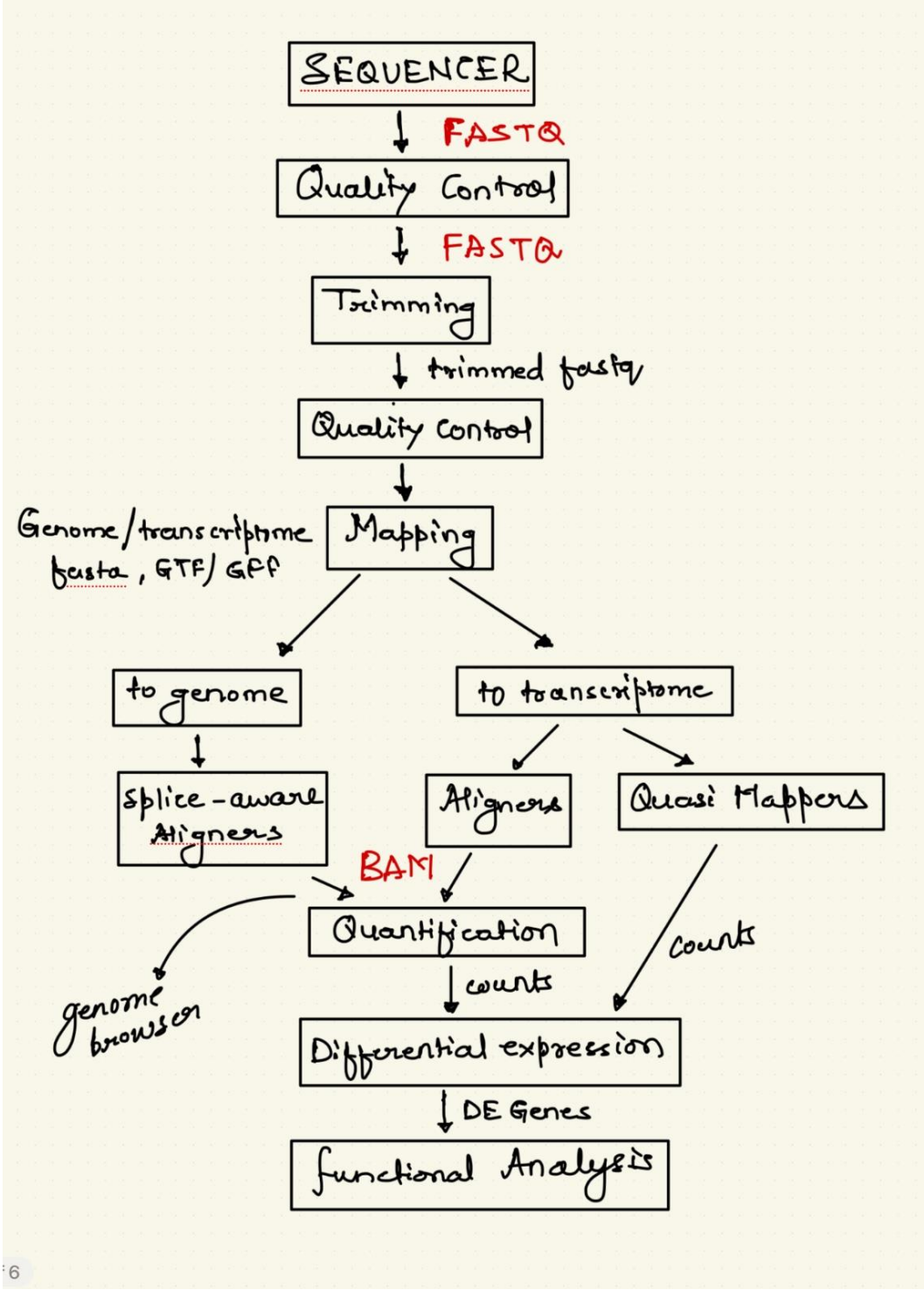
Bio-informatics Tasks

Tasks:

- Downloading a FASTQ File - https://drive.google.com/file/d/1DGHjbhcRy_zTm6H9C_AUpkzBML-JhtA3/view
- Converting FASTQ File to SAM and BAM Format.
- SAM format is known as **Sequence Alignment Map** and BAM files are **Binary representation of SAM Files** which are compressed SAM Files.

General Workflow:



We did work till converting FASTQ to a SAM & BAM file.

