Functional Genomics Final Project

Given the sample fastq file. (392_1.fastq.gz on the server /mnt/gkhazen/NGS-Fall2020/FinalProject/)
Build a complete pipeline for the identification and annotation of variants. (You should identify both SNPs and INDEL).
You should explain every step you use in the pipeline and justify the arguments used, from quality control till variant annotation. All codes should be uploaded to your gitlab folder. Even if you run the code on the server, you should upload your code and commands used to your gitlab. Each student will use a different chromosome for mapping. You should download it from the proper resource and index it.

Charbel El Gemayel: Chromosome 1	Mohamed Mehdi Merbah: Chromosome 7
Auriane Mahfouz: Chromosome 2	Kelven Rahy: Chromosome 8
Bilal Hamdanieh: Chromosome 3	Melissa El Feghali: Chromosome 9
Elissa Lichaa El Khoury: Chromosome 4	Nesrine Naaman: Chromosome 10
Aquilina Barbouche: Chromosome 5	Elissa Younan: Chromosome 11
Margueritta Abi Younes: Chromosome 6	Lynn El Khoury: Chromosome 12
	Pia Chouaifati: Chromosome 13

Mandatory information and statistics to report:

- Based on the read which Illumina version was used
- Tiles with bad quality if any. (Number and IDs)
- Trimming arguments used and why
- Validation of trimming plots
- Length of remaining reads after trimming.
- Maximum average read phred score.
- IDs of the reads with the maximum average phred score.
- IDs of the 10 shortest reads.
- Percentage and Number of reads mapped.
- Number of reads without a pair complement.
- Use the CIGAR string, to compute the number of reads without any Insertion or Deletion.
- Average Mapping score/quality for the mapped reads.
- Number of Duplicates.
- Number of supplementary and secondary read.
- Total Number of variants, number of SNPs and INDELs.
  Number of Homozygotes wild type, Heterozygotes, Homozygotes mutant.