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

W. Richard McCombie

Disclosures

CSH

article

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[2], S. Cheng[2], J. Gusella[3], A.-M. Frischauf[4], A. Poustka[4], 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?

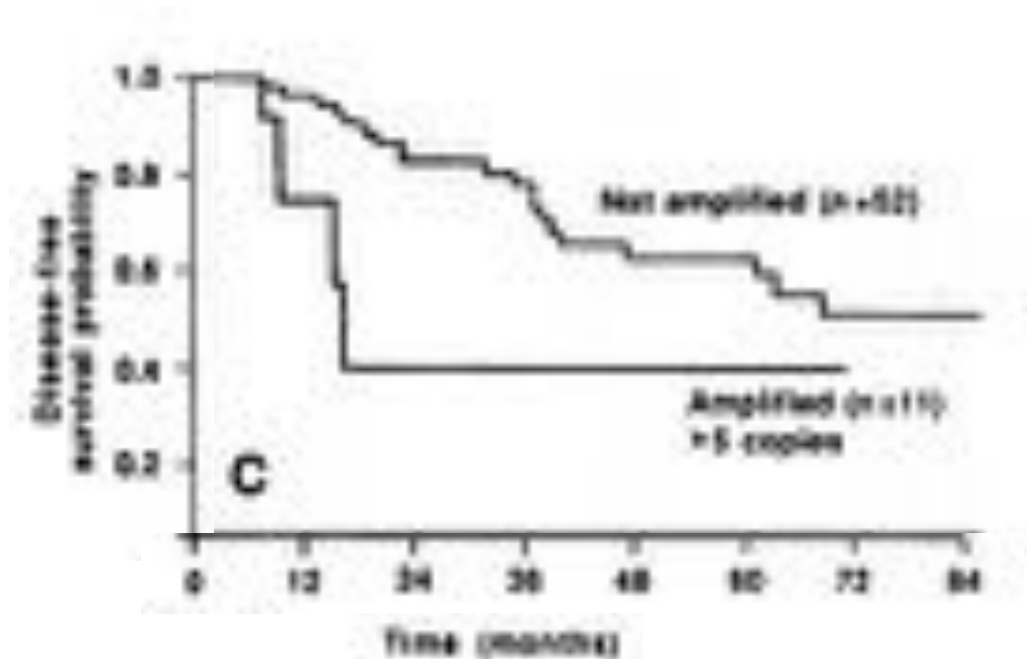# Her2 amplified breast cancer

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.

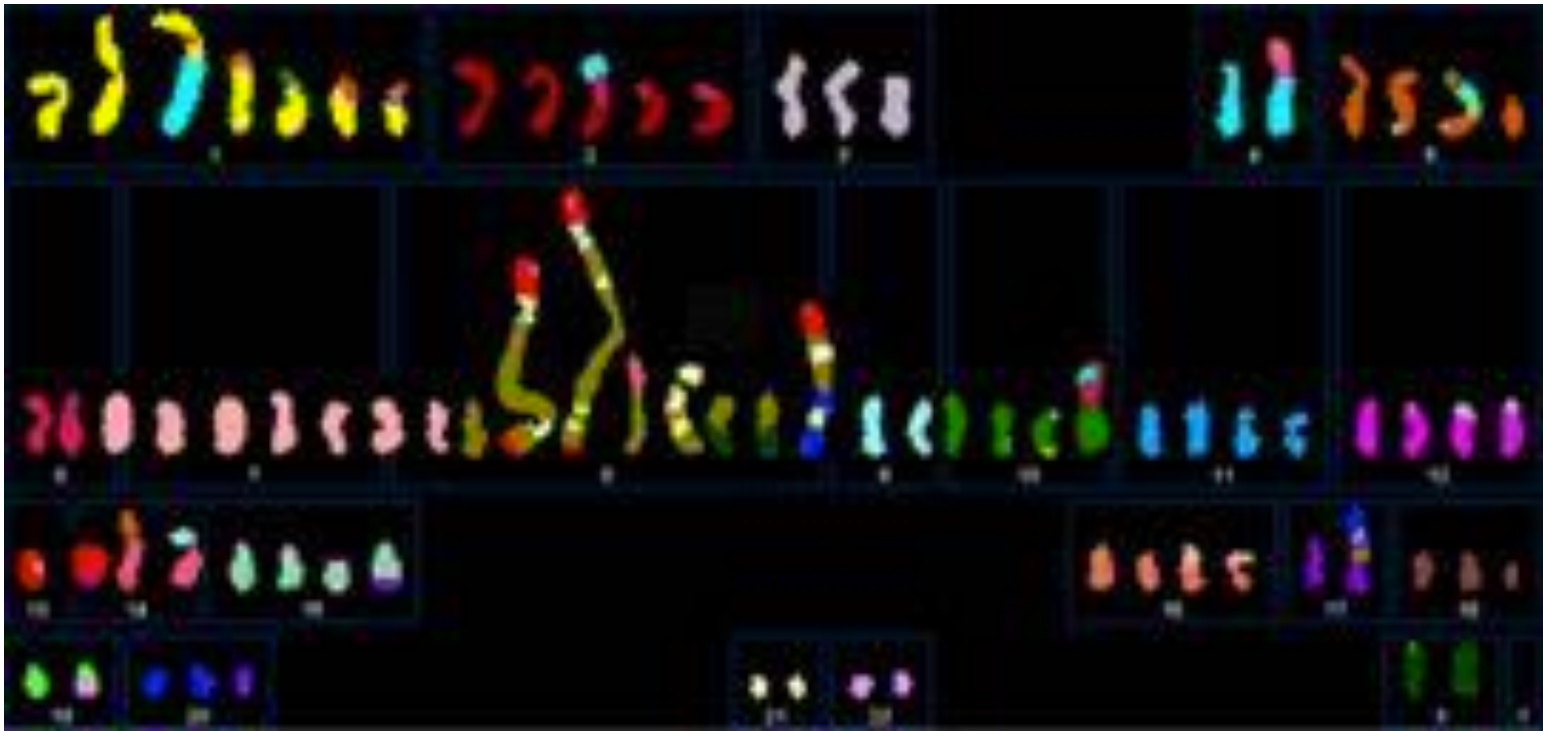## Her2 amplified breast cancer

- 20% of breast cancers

- 2-3X recurrence risk

- 5X metastasis risk

(Adapted from Slamon et al, 1987)

