Detection and interpretation of structural variation from whole genome sequence data

Ira Hall
Associate Professor
The Genome Institute
Washington University School of Medicine

CSHL Sequencing Course, 2014

Genome structural variation

<u>Structural Variation (SV):</u> Differences in the copy number, orientation or location of "large" genomic segments (e.g., >50 bp)

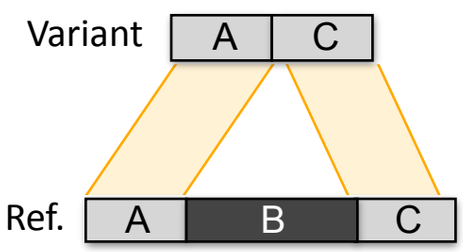
<u>Other terms:</u>

Genomic Rearrangements

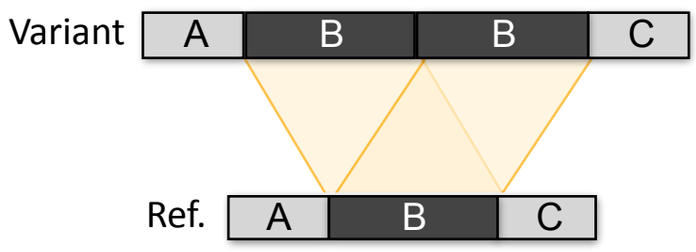
Copy Number Variants (CNVs)

Mobile Element Insertions (MEIs)

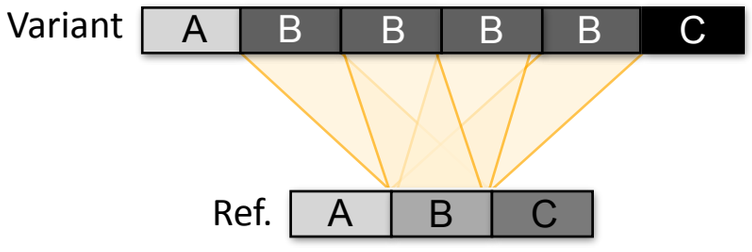
Deletion



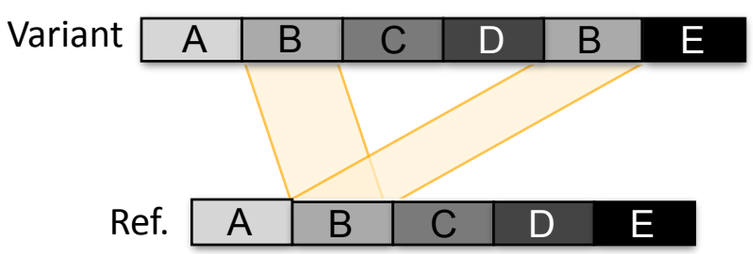
Tandem Duplication



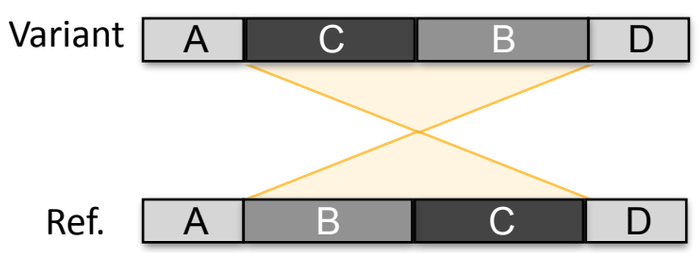
Amplification



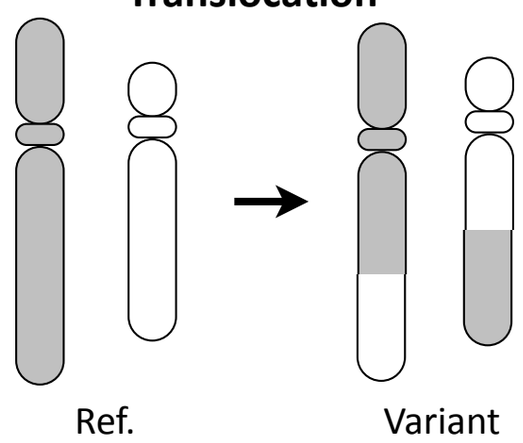
Interspersed duplication



Inversion



Translocation



Why map structural variation?

1) SV is a common form of genetic variation

- 5,000 variants between two humans; SVs are present in all species.
- Large functional potential relative to point mutations.
- Major source of mutation in cancer genomes: gene amplification, gene deletion, gene fusion.

2) To map traits

- Inherited causal variants may not be well-tagged by SNPs.
- Some traits are caused by spontaneous mutation (e.g., cancer, various sporadic human disorders).

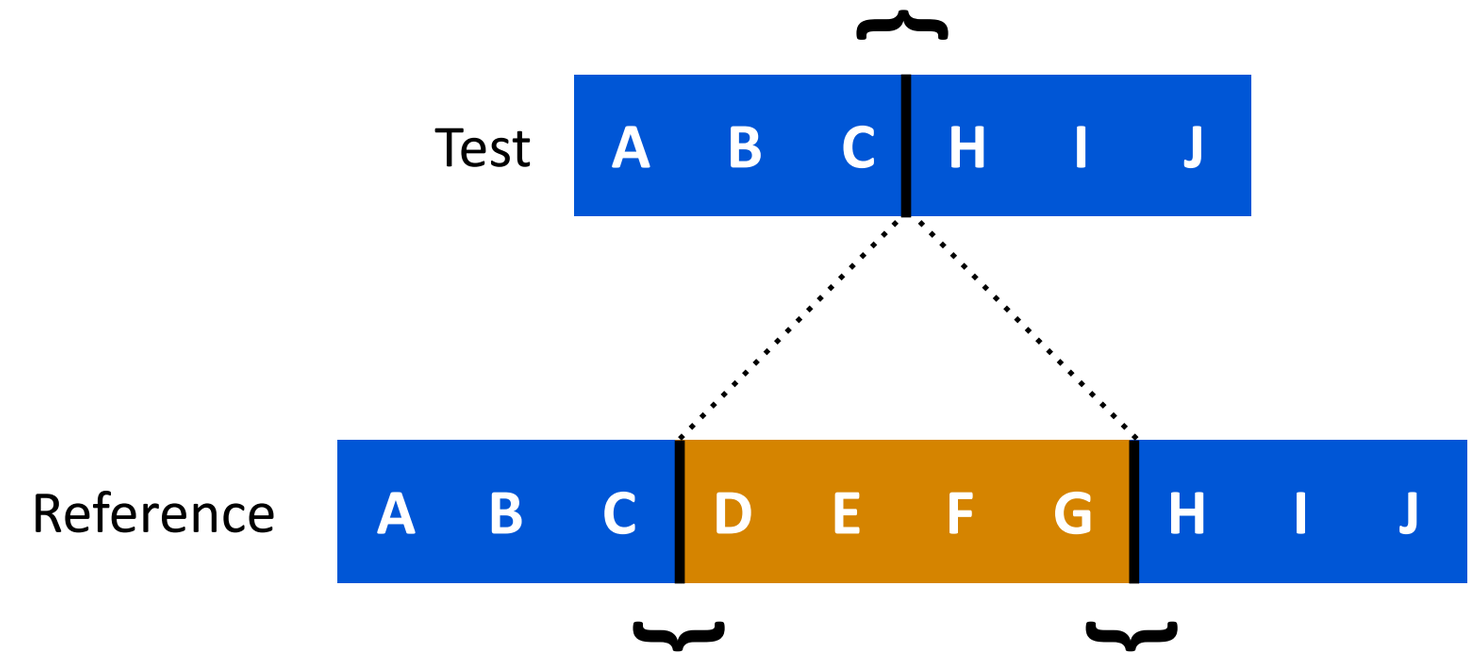
3) To understand genome/tumor/species evolution

- Gene birth, gene dosage, effects on recombination.
- Evolution of genome architecture.
- Local and global genomic instability.
- Punctuated genome evolution.

SV breakpoints defined

Breakpoints are the junctions that define structurally variable genomic segments. SVs are usually detected based on the presence of these junctions the experimentally sequenced "test" genome, but not the reference genome.

However, "breakpoint" is an ambiguous term because it can simultaneously describe one junction in the test genome, and two junctions in the reference genome.

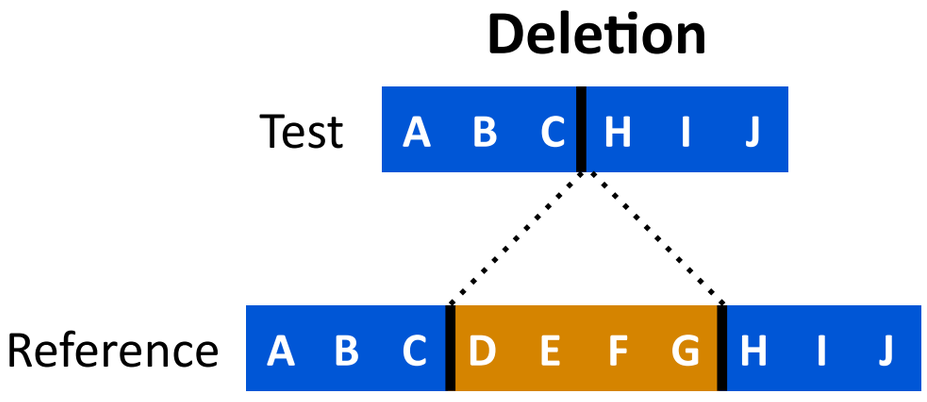


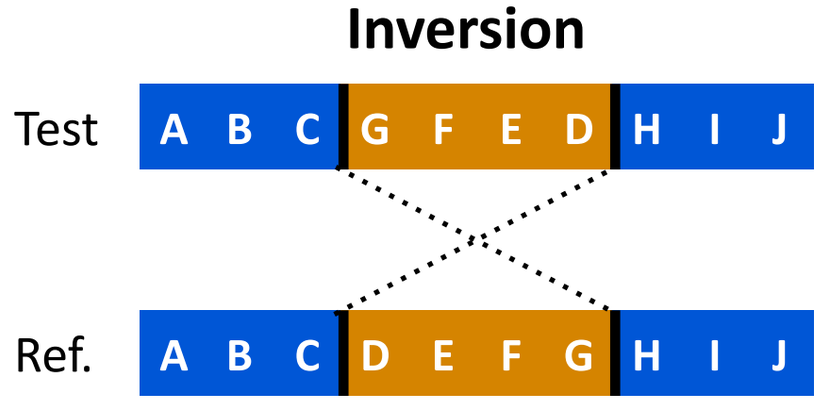
The VCF file format accounts for this ambiguity by introducing two new terms:

"**novel adjacency**": the breakpoint in the test genome

"**breakends**": the two breakpoints in the reference genome

Visualizing SV breakpoints

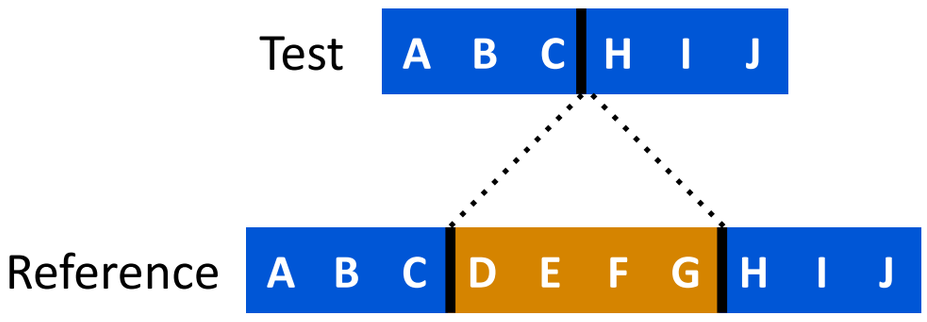




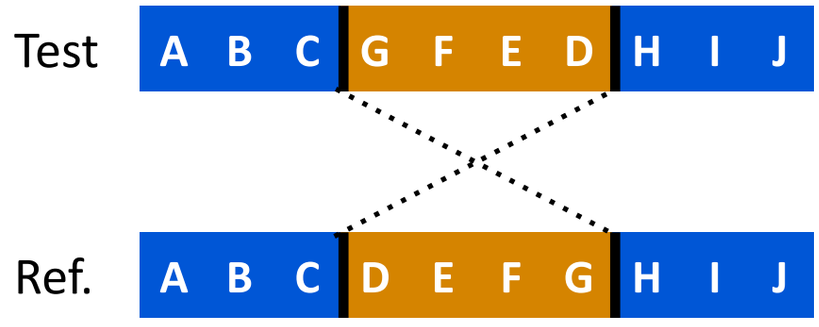
<u>NOTE</u>: Deletions produce one breakpoint (novel adjacency) in the test genome and two in the reference, whereas inversions produce two breakpoints in both genomes.

Visualizing SV breakpoints

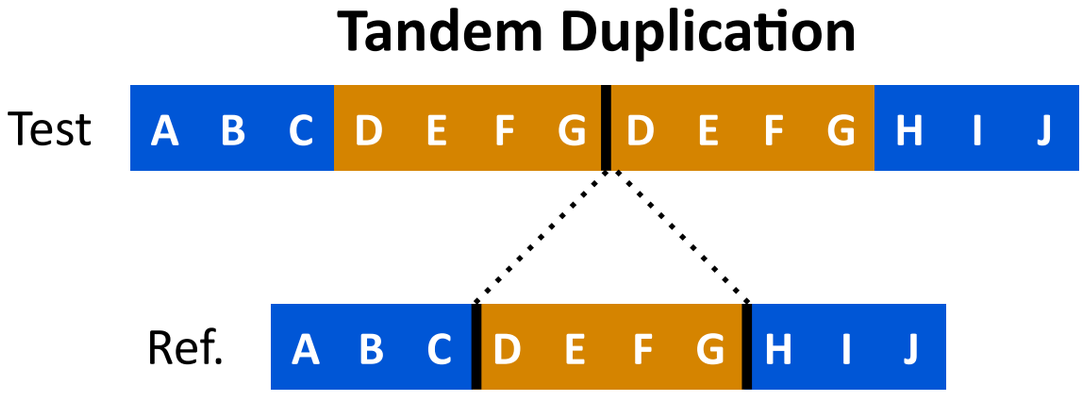
Deletion



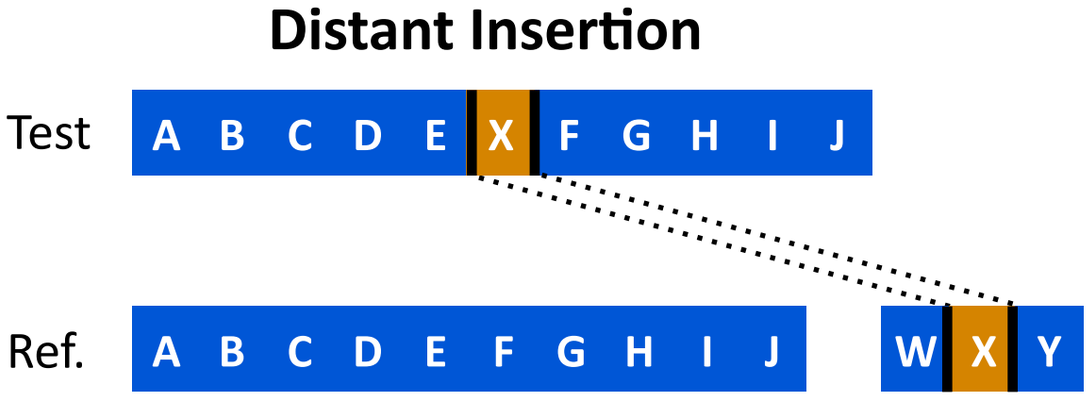
Inversion



Tandem Duplication

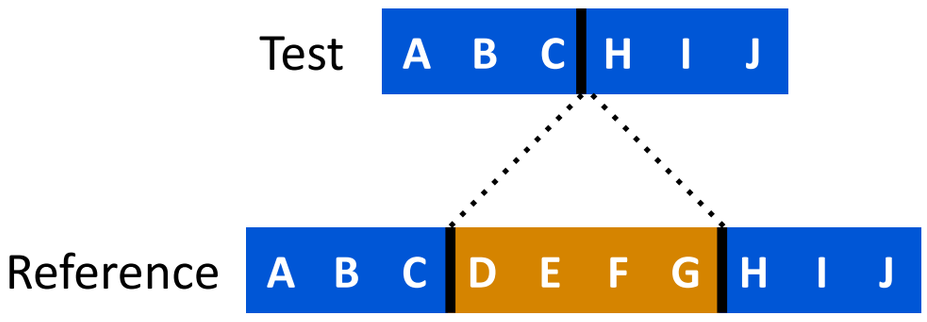


Distant Insertion

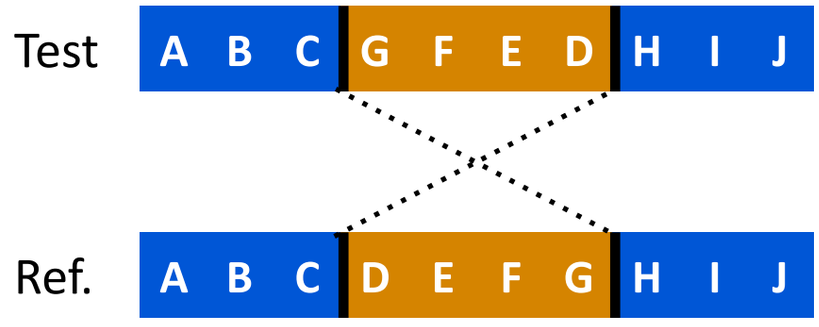


Visualizing SV breakpoints

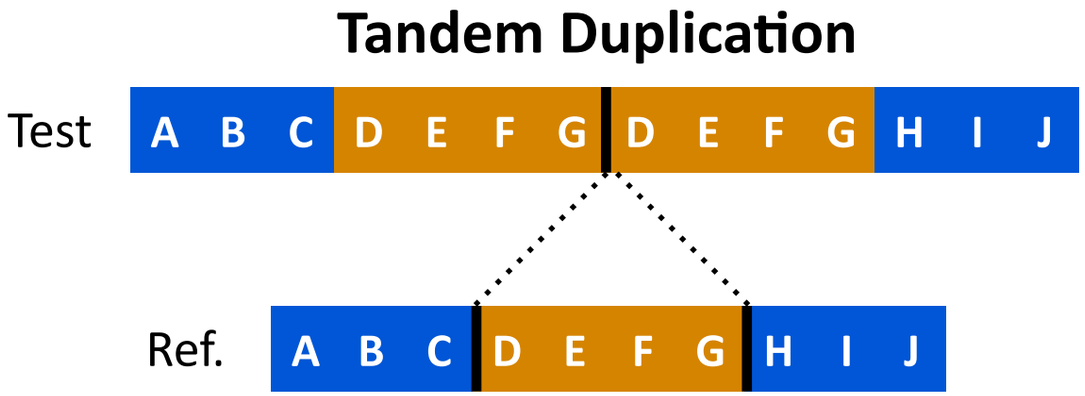
Deletion



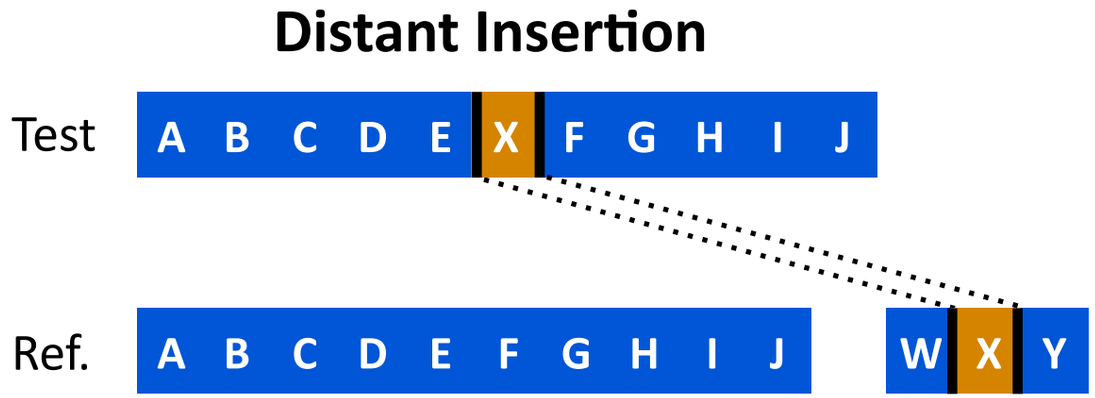
Inversion



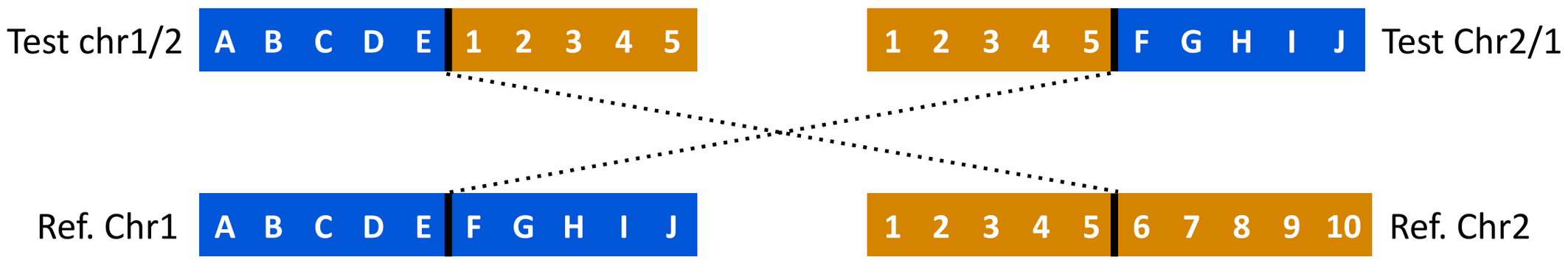
Tandem Duplication



Distant Insertion



Reciprocal translocation



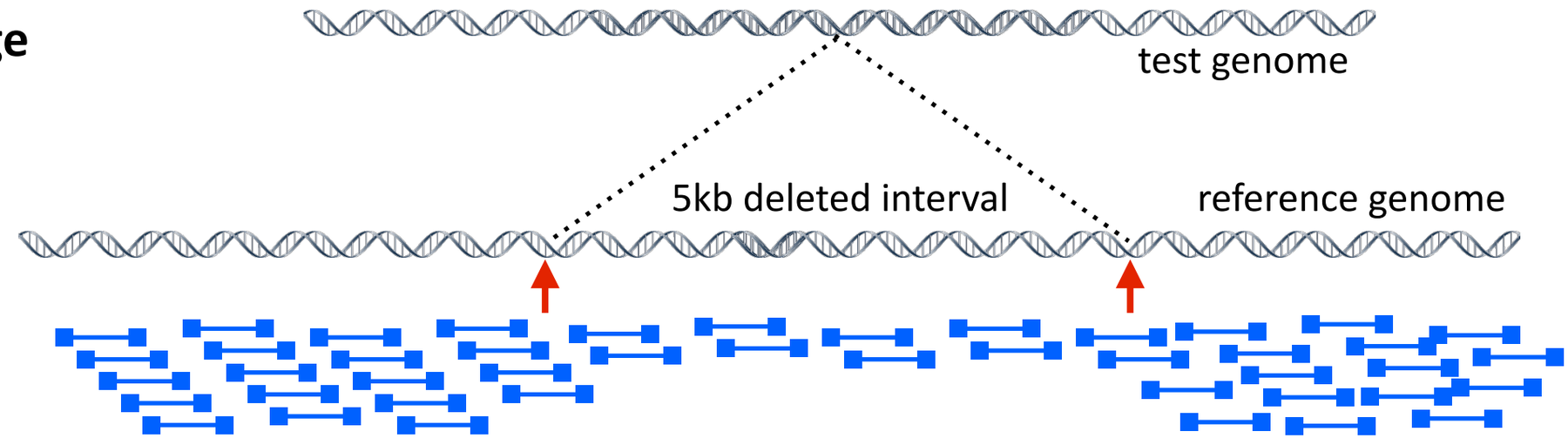
The ultimate goal: Assemble whole genomes, align whole genomes, reconstruct architecture

But we can't routinely do that yet. We need better DNA sequencing technologies

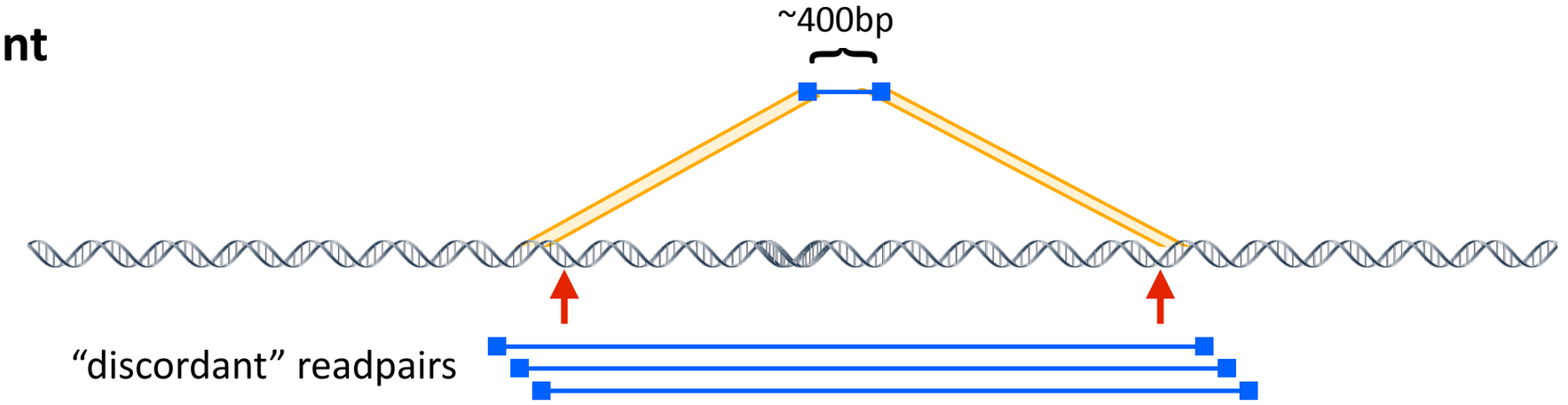
What can we do now?

