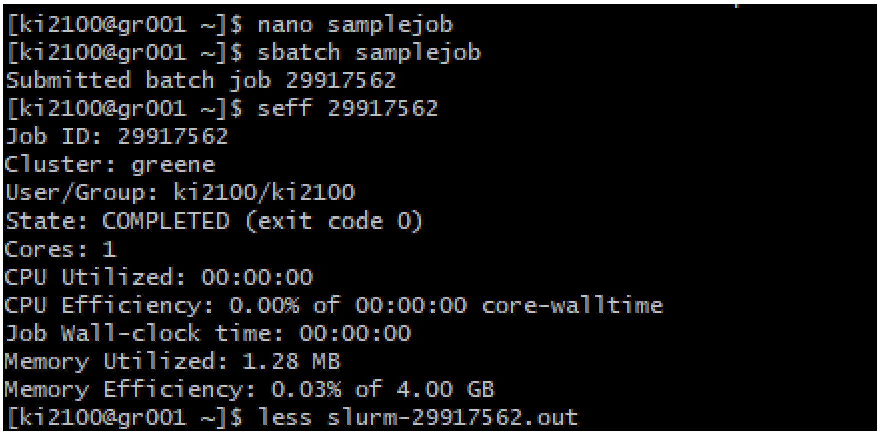
Khan Inan
Transcriptomics 7653
Homework #2

Q1.1 For your answer report the following [1 point].

Q1.1a.Paste the contents of your job script into your homework file. Please use an RMarkdown code block or equivalent (see Week 2 webinar when we will discuss RMarkdown) if possible.



```
[ki2100@gr001 ~]$ nano samplejob
[ki2100@gr001 ~]$ sbatch samplejob
Submitted batch job 29917562
[ki2100@gr001 ~]$ seff 29917562
Job ID: 29917562
Cluster: greene
User/Group: ki2100/ki2100
State: COMPLETED (exit code 0)
Cores: 1
CPU Utilized: 00:00:00
CPU Efficiency: 0.00% of 00:00:00 core-walltime
Job Wall-clock time: 00:00:00
Memory Utilized: 1.28 MB
Memory Efficiency: 0.03% of 4.00 GB
[ki2100@gr001 ~]$ less slurm-29917562.out
```

Q1.1b Report your job id.

My job ID is 29917562

Q1.1c Check the exit status of your job usig the seff command. What is the exit code and what does it mean?

The exit code is 0 and this essentially means that it was ran successfully

Q1.2 A common source of confusion is the distinction between relative and absolute file paths. [1 point] Q1.2a What is the difference?

The difference between the two is that a relative path is the path in relation to the current directory, while the absolute is the path in regards to the root directory

Q1.2b Did you use absolute or relative paths in your fastp in the course directory? Did you use an absolute or relative path to write the processed fastq files?

To read the fastq files and access them I had to use an absolute path. While I used a relative path for the output files to be directly written to the current directory.

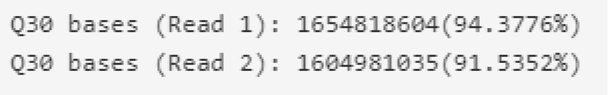
Q1.3 Confirm your job has completed and that you are working interactively at a Greene compute node. You may then answer the following questions using a combination of gunzip and the STDOUT of fastp. [1 point].

Q1.3a What is the name of the STDOUT file for your job?

The name of the STDOUT file for my job is slurm-29917562.out

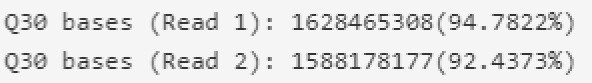
Q1.3b Review the STDOUT file from Q1.3a. What percentage of the bases were Phred quality of Q30 or above in each of the original and processed fastqs?

