# EN5636
# Atelier NGS
## Génomique fonctionnelle appliquée

**Enseignants: Daniel Gautheret, Gaëlle Lelandais**
**Remerciements: Yannick Boursin (Gustave Roussy), Frédéric Lemoine (Pasteur)**

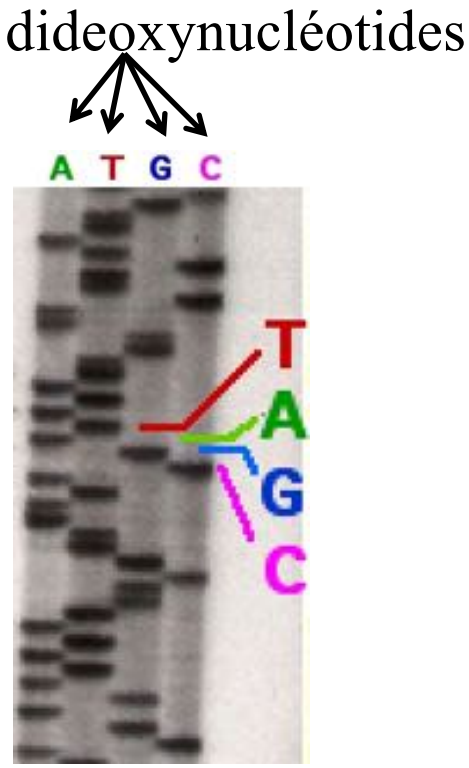
# Plan 2018

- 19/12 09:00-12:30 intro NGS + prise en main Cloud IFB (DG)
- 20/12 14:00-17:30 RNA-seq presentation + pipeline (DG)
- 8/1 9:00-12:30 RNA-seq visu IGV (DG) + debut analyse (GL)
- 09/01 -- (meetU) ---
- 10/01 14:00-17:30 Exome-seq pipeline (DG)
- 17/01 14:00-17:30 Exome-seq pipeline fin + analyse (DG)
- 18/01 14:00-17:30 RNA-seq analyse fin (GL)

# NGS data

# Le séquençage de Sanger (1977)

- Séquençage par terminaison de chaine
  - Synthèse interrompue à un certain type de base.
  - 4 types= 4 réactions
- Amélioré en 1987 par l'introduction de marqueurs flurorescents (1 seule réaction) et l'automatisation.



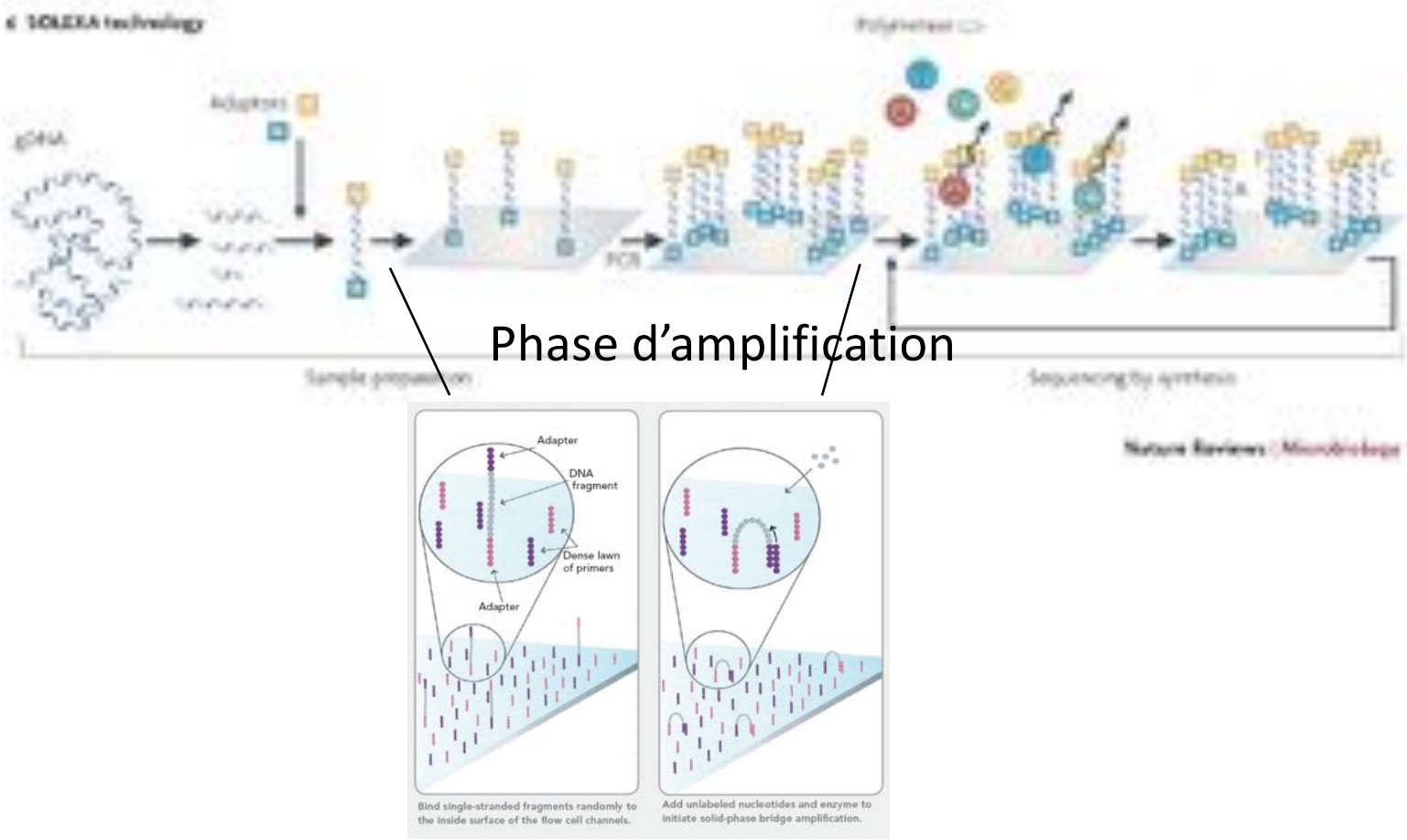
Wikipedia

# NGS : Next Generation Sequencing (2005-)

**Faster, Cheaper, Deeper**

# NGS: parallélisation du séquençage

Solexa/Illumina




Phase d'amplification

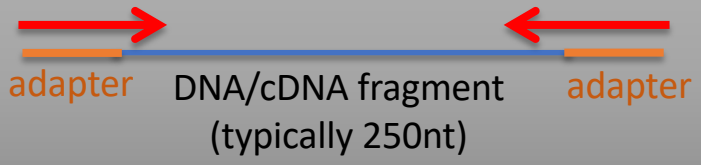
# Sequencers & output






| | Nanopore MinIon | Lifetech Ion torrent PGM | Illumina MySeq | Lifetech Ion proton | Illumina Hi-Seq 2000 | Illumina NovaSeq |
|---|---|---|---|---|---|---|
| | 1Gb | 400 Mb | 4 Gb | 20 Gb | 300 Gb | 3Tb |
| Read size | >100kb | 35-400 | 2x200 | 35-400 | 2x150 | 2x150 |

**Single** vs. **paired end** sequencing



| Instrument | Durée | Millions de Reads/run | Bases/read | Gb/Run |
|---|---|---|---|---|
| **Applied Biosystems 3730** | 2h | 0,000096 | 650 | 0,00006 |
| **454 GS Jr. Titanium** | 10h | 0,1 | 400 | 0,1 |
| **454 FLX Titanium** | 10h | 1 | 400 | 0,4 |
| **454 FLX+** | 20h | 1 | 650 | 0,7 |
| **Illumina GA IIx v5 SE** | 2j | 640 | 36 | 23 |
| **Illumina GA IIx v5 PE** | 14j | 640 | *288* | 184,3 |
| **Illumina MiSeq v2 Nano** | 17h | 1 | *300* | 0,3 |
| **Illumina MiSeq v2 Micro** | 19h | 4 | *300* | 1,2 |
| **Illumina MiSeq v3** | 20h | 22 | *150* | 3,3 |
| **Illumina MiSeq v3** | 55h | 22 | *600* | 13,2 |
| **Illumina NextSeq 500 Mid** | 15h | 130 | *150* | 19,5 |
| **Illumina NextSeq 500 High** | 18h | 400 | *150* | 60 |
| **Illumina HiSeq 2500 Rapid run** | 27h | 300 | *200* | 60 |
| **Illumina HiSeq 2500 v3** | 11j | 1500 | *200* | 300 |
| **Illumina HiSeq X (2 flow cells)** | 3j | 6000 | *300* | 1800 |
| **Ion Torrent – PGM 314 chip** | 2,3h | 0,475 | 200 | 0,1 |
| **Ion Torrent – PGM 314 chip** | 3,7h | 0,475 | 400 | 0,2 |
| **Ion Torrent – PGM 316 chip** | 3h | 2,5 | 200 | 0,5 |
| **Ion Torrent – PGM 316 chip** | 4,9h | 2,5 | 400 | 1 |
| **Ion Torrent – PGM 318 chip** | 4,4h | 4,75 | 200 | 1 |
| **Ion Torrent – PGM 318 chip** | 7,3h | 4,75 | 400 | 1,9 |
| **Ion Torrent - Proton I** | 4h | 70 | 175 | 12,3 |
| **Ion Torrent - Proton II** | 5h | 280 | 175 | 49 |
| **Ion Torrent - Proton III** | 6h | 500 | 175 | 87,5 |
| **Life Technologies SOLiD 5500xl** | 8j | 1410 | 110 | 155 |
| **Pacific Biosciences RS II** | 2h | 0,03 | 3000 | 0,1 |
| **Oxford Nanopore MinION** | ≤6h | 0,1 | 9000 | 0,9 |

*[ANPGM GT1] Recommandations informatiques pour les analyses NGS, 2016*

# Les grandes applications des NGS

- DNA-seq (variants génomiques)
- RNA-seq (transcriptome)
- ChiP-Seq (sites de liaisons à l'ADN)
- Autres applications
  - Hi-C, clip-seq, net-seq, ribosome profiling etc.

