APPLIED SEQUENCING INFORMATICS-Assignment-2

Step1- Download the Dataset(s):

Code
:#Run an interactive session and make a directory called Assignment2 and submit the bash script shown.

```
[as18818@bigpurple-ln2 ~]$ srun --mem=16gb --cpus-per-task=1 --time=4:00:00 --pty /bin/bash
[as18818@cn-0012 as18818]$ mkdir Assignment2
[as18818@cn-0012 as18818]$ cd Assignment2
[as18818@cn-0012Assignment2]$ emacs Assignment2
```

BASH SCRIPT

```
#!/bin/bash
#SBATCH --job-name=Assignment2 # Job name
#SBATCH --mail-type=END,FAIL # Mail events (NONE, BEGIN, END, FAIL, ALL)
#SBATCH --mail-user=Apoorva.Sharma@nyulangone.org # Where to send mail
#SBATCH --ntasks=1 # Run on a single node
#SBATCH --mem=8gb # Job memory request
#SBATCH --time=04:00:00 # Time limit hrs:min:sec
#SBATCH --output=Assignment2_%j.log # Standard output and error log
#SBATCH -p cpu_short # Specifies location to submit job

## [Assignment 2 ] ###
#Load sratoolkit module
module load sratoolkit/2.9.1

## Run fastq-dump to download a specified file from the SRA, split into R1 & R2 file (if appropriate), and compress with gzip. Note --origfmt restores fastq header sequences to original form ##

fastq-dump --split-files SRR1523657 --gzip -O /gpfs/scratch/as18818/Assignment2
--origfmt

#for SRX4037309
fastq-dump --split-files SRR7109502 --gzip -O /gpfs/scratch/as18818/Assignment2
--origfmt

#for SRX4146457(2 files)
fastq-dump --split-files SRR7240634 --gzip -O /gpfs/scratch/as18818/Assignment2
--origfmt
fastq-dump --split-files SRR7240635 --gzip -O /gpfs/scratch/as18818/Assignment2
--origfmt

# remove temporary directory that is utilised by fastq-dump
rm -r ~/ncbi
#######################
[as18818@cn-0012 Assignment2]$ sbatch Assignment2
```

2) What sequencing methodologies were employed? Single or paired-end?
SRR1523657: RNA-seq, Illumina HiSeq 2500, Paired end Sequencing
SRX4037309: RNA-seq, NextSeq 500, Paired end Sequencing
SRX4146457: RNA-seq, Illumina HiSeq 2000, Paired end Sequencing

3) Examine datasets with FASTQC

Code:

```
[as18818@cn-0013 Assignment2]$ emacs FASTQC.sh
#!/bin/bash
#SBATCH --job-name=Exercise2 # Job name
#SBATCH --mail-type=END,FAIL # Mail events (NONE, BEGIN, END, FAIL, ALL)
#SBATCH --mail-user=Assignment2@nyulangone.org # Where to send mail
#SBATCH --ntasks=1 # Run on a single node
#SBATCH --mem=8gb # Job memory request
#SBATCH --time=03:00:00 # Time limit hrs:min:sec
#SBATCH --output=Assignment2fq_%j.log # Standard output and error log
#SBATCH -p cpu_short # Specifies location to submit job
#Load fastqc module
module load fastqc/0.11.7
#run fastqc on test files
fastqc -o /gpfs/scratch/as18818/Assignment2
/gpfs/scratch/as18818/Assignment2/SRR1523657_1.fastq.gz
/gpfs/scratch/as18818/Assignment2/SRR1523657_2.fastq.gz

fastqc -o /gpfs/scratch/as18818/Assignment2
/gpfs/scratch/as18818/Assignment2/SRR7109502_1.fastq.gz
/gpfs/scratch/as18818/Assignment2/SRR7109502_2.fastq.gz
fastqc -o /gpfs/scratch/as18818/Assignment2
/gpfs/scratch/as18818/Assignment2/SRR7240634_1.fastq.gz
/gpfs/scratch/as18818/Assignment2/SRR7240634_2.fastq.gz
fastqc -o /gpfs/scratch/as18818/Assignment2
/gpfs/scratch/as18818/Assignment2/SRR7240635_1.fastq.gz
/gpfs/scratch/as18818/Assignment2/SRR7240635_2.fastq.gz
######################
[as18818@cn-0013 Assignment2]$ sbatch FASTQC.sh
```

ANALYSIS OF FASTQC:(troublesome parameters are color coded.)

