The resurgence of reference quality genomes

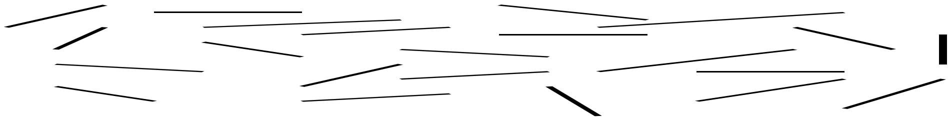
Michael Schatz

May 22, 2015

NYU Genomics Symposium

Sequence Assembly Problem

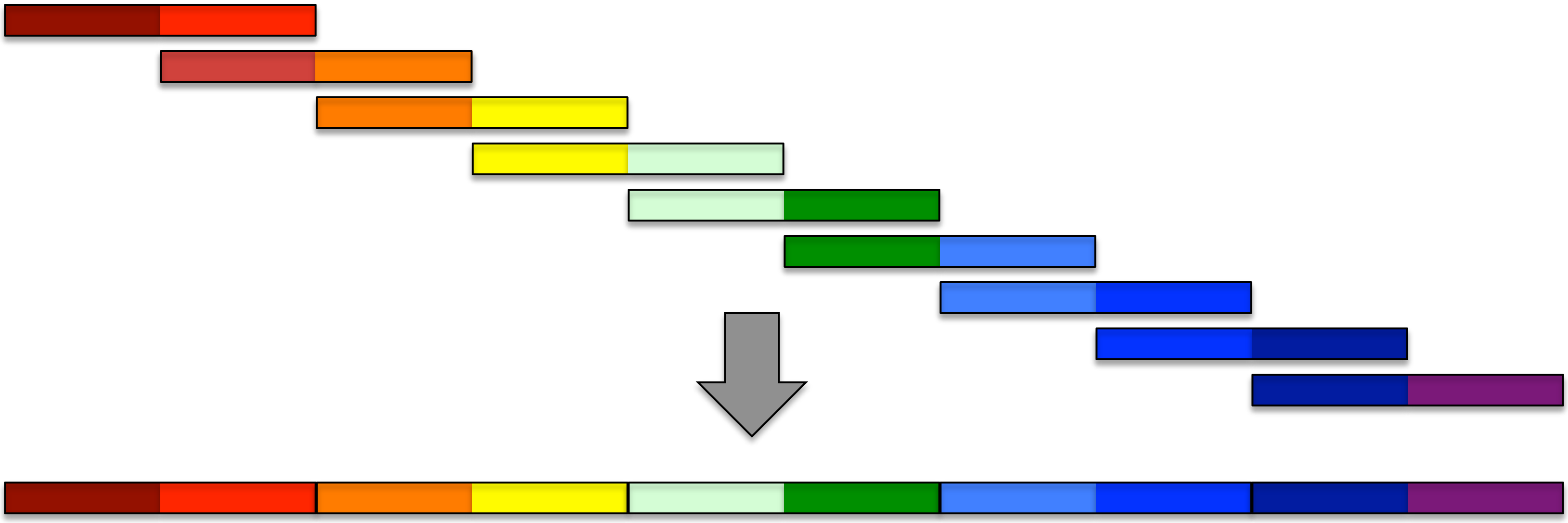
1. Shear & Sequence DNA



2. Construct assembly graph from overlapping reads

```
...AGCCTAGGGATGCGCGACACGT
          GGATGCGCGACACGTCGCATATCCGGTTTGGTCAACCTCGGACGGAC
                                          CAACCTCGGACGGACCTCAGCGAA...
```

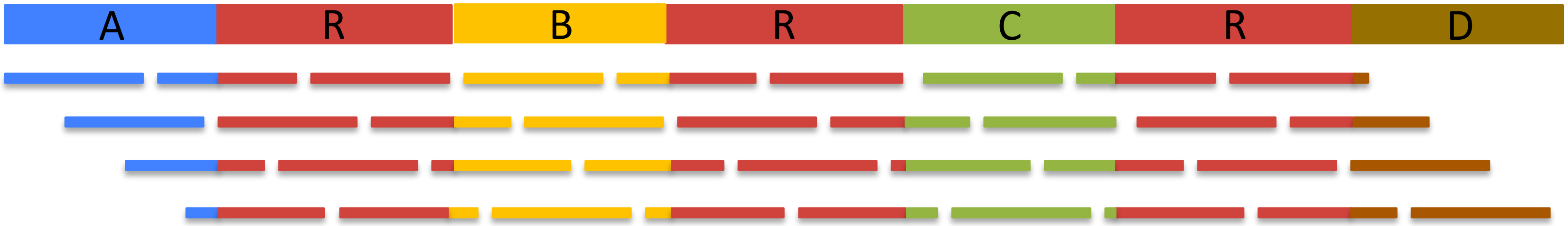
3. Simplify assembly graph

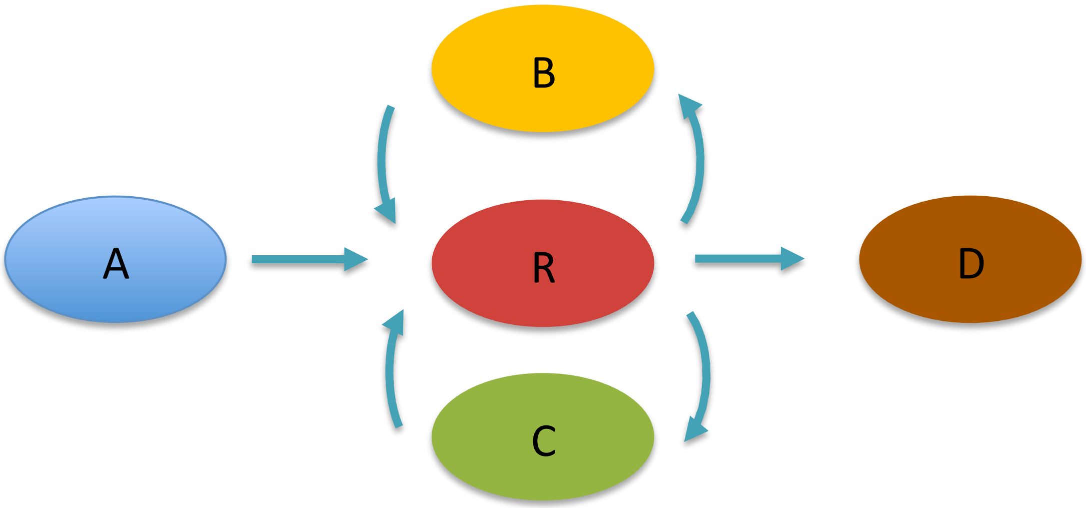


On Algorithmic Complexity of Biomolecular Sequence Assembly Problem

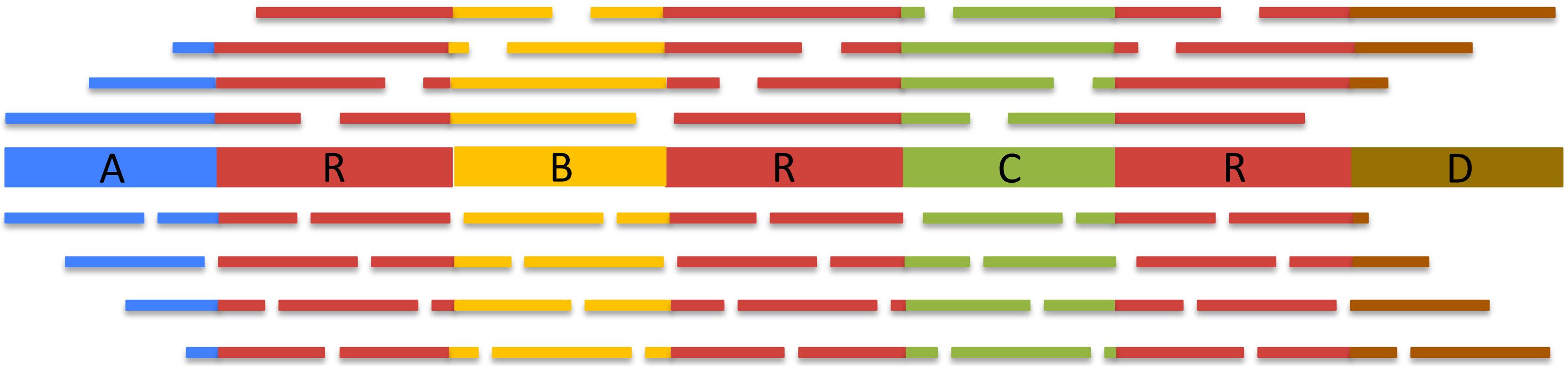
Narzisi, G, Mishra, B, Schatz, MC (2014) *Algorithms for Computational Biology.* Lecture Notes in Computer Science. *Vol. 8542*

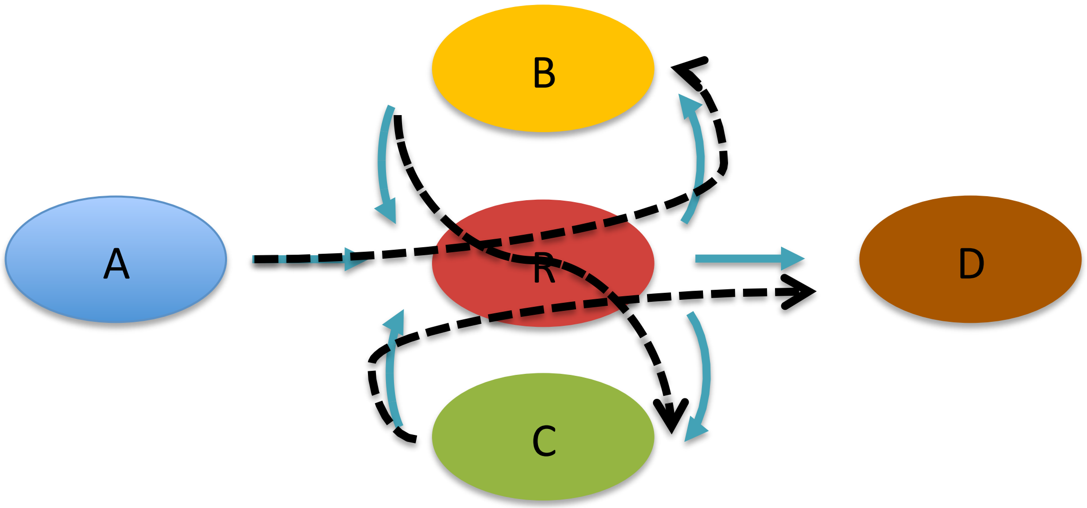
Assembly Complexity



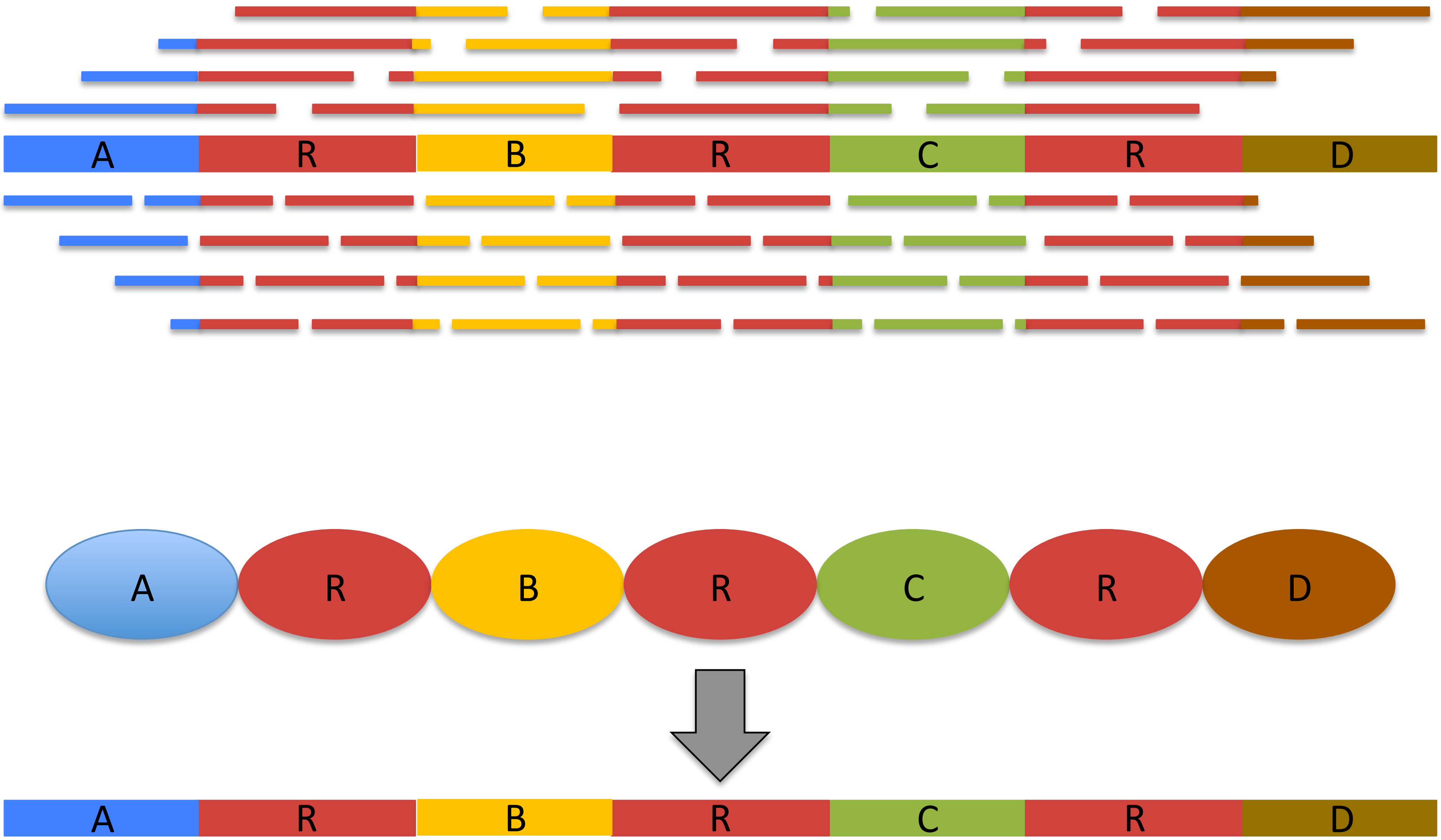


Assembly Complexity





Assembly Complexity



The advantages of SMRT sequencing
Roberts, RJ, Carneiro, MO, Schatz, MC (2013) *Genome Biology.* 14:405