SK-BR-3

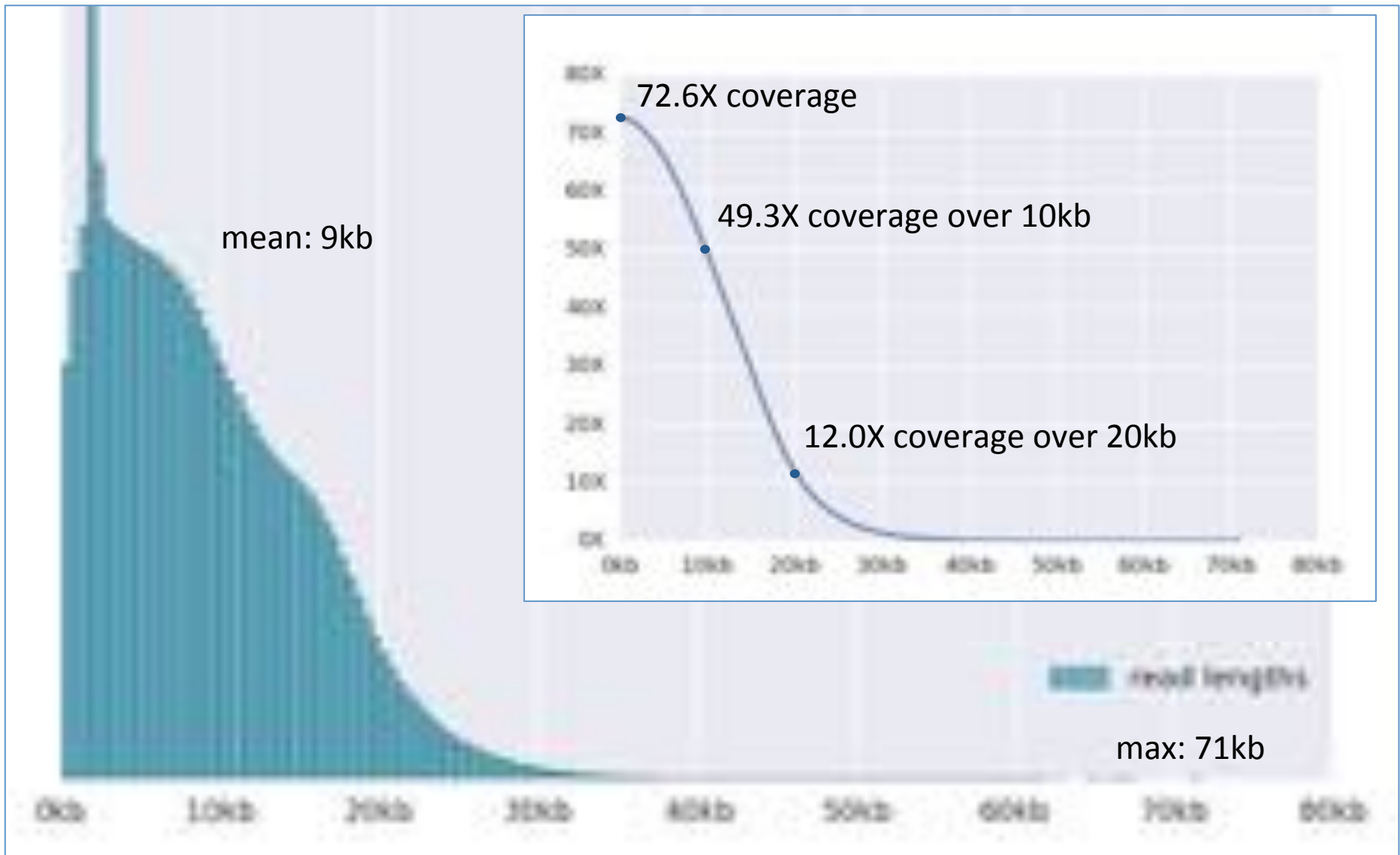Most commonly used Her2-amplified breast cancer cell line

Often used for pre-clinical research on Her2-targeting therapeutics such as Herceptin (Trastuzumab) and resistance to these therapies.
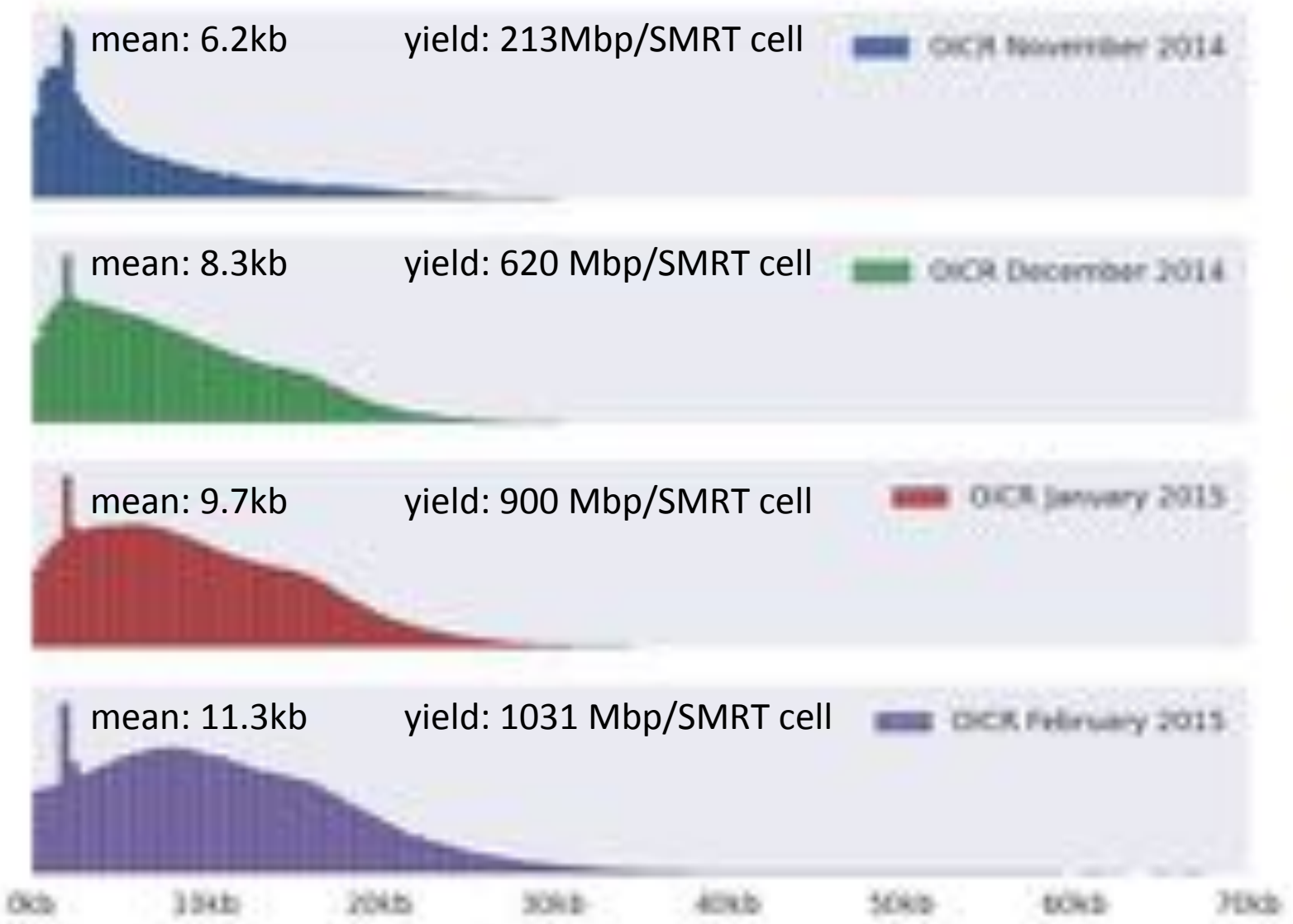
(Davidson et al, 2000)

