Assignment 1 – Margarida Pinheiro (201805012)

1.

The sequence reads retrieved from the SRA Toolkit using `fastq-dump` will be downloaded as FASTQ files.

The "`--split-files`" option splits the FASTQ reads into two files in order to obtain pair-ended reads in separate files.

The "`--split-3`" separates the reads into left and right ends. After this command, single and paired-end data will produce one or two FASTQ files, respectively. For paired-end data, the file name will be suffixed '*_1.fastq', and the second will be suffixed '*_2.fastq', otherwise, a single file with extension '*.fastq' will be produced.

2.

a) The quality scores are encoded into a compact form in FASTQ files. In this encoding, the character with an ASCII code is equal to the value of the quality score + 33.

b)

Symbol	ASCII Code	Quality score
D	68	35
E	69	36
1	49	16
?	63	30
<	60	27

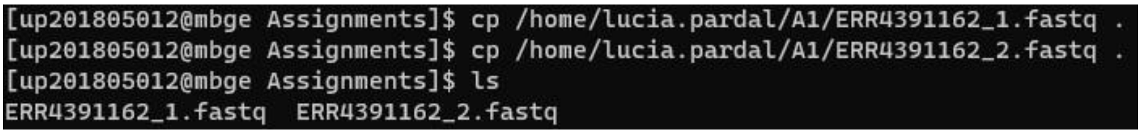
Sequence	G	G	G	C	A	A	T
Quality Score	35	35	30	35	36	16	27
Total	214						

The mean quality of the read is calculated by dividing the sum of the quality scores (214) by the number of characters in the sequence (7): 214/7 = 30.57

3.

Copy files **ERR4391162_1.fastq** and **ERR4391162_2.fastq** stored at /home/lucia.pardal/A1 using `cp`.

`ls` display the list of files in the current directory after copying the files.



```
[up201805012@mbge Assignments]$ cp /home/lucia.pardal/A1/ERR4391162_1.fastq .
[up201805012@mbge Assignments]$ cp /home/lucia.pardal/A1/ERR4391162_2.fastq .
[up201805012@mbge Assignments]$ ls
ERR4391162_1.fastq  ERR4391162_2.fastq
```

Through the search of this sample in the Sequence Read Archive (SRA) the following information was obtained:

Regarding the run this sample has 4,079 spots, 2M bases, a GC Content of 39.9% and a size of 1.4Mb.

The library name is CRAY-SWB2_P20-EZNA-2-trnL-HSBAS:trnL, the instrument used for the sequencing was the Illumina HiSeq 2500, the strategy was AMPLICON, the source METAGENOMIC, the Selection: PCR, the Layout: PAIRED and the Construction protocol: Primer_F=GGGCAATCCTGAGCCAA;Primer_R=CCATTGAGTCTCTGCACCTATC.

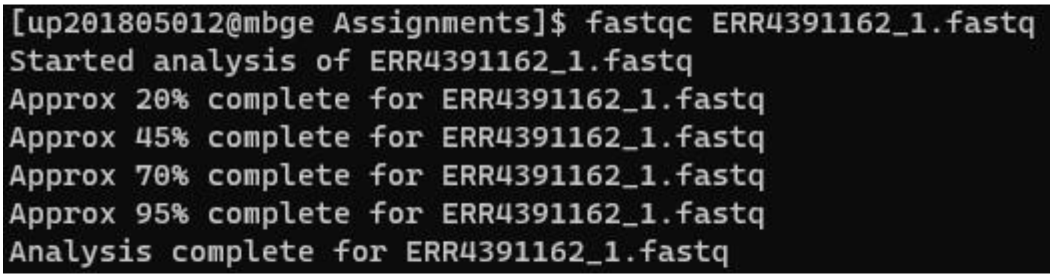
The organism is a freshwater metagenome and the sample was collected by Guilherme Buzzo in Portugal.

The study conducted on this sample is referred to as CRAYFISH:DNA-based diet analysis of invasive crayfish in Portugal and the sample was submitted by the CIBIO-INBIO RESEARCH CENTER IN BIODIVERSITY AND GENETIC RESOURCES.

4.

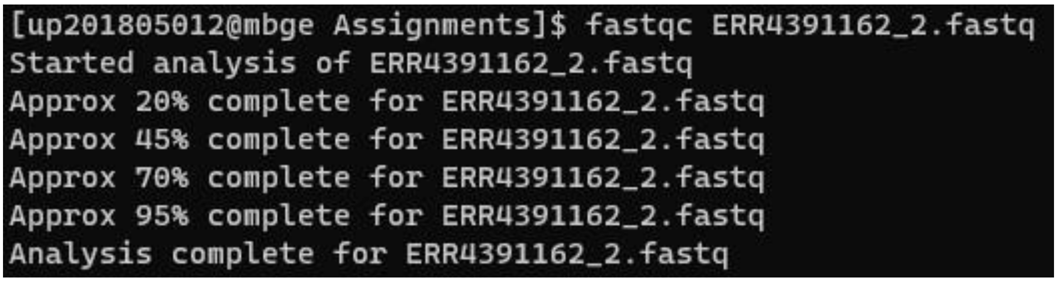
a)

Quality of ERR4391162_1.fastq using **`fastqc`**



```
[up201805012@mbge Assignments]$ fastqc ERR4391162_1.fastq
Started analysis of ERR4391162_1.fastq
Approx 20% complete for ERR4391162_1.fastq
Approx 45% complete for ERR4391162_1.fastq
Approx 70% complete for ERR4391162_1.fastq
Approx 95% complete for ERR4391162_1.fastq
Analysis complete for ERR4391162_1.fastq
```

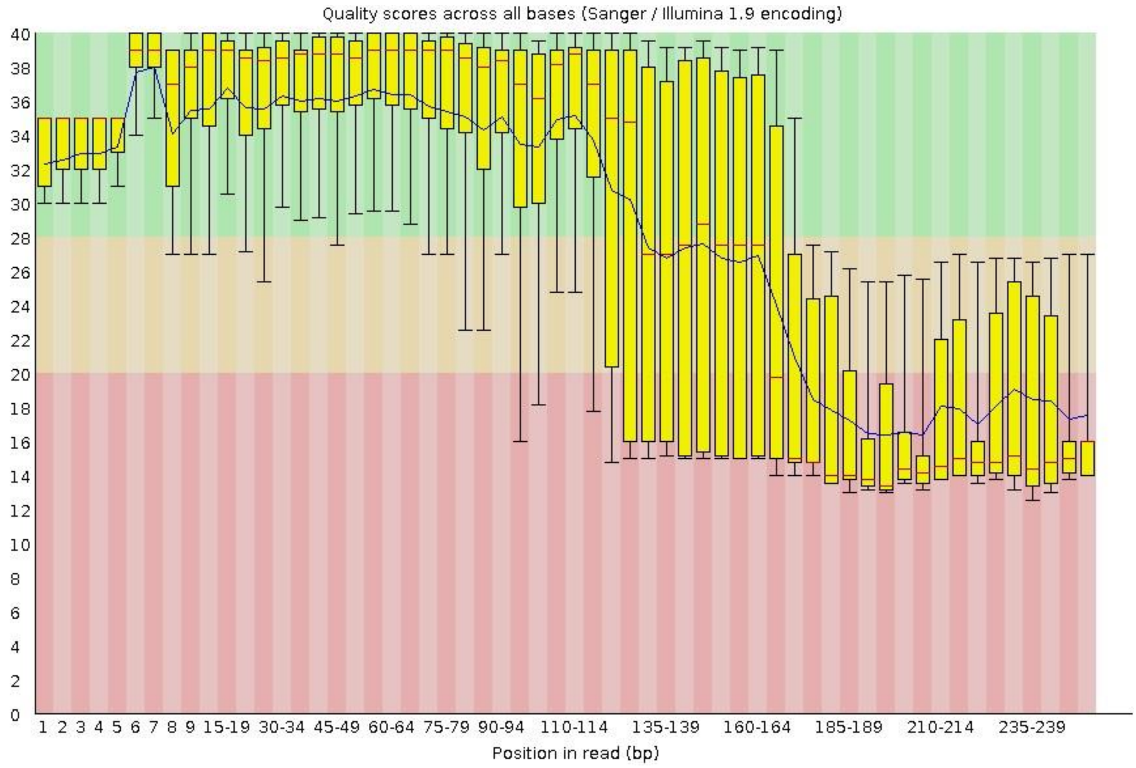
Quality of ERR4391162_2.fastq using **`fastqc`**



