PacBio Long Read Sequencing and Structural Analysis of a Breast Cancer Cell Line

W. Richard McCombie

Disclosures

Orion Genomics – Founder and Shareholder
Cancer epigenetics and plant genomics

Previously Compensated Speaker for Illumina, Inc.

Previously Compensated Speaker for Pacific Biosciences, Inc.

Expressed genes, *Alu* repeats and polymorphisms in cosmids sequenced from chromosome 4p16.3

W.R. McCombie[1], A. Martin-Gallardo[1], J.D. Gocayne[1], M. FitzGerald[1], M. Dubnick[1], J.M. Kelley[1], L. Castilla[2], L.I. Liu[1], S. Wallace[1], S. Trapp[1], D. Tagle[2], W.L. Whaley[3], S. Cheng[3], J. Gusella[3], A.-M. Frischauf[4], A. Poustka[5], H. Lehrach[4], F. S. Collins[2], A. R. Kerlavage[1], C. Fields[1] & J.C. Venter[1]

The sequences of three cosmids (90 kilobases) from the Huntington's disease region in chromosome 4p16.3 have been determined. A 30,837 base overlap of DNA sequenced from two individuals was found to contain 72 DNA sequence polymorphisms, an average of 2.3 polymorphisms per kilobase (kb). The assembled 58 kb contig contains 62 *Alu* repeats, and eleven predicted exons representing at least three expressed genes that encode previously unidentified proteins. Each of these genes is associated with a CpG island. The structure of one of the new genes, *hda1-1*, has been determined by characterizing cDNAs from a placental library. This gene is expressed in a variety of tissues and may encode a novel housekeeping gene.

Evolution of genome assemblies

- Initial references – very high quality – extremely expensive
- Period of lower quality Sanger assemblies (~2001-2007)
- Next gen assemblies (short read) – 2007- now
- Third generation – long read assemblies -2013/2014 –now – what can we do currently?

Her2 amplified breast cancer

Breast cancer

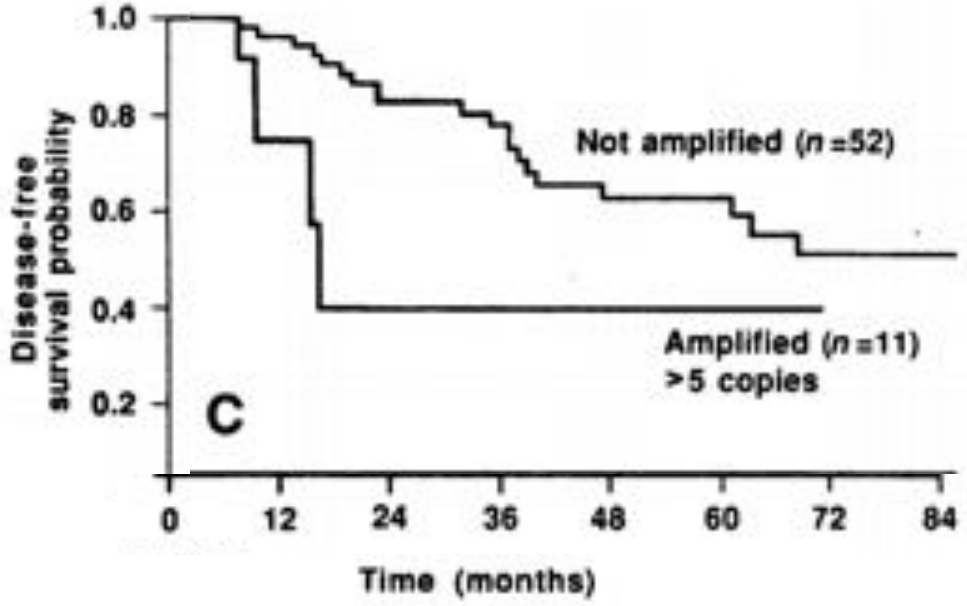
- About 12% of women will develop breast cancer during their lifetimes
- ~230,000 new cases every year (US)
- ~40,000 deaths every year (US)

Statistics from American Cancer Society and Mayo Clinic.

Recurrence and metastasis from Gonzalez-Angulo, et al, 2009.

Her2 amplified breast cancer

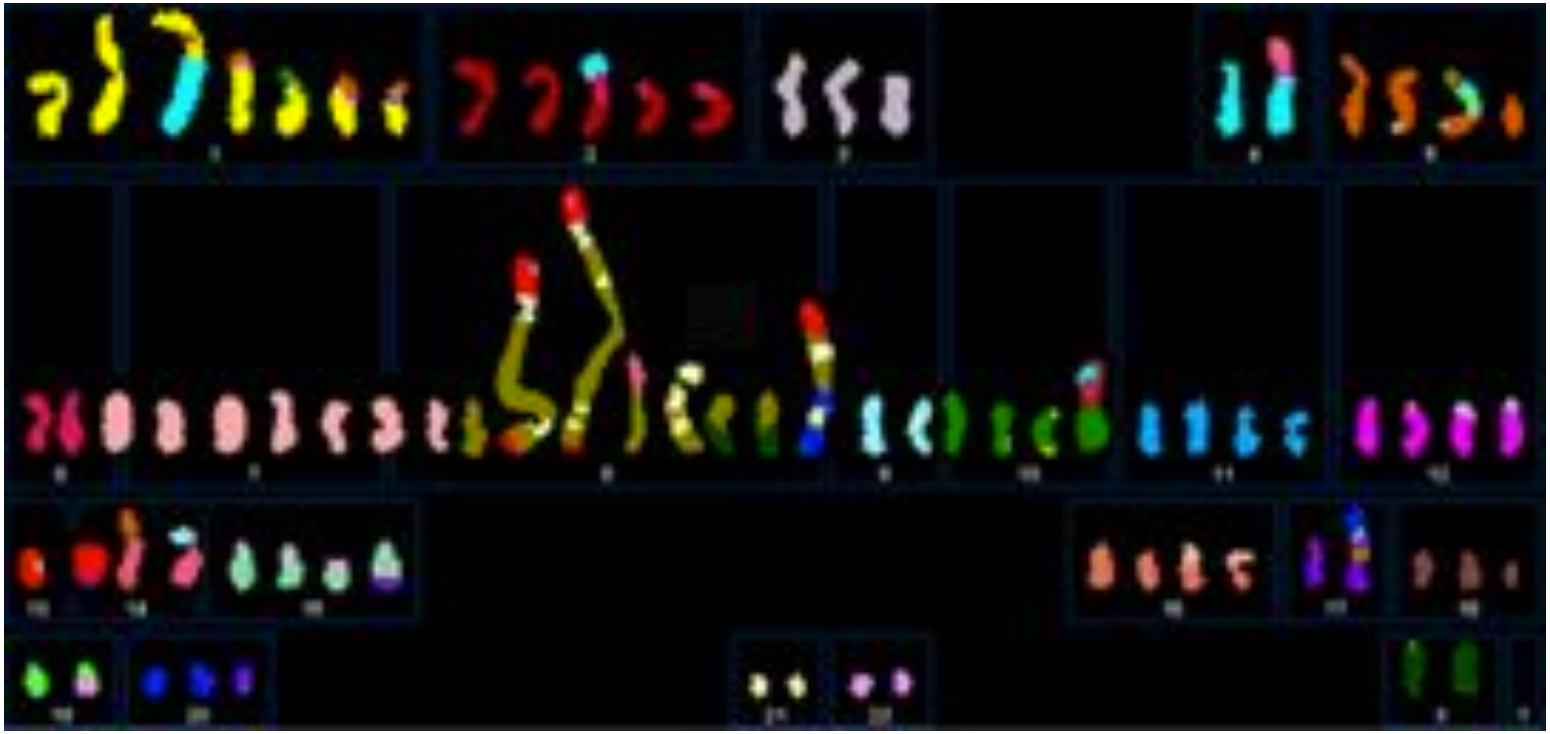
- 20% of breast cancers
- 2-3X recurrence risk
- 5X metastasis risk



