Formation
« Bioinformatique pour le traitement de données deséquençage (NGS) »

Annabelle Haudry

Equipe Le Cocon

LBBE

cnrs

Université Claude Bernard Lyon 1

Récupération des données

1. données privées

lien fourni par le service de séquençage,

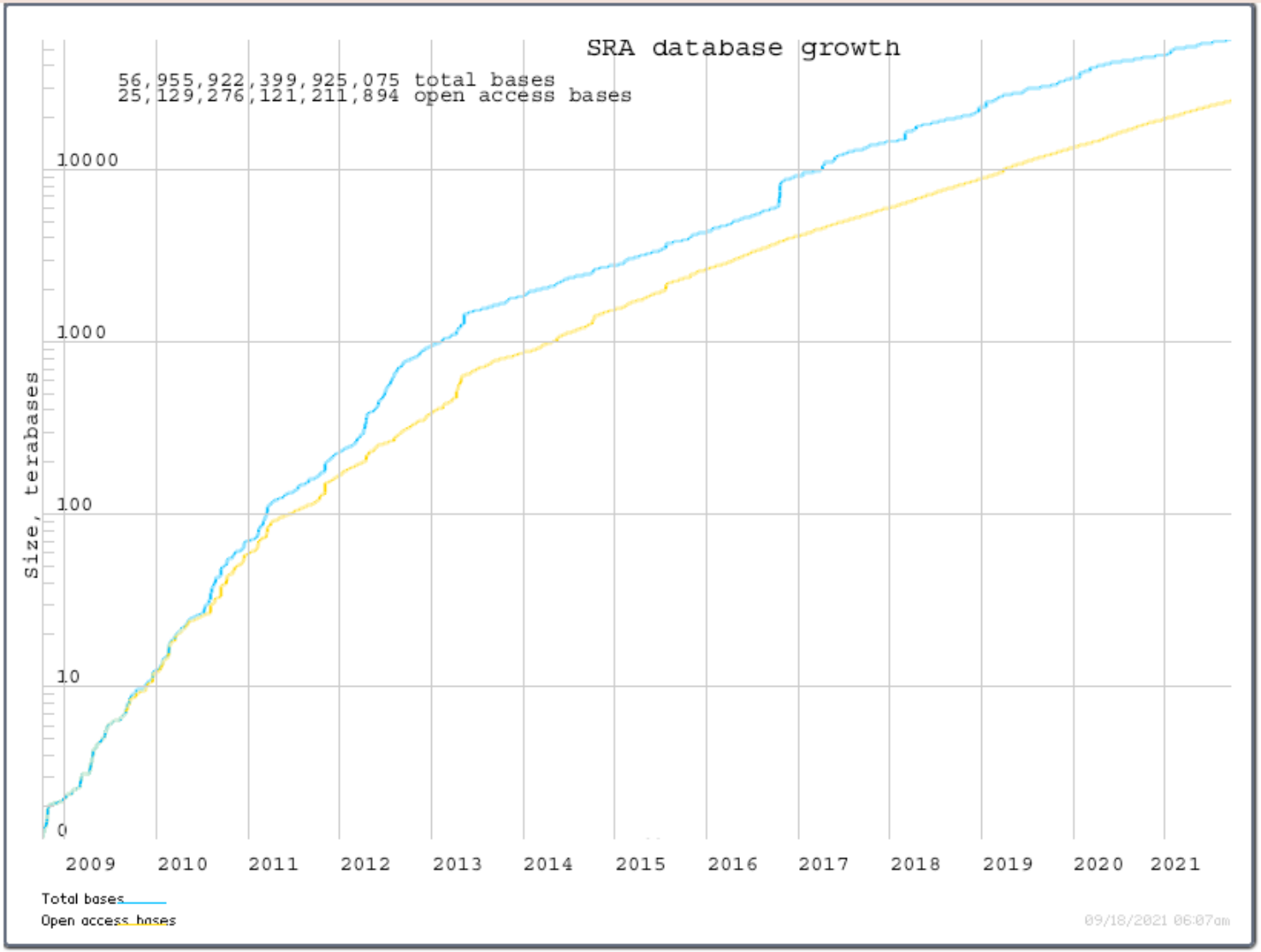
généralement des fichiers fastq ou fastq.gz

2. données publiques

bases de données SRA (Sequence Read Archive)

https://www.ncbi.nlm.nih.gov/sra/docs/

https://www.ncbi.nlm.nih.gov/sra/docs/sragrowth/


SRA database growth
56,955,922,399,925,075 total bases
25,129,276,121,211,894 open access bases
Size, terabases
10000
1000
100
10
0
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
Total bases
Open access bases
09/18/2021 06:07am

