FILTERING BASED ON QUALITY PARAMETERS, TRIMMING, DOWNLOADING REFERENCE GENOME AND ANNOTATION FILE, MAPPING RNA-SEQ READS ON THE GENOME.

In this step you will check the presence of primer, adapters and remove bases with low quality scores. Three frequently used tools are Cutadapt, fastp and Trimmomatic.

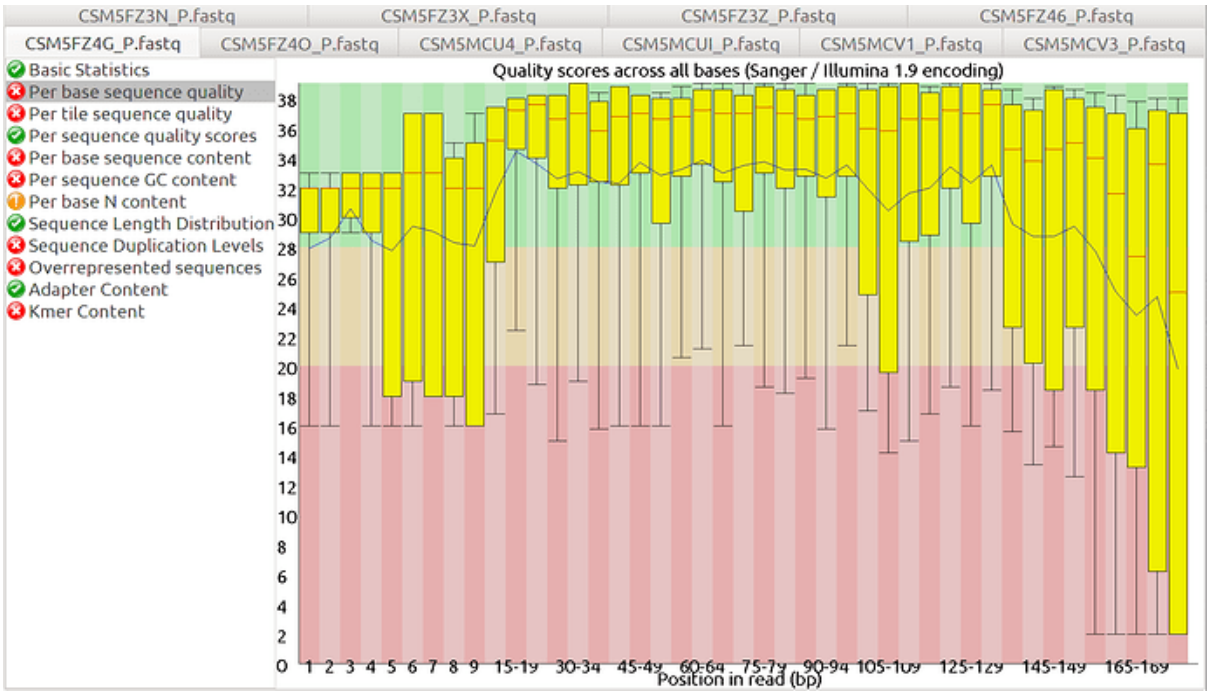
FASTP

https://github.com/OpenGene/fastp
https://academic.oup.com/bioinformatics/article/34/17/i884/5093234

```
# conda create -n fastp
# conda install -c bioconda fastp
```

example
fastp -i file_R1.fastq -I file_R2.fastq -o file_R1.trimmed.fq -O file_R2.trimmed.fq -V --detect_adapter_for_pe --html report.html -f 10 -t 5 -F 10 -T 5 -g -x -M 26 -w 4 -n 3 -l 75 -5 -3

Run FastQC analysis again after trimming and compare the results. Do you see any improvements in your dataset? Describe differences.



How would you trim a file if you see the following result in a fastqc file? Put the command here. It's a single end file.

DOWNLOAD REFERENCE GENOME AND ANNOTATION FILES

Download reference genome and GTF or GFF file, which will be used for creating an index prior to alignment. STAR supports GFF and GTF files. GFF can be converted to GTF with Cufflinks.

Genomic sequence in fasta file and annotated features (GTF or GFF3) can be downloaded from:
https://www.ncbi.nlm.nih.gov/genome
https://ftp.ncbi.nlm.nih.gov/genomes/
https://www.ncbi.nlm.nih.gov/datasets/genomes/

Or from Ensembl database:

https://www.ensembl.org/index.html

Looking for reference files using NCBI dataset

`https://www.ncbi.nlm.nih.gov/datasets/docs/v2/download-and-install/`

`conda install conda-forge::ncbi-datasets-cli`
We will create a for loop to download in one command all the possible accessions.

```
for element in $( cat list_of_accessions.txt); do  datasets download genome accession $element --include genome,rna,protein,cds,gff3,gtf --filename $element.zip; done
```

Questions:

a. Describe the GTF.
b. Examine GTF files. Which information can be found in these files?
c. How many genes are present?
d. Provide me the commands and results for counting the number of sequences in the various fasta files (DNA, RNA, protein).
e. Describe differences between different genomic fasta files.

ALIGNMENT USING HISAT2

```
# conda create -n hisat2
# conda activate hisat2
# conda install -c bioconda hisat2
```

https://www.nature.com/articles/s41587-019-0201-4

Hisat2 is a splice-aware mapper (meaning that it aligns RNA-Seq data on genomes of eukaryotes considering introns). STAR is another splice-aware mapping tool.

Practical work:

1) First step: Index of the genomic sequence should be created.
2) Second step: mapping can be performed.

Index building consists of pre-processing the reference genome so that the program can search in it efficiently. Each program will build a different type of index.

```
# Activate environment
conda activate hisat2

# create an index with hisat2
hisat2-build  reference_genome.fasta genome

# Mapping
hisat2  -x genome -1 SRR7503185_1.trimmed.fastq -2 SRR7503185_2.trimmed.fastq  -p 4
-S SRR7503185.sam
```

How many input reads were mapped uniquely?
How many reads were mapped to multiple loci?

QUANTIFYING THE EXPRESSION OF TRANSCRIPTS USING RNA-SEQ DATA WITH SALMON

Salmon is a tool for quantifying the expression of transcripts using RNA-seq data. Salmon uses new algorithms (specifically, coupling the concept of quasi-mapping with a two-phase inference procedure) to provide accurate expression estimates very quickly (i.e. wicked-fast) and while using little memory.
https://combine-lab.github.io/salmon/
https://translational-medicine.biomedcentral.com/articles/10.1186/s12967-021-02936-w
https://www.rna-seqblog.com/rpkm-fpkm-and-tpm-clearly-explained/

1. Create the index

salmon index -t <cds.fasta> -i <index_dir>

2. Quantify Transcript Abundances

salmon quant -i <index_dir> -l A -1 <reads_R1.fastq> -2 <reads_R2.fastq> -o <output_dir>