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"
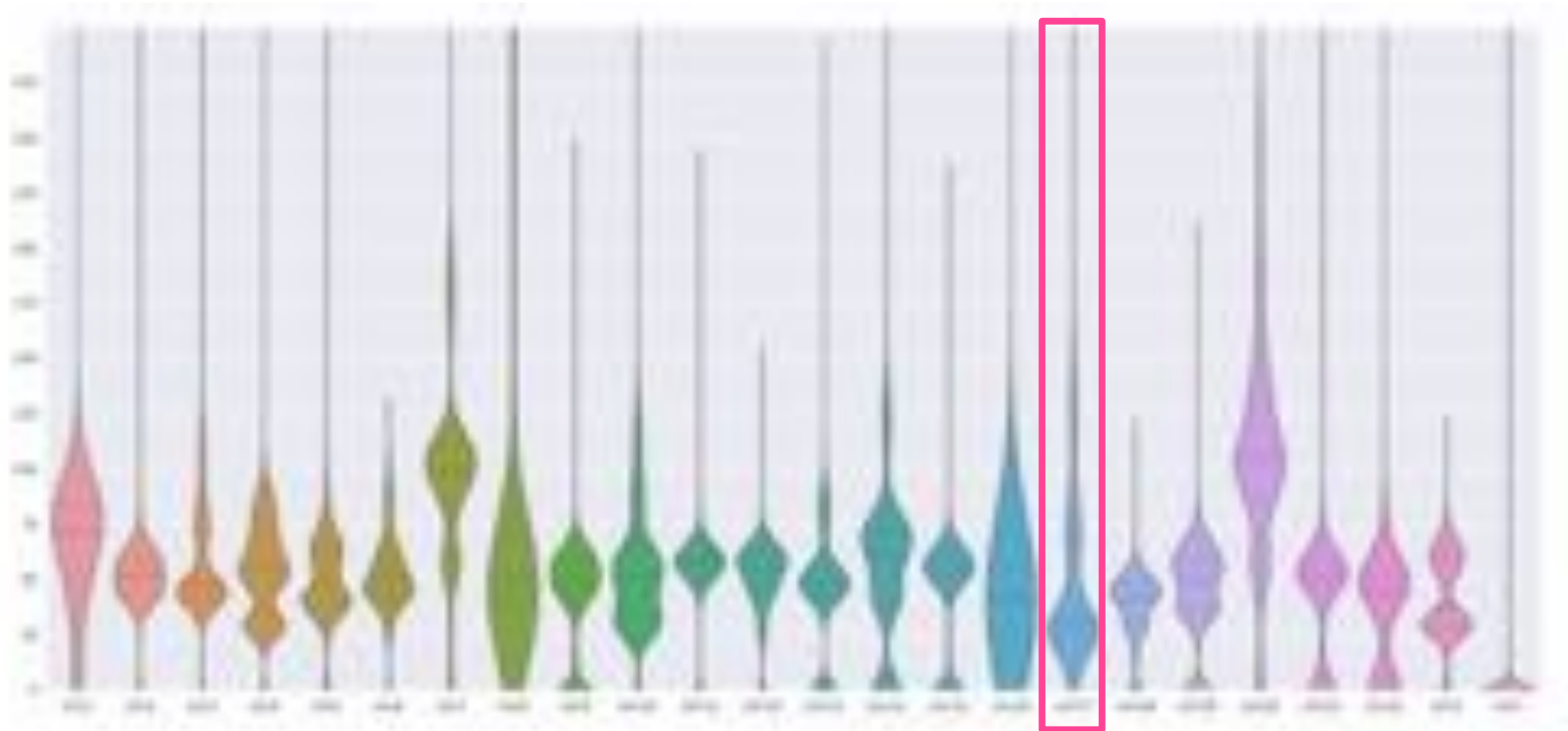- Ongoing project

