Genome Sequencing & Assembly

Michael Schatz

March 31, 2015

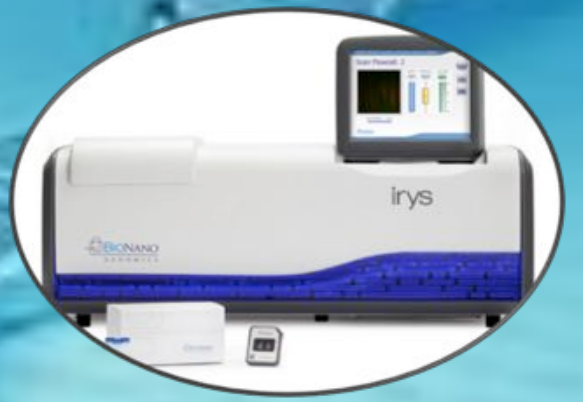
FDA

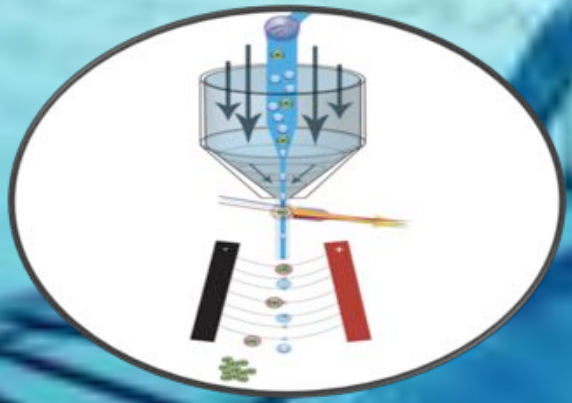
Genomics Arsenal in the year 2015

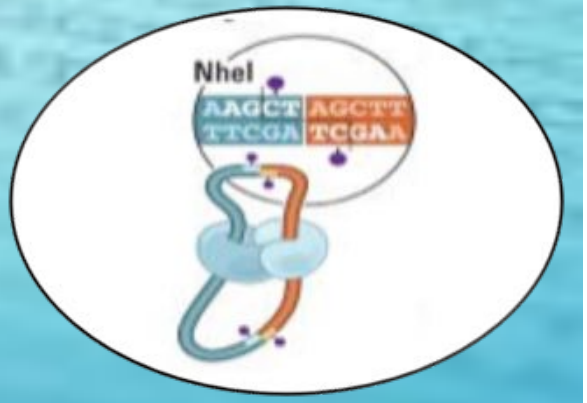
Sample Preparation

Sequencing

Chromosome Mapping

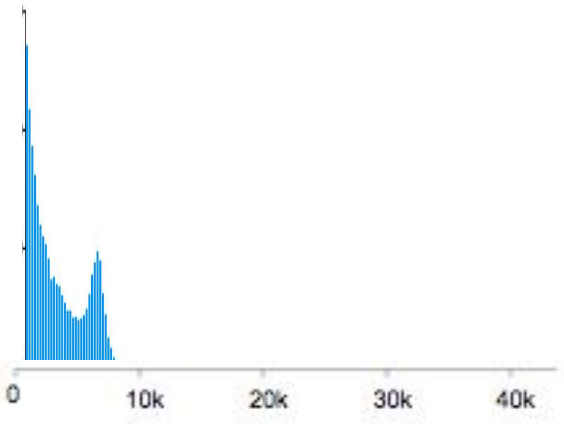






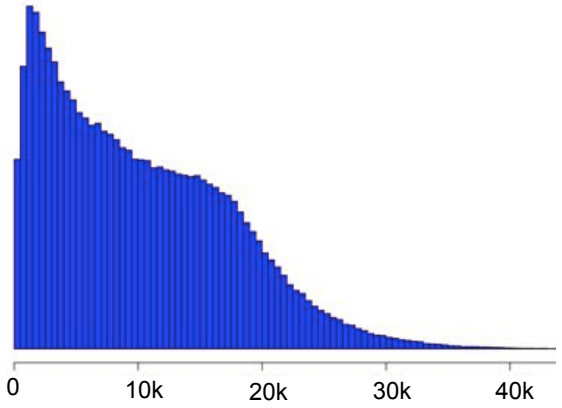
Long Read Sequencing Technology

Moleculo



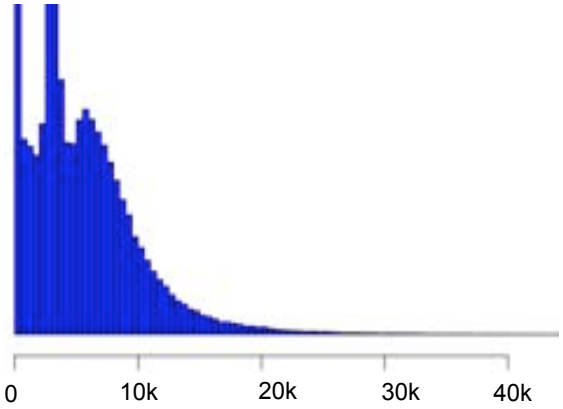
(Voskoboynik et al. 2013)

PacBio RS II



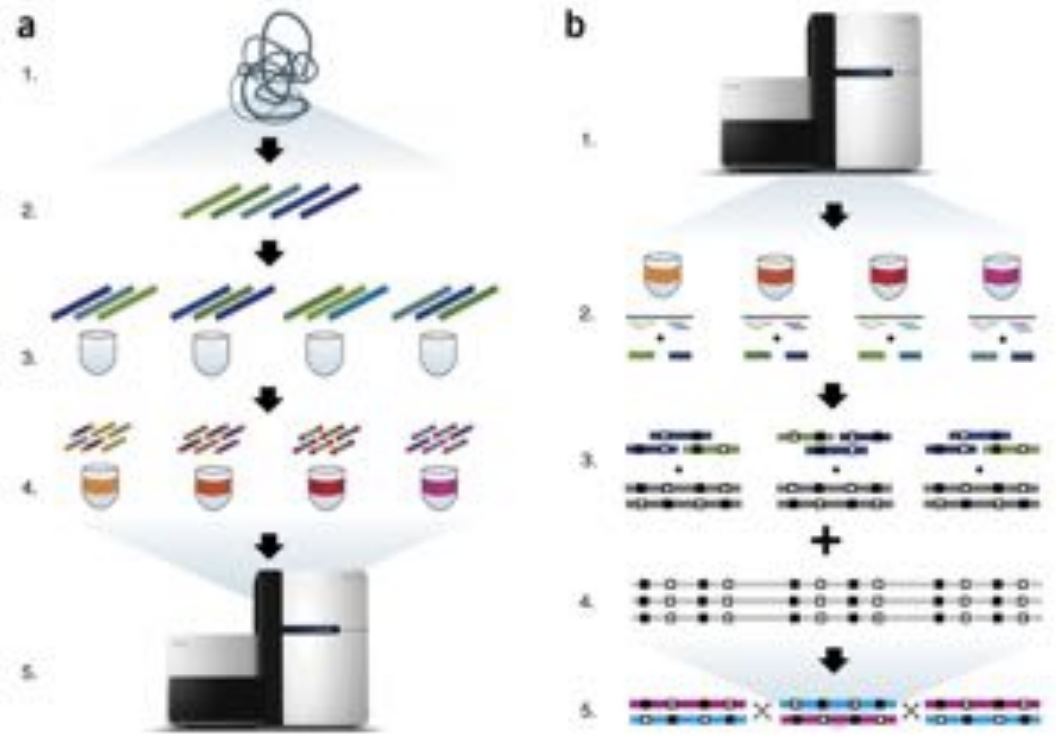
CSHL/PacBio

Oxford Nanopore



CSHL/ONT

Moleculo Sequencing

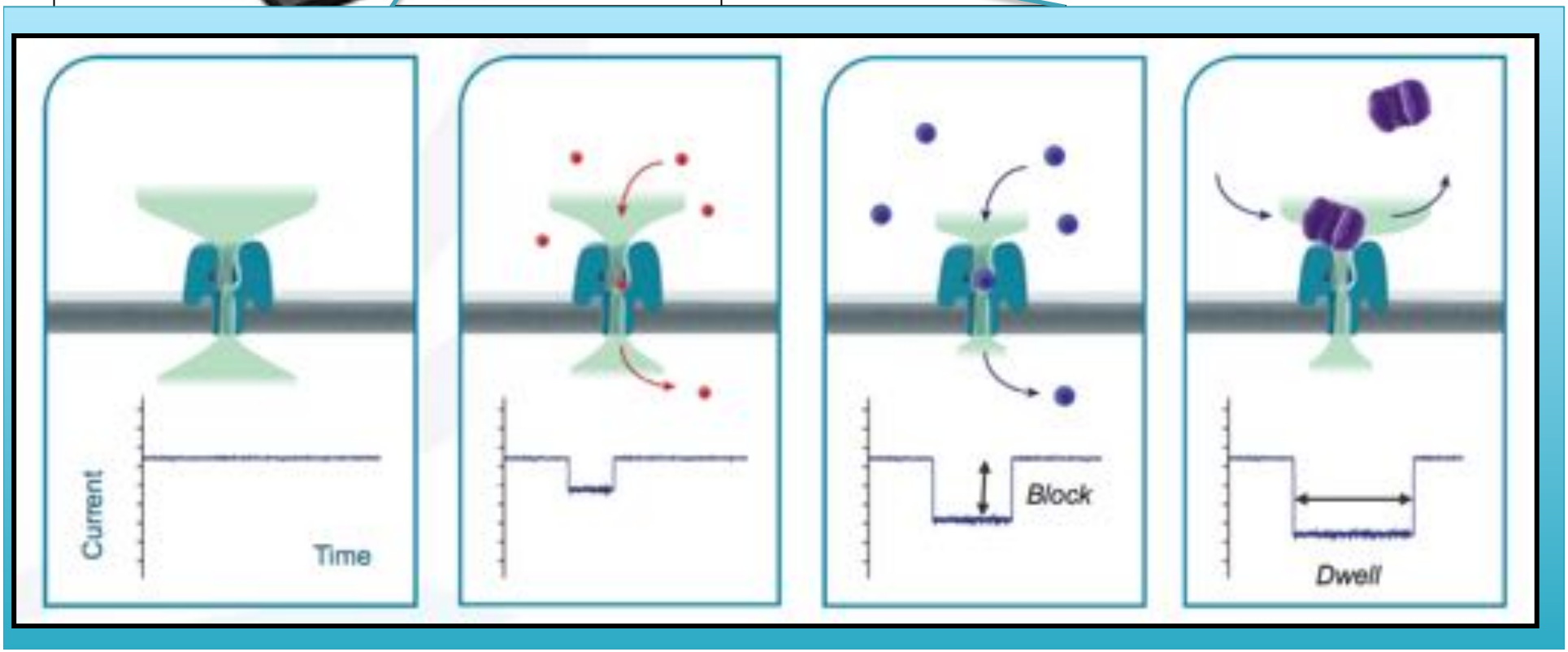


Clever library preparation technique to turn a short read sequencer into a quazi-long read sequencer

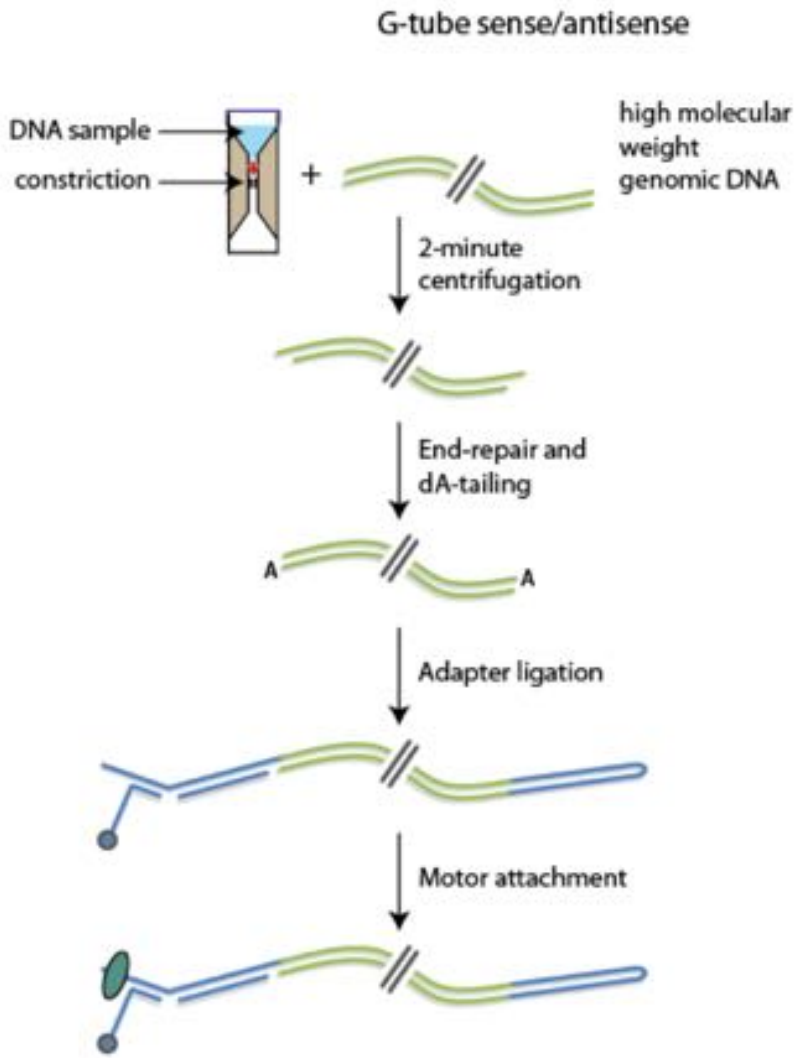
- Very high quality reads, excellent data for phasing
- Restricted to ~10kbp max read lengths
- Excited for future advances, 10X genomics

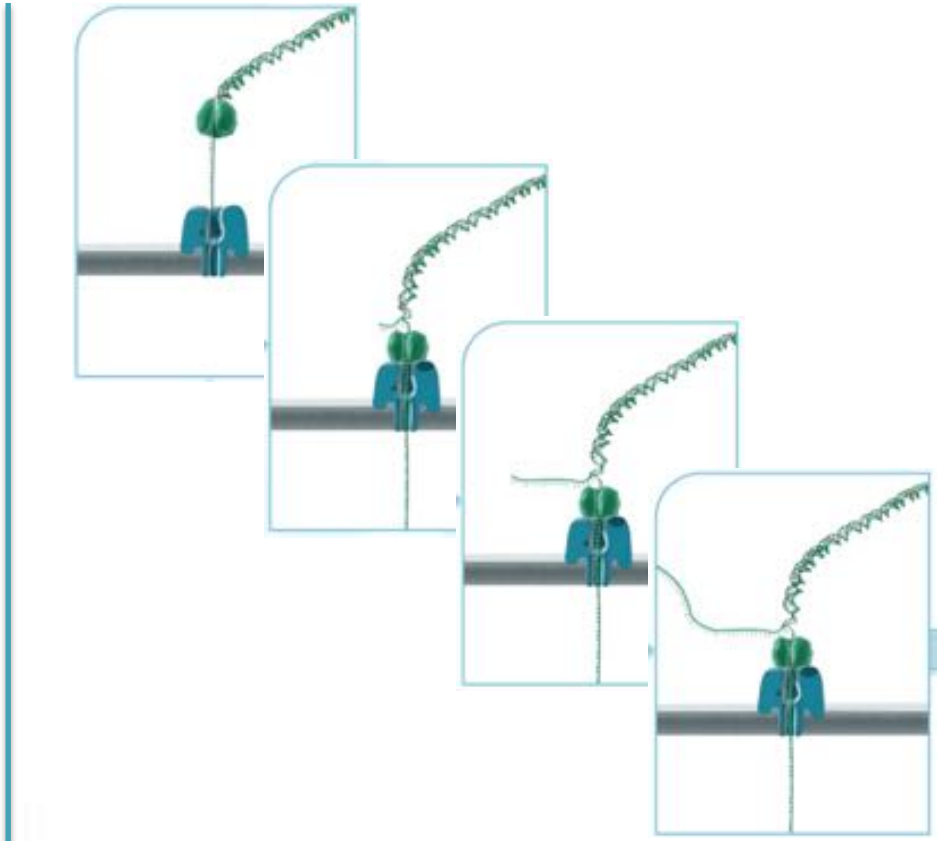
Oxford Nanopore MinION

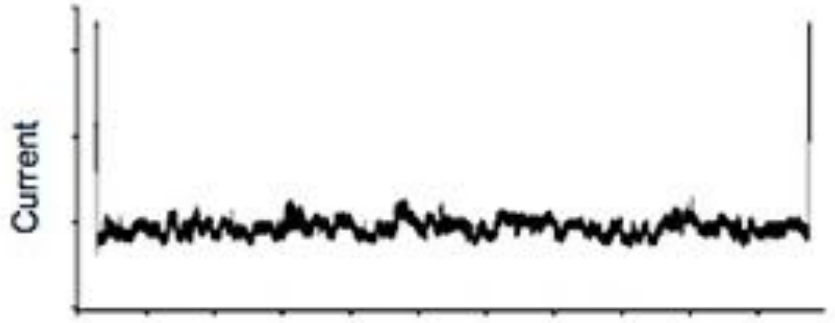
- Thumb drive sized sequencer powered over USB
- Capacity for 512 reads at once
- Senses DNA by measuring changes to ion flow