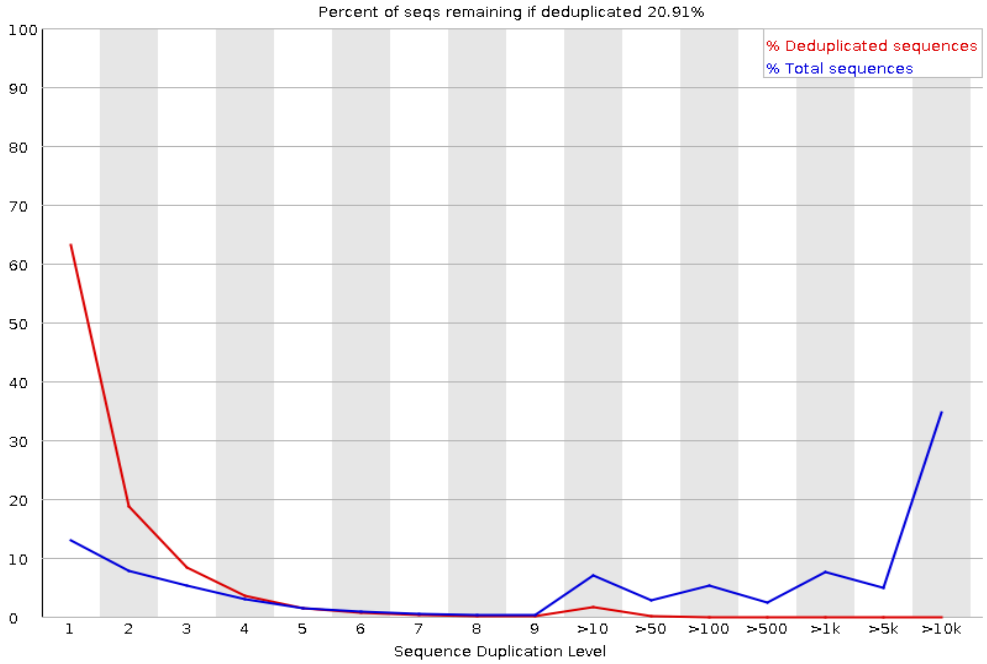
	Read	FastQC Report Summary: Warning and Failing Items
SRR1523657	Pair 1	per base sequence content, per sequence GC content, sequence duplication levels, overrepresented sequences, adapter content, Per tile sequence quality
	Pair 2	per sequence GC content, sequence duplication levels, adapter content, Per base sequence content, Overrepresented sequences, Per tile sequence quality
SRR7109502	Pair 1	sequence duplication levels, overrepresented sequences, Per base sequence content, Per sequence GC content
	Pair 2	sequence duplication levels, overrepresented sequences, Per base sequence content
SRR7240634	Pair 1	per sequence GC content, sequence duplication levels
	Pair 2	sequence duplication levels, Per base sequence content
SRR7240635	Pair 1	per sequence GC content, sequence duplication levels
	Pair 2	sequence duplication levels

Certainly, these datasets require quality assurance measures. Particularly, SRR1523657 stands out for needing both adaptor trimming and quality adjustment. Ideally, adaptor content should be absent throughout all read positions. However, it's observed that the presence of the Illumina Universal Adapter increases towards the ends of the reads, as shown in both read 1 and read 2. Despite the relatively high "per base sequence quality" throughout the reads (with a Phred score > 25 for all reads, both forward and reverse), the quality peaks in the middle and diminishes towards the ends, especially noticeable in read 2. Since the Phred score reflects the accuracy of base identification, implementing a stricter quality control filter (e.g., Phred score > 30) could enhance our confidence in base calls across the dataset.

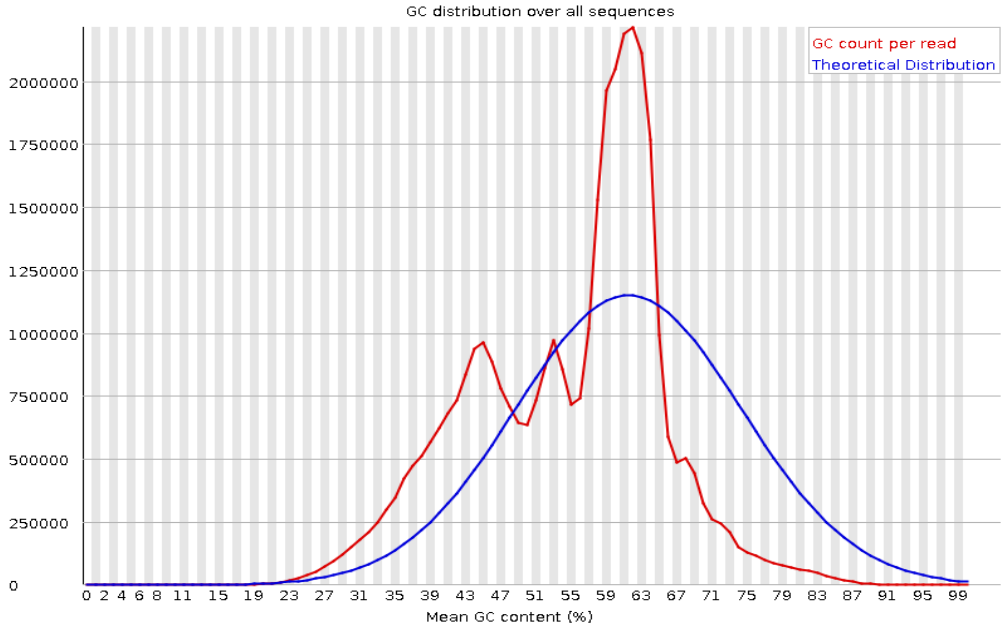
Dataset SRR1523657 requires the most attention in quality control. In contrast, in the other datasets, adaptor content is absent, and the base sequence read quality is already consistently high across all positions. All the datasets also represent a high percentage of duplicate reads.

