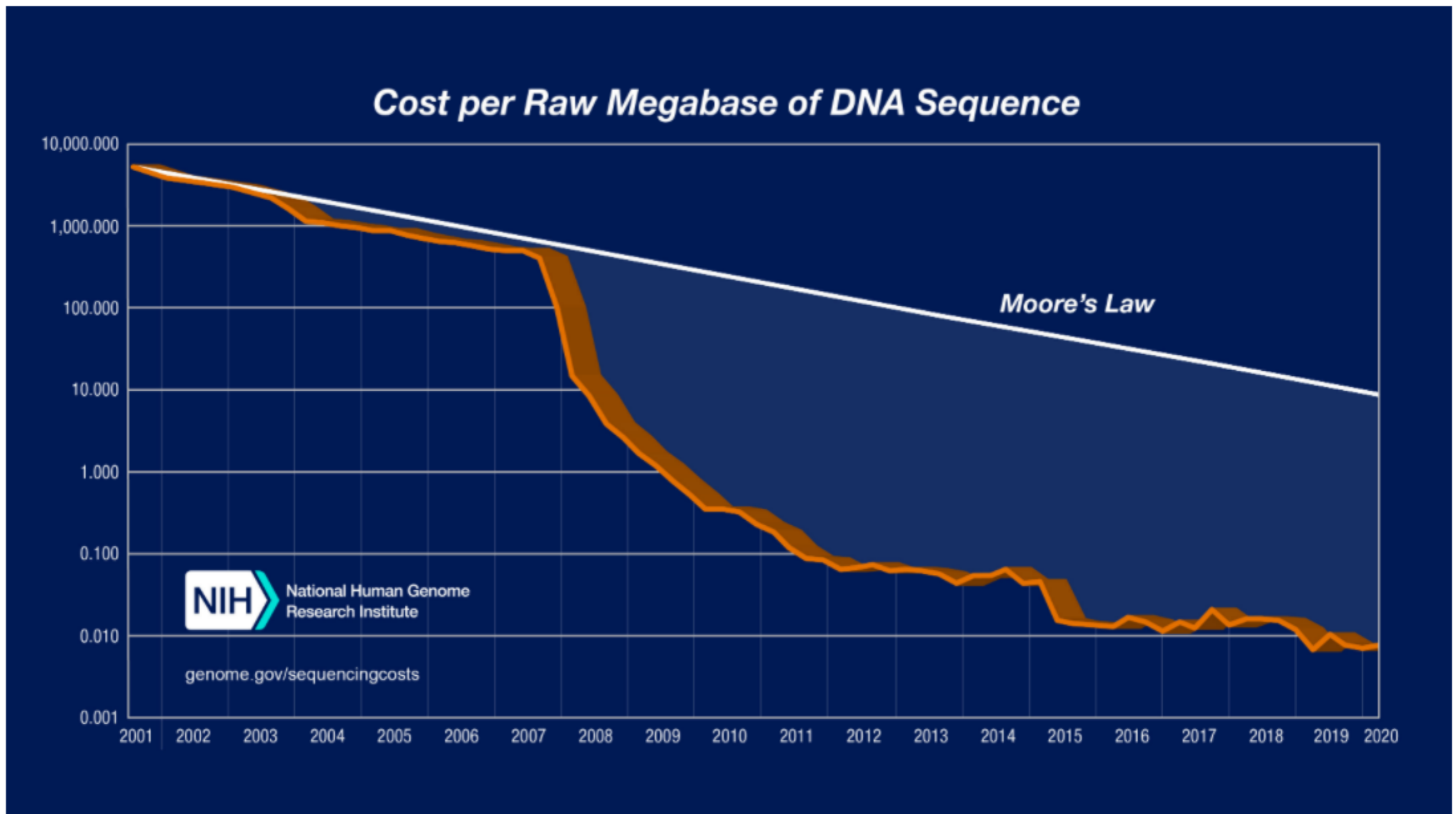


# Long Read Sequencing

Dick McCombie

Advanced Sequencing Technologies and Applications course  
Cold Spring Harbor Laboratory  
2020

# 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?

??





# 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

Enables a broader set of applications

Full scale of genetic variation

Repetitive regions

Structural variants

Enables higher quality alignments and assembly

Less fold coverage required

Finished genomes

# Limitations of long reads

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

# Two “flavors” of long read sequencing

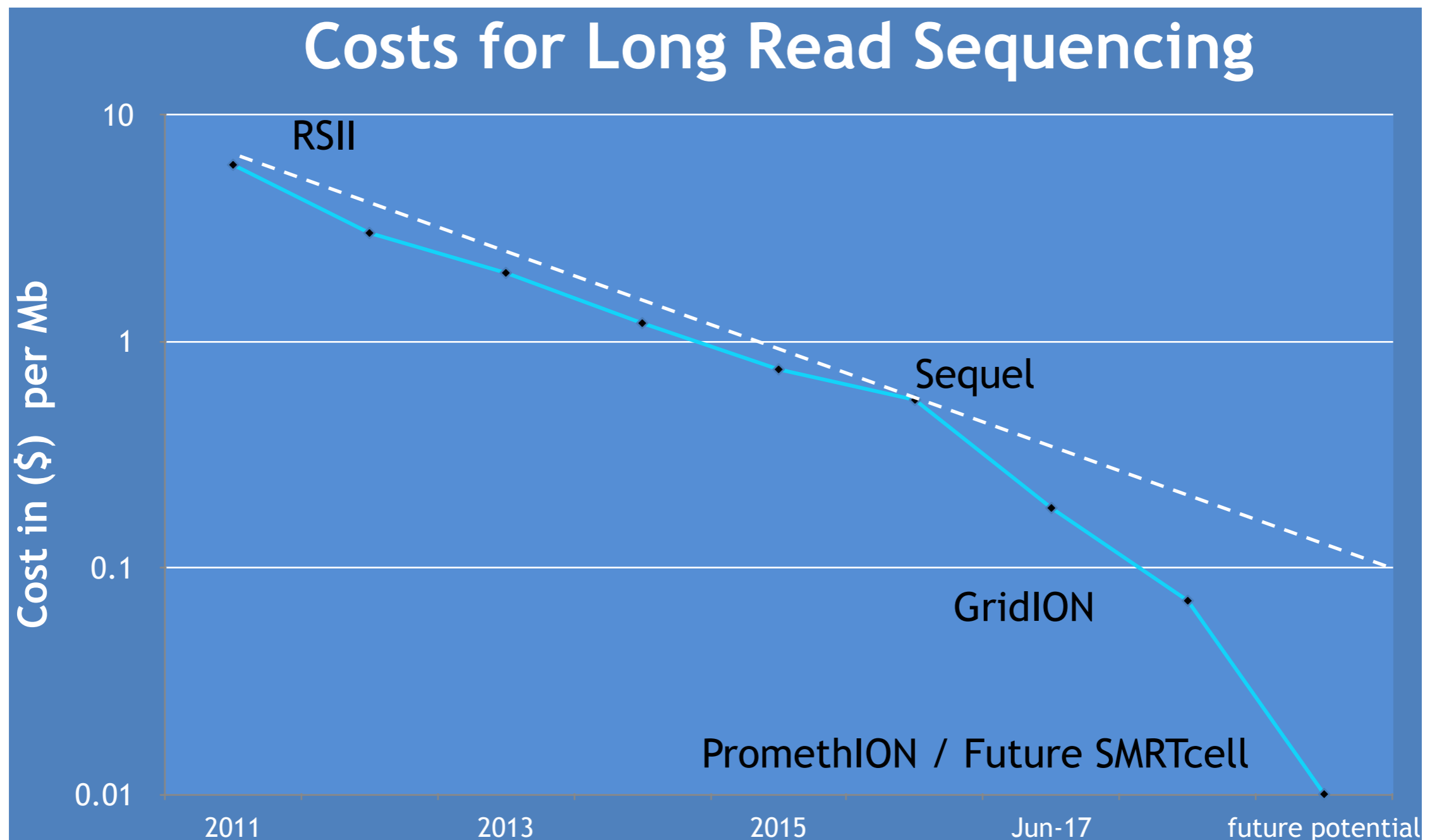


PACIFIC  
**BIOSCIENCES**<sup>®</sup>



Oxford  
**NANOPORE**  
Technologies

# 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

1 Million ZMW (155k RSII)

Early runs were rocky

Substantial recent improvement in  
performance

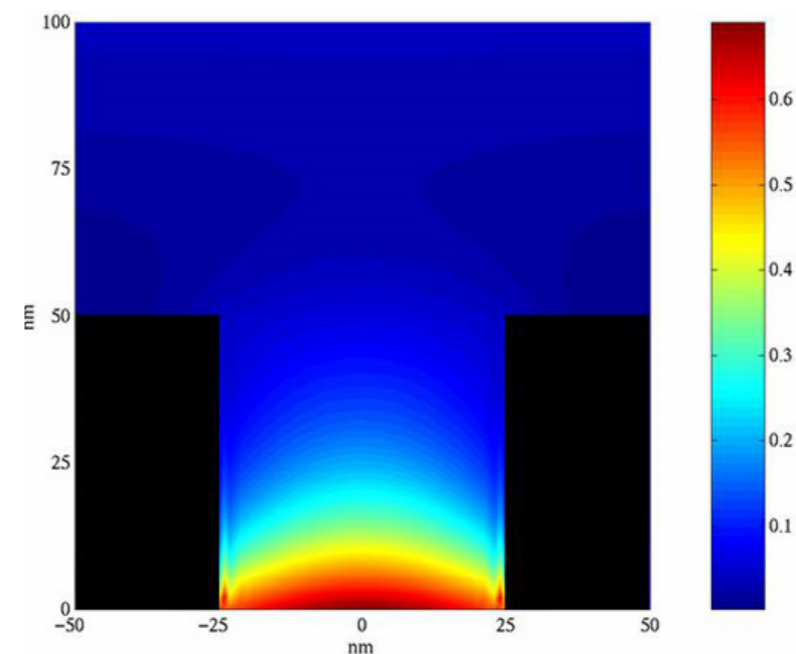
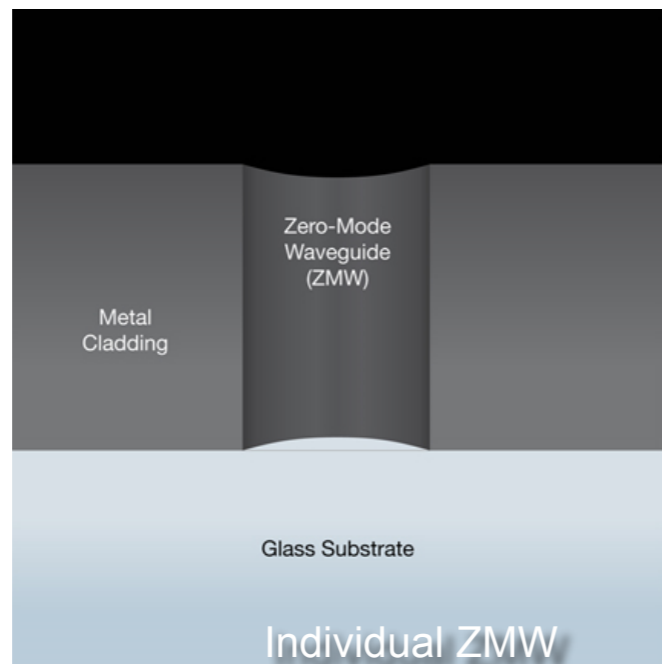