# DNA-seq (variants génomique)

- Panel de gènes
  - Une série d'exons d'intérêt (gènes de cancer= 100kb)
- Exome
  - Tous les exons du génome (30 Mb)
- Whole genome
  - Le génome complet (3 Gb)

# DNA-seq pour la Génétique constitutionnelle

At hospital

Blood sample → DNA-seq → Look for specific alteration (BRCA)

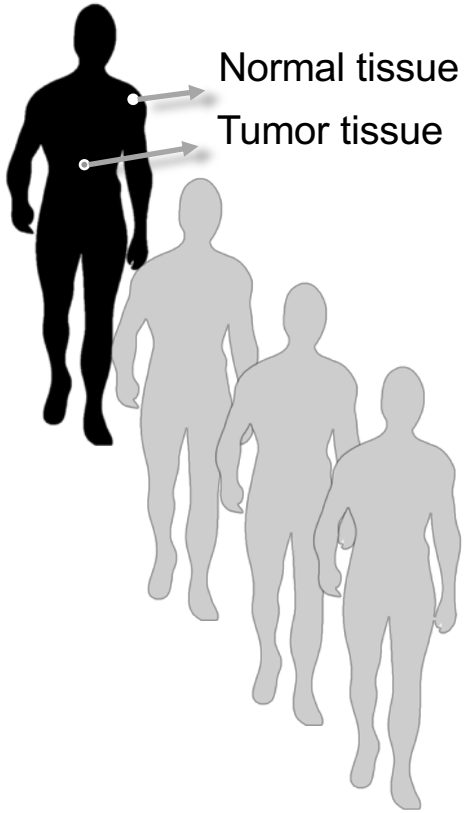
Research

DNA-seq → Compare disease and healthy cohorts

GWAS studies, 1000 Genome Project…

# Et pour la génétique somatique

## Finding somatic mutations in the tumor genome



Normal tissue (blood) → Extract and sequence DNA

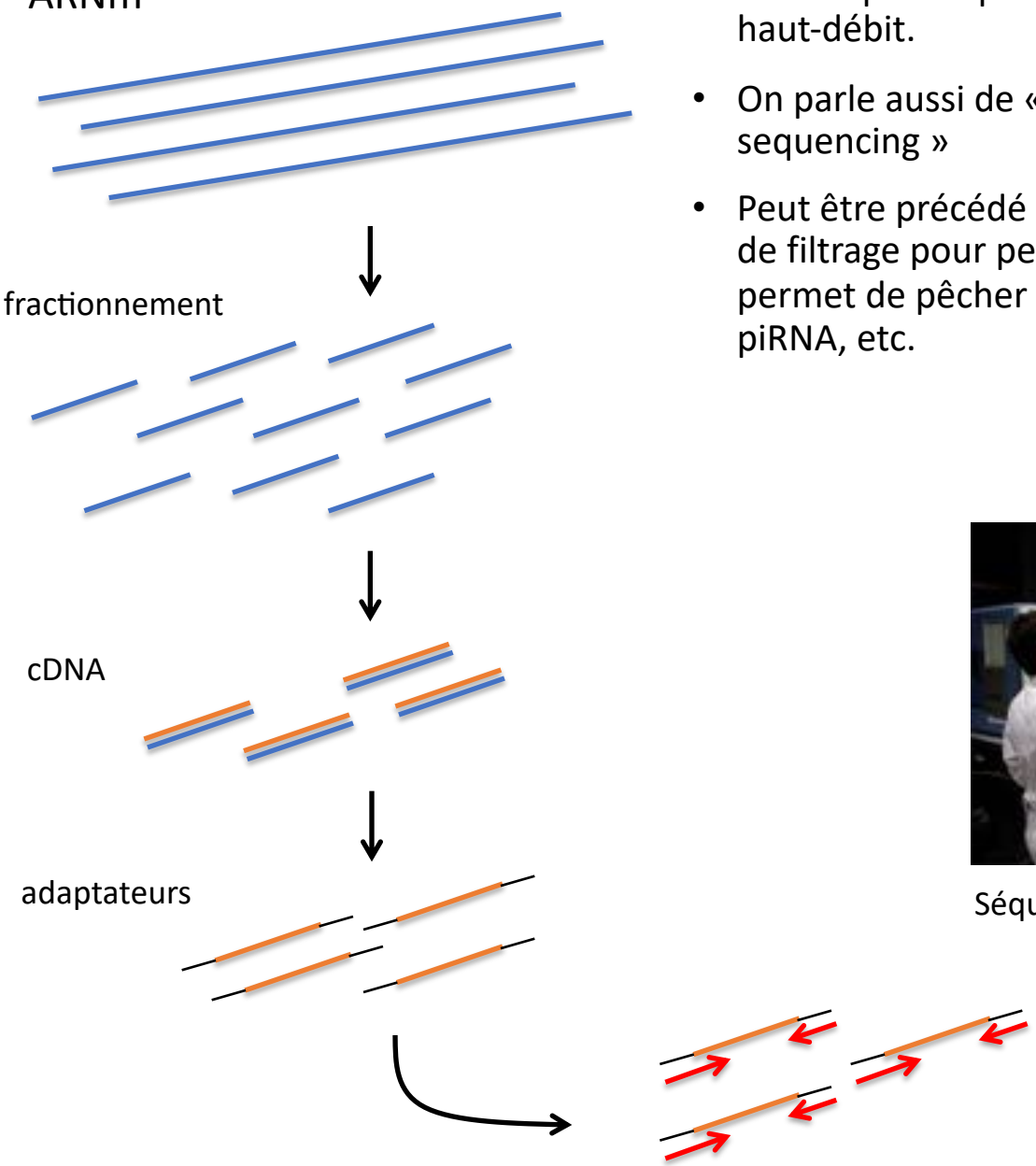
Tumor tissue → Look for mutations, structural variations

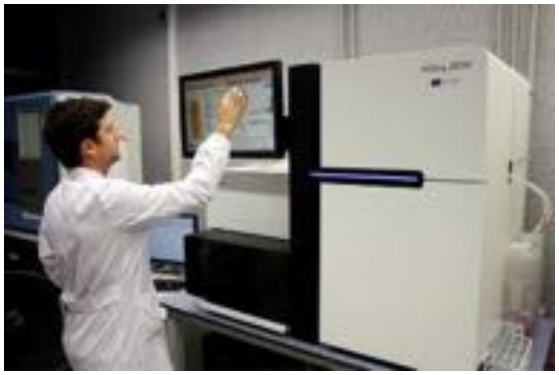
Repeat for several patients → Look for recurrent events

- 100-10,000 differences between a tumor and normal tissue from the same person

# RNA-Seq

- Transcriptome par séquençage haut-débit.
- On parle aussi de « deep sequencing »
- Peut être précédé d'une étape de filtrage pour petits ARN, permet de pêcher les miRNA, piRNA, etc.