SRR1523657_Read 1

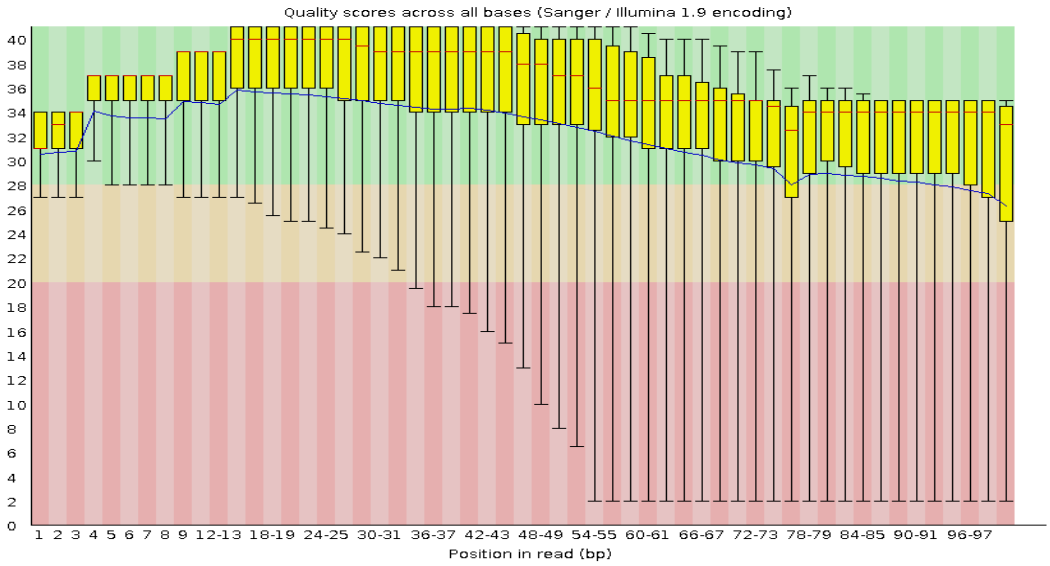
Sequence Duplication levels



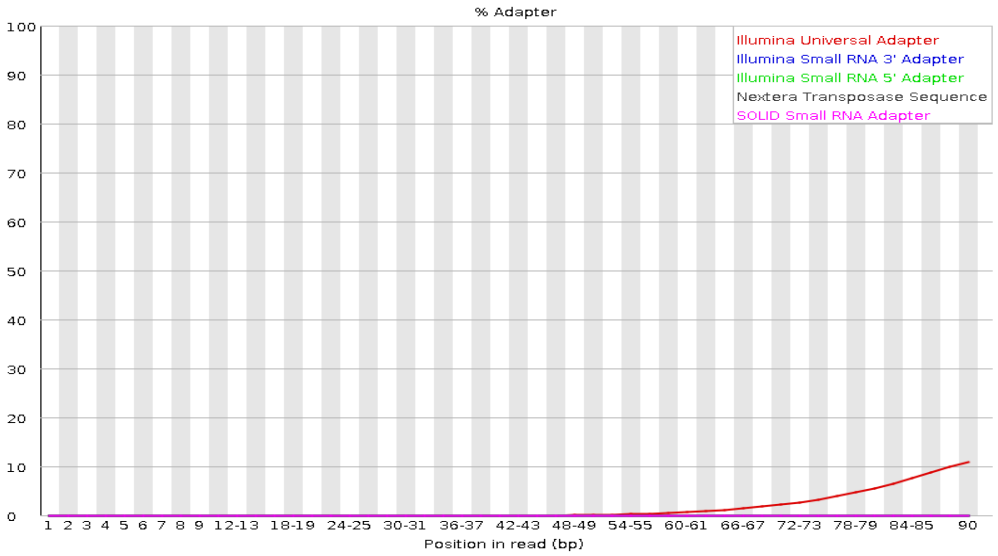
Per sequence GC content:



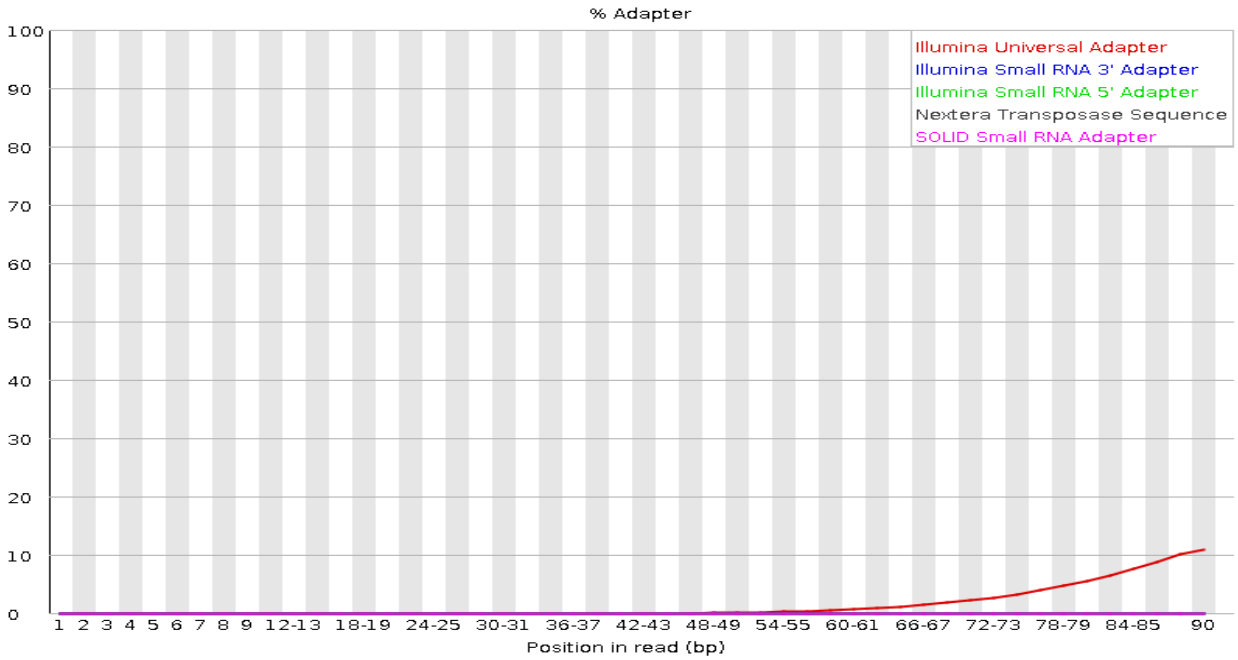
Per base Sequence Quality:

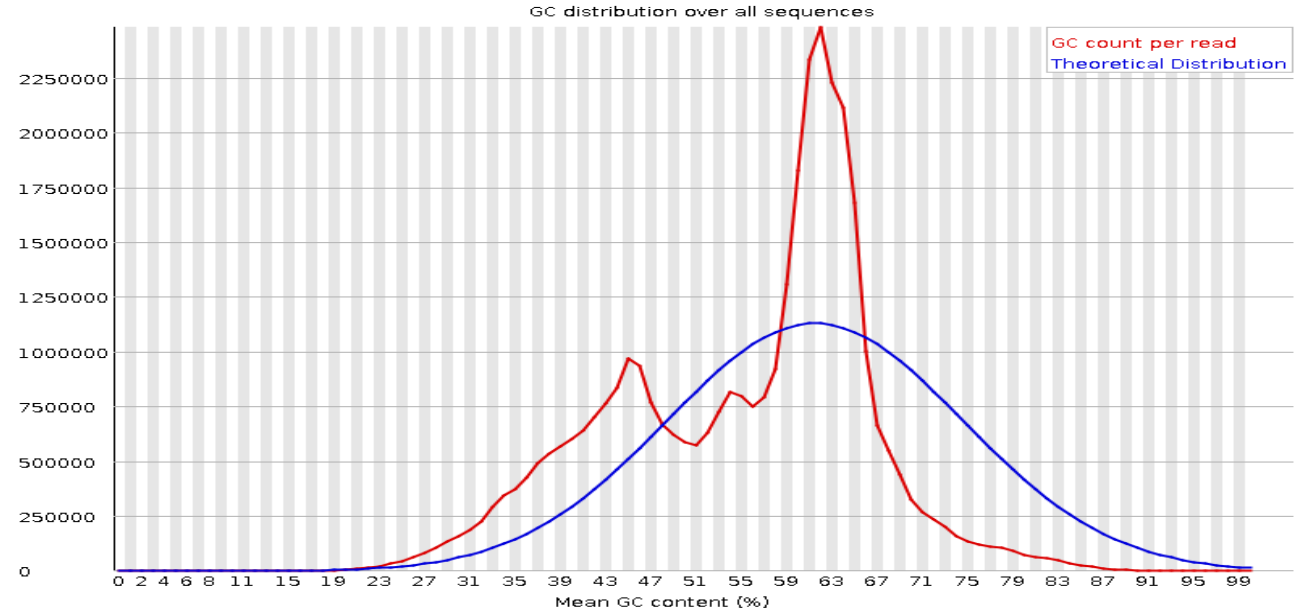


Adapter content:



SRR1523657_Read-2
Adapter Content:



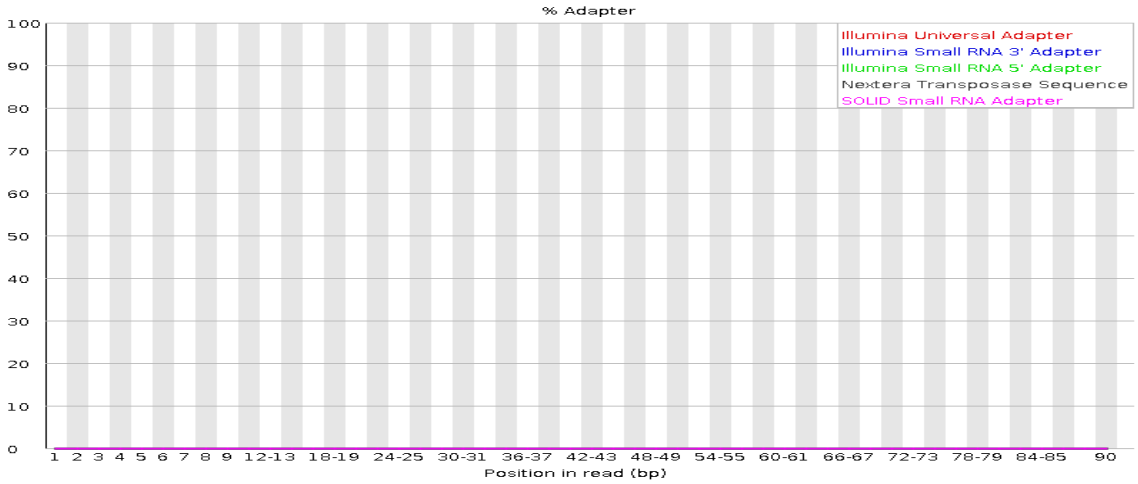


4) Trim data with Trim Galore AND Trimmomatic
Code: Trim Galore
Making a directory called trim in Assignment2

