Single Cell and Single Molecule Analysis of Cancer

Michael Schatz

April 20, 2015
Laufer Center Retreat
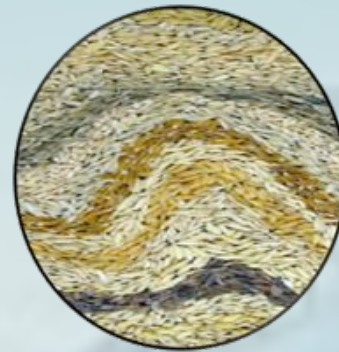
Schatzlab Overview

Human Genetics

Role of mutations in disease

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

Plant Biology
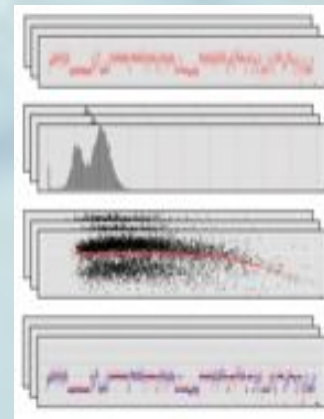
Genomes & Transcriptomes

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

Algorithmics & Systems Research

Ultra-large scale biocomputing

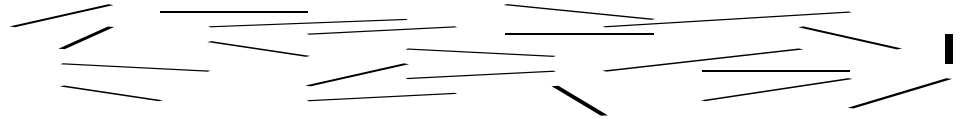Marcus *et al.* (2014)
Schatz *et al.* (2013)

Single Cell & Single Molecule

CNVs, SVs, & Cell Phylogenetics

Garvin *et al.* (2014)
Roberts *et al.* (2013)
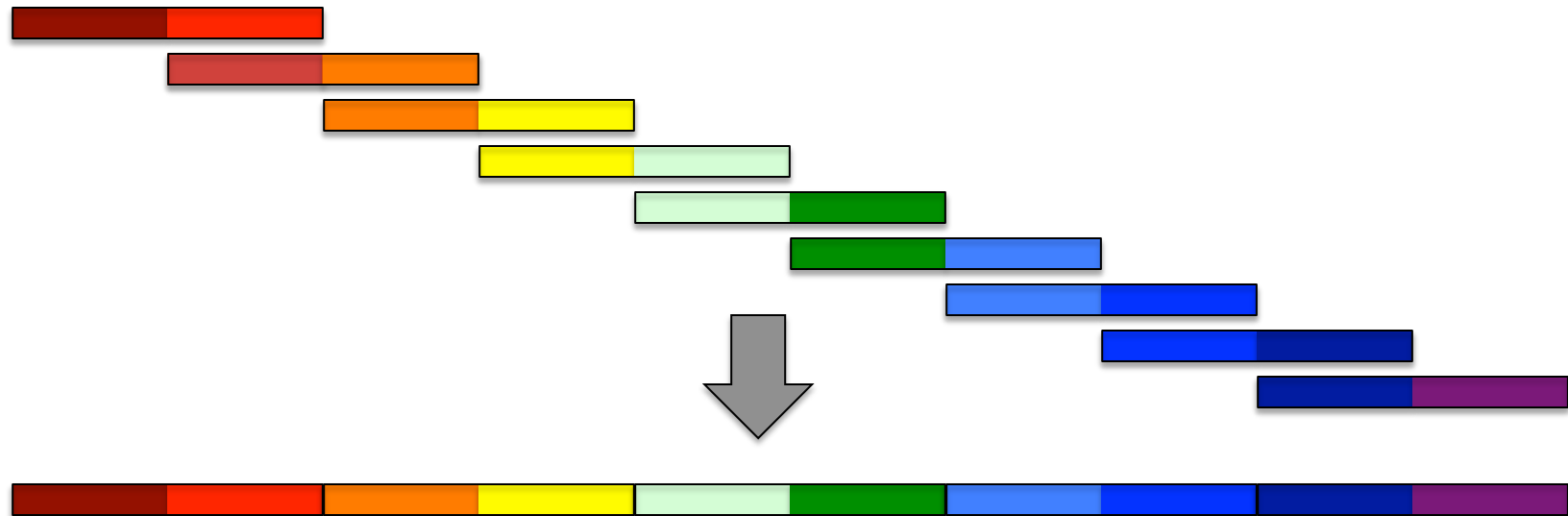
Sequence Assembly Problem

1. Shear & Sequence DNA

2. Construct assembly graph from overlapping reads

...AGCCTAGGGATGCGCGACACGT

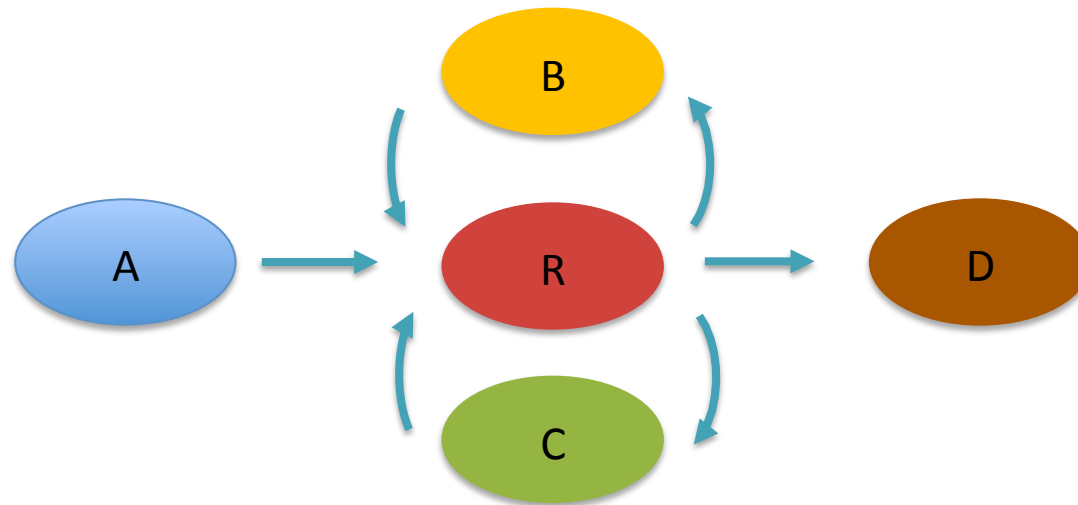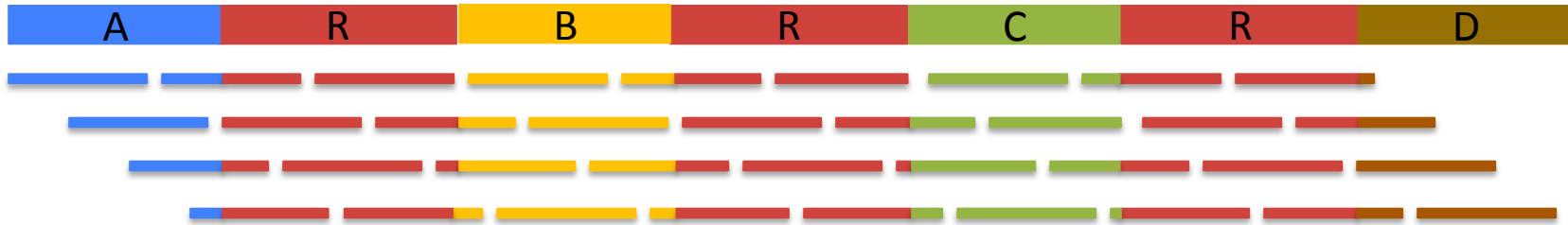GGATGCGCGACACGTCGCATATCCGGTTTGGTCAACCTCGGACGGAC
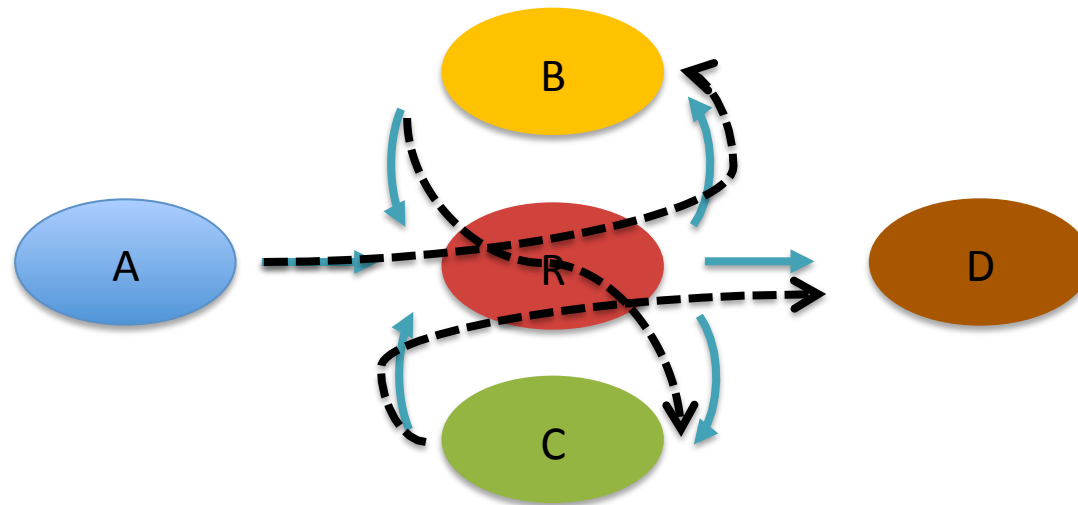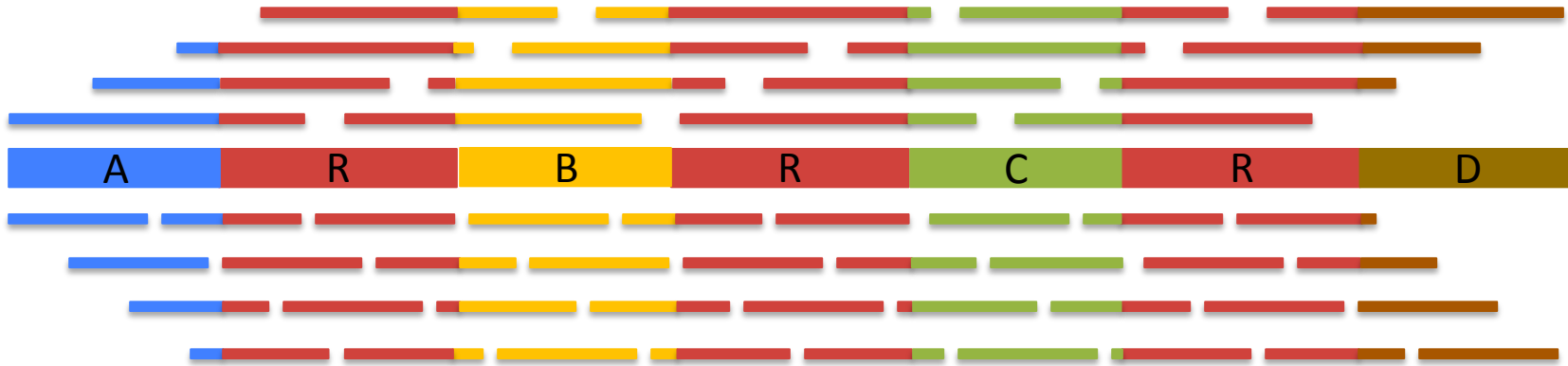
CAACCTCGGACGGACCTCAGCGAA...

