LEIDEN UNIVERSITY MEDICAL CENTER

# Introduction to NGS data analysis

**Jeroen F. J. Laros**

**Leiden Genome Technology Center**

**Department of Human Genetics**

**Center for Human and Clinical Genetics**

# Sequencing

## *Illumina platforms*

Figure 1: HiSeq 2500.

Characteristics:

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

# Sequencing

## *Illumina platforms*

Figure 2: MiSeq.

Characteristics:

- Moderate throughput (3 exomes).
- Paired end.
- High accuracy.
- Read length $2 \times 300$bp.
- Relatively short run time (3 days).
- Relatively expensive.

# Sequencing

## *Illumina platforms*

Figure 3: Flowcell.

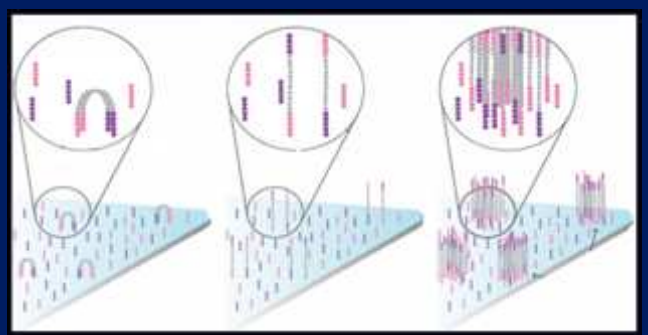


Figure 4: Amplification.

# Sequencing

## *Illumina platforms*

Figure 3: Flowcell.

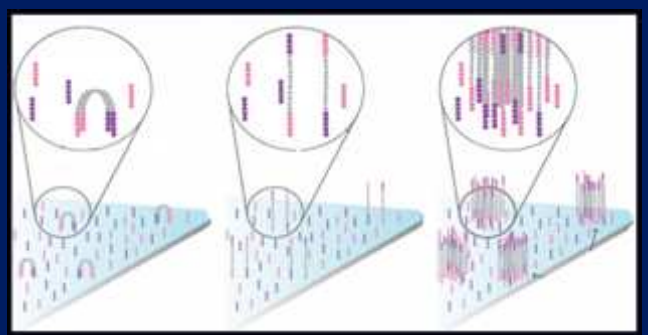


Figure 4: Amplification.

Optical system.

- DNA fragments are loaded on a flowcell.
- Fragments are amplified.
- Clusters of fragments give the same signal.

# Sequencing

## *Illumina platforms*

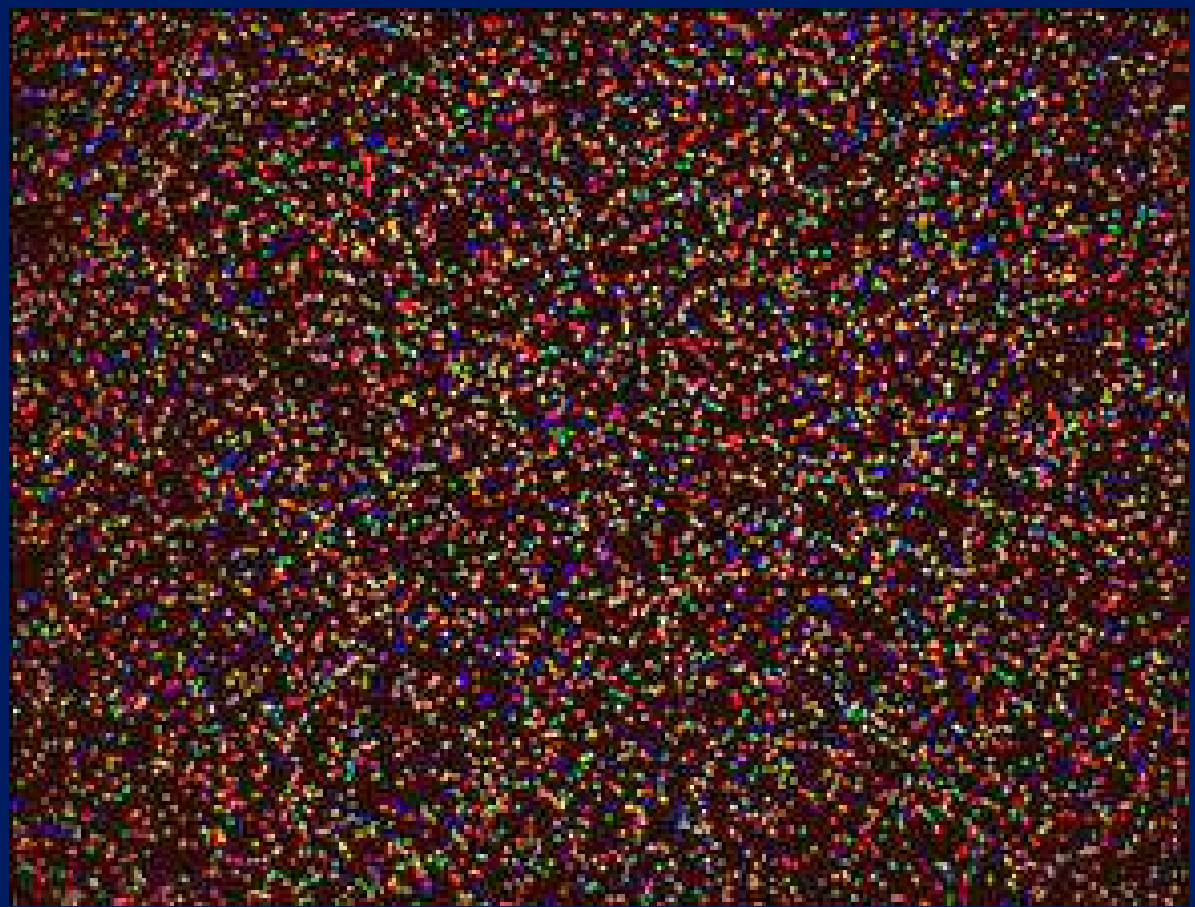


Figure 5: Image of tile (part of the flowcell).

