

# Single Cell and Single Molecule Analysis of Cancer

Michael Schatz

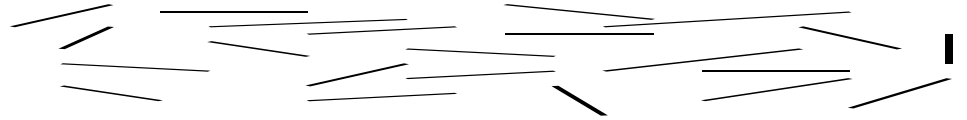
April 20, 2015

Laufer Center Retreat



# Sequence Assembly Problem

1. Shear & Sequence DNA



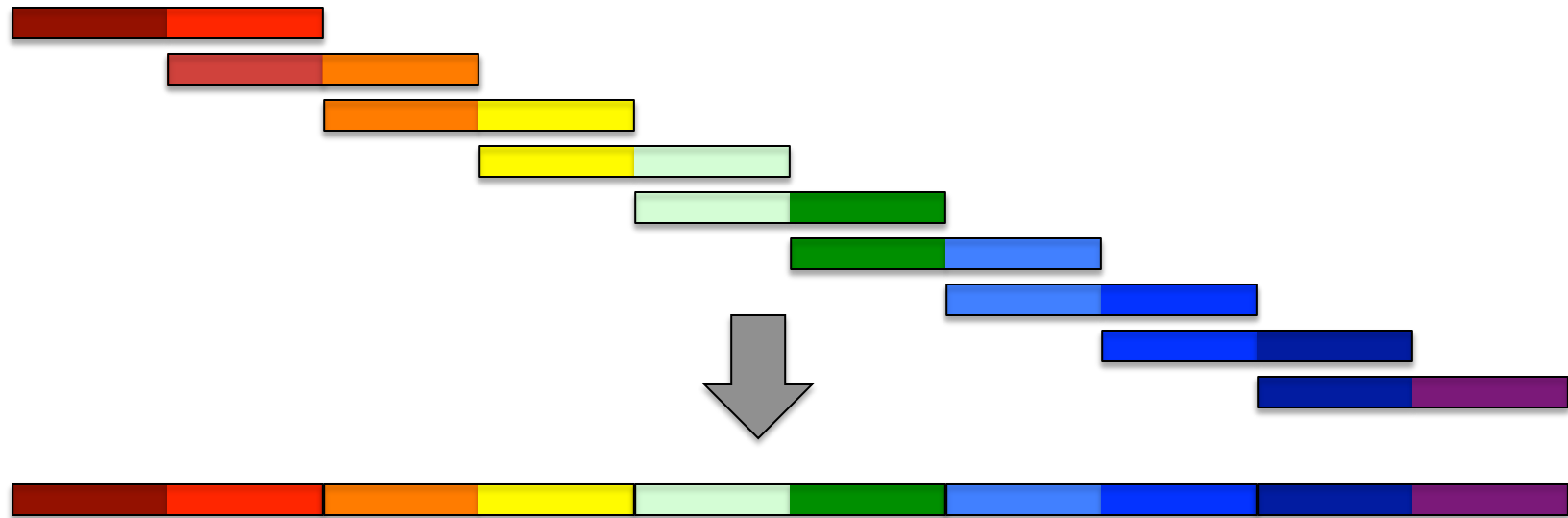
2. Construct assembly graph from overlapping reads

...AGCCTAGGGATGCGCGACACGT

GGATGCGCGACACGT CGCATATCCGGTTTGGT CAACCTCGGACGGAC

CAACCTCGGACGGACCTCAGCGAA...

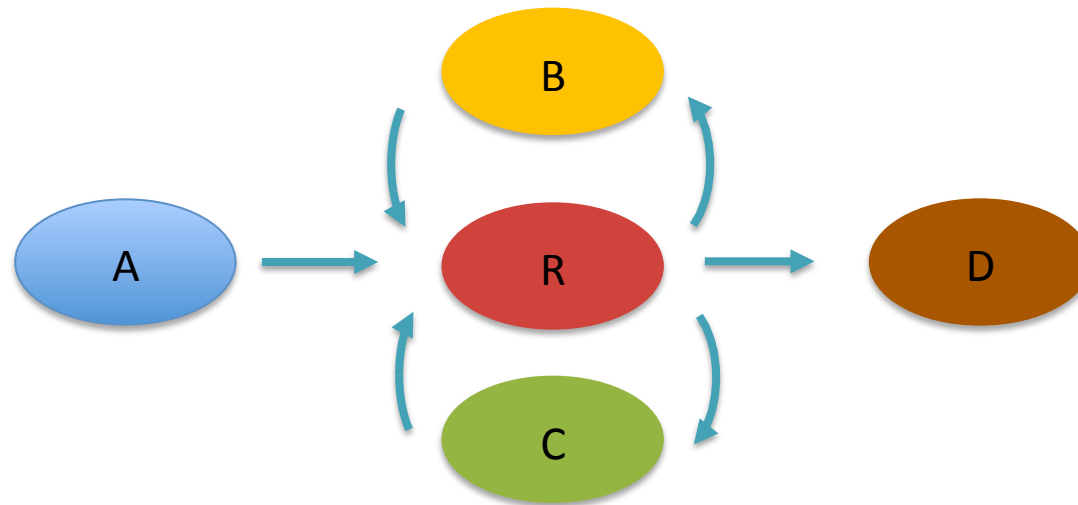
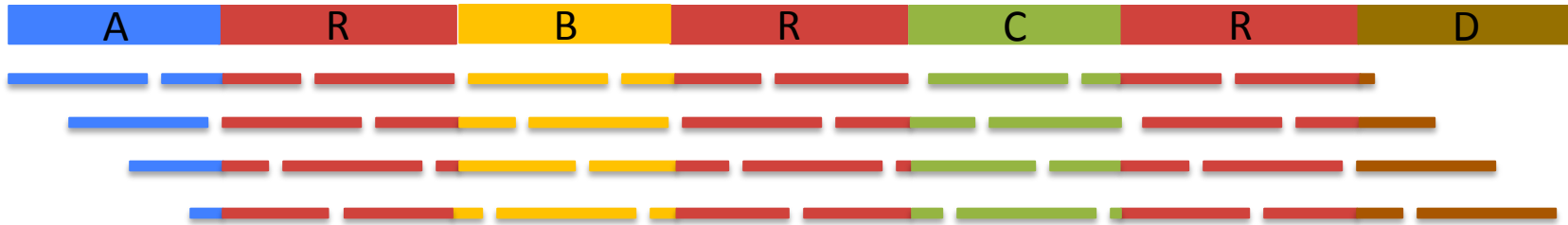
3. Simplify assembly graph



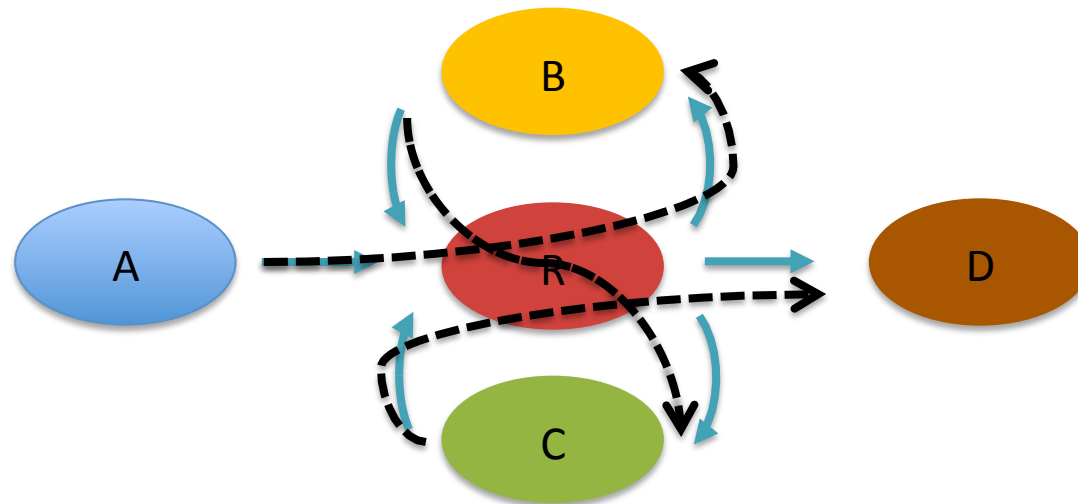
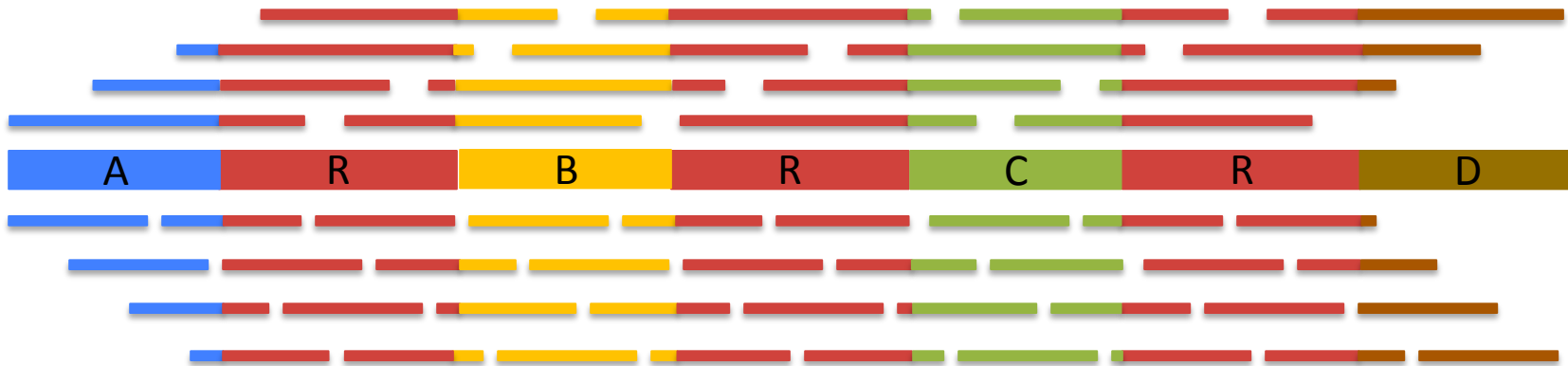
**On Algorithmic Complexity of Biomolecular Sequence Assembly Problem**

