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

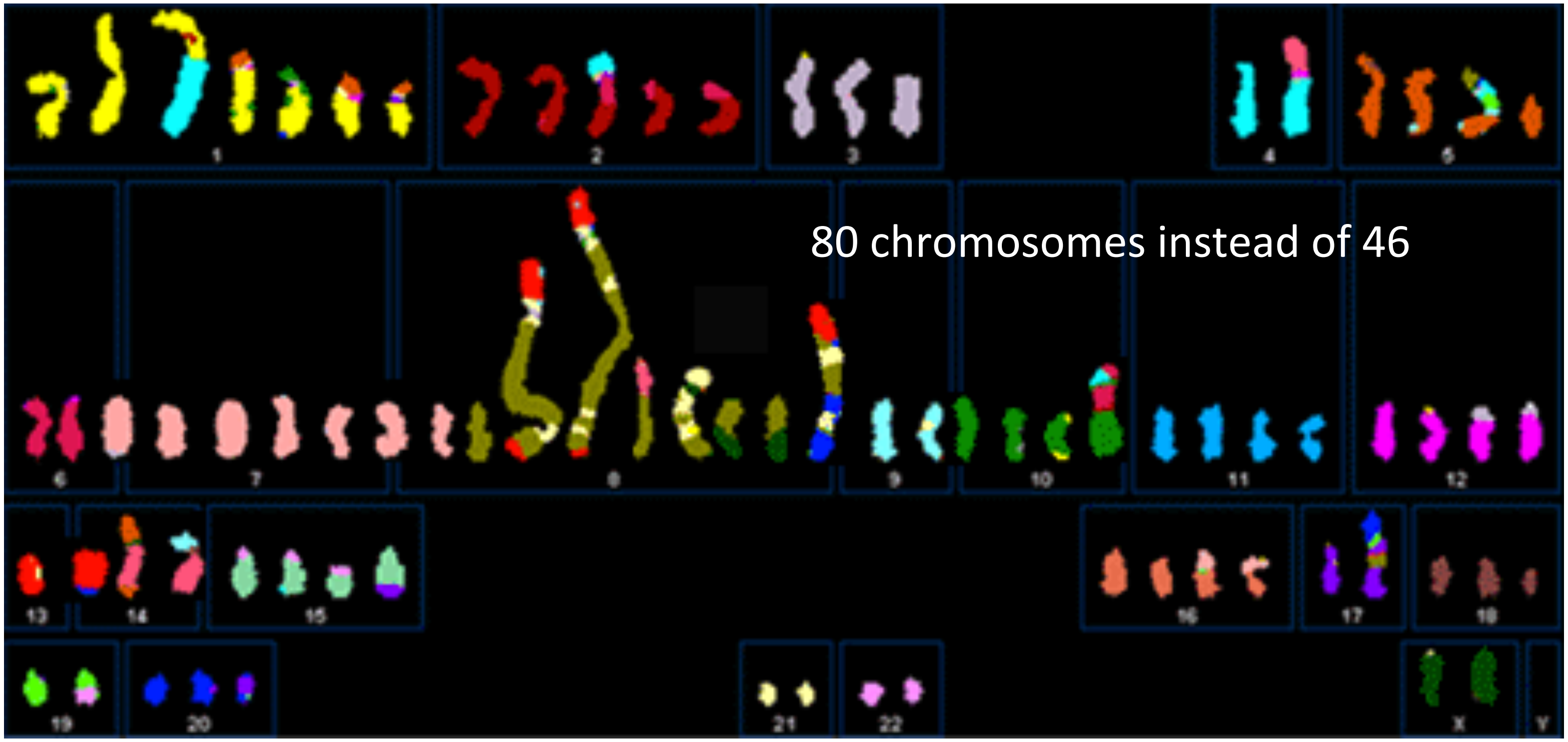
McPherson + Beck at the Ontario Institute for Cancer Research

Pacific Biosciences

DNAnexus

SK-BR-3

Most commonly used Her2-amplified breast cancer cell line

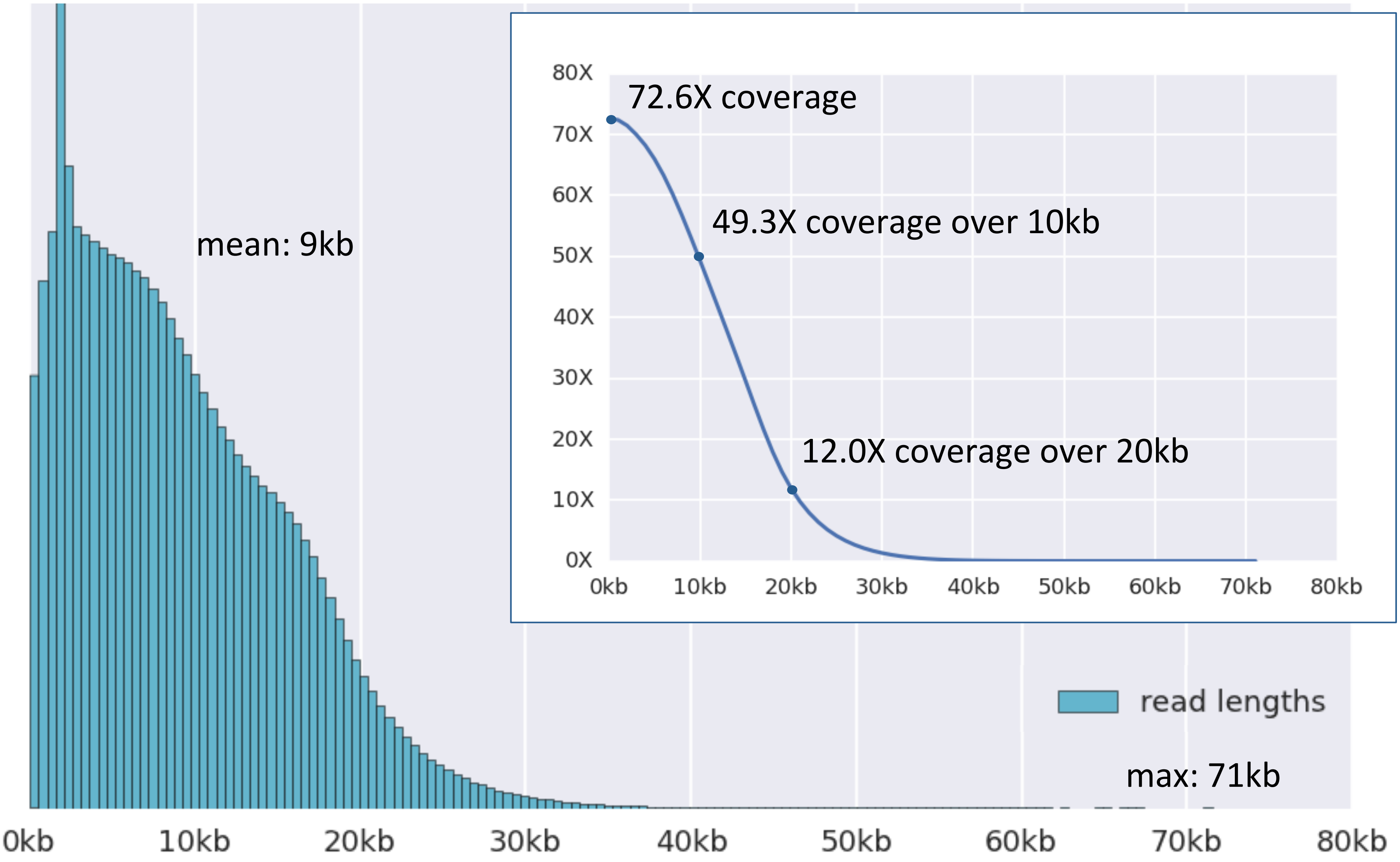


80 chromosomes instead of 46

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

(Davidson et al, 2000)

Sequencing SK-BR-3:
PacBio read length distribution



mean: 9kb

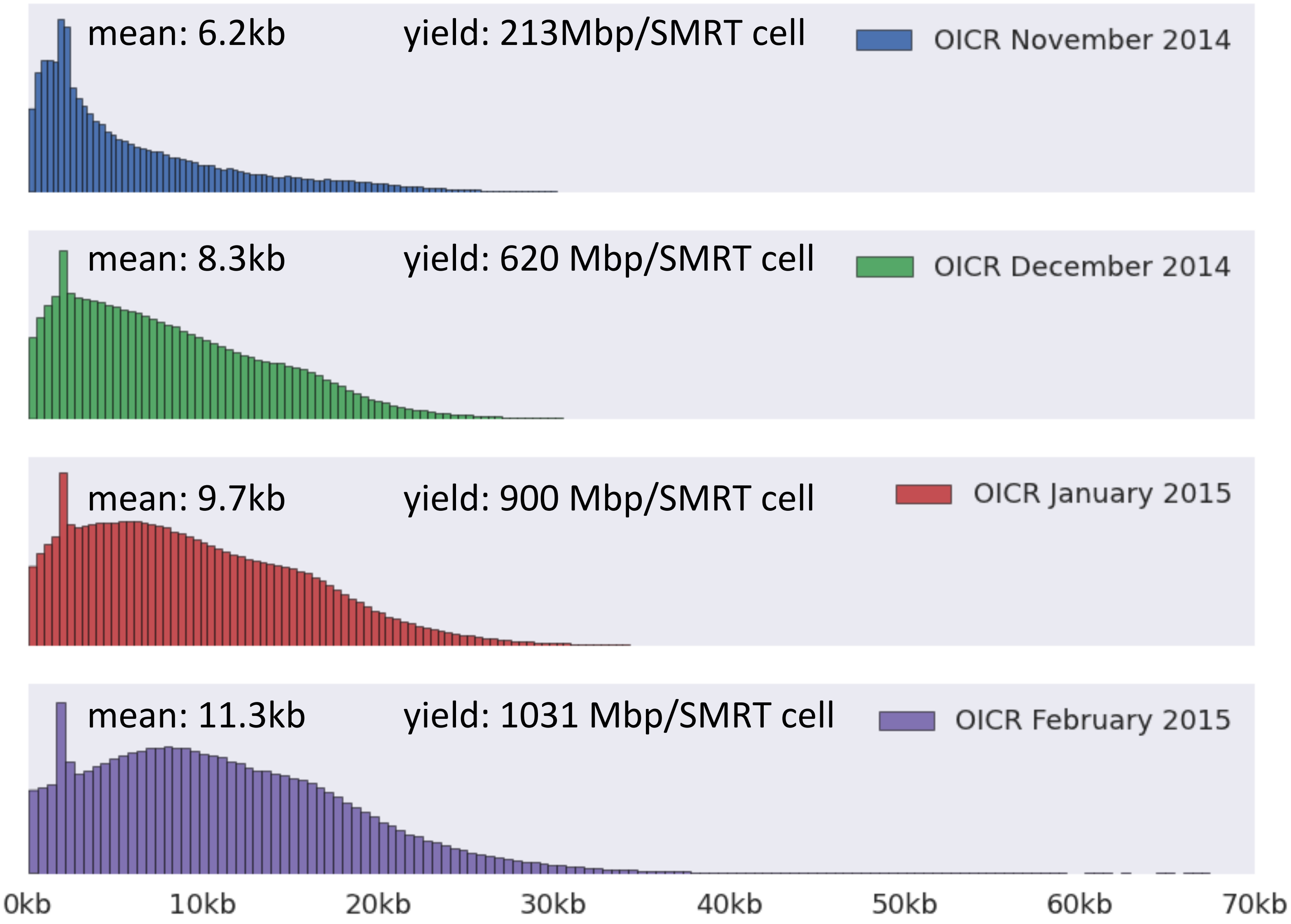
72.6X coverage

49.3X coverage over 10kb

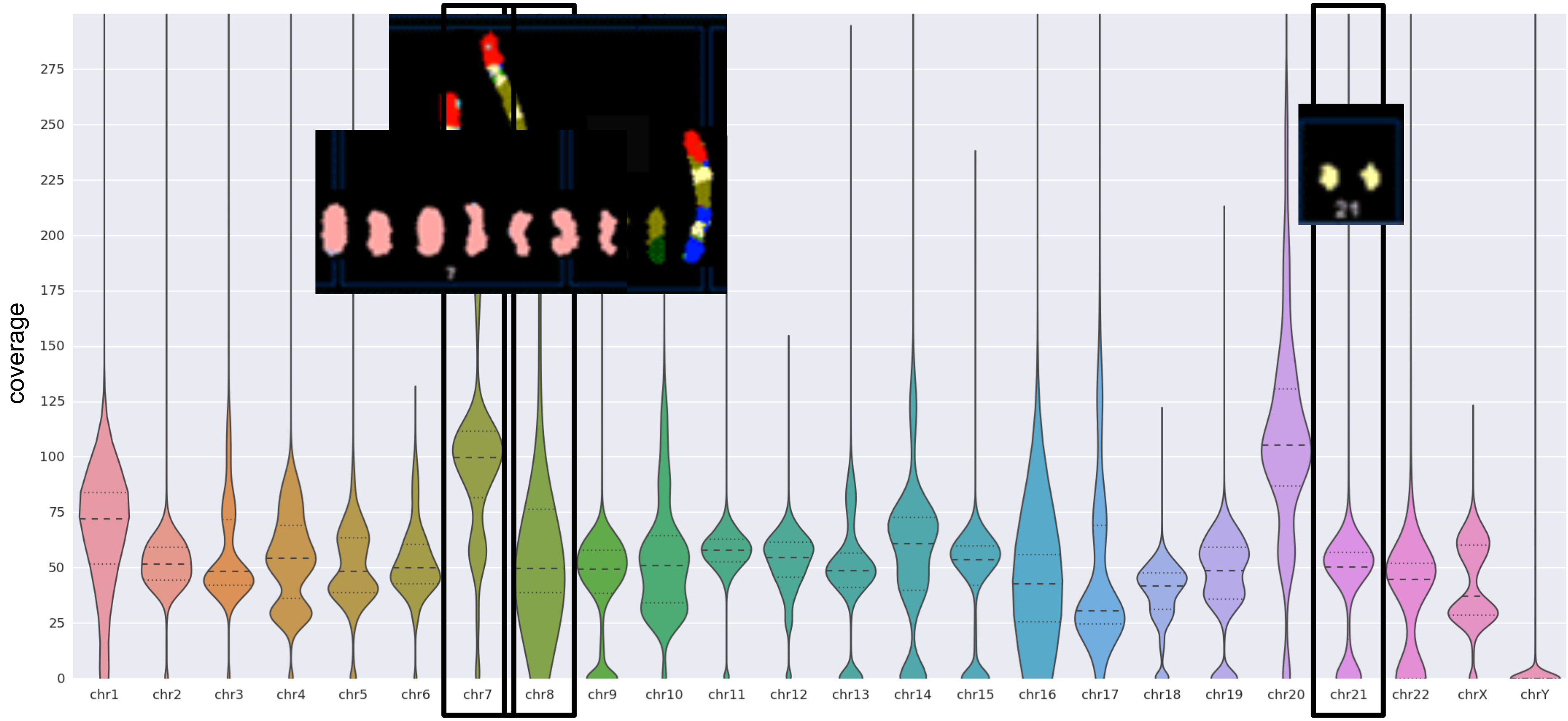
12.0X coverage over 20kb

max: 71kb

Dramatic changes just by experimenting with library preparation



Copy-number analysis is consistent with karyotype results

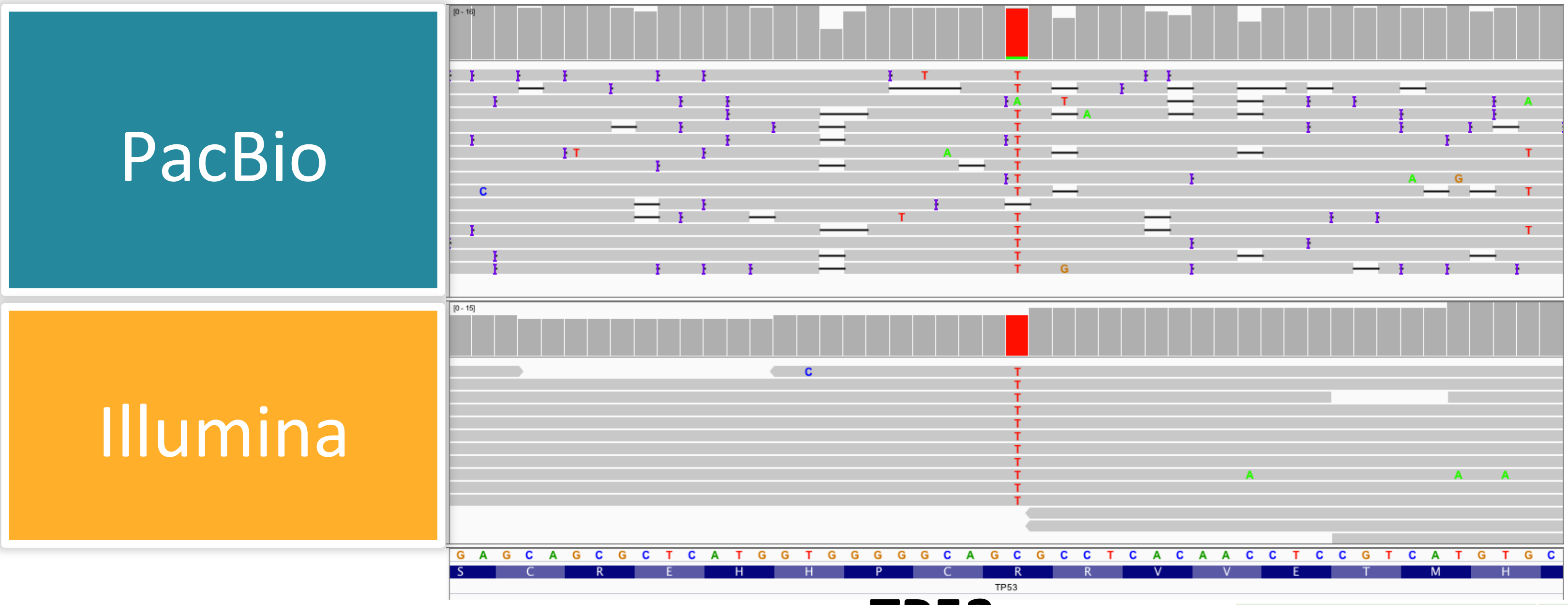


Genome-wide coverage averages around 54X
Coverage per chromosome varies greatly as expected from previous karyotyping results