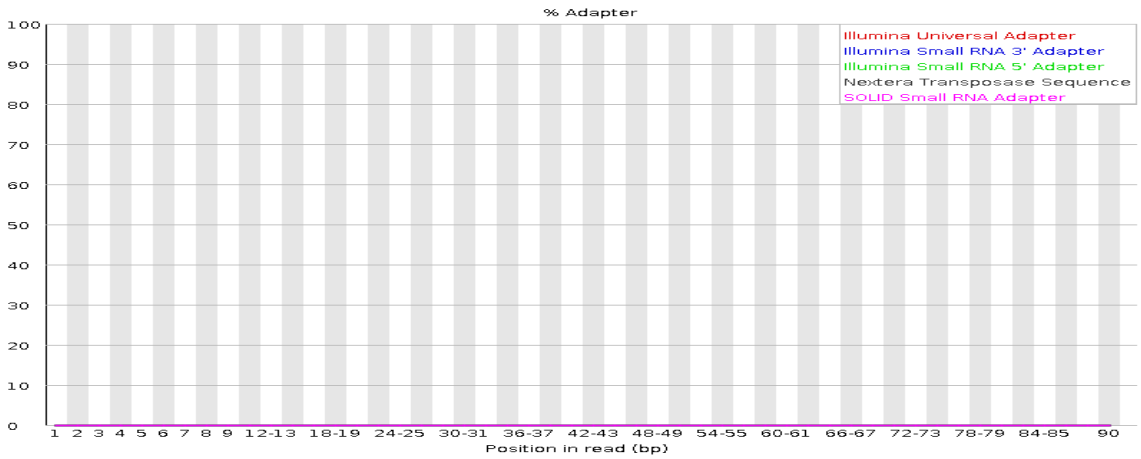
```
#!/bin/bash
#SBATCH --job-name=Assignment2Trim # Job name
#SBATCH --mail-type=END,FAIL # Mail events (NONE, BEGIN, END, FAIL, ALL)
#SBATCH --mail-user=Apoorva.Sharma@nyulangone.org # Where to send mail
#SBATCH --ntasks=1 # Run on a single node
#SBATCH --mem=8gb # Job memory request
#SBATCH --time=03:00:00 # Time limit hrs:min:sec
#SBATCH --output=TrimGalore_%j.log # Standard output and error log
#SBATCH -p cpu_short # Specifies location to submit job
#load modules
module load trimgalore/0.5.0 # Load trimgalore module
module load python/cpu/2.7.15-ES # TrimGalore requires a python module called CutAdapt that is loaded as  part of a the python module 'python/cpu/2.7.15-ES'
module load fastqc/0.11.7
## Run TrimGalore on a paired-end dataset (standard parameters except for raising Q value from 20 to 30) and  request a new fastqc analysis
trim_galore --paired -o /gpfs/scratch/as18818/Assignment2/trim
/gpfs/scratch/as18818/Assignment2/SRR1523657_1.fastq.gz
/gpfs/scratch/as18818/Assignment2/SRR1523657_2.fastq.gz --fastqc --fastqc_args "-o
/gpfs/scratch/as18818/Assignment2/trim" --q 30 --gzip
trim_galore --paired -o /gpfs/scratch/as18818/Assignment2/trim
/gpfs/scratch/as18818/Assignment2/SRR7109502_1.fastq.gz
/gpfs/scratch/as18818/Assignment2/SRR7109502_2.fastq.gz --fastqc --fastqc_args "-o
/gpfs/scratch/as18818/Assignment2/trim" --q 30 --gzip
trim_galore --paired -o /gpfs/scratch/as18818/Assignment2/trim
/gpfs/scratch/as18818/Assignment2/SRR7240634_1.fastq.gz
/gpfs/scratch/as18818/Assignment2/SRR7240634_2.fastq.gz --fastqc --fastqc_args "-o
/gpfs/scratch/as18818/Assignment2/trim" --q 30 --gzip
trim_galore --paired -o /gpfs/scratch/as18818/Assignment2/trim
/gpfs/scratch/as18818/Assignment2/SRR7240635_1.fastq.gz
/gpfs/scratch/as18818/Assignment2/SRR7240635_2.fastq.gz --fastqc --fastqc_args "-o
/gpfs/scratch/as18818/Assignment2/trim" --q 30 --gzip
######################
[as18818@cn-002 trim]$ ls
SRR1523657_1.fastq.gz_trimming_report.txt
SRR1523657_1_val_1.fq.gz
SRR1523657_2.fastq.gz_trimming_report.txt
SRR1523657_2_val_2.fq.gz
SRR1523671_1_val_1_fastqc.html
SRR1523671_2_val_2_fastqc.html
SRR7109502_1.fastq.gz_trimming_report.txt
SRR7109502_1_val_1_fastqc.html
SRR7109502_1_val_1.fq.gz
SRR7109502_2.fastq.gz_trimming_report.txt
SRR7109502_2_val_2_fastqc.html
SRR7109502_2_val_2.fq.gz
```

SRR7240634_1.fastq.gz_trimming_report.txt
SRR7240634_1_val_1.fq.gz
SRR7240634_2.fastq.gz_trimming_report.txt
SRR7240634_2_val_2.fq.gz
SRR7240635_1.fastq.gz_trimming_report.txt
SRR7240635_1_val_1.fq.gz
SRR7240635_2.fastq.gz_trimming_report.txt
SRR7240635_2_val_2.fq.gz

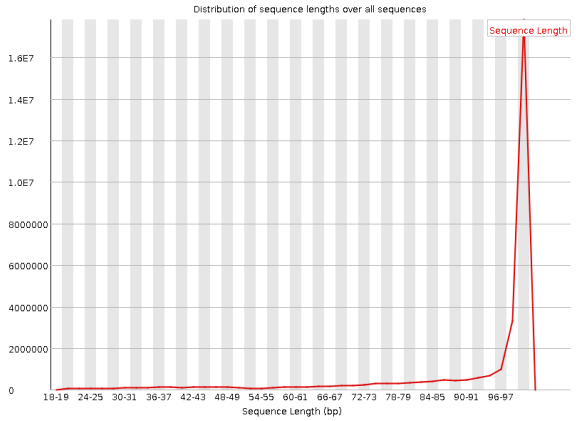
Following the execution of Trim Galore, the adapter content has been successfully removed (adapter % = 0). As expected, the application of Trim Galore had the most significant effect on the paired reads of SRR1532657 compared to the other datasets that were downloaded.