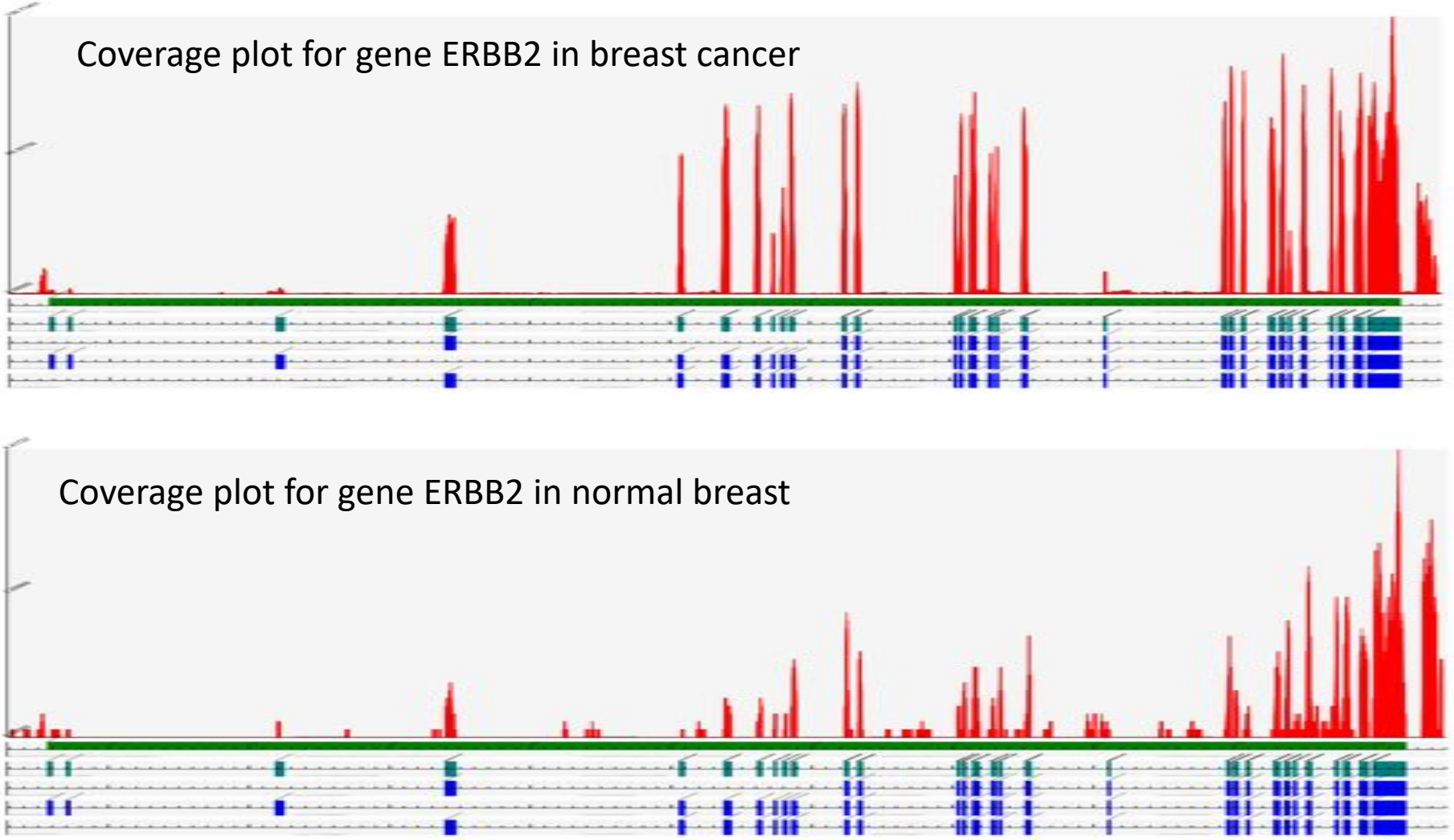


Séquençage

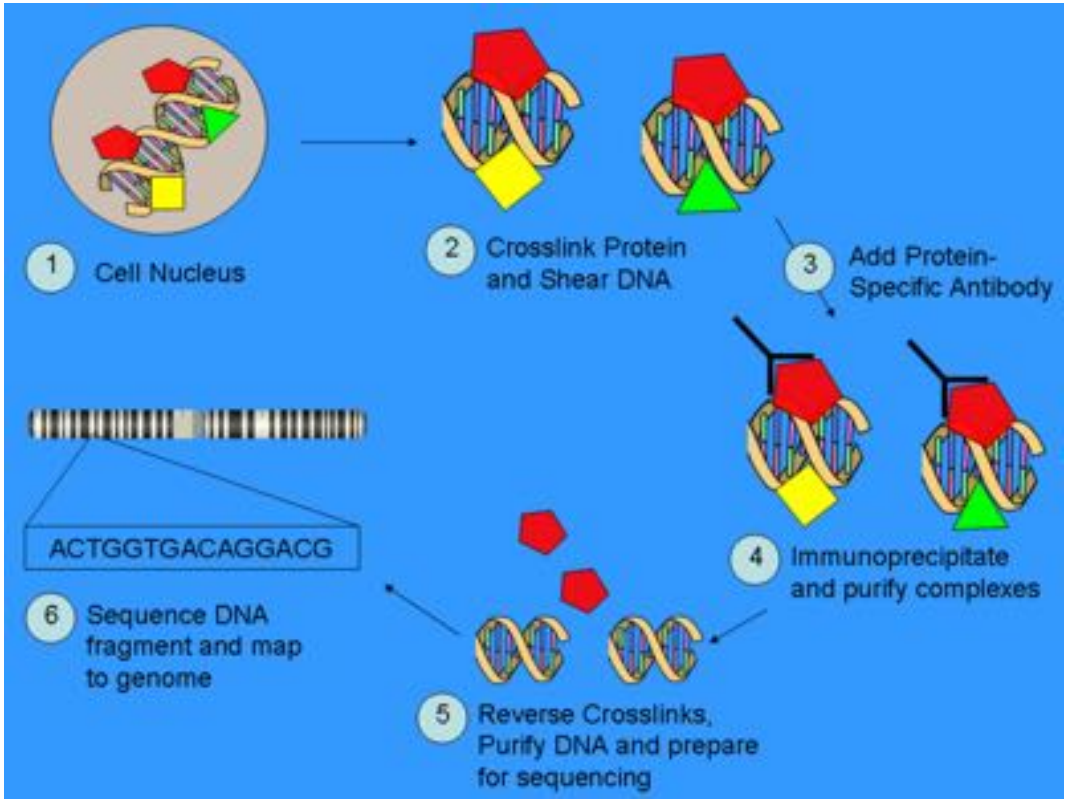

13

# Differential expression



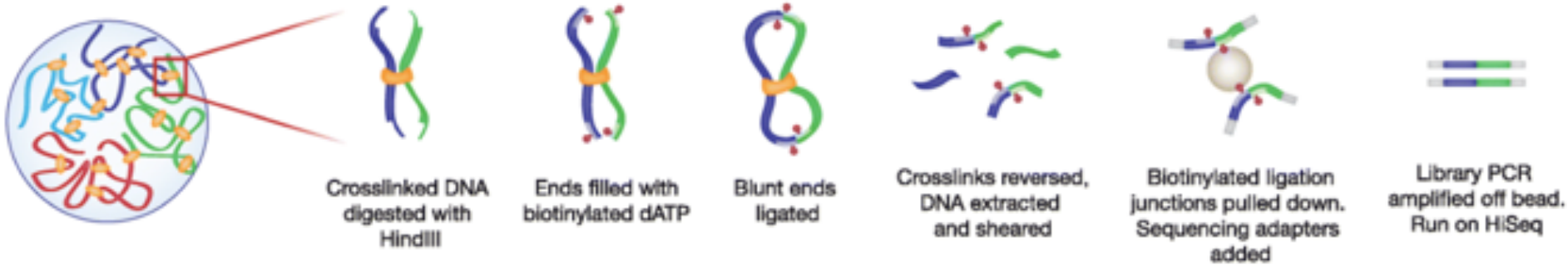
# ChIP-Seq

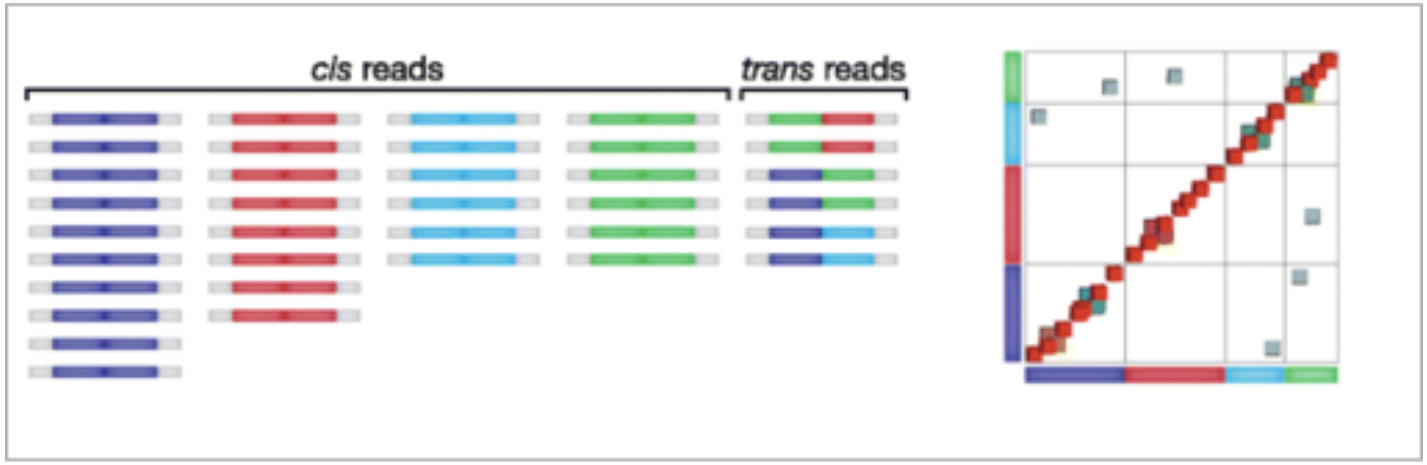
- ChIP=Chromatin immunoprecipitation
- Permet d'identifier les sites de liaison de protéines (facteurs de transcription, represseurs, enhancers, etc.) sur l'ADN génomique




Wikipedia

# Hi-C




Harewood et al. Genome Biol 2017

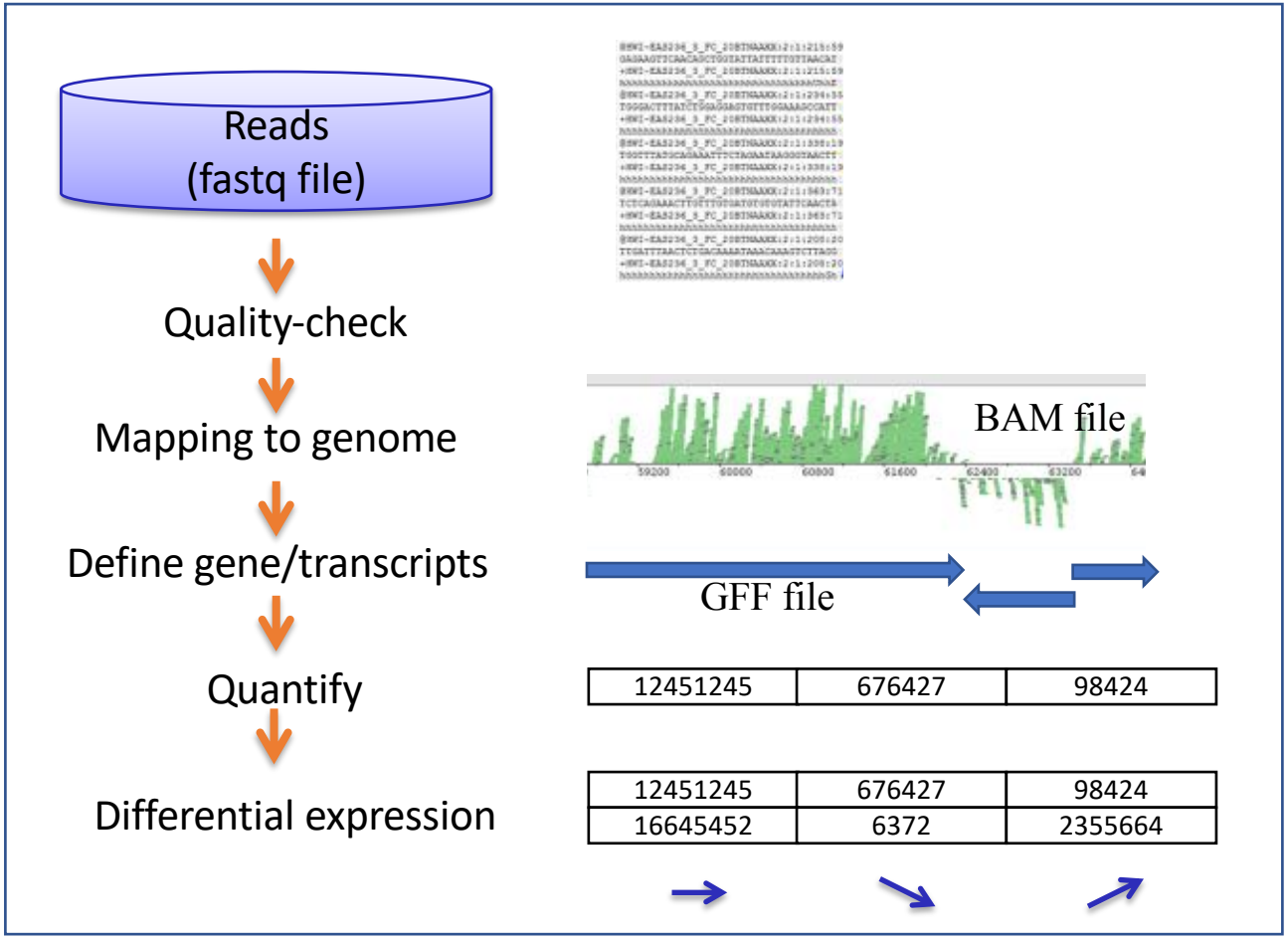
# NGS pipelines

# « Pipelines » & « workflows »

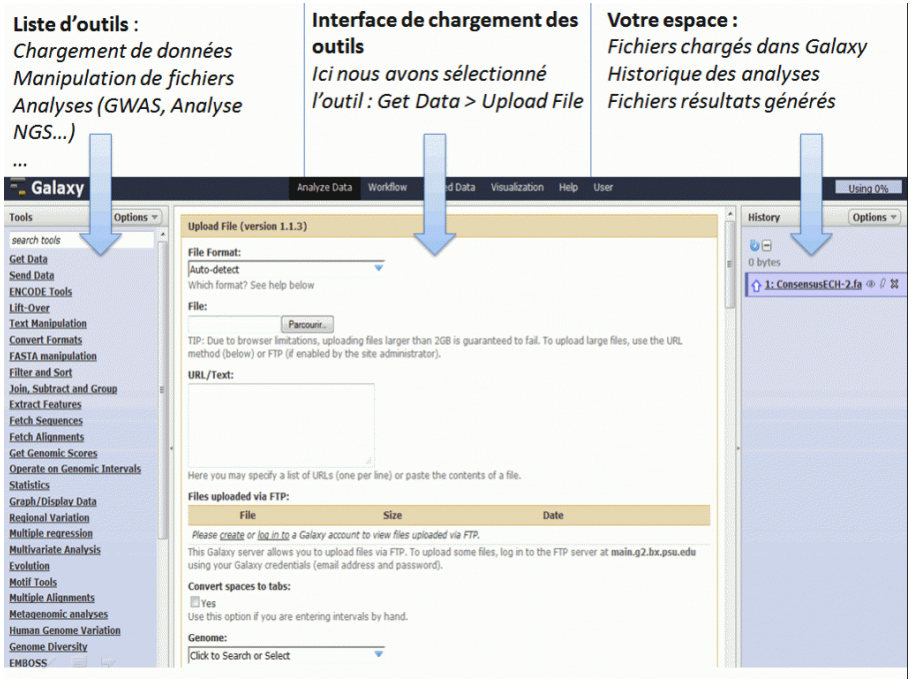
« Bricks » from Unix open source programs

Combined into pipelines (typically a few hours to days to run)

