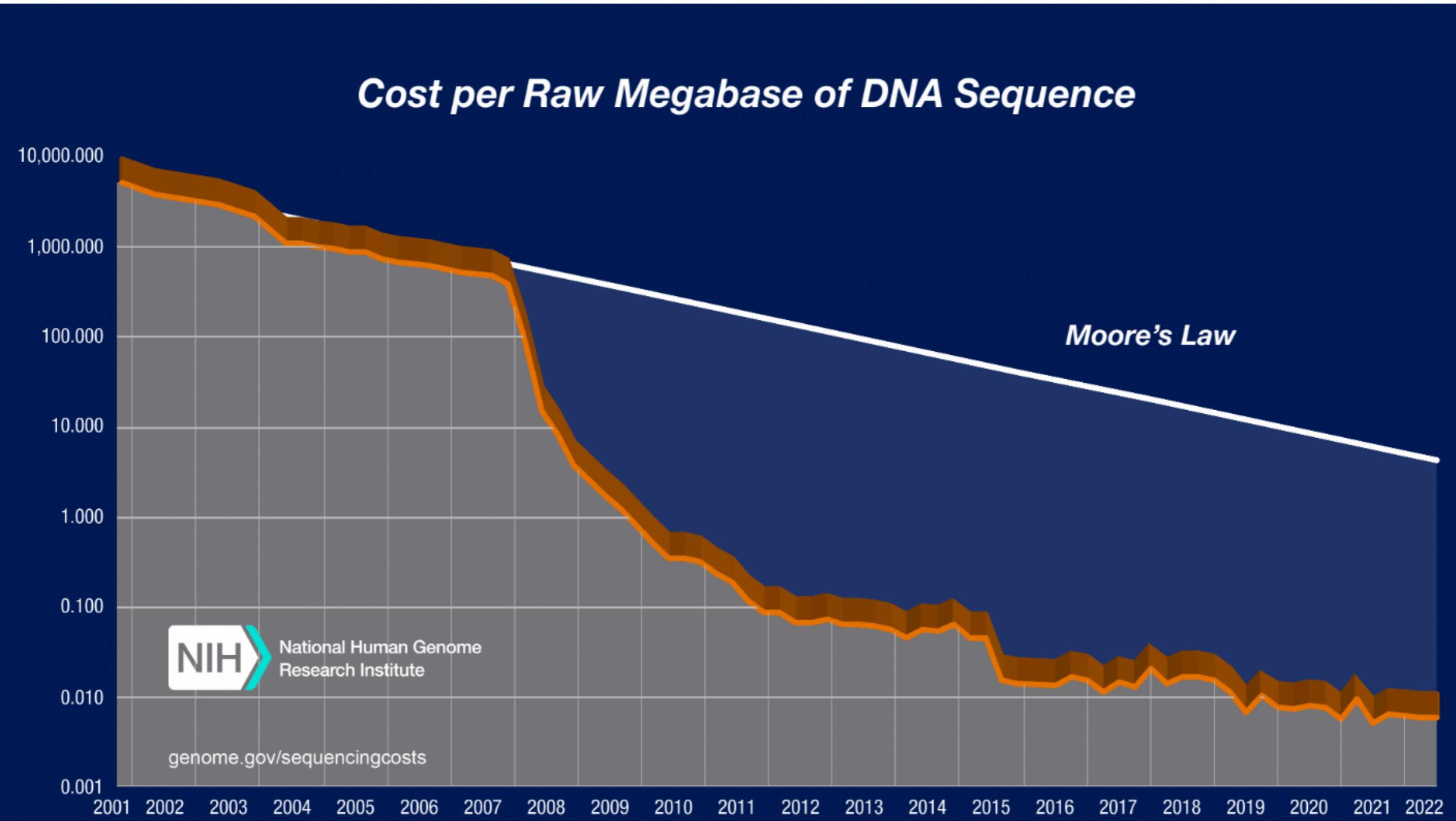


# Long Read Sequencing

Dick McCombie  
Davis Family Professor of Human Genetics  
Cold Spring Harbor Laboratory

Advanced Sequencing Technologies and Applications course  
Cold Spring Harbor Laboratory  
2025

# Significant advances in genome sequencing over last 20 years



## Evolution of genome assemblies

- Initial references – very high quality – extremely expensive
- Period of lower quality Sanger assemblies (~2001-2007)
- Next gen assemblies (short read) – 2007- now
- Third generation – long read assemblies -2013/2014 –now – what can we do currently?
- T2T extremely complete genomes

# Short vs long reads

- Short read NGS has revolutionized resequencing
- *De novo* assembly is possible but not optimal with short reads
- Long reads improve the ability to do *de novo* assembly dramatically
- Even in organisms with a good reference, such as humans, resequencing misses many structural differences relative to the reference
- Plant genomes are very large in general
- There are significant structural differences between different strains of the same plant such as rice
- These structural differences contribute to salient biological differences

# Advantages of Long Read length

Full scale of genetic variation

Repetitive regions

Structural variants

Enables higher quality alignments and assembly

Gapless genomes - T2T

??



STANLEY INSTITUTE FOR  
COGNITIVE GENOMICS  
COLD SPRING HARBOR LABORATORY

# The Telomere-to-Telomere Consortium

Long read sequencing  
of the hydatidiform  
mole CHM13 with  
multiple technologies

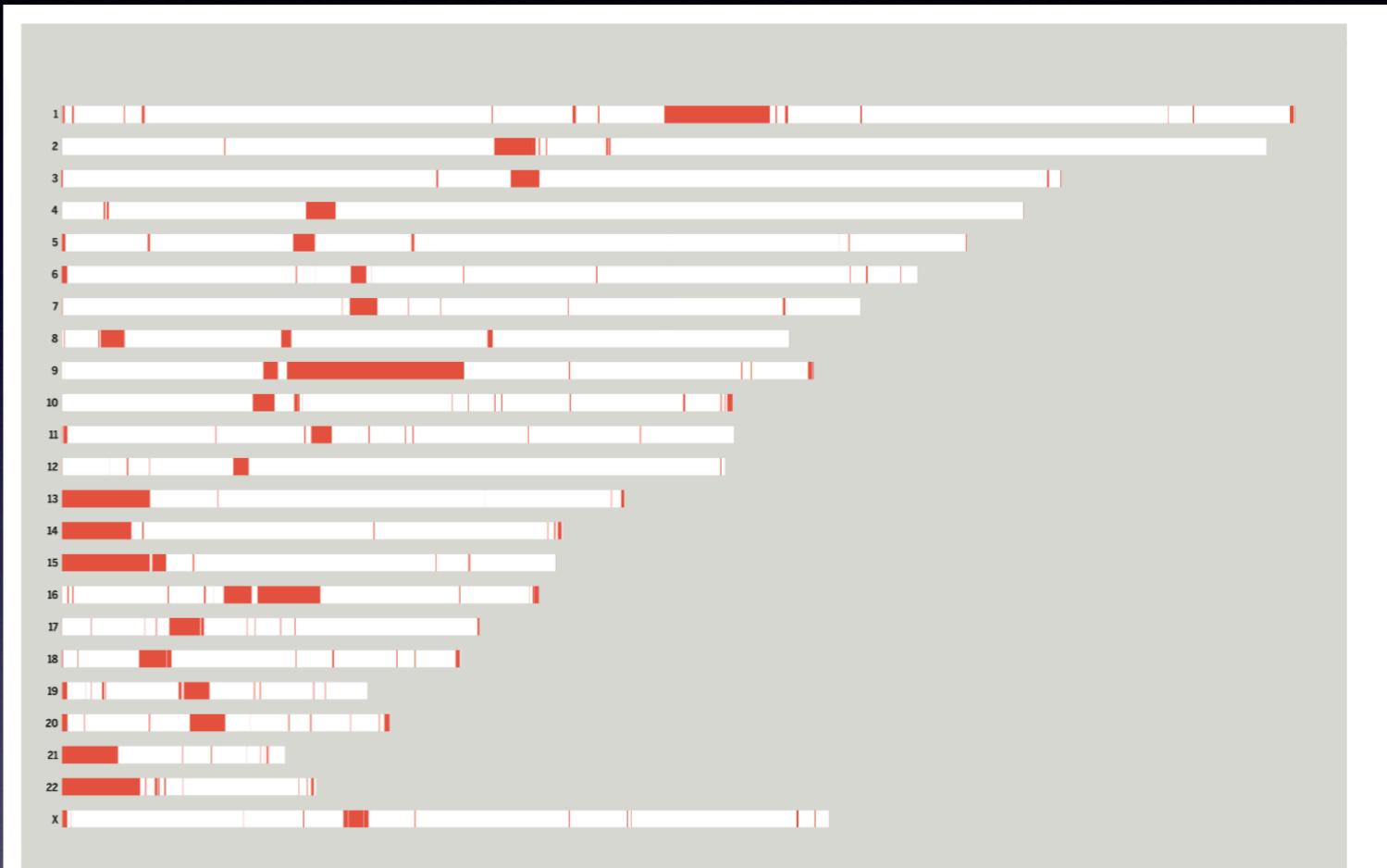
- PacBio HiFi
- ONT ultralong reads
  - Illumina Arima Genomics Hi-C (Hi-C)
- BioNano optical maps
  - single-cell DNA template strand sequencing (Strand-seq)



# CHM13 reference

Vastly improves upon the previous "gold standard" reference genome GRCh38

- Introduces nearly 200 million base pairs of sequence
- 1956 new gene predictions, 99 of which are predicted to be protein coding
- Gapless assemblies for all chromosomes except Y
- Corrects errors in the prior reference
- Resolves highly repetitive/complex regions



GRAPHIC: V. ALTOUNIAN/SCIENCE; DATA: T2T CONSORTIUM

Filling the gaps

Laura M. Zahn

Science, 376 (6588), • DOI: 10.1126/science.abp8653

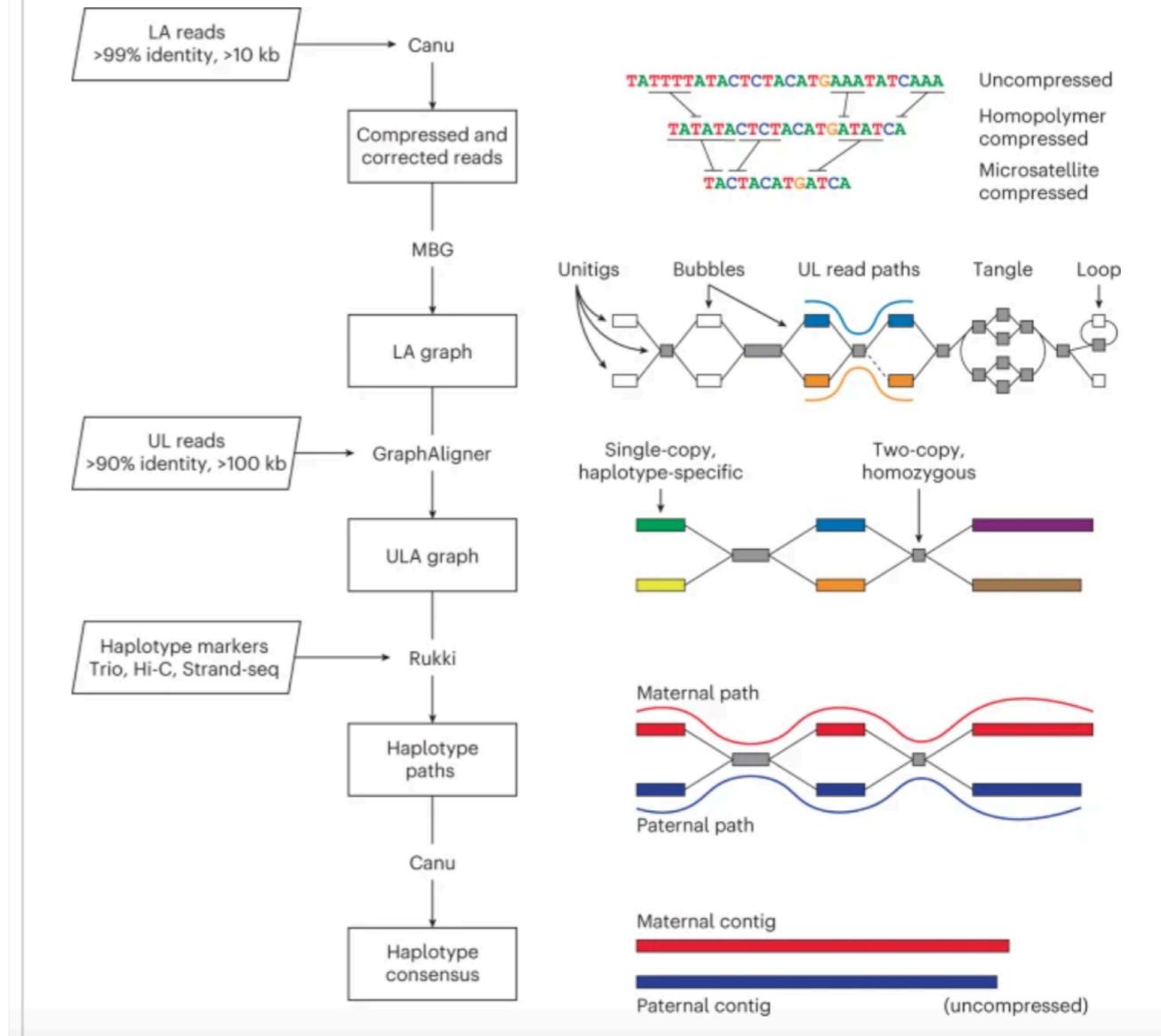
# T2T phased diploid genomes

Improve variant detection compared to previous haploid references, simplify genotyping

Pan Genomes provide greater representation of diversity

Liao, WW., Asri, M., Ebler, J. et al. A draft human pangenome reference. *Nature* 617, 312–324 (2023). <https://doi.org/10.1038/s41586-023-05896-x>

**Fig. 1: Verkko assembly workflow.**



Rautiainen, M., Nurk, S., Walenz, B.P. et al. Telomere-to-telomere assembly of diploid chromosomes with Verkko. *Nat Biotechnol* (2023). <https://doi.org/10.1038/s41587-023-01662-6>

# Limitations of long reads

- Cost\*
- Throughput\*
- Accuracy\*
- DNA amount required\*
- DNA quality required
- Data storage footprint

\*This is rapidly changing

## **Sample Input Requirements For Long Reads**

<b>Instrument</b>	<b>Initial DNA input</b>	<b>Latest DNA input</b>
Pacbio Sequel HiFi	10ug	1ug
PacBio Revio	2ug	500ng (SPRQ Dec 2024)
Oxford Nanopore PromethION	5ug	1ug

For smaller genomes <1Gb less DNA (300-400ng) may be sufficient

If PCR amplification is used, input may be as low as 5ng

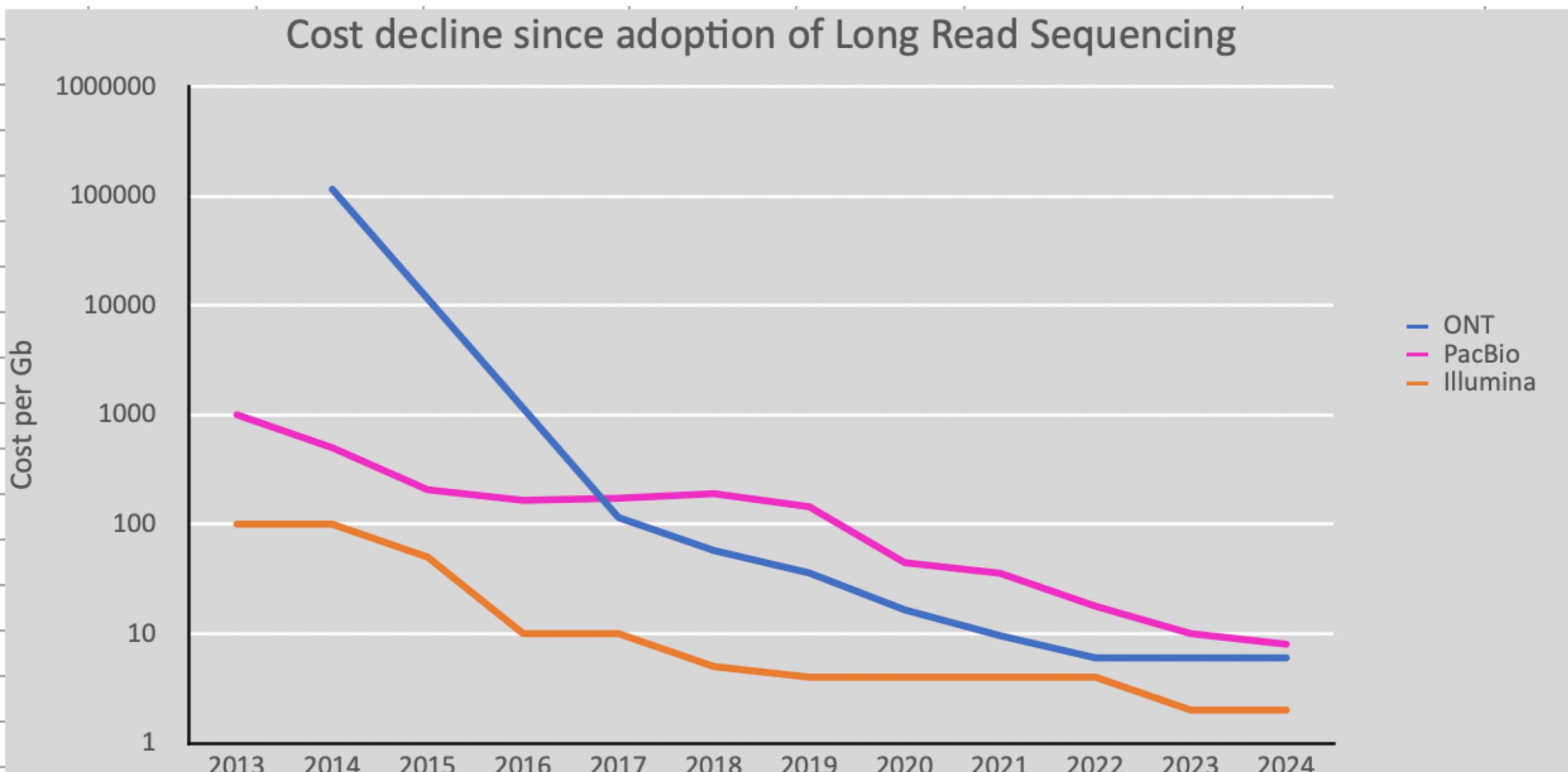
# Two “flavors” of long read sequencing



PACIFIC  
BIOSCIENCES®



# Significant advances in long read sequencing over last 11 years



## Platform Comparisons

Instrument	Avg Readlength	Yield (Gb)	Cost per Gb
Illumina NovaSeq X	300bp	8000	\$2.00
PacBio Sequel II	15-20kb	30	\$18.00
PacBio Revio	15-20kb	120	\$8.00
Oxford Nanopore PromethION	40-80kb	150	\$5.00



PACIFIC  
**BIOSCIENCES**<sup>®</sup>



# Pacific Biosciences Revio

Similar in size to Sequel

25M ZMW (1M Sequel I, 8M Sequel II)