Steps taken:

- As we cannot directly convert the FASTQ Files into SAM and BAM formats, we have to first create a pipeline to do this task where we first get the FASTQ file upon which we do the **"Quality Control"** of reads from sequencer which are in FASTQ files. To perform we used FASTQC as a tool to check quality control reports. Goal of this step: We did quality control to look for bases from the reads that are of poor quality and also to check if any adapter sequences which are present.
- To remove such poor-quality reads and to remove adapter sequences we performed "Trimming". For this task we used **Trimmomatic tool** to perform this task.
- After getting good quality reads from the above step, we performed **"Mapping"** to a genome file which was downloaded from this link - https://genome-idx.s3.amazonaws.com/hisat/grch38_genome.tar.gz
- Using HISAT we mapped this genome files to the Trimmed FASTQ file which contains RNA Sequences, here using HISAT2 and samtools we outputted the BAM and SAM file using Bash pipeline script.
- This study helped us to understand about the genome file system and on this basis, we also understand that using this BAM file we can do quantification such as sequence counts and also this can be used to create Differential Expressions which are further used to do functional analysis depending on the user-based research scenarios.

Shell File Commands: -

```
SECONDS=0

# change working directory

cd /Desktop/Fastq/

# STEP 1: Running fastqc

fastqc data/demo.fastq -o data/

# run trimmomatic to trim reads with poor quality

java -jar ~ /Trimmomatic-0.39/trimmomatic-0.39.jar SE -threads 4 data/demo.fastq data/demo_trimmed.fastq
TRAILING:10 -phred33

echo "Trimmomatic finished running!"

fastqc data/demo_trimmed.fastq -o data/

# STEP 2: Running HISAT2

mkdir HISAT2

# get the genome indices

wget https://genome-idx.s3.amazonaws.com/hisat/grch38_genome.tar.gz

# run alignment

hisat2 -q --rna-strandness R -x HISAT2/grch38/genome -U data/demo_trimmed.fastq | samtools sort -o
HISAT2/demo_trimmed.bam

hisat2 -q --rna-strandness R -x HISAT2/grch38/genome -U data/demo_trimmed.fastq -S
HISAT2/demo_trimmed.sam

echo "HISAT2 finished running!"

duration=$SECONDS
```

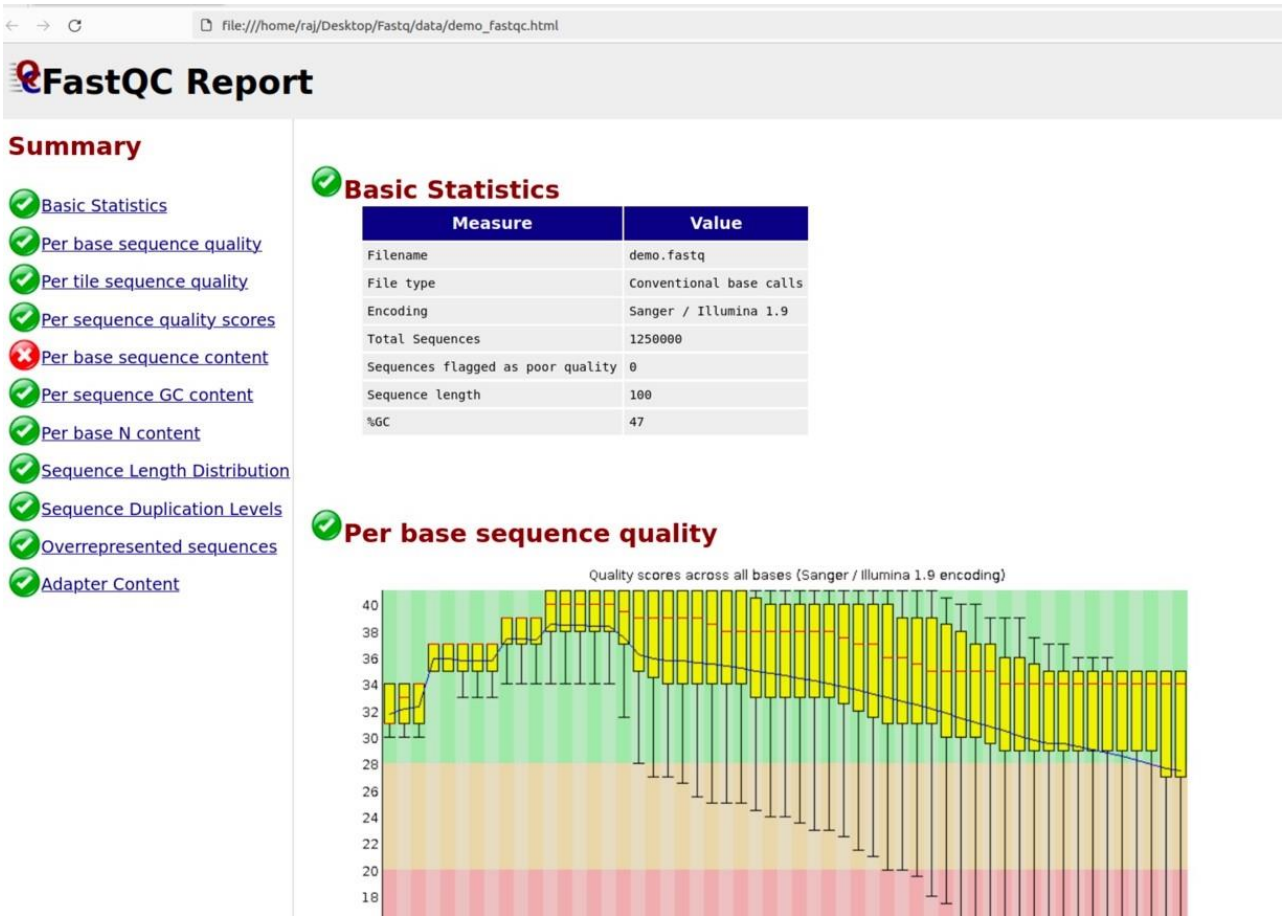
echo "$(($duration / 60)) minutes and $(($duration % 60)) seconds elapsed."