3 ways to detect a structural variant (SV)

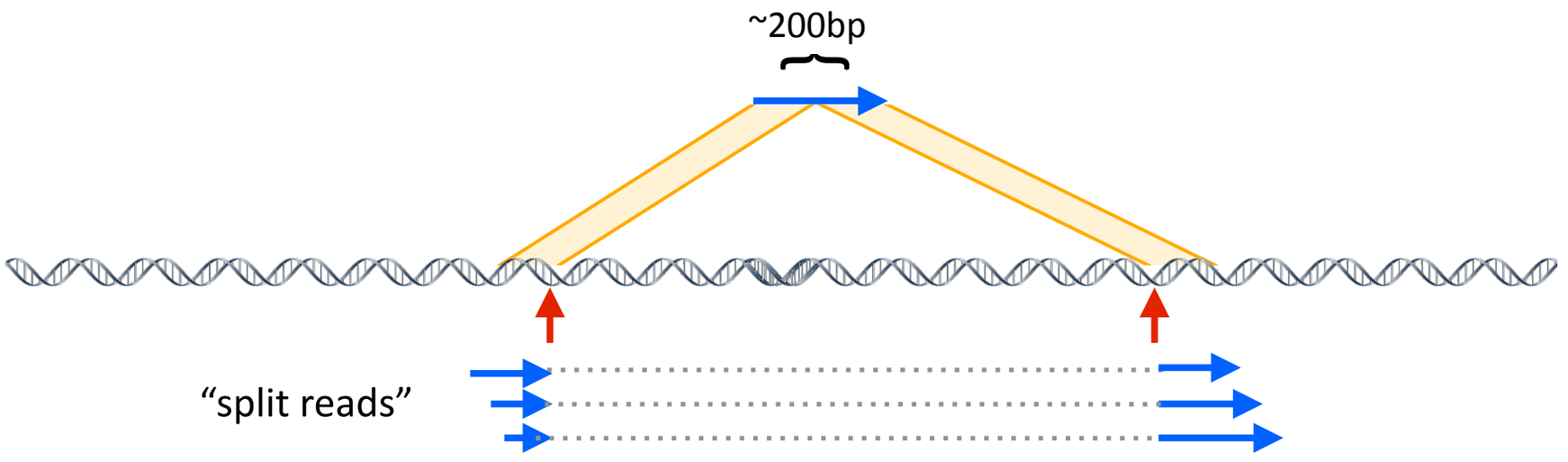
1) depth of sequence coverage
= "read-depth analysis"
(copy number alterations)



2) "readpairs" span a breakpoint
= "paired-end mapping"
(all classes of SV)



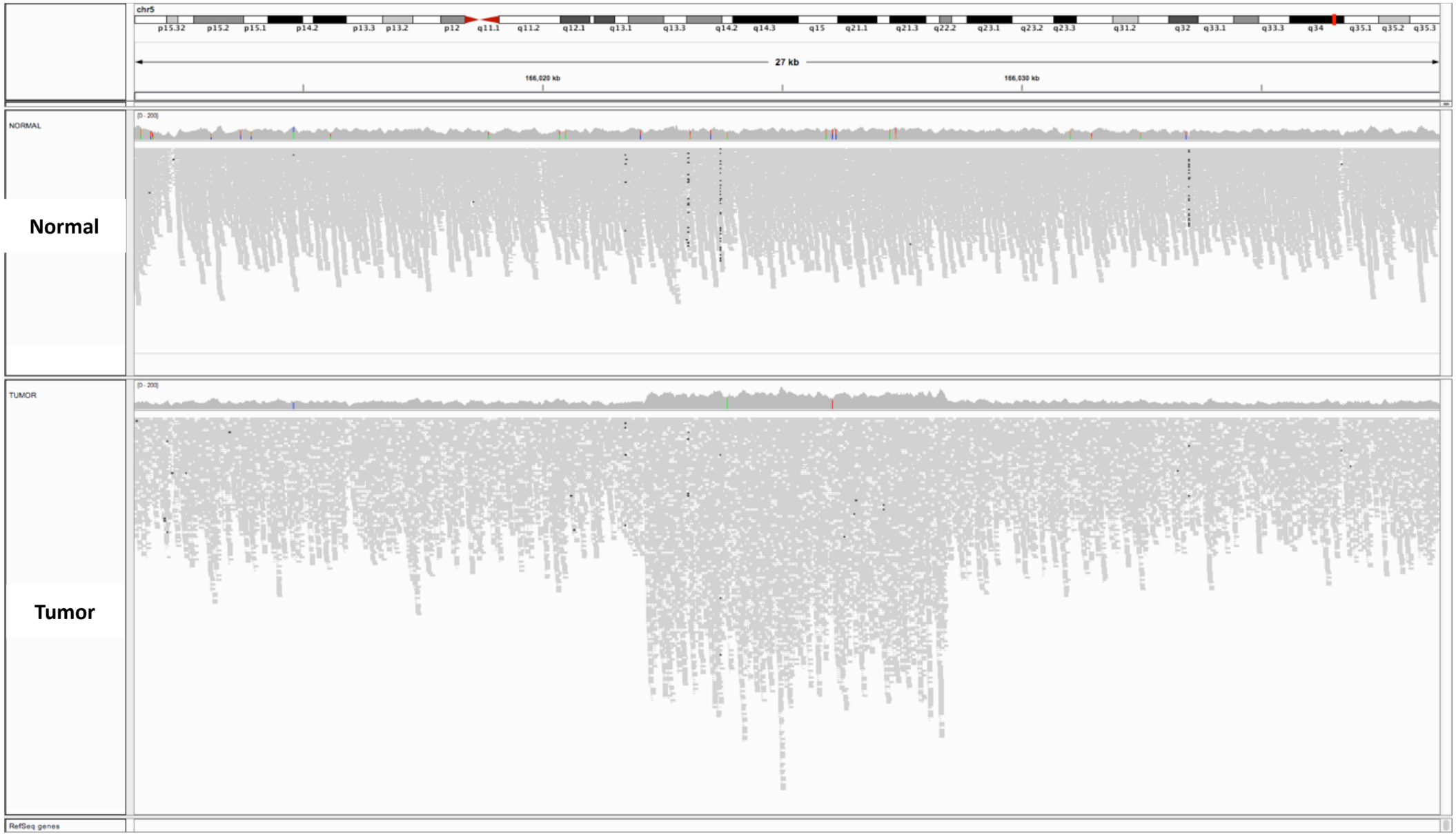
3) read contains a breakpoint
= "split-read mapping"
(all classes of SV)



1) Read-depth analysis

(a.k.a., depth of coverage)

Detecting CNVs with read-depth analysis



Basic approach:

1) Count reads in sliding windows (e.g., 1 kb) of uniquely mappable genome sequence.

2) Normalize for GC bias.

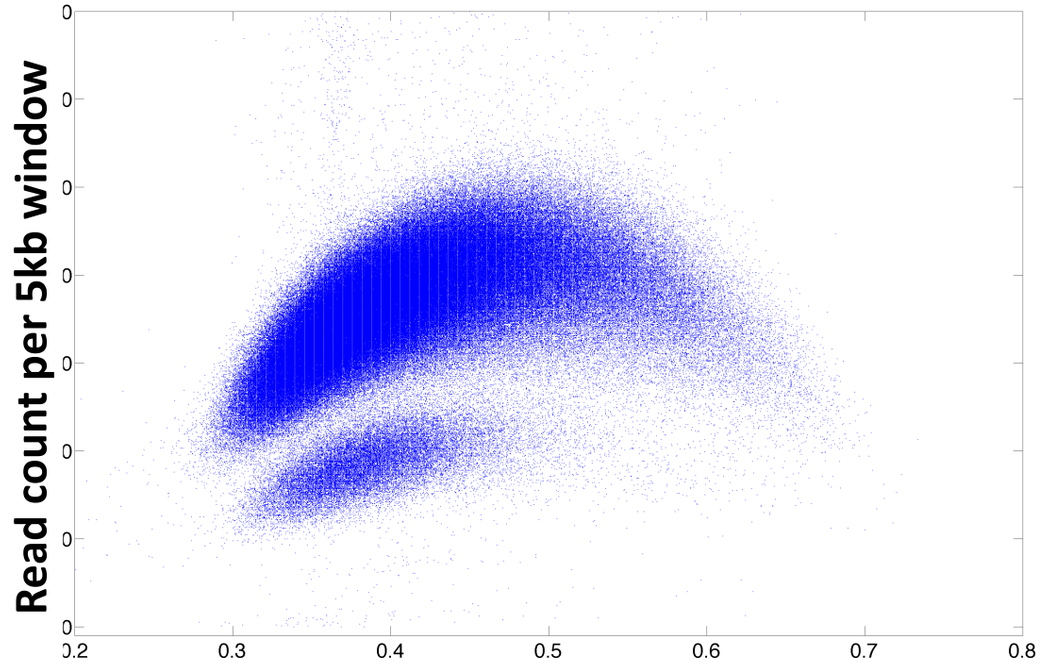
3) Use segmentation to define CNVs (similar to array-CGH data).

4) Some methods predict absolute copy number, others compare two samples.

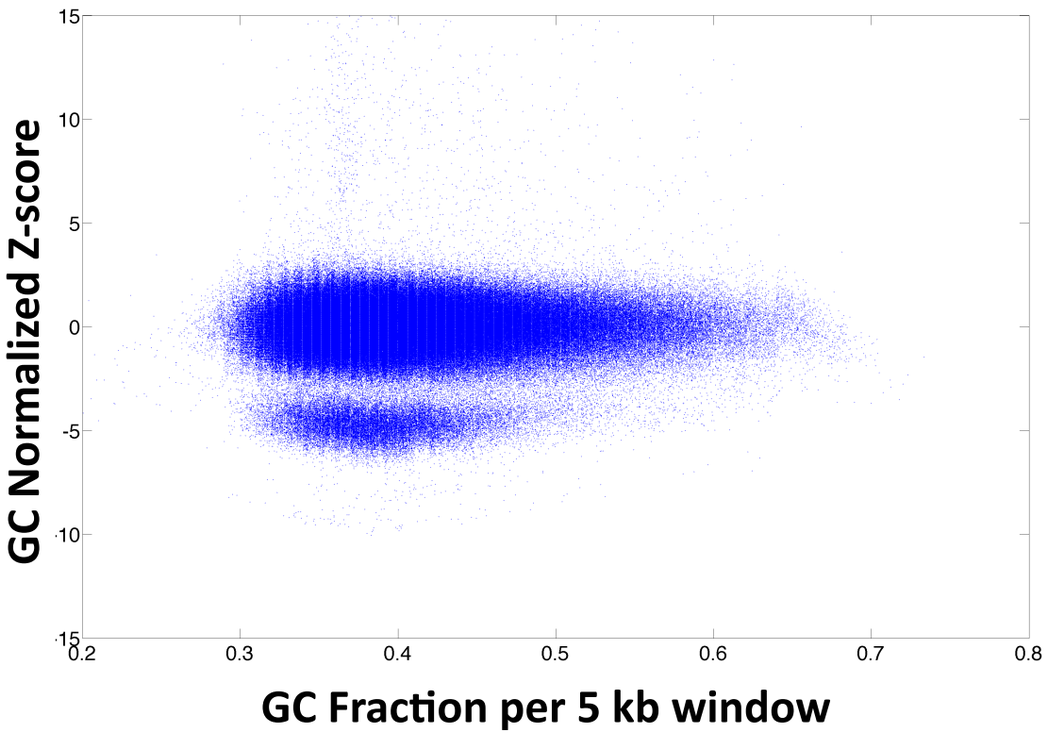
5) Lots of read-depth methods. What do we use? CNVnator + in-house tools.

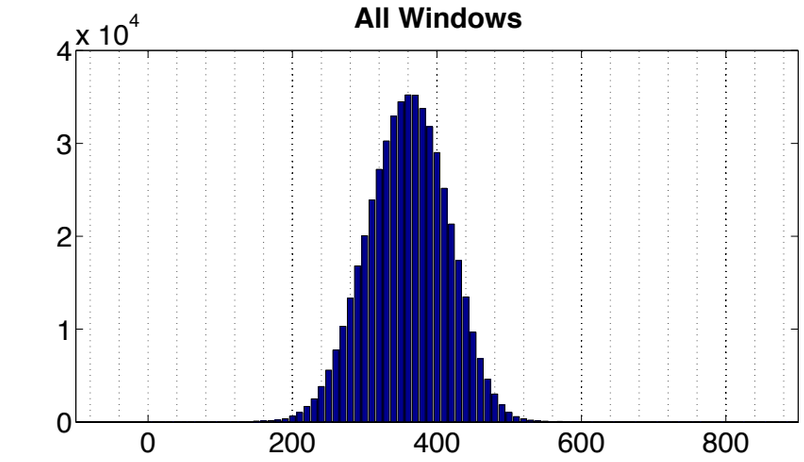
GC normalization of Illumina Data

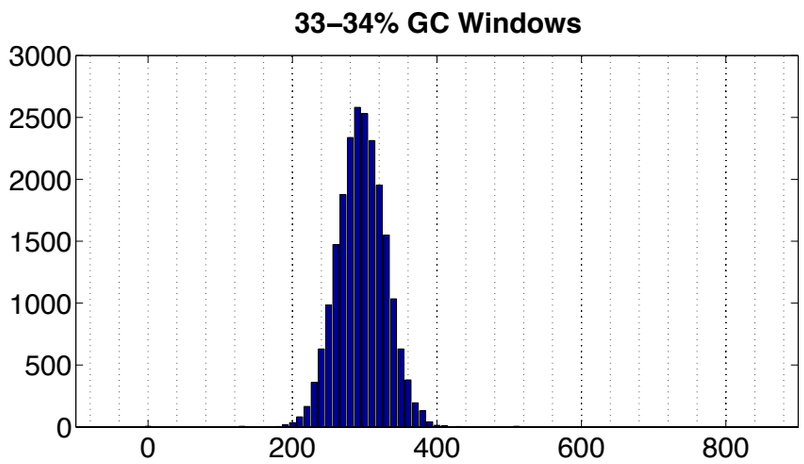
Raw Data

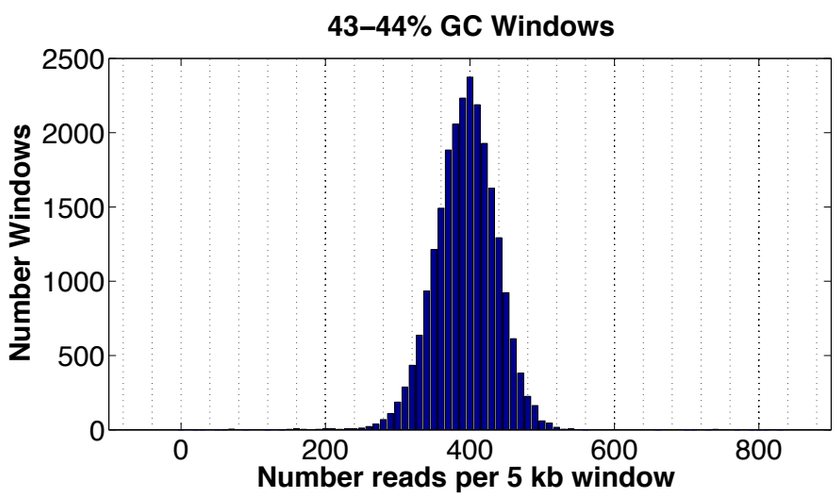


Normalized Data

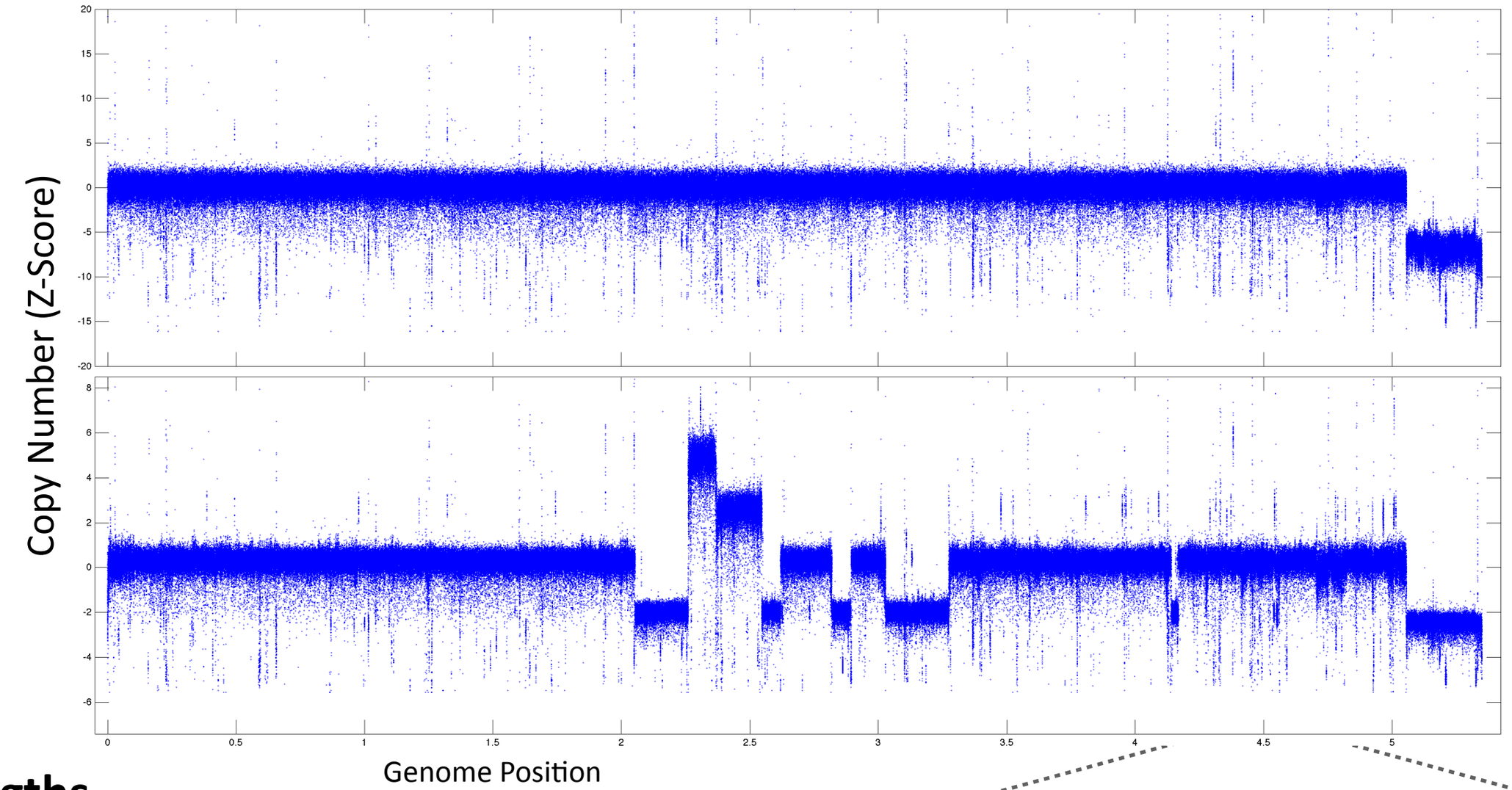






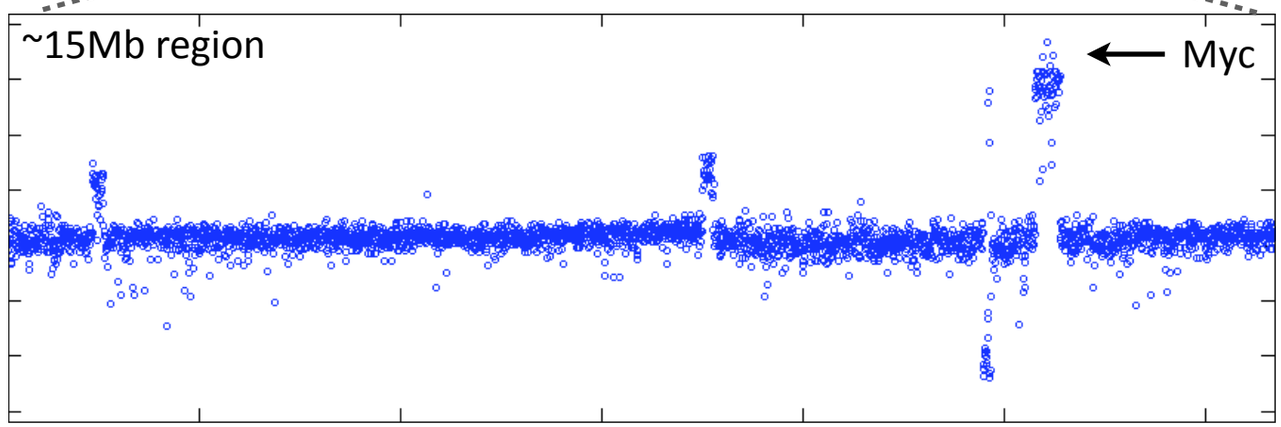


Detecting CNVs with read-depth analysis



Strengths

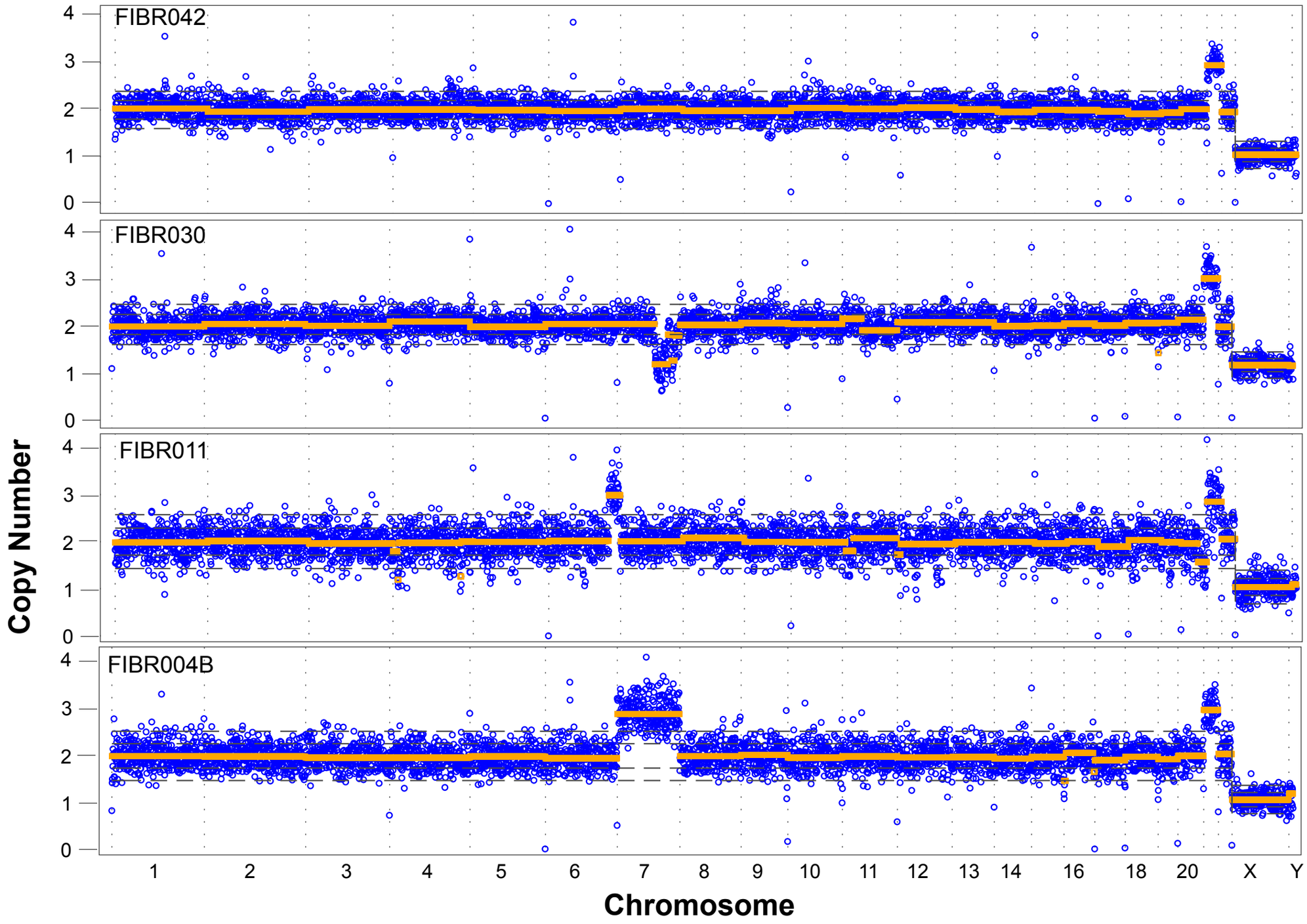
1) Fast and simple.
2) Directly measures copy number
3) Relatively straightforward interpretation: is gene X duplicated, deleted, or amplified?



Weaknesses:

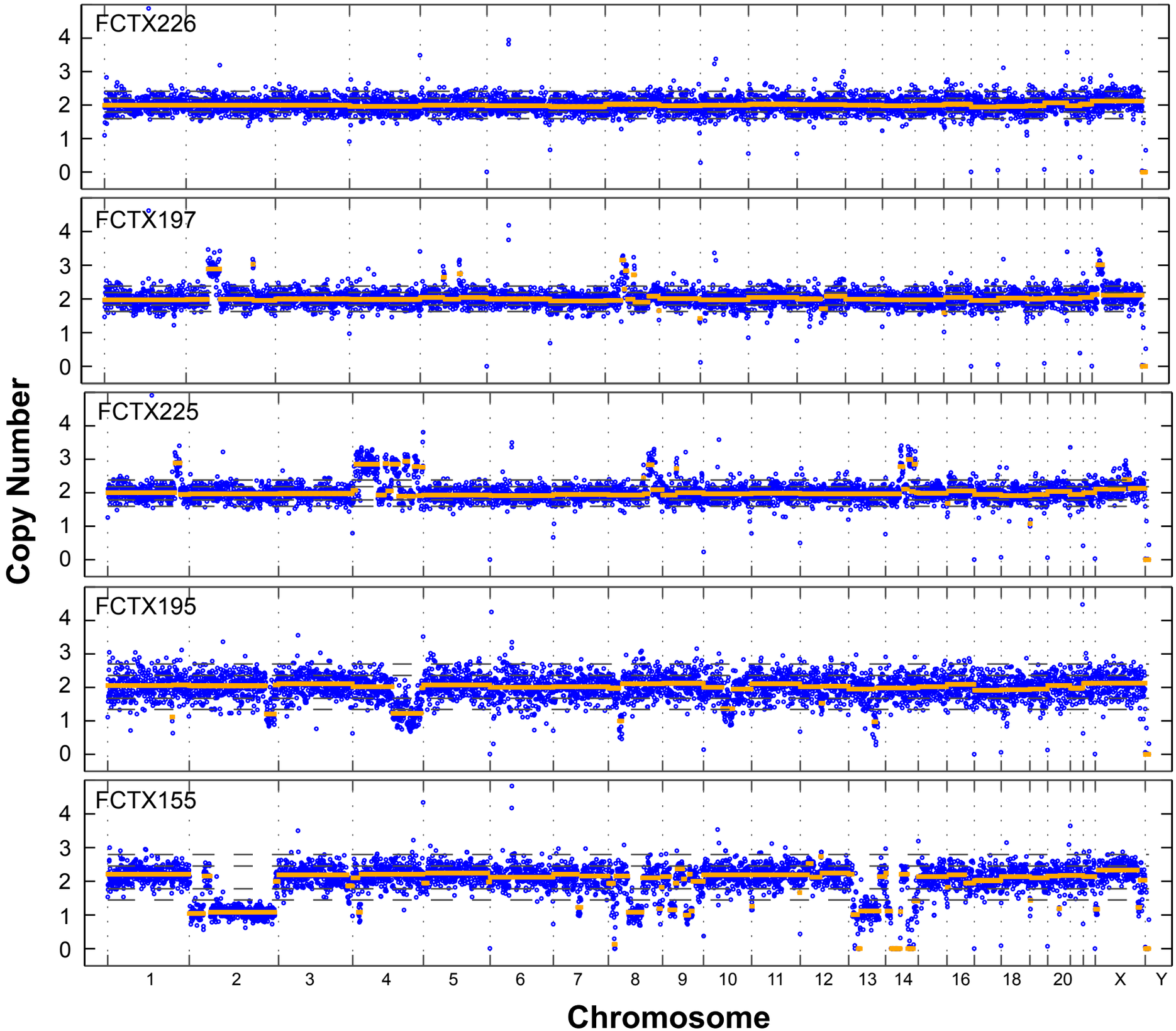
1) Limited resolution (1-10kb) = imprecise boundaries
2) Cannot detect balanced events or reveal variant architecture.

