Comprehensive Genome and Transcriptome Structural Analysis of a Breast Cancer Cell Line using PacBio Long Read Sequencing

Maria Nattestad

Schatz + McCombie + Hicks at Cold Spring Harbor Laboratory

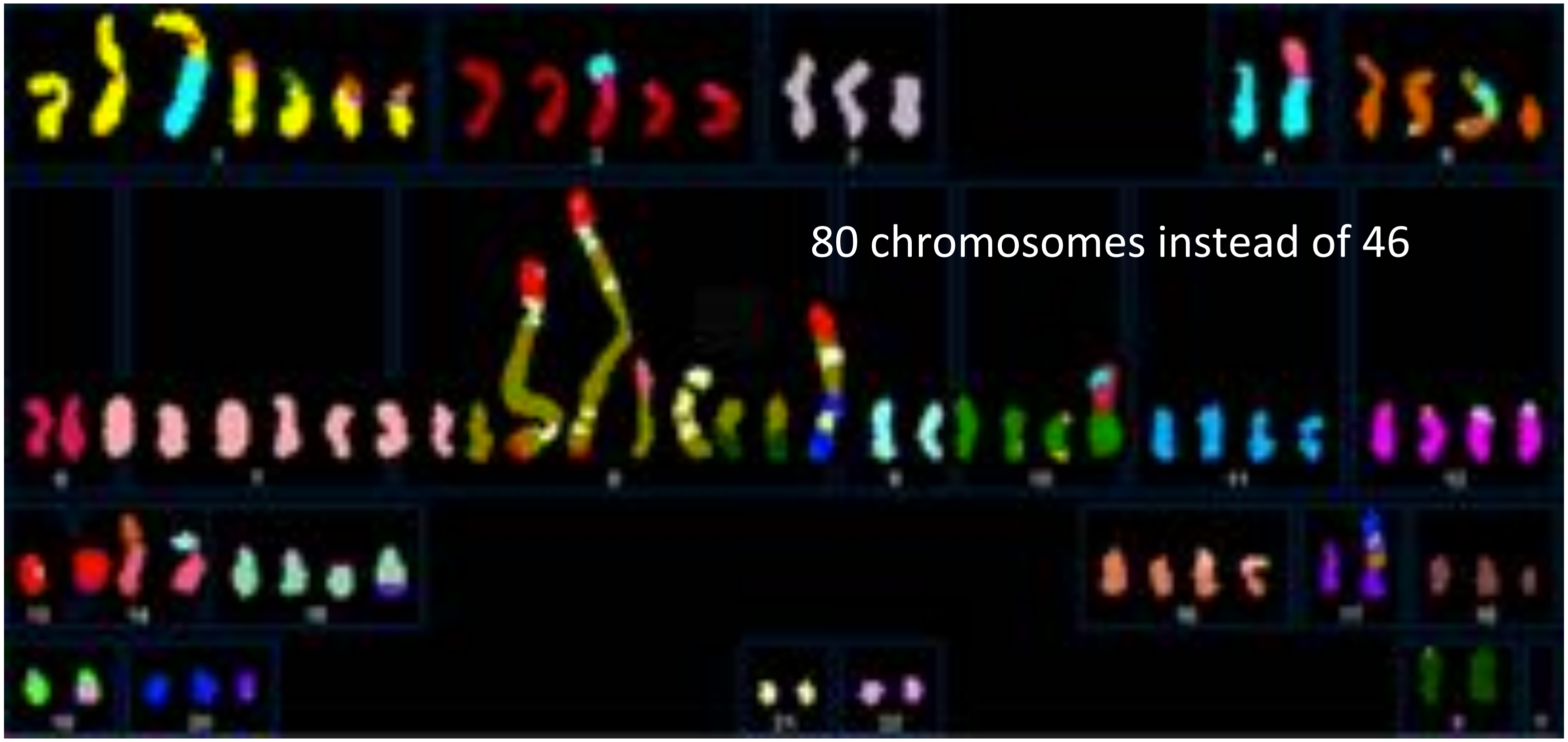
McPherson + Beck at the Ontario Institute for Cancer Research

Pacific Biosciences

DNAnexus

SK-BR-3

Most commonly used Her2-amplified breast cancer cell line



80 chromosomes instead of 46

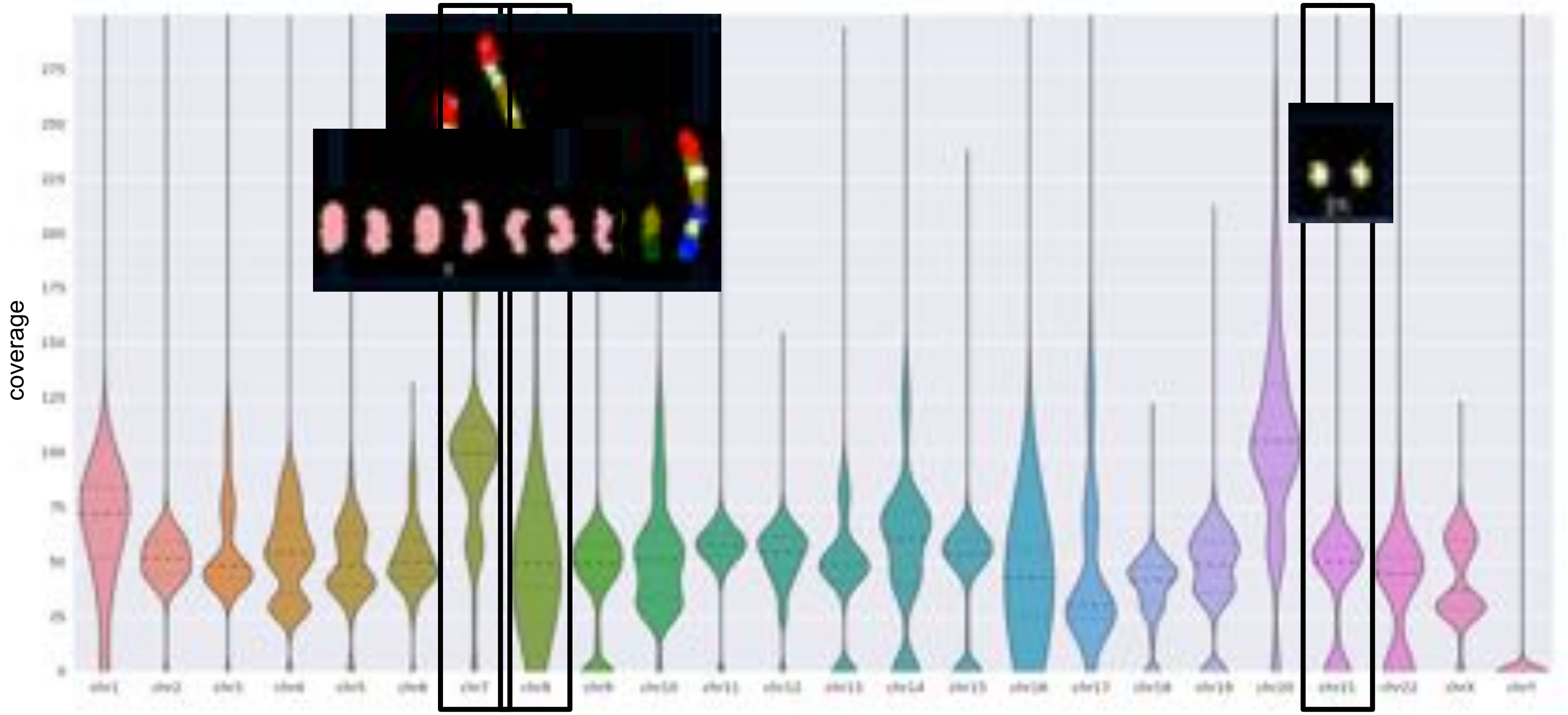
Often used for pre-clinical research on Her2-targeting therapeutics such as Herceptin (Trastuzumab) and resistance to these therapies.

(Davidson et al, 2000)

