Answers to QAA

Part 1

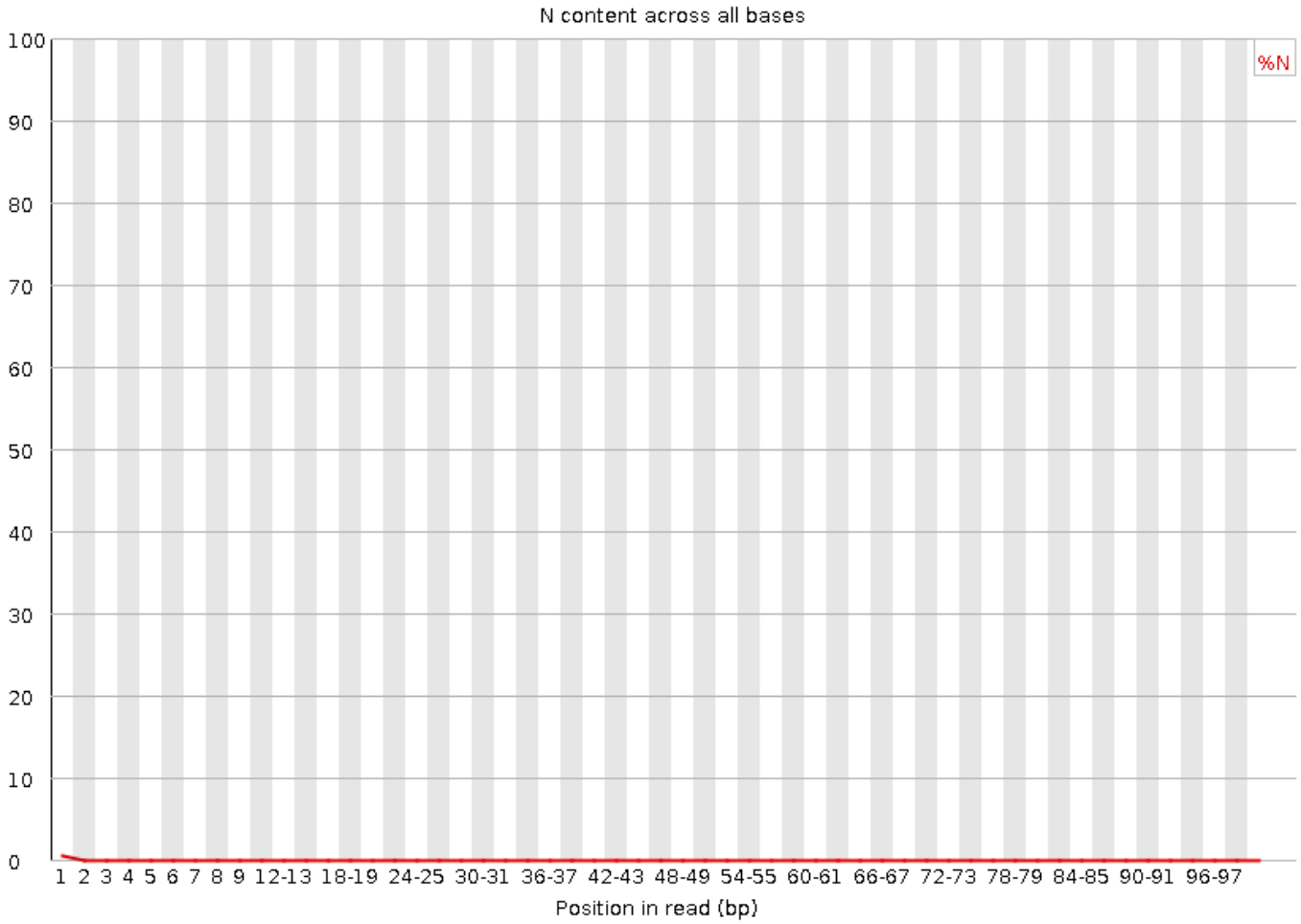
Question 1

I believe that the data looks pretty consistent across the Fastqc. A little bit of Ns at the beginning of the reads. Also, R1 always has a better quality overall then R2 which makes sense.

