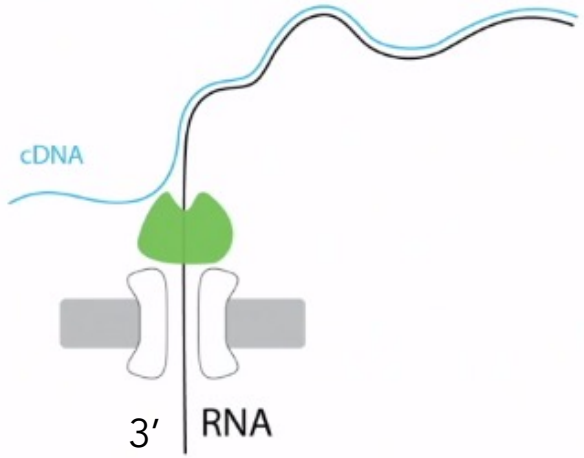
Same-day detection of :
- structural variants
- point mutations
- CpG methylation profiling

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



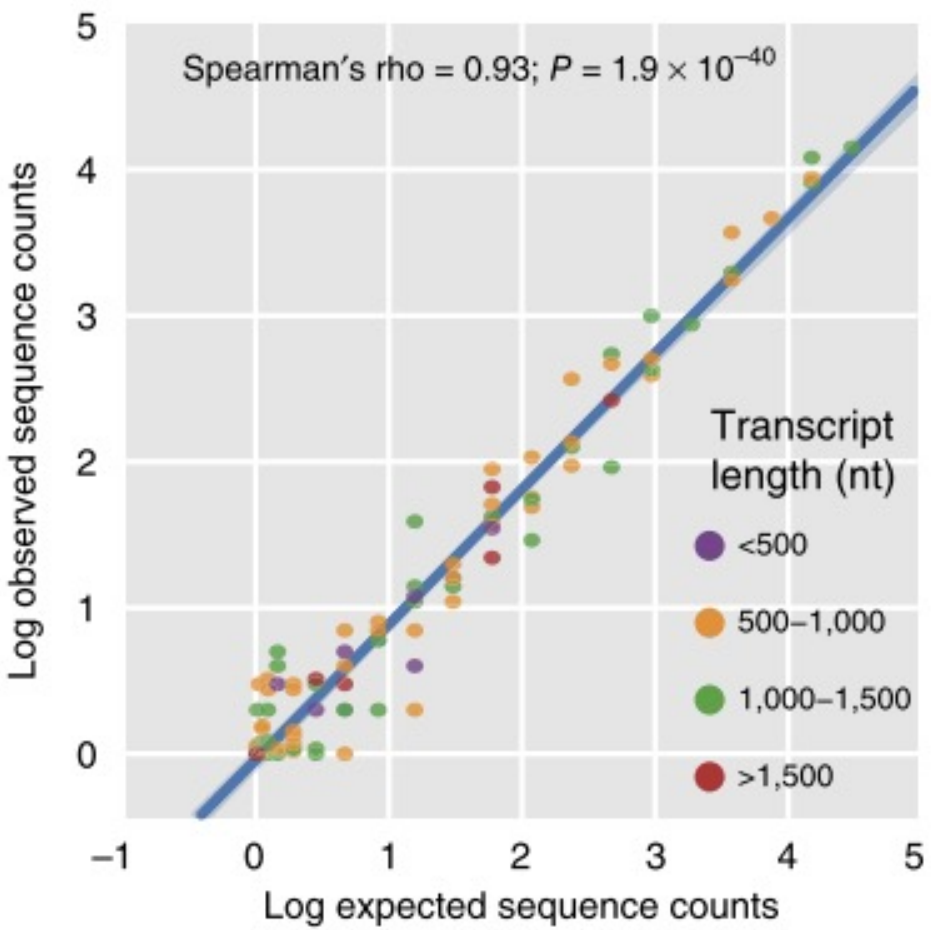


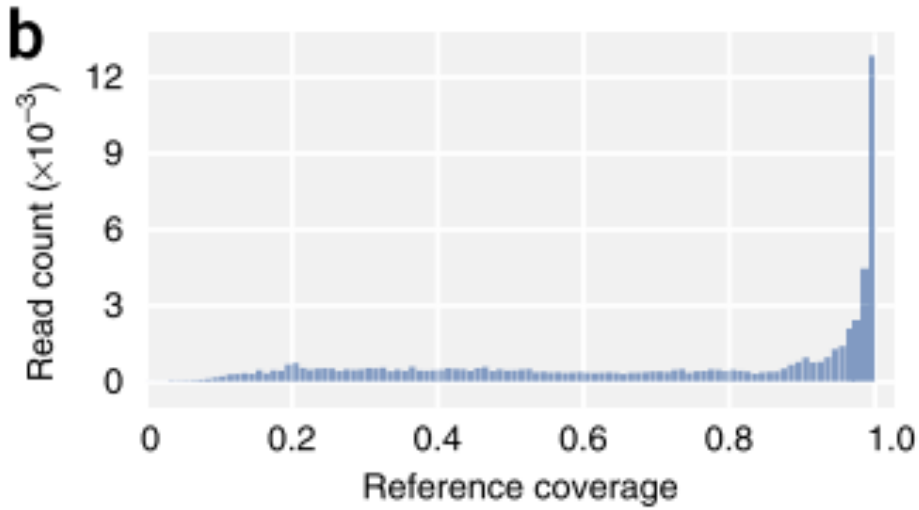
RNA directly sequenced in nanopore

- No PCR bias
- Quantitative

DIRECT RNA SEQUENCING : CONTROLS

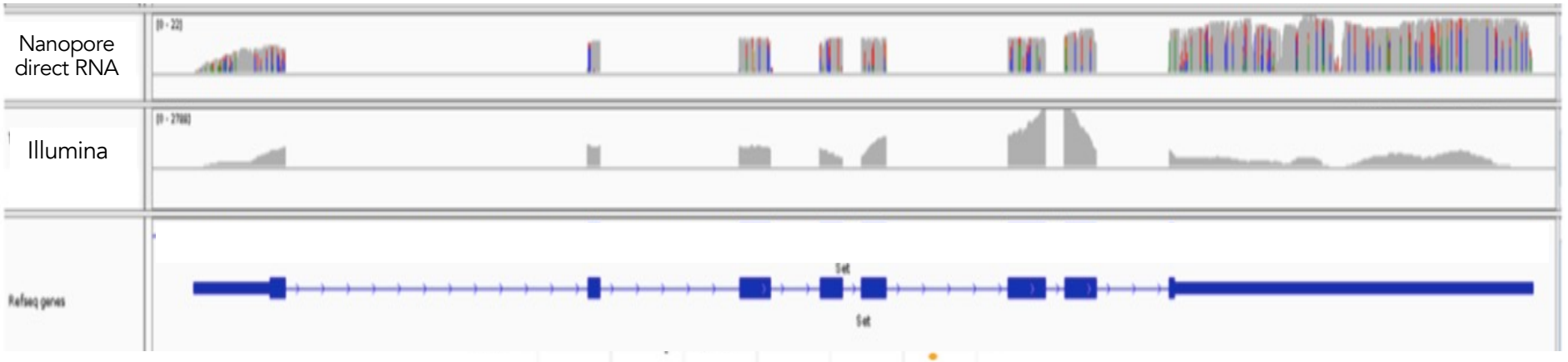
RNA spike in



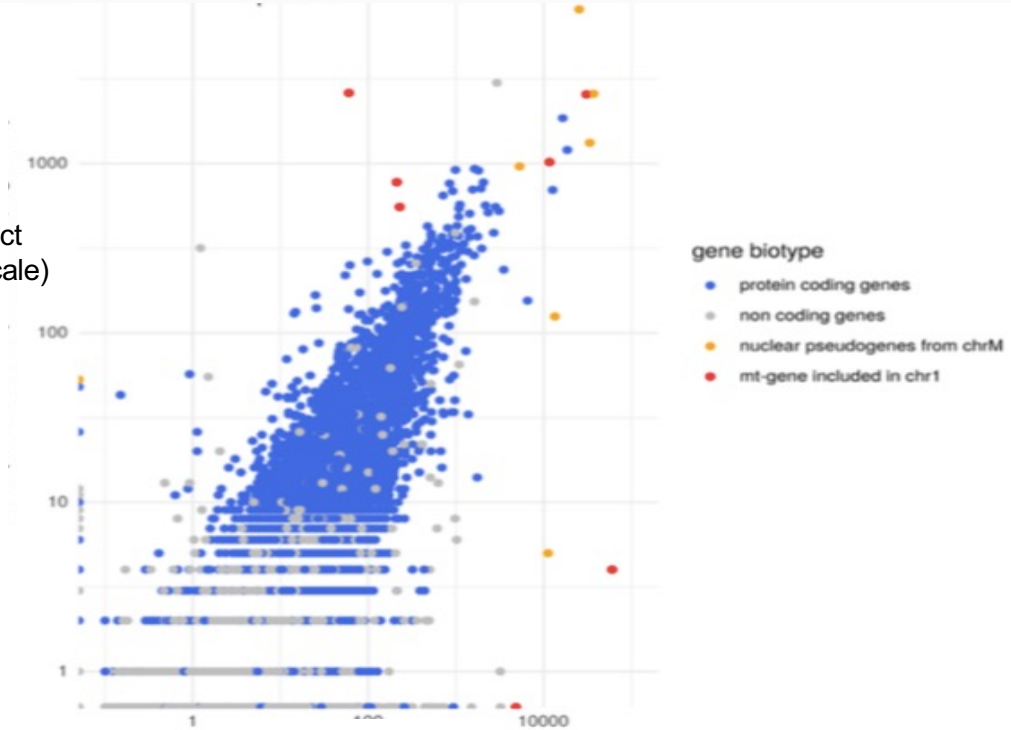


Garalde et al. *Nat. Methods* 2018

DIRECT RNA SEQUENCING vs ILLUMINA



Nanopore RNA direct
(read number ; Log scale)

