Algorithms for studying the structure and function of genomes

Michael Schatz

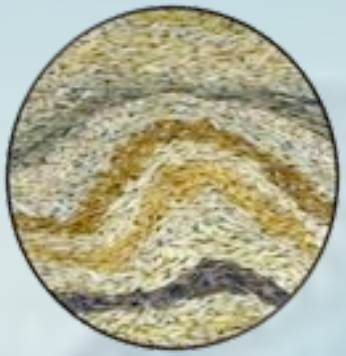
April 7, 2015
LIIGH UNAM

Schatzlab Overview

Human Genetics

Role of mutations in disease

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



Plant Biology

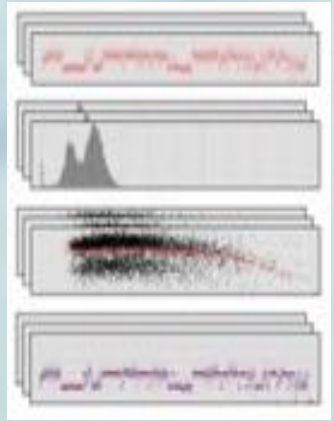
Genomes & Transcriptomes

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

Algorithmics & Systems Research

Ultra-large scale biocomputing

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



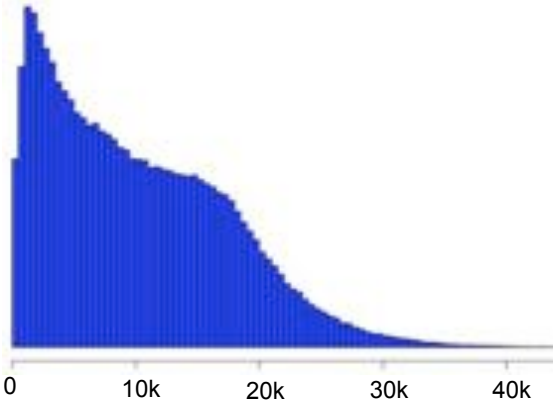
Single Cell & Single Molecule

CNVs, SVs, & Cell Phylogenetics

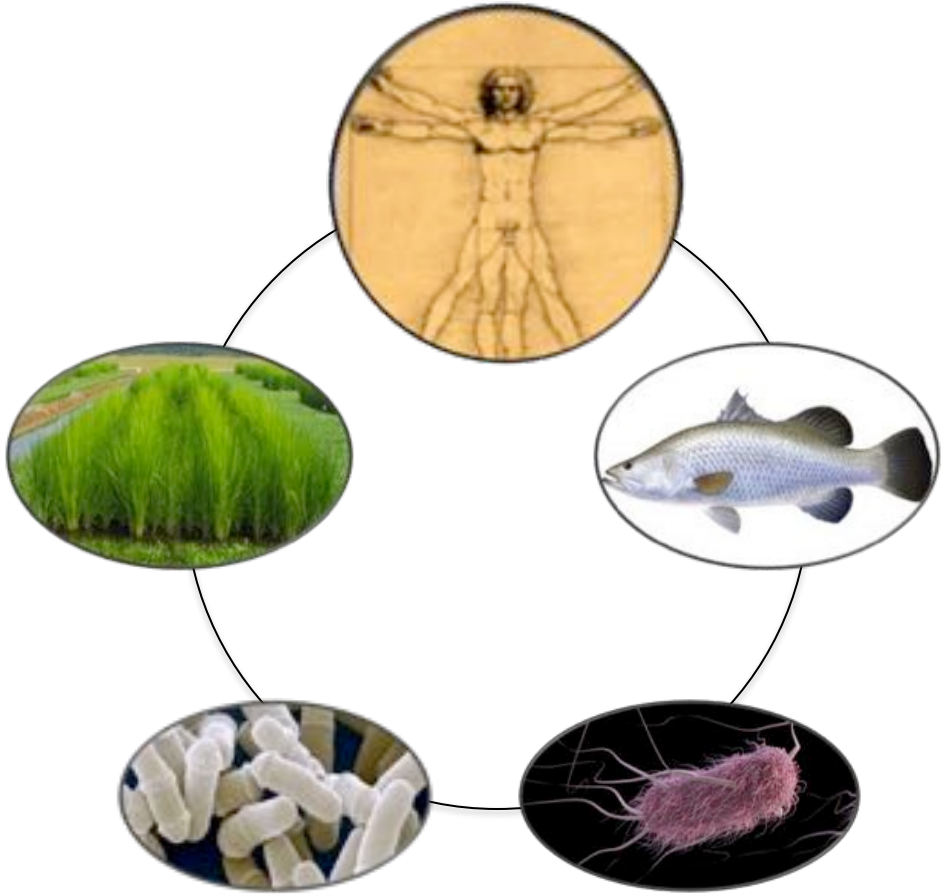
Garvin *et al.* (2014)
Roberts *et al.* (2013)

3rd Gen Long Read Sequencing

PacBio RS II

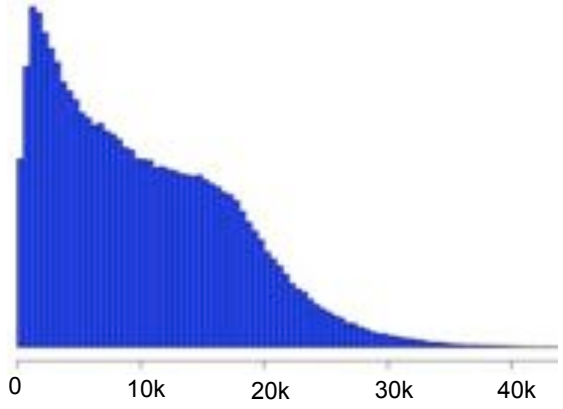


CSHL/PacBio

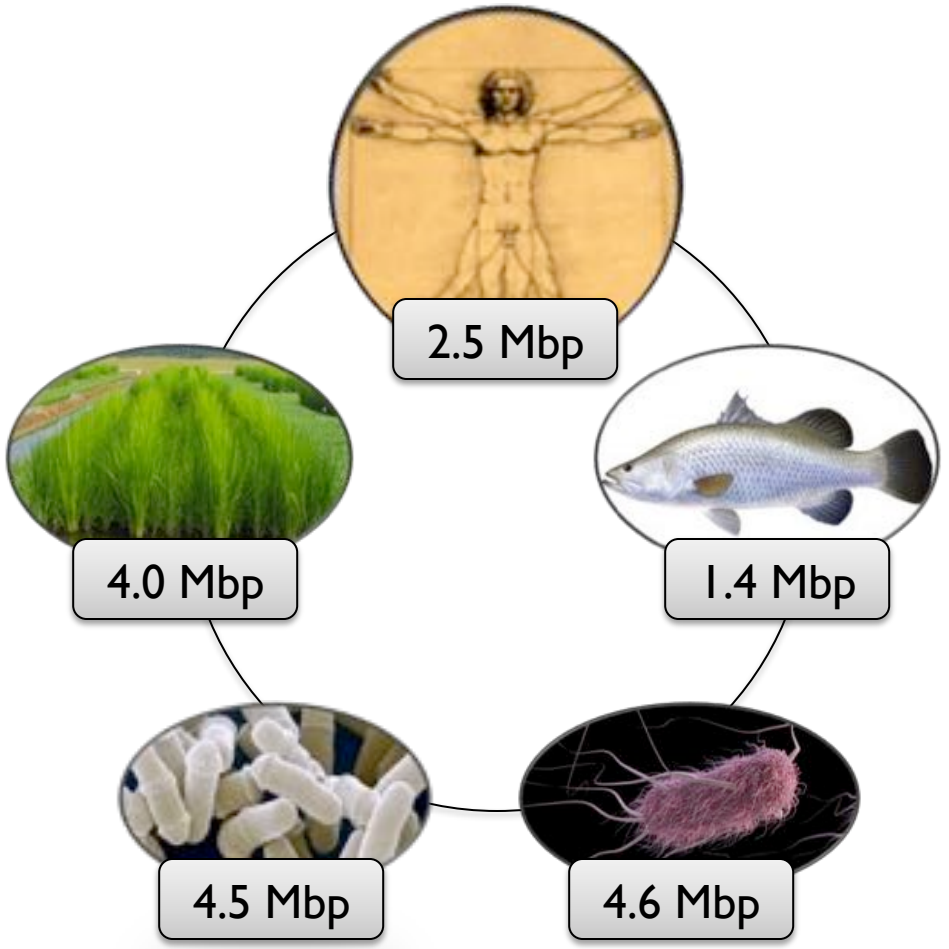


3rd Gen Long Read Sequencing

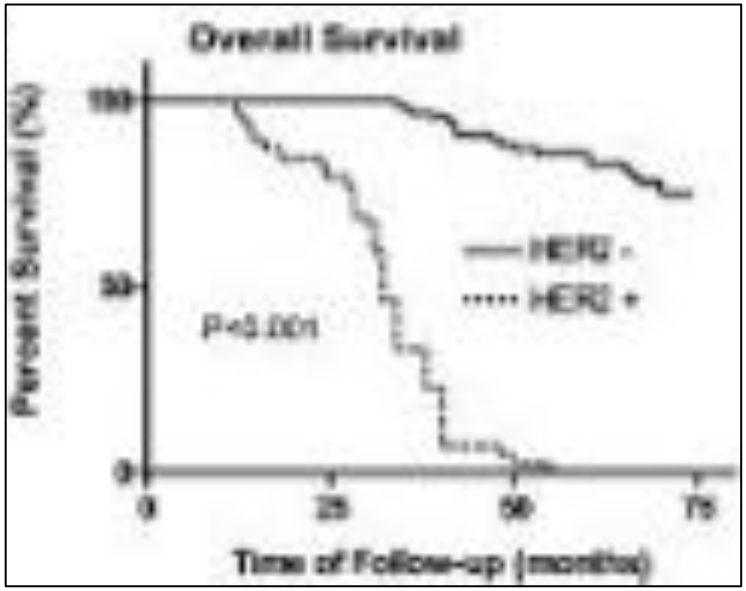
PacBio RS II



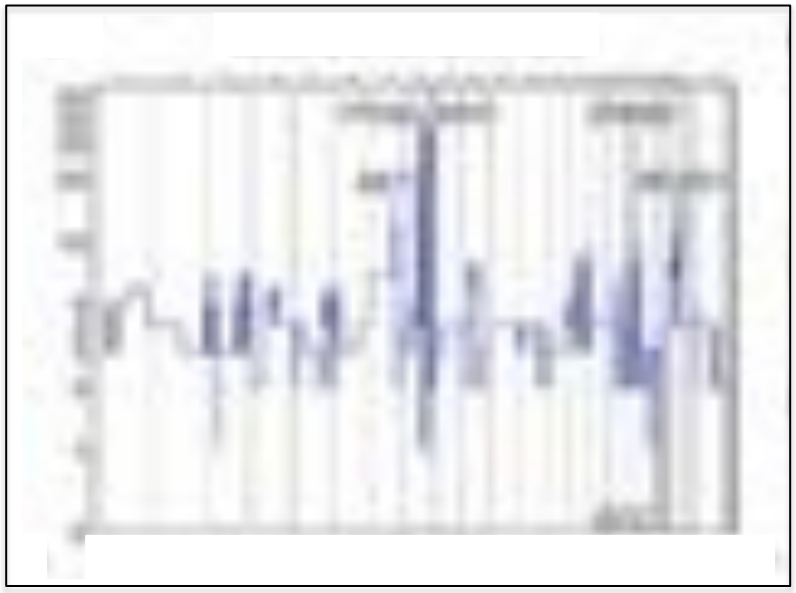
CSHL/PacBio



Long Read Sequencing of SK-BR-3



(Wen-Sheng et al, 2009)



(Navin et al, 2011)

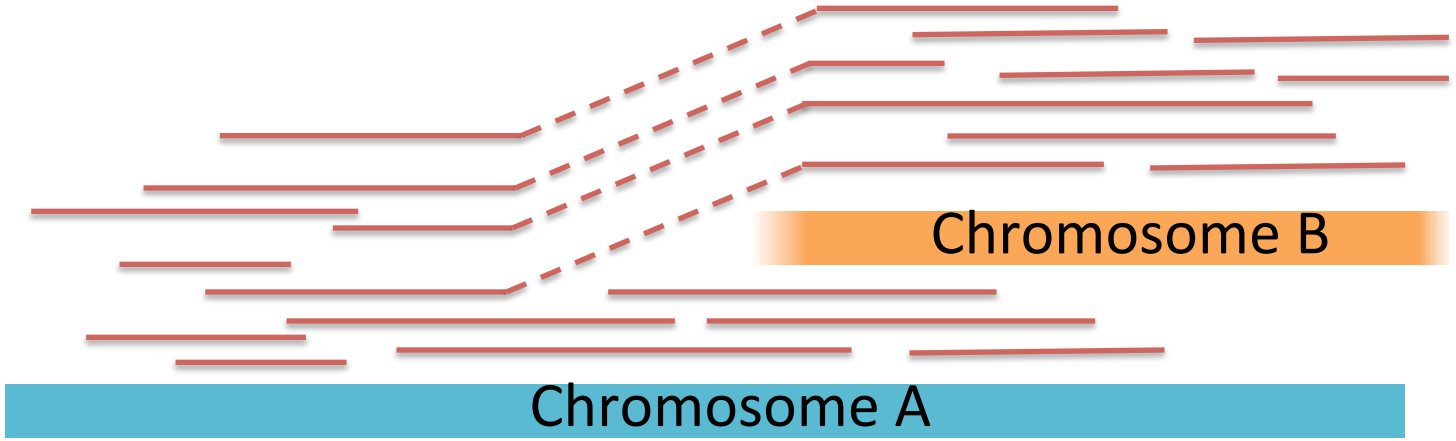
Long read PacBio sequencing of SK-BR-3 breast cancer cell line

- Her2+ breast cancer is one of the most deadly forms of the disease
- SK-BR-3 is one of the most important models, known to have widespread CNVs

- Currently have 72x coverage with long read PacBio sequencing (mean: ~10kbp)
- Analyzing breakpoints in an attempt to infer the mutation history, especially around HER2

In collaboration with McCombie (CSHL) and McPherson (OICR) labs

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*

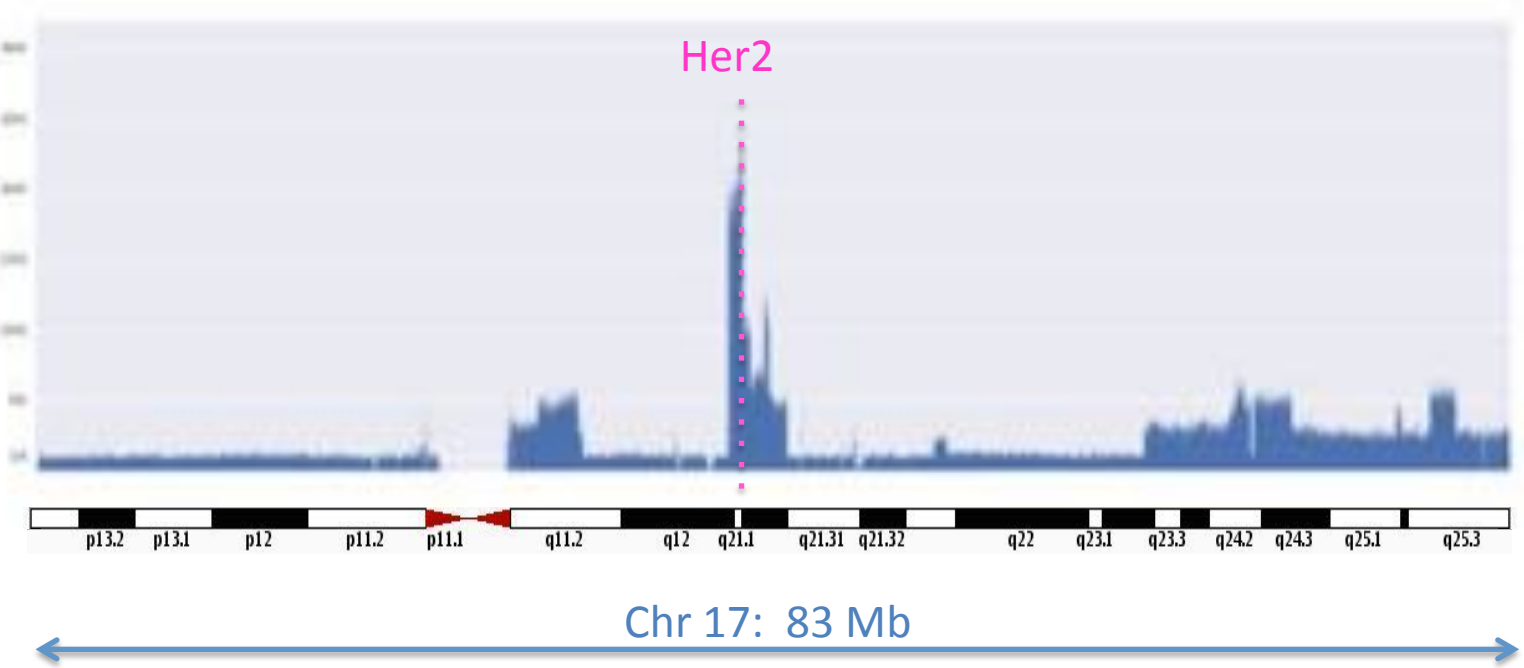
Localized HGAP + Celera Assembler + MUMmer

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

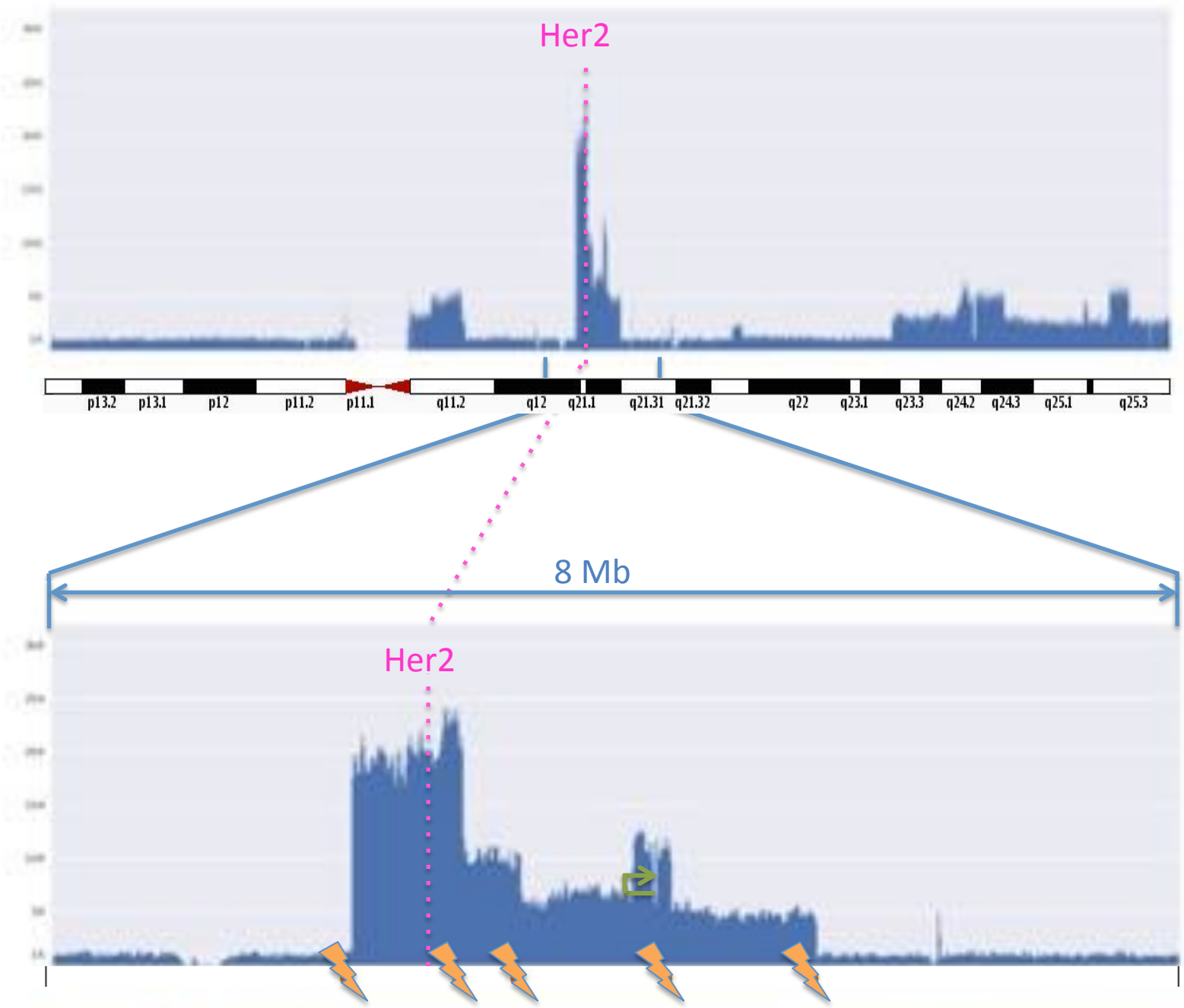
DNAnexus-enabled version of Falcon

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

PacBio

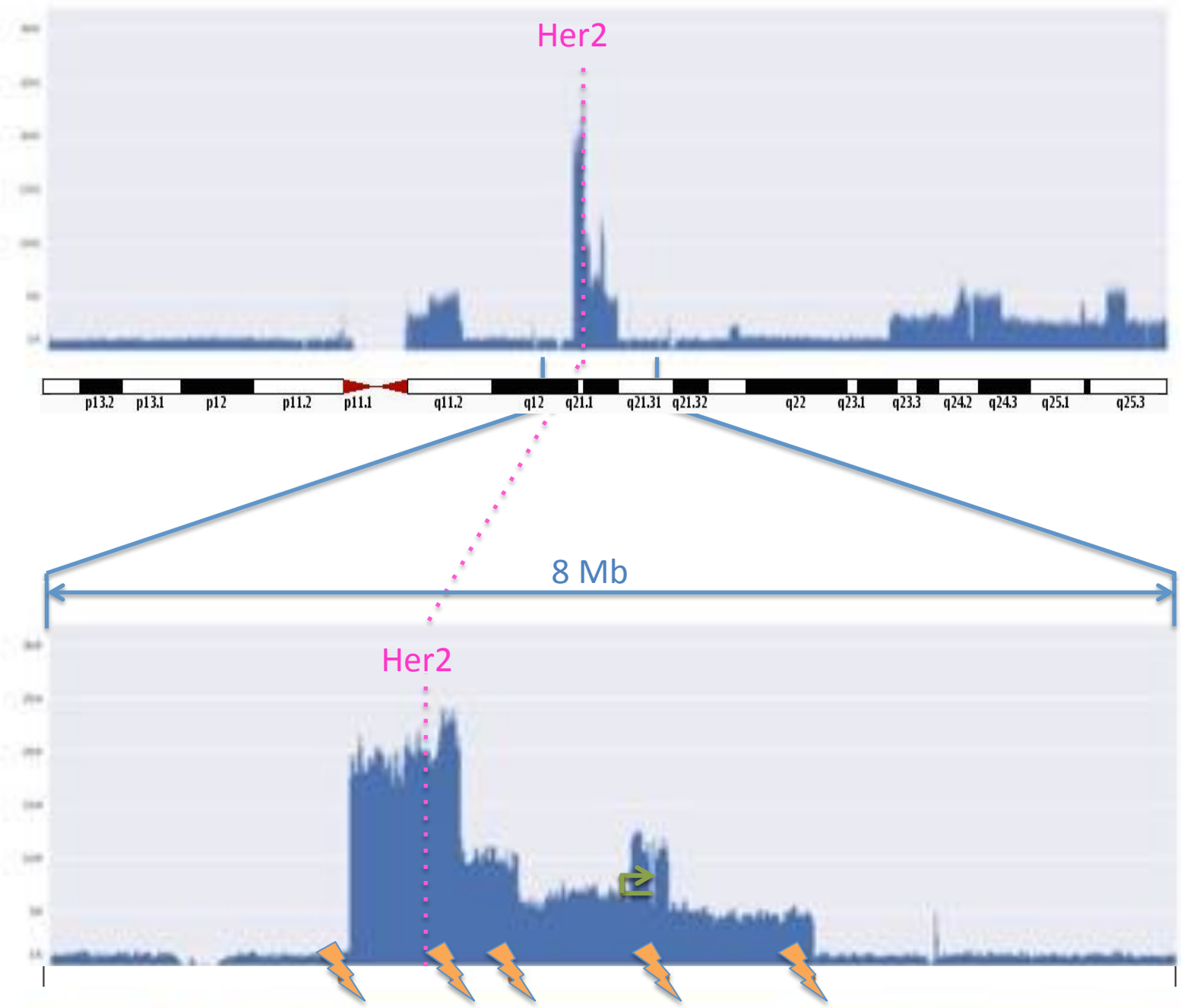

Her2
p13.2
p13.1
p12
p11.2
p11.1
q11.2
q12
q21.1
q21.31
q21.32
q22
q23.1
q23.3
q24.2
q24.3
q25.1
q25.3
Chr 17: 83 Mb

PacBio

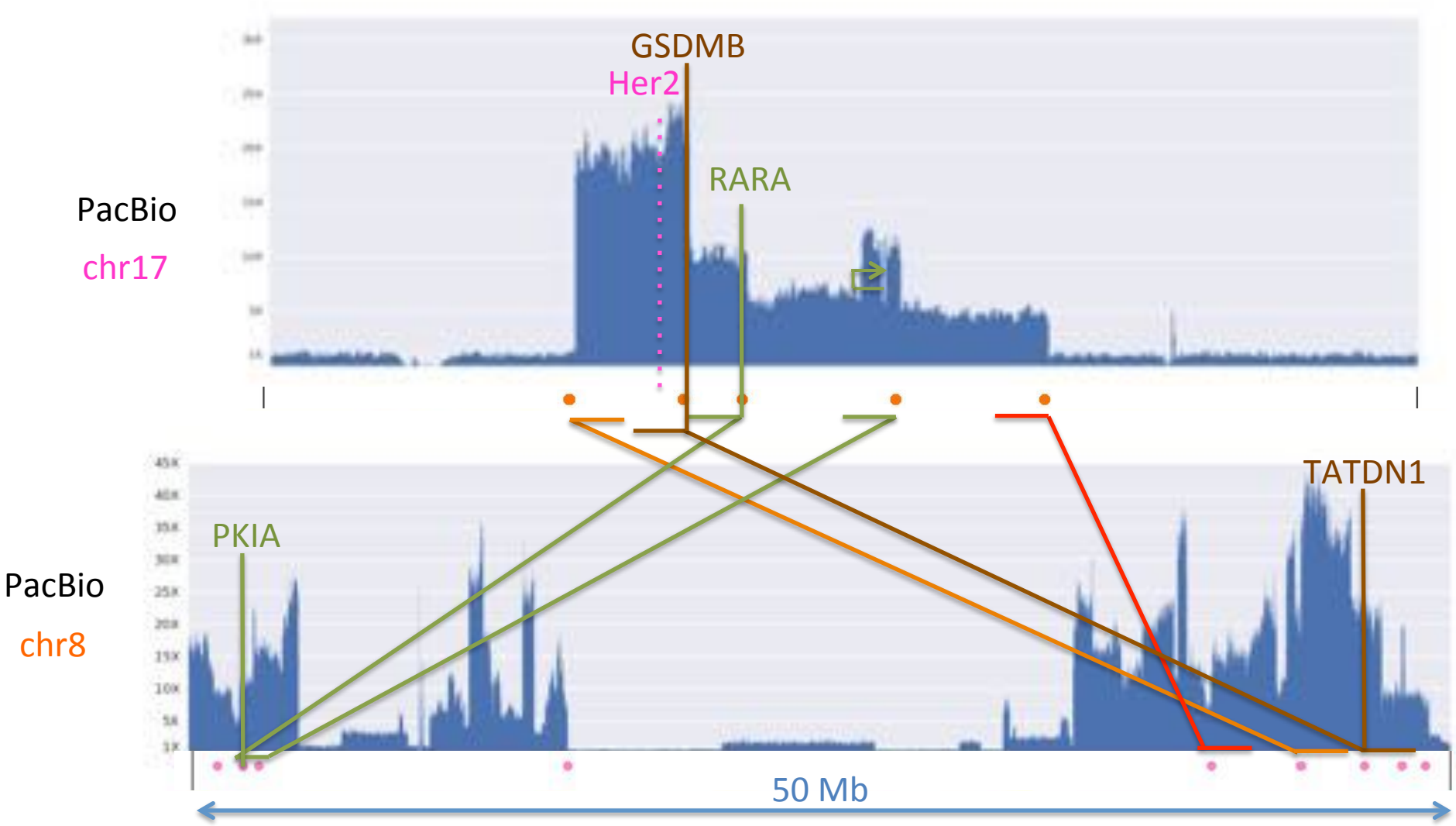


PacBio
chr17

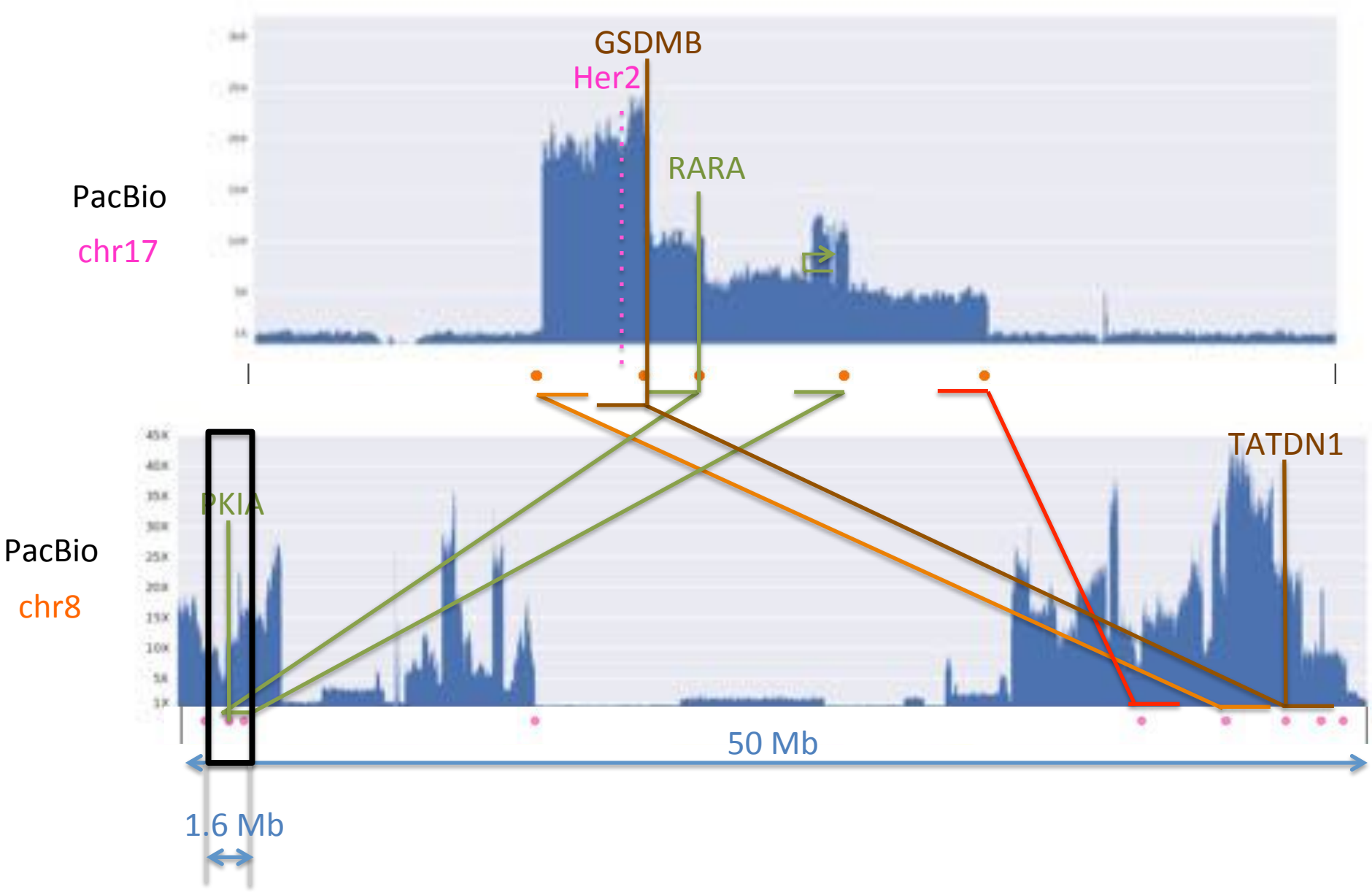
PacBio



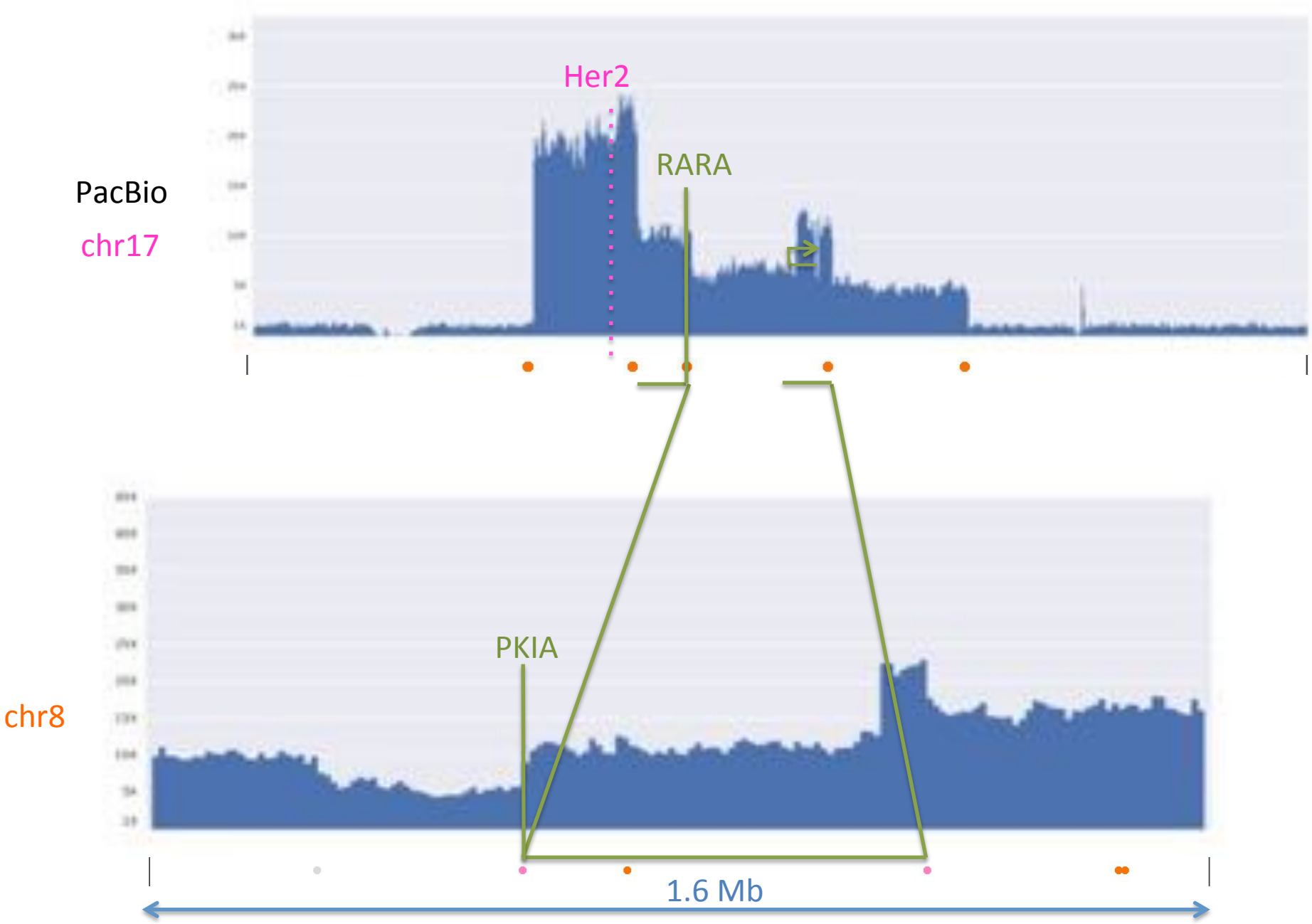
PacBio
chr17



Confirmed both known gene fusions in this region



Confirmed both known gene fusions in this region

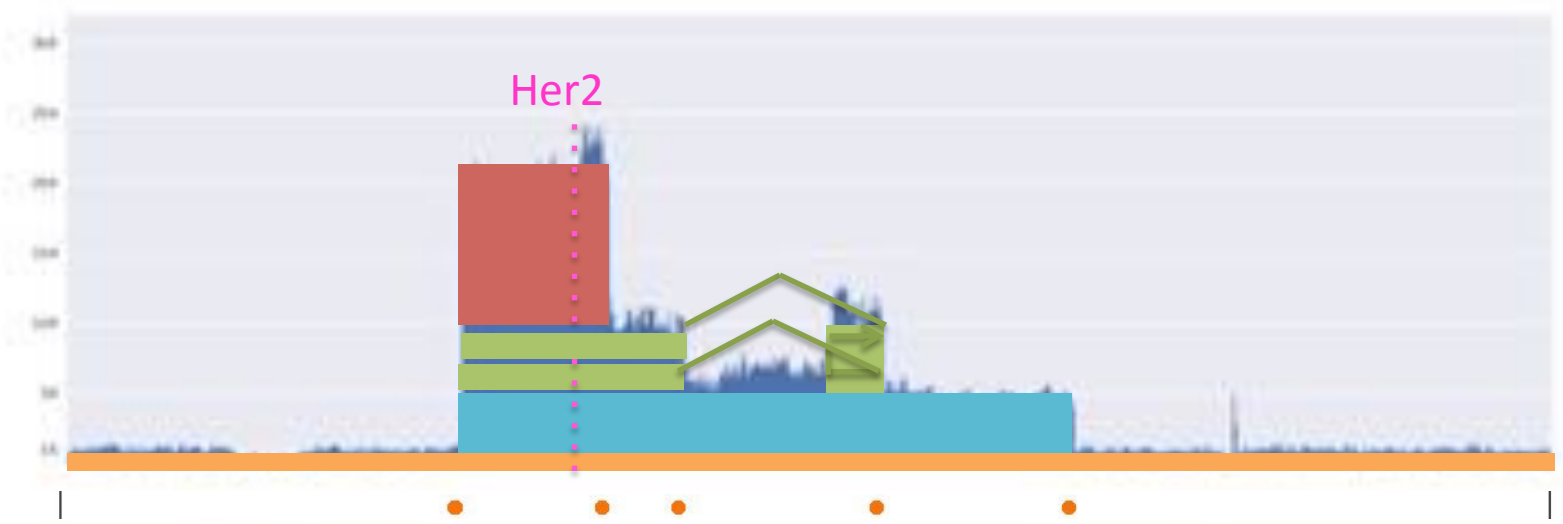


Joint coverage and breakpoint analysis to discover underlying events

Cancer lesion Reconstruction

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

Cancer lesion Reconstruction

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

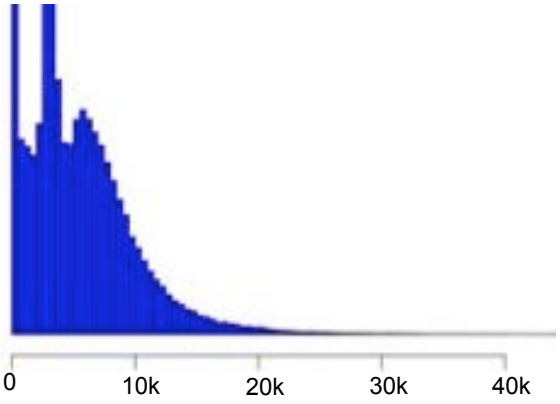
http://schatzlab.cshl.edu

4. Final duplication from within chromosome 8

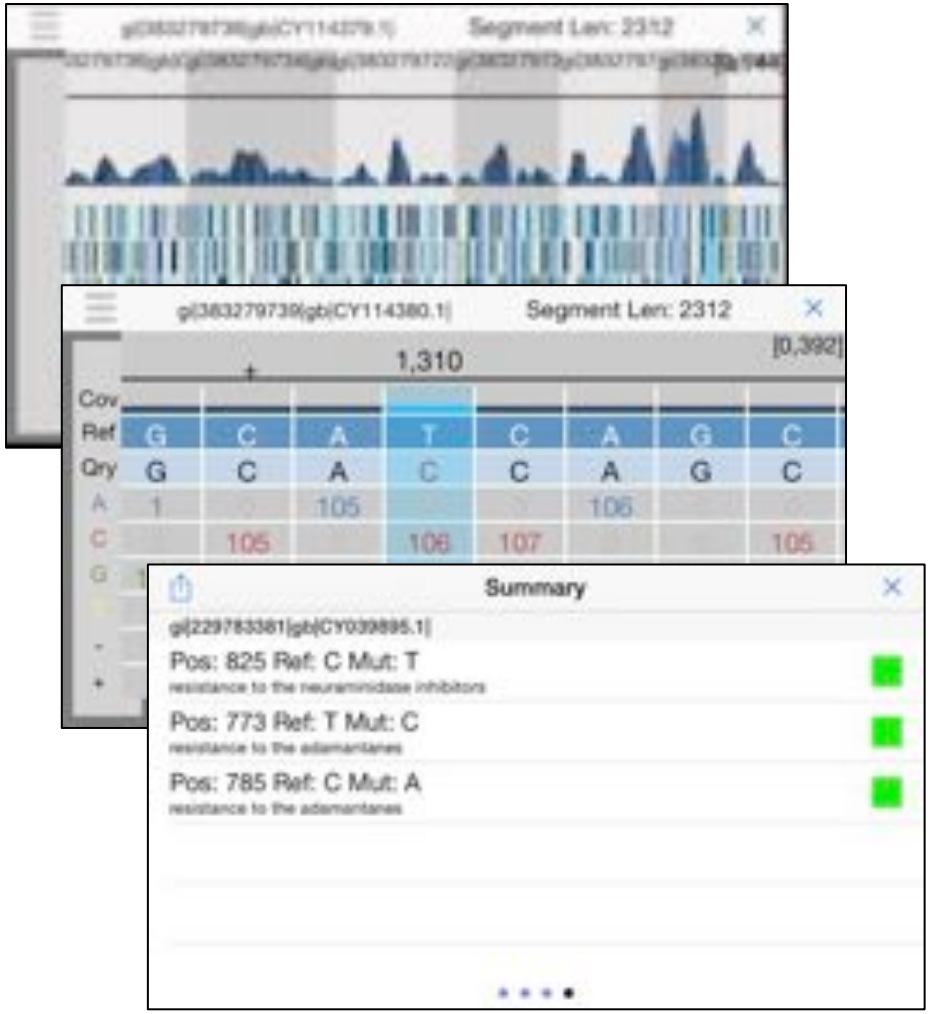
Genomic Futures?

Mobile Sequencing

Oxford Nanopore



CSHL/ONT



Understanding Genome Structure & Function

Reference quality genome assembly is here

- Driven by new technologies for long read sequencing
- Provide us new insights into the origins of disease, the stages of development, and the forces of evolution

Focus on population analysis

- Large scale sequencing of many individuals, many cells, & many assays
- Shift from relatively straightforward analysis of protein coding changes into more and more subtle signals across the genome and environment
- Informatics is the key for integrating these data all together

Ultimately the discoveries will come from the next generation of students and researchers!

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®

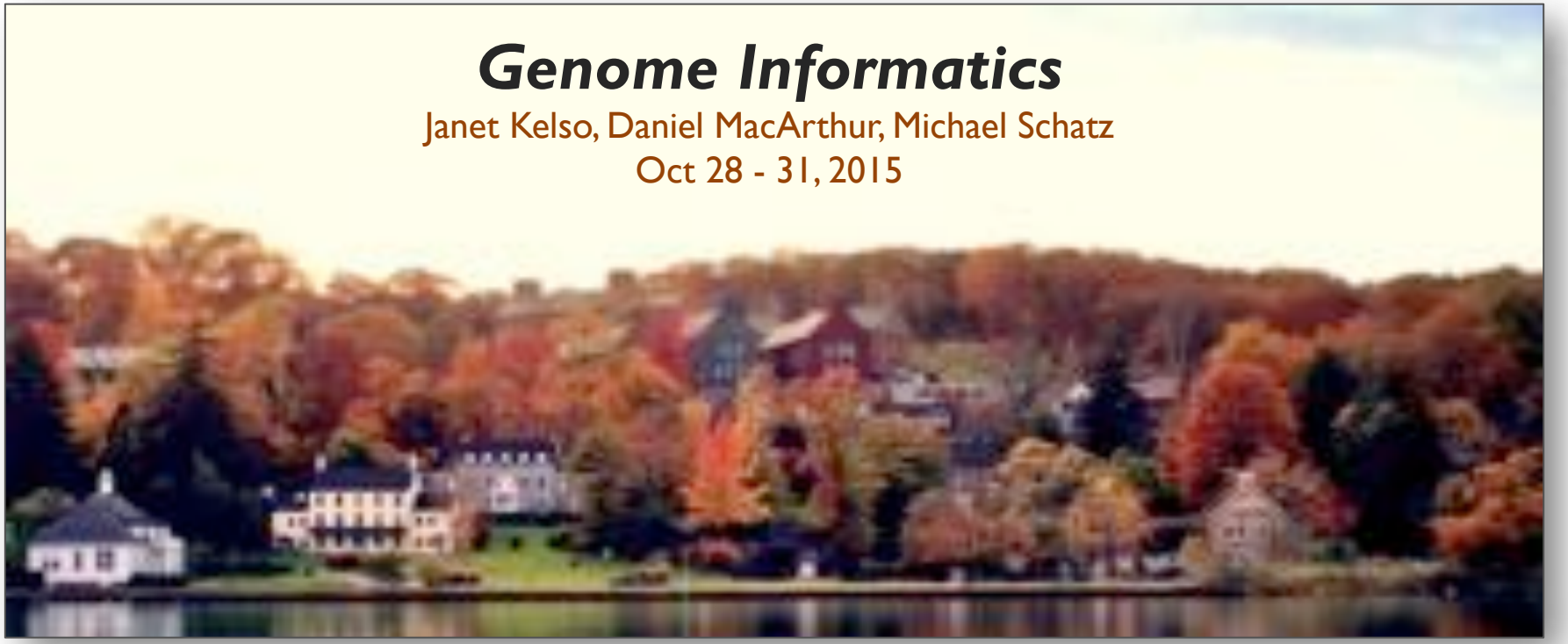
NSF

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