Example: an RNA-seq pipeline



# Galaxy: user-friendly interface to NGS pipelines



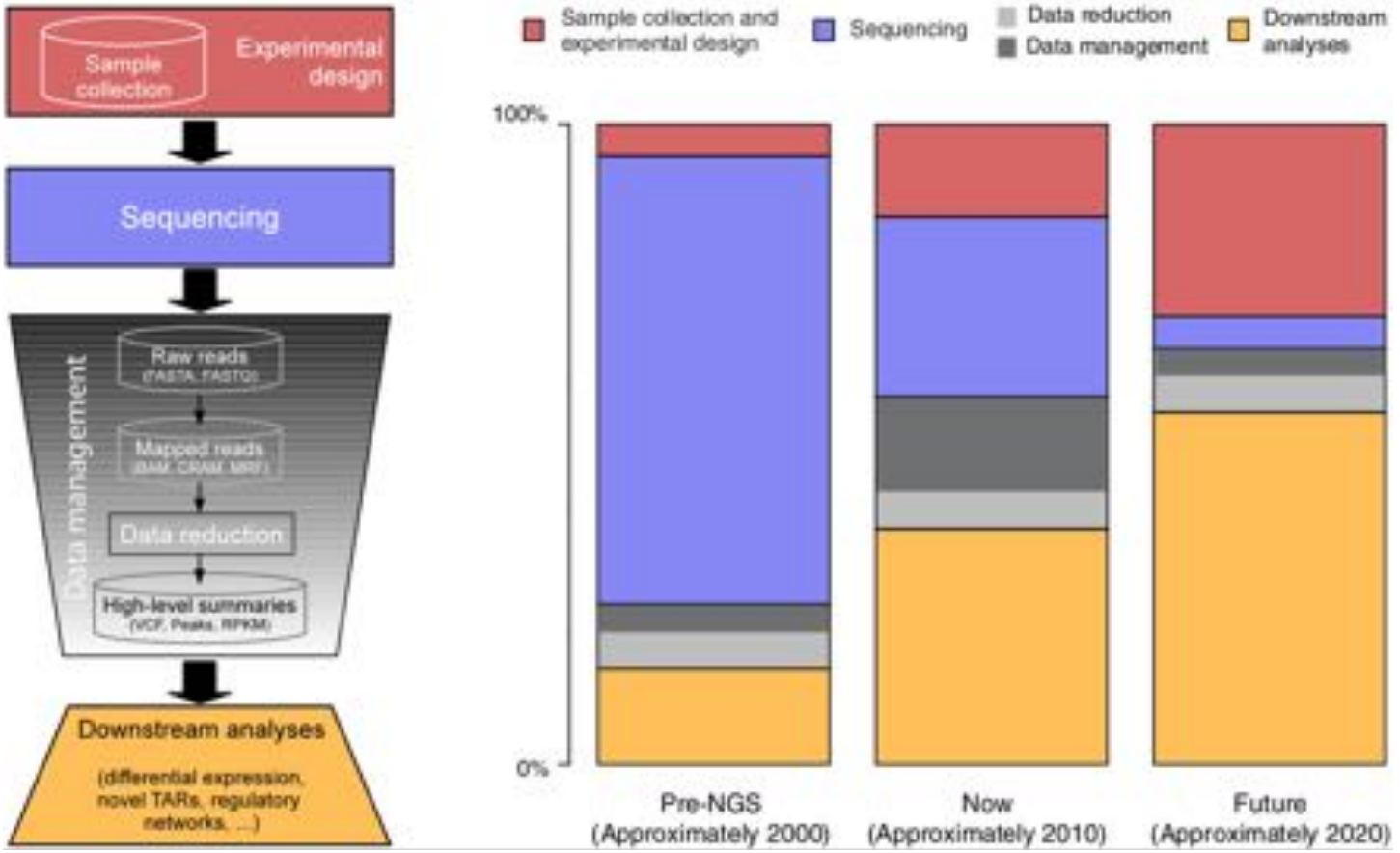

Credit: Biorigami

- Interest: avoiding Unix command line + traçability
- But: running NGS workflow on real human data often requires a computer cluster (will not run on a single-node Galaxy server)

# Volume des données NGS

- Un exome humain avec fichiers de mapping et analyse: 50 Go
- Données génomiques produites annuellement dans un hôpital universitaire: 100-500 To
- La banque TCGA complete: 1 Po

# Evolution of sequencing cost *vs.* bioinformatics cost

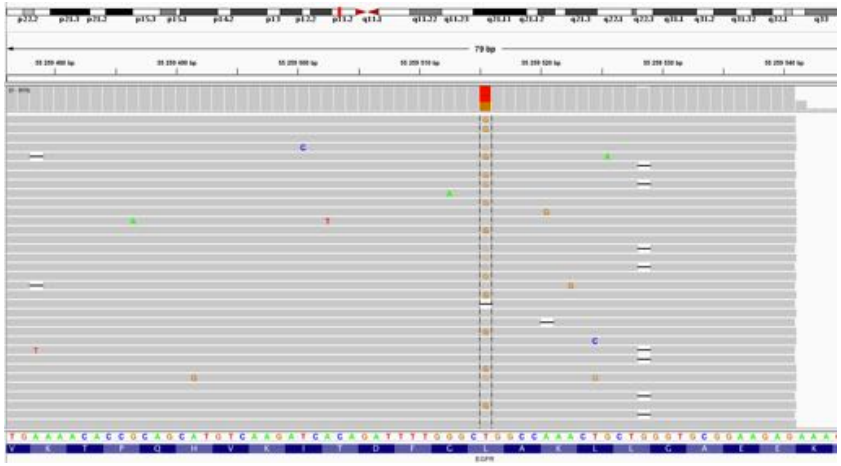



Sboner et al. 2011. Genome Biology, **12**:125

# NGS: Mastering the esser

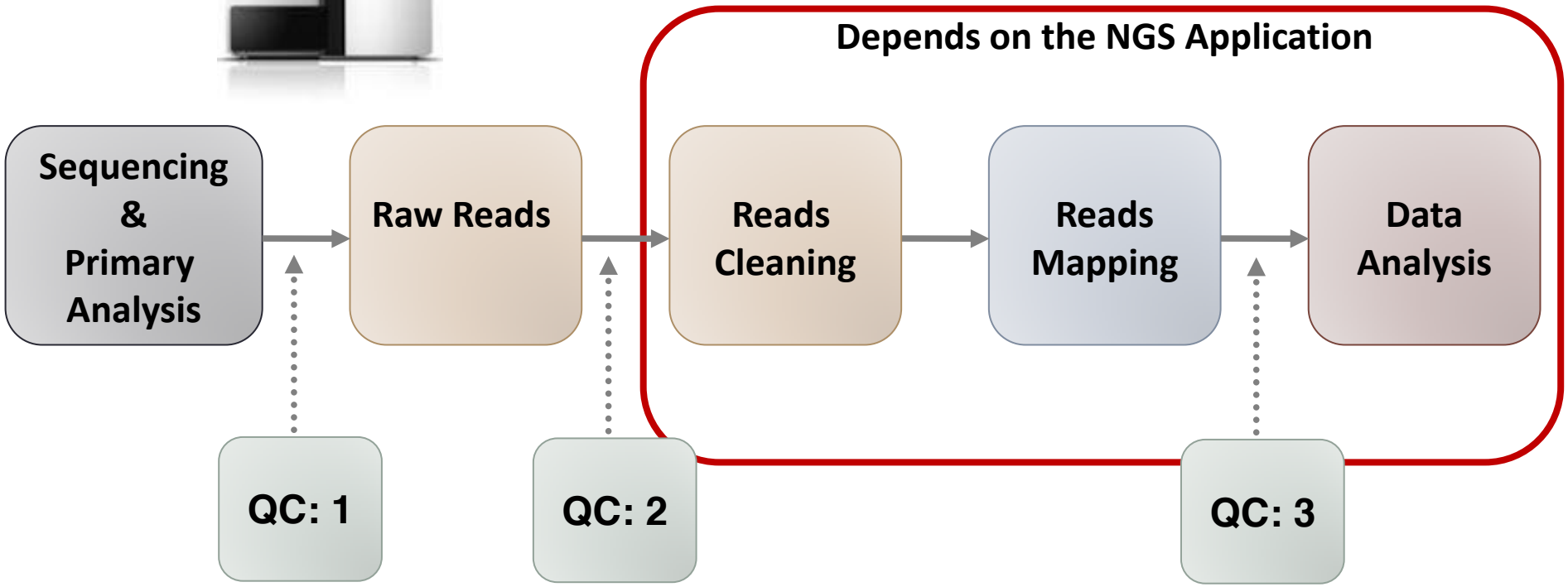
LINUX Shell

File formats: Fastq, BAM, VCF, BED, GFF



Browsers

# Standard Workflow for NGS Analysis

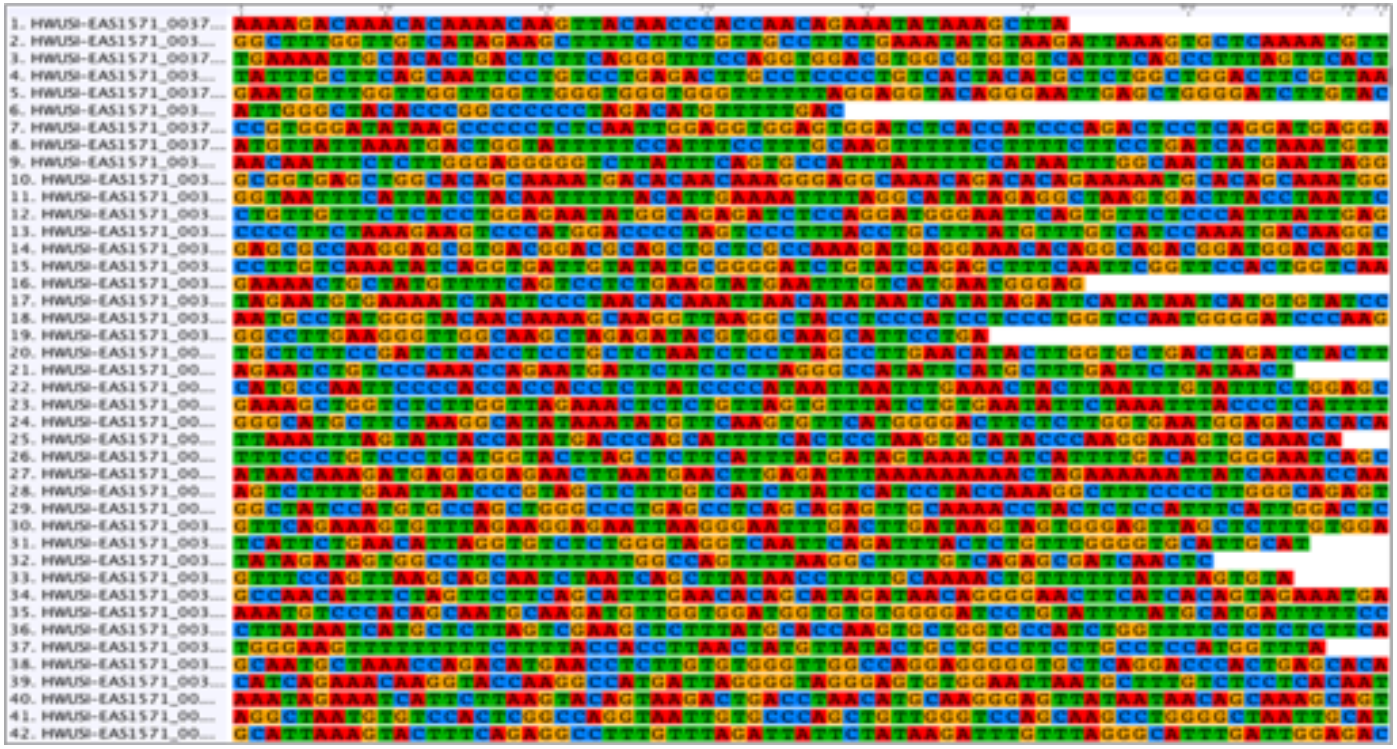



A typical NGS workflow

# Step 1: Quality Check and Cleaning

# NGS Data: what do they look like ?

A **raw** data file (.fastq, .sff, .fa, .csfasta/.qual)
with **millions** of short reads of the **same size** (SOLiD, HiSeq) or reads of different size (Ion PGM/Proton)



Enhanced view of the reads in a fastq file

# Format fasta

*.fa , *.fasta

```
>identifiant1 commentaire libre
CAGCATCGATCGTCGGCGATGCATGCGGATGCTAGCTGATCACGATGC
CGCATGCTAGTCAGGCGGGAGGGATATTATTAGCGGCGTATCGGATGA
CAGCATTTACGGCGGGAGTGCTATTATTTATGAGCGGCGAT
>identifiant2 commentaire libre
CAGGCGGAGGTTCTTTATTATATCGGCGGGCGGAGGCGGCGATGCATC
CAGTGCACTGACGTAGTCAGCGATGCATTTTATGACTGACTCAGTTTT
CCCGCTAGCTATGCTATGCTATTCGATCGATTCGTGAGCTGATCTGGC
CAGCTATGCTTTAGTA
```

# Format FASTQ

- 1 sequence = 1 read = 4 lines in the file

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

- First line = sequence identifier

```
@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG
```

| | |
|---|---|
| **EAS139** | the unique instrument name |
| **136** | the run id |
| **FC706VJ** | the flowcell id |
| **2** | flowcell lane |
| **2104** | tile number within the flowcell lane |
| **15343** | 'x'-coordinate of the cluster within the tile |
| **197393** | 'y'-coordinate of the cluster within the tile |
| **1** | the member of a pair, 1 or 2 *(paired-end or mate-pair reads only)* |
| **Y** | Y if the read fails filter (read is bad), N otherwise |
| **18** | 0 when none of the control bits are on, otherwise it is an even number |
| **ATCACG** | index sequence |

# Format FASTQ

- Fourth line = Quality

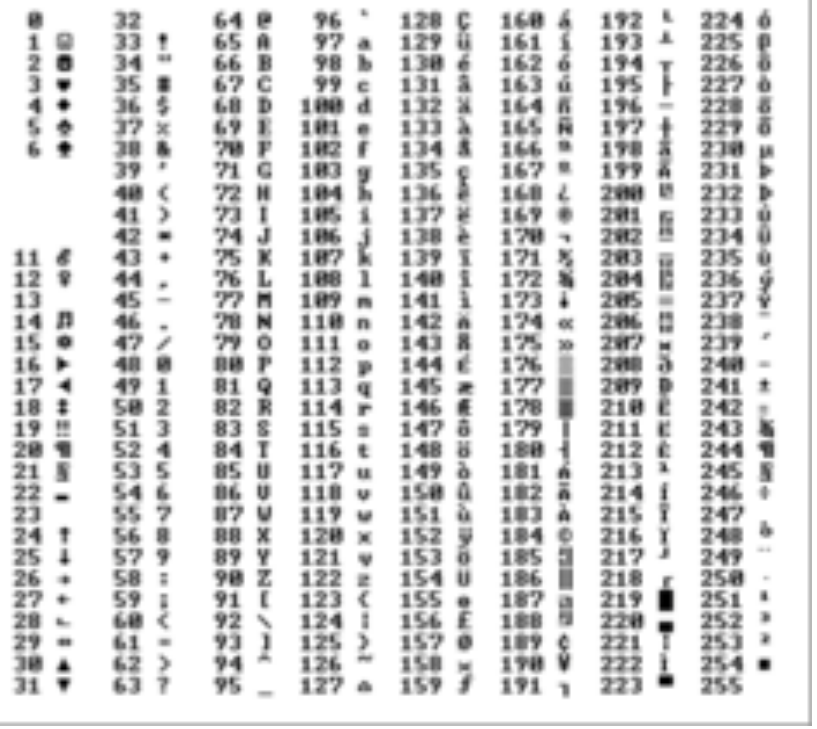
```
!''*((((****+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

# Qualité dans le format fastq

Qualité=-10 $\log_{10}(P_{erreur})$

@SEQ_ID1
GATTTGGGGT
+
42,40,134,47,36 35 40 40 40

@SEQ_ID1
GATTTGGGGT
+
⋆(å/$#((((



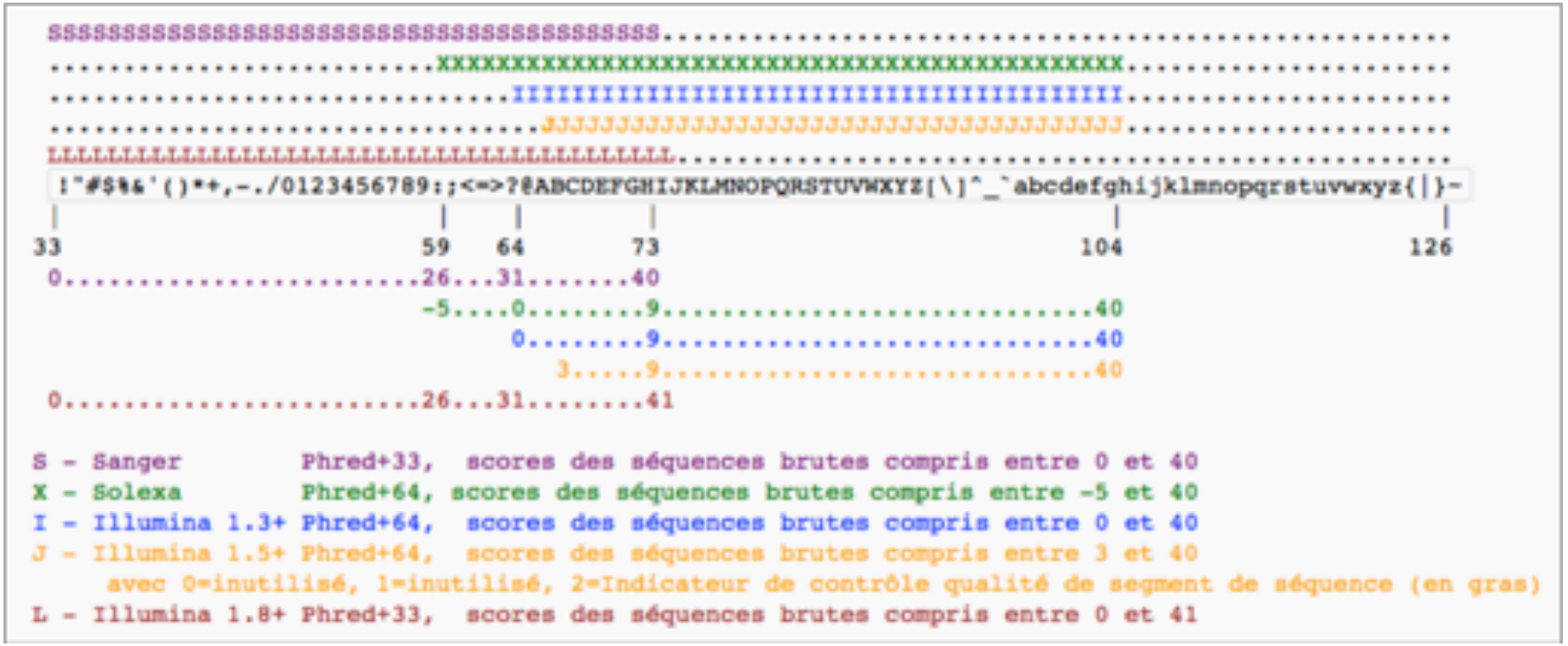
# Sequence quality encoding

*Phred scores Q* : Q scores are defined as a property that is logarithmically related to the base-calling error probabilities (P).

$$Q = -10 \log_{10} P$$

| Phred Quality Score | Probability of Incorrect Base Call | Base Call Accuracy |
|---|---|---|
| 10 | 1 in 10 | 90% |
| 20 | 1 in 100 | 99% |
| 30 | 1 in 1,000 | 99.9% |
| 40 | 1 in 10,000 | 99.99% |
| 50 | 1 in 100,000 | 99.999% |

# Différents encodages ASCII



# Why looking at sequencing quality ?

- Quality of data is very important for various downstream analyses:
    - Sequence assembly or mapping
    - Variants detection
    - Gene expression studies
    - …
- Quality of data = poor
    - Try to find a reason
    - Can we correct/improve the quality ?
    - May lead to erroneous conclusions

# Quality controls on raw reads: which metrics to check ?

Mainly:

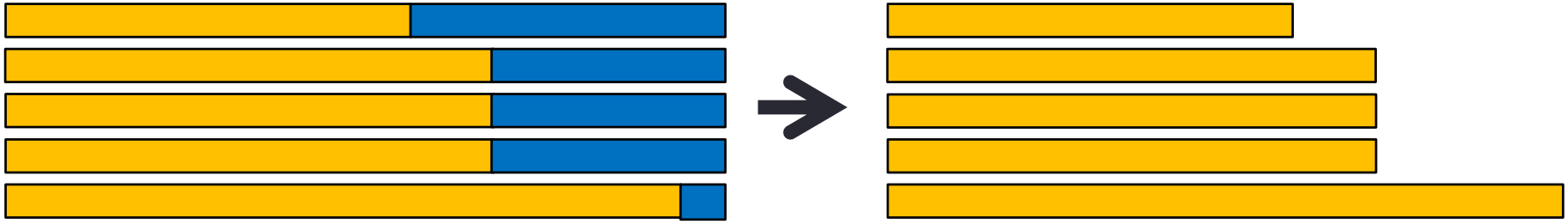
- Quality score per base and over the reads

But also:

- Read length distribution
- Sequence content per base and % of GC
- Kmers content
- Overrepresented sequences
- Duplicated reads

# Quality control on raw reads: adapters removal

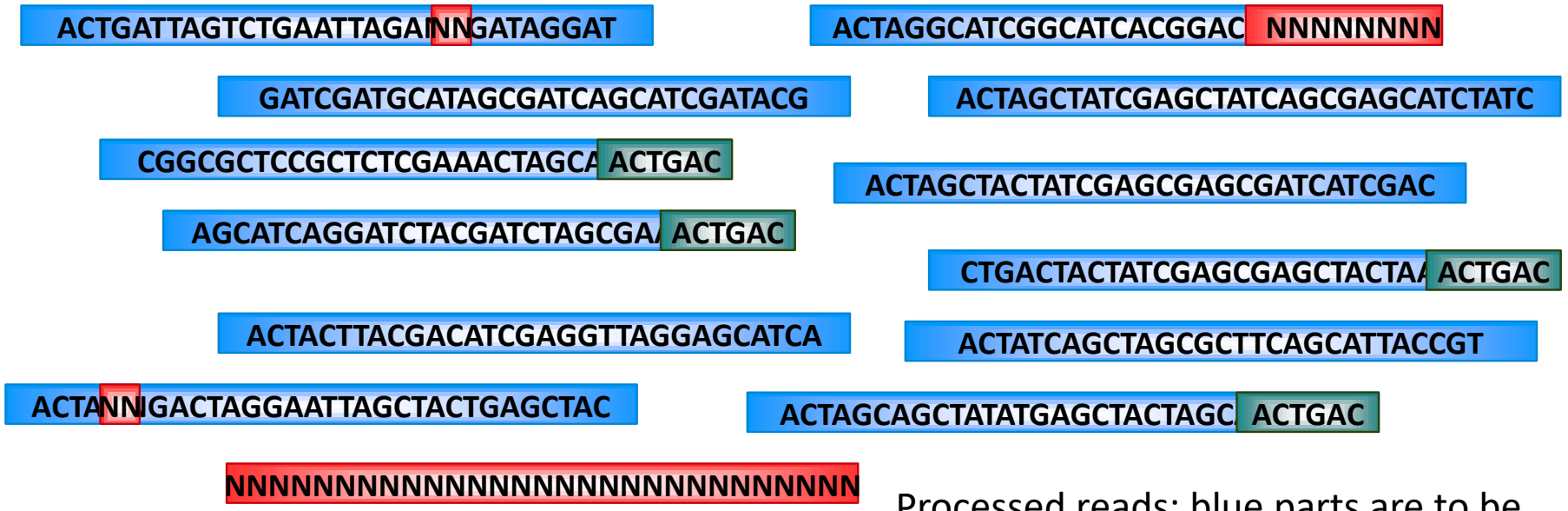
- An adapter is a small piece of known DNA located at the end of the reads
- Adapters roles:
    - Hang read to the sequencer flowcell
    - Allows a specific PCR enrichment of reads having adapter
    - Use in multiplex sequencing (samples in mix)
- Available tools to trim adapters:
    - Cutadapt
    - Trimmomatic



**In blue**: adapters. **In orange**: informative part of the read.

# Quality controls on raw reads : let's start after sequencing

A first Quality Control of raw reads is mandatory and can be established according to the application ('N', adapter sequences, barcode, contamination, etc.)




Processed reads: blue parts are to be kept, green and red parts to be removed

# Quality scores

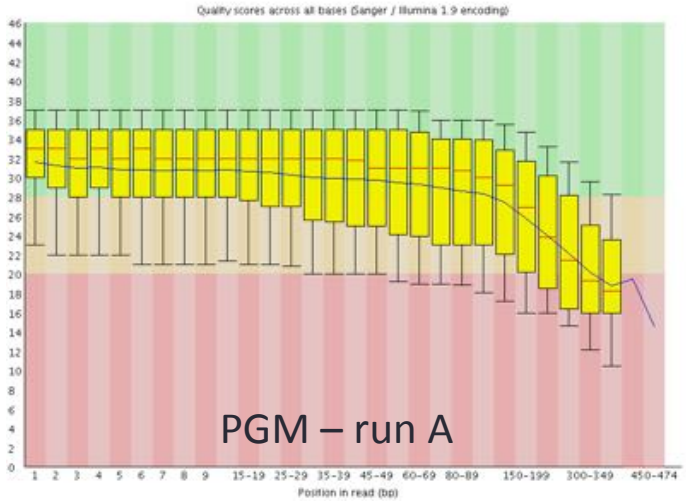
- Per base (Box Whisker type plot)

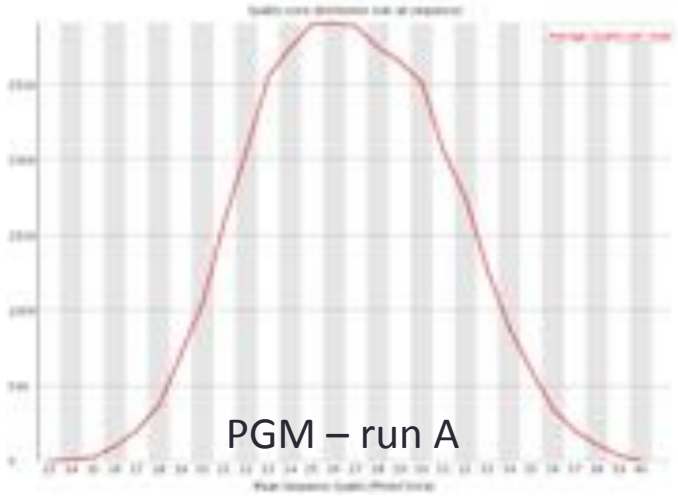
-> to see wether base calls falls into low quality (commonly towards the end of a read)

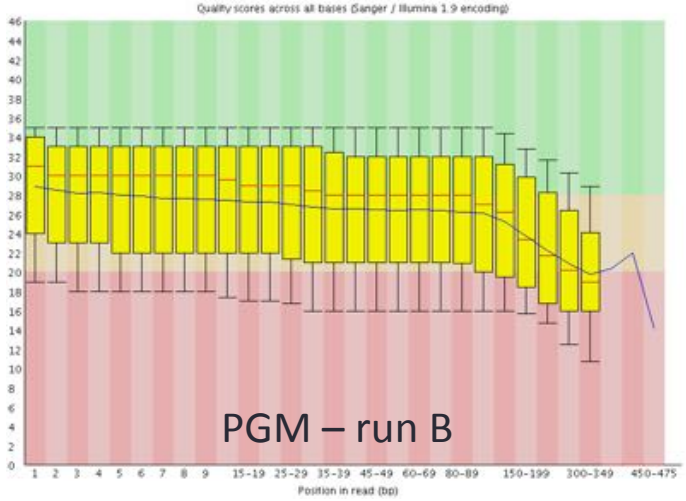
- Per sequence (mean quality distribution)

-> to see if a subset of your sequences have universally low quality values

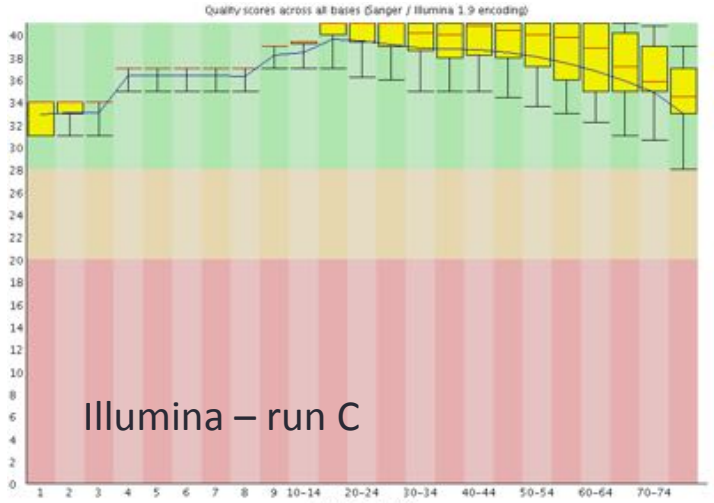
# Quality scores

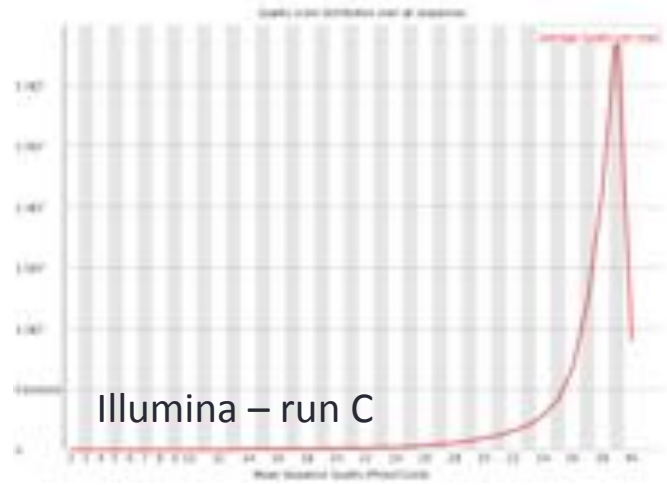


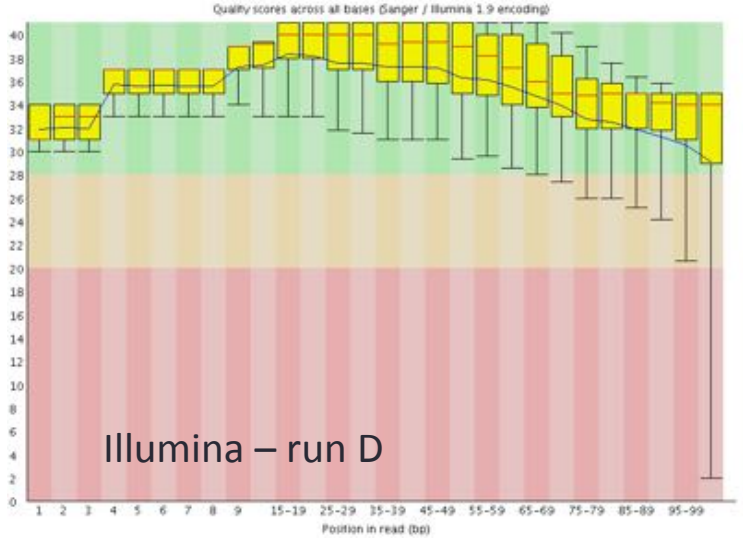



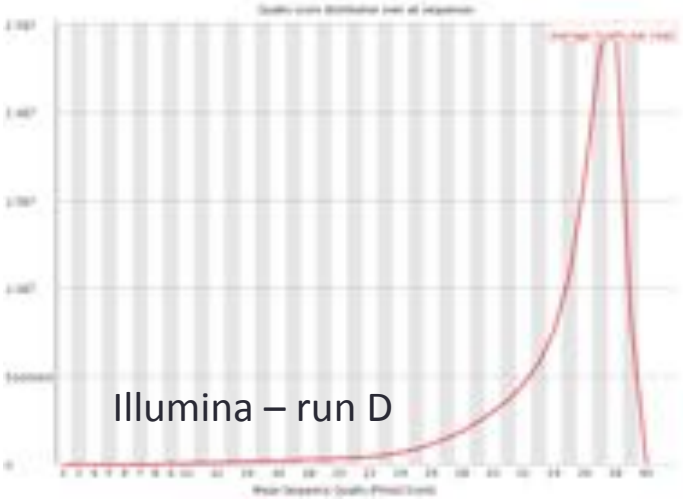




# Quality scores

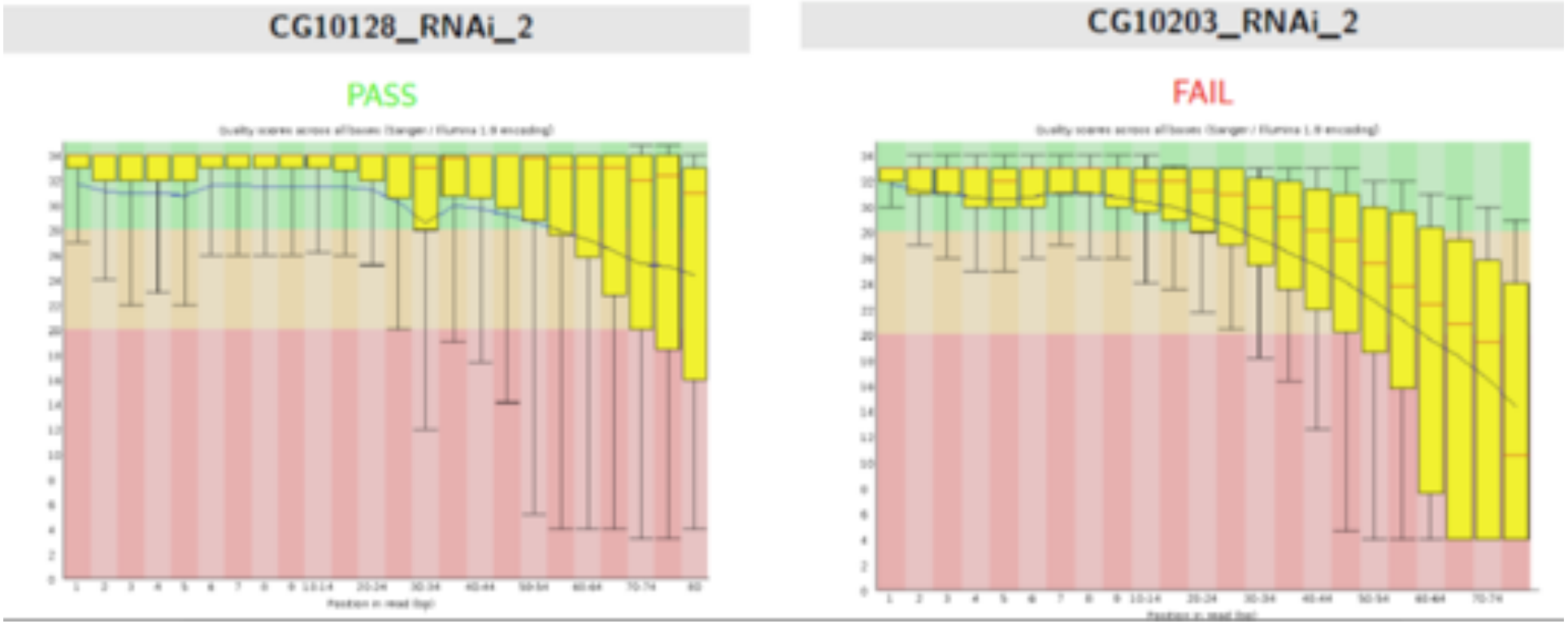


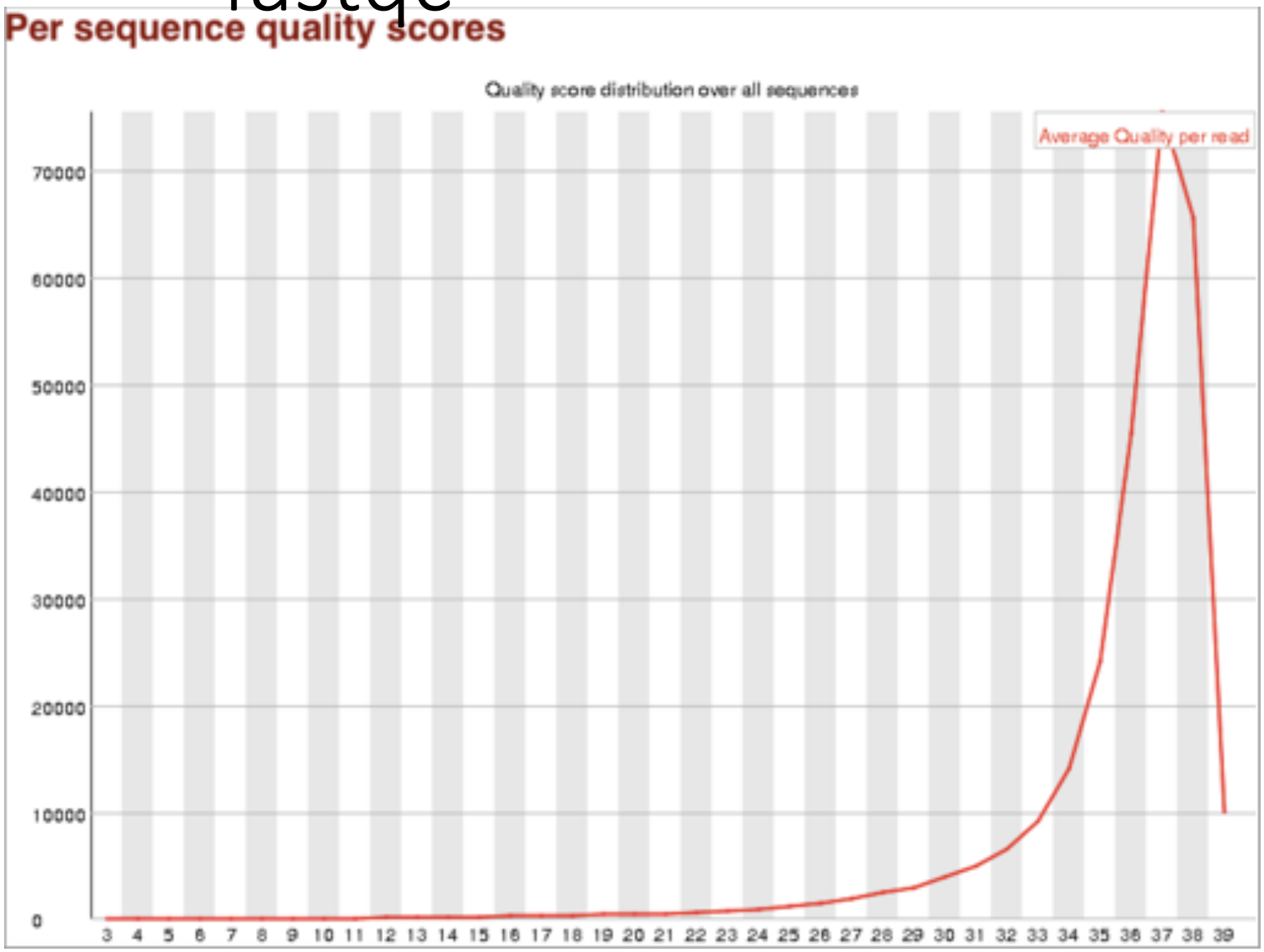



# Fastqc

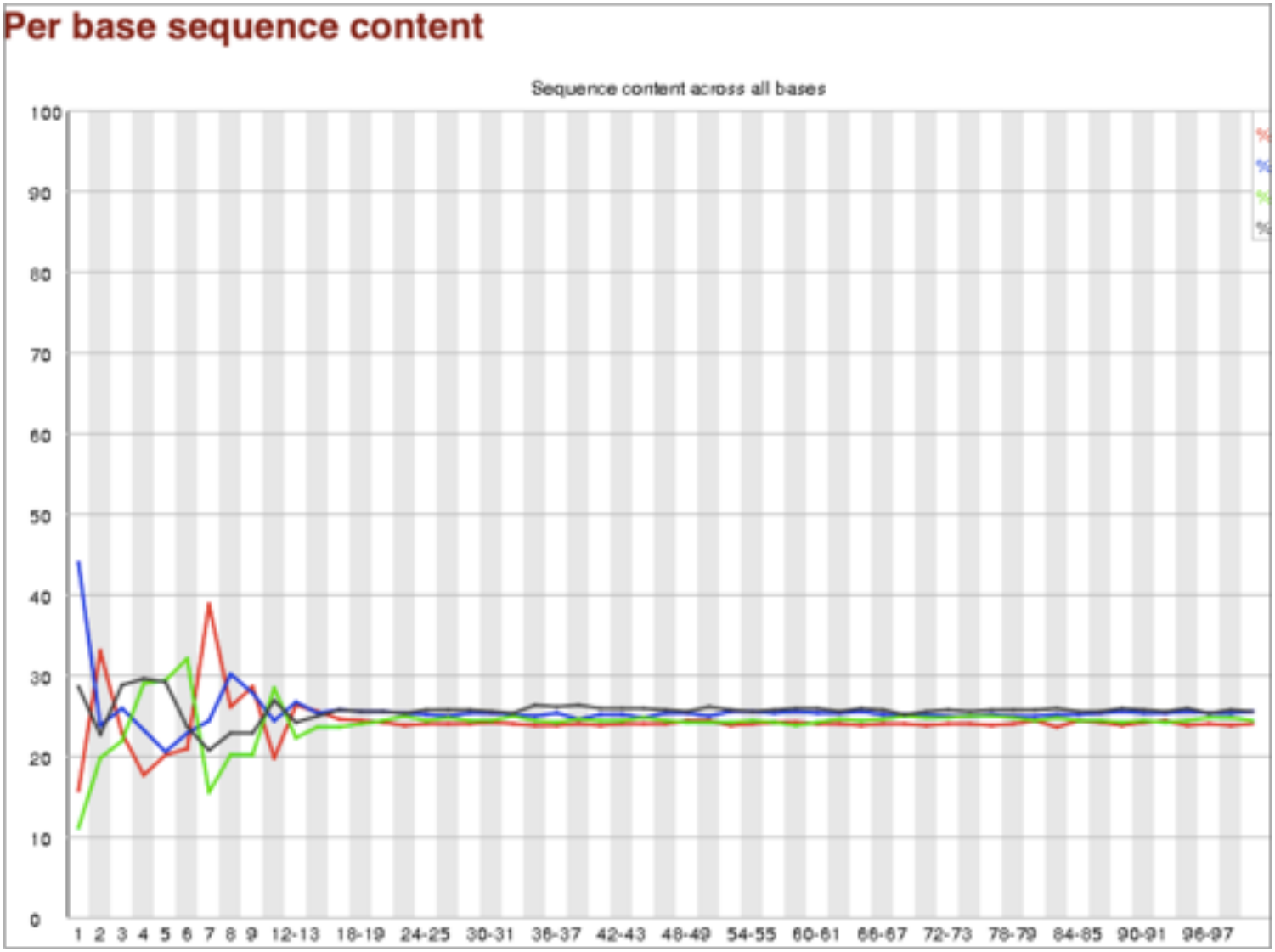



Pour couper les extrémités de basse qualité: FastqTrimmer

# fastqc


Per sequence quality scores
Quality score distribution over all sequences
Average Quality per read
70000
60000
50000
40000
30000
20000
10000
0
3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39

# fastqc


Per base sequence content
Sequence content across all bases
100
90
80
70
60
50
40
30
20
10
0
1 2 3 4 5 6 7 8 9 12-13 18-19 24-25 30-31 36-37 42-43 48-49 54-55 60-61 66-67 72-73 78-79 84-85 90-91 96-97

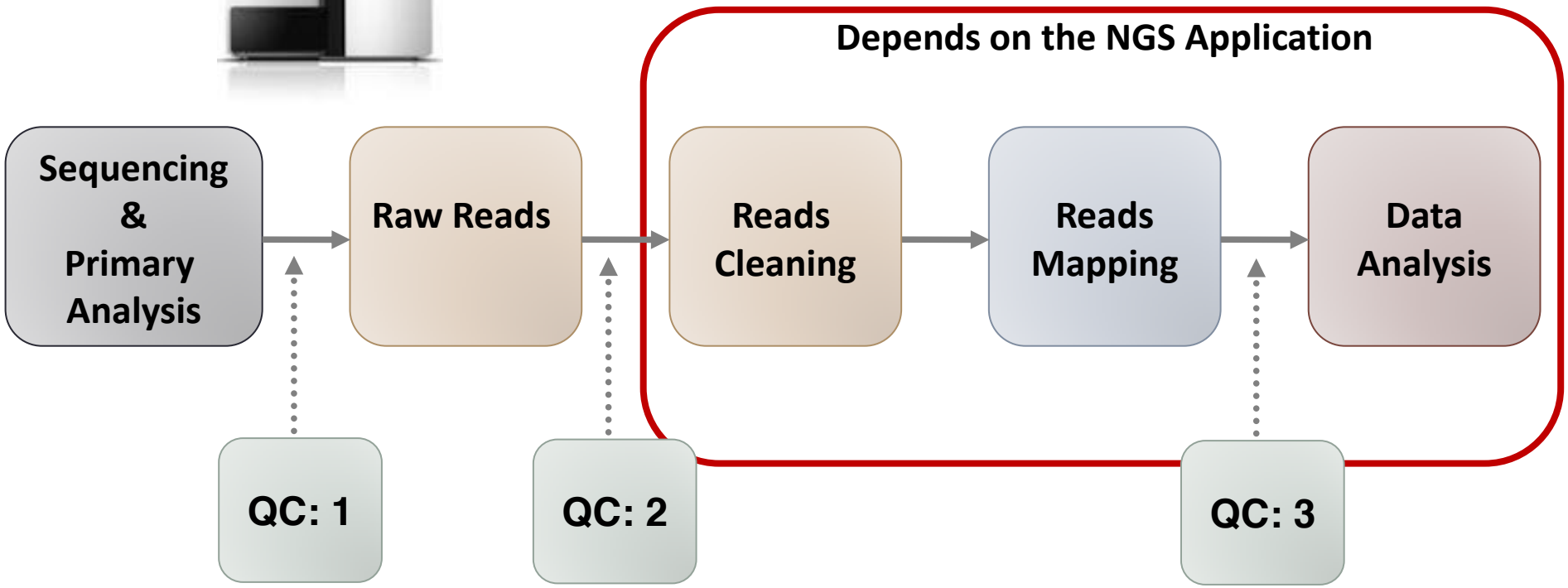
# fastqc

- Adaptateurs et séquences surreprésentées

## Overrepresented sequences

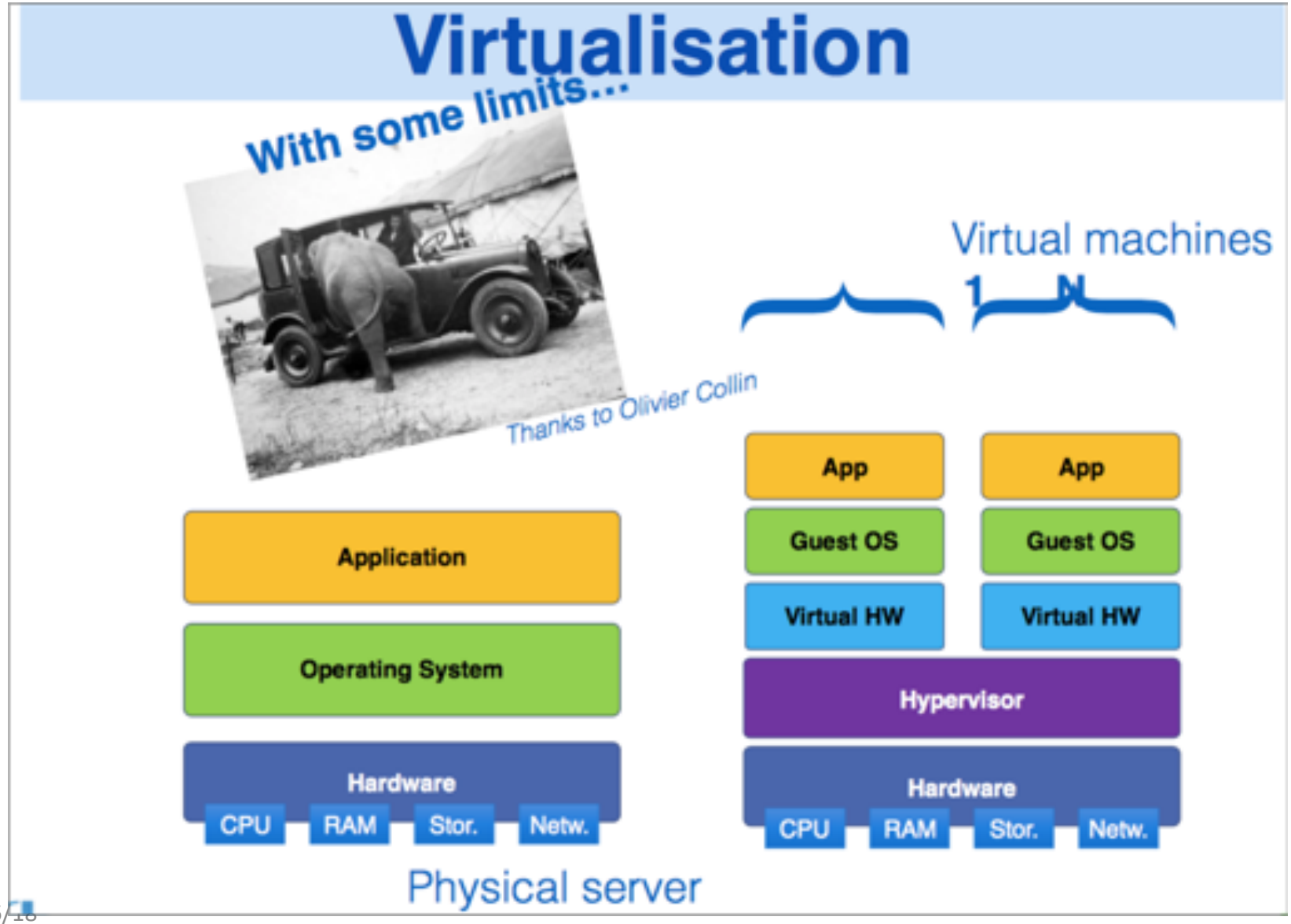
| Sequence | Count | Percentage | Possible Source |
|---|---|---|---|
| TTTTTTTGGAAACCTCTGCGCCATGAGAGCCAAGTGGAGGAAGAAGCGAA | 608 | 0.22002120599123534 | No Hit |
| TTTTTGGAAACCTCTGCGCCATGAGAGCCAAGTGGAGGAAGAAGCGAATG | 478 | 0.17297719813126725 | No Hit |
| CTCCAGTCAAAAGTTCTTTGAGACGATGCCATCGGCCTTGGCCAATCGGA | 411 | 0.14873144023420679 | No Hit |
| TTTTTTGGAAACCTCTGCGCCATGAGAGCCAAGTGGAGGAAGAAGCGAAT | 356 | 0.1288282061396049 | No Hit |
| GCAGGCGCAGCCCAGCCTCGAAATGCAGAACGACGCCGGCGAGTTCGTGG | 337 | 0.1219525434523788 | No Hit |
| CAGGCGCAGCCCAGCCTCGAAATGCAGAACGACGCCGGCGAGTTCGTGGA | 308 | 0.11145811092977054 | No Hit |
| CGCAGATAGCATAAGTTTTAAACTGGCCATTAAACCTGCCTGTGACCTTG | 288 | 0.1042205712590062 | No Hit |

# Quality controls : Standard Workflow for NGS Analysis




A typical NGS workflow

# Cloud Computing: just virtual machines on the net

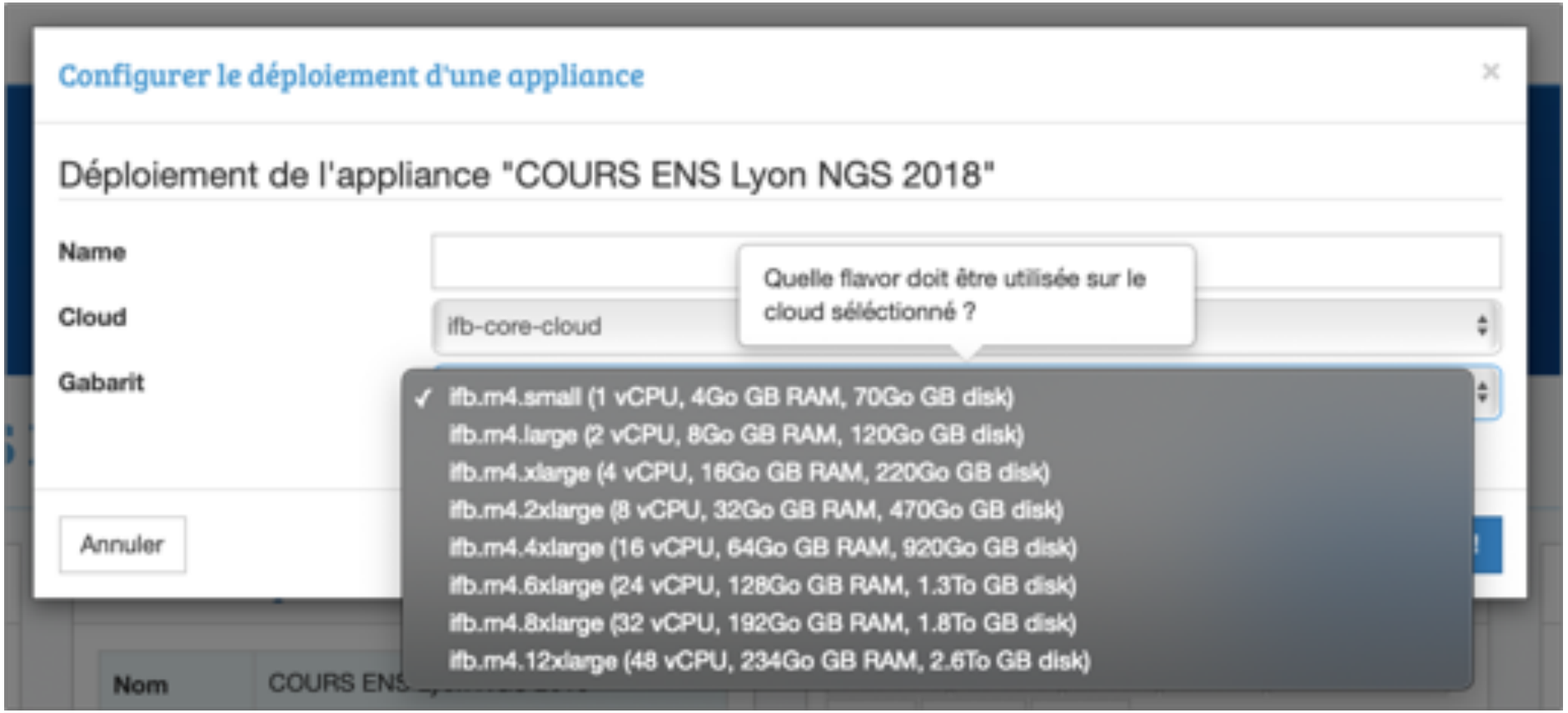


# Le cloud Biosphere de l'IFB

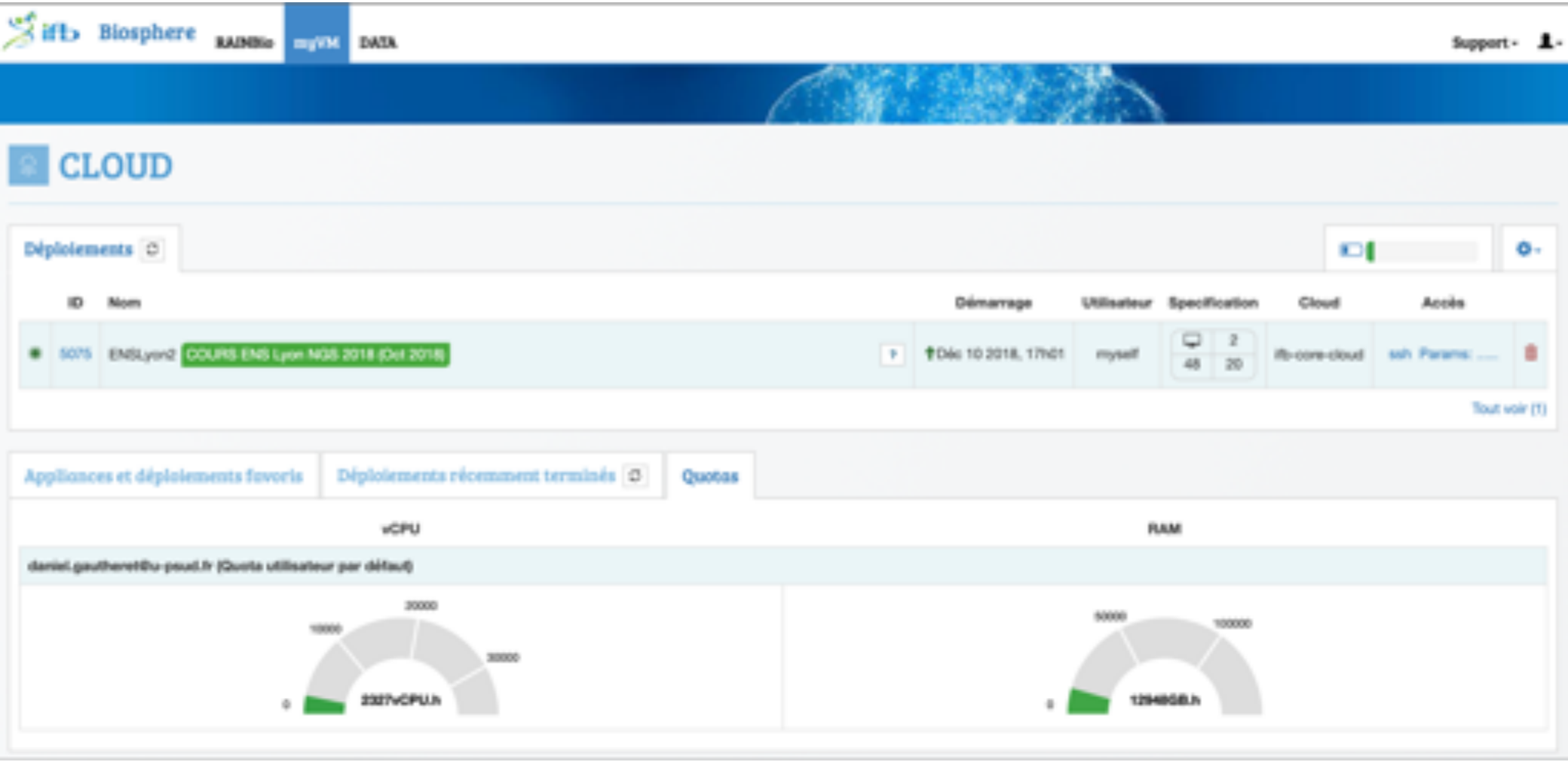
- Documentation pour déploiement d'une VM:
  https://ifb-elixirfr.github.io/biosphere/

# Les appliances (des images disque)…

# Des VM de différentes capacités...



# Tableau de bord de vos VM Biosphere



# TP RNA-seq

Créez une VM IFB de type: Ubuntu 16.04 (16.04)

Voir Documentation pour créer clé privée/publique:
https://ifb-elixirfr.github.io/biosphere/

Connectez-vous à votre VM par ssh

Récupérez les données dans :
http://rssf.i2bc.paris-saclay.fr/X-fer/AtelierNGS/ TPrnaseq.tar.gz
(par wget depuis la VM!)

Démarrage: installer fastqc avec conda et lancer fastqc sur toutes les banques.