Bioinformatics project 1 semester 2

A. Libraries installation

Opening a command prompt

```
# Windows:
# Click on the start button and type `cmd` in the search bar. Press Enter.
# macOS:
# Press Command (⌘) + space to open Spotlight Search. Type "Terminal" and press Enter.
```

Biopython & Numpy & Matplotlib

```
# Windows and macOS: in your command prompt type:
python3 -m ensurepip --upgrade
python3 -m pip install --upgrade pip
python3 -m pip install Bio Numpy Matplotlib
# Linux: in your command prompt type:
sudo apt update
sudo apt install python3-pip
pip install Bio Numpy Matplotlib
```

B. Dowloading the files

1. What sequencing platform was used to do the sequencing run?

   Illumina

2. What is the instrument model used?

   Illumina HiSeq 2000

3. What is the library type used for the sequencing? Paired-end? Single-end or mate-pair?

   Paired-end (Library layout)

4. What is the type of molecules that have been sequenced?

   Library source genomic --> DNA

5. What is the extension of these files? What does it indicate?

   The files are different because they come from the ends of the same fragments

6. Are these files expected to be different in their contents? Justify your answer.

   The files are different because they come from the ends of the same fragments

7. Select both of the files and click on **Get Download script**.✅

8. What are the sizes of both files?

   While clicking on the files we can clearly see that they are 388Mb each. In the other hands, the sub's one are less bigger: 22Mb.

C. Analyzing the files

Importing the installed libraries in Python

Open a Python terminal. Do not forget to type all your commands in a text file with a .py extension. Here are the libraries that need to be imported at the beginning of your python script:

```
from Bio import SeqIO
import gzip
import matplotlib.pyplot as plt
import numpy as np
```

Reading fastq files

Make sure that the Python session's working directory aligns with the location of this fatsq files. For that you can use the Python library os as follows:

```
import os
os.getcwd()
```

In case your fastq files are not located in the Python current working directory, you have two options: either copy them into the Python current directory, or modify the Python current directory to that of the fastq files using the following code:

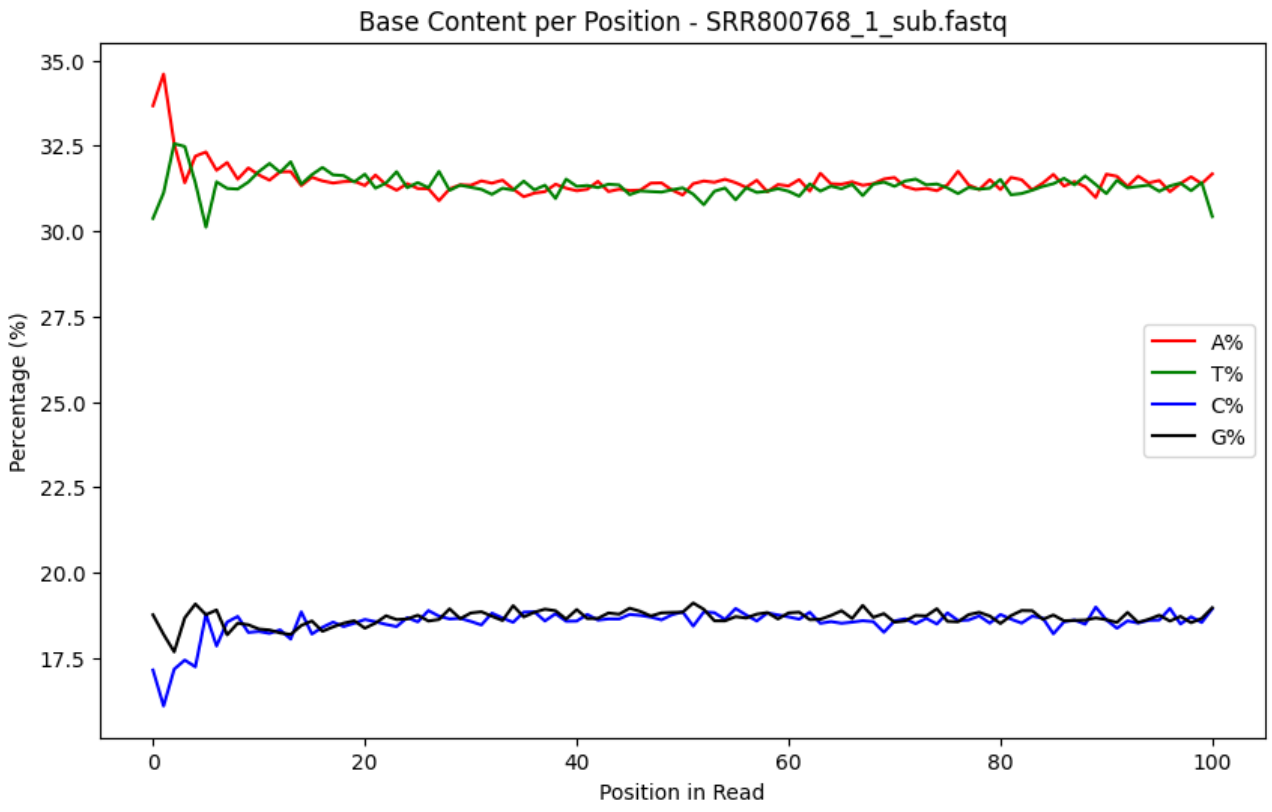
```
new_directory = "/path/to/your/new/directory" # better if absolute path
os.chdir(new_directory)
```