PacBio long-read DNA sequencing

mean read length: 9 kb
max read length: 71 kb

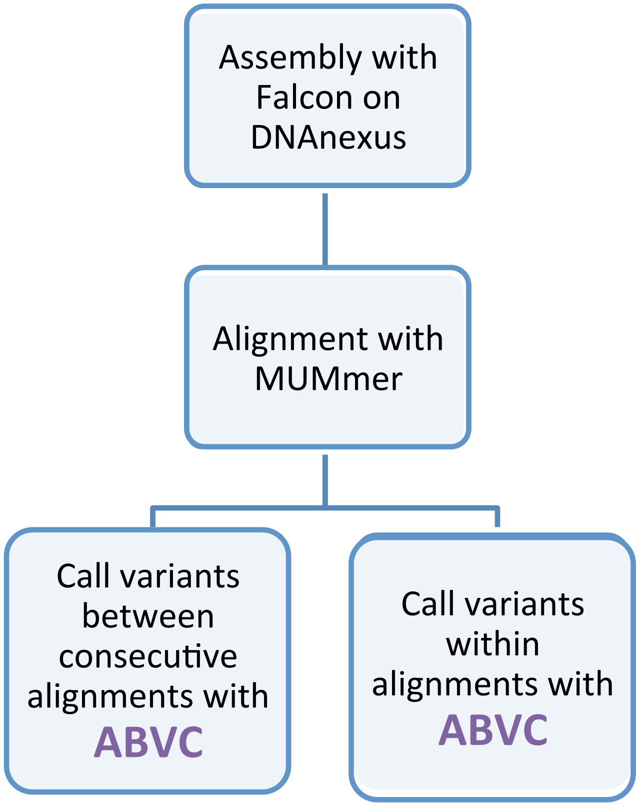
72X coverage



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

Genome structural analysis

Assembly-based



Assembly with Falcon on DNAnexus

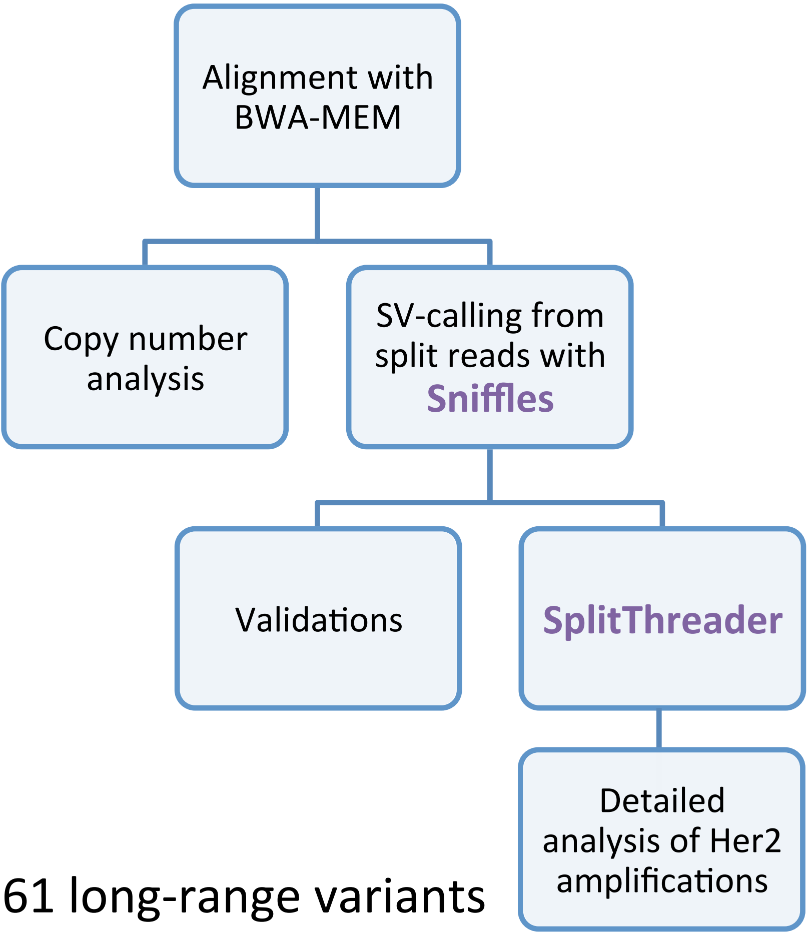
Alignment with MUMmer

Call variants between consecutive alignments with **ABVC**

Call variants within alignments with **ABVC**

~ 11,000 local variants
50 bp < size < 10 kbp

Alignment-based



Alignment with BWA-MEM

Copy number analysis

SV-calling from split reads with **Sniffles**

Validations

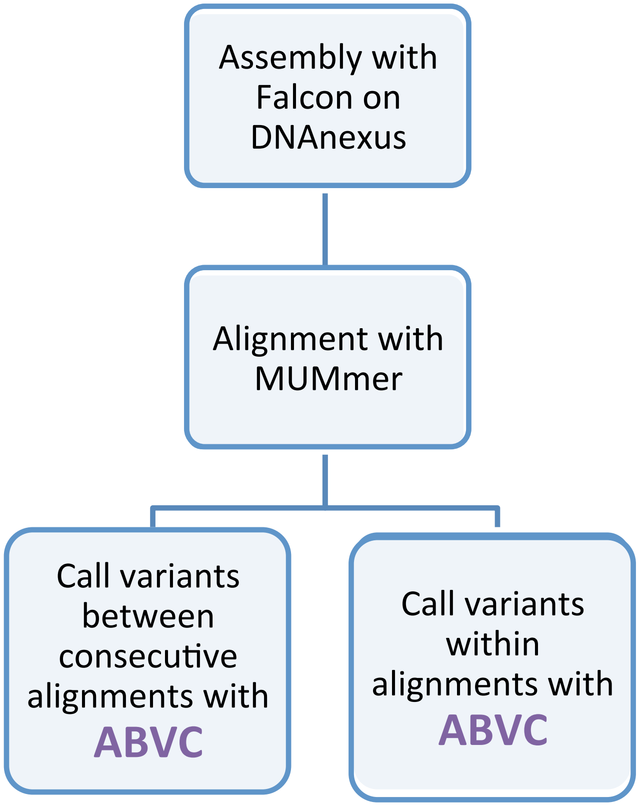
SplitThreader

Detailed analysis of Her2 amplifications

661 long-range variants
(>10kb distance)

Genome structural analysis

Assembly-based



Assembly with Falcon on DNAnexus

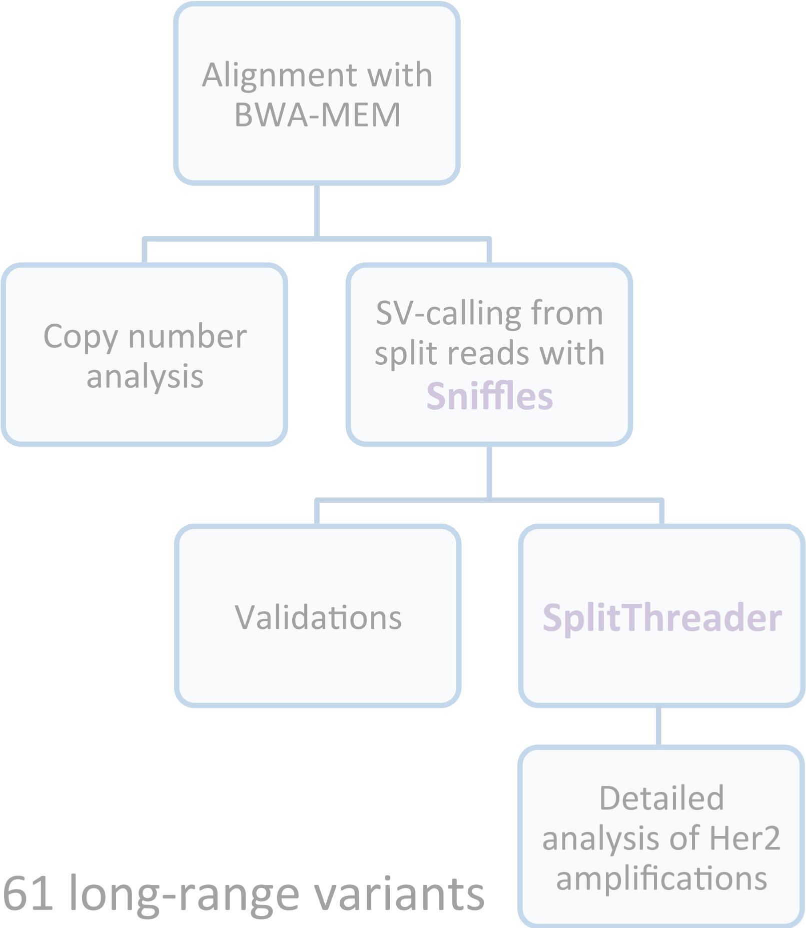
Alignment with MUMmer

Call variants between consecutive alignments with **ABVC**

Call variants within alignments with **ABVC**

~ 11,000 local variants
50 bp < size < 10 kbp

Alignment-based



Alignment with BWA-MEM

Copy number analysis

SV-calling from split reads with **Sniffles**

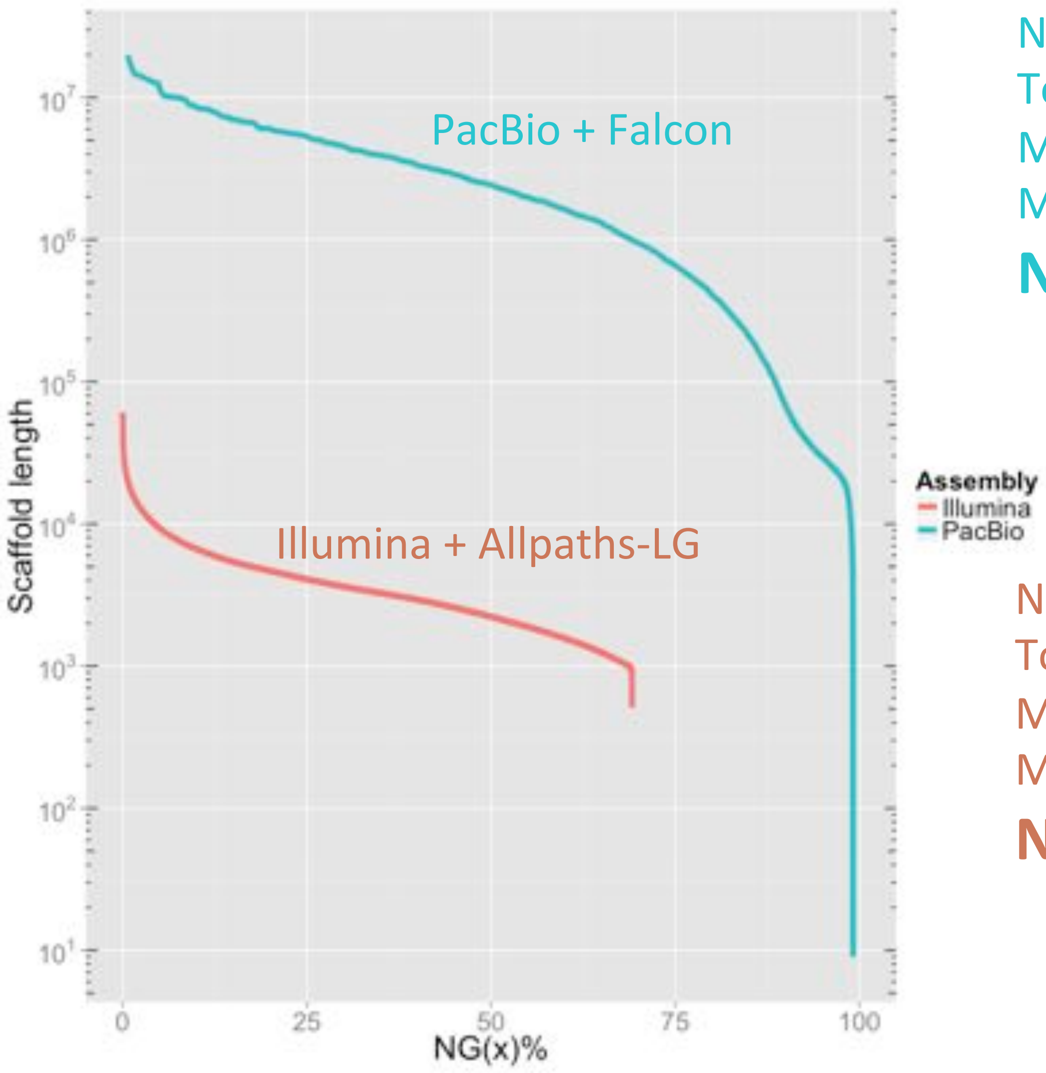
Validations

SplitThreader

Detailed analysis of Her2 amplifications

661 long-range variants
(>10kb distance)

