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

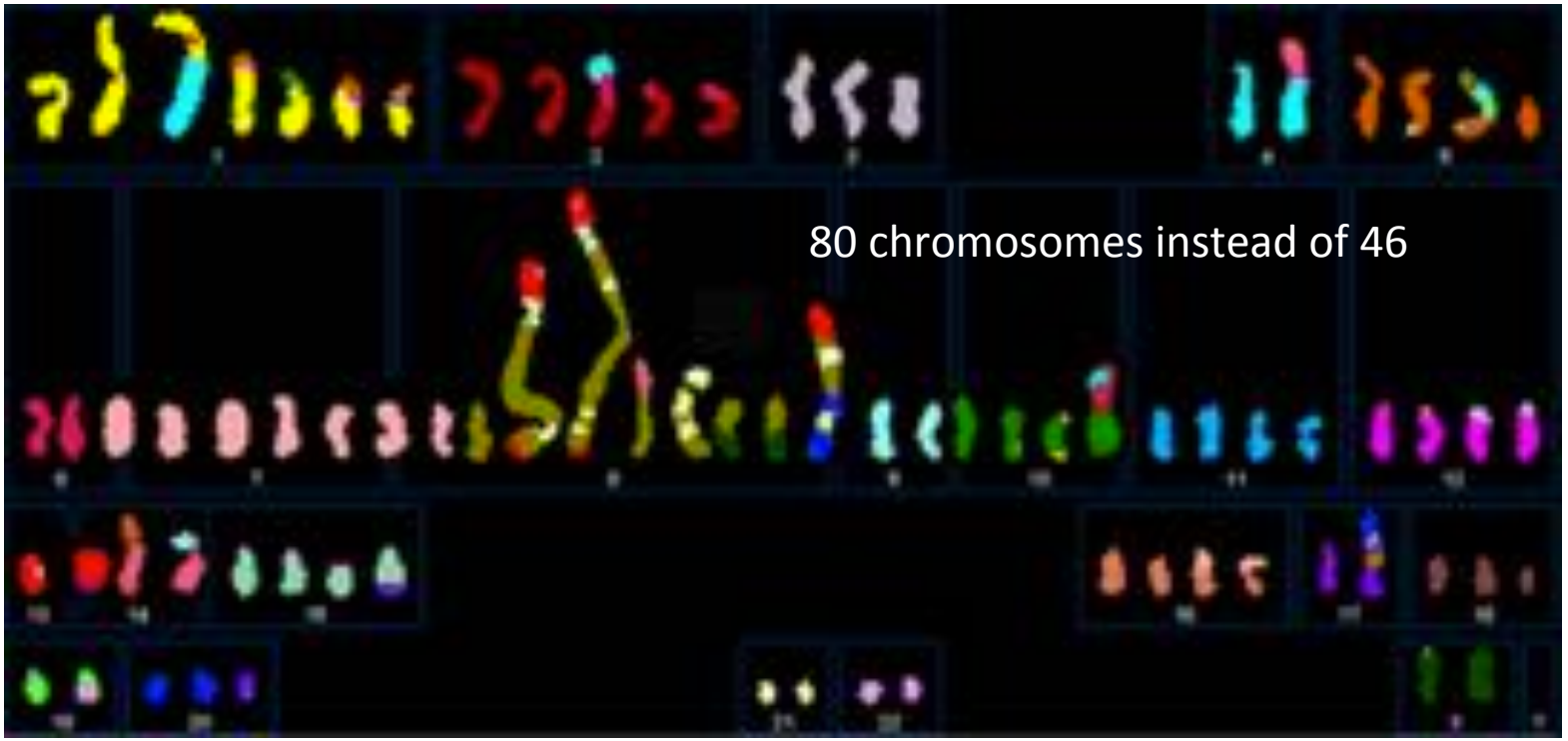
McPherson + Beck at the Ontario Institute for Cancer Research

Pacific Biosciences

DNAnexus

SK-BR-3

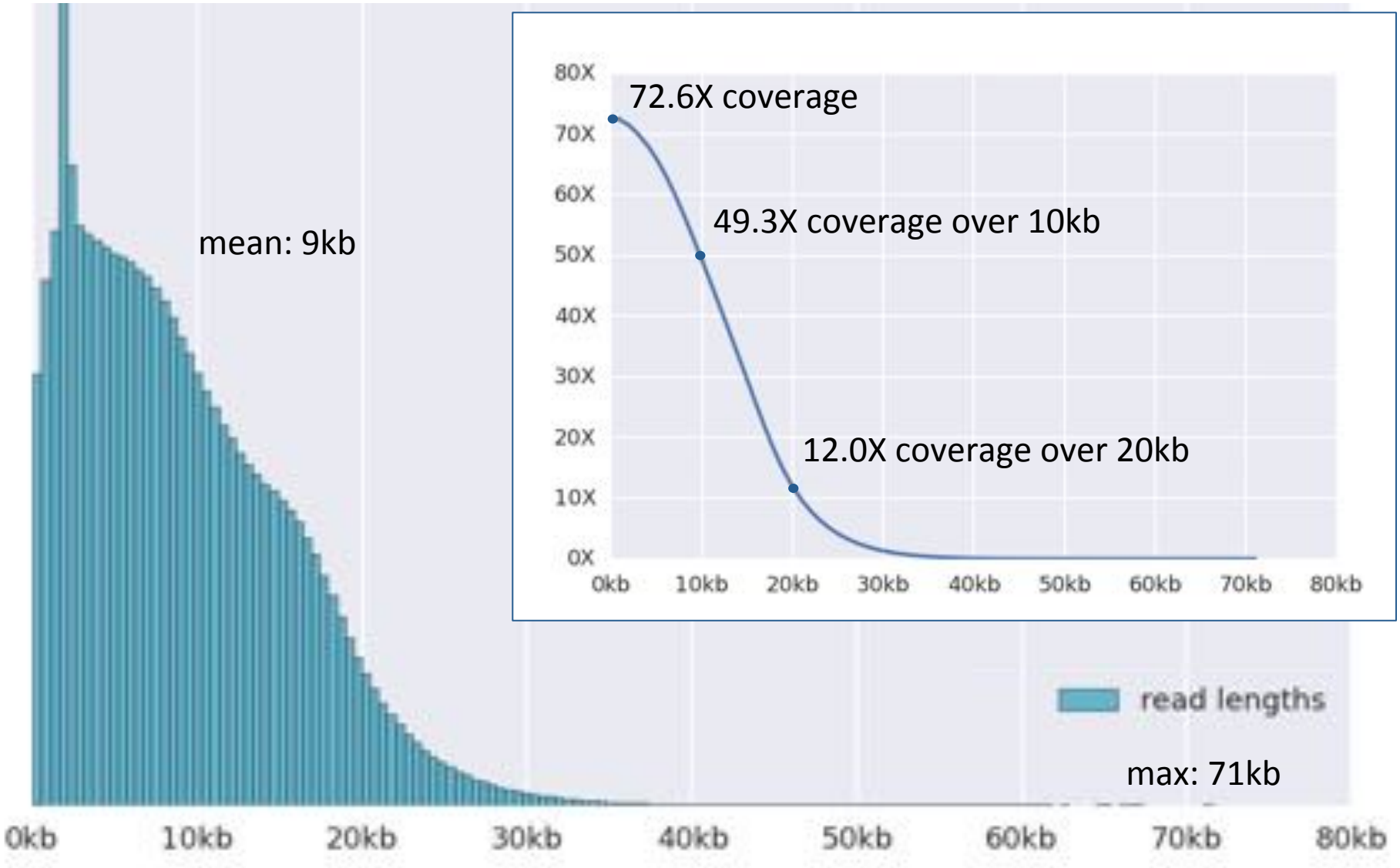
Most commonly used Her2-amplified breast cancer cell line



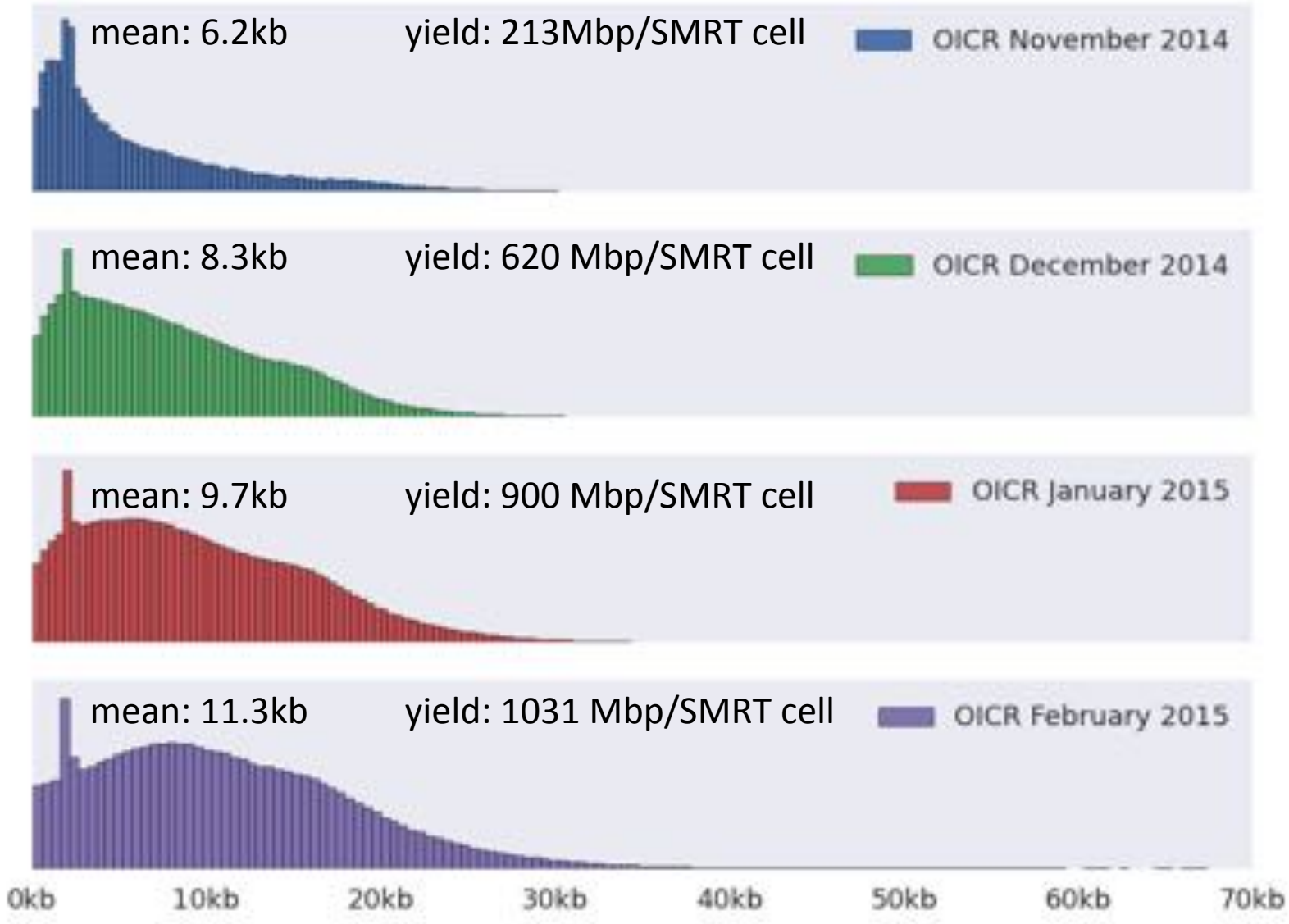
Often used for pre-clinical research on Her2-targeting therapeutics such as Herceptin (Trastuzumab) and resistance to these therapies.

(Davidson et al, 2000)

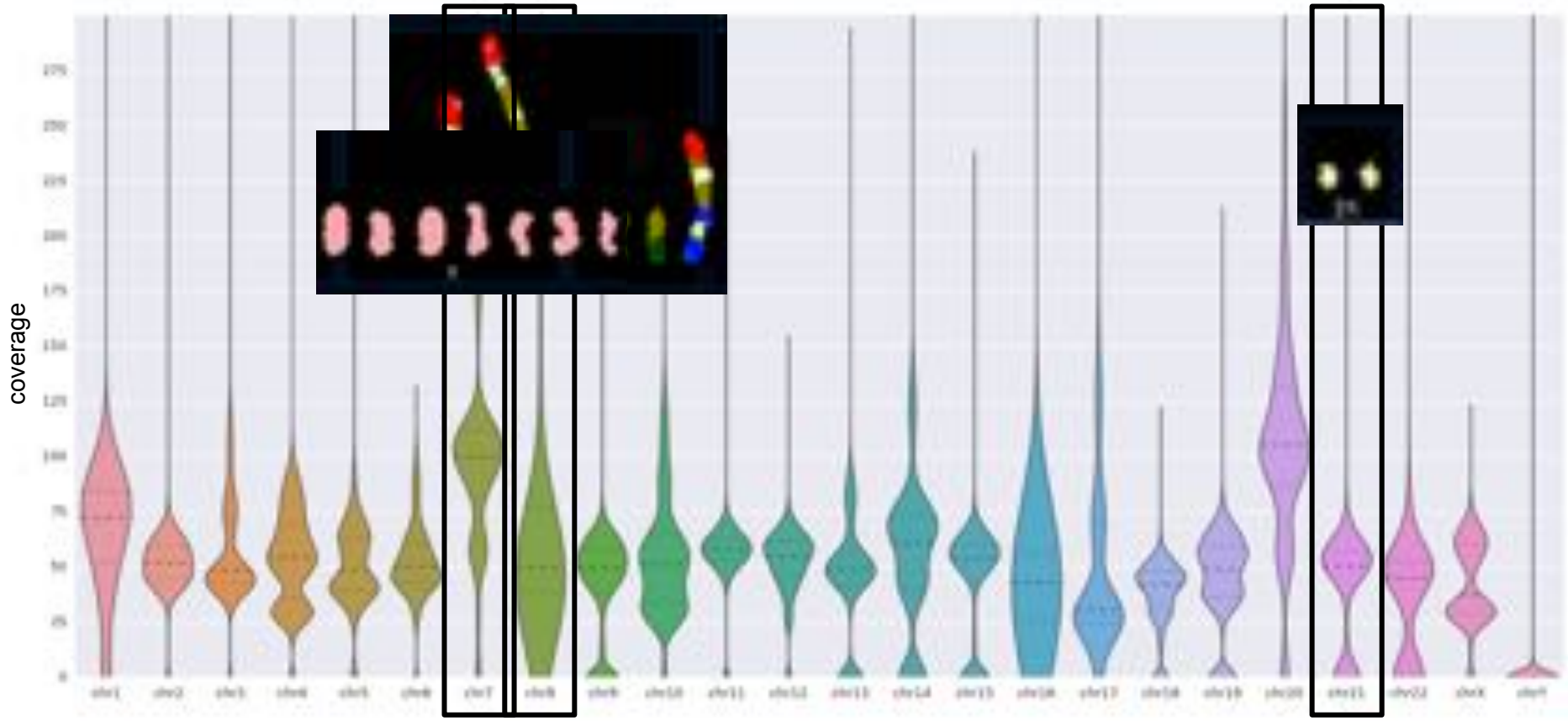
Sequencing SK-BR-3:
PacBio read length distribution



Dramatic changes just by experimenting with library preparation



Copy-number analysis is consistent with karyotype results



Genome-wide coverage averages around 54X

Coverage per chromosome varies greatly as expected from previous karyotyping results