PacBio read length distribution

mean: 9kb

72.6X coverage

49.3X coverage over 10kb

12.0X coverage over 20kb

read lengths

max: 71kb

Improving SMRTcell Performance

mean: 6.2kb   yield: 213Mbp/SMRT cell   OiCR November 2014

mean: 8.3kb   yield: 620 Mbp/SMRT cell   OiCR December 2014

mean: 9.7kb   yield: 900 Mbp/SMRT cell   OiCR January 2015

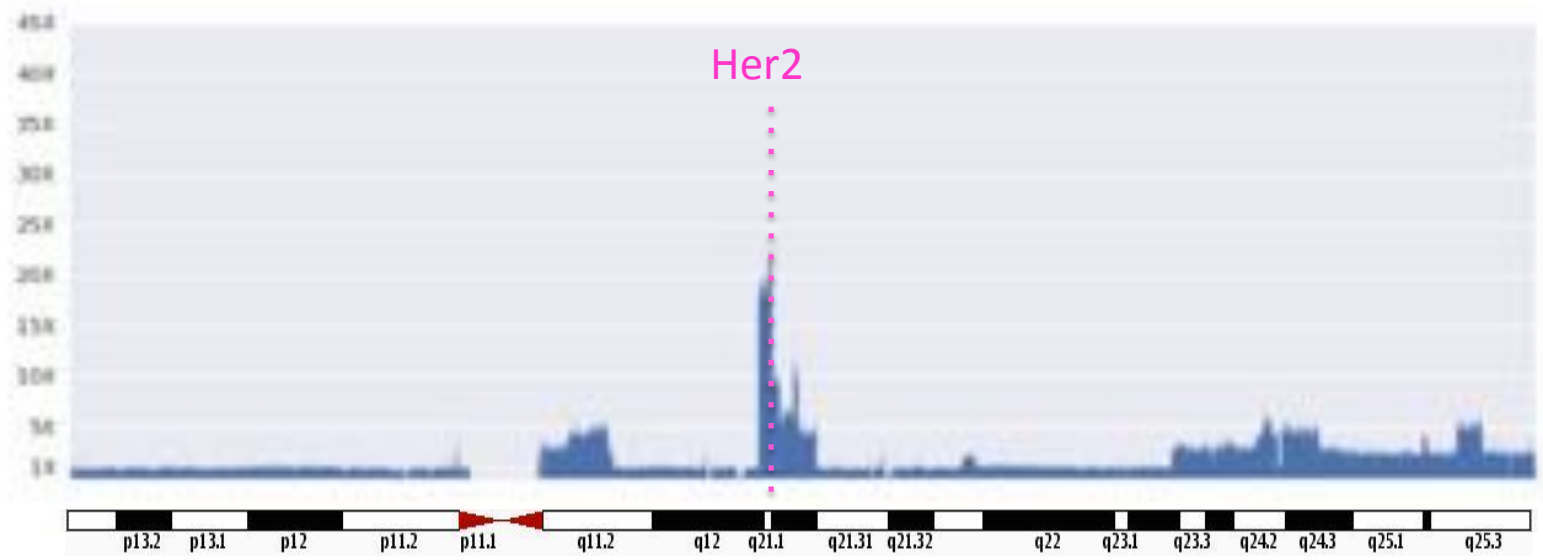mean: 11.3kb   yield: 1031 Mbp/SMRT cell   OiCR February 2015

Genome-wide alignment coverage

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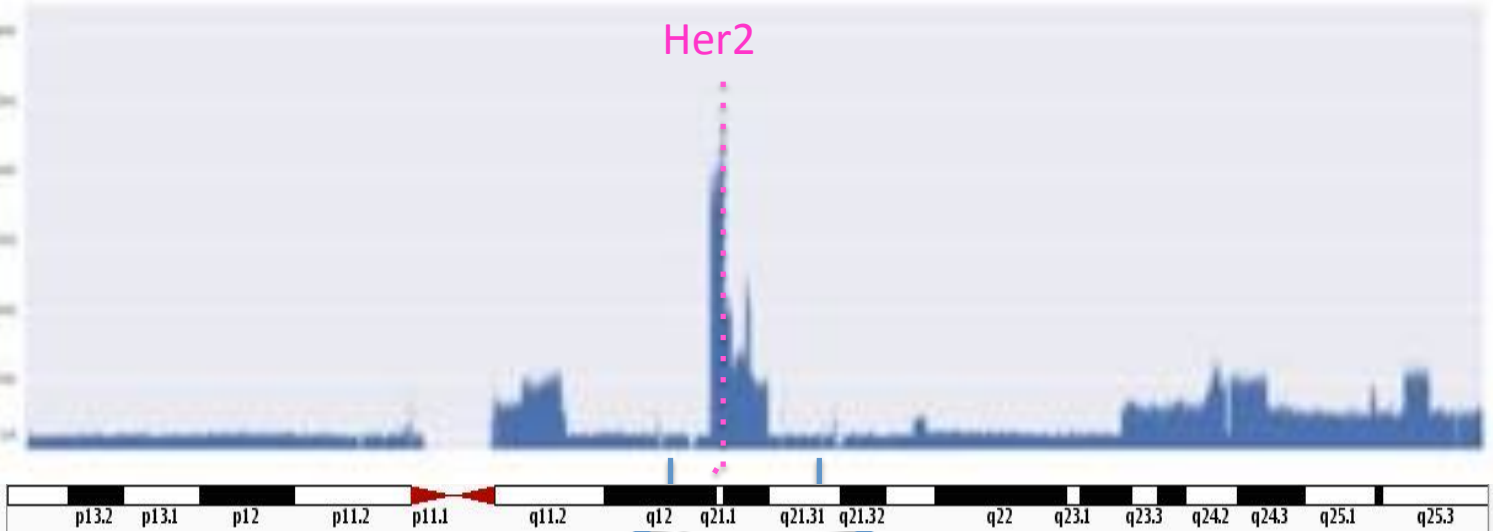Coverage per chromosome varies greatly as expected from previous karyotyping results