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*
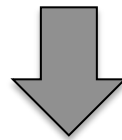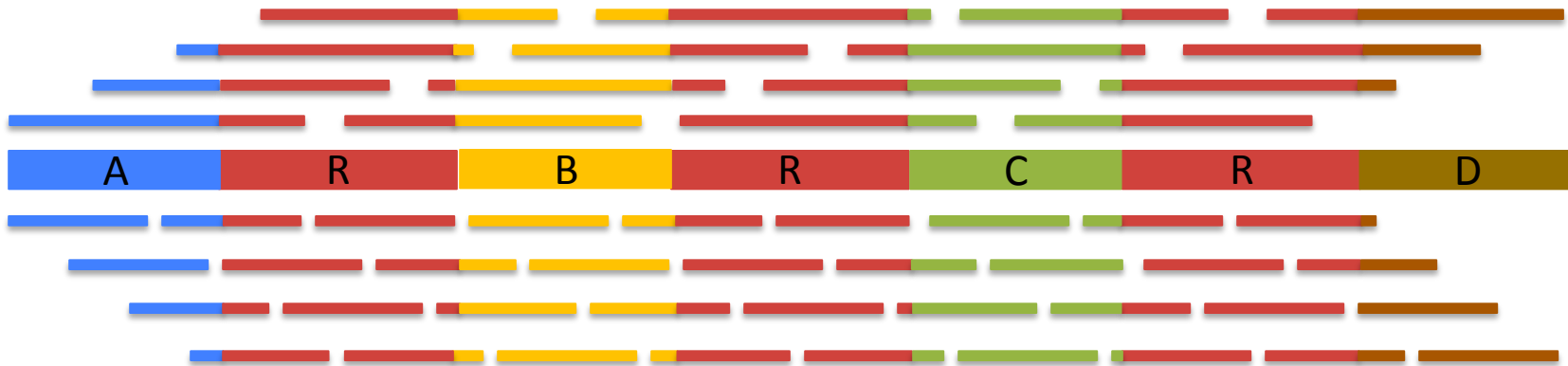
Assembly Complexity

Assembly Complexity

Assembly Complexity

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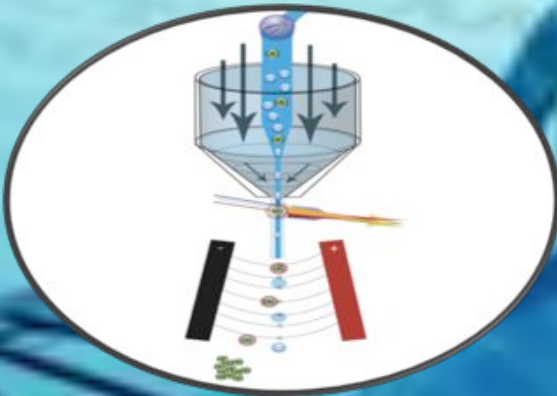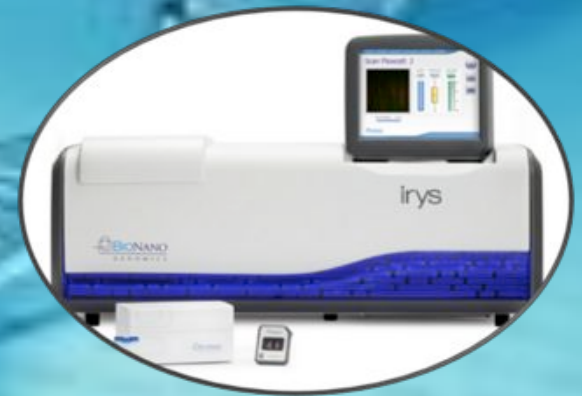Roberts, RJ, Carneiro, MO, Schatz, MC (2013) *Genome Biology.* 14:405