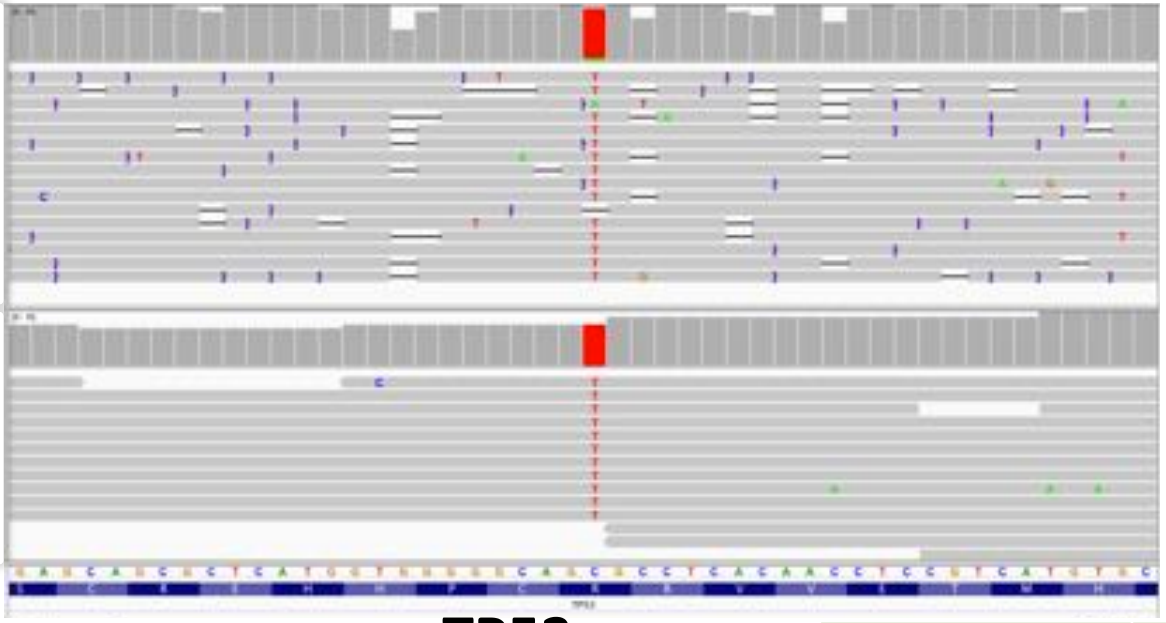
We could call SNPs if we wanted to

We recovered a known missense mutation in p53: **R175H**

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

PacBio

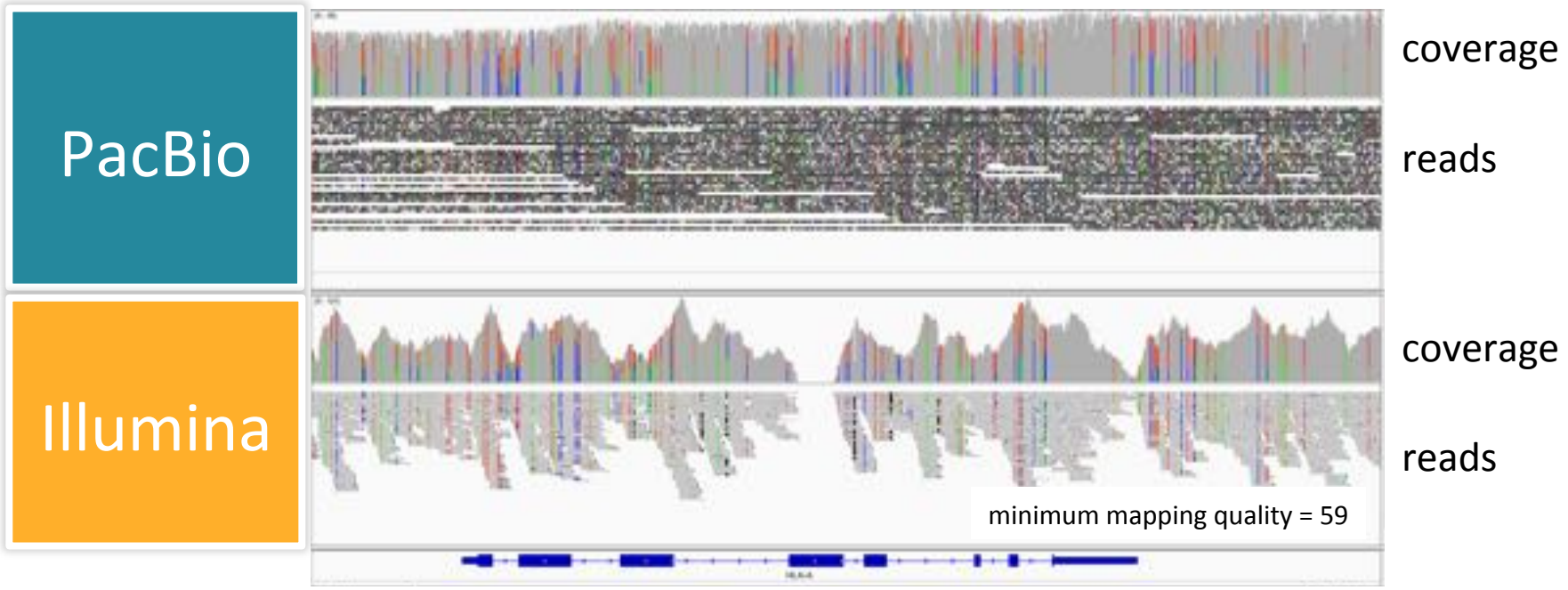
Illumina



TP53 gene

Insertion rate	11.5%
Deletion rate	3.4%
Mismatch rate	1.4%

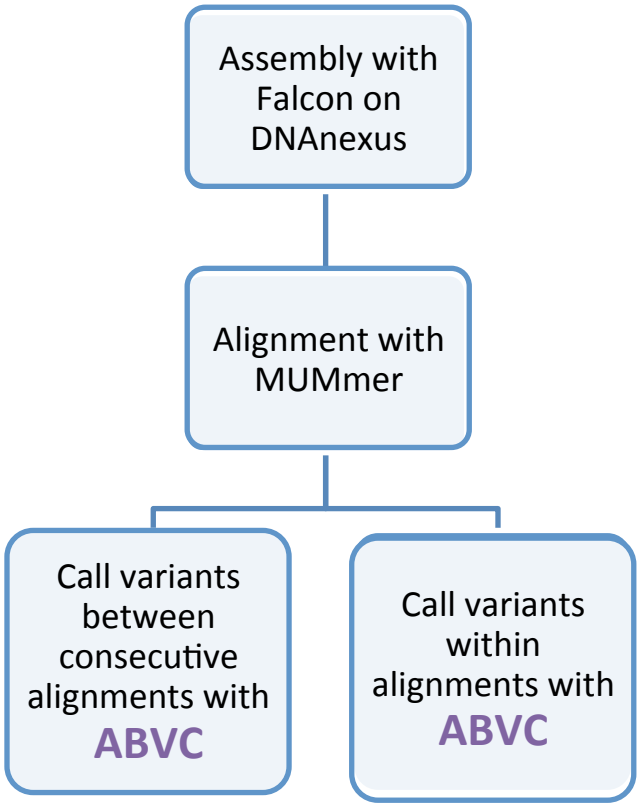
PacBio reads are longer and less susceptible to mapping issues



HLA-A gene

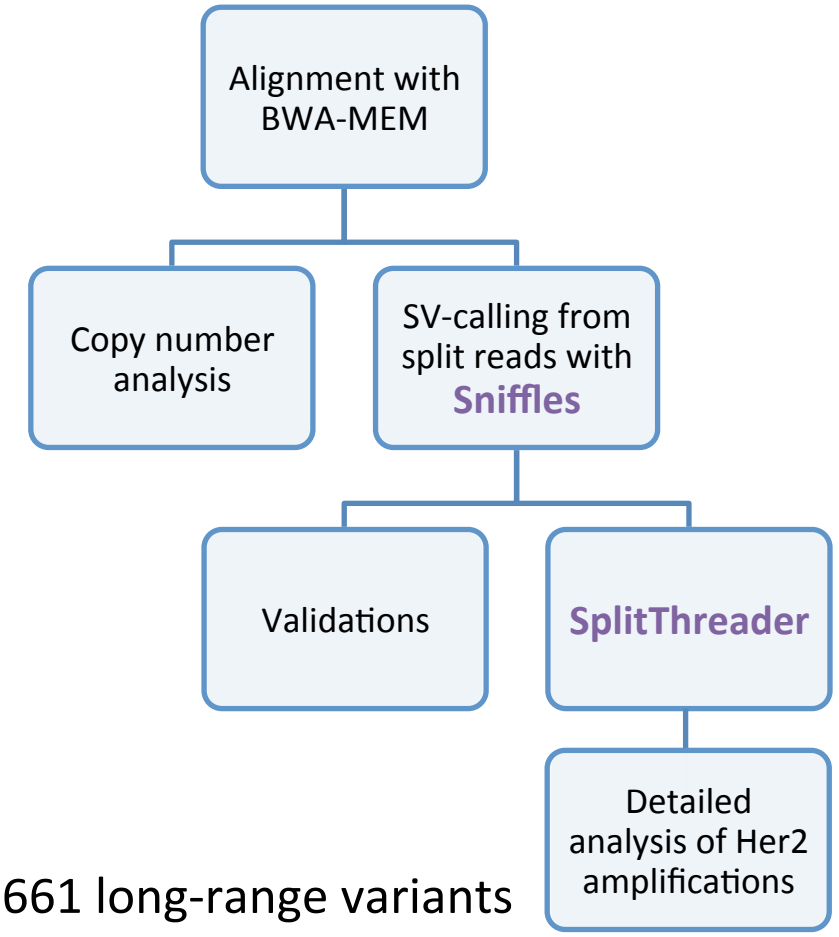
Genome structural analysis

Assembly-based



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

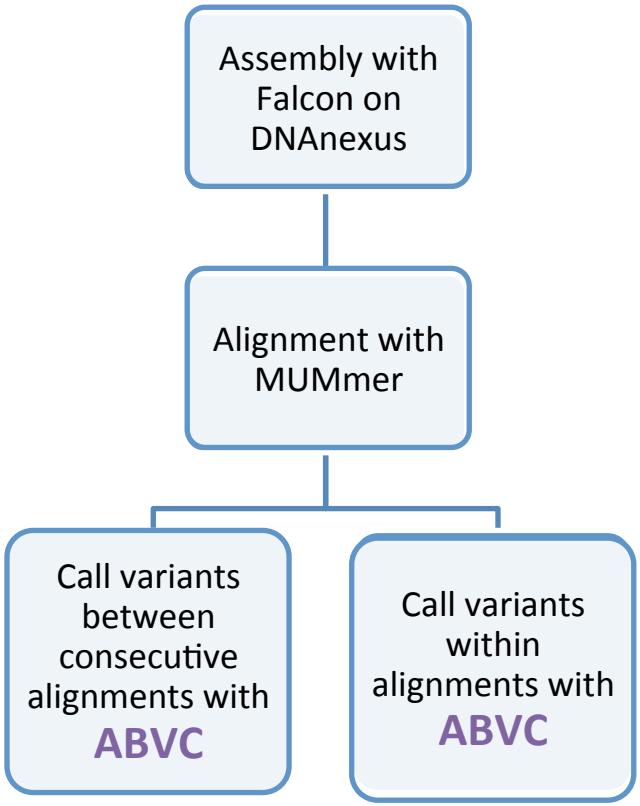
Alignment-based



661 long-range variants
(>10kb distance)

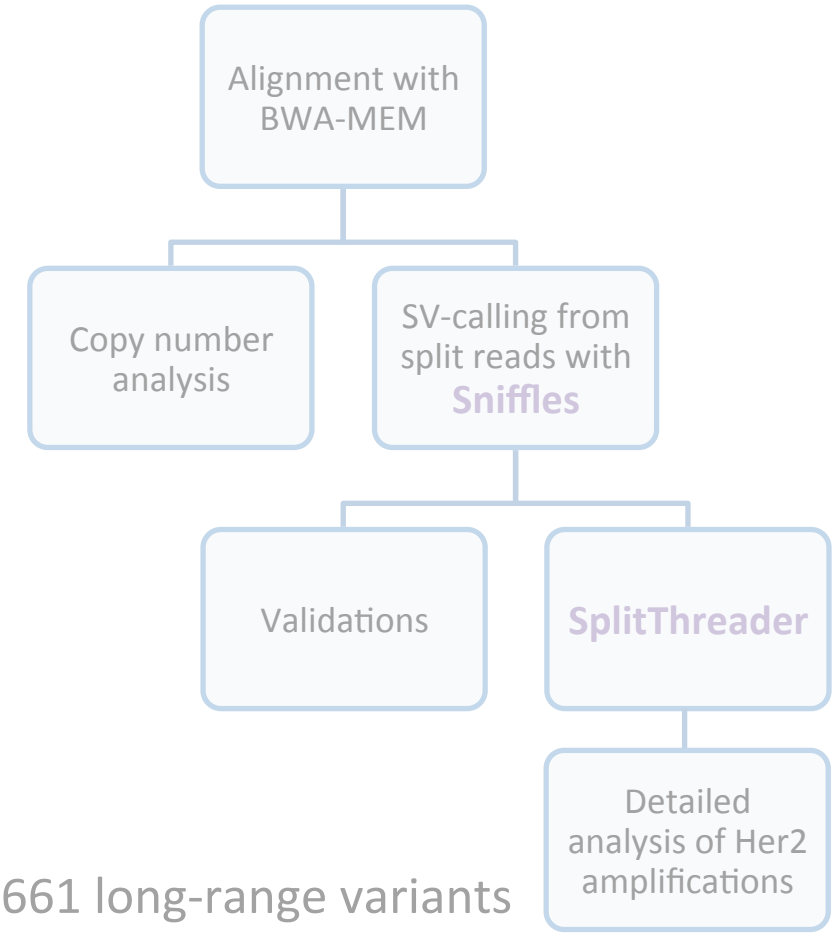
Genome structural analysis

Assembly-based



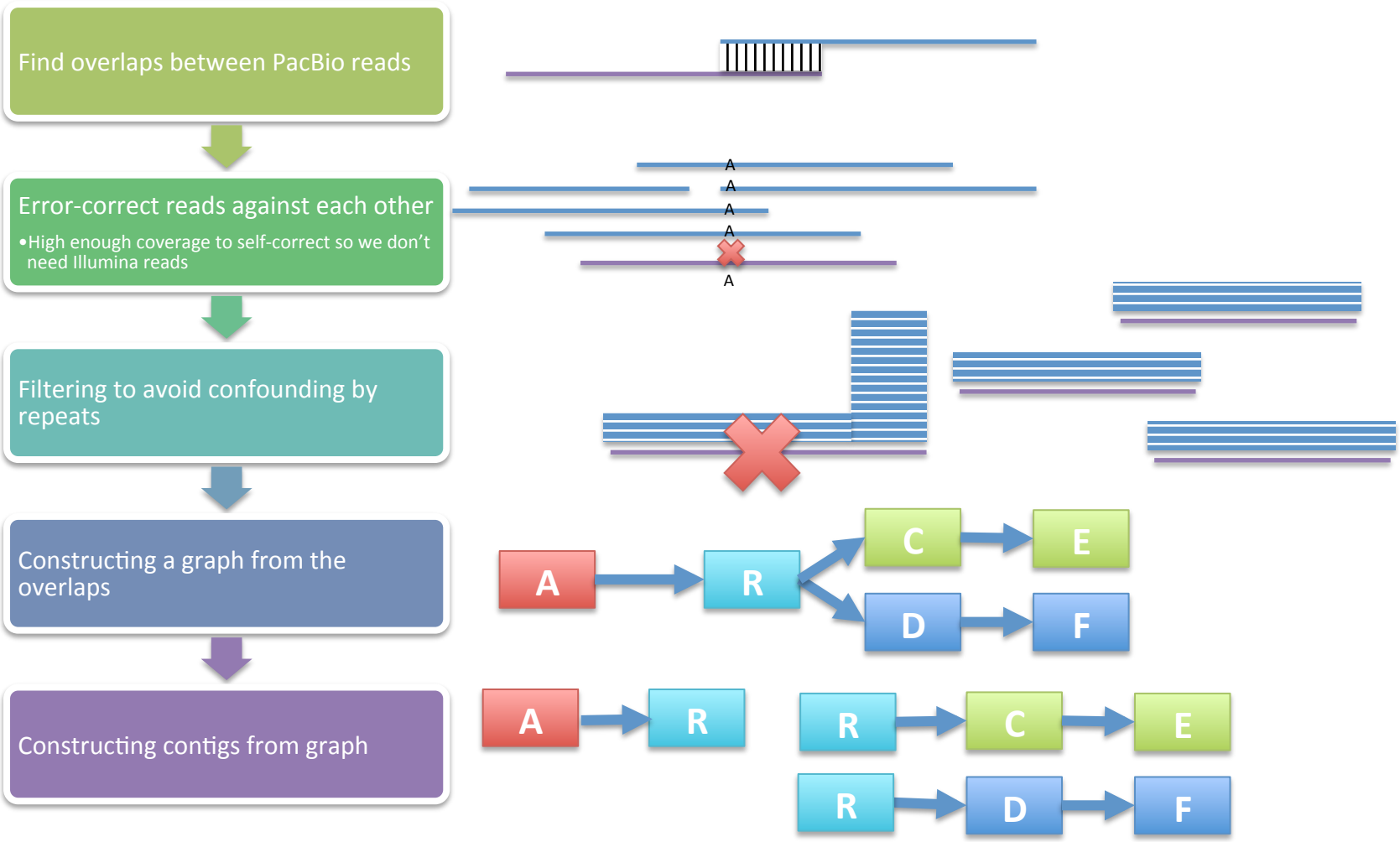
~ 11,000 local variants
50 bp < size < 10 kbp

Alignment-based



661 long-range variants
(>10kb distance)

