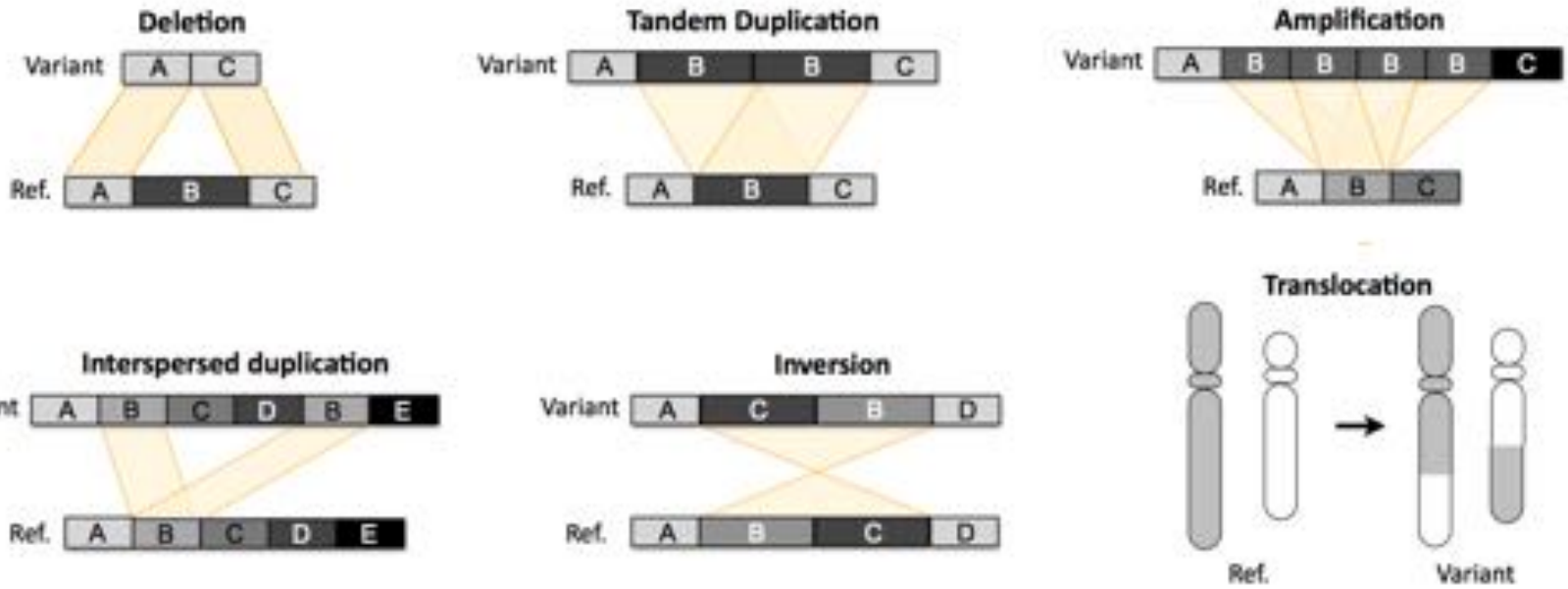
Structural variation

Advanced Sequencing Course
CSHL 2015

Maria Nattestad

What is structural variation?

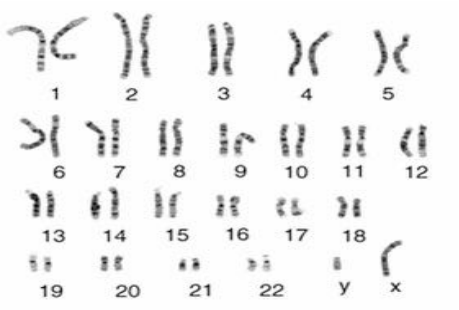
Difference in copy number, orientation, or location of any genomic sequence over 50 bp in size



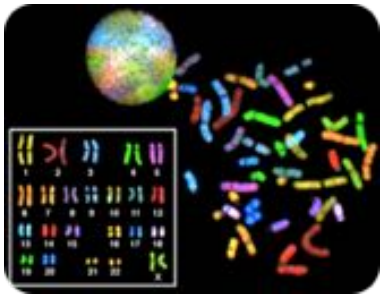
Why are structural variants important?

- They are common and affect a large fraction of the genome
  - Any two humans differ by 3,000 - 10,000 SVs
  - In total, they impact more base pairs than all single-nucleotide differences.
- They are a major driver of genome evolution
  - Speciation can be driven by rapid changes in genome architecture
  - Genome instability and aneuploidy: hallmarks of solid tumor genomes
- Genetic basis of traits
  - Gene dosage effects.
  - Neuropsychiatric disease (e.g., autism, schizophrenia)
  - Spontaneous SVs implicated in so-called “genomic” and developmental disorders
  - Common SNPs are not proxies for all forms of variation.
  - Somatic genome instability; age-dependent disease

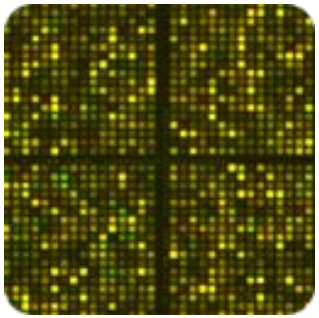
Our understanding is driven by technology



1940s - 1980s
Cytogenetics / Karyotyping



1990s
CGH / FISH /
SKY / COBRA



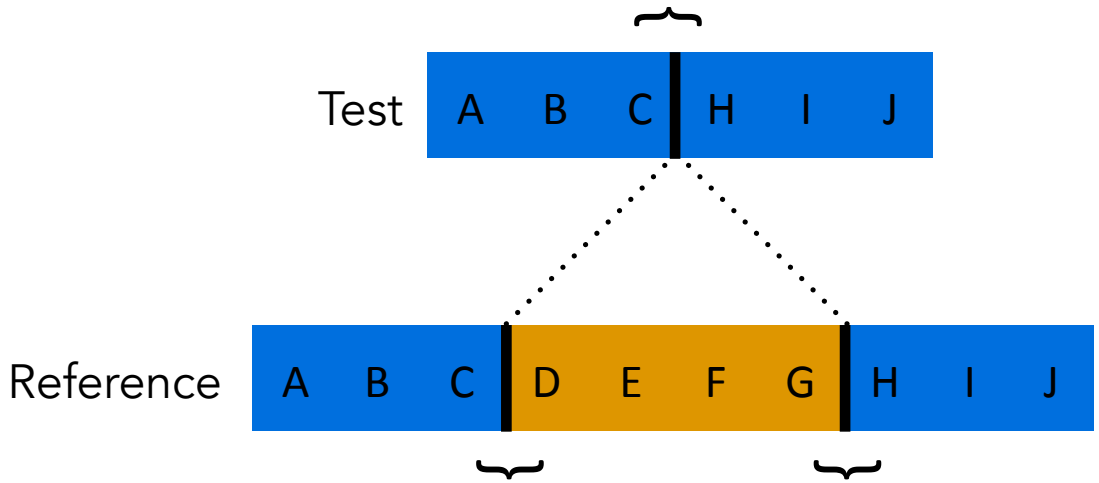
2000s
Genomic microarrays
BAC-aCGH / oligo-aCGH

Today
High throughput
DNA sequencing

SV breakpoints defined

Breakpoints are the junctions that define structurally variable genomic segments

Breakpoint is an ambiguous term because it simultaneously describes one junction in the test genome, and two junctions in the reference genome. Not surprisingly, this leads to confusion.

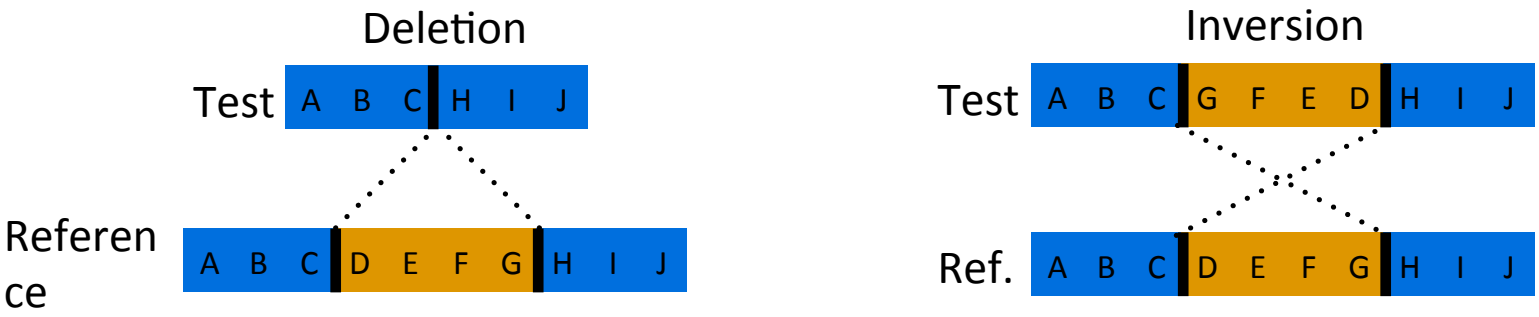


The VCF file format accounts for this ambiguity by introducing two new terms:

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

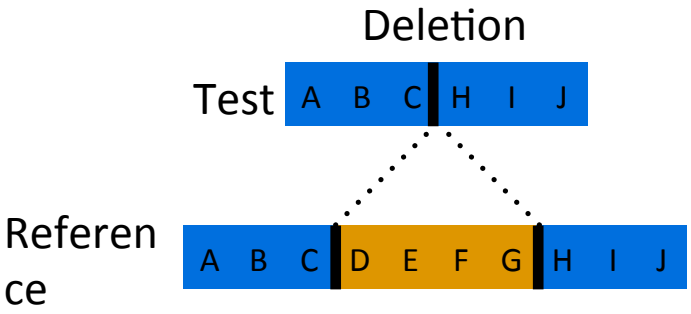
breakends: the two breakpoints in the reference genome

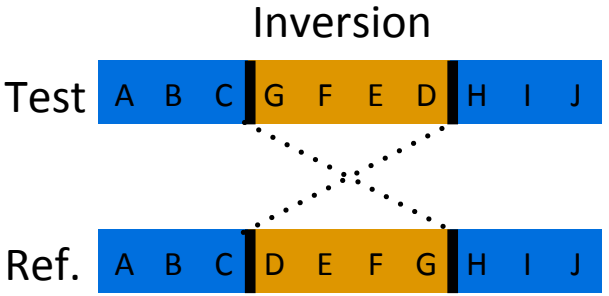
Visualizing SV breakpoints

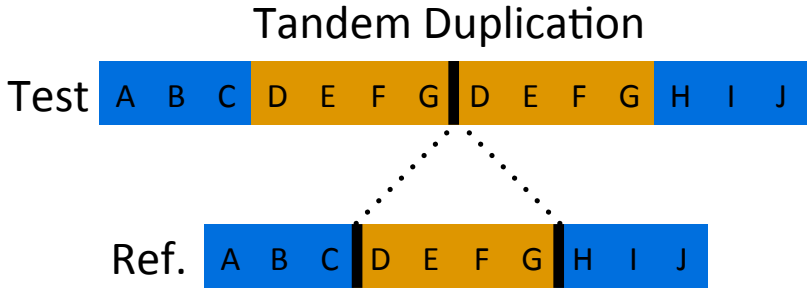


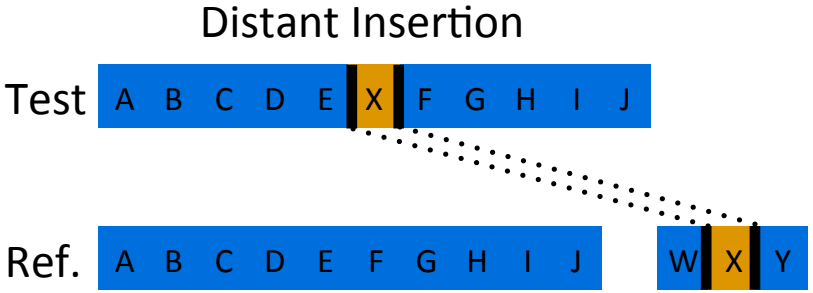
<u>NOTE</u>: Deletions produce one breakpoint in the test genome and two in the reference, whereas inversions produce two breakpoints in both genomes.

Visualizing SV breakpoints

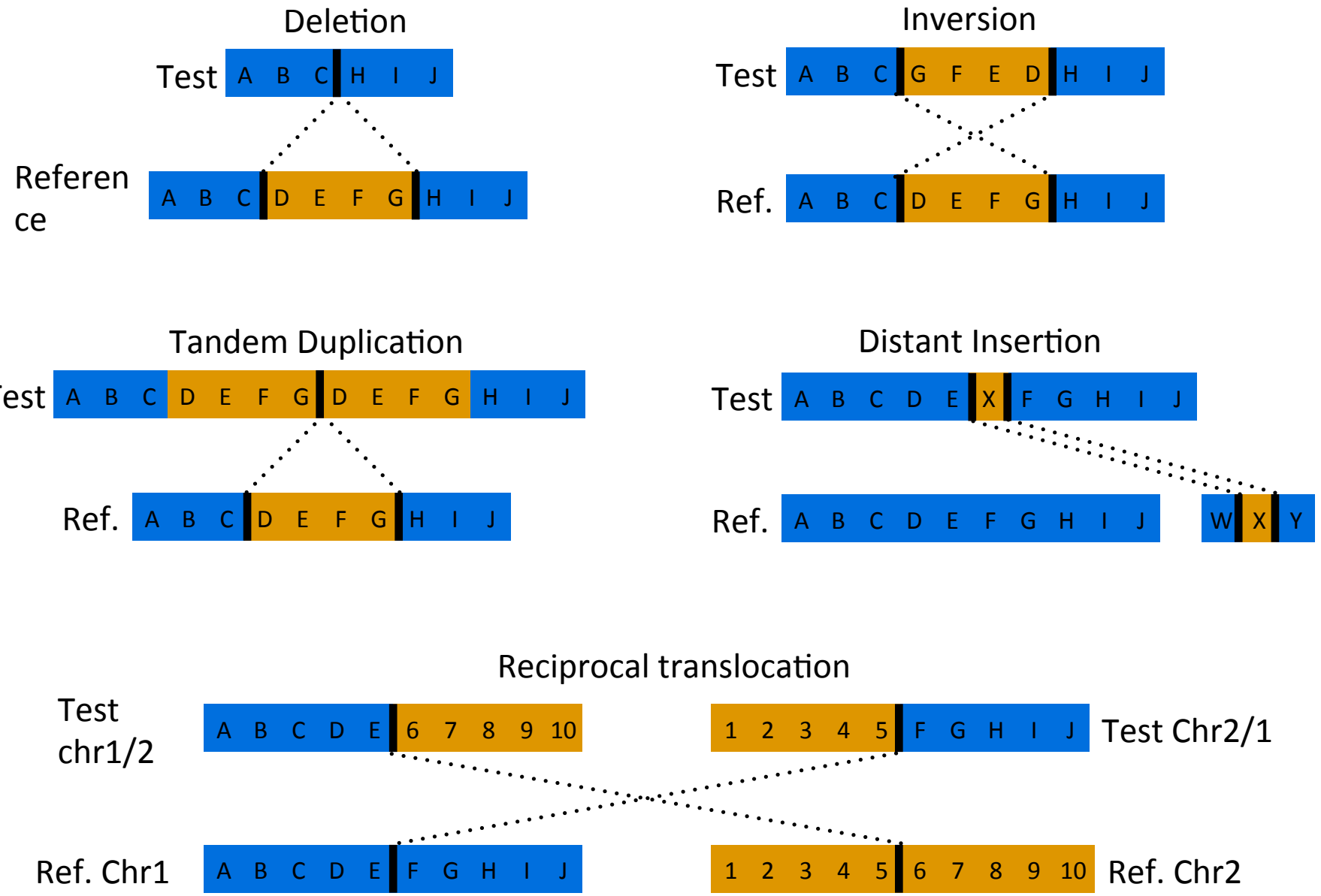






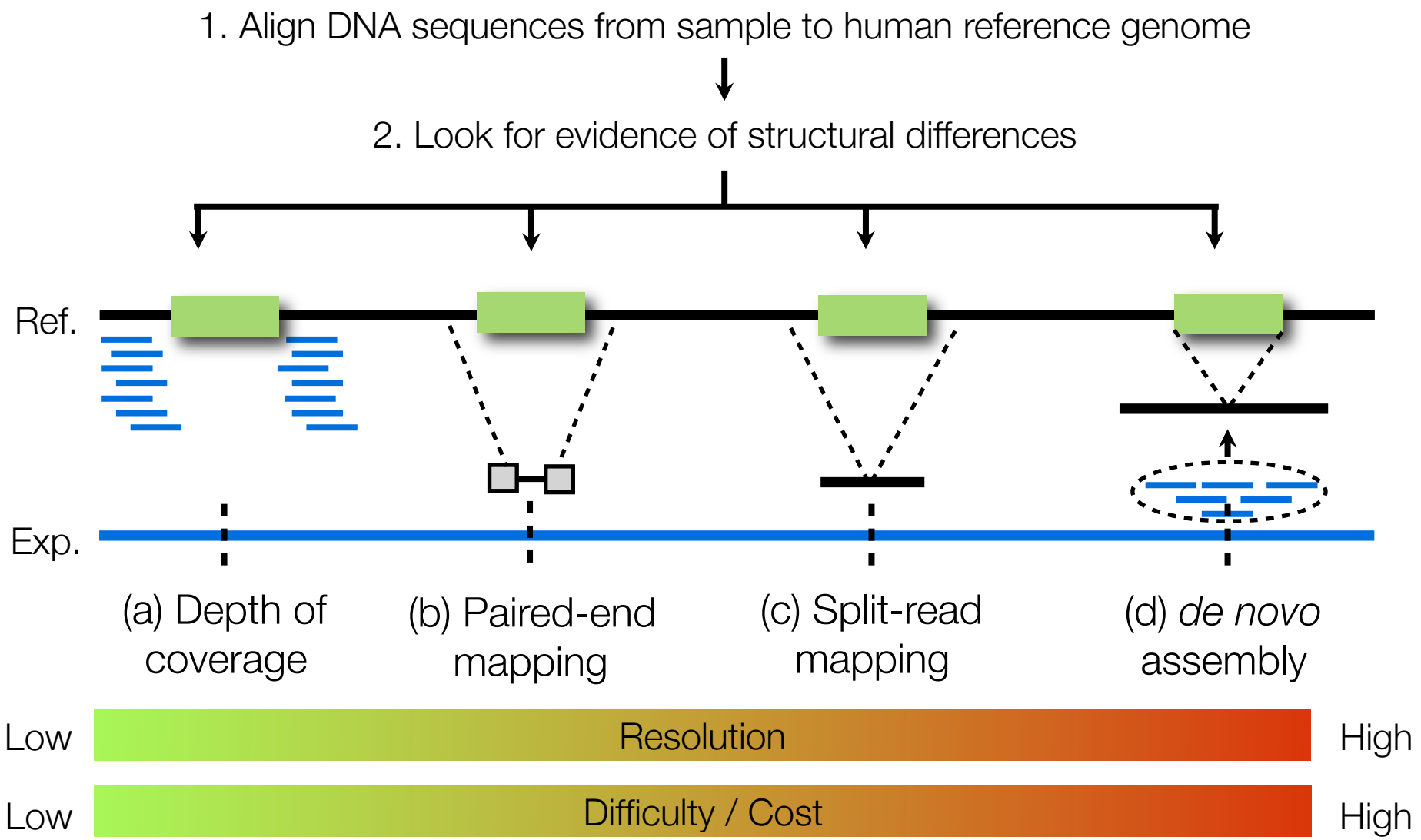


Visualizing SV breakpoints



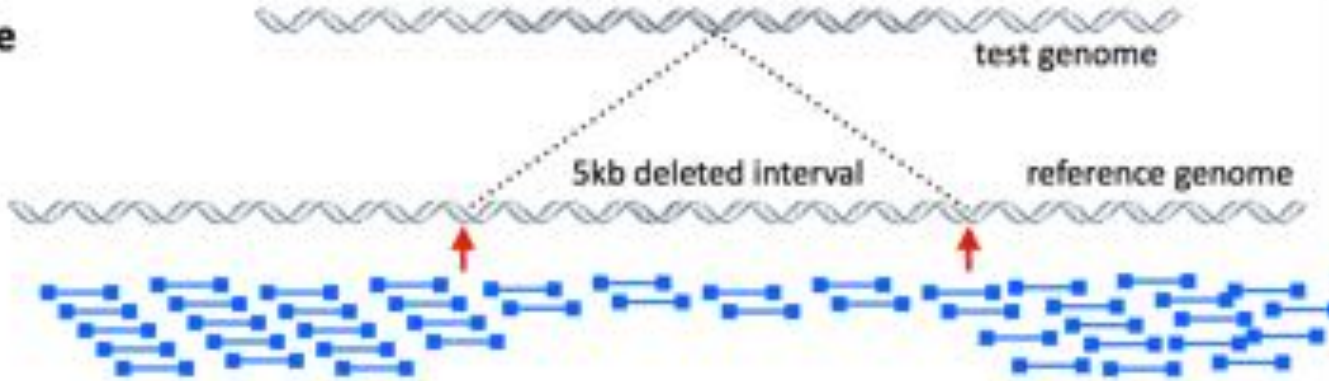
A brief primer of structural variation

SV discovery with modern sequencing

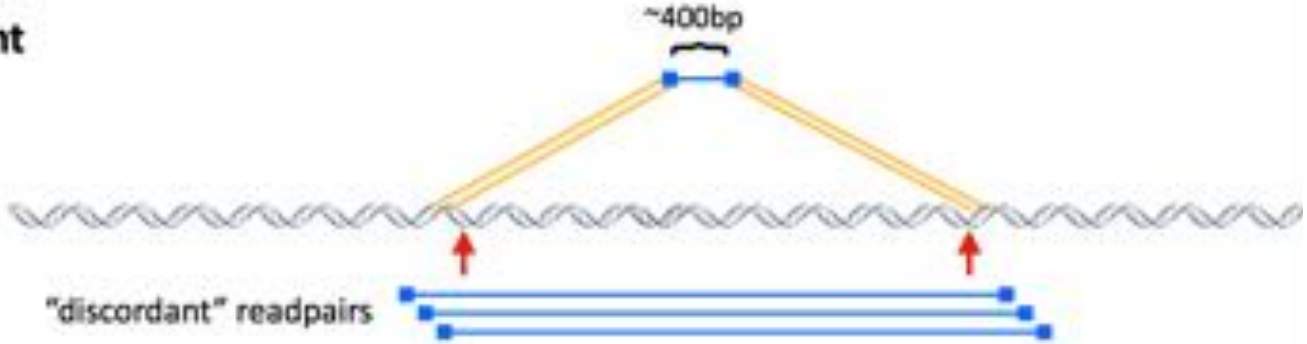


3 ways to detect a structural variant (SV)

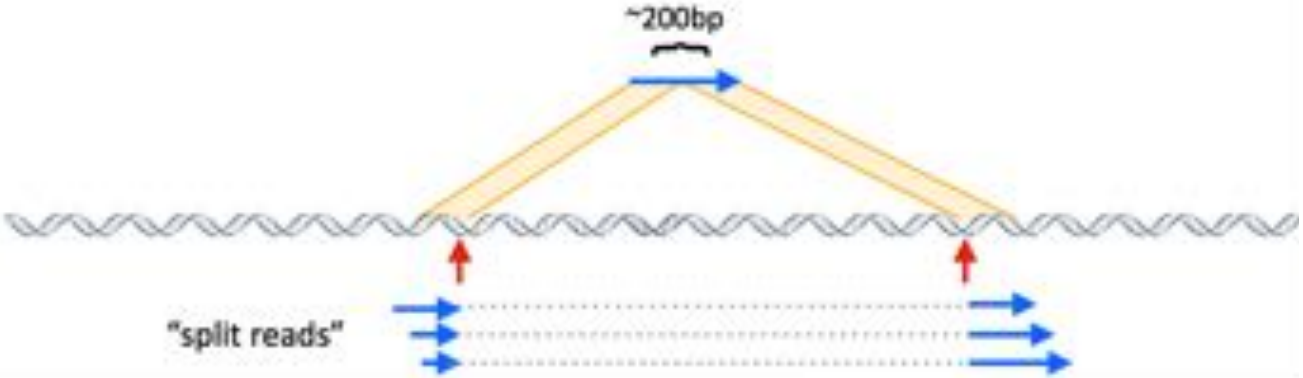
1) depth of sequence coverage
= "read-depth analysis"
(copy number alterations)



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



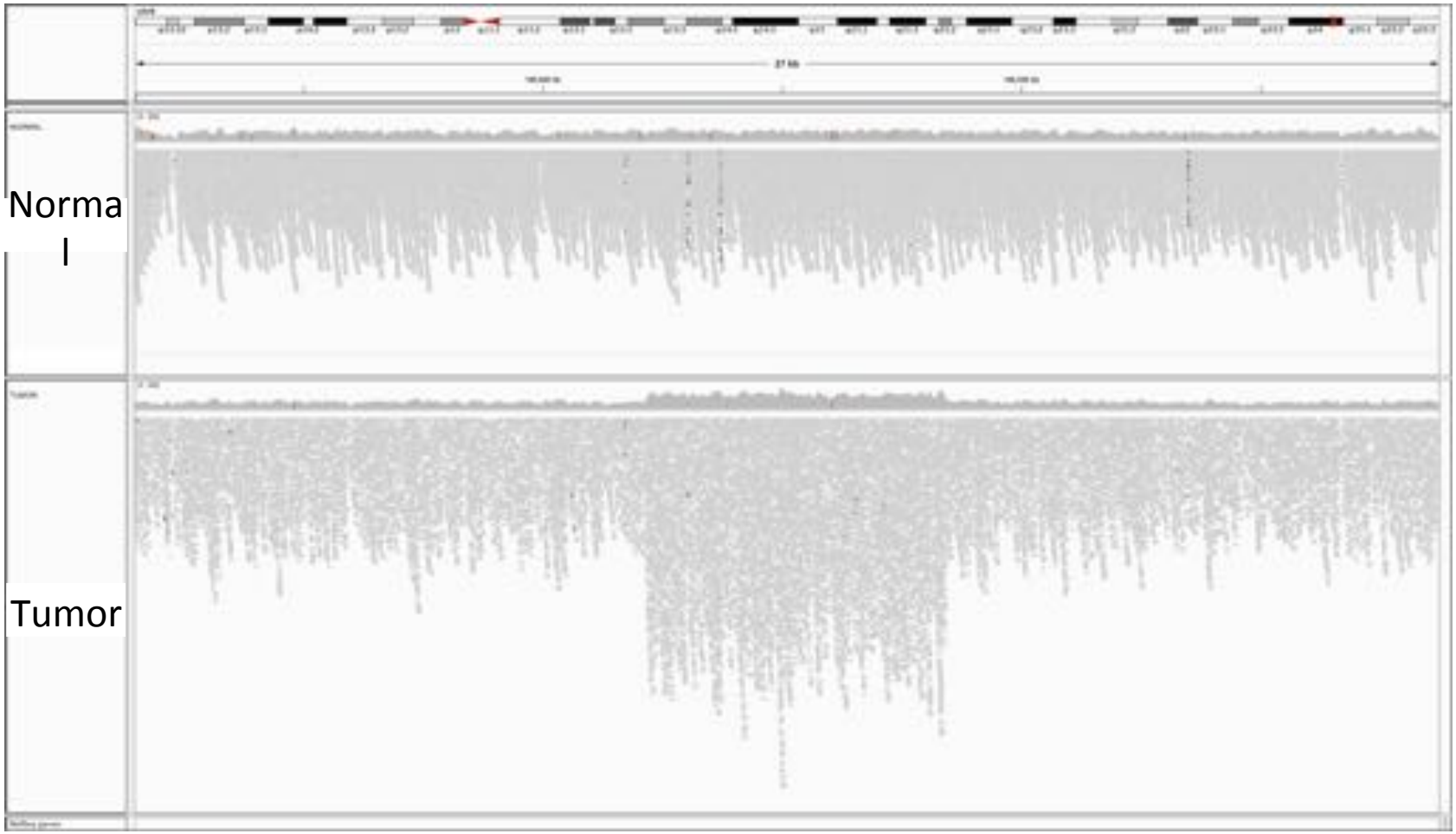
3) read contains a breakpoint
= "split-read mapping"
(all classes of SV)



SV-calling method 1: Read-depth analysis

Copy number variants only

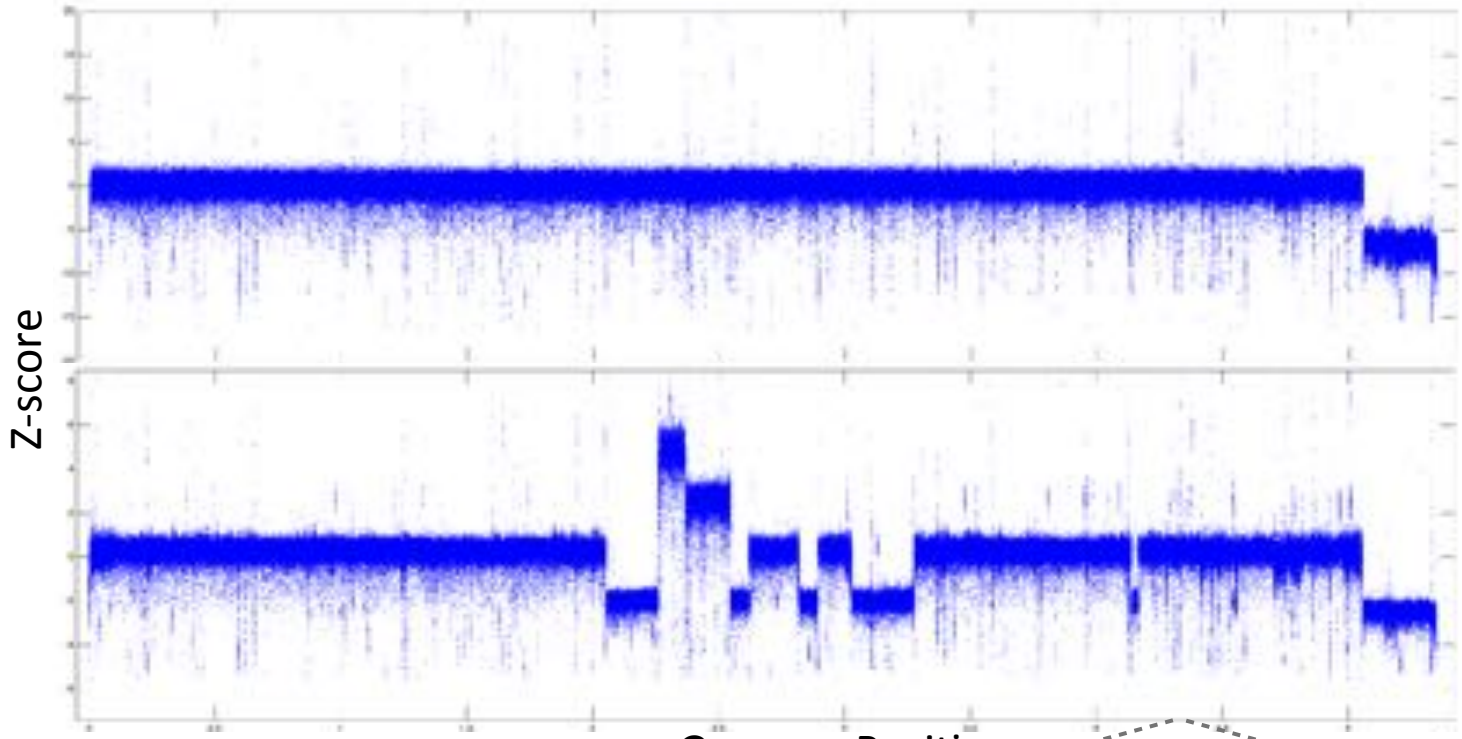
Detecting CNAs with read-depth analysis



Basic approach:

1) Count reads in sliding windows (e.g., 5kb) across the genome.
2) Normalize for GC bias.
3) Use segmentation to define CNAs (similar to array-CGH data).

Detecting CNAs with read-depth analysis



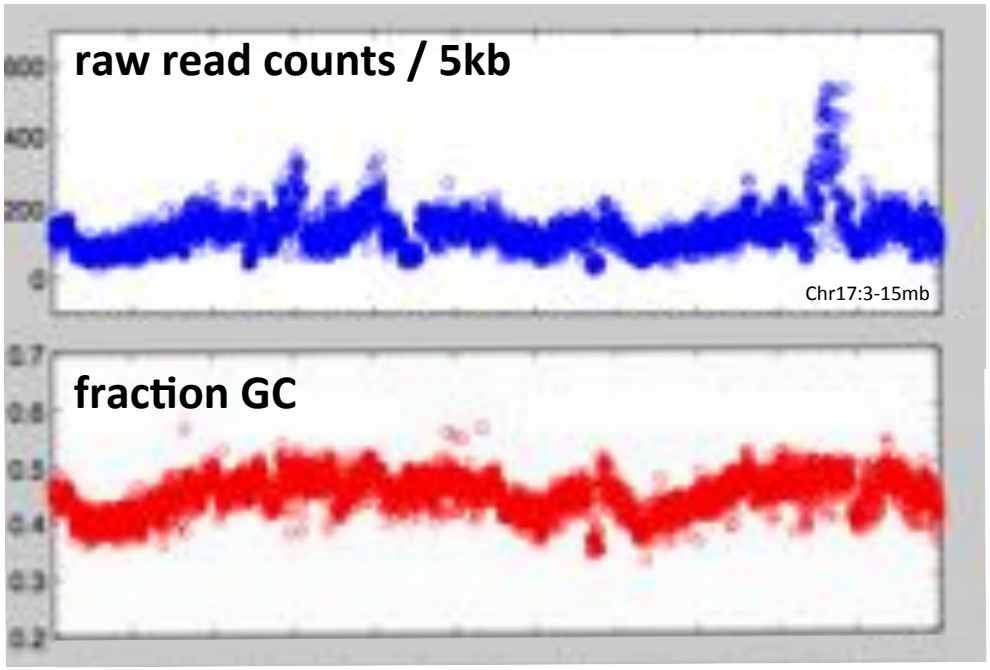
Strengths:

1) Fast and simple.
2) Easy to identify gene amplifications.
3) Relatively straightforward interpretation: is gene X amplified or deleted?

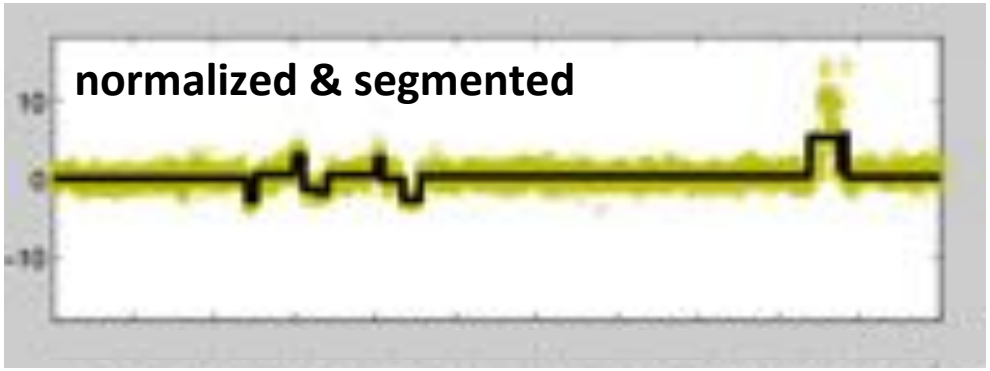
Weaknesses:

1) Limited resolution (5-10kb) = imprecise boundaries
2) Cannot detect balanced events or reveal variant architecture.

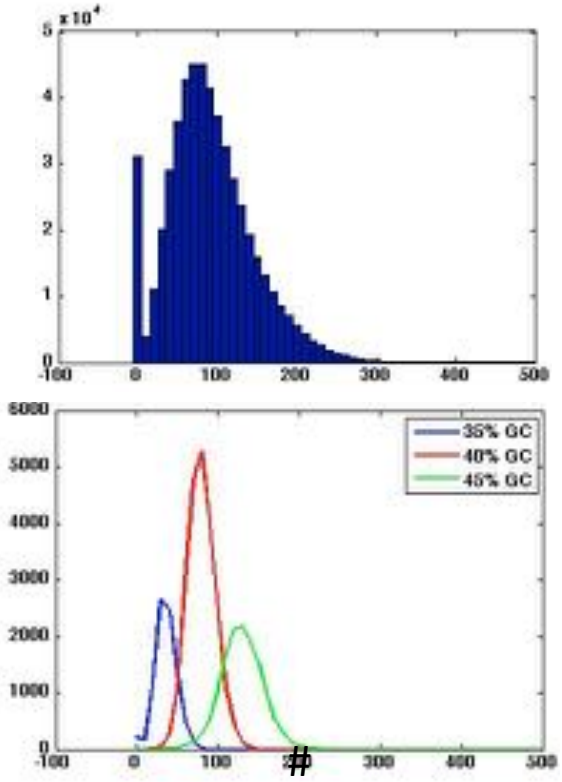
Normalization of read-depth data



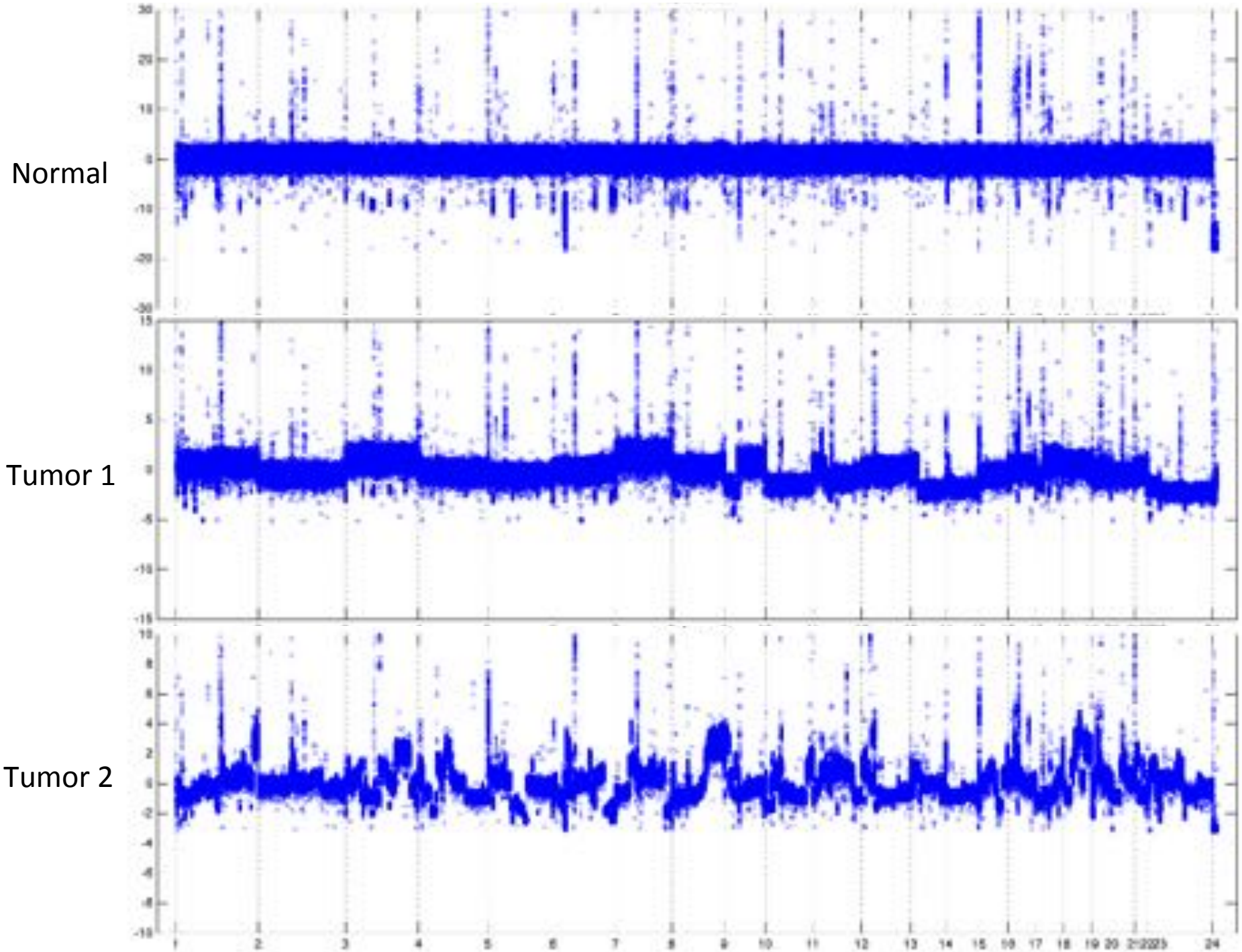
GC normalization (Z-score)
Copy number segmentation



Coverage (5kb windows)

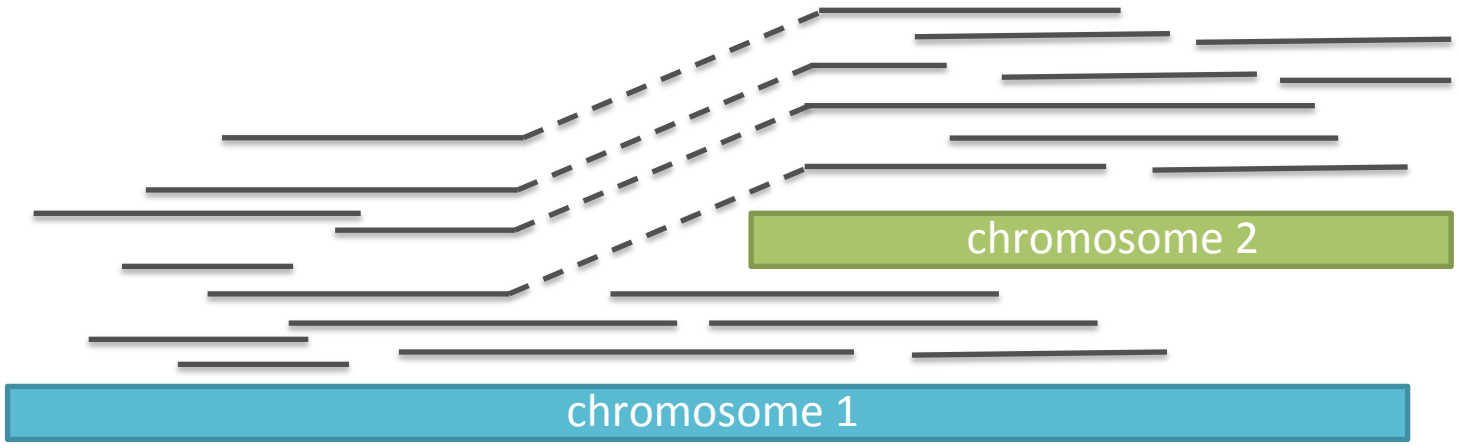


- **Use variably-sized windows, masked for repeats**
  - repeatMasker, SSRs, "mapability"
- **Window size should yield >100 reads (median)**
- **With all alignments, absolute copy number can be discerned (Studmant et al., 2011)**
- **Publicly available tools:**
  - ReadDepth, RDXplorer, cnvSeq, CNVer, CopySeq, GenomeSTRiP, CNVnator

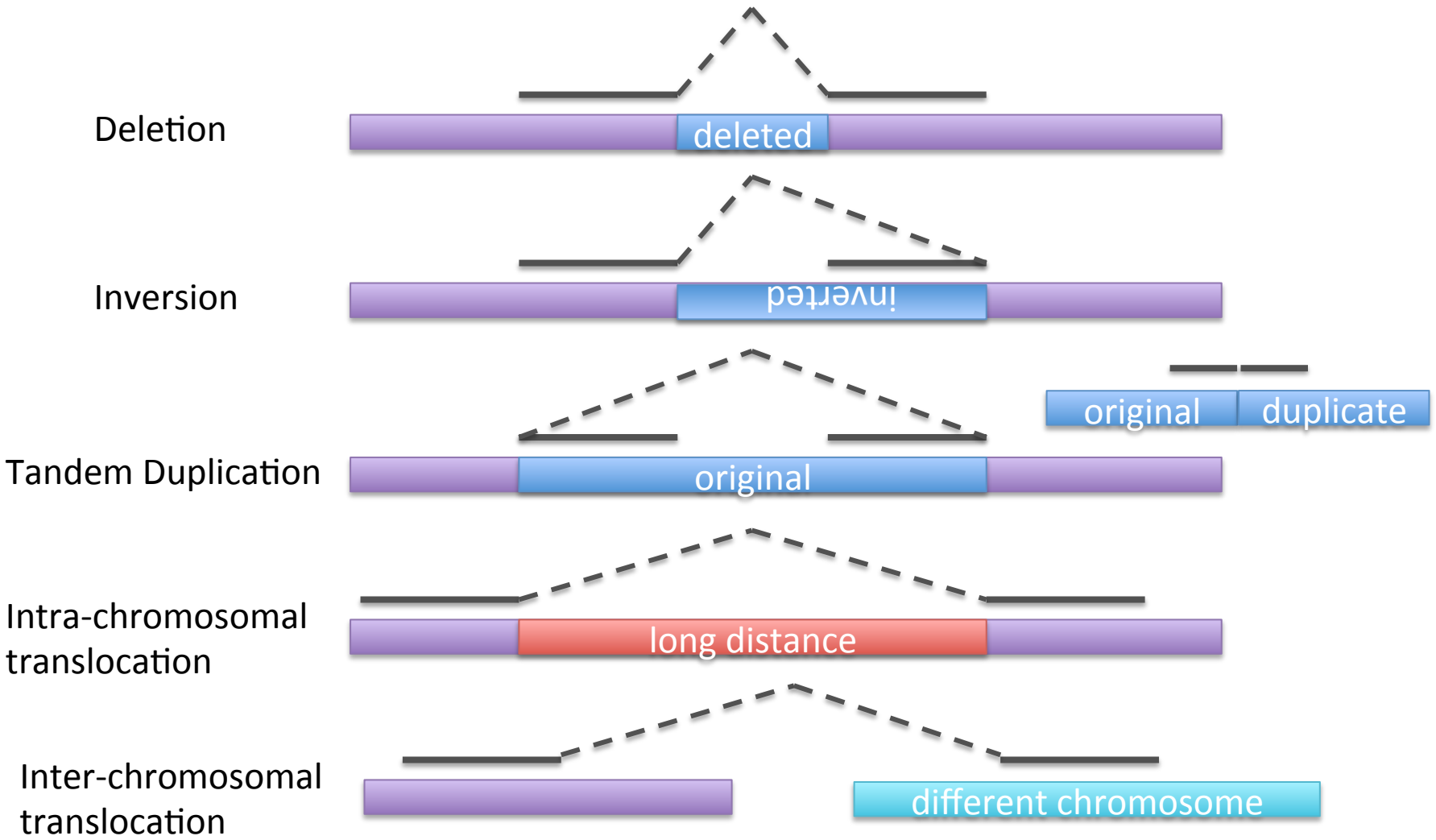

Normal
Tumor 1
Tumor 2

SV-calling method 2: Split-read mapping

An interchromosomal translocation found by 4 split reads

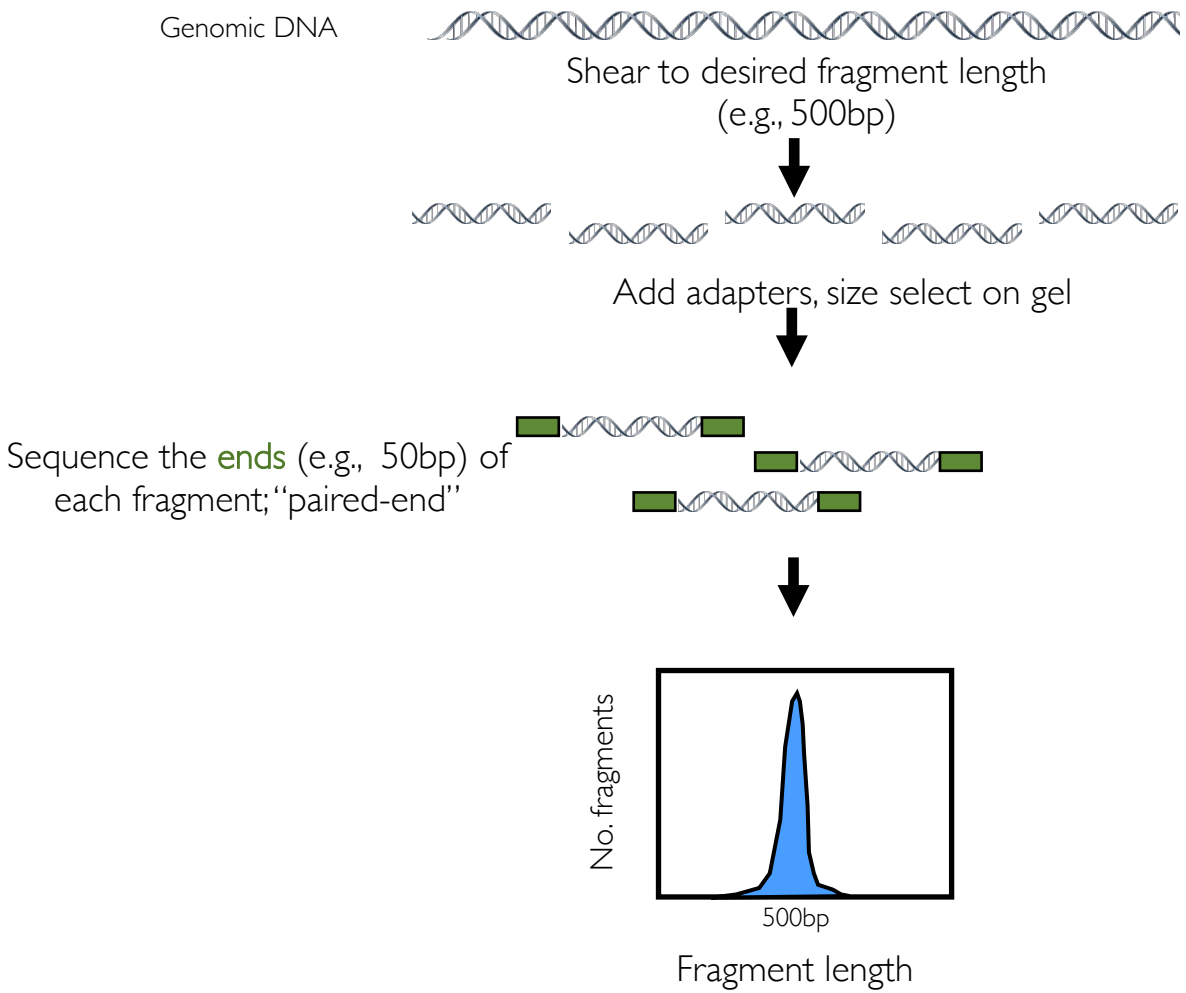


Detecting structural variants with split-read mapping

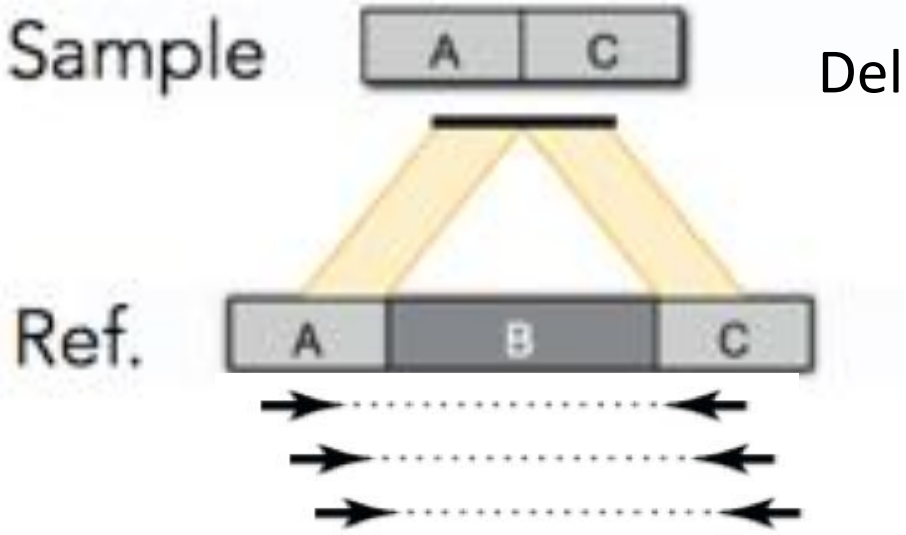


SV-calling method 3: Paired-end mapping

Paired-end sequencing



Paired-end reads for variant calling



Compared to split reads:

Same: Paired-end reads can detect same types as split reads.

Better: Longer fragments can span across some small repeats.

Worse: Cannot narrow down the breakpoint to base-pair level without also using split reads.

SV-calling method 4: Assembly-based variant calling

Assembly-based variant calling

It takes a lot more compute power to do an assembly, so this approach is rarely used for large projects with many samples

The dirty secrets of SV discovery

Secret #1.
Often many false positives

- Short reads + heuristic alignment + repetitive genome
  = systematic alignment artifacts (false calls)
- Chimeras and duplicate molecules
- Ref. genome errors (e.g., gaps, mis-assemblies)
- **ALL** SV mapping studies use strict filters for above

Secret #2.

The false negative rate is also generally very high.

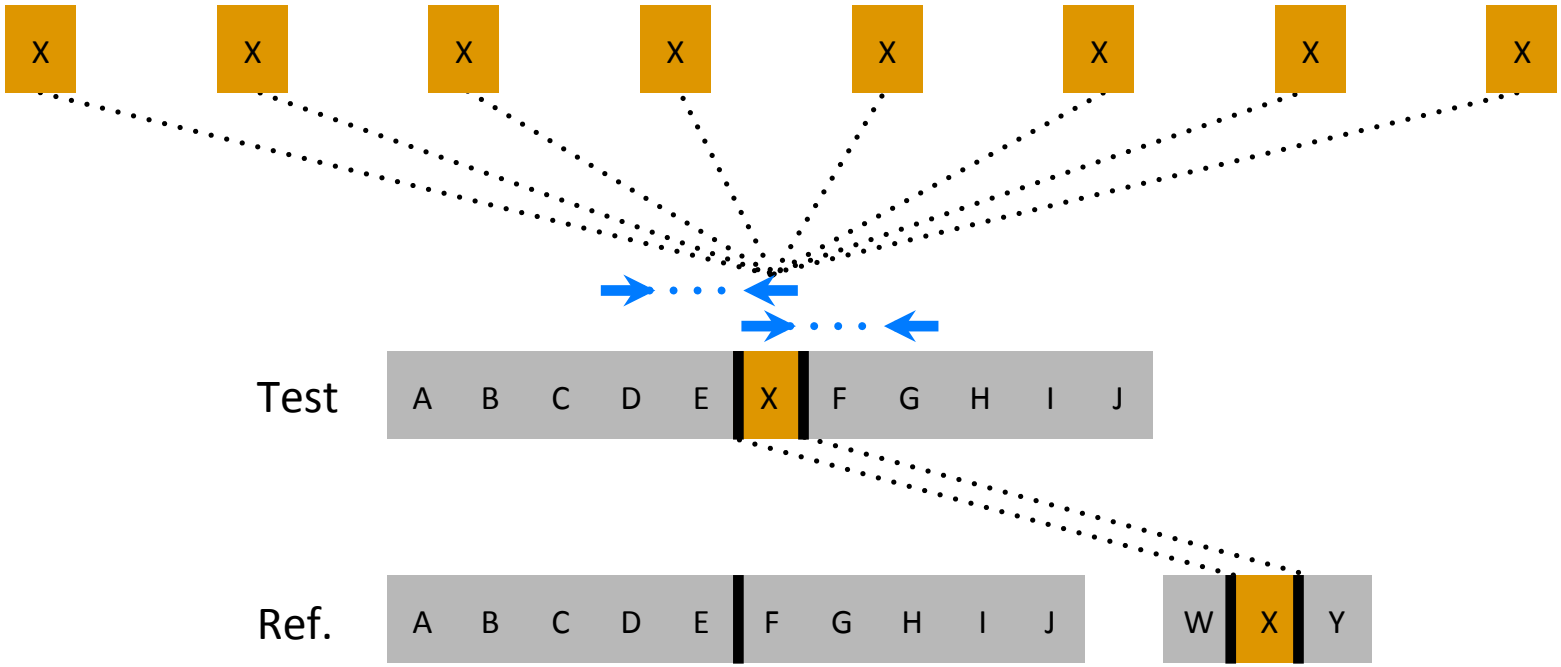
- Most current datasets have low to moderate physical coverage due to small insert size (~10-20X)
- Breakpoints are enriched in repetitive genomic regions that pose problems for sensitive read alignment
- FILTERING!
- The false negative rate is usually hard to measure, but is thought to be extremely high for most PEM studies (>30%)
- When searching for spontaneous mutations in a family or a tumor/normal comparison, a false negative call in one sample can be a false positive somatic or de novo call in another.

Secret #3.

SVs involving repetitive sequence are very difficult to detect.

- Many variants involve repetitive sequence such as transposons and segmental duplications.
- The reads that define these breakpoints will align to multiple loci. To cluster these properly, one must keep track of all mappings, not just one.
- This is a complicated analysis. To our knowledge, only two algorithms can do this: Hydra and VariationHunter

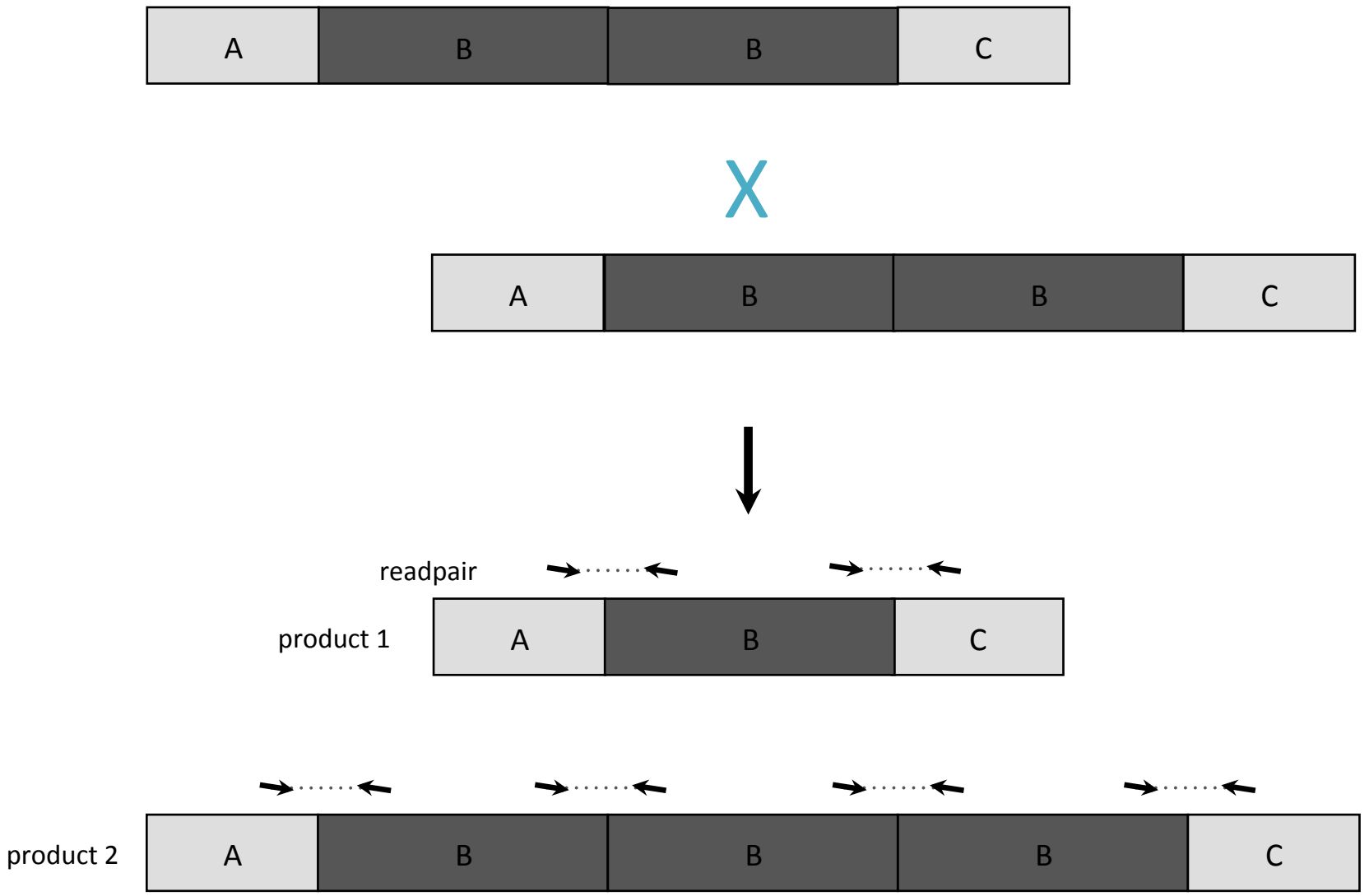
A challenging case: repetitive elements



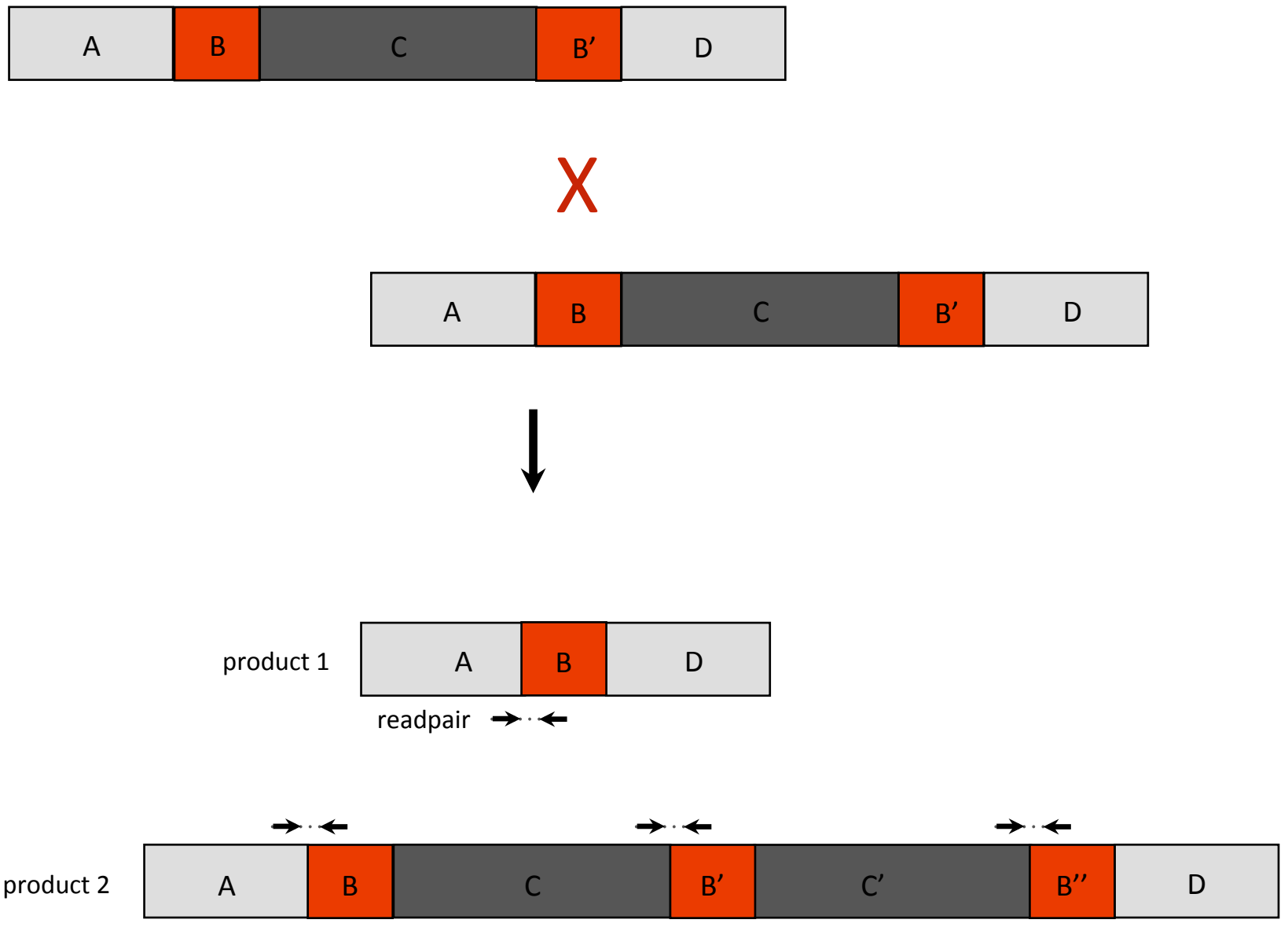
Secret #4.

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

NAHR within a tandem array



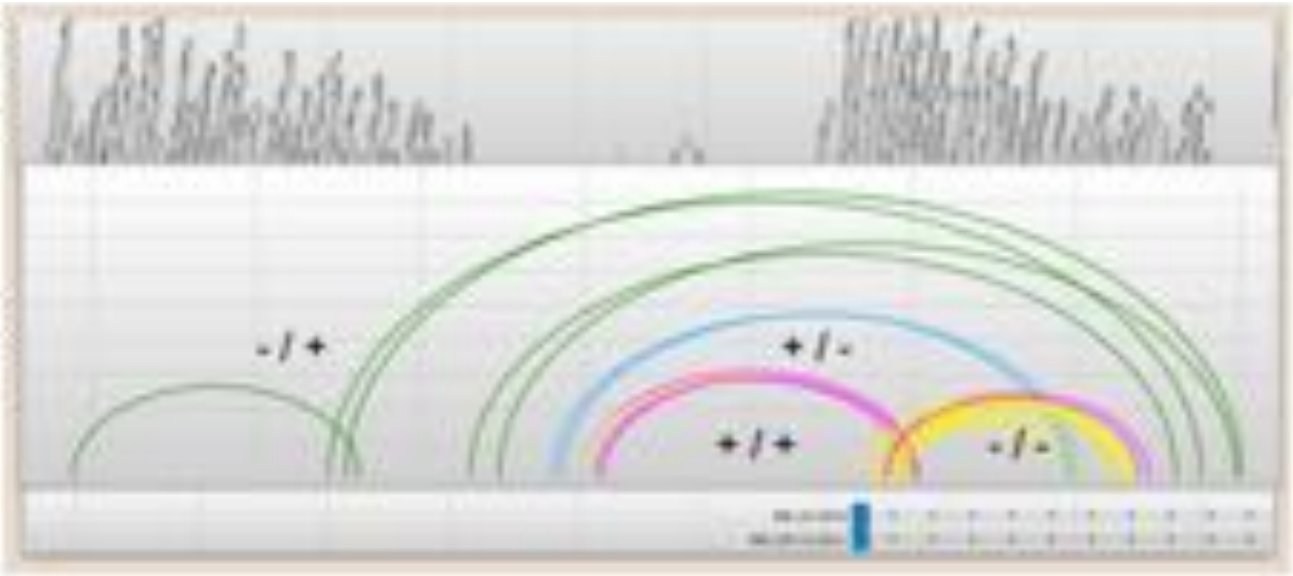
NAHR between large flanking repeats



Secret #5.

Complex structural variants generate confusing breakpoint patterns

Complex variants



Quinlan et al., 2010 (Genome Research)

Tools

- Copy number variants using read depth: CNVnator, Ginkgo***: qb.cshl.edu/ginkgo

- Split-read or paired-end reads from Illumina sequencing: DELLY, PINDEL, GASV-PRO, LUMPY

- Split-read and within-read variants from long-read sequencing: Coming soon