Main focus is HiFi data

Runs 4 chips in parallel

Estimated up to 3Tb of HiFi data per week

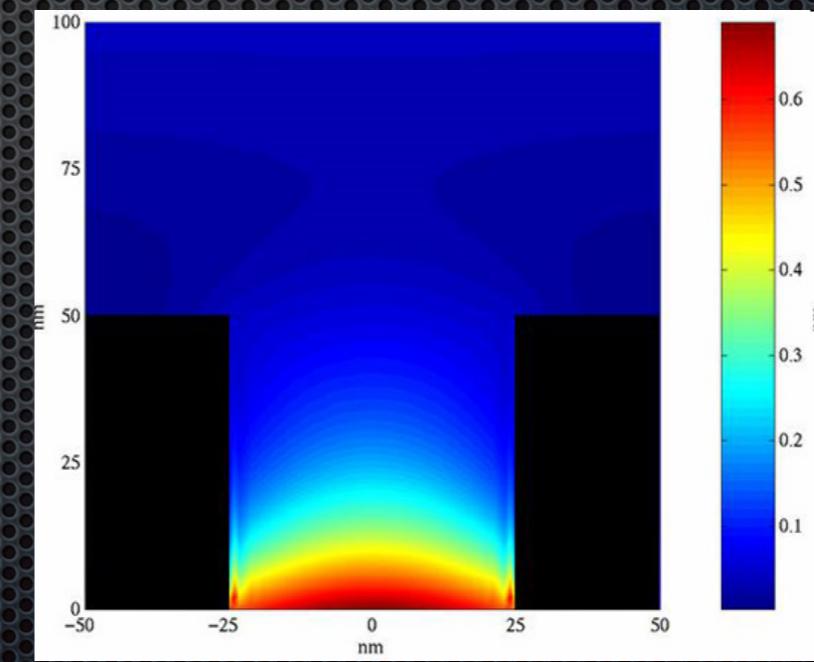
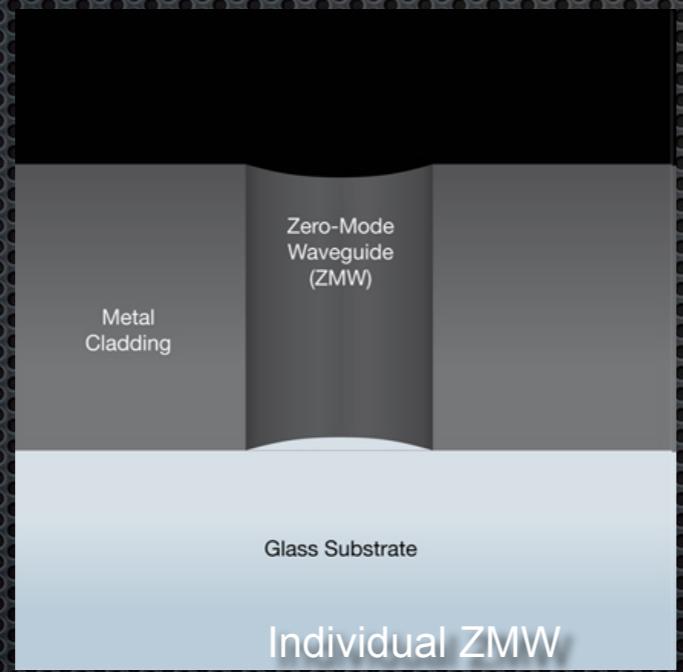


# Zero-Mode Waveguides Are the Observation Windows

DNA sequencing is performed on SMRT™ Cells, each containing tens of thousands of zero-mode waveguides (ZMWs)

A ZMW is a cylindrical hole, hundreds of nanometers in diameter, perforating a thin metal film supported by a transparent substrate

The ZMW provides a window for observing DNA polymerase as it performs sequencing by synthesis

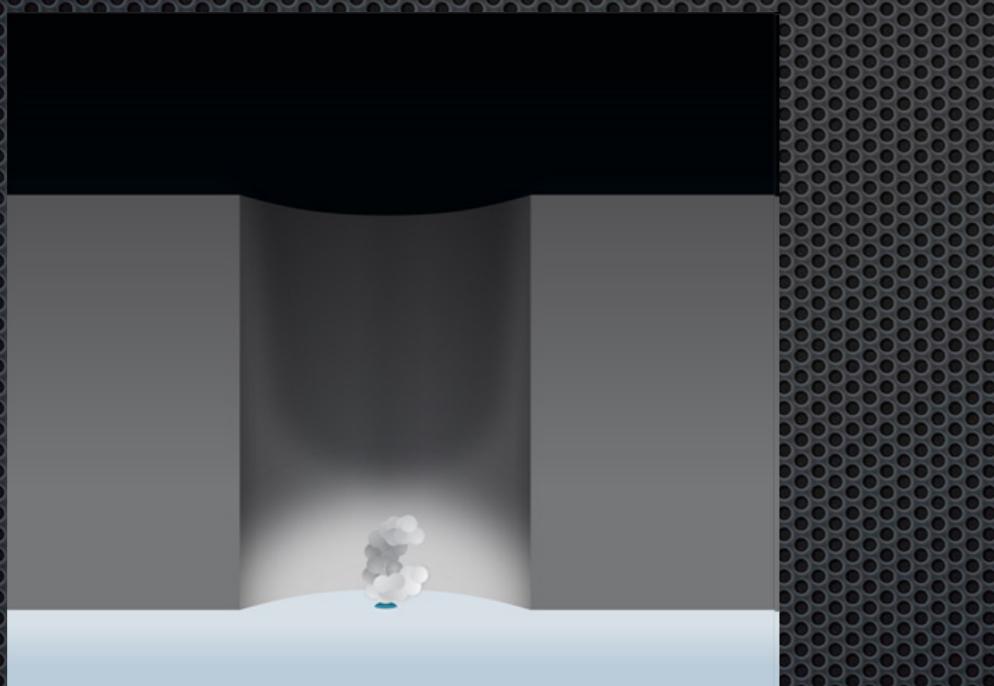


Laser light illuminates the ZMW

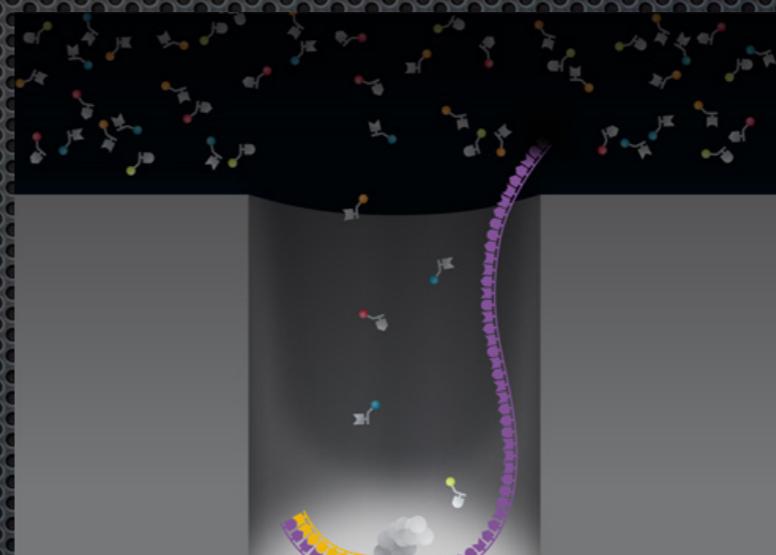
# DNA Polymerase as a Sequencing Engine

A single DNA polymerase molecule is attached to the bottom of the ZMW

A single incorporation event can be identified against the background of fluorescently labeled nucleotides



ZMW with DNA polymerase

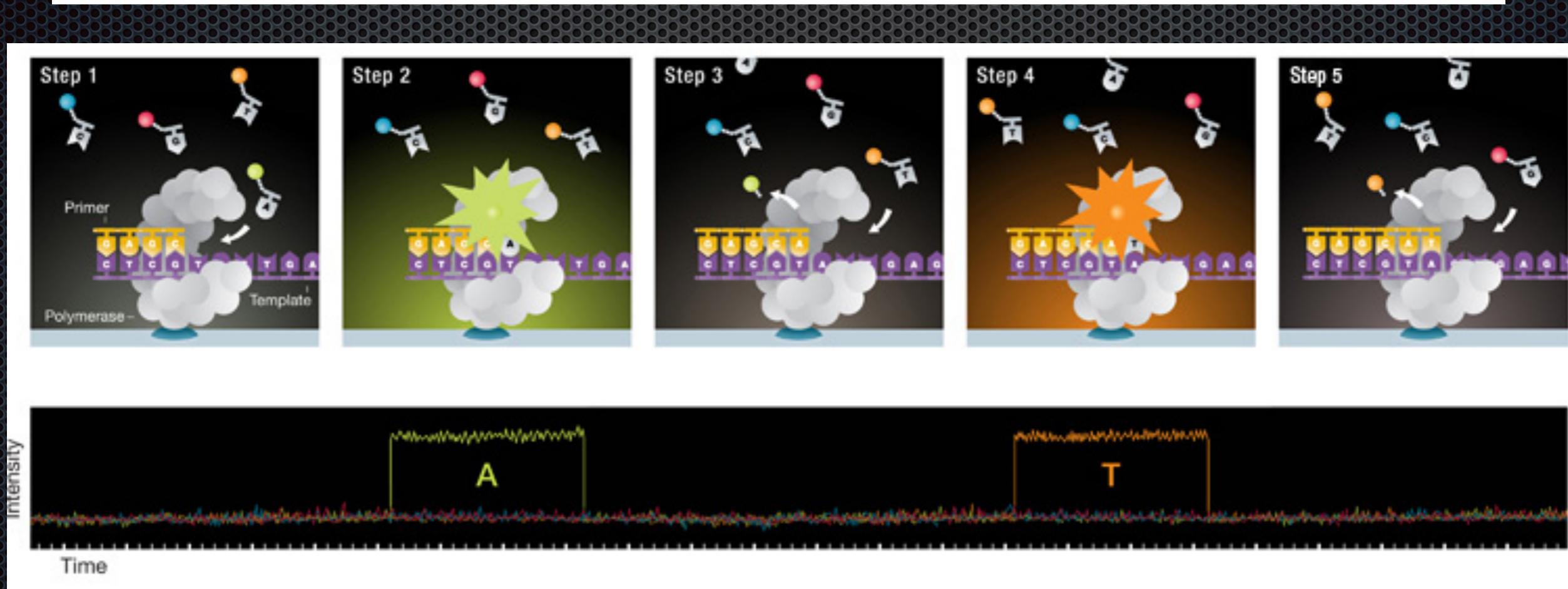


ZMW with DNA polymerase and phospholinked nucleotides

# Processive Synthesis with Phospholinked Nucleotides

Enzymatic incorporation of the labeled nucleotide creates a flash of light, which is captured by the optics system and converted into a base call with associated quality metrics using optimized algorithms

To generate consensus sequence from the data, an assembly process aligns the different fragments based on common sequences



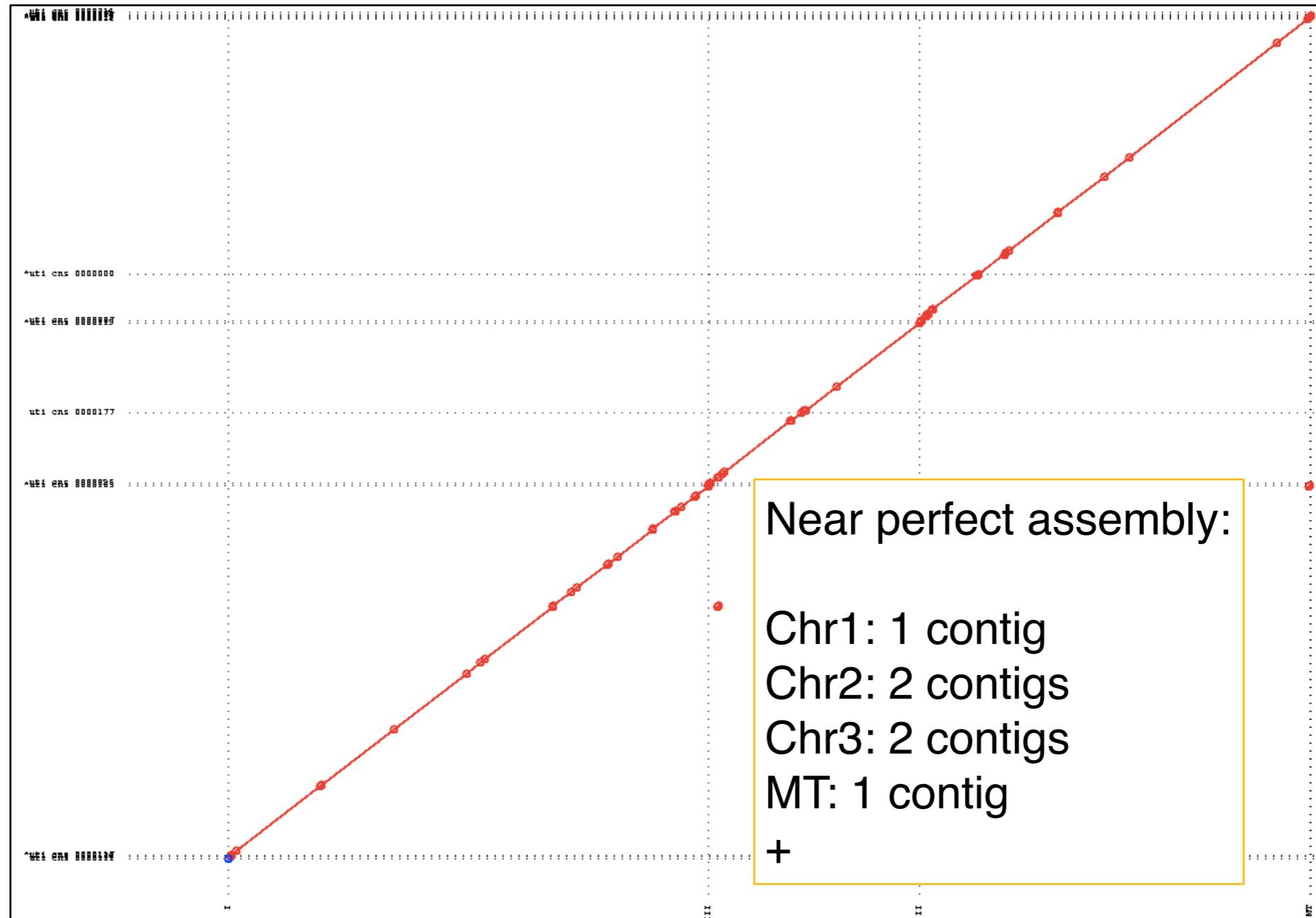
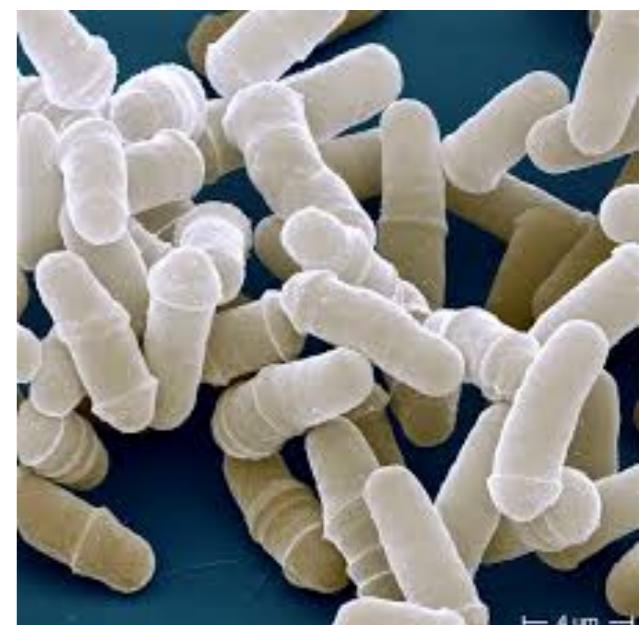
# S. pombe dg21

ASM294 Reference sequence

- 12.6Mbp; 3 chromo + mitochondria; N50: 4.53Mbp

PacBio assembly using HGAP + Celera Assembler

- 12.7Mbp; 13 non-redundant contigs; N50: 3.83Mbp; >99.98% id

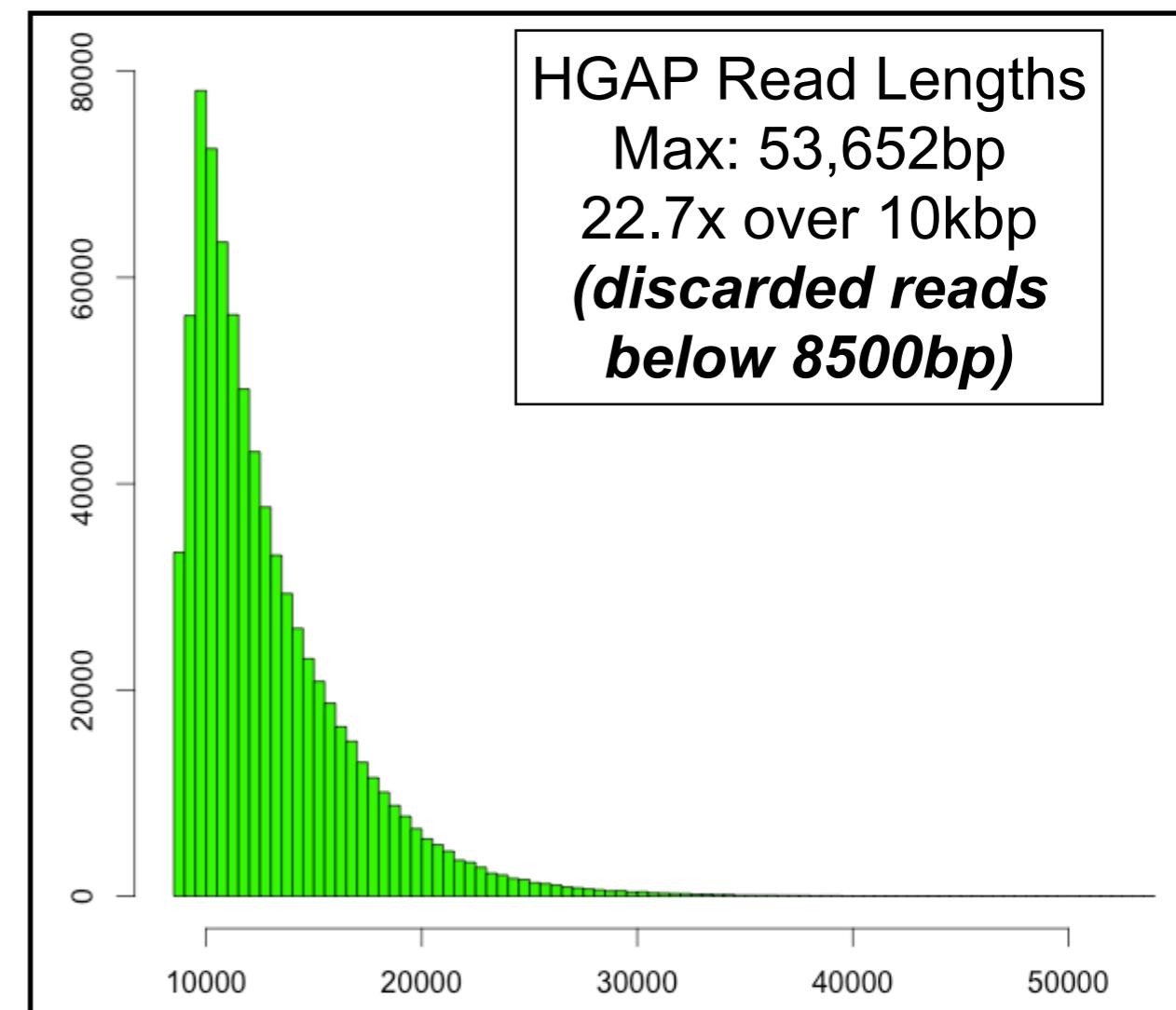


# O. sativa pv Indica (IR64)



Genome size: ~370 Mb  
Chromosome N50: ~29.7 Mbp

Assembly	Contig NG50
MiSeq Fragments 25x 456bp (3 runs 2x300 @ 450 FLASH)	19 kbp
“ALLPATHS-recipe” 50x 2x100bp @ 180 36x 2x50bp @ 2100 51x 2x50bp @ 4800	18 kbp
HGAP + CA 22.7x @ 10kbp	4.0 Mbp
Nipponbare BAC-by-BAC Assembly	5.1 Mbp



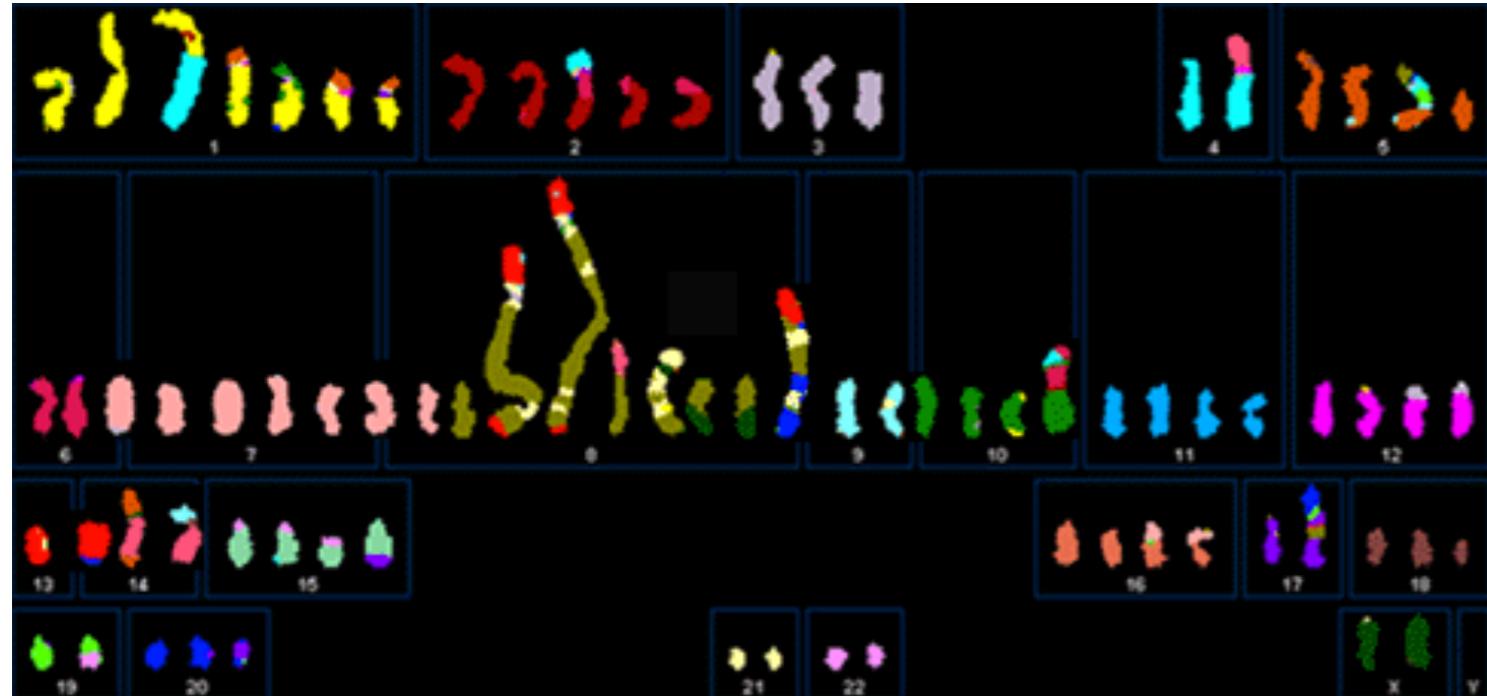
# Structural Variations in SKBR3

SKRB3 cell line was derived by G. Trempe and L. J. Old in 1970 from pleural effusion cells of a patient, a white, Caucasian female

Most commonly used Her2-amplified breast cancer cell line

Often used for pre-clinical research on Her2-targeting

therapeutics such as Herceptin (Trastuzumab) and resistance to these therapies.