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

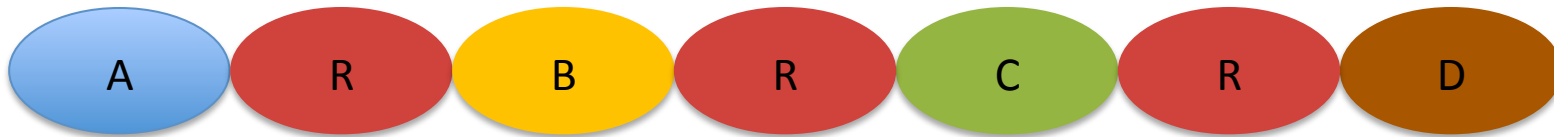
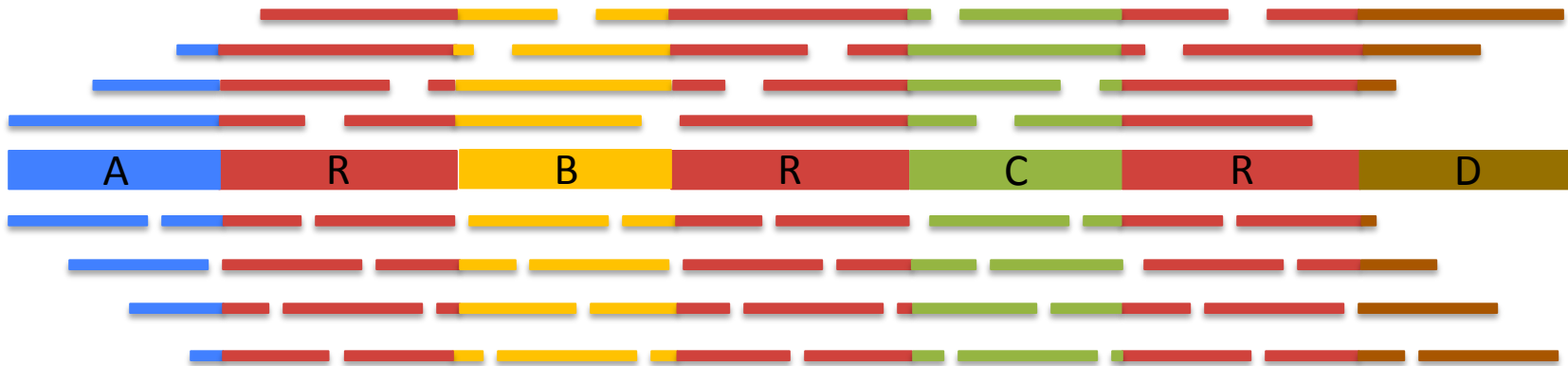
# Assembly Complexity



# Assembly Complexity



# Assembly Complexity

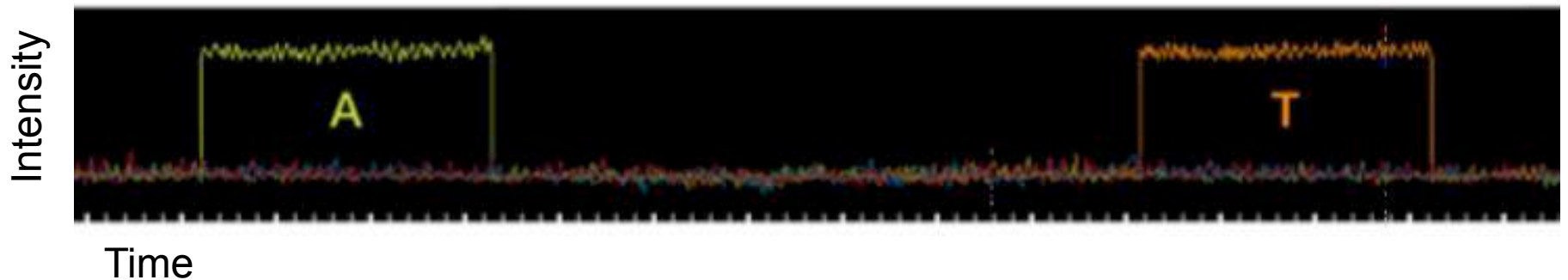
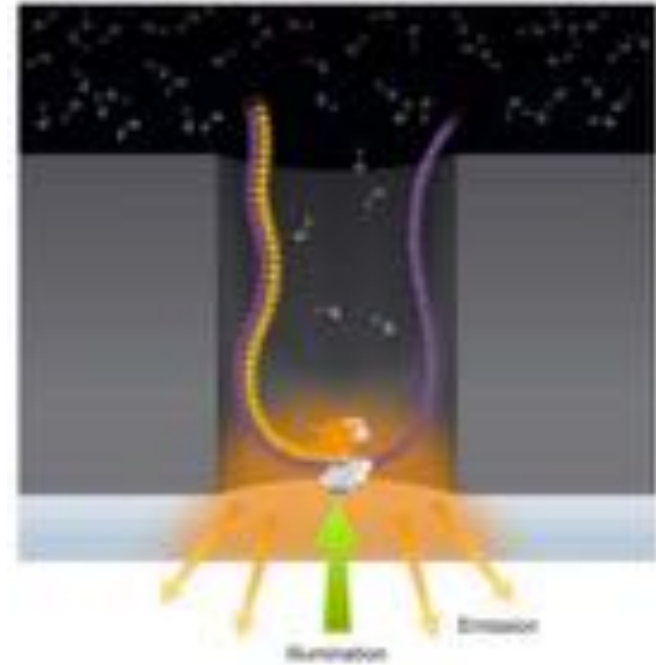
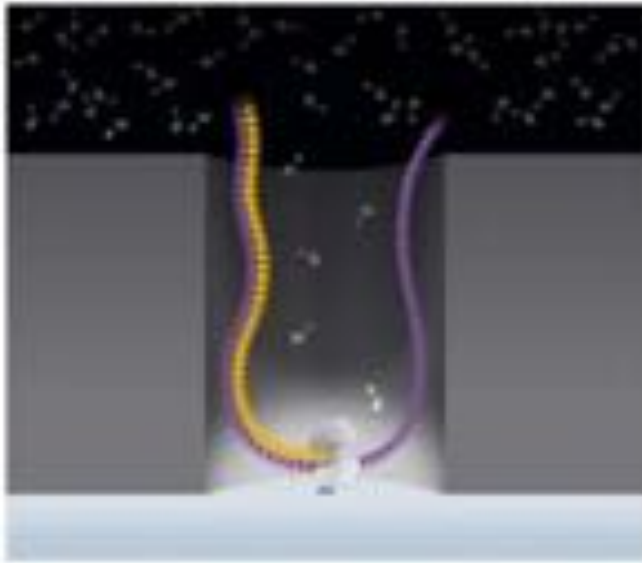


## The advantages of SMRT sequencing

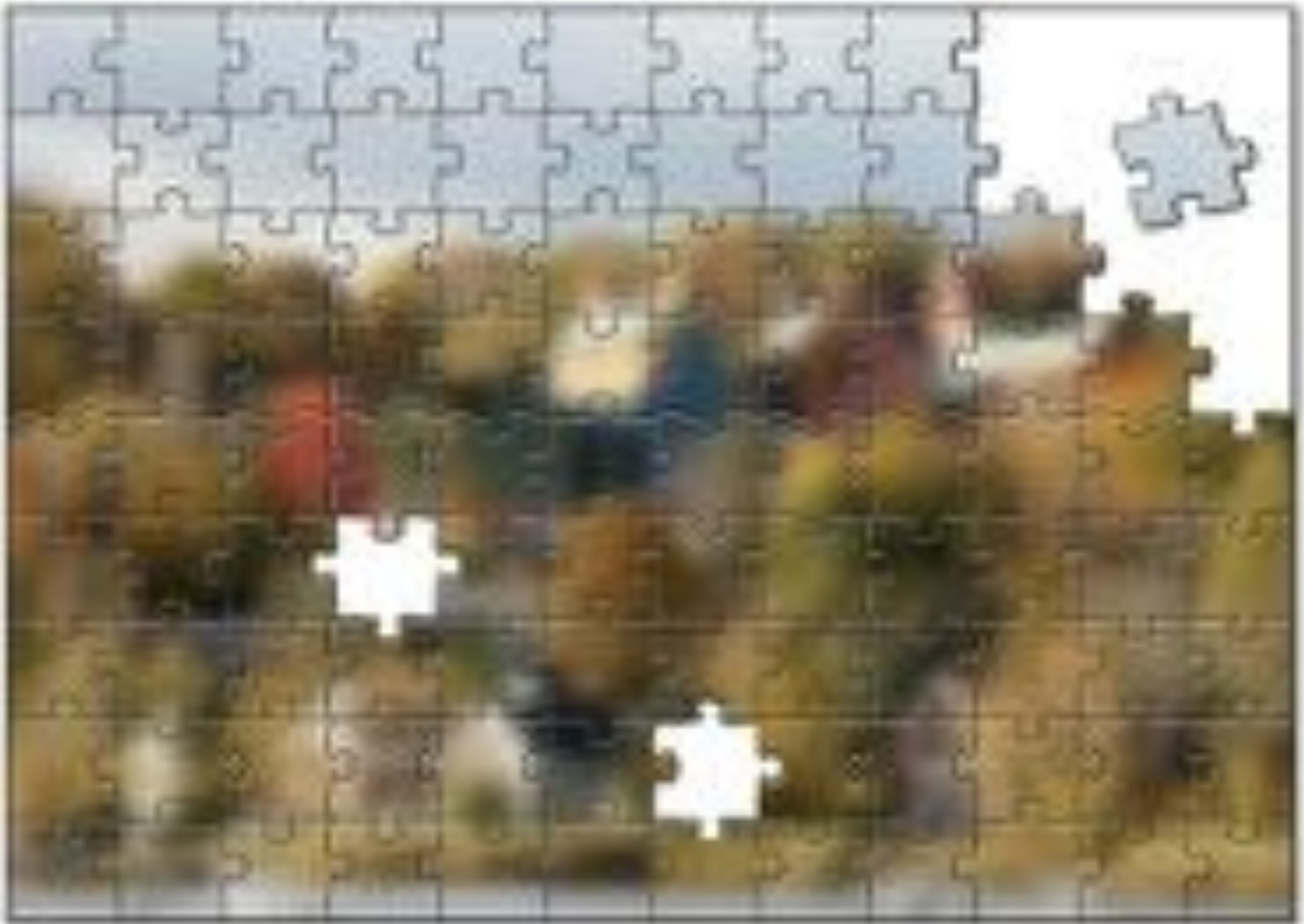
Roberts, RJ, Carneiro, MO, Schatz, MC (2013) *Genome Biology*. 14:405

