LONG-READ SEQUENCING

Claude THERMES

PLATEFORME DE SÉQUENÇAGE I2BC

INSTITUT DE BIOLOGIE INTÉGRATIVE DE LA CELLULE

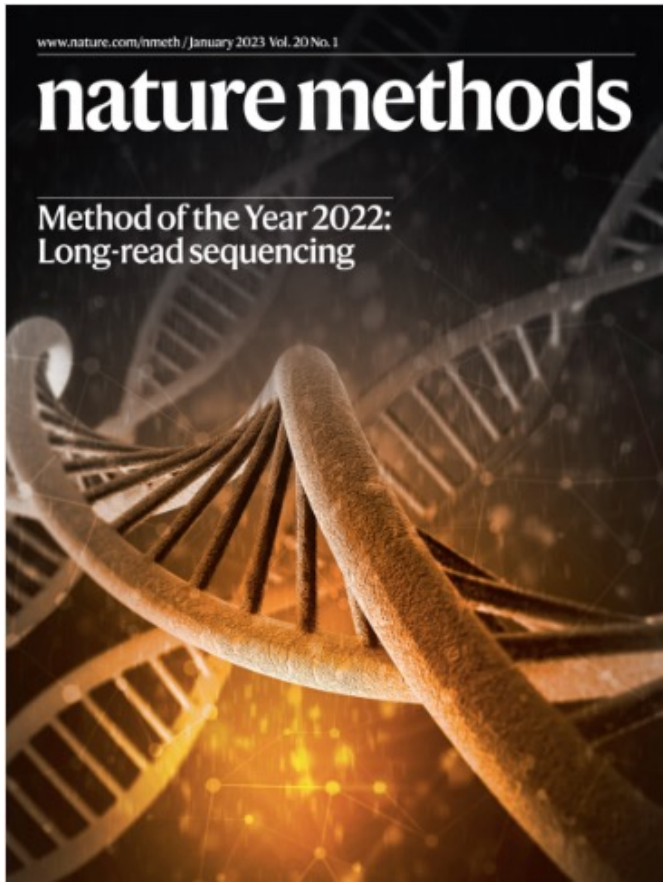
GIF-SUR-YVETTE

13ème ÉCOLE DE BIOINFORMATIQUE EBAII - 18/11/2024

LONG-READ SEQUENCING

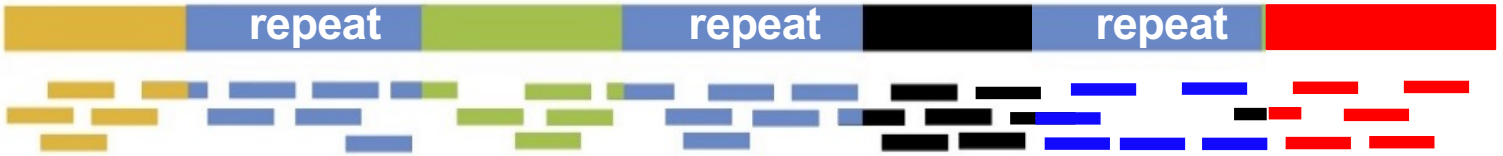
Long-reads: ≈1 kb to ≈100 kb (ultra-long reads: > 100 kb)

- Genome assembly
- Haplotype phasing
- Splicing isoforms

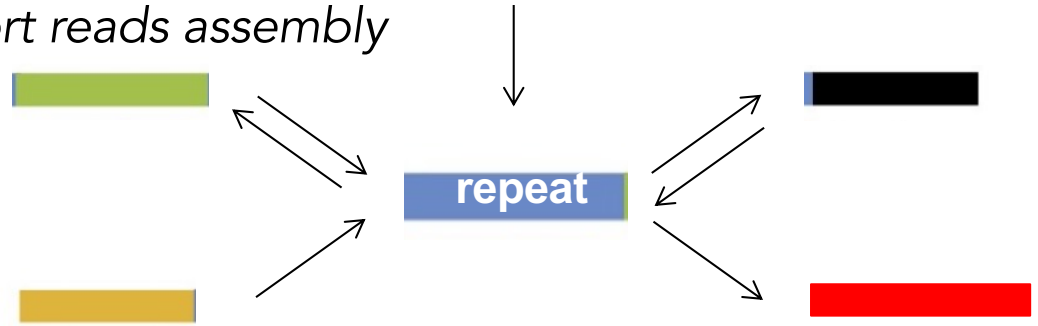


LONG-READS VERSUS SHORT-READS : GENOME ASSEMBLY

Assembly of DNA fragments with repeated sequences



NGS short reads assembly



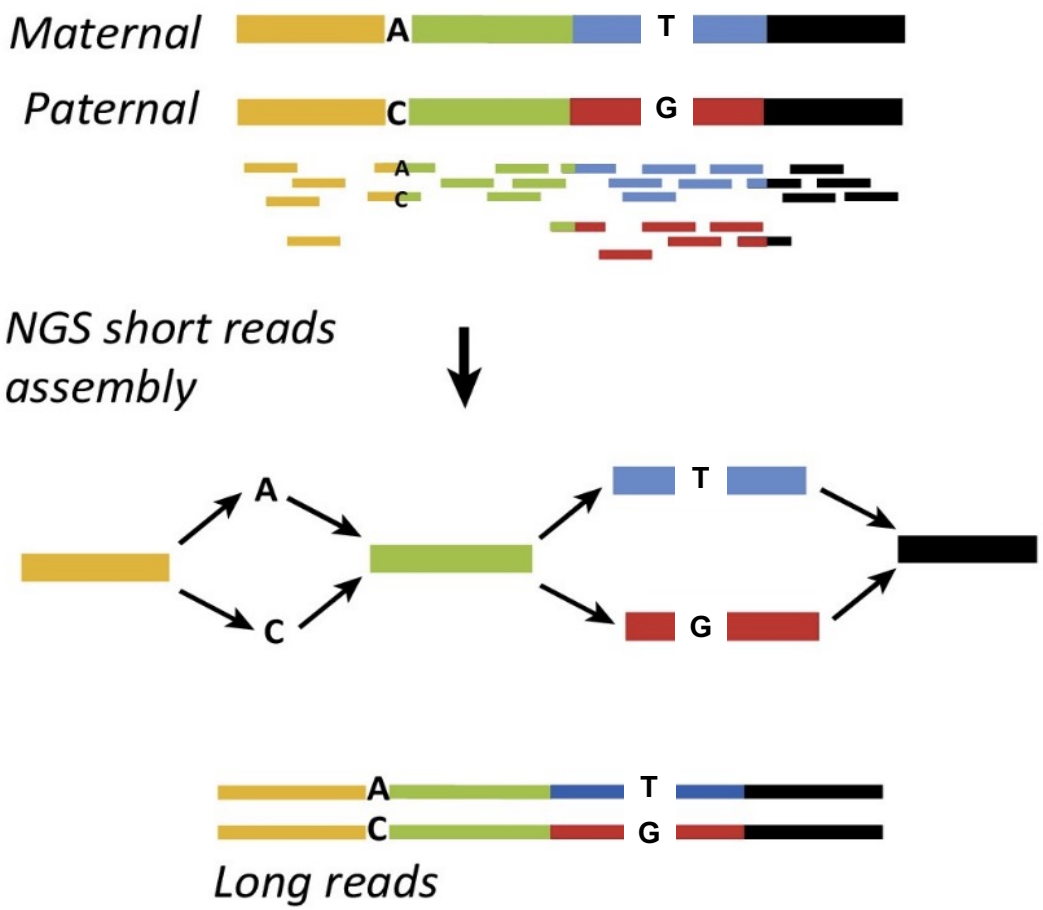
Several contigs ⟶ incomplete assembly, underestimation of repeats

Long reads assembly

Long-reads (1- 200 kb) allow assembly of large repeat-rich regions (centromeres, telomeres…)