PacBio

Her2

Chr 17: 83 Mb

8 Mb

PacBio

PacBio

Her2

Her2

PacBio
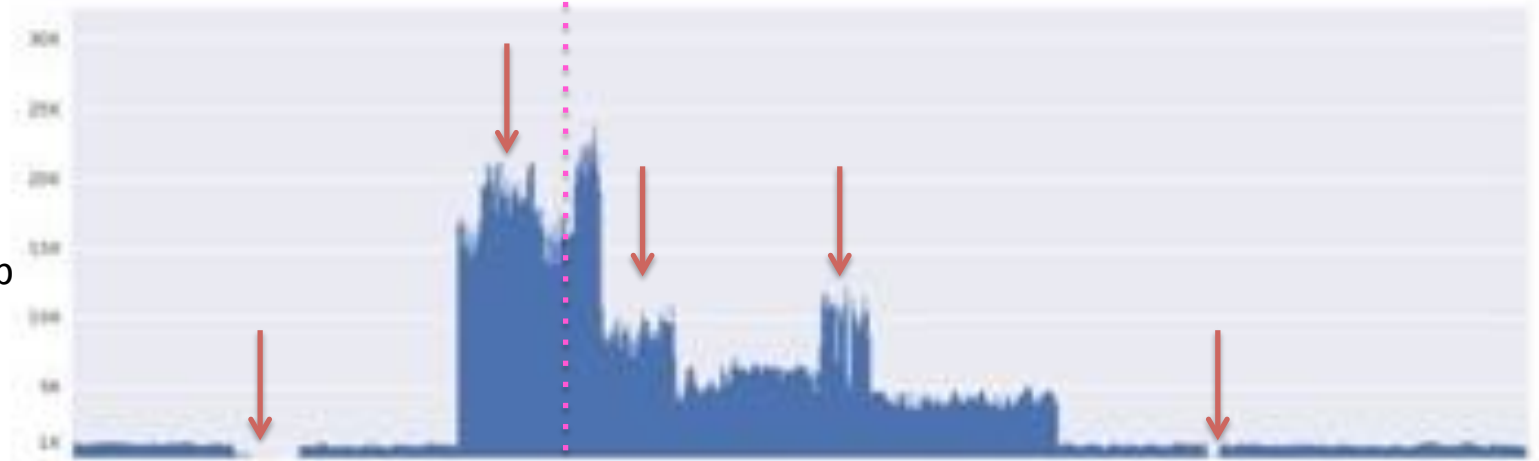67X @ 10kb

Illumina
120X @ 100bp

Her2

8 Mb

PacBio and Illumina coverage values are highly correlated
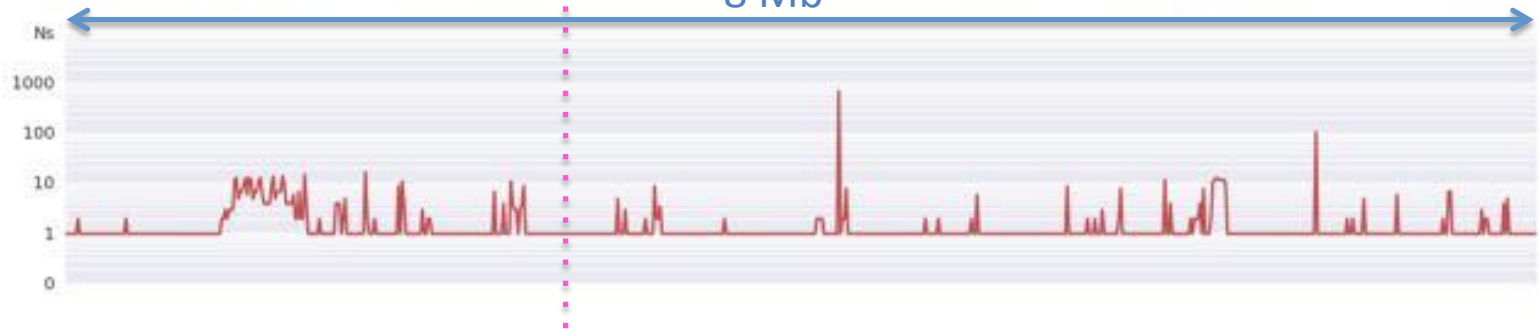but Illumina shows greater variance because of poorly mapping reads
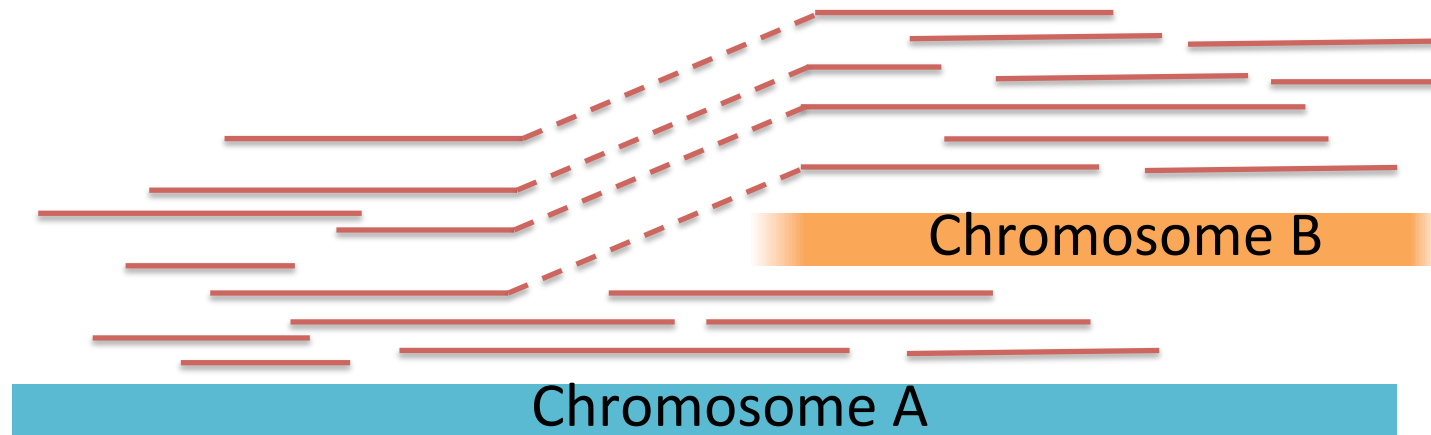
PacBio
67X @ 10kb