(Adapted from Slamon et al, 1987)

SK-BR-3

Most commonly used Her2-amplified breast cancer cell line



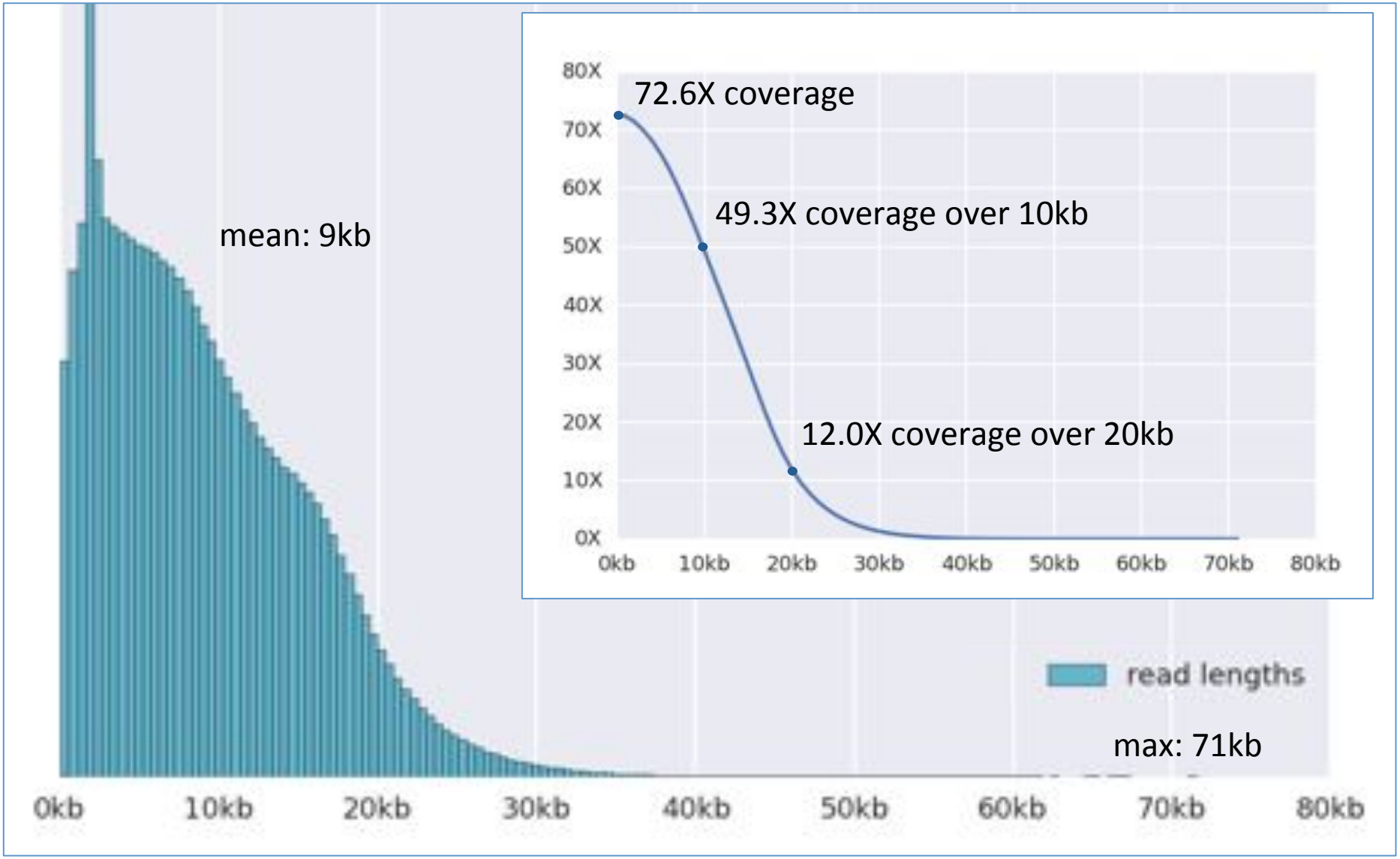
Often used for pre-clinical research on Her2-targeting therapeutics such as Herceptin (Trastuzumab) and resistance to these therapies.

(Davidson et al, 2000)