Genome assembly with FALCON

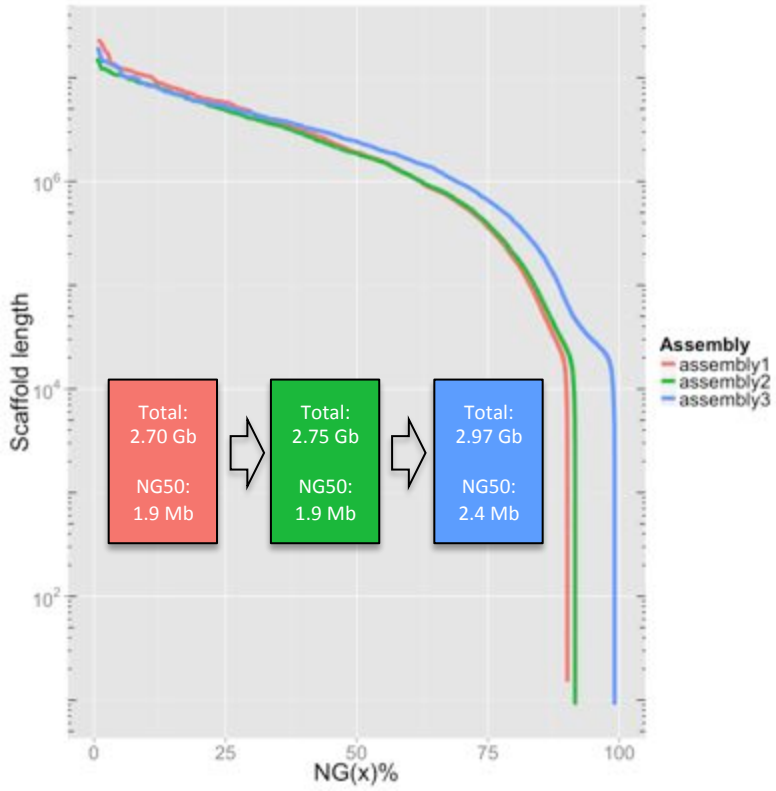


Iterations of Falcon assembly on DNAnexus

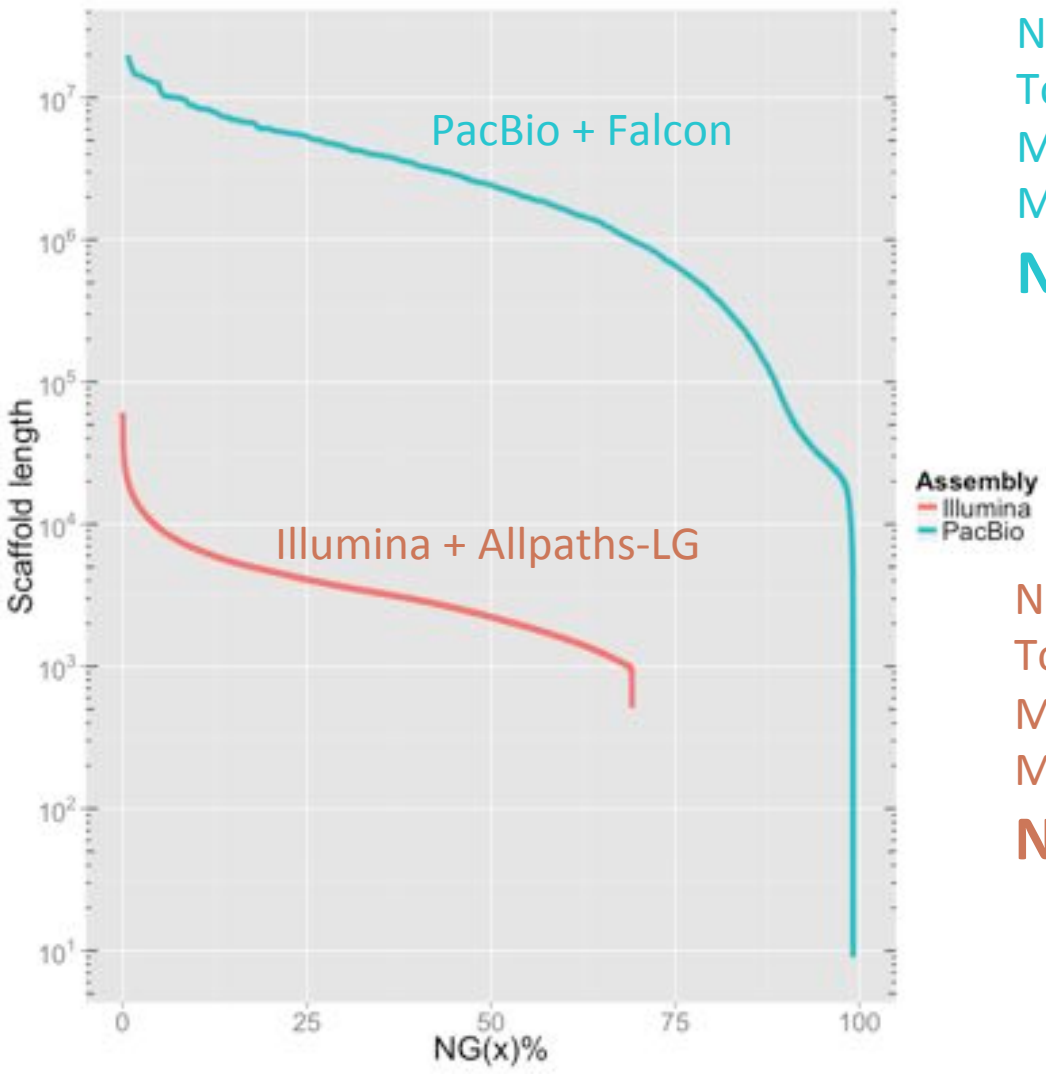
Standard FALCON parameters → Higher allowance for coverage differences → Require only 2X error-corrected coverage

Repeat filters in FALCON also remove aneuploid regions

Diploid regions have ~25X coverage --> low after error correction, especially if heterozygous



Assembly using PacBio yields far better contiguity



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

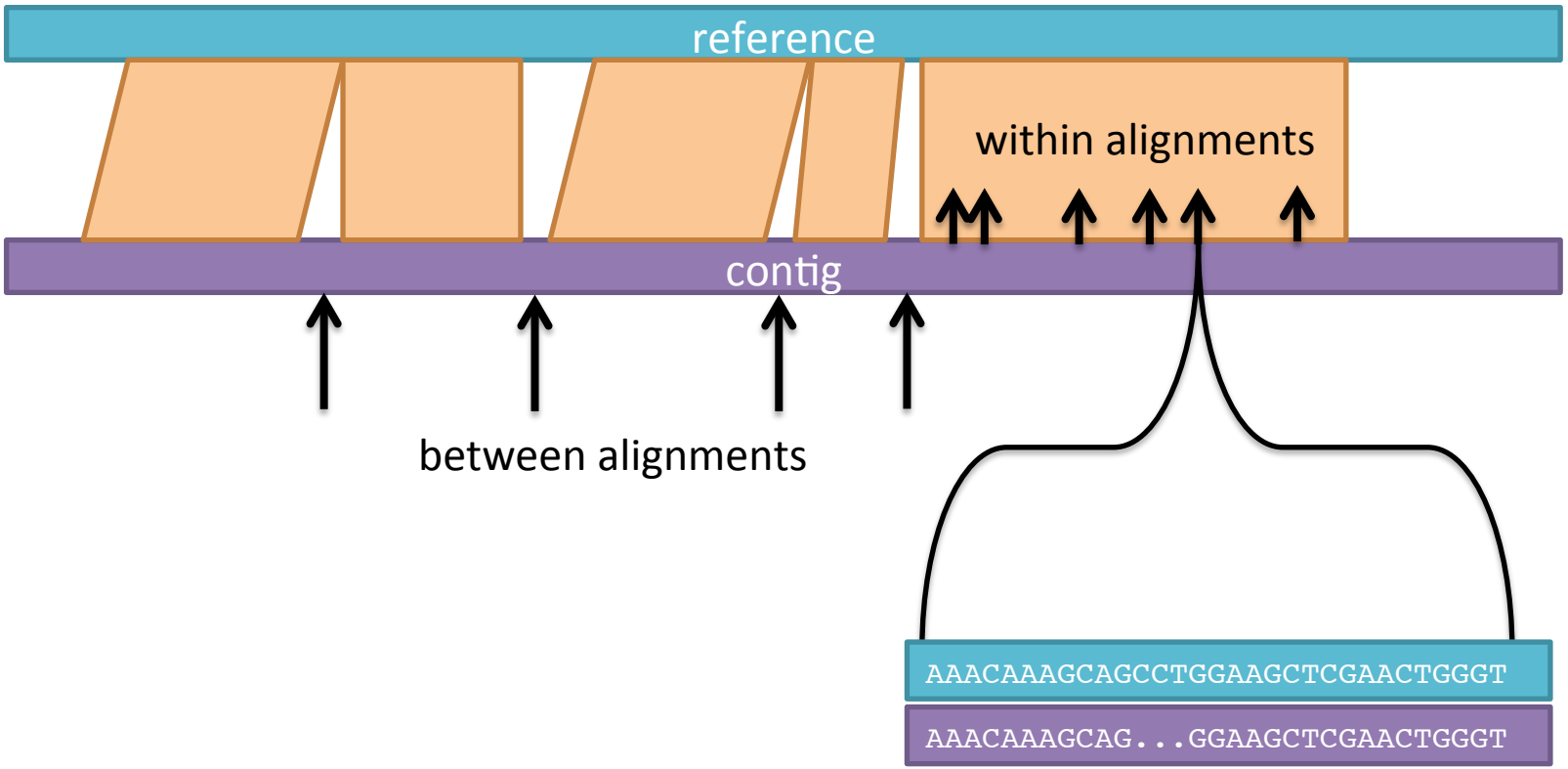
N50: 2.46 Mb

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

Variant detection from a genome assembly



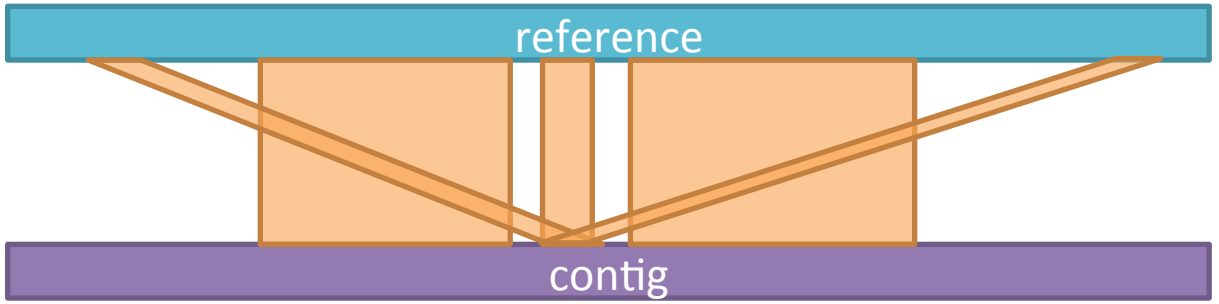
ABVC: Variants within alignments

Insertion

Deletion

ABVC: Unique length filtering is needed to prevent false positives due to repetitive elements

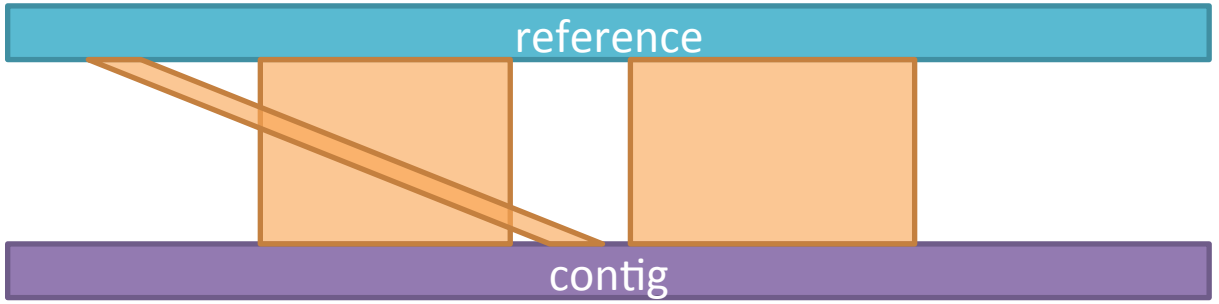
All alignments



Repeat maps in multiple locations

Traditional filtering by MUMmer

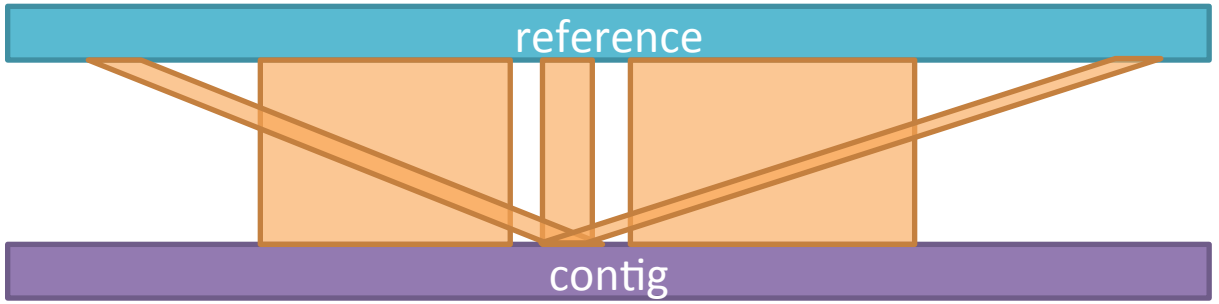
- choose the best alignment for each query
- random choice if multiple alignments with the same score



May be falsely called as a translocation

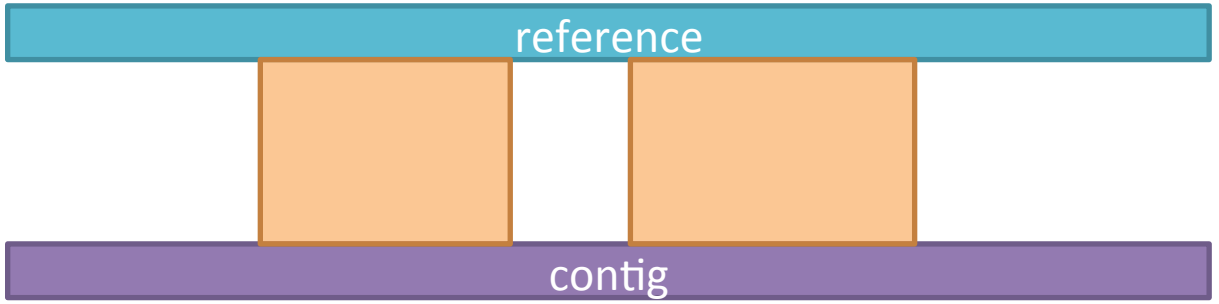
ABVC: Unique length filtering is needed to prevent false positives due to repetitive elements

All alignments



Repeat maps in multiple locations

Filtering based on requiring a certain length of unique alignment

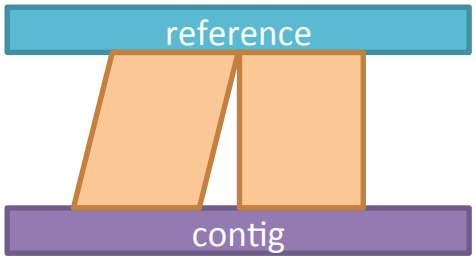


All repeats with less than 10kb of unique sequence are filtered out

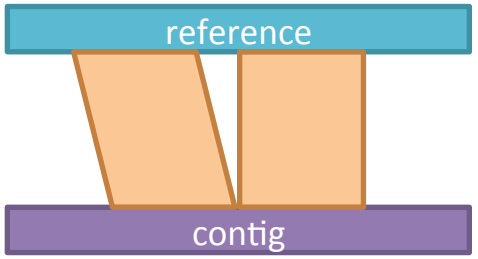
Types of variants detected by ABVC

Defined point

Insertion

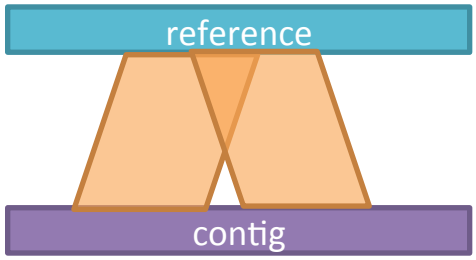


Deletion

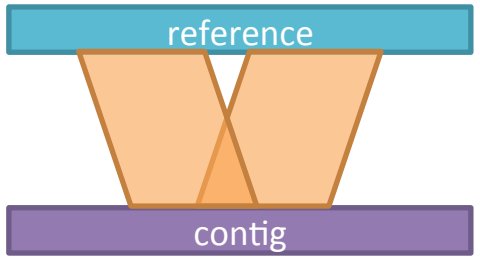


Overlapping alignments suggest tandem repeat

Tandem Expansions

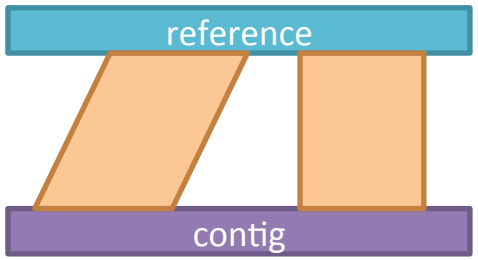


Tandem Contractions

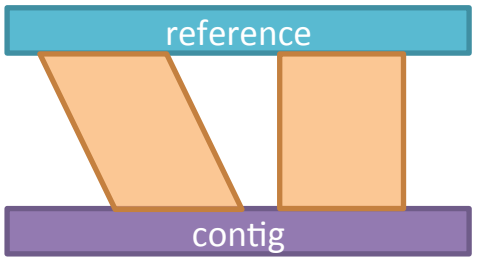


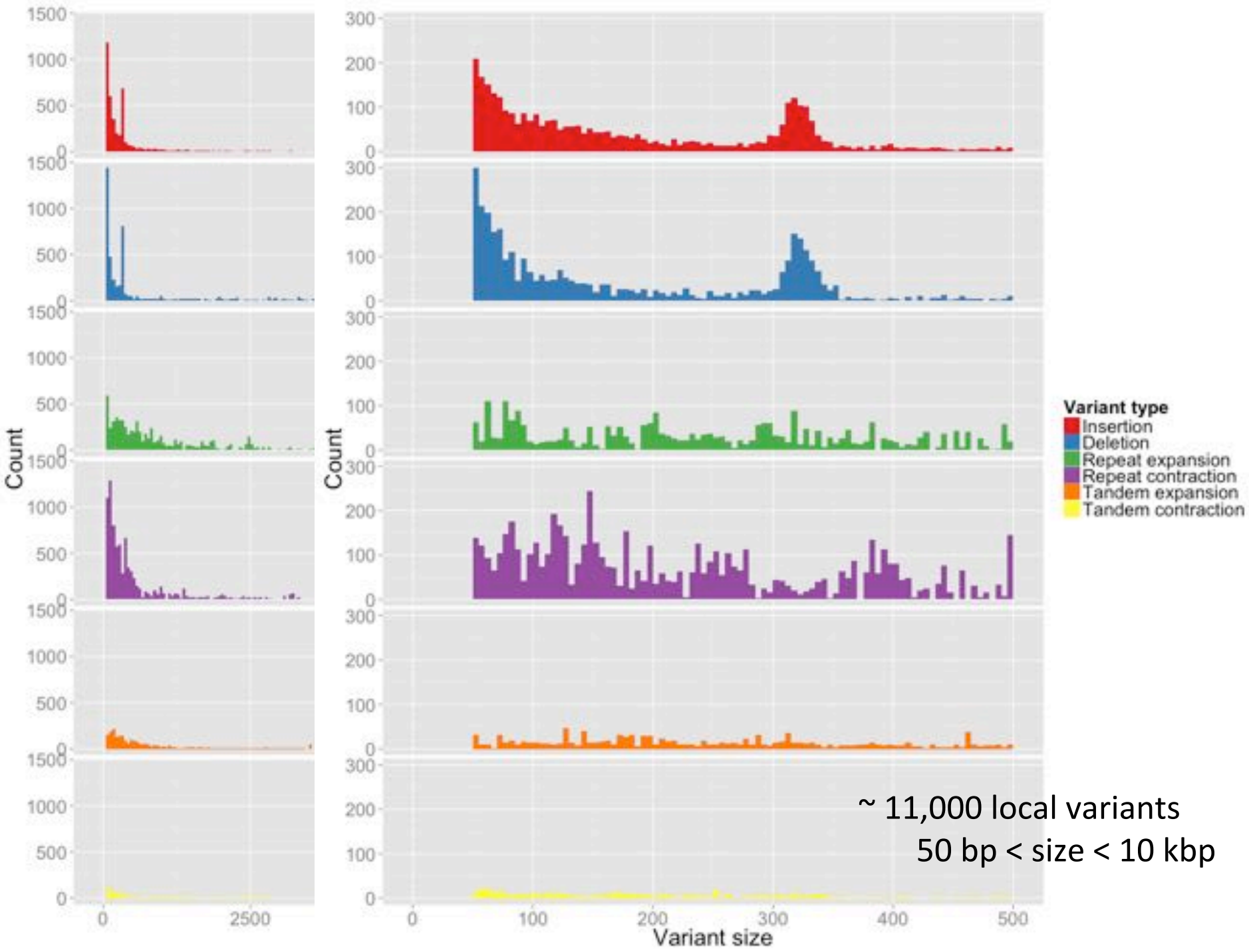
Gap where sequences do not align uniquely suggests a repeat