# Zero-Mode Waveguides Are the Observation Windows

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

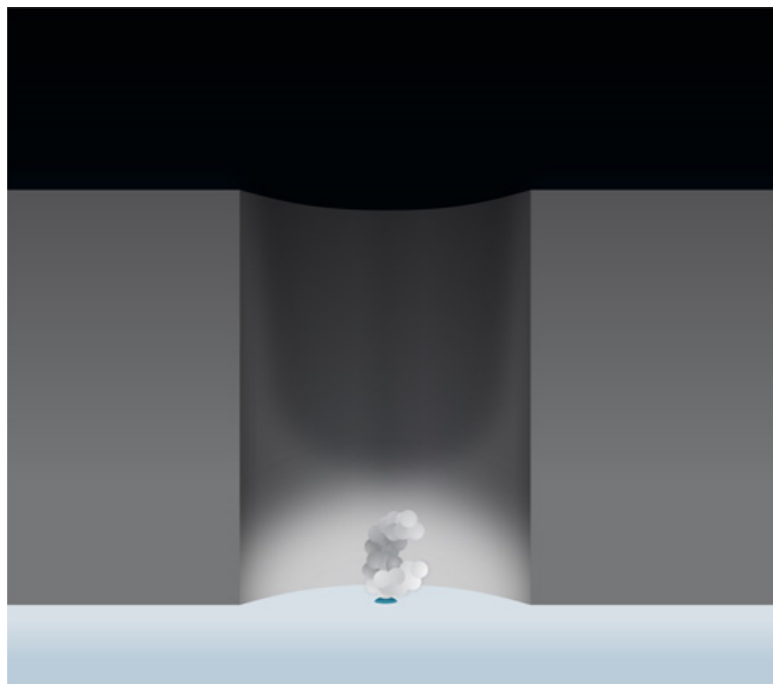


Laser light illuminates the ZMW

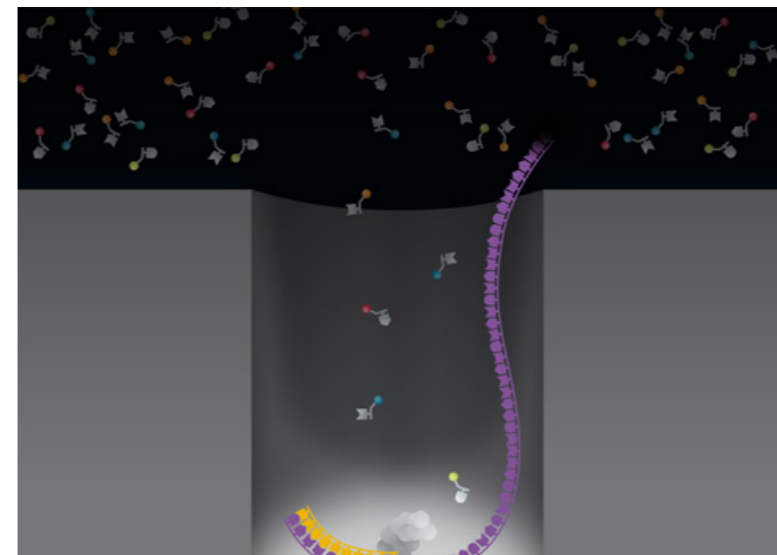
# DNA Polymerase as a Sequencing Engine

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

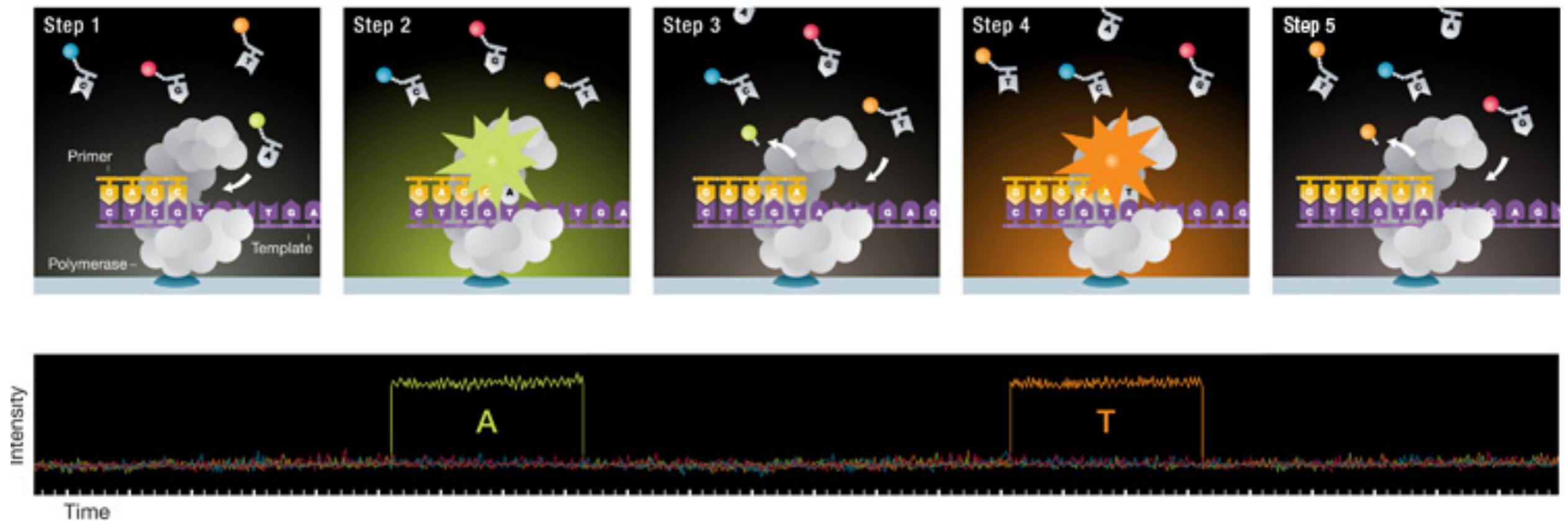


ZMW with DNA  
polymerase and  
phospholinked  
nucleotides

# 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



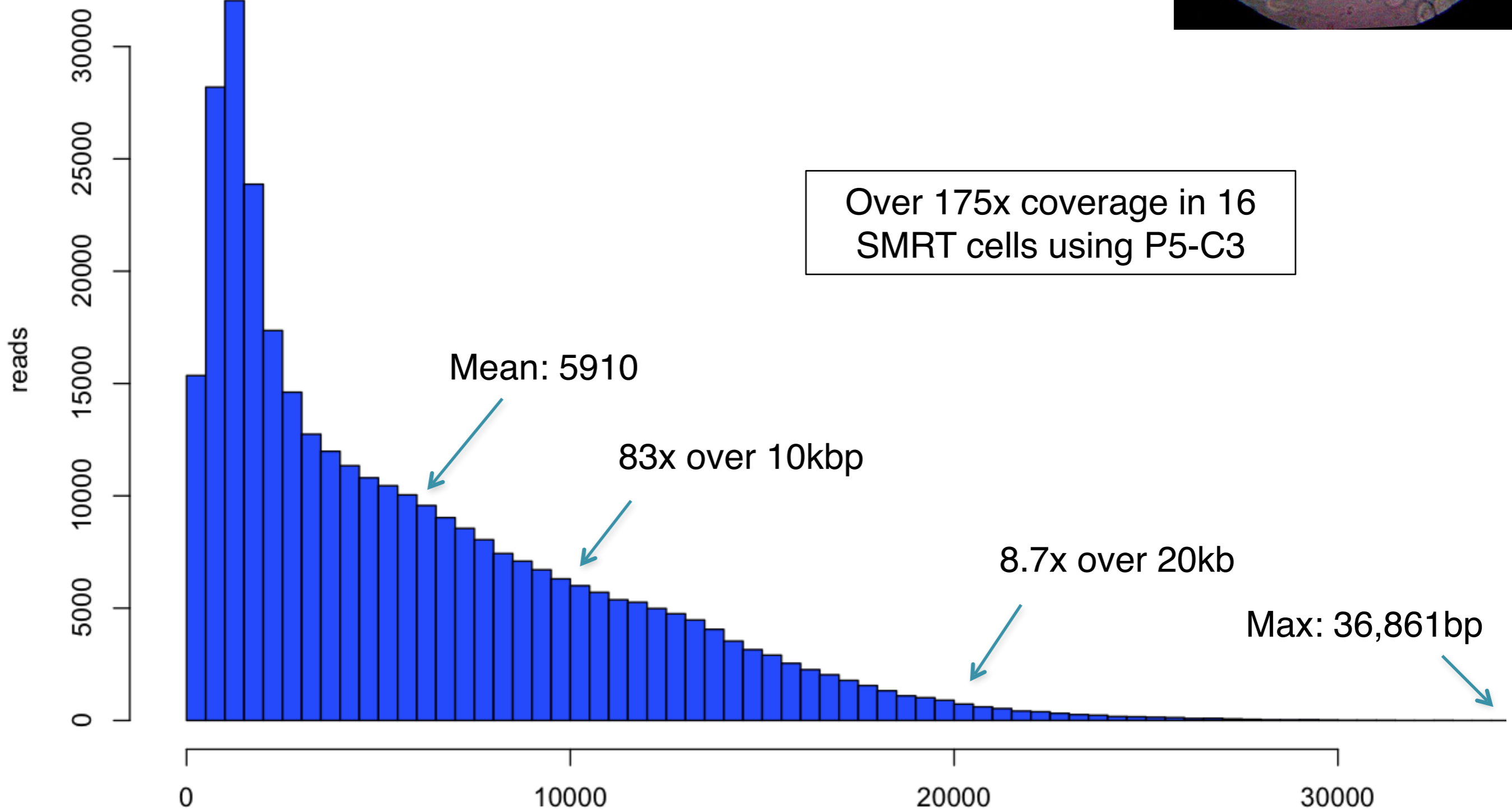
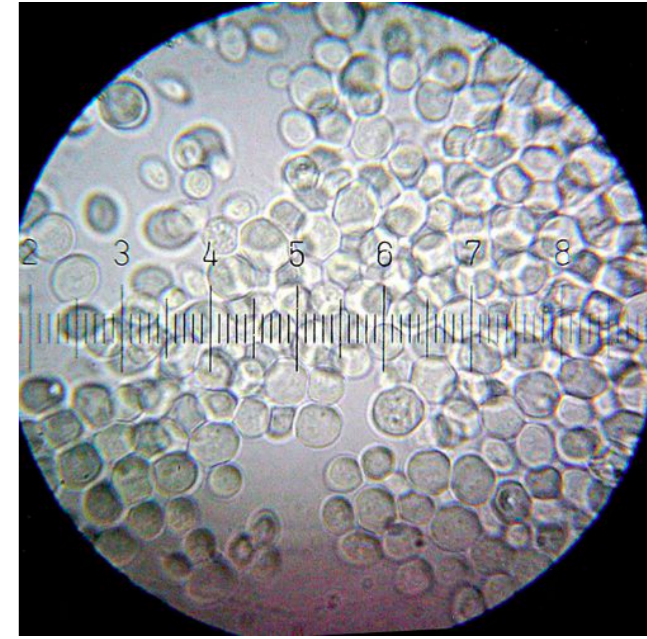
**LIGHTS ALL ASKEW IN THE HEAVENS;  
Men of Science More or Less Agog Over Results  
of Eclipse Observations. EINSTEIN THEORY  
TRIUMPHS Stars Not Where They Seemed or  
Were Calculated to be, but Nobody Need Worry.  
A BOOK FOR 12 WISE MEN No More in All  
the World Could Comprehend It, Said Einstein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



# 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



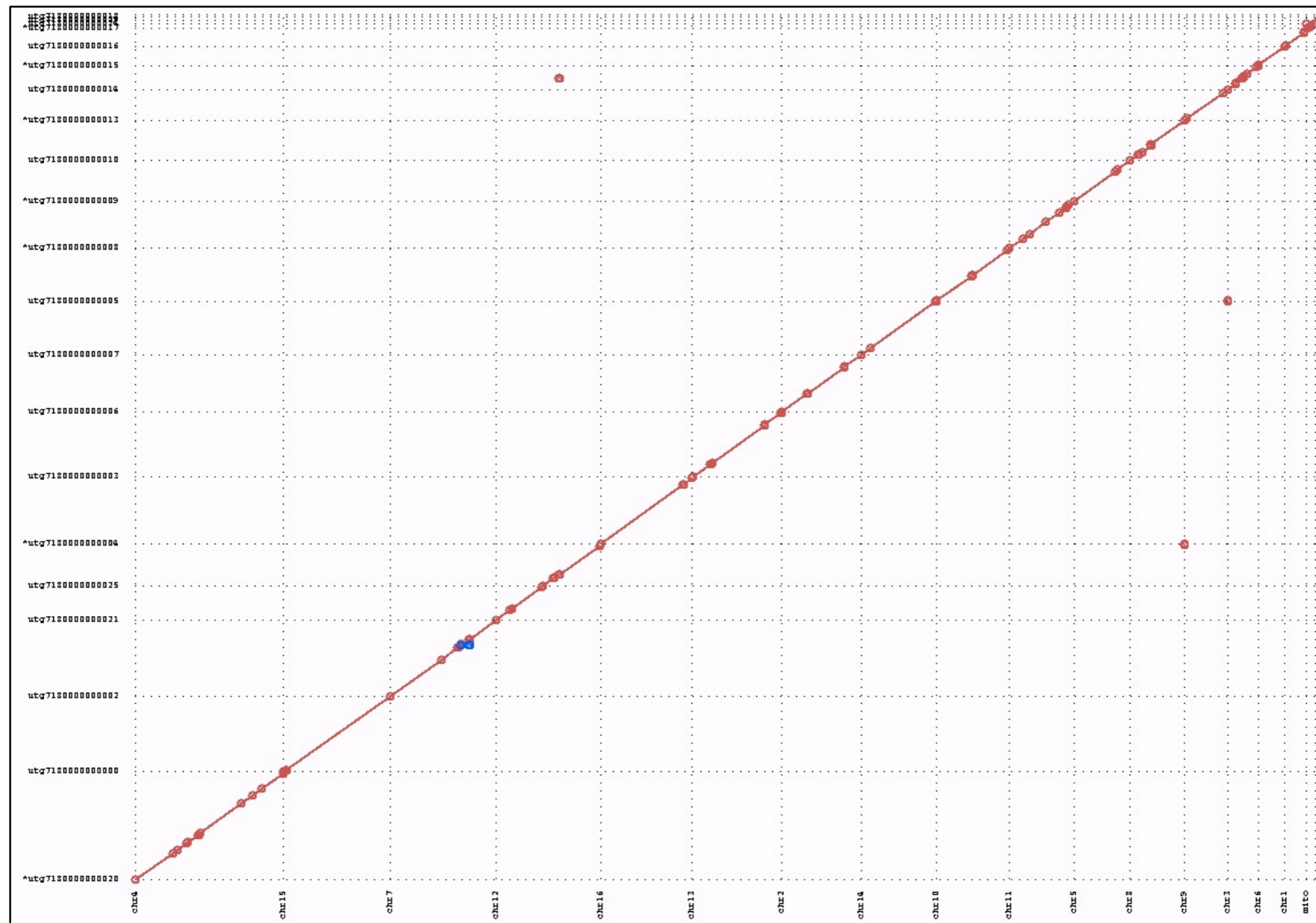
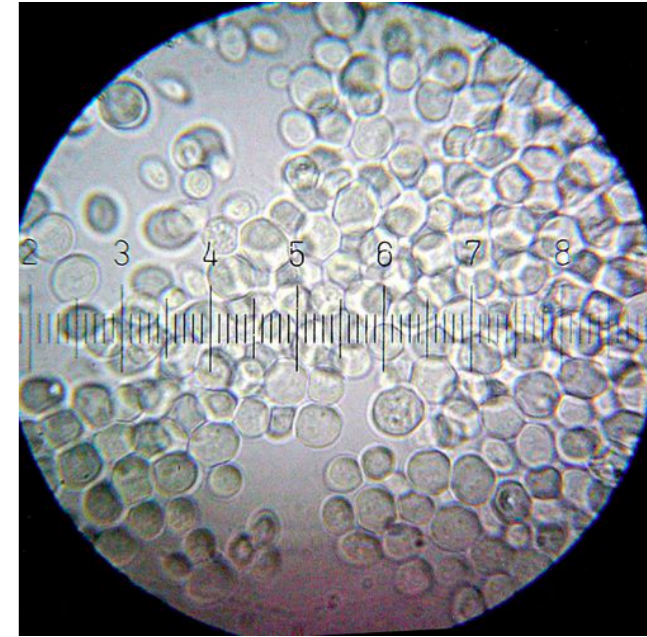
# *S. cerevisiae* W303