PacBio SMRT Sequencing

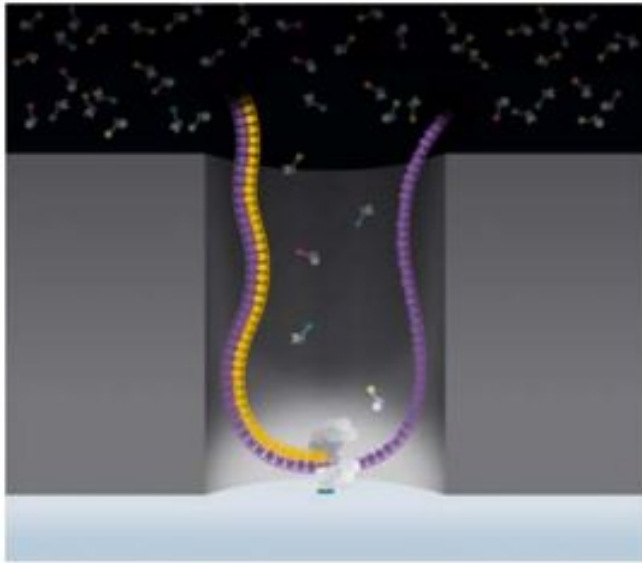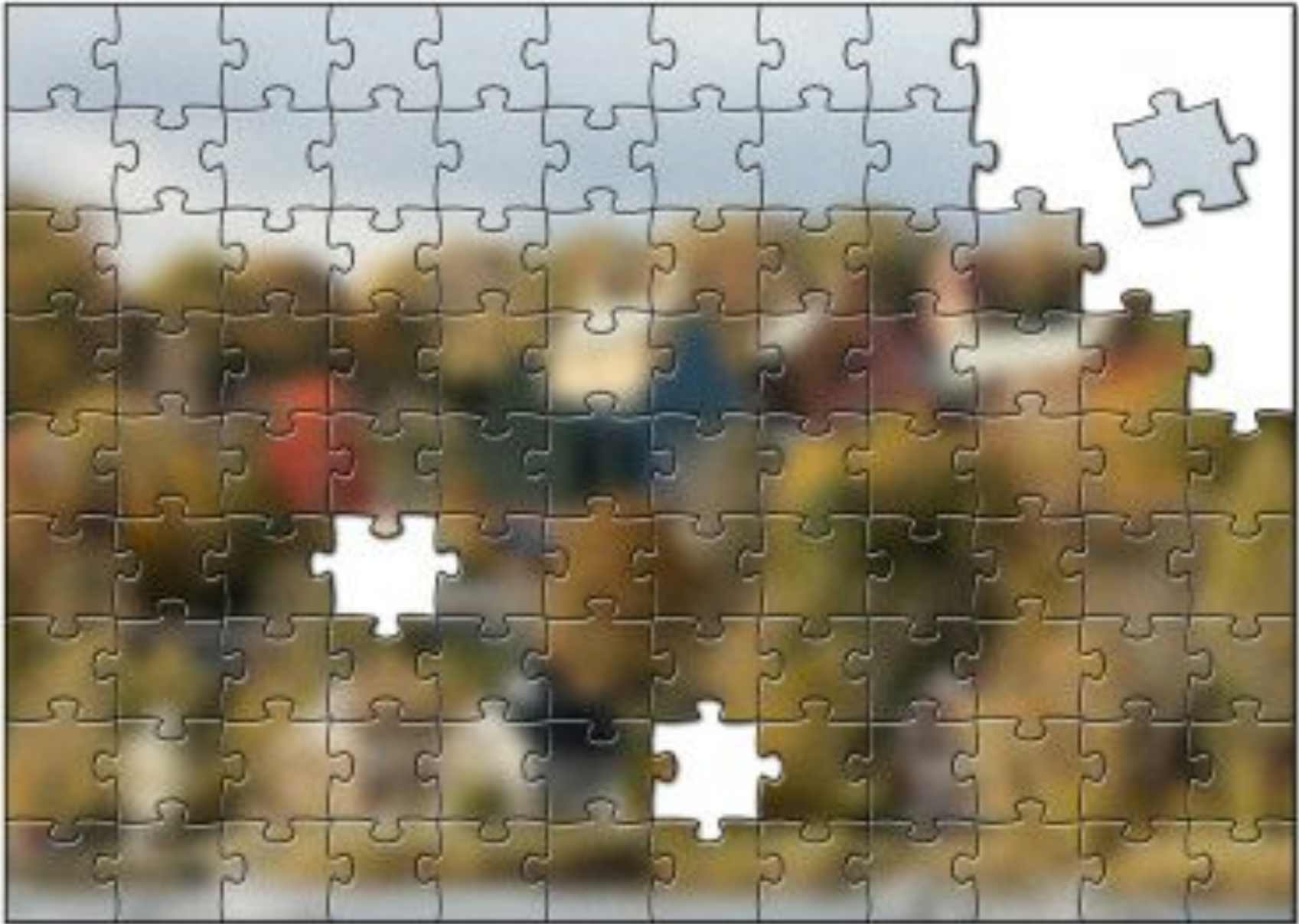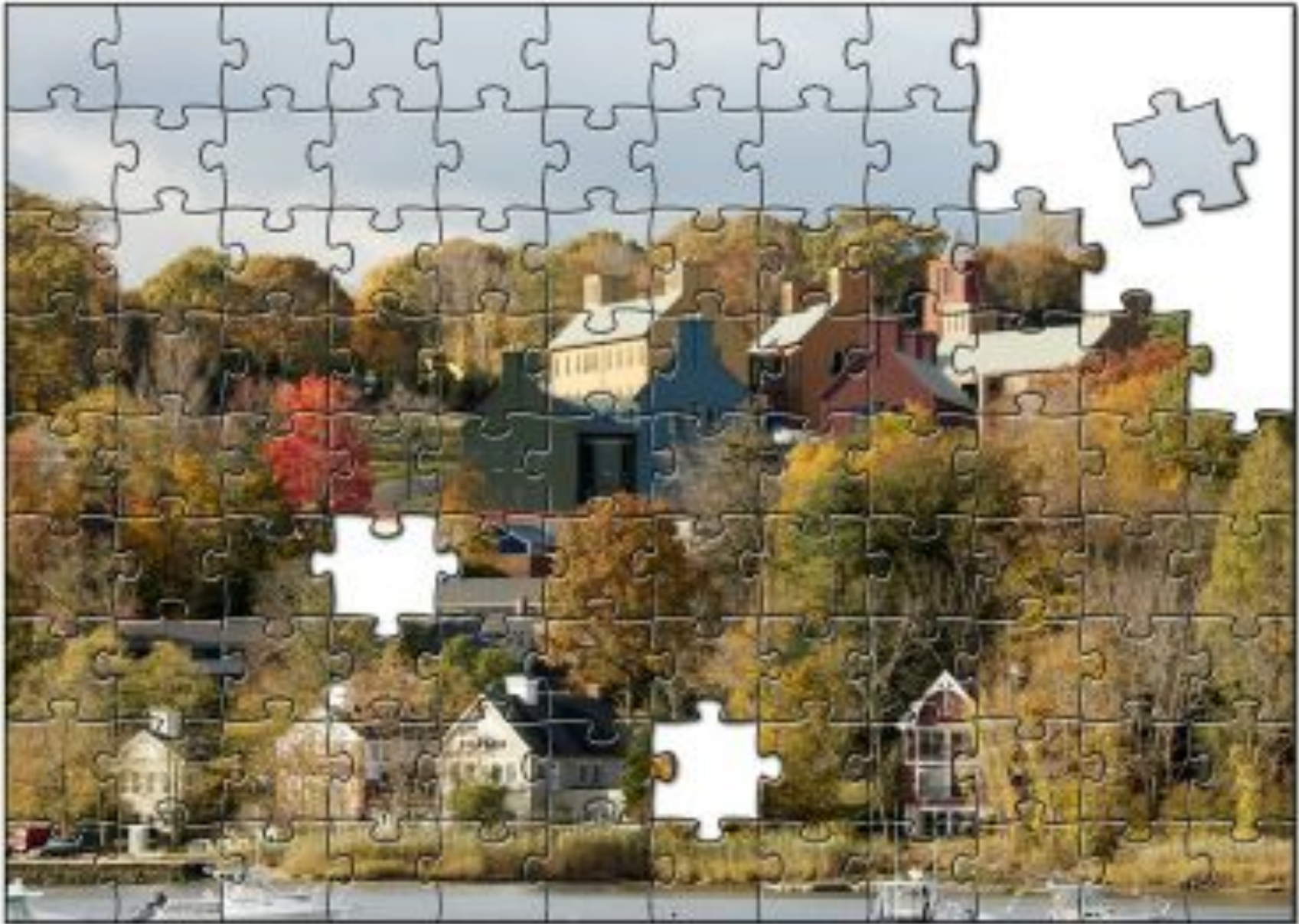Imaging of fluorescently phospholinked labeled nucleotides as they are incorporated by a polymerase anchored to a Zero-Mode Waveguide (ZMW).
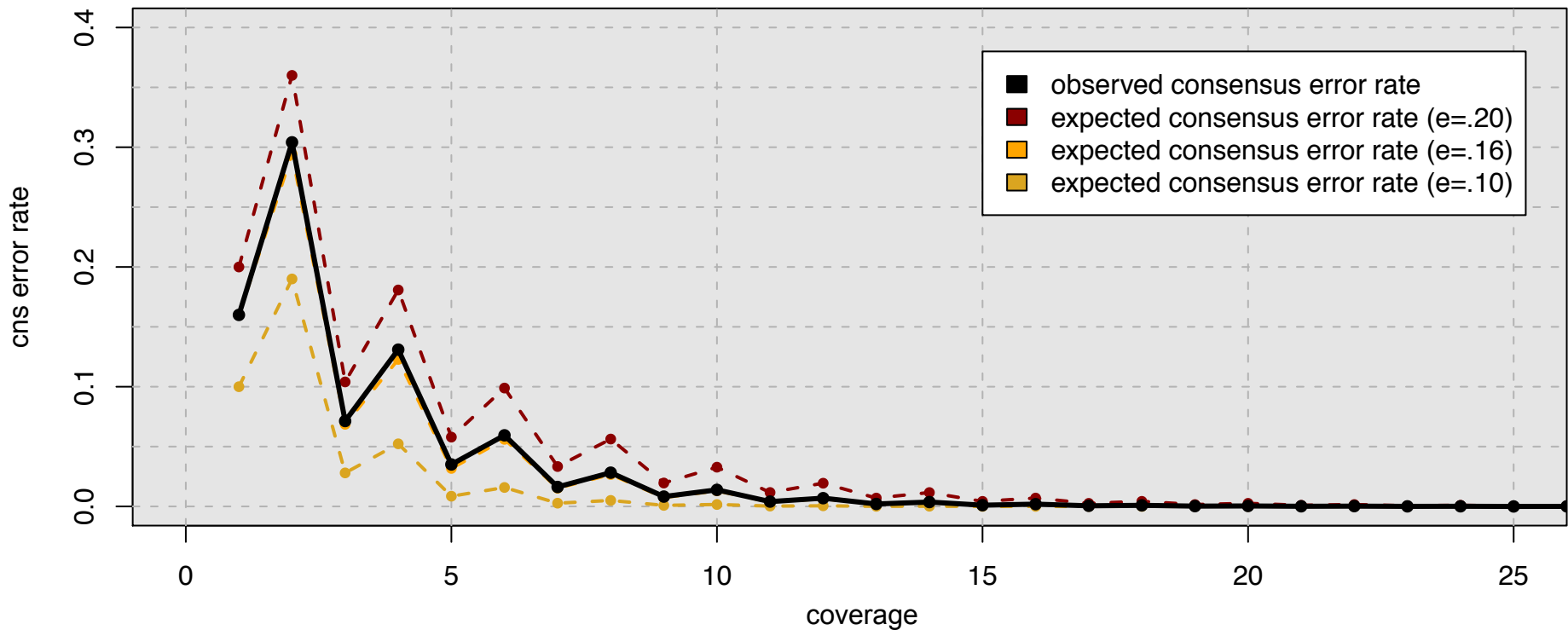

http://www.pacificbiosciences.com/assets/files/pacbio_technology_backgrounder.pdf
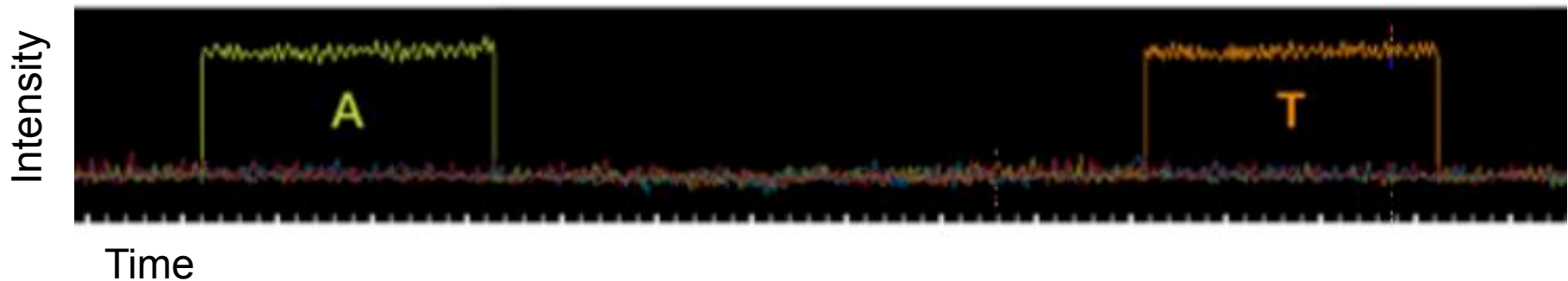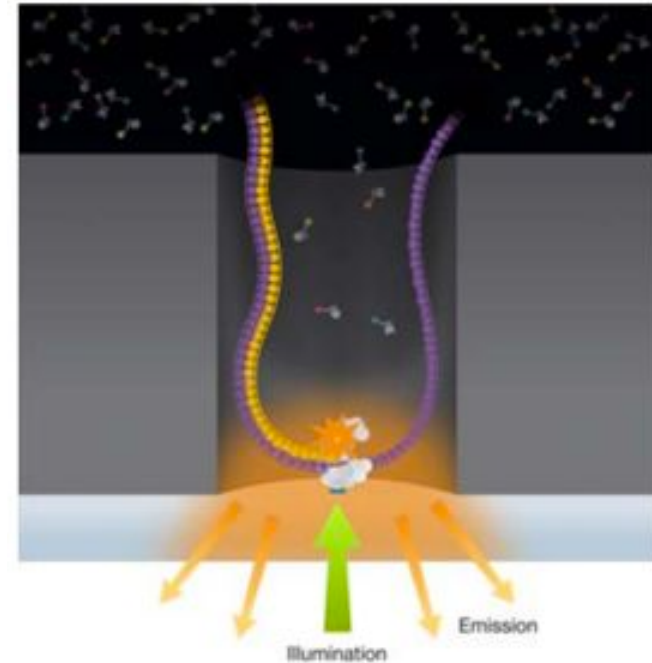
Single Molecule Sequences