We could call SNPs if we wanted to

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

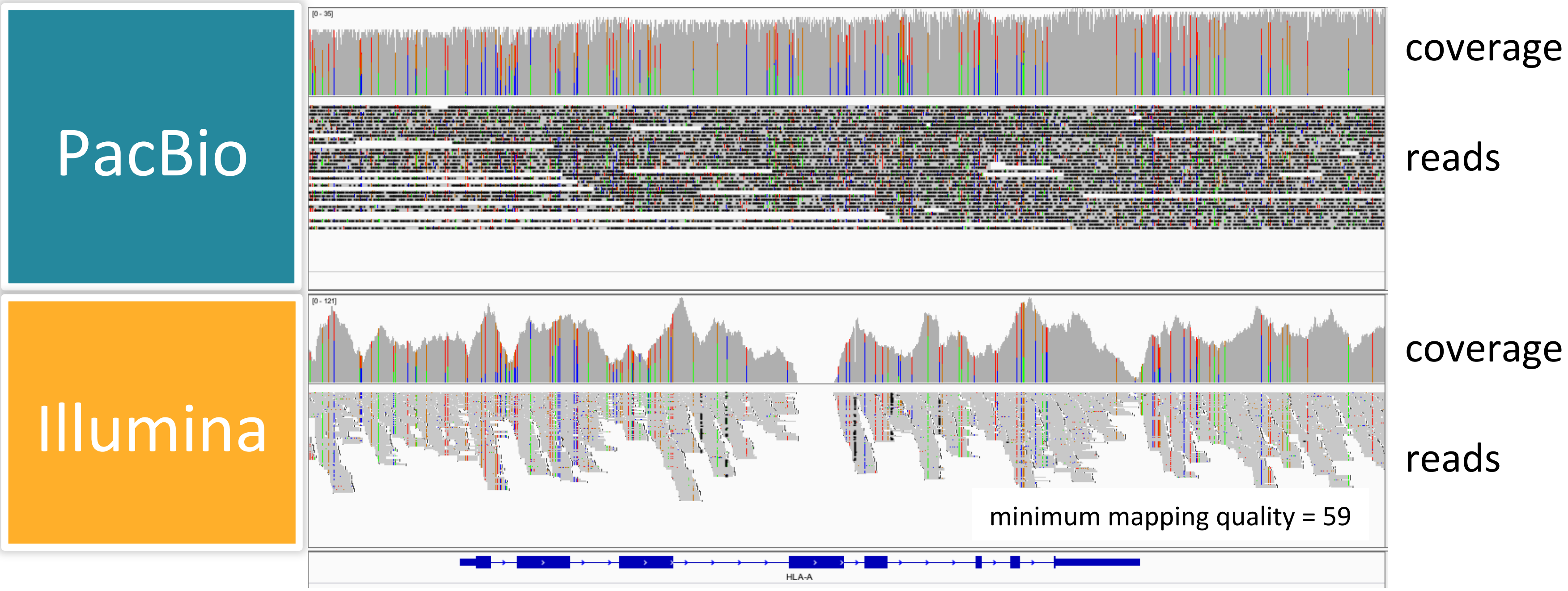
```
                                        Arg
Reference   ATCTGAGCAGCGCTCATGGTGGGGGCAGCGCCTCACAACCTCCGTCATGTGCTGTGACTGCTT
Illumina    ATCTGAGCAGCGCTCATGGTGGGGGCAGTGCCTCACAACCTCCGTCATGTGCTGTGACTGCTT
PacBio      ATCTGAGCAGCGCTCATGGTGGGGGCAGTGCCTCACAACCTCCGTCATGTGCTGTGACTGCTT
                        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


PacBio
Illumina
[0 - 16]
[0 - 15]
G A G C A G C G C T C A T G G T G G G G G C A G C G C C T C A C A A C C T C C G T C A T G T G C
S C R E H H P C R R V V E T M H
TP53

TP53 gene

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

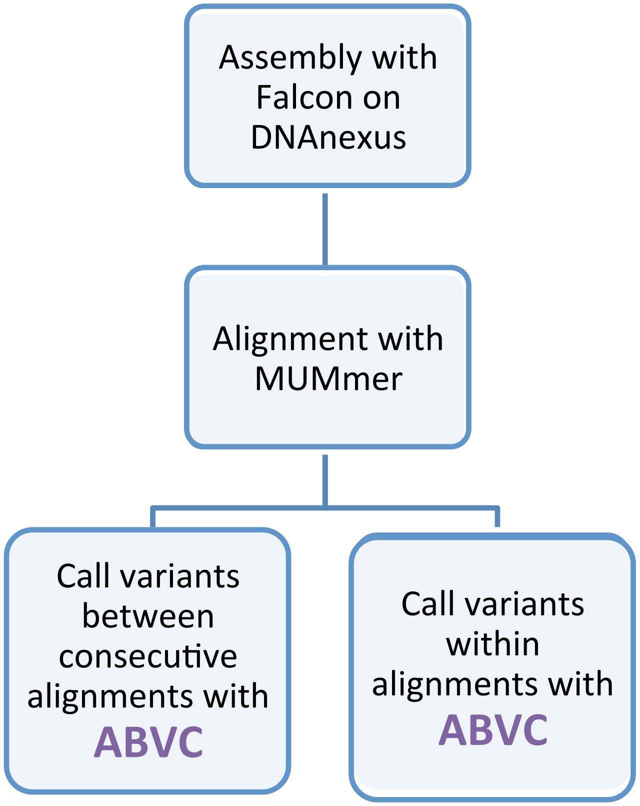
PacBio reads are longer and less susceptible to mapping issues



HLA-A gene

Genome structural analysis

Assembly-based



Assembly with Falcon on DNAnexus

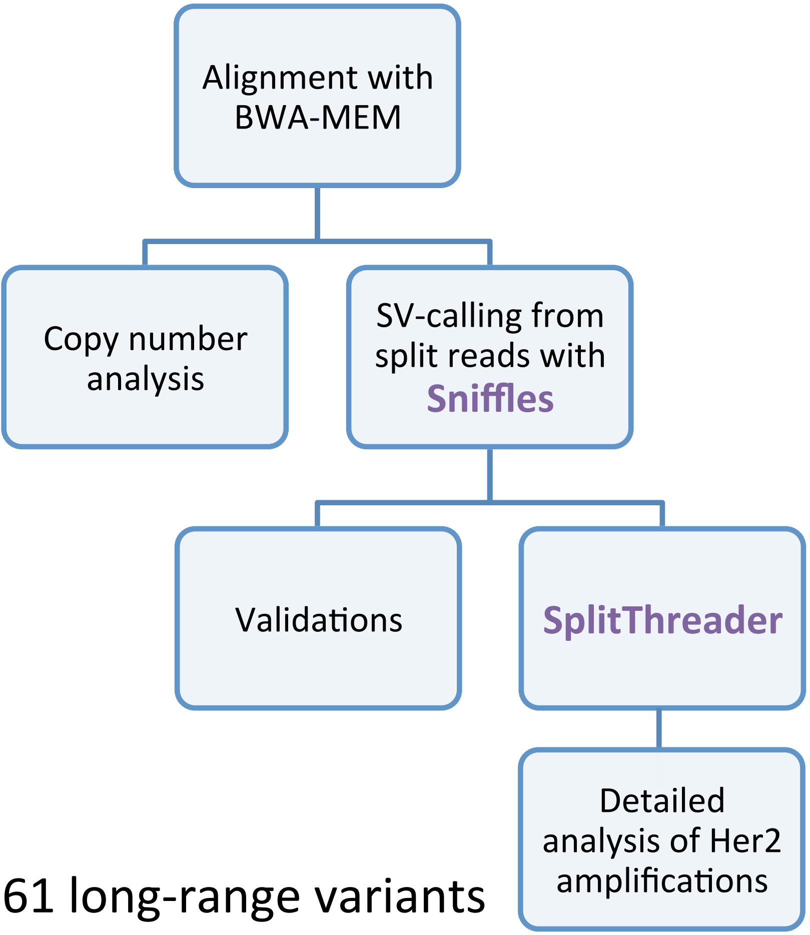
Alignment with MUMmer

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

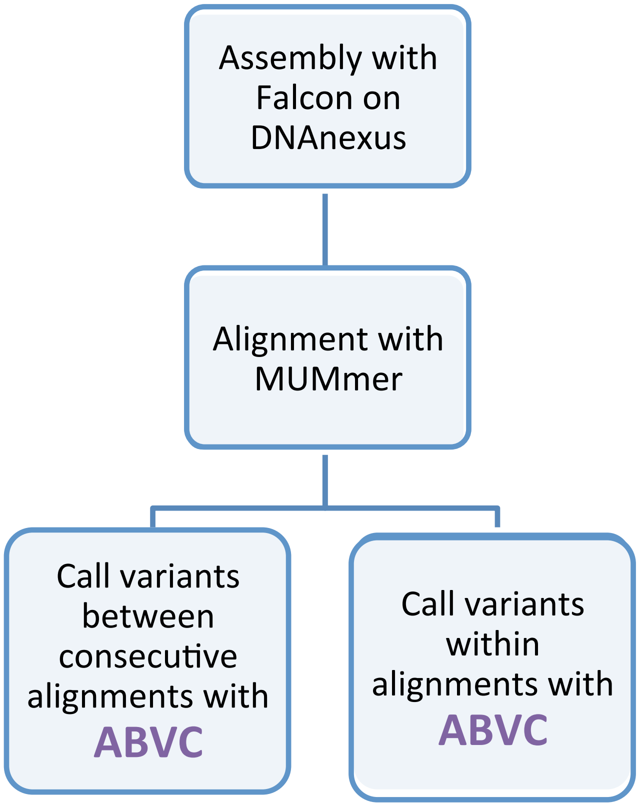
SplitThreader

Detailed analysis of Her2 amplifications

661 long-range variants
(>10kb distance)

Genome structural analysis

Assembly-based



Assembly with Falcon on DNAnexus

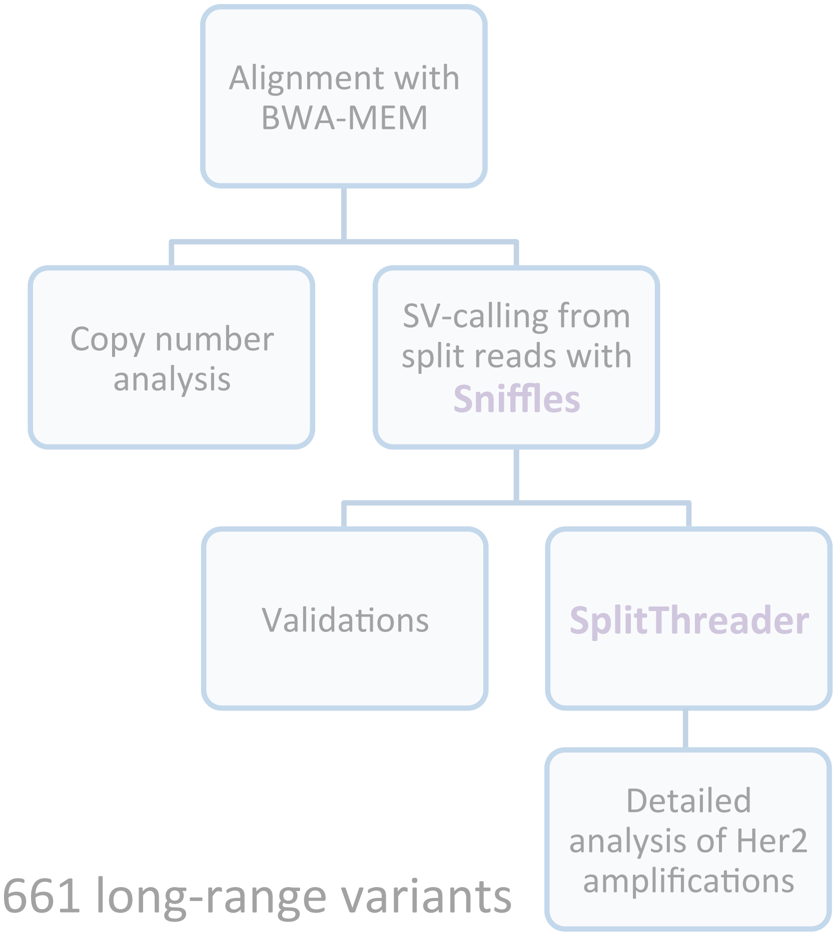
Alignment with MUMmer

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)