Working Environment-

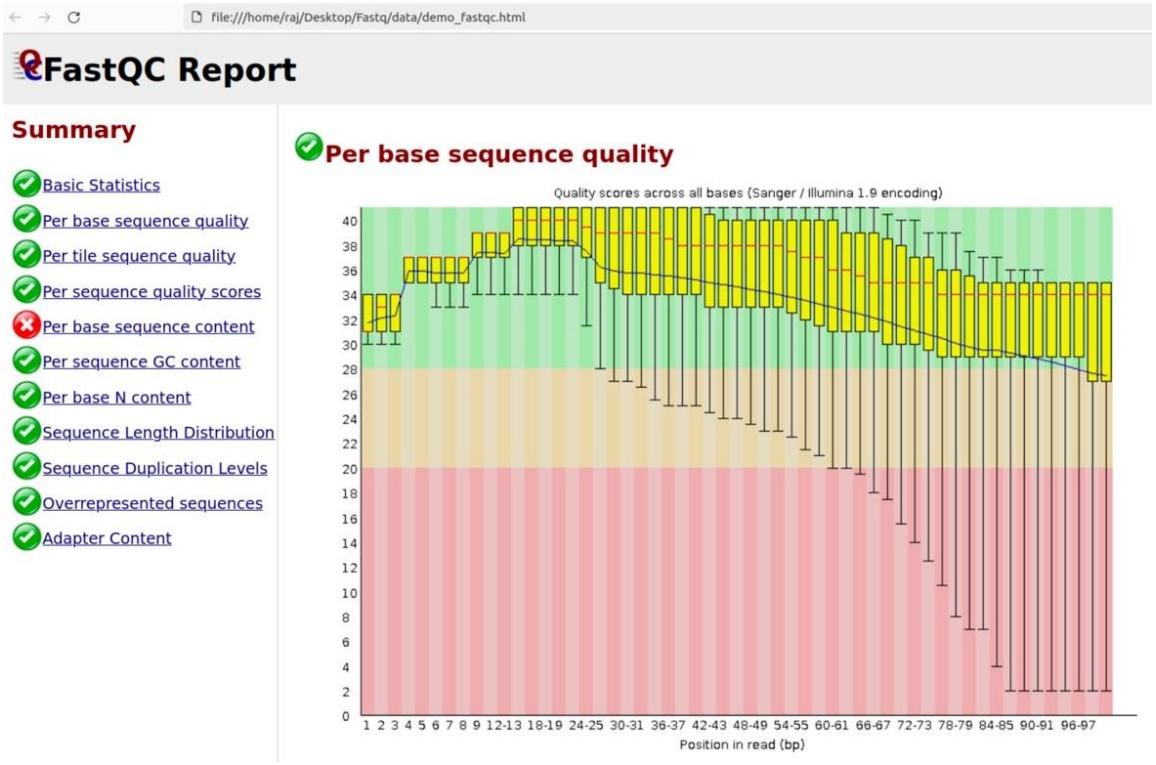
I downloaded Oracle VM VirtualBox on laptop and installed Ubuntu to do this task as my MacBook did not gave me permissions to download third-party apps and was throwing errors with wget commands as I was not able to download the wget on the terminal. Using ubuntu (Linux) terminal the task was successfully performed and executed.