"Corrective Lens" for Sequencing

Consensus Accuracy and Coverage

Legend:
- observed consensus error rate
- expected consensus error rate (e=.20)
- expected consensus error rate (e=.16)
- expected consensus error rate (e=.10)

x-axis: coverage; y-axis: cns error rate

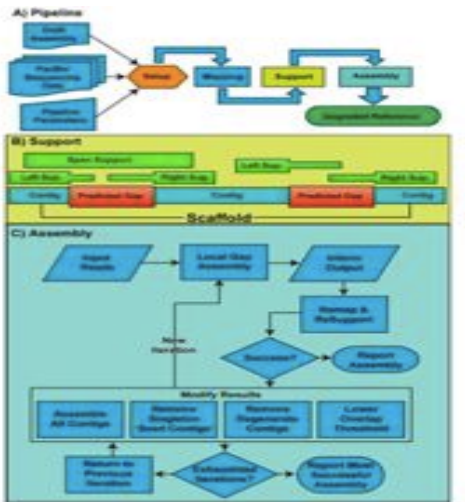## Coverage can overcome random errors

- Dashed: error model from binomial sampling

- Solid: observed accuracy

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

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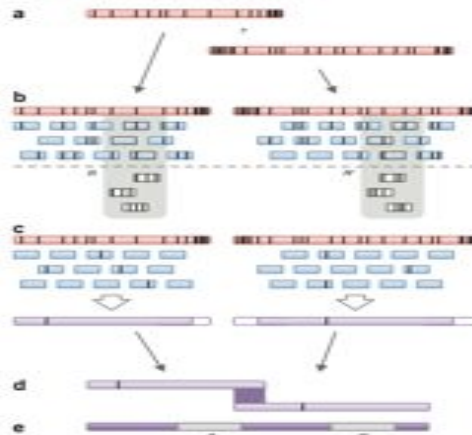
PacBio Assembly Algorithms

PBJelly	PacBioToCA & ECTools	HGAP & Quiver
Gap Filling and Assembly Upgrade	**Hybrid/PB-only Error Correction**	**PB-only Correction & Polishing**
English *et al* (2012) *PLOS One.* 7(11): e47768	Koren, Schatz, *et al* (2012) *Nature Biotechnology.* 30:693–700	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
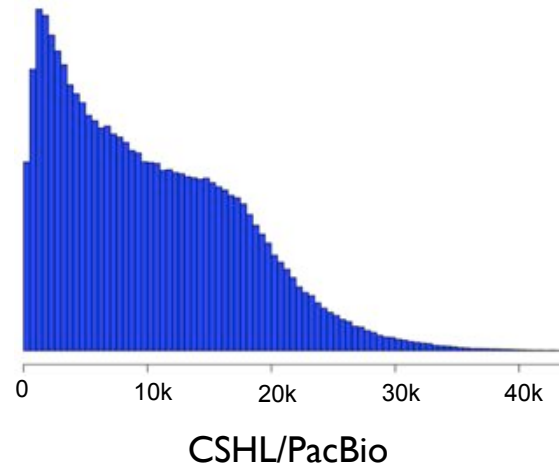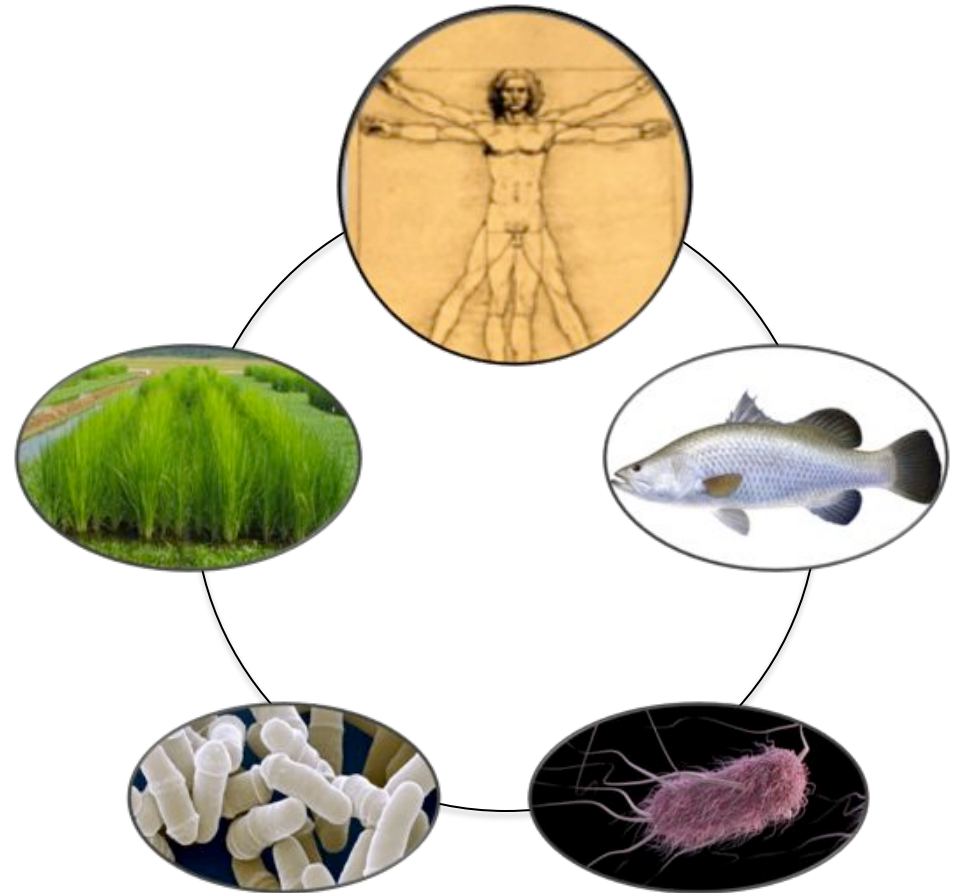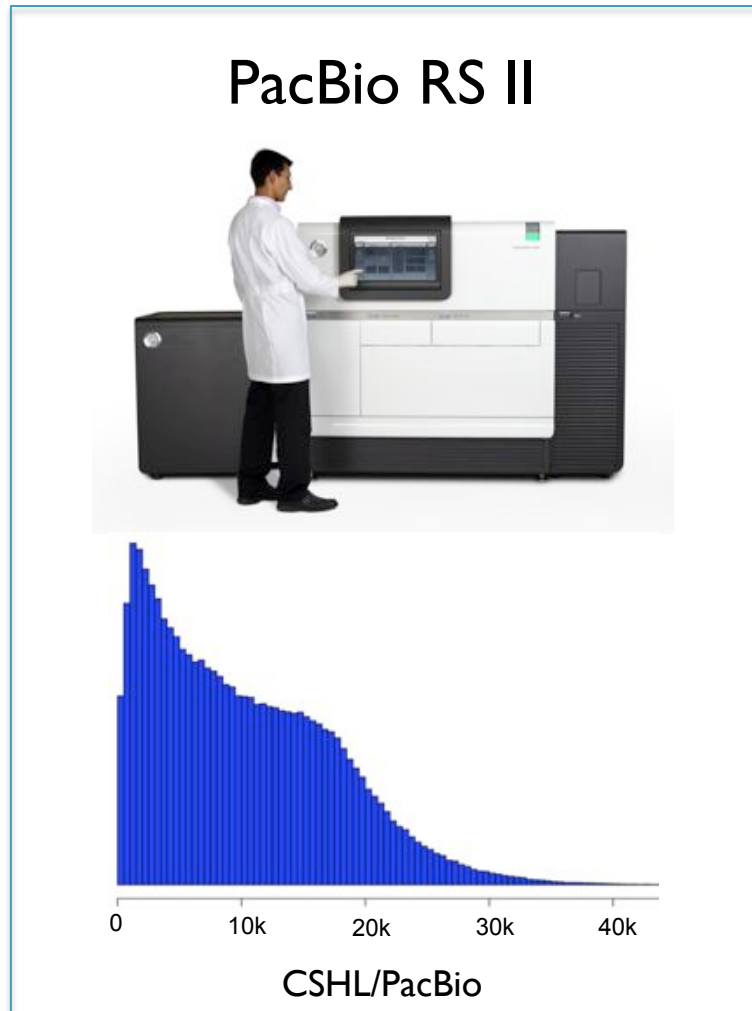
3rd Gen Long Read Sequencing
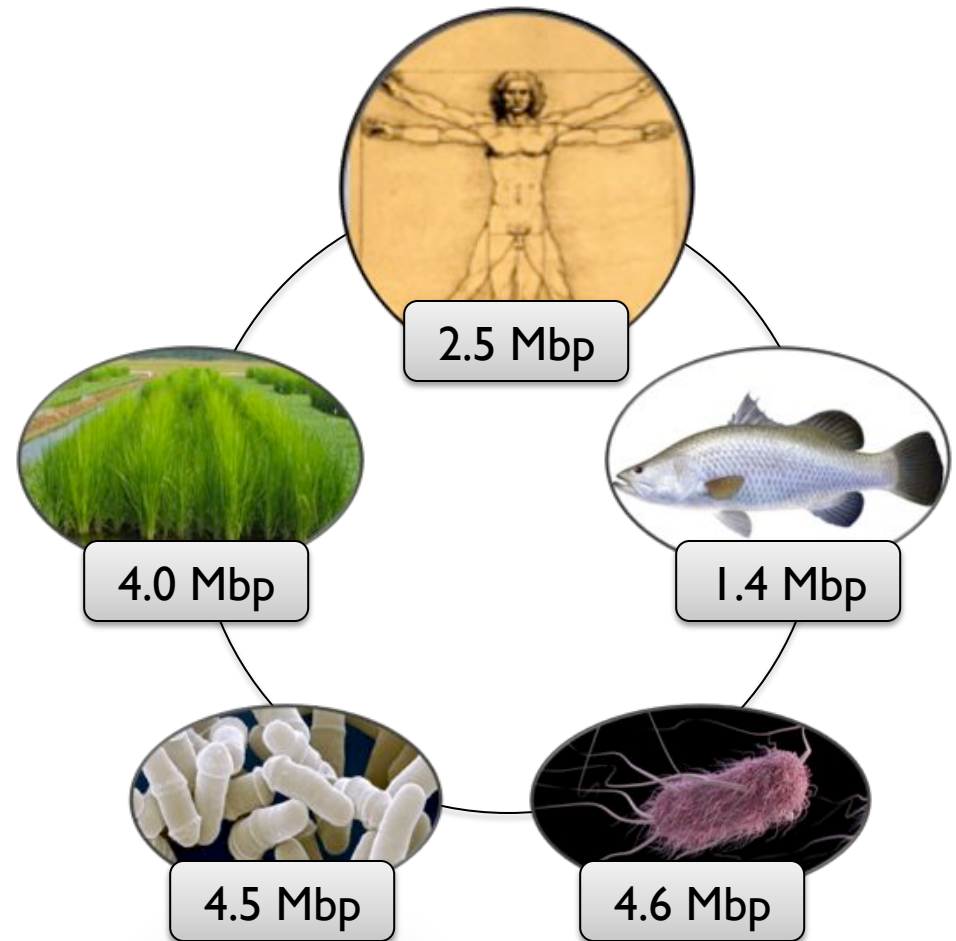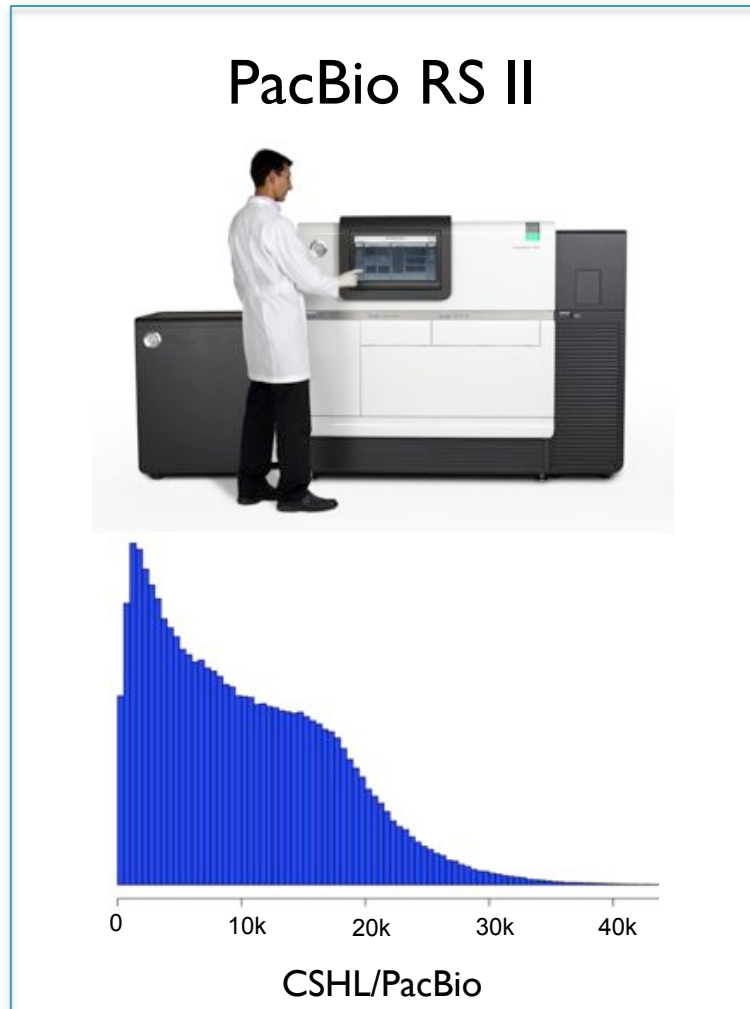
SK-BR-3