Assembly using PacBio yields far better contiguity



Number of sequences: 13,532
Total sequence length: 2.97Gb
Mean: 266 kb
Max: 19.9 Mb

N50: 2.46 Mb

Relative to a genome size of 3 Gb

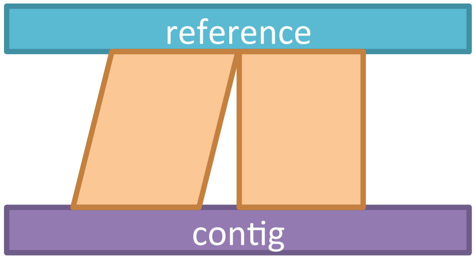
Number of sequences: 748,955
Total sequence length: 2.07 Gb
Mean: 2.8 kb
Max: 61 kb

N50: 3.3 kb

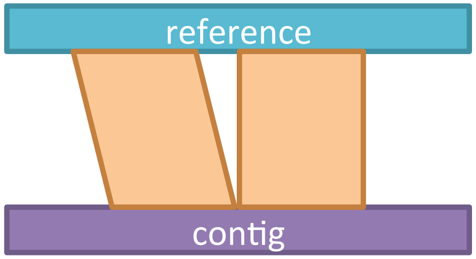
ABVC: Assembly-Based Variant-Caller

Defined point

Insertion

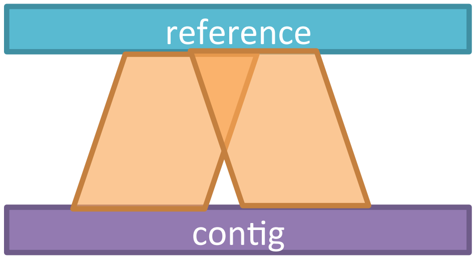


Deletion

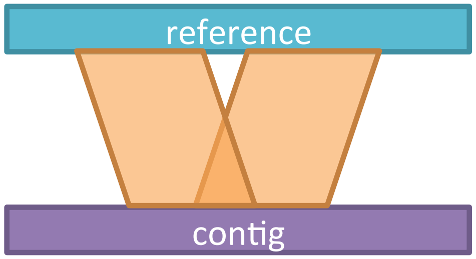


Overlapping alignments suggest tandem repeat

Tandem Expansions

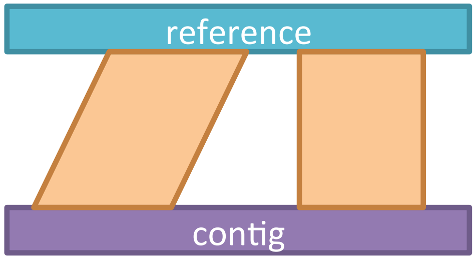


Tandem Contractions

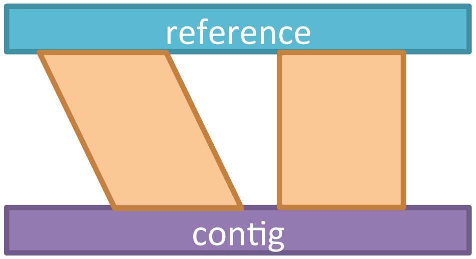


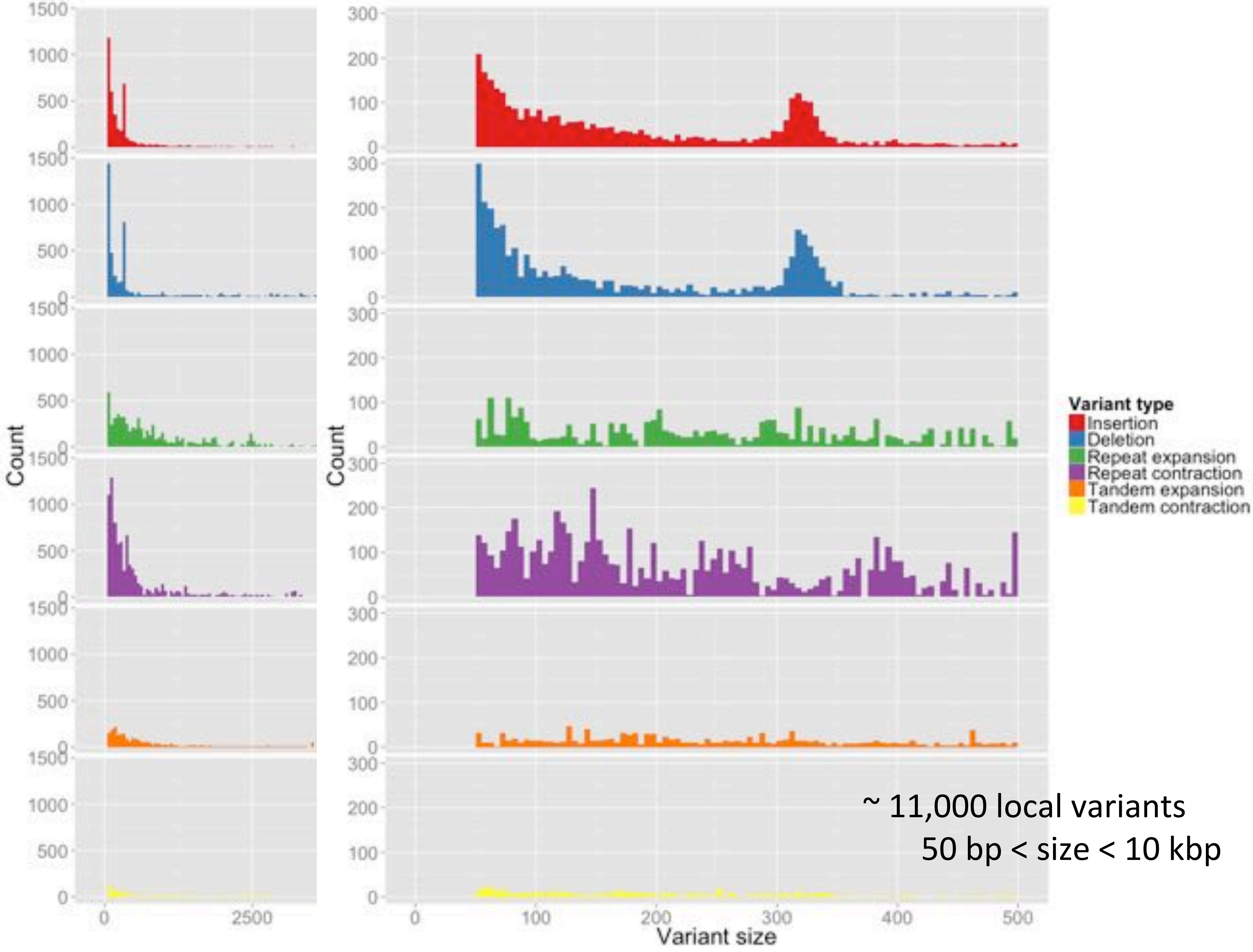
Gap where sequences do not align uniquely suggests a repeat