Before I processed the data



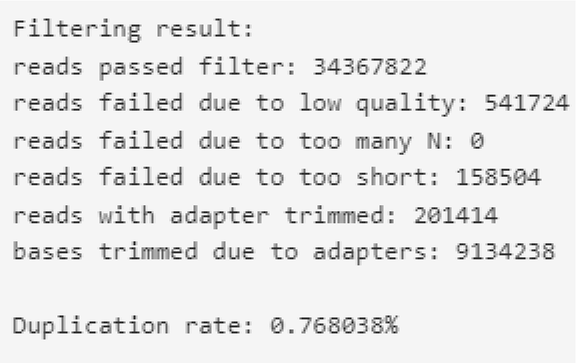
```
Q30 bases (Read 1): 1654818604(94.3776%)
Q30 bases (Read 2): 1604981035(91.5352%)
```

After the processing



```
Q30 bases (Read 1): 1628465308(94.7822%)
Q30 bases (Read 2): 1588178177(92.4373%)
```

Q1.3c Report the “Filtering result” section of the output and the duplication rate. Notice the different filters that are applied by fastp.



```
Filtering result:
reads passed filter: 34367822
reads failed due to low quality: 541724
reads failed due to too many N: 0
reads failed due to too short: 158504
reads with adapter trimmed: 201414
bases trimmed due to adapters: 9134238

Duplication rate: 0.768038%
```

Q1.4. fastp produces an .html report fastp.html by default. Please download it and identify something interesting or unexpected to you and upload along with your homework document [1 point].

The most interesting to me is how the data looks relatively scattered and mixed prior to filtering, but after filtering the base content ration plots were much more consistent and smoothly displayed

Q2.1. Each array index will have its own STDERR and STDOUT which by default are written to a single file with naming convention slurm-<job id>_<index>.out. Please review the contents of the output for index 1 and answer the following [1 point].

Q2.1a Were any adapter sequences detected?

There were no adapter sequences detected

Q2.1b How many reads were in the read 1 set before filtering? Read 2?

There were about 58521629 reads for read 1 and read 2 has 58521629 reads

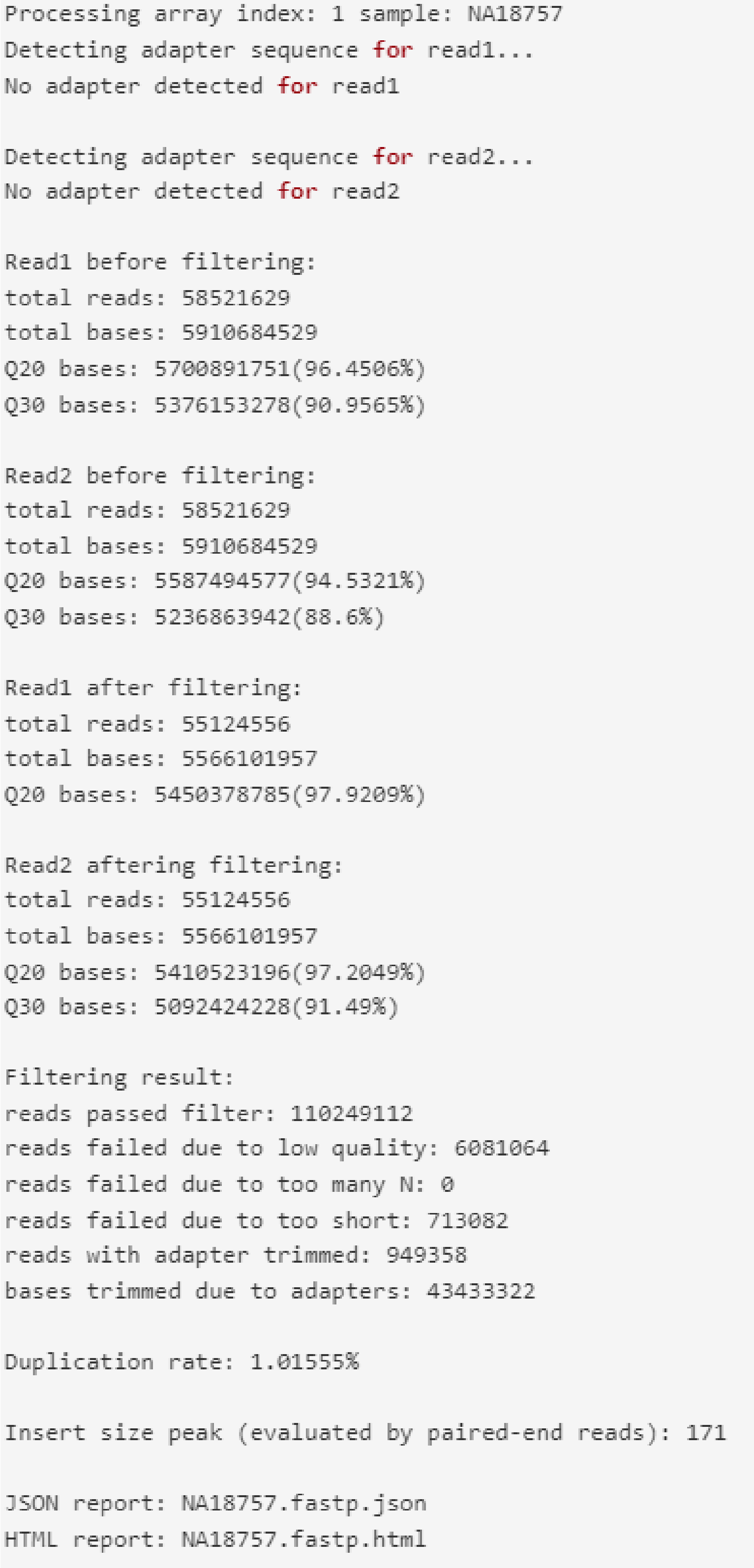
Q2.1c How many reads survived filtering in Read 1 set? Read 2?