Illumina
120X @ 100bp

Her2

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*
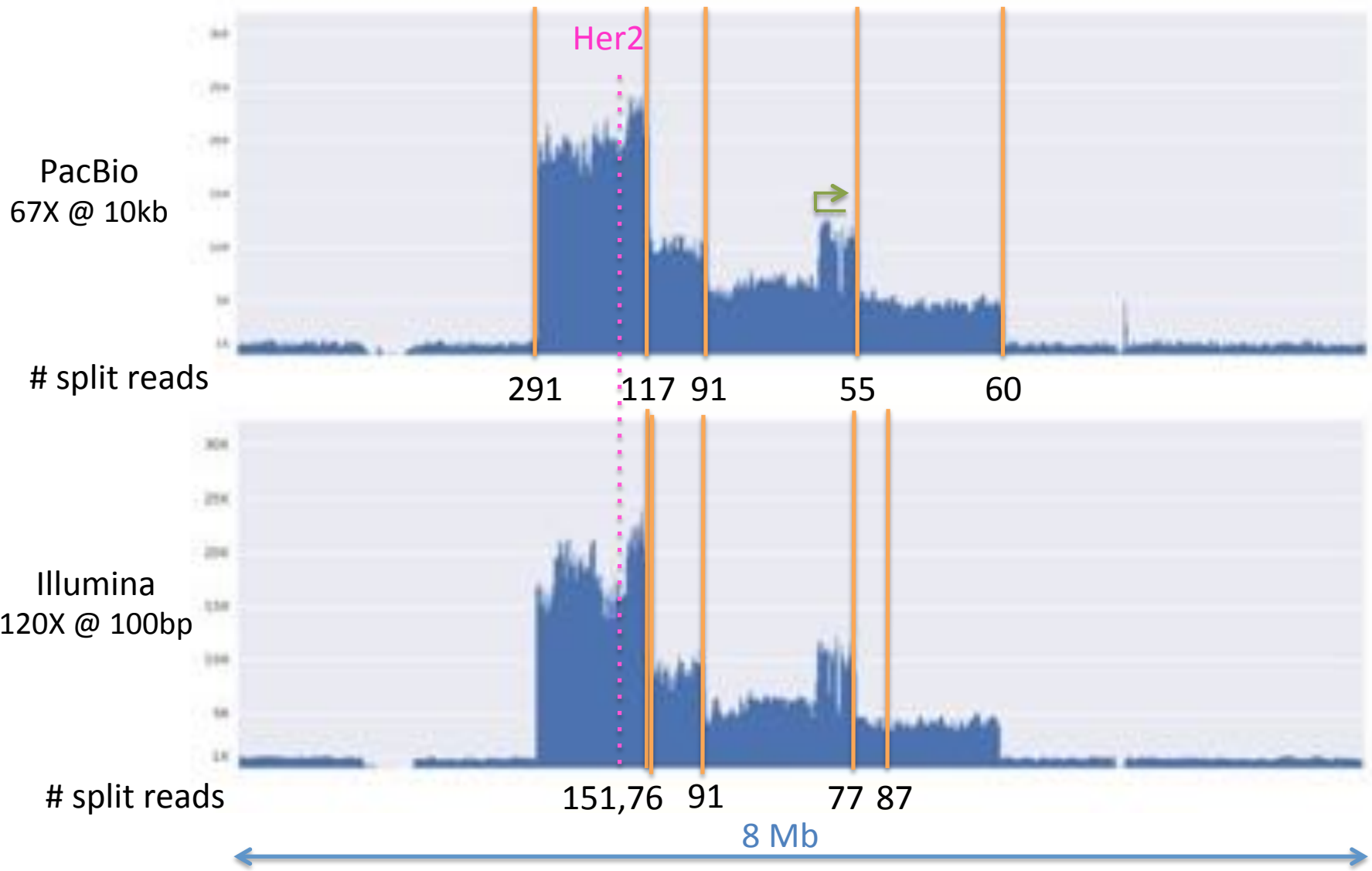   Localized HGAP + Celera Assembler + MUMmer

3. Whole genome assembly: In-depth analysis including *novel sequences*
   DNAnexus-enabled version of Falcon
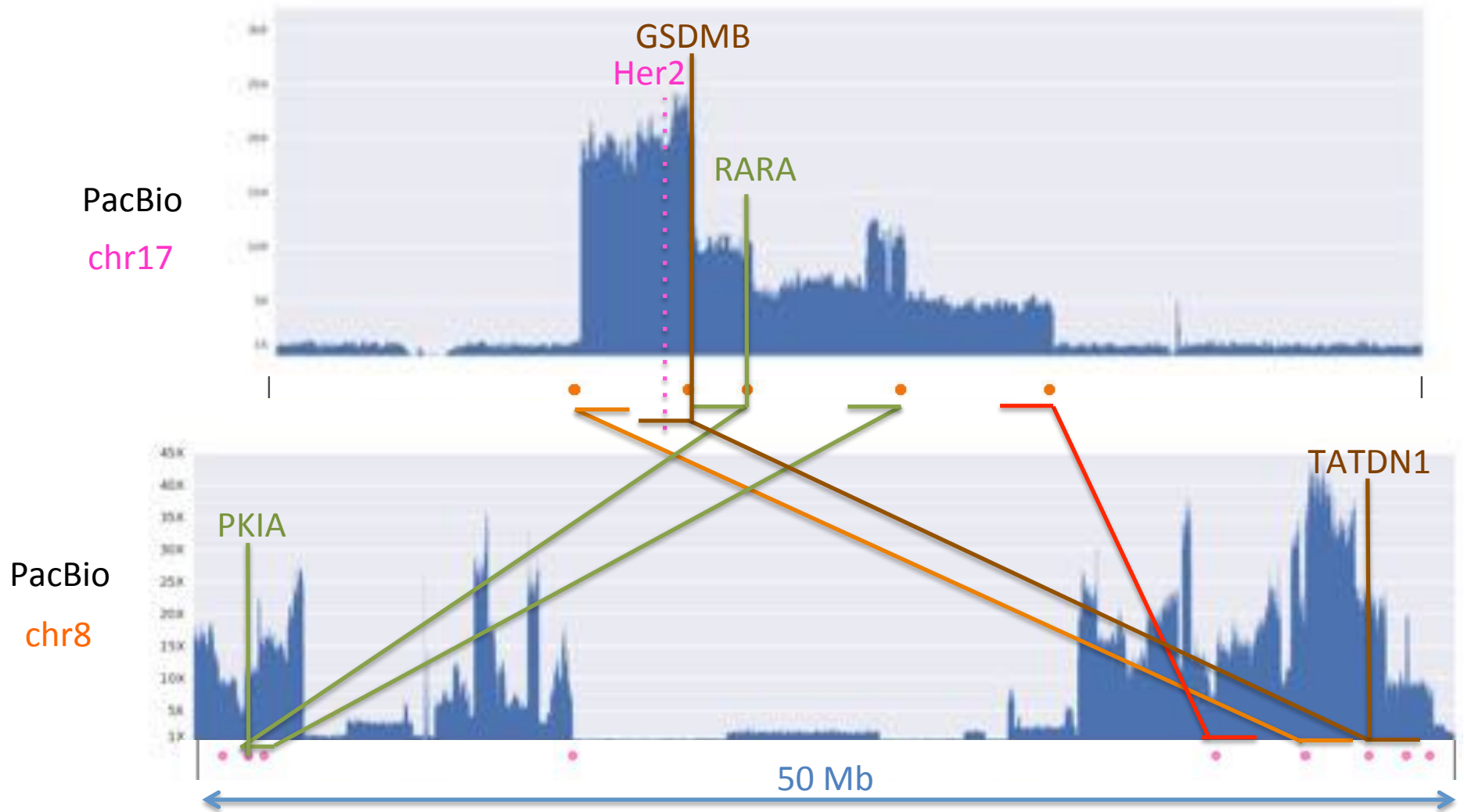
Total Assembly: 2.64Gbp   Contig N50: 2.56 Mbp   Max Contig: 23.5Mbp

