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.



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?

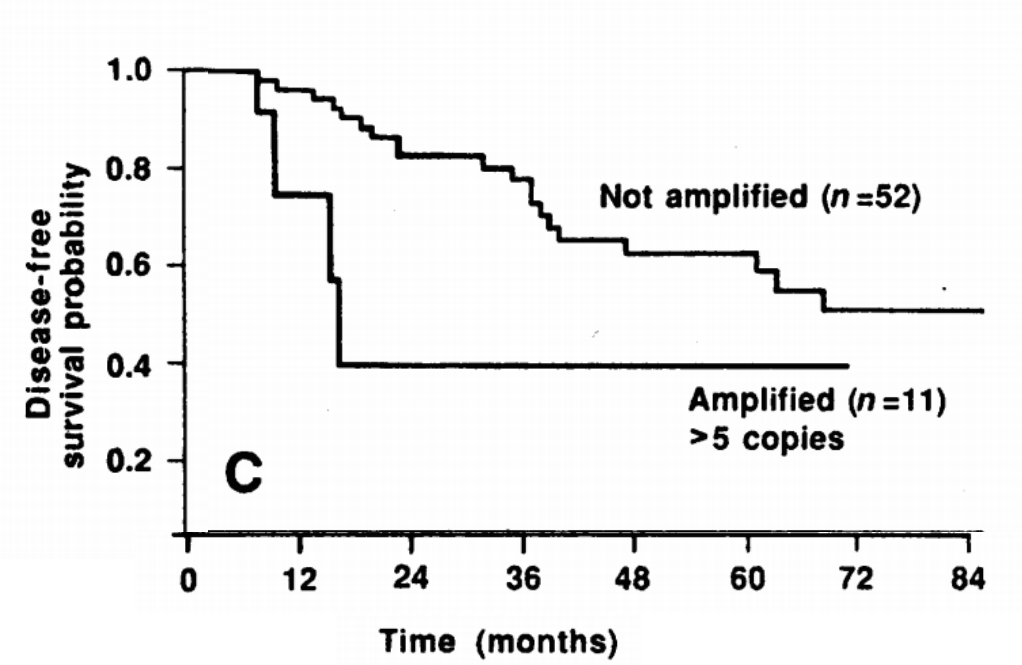
Her2 amplified breast cancer

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)

Her2 amplified breast cancer

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



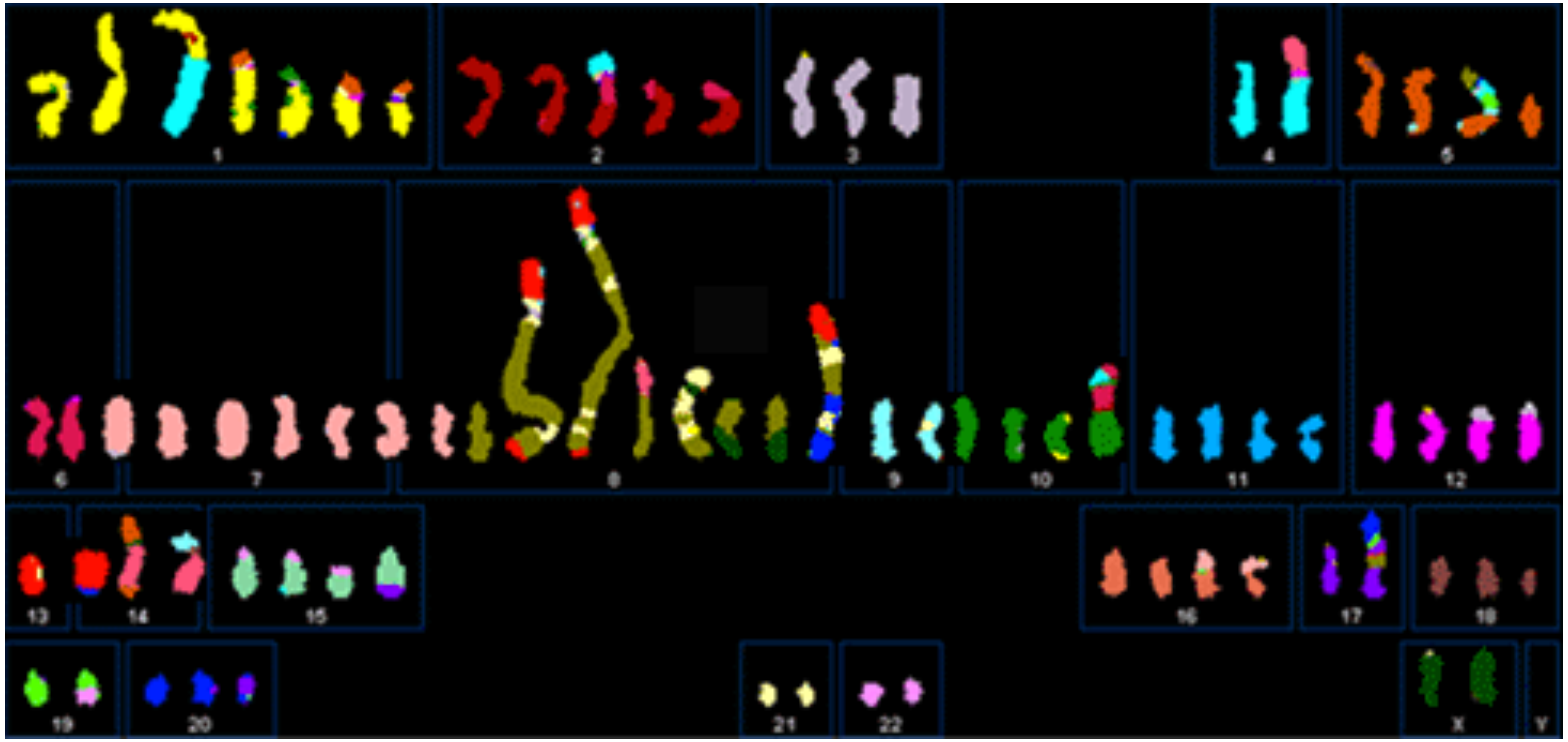
(Adapted from Slamon et al, 1987)

Statistics from American Cancer Society and Mayo Clinic.

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

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

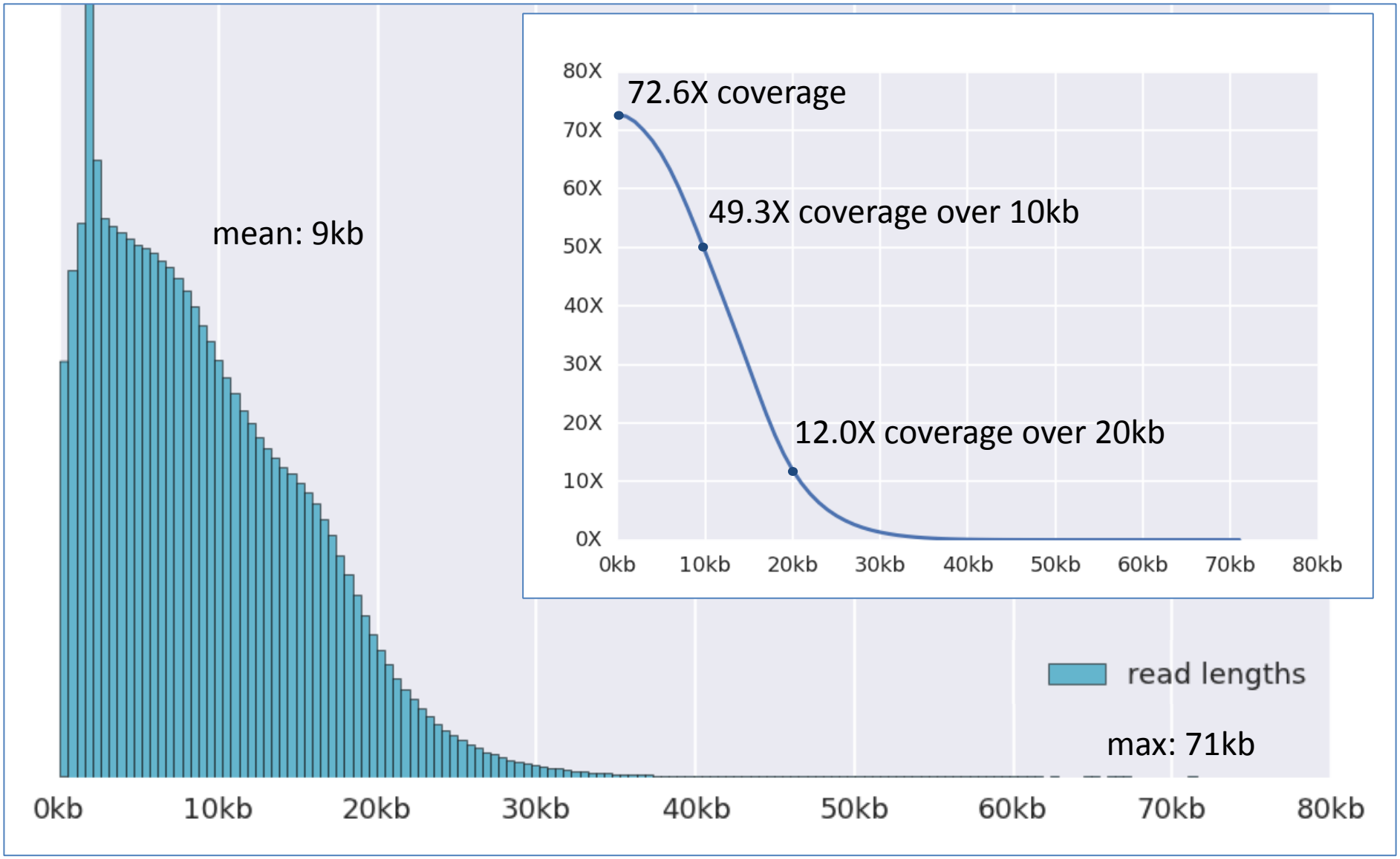
- Just long read data
- High coverage
- Map back to human reference
- *De novo* assembly
- Characterize the view of the SKBR3 genome presented by different assemblies as well as determine “truth”
- Ongoing collaboration with OICR, Hicks & Schatz labs at CSHL

High quality sequencing with Pac Bio

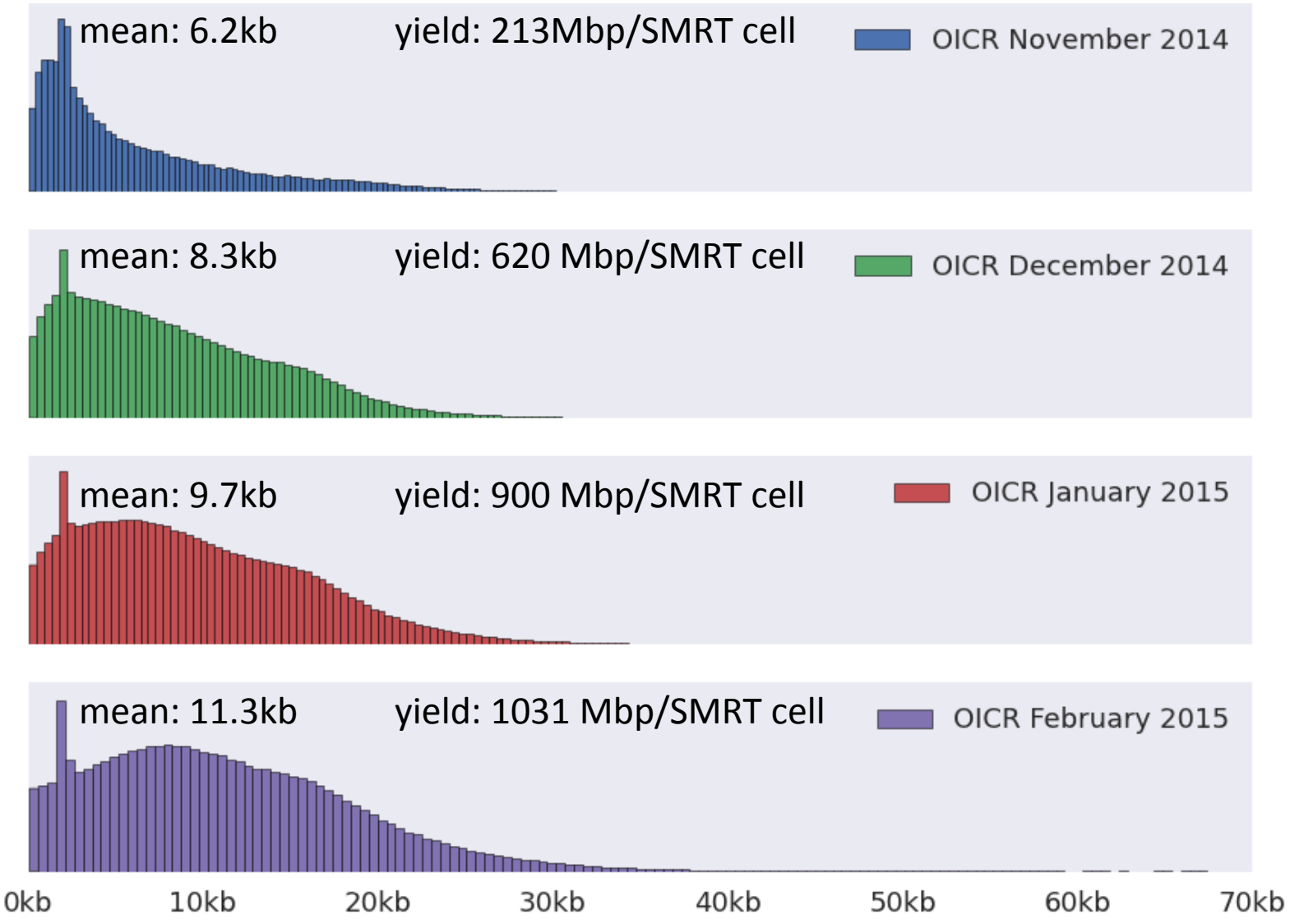

Maria Nattestad

- 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 collaboration with OICR, Hicks & Schatz labs at CSHL

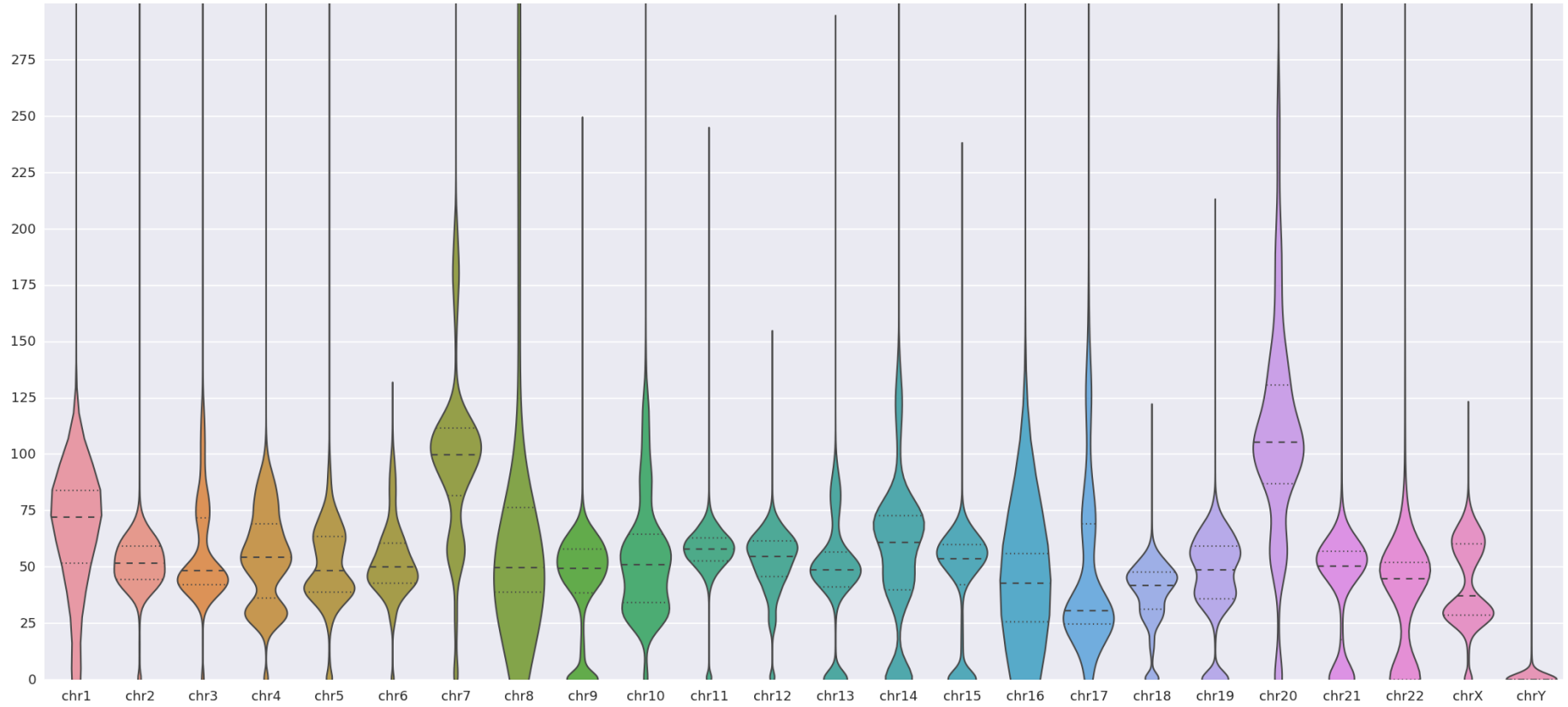
PacBio read length distribution



Improving SMRTcell Performance



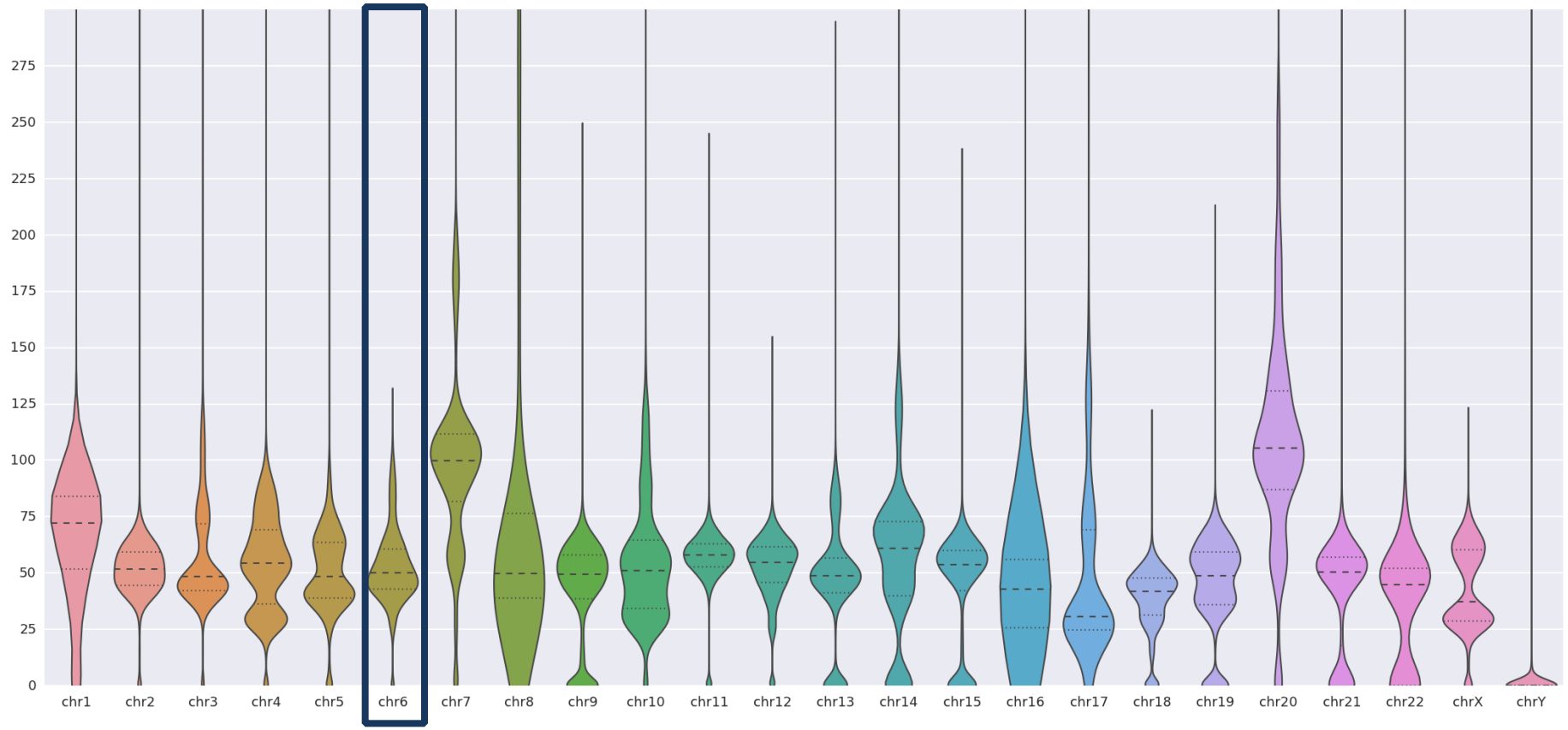
Genome-wide alignment coverage



Genome-wide coverage averages around 54X

Coverage per chromosome varies greatly as expected from previous karyotyping results