55124556 reads survived filtering in both of the reads

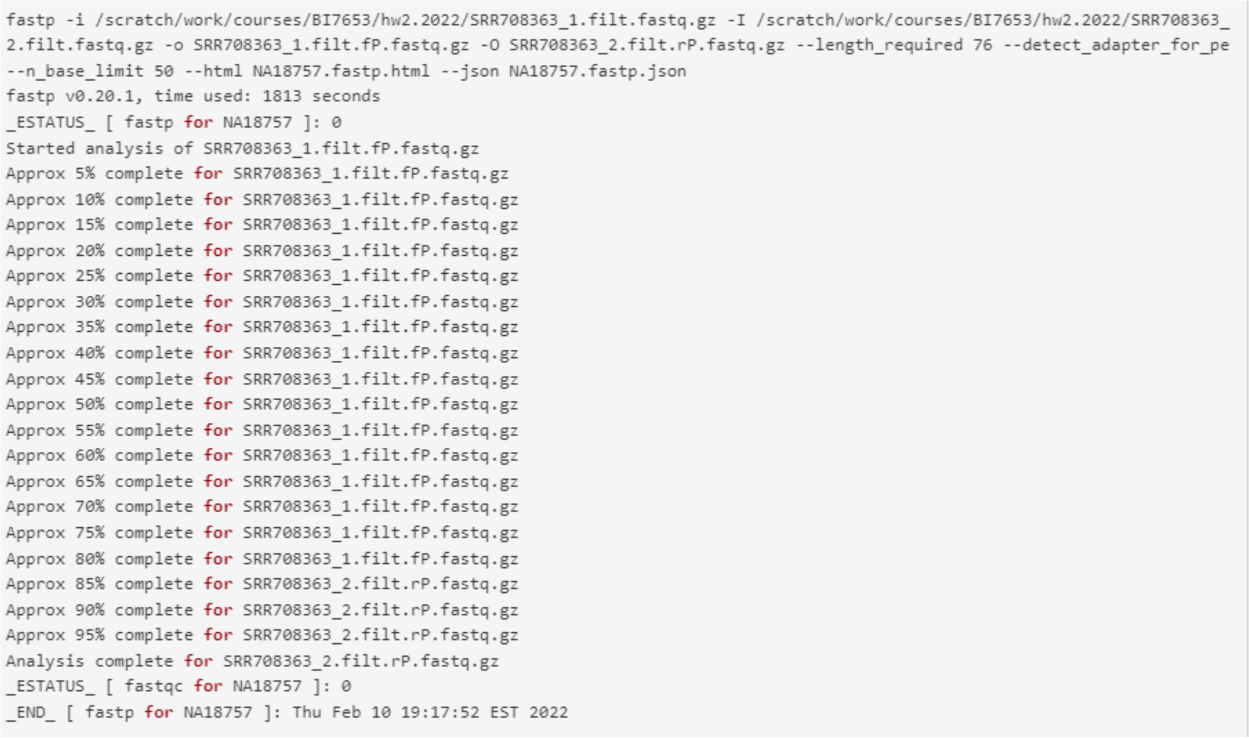
Q2.1d What percentage of reads survived filtering in Read 1 set?