LONG-READS VERSUS SHORT-READS : HAPLOTYPES

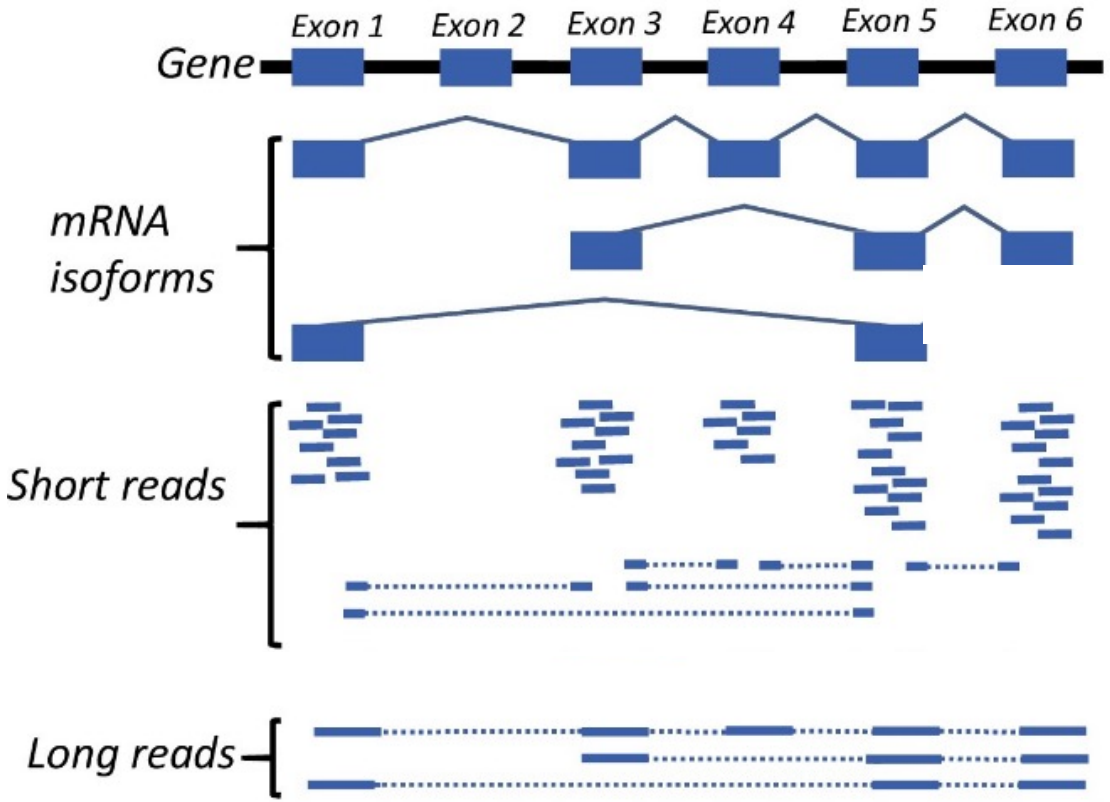
Haplotype phasing



Long-reads allow phasing of maternal and paternal haplotypes

LONG-READS VERSUS SHORT-READS : SPLICING ISOFORMS

Detection of splicing isoforms



Long-reads allow identification of multiple splicing events

along each mRNA molecule

The 3rd generation winning technologies

Pacific Biosciences

Vega

Revio

Single molecules
Up to 200 kbp long

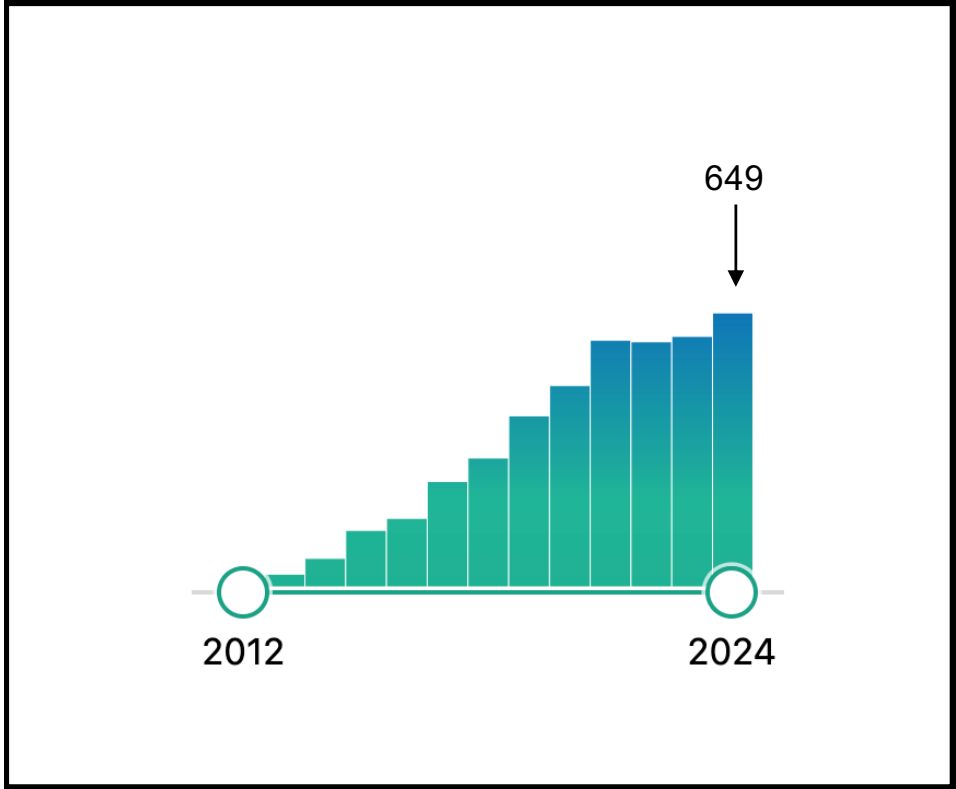
Oxford Nanopore

MinION – PromethION

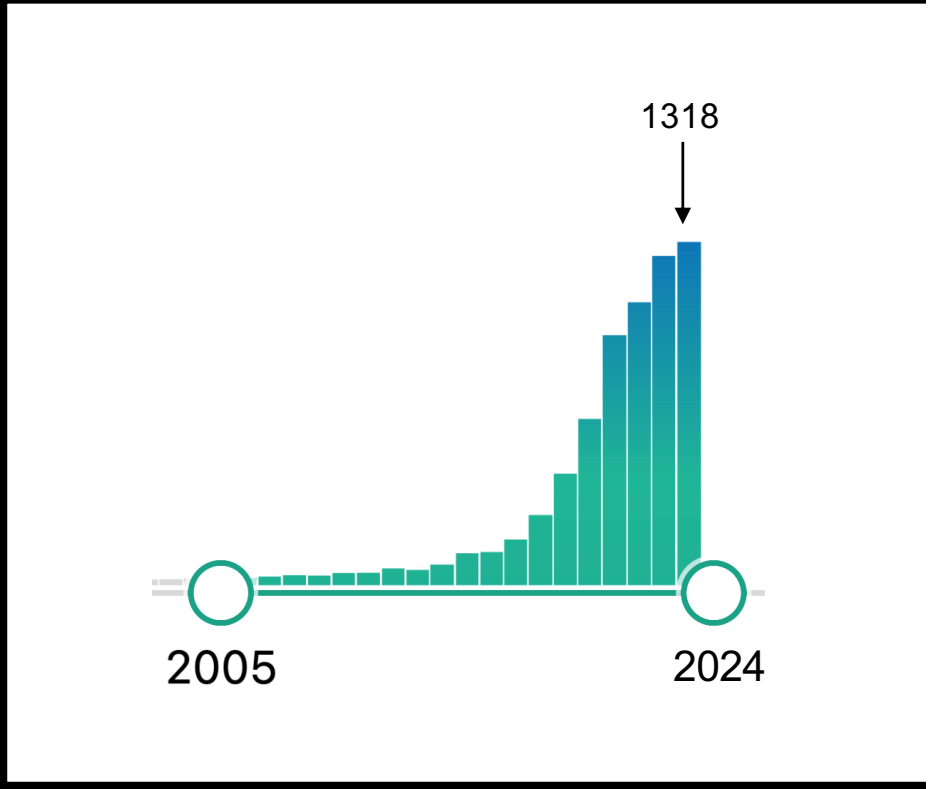
Single molecules
Up to 1 Mbp long

3rd generation technologies

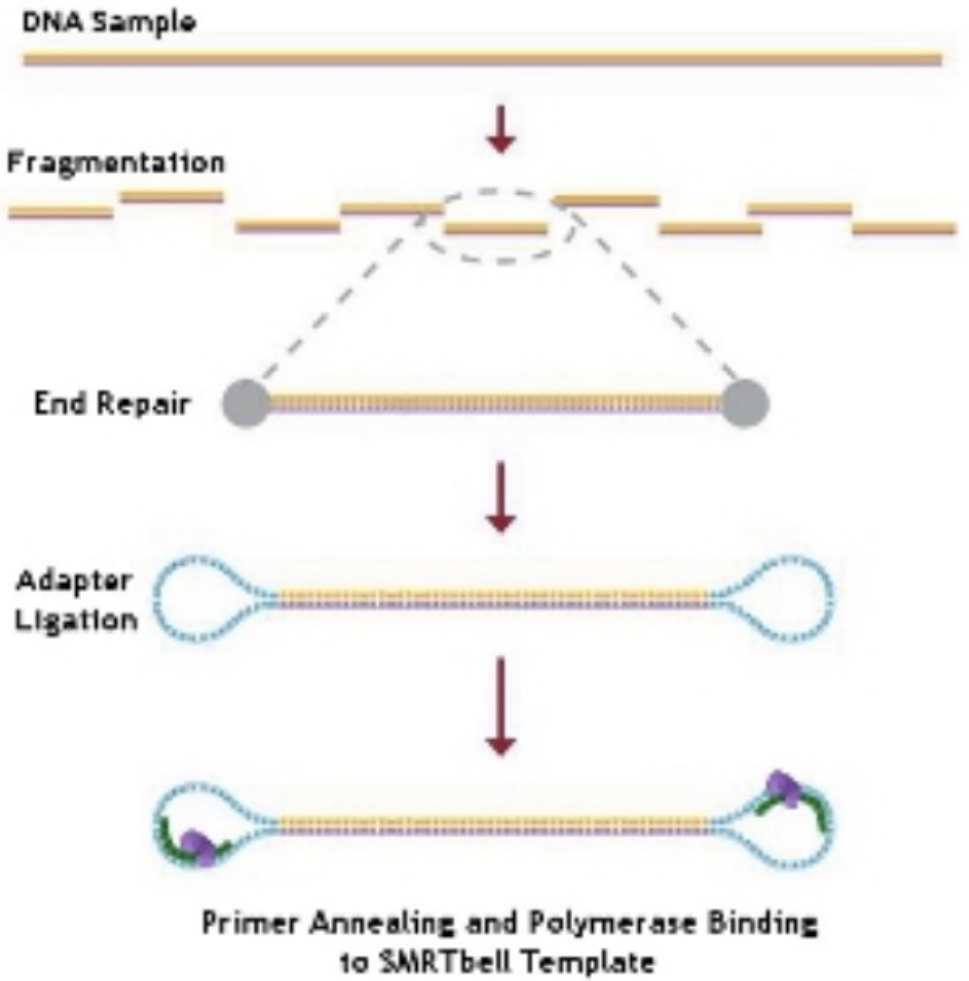
Pacific Biosciences



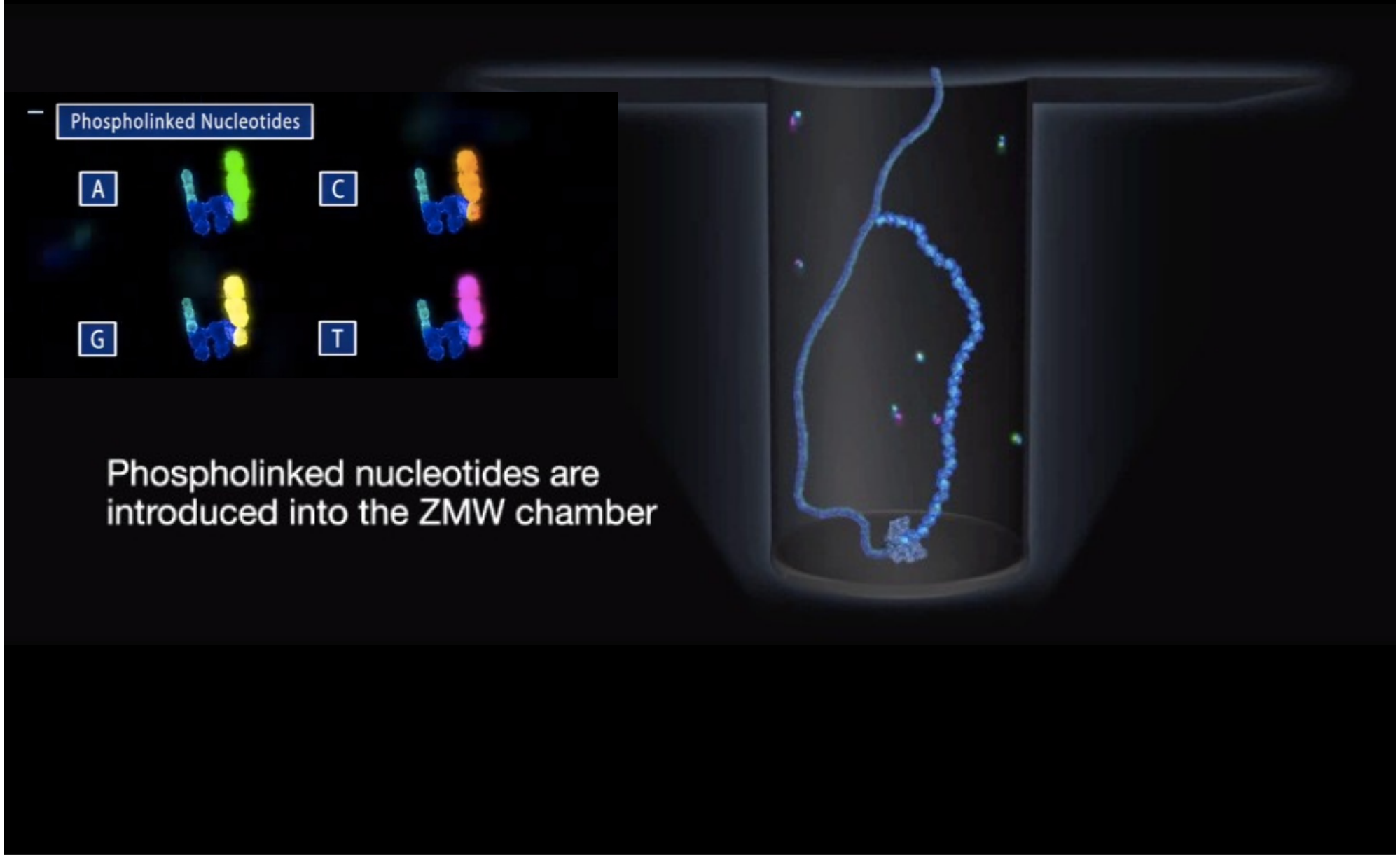
Oxford Nanopore



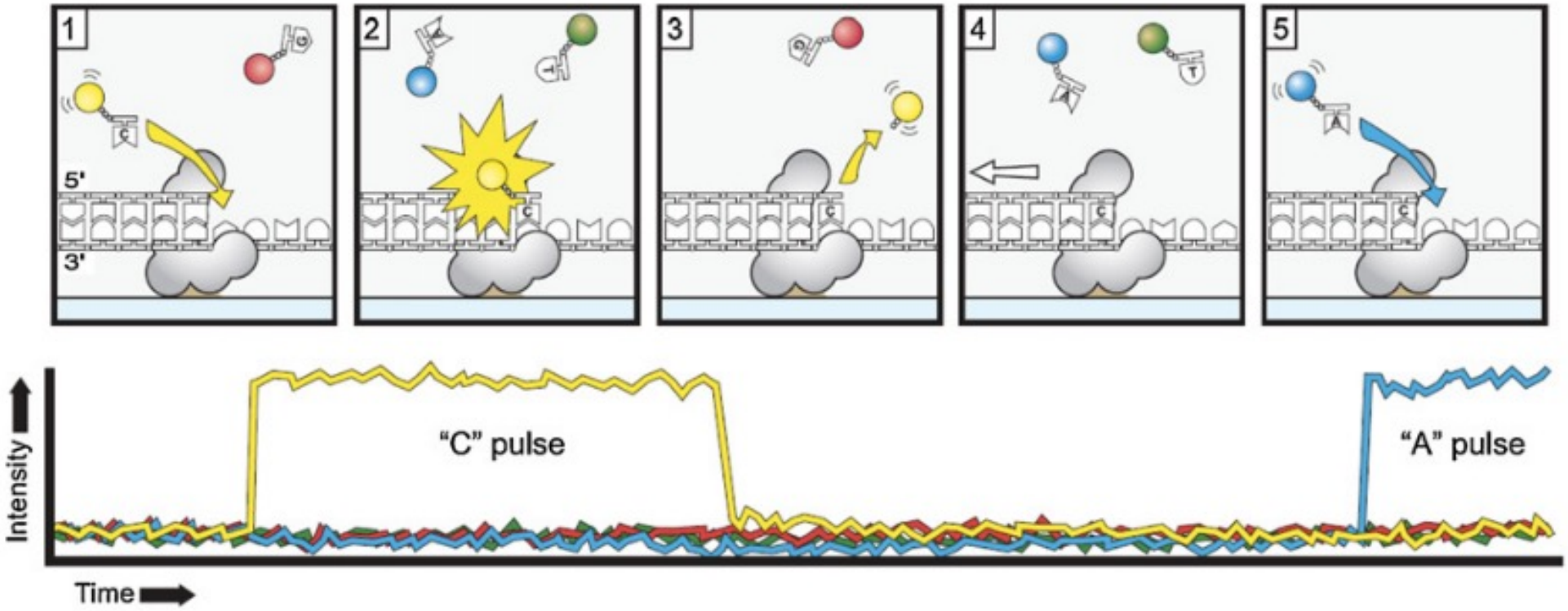
PacBio DNA-seq library



PACIFIC BIOSCIENCES

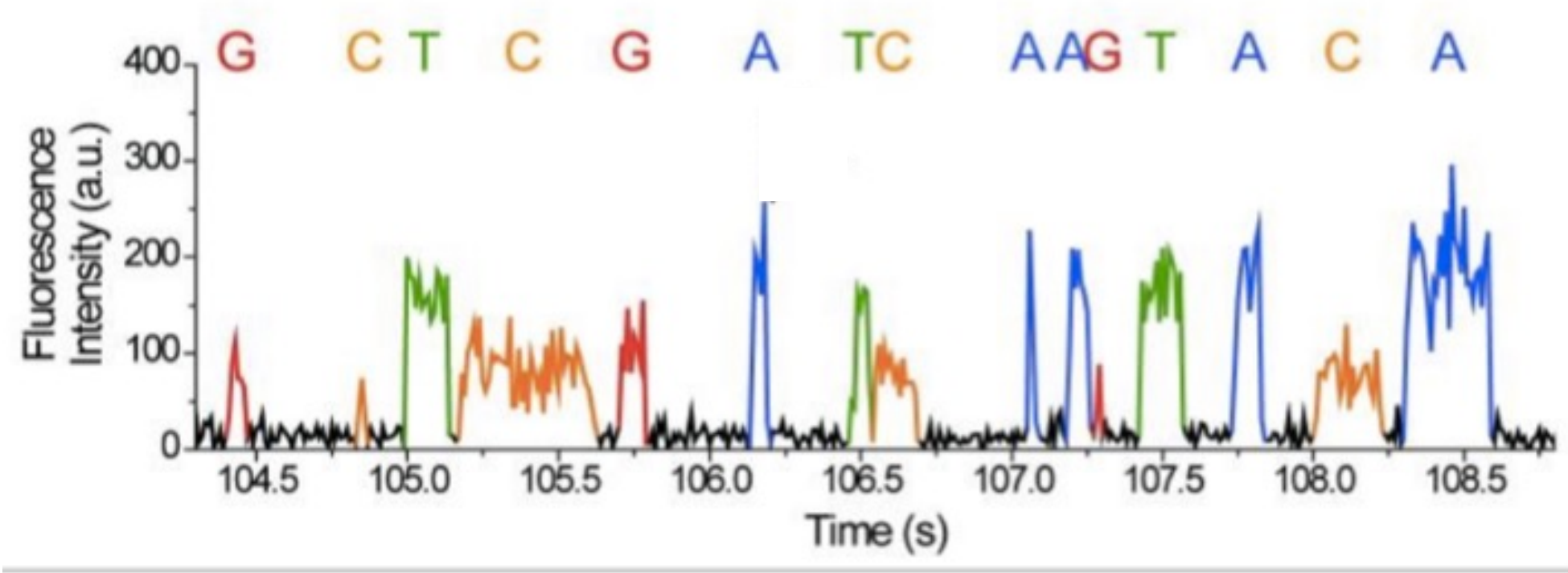


PACIFIC BIOSCIENCES

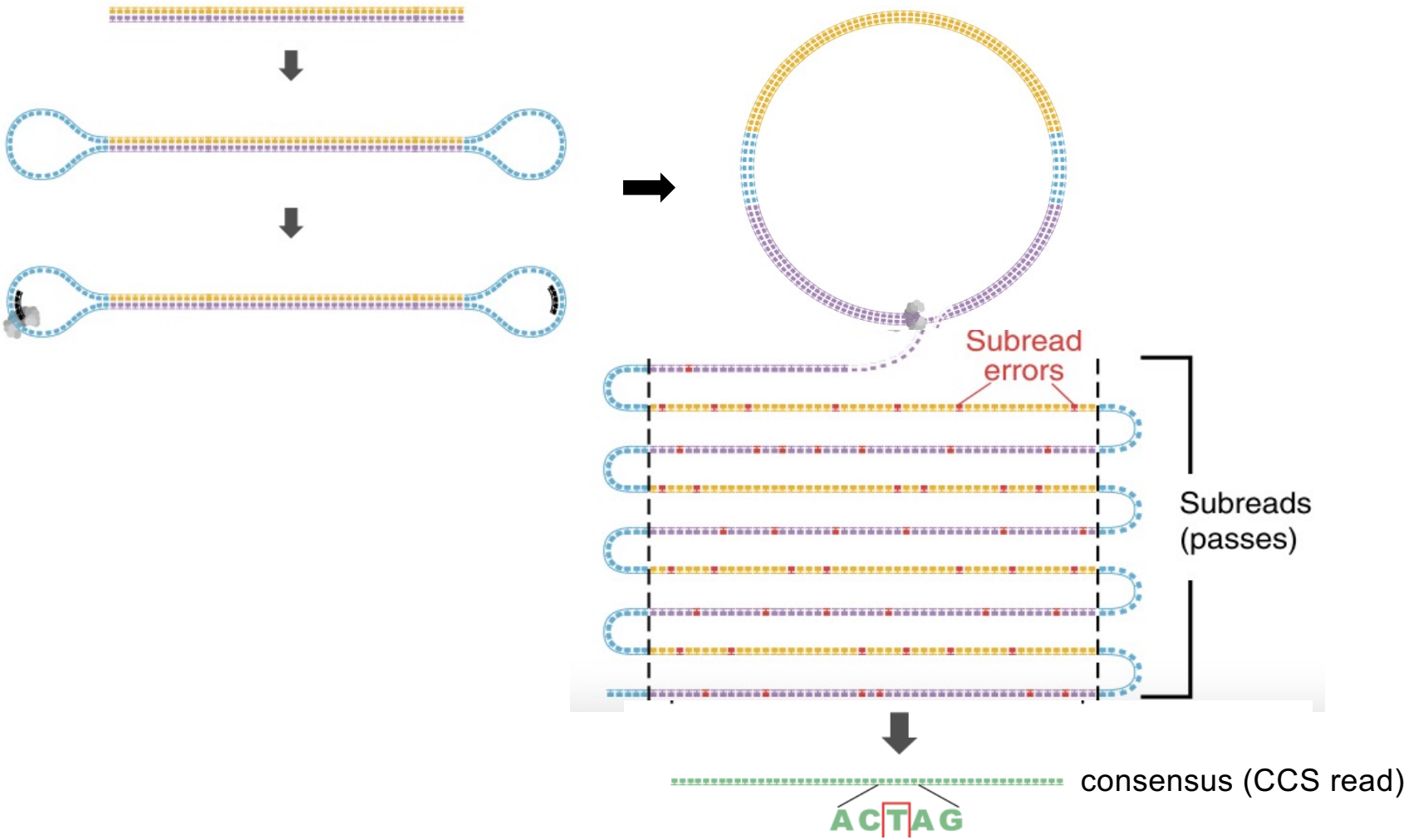


Eid, J., et al. Science (2009)

PACIFIC BIOSCIENCES

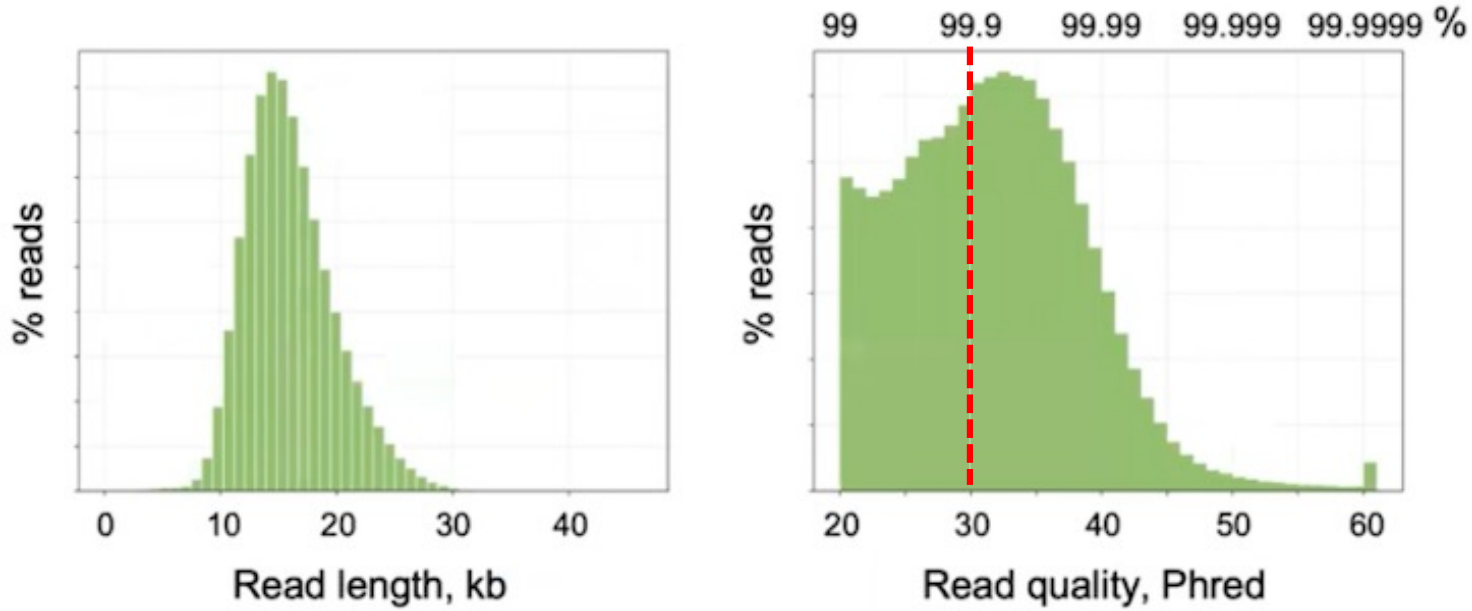


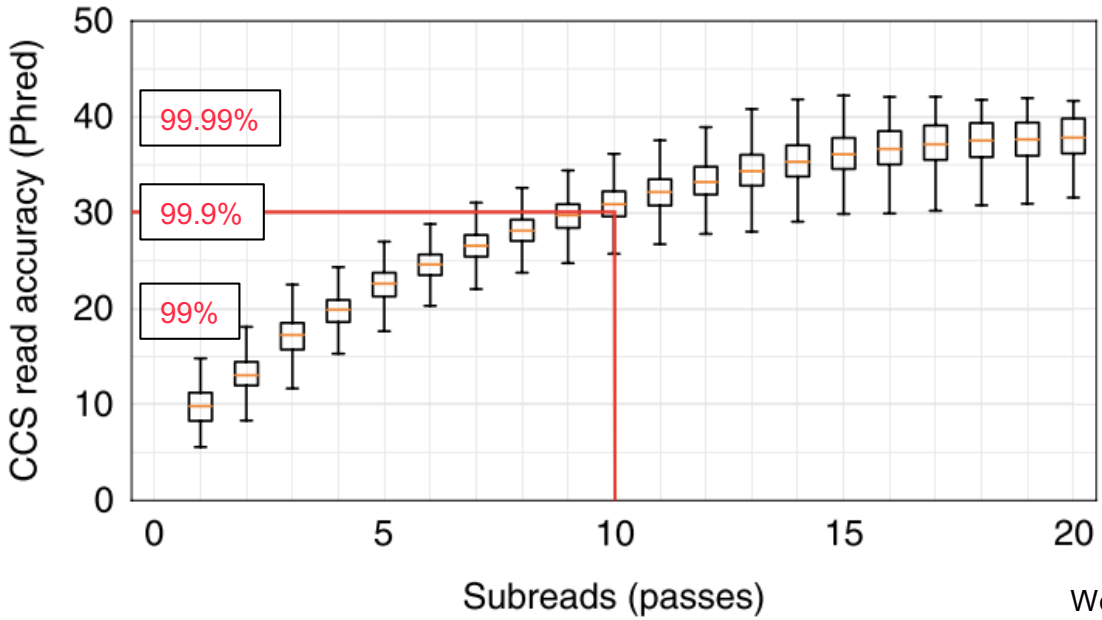
CIRCULAR CONSENSUS SEQUENCES (CCS) : HIFI READS



Randomly positioned errors ➡ they can be corrected

CIRCULAR CONSENSUS SEQUENCES (CCS) : HIFI READS

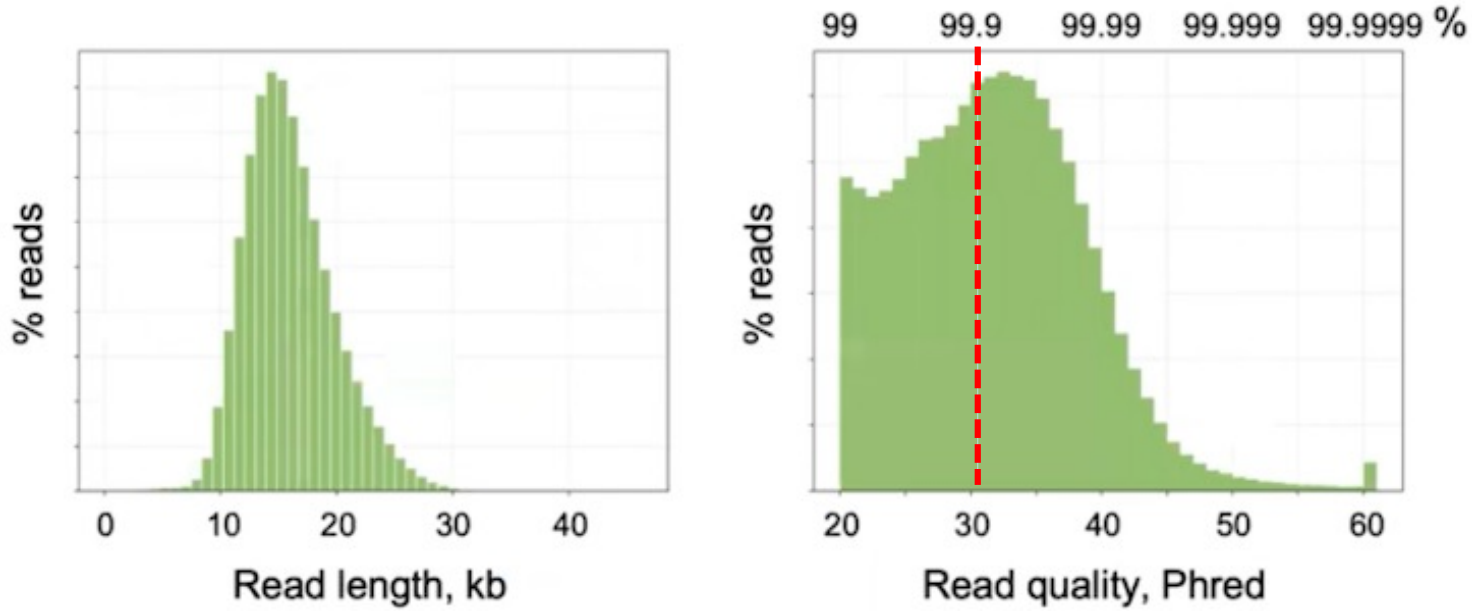




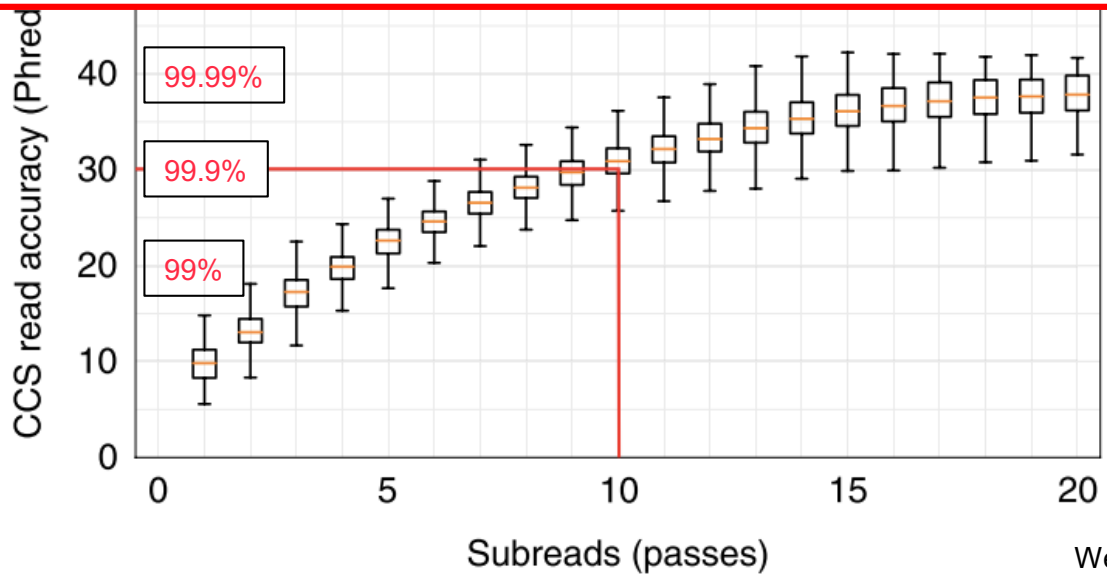
$$Q = -10 \log_{10} P$$

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

CIRCULAR CONSENSUS SEQUENCES (CCS) : HIFI READS



Mean accuracy > 99.9 % (Q > 30 ; < 1 err/kb)



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

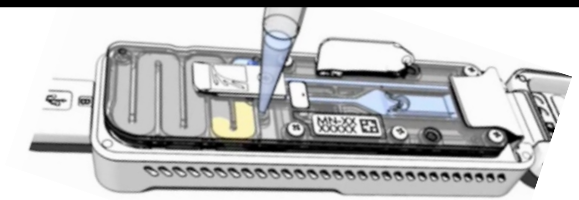
Next Generation Sequencing

Oxford NANOPORE Technologies®

Oxford NANOPORE Technologies

GridION

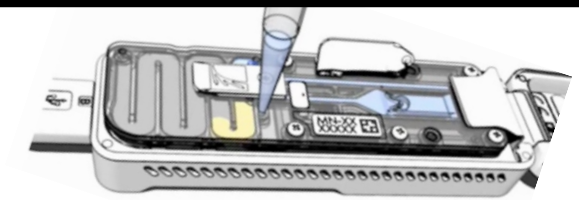
MinION flowcell
Maximum throughput 50 Gb