(Davidson et al, 2000)



Nattestad, et al, Gen. Res. 2018

Number of sequences:

10,304

Total sequence length:

2.75 Gb

Mean: 266 kb

Max: 15 Mb

N50: 2.17 Mb

**NG50: 1.86 Mb**



Number of sequences:

748,955

Total sequence length:

2.07 Gb

Mean: 2.8 kb

Max: 61 kb

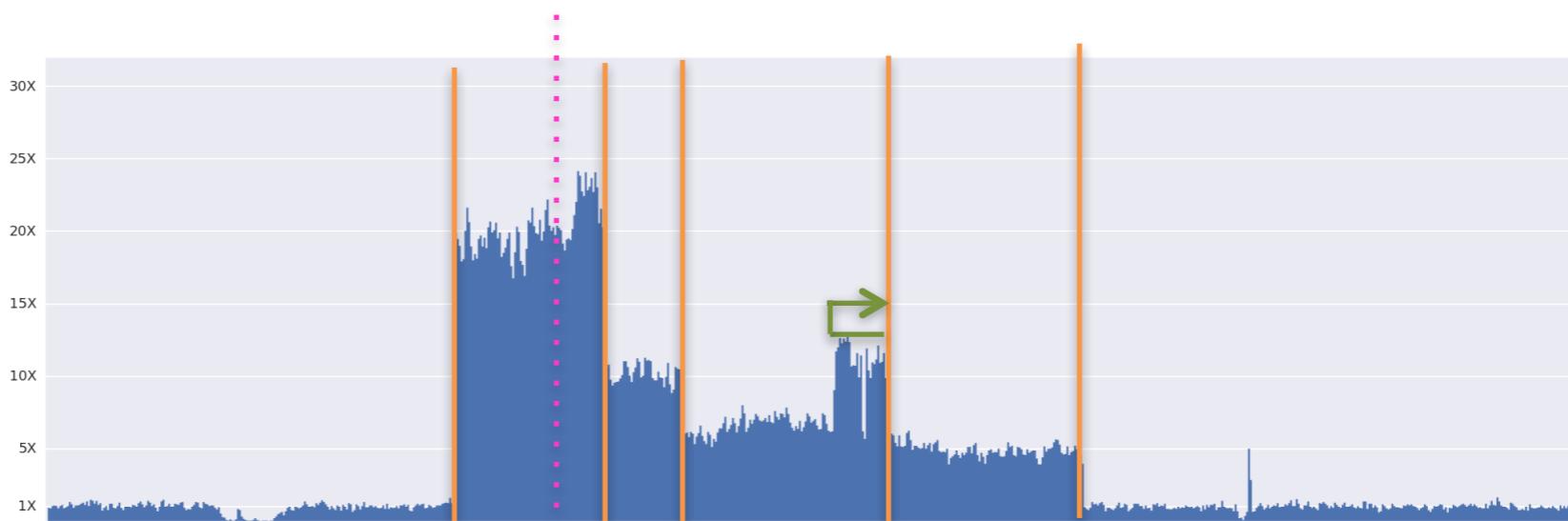
N50: 3.3 kb

**NG50: 1.9 kb**



# Her2

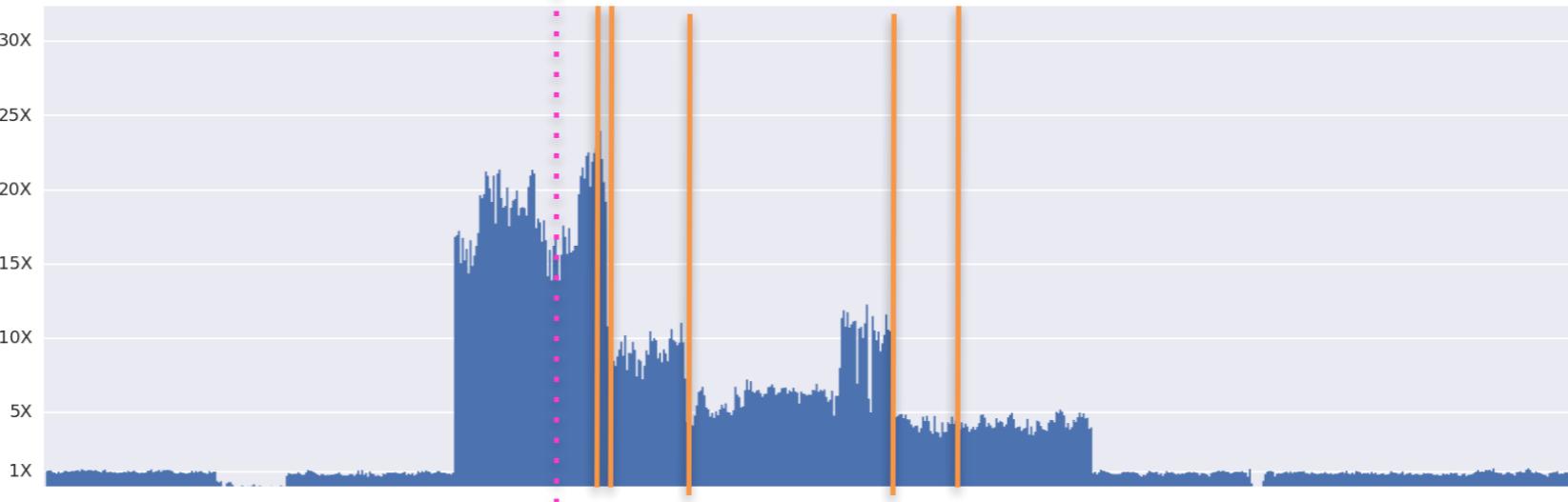
PacBio  
73X @ 10kb



# split reads

Illumina  
120X @ 100bp

295 187 91 87 60



# split reads

151,76 91 77 87

8 Mb

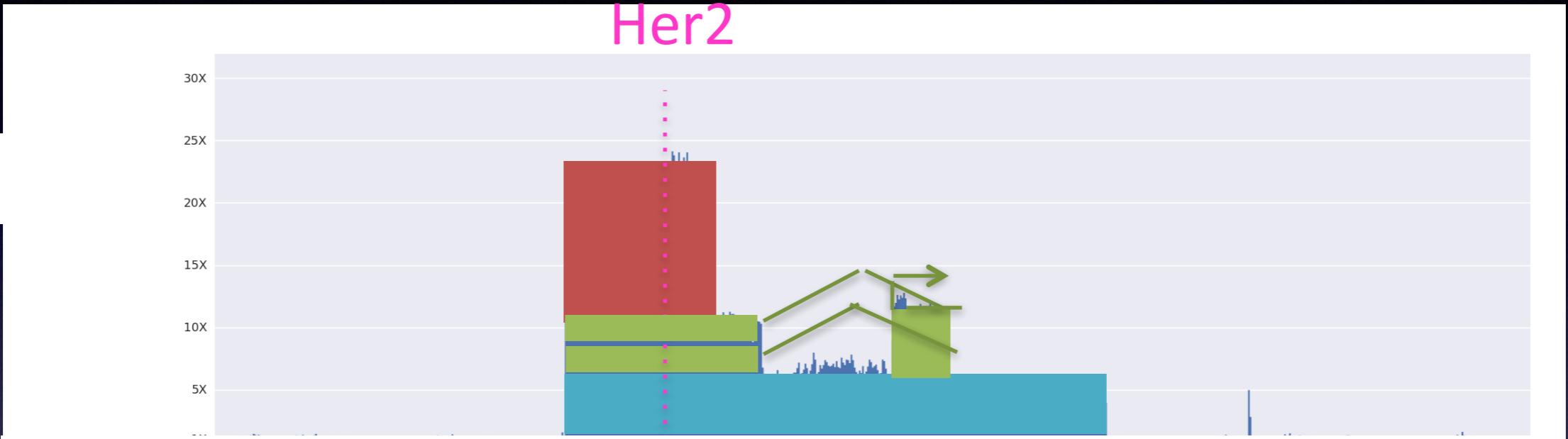
Green arrow indicates an inverted duplication.

False positive and missing Illumina calls due to mis-mapped reads (especially low complexity).

# Cancer lesion reconstruction from genomic threads

PacBio

chr17



By comparing the proportion of reads that are spanning or split at breakpoints we can begin to infer the history of the genetic lesions.

1. Healthy diploid genome
2. Original translocation into chromosome 8
3. Duplication, inversion, and inverted duplication within chromosome 8
4. Final duplication from within chromosome 8