Screenshots –

- FASTQC Report after doing initial quality control.

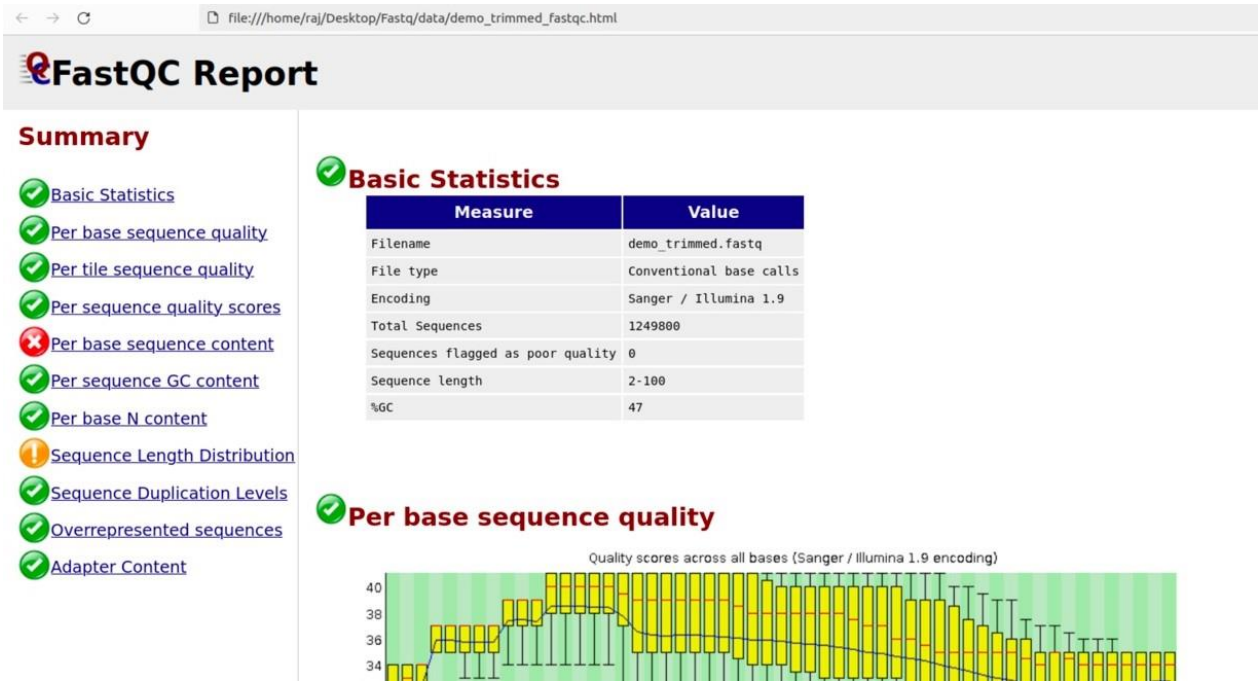


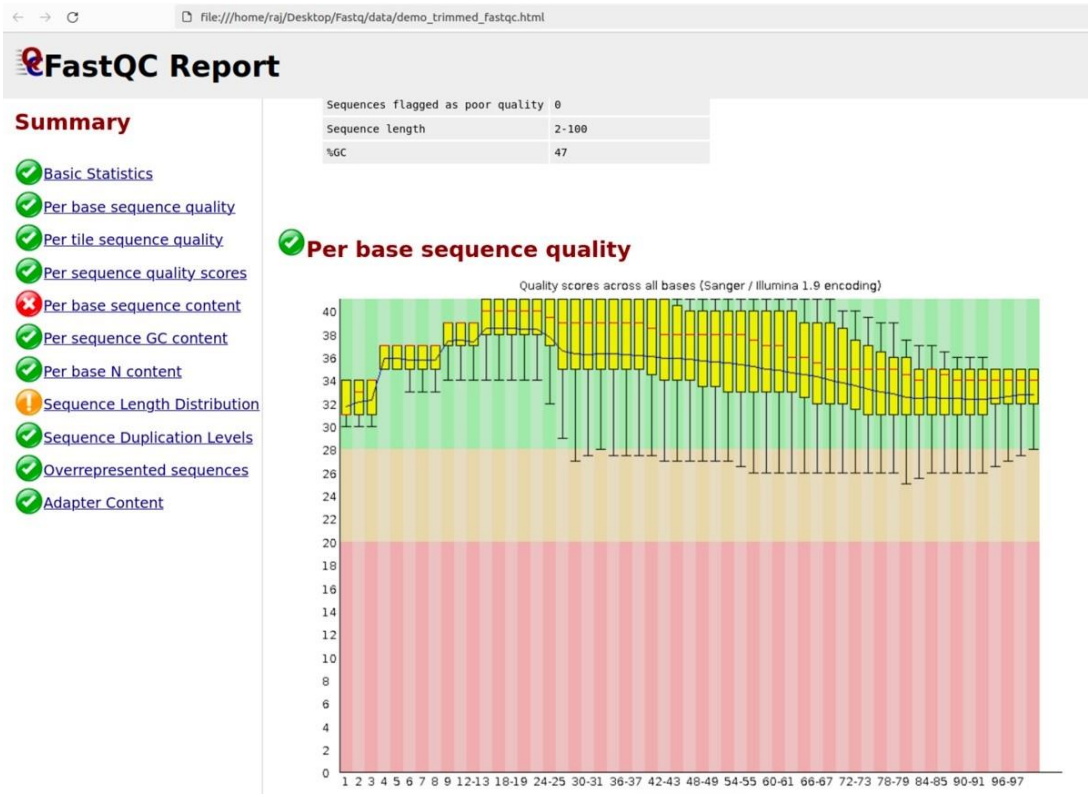
The intial FASTQ file has 1,250,000 Total squences which we can