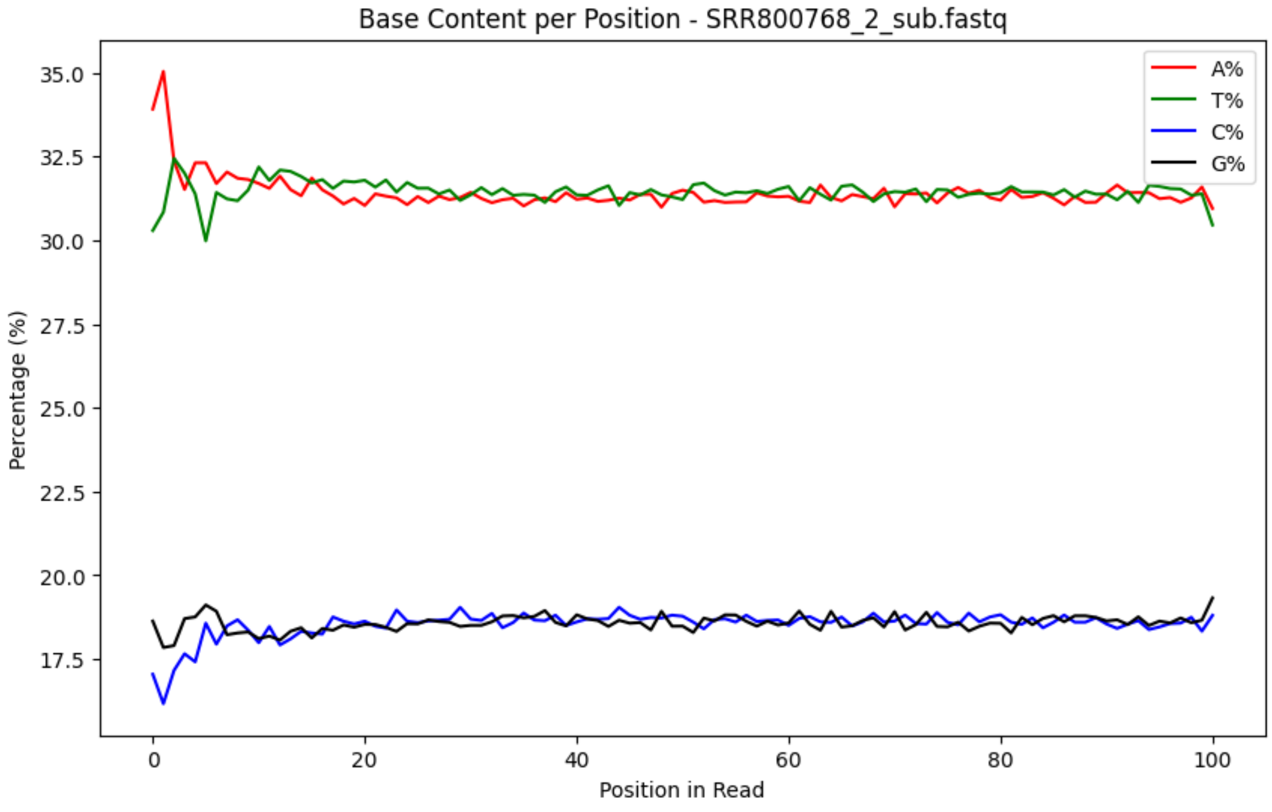
Then, you will have to read both files using the following syntax.

```
# For gzipped files
records = SeqIO.parse(gzip.open('filename.fastq.gz', 'rt', encoding='utf-8'), 'fastq')
# For non-gzipped files
records = SeqIO.parse('filename.fastq', 'fastq')

# You can iterate over my file as follow:
for record in records:
  seq = record.seq
  phred_scores = record.letter_annotations[‘phred_quality']
```