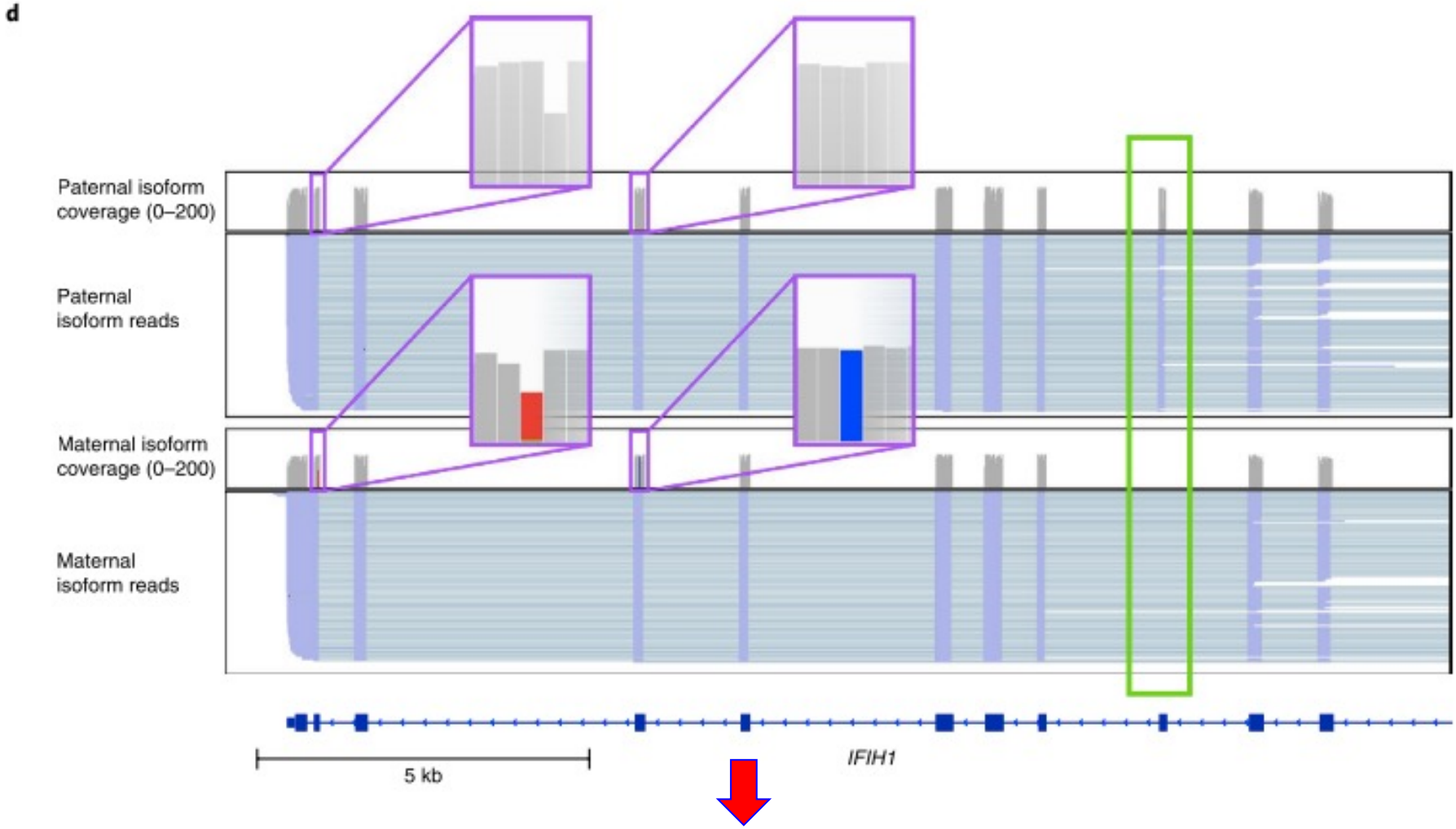


Illumina Truseq

Sessegolo et al. *Sci. Reports* 2019

DIRECT RNA SEQUENCING: TRANSCRIPT HAPLOTYPE

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



34 genes with discordant allele specificity in two isoforms

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

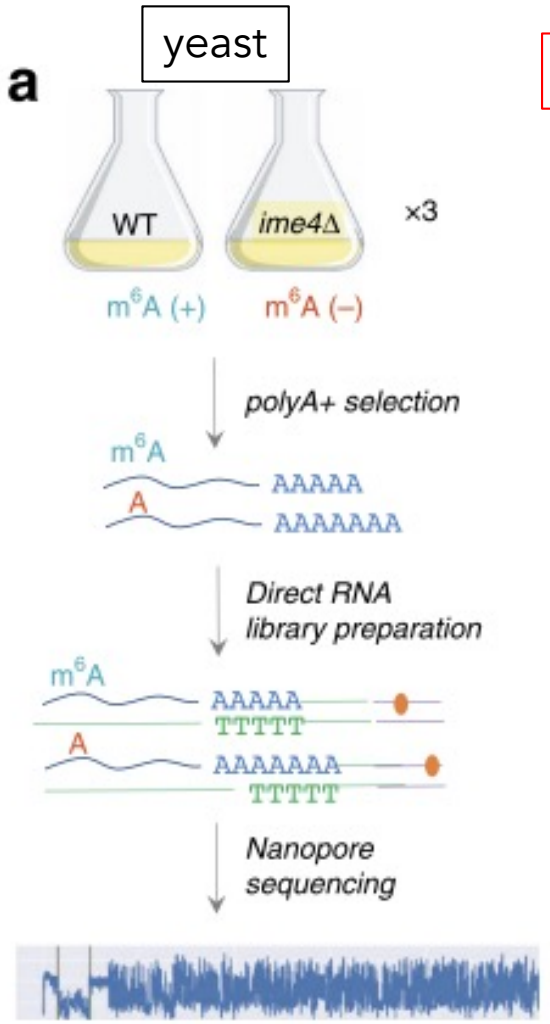
- microRNAs (miRNAs)
- long non-coding RNAs (lncRNAs)
- circular RNAs (circRNAs)

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

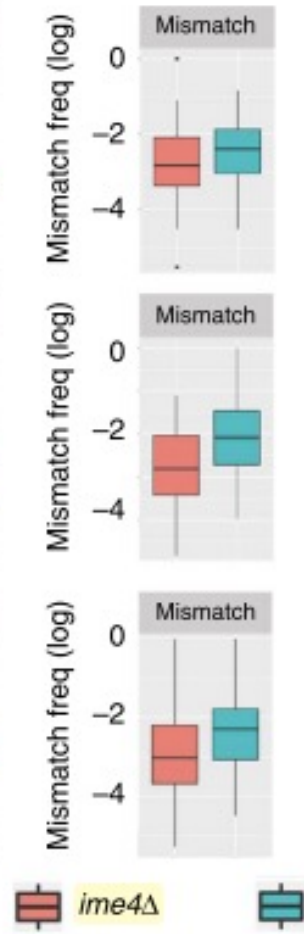
- HIV RNA rich in :
  - *6mA*
  - *5mC*
  - *2'O-methyl*