Genome assembly with FALCON

Find overlaps between PacBio reads

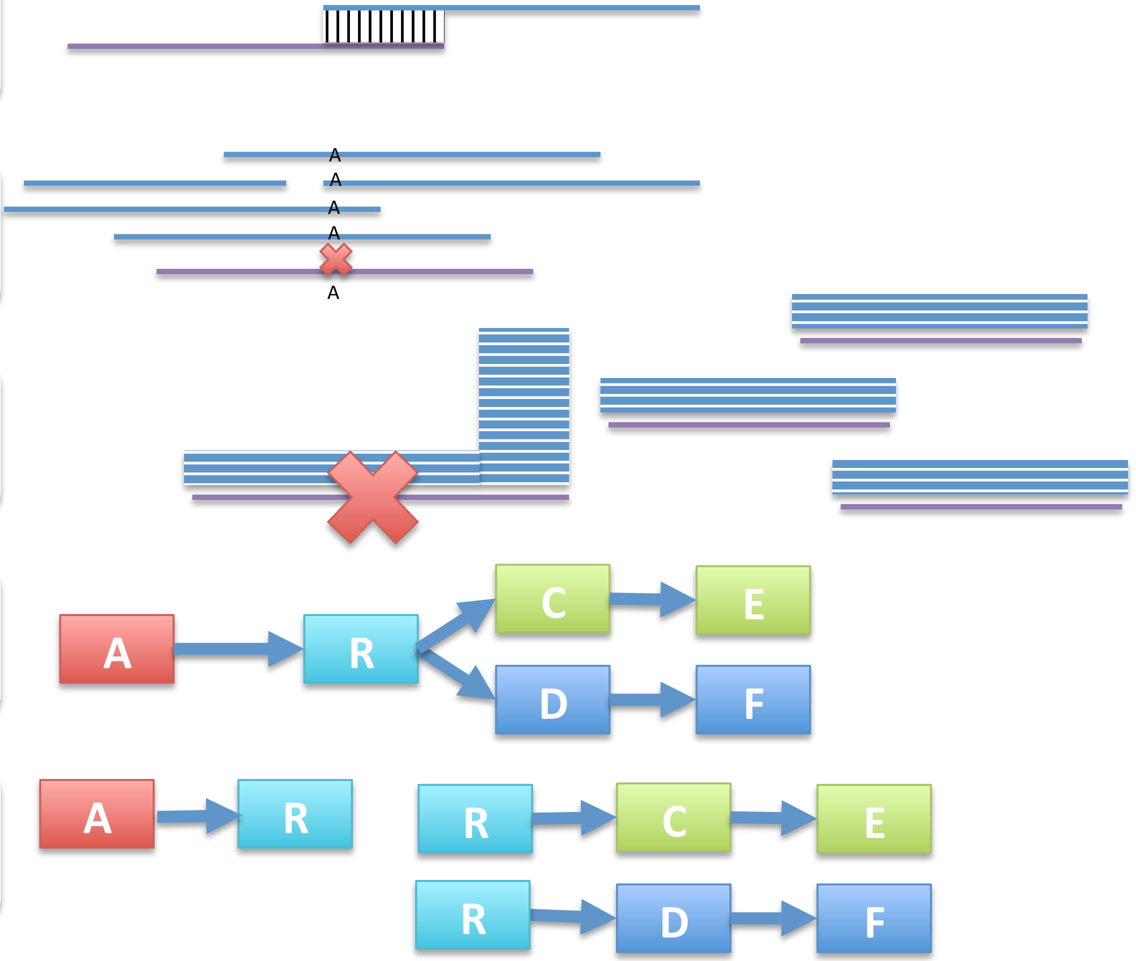
Error-correct reads against each other

- High enough coverage to self-correct so we don't need Illumina reads

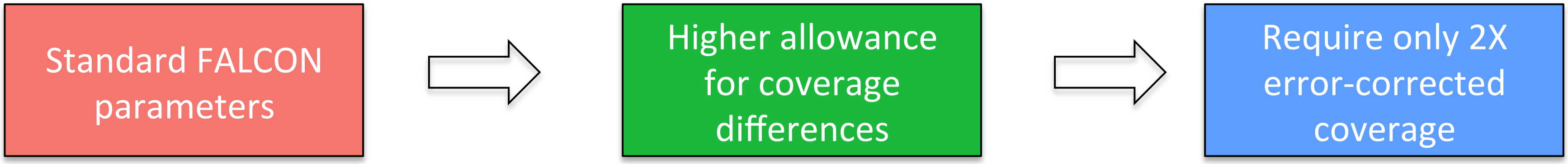
Filtering to avoid confounding by repeats

Constructing a graph from the overlaps

Constructing contigs from graph



Iterations of Falcon assembly on DNAnexus



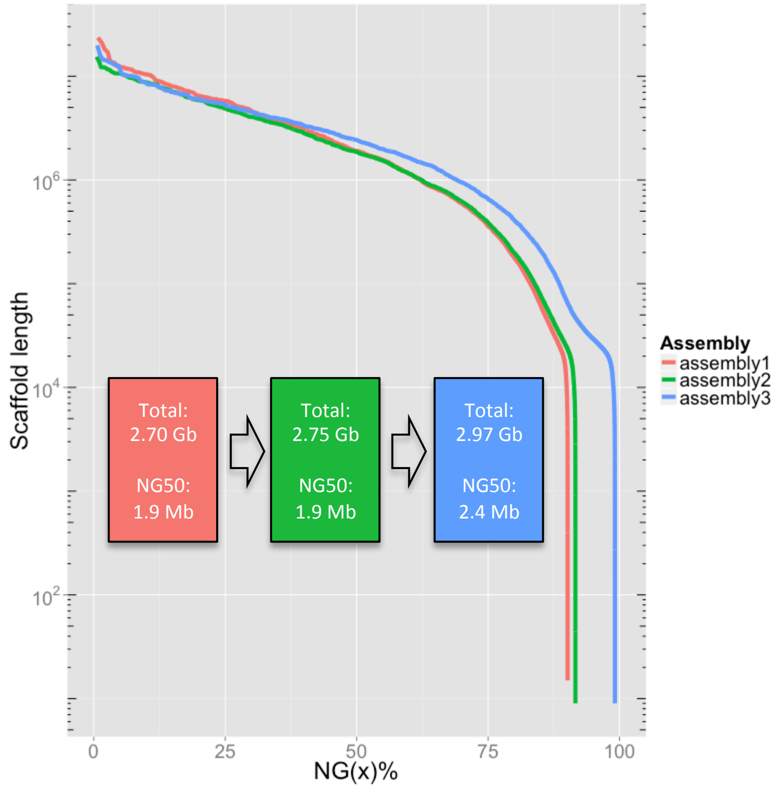
Standard FALCON parameters

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



Total: 2.70 Gb

NG50: 1.9 Mb

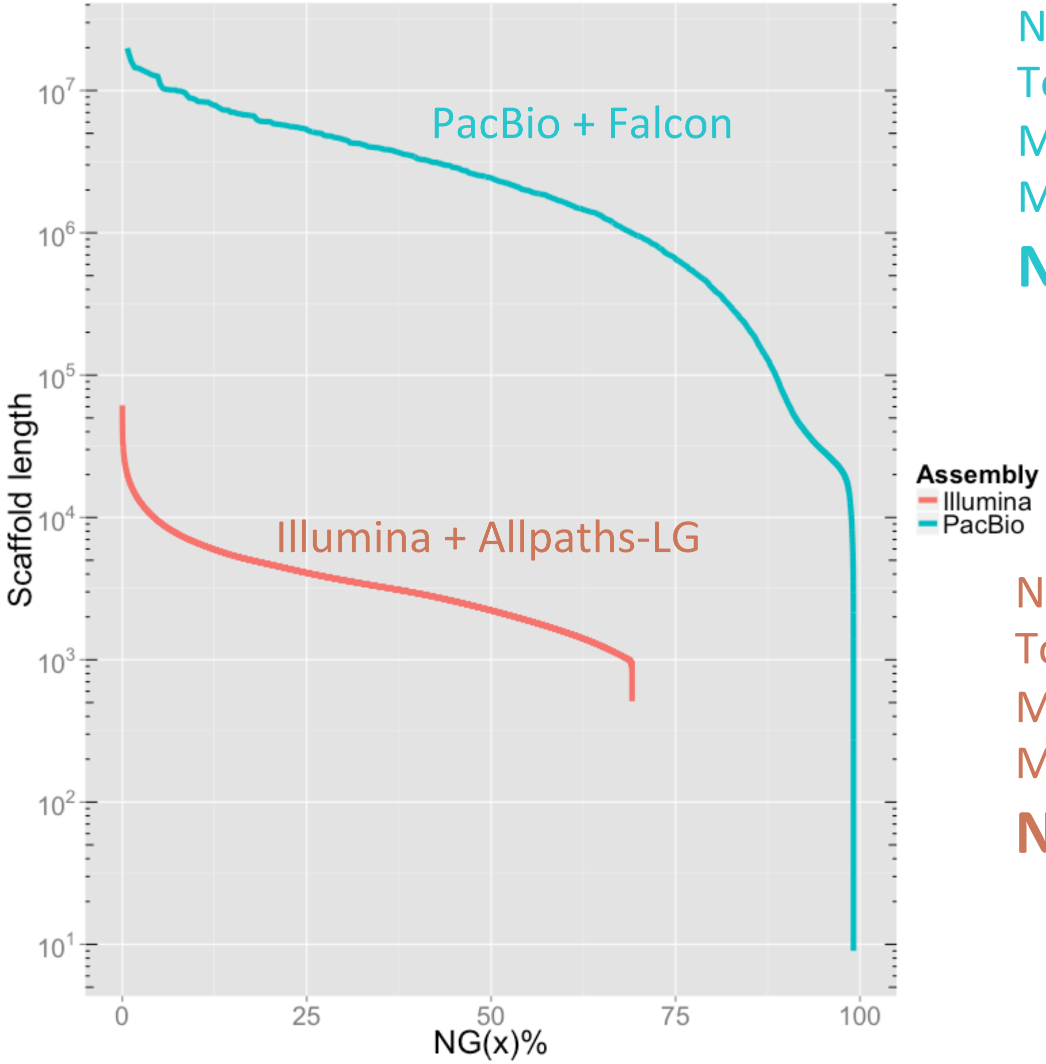
Total: 2.75 Gb

NG50: 1.9 Mb

Total: 2.97 Gb

NG50: 2.4 Mb

Assembly using PacBio yields far better contiguity



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

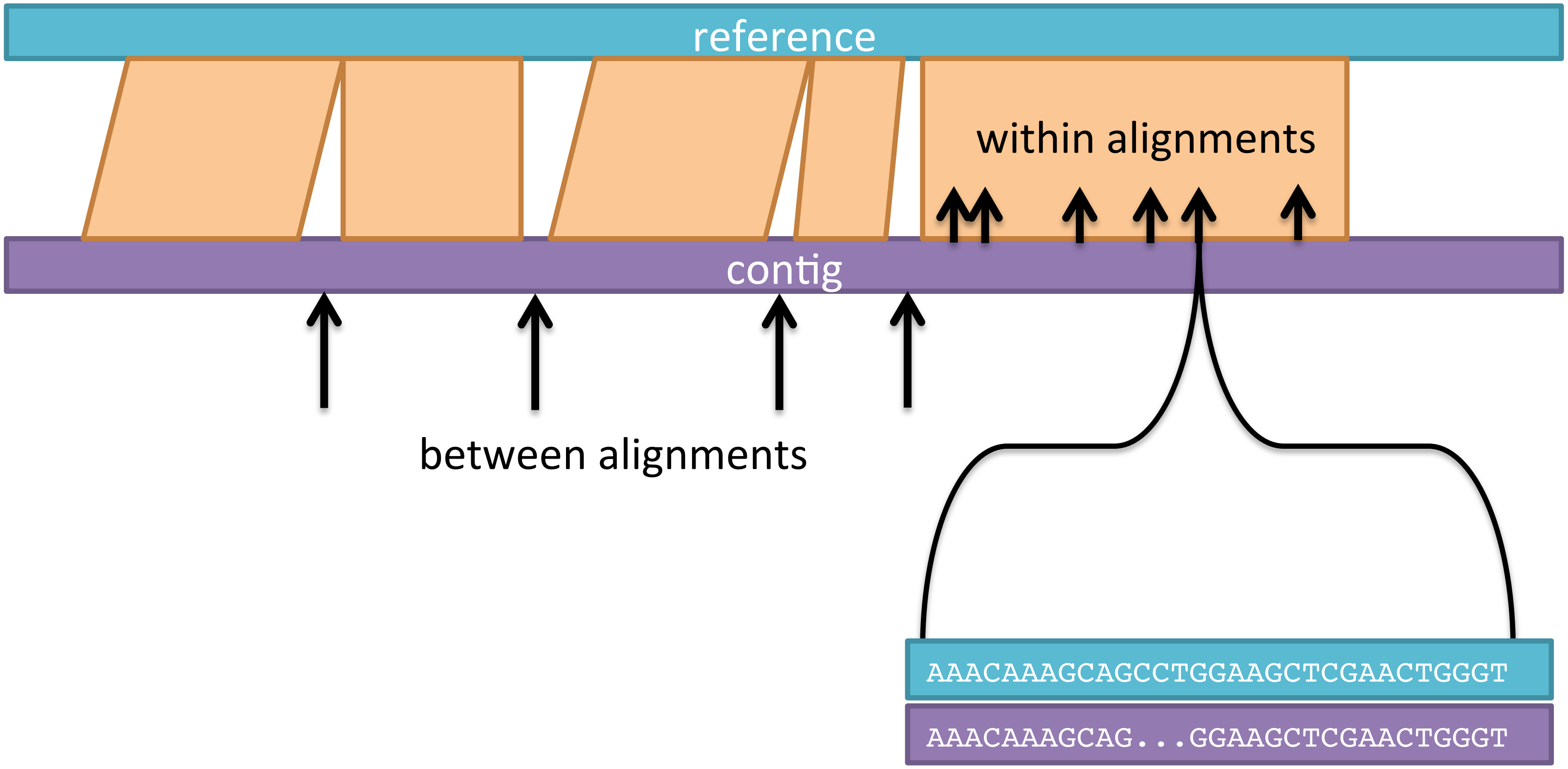
N50: 2.46 Mb

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



reference

within alignments

contig

between alignments

AAACAAAGCAGCCTGGAAGCTCGAACTGGGT

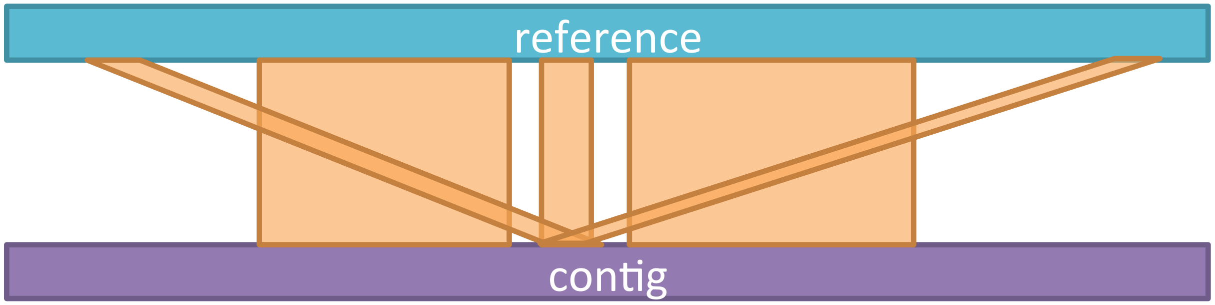
AAACAAAGCAG...GGAAGCTCGAACTGGGT

ABVC: Variants within alignments

Insertion	reference	AAACAAAGCAG...CCTGGGAAGCTCGAACTGGGT
	contig	AAACAAAGCAGTACCCTGGGAAGCTCGAACTGGGT
Deletion	reference	AAACAAAGCAGCCTGGAAGCTCGAACTGGGT
	contig	AAACAAAGCAG...GGAAGCTCGAACTGGGT

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

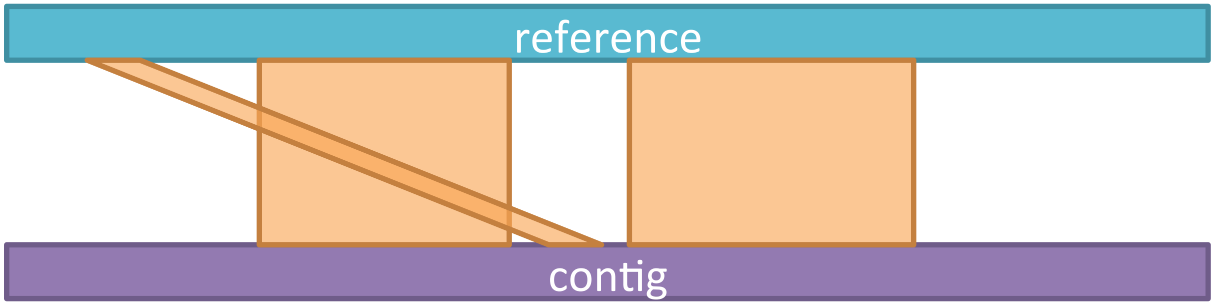