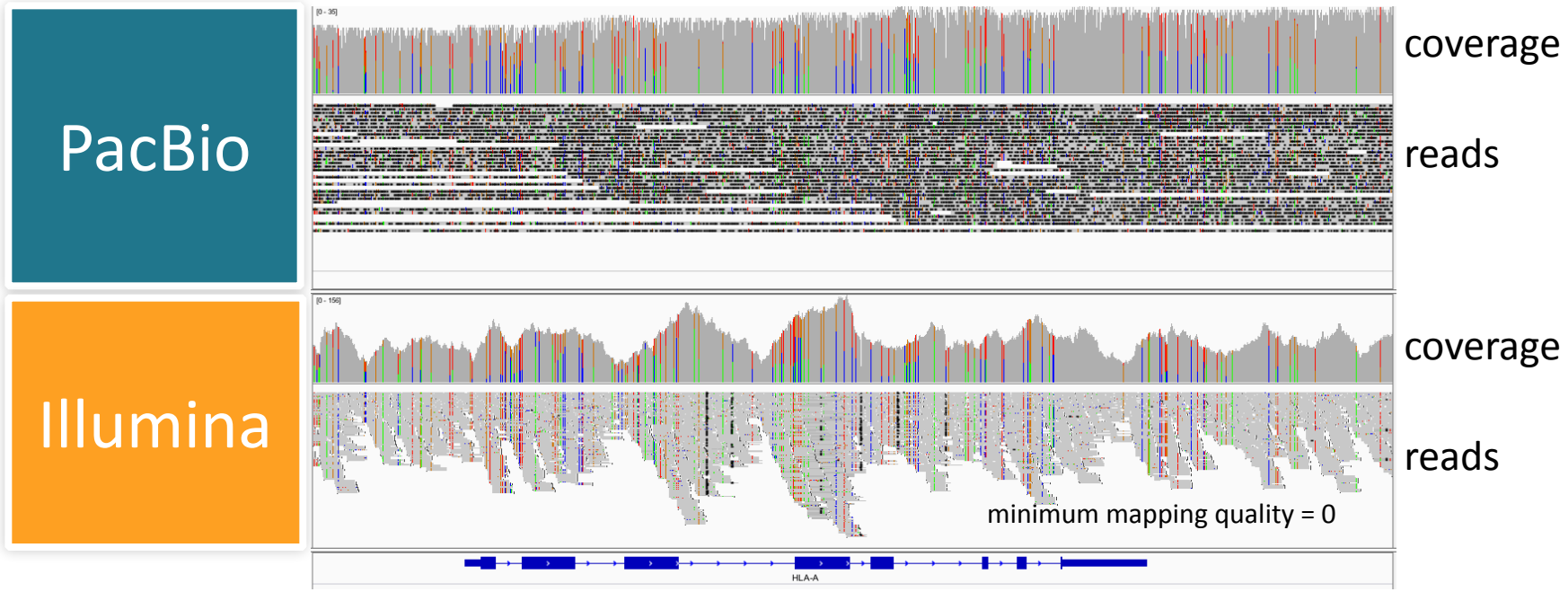
Genome-wide alignment coverage



Genome-wide coverage averages around 54X

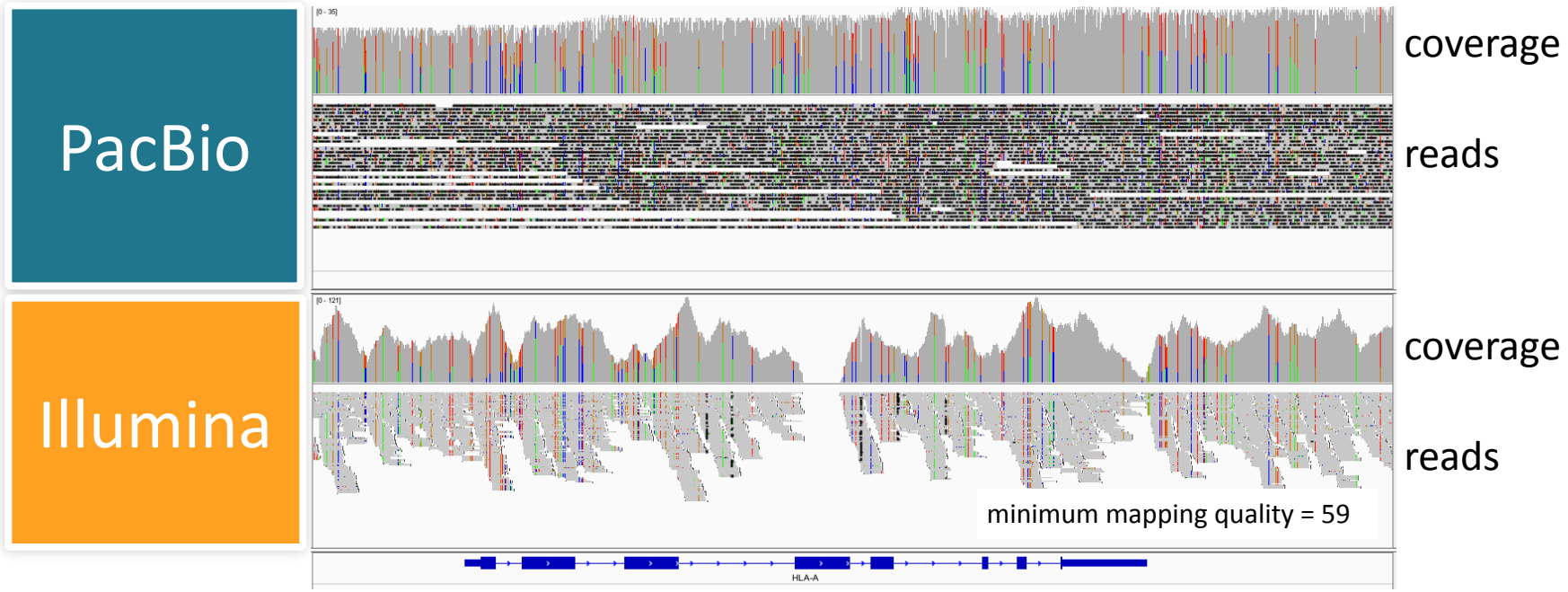
Coverage per chromosome varies greatly as expected from previous karyotyping results

PacBio coverage is more stable than Illumina coverage in repetitive regions



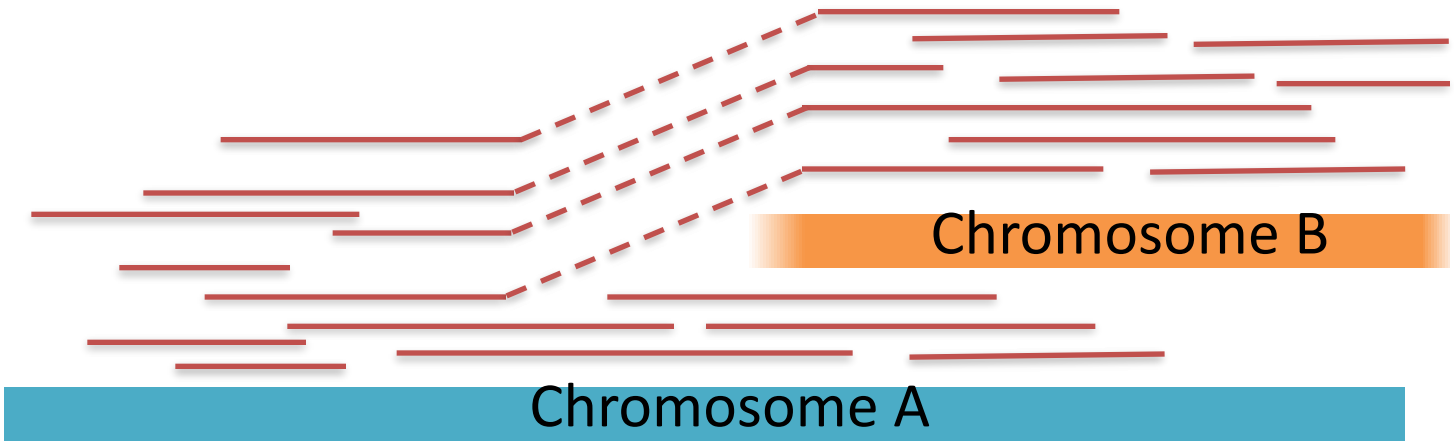
HLA-A gene

PacBio coverage is more stable than Illumina coverage in repetitive regions



HLA-A gene

Structural variant discovery with long reads



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

BWA-MEM + Lumpy

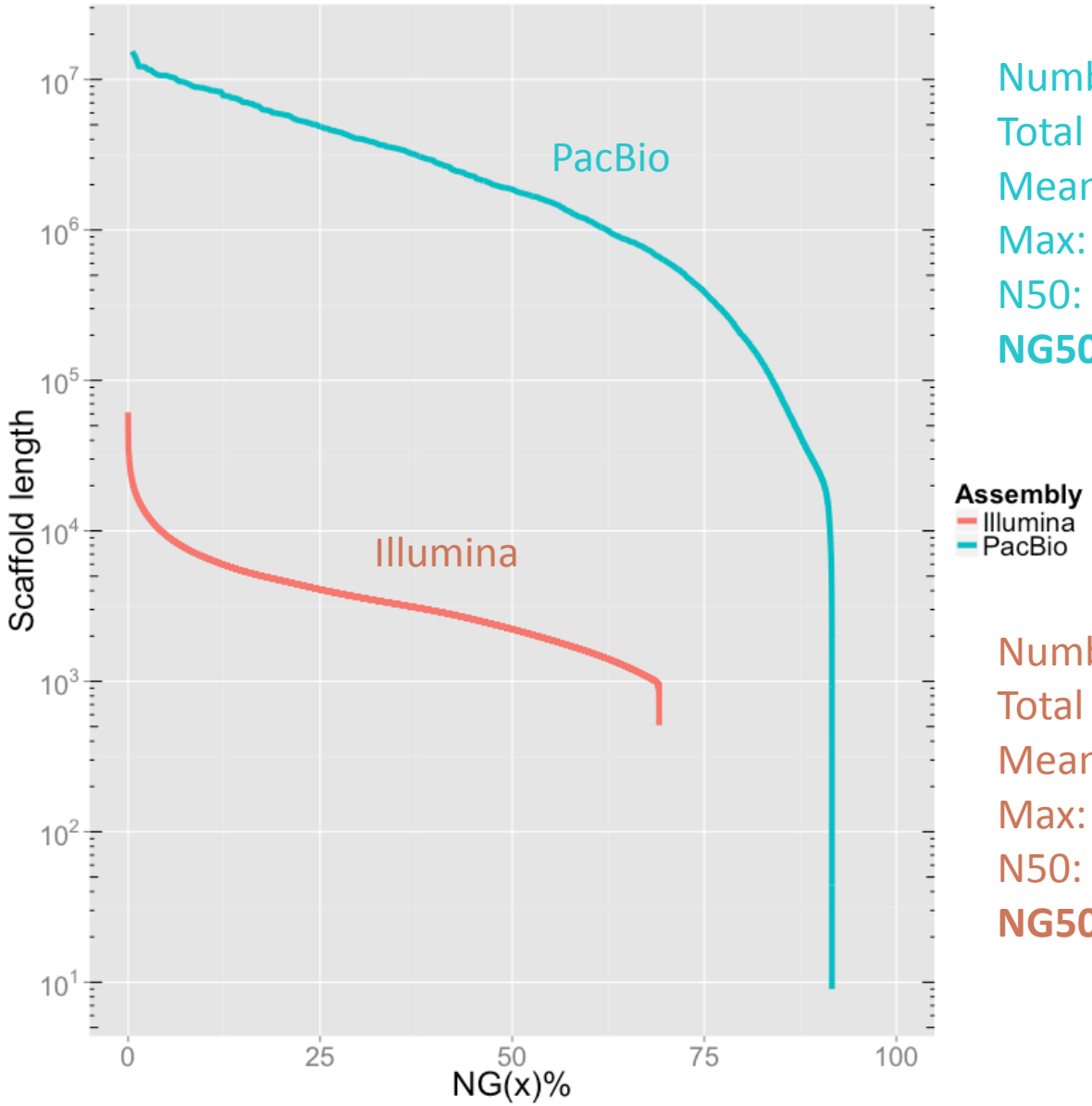
2. Local assembly of regions of interest: In-depth analysis with *base-pair precision*

Localized HGAP + Celera Assembler + MUMmer

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

DNAnexus-enabled version of Falcon

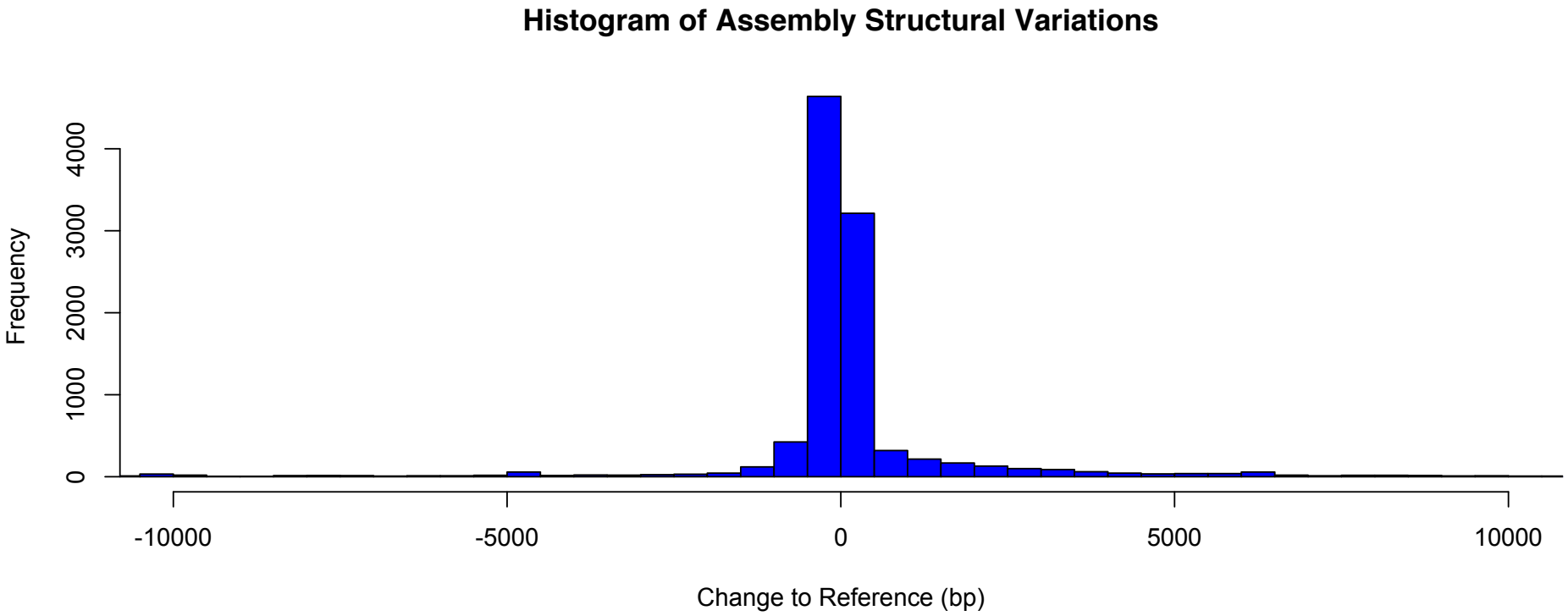
Assembly using PacBio yields far better contiguity



Number of sequences: 10,304
Total sequence length: 2.75 Gb
Mean: 266 kb
Max: 15 Mb
N50: 2.17 Mb
NG50: 1.86 Mb

Number of sequences: 748,955
Total sequence length: 2.07 Gb
Mean: 2.8 kb
Max: 61 kb
N50: 3.3 kb
NG50: 1.9 kb

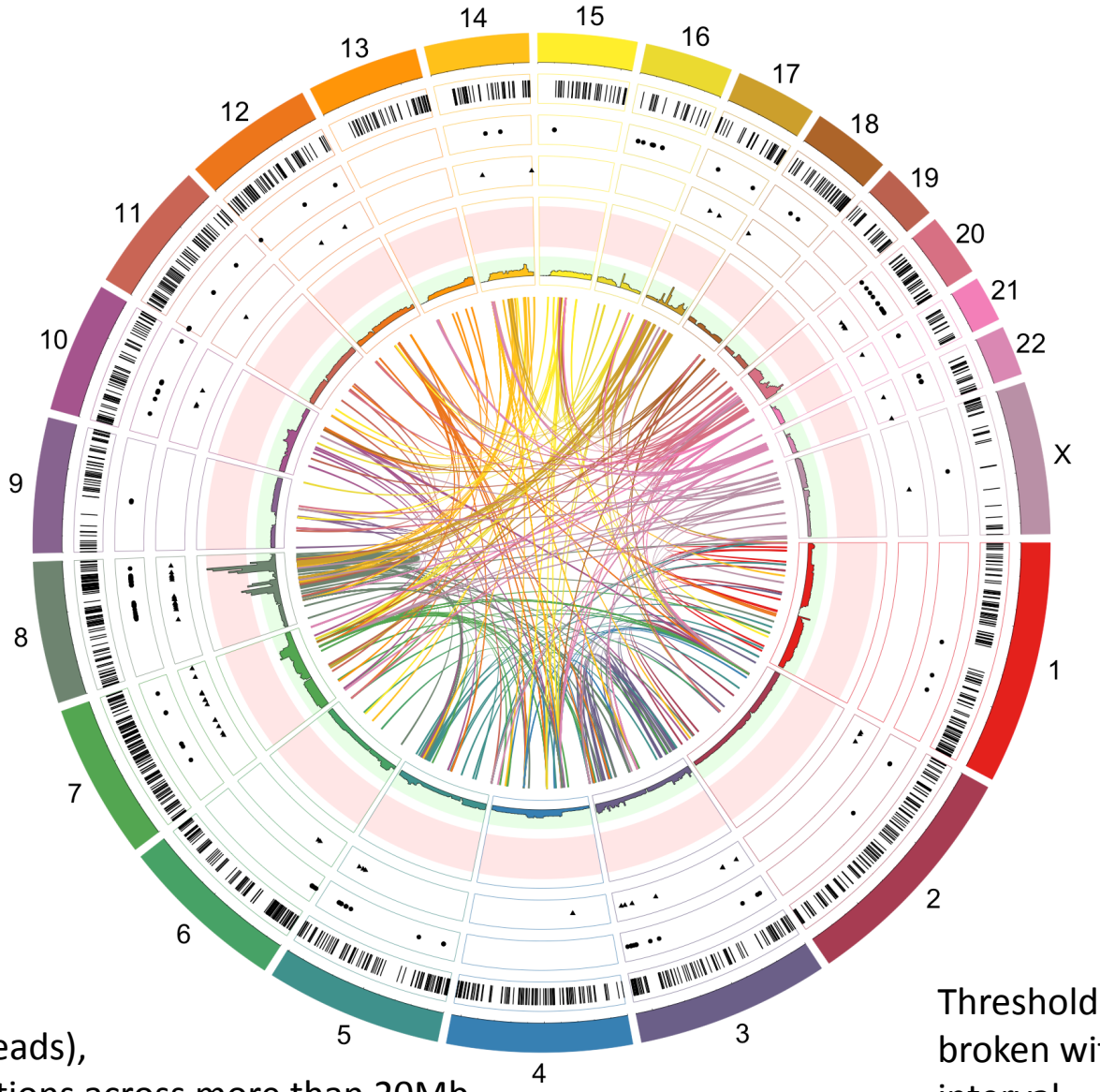
Indels in the Assembly



Assembly reveals >10,000 indels in the genome

- Whole genome alignment of assembly to reference genome with MUMmer/nucmer to discover insertions (>0) and deletions (<0)
- Enriched for SINE (+/- 300bp) and LINE(+/-6000bp) changes
- Over-representation of insertions, suggest reference genome systematically collapsed repetitive elements
- Many apparent changes to protein coding genes

