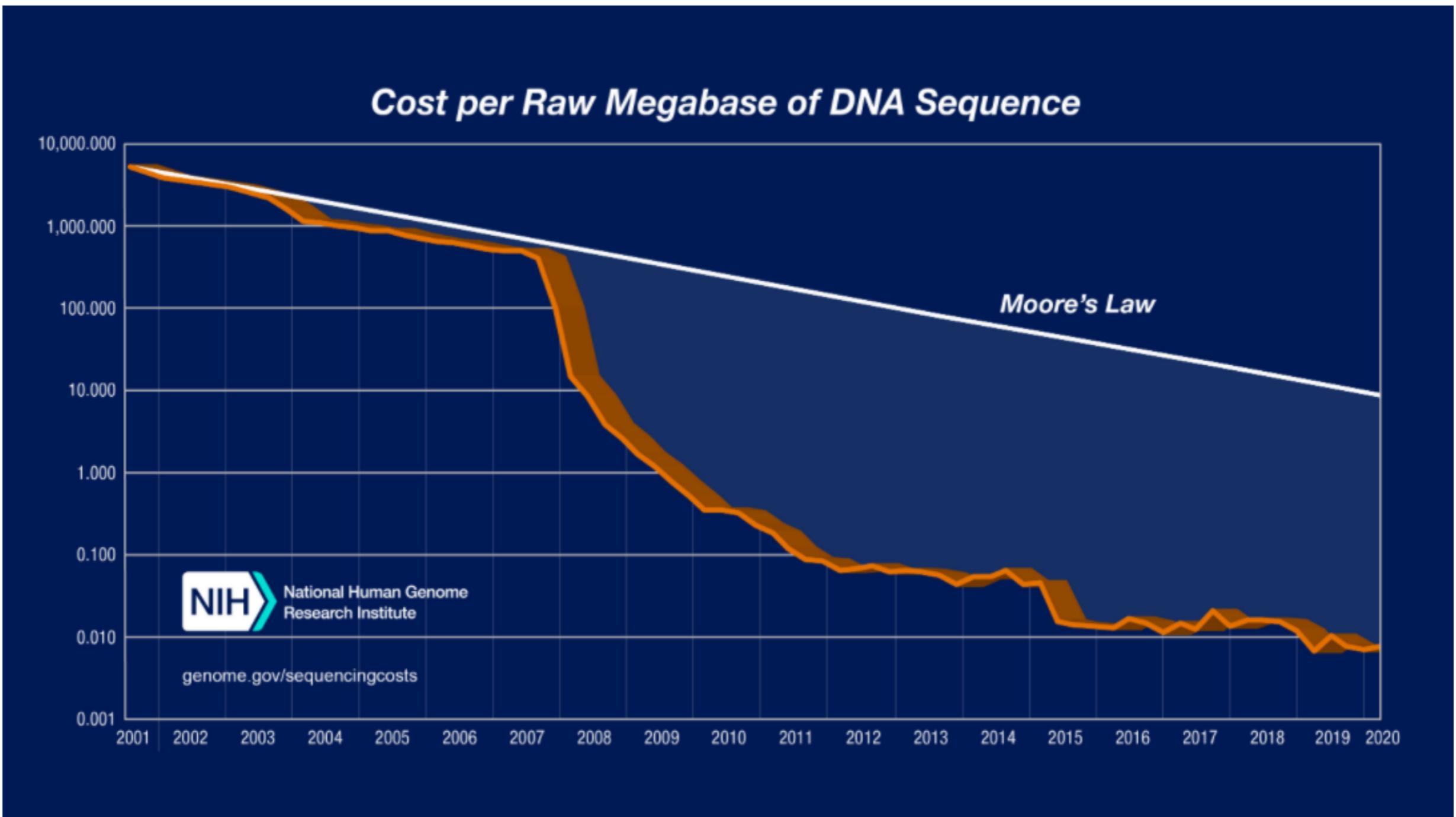


# Long Read Sequencing

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

Advanced Sequencing Technologies and Applications course  
Cold Spring Harbor Laboratory  
2022

# Significant advances in genome sequencing over last 16 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

??



STANLEY INSTITUTE FOR  
COGNITIVE GENOMICS  
COLD SPRING HARBOR LABORATORY

# 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

# The Telomere-to-Telomere Consortium

Long read sequencing  
of the hydatidiform  
mole CHM13 with  
multiple technologies

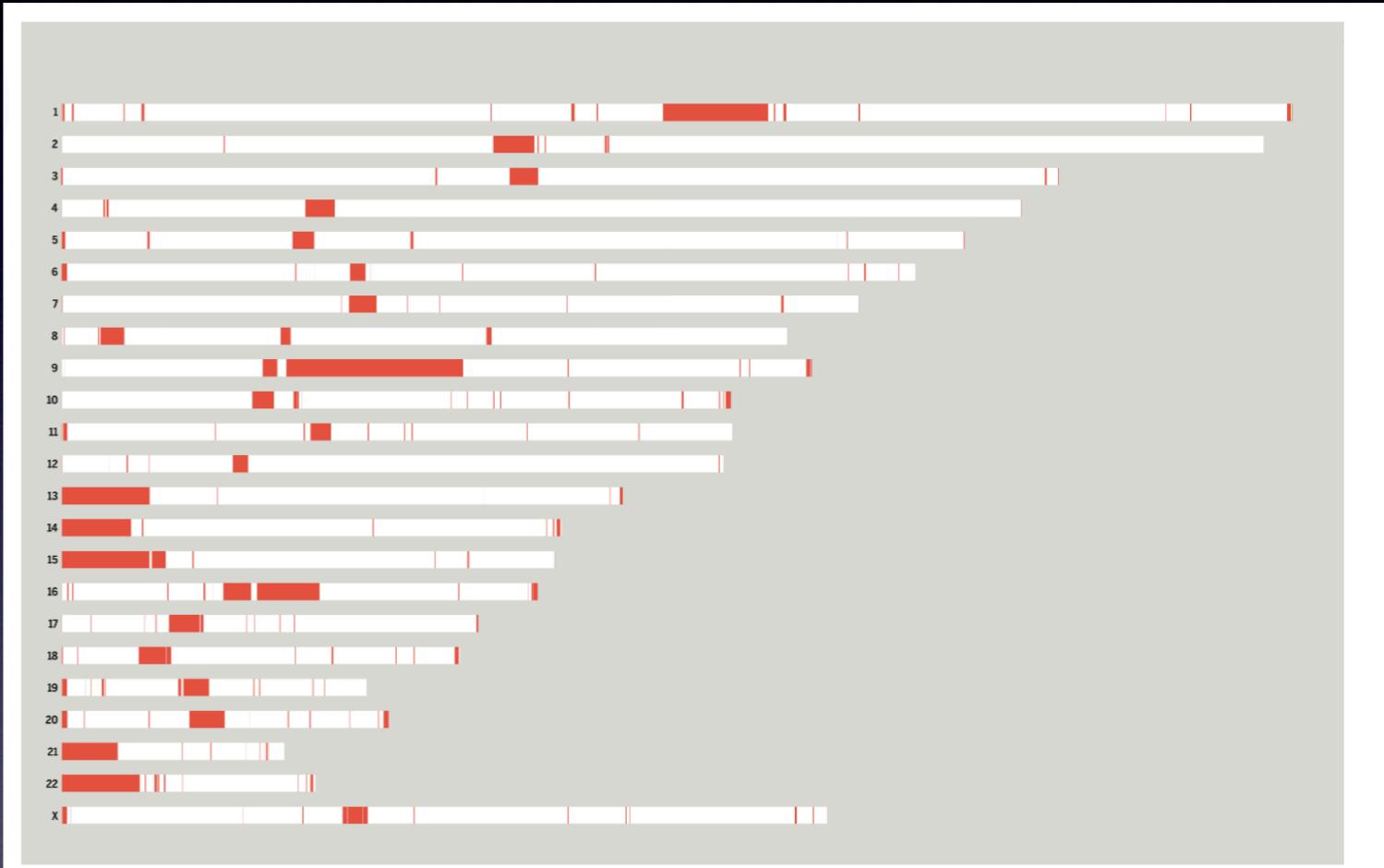
- PacBio HiFi
- ONT ultralong reads
  - Illumina Arima Genomics Hi-C (Hi-C)
- BioNano optical maps
  - single-cell DNA template strand sequencing (Strand-seq)



# 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

# Limitations of long reads

- Cost
- Throughput\*
- Accuracy\*
- DNA amount required
- DNA quality required

\*This is rapidly changing

# Two “flavors” of long read sequencing

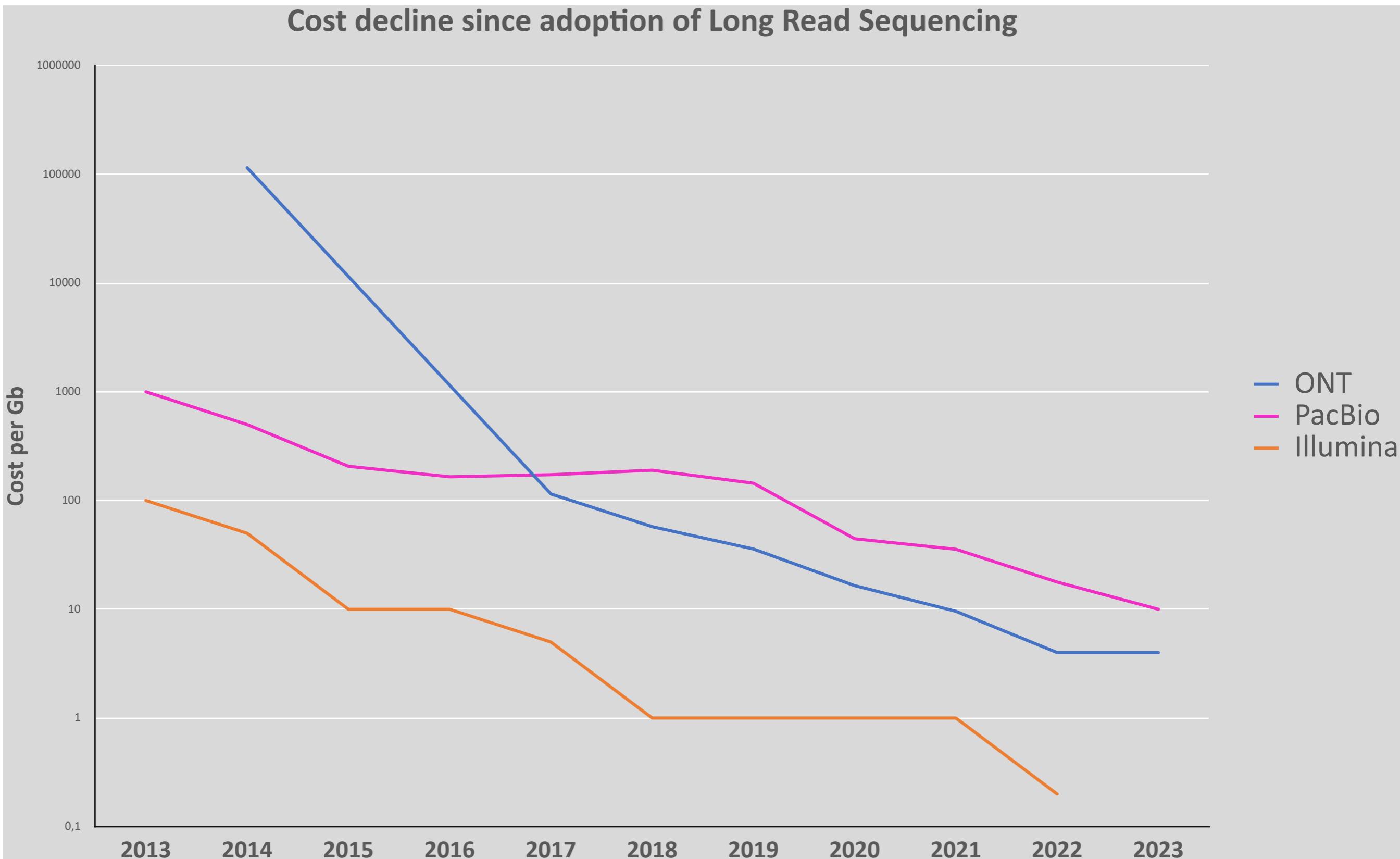


PACIFIC  
BIOSCIENCES®



# Significant advances in long read sequencing over last 9 years

Cost decline since adoption of Long Read Sequencing





PACIFIC  
**BIOSCIENCES**<sup>®</sup>

# PacBio



RSII

- ~85% single pass accuracy
- “short read” CCS accuracy >99.999%
- Up to 2Gb per SMRTcell
- Read lengths up to 60kb

# Pacific Biosciences Sequel II

Released in 2018

Smaller, lower cost instrument

8 Million ZMW (155k RSII, 1M Sequel I)

Early runs were rocky

Substantial recent improvement in performance up to 200Gb of CLR data or 30Gb of HiFi data

Up to 800Gb CLR or 120Gb HiFi in one week



# Pacific Biosciences Revio (Available 2023)

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

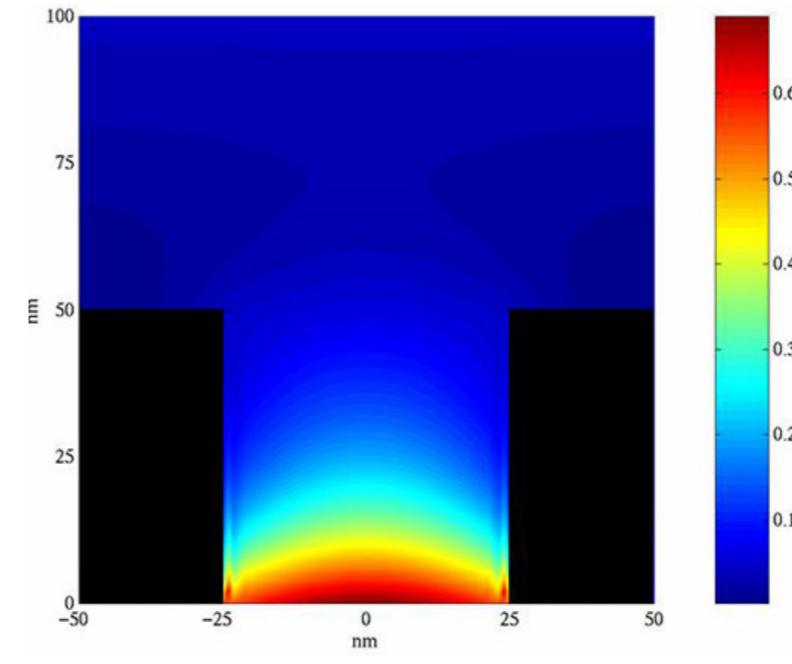
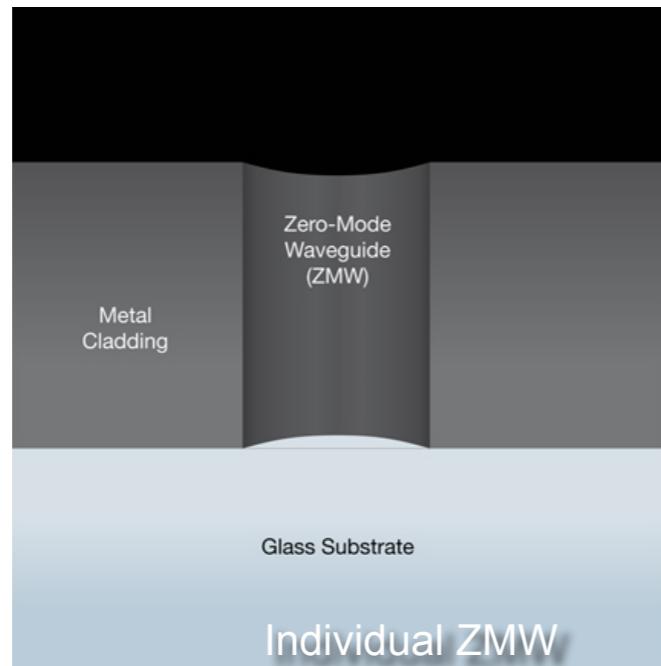