# PacBio SMRT Sequencing

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



# Single Molecule Sequences



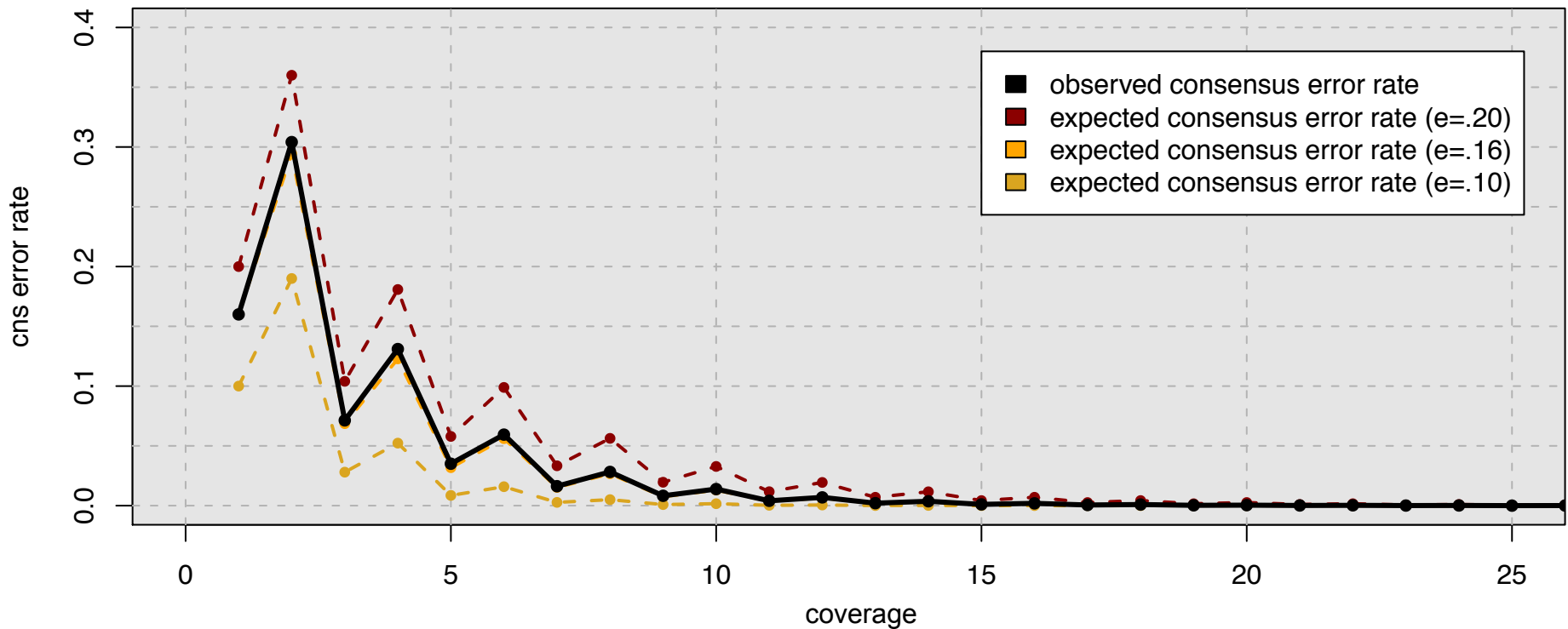


# “Corrective Lens” for Sequencing





# Consensus Accuracy and Coverage



## Coverage can overcome random errors

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

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

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

# PacBio Assembly Algorithms

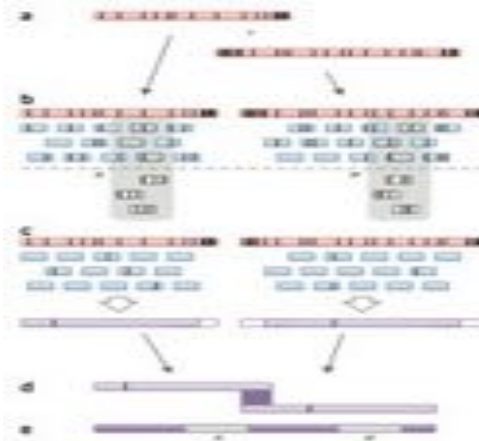
## PBJelly



**Gap Filling  
and Assembly Upgrade**

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

## PacBioToCA & ECTools



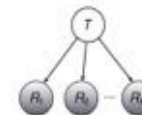
**Hybrid/PB-only Error  
Correction**

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

## HGAP & Quiver



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



Quiver Performance Results Comparison to Reference Genome ( <i>M. ruber</i> ; 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

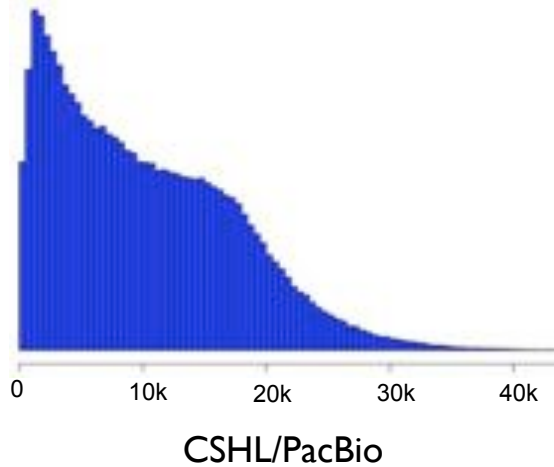
< 5x

PacBio Coverage

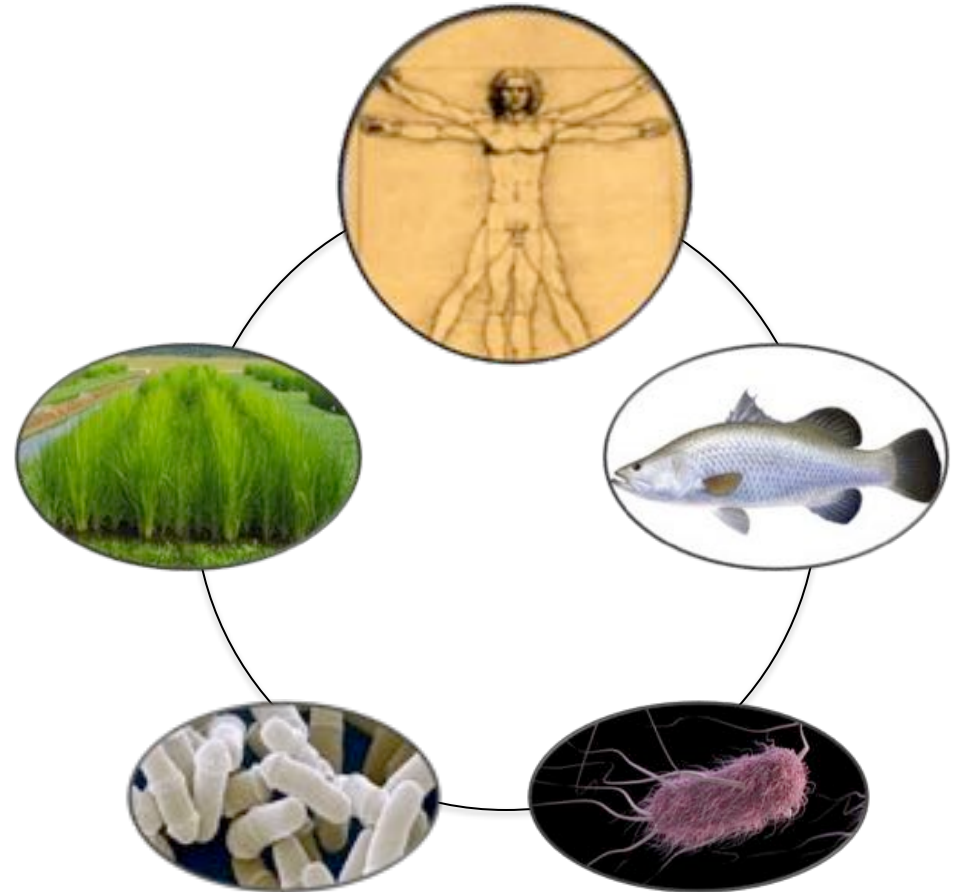
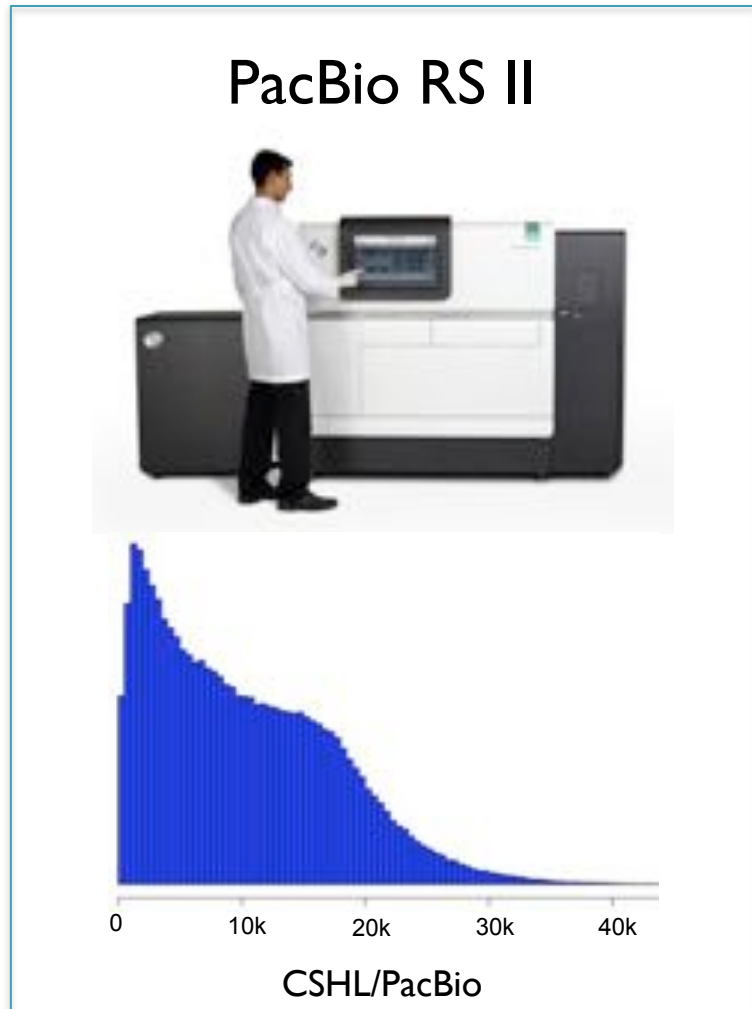
> 50x

