LEIDEN UNIVERSITY MEDICAL CENTER

Analyse what others throw away

Jeroen F.J. Laros

Leiden Genome Technology Center

Department of Human Genetics

Center for Human and Clinical Genetics

Introduction

Sequencing

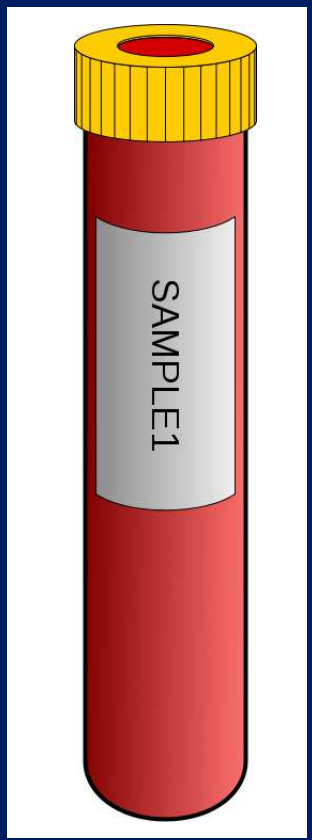

Figure 1: Blood sample.

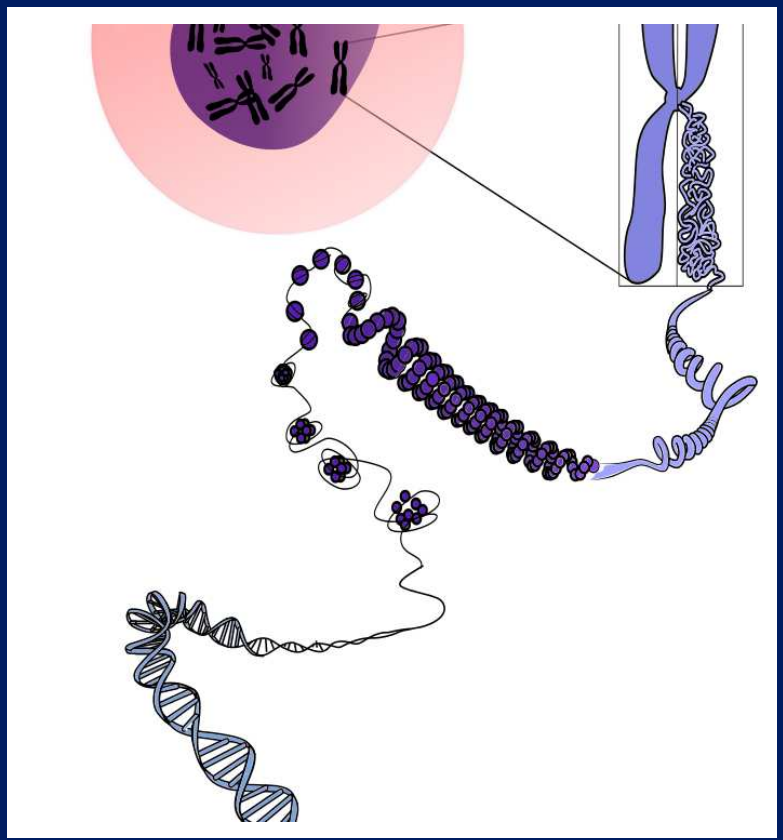

Figure 2: DNA.

Introduction

Sequencing

Figure 3: HiSeq 2500.

Characteristics:

- High throughput (3 genomes).
- Paired end.
- High accuracy.
- Read length 2×125bp.
- Relatively long run time (6 days).
- Relatively expensive.

Introduction

Next generation sequencing data

```
1  @SGGPP:4:101
2  TTCGGGGGCTGGCAAATCCACTTCCGTGACACGCTACCATTCGCTGGTGGT
3  +
4  −’+4589,53330−0&07+03:54/2362−+.488587>@/25440++0(+
5  @SGGPP:4:102
6  CGGTAAACCACCCTGCTGACGGAACCCTAATGCGCCTGAAAGACAGCGTTC
7  +
8  34/−−0’+.000(.55:;:99(0(+2(22(0316;185;;0;:<<>=AA59
9  @SGGPP:4:106
10 TCGTTAACGACTTTGTTCGCCACCGCAACCGCCTGTTTCGGGTCACAGGCA
11 +
12 09875;5?<;?@A4?B:BBB<AA>CCC>C>BB0.−>=0488+3444:@5@<
13 @SGGPP:4:112
14 TTGATGAATATATTATTTCAGGGAATAATTATGACACCTTTAGAACGCATT
15 +
16 70<<@::5:<;==7;>>/79<:.:494.8(,,8:753/5@5??C>B???B7
```

Listing 1: A FastQ file.

Introduction

Data analysis

Resequencing pipelines can roughly be divided in five steps.

Introduction

Data analysis

Resequencing pipelines can roughly be divided in five steps.

1. Pre-alignment.
    - Quality control.
    - Data cleaning.

Introduction

Data analysis

Resequencing pipelines can roughly be divided in five steps.

1. Pre-alignment.
    - Quality control.
    - Data cleaning.
2. Alignment.
    - Post-alignment quality control.

Introduction

Data analysis

Resequencing pipelines can roughly be divided in five steps.

1. Pre-alignment.
    - Quality control.
    - Data cleaning.
2. Alignment.
    - Post-alignment quality control.
3. Variant calling.

Introduction

Data analysis

Resequencing pipelines can roughly be divided in five steps.

1. Pre-alignment.
    - Quality control.
    - Data cleaning.
2. Alignment.
    - Post-alignment quality control.
3. Variant calling.
4. Filtering.
    - Post-variant calling quality control.

Introduction

Data analysis

Resequencing pipelines can roughly be divided in five steps.

1. Pre-alignment.
    - Quality control.
    - Data cleaning.
2. Alignment.
    - Post-alignment quality control.
3. Variant calling.
4. Filtering.
    - Post-variant calling quality control.
5. Annotation.

Alignment

The best match to the reference genome

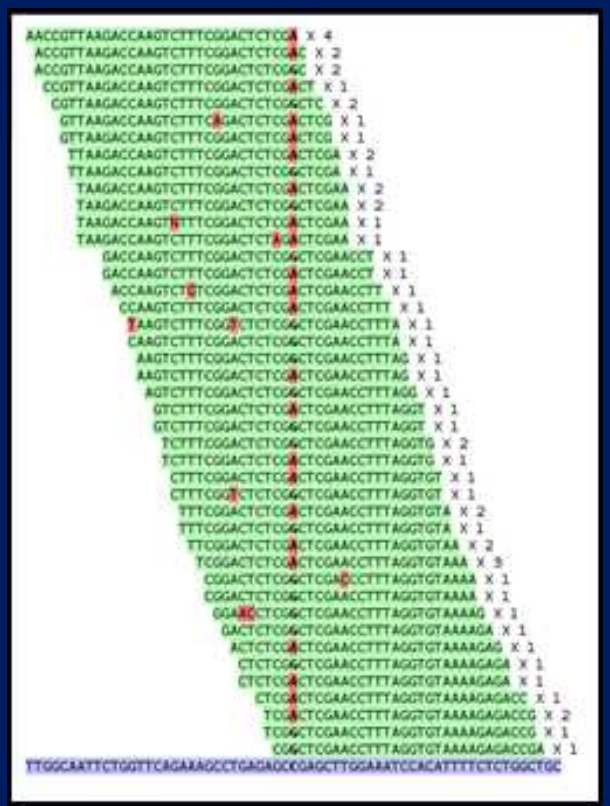

Figure 4: Visualisation of an alignment.

Very efficient.

- The reference genome needs to be *indexed*.
- Finding an alignment is as easy as looking up a word in a dictionary.

Full genome sequencing

Copy number variation

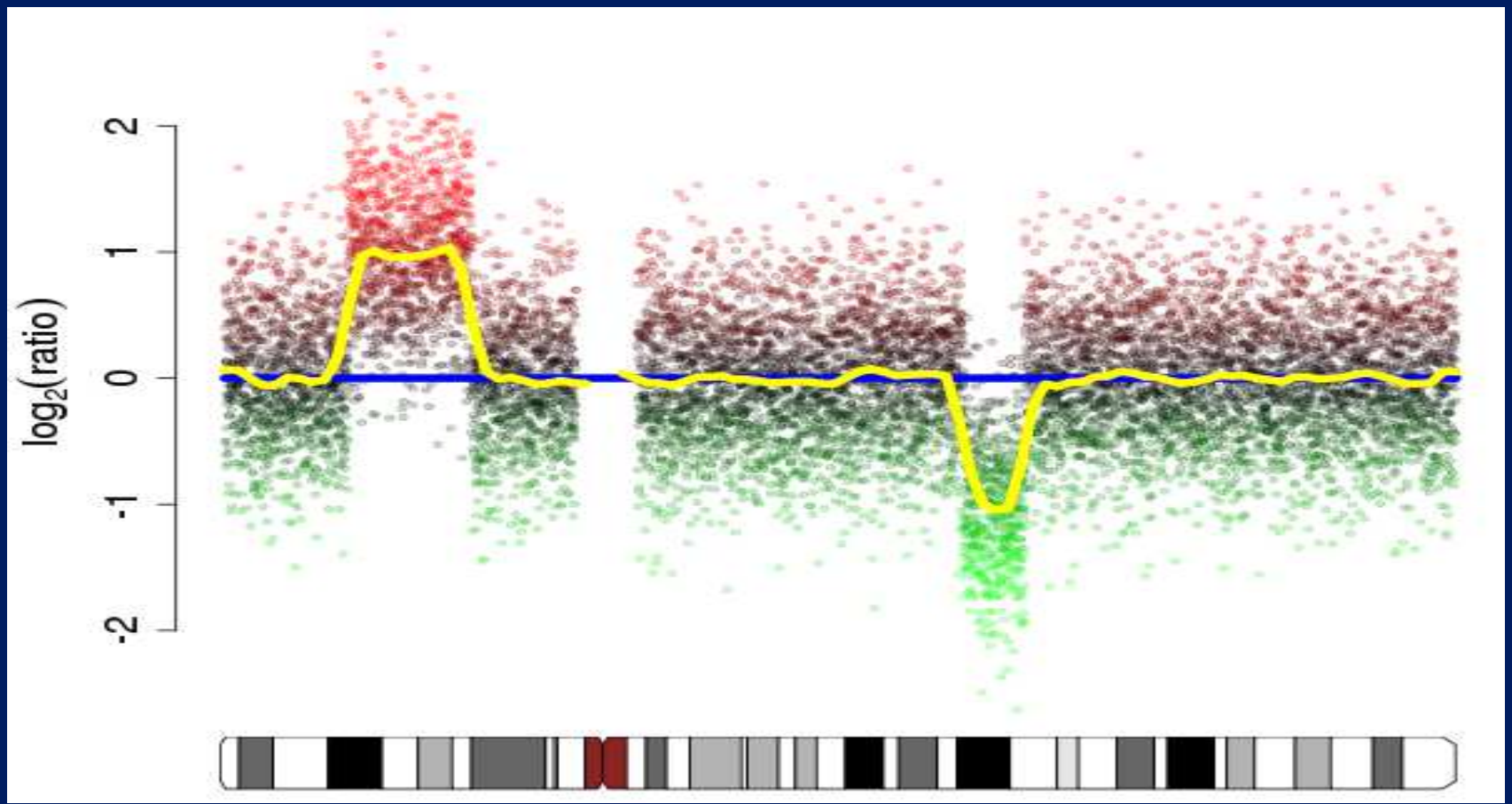

Figure 5: Coverage patterns over a whole chromosome.

Full genome sequencing

Structural variation

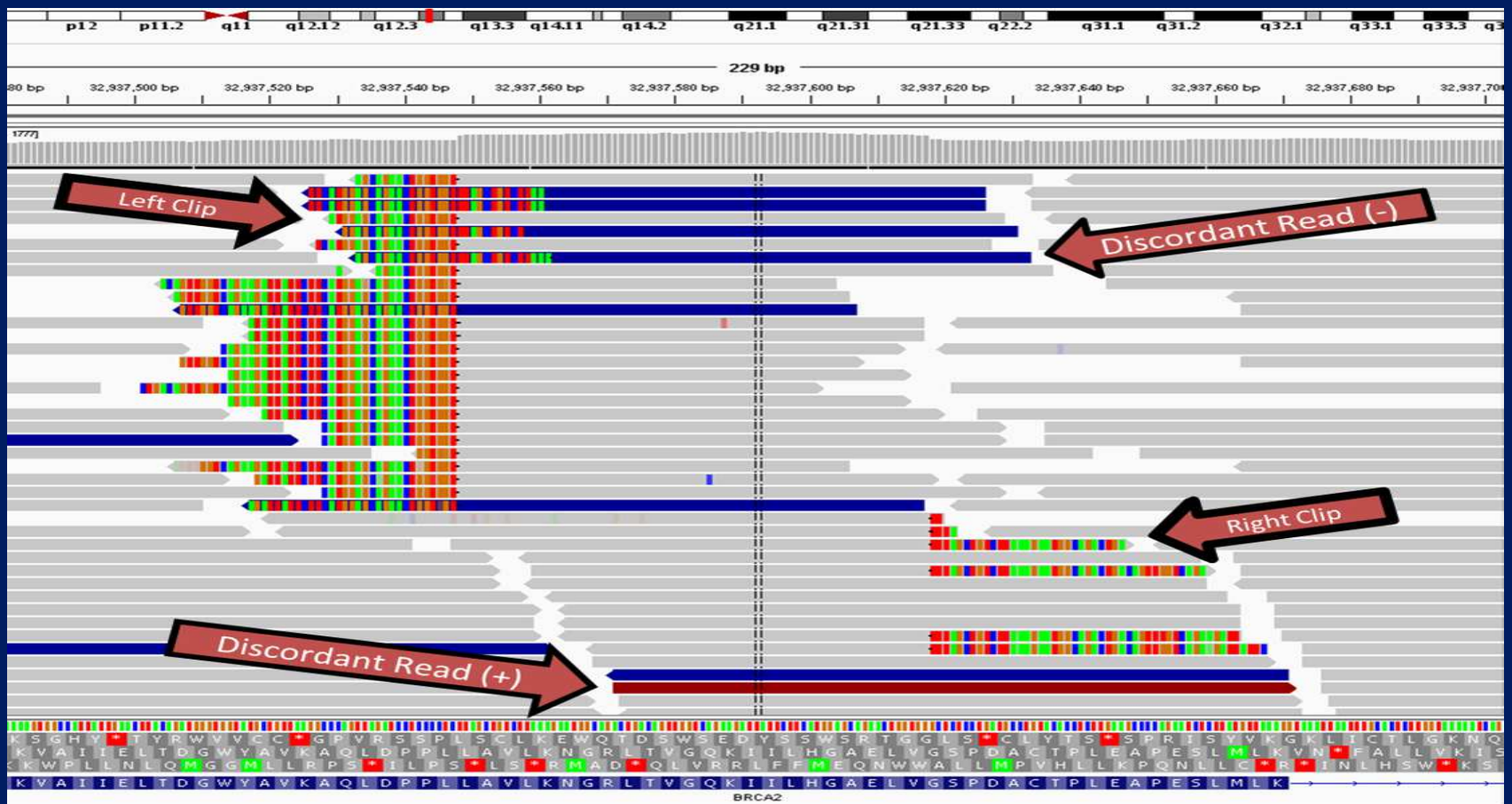

Figure 6: Multiple issues while mapping.

Full genome sequencing

Structural variation

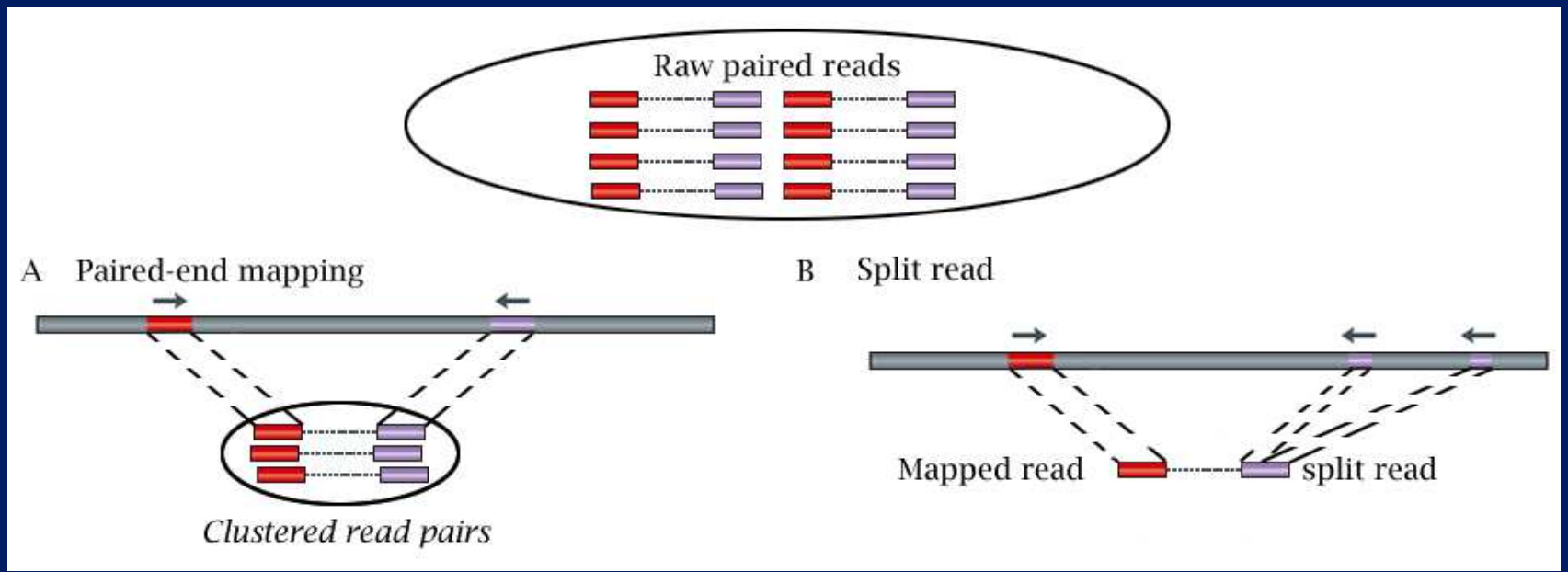

Figure 7: Discordant and split reads.

http://breakdancer.sourceforge.net/

http://sourceforge.net/projects/pindel/

Full genome sequencing

Structural variation

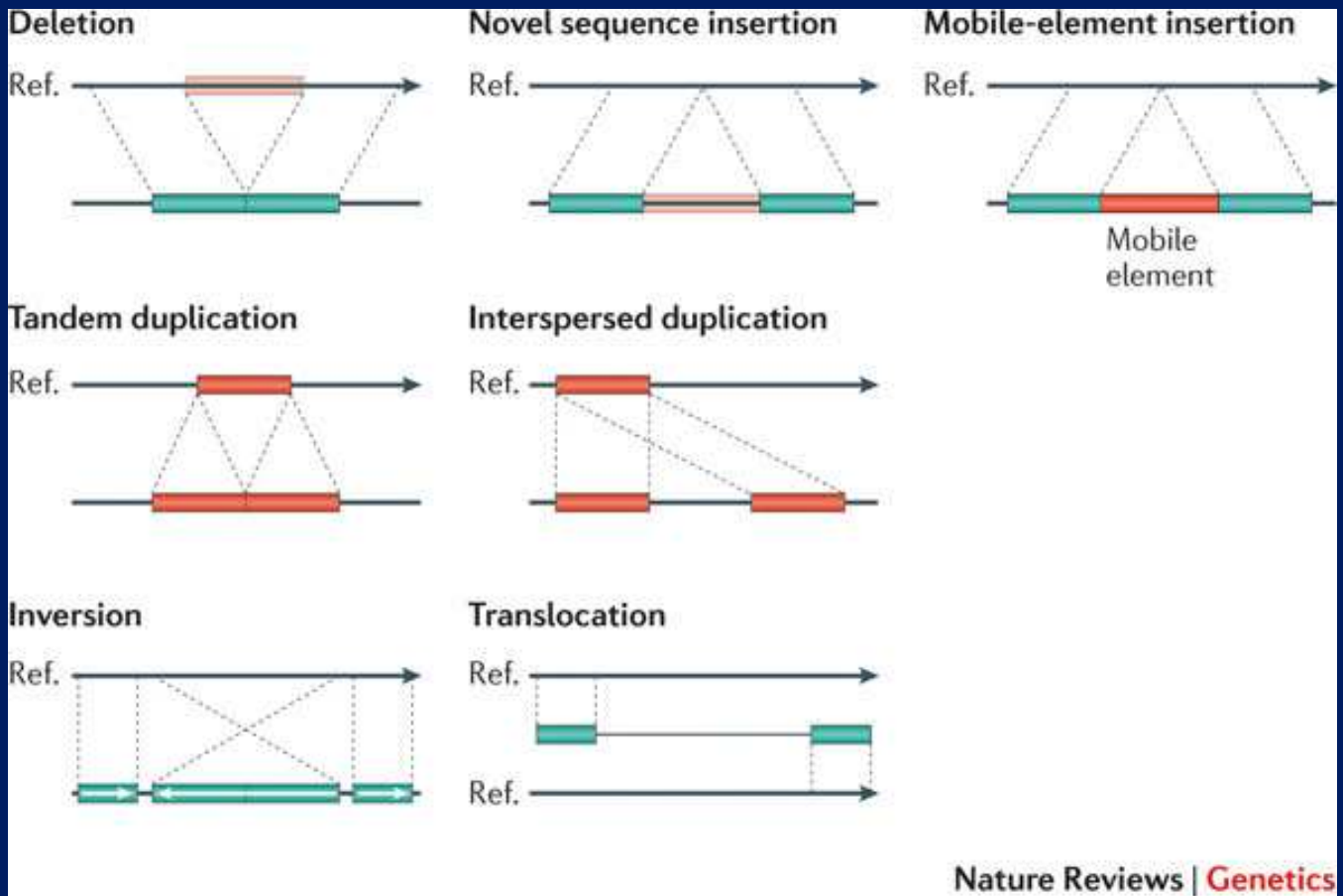

Figure 8: Different types of structural variation.

Outline

Candidate analyses

We propose a number of analyses:

- k-mer analysis.
  - Catalogue the reads, look for shifts.
- Map to repetitive elements, look for shifts.
- *De novo* assembly of unmapped reads.
  - BLAST the assembled contigs to determine their origin.
- Analysis of *partially mapped* reads.
  - Breakpoint sequences?
  - Transposable elements, large genomic rearrangements?

Outline

Candidate analyses

We propose a number of analyses:

- k-mer analysis.
  - Catalogue the reads, look for shifts.
- Map to repetitive elements, look for shifts.
- *De novo* assembly of unmapped reads.
  - BLAST the assembled contigs to determine their origin.
- Analysis of *partially mapped* reads.
  - Breakpoint sequences?
  - Transposable elements, large genomic rearrangements?
- Your own ideas.

Questions?

Jeroen Laros
Johan den Dunnen