# Zero-Mode Waveguides Are the Observation Windows

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

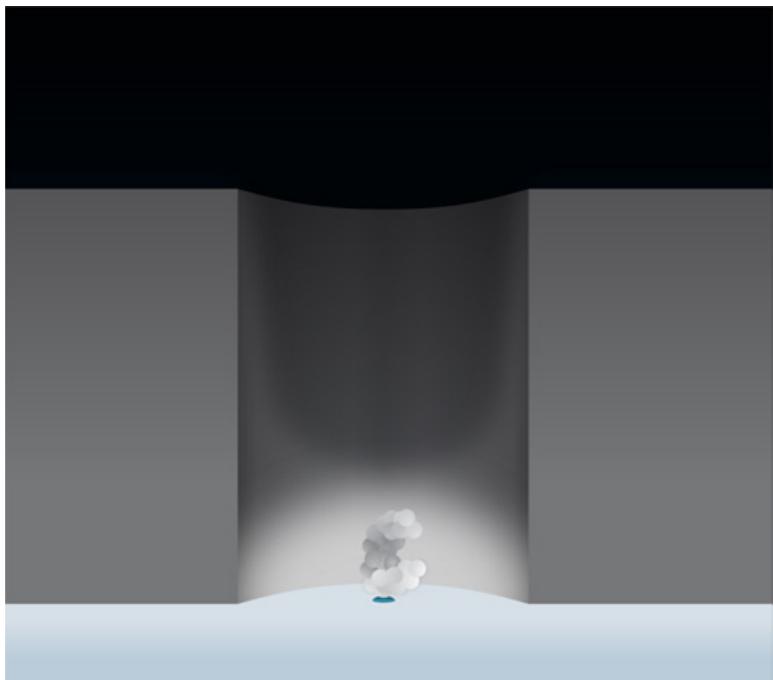


Laser light illuminates the ZMW

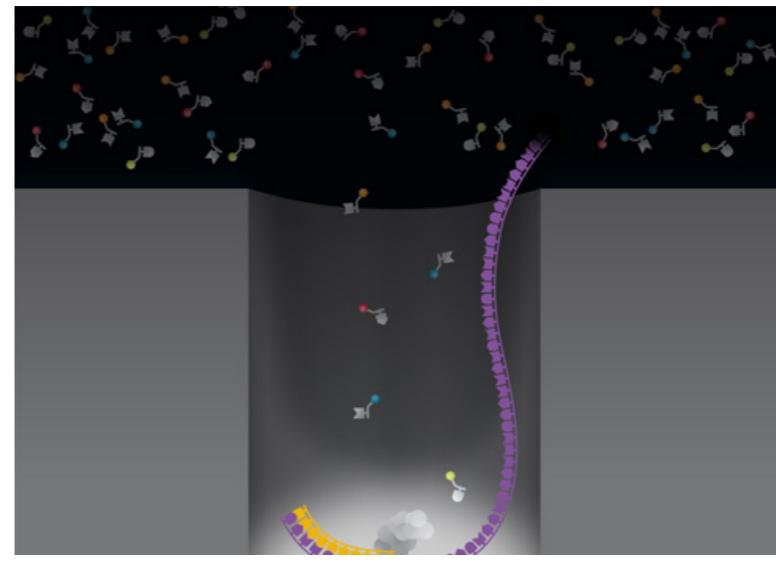
# DNA Polymerase as a Sequencing Engine

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

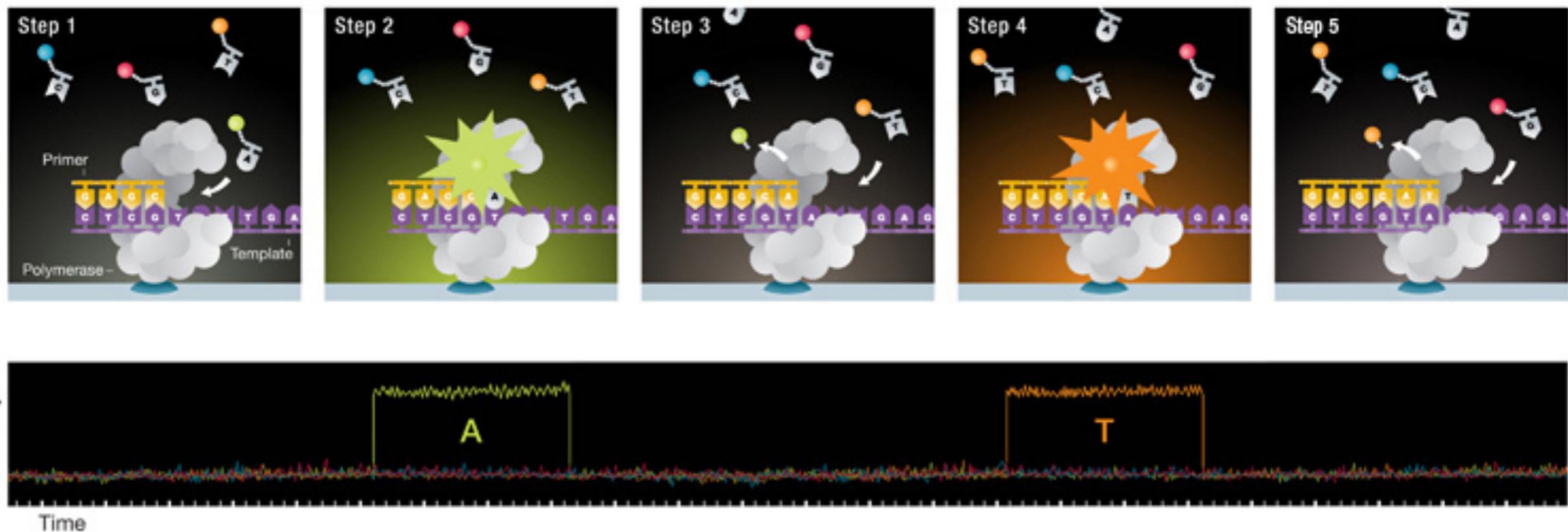


ZMW with DNA polymerase and phospholinked nucleotides

# 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



# Sample Loading May Require Titration

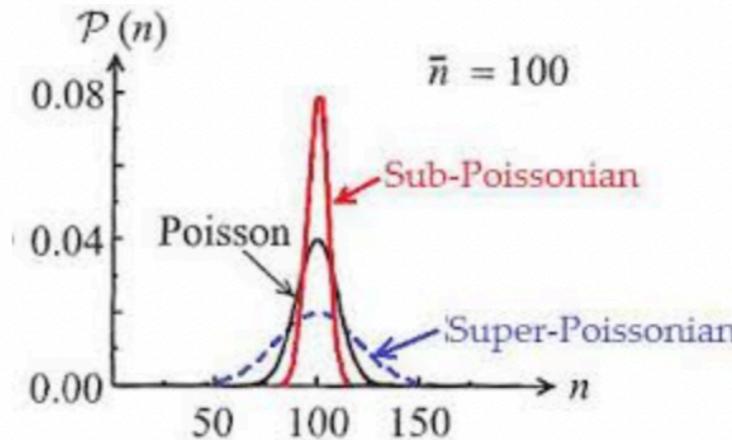
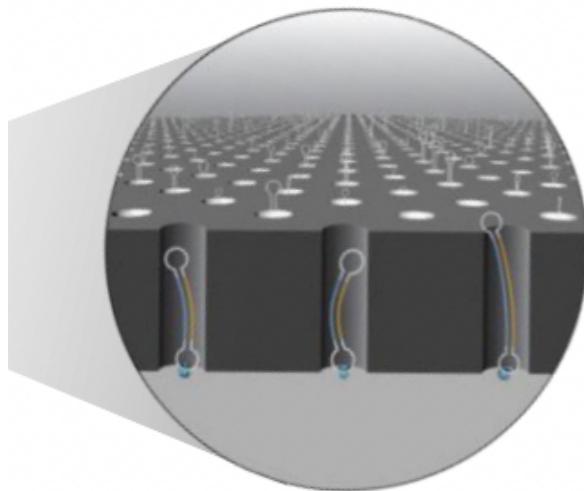


Figure 5.26 from Adrian Jeantet, Cavity quantum electrodynamics with carbon nanotubes, 2017.

## New General Loading Guidance

For best results using Sequel System Software Suite (v5.1.0), we recommend that you load higher than classic Poisson distribution (e.g. 37%). Please refer to PacBio's Quick Reference Card, [Diffusion Loading and Pre-Extension Time Recommendations for the Sequel System](#), for more information including starting loading concentrations.

1. We recommend that for most applications and sample types, set target P1 value at >50%. Poisson statistics still apply, and we want to target only 1 active polymerase per ZMW. Pre-extension can help eliminate some >1 sequencing polymerase/ZMW to allow the target loading to increase from P1 ~37 to >50%.
2. As P1's increase, there may be some decrease in read length and this should be monitored.
3. We recommend monitoring the P0 value for sample overloading. We recommend that you set target P0 values at ~20%. **Note:** If the P0 values are <10%, then the SMRT® Cell is overloaded.
4. For application-based loading, we recommend the following:
  - Iso-Seq® libraries, and amplicons with pre-extension, will benefit if you target P1 at ~70% and keep P2 <20%.
  - For *de novo* libraries generated from the SMRTbell® Express Template Prep Kit, we recommend targeting P1 ~50%.
  - For microbial multiplex samples, we recommend targeting P1 ~50-65%.

PacBio terminology:

- P0: The Percentage of ZMWs that are Empty.
- P1: The Percentage of ZMWs that are Productive.
- P2: The Percentage of ZMWs that are not P1 or P0.

## NEW ADAPTIVE LOADING FEATURE FOR SEQUEL II AND IIE SYSTEMS

Adaptive Loading reduces sample overloading, allowing users to load higher with confidence

- Adaptive loading technology actively monitors polymerase complex binding to the bottom of ZMWs during the sample immobilization step.
- Detection of these active polymerase complexes allows the system to terminate the immobilization step when the desired loading target has been achieved.
  - This approach can help reduce sample overloading and run-to-run yield variability

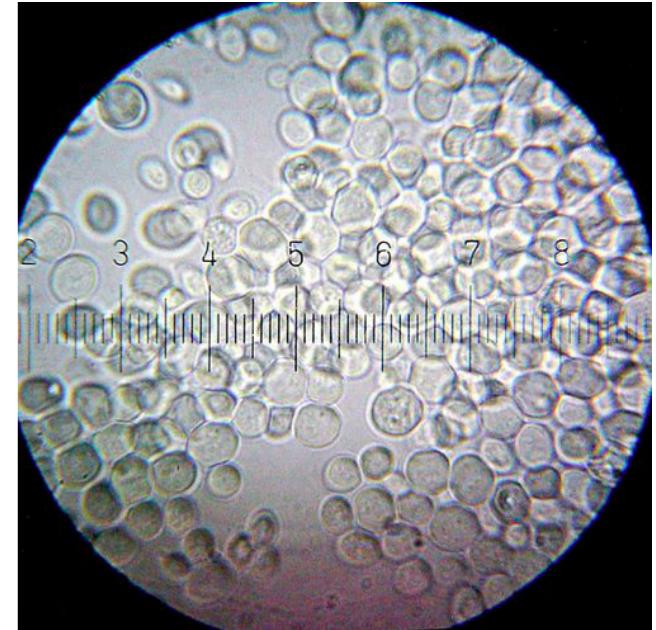


**Adaptive Loading (AL)** uses active monitoring of polymerase binding to the bottom of the ZMW during loading to reduce variability and the risk of overloading with high-concentration samples

**LIGHTS ALL ASKEW IN THE HEAVENS;  
Men of Science More or Less Agog Over Results  
of Eclipse Observations. EINSTEIN THEORY  
TRIUMPHS Stars Not Where They Seemed or  
Were Calculated to be, but Nobody Need Worry.  
A BOOK FOR 12 WISE MEN No More in All  
the World Could Comprehend It, Said Einstein  
When His Daring Publishers Accepted It.**

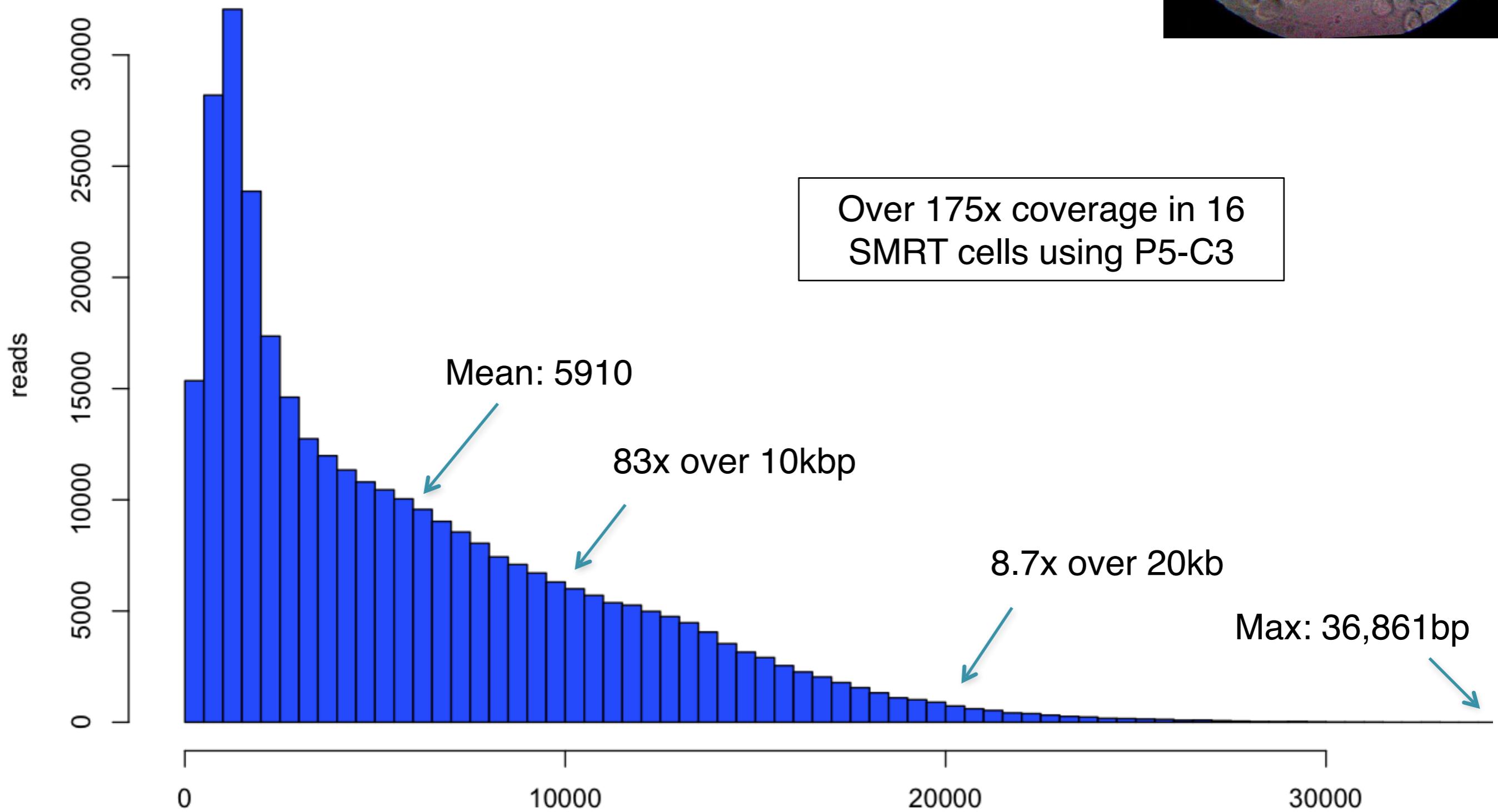
New York Times Nov. 9, 1919.

# Yeast: *S. cerevisiae* W303



PacBio RS II sequencing at CSHL

Size selection using an 7 Kb elution window on a BluePippin™ device from Sage Science



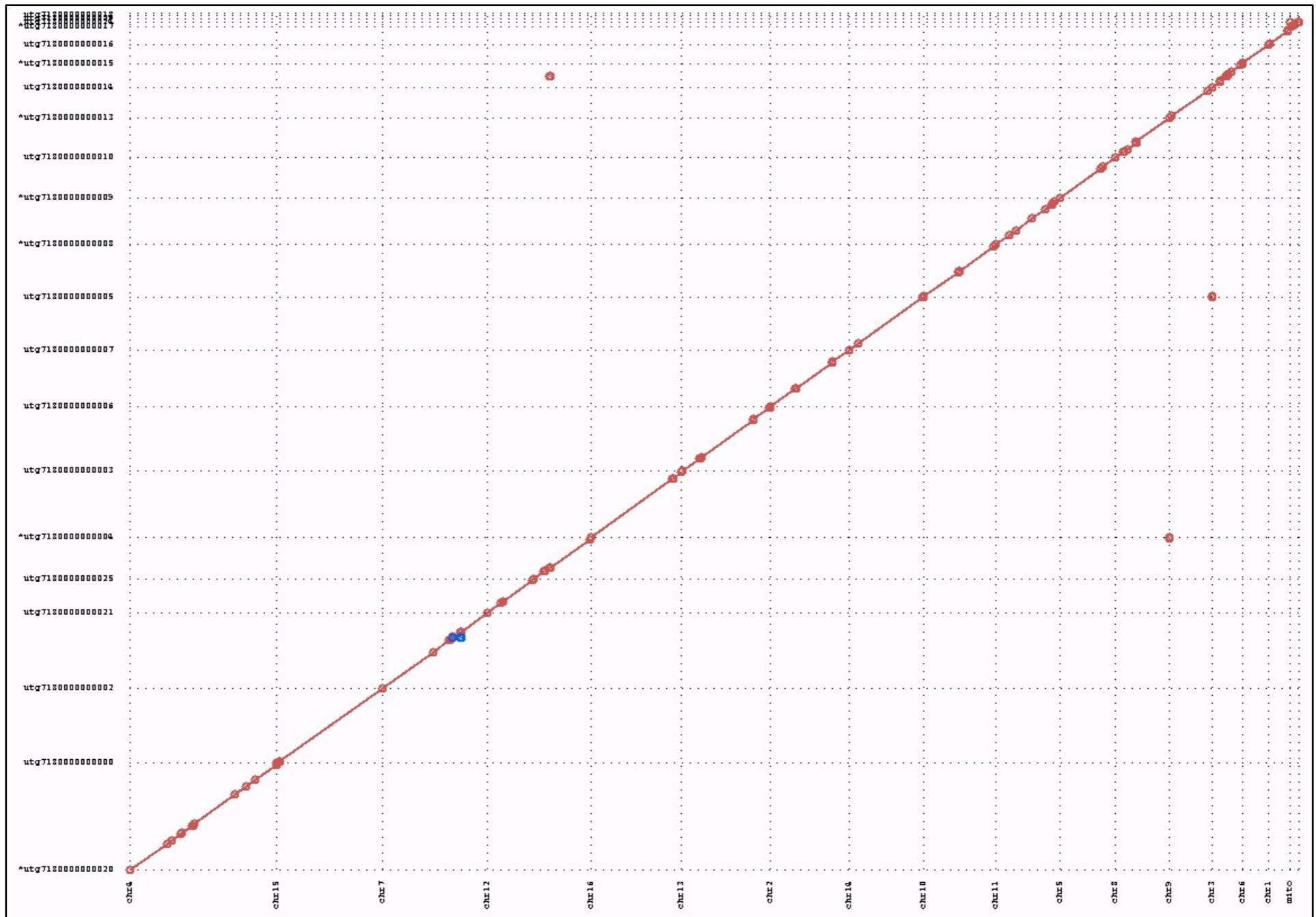
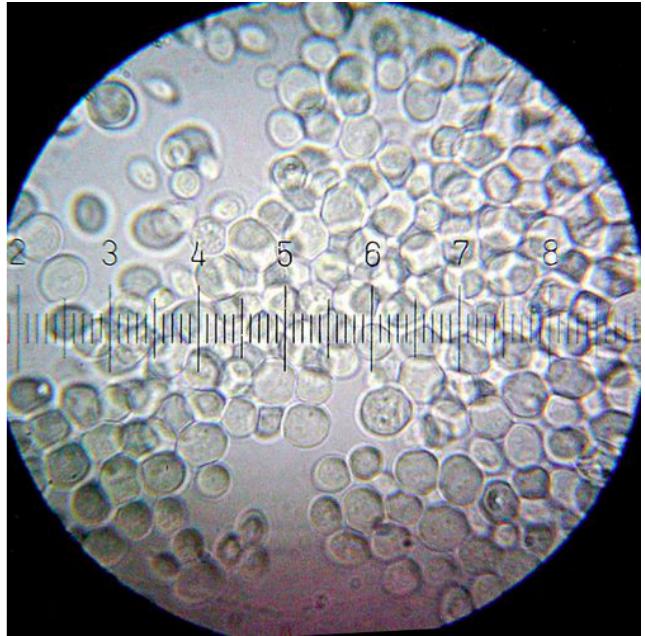
# *S. cerevisiae* W303