# PacBio errors are randomly distributed

ATGCTCTCGATCGATGCTGCTAGCTAGCTACTAGCTATCCGATCCTACTGACTTACTATGCT

ATGCTGTTCGATCGATGCTGCTAGCTAGCTACTAGCTATCCGATCCTACTGACTTACTATGCT

ATGCTCTCGATCGATGCTGCTCGCTAGCTAGCTACTAGCTATCCGATCCTACTGACTTACTATGCT

ATGCTCTCGATCGATGCTGCTAGCTAGCTACTAGCTATCAGATCCTACTGACTTACTATGCT

ATGCTCTCGATCGATGCTGCTAGCTAGCTACTAGCTATCGGATCCTACTGACTTACTATGCT

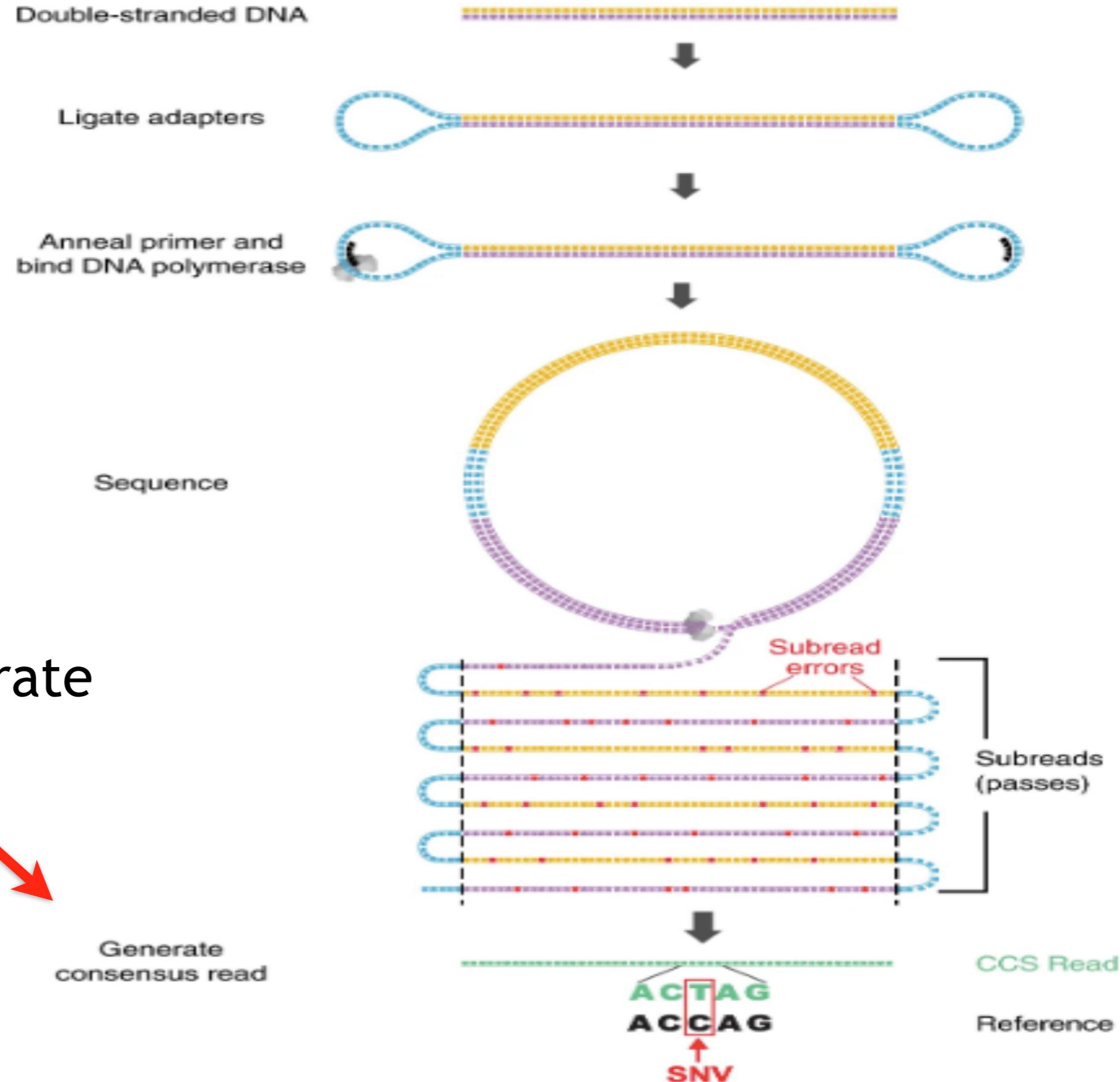
ATGCTCTCGATCGATGCTGCTAGCTAGCTACTAGCTATCCGATCCTACTGACTTACTATGGT



ATGCTCTCGATCGATGCTGCTAGCTAGCTACTAGCTATCCGATCCTACTGACTTACTATGCT

Enough coverage makes error drop out

PacBio CCS  
“HiFi” for longer  
(~15-20kb)  
fragments

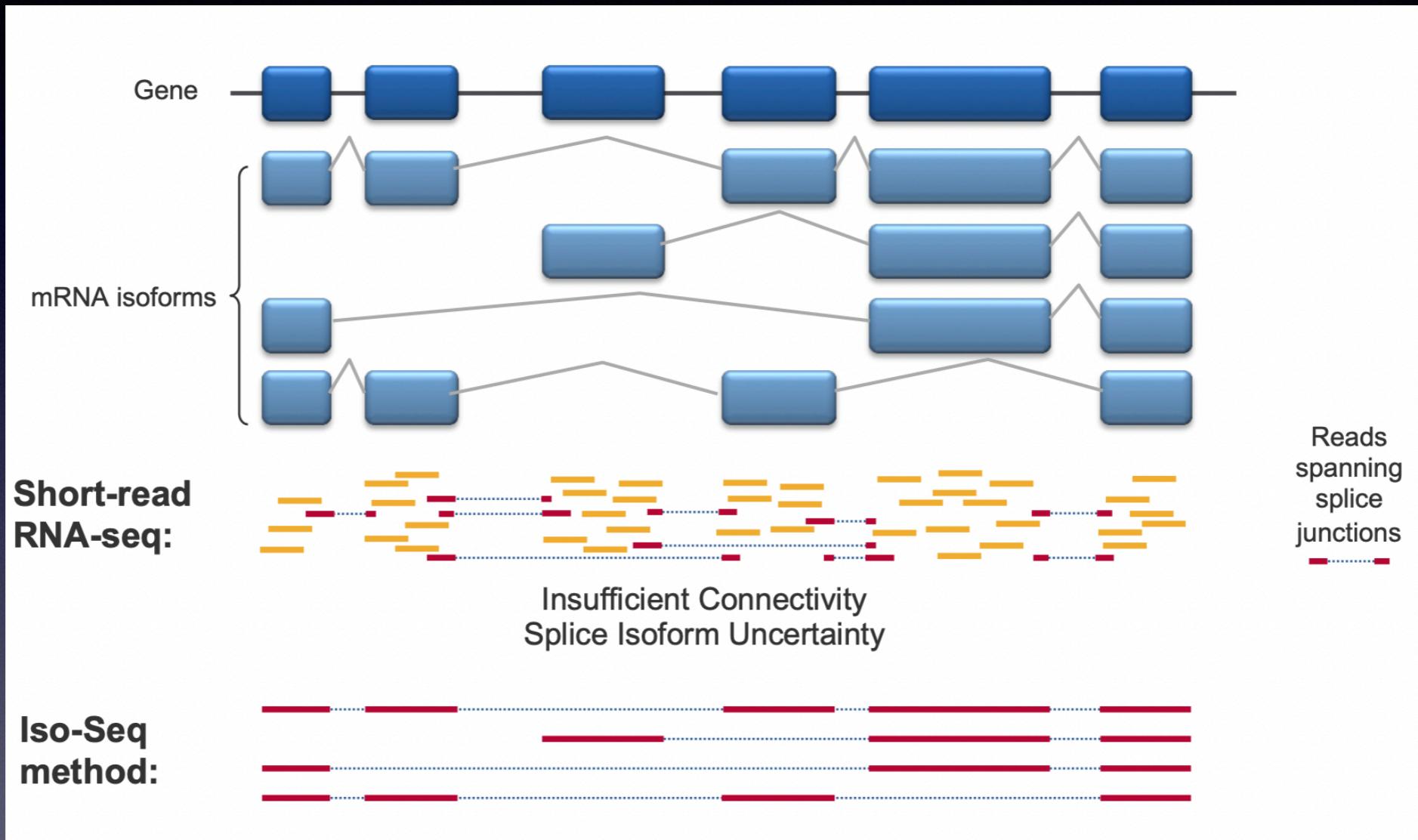


99.99% Accurate

Generate  
consensus read

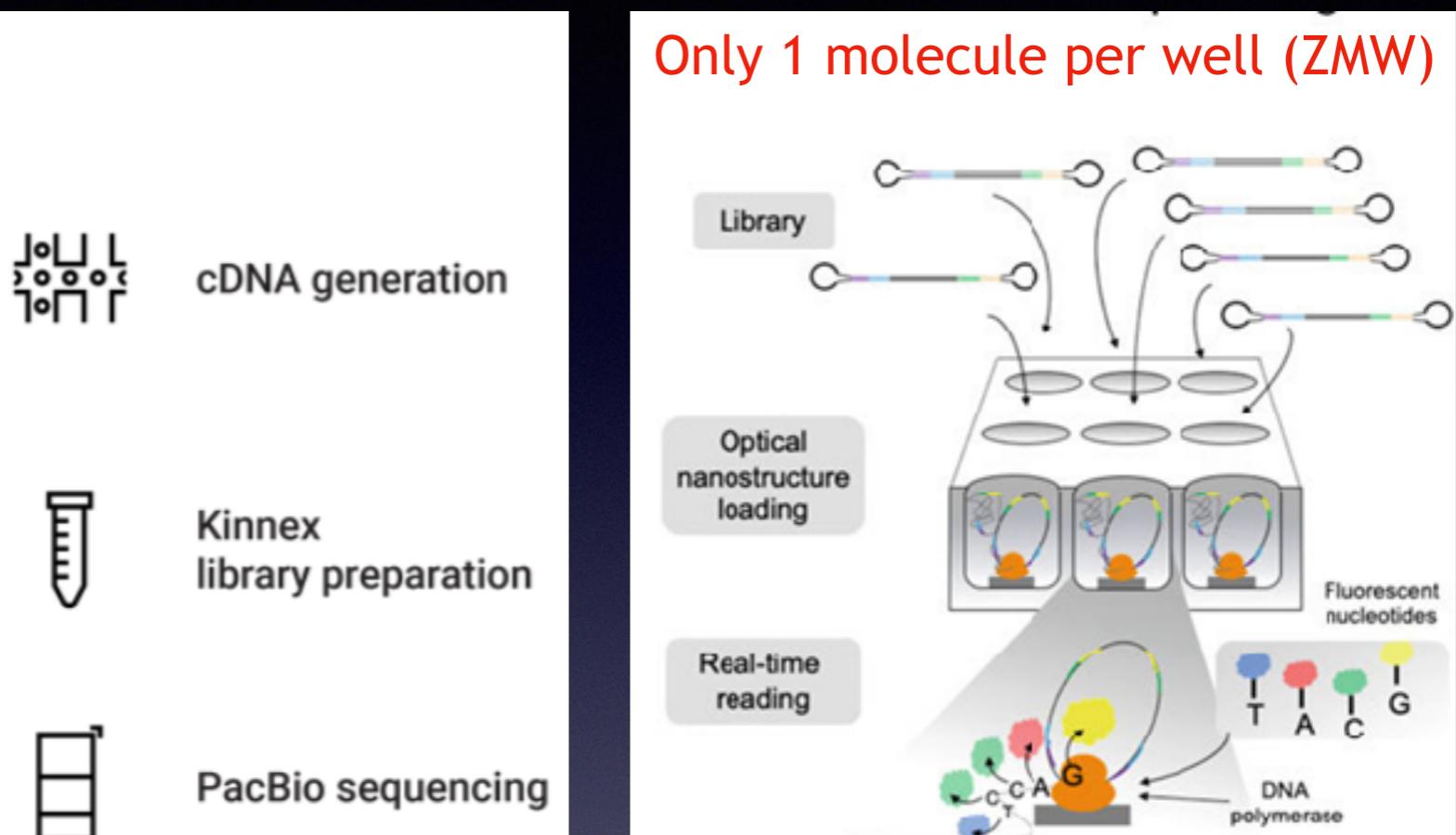
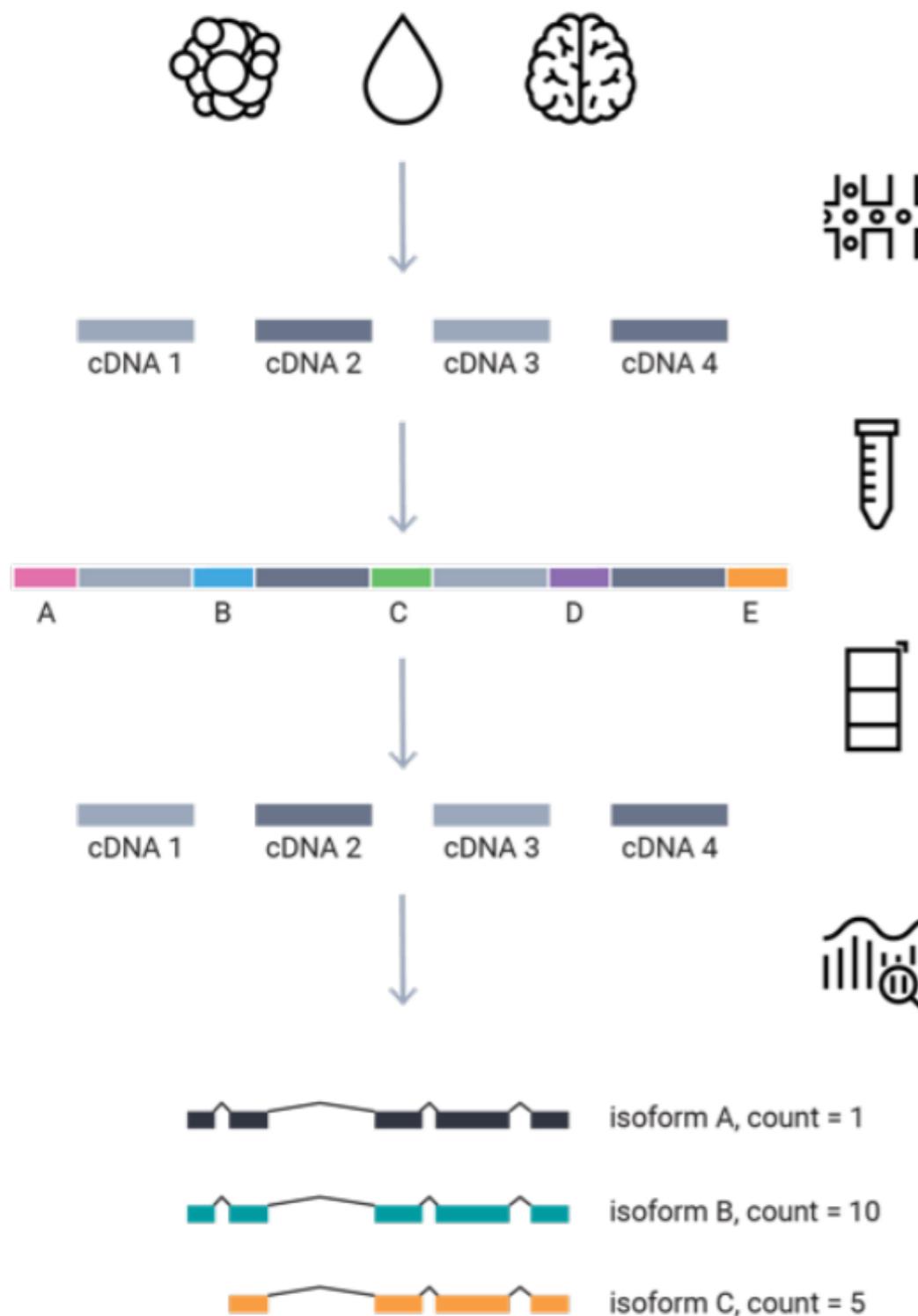
From Wenger et al (2019) Nature Biotechnology

# Benefits of long read transcripts



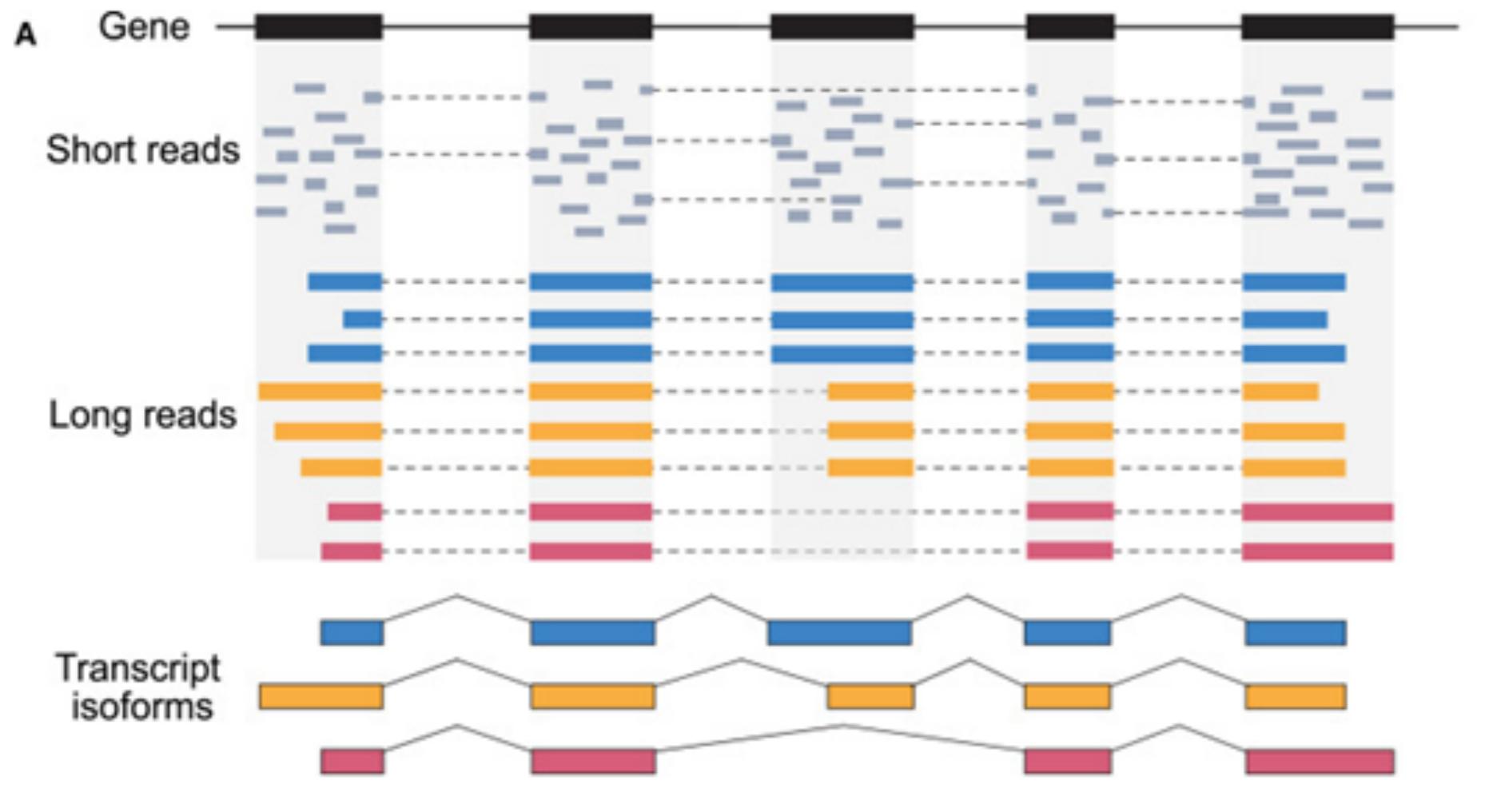
Long read transcripts provide complete isoform information, enables identification of alternative splicing, fusion events, and allows for isoform level phasing

# PacBio Kinnex transcript sequencing



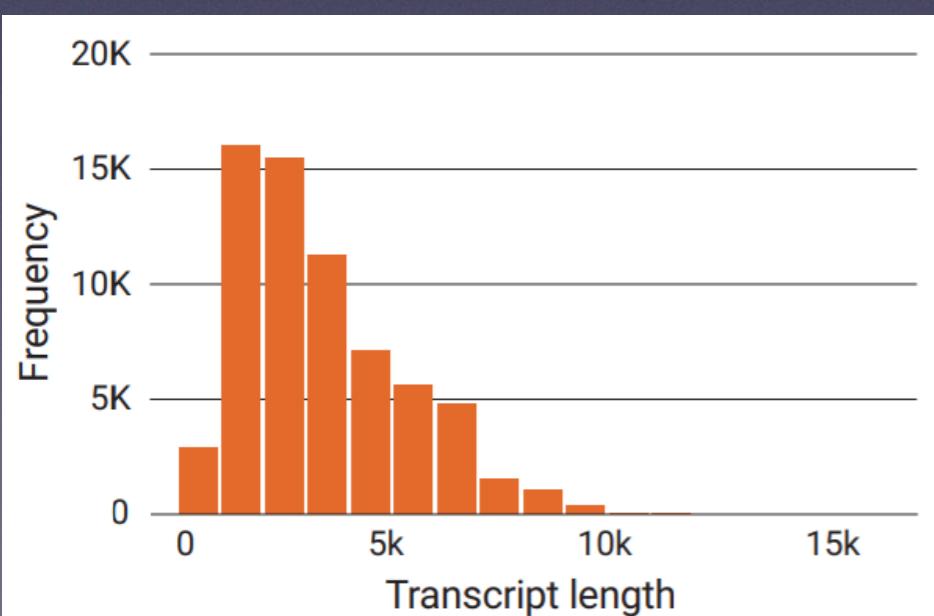
**25M wells (ZMW) per flowcell**

# PacBio Kinnex transcript sequencing



Fewer reads are needed to construct introns

Thus fewer total reads are needed to discover and assemble genes



Full length transcripts are preserved  
Extra long RNAs are discoverable



# PromethION



24 independent flowcells

500bp/s sequencing speed

3000 pores per flowcells = 144,000 pores (fully loaded) (MinION cells 512 pores)

On board single or duplex basecalling

>140Gb in CSHL hands

>100M cDNA reads

Up to ~5 Tb fully loaded in one week

Sequencing "flavors" include:

Ligation based - standard methods for gDNA

Q20 - enables higher accuracy including duplex

Barcodeing - allows multiplexing up to 96 sample

16S - enables 16S metagenome sequencing

PCR sequencing - long-range PCR for low mass samples

Cas9 - enables Cas9 mediated target enrichment

Rapid - enables library prep <2hrs w/o mechanical shearing

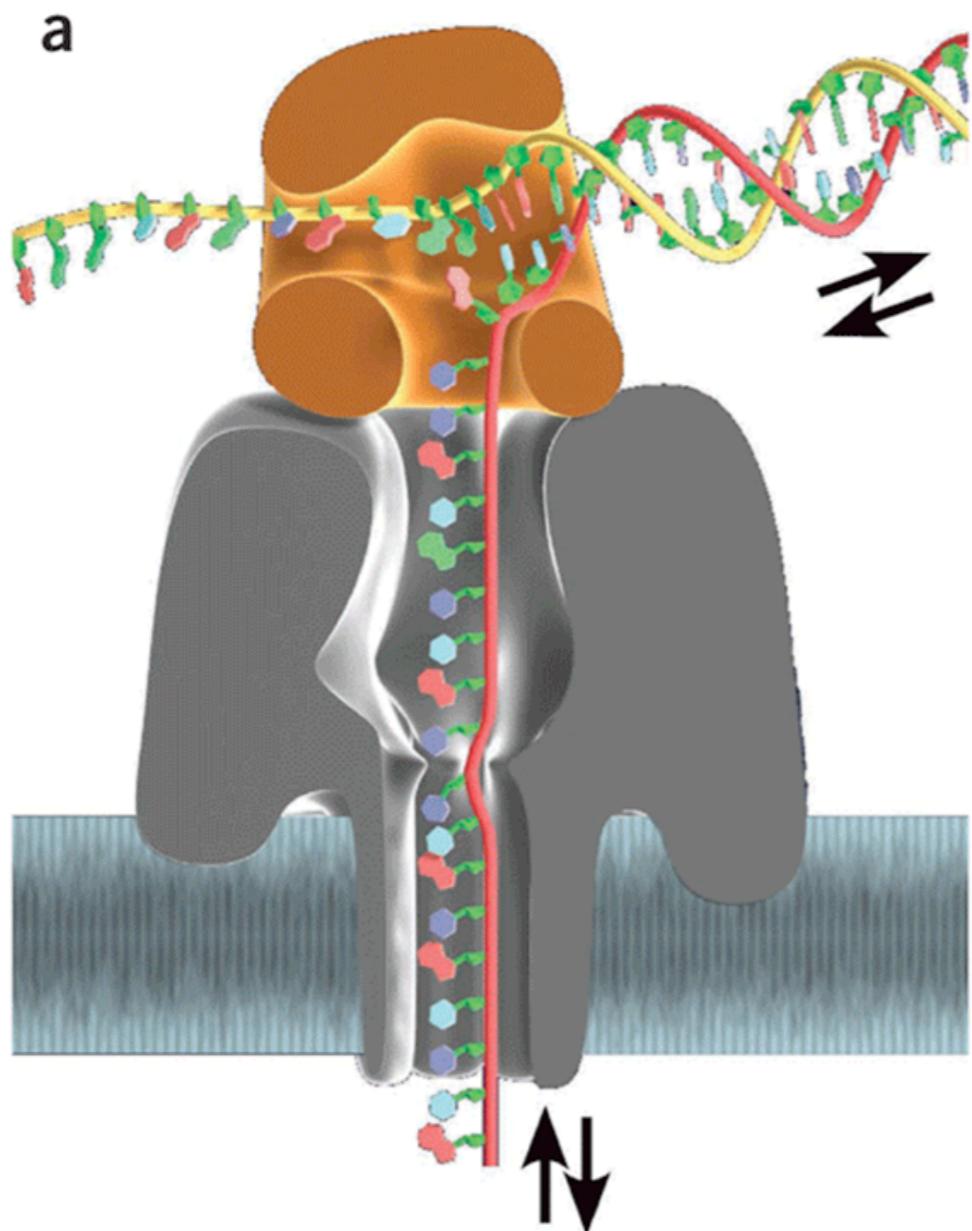
Ultra-long - enables N50s up to 100kb

Native barcodeing - PCR-free barcodeing to preserve epigenetic marks

Field kit - enables sequencing in the field w/o cold chain

Short fragment - enables sequencing of fragments <1000bp

# Oxford Nanopore relies on CsgG and a non-destructive motor protein



Cis side voltage drives DNA through pore

Motor protein mediates DNA unwinding  
and translocation speed

Ions flow through the pore to change  
membrane potential

Small changes in measured voltage are  
translated into k-mers

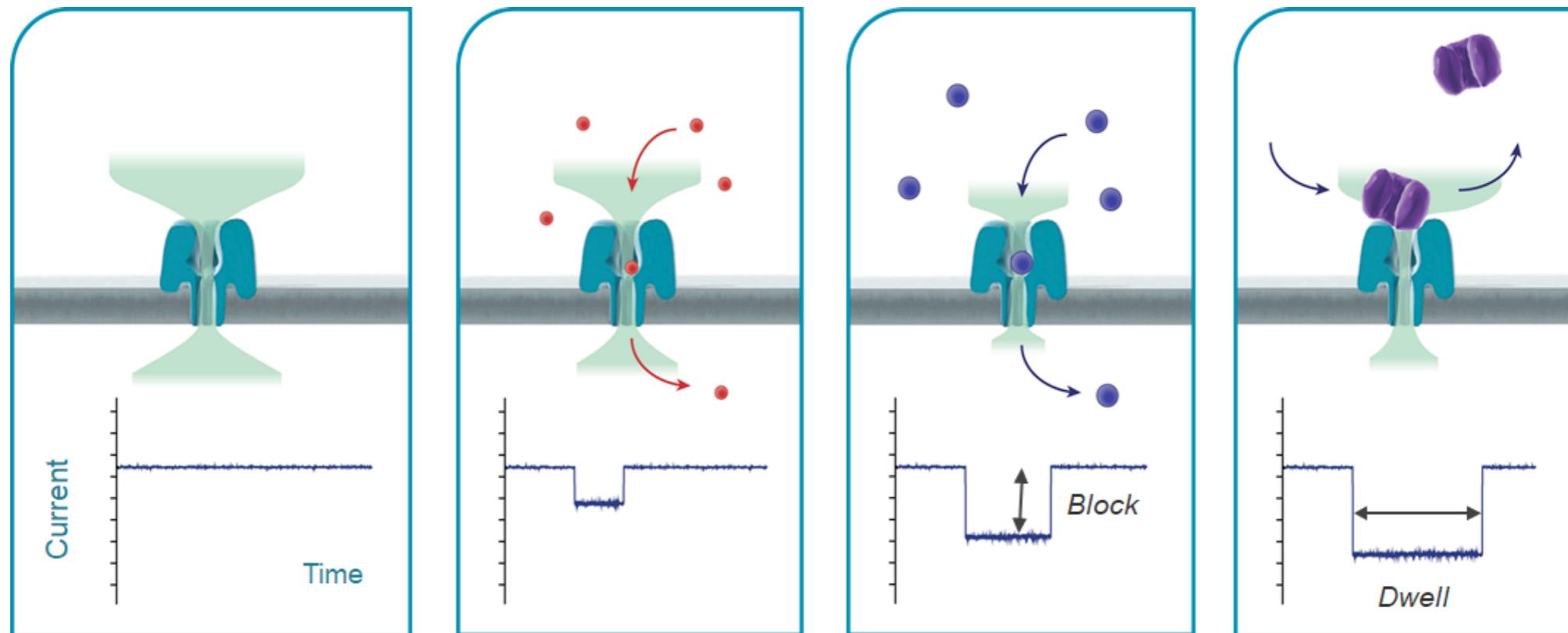
# Nanopore Sensing Summary

Nanopore = ‘very small hole’