All alignments



Repeat maps in multiple locations

Traditional filtering by MUMmer

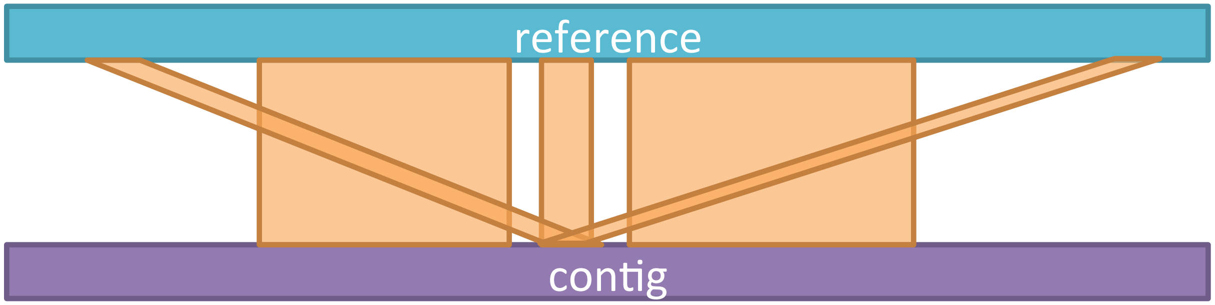
- choose the best alignment for each query
- random choice if multiple alignments with the same score



May be falsely called as a translocation

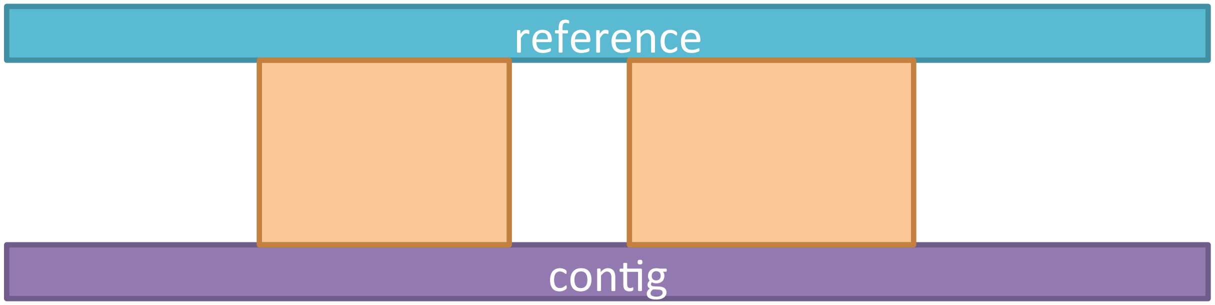
ABVC: Unique length filtering is needed to prevent false positives due to repetitive elements

All alignments



Repeat maps in multiple locations

Filtering based on requiring a certain length of unique alignment

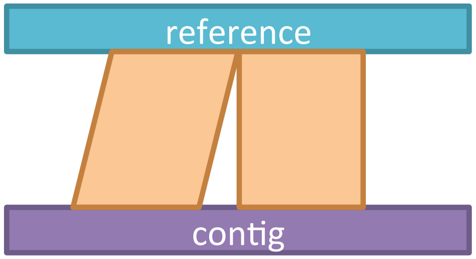


All repeats with less than 10kb of unique sequence are filtered out

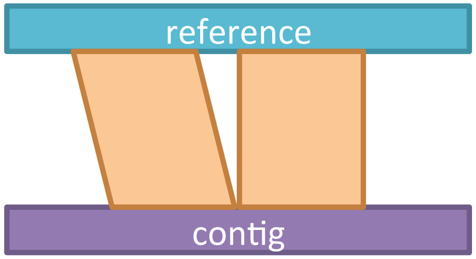
Types of variants detected by ABVC

Defined point

Insertion

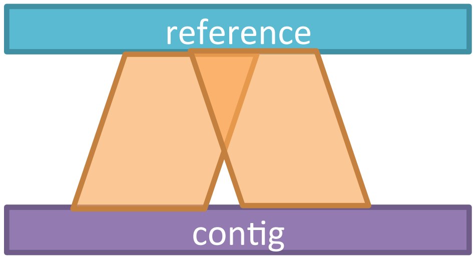


Deletion

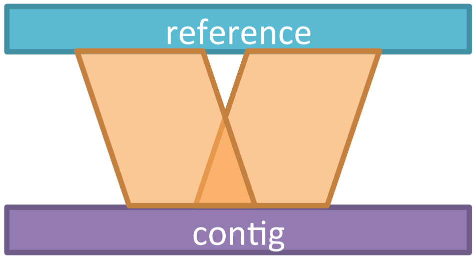


Overlapping alignments suggest tandem repeat

Tandem Expansions

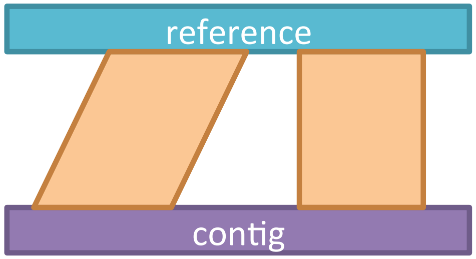


Tandem Contractions

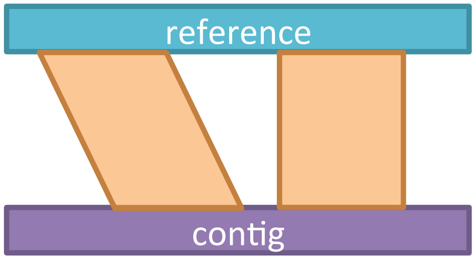


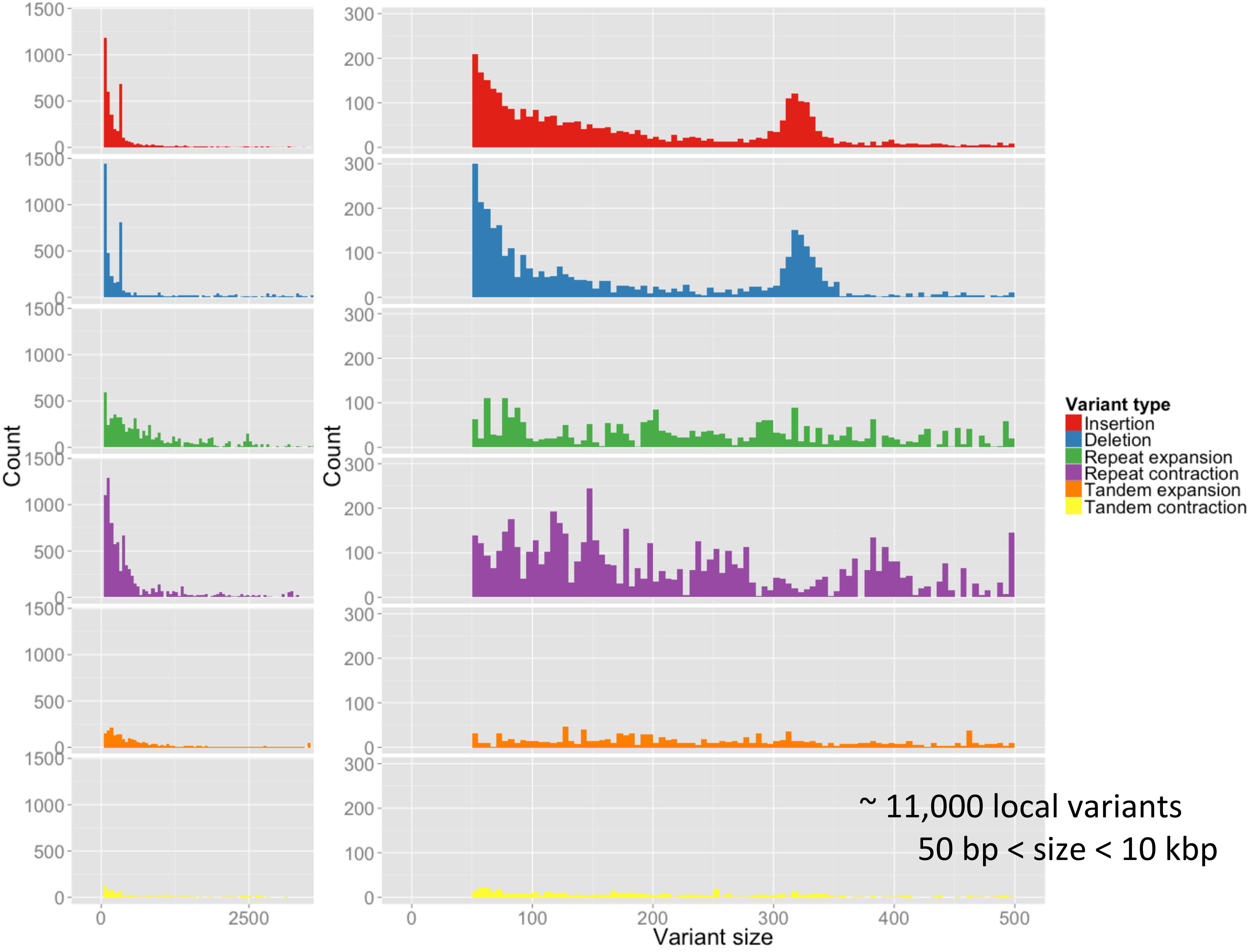
Gap where sequences do not align uniquely suggests a repeat

Repeat Expansions



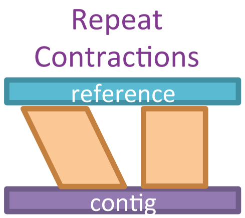
Repeat Contractions

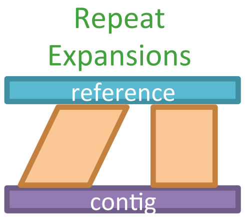


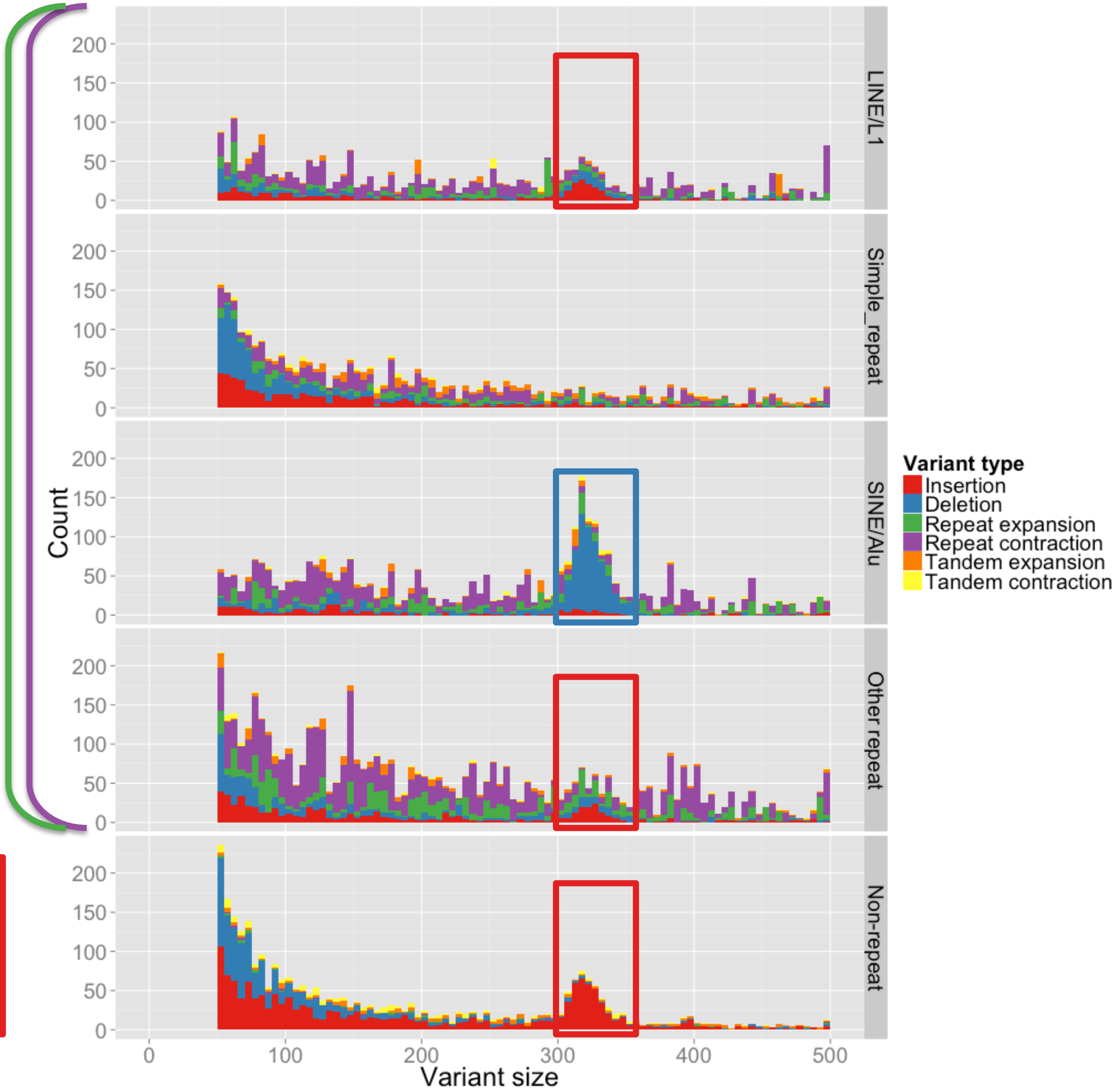


~ 11,000 local variants

50 bp < size < 10 kbp

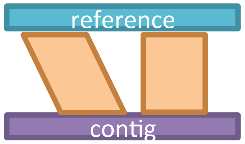