Oxford NANOPORE Technologies

GridION

MinION flowcell
Maximum throughput 50 Gb



MGI

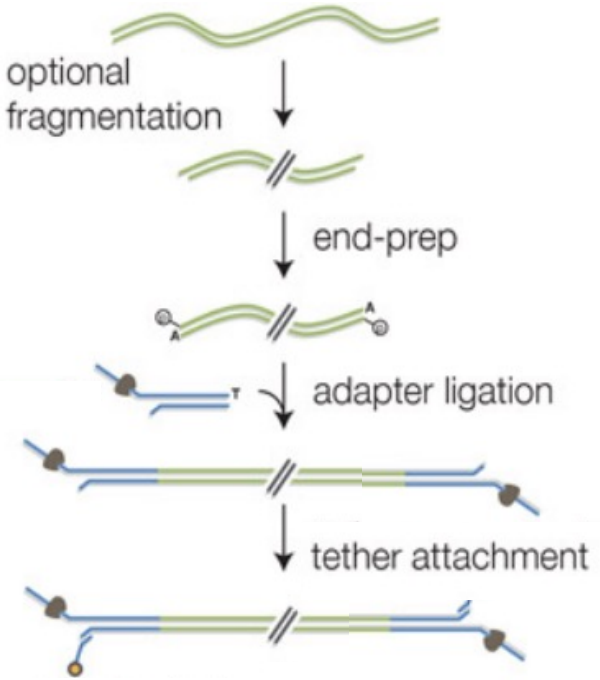
Leading Life Science Innovation

CycloneSEQ-WT02 Nanopore Gene Sequencer

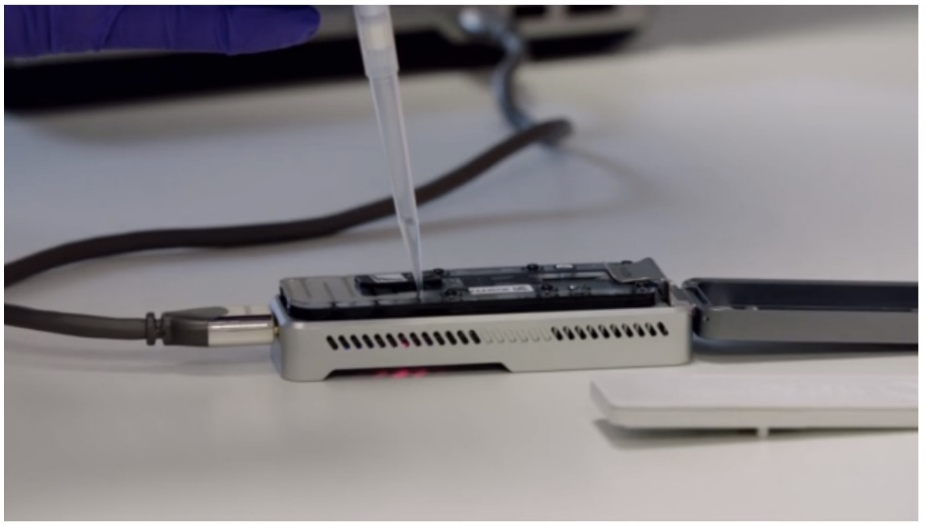
Flowcell
Maximum throughput 50 Gb

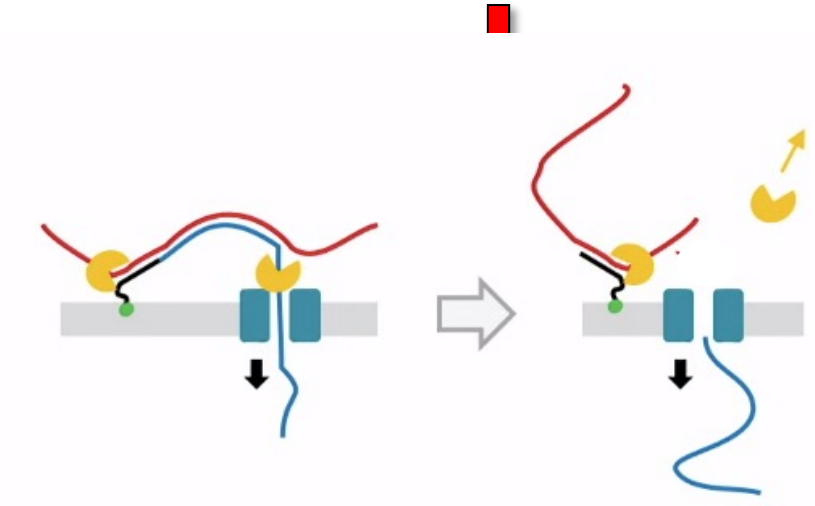
SEQUENCING PROCESS

Library preparation

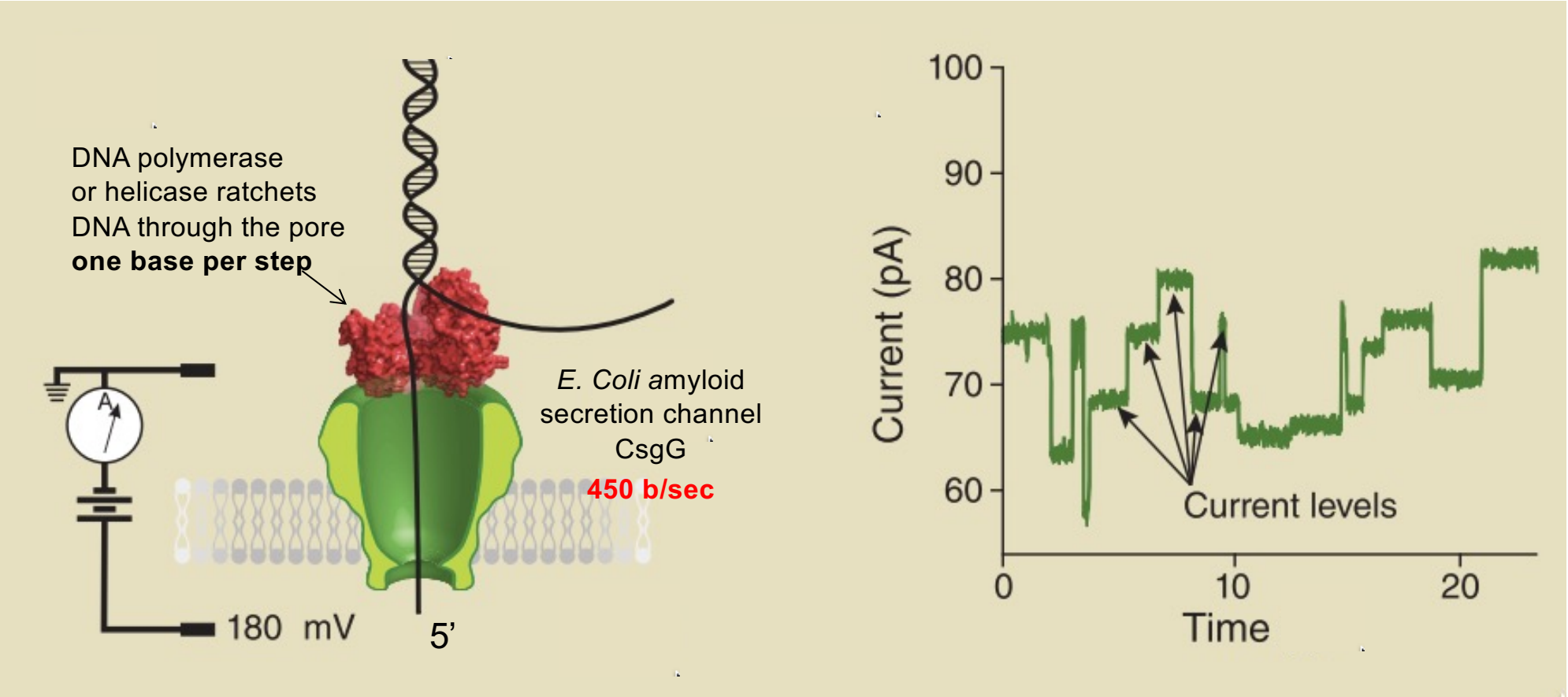


SEQUENCING

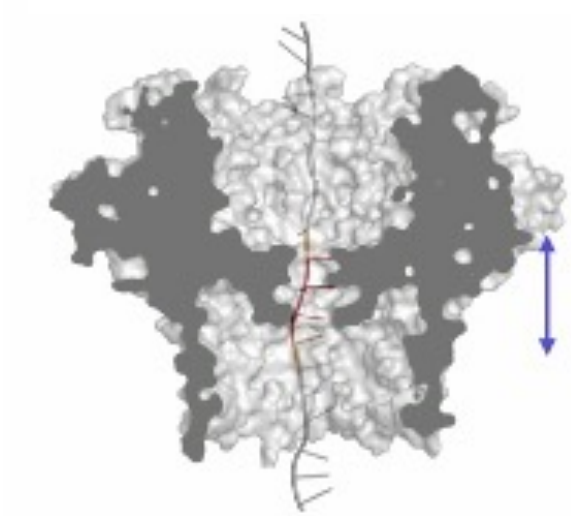




BASIC CONCEPTS

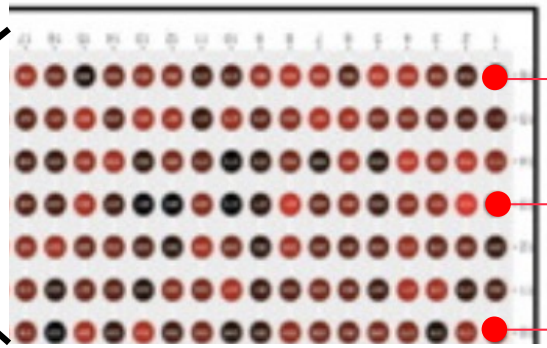


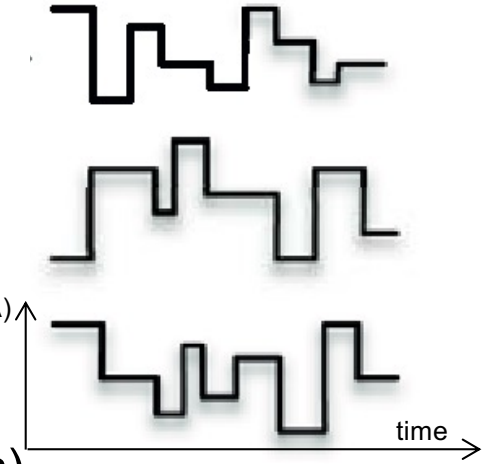
SEQUENCING PROCESS



5-6 bases
dominate the
current signal

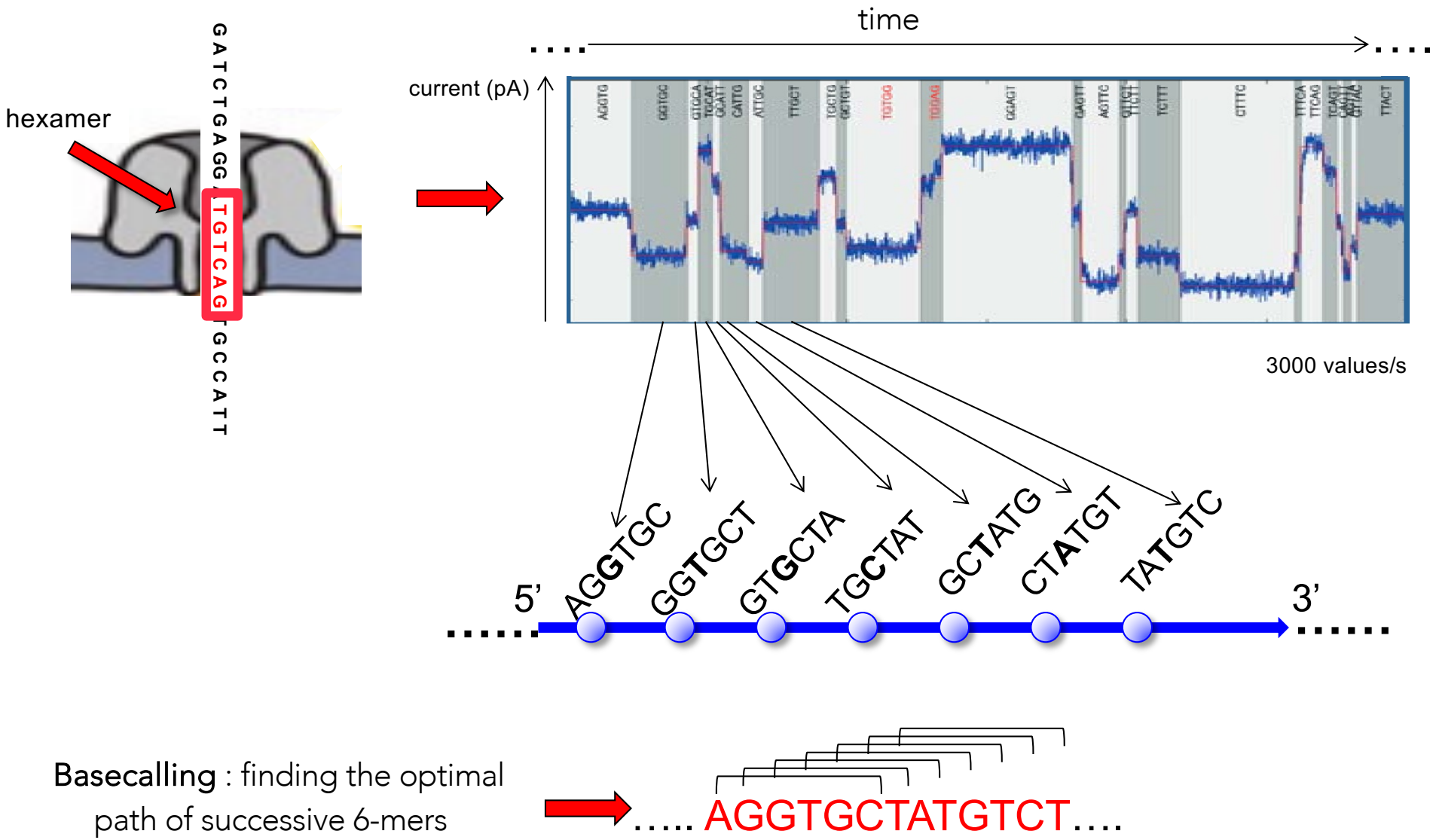
MinION : 1 flow cell → 512 pores





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

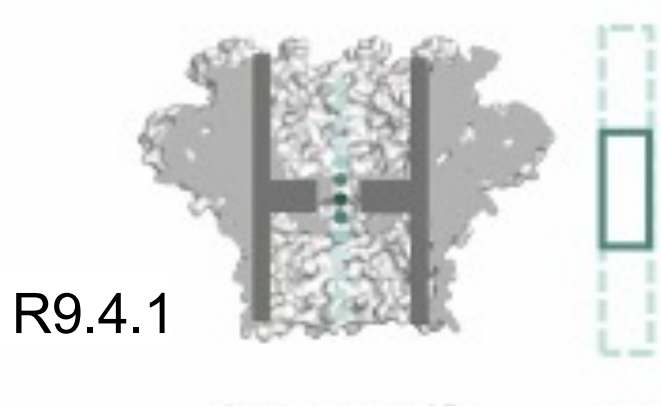
BASE CALLING

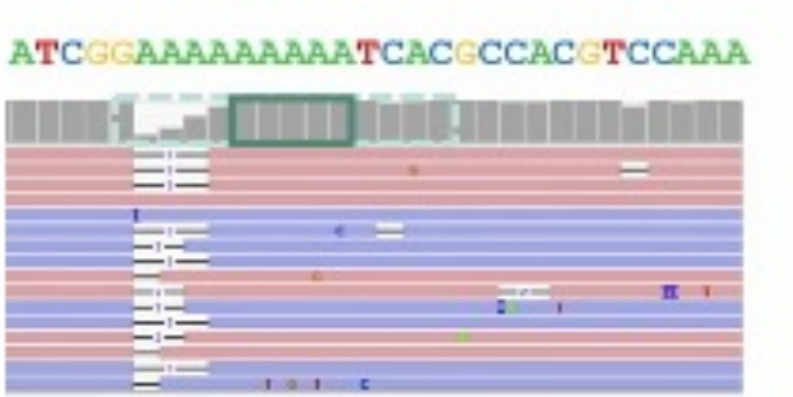


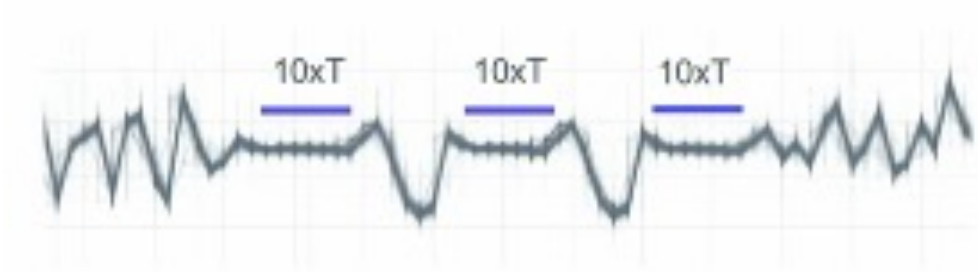
Basecalling : finding the optimal path of successive 6-mers

….. AGGTGCTATGTCT….

“One-reader” pore has difficulty to read homopolymers

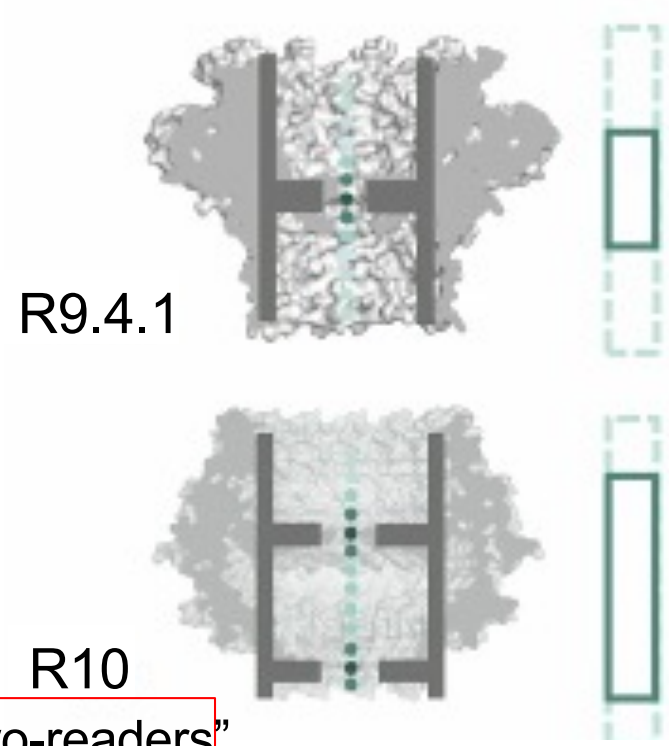






Homopolymers difficult to sequence

”TWO READERS” NANOPORE



2019

“two-readers”

Sereika et al. *Nature Methods,* 2022

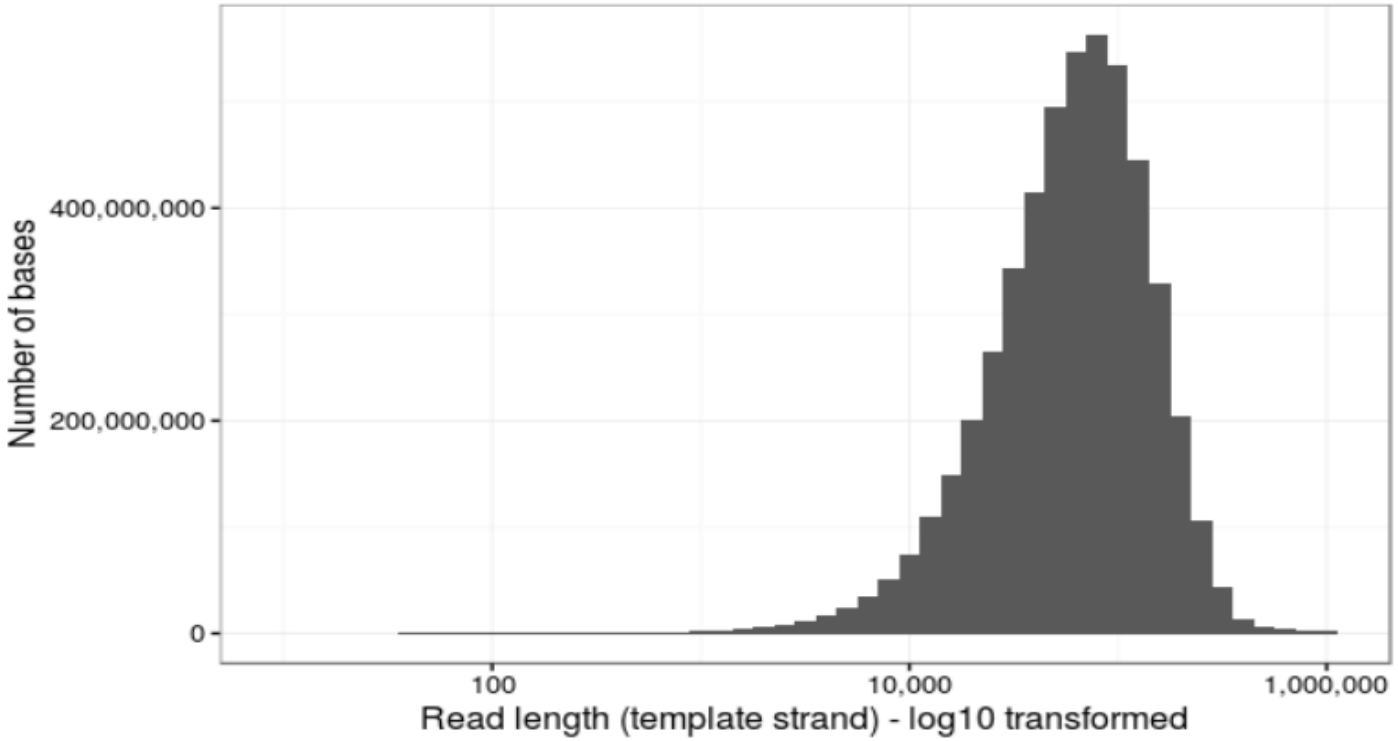
Long homopolymers are better “seen” by the pore and can be decoded with higher accuracy

Mean accuracy (R10) > 99% Q>20 <1% errors

$Q = -10 \log_{10} P$

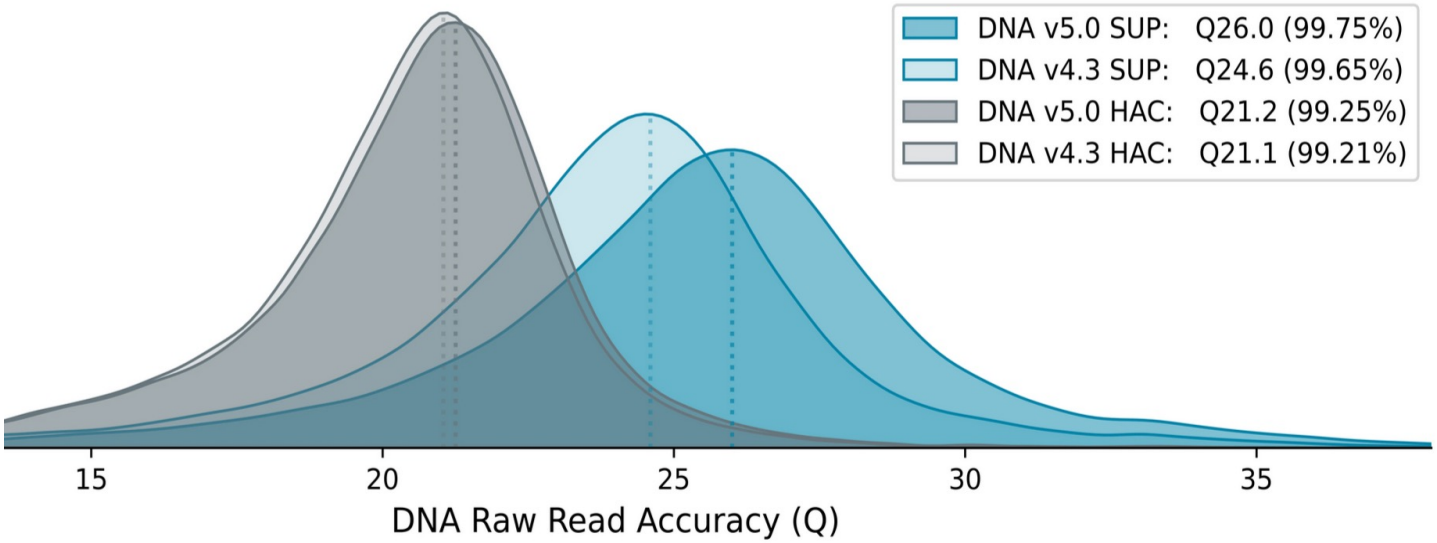
LENGTH OF NANOPORE READS

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



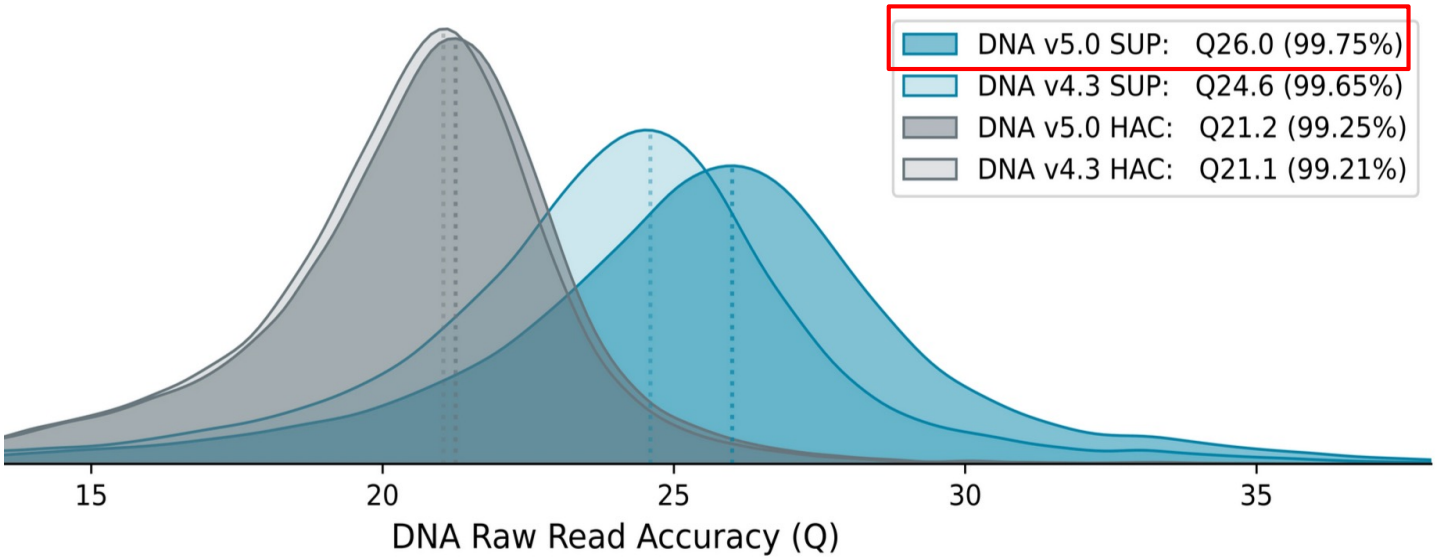
Size of the longest read > 1 Mb

READ ACCURACY



Nanopore website 2024

READ ACCURACY



Nanopore website 2024

PacBio

SEQUENCING SYSTEMS

WHOLE GENOME SEQUENCING — HOW PACBIO COMPARES

	PacBio HiFi	Illumina	Oxford Nanopore
Average read length[1]	15–20 kb	2 x 150 bp	10–100 kb
Average read accuracy[1]	99.95% (Q33)	99.92% (Q31)	99.26% (Q21)
Coverage[2]	0.5 error/kb	0.8 error/kb	8 error/kb
Variant calling: SNVs	✓	✓	✓
Variant calling: indels	✓	✓	✗
Variant calling: SVs	✓	✗	✓
Genome assembly: contiguity	✓	✗	✓
Genome assembly: accuracy	✓	✓	✗
Epigenetics: 5mC	✓	✗	✓

1. PacBio HiFi: HG003 18 kb library, Sequel II system chemistry 2.0, precisionFDA *Truth Challenge* V2 (https://doi.org/10.1101/2020.11.13.380741), Illumina: HG002 2×150 bp NovaSeq library, precisionFDA *Truth Challenge* V2 (https://doi.org/10.1101/2020.11.13.380741), ONT: Q20+ chemistry (R10.4, Kit 12), Oct 2021 GM24385 Q20+ Simplex Dataset Release (https://labs.epi2me.io/gm24385_q20_2021.10/)

2. HiFi+ONT: Nurk 2021 https://doi.org/10.1101/2021.05.26.445798, HiFi+Illumina: Logsdon 2020 https://doi.org/10.1038/s41576-020-0236-x, ONT: Tan 2022 https://doi.org/10.1101/2022.01.11.475254

PacBio website nov. 2024

PacBio

SEQUENCING SYSTEMS

WHOLE GENOME SEQUENCING — HOW PACBIO COMPARES

	PacBio HiFi	Illumina	Oxford Nanopore
Average read length[1]	15–20 kb	2 x 150 bp	10–100 kb
Average read accuracy[1]	99.95% (Q33)	99.92% (Q31)	99.75% (Q26)
Coverage[2]	0.5 error/kb	0.8 error/kb	2.5 error/kb
Variant calling: SNVs	✓	✓	✓
Variant calling: indels	✓	✓	✗
Variant calling: SVs	✓	✗	✓
Genome assembly: contiguity	✓	✗	✓
Genome assembly: accuracy	✓	✓	✗
Epigenetics: 5mC	✓	✗	✓

1. PacBio HiFi: HG003 18 kb library, Sequel II system chemistry 2.0, precisionFDA *Truth Challenge* V2 (https://doi.org/10.1101/2020.11.13.380741), Illumina: HG002 2×150 bp NovaSeq library, precisionFDA *Truth Challenge* V2 (https://doi.org/10.1101/2020.11.13.380741), ONT: Q20+ chemistry (R10.4, Kit 12), Oct 2021 GM24385 Q20+ Simplex Dataset Release (https://labs.epi2me.io/gm24385_q20_2021.10/)

2. HiFi+ONT: Nurk 2021 https://doi.org/10.1101/2021.05.26.445798, HiFi+Illumina: Logsdon 2020 https://doi.org/10.1038/s41576-020-0236-x, ONT: Tan 2022 https://doi.org/10.1101/2022.01.11.475254

PacBio website nov. 2024

GENOME ASSEMBLY

SMALL GENOMES

PacBio vs Nanopore

Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing

Sereika et al. *Nature Methods*, 2022

Samples :

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

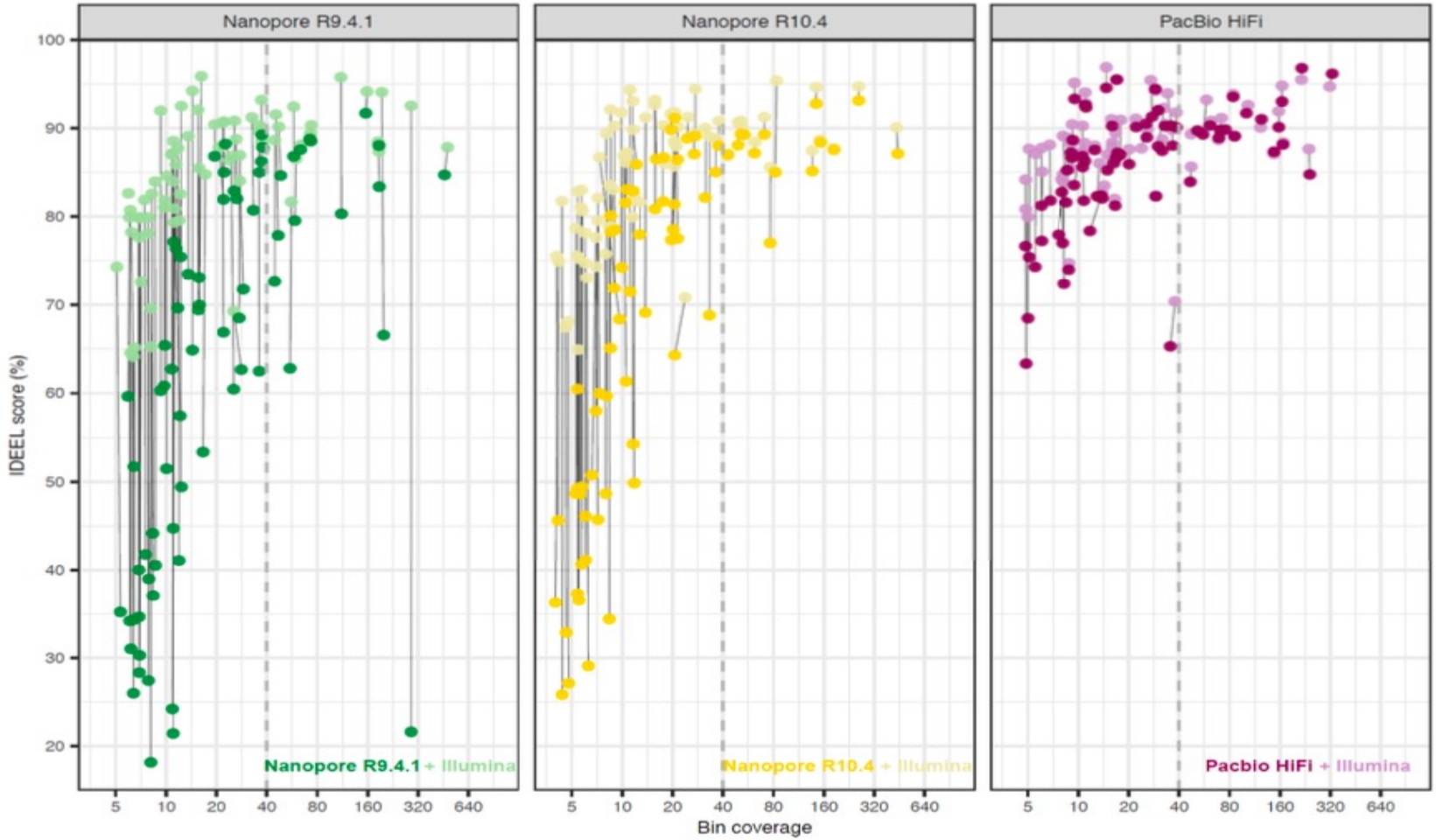
Sequenced with :