S288C Reference sequence

- 12.1Mbp; 16 chromo + mitochondria; N50: 924kbp

PacBio assembly using HGAP + Celera Assembler

- 12.4Mbp; 21 non-redundant contigs; N50: 811kbp; >99.8% id



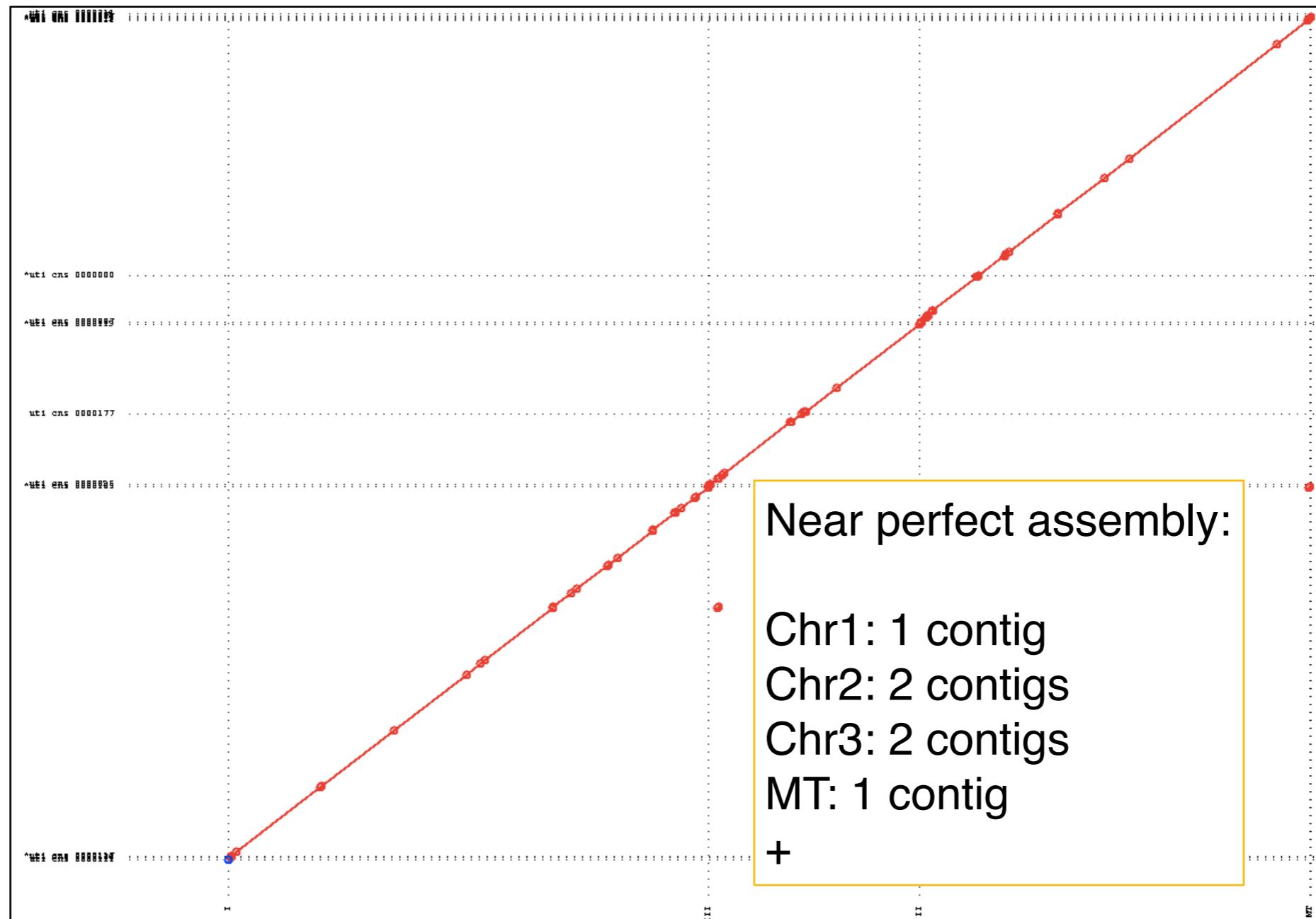
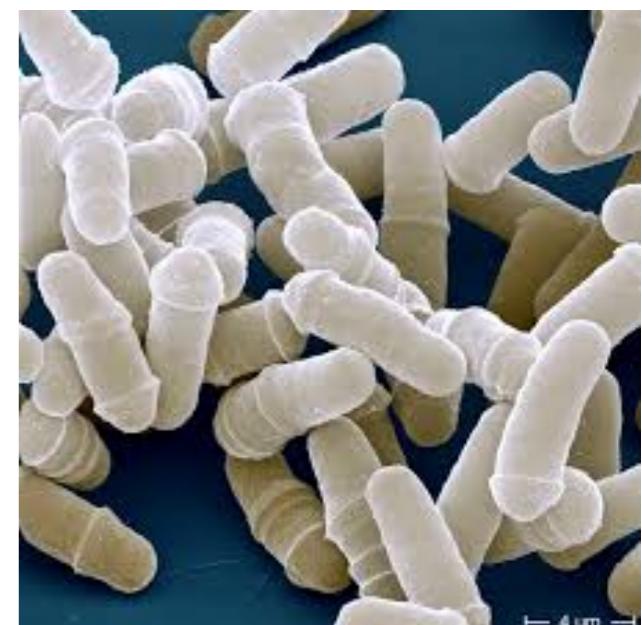
# S. pombe dg21

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

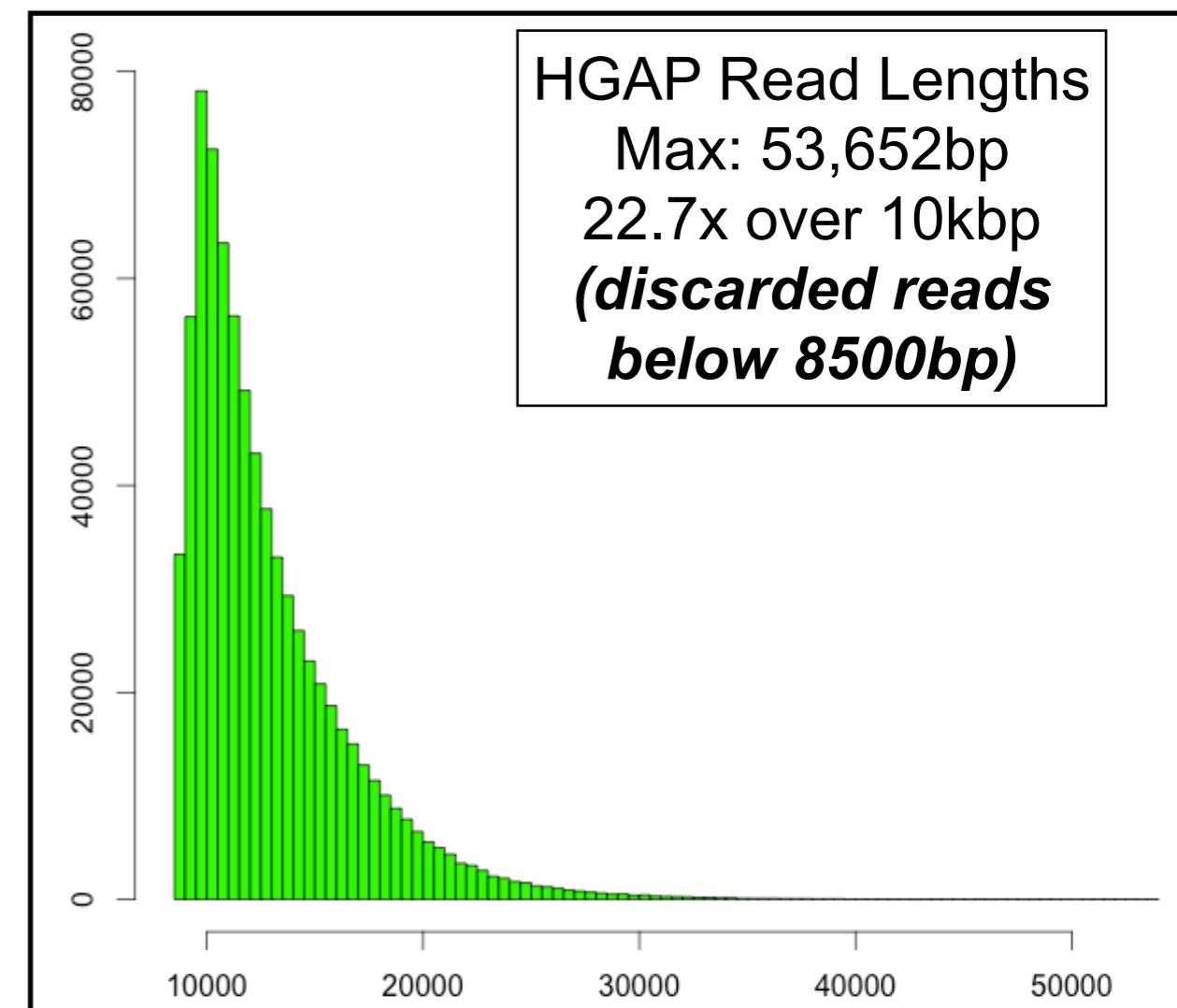


# O. sativa pv Indica (IR64)



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

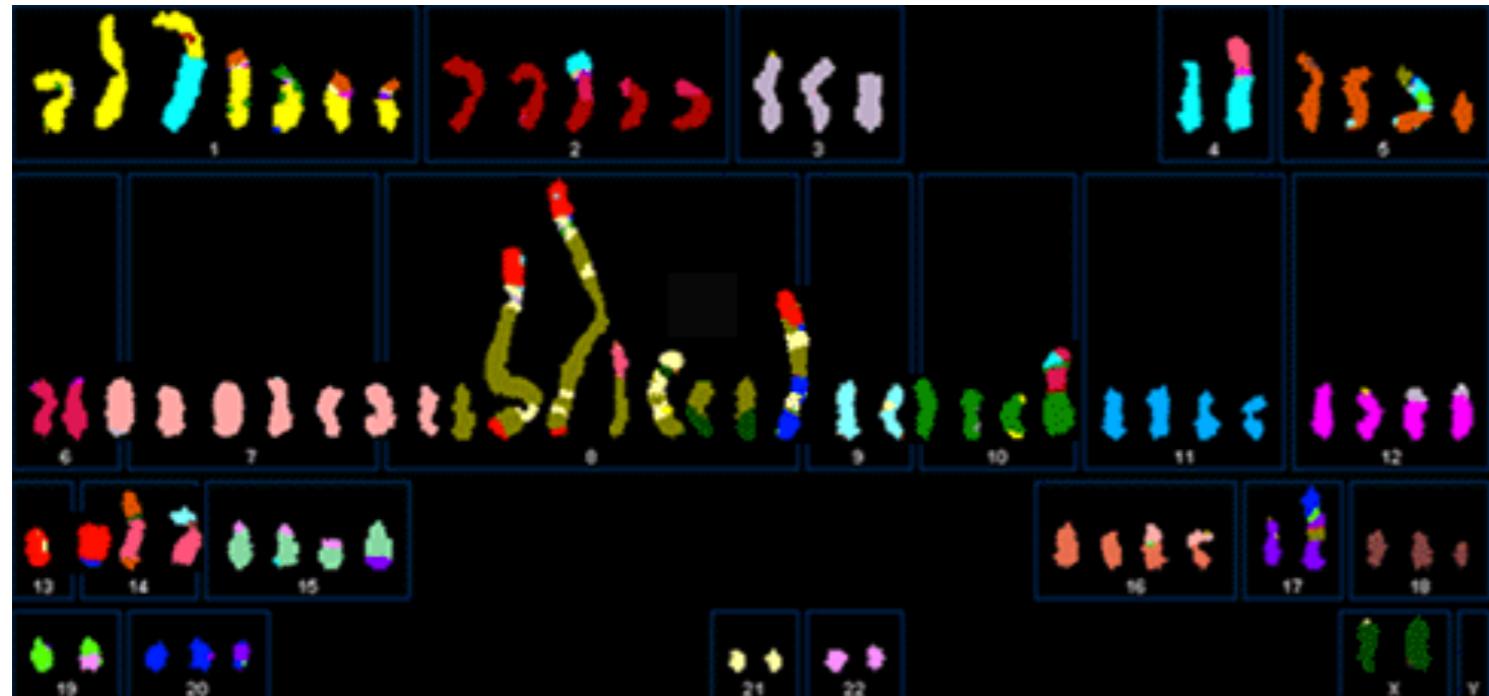


# Structural Variations in SKBR3

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.



Nattestad, et al, Gen. Res. 2018

(Davidson et al, 2000)