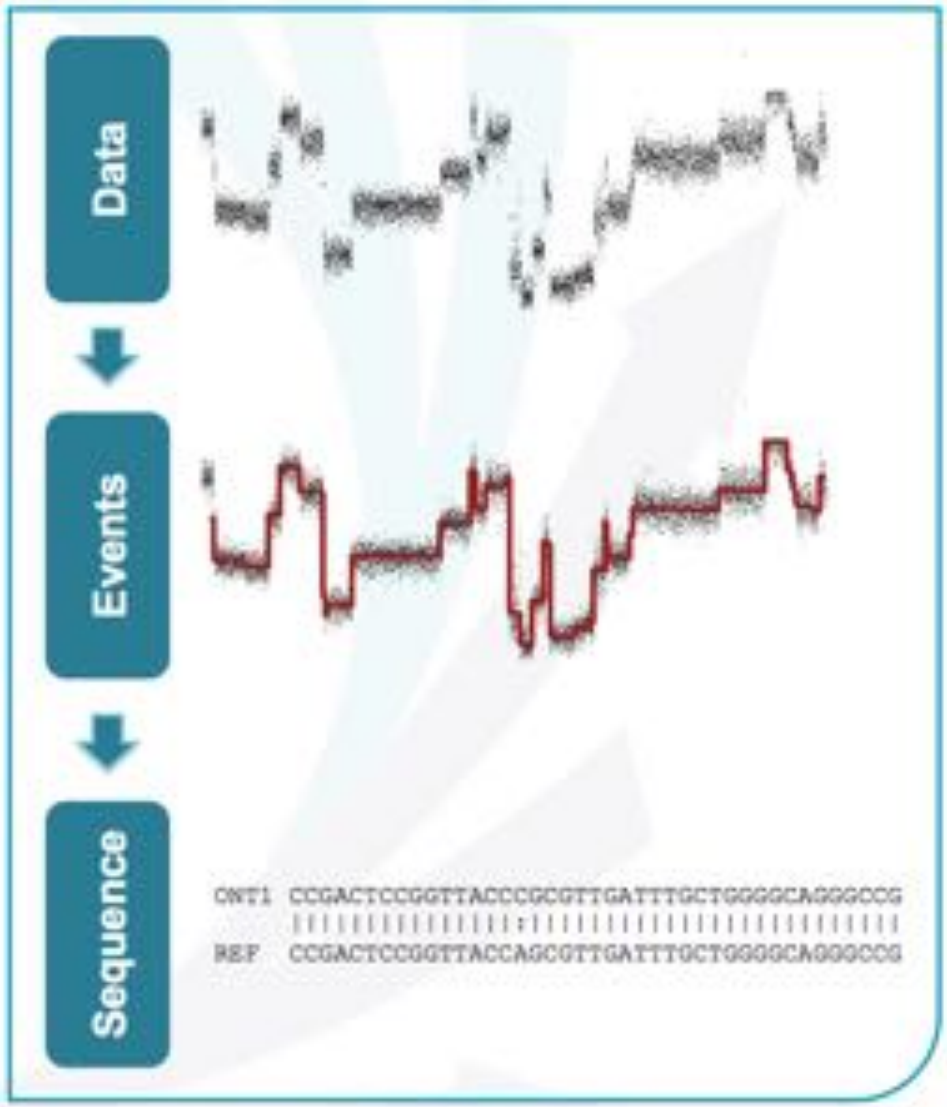
Nanopore Sequencing







Nanopore Basecalling

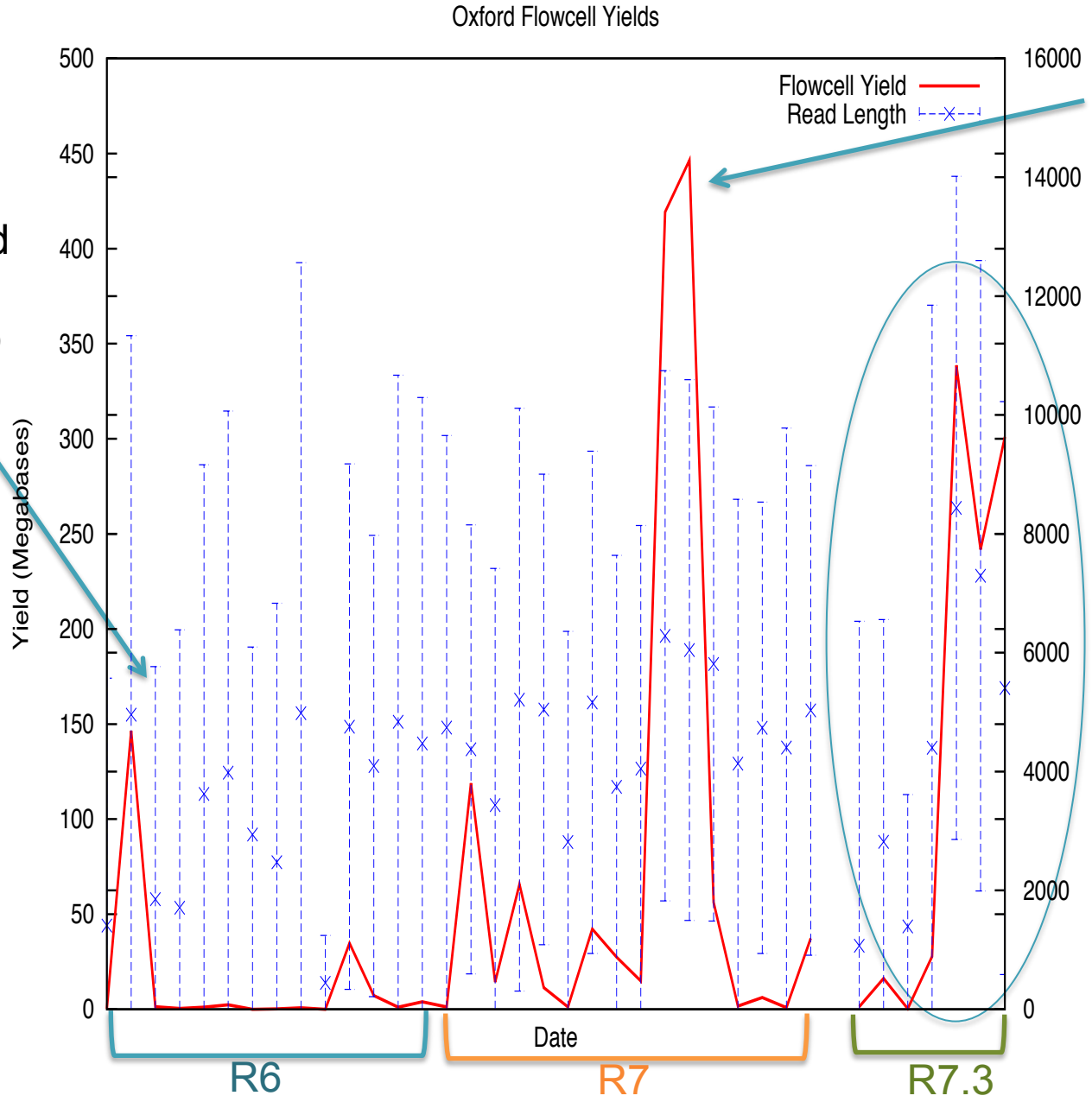


- Hidden Markov model
- Only four options per transition
- Pore type = distinct kmer length

- Form probabilistic path through measured states currents and transitions
  - e.g. Viterbi algorithm

Basecalling currently performed at Amazon with frequent updates to algorithm

Our Data - Yeast W303

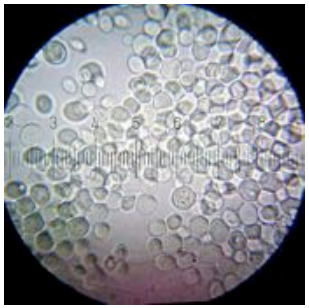


Mean Read Length: ~6kb (10kb shear)

Best: 446Mb

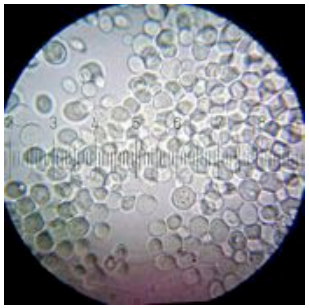
New Flowcells 7k Avg Read Len

Nanopore Readlengths

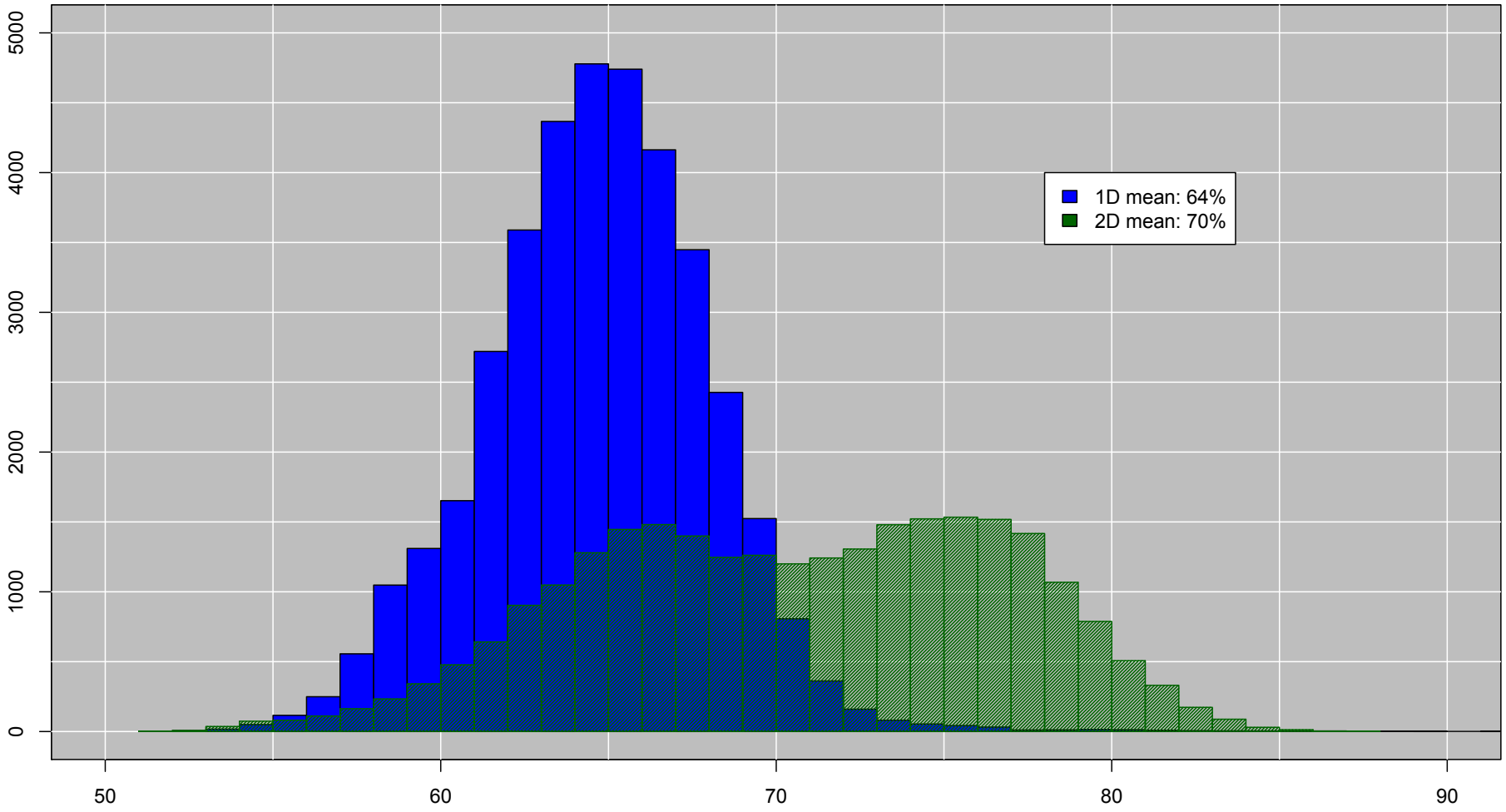


Oxford Nanopore Sequencing at CSHL
30 runs, 267k reads, 122x total coverage
Between 11 and 73k reads per run!
Mean flow cell: 50 Mbp in 2 days
Max flow cell: 446Mbp in 2 days

Nanopore Accuracy



Alignment Quality (BLASTN)
Of reads that align, average ~64% identity
"2D base-calling" improves to ~70% identity



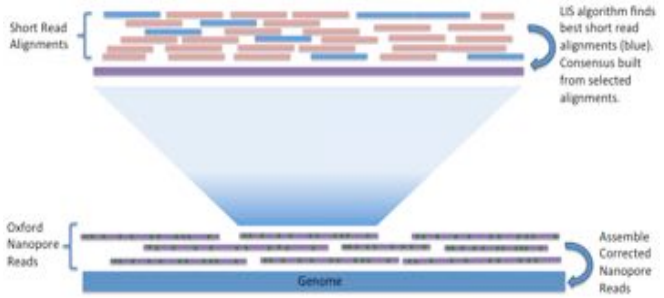
NanoCorr: Nanopore-Illumina Hybrid Error Correction

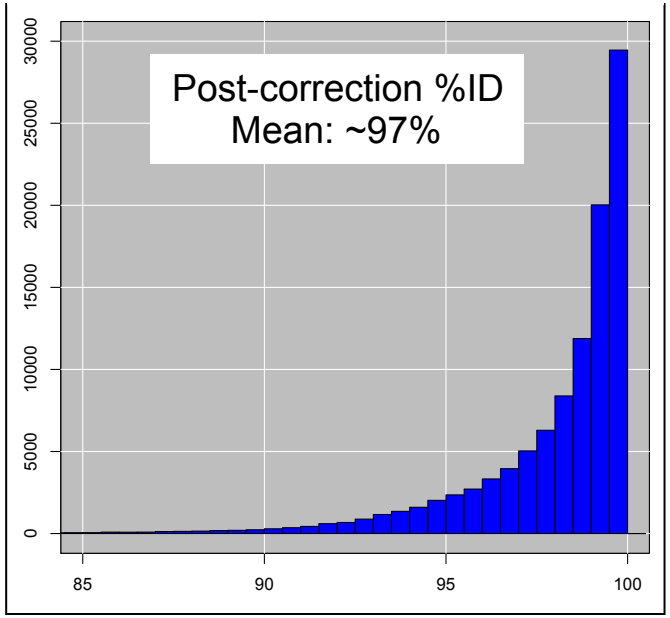
https://github.com/jgurtowski/nanocorr

1. BLAST Miseq reads to all raw Oxford Nanopore reads

2. Select non-repetitive alignments
   - First pass scans to remove "contained" alignments
   - Second pass uses Dynamic Programming (LIS) to select set of high-identity alignments with minimal overlaps

3. Compute consensus of each Oxford Nanopore read
   - State machine of most commonly observed base at each position in read