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

Read processing

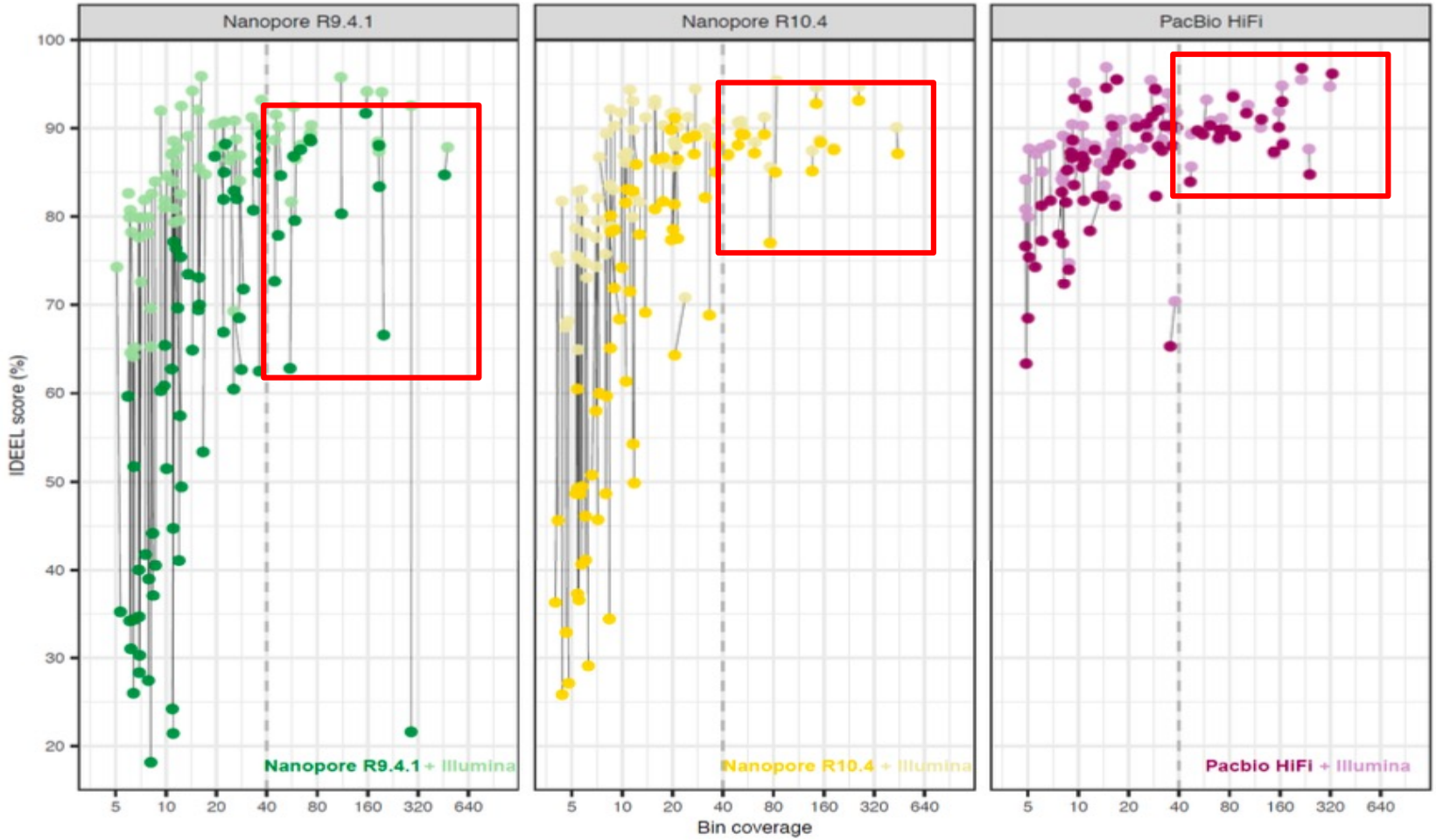
- reads assembled with Flye

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



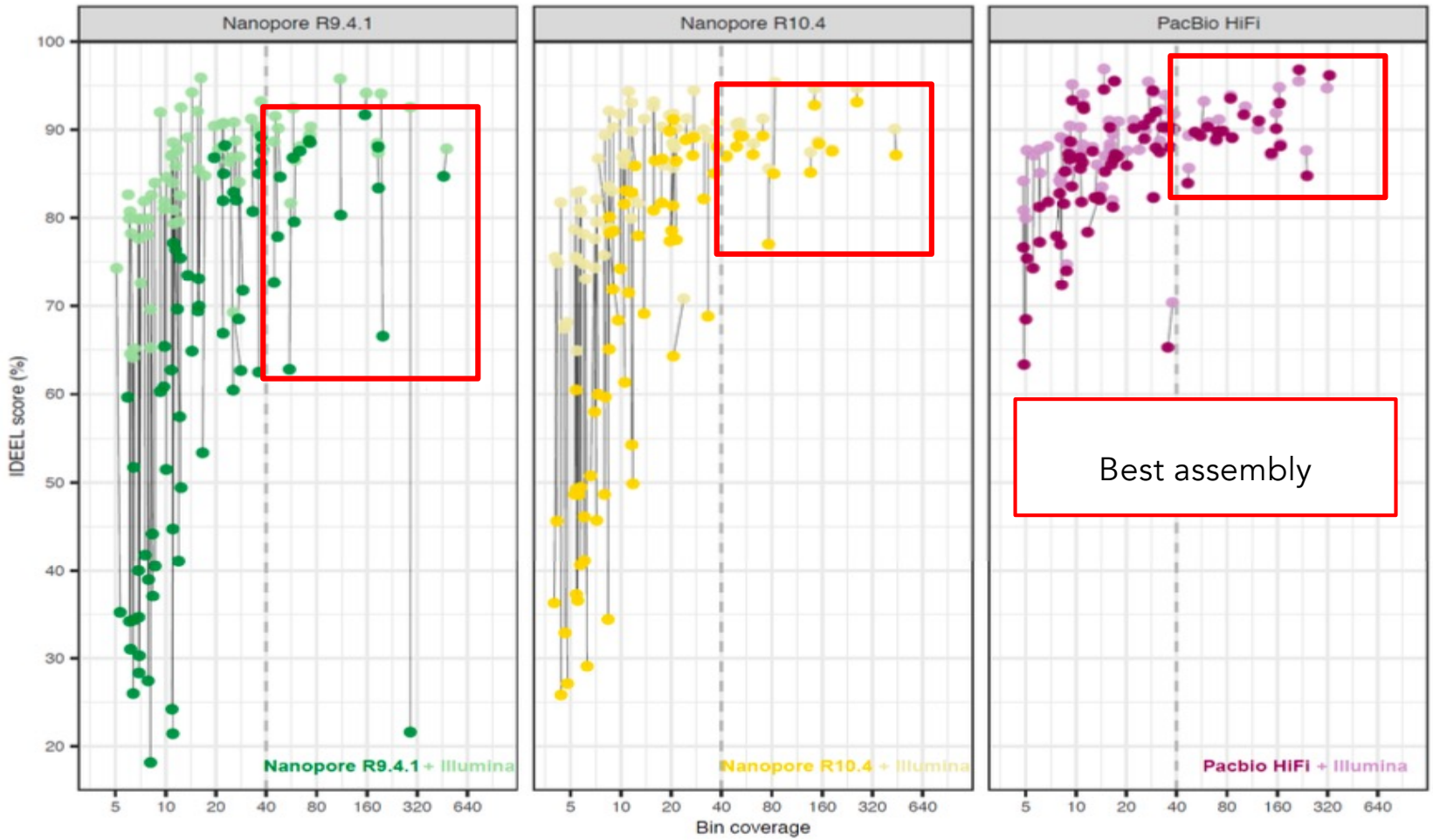
IDEEL score : proportion of predicted proteins that are ≥95% the length of their best-matching known protein in a database

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



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

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



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

Conclusions

- **HiFi reads** : best microbial genome assembly
- **Nanopore** : "Near-finished assembly" of microbial genomes with R10.4 data alone

Nanopore only

Benchmarking reveals superiority of deep learning variant callers on bacterial nanopore sequence data

Hall et al. *eLife* oct. 2024

Objectives

Create a benchmark of the performance of seven variant callers using ONT and Illumina sequencing data :

- deepvariant, clair3, bcftools, freebayes, longshot, medaka, nanocaller
- Creation of a variant truthset for benchmarking
- Analysis of 14 samples from different bacterial species spanning a wide range of GC content (30–66%)
- ONT data basecalled with three different accuracy models – fast, high accuracy (hac), super-accuracy (sup)

Nanopore only

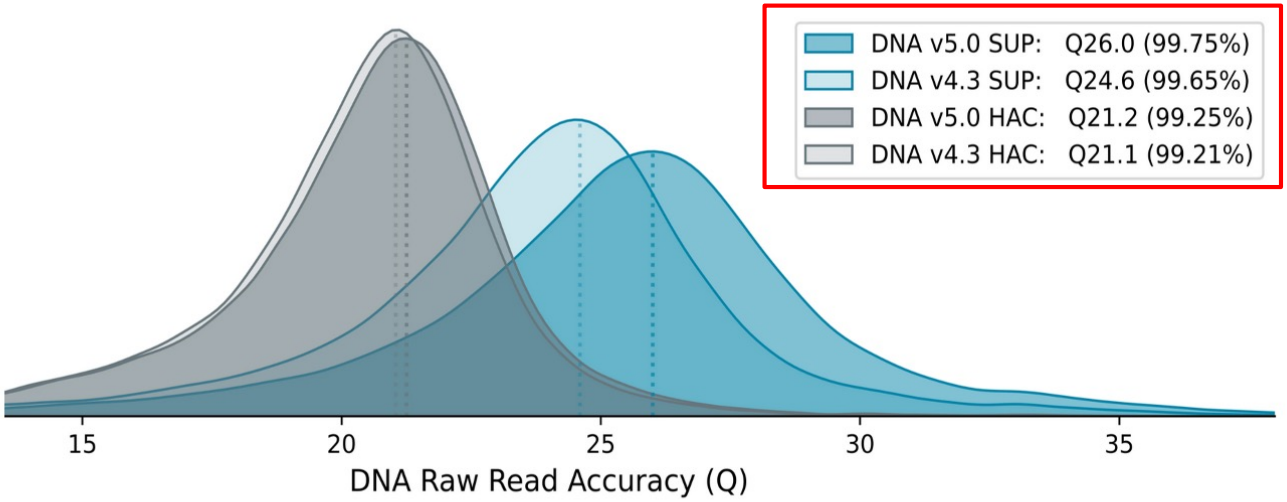
Benchmarking reveals superiority of deep learning variant callers on bacterial nanopore sequence data

Hall et al. *eLife* oct. 2024

Objectives

Create a benchmark of the performance of seven variant callers using ONT and Illumina sequencing data :

- deepvariant, clair3, bcftools, freebayes, longshot, medaka, nanocaller
- Creation of a variant truthset for benchmarking
- Analysis of 14 samples from different bacterial species spanning a wide range of GC content (30–66%)
- ONT data basecalled with three different accuracy models – fast, high accuracy (hac), super-accuracy (sup)



Nanopore only

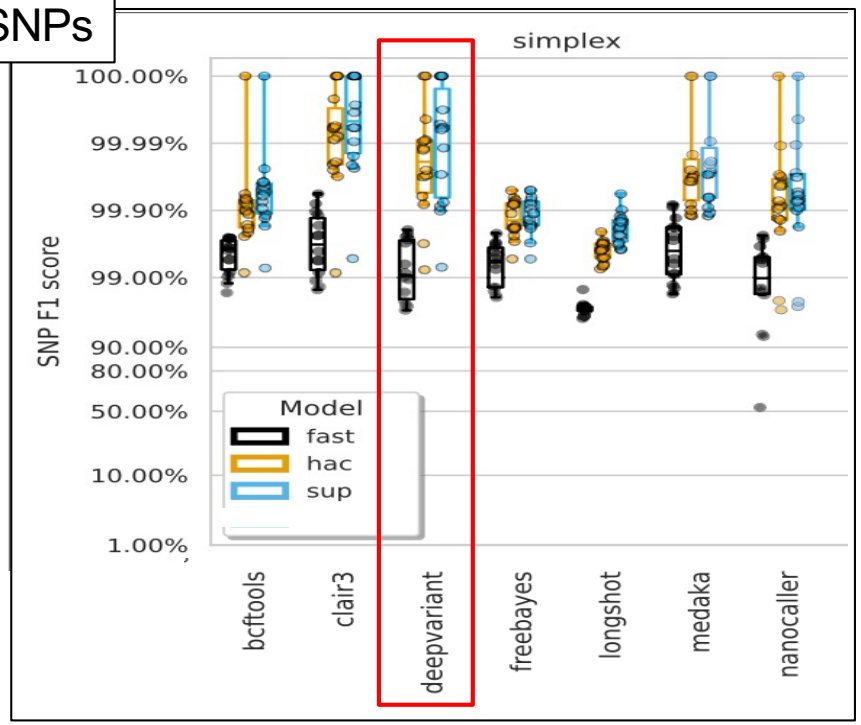
Benchmarking reveals superiority of deep learning variant callers on bacterial nanopore sequence data
Hall et al. *eLife* oct. 2024

Objectives

Create a benchmark of the performance of seven variant callers using ONT and Illumina sequencing data :

- deepvariant, clair3, bcftools, freebayes, longshot, medaka, nanocaller
- Creation of a variant truthset for benchmarking
- Analysis of 14 samples from different bacterial species spanning a wide range of GC content (30–66%)
- ONT data basecalled with three different accuracy models – fast, high accuracy (hac), super- accuracy (sup)

SNPs



F1 = 2TP/(2TP+FP+FN) = harmonic mean of precision and recall

Nanopore only

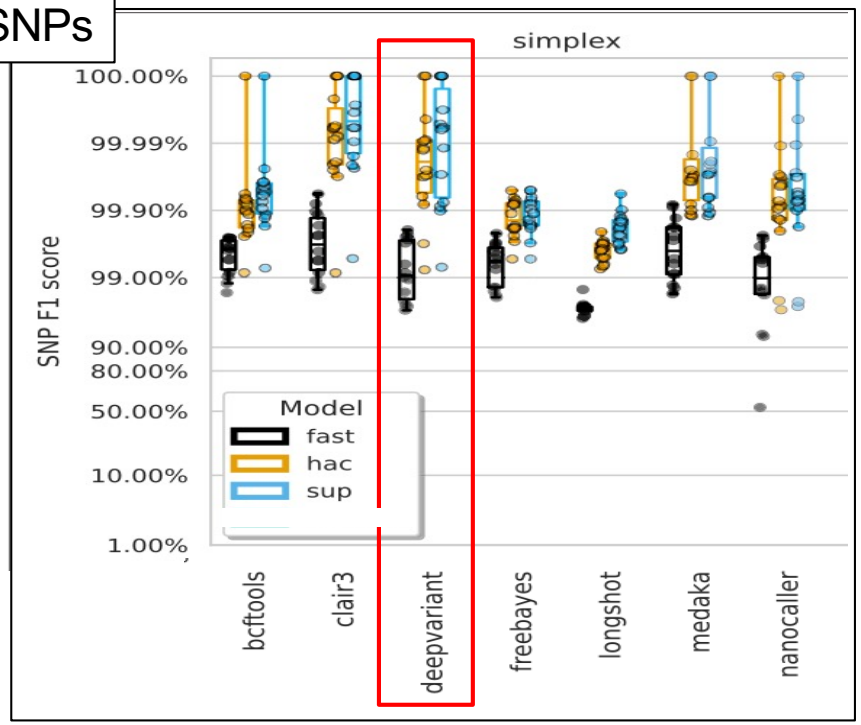
Benchmarking reveals superiority of deep learning variant callers on bacterial nanopore sequence data
Hall et al. *eLife* oct. 2024

Objectives

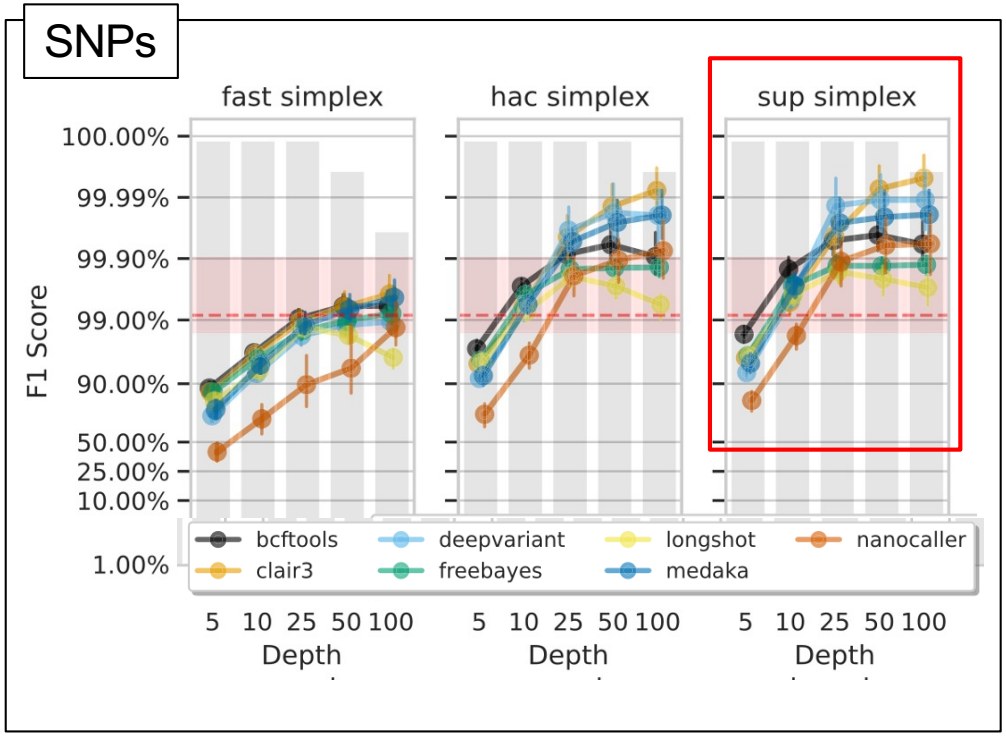
Create a benchmark of the performance of seven variant callers using ONT and Illumina sequencing data :

- deepvariant, clair3, bcftools, freebayes, longshot, medaka, nanocaller
- Creation of a variant truthset for benchmarking
- Analysis of 14 samples from different bacterial species spanning a wide range of GC content (30–66%)
- ONT data basecalled with three different accuracy models – fast, high accuracy (hac), super- accuracy (sup)

SNPs



SNPs



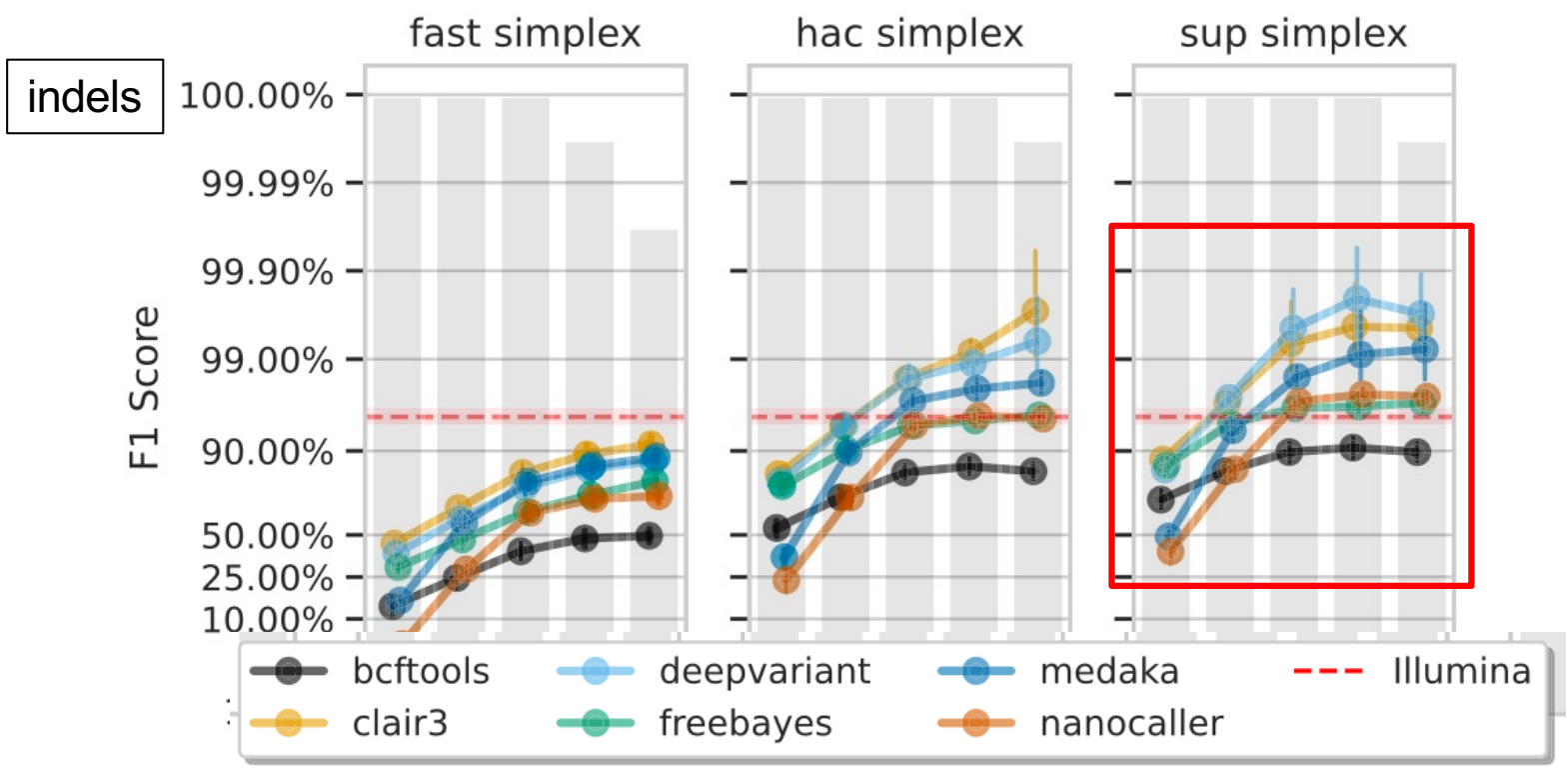
Nanopore only

Benchmarking reveals superiority of deep learning variant callers on bacterial nanopore sequence data
Hall et al. *eLife* oct. 2024

Objectives

Create a benchmark of the performance of seven variant callers using ONT and Illumina sequencing data :

- deepvariant, clair3, bcftools, freebayes, longshot, medaka, nanocaller
- Creation of a variant truthset for benchmarking
- Analysis of 14 samples from different bacterial species spanning a wide range of GC content (30–66%)
- ONT data basecalled with three different accuracy models – fast, high accuracy (hac), super- accuracy (sup)



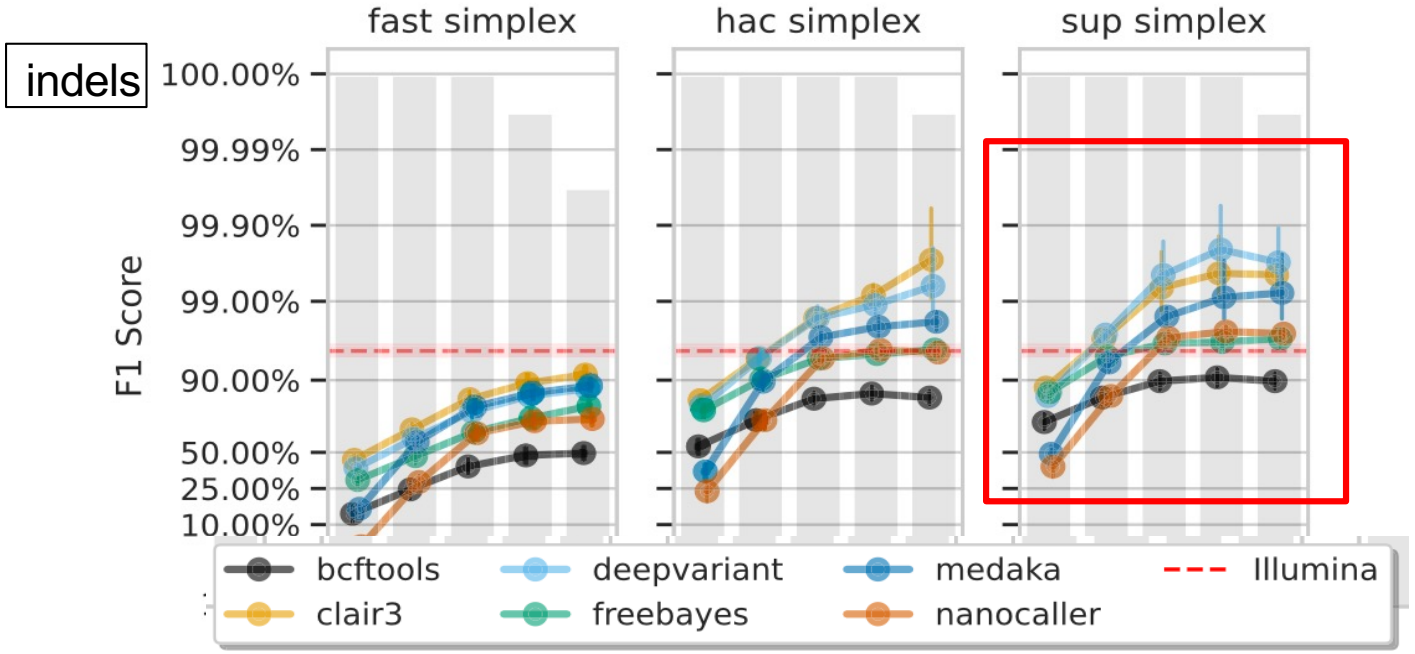
Nanopore only

Benchmarking reveals superiority of deep learning variant callers on bacterial nanopore sequence data
Hall et al. *eLife* oct. 2024

Objectives

Create a benchmark of the performance of seven variant callers using ONT and Illumina sequencing data :

- deepvariant, clair3, bcftools, freebayes, longshot, medaka, nanocaller
- Creation of a variant truthset for benchmarking
- Analysis of 14 samples from different bacterial species spanning a wide range of GC content (30–66%)
- ONT data basecalled with three different accuracy models – fast, high accuracy (hac), super- accuracy (sup)



Clair and Deepvariant → real improvements of Nanopore read accuracy (sup models)
→ SNP F1 score ≈ 99.99%
→ Indel F1 score ≈ 99.5%

GENOME ASSEMBLY

LARGE GENOMES

PacBio only

Long-read genome sequencing and variant reanalysis increase diagnostic yield in neurodevelopmental disorders
Hiatt et al., *Genome Research,* oct. 2024

Cohort :

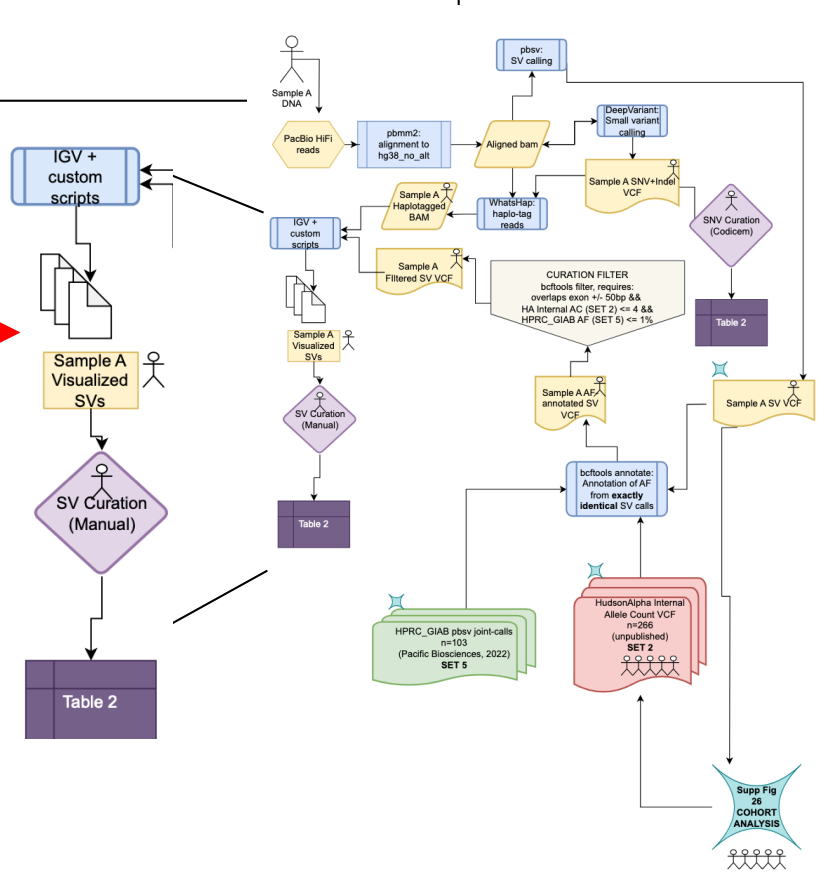
- 96 cases children with rare diseases : suspected to present genetic neurodevelopmental disorders (NDDs)
- Illumina sequencing ⇔ no pathogenic variant

- PacBio HiFi sequencing (median: 27X)
- Genome assembly, phasing
- Variant calling (DeepVariant, pbsv)