Repeat Expansions

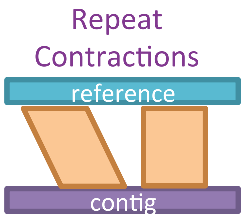


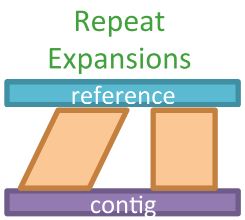
Repeat Contractions

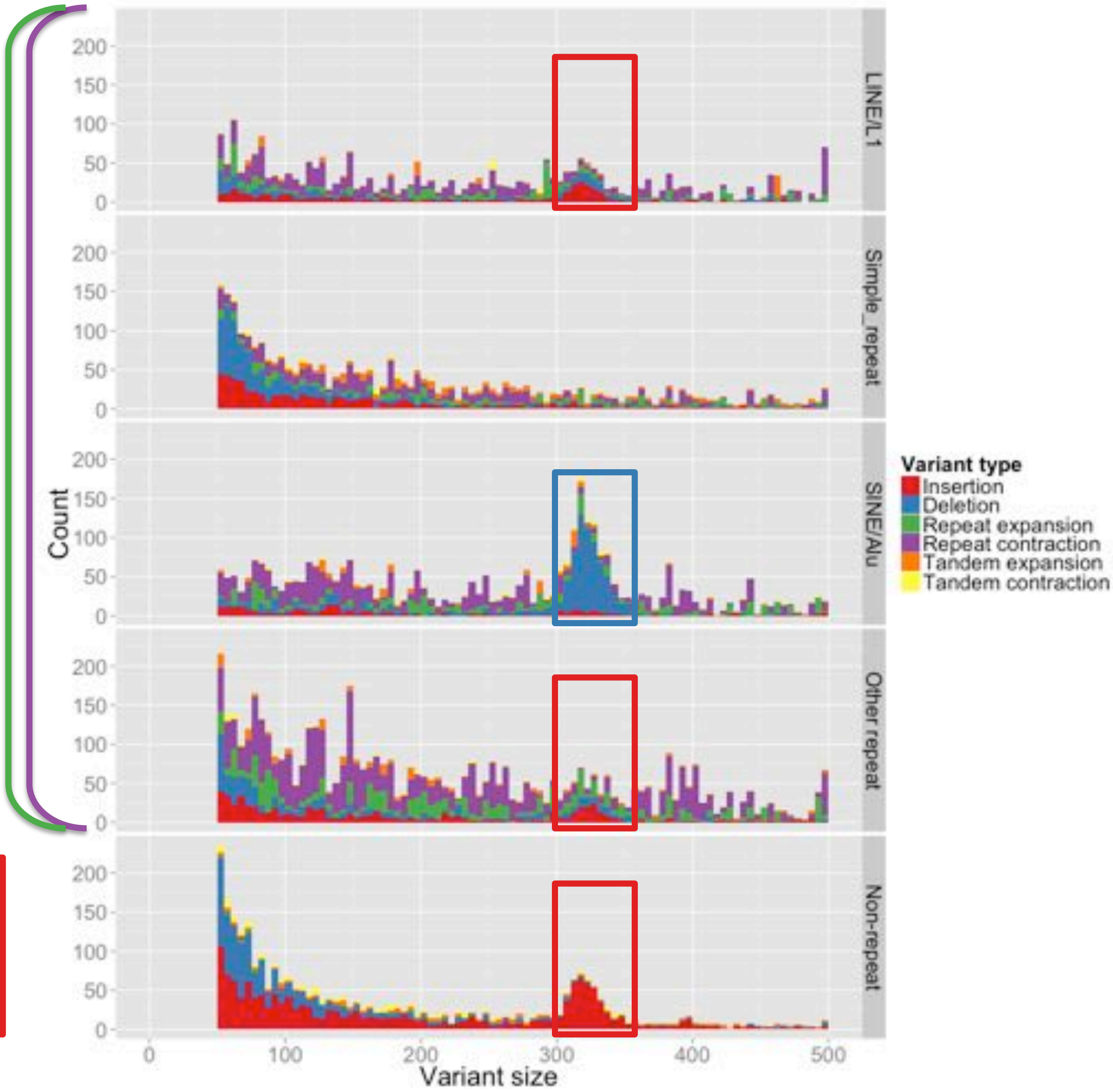




~ 11,000 local variants
50 bp < size < 10 kbp



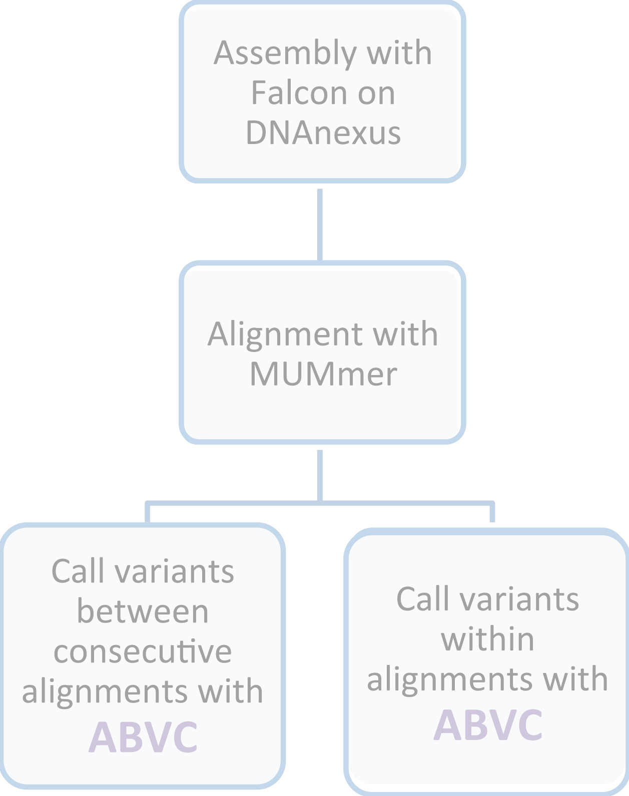




BLASTed 515 insertions:
427 (83%) of them matched Alu elements

Genome structural analysis

Assembly-based



Assembly with Falcon on DNAnexus

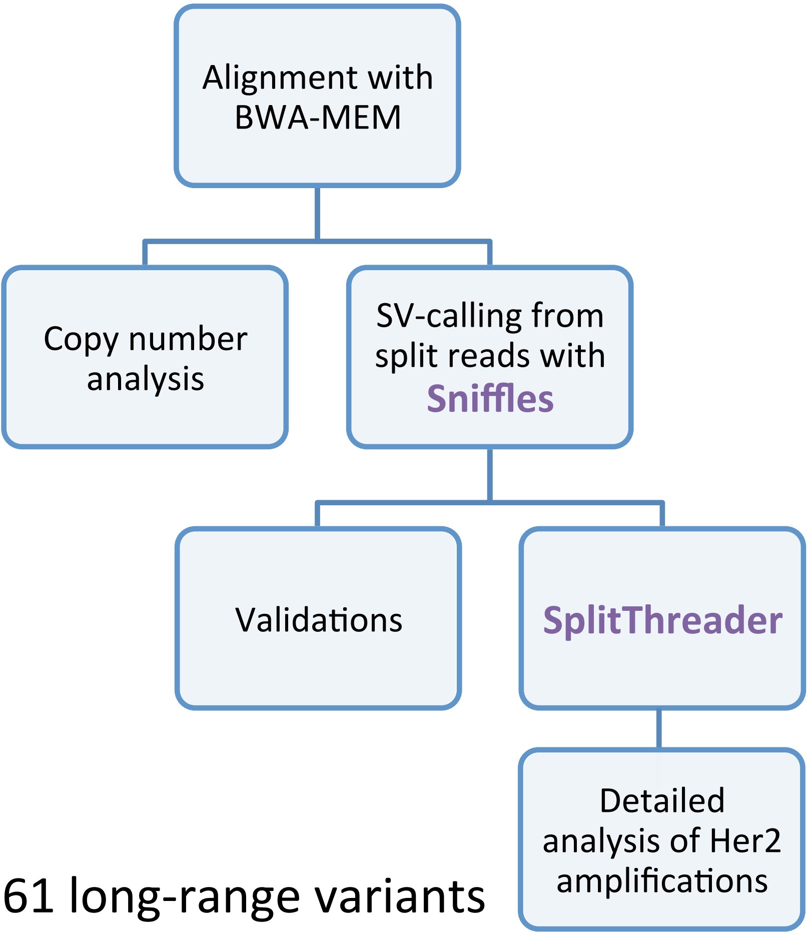
Alignment with MUMmer

Call variants between consecutive alignments with ABVC

Call variants within alignments with ABVC

~ 11,000 local variants
50 bp < size < 10 kbp

Alignment-based



Alignment with BWA-MEM

Copy number analysis

SV-calling from split reads with Sniffles

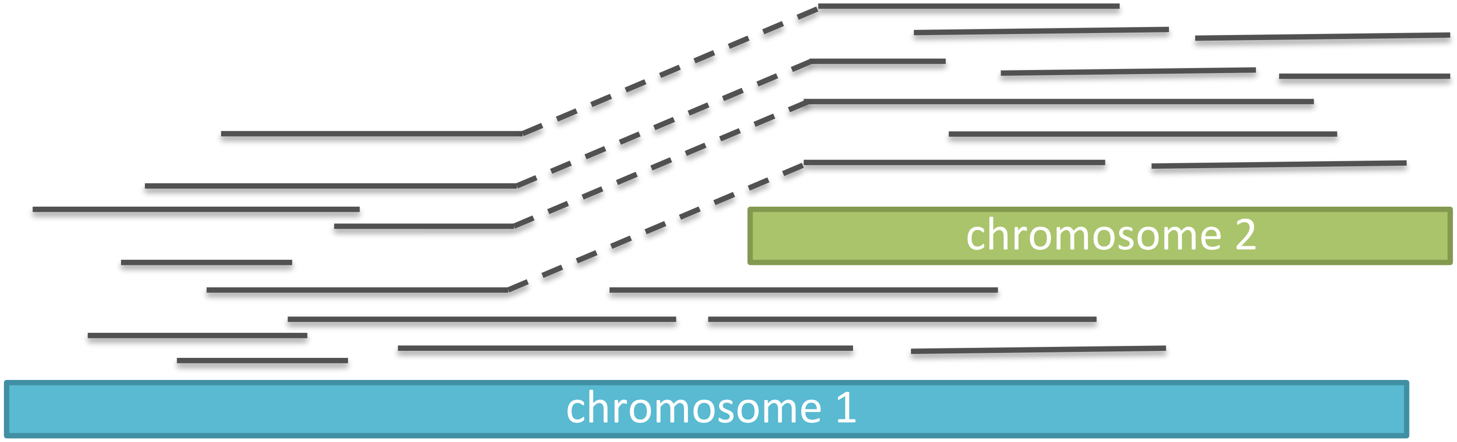
Validations

SplitThreader

Detailed analysis of Her2 amplifications

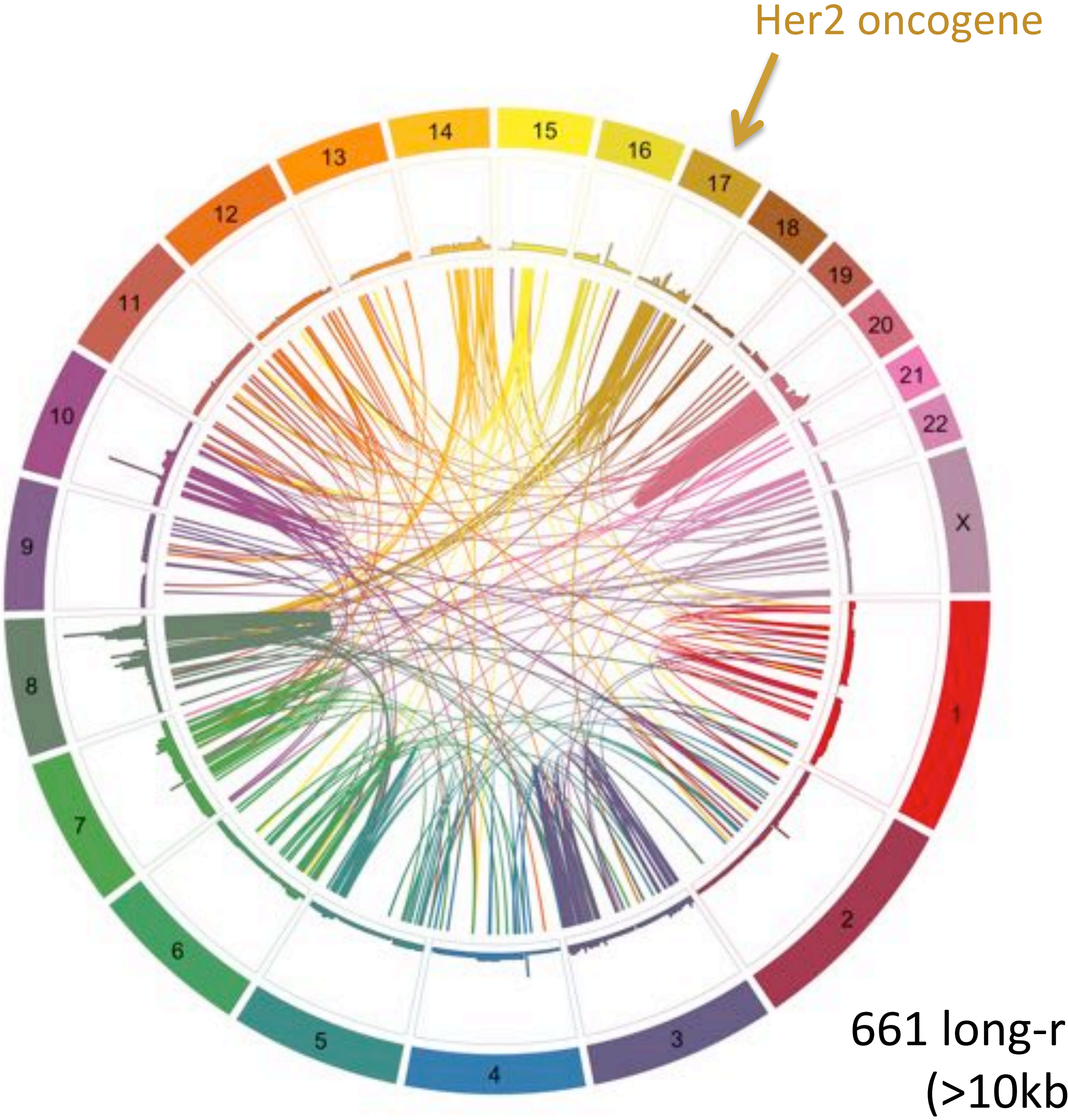
661 long-range variants
(>10kb distance)

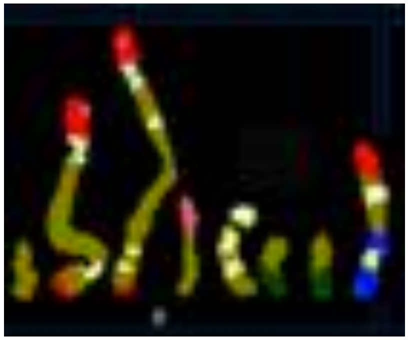
Split-read variant calling with Sniffles to capture the long-range variants