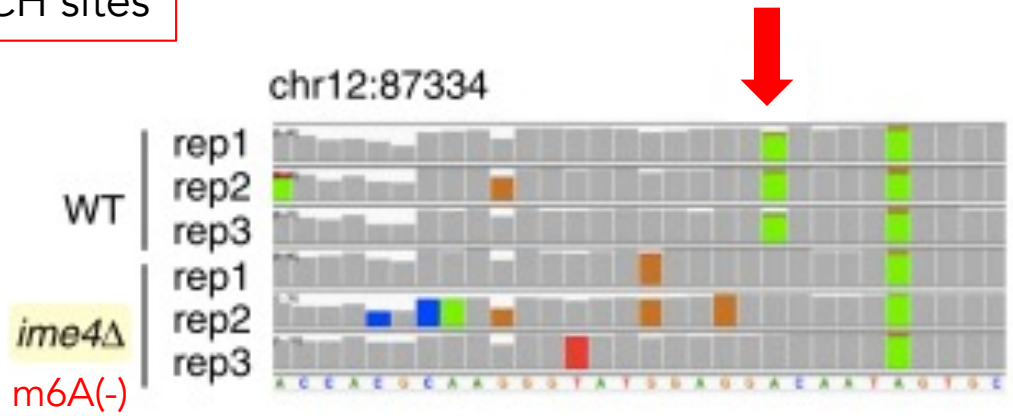
DIRECT RNA SEQUENCING: DETECTION OF m6A

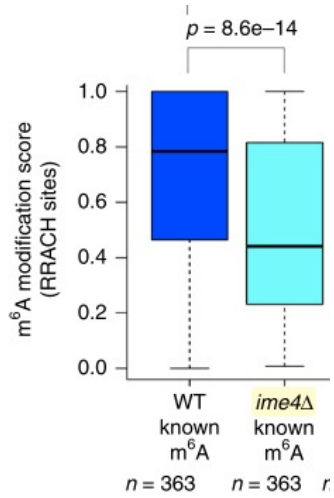
Accurate detection of m6A RNA modifications in native RNA sequences
Liu et al. *Nat. Comm.* 2019

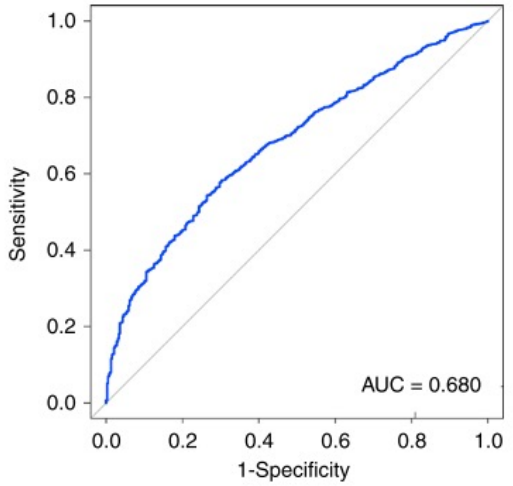


m6A-modified RRACH sites









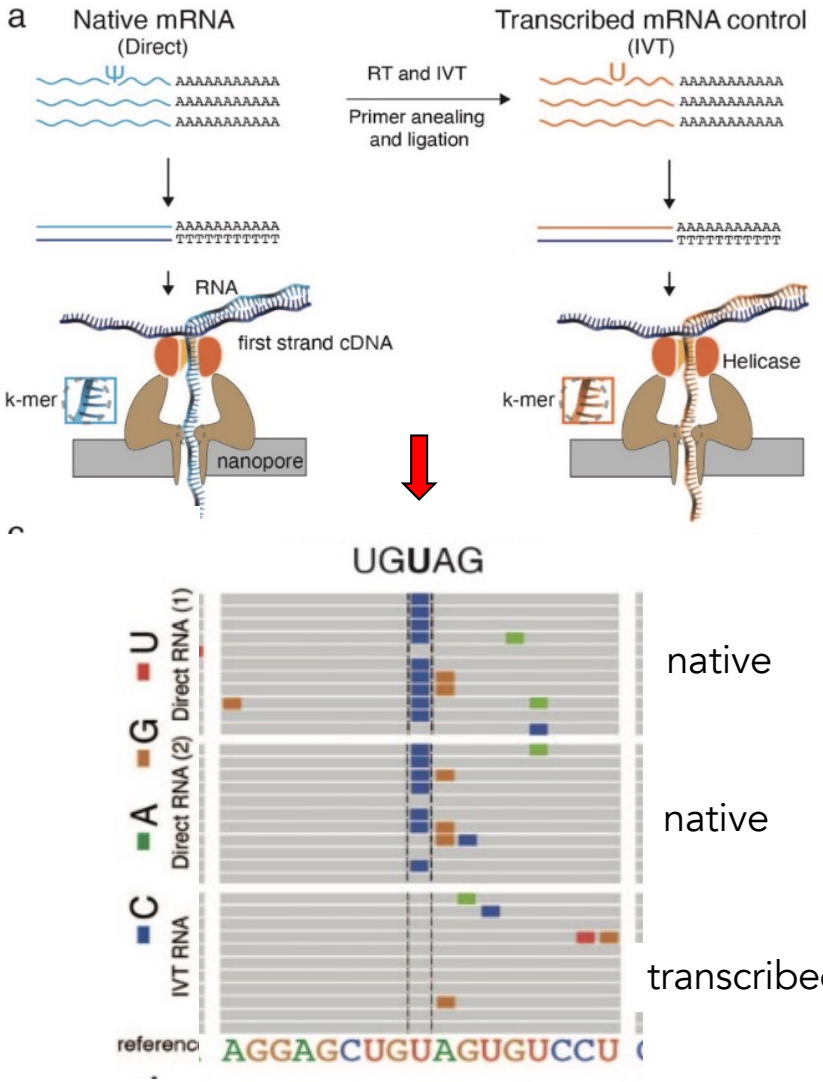
DIRECT RNA SEQUENCING : DETECTION OF pseudo-U

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

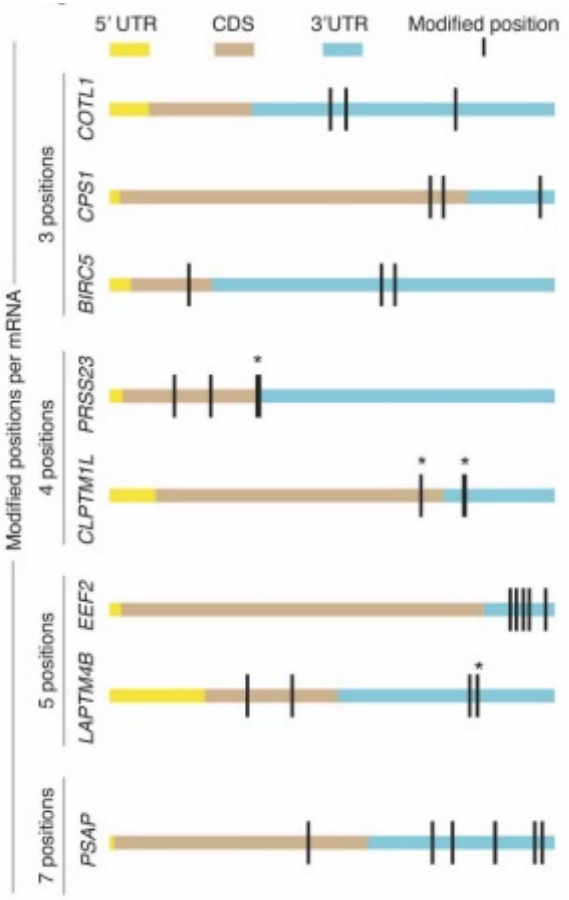
- Pipeline for direct identification, quantification, and detection of pseudouridine modifications and
- Controls :
  - 1000mer synthetic RNA with single pseudouridine in center position
  - U-to-C occurs at the site of pseudouridylation
- Discovery of human mRNAs with up to 7 unique sites of pseudouridine 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 ?

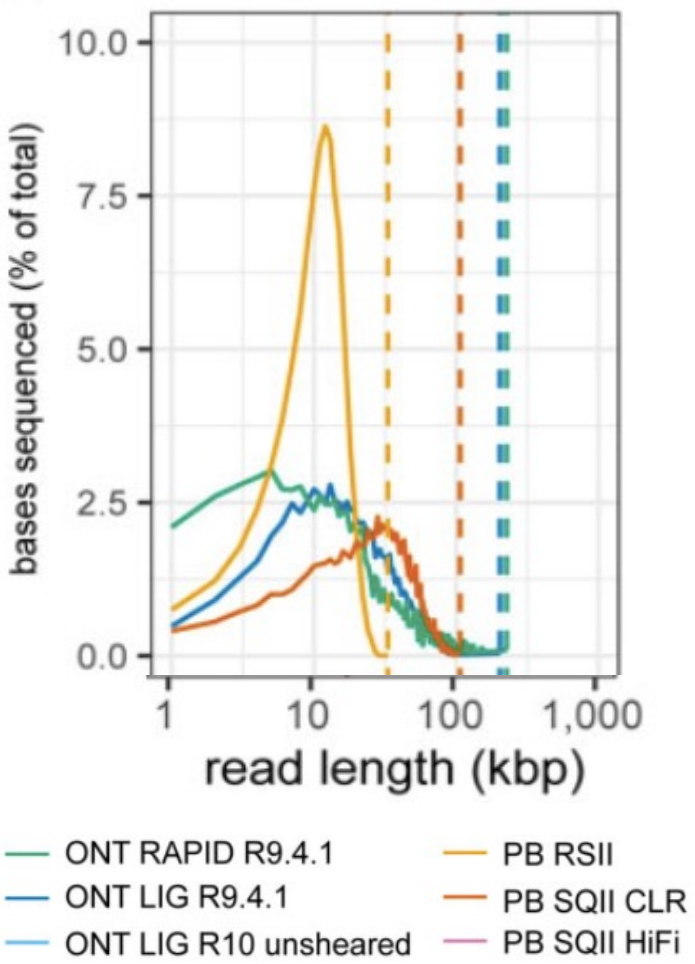
Assembly of small genomes : Nanopore vs PacBio