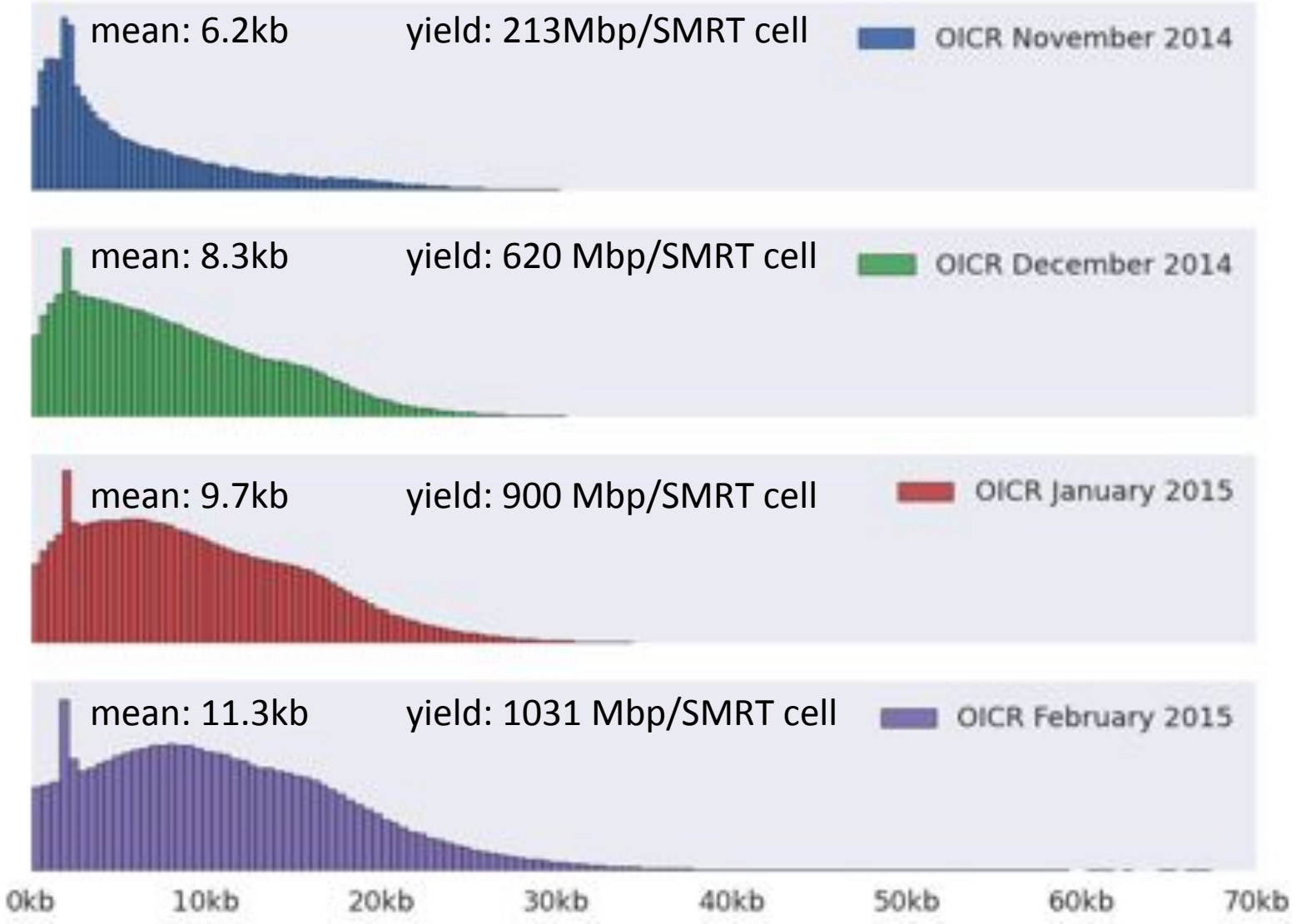
High quality sequencing with Pac Bio

- Just long read data
- High coverage
- Map back to human reference
- *De novo* assembly
- Characterize the view of the SKBR3 genome presented by different assemblies as well as determine “truth”
- Ongoing project