# Sequencing

## *Illumina platforms*

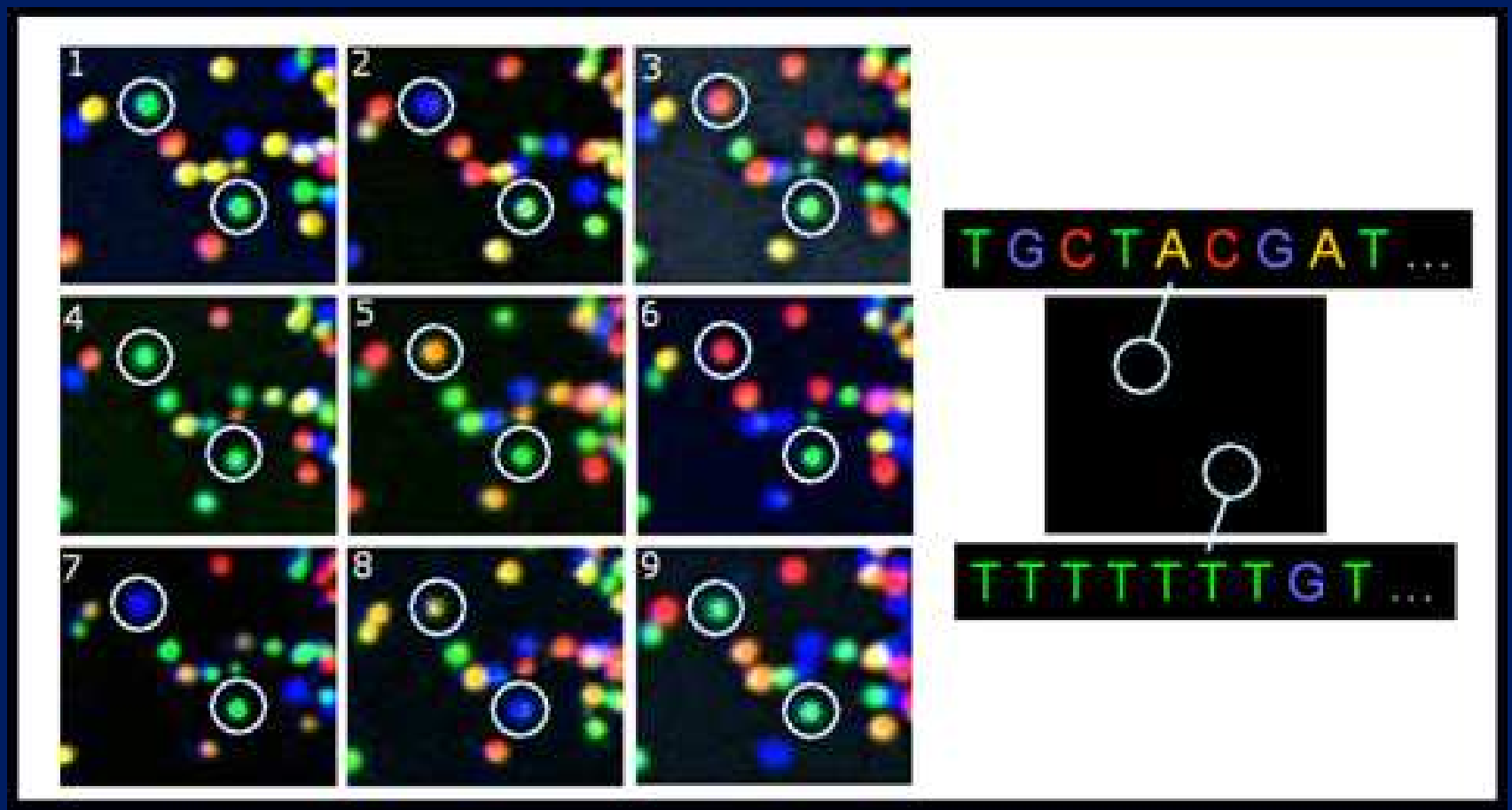



Figure 6: Base calling on Illumina systems.

# Sequencing

## *Illumina platforms*

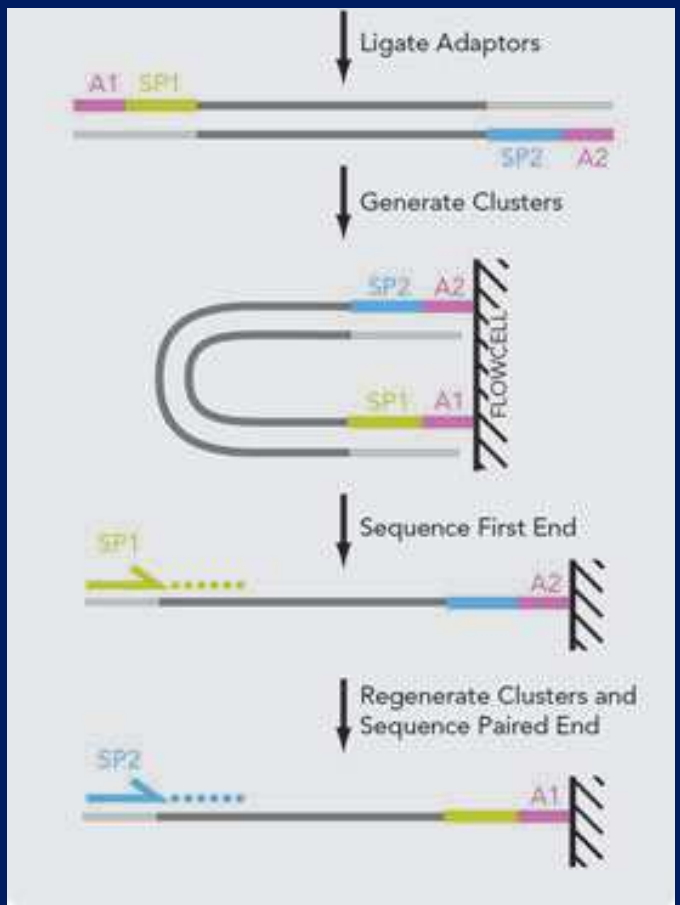



Figure 7: Paired end.

# Sequencing

## *Illumina platforms*

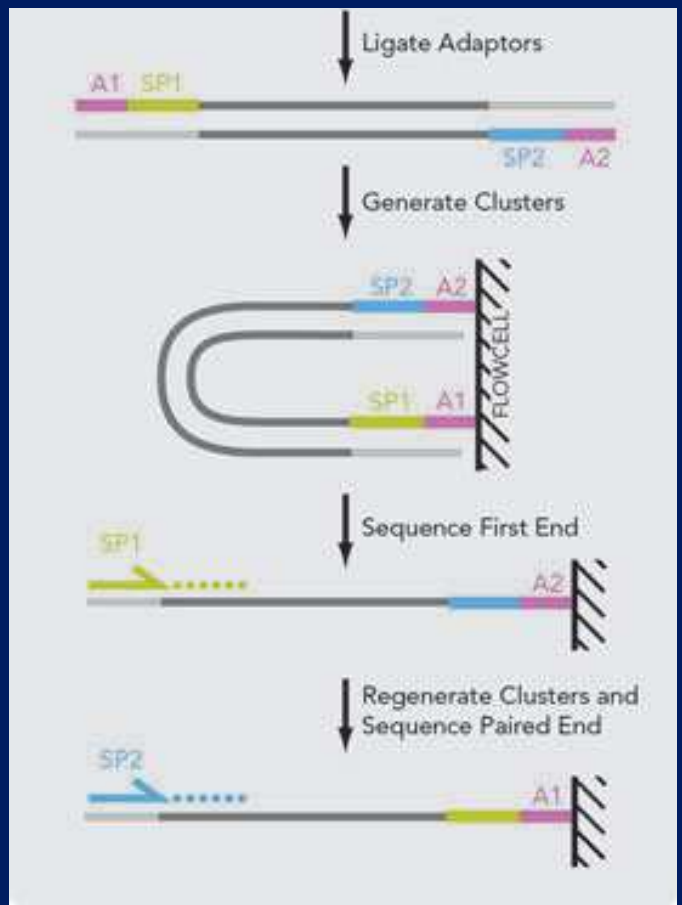



Figure 7: Paired end.

