COMP90016 Computational Genomics: Structural Variations in DNA and Bioinformatics Detection Methods Part II

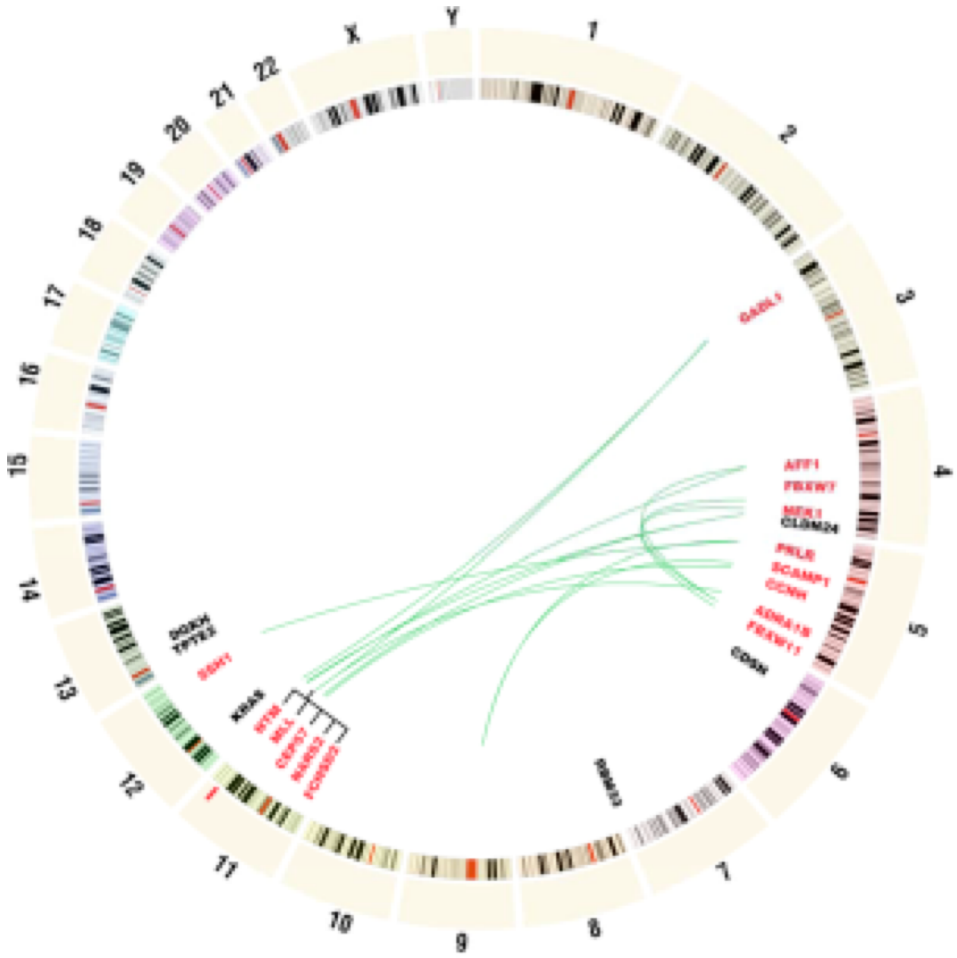


Image from: Dobbins et al 2013: **silent mutational landscape of infant *MLL-AF4* pro-B acute lymphoblastic leukemia**

Overview

- Motivation
- SV detection with paired-end reads
- SV detection with split (soft-clipped) reads
- SV detection

SV Detection with Anomalous Paired-end Reads

- What kind of SVs create signal in paired mapping of reads?
  - Deletions
  - Duplications
  - Translocations
  - Inversions
  - … and more.

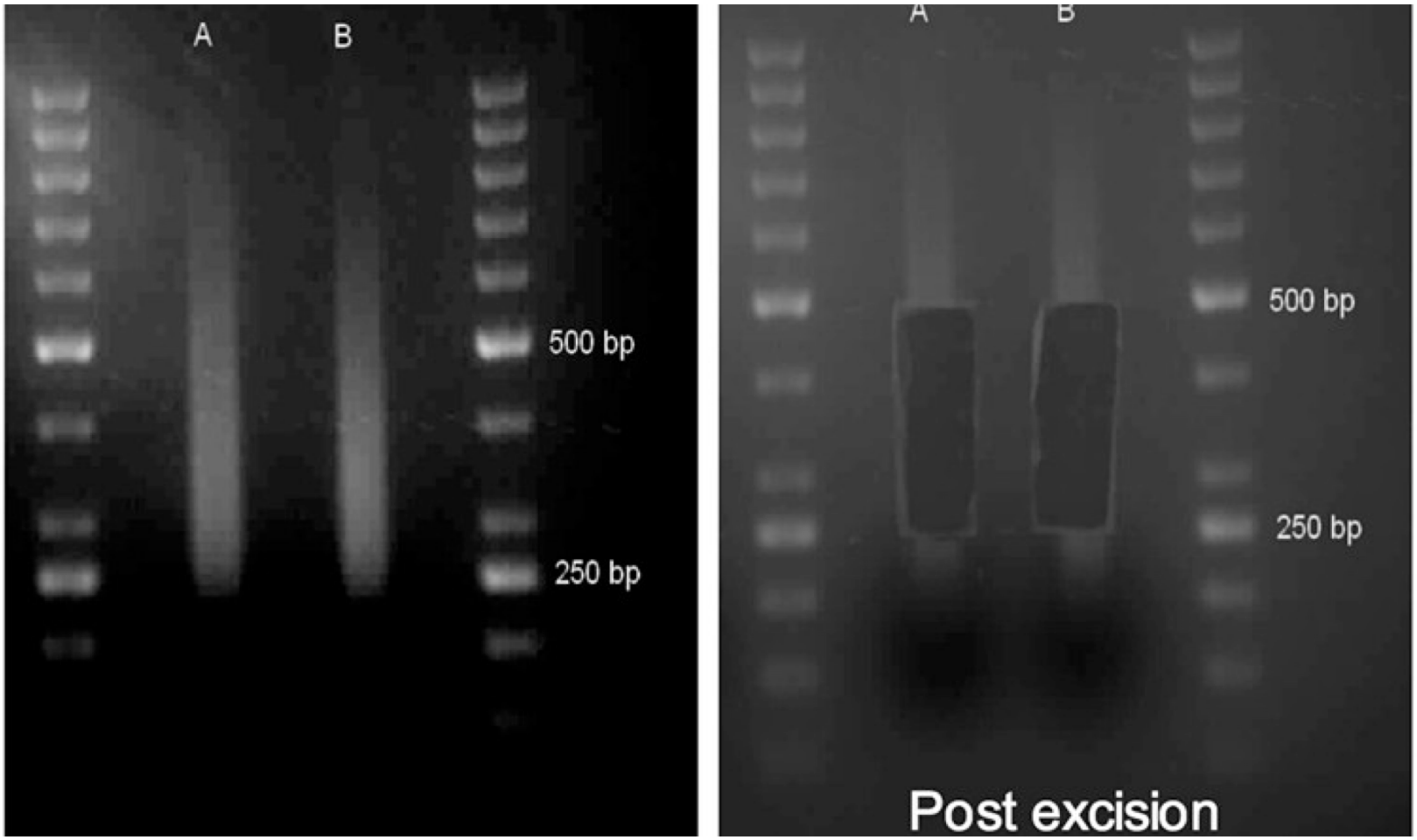
- Methods using this signal
  - Clustering of Anomalous paired-end reads

Paired-End Sequencing Recap

1. The Genome gets amplified and randomly fragmented.
2. The fragments are run on a gel (gel-electrophoreses) to separate by size.
3. A subset of DNA fragments is extracted from the gel (see next slide).
4. Fragments are sequenced from both ends.

The length of a fragment is referred to as the fragment length or, curiously, as the insert size.

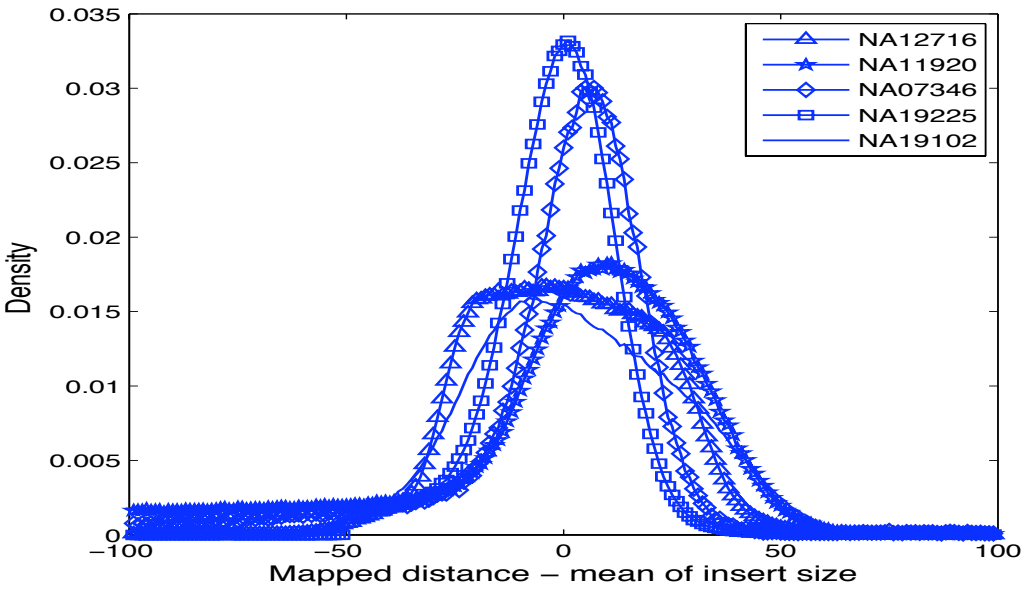
Size Selection For Paired-end Sequencing



Size selection ensures an approximately known distance between each two reads in a read pair.

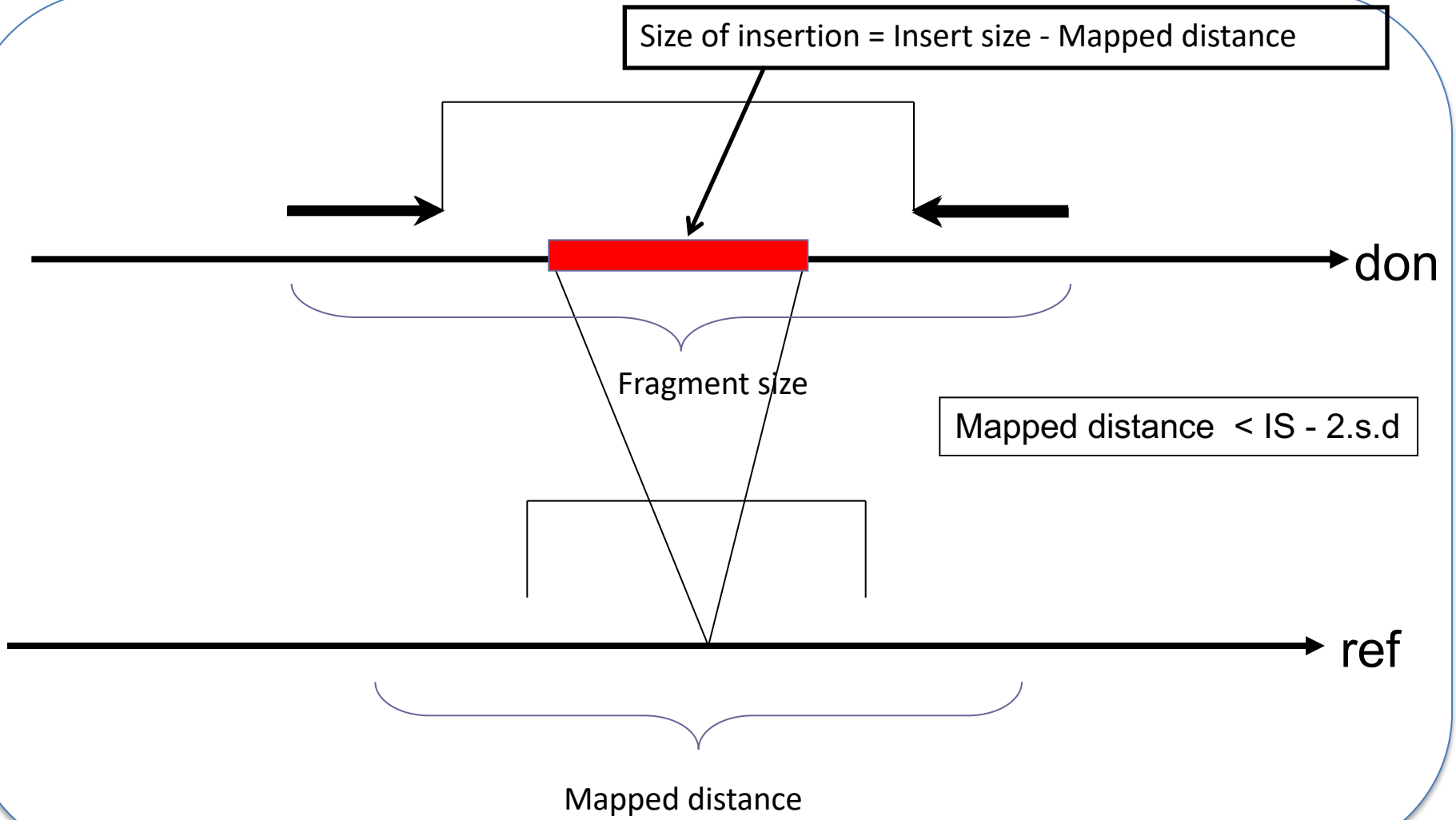
Insert Sizes and the PE Strategy

Examples of normalised insert size distributions

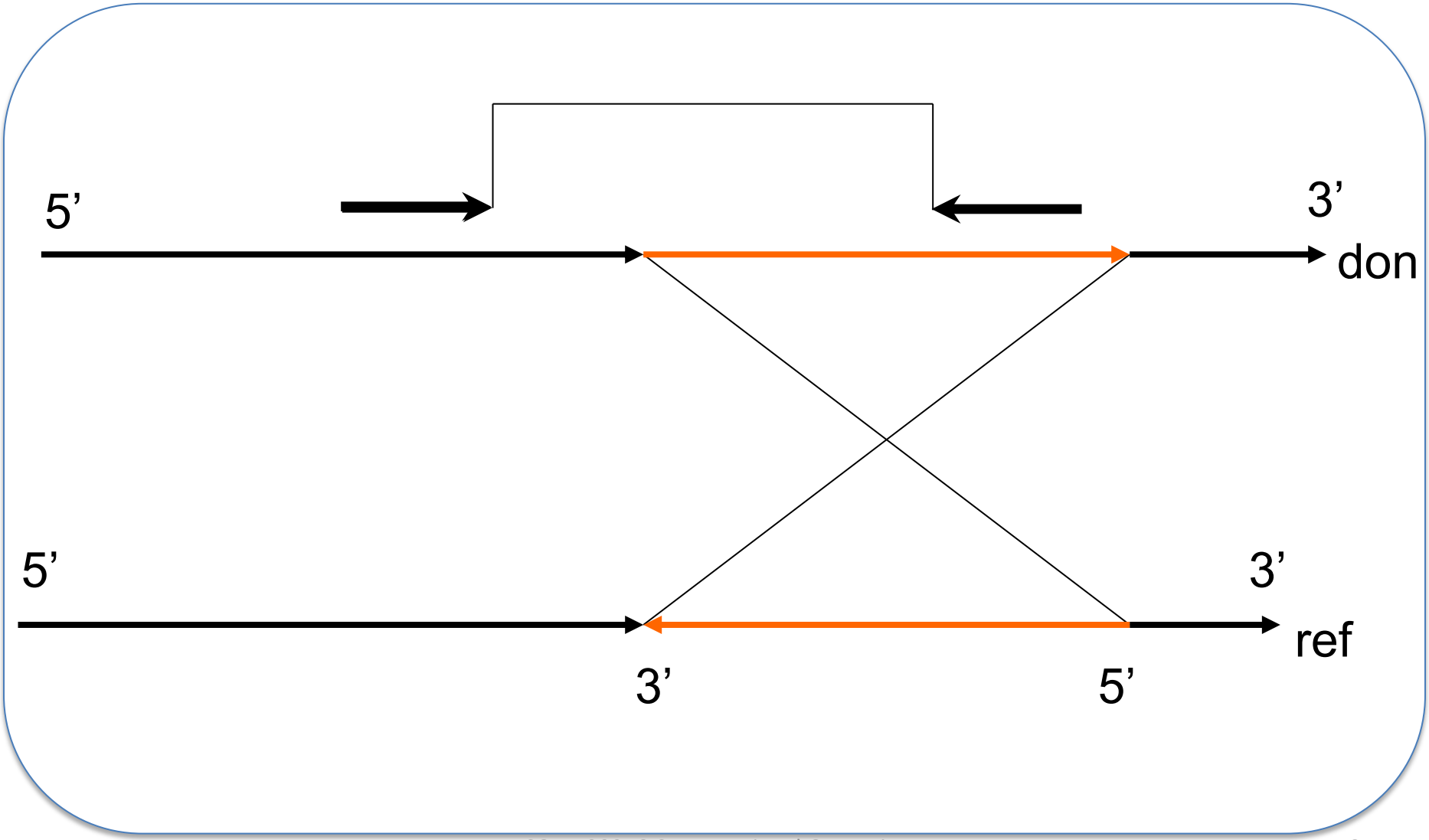


Any pair mapping within the expected distance and orientation is referred to as mapping concordantly – otherwise as anomalous.

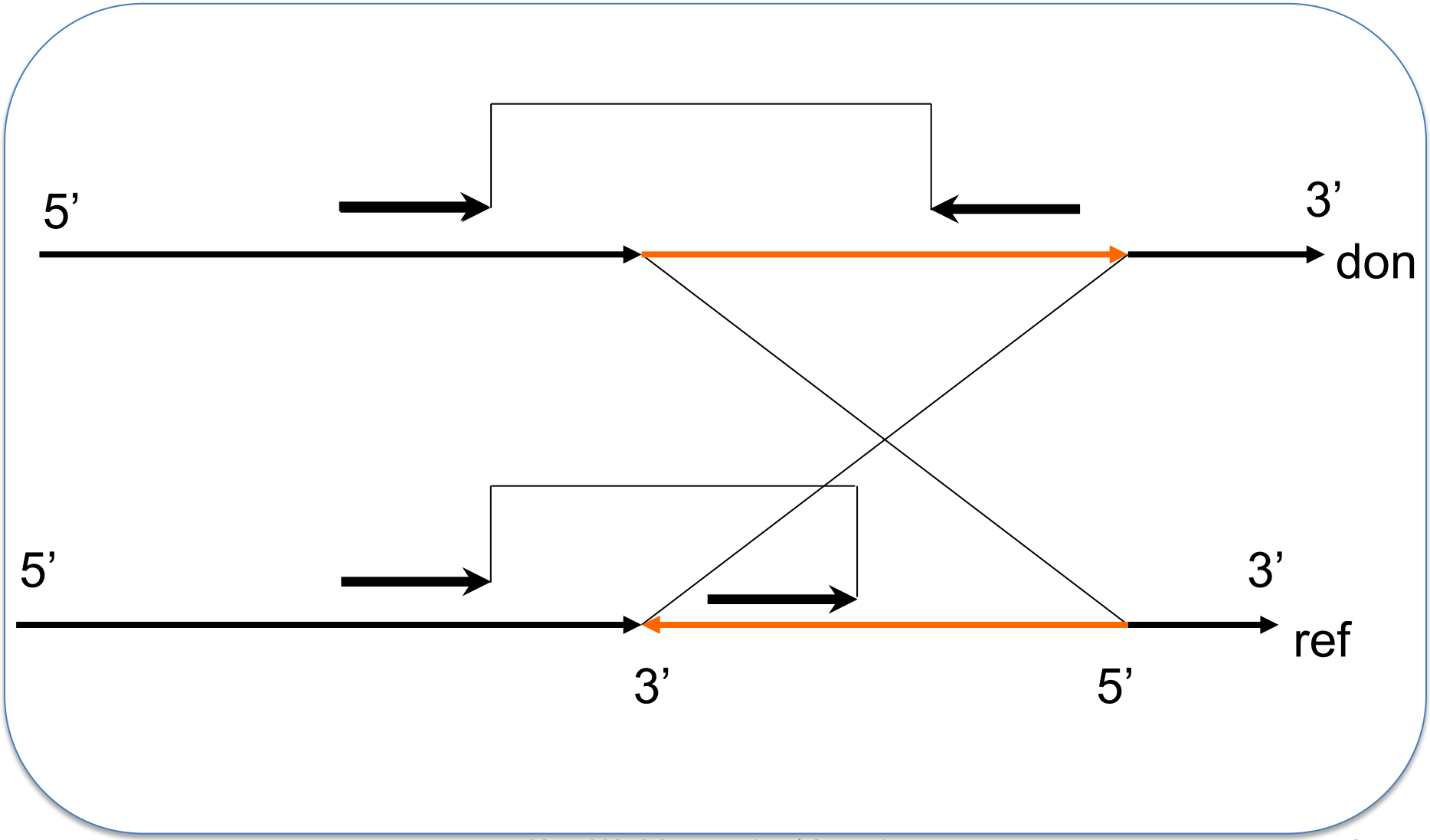
Insertion: signature



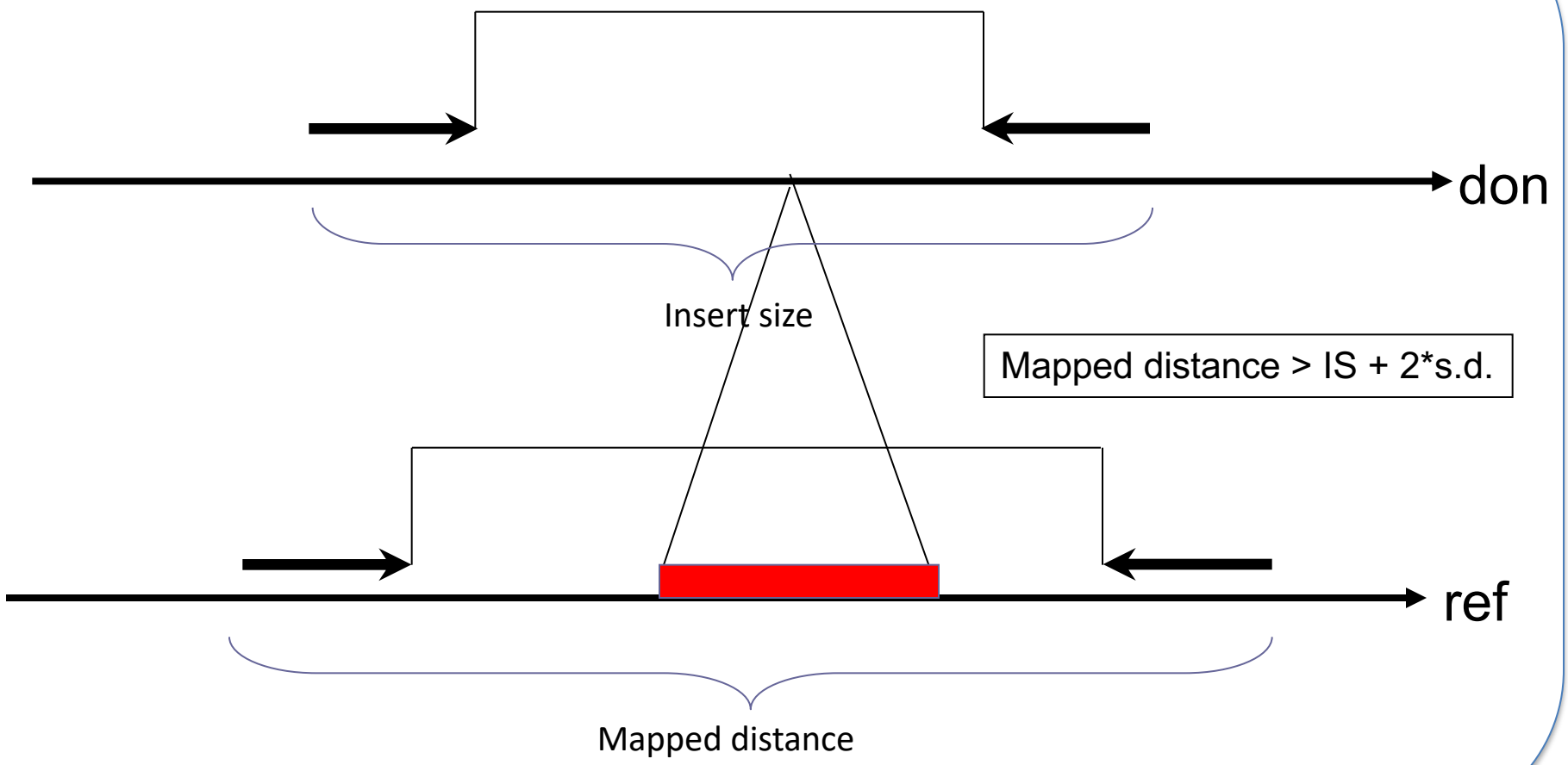
Inversion: signature



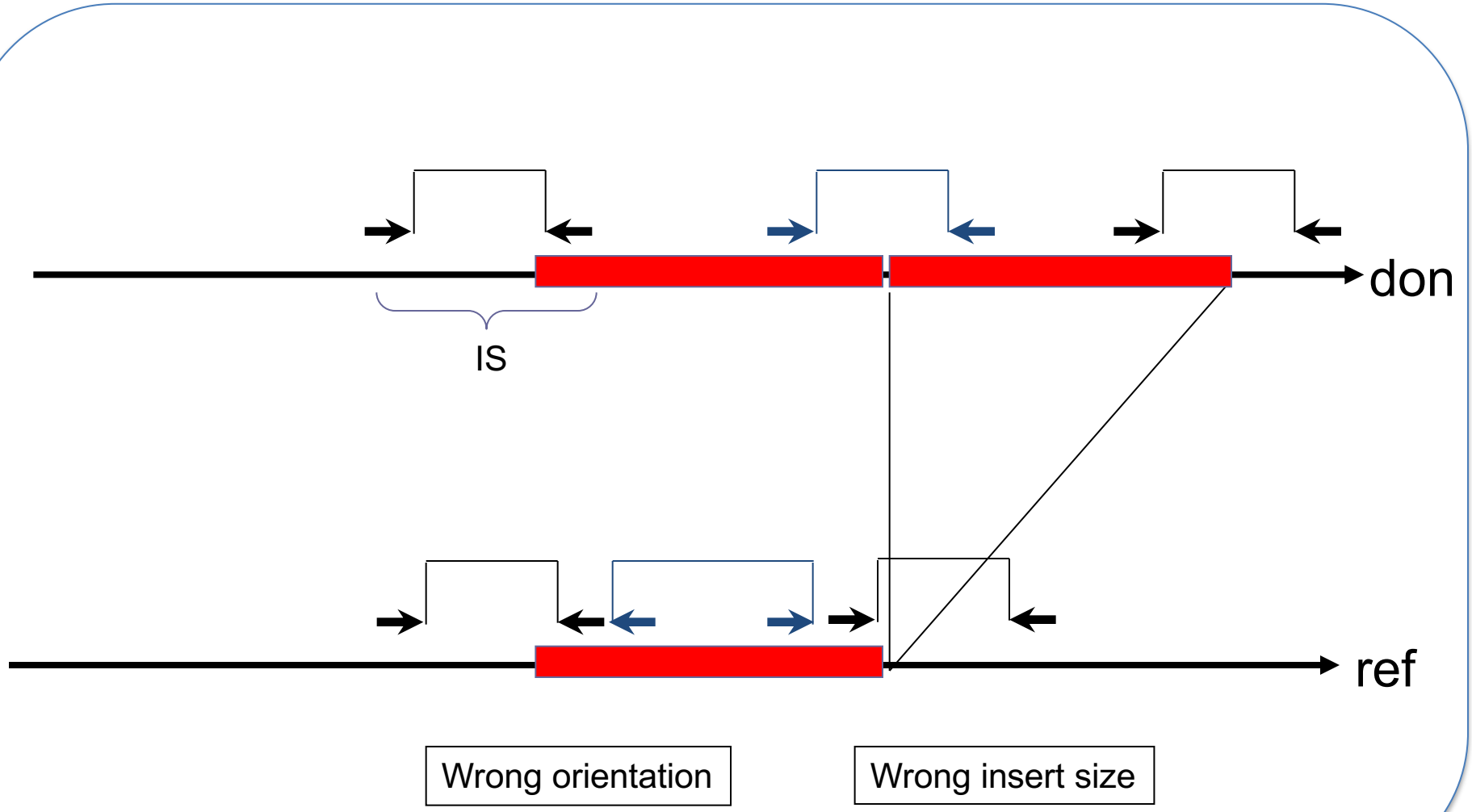
Inversion: signature



Deletion: signature



Tandem duplication: signature



PE SV Signatures

- There are more signatures than inversion, insertion, tandem duplication, and deletion.
- For instance, there are inter-chromosomal SV events that join DNA from two different chromosomes.
- Anything that does not map in the correct distance and/or orientation bears a SV signature.
- How many are there?
- We will explore more a bit later.

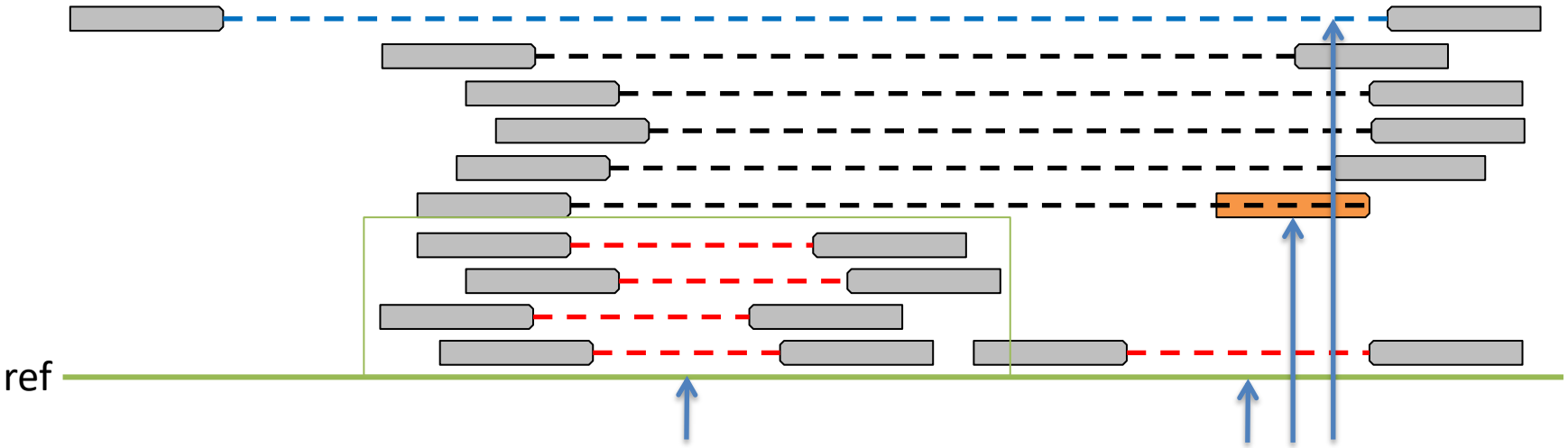
SV Detection Methods with PE reads

- Clustering of anomalous PE reads that have the same signature.
  - Around 10% of the pairs map anomalously* – but only a small fraction actually support real SV events.
- For further reading:
  - **BreakDancer: an algorithm for high-resolution mapping of genomic structural variation**, Chen et al, 2009.

*In real data. This week's workshop shows different numbers, but on simulated data

PE Read Clustering Example

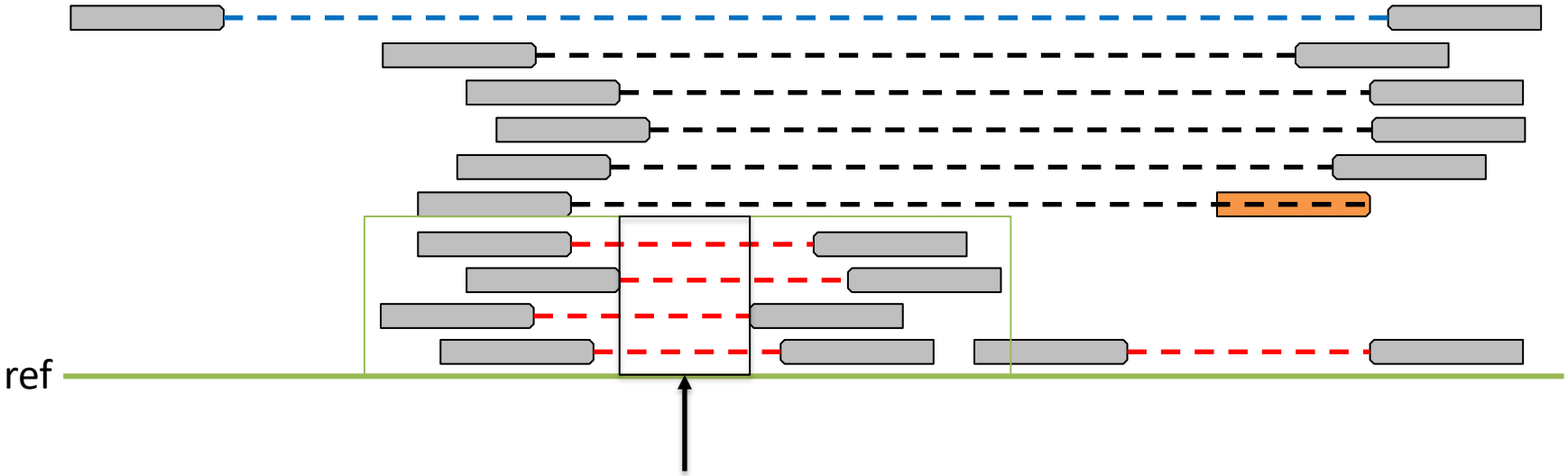
PE reads represented by dashed lines between mapped reads.
Red connection indicates closer than expected alignment.
Blue connection indicates large mapping distance.
Orange read to mark unexpected mapping orientation



Cluster of reads mapping too close to each other demonstrate compound evidence of a heterozygous insertion event.

Isolated read pairs show only weak evidence of a real event in the data

PE Read Clusters to Location



Approximate location of insertion in the reference genome

- If the above cluster indicates an insertion event, where exactly does this insertion take place?
- What is the size (length) of the inserted sequence?

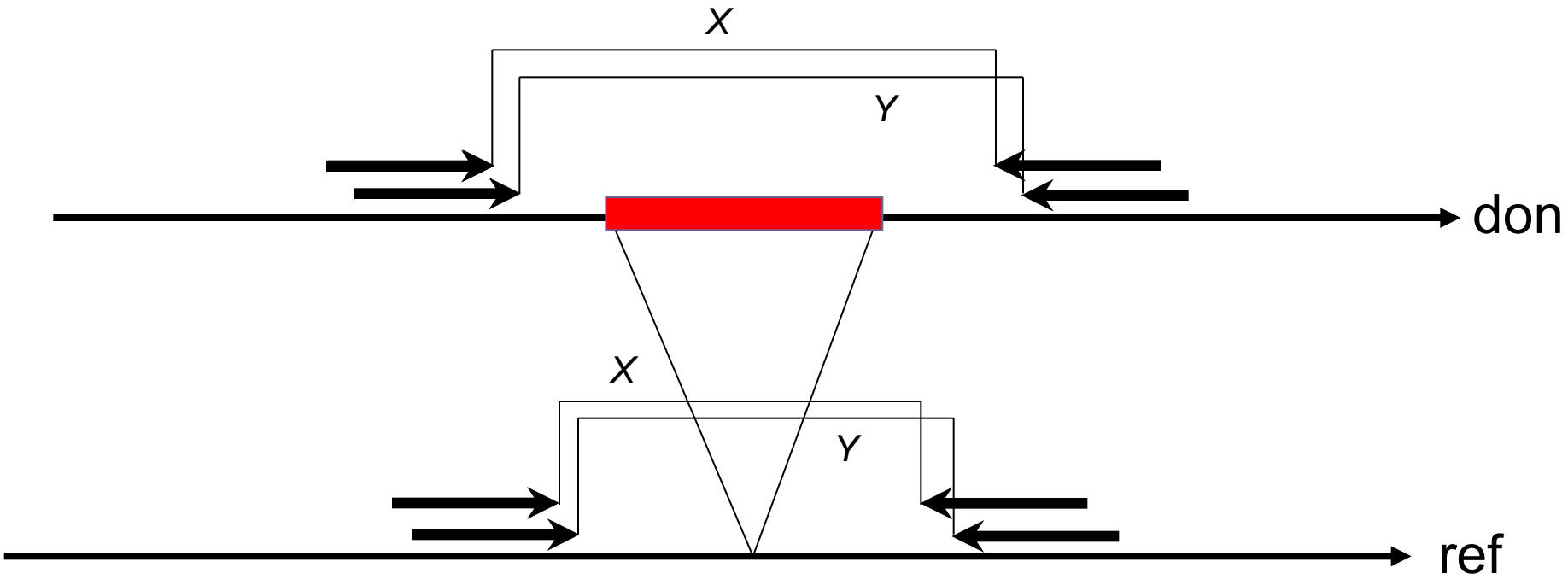
The difference insert size in the red pairs vs the black pairs informs about the additional DNA content.

PE Detection Accuracy

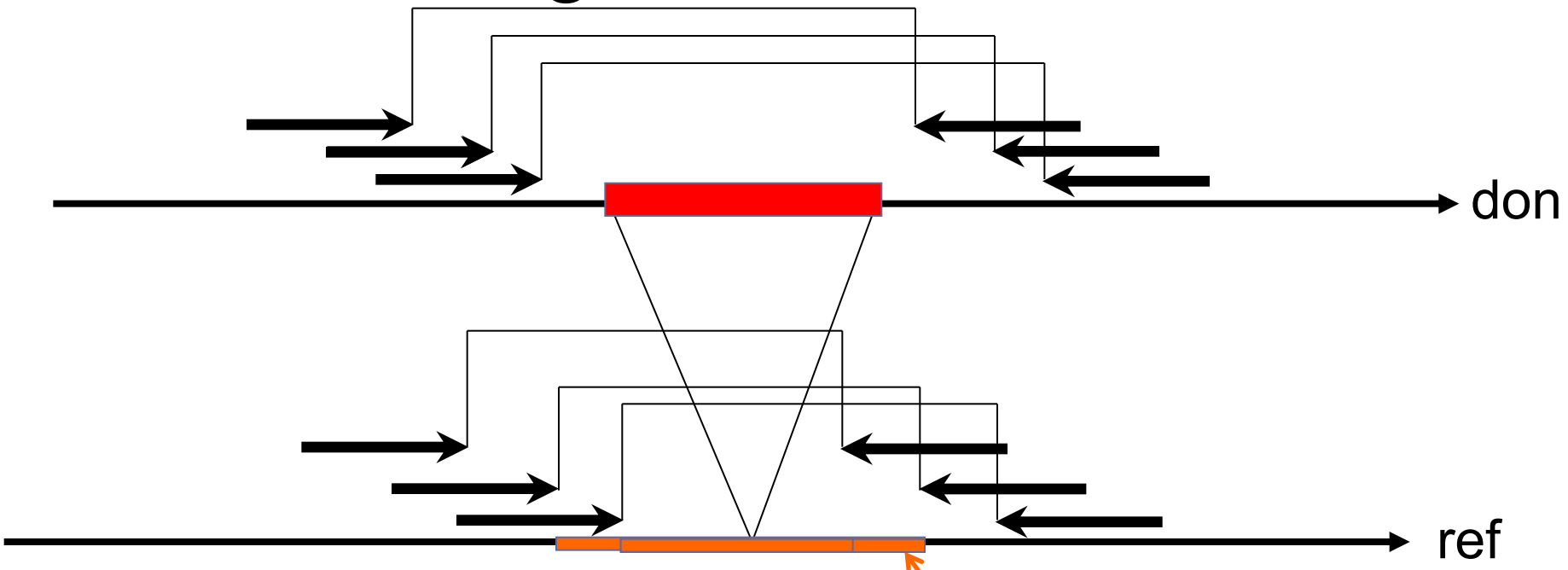
- We have observed how patterns of anomalously mapping read pairs indicate rearrangements.
- Where is the exact location of the event?
  - Paired-end reads only indicate SVs if the two reads map on either side of the rearrangement.
  - This only tells us that the event is somewhere in between the two reads.
- What is the exact size of the event (for insertions, deletions, inversions, or tandem duplications)?
  - Approximately known, but still inexact, insert size prohibits single-nucleotide precision: We don't know the exact insert size, but only an expected, average value.

Insertion: Consistency

1. Clustering
2. Size of insertion explained by *X* == size of insertion explained by Y.
3. It is important to analyse the consistency of anomalous insert sizes within the cluster.



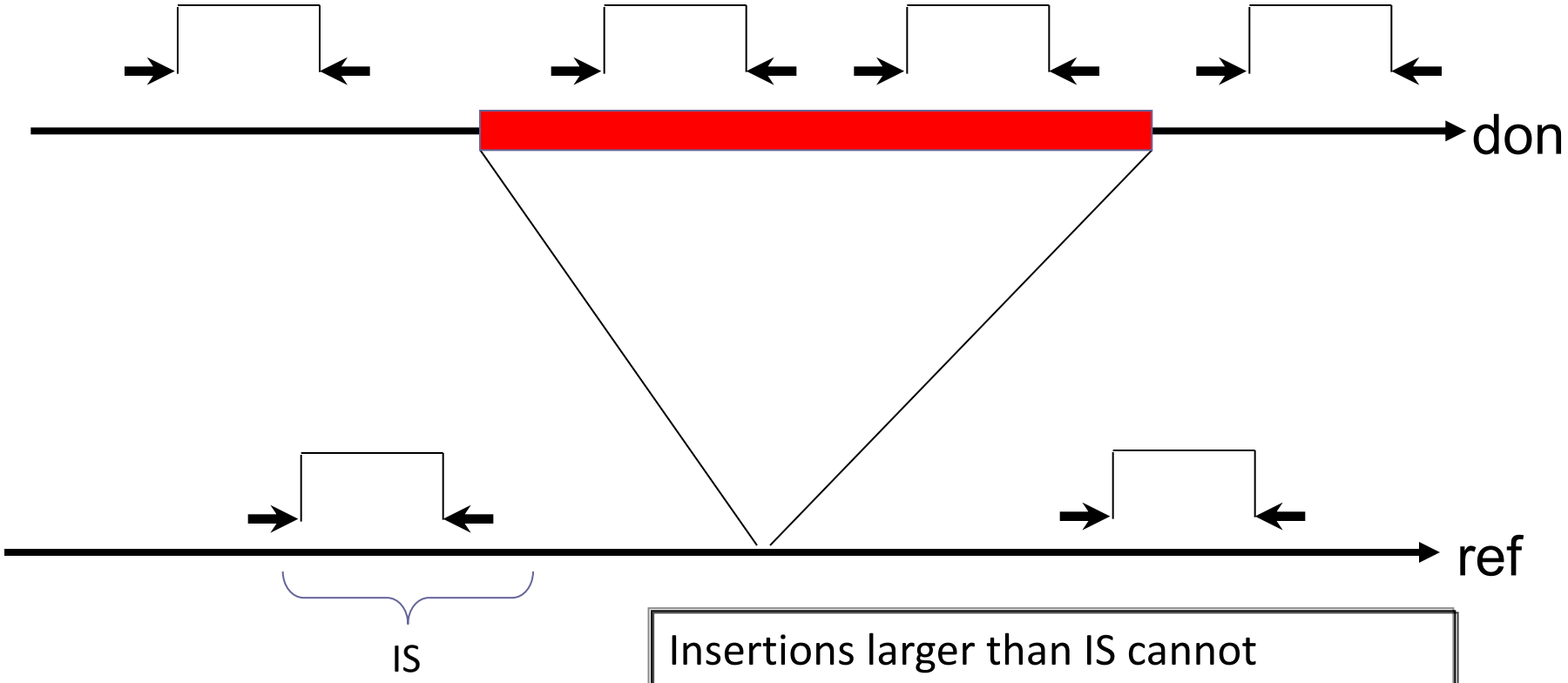
Insertion: Narrowing Down the Location



- Insertion lies within spanning region of matepair
- For clusters, it lies within the intersection

Possible location of insertion

Where can we go wrong: missed insertion



Problems with PE SV detection

- Short insert size libraries:
  - Overlapping reads due to fragments not longer than twice the read length.
    => Low insert coverage
- High variance insert size libraries:
  - To get more yield out of low DNA samples, the library cannot be cut very thinly.
    => Diminishing sensitivity to small event sizes.
    => Lowered confidence in event calls.
- Non-normal insert size distributions:
  - Algorithms assuming normality of insert sizes are likely to fail on the data.
- AND: Repeats!

PE SV Detection - Summary

- Better resolution than read depth methods (dictated by the variance of the insert size distribution), usually 10s of nucleotides.
- Can detect copy number neutral events, such as inversions.
- Delivers association between loci in the genome (compared to read depth's locus independent copy number changes).