see in the basic statistics above.



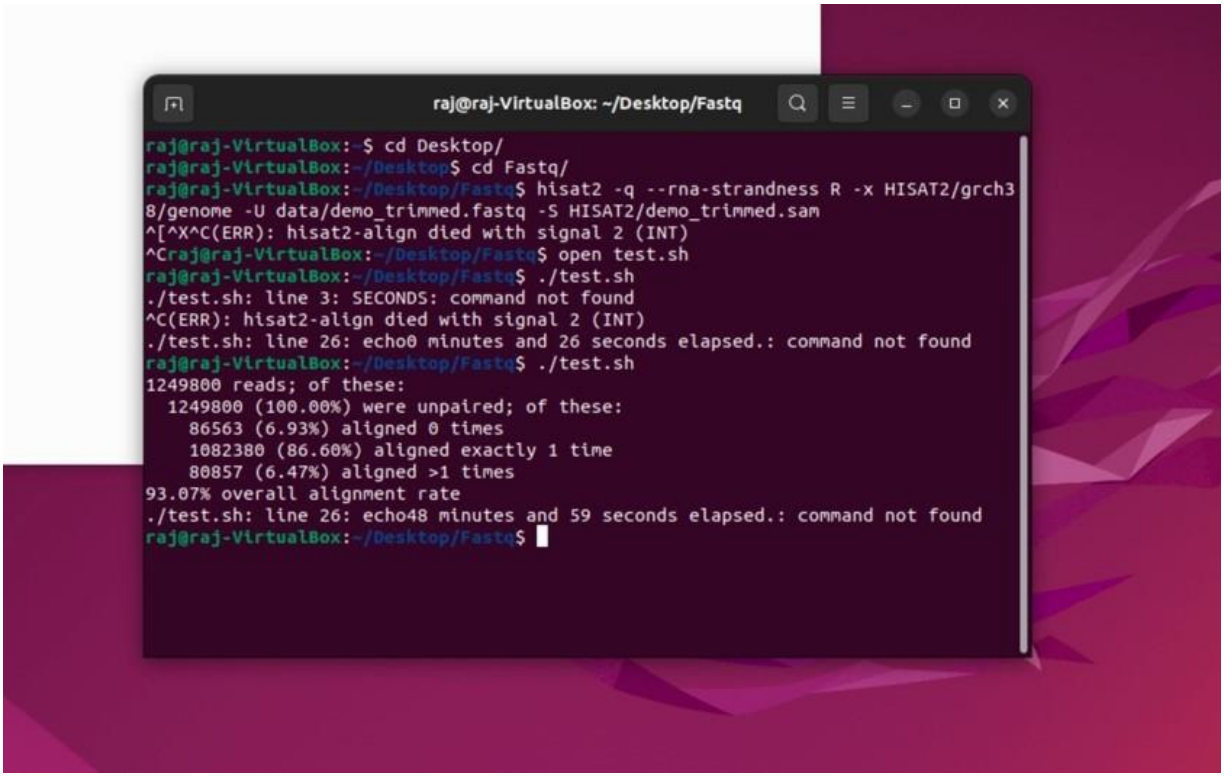
As we can see from the base sequence quality above, the end sequence quality scores or phred are not that good, hence we did trimming.