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



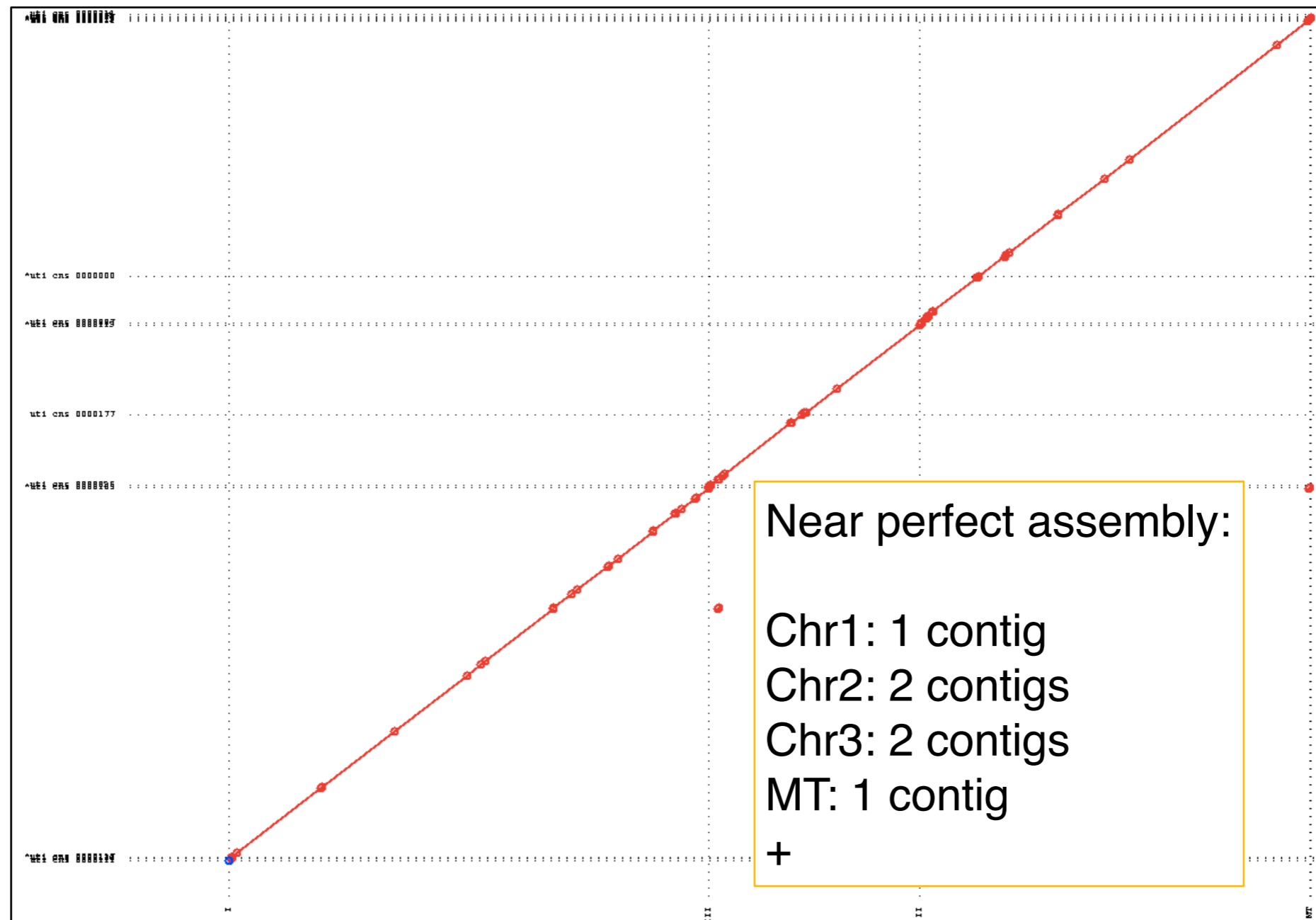
# *S. pombe* dg21

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



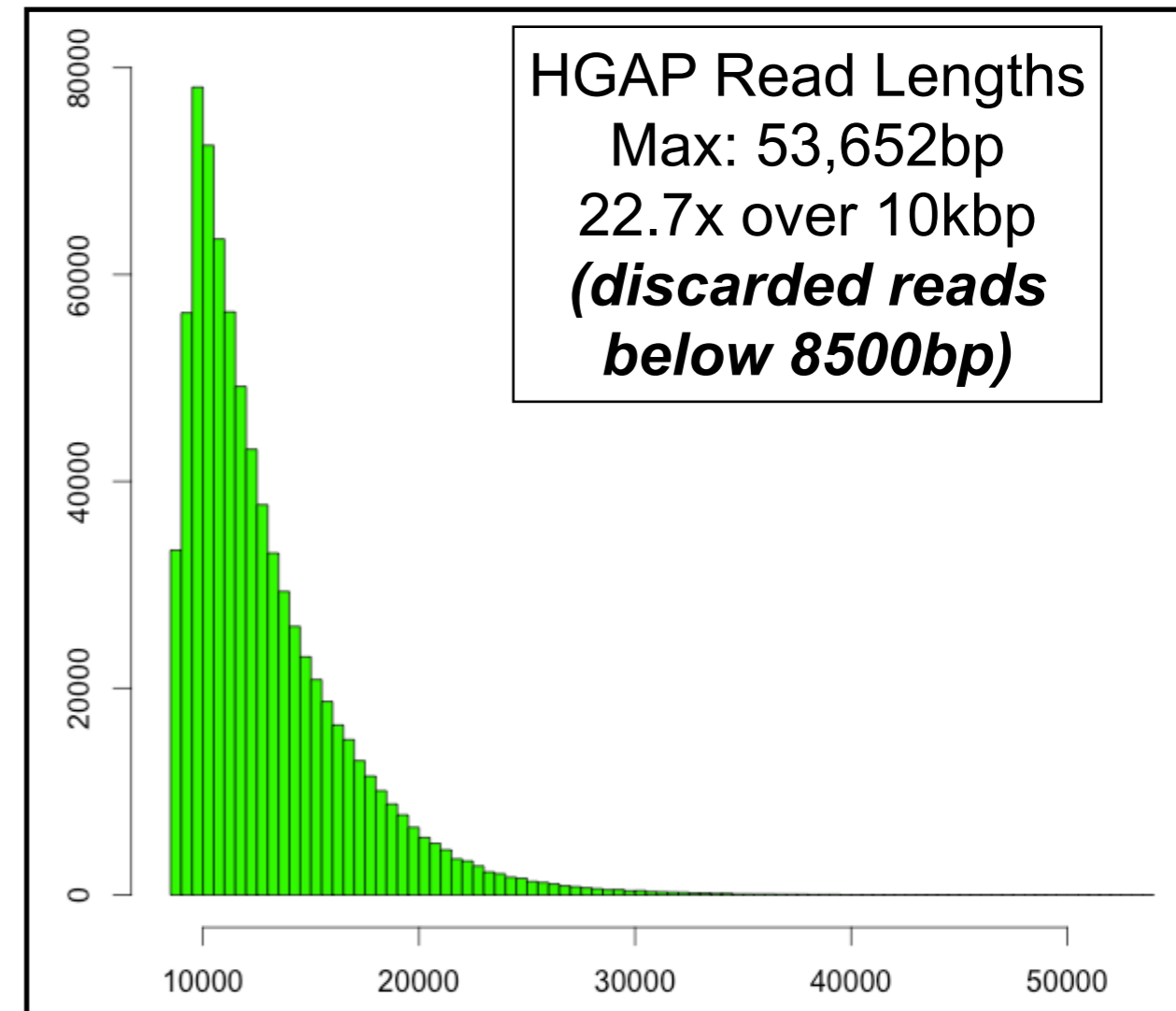


# O. sativa pv Indica (IR64)



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

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

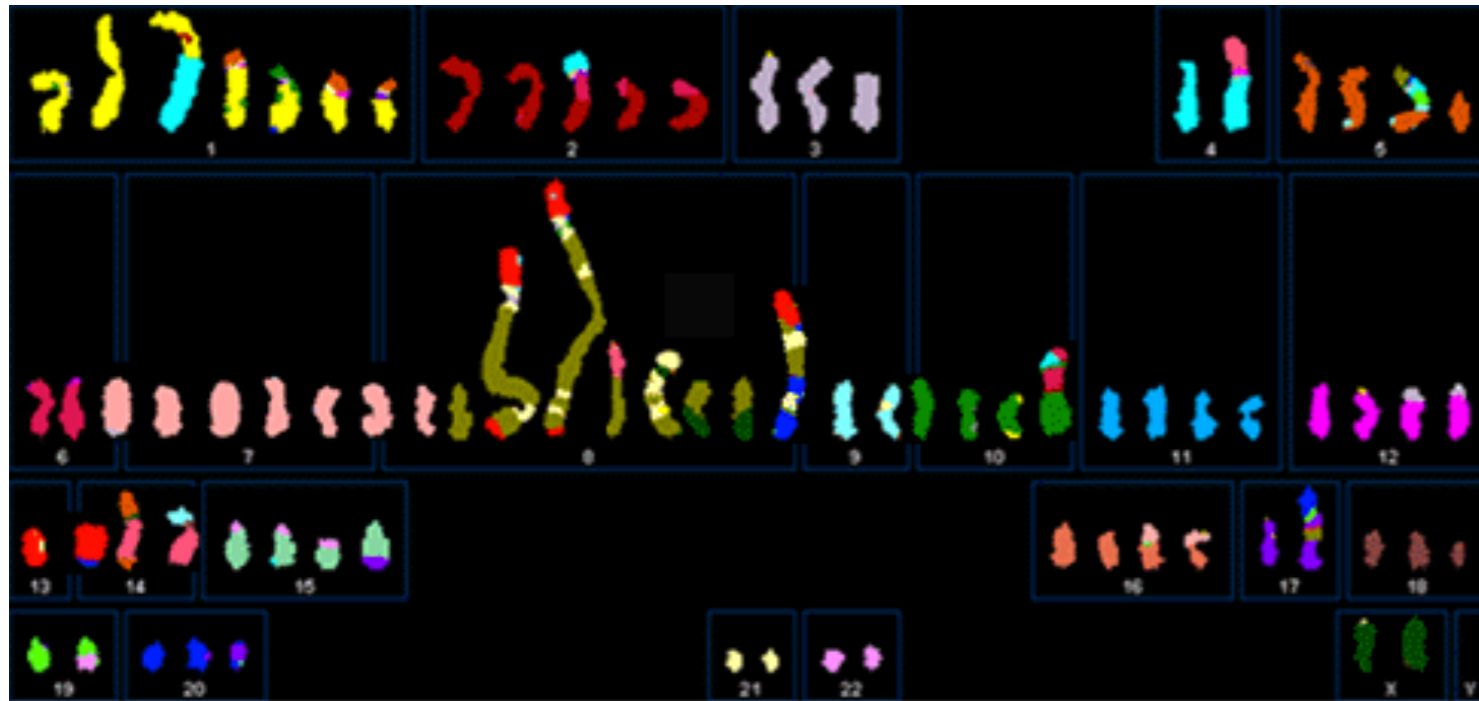


# Structural Variations in SKBR3

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)



# Importance of Structural Variations in Cancer

## ***Copy number changes***

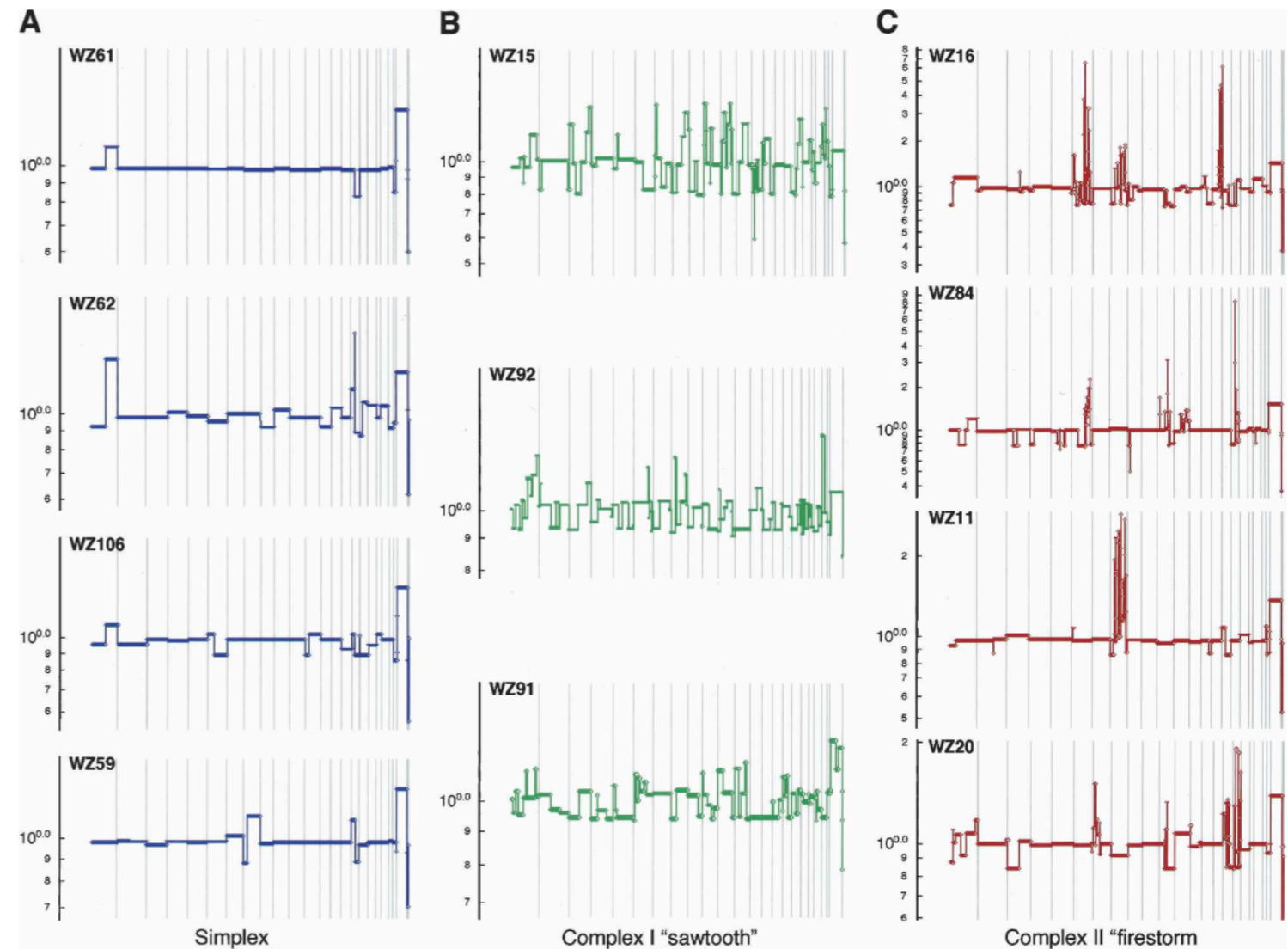
Especially amplification & deletions of oncogenes and tumor suppressors

## ***Gene Fusions***

Modifies protein sequence & function, potentially alters gene expression by fusing highly expressed transcript with lowly expressed transcript

## ***Prognostic indicator***

Greater genome instability generally leads to worse patient outcomes



**Figure 2.** Major types of tumor genomic profiles. Segmentation profiles for individual tumors representing each category: (A) simplex; (B) complex type I or sawtooth; (C) complex type II or firestorm. Scored events consist of a minimum of six consecutive probes in the same state. The y-axis displays the geometric mean value of two experiments on a log scale. Note that the scale of the amplifications in C is compressed relative to A and B owing to the high levels of amplification in firestorms. Chromosomes 1–22 plus X and Y are displayed in order from left to right according to probe position.

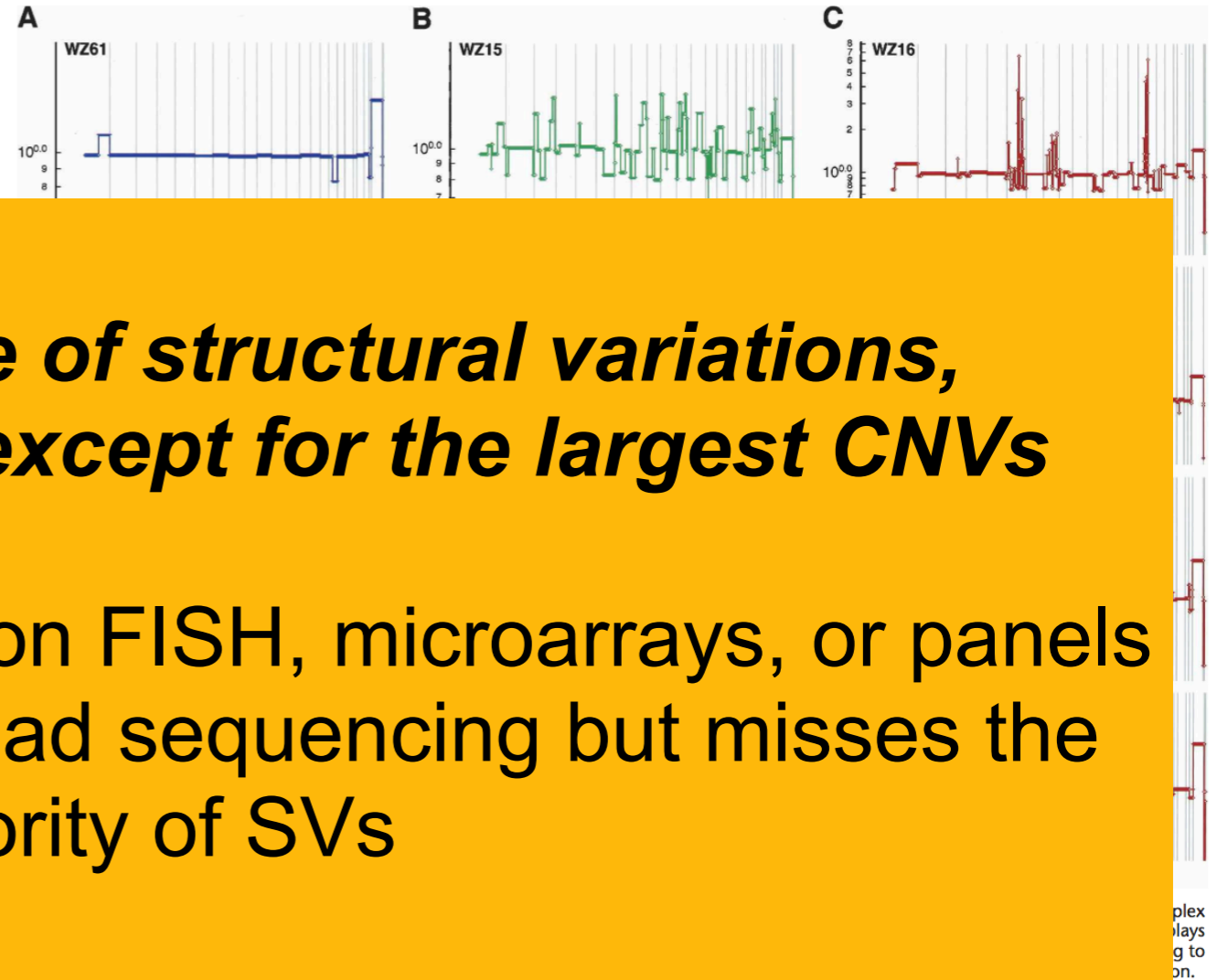
(Hicks *et al*, 2006, Genome Research)