Advantages:

- Mapping to non-unique regions.
- Structural variation.
- *De novo* assembly (scaffolding).

# Sequencing

## *Life-Tech*

Figure 8: Ion torrent.

Characteristics:

- Low throughput.
- Single end (for now).
- High accuracy.
- Read length $\pm$400bp.
- Short run time (1 day).
- Cheap runs.

# Sequencing

## *Life-Tech*

Figure 9: Ion proton.

Characteristics:

- Moderate throughput (1 exome).
- Single end (for now).
- High accuracy.
- Read length $\pm$175bp.
- Short run time (1 day).
- Cheap runs.

# Sequencing

## *Life-Tech*

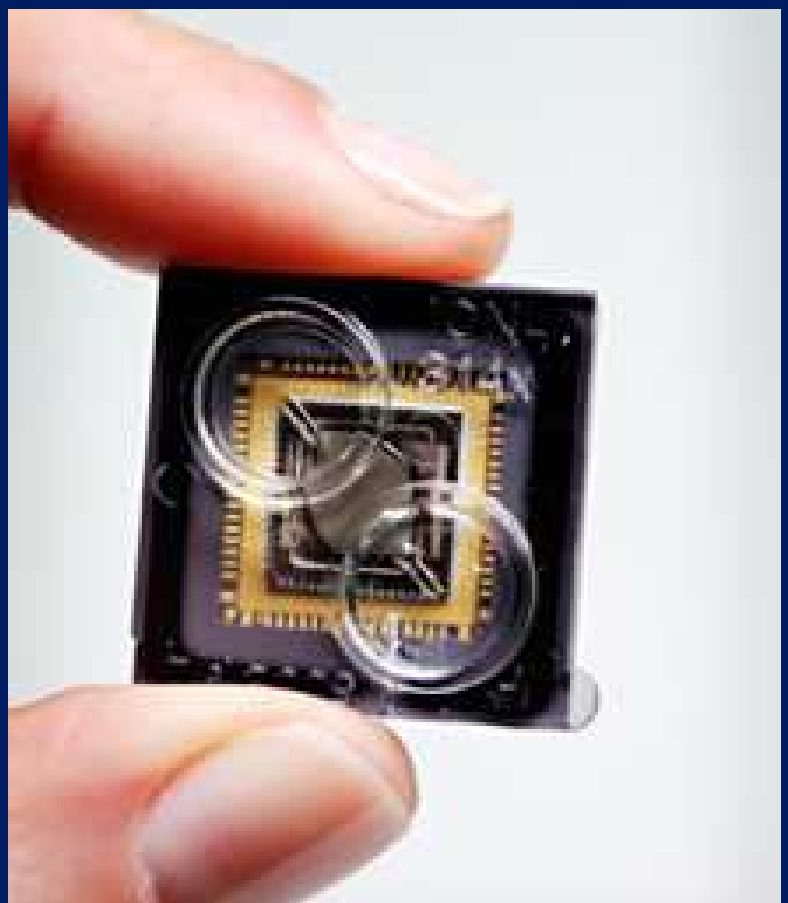


Figure 10: Ion Torrent chip.

Figure 11: Ion Proton chip.

# Sequencing

## *Life-Tech*

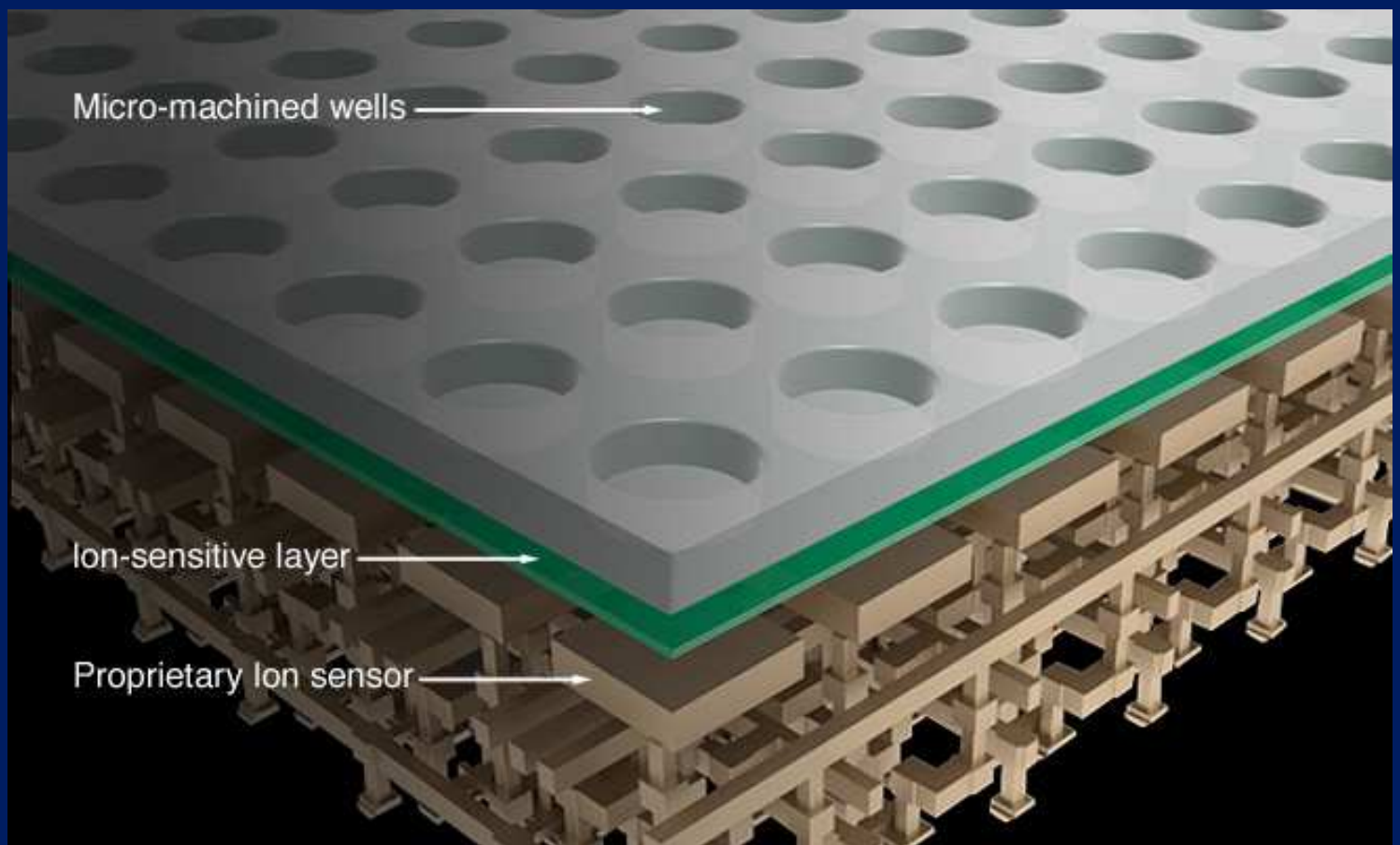



Figure 12: Close up of an Ion chip.

# Sequencing

## *Life-Tech*

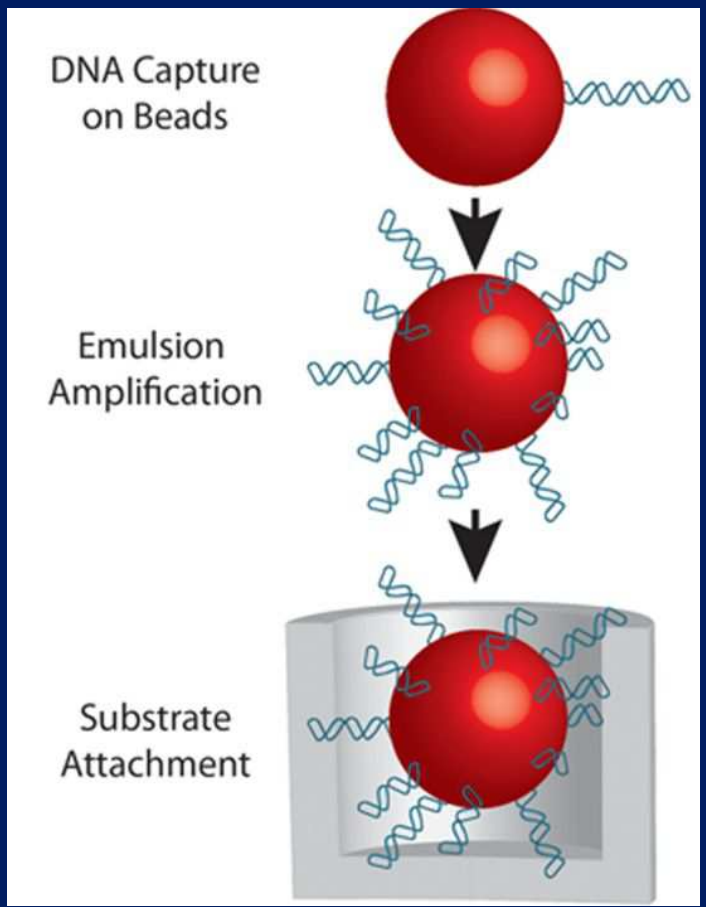



Figure 13: Loading a sample.

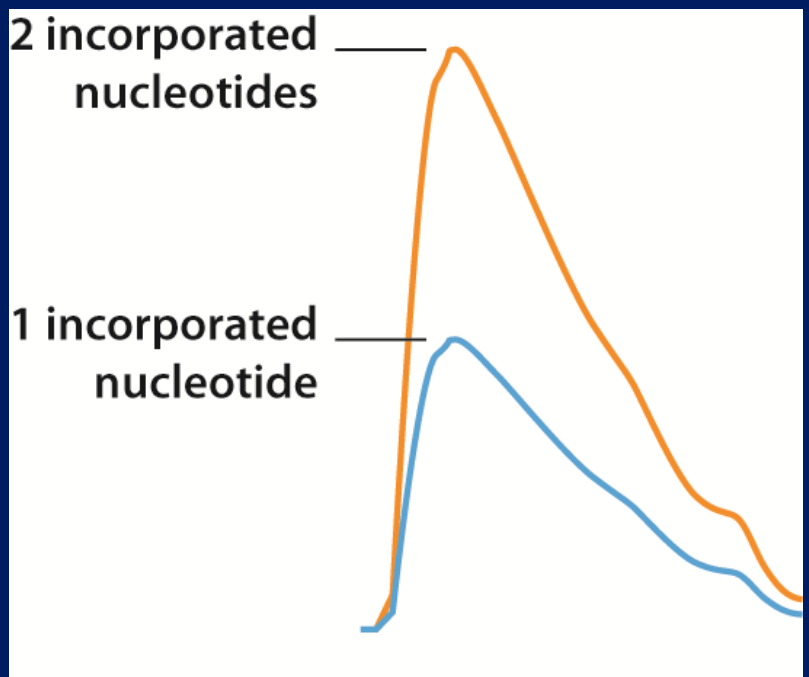



Figure 14: Base calling.

Problems with mono nucleotide stretches.

# Sequencing

## *Pacific Biosciences*

Figure 15: PacBio RS.

# Sequencing

## *Pacific Biosciences*

Characteristics:

- Long reads (50% is longer than 10,000 bp).
- High error rate (15-20%).
- Low throughput (comparable with the Roche 454).

# Sequencing

## *Pacific Biosciences*

Characteristics:

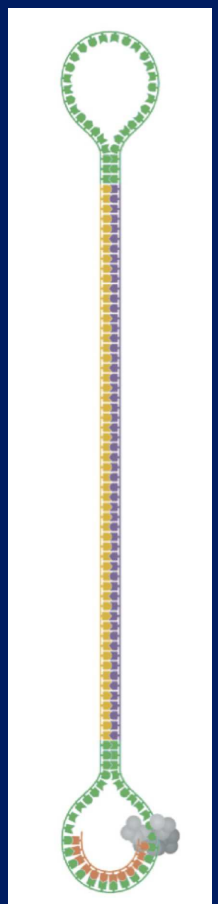
- Long reads (50% is longer than 10,000 bp).
- High error rate (15-20%).
- Low throughput (comparable with the Roche 454).

Circular consensus sequencing.

- Sequence the same molecule several times.
- Extremely high accuracy.
- Acceptable read length ($\pm$250bp).

# Sequencing

## *Pacific Biosciences*



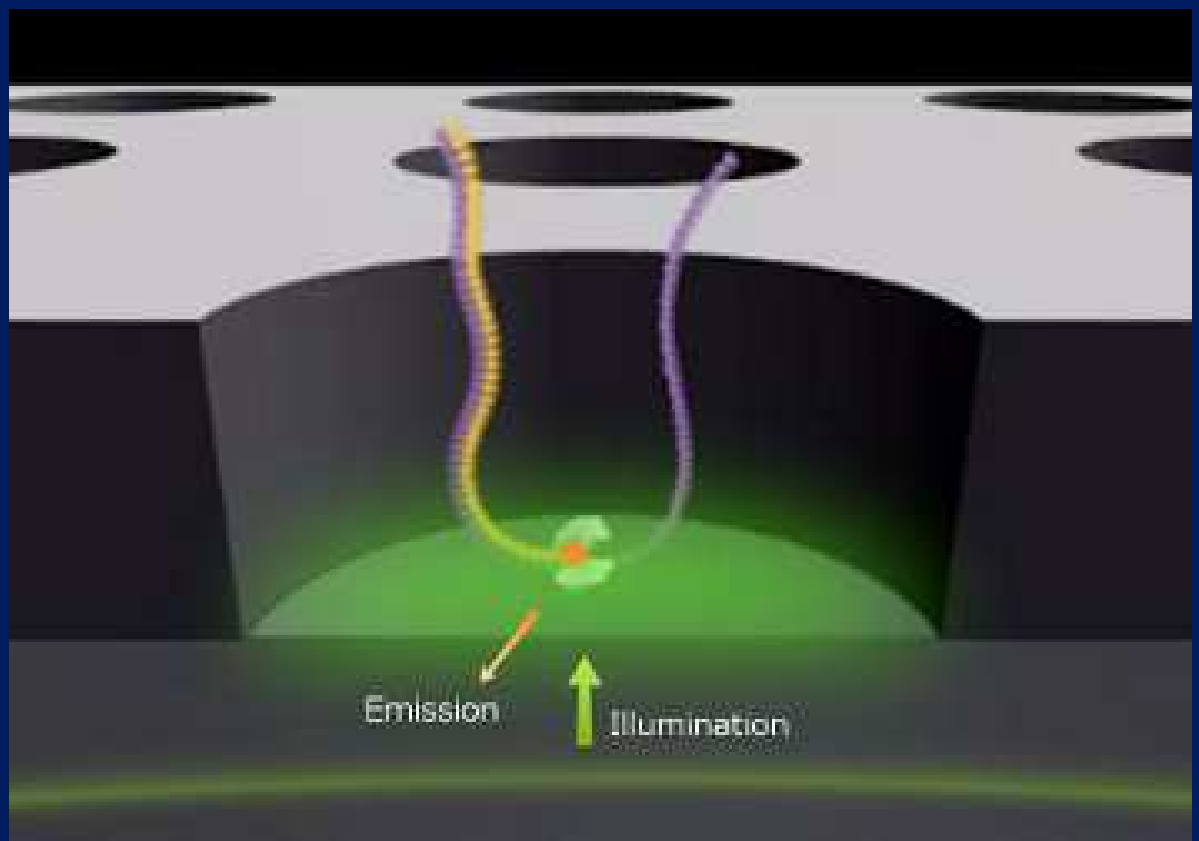



Figure 16: Real time polymerase monitoring.

# Sequencing

## *Pacific Biosciences*

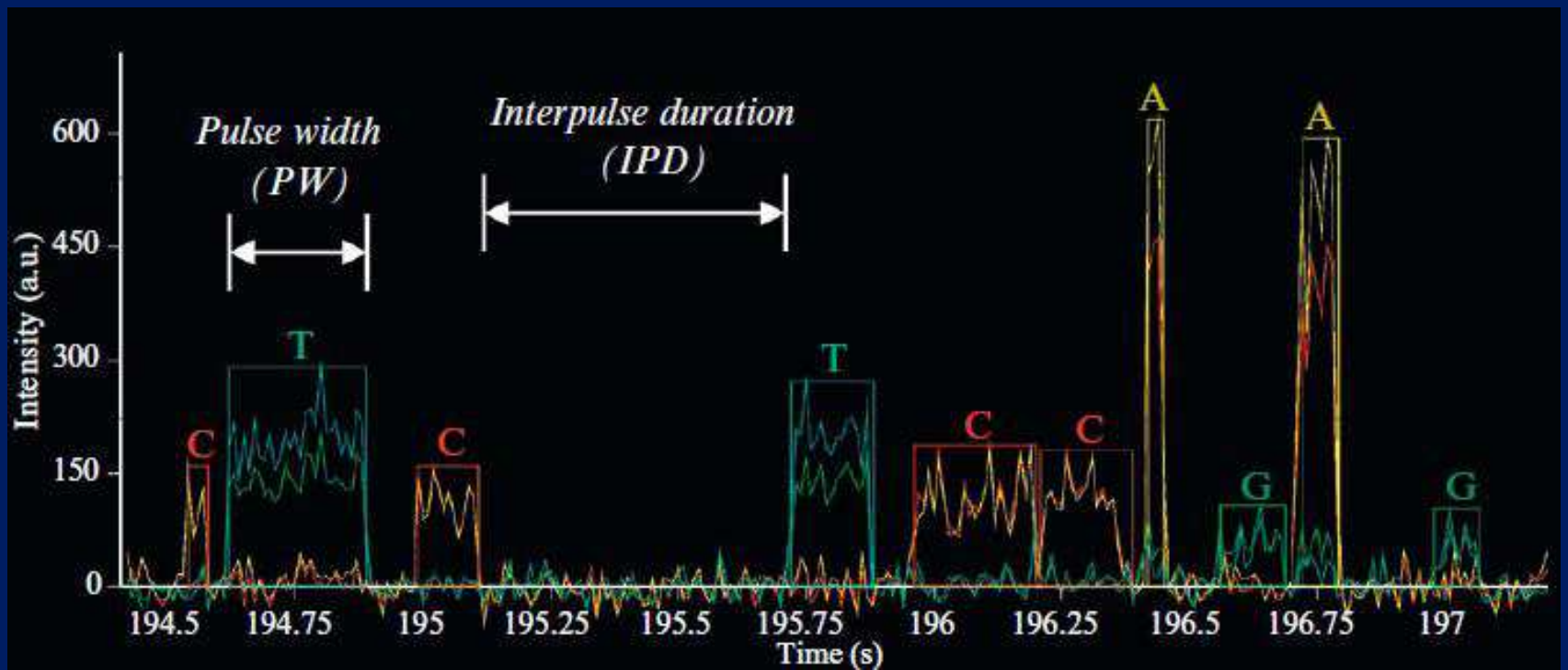



Figure 17: Base calling on the PacBio.

# Data analysis

## *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.

# Data analysis

## *Common data analysis techniques*

Reference based techniques:

- Resequencing.
  - Exome sequencing.
  - Transcriptome sequencing (RNA).
  - Full genome sequencing.
- Structural variation.
- Copy number variation.

Not reference based:

- *De novo* assembly.
- *k*-mer profiling.

## *Quality control*

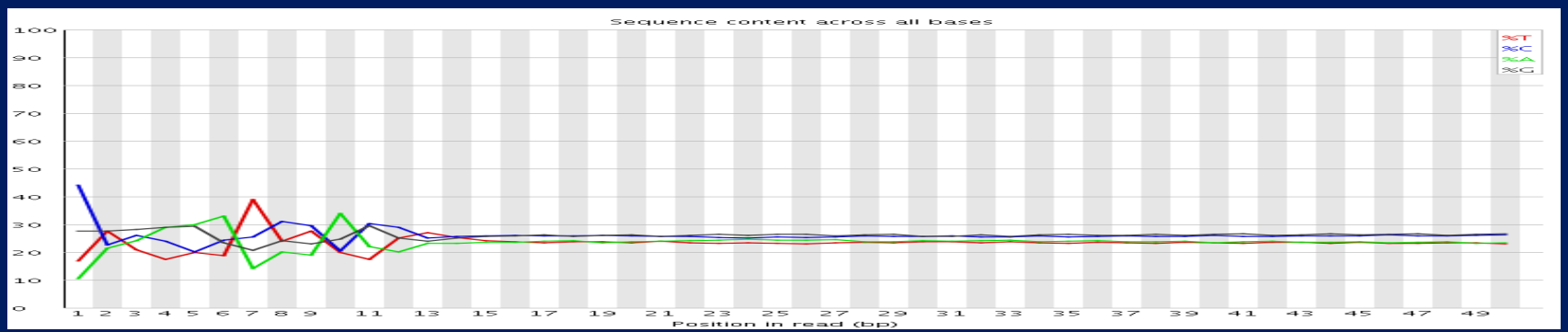



Figure 18: Per base sequence content.

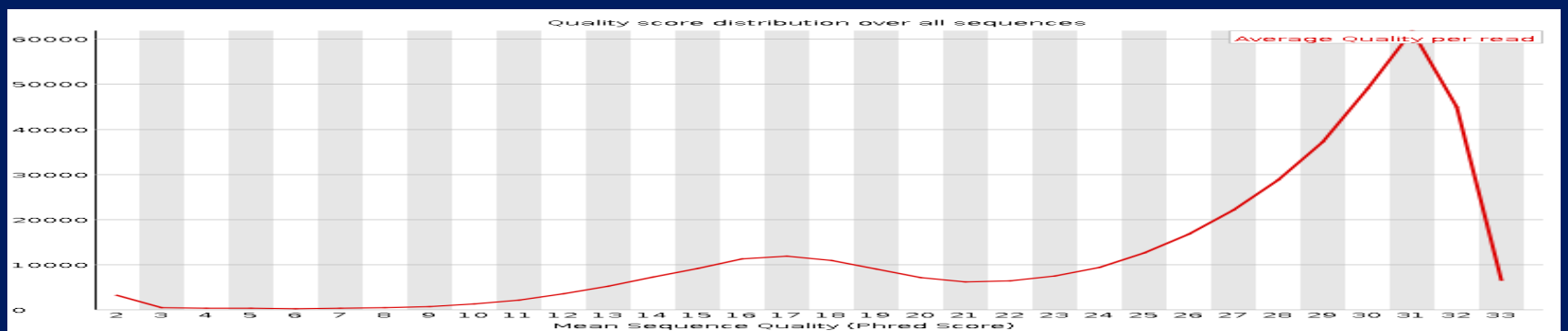



Figure 19: Per sequence quality.

# Alignment

***The best match to the reference genome***

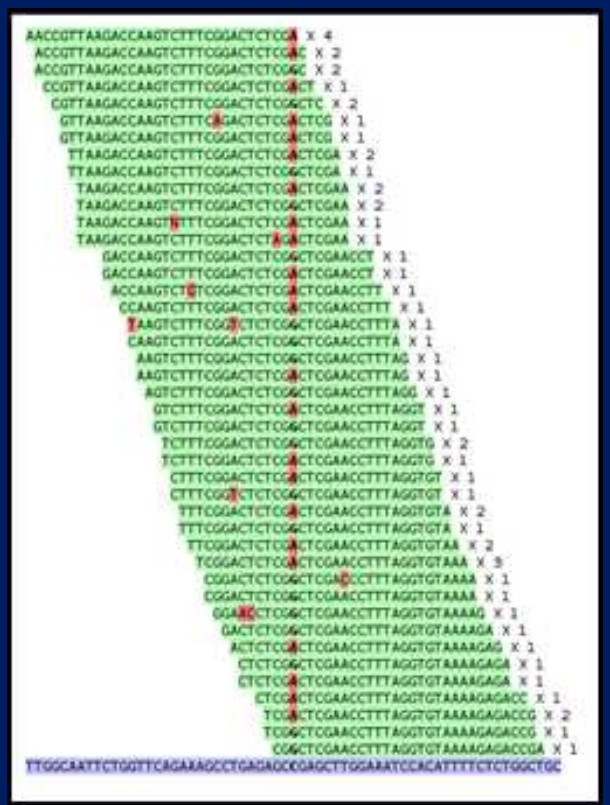


Figure 20: Visualisation of an alignment.

Very efficient.

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

# Variant calling

## *Consistent deviations from the reference*

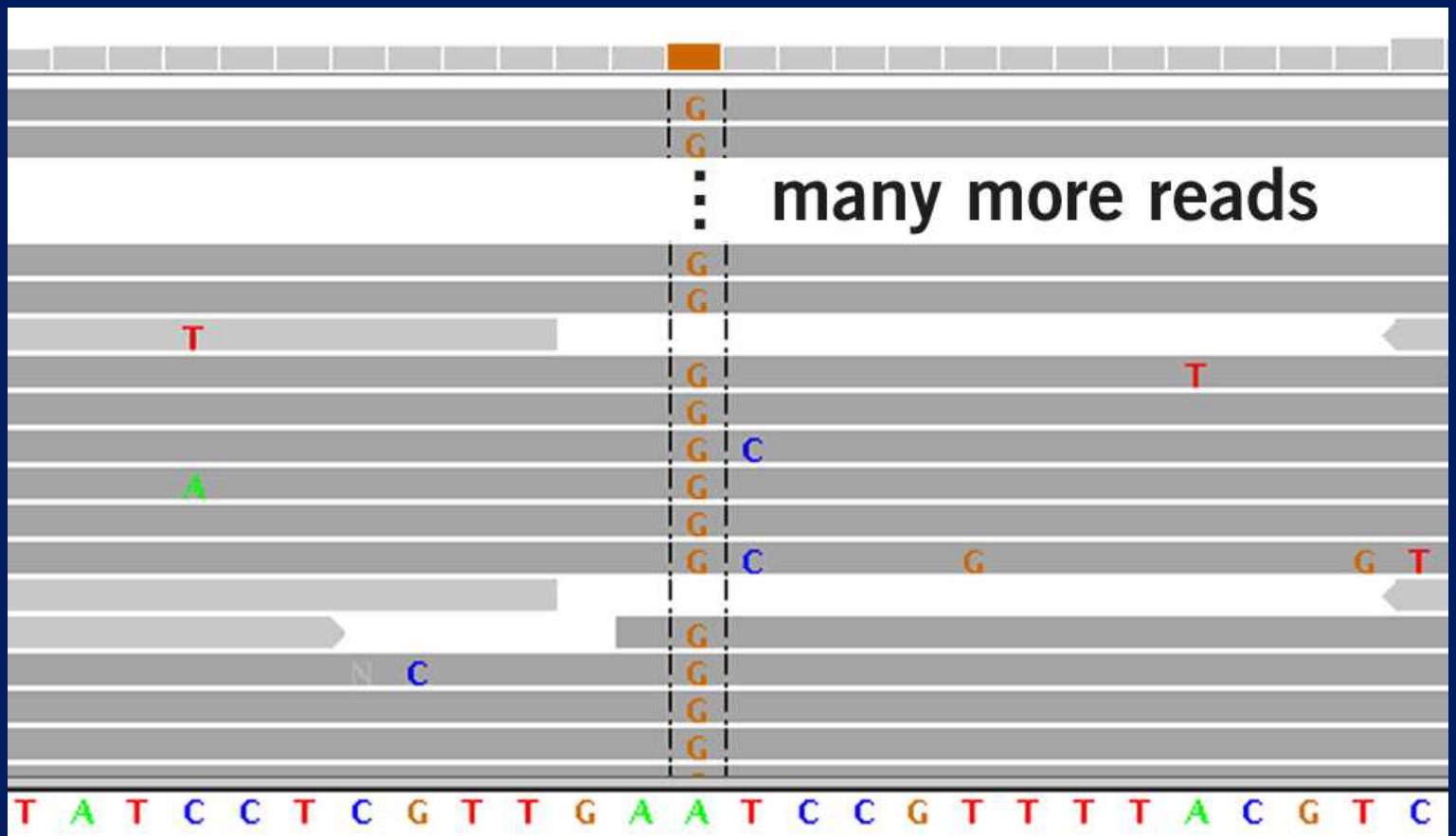



Figure 21: Result of an alignment.

# De Novo assembly

## *Assesmbly*

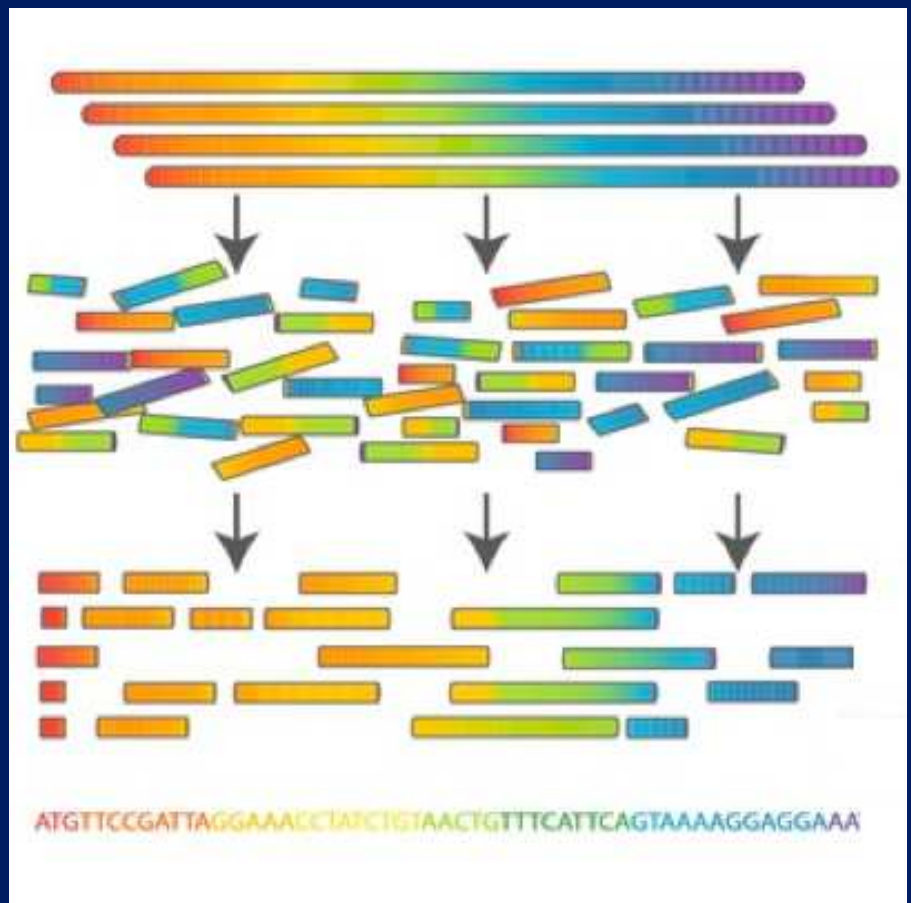



Figure 22: General overview of *de novo* assembly.

# Pipelines

## *Pipelines*

Figure 23: A real-life pipeline.

## *Pipelines*

Figure 24: Scene from “Modern times”.

# Pipelines

## *Pipelines*

Combining tools:

- The output of one tool can serve as the input for another.
- Not necessarily linear.
- . . .

# Pipelines

## *Pipelines*

Combining tools:

- The output of one tool can serve as the input for another.
- Not necessarily linear.
- ...

Running various different tools:

- Two or three different aligners.
- A couple of variant callers.
- ...

# Pipelines

## *Combining tools*

```
1  bwa aln -t 8 $reference $i > $i.sai
2  bwa samse $reference $i.sai $i > $i.sam
3  samtools view -bt $reference -o $i.bam $i.sam
```

Listing 2: Shell script

# Pipelines

## *Combining tools*

```
1  bwa aln -t 8 $reference $i > $i.sai
2  bwa samse $reference $i.sai $i > $i.sam
3  samtools view -bt $reference -o $i.bam $i.sam
```

Listing 2: Shell script

```
1  %.sai: %.fq
2    $(BWA) aln -t $(THREADS) $(call MKREF, $@) $< > $@
3
4  %.sam: %.sai %.fq
5    $(BWA) samse $(call MKREF, $@) $^ > $@
6
7  %.bam: %.sam
8    $(SAMTOOLS) view -bt $(call MKREF, $@) -o $@ $<
```