Around 94.1952% survived after filtering the data set

Q2.1e Copy the contents of your output for index 1 of your job array to your homework file



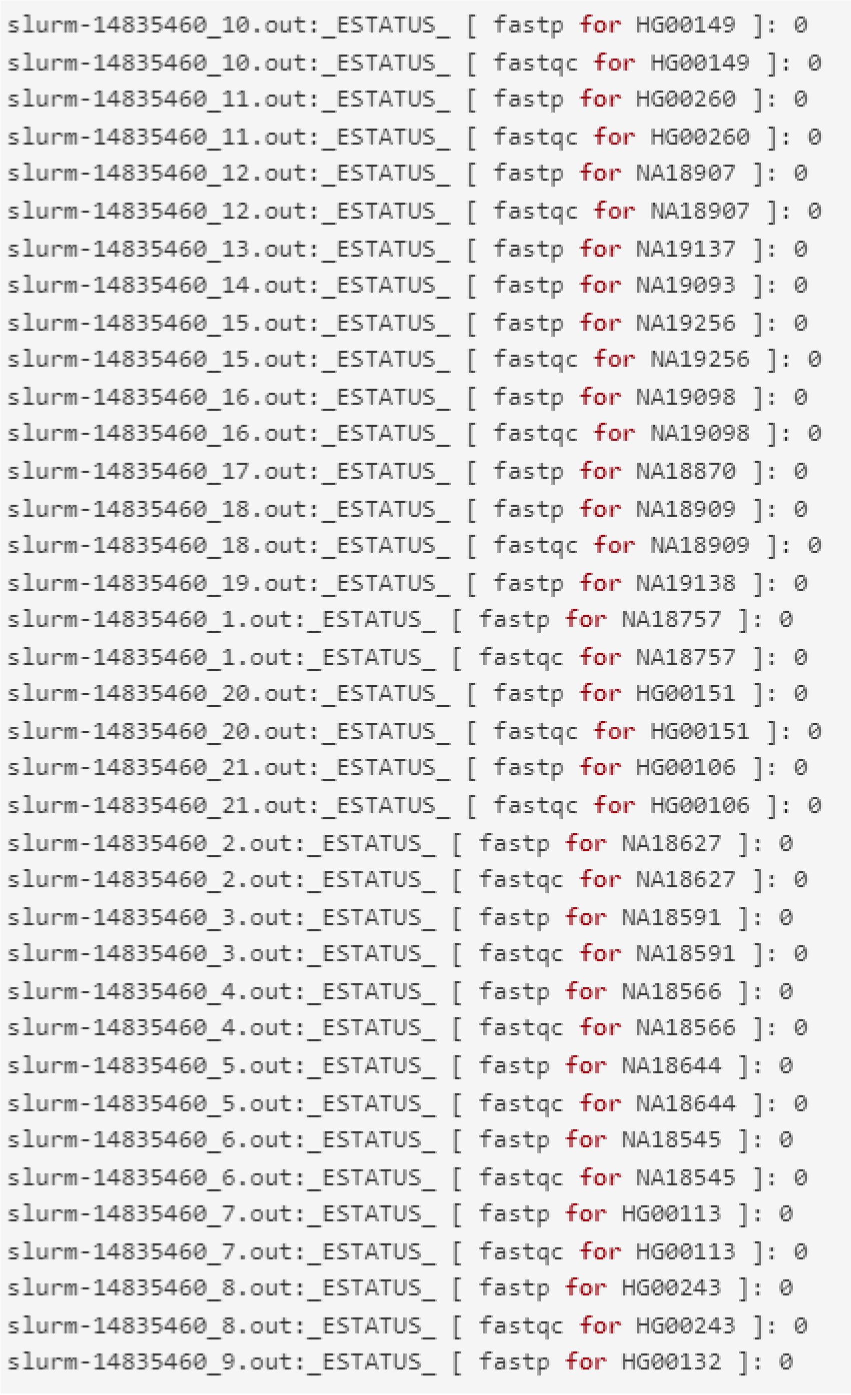
```
Processing array index: 1 sample: NA18757
Detecting adapter sequence for read1...
No adapter detected for read1

Detecting adapter sequence for read2...
No adapter detected for read2

Read1 before filtering:
total reads: 58521629
total bases: 5910684529
Q20 bases: 5700891751(96.4506%)
Q30 bases: 5376153278(90.9565%)

Read2 before filtering:
total reads: 58521629
total bases: 5910684529
Q20 bases: 5587494577(94.5321%)
Q30 bases: 5236863942(88.6%)

Read1 after filtering:
total reads: 55124556
total bases: 5566101957
Q20 bases: 5450378785(97.9209%)

Read2 aftering filtering:
total reads: 55124556
total bases: 5566101957
Q20 bases: 5410523196(97.2049%)
Q30 bases: 5092424228(91.49%)

Filtering result:
reads passed filter: 110249112
reads failed due to low quality: 6081064
reads failed due to too many N: 0
reads failed due to too short: 713082
reads with adapter trimmed: 949358
bases trimmed due to adapters: 43433322

Duplication rate: 1.01555%

Insert size peak (evaluated by paired-end reads): 171

JSON report: NA18757.fastp.json
HTML report: NA18757.fastp.html
```



```
fastp -i /scratch/work/courses/BI7653/hw2.2022/SRR708363_1.filt.fastq.gz -I /scratch/work/courses/BI7653/hw2.2022/SRR708363_