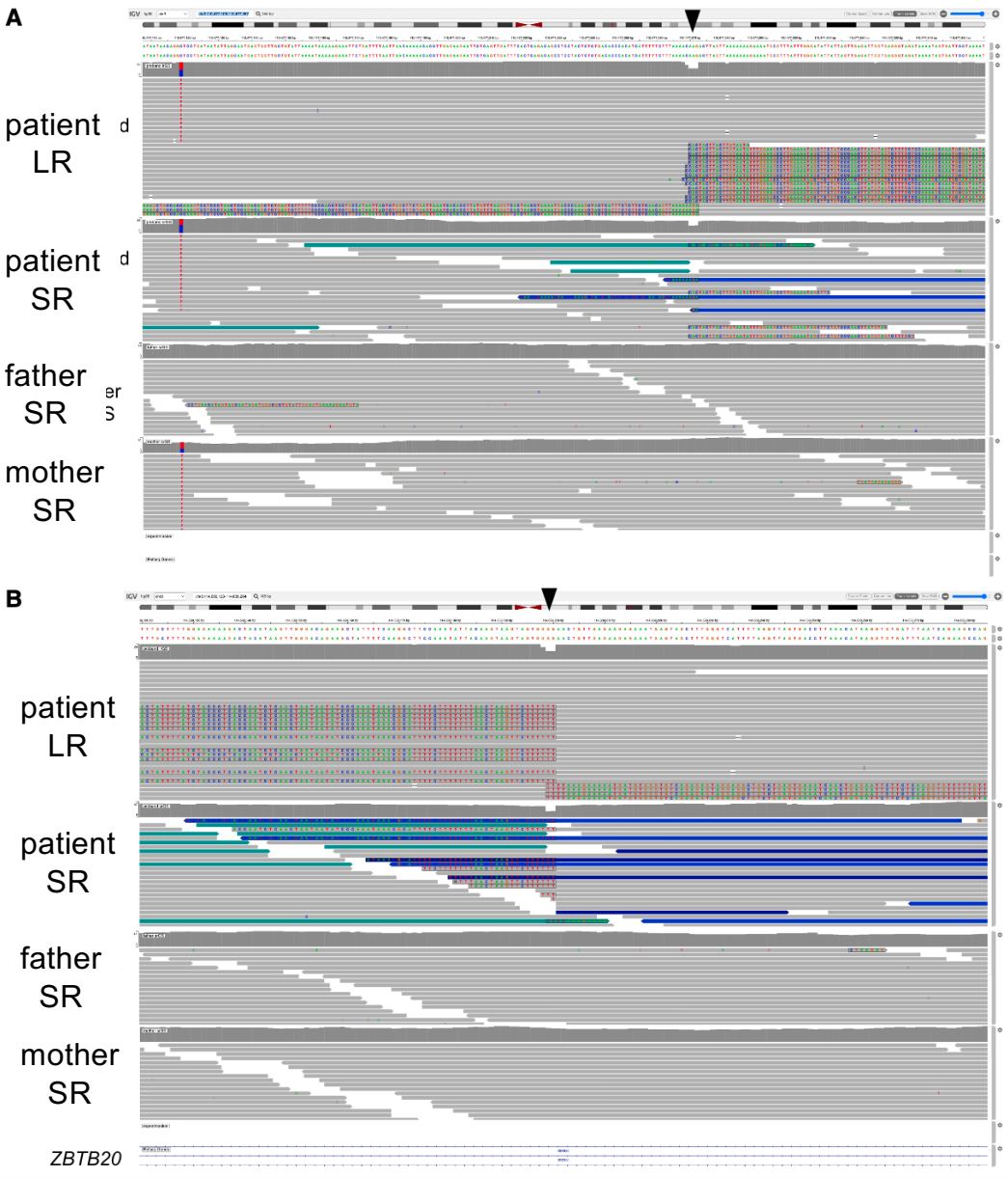
  → median of 25 000 structural variants (>50 nt) per genome

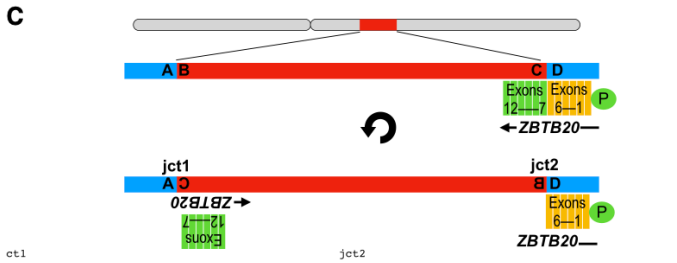
- annotation, filtering and curation pipeline combining multiple variant databases

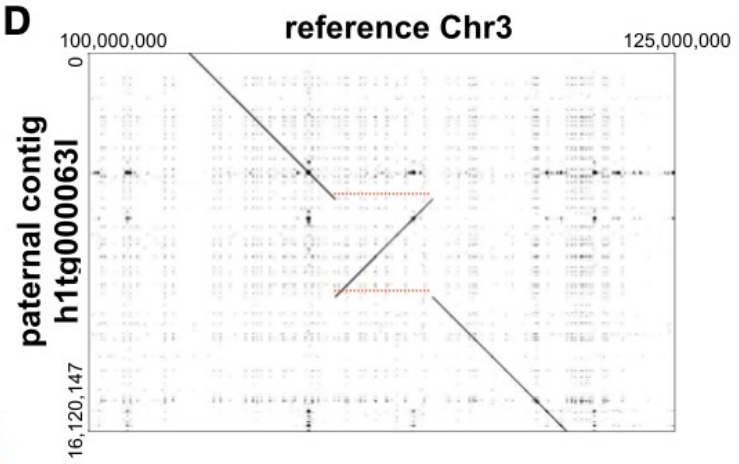
16 “clinically relevant” new variants



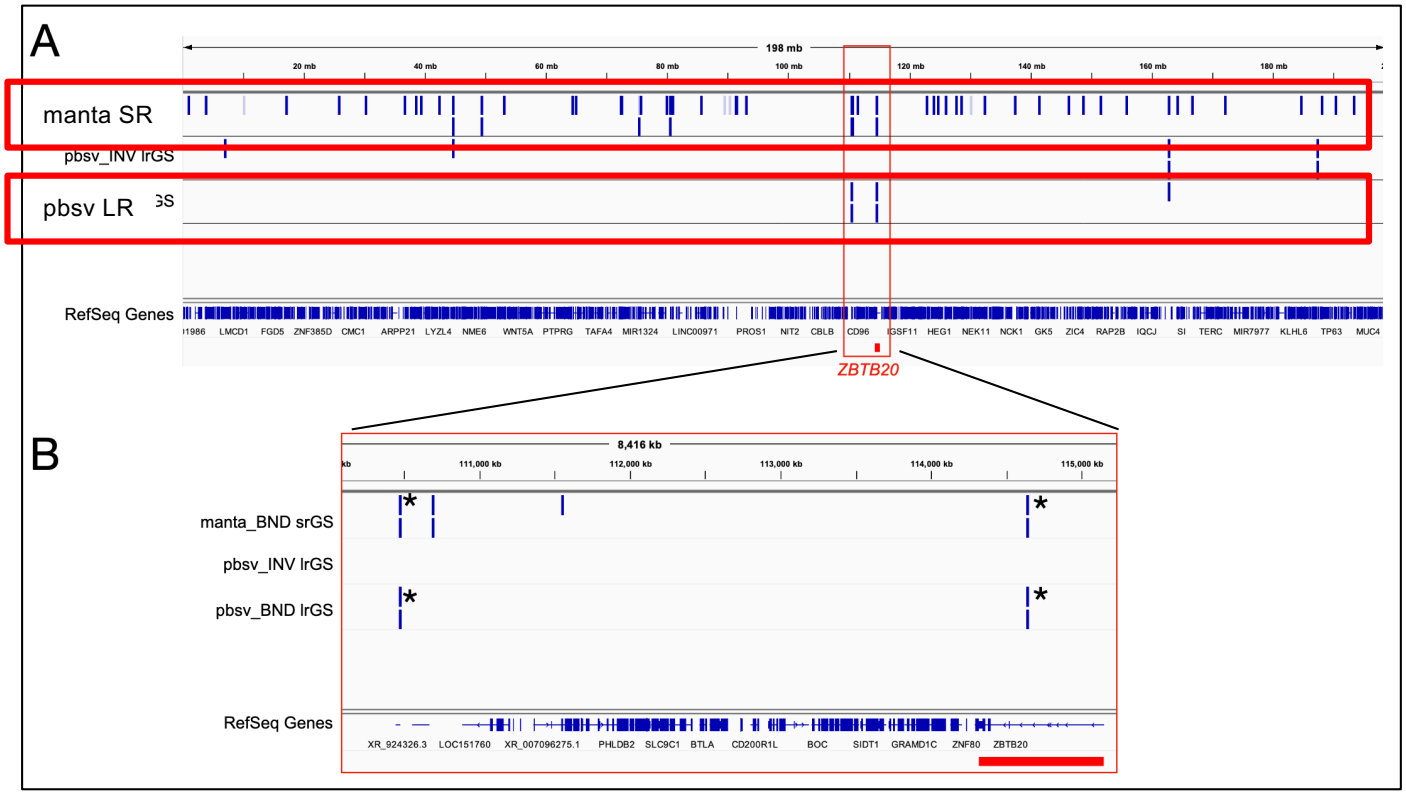
One example of variant







Loss-of-function for ZBTB20
Gene associated with “Primrose Syndrome”



CONCLUSIONS

Difficulties with long-reads :

- Discriminate genuine variants from background benign alleles depends on their annotation in databases
- But main variant databases are built from short-read data
- Strength of long-reads : they see variants invisible for short reads ---> difficult to filter benign alleles among long-read-only variants

Advantages of long-reads :

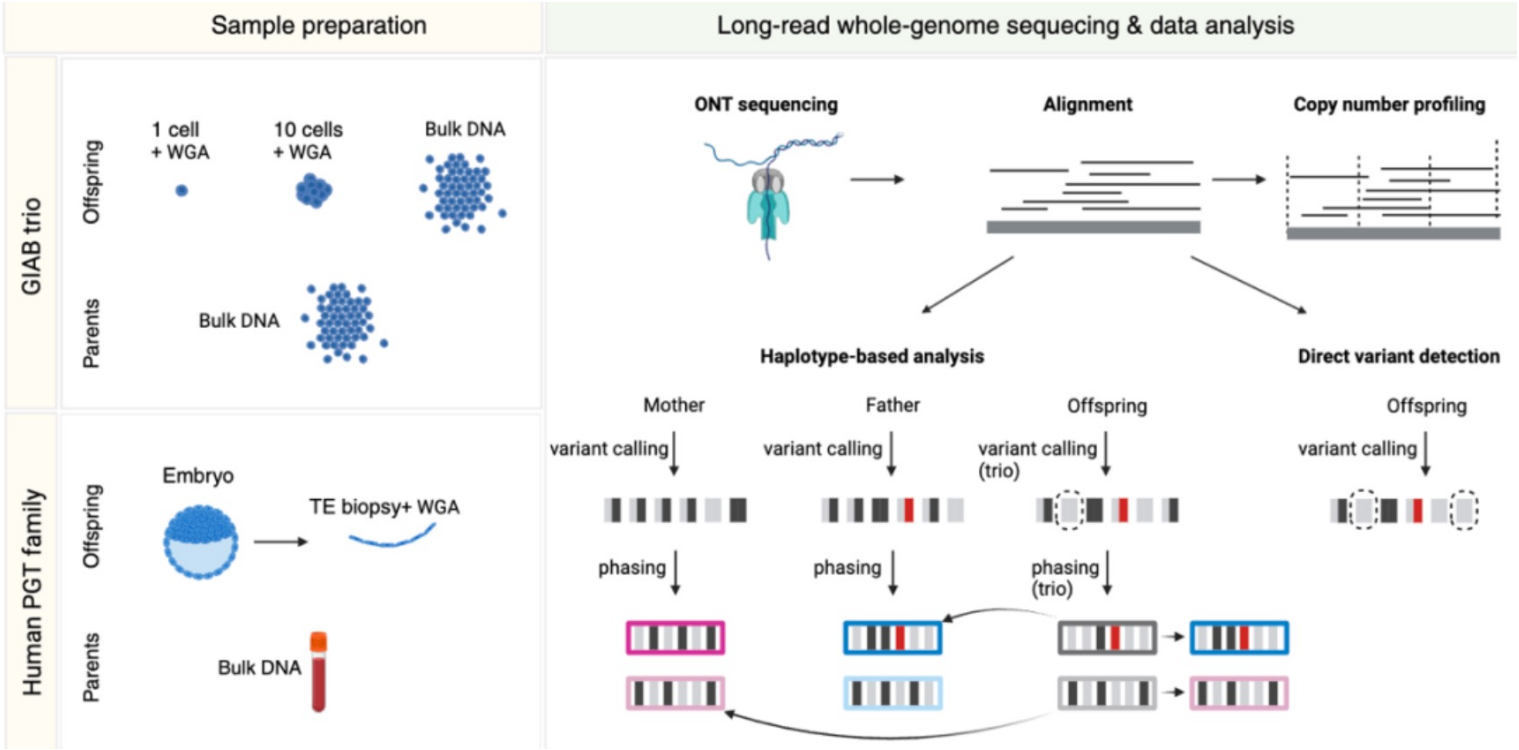
- PacBio long-read sequencing has clear benefits to variant detection specificity and sensitivity
- As long-read data sets grow → annotation of long-read-only rare disease will grow over time

Nanopore only

Long-read whole-genome sequencing-based concurrent haplotyping and aneuploidy profiling of single cells
Zhao et al., *bioRxiv* sept. 2024

Feasibility of lrWGS for haplotyping of single cells without requiring additional phasing references ?

Nanopore sequencing was performed on single-cell (1 cell) and multi-cell (10 cells) from the offspring



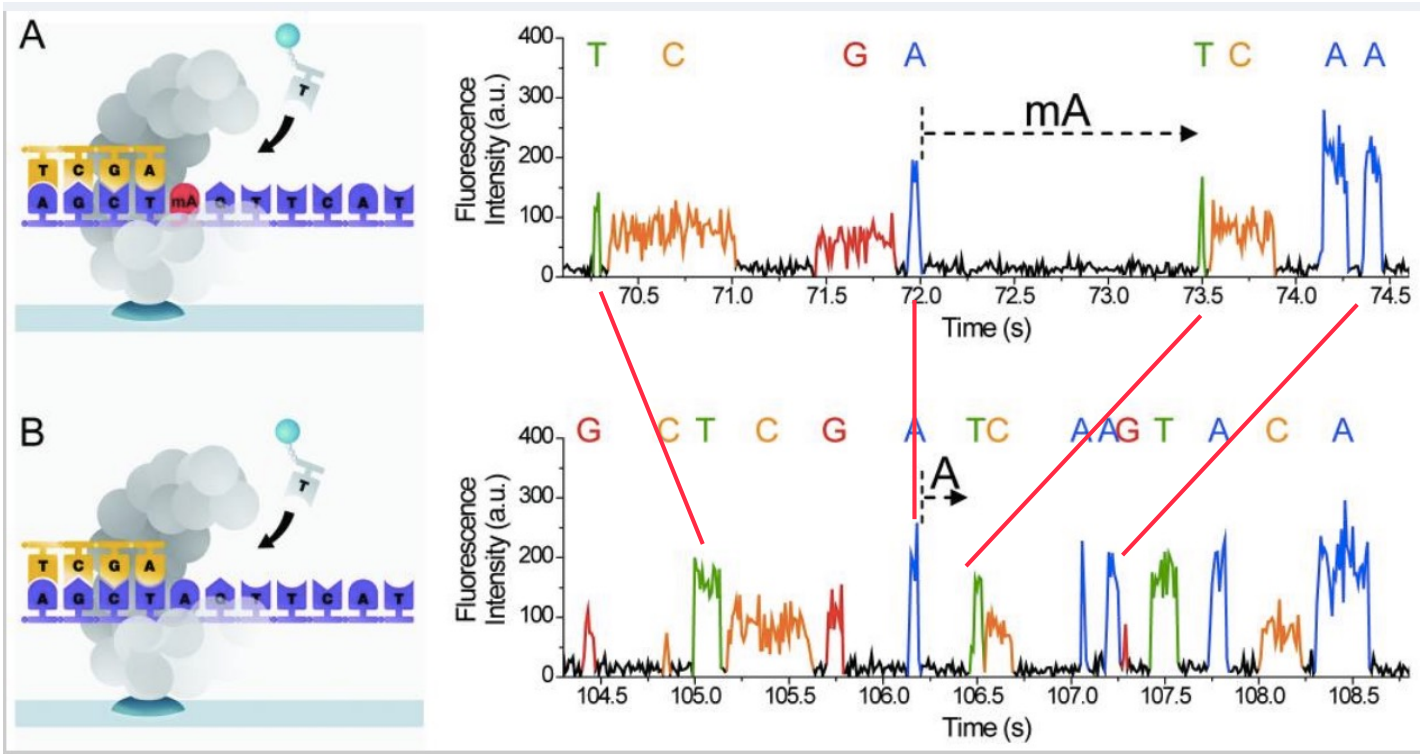
Bioinformatics pipeline that enables haplotyping of single cells using Nanopore sequencing

Effectiveness for genome-wide, reference-free comprehensive pre-implantation genetic testing

→ Cell-based noninvasive prenatal testing by analyzing single circulating trophoblast cells in maternal blood

DNA MODIFICATIONS

PacBio DETECTION OF MODIFIED DNA BASES



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

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

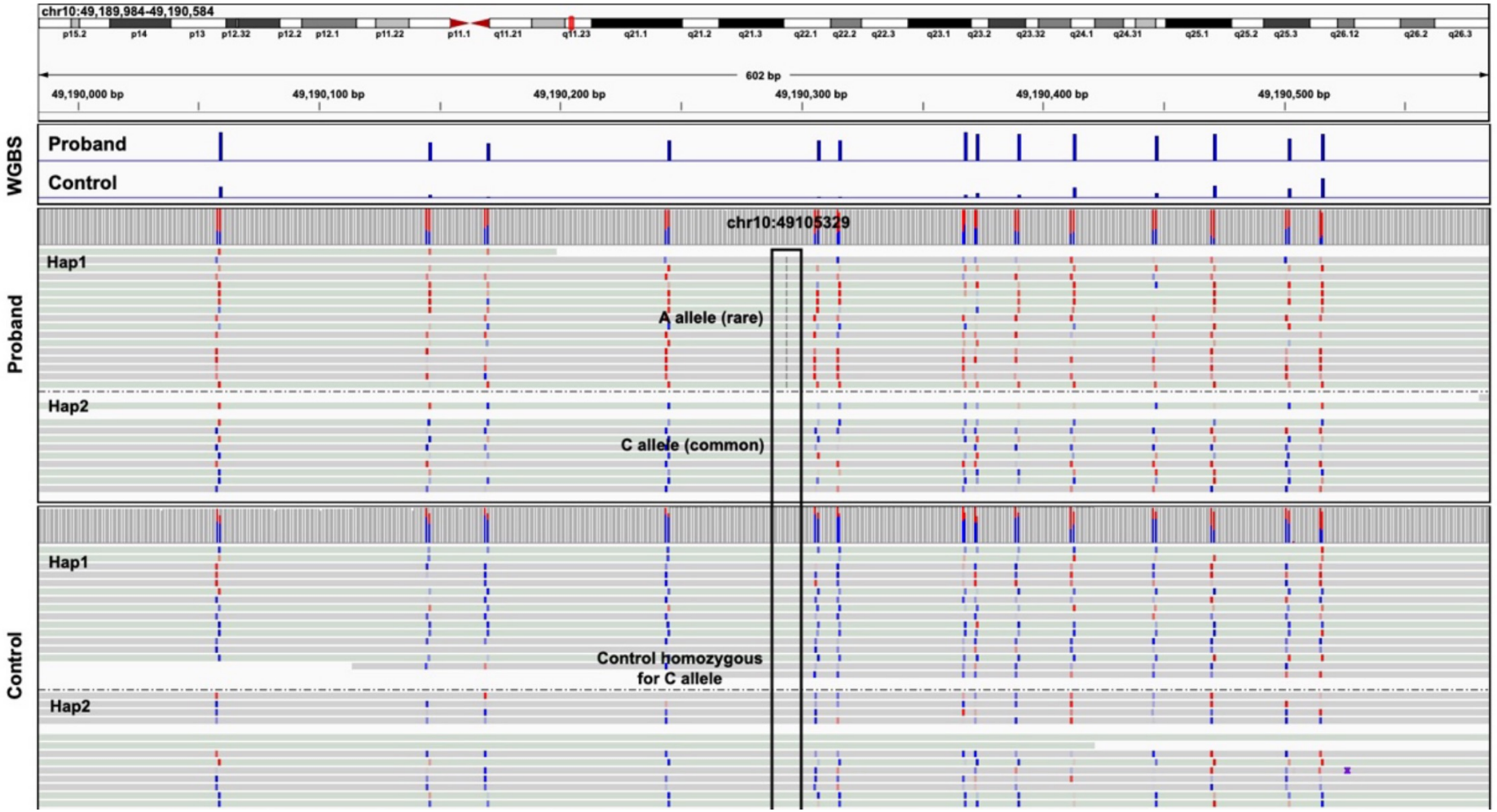
Feng et al. *PLOS Comput Biol* 2013

PacBio only

Direct haplotype-resolved 5-base HiFi sequencing for genome-wide profiling of hypermethylation outliers in a rare disease cohort
Cheung et al. *Nature Comm.* 2023

Cohort:
- rare disease cohort of 276 samples in 152 families
- sequencing data set : haplotype resolved 5-base HiFi

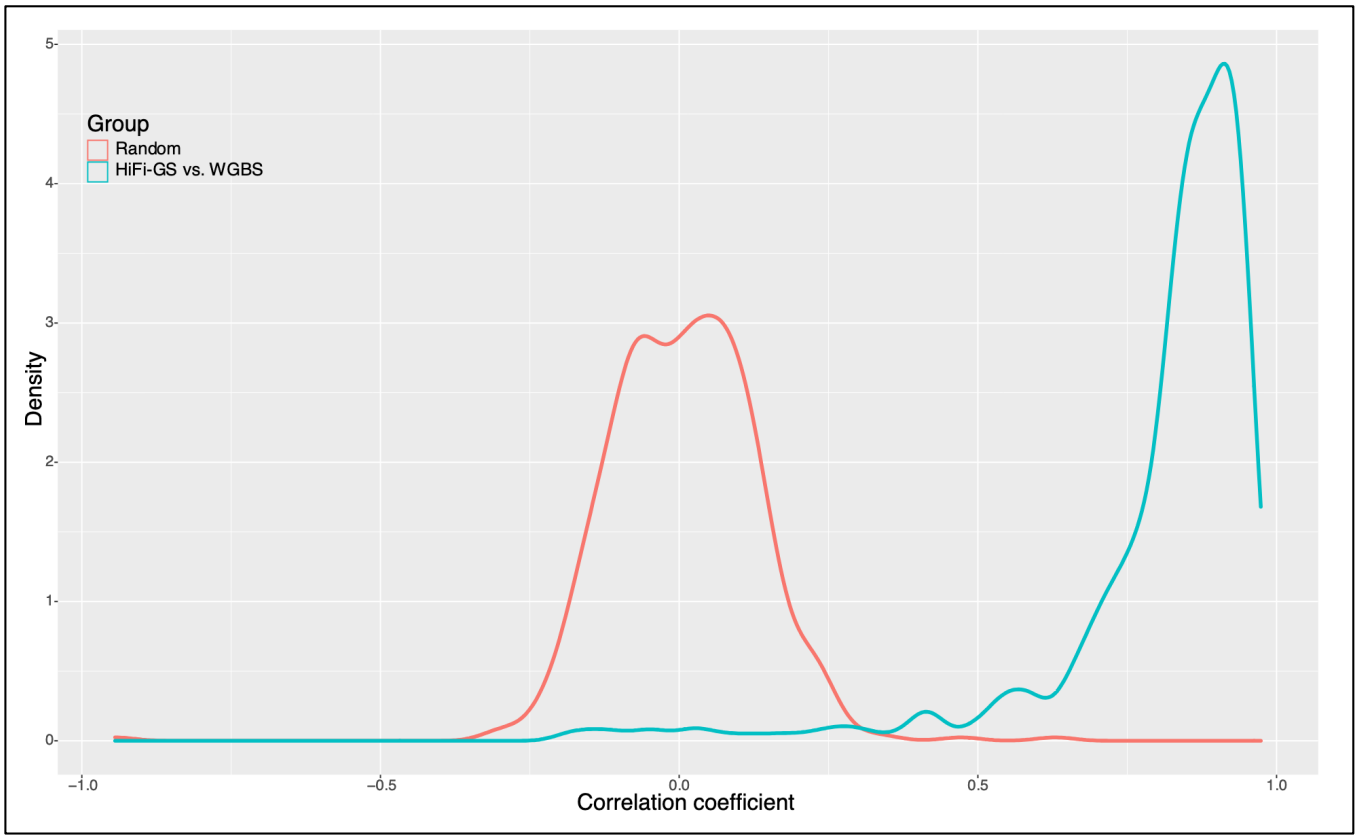
Objectives:
- Identify rare (~0.5%) hypermethylation events



Correlation (Pearson R, x-axis) of 500 CpGs from 93 samples profiled by HiFi-GS and WGBS.