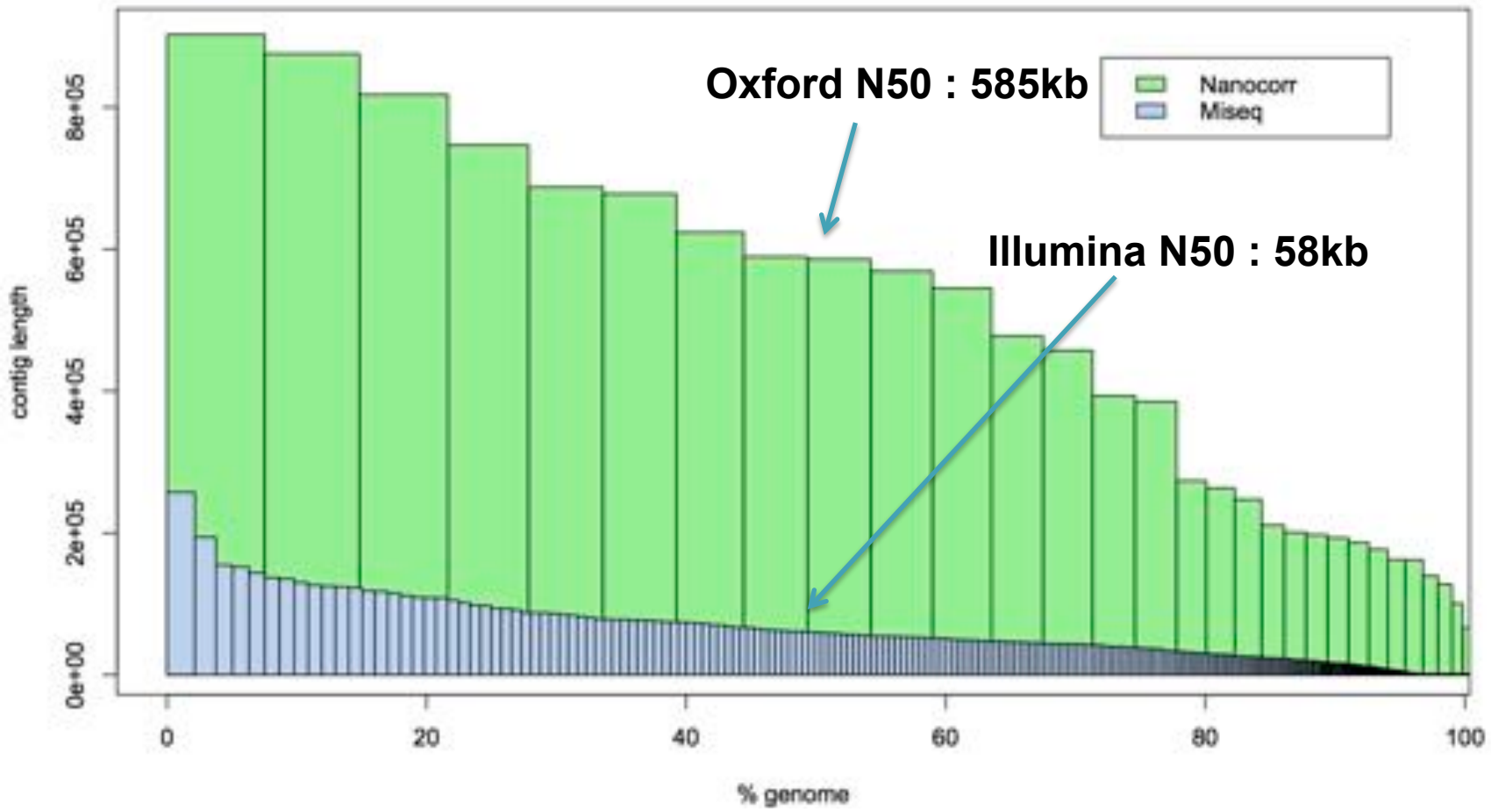




Oxford Nanopore Sequencing and de novo Assembly of a Eukaryotic Genome
Goodwin, S, Gurtowski, J *et al.* (2015) bioRxiv doi: http://dx.doi.org/10.1101/013490

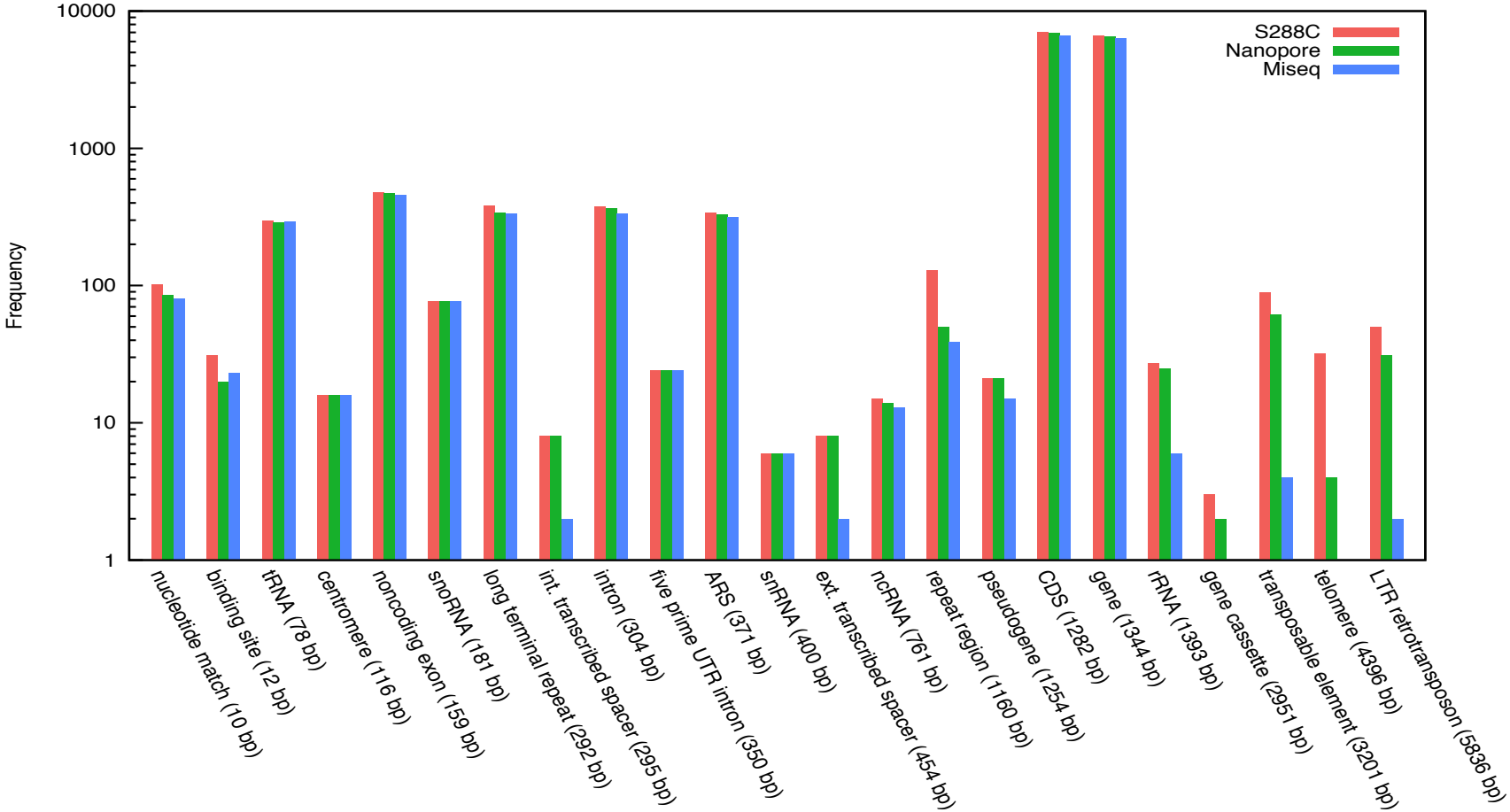
Advantages of Long Reads

In yeast, Nanopore-based assembly is 10x more contiguous
In E. coli, Nanopore-based assembly is basically perfect



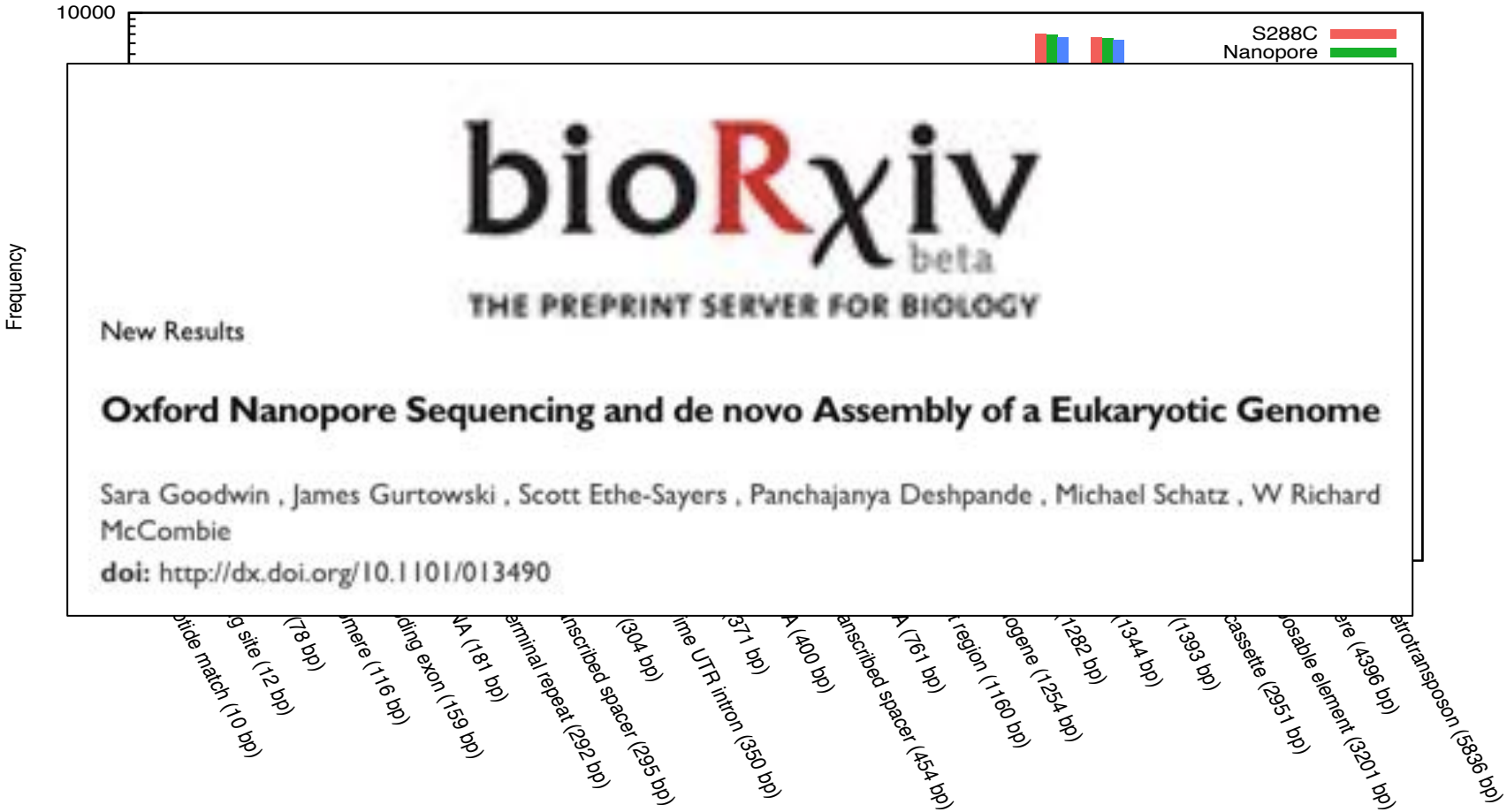
Oxford Nanopore Sequencing and de novo Assembly of a Eukaryotic Genome.
Goodwin, S*, Gurtowski, J*, Ethe-Sayers, S, Deshpande, P, Schatz, MC†, McCombie WR† (2015) *Under review.*

Advantages of Long Reads



Oxford Nanopore Sequencing and de novo Assembly of a Eukaryotic Genome.
Goodwin, S*, Gurtowski, J*, Ethe-Sayers, S, Deshpande, P, Schatz, MC†, McCombie WR† (2014) *Under review.*

Advantages of Long Reads



Oxford Nanopore Sequencing and de novo Assembly of a Eukaryotic Genome.
Goodwin, S*, Gurtowski, J*, Ethe-Sayers, S, Deshpande, P, Schatz, MC†, McCombie WR† (2014) *Under review.*

Genomic Futures?

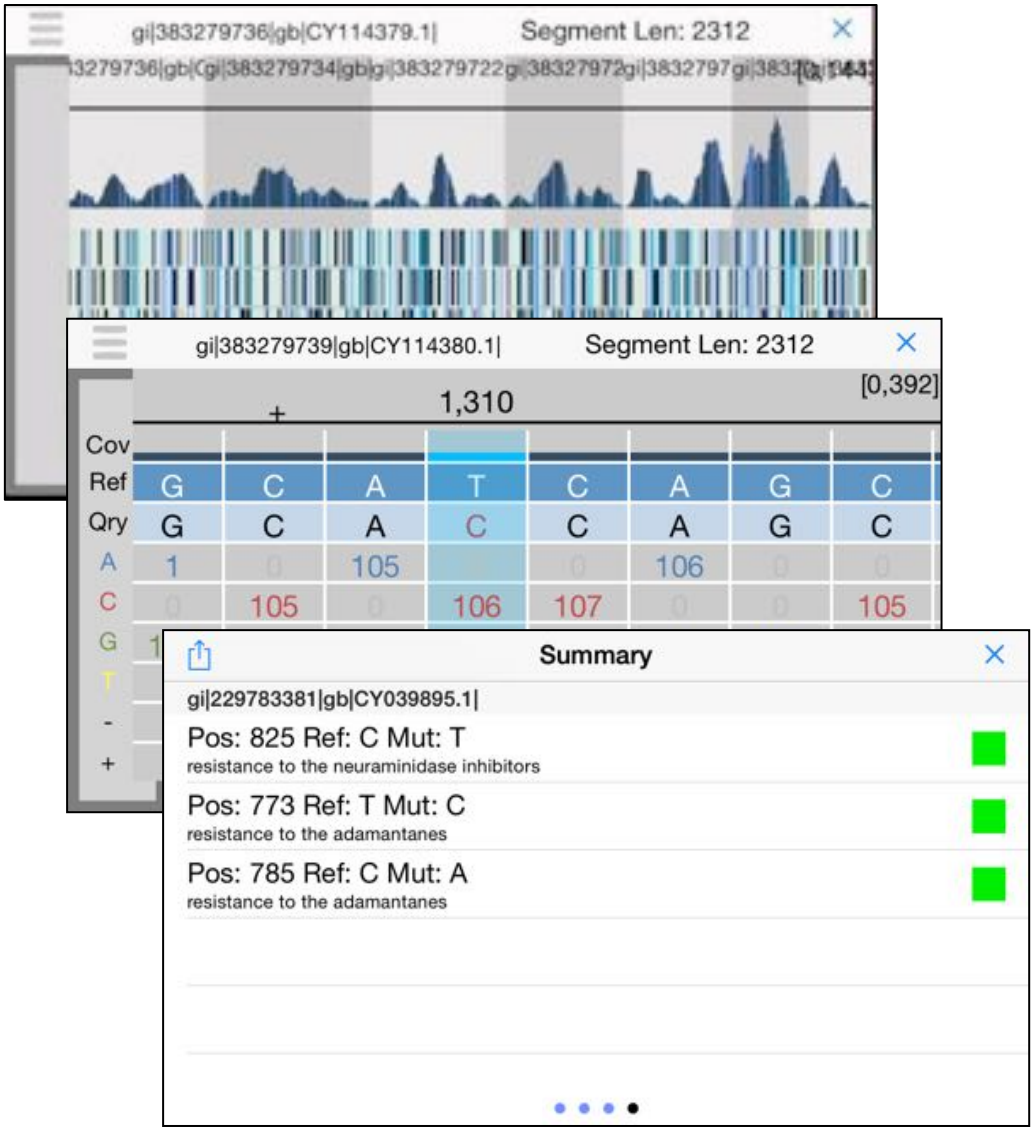
Zamin Iqbal and 5 others retweeted

GenomeWeb InSequence @InSequence · Oct 20

Oxford Nanopore shows off **PromethION** at **ASHG**. #ASHG14 #nanopore

iGenomics: Mobile Sequence Analysis

Aspyn Palatnick, Elodie Ghedin, Michael Schatz



The worlds first genomics analysis app for iOS devices

First application:

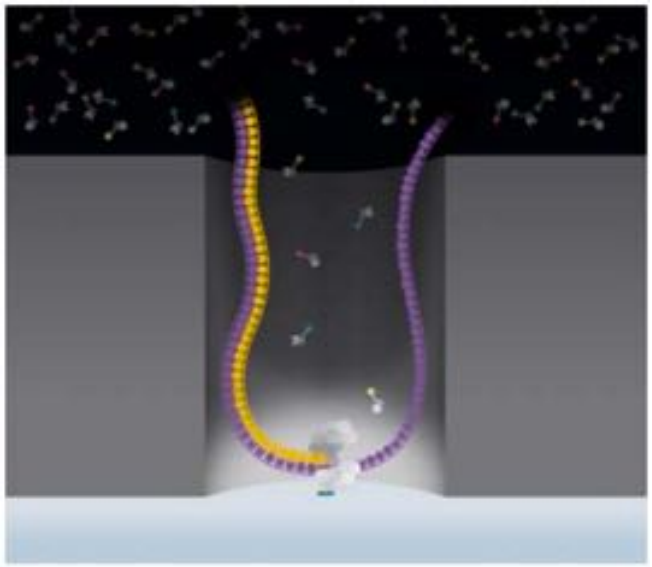
- Handheld diagnostics and therapeutic recommendations for influenza infections
- In a few seconds, iGenomics tells you which antivirals to take or avoid
- Currently in the App Store

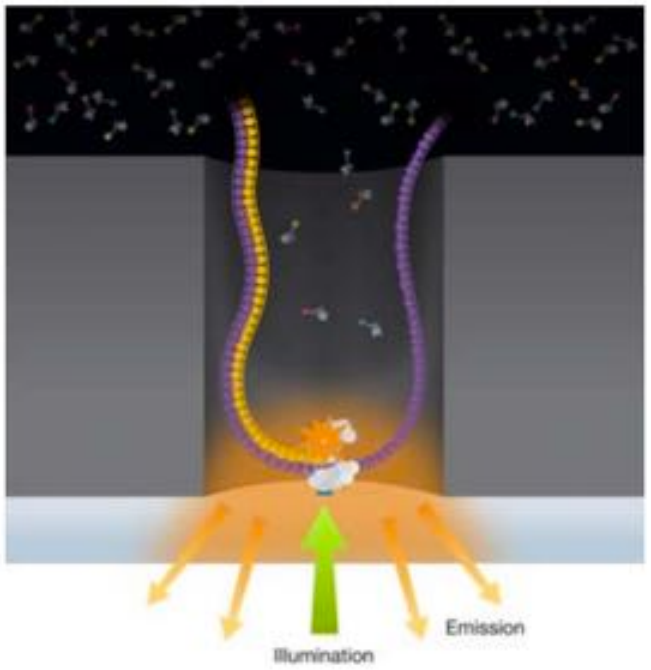
Future applications

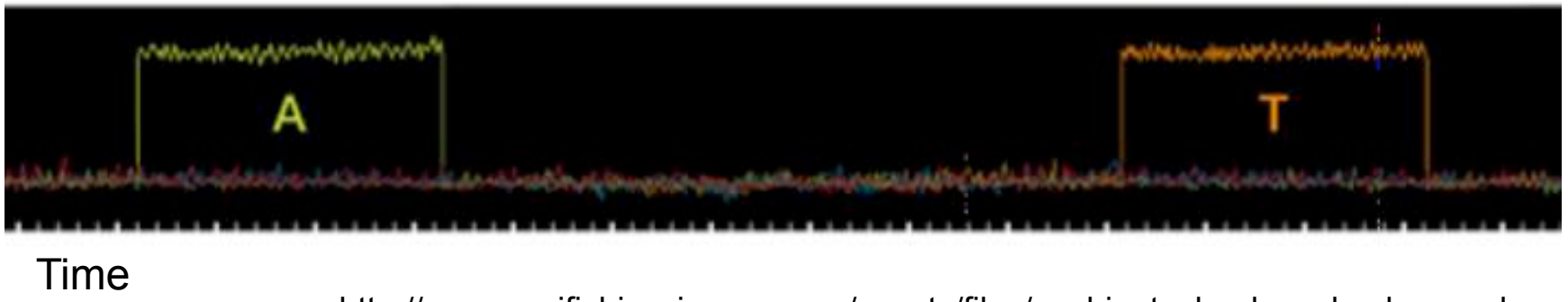
- Pathogen detection
- Food safety
- Biomarkers
- etc..

PacBio SMRT Sequencing

Imaging of fluorescently phospholinked labeled nucleotides as they are incorporated by a polymerase anchored to a Zero-Mode Waveguide (ZMW).

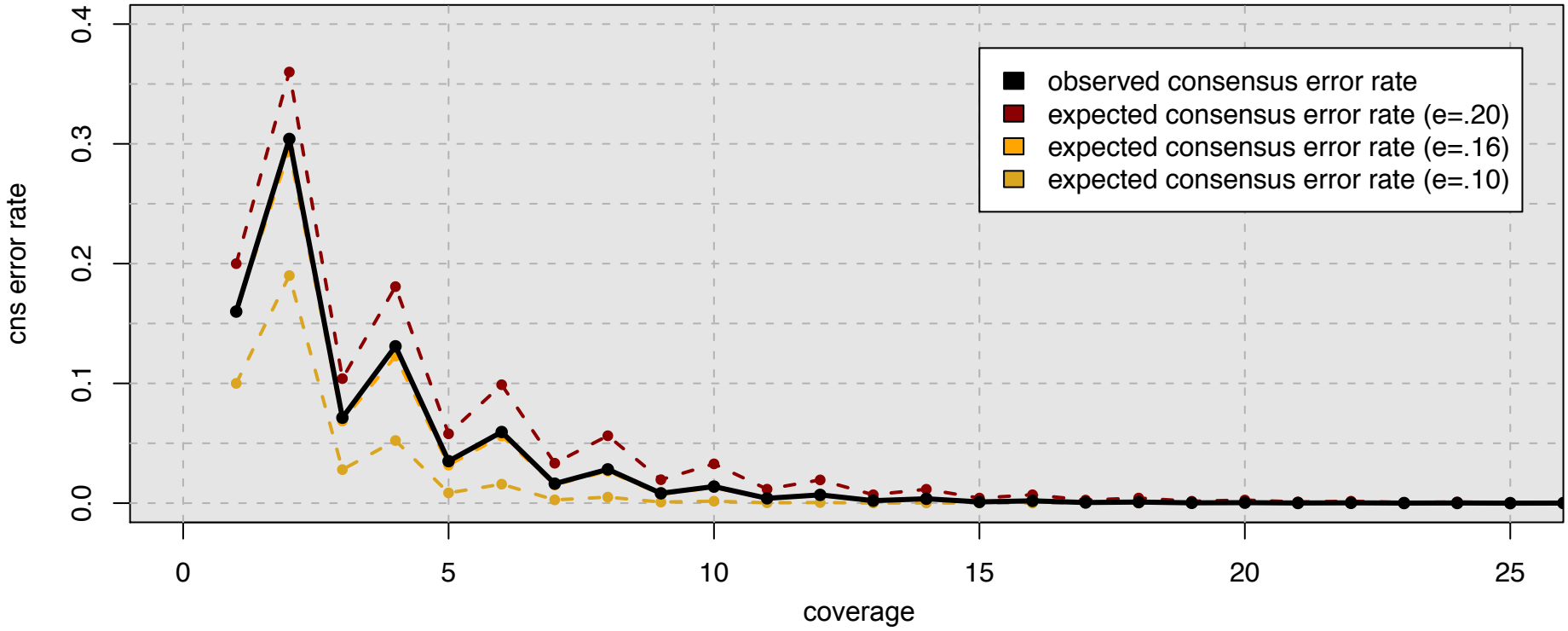






http://www.pacificbiosciences.com/assets/files/pacbio_technology_backgrounder.pdf

Consensus Accuracy and Coverage



Coverage can overcome random errors

- Dashed: error model from binomial sampling
- Solid: observed accuracy

$$CNS\ Error = \sum_{i=\lceil c/2 \rceil}^{c} \binom{c}{i} (e)^{i} (1-e)^{n-i}$$

Koren, Schatz, *et al* (2012)
Nature Biotechnology. 30:693–700

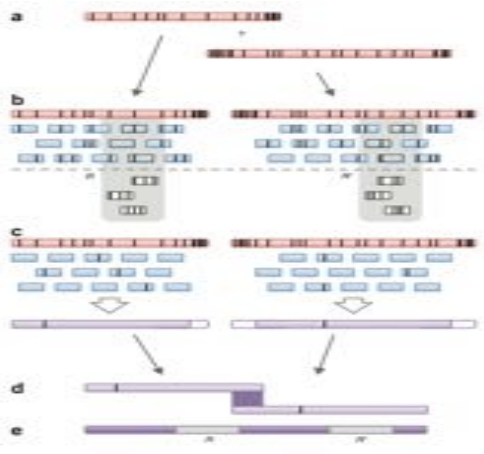
PacBio Assembly Algorithms

PBJelly

Gap Filling and Assembly Upgrade

English *et al* (2012)
PLOS One. 7(11): e47768

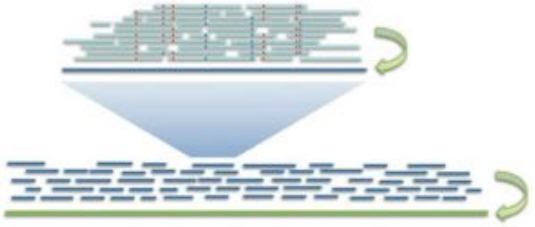
PacBioToCA & ECTools



Hybrid/PB-only Error Correction

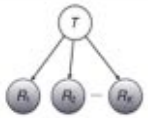
Koren, Schatz, *et al* (2012)
Nature Biotechnology. 30:693–700

HGAP & Quiver



$$\Pr(\mathbf{R} \mid T)$$

$$\Pr(\mathbf{R} \mid T) = \prod_k \Pr(R_k \mid T)$$



Quiver Performance Results
Comparison to Reference Genome
(M. ruber ; 3.1 MB ; SMRT® Cells)

	Initial Assembly	Quiver Consensus
QV	43.4	54.5
Accuracy	99.99540%	99.99964%
Differences	141	11