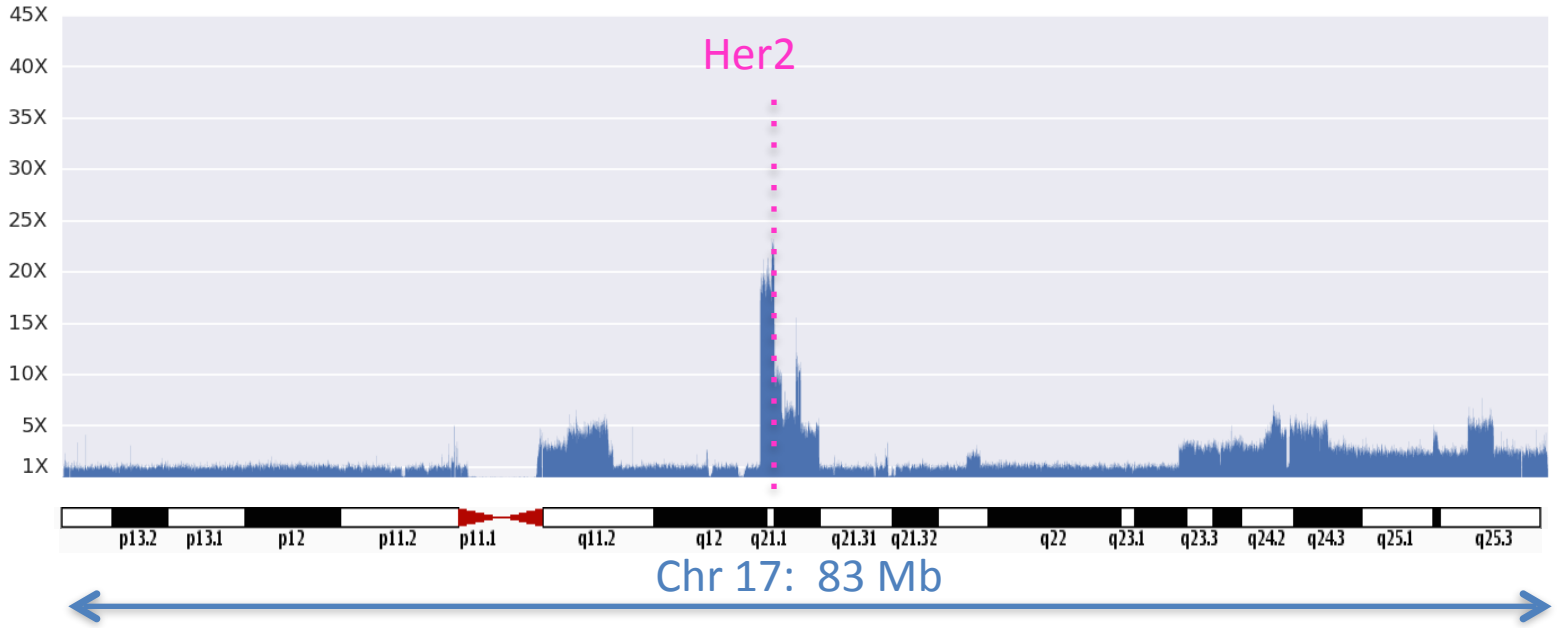
Structural variants found by split-read mapping



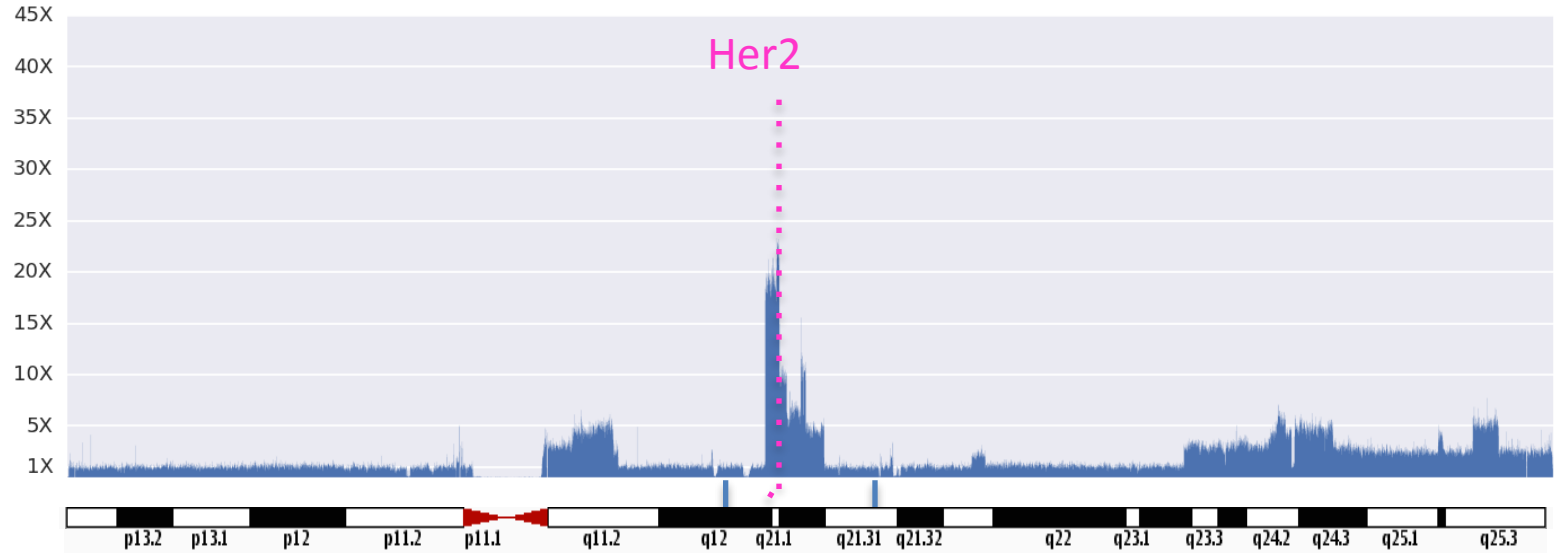
From the outside:
deletions,
inversions,
duplications,
coverage (pink: 300+ reads),
connections: translocations across more than 20Mb

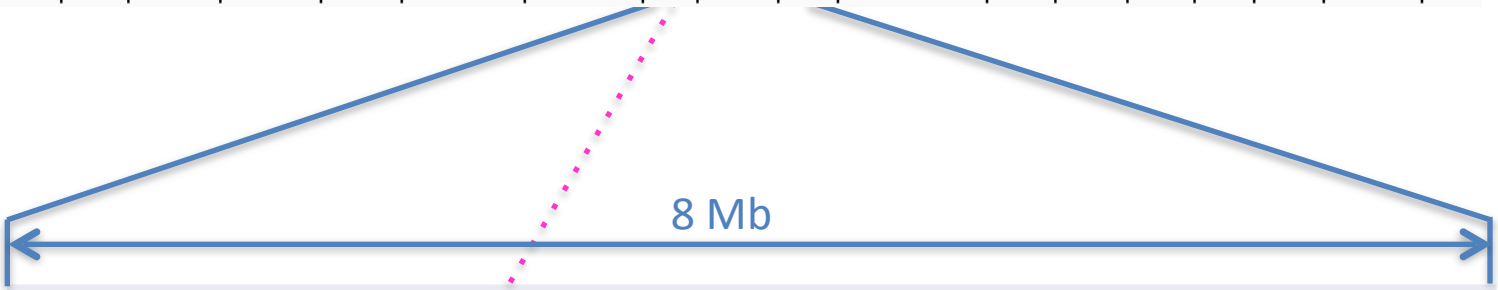
Threshold: 10 split reads broken within 200 bp interval

PacBio

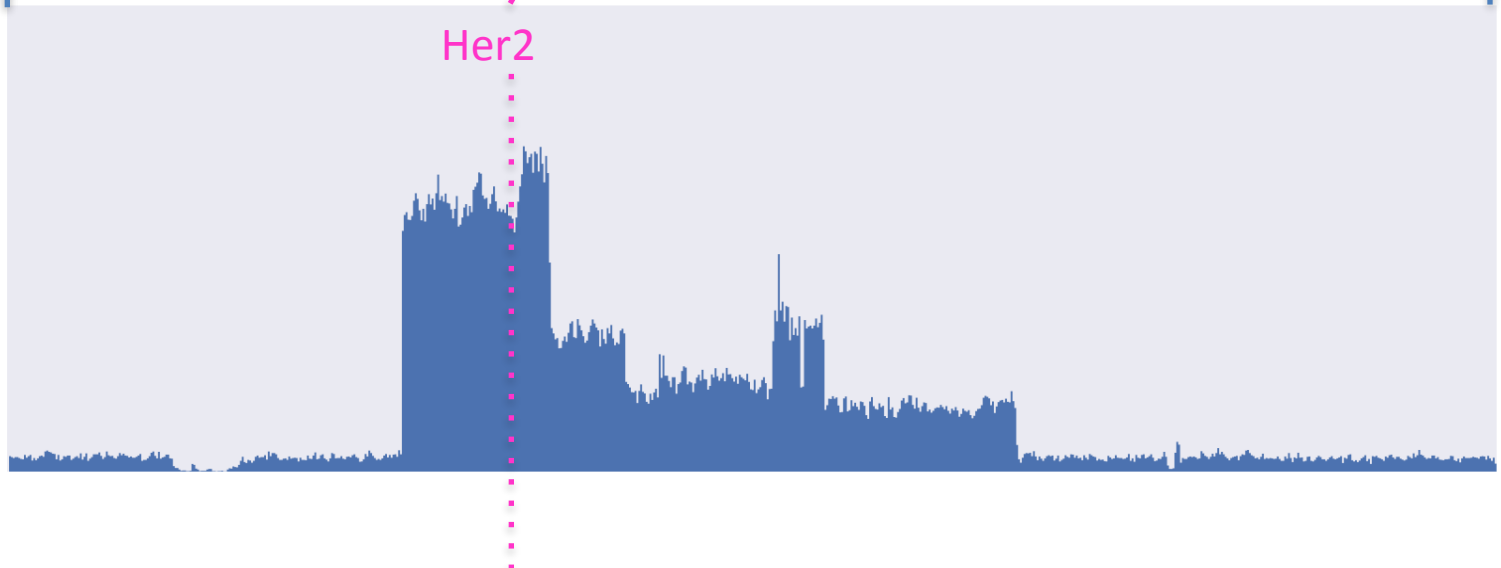

45X
40X
35X
30X
25X
20X
15X
10X
5X
1X
Her2
p13.2
p13.1
p12
p11.2
p11.1
q11.2
q12
q21.1
q21.31
q21.32
q22
q23.1
q23.3
q24.2
q24.3
q25.1
q25.3
Chr 17: 83 Mb