Listing 3: Makefile

# Pipelines

## *Workflow of a parallel pipeline*

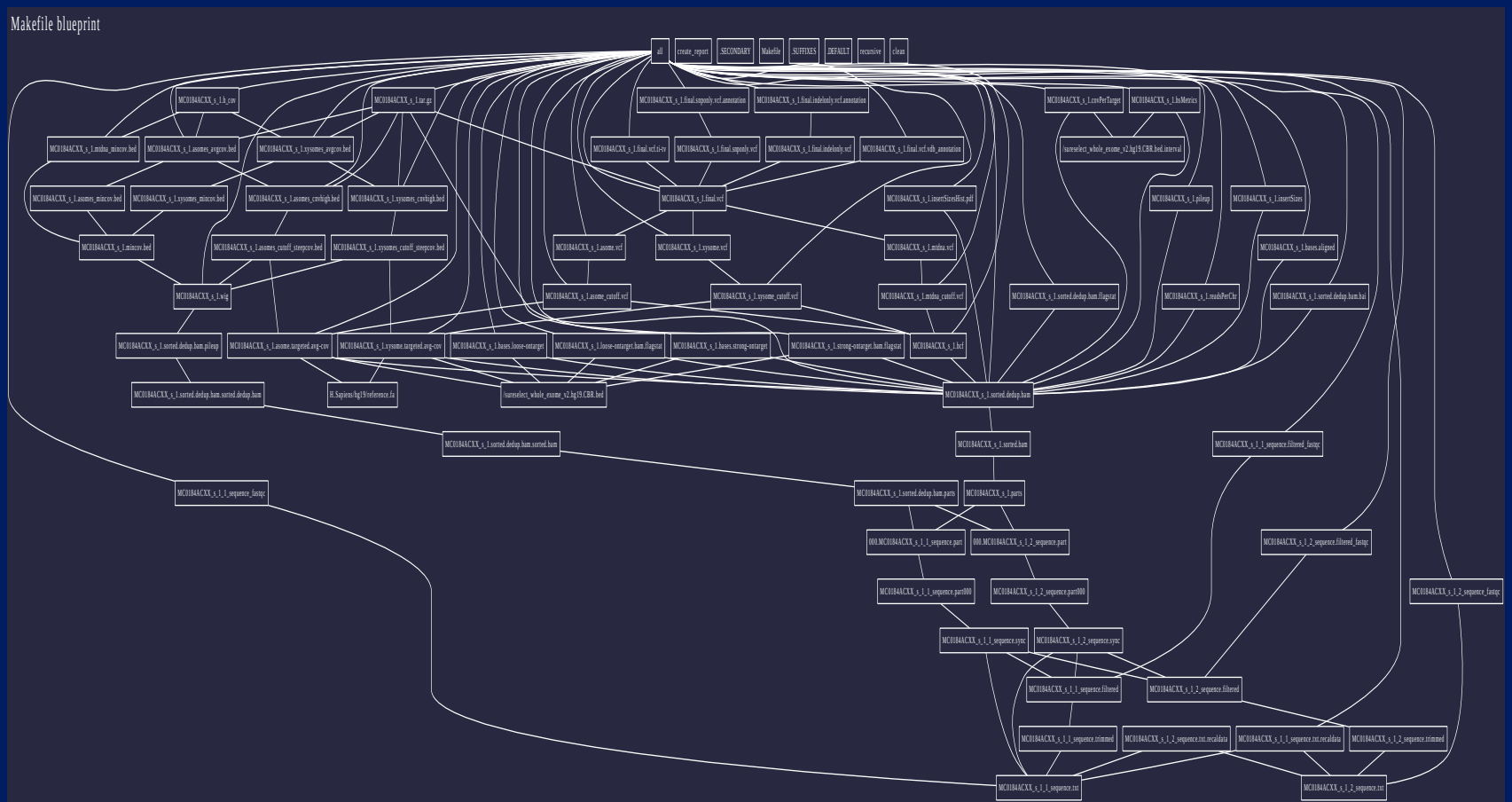



Figure 25: Dependency diagram.

# Pipelines

## *Workflow of a parallel pipeline*

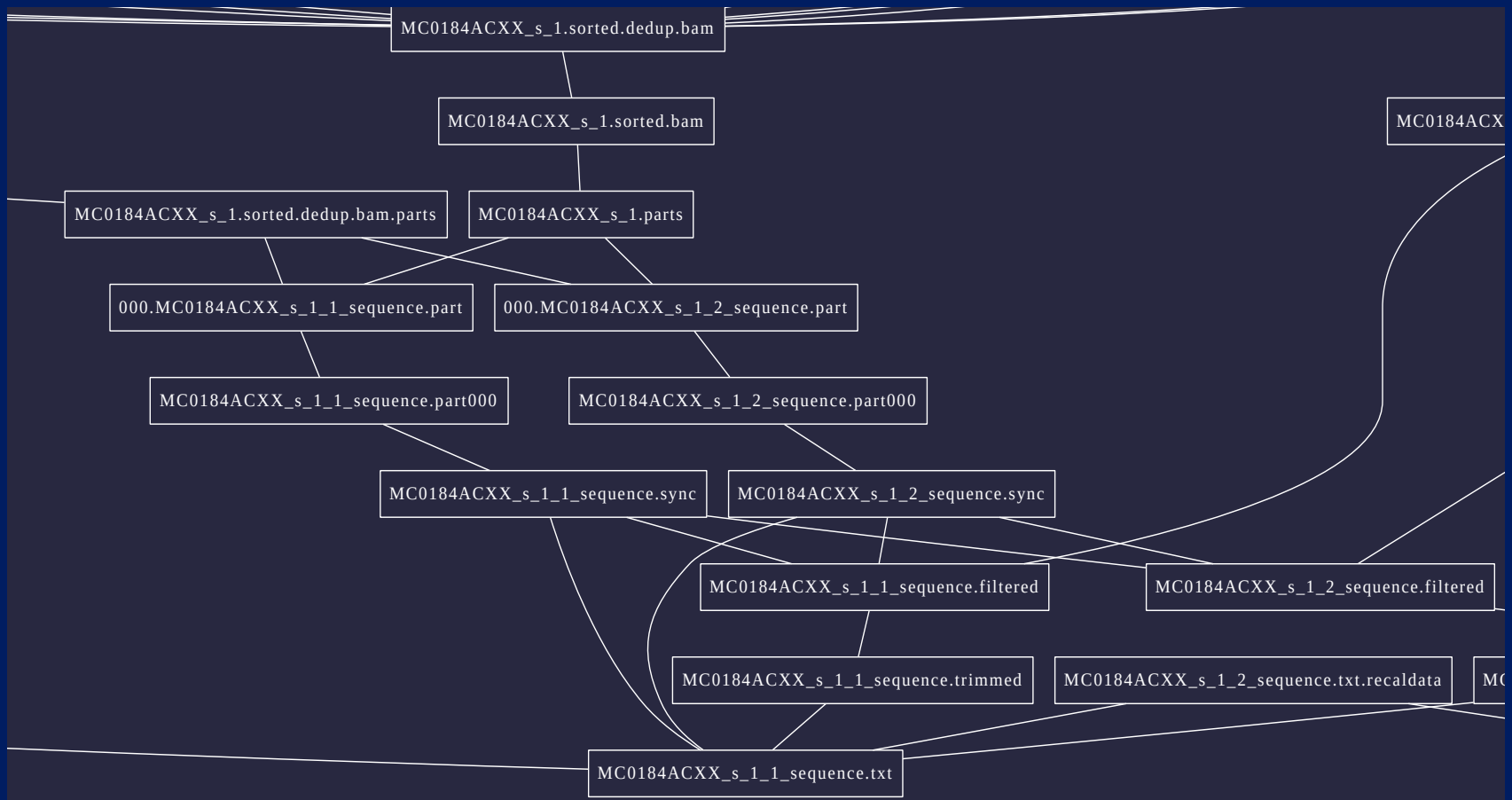



Figure 26: Zoomed in.

# Questions?

Acknowledgements:

Michiel van Galen
Martijn Vermaat
Johan den Dunnen