PacBio
67X @ 10kb

split reads

| 291 | 117 | 91 | 55 | 60 |

Her2

Illumina
120X @ 100bp

split reads

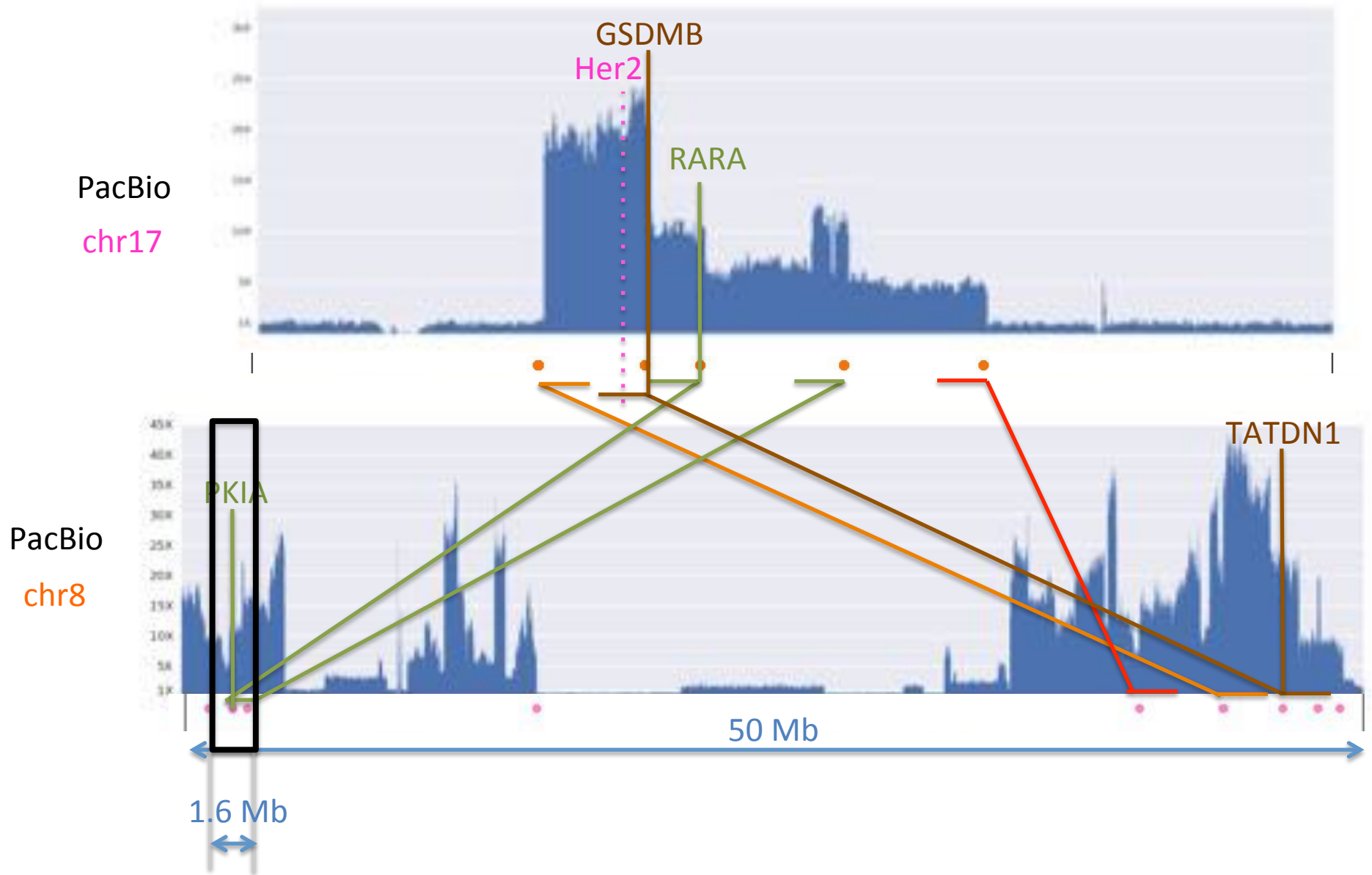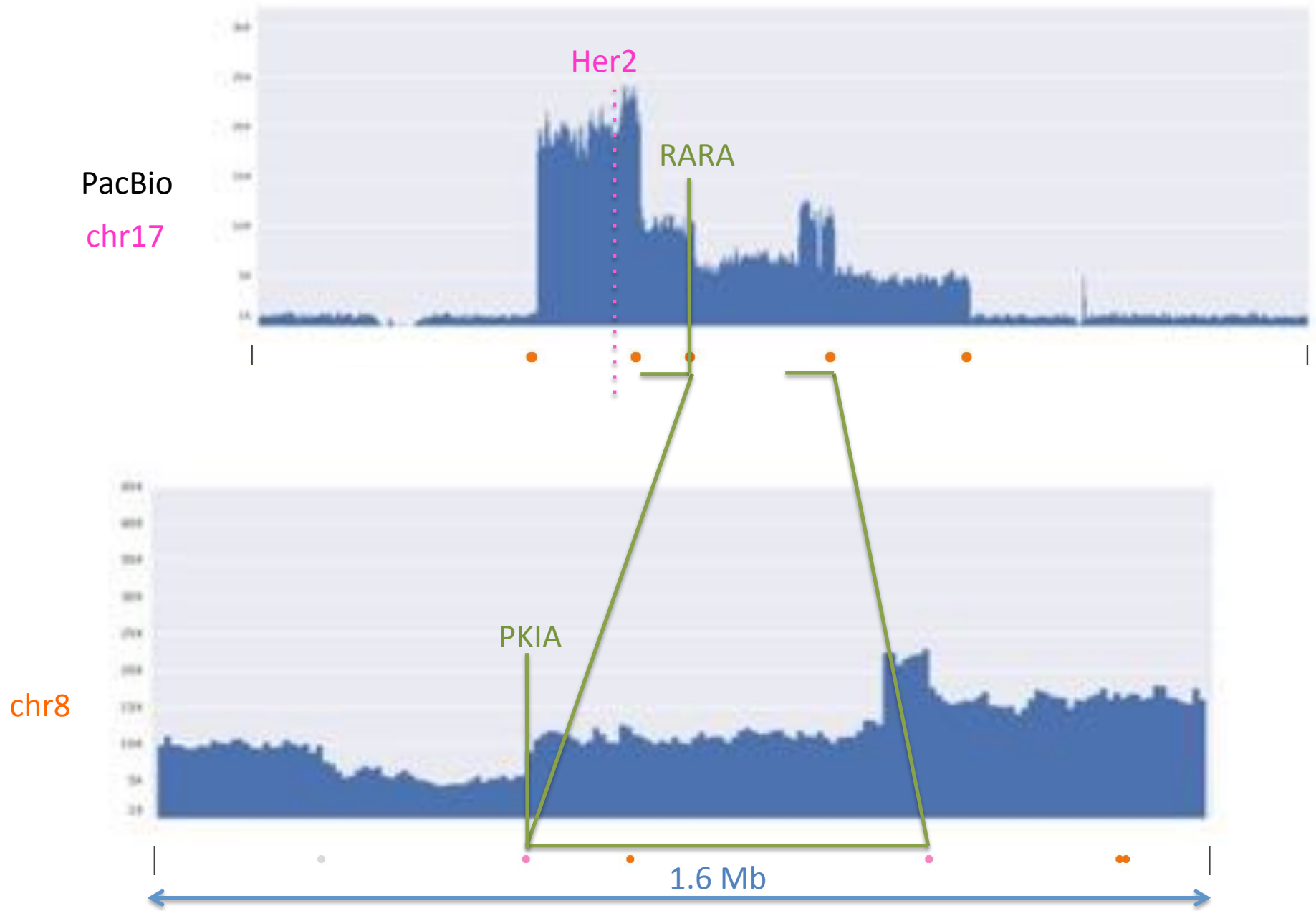| 151,76 | 91 | 77 | 87 |