PacBio

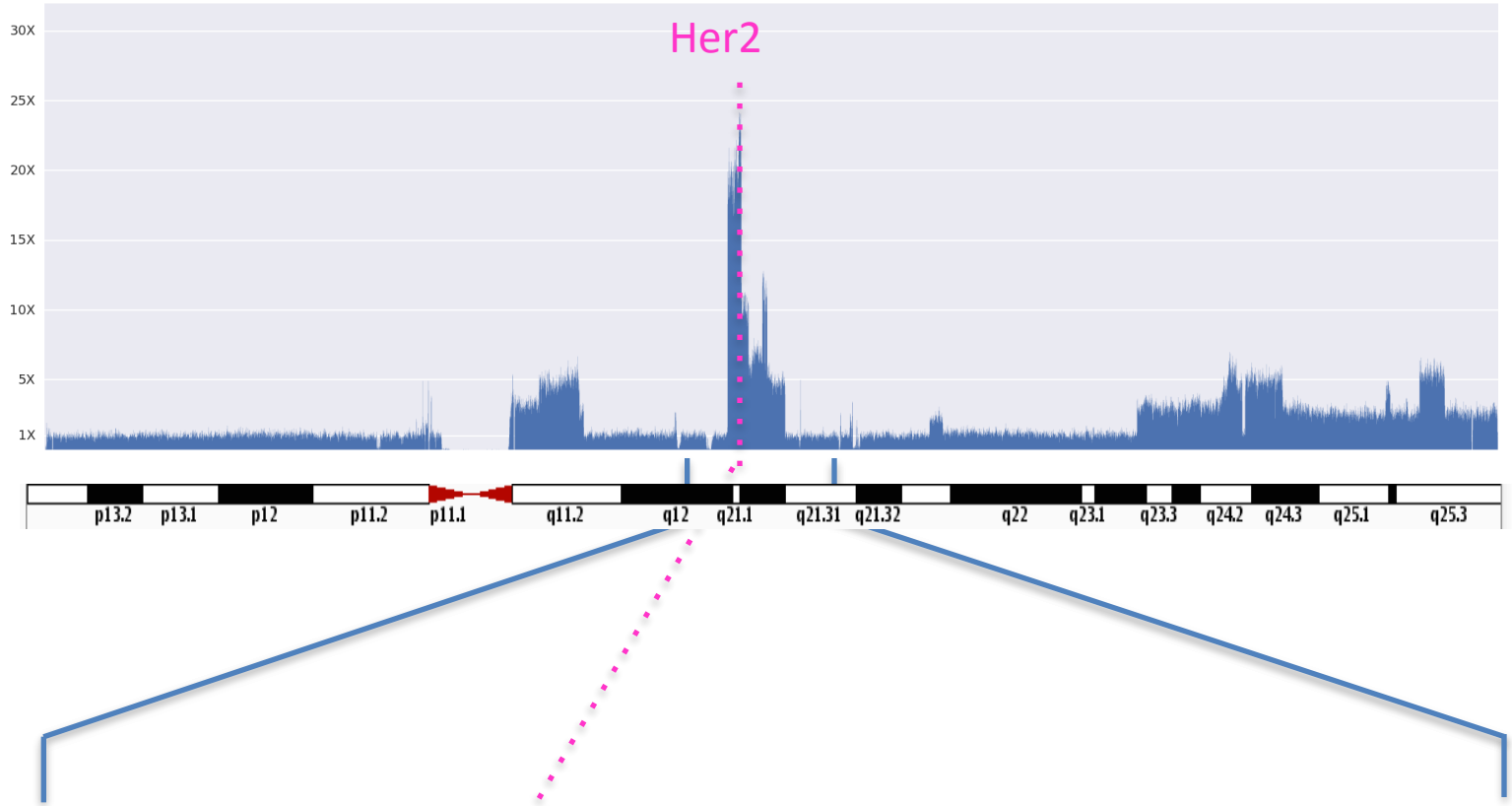




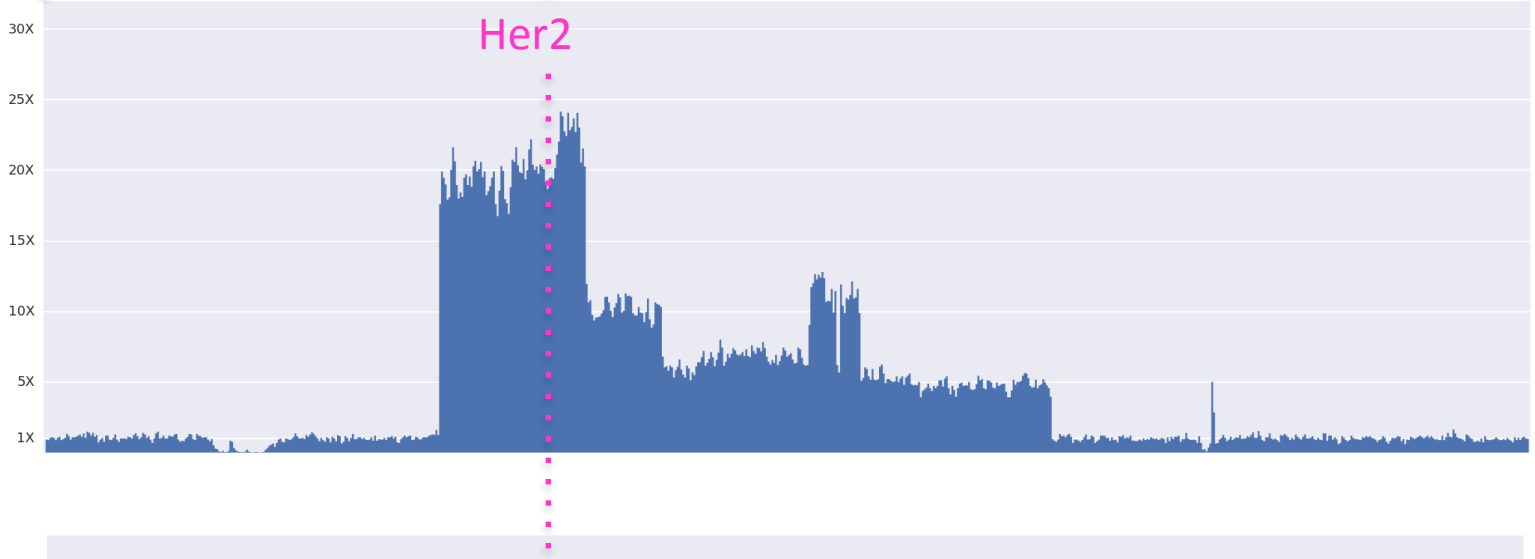
PacBio

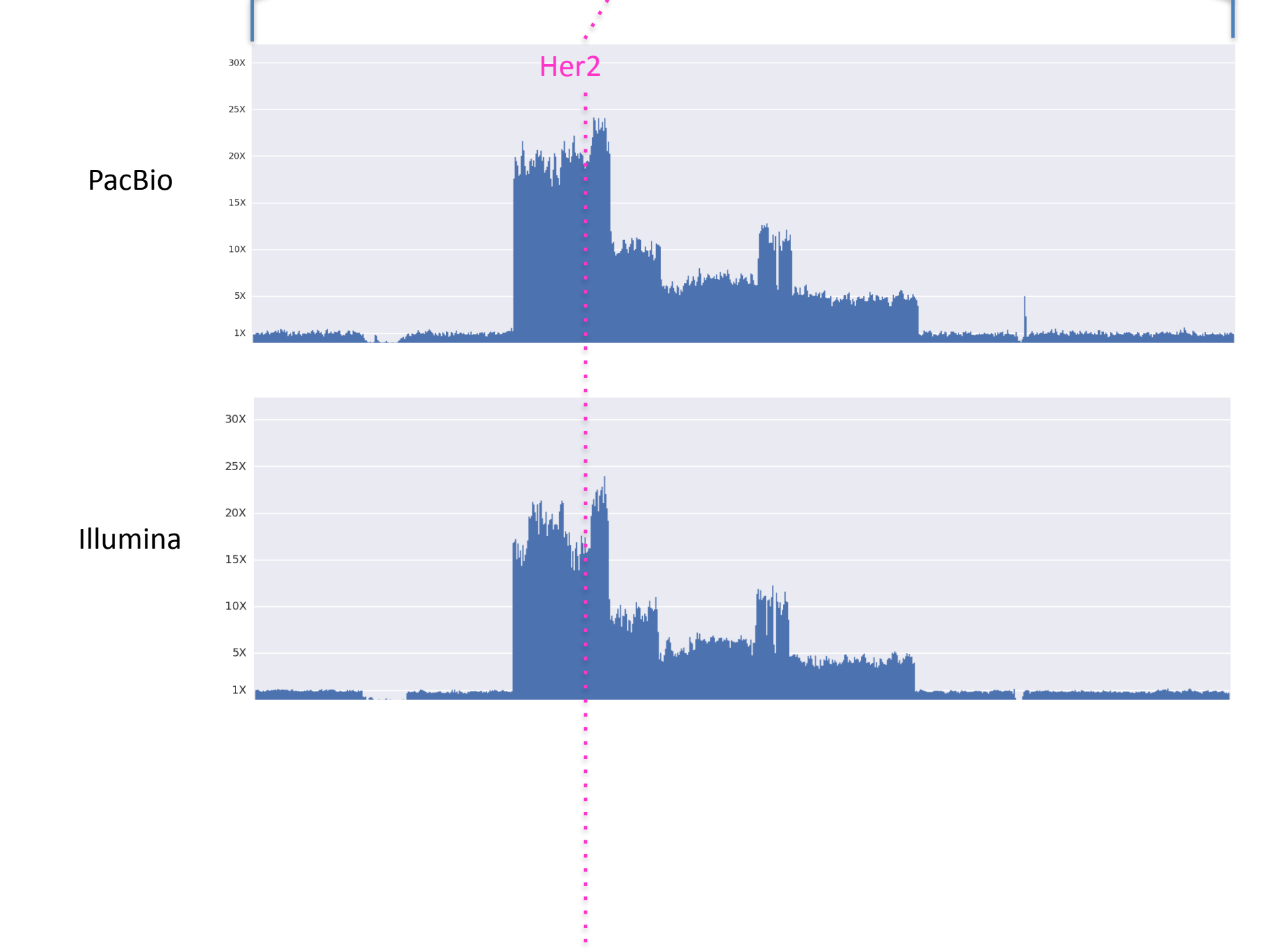


PacBio



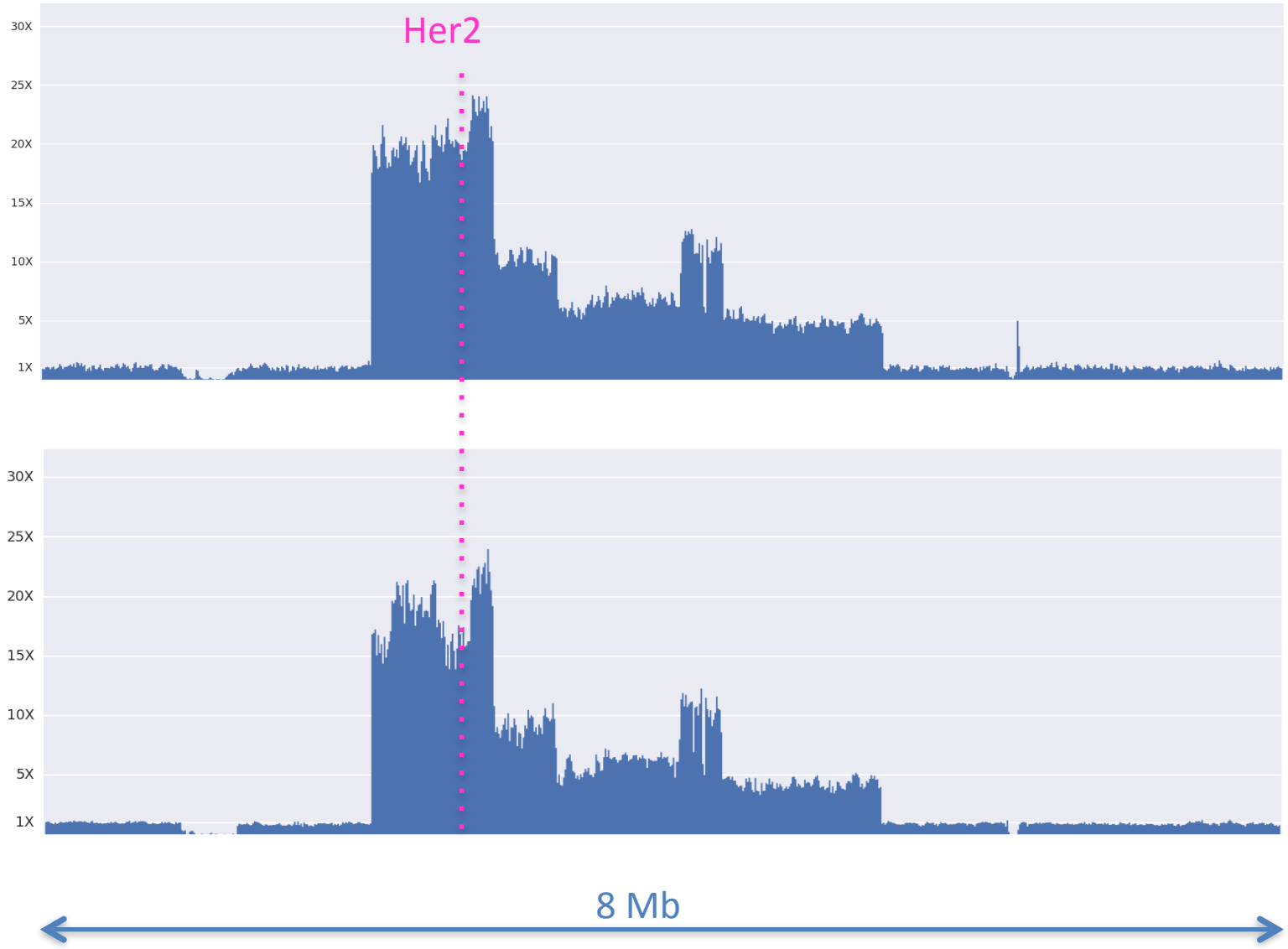
PacBio

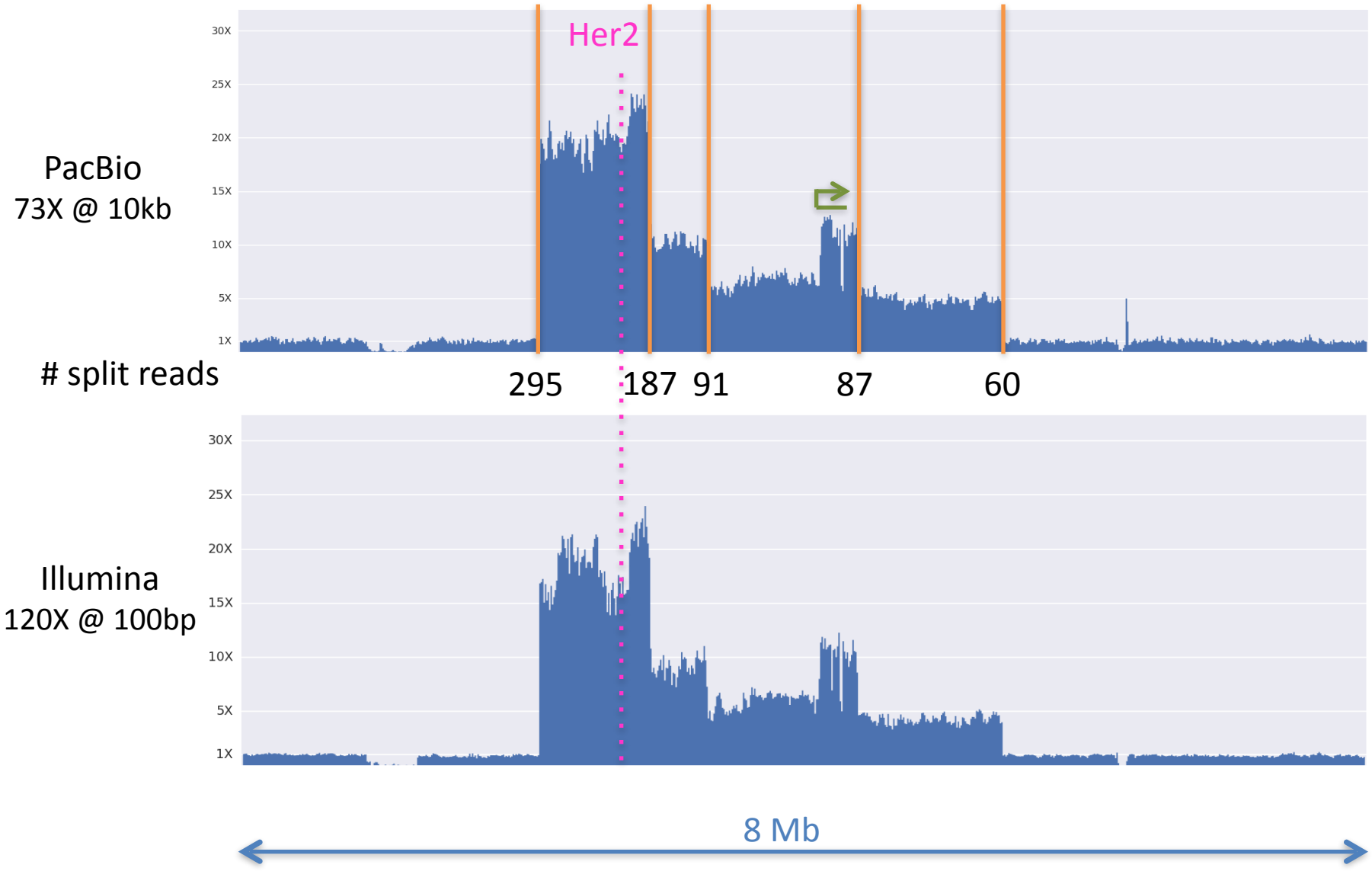



Her2
PacBio
30X
25X
20X
15X
10X
5X
1X
Illumina
30X
25X
20X
15X
10X
5X
1X

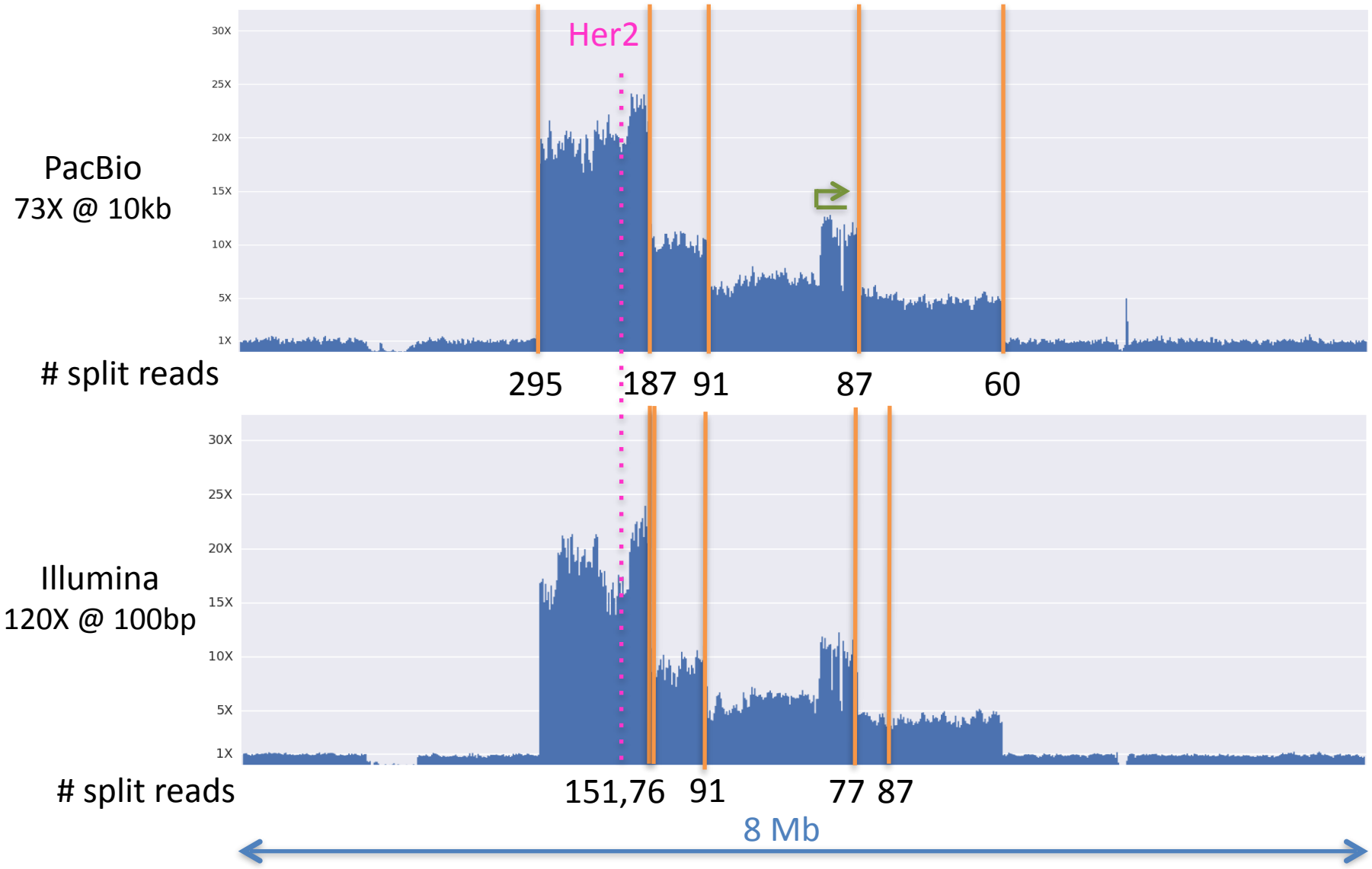
PacBio
73X @ 10kb

Illumina
120X @ 100bp





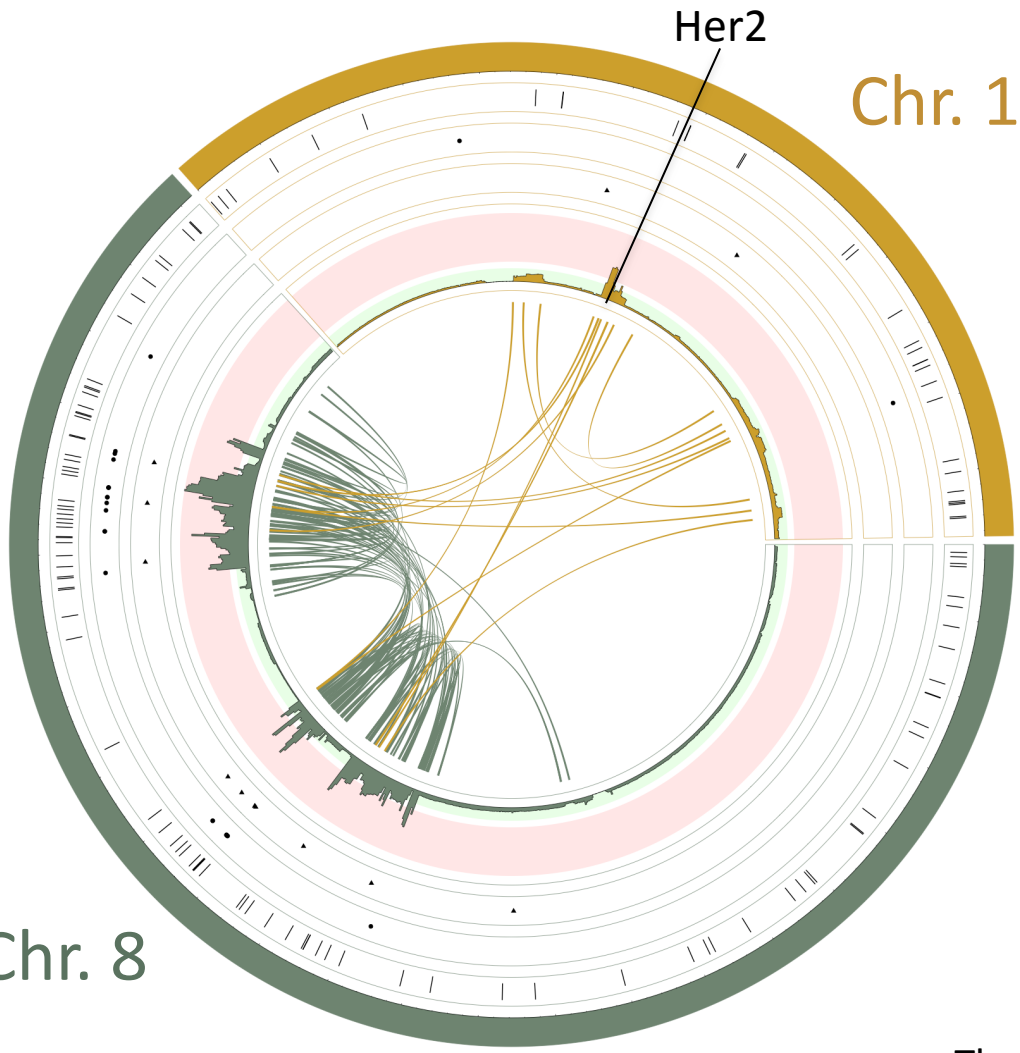
Green arrow indicates an inverted duplication.



Green arrow indicates an inverted duplication.

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

All five translocations near Her2 go to chromosome 8

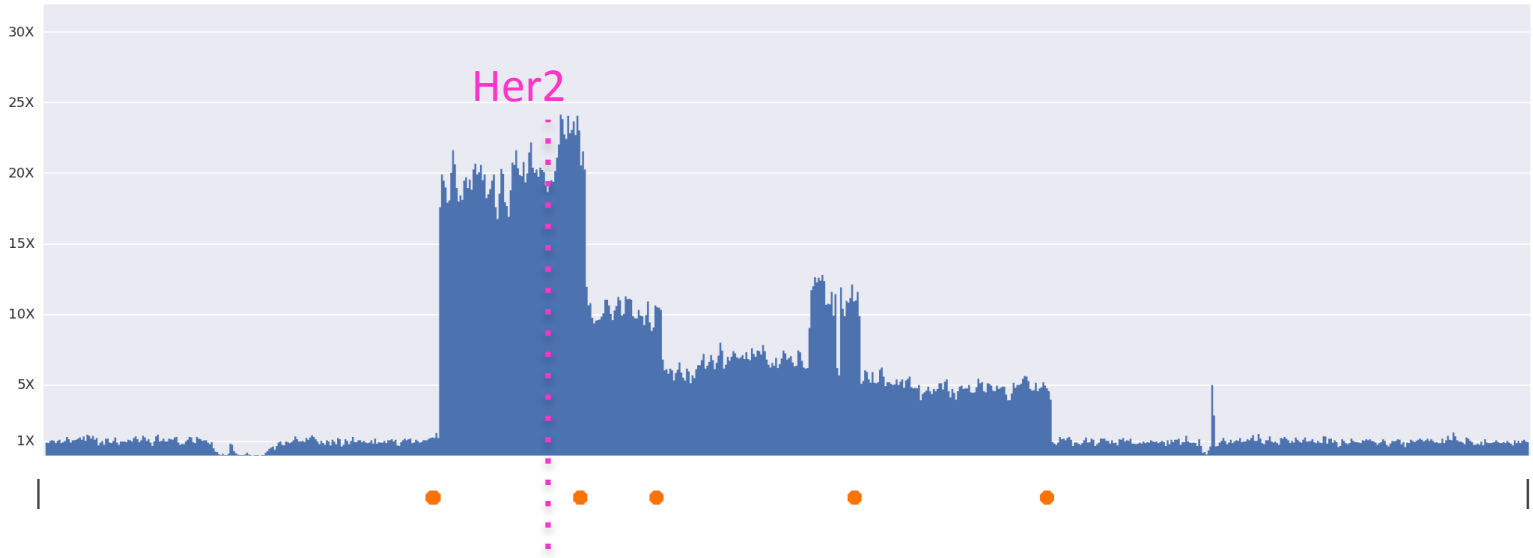


From the outside:
deletions,
inversions,
duplications,
coverage (pink: 300+ reads),
connections: translocations across more than **2 Mb**