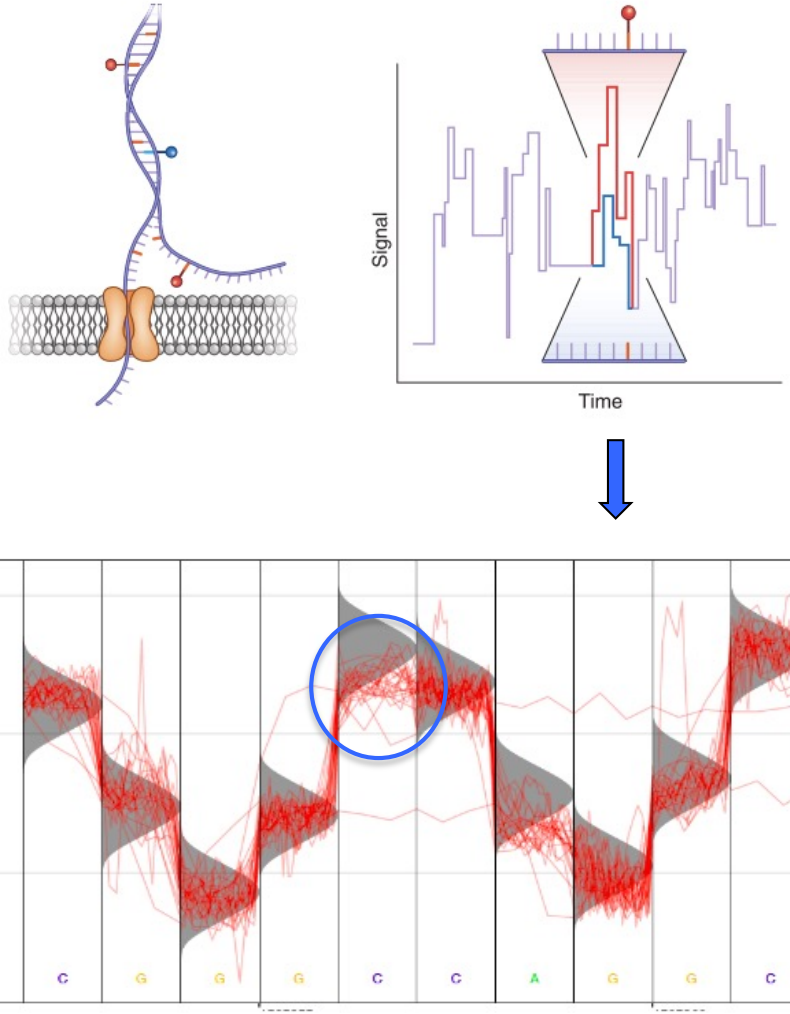
Blue line : correlation between HiFi-GS and WGBS measures of the 500 CpGs

Red line : similar but when WGBS values are permuted



Recent algorithmic development enables simultaneous detection of CpG methylation directly in HiFi reads

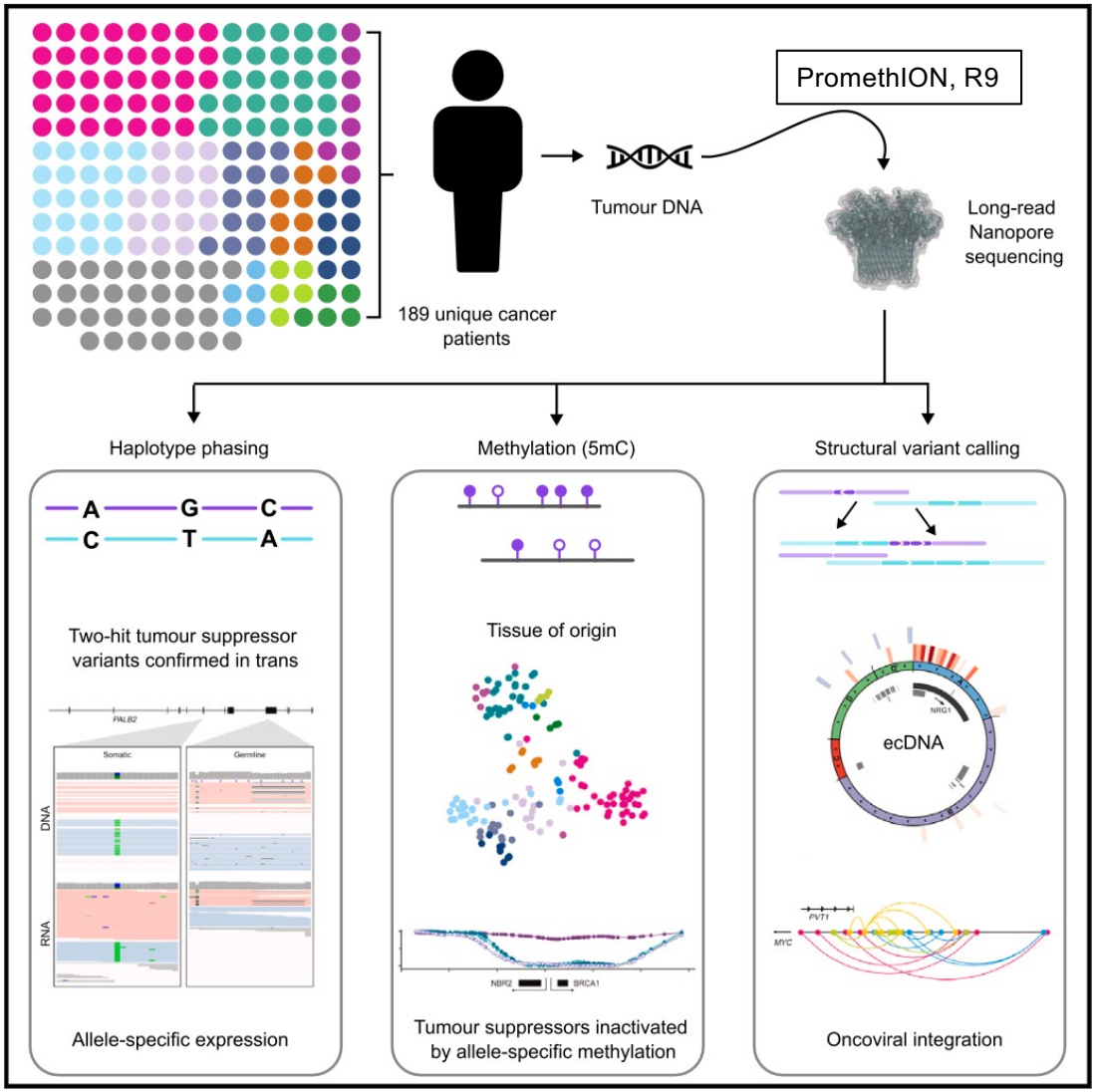
NANOPORE DETECTION OF MODIFIED DNA BASES



Nanopore only

Long-read sequencing of an advanced cancer cohort resolves rearrangements, unravels haplotypes, and reveals methylation landscapes
O'Neill et al. Cell Genomics Nov. 2024 (39 authors)

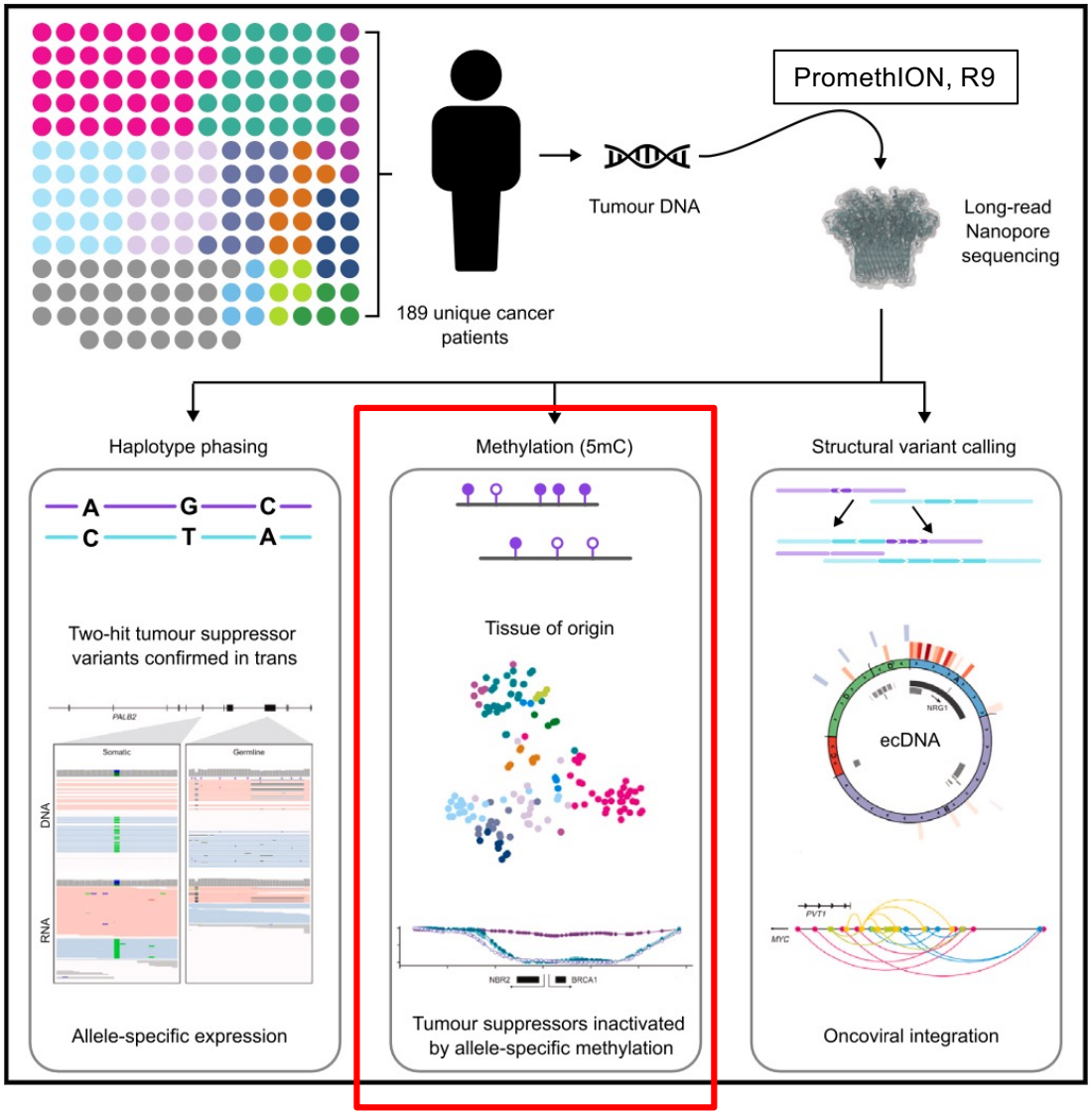
- Rich data resource of 189 long-read sequenced patient tumors
- Long-read sequencing allows detection of features not detectable with short-reads
  - phasing
  - complex rearrangements
  - methylation



Nanopore only

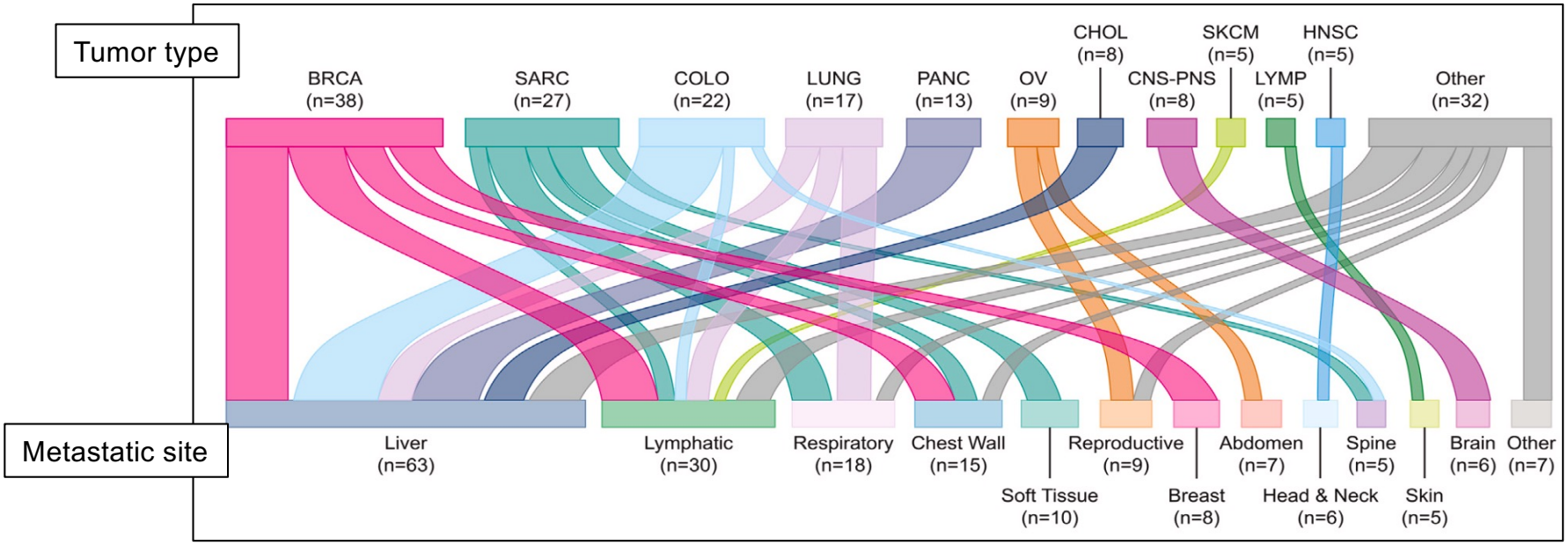
Long-read sequencing of an advanced cancer cohort resolves rearrangements, unravels haplotypes, and reveals methylation landscapes
O'Neill et al. Cell Genomics Nov. 2024 (39 authors)

- Rich data resource of 189 long-read sequenced patient tumors
- Long-read sequencing allows detection of features not detectable with short-reads
  - phasing
  - complex rearrangements
  - methylation



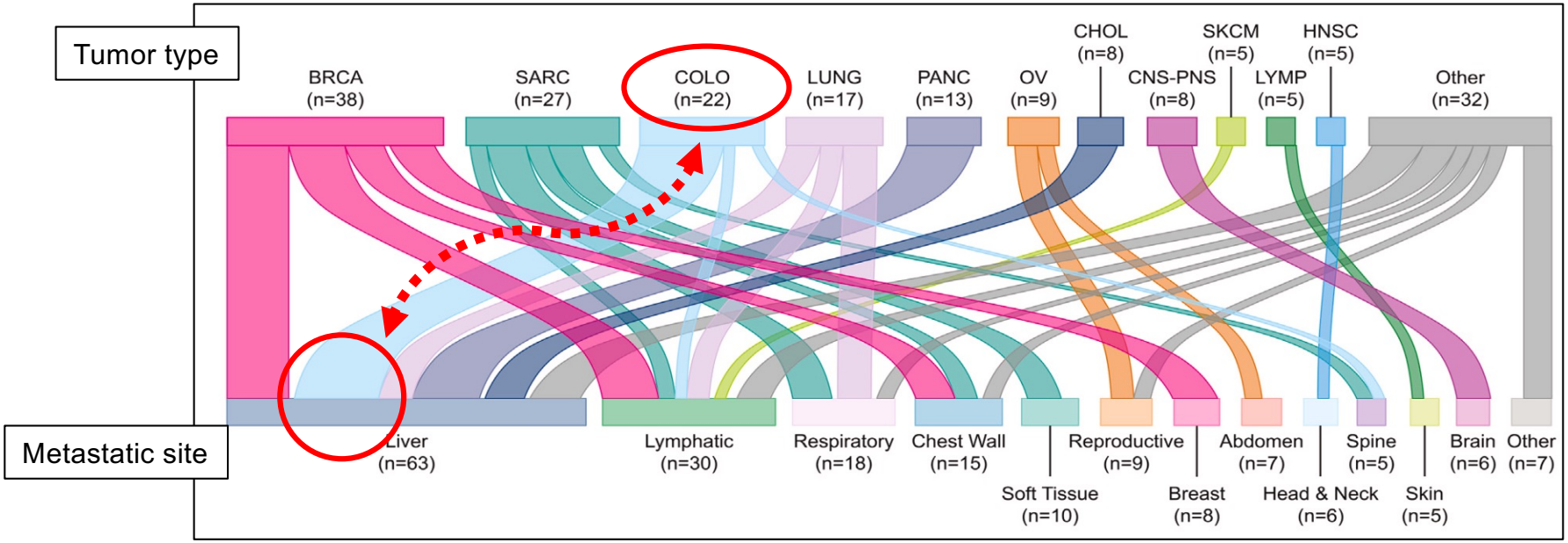
Long-Read Personalized OncoGenomics (POG) cohort :

- 189 tumor samples
- 26 cancer types
- majority from biopsies of metastatic sites

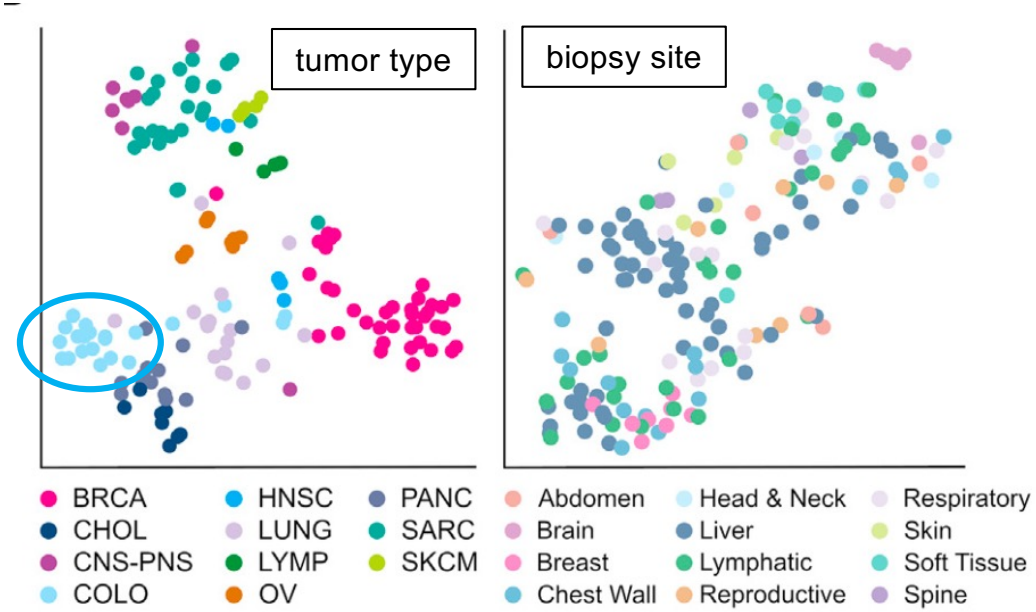


Long-Read Personalized OncoGenomics (POG) cohort :

- 189 tumor samples
- 26 cancer types
- majority from biopsies of metastatic sites

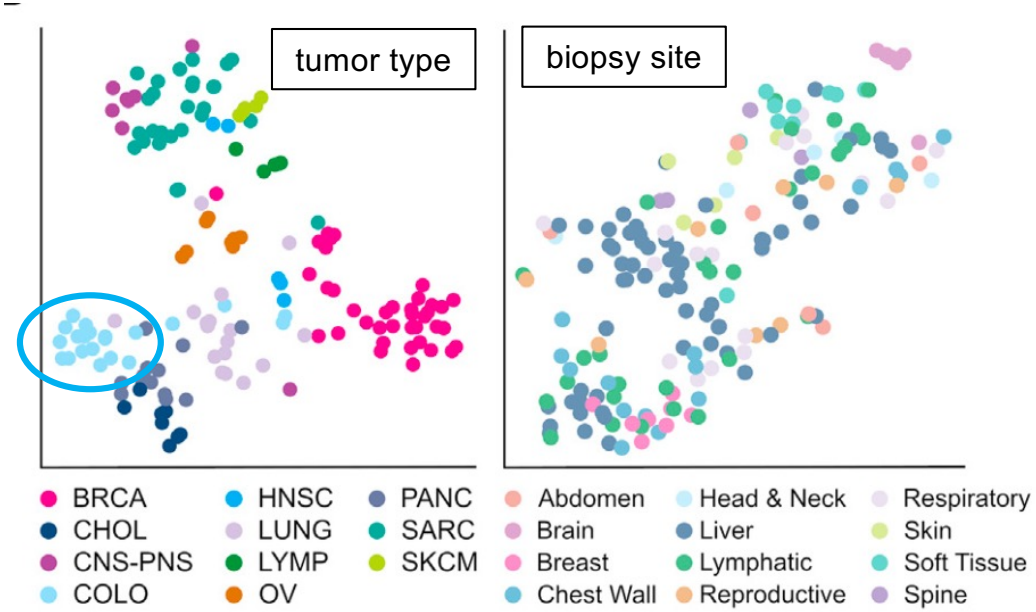


tSNE plots based on DNA methylation at regulatory regions compared with tumor type and biopsy site



tSNE: t-distributed stochastic neighbor embedding

tSNE plots based on DNA methylation at regulatory regions compared with tumor type and biopsy site



tSNE: t-distributed stochastic neighbor embedding

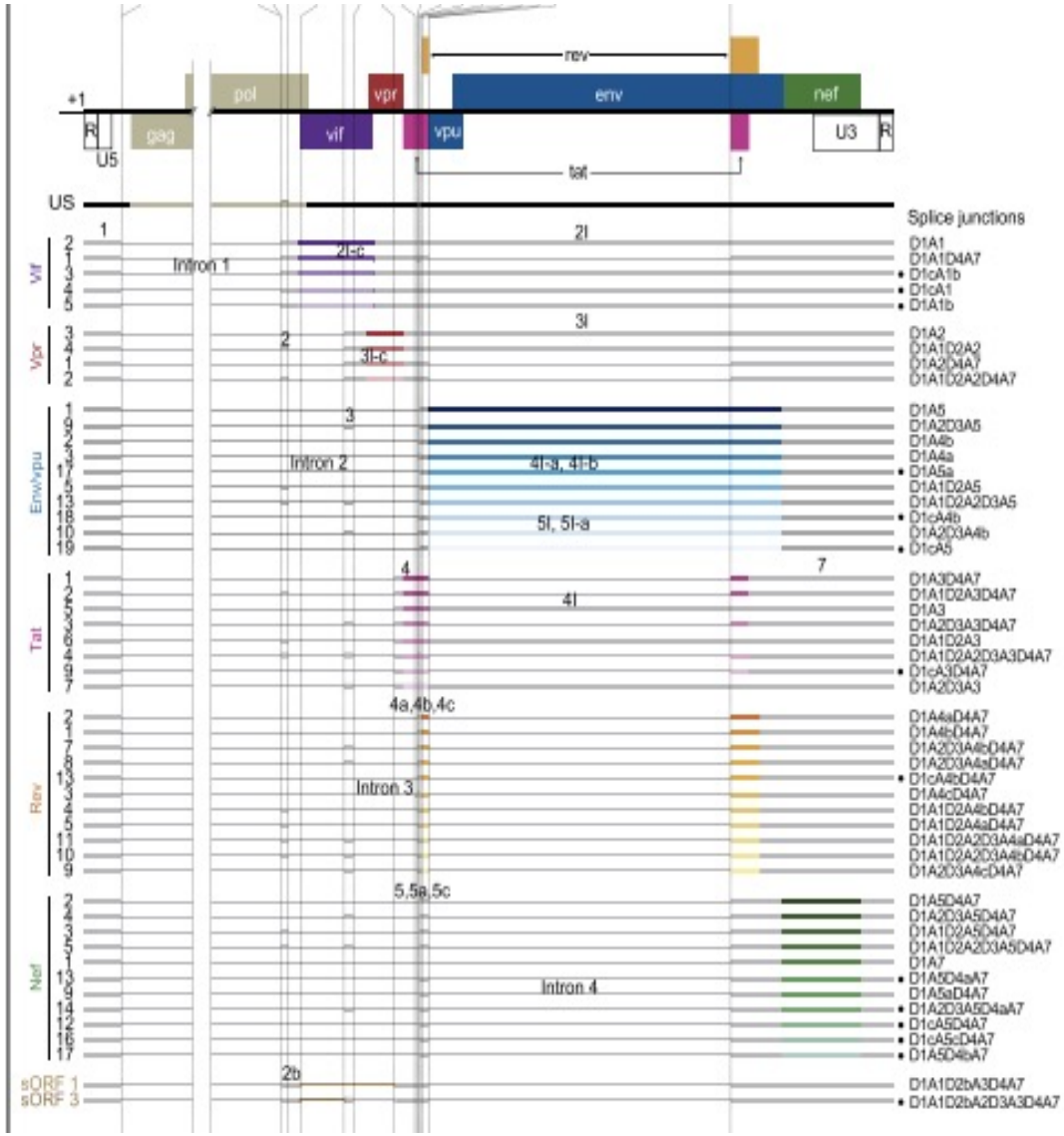
- Samples tend to group by tumor tissue of origin, irrespective of metastatic biopsy site
- Potential utility of DNA methylation for detecting or confirming tissue of origin in advanced and metastatic cancers, in addition to RNA-seq

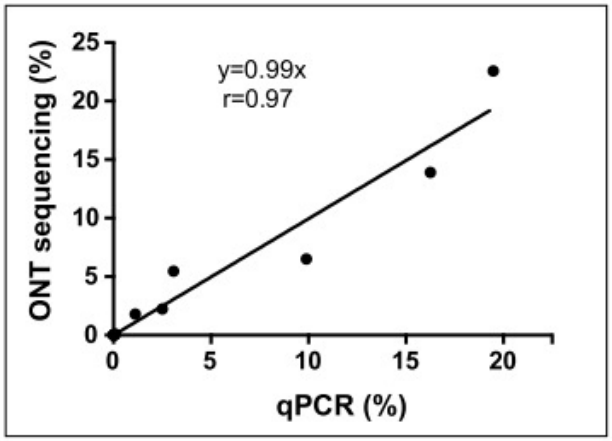
DETECTION OF SPLICING ISOFORMS

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

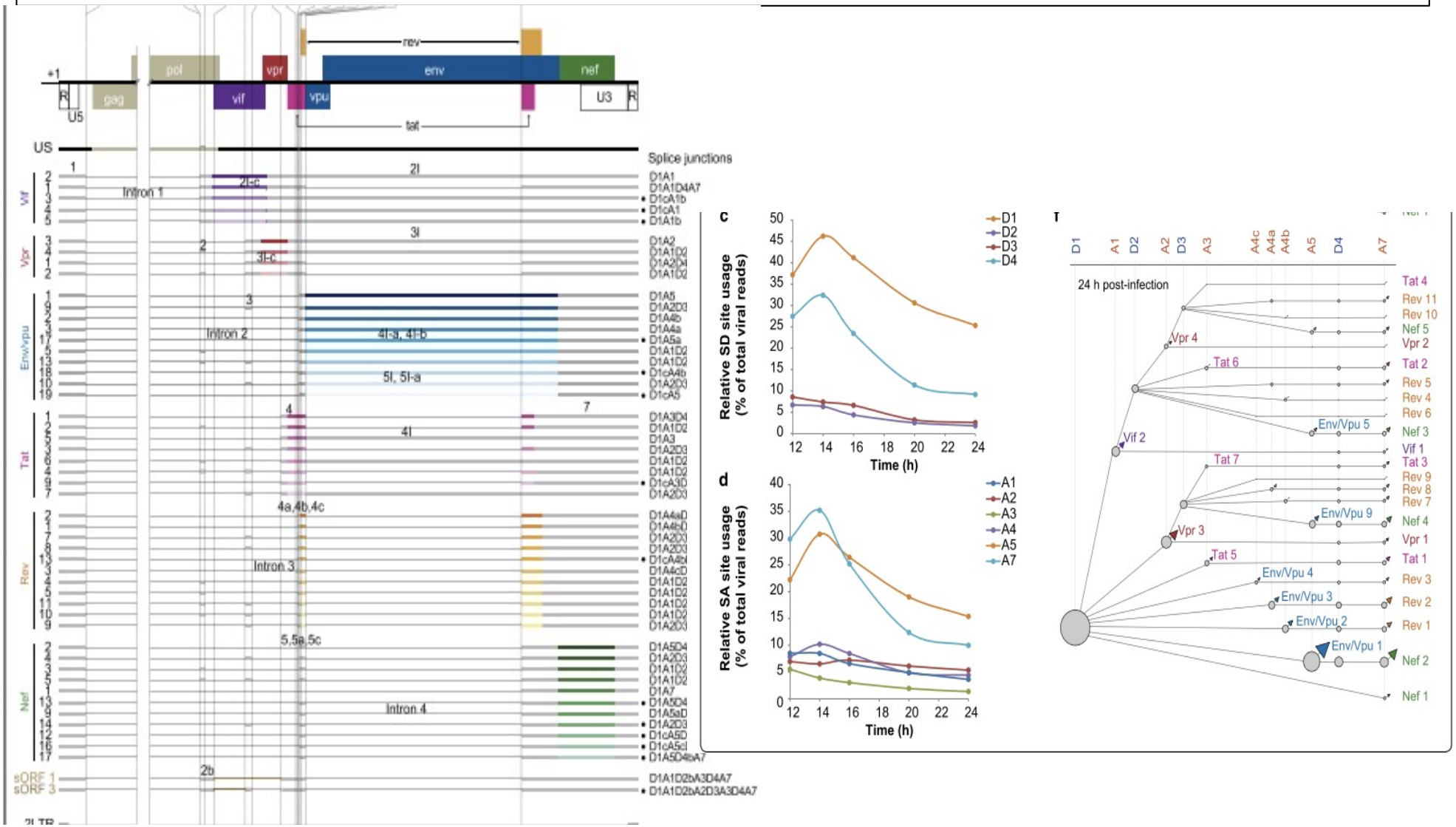




cDNA Nanopore sequencing

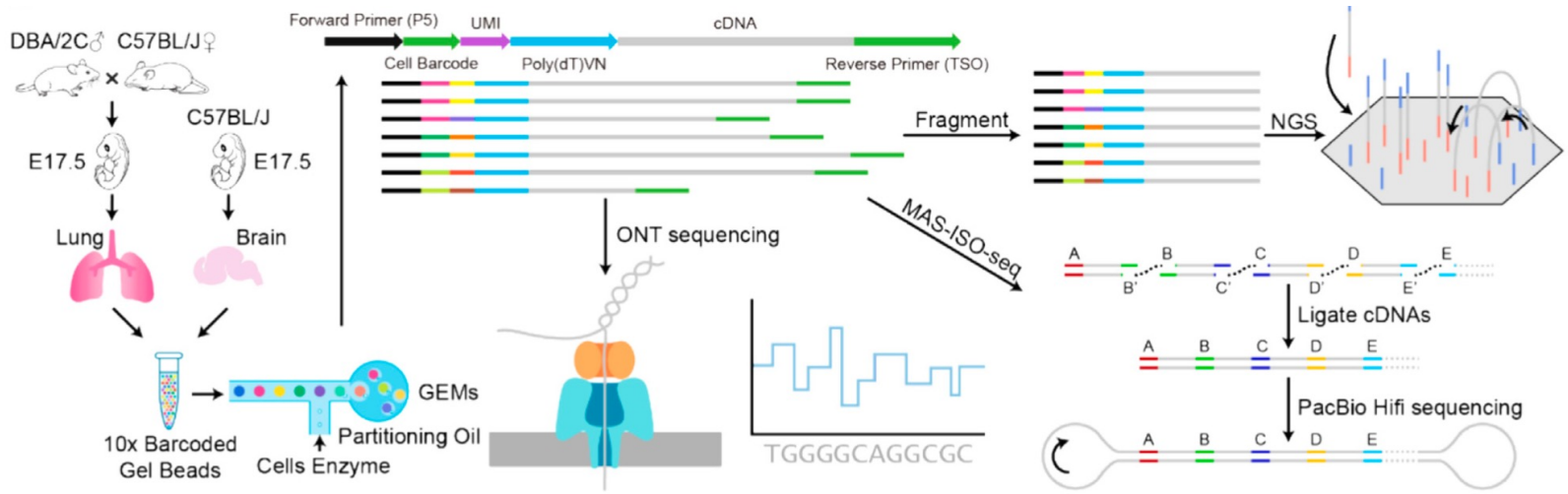
Dynamic nanopore long-read sequencing analysis of HIV-1 splicing events during the early steps of infection
Quang et al. *Retrovorology* 2020