CNV detection in single cells (male trisomy 21 fibroblasts)



McConnell et al., *Science* (2013)

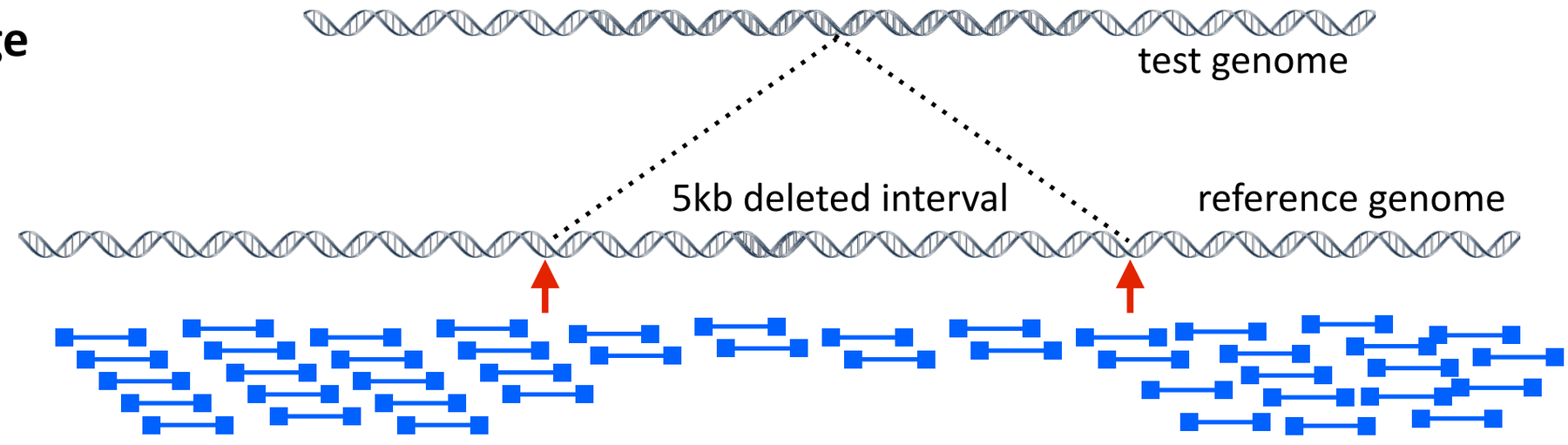
Copy number variation in single neurons



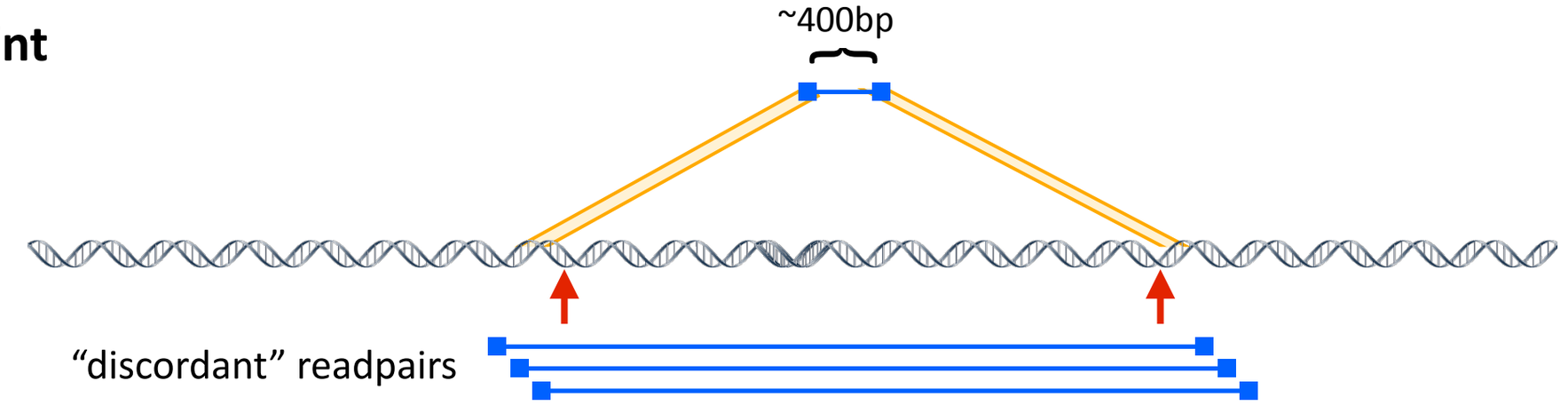
McConnell et al., *Science* (2013)

3 ways to detect a structural variant (SV)

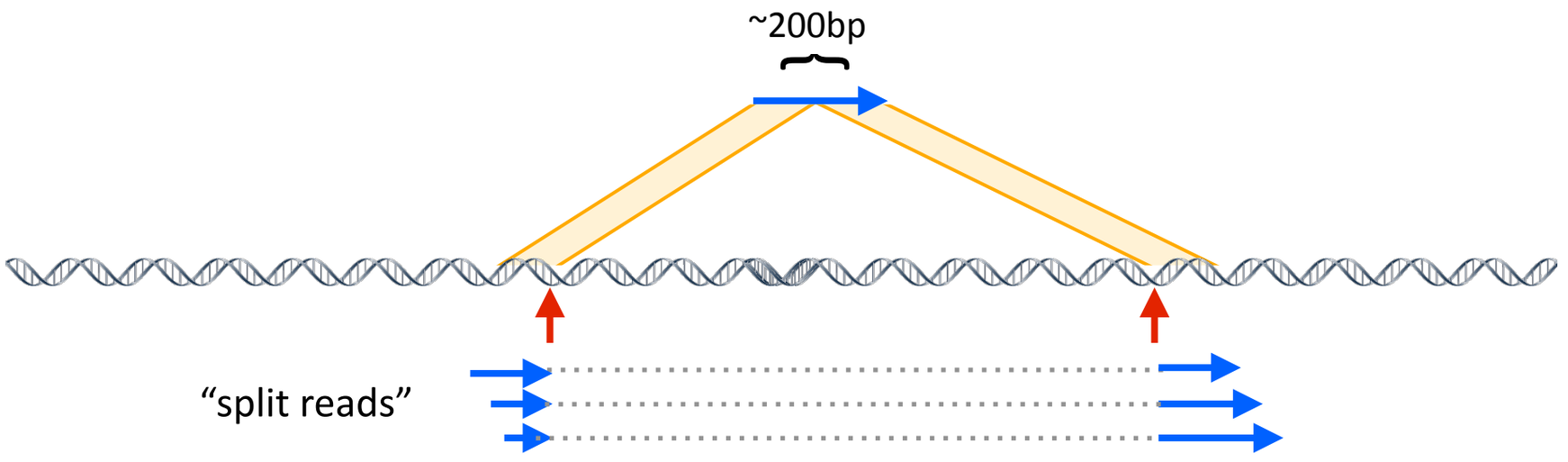
1) depth of sequence coverage
= "read-depth analysis"
(copy number alterations)



2) "readpairs" span a breakpoint
= "paired-end mapping"
(all classes of SV)

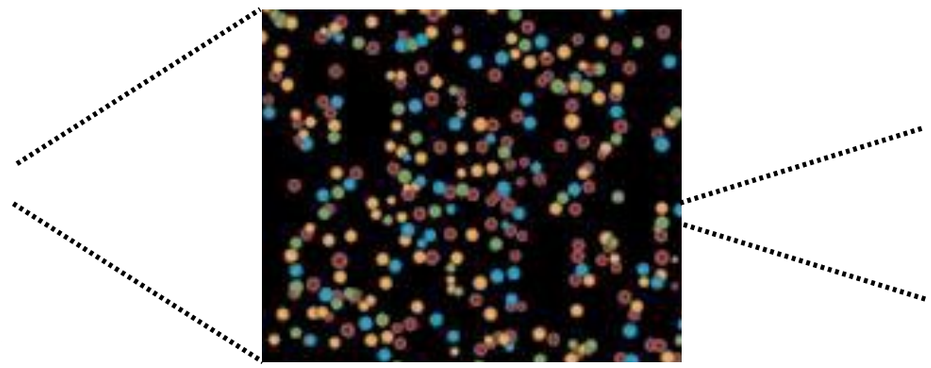


3) read contains a breakpoint
= "split-read mapping"
(all classes of SV)



Discordant paired-end alignments suggest breakpoints

Illumina HiSeq

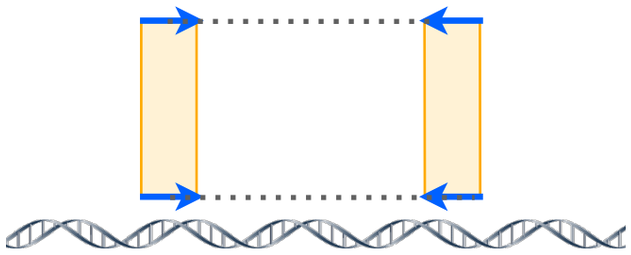


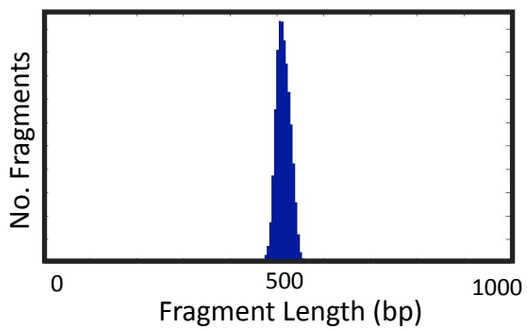
paired-end reads (readpairs)

5’ GGTGTACGAATAGTTTCCTTTTACACTCCTTGACCATCCTAGC ----------//---------- ----------//---------- GGACTGAAACTTCATCTGTCTTTATAGATATGCGTGCAGCAGC 5’

align reads

concordant readpairs
(>99%)





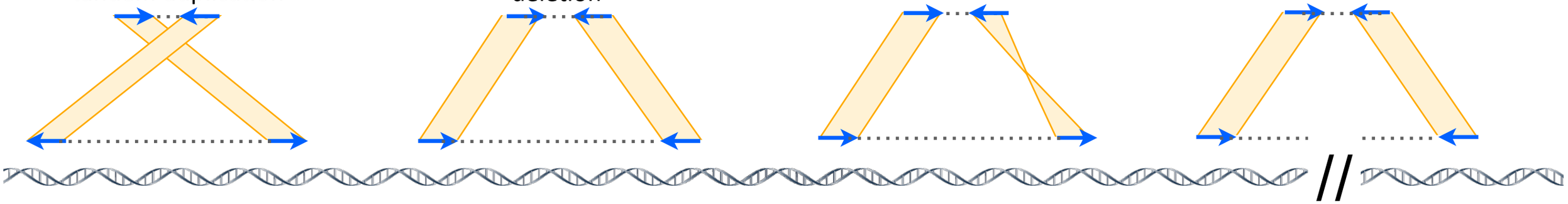
discordant read-pairs (<1%)

everted orientation
= tandem duplication

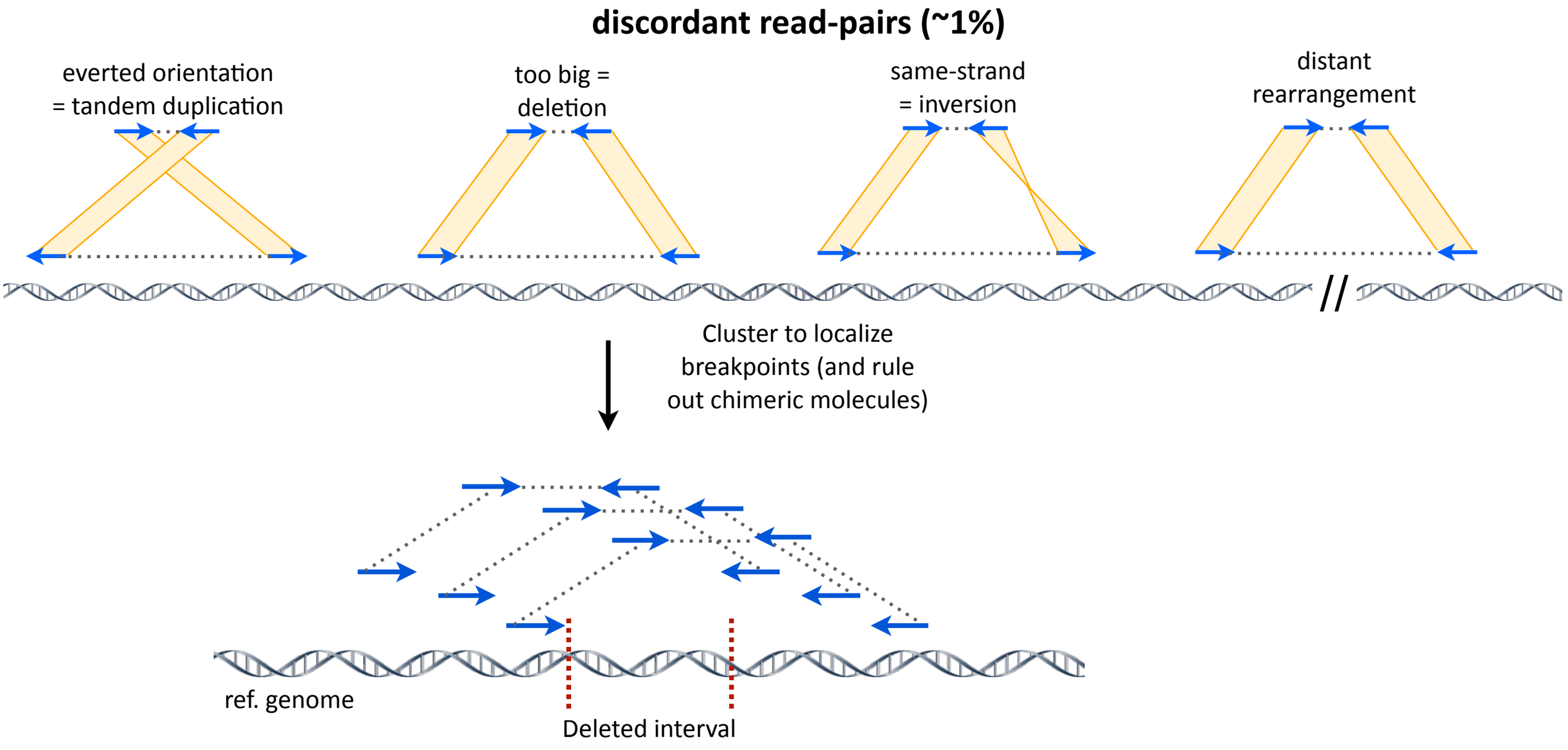
too big =
deletion

same-strand
= inversion

distant
rearrangement



Paired-end mapping algorithms cluster discordant alignments that “agree” with each other (support the same breakpoint)



Paired-end mapping signatures

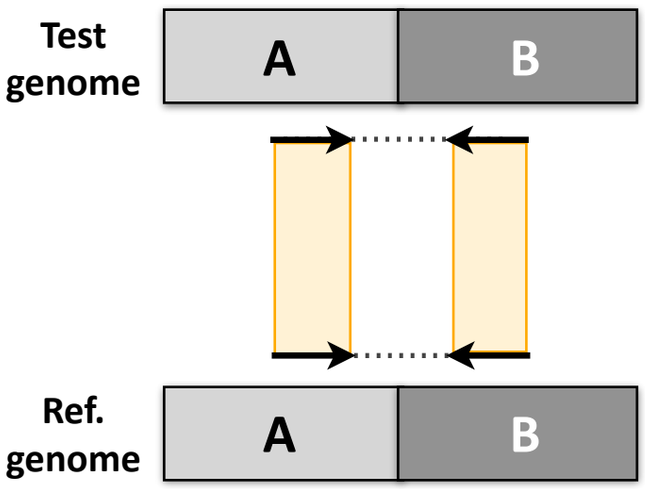
RULES:

1) The “test” genome is shown above the reference genome

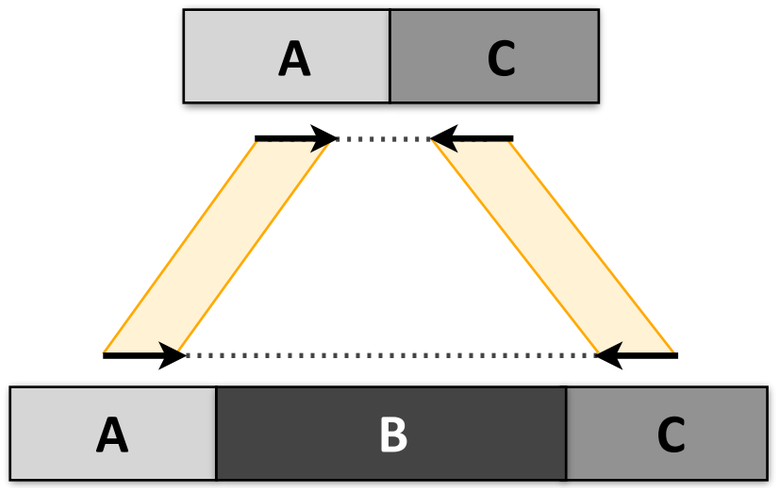
2) A rightward facing arrow denotes a “+” strand alignment, leftward a “-” strand alignment

3) Always give the orientation of the leftmost read mapping first when describing a mapping pattern in a file format.

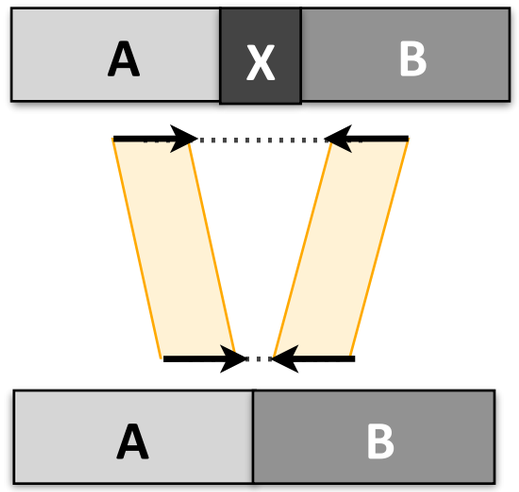
concordant (+/-)



too big (+/-) = deletion



too small (+/-) = spanned insertion



Paired-end mapping signatures

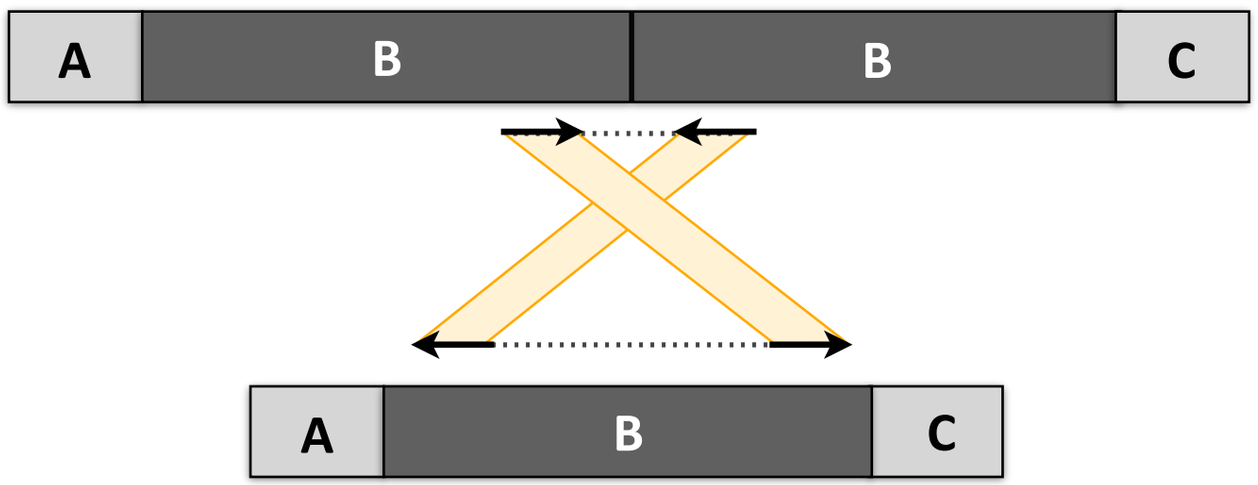
RULES:

1) The “test” genome is shown above the reference genome

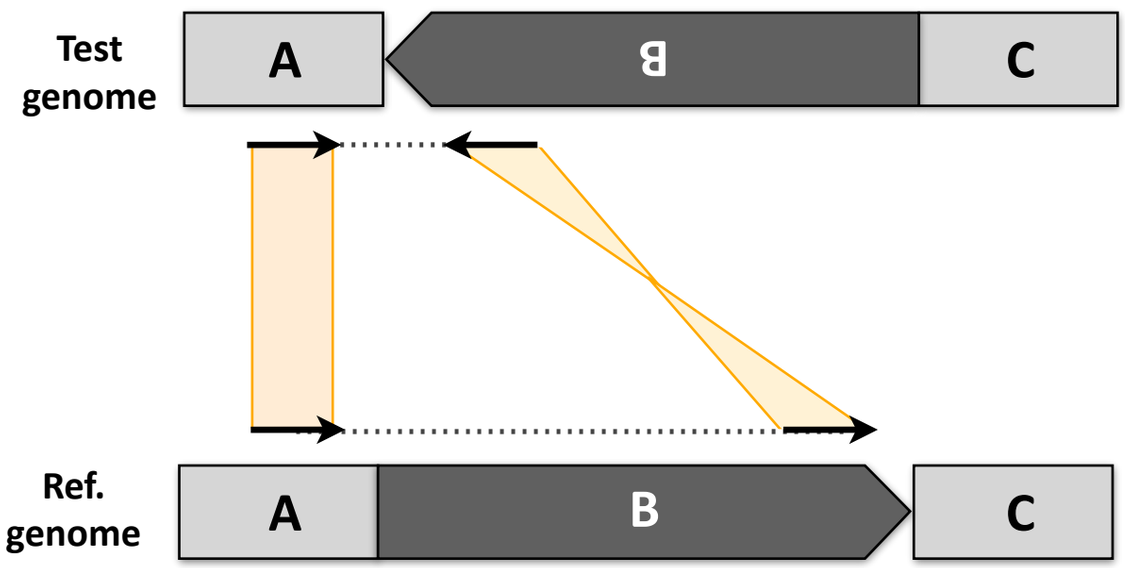
2) A rightward facing arrow denotes a “+” strand alignment, leftward a “-” strand alignment

3) Always give the orientation of the leftmost read mapping first when describing a mapping pattern in a file format.

everted (-/+)
= tandem duplication



same strand (+/+ or -/-)
= inversion



Paired-end mapping signatures

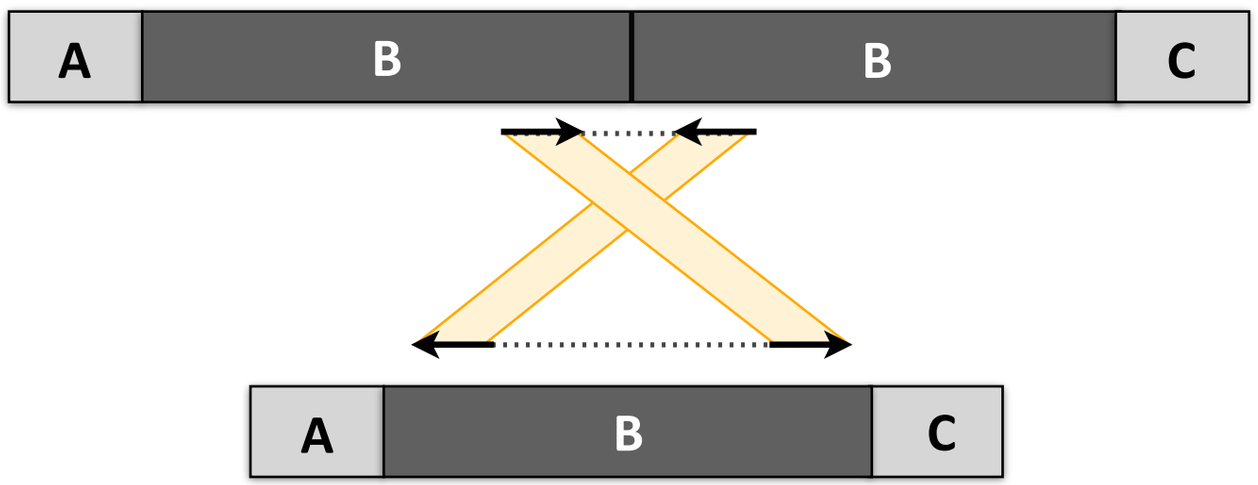
RULES:

1) The “test” genome is shown above the reference genome

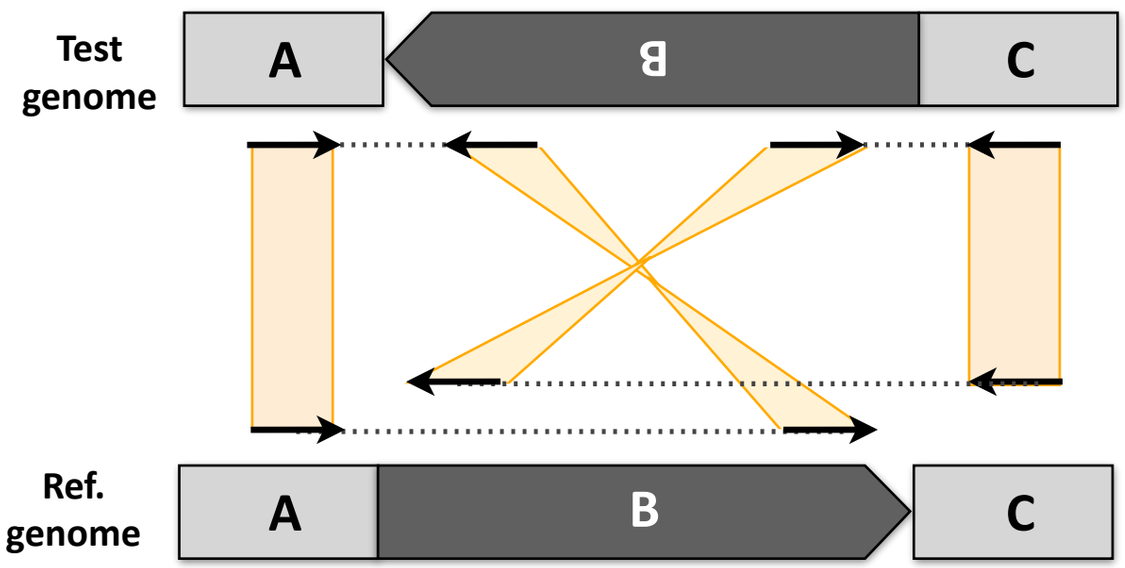
2) A rightward facing arrow denotes a “+” strand alignment, leftward a “-” strand alignment

3) Always give the orientation of the leftmost read mapping first when describing a mapping pattern in a file format.

everted (-/+)
= tandem duplication



same strand (+/+ or -/-)
= inversion



Paired-end mapping signatures

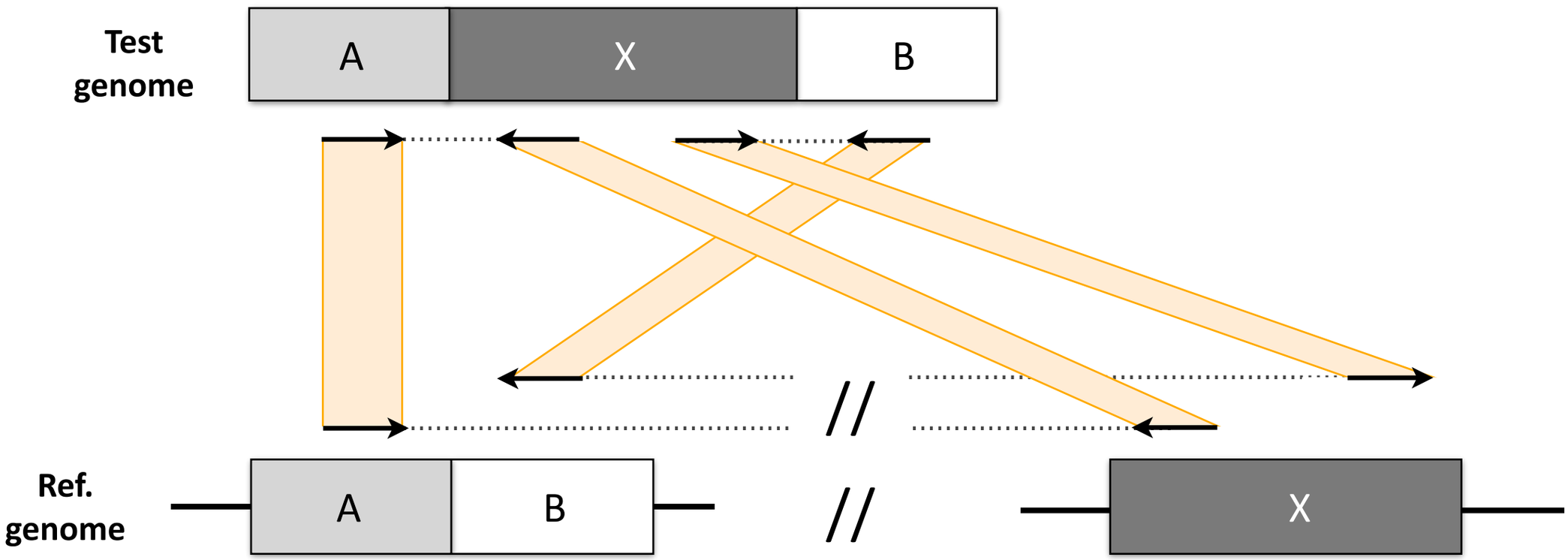
RULES:

1) The “test” genome is shown above the reference genome

2) A rightward facing arrow denotes a “+” strand alignment, leftward a “-” strand alignment

3) Always give the orientation of the leftmost read mapping first when describing a mapping pattern in a file format.