Repeat Expansions

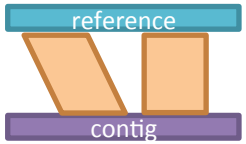


Repeat Contractions

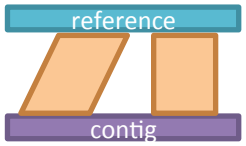



Count
Variant size
Variant type
Insertion
Deletion
Repeat expansion
Repeat contraction
Tandem expansion
Tandem contraction
~ 11,000 local variants
50 bp < size < 10 kbp

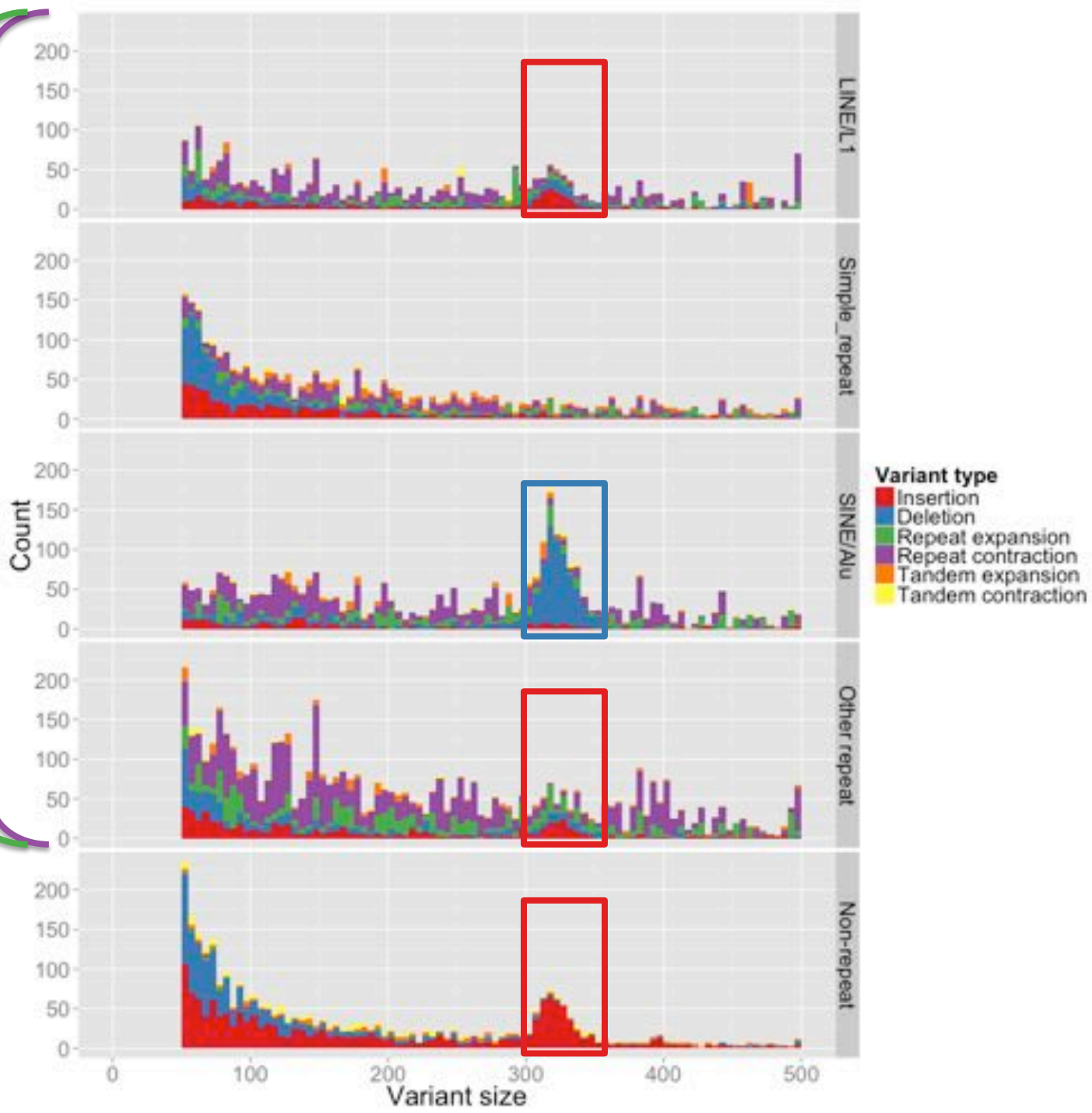
Repeat Contractions



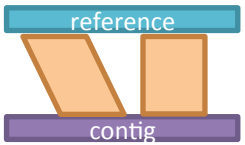
Repeat Expansions



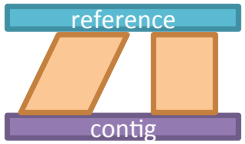
BLASTed 515 insertions:
427 (83%) of them matched Alu elements

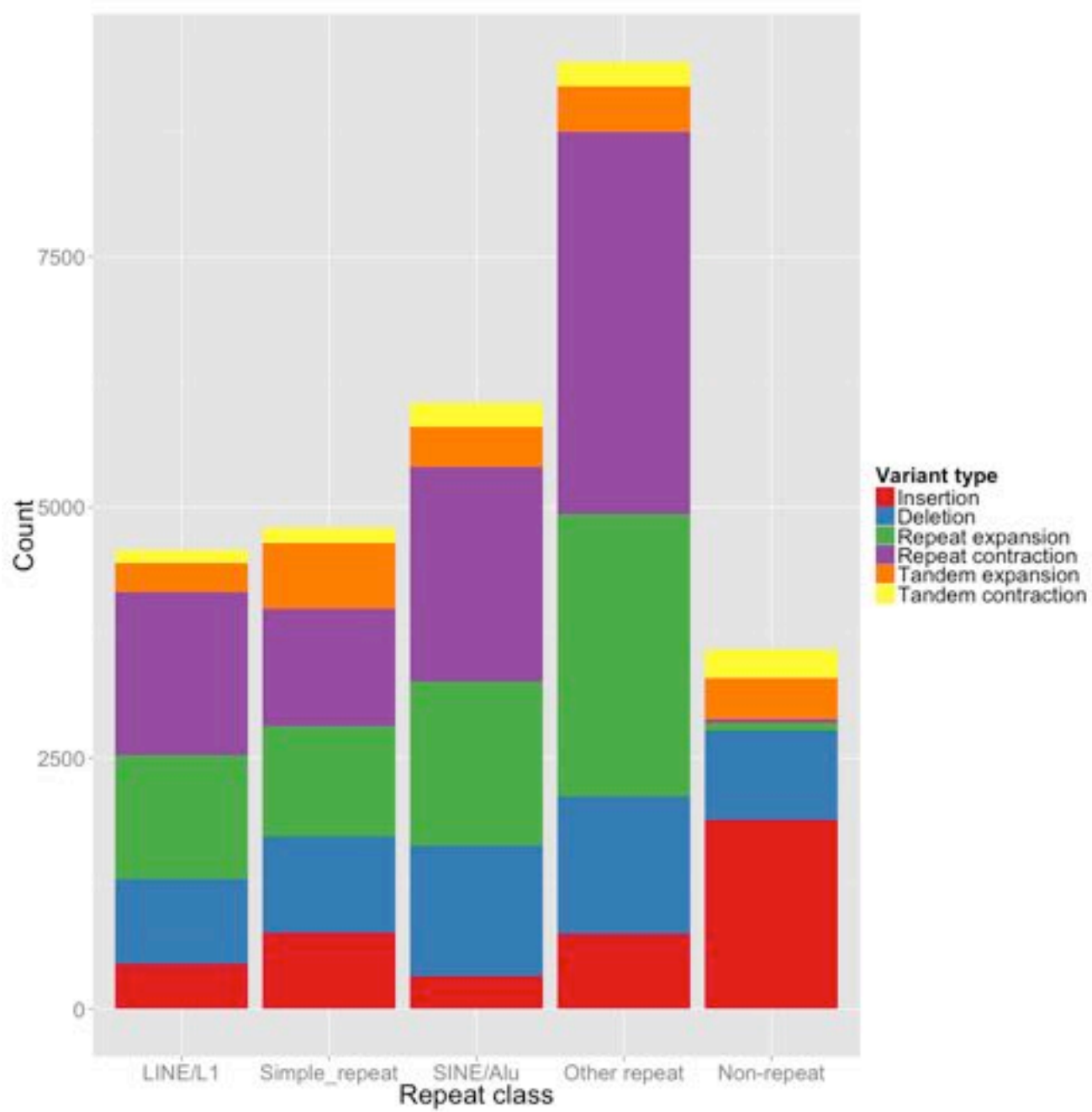


Repeat Contractions

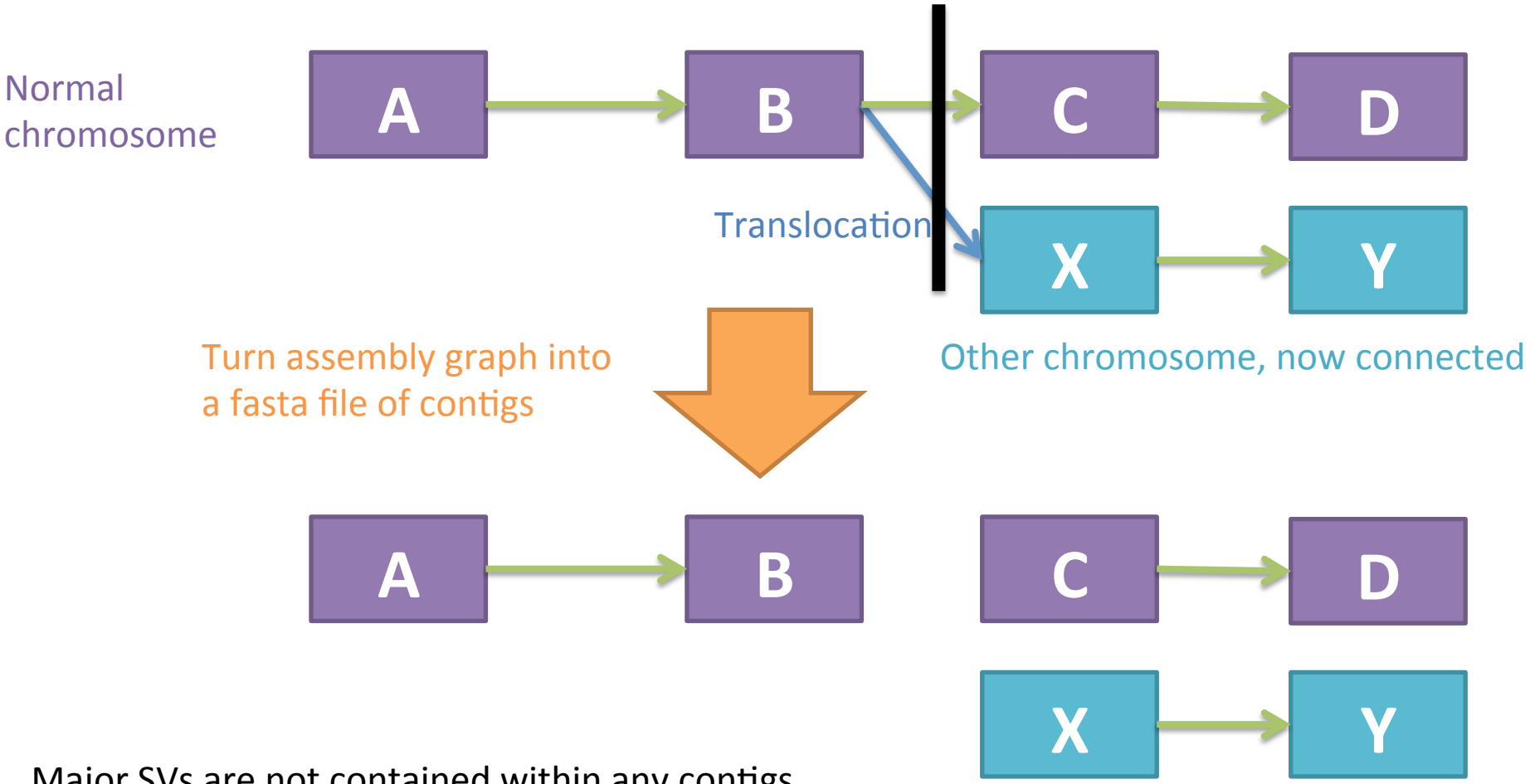


Repeat Expansions





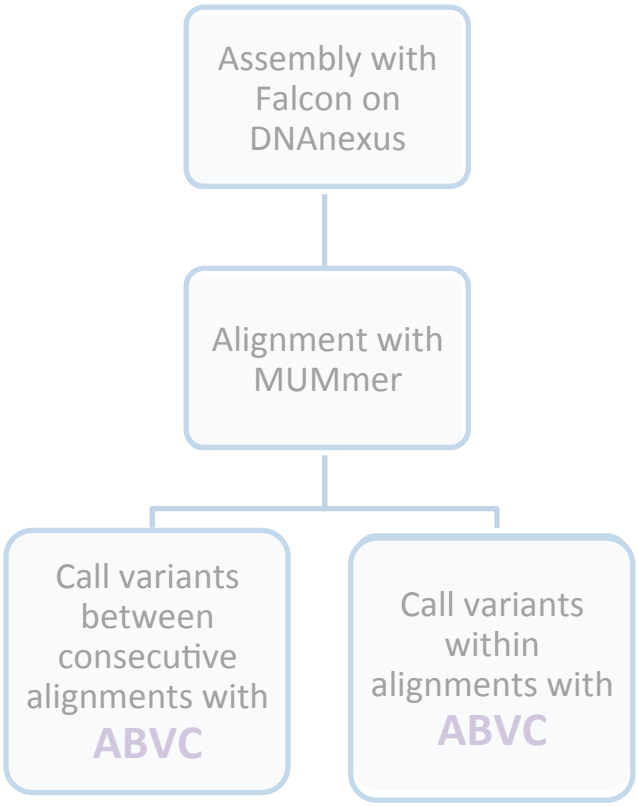
Why assembly doesn't capture long-range variants



