Work flow for Variant Calling

1)Data Collection
2)Data processing
3)Alignment
4) Variant calling and data analysis

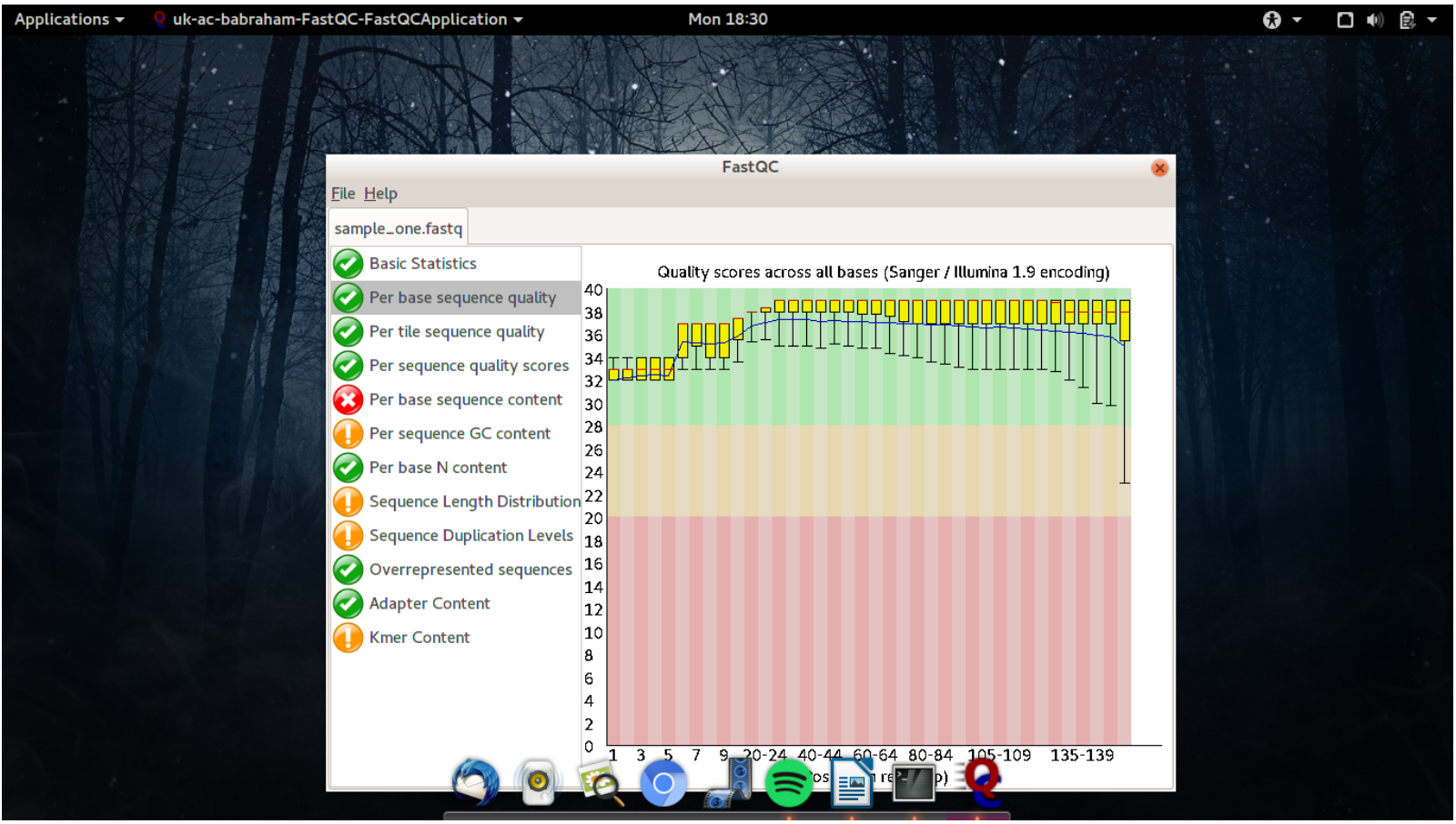
Data Collection
The sequence reads were provided in fastQ format.
The reference genome was downloaded from USCS
Programming language : Python and shell scripting.
Tools and packages : FastQC, bowtie2 , Biopython, Homer, Cutadpt.

Data Processing

Step 1 – Perform Quality control for FastaQ files
I performed the Quality control by looking at the Phred scores and with help of visualization with FastaQc tool

Calculating phred score;
The python script ‘phredscore.py’ would calculate the score and return values,
Generally we consider reads to be good if the phred score is above 20 and for the reads given we have most of them above 20 and hence we can bypass this step

Trimming and Kmer Content;
By looking at the plots from FastQc we can determine if we need to trim our sequences, since we dont have any overrepresented sequences we can avoid trimming of our reads .
Also our ‘Per base Seq quality’ seems to be on the higher side.



Step -2 : Mapping of the sequence with reference genome. Alignment.
The reference genome or wholegenome fasta for humans hg19 was obtained with the help of the script

‘rsync -avzP rsync://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/ .’

sort the required chromosomes, based on the chromosome name.
For our project we required chr1 – chr22 chromosome.
Unmask the chromosome and convert all the lower letter sequence to capital seq.
Concatenate all the chr * files, make sure to remove the random.fa files, for better alignment.

Alignment and Variant calling,
The fastq files given for analysis and the reference genome files given amounted to 500 Mb and 3 Gb files, unfortunately my system could not handle these files and I was unsuccessful in making my alignments and getting the output for performing Variant calling.

** It could also be that my approach was wrong from step 1 and hence could not perform a proper analysis.**