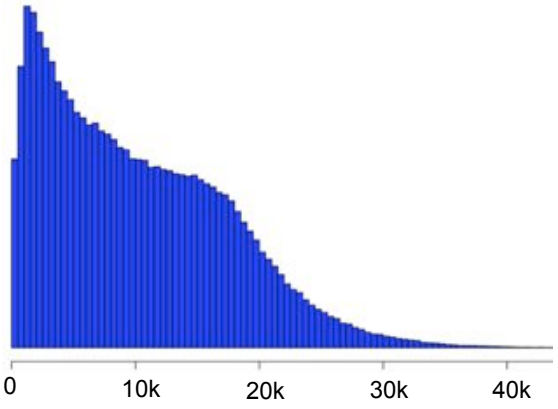
PB-only Correction & Polishing

Chin *et al* (2013)
Nature Methods. 10:563–569

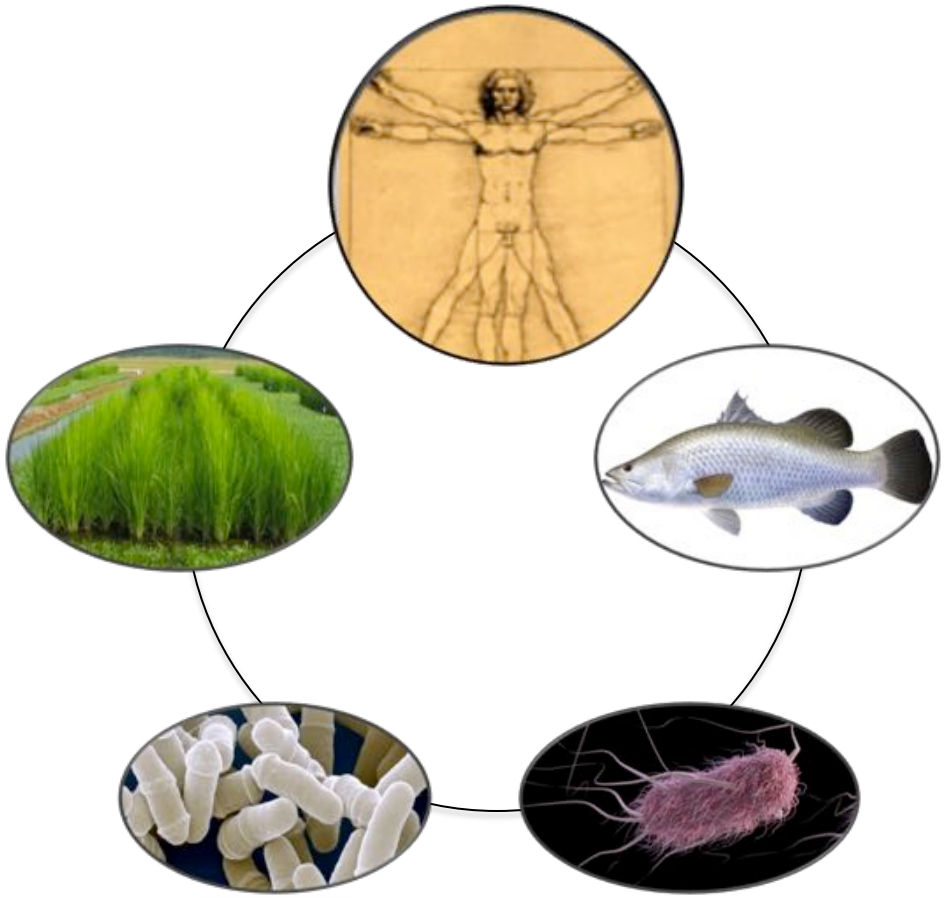
< 5x PacBio Coverage > 50x

3rd Gen Long Read Sequencing

PacBio RS II

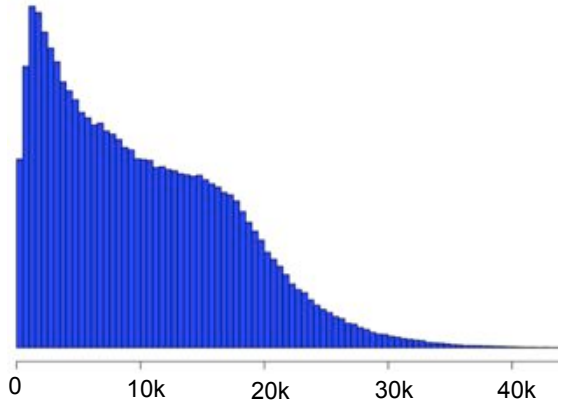


CSHL/PacBio

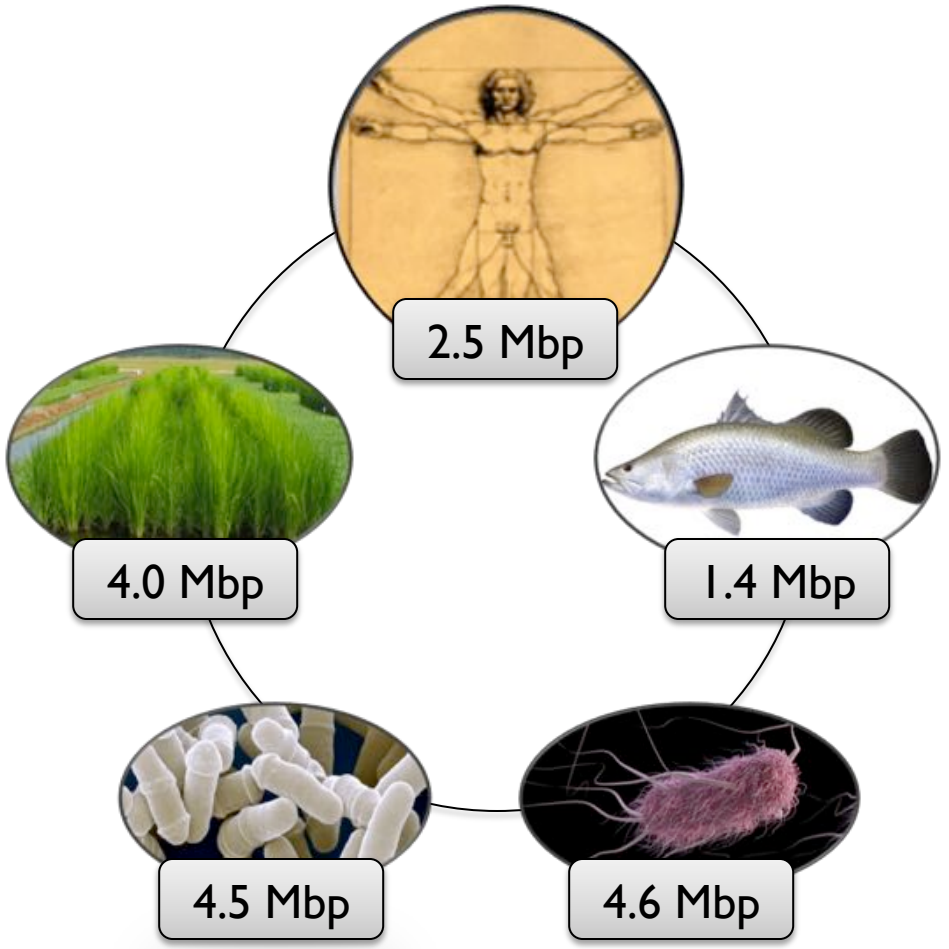


3rd Gen Long Read Sequencing

PacBio RS II



CSHL/PacBio



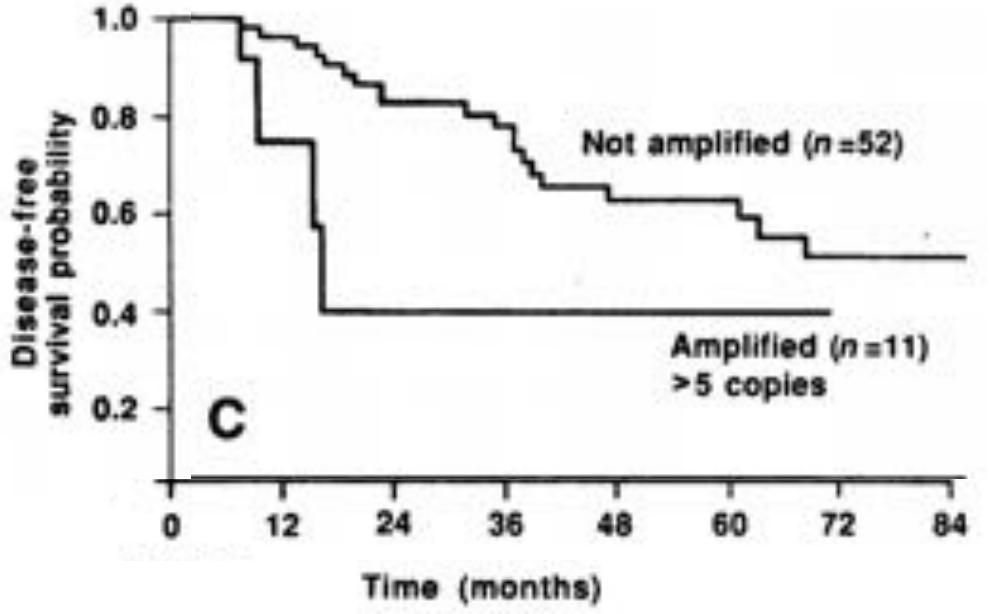
Her2 amplified breast cancer

Breast cancer

- About 12% of women will develop breast cancer during their lifetimes
- ~230,000 new cases every year (US)
- ~40,000 deaths every year (US)

Her2+ breast cancer

- 20% of breast cancers
- 2-3X recurrence risk
- 5X metastasis risk



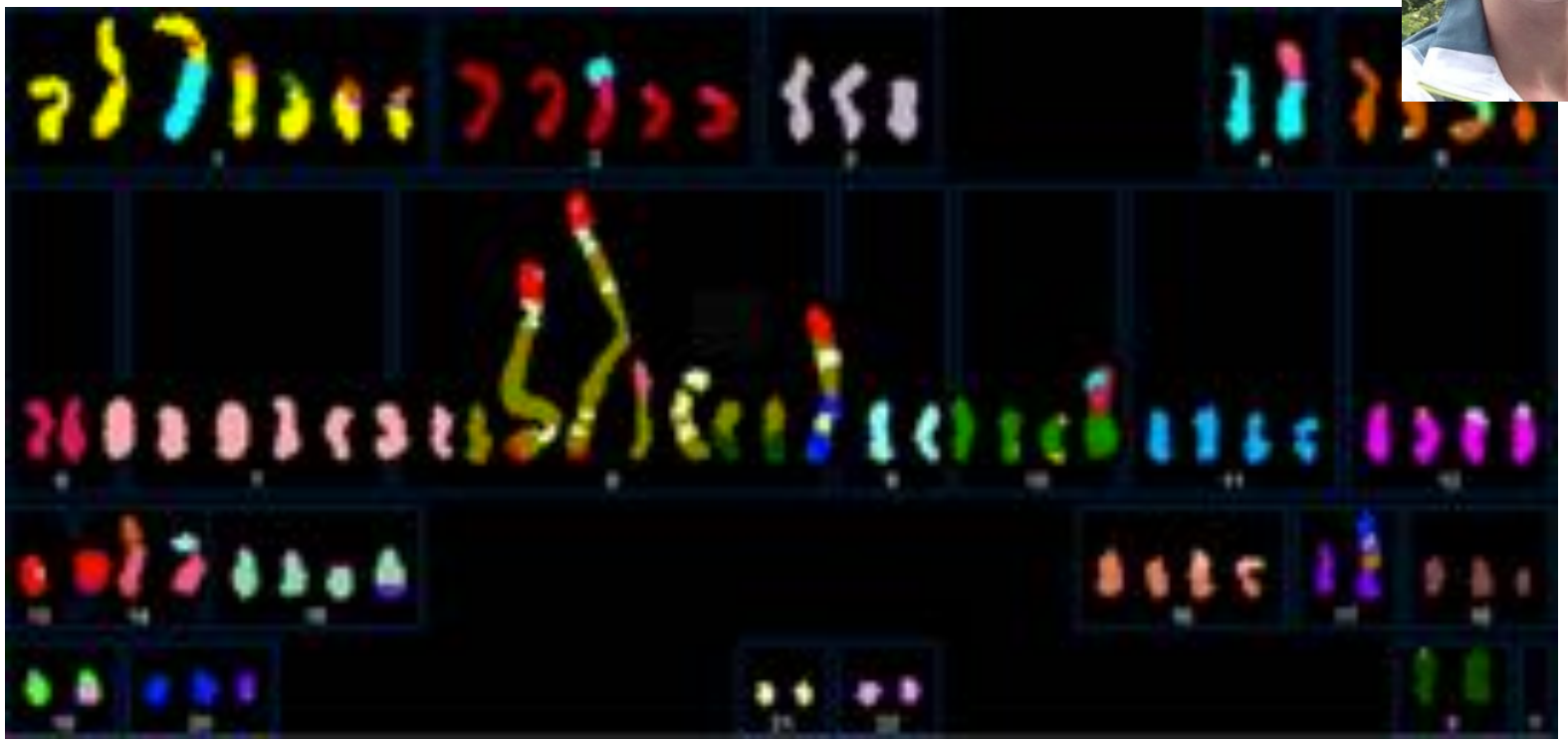
(Adapted from Slamon et al, 1987)

Statistics from American Cancer Society and Mayo Clinic.

Recurrence and metastasis from Gonzalez-Angulo, et al, 2009.

SK-BR-3

Most commonly used Her2-amplified breast cancer ce

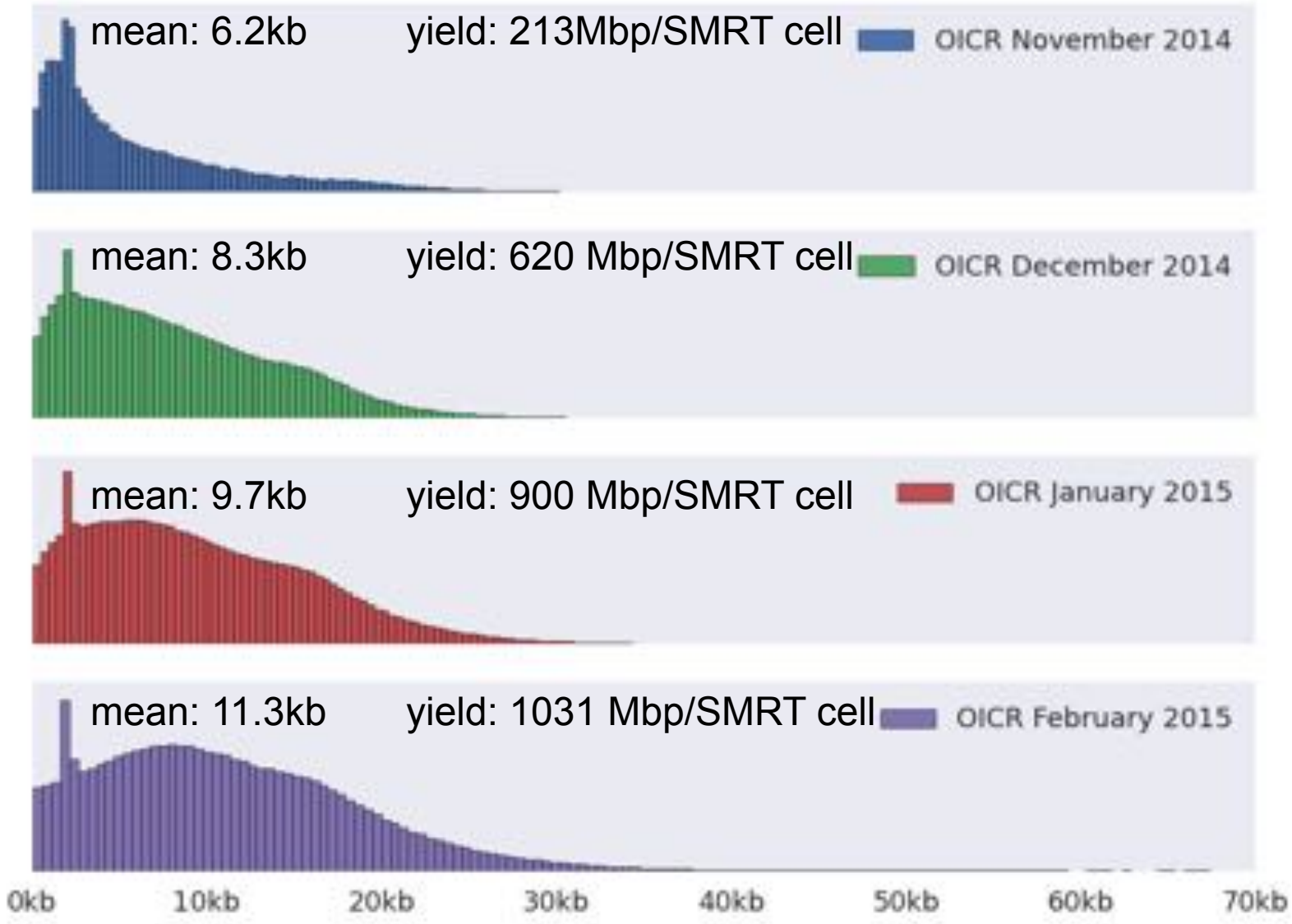


(Davidson et al, 2000)

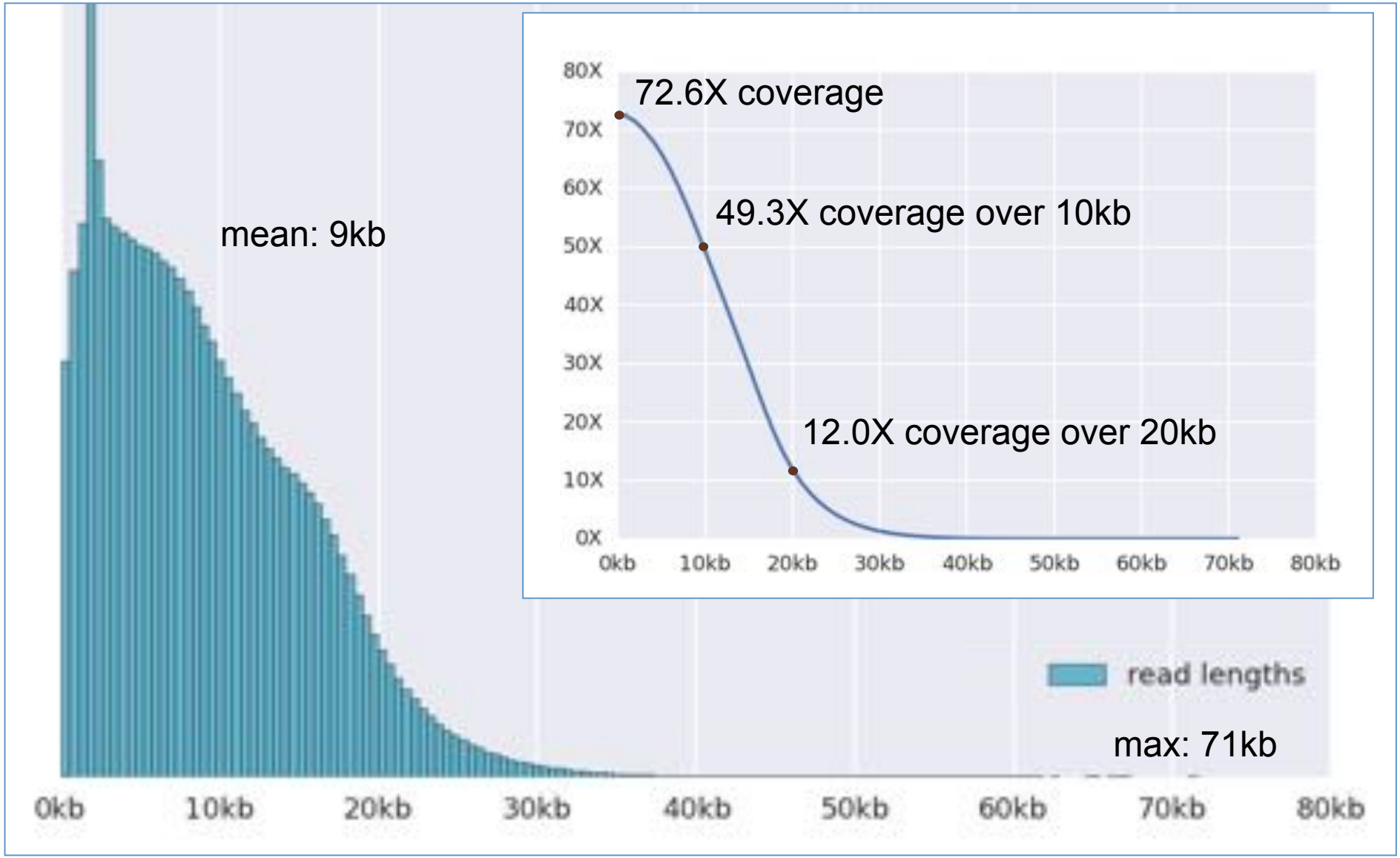
Can we resolve the complex structural variations, especially around Her2?

Ongoing collaboration between CSHL and OICR to *de novo* assemble the complete cell line genome with PacBio long reads

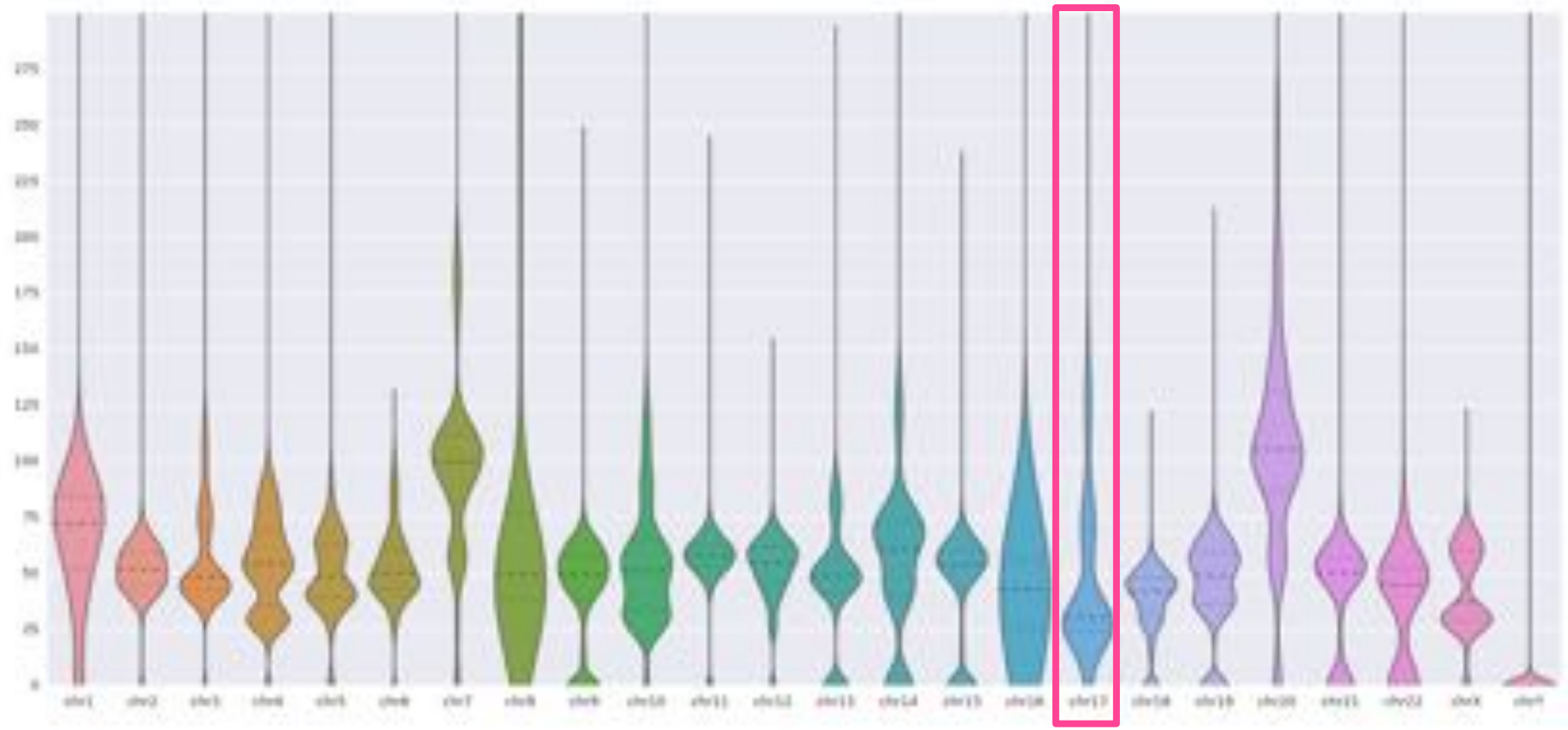
Improving SMRTcell Performance



PacBio read length distribution

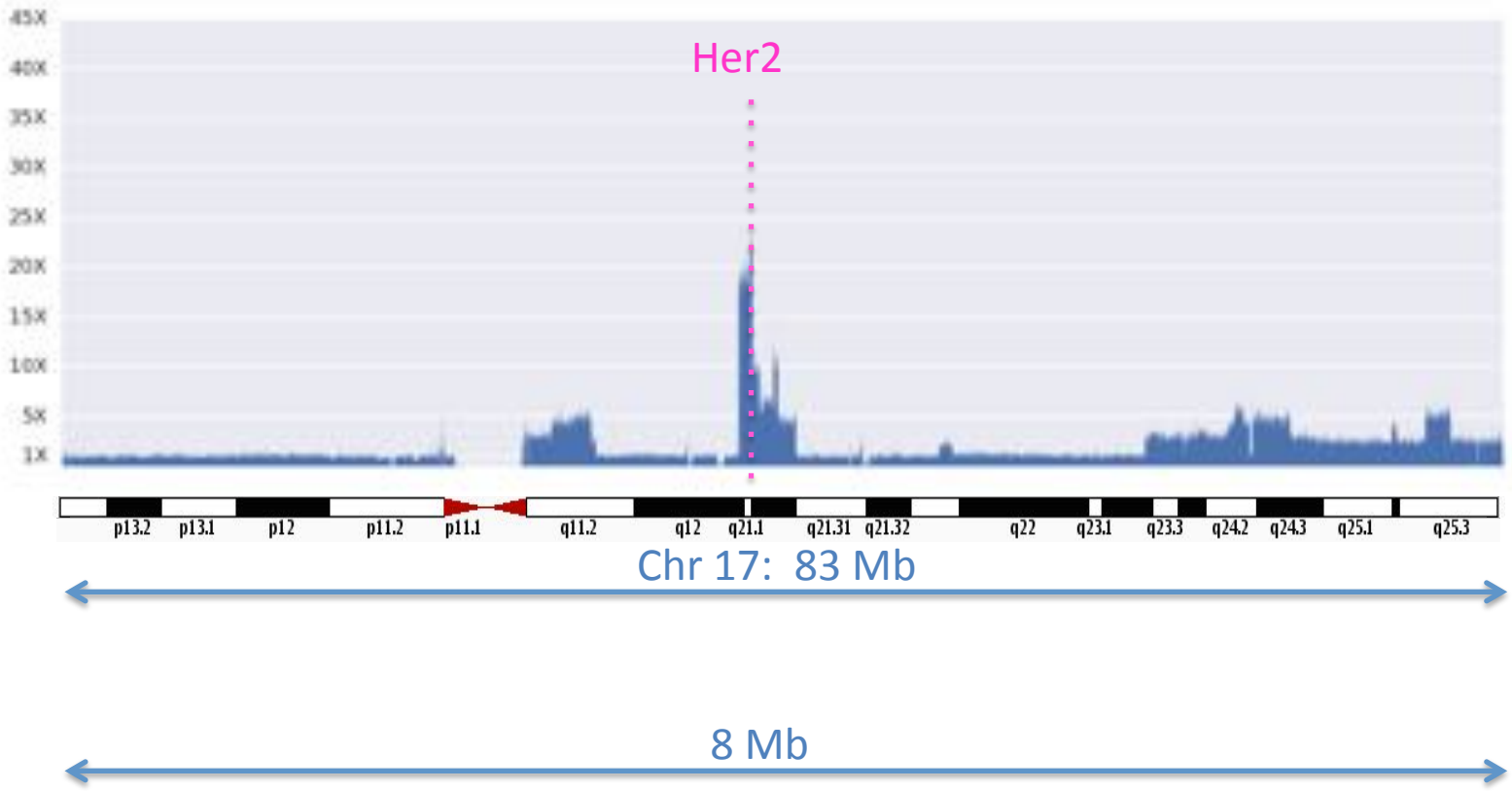


Genome-wide alignment coverage

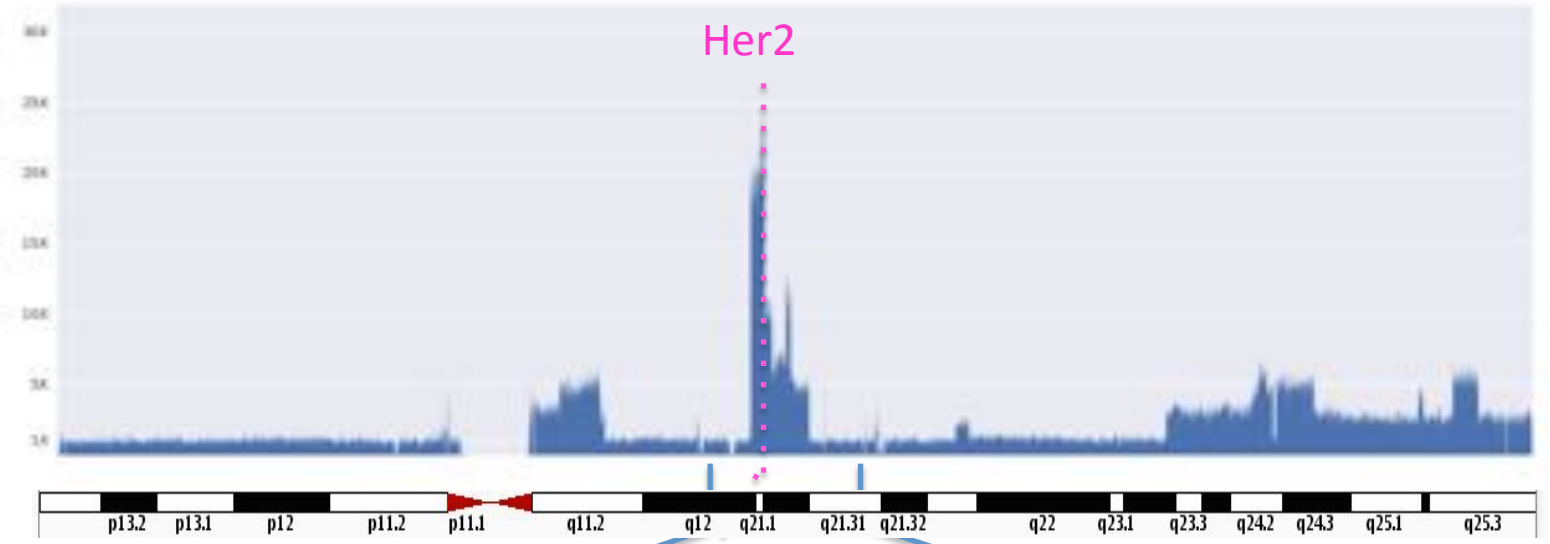


Genome-wide coverage averages around 54X
Coverage per chromosome varies greatly as expected from previous karyotyping results

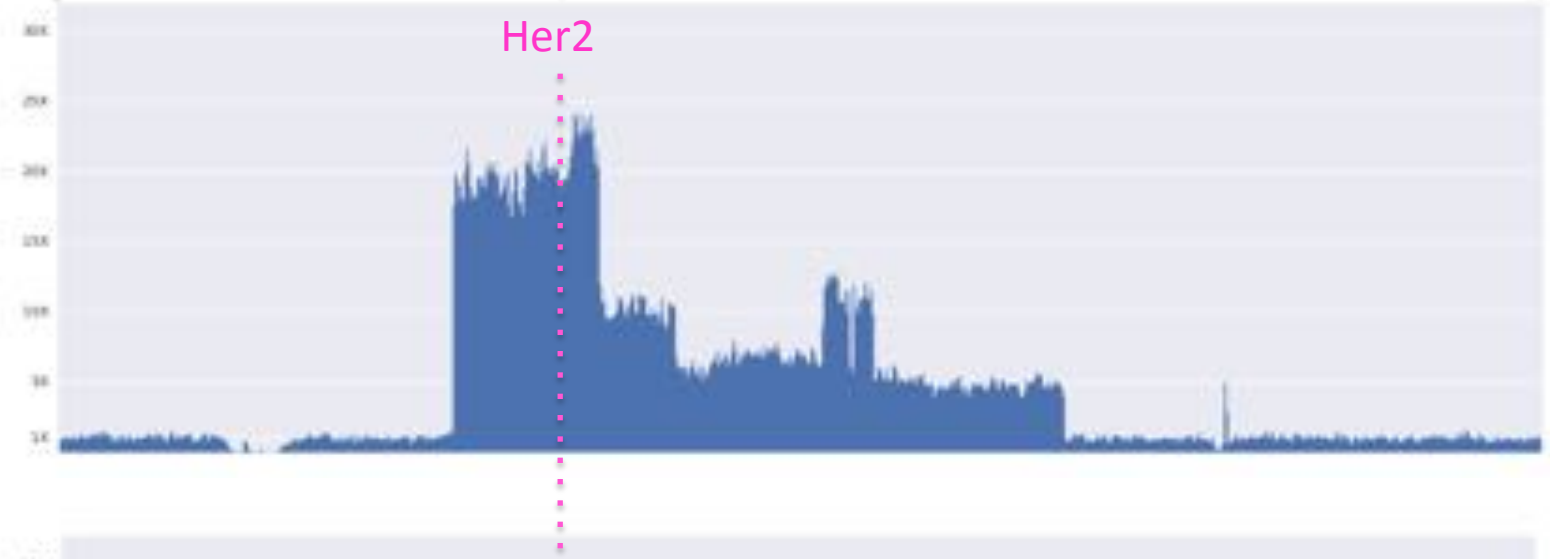
PacBio

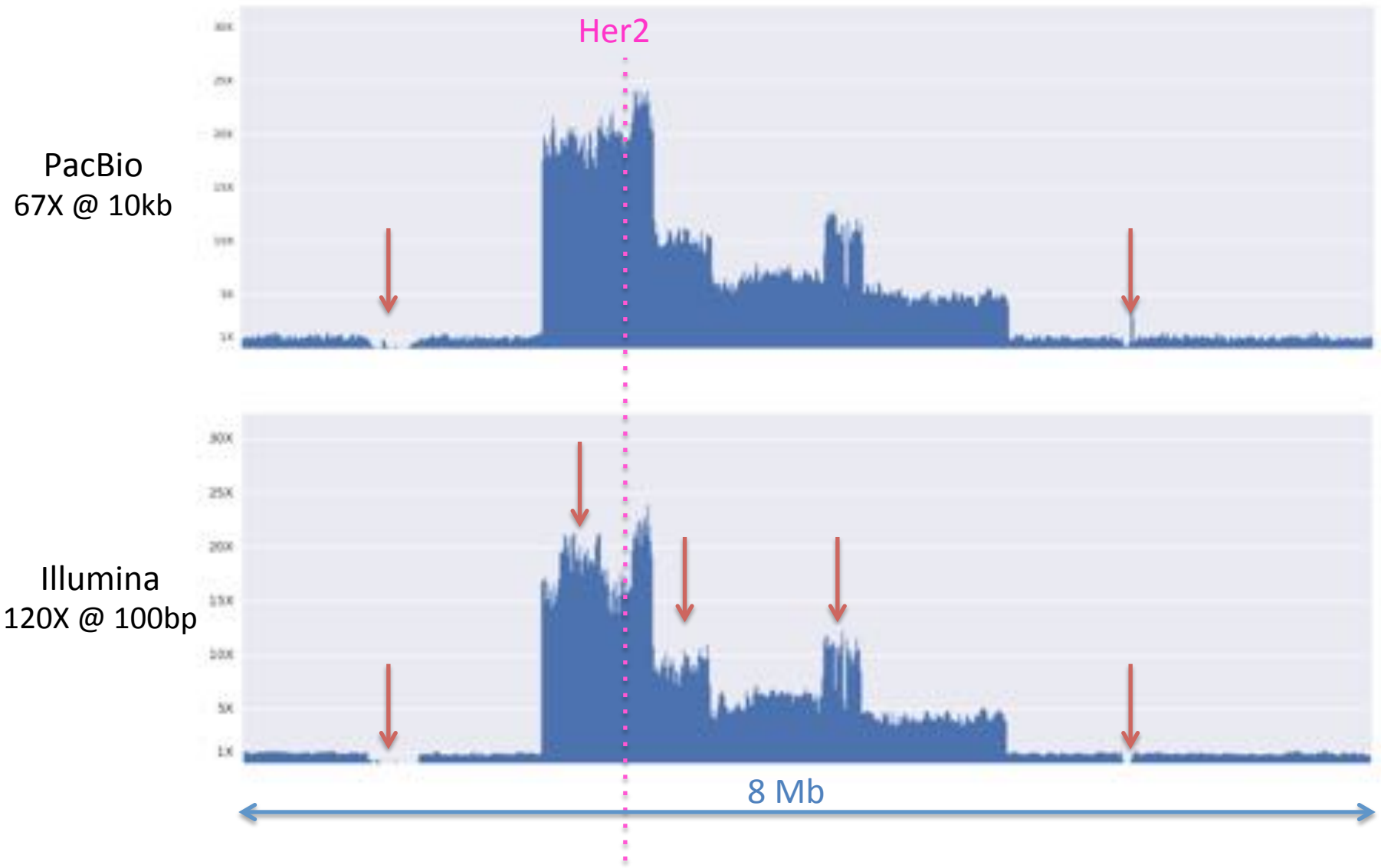

Her2
45X
40X
35X
30X
25X
20X
15X
10X
5X
1X
p13.2
p13.1
p12
p11.2
p11.1
q11.2
q12
q21.1
q21.31
q21.32
q22
q23.1
q23.3
q24.2
q24.3
q25.1
q25.3
Chr 17: 83 Mb
8 Mb

PacBio

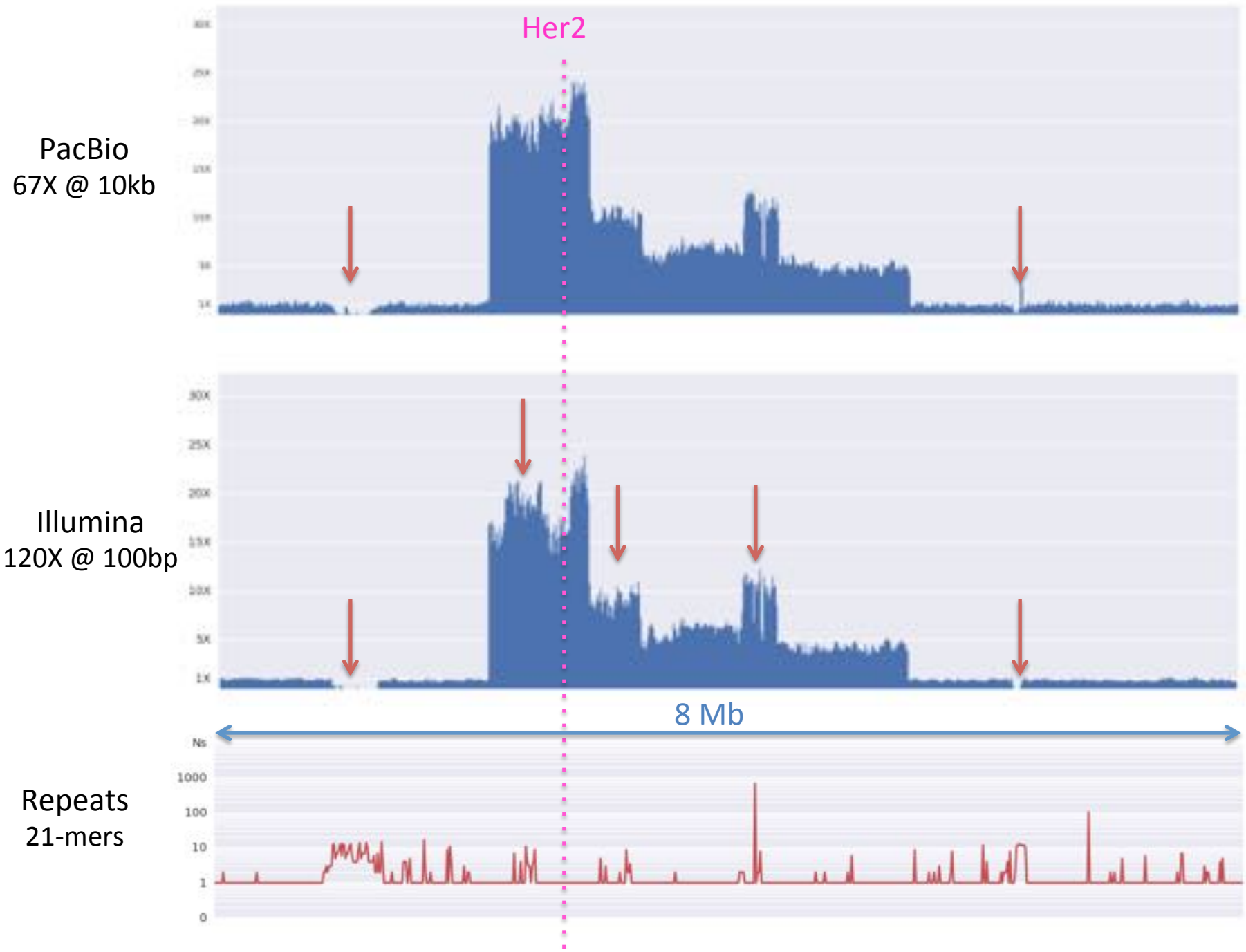


PacBio

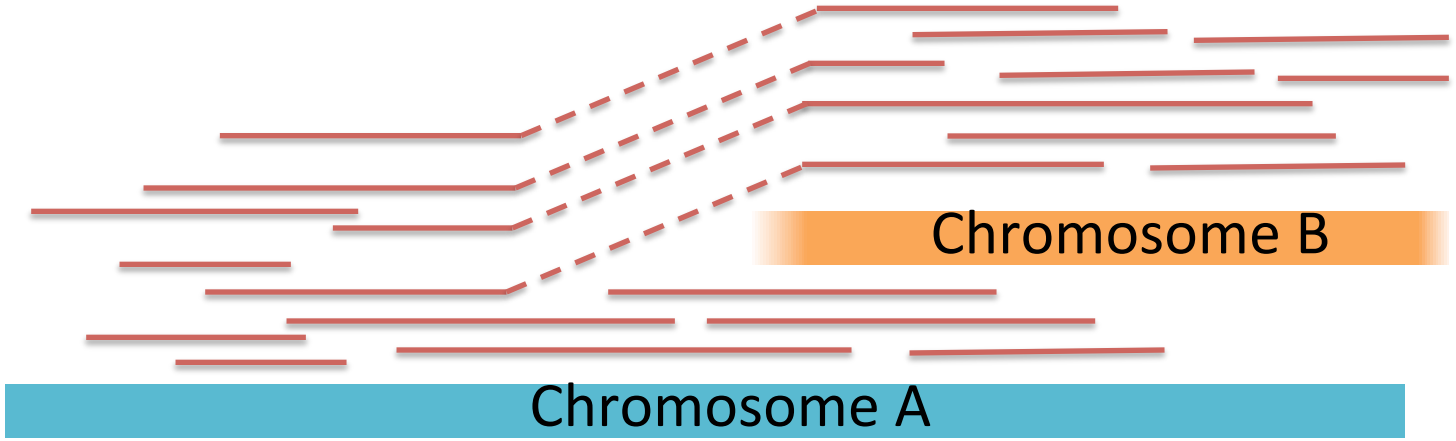




PacBio and Illumina coverage values are highly correlated
but Illumina shows greater variance because of poorly mapping reads


PacBio
67X @ 10kb
Her2
Illumina
120X @ 100bp
8 Mb
Repeats
21-mers
Ns
1000
100
10
1
0

Structural variant discovery with long reads



1. Alignment-based split read analysis: Efficient capture of most events

BWA-MEM + Lumpy

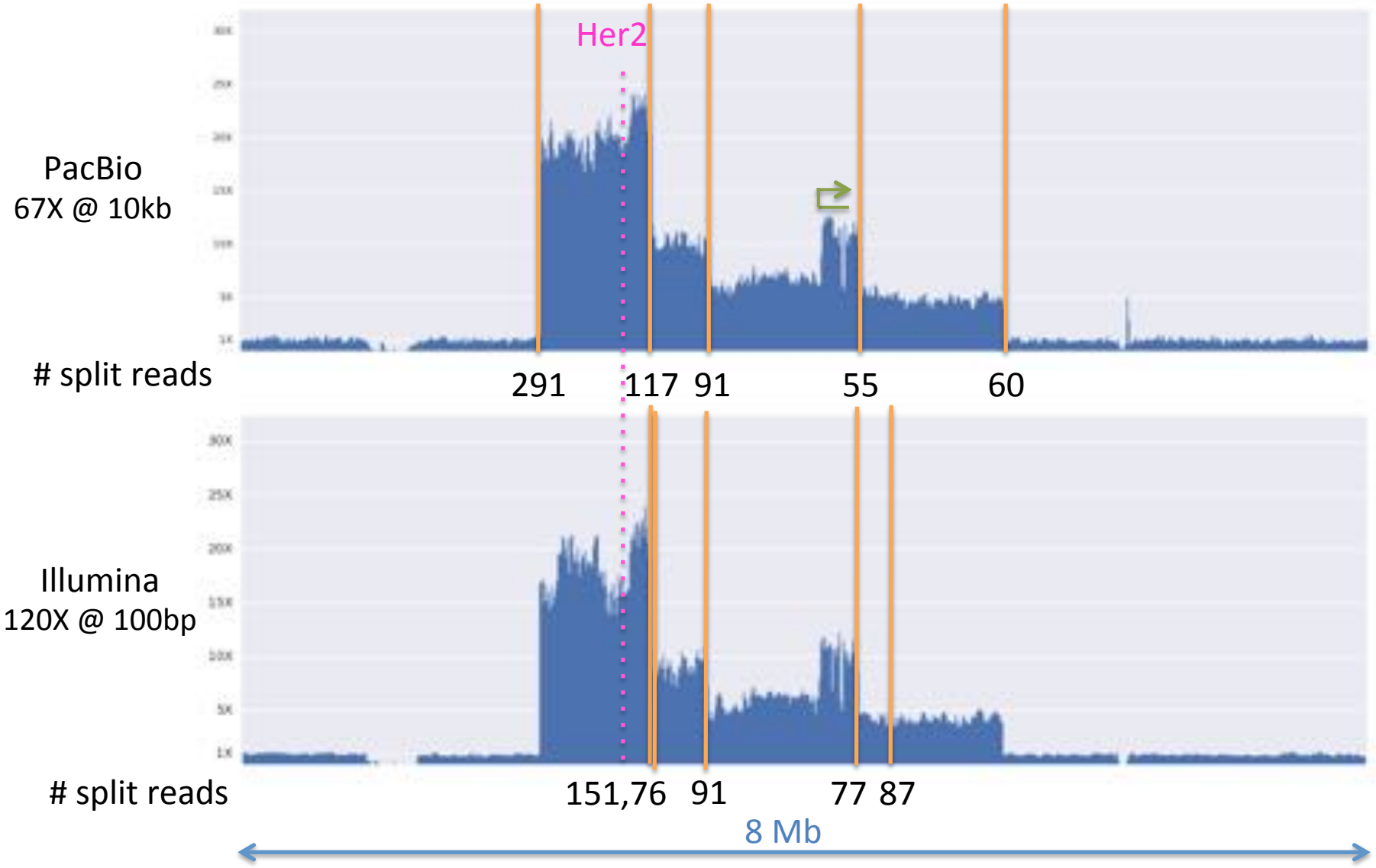
2. Local assembly of regions of interest: In-depth analysis with *base-pair precision*

Localized HGAP + Celera Assembler + MUMmer

3. Whole genome assembly: In-depth analysis including *novel sequences*

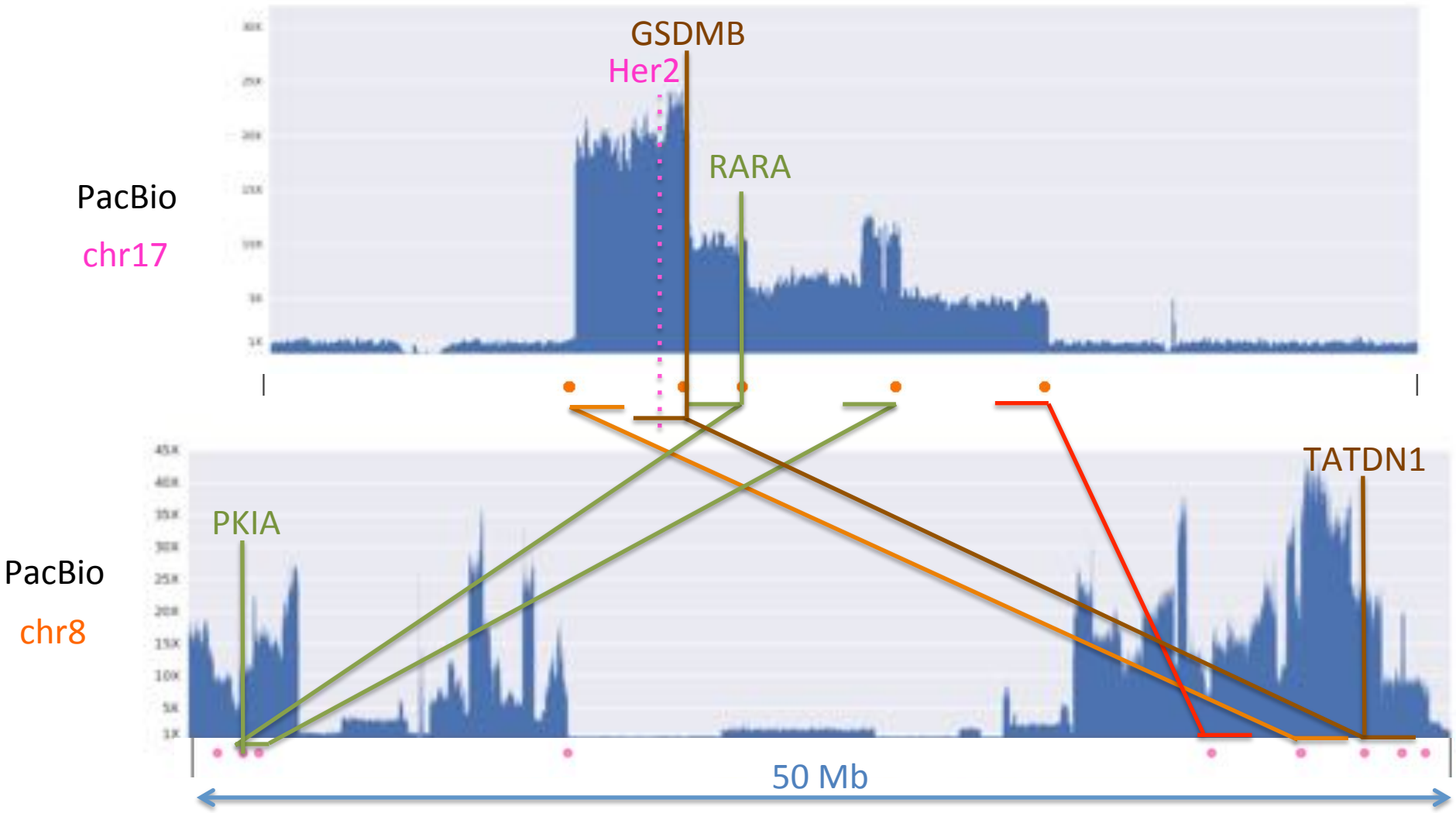
DNAnexus-enabled version of Falcon

Total Assembly: 2.64Gbp **Contig N50: 2.56 Mbp** **Max Contig: 23.5Mbp**

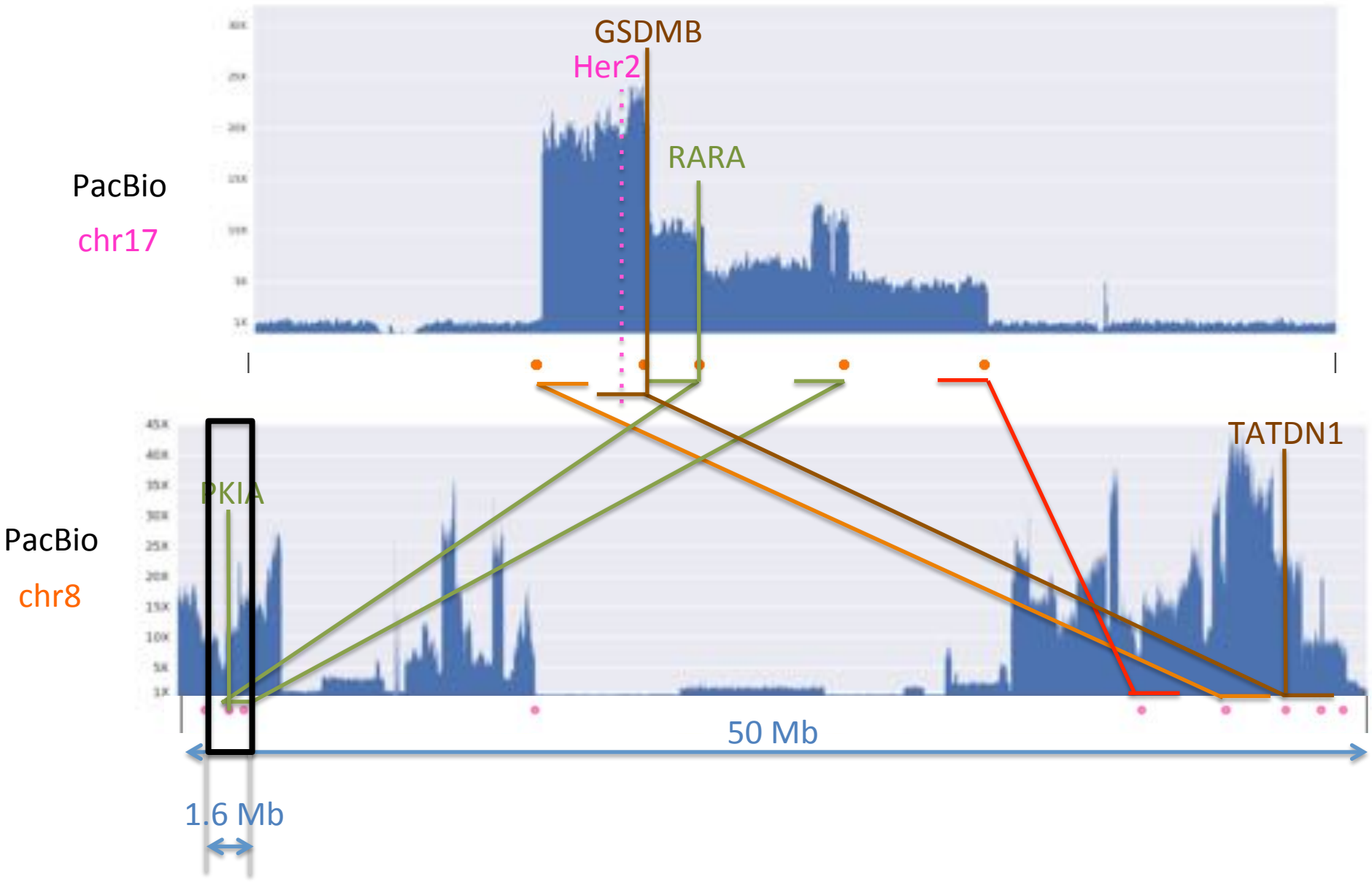


Green arrow indicates an inverted duplication.

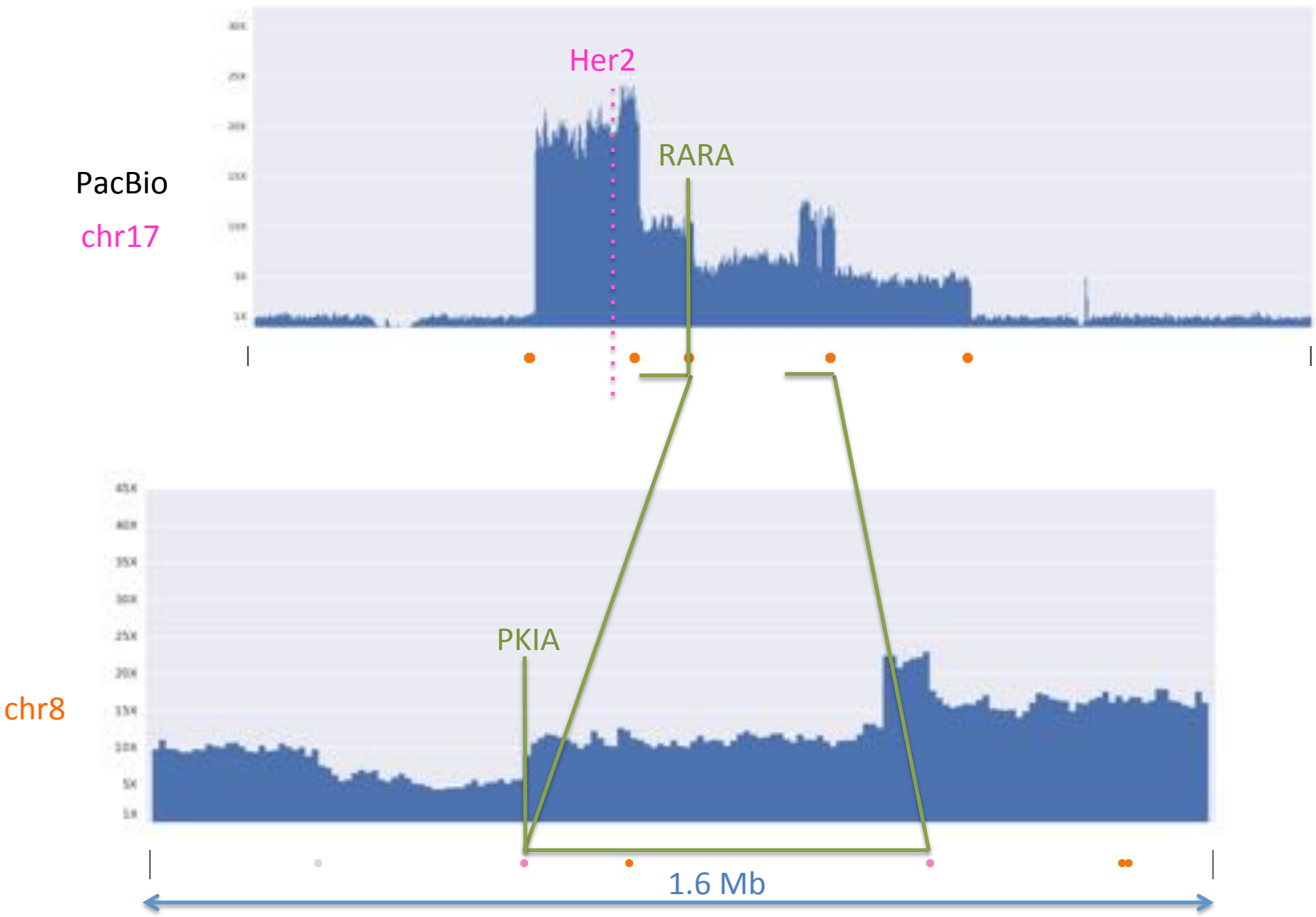
False positive and missing Illumina calls due to mis-mapped reads (especially low complexity).



Confirmed both known gene fusions in this region



Confirmed both known gene fusions in this region

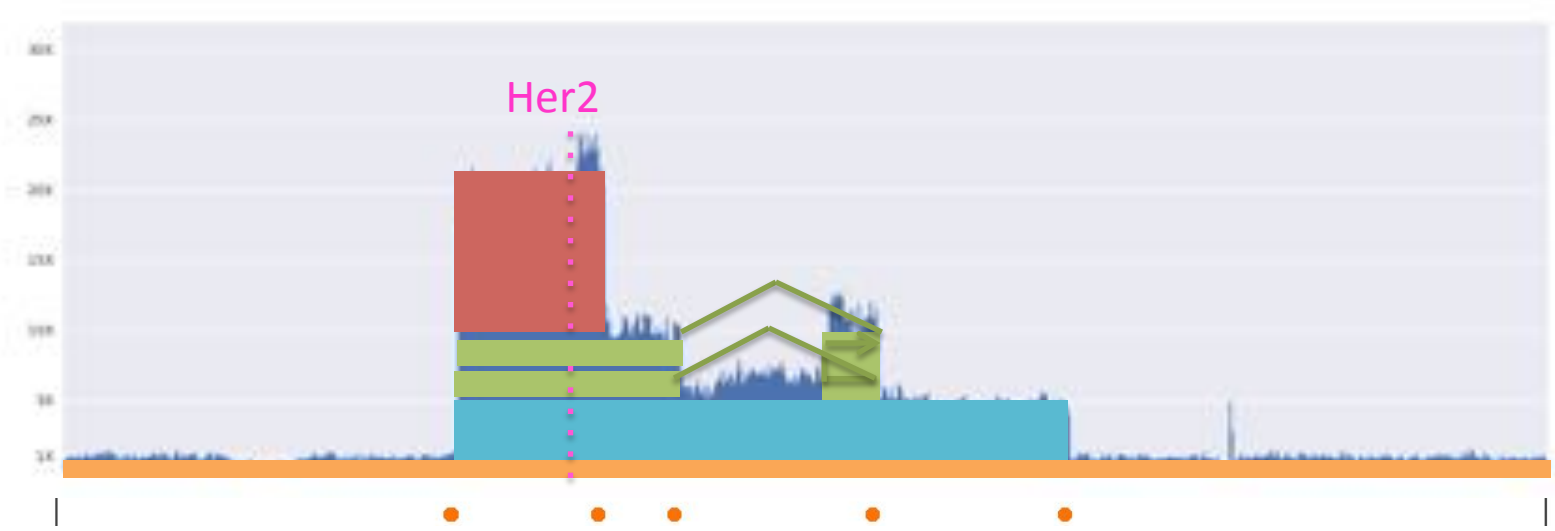


Joint coverage and breakpoint analysis to discover underlying events

Cancer lesion Reconstruction

PacBio

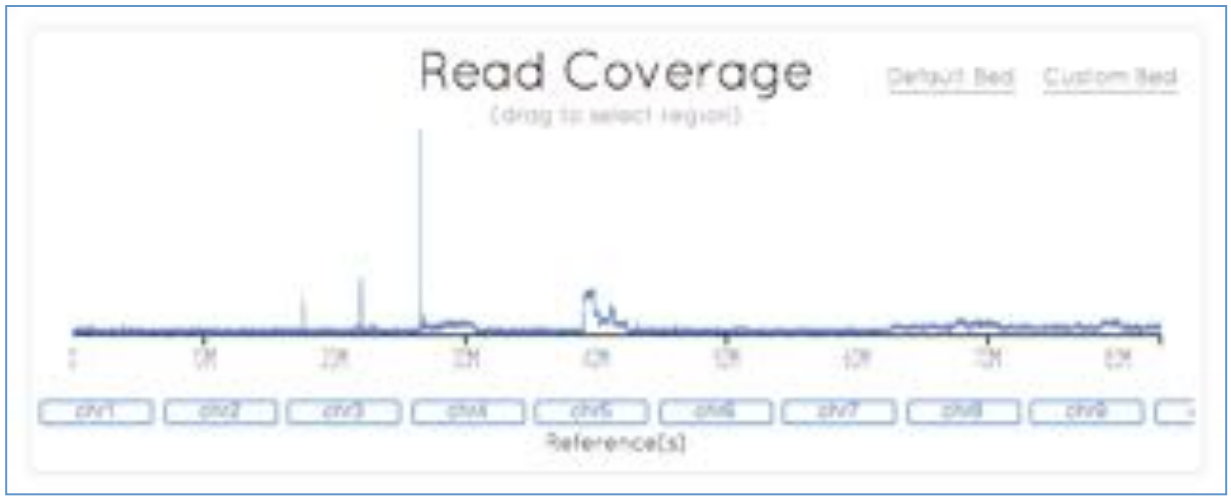
chr17



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

Her2+ Breast Cancer Reference Genome



Available *today* under the Toronto Agreement:

- Fastq & BAM files of aligned reads
- Interactive Coverage Analysis with BAM.IOBIO
- Whole genome assembly

Available soon

- Whole genome methylation analysis
- Full length cDNA transciptome analysis
- Comparison to single cell analysis of >100 individual cells

http://schatzlab.cshl.edu/data/skbr3/

What should we expect from an assembly?

The resurgence of reference quality genomes

Summary & Recommendations

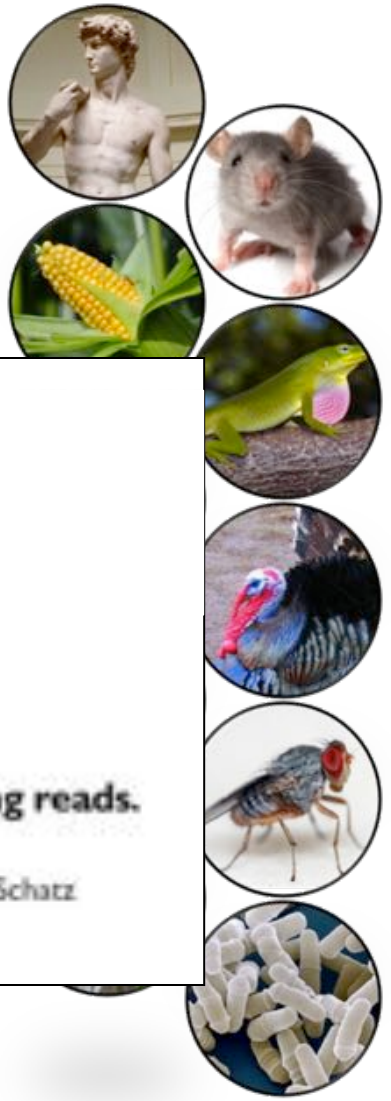
< 100 Mbp: HGAP/PacBio2CA @ 100x PB C3-P5

expect near perfect chromosome arms

< 1GE

> 1GE

> 5GE



bioRxiv beta

THE PREPRINT SERVER FOR BIOLOGY

New Results

Error correction and assembly complexity of single molecule sequencing reads.

Hayan Lee , James Gurtowski , Shinjae Yoo , Shoshana Marcus , W. Richard McCombie , Michael Schatz

doi: http://dx.doi.org/10.1101/006395

Caveats

Model only as good as the available references (esp. haploid sequences)

Technologies are quickly improving, exciting new scaffolding technologies

Acknowledgements

Schatz Lab
Rahul Amin
Eric Biggers
Han Fang
Tyler Gavin
James Gurtowski
Ke Jiang
Hayan Lee
Zak Lemmon
Shoshana Marcus
Giuseppe Narzisi
Maria Nattestad
Aspyn Palatnick
Srividya Ramakrishnan
Fritz Sedlazeck
Rachel Sherman
Greg Vurture
Alejandro Wences

CSHL
Hannon Lab
Gingeras Lab
Jackson Lab
Hicks Lab
Iossifov Lab
Levy Lab
Lippman Lab
Lyon Lab
Martienssen Lab
McCombie Lab
Tuveson Lab
Ware Lab
Wigler Lab

OICR
Karen Ng
Timothy Beck
Yogi Sundaravadanam
John McPherson

NBACC
Adam Phillippy
Serge Koren

Oxford NANOPORE Technologies

PACIFIC BIOSCIENCES®

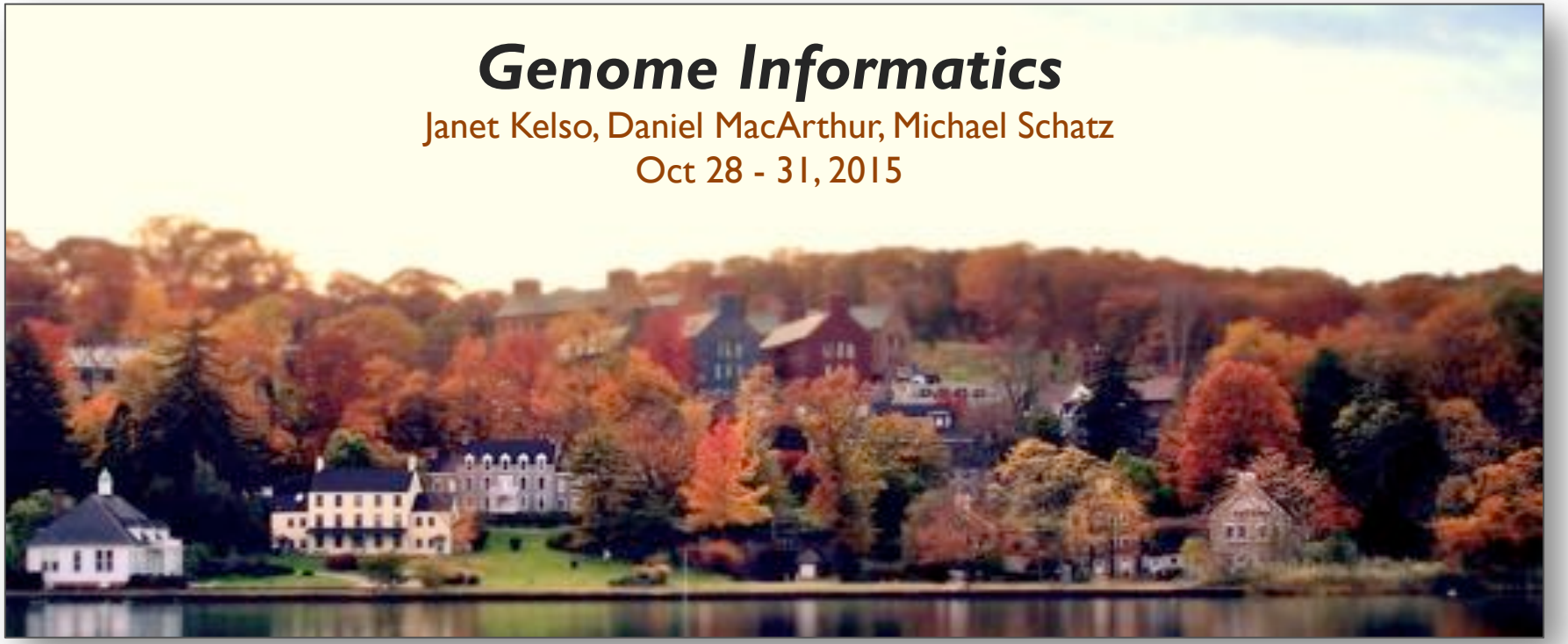
NSF

National Human Genome Research Institute

U.S. DEPARTMENT OF ENERGY

SFARI
SIMONS FOUNDATION AUTISM RESEARCH INITIATIVE

ALFRED P. SLOAN FOUNDATION



Thank you

http://schatzlab.cshl.edu

@mike_schatz