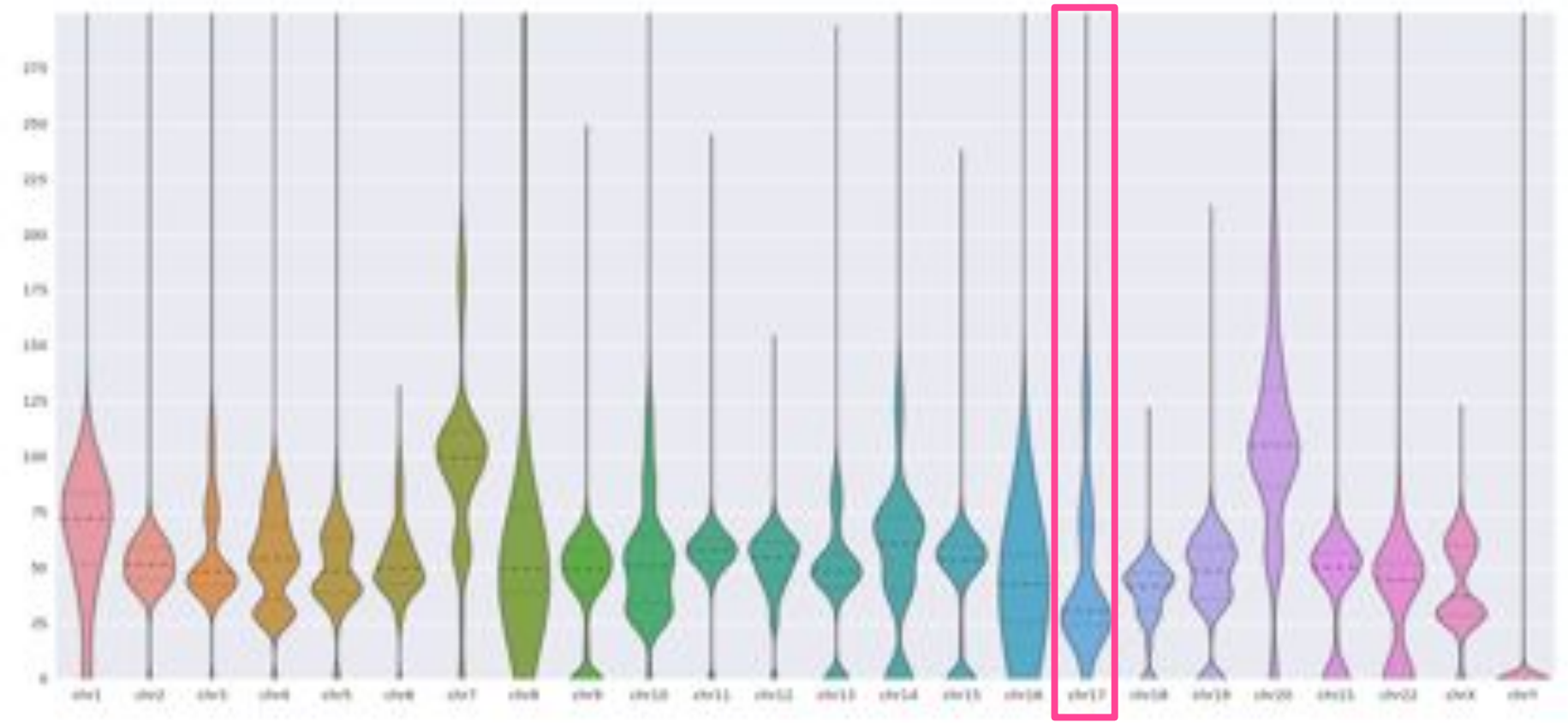
PacBio read length distribution



Improving SMRTcell Performance

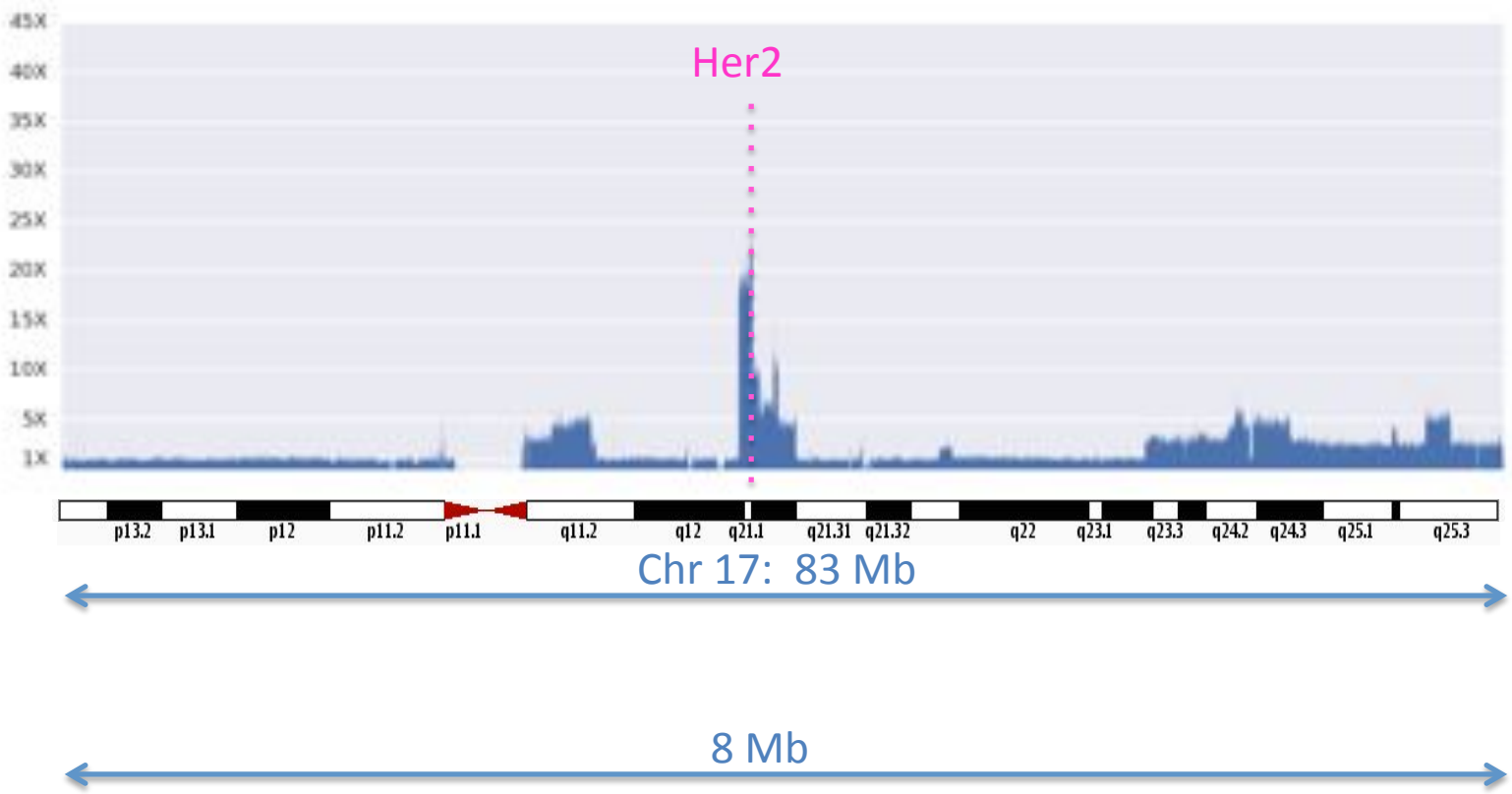


Genome-wide alignment coverage

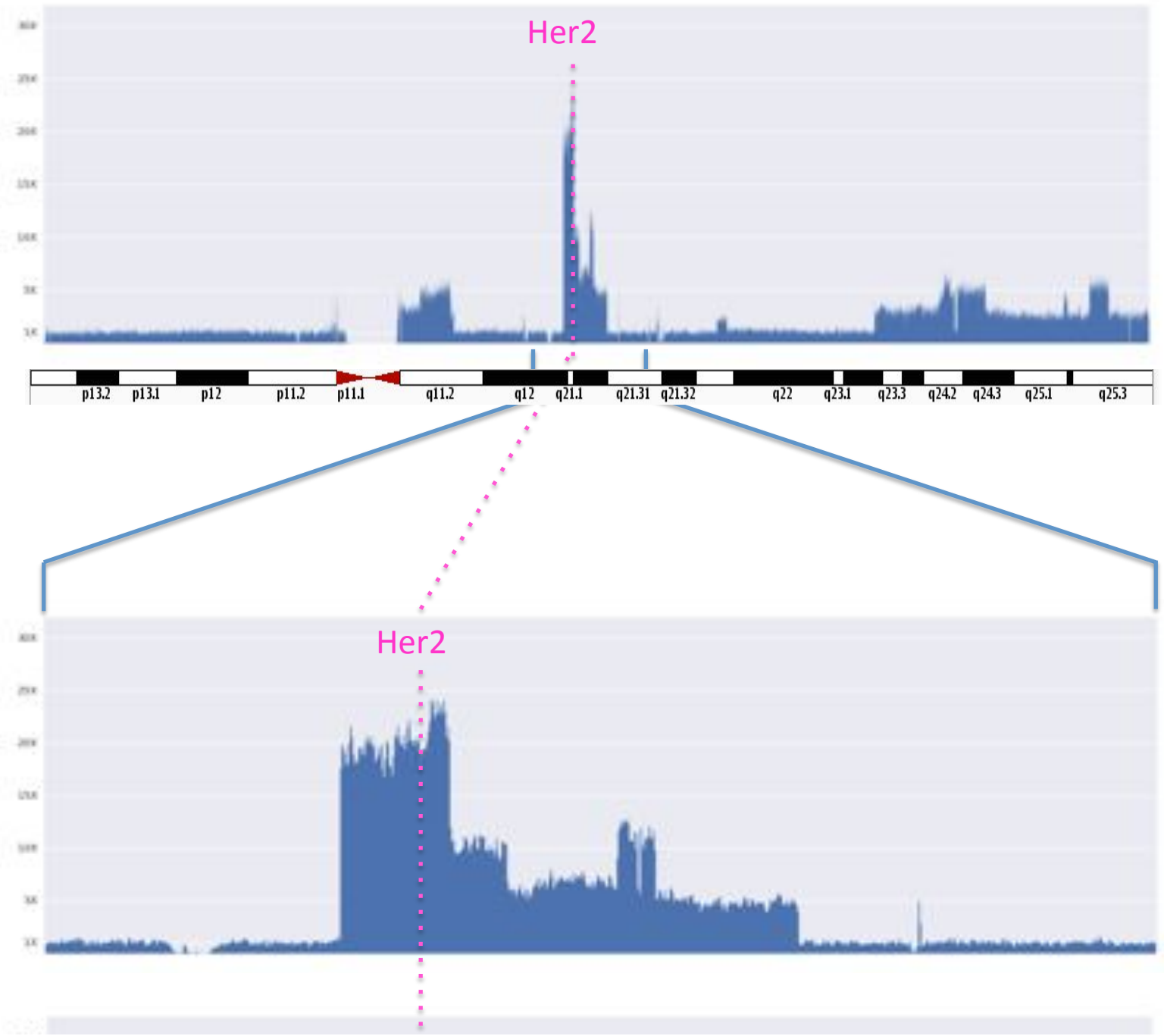


Genome-wide coverage averages around 54X
Coverage per chromosome varies greatly as expected from previous karyotyping results

PacBio


45X
40X
35X
30X
25X
20X
15X
10X
5X
1X
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
8 Mb