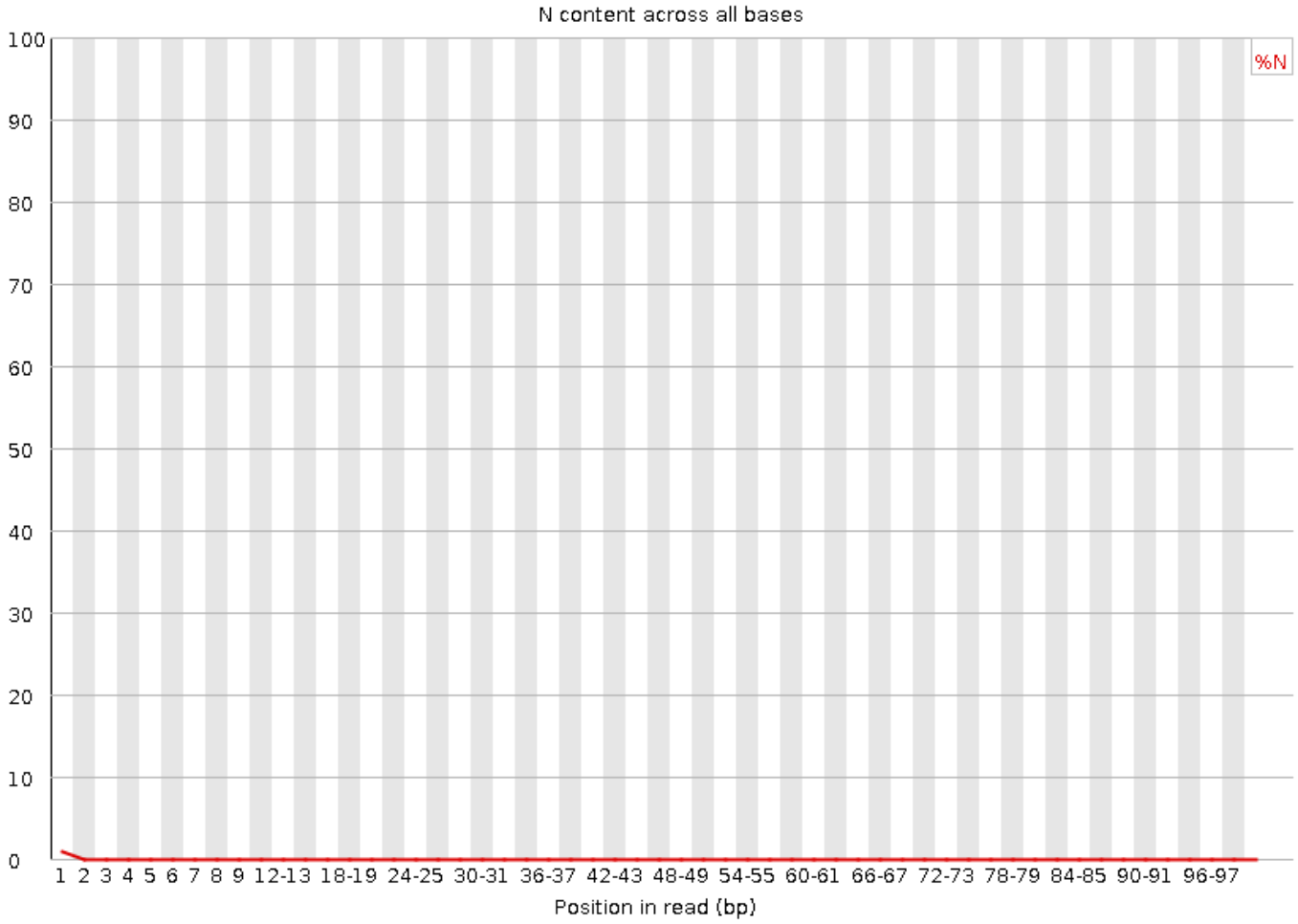
Fastqc Data