1 - Comparison of long-read sequencing technologies ininterrogating 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)



Assembly of small genomes : Nanopore vs PacBio

1 - Comparison of long-read sequencing technologies ininterrogating 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 ininterrogating 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

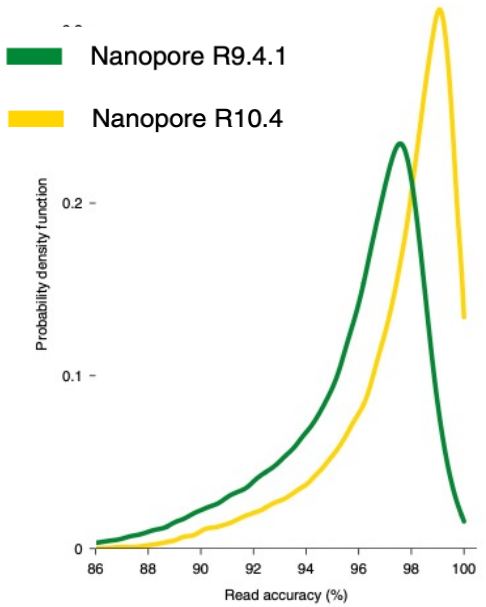
- All libraries
  - High genome contiguity (longer reads)
  - Better detection of DNA methylation motifs than PacBio (HiFi and CLR)
- ONT Rapid
  - Lowest percentage of chimeric reads (0.02%)
  - Detection of the 2 plasmids
- ONT : more cost-effective

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

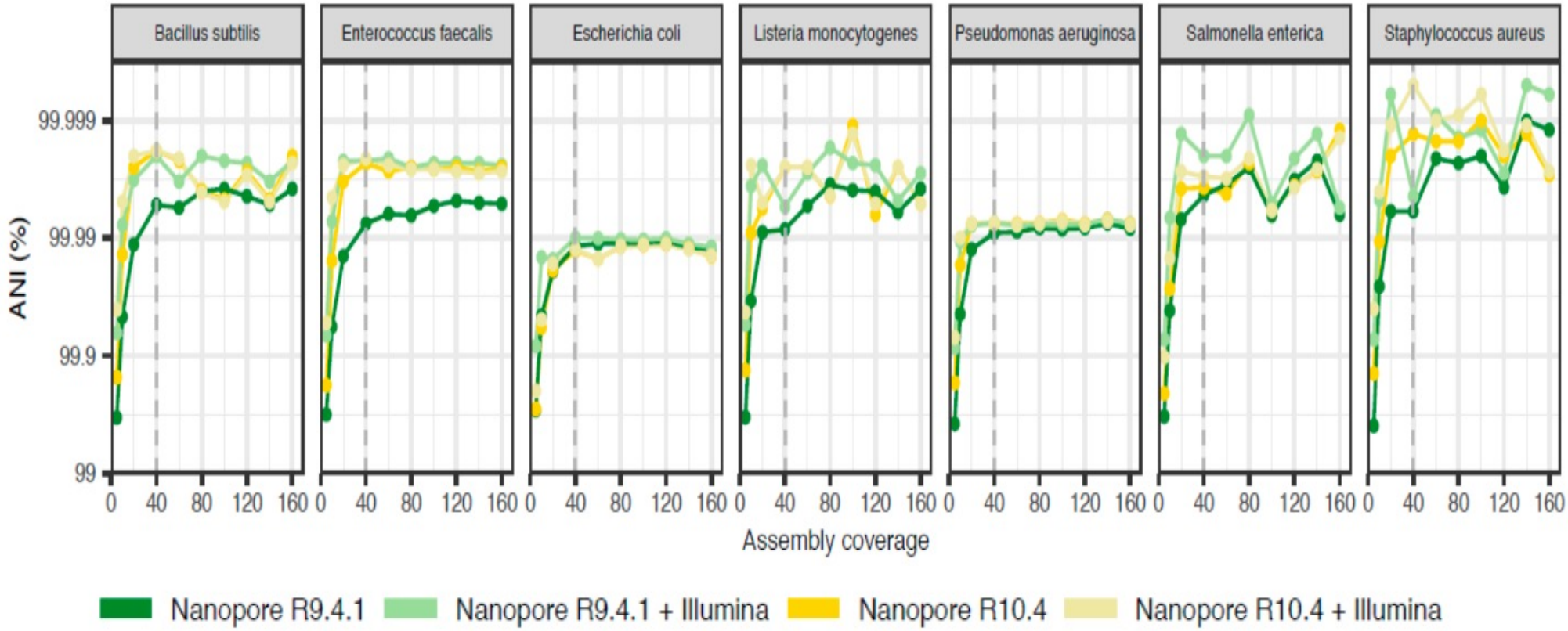
Read accuracies



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)



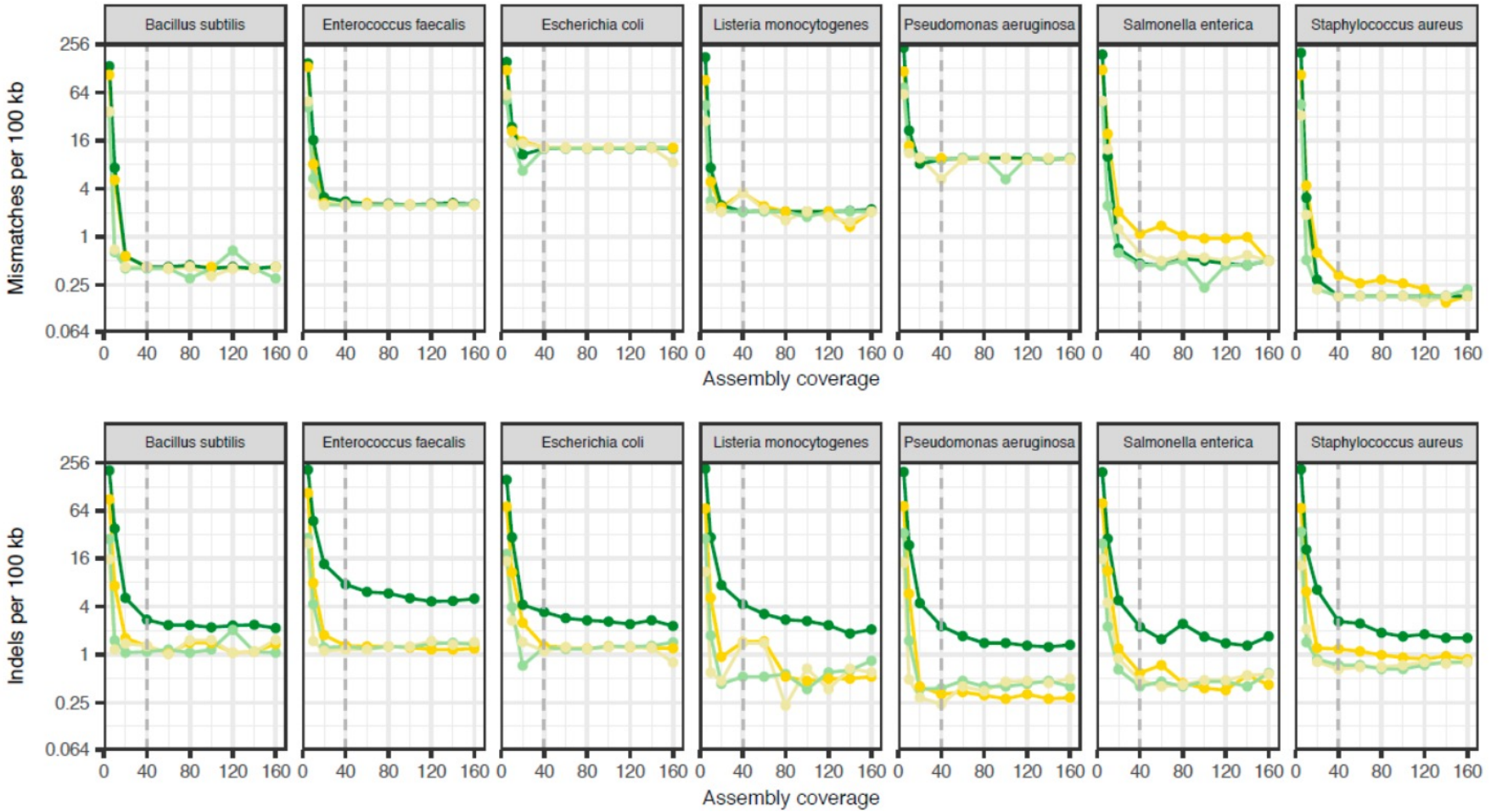
$$gANI_{G1 \rightarrow G2} = \frac{\sum_{bbh} (\text{Percent Identity} * \text{Alignment length})}{\text{lengths of BBH genes}}$$