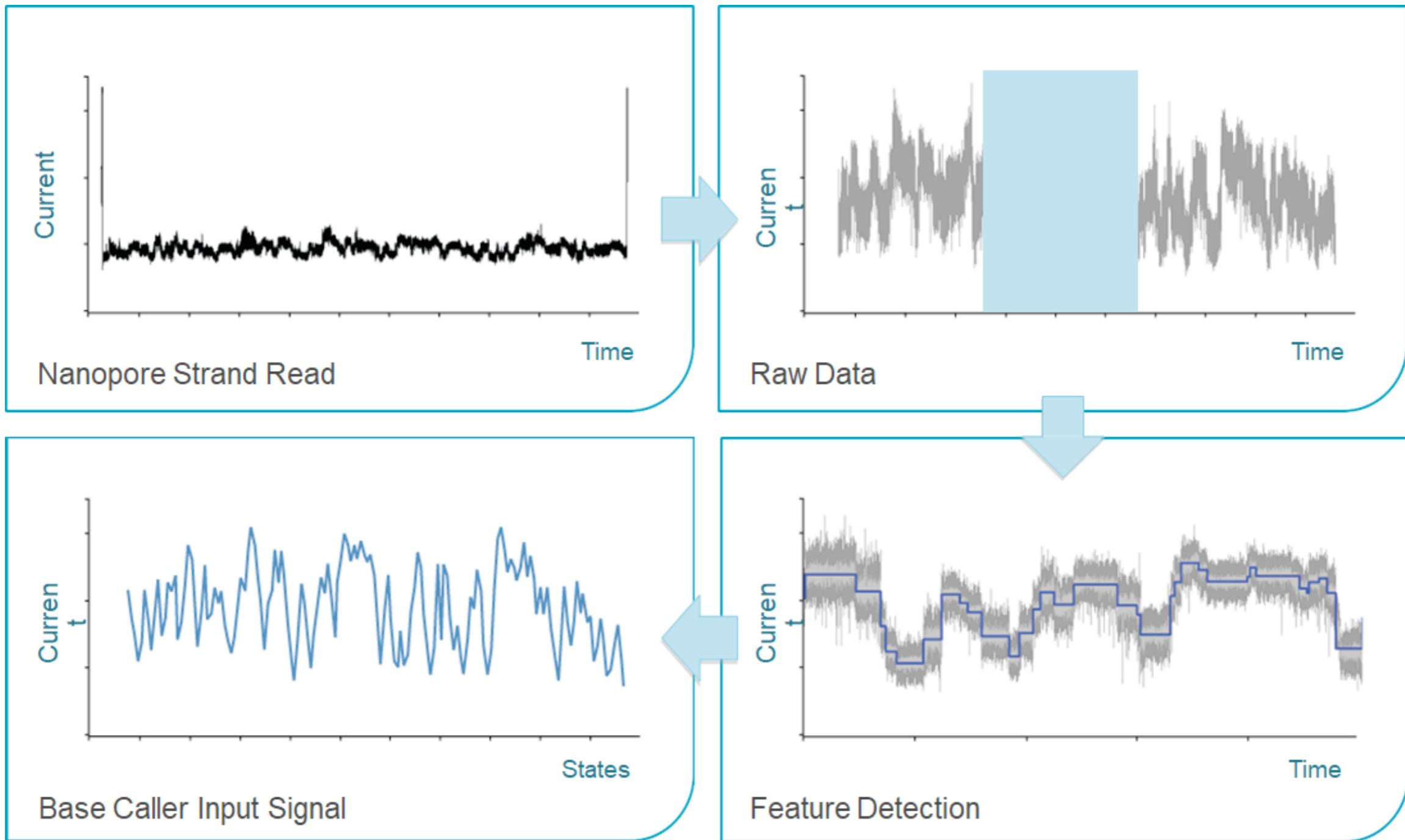
Ionic current flows through the pore Introduce analyte of interest into the pore

Identify target analyte by the characteristic disruption or block to the electrical current

Block or ‘State’, Dwell, Noise



# Raw Data and Data Reduction



# Nanopore errors are (mostly) randomly distributed

ATGCTCTCGATCGATGCTGCTAGCTAGCTAGCTTTTTCCGATCCTACTGACTTACTATGCT

ATGCTGTTCGATCGATGCTGCTAGCTAGCTAGCTTTTT CCGATCCTACTGACTTACTATGCT

ATGCTCTCGATCGATGCTGCTCGCTAGCTAGCTAGCTTTTTTT CCGATCCTACTGACTTACTATGCT

ATGCTCTCGATCGATGCTGCTAGCTAGCTAGCTAGCTTTTTTCAGATCCTACTGACTTACTATGCT

ATGCTCTCGATCGATGCTGCTAGCTAGCTAGCTAGCTTTTT CCGATCCTACTGACTTACTATGCT

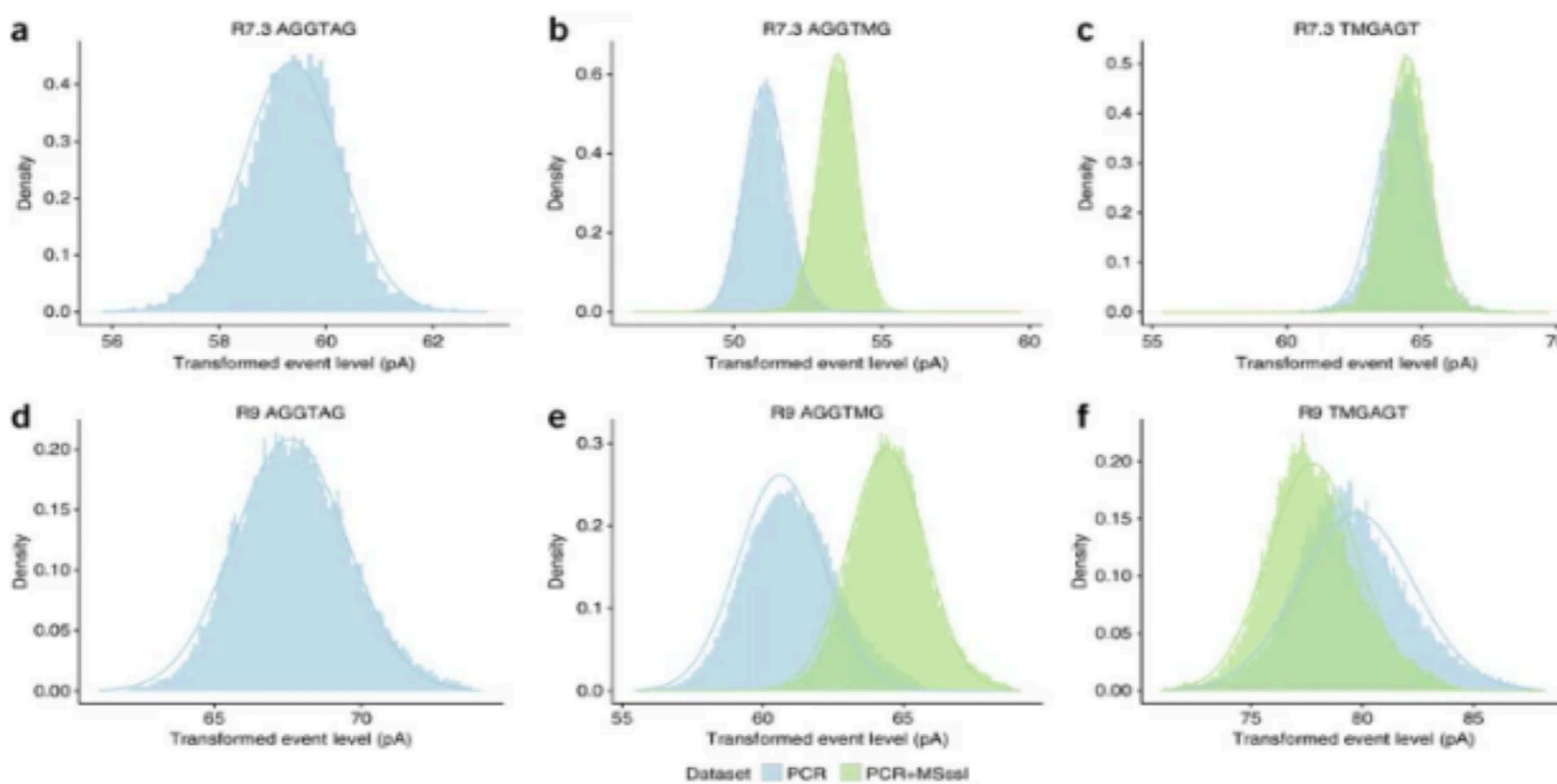
ATGCTCTCGATCGATGCTGCTAGCTAGCTAGCTAGCTTTTT CCGATCCTACTGACTTACTATGGT



ATGCTCTCGATCGATGCTGCTAGCTAGCTAGCTTTTT CCGATCCTACTGACTTACTATGCT

Enough coverage makes error (mostly) drop out

### Figure 1: Differences in event distribution between methylated and unmethylated 6-mers in nanopore sequencing data.



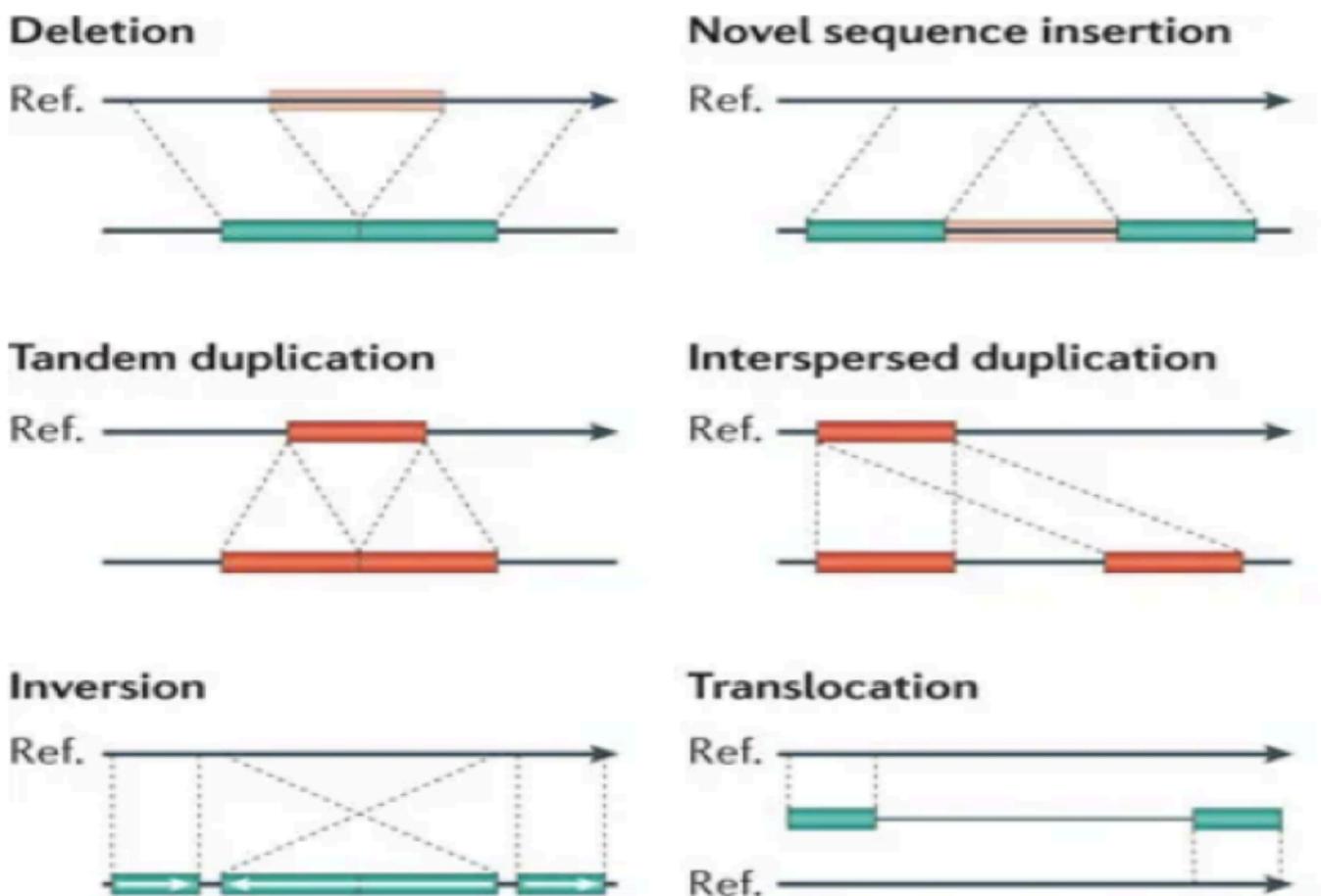
(a–c) Event distribution for the three sample 6-mers AGGTAG (a), AGGTMG (b) and TMGAGT (c) using R7.3 pore data from PCR-amplified *E. coli* DNA. 'M' represents 5-mC. (d–f) Data from the R9 pore for the AGGTAG (d), AGGTMG (e) and TMGAGT (f) 6-mers from PCR-amplified *E. coli* DNA. Data from M.Sssi-methylated DNA are in green, and data from untreated DNA are in blue. Solid lines indicate Gaussian fits. Events are from the template strand and have been transformed to account for per-read differences with respect to the model (Supplementary Note).

Nanopores are sensitive to DNA modifications which alter the electrical current as the nucleotide transduces the pore. Simpson et al trained a HMM to distinguish signal differences between 5methylC and unmethylated cytosines. They used synthetically modified 6 mers from bacteria to test the shift of methylation at different positions.

# Significance of Structural Variants

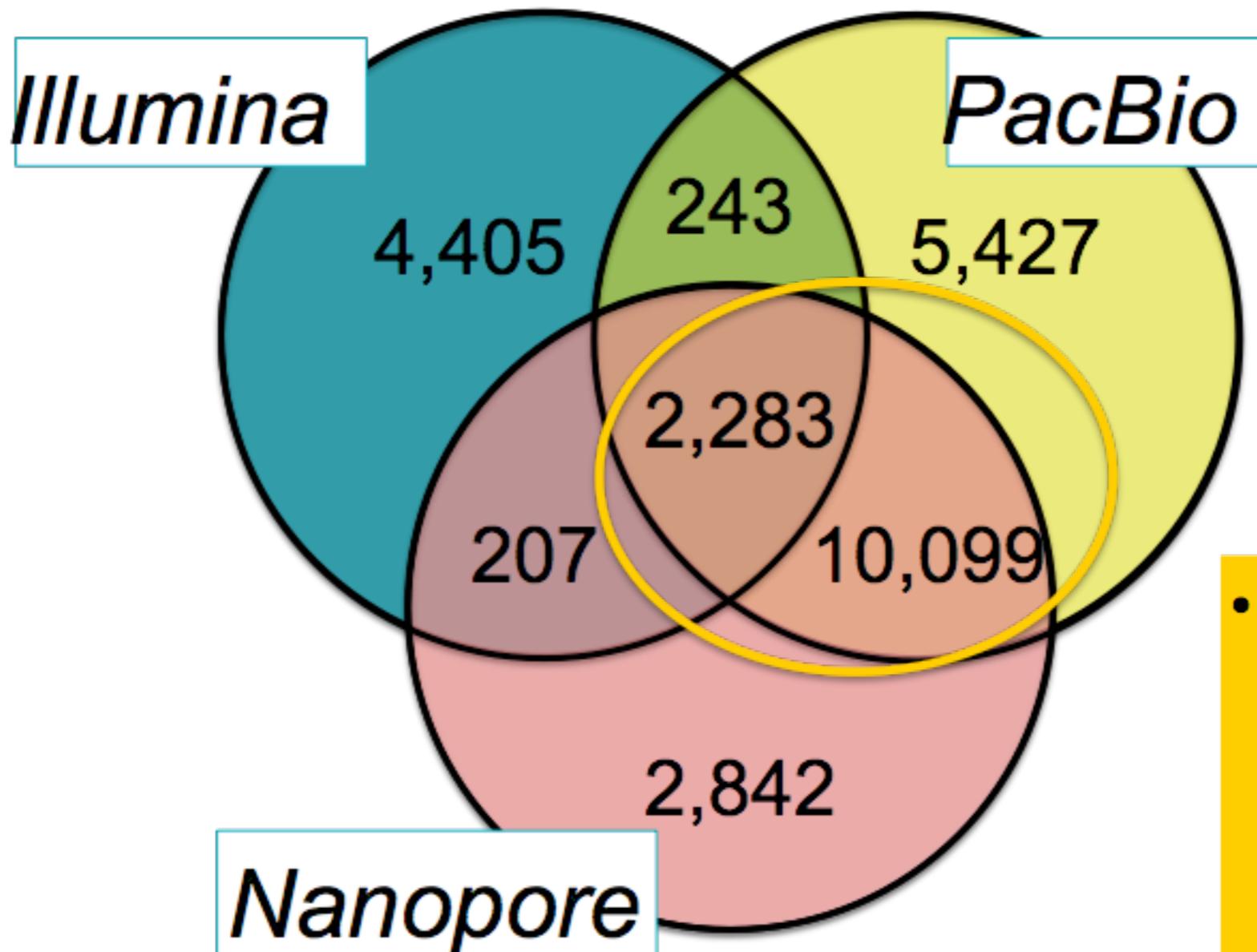
- Often missed by short read approaches
- May uncover causative mutations underlying disease, especially where typical screening has failed
- SVs account for much more variation in the genome than SNPs
- SVs can alter methylation and gene expression
- However, the functional impact of most SVs remains unknown

**Figure 1: Classes of structural variation.**



Alkan, C., Coe, B. & Eichler, E.  
Genome structural variation  
discovery and genotyping. Nat Rev  
Genet 12, 363–376 (2011)  
doi:10.1038/nrg2958

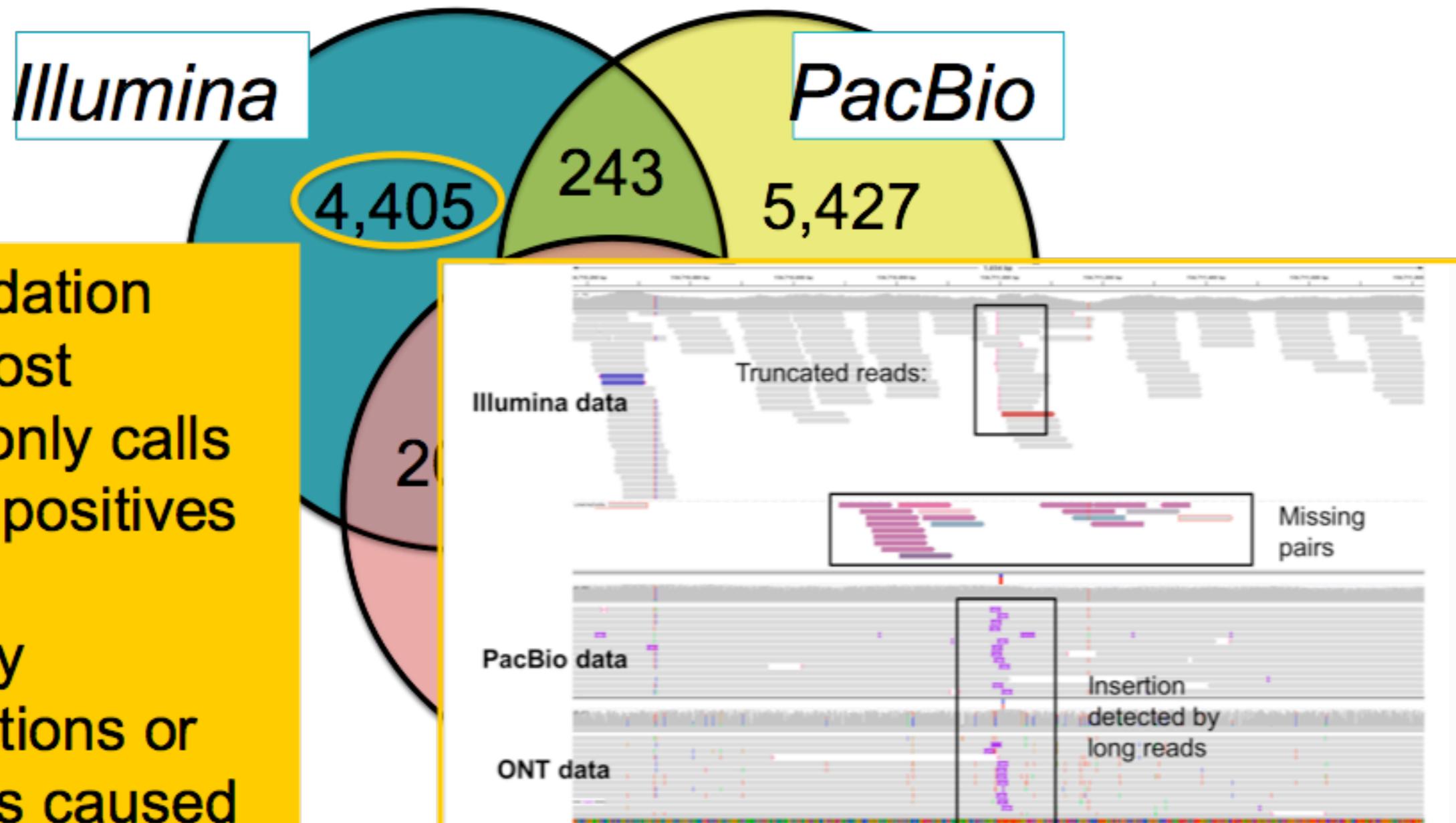
# Structural Variant Comparison of SKBR3