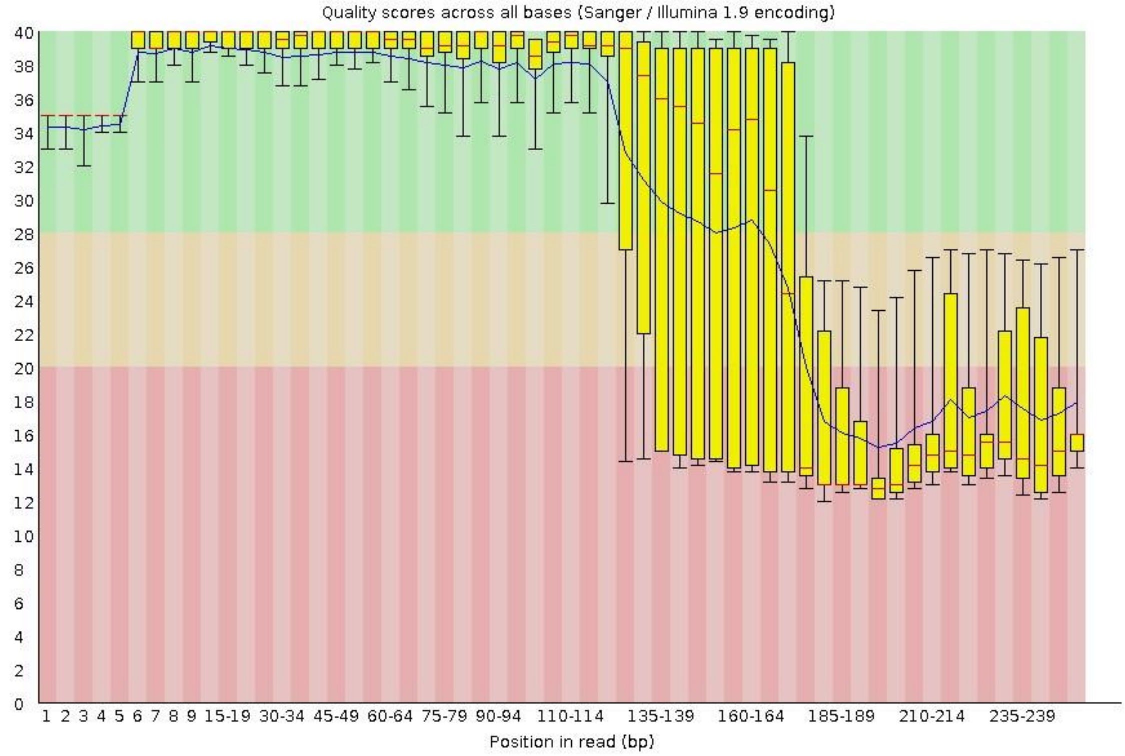
```
[up201805012@mbge Assignments]$ fastqc ERR4391162_2.fastq
Started analysis of ERR4391162_2.fastq
Approx 20% complete for ERR4391162_2.fastq
Approx 45% complete for ERR4391162_2.fastq
Approx 70% complete for ERR4391162_2.fastq
Approx 95% complete for ERR4391162_2.fastq
Analysis complete for ERR4391162_2.fastq
```

FastQC Report

Per base sequence quality - ERR4391162_1.fastq



Per base sequence quality - ERR4391162_2.fastq



The Per base sequence quality plot provides the distribution of quality scores across all bases at each position in the reads.

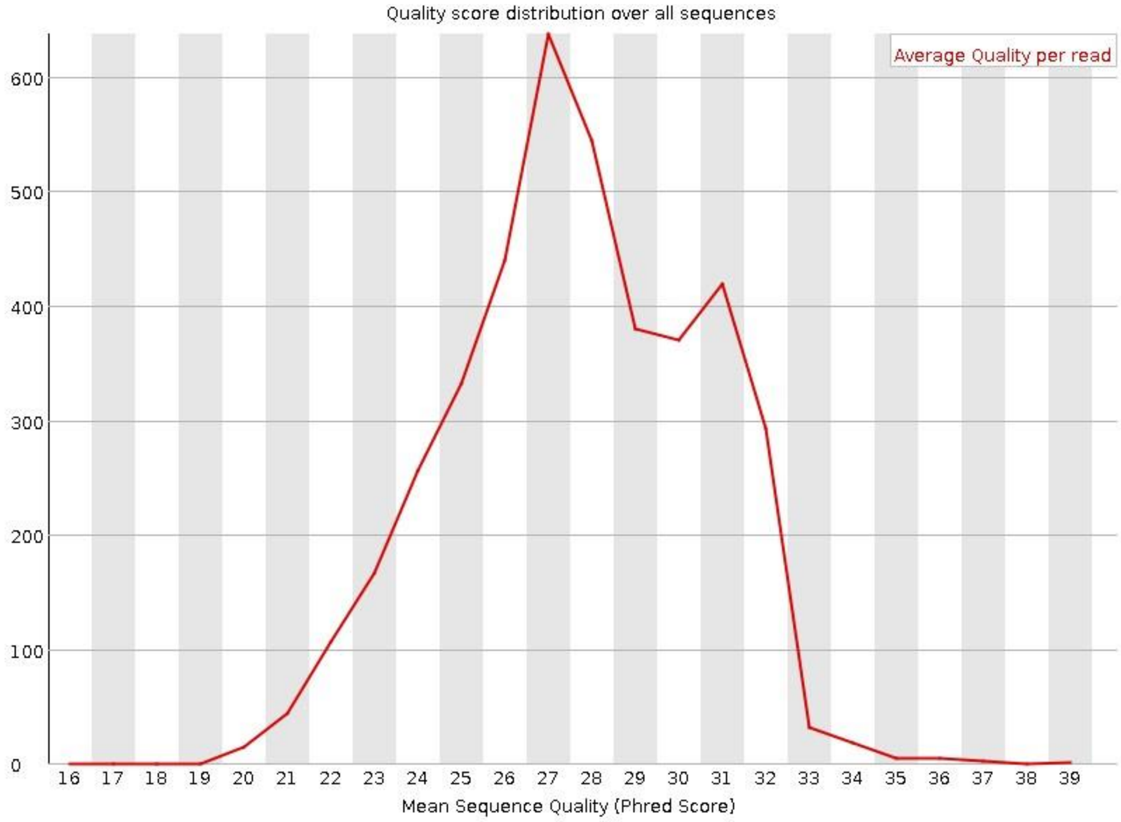
The y-axis on the graph shows the quality scores. The higher the score the better the base call. In the graph of the file ERR4391162_1.fastq we can observe that the quality of the reads starts to decrease from position 115, approximately. In the graph of file ERR4391162_2.fastq we can see that this decrease in the quality scores occurs a little later. Besides that in this graph we can also see that the quality scores are higher.