Genomics Arsenal in the Year 2015

Long Read Sequencing: De novo assembly, SV analysis, phasing

Illumina/Moleculo

(Kuleshov et al. 2014)

Pacific Biosciences

(Berlin et al, 2014)

Oxford Nanopore

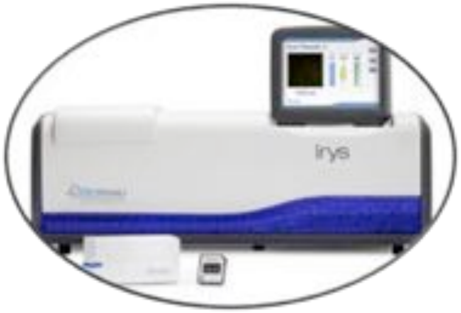
(Quick et al, 2014)

Long Span Sequencing: Chromosome Scaffolding, SV analysis, phasing

Molecular Barcoding

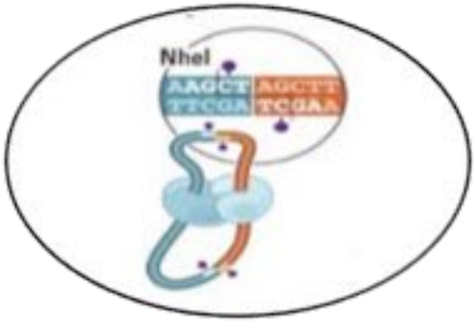
(10Xgenomics.com)

Optical Mapping



(Cao et al, 2014)

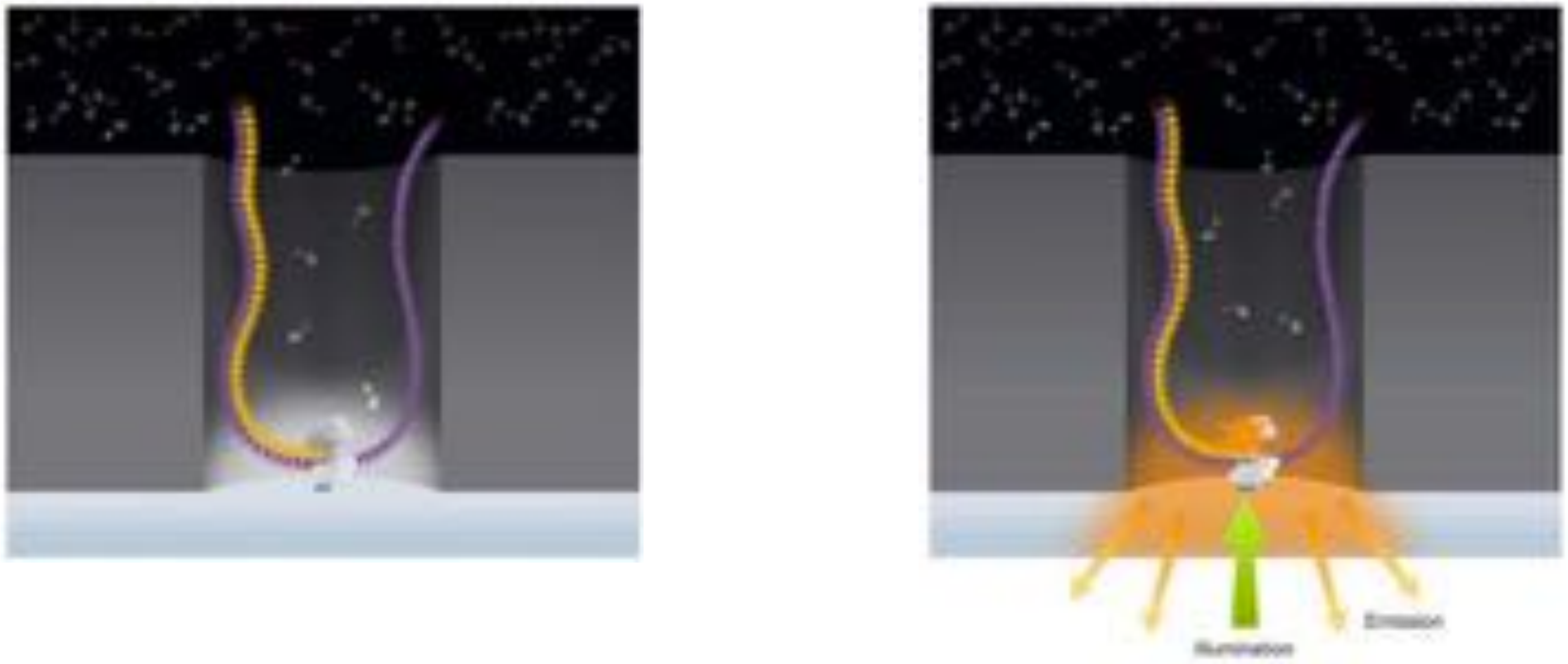
Chromatin Assays

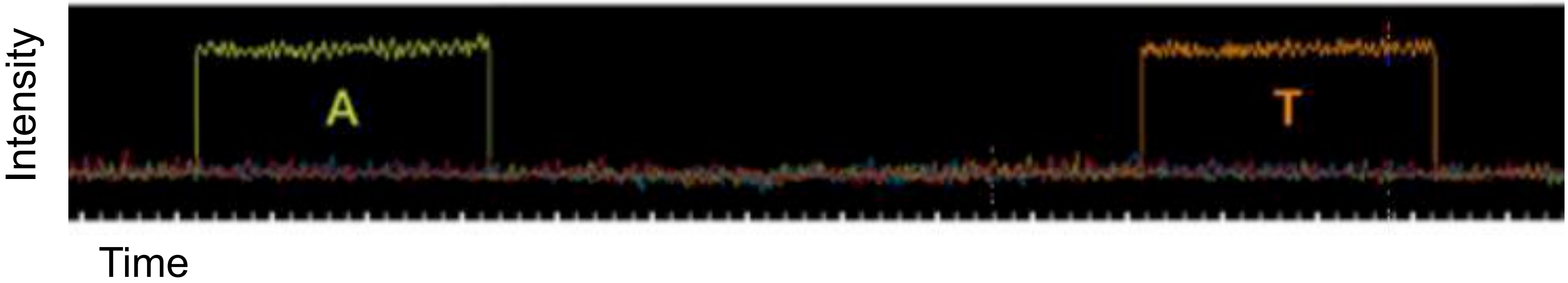


(Putnam et al, 2015)

PacBio SMRT Sequencing

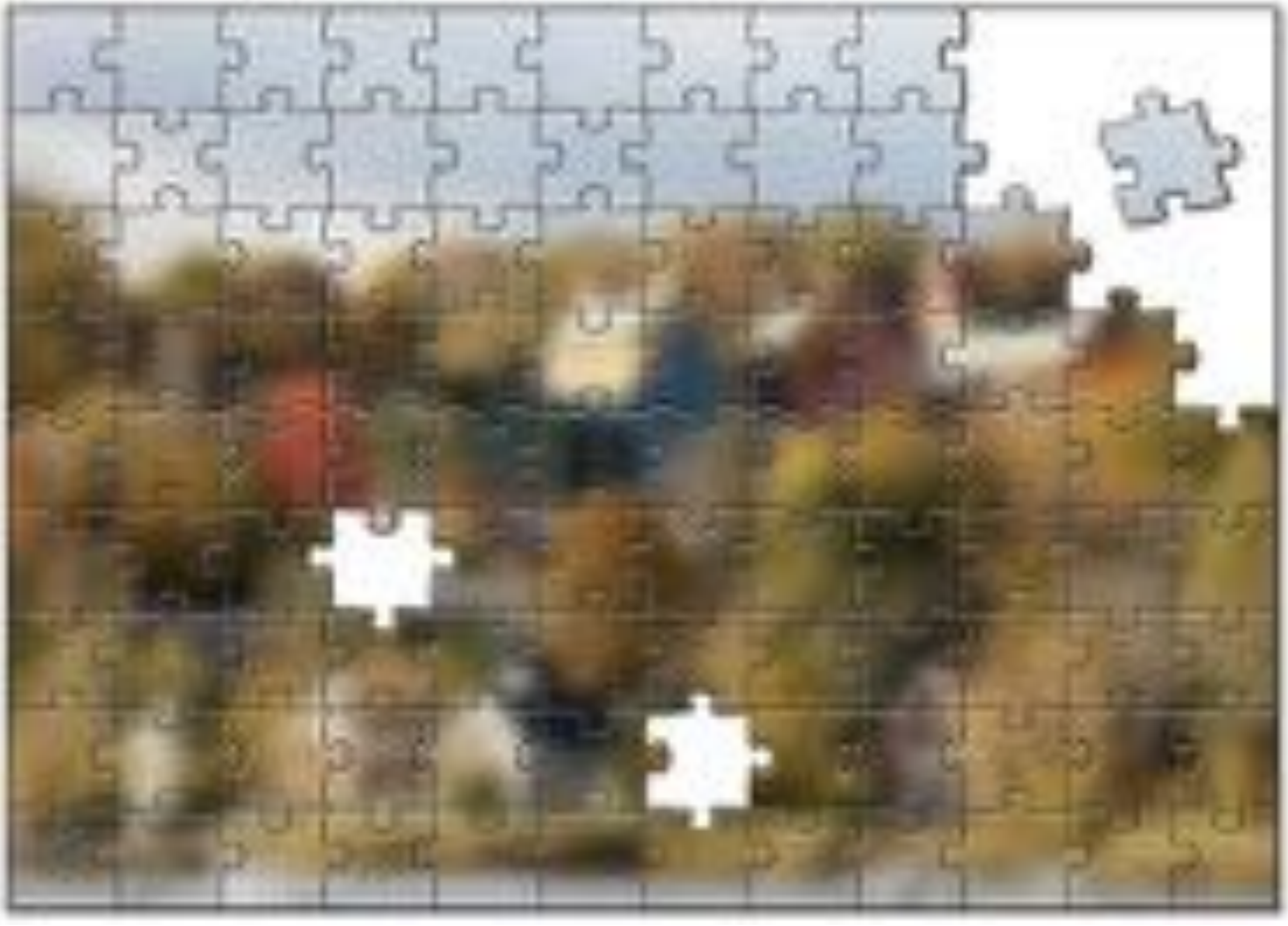
Imaging of fluorescently phospholinked labeled nucleotides as they are incorporated by a polymerase anchored to a Zero-Mode Waveguide (ZMW).



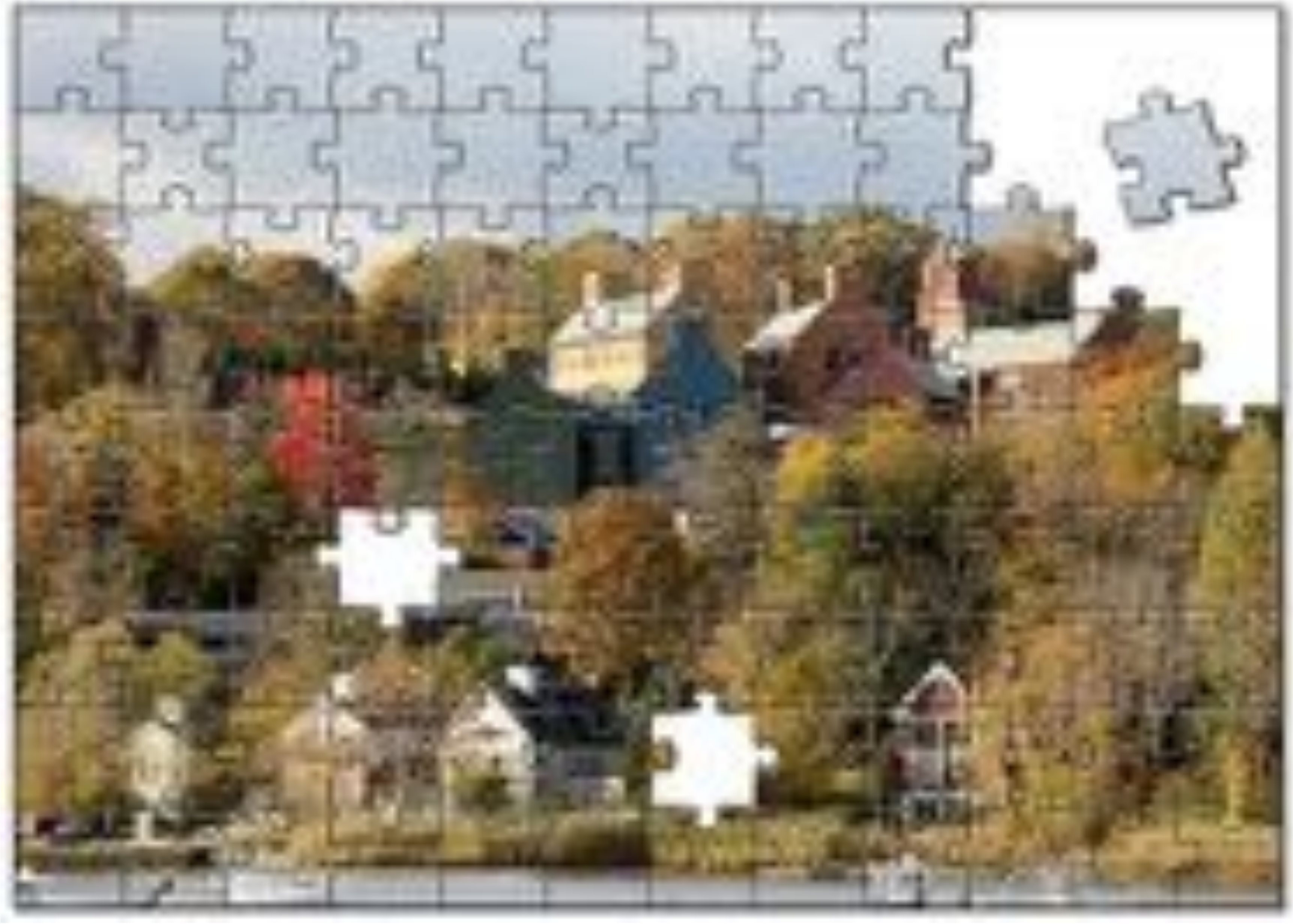


http://www.pacificbiosciences.com/assets/files/pacbio_technology_backgrounder.pdf