1. How many reads do we have in each of the two files? Verify with a Python script that they are correctly paired.
   100 000
2. What is the read length?
   101
3. Plot the reads per base content for each of the fastq files: the x axis is the position on the read (`0` to `read length - 1`) and on the y axis the percentages of each base (A in red, T in green, C in blue and G in black).


Base Content per Position - SRR800768_1_sub.fastq
Percentage (%)
35.0
32.5
30.0
27.5
25.0
22.5
20.0
17.5
0
20
40
60
80
100
Position in Read
A%
T%
C%
G%


Base Content per Position - SRR800768_2_sub.fastq
Percentage (%)
35.0
32.5
30.0
27.5
25.0
22.5
20.0
17.5
0
20
40
60
80
100
Position in Read
A%
T%
C%
G%

4. What do you notice regarding the relative quantity of the different bases? Is it an expected result? Justify your answer.

   In the first few positions (0-10) of the sequencing reads, there is a noticeable spike in the percentages of A and T bases, which gradually stabilize as sequencing progresses. Beyond these initial positions, the base composition tends to reach a more balanced state, with A, T, C, and G occurring within a relatively stable range, though A and T slightly dominate over G and C. Additionally, some fluctuations across positions indicate sequencing quality variations or biases. This pattern is both expected and partially unexpected, depending on the context. In Illumina sequencing, biases at the start of reads are common due to factors such as adapter or primer sequences influencing base content, sequencing chemistry artifacts introducing early errors, or the machine's calibration during the initial cycles. However, if the sample were a randomly distributed genomic DNA, a more uniform distribution of A, T, C, and G across all positions would be expected. Any consistent bias toward A/T over G/C might indicate either a sequencing artifact or an intrinsic genomic feature. Since this data originates from *Saccharomyces cerevisiae* (yeast), the observed AT dominance may be reflective of its genomic characteristics. Yeast genomes often contain regions with higher AT content, particularly in non-coding regions, though potential sequencing biases should not be overlooked.

5. The GC-content (or guanine-cytosine content) is the percentage of nucleobases in a DNA or RNA molecule that are either guanine (G) or cytosine (C). It is computed as follows:

$$\%GC = \frac{G+C}{A+T+G+C} \times 100$$

   Estimate the GC content of the *Saccharomyces cerevisiae* sequenced genome and compare it with the known GC content of the species.