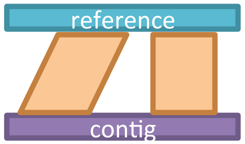


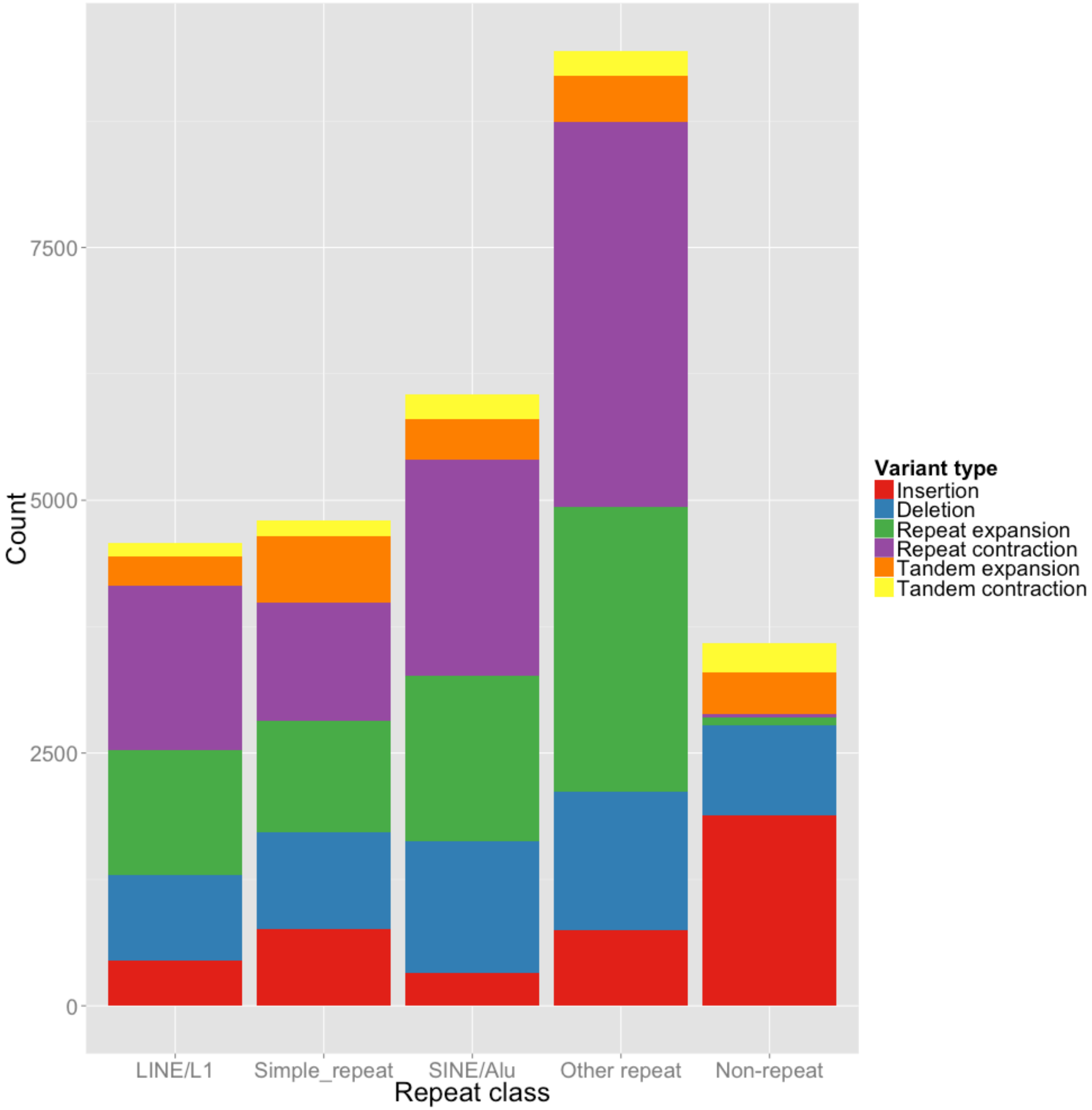
BLASTed 515 insertions:
427 (83%) of them
matched Alu elements

Repeat Contractions

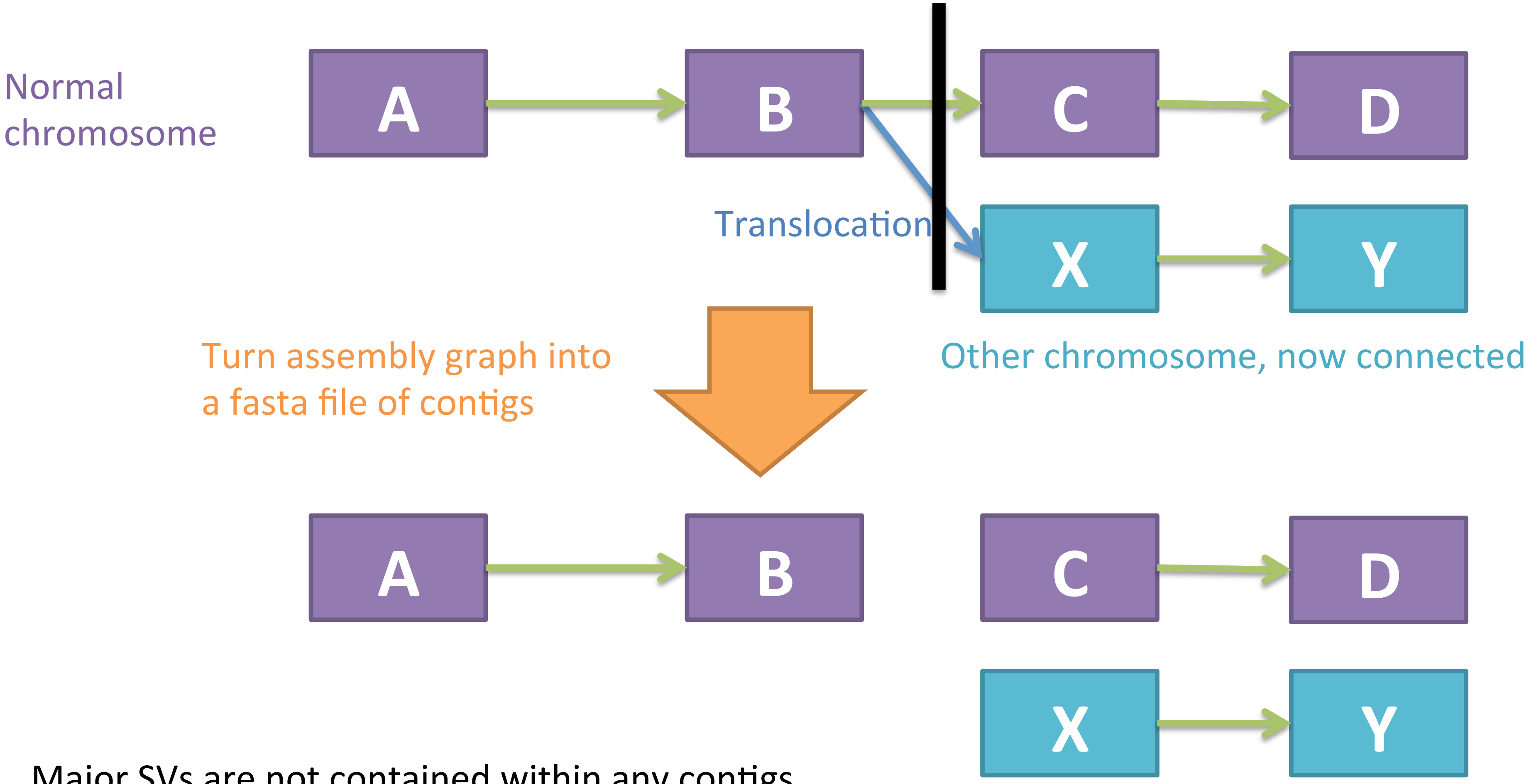


Repeat Expansions





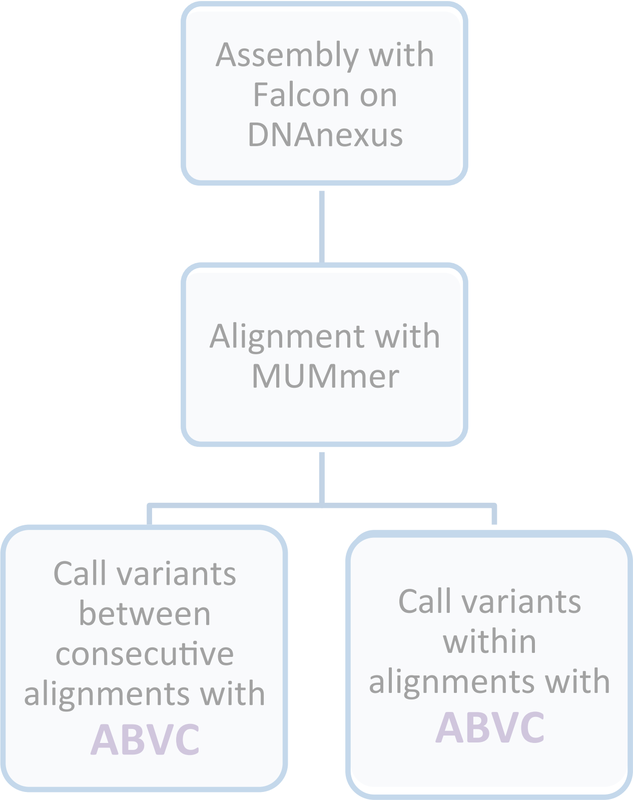
Why assembly doesn't capture long-range variants


Normal chromosome
A
B
C
D
Translocation
X
Y
Other chromosome, now connected
Turn assembly graph into a fasta file of contigs
A
B
C
D
X
Y

Major SVs are not contained within any contigs

Genome structural analysis

Assembly-based



Assembly with Falcon on DNAnexus

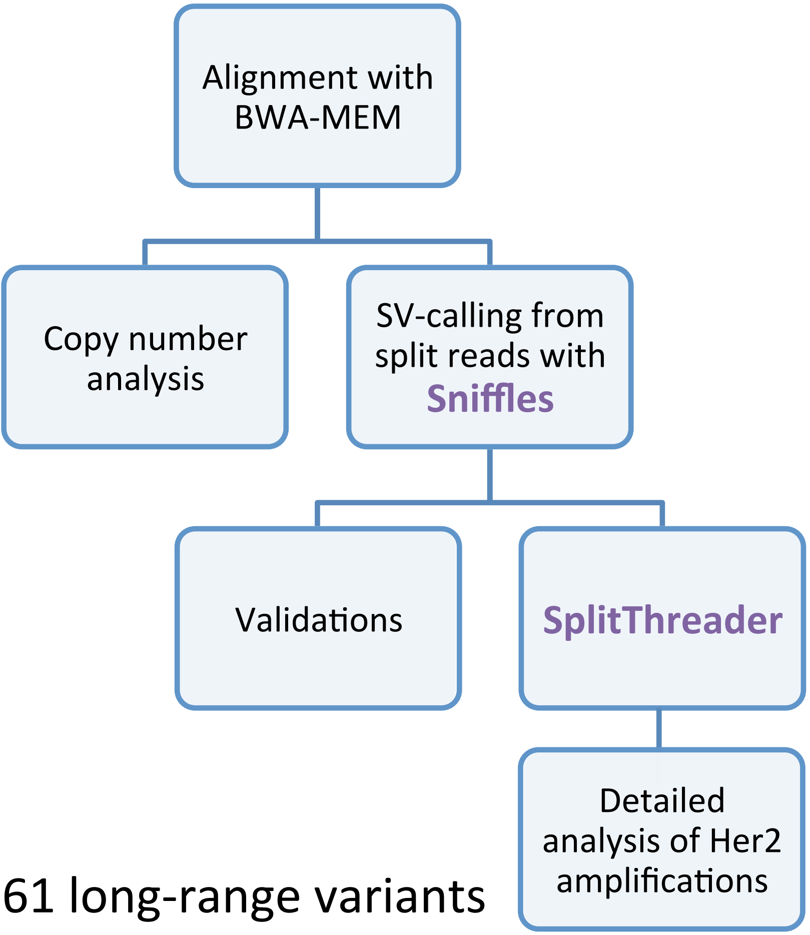
Alignment with MUMmer

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

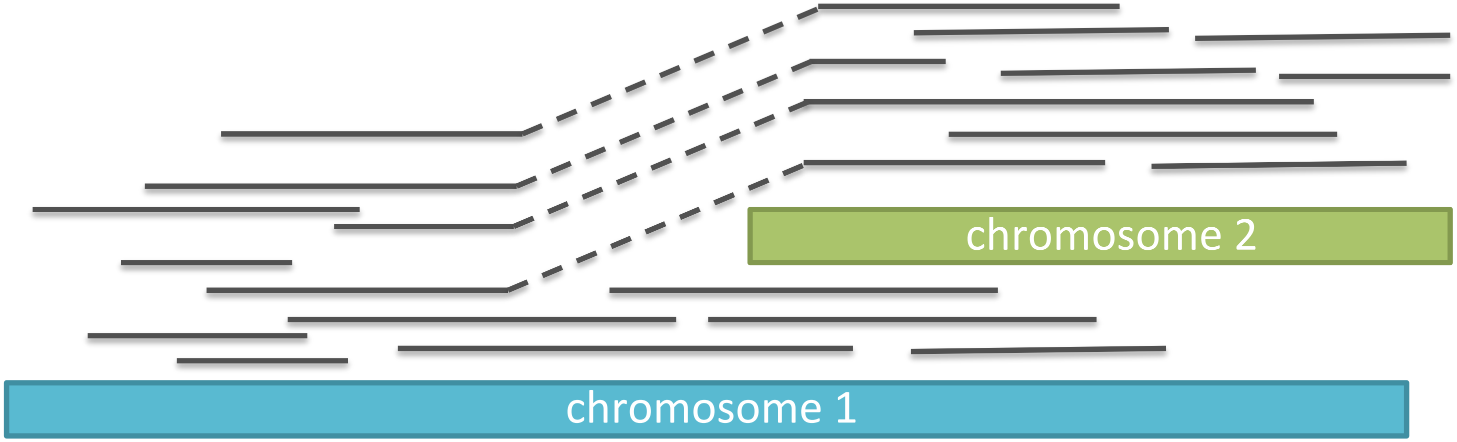
Validations

SplitThreader

Detailed analysis of Her2 amplifications

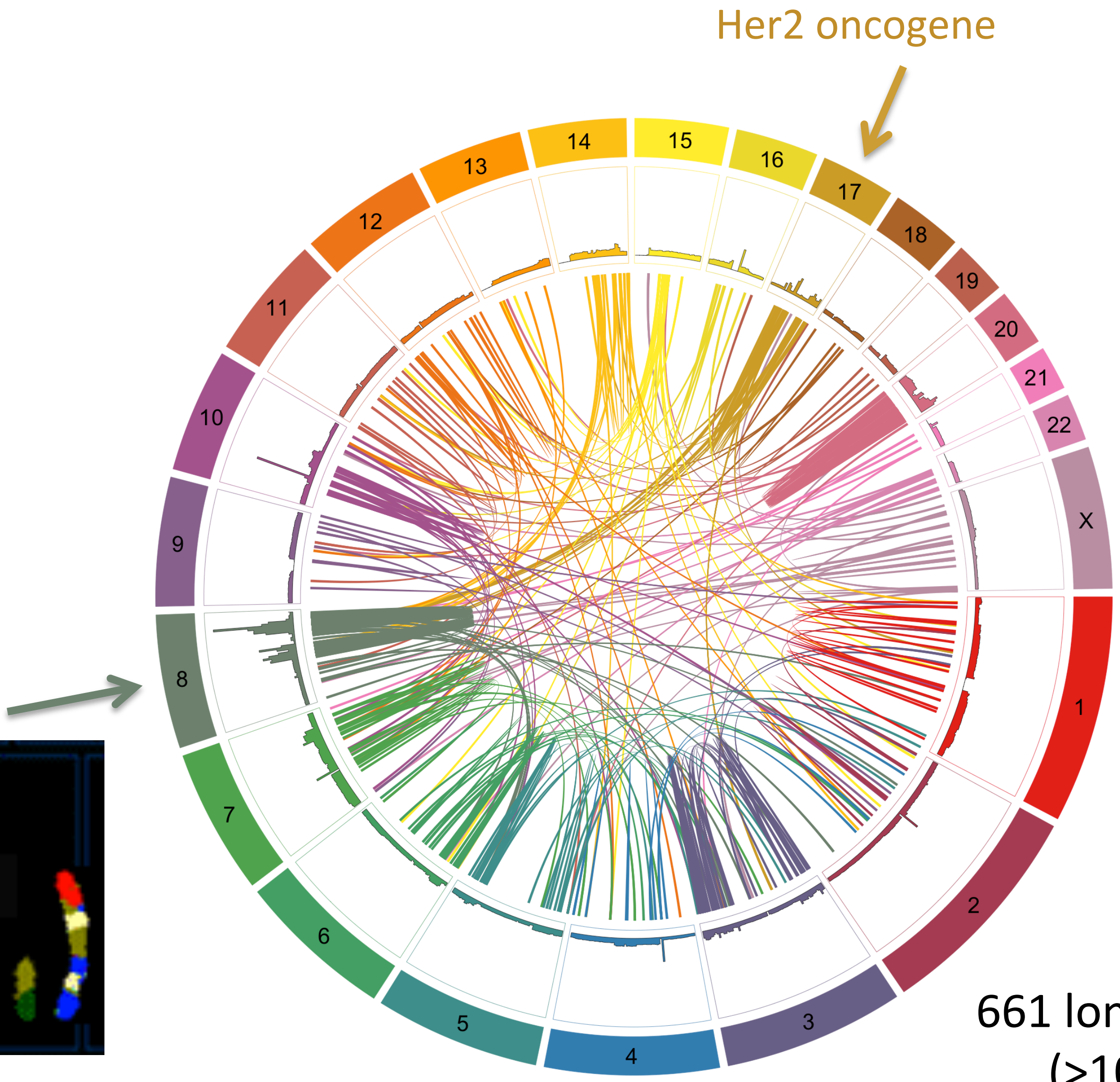
661 long-range variants
(>10kb distance)

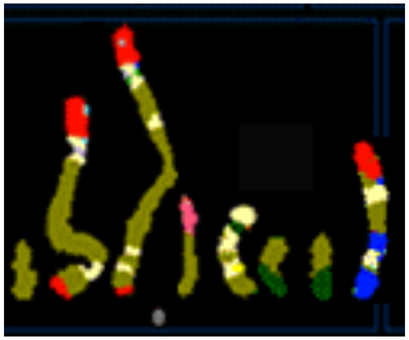
Variant-calling from split-read alignment



Software: Sniffles by Fritz Sedlazeck

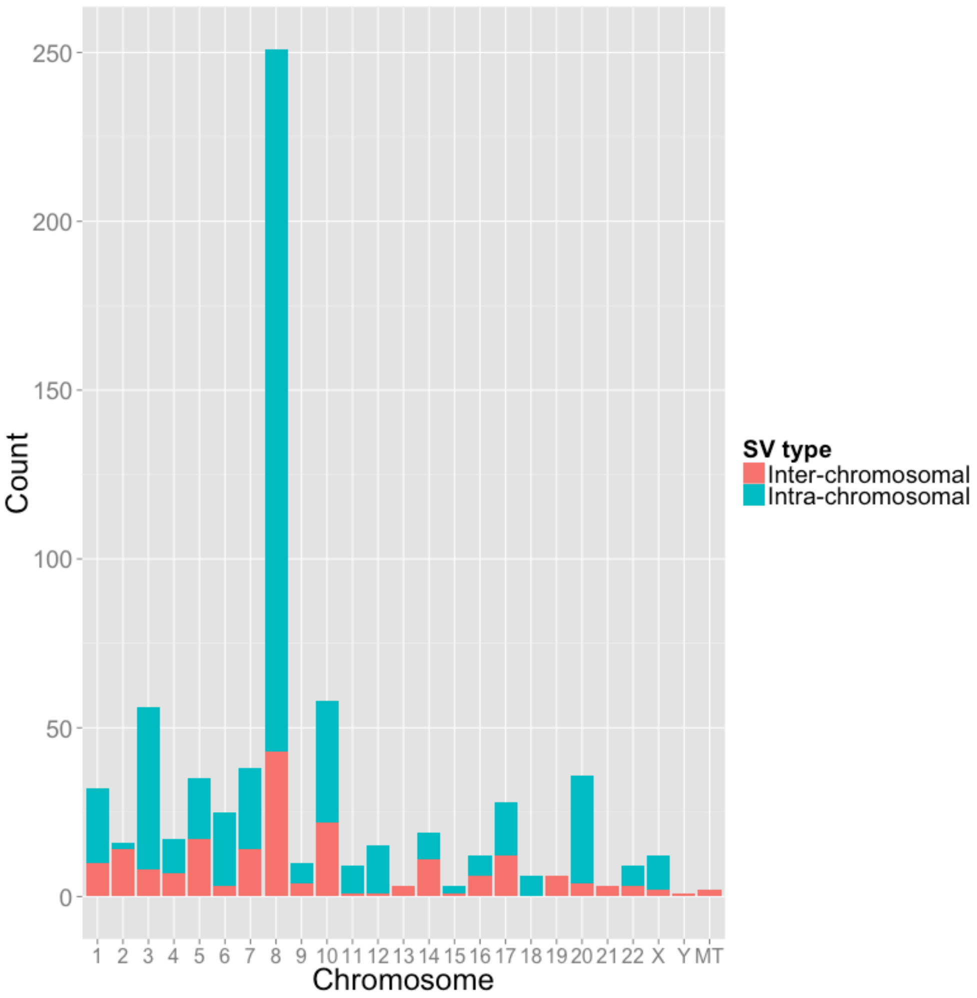
Long-range structural variants found by Sniffles



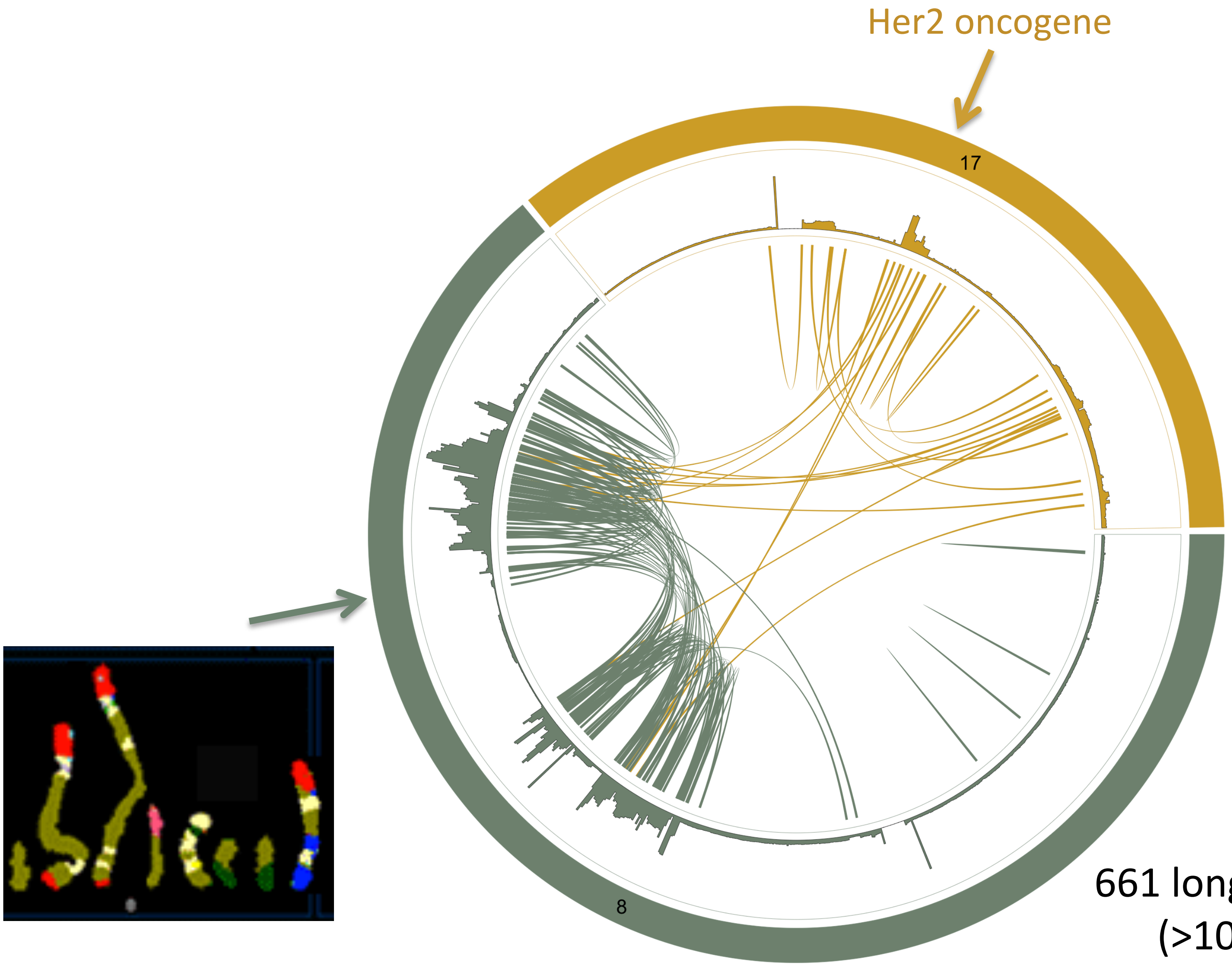


661 long-range variants
(>10kb distance)

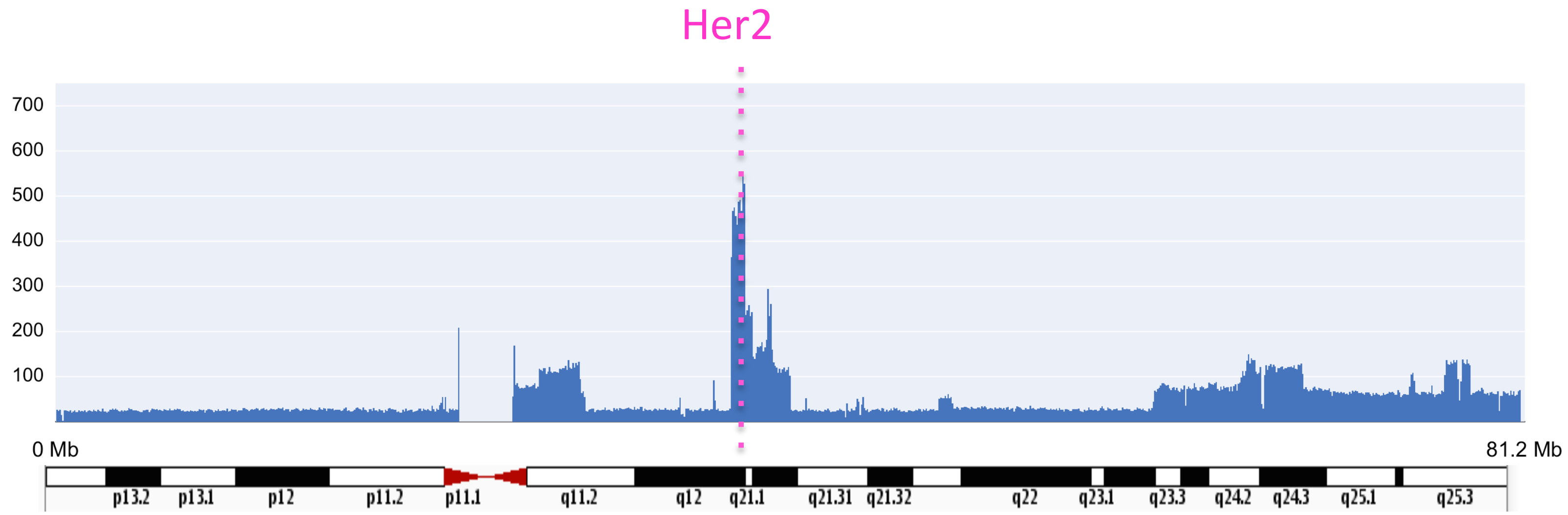
Chromosome 8 has the most intra- and inter-chromosomal long-range variants



Long-range structural variants found by Sniffles