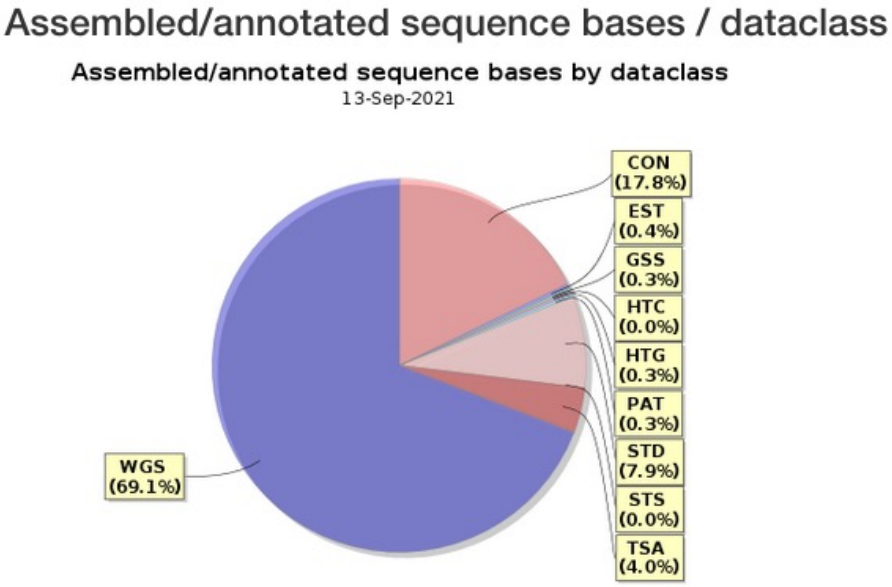
Récupération des données

2. données publiques

ENA (European Nucleotide Archive)

https://www.ebi.ac.uk/ena

https://www.ebi.ac.uk/ena/browser/about/content



Récupération des données

Exercices

1. Choisissez votre espèce favorite

Recherchez les données génomiques disponibles pour cette

espèce dans chacune des deux bases (SRA et ENA). Explorez et utilisez les modes de recherche ciblée (organisme, ADN/ARN, Illumina ou 454,etc.)

1. Récuperez les trois jeux de données suivants de séquencage de génome de *Staphylococcus aureus*

Récupération des données

Exercices

2. Récuperez 1 jeu de données suivants de séquencage de génome de *Staphylococcus aureus*

- *HiSeq4000 (_SRR7748059_)*

Format Fastq

Exercice

Visualisez un fichier fastq.

Attention !! Ce sont de très gros fichiers !

Format Fastq

Format qui permet de représenter une ou plusieurs séquences avec leurs scores de qualités par base. Une séquence est représentée par 4 lignes.

```
@HWI-QMN273:4:1:2:779#0/1
ANCAAAATCTGCATTACCTCCTCGGCTGGGACAACTTTATTC
+HWI-QMN273:4:1:2:779#0/1
\D[aaaab_aaaabba__a^Za^aaZa`]a__a_Z_aaaa`\
```

Format Fastq

La 1ère ligne commence par le symbole '@' suivi d'un identifiant de séquence.
La 2ème ligne correspond à la séquence.
La 3ème ligne commence par '+' suivi d'éventuelles autres infos.
La 4ème ligne correspond à la séquence qualité de la 2e ligne.

```
@HWI-QMN273:4:1:2:779#0/1
ANCAAAATCTGCATTACCTCCTCGGCTGGGACAACTTTATTC
+HWI-QMN273:4:1:2:779#0/1
\D[aaaab_aaaabba__a^Za^aaZa`]a__a_Z_aaaa`\
```

Format Fastq

Avantages :

Est un format simple et universel pour partager des séquences et le score de qualité

Beaucoup de logiciels reconnaissent et exploitent ce format

Ce format peut contenir plus ou moins d'informations (en les ajoutant sur la ligne commençant par '@').

Inconvénients :

Il n'y a pas un standard rigoureux et Illumina en a fait des variants

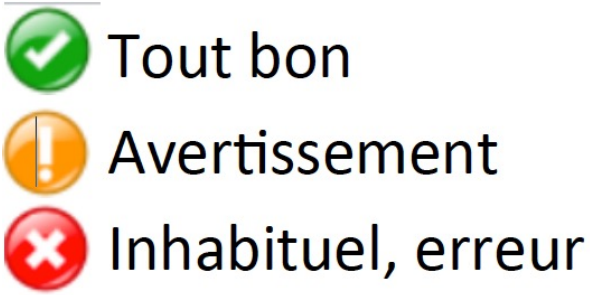
Le score de qualité n'est pas encodé de la même façon selon l'éditeur du fichier

Le fichier est volumineux.

Analyse de la qualité des reads

Interprétation des résultats FastQC

Plusieurs modules d'évaluation:



- Basic statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

Basic statistics

Measure	Value
Filename	SRR6322985_1.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	26039328
Sequences flagged as poor quality	0
Sequence length	100
%GC	38