(Hicks et al, 2006)

- Strong concordance between long read platforms
- Substantially more variants than detected by short reads

# Structural Variant Comparison of SKBR3



- PCR validation shows most Illumina-only calls are false positives
- Especially translocations or inversions caused by smaller insertions or deletions

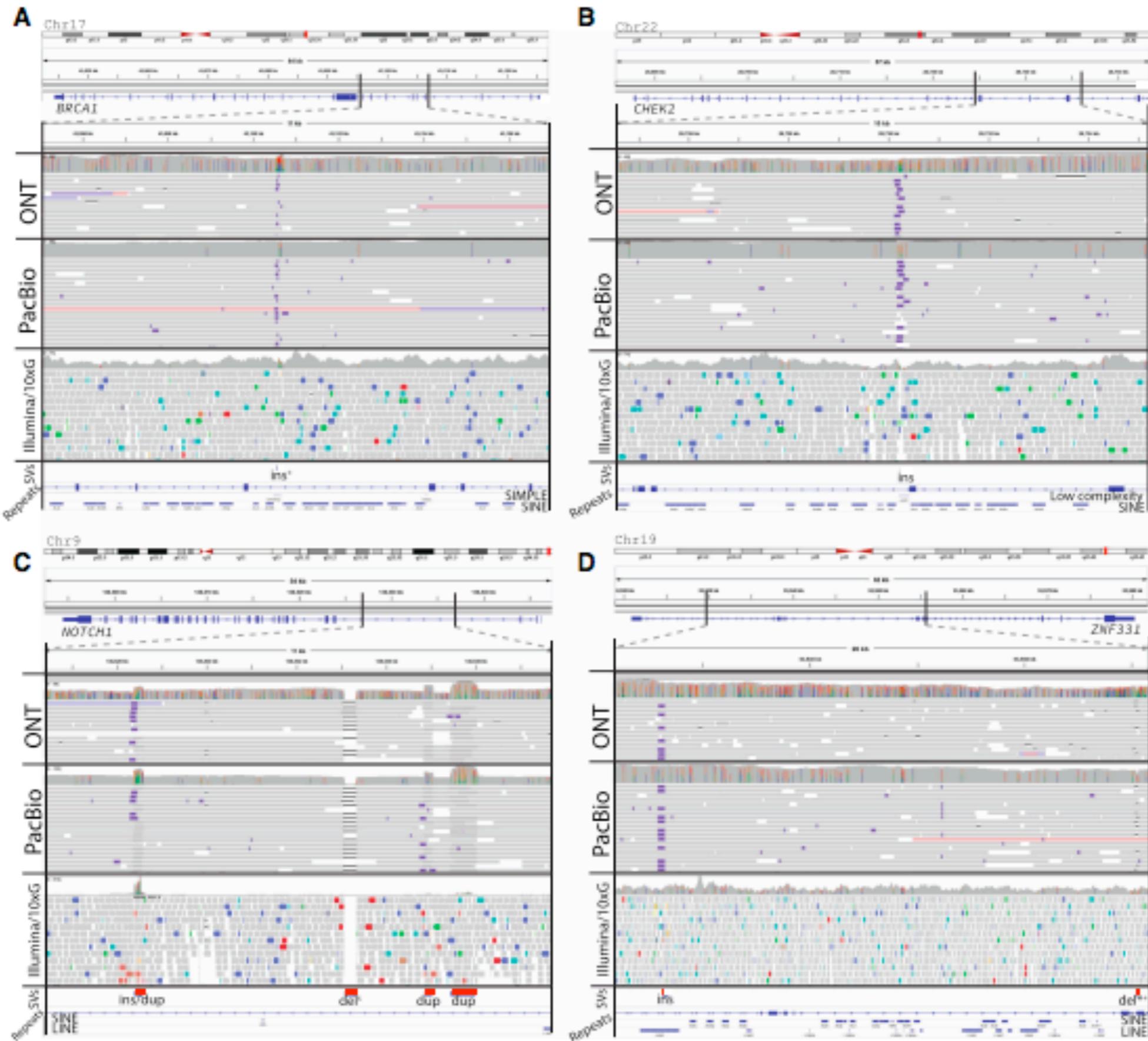
# Preliminary Structural Variations Analysis



	Total	Deletions	Duplications	Insertions	Inversions	Translocations
All SVs in normal	9816	5225	578	3727	130	156
All SVs in tumor	13737	7020	988	5292	202	235
SVs only in tumor (Also exclude NA12878)	3662	1805	420	1250	98	89

# SVs in sample 51 not detected by short reads.

Insertions found in BRCA1 and CHEK2. Insertions and duplications found in NOTCH1.



## JOURNAL ARTICLE

# Long-Read Sequencing Reveals Rapid Evolution of Immunity- and Cancer-Related Genes in Bats

Armin Scheben, Olivia Mendivil Ramos, Melissa Kramer, Sara Goodwin, Sara Oppenheim, Daniel J Becker, Michael C Schatz, Nancy B Simmons, Adam Siepel , W Richard McCombie  Author Notes

*Genome Biology and Evolution*, Volume 15, Issue 9, September 2023, evad148,

<https://doi.org/10.1093/gbe/evad148>

Published: 20 September 2023 Article history ▾



PDF

Split View

Cite



Permissions

Share ▾

## Abstract

Bats are exceptional among mammals for their powered flight, extended lifespans, and robust immune systems and therefore have been of particular interest in comparative genomics. Using the Oxford Nanopore Technologies long-read platform, we sequenced the genomes of two bat species with key phylogenetic positions, the Jamaican fruit bat (*Artibeus jamaicensis*) and the Mesoamerican mustached bat (*Pteronotus mesoamericanus*), and carried out a comprehensive comparative genomic analysis with a diverse collection of bats and other mammals. The high-quality, long-read genome assemblies revealed a contraction of interferon (IFN)- $\alpha$  at the immunity-related type I IFN locus in bats, resulting in a shift in relative IFN- $\omega$  and IFN- $\alpha$  copy numbers.

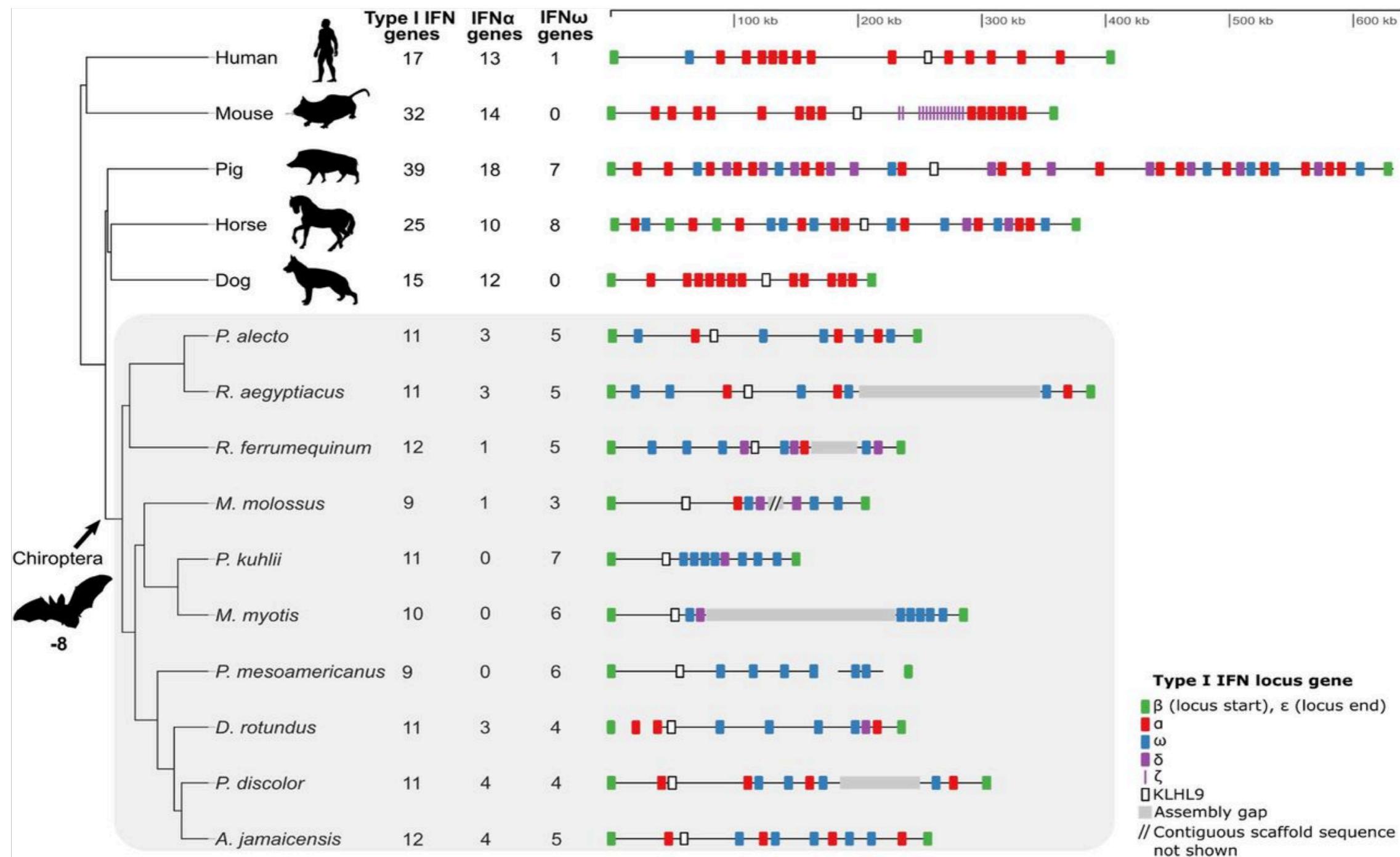
Contradicting previous hypotheses of constitutive expression of IFN- $\alpha$  being a feature of the bat immune system, three bat species lost all IFN- $\alpha$  genes. This shift to IFN- $\omega$  could contribute to the increased viral tolerance that has made bats a common reservoir for viruses that can be transmitted to humans.

Antiviral genes stimulated by type I IFNs also showed evidence of rapid evolution, including a lineage-specific duplication of IFN-induced transmembrane genes and positive selection in *IFIT2*. In addition, 33 tumor suppressors and 6 DNA-repair genes showed signs of positive selection, perhaps contributing to increased longevity and reduced cancer rates in bats. The robust immune systems of bats rely on both bat-wide and lineage-specific evolution in the immune gene repertoire, suggesting diverse immune



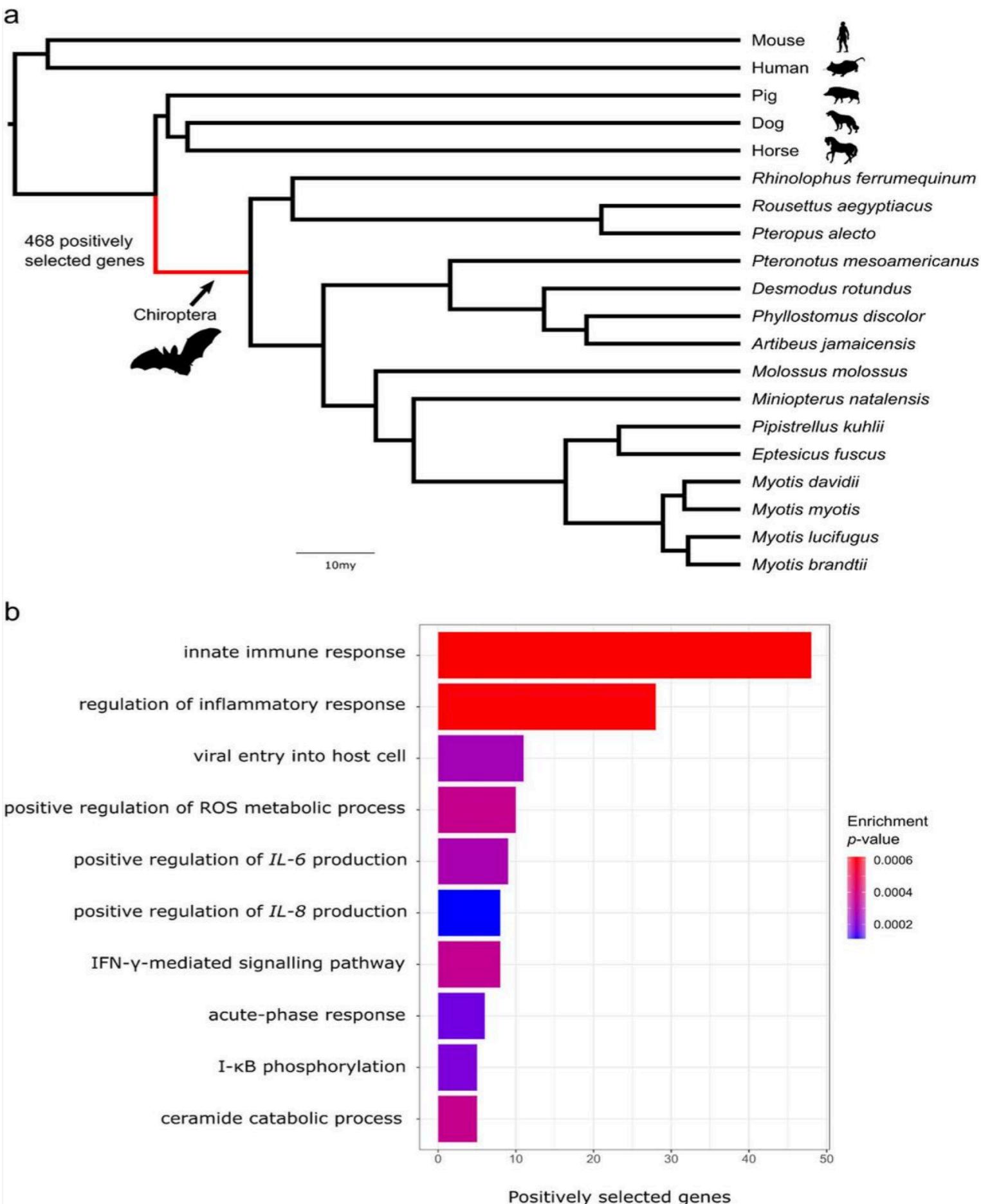
© Brock & Sherri Fenton

# The type I Interferon (IFN) locus is contracted in bats compared to other mammals



Positive selection  
on the bat  
ancestral branch  
suggests strong  
enrichment of  
innate immunity  
genes

Eight of the ten  
most significant  
GO terms are  
related to  
immunity



# DNA repair genes and tumor suppressors are positively selected in bats

Symbol	Name	aBSREL P-Value	Sites under Selection	Bat Branch ω	Outgroup Branch ω
CDH1	<i>Cadherin-1</i>	2.0e-6	3	0.17	0.29
CAT	<i>Catalase</i>	7.0e-6	1	0.38	>10
BIK	<i>BCL2 interacting killer</i>	3.7e-5	1	1.92	0.89
PALB2	<i>Partner and localizer of BRCA2</i>	1.8e-4	3	0.67	0.43
LATS2	<i>Large tumor suppressor kinase 2</i>	2.5e-4	10	0.23	0.11
SLC39A4	<i>Solute carrier family 39 member 4</i>	3.8e-4	3	0.55	0.42
SPARCL1	<i>SPARC like 1</i>	4.0e-4	4	0.88	0.95
PLA2G7	<i>Phospholipase A2 group VII</i>	8.5e-4	5	>10	1.70
GALR1	<i>Galanin receptor 1</i>	8.7e-4	1	0.10	0.08
CD79A	<i>CD79a molecule</i>	9.0e-4	1	>10	0.37
MYO1A	<i>Myosin IA</i>	9.8e-4	7	0.94	0.14

uth about  
rbation nobody  
to hear



Dear Abby: I want to  
bring my girlfriend to  
my son's wedding, do...



Gen Z and Gen Alpha  
are already beefing:  
'We're gonna be made...



16 Comments

**HEALTH**

# Bats could play vital role in preventing and treating cancer: 'first step' discovery

By Rob Bailey-Millado

Published Sep. 22, 2023, 12:28 p.m. ET

Bats get a deadly bad rap — but they could play a lifesaving role in cancer treatment, according to a "fascinating new study" that examines their remarkable immune systems.

As documented in peer-reviewed research in [Genome Biology and Evolution](#), published Thursday by Oxford University Press, scientists from New York hope that by better [understanding the winged creatures'](#) extraordinary ability to both host and survive infections, they can develop ways to treat — and prevent — cancer in humans.

[Privacy Policy](#) | [Feedback](#) [Follow 22.6M](#)**Daily Mail**.com[Home](#) | [Showbiz](#) | [Femail](#) | [Royals](#) | [Health](#) | **Science** | [Sports](#) | [Politics](#) | [Money](#) | [U.K.](#) | [Video](#) | [Travel](#) | [Puzzles](#) | [Shop](#)

Tuesday, Oct 24th 2023 4PM 58°F 7PM 56°F 5-Day Forecast

**Science & Tech**

## It might sound bat crazy, but these disease-riddled cave-dwelling creatures of the night might hold the key to curing CANCER

- Bats are recognized for their ability to tolerate viruses and low rates of cancer
- **READ MORE:** Now the CDC wants to monitor your POOP to track flu outbreaks

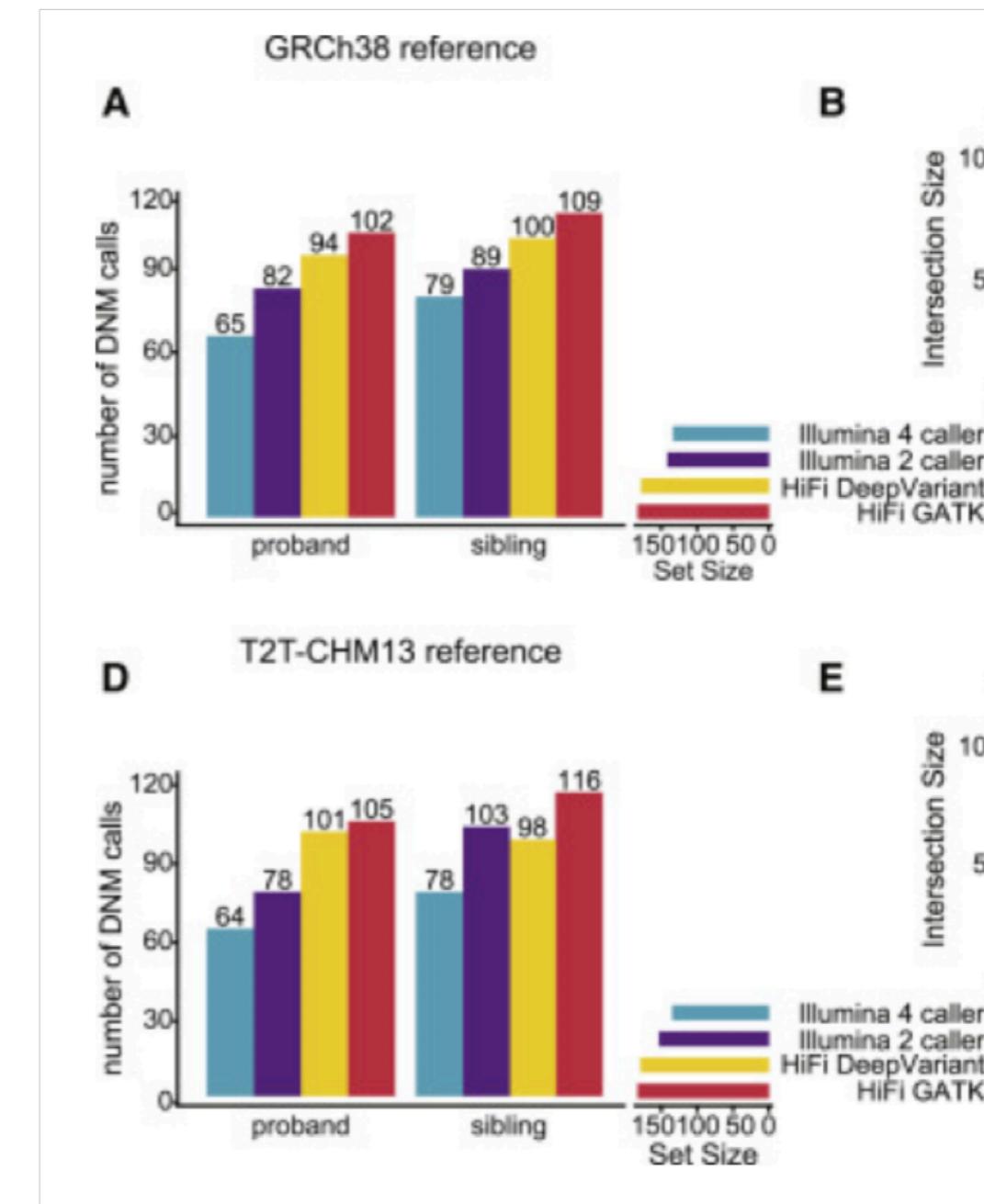
By [BRITNEY NGUYEN FOR DAILYMAIL.COM](#)