Most commonly used Her2-amplified breast cancer ce
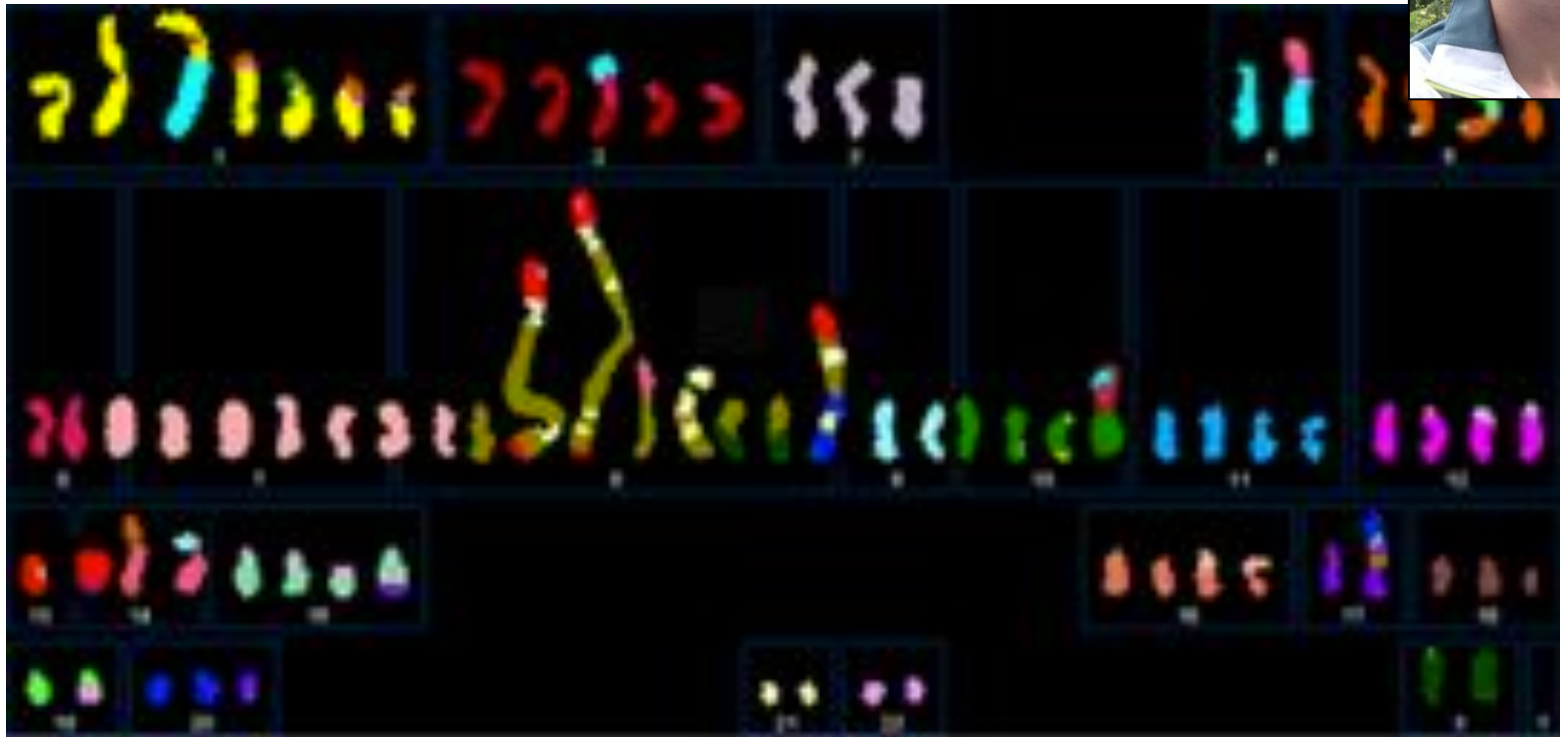
(Davidson et al, 2000)