- After trimming



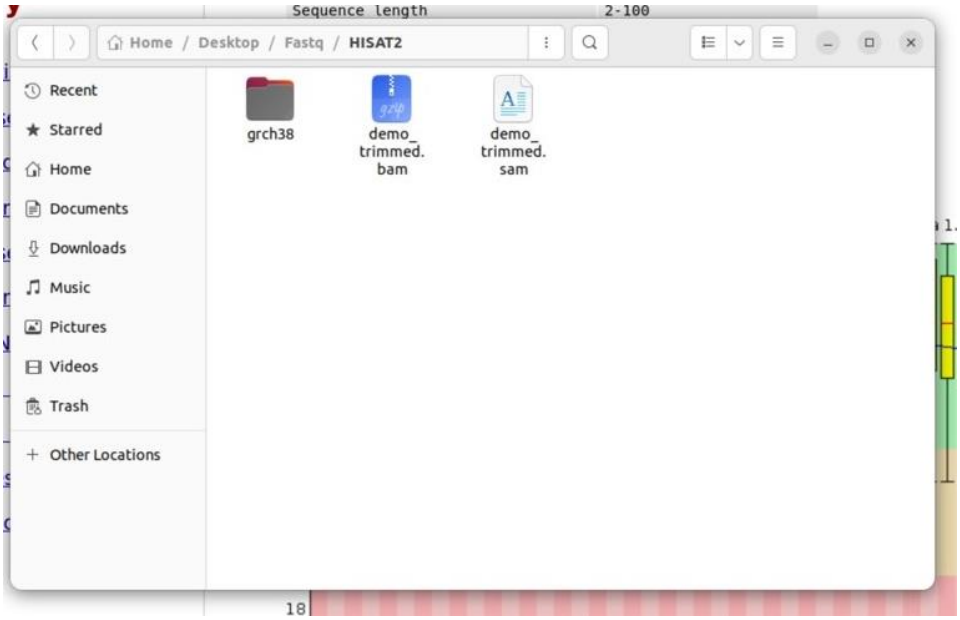


After trimming we again did quality control to check the results of the trimming, as you can see above all sequences are of good quality and it got improved from the initial fastq file.