PUBLISHED: 07:31 EDT, 21 September 2023 | UPDATED: 08:47 EDT, 21 September 2023

# Long read sequencing uncovers DNMs in Autism quad

Long reads, increased detection of De Novo variants including SNPs, InDels and SVs

Further improvement when mapping to the more complete CHM13 reference

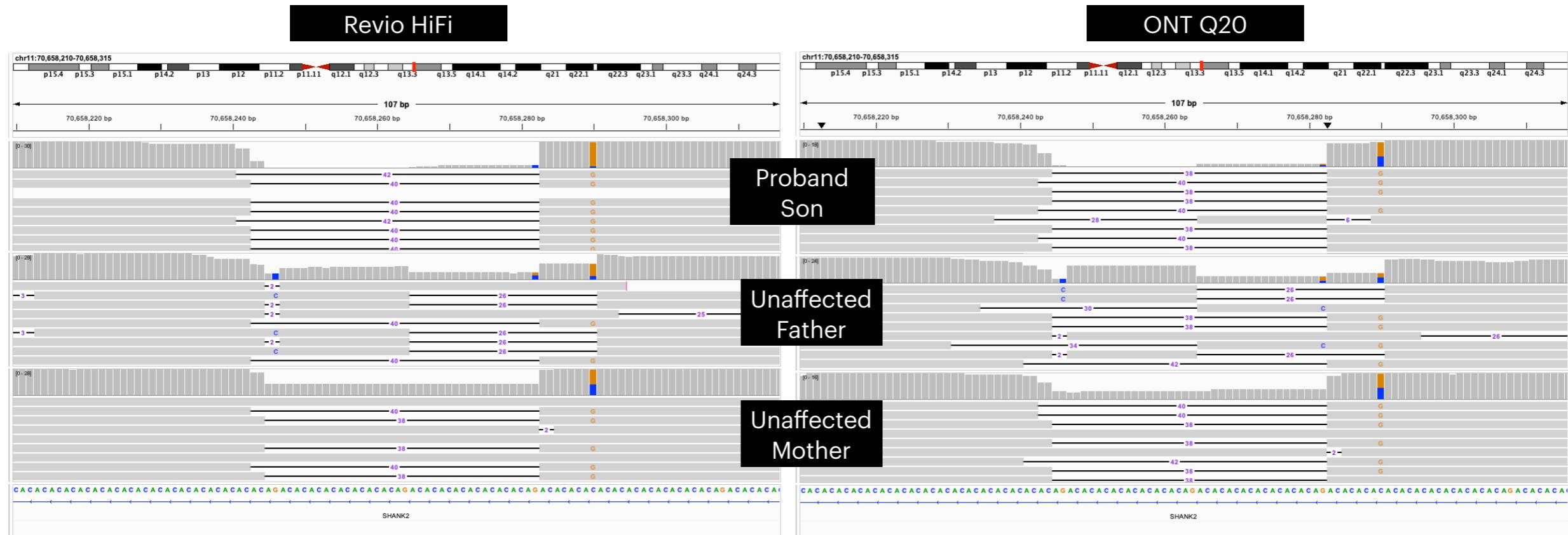


Michelle D. Noyes, William T. Harvey, David Porubsky, Arvis Sulovari, Ruiyang Li, Nicholas R. Rose, Peter A. Audano, Katherine M. Munson, Alexandra P. Lewis, Kendra Hoekzema, Tuomo Mantere, Tina A. Graves-Lindsay, Ashley D. Sanders, Sara Goodwin, Melissa Kramer, Younes Mokrab, Michael C. Zody, Alexander Hoischen, Jan O. Korbel, W. Richard McCombie, Evan E. Eichler

Familial long-read sequencing increases yield of de novo mutations

The American Journal of Human Genetics, Volume 109, Issue 4

# Long Read Sequencing of Autism trio Cell Line DNA



Homozygous deletion in an intron of the gene *SHANK2*, which has been associated with ASD

# Long Read Germline Sequencing of Early Onset Cancer

## SV Filtering Workflow

- Each individual will have ~25000 SV calls per genome
- Merge and genotype all calls across family members (~34000)
- Filter by family structure (~1400)
- Pull variants in/near genes (~450)
- Filter common events (~300)
- Filter false positives / ambiguous events/ select likely genes (~<100)

- No family history of cancer
- Standard IMPACT panel did not detect germline drivers



**bioRxiv**  
THE PREPRINT SERVER FOR BIOLOGY

HOME | SUBM

New Results

Follow this preprint

### Exploring the genetic and epigenetic underpinnings of early-onset cancers: Variant prioritization for long read whole genome sequencing from family cancer pedigrees

Melissa Kramer, Sara Goodwin, Robert Wappel, Matilde Borio, Kenneth Offit, Darren R. Feldman, Zsofia K. Stadler, W. Richard McCombie

**doi:** <https://doi.org/10.1101/2024.06.27.601096>

This article is a preprint and has not been certified by peer review [what does this mean?].



[Abstract](#) [Full Text](#) [Info/History](#) [Metrics](#) [Preview PDF](#)

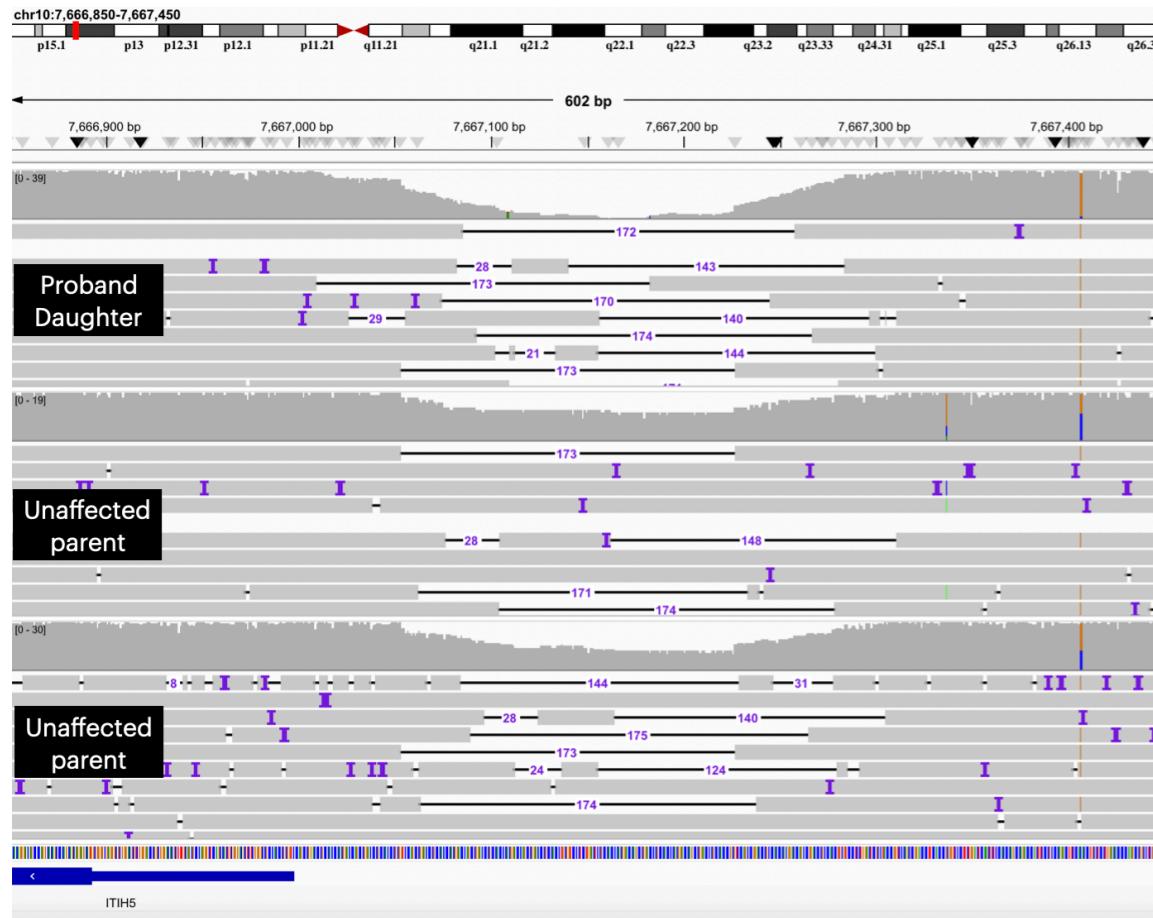
#### Abstract

Despite significant advances in our understanding of genetic cancer susceptibility, known inherited cancer predisposition syndromes explain at most 20% of early-onset cancers. As early-onset cancer prevalence continues to increase, the need to assess previously inaccessible areas of the human genome, harnessing a trio or quad family-based architecture for variant filtration, may reveal further insights into cancer susceptibility. To assess a broader spectrum of variation than can be ascertained by multi-gene panel sequencing, or even whole

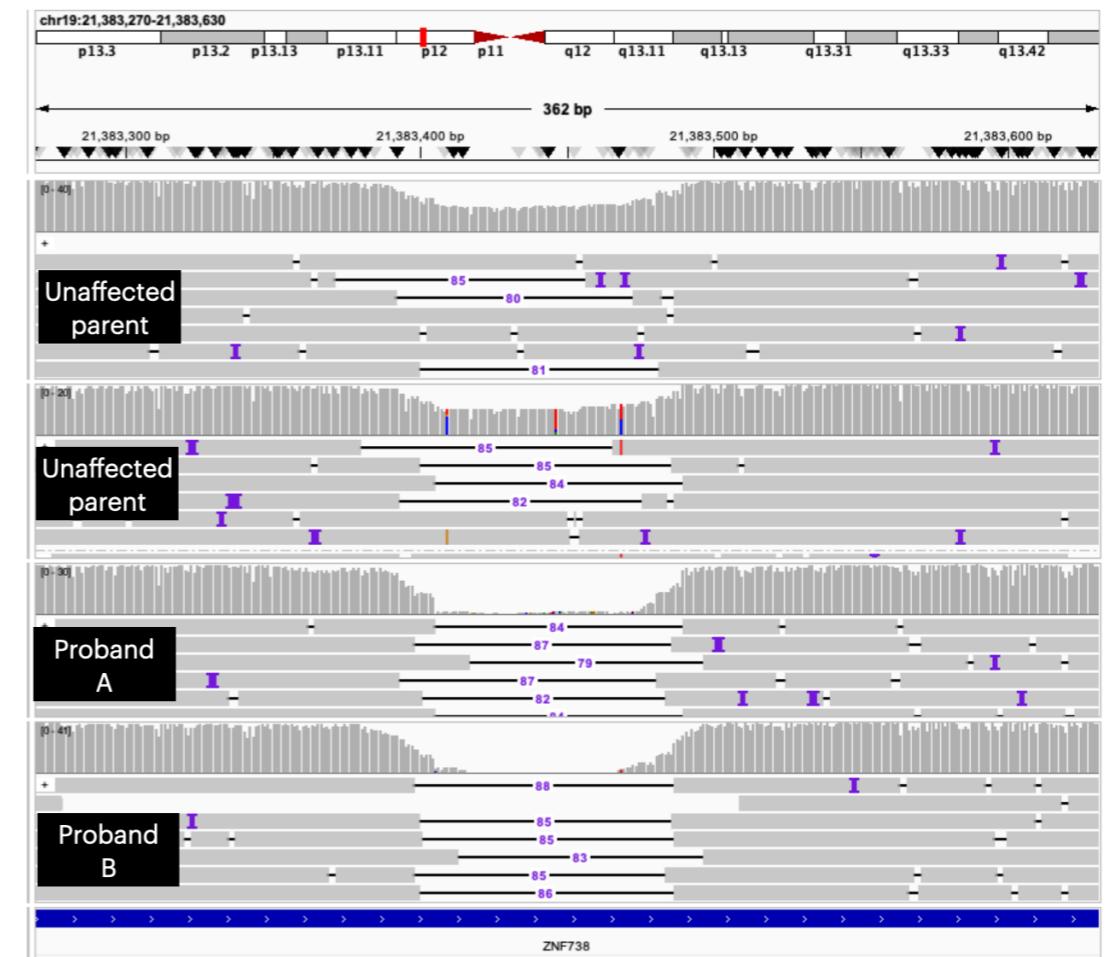
Collaboration with Zsofia Stadler MSKCC

# Long Read Sequencing of Early Onset Cancer Pedigrees

**Example of a structural variant which differs in the proband compared to the parents of EO-CRC trio #2, and overlaps a regulatory promoter region in ITIH5.**



**IGV screenshot of the last exon of gene ZNF738 in the TGCT quad. Aligned reads are displayed with heterozygous deletions in the unaffected parents (top panels), but homozygous deletions in both probands (bottom panels).**



Collaboration with Zsofia Stadler MSKCC

# Long-read DNA and cDNA sequencing identify cancer-predisposing deep intronic variation in tumor-suppressor genes

Suleyman Gulsuner,<sup>1,4</sup> Amal AbuRayyan,<sup>1,4</sup> Jessica B. Mandell,<sup>1</sup> Ming K. Lee,<sup>1</sup> Greta V. Bernier,<sup>2</sup> Barbara M. Norquist,<sup>3</sup> Sarah B. Pierce,<sup>1</sup> Mary-Claire King,<sup>1</sup> and Tom Walsh<sup>1</sup>

<sup>1</sup>Departments of Medicine (Medical Genetics) and Genome Sciences, University of Washington, Seattle, Washington 98195-7720, USA; <sup>2</sup>UW Medicine–Valley Medical Center, Renton, Washington 98055, USA; <sup>3</sup>Division of Gynecologic Oncology, University of Washington, Seattle, Washington 98195, USA

The vast majority of deeply intronic genomic variants are benign, but some extremely rare or private deep intronic variants lead to exonification of intronic sequence with abnormal transcriptional consequences. Damaging variants of this class are likely underreported as causes of disease for several reasons: Most clinical DNA and RNA testing does not include full intronic sequences; many of these variants lie in complex repetitive regions that cannot be aligned from short-read whole-genome sequence; and, until recently, consequences of deep intronic variants were not accurately predicted by *in silico* tools. We evaluated the frequency and consequences of rare deep intronic variants for families severely affected with breast, ovarian, pancreatic, and/or metastatic prostate cancer, but with no causal variant identified by any previous genomic or cDNA-based approach. For 10 tumor-suppressor genes, we used multiplexed adaptive sampling long-read DNA sequencing and cDNA sequencing, based on patient-derived DNA and RNA, to systematically evaluate deep intronic variation. We identified all variants across the full genomic loci of targeted genes, applied the *in silico* tools SpliceAI and Pangolin to predict variants of functional consequence, and then carried out long-read cDNA sequencing to identify aberrant transcripts. For eight of the 120 (6%) previously unsolved families, rare deep intronic variants in *BRCA1*, *PALB2*, and *ATM* create intronic pseudoexons that are spliced into transcripts, leading to premature truncations. These results suggest that long-read DNA and cDNA sequencing can be integrated into variant discovery, with strategies for accurately characterizing pathogenic variants.

> Genome Res. 2024 Nov 20;34(11):1825-1831. doi: 10.1101/gr.279158.124.

## Long-read DNA and cDNA sequencing identify cancer-predisposing deep intronic variation in tumor-suppressor genes

Suleyman Gulsuner <sup># 1</sup>, Amal AbuRayyan <sup># 1</sup>, Jessica B Mandell <sup>1</sup>, Ming K Lee <sup>1</sup>,  
Greta V Bernier <sup>2</sup>, Barbara M Norquist <sup>3</sup>, Sarah B Pierce <sup>1</sup>, Mary-Claire King <sup>4</sup>, Tom Walsh <sup>1</sup>

Affiliations + expand

PMID: 39271294 PMCID: PMC11610570 (available on 2025-05-01) DOI: [10.1101/gr.279158.124](https://doi.org/10.1101/gr.279158.124)

- Long read sequencing uncovers driver variants in introns of tumor suppressor genes in previously unexplained high burden cancer pedigrees
- Intronic variants outside typical splice regions cause aberrant transcripts