Per Base N Content

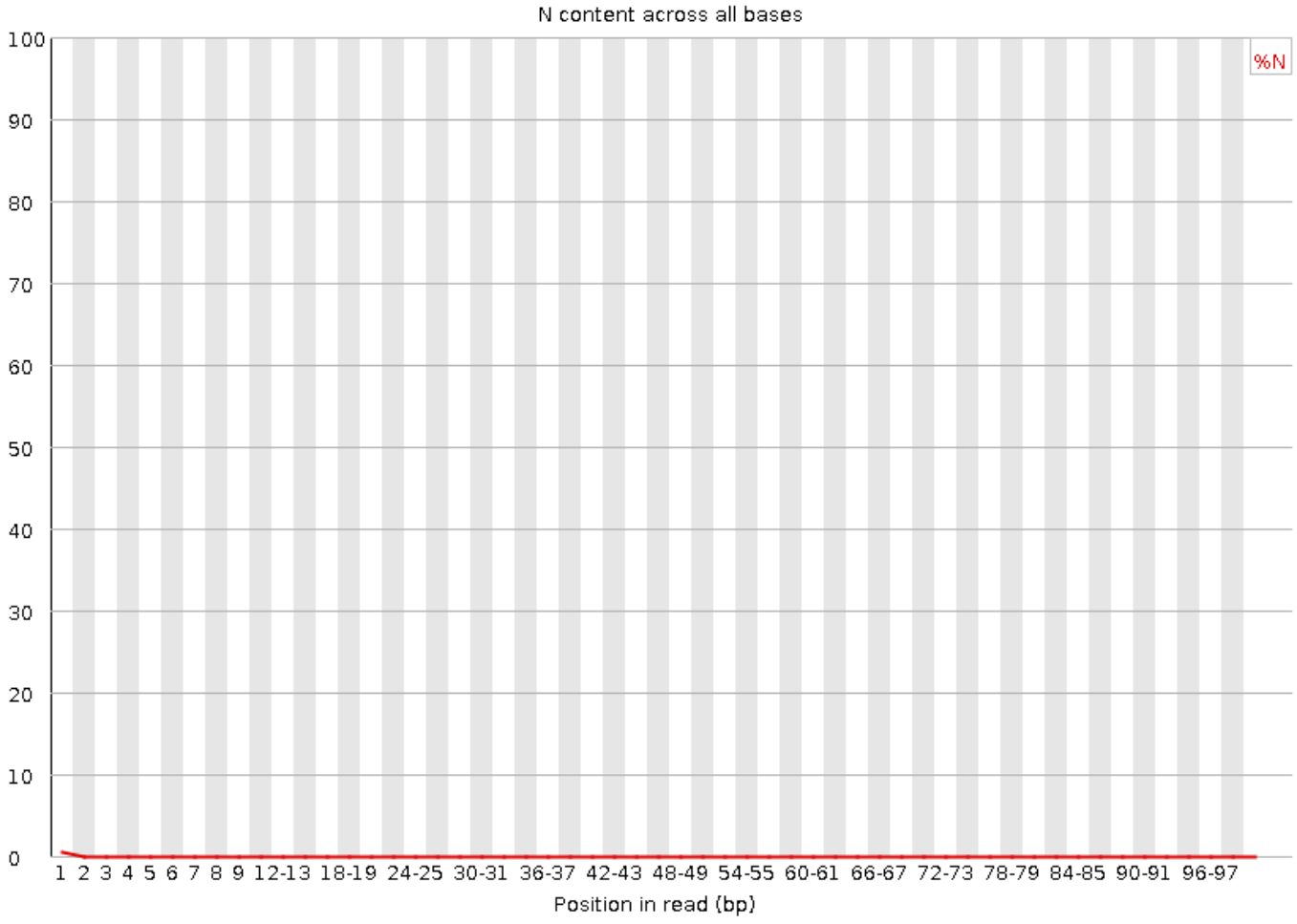
19_3F_R1