# Importance of Structural Variations in Cancer

## ***Copy number changes***

Especially amplification & deletions of oncogenes and



***Despite the importance of structural variations, relatively little is known except for the largest CNVs***

Clinical standard: low resolution FISH, microarrays, or panels

Research standard: Short read sequencing but misses the vast majority of SVs

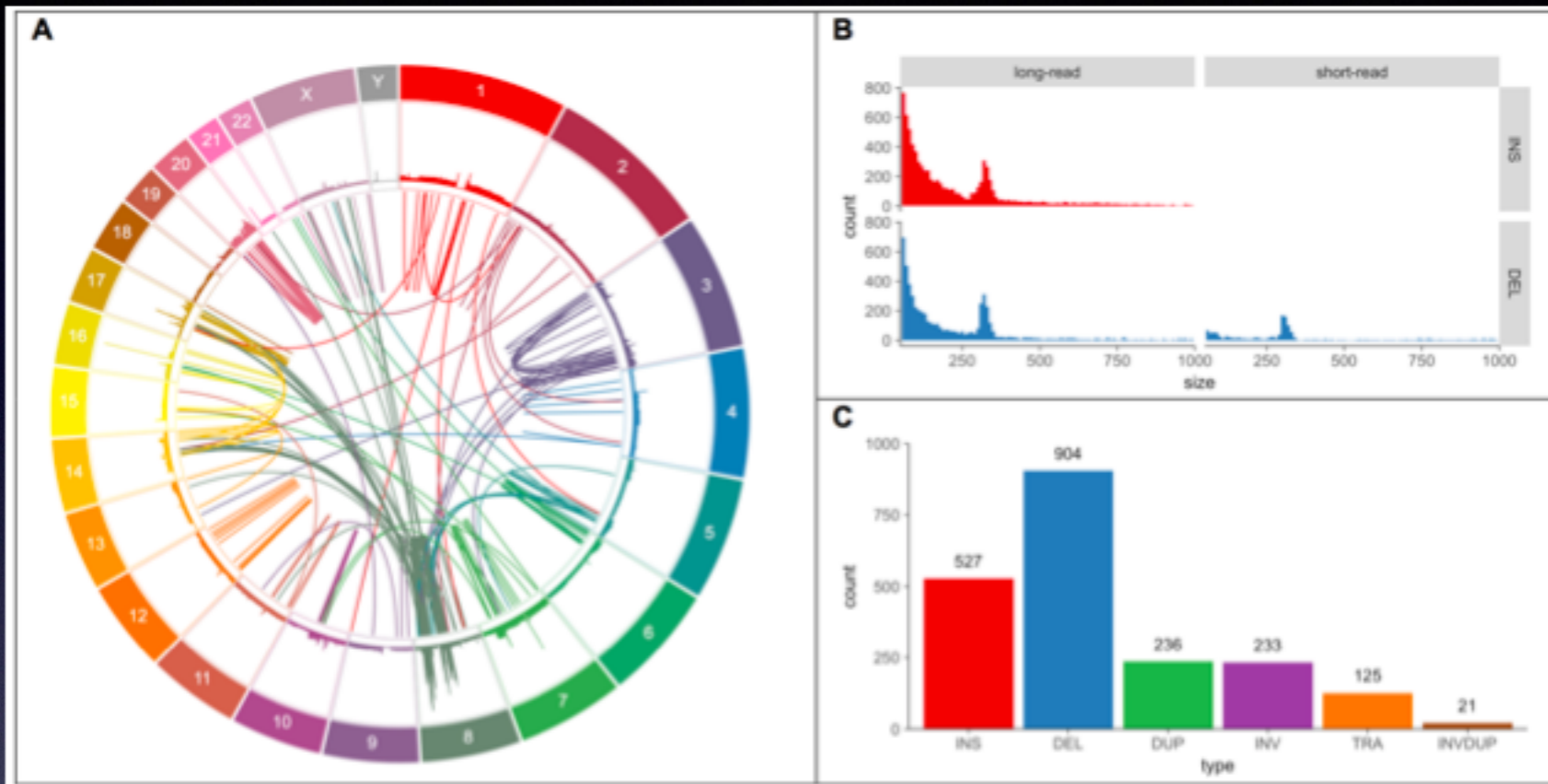
## ***Prognostic indicator***

Greater genome instability generally leads to worse patient outcomes

(Hicks *et al*, 2006, Genome Research)



# Structural Variations in SKBR3



**Figure 1** | Variants found in SK-BR-3 with PacBio long-read sequencing. (A) Circos plot showing long-range (larger than 10 kbp or interchromosomal) variants found by Sniffles from split-read alignments, with read coverage shown in the outer track. (B) Variant size histogram of deletions and insertions from size 50 bp up to 1 kbp found by long-read (Sniffles) and short-read (Survivor 2-caller consensus) variant-calling, showing similar size distributions for insertions and deletions from long reads but not for short reads where insertions are entirely missing. (C) Sniffles variant counts by type for variants above 1 kbp in size, including translocations and inverted duplications.

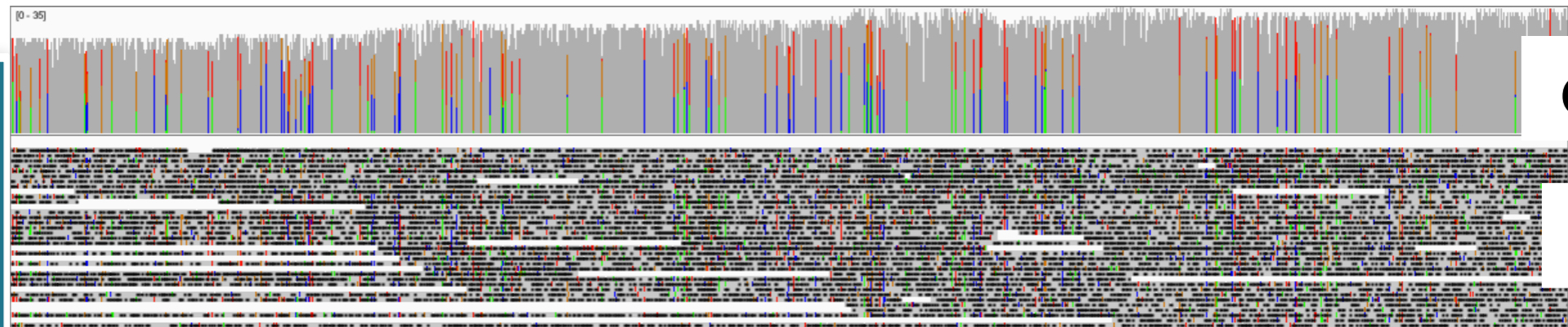
- Finding 10s of thousands of additional variants in the cancer
- PCR validation confirms high accuracy of long read calls
- With improved SV analysis, can infer the progression of the cancer
- Detect many novel gene fusions

***Complex rearrangements and oncogene amplifications revealed by long-read DNA and RNA sequencing of a breast cancer cell line***  
Nattestad, M et al (2018) Genome Research



# PacBio coverage is more stable than Illumina coverage in repetitive regions

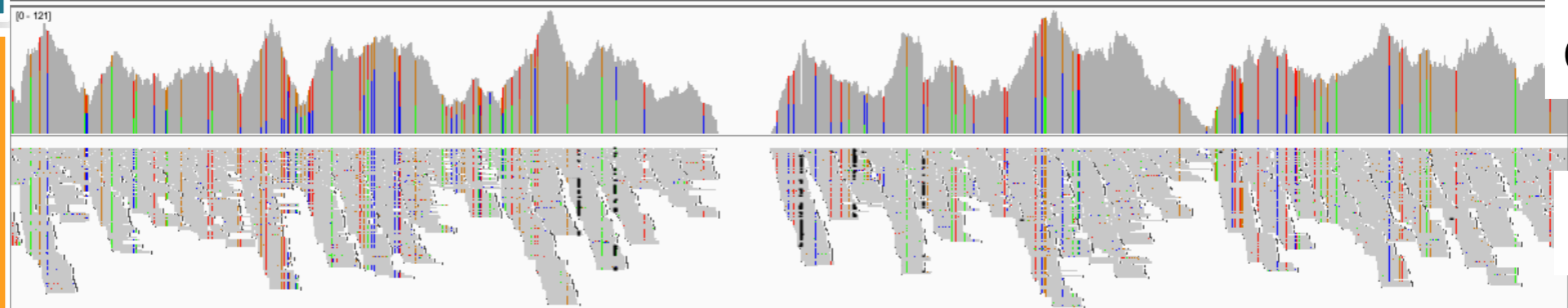
PacBio



coverage

reads

Illumina



coverage

reads



HLA-A

# 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**



PACIFIC  
BIOSCIENCES®

Number of sequences:

748,955

Total sequence length:

2.07 Gb

Mean: 2.8 kb

Max: 61 kb

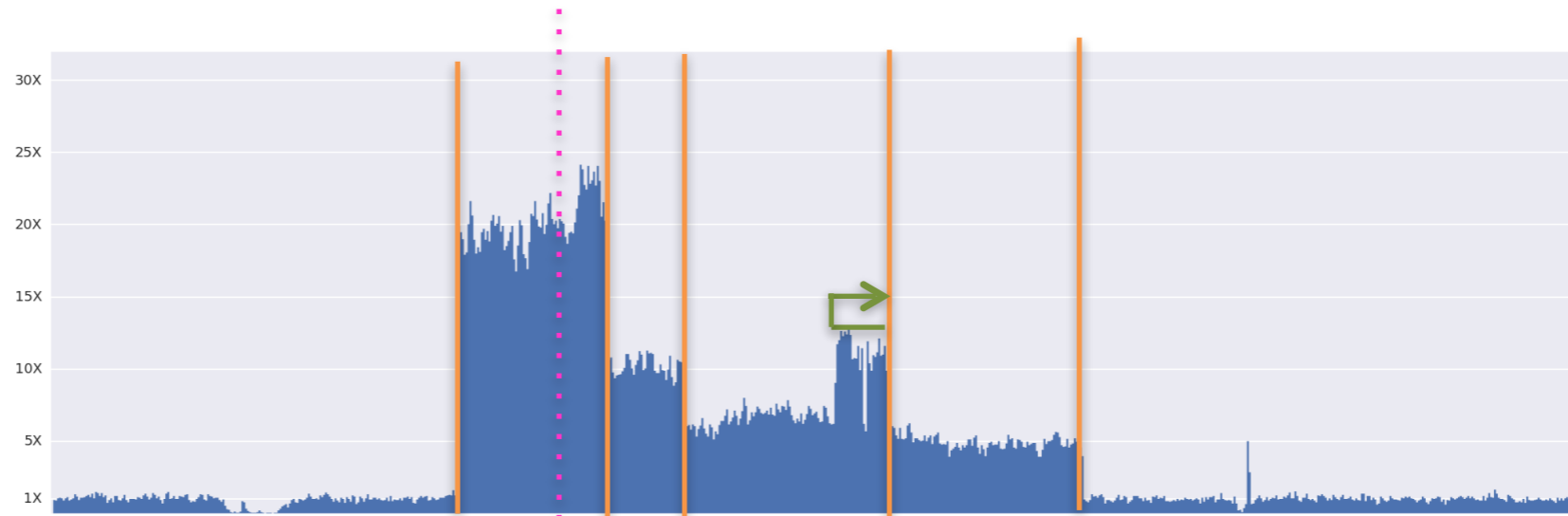
N50: 3.3 kb

**NG50: 1.9 kb**

illumina®

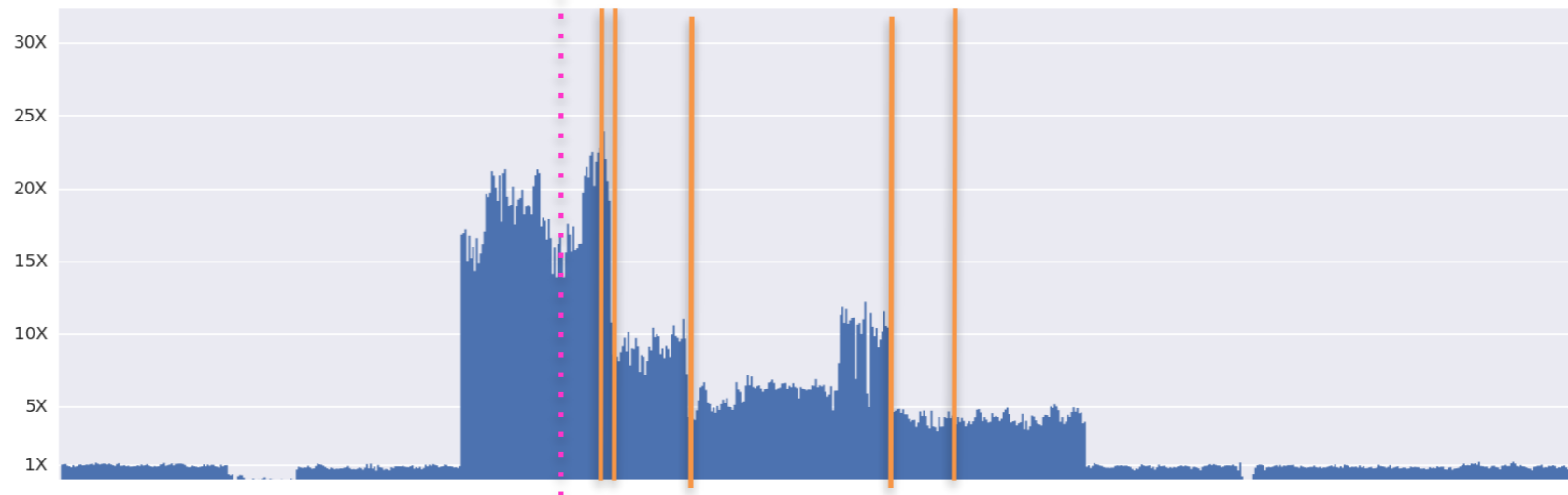
# Her2

PacBio  
73X @ 10kb



# split reads

Illumina  
120X @ 100bp



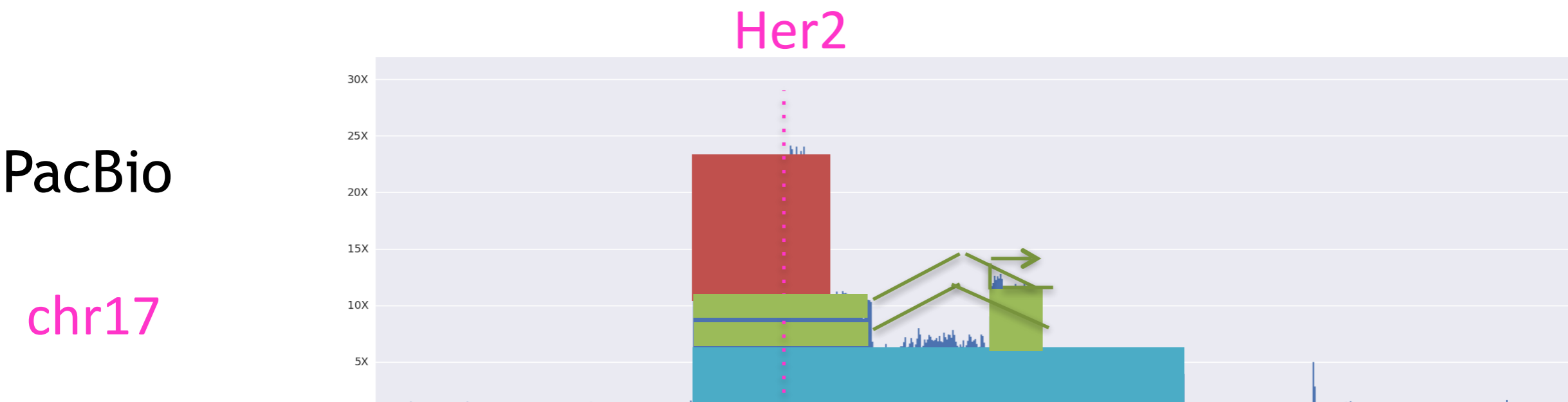
# split reads



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



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



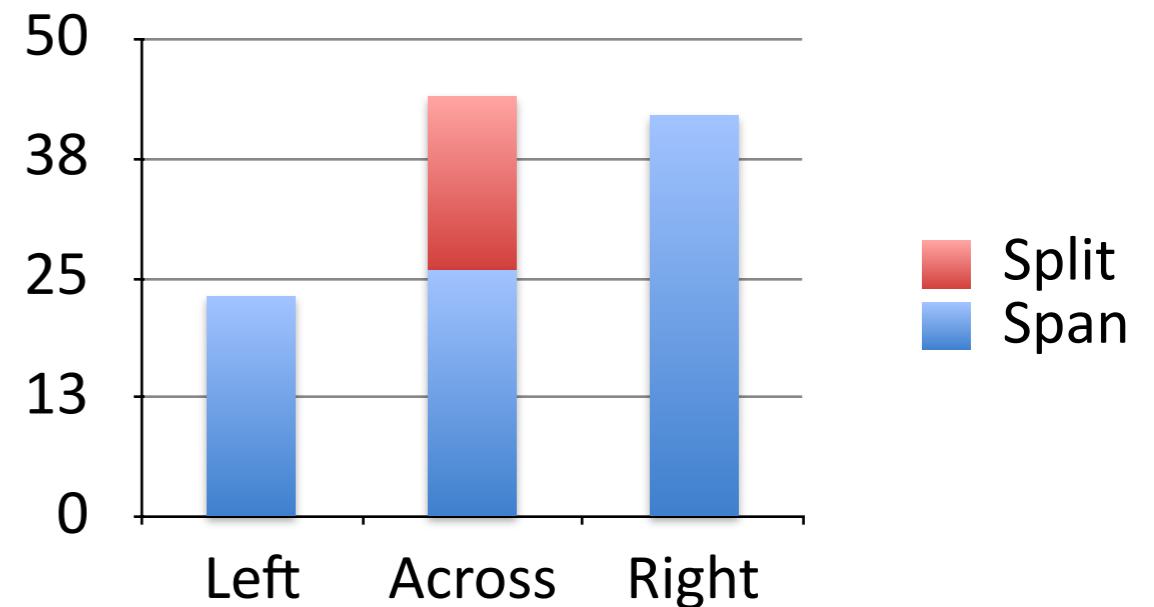
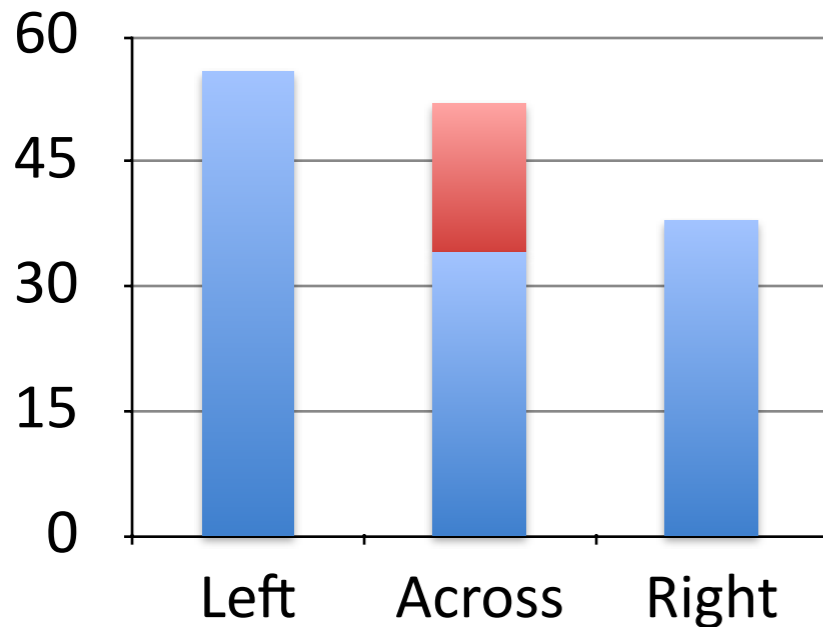
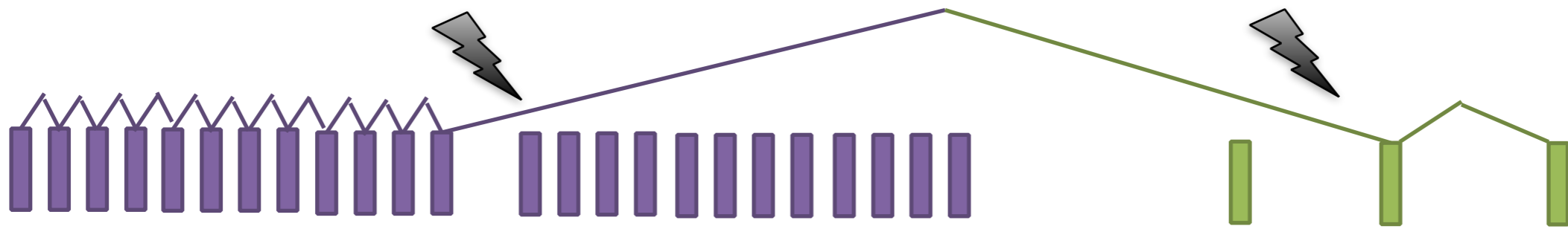
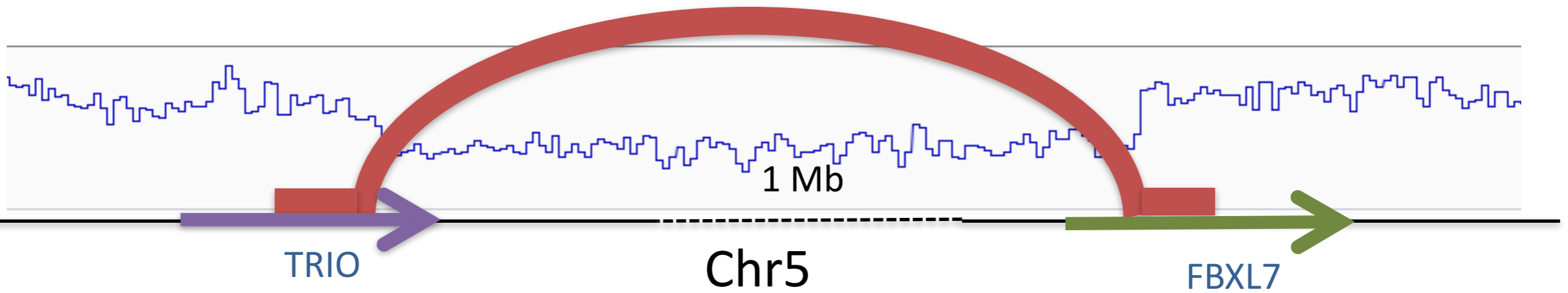
# Combined genome and transcriptome analysis

- 143,532 distinct isoforms
  - 18,186 overlapping groups
- 7 of 9 known gene fusions represented

Known Gene fusions		Confirmed by PacBio DNA?	Confirmed by PacBio Iso-Seq
<i>TATDN1</i>	<i>GSDMB</i>	Yes	Yes
<i>RARA</i>	<i>PKIA</i>	Yes	Yes
<i>ANKHD1</i>	<i>PCDH1</i>	Yes	No
<i>CCDC85C</i>	<i>SETD3</i>	Yes	No
<i>SUMF1</i>	<i>LRRFIP2</i>	Yes	Yes
<i>WDR67 (TBC1D31)</i>	<i>ZNF704</i>	Yes	Yes
<i>DHX35</i>	<i>ITCH</i>	Yes	Yes
<i>NFS1</i>	<i>PREX1</i>	Yes <i>*if allowing for 3 translocations</i>	Yes
<i>CYTH1</i>	<i>EIF3H</i>	Yes <i>*if allowing for 2 translocations</i>	Yes