Per base sequence quality

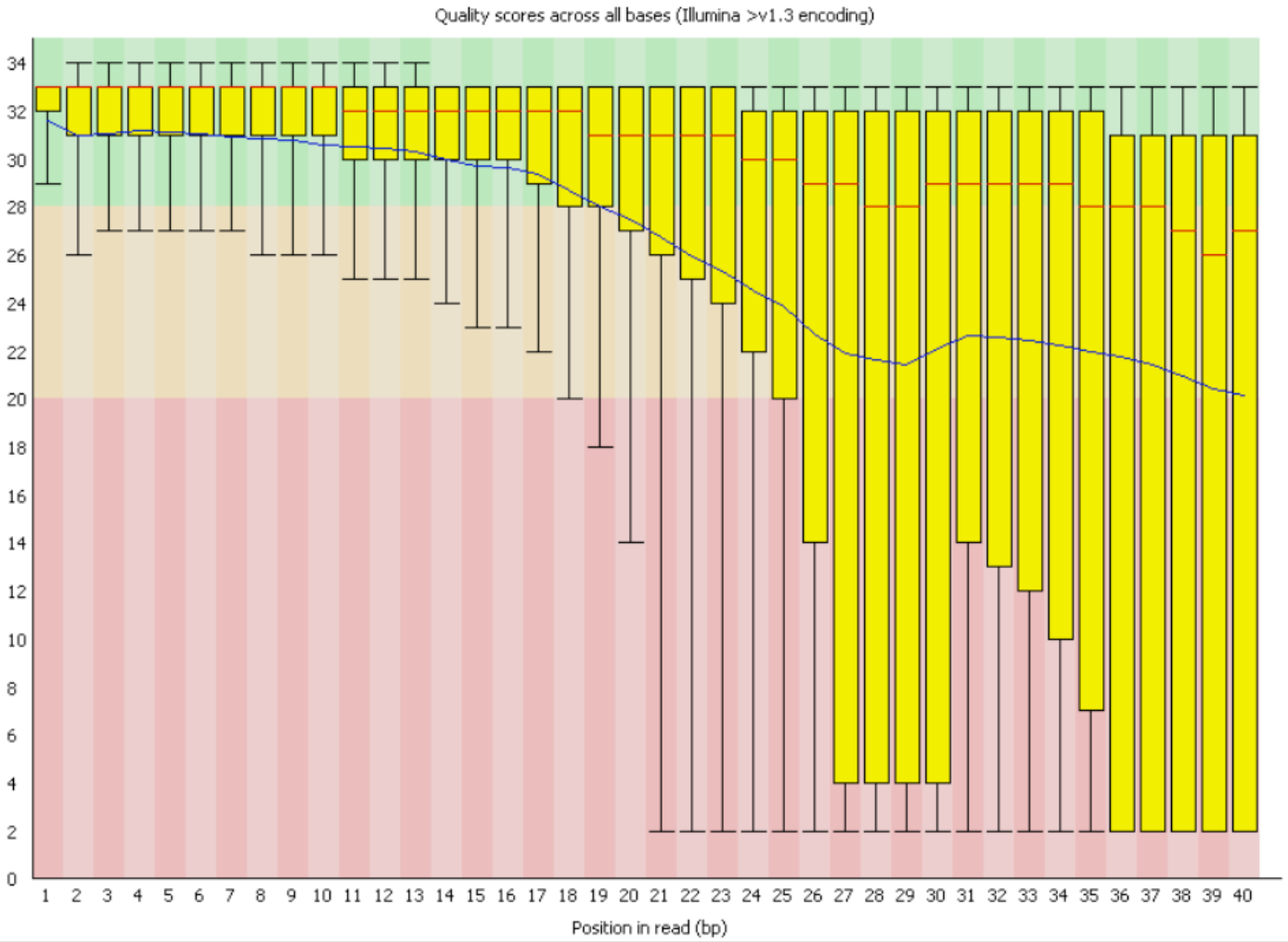
Distribution of quality values for each position in read

moyenne

médiane

25-75 percentiles

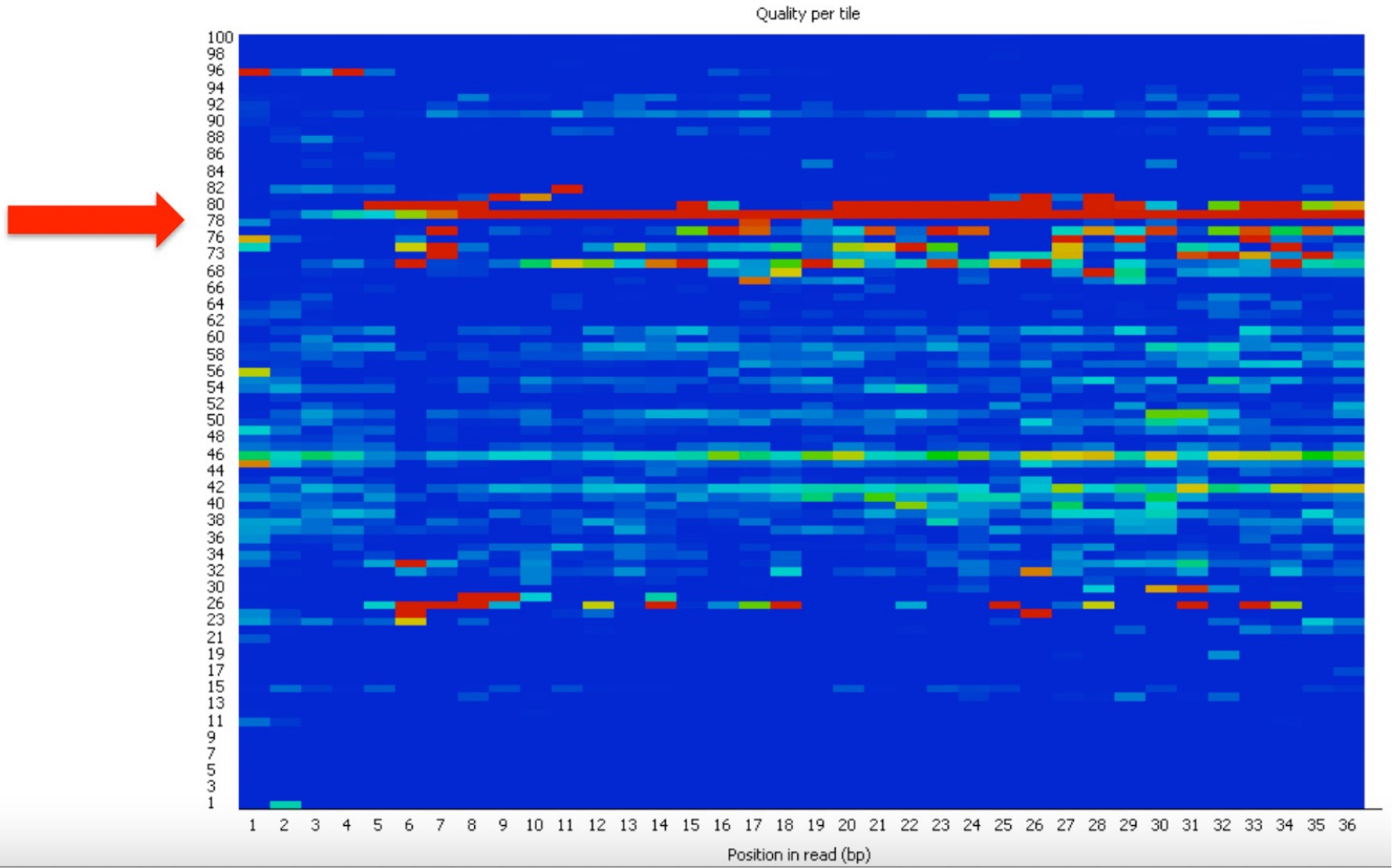
10-90 percentiles



Per tile sequence quality

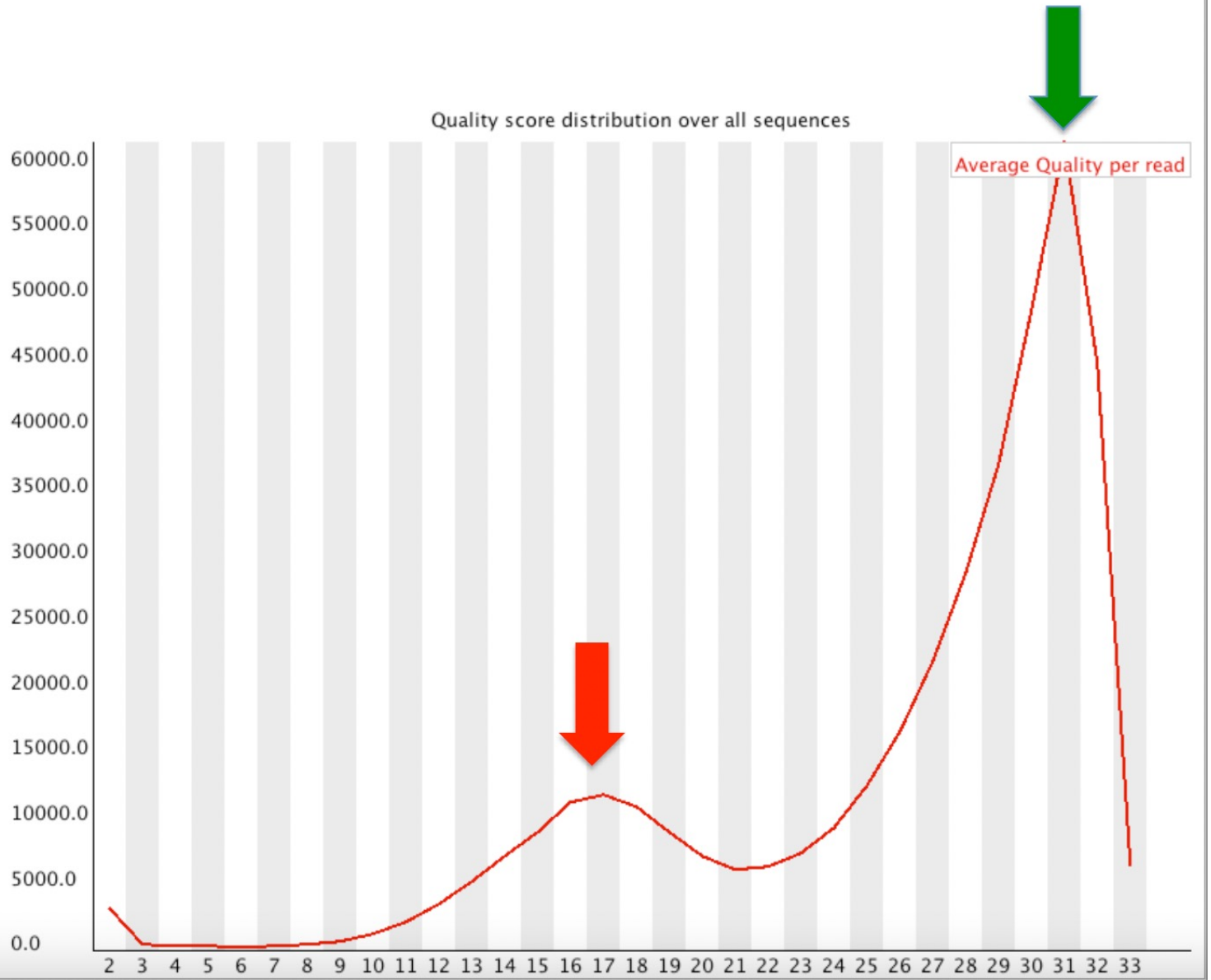
Seulement avec des données Illumina

Permet de vérifier que la qualité des reads est homogène dans la cellule



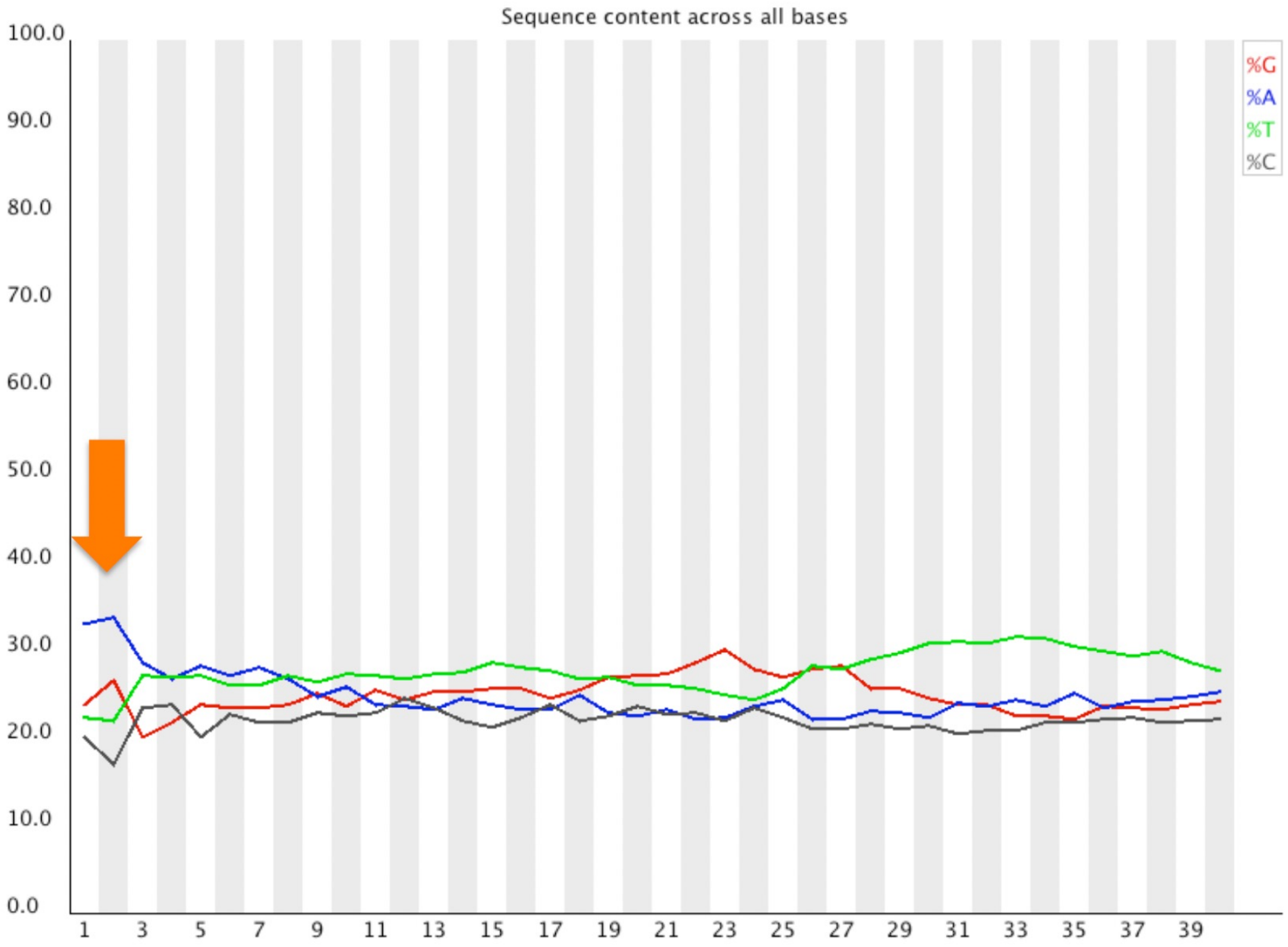
Per sequence quality scores

Distribution des valeurs de qualité moyennes par read



Per base sequence content

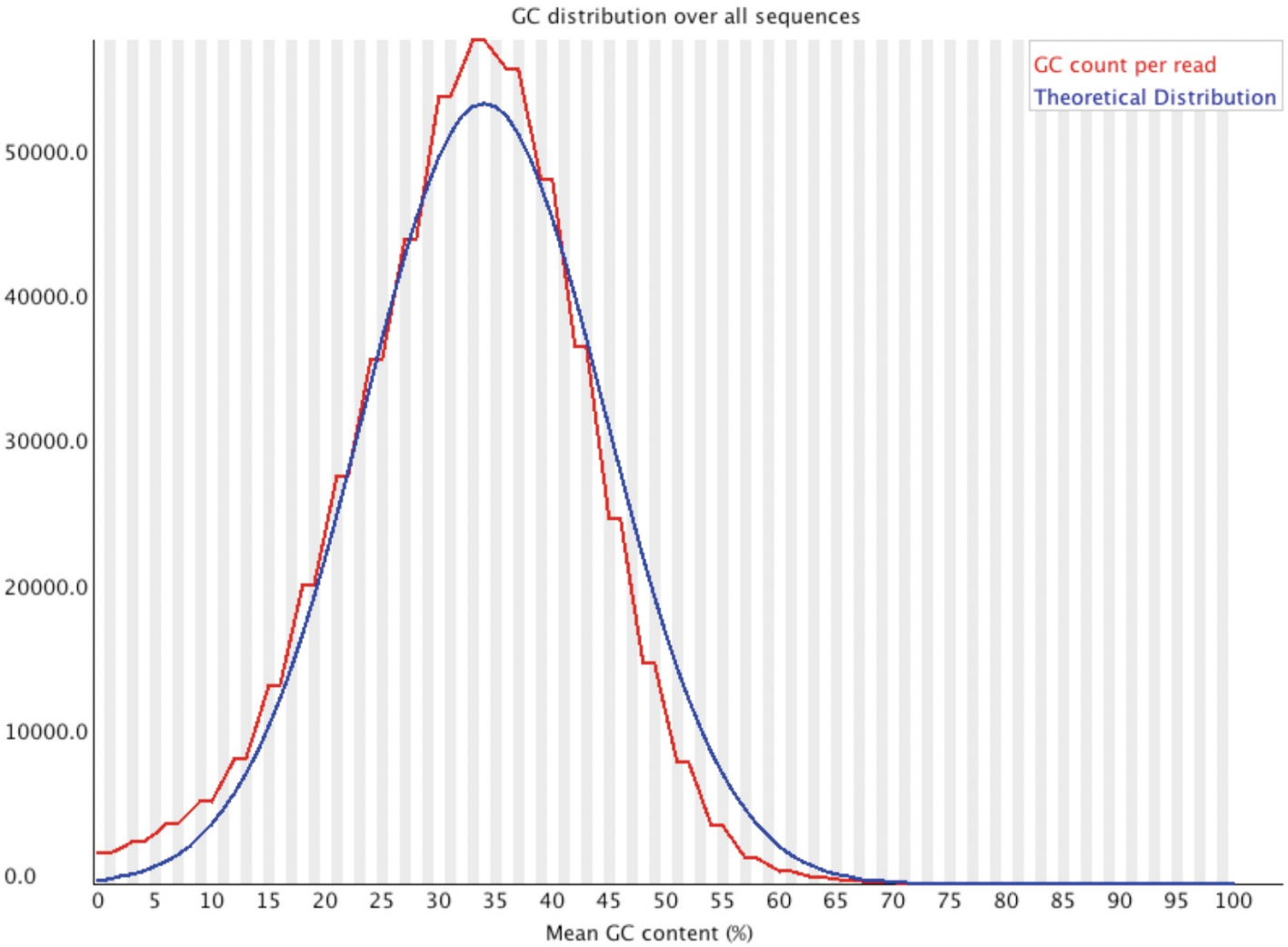
La teneur en base ne devrait pas varier en fonction de la position sur le read, sauf s'il y a un biais dans la librairie



Per sequence GC content

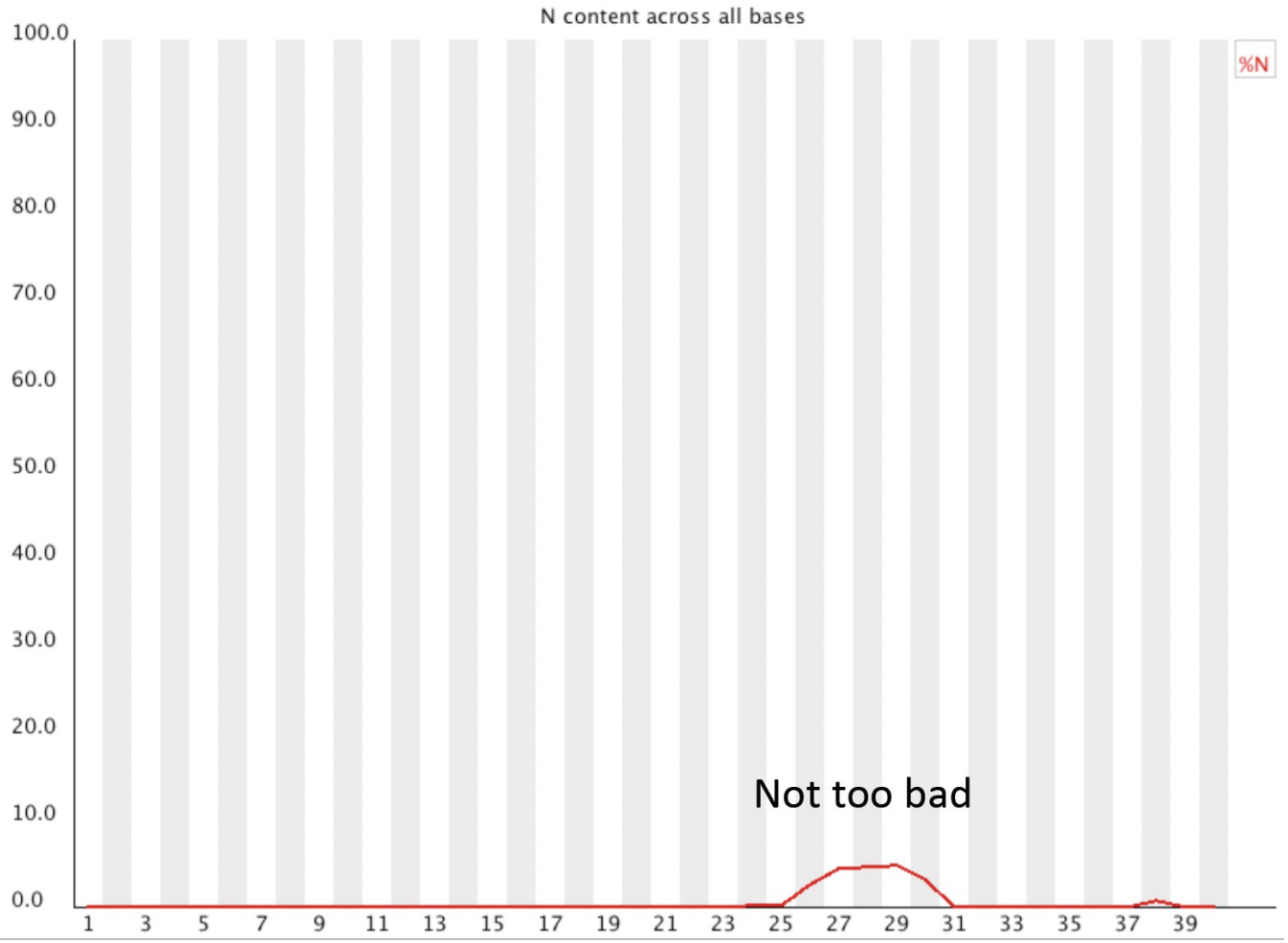
Attendu: distribution normale autour de la valeur moyenne

Distribution observée dans le run

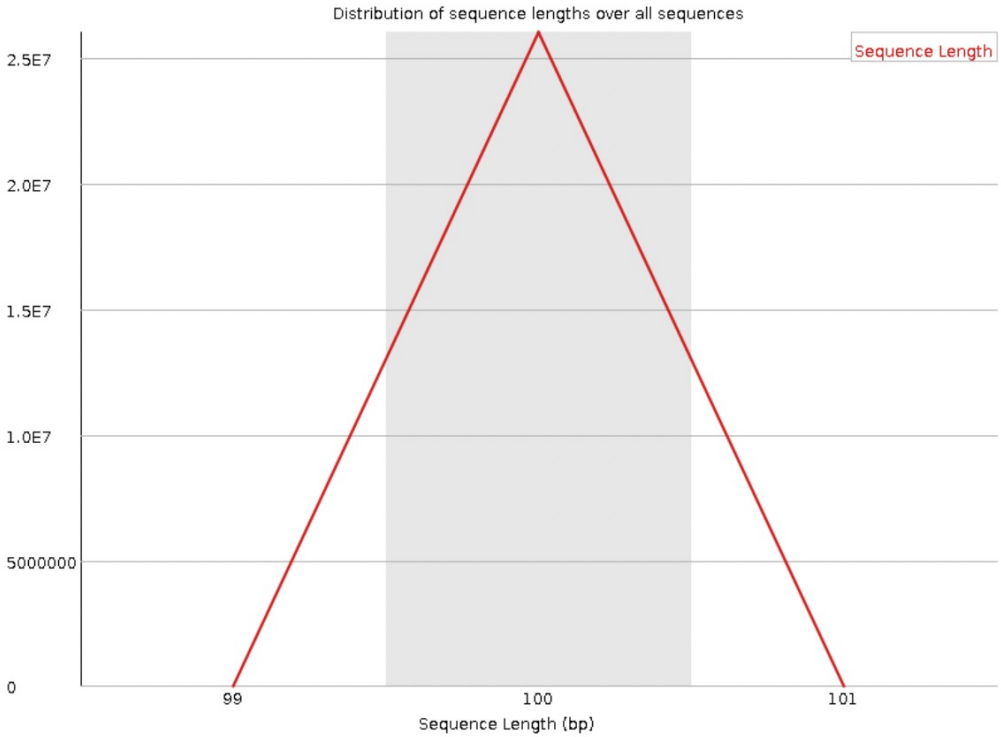


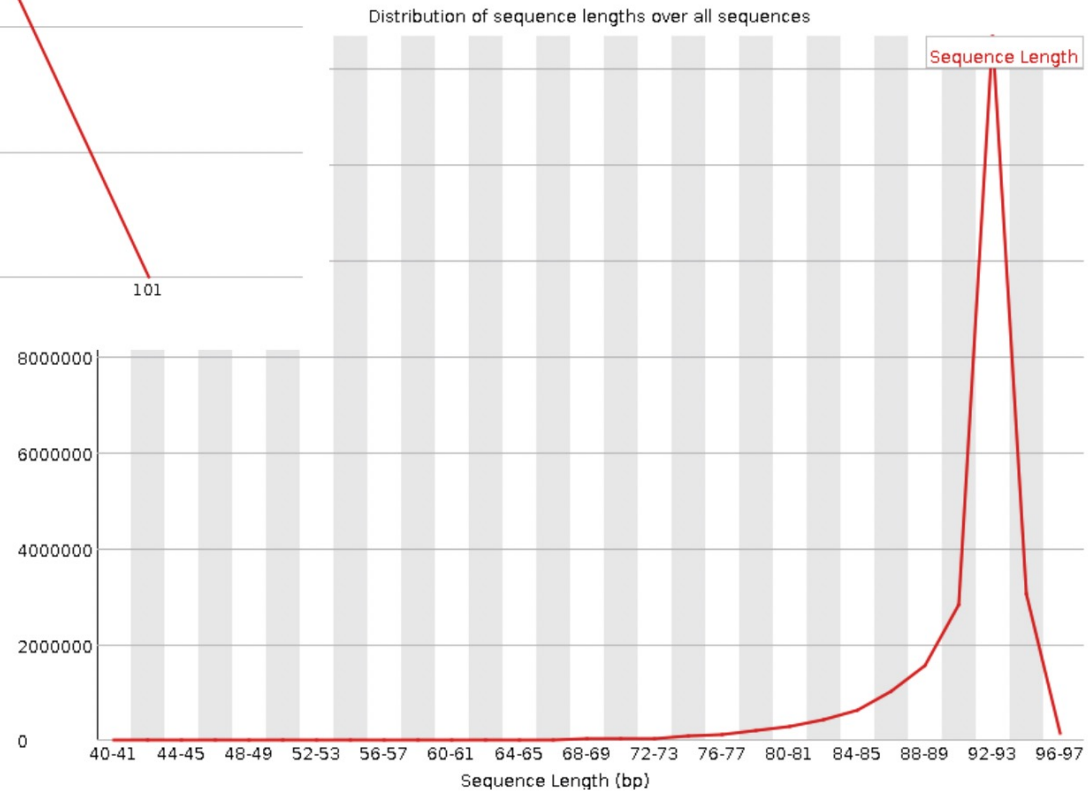
Per base N content

Présence de bases indéterminées?