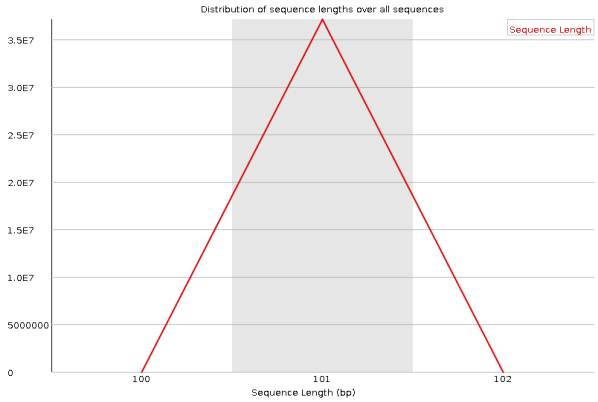


For read2



Though the sequence duplication rate in read 2 has changed significantly:





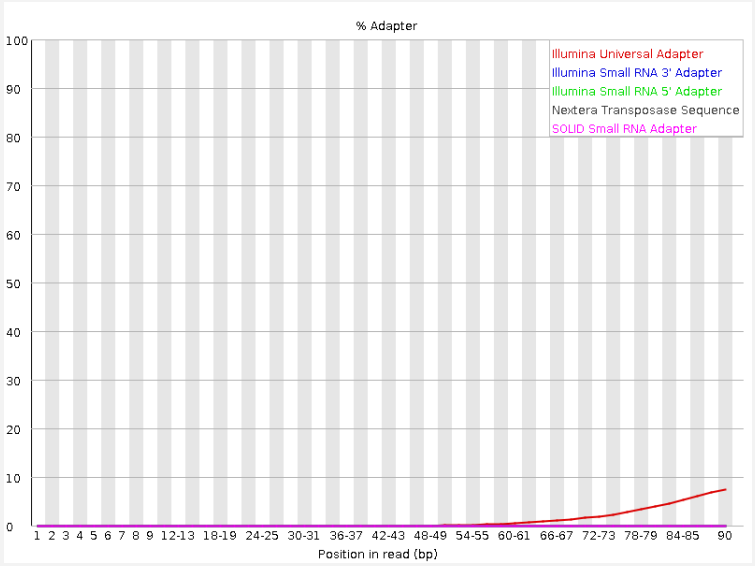
Now Using Trimmomatic:

Code:

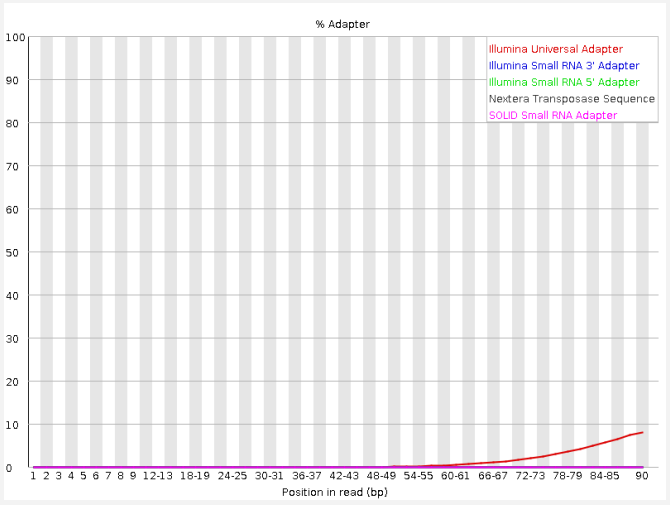
```
#!/bin/bash
#SBATCH --job-name=Assignment2Trimmomatic # Job name
#SBATCH --mail-type=END,FAIL # Mail events (NONE, BEGIN, END, FAIL, ALL)
#SBATCH --mail-user=email@nyulangone.org # Where to send mail
#SBATCH --ntasks=1 # Run on a single node
#SBATCH --mem=8gb # Job memory request
#SBATCH --time=03:00:00 # Time limit hrs:min:sec
#SBATCH --output=A2_Trimmomatic_%j.log # Standard output and error log
#SBATCH -p cpu_short # Specifies location to submit job
#load modules
module load trimmomatic/0.36
java -jar /gpfs/share/apps/trimmomatic/0.36/trimmomatic-0.36.jar PE SRR1523657_1.fastq.gz
SRR1523657_2.fastq.gz SRR1523657_1.trimmed.fastq.gz SRR152367_1un.trimmed.fastq.gz
SRR1523657_2.trimmed.fastq.gz SRR1523657_2un.trimmed.fastq.gz
ILLUMINACLIP:contaminant.fa:2:30:10  SLIDINGWINDOW:4:30
java -jar /gpfs/share/apps/trimmomatic/0.36/trimmomatic-0.36.jar PE SRR7109502_1.fastq.gz
SRR7109502_2.fastq.gz SRR7109502_1.trimmed.fastq.gz SRR7109502_1un.trimmed.fastq.gz
SRR7109502_2.trimmed.fastq.gz SRR7109502_2un.trimmed.fastq.gz
ILLUMINACLIP:contaminant.fa:2:30:10  SLIDINGWINDOW:4:30
java -jar /gpfs/share/apps/trimmomatic/0.36/trimmomatic-0.36.jar PE SRR7240634_1.fastq.gz
SRR7240634_2.fastq.gz SRR7240634_1.trimmed.fastq.gz SRR7240634_1un.trimmed.fastq.gz
SRR7240634_2.trimmed.fastq.gz SRR7240634_2un.trimmed.fastq.gz
ILLUMINACLIP:contaminant.fa:2:30:10  SLIDINGWINDOW:4:30
java -jar /gpfs/share/apps/trimmomatic/0.36/trimmomatic-0.36.jar PE SRR7240635_1.fastq.gz
SRR7240635_2.fastq.gz SRR7240635_1.trimmed.fastq.gz SRR7240635_1un.trimmed.fastq.gz
SRR7240635_2.trimmed.fastq.gz SRR7240635_2un.trimmed.fastq.gz
ILLUMINACLIP:contaminant.fa:2:30:10  SLIDINGWINDOW:4:30

######################
```