Threshold: 10 split reads broken within 200 bp interval

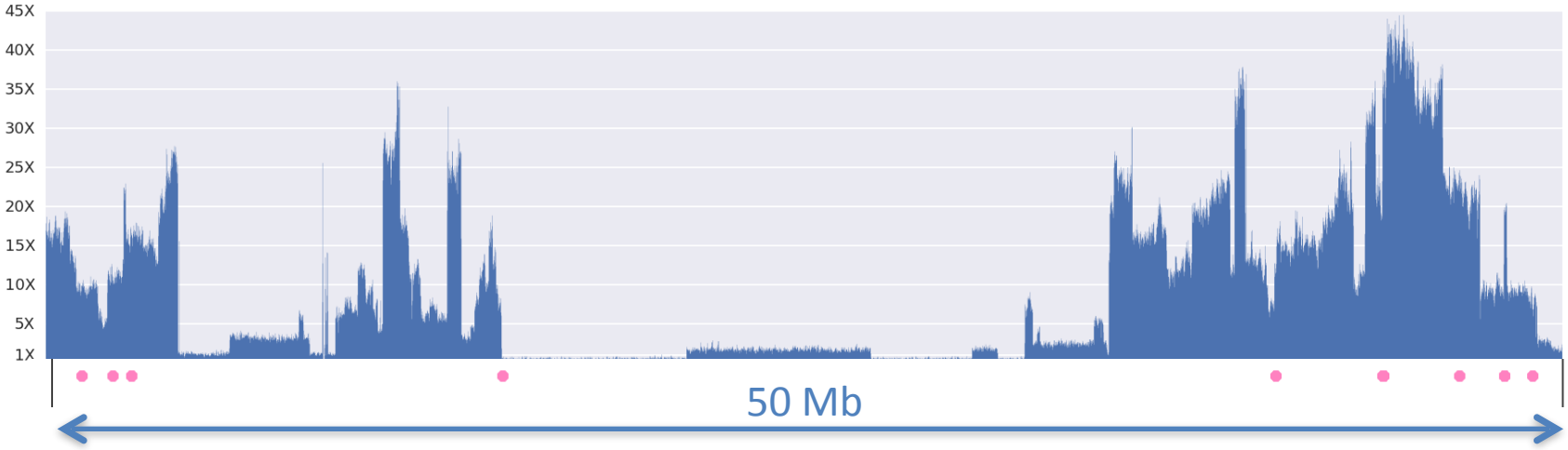
PacBio

chr17



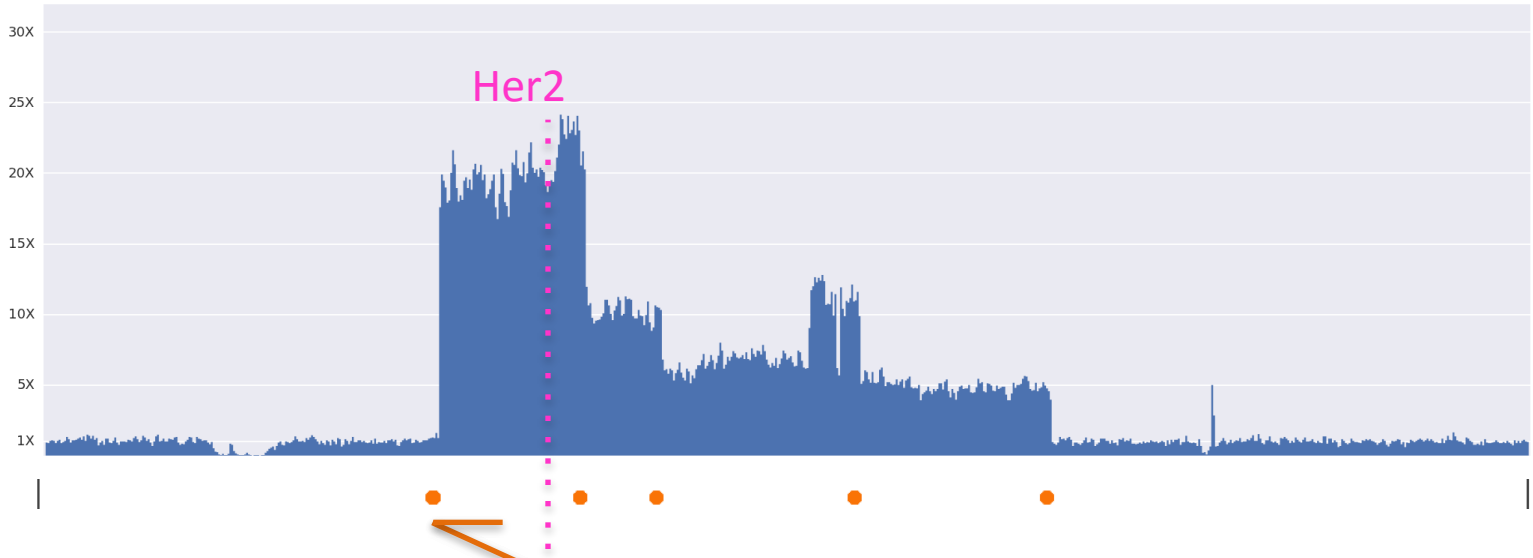
PacBio

chr8



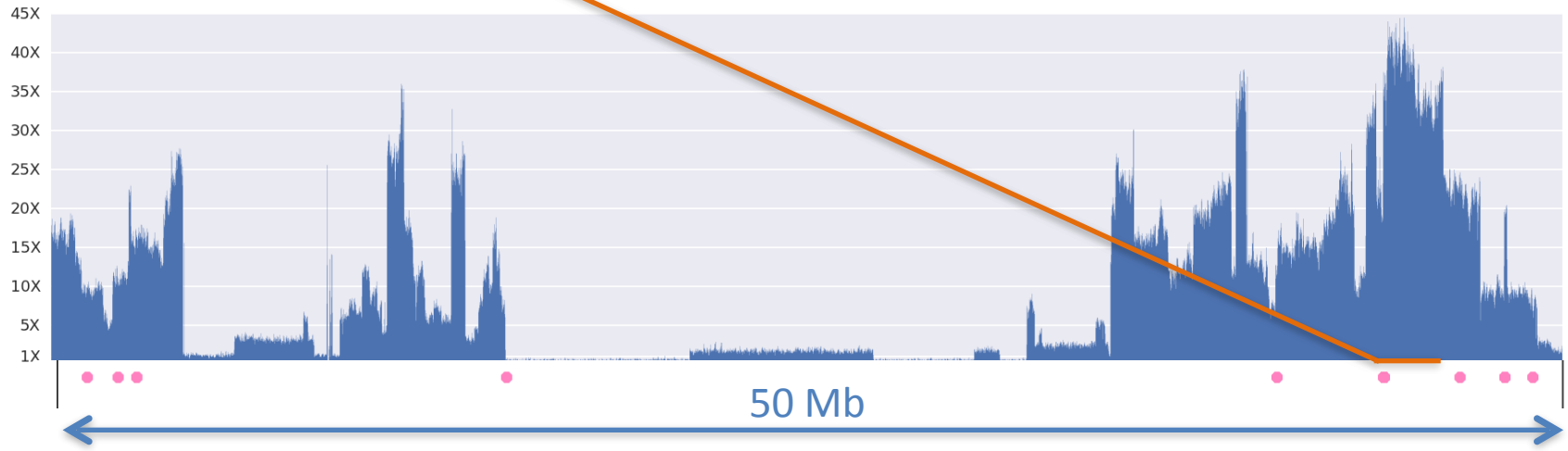
PacBio

chr17



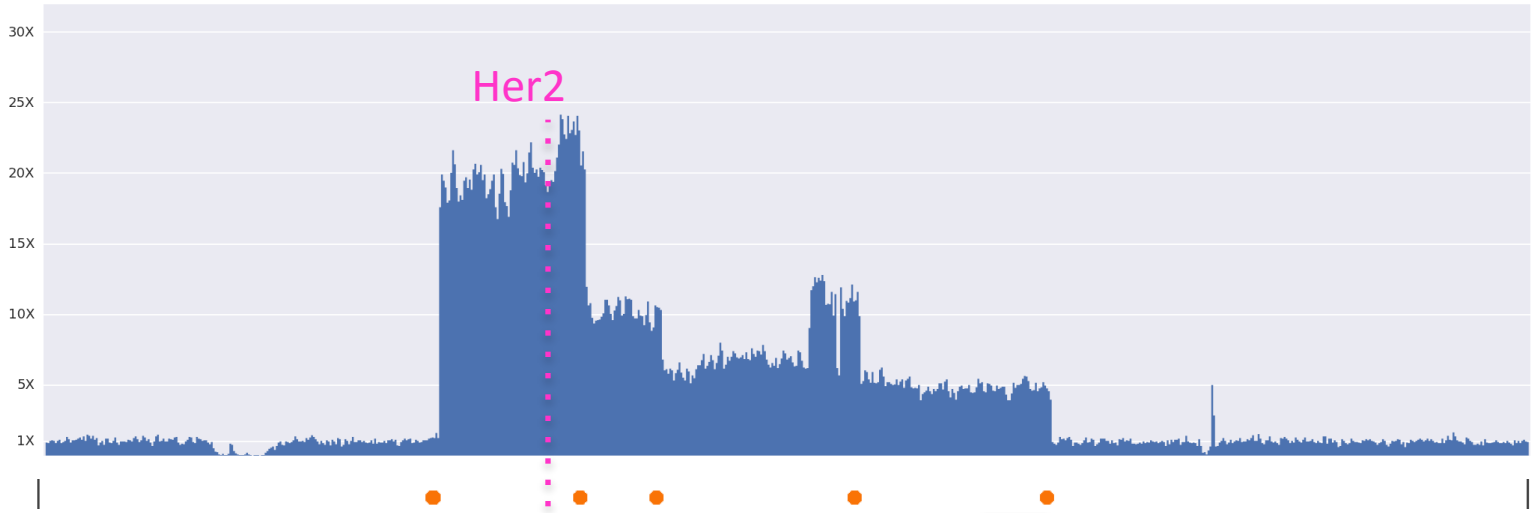
PacBio

chr8



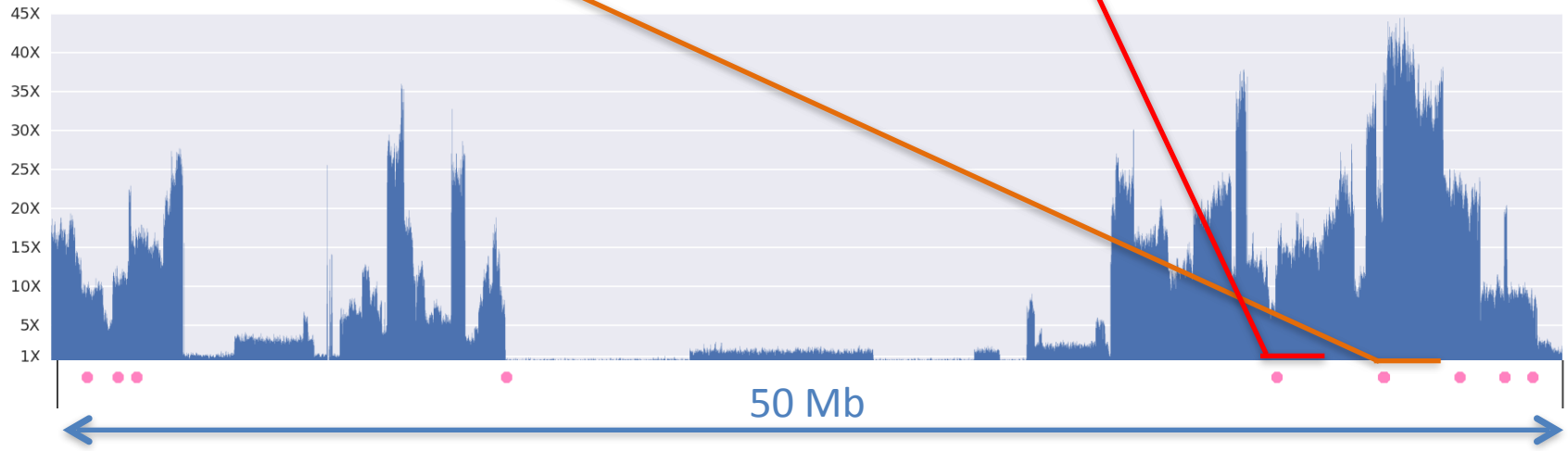
PacBio

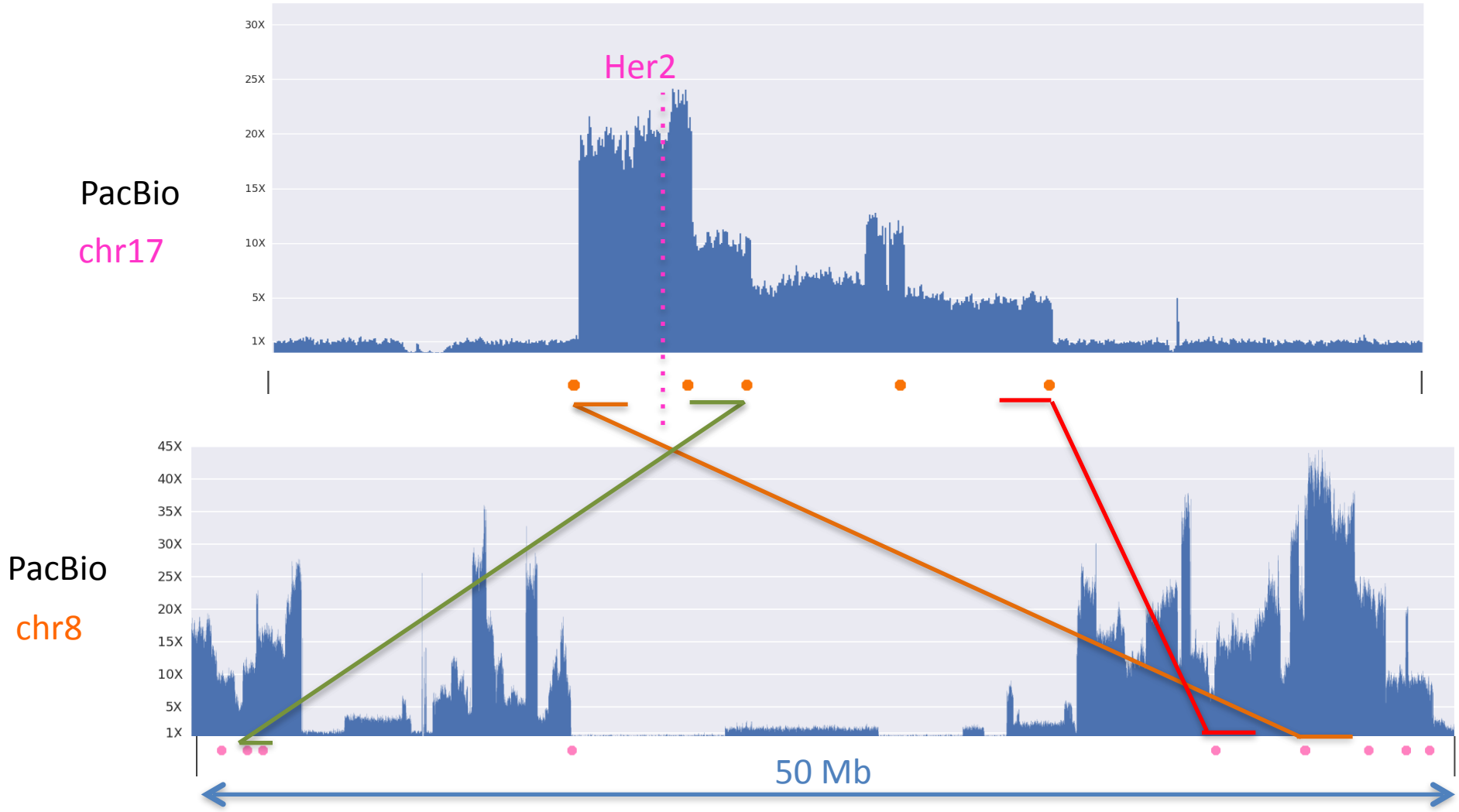
chr17

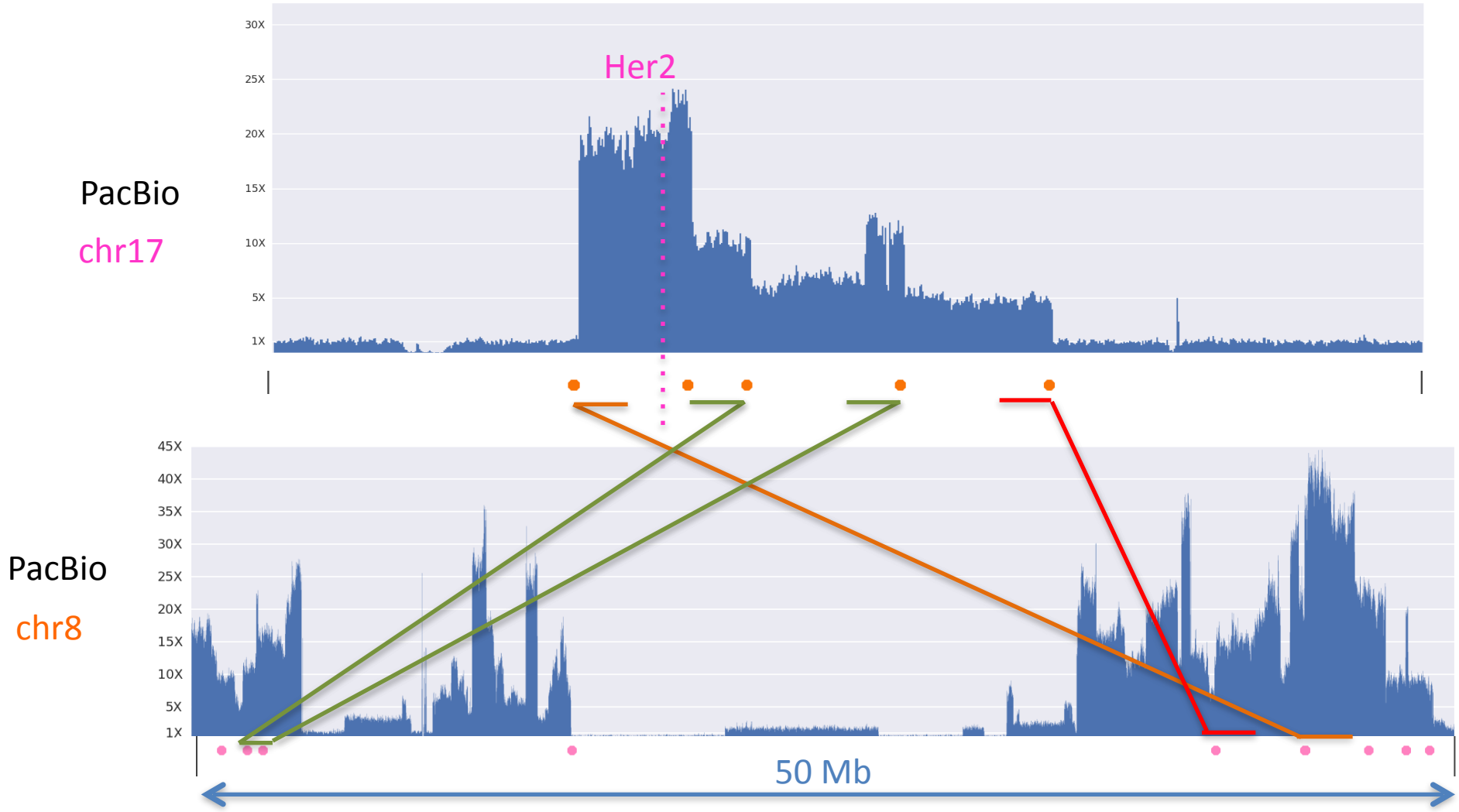


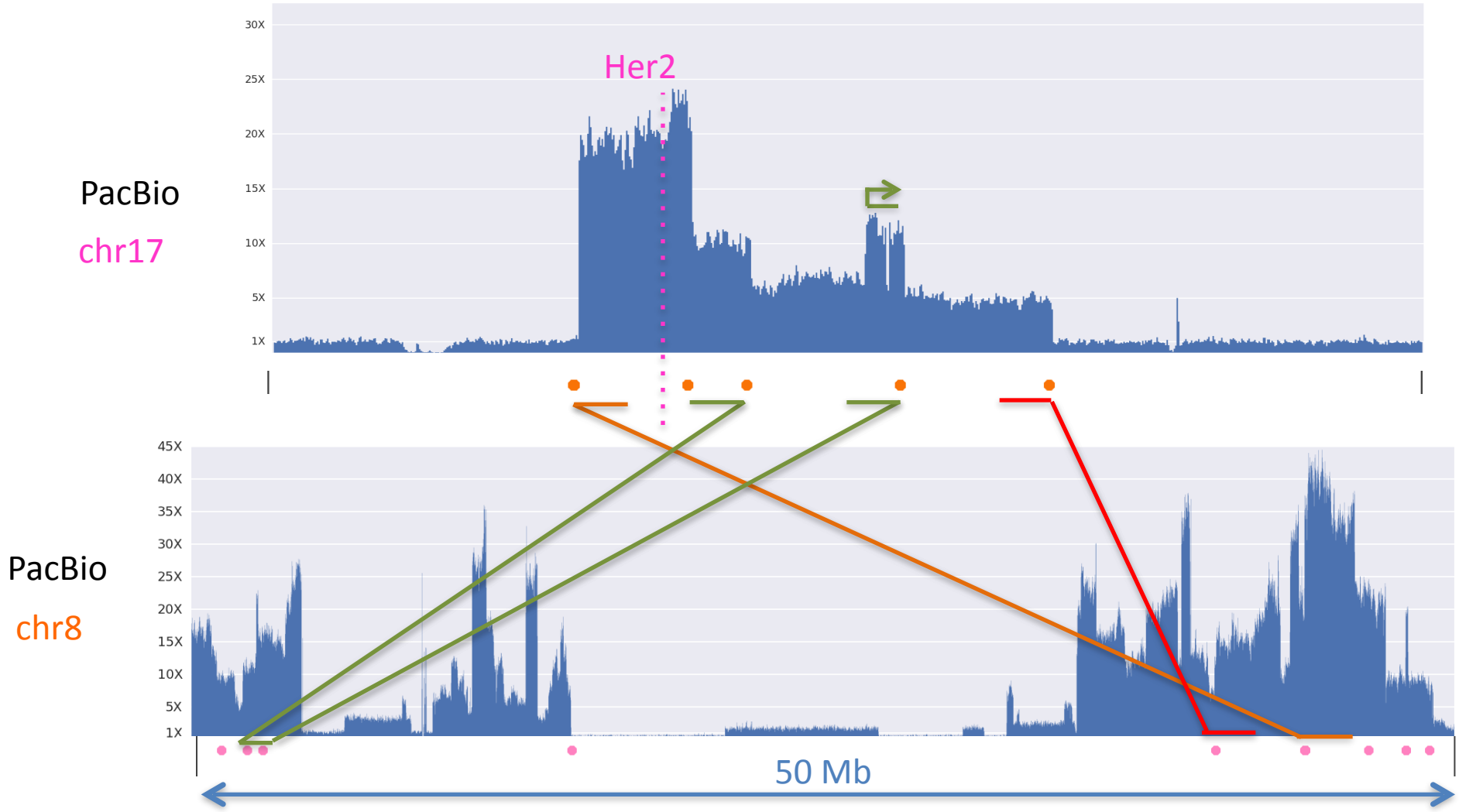
PacBio

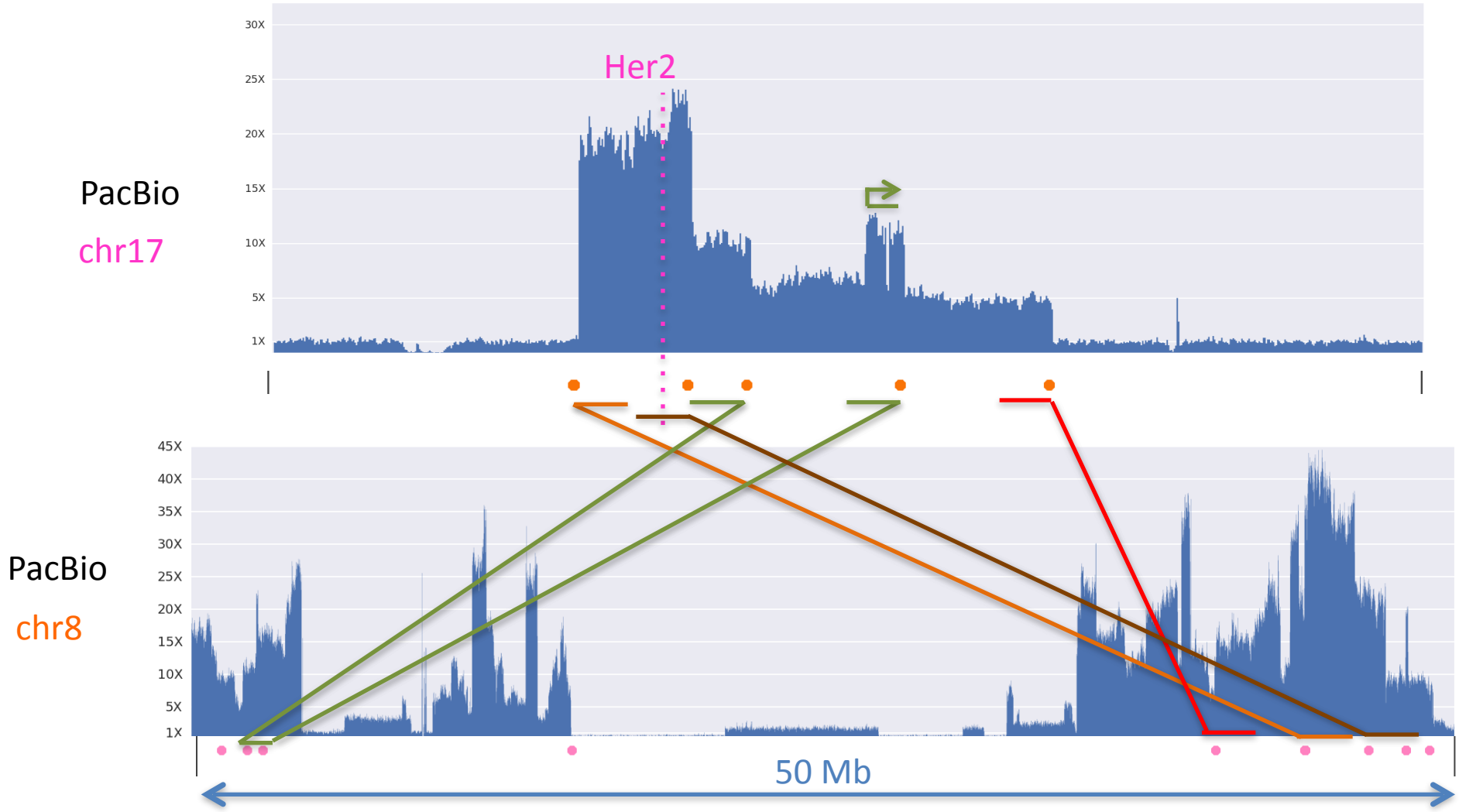
chr8

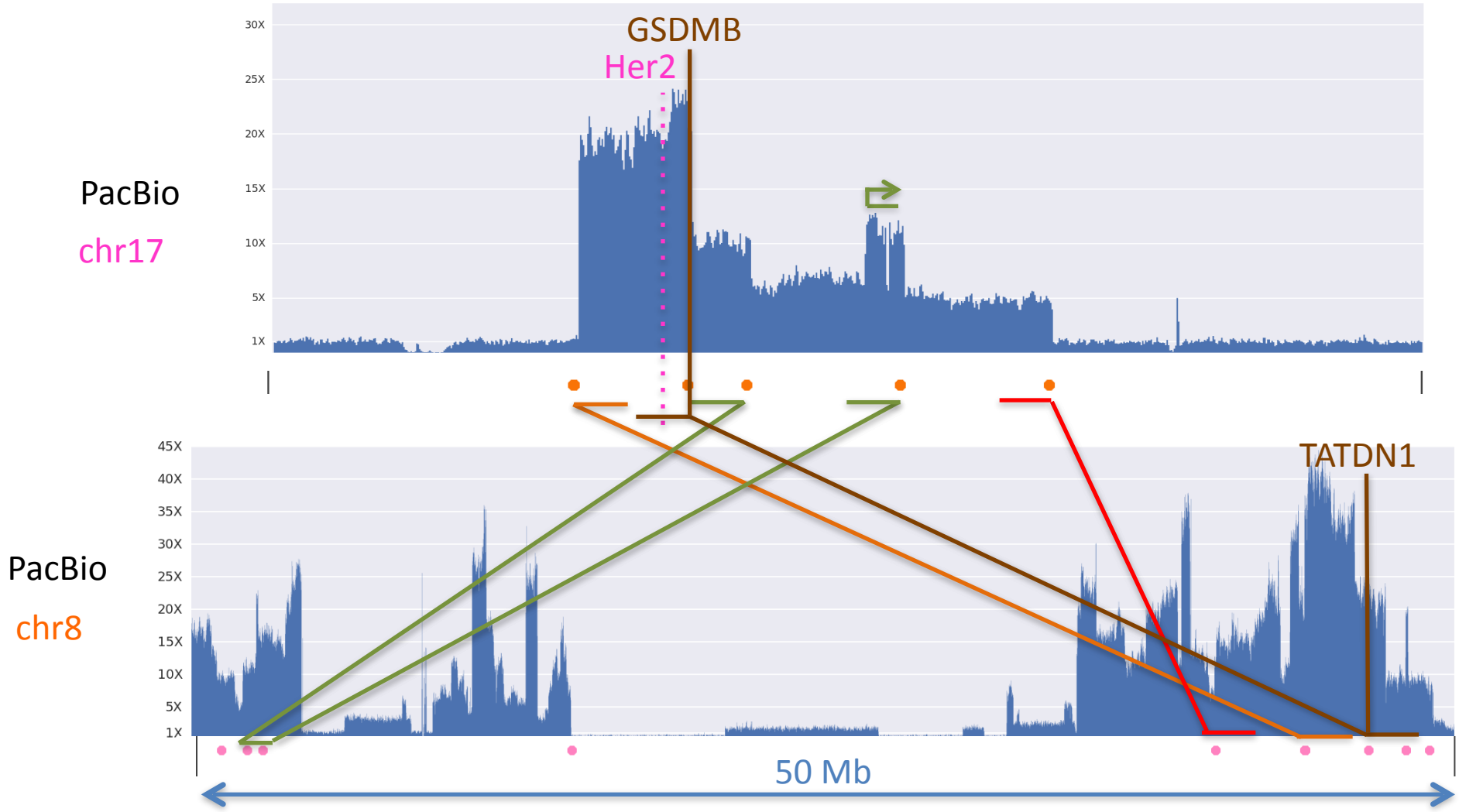


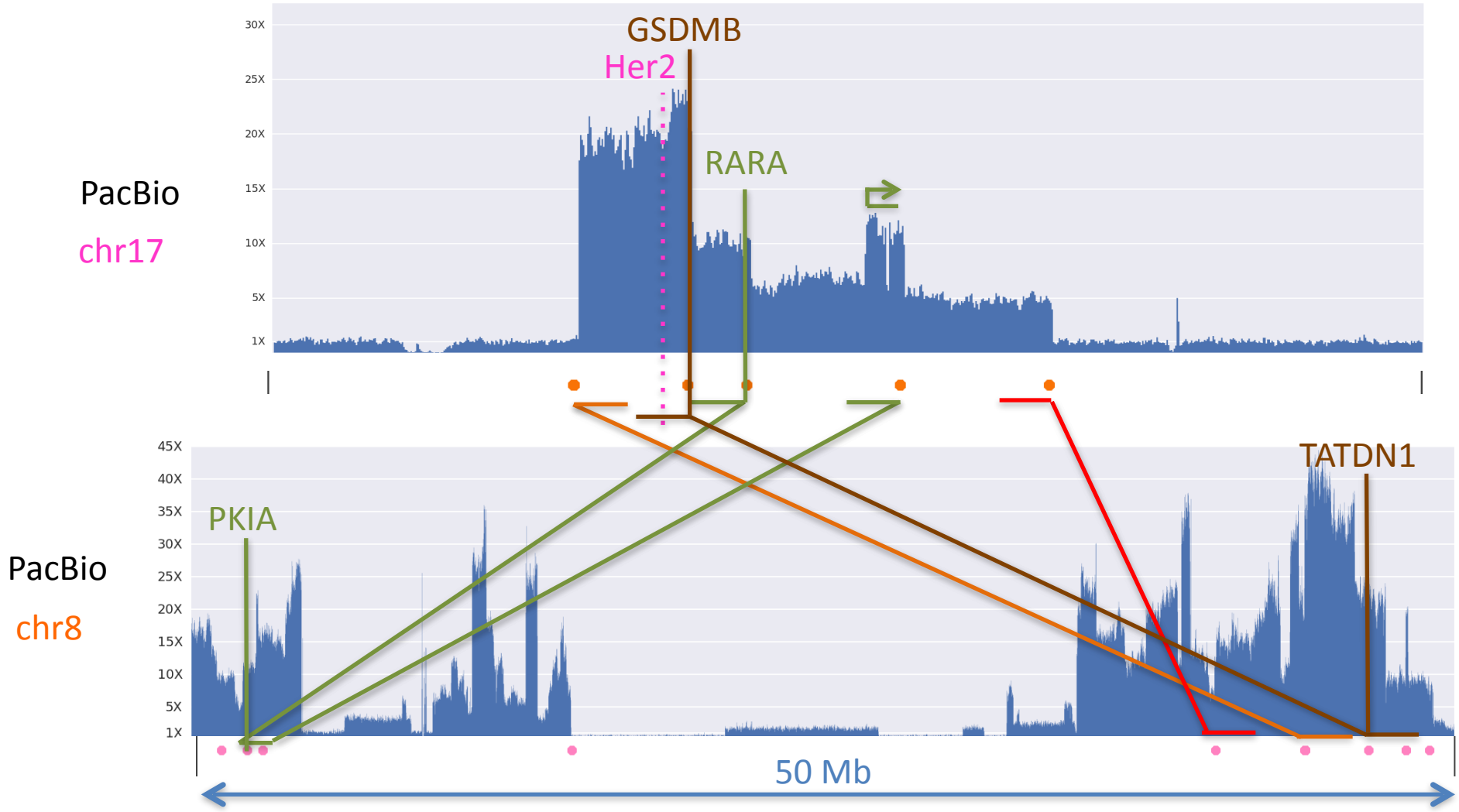

PacBio
chr17
30X
25X
20X
15X
10X
5X
1X
Her2
PacBio
chr8
45X
40X
35X
30X
25X
20X
15X
10X
5X
1X
50 Mb


PacBio
chr17
30X
25X
20X
15X
10X