Major SVs are not contained within any contigs

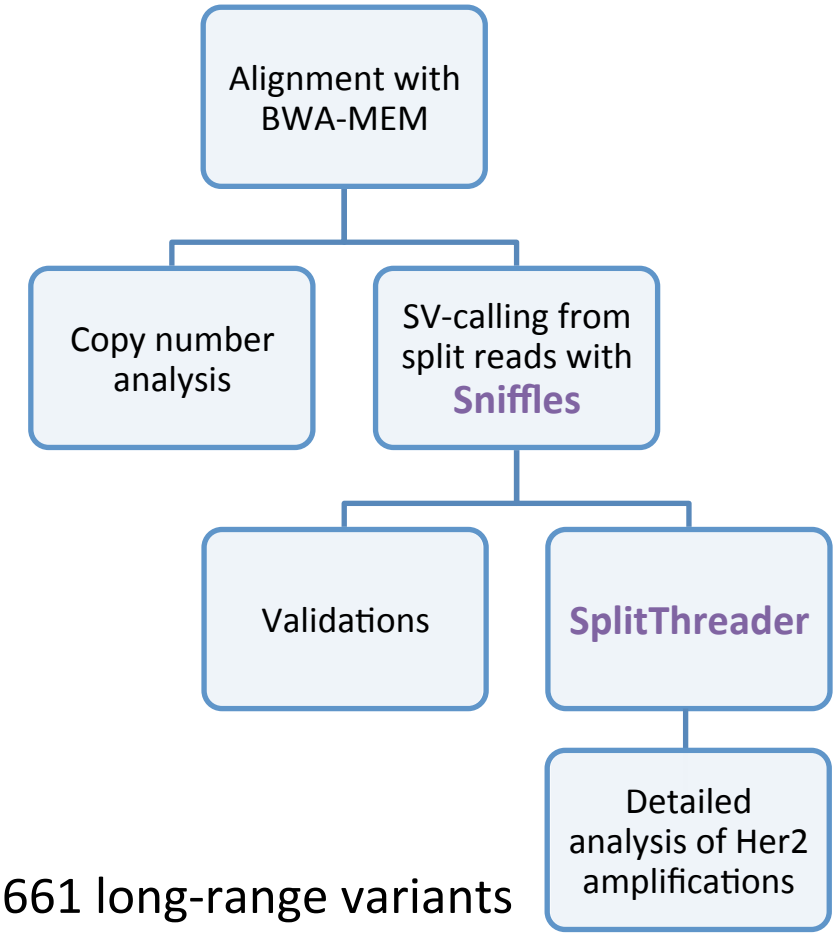
Genome structural analysis

Assembly-based



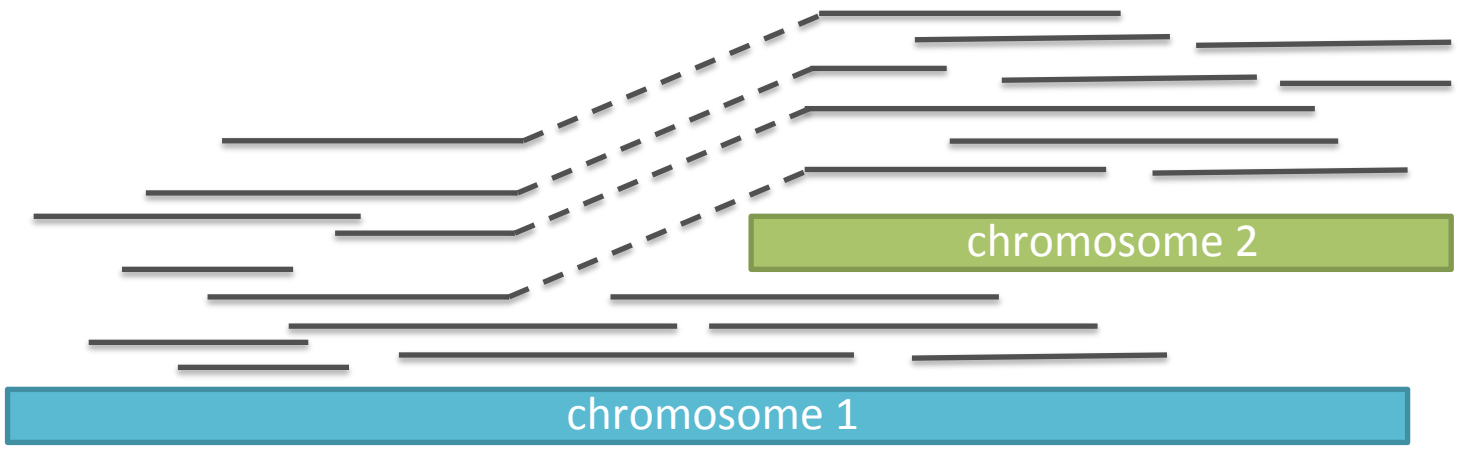
~ 11,000 local variants
50 bp < size < 10 kbp

Alignment-based



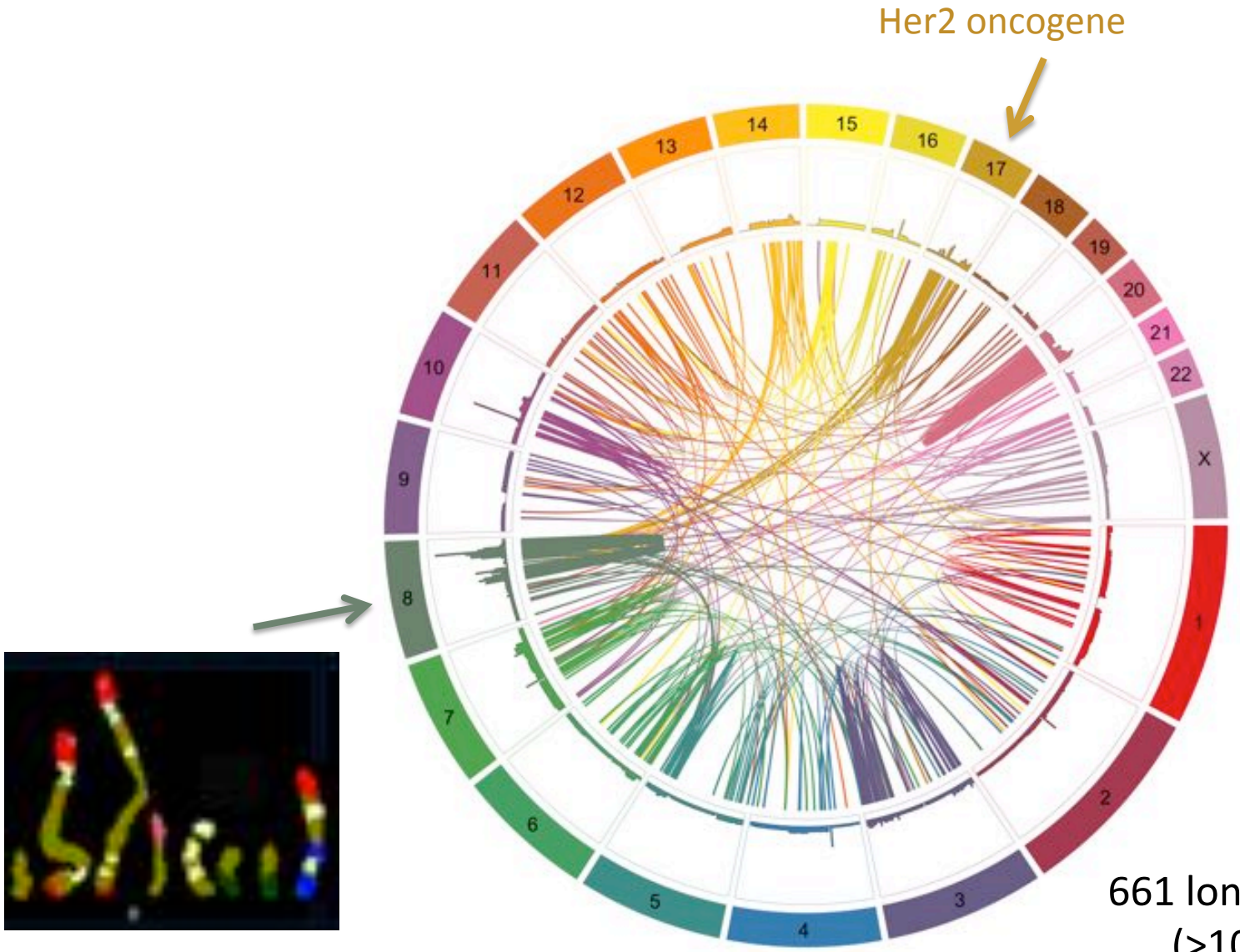
661 long-range variants
(>10kb distance)

Variant-calling from split-read alignment



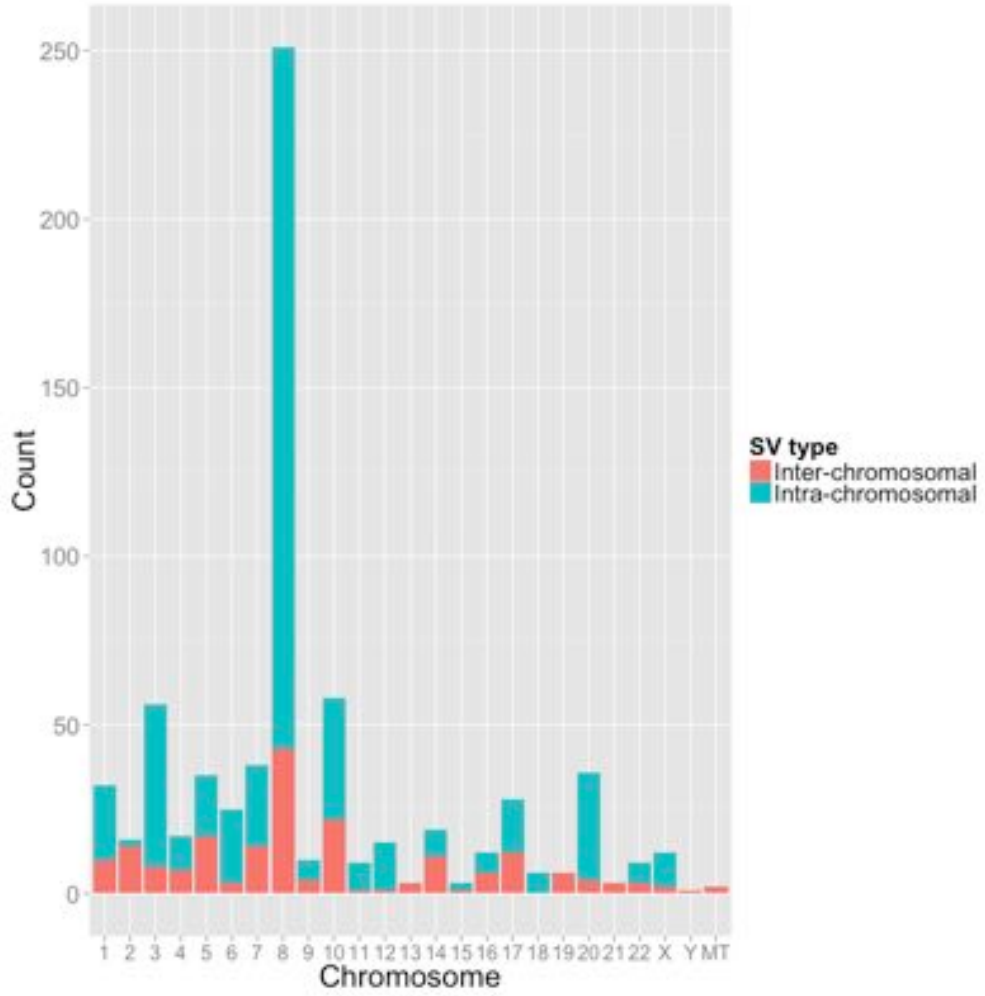
Software: Sniffles by Fritz Sedlazeck

Long-range structural variants found by Sniffles

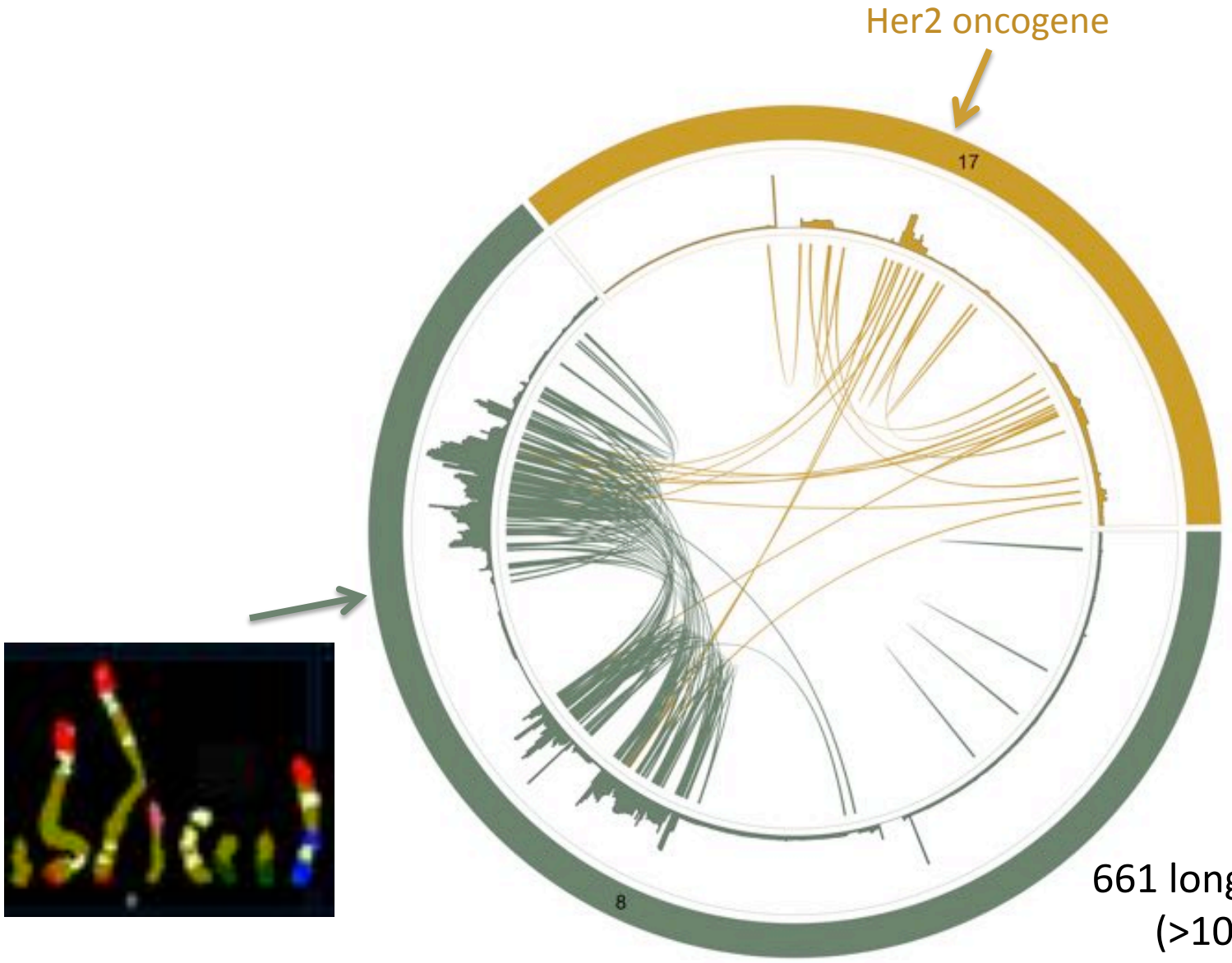


661 long-range variants
(>10kb distance)

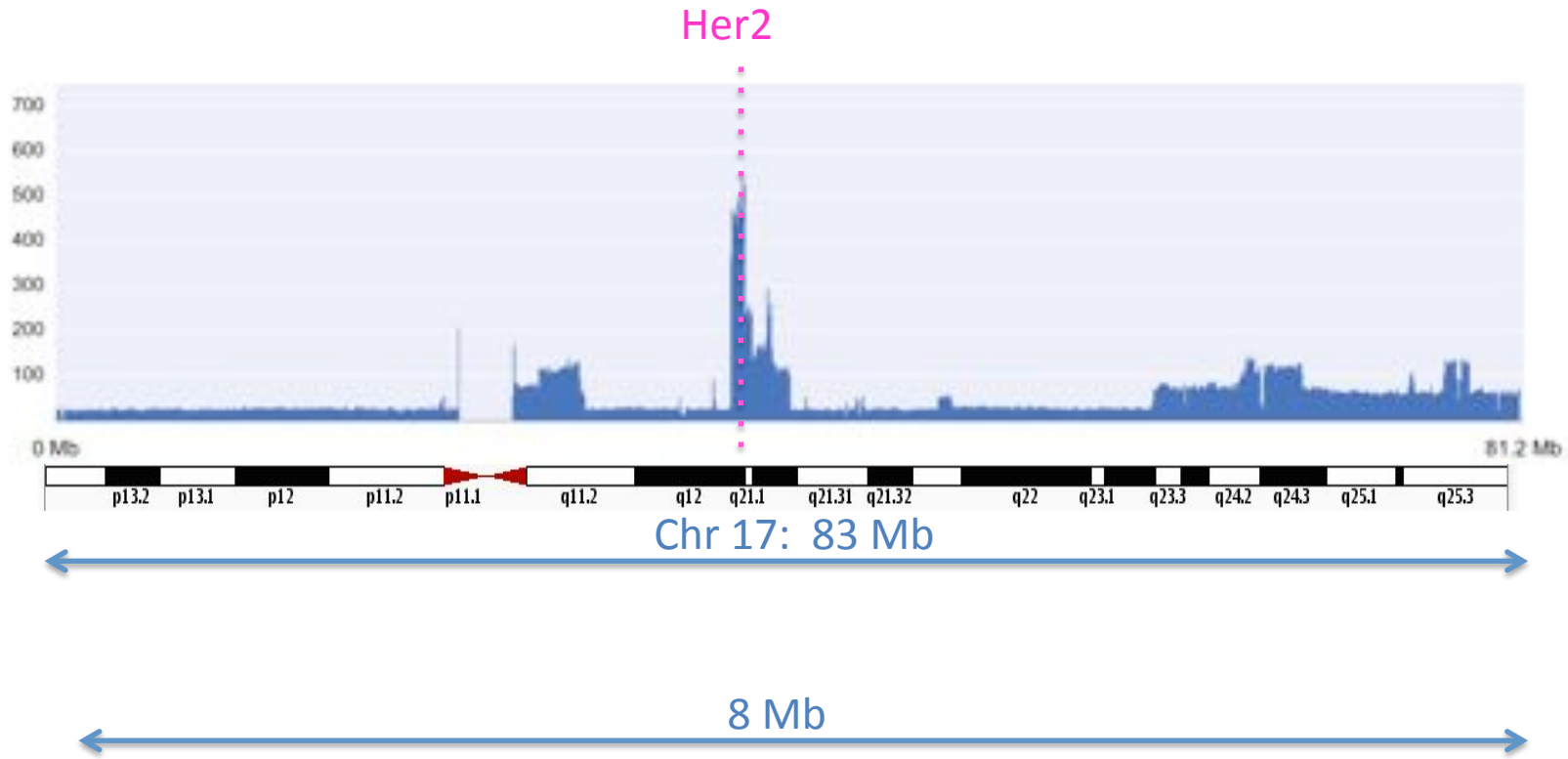
Chromosome 8 has the most intra- and inter-chromosomal long-range variants