2.filt.fastq.gz -o SRR708363_1.filt.fP.fastq.gz -O SRR708363_2.filt.rP.fastq.gz --length_required 76 --detect_adapter_for_pe
--n_base_limit 50 --html NA18757.fastp.html --json NA18757.fastp.json
fastp v0.20.1, time used: 1813 seconds
_ESTATUS_ [ fastp for NA18757 ]: 0
Started analysis of SRR708363_1.filt.fP.fastq.gz
Approx 5% complete for SRR708363_1.filt.fP.fastq.gz
Approx 10% complete for SRR708363_1.filt.fP.fastq.gz
Approx 15% complete for SRR708363_1.filt.fP.fastq.gz
Approx 20% complete for SRR708363_1.filt.fP.fastq.gz
Approx 25% complete for SRR708363_1.filt.fP.fastq.gz
Approx 30% complete for SRR708363_1.filt.fP.fastq.gz
Approx 35% complete for SRR708363_1.filt.fP.fastq.gz
Approx 40% complete for SRR708363_1.filt.fP.fastq.gz
Approx 45% complete for SRR708363_1.filt.fP.fastq.gz
Approx 50% complete for SRR708363_1.filt.fP.fastq.gz
Approx 55% complete for SRR708363_1.filt.fP.fastq.gz
Approx 60% complete for SRR708363_1.filt.fP.fastq.gz
Approx 65% complete for SRR708363_1.filt.fP.fastq.gz
Approx 70% complete for SRR708363_1.filt.fP.fastq.gz
Approx 75% complete for SRR708363_1.filt.fP.fastq.gz
Approx 80% complete for SRR708363_1.filt.fP.fastq.gz
Approx 85% complete for SRR708363_2.filt.rP.fastq.gz
Approx 90% complete for SRR708363_2.filt.rP.fastq.gz
Approx 95% complete for SRR708363_2.filt.rP.fastq.gz
Analysis complete for SRR708363_2.filt.rP.fastq.gz
_ESTATUS_ [ fastqc for NA18757 ]: 0
_END_ [ fastp for NA18757 ]: Thu Feb 10 19:17:52 EST 2022
```