# 3<sup>rd</sup> Gen Long Read Sequencing

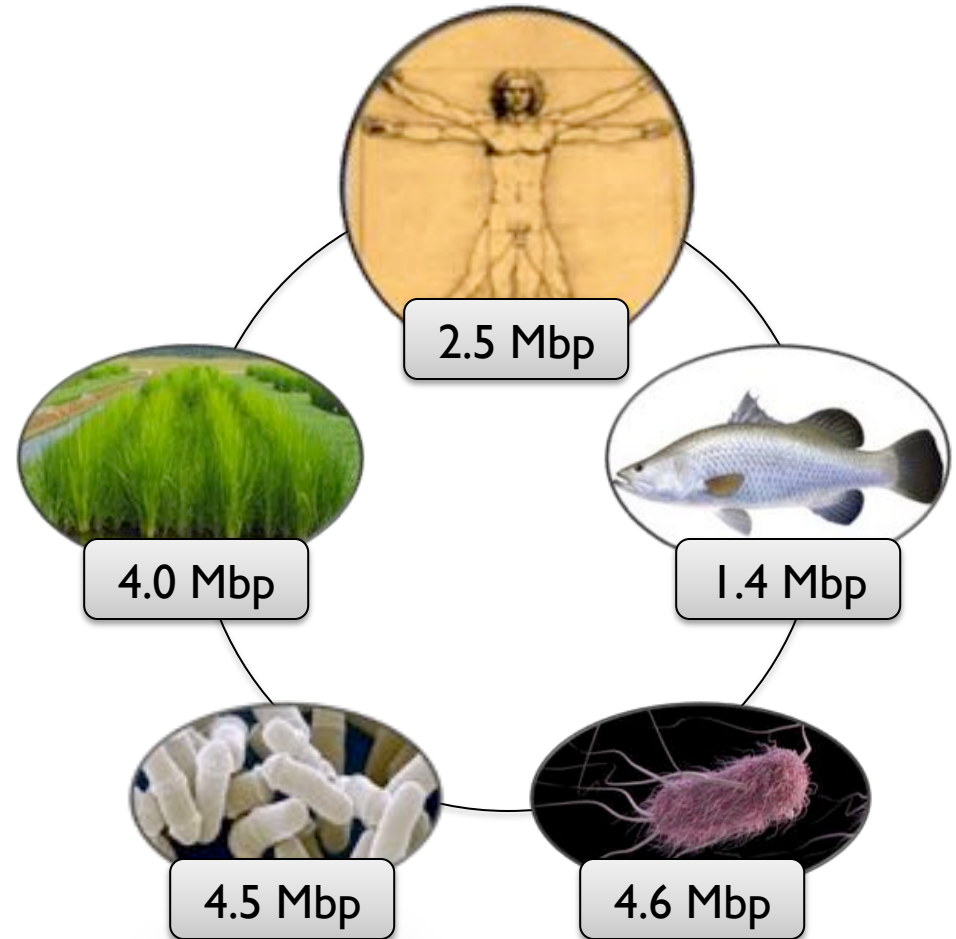
PacBio RS II



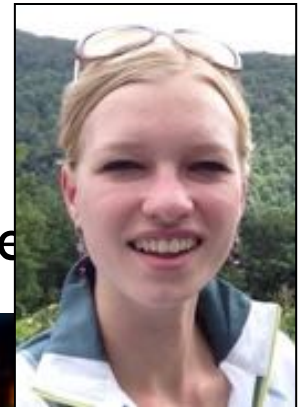
# 3<sup>rd</sup> Gen Long Read Sequencing



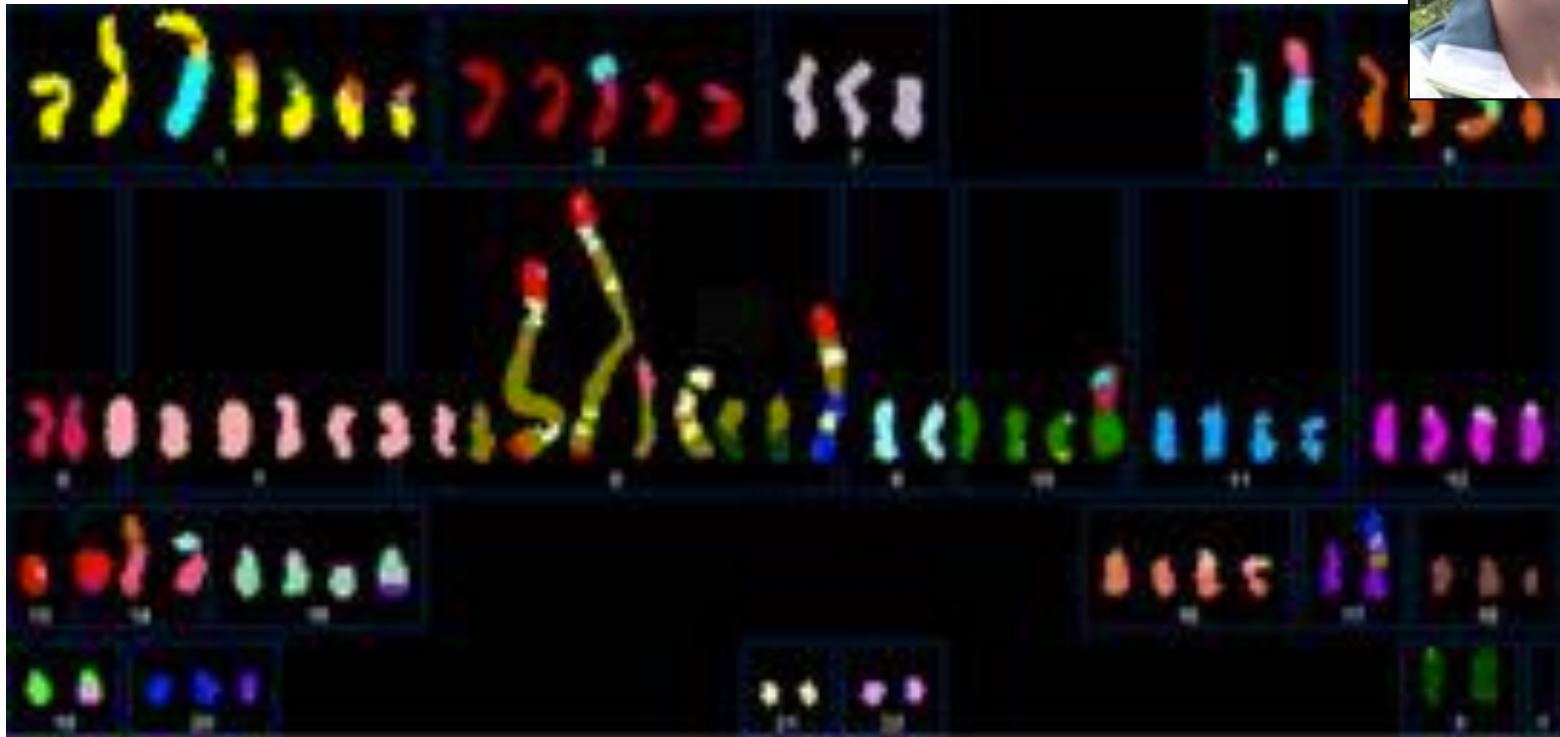
# 3<sup>rd</sup> Gen Long Read Sequencing



# SK-BR-3



Most commonly used Her2-amplified breast cancer cell line

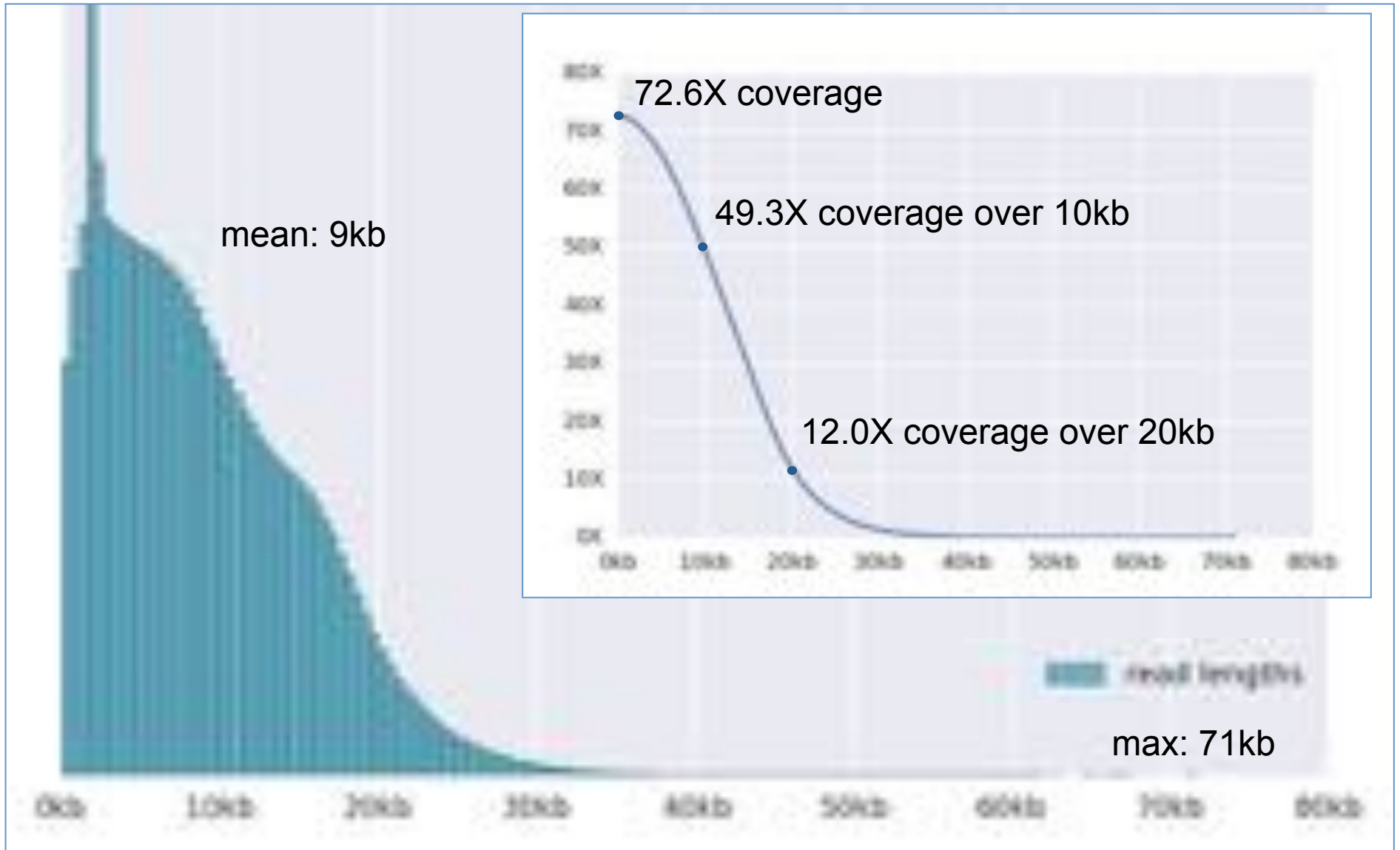


(Davidson et al, 2000)

