Single Cell and Single Molecule Analysis of Cancer

Michael Schatz

April 20, 2015

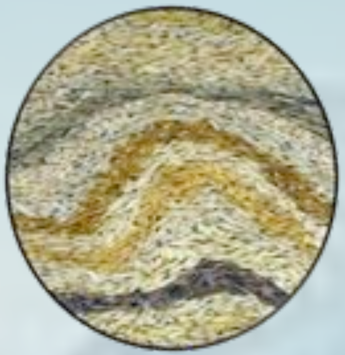
Laufer Center Retreat

Schatzlab Overview

Human Genetics

Role of mutations in disease

Narzisi *et al.* (2014)
Iossifov *et al.* (2014)



Plant Biology

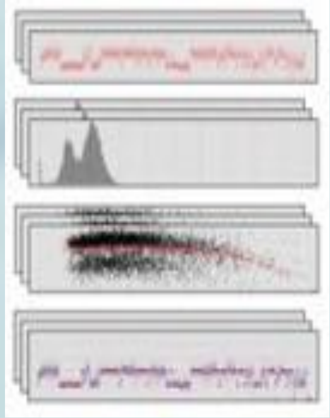
Genomes & Transcriptomes

Schatz *et al.* (2014)
Ming *et al.* (2013)

Algorithmics & Systems Research

Ultra-large scale biocomputing

Marcus *et al.* (2014)
Schatz *et al.* (2013)



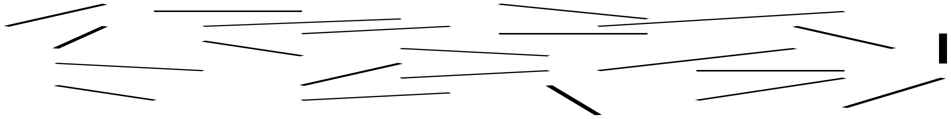
Single Cell & Single Molecule

CNVs, SVs, & Cell Phylogenetics

Garvin *et al.* (2014)
Roberts *et al.* (2013)

Sequence Assembly Problem

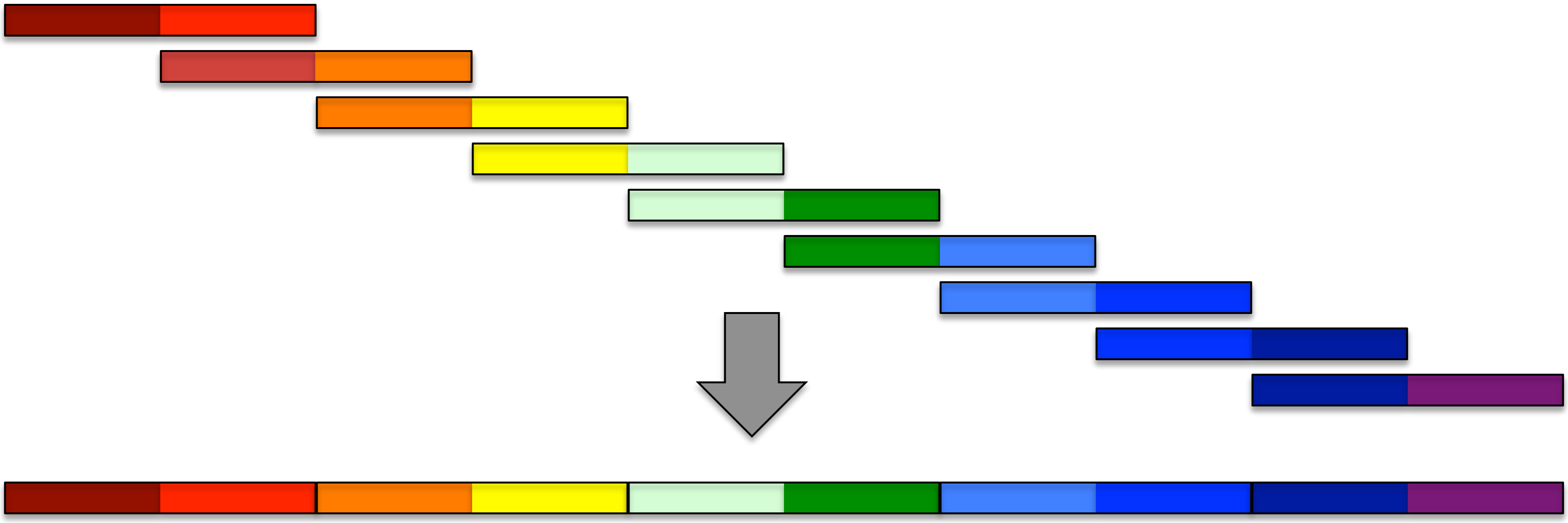
1. Shear & Sequence DNA



2. Construct assembly graph from overlapping reads

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

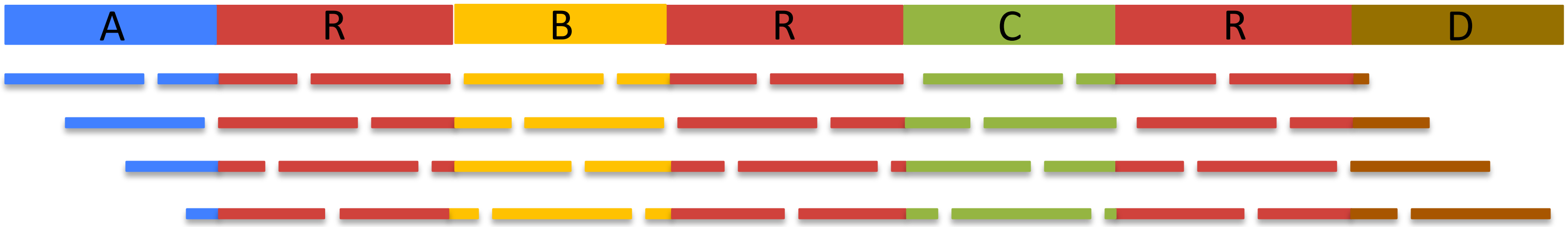
3. Simplify assembly graph

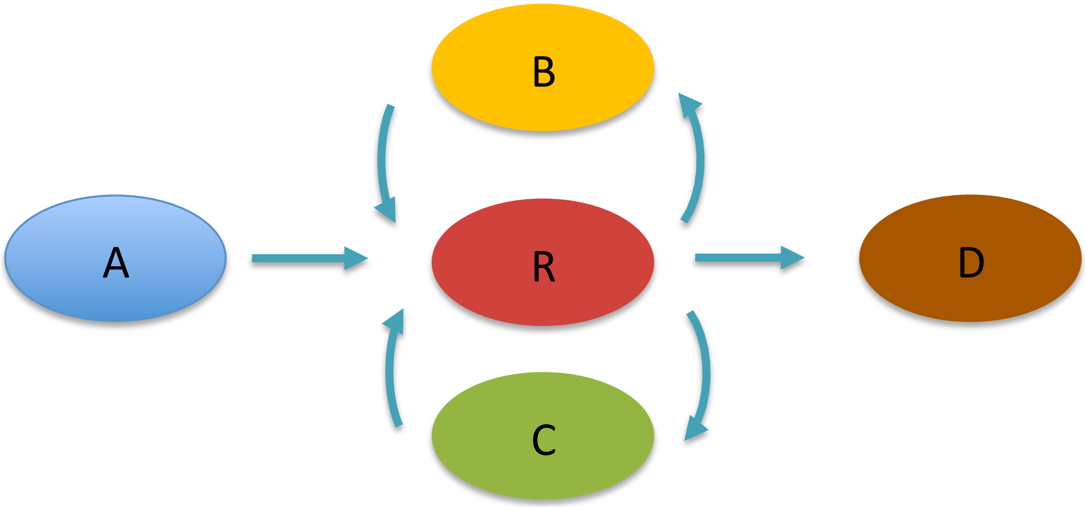


On Algorithmic Complexity of Biomolecular Sequence Assembly Problem

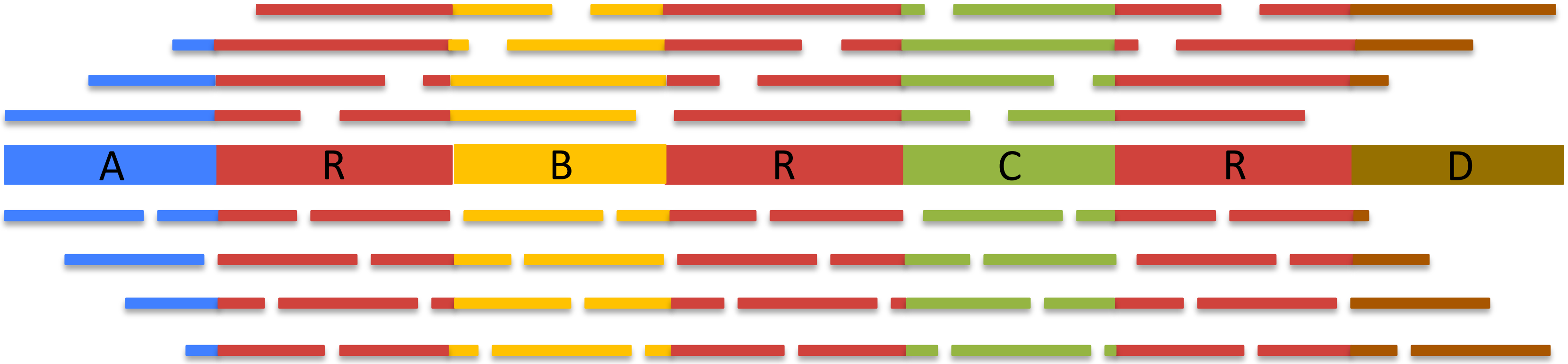
Narzisi, G, Mishra, B, Schatz, MC (2014) *Algorithms for Computational Biology.* Lecture Notes in Computer Science. *Vol. 8542*