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

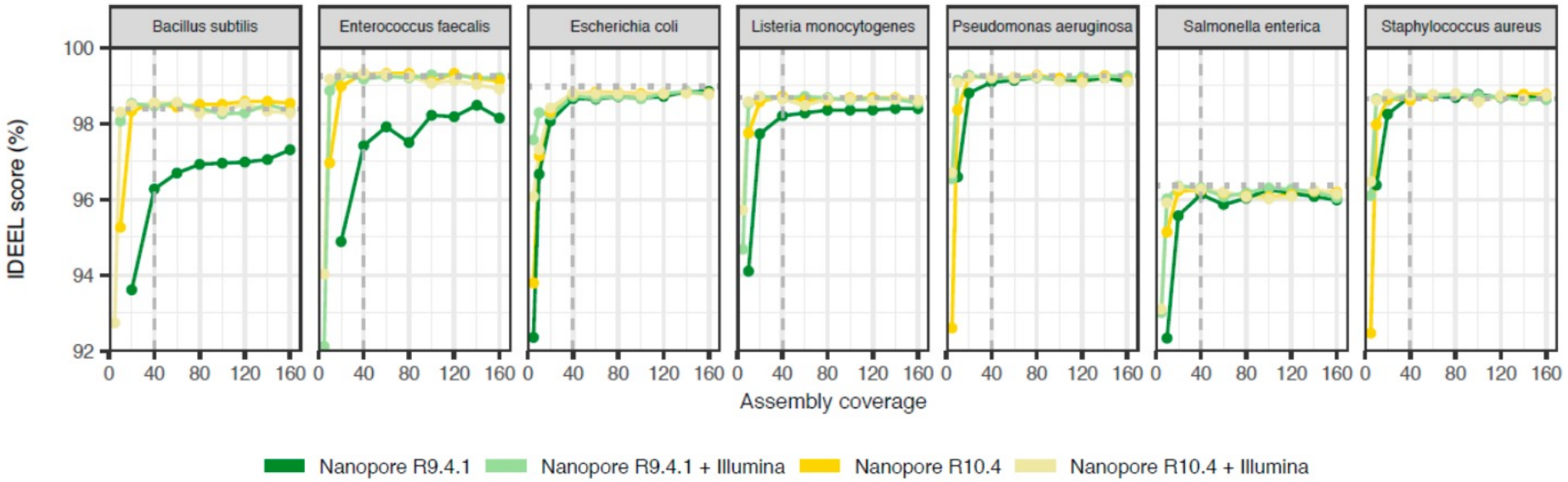
Sequencing and assembly statistics for the bacterial species (*n* = 7)



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)



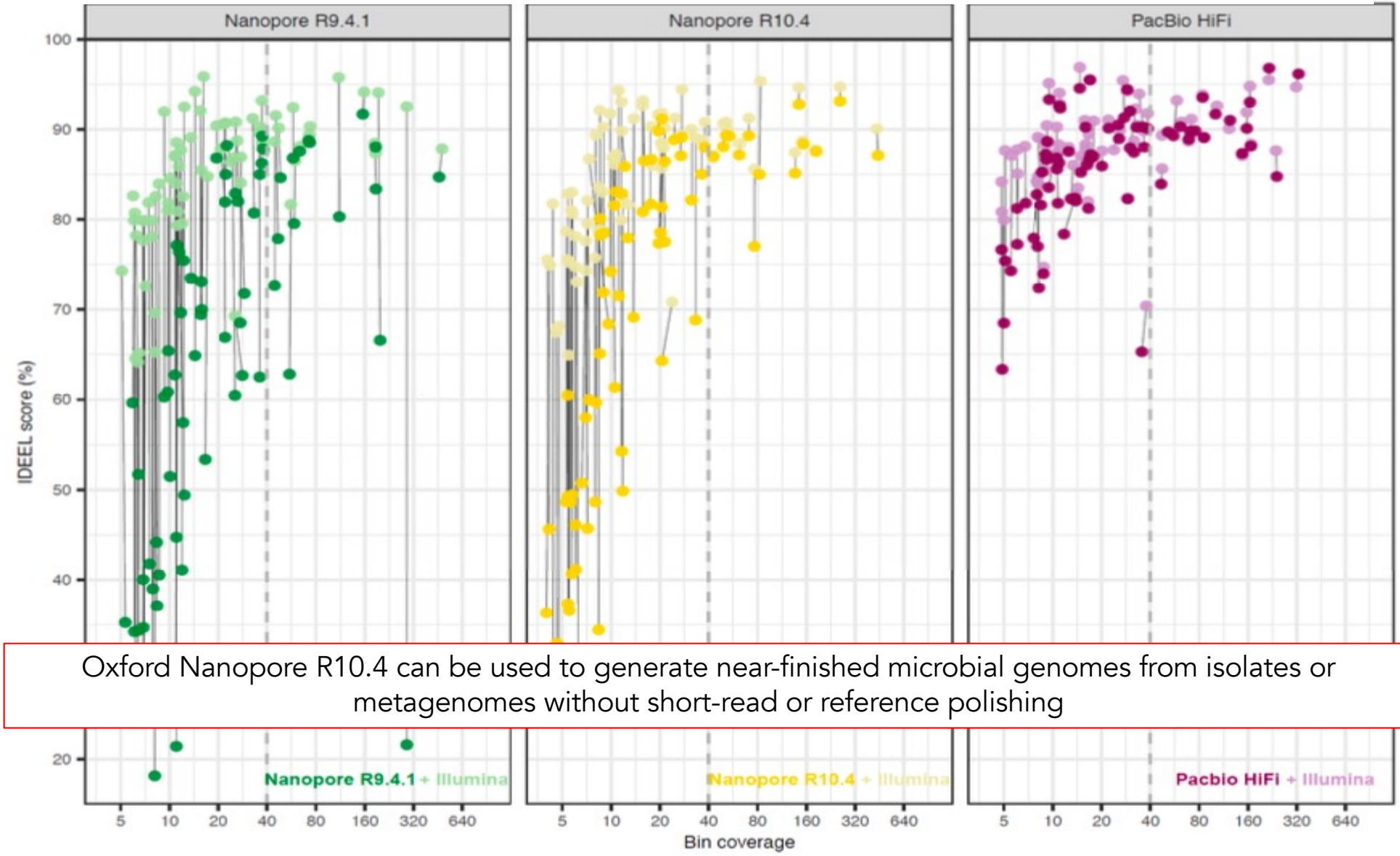
IDEEL score : proportion of predicted proteins that are ≥95% the length of their best-matching known protein in a database

➢ 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

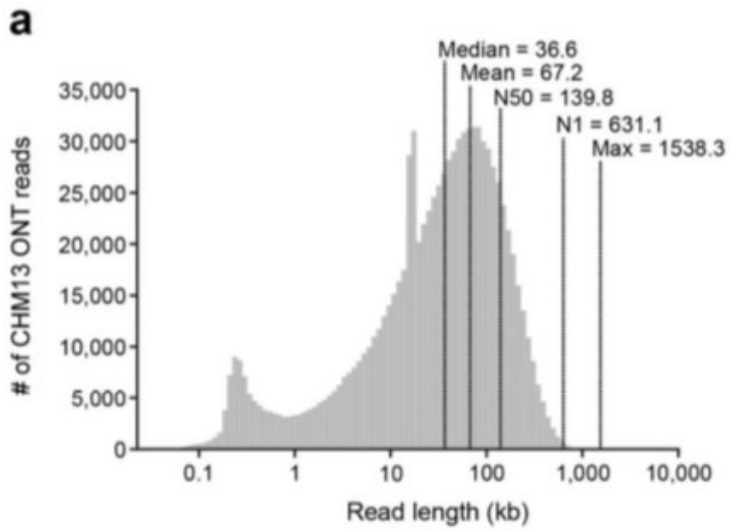
2 – Human Pangenome Project :
- Semi-automated assembly of high-quality diploid human reference genomes. Jarvis et al. *Nature* July 2022