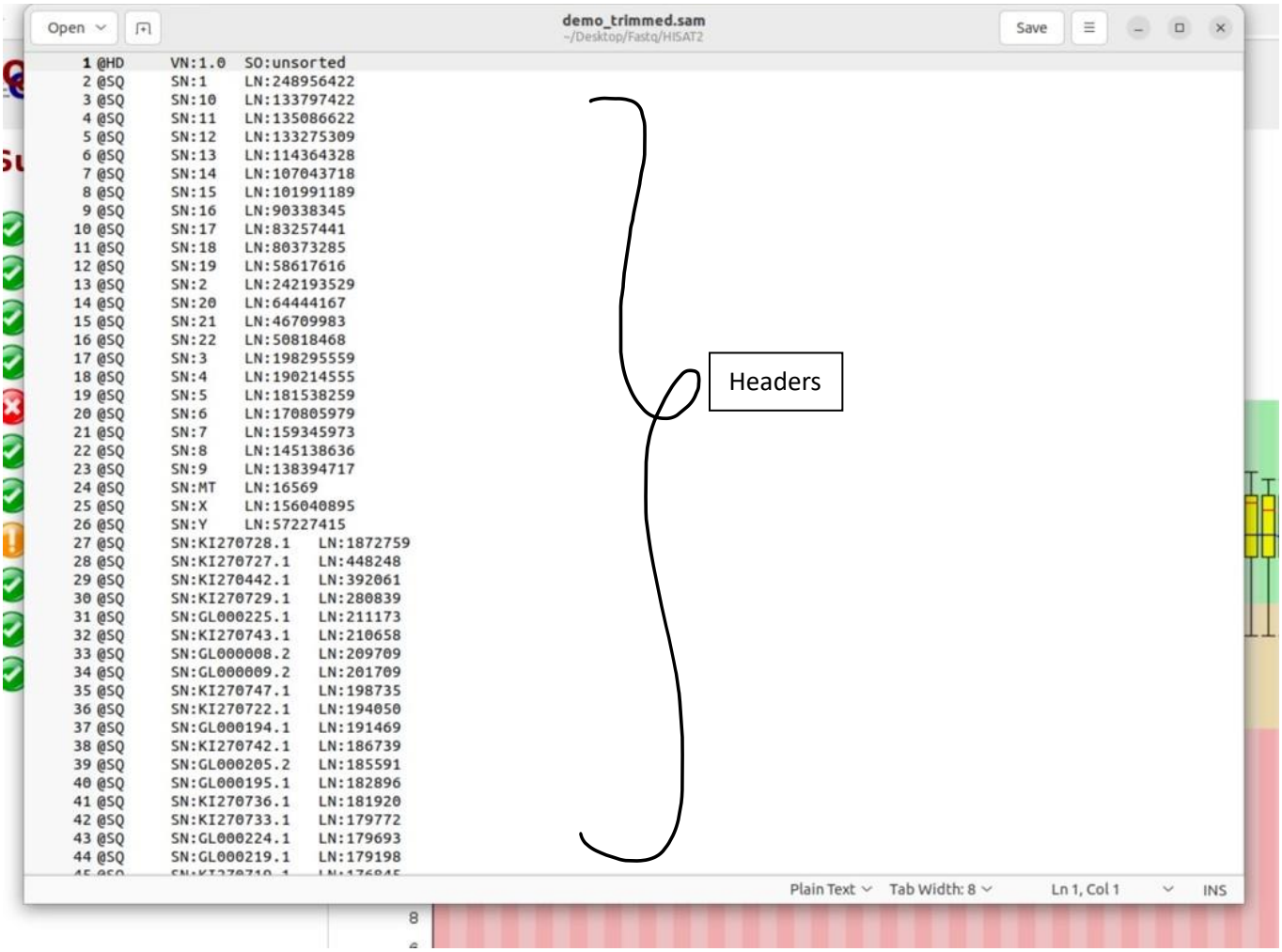


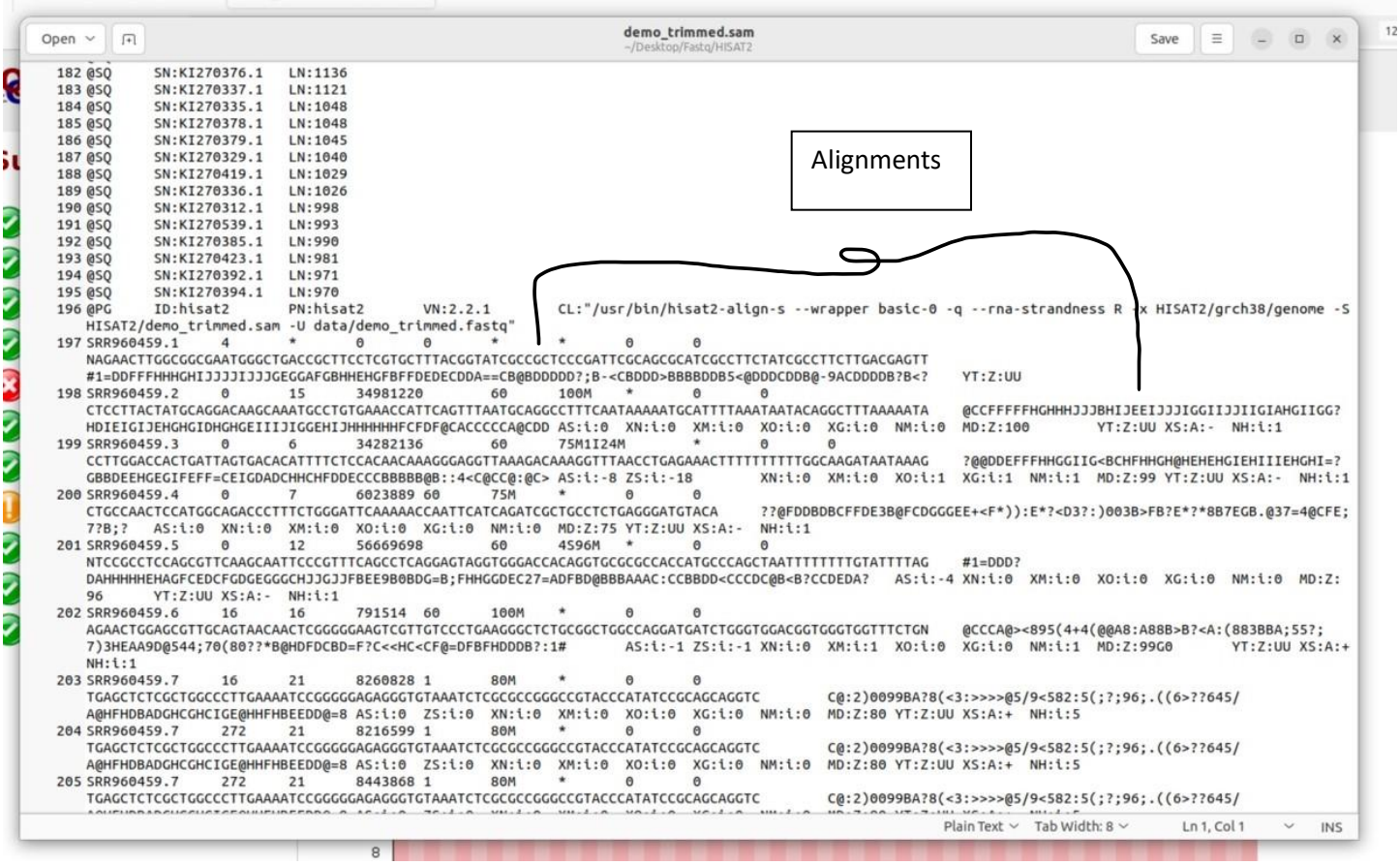
After this we mapped the trimmed file to the human genome data, and using HISAT2 we outputted the trimmed fastq file as a SAM file and using HISAT2 and samtools we outputted the trimmed fastq file as a BAM