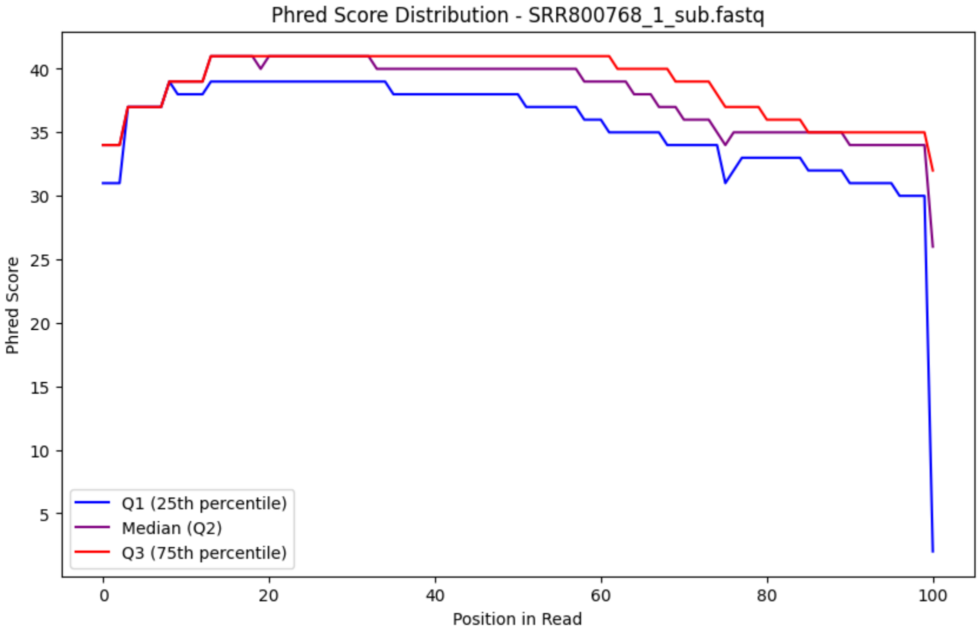
   GC Content of SRR800768_1_sub.fastq: 37.20%
   GC Content of SRR800768_2_sub.fastq: 37.12%
   Known GC content of Saccharomyces cerevisiae: ~38-40%
   Difference for SRR800768_1_sub.fastq: 1.80%
   Difference for SRR800768_2_sub.fastq: 1.88%

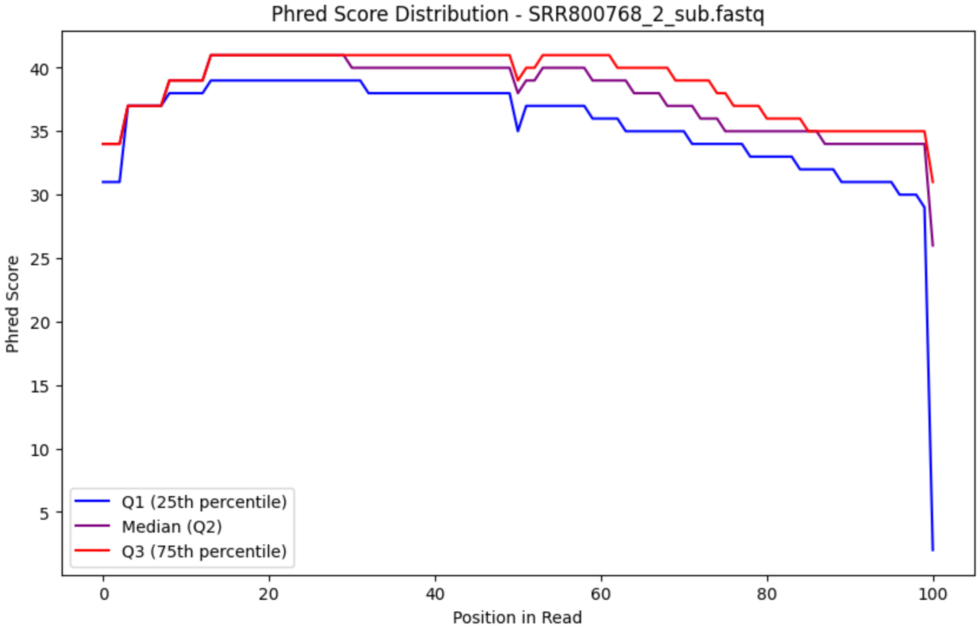
   The GC content results obtained for both files appear to be consistent with the known GC composition of *Saccharomyces cerevisiae*. This correlation suggests that the sequencing data accurately reflects the expected genomic characteristics of yeast. *Saccharomyces cerevisiae* typically exhibits a moderate GC content, which varies across coding and non-coding regions, with coding sequences generally having a higher GC percentage. The observed GC content in the sequencing data reinforces the reliability of the dataset and aligns with previously documented genomic studies of yeast.

6. For each file, generate a plot illustrating the median, first quartile (Q1) and third quartile (Q3) of the Phred scores for each position in the read. The x-axis represents the position on the read, and the y-axis depicts the median (in purple), Q1 (in blue), and Q3 (in red) of the Phred scores.

   Tip: you can use `numpy` to compute the Q1, Q2 and Q3 of a list:

   ```
   my_list = [1, 2, 3, 3, 3, 5, 8]
   # Q2 = median
   my_median = np.median(my_list)
   # Q1
   Q1 = np.percentile(my_list, 25)
   # Q3
   Q3 = np.percentile(my_list, 75)
   ```





7. Is the quality of the bases homogeneous across the read? Propose a hypothesis explaining that.
   The Phred score distribution plots indicate that base quality is not uniform across the sequencing reads. Initially, the Q1, median (Q2), and Q3 values are relatively high (~35-40), reflecting high confidence in base calls for the first 10-20 positions. However, a gradual decline in quality is observed beyond positions 60-80, with an increasing spread between Q1 and Q3, indicating greater variability in sequencing accuracy. A sharp drop around position 100 highlights that 3' end bases are more prone to errors and less reliable. This decline in sequencing quality is a well-known phenomenon in Next-Generation Sequencing (NGS) and can be attributed to multiple factors. Signal decay and dye degradation progressively weaken fluorescence signals, reducing base-calling accuracy. Phasing and pre-phasing effects, where some DNA strands lose synchronization during sequencing cycles, further contribute to errors at later positions. Additionally, accumulated base-calling errors and PCR amplification bias can amplify sequencing noise, further reducing confidence in bases at later positions. To mitigate these issues, quality control measures such as trimming low-quality bases (removing bases with Phred scores <20-25), filtering out low-quality reads (removing reads with an average Phred score <30), and applying error correction tools like **SPAdes** or **BFC** can be implemented. These steps enhance data quality and improve downstream analyses by retaining only high-confidence sequences. If necessary, an automated algorithm can be designed to trim low-quality bases and filter poor-quality reads, ensuring a more reliable dataset for analysis.

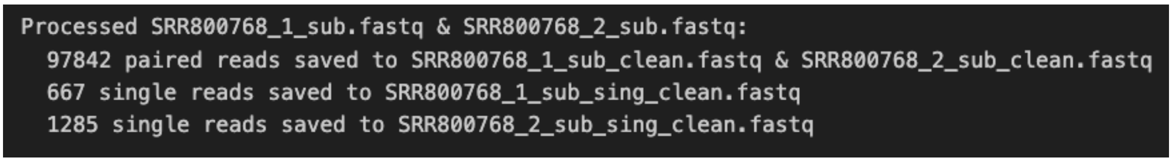
8. Suggest an algorithm for cleaning reads and implement it in Python. This algorithm should allow trimming low-quality bases at the ends of the reads and then keep reads whose length is above a threshold. As pairing may be disturbed by cleaning, 4 output files are expected:

- two paired files: `SRR800768_1_sub_clean.fastq` and `SRR800768_2_sub_clean.fastq`.
- two files with singletons: `SRR800768_1_sub_sing_clean.fastq` and `SRR800768_2_sub_sing_clean.fastq`.

Tip: Here is a code snippet to create and fill a fastq file:

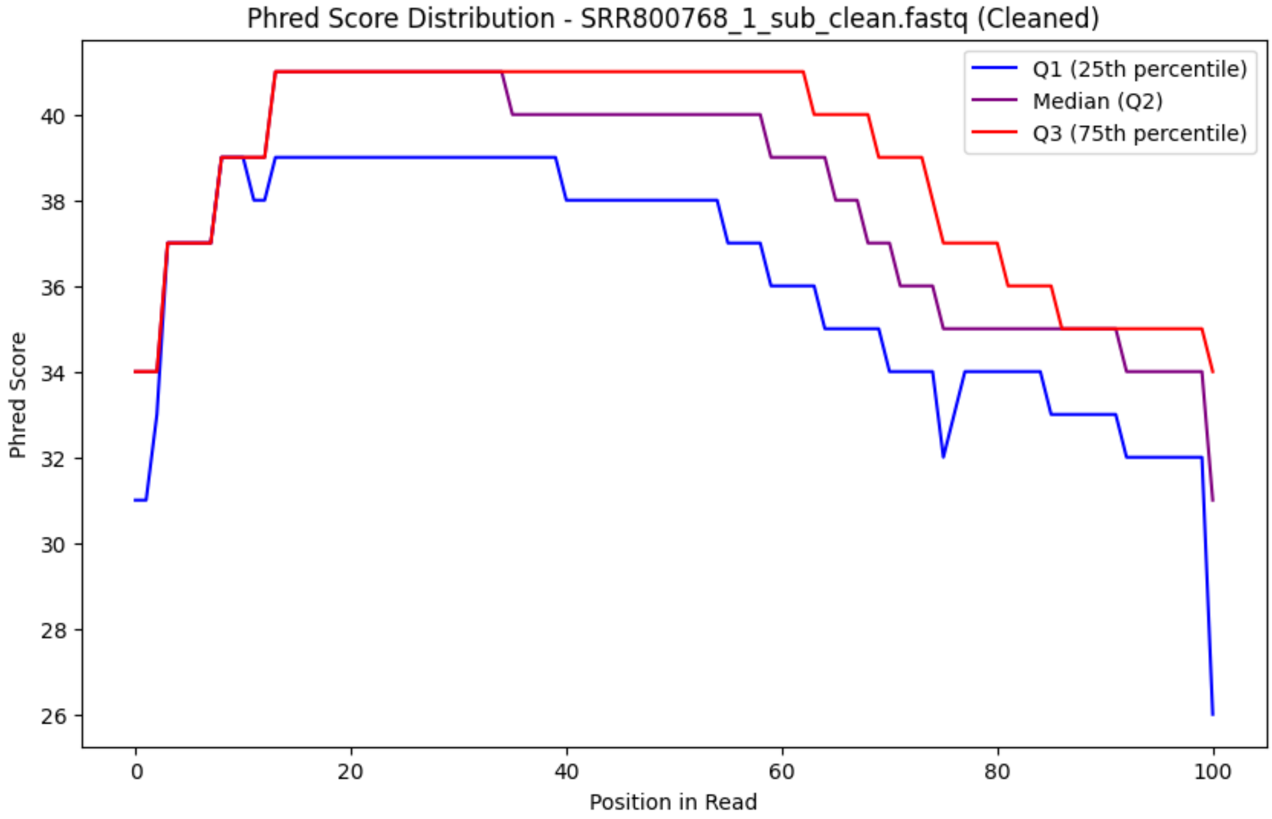
```
with open(path/to/my_new_file, 'w') as f:
  for record in records :
    SeqIO.write(record, f, ‘fastq’)
```

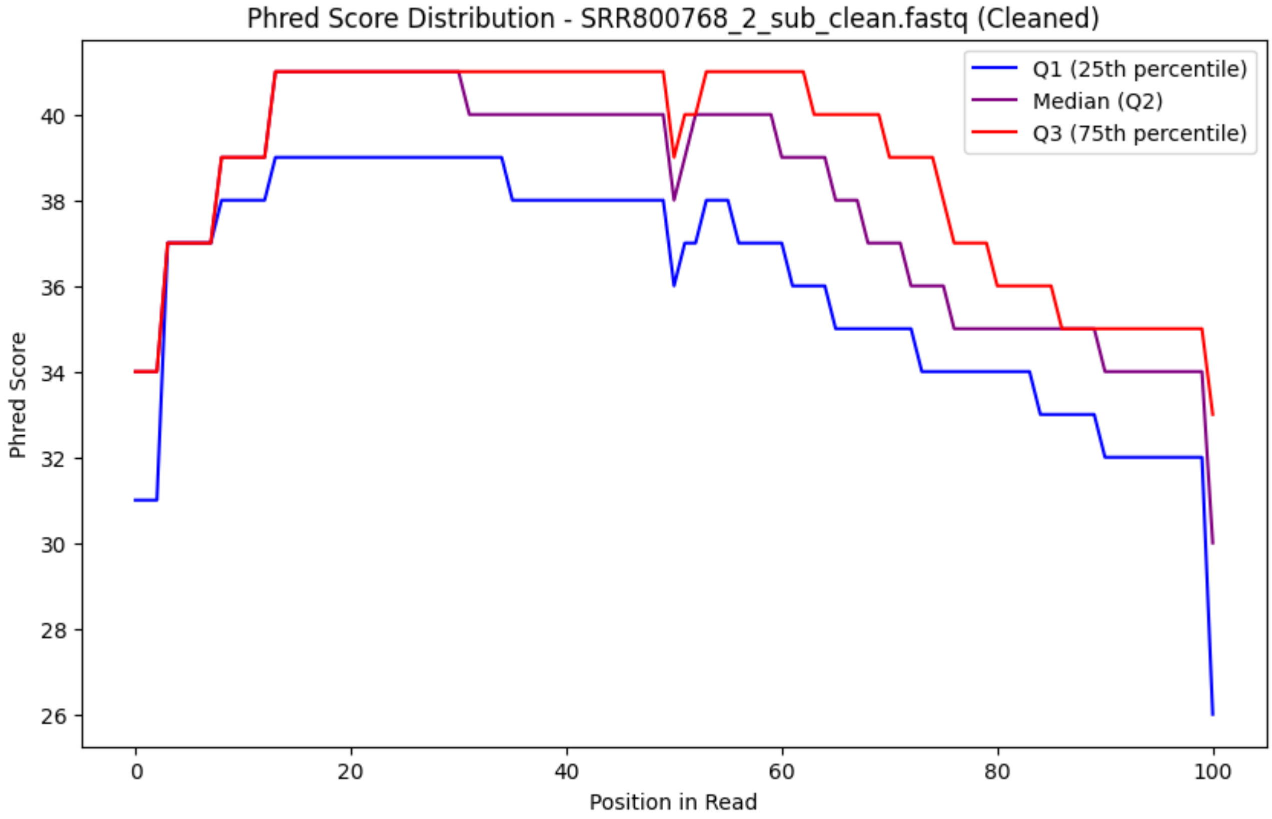
Look code.
Output :



```
Processed SRR800768_1_sub.fastq & SRR800768_2_sub.fastq:
  97842 paired reads saved to SRR800768_1_sub_clean.fastq & SRR800768_2_sub_clean.fastq
  667 single reads saved to SRR800768_1_sub_sing_clean.fastq
  1285 single reads saved to SRR800768_2_sub_sing_clean.fastq
```

9. To check the validity of the cleaning algorithm, plot the qualities of the clean paired files as done in part `C.6`.





Summary

1. Illumina Sequencing Biases

- **Kozerewa et al., 2009**: *Illumina library preparation artifacts lead to AT bias in initial cycles.*

  Link
- **Dohm et al., 2008**: *Systematic biases in Illumina sequencing, including early sequence errors.*

  Link

2. Base Composition Bias in High-Throughput Sequencing

- **Chen et al., 2013**: *Effects of sequence context and base composition on sequencing errors.*

  Link

3. Genomic AT Bias in Saccharomyces cerevisiae

- **Liti et al., 2009**: *Yeast genome evolution and variation in AT-rich non-coding regions.*

  Link

4. GC Content in Saccharomyces cerevisiae

- **Liti et al., 2009**: *Variation in GC content across yeast genomes and its evolutionary significance.*

  Link
- **Sharp et al., 1986**: *GC composition bias in yeast coding sequences.*

  Link

5. Genome-Wide Analysis of Yeast GC Content

- **Goffeau et al., 1996**: *Complete sequence of Saccharomyces cerevisiae genome, including GC distribution.*

  Link

6. Phred Score and Base Quality in NGS

- **Ewing & Green, 1998**: *Phred: base-calling accuracy and sequencing quality scores.*

  Link
- **Cock et al., 2010**: *FASTQ format and quality control in sequencing.*

  Link

7. Common NGS Sequencing Errors and Solutions

- **Schirmer et al., 2016**: *Phasing, signal decay, and sequencing error biases in Illumina data.*

  Link
- **Nikolenko et al., 2013**: *BFC: Bayesian error correction for Illumina sequencing.*

  Link