Large insertion from a distant genomic location



Paired-end mapping signatures

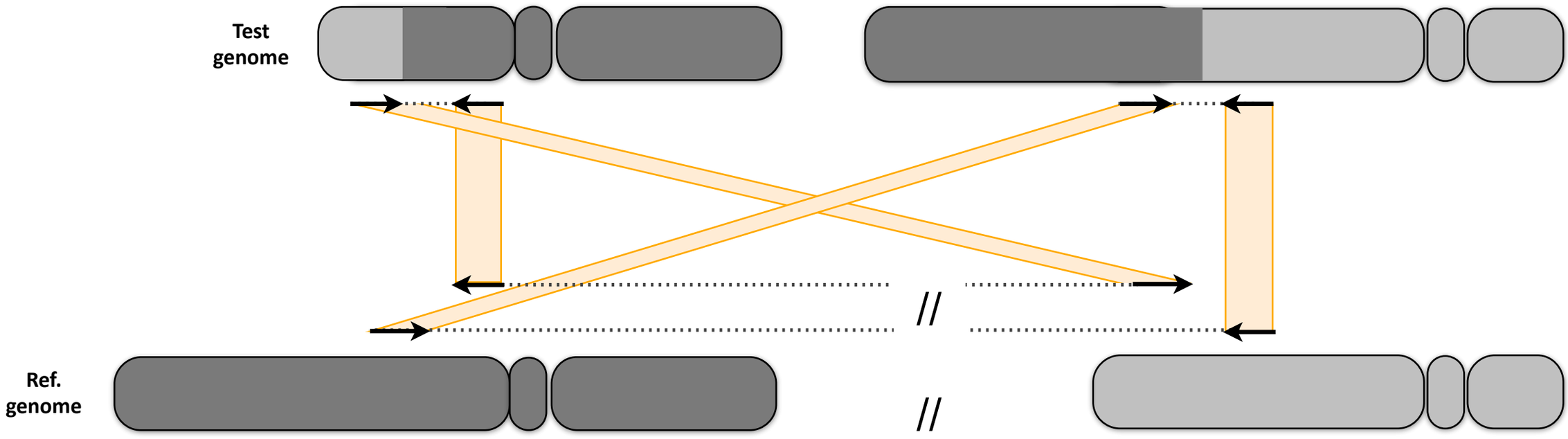
RULES:

1) The “test” genome is shown above the reference genome

2) A rightward facing arrow denotes a “+” strand alignment, leftward a “-” strand alignment

3) Always give the orientation of the leftmost read mapping first when describing a mapping pattern in a file format.

Reciprocal Translocation

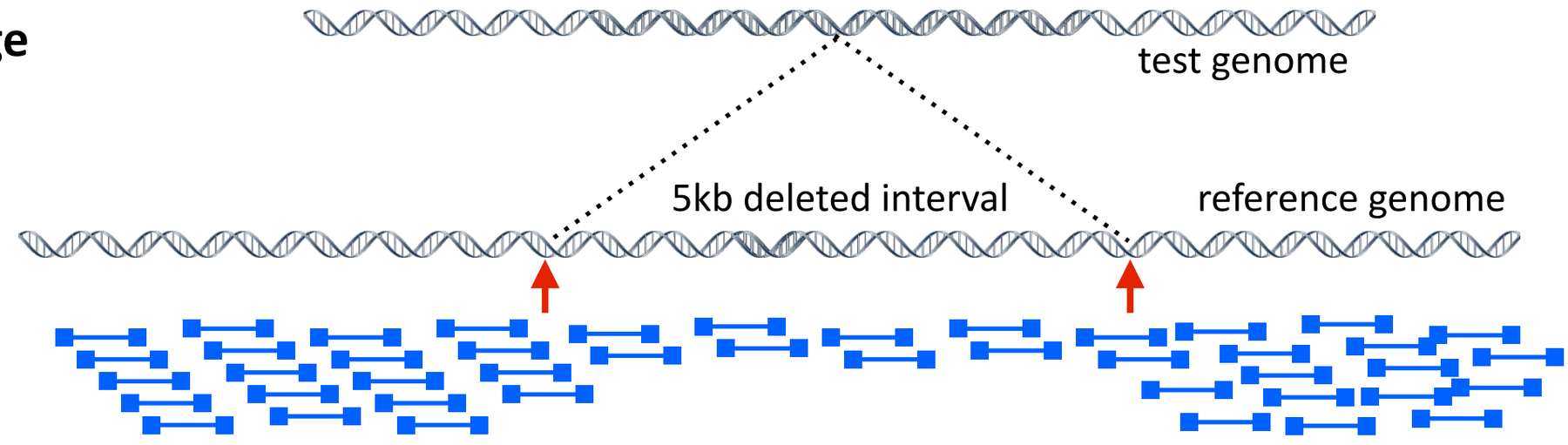


Paired-end mapping mania!

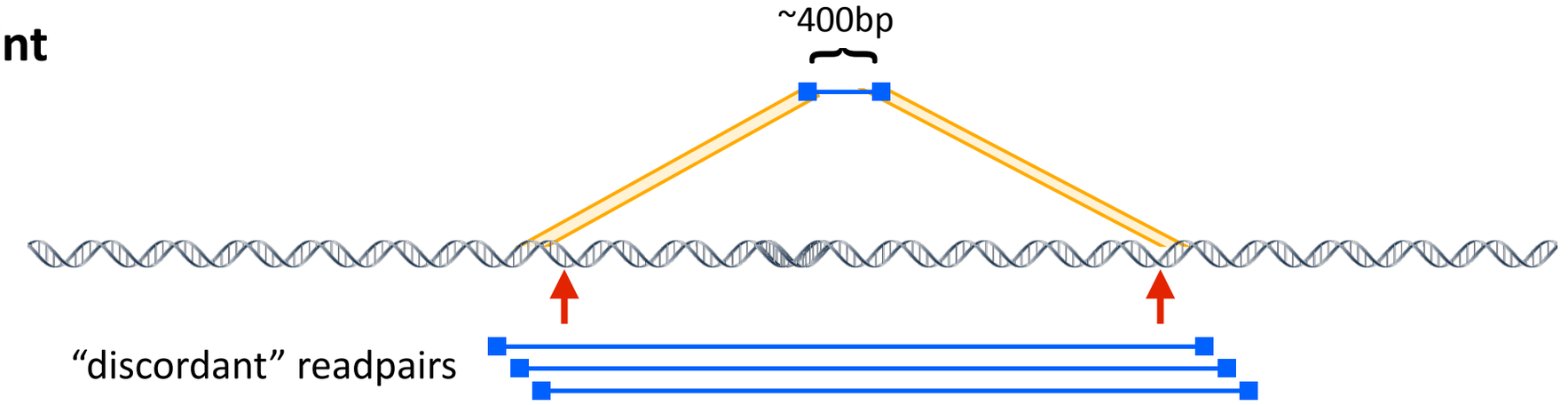
- **There are a ton of paired-end mapping algorithms and new ones keep coming out all the time.**
- **There are only five that I can personally recommend:**
  - LUMPY (Hall and Quinlan labs)
  - DELLY (Korbel lab)
  - GenomeStrip (McCarroll lab)
  - GASV-PRO (Raphael lab)
  - HYDRA (Hall & Quinlan labs)
- **Others may be excellent but we haven't tested them.**
- **For high quality analysis, parameter selection and filtering are more important than the algorithm.**

3 ways to detect a structural variant (SV)

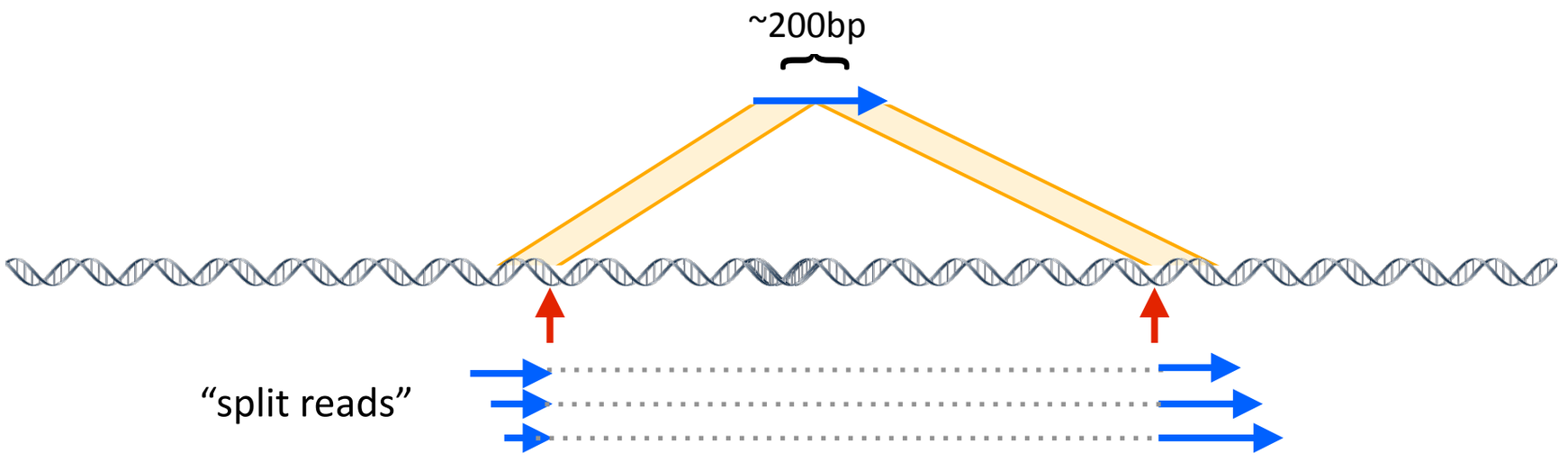
1) depth of sequence coverage
= "read-depth analysis"
(copy number alterations)



2) "readpairs" span a breakpoint
= "paired-end mapping"
(all classes of SV)



3) read contains a breakpoint
= "split-read mapping"
(all classes of SV)



Split-read mapping localizes breakpoints to base-pair resolution (more or less)

Sanger Sequencing

~500 bp read

Roche/454

~500 bp read

Pacific Biosciences

~5 kb read

Illumina HiSeq

paired ~100 bp reads

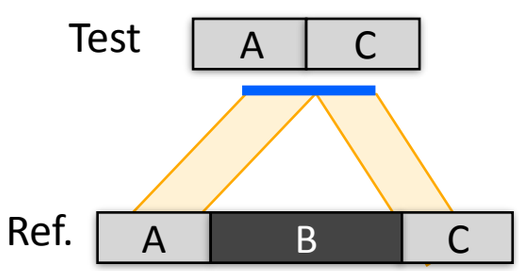
Traditional split-read mapping

- align to genome
- Cluster reads (+/- a few bp)

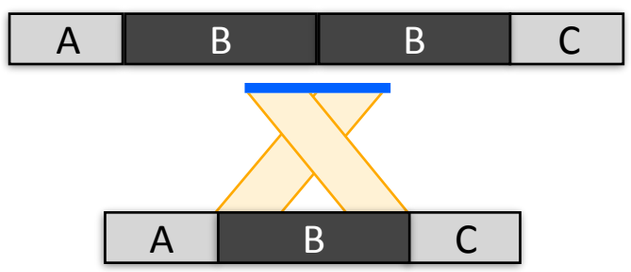
PINDEL (Ye et al., 2009)

- align reads to genome
- sensitive split-read alignment at loci with hints of SV

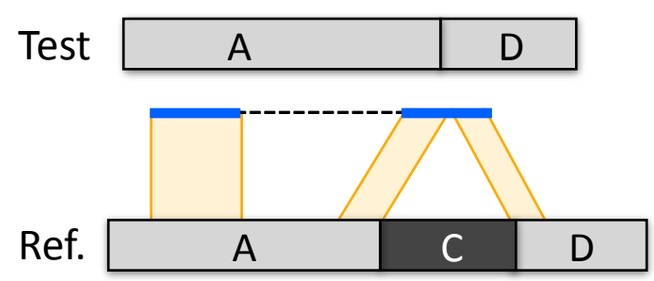
Deletion



Duplication



Deletion



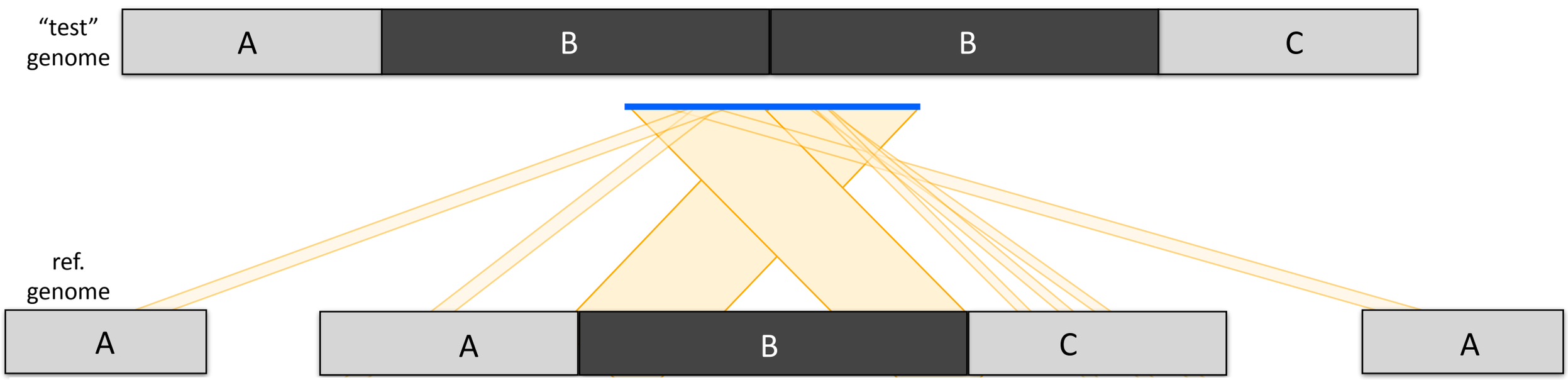
NOTE: BWA-MEM now does joint paired-end and split-read alignment

Split-read mapping challenges

1) It is difficult to accurately align short breakpoint-containing reads to the reference genome, especially when repeats are involved.

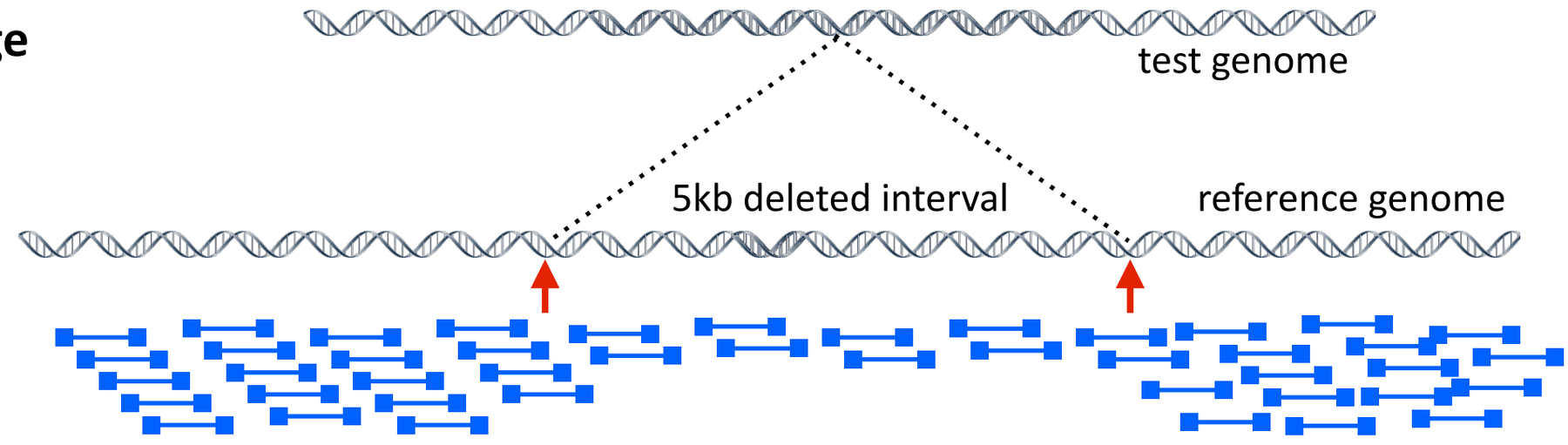
2) With long reads, one must select which "sub-alignments" to retain. Most aligners don't do this properly, or at all. The three best aligners that automatically select sub-alignments for you are BWA-SW (Li et al., 2011), YAHA (Faust et al., 2012) and BWA-MEM (Li, 2013).

3) Few published tools do "pure" split-read mapping, in the absence of other evidence types such as paired-end alignments. LUMPY does.

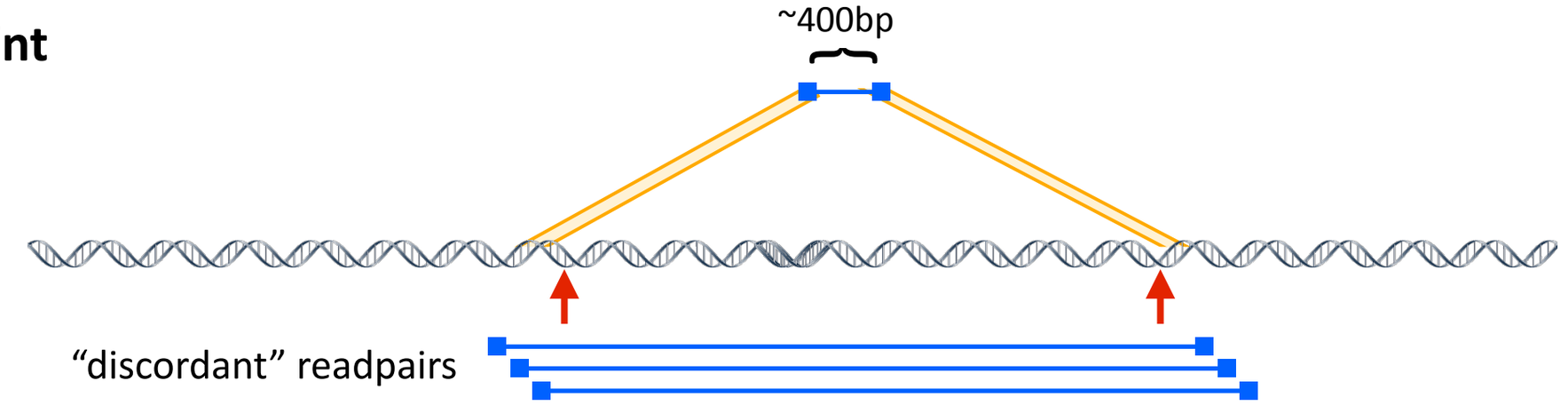


3 ways to detect a structural variant (SV)

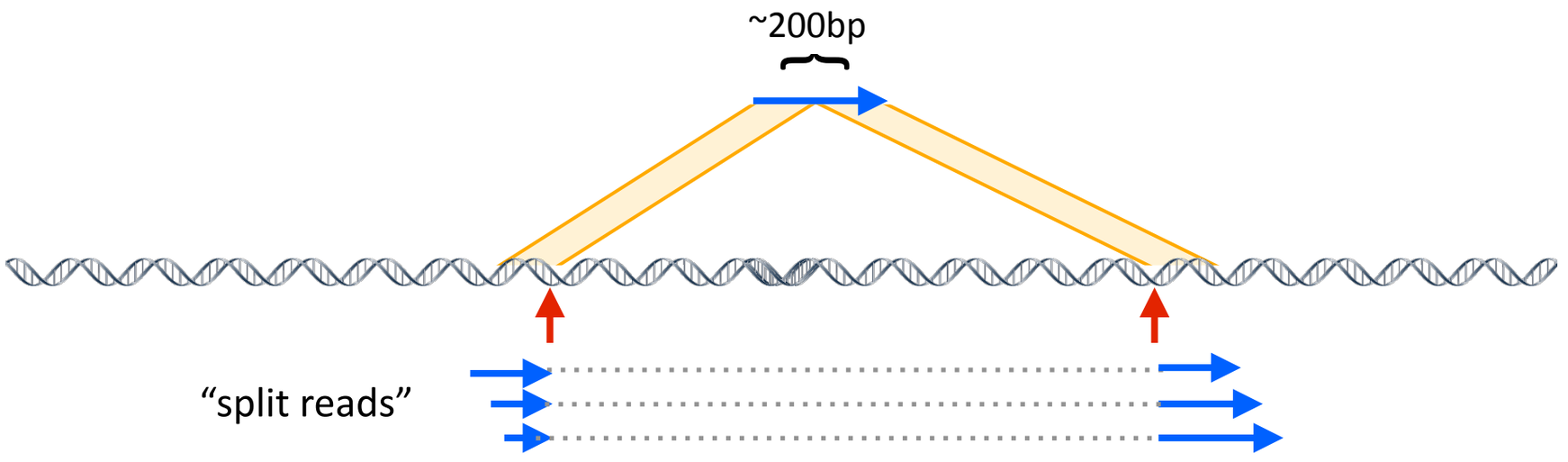
1) depth of sequence coverage
= "read-depth analysis"
(copy number alterations)



2) "readpairs" span a breakpoint
= "paired-end mapping"
(all classes of SV)



3) read contains a breakpoint
= "split-read mapping"
(all classes of SV)



Some algorithms that use multiple signals

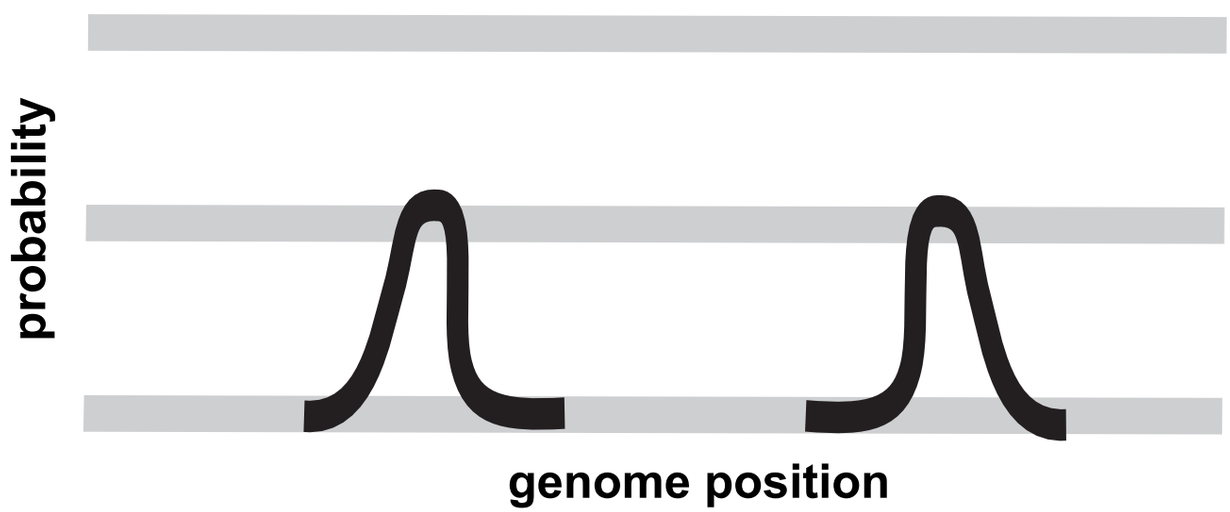
- **GenomeStrip (McCarroll Lab)**
  - Originally, used read-depth to genotype deletions discovered by paired-end mapping. Much development recently…….
- **GASV-PRO (Raphael Lab)**
  - Identifies SVs with read-pair information; uses read-depth to refine calls and add confidence.
- **DELLY (Korbel Lab)**
  - Identifies breakpoints with paired-end mapping, refines coordinate positions and adds confidence with split-read mapping.
- **PINDEL (Kai Ye)**
  - Uses discordant paired-end alignments to decide at which loci to perform sensitive split-read alignment.

Note: To my knowledge, all of these tools use two signals *sequentially*

LUMPY: a breakpoint detection framework that naturally integrates multiple signals

The core idea: For each *pair* of coordinates in the reference genome, calculate the probability that they are a novel adjacency based on all available measurements.