# TRIO-FBXL7

18 split DNA reads + PCR validation



# PacBio errors are randomly distributed

ATGCTCTTCGATCGATGCTGCTAGCTAGCTAGCTACTAGCTATCCGATCCTACTGACTTACTATGCT

ATGCTGTTTCGATCGATGCTGCTAGCTAGCTAGCTACTAGCTATCCGATCCTACTGACTTACTATGCT

ATGCTCTTCGATCGATGCTGCTCGCTAGCTAGCTACTAGCTATCCGATCCTACTGACTTACTATGCT

ATGCTCTTCGATCGATGCTGCTAGCTAGCTAGCTACTAGCTATCAGATCCTACTGACTTACTATGCT

ATGCTCTTCGATCGATGCTGCTAGCTAGCTAGCTACTAGCTATCGGATCCTACTGACTTACTATGCT

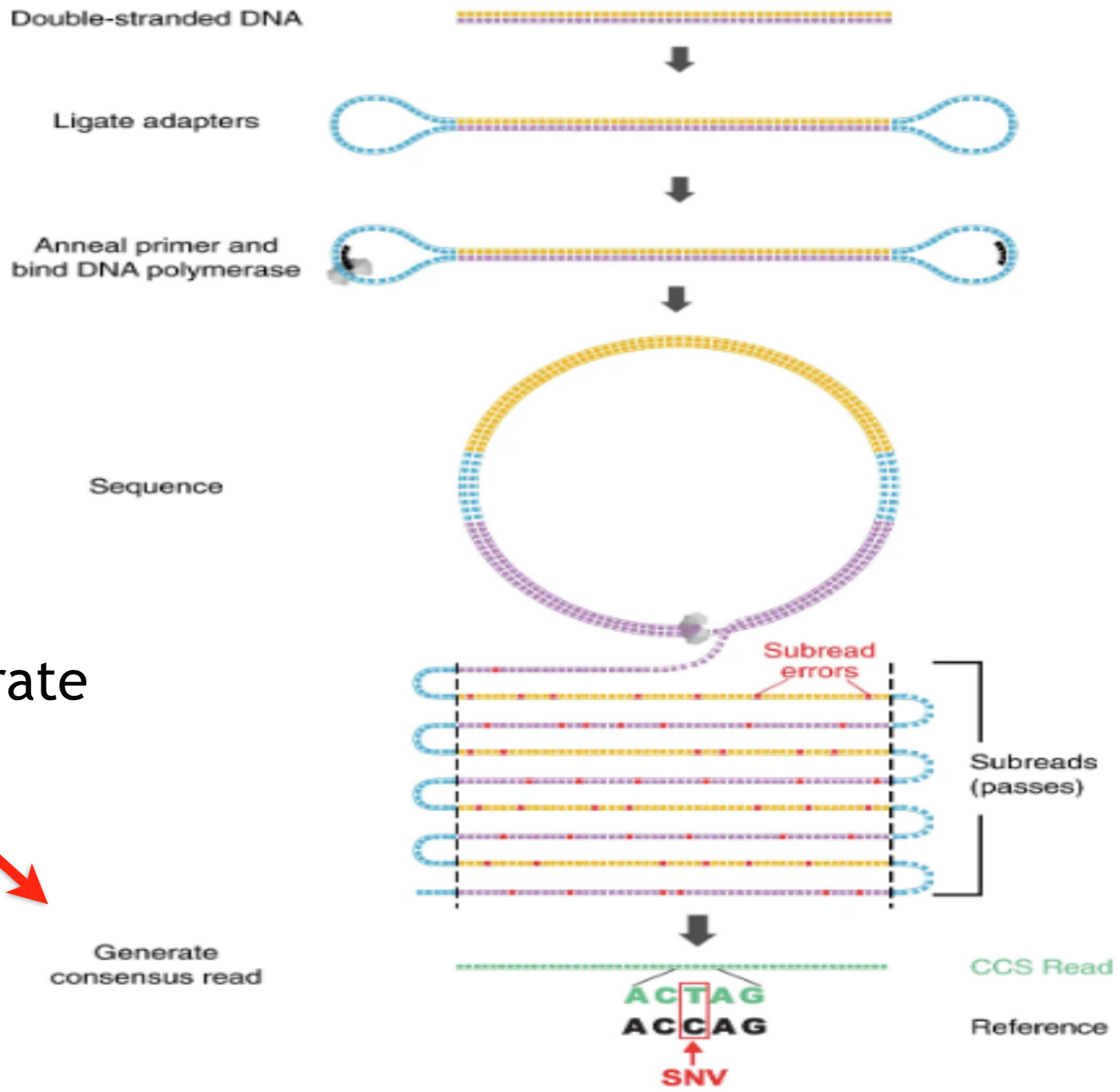
ATGCTCTTCGATCGATGCTGCTAGCTAGCTAGCTACTAGCTATCCGATCCTACTGACTTACTATGGT



ATGCTCTTCGATCGATGCTGCTAGCTAGCTAGCTACTAGCTATCCGATCCTACTGACTTACTATGCT

Enough coverage makes error drop out

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



99.99% Accurate





Oxford

**NANOPORE**

Technologies

# PromethION

48 independent flowcells

500bp/s sequencing speed

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

On site 1D basecalling

>140Gb in CSHL hands

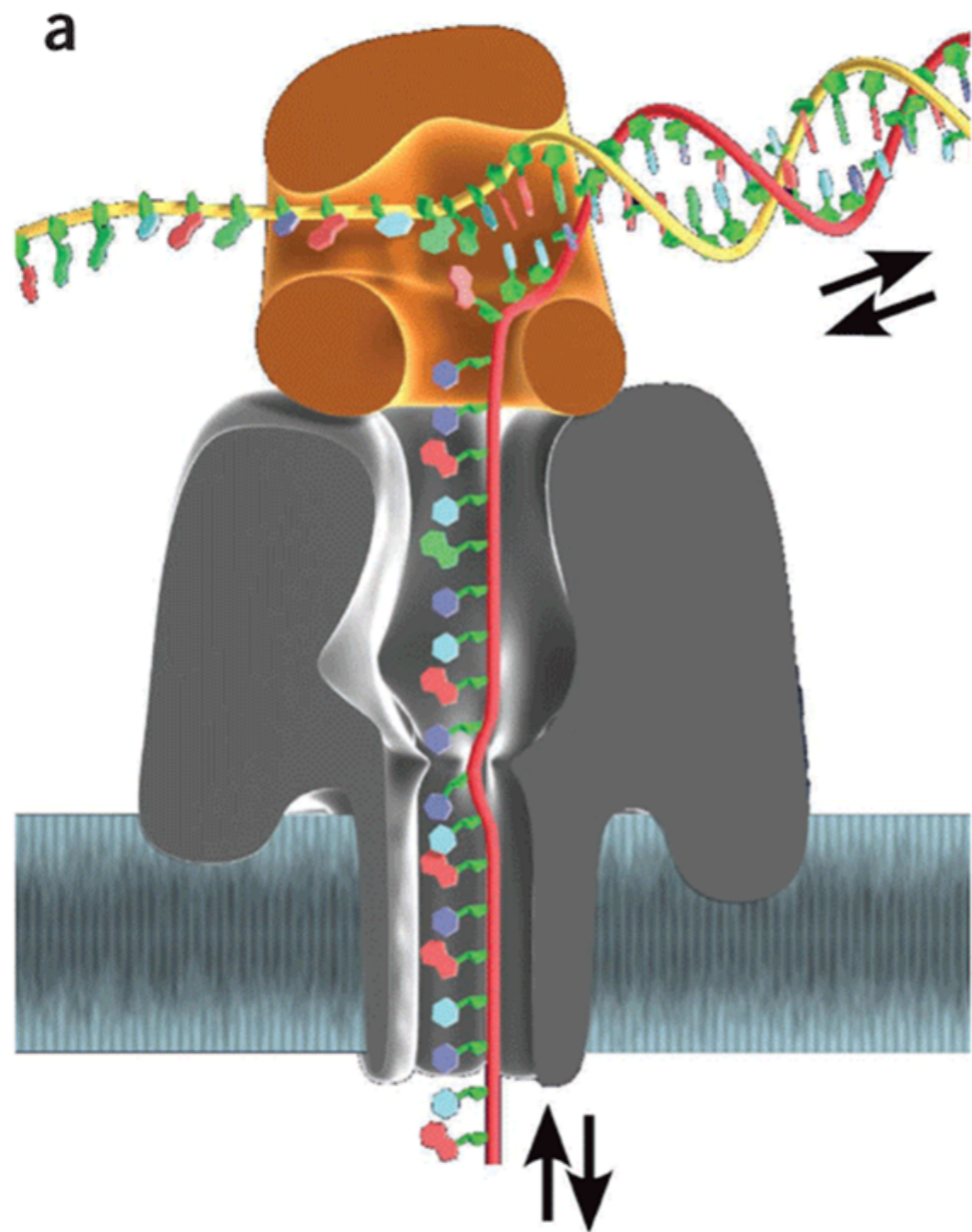
>100M cDNA reads

Up to ~7Tb fully loaded on 60 hours





# 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

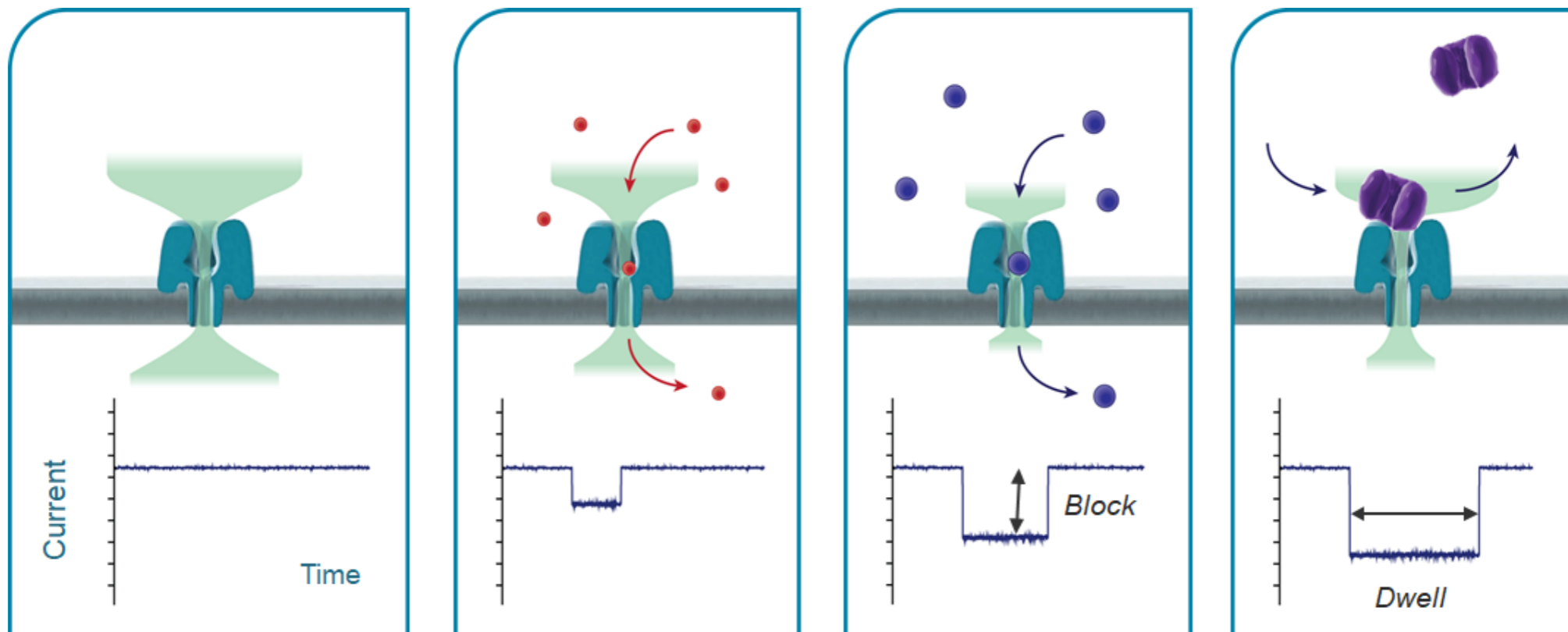
# Nanopore Sensing Summary

Nanopore = 'very small hole'

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

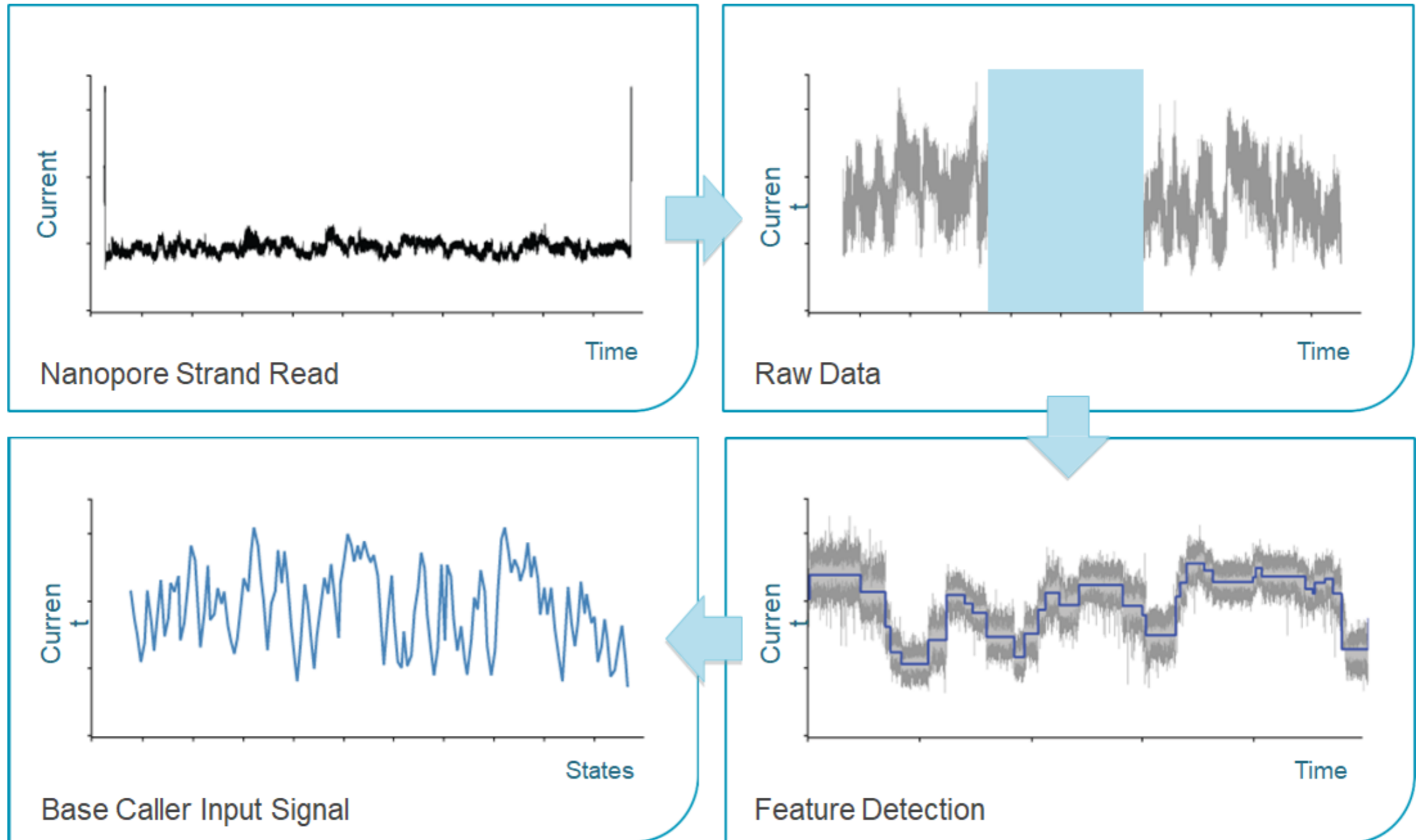
Identify target analyte by the characteristic disruption or block to the electrical current