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

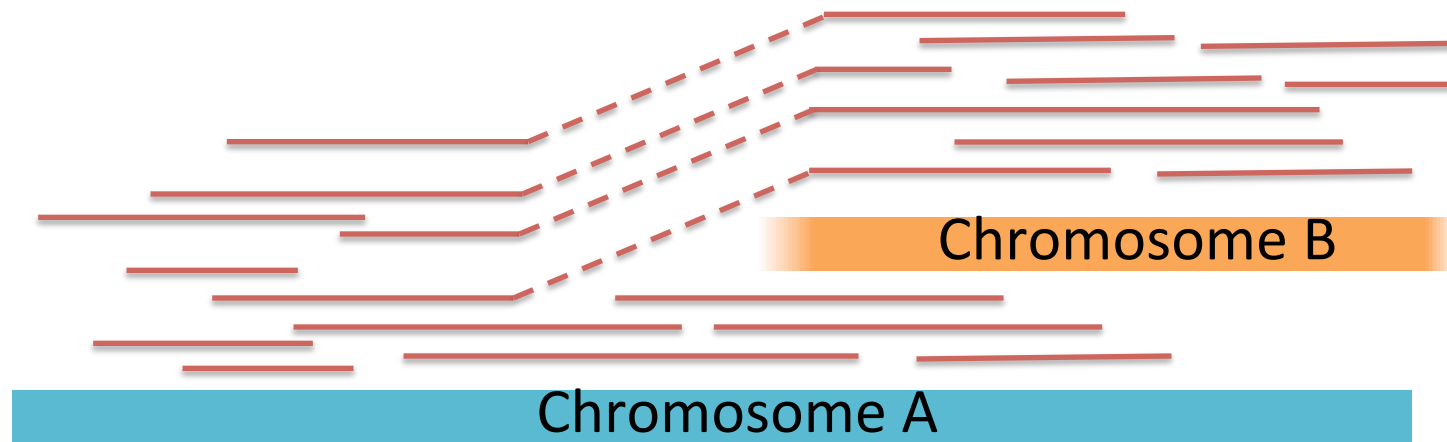
Ongoing collaboration between CSHL and OICR to *de novo* assemble the complete cell line genome with PacBio long reads