# Assembly using PacBio yields far better contiguity

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

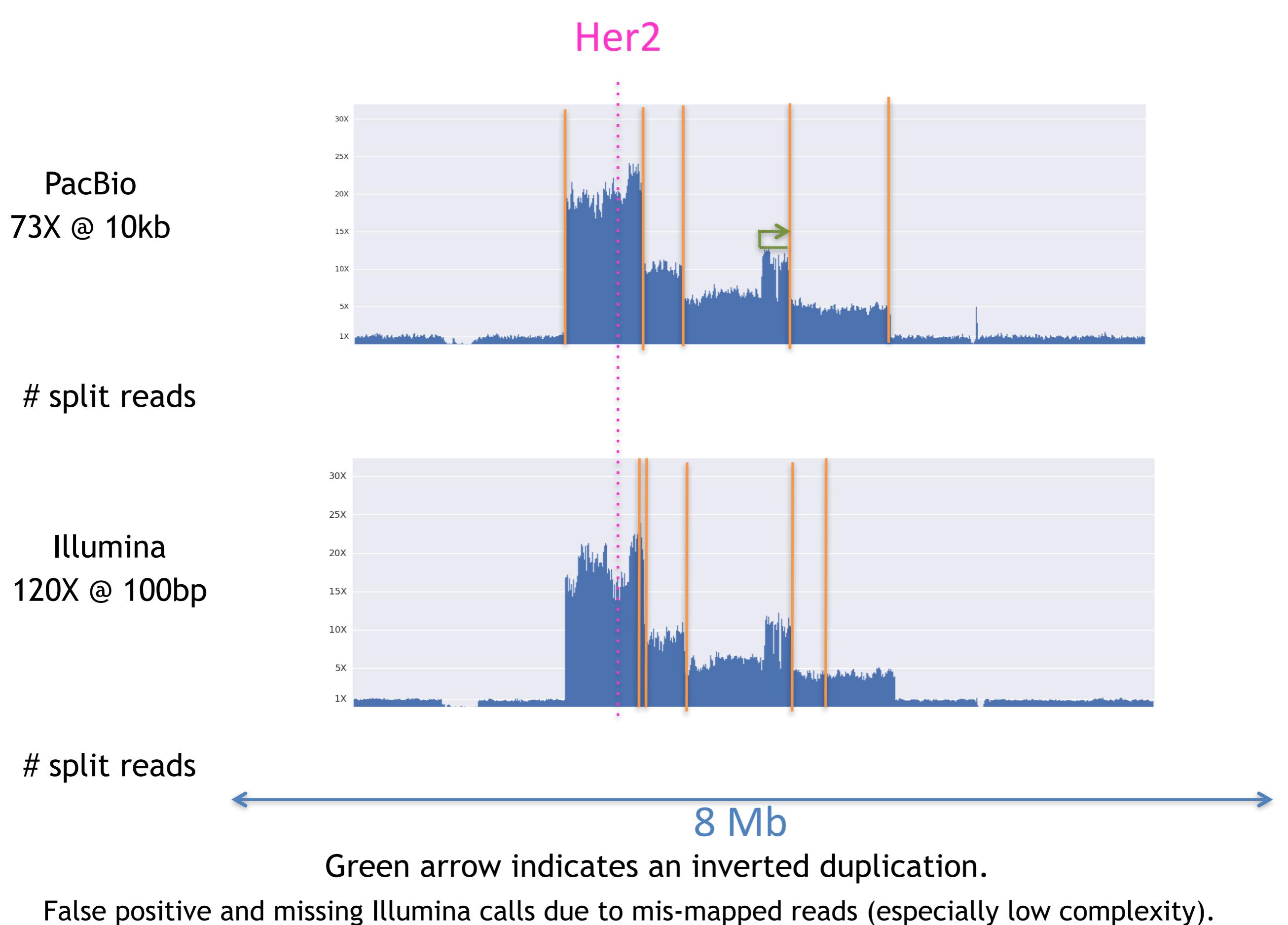
Mean: 2.8 kb

Max: 61 kb

N50: 3.3 kb

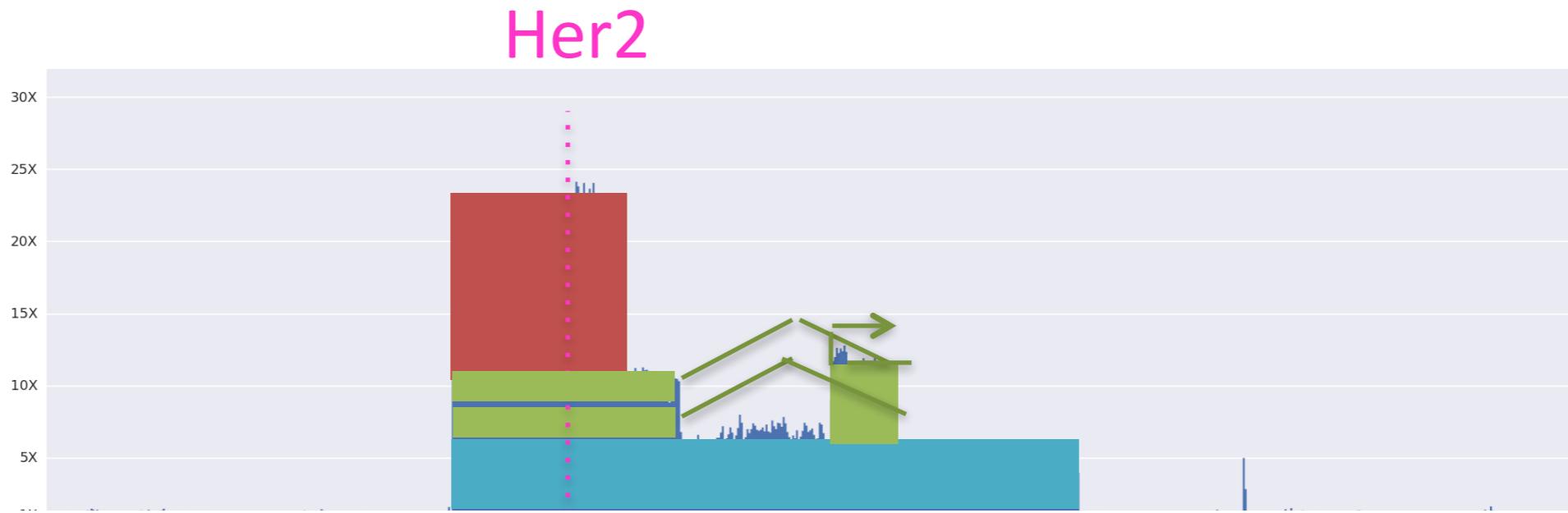
**NG50: 1.9 kb**





# 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

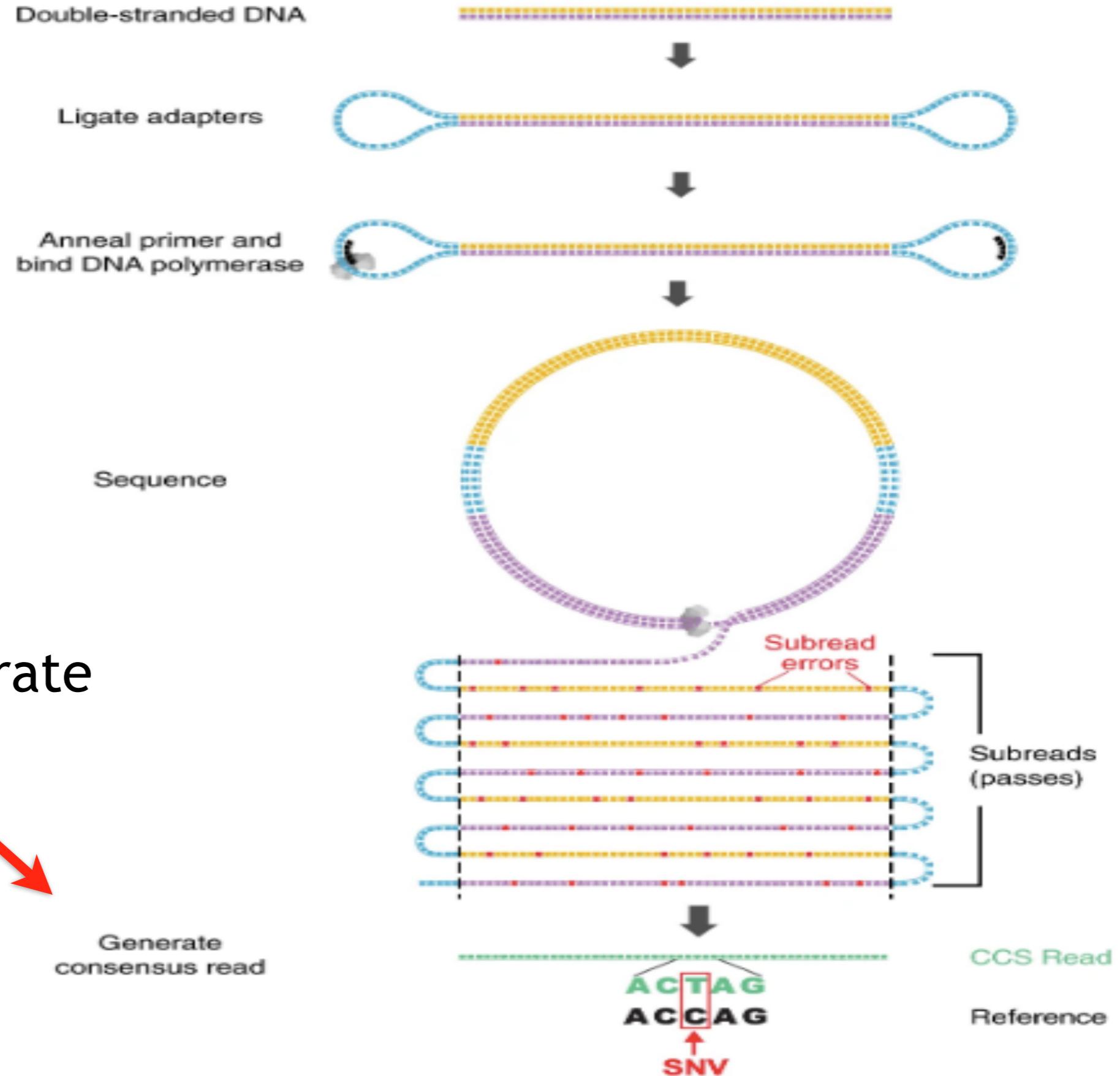
ATGCTCTCGATCGATGCTGCTAGCTAGCTACTAGCTATCCGATCCTACTGACTTACTATGGT



ATGCTCTCGATCGATGCTGCTAGCTAGCTACTAGCTATCCGATCCTACTGACTTACTATGCT

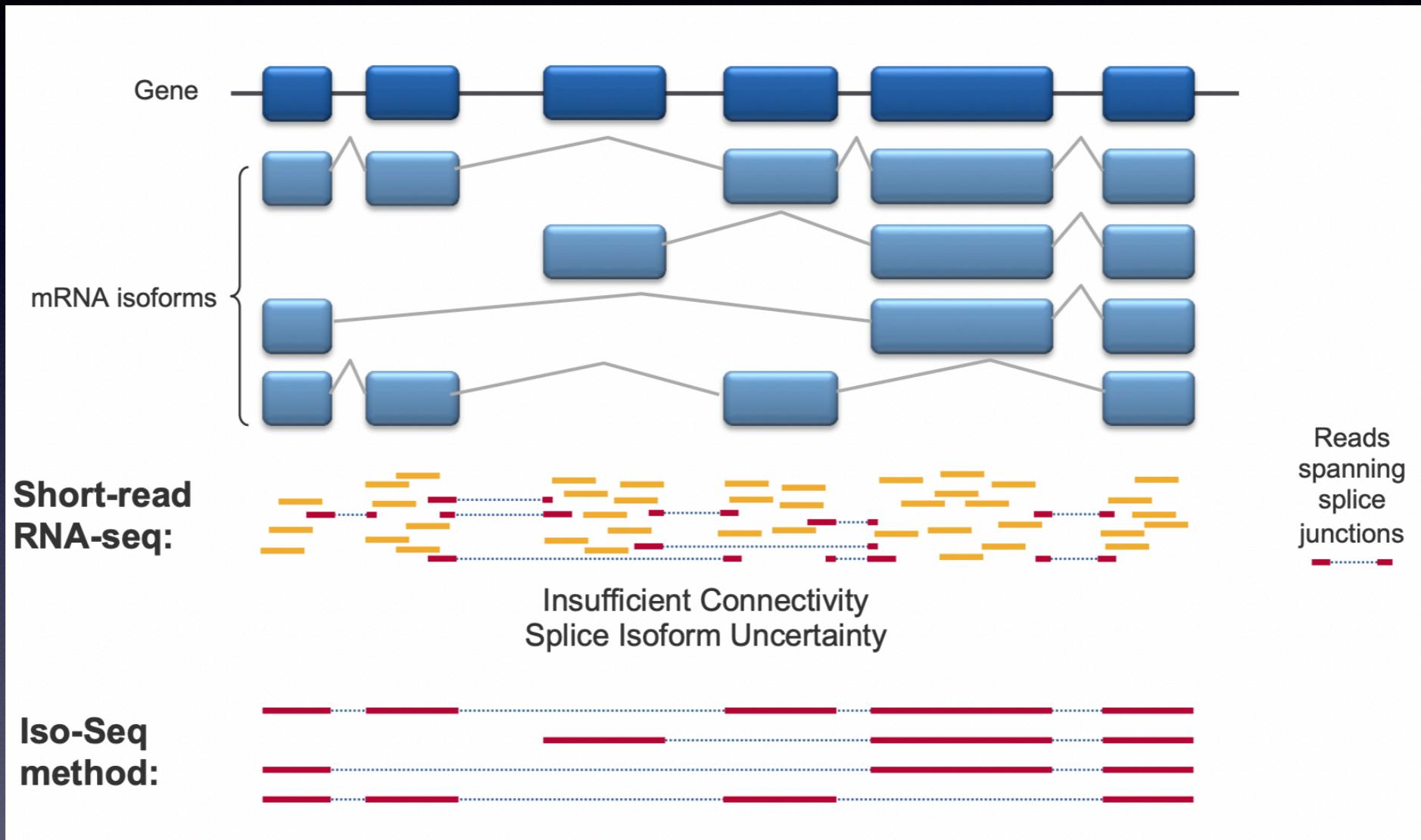
Enough coverage makes error drop out

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



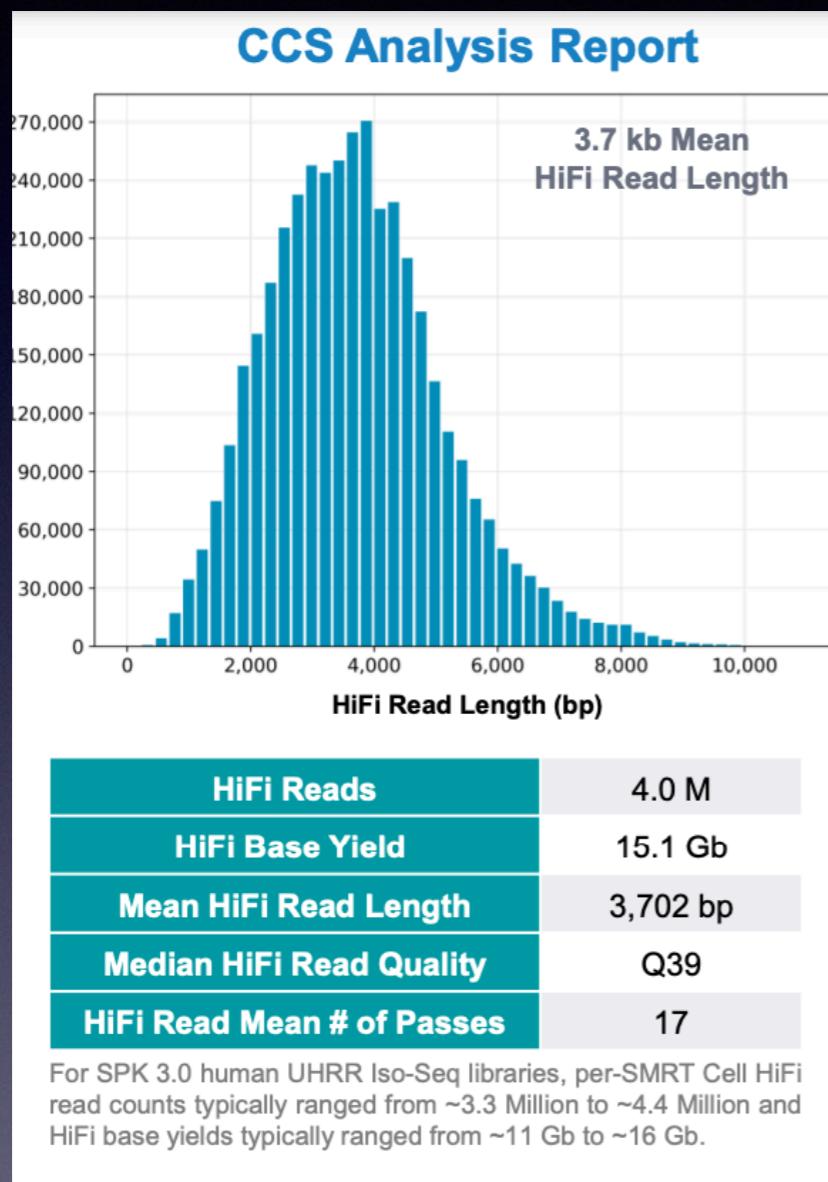
From Wenger et al (2019) Nature Biotechnology

# 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 IsoSeq transcript sequencing



High accuracy, low throughput

>300ng total RNA input

Analysis Metric	Value
HiFi Reads	3,701,876
HiFi Yield (bp)	7,241,572,252
HiFi Read Length (mean, bp)	1,956
HiFi Read Quality (median)	Q40
HiFi Number of Passes (mean)	19
<Q20 Reads	386,924
<Q20 Yield (bp)	887,074,413
<Q20 Read Length (mean, bp)	2,292
<Q20 Read Quality (median)	Q15

CSHL run



# 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

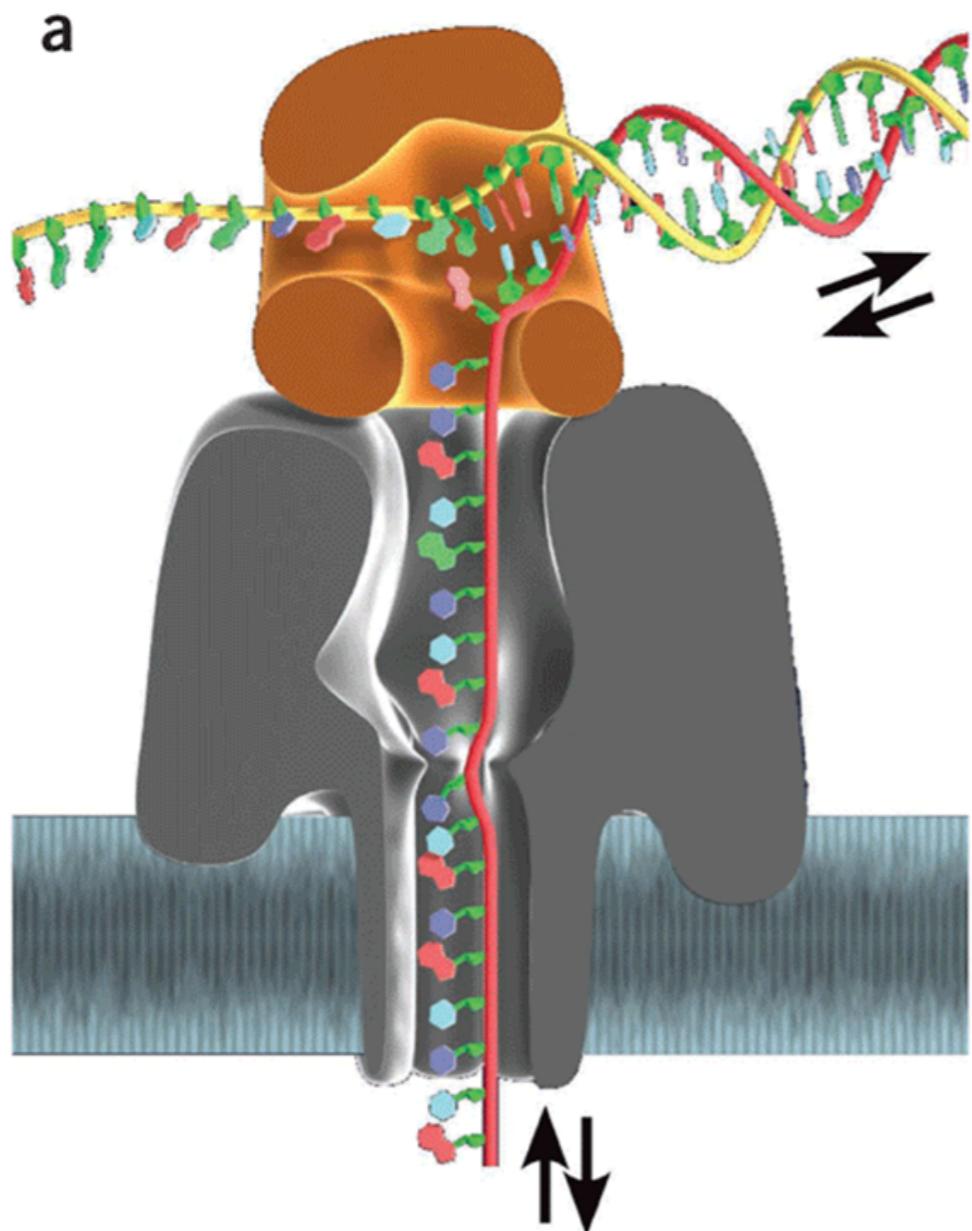
Ultra-long - enables N50s up to 100kb

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

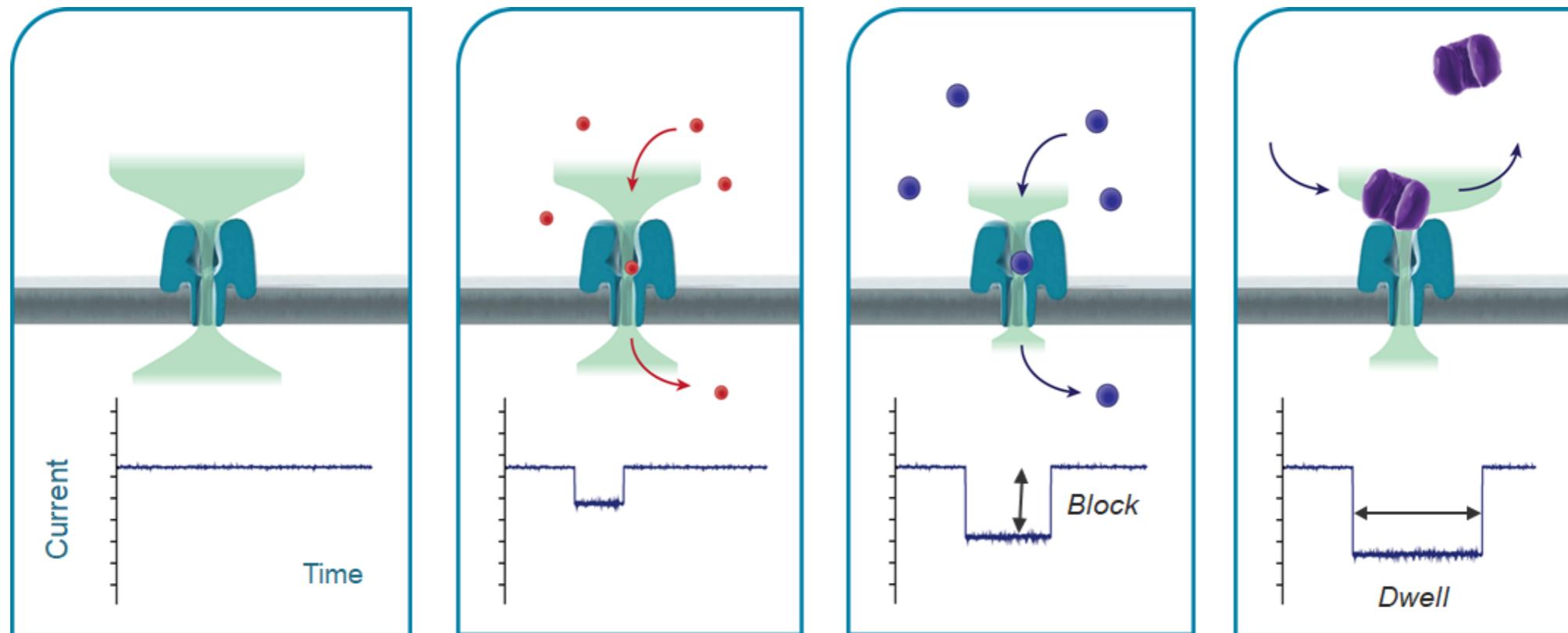
# Nanopore Sensing Summary

Nanopore = ‘very small hole’

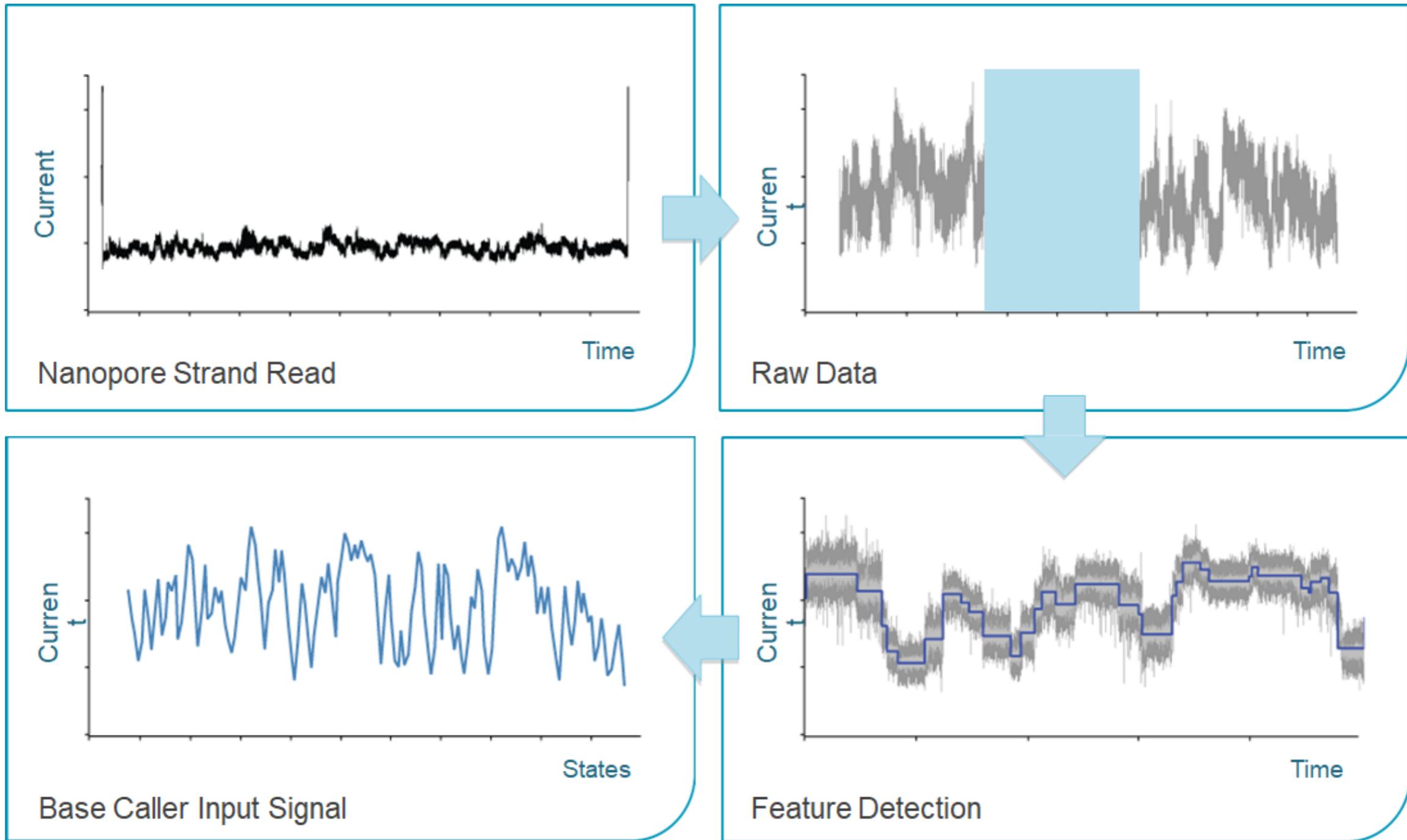
Ionic current flows through the pore Introduce analyte of interest into the pore

