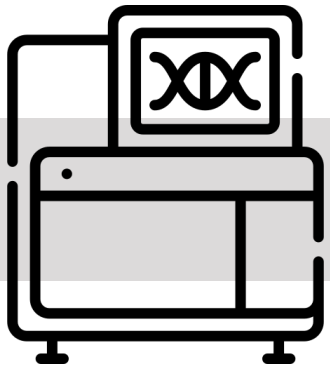
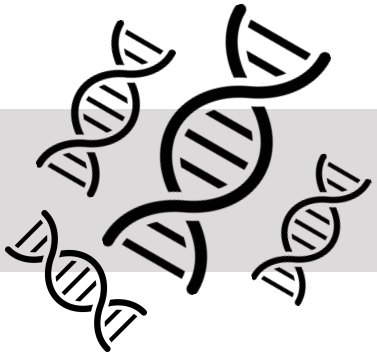
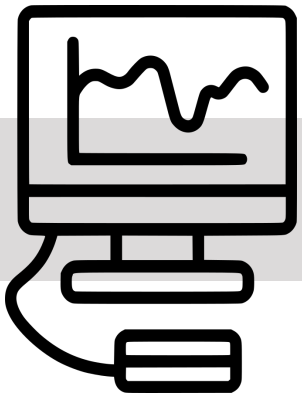
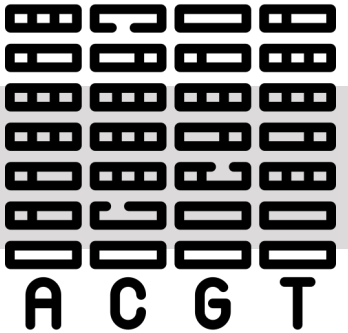
Indels and Structural variants









Learning objectives

1. Understand the different types of structural variants
2. Learn the different algorithms to detect structural variants in each sequencing technology
3. Annotate structural variants in VCF format
4. Visualize structural variants in IGV
5. Merge structural variants in VCF format
6. Benchmark structural variants

Types of genetic variation

Single-nucleotide polymorphisms (**SNPs**)

Single-nucleotide variants (**SNVs**)

Types of genetic variation

Single-nucleotide polymorphisms (**SNPs**)

Single-nucleotide variants (**SNVs**)

Insertion-deletion polymorphisms (**INDELs**)

Types of genetic variation

ctccgag
ctctgag

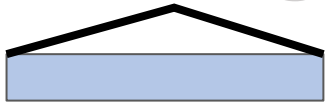
Single-nucleotide polymorphisms (**SNPs**)

Single-nucleotide variants (**SNVs**)

ctc--ag
ctctgag

Insertion-deletion polymorphisms (**INDELs**)

ctcaag
ctc ag

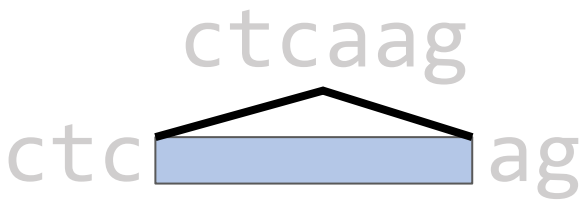


Structural variants (**SVs**)

Types of genetic variation

ctc--ag
ctctgag

Insertion-deletion polymorphisms (**INDELs**)



Structural variants (**SVs**)

Differences between Indels and SVs are often blurry. In general:

Biology

- Indels – Replication slippage
- SVs
  - Recombination issues: Nonallelic homologous recombination
  - DNA double strand break repair: Non-homologous end-joining (NHEJ)

Size (historical alignment of short-reads)

- Indels < 50bp
- SVs > 50bp

Detection

- Indels – Gaps and insertions in the alignment process, detected by variant callers
- SVs – Specialized tools to detect signals of SV (although long read mappers can produce large gaps)

Indels

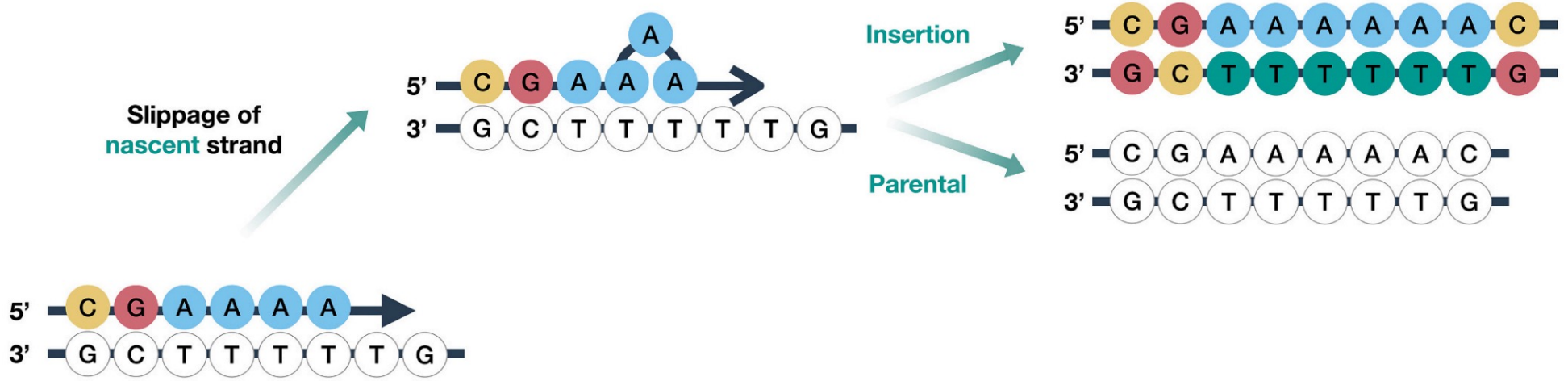
- In(sertion)Del(eletion)
  - A mutation that results from the gain or loss of a sequence

```
Reference TCCAGCAATCAGCGTCAAGCTT
Sample    TCAAGCAA---GCGTCAAGCAA
```

```
Reference TCCAGCAA   GCGTCAAGCTT
Sample    TCAAGCAA(TCA)GCGTCAAGCAA
```

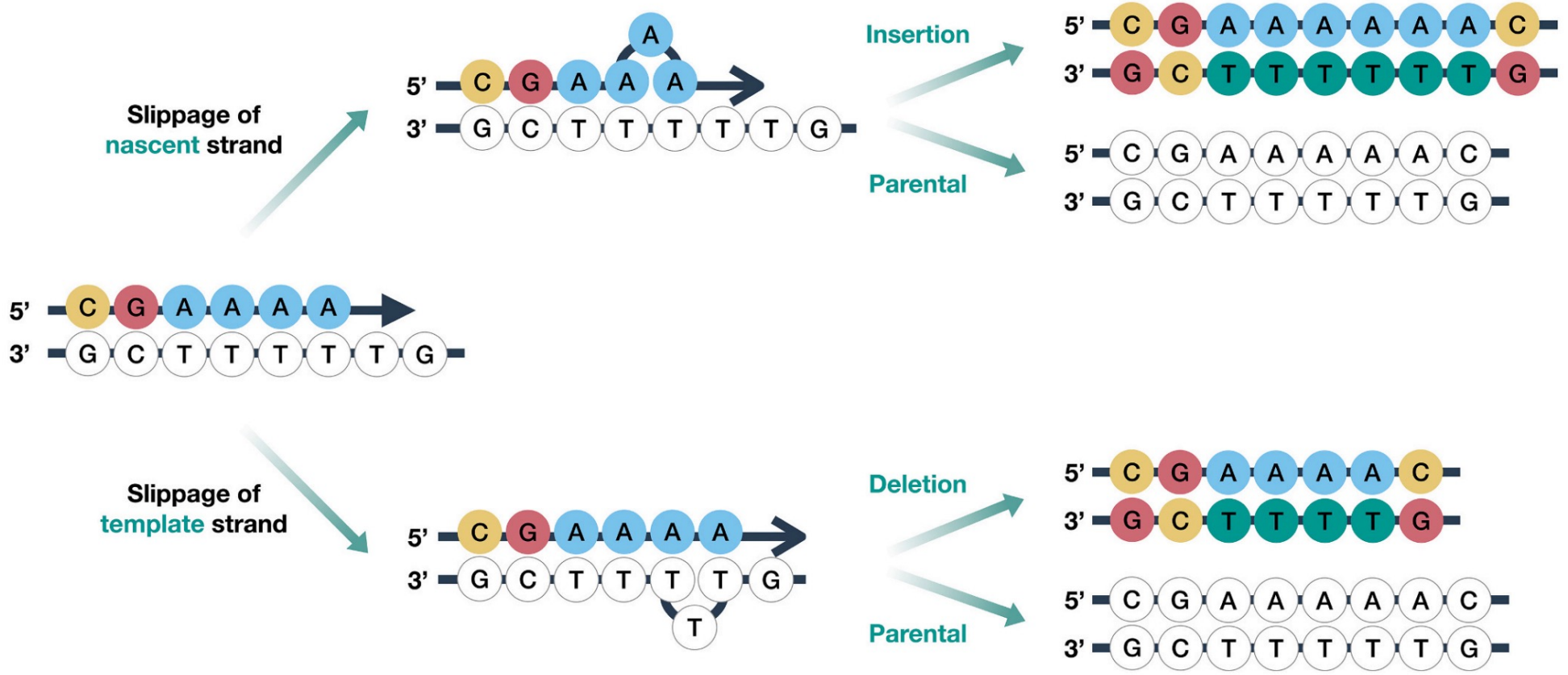
Indels

- Replication slippage



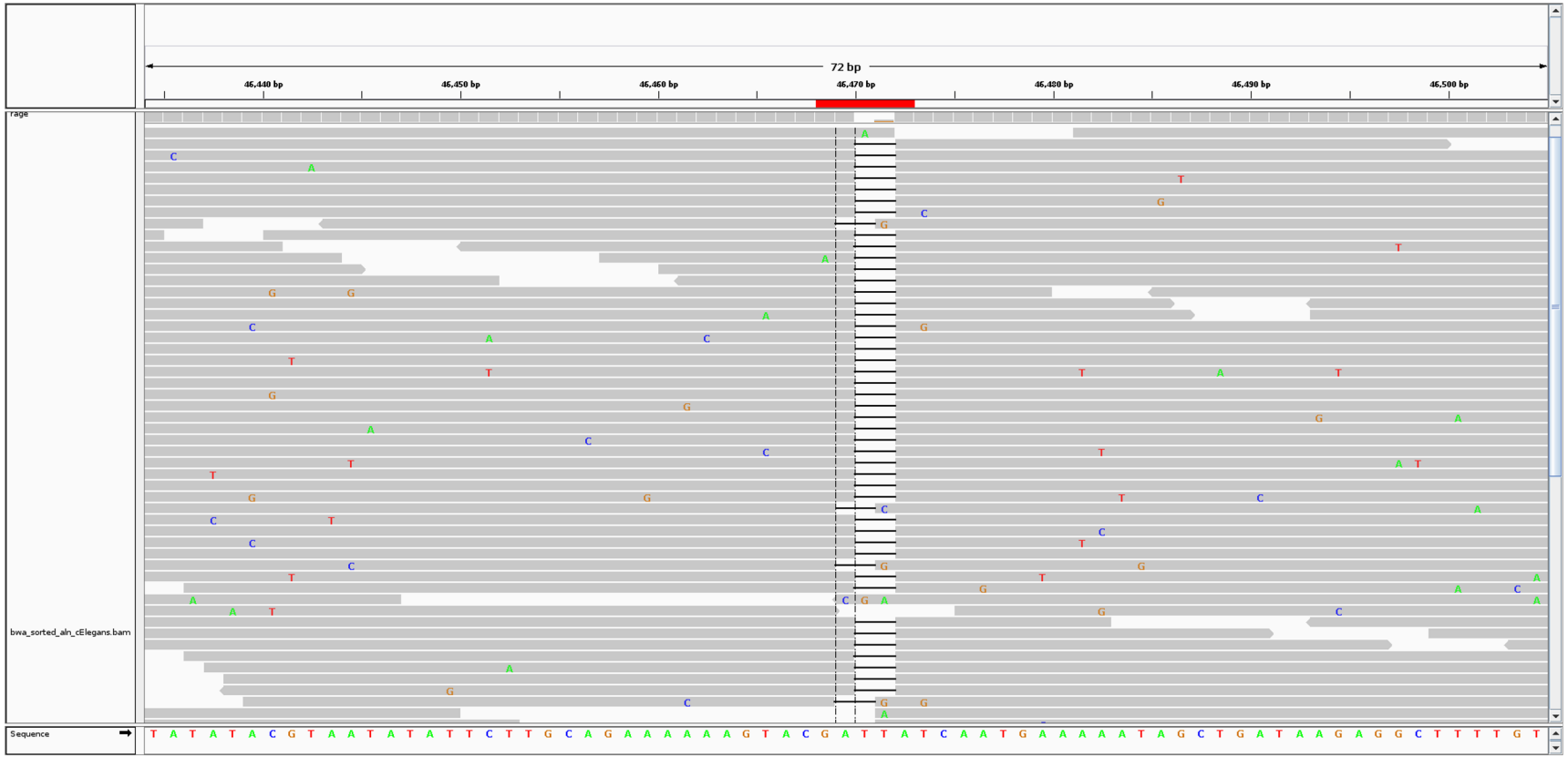
Indels

- Replication slippage



Indels

- In(sertion)Del(eletion)

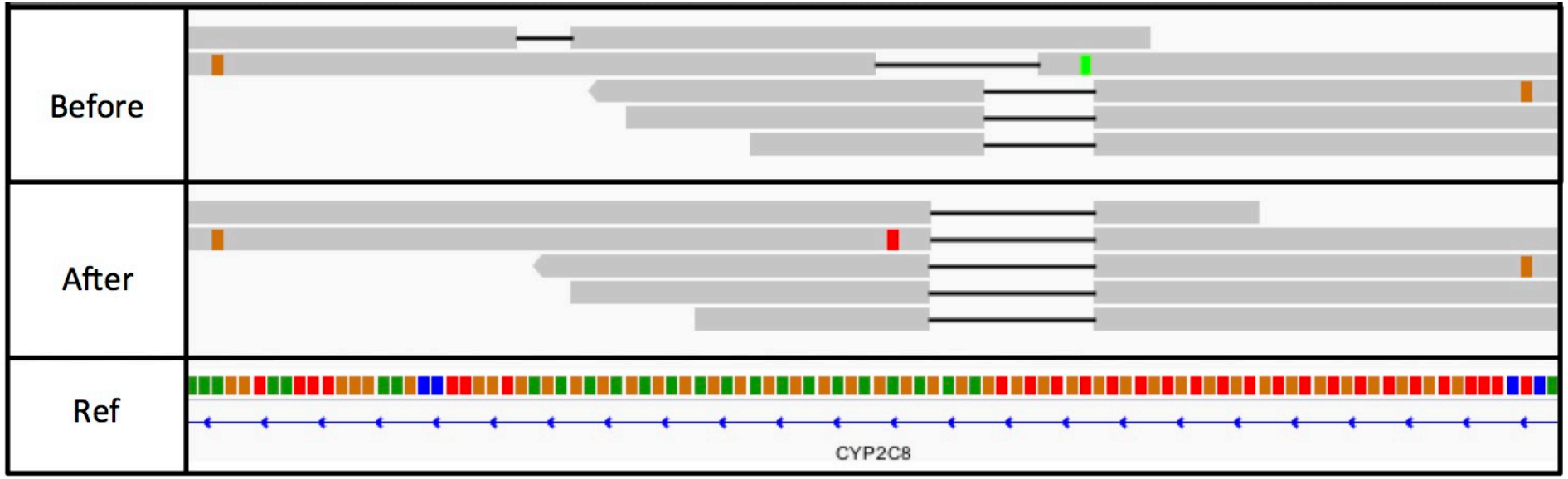


Indels

- Left alignment of the variants (bcftools norm)

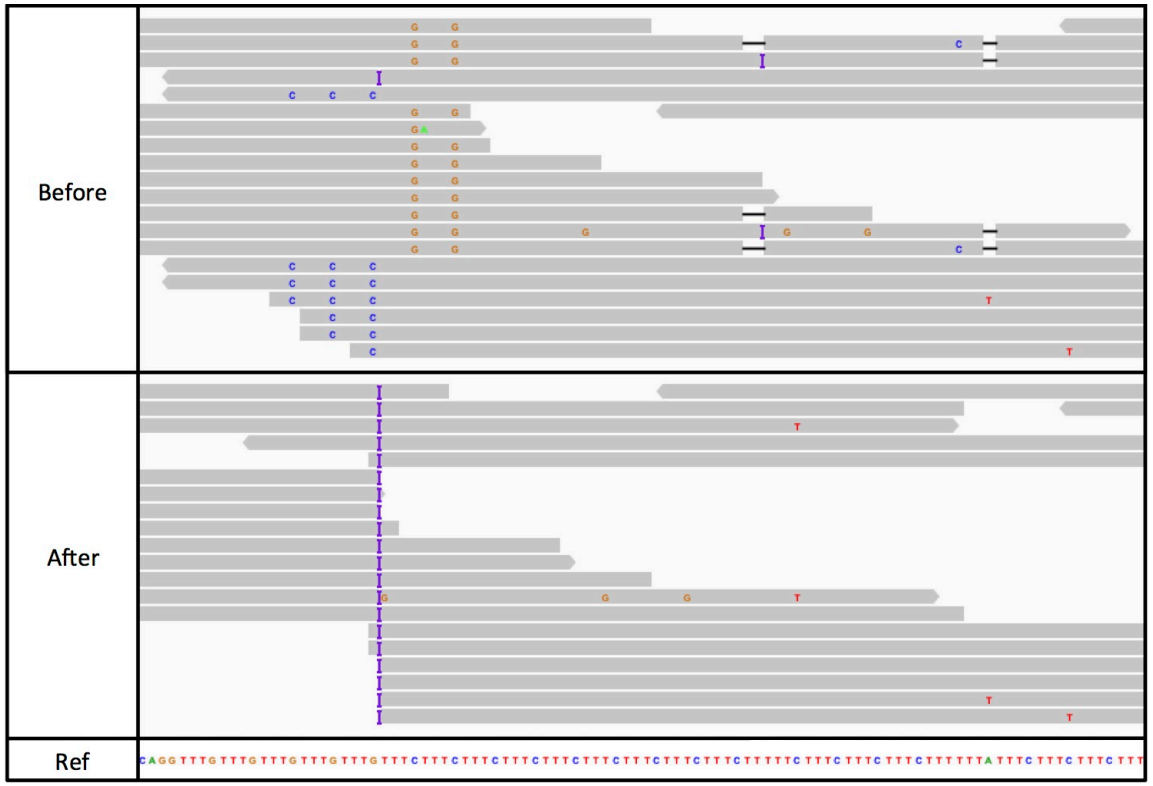
Indel realignment

- Local realignment around indels
- Correct mapping errors – more precise indel discovery



Indel realignment

- Local realignment around indels
- Correct mapping errors – more precise indel discovery



Indel realignment

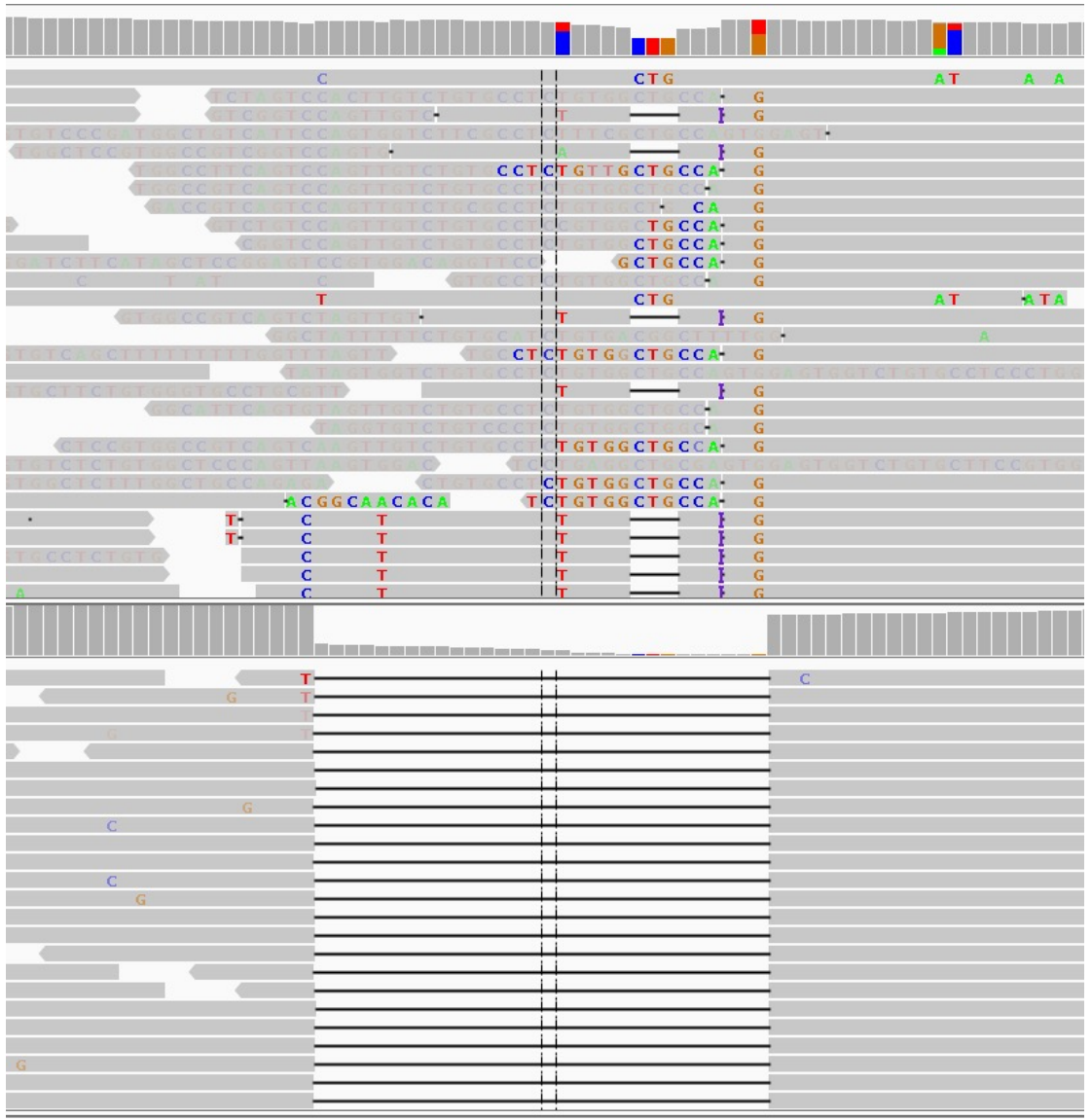
Before

After

Indel realignment

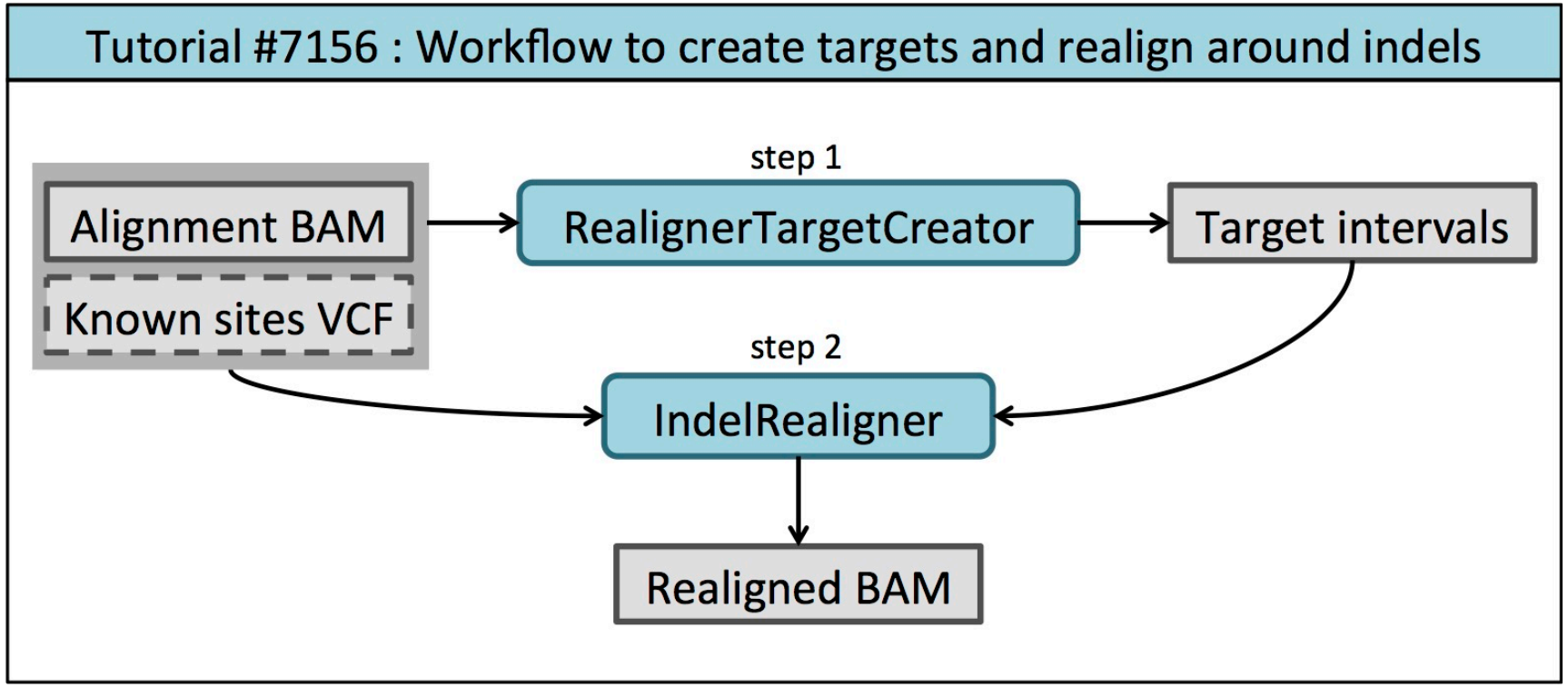
Before

After



Indel realignment

- GATK

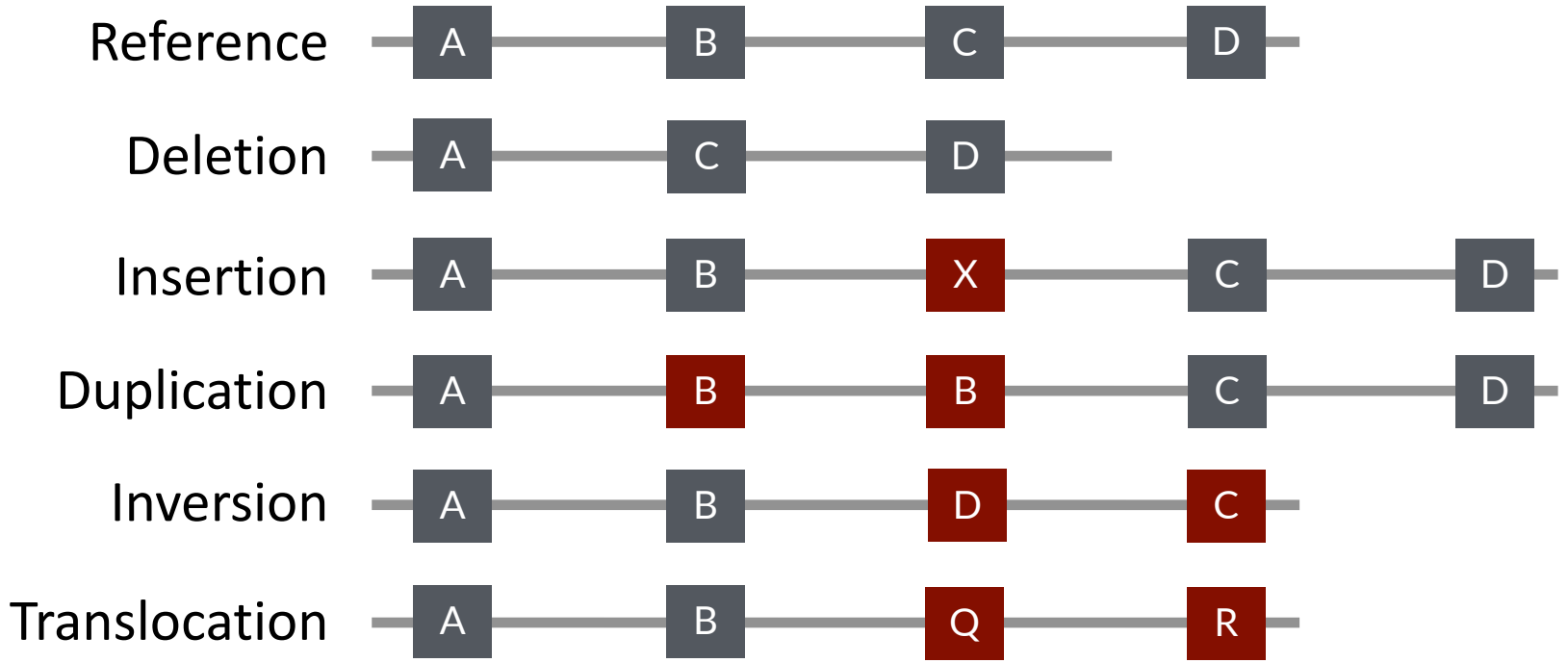


Indel realignment

- ABRA

```
java -Xmx16G -jar $abra_jar
--in <bam_file>
--out <bam_file>
--ref <reference_genome>
--threads <n>
--tmpdir <dir>
> <file>.log
```

Structural variants

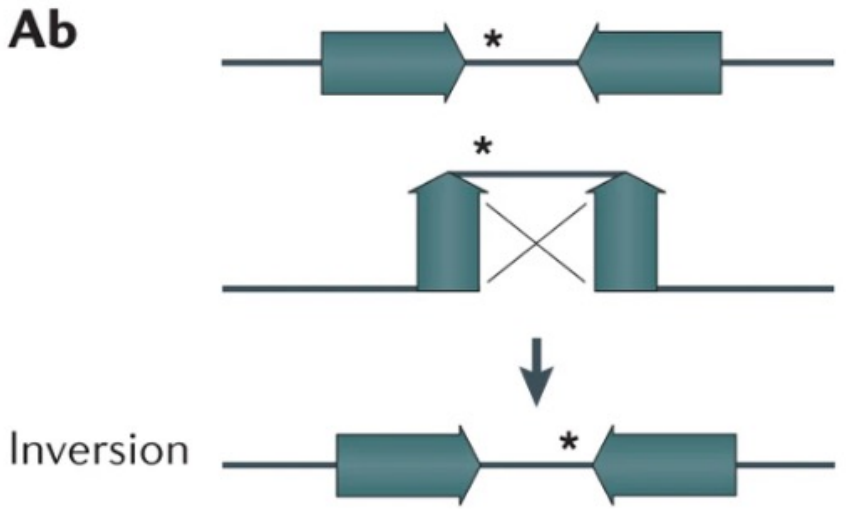


Structural variants

Nonallelic homologous recombination

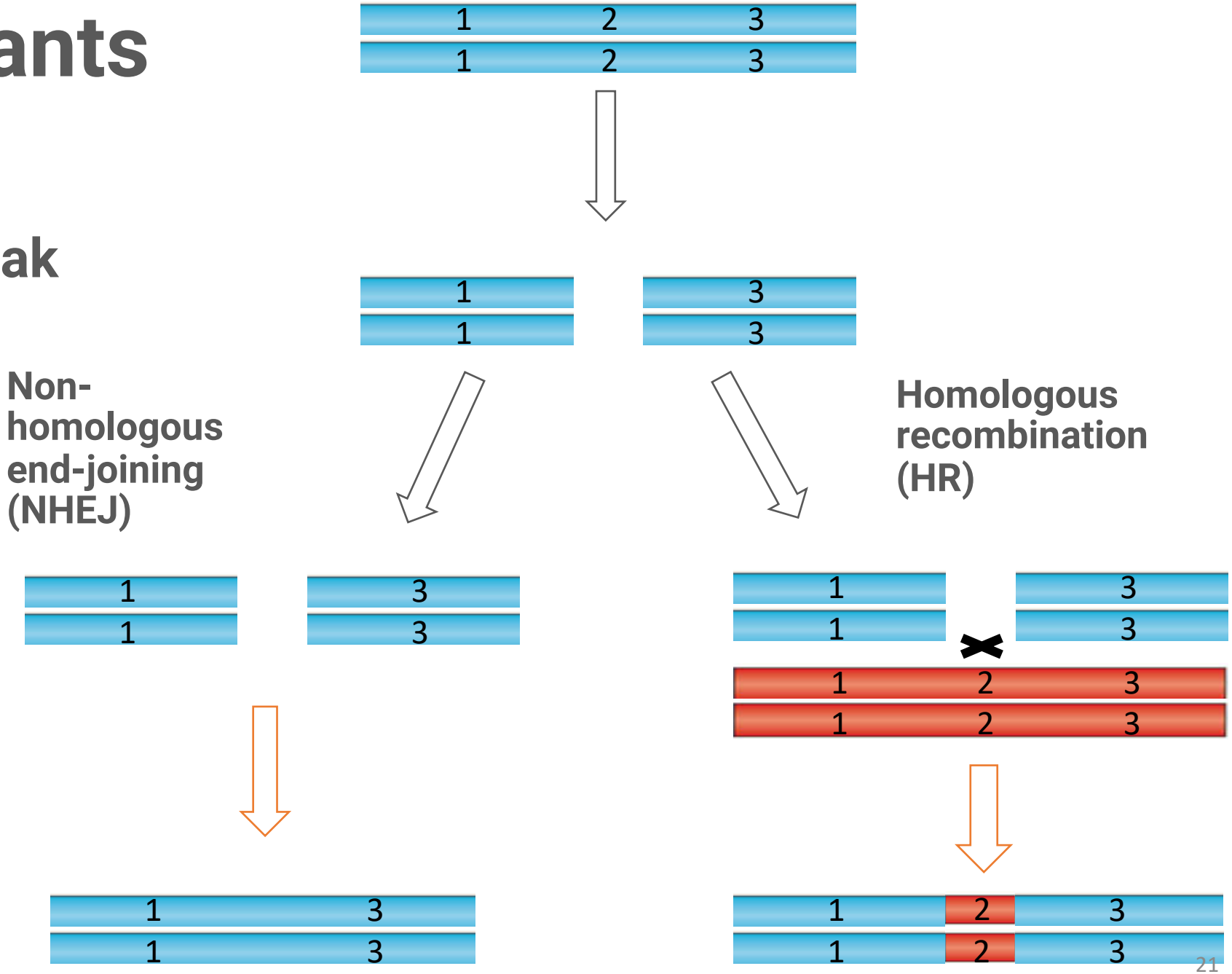
Structural variants

Nonallelic homologous recombination



Structural variants

Double strand break

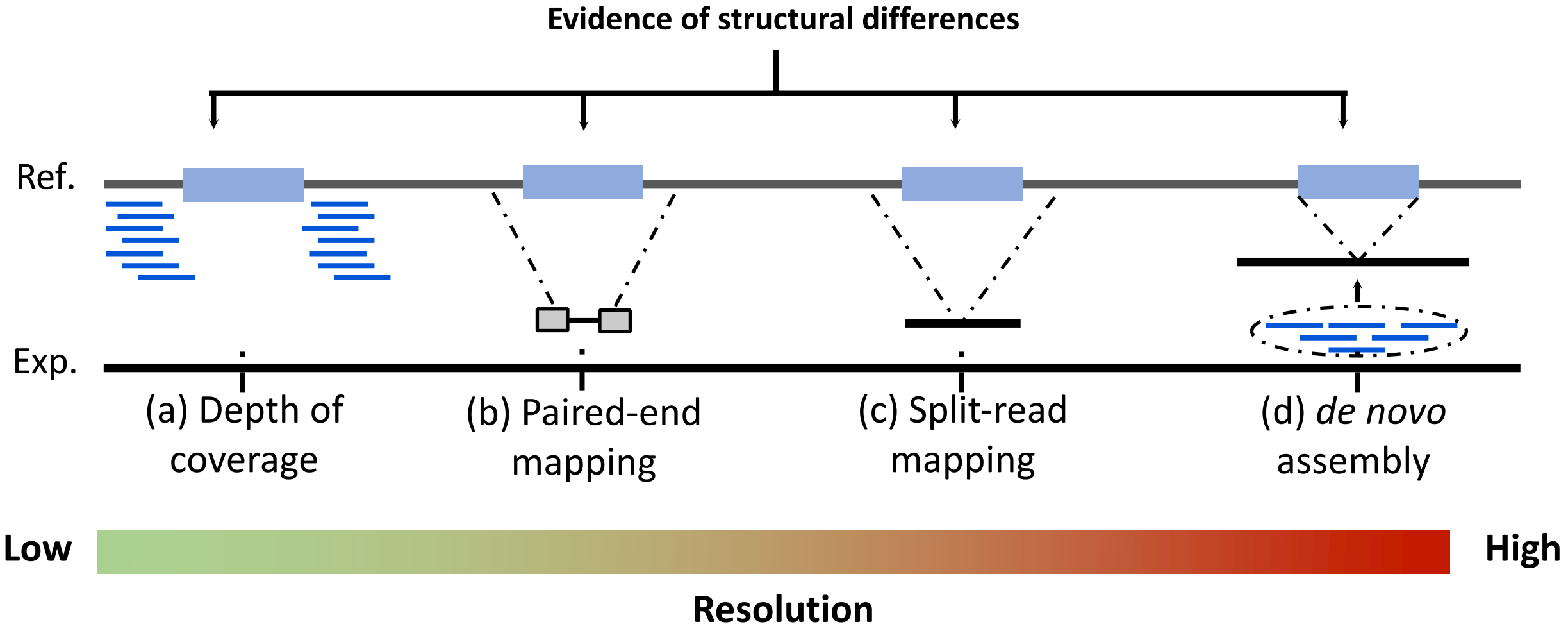


Structural variants

Table 2 Examples of copy number variations (CNVs) and conveyed genomic disorders[a]

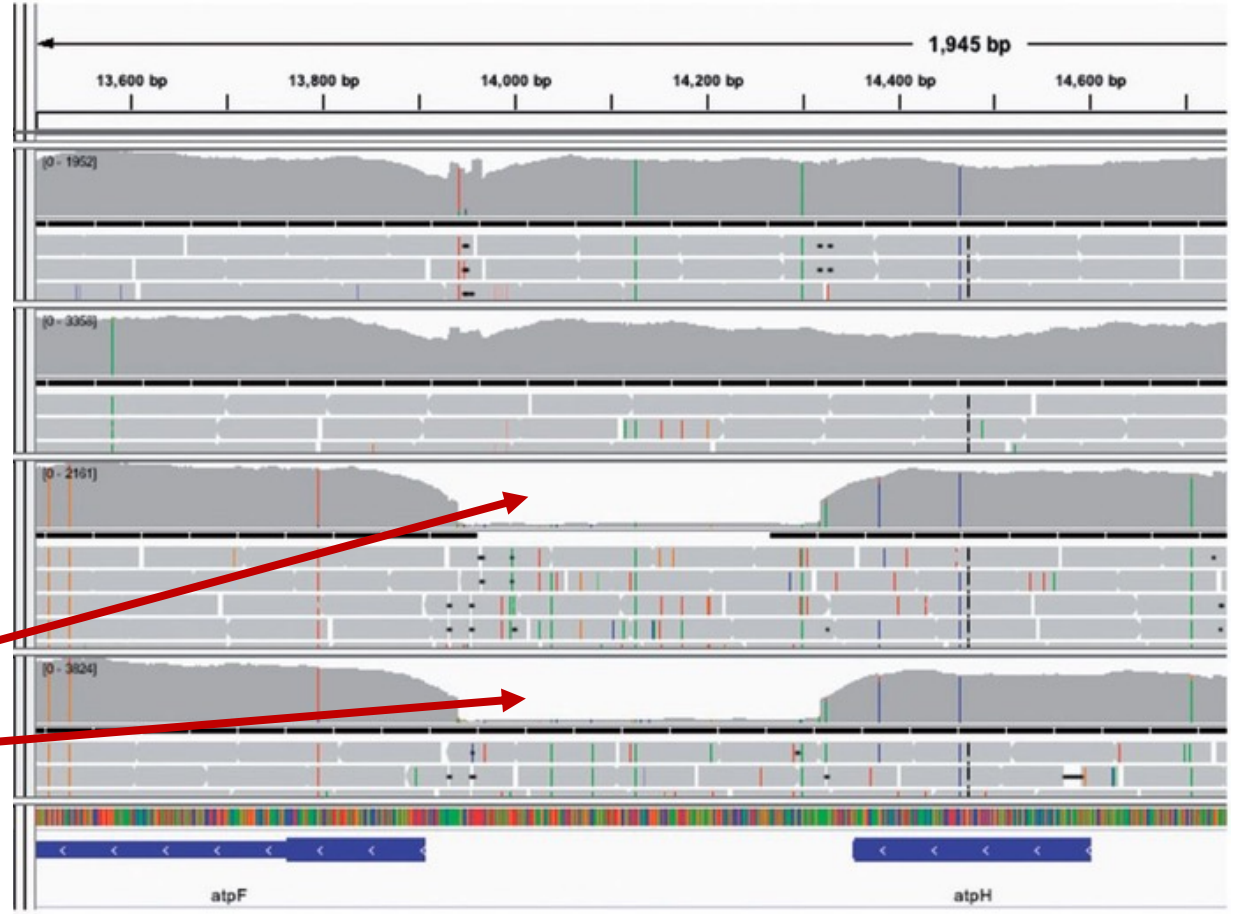
Phenotype	OMIM	Locus	CNV
Mendelian (autosomal dominant)[b]			
Williams-Beuren syndrome	194050	7q11.23	del
7q11.23 duplication syndrome	609757	7q11.23	dup
Spinocerebellar ataxia type 20	608687	11q12	dup
Smith-Magenis syndrome	182290	17p11.2/*RAI1*	del
Potocki-Lupski syndrome	610883	17p11.2	dup
HNPP	162500	17p12/*PMP22*	del
CMT1A	118220	17p12/*PMP22*	dup
Miller-Dieker lissencephaly syndrome	247200	17p13.3/*LIS1*	del
Mental retardation	601545	17p13.3/*LIS1*	dup
DGS/VCFS	188400/192430	22q11.2/*TBX1*	del
Microduplication 22q11.2	608363	22q11.2	dup
Adult-onset leukodystrophy	169500	*LMNB1*	dup
Mendelian (autosomal recessive)			
Familial juvenile nephronophthisis	256100	2q13/*NPHP1*	del
Gaucher disease	230800	1q21/*GBA*	del
Pituitary dwarfism	262400	17q24/*GH1*	del
Spinal muscular atrophy	253300	5q13/*SMN1*	del
beta-thalassemia	141900	11p15/*beta-globin*	del
alpha-thalassemia	141750	16p13.3/*HBA*	del
Mendelian (X-linked)			
Hemophilia A	306700	*F8*	inv/del
Hunter syndrome	309900	*IDS*	del/inv
Ichthyosis	308100	*STS*	del
Mental retardation	300706	*HUWE1*	dup
Pelizaeus-Merzbacher disease	312080	*PLP1*	del/dup/tri
Progressive neurological symptoms (MR+SZ)	300260	*MECP2*	dup
Red-green color blindness	303800	opsin genes	del
Complex traits			
Alzheimer disease	104300	*APP*	dup
Autism	612200	3q24	inherited homozygous del
	611913	16p11.2	del/dup
Crohn disease	266600	*HBD-2*	copy number loss
	612278	*IRGM*	del
HIV susceptibility	609423	*CCL3L1*	copy number loss
Mental retardation	612001	15q13.3	del
	610443	17q21.31	del
	300534	Xp11.22	dup
Pancreatitis	167800	*PRSS1*	tri
Parkinson disease	168600	*SNCA*	dup/tri
Psoriasis	177900	*DEFB*	copy number gain
Schizophrenia	612474	1q21.1	del
	181500	15q11.2	del
	612001	15q13.3	del
Systemic lupus erythematosus	152700	*FCGR3B*	copy number loss
	120810	*C4*	copy number loss

Structural variants



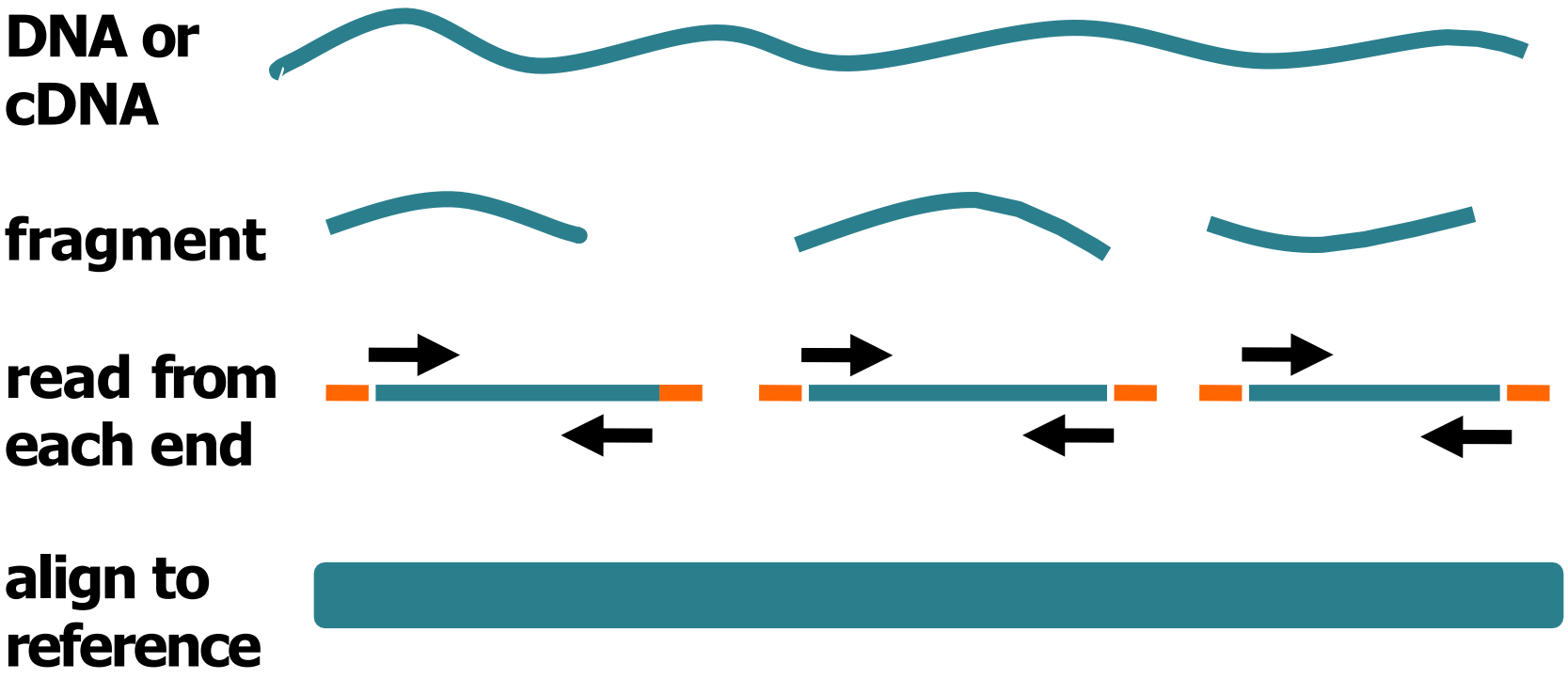
Structural variants - Deletions

Depth of Coverage

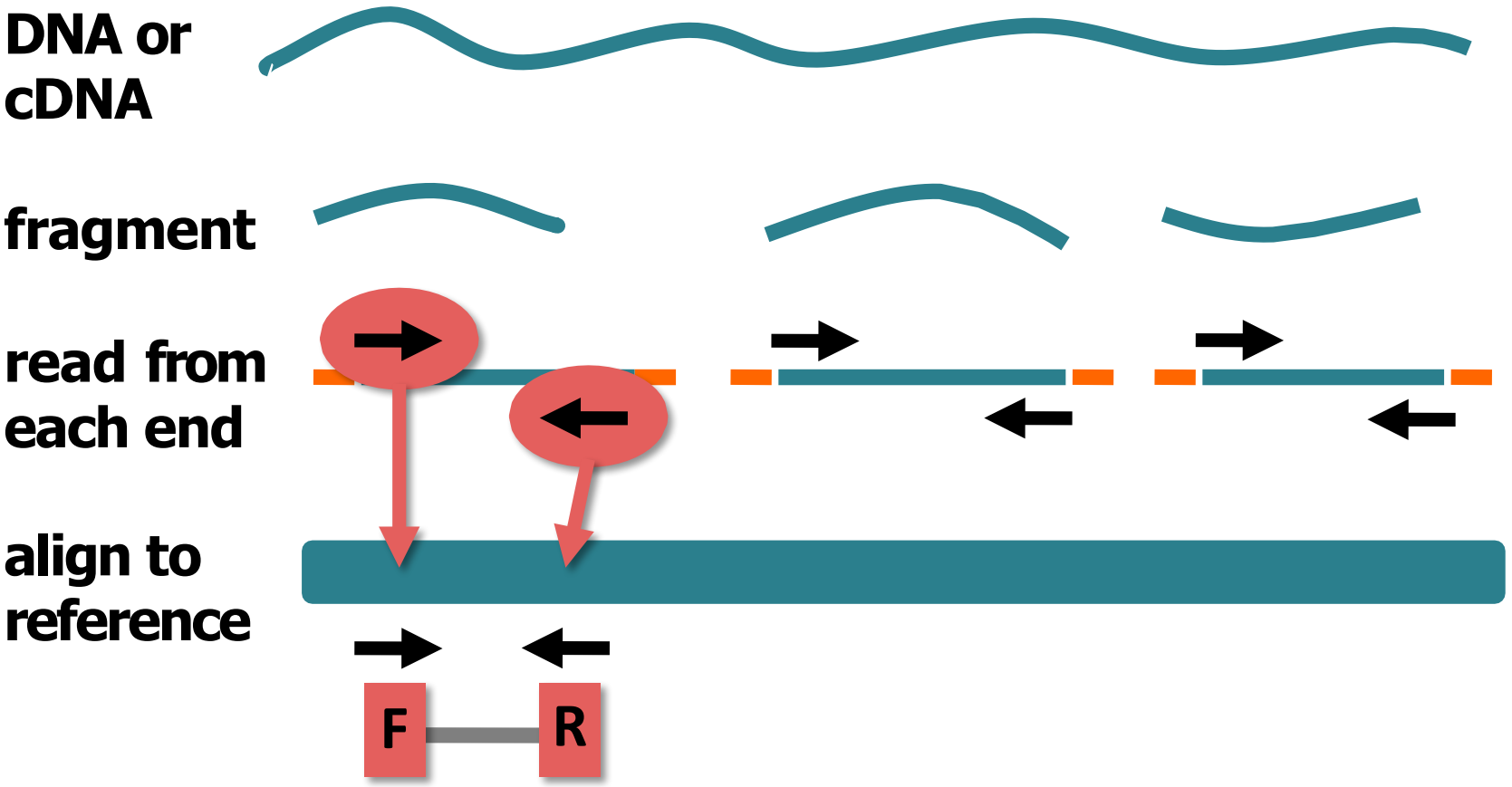


Drop in coverage is a sign of a deletion

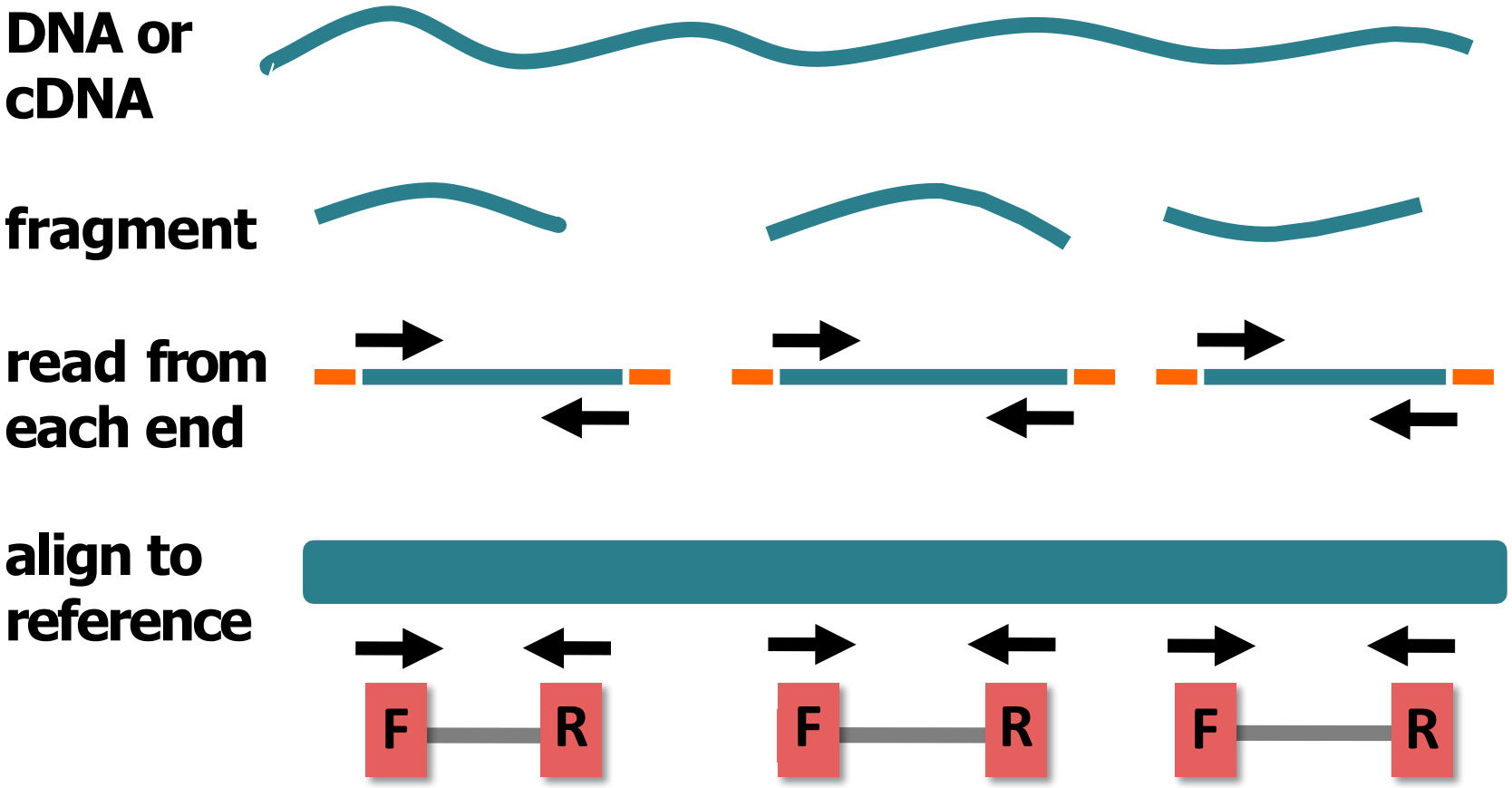
Structural variants – Paired end sequencing



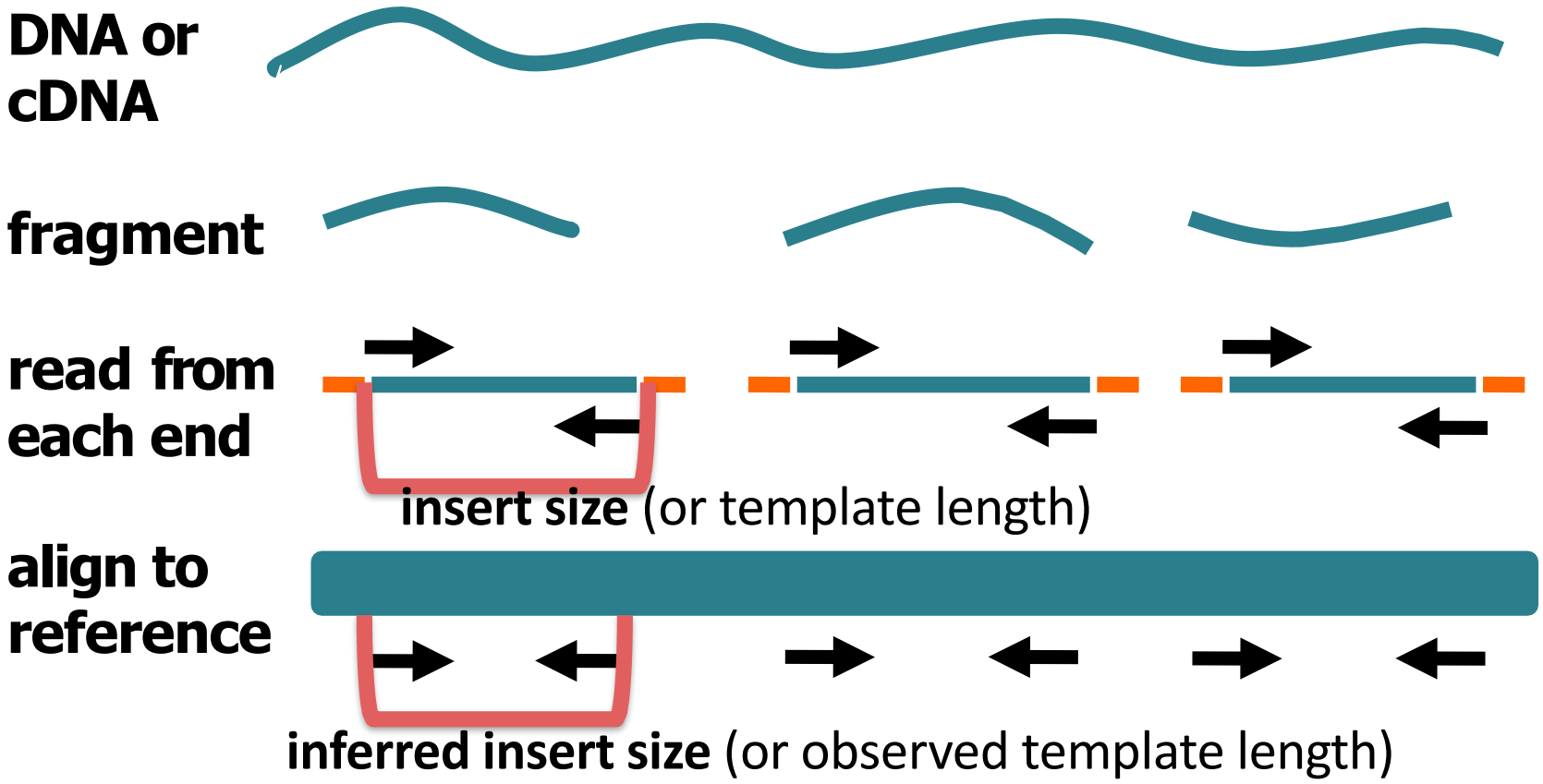
Structural variants – Paired end sequencing



Structural variants – Paired end sequencing

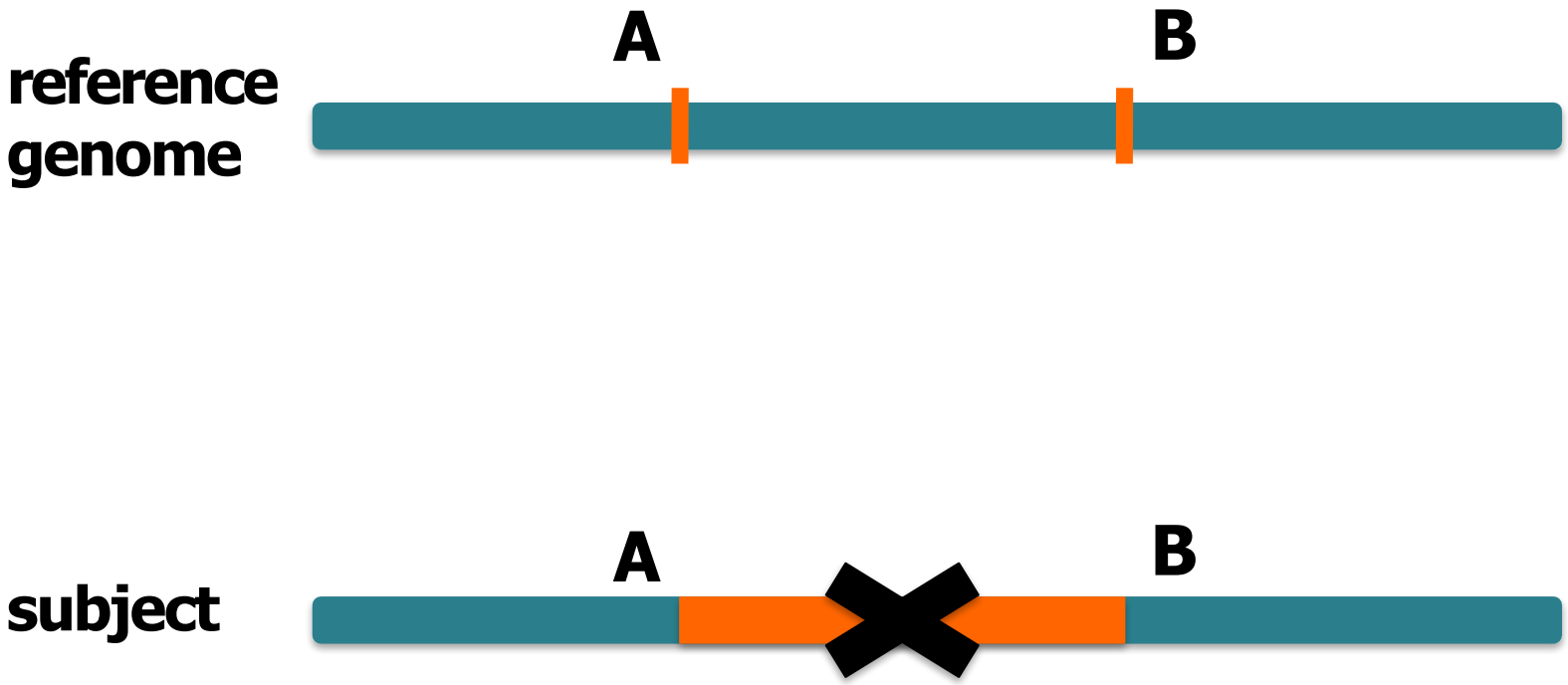


Structural variants – Paired end sequencing



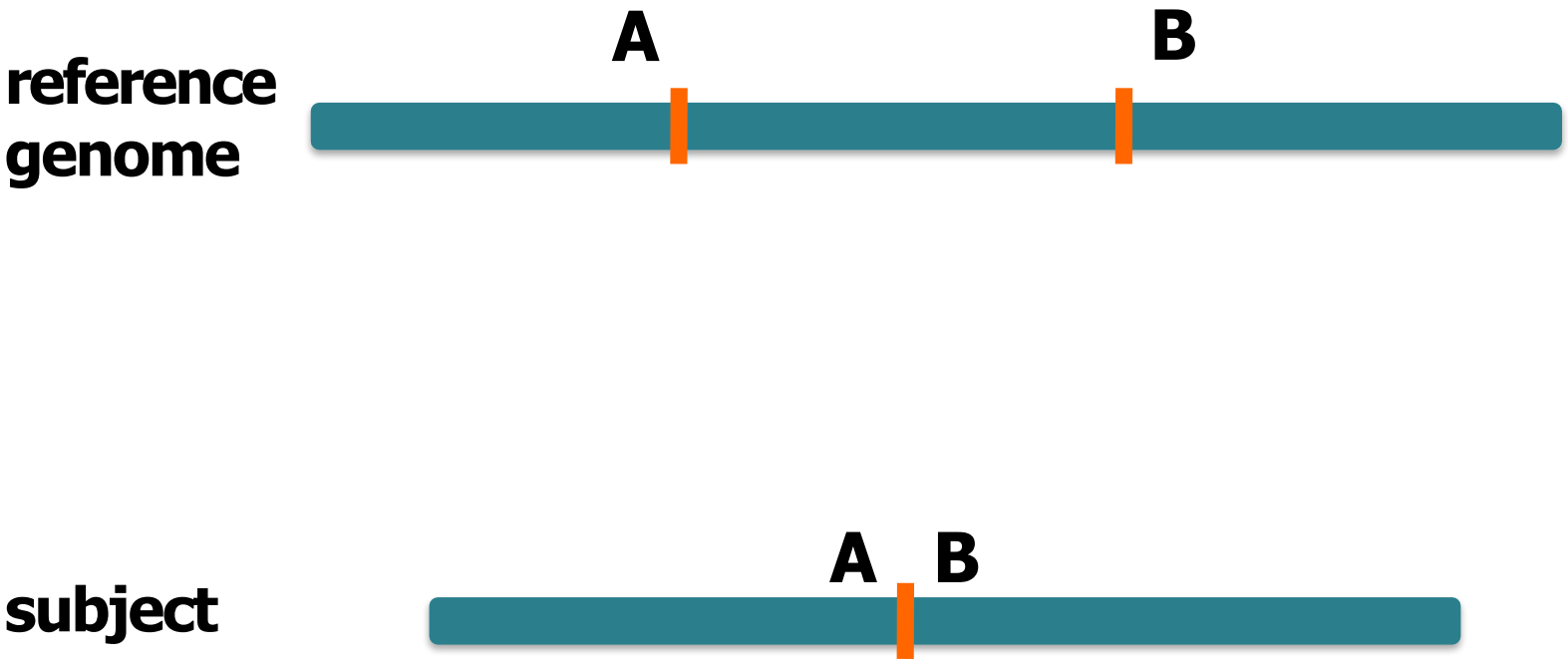
Structural variants – Deletions

Insert size



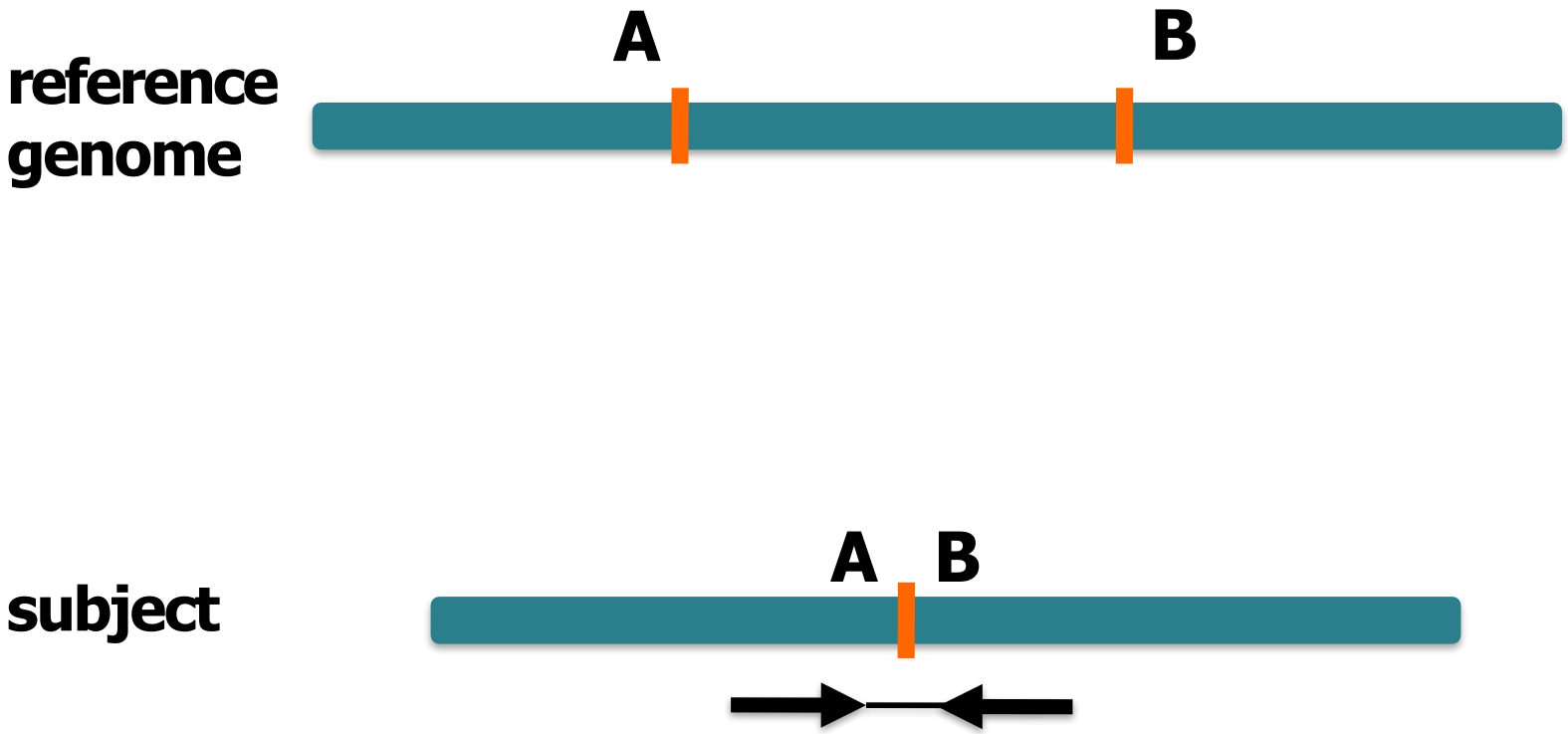
Structural variants – Deletions

Insert size



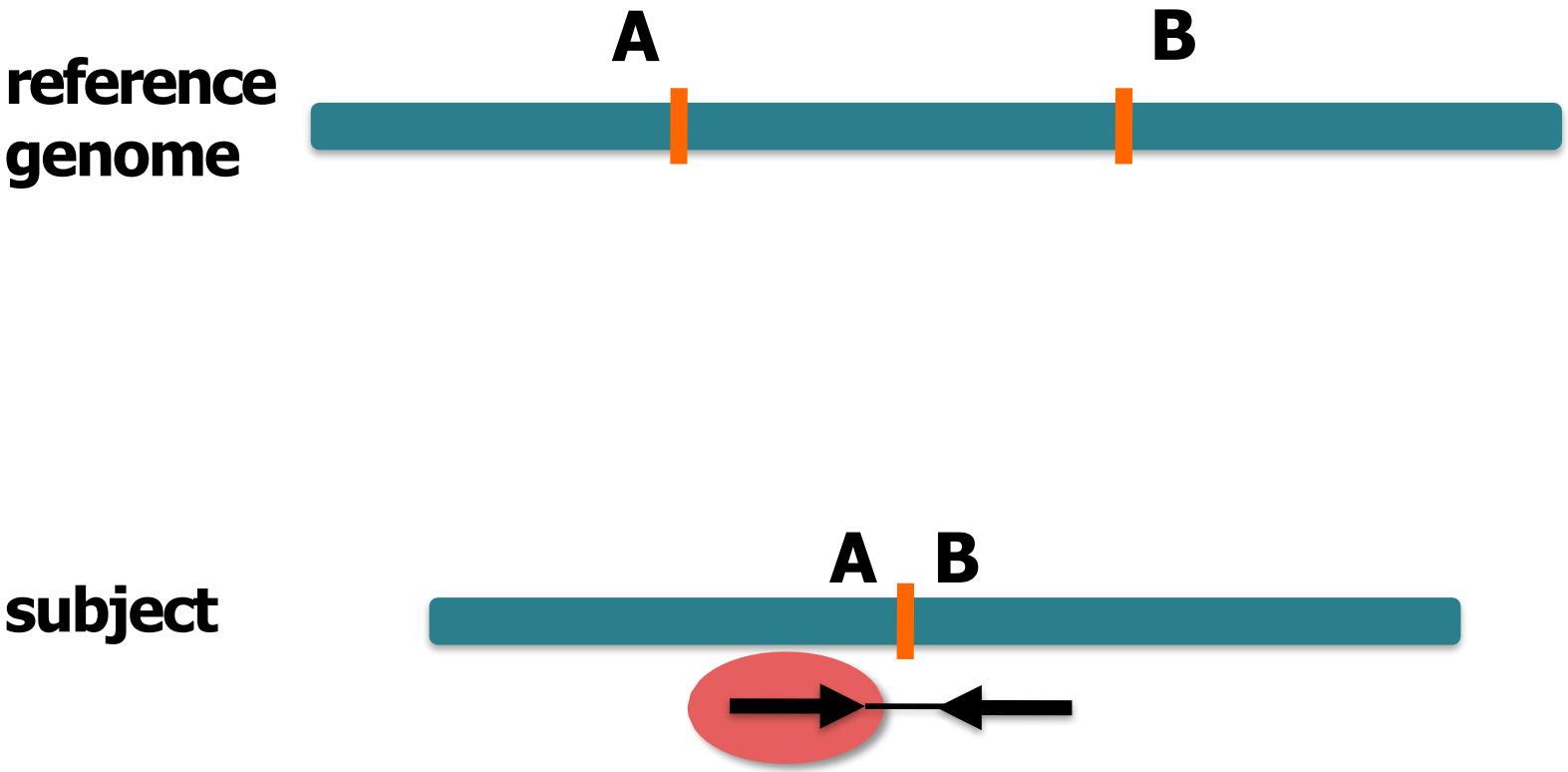
Structural variants – Deletions

Insert size



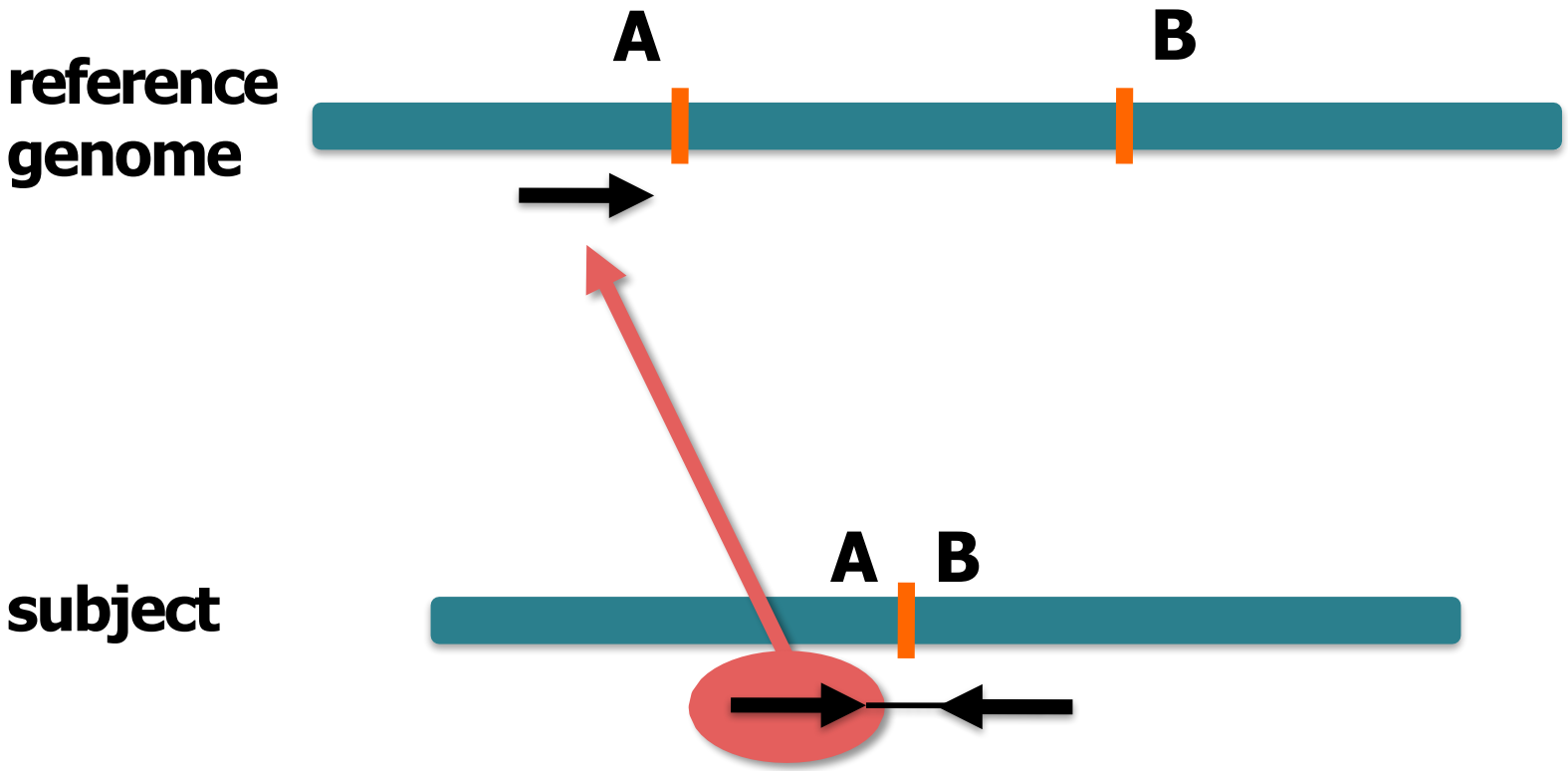
Structural variants – Deletions

Insert size



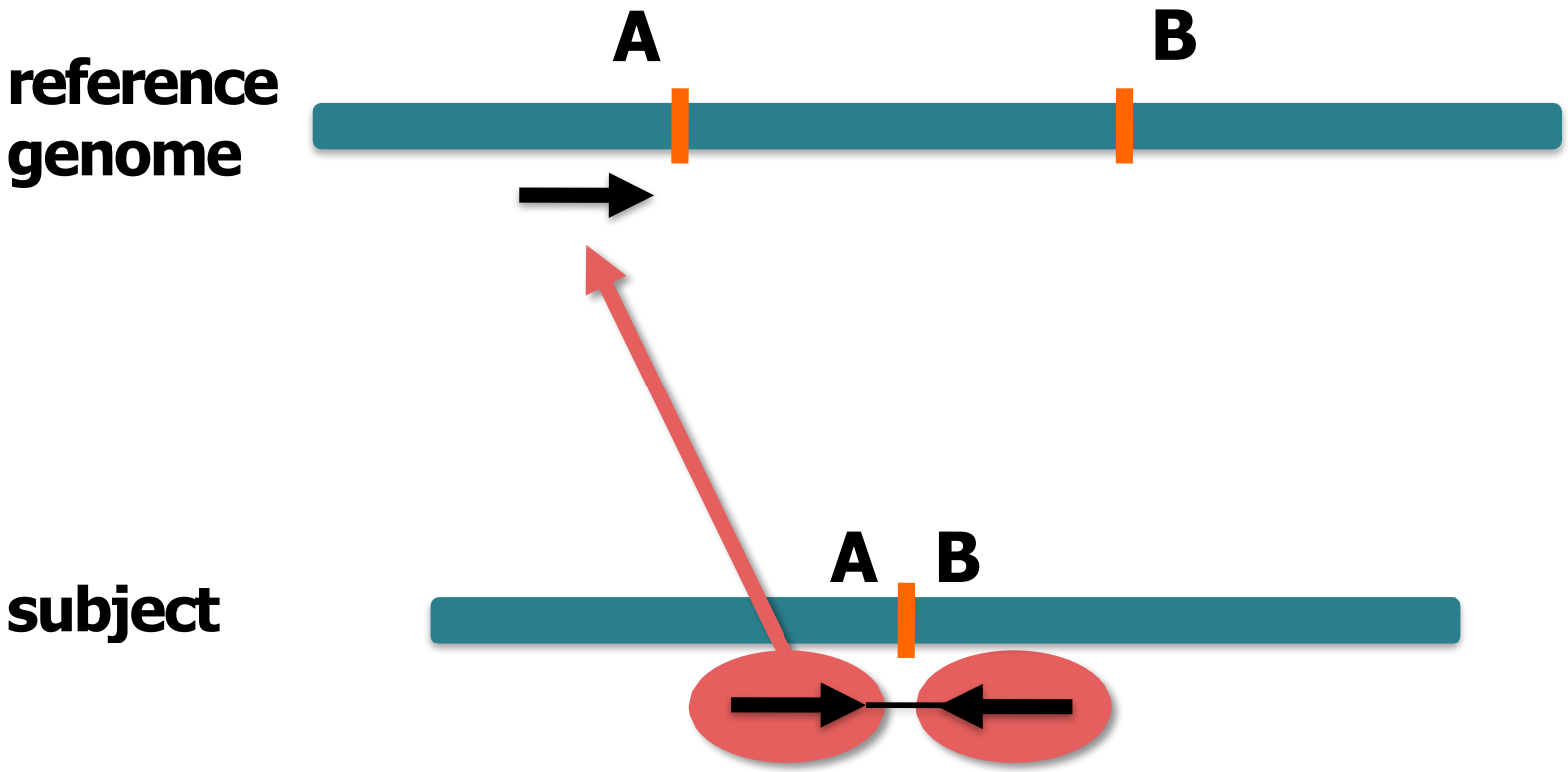
Structural variants – Deletions

Insert size



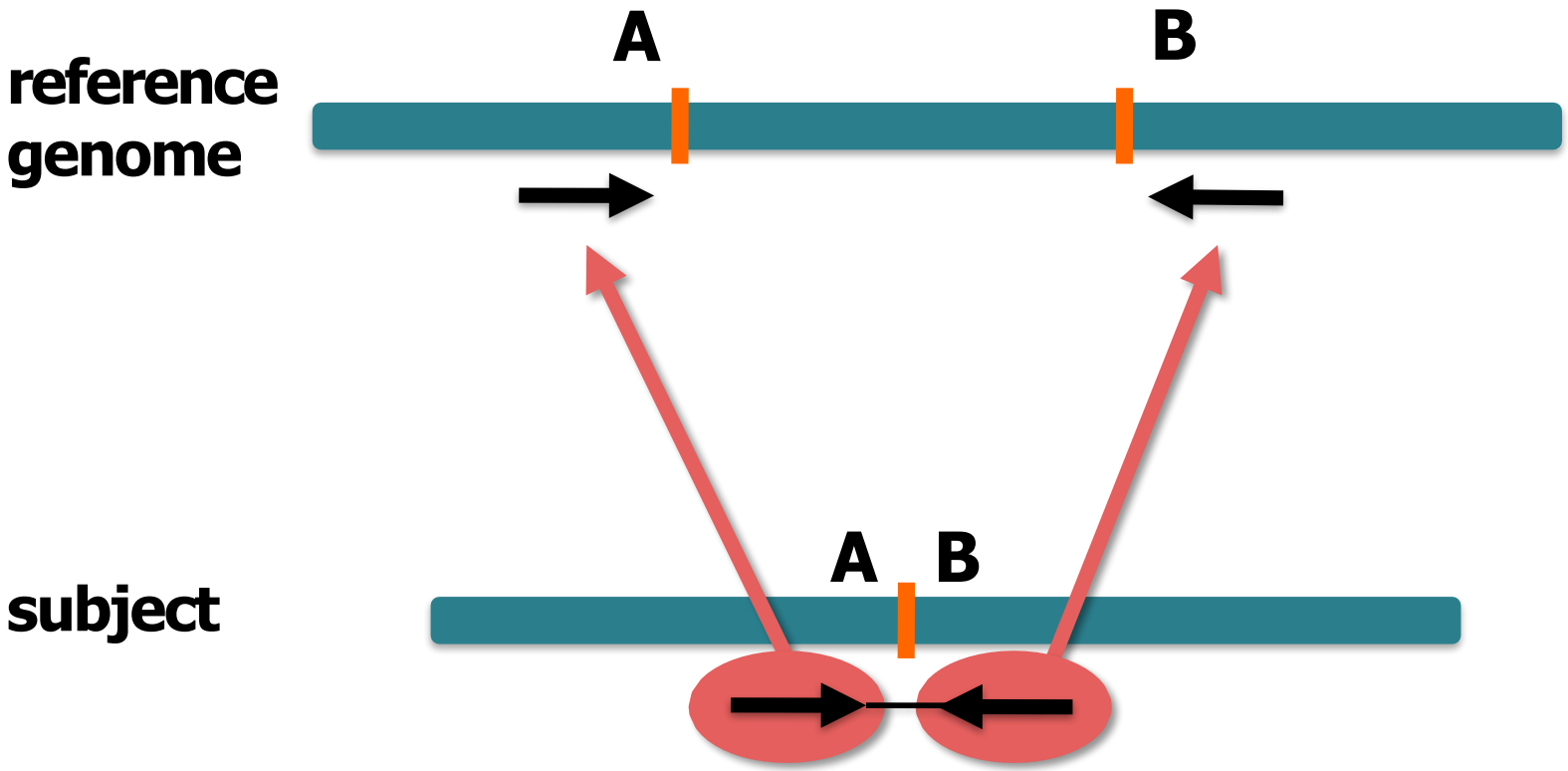
Structural variants – Deletions

Insert size



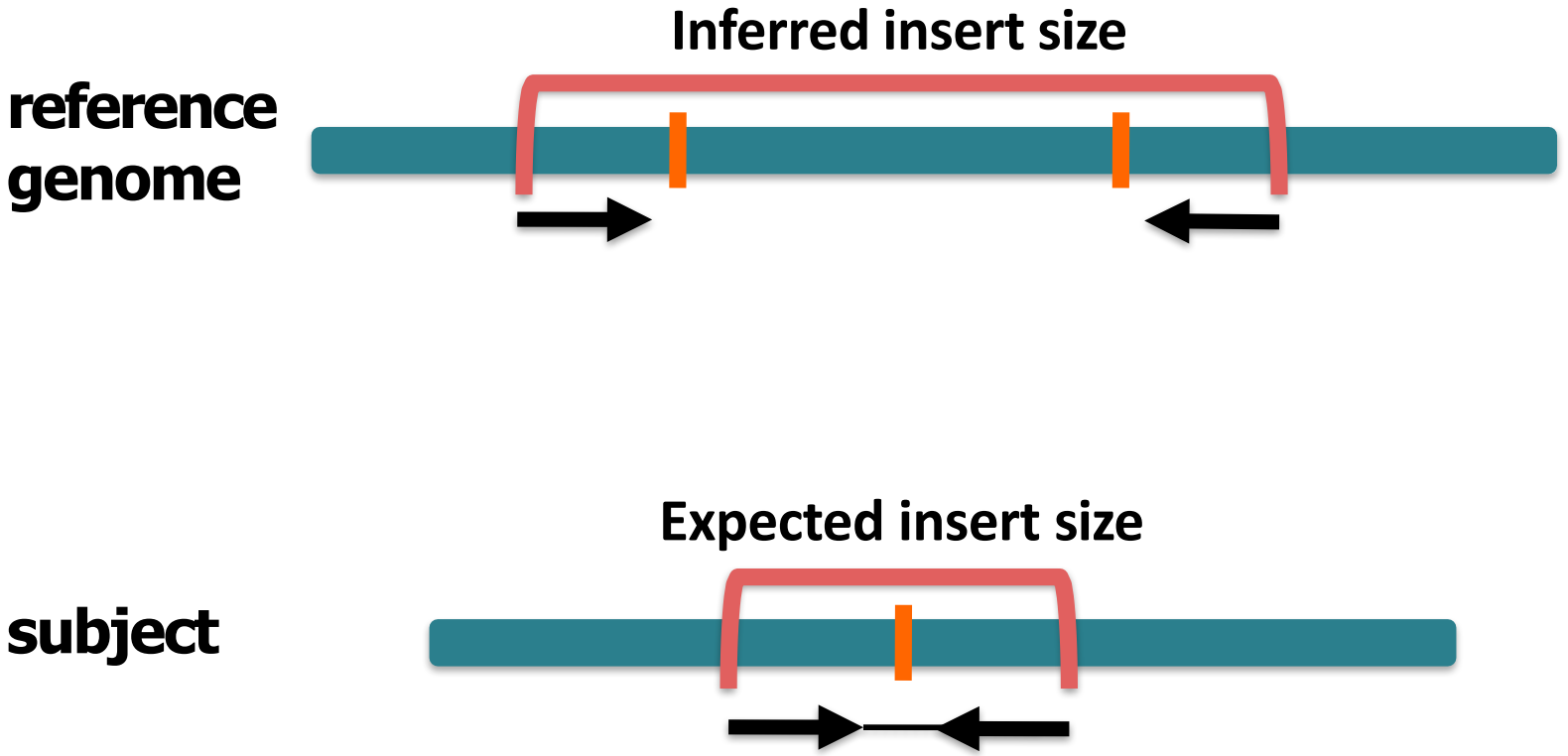
Structural variants – Deletions

Insert size



Structural variants – Deletions

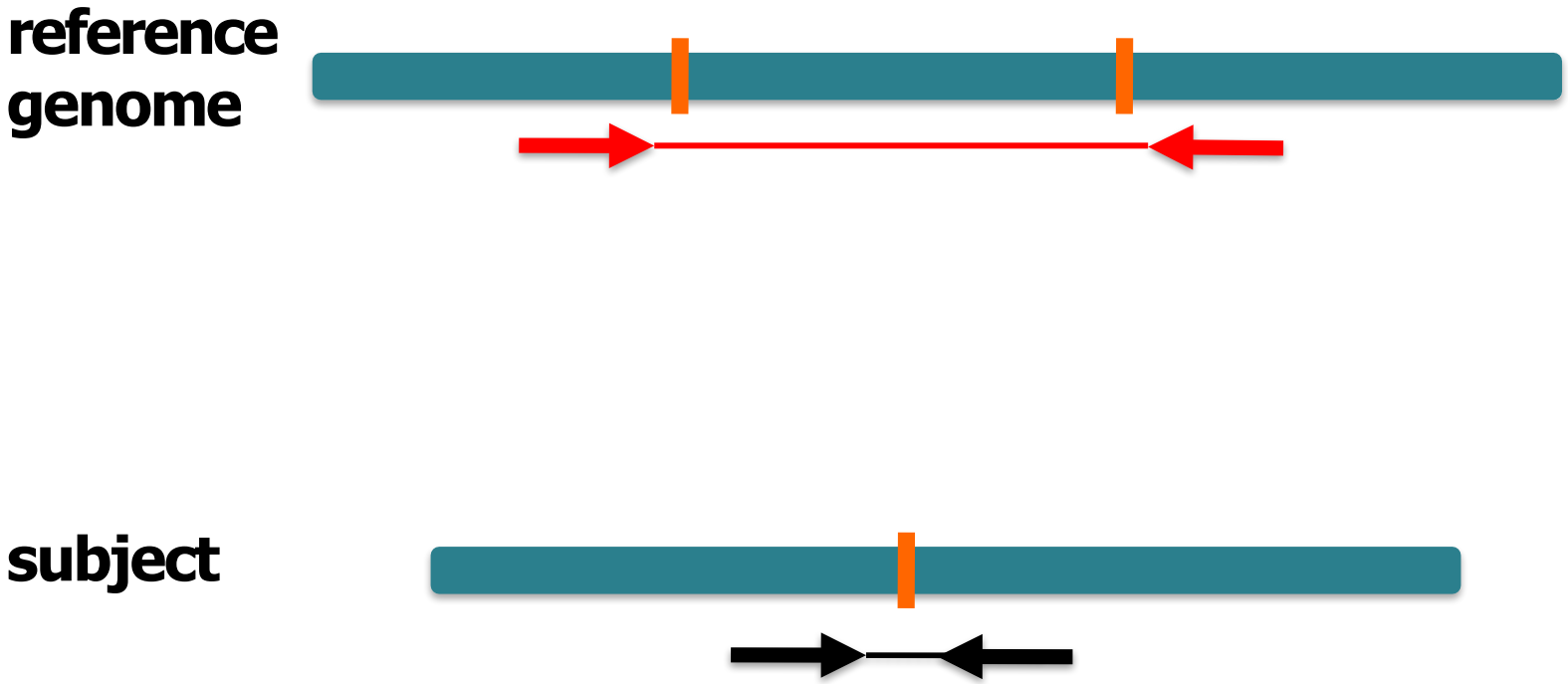
Insert size



Inferred insert size is greater than expected value

Structural variants – Deletions

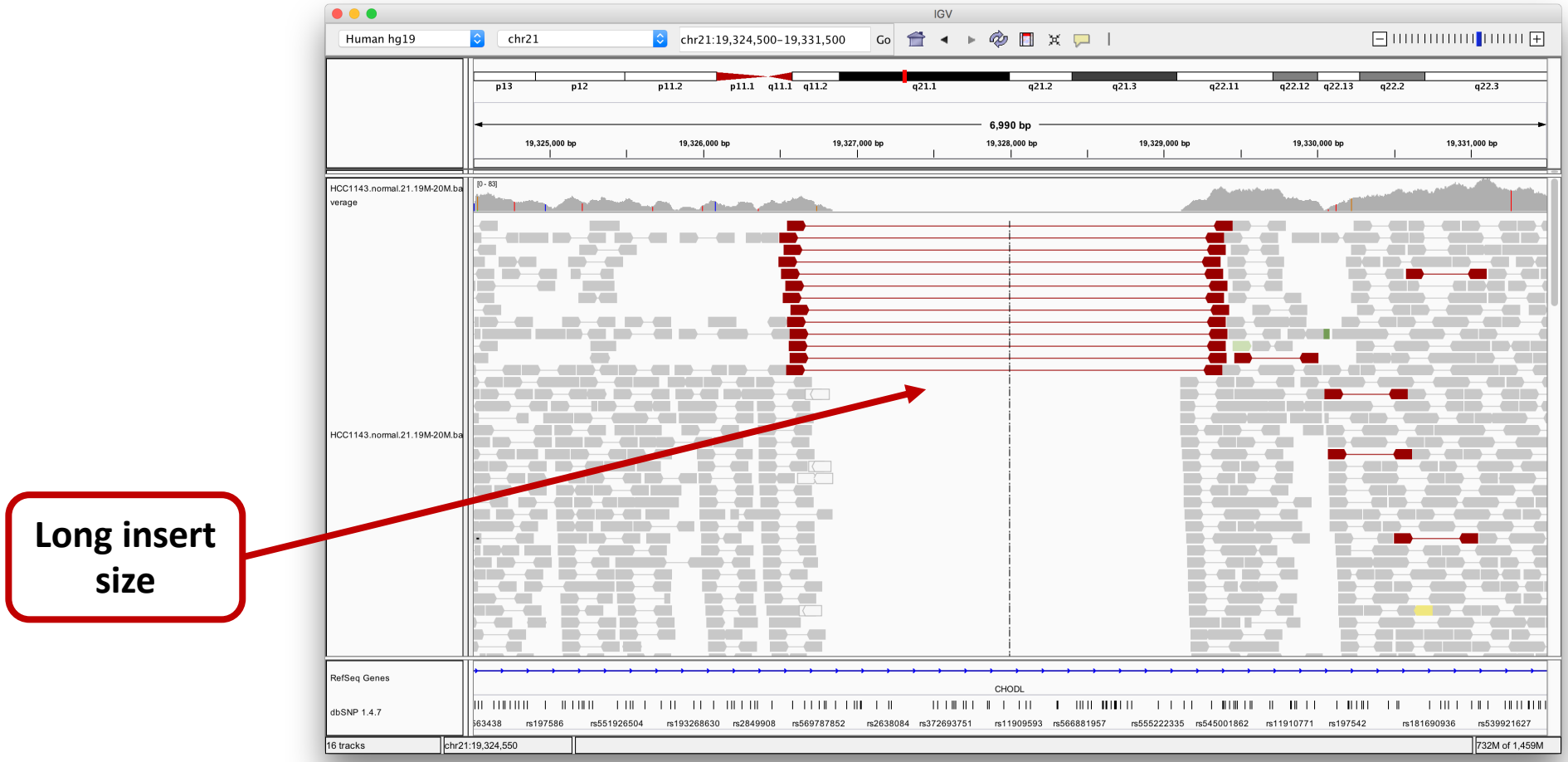
Insert size



Pairs with larger than expected insert size are colored red in IGV

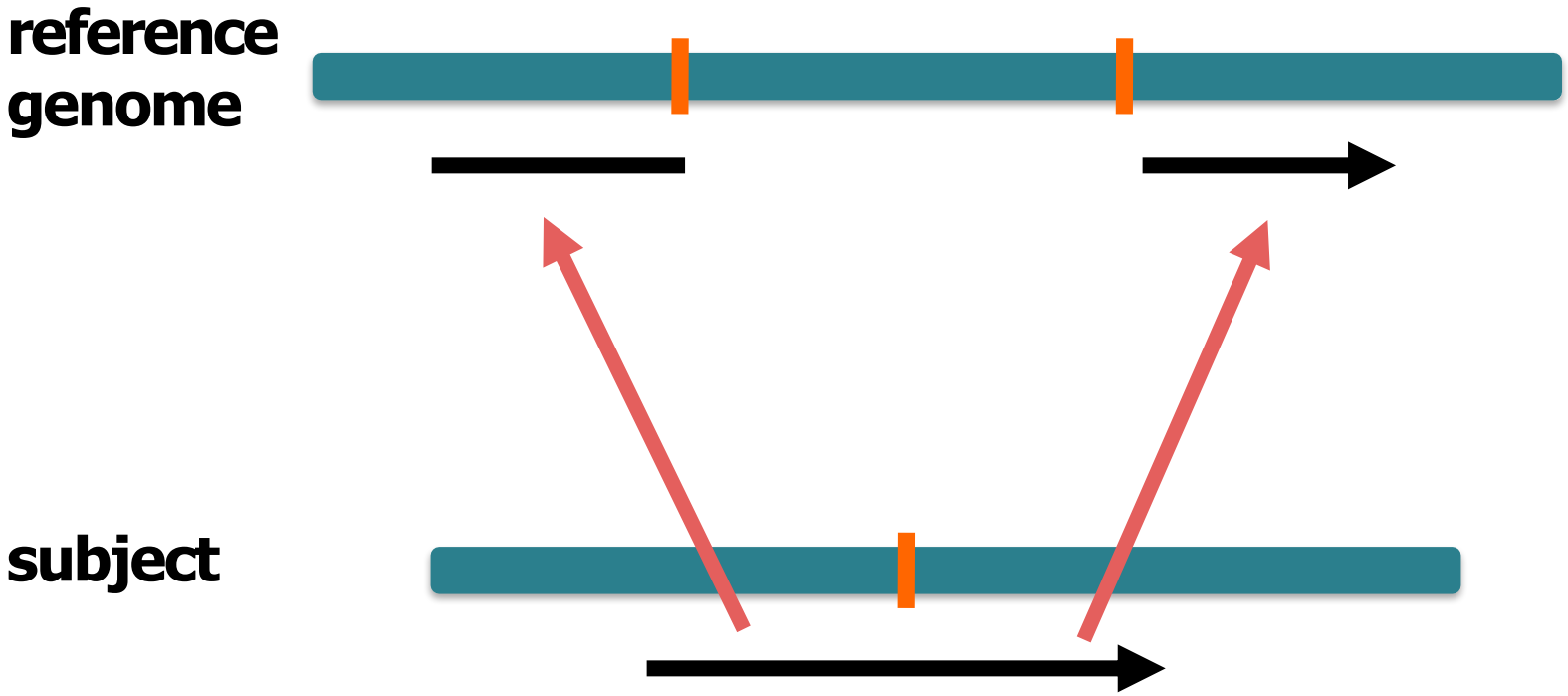
Structural variants – Deletions

Insert size



Structural variants – Deletions

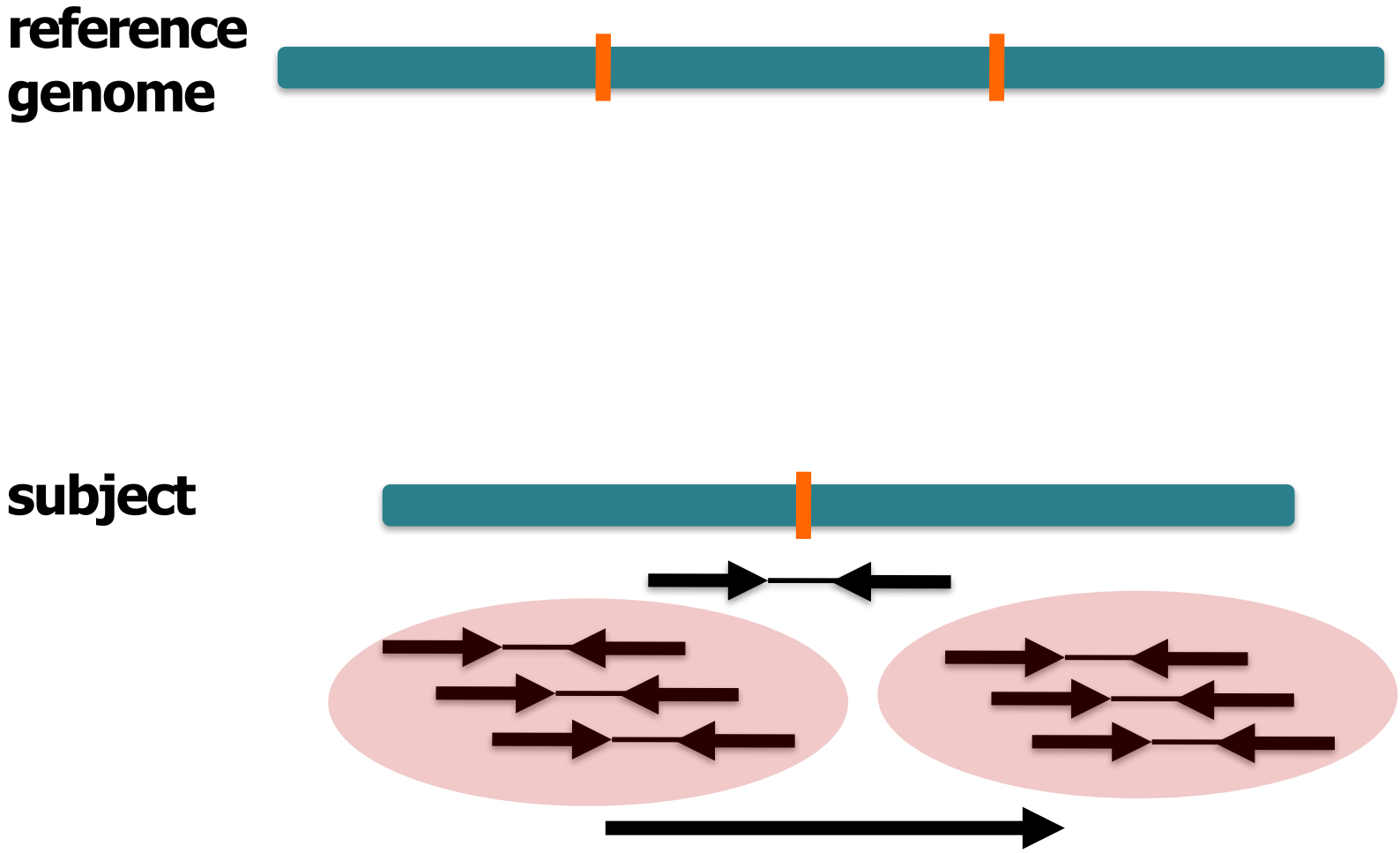
Split reads



Reads spanning the breakpoints will get split when mapped to the reference genome and mapped to both sides of the breakpoints – precise breakpoint detection

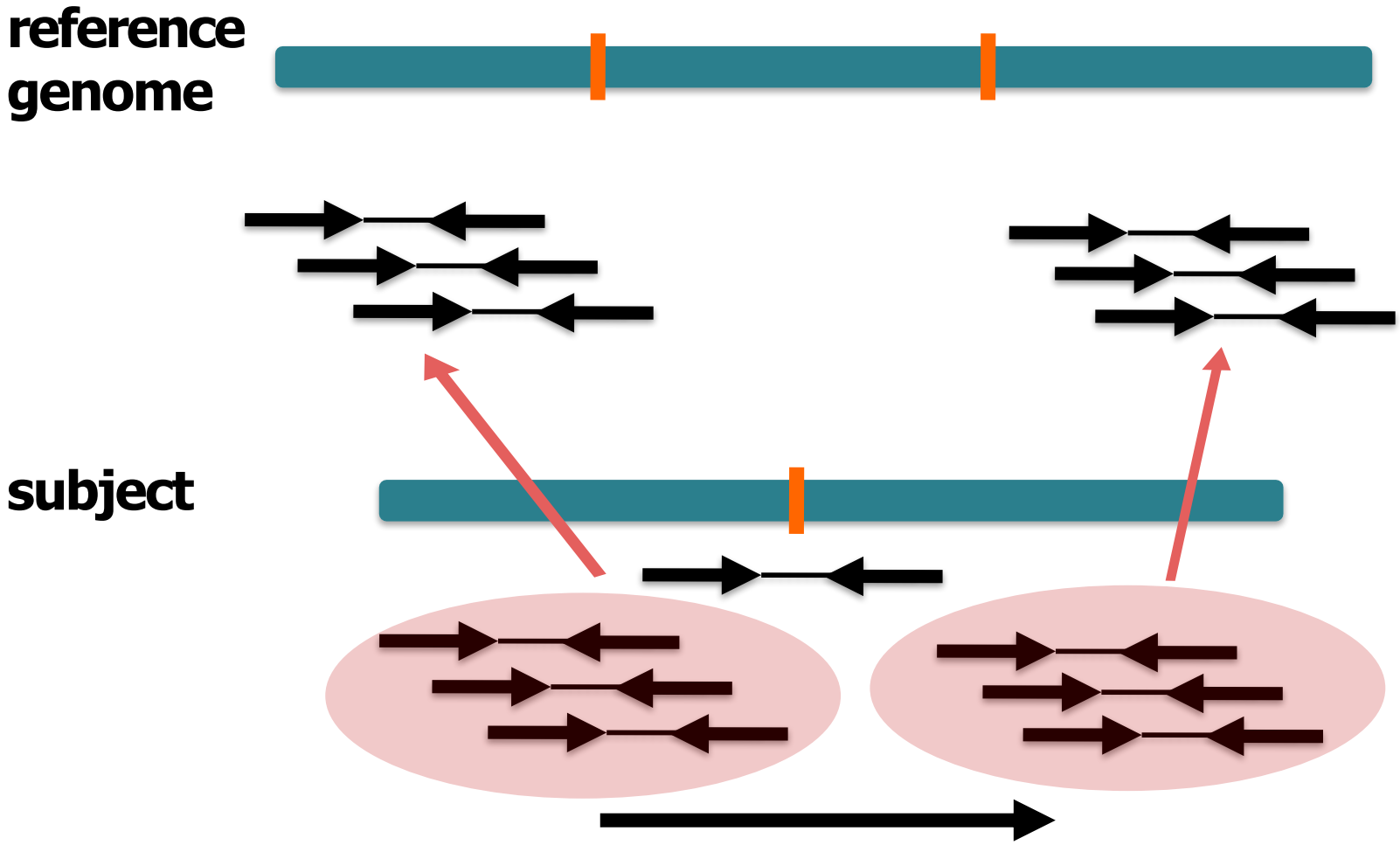
Structural variants – Deletions

1. Depth of coverage



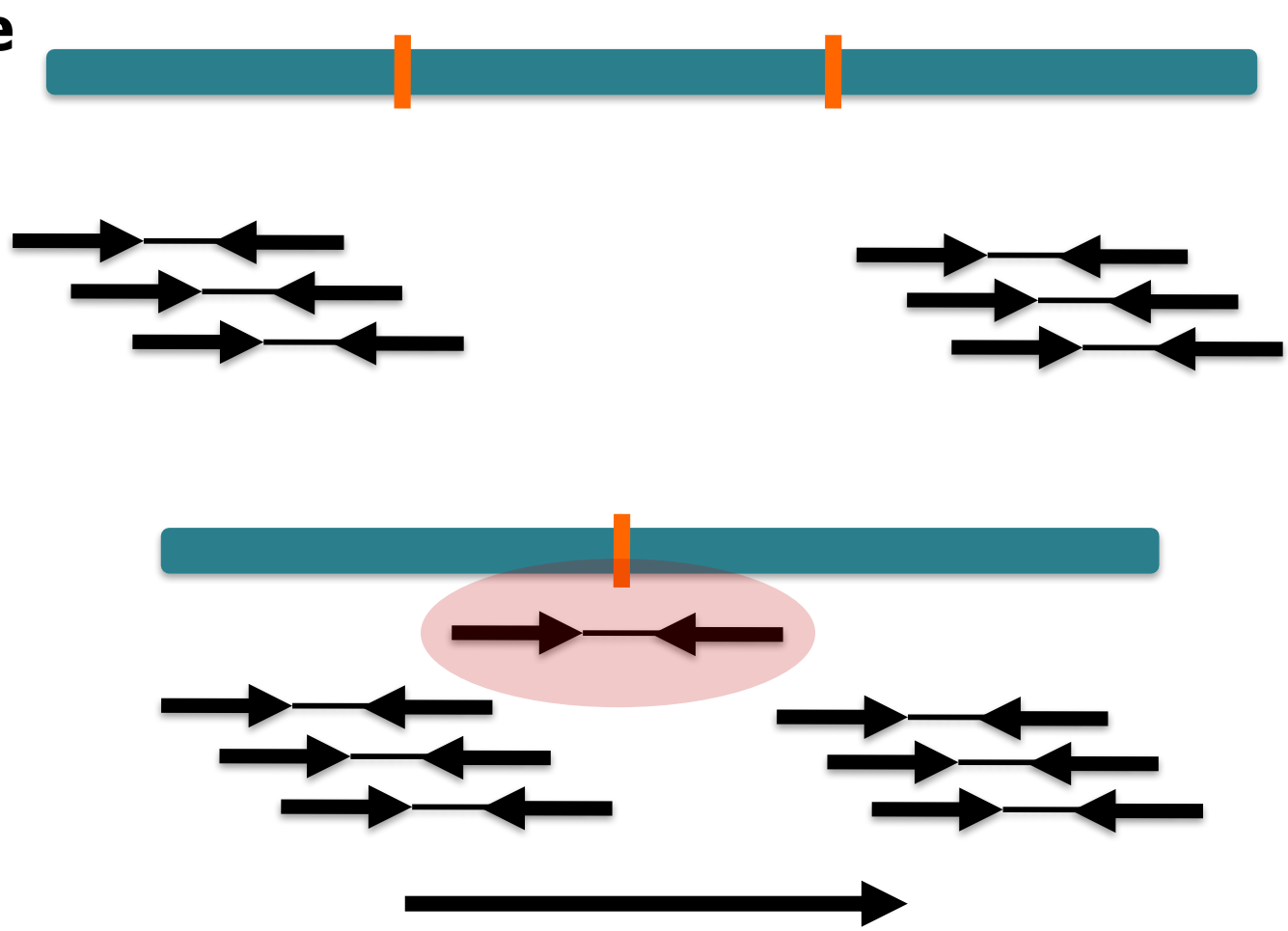
Structural variants – Deletions

1. Depth of coverage



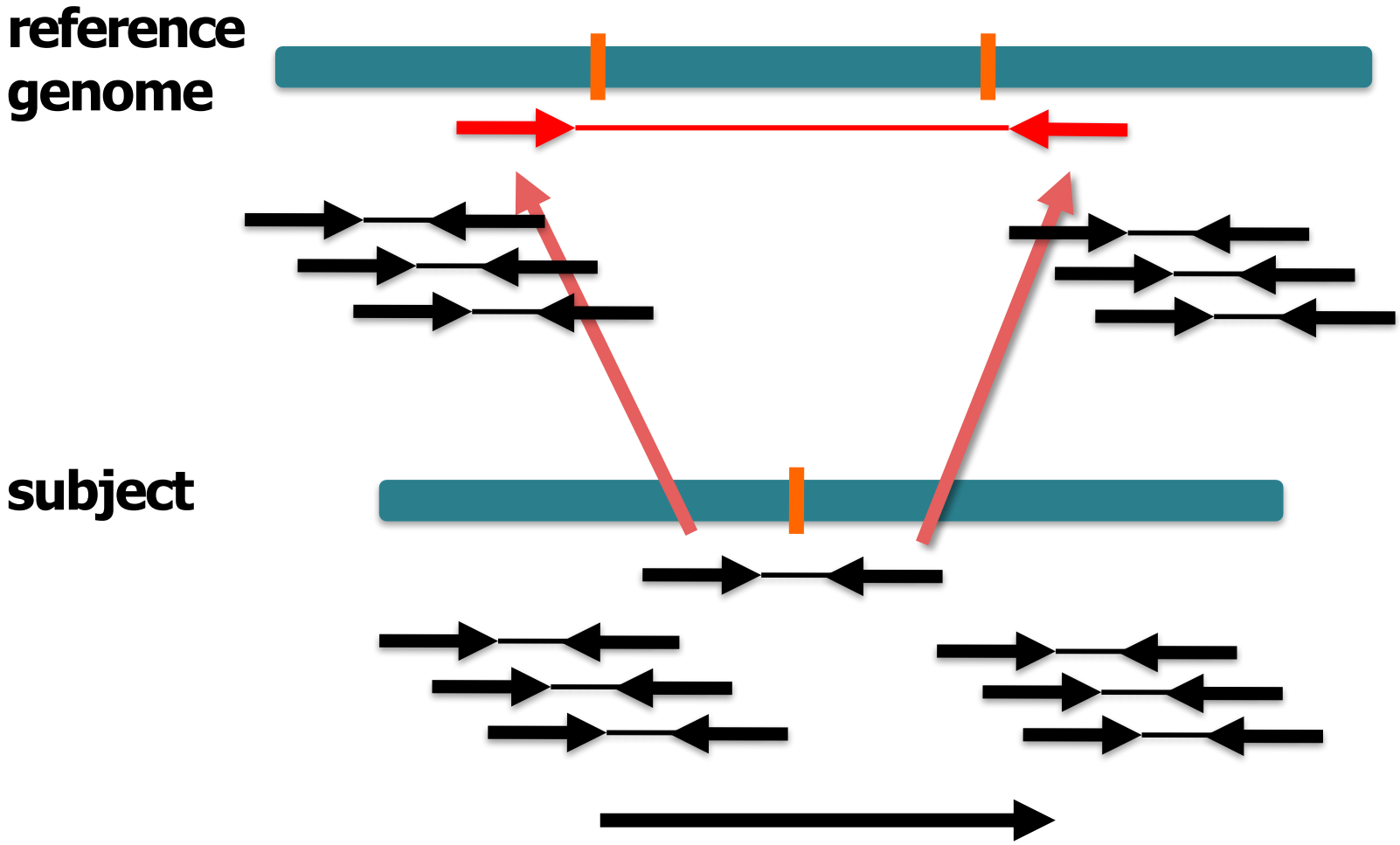
Structural variants – Deletions

1. Depth of coverage
2. Paired-end mapping



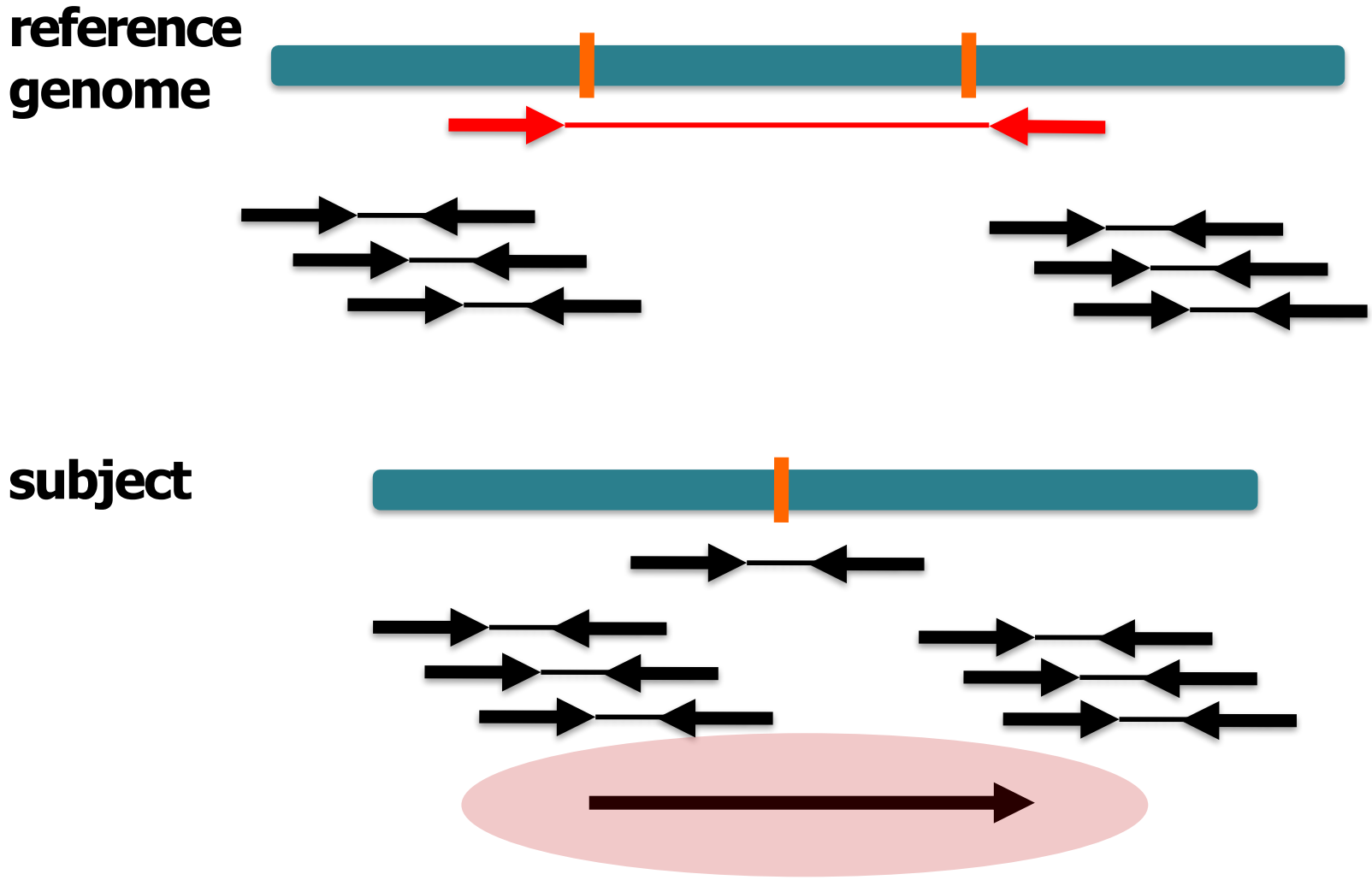
Structural variants – Deletions

1. Depth of coverage
2. Paired-end mapping



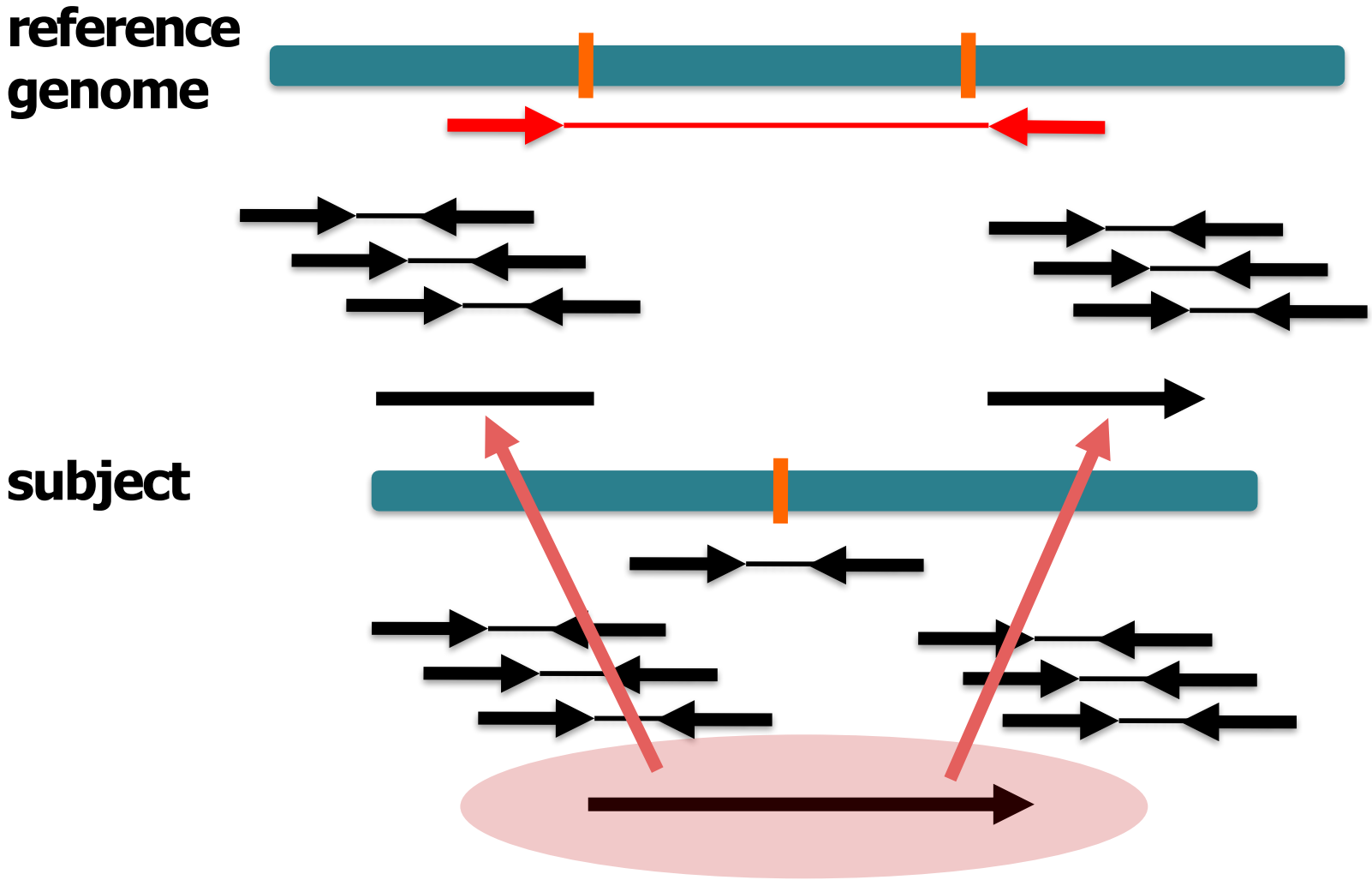
Structural variants – Deletion

1. Depth of coverage
2. Paired-end mapping
3. Split read



Structural variants – Deletion

1. Depth of coverage
2. Paired-end mapping
3. Split read



Structural variants – Insertion

1. Paired-end mapping

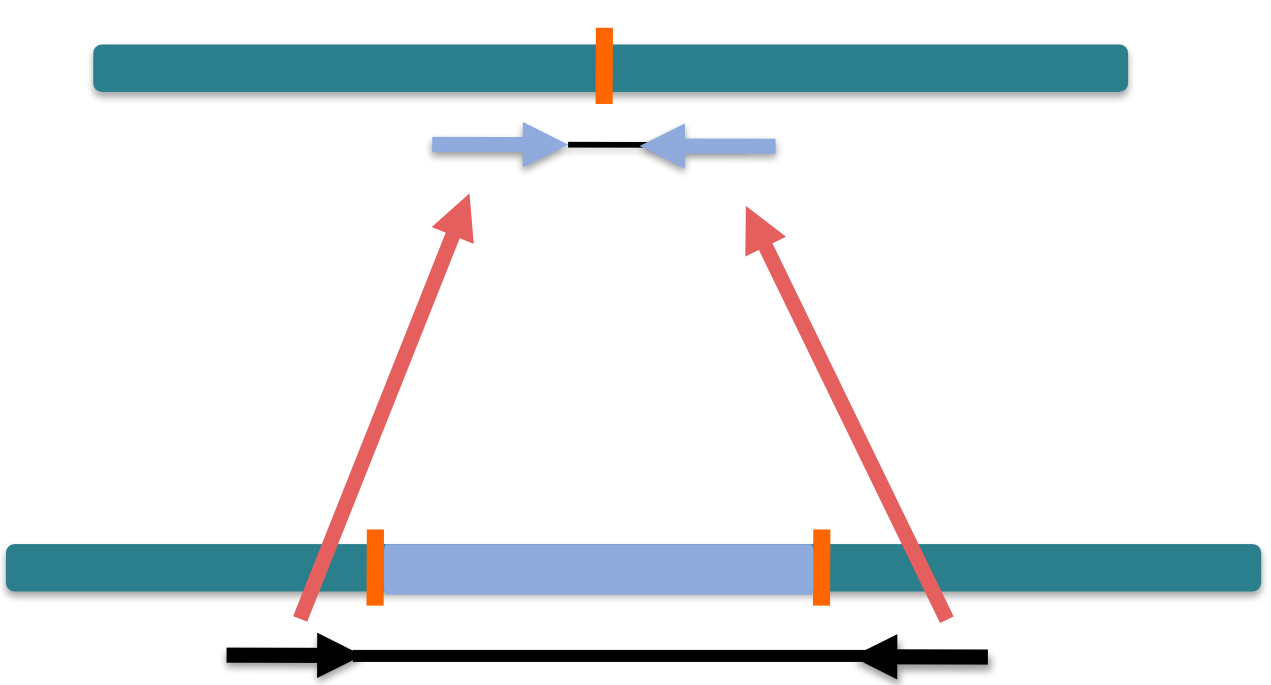
reference genome

subject

Structural variants – Insertion

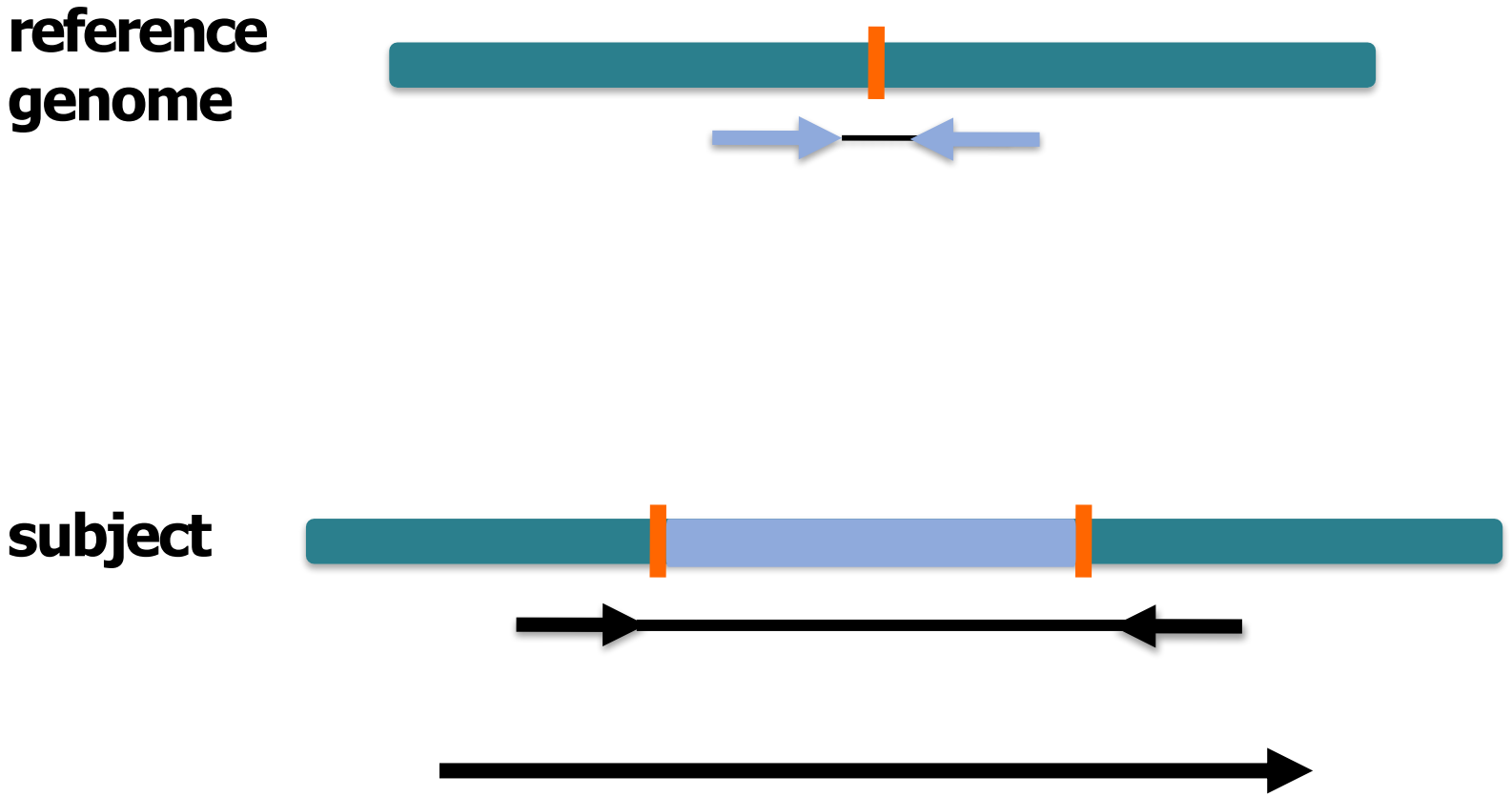
1. **Paired-end mapping**

Shorter insert size than expected is a sign of an insertion



Structural variants – Insertion

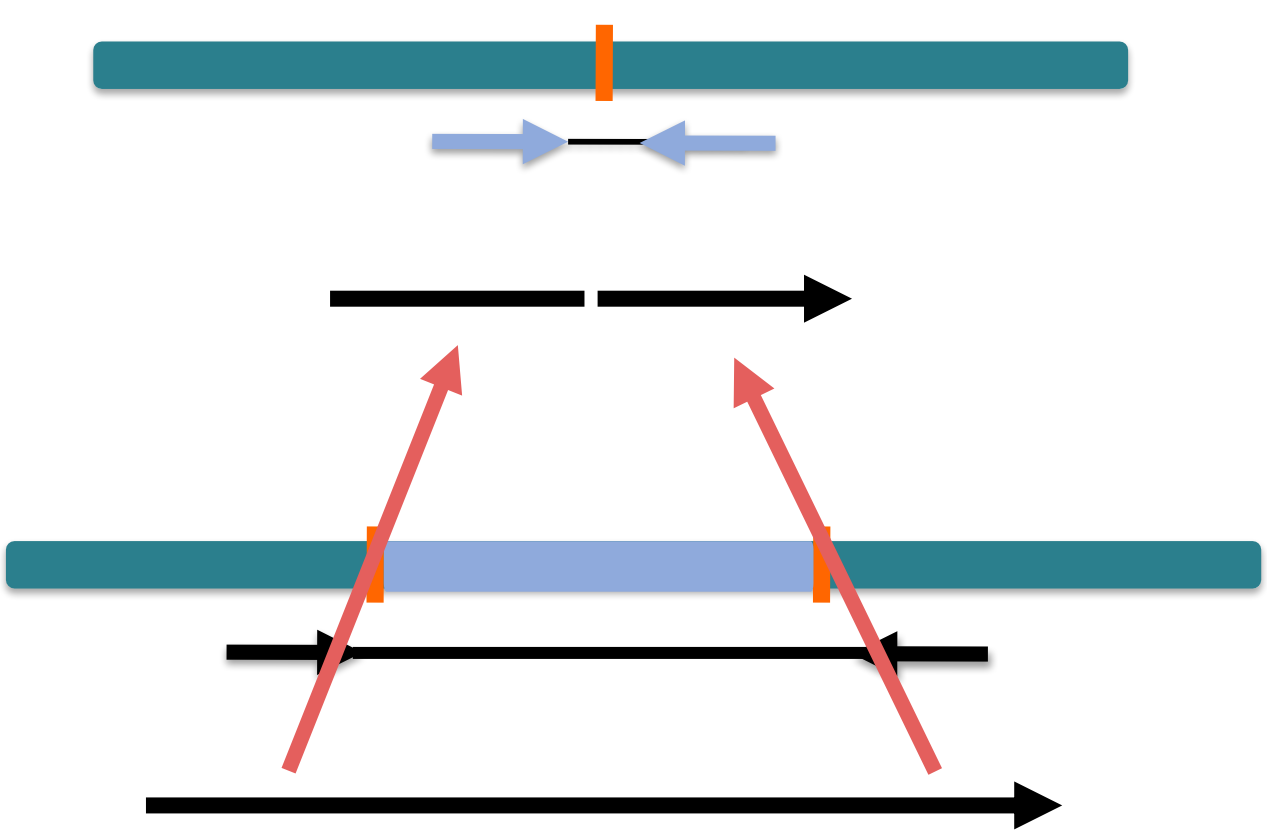
1. Paired-end mapping
2. Split read



Structural variants – Insertion

1. **Paired-end mapping**
2. **Split read**

Split reads with an unmapped region is a sign of a read spanning an insertion

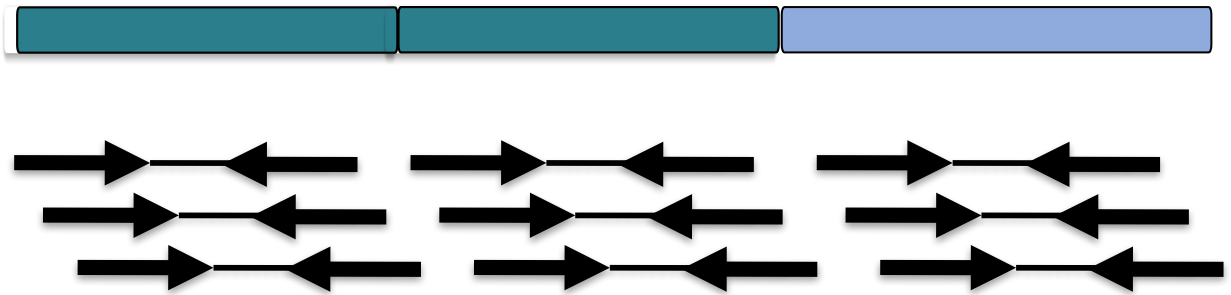


Structural variants - Duplication

1. Coverage

reference genome

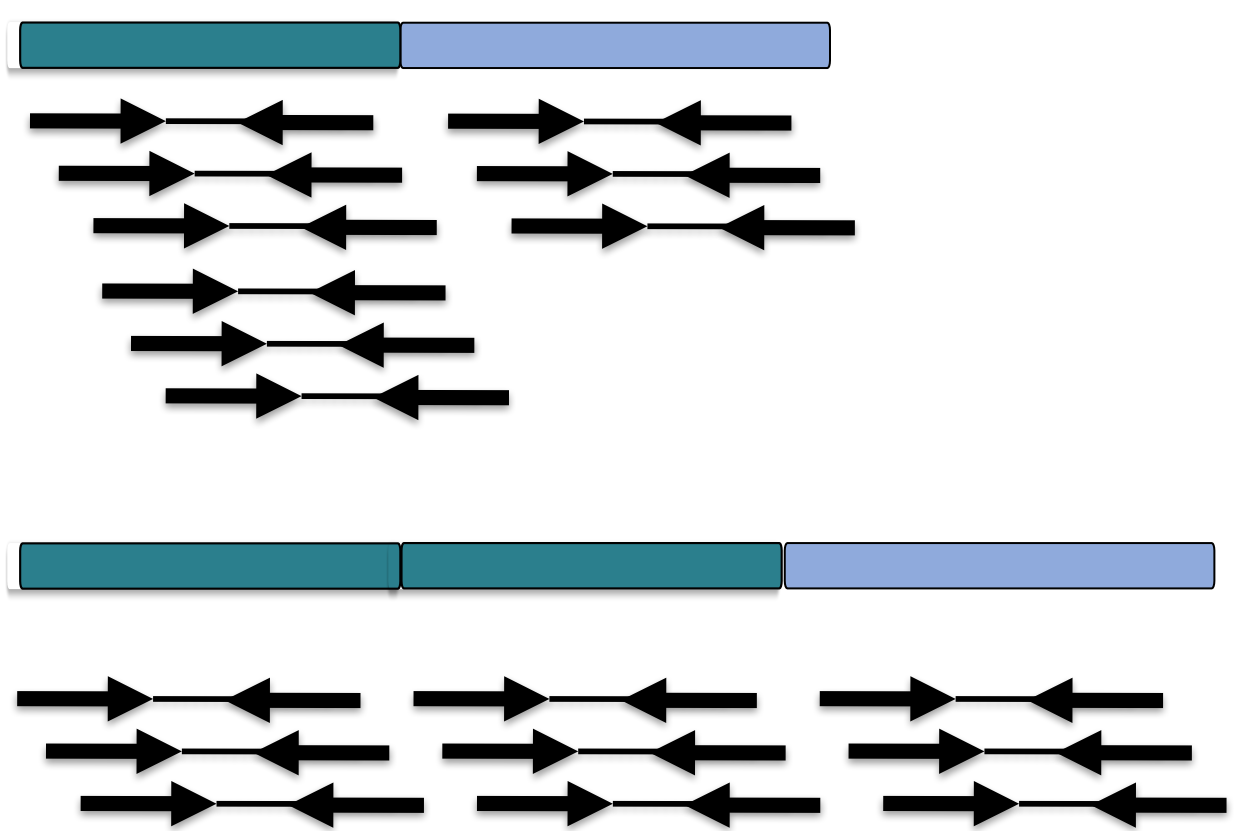
subject



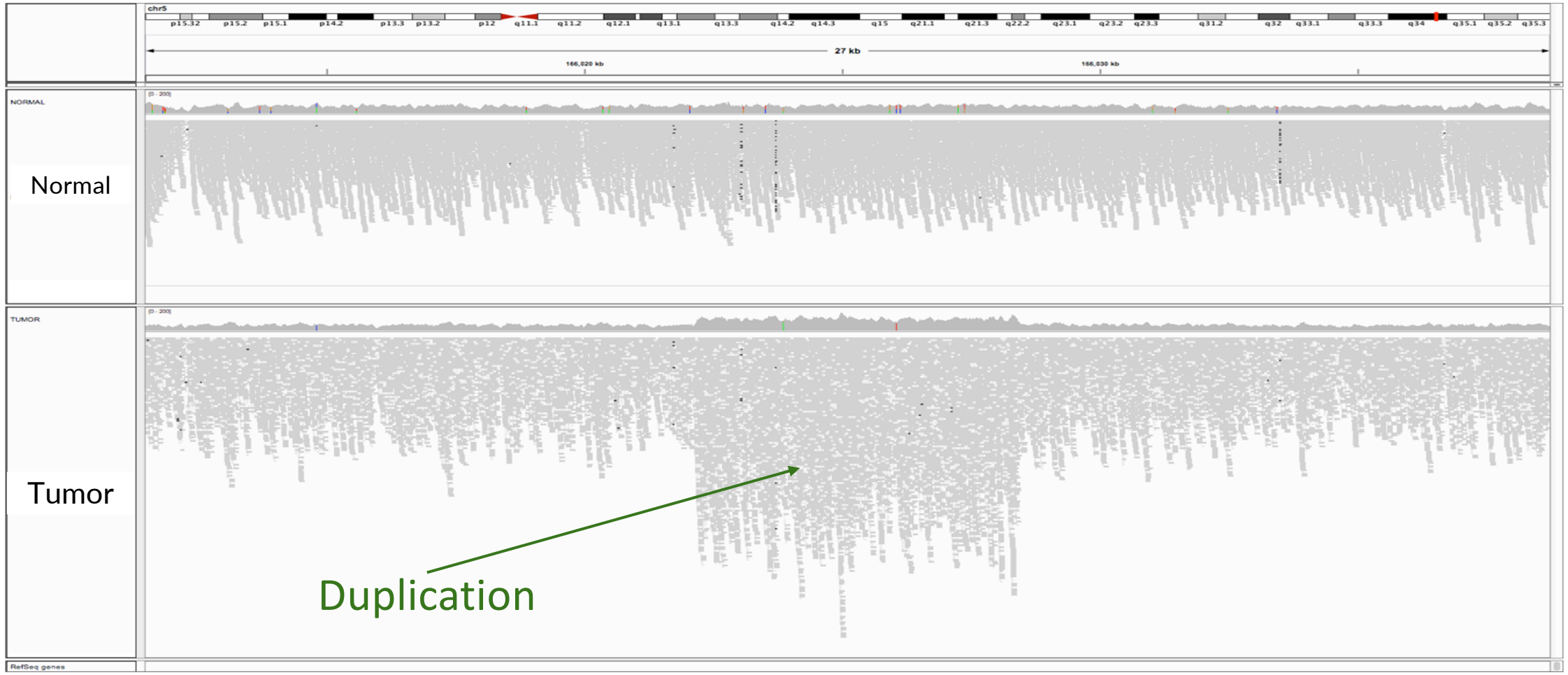
Structural variants - Duplication

1. **Coverage**

Increase in coverage is a sign of a duplication



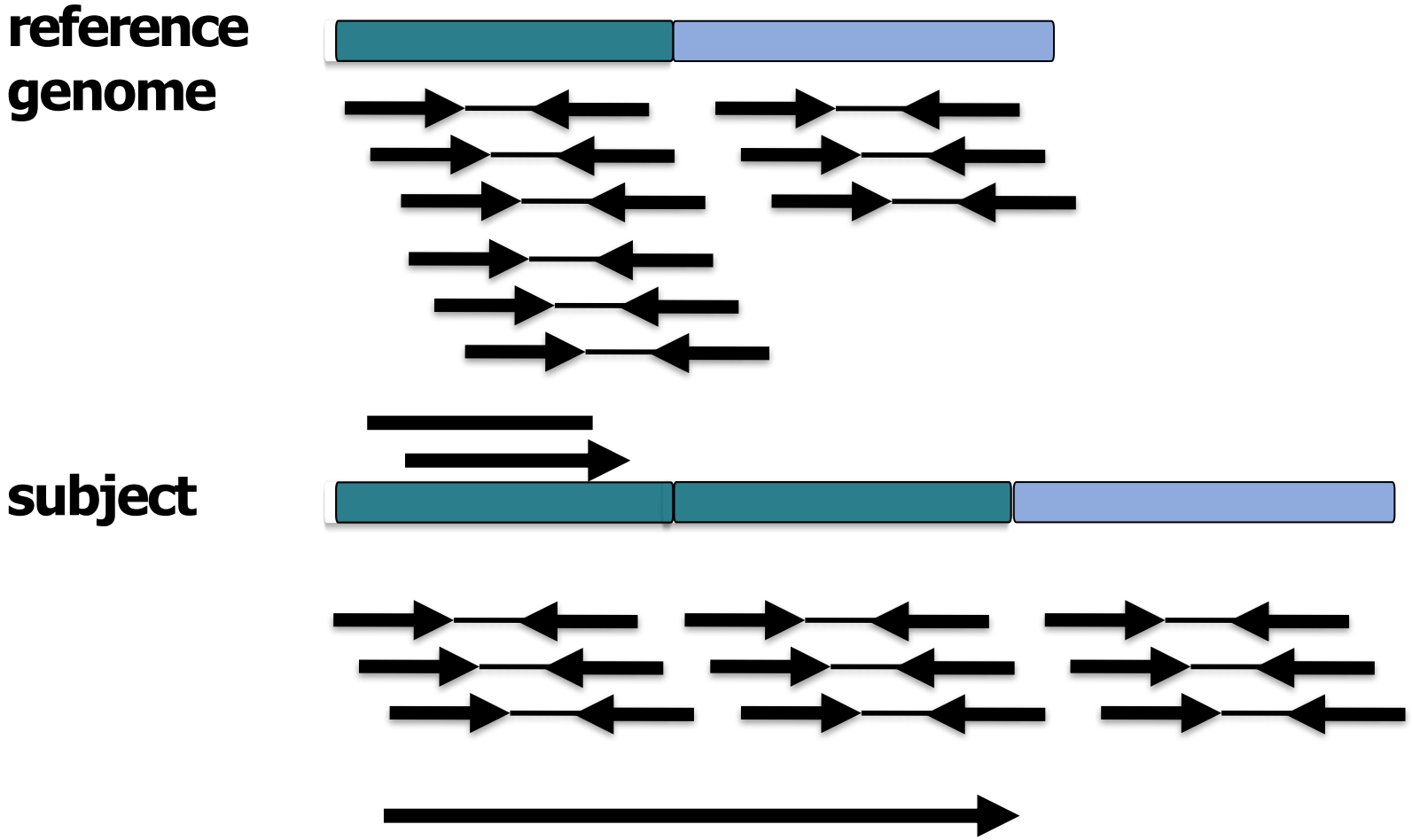
Structural variants – Duplication



Structural variants – Duplication

1. **Coverage**
2. **Split read**

Split reads with overlapping sequences is a sign of a tandem duplication

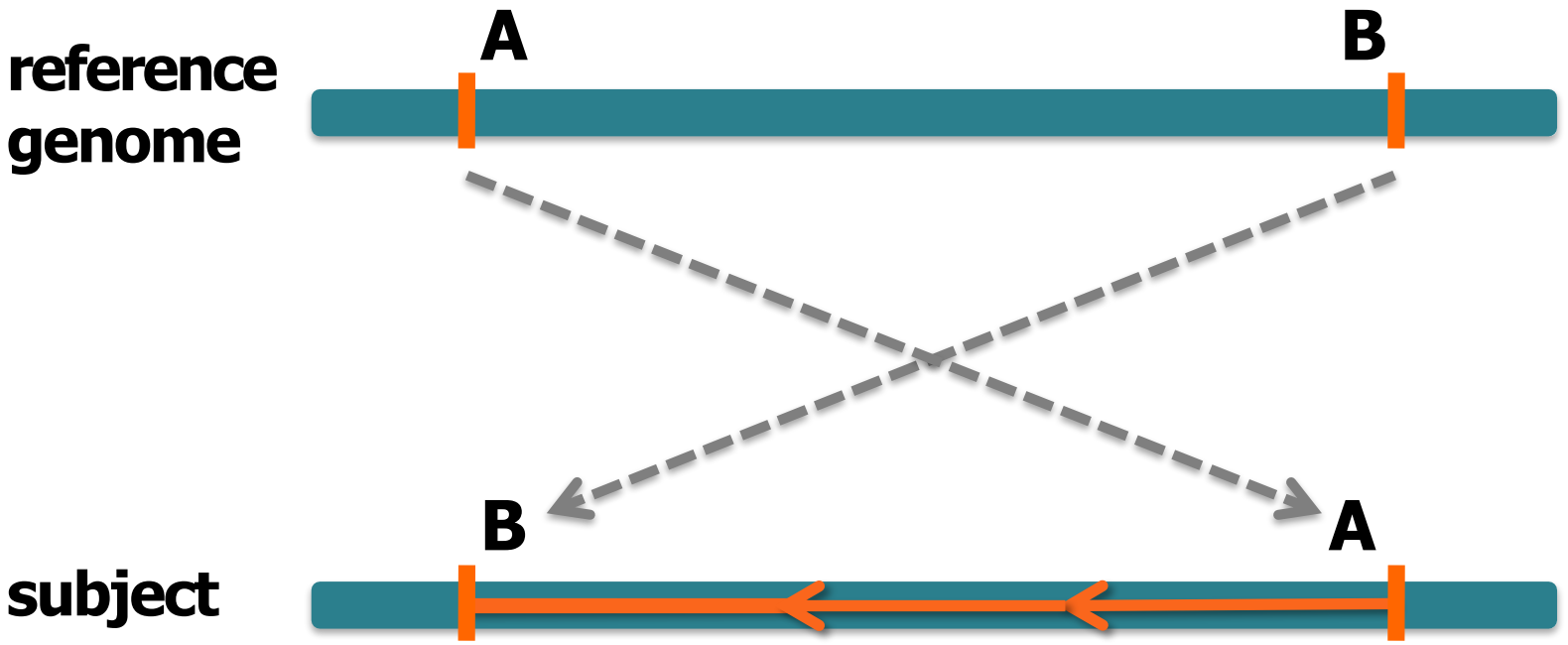


Structural variants

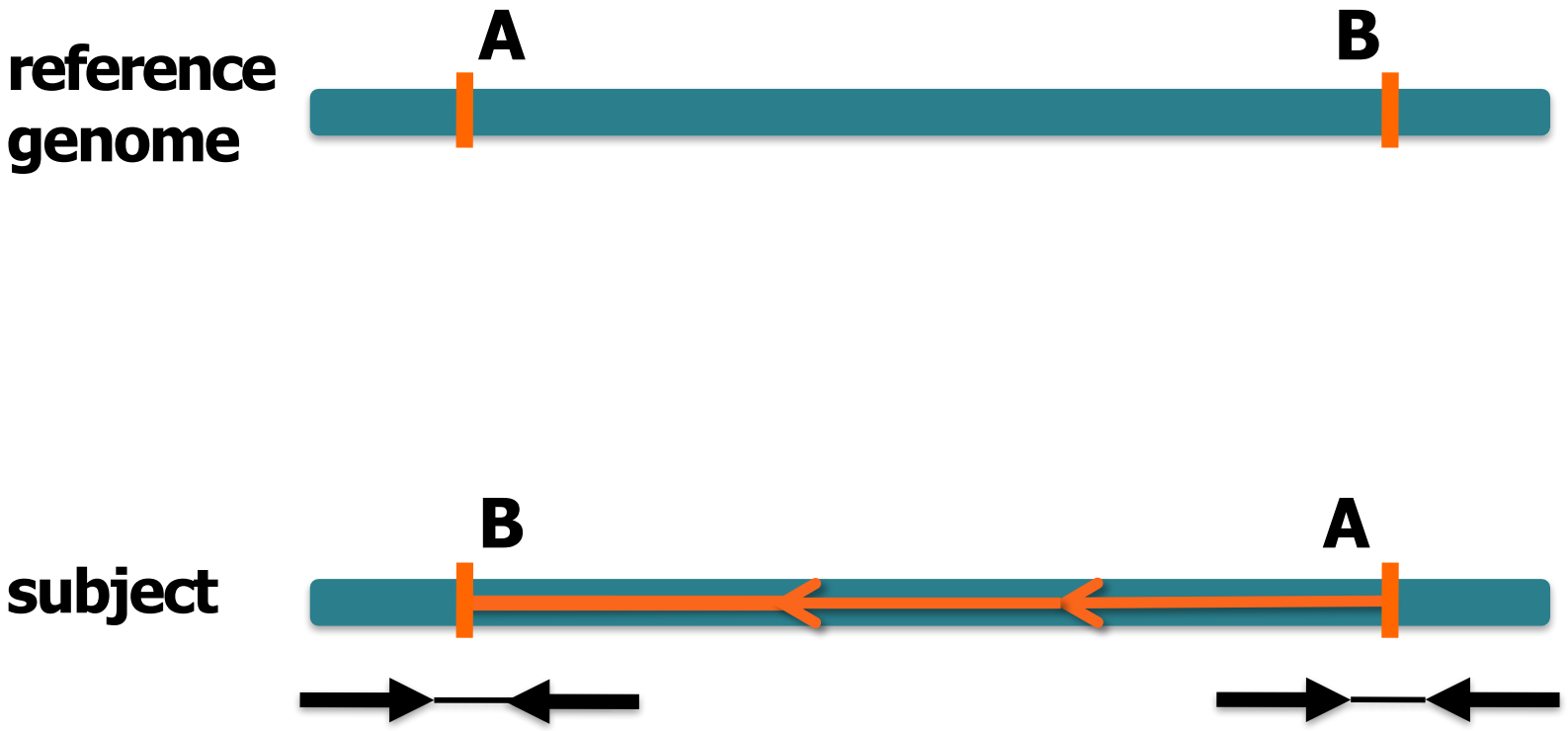
Orientation of paired reads can reveal evidence of structural events, including:

- Inversions
- Duplications
- Translocations

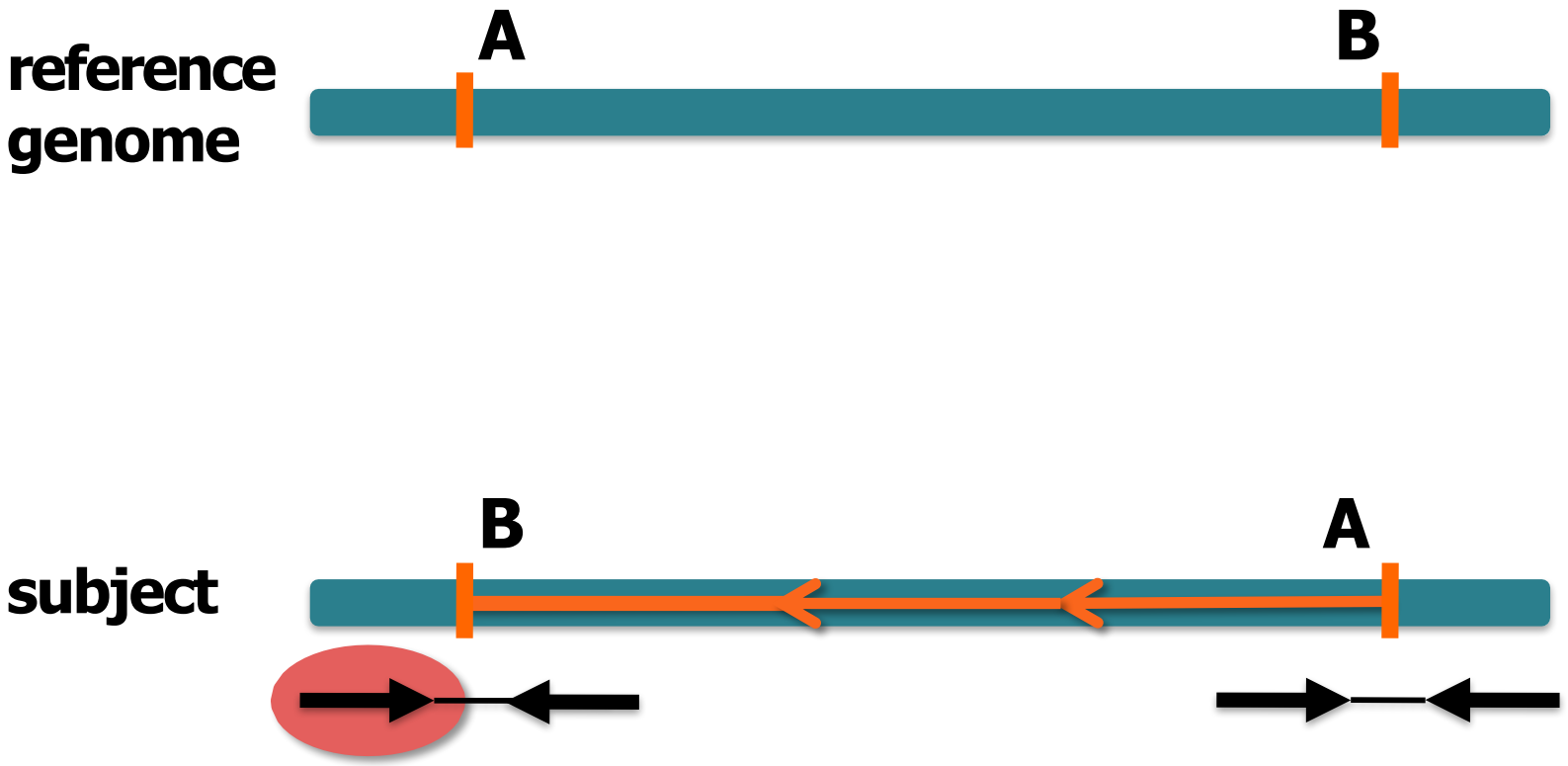
Structural variants - Inversions



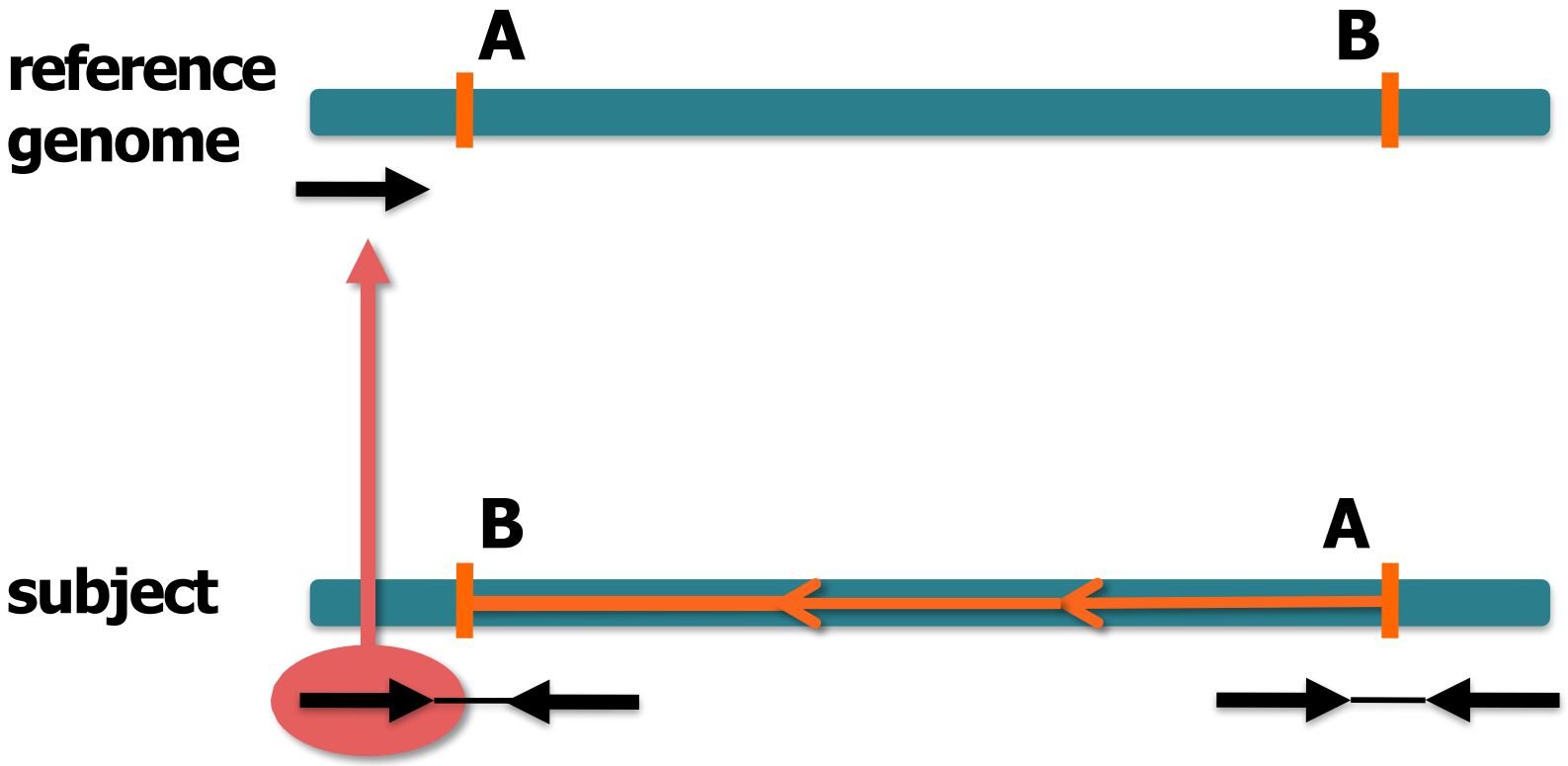
Structural variants - Inversions



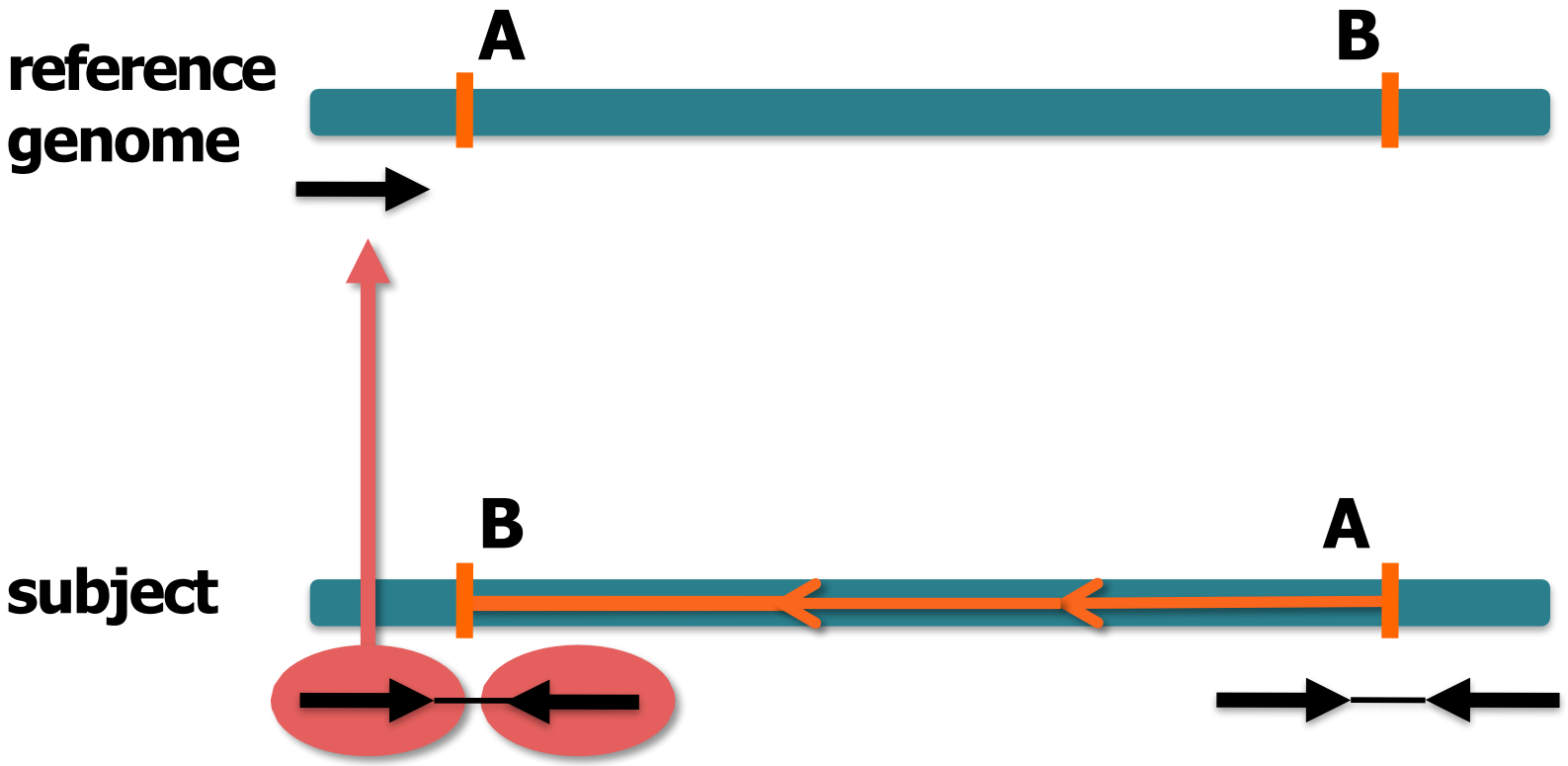
Structural variants - Inversions



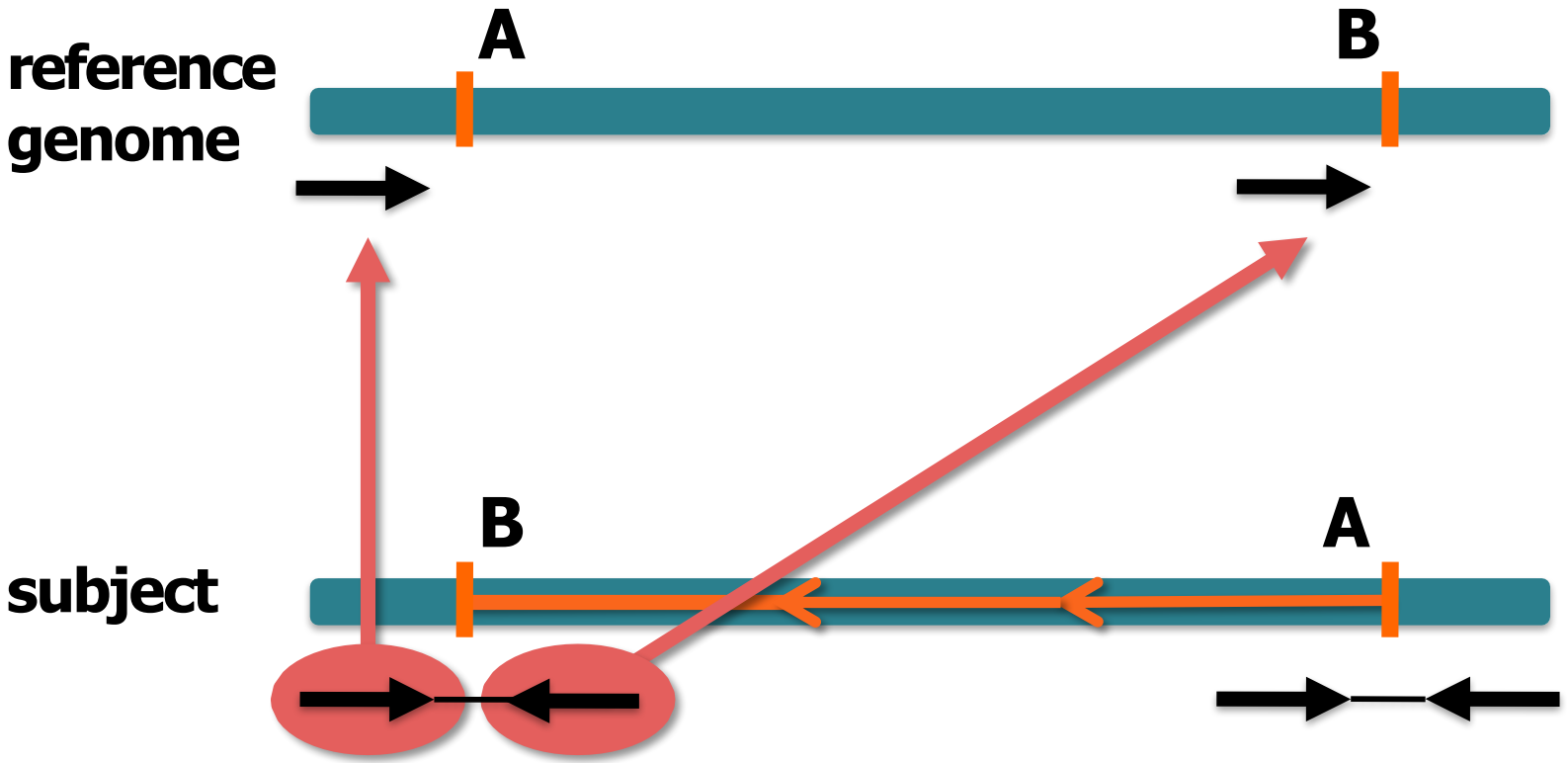
Structural variants - Inversions



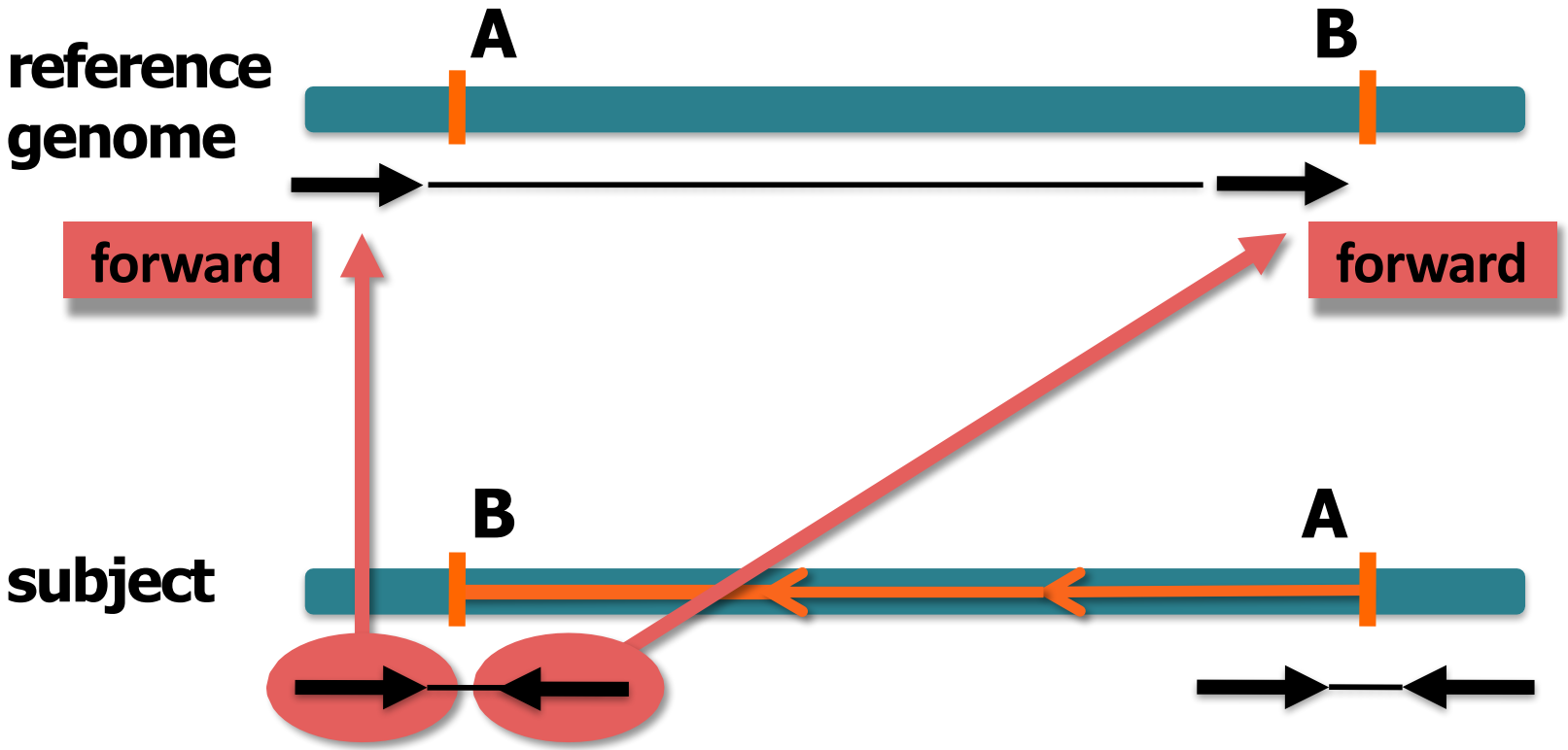
Structural variants - Inversions



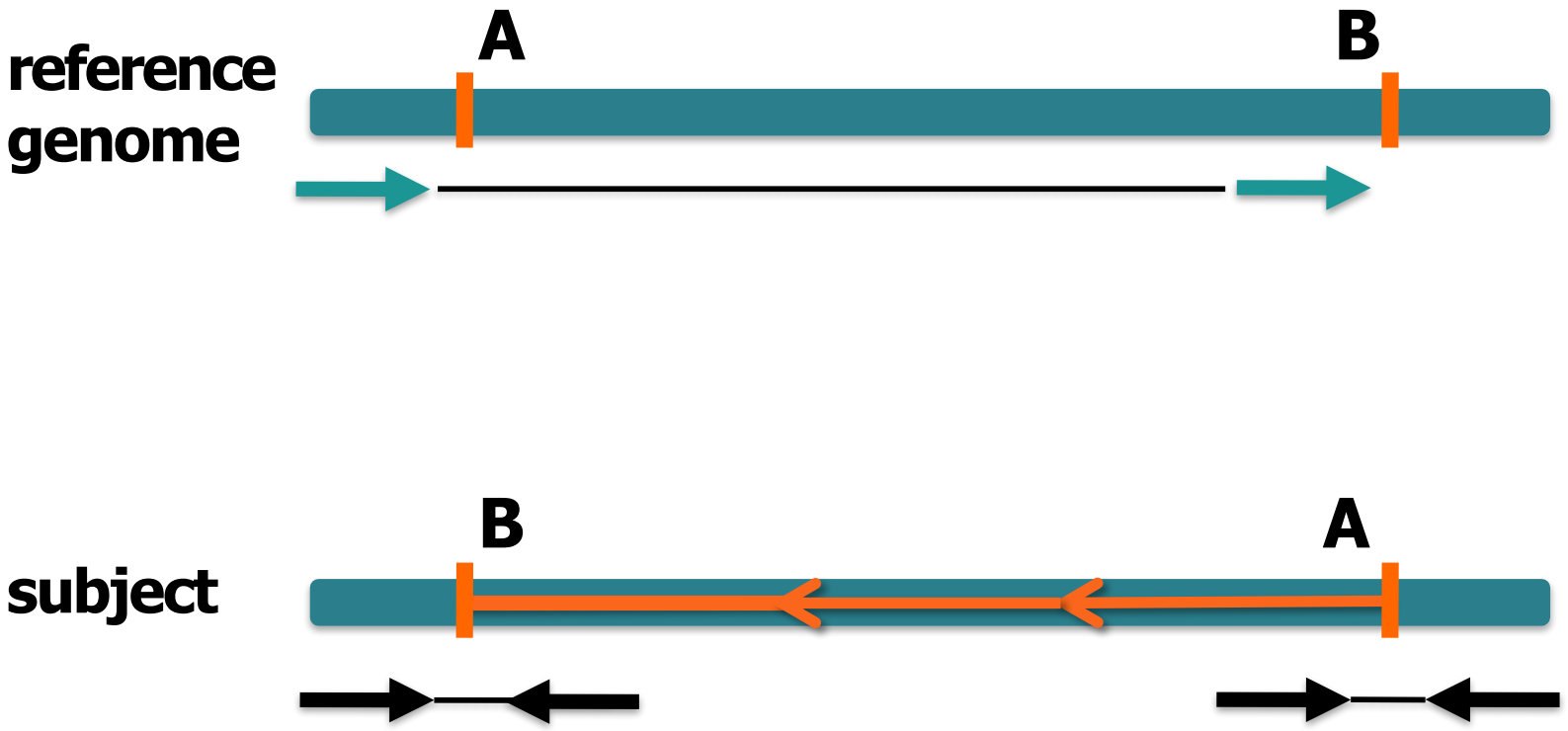
Structural variants - Inversions



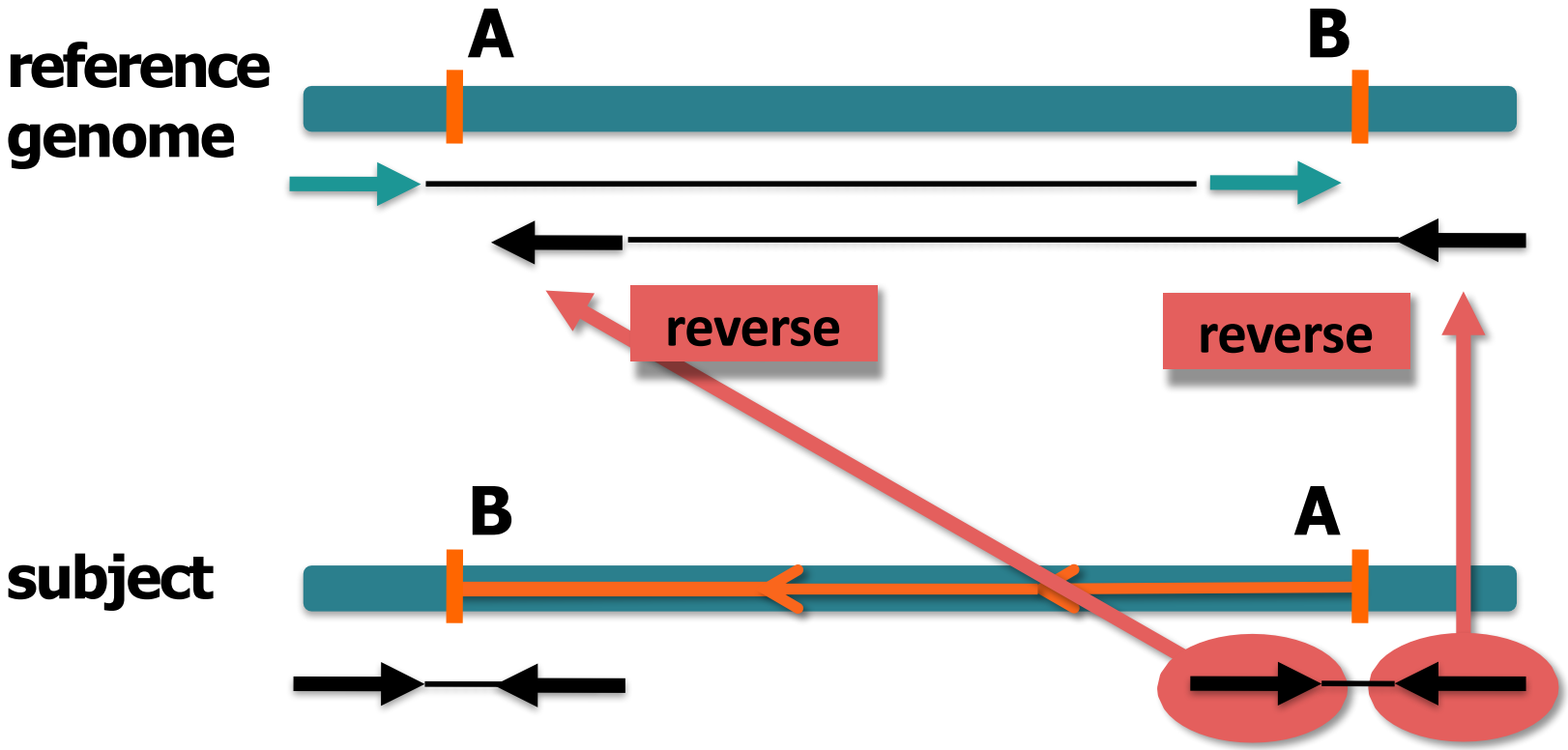
Structural variants - Inversions



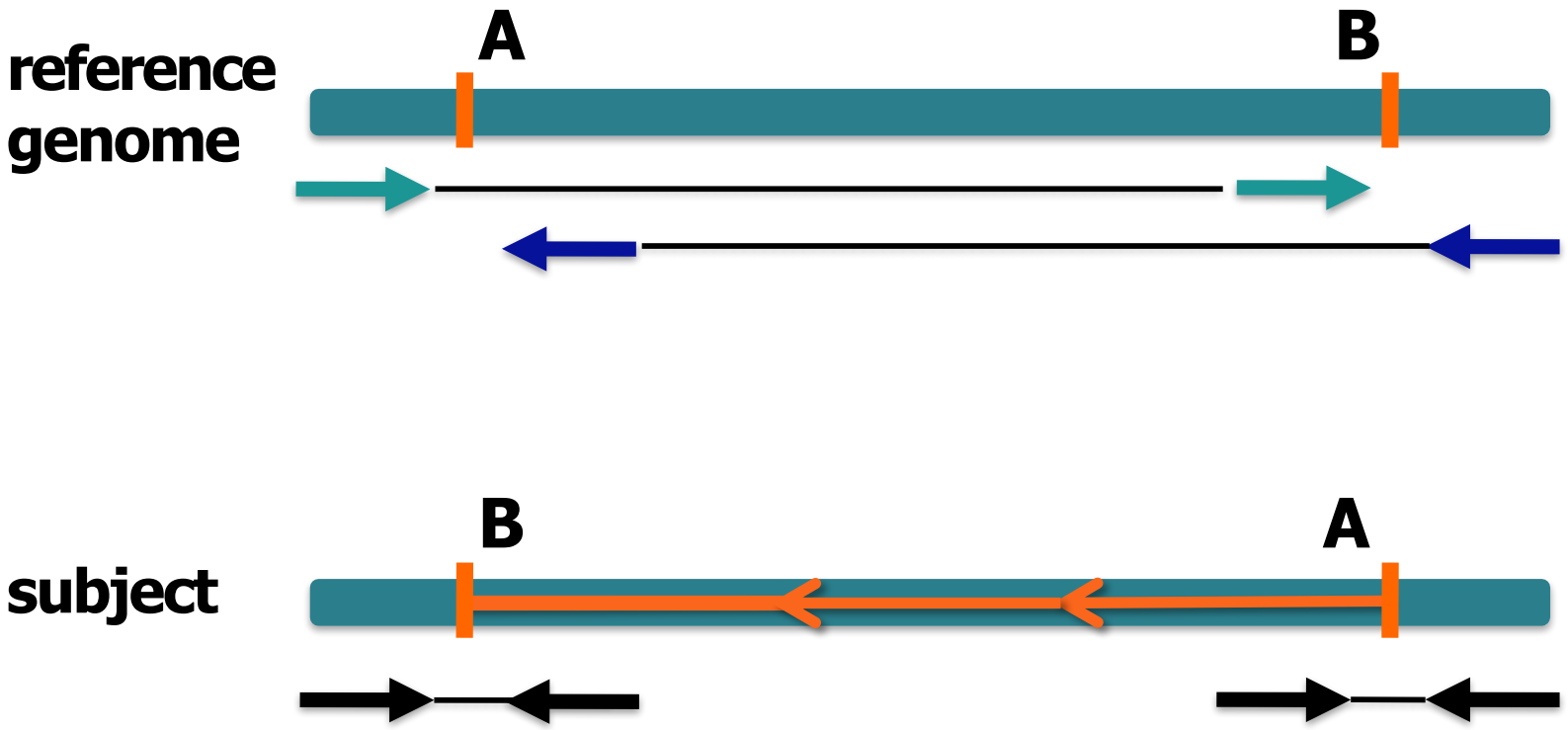
Structural variants - Inversions



Structural variants - Inversions

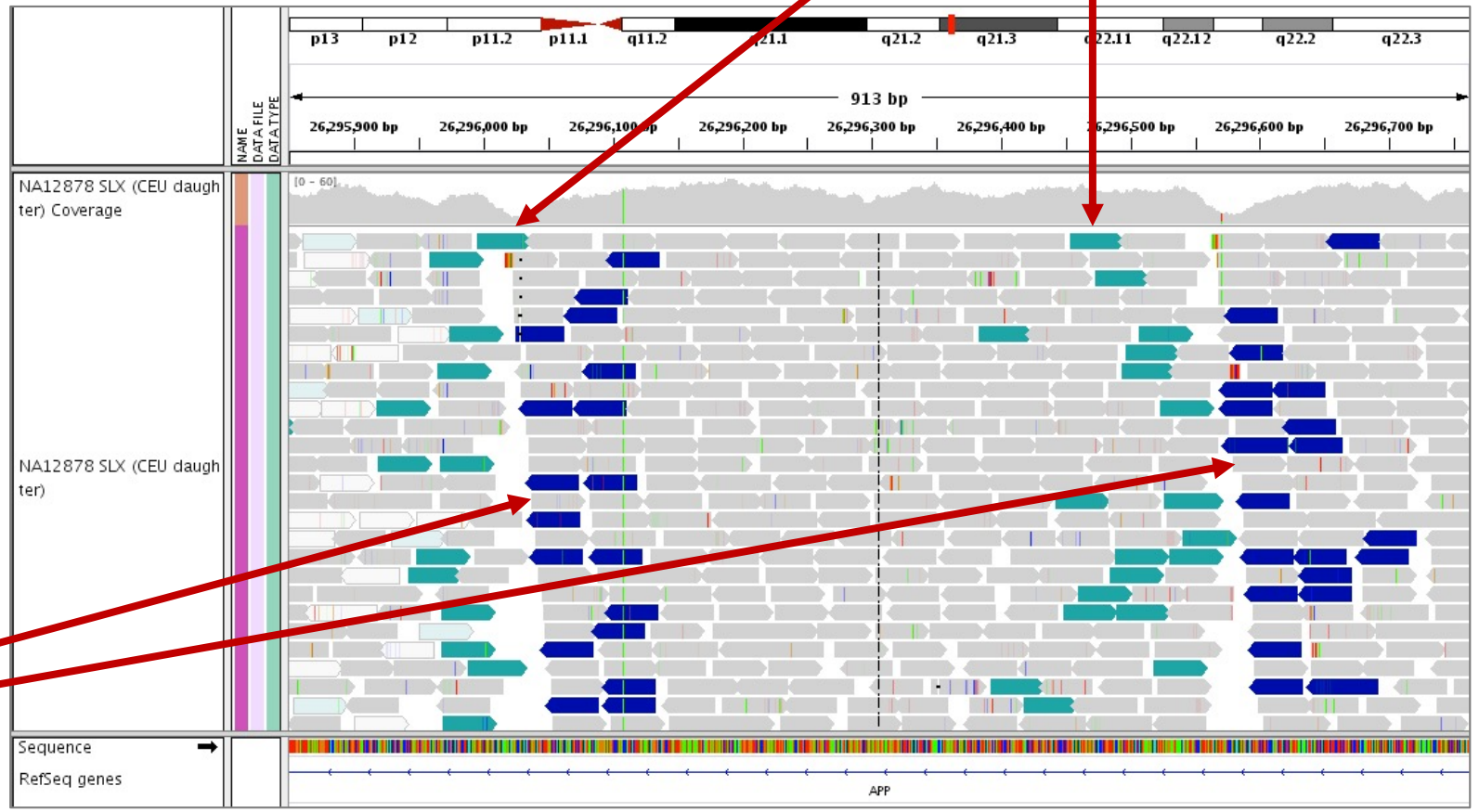


Structural variants - Inversions

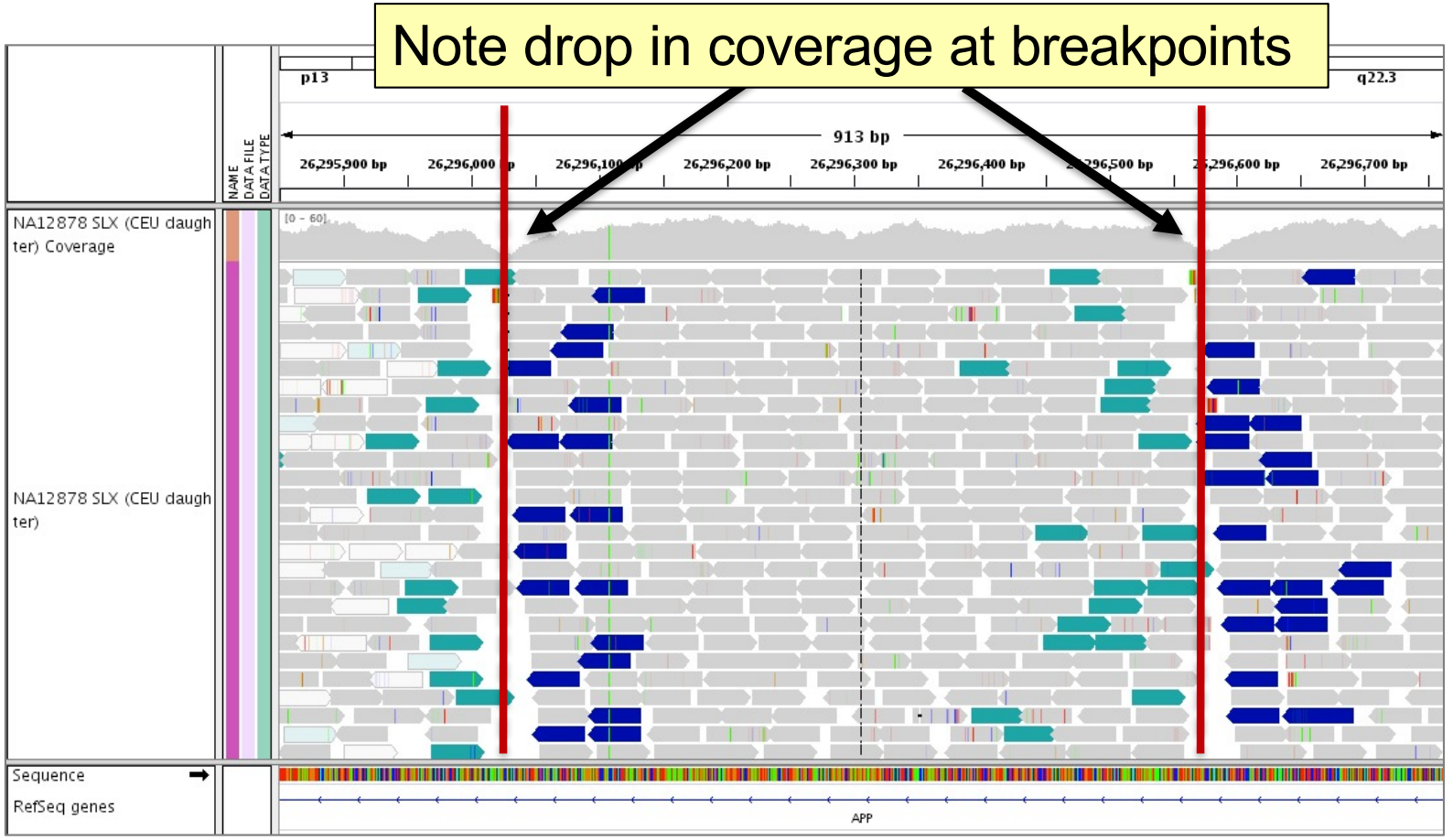


Structural variants - Inversions

Paired reads with the same orientation indicate the breakpoint of an inversion



Structural variants - Inversions

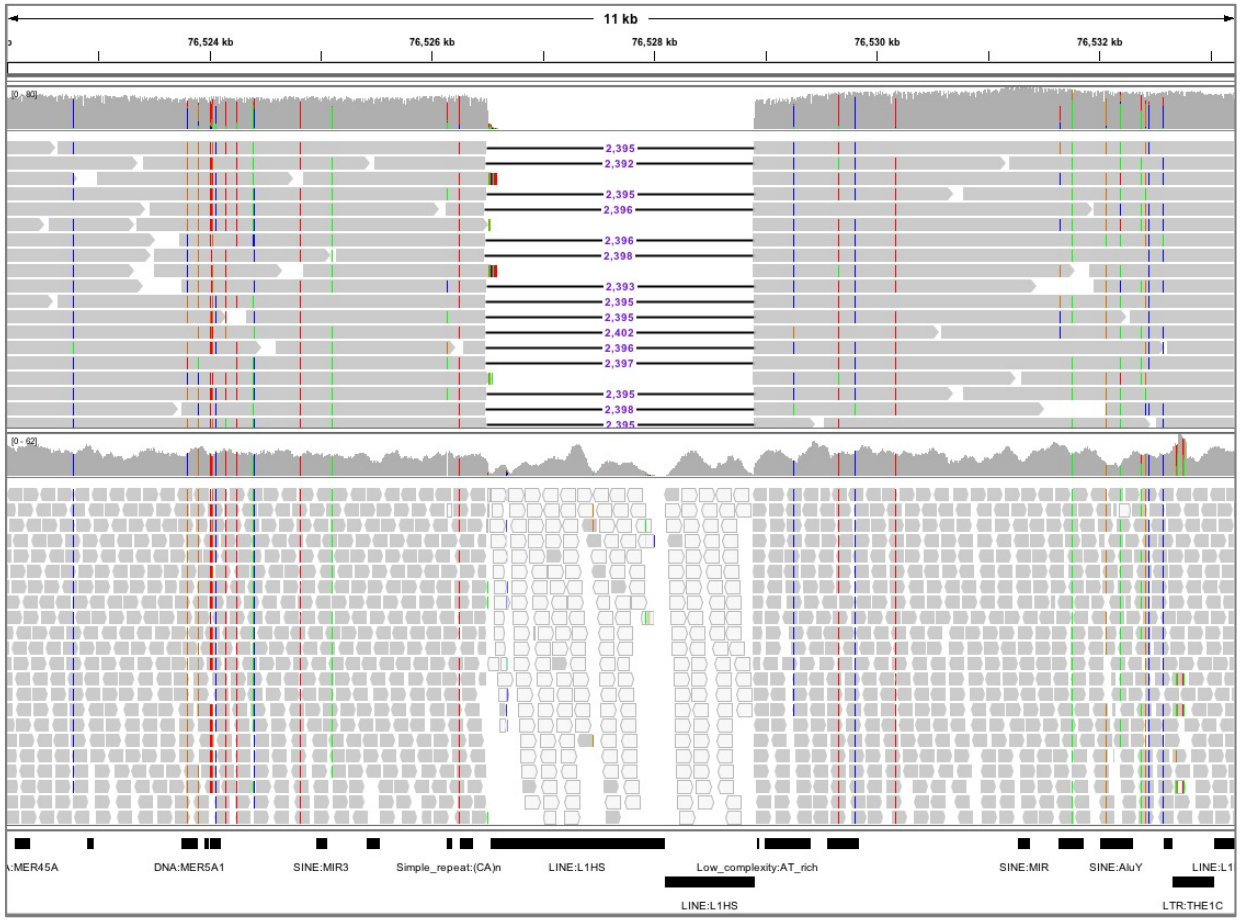


Structural variants - Inversions

Long reads can span long deletions without splitting the read! They will show long gaps

Pacbio

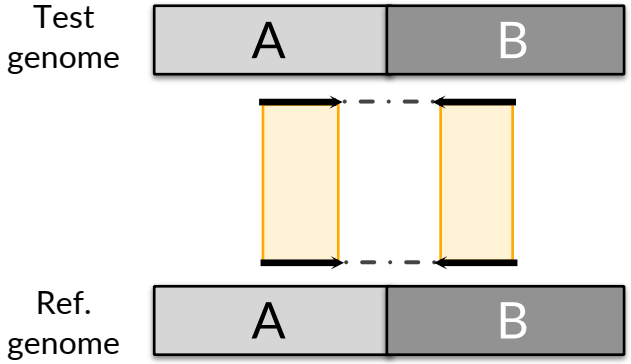
Illumina



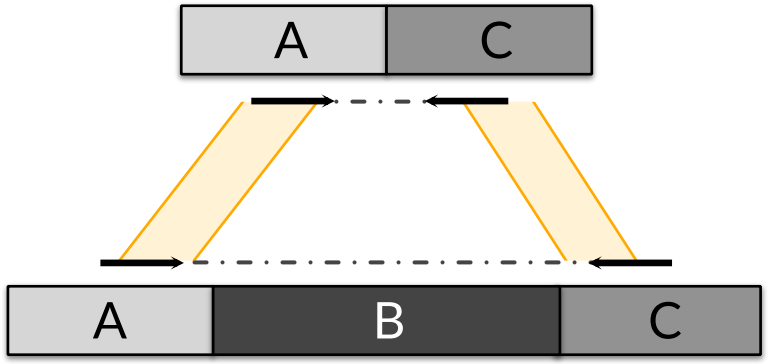
Aaron Wenger, PacBio

Structural variants

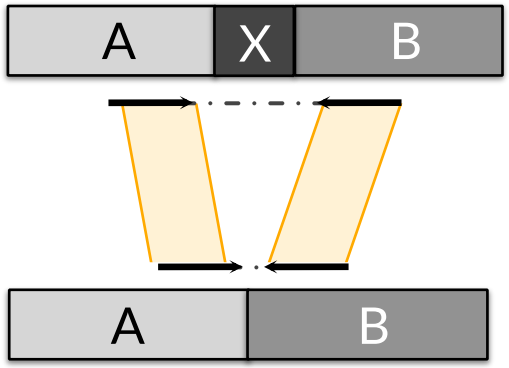
concordant (+/-)



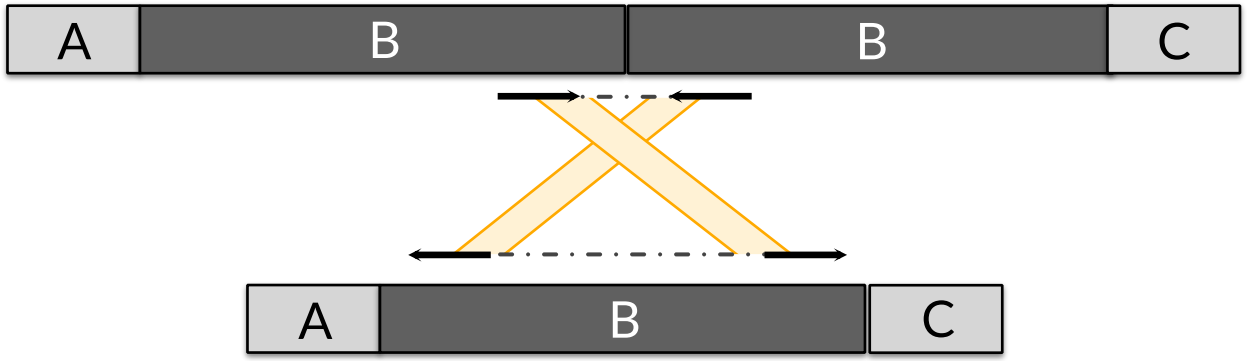
too big (+/-)
= deletion



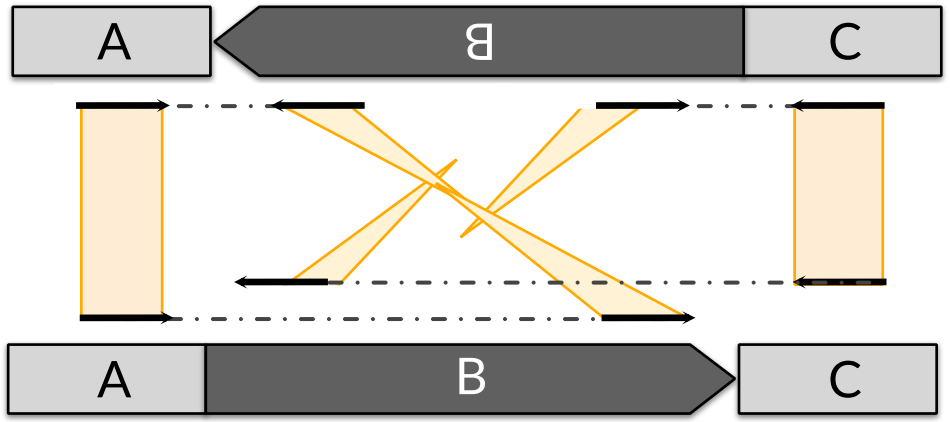
too small (+/-)
= spanned insertion



everted (-/+)
= tandem duplication

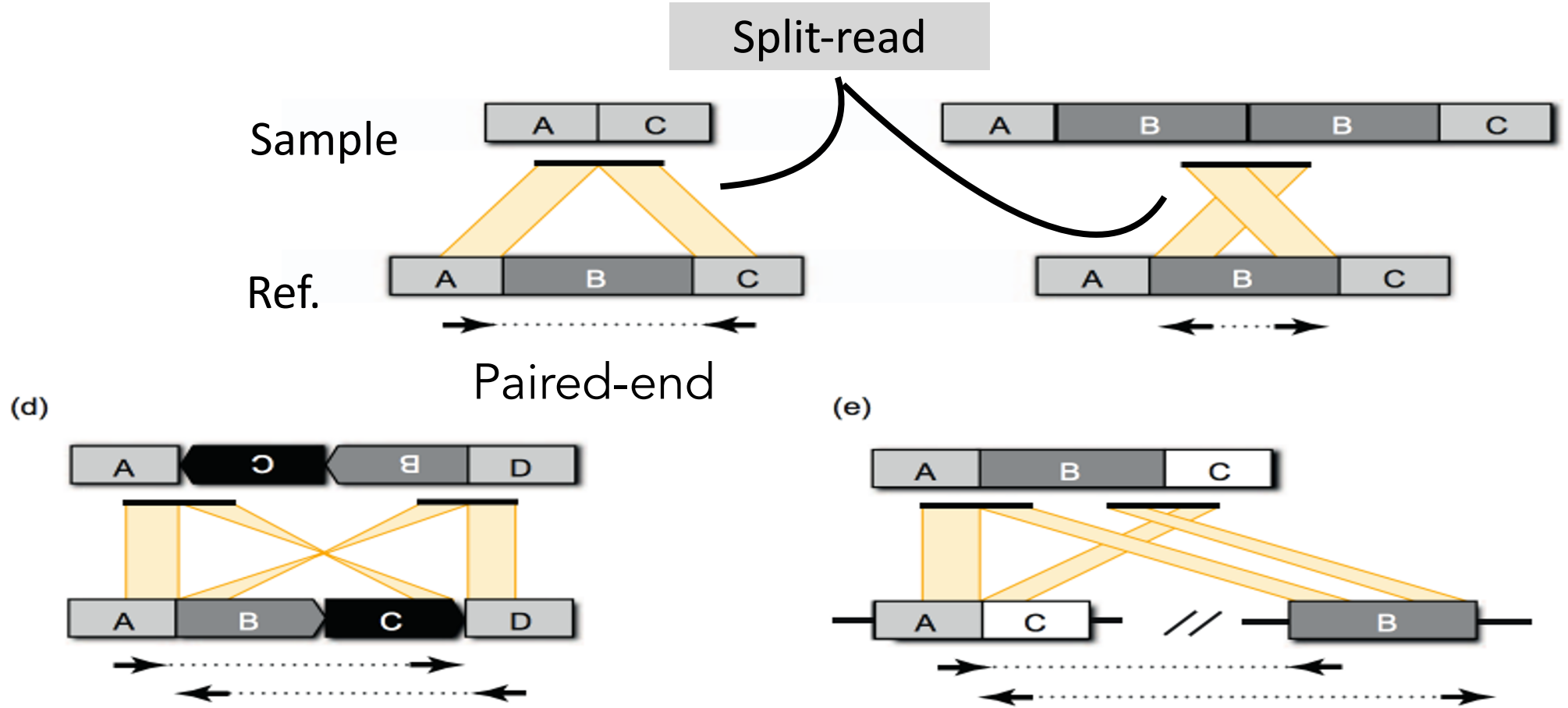


same strand (+/+ or -/-)
= inversion



Quinlan and Hall, 2012

Structural variants



Quinlan and Hall, 2012

Structural variant annotation

```
CGTGTtgtagtaCCGTAA Reference
CGTGT-------CCGTAA Sample
```

Structural variant annotation

```
CGTGTtgtagtaCCGTAA Reference
CGTGT-------CCGTAA Sample
```

1. Direct sequence notation

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SAMPLE
chr1	5	.	TTGTAGTA	T	60	PASS	…	GT	1/1

Structural variant annotation

```
CGTGTtgtagtaCCGTAA Reference
CGTGT-------CCGTAA Sample
```

1. Direct sequence notation

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SAMPLE
chr1	5	.	TTGTAGTA	T	60	PASS	…	GT	1/1

2. Symbolic notation

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SAMPLE
chr1	5	.	T	<DEL>	60	PASS	…	GT	1/1

Structural variant annotation

```
CGTGTtgtagtaCCGTAA Reference
CGTGT-------CCGTAA Sample
```

1. Direct sequence notation

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SAMPLE
chr1	5	.	TTGTAGTA	T	60	PASS	…	GT	1/1

2. Symbolic notation

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SAMPLE
chr1	5	.	T	<DEL>	60	PASS	SVLEN=7	GT	1/1

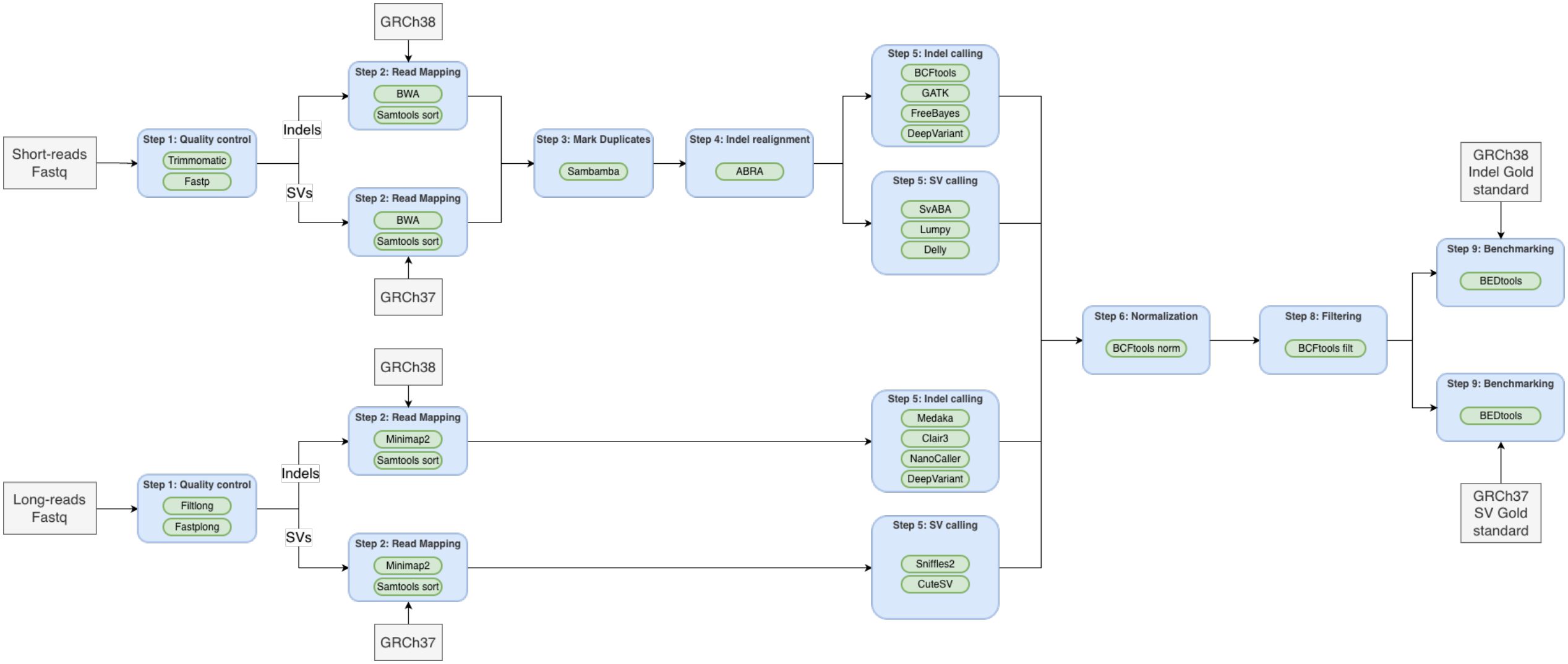
Assignment 2

We will compare Indel and SV calling for three sequencing technologies:

1. Illumina
2. PacBio
3. Nanopore

The benchmarking will be done using the well-characterized HG002

Assignment 2



Assignment 2

Chromosome	Position	ID	Reference	Alternate	Quality	Filter	Info	Format	Genotype
1	10415		ACCCTAACCCTAACCCTAACCCTAAC	A	.	.	…	GT	1/1
1	62297		T	TCTTC	.	.	…	GT	1/1

Assignment 2 – Intersecting SV

Step 1: Get VCF file to BED file

VCF

Chromosome	Position	ID	Reference	Alternate	Quality	Filter	Info	Format	Genotype
1	10415		ACCCTAACCCTAACCCTAACCCTAAC	A	.	.	…	GT	1/1
1	62297		T	TCTTC	.	.	…	GT	1/1

BED

10414	10440
62296	62297

Assignment 2 – Intersecting SV

Step 2: Get benchmark VCF file to BED file

VCF

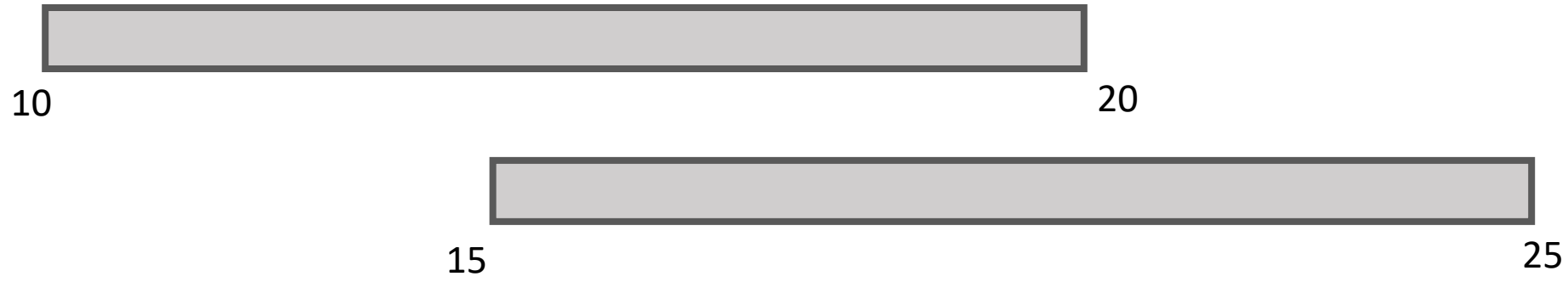
Chromosome	Position	ID	Reference	Alternate	Quality	Filter	Info	Format	Genotype
1	10415		ACCCTAACCCTAACCCTAACCCTAAC	A	.	.	…	GT	1/1
1	62297		T	TCTTC	.	.	…	GT	1/1

BED

10414	10440
62296	62297

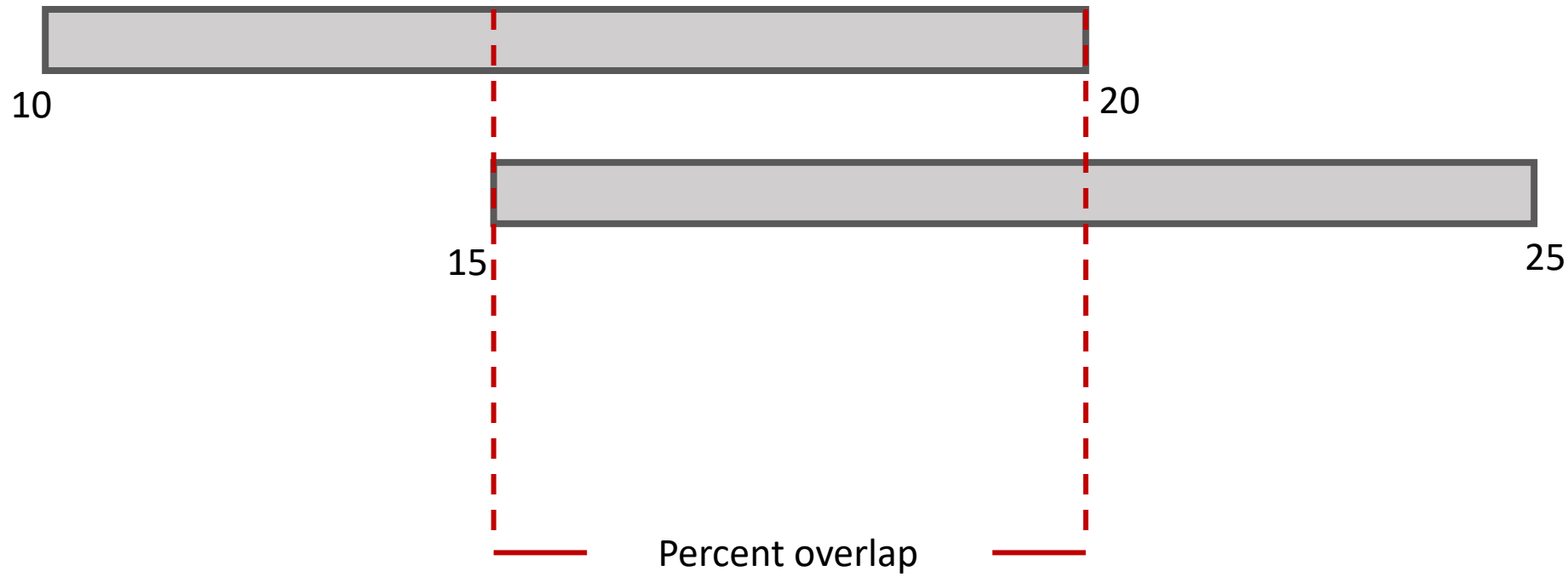
Assignment 2 – Intersecting SV

Step 3: Intersect coordinates of SV



Assignment 2 – Intersecting SV

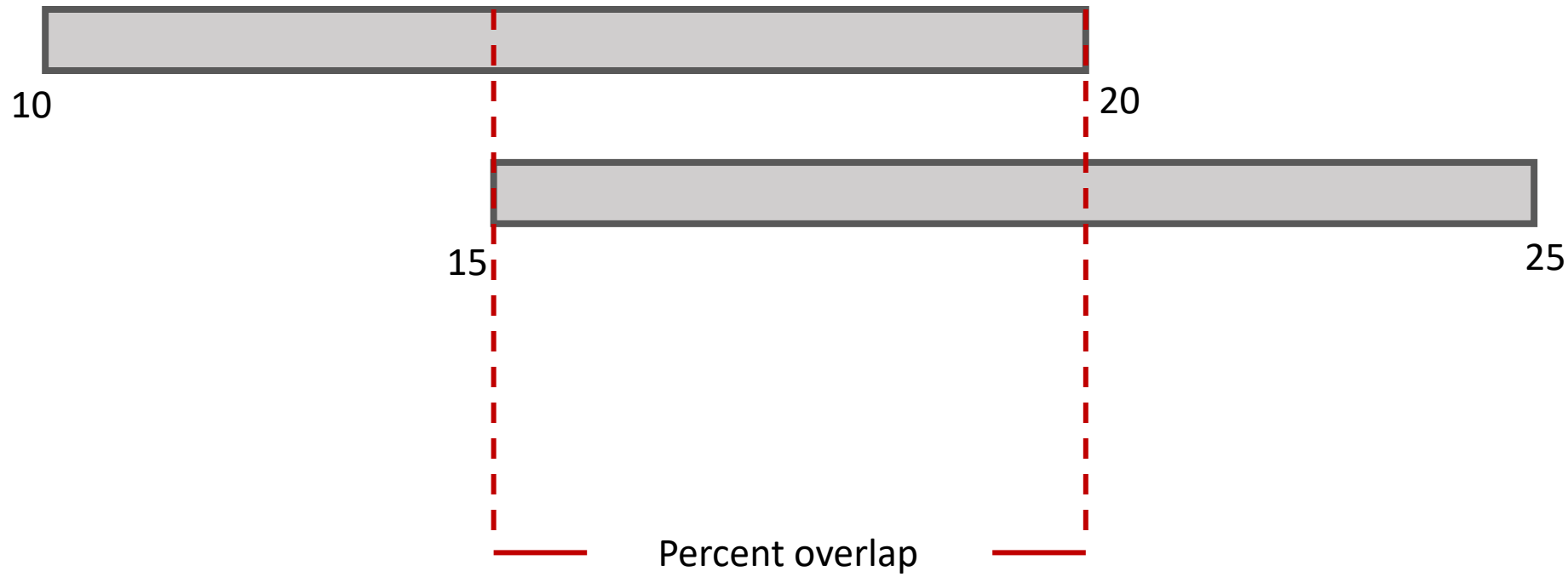
Step 3: Intersect coordinates of SV



If percent overlap > x% -> Same SV

Assignment 2 – Intersecting SV

Step 3: Intersect coordinates of SV

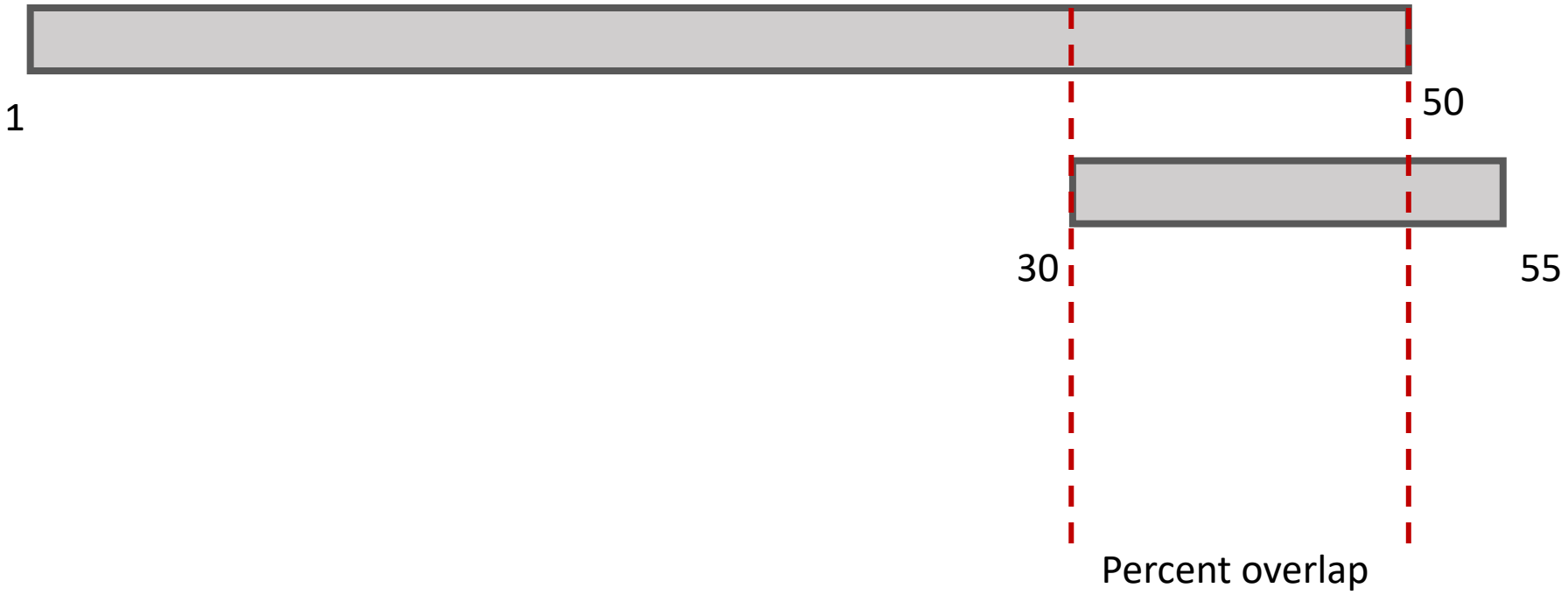


If percent overlap > x% -> Same SV

* Reciprocal overlap – Both Variants need to overlap with each other

Assignment 2 – Intersecting SV

Step 3: Intersect coordinates of SV

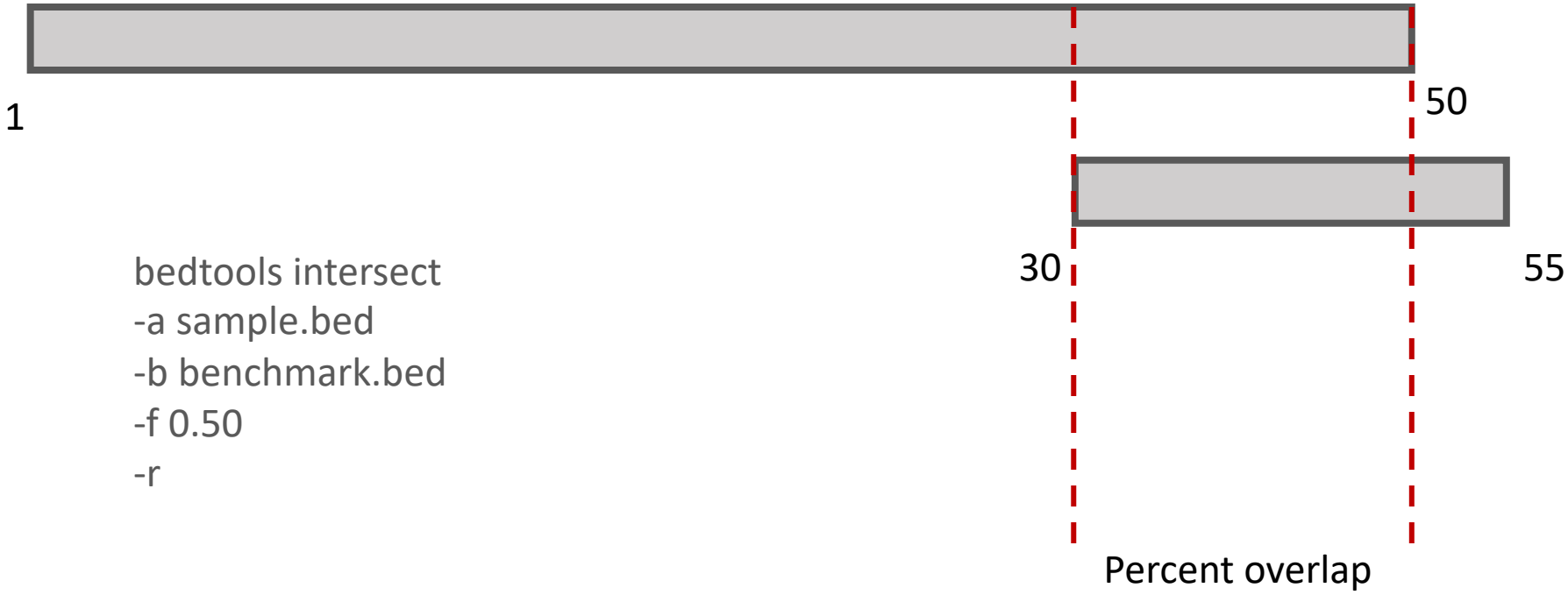


If percent overlap > x% -> Same SV

* Reciprocal overlap – Both Variants need to overlap with each other

Assignment 2 – Intersecting SV

Step 3: Intersect coordinates of SV



bedtools intersect
-a sample.bed
-b benchmark.bed
-f 0.50
-r

If percent overlap > x% -> Same SV

* Reciprocal overlap – Both Variants need to overlap with each other

Filtering SV by length

- Most structural variant callers will add an INFO/SVLEN field that can be used for filtering SVs by length

Filtering Indels by length

- BCFtools can be used to easily filter Indels by length using ”ILEN”
- ILEN will count up the length of the indel for you: positive numbers are for insertions, and negative are deletions.

```
# To remove indels smaller than 5bp:
bcftools view -i '(ILEN >= -5 && ILEN <= 5)’
```

TYPE is a built-in that will determine whether a variant is an indel, snp, etc. on-the-fly, so you can use the OR logic to include variants that aren't indels (i.e. snps) using the second command. Note that otherwise, ILEN explicitly excludes anything that's not an indel.

```
# To remove indels smaller than 5bp and preserve SNPs
Bcftools view -i '(ILEN >= -5 && ILEN <= 5) || TYPE!="INDEL”’
```