The background of the graph divides the y axis into very good quality calls (green), calls of reasonable quality (orange), and calls of poor quality (red). The quality of the calls in both files degrades over the course of the process , so we observe the base calls falling into the orange and red area towards the middle and the end of a read, respectively.

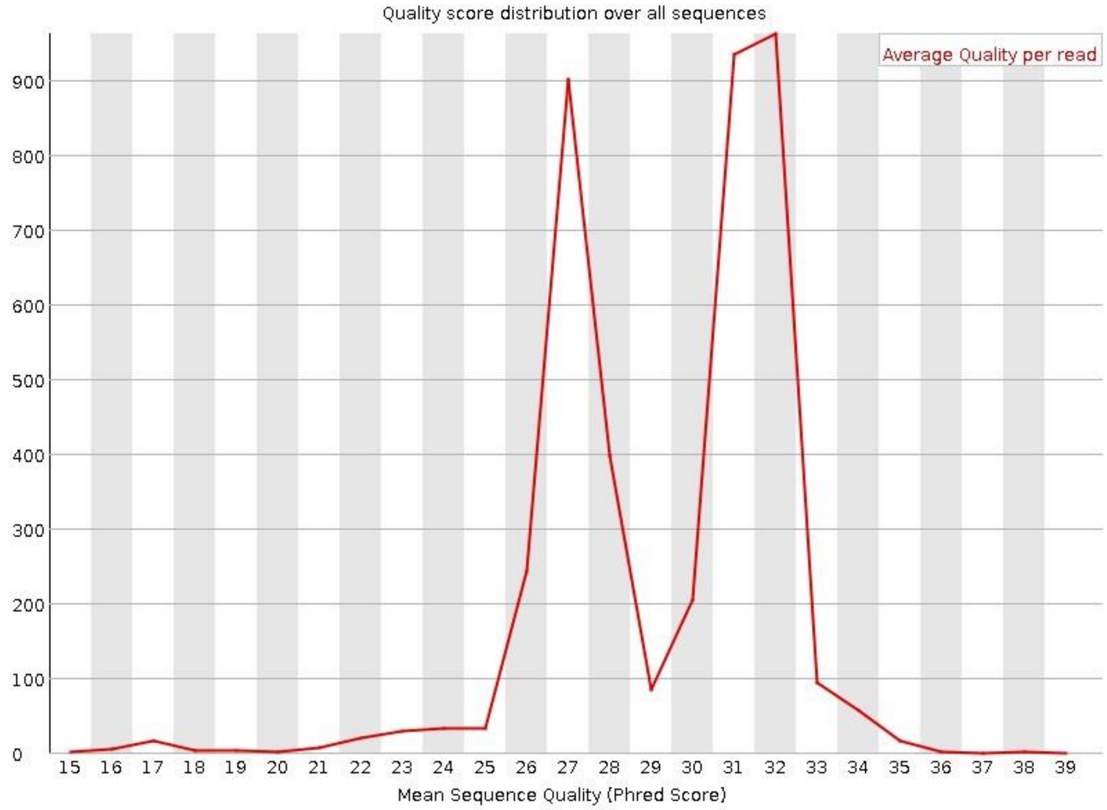
The blue line in the graphics represents the mean quality. We can see that the mean quality is much more uniform at the beginning of the graph of the file ERR4391162_2.fastq than the graph of the file ERR4391162_1.fastq. Towards the end of a read, the lines decline representing a drop of the mean quality in both files.

The inter-quartile range represented in the BoxWhisker plot, is much higher in file ERR4391162_1.fastq indicating that data points are more spread out so this file has lower quality scores than ERR4391162_2.fastq.

Per sequence quality scores - ERR4391162_1.fastq



Per sequence quality scores - ERR4391162_2.fastq



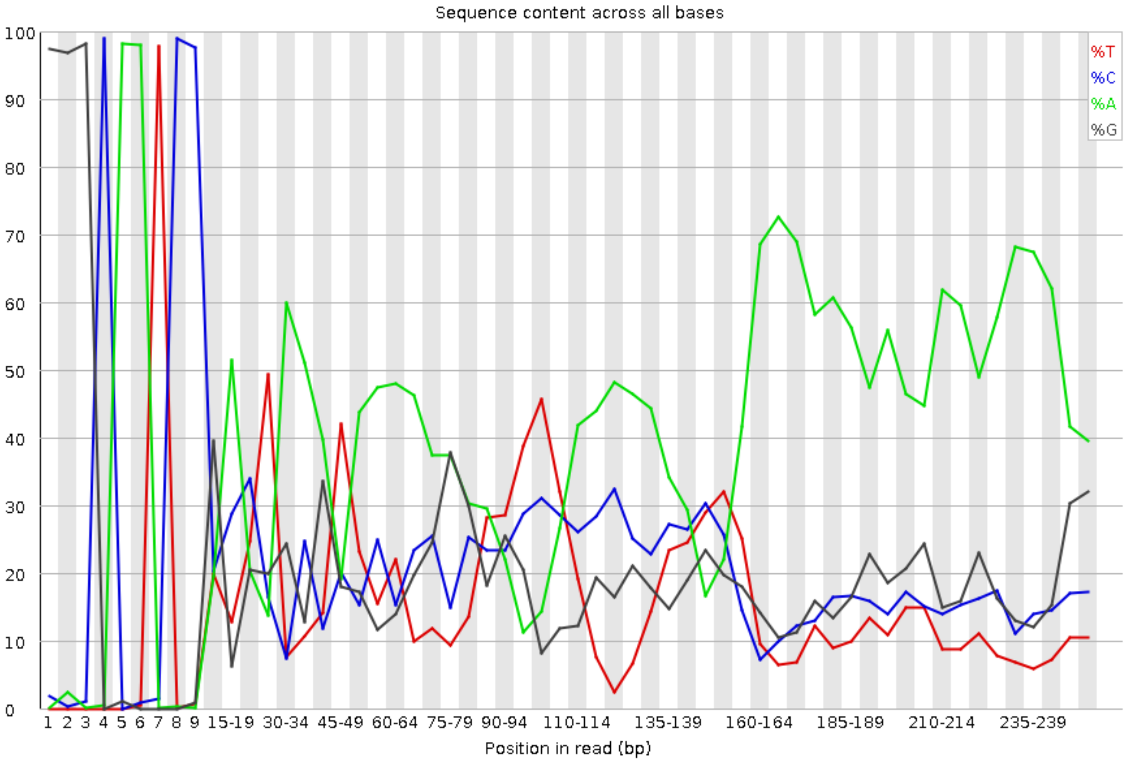
The per sequence quality score report allows us to see if a subset of sequences have universally low quality values. In both graphs we observe a binominal distribution of the quality of reads which indicates a poor quality of sequencing. Nevertheless the graph of the file ERR4391162_2.fastq shows higher average quality values per read (27 and 32) than the graph of the file ERR4391162_1.fastq (27 and 31).

Per Base Sequence Content

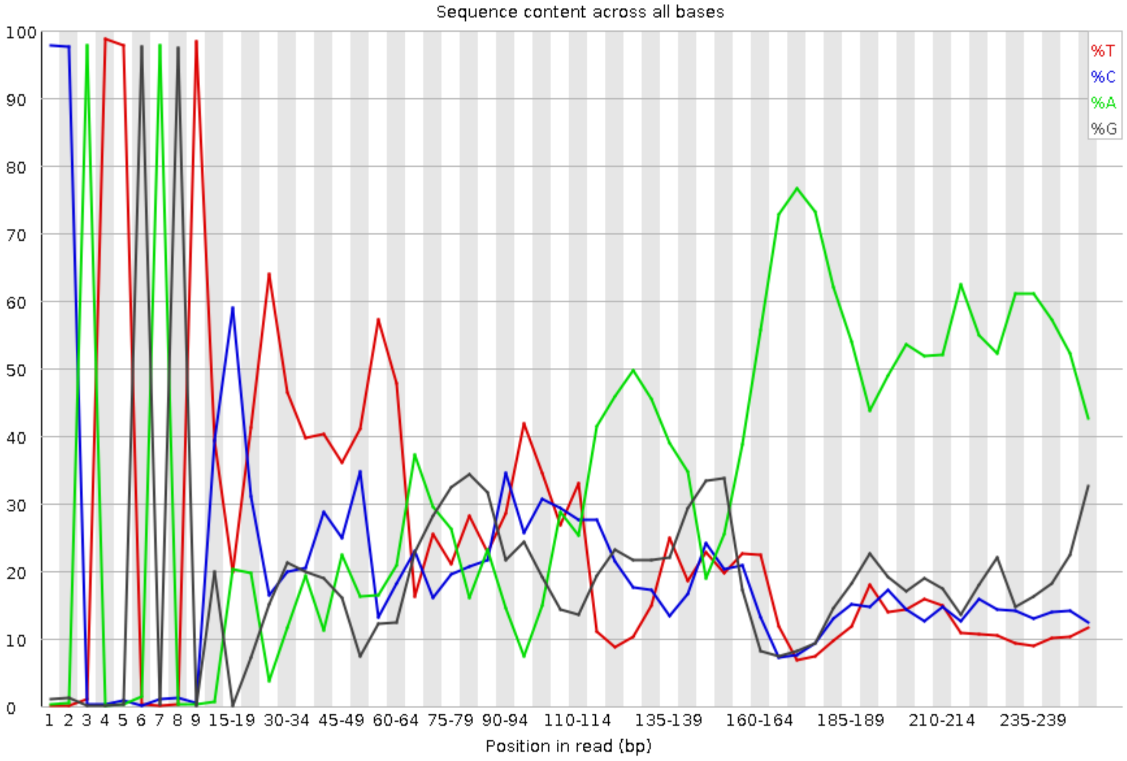
Per Base Sequence Content plots the proportion of the four normal DNA bases positions in a file.

Is expect that there would be little to no difference between the different bases of a sequence run, so the lines in this plot should run parallel with each other. This does not happen in any of the files, meaning the reads are compromised and indicating an overrepresented sequence.

Per Base Sequence Content - ERR4391162_1.fastq

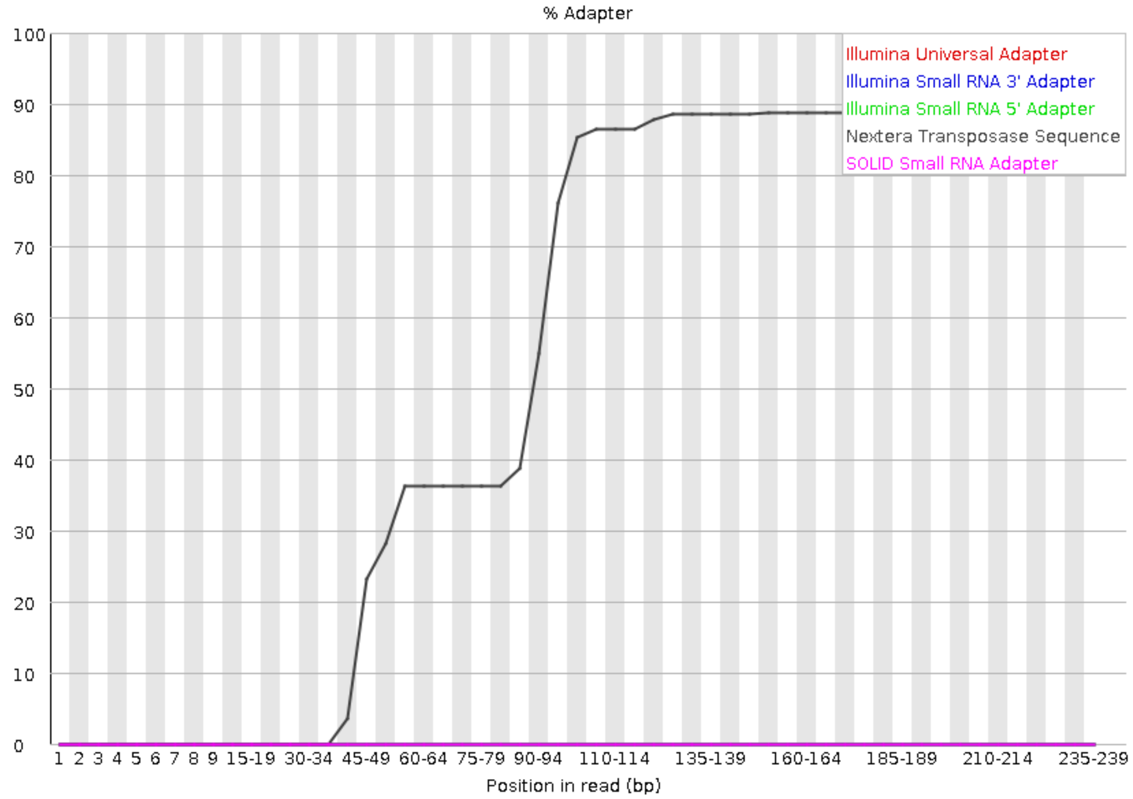


Per Base Sequence Content - ERR4391162_2.fastq

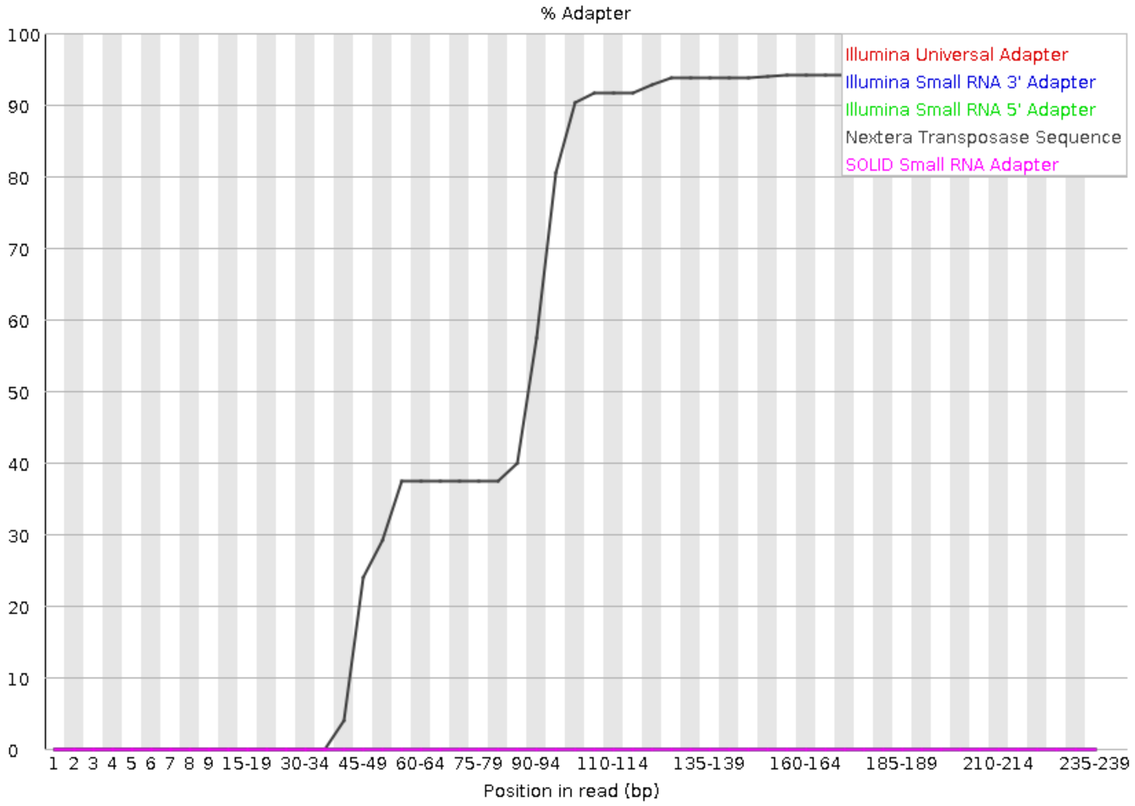


After the analysis of the previous graphics it can be concluded that the file with the highest quality is the ERR4391162_2.fastq.

Adapter Content - ERR4391162_1.fastq



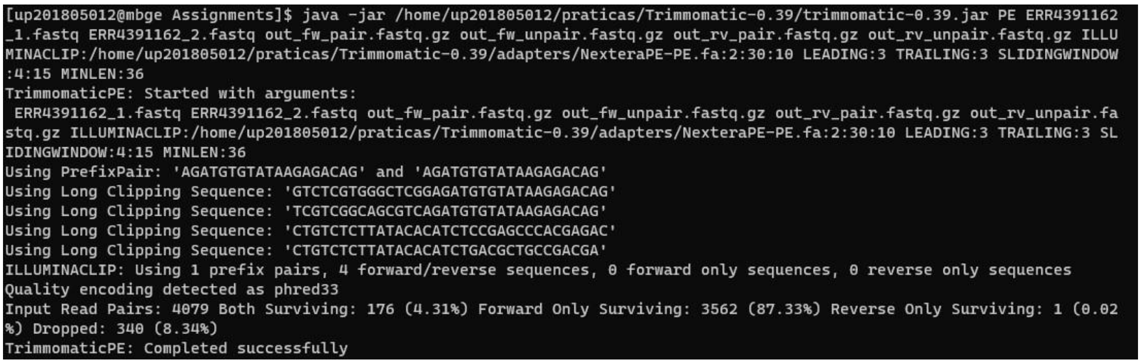
Adapter Content - ERR4391162_2.fastq



From the plot Adapter Content it can be seen that both files have adapters. The adaptor Nextera Transposase Sequence is present similarly and in large quantities in both files starting at position 34 of the read.

4. b)

In order to get clean reads without adapters and low quality bases, Trimmomatic was used with a sliding window trimming approach (SLIDINGWINDOW: 4:15) to remove reads with length smaller than 36 (MINLEN:36). If below a threshold quality of 3, the program also cut bases off the start (LEADING: 3) and end (TRAILING: 3) of the read.

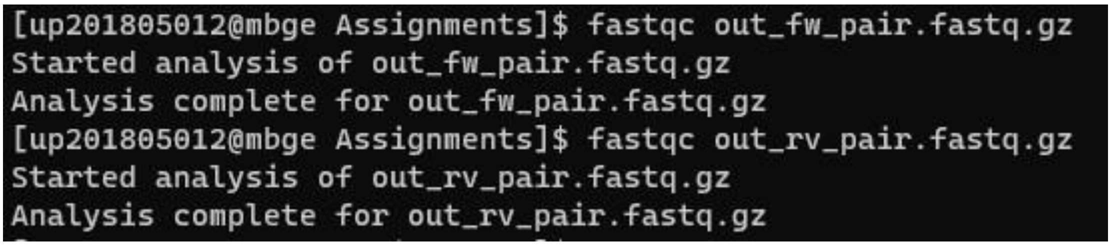


```
[up201805012@mbge Assignments]$ java -jar /home/up201805012/praticas/Trimmomatic-0.39/trimmomatic-0.39.jar PE ERR4391162
_1.fastq ERR4391162_2.fastq out_fw_pair.fastq.gz out_fw_unpair.fastq.gz out_rv_pair.fastq.gz out_rv_unpair.fastq.gz ILLU
MINACLIP:/home/up201805012/praticas/Trimmomatic-0.39/adapters/NexteraPE-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW
:4:15 MINLEN:36
TrimmomaticPE: Started with arguments:
 ERR4391162_1.fastq ERR4391162_2.fastq out_fw_pair.fastq.gz out_fw_unpair.fastq.gz out_rv_pair.fastq.gz out_rv_unpair.fa
stq.gz ILLUMINACLIP:/home/up201805012/praticas/Trimmomatic-0.39/adapters/NexteraPE-PE.fa:2:30:10 LEADING:3 TRAILING:3 SL
IDINGWINDOW:4:15 MINLEN:36
Using PrefixPair: 'AGATGTGTATAAGAGACAG' and 'AGATGTGTATAAGAGACAG'
Using Long Clipping Sequence: 'GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG'
Using Long Clipping Sequence: 'TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG'
Using Long Clipping Sequence: 'CTGTCTCTTATACACATCTCCGAGCCCACGAGAC'
Using Long Clipping Sequence: 'CTGTCTCTTATACACATCTGACGCTGCCGACGA'
ILLUMINACLIP: Using 1 prefix pairs, 4 forward/reverse sequences, 0 forward only sequences, 0 reverse only sequences
Quality encoding detected as phred33
Input Read Pairs: 4079 Both Surviving: 176 (4.31%) Forward Only Surviving: 3562 (87.33%) Reverse Only Surviving: 1 (0.02
%) Dropped: 340 (8.34%)
TrimmomaticPE: Completed successfully
```

In both pair were kept 4.31% of reads.

c)

Re-run **`fastqc`** of ERR4391162_1.fastq (out_fw_pair.fastq.gz) and ERR4391162_2.fastq (out_rv_pair.fastq.gz).



```
[up201805012@mbge Assignments]$ fastqc out_fw_pair.fastq.gz
Started analysis of out_fw_pair.fastq.gz
Analysis complete for out_fw_pair.fastq.gz
[up201805012@mbge Assignments]$ fastqc out_rv_pair.fastq.gz
Started analysis of out_rv_pair.fastq.gz
Analysis complete for out_rv_pair.fastq.gz
```

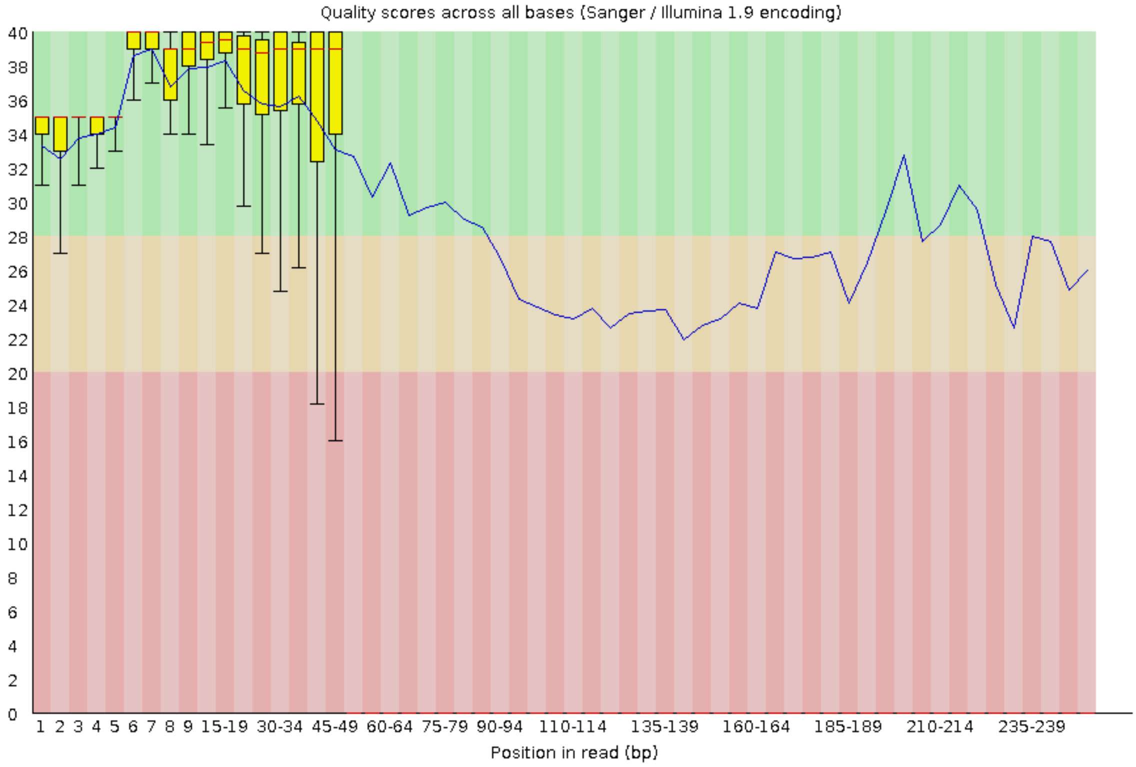
FastQC Report

Per base sequence quality

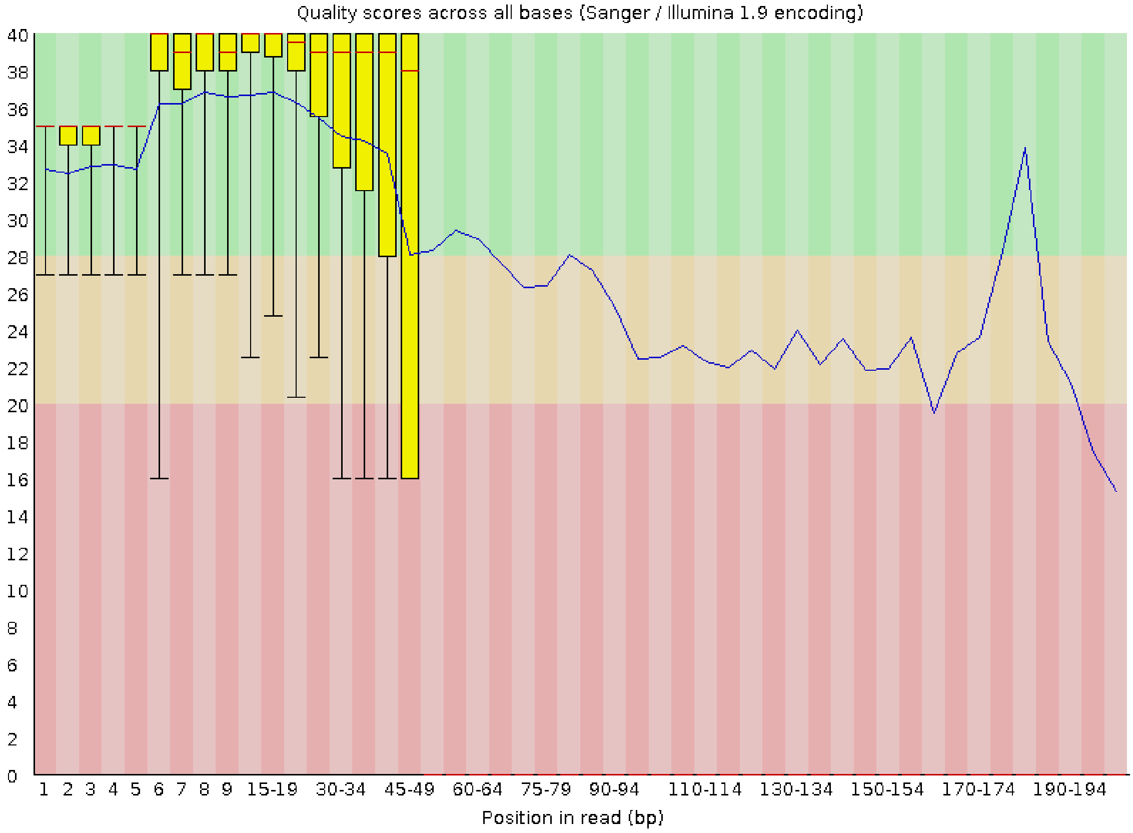
The loss of reads in both files can be seen in the Per base sequence quality plot, as only 4.31% of reads were kept.

In general, the quality of the reads has improved, and in the file out_fw_pair.fastq.gz the reads remained in the green zone. On the other hand, in the file out_rv_pair.fastq.gz the quality of the calls improved but a small number of calls kept falling into the red zone (position 45-49).

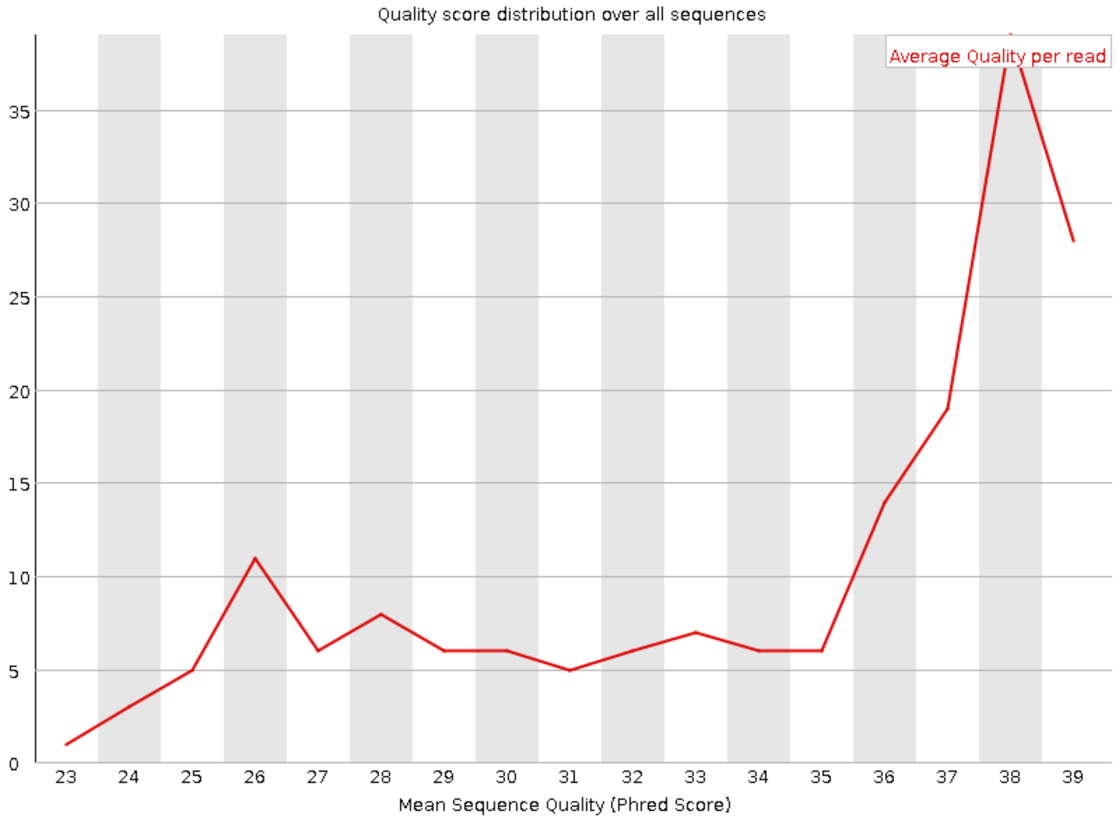
Per base sequence quality - out_fw_pair.fastq.gz



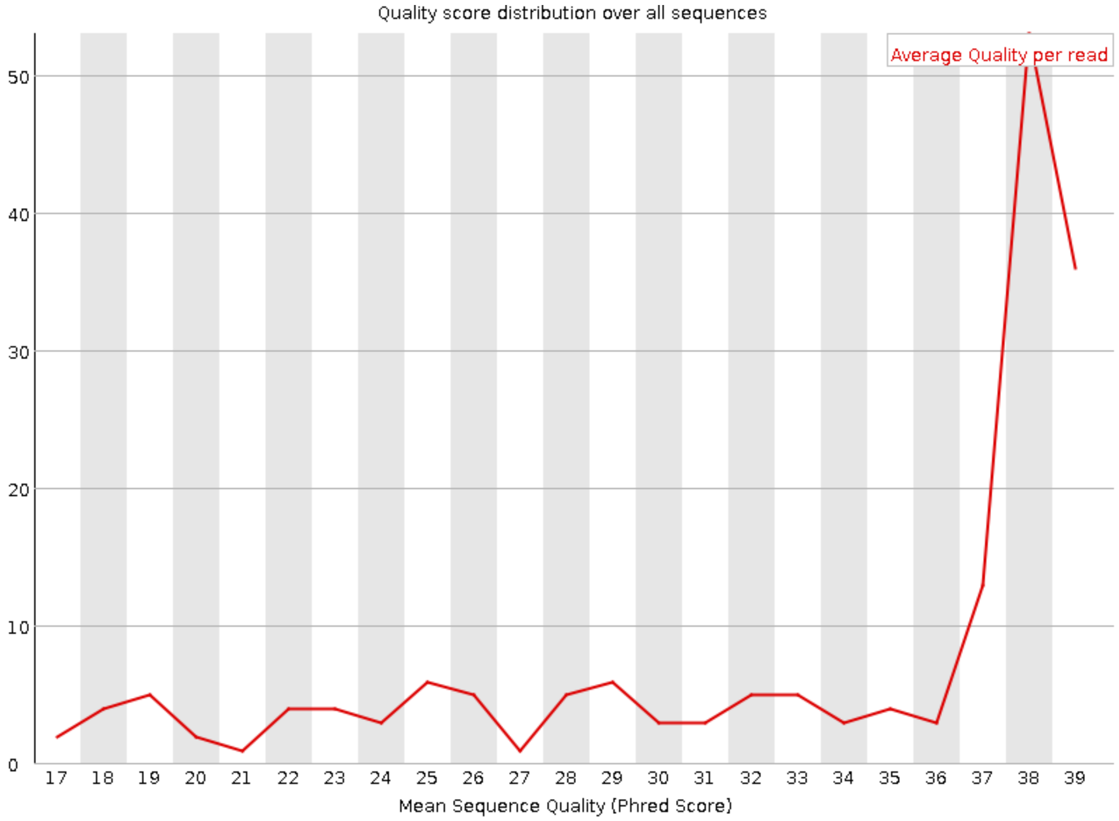
Per base sequence quality - out_rv_pair.fastq.gz