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

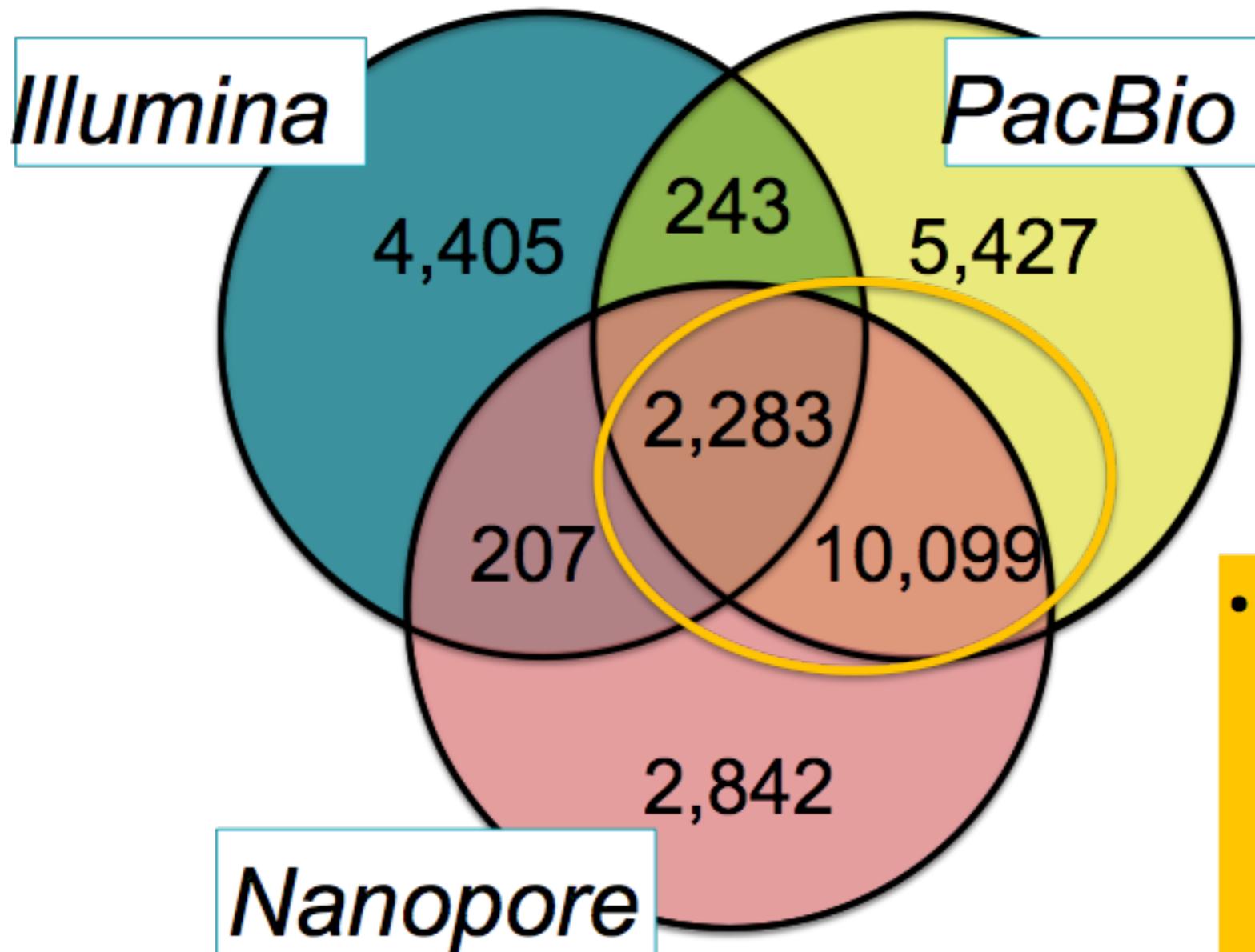
ATGCTCTCGATCGATGCTGCTAGCTAGCTAGCTAGCTTTTT CCGATCCTACTGACTTACTATGGT



ATGCTCTCGATCGATGCTGCTAGCTAGCTAGCTTTTT CCGATCCTACTGACTTACTATGCT

Enough coverage makes error (mostly) drop out

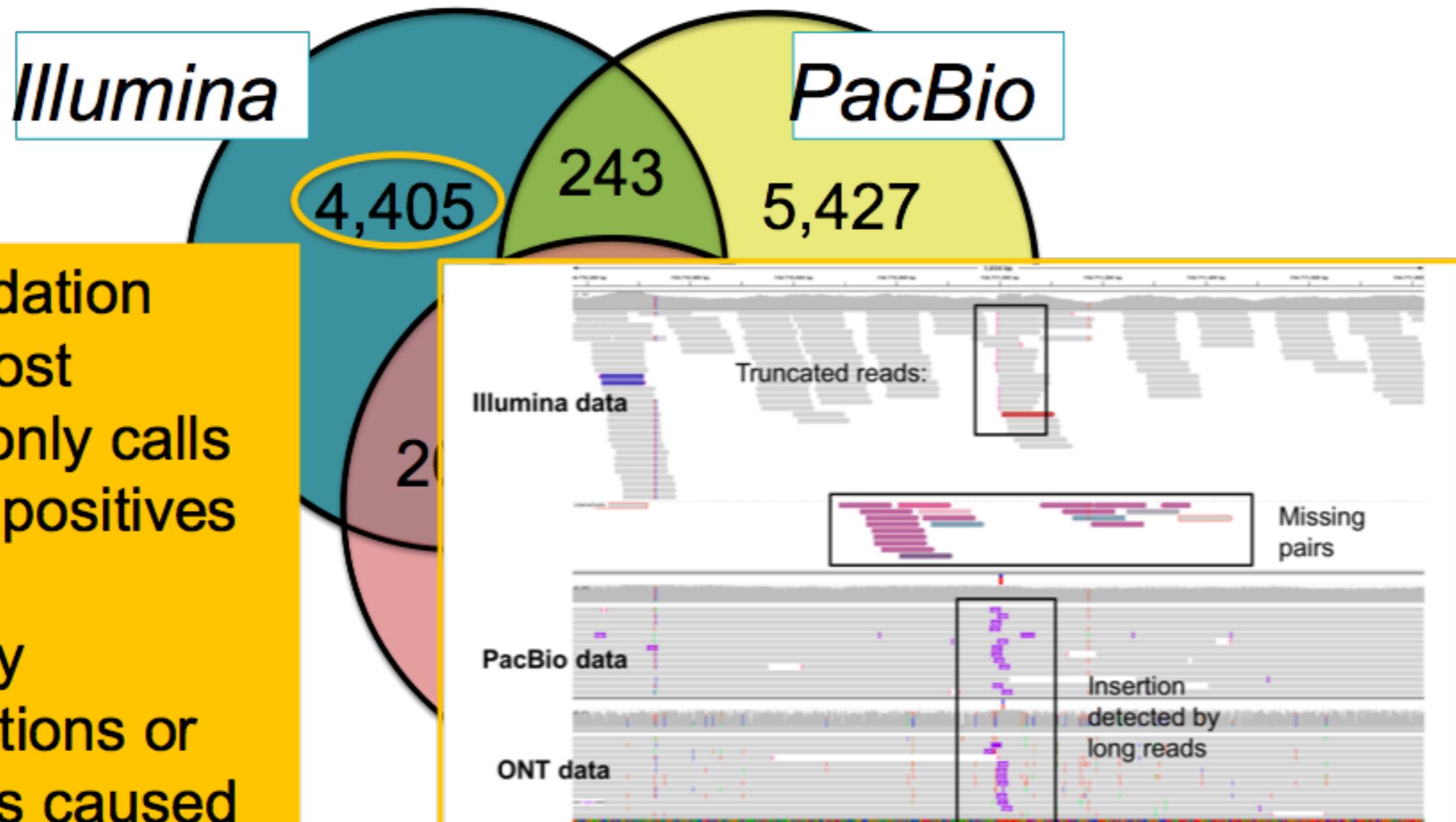
# Structural Variant Comparison of SKBR3



(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

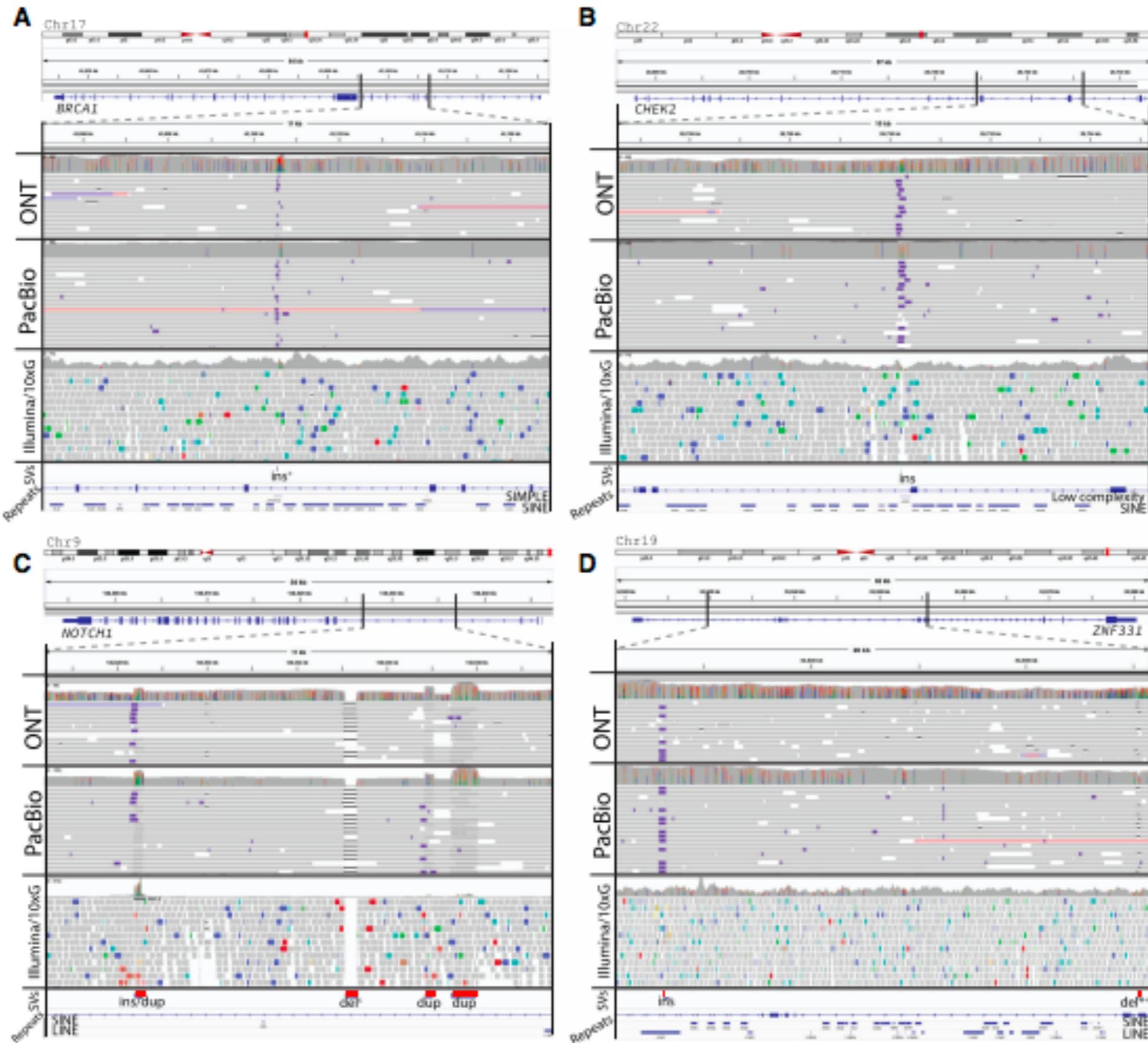
# Preliminary Structural Variations Analysis



	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.



# **Wollemia nobilis Genome Assembly**

## **Previous Assembly with GuppyV3 and wtbg2 assembler**

---

Genome size	15.6 Gbp
No of Contigs	223,812
N50 Contig-size	312 Kbp
Max Contig-size	7 Mbp
Assembly Quality	Q20 (99%)

## **Current Assembly with GuppyV4 and Flye assembler**

---

Genome size	11.56 Gbp
No of Contigs	17,294
N50 Contig-size	9.21 Mbp
Max Contig-size	54.83 Mbp
Assembly Quality	Q31 (99.9%)

Recently published 25Gb Chinese pine genome:  
contig N50 of 2.6 Mb

Niu et al 2022 Cell

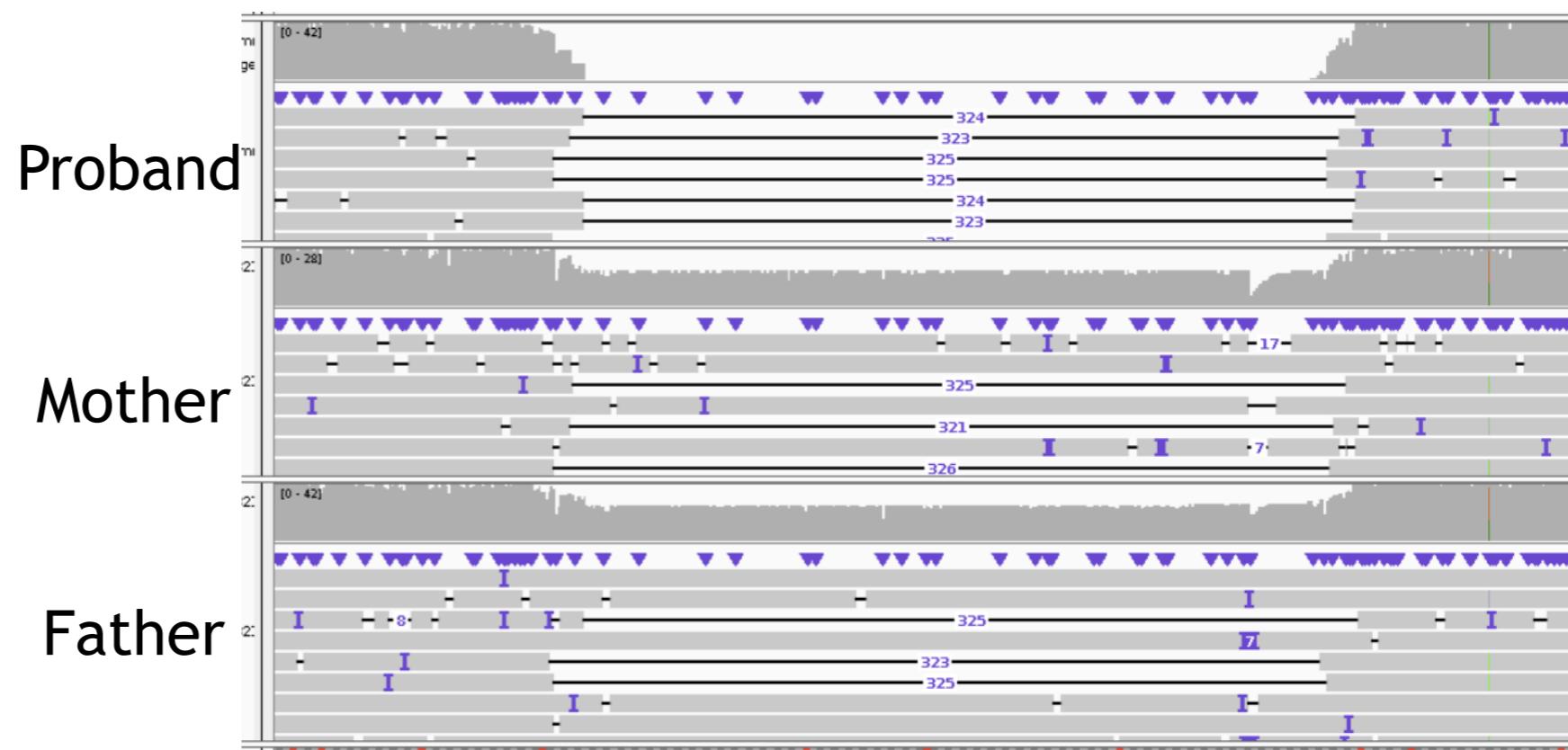
<https://doi.org/10.1016/j.cell.2021.12.006>

# Long Read Sequencing of Early Onset Cancer Pedigrees

## SV Filtering Workflow

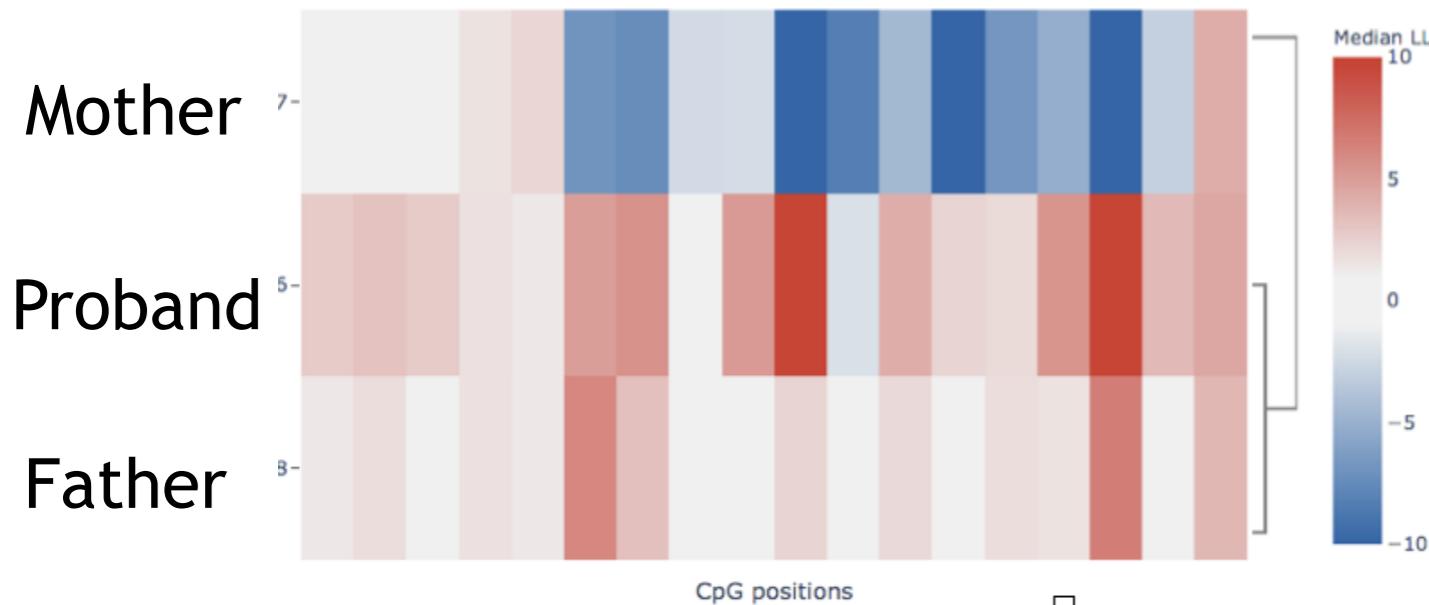
- Each individual will have ~25000 SV calls per genome
- Merge and genotype all calls across family members (~34000)
- Filter by family structure (~1400)
- Pull variants in/near genes (~450)
- Filter common events (~300)
- Filter false positives / ambiguous events/ select likely genes (~<100)