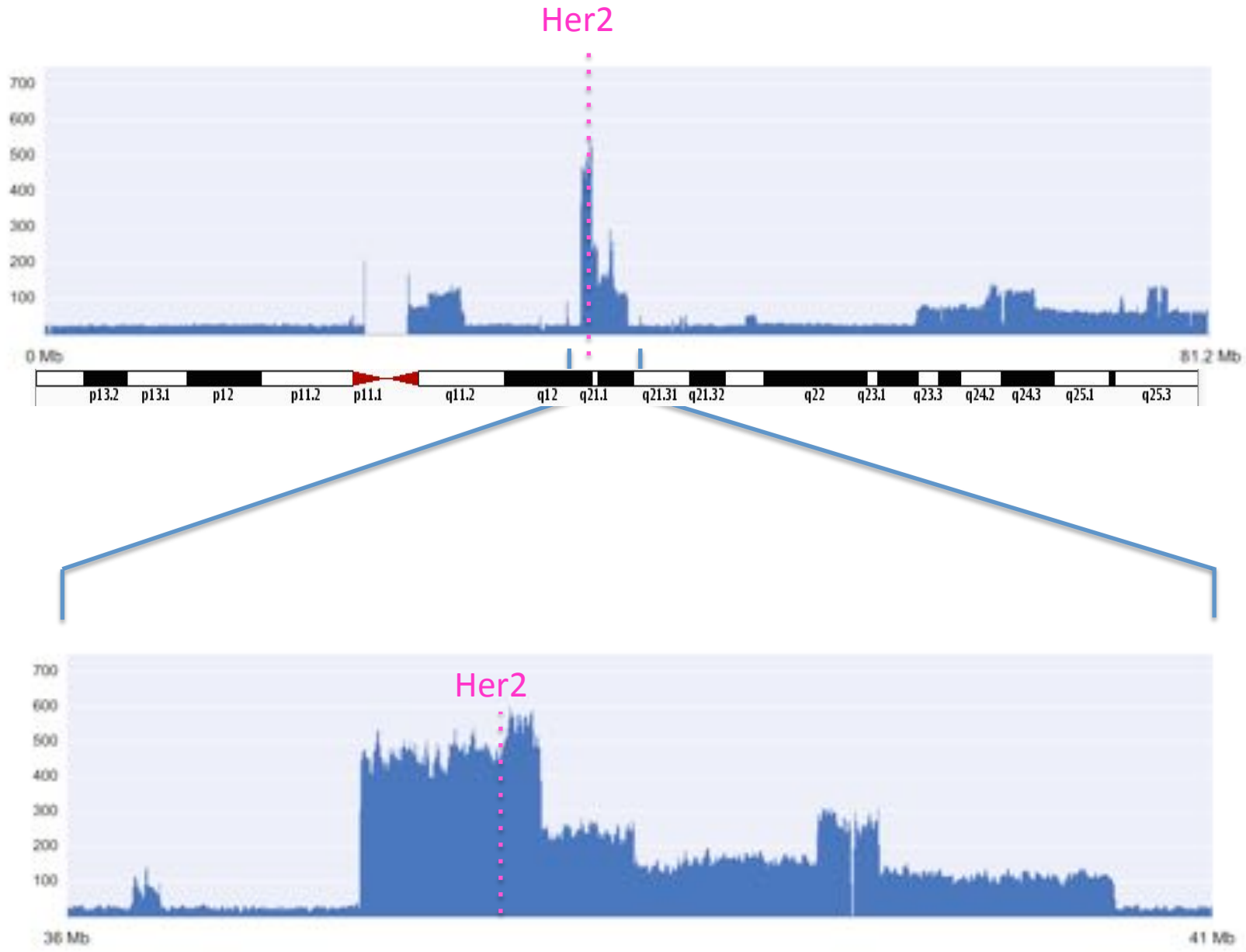


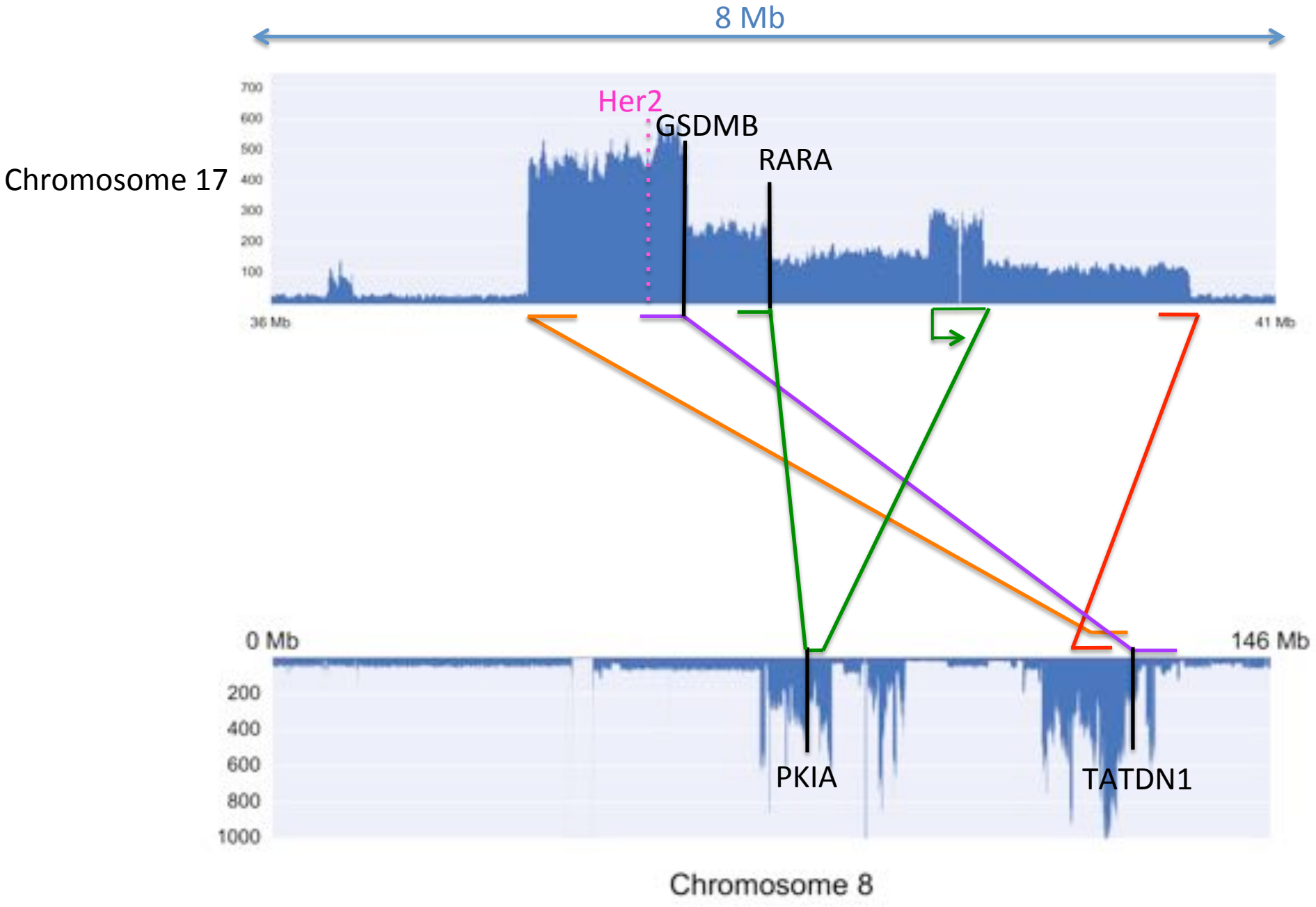
Long-range structural variants found by Sniffles



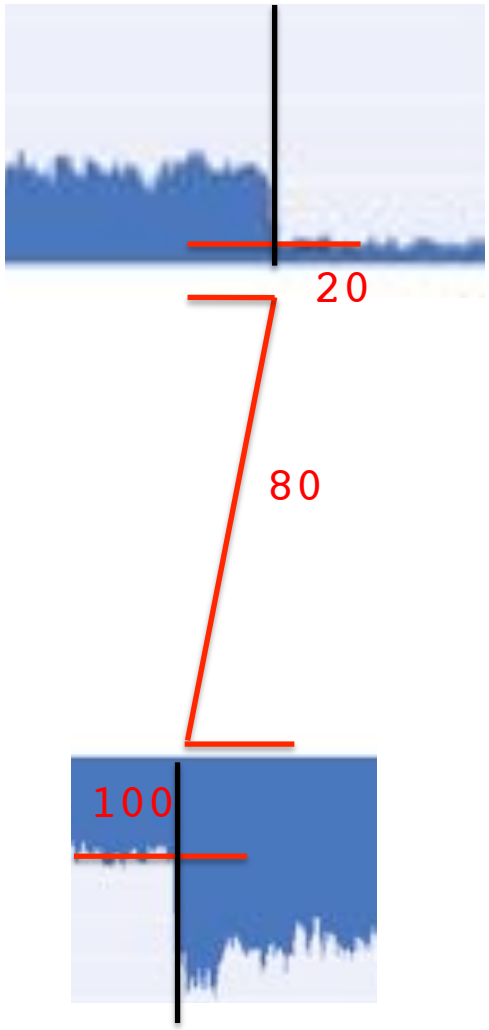
661 long-range variants
(>10kb distance)

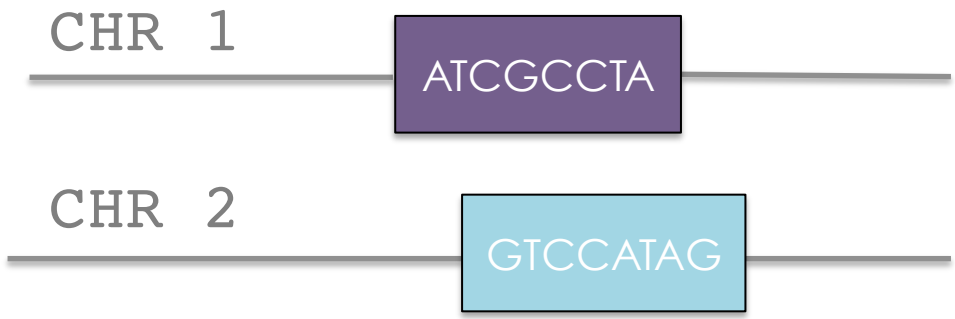

Her2
700
600
500
400
300
200
100
0 Mb
81.2 Mb
p13.2
p13.1
p12
p11.2
p11.1
q11.2
q12
q21.1
q21.31
q21.32
q22
q23.1
q23.3
q24.2
q24.3
q25.1
q25.3
Chr 17: 83 Mb
8 Mb

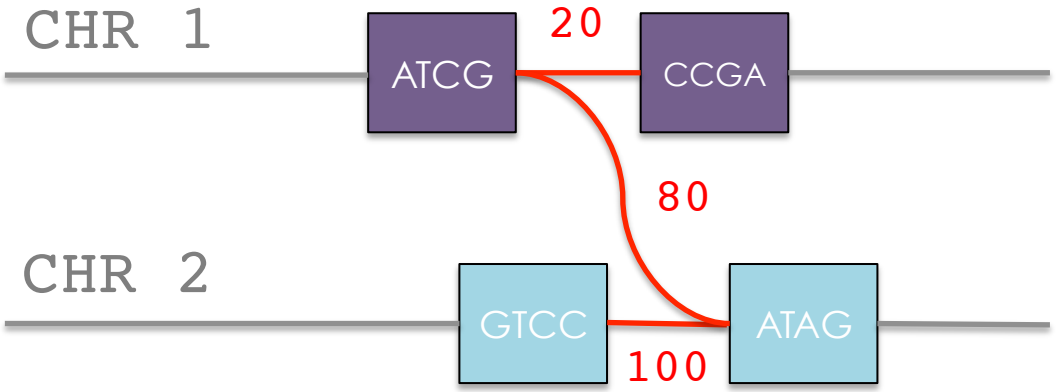

Her2
700
600
500
400
300
200
100
0 Mb
81.2 Mb
p13.2
p13.1
p12
p11.2
p11.1
q11.2
q12
q21.1
q21.31
q21.32
q22
q23.1
q23.3
q24.2
q24.3
q25.1
q25.3
Her2
700
600
500
400
300
200
100
36 Mb
41 Mb

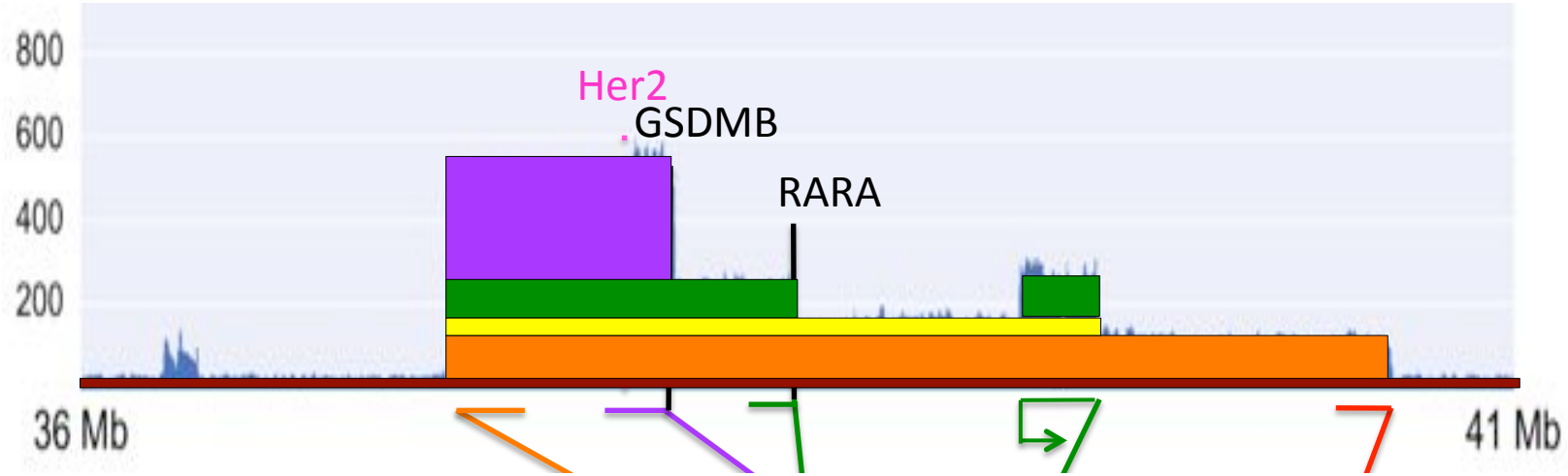

8 Mb
Chromosome 17
700
600
500
400
300
200
100
36 Mb
41 Mb
Her2
GSDMB
RARA
0 Mb
146 Mb
200
400
600
800
1000
PKIA
TATDN1
Chromosome 8

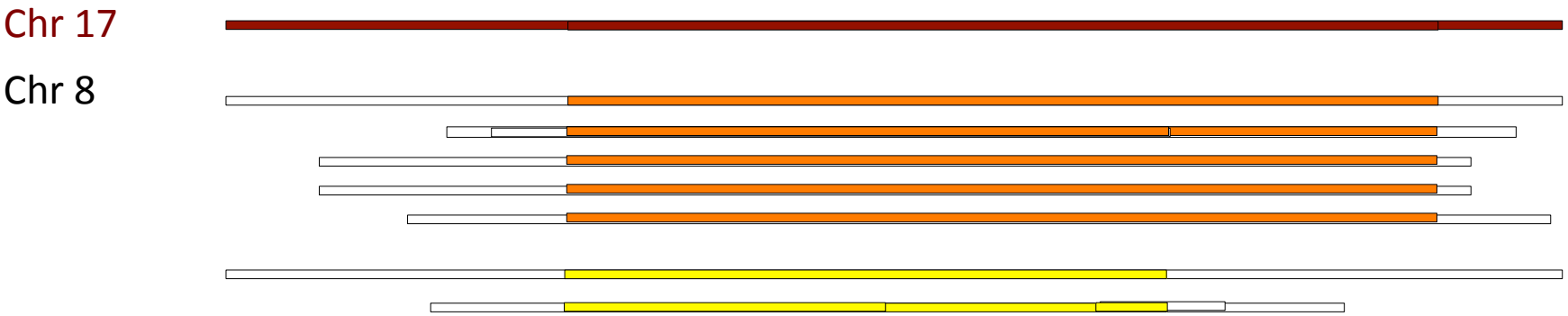
SplitThreader:
Graphical threading to retrace complex history of rearrangements in cancer genomes





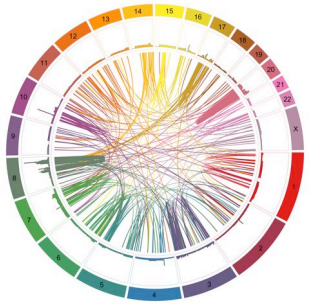






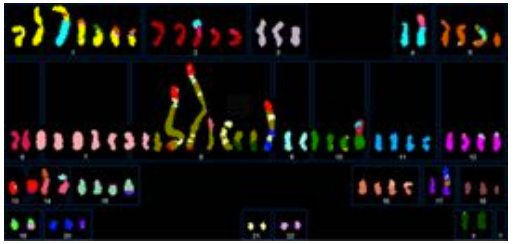
1. Healthy chromosome 17
2. Translocation into chromosome 8
3. Translocation within chromosome 8
4. Complex variant and inverted duplication within chromosome 8
5. Translocation within chromosome 8

Bridging the gap



variant-calling from sequencing

?