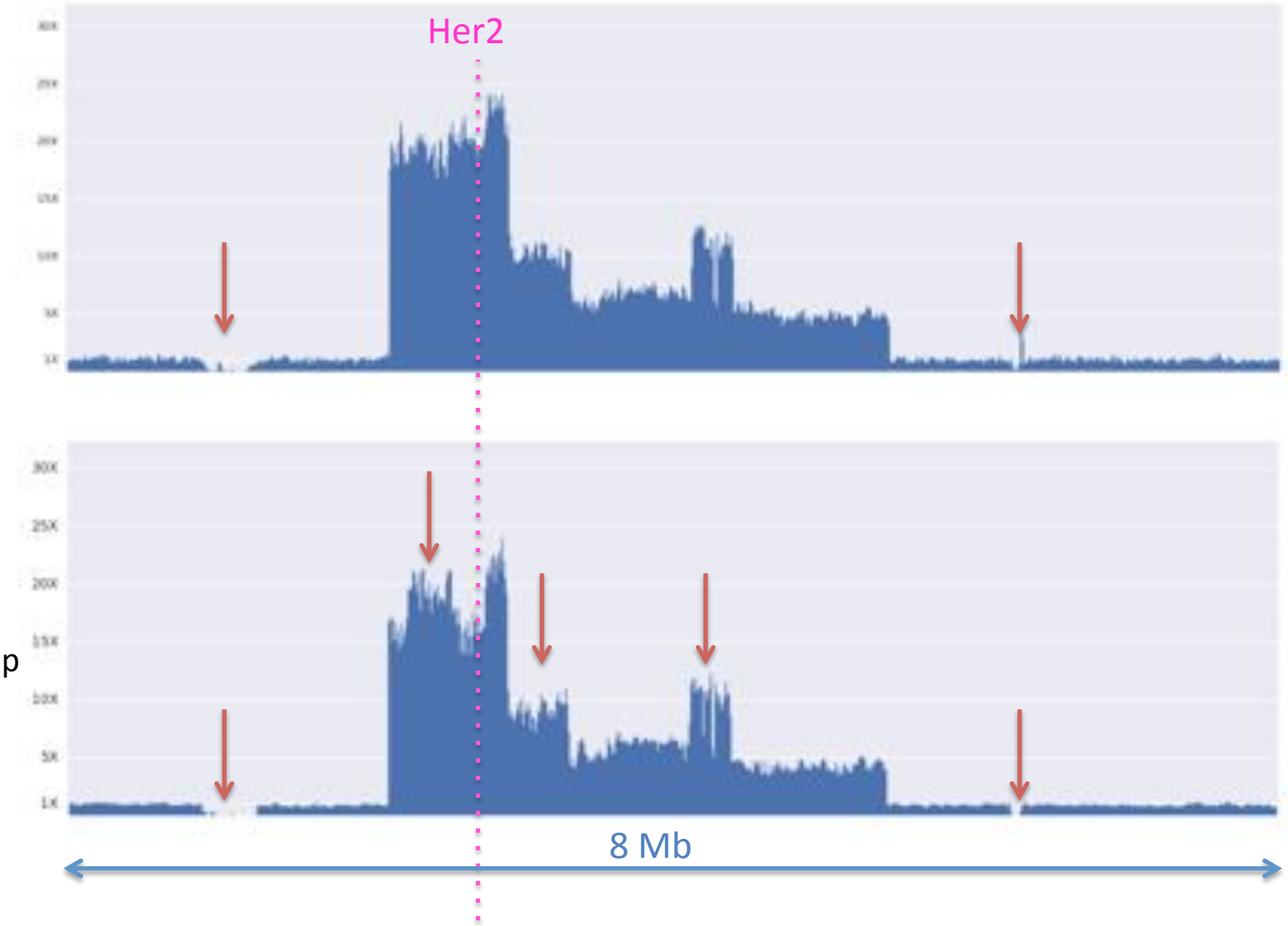
PacBio



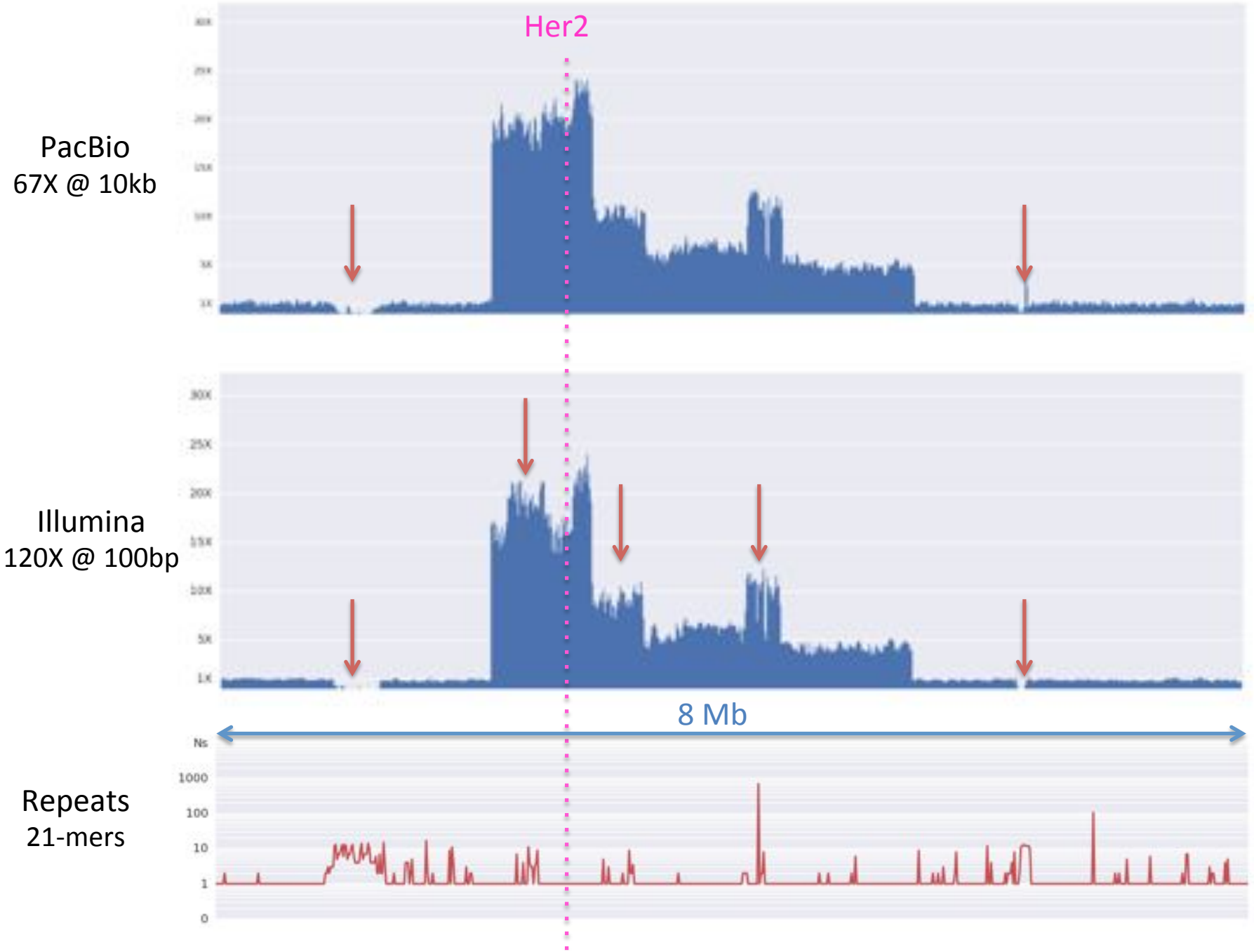
PacBio

PacBio
67X @ 10kb

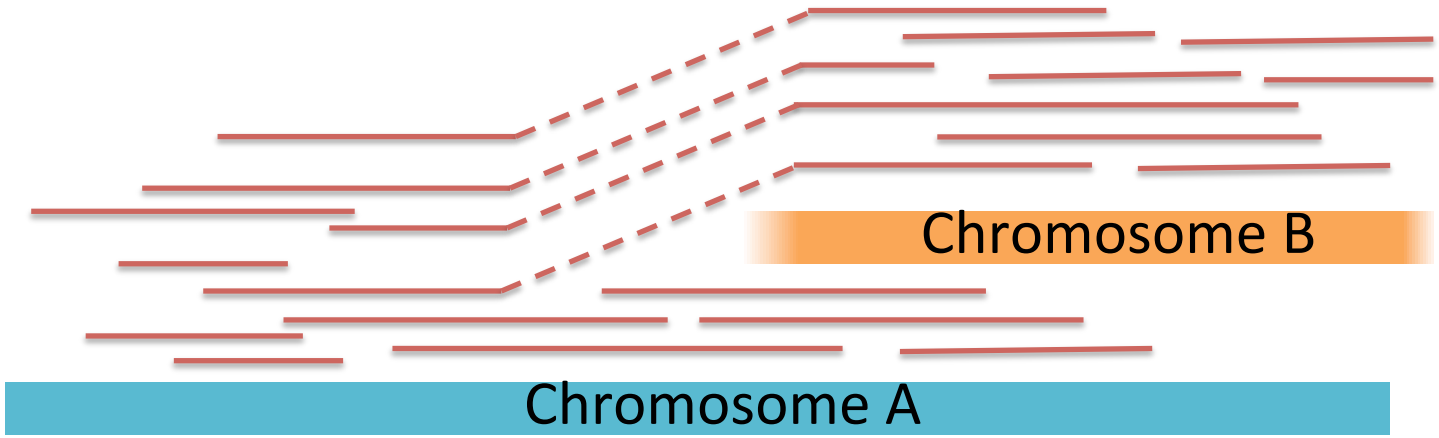
Illumina
120X @ 100bp



PacBio and Illumina coverage values are highly correlated
but Illumina shows greater variance because of poorly mapping reads


Her2
PacBio
67X @ 10kb