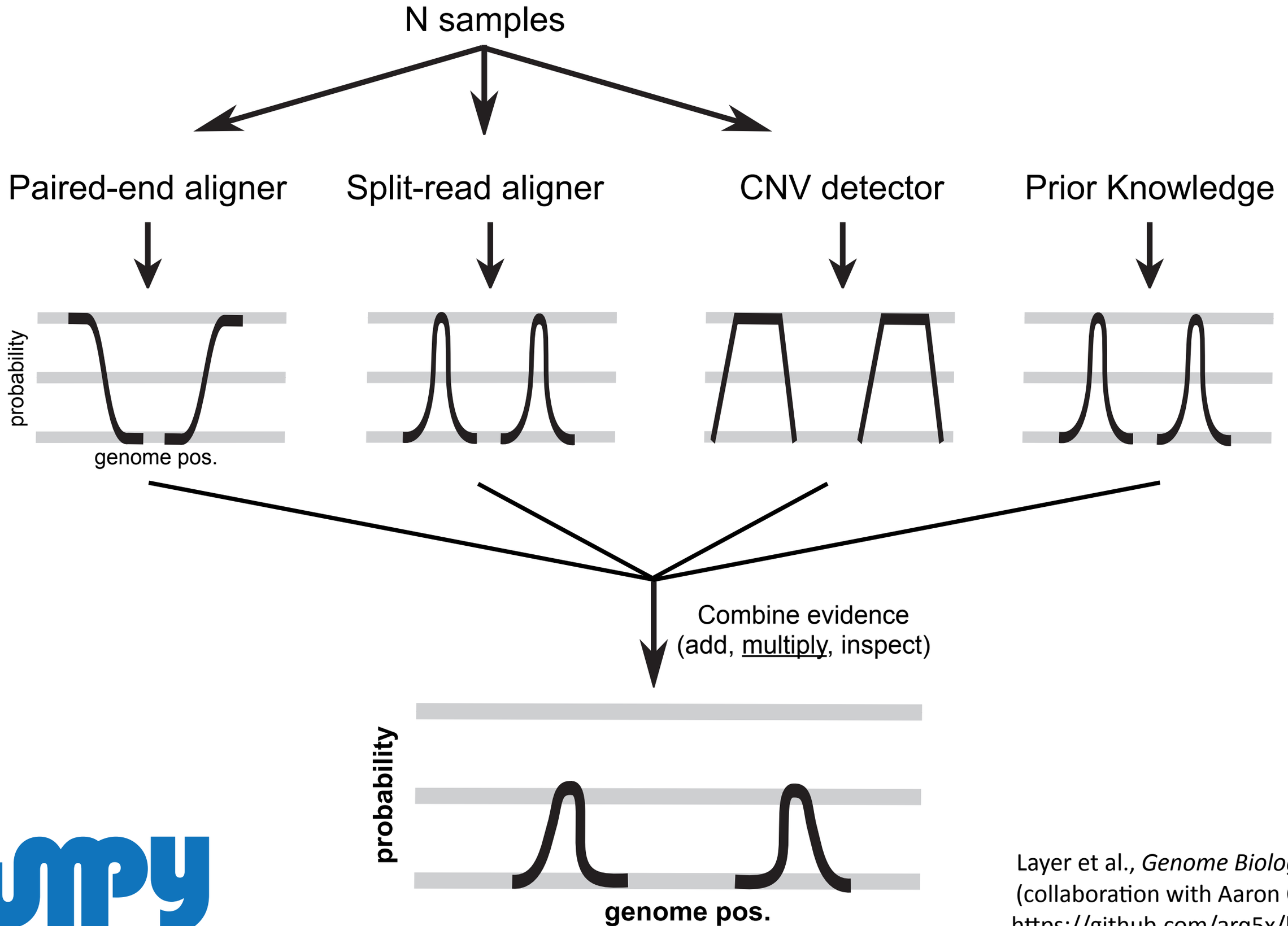
One could store this information in a 3 billion X 3 billion matrix. We represent it more succinctly as paired probability distributions for informative genomic intervals



lumpy

Layer et al., *Genome Biology*, 2014

LUMPY clusters breakpoint probability distributions

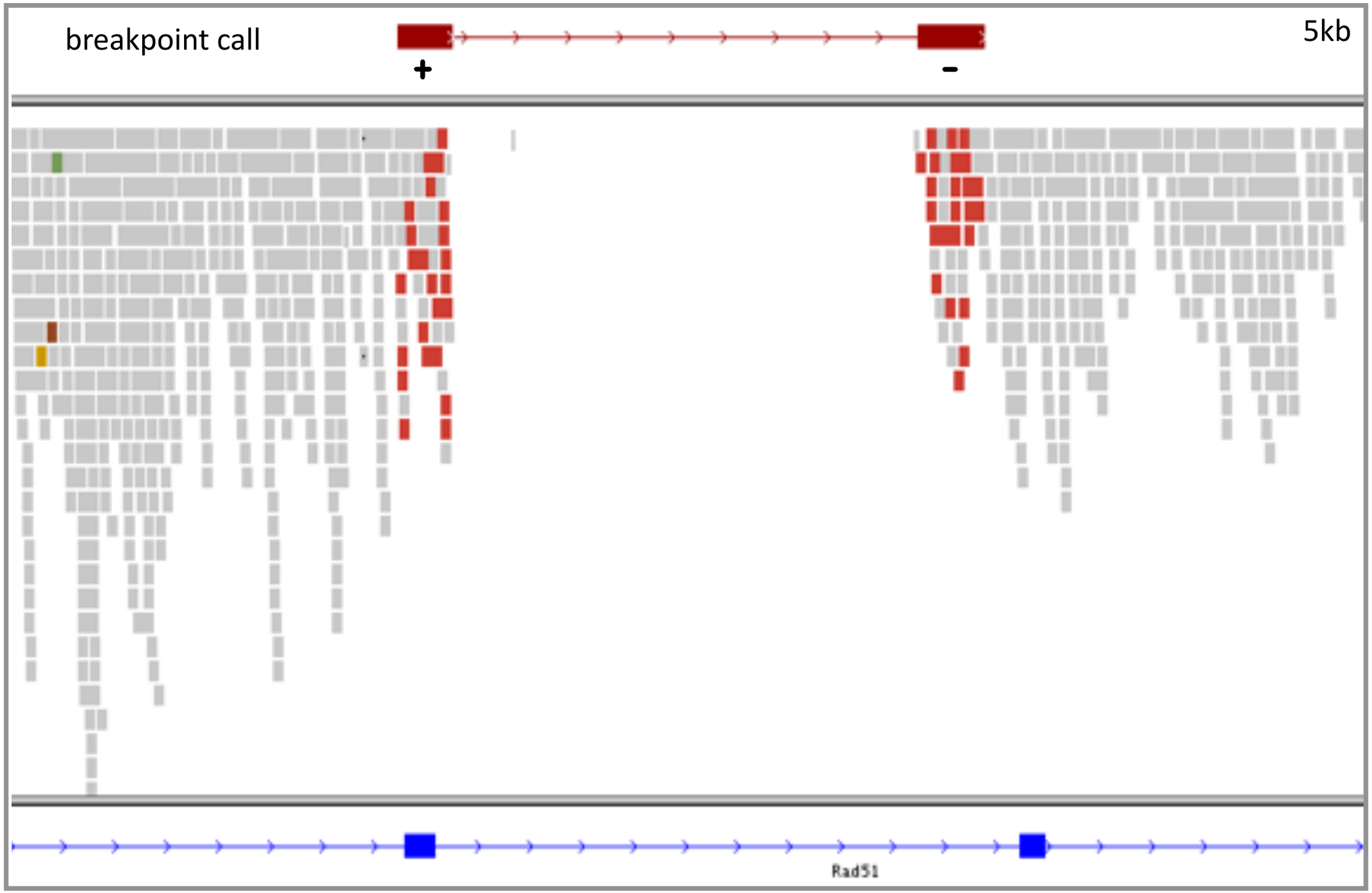


lumpy

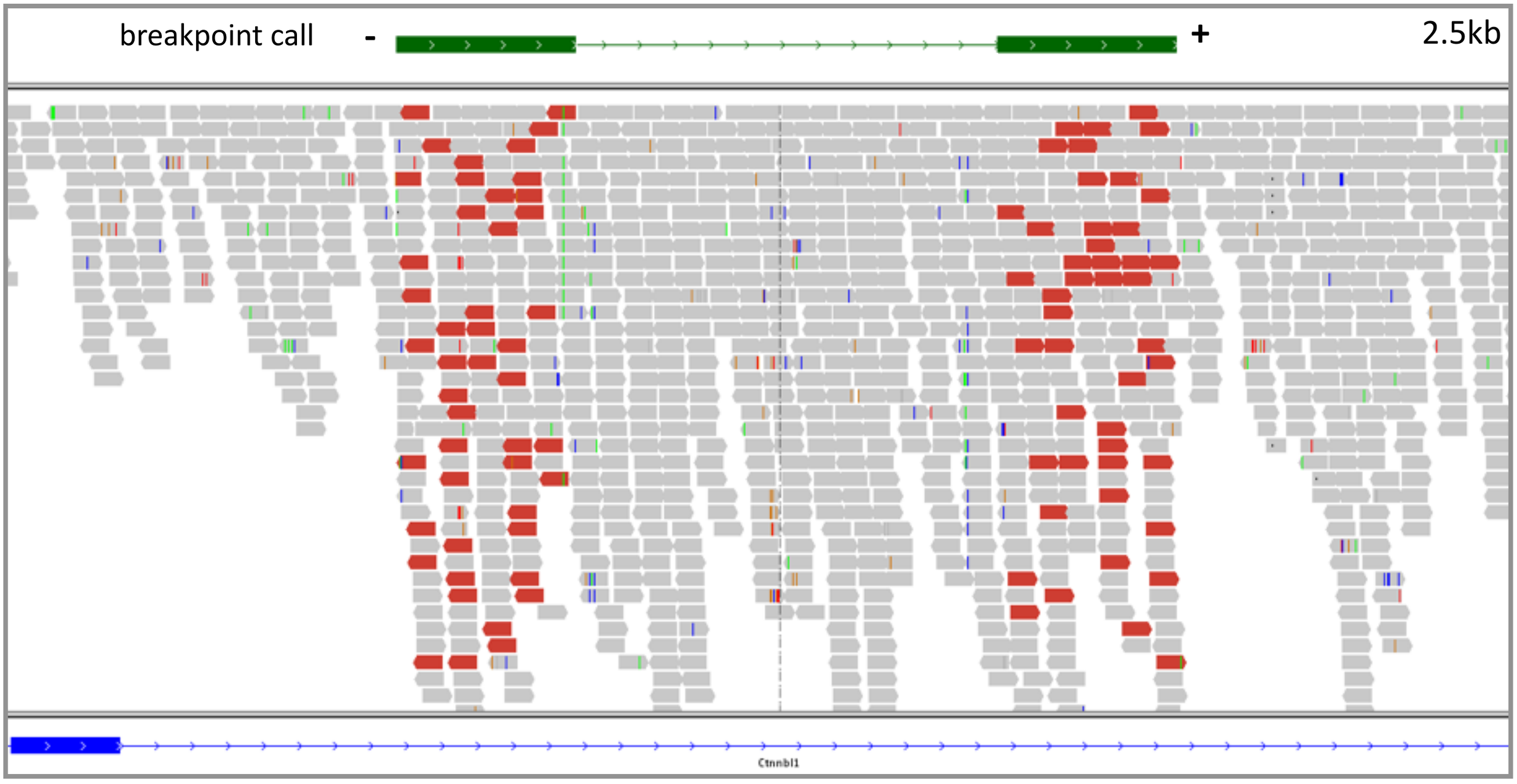
Layer et al., *Genome Biology*, 2014
(collaboration with Aaron Quinlan)
https://github.com/arq5x/lumpy-sv

The structural variation landscape between two “normal” humans

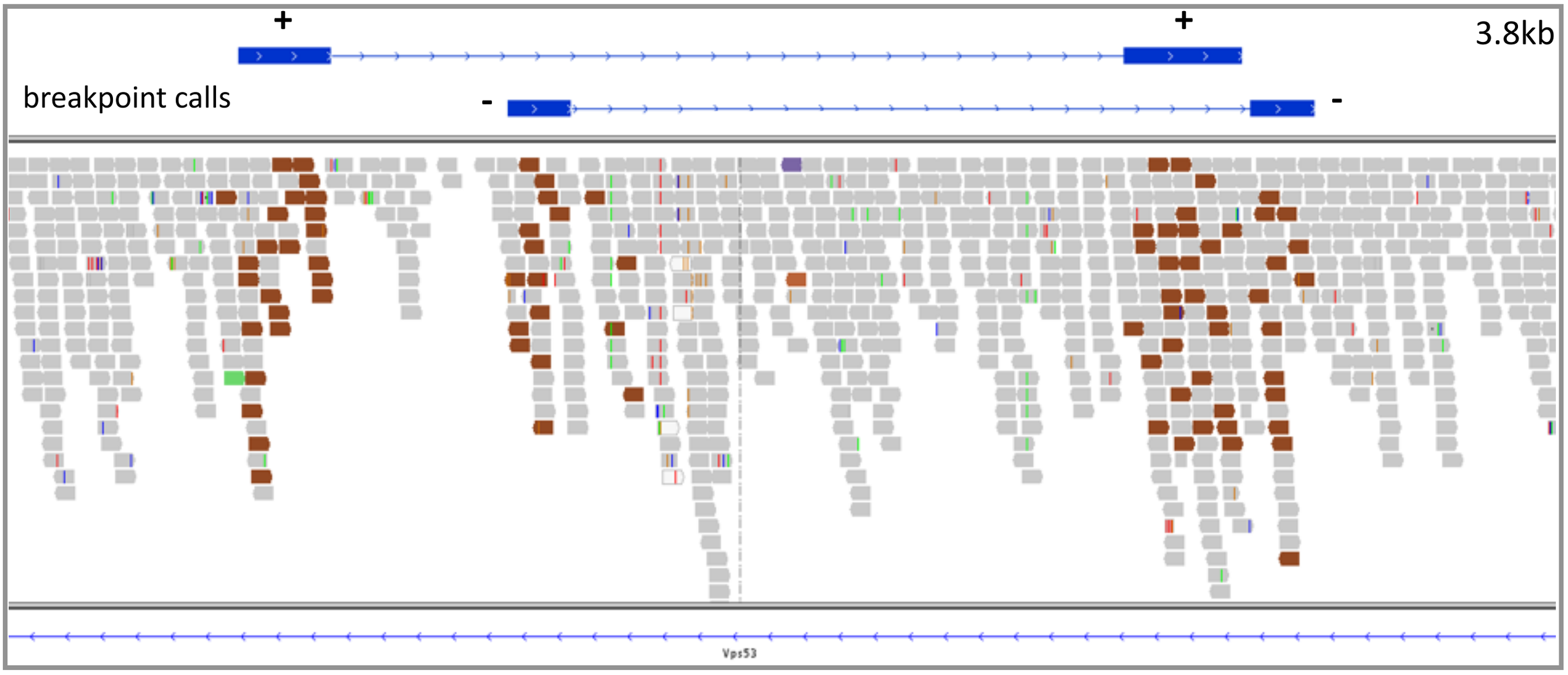
Several thousand deletions



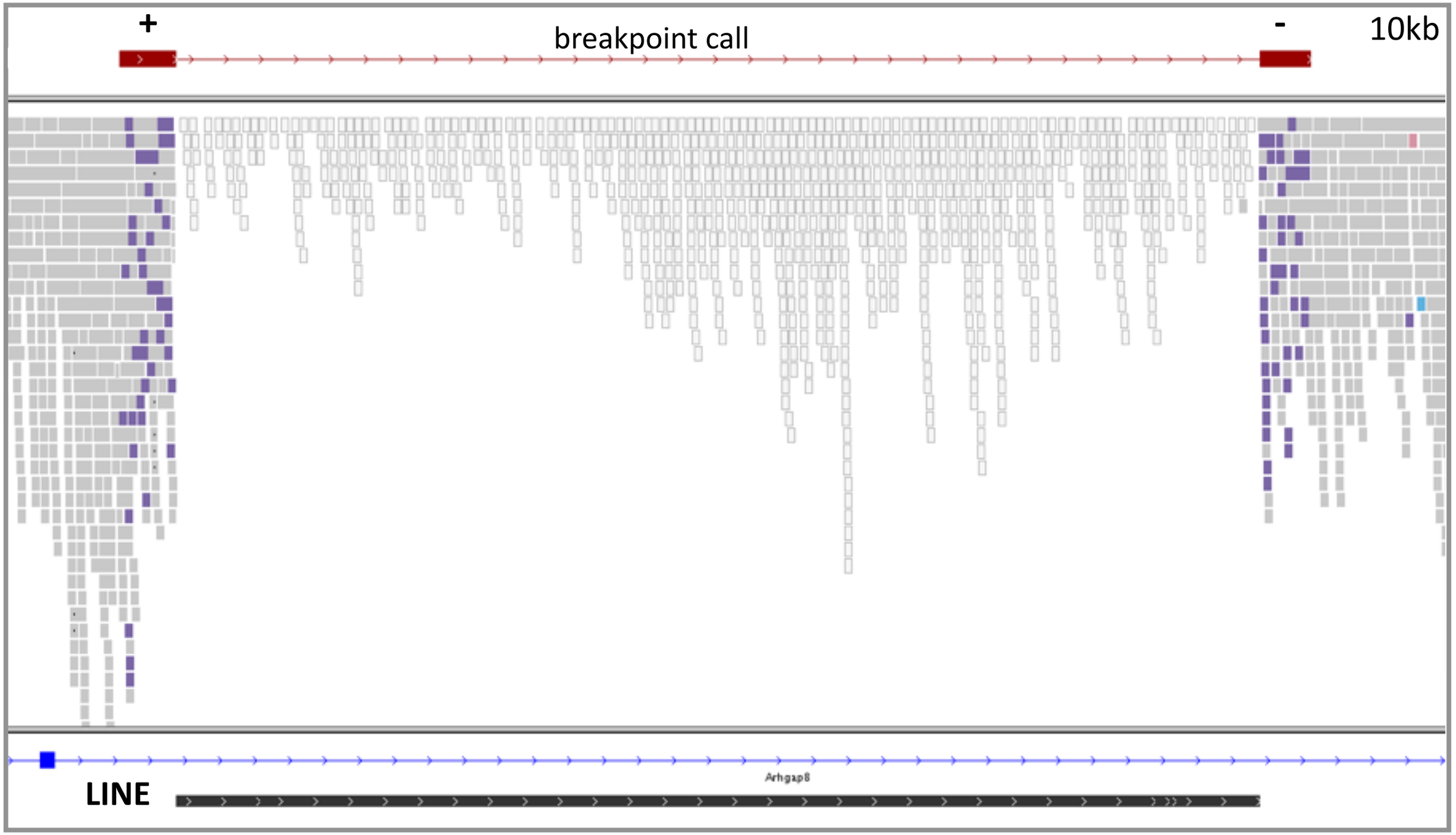
Several hundred duplications



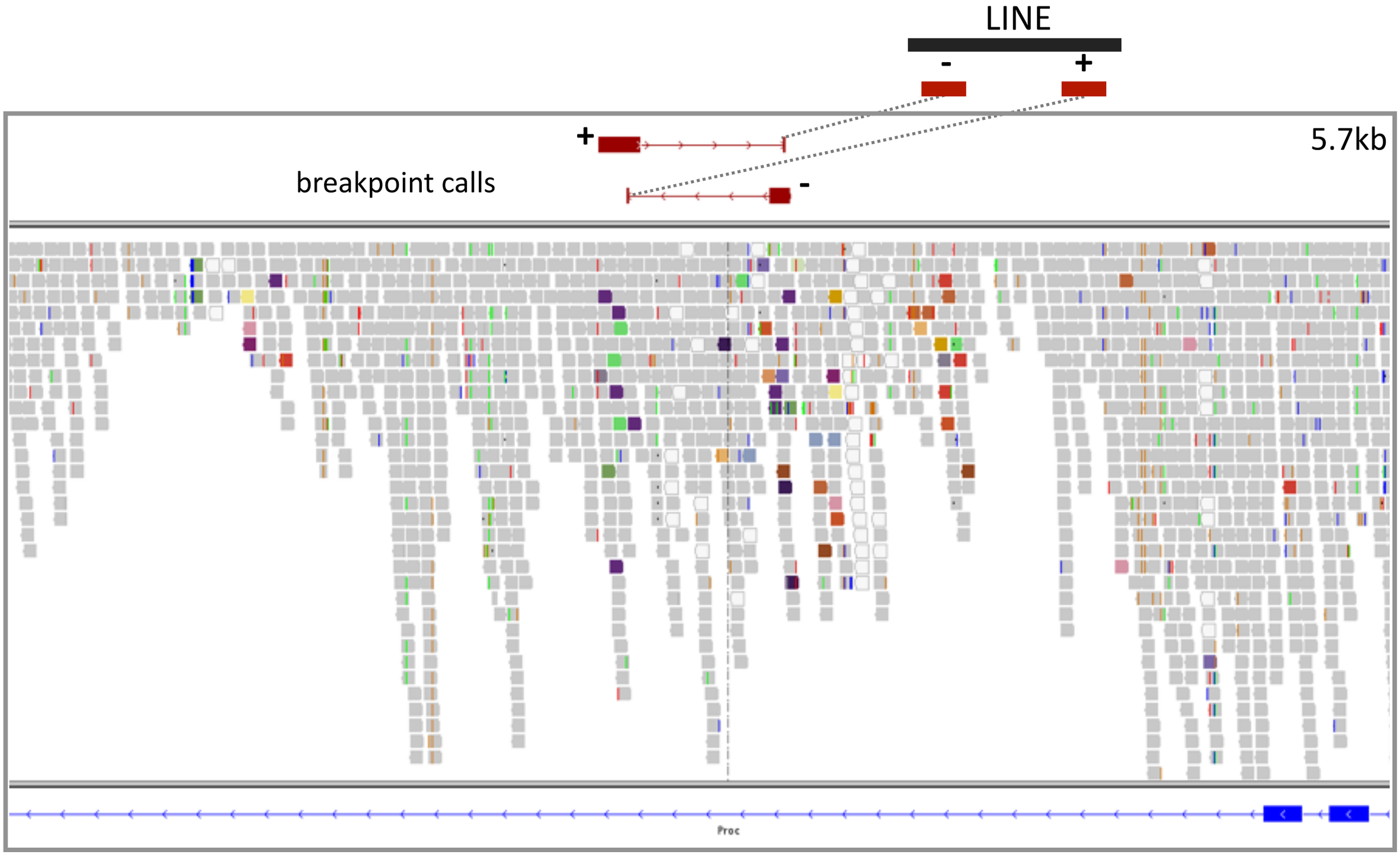
Several hundred inversions



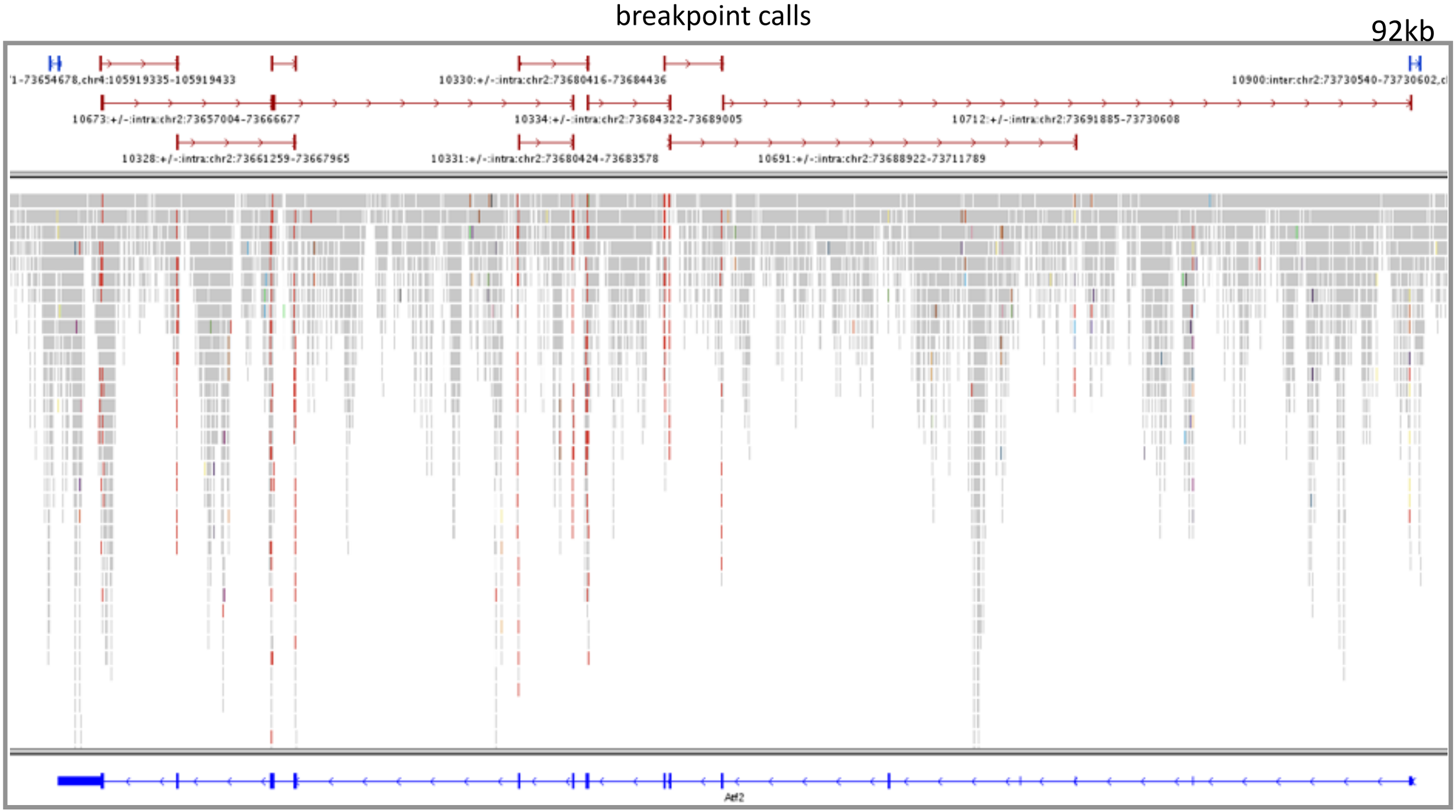
Several hundred transposon insertions in the reference genome



Several hundred transposon insertions in the test genome

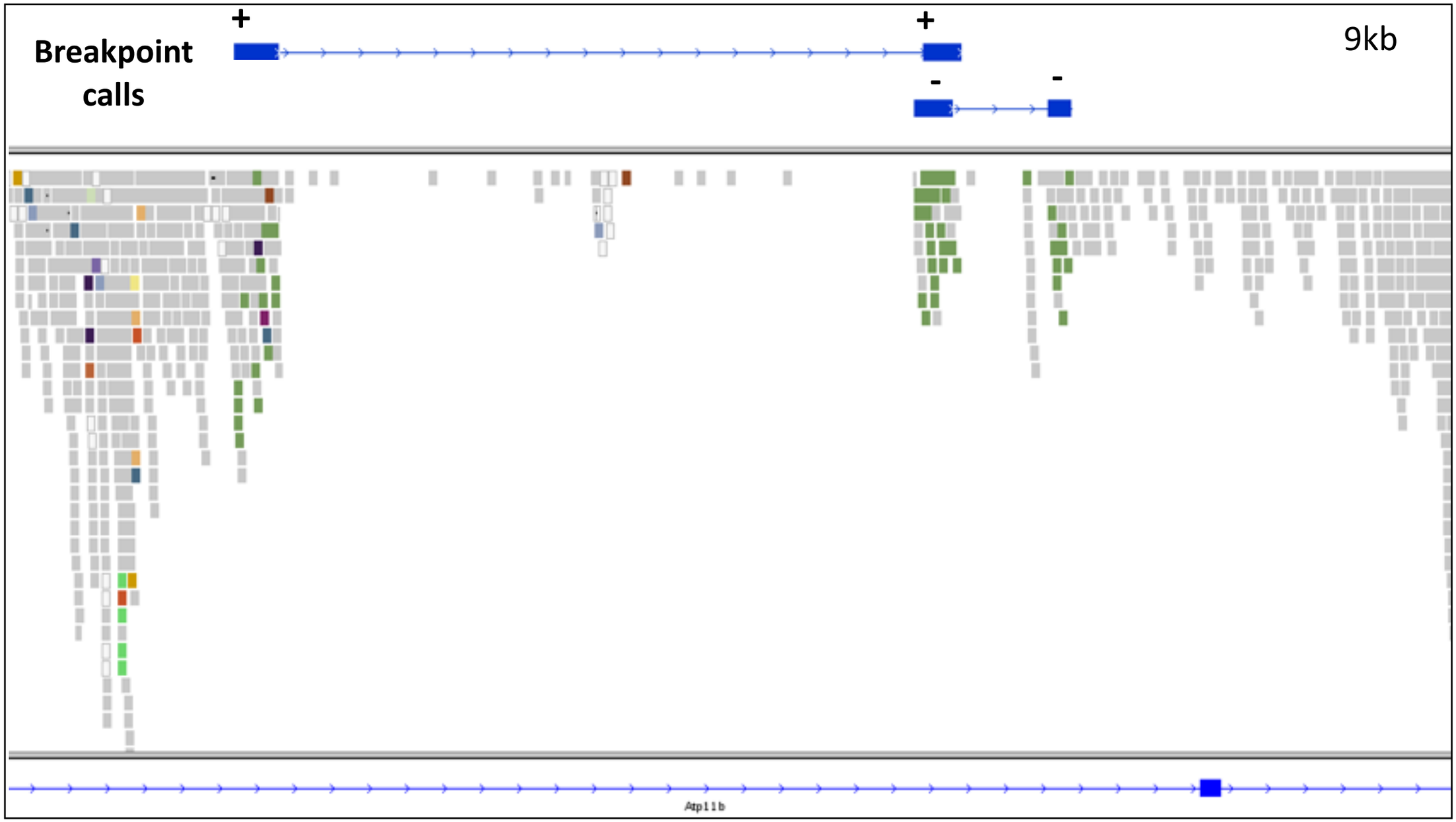


Tens of retroposed genes



Also: There also exist tens to hundreds of additional insertions, which are difficult to classify. These can be caused by un-annotated transposons, retroposed non-coding RNAs, and DNA insertion caused by 3 DNA breakages.

Tens to hundreds of complex variants

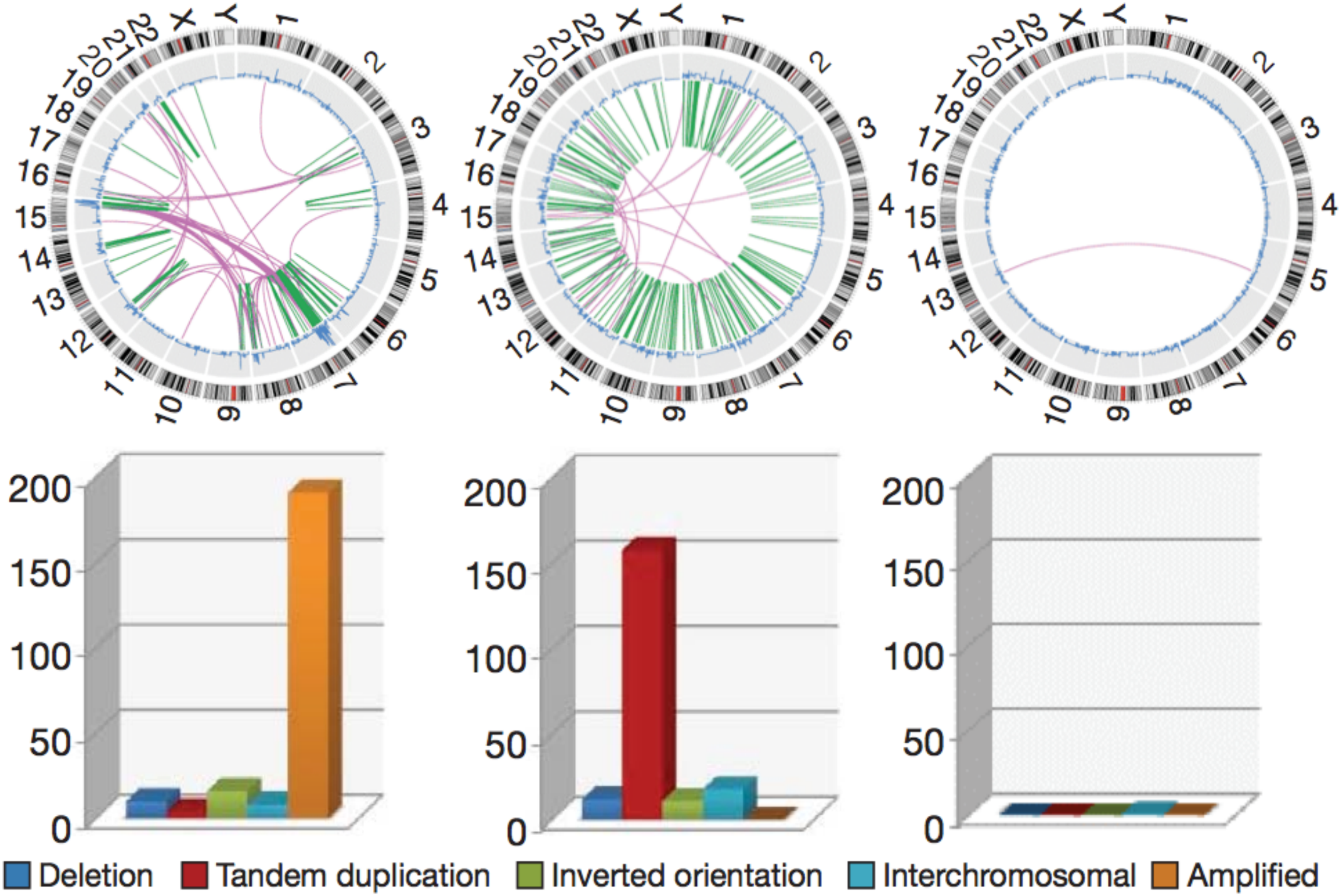


Important: we cannot infer variant class based solely upon relative read orientation; e.g., an apparent deletion may really be part of a complex rearrangement

Quinlan et al., 2010 (Genome Research)

Tumor genome landscapes are extremely diverse

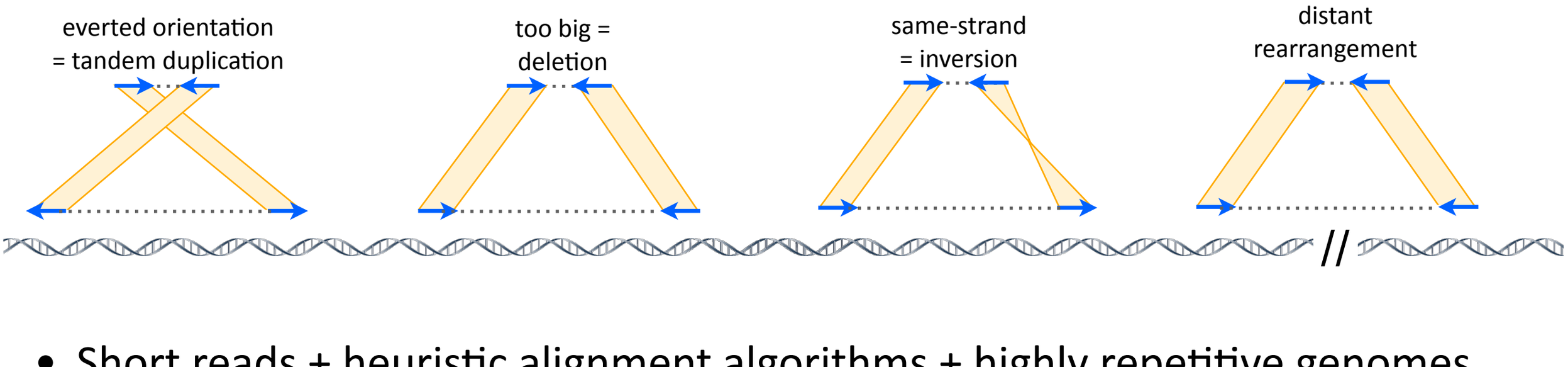
genomic rearrangements in 3 breast cancer genomes



Stephens et al., Nature (2009)

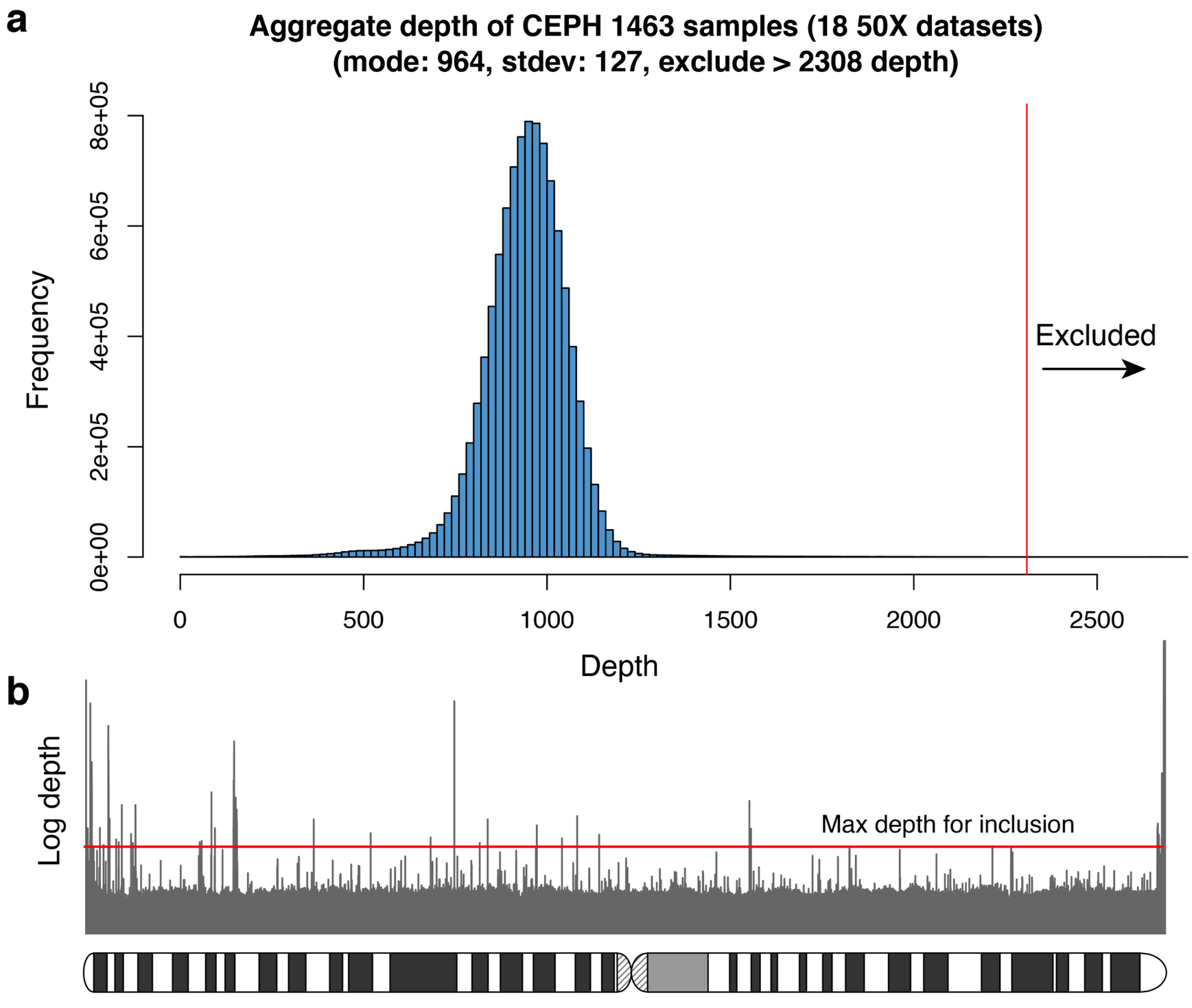
So, why is breakpoint mapping so #$%^&#@ hard?

1) Lots of false positives



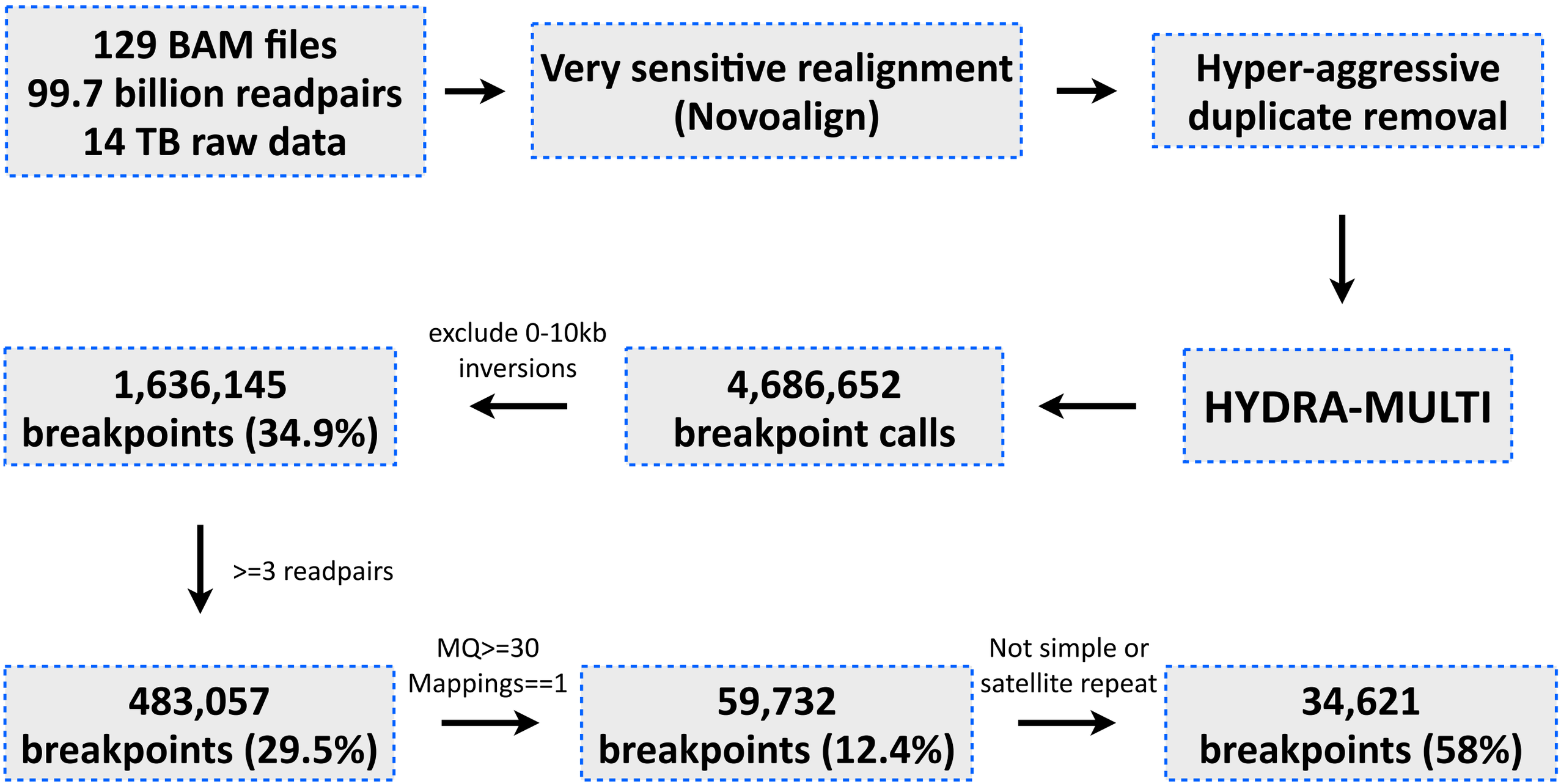
- Short reads + heuristic alignment algorithms + highly repetitive genomes = systematic alignment artifacts.
- Chimeric molecules + duplicate reads = clusters of discordant readpairs.
- Mis-assembled regions: assembly gaps, simple sequence repeats, segmental duplications, etc.
- Alignment of reads from unassembled regions (e.g., satellite repeats)
- General weirdness
- <u>FOR THESE REASONS, ALL SV MAPPING STUDIES USE FILTERING SCHEMES</u>

Identification of mis- or un-assembled genomic regions showing abnormally high read-depth



This identifies 10,701 loci encompassing ~10 Mb (0.34%) of the genome.

A semi-old data processing pipeline (128 genomes)



- **Only 0.4% of total breakpoint calls retained (2.1% not counting inversion artifact)!**
- **Filtering is a major source of false negatives (along with coverage); our FNR was >40%**

Malhotra et al., *Genome Research* (2013)

2) Lots of false negatives

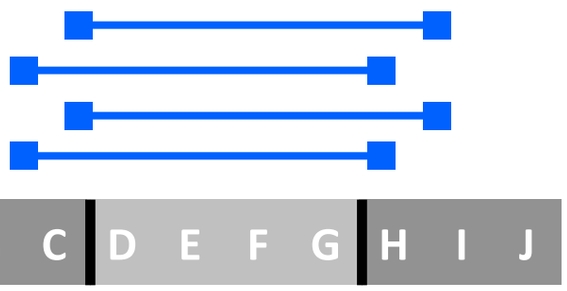
- Many current datasets have inadequate coverage for robust SV breakpoint detection.
- SVs are enriched in repetitive regions (e.g., transposons, segmental duplications). These are difficult to detect using standard read-alignment strategies.
- FILTERING!
- The false negative rate is impossible to measure properly, but is thought to be extremely high for most studies (>50%)
- When searching for spontaneous mutations in a family or a tumor/normal comparison, a false negative call in one sample is a false positive somatic call in another.

Joint variant calling allows for evidence to be "borrowed" across samples

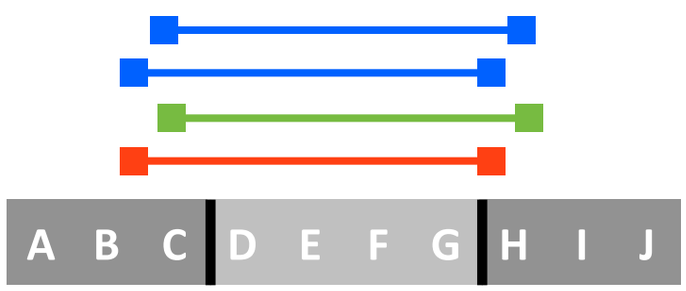
sample 1 | sample 2 | sample 3

Hydra Algorithm

private mutations



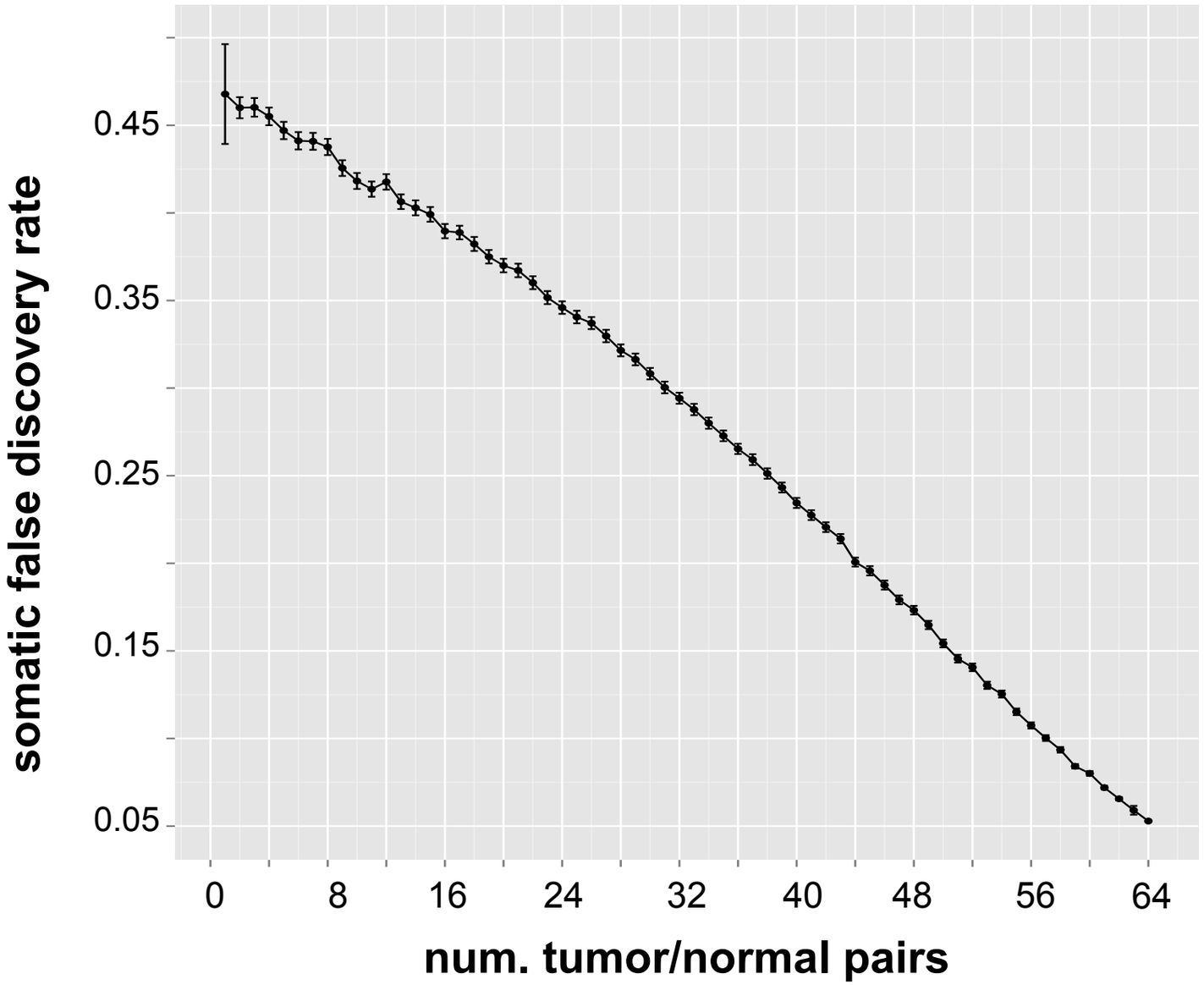
common variants
(& systematic false positives)



Quinlan et al., Cell Stem Cell (2011);

Note: GATK pioneered population-based SNP and INDEL detection; GenomeSTRiP and VariationHunter use a conceptually similar approach

Somatic SV misclassification rate (due to false negatives) decreases with more samples

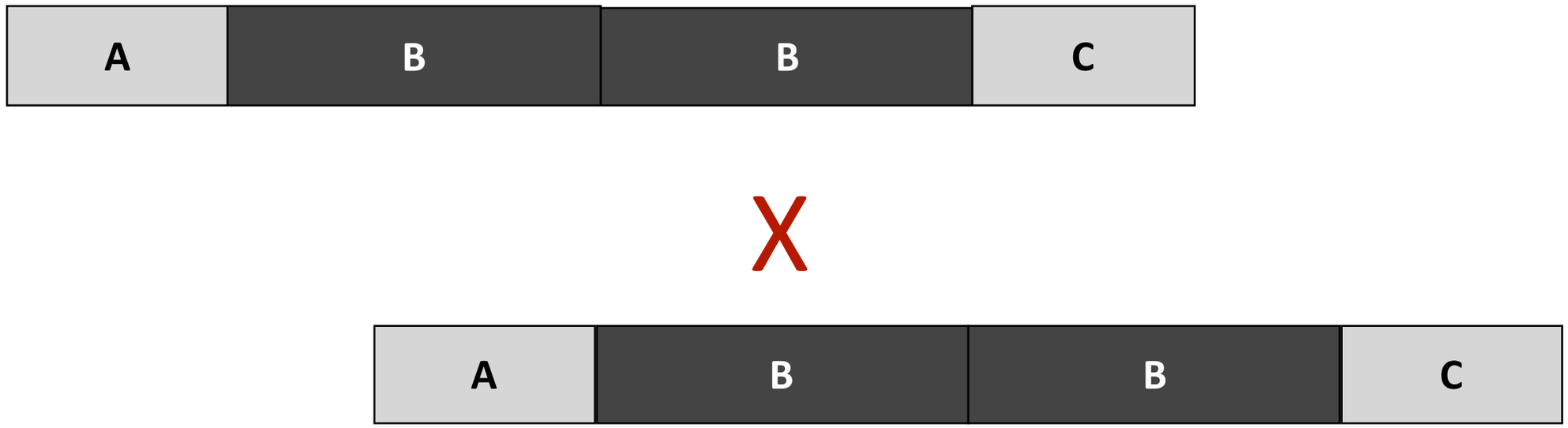


The N+1 problem: do we have to analyze hundreds of genomes each time we sequence a single new sample?

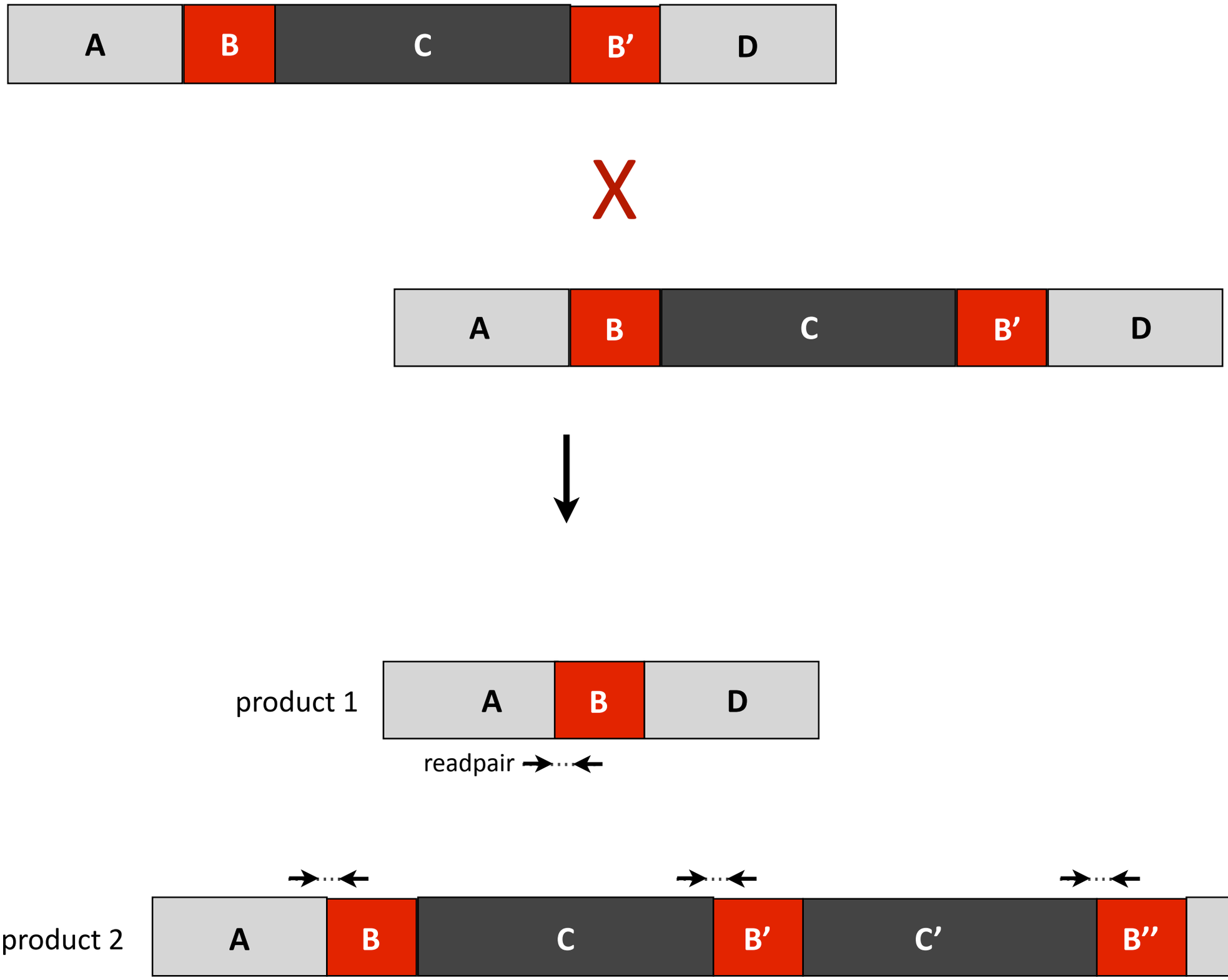
M. Lindberg & A. Quinlan

3) SVs formed by non allelic homologous recombination (NAHR) between large, highly similar repeats are virtually impossible to detect

Recombination within a tandem array

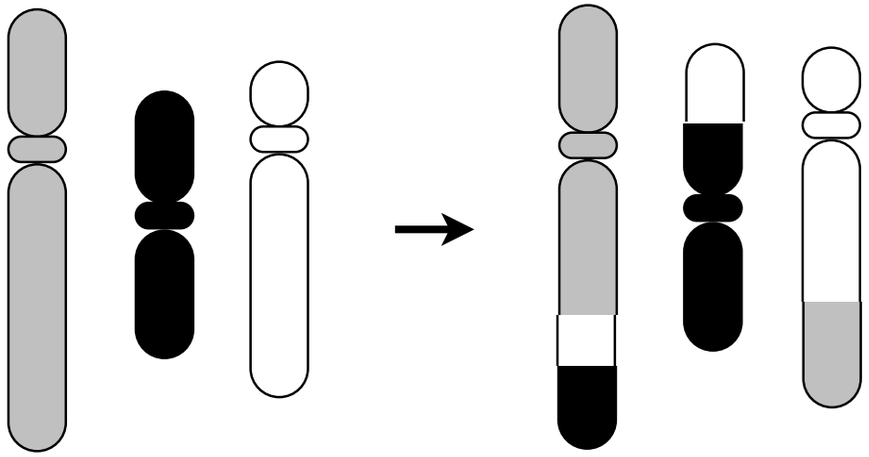


Recombination between large flanking repeats



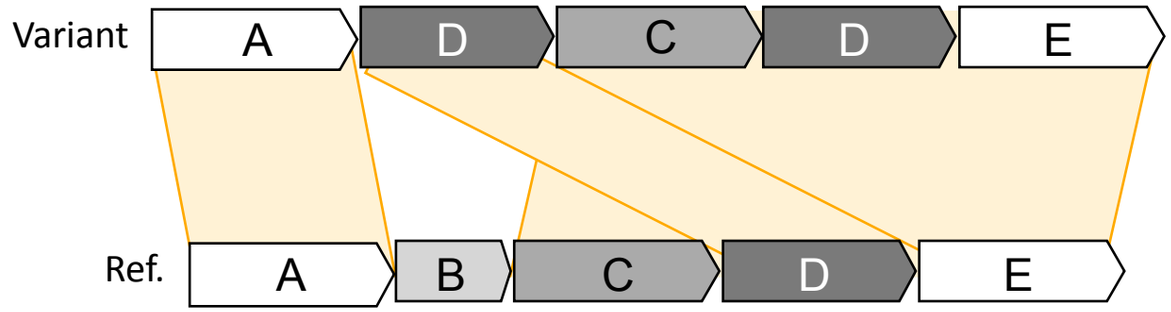
4) Complex rearrangements can produce extremely confusing breakpoint patterns

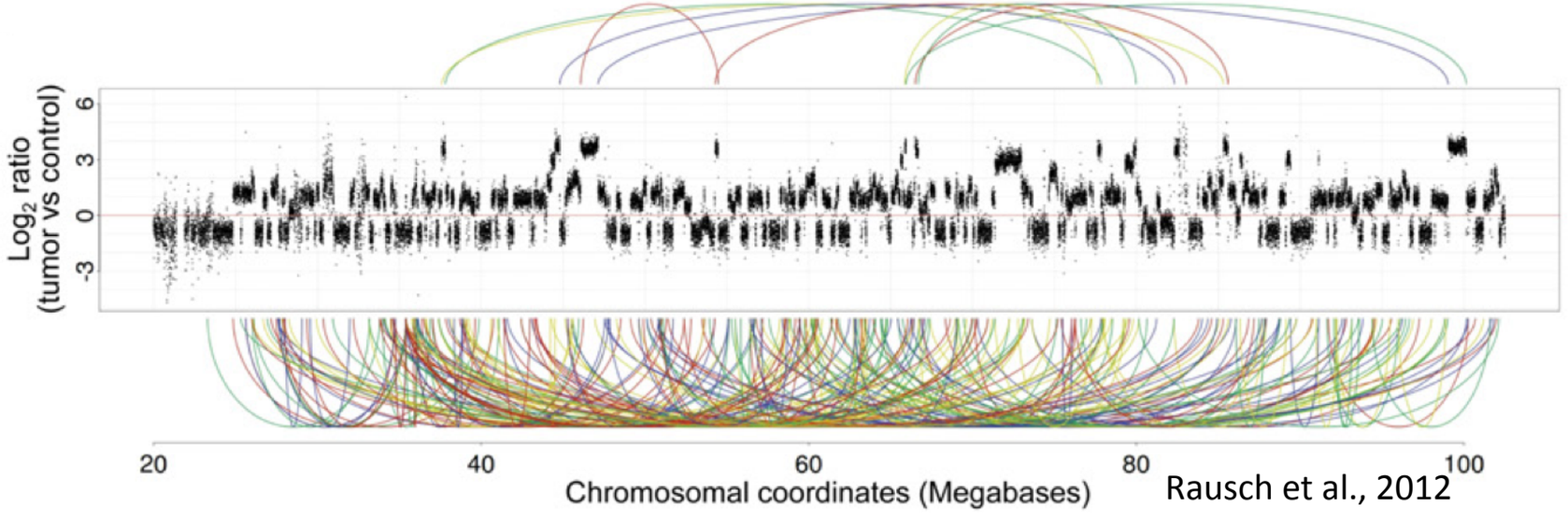
complex chromosomal rearrangement (CCR)



Chromothripsis (Stephens et al., 2011)

mildly complex germline SV





Rausch et al., 2012

<u>Definition</u>: a variant with multiple clustered breakpoints that most likely arose through a single mutation, but cannot be explained by a simple end-joining or recombination event.

<u>Implications:</u>

- multiple simultaneous mutations = punctuated evolution
- novel mechanisms
- difficult to detect and interpret

Acknowledgements

Current Lab Members

Colby Chiang

Ryan Layer

Greg Faust

Mike Lindberg

Collaborator

Aaron Quinlan (UVA)

Funding

NIH New Innovator Award
Burroughs Wellcome Fund Career Award
NIH (5R01MH102698-02)
March of Dimes Basil O'Connor Award

Contributing Ex-Lab Members

Svetlana Shumilina

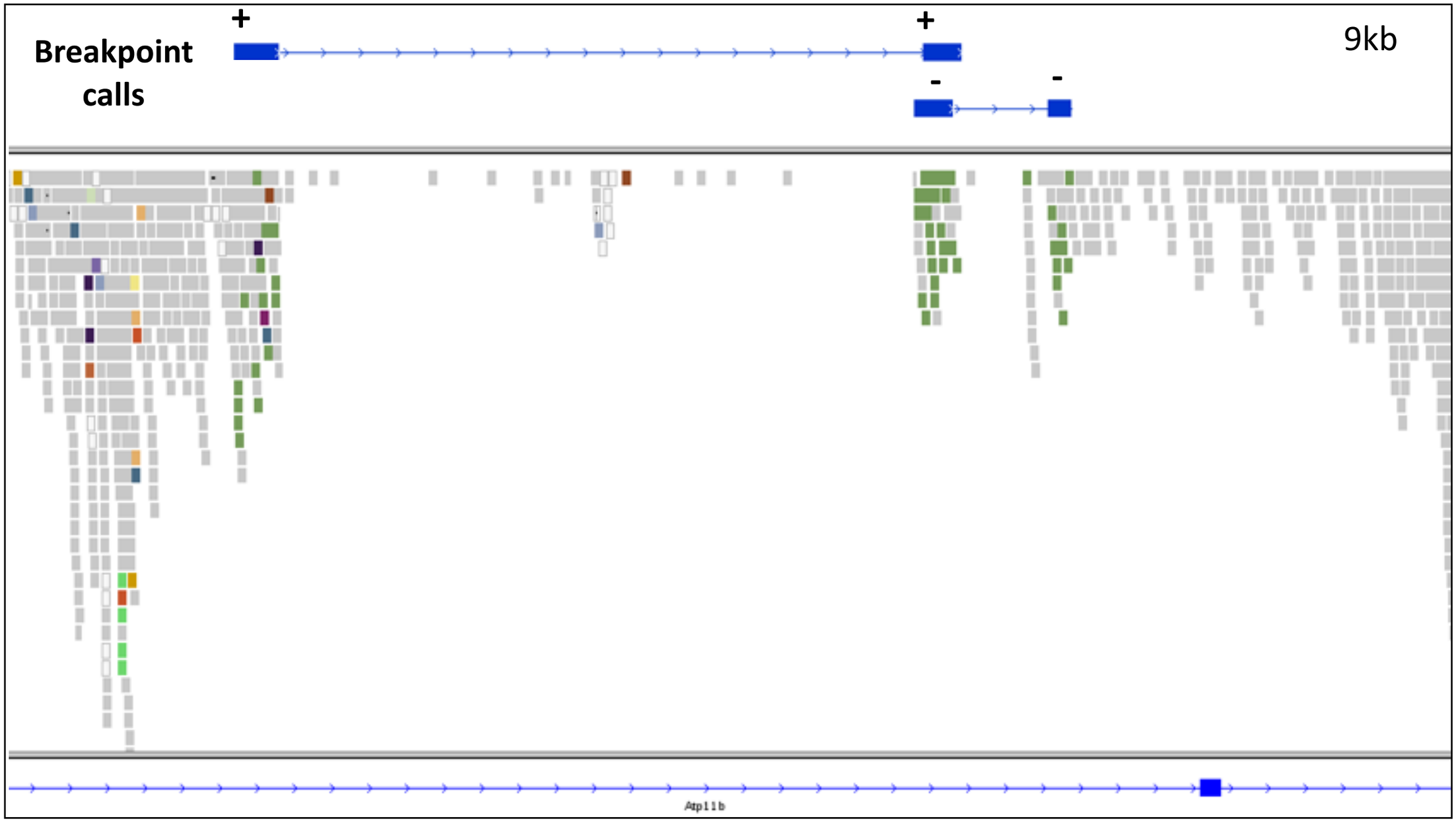
Ankit Malhotra

Mitchell Leibowitz

Royden Clark

Some examples of complex variants

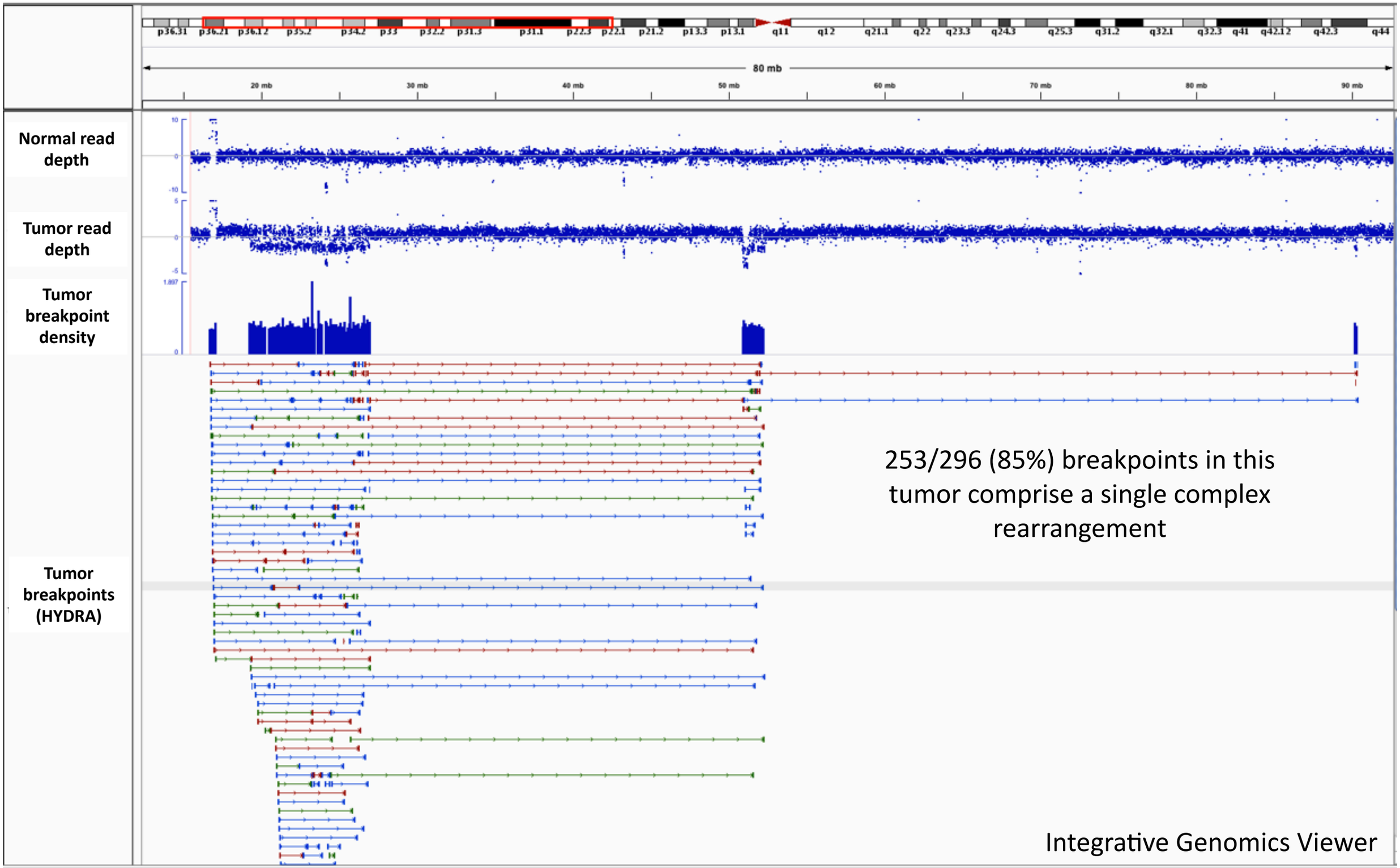
A mildly complex germline SV



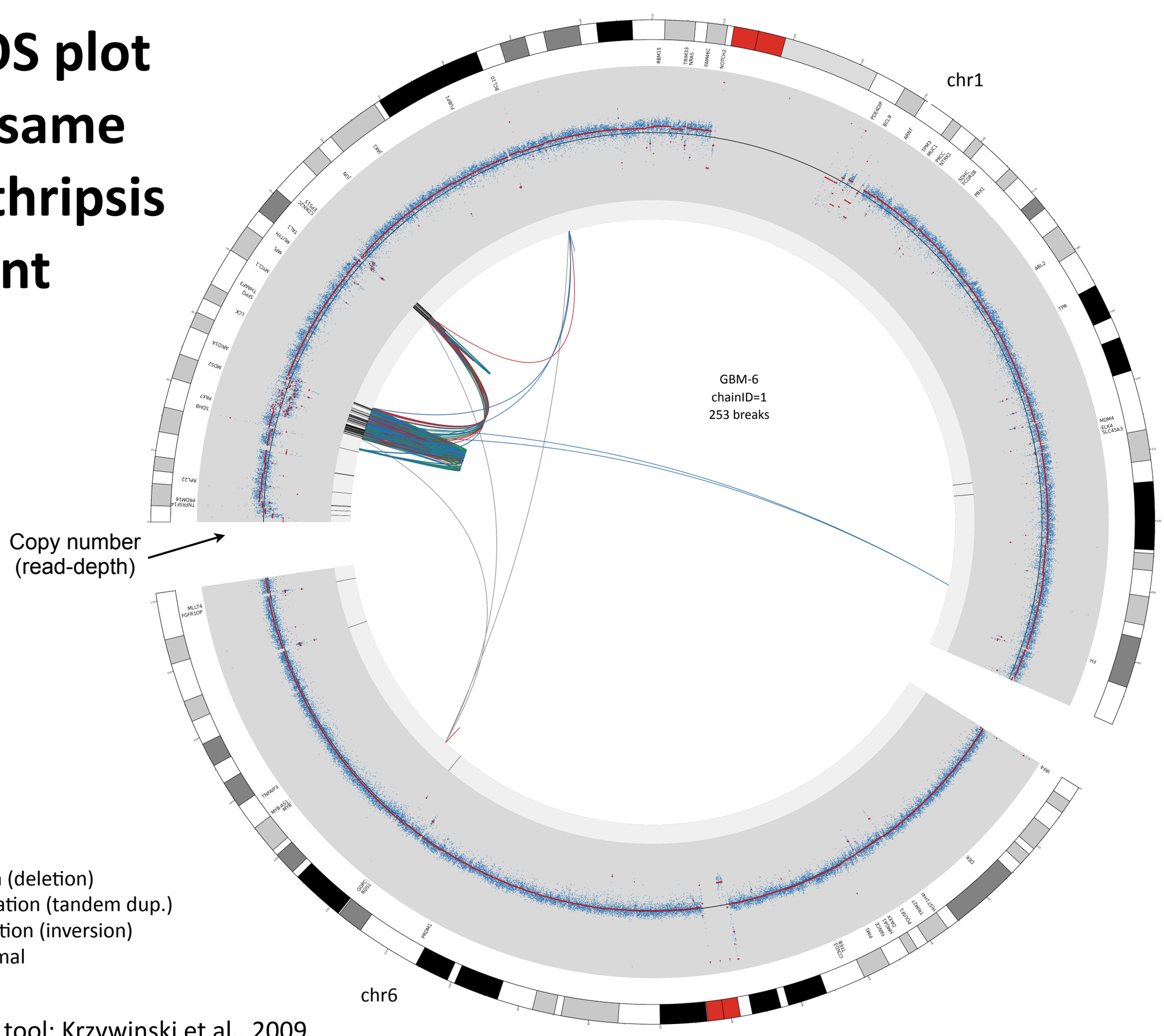
Important: we cannot infer variant class based solely upon relative read orientation; e.g., an apparent deletion may really be part of a complex rearrangement

Quinlan et al., 2010 (Genome Research)

A chromothripsis event shown in IGV



A CIRCOS plot of the same chromothripsis event

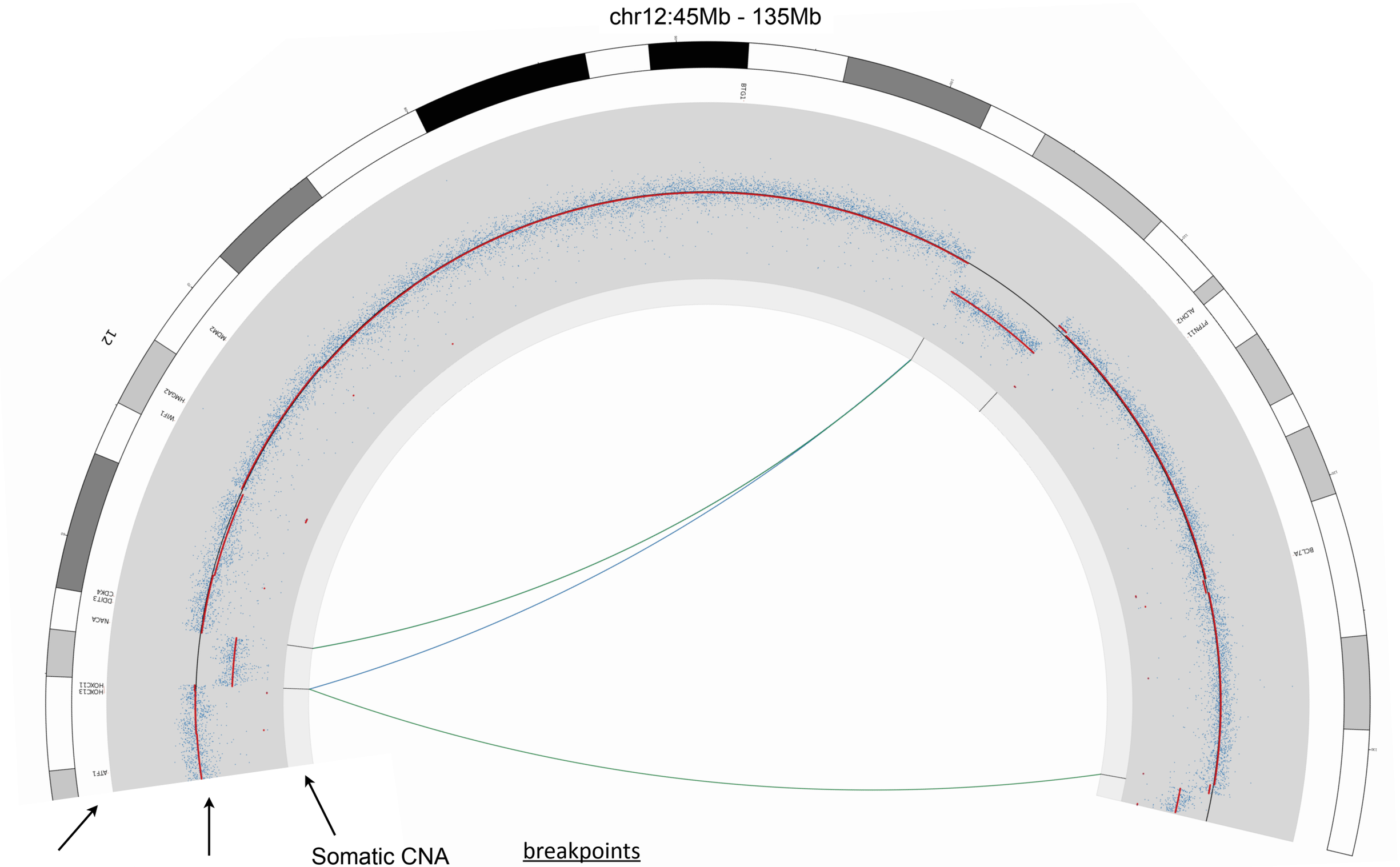


breakpoints

red : direct orientation (deletion)
green : everted orientation (tandem dup.)
blue : inverted orientation (inversion)
grey : inter-chromosomal

Circos visualization tool: Krzywinski et al., 2009

CNVs are often linked by complex rearrangements



COSMIC genes

Copy number (read-depth)

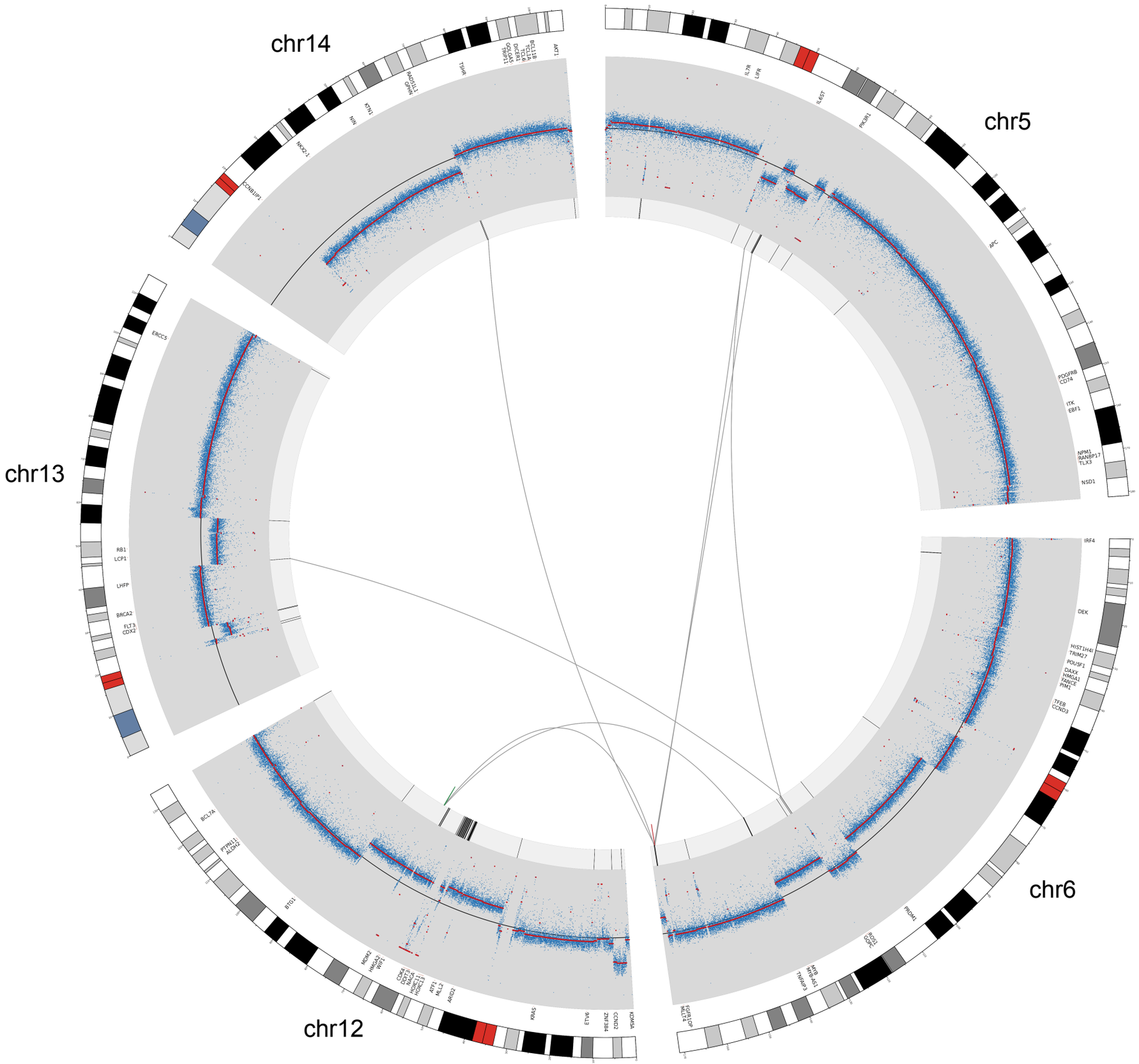
Somatic CNA "change-points"

<u>breakpoints</u>

red : direct orientation (deletion)
green : everted orientation (tandem dup.)
blue : inverted orientation (inversion)
grey : inter-chromosomal

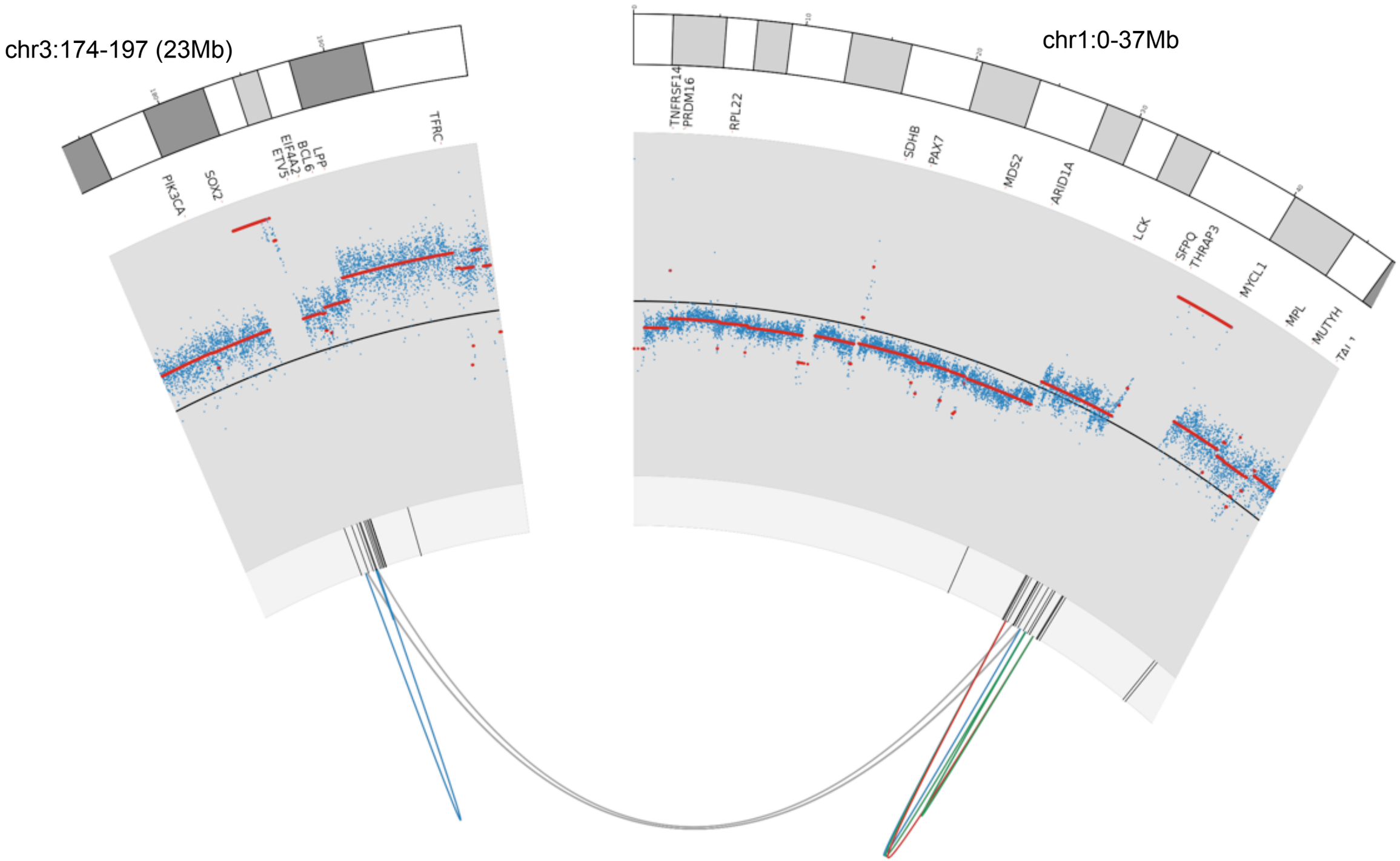
Circos visualization tool: Krzywinski et al., 2009