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

read lengths

max: 71kb

Structural variant discovery with long reads

1. Alignment-based split read analysis: Efficient capture of most events
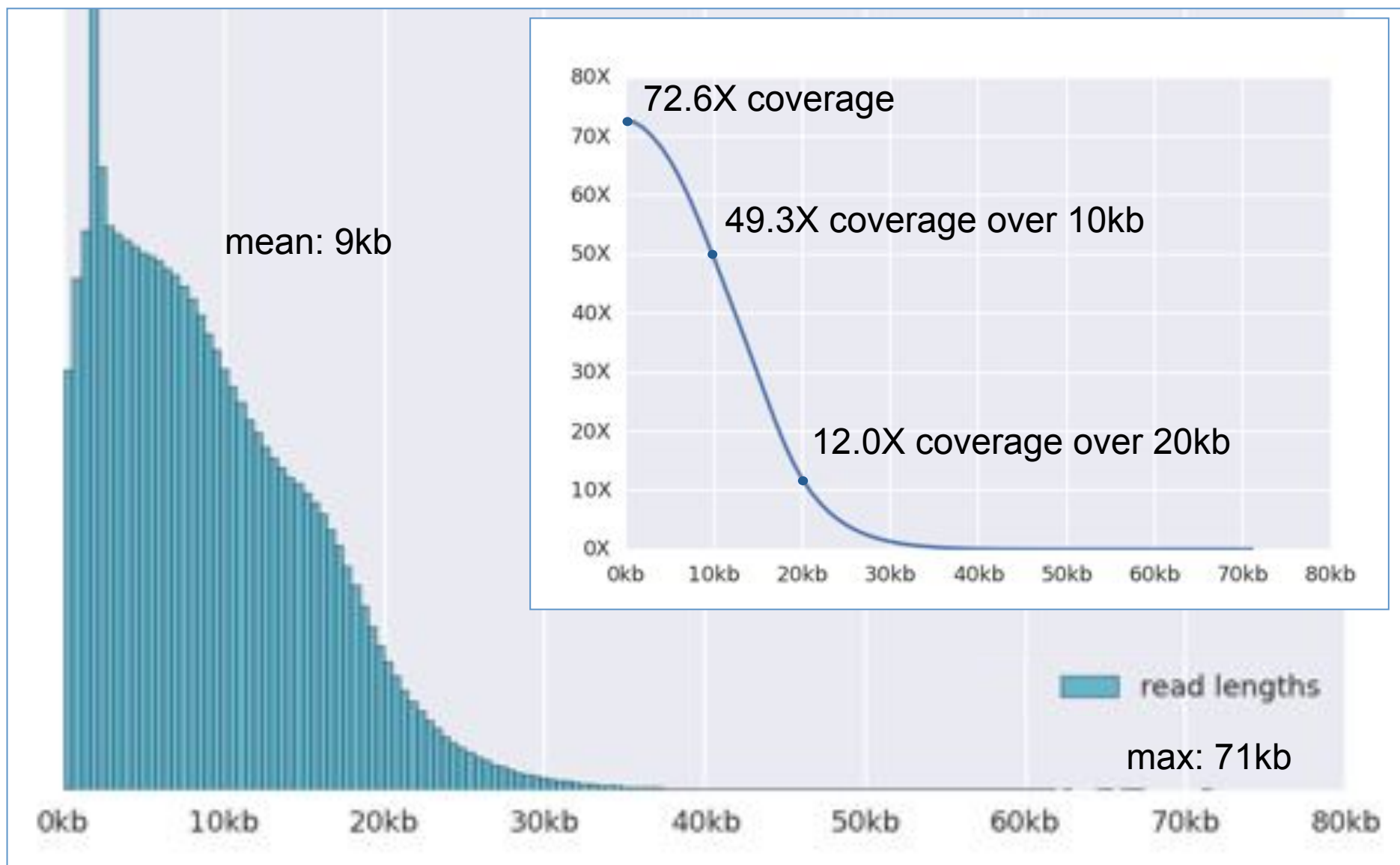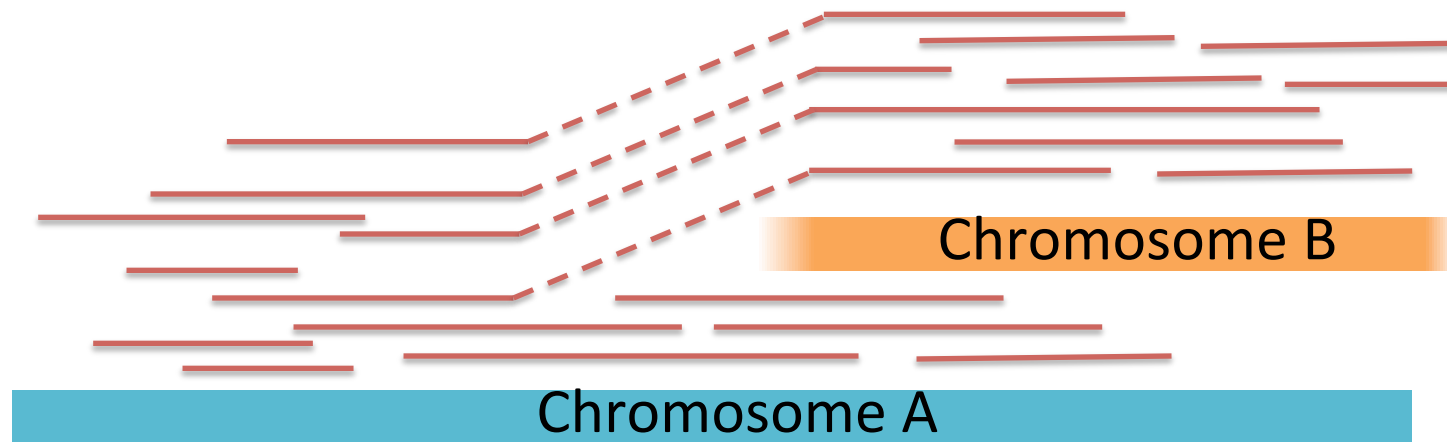   BWA-MEM + Lumpy

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

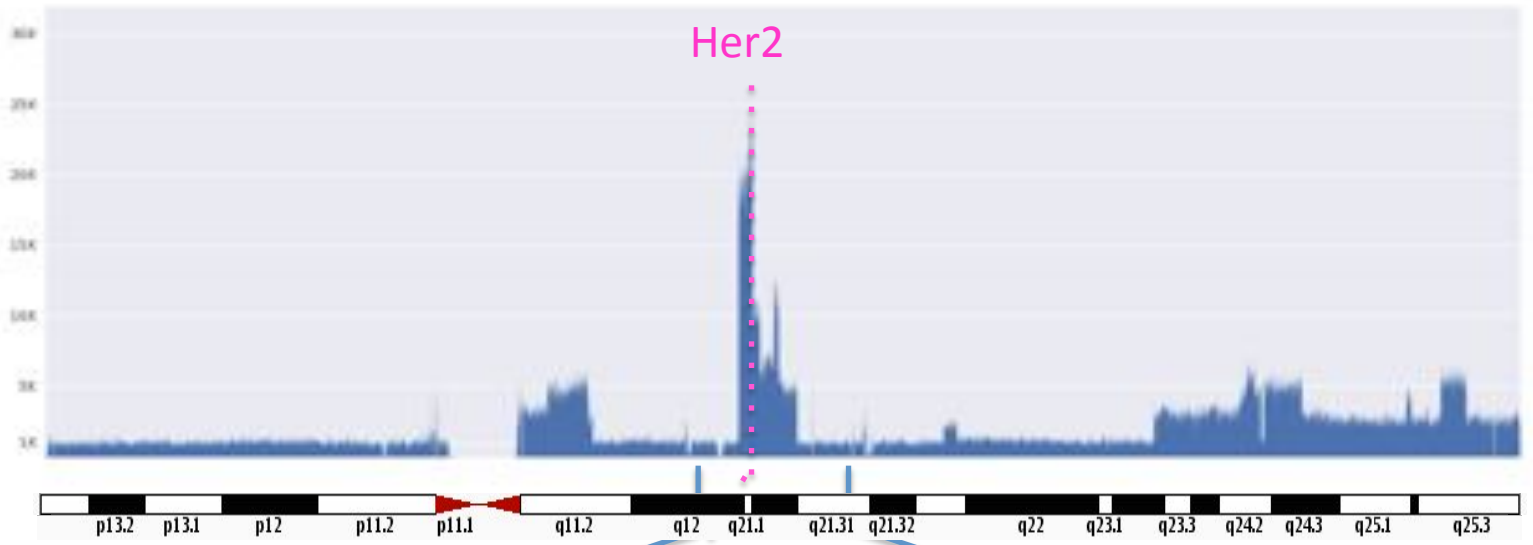   **Total Assembly: 2.64Gbp   Contig N50: 2.56 Mbp   Max Contig: 23.5Mbp**