Illumina
120X @ 100bp
8 Mb
Repeats
21-mers
Ns
1000
100
10
1
0

Structural variant discovery with long reads



1. Alignment-based split read analysis: Efficient capture of most events

BWA-MEM + Lumpy

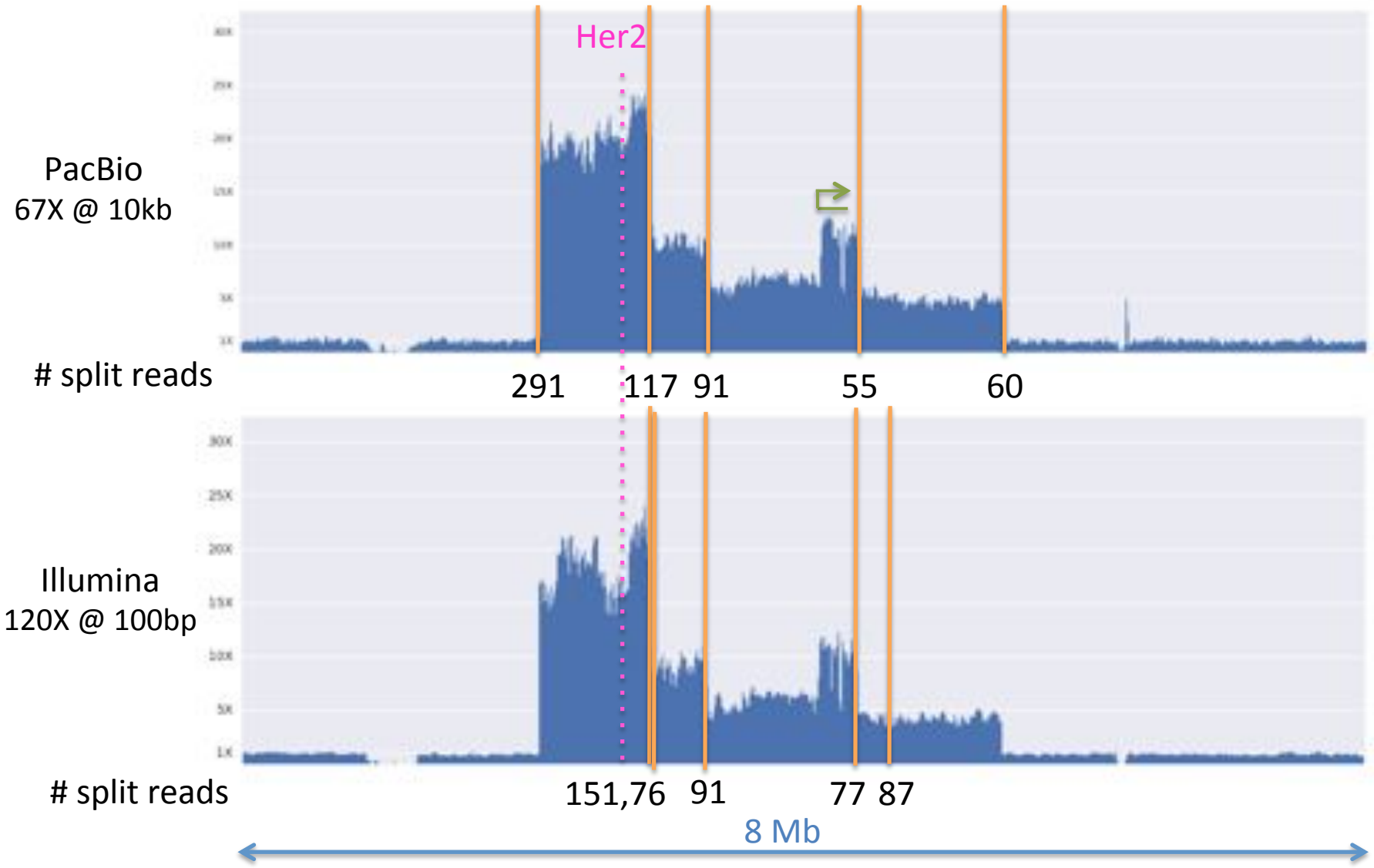
2. Local assembly of regions of interest: In-depth analysis with *base-pair precision*