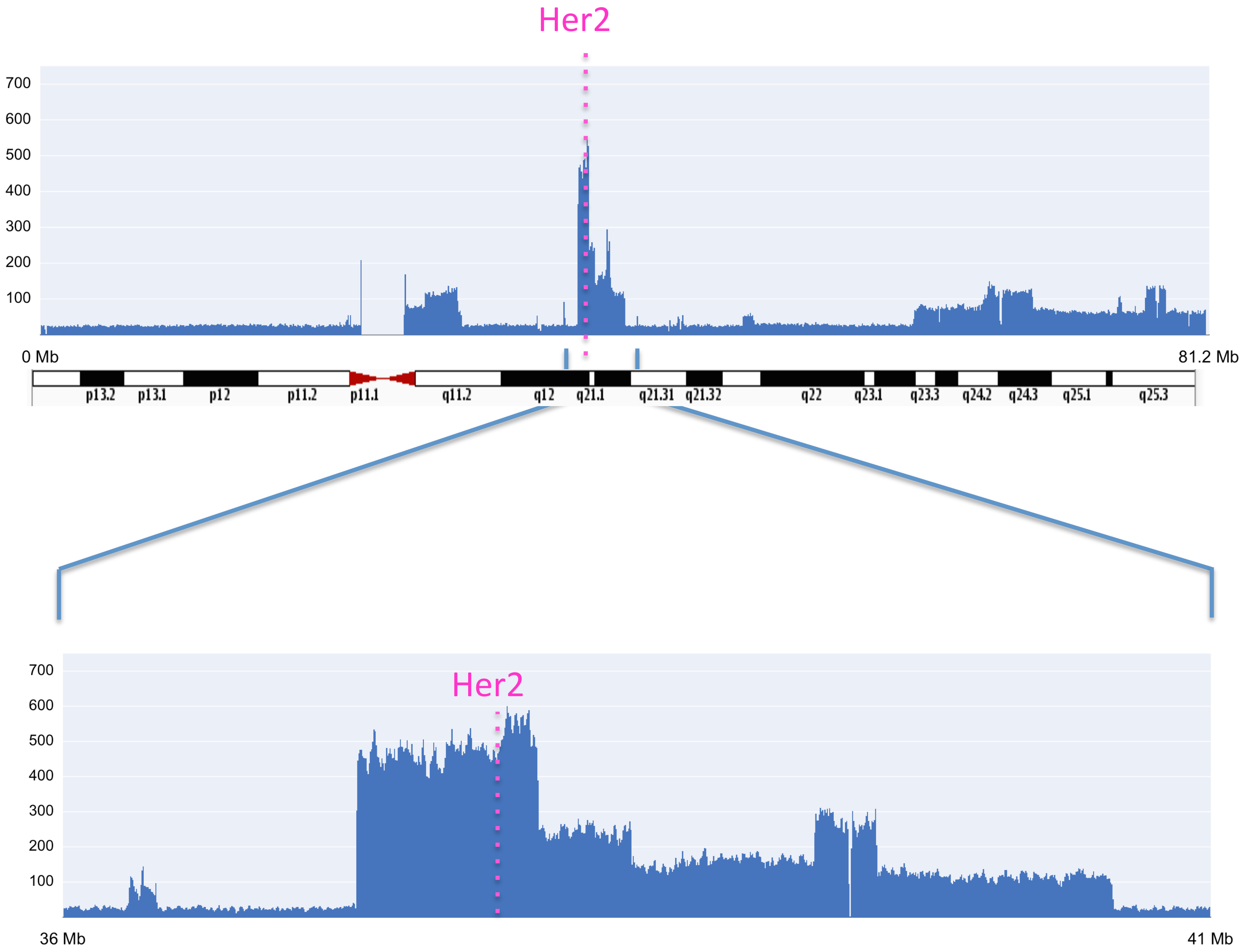


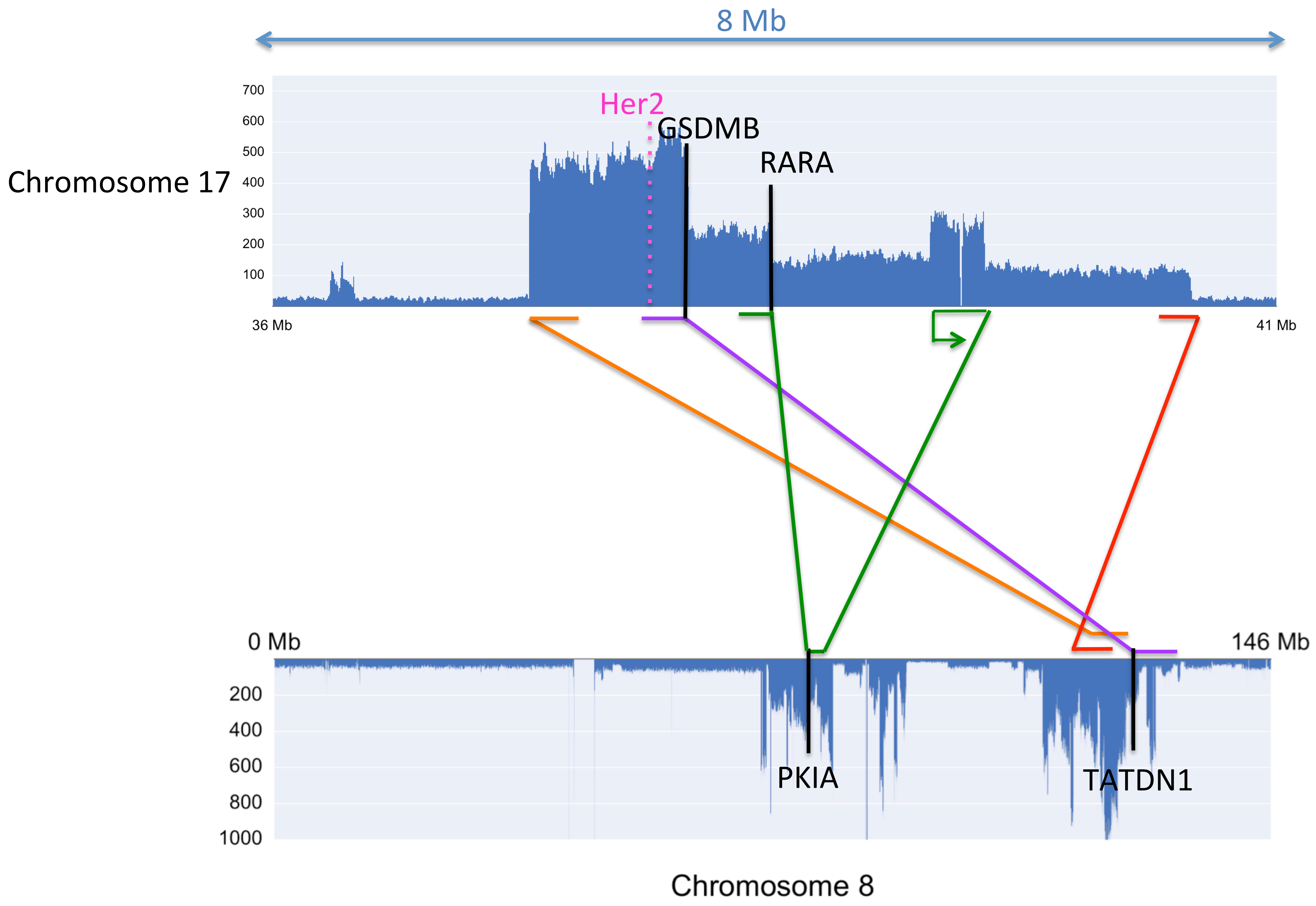
661 long-range variants
(>10kb distance)



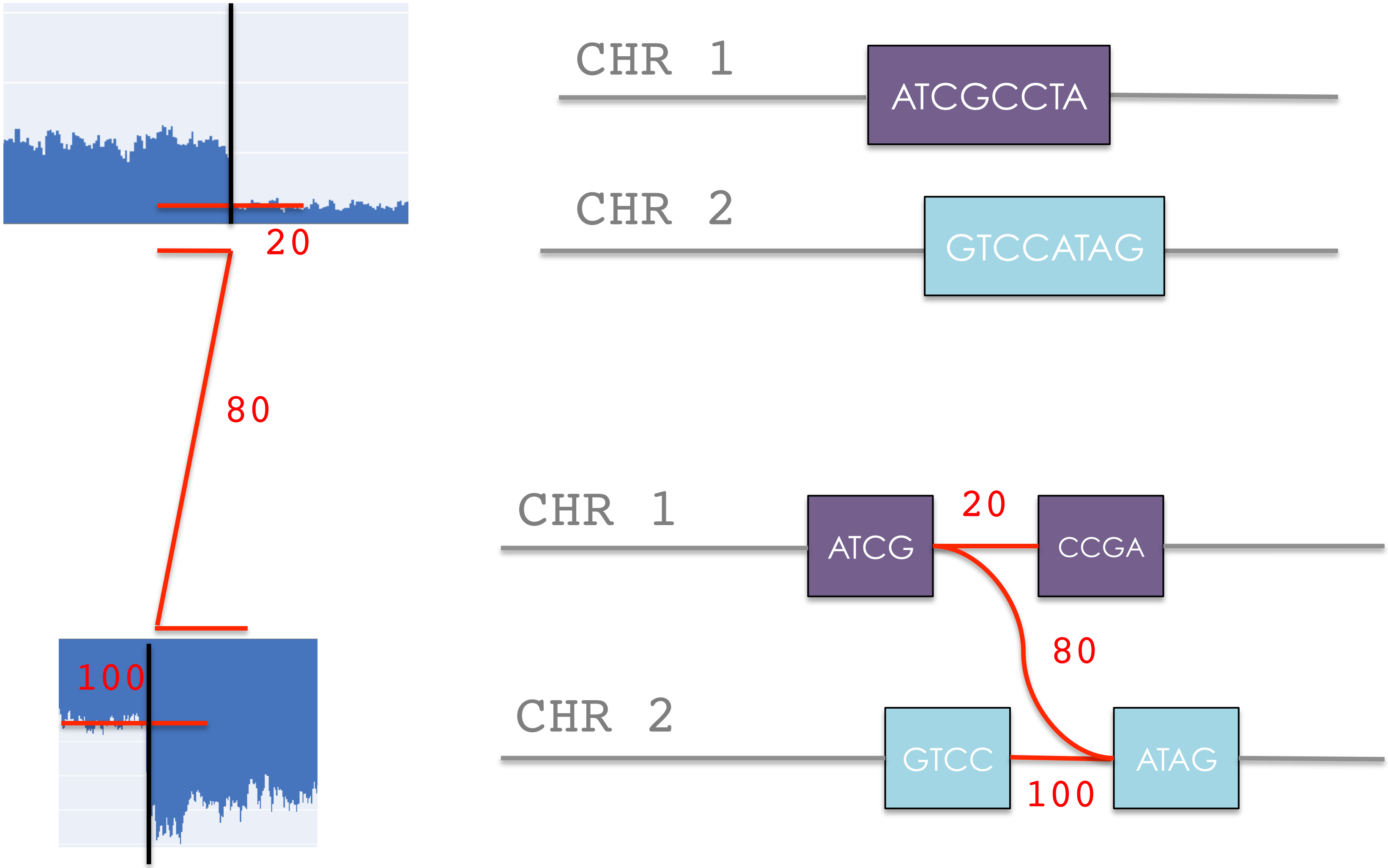
Chr 17: 83 Mb

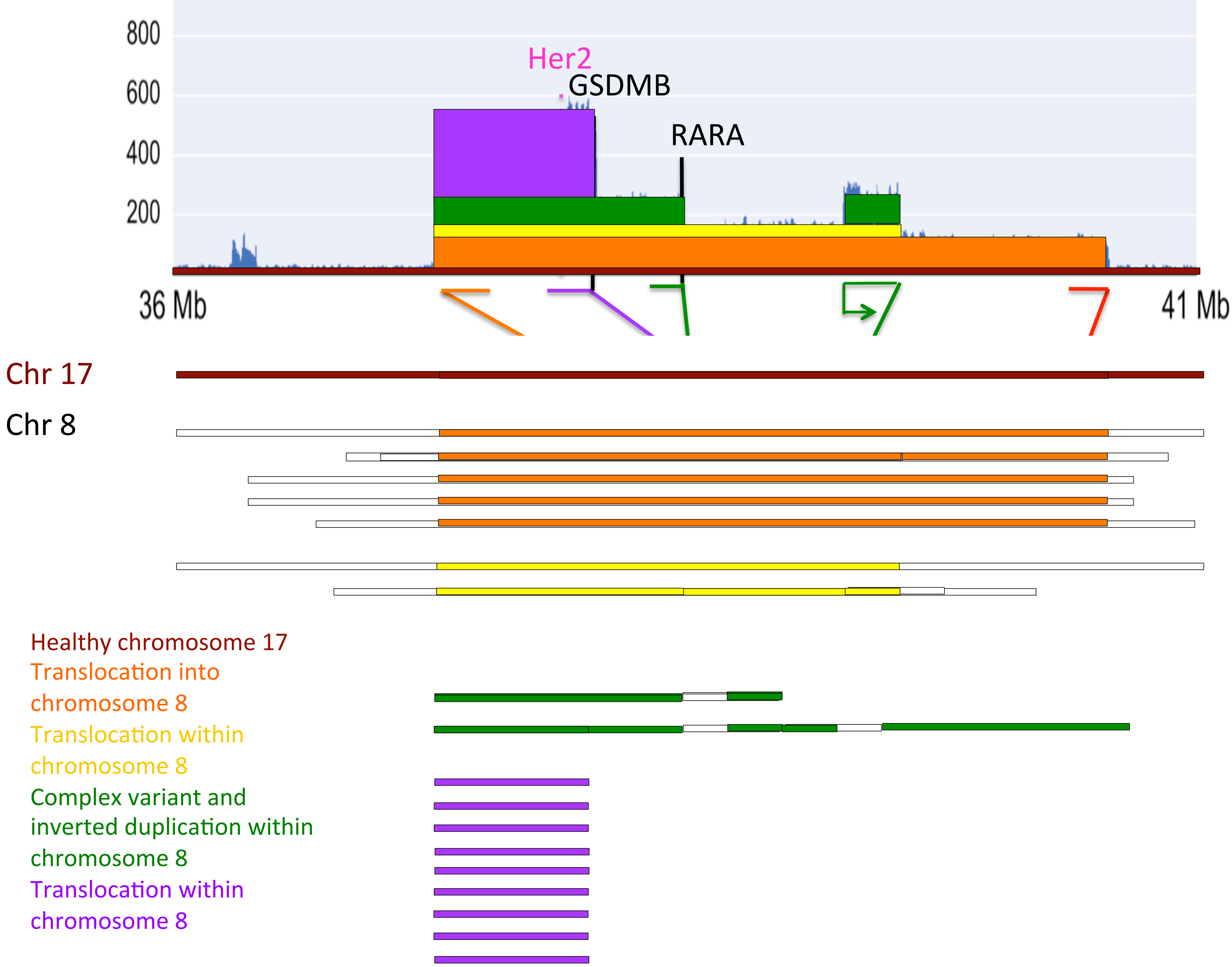
8 Mb


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
700
600
500
400
300
200
100
Her2
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

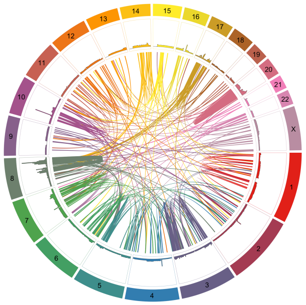
SplitThreader:
Graphical threading to retrace complex history of rearrangements in cancer genomes





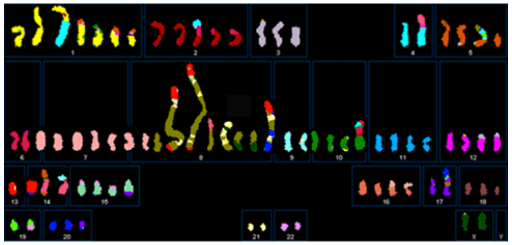
1. Healthy chromosome 17
2. Translocation into chromosome 8
3. Translocation within chromosome 8
4. Complex variant and inverted duplication within chromosome 8
5. Translocation within chromosome 8

Bridging the gap



variant-calling from sequencing

?

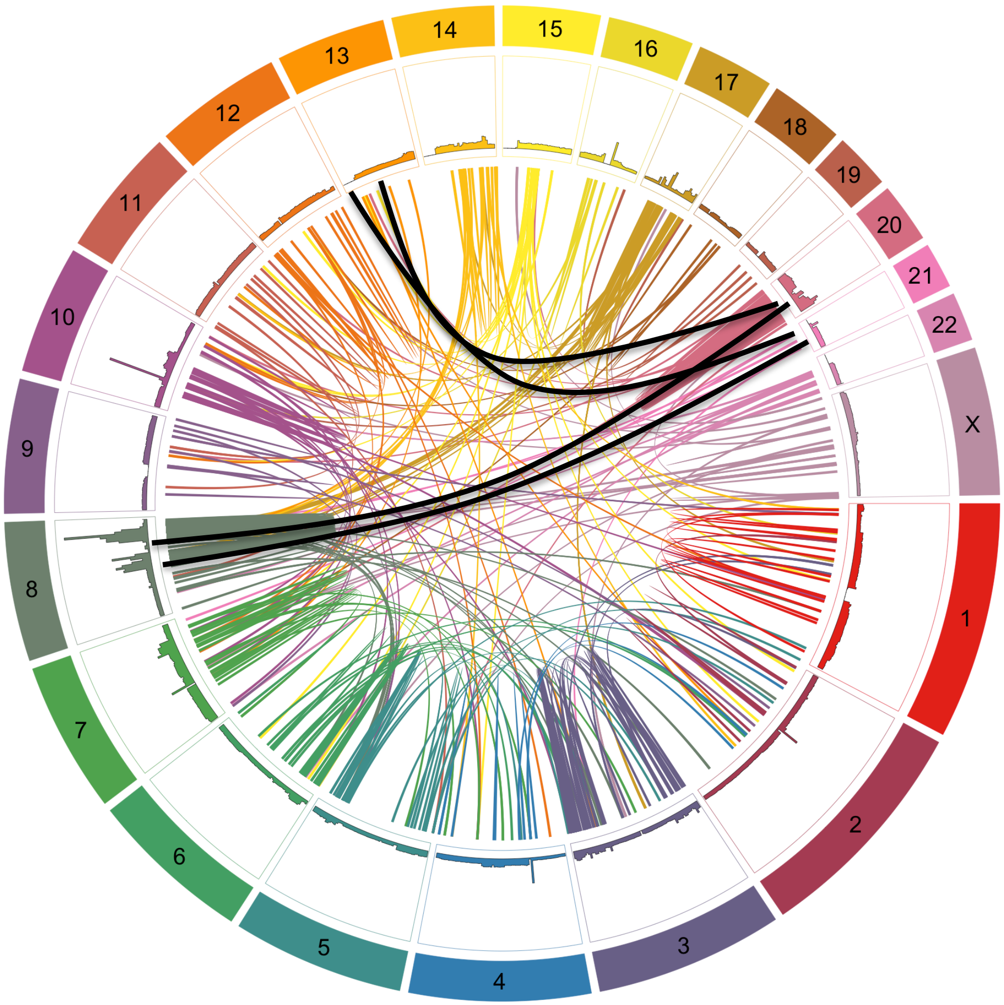


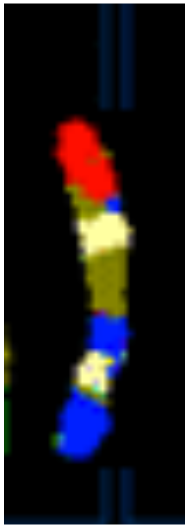
big-picture view from karyotyping

Context

Resolution

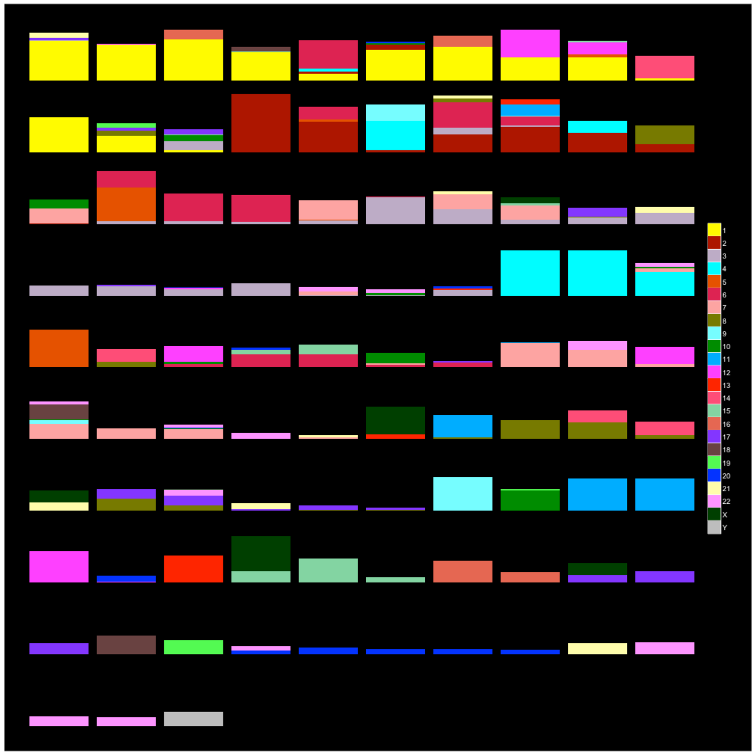
Threading through the whole-genome graph to produce a synthetic karyotype



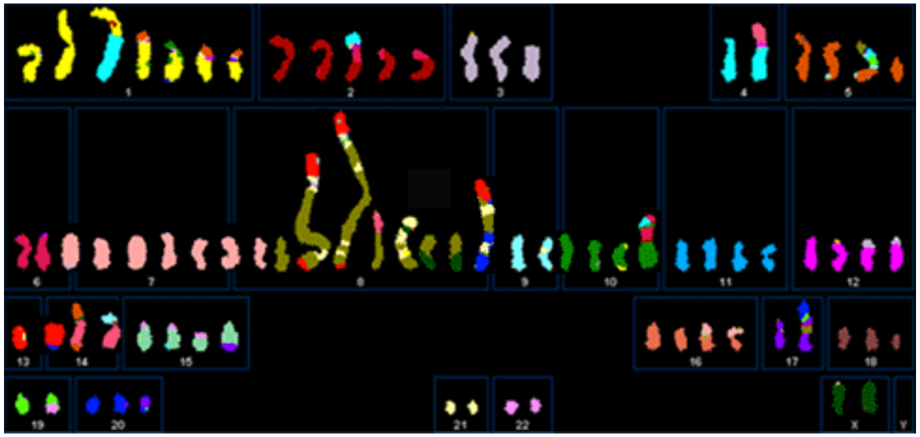


Synthetic karyotype with SplitThreader

Preliminary SplitThreader synthetic karyotype



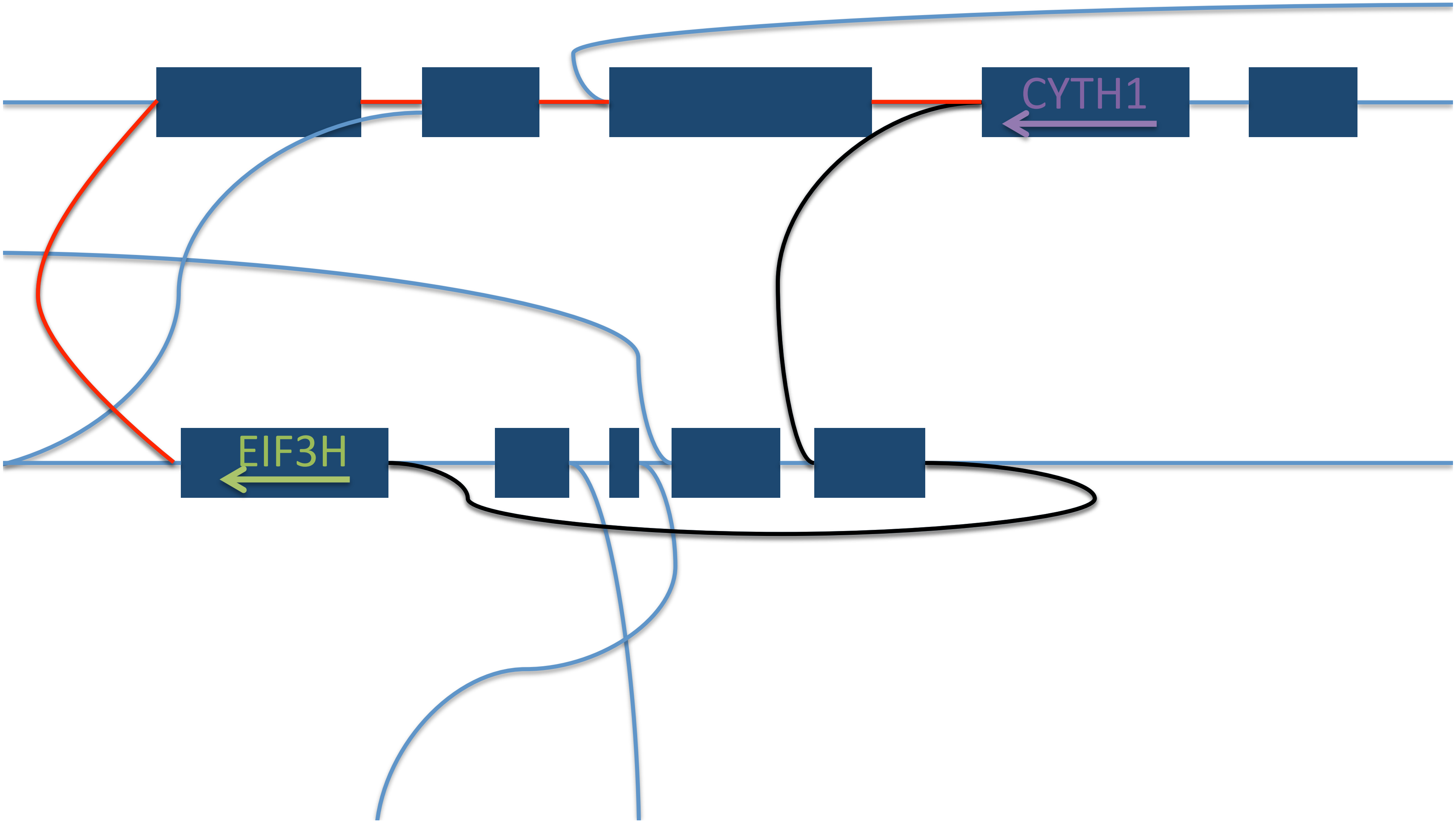
Real karyotype



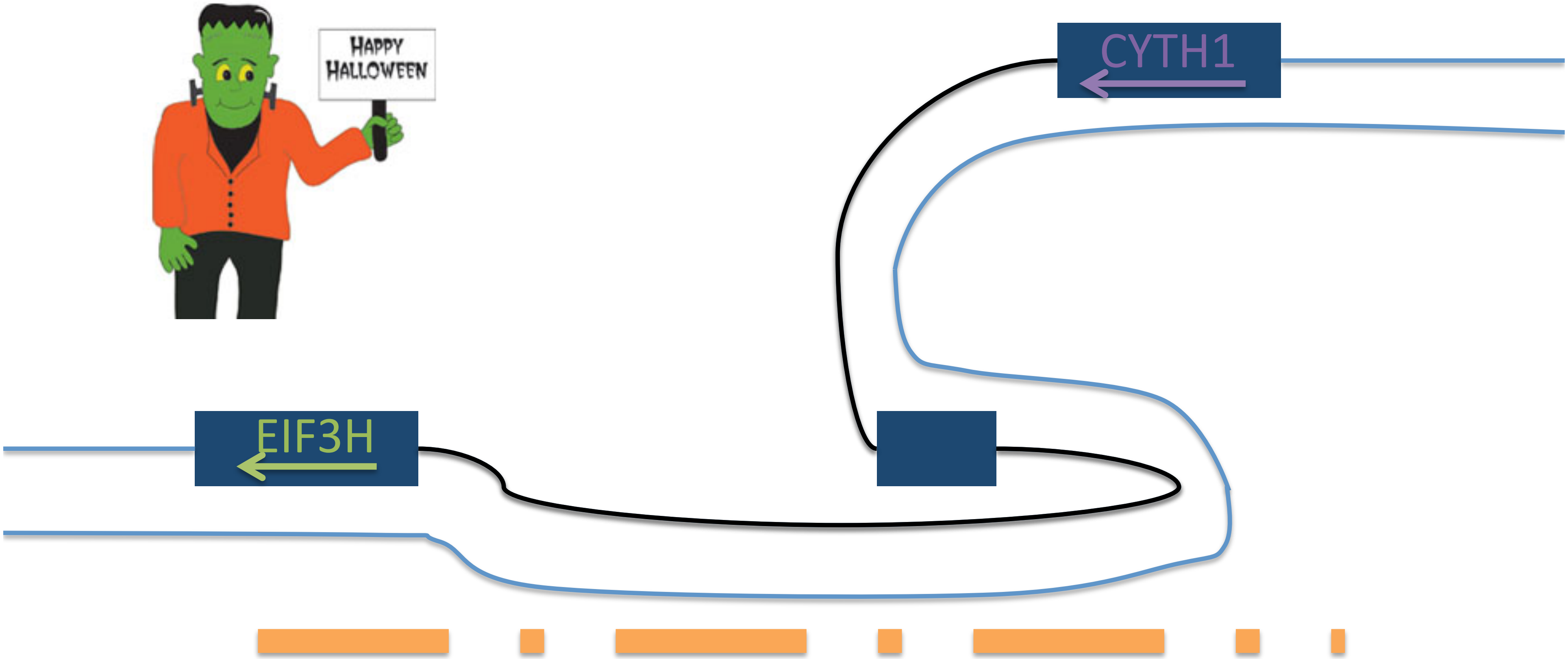
Transcriptome analysis with IsoSeq: Long-read RNA sequencing

- 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

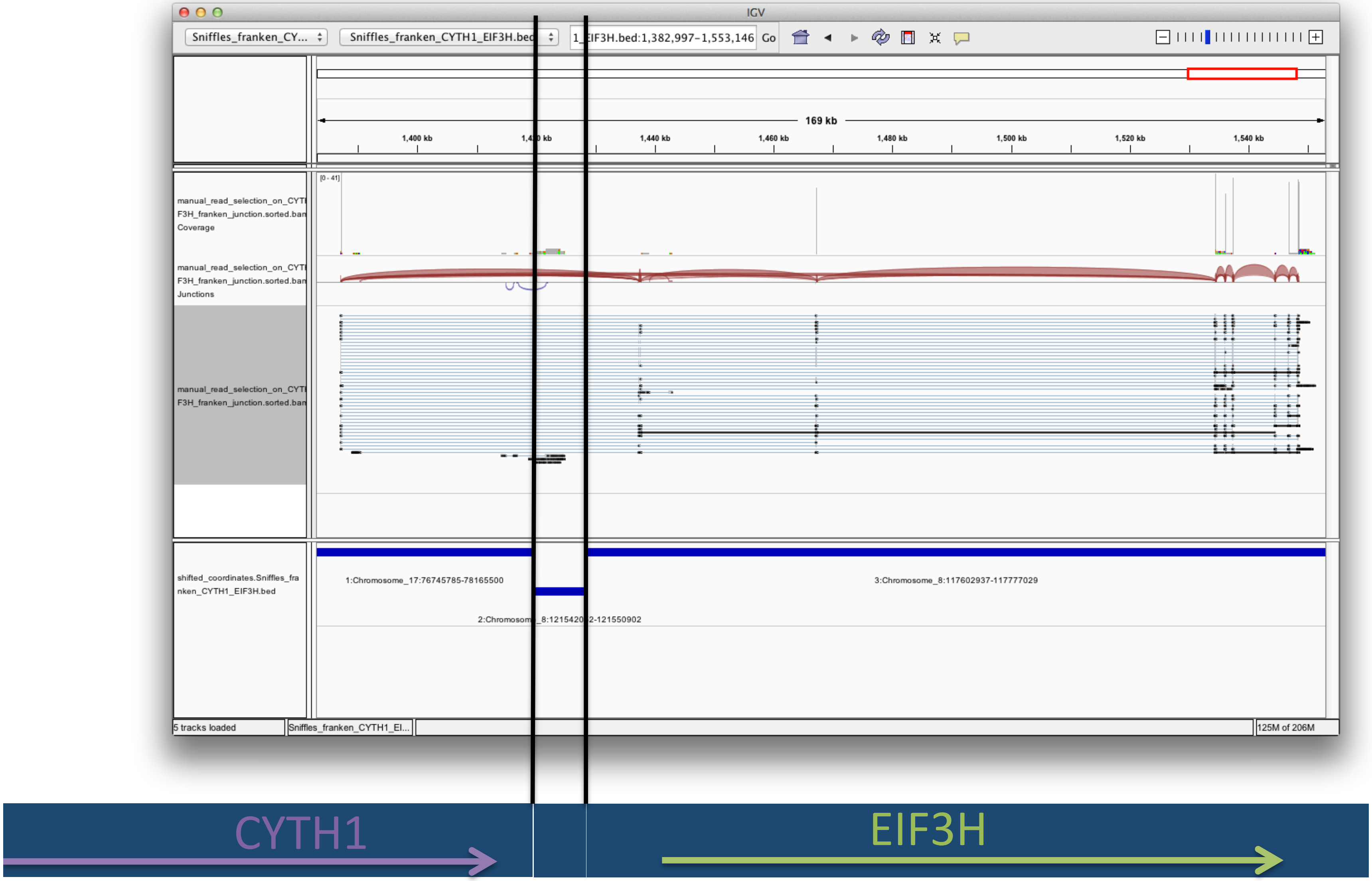
CYTH1-EIF3H gene fusion in the SplitThreader graph



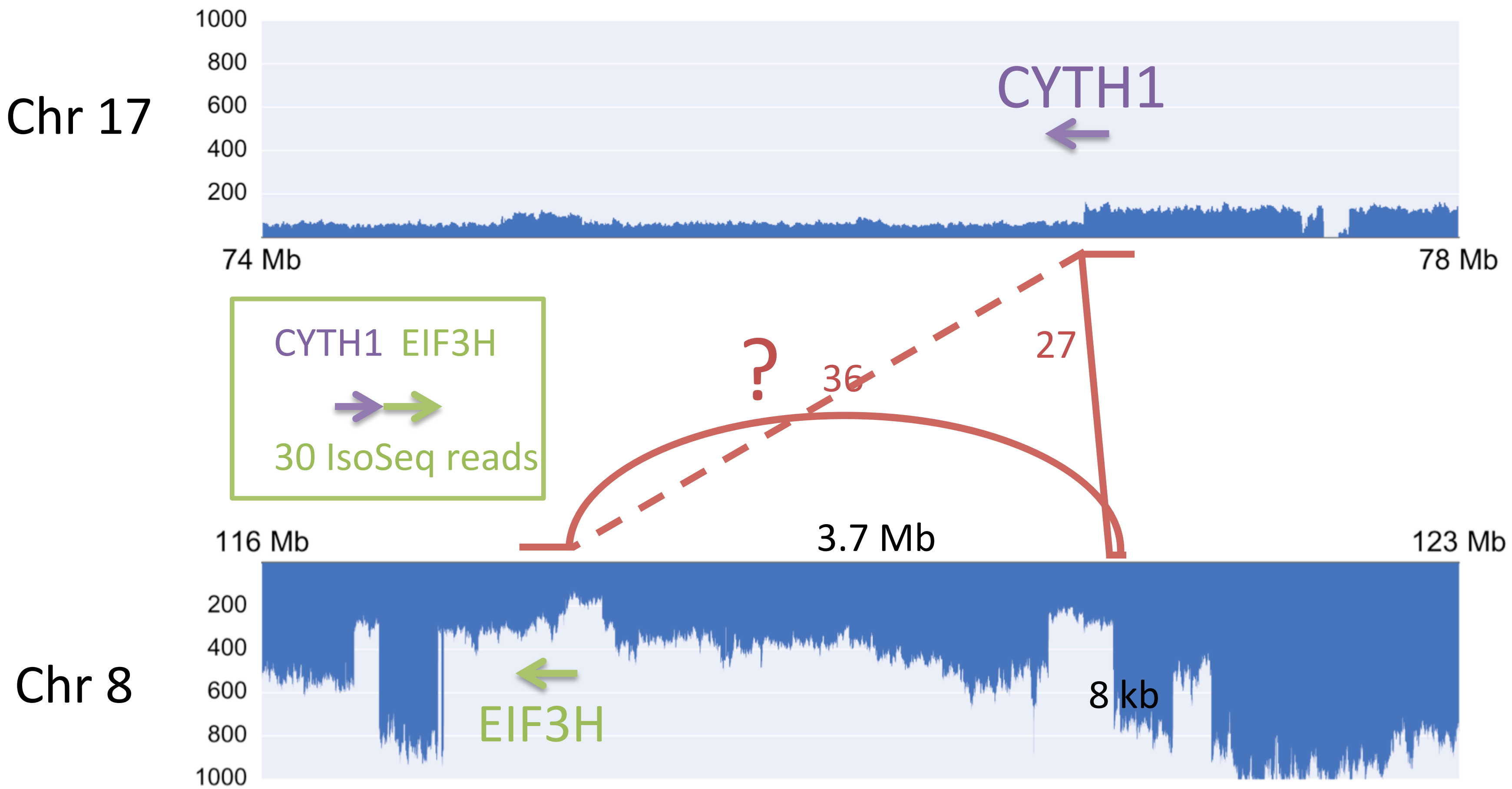
CYTH1-EIF3H gene fusion in the SplitThreader graph



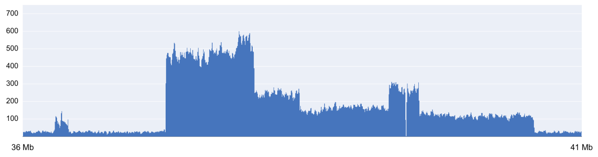
Frankensteining the CYTH1-EIF3H gene fusion


IGV
Sniffles_franken_CY...
Sniffles_franken_CYTH1_EIF3H.bed
1_EIF3H.bed:1,382,997-1,553,146
Go
169 kb
1,400 kb
1,4 0 kb
1,440 kb
1,460 kb
1,480 kb
1,500 kb
1,520 kb
1,540 kb
[0 - 41]
manual_read_selection_on_CYTH1_EIF3H_franken_junction.sorted.bam Coverage
manual_read_selection_on_CYTH1_EIF3H_franken_junction.sorted.bam Junctions
manual_read_selection_on_CYTH1_EIF3H_franken_junction.sorted.bam
shifted_coordinates.Sniffles_franken_CYTH1_EIF3H.bed
1:Chromosome_17:76745785-78165500
3:Chromosome_8:117602937-117777029
2:Chromosom _8:1215420 2-121550902
5 tracks loaded
Sniffles_franken_CYTH1_EI...
125M of 206M
CYTH1
EIF3H

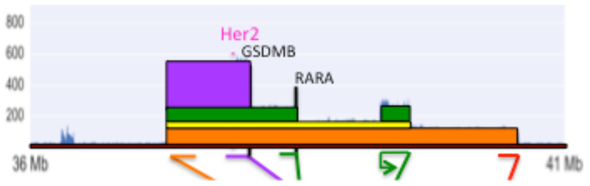
CYTH1-EIF3H gene fusion

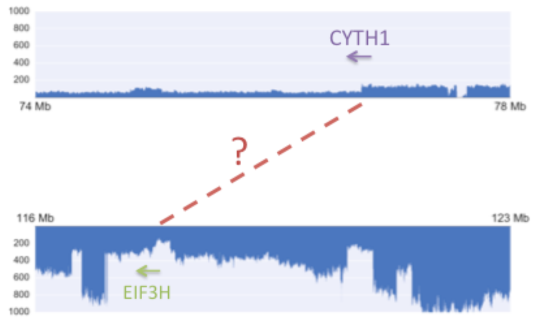


The genome informs the transcriptome

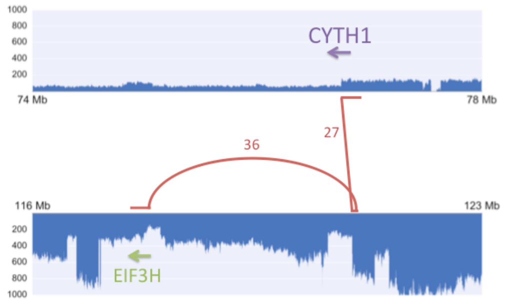


Explain amplifications





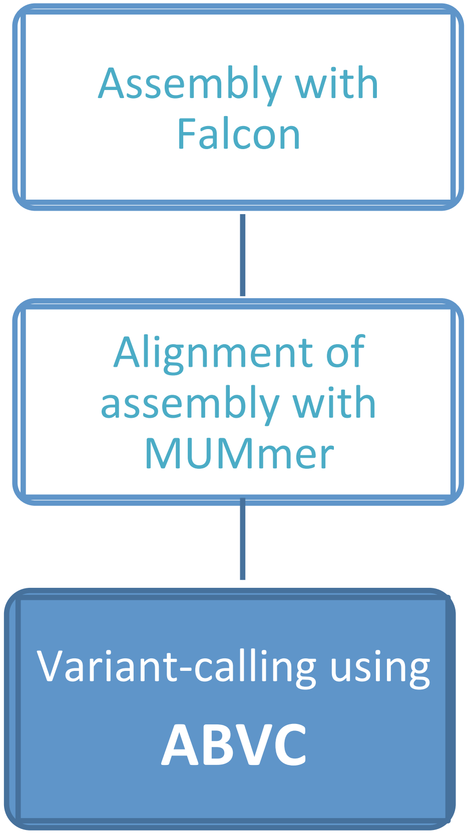
Trace gene fusions



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

New software in development for long-read genome analysis

Assembly-based analysis

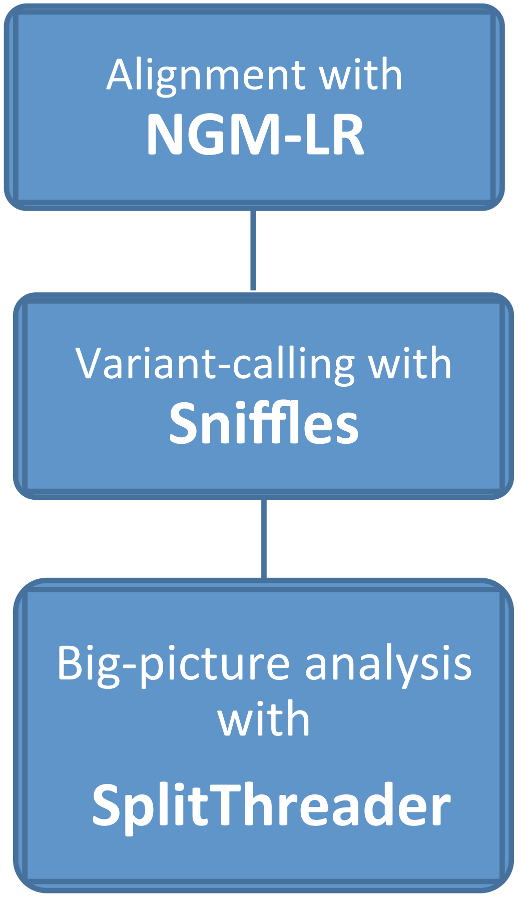


Assembly with Falcon

Alignment of assembly with MUMmer

Variant-calling using **ABVC**

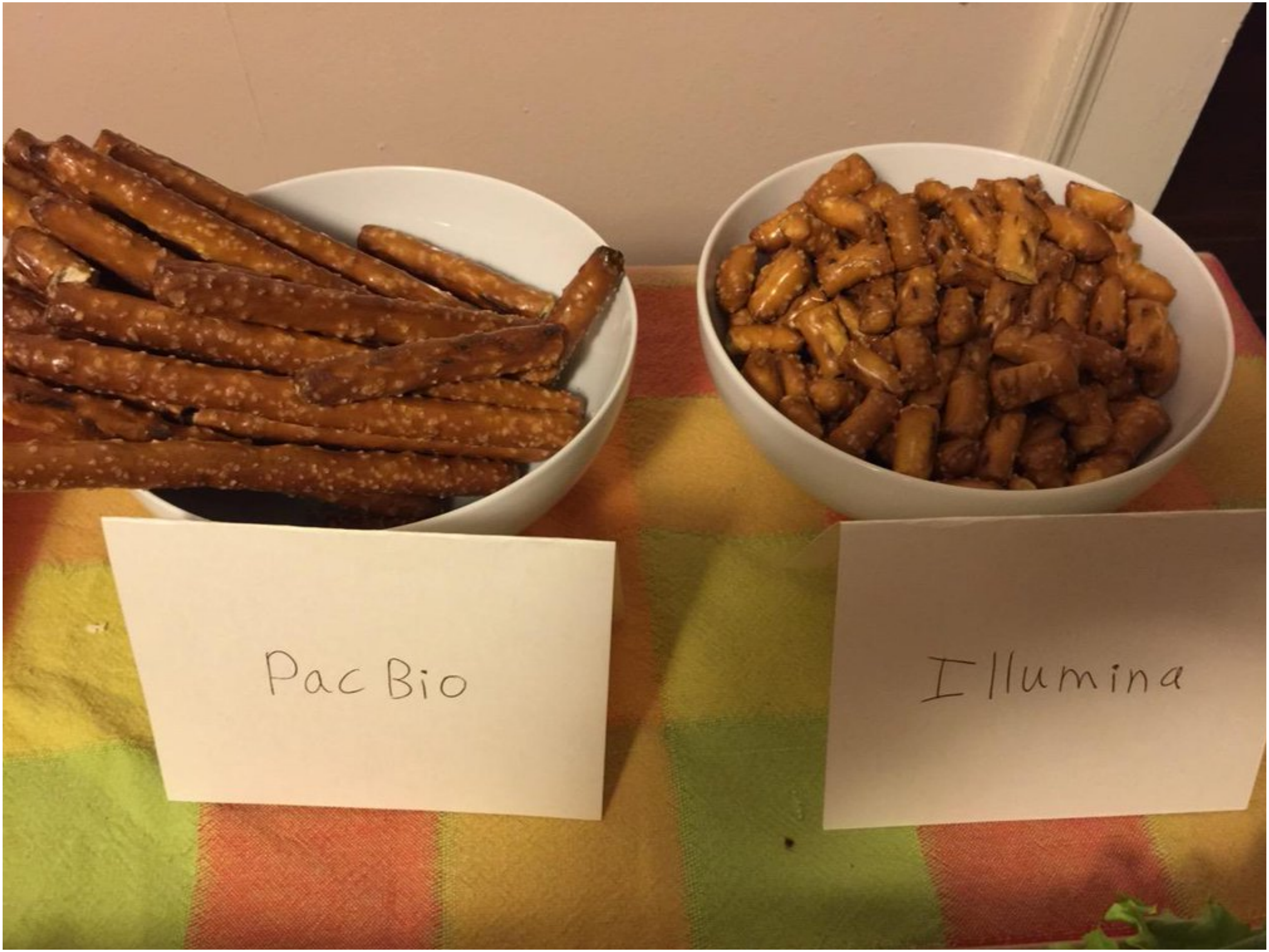
Alignment-based analysis



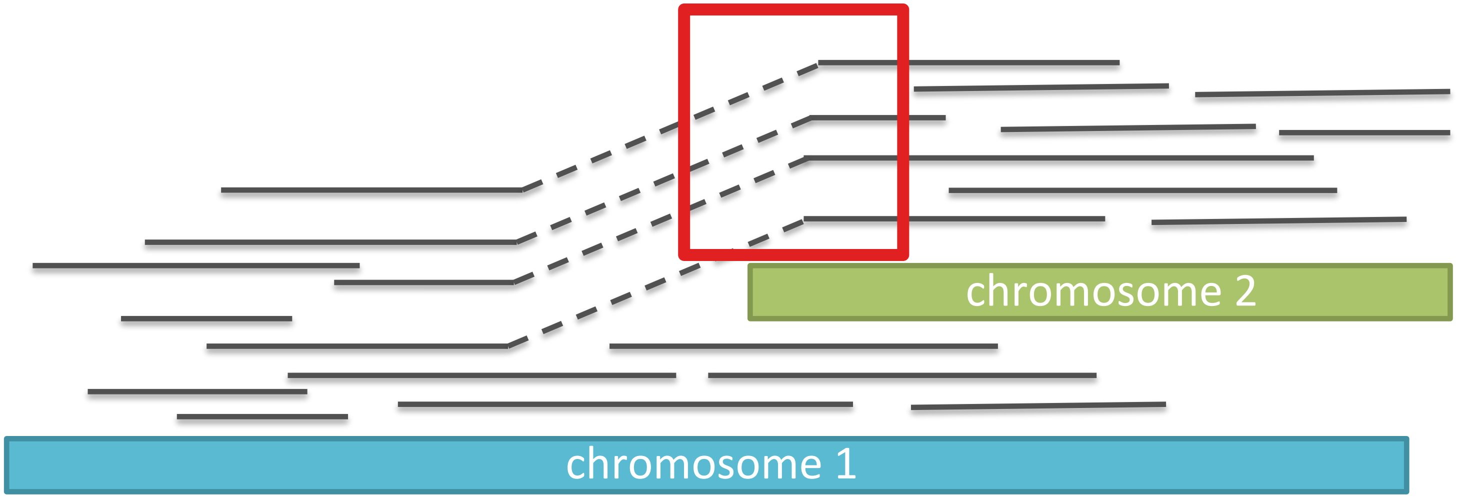
Alignment with **NGM-LR**

Variant-calling with **Sniffles**

Big-picture analysis with **SplitThreader**


PacBio
Illumina

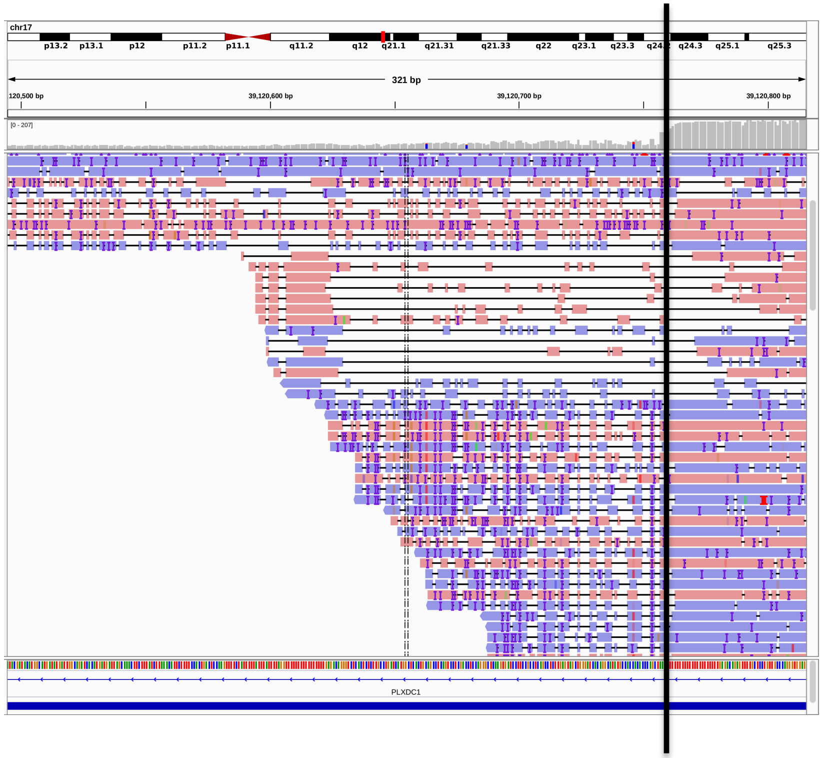
Zooming in on the breakpoint



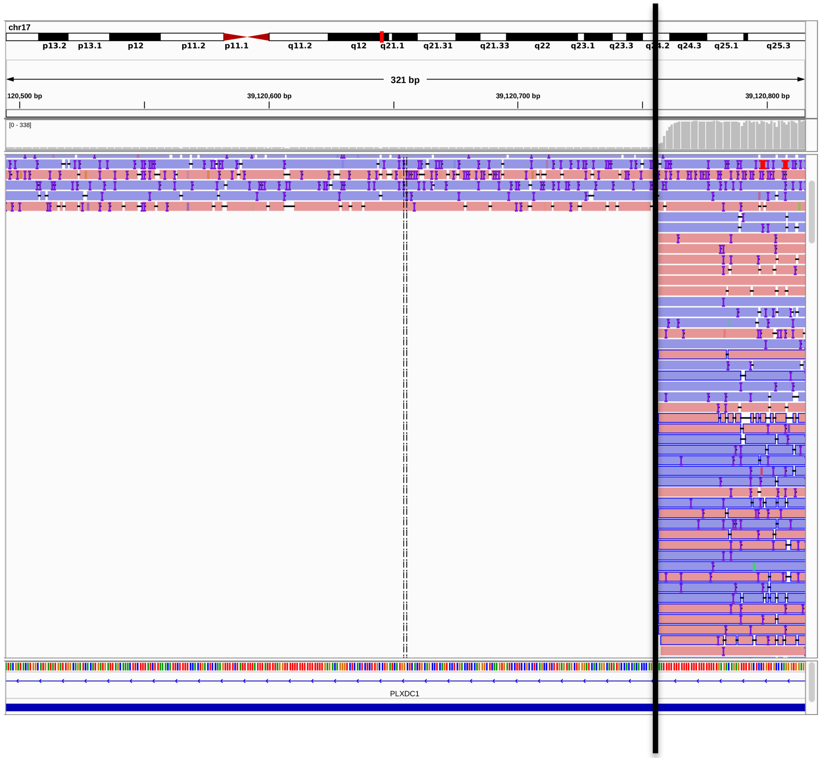
New aligner NGM-LR narrows down the breakpoint to base-pair resolution

Philipp Rescheneder

BWA-MEM



NGM-LR

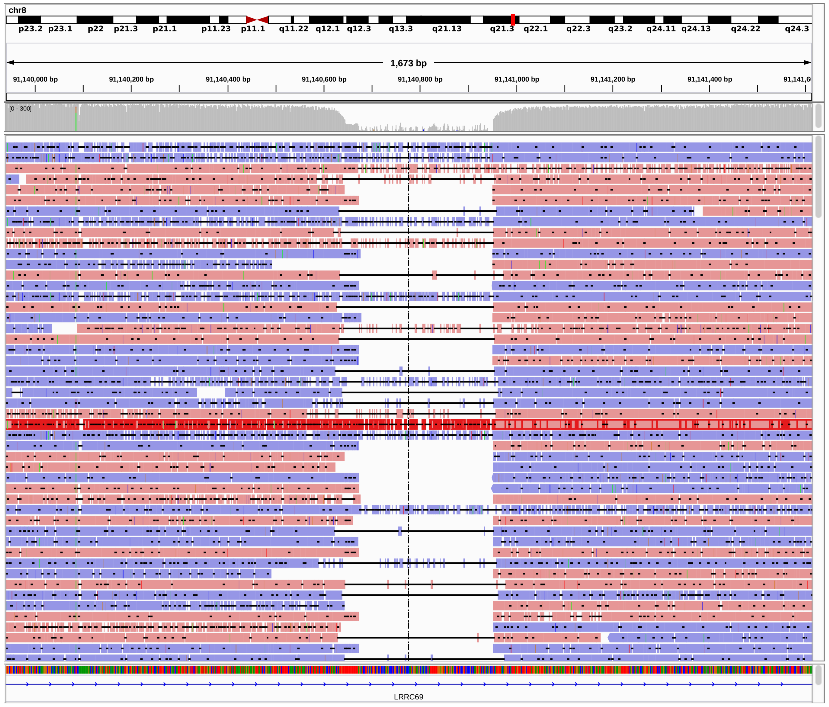


One side of an interchromosomal translocation

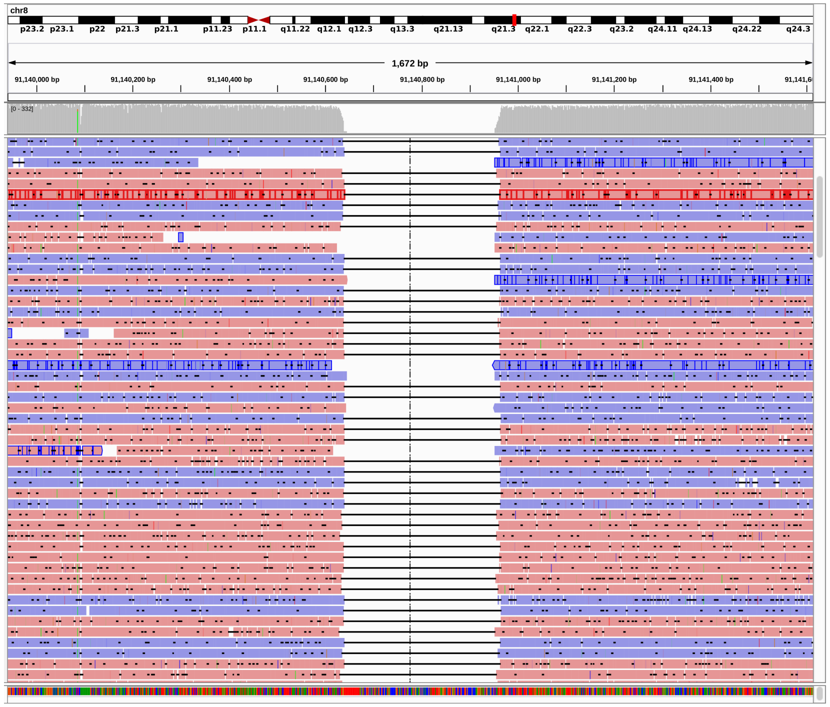
NGM-LR also enables better small variant calling

BWA-MEM

NGM-LR



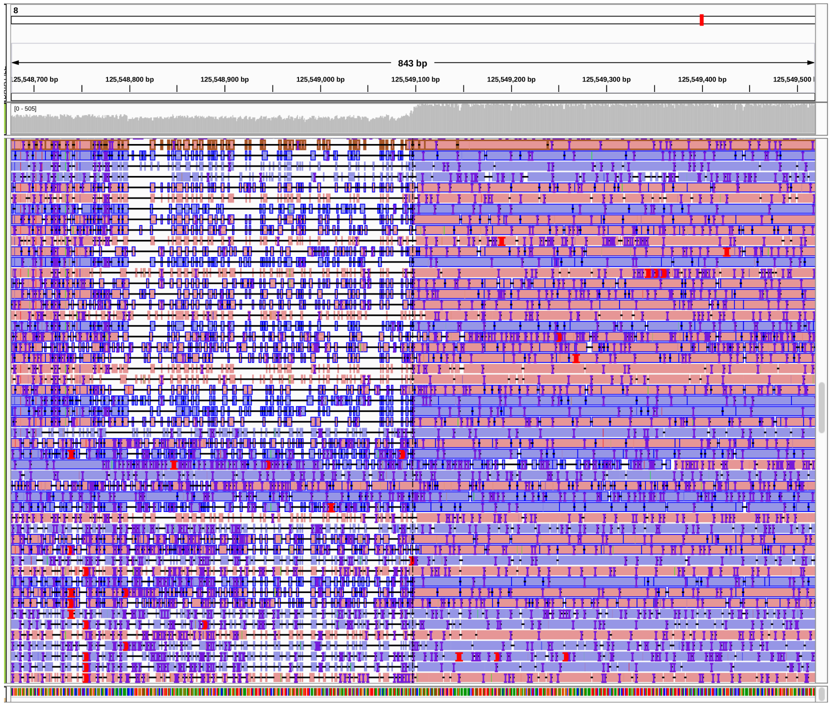
deletion



deletion

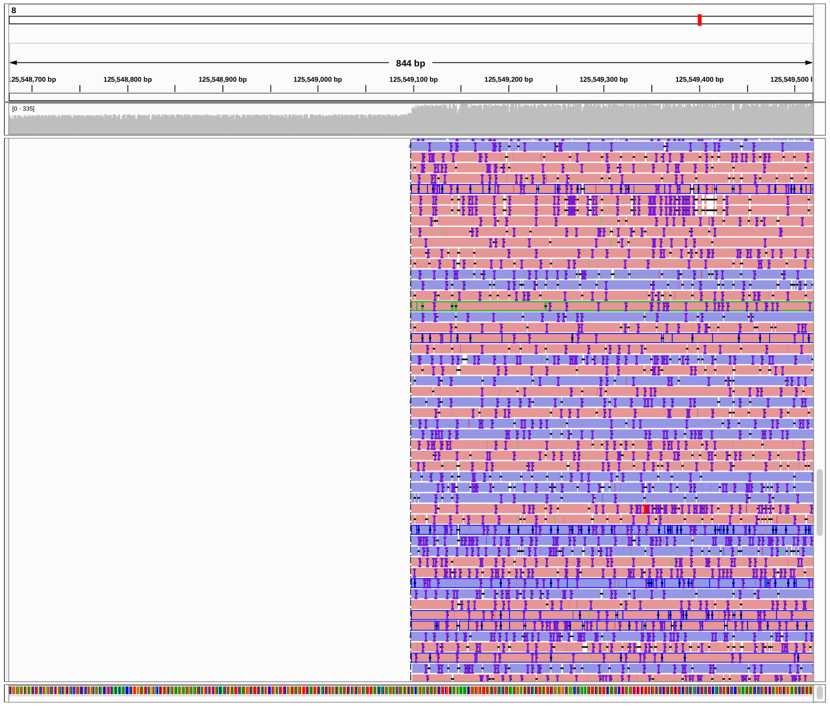
Without NGM-LR, alignments can be smudged over hundreds of base-pairs away from the breakpoint

BWA-MEM



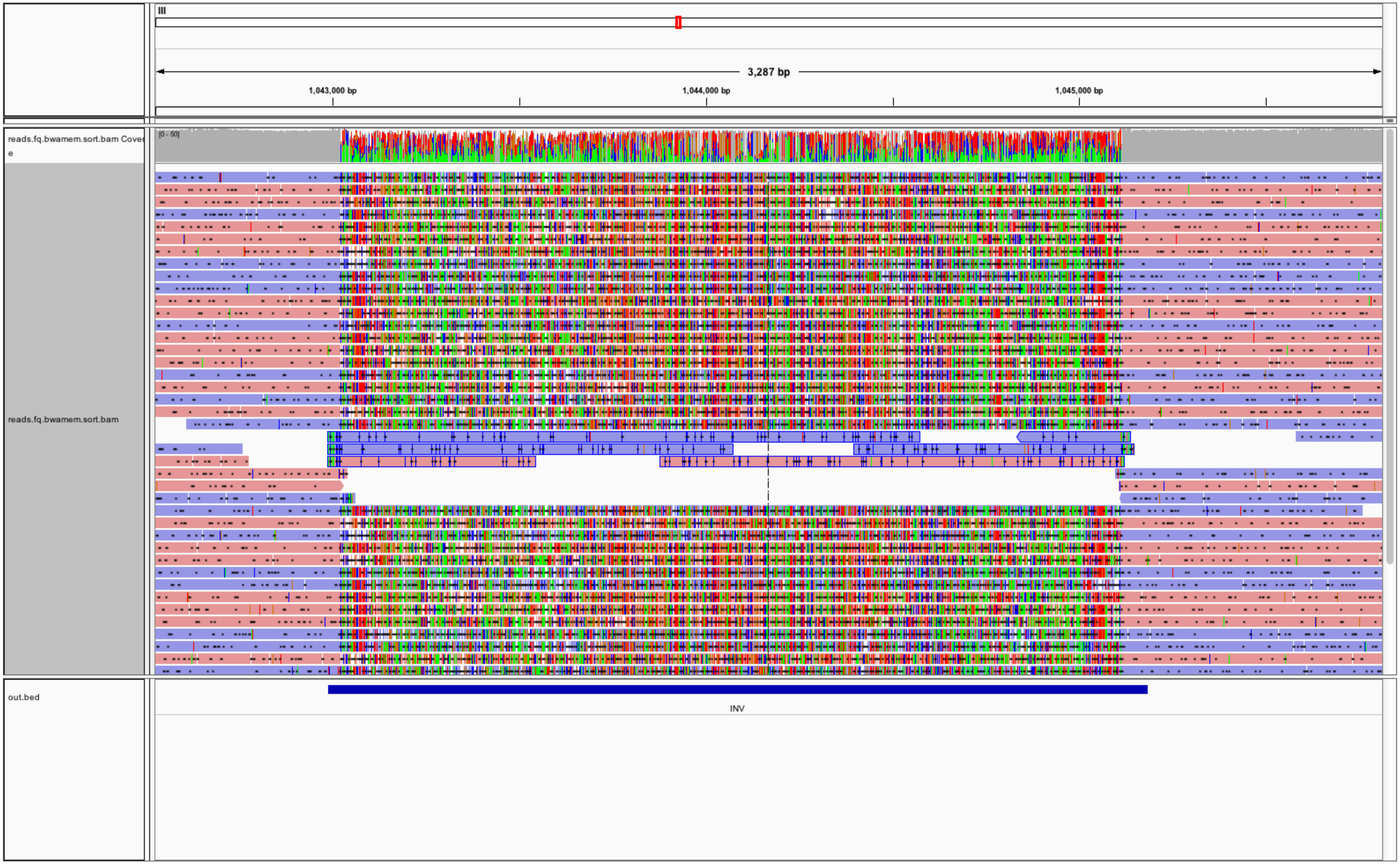
translocation

NGM-LR



translocation

Inversion in BWA-MEM



Acknowledgments

CSH 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

OICR Ontario Institute for Cancer Research

Karen Ng
Timothy Beck
Yogi Sundaravadanam
John McPherson

DNAnexus

PACIFIC BIOSCIENCES®

Elizabeth Tseng
Jason Chin

Genentech FOUNDATION