Block or 'State', Dwell, Noise





# Raw Data and Data Reduction



# Nanopore errors are (mostly) randomly distributed

ATGCTCTTCGATCGATGCTGCTAGCTAGCTAGCTTTTTTTTTTTCCGATCCTACTGACTTACTATGCT

ATGCTGTTTCGATCGATGCTGCTAGCTAGCTAGCTTTTTTT CCGATCCTACTGACTTACTATGCT

ATGCTCTTCGATCGATGCTGCTCGCTAGCTAGCTTTTTTTTTT CCGATCCTACTGACTTACTATGCT

ATGCTCTTCGATCGATGCTGCTAGCTAGCTAGCTTTTTTTTTTTCAGATCCTACTGACTTACTATGCT

ATGCTCTTCGATCGATGCTGCTAGCTAGCTAGCTTTTTTTT CCGATCCTACTGACTTACTATGCT

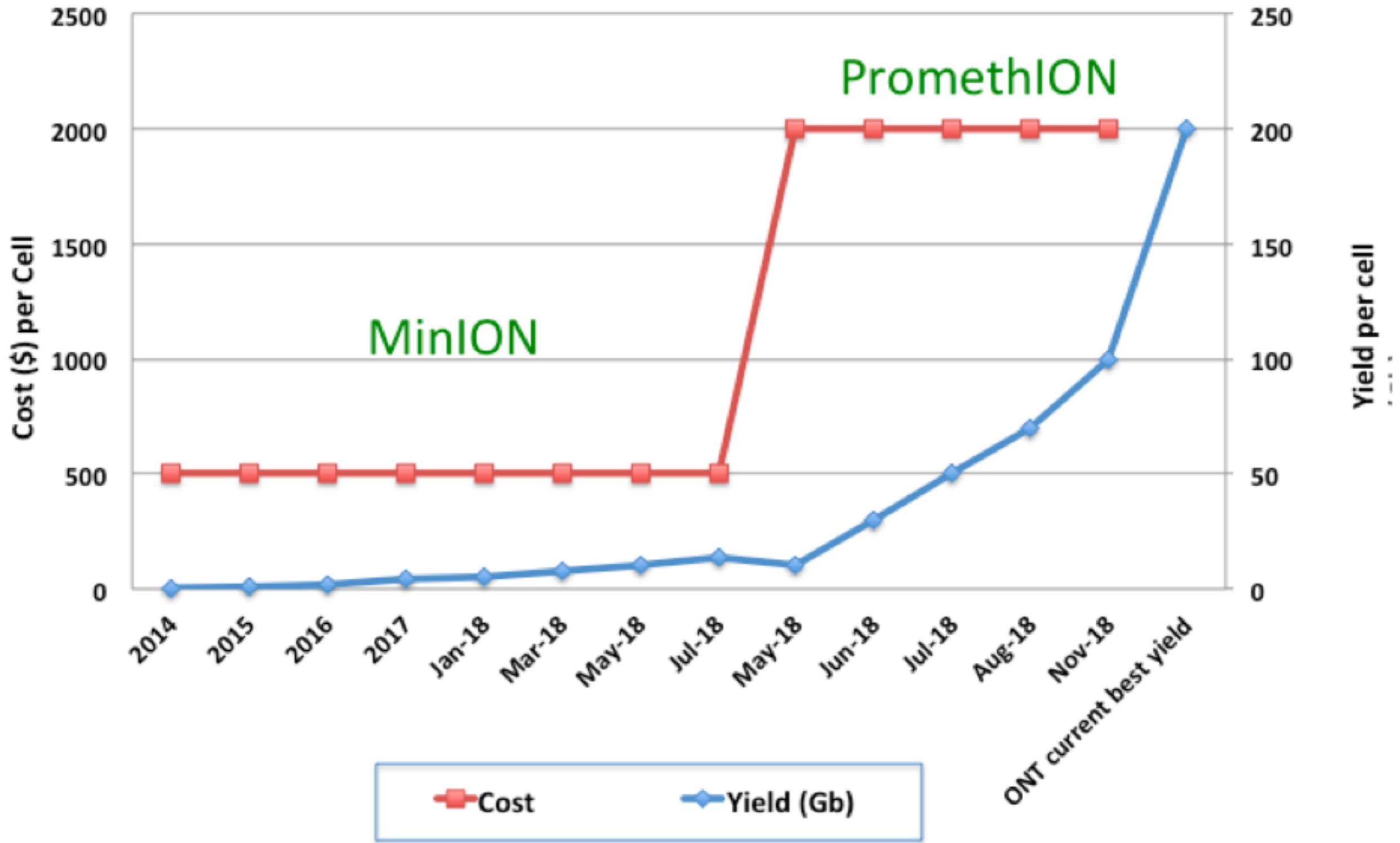
ATGCTCTTCGATCGATGCTGCTAGCTAGCTAGCTTTTTTTTT CCGATCCTACTGACTTACTATGGT



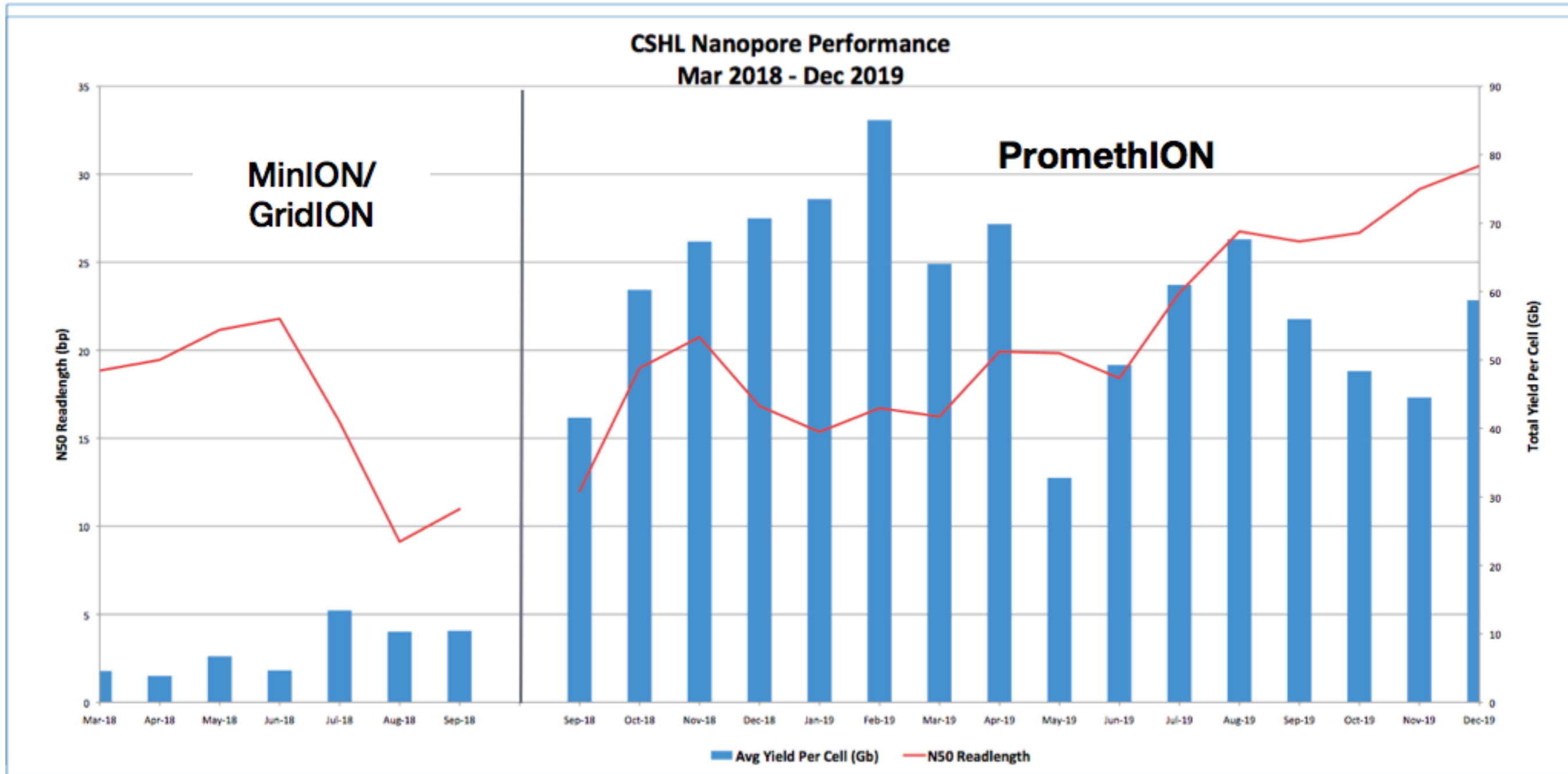
ATGCTCTTCGATCGATGCTGCTAGCTAGCTAGCTTTTTTTT CCGATCCTACTGACTTACTATGCT

Enough coverage makes error (mostly) drop out

# Oxford Nanopore Cost vs Yield

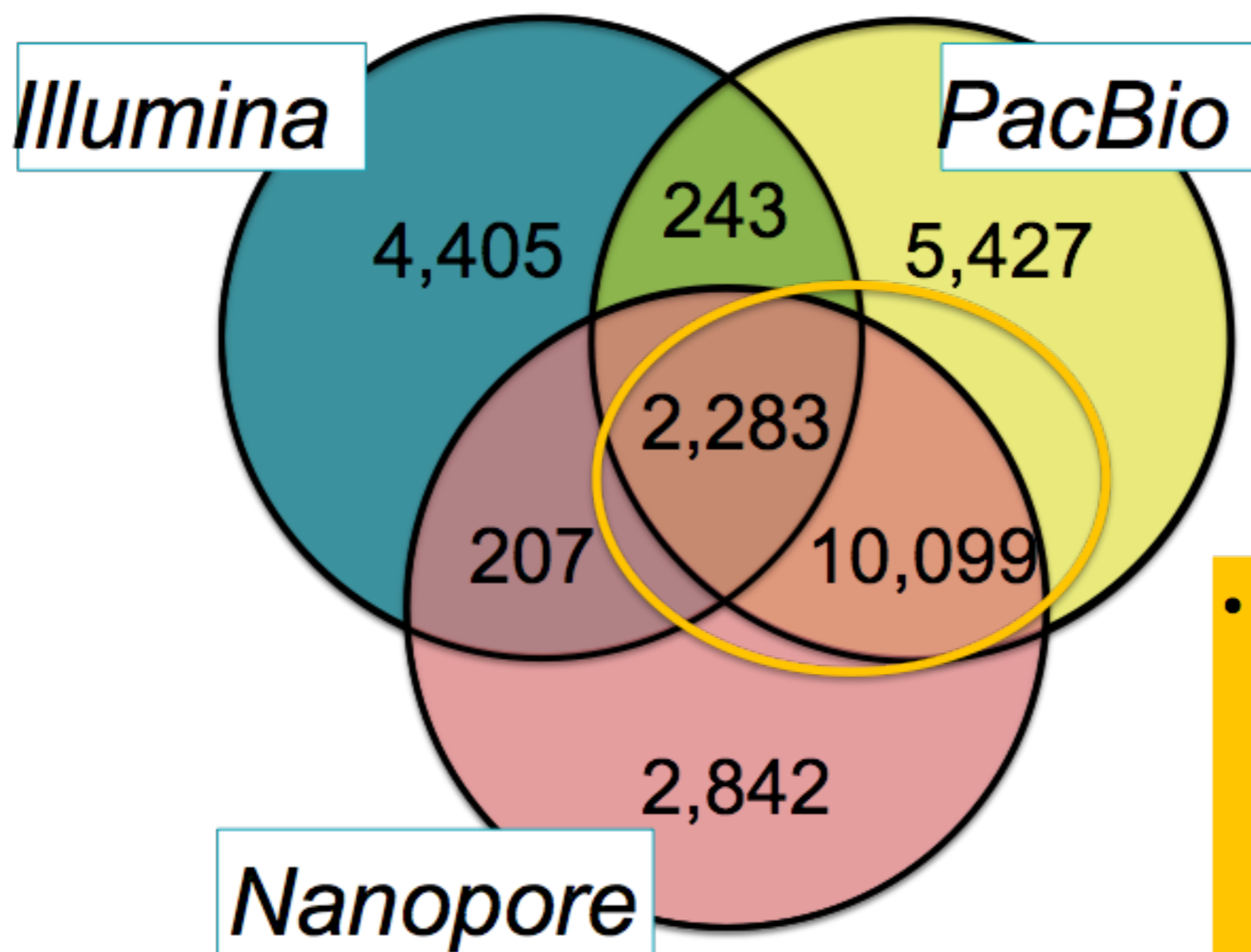


# Oxford Nanopore Sequencing at CSHL



PromethION yields have declined as we have targeted longer fragments, but further optimization to increase yield is underway

# Structural Variant Comparison of SKBR3

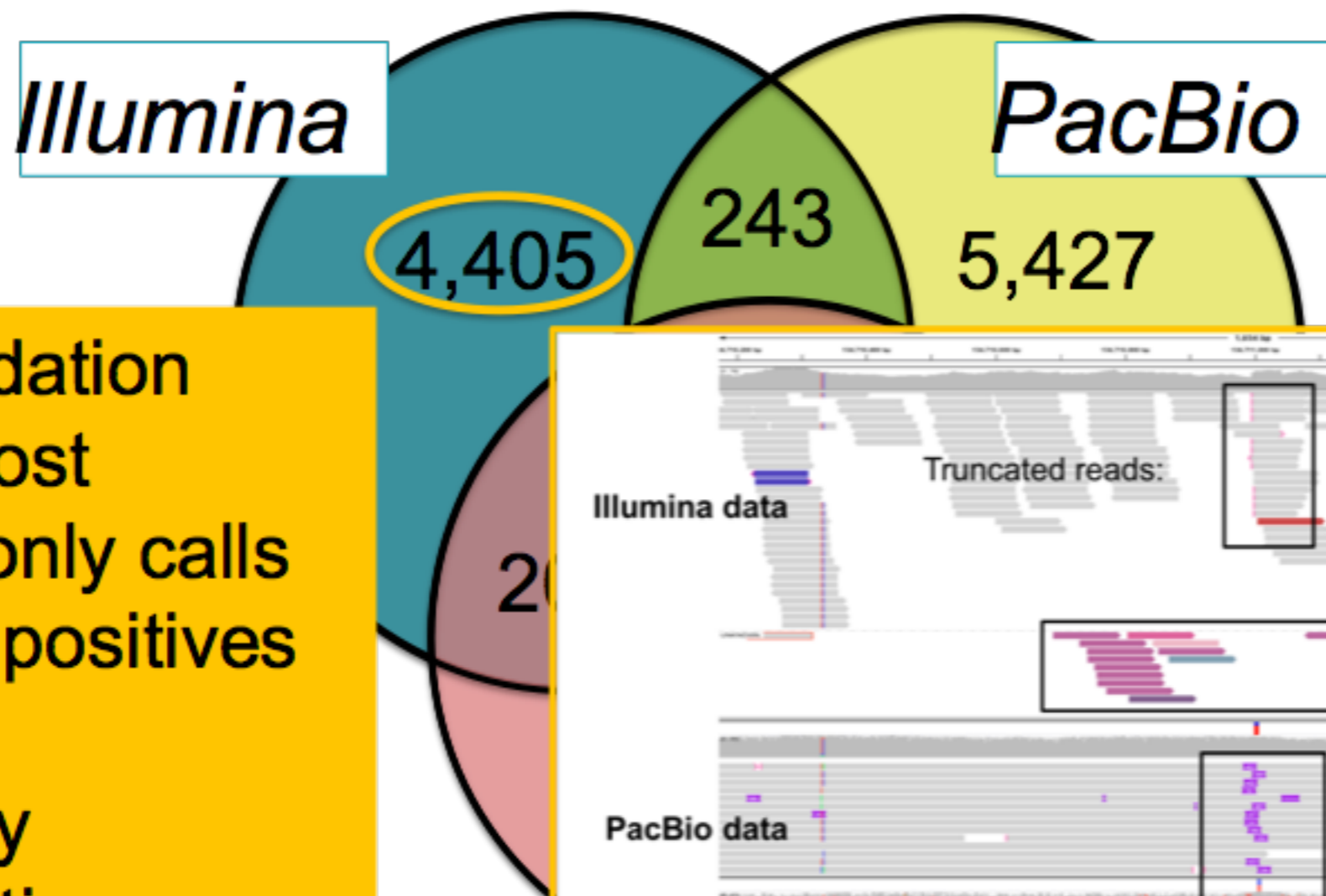


- Strong concordance between long read platforms
- Substantially more variants than detected by short reads

(Hicks et al, 2006,

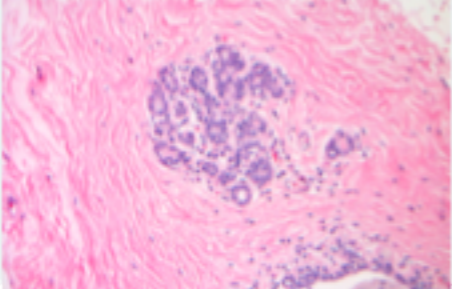

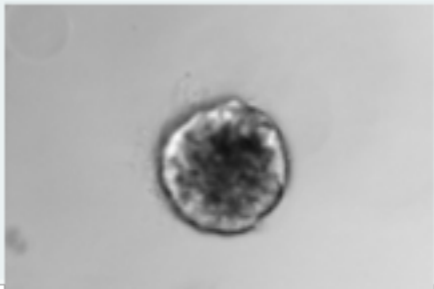
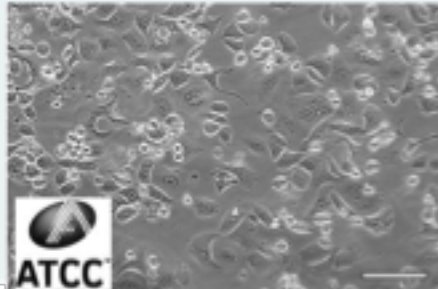


# Structural Variant Comparison of SKBR3

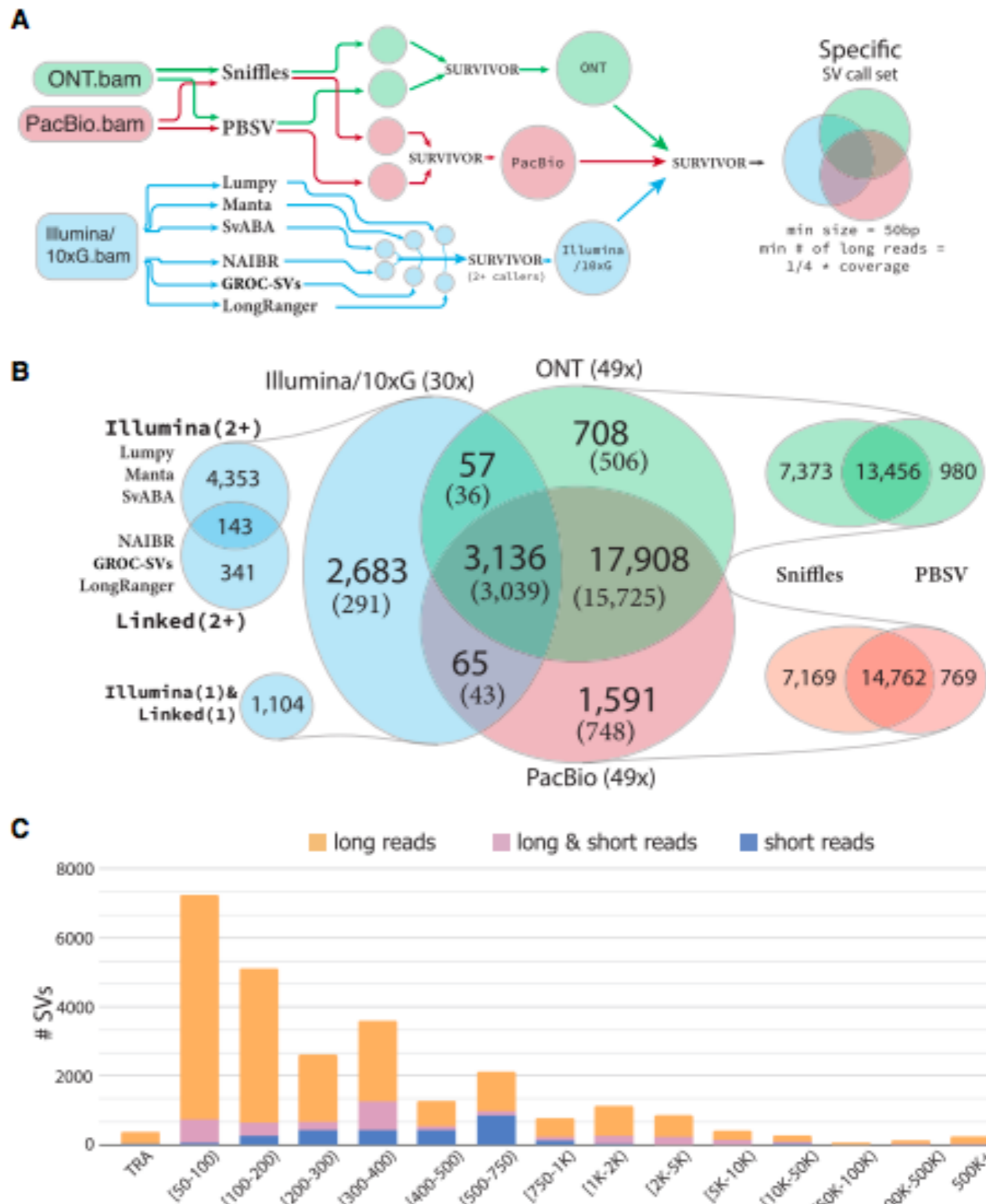


- PCR validation shows most Illumina-only calls are false positives
- Especially translocations or inversions caused by smaller insertions or deletions

# Multi-omics Long Read Analysis of Cancer

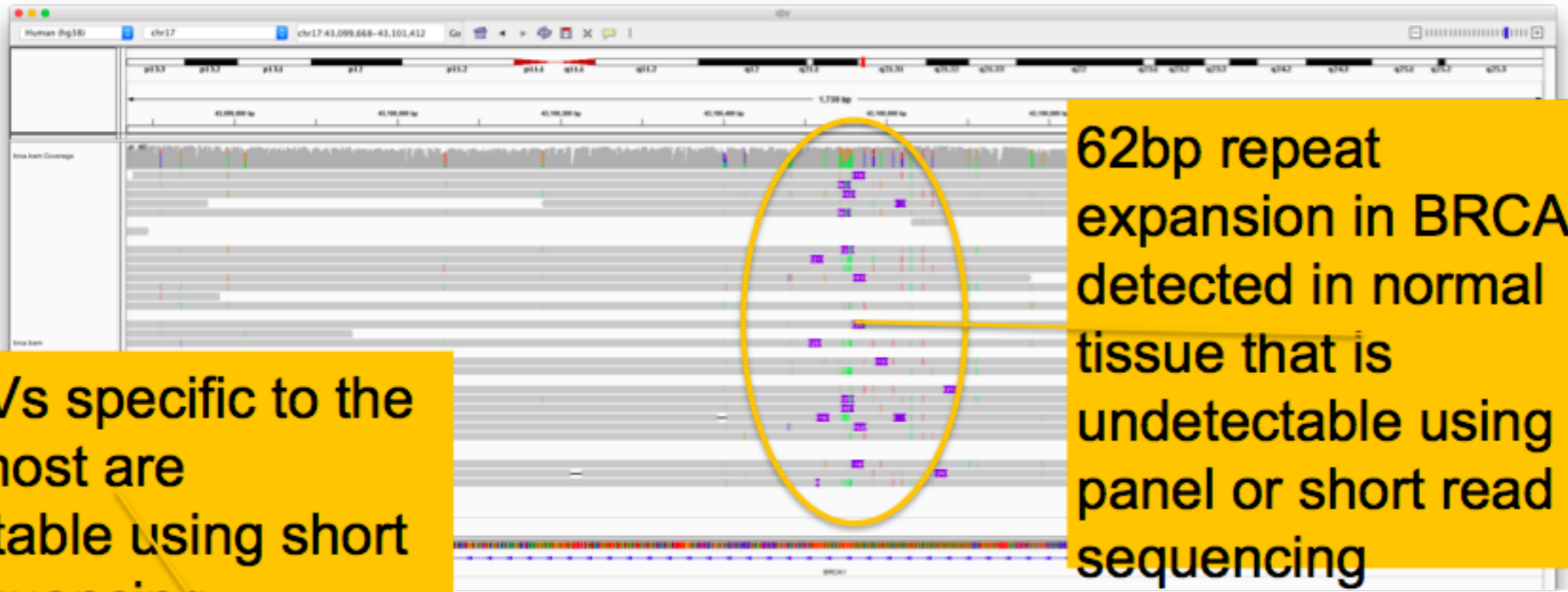
	Normal Breast Tissue	Normal Breast Organoid	Tumor Breast Organoid	SK-BR-3 Breast Cancer Cell Line
<b>Oxford Nanopore WGS</b>	Y	N	Y	Y
<b>PacBio WGS</b>	N	N	N	Y
<b>ONT Methylation</b>	Y	N	Y	Y
<b>Illumina Methylation</b>	Y	N	Y	Y
<b>Illumina RNA-seq</b>	N	Y	Y	Y
<b>PacBio RNA-seq</b>	N	N	N	Y
<b>Pathology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

# Cross Platform SV comparison for sample 51



From Aganezov 2020

# Preliminary Structural Variations Analysis



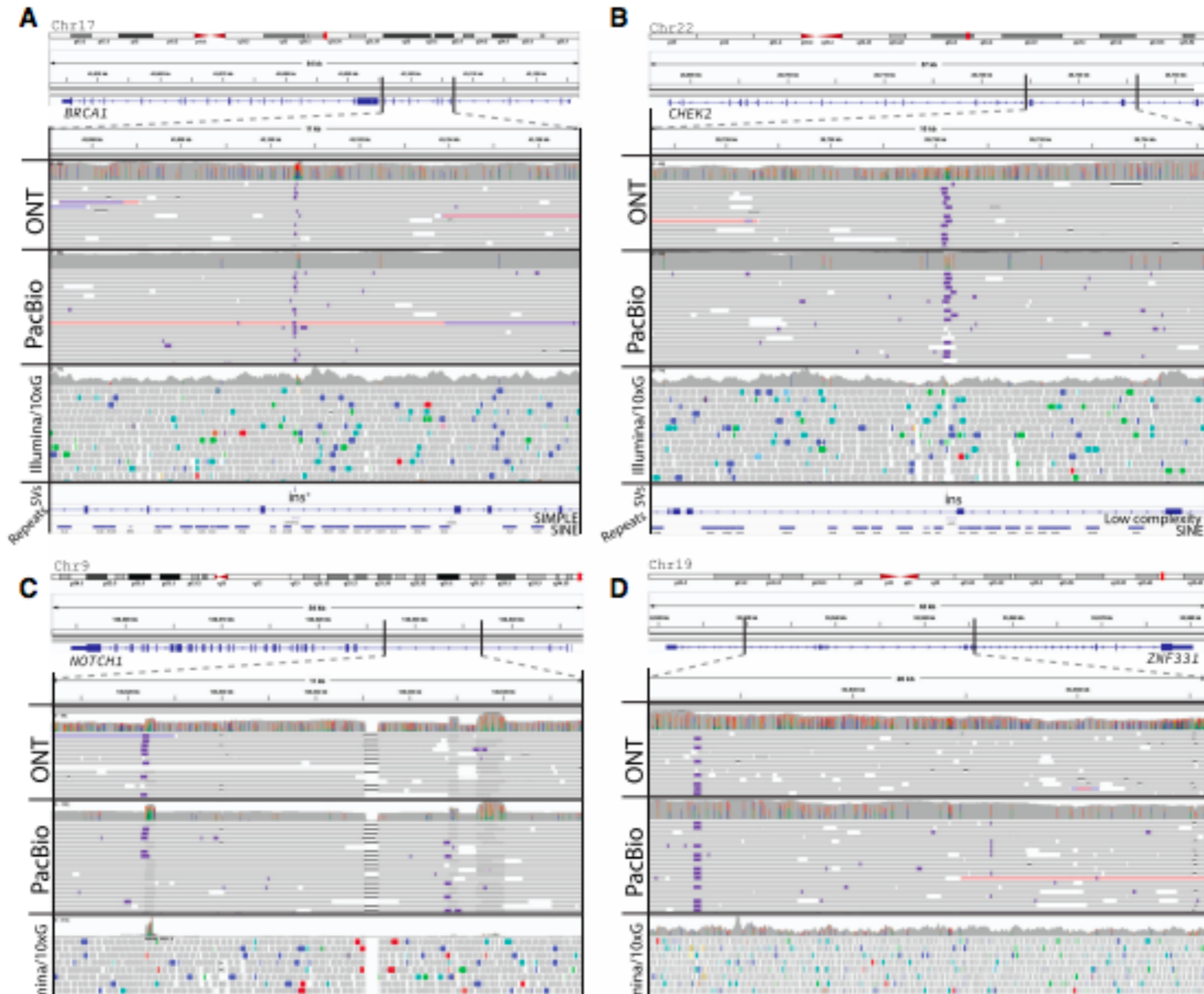
3,662 SVs specific to the tumor, most are undetectable using short read sequencing

	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.



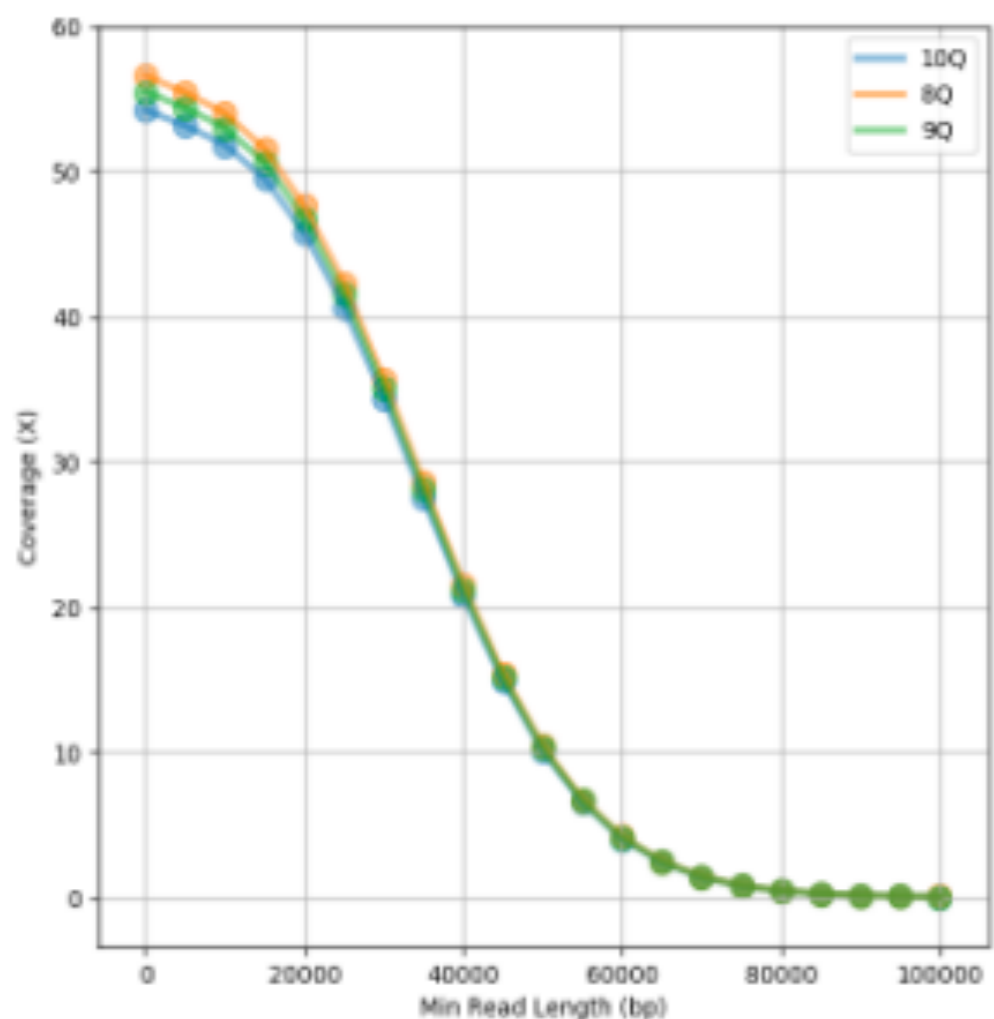


# Living Fossils Oxford Nanopore Sequencing

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 ula</i>	2.25	2.2	Oxford Nanopore*

# Collaboration with Srividya Ramakrishnan and Mike Schatz

## Wollemia Nanopore Assembly with wtdbg2



### Assembled reads >Q10 & >40kb

- Required 10 days with 1TB RAM
- Assembly with 30kbp reads produced worse assembly

### Assembly Stats:

- Total Span: 15,659,209,344 bp
- Contig N50: 312,370 bp
- Max contig len: 7,090,464bp
- Number contigs: 223,812

### Comparisons:

- 22 Gbp loblolly pine: contig N50=25kbp
- <https://academic.oup.com/gigascience/article/6/1/giw016/2865215>
- 15.3 Gbp hexaploid wheat: contig N50=232kbp
- <http://academic.oup.com/gigascience/article/6/11/gix097/4561661>

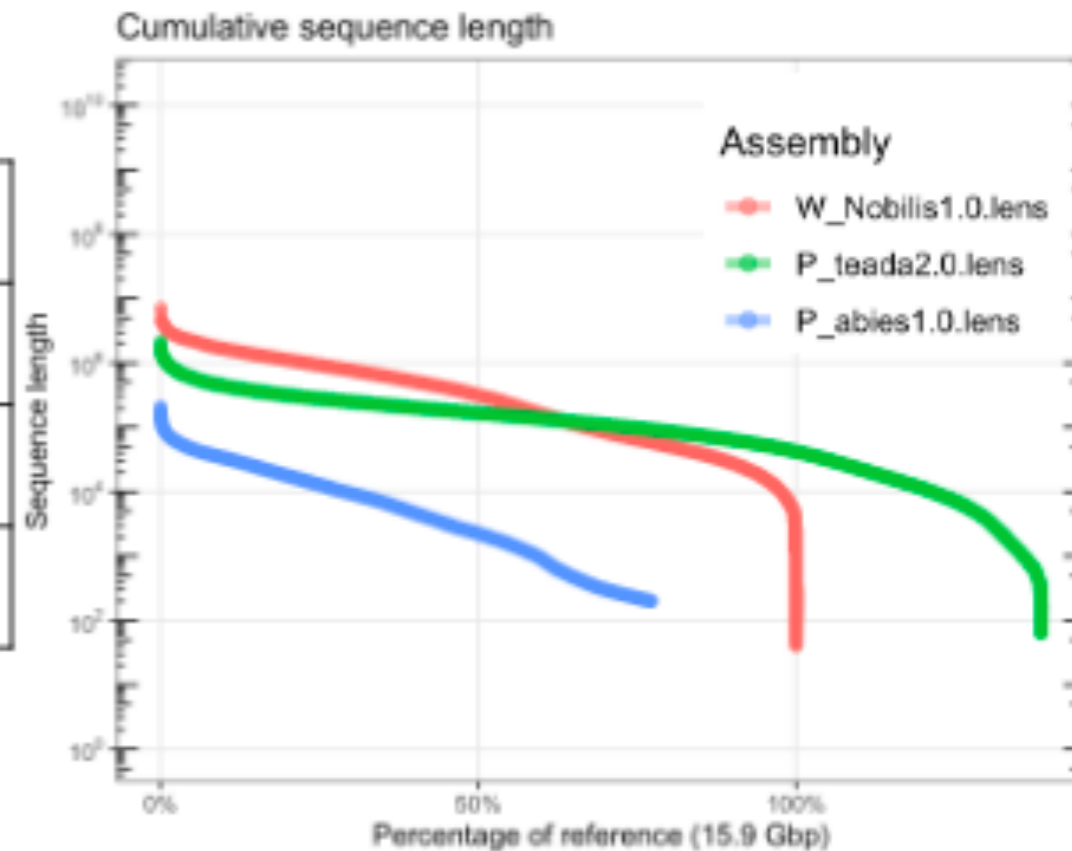
# Assembly comparison to large plant genomes

## Wollemia Polished Assembly Stats

### Comparison to Loblolly Pine and Norwegian Spruce genomes

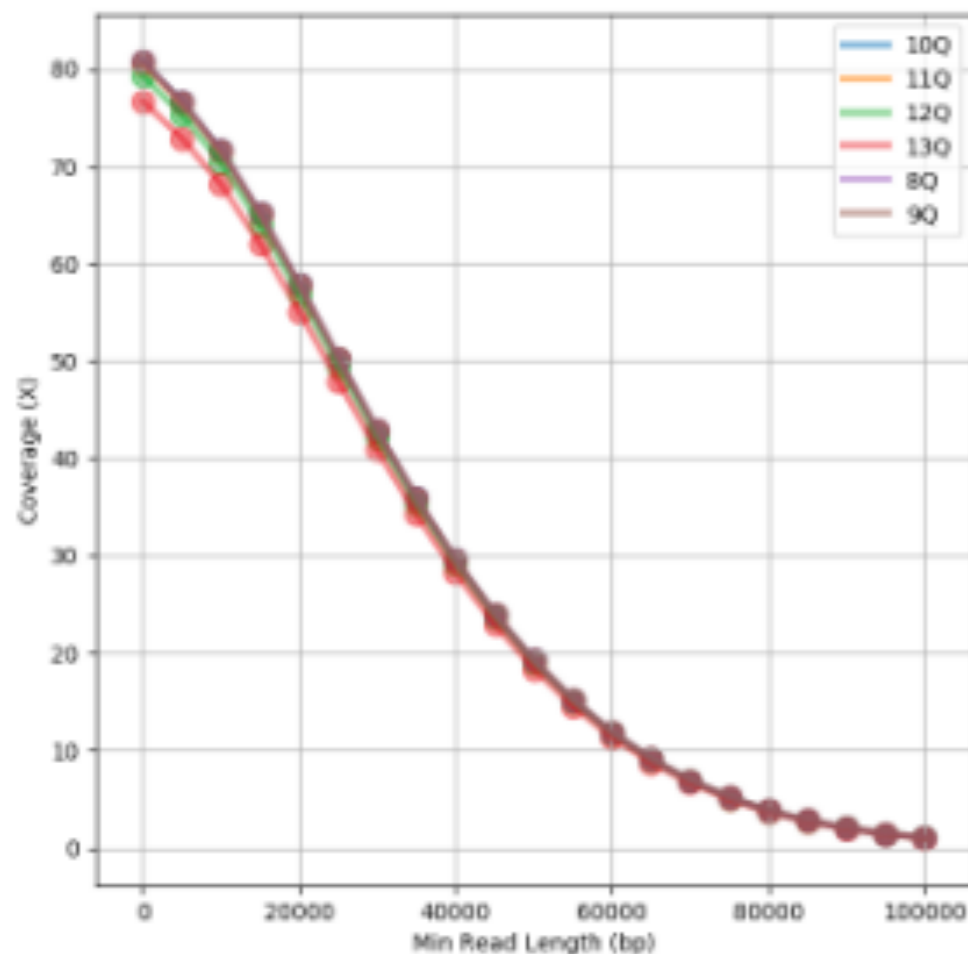
Assembly	Total bps	# Contigs	N50	Mean	Min	Max
W.Nobilis v1.0 (wtdbg2-racon1-medaka1)	15.94 Gbp	243,696	314.09 Kbp	190.15 5 Kbp	41 bp	7.21 Mbp
P.taeda v2.0	22.10 Gbp	1,755,249	110.55 Kbp	43.57 Kbp	64 bp	2.14 Mbp
P.abies v1.0	12.30 Gbp	10253694	5.2 kbp	3810.8	200 bp	208.09 Kbp

### Assembly Contiguity



# Largest genome of the Living Fossils project - estimated 22Gb genome

## Araucaria Nanopore Assembly with wtdbg2



### Assembled reads >Q12 & >45kb

- Required 1 month with about 1.6 TB RAM

### Assembly Stats:

- Total Span: 32,168,661,985 bp
- Contig N50: 126,834 bp
- Max contig len: 2,932,577 bp
- Number contigs: 561,509

### Comparisons:

- 22 Gbp loblolly pine: contig N50=25kbp
- <https://academic.oup.com/gigascience/article/6/1/giw016/2865215>
- 15.3 Gbp hexaploid wheat: contig N50=232kbp
- <http://academic.oup.com/gigascience/article/6/11/gix097/4561661>

# 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  
Fix genomics



# 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

Medhat Helmy



## Karen Kostroff

## Funding

MaizeCODE consortium  
Living Fossils consortium

AMNH  
Nancy Simmons  
Sara Oppenheim

NCI  
NSF  
NHGRI  
Northwell  
Health

Thank you!