Localized HGAP + Celera Assembler + MUMmer

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

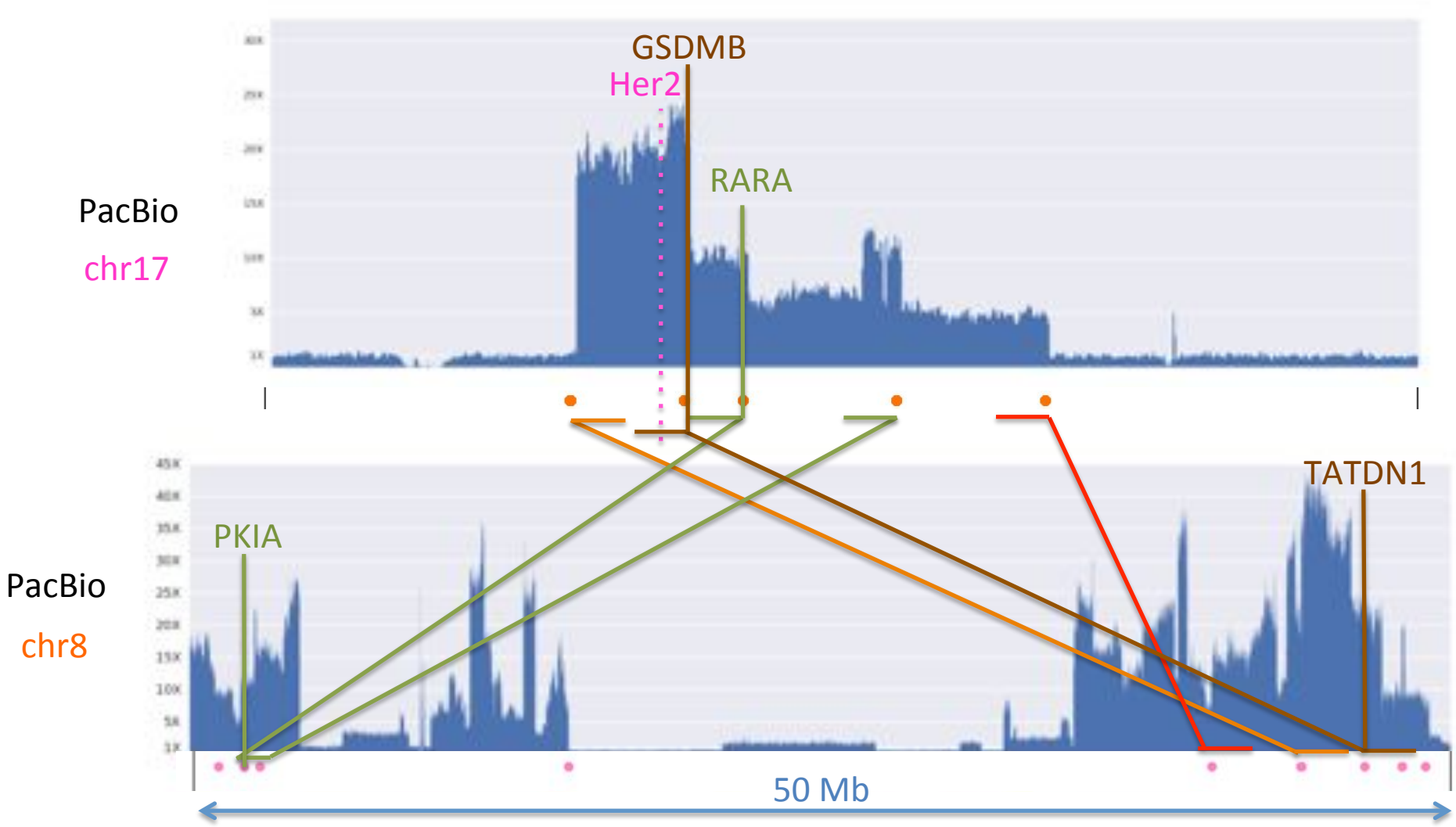
DNAnexus-enabled version of Falcon

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

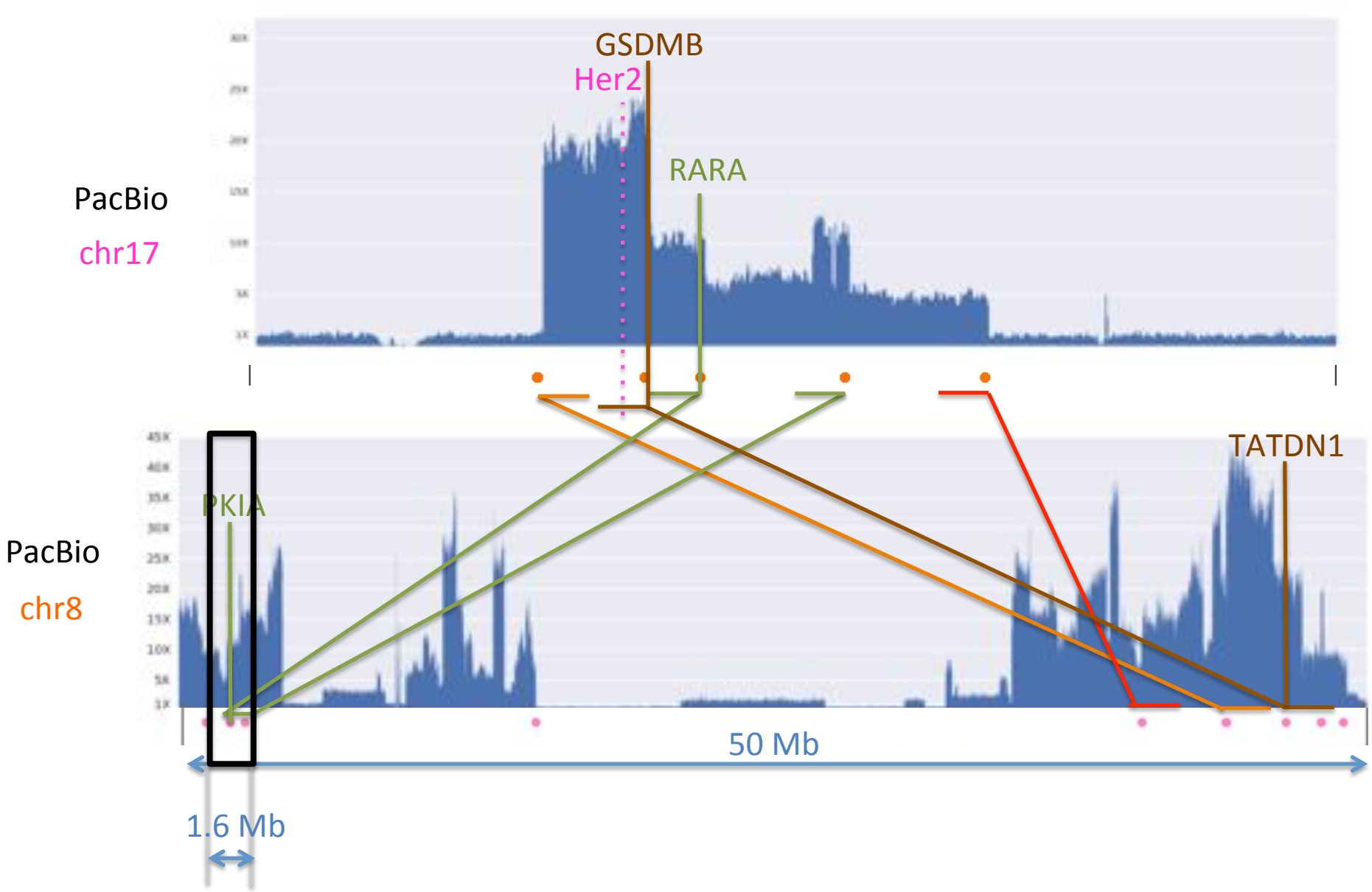


Green arrow indicates an inverted duplication.