Sequence Length Distribution



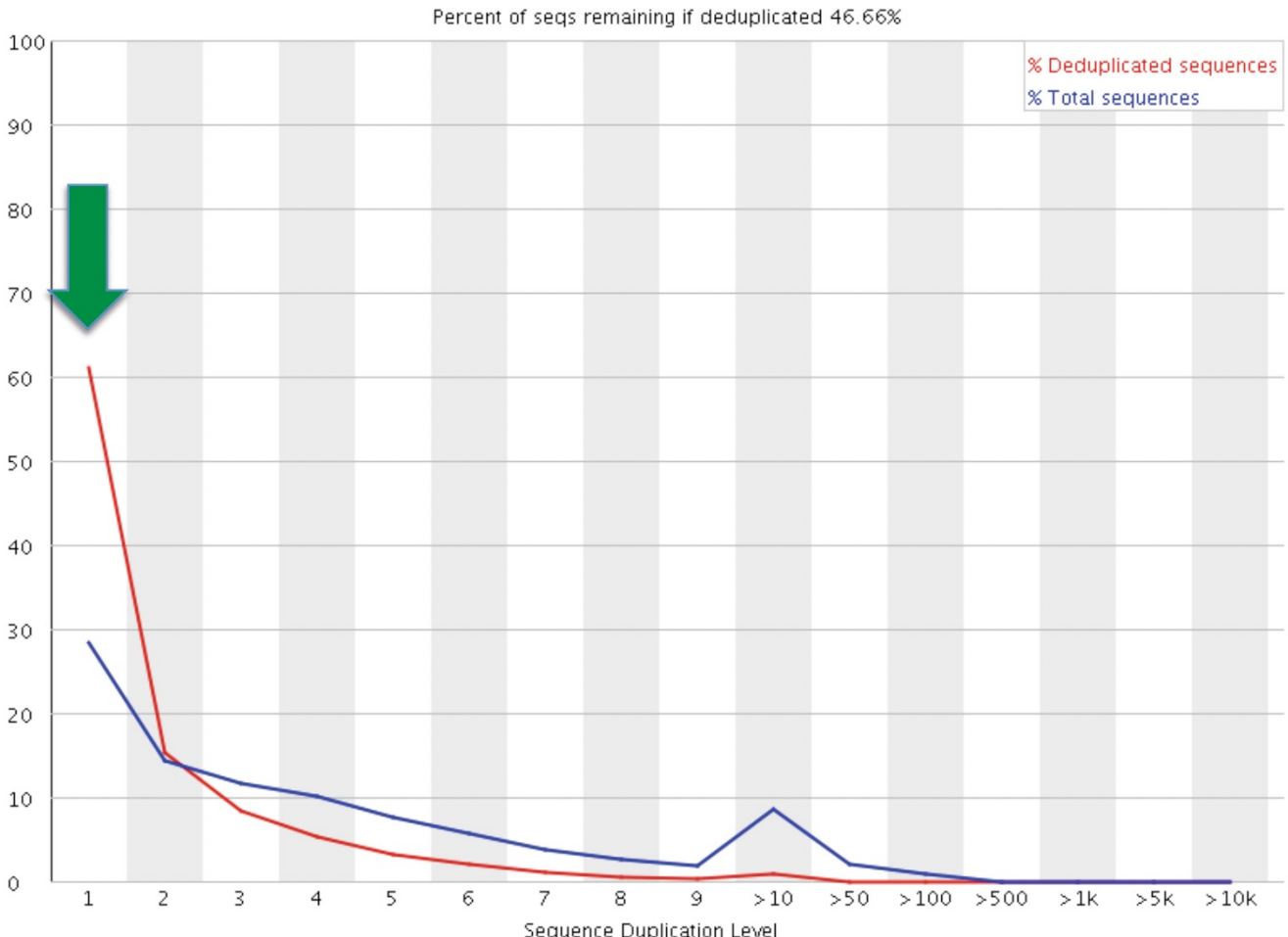


Sequence Duplication Levels

La plupart des séquences devraient être uniques dans la librairie finale. Un faible taux de duplication peut indiquer une forte couverture de séquençage. Un fort taux de duplication suggère plutot un biais d'enrichissement (PCR).

non-unique sequences
> 20% of the total

non-unique sequences
> 50% of the total



Overrepresented sequences & Adapter Content

Liste des séquences sur représentées, leur % dans la librairie.

Et recherche des séquences d'adaptateurs et primers connus.

Retirer les adaptateurs

Cutadapt

cutadapt -a AACCGGTT -o output.fastq input.fastq
https://cutadapt.readthedocs.io/en/stable/guide.html

Sinon, comment identifier une séquence inconnue?

Overrepresented sequences & Adapter Content

Liste des séquences sur représentées, leur % dans la librairie.

Et recherche des séquences d'adaptateurs et primers connus.

Retirer les adaptateurs

Cutadapt

cutadapt -a AACCGGTT -o output.fastq input.fastq
https://cutadapt.readthedocs.io/en/stable/guide.html

Sinon, comment identifier une séquence inconnue?

=> recherche blast dans « nr »

Interprétation des résultats FastQC

FastQC

Good sequence

https://www.bioinformatics.babraham.ac.uk/projects/fastqc/good_sequence_short_fastqc.html

Bad sequence

https://www.bioinformatics.babraham.ac.uk/projects/fastqc/bad_sequence_fastqc.html

Filtre des reads pour la qualité

Trimmomatic

http://www.usadellab.org/cms/?page=trimmomatic

FastP

https://github.com/OpenGene/fastp

NB : Refaire un FastQC après le filtrage des reads