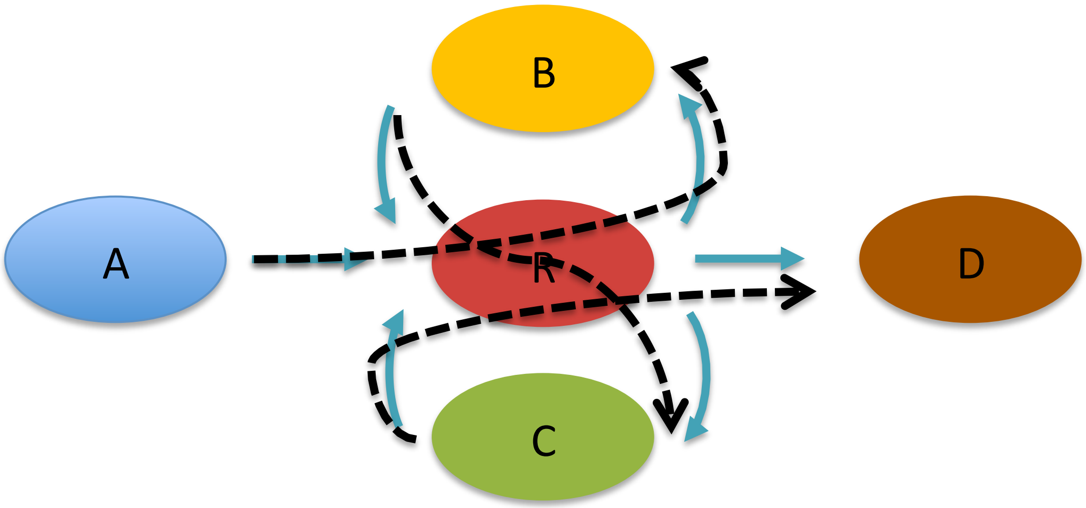
Assembly Complexity



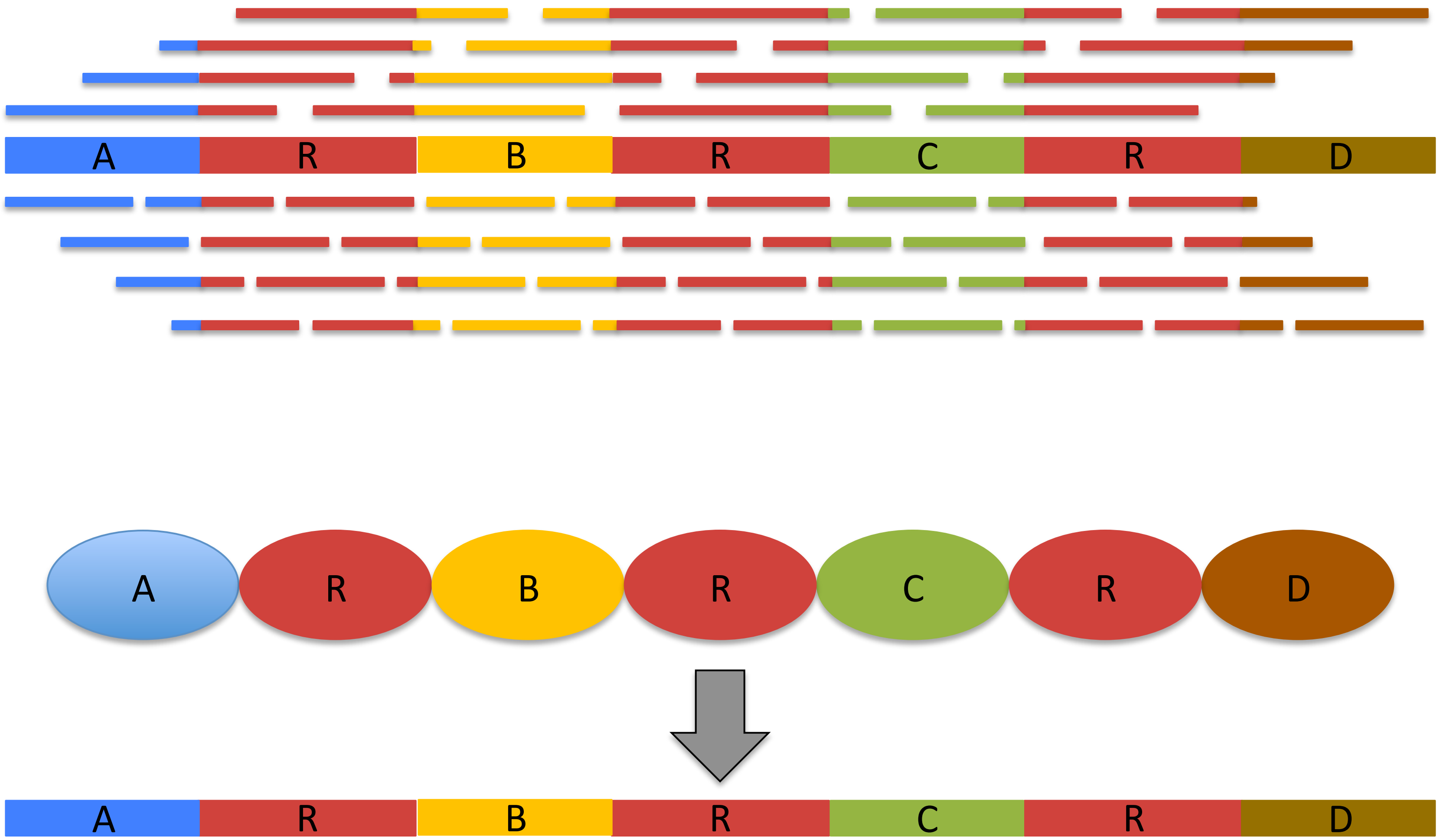


Assembly Complexity





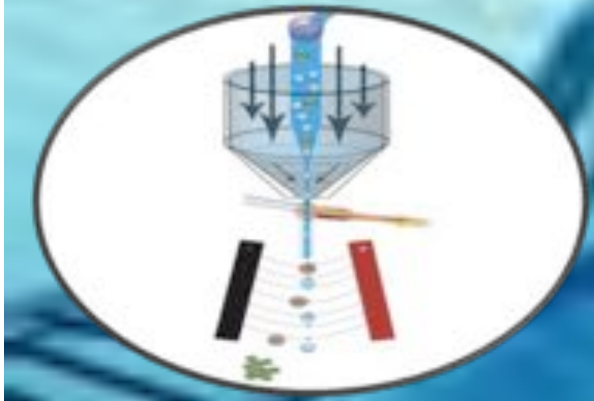
Assembly Complexity



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

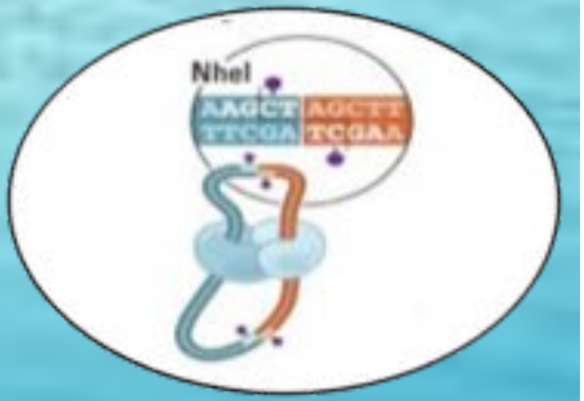
Genomics Arsenal in the year 2015

Sample Preparation



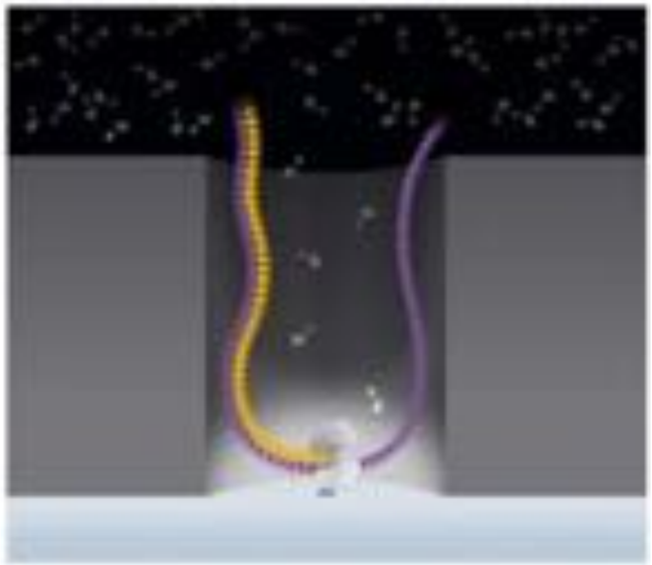