file. From the above terminal snippet, you can see that it took 48 minutes and 59 seconds for this mapping and conversion.

- **Snippet of the files converted:**

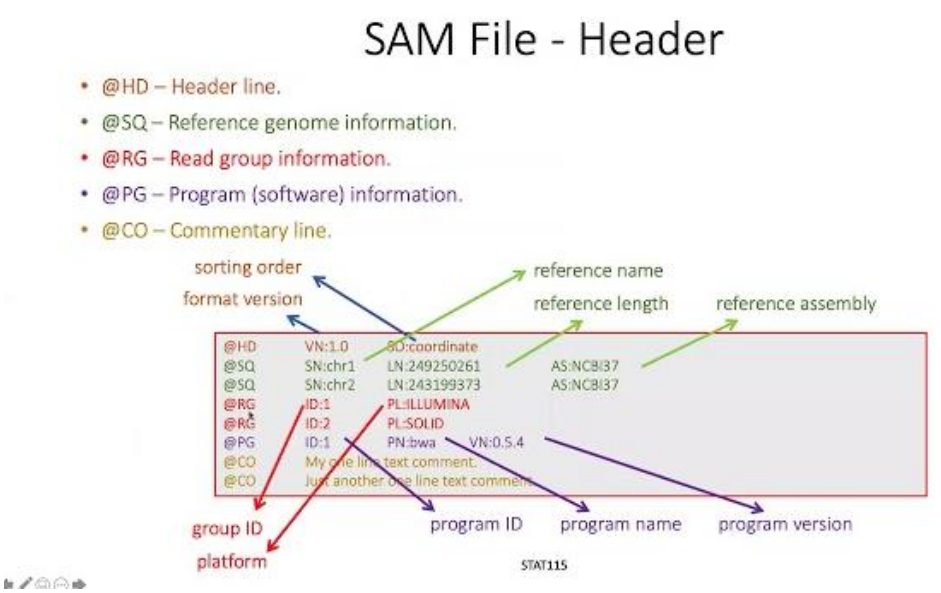


Output of the SAM file:





Header understanding of the SAM file:



Steps Followed: