HEIDELBERG
UNIVERSITY
HOSPITAL

GERMAN
CANCER RESEARCH CENTER
IN THE HELMHOLTZ ASSOCIATION

Research for a Life without Cancer

Key Terms in Paired-End Sequencing and Structural Variation Analysis

PhD Student Quynh Nhu Nguyen

04.05.2025

Primary alignment, Secondary alignments, Supplement alignments

- **Primary Alignment**: The best or most confident alignment of a read to the reference genome. Only one per read.

- **Secondary Alignment**: An alternative alignment for a read, usually when it maps to multiple locations with similar quality.

- **Supplementary Alignment**: Part of a read that aligns to different genomic locations, typically for split reads or structural variations.

Soft Clipping and Hard Clipping

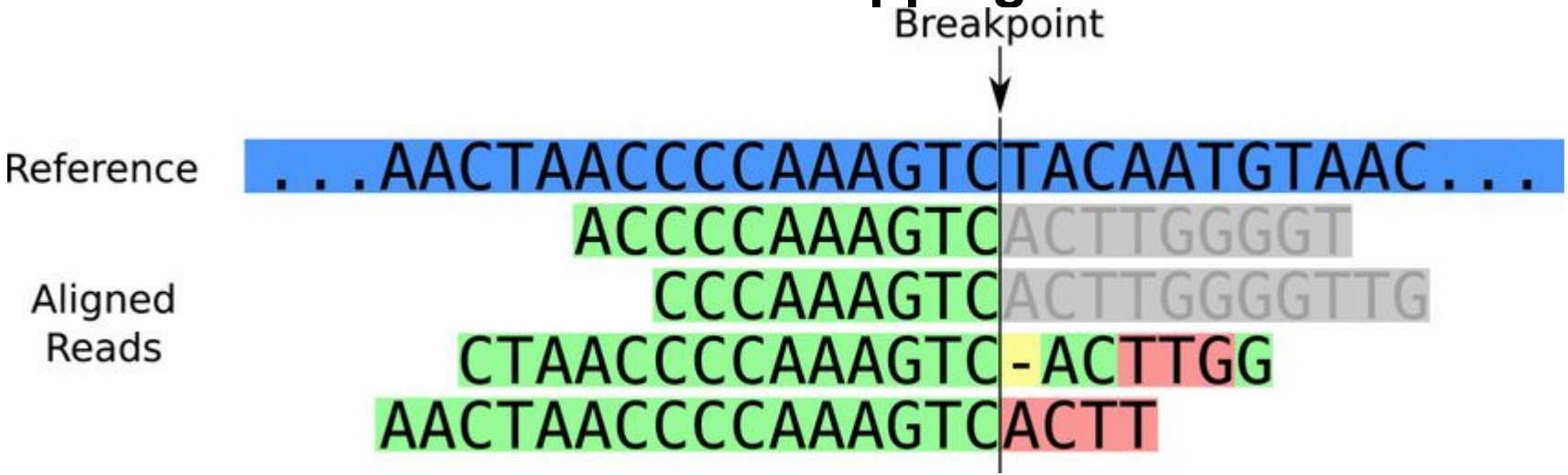
Soft Clipping:

- Refers to trimming bases from the ends of a read that do not align to the reference genome.
- The clipped bases are kept in the alignment file but marked as unaligned. They are not considered for downstream analysis like variant calling.
- Common in situations where part of the read aligns well but the ends do not (e.g., in cases of sequencing errors or low-quality regions).

Hard Clipping:

- Refers to removing bases from the ends of a read that do not align to the reference genome.
- The clipped bases are completely discarded from the alignment, meaning they are not present in the alignment file.
- Often used when the clipped sequence is considered irrelevant or unreliable for analysis.

Soft Clipping



- Blue: reference, Green: match, Pink: mismatch, Yellow: skip. Gray: soft-clipped
- Breakpoint: the exact position where a structural variant (like insertion, deletion, or translocation) occurs in the genome.
- Top two reads are correctly soft-clipped at the breakpoint. Bottom two misalign with mismatches/skips due to short overhangs and global alignment.

Landman et al., BMC Genomics, 2014; Fiedler et al., BMC Bioinformatics, 2023

Soft Clipping in CIGAR string

```
Ref   CTGGCCATTATCTC--GGTGGTAGGACATGGCATGCCC
Read        aaATGTCGCGGTG.TAGGAggatcc
```

2S5M2I4M1D5M6S

Op	BAM	Description
M	0	alignment match (can be a sequence match or mismatch)
I	1	insertion to the reference
D	2	deletion from the reference
N	3	skipped region from the reference
S	4	soft clipping (clipped sequences present in SEQ)
* H	5	hard clipping (clipped sequences NOT present in SEQ)
* P	6	padding (silent deletion from padded reference)
* =	7	sequence match
* X	8	sequence mismatch

"N" indicates splicing event in RNAseq BAMs

*Rarer / newer

CIGAR = "Concise Idiosyncratic Gapped Alignment Report"

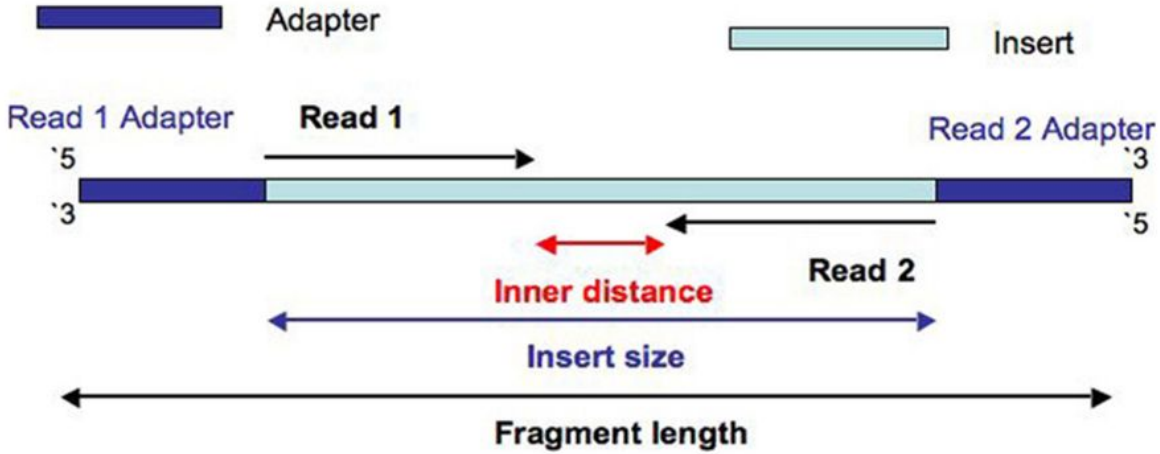
Giaever, G. (2025). Wrangling Genomics: Variant Calling Workflow. Bookdown.org.

Soft Clipping in IGV

Read pairs

- **R+ (Read Positive)**: Refers to a read mapped to the positive strand of DNA in sequencing.
- **Pairs span**: Refers to the distance between paired-end reads in sequencing (especially in long-range sequencing). This is the span between two paired reads in terms of physical distance in the genome.
- **F+ (Forward Positive)**: Refers to a forward read mapped on the positive strand.
- **F- (Forward Negative)**: Refers to a forward read mapped on the negative strand.
- **R+ (Reverse Positive)**: Refers to a reverse read mapped on the positive strand.
- **R- (Reverse Negative)**: Refers to a reverse read mapped on the negative strand.

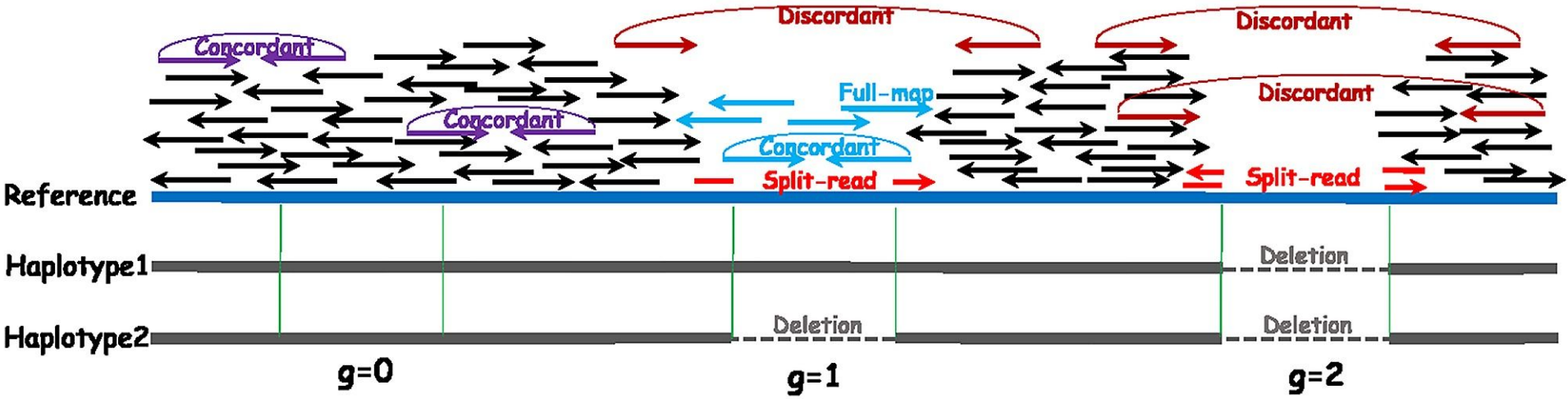
Insert size



Turner et al., Front Genet, 2014

- **Insert size** refers to the length of the DNA fragment between the paired-end reads. It is the distance between the start of one read and the start of the other read in a paired-end sequencing run.
- For Illumina systems DNA insert size range of 200–800 bp (ecseq.com).

Concordant, discordant, split reads



Chu et al., PLoS ONE, 2015

Concordant reads: Both reads are aligned as expected based on the insert size and orientation.

Discordant reads: The reads' alignment does not fit the expected pattern, possibly indicating structural variations or errors.

Split reads: Reads that align to different genomic regions, often due to structural variations such as deletions, insertions, or translocations.

Discordant reads

Definition: Discordant reads are paired-end reads where the alignment of the two reads does not match the expected pattern based on their insert size or orientation.

Characteristics:

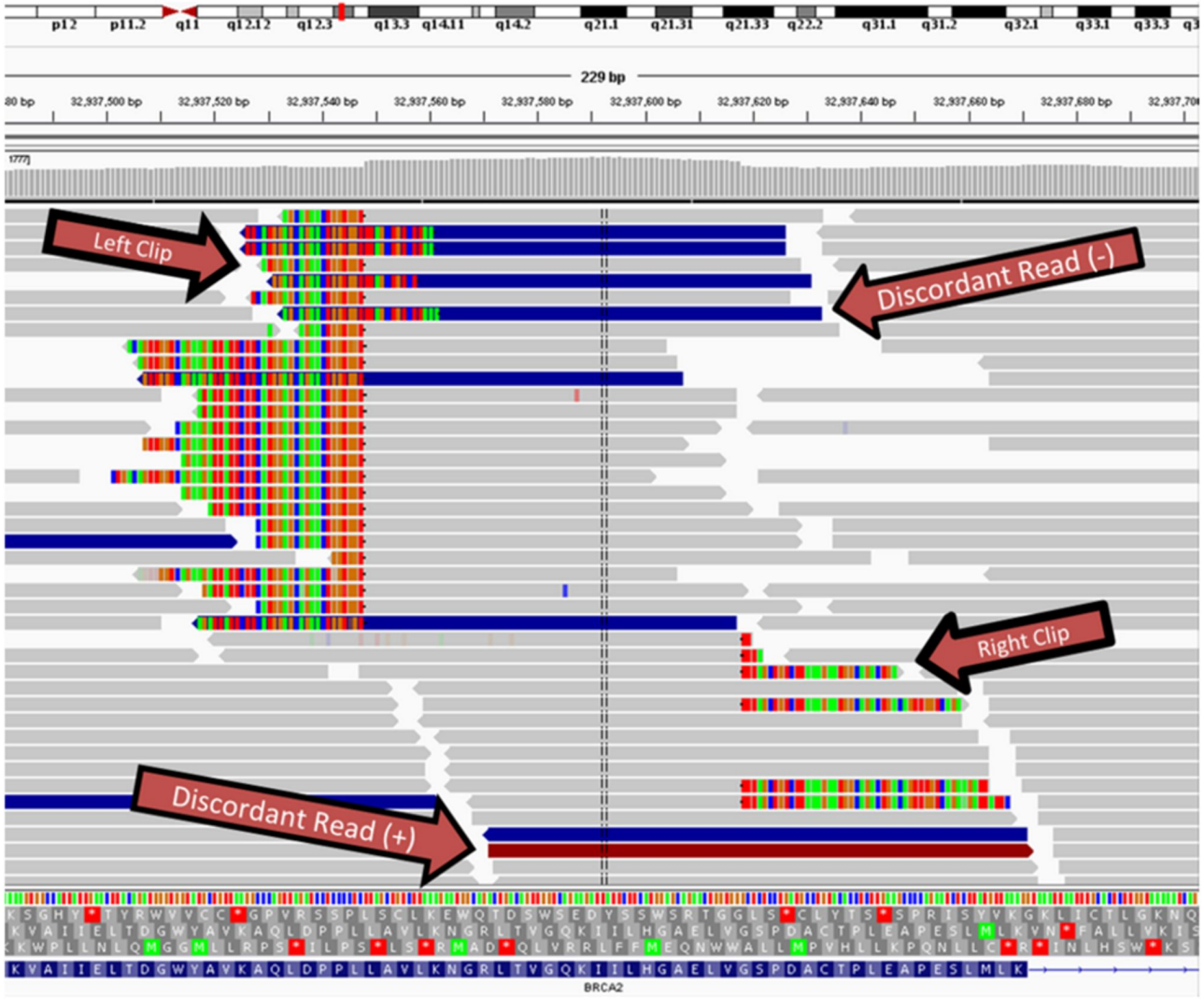
The orientation of the reads is inconsistent with what is expected.

The distance between the reads is too large or too small compared to the expected insert size distribution.

Discordant reads can indicate **structural variations** (such as deletions, insertions, inversions, or translocations) or sequencing errors.

Example: If you have a forward read (F+) and a reverse read (R+) that are supposed to be paired, but the distance between them is too large (or too small), or their orientation is incorrect, these would be considered discordant.

Discordant reads and soft clipping



Hart et al., PLoS ONE, 2013

Split reads

Definition: Split reads are reads that map to two different locations in the genome. This typically happens when a read spans a structural variation, such as a **deletion**, **insertion**, or **translocation**.

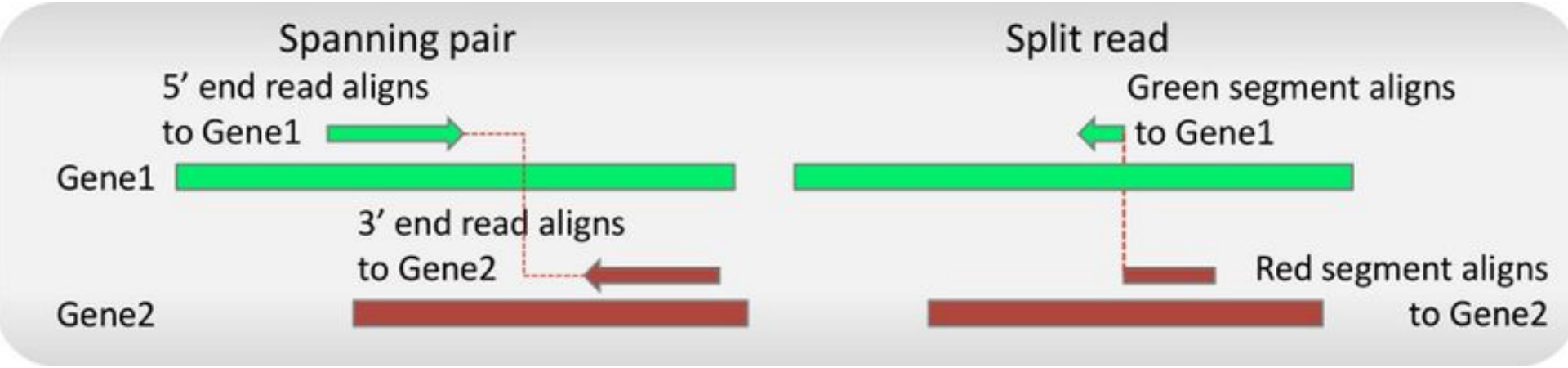
Characteristics:

The read aligns to one part of the reference genome but contains part of the sequence that aligns to a different region.

- **Deletions**: A part of the read aligns to a region where there is a gap in the genome.
- **Insertions**: Part of the read aligns to a region with extra bases inserted.
- **Translocations**: Part of the read aligns to a region that has been moved to a different location in the genome.

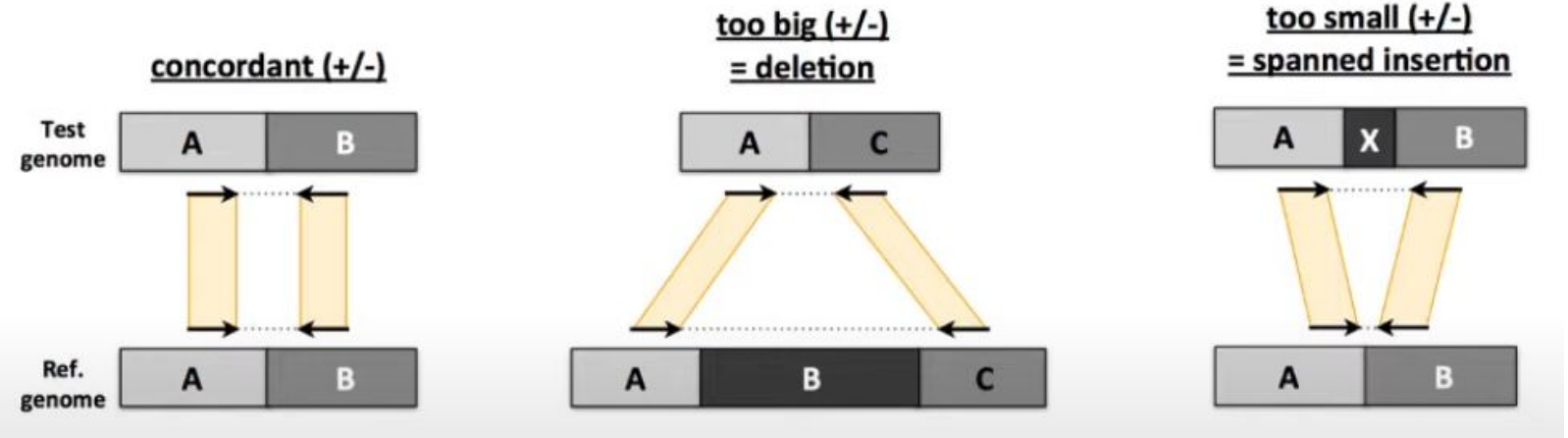
Example: A read might align perfectly to one part of the genome, but the second part of the read aligns to a different chromosome or a different region of the same chromosome, indicating a potential structural variant.

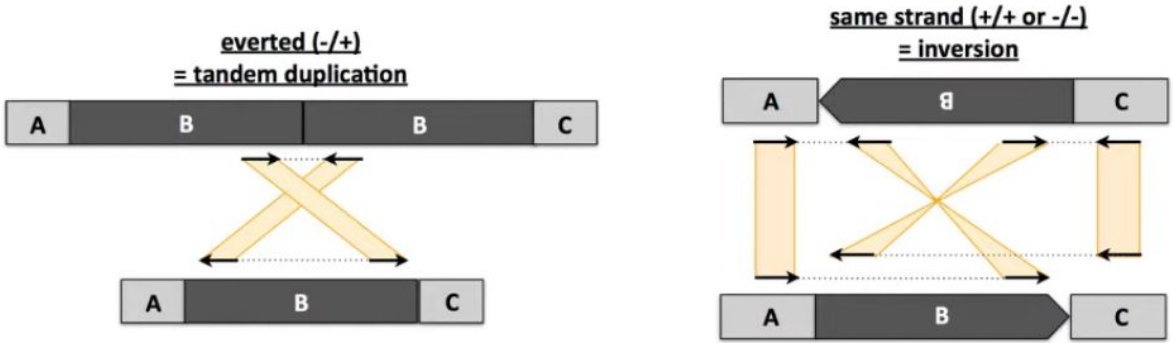
Detection gene fusion



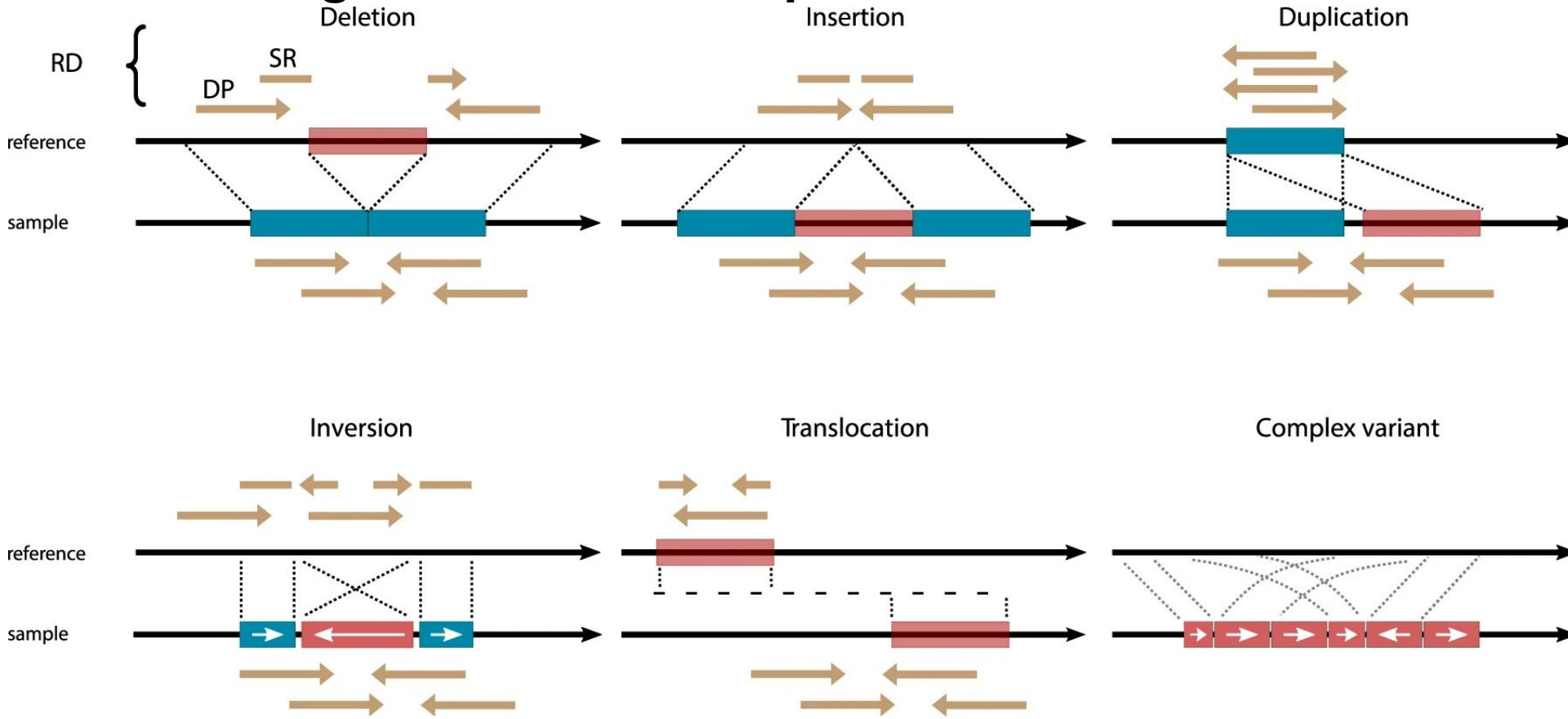
Liu et al., BMC Bioinformatics, 2013

Using discordant reads to detect SVs





Using discordant and split reads to detect SVs



van Belzen et al., NPJ Precis Oncol., 2021