A more severe 9-breakpoint event



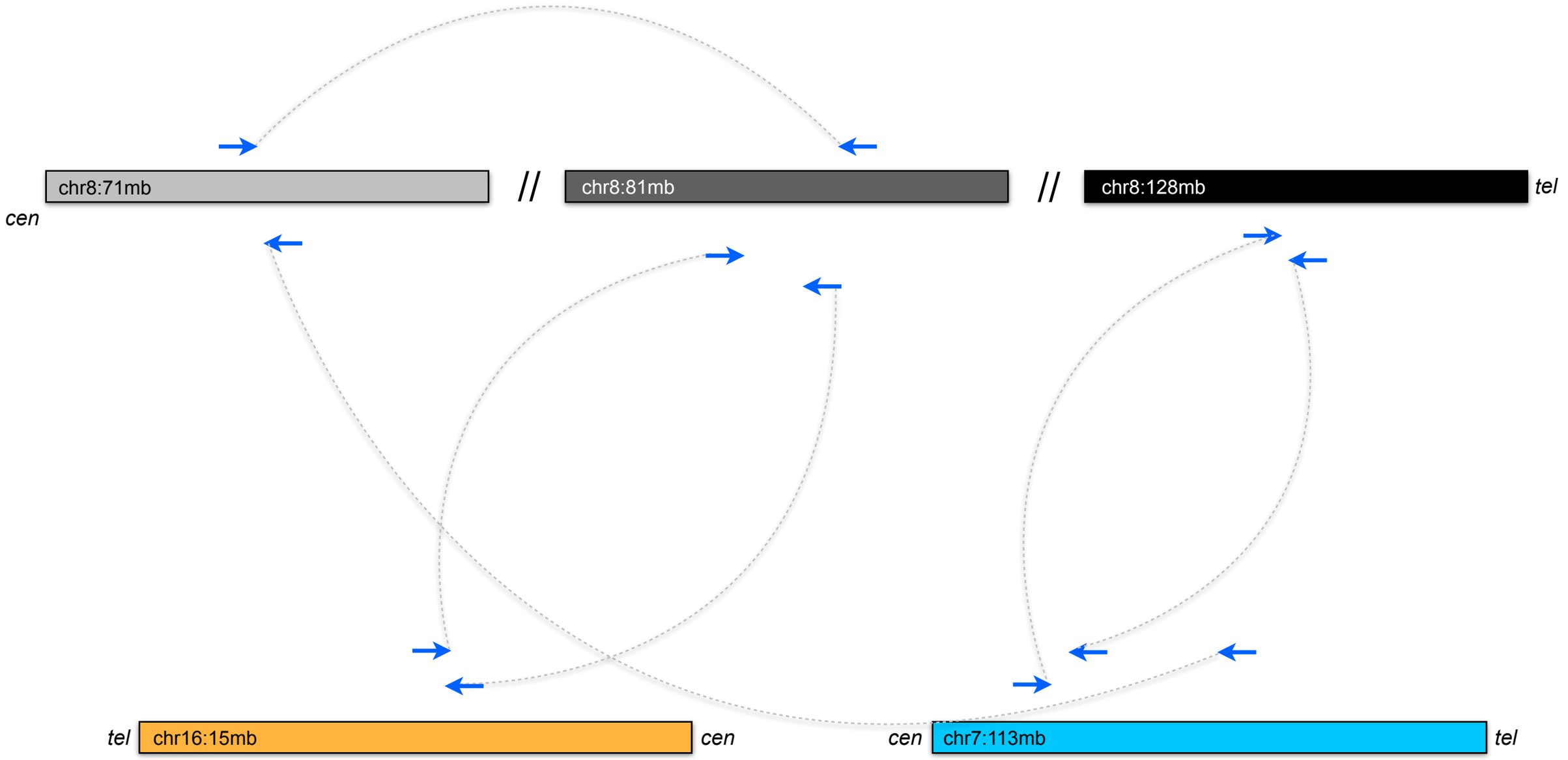
GBM-19
chainID: 2
9 breaks

Co-amplification of genes from different chromosomes

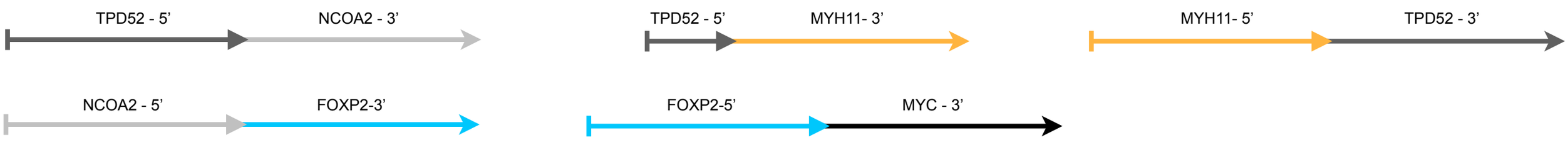


LUSC-7
chainID: 1
12 breaks

One complex balanced rearrangement, 5 fusion products



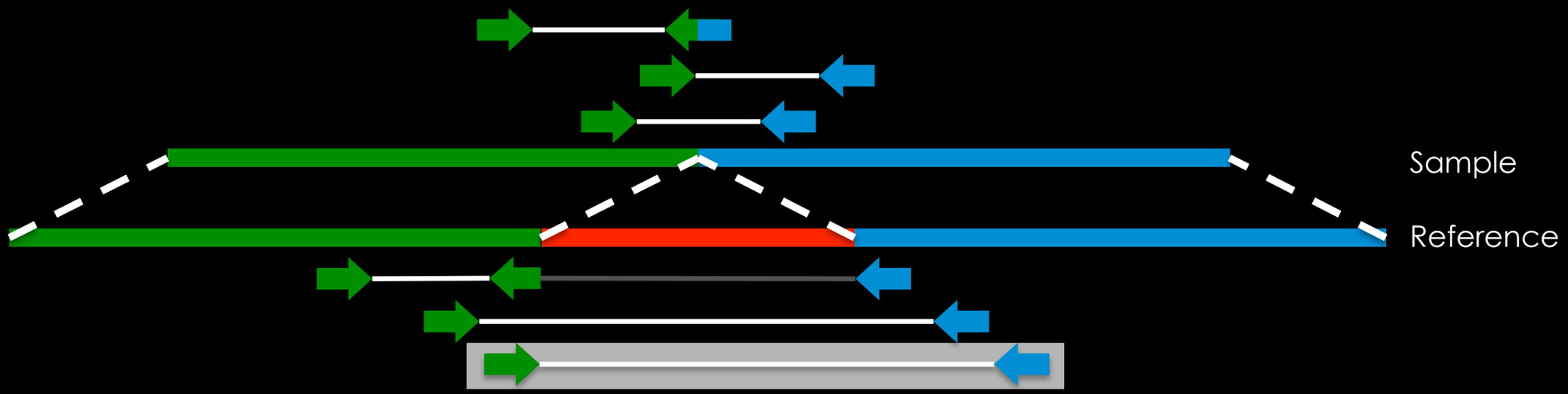
fusion products

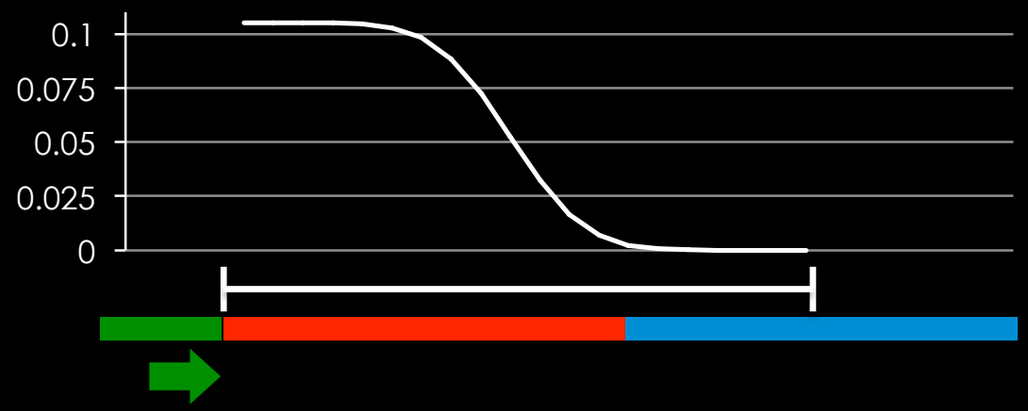


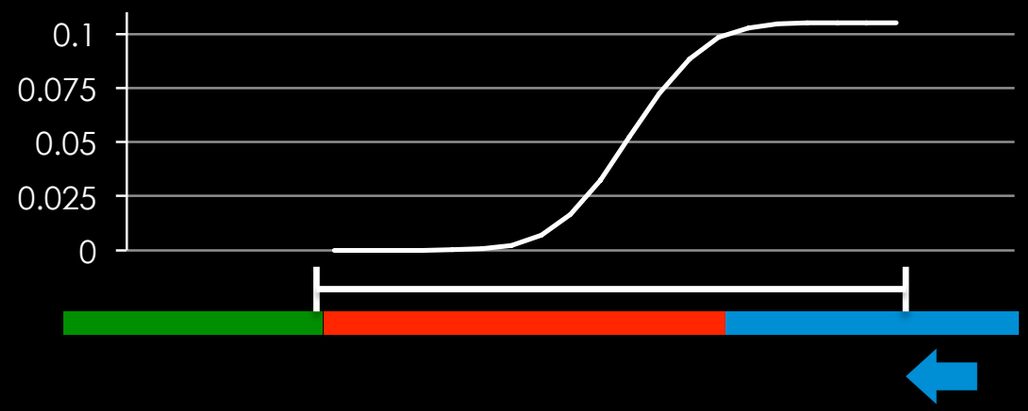
w/ Elaine Mardis & Co.

An example of how LUMPY clusters discordant and split-read mappings

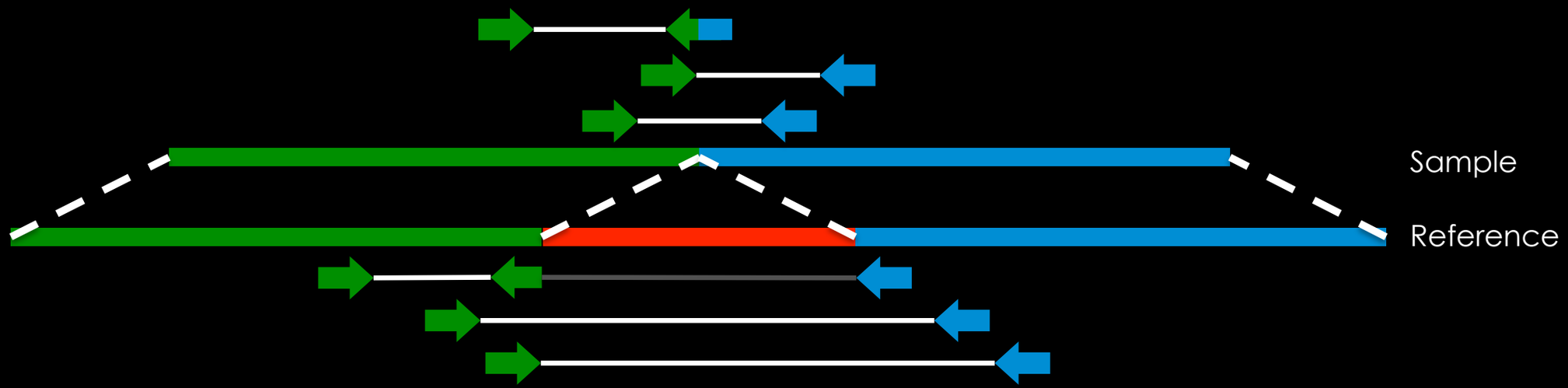
Pooling Evidence

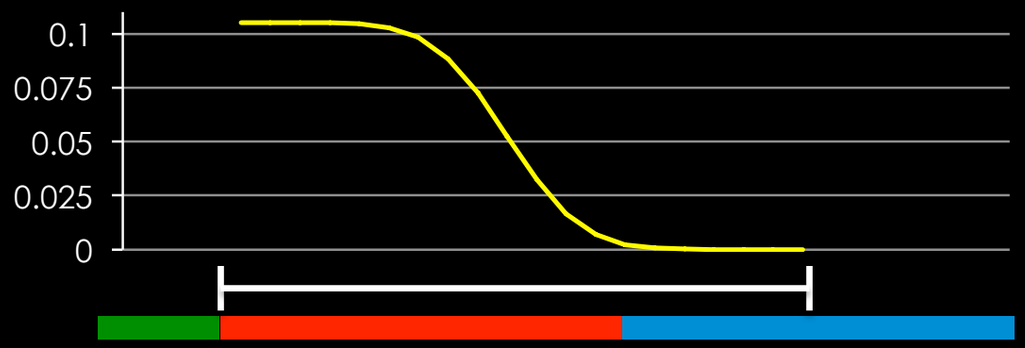


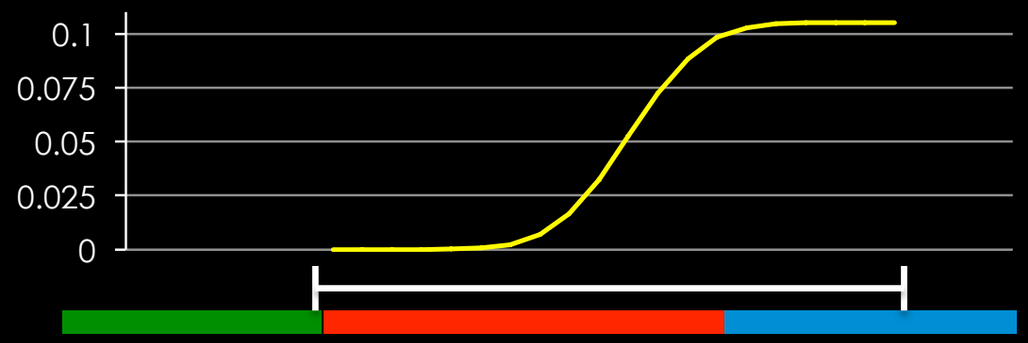




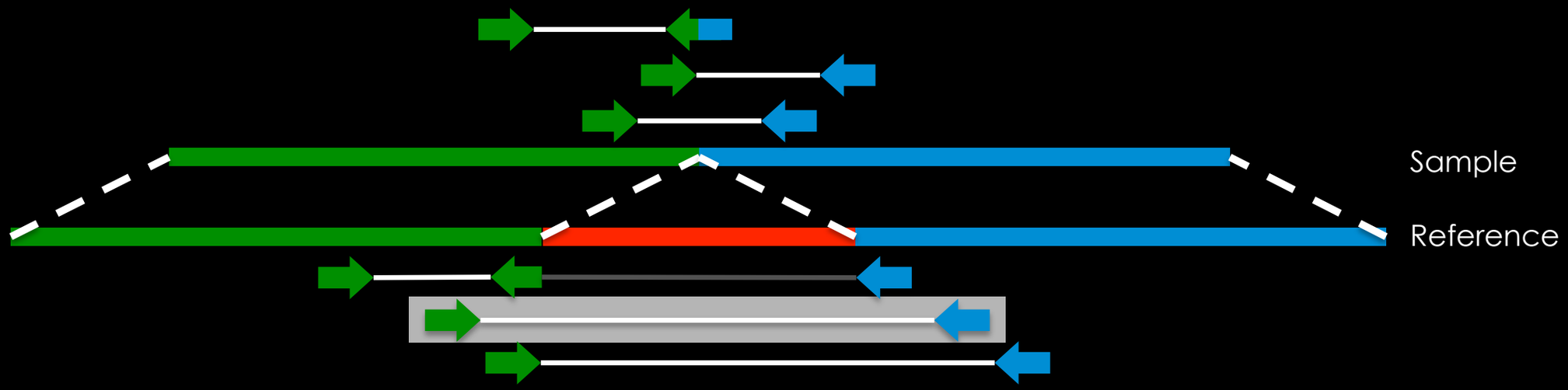
Pooling Evidence

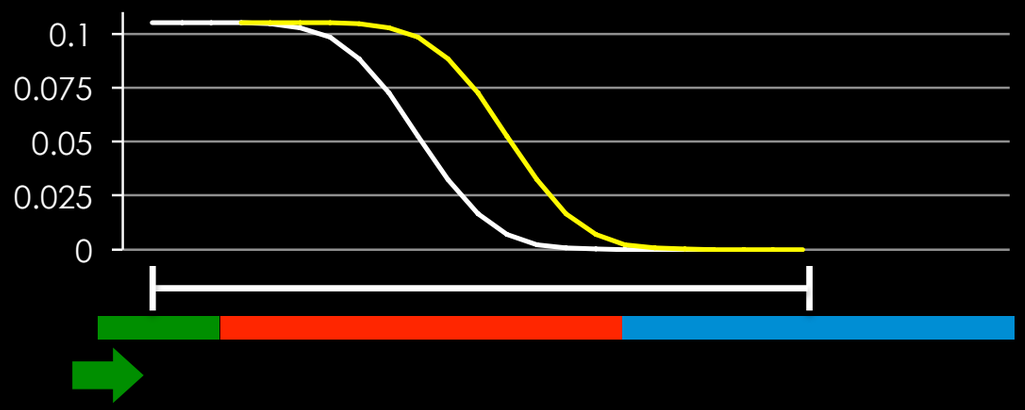


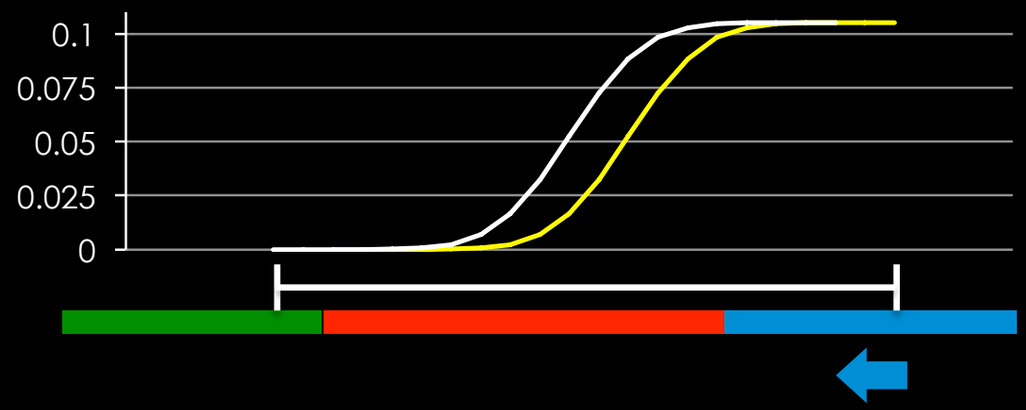




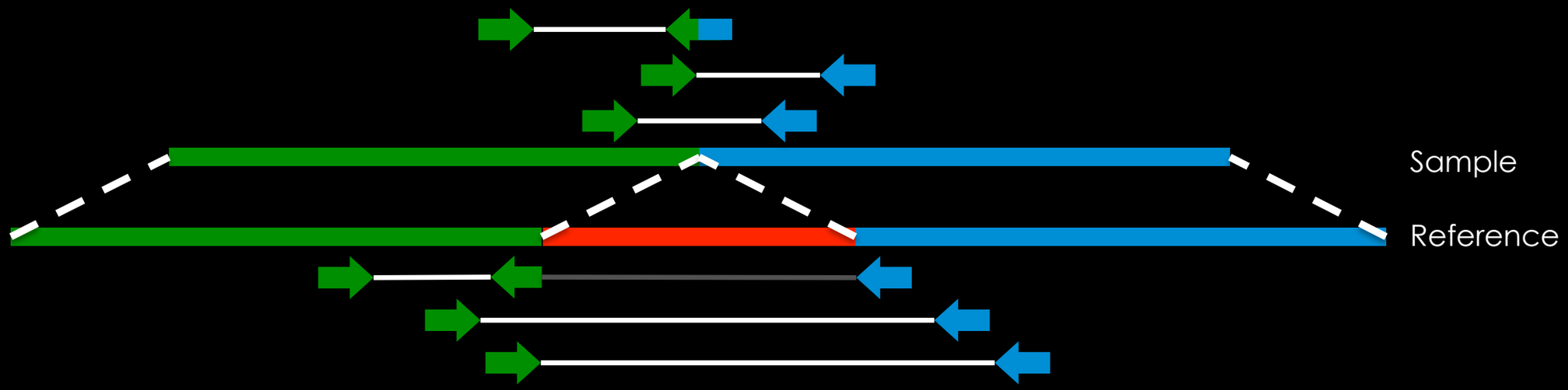
Pooling Evidence

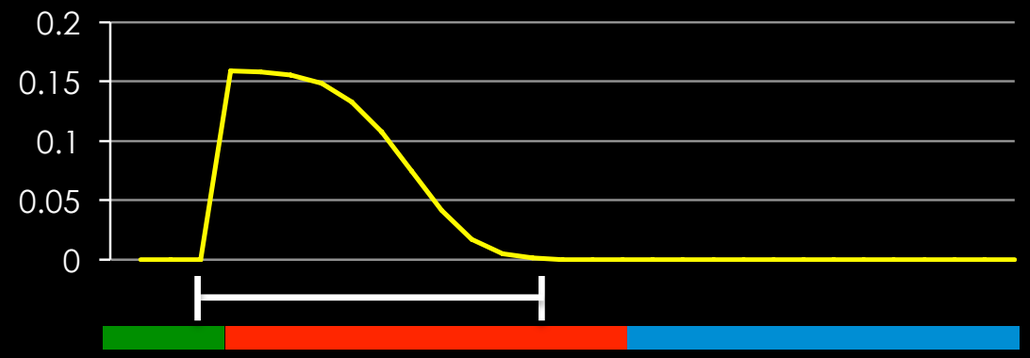


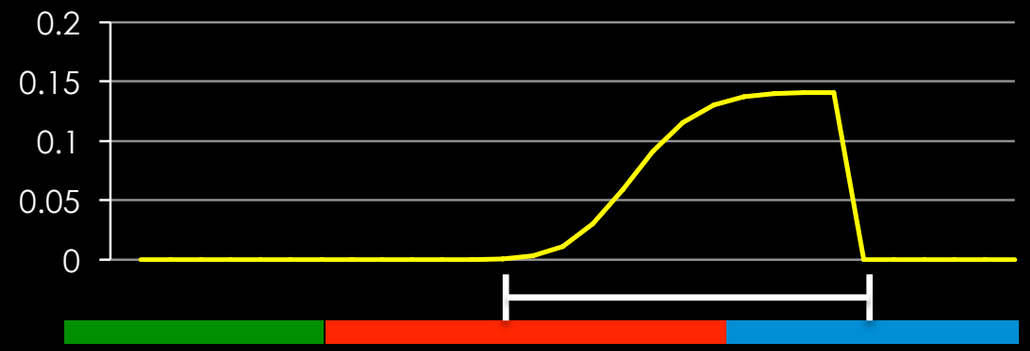




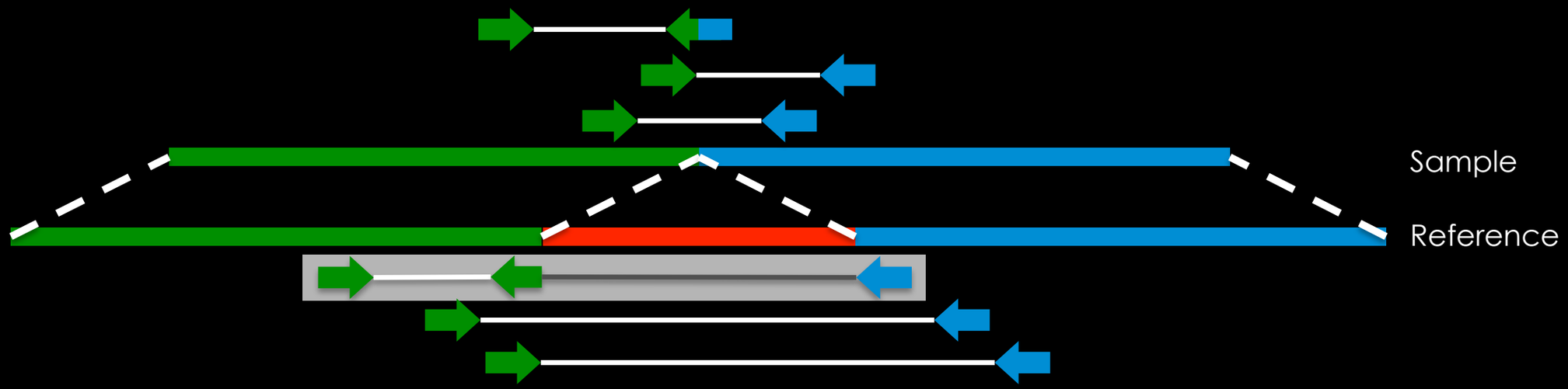
Pooling Evidence

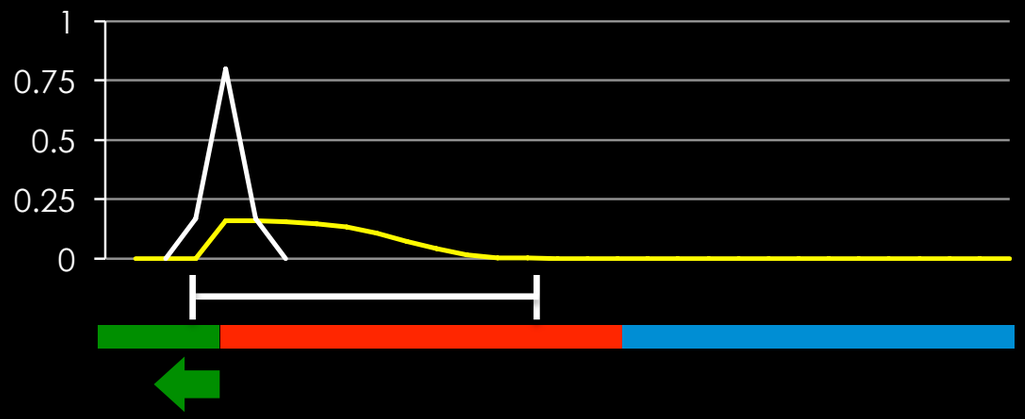


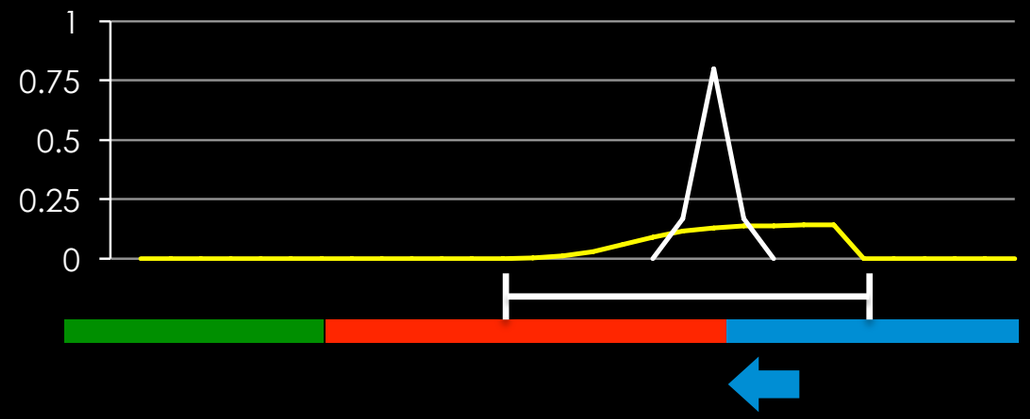




Pooling Evidence







Pooling Evidence