# PacBio read length distribution





# 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*

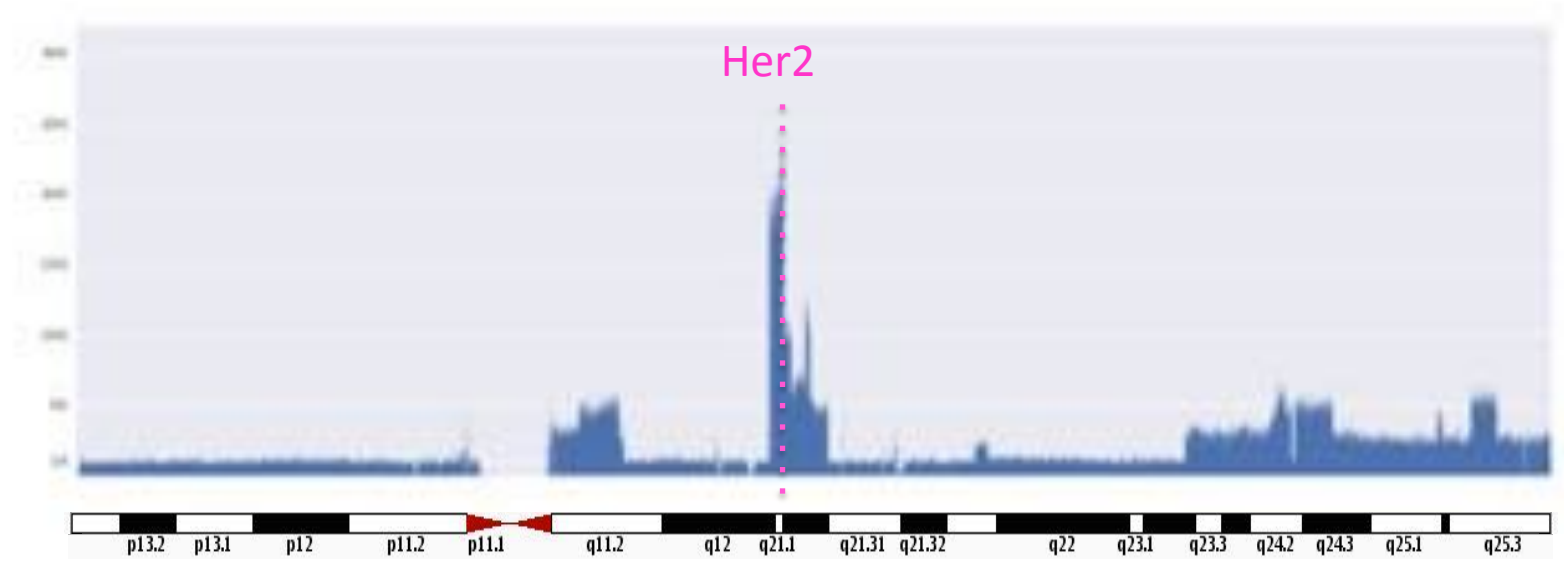
DNAexus-enabled version of Falcon

**Total Assembly: 2.64Gbp**

**Contig N50: 2.56 Mbp**

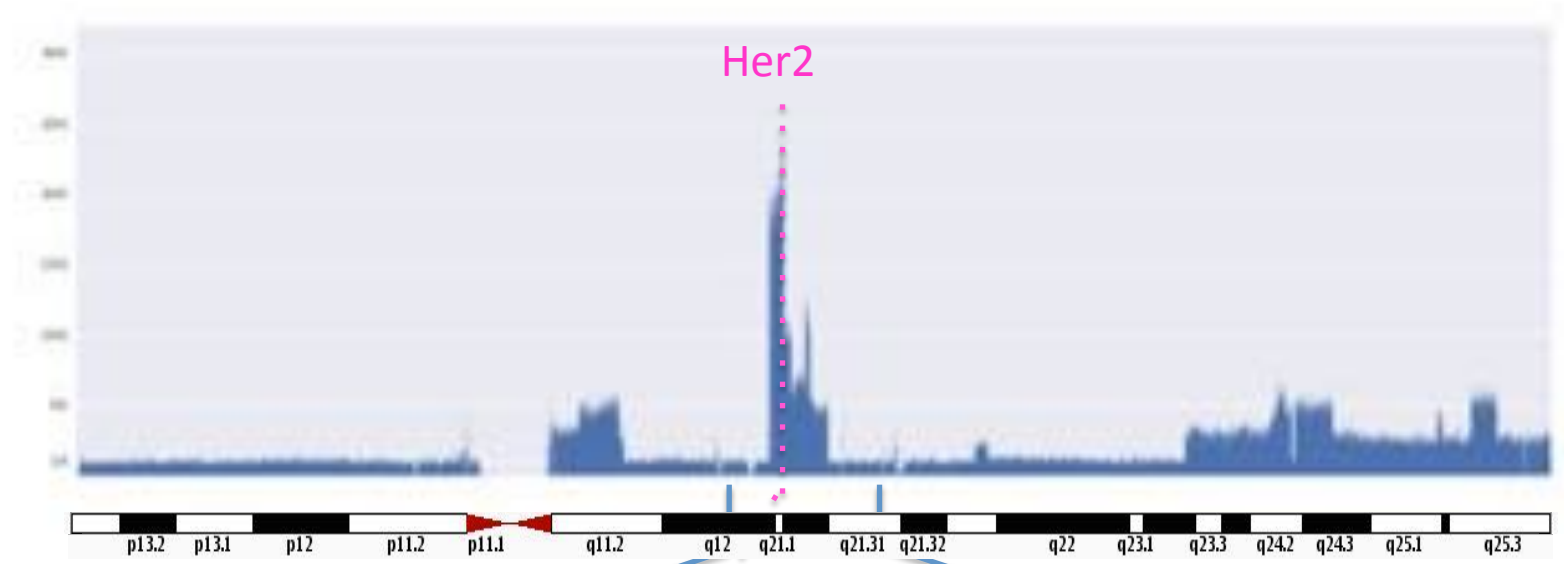
**Max Contig: 23.5Mbp**

PacBio

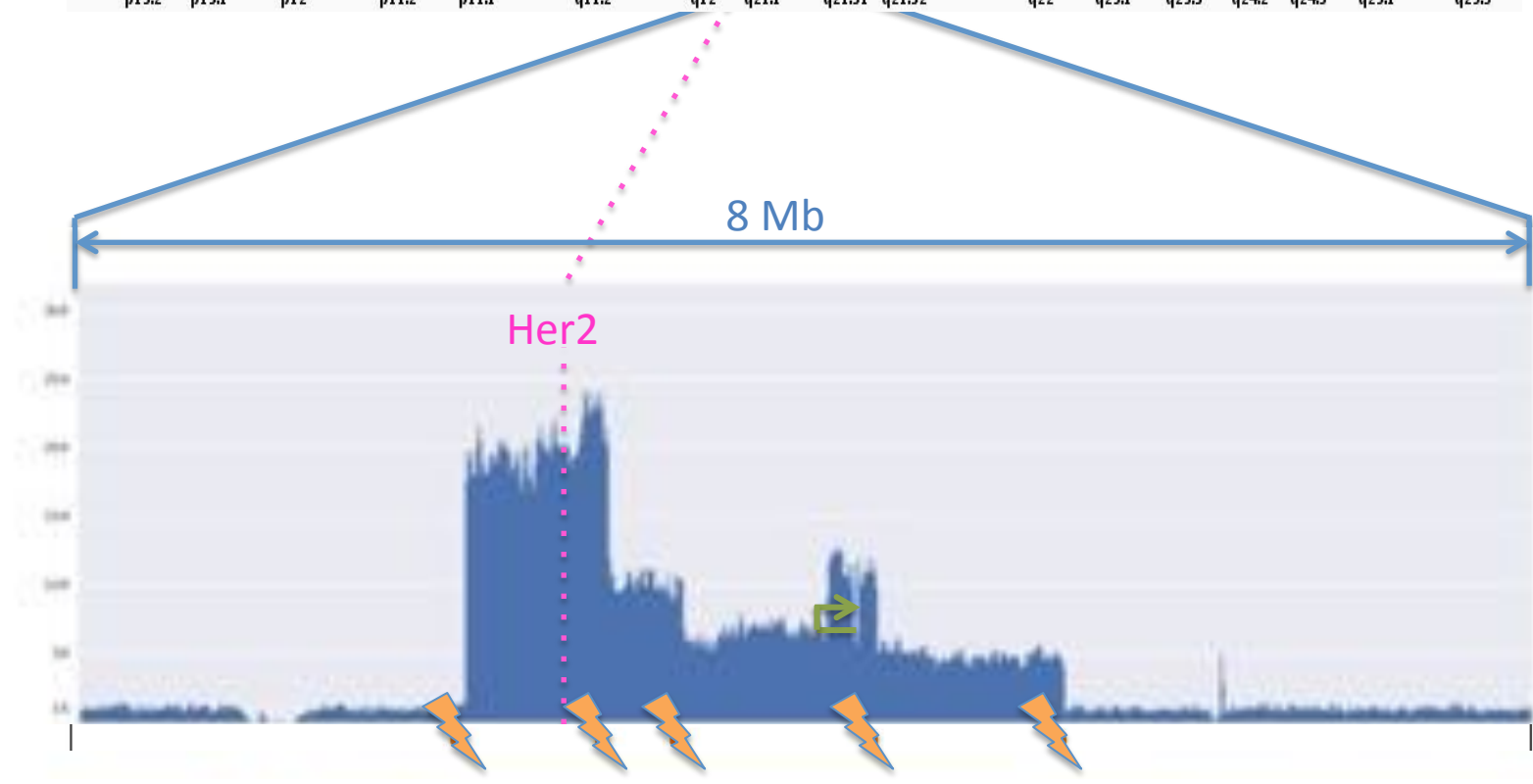


Chr 17: 83 Mb

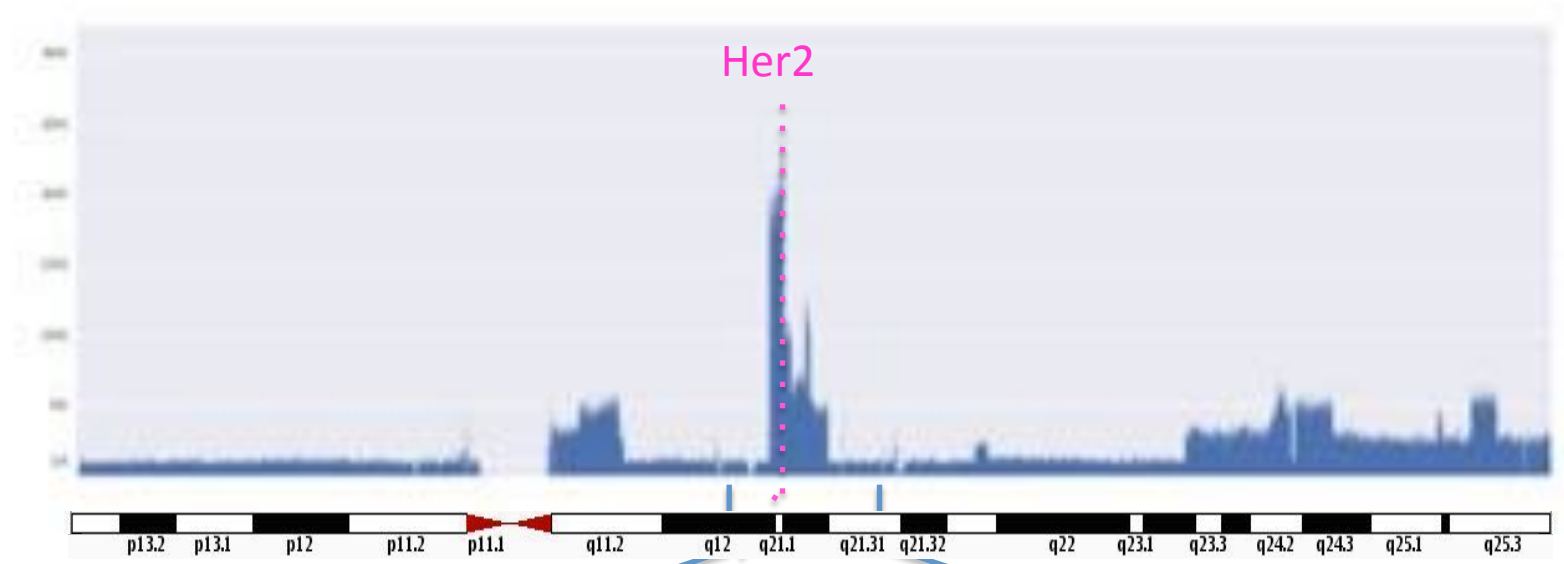