PacBio
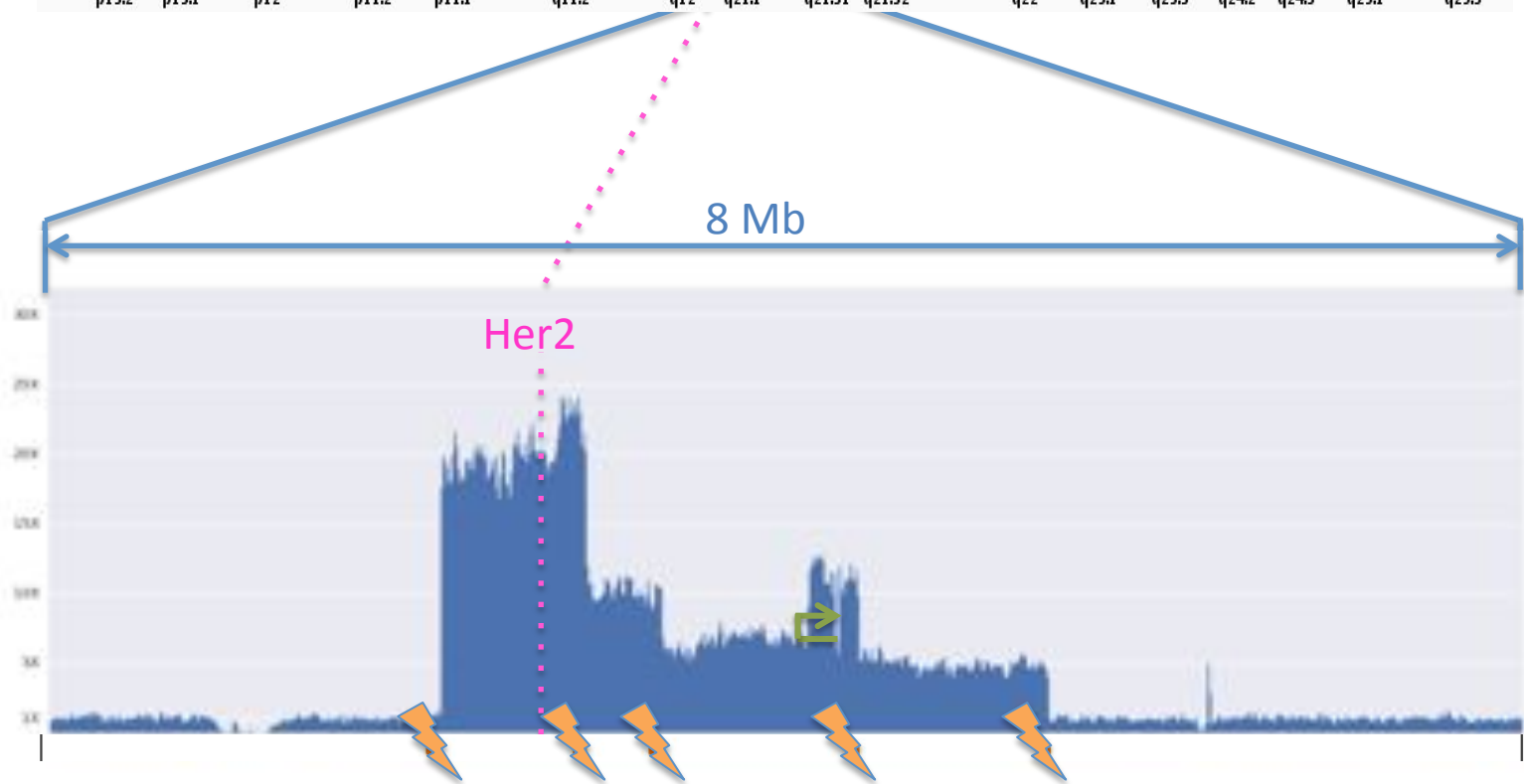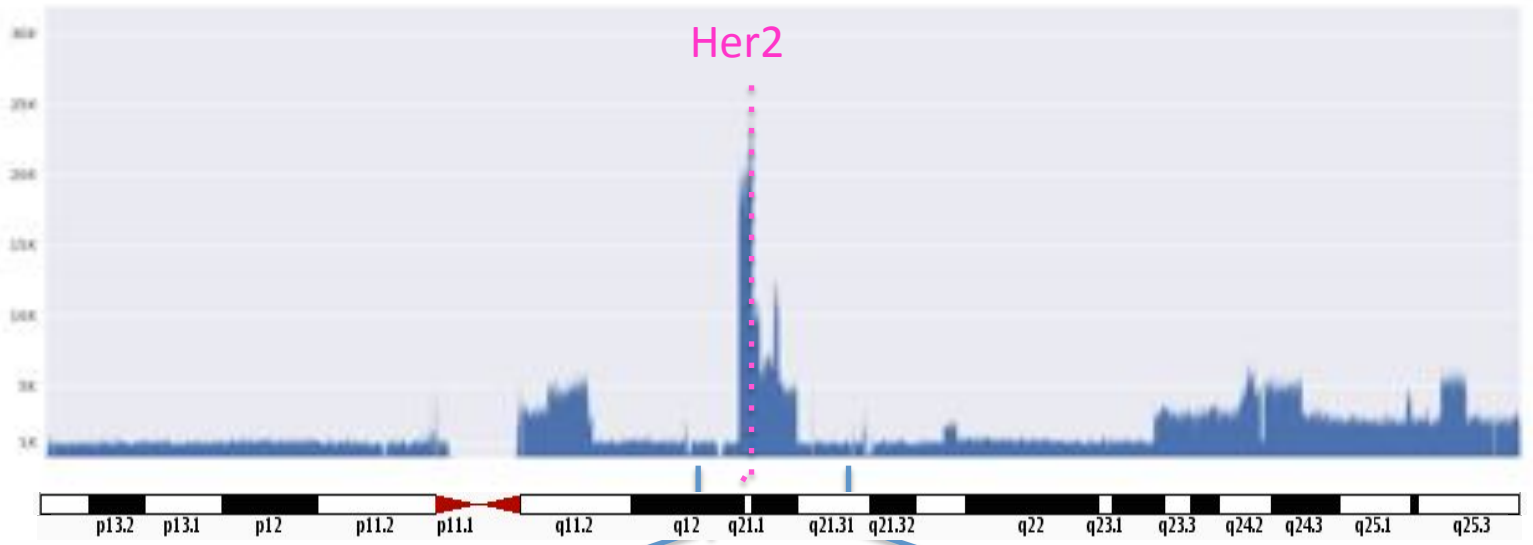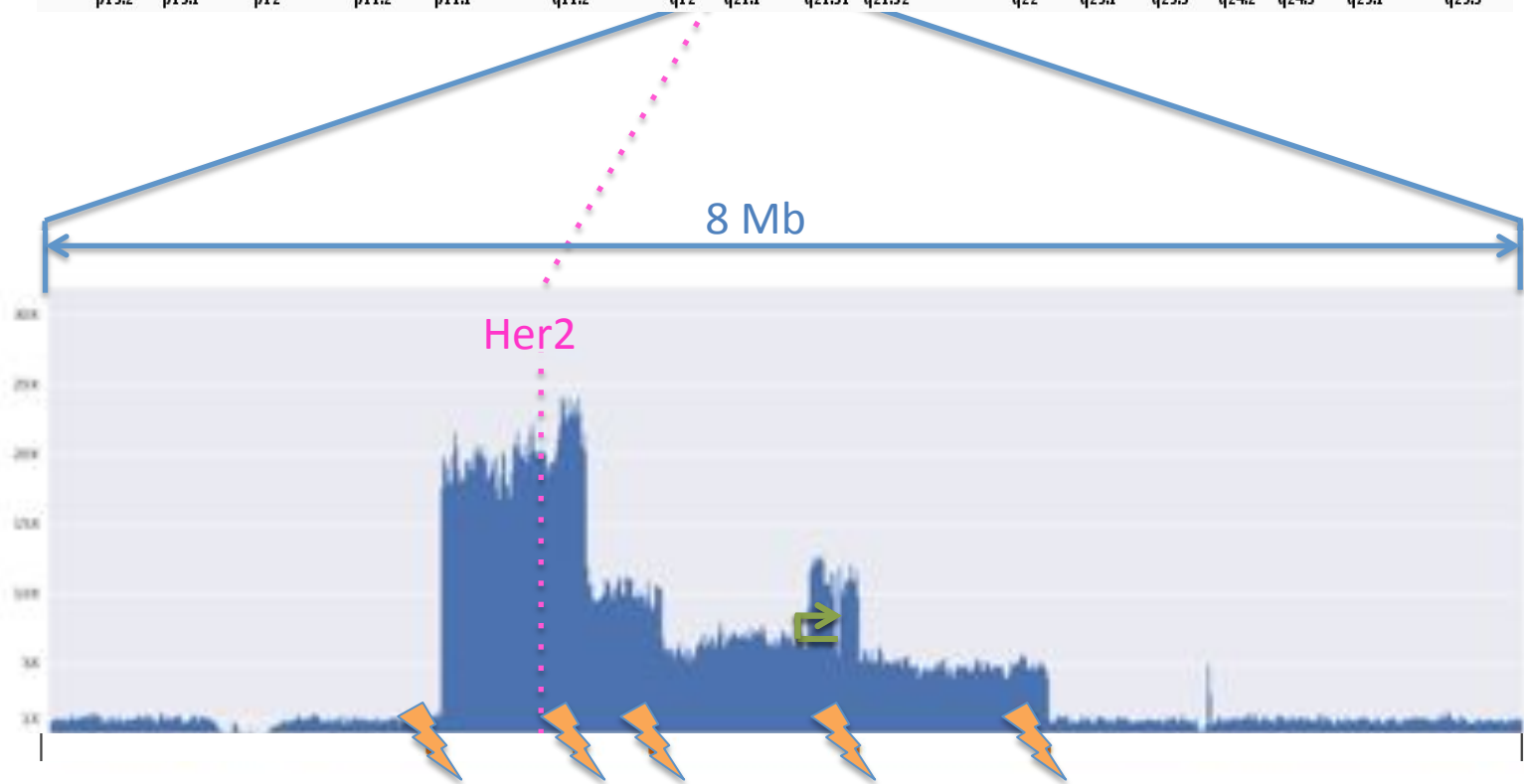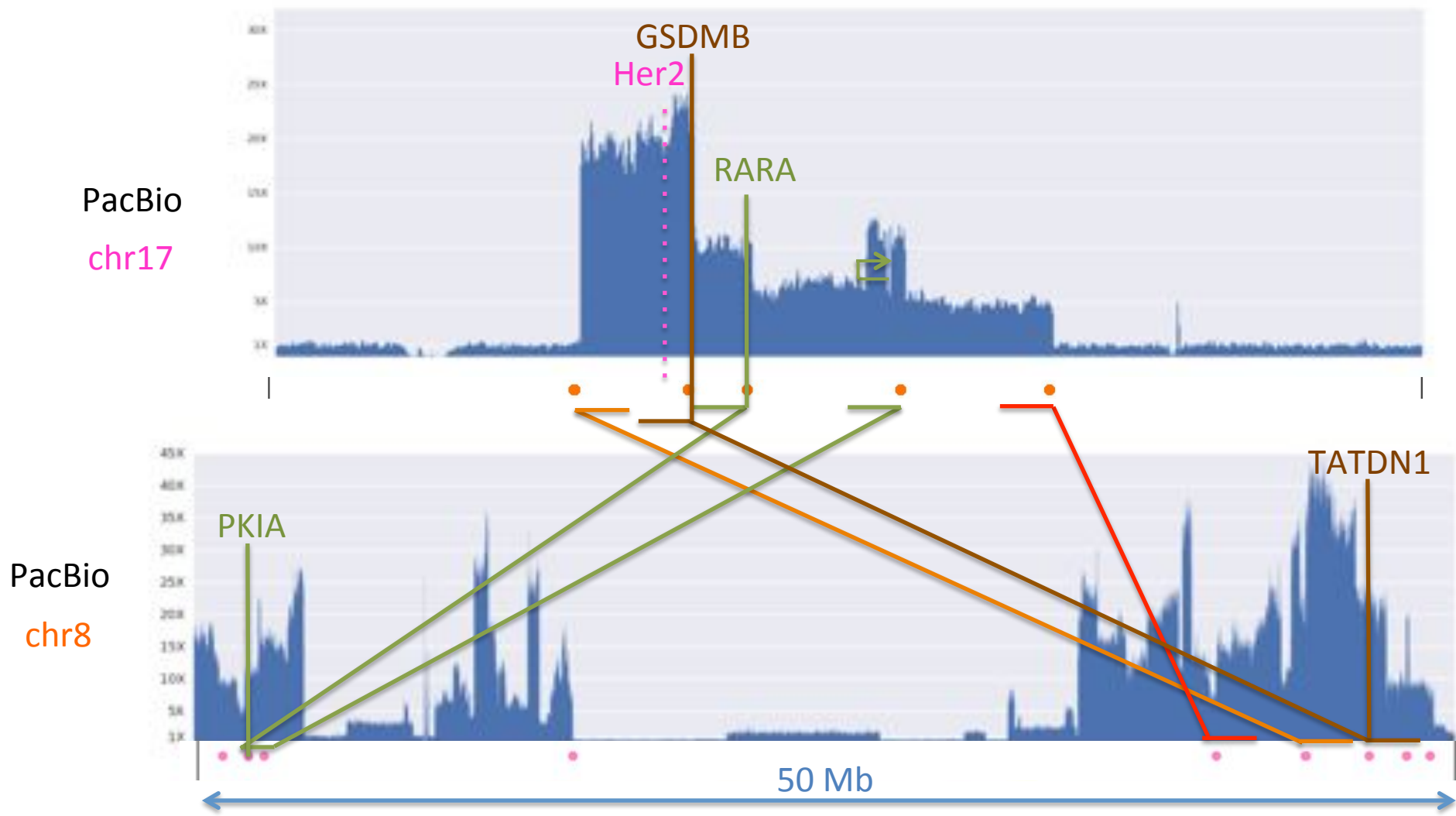
Her2