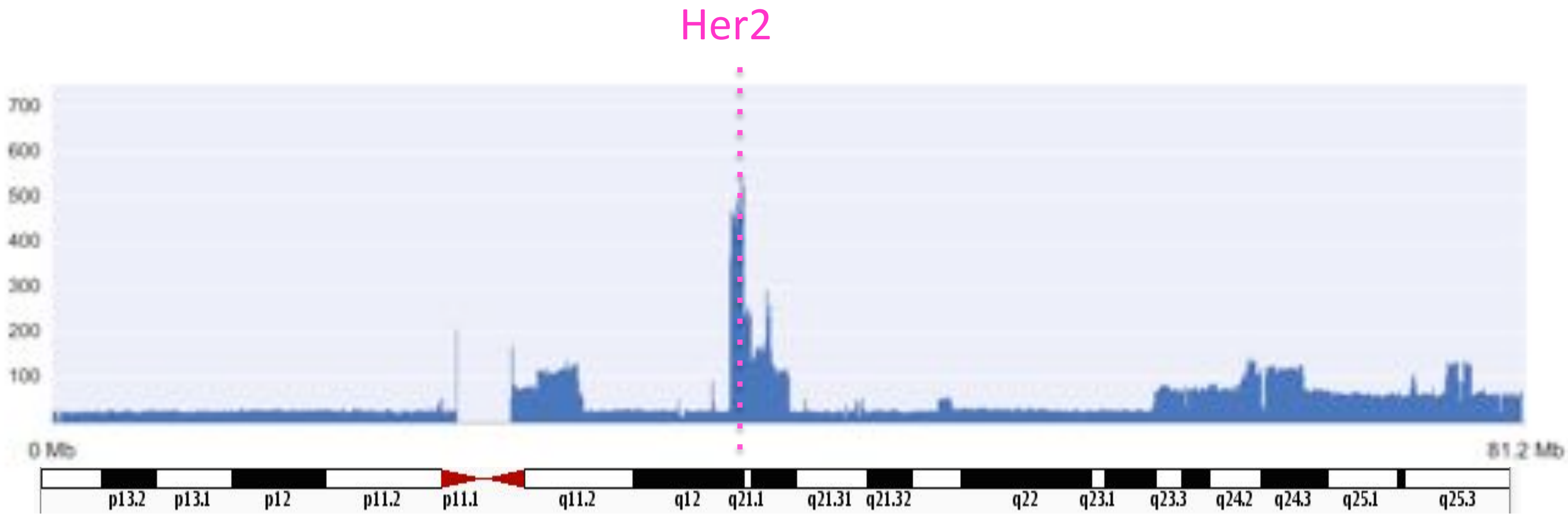
See Fritz at Poster 183

Long-range structural variants found by Sniffles



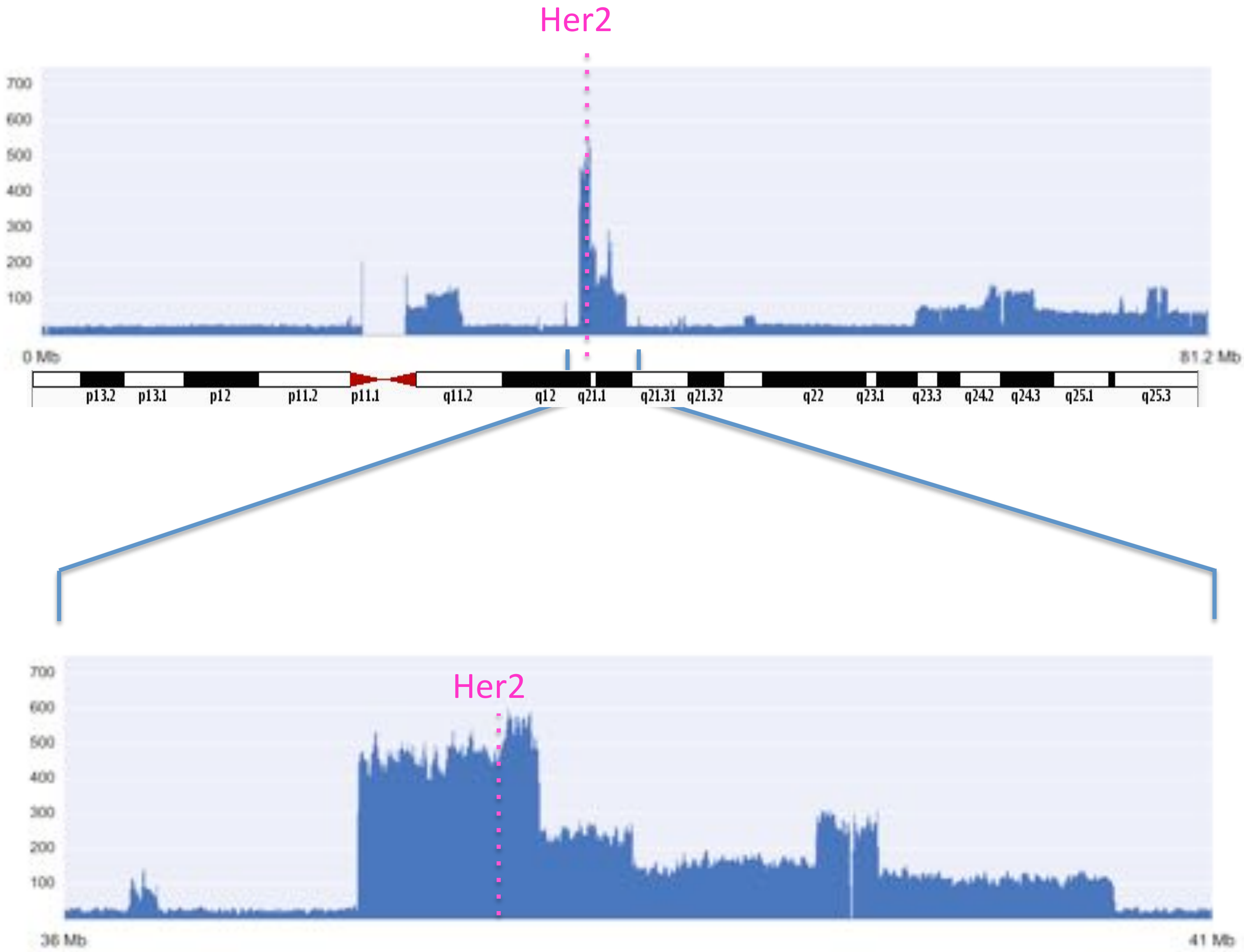


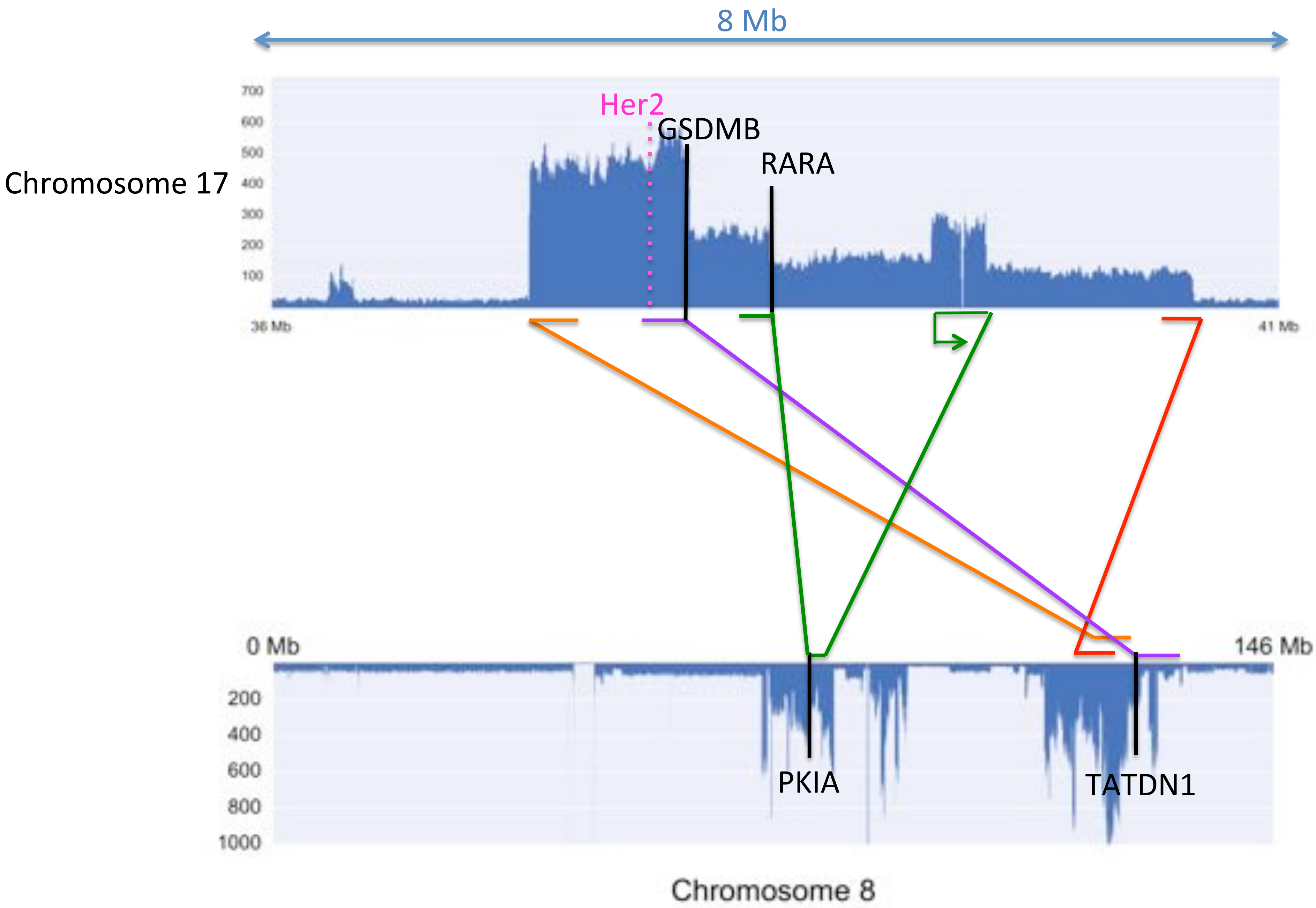
661 long-range variants
(>10kb distance)