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



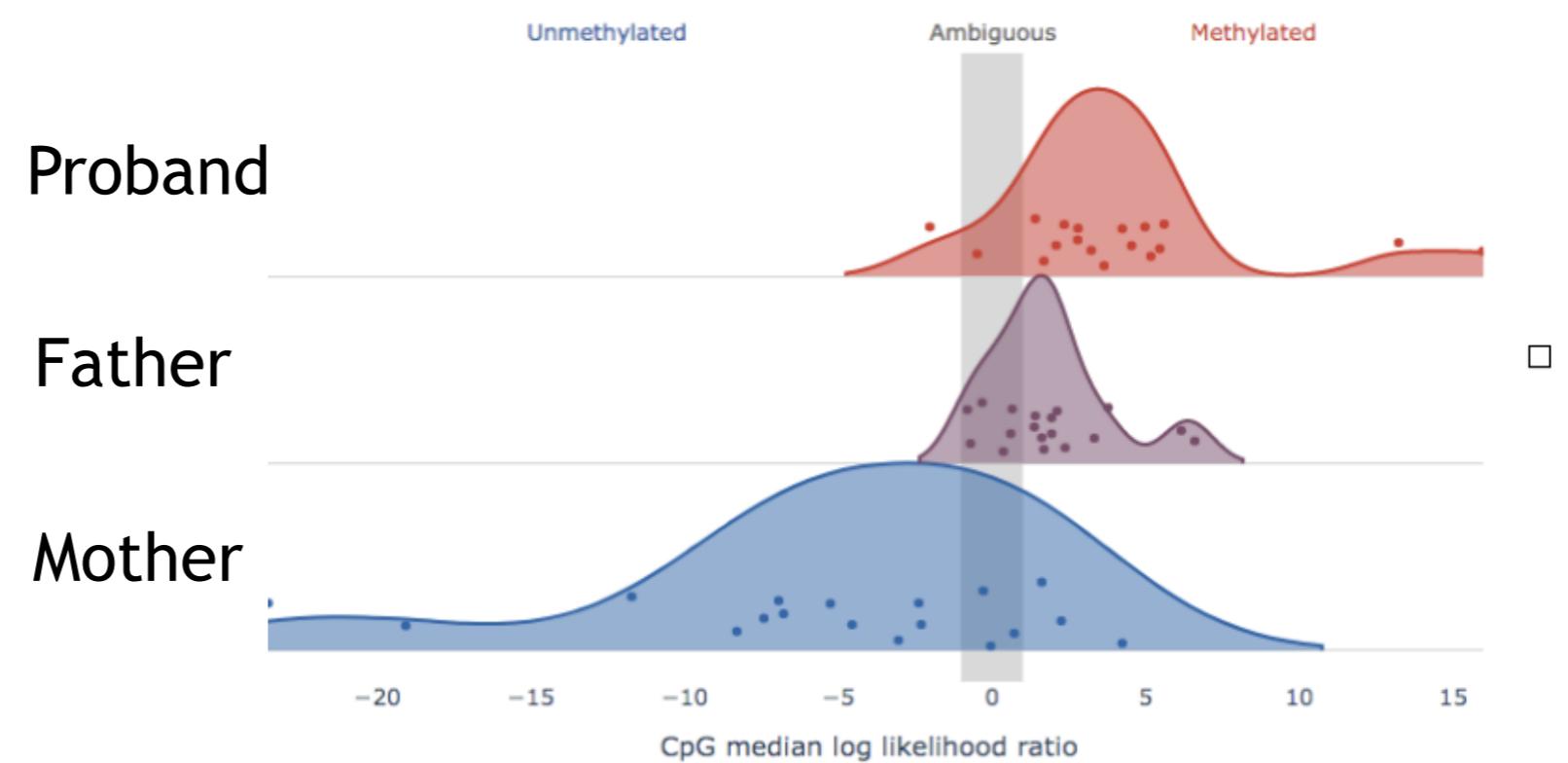
Homozygous intronic gene deletion in proband, heterozygous in healthy parents

# Long Read Sequencing of Early Onset Cancer Pedigrees



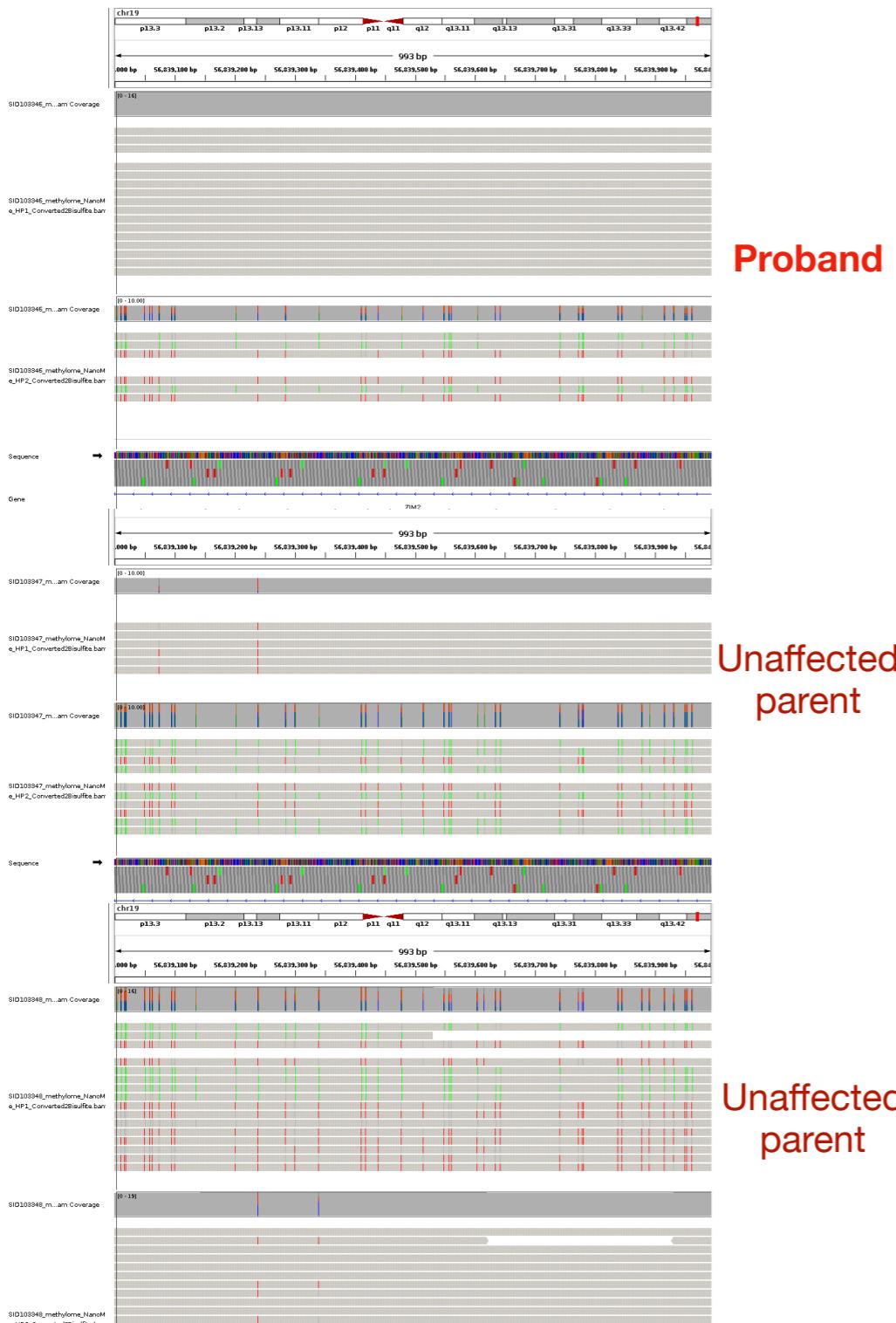
Hypermethylation of promoter region of tumor suppressor in proband compared to healthy parents

ONT signal data allows for direct detection of methylation state

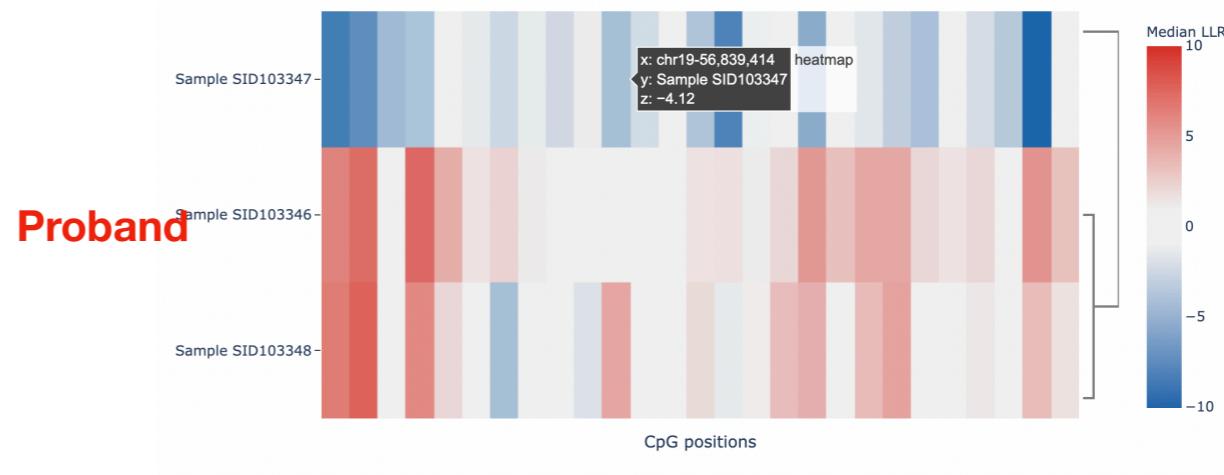


Collaboration with Zsofia Stadler MSKCC

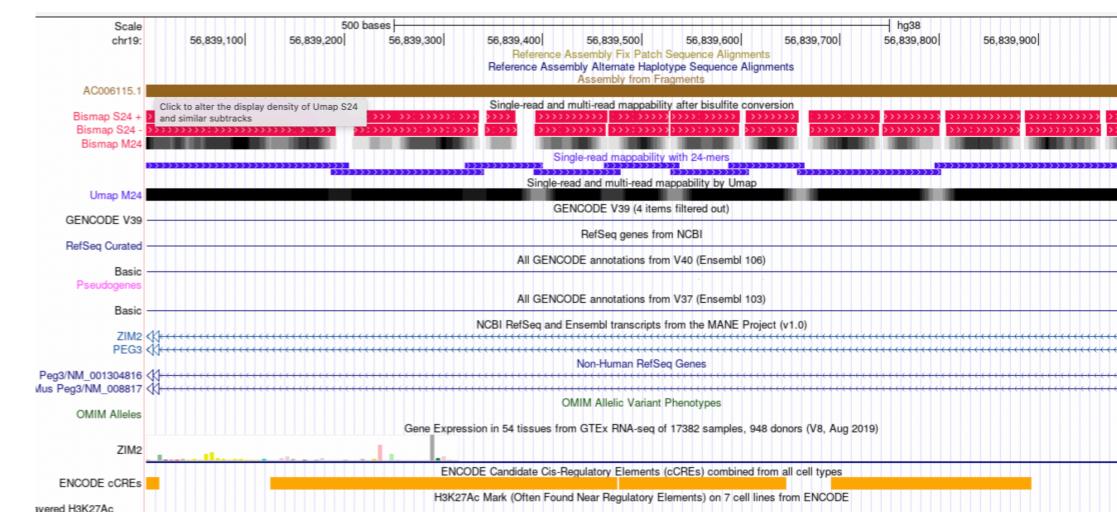
# Phasing methylation provides allele specific context



IGV showing increased methylation in proband (colored bases are “unprotected” converted bases)



# PEG3 downregulated in colon cancer



## Overlaps Enhancer region

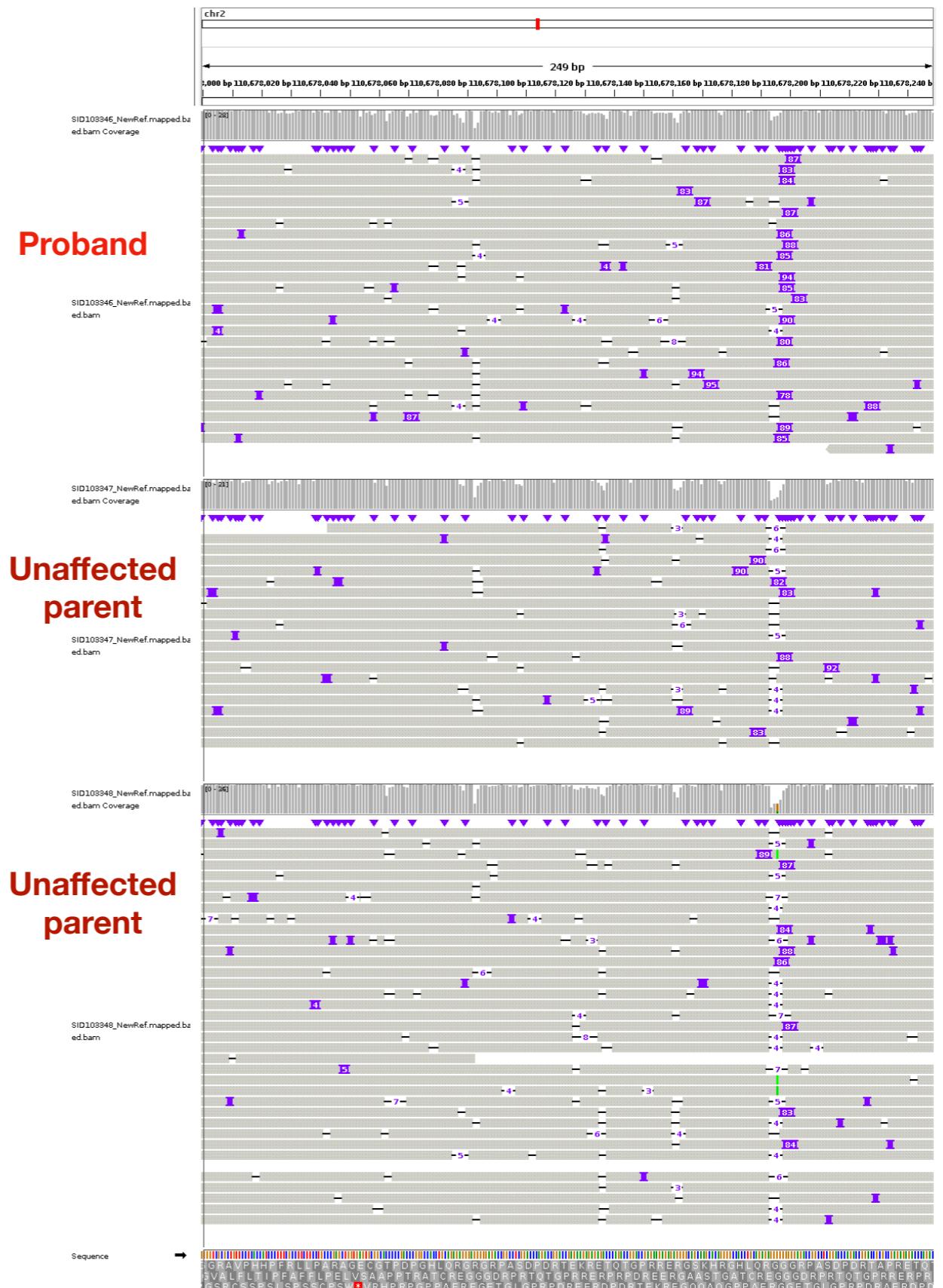
# Improved T2T reference genome uncovers new variants

## Early Onset ColorectalCancer Trio

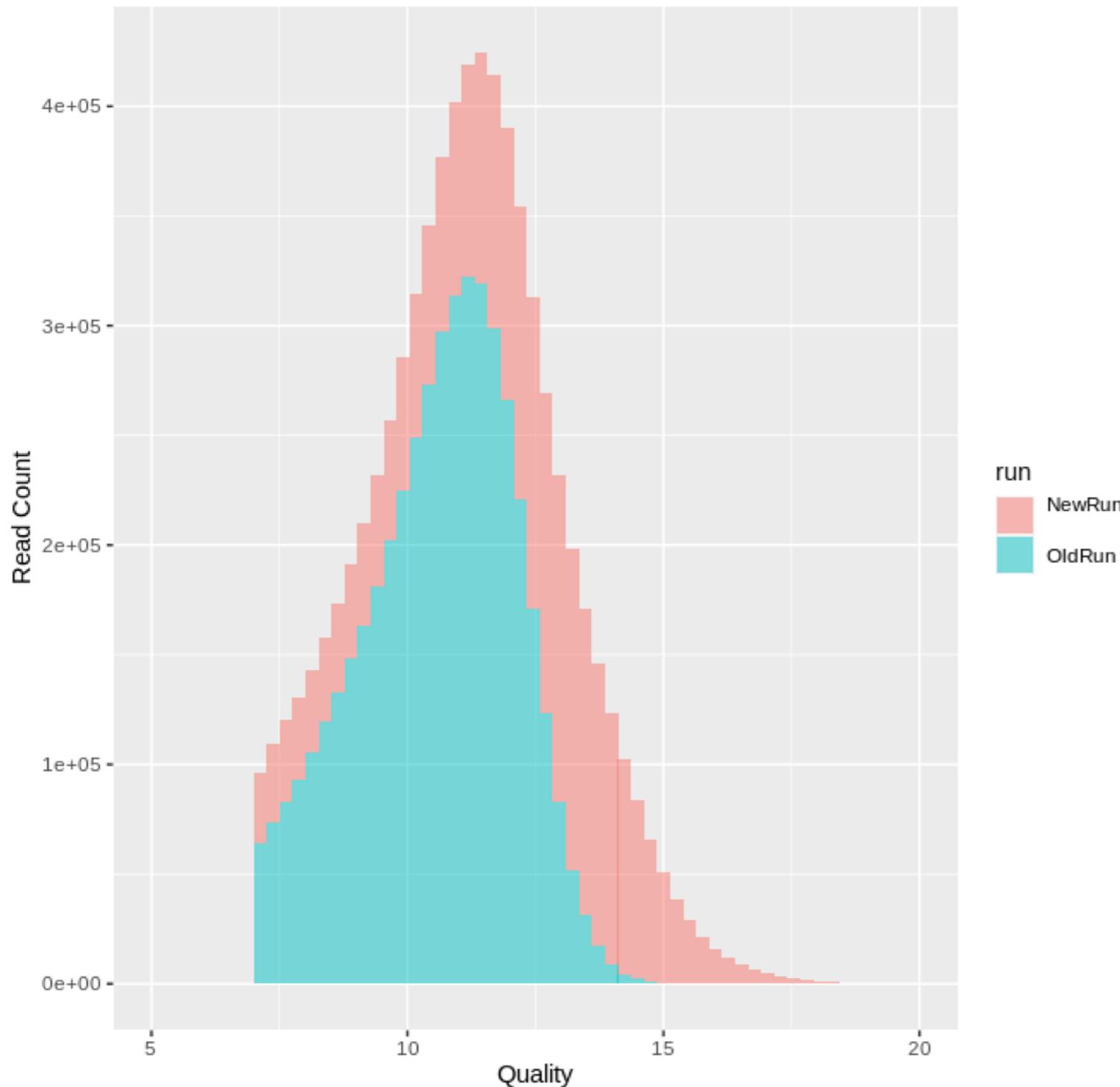
Intronic insertion in MALL gene  
(homozygous in proband)

Region unique to CHM13 compared  
to hg38

MALL expression is reduced in colon  
tumor tissue



## Nanopore Quality Distribution

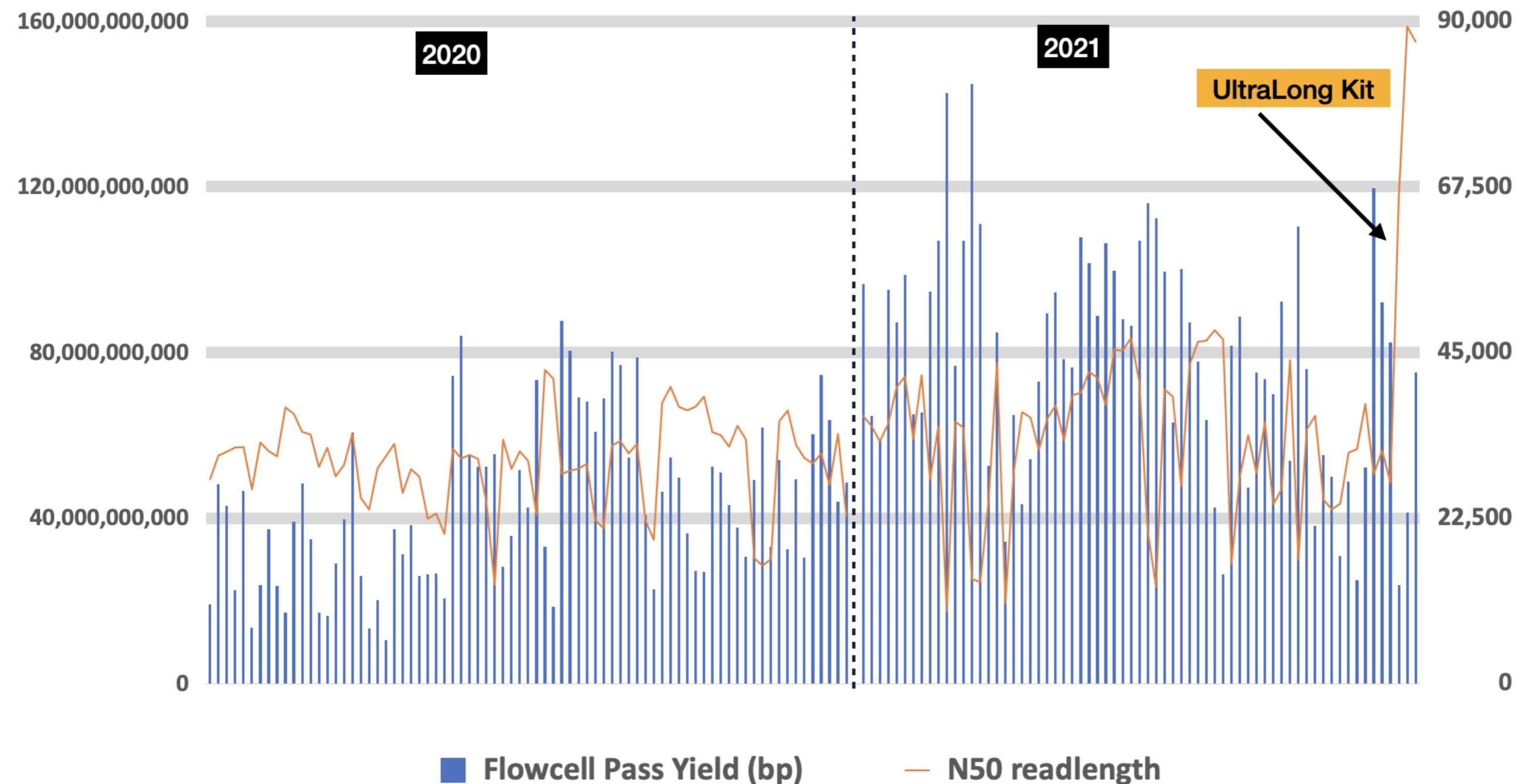


**ONT software  
improvements are  
increasing base quality**

**“Guppy V5” sup accuracy model**

**“Guppy V4” high accuracy model**

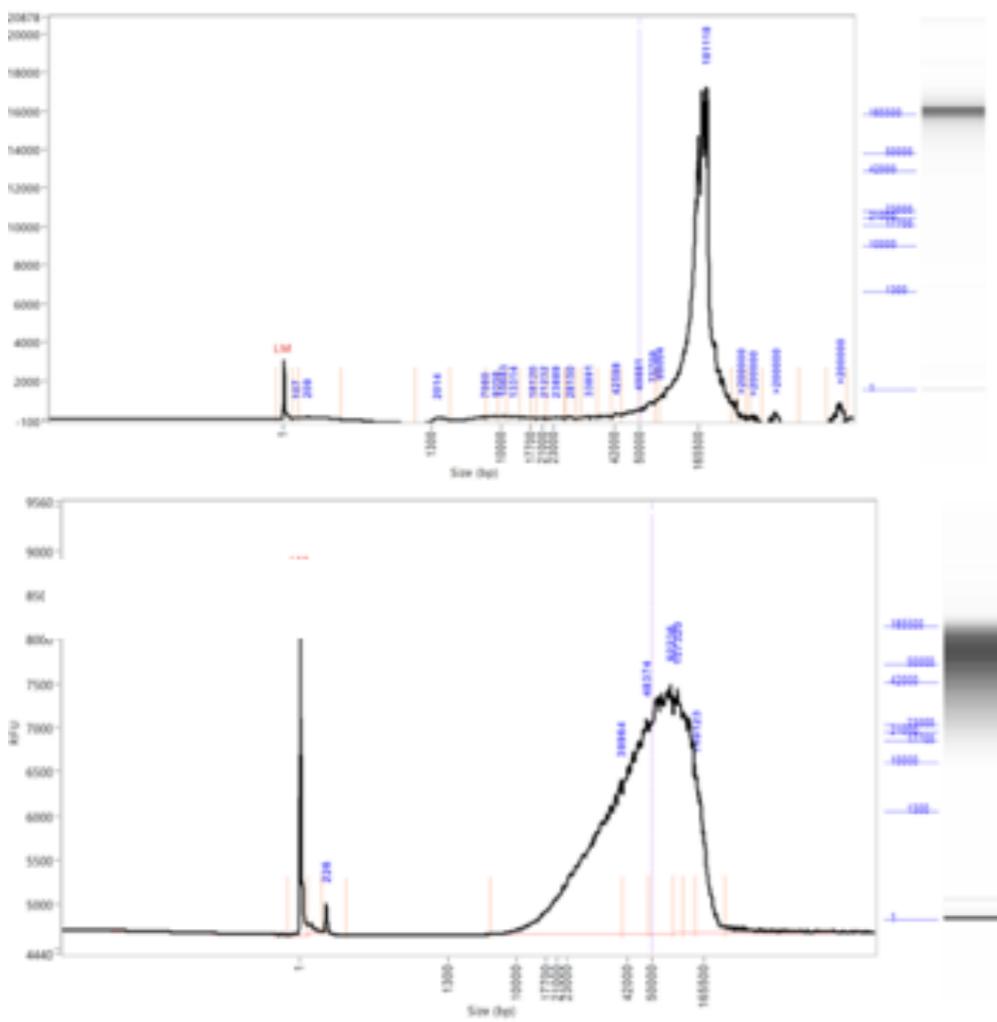
## Yield vs N50



# Optimization of long read sequencing on the PromethION

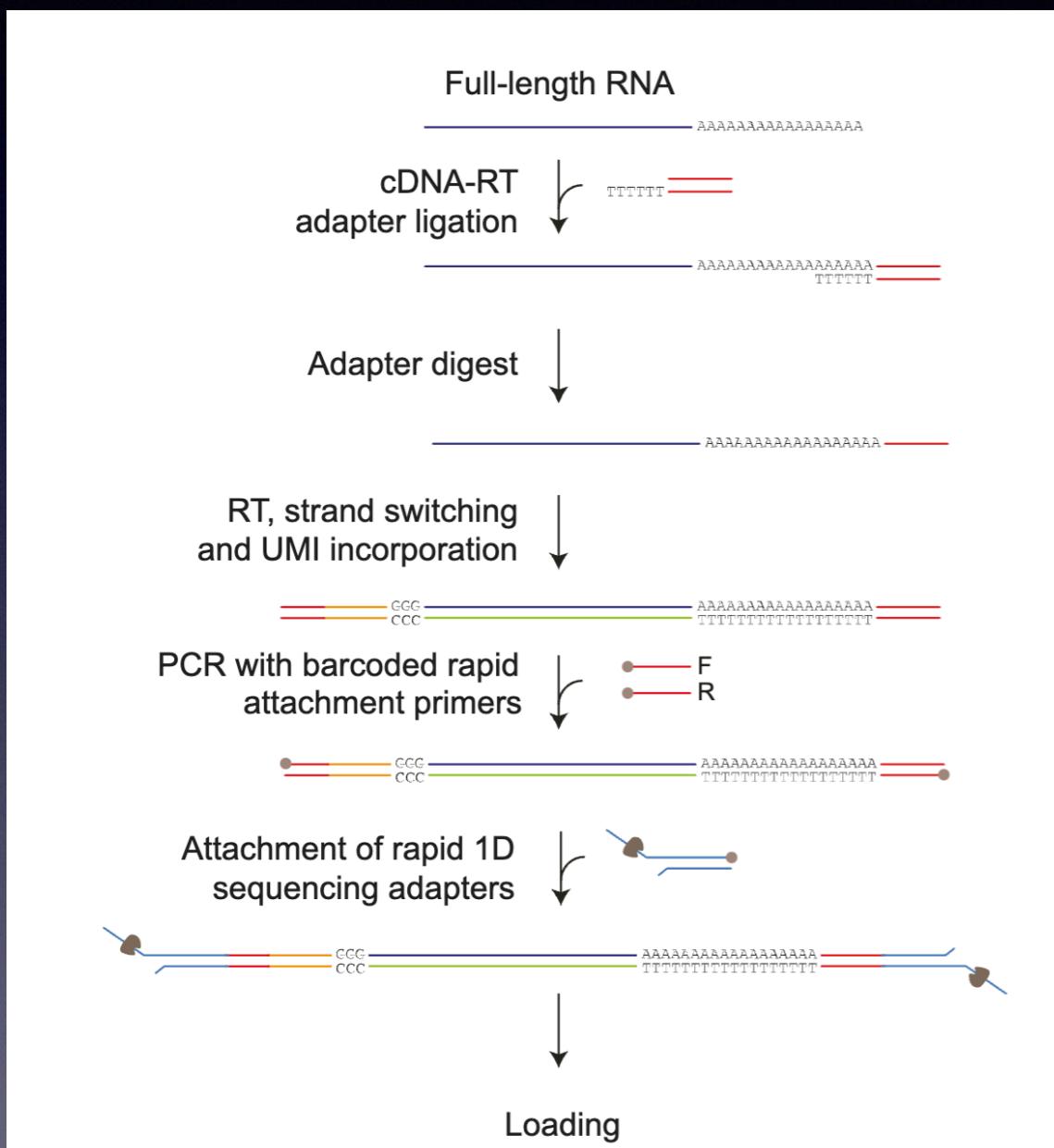


Femto Pulse Fragment Size Estimations  
before and after protocol adjustments for  
shearing and application of SRE



# Transcriptome Sequencing on Oxford Nanopore

## ONT PCR cDNA



Low input (~1ng poly A+)

PCR may introduce biases

Enriched for full length cDNA  
(template switching)

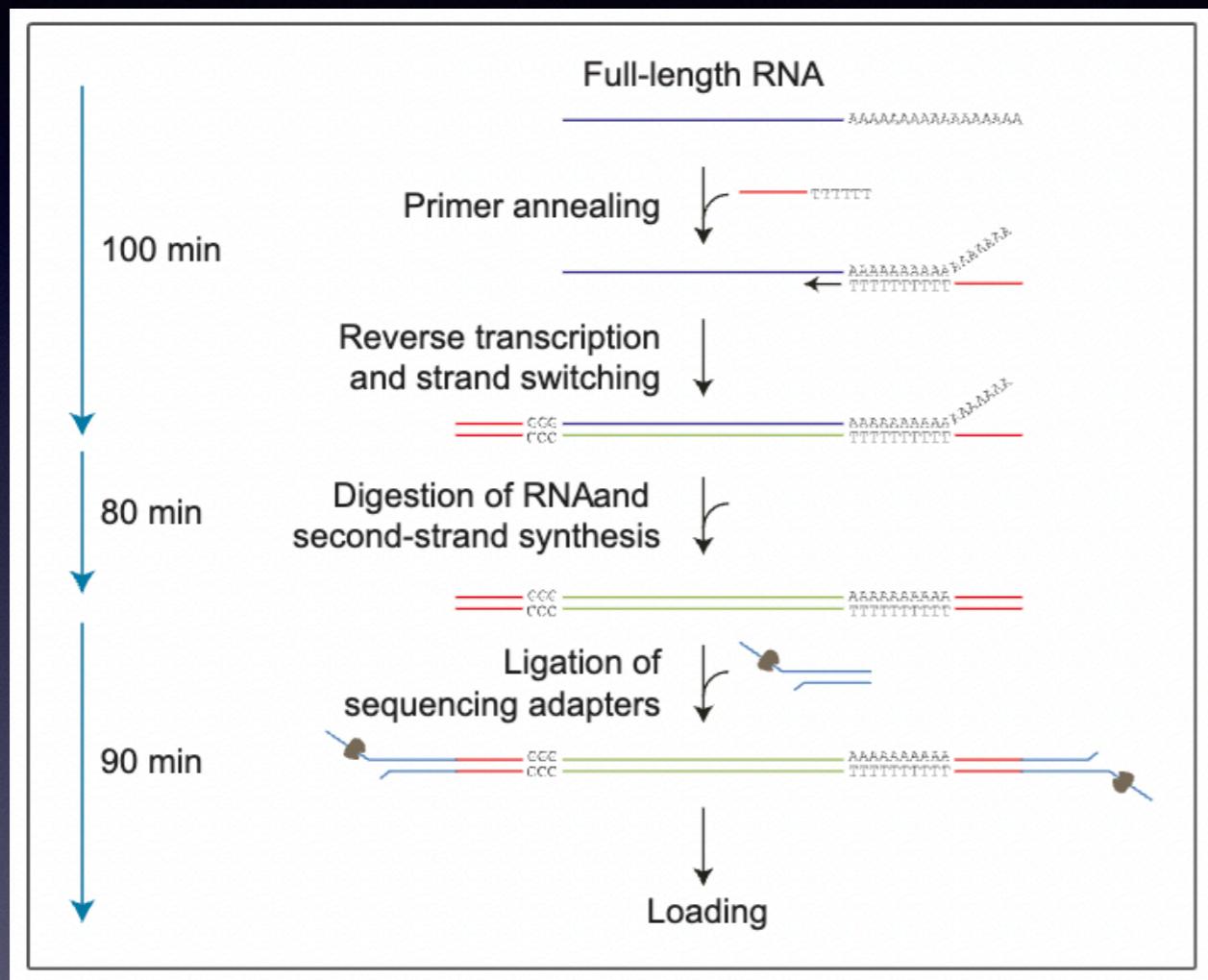
Multiplex up to 24 samples

Much higher throughput (>60 million reads per PromethION cell, up to ~180M)

Lengths ~700bp

Recent paper shows 40 fold fewer long reads/8 fold fewer bases required to cover 6000 genes across 95%

## ONT direct cDNA



Requires more input (~100ng poly A+)

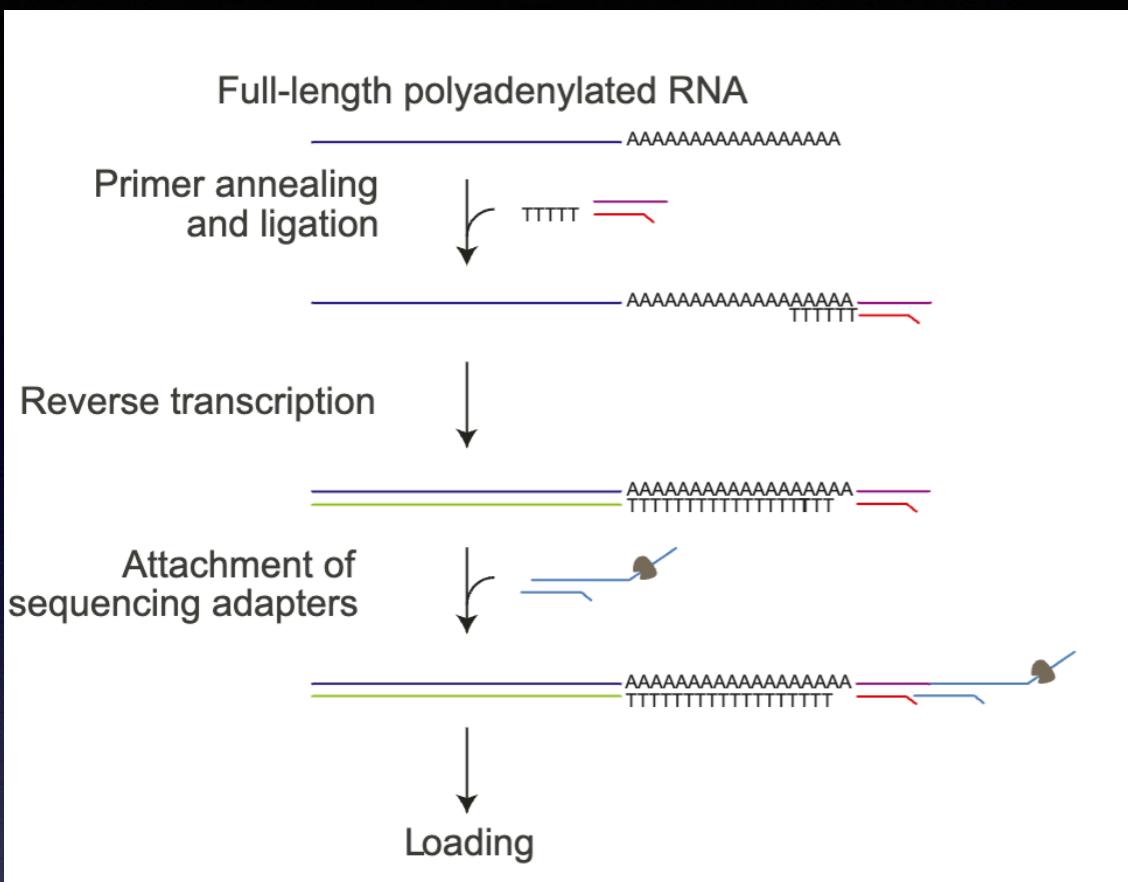
Does not use PCR

Enriched for full length cDNA  
(template switching)