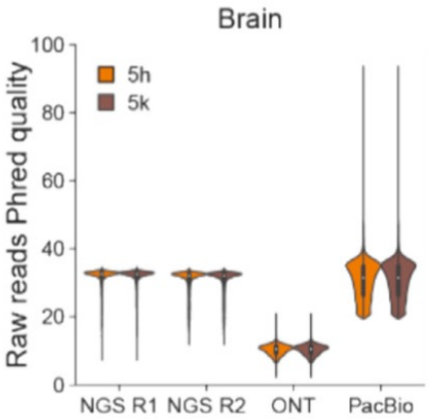
- 53 viral RNA isoforms, including 14 new ones
- Relative levels highly correlated with qPCR
- First dynamic picture of the cascade of events occurring between 12 and 24 h of viral infection



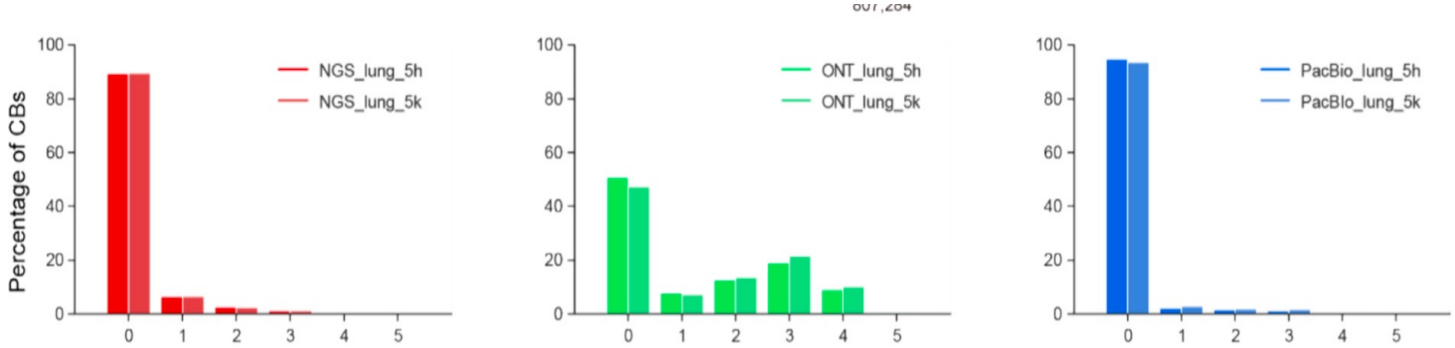
PacBio vs Nanopore

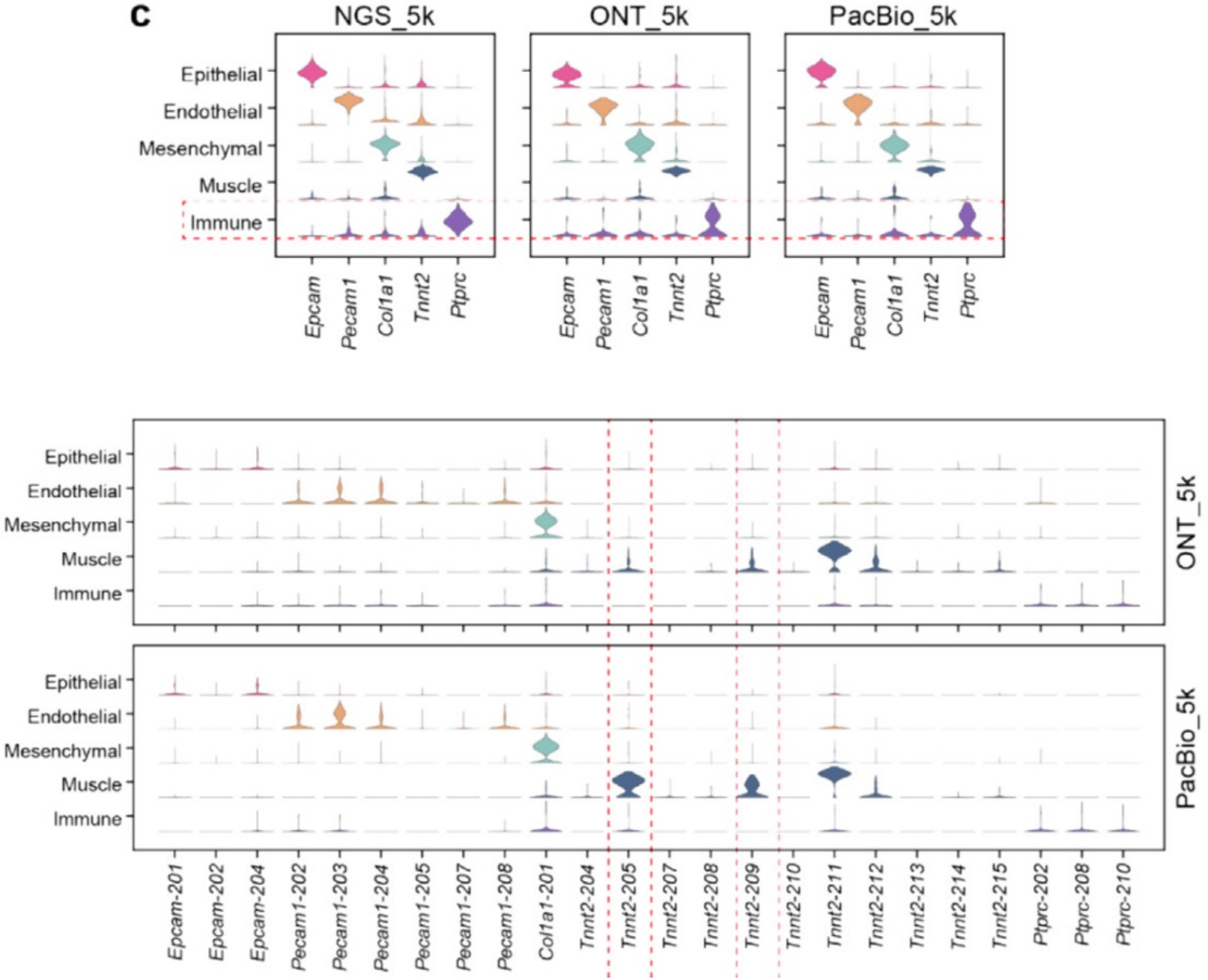
Systematic evaluation of single-cell RNA-seq analyses performance based on long-read sequencing platforms
Deng et al. *J. Adv. Res.* may 2024



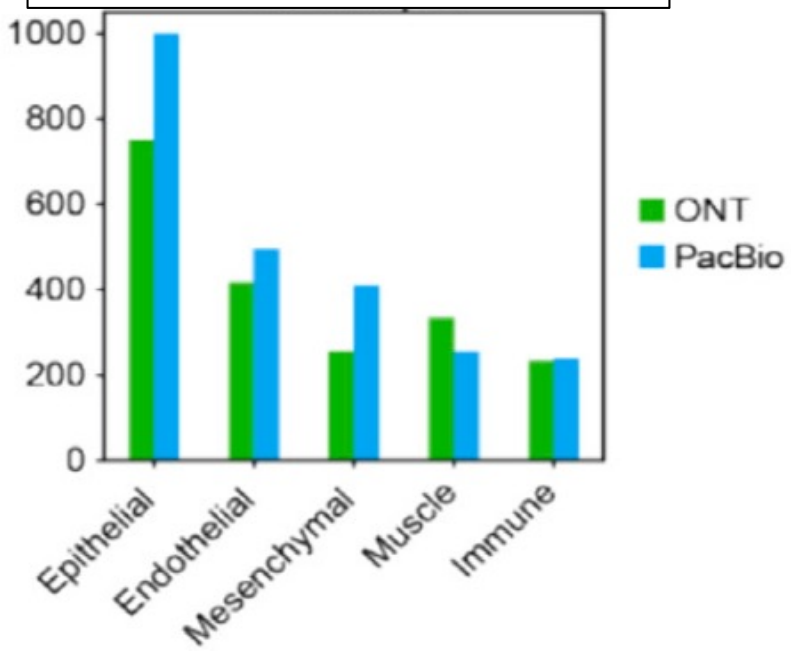


cell barcodes

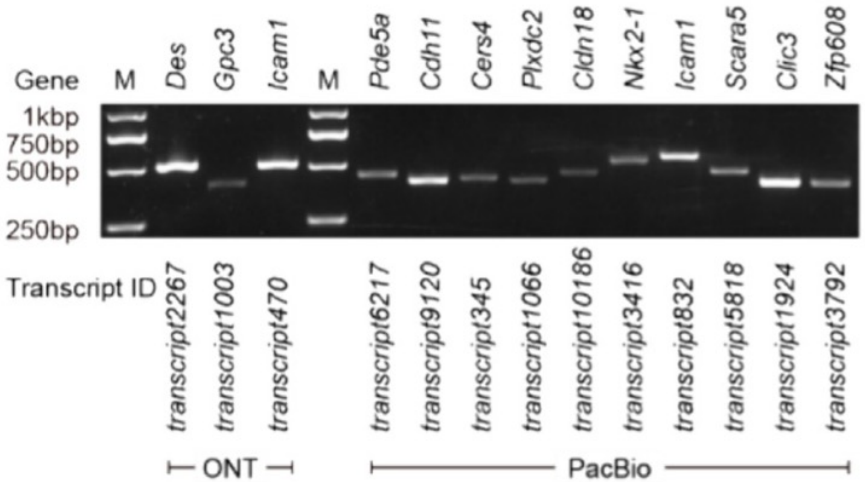



C
NGS_5k
ONT_5k
PacBio_5k
Epithelial
Endothelial
Mesenchymal
Muscle
Immune
Epcam
Pecam1
Col1a1
Tnnt2
Ptprc
ONT_5k
PacBio_5k
Epcam-201
Epcam-202
Epcam-204
Pecam1-202
Pecam1-203
Pecam1-204
Pecam1-205
Pecam1-207
Pecam1-208
Col1a1-201
Tnnt2-204
Tnnt2-205
Tnnt2-207
Tnnt2-208
Tnnt2-209
Tnnt2-210
Tnnt2-211
Tnnt2-212
Tnnt2-213
Tnnt2-214
Tnnt2-215
Ptprc-202
Ptprc-208
Ptprc-210

Differentially expressed isoforms



Novel isoforms



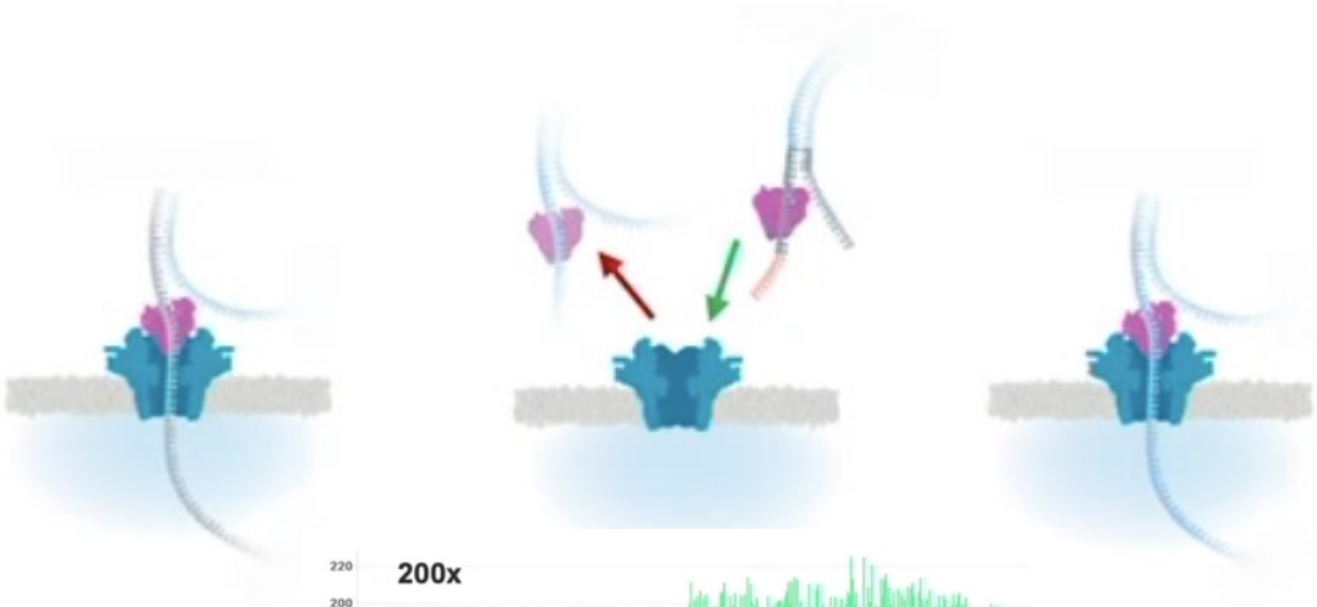
Conclusions

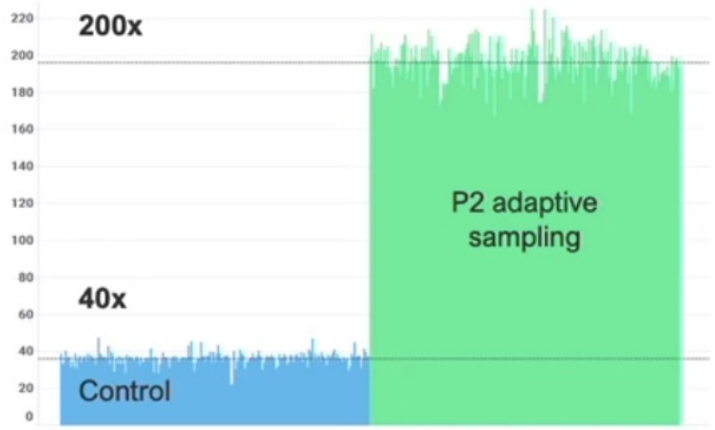
- Both platforms achieved comparable accuracy on cell type identification
- but PacBio identified a greater amount of cell type specific isoforms and of novel isoforms

TARGETED SEQUENCING

NANOPORE ADAPTIVE SAMPLING

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





Cancer gene panel – 202 target regions

Nanopore only

Nanopore-targeted sequencing (NTS) for intracranial tuberculosis: a promising and reliable approach
Yang et al. *Ann Clin Microbiol Antimicrob.* oct. 2024

Cohort

- 100 patients with intracranial tuberculosis ; the diagnosis was based on :
  - their clinical features
  - micro-biological and cerebrospinal fluid cytology
  - radiological findings, etc.
- 22 patients with other brain diseases

Table 2 Diagnostic efficiency of the five tests for intracranial tuberculosis

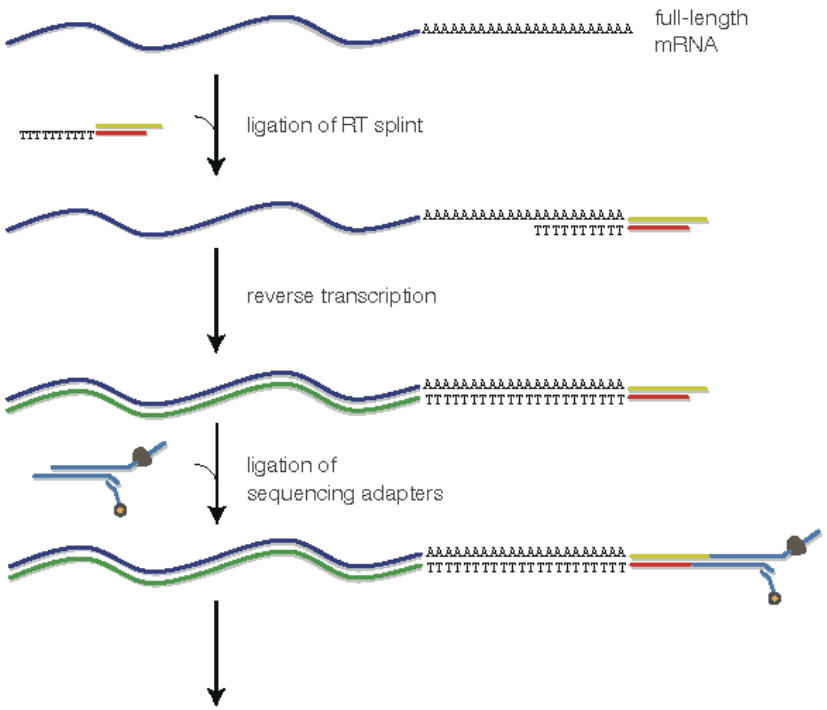
Test	Sensitivity(%,95%CI)	Specificity(%,95%CI)	PPV(%,95%CI)	NPV(%,95%CI)	AUC(95%CI)
NTS	60.0(49.7–69.5)	95.5(75.1–99.8)	98.4(90.0-99.9)	34.4 (23.0-47.8)	0.78(0.71–0.84)
Xpert	5.0(1.9–11.8)	95.5(75.1–99.8)	83.3(36.5–99.1)	18.1(11.8–26.6)	0.50(0.45–0.55)
MTB culture	2.0(0.3–7.7)	100.0(81.5–100.0)	100.0(19.8–100.0)	18.3(12.1–26.7)	0.51(0.50–0.52)
PCR	1.0(0.1–6.2)	100.0(81.5–100.0)	100.0(5.5–100)	18.2(12.0-26.5)	0.51(0.50–0.51)
AFB smear	0.0(0.0-4.6)	100.0(81.5–100.0)	/	18.0(11.9–26.3)	0.50(0.50–0.50)

PPV: positive predictive value; NPV: negative predictive value; AUC: area under the curve; MTB: Mycobacterium tuberculosis; AFB: acid-fast bacilli

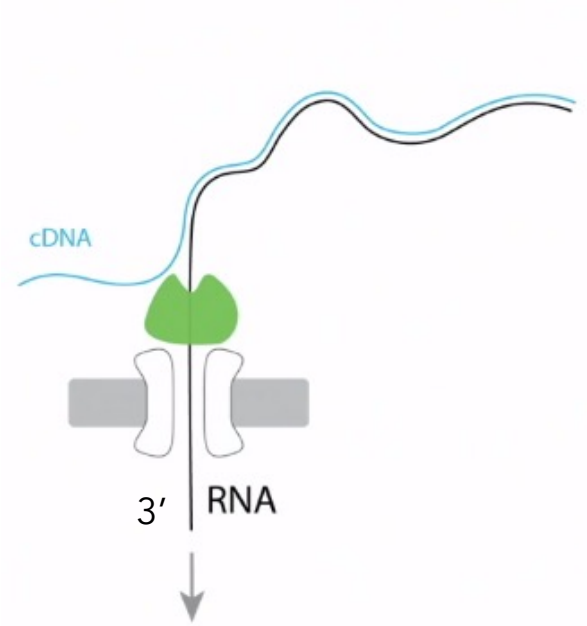
DIRECT RNA SEQUENCING

DIRECT RNA SEQUENCING

Library preparation



sequencing



RNA directly sequenced in nanopore

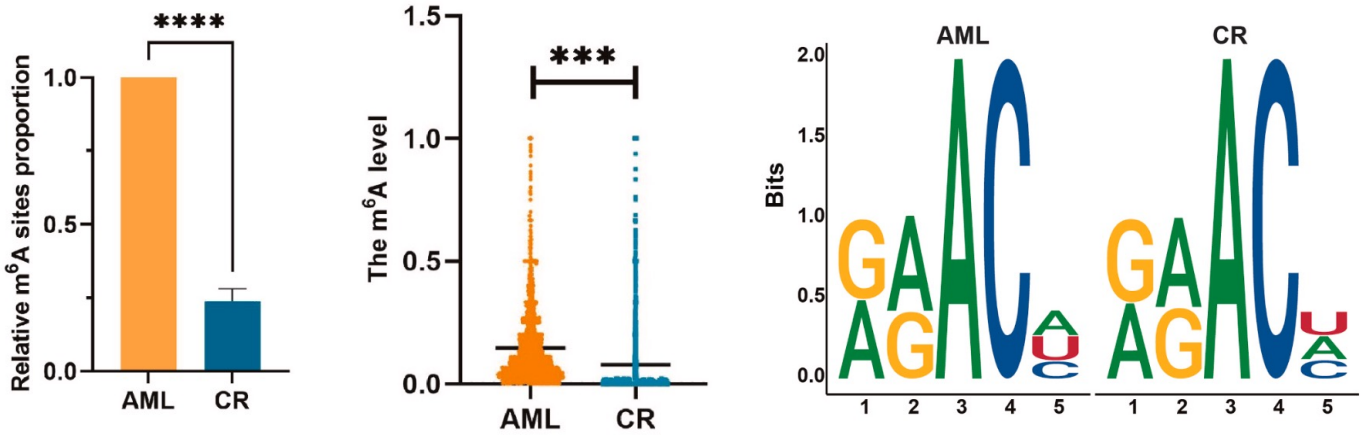
- No PCR bias
- Quantitative

Nanopore only

Comprehensive analysis of m6A methylome alterations after azacytidine plus venetoclax treatment for acute myeloid leukemia by nanopore sequencing
Zhang et al. Computational and Structural Biotechnology Journal, 2024

Objectives :
- Growing evidence of functional role of RNA-m6A in acute myeloid leukemia (AML)
- However the global m6A level after azacytidine plus venetoclax treatment is unclear
- Determine the m6A levels in bone marrow samples before and after treatment

Sequencing and detection of m6A methylation :
- Nanopore direct RNA sequencing with GridION, R9.4.1 flow cells, **RNA002 chemistry**
- Determination of m6A levels at nucleotide precision with Tombo and DENA



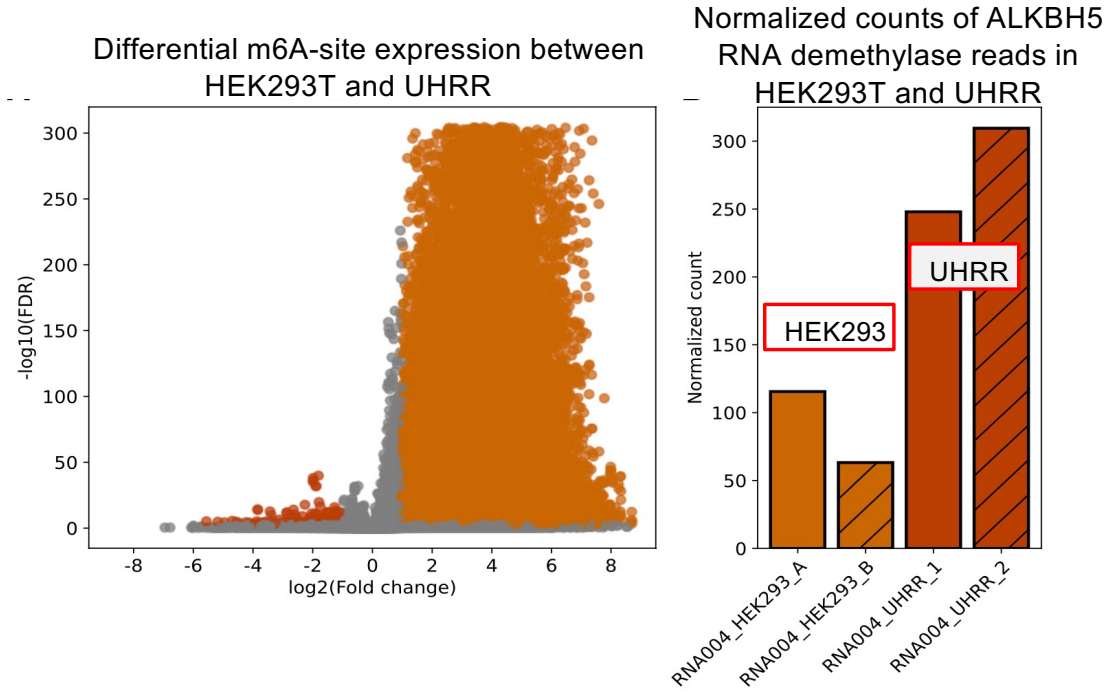
CONCLUSIONS
- Ilustration for the first time of the global landscape of m6A levels in AZA plus VEN treated patients
- → AZA plus VEN treatment has a significant demethylation effect at the RNA level in AML patients

Nanopore only

Direct RNA sequencing (RNA004) allows for improved transcriptome assessment and near real-time tracking of methylation for medical applications.
Hewel et al. *bioRxiv* jul. 2024

Previous chemistry for direct RNA sequencing : RNA002
Major points of concern : low throughput, low accuracy, modification basecalling was not enabled by default
New chemistry : RNA004 :

- new nanopore
- new motor-protein
- new base-calling
- models for both standard nucleotides and modifications (m6A, pseudo U)



Global upregulation of m6A in HEK293T correlates with decreased expression of the ALKBH5 demethylase