Single Molecule Sequences



“Corrective Lens” for Sequencing



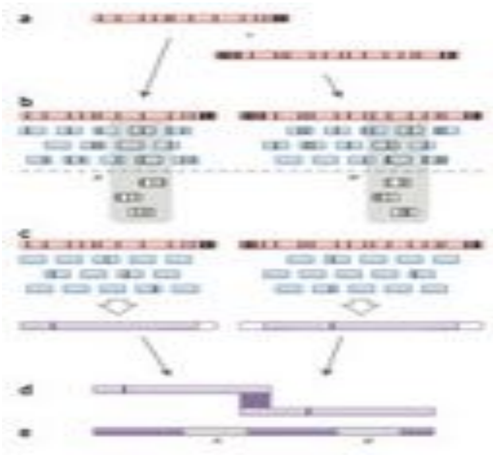
PacBio Assembly Algorithms

PBJelly

Gap Filling and Assembly Upgrade

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

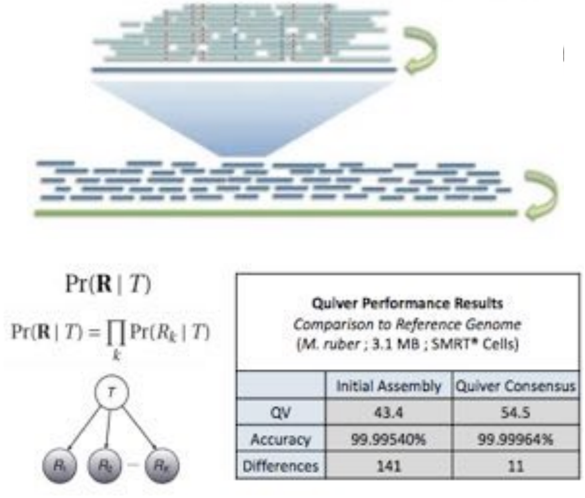
PacBioToCA & ECTools



Hybrid/PB-only Error Correction

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

HGAP/MHAP & 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

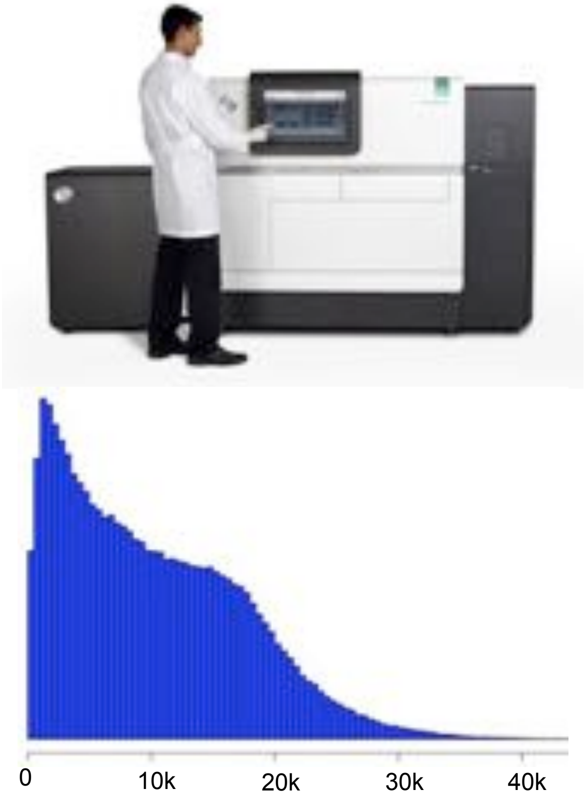
PB-only Correction & Polishing

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

< 5x PacBio Coverage > 50x

3rd Gen Long Read Sequencing

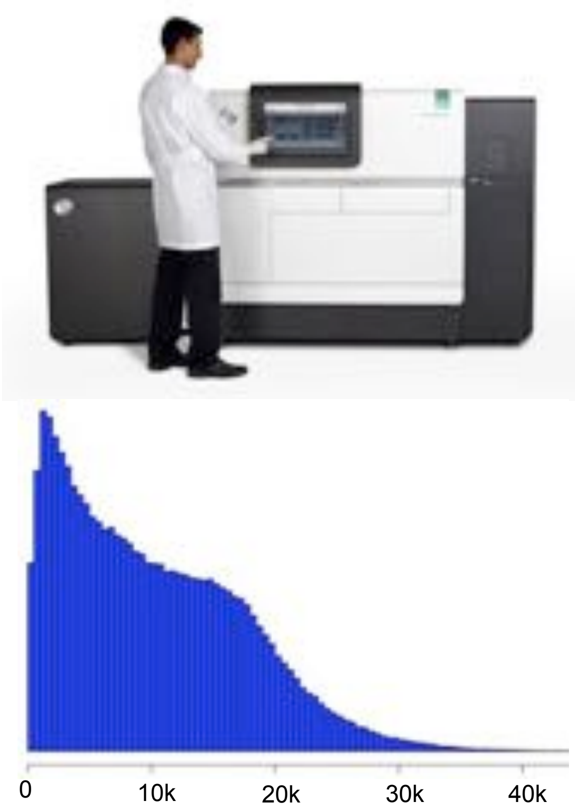
PacBio RS II



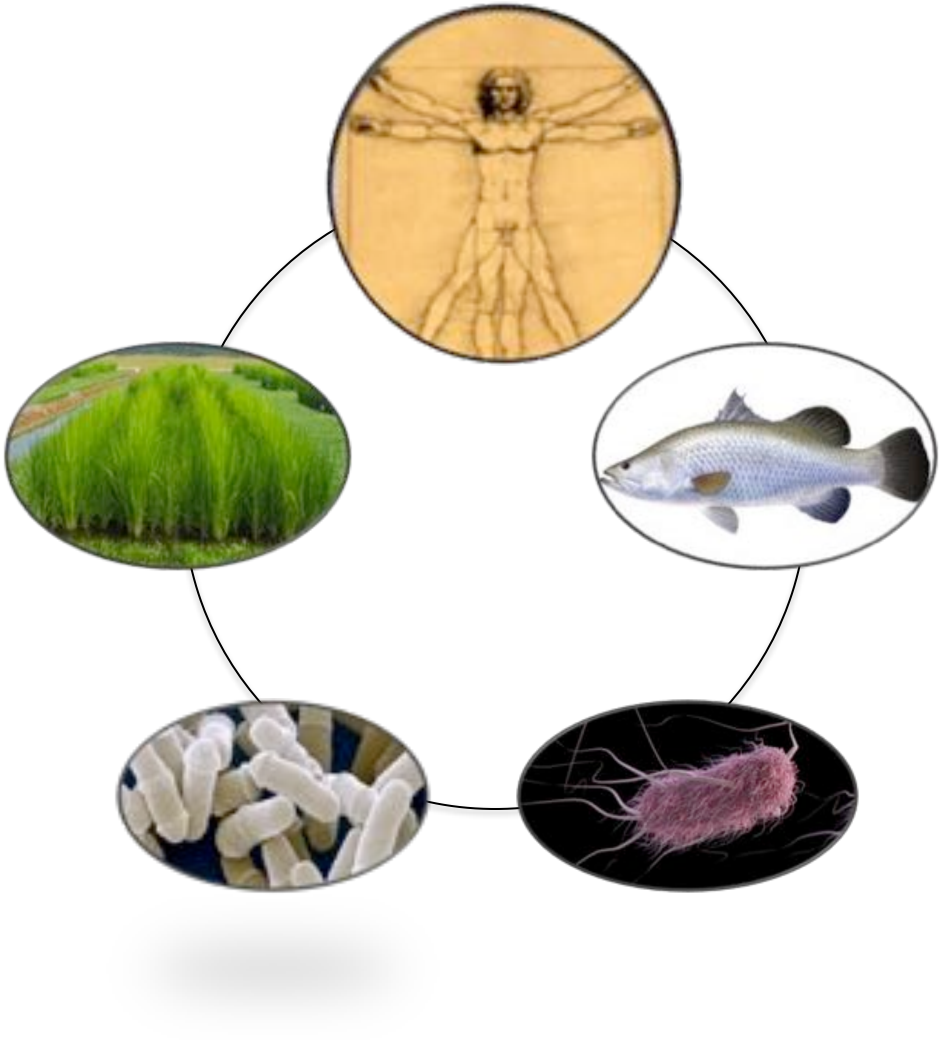
CSHL/PacBio

3rd Gen Long Read Sequencing

PacBio RS II

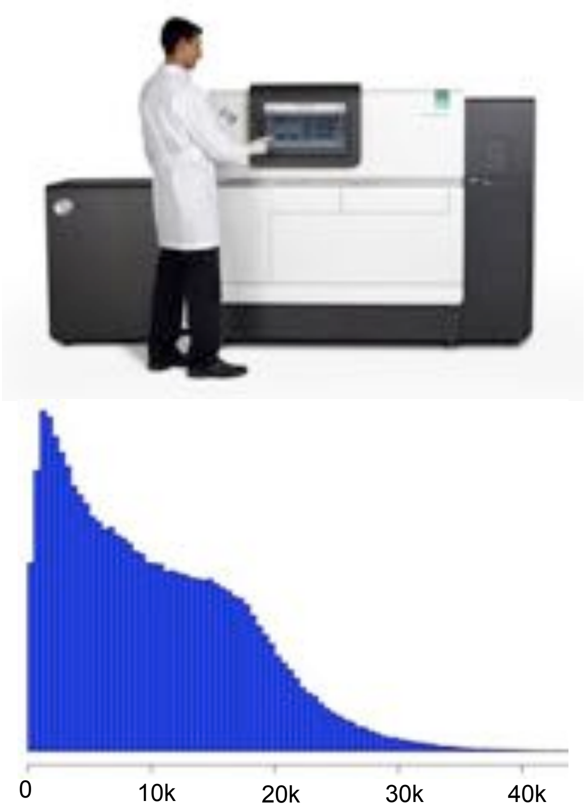


CSHL/PacBio

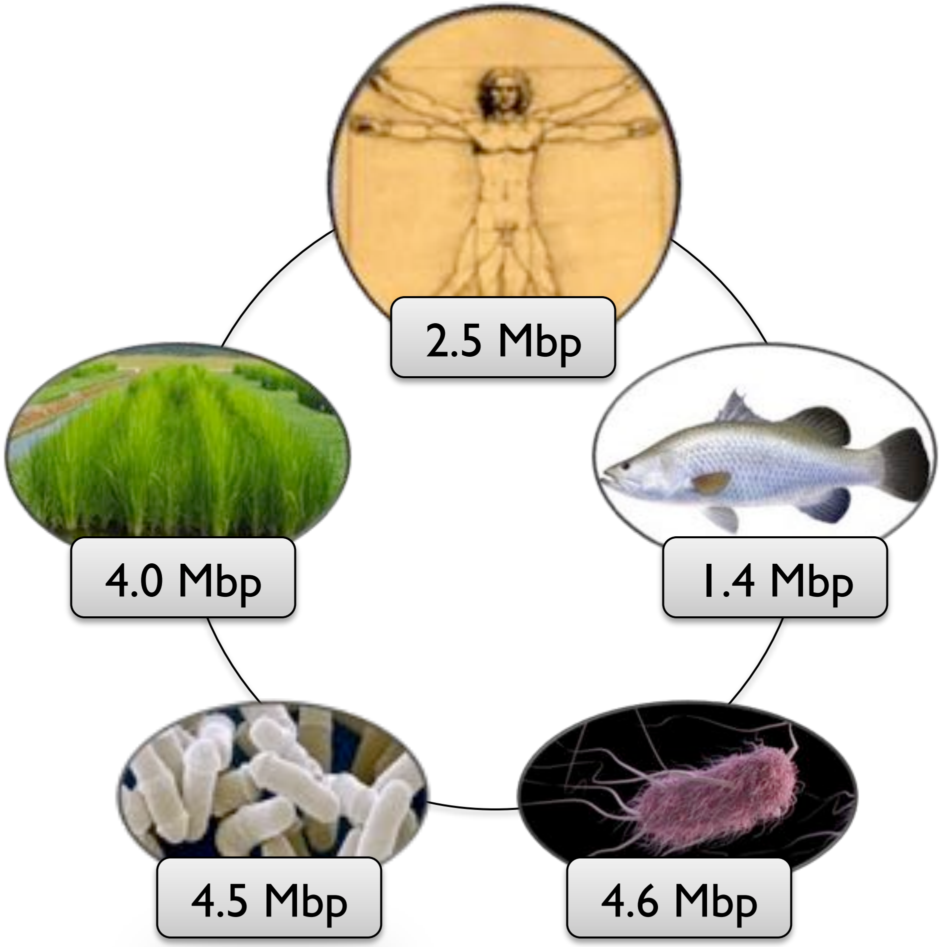


3rd Gen Long Read Sequencing

PacBio RS II



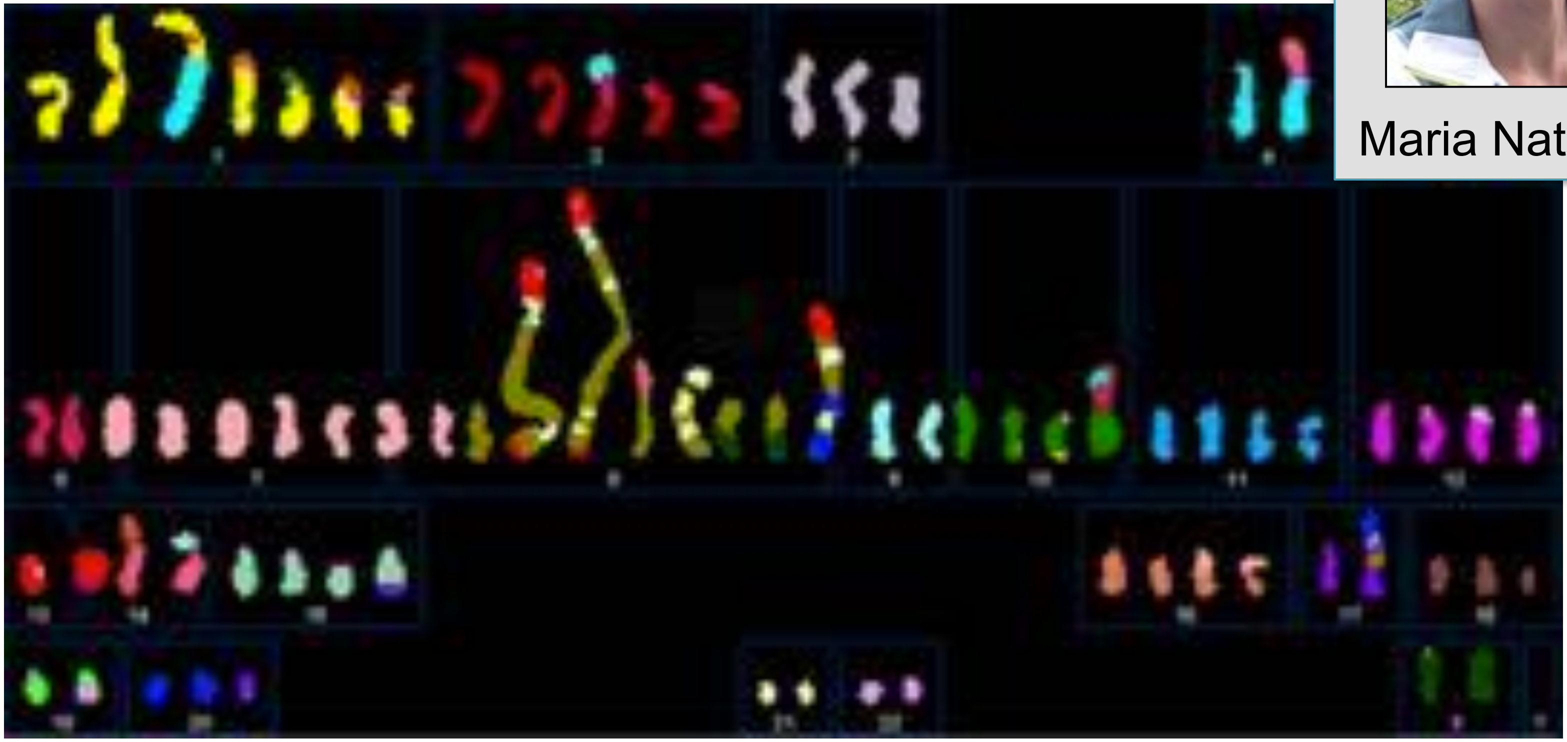
CSHL/PacBio



SK-BR-3

Maria Nattestad

Most commonly used Her2-amplified breast cancer

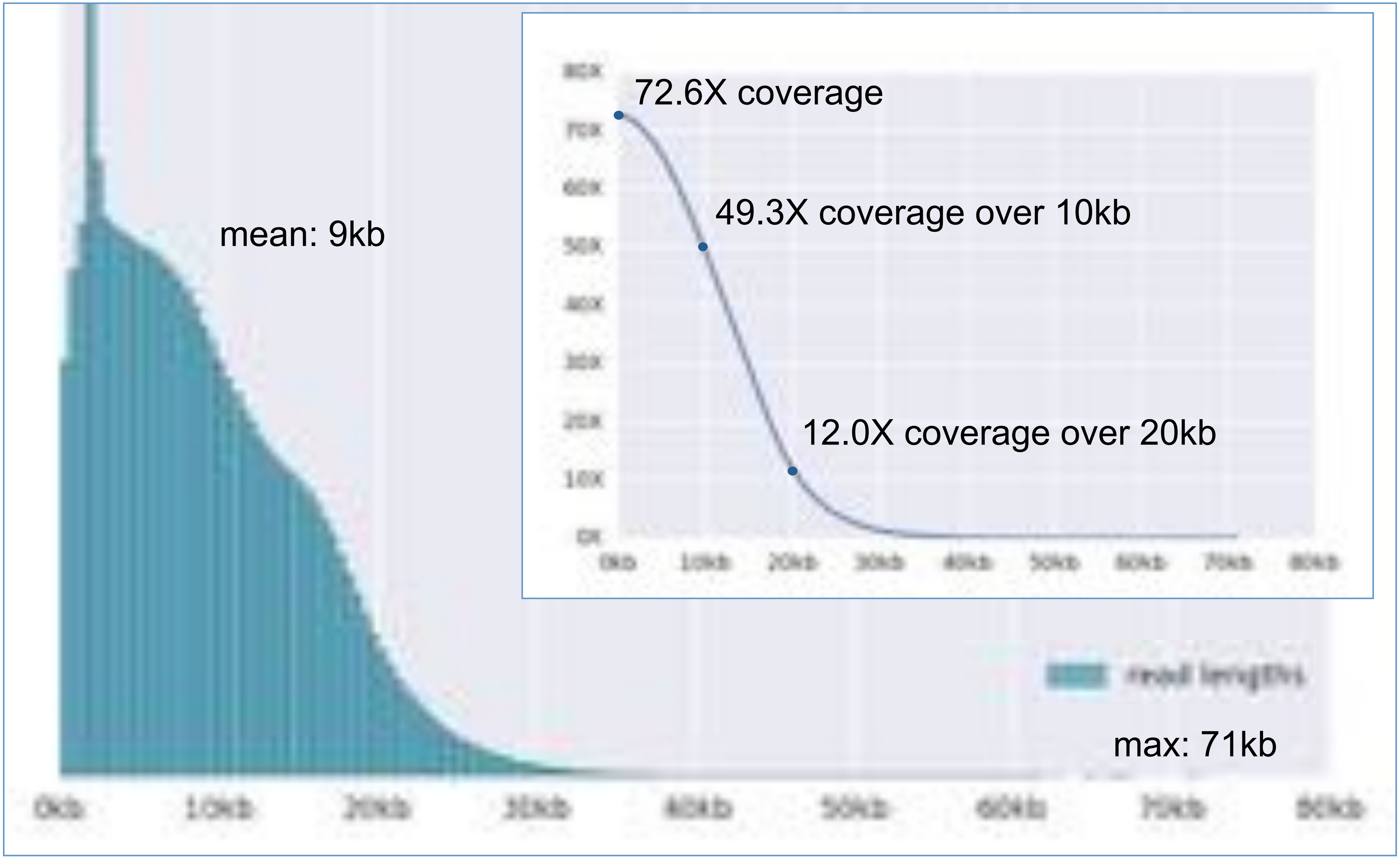


(Davidson et al, 2000)

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

PacBio read length distribution



mean: 9kb

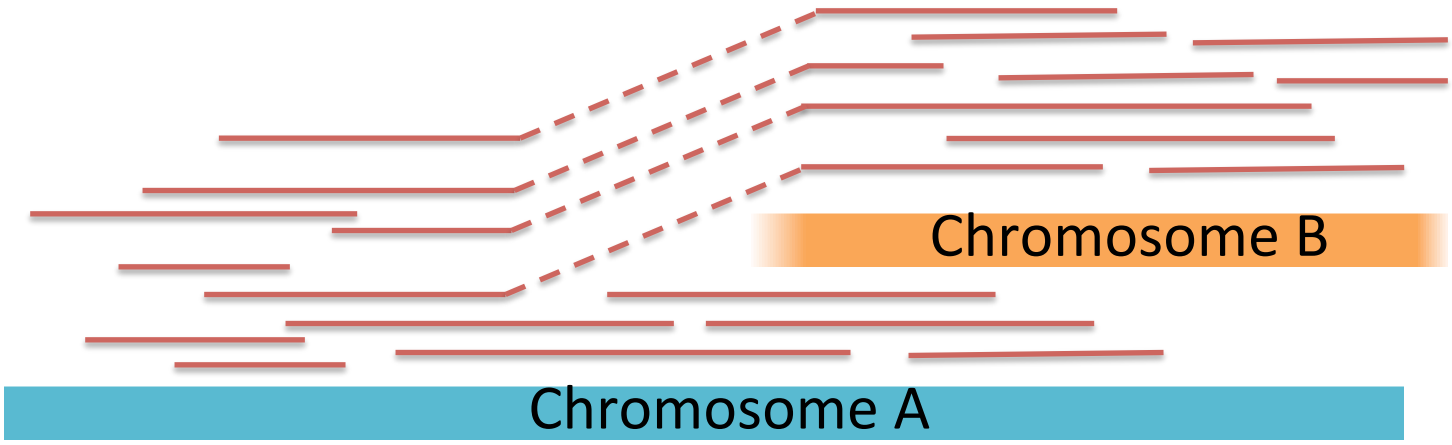
72.6X coverage

49.3X coverage over 10kb

12.0X coverage over 20kb

max: 71kb

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*

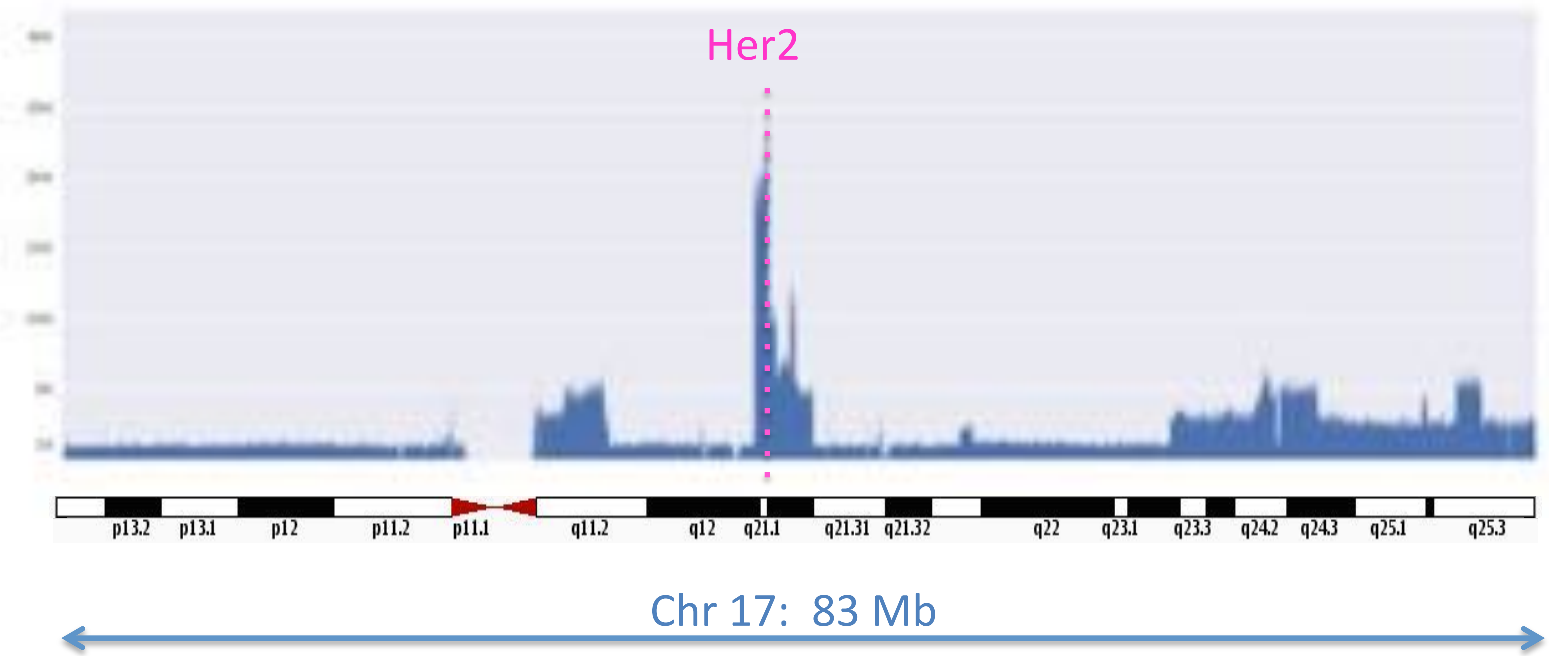
Localized HGAP + Celera Assembler + MUMmer

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

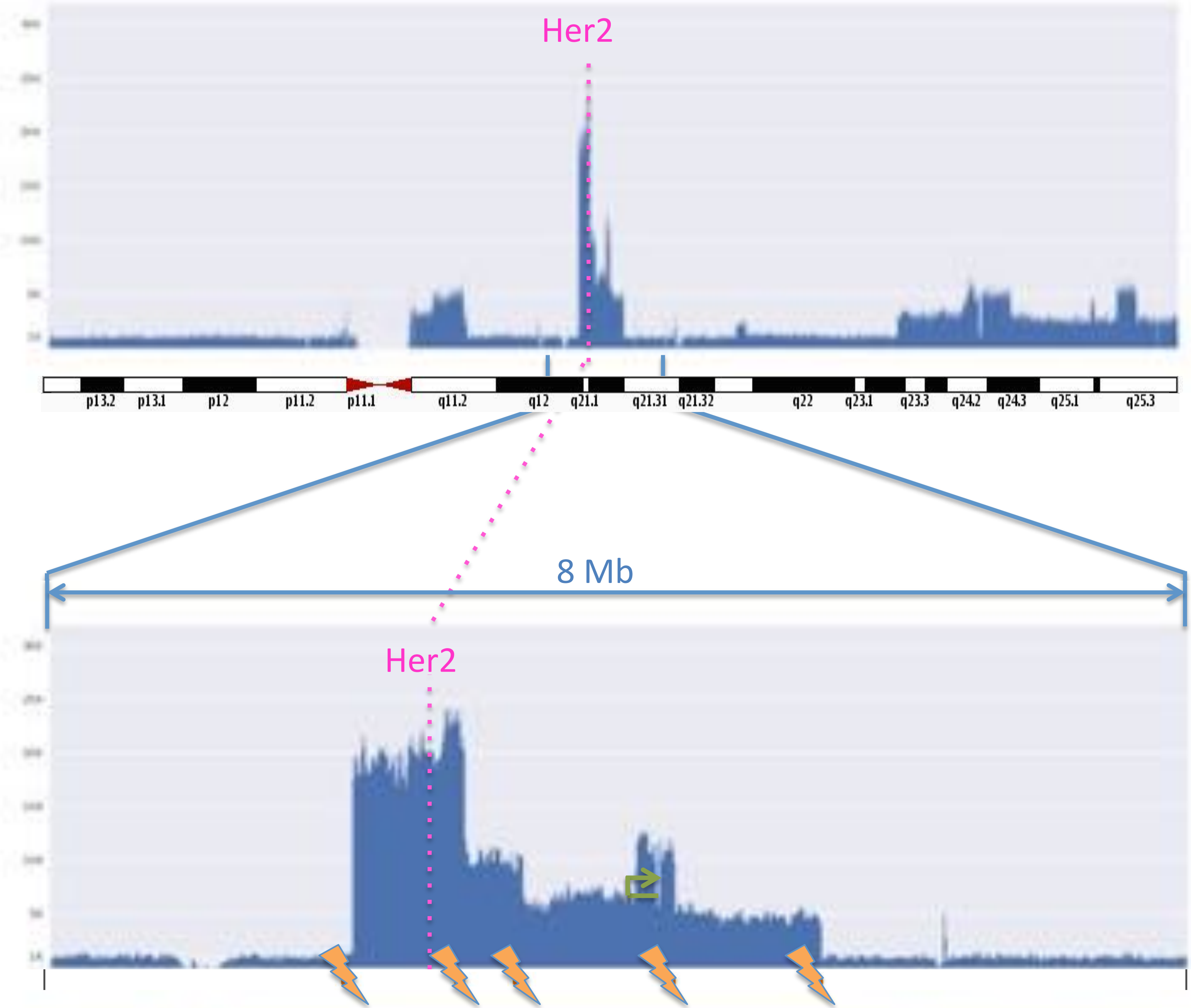
DNAnexus-enabled version of Falcon

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

PacBio

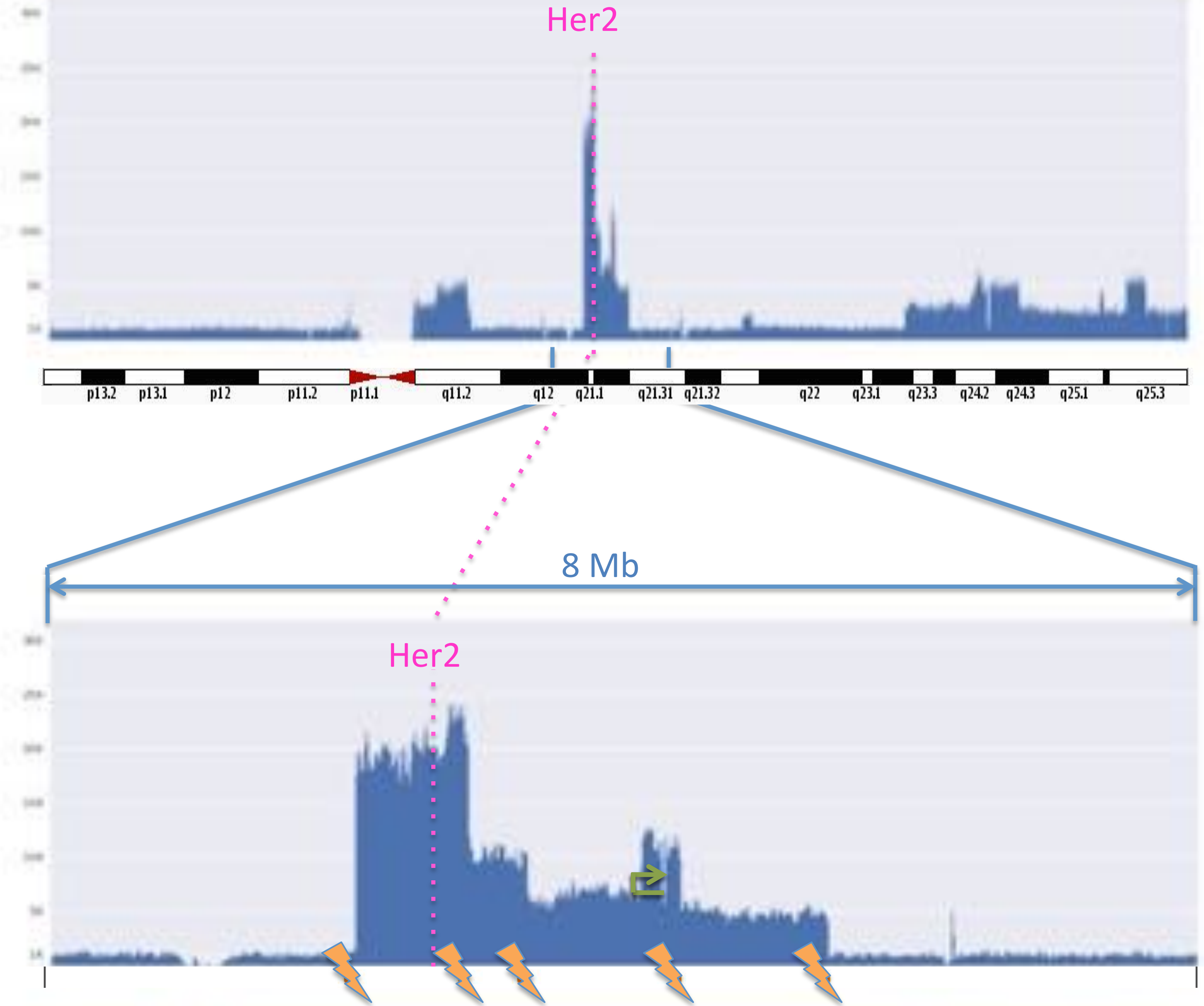

Her2
p13.2
p13.1
p12
p11.2
p11.1
q11.2
q12
q21.1
q21.31
q21.32
q22
q23.1
q23.3
q24.2
q24.3
q25.1
q25.3
Chr 17: 83 Mb

PacBio

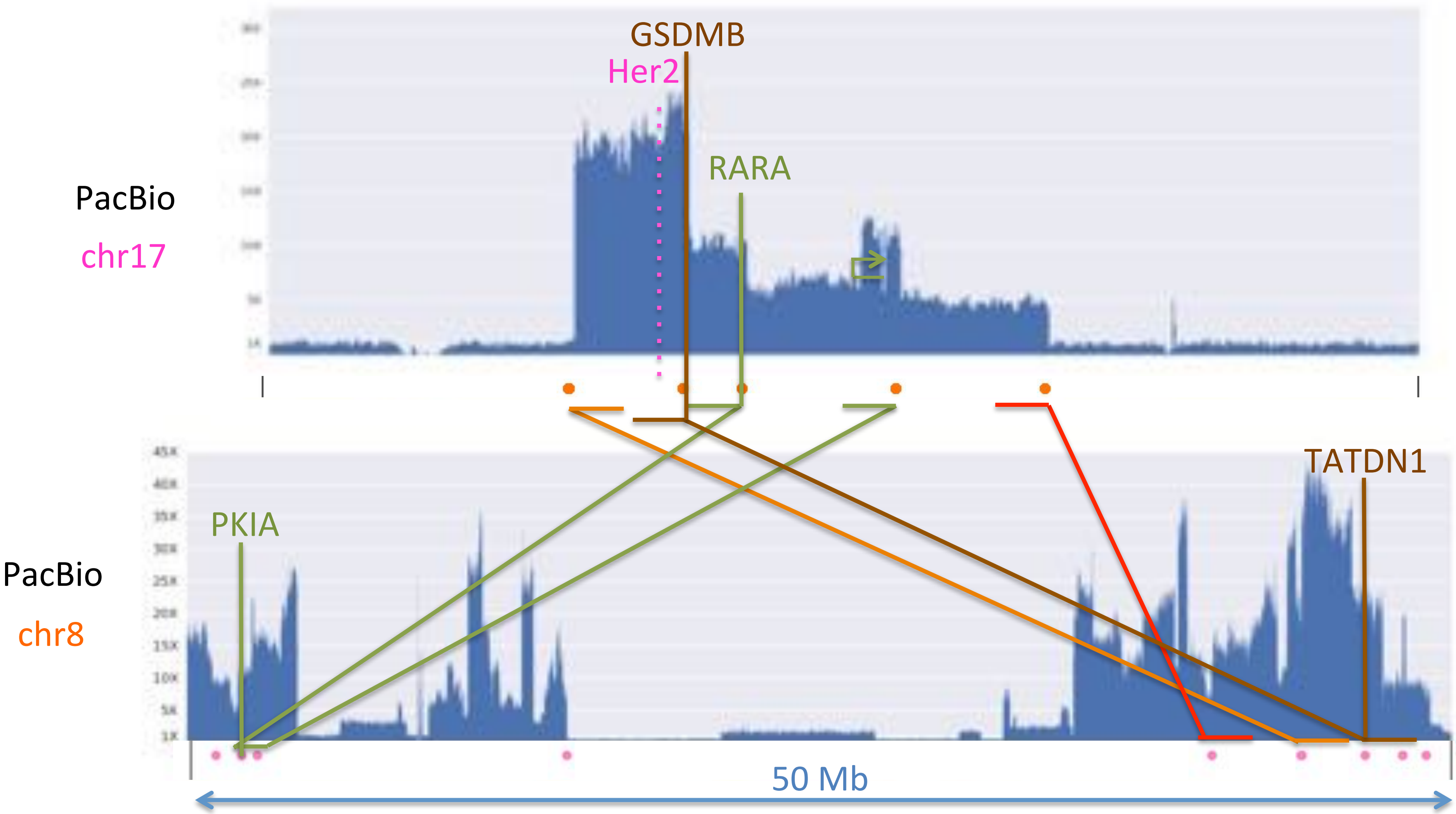


PacBio
chr17

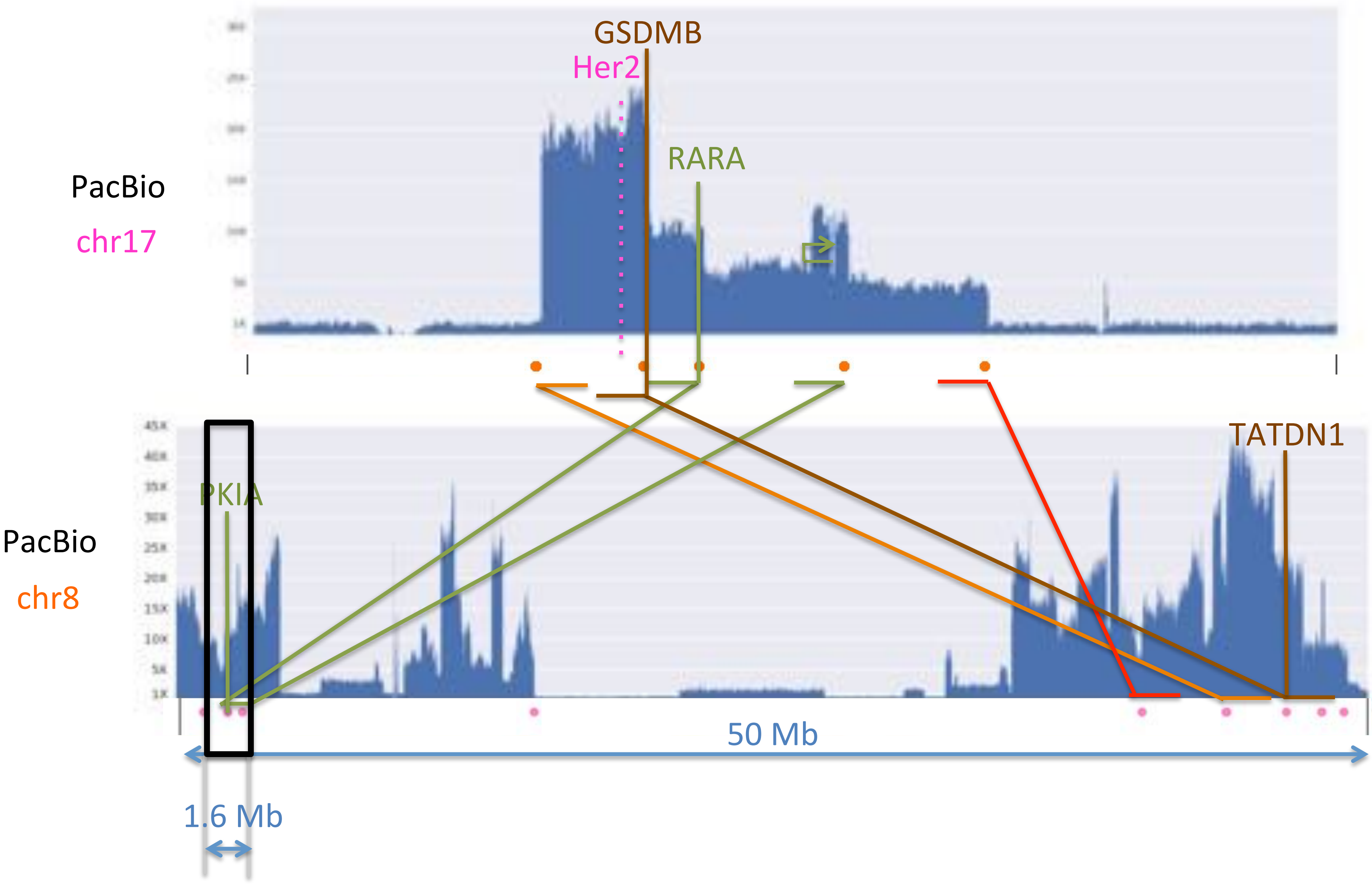
PacBio



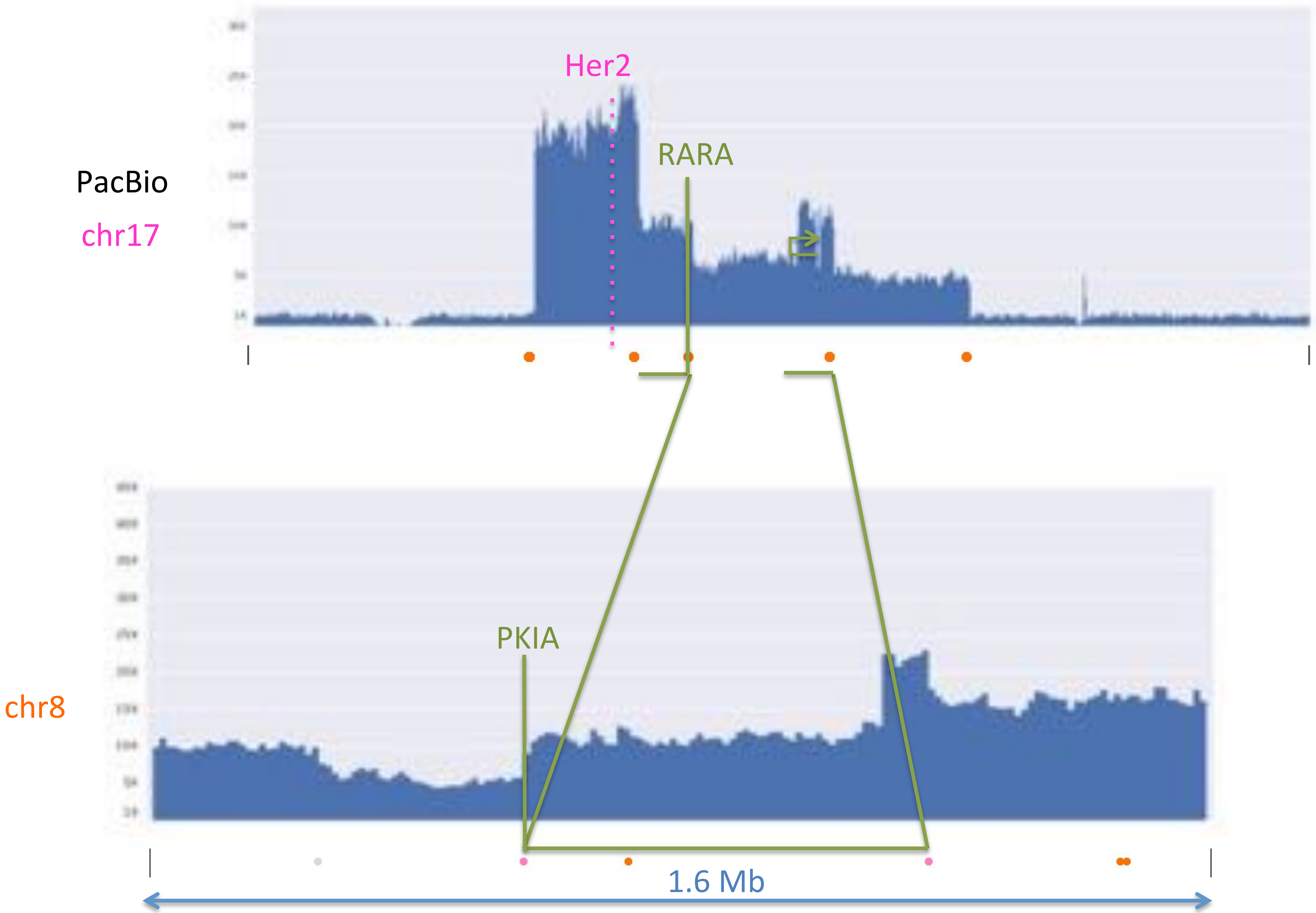
PacBio
chr17



Confirmed both known gene fusions in this region

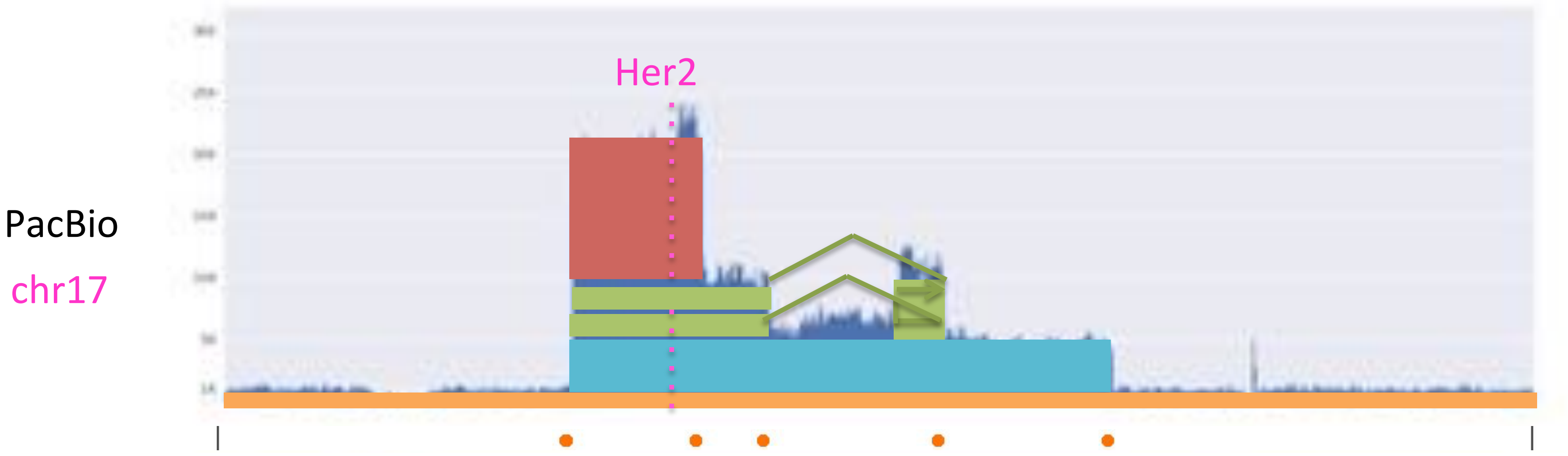


Confirmed both known gene fusions in this region



Joint coverage and breakpoint analysis to discover underlying events

Cancer lesion Reconstruction



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

Cancer lesion Reconstruction

Available *today* under the Toronto Agreement:

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

Available soon

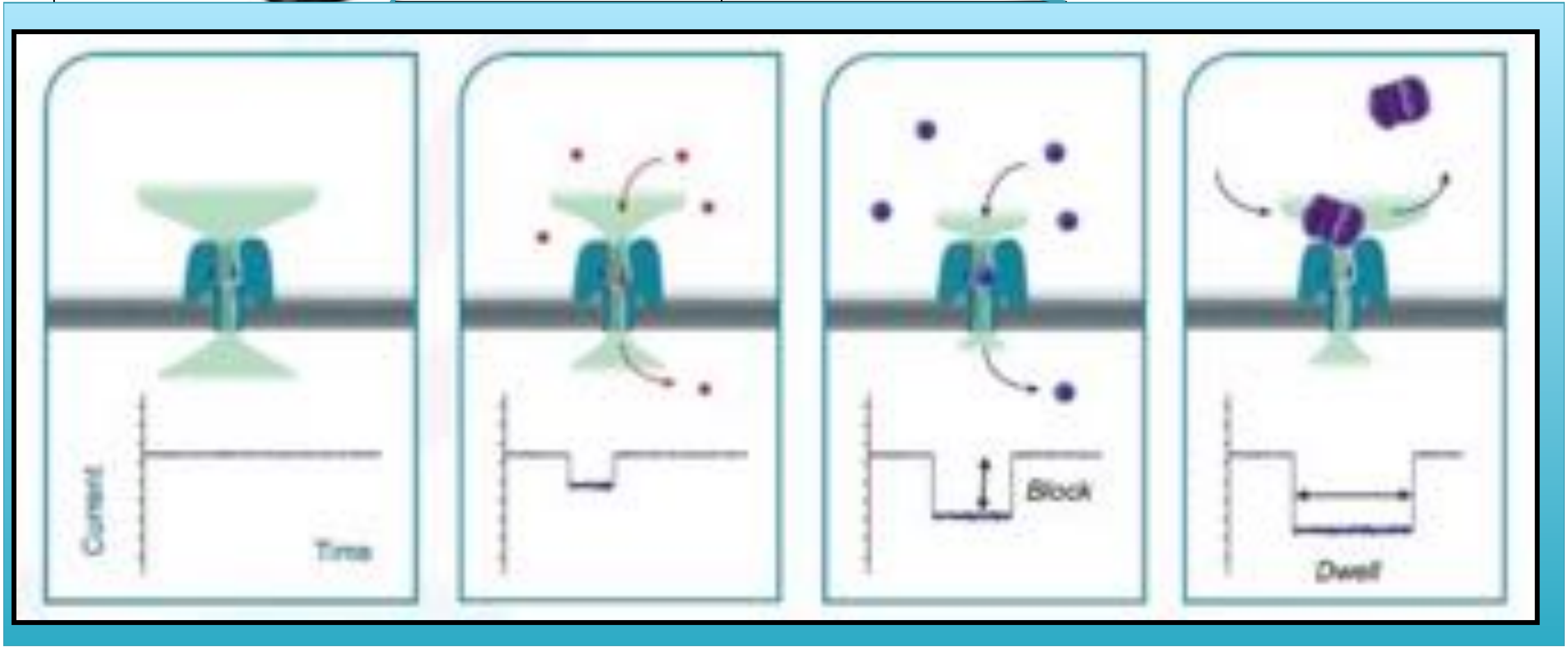
- Whole genome methylation analysis
- Full length cDNA transciptome analysis
- Comparison to single cell analysis of >100 individual cells

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

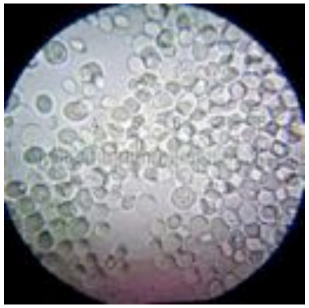
4. Final duplication from within chromosome 8

Oxford Nanopore MinION

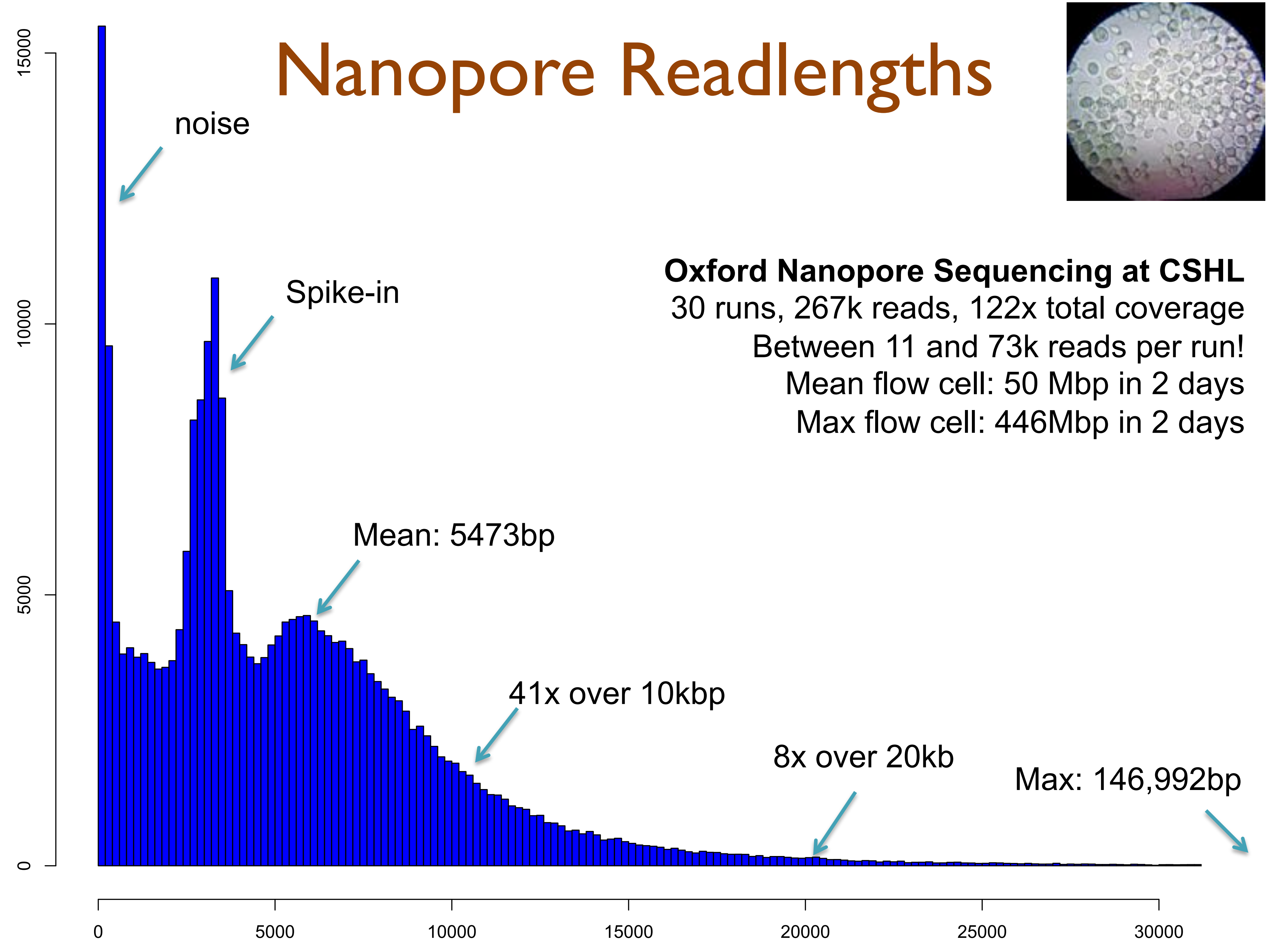
- Thumb drive sized sequencer powered over USB
- Capacity for 512 reads at once
- Senses DNA by measuring changes to ion flow



Nanopore 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



noise

Spike-in

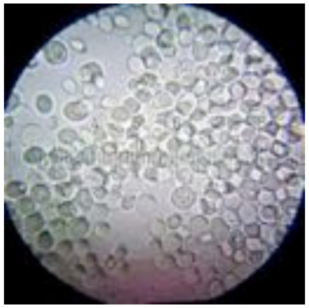
Mean: 5473bp

41x over 10kbp

8x over 20kb

Max: 146,992bp

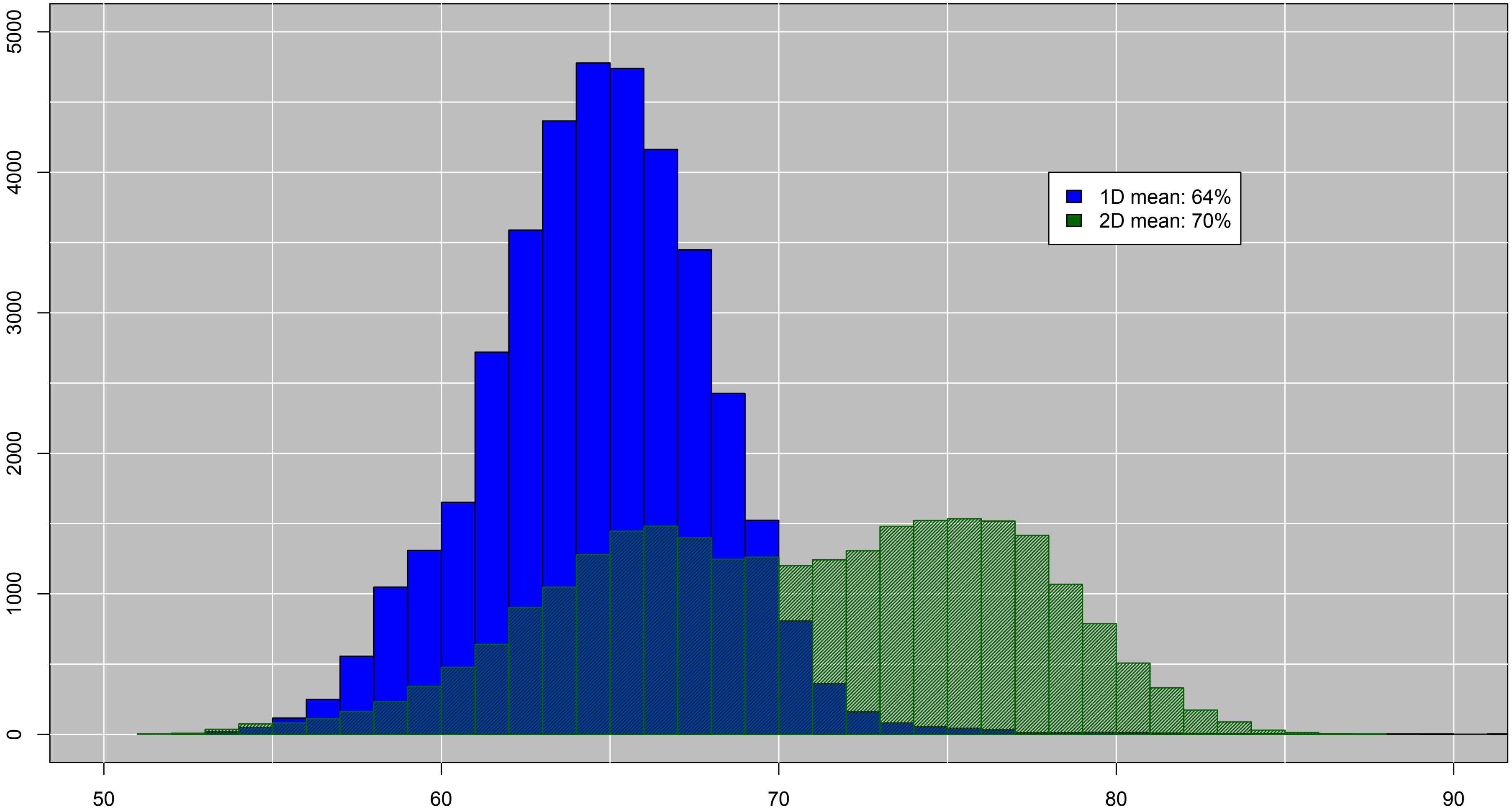
Nanopore Accuracy



Alignment Quality (BLASTN)

Of reads that align, average ~64% identity

“2D base-calling” improves to ~70% identity



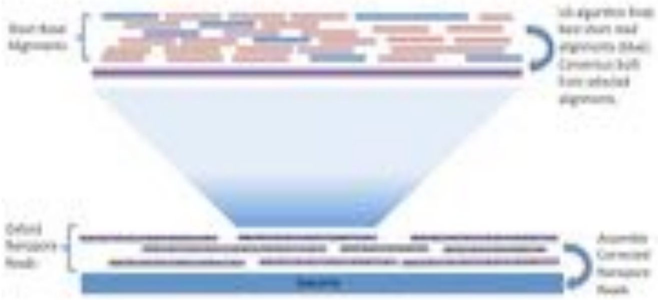
NanoCorr: Nanopore-Illumina Hybrid Error Correction

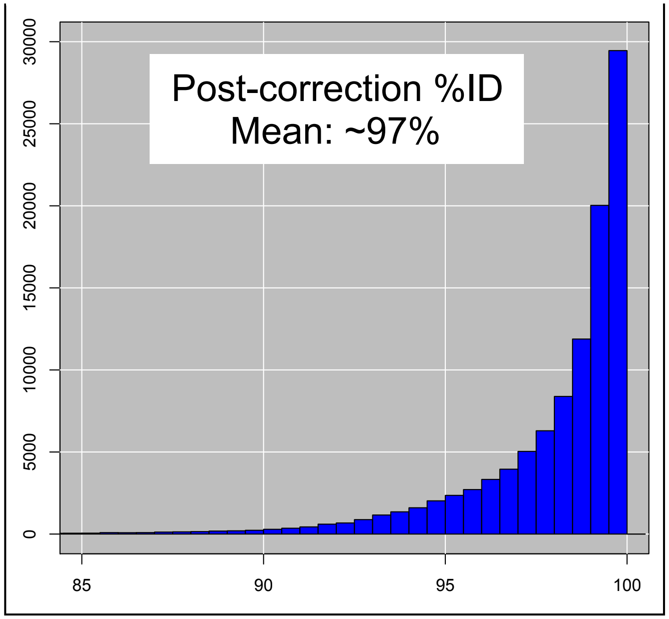
https://github.com/jgurtowski/nanocorr

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

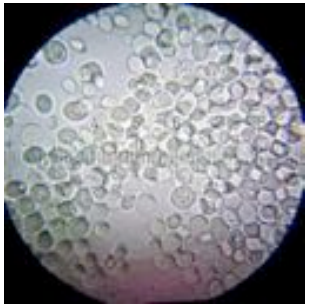
3. Compute consensus of each Oxford Nanopore read
   - ○ State machine of most commonly observed base at each position in read





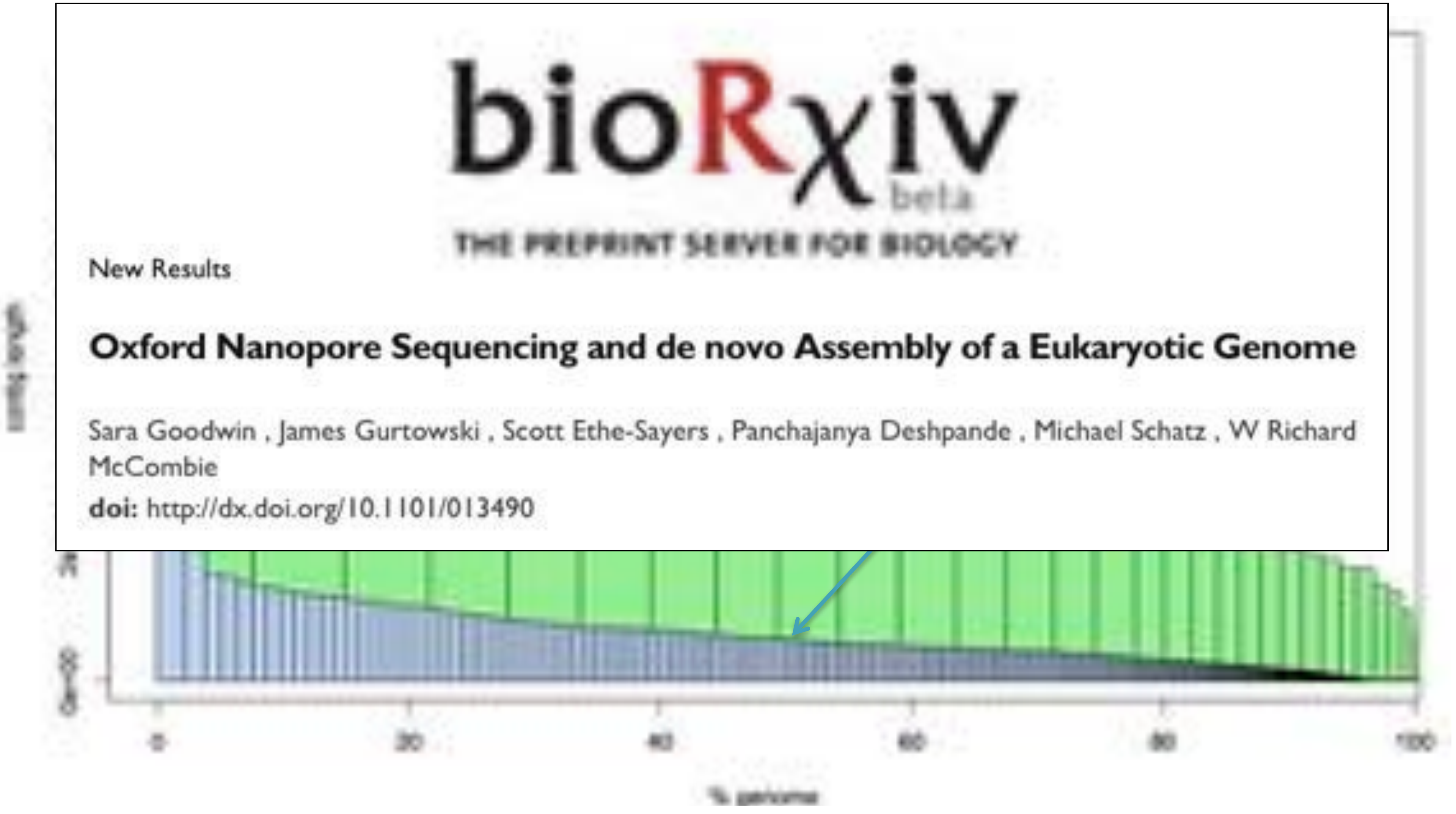
Oxford Nanopore Sequencing and de novo Assembly of a Eukaryotic Genome
Goodwin, S, Gurtowski, J *et al.* (2015) bioRxiv doi: http://dx.doi.org/10.1101/013490

NanoCorr Yeast Assembly



S288C Reference sequence

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



bioRxiv beta

THE PREPRINT SERVER FOR BIOLOGY

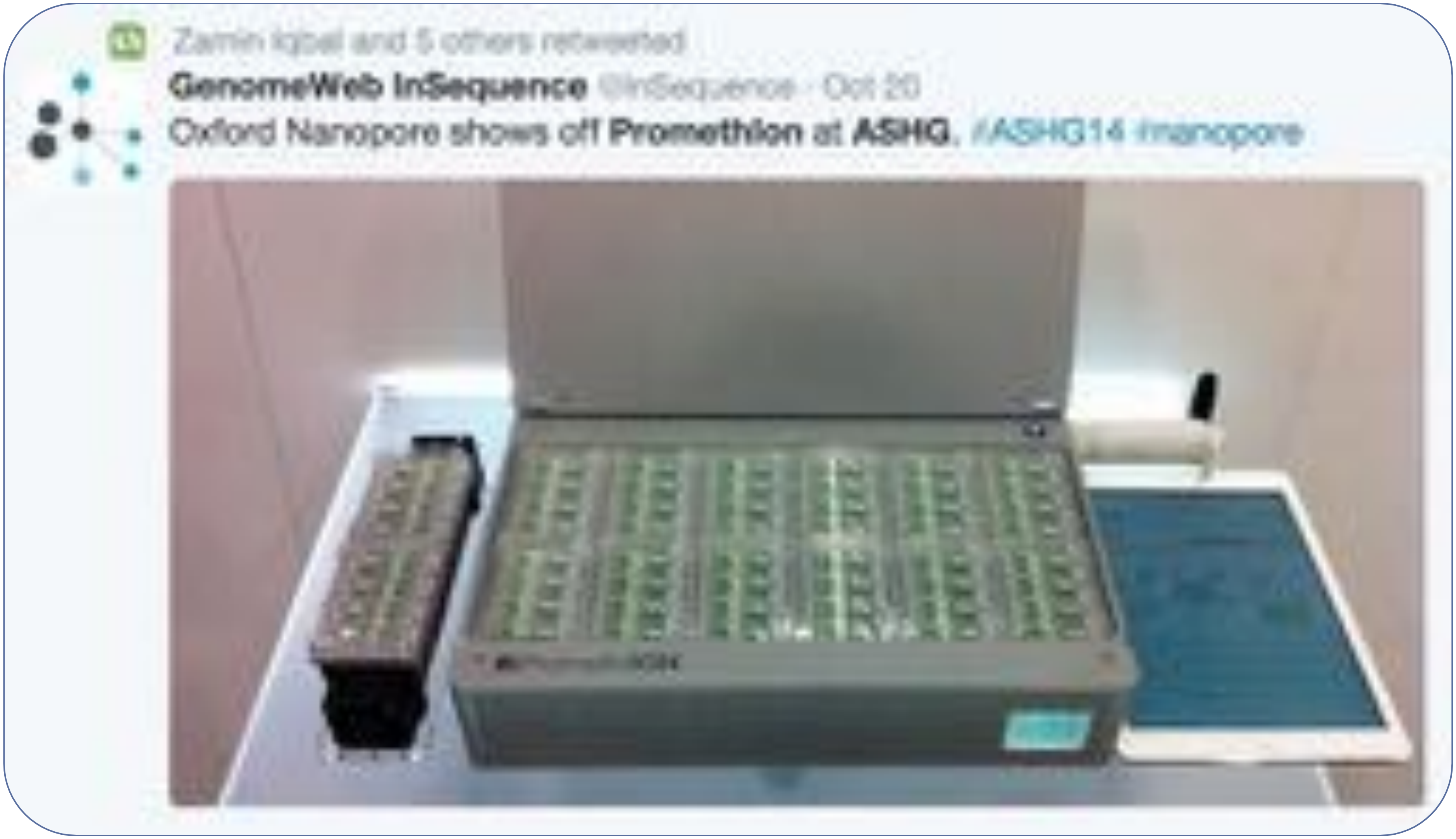
New Results

Oxford Nanopore Sequencing and de novo Assembly of a Eukaryotic Genome

Sara Goodwin , James Gurtowski , Scott Ethe-Sayers , Panchajanya Deshpande , Michael Schatz , W Richard McCombie

doi: http://dx.doi.org/10.1101/013490

Genomic Futures?



Zamin Iqbal and 5 others retweeted

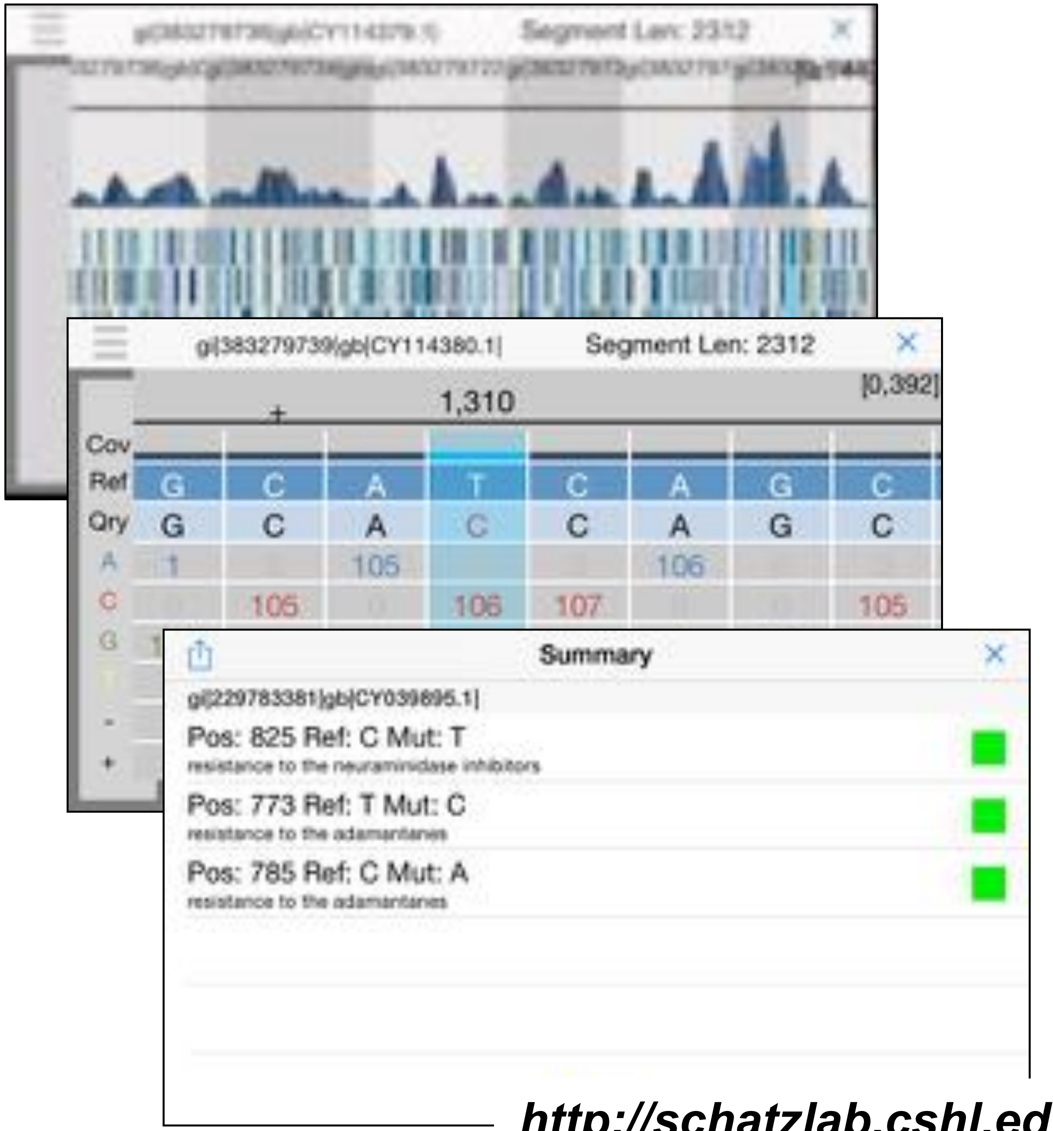
GenomeWeb InSequence @InSequence · Oct 20

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

Genomic Futures?

iGenomics: Mobile Sequence Analysis

Aspyn Palatnick, Elodie Ghedin, Michael Schatz



gi|383279739|gb|CY114380.1| Segment Len: 2312

1,310 [0,392]

Cov								
Ref	G	C	A	T	C	A	G	C
Qry	G	C	A	C	C	A	G	C
A	1		105			106		
C		105		106	107			105

Summary

gi|229783381|gb|CY039895.1|

Pos: 825 Ref: C Mut: T
resistance to the neuraminidase inhibitors

Pos: 773 Ref: T Mut: C
resistance to the adamantanes

Pos: 785 Ref: C Mut: A
resistance to the adamantanes

The worlds first genomics analysis app for iOS devices

BWT + Dynamic Programming + UI

First application:

- Handheld diagnostics and therapeutic recommendations for influenza infections

- In the iOS AppStore now!

Future applications

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

http://schatzlab.cshl.edu/iGenomics

What should we expect from an assembly?

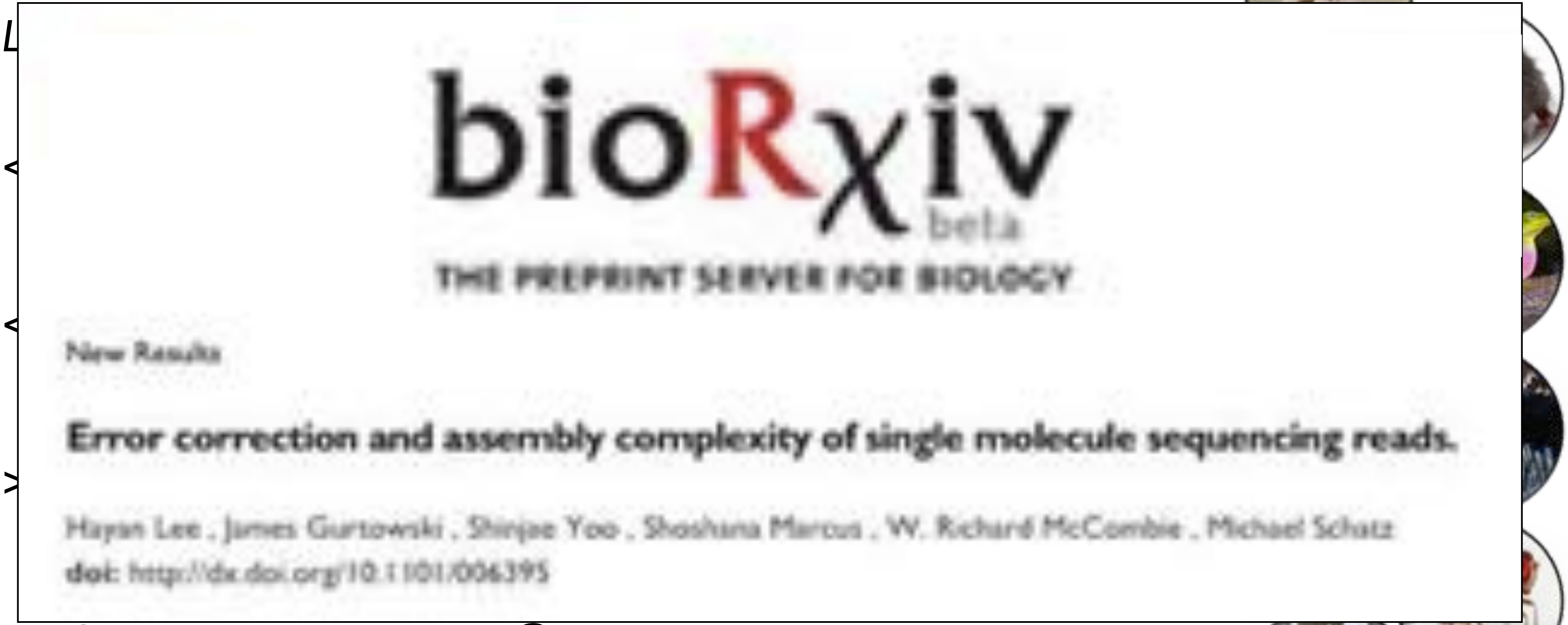
Summary & Recommendations

L

<

<

>



bioRχiv 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

> 5GB: Email mschatz@cshl.edu

C

The year 2015 will mark the return to reference quality genome sequence

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

SBU
Skiena Lab
Patro Lab

Cornell
Susan McCouch
Lyza Maron
Mark Wright

OICR
John McPherson
Karen Ng
Timothy Beck
Yogi Sundaravadanam

NYU
Jane Carlton
Elodie Ghedin

Oxford NANOPORE Technologies

PACIFIC BIOSCIENCES®

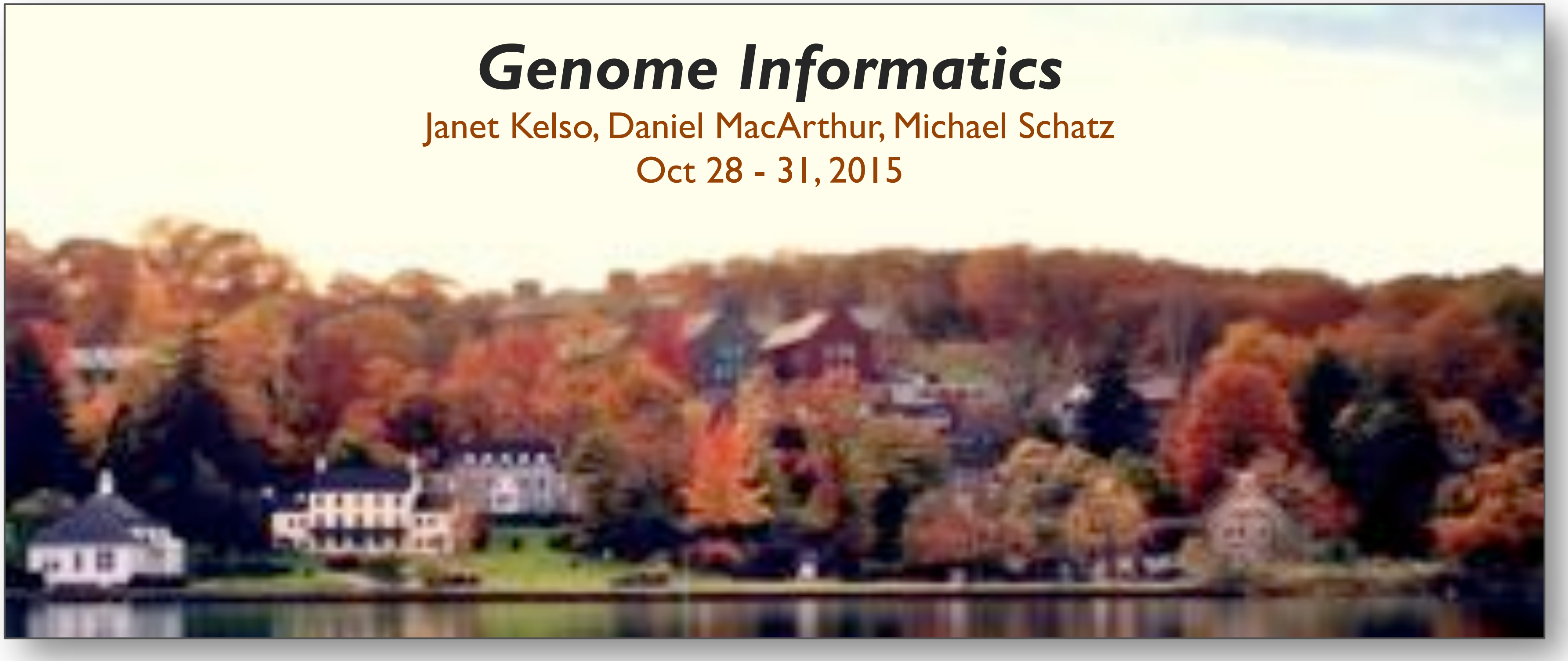
NSF

National Human Genome Research Institute

U.S. DEPARTMENT OF ENERGY

SFARI
SIMONS FOUNDATION AUTISM RESEARCH INITIATIVE

ALFRED P. SLOAN FOUNDATION



Genome Informatics

Janet Kelso, Daniel MacArthur, Michael Schatz

Oct 28 - 31, 2015

Thank you

http://schatzlab.cshl.edu

@mike_schatz