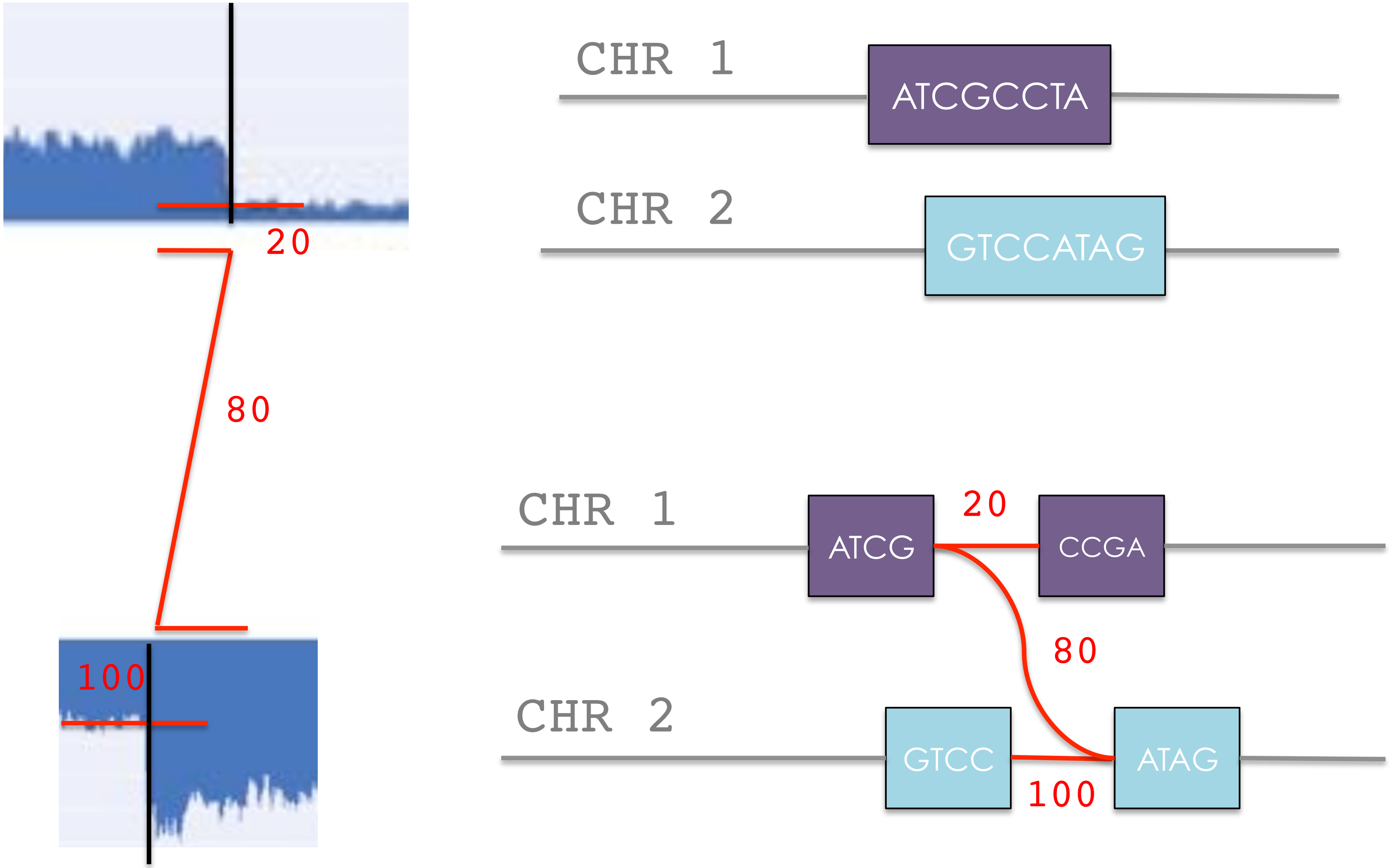
Chr 17: 83 Mb

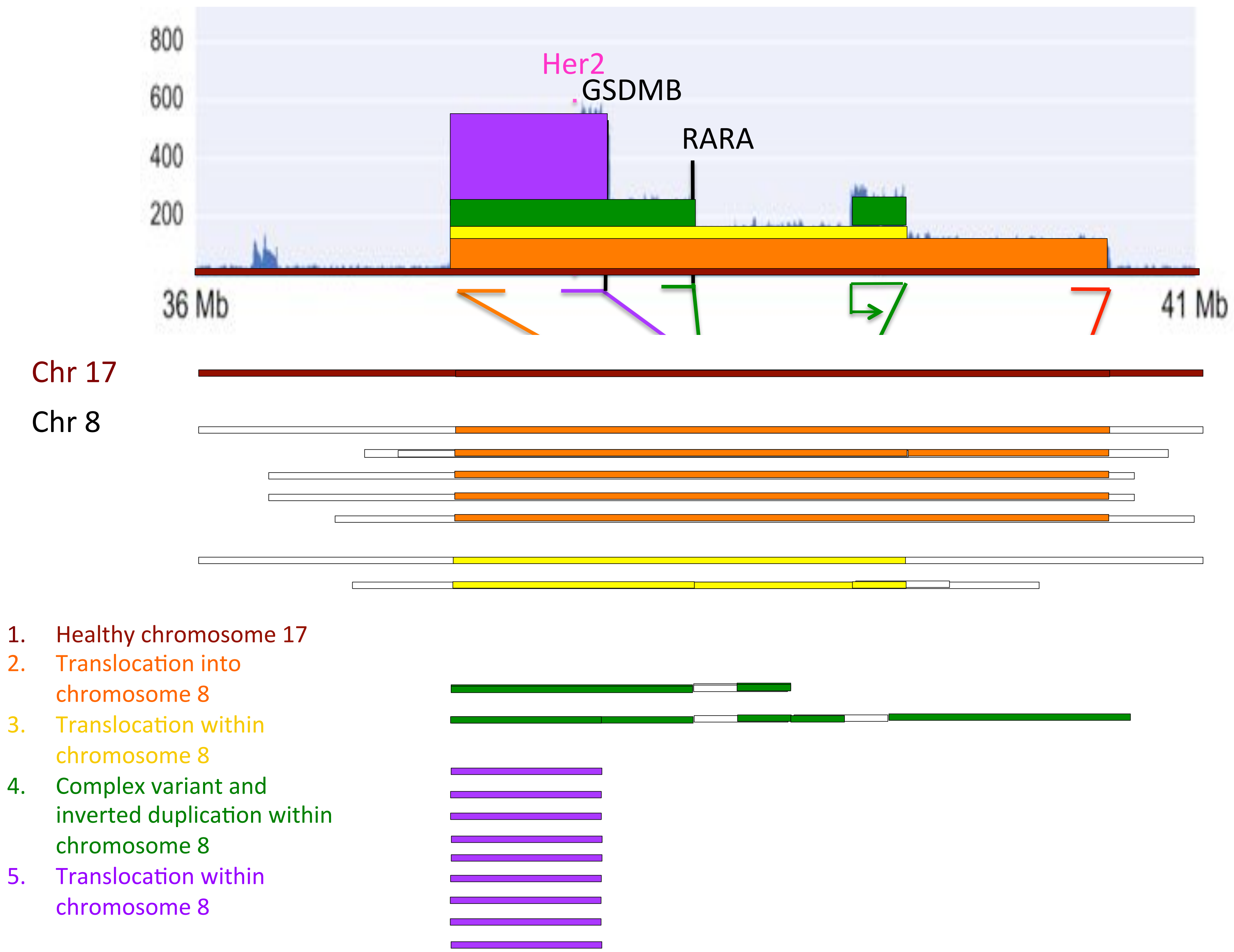
8 Mb


Her2
700
600
500
400
300
200
100
0 Mb
81.2 Mb
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
Her2
700
600
500
400
300
200
100
36 Mb
41 Mb