8 Mb

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

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

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**

```
                                         Arg
Reference   ATCTGAGCAGCGCTCATGGTGGGGGCAGCGCCTCACAACCTCCGTCATGTGCTGTGACTGCTT
Illumina    ATCTGAGCAGCGCTCATGGTGGGGGCAGTGCCTCACAACCTCCGTCATGTGCTGTGACTGCTT
PacBio      ATCTGAGCAGCGCTCATGGTGGGGGCAGTGCCTCACAACCTCCGTCATGTGCTGTGACTGCTT
                                         His
```

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

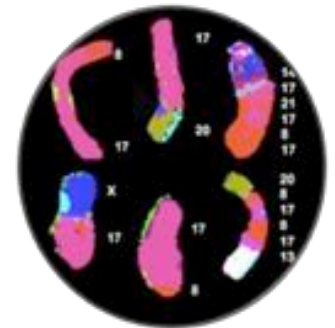**http://schatzlab.cshl.edu/data/skbr3/**
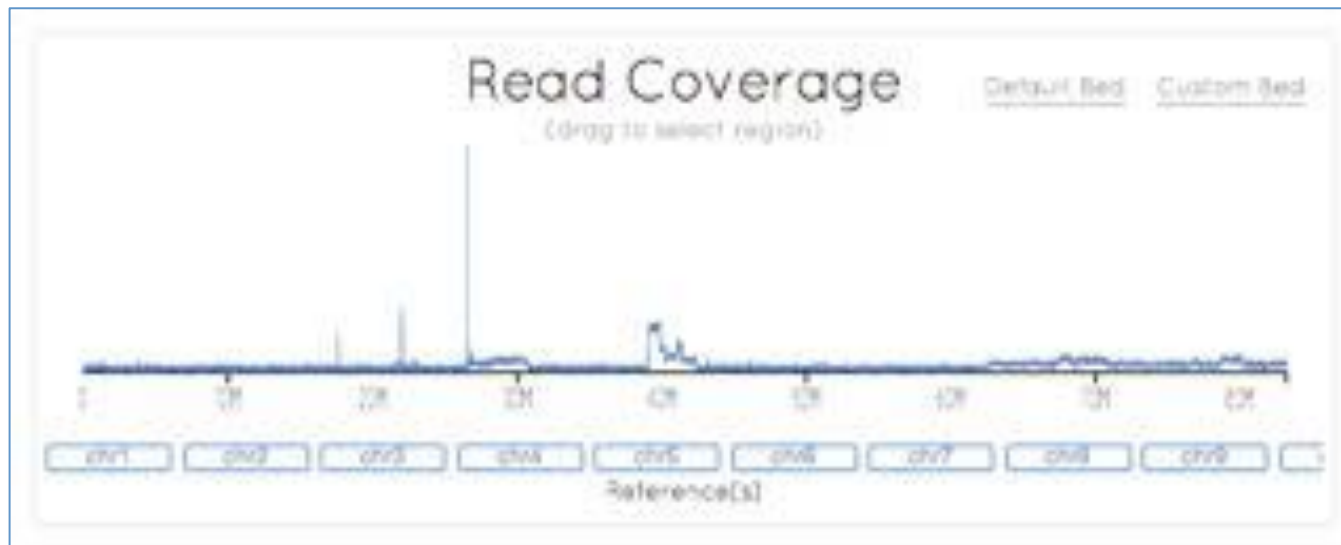
Releasing all data pre-publication to accelerate breast cancer research

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

Thank you!

May the force be with you!

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