5X
1X
Her2
PacBio
chr8
45X
40X
35X
30X
25X
20X
15X
10X
5X
1X
50 Mb

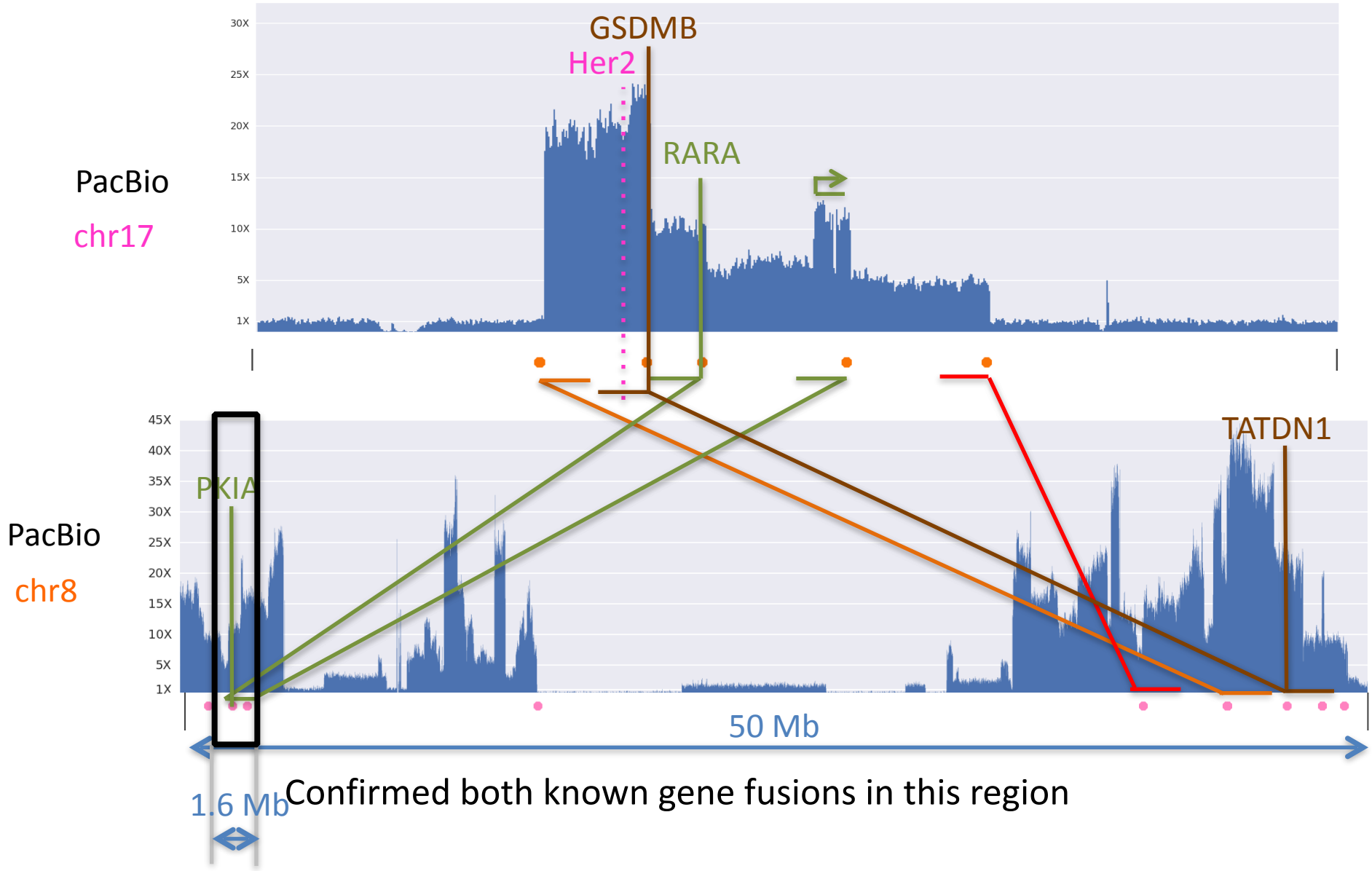

PacBio
chr17
30X
25X
20X
15X
10X
5X
1X
Her2
PacBio
chr8
45X
40X
35X
30X
25X
20X
15X
10X
5X
1X
50 Mb


PacBio
chr17
30X
25X
20X
15X
10X
5X
1X
Her2
PacBio
chr8
45X
40X
35X
30X
25X
20X
15X
10X
5X
1X
50 Mb


PacBio
chr17
30X
25X
20X
15X
10X
5X
1X
GSDMB
Her2
PacBio
chr8
45X
40X
35X
30X
25X
20X
15X
10X
5X
1X
TATDN1
50 Mb

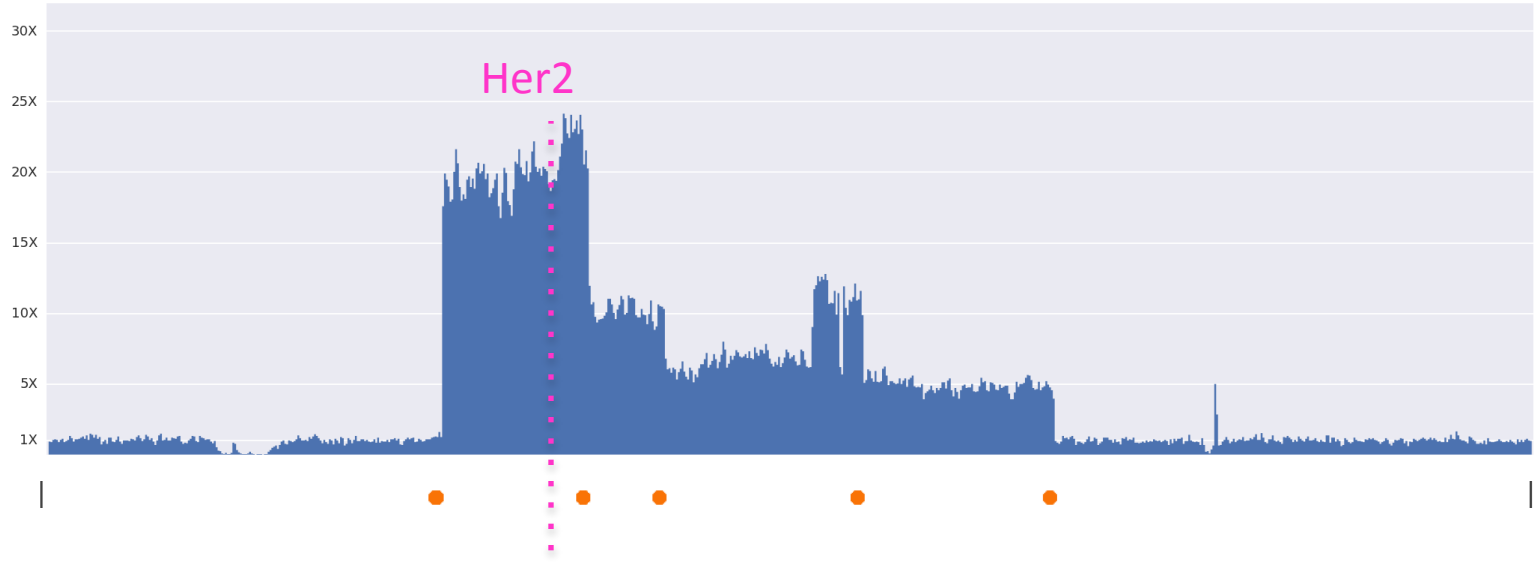


Confirmed both known gene fusions in this region

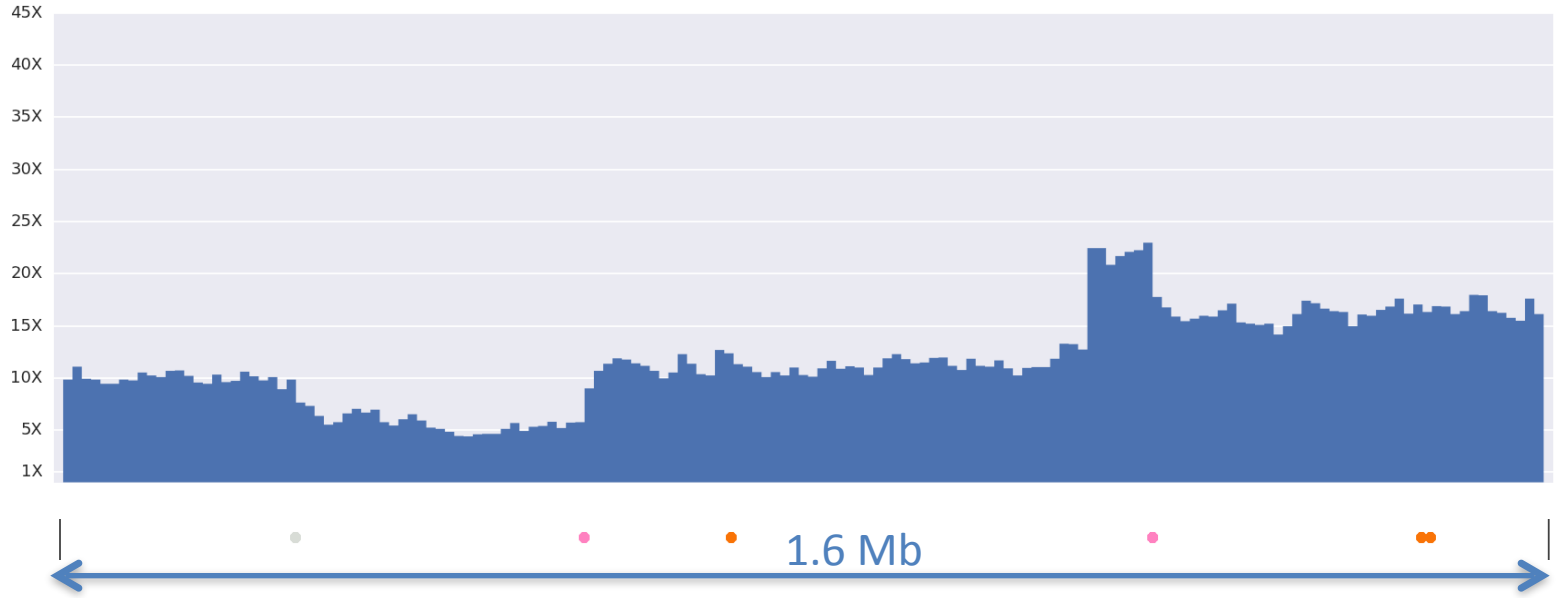

PacBio
chr17
30X
25X
20X
15X
10X
5X
1X
GSDMB
Her2
RARA
PacBio
chr8
45X
40X
35X
30X
25X
20X
15X
10X
5X
1X
PKIA
TATDN1
50 Mb
1.6 Mb
Confirmed both known gene fusions in this region