8 Mb
Chromosome 17
700
600
500
400
300
200
100
Her2
GSDMB
RARA
36 Mb
41 Mb
0 Mb
146 Mb
200
400
600
800
1000
PKIA
TATDN1
Chromosome 8

SplitThreader: Graphical threading to retrace complex history of rearrangements in cancer genomes



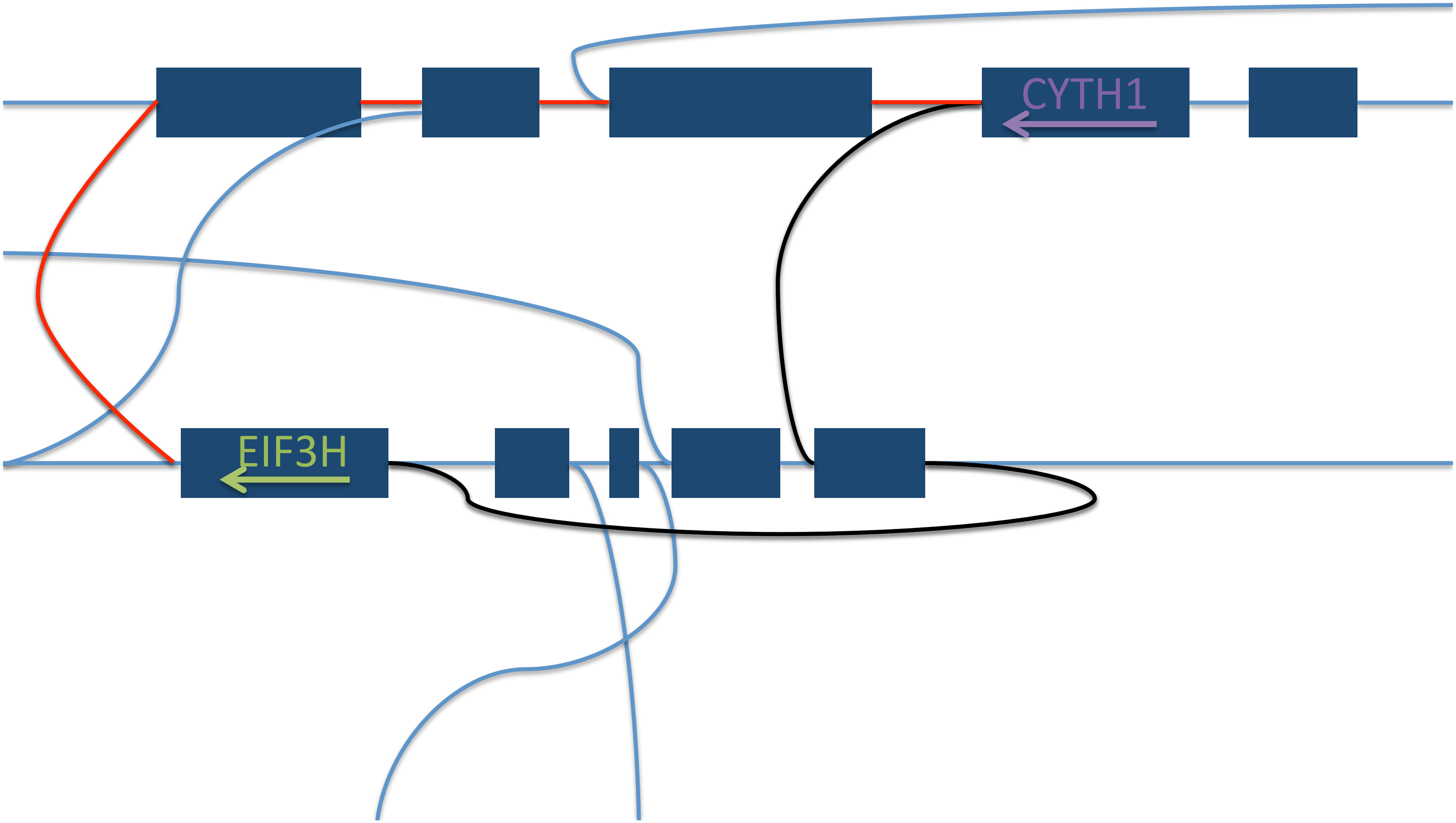


1. Healthy chromosome 17
2. Translocation into chromosome 8
3. Translocation within chromosome 8
4. Complex variant and inverted duplication within chromosome 8
5. Translocation within chromosome 8

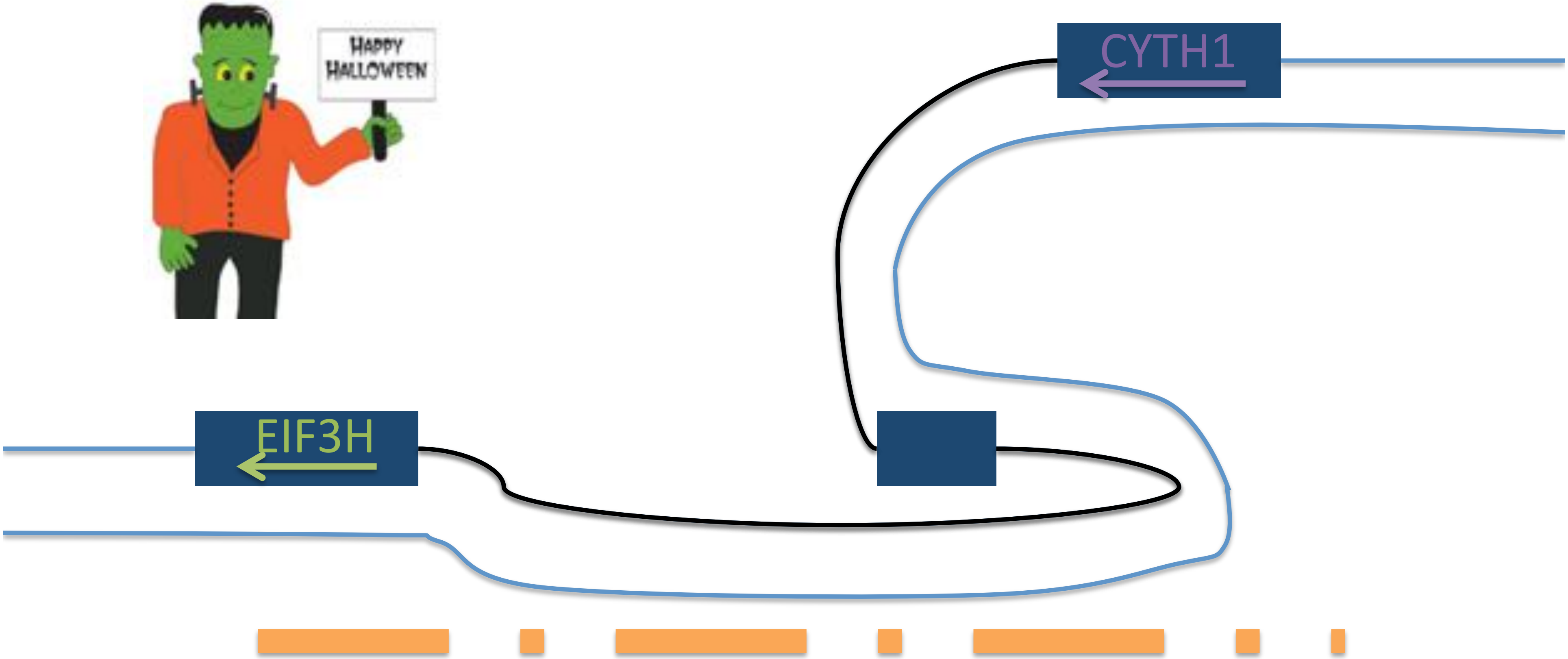
Transcriptome analysis with IsoSeq: Long-read RNA sequencing

- Full-length transcripts
- Found 17 gene fusions with both DNA and RNA evidence
  - 13 seen in previous RNA-seq literature
  - 4 novel fusions
- 2 previously observed fusions had RNA evidence but no direct link in the DNA
  - Confirmed using SplitThreader