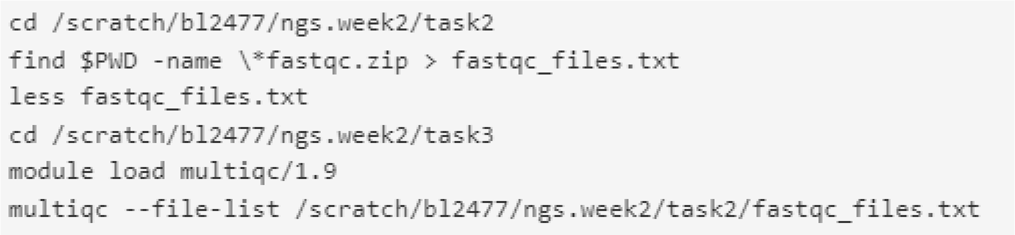
Q2.2 Did all commands have an exit status of zero? Please copy the result of the grep command into your homework document [1 point].

Yes they did as is shown down below



```
slurm-14835460_10.out:_ESTATUS_ [ fastp for HG00149 ]: 0
slurm-14835460_10.out:_ESTATUS_ [ fastqc for HG00149 ]: 0
slurm-14835460_11.out:_ESTATUS_ [ fastp for HG00260 ]: 0
slurm-14835460_11.out:_ESTATUS_ [ fastqc for HG00260 ]: 0
slurm-14835460_12.out:_ESTATUS_ [ fastp for NA18907 ]: 0
slurm-14835460_12.out:_ESTATUS_ [ fastqc for NA18907 ]: 0
slurm-14835460_13.out:_ESTATUS_ [ fastp for NA19137 ]: 0
slurm-14835460_14.out:_ESTATUS_ [ fastp for NA19093 ]: 0
slurm-14835460_15.out:_ESTATUS_ [ fastp for NA19256 ]: 0
slurm-14835460_15.out:_ESTATUS_ [ fastqc for NA19256 ]: 0
slurm-14835460_16.out:_ESTATUS_ [ fastp for NA19098 ]: 0
slurm-14835460_16.out:_ESTATUS_ [ fastqc for NA19098 ]: 0
slurm-14835460_17.out:_ESTATUS_ [ fastp for NA18870 ]: 0
slurm-14835460_18.out:_ESTATUS_ [ fastp for NA18909 ]: 0
slurm-14835460_18.out:_ESTATUS_ [ fastqc for NA18909 ]: 0
slurm-14835460_19.out:_ESTATUS_ [ fastp for NA19138 ]: 0
slurm-14835460_1.out:_ESTATUS_ [ fastp for NA18757 ]: 0
slurm-14835460_1.out:_ESTATUS_ [ fastqc for NA18757 ]: 0
slurm-14835460_20.out:_ESTATUS_ [ fastp for HG00151 ]: 0
slurm-14835460_20.out:_ESTATUS_ [ fastqc for HG00151 ]: 0
slurm-14835460_21.out:_ESTATUS_ [ fastp for HG00106 ]: 0
slurm-14835460_21.out:_ESTATUS_ [ fastqc for HG00106 ]: 0
slurm-14835460_2.out:_ESTATUS_ [ fastp for NA18627 ]: 0
slurm-14835460_2.out:_ESTATUS_ [ fastqc for NA18627 ]: 0
slurm-14835460_3.out:_ESTATUS_ [ fastp for NA18591 ]: 0
slurm-14835460_3.out:_ESTATUS_ [ fastqc for NA18591 ]: 0
slurm-14835460_4.out:_ESTATUS_ [ fastp for NA18566 ]: 0
slurm-14835460_4.out:_ESTATUS_ [ fastqc for NA18566 ]: 0
slurm-14835460_5.out:_ESTATUS_ [ fastp for NA18644 ]: 0
slurm-14835460_5.out:_ESTATUS_ [ fastqc for NA18644 ]: 0
slurm-14835460_6.out:_ESTATUS_ [ fastp for NA18545 ]: 0
slurm-14835460_6.out:_ESTATUS_ [ fastqc for NA18545 ]: 0
slurm-14835460_7.out:_ESTATUS_ [ fastp for HG00113 ]: 0
slurm-14835460_7.out:_ESTATUS_ [ fastqc for HG00113 ]: 0
slurm-14835460_8.out:_ESTATUS_ [ fastp for HG00243 ]: 0
slurm-14835460_8.out:_ESTATUS_ [ fastqc for HG00243 ]: 0
slurm-14835460_9.out:_ESTATUS_ [ fastp for HG00132 ]: 0
```

Q3.1 Report your multiqc command [1 point].



```
cd /scratch/bl2477/ngs.week2/task2
find $PWD -name \*fastqc.zip > fastqc_files.txt
less fastqc_files.txt
cd /scratch/bl2477/ngs.week2/task3
module load multiqc/1.9
multiqc --file-list /scratch/bl2477/ngs.week2/task2/fastqc_files.txt
```

Q3.2. Download the MultiQC output (multiqc_report.html by default) to your personal computer, open the report in your browser, and answer the following questions.

You can review the following short tutorial on working with MultiQC reports https://www.youtube.com/watch?v=qPbIlO_KWN0. You can hover your arrow over features in the interactive .html report to determine information like the fastq file that is represented by the feature.

Q3.2a Which fastq file has the greatest decline in base quality with increasing sequencing cycle ("the dephasing problem")? [1 point]

The file that is labeled ERR252551_1.fP has the greatest decline in the quality along with the cycle

Q3.2b Two samples (four fastqs) appear to have unusually high GC content and unusually high duplication levels? Which samples are they? [1 point]

The samples with high GC content and duplication levels seem to be SRR766045 and SRR702073

Q3.2c Was there any residual adapter contamination in any fastq file after processing with reads with fastp [1 point]?

There wasn't any residual adapter that was contaminating the fastq files. The multiQC report states that there were no samples found with contamination greater than 0.1%