Large genome assembly : Nanopore + PacBio

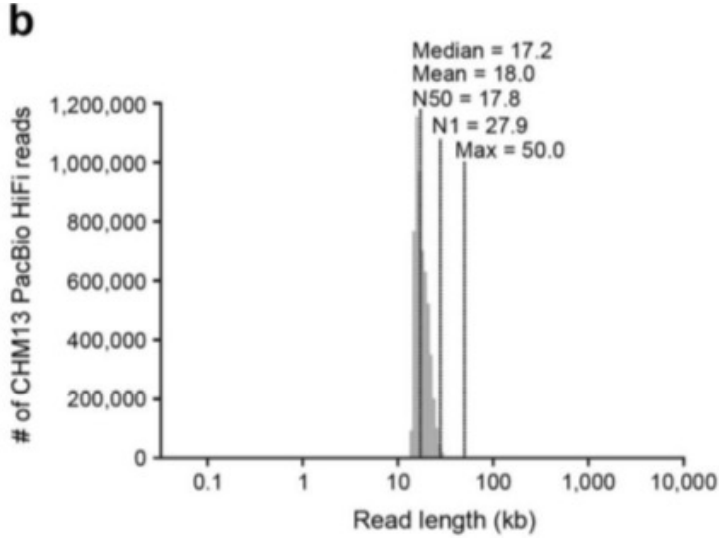
1 - The structure, function, and evolution of a complete human chromosome 8
Logsdon et al., *Nature*, May 2021

- 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



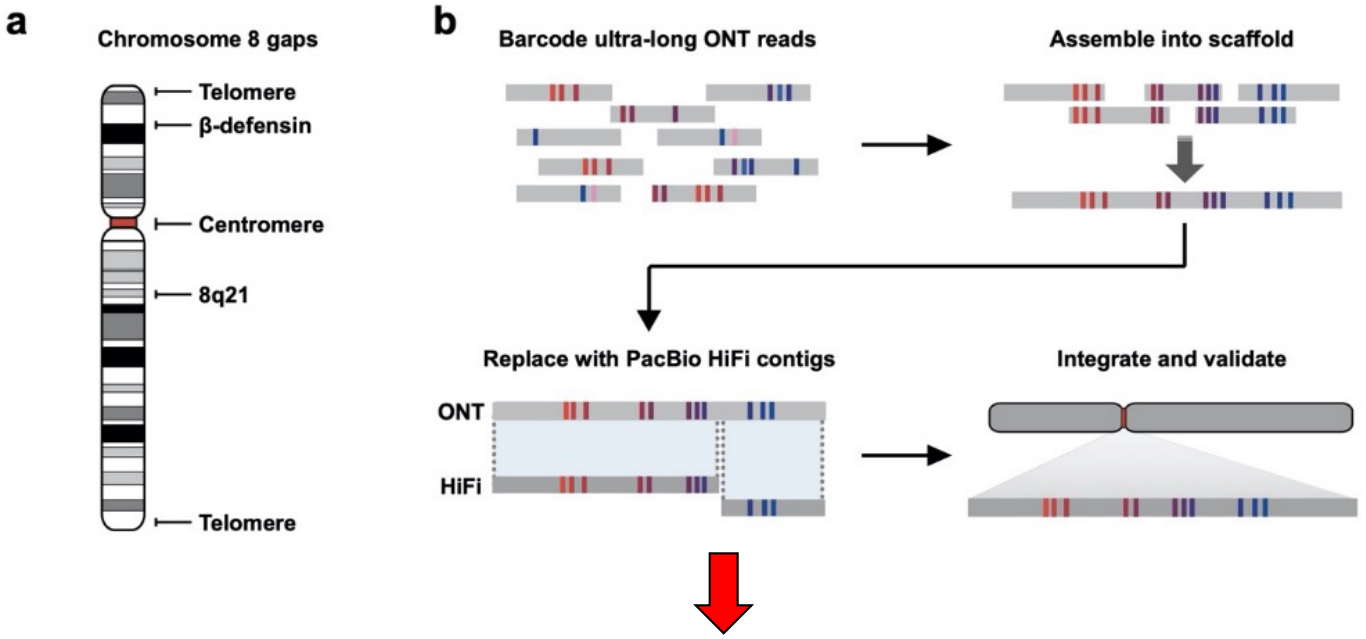
PacBio HiFi reads



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..

Large genome assembly : Nanopore + PacBio

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. 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*]

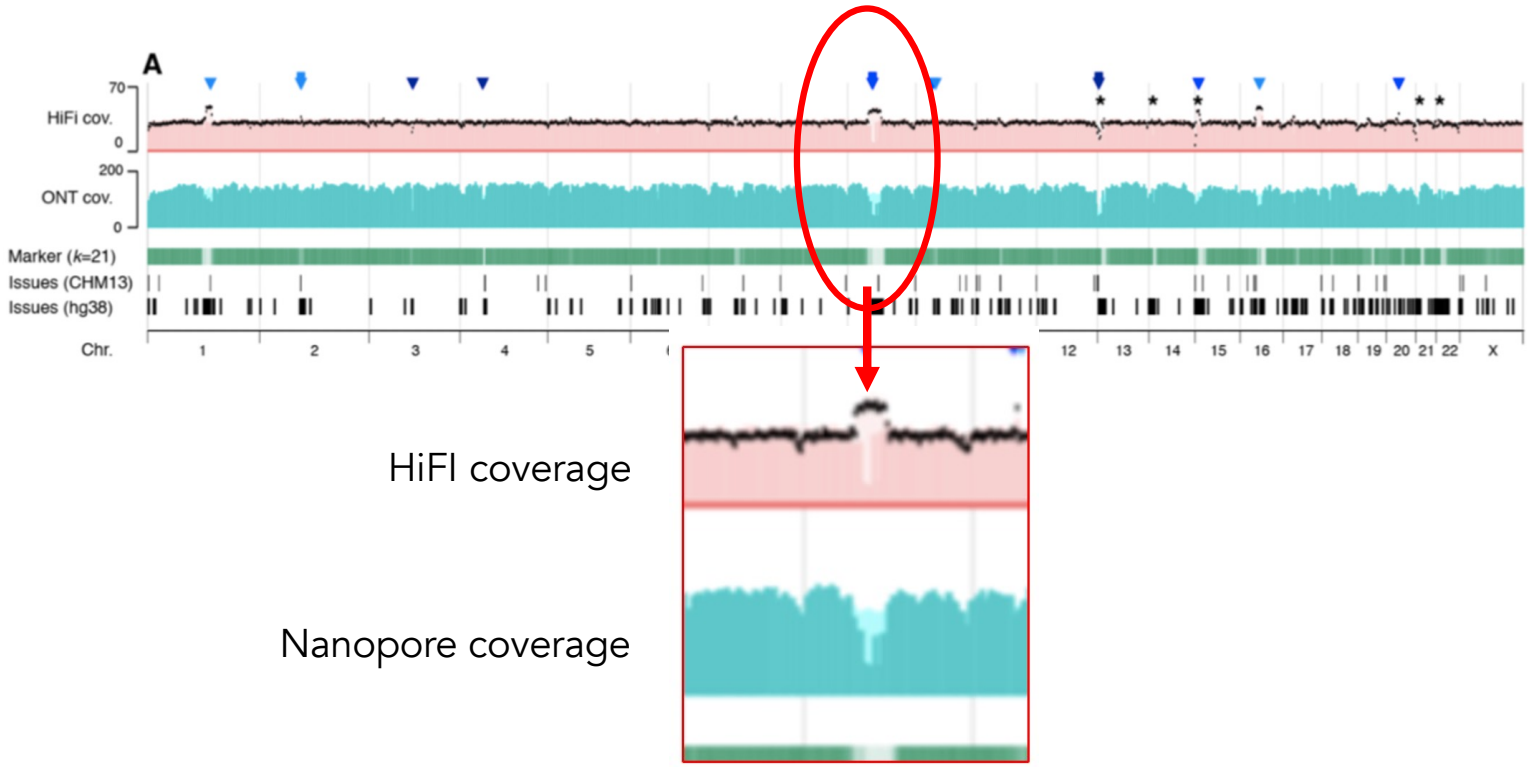
Large genome assembly : Nanopore + PacBio