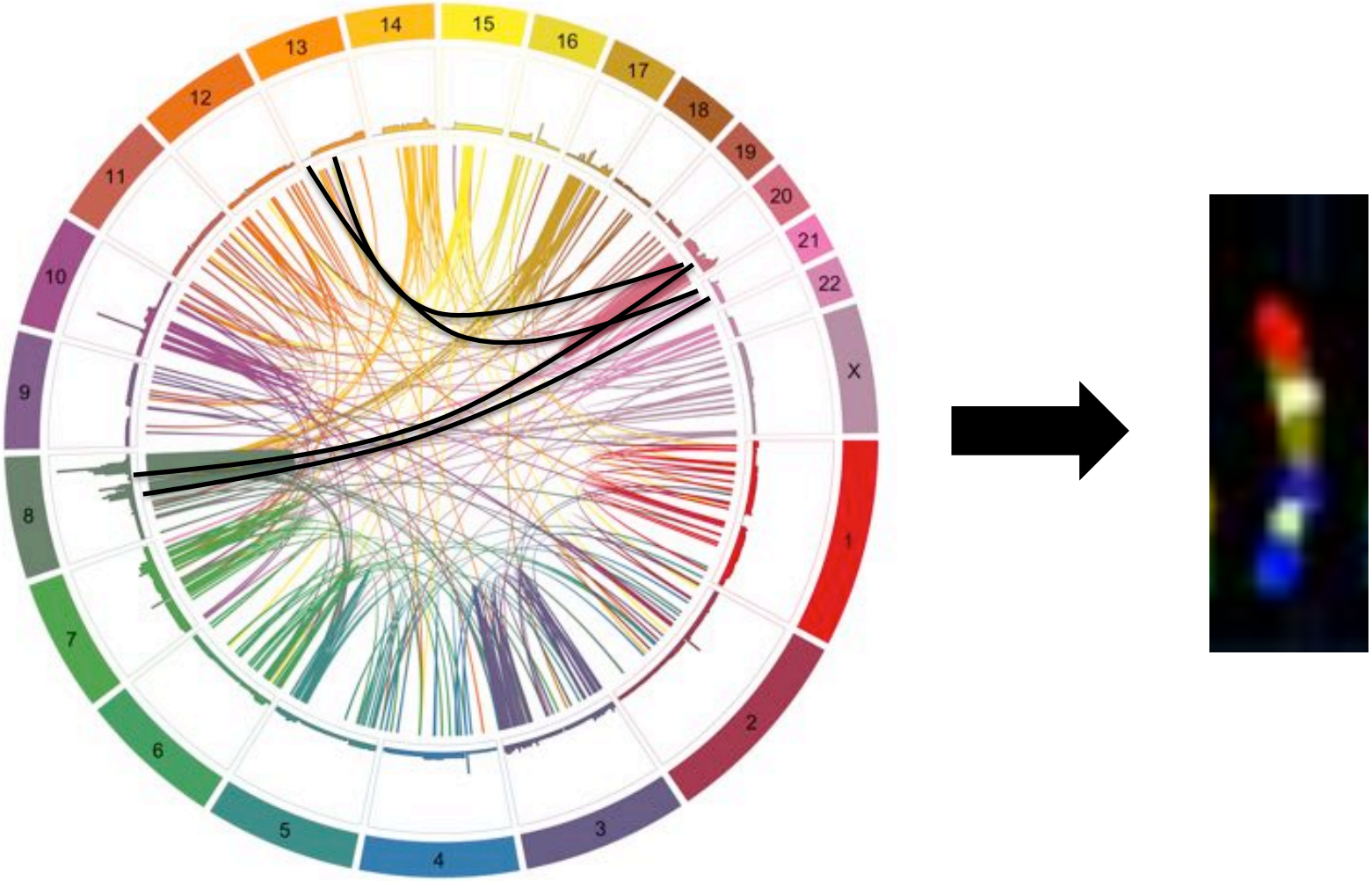


big-picture view from karyotyping

Context

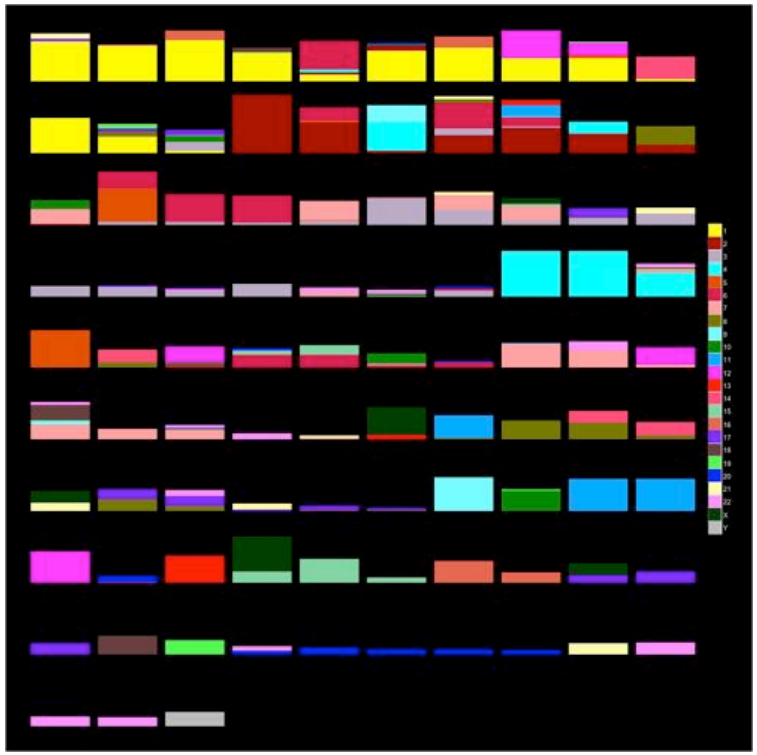
Resolution

Threading through the whole-genome graph to produce a synthetic karyotype

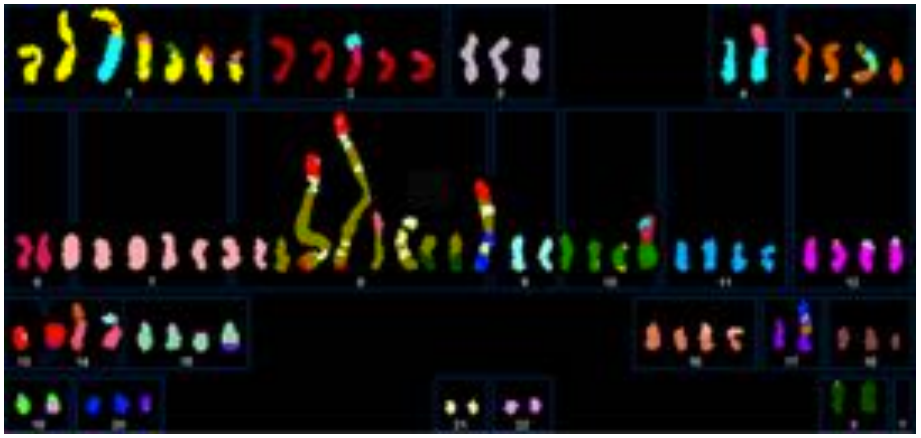


Synthetic karyotype with SplitThreader

Preliminary SplitThreader synthetic karyotype



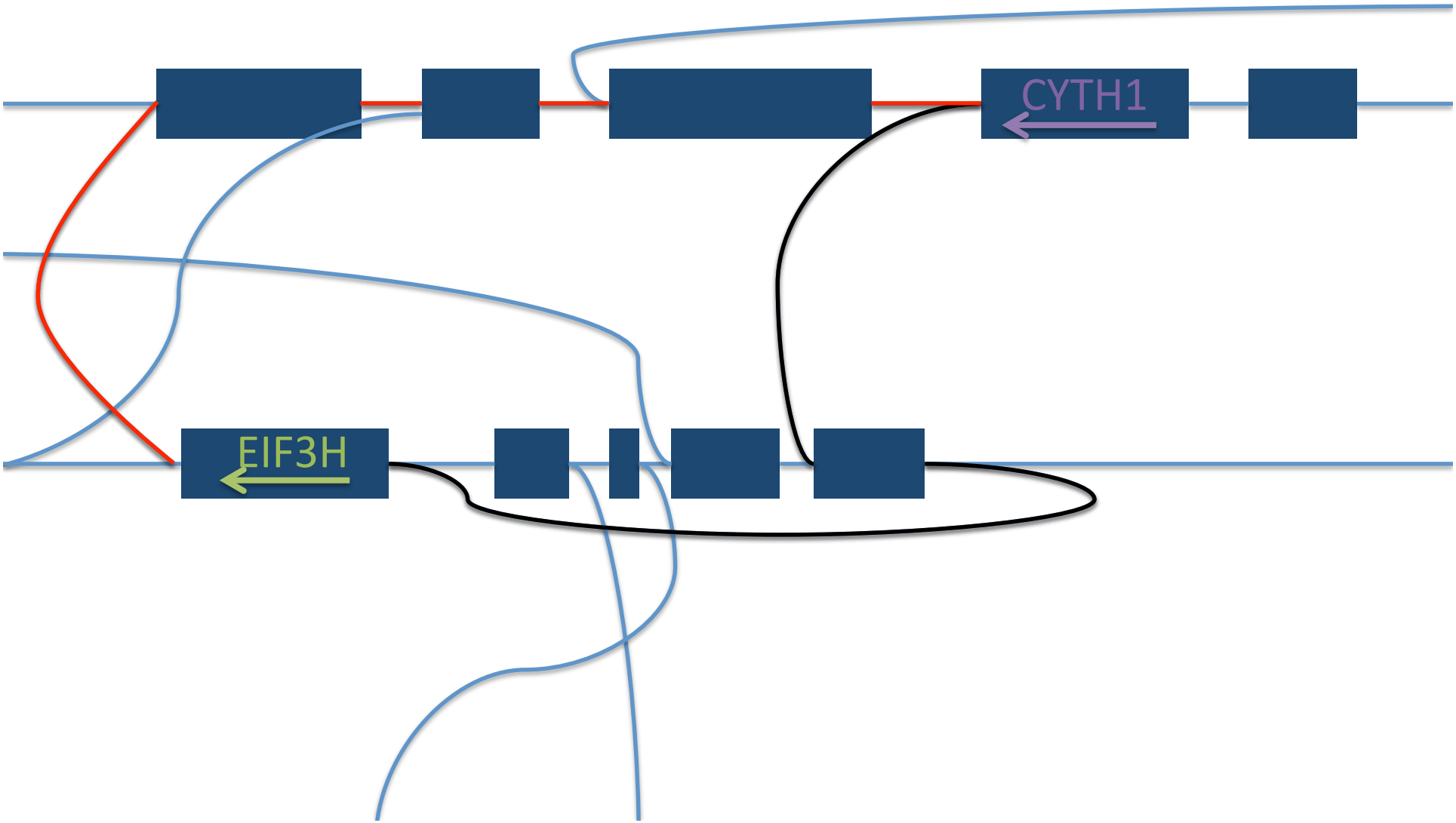
Real karyotype



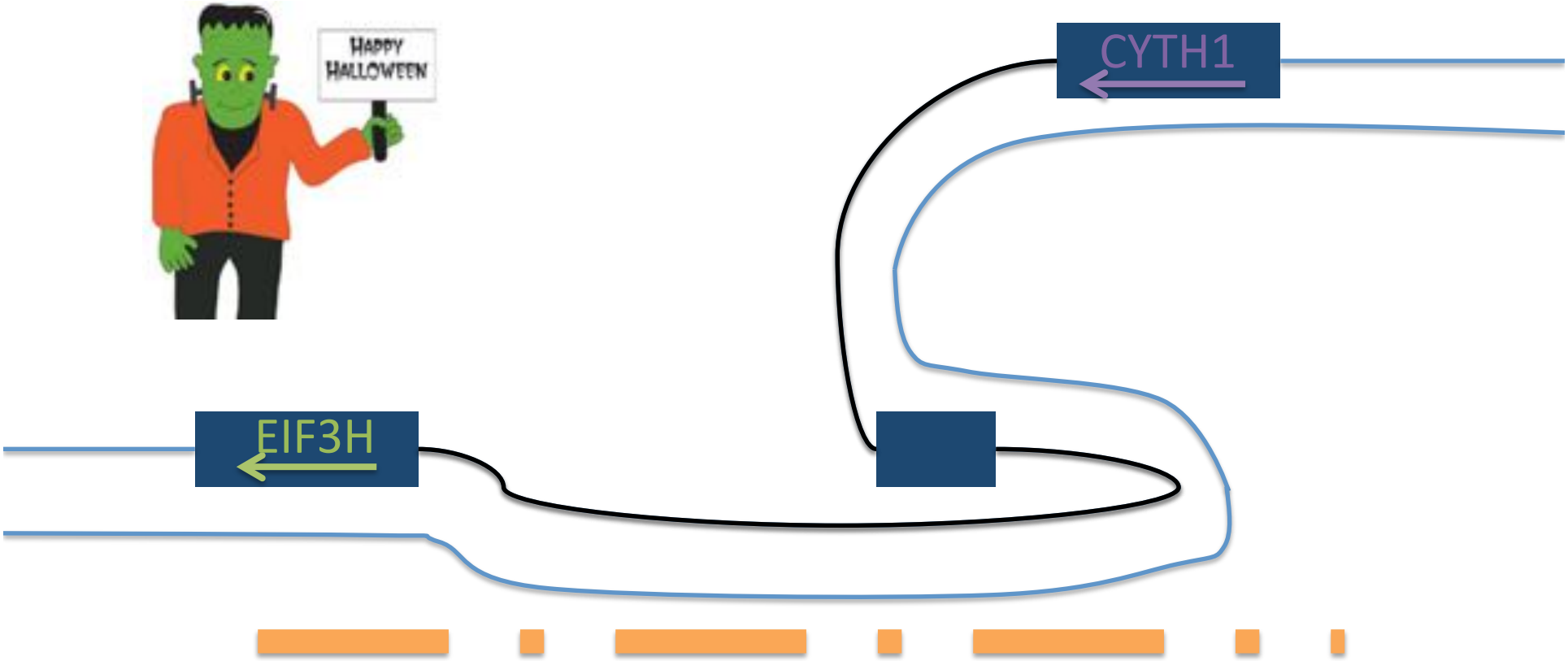
Transcriptome analysis with IsoSeq: Long-read RNA sequencing

- 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

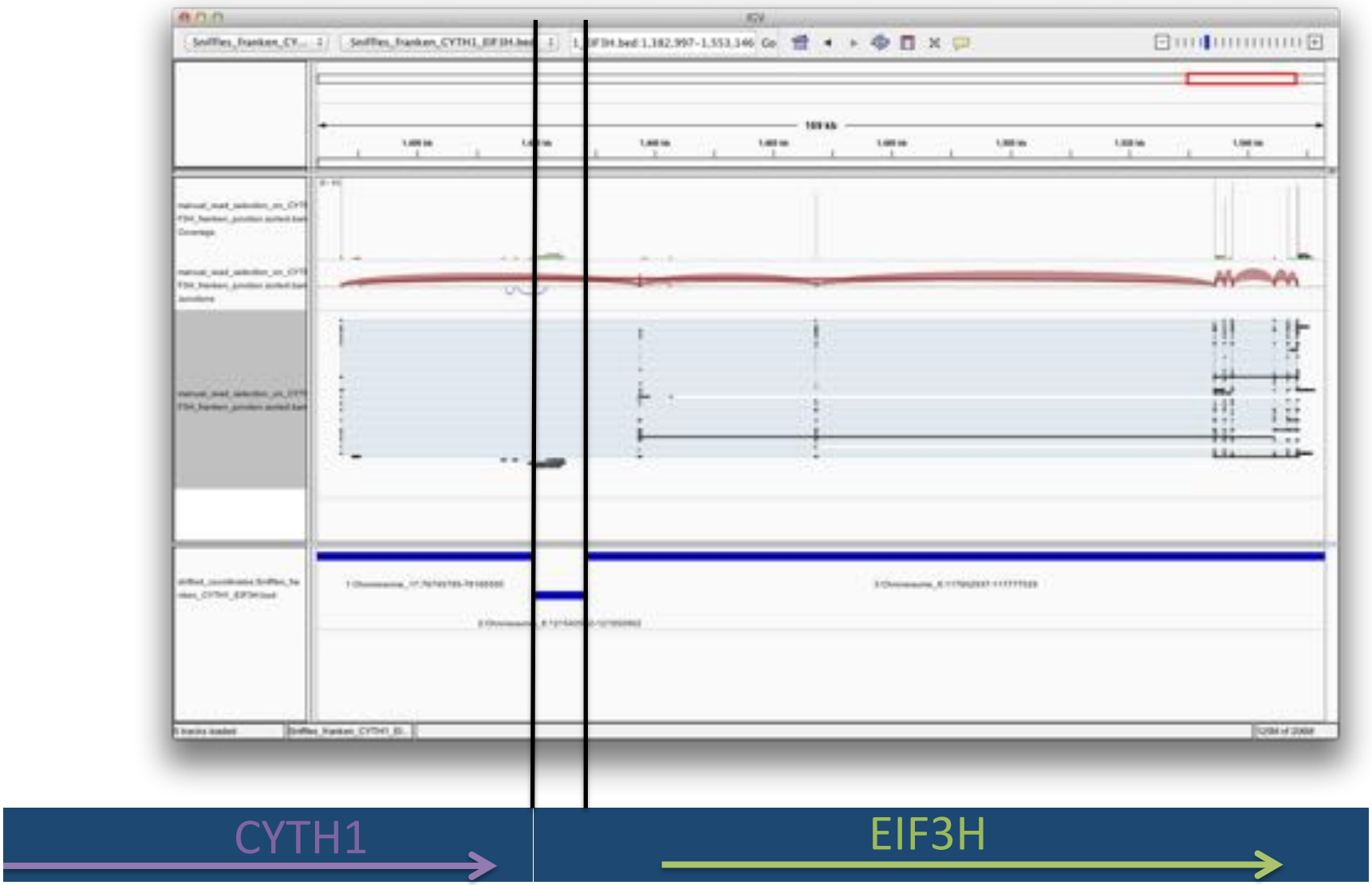
CYTH1-EIF3H gene fusion in the SplitThreader graph