PacBio

chr17

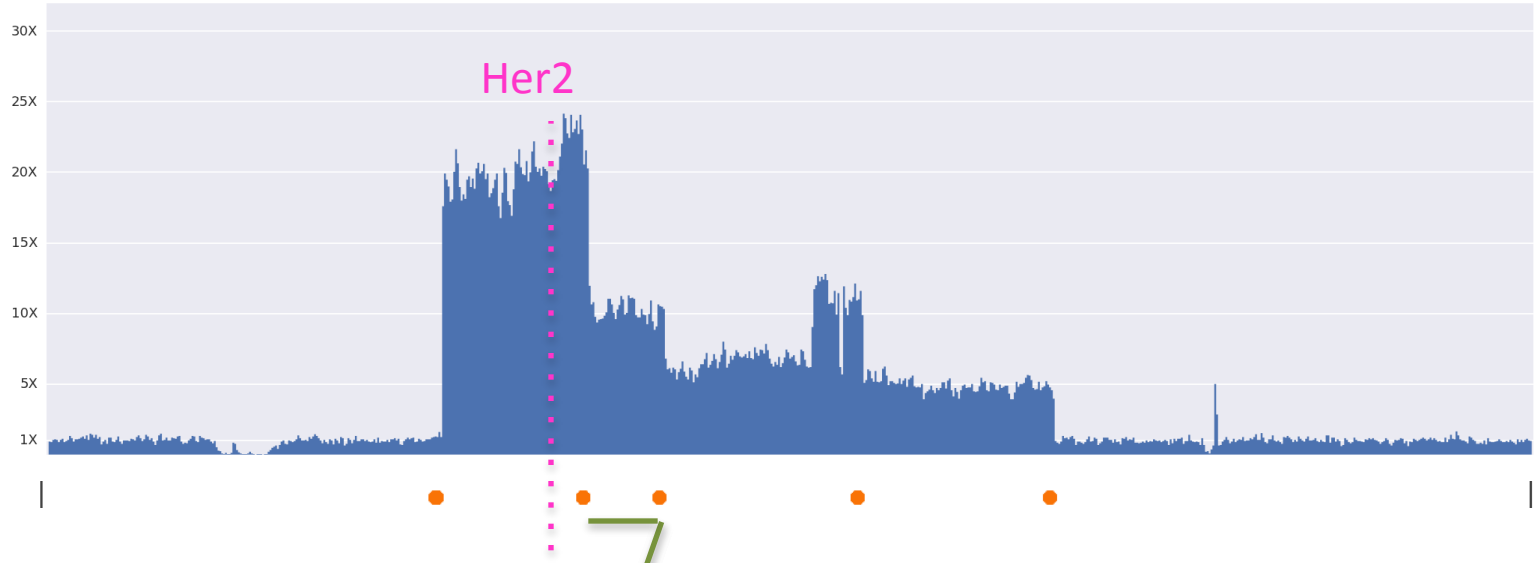


chr8

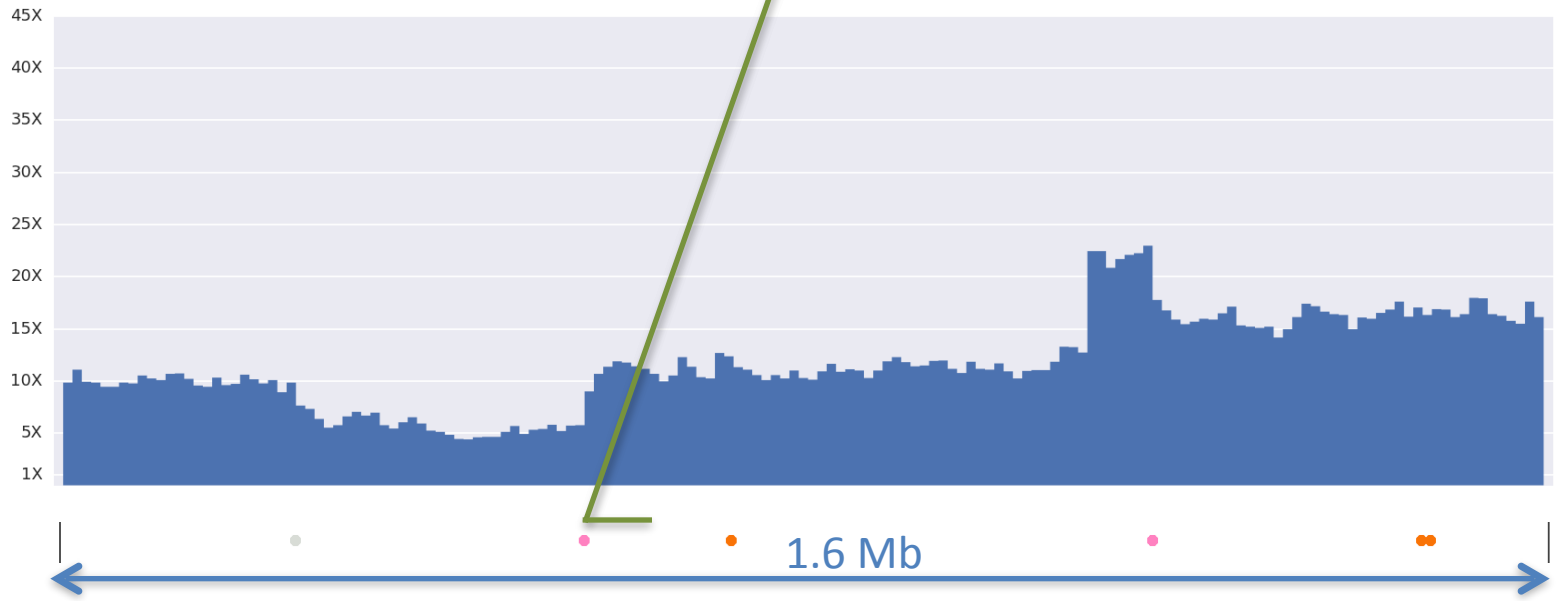


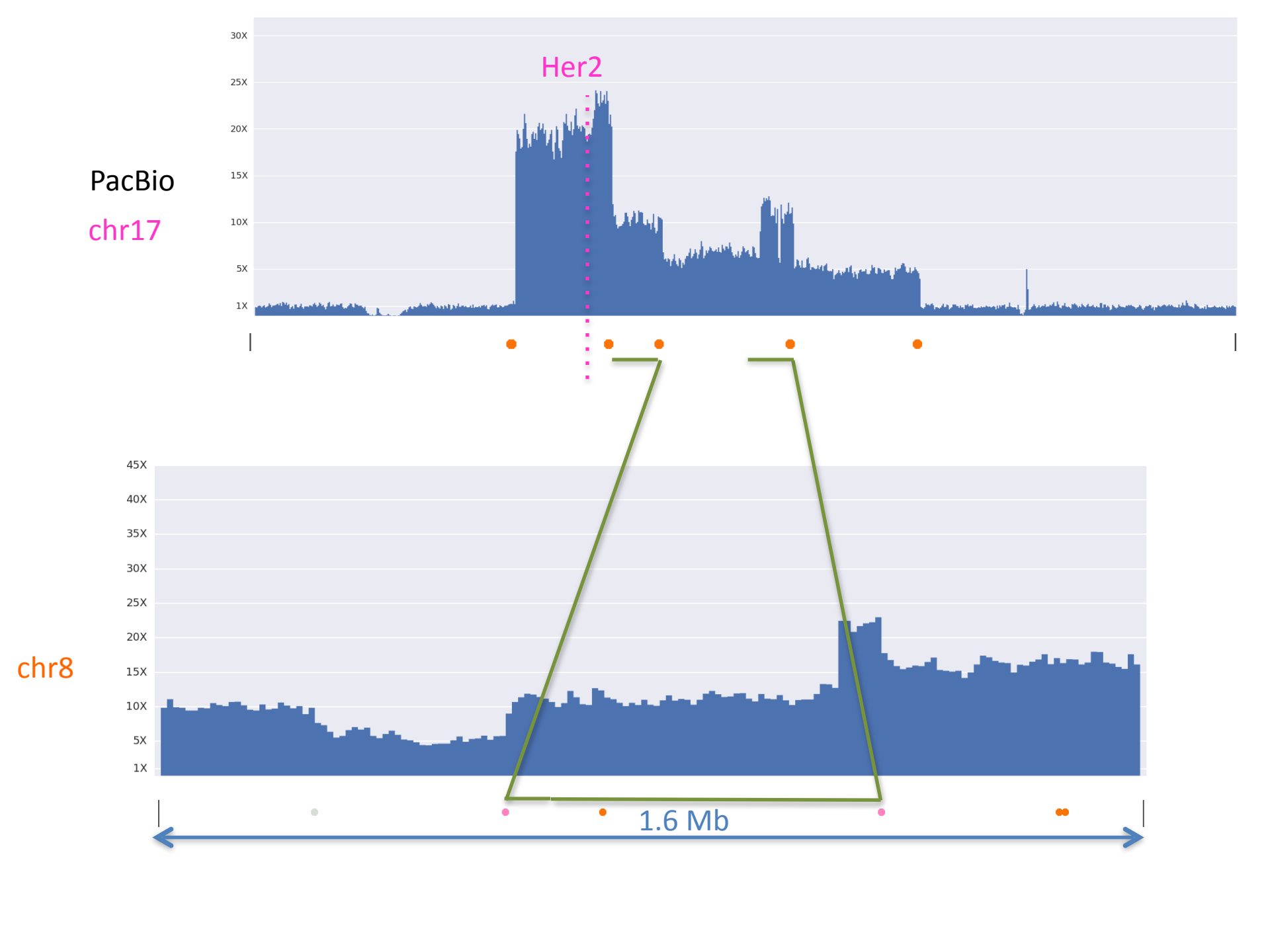
PacBio

chr17



chr8

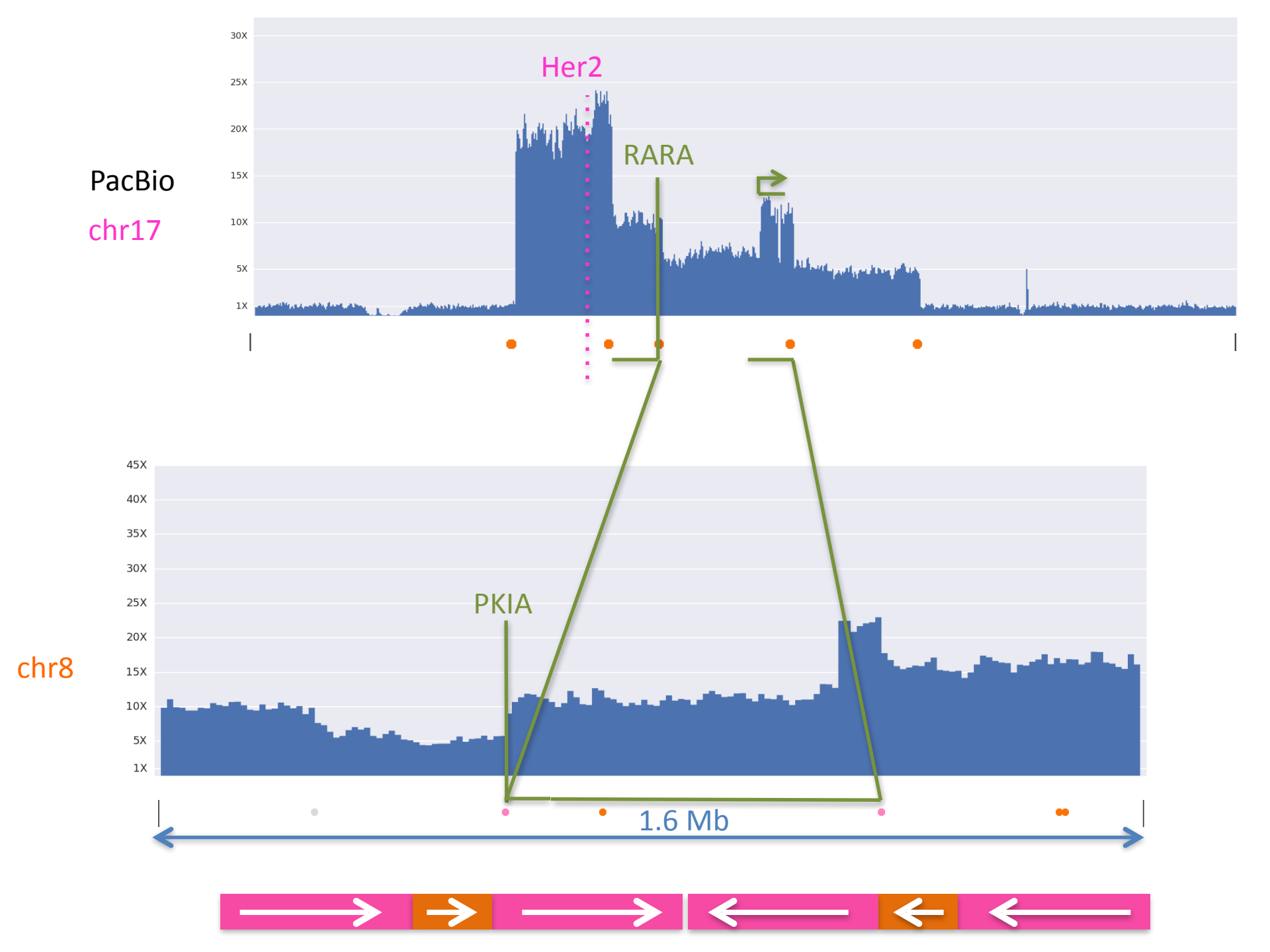



PacBio
chr17
30X
25X
20X
15X
10X
5X
1X
Her2
chr8
45X
40X
35X
30X
25X
20X
15X
10X
5X
1X
1.6 Mb


PacBio
chr17
30X
25X
20X
15X
10X
5X
1X
Her2
chr8
45X
40X
35X
30X
25X
20X
15X
10X
5X
1X
1.6 Mb


PacBio
chr17
30X
25X
20X
15X
10X
5X
1X
Her2
chr8
45X
40X
35X
30X
25X
20X
15X
10X
5X
1X
1.6 Mb

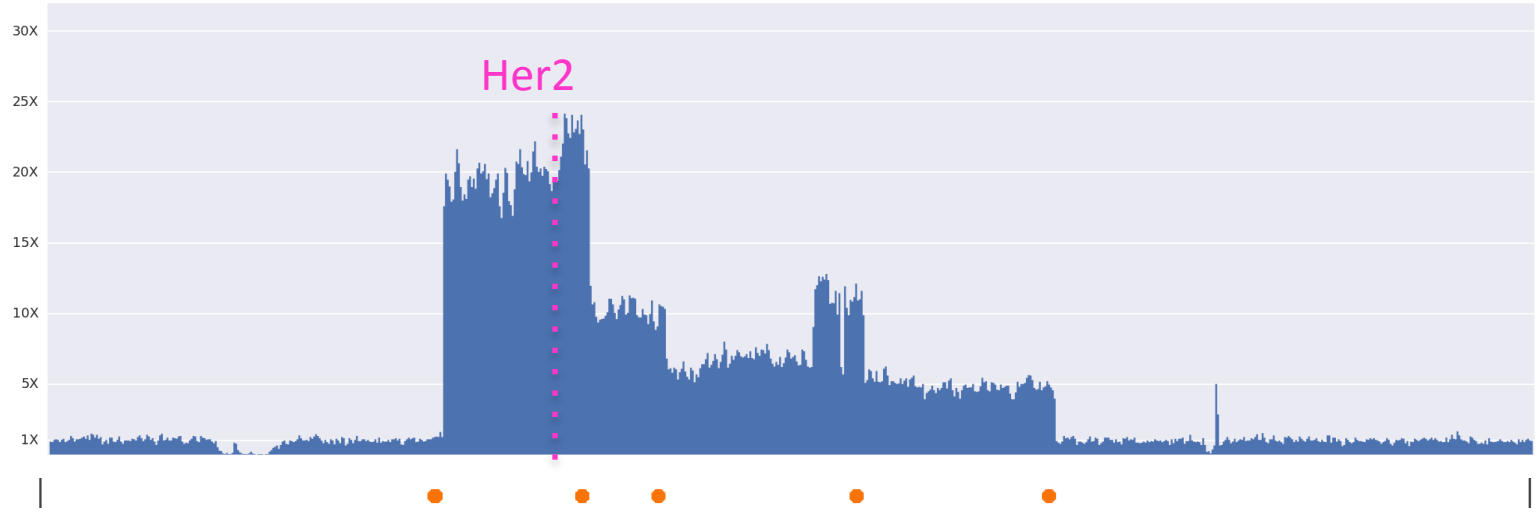

PacBio
chr17
30X
25X
20X
15X
10X
5X
1X
Her2
chr8
45X
40X
35X
30X
25X
20X
15X
10X
5X
1X
1.6 Mb


PacBio
chr17
30X
25X
20X
15X
10X
5X
1X
Her2
RARA
chr8
45X
40X
35X
30X
25X
20X
15X
10X
5X
1X
PKIA
1.6 Mb

Cancer lesion reconstruction from genomic threads

PacBio

chr17

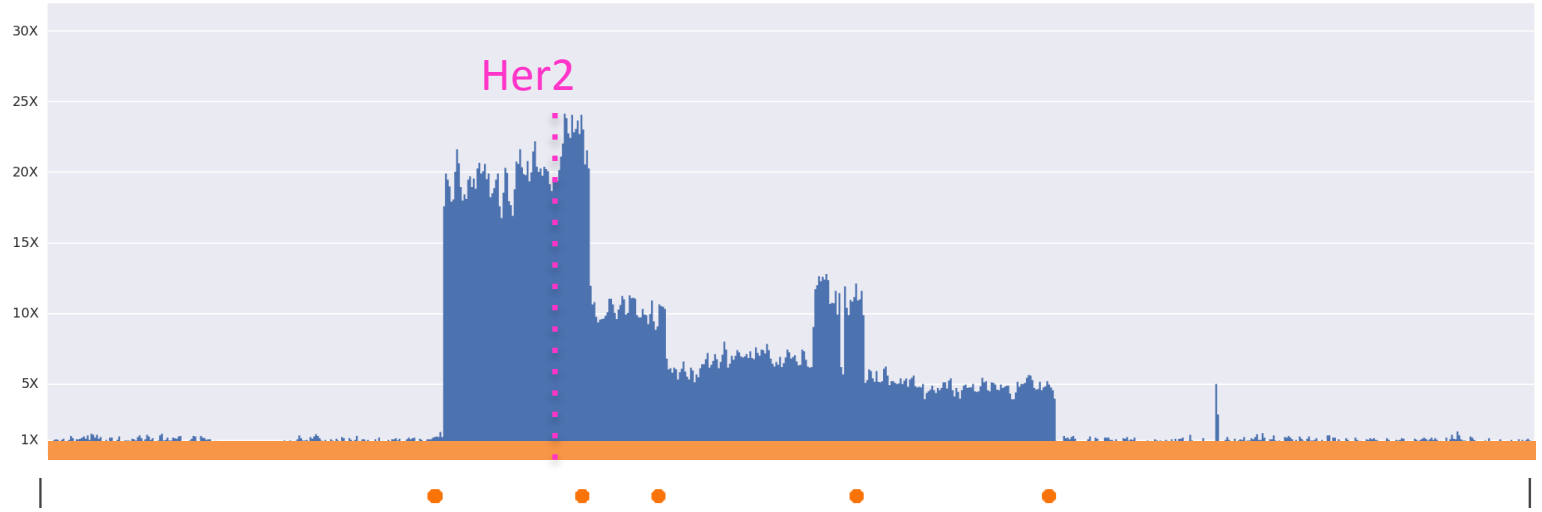


By comparing the proportion of reads that are spanning or split at breakpoints we can begin to infer the history of the genetic lesions.

Cancer lesion reconstruction from genomic threads

PacBio

chr17



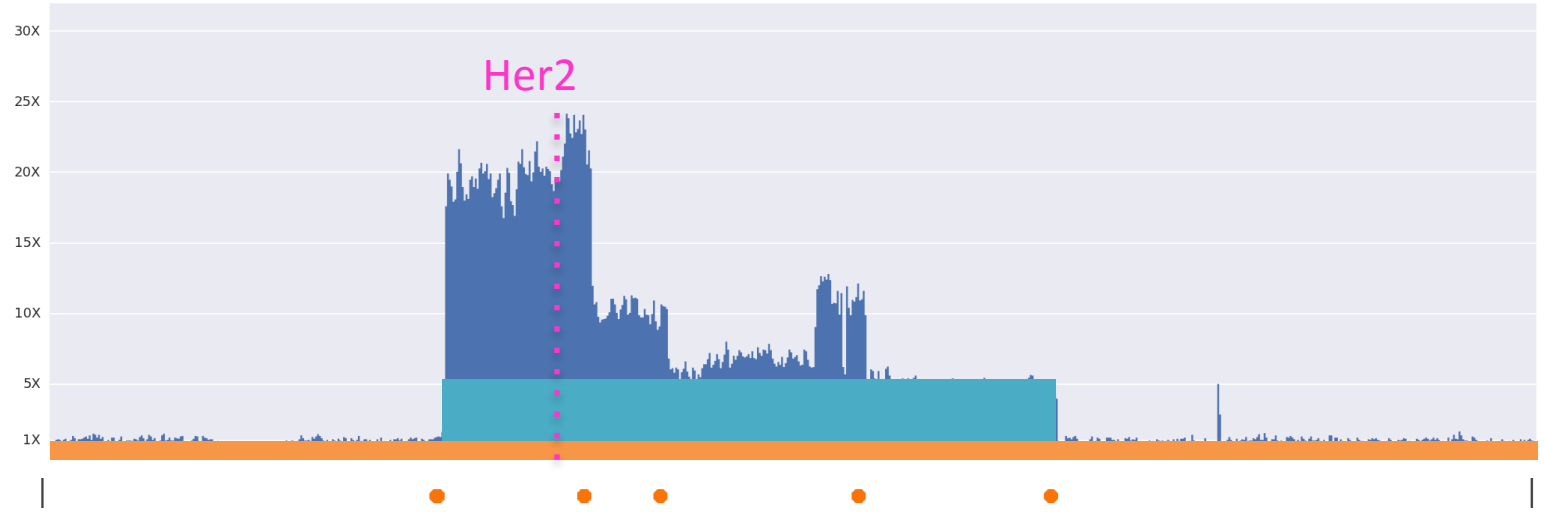
By comparing the proportion of reads that are spanning or split at breakpoints we can begin to infer the history of the genetic lesions.

1. Healthy diploid genome

Cancer lesion reconstruction from genomic threads

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.

