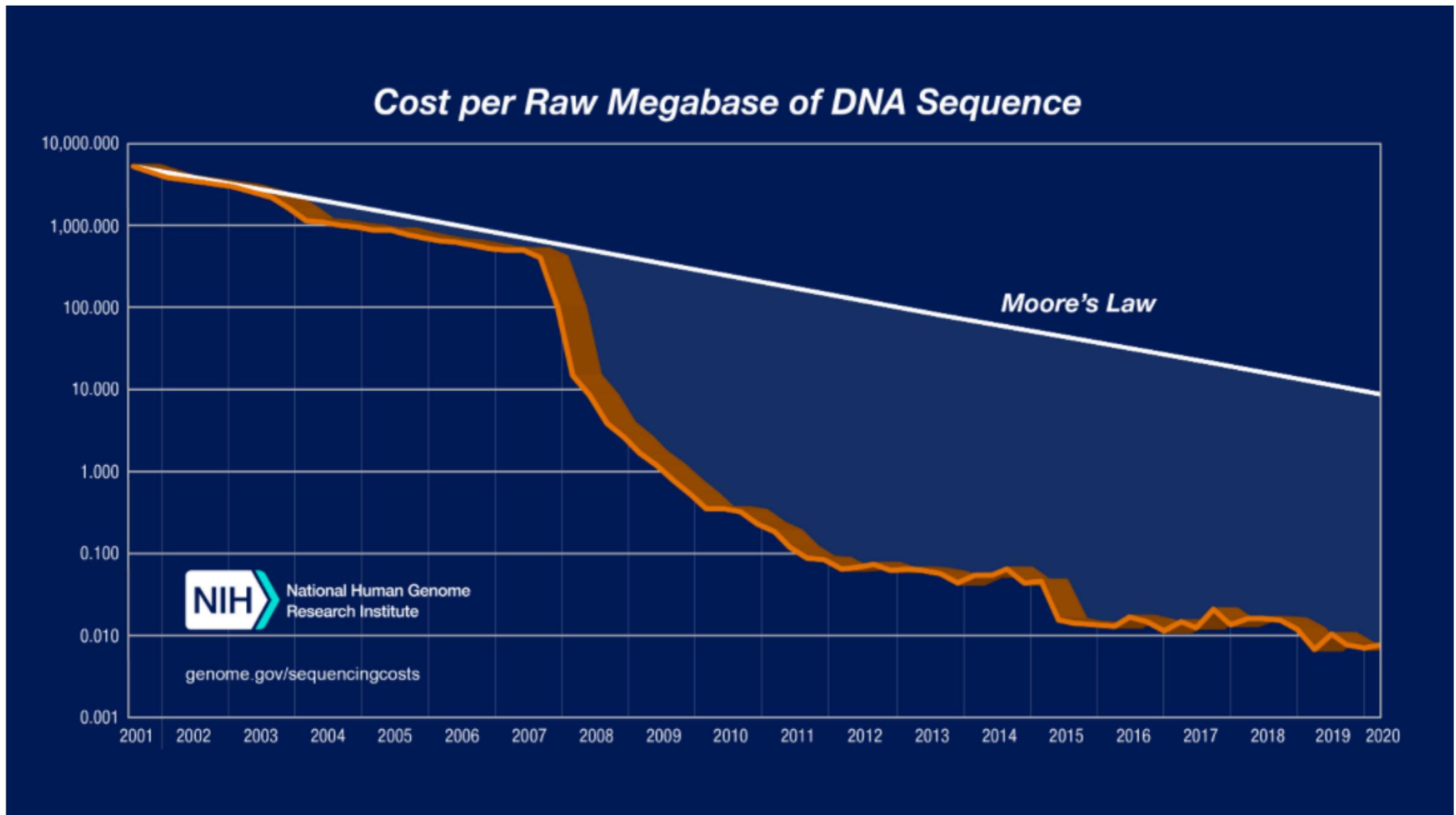


Long Read Sequencing

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

Advanced Sequencing Technologies and Applications course
Cold Spring Harbor Laboratory
2021

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 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

Less fold coverage required?

Finished genomes - T2T

Limitations of long reads

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

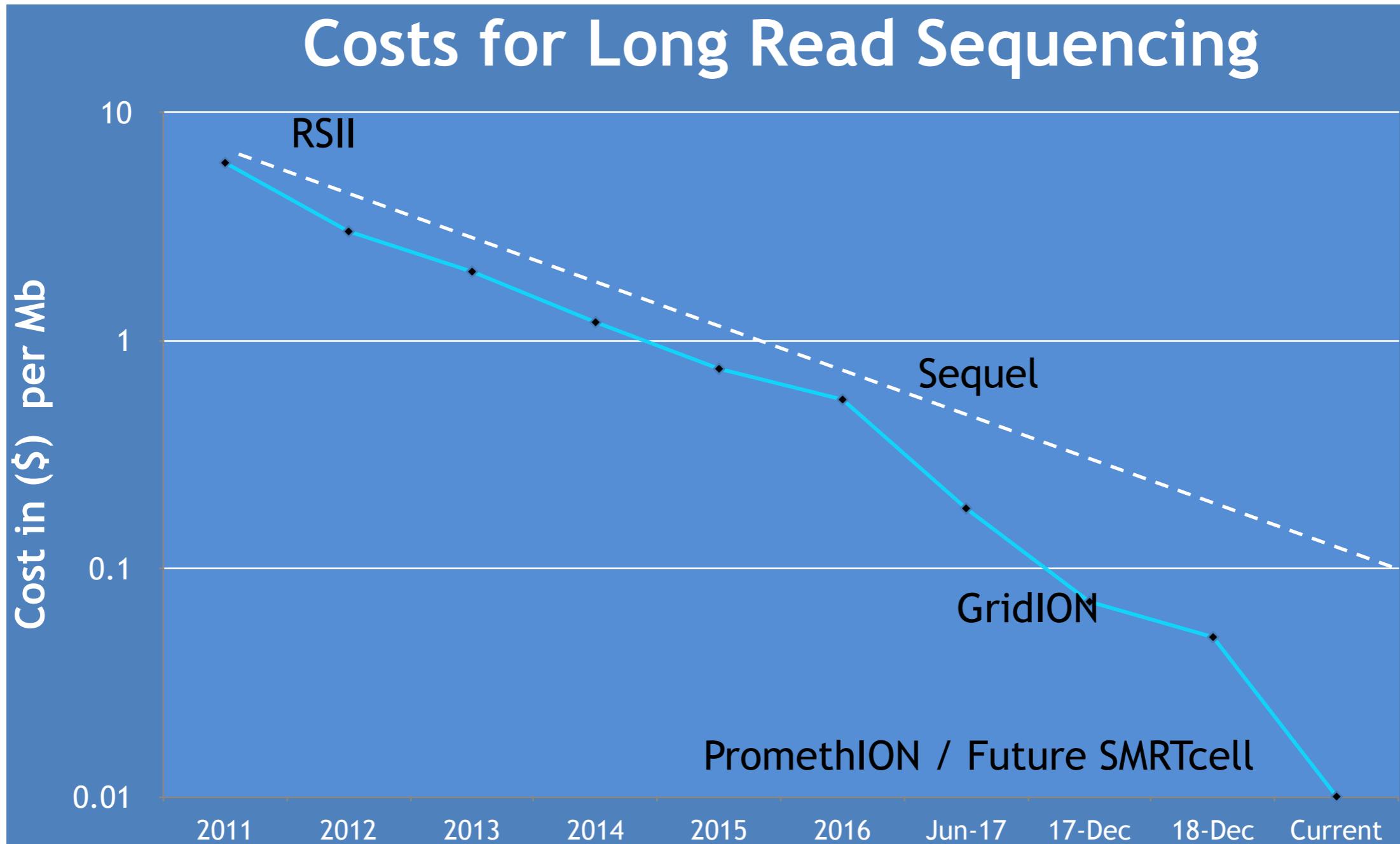
Two “flavors” of long read sequencing



PACIFIC
BIOSCIENCES®



Significant advances in long read sequencing over last 6 years





PACIFIC
BIOSCIENCES[®]

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

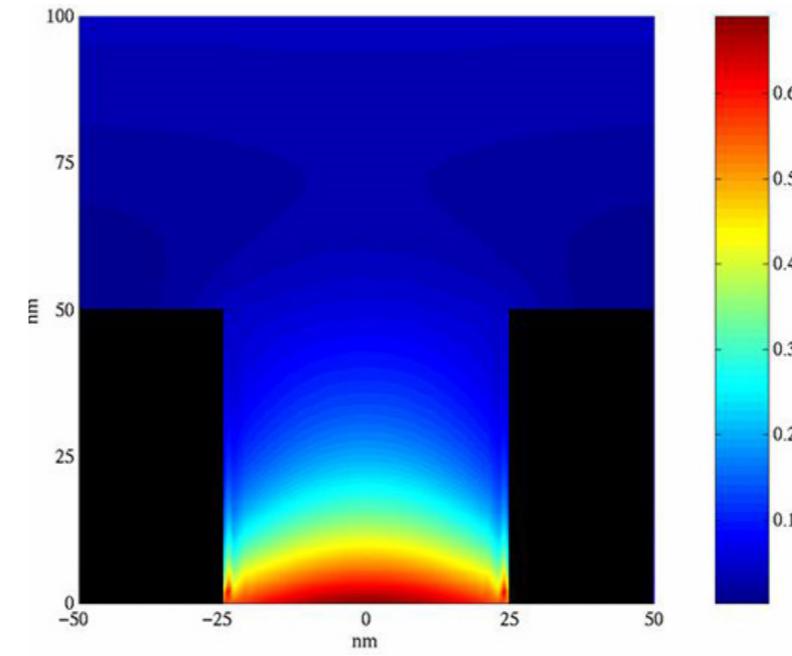
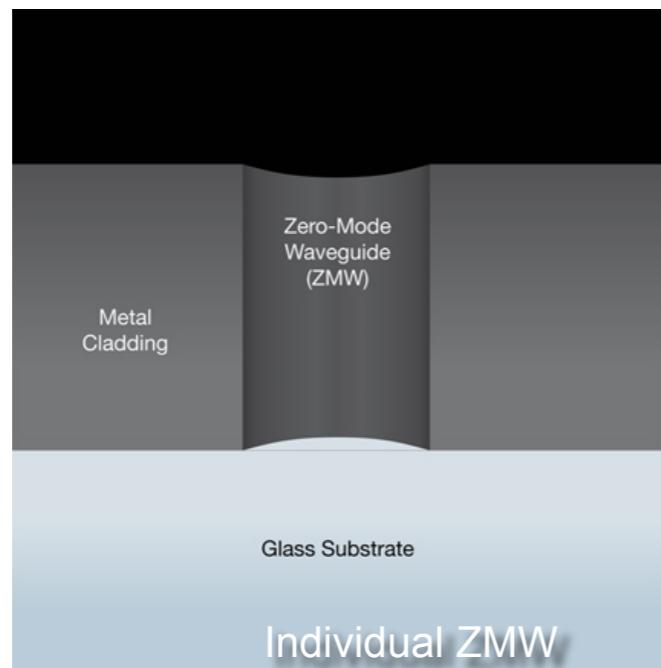


Zero-Mode Waveguides Are the Observation Windows

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

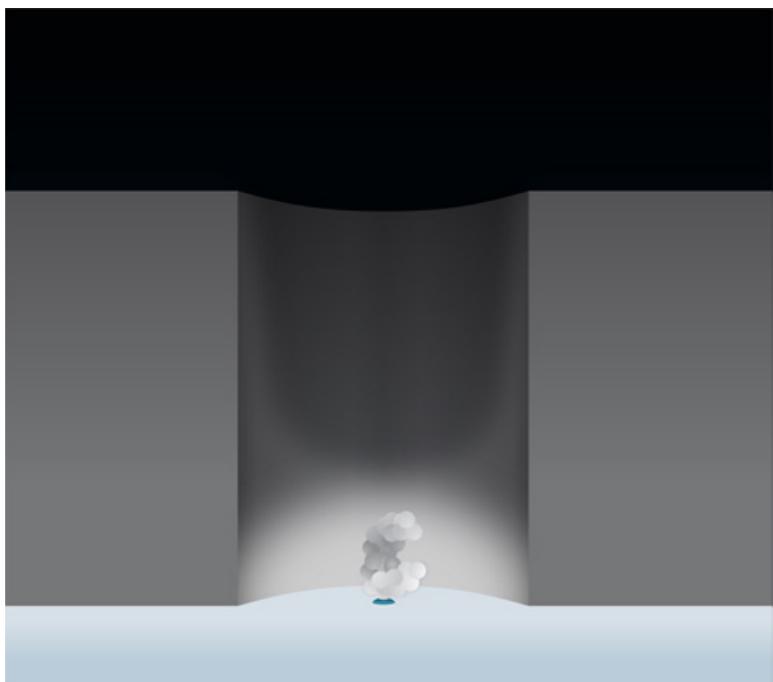
The ZMW provides a window for observing DNA polymerase as it performs sequencing by synthesis



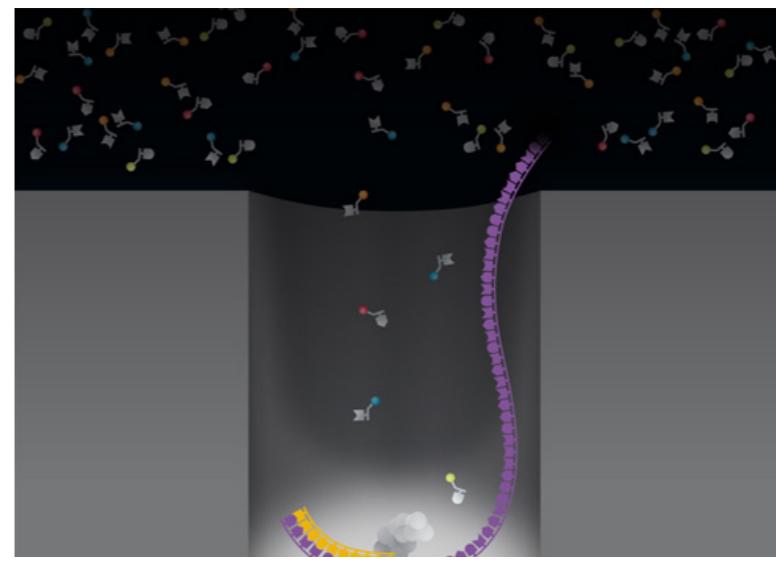
DNA Polymerase as a Sequencing Engine

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

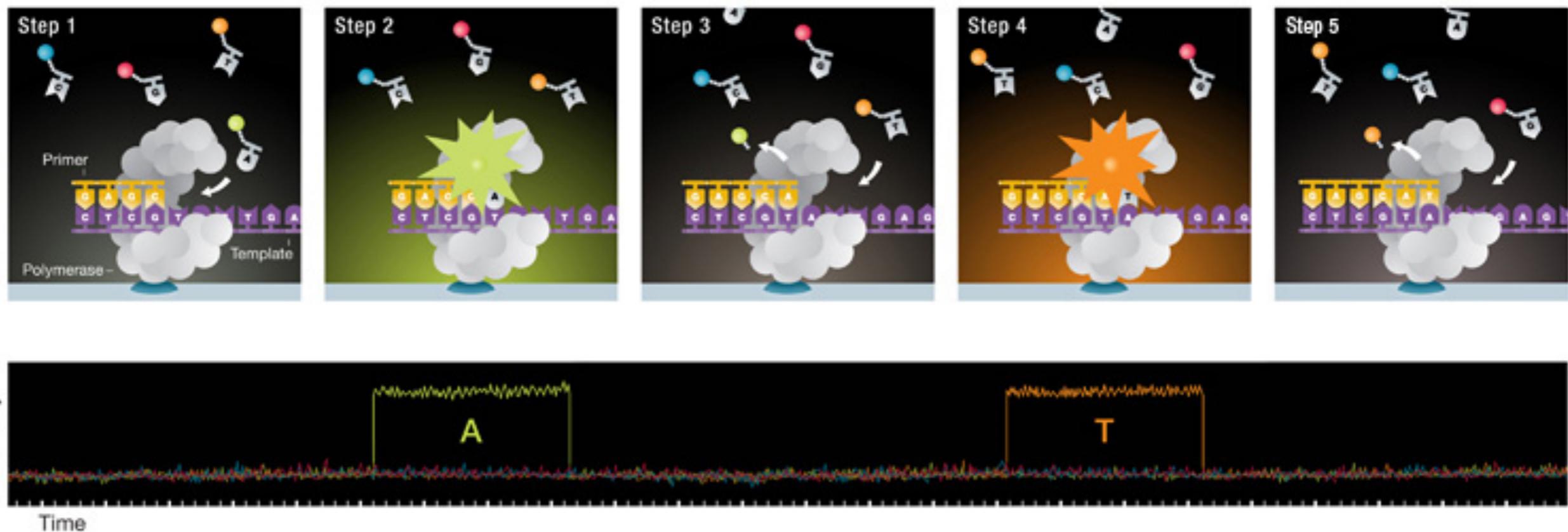


ZMW with DNA polymerase and phospholinked nucleotides

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

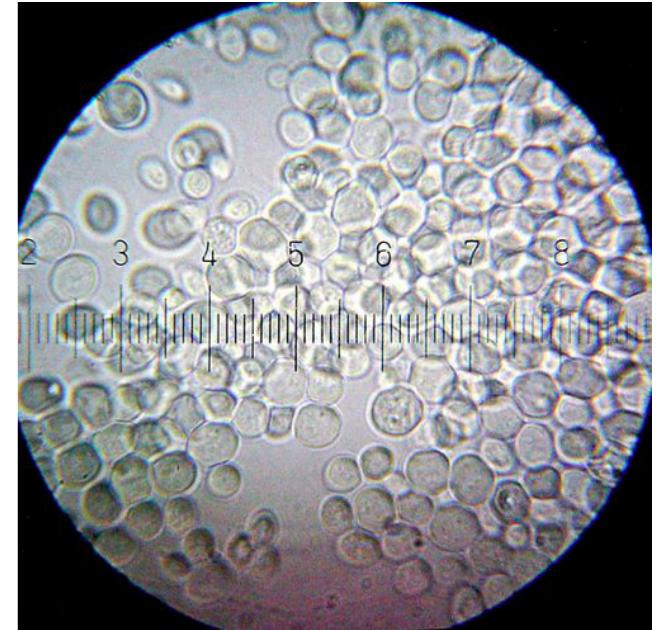
To generate consensus sequence from the data, an assembly process aligns the different fragments based on common sequences



**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.**

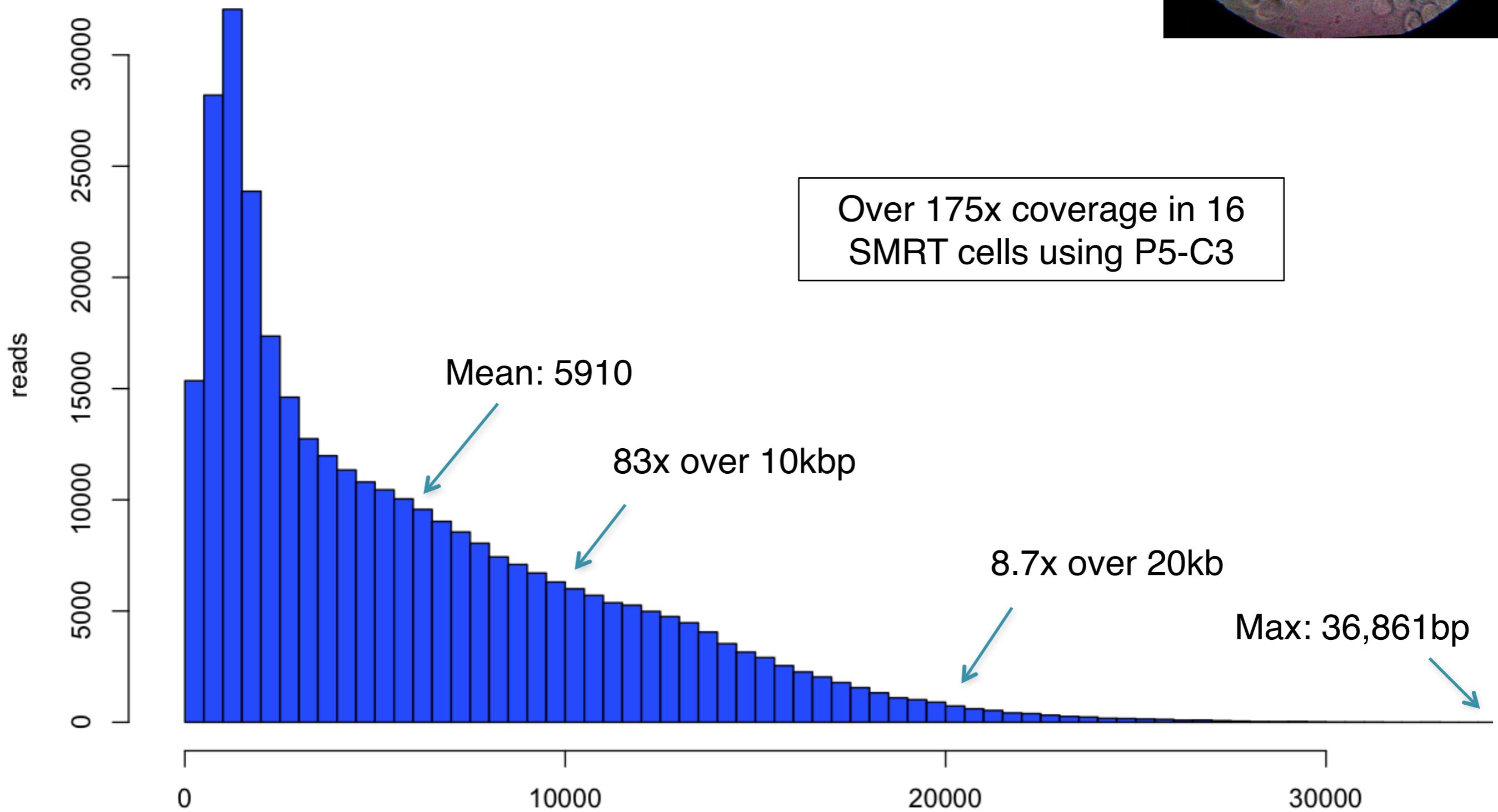
New York Times Nov. 9, 1919.

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



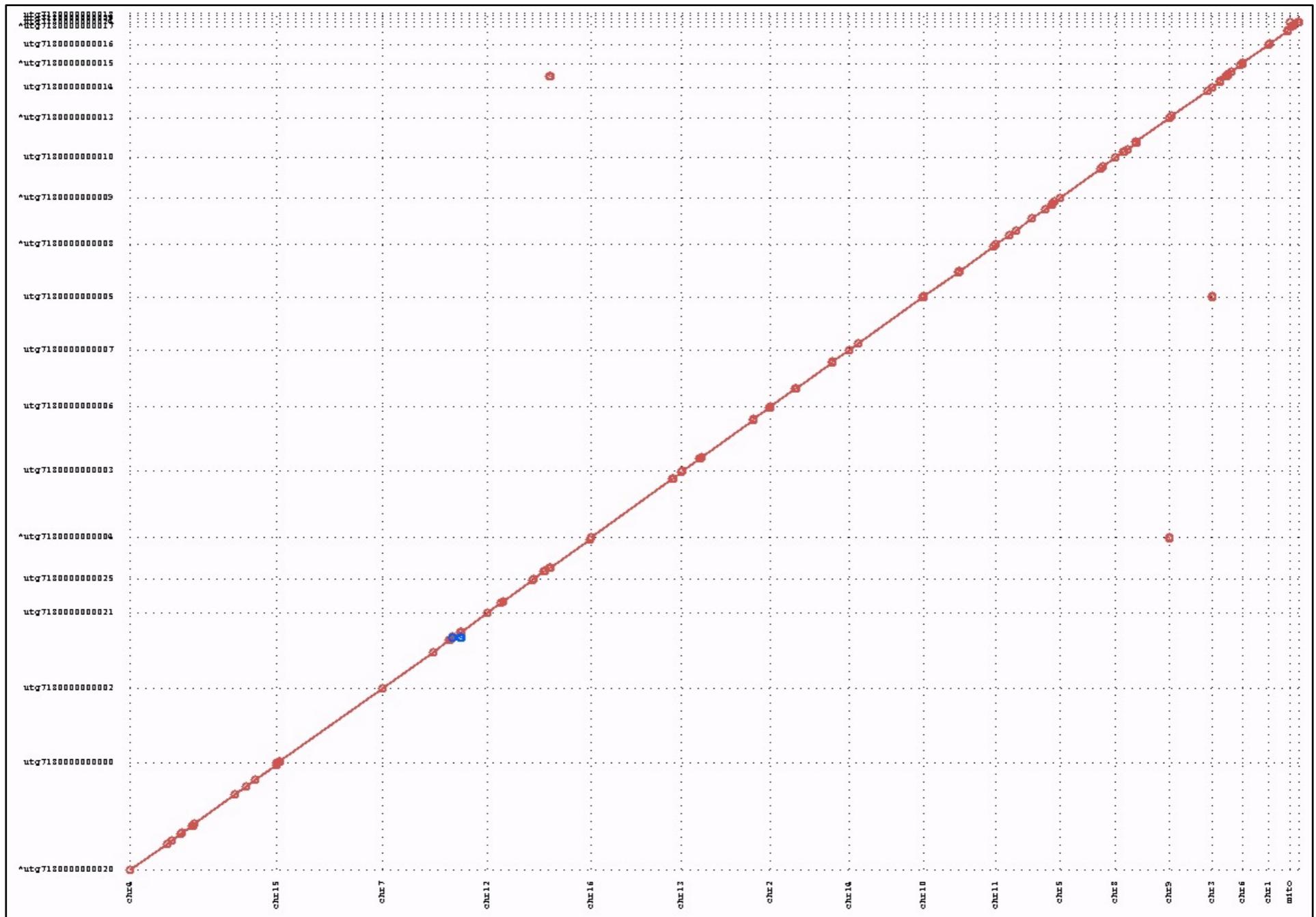
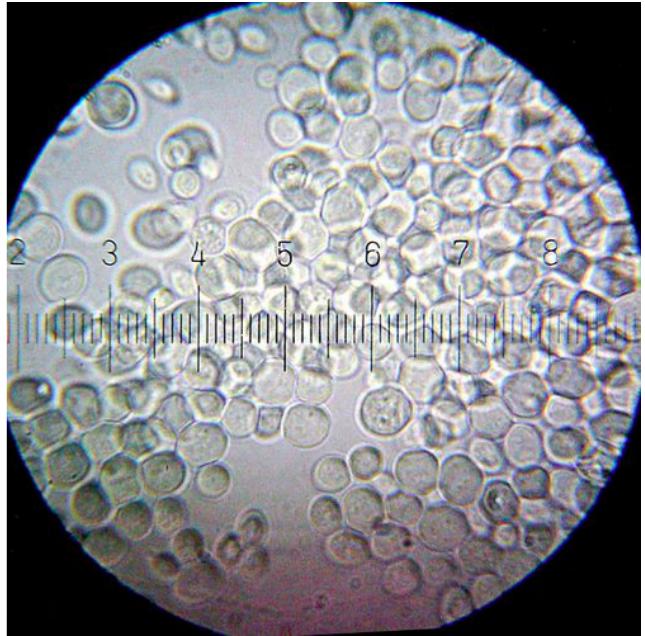
S. cerevisiae W303

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



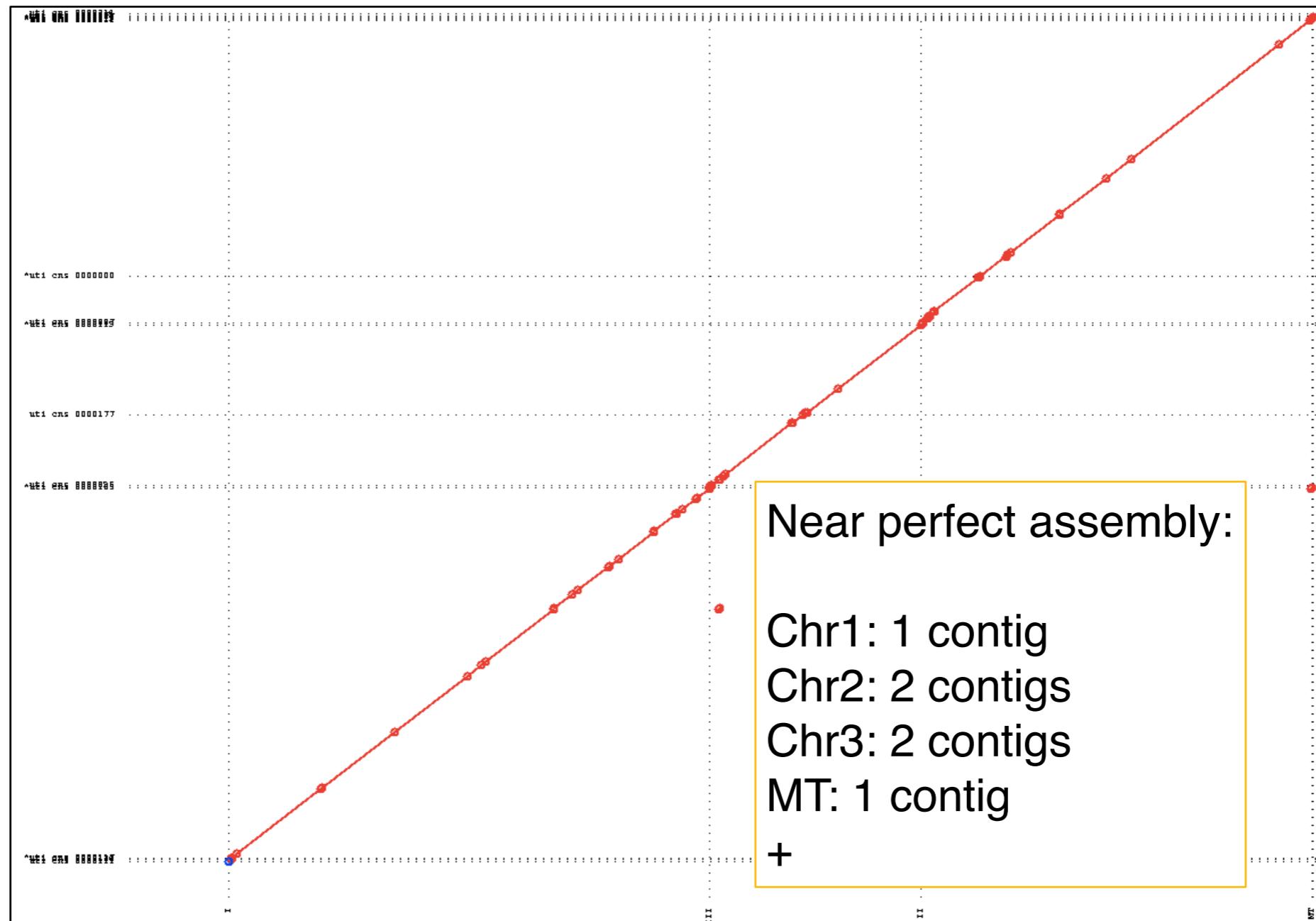
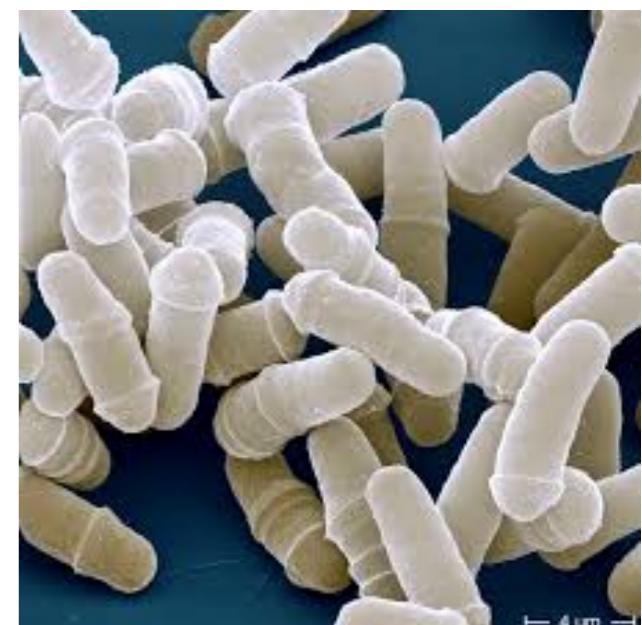
S. pombe dg21

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

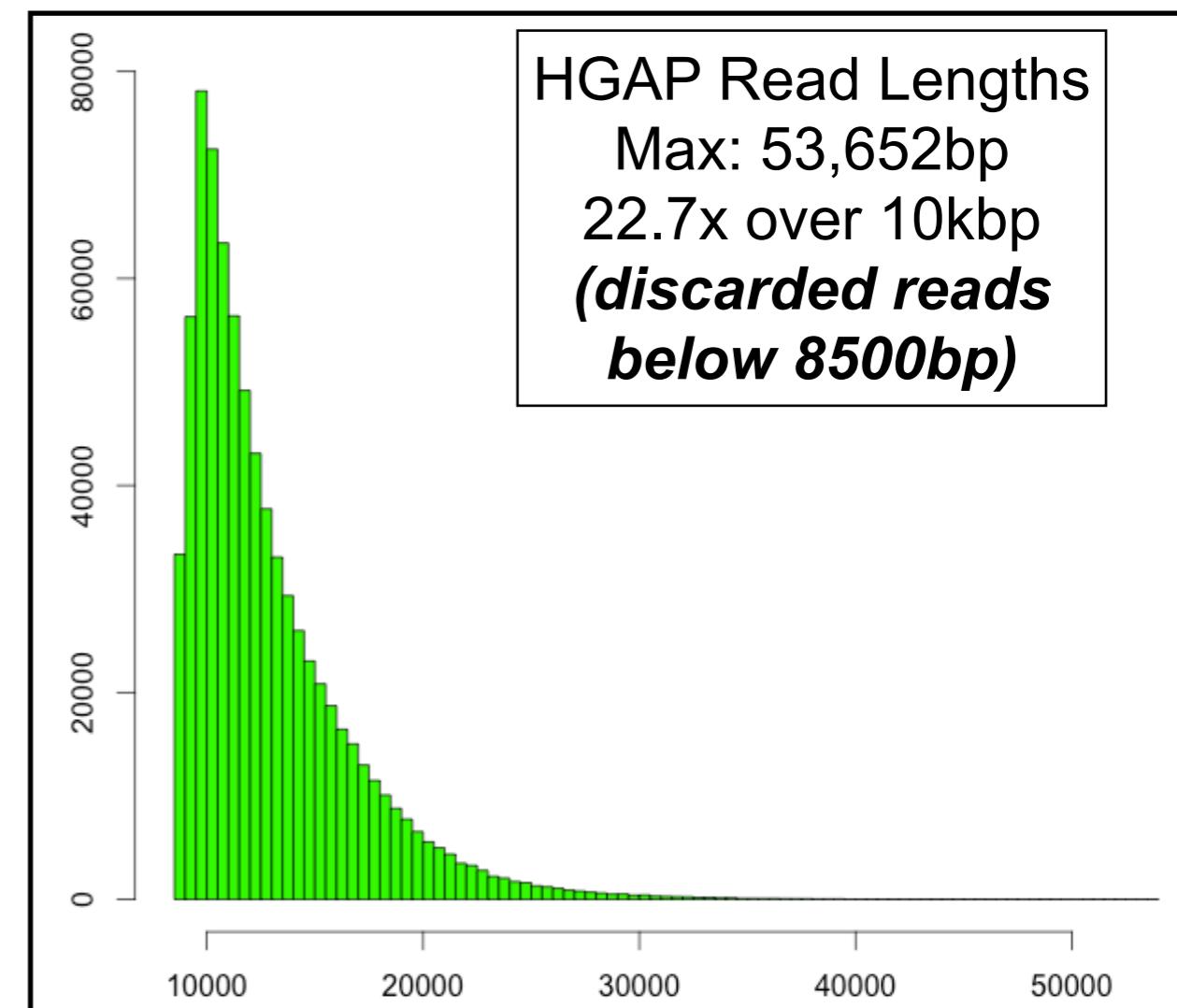


O. sativa pv Indica (IR64)



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

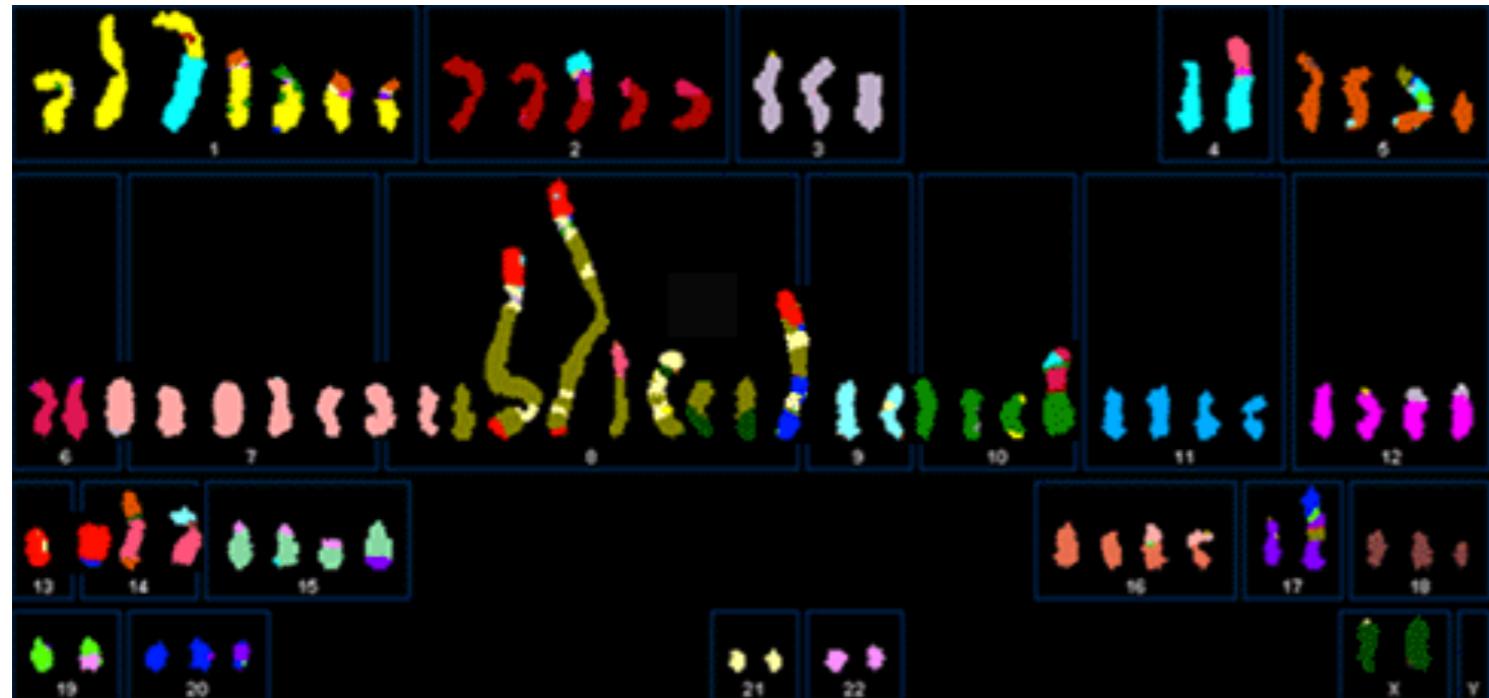


Structural Variations in SKBR3

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

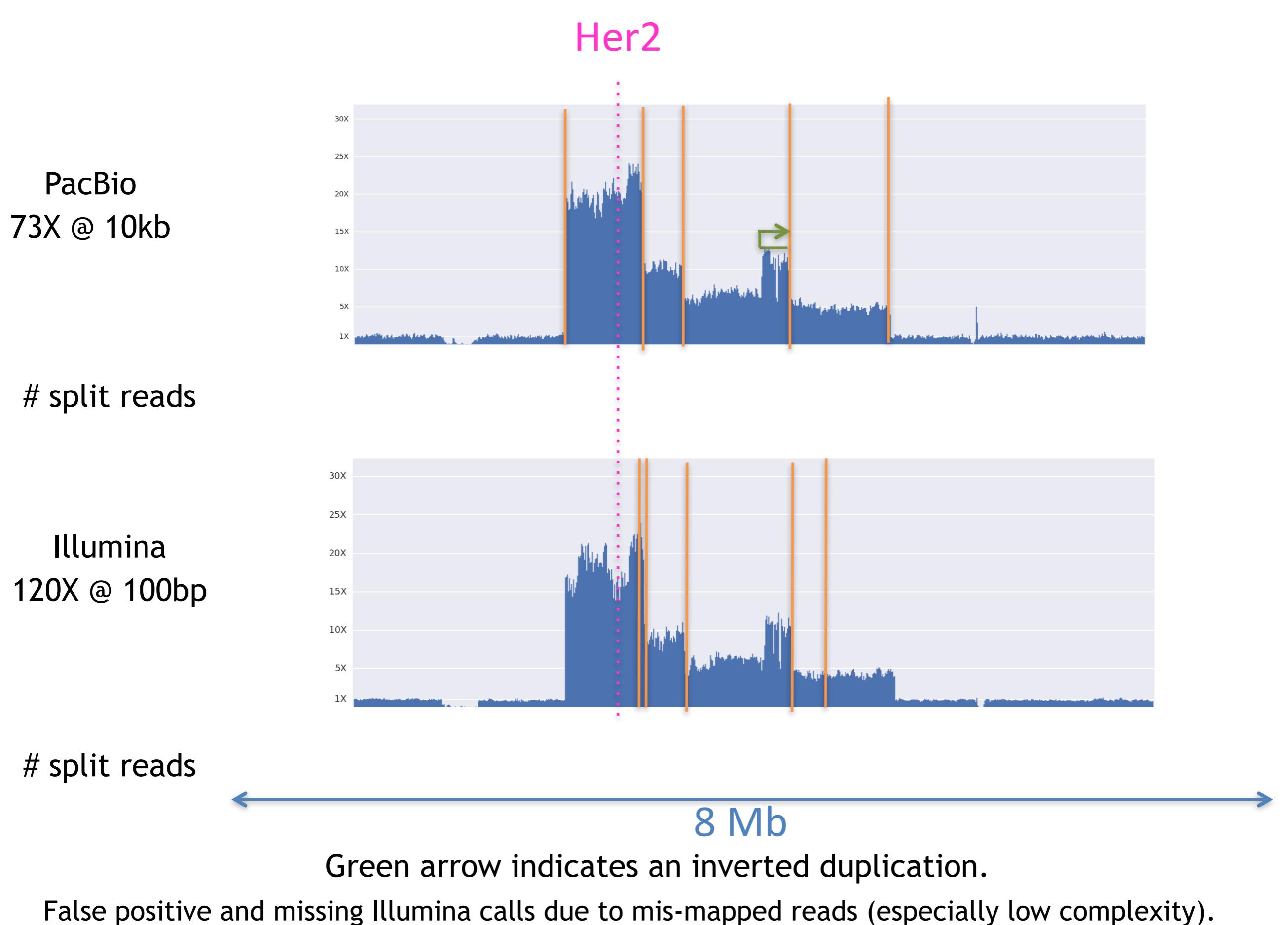
Mean: 2.8 kb

Max: 61 kb

N50: 3.3 kb

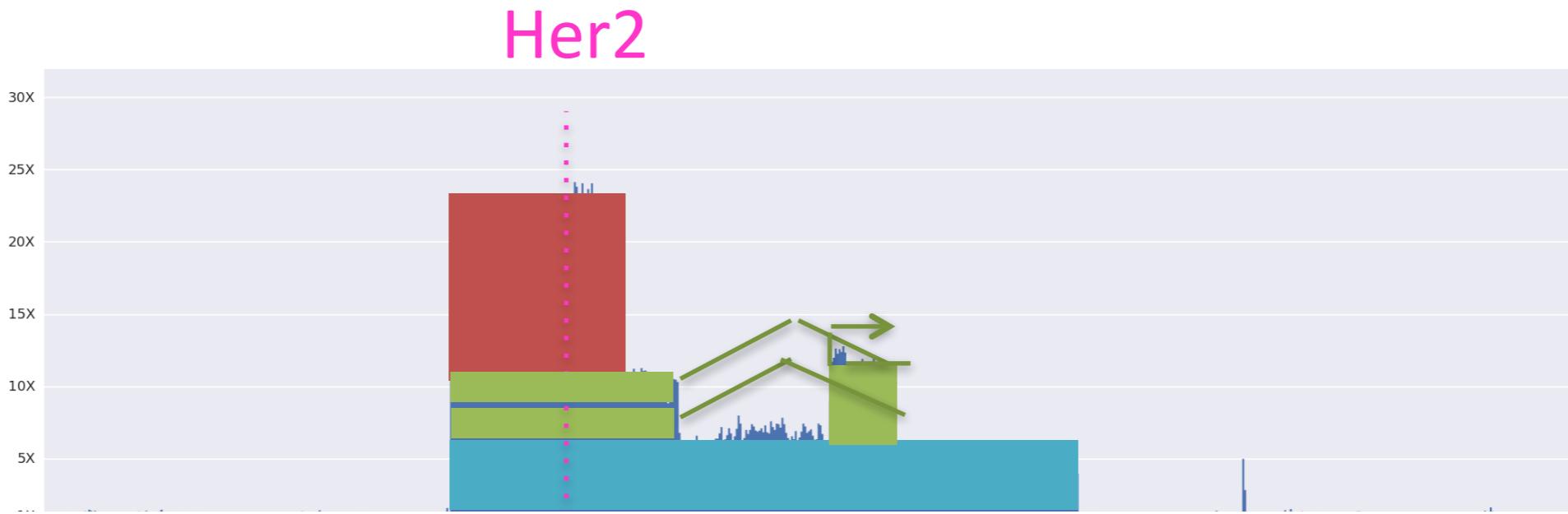
NG50: 1.9 kb





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

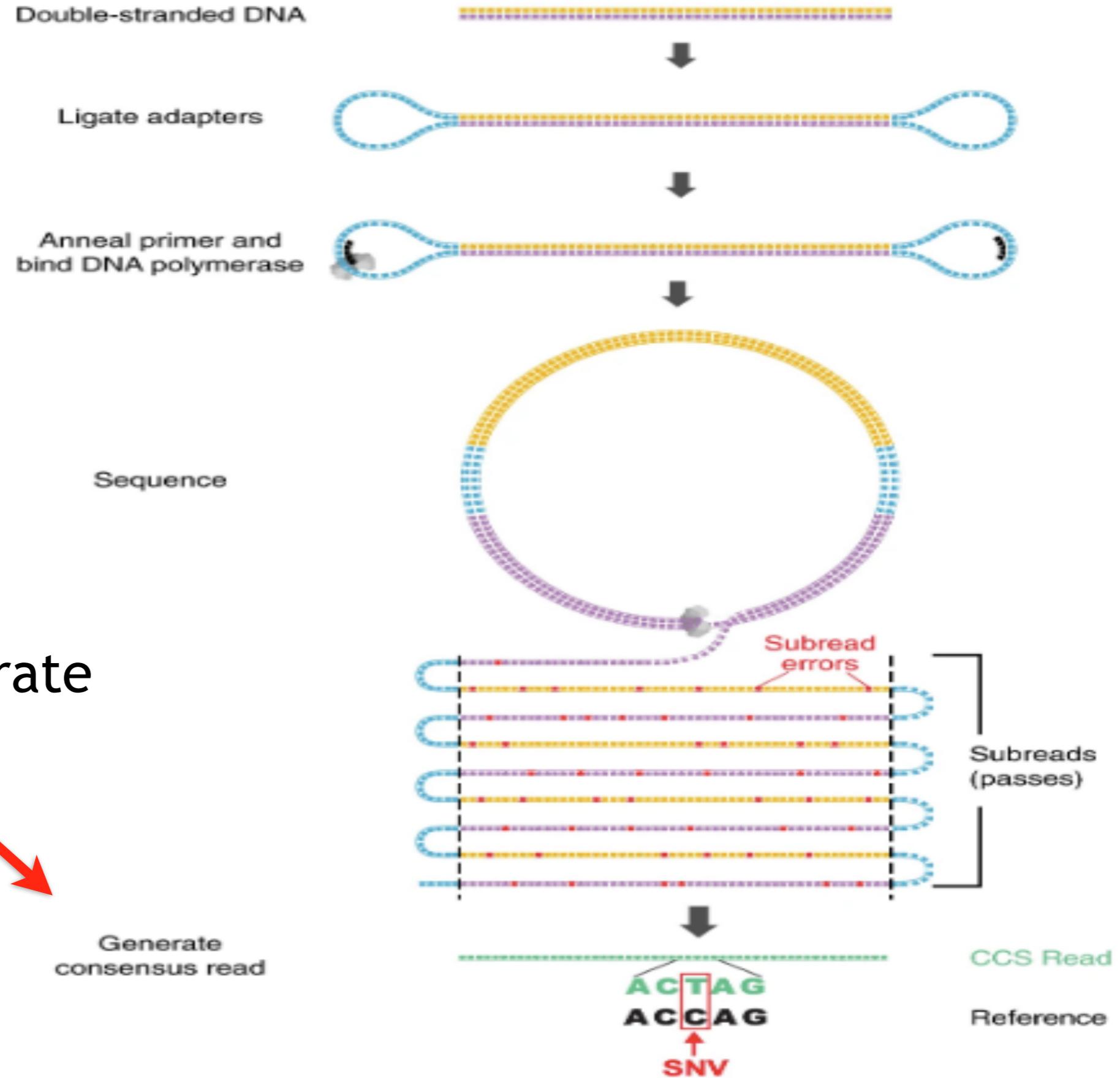
ATGCTCTCGATCGATGCTGCTAGCTAGCTACTAGCTATCCGATCCTACTGACTTACTATGGT



ATGCTCTCGATCGATGCTGCTAGCTAGCTACTAGCTATCCGATCCTACTGACTTACTATGCT

Enough coverage makes error drop out

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



From Wenger et al (2019) Nature Biotechnology



PromethION



24 independent flowcells

500bp/s sequencing speed

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

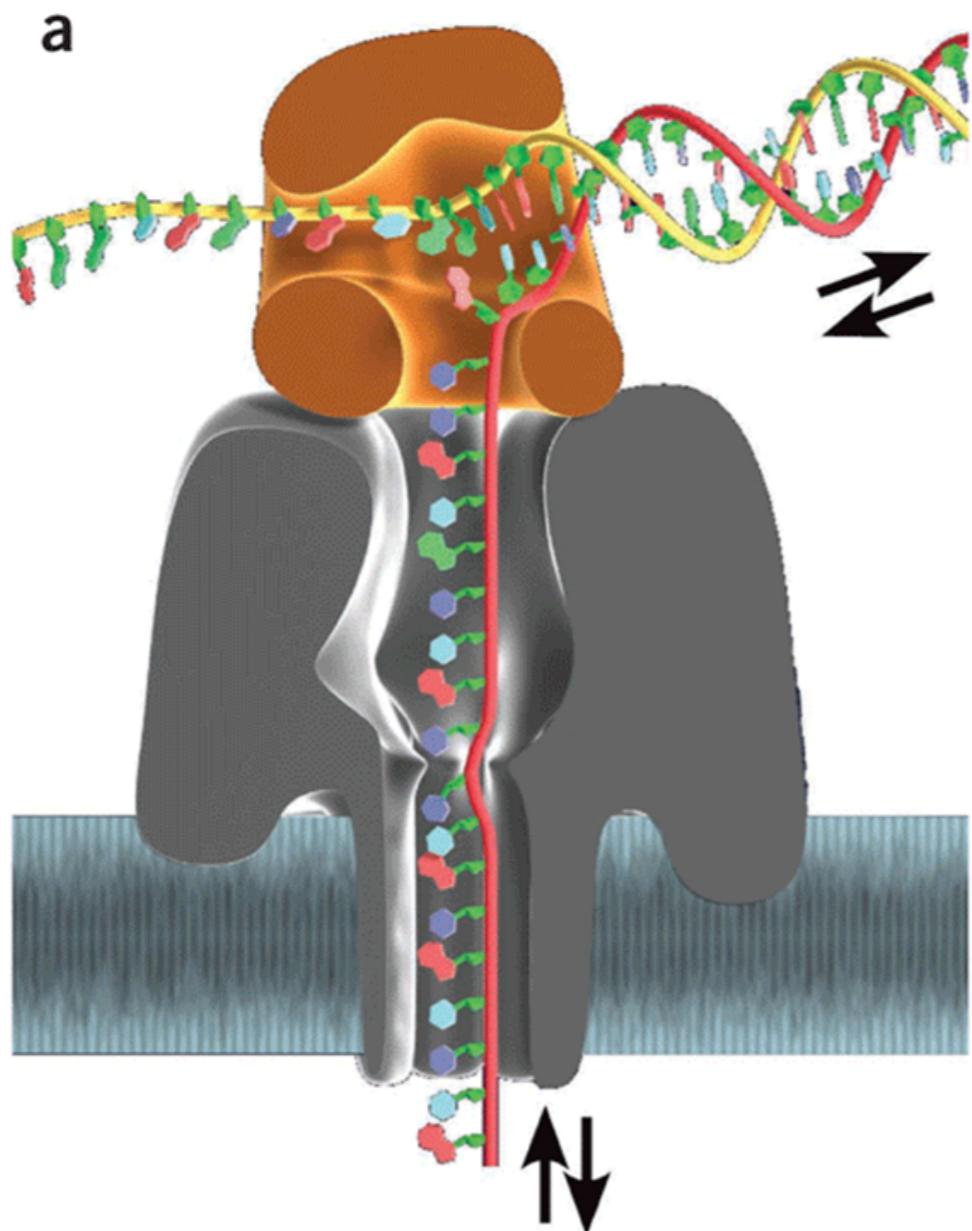
On site 1D basecalling

>140Gb in CSHL hands

>100M cDNA reads

Up to ~5 Tb fully loaded in one week

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

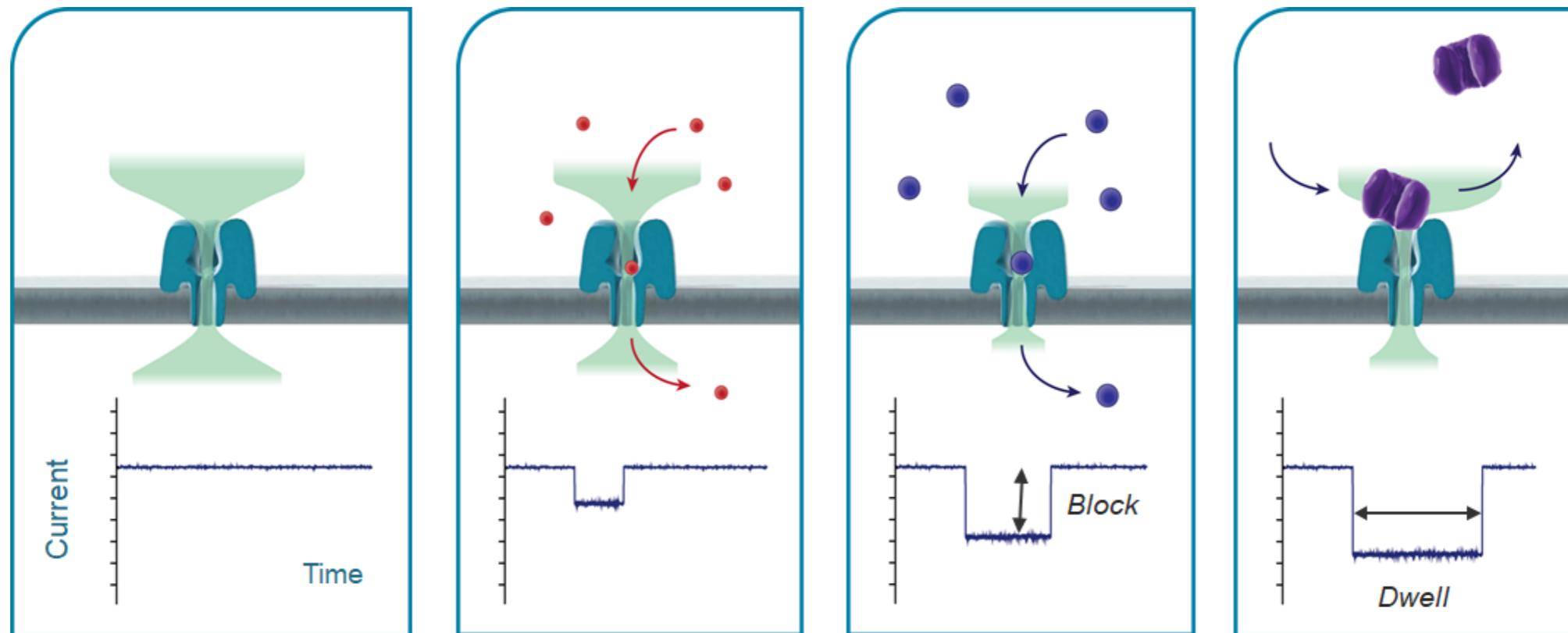
Nanopore Sensing Summary

Nanopore = ‘very small hole’

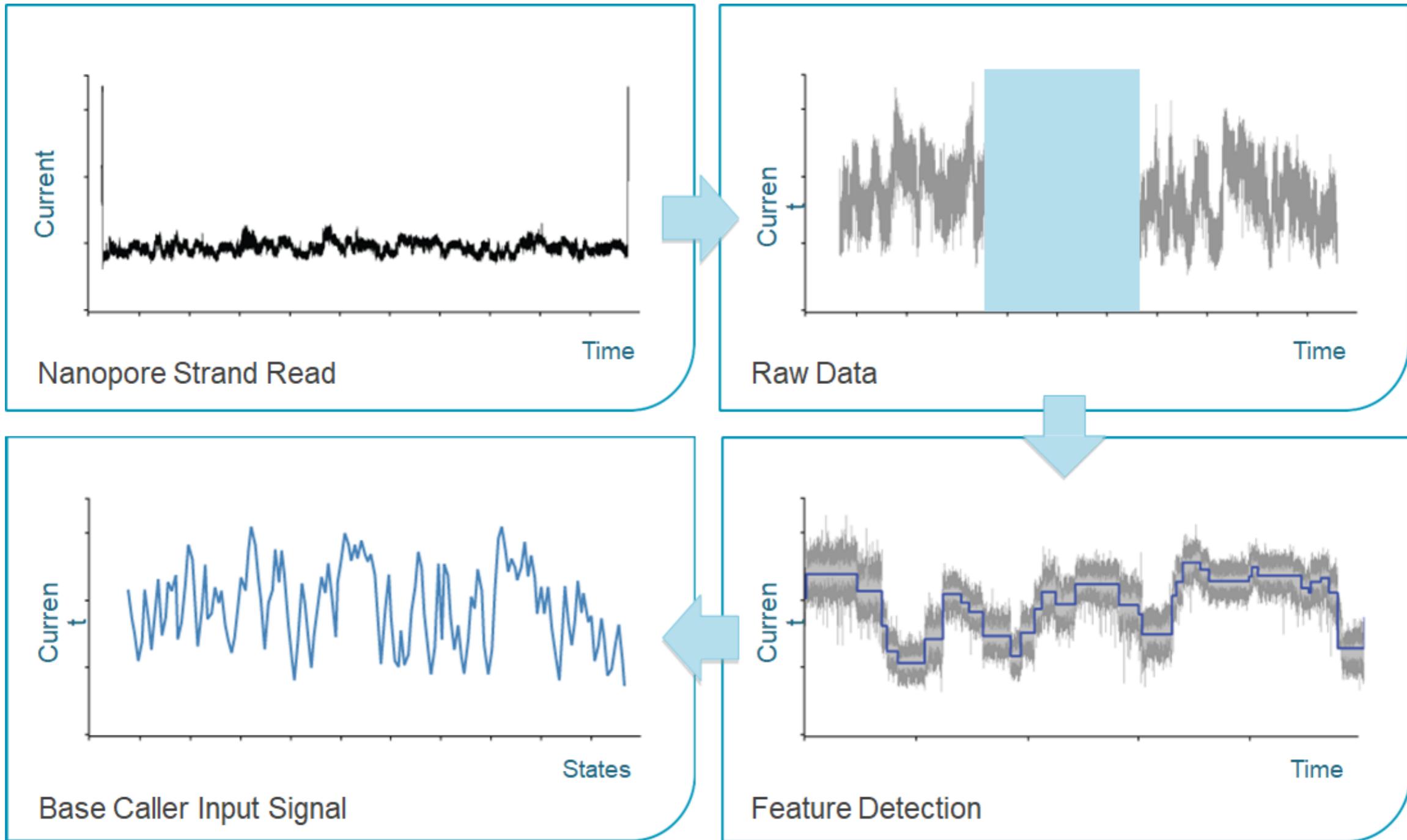
Ionic current flows through the pore Introduce analyte of interest into the pore

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

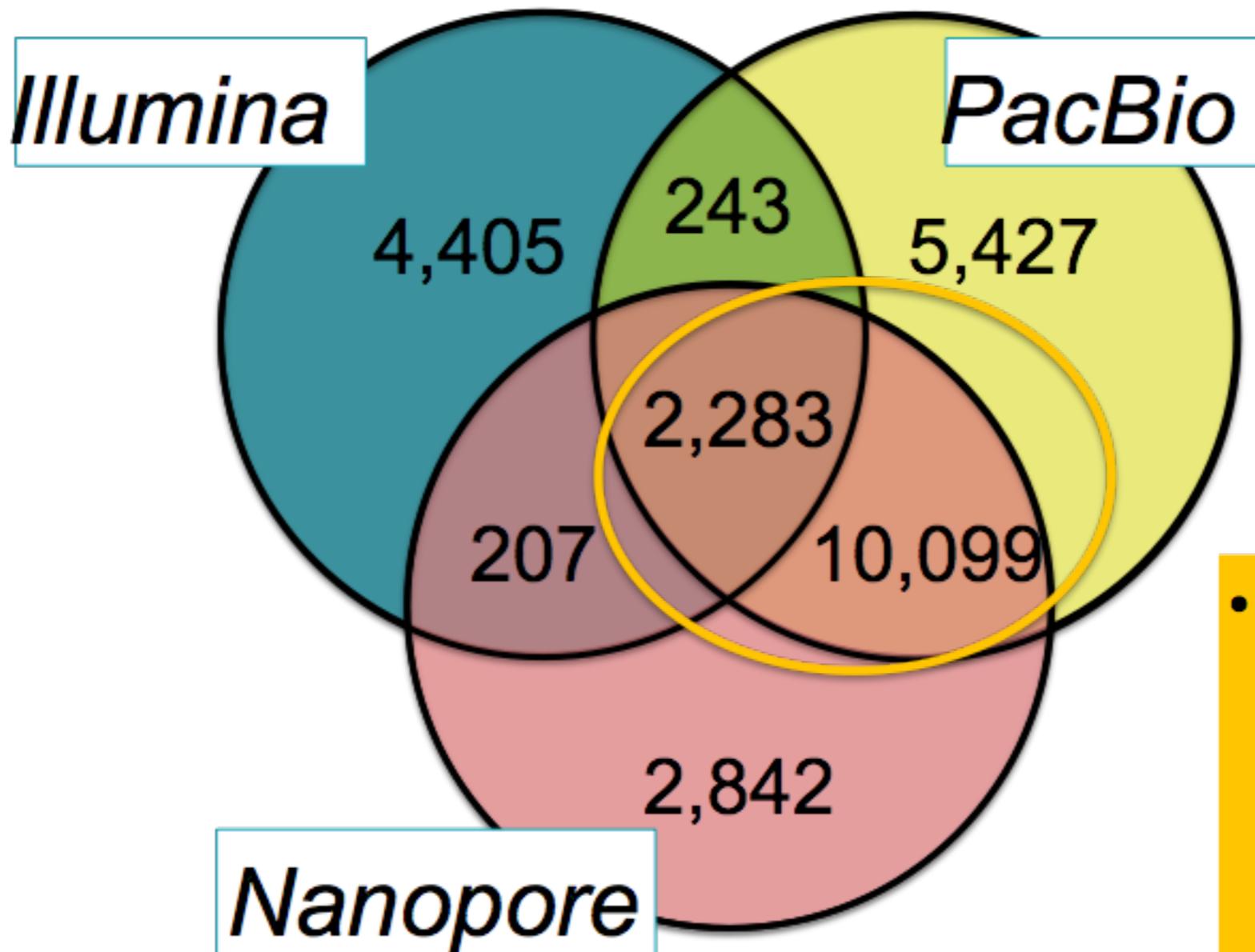
ATGCTCTCGATCGATGCTGCTAGCTAGCTAGCTAGCTTTTT CCGATCCTACTGACTTACTATGGT



ATGCTCTCGATCGATGCTGCTAGCTAGCTAGCTTTTT CCGATCCTACTGACTTACTATGCT

Enough coverage makes error (mostly) drop out

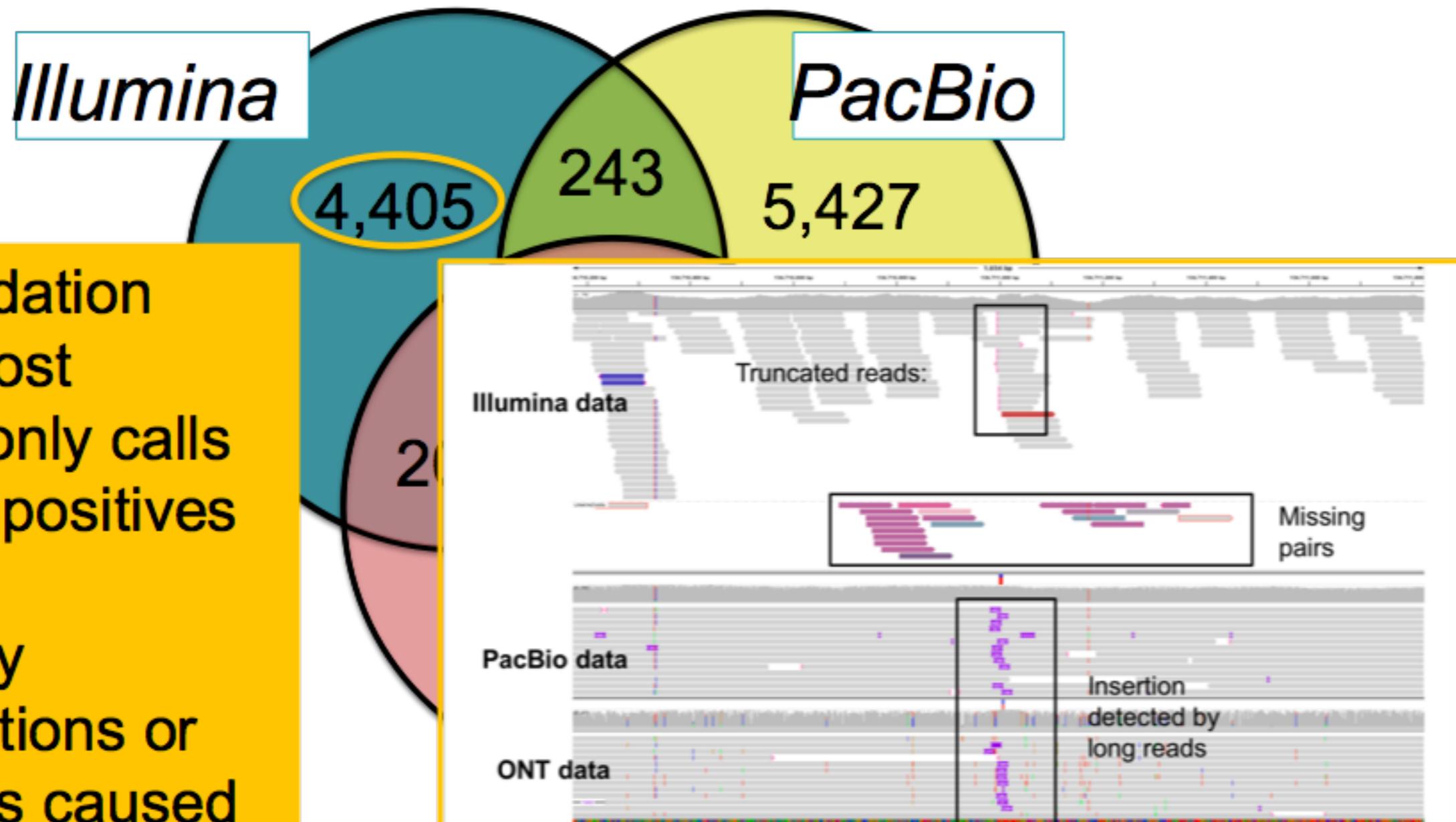
Structural Variant Comparison of SKBR3



(Hicks et al, 2006)

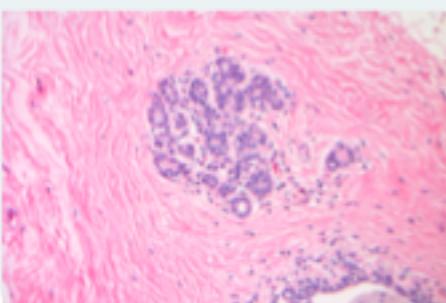
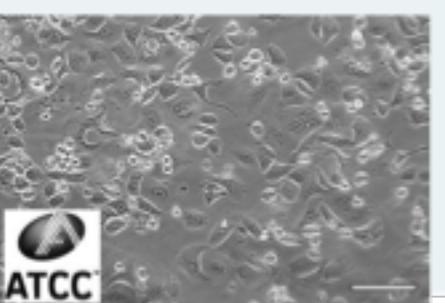
- Strong concordance between long read platforms
- Substantially more variants than detected by short reads

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

Multi-omics Long Read Analysis of Cancer

	Normal Breast Tissue	Normal Breast Organoid	Tumor Breast Organoid	SK-BR-3 Breast Cancer Cell Line
Oxford Nanopore WGS	Y	N	Y	Y
PacBio WGS	N	N	N	Y
ONT Methylation	Y	N	Y	Y
Illumina Methylation	Y	N	Y	Y
Illumina RNA-seq	N	Y	Y	Y
PacBio RNA-seq	N	N	N	Y
Pathology	NA	NA	ER+, PR+, Her2-	ER-, PR-, Her2+
Histology	Digital Atlas of Breast Pathology	David Spector, CSHL	David Spector, CSHL	ATCC
Image Source				

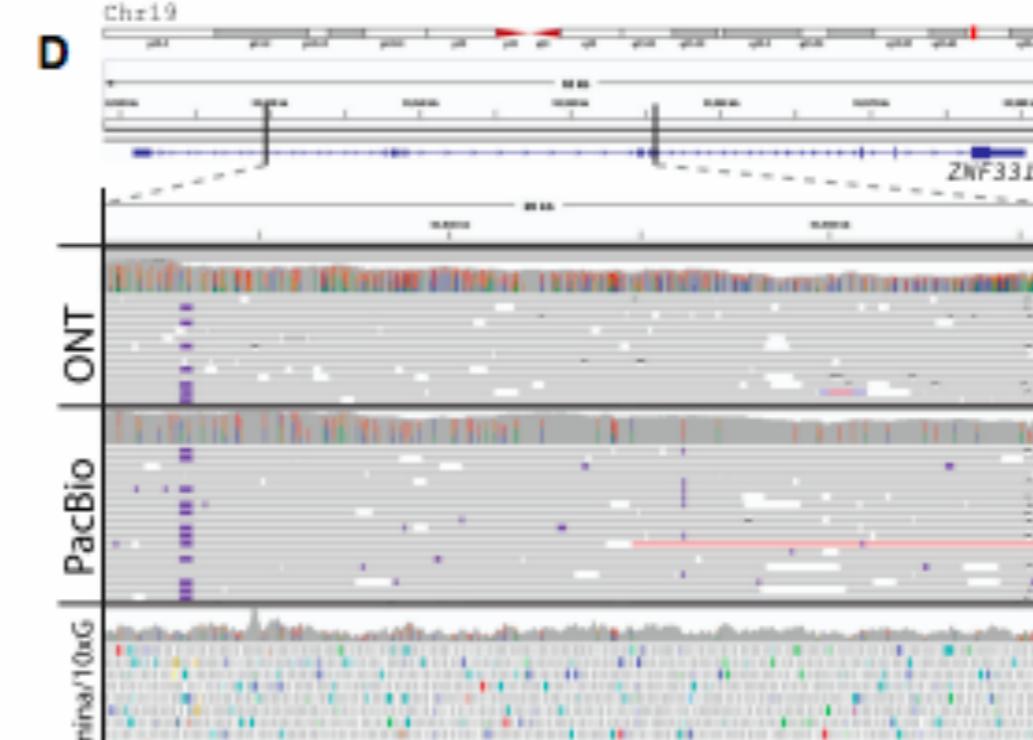
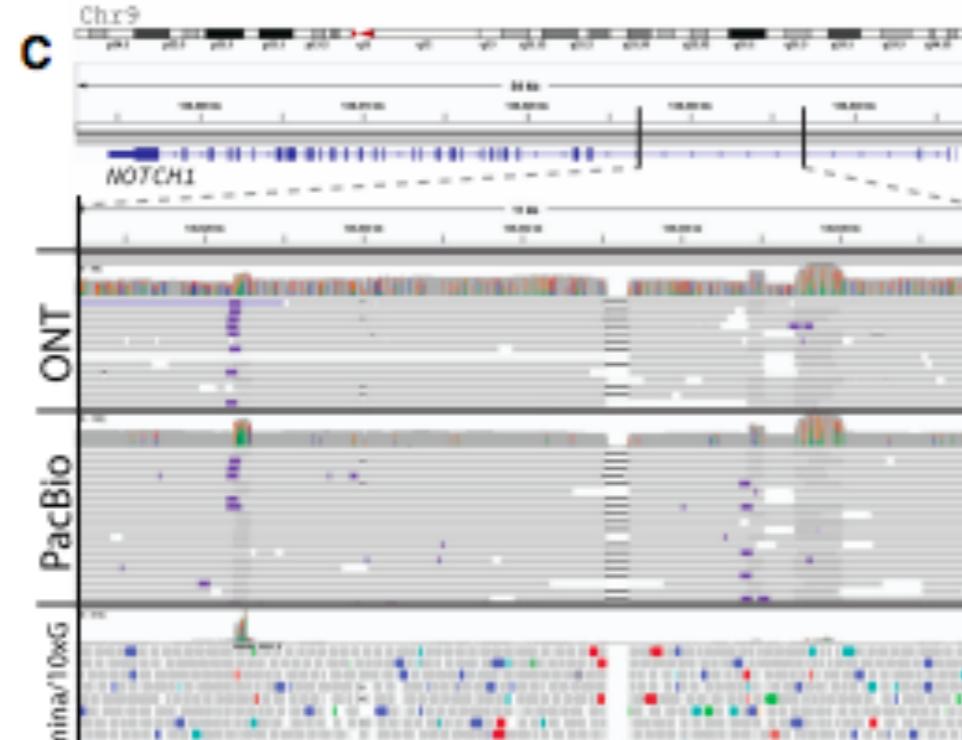
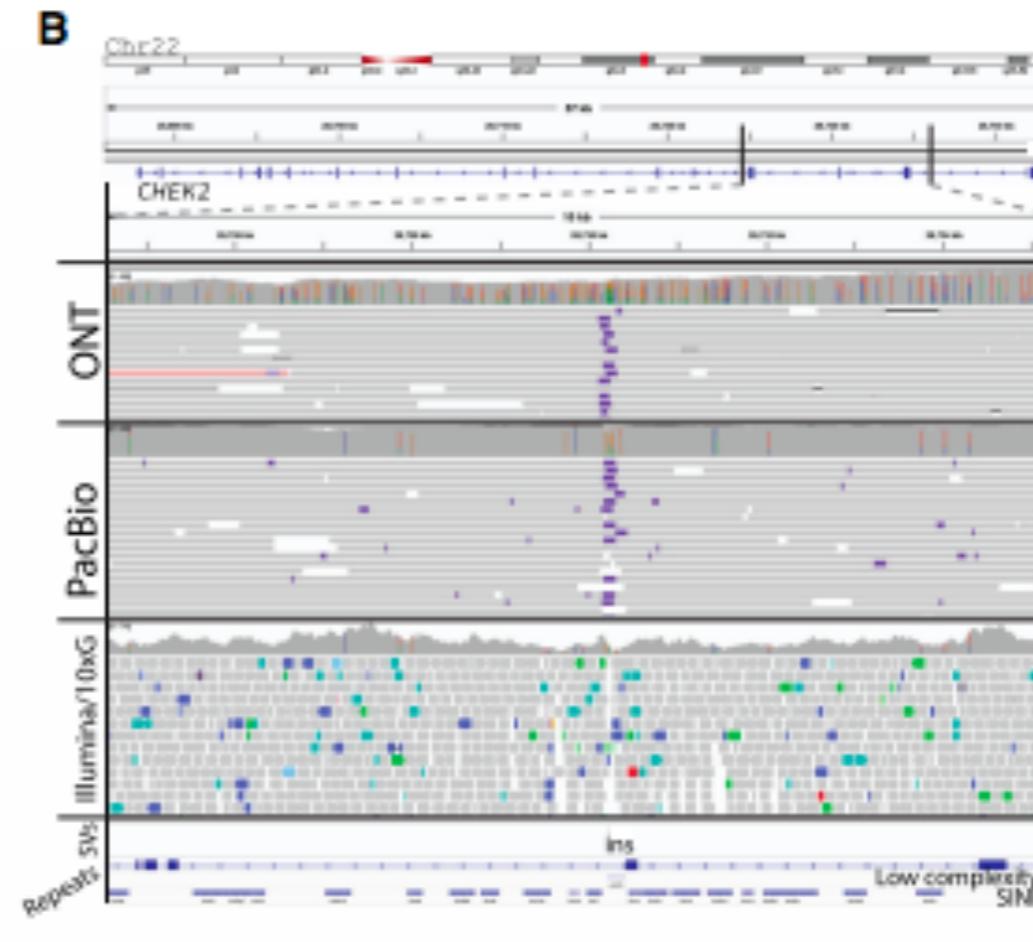
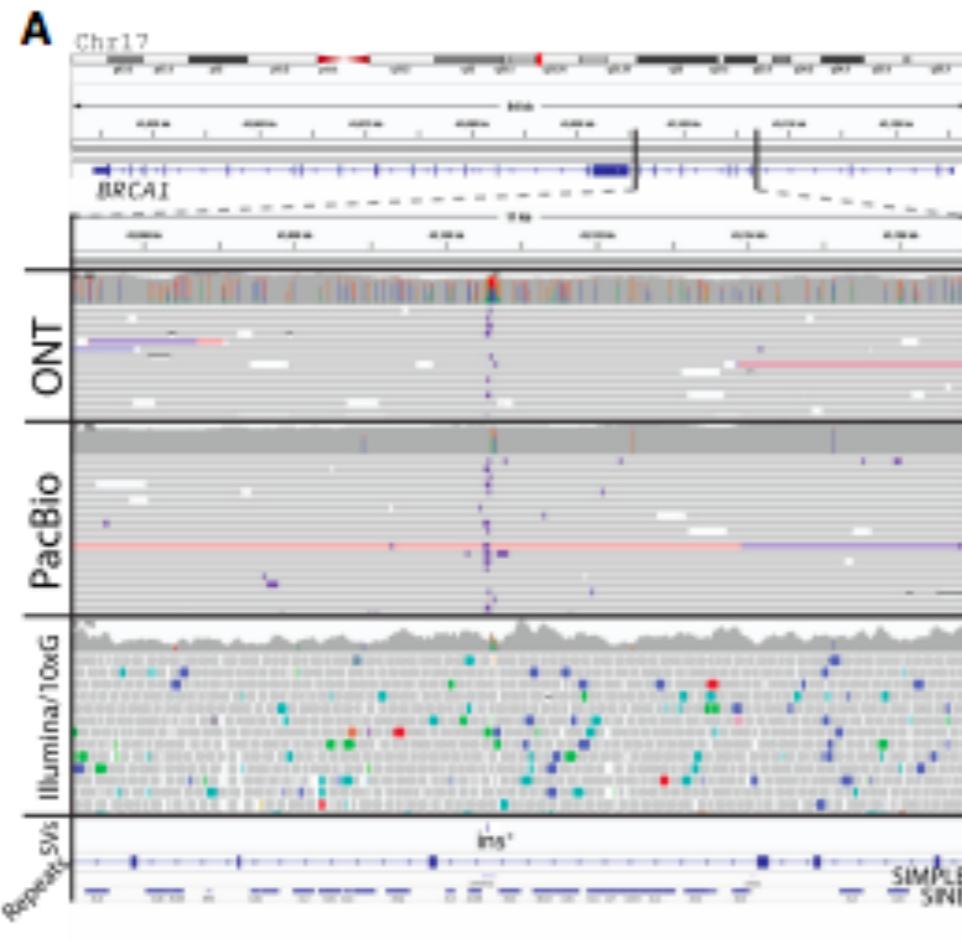
Preliminary Structural Variations Analysis



	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.



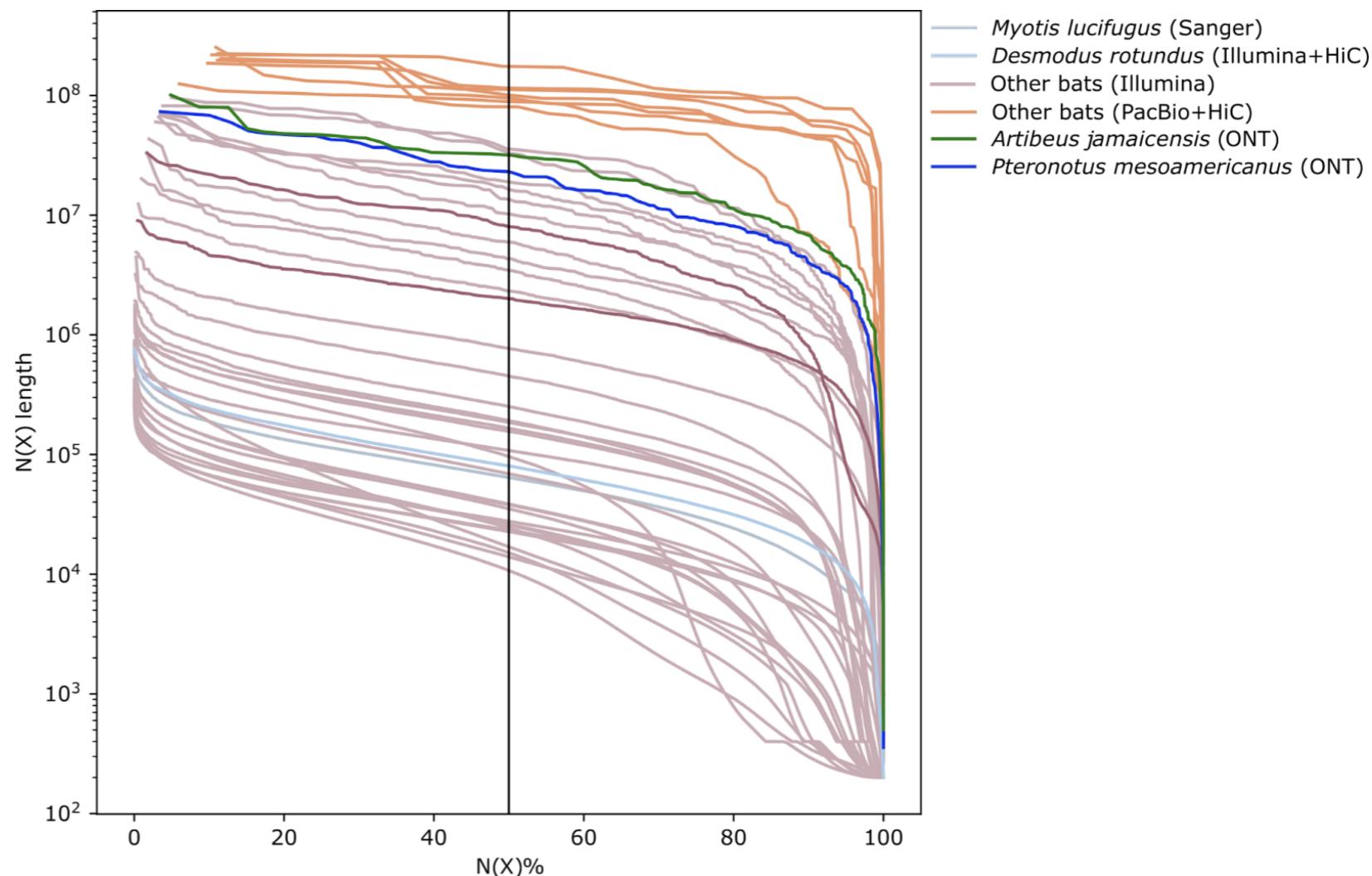
ONT long read sequencing of neo-tropical bats



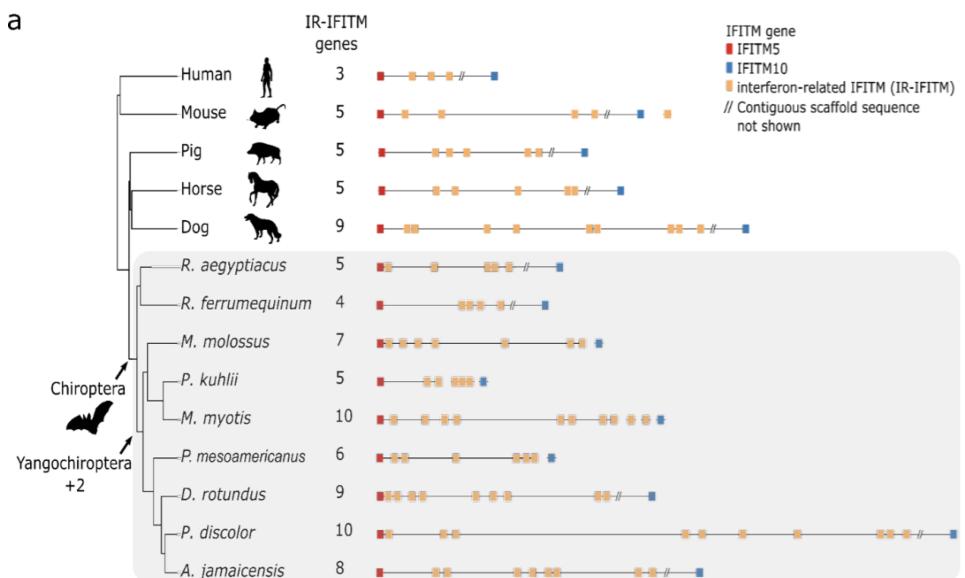
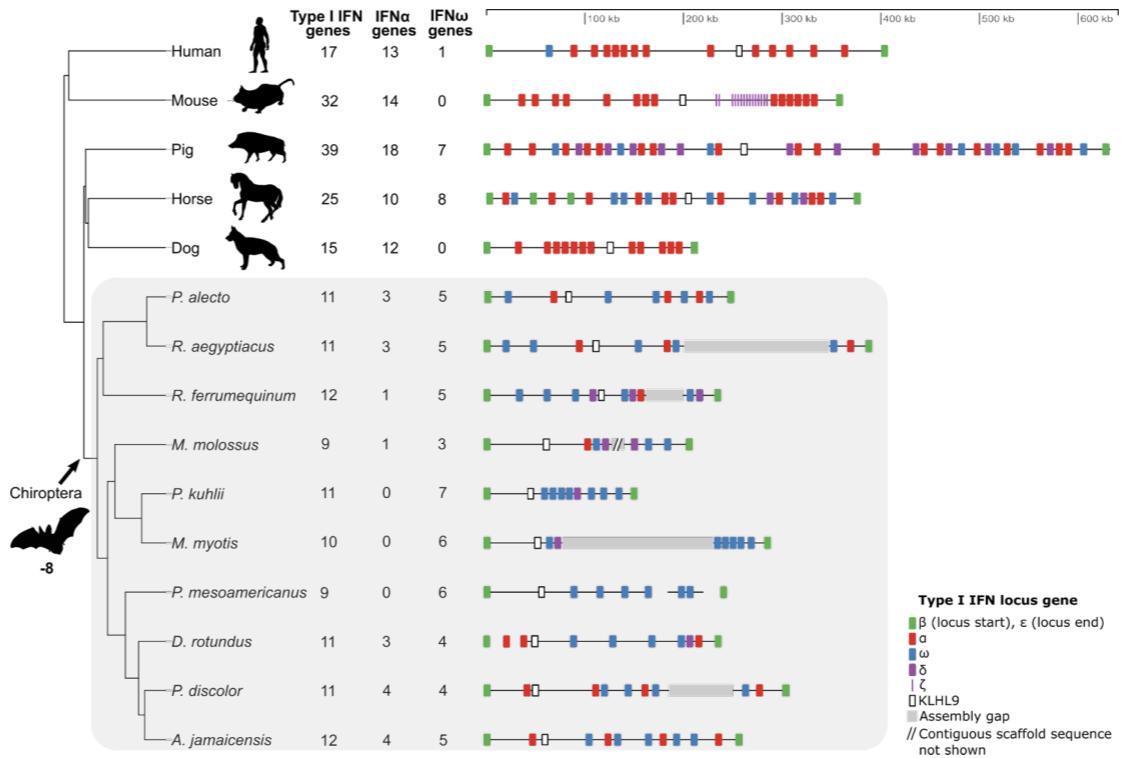
We have sequenced the genomes of 2 bat species (*Artibeus jamaicensis* and *Pteronotus mesoamericanus*) using Oxford Nanopore PromethION long reads (with Illumina short reads for error correction) to fully assess the spectrum of genomic variations which may contribute to longevity and cancer suppression.

*Collaboration with Bat1K and AMNH (Nancy Simmons and Sara Oppenheim)

Highly contiguous assemblies with contig N50 of 28-29Mb and consensus quality of >99.99%

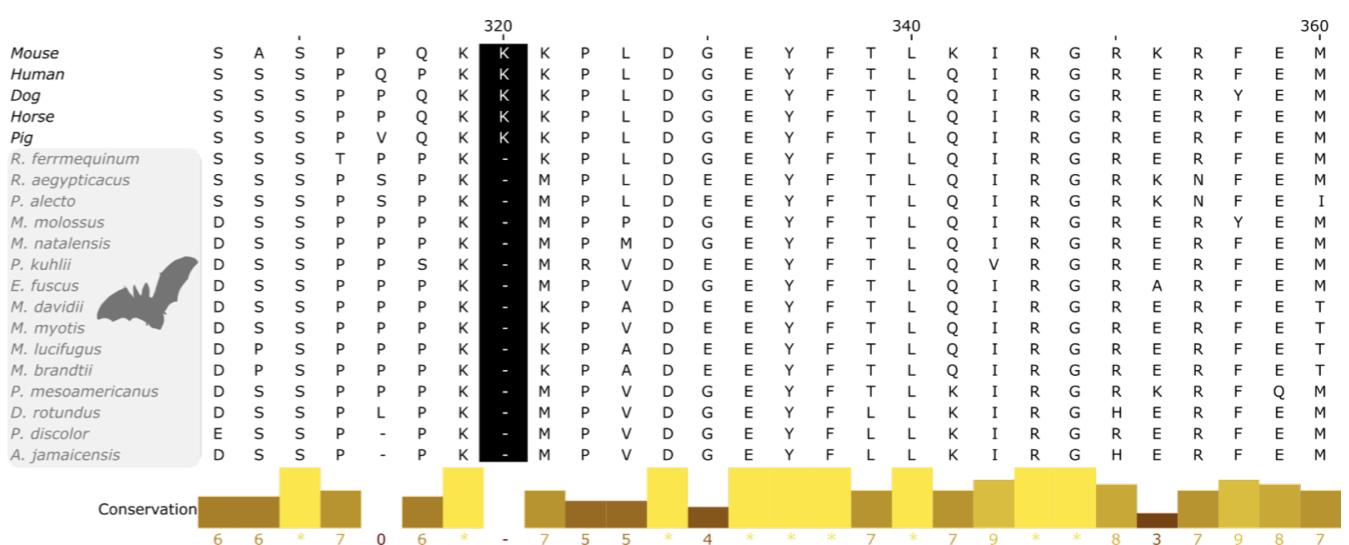


Contraction of type I IFN locus in bats compared to other mammals



Expansion of IFITM gene family and positive selection associated with antiviral immune response in bats

Bat-specific deletion in TP53



Living Fossils Oxford Nanopore Sequencing

Node	Gymnosperm species	1C (pg)	1C (Gbp)	Sequencing strategy * = this project
1	<i>Ginkgo biloba</i> ("living fossil")	11.75	11.5	NGS [1]
1	<i>Cycas revoluta</i>	13.70	13.4	NGS [2]
2	<i>Pinus taeda</i>	22.10	21.6	NGS [3]
2	<i>Picea abies</i> ("living fossil")	20.01	19.6	NGS [4]
3	<i>Juniperus communis</i>	9.84	9.6	Oxford Nanopore*
3	<i>Thuja plicata</i>	12.84	12.6	NGS [2]
3	<i>Metasequoia glyptostroboides</i> ("living fossil")	11.04	10.8	Oxford Nanopore*
4	<i>Wollemia nobilis</i> ("living fossil")	11.04	10.8	Oxford Nanopore*
4	<i>Agathis vitiensis</i>	15.80	15.5	Oxford Nanopore*
5	<i>Welwitschia mirabilis</i>	7.20	7.0	NGS [2]
5	<i>Gnetum ulna</i>	2.25	2.2	Oxford Nanopore*

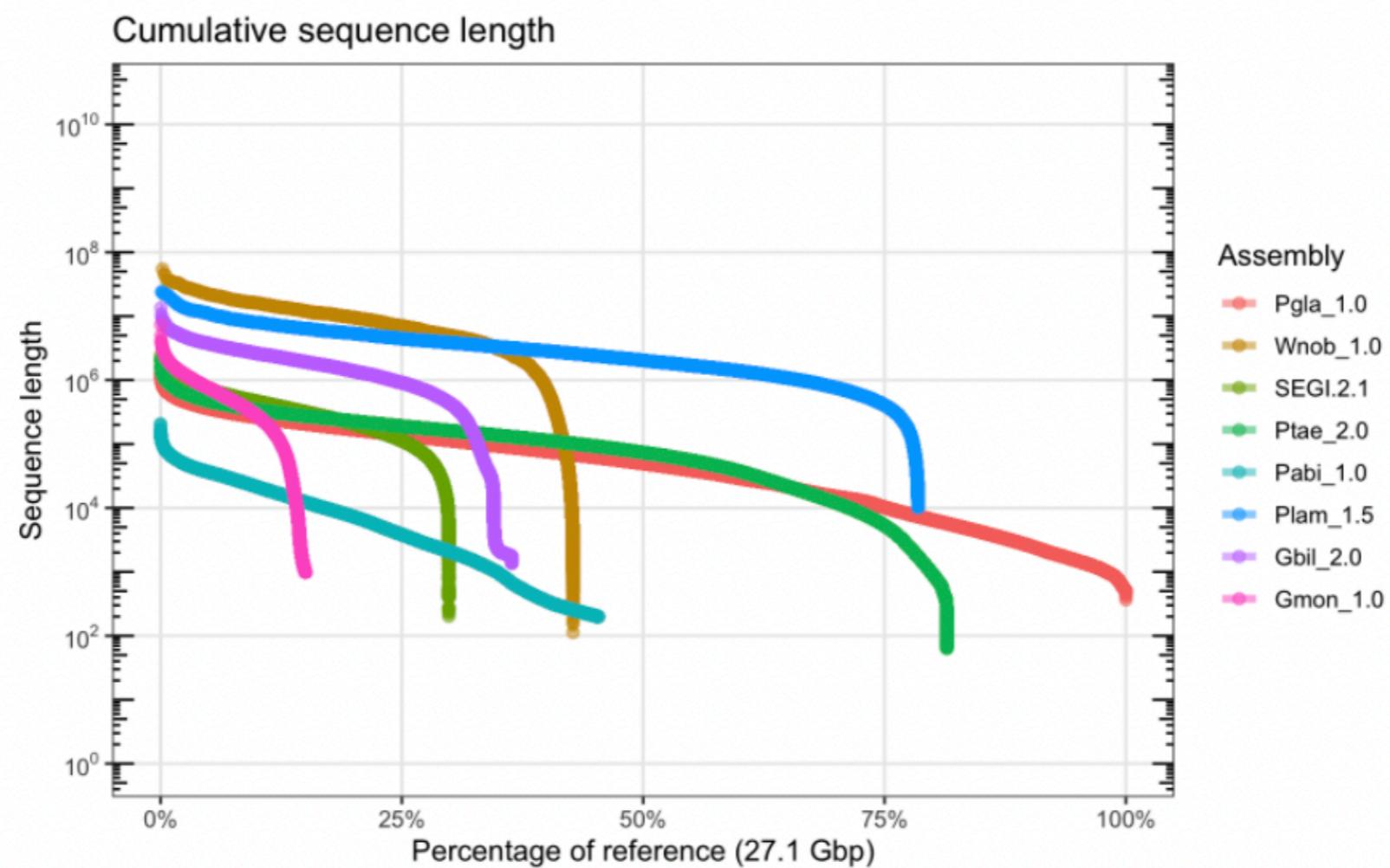
Wollemia nobilis Genome Assembly

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%)

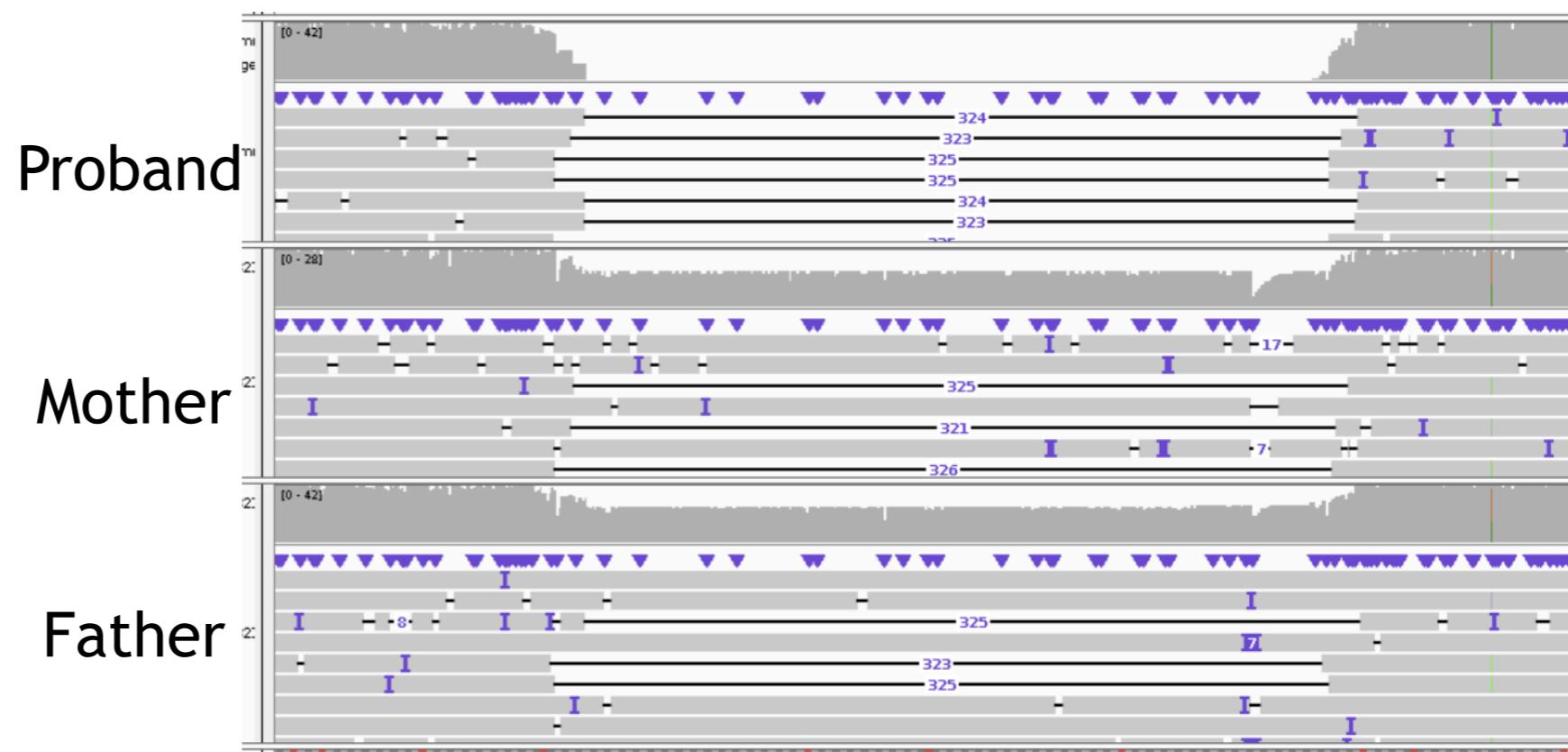


Long Read Sequencing of Early Onset Cancer Pedigrees

SV Filtering Workflow

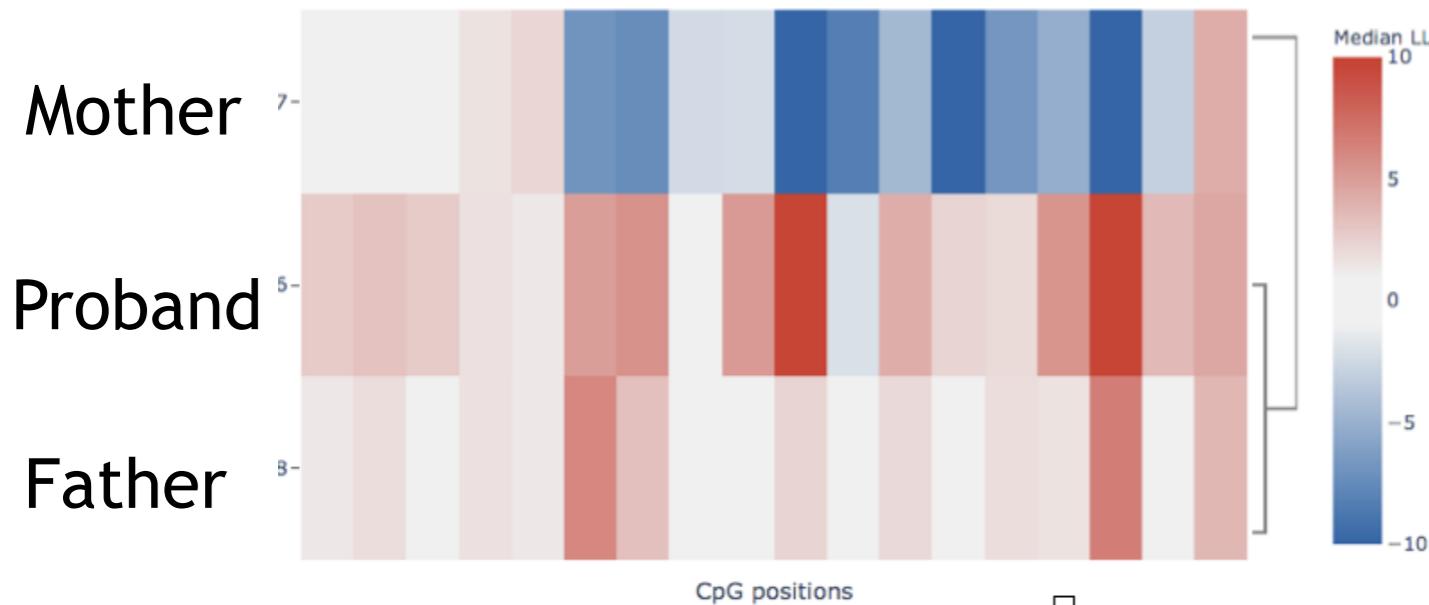
- Each individual will have ~25000 SV calls per genome
- Merge and genotype all calls across family members (~34000)
- Filter by family structure (~1400)
- Pull variants in/near genes (~450)
- Filter common events (~300)
- Filter false positives / ambiguous events/ select likely genes (~<100)

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



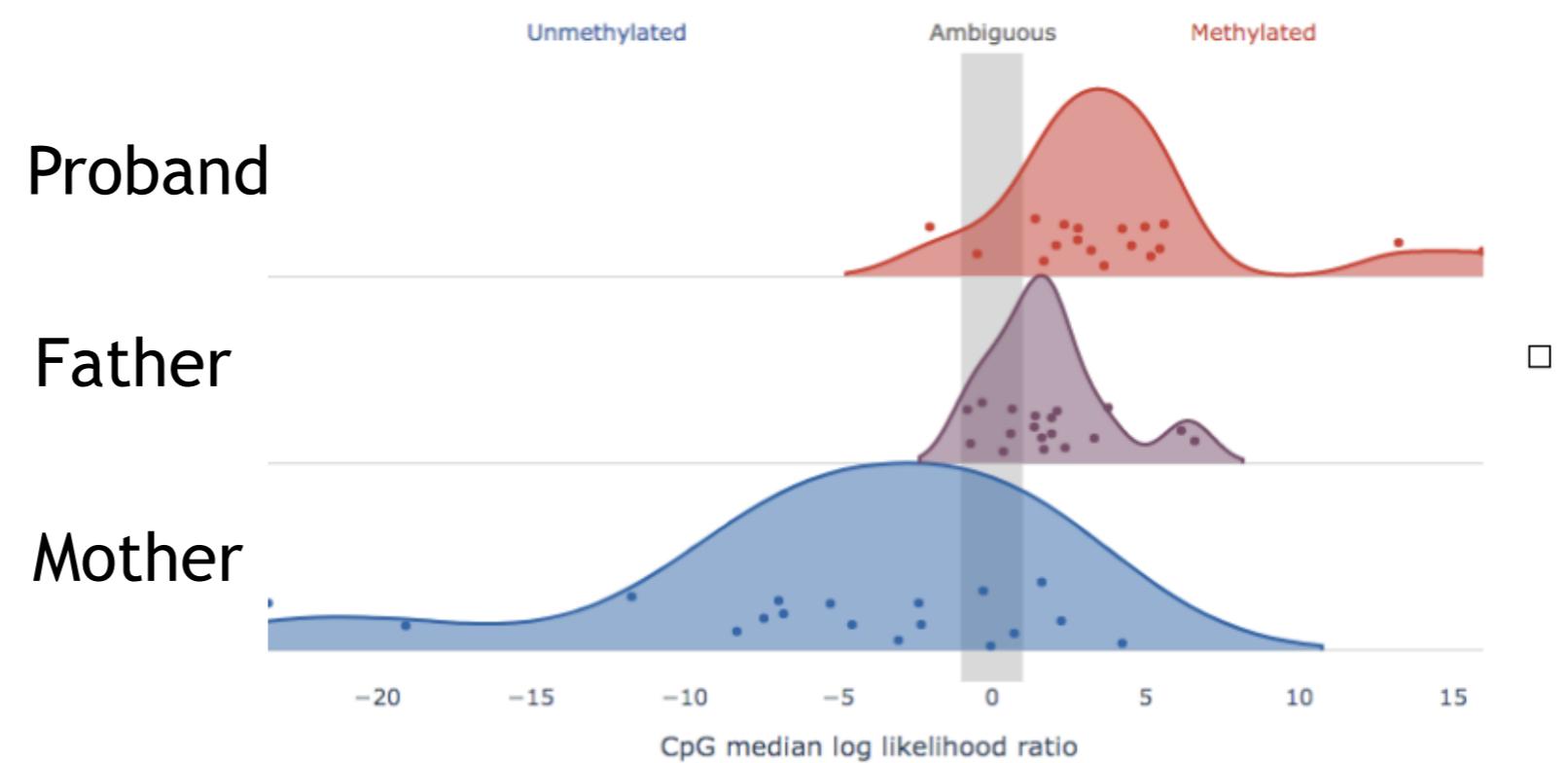
Homozygous intronic gene deletion in proband, heterozygous in healthy parents

Long Read Sequencing of Early Onset Cancer Pedigrees



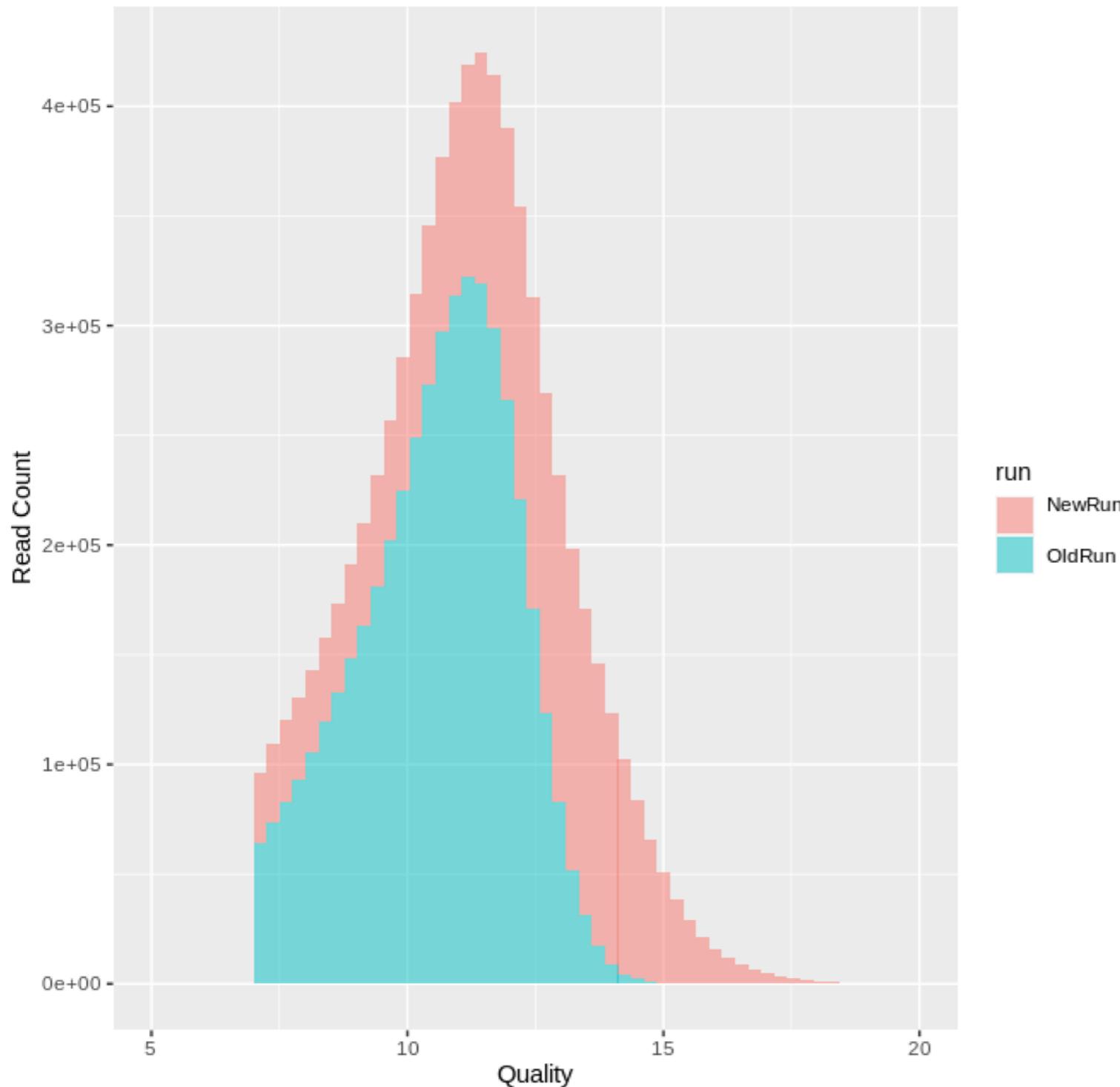
Hypermethylation of promoter region of tumor suppressor in proband compared to healthy parents

ONT signal data allows for direct detection of methylation state



Collaboration with Zsofia Stadler MSKCC

Nanopore Quality Distribution

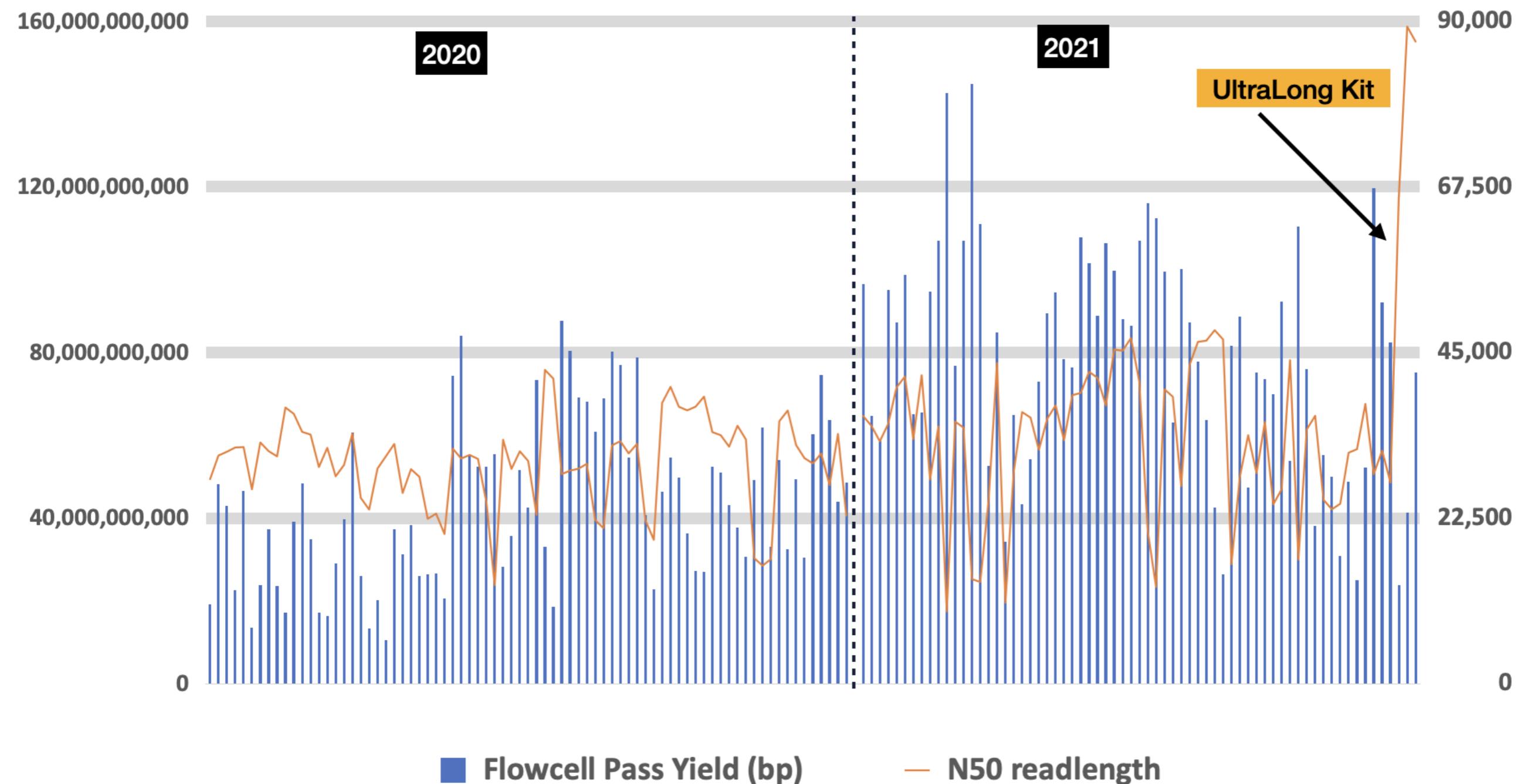


**ONT software
improvements are
increasing base quality**

“Guppy V5” sup accuracy model

“Guppy V4” high accuracy model

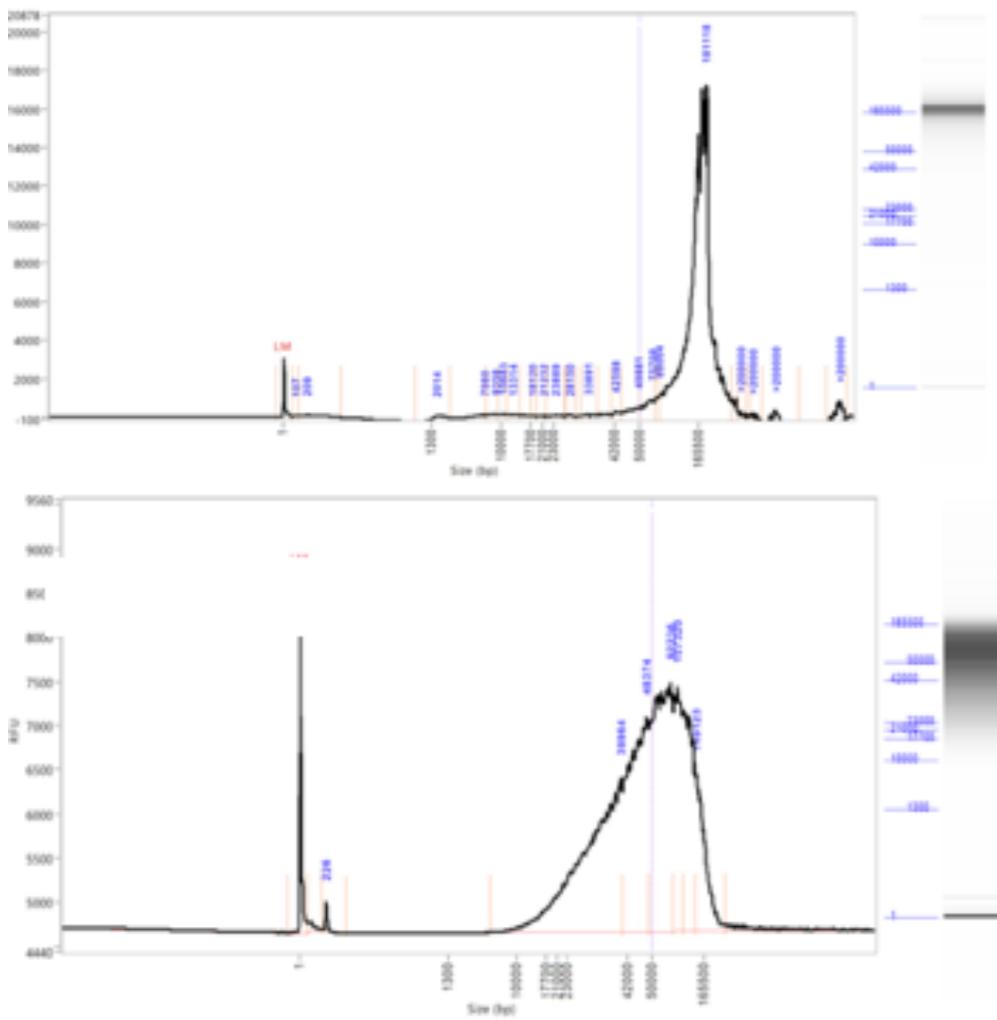
Yield vs N50



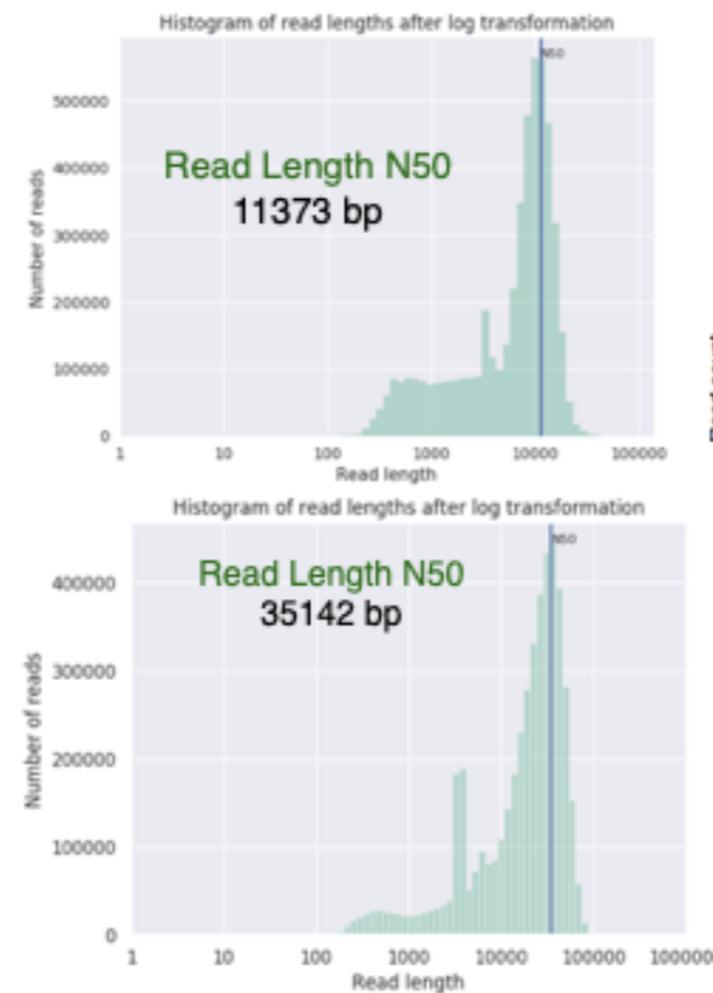
Optimization of long read sequencing on the PromethION



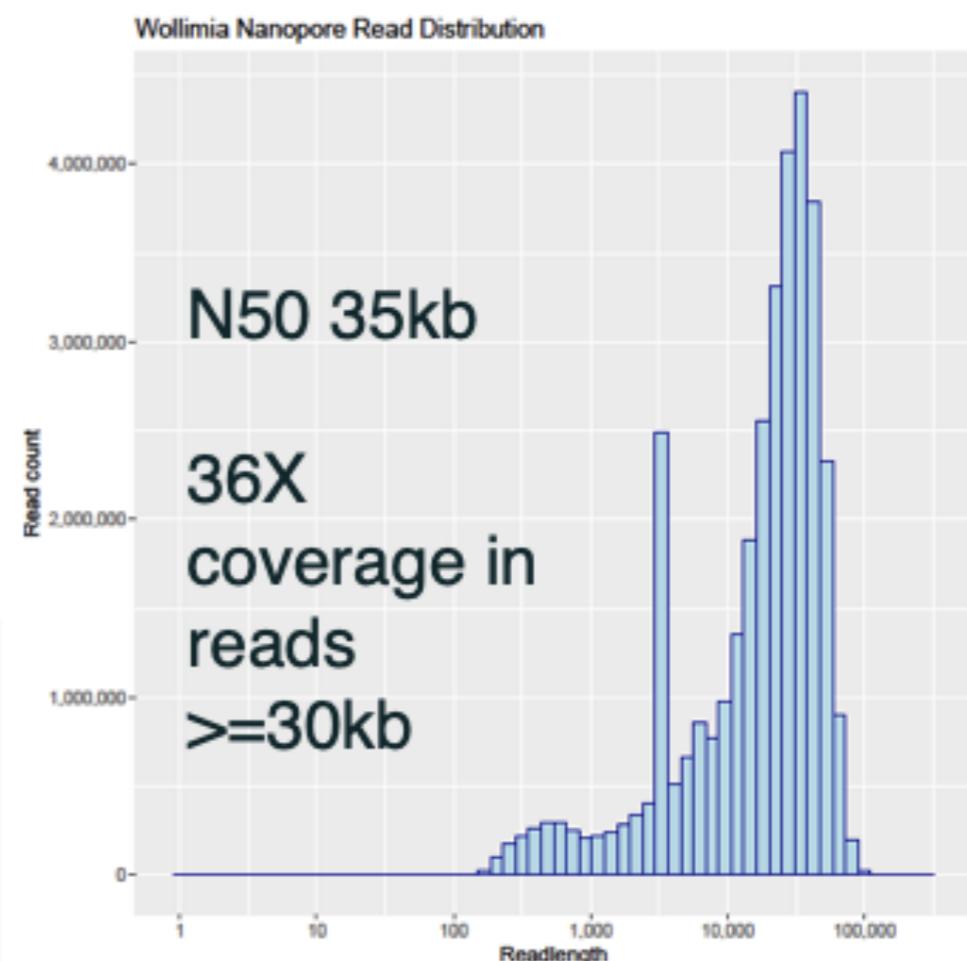
Femto Pulse Fragment Size Estimations
before and after protocol adjustments for
shearing and application of SRE



Read Length Distributions before and
after protocol adjustments for shearing
and application of SRE



Final ONT read length distribution

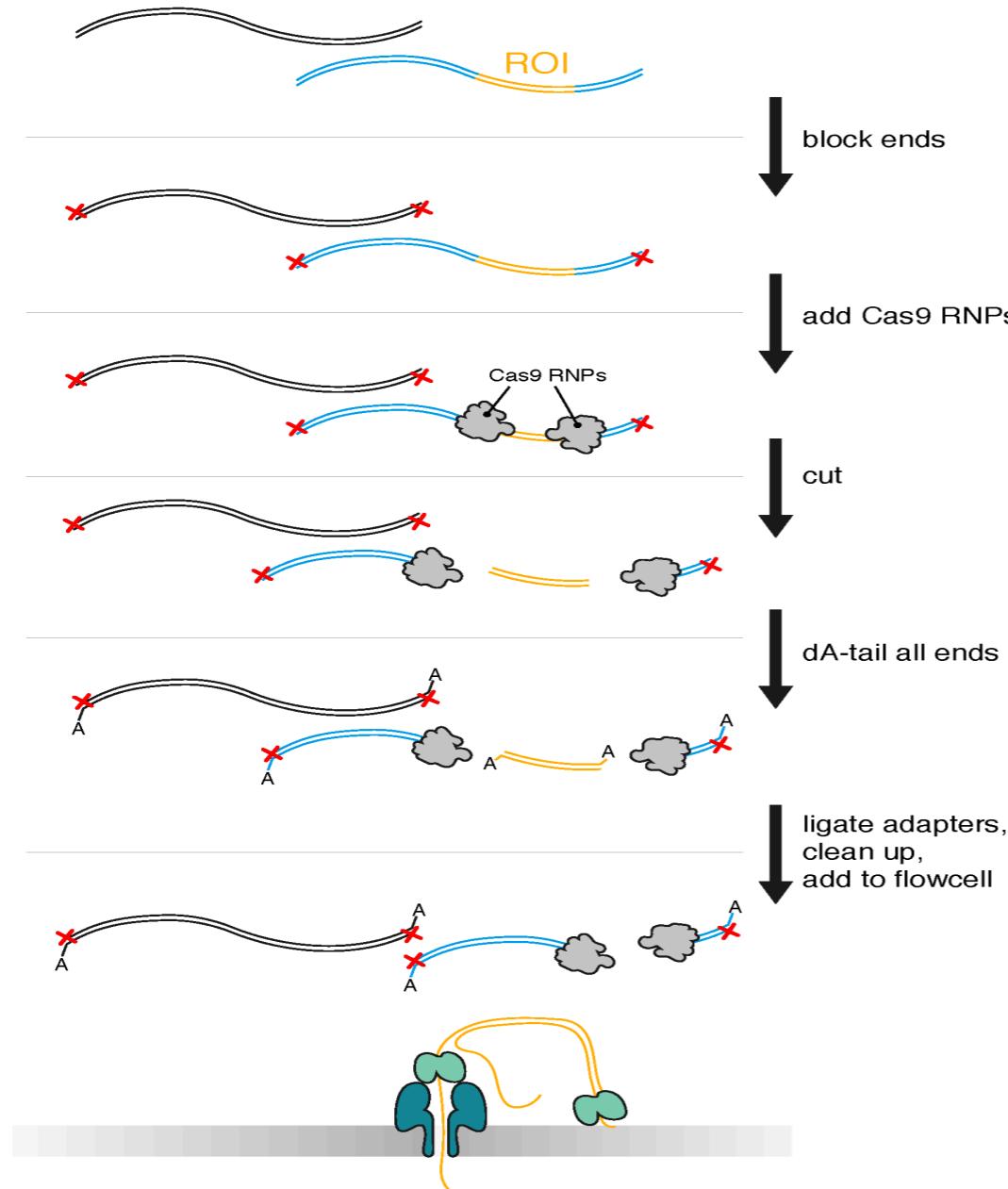


Oxford
NANOPORE
Technologies

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
 -

Cas9-mediated PCR-free enrichment (nCATS)



Targeted long reads

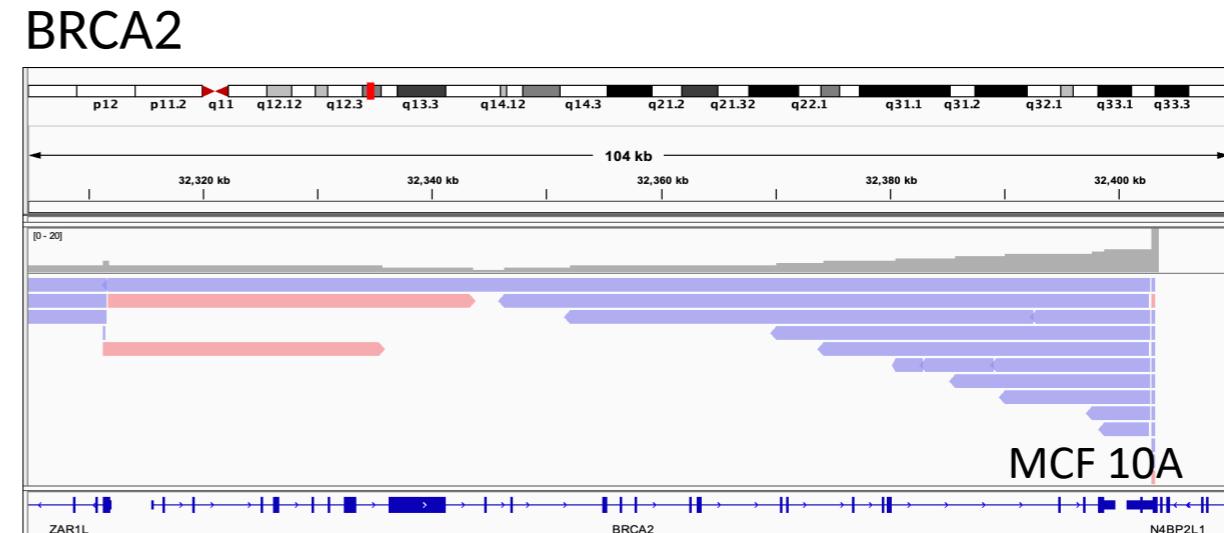
- Cover entire target region with a single read
- Reads covering entire target regions minimize mapping errors due to SVs

MCF 10A & SK-BR-3

- Amplification panel - can get methylation calls

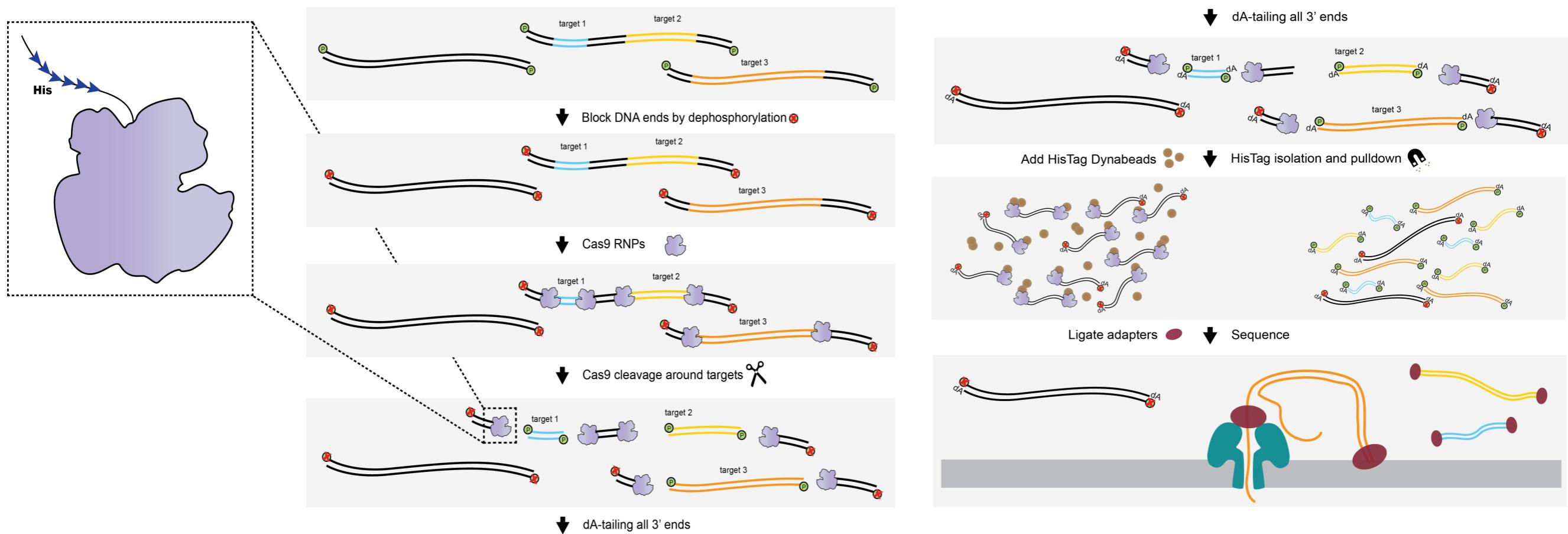
Genes targeted and coverage obtained from nCATS runs

Gene	Target size (bp)	MCF10A
		Coverage nCATS
MYC	12565	141
HOXA9	18506	40
FGFR4	19916	69
STK11	30286	36
CDKN2A	30774	NA
TERT	44787	13
KRAS	50955	23
BRCA2	91218	3
PAX7	120934	5
APC	144265	4
Total target	564206	15



Gene	Number of reads covering X% of gene			
	100 %	80%	60%	40%
MYC	151	165	178	185
FGFR4	79	82	86	99
HOXA9	42	44	46	49
STK11	32	35	42	52
TERT	5	8	8	15
KRAS	11	18	28	39
BRCA2	1	1	2	3
PAX7	1	1	1	4
APC	0	0	0	2

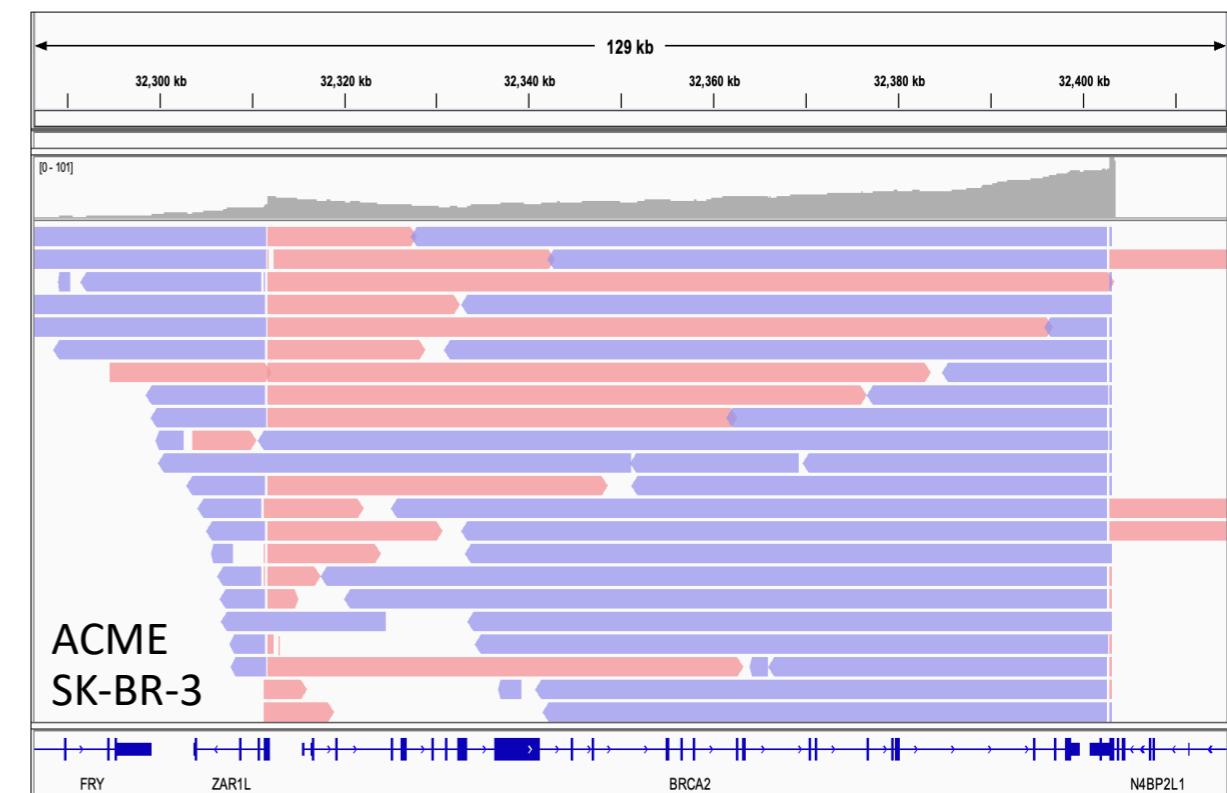
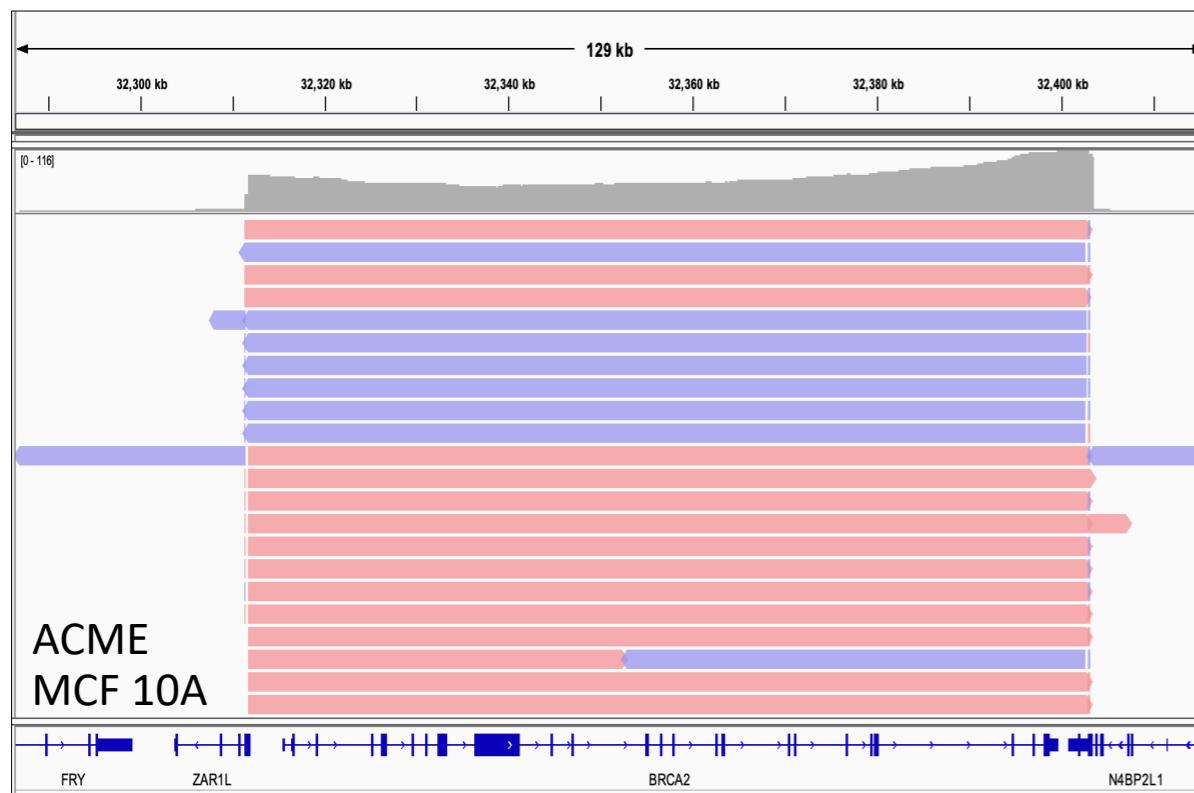
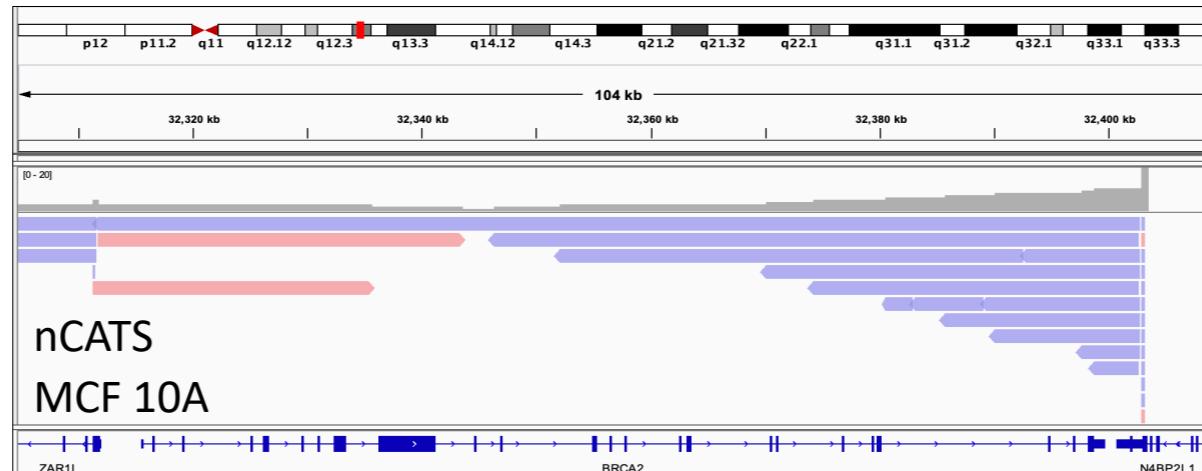
Affinity-based Cas9-Mediated Enrichment (ACME)



Coverage of genes targeted using nCATS vs ACME

Gene	Target size (bp)	Coverage		
		MCF10A nCATS	MCF 10A ACME	SK-BR-3 ACME
MYC	12565	141	1025	2274
HOXA9	18506	40	246	223
FGFR4	19916	69	467	124
STK11	30286	36	175	93
CDKN2A	30774	NA	NA	55
TERT	44787	13	105	90
KRAS	50955	23	158	57
BRCA2	91218	3	55	35
PAX7	120934	5	31	25
APC	144265	4	22	13
Total target	564206	15	101	97

BRCA2 (~90 kb target size)



Spanning read counts from nCATS and ACME runs

MCF10A nCATS

Gene	Number of reads covering X% of gene			
	100 %	80%	60%	40%
MYC	151	165	178	185
FGFR4	79	82	86	99
HOXA9	42	44	46	49
STK11	32	35	42	52
CDKN2 A	NA	NA	NA	NA
TERT	5	8	8	15
KRAS	11	18	28	39
BRCA2	1	1	2	3
PAX7	1	1	1	4
APC	0	0	0	2

MCF10A ACME

Gene	Number of reads covering X% of gene			
	100%	80%	60%	40%
MYC	993	1036	1055	1075
FGFR4	425	444	471	489
HOXA9	233	242	253	263
STK11	152	161	168	183
CDKN2 A	NA	NA	NA	NA
TERT	61	74	87	119
KRAS	111	123	142	168
BRCA2	20	26	36	54
PAX7	2	4	7	26
APC	0	1	2	16

SKBR3 ACME

Gene	Number of reads covering X% of gene			
	100%	80%	60%	40%
MYC	2057	2210	2300	2424
FGFR4	47	106	118	133
HOXA9	186	206	227	239
STK11	64	79	89	102
CDKN2 A	35	41	54	60
TERT	0	6	7	106
KRAS	27	37	43	60
BRCA2	3	10	21	34
PAX7	1	3	7	18
APC	0	1	1	7

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

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