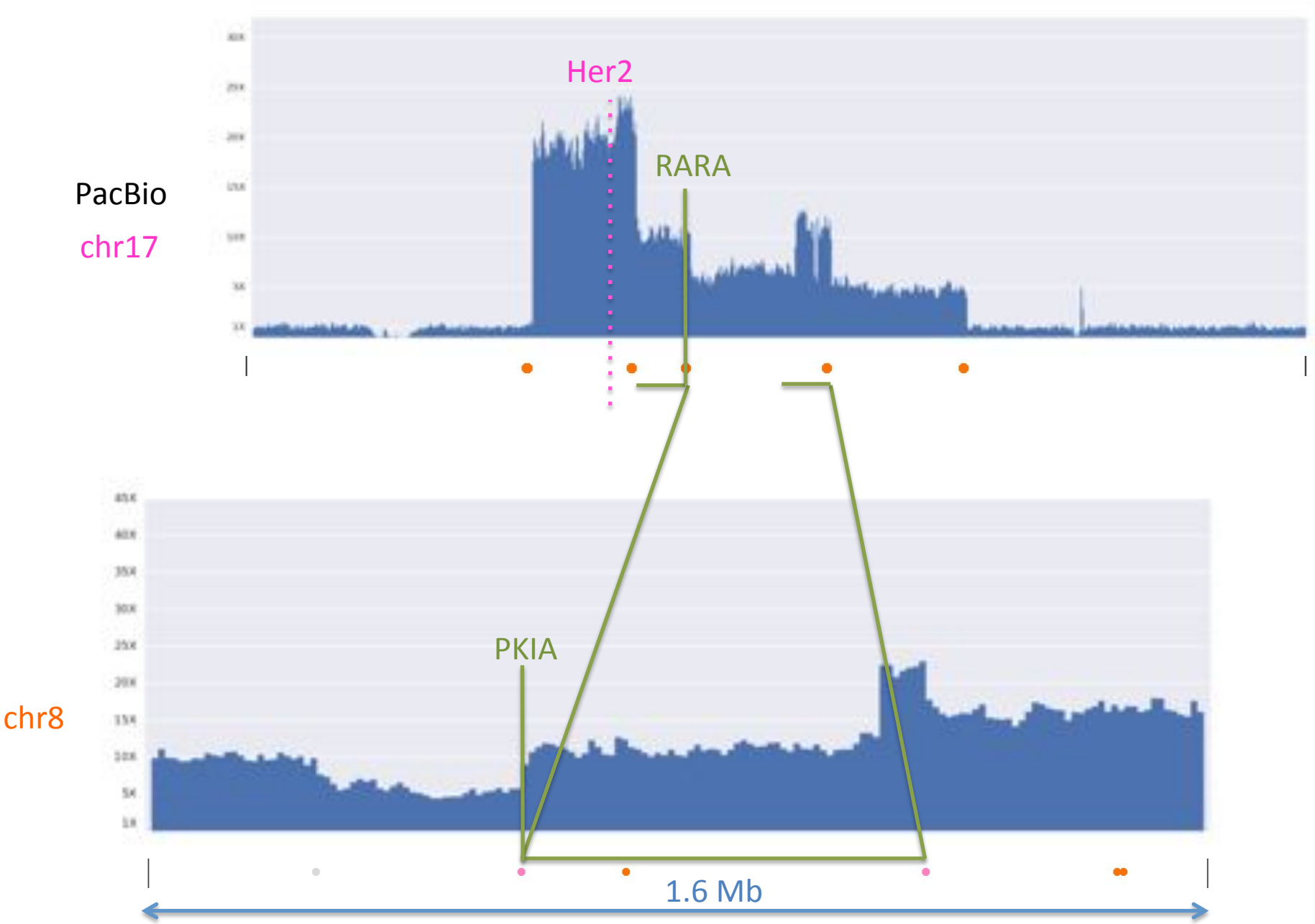
False positive and missing Illumina calls due to mis-mapped reads (especially low complexity).



Confirmed both known gene fusions in this region



Confirmed both known gene fusions in this region

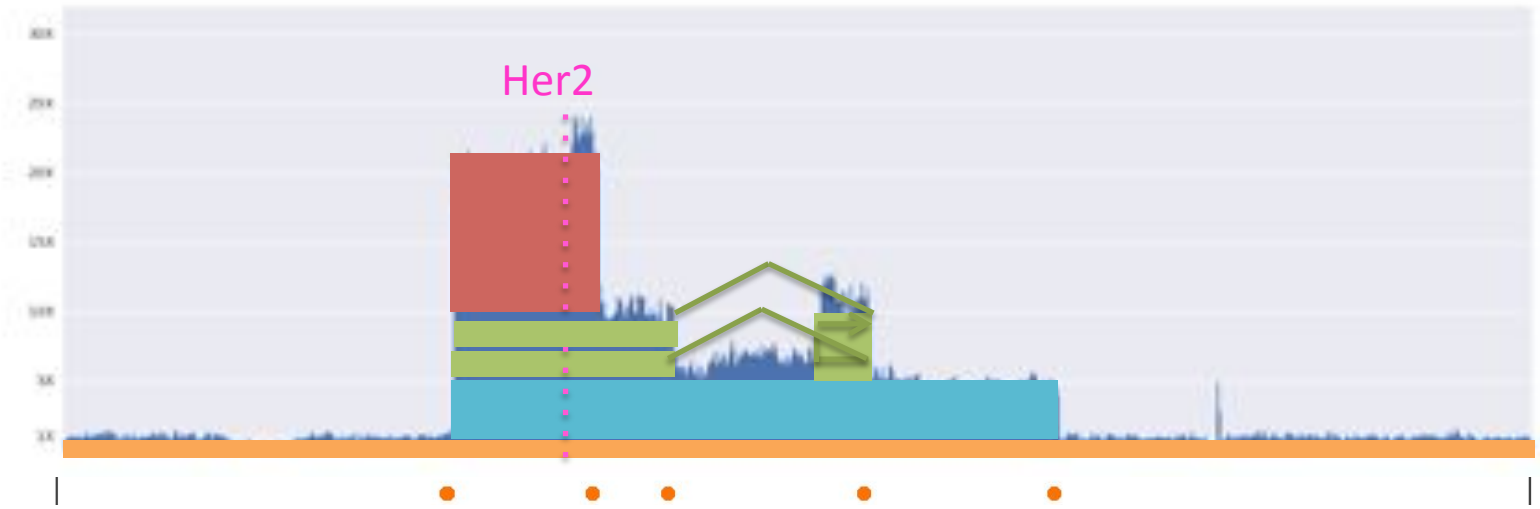


Joint coverage and breakpoint analysis to discover underlying events

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

SKBR3 Oncogene Analysis

Known missense mutation in p53: **R175H**

Oncogene amplifications	
ErbB2 (Her2/neu)	≈20X
MYC	≈27X
MET	≈8X

Genetic Lesion History Analysis Underway

Known Gene fusions		Confirmed by PacBio reads?
TATDN1	GSDMB	**Yes**
RARA	PKIA	**Yes**
ANKHD1	PCDH1	**Yes**
CCDC85C	SETD3	**Yes**
SUMF1	LRRFIP2	**Yes**
WDR67 (TBC1D31)	ZNF704	**Yes**
DHX35	ITCH	**Yes**
NFS1	PREX1	**Yes *read-through transcription**
CYTH1	EIF3H	**Yes *nested inside 2 translocations**

Her2+ Breast Cancer Reference Genome

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

Releasing all data pre-publication to accelerate breast cancer research

Available *today* under the Toronto Agreement:

- Fastq & BAM files of aligned reads
- Interactive Coverage Analysis with BAM.IOBIO

Available soon

- Whole genome assembly and methylation analysis
- Comparison to single cell analysis of >100 individual cells

Acknowledgments

Cold Spring Harbor Laboratory

Maria Nattestad
Sara Goodwin
Timour Baslan
James Gurtowski
Melissa Kramer
Panchajanya Deshpande
Senem Mavruk Eskipehlivan
Eric Antoniou
James Hicks
Michael Schatz

Karen Ng
Timothy Beck
Yogi Sundaravadanam
John McPherson

LUMPY technical support: Ryan Layer and Aaron Quinlan

Thank you!

May the force be with you!

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