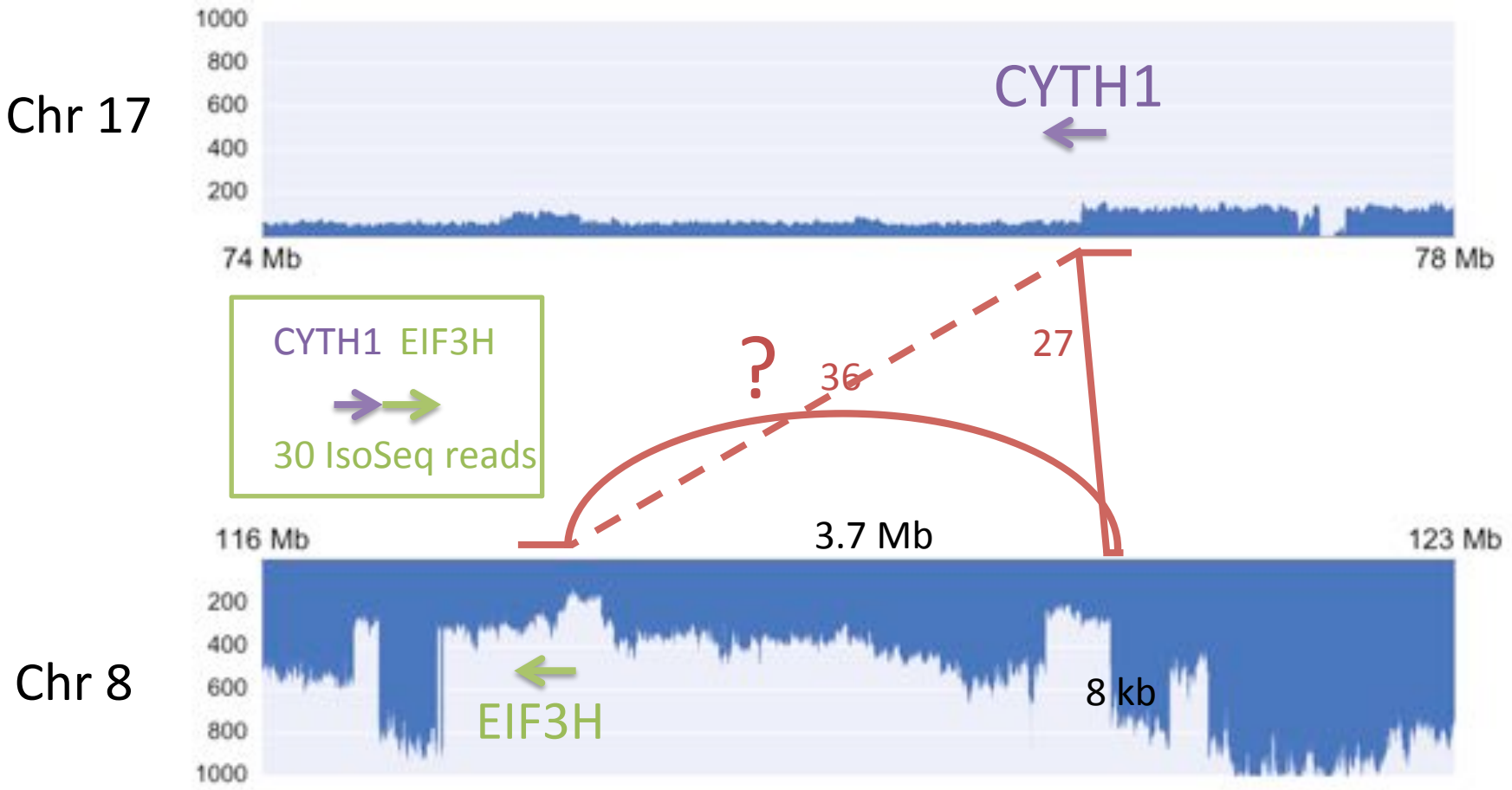
CYTH1-EIF3H gene fusion in the SplitThreader graph



Frankensteining the CYTH1-EIF3H gene fusion

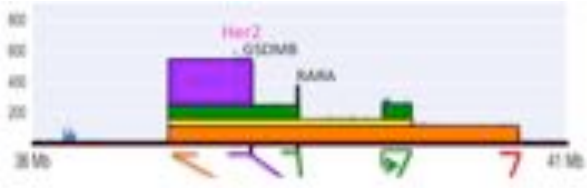


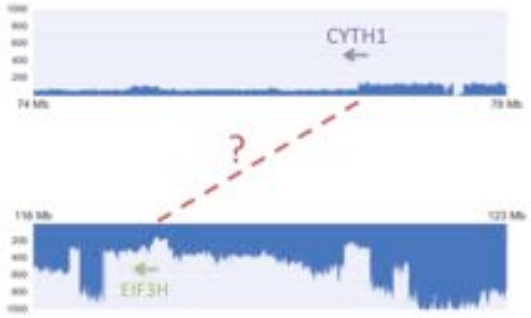
CYTH1-EIF3H gene fusion



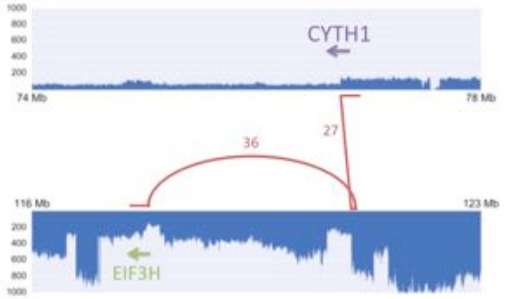
The genome informs the transcriptome

Explain amplifications





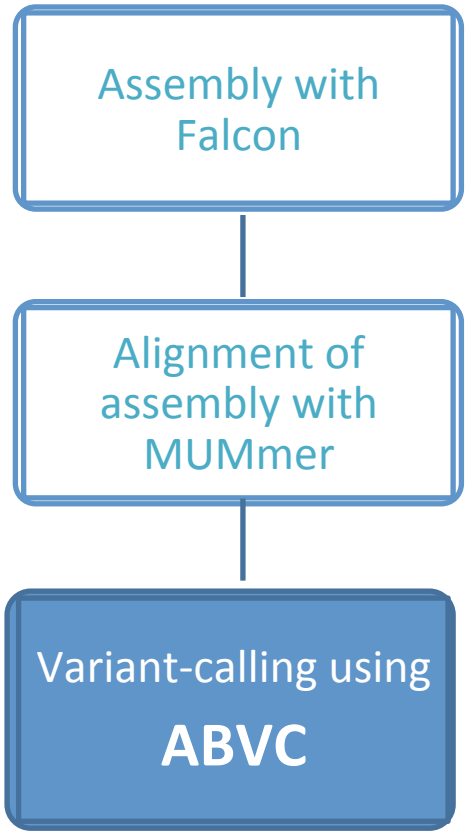
Trace gene fusions



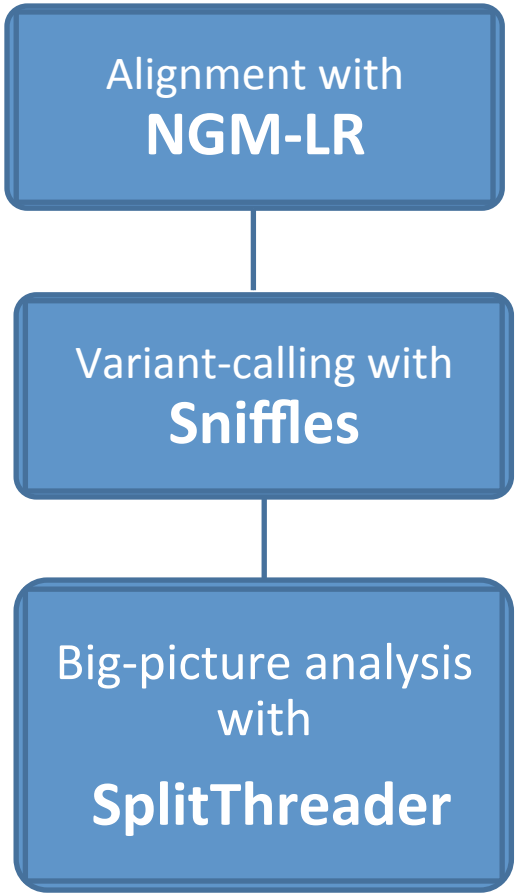
Data and additional results: http://schatzlab.cshl.edu/data/skbr3/

New software in development for long-read genome analysis

Assembly-based analysis

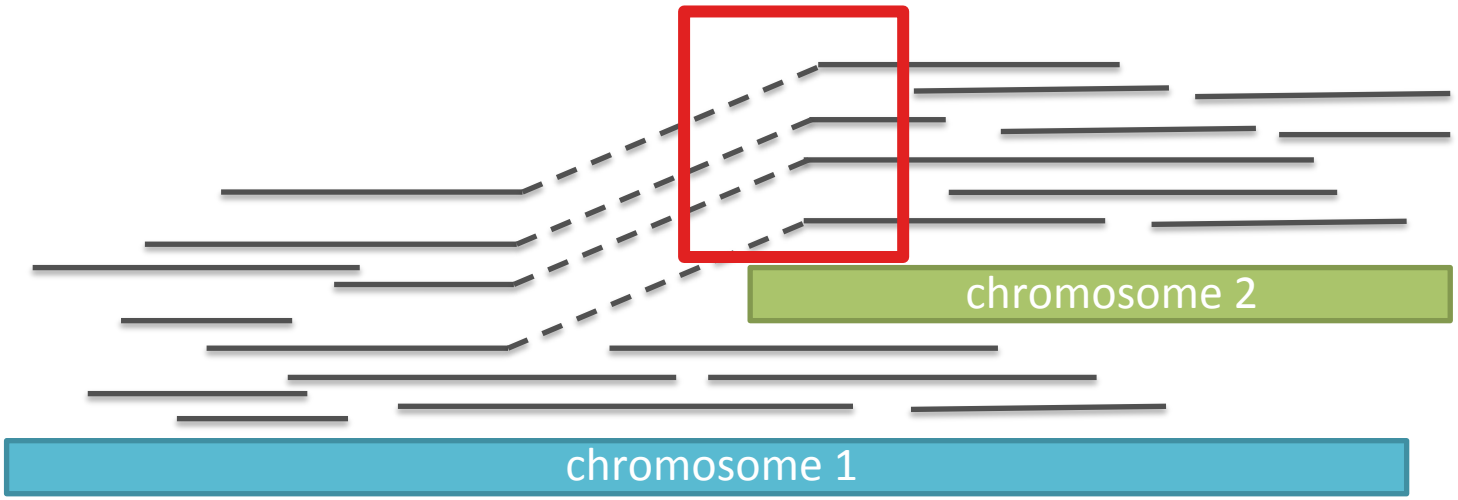


Alignment-based analysis




Pac Bio
Illumina

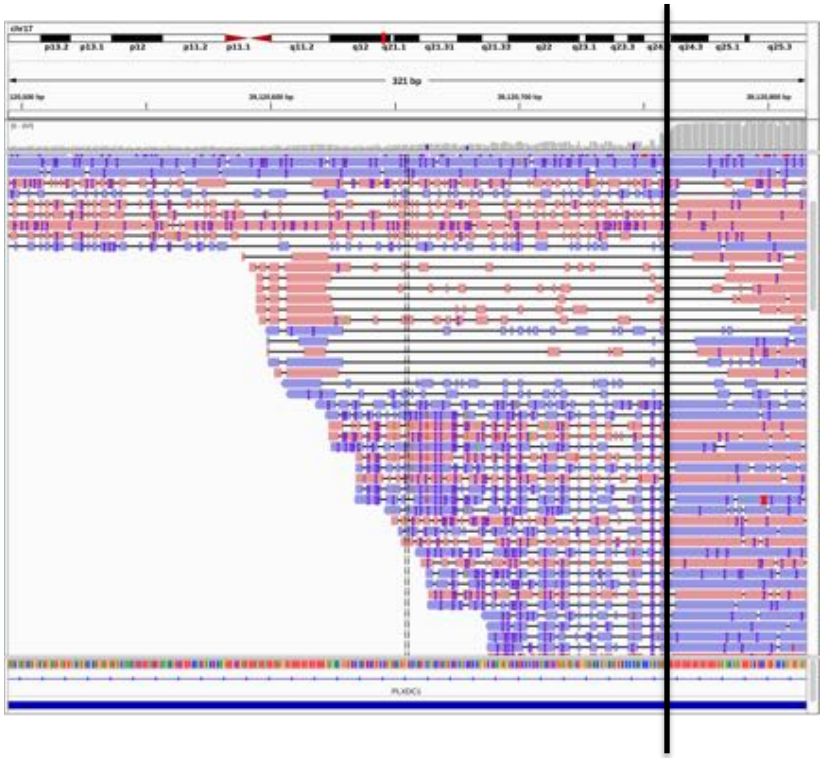
Zooming in on the breakpoint



New aligner NGM-LR narrows down the breakpoint to base-pair resolution

Philipp Rescheneder

BWA-MEM



NGM-LR

One side of an interchromosomal translocation

NGM-LR also enables better small variant calling

BWA-MEM

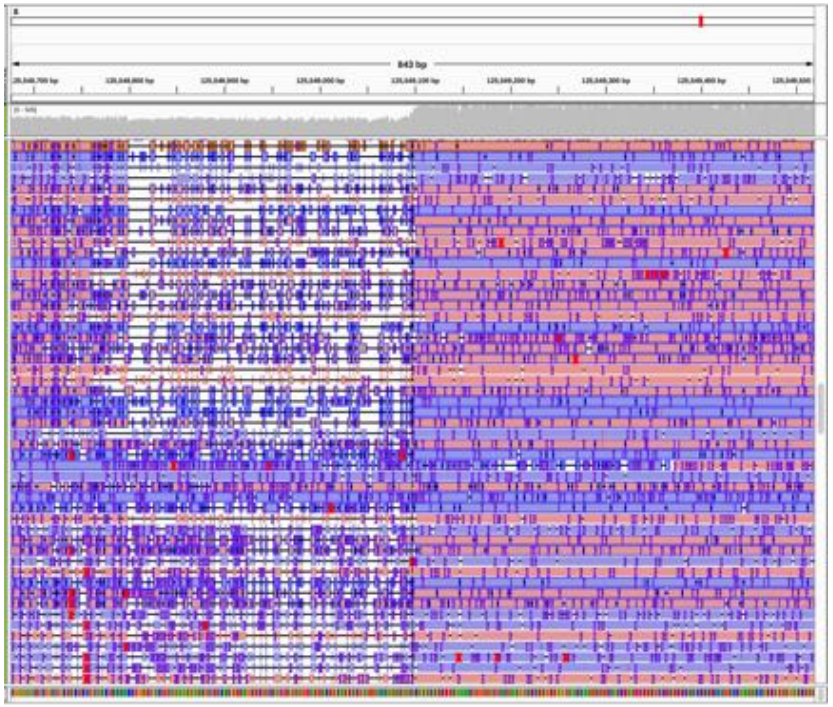
deletion

NGM-LR

deletion

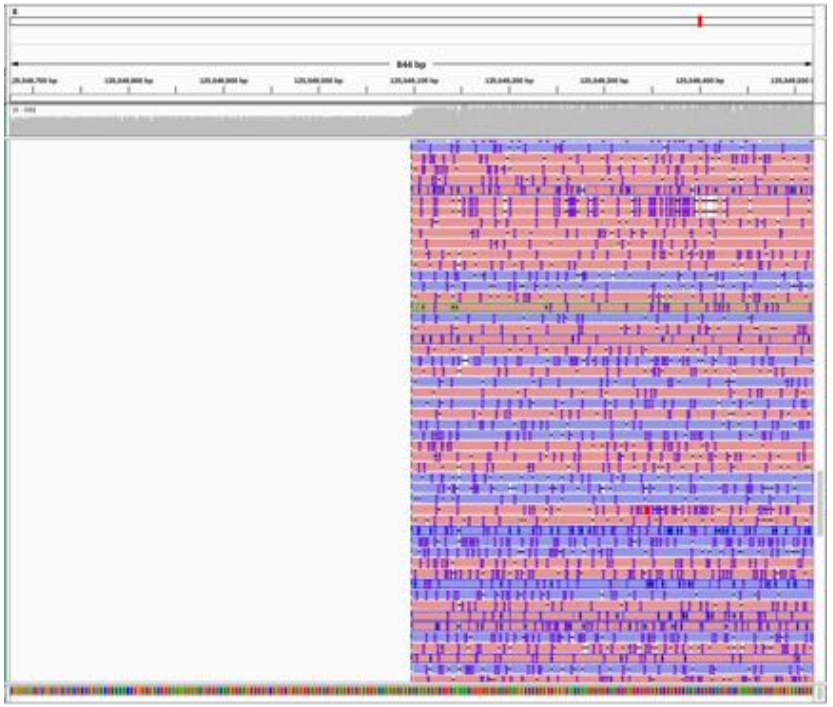
Without NGM-LR, alignments can be smudged over hundreds of base-pairs away from the breakpoint

BWA-MEM



translocation

NGM-LR



translocation

Inversion in BWA-MEM

Acknowledgments

Cold Spring Harbor Laboratory

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

Karen Ng
Timothy Beck
Yogi Sundaravadanam
John McPherson

DNAnexus

PACIFIC BIOSCIENCES®

Elizabeth Tseng
Jason Chin

Genentech FOUNDATION