PacBio



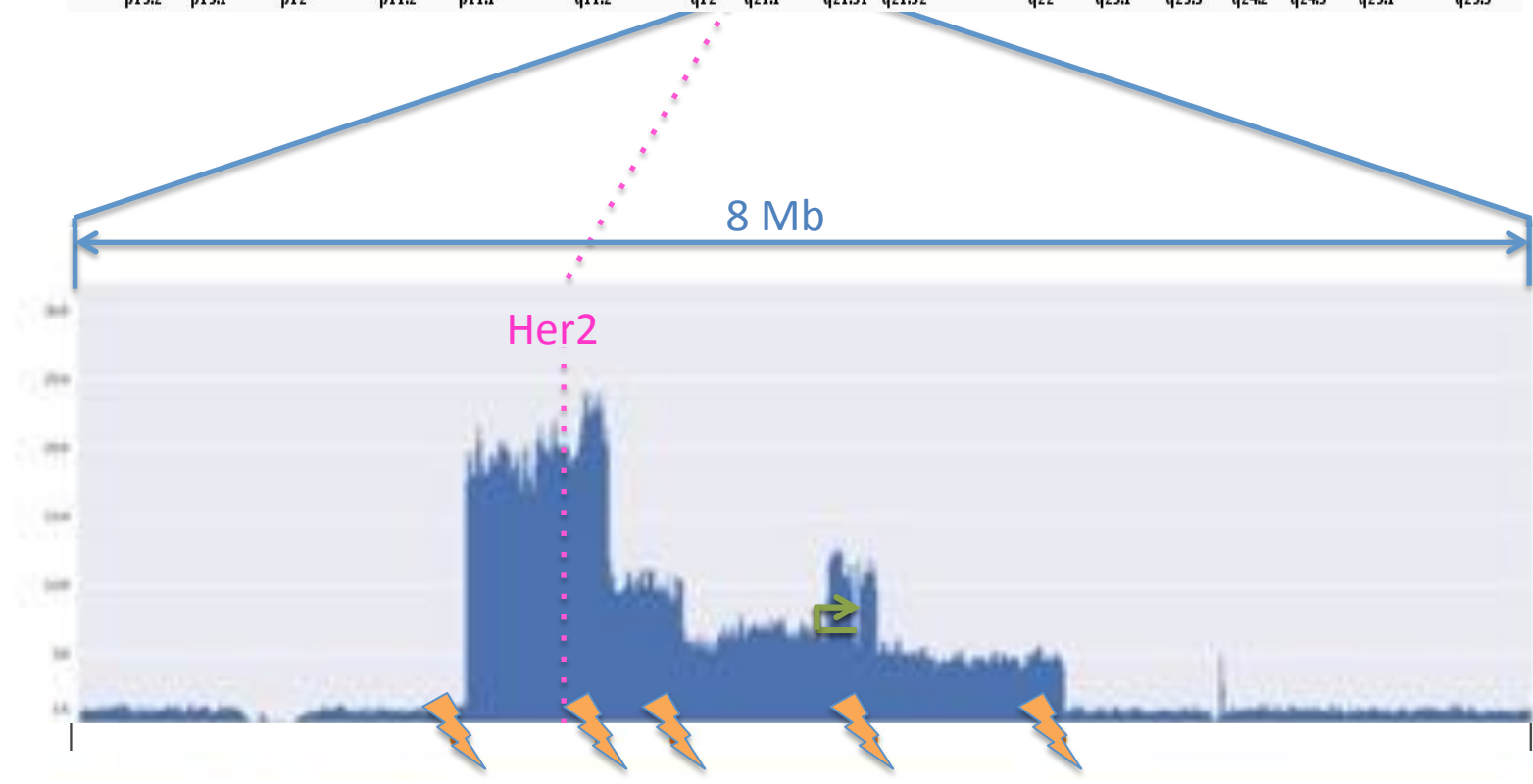
PacBio  
chr17

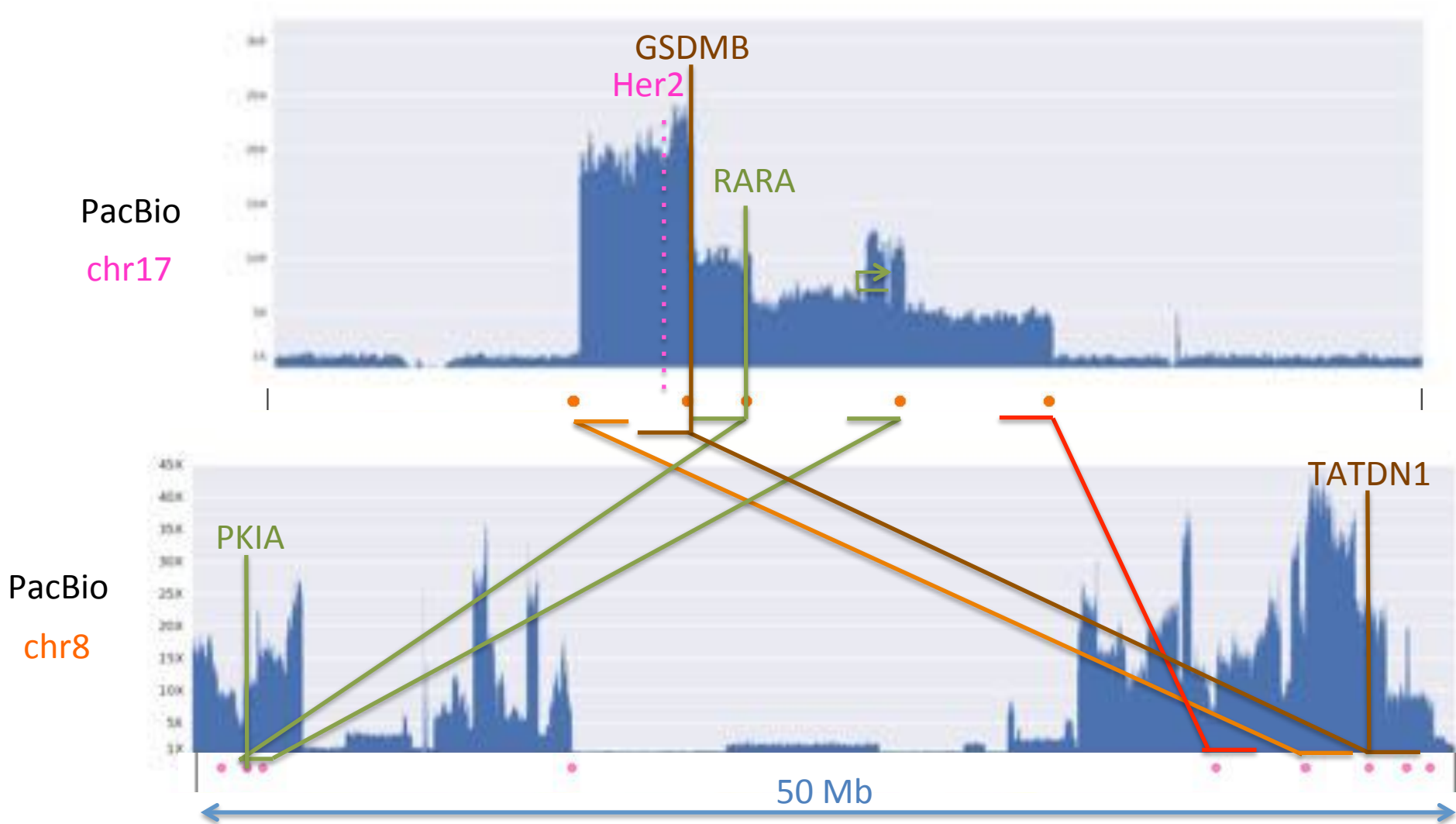


PacBio

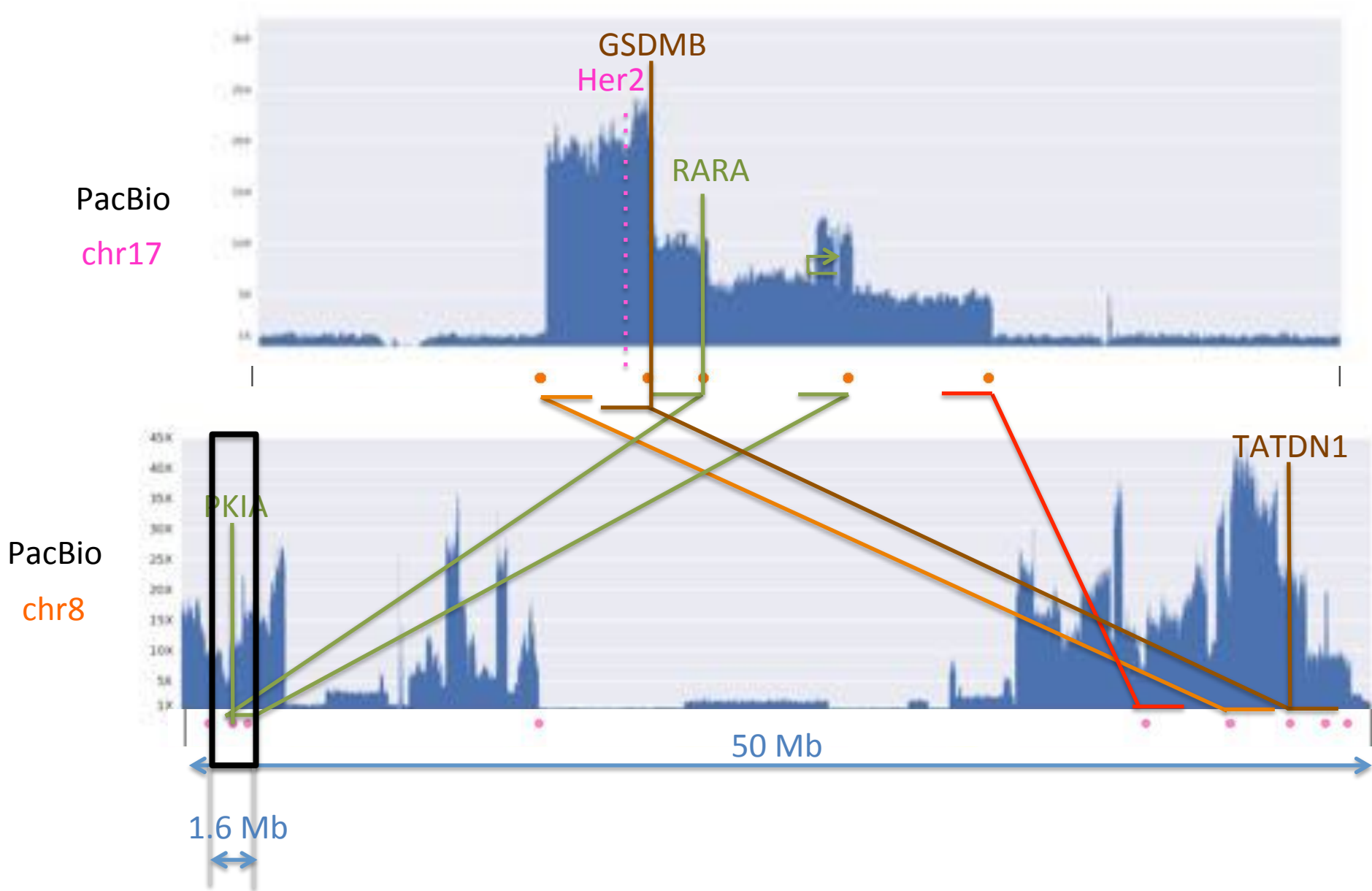


PacBio  
chr17

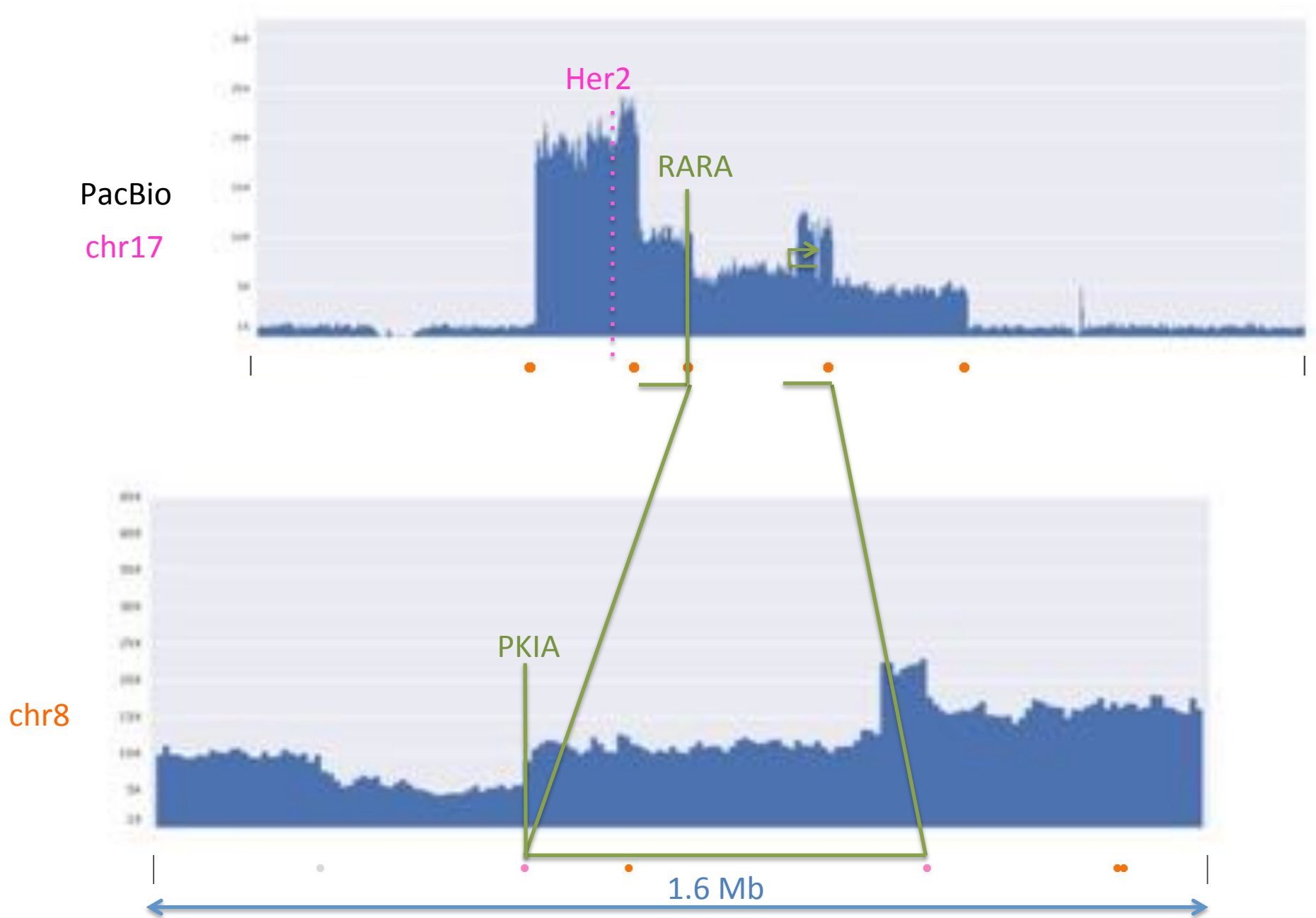




Confirmed both known gene fusions in this region



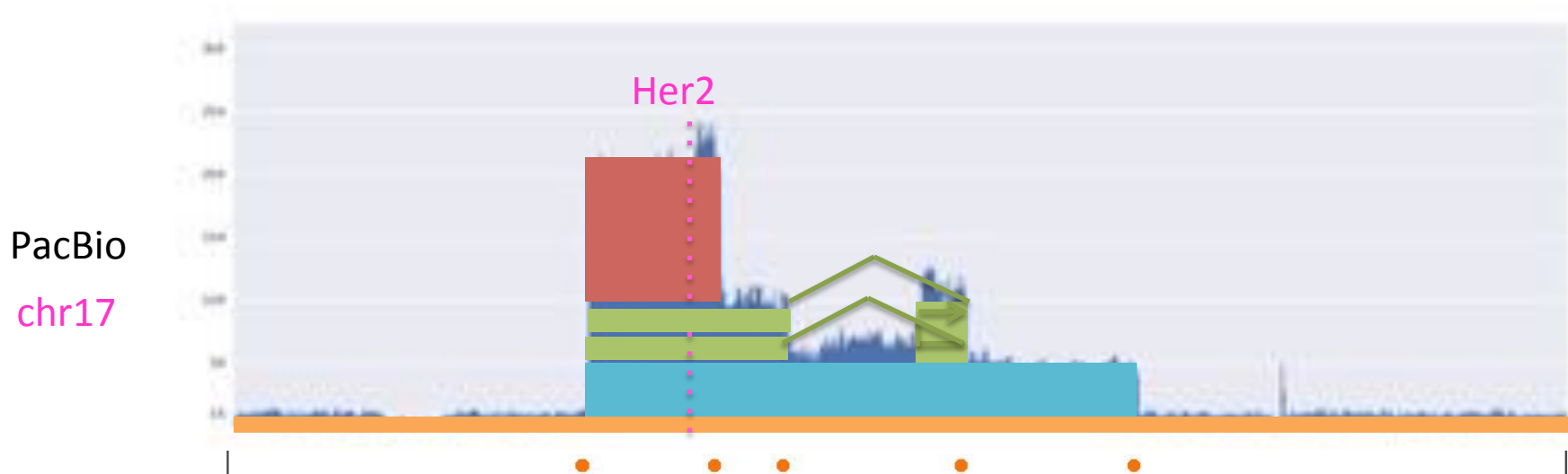
Confirmed both known gene fusions in this region



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 transcriptome analysis
- Comparison to single cell analysis of >100 individual cells

***Go see Maria's poster!***

4. Final duplication from within chromosome 8

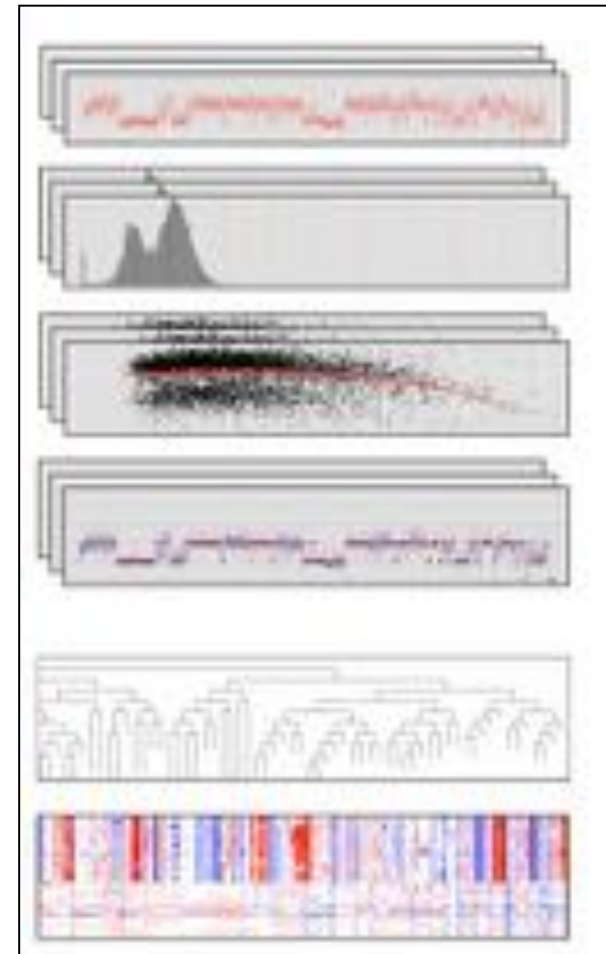
# Ginkgo

<http://qb.cshl.edu/ginkgo>



## **Single-Cell Copy Number Analysis**

- Extremely low coverage sequencing ( $\sim 1\times$ ) from amplified cells is sufficient to determine large copy number changes ( $>50\text{kbp}$ )
- 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

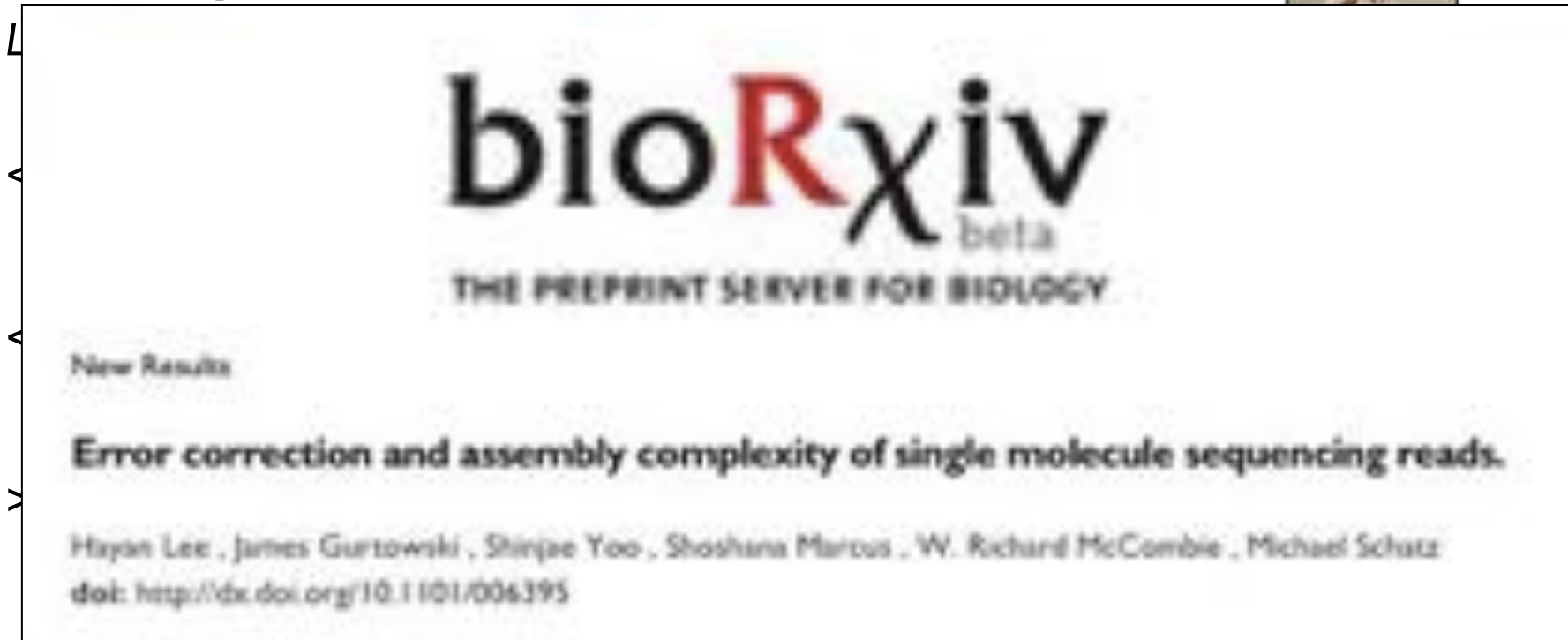


## **Interactive analysis and quality assessment of single-cell copy-number variations.**

Garvin, T., Aboukhalil, et al. (2015) *Under review*

# What should we expect from an assembly?

## Summary & Recommendations

A screenshot of a bioRxiv preprint announcement. The bioRxiv logo is at the top, with 'beta' and 'THE PREPRINT SERVER FOR BIOLOGY' below it. Under 'New Results', the title 'Error correction and assembly complexity of single molecule sequencing reads.' is displayed. The authors listed are Hyeon Lee, James Gurtowski, Shinjae Yoo, Shoshana Marcus, W. Richard McCombie, and Michael Schatz. The DOI is <http://dx.doi.org/10.1101/006395>.

bioRxiv  
beta  
THE PREPRINT SERVER FOR BIOLOGY

New Results

**Error correction and assembly complexity of single molecule sequencing reads.**

Hyeon 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](mailto:mschatz@cshl.edu)

***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 Vulture  
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**

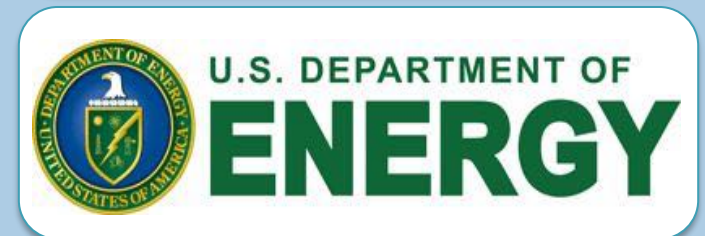
Susan McCouch  
Lyza Maron  
Mark Wright

## **OICR**

John McPherson  
Karen Ng  
Timothy Beck  
Yogi Sundaravadanam

## **NBACC**

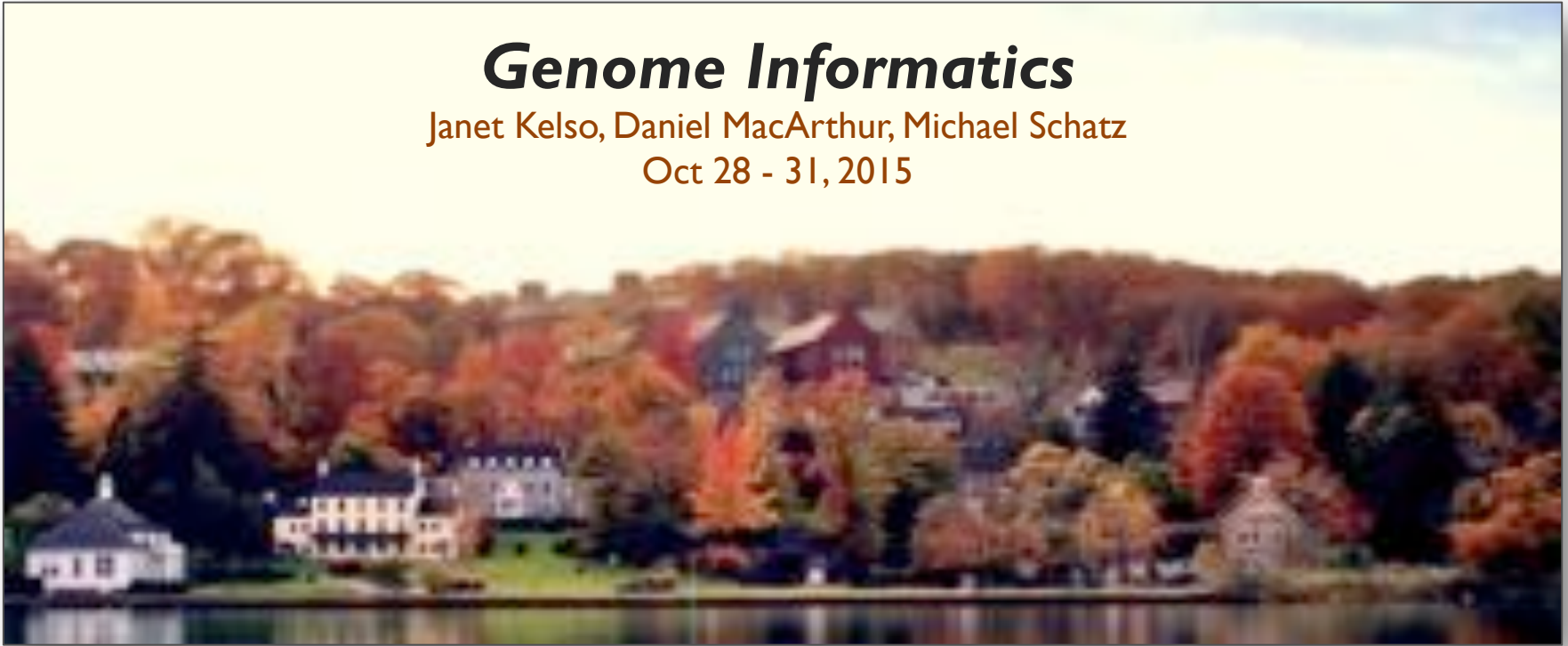
Adam Phillippy  
Serge Koren



# ***Genome Informatics***

Janet Kelso, Daniel MacArthur, Michael Schatz

Oct 28 - 31, 2015



# Thank you

<http://schatzlab.cshl.edu>

@mike\_schatz