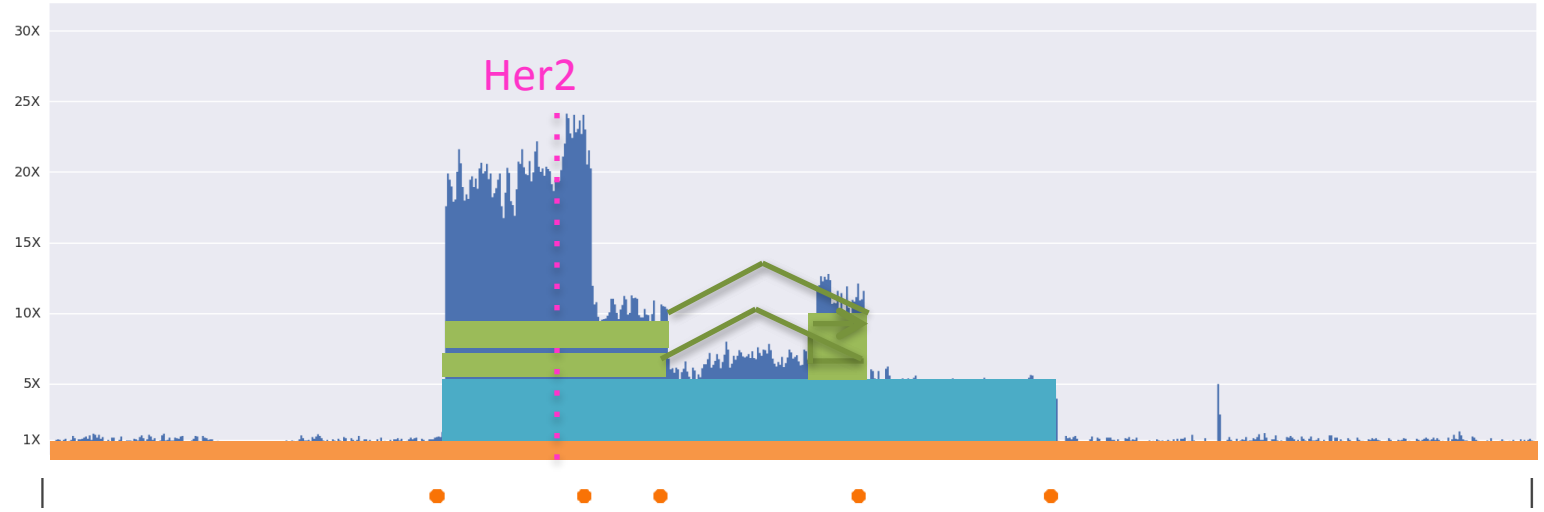
1. Healthy diploid genome

2. Original translocation into chromosome 8

Cancer lesion reconstruction from genomic threads

PacBio

chr17



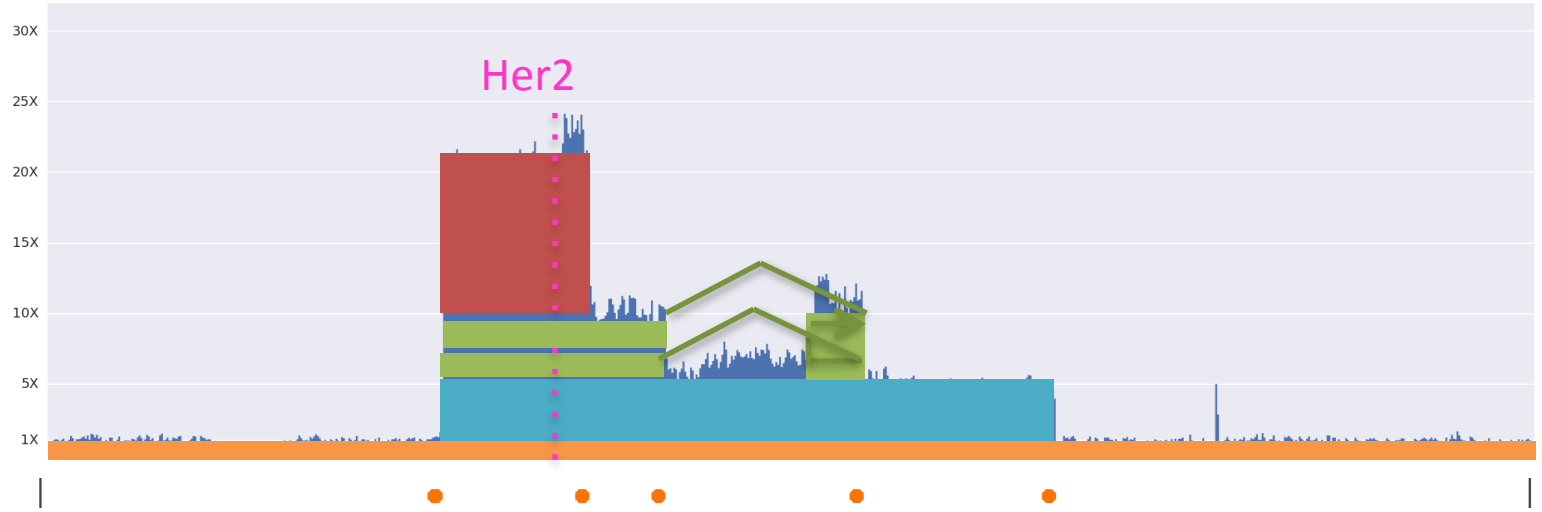
By comparing the proportion of reads that are spanning or split at breakpoints we can begin to infer the history of the genetic lesions.

1. Healthy diploid genome
2. Original translocation into chromosome 8
3. Duplication, inversion, and inverted duplication within chromosome 8

Cancer lesion reconstruction from genomic threads

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

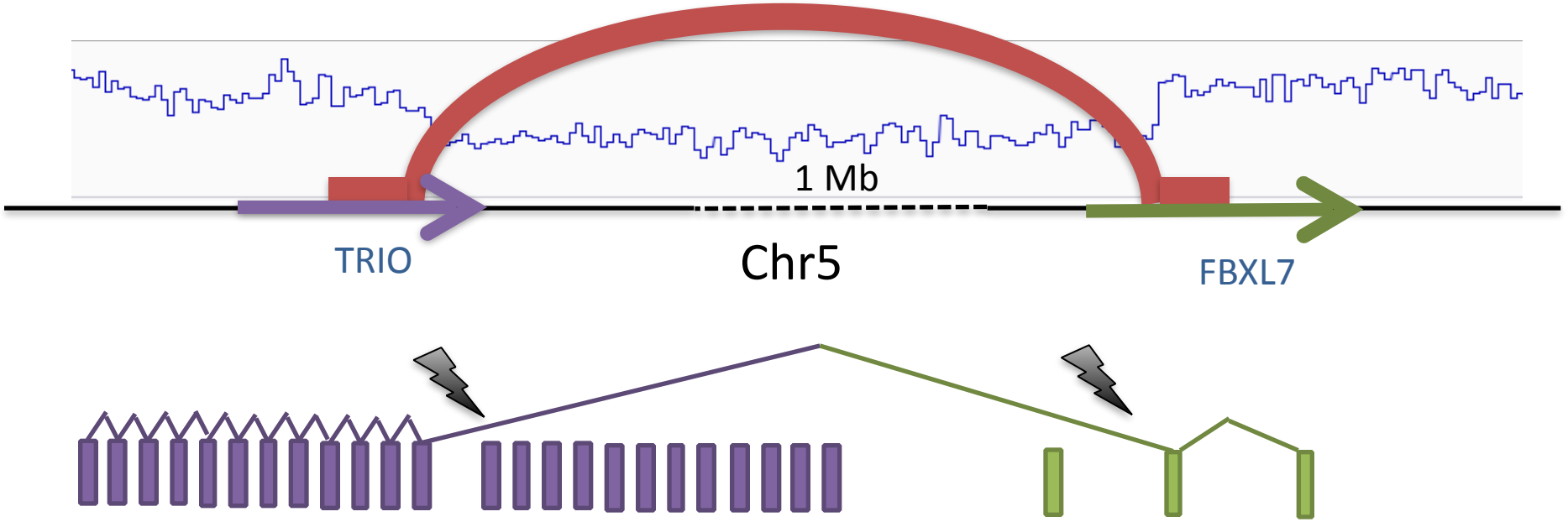
Combined genome and transcriptome analysis

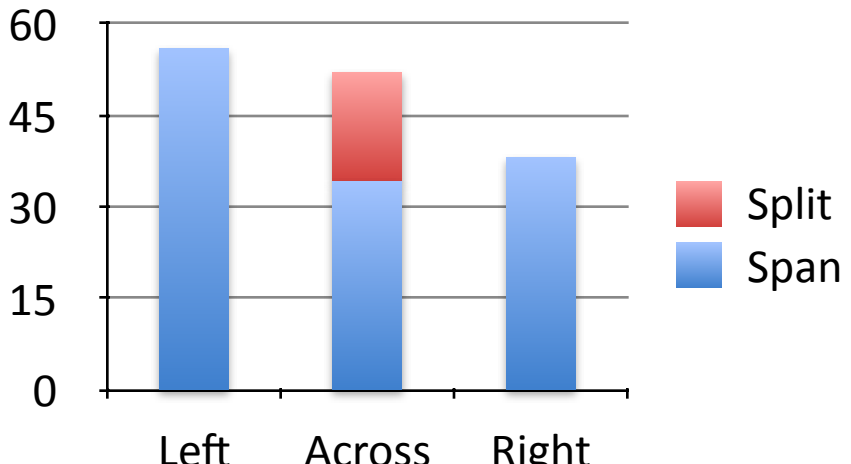
- 143,532 distinct isoforms
  - 18,186 overlapping groups
- 7 of 9 known gene fusions represented

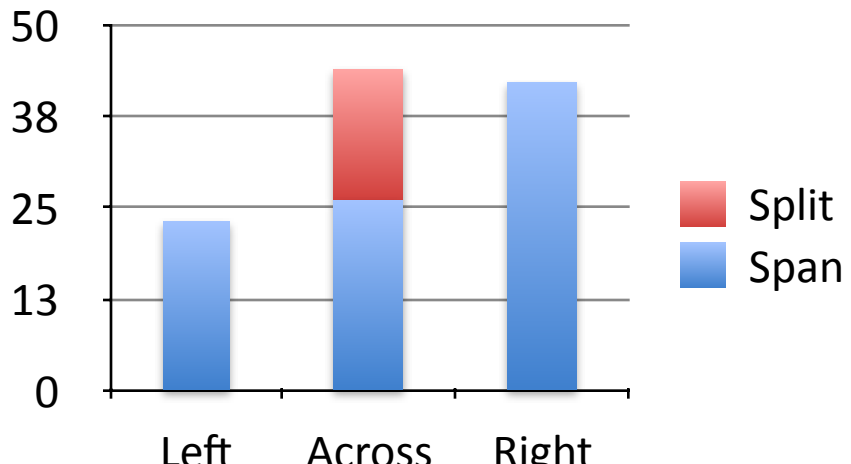
Known Gene fusions		Confirmed by PacBio DNA?	Confirmed by PacBio Iso-Seq
TATDN1	*GSDMB*	*Yes*	*Yes*
RARA	*PKIA*	*Yes*	*Yes*
ANKHD1	*PCDH1*	*Yes*	*No*
CCDC85C	*SETD3*	*Yes*	*No*
SUMF1	*LRRFIP2*	*Yes*	*Yes*
WDR67 (TBC1D31)	*ZNF704*	*Yes*	*Yes*
DHX35	*ITCH*	*Yes*	*Yes*
NFS1	*PREX1*	*Yes *if allowing for 3 translocations*	*Yes*
CYTH1	*EIF3H*	*Yes *if allowing for 2 translocations*	*Yes*

TRIO-FBXL7

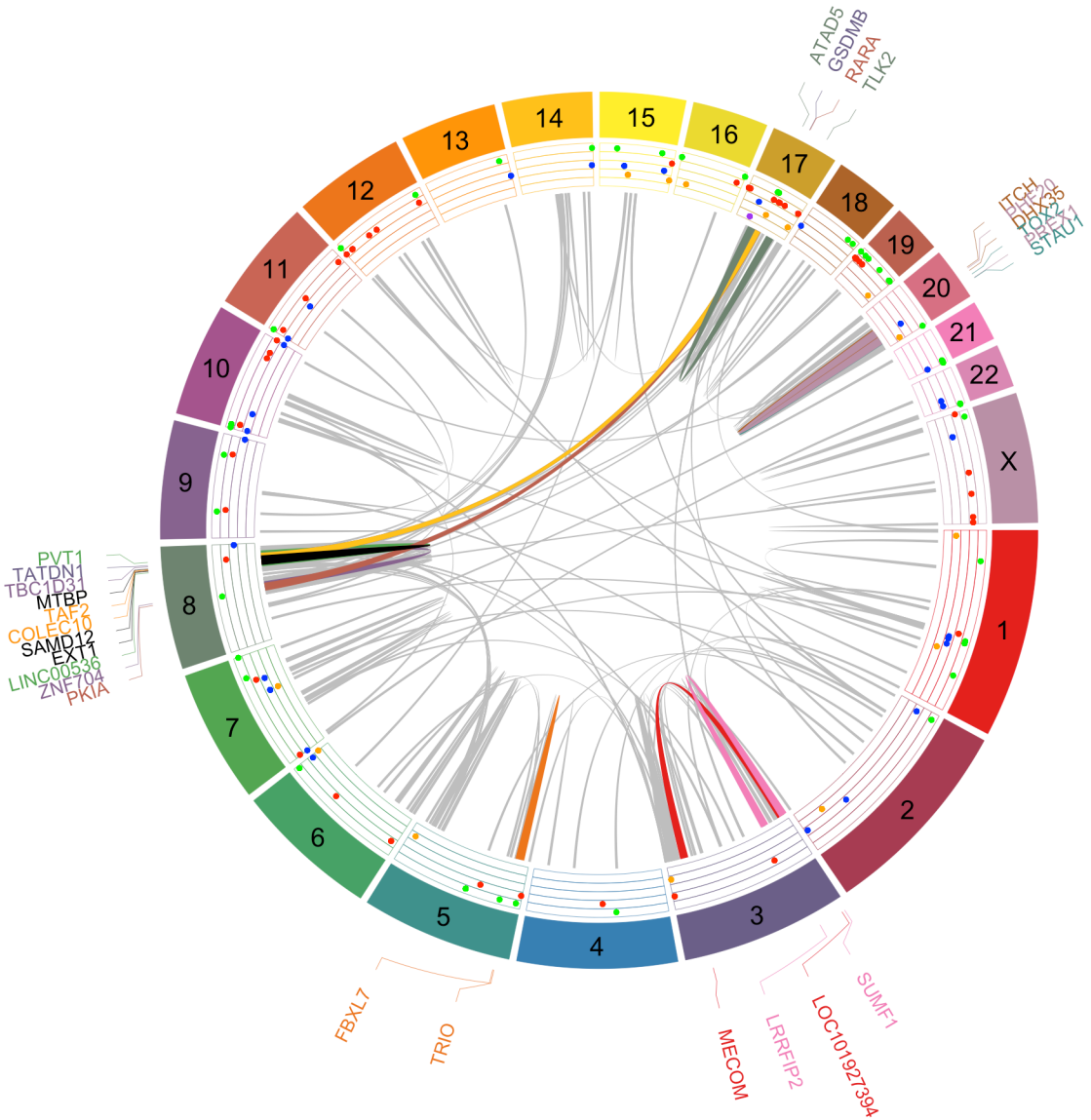
18 split DNA reads + PCR validation







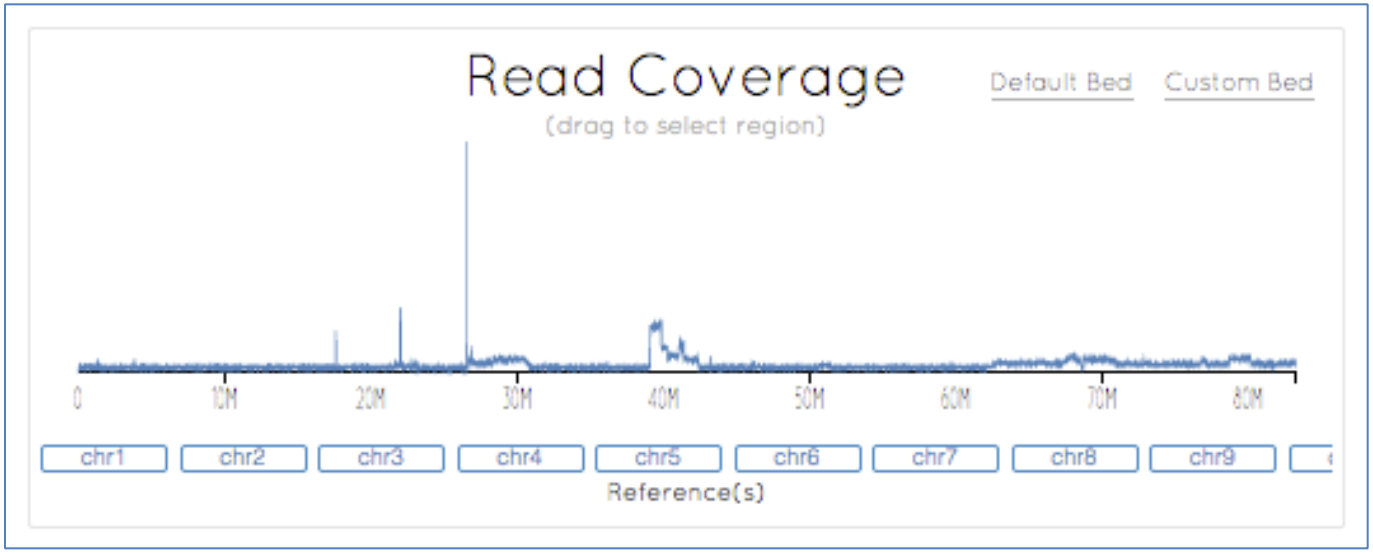
Gene fusions and gene interruptions



- Dots from the outside: insertions, deletions , tandem expansions, tandem contractions, inversions (<10kb) that intersect exons of genes.
- Connecting lines: translocations (>10kb distance) that intersect at least one gene
- Colored lines and named genes: Fusions indicated by both Iso-seq and genomic split read evidence.

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

Conclusions

High coverage, long read sequencing may yield dramatically different views of complex genomes

This may be seen through both reassembly and de novo assembly

Long read transcript analysis may enhance this as well

The next challenges are:

- **Making this process more efficient**
- **Understanding the biological implications of the newly discovered genetic variation**



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





Elizabeth Tseng
Jason Chin

LUMPY technical support: Ryan Layer and Aaron Quinlan

Thank you!

May the force be with you!

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

Supplemental slides

SKBR3 Oncogene Analysis

Known missense mutation in p53: **R175H**



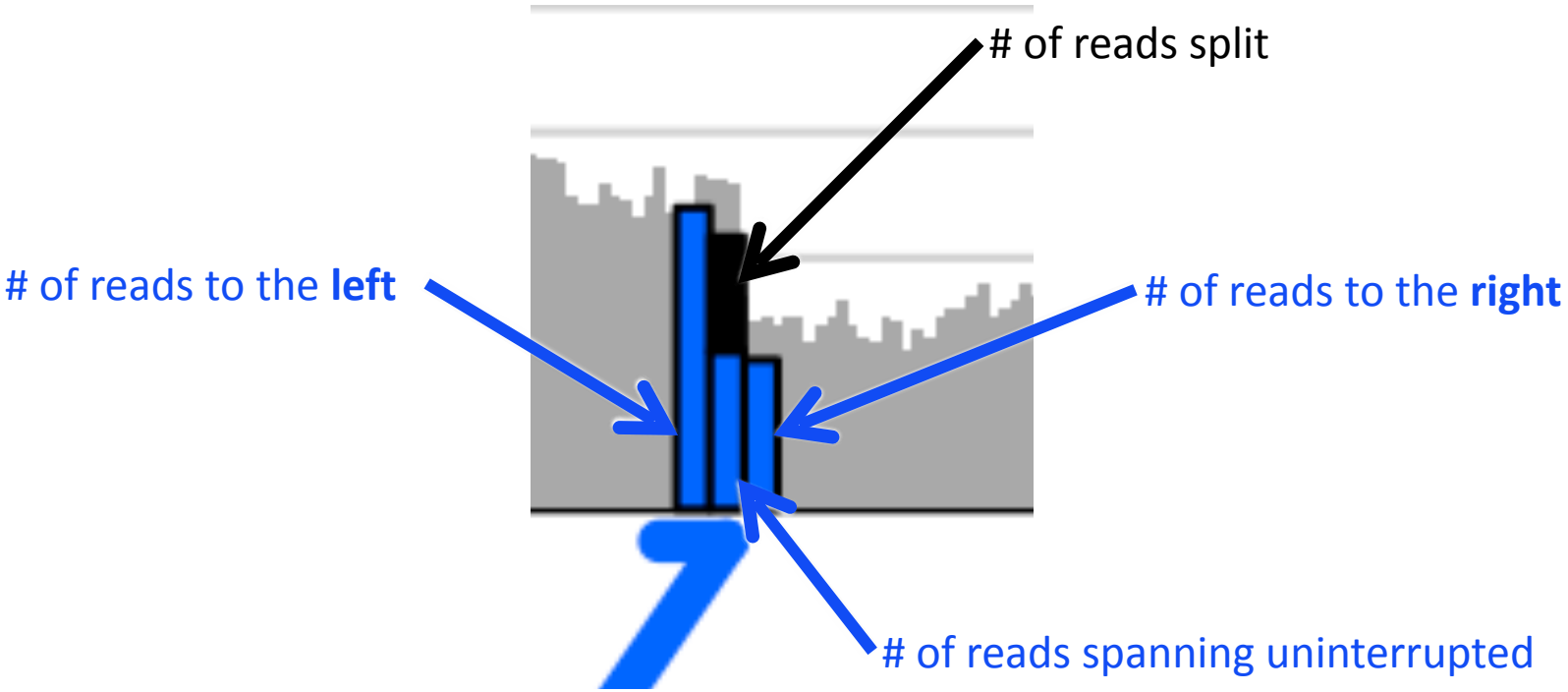
Oncogene amplifications	
ErbB2	*≈20X*
MYC	*≈27X*
MET	*≈8X*

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 *nested inside 3 translocations*
CYTH1	*EIF3H*	*Yes *nested inside 2 translocations*

Using joint coverage and breakpoint analysis to discover underlying events

Coverage + Structural variants = Genomic threads



Using joint coverage and breakpoint analysis to discover underlying events