Can multiplex with native barcoding kit

Less throughput (20-30 million reads per PromethION cell), lengths a bit longer ~1.5kb

## ONT direct RNA



Many tools have been/are being developed to use the raw ONT signal data to detect modifications

Tombo (Stoiber et al 2017)  
Nanocompose (Leger et al 2021)  
xPore (Pratanwanich et al 2021)  
nanoRMS (Begik et al 2021)

**Input requirement (500 ng total RNA or 50 ng poly-A+ RNA)**

**RNA length preserved**

**No PCR, RT is optional**

**Can detect base modifications (6mA, 5mC)**

**Output is much lower, 6-8 million reads on PromethION**

**Lengths 1.5-2kb**

**Ribosomal RNA depletion is an issue**

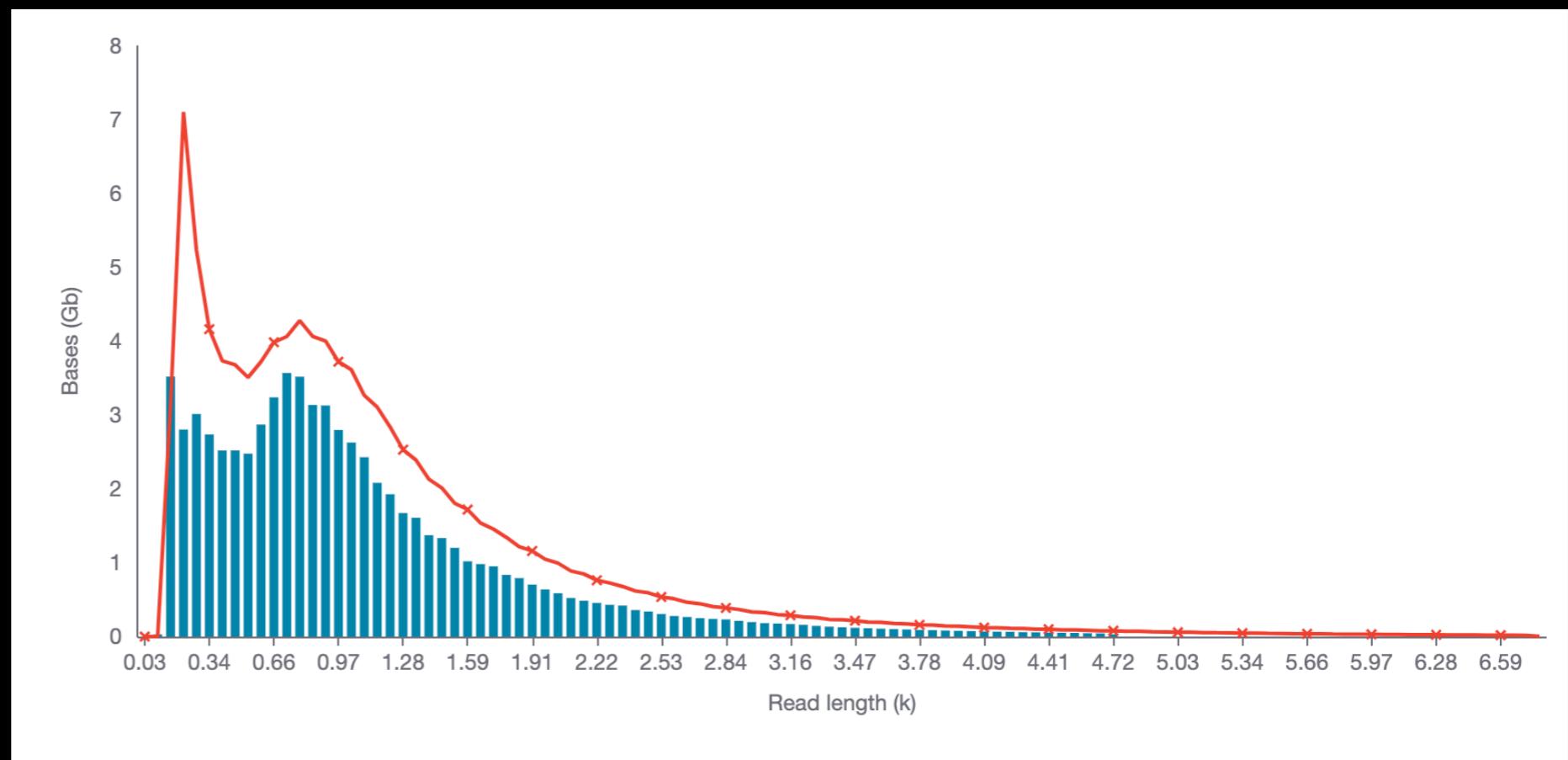
# Recent PromethION sequencing metrics

## PCR cDNA cell

169 million reads

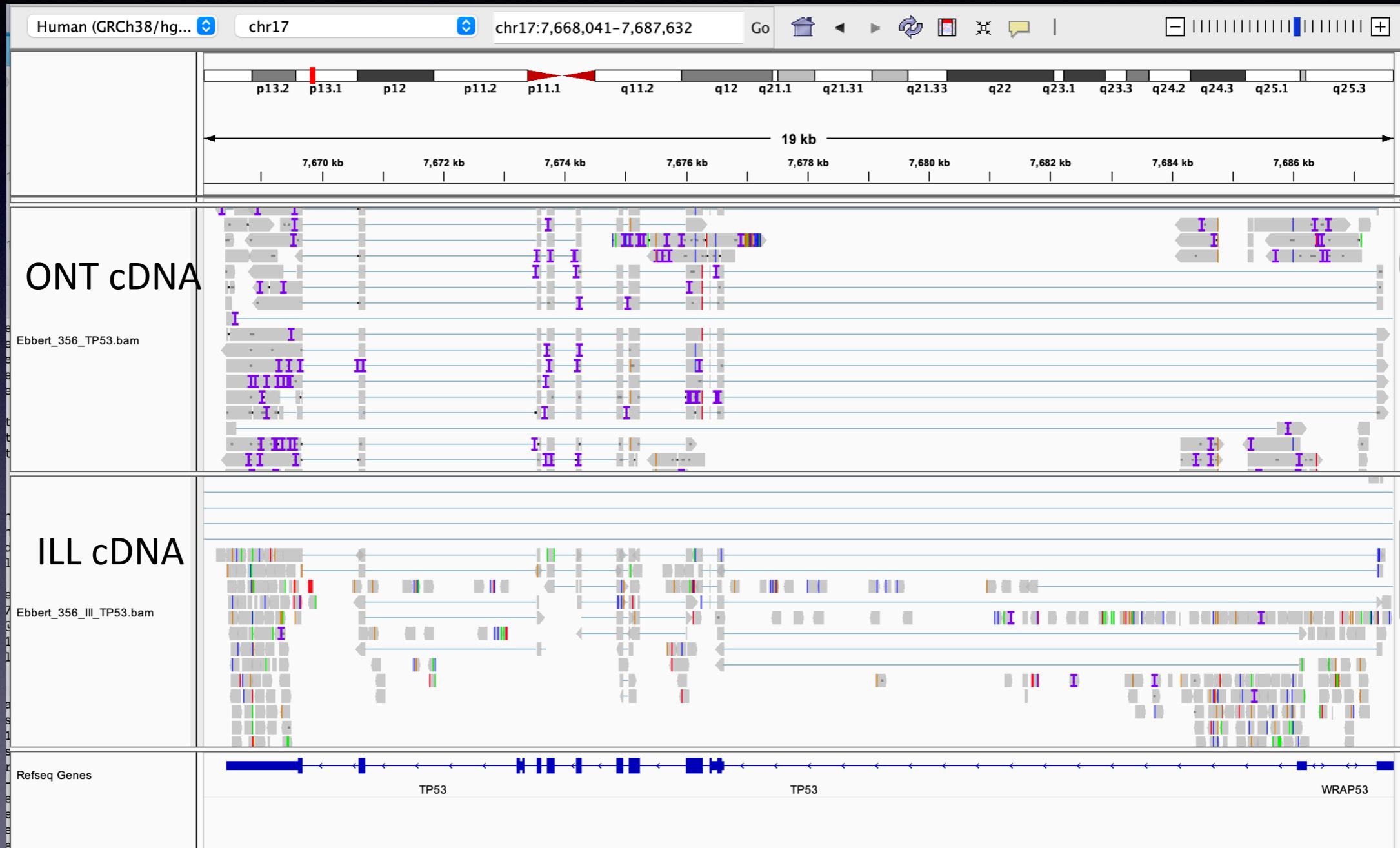
N50 readlength 899bp

108Gb total yield



Average human transcript ~2kb

# Long reads span junctions and provide connection



# Transcript Coverage

StringTie2 (Kovaka 2019) used to assemble transcripts and detect genes where transcripts were fully covered end to end by reads

Illumina data - 16,476 genes

ONT full data set - 25,478 genes

ONT 50pct downsample - 21,322 genes

ONT 75pct downsample - 19,088 genes

# 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  
\$600

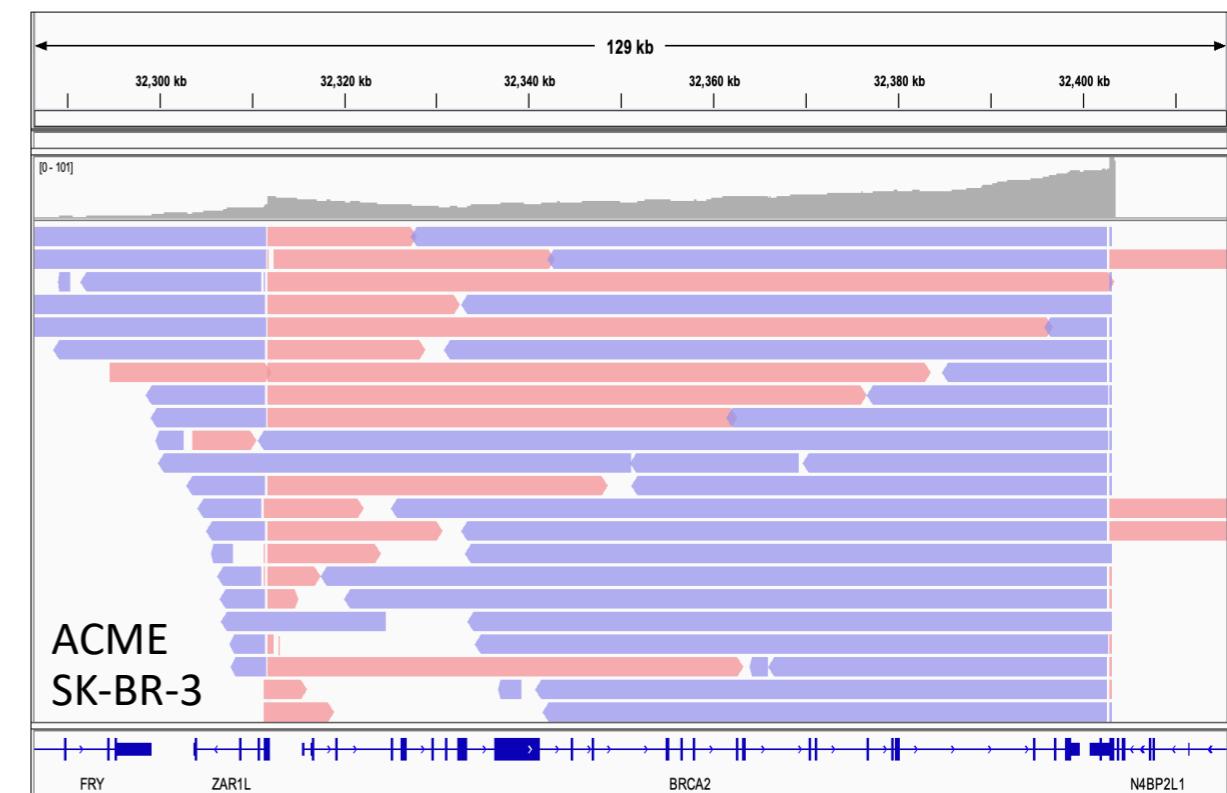
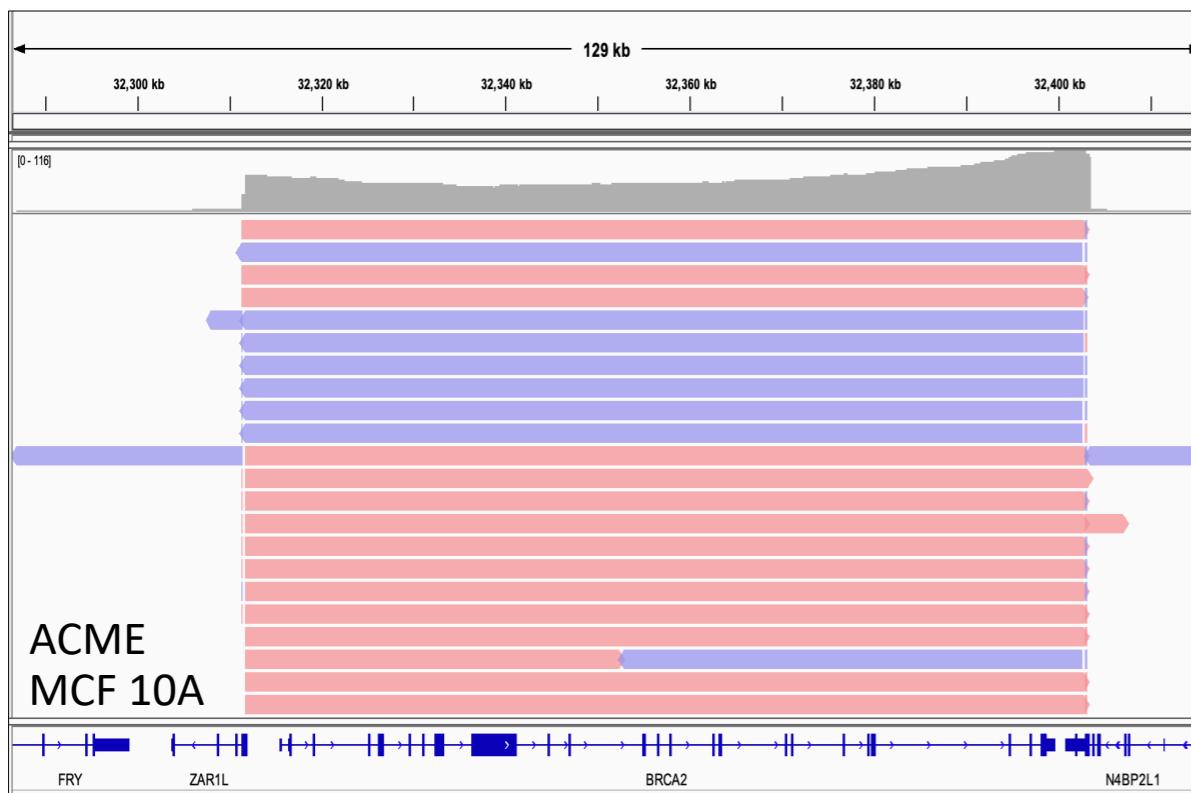
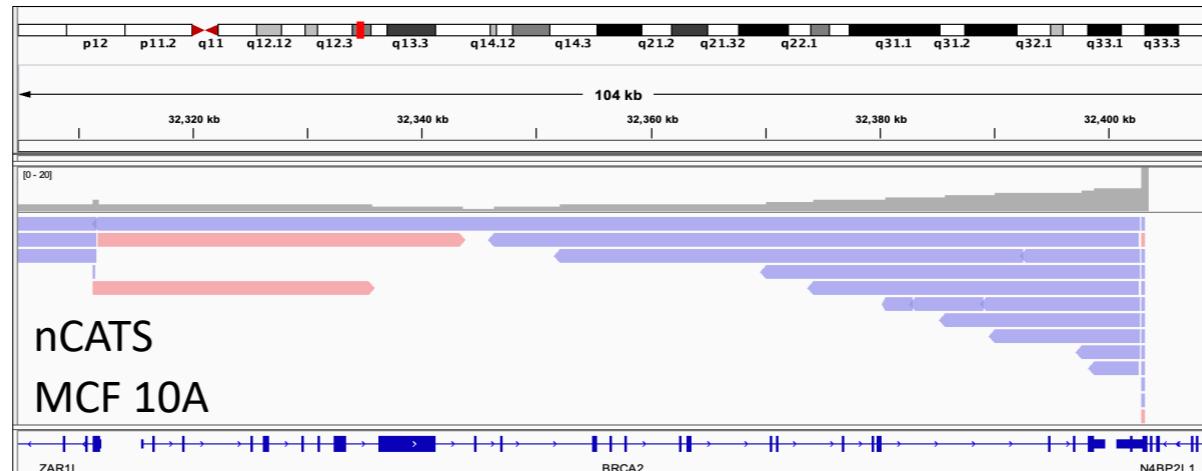
Tradeoffs on accuracy and length, so it  
is key to assess the method that will  
address critical questions of your  
experiment

# Long Read Sequence Capture - Shruti Iyer

---

- Original sequence capture with Illumina used hybridization methods to target exomes or other regions of the genome for Illumina sequencing with very short reads (Hodges et al, Nat Gen. 2007)
- Cancer cells within same sample can be heterogeneous
- Malignant cells can be as low as 10%
  - Subpopulations exhibit different alleles / genomic features
- Detecting subpopulations difficult with 30x WGS
  - Targeted sequencing, exome capture -200 to 500 fold coverage is possible with Illumina sequencing
  - Relative coverage for same cost is higher
  -

# BRCA2 (~90 kb target size)



# Targeting Sequencing Methods for ONT

## Molecular Method

- Enrich for regions of interest from each sample prior to sequencing
- ACME capture, nCATS, PCR

## Computational Method

- Utilize “read until” capability to selectively eject molecules from the pore
- Reads are mapped to a reference in real time
- Enrich for targets or deplete unwanted regions

These methods may be used separately or combined for further enrichment

# 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

Analyse 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  
Elena Ghiban  
Senem Mavruk  
Shruti Iyer

## Spector Lab

Gayatri Arun  
Sonam Bhatia

## Siepel Lab

Armin Scheben

## Schatz Lab

Sam Kovaka  
Michael Kirsche  
Rachel Sherman  
Katie Jenike  
Sergey Aganezov  
Srividya Ramakrishnan

## Timp Lab

Isac Lee

## **Fritz Sedlazeck   Karen Kostroff**

Medhat Helmy

## Funding

MaizeCODE consortium  
Living Fossils consortium

AMNH  
Nancy Simmons  
Sara Oppenheim

NCI

NSF

NHGRI

Northwell  
Health