After running Trimmomatic, the presence of adapter sequence has decreased to less than 10%, yet there remains some contamination among the SRR1523657 R_1 & R_2 datasets. Phred scores have been capped at >30 as instructed. Additionally, as part of the Trimmomatic process, reads were discarded if their partner paired read failed to meet the QC parameters. For instance, if both paired reads didn't boast Phred scores >30, they were both discarded, irrespective of whether the forward or reverse paired read met the filter.

Read1



Read2



Log excerpt:
TrimmomaticPE: Started with arguments:
SRR1523657_1.fastq.gz SRR1523657_2.fastq.gz SRR1523657_1.trimmed.fastq.gz SRR152367_1un.trimmed.fastq.gz SRR1523657_2.trimmed.fastq.gz SRR1523657_2un.trimmed.fastq.gz ILLUMINACLIP:contaminant.fa:2:30:10 SLIDINGWINDOW:4:20
Using Short Clipping Sequence: ''
Skipping duplicate Clipping Sequence: ''
Skipping duplicate Clipping Sequence: ''
Skipping duplicate Clipping Sequence: ''
Skipping duplicate Clipping Sequence: ''
ILLUMINACLIP: Using 0 prefix pairs, 1 forward/reverse sequences, 0 forward only sequences, 0 reverse only sequences
Quality encoding detected as phred33
Input Read Pairs: 37124453 Both Surviving: 28558308 (76.93%) Forward Only Surviving: 2811755 (7.57%) Reverse Only Surviving: 1171016 (3.15%) Dropped: 4583374 (12.35%)
TrimmomaticPE: Completed successfully

COMPARING BOTH THE TRIMMING SOFTWARE:
Let's talk about the face-off: Trim Galore vs Trimmomatic. In general, the superiority of read trimming software likely hinges on the nature of your data – In this scenario, particularly with the SRR152357 data, Trim Galore emerges as the winner over Trimmomatic. Why? Well, Trim Galore excelled in adapter trimming. It's a wrapped software utilizing CutAdapt, equipped with the ability to automatically detect standard adapters like Illumina. Moreover, it can handle manual adapter specifications. Trimmomatic, on the other hand, necessitates a fasta file containing the specific adapter sequences, and it didn't completely eradicate all adapter content.

Trimmomatic and Trim Galore both tools were configured to employ a Phred score cutoff of 30 for base call quality. However, it's worth noting that Trimmomatic discarded reads if one of the pairs failed this parameter, unlike Trim Galore. Hence, Trim Galore could be deemed the stricter program in this regard, potentially resulting in a higher-quality trimmed dataset.

Both low-quality base calls and adapter contamination could impede mapping efficiency. In most datasets here (SRR7109502, SRR7240635), base quality trimming had a more significant impact due to minimal or absent adapter contamination. However, in SRR152357, adapter trimming had a larger impact than base quality trimming, especially considering the absence of very low-quality base calls initially. Generally, the magnitude of impact on overall data quality would depend on various factors such as library preparation, sequencing run conditions, etc. Finally, the decision to utilize these features depends on the initial state of your data and your downstream analysis objectives.