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

Her2

PacBio
chr17

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

Confirmed both known gene fusions in this region

PacBio
chr17

GSDMB
Her2
RARA

PacBio
chr8

PKIA

TATDN1

50 Mb

1.6 Mb

Confirmed both known gene fusions in this region

PacBio

chr17

Her2

RARA

chr8

PKIA

1.6 Mb
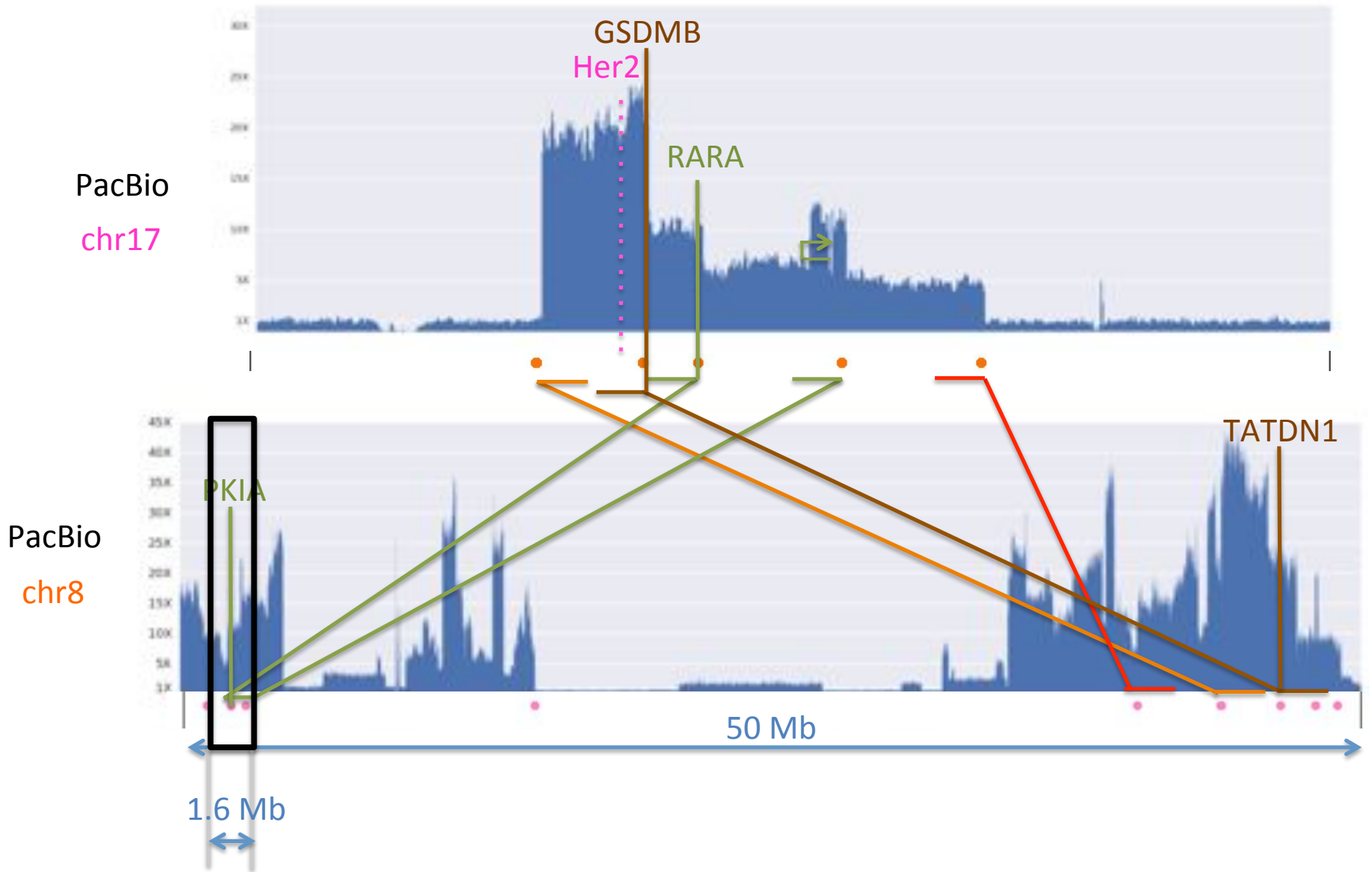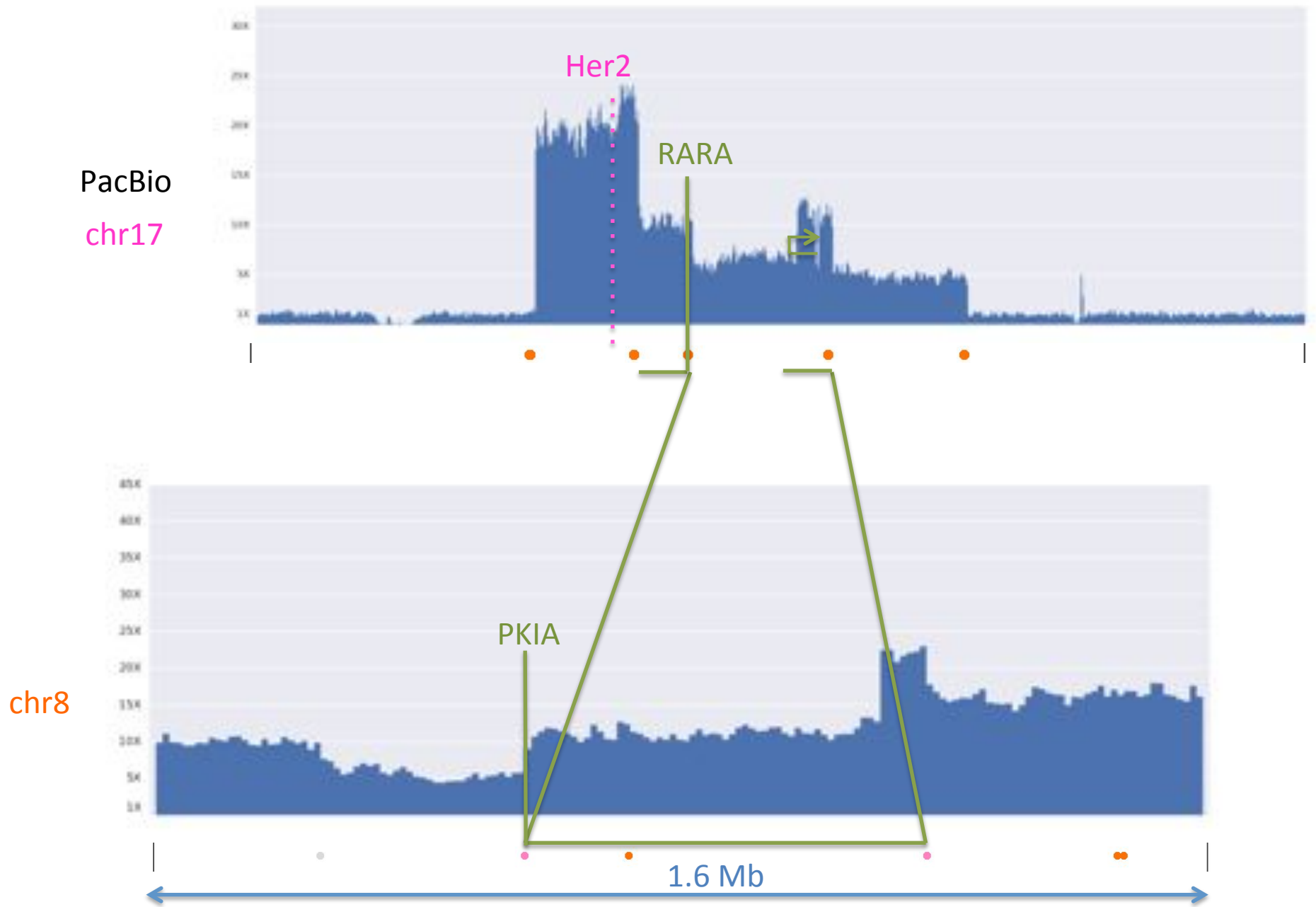
Joint coverage and breakpoint analysis to discover underlying events

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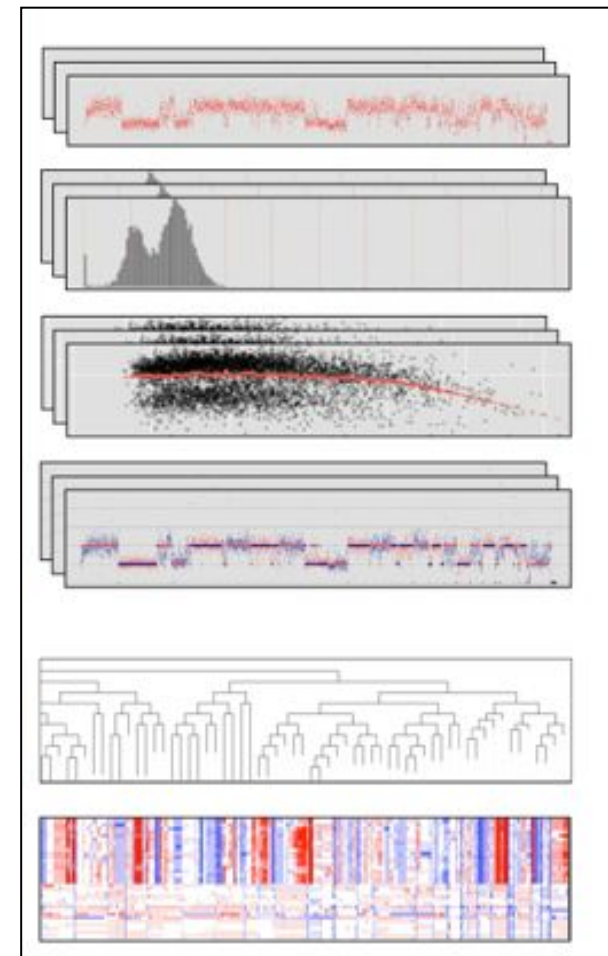
Interactive analysis and quality assessment of single-cell copy-number variations.
Garvin, T., Aboukhalil, *et al.* (2015) *Under review*

What should we expect from an assembly?

Summary & Recommendations

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

Acknowledgements

Genome Informatics
Janet Kelso, Daniel MacArthur, Michael Schatz
Oct 28 - 31, 2015

Thank you

http://schatzlab.cshl.edu

@mike_schatz