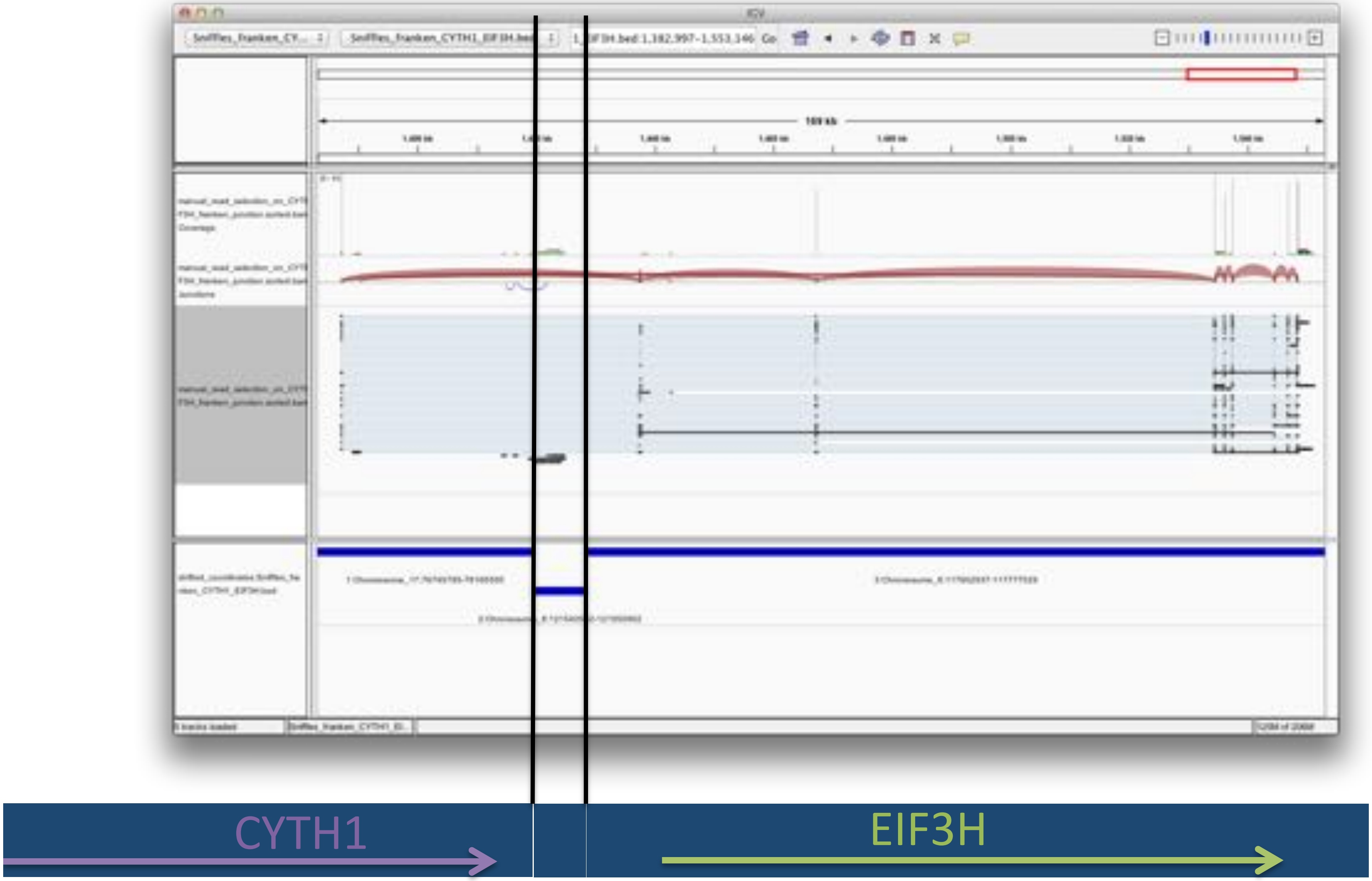
CYTH1-EIF3H gene fusion in the SplitThreader graph



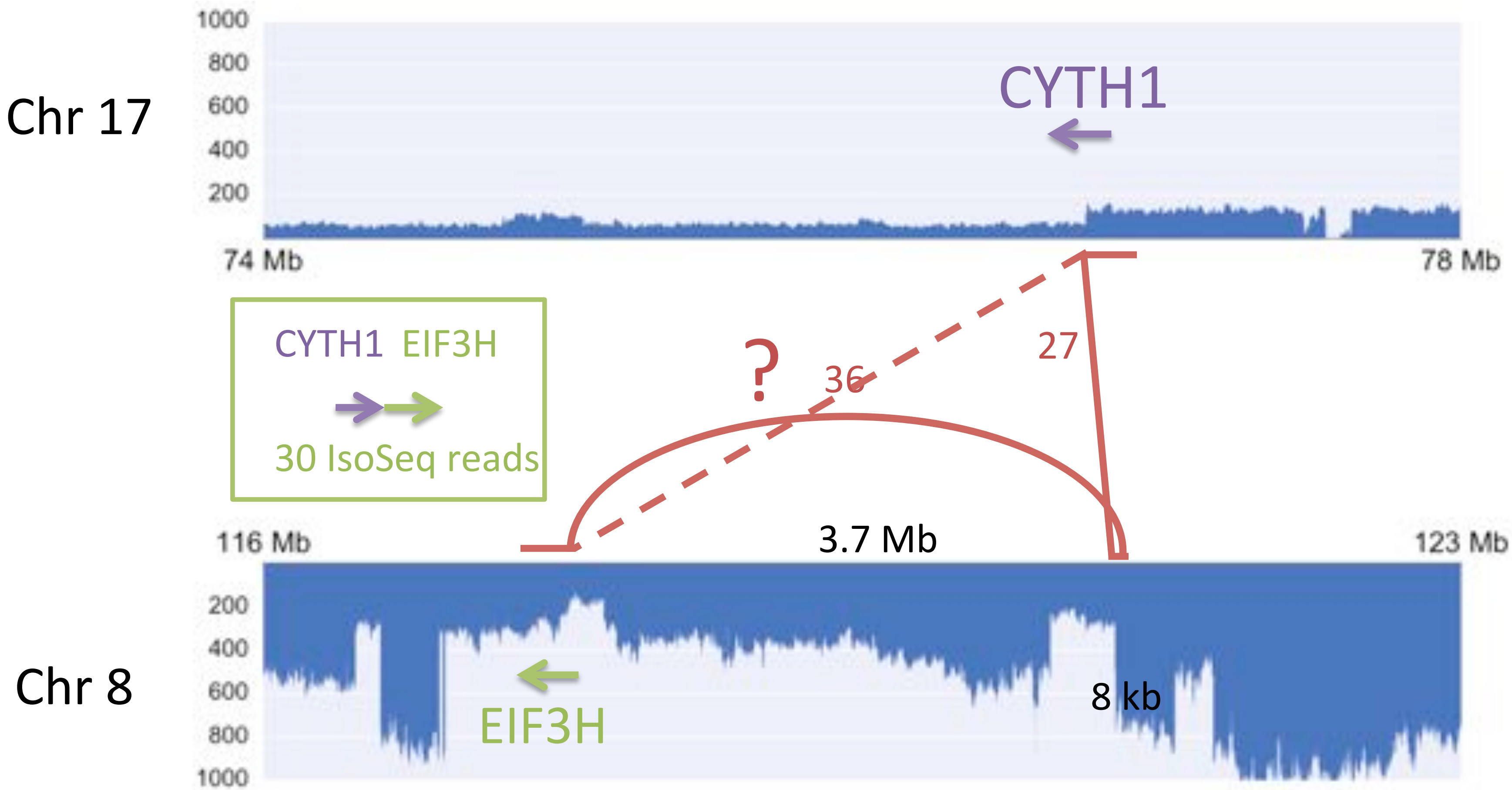
CYTH1-EIF3H gene fusion in the SplitThreader graph



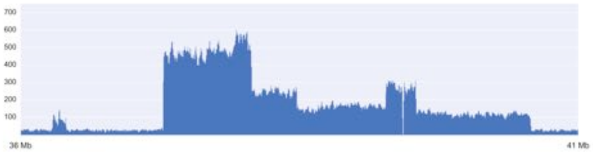
Frankensteining the CYTH1-EIF3H gene fusion


IGV
Sniffles_Franken_CY...
Sniffles_Franken_CYTH1_EIF3H.bed
1_EIF3H.bed:1,382,997-1,553,146
Go
169 kb
CYTH1
EIF3H

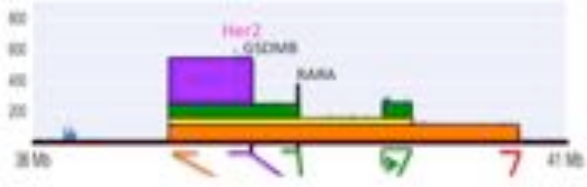
CYTH1-EIF3H gene fusion

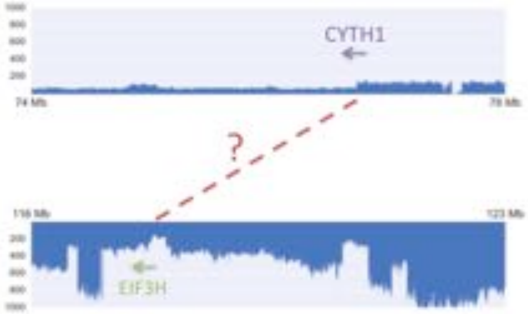


The genome informs the transcriptome

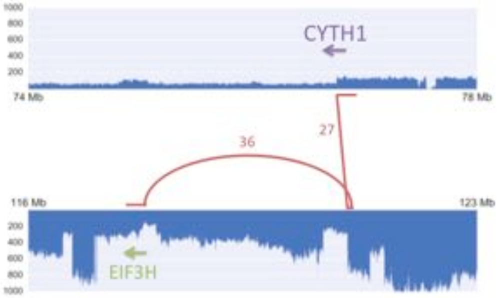


Explain amplifications





Trace gene fusions



More genomes coming soon!

Data and additional results: http://schatzlab.cshl.edu/data/skbr3/

Acknowledgments

CSH Cold Spring Harbor Laboratory

Sara Goodwin
Timour Baslan
Fritz Sedlazeck
Tyler Garvin
Han Fang
James Gurtowski
Philipp Rescheneder
Elizabeth Hutton
Marley Alford
Melissa Kramer
Eric Antoniou
James Hicks
Michael Schatz
W. Richard McCombie

OICR Ontario Institute for Cancer Research

Karen Ng
Timothy Beck
Yogi Sundaravadanam
John McPherson

DNAnexus

PACIFIC BIOSCIENCES®

Elizabeth Tseng
Jason Chin

Genentech FOUNDATION