Per sequence quality scores - out_fw_pair.fastq.gz

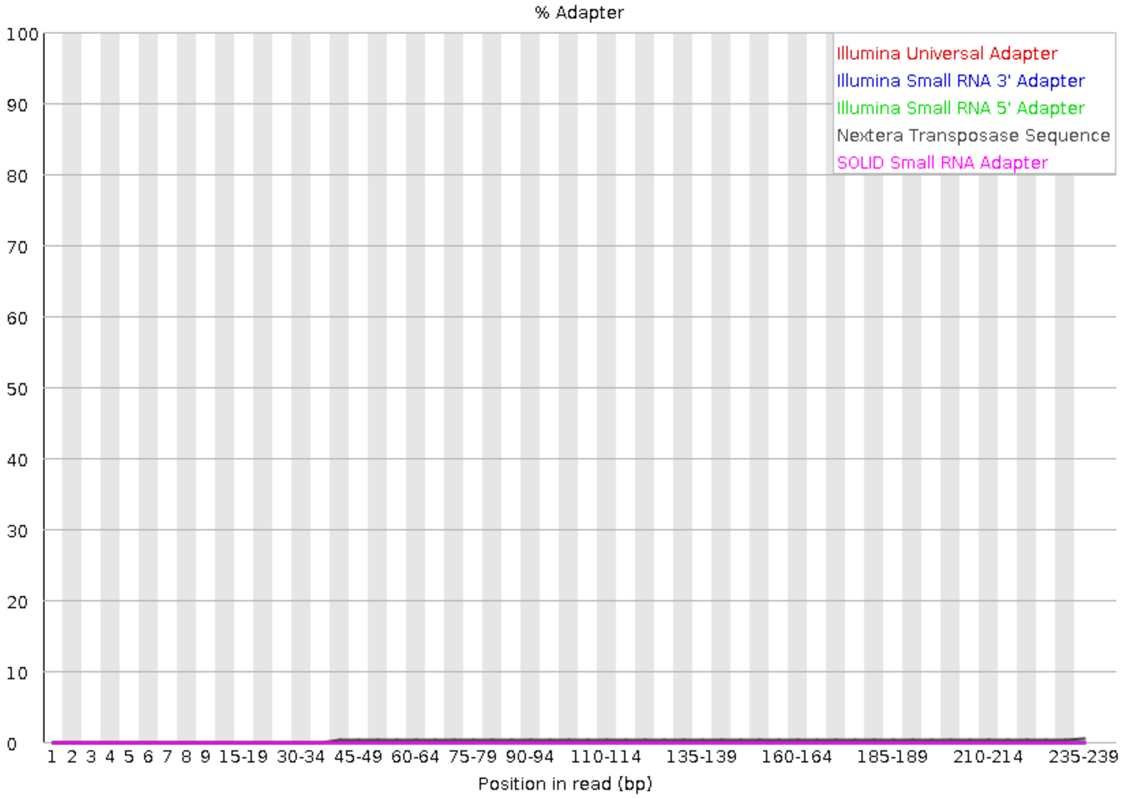


Per sequence quality scores - out_rv_pair.fastq.gz

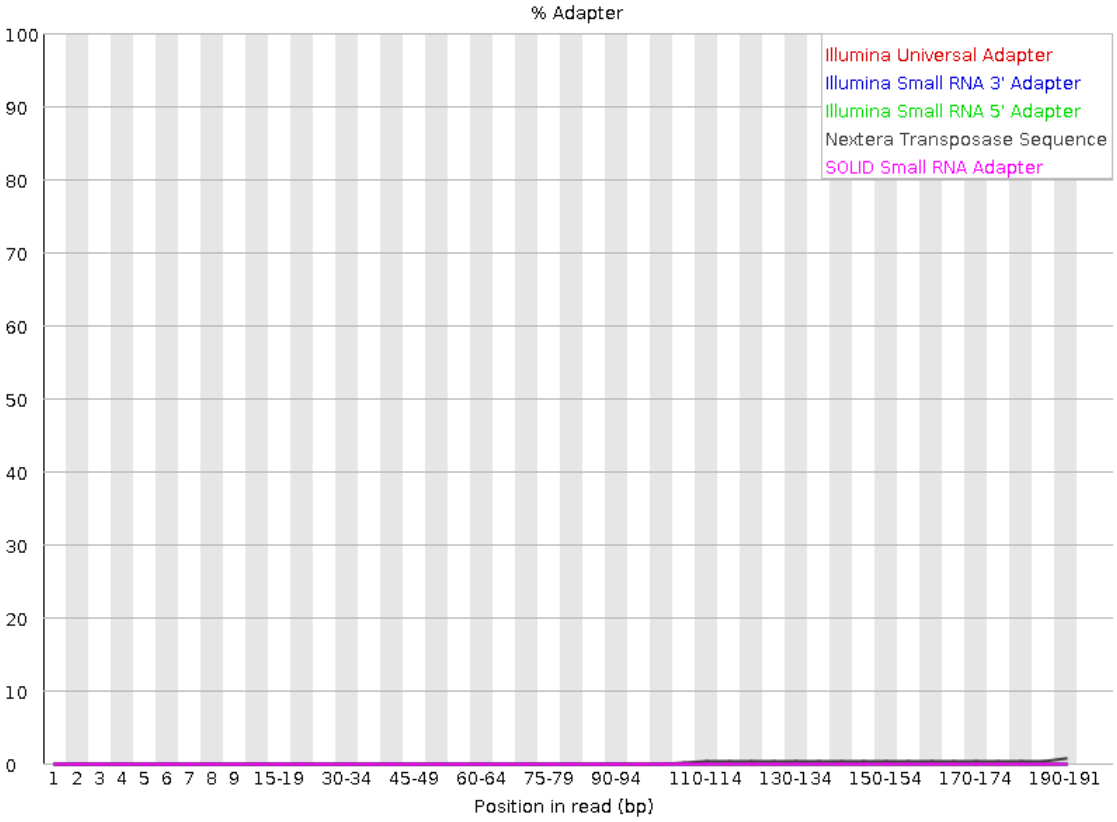


The per sequence quality score report shows an increase in the average quality of the reads in both files (Average quality per read is 38).

Adapter Content - out_fw_pair.fastq.gz



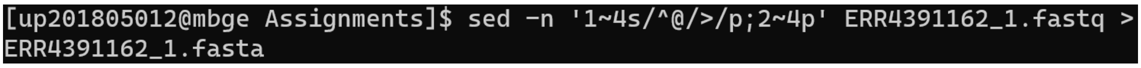
Adapter Content - out_rv_pair.fastq.gz



From the plot Adapter Content it can be concluded that the adapters have been successfully removed in both files.

5.

The **sed** command can be used to convert a FASTQ format file to a fasta format file. This command suppresses all output (**-n**) , then explicitly prints every 4 lines, starting from the first line (**1~4**) , replacing every leading **@** by a **>** . Then it prints every 4 lines, beginning from the second line (**2~4**) and we get the sequence header and the sequence needed for fasta format. Finally input the data file to be converted and its final format (**INFILE.fastq > OUTFILE.fasta**).



```
[up201805012@mbge Assignments]$ sed -n '1~4s/^@/>/p;2~4p' ERR4391162_1.fastq >
ERR4391162_1.fasta
```

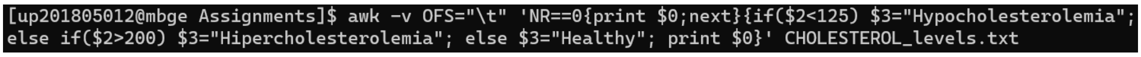
6.

The awk command is an effective command-line tool for text processing that works well with column-based data.

The command built-in variables **OFS** (Output Field Separator) is used to set the output field separator which is a space by default. **'\t'** specifies that the delimiter is the tab character.

Starting at line 0 (**NR==0**), a new column is created where **$0** represents the value of the entire record the **awk** program read on standard input. Finally a list of conditions **if else** is created according to the date received and using the values in the second column (**$2**).

- 125 – 200 mg/dL - "Healthy"
- < 125 mg/dL - "Hypocholesterolemia"
- > 200 mg/dL - "Hipercholesterolemia"



```
[up201805012@mbge Assignments]$ awk -v OFS="\t" 'NR==0{print $0;next}{if($2<125) $3="Hypocholesterolemia";
else if($2>200) $3="Hipercholesterolemia"; else $3="Healthy"; print $0}' CHOLESTEROL_levels.txt
```