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

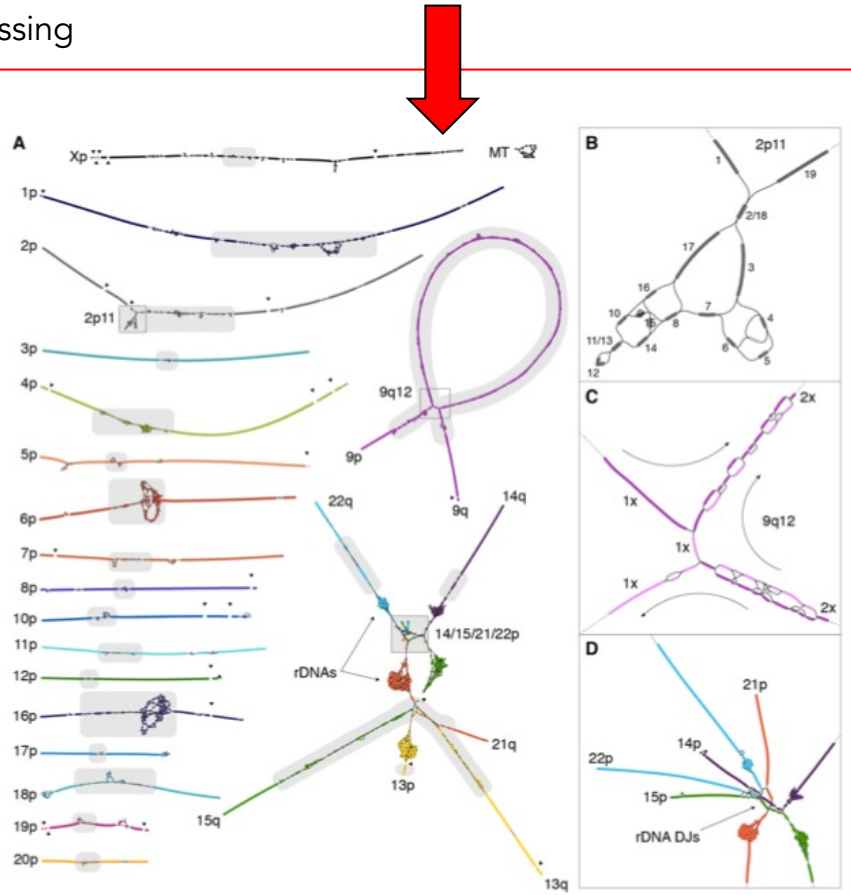


Large genome assembly : Nanopore + PacBio

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

ASSEMBLY

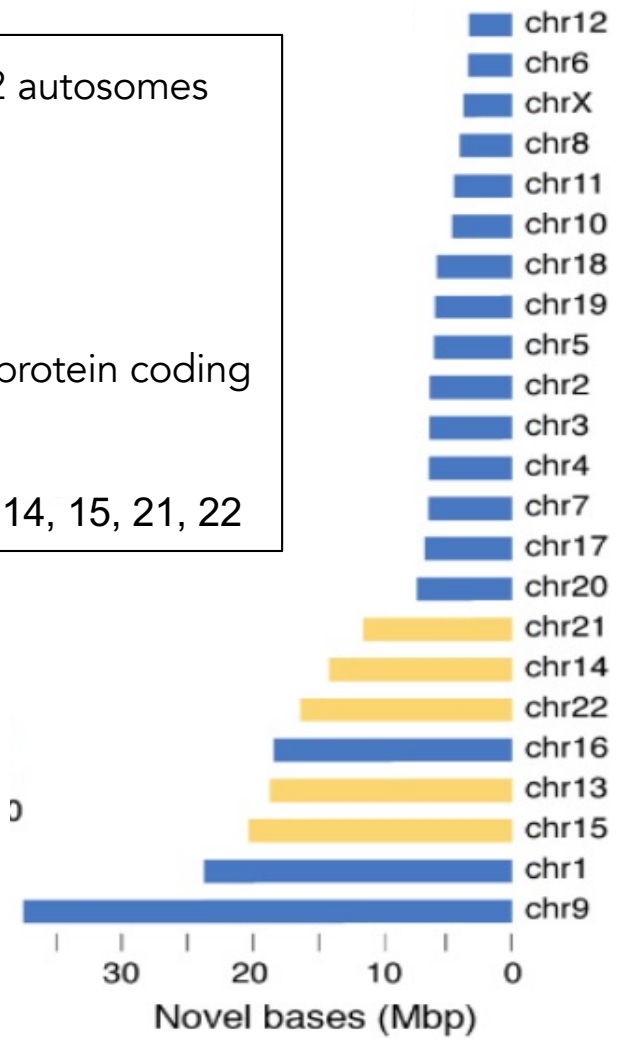
- HiFi-based string graph constructed using a purpose-built method that combines components from
  - HiCanu
  - Miniasm
  - specialized graph processing



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



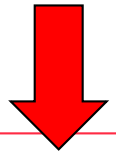
Large genome assembly : Nanopore + PacBio

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 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 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[1,103], 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[40,53], 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[63], 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)

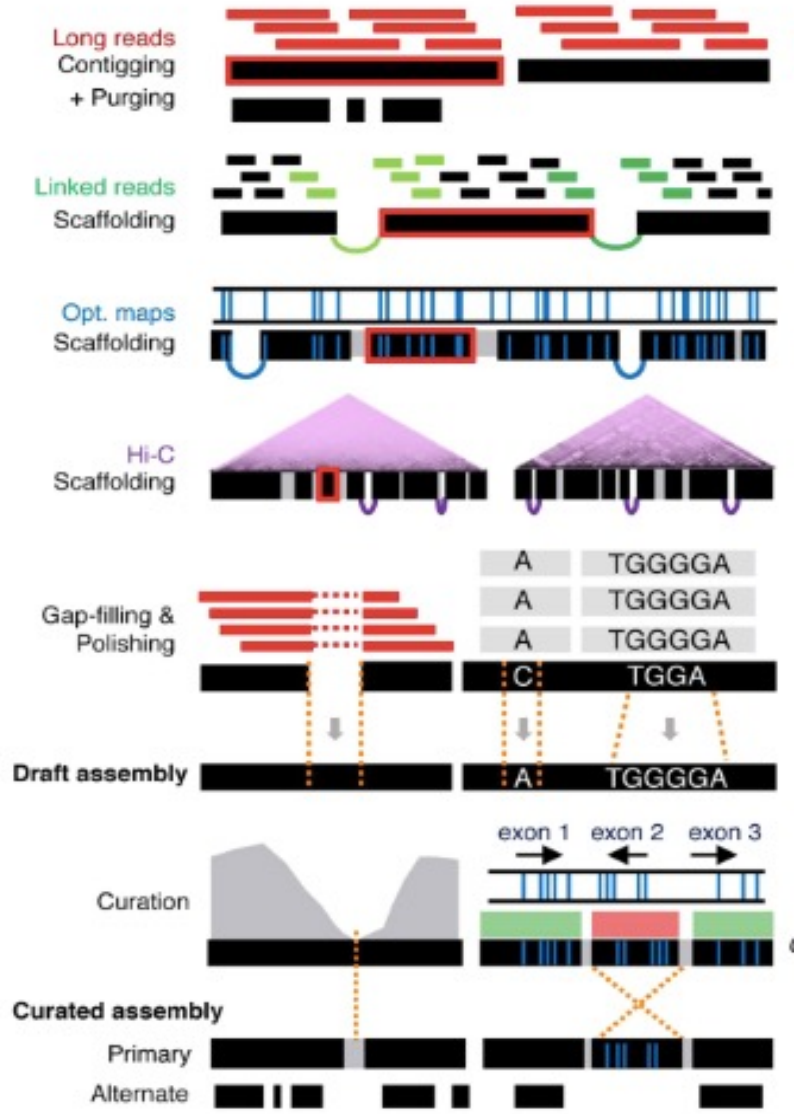
Towards complete and error-free genome assemblies of all vertebrate species
Rhie et al. *Nature* 2021

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)

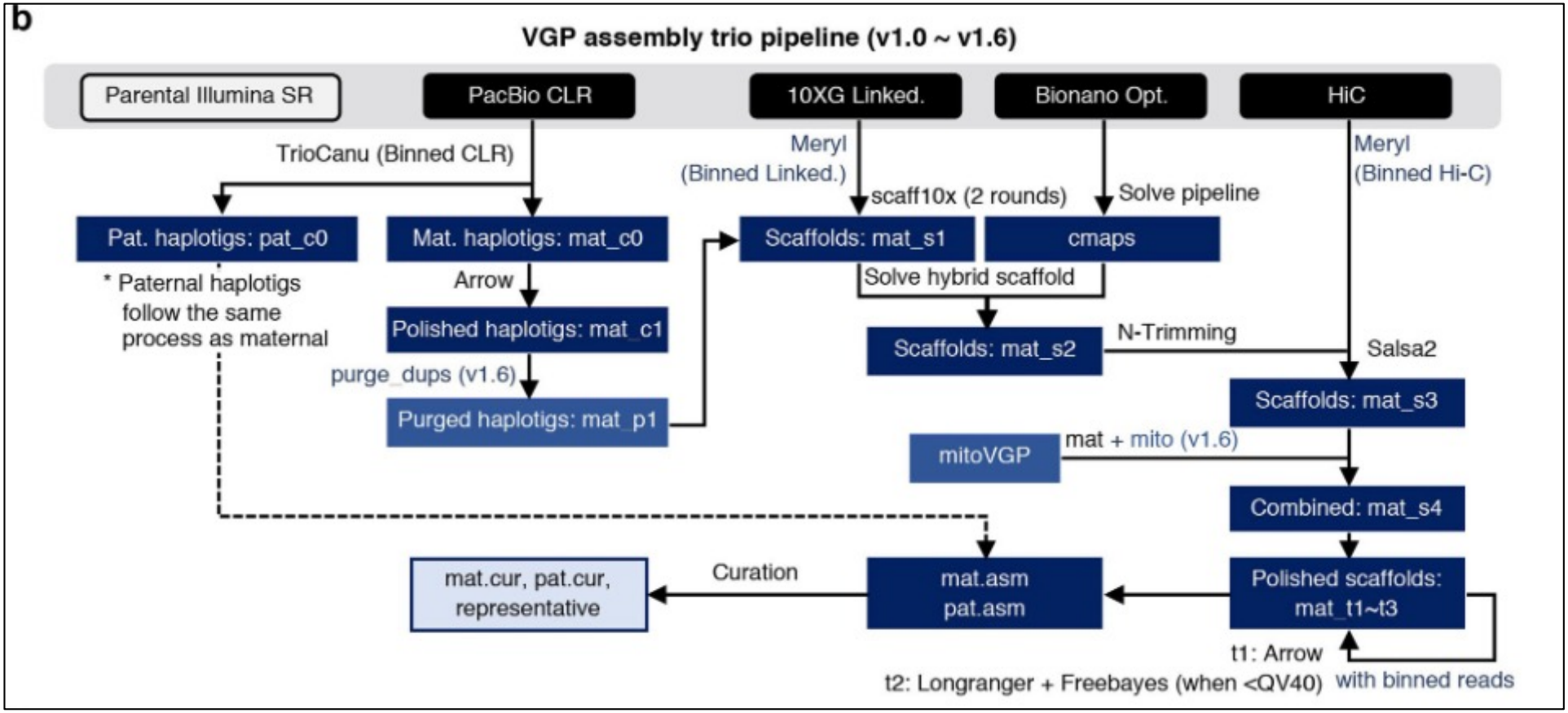
We will take advantage of continuing improvements in genome sequencing technology, assembly, and annotation, including advances in PacBio HiFi reads, Oxford Nanopore reads, and replacements for 10XG reads