New step in genome assembly :

the T2T era

PacBio+Nanopore+Illumina

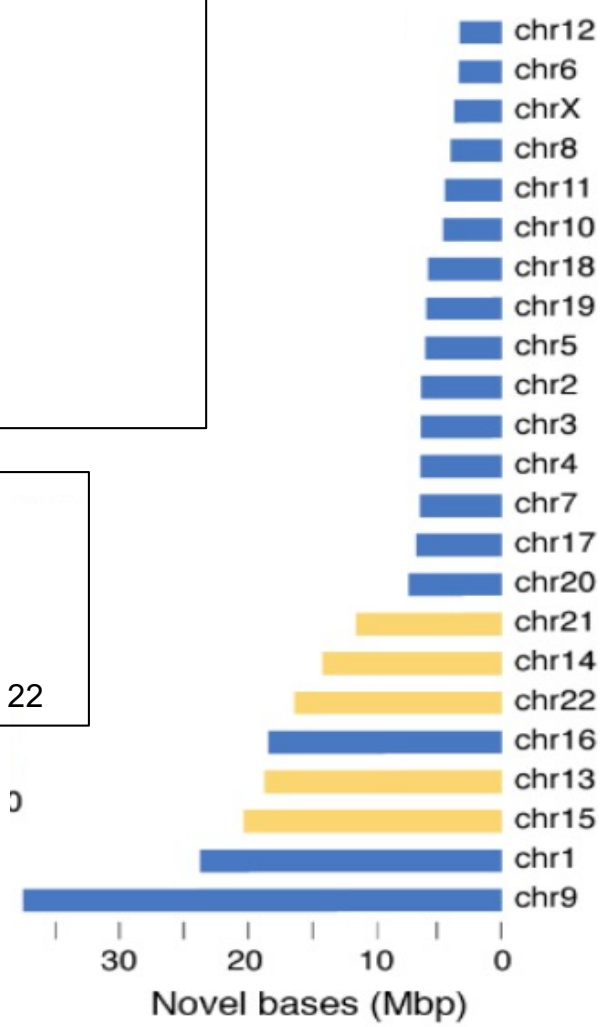
The complete sequence of a human genome
Nurk et al. *Science* 2022

SEQUENCING

Data were obtained with a “complete hydatidiform mole” (CHM13) cell line (haploid) :

- 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

- T2T assembly :including all 22 autosomes plus Chromosome X :
  - Introduces 200 million bp of novel sequence
  - all centromeric regions
  - entire short arms (p-arms) of 5 acrocentric chromosomes : 13, 14, 15, 21, 22

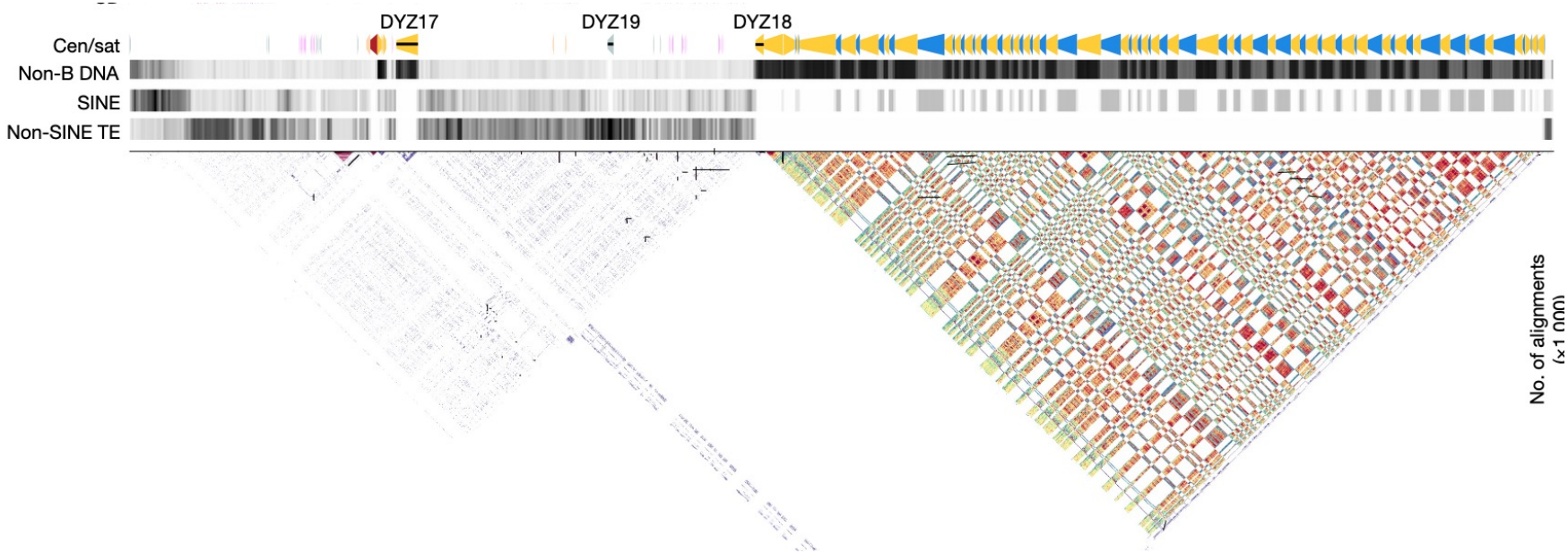


PacBio+Nanopore+Illumina

The complete sequence of a human Y chromosome
Rhie et al. *Nature* 2023 (88 authors)

HG002 diploid genome

- Y chromosome -> last chromosome completed from telomere to telomere
- PacBio HiFi reads (60× haploid genome coverage)
- ONT ultralong reads (90× in reads > 100 kb)
- Strand-seq
- combined T2T-Y with CHM13 to produce a complete reference sequence for all 24 human chromosomes

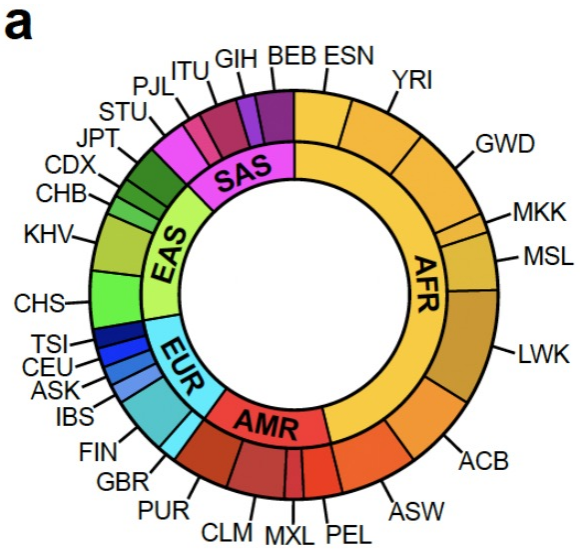


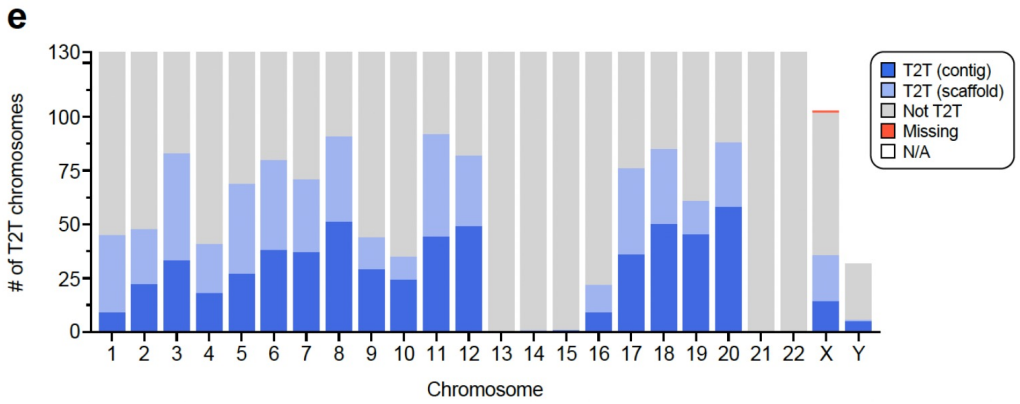
PacBio+Nanopore+Illumina

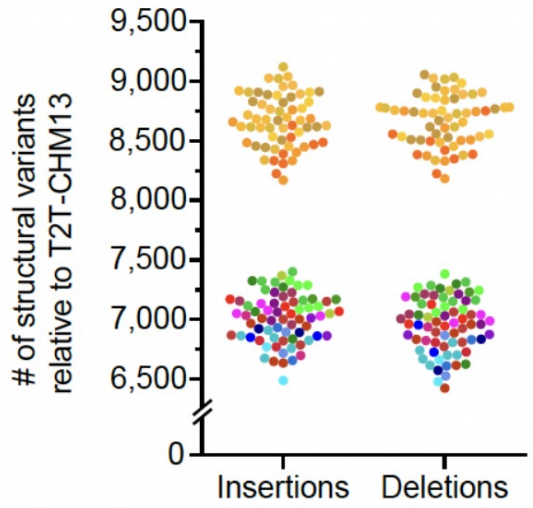
Complex genetic variation in nearly complete human genomes
Logsdon et al. *bioRxiv* sept. 2024. (39 authors)

Objectives
- construct a human pangenome reference
- understand the extent of complex structural variation

Results
- sequence 65 diverse human genomes from 28 population groups
- build 130 haplotype-resolved assemblies (130 Mbp median continuity)
- reaching telomere-to-telomere (T2T) for 39% of the chromosomes
- completely assemble and validate 1,246 human centromeres
- whole-genome inference to a median quality value QV = 45

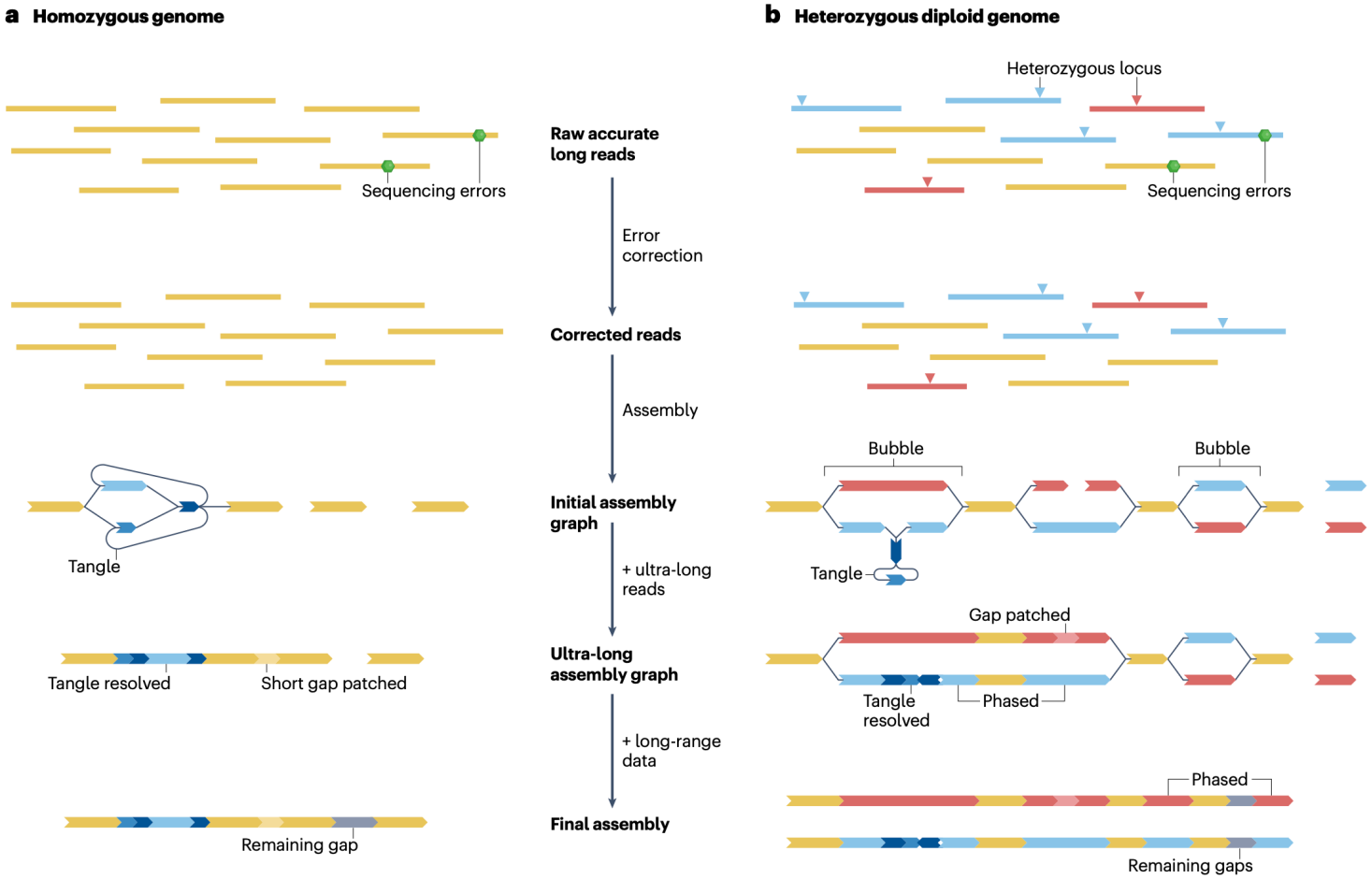






Genome assembly in the telomere-to-telomere era
Li H and Durbin R. *Nature Reviews Genetics*, sept 2024

Strategy for near-telomere-to-telomere assembly



Data type	Technologies	Description	Roles
Accurate long reads	PacBio HiFi, ONT duplex	>10 kb in length; error rate <0.5%	Initial assembly graph construction; phasing over heterozygous variants that are less than 10 kb apart
Ultra-long reads	ONT ultra-long	>100 kb in length; error rate <10%	Resolving tangles; phasing through homozygous regions over 100 kb in length
Trio data	Short-read	Standard whole-genome shotgun sequencing of parents	Whole-genome phasing
Long-range data	Hi-C, Pore-C, Strand-seq	Information over 1 kb to over 10 Mb in length	Chromosomal phasing; chromosome-scale scaffolding

Nanopore unltra-long reads only

Telomere-to-Telomere Phased Genome Assembly Using HERRO-Corrected Simplex Nanopore Reads
Stanojević et al. *bioRxiv* oct. 2024

- Are T2T phased assemblies with unltra-long nanopore reads-only ?
- Development of HERRO :
  - deep learning-based framework :
  - corrects nanopore ultra-long reads
  - while carefully preserving informative positions that differentiate the haplotypes or repeat copies

Dataset	**Correction**	**Mismatch per 10 kbp**	**Non-Hp Ins per 10 kbp**	**Non-Hp Del per 10 kbp**	**Hp-Ins per 10 kbp**	**Hp-Del per 10 kbp**	**Errors per 10 kbp**
HG002	Before	133.15	45.56	58.54	26.22	71.02	334.48
	After (66x)	0.23	0.63	0.74	0.41	1.77	3.78
I002C	Before	186.86	58.76	79.74	33.41	99.44	458.21
	After (63x)	0.22	1.01	1.72	0.50	2.18	5.64
CHM13	Before	158.63	74.75	91.23	46.91	128.49	500.01
	After	0.38	0.34	0.85	1.23	1.60	4.39
A. thaliana	Before	38.59	27.84	22.82	10.03	29.09	128.37
	After	0.70	1.37	1.92	0.05	2.85	6.89
D. rerio	Before	71.84	29.64	50.04	17.27	41.64	210.43
	After	0.41	2.23	15.15	0.10	3.88	21.77

Table 1 ***Counts of errors before and after correction, by type.***

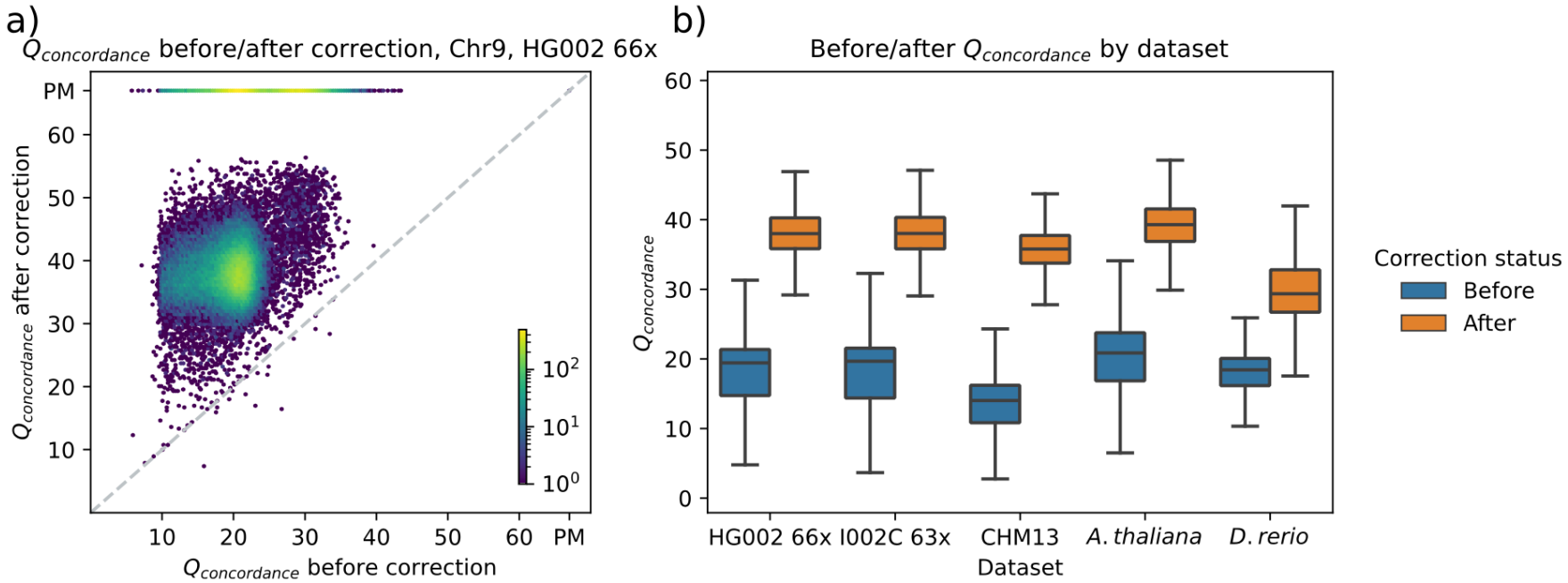
Nanopore only

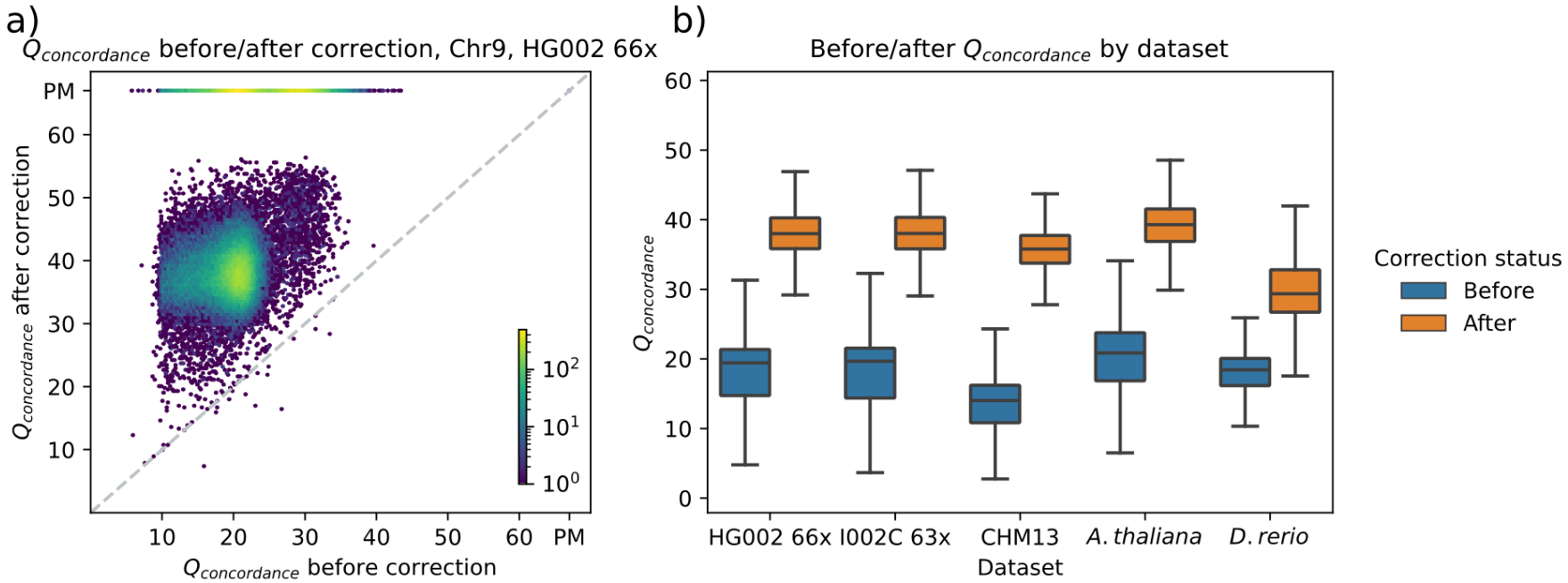
Telomere-to-Telomere Phased Genome Assembly Using HERRO-Corrected Simplex Nanopore Reads
Stanojević et al. *bioRxiv* oct. 2024

- Telomere-to-telomere phased assemblies involve HiFi reads and unltra-long nanopore reads
- Development of HERRO :
  - deep learning-based framework :
  - corrects ultra-long nanopore reads
  - while carefully preserving informative positions that differentiate the haplotypes or repeat copies

Dataset	**Correction**	**Mismatch per 10 kbp**	**Non-Hp Ins per 10 kbp**	**Non-Hp Del per 10 kbp**	**Hp-Ins per 10 kbp**	**Hp-Del per 10 kbp**	**Errors per 10 kbp**	
HG002	Before	133.15 **Q18**	45.56	58.54	26.22	71.02	334.48	**Q15**
	After (66x)	0.23 **Q46**	0.63	0.74	0.41	1.77	3.78	**Q34**
I002C	Before	186.86	58.76	79.74	33.41	99.44	458.21	
	After (63x)	0.22	1.01	1.72	0.50	2.18	5.64	
CHM13	Before	158.63	74.75	91.23	46.91	128.49	500.01	
	After	0.38	0.34	0.85	1.23	1.60	4.39	
A. thaliana	Before	38.59	27.84	22.82	10.03	29.09	128.37	
	After	0.70	1.37	1.92	0.05	2.85	6.89	
D. rerio	Before	71.84	29.64	50.04	17.27	41.64	210.43	
	After	0.41	2.23	15.15	0.10	3.88	21.77	

Table 1 ***Counts of errors before and after correction, by type.***


a)
$Q_{concordance}$ before/after correction, Chr9, HG002 66x
$Q_{concordance}$ after correction
$Q_{concordance}$ before correction
PM
10
20
30
40
50
60
10^2
10^1
10^0
b)
Before/after $Q_{concordance}$ by dataset
$Q_{concordance}$
0
10
20
30
40
50
60
HG002 66x
I002C 63x
CHM13
A. thaliana
D. rerio
Dataset
Correction status
Before
After

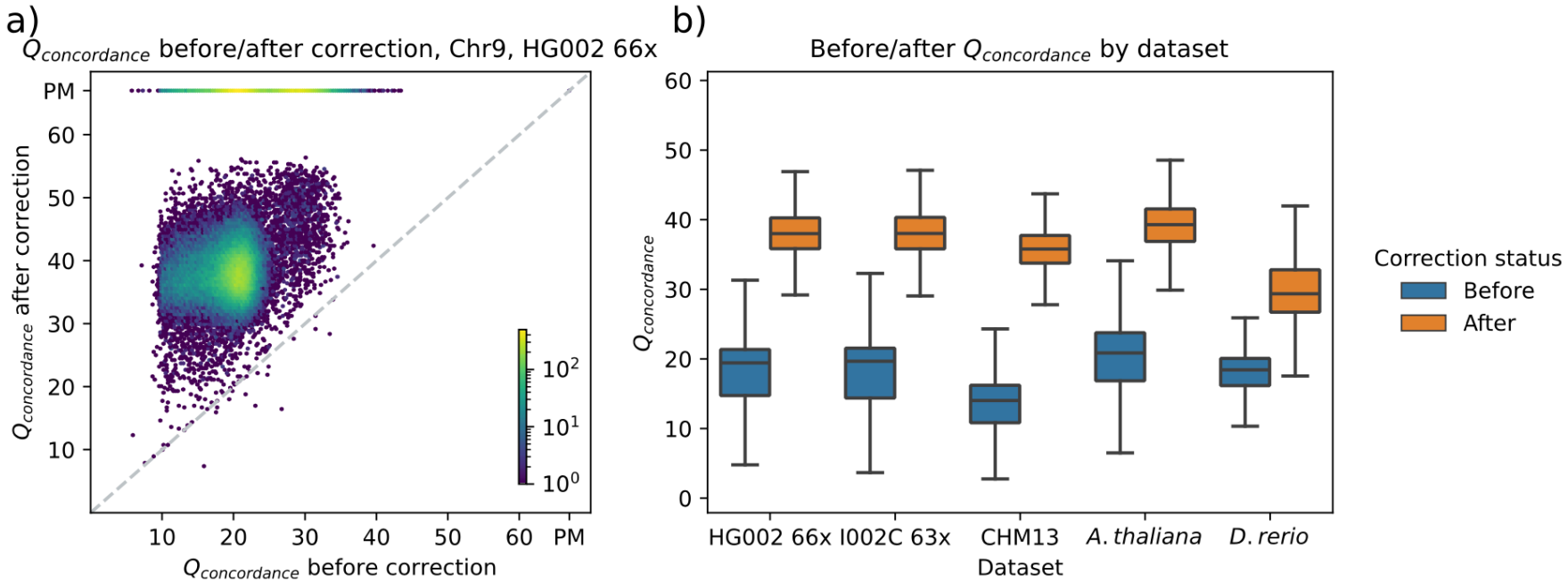


De novo assembly of HERRO-corrected ONT reads :

- more than half of the 46 chromosomes were obtained as T2T contigs or scaffolds
- Example of X and Y chromosomes :

Assembly	**Genome fraction (%)**	**Misassemblies**	**Mismatches per 100kbp**	**Total mismatches**	**Indels per 100kbp**	**Total Indels**	**Indel <=5 bp**	**Indel >5bp**
ChrX, trio	100	2	0.19	286	8.43	13018	12672	346
ChrY, trio	99.98	5	1.18	736	4.08	2544	2410	134

Table 3 ***Quast evaluation results of assembled T2T contigs of chromosomes X and Y.***



De novo assembly of HERRO-corrected ONT reads :

- more than half of the 46 chromosomes were obtained as T2T contigs or scaffolds
- Example of X and Y chromosomes :

Assembly	**Genome fraction (%)**	**Misassemblies**	**Mismatches per 100kbp**	**Total mismatches**	**Indels per 100kbp**	**Total Indels**	**Indel <=5 bp**	**Indel >5bp**
ChrX, trio	100	2	0.19	286	8.43	13018	12672	346
ChrY, trio	99.98	5	1.18	736	4.08	2544	2410	134

Table 3 ***Quast evaluation results of assembled T2T contigs of chromosomes X and Y.***

- HERRO enables high-quality assembly with corrected reads from a unique long-read technology
- It provides an opportunity to reduce the cost of genome sequencing and to analyze even more complex genomes with different levels of ploidy

Summary

PacBio

- Maximum read length : 200 kb
- CCS sequencing (HiFI reads) :
  - Very low error rate (Q33)
  - Best bacterial genome assembly
  - cDNA :
    - RNA-seq
    - Best for new splicing isoforms detection

Nanopore

- Very light sequencing system → portability → sequencing “in the field”
- Very long reads : maximum length > 1 Mb
- 10.4.1 flow cells: low error rate, accurate bacterial genome assembly
- Detection of modified DNA (5mC, 6mA)
- Direct sequencing of RNA (RNA004) :
  - RNA-seq
  - splicing isoforms detection
  - Direct base-calling of modified RNA nucleotides (6mA, pseudo U, etc..)
- Future improvements :
  - read correction → enables high-quality assembly, even T2T (?) from a unique long-read technology