Figure 1. BRCA1

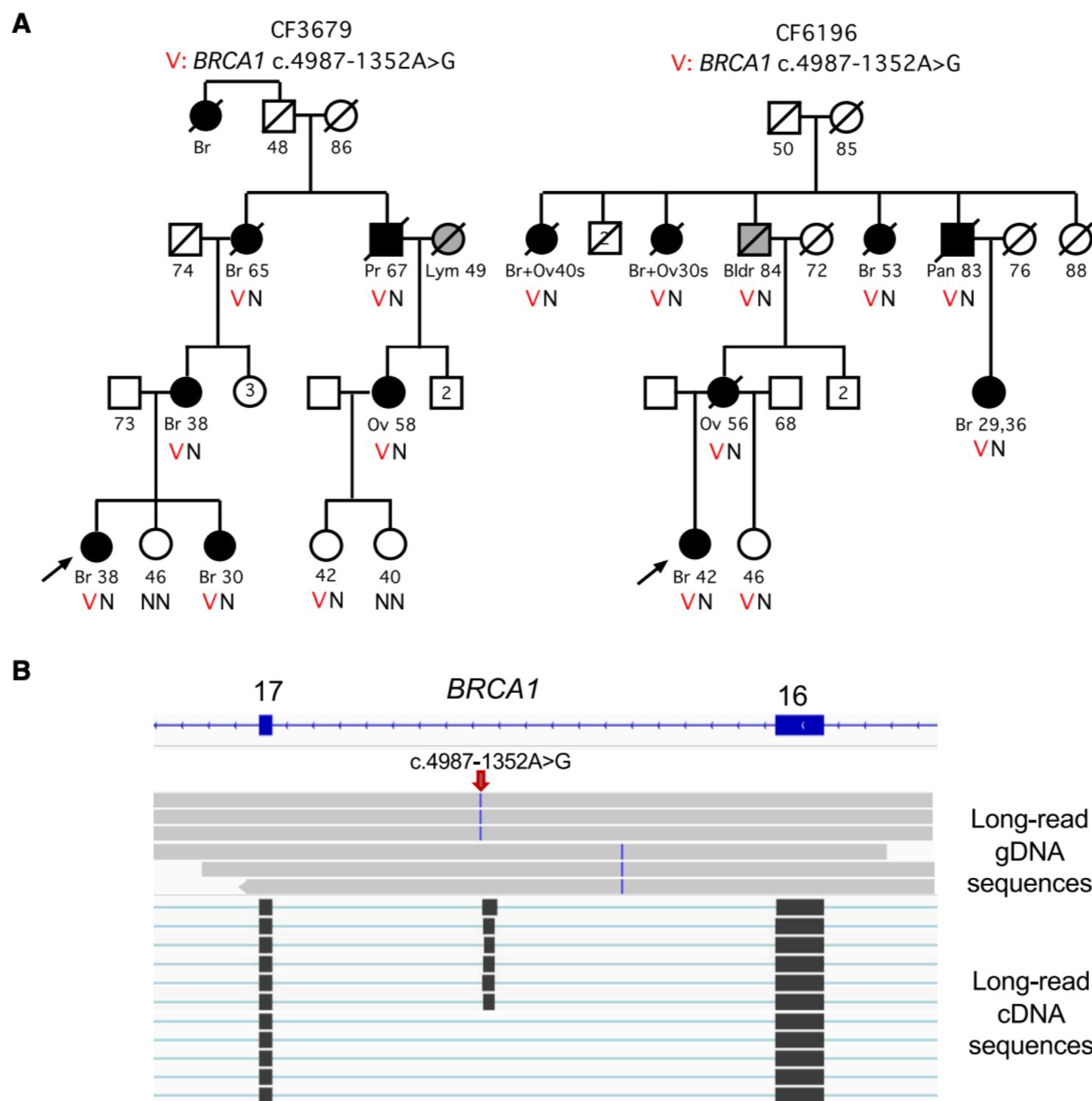


Figure 2. BRCA1 and PALB2

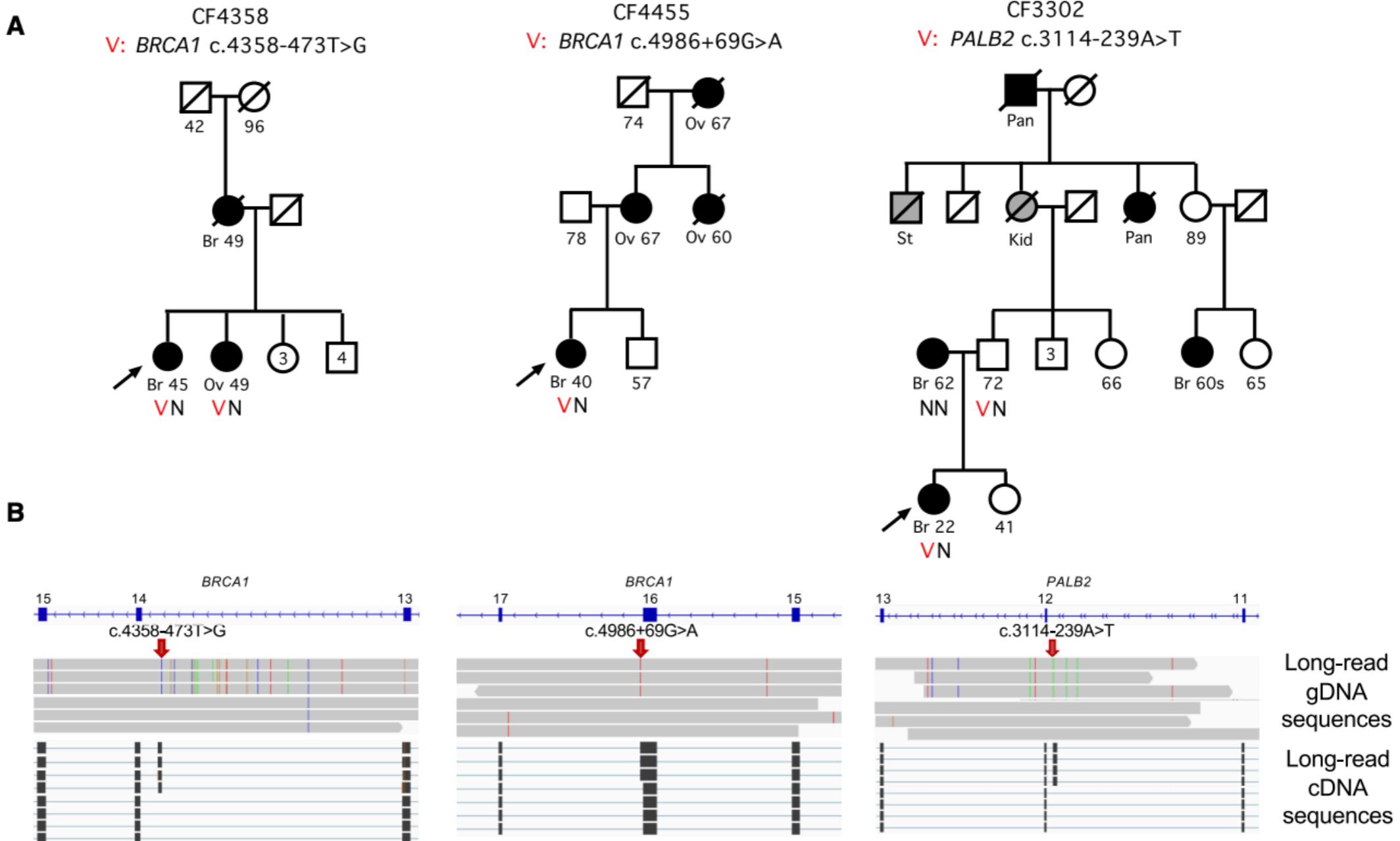
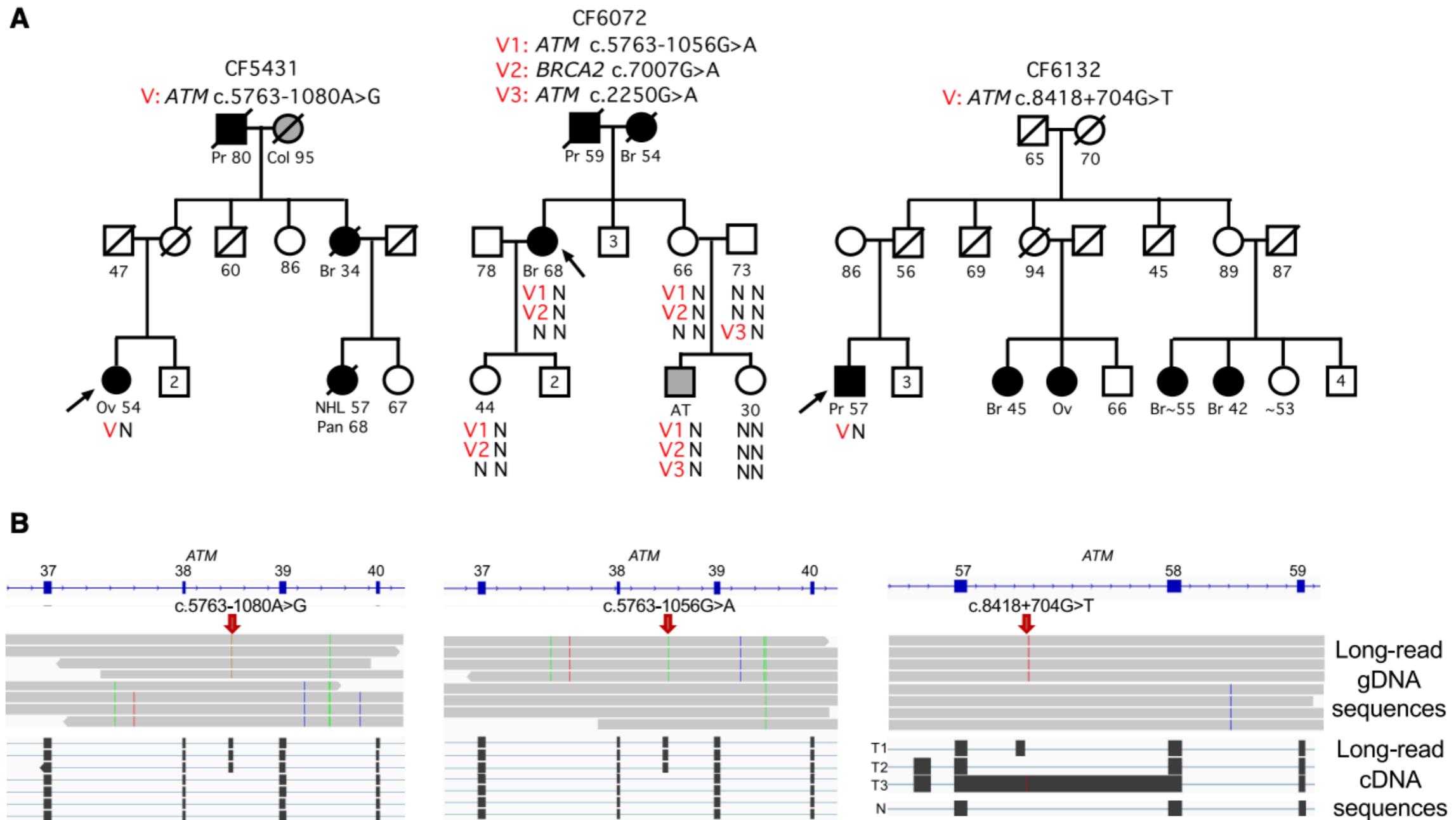


Figure 3. ATM

The 3rd variant causes 3 different aberrant transcripts



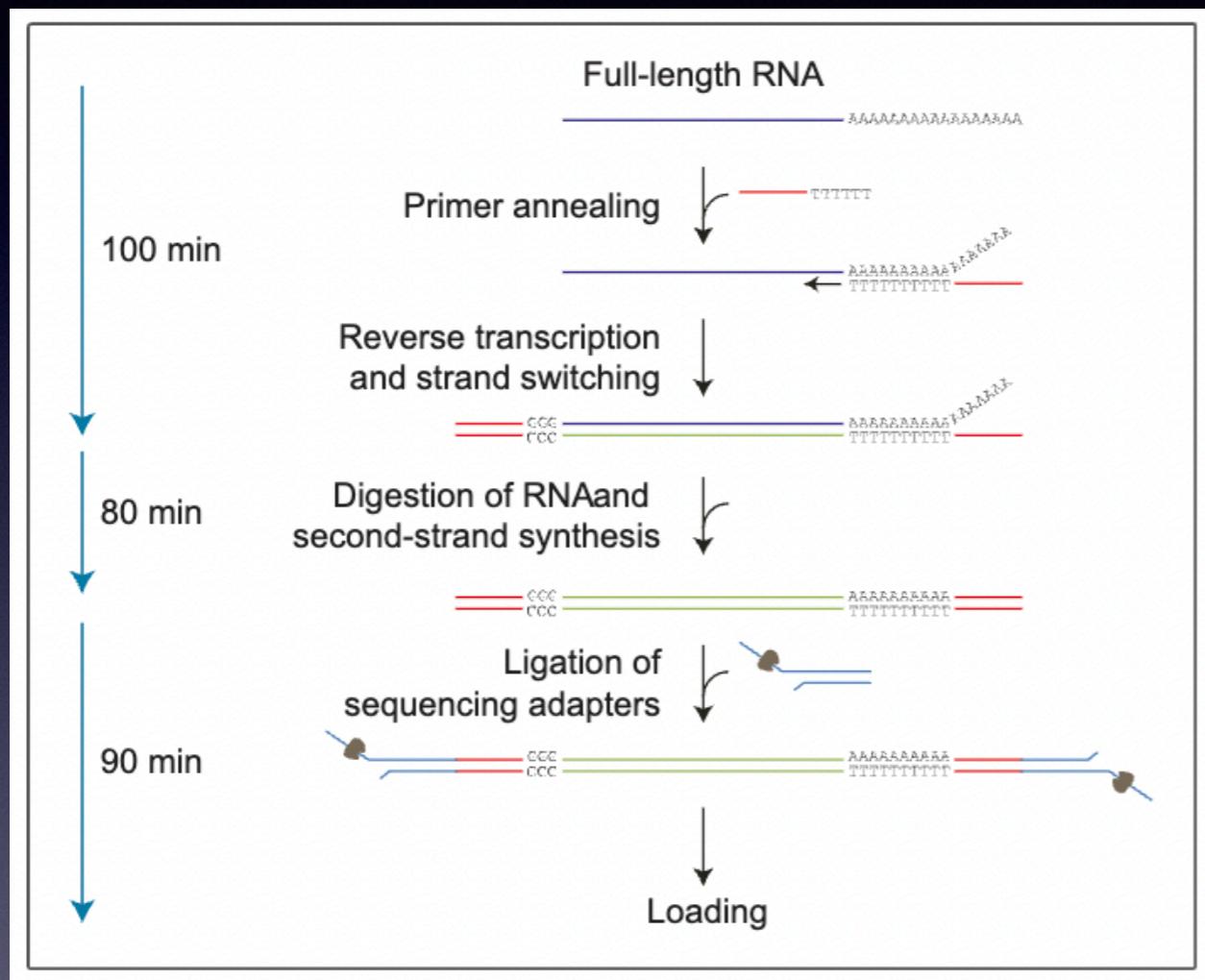
Nanopore error rates are improving

Chemistry and software improvements are increasing accuracy

ONT simplex reads with Dorado (single read):  
~1% error (Q20)

ONT now recommends HERRO read correction to drive quality to  
~Q30-Q40

# ONT direct cDNA



Requires more input (~50 poly A+)

Does not use PCR

Enriched for full length cDNA  
(template switching)

Can multiplex with native barcoding kit

Less throughput (40-50 million reads per PromethION cell), lengths a bit longer ~1.5kb

# Transcriptome Sequencing Cost Comparisons Across Platforms

Illumina cost per  
million reads  
\$6

ONT PCR cDNA cost  
per million reads  
\$10

PacBio IsoSeq cost  
per million reads  
\$20

Tradeoffs on accuracy and length, so it  
is key to assess the method that will  
address critical questions of your  
experiment



AA

biorxiv.org



x

bR Gapless assembly of complete human and plant chromosomes using only nanopore sequencing | bioRxiv

Cold  
Spring  
Harbor  
Laboratory**bioRxiv**

THE PREPRINT SERVER FOR BIOLOGY

HOME | SUBMIT | FAQ | BLOG | ALERTS / RSS | RESOURCES

| ABOUT | CHANNELS

Search



Advanced Search

New Results

Follow this preprint

Previous

Next

## Gapless assembly of complete human and plant chromosomes using only nanopore sequencing

Sergey Koren, Zhigui Bao, Andrea Guerracino, Shujun Ou, Sara Goodwin, Katharine M. Jenike, Julian Lucas, Brandy McNulty, Jimin Park, Mikko Rautiainen, Arang Rhee, Dick Roelofs, Harrie Schneiders, Ilse Vrijenhoek, Koen Nijbroek, Doreen Ware, Michael C. Schatz, Erik Garrison, Sanwen Huang, W. Richard McCombie, Karen H. Miga, Alexander H.J. Wittenberg, Adam M. Phillippy

**doi:** <https://doi.org/10.1101/2024.03.15.585294>

This article is a preprint and has not been certified by peer review [what does this mean?].



Abstract

Full Text

Info/History

Metrics

Preview PDF

Download PDF

**COVID-19 SARS-CoV-2  
preprints from medRxiv and  
bioRxiv**

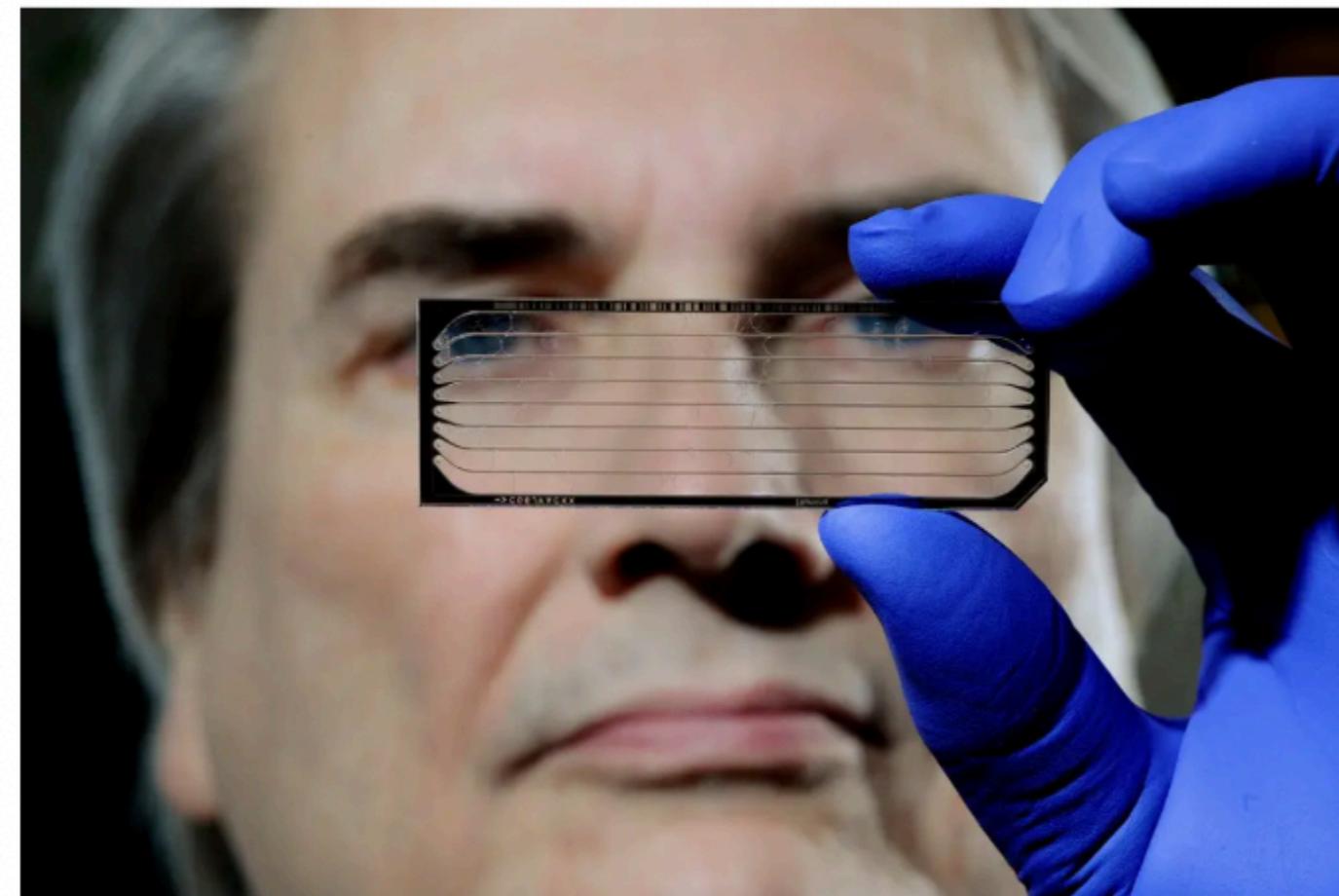
Subject Area

Bioinformatics

Subject Areas

# DNA Sequencing Caught in Deluge of Data

Share full article



W. Richard McCombie, a professor of human genetics at the Cold Spring Harbor Laboratory, examining DNA samples. Kathy Kmonicek for The New York Times

# Conclusions

- Substantial variation is missed by short reads
- Phasing can reveal more detailed context of variation
- Massive data storage problems are worse for long read sequencing
- Interpretation of SVs must be optimized to understand regulatory effects of widespread variation

# Summary

Long read platforms have matured significantly in the last few years  
PacBio and Oxford Nanopore producing similar length distributions  
Overcome high error sequencing with improved informatics  
Oxford Nanopore exciting for methylation & direct RNA capabilities

Long reads are crucial for accurate SV calling  
Finding thousands to tens of thousands of additional SVs over short reads  
Resolves the false positives observed with short reads  
Detecting potential cancer risk factors that would otherwise go unnoticed

Sample & DNA requirements one of the largest barriers for clinical application  
Continue to advance protocols for extracting, preparing samples  
Organoids (as opposed to primary tumors) enable large DNA amounts for long read sequencing, though it remains much more difficult than cell culture  
Organoids also enable application and profiling of other molecular and pharmaceutical assays

## Future goals

Reduce sample DNA input - tumors, single cell, targeting - Shruti Iyer  
Analyze data from projects for relevant genome properties  
Improve long read sequencing efficiency - read length, yield, combination of input data types  
Optimum cost benefit analyses of different long read approaches and coverage  
Optimize long read transcriptome sequencing

# Acknowledgements



## McCombie Lab

Sara Goodwin  
Melissa Kramer  
Olivia Mendivil Ramos  
Stephanie Muller  
Robert Wappel  
Senem Mavruk  
Elena Ghiban  
Shruti Iyer

## Siepel Lab

Armin Scheben

## Spector Lab

Sonam Bhatia  
Gayatri Arun



## Schatz Lab

Sam Kovaka  
Melanie Kirsche  
Rachel Sherman  
Katie Jenike  
Sergey Aganeov  
Srividya Ramakrishnan

## Timp Lab

Isac Lee

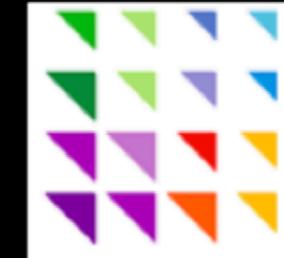
## AMNH

Nancy Simmons  
Sara Oppenheim



## **Fritz Sedlazeck   Karen Kostroff**

Medhat Helmy



Zsofia Stadler  
Matilde Borio  
Zalak Patel  
Sami Belhadj

## Mayo Clinic Mark Ebbert

## Living Fossils Consortium

Funding  
NCI  
NSF  
NHGRI  
Northwell Health