VERTEBRATE GENOMES PROJECT (VGP)

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

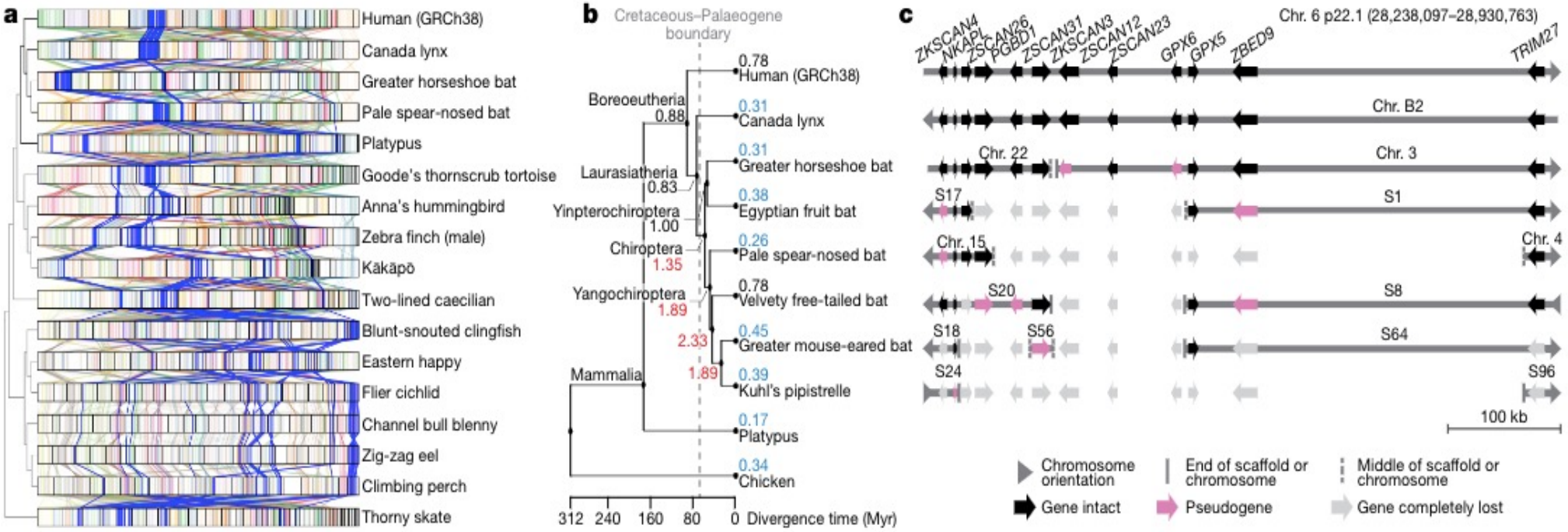
Standard VGP trio assembly pipeline when DNA is available for a child and parents20.



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)

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

- 23 genome assemblies, generated with 23 different methods

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

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-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]

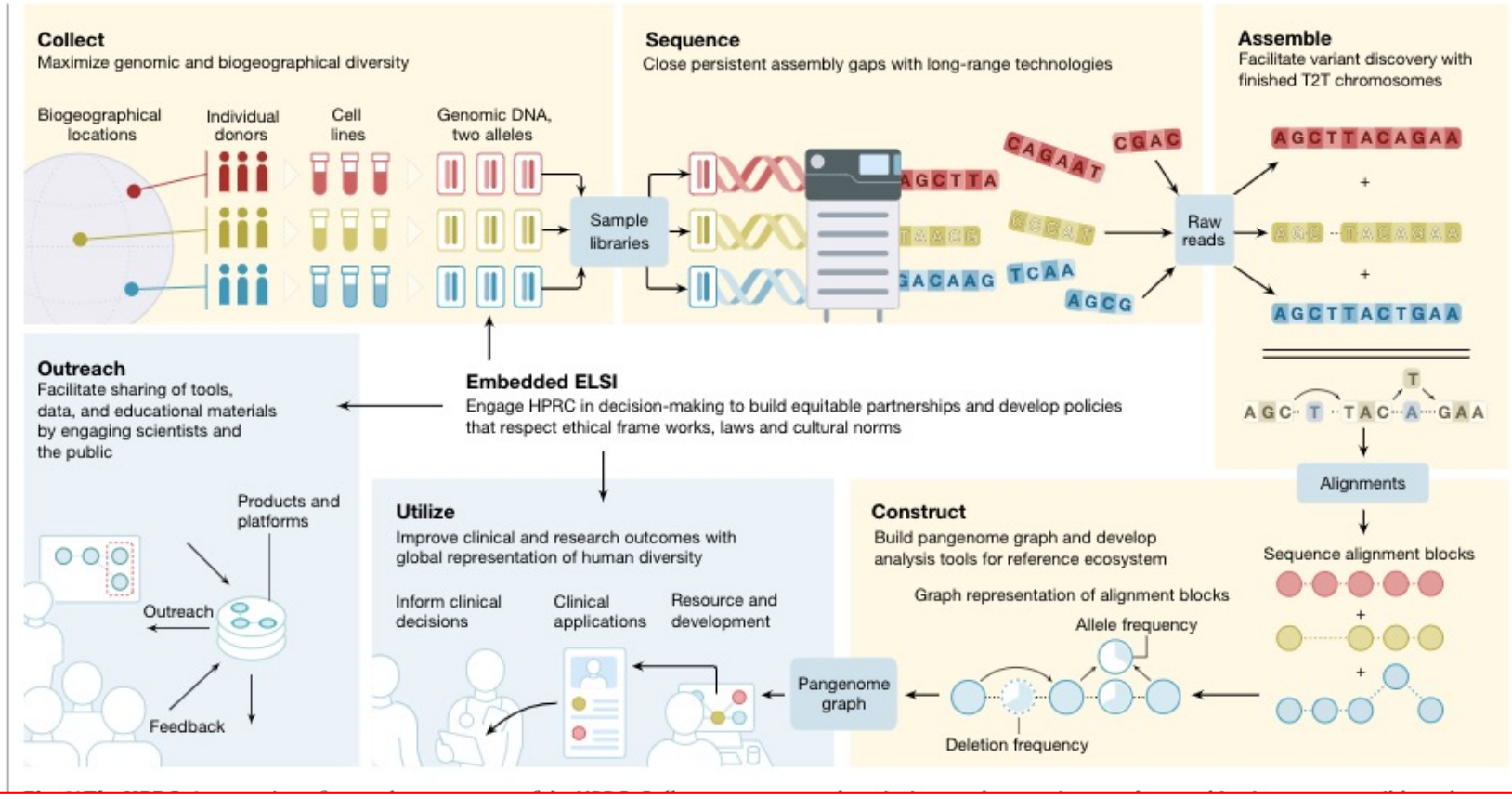
Sample Working Group including Teams for Population Genetics and Ethical, Legal, and Social Issues
Carlos Bustamante[8], Judy Cho[9,10,11], Robert Cook-Deegan[12], Jean-Francois Deleuze[13], Richard Durbin[14,15], Simon Easteal[16], Evan E. Eichler[17,18], Xiaowen Feng[19,20], Nanibaa Garrison[21,22,23], 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 Meredith[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



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 leve.
- 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
- 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