Sequencing

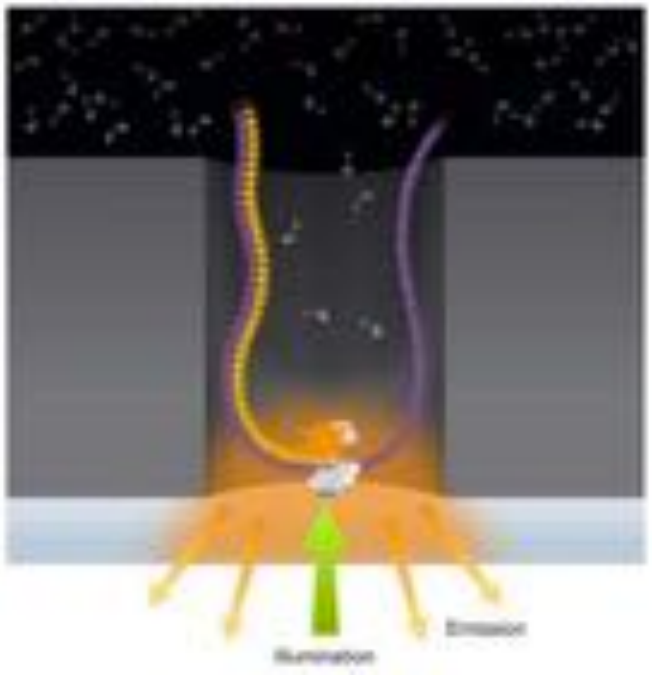
Chromosome Mapping

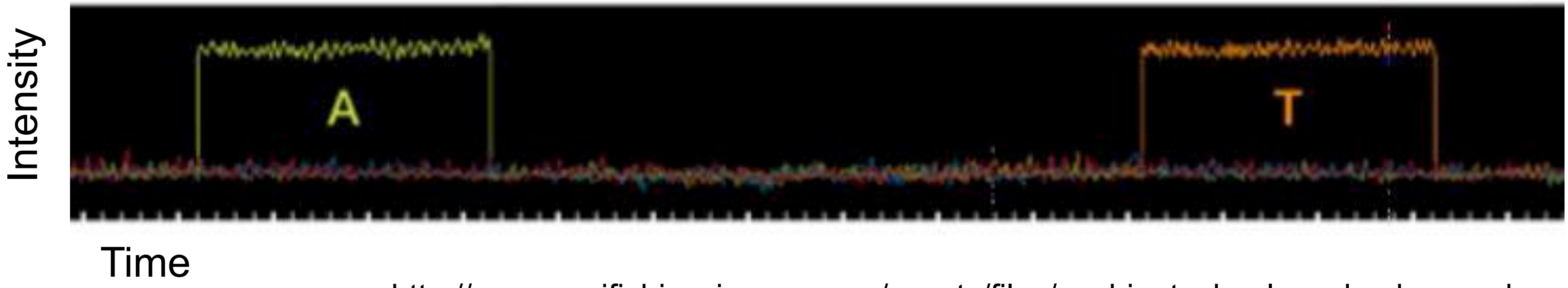


PacBio SMRT Sequencing

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

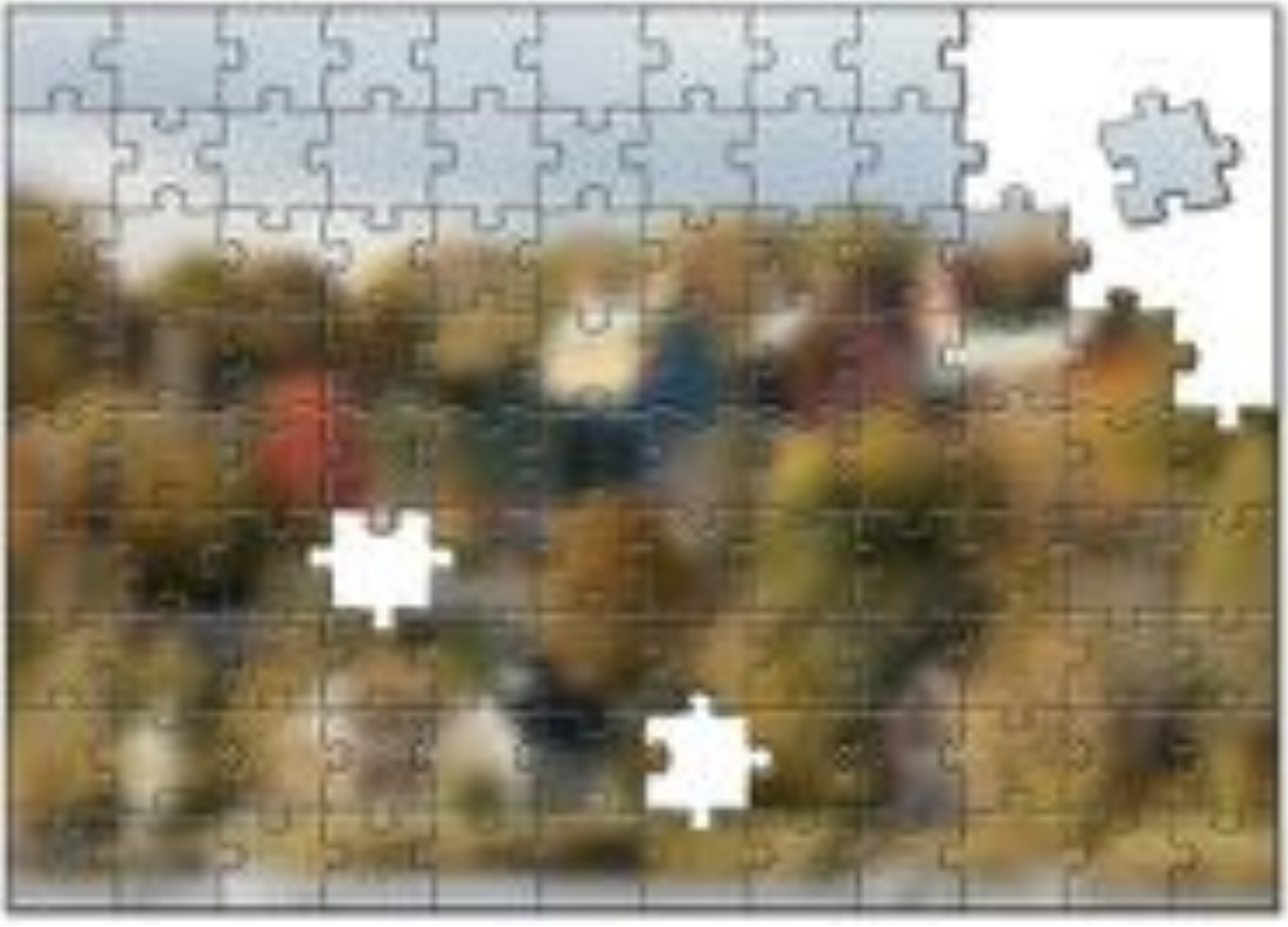




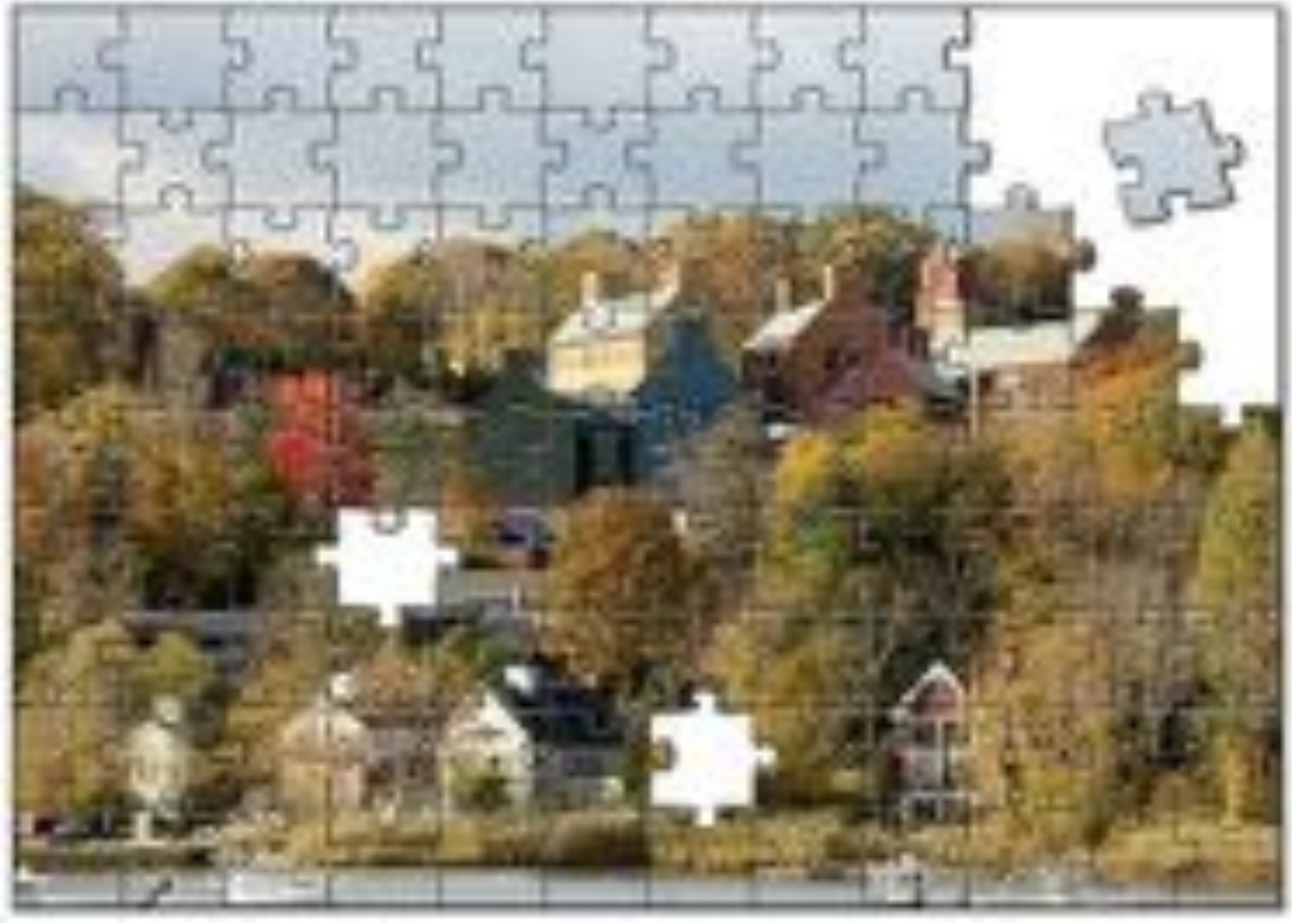


http://www.pacificbiosciences.com/assets/files/pacbio_technology_backgrounder.pdf

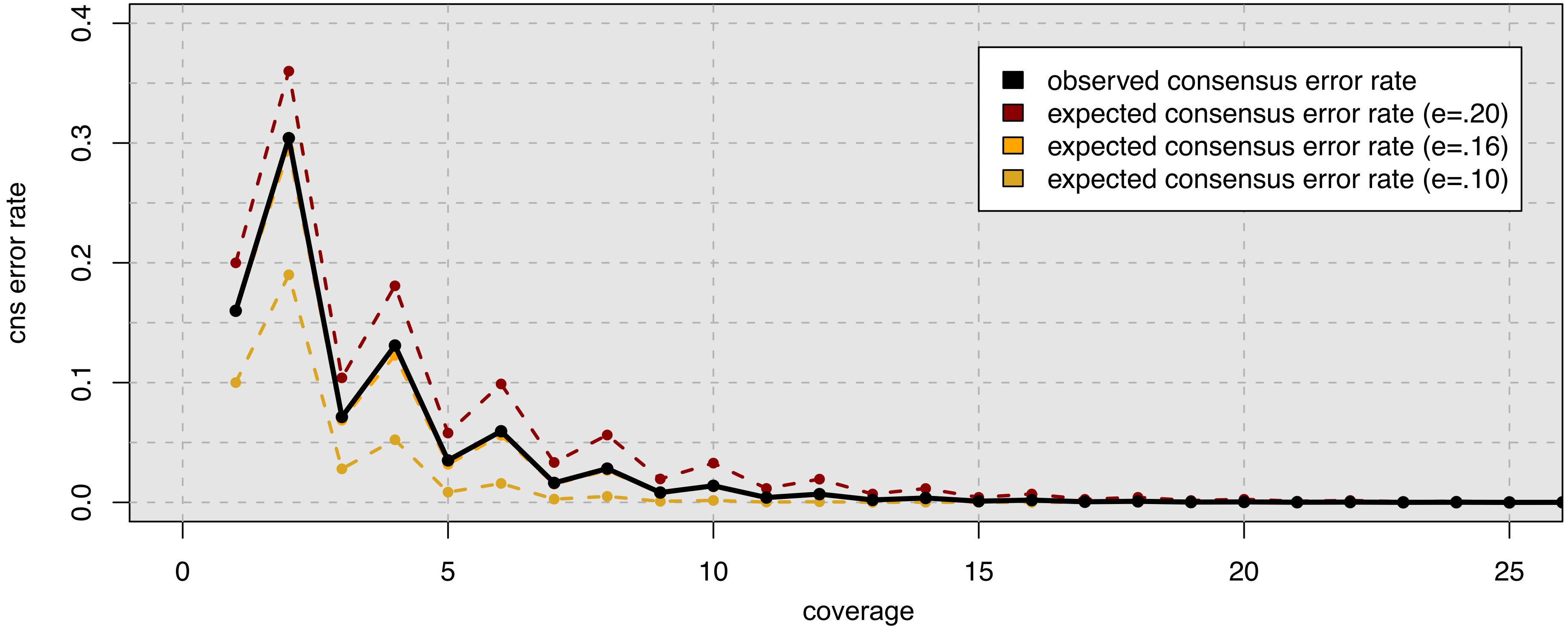
Single Molecule Sequences



“Corrective Lens” for Sequencing



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

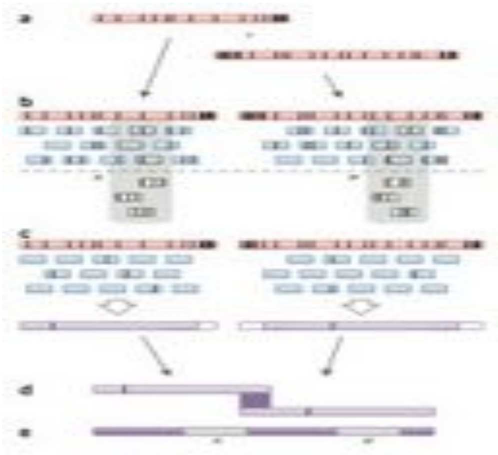
PacBio Assembly Algorithms

PBJelly

Gap Filling and Assembly Upgrade

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

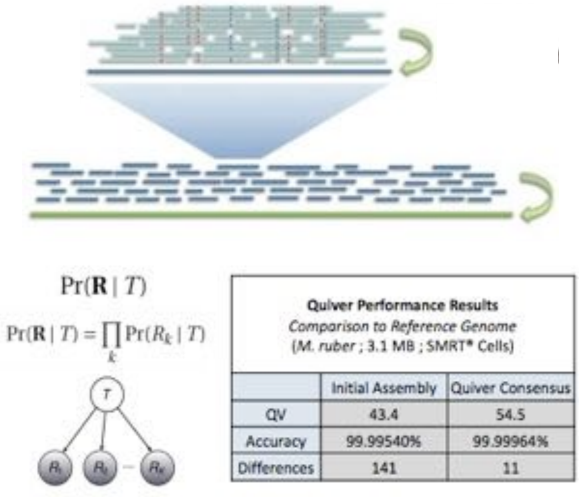
PacBioToCA & ECTools



Hybrid/PB-only Error Correction

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

HGAP & Quiver



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

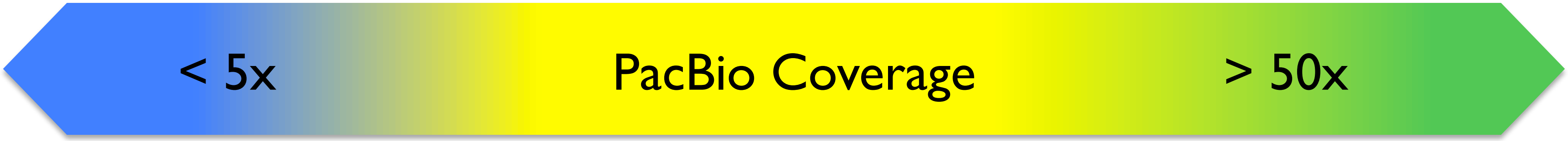
$$\Pr(\mathbf{R} \mid T) = \prod_k \Pr(R_k \mid T)$$

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

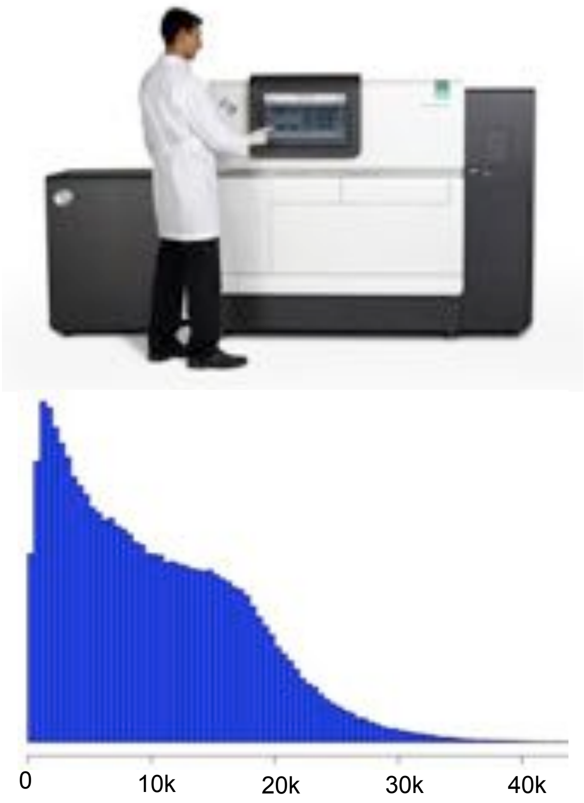
	Initial Assembly	Quiver Consensus
QV	43.4	54.5
Accuracy	99.99540%	99.99964%
Differences	141	11

PB-only Correction & Polishing

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



3rd Gen Long Read Sequencing

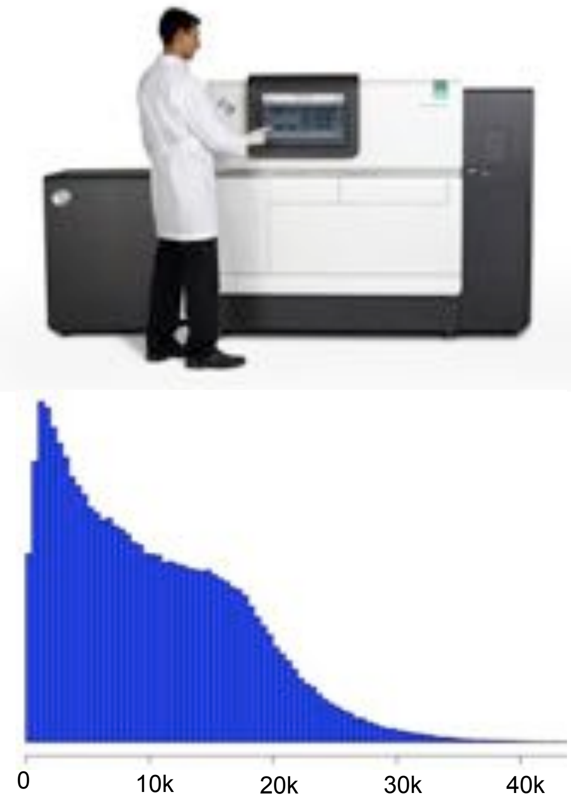


PacBio RS II

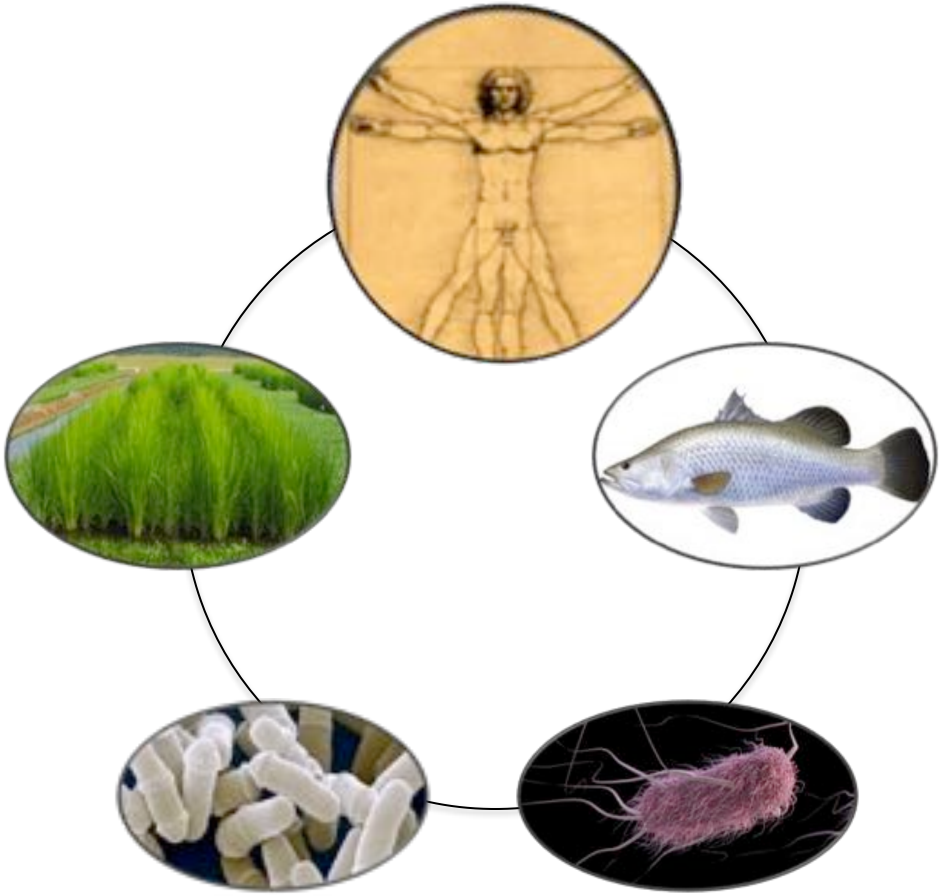
CSHL/PacBio

3rd Gen Long Read Sequencing

PacBio RS II

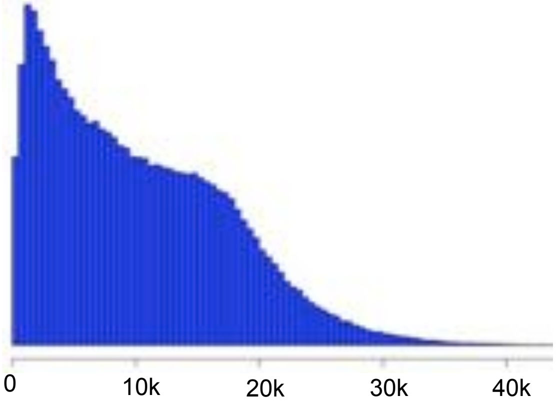


CSHL/PacBio

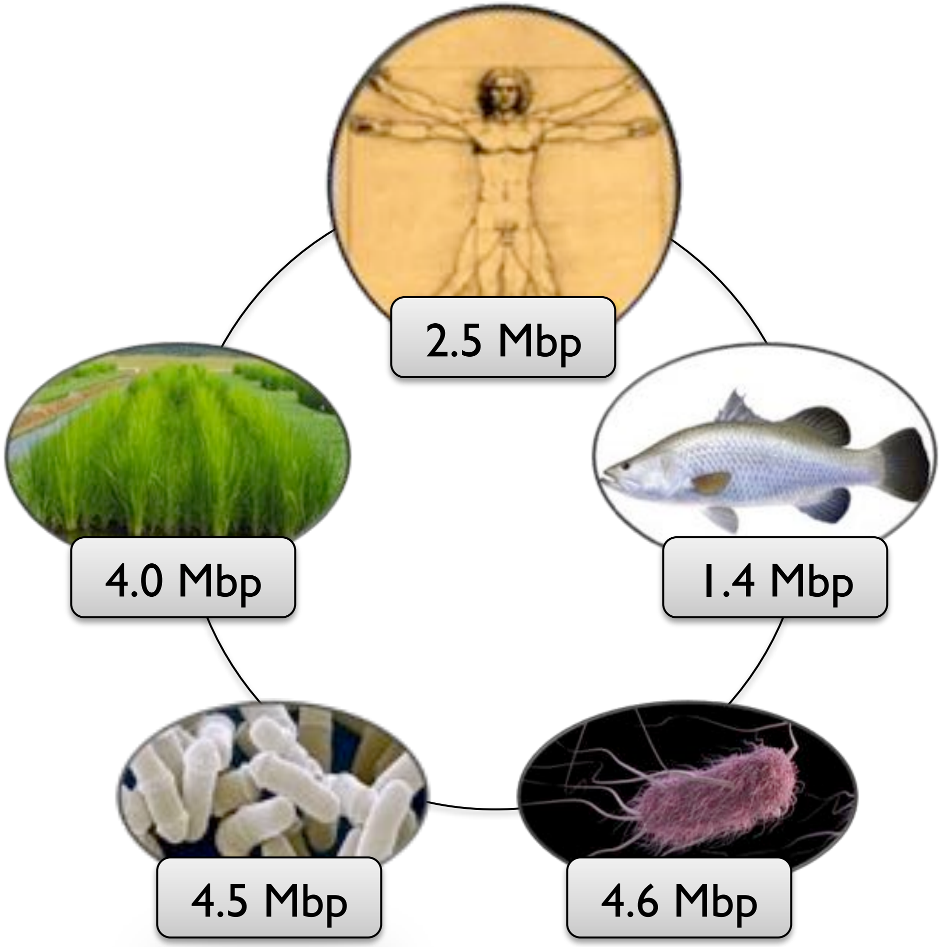


3rd Gen Long Read Sequencing

PacBio RS II

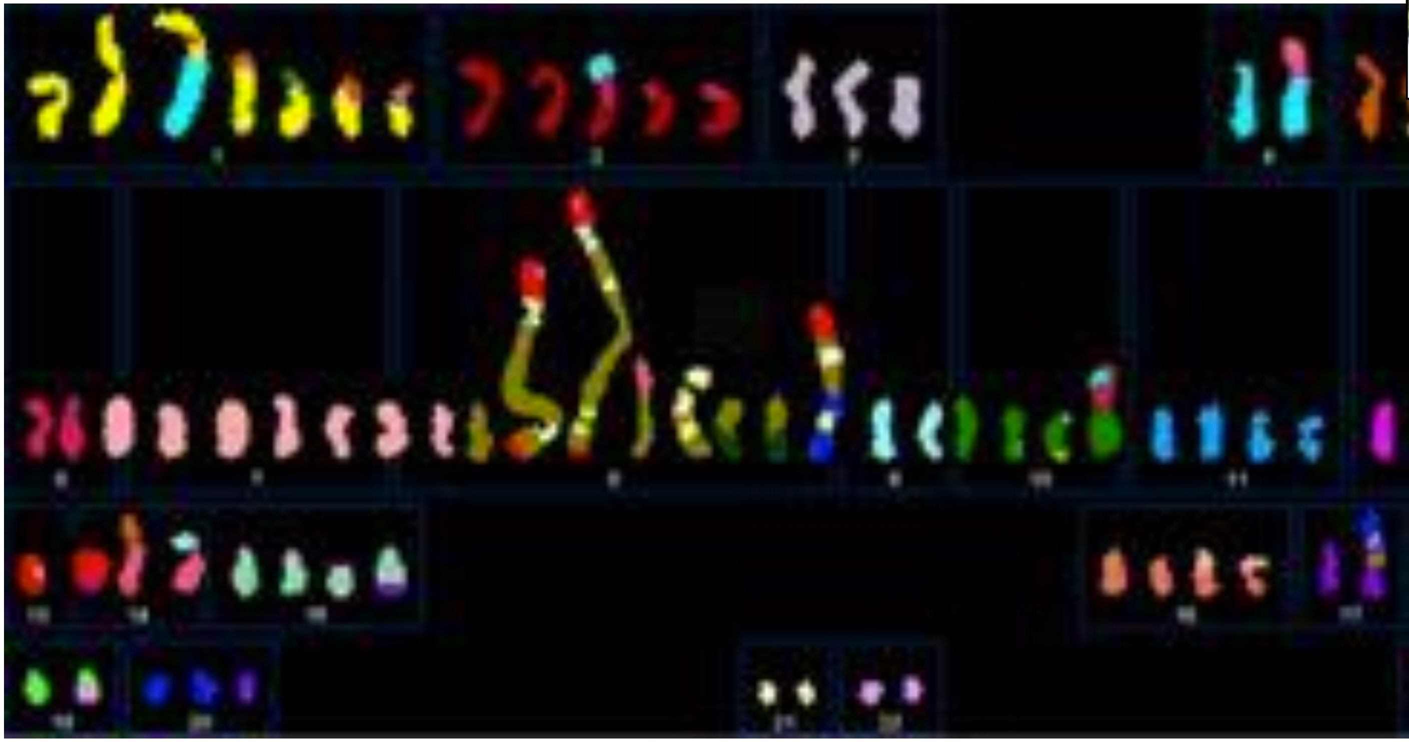


CSHL/PacBio



SK-BR-3

Most commonly used Her2-amplified breast cancer ce

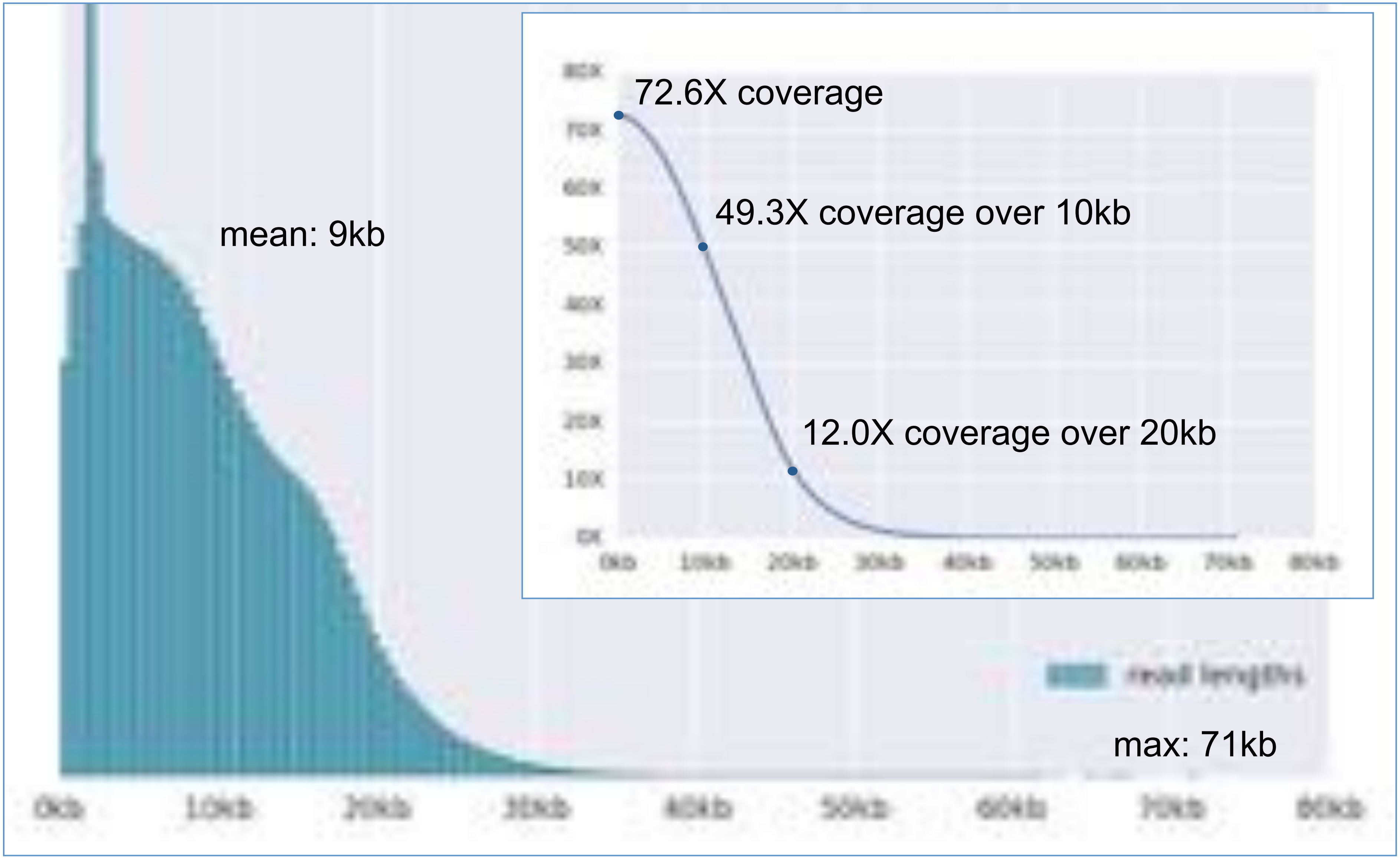


(Davidson et al, 2000)

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

PacBio read length distribution



mean: 9kb

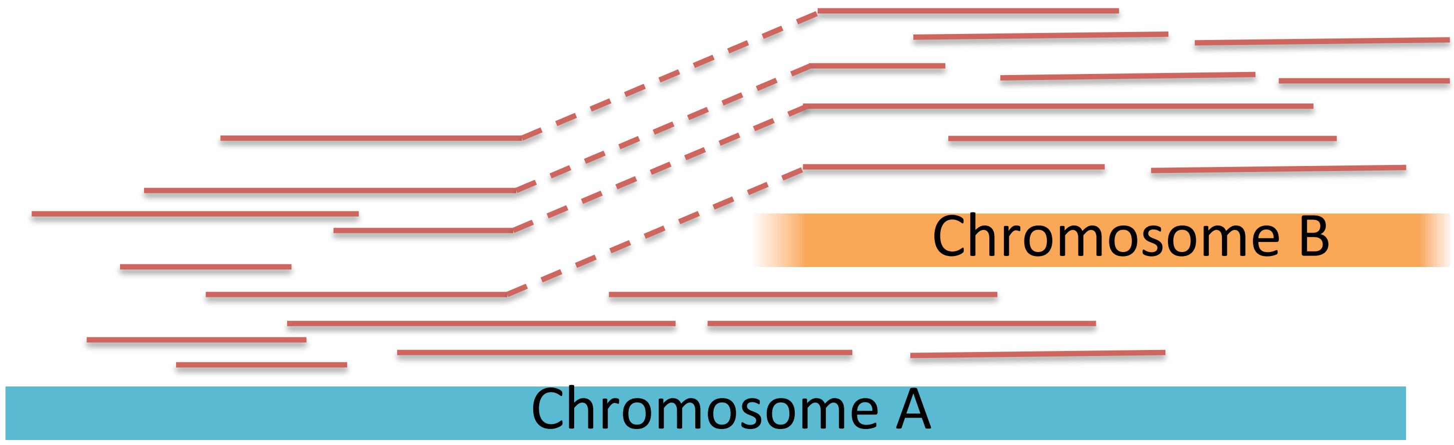
72.6X coverage

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*

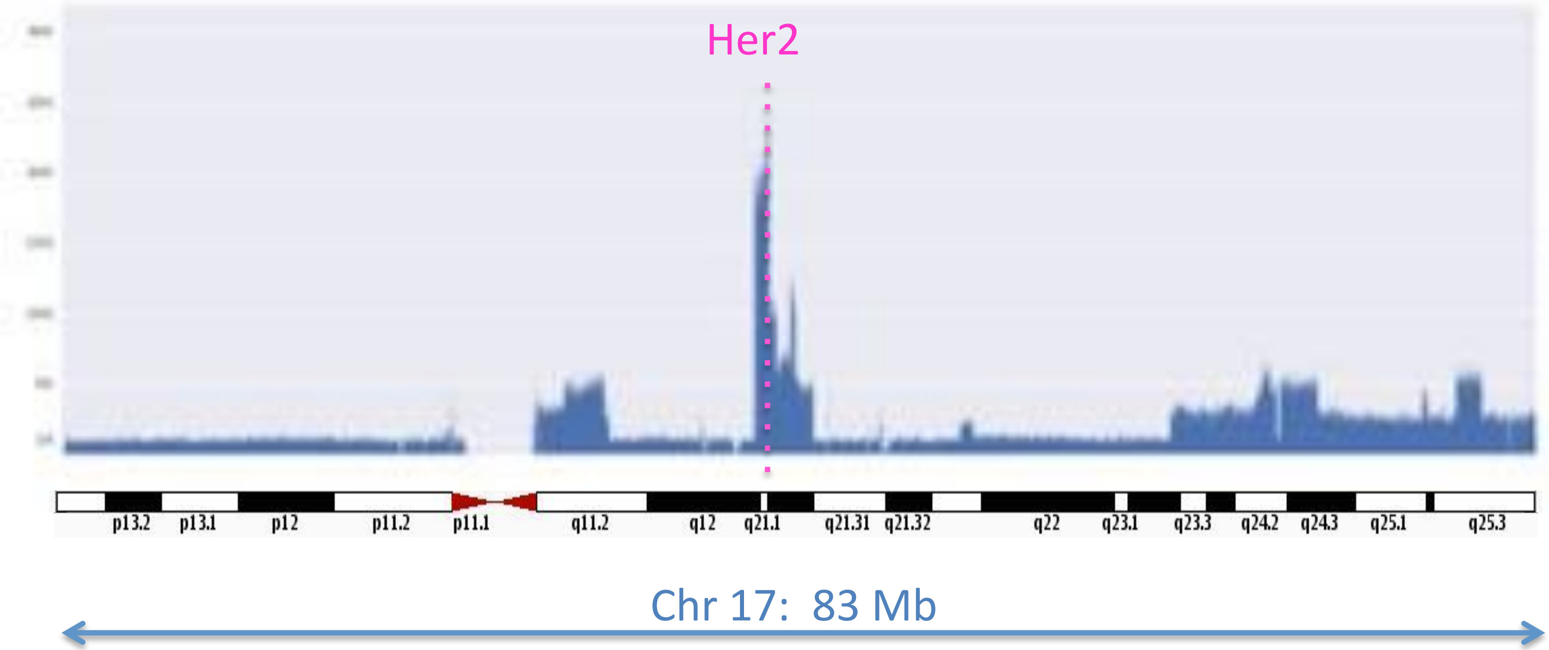
Localized HGAP + Celera Assembler + MUMmer

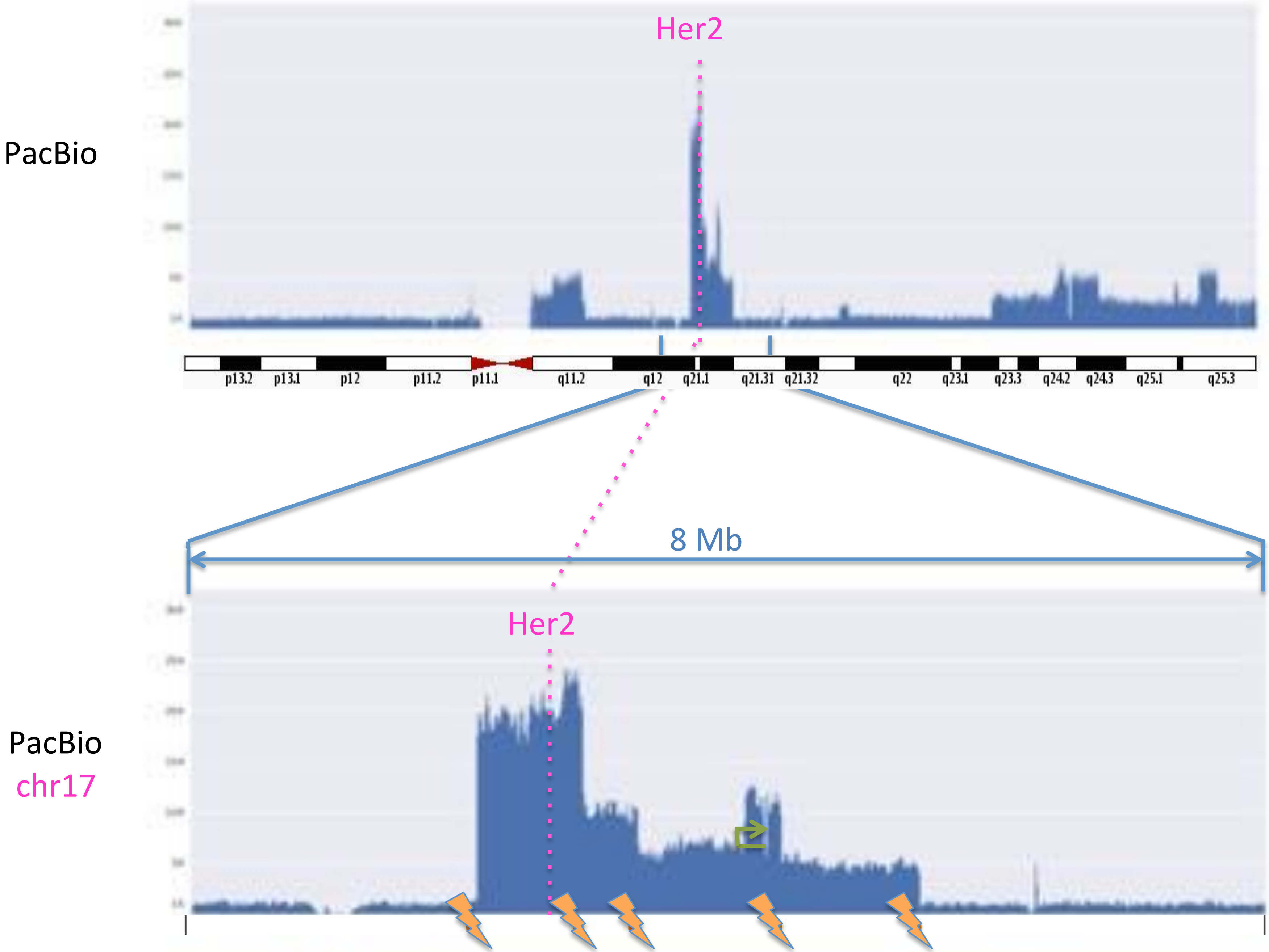
3. Whole genome assembly: In-depth analysis including *novel sequences*

DNAnexus-enabled version of Falcon

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

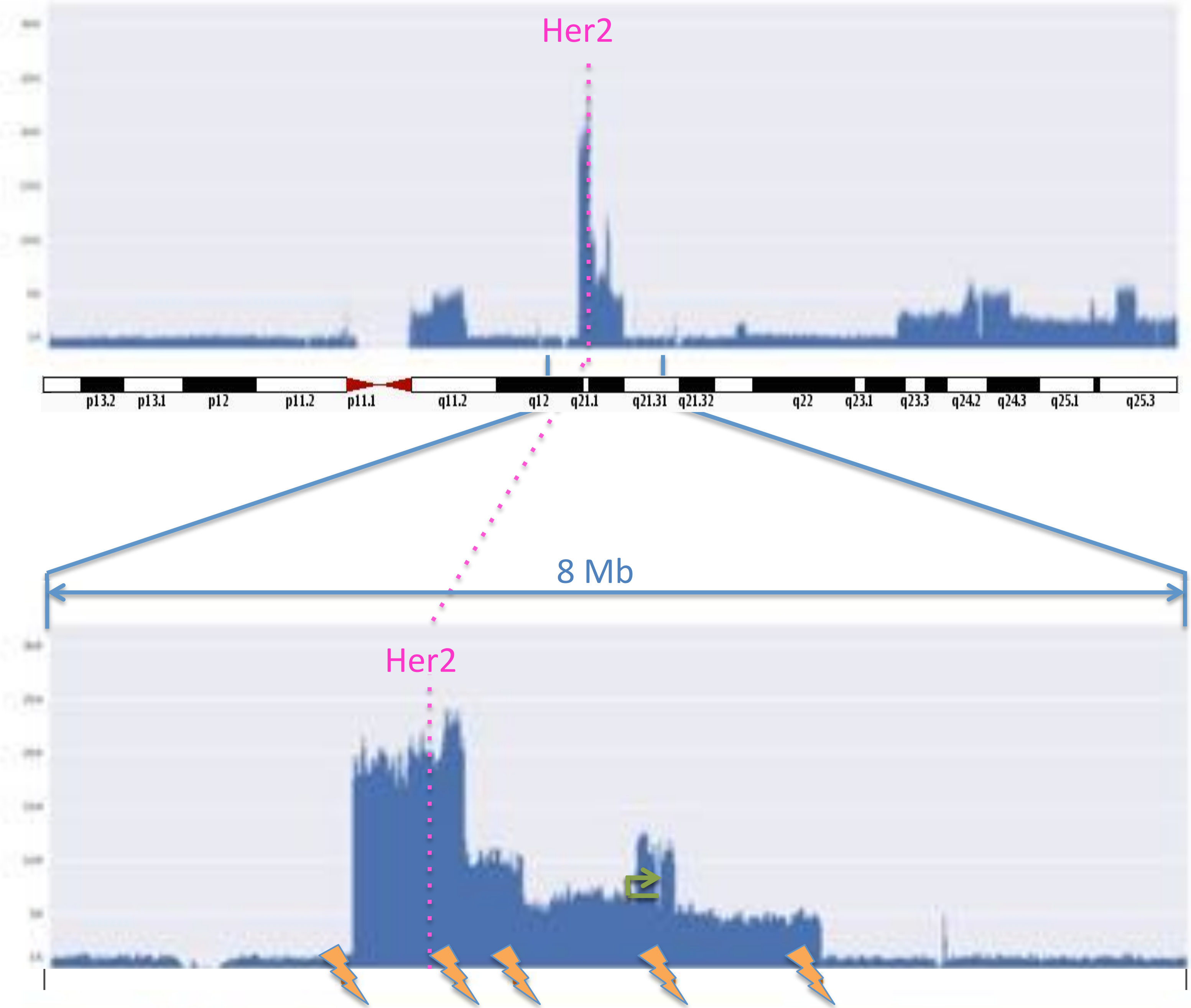
PacBio

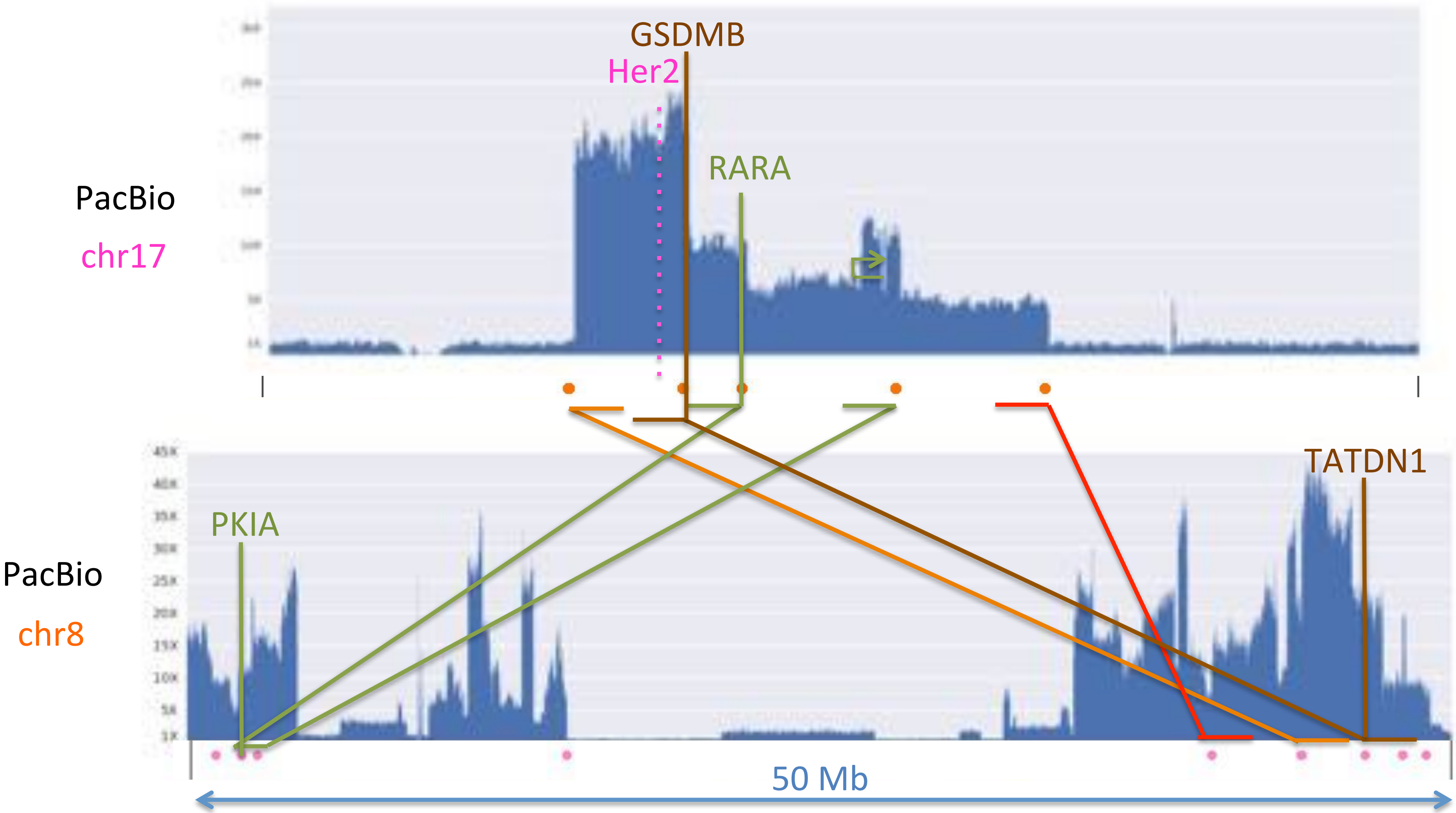

Her2
p13.2
p13.1
p12
p11.2
p11.1
q11.2
q12
q21.1
q21.31
q21.32
q22
q23.1
q23.3
q24.2
q24.3
q25.1
q25.3
Chr 17: 83 Mb


PacBio
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
8 Mb
PacBio
chr17
Her2

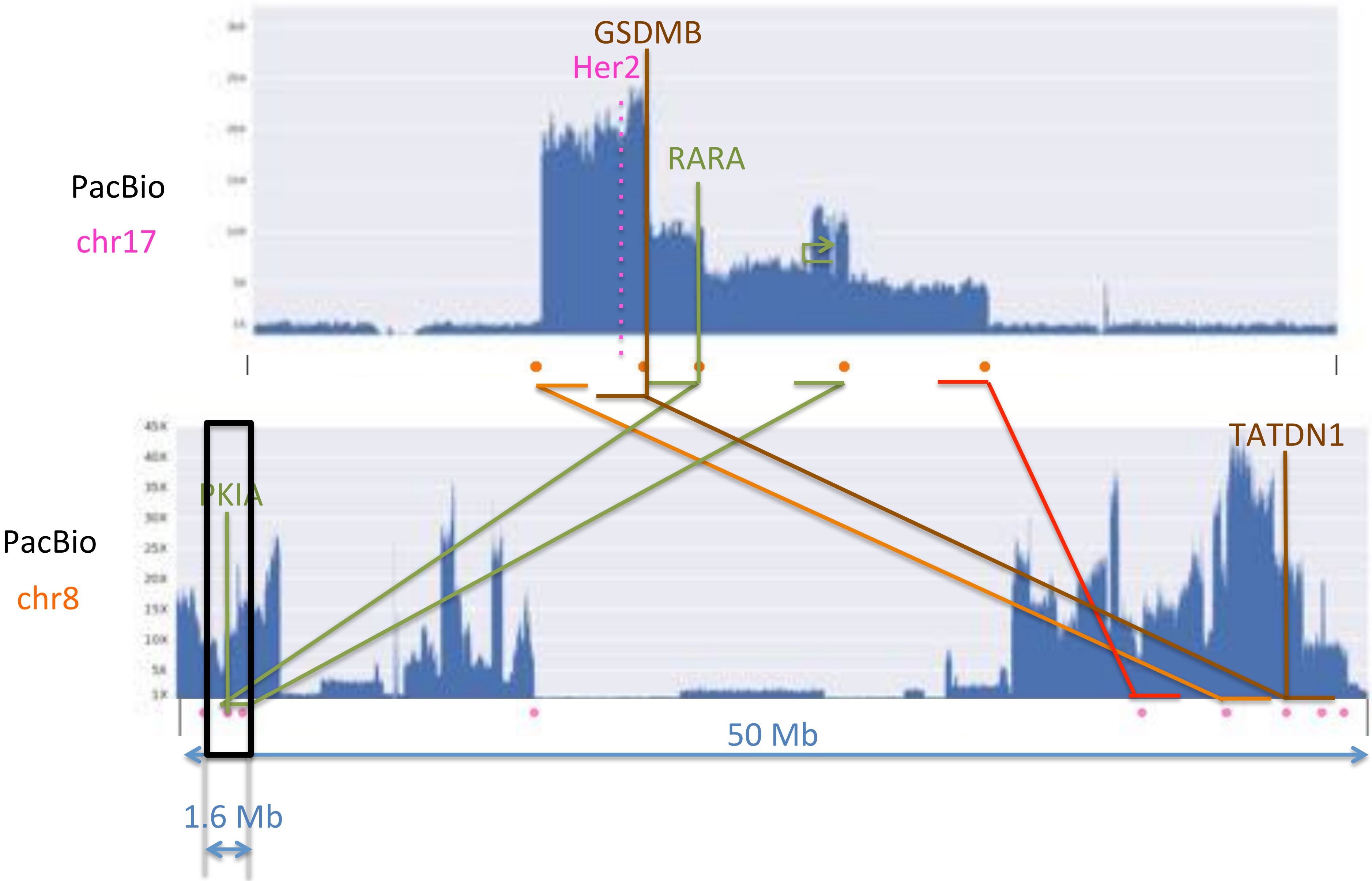
PacBio

PacBio
chr17

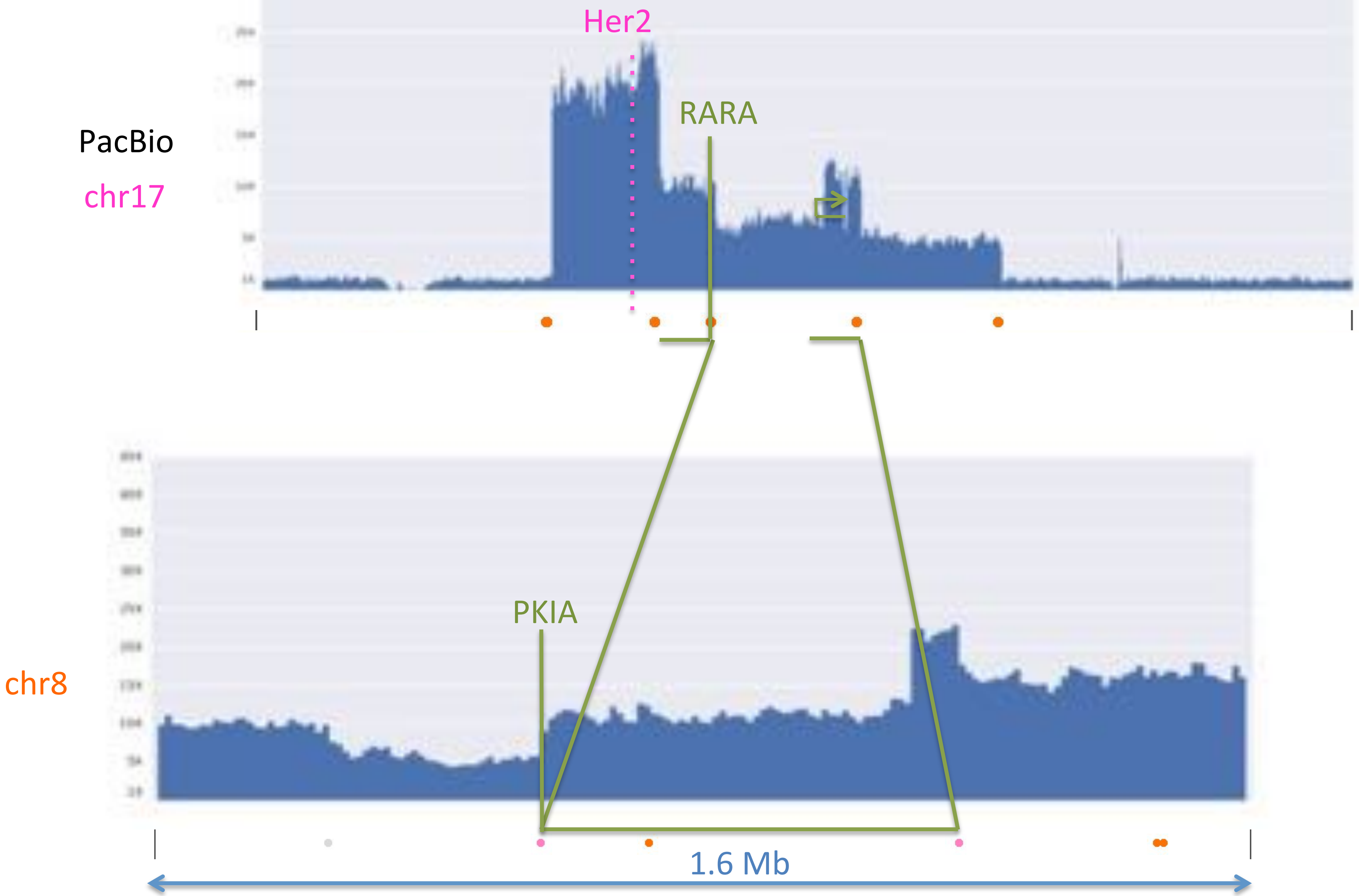




Confirmed both known gene fusions in this region

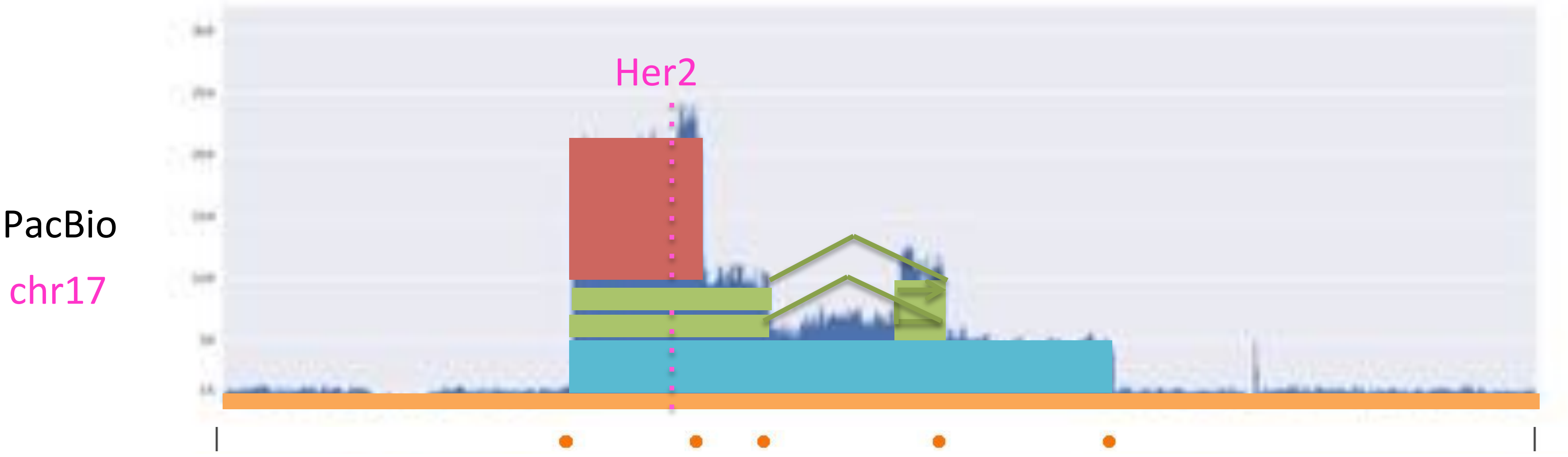


Confirmed both known gene fusions in this region



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

Go see Maria's poster!

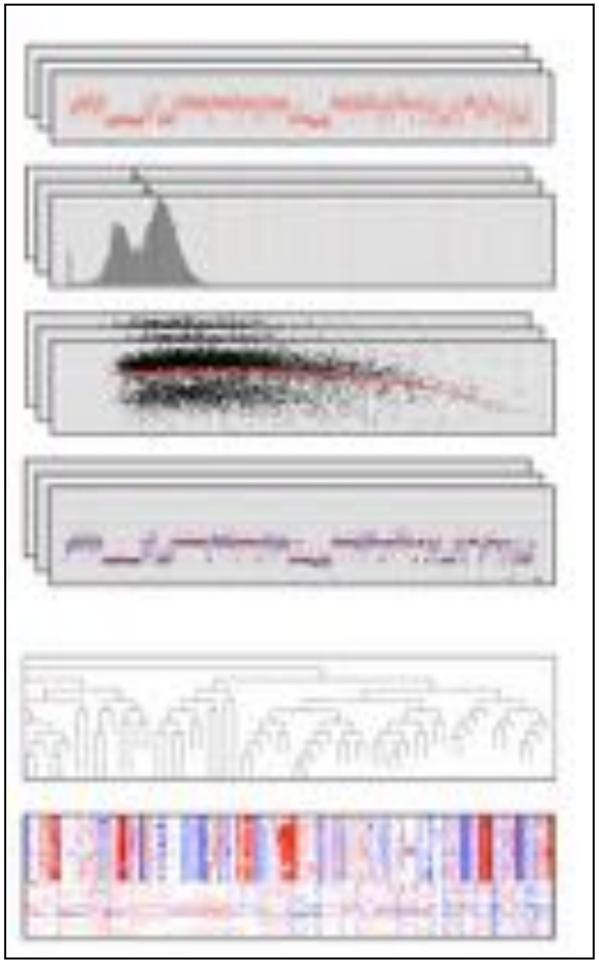
4. Final duplication from within chromosome 8

Gingko

http://qb.cshl.edu/ginkgo

Single-Cell Copy Number Analysis

- Extremely low coverage sequencing (~.1x) from amplified cells is sufficient to determine large copy number changes (>50kbp)
- Use this technique to discover CNVs in multiple cells from the same tumor to map its progress
- Implemented a new analysis suite (Ginkgo) to carry out the highly specialized processing



Interactive analysis and quality assessment of single-cell copy-number variations.
Garvin, T., Aboukhalil, *et al.* (2015) *Under review*

What should we expect from an assembly?

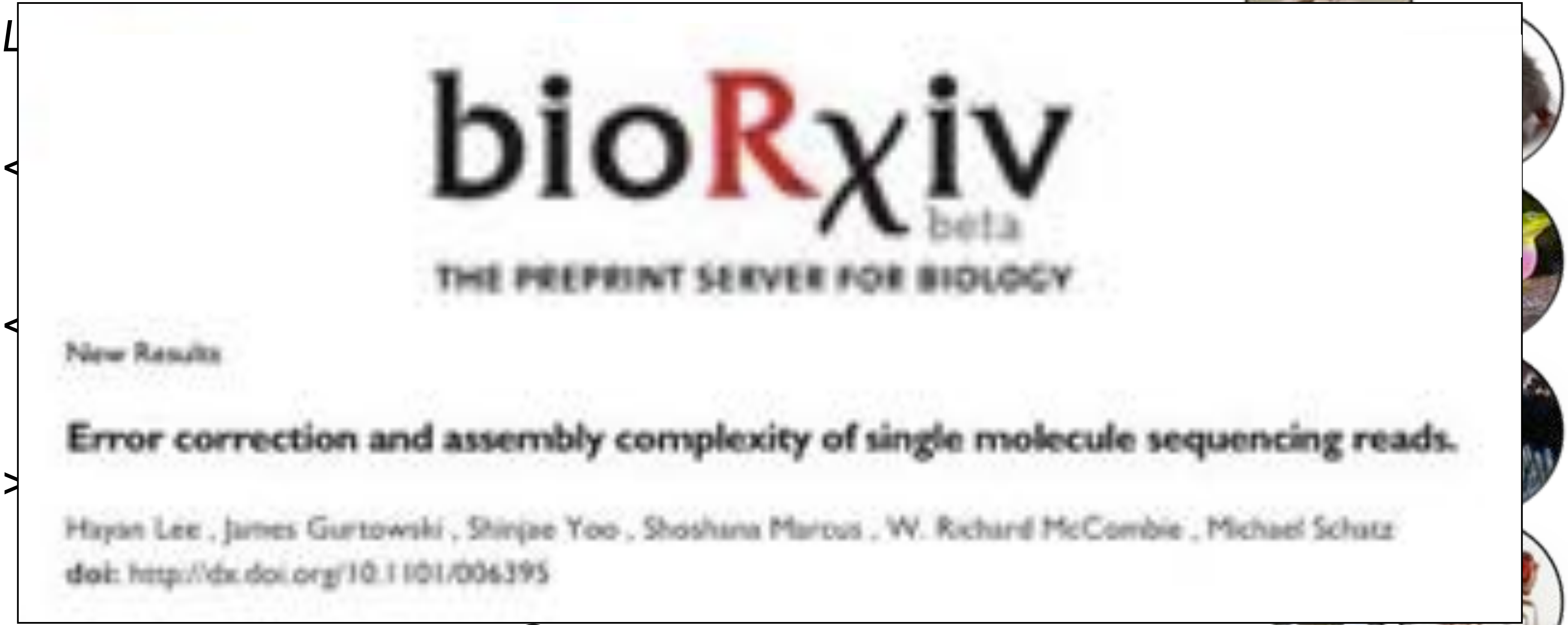
Summary & Recommendations

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

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

C

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

NBACC

Adam Phillippy
Serge Koren

Oxford NANOPORE Technologies

PACIFIC BIOSCIENCES®

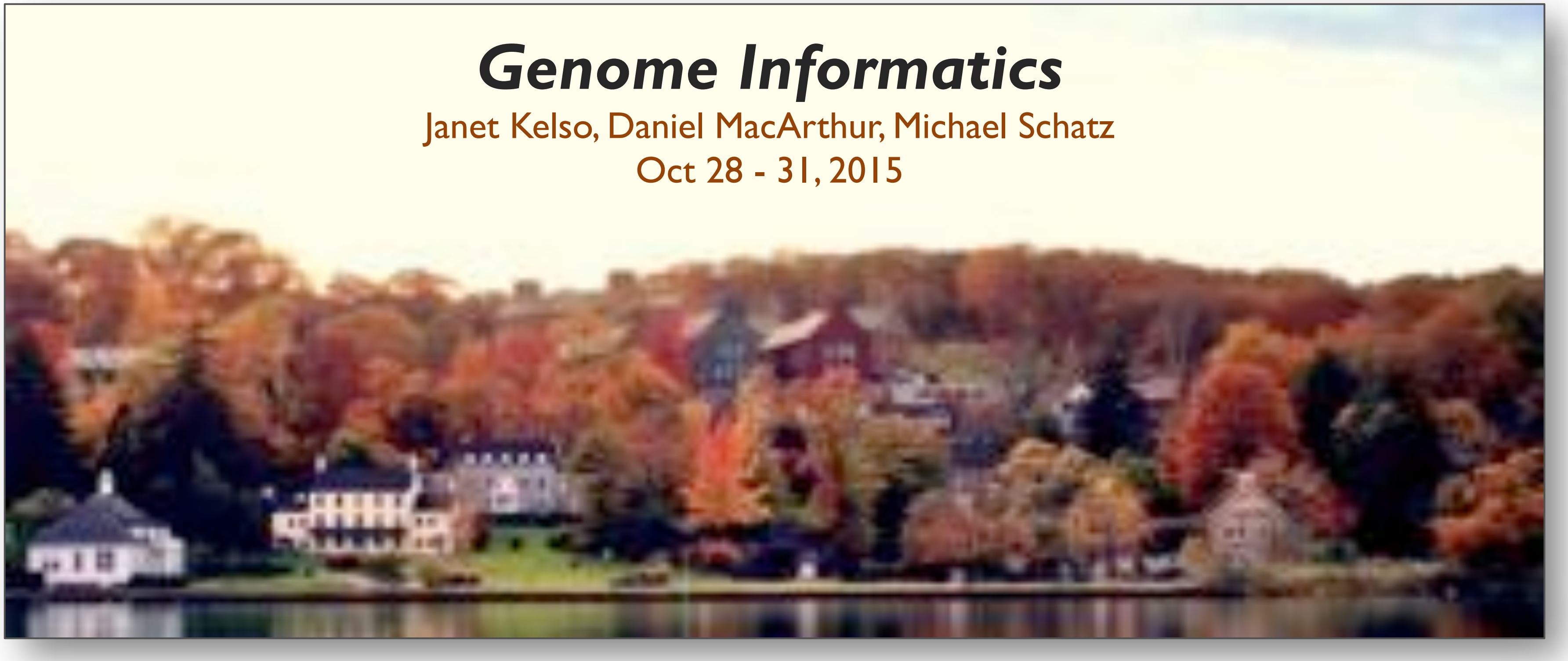
NSF

National Human Genome Research Institute

U.S. DEPARTMENT OF ENERGY

SFARI
SIMONS FOUNDATION AUTISM RESEARCH INITIATIVE

ALFRED P. SLOAN FOUNDATION



Genome Informatics

Janet Kelso, Daniel MacArthur, Michael Schatz

Oct 28 - 31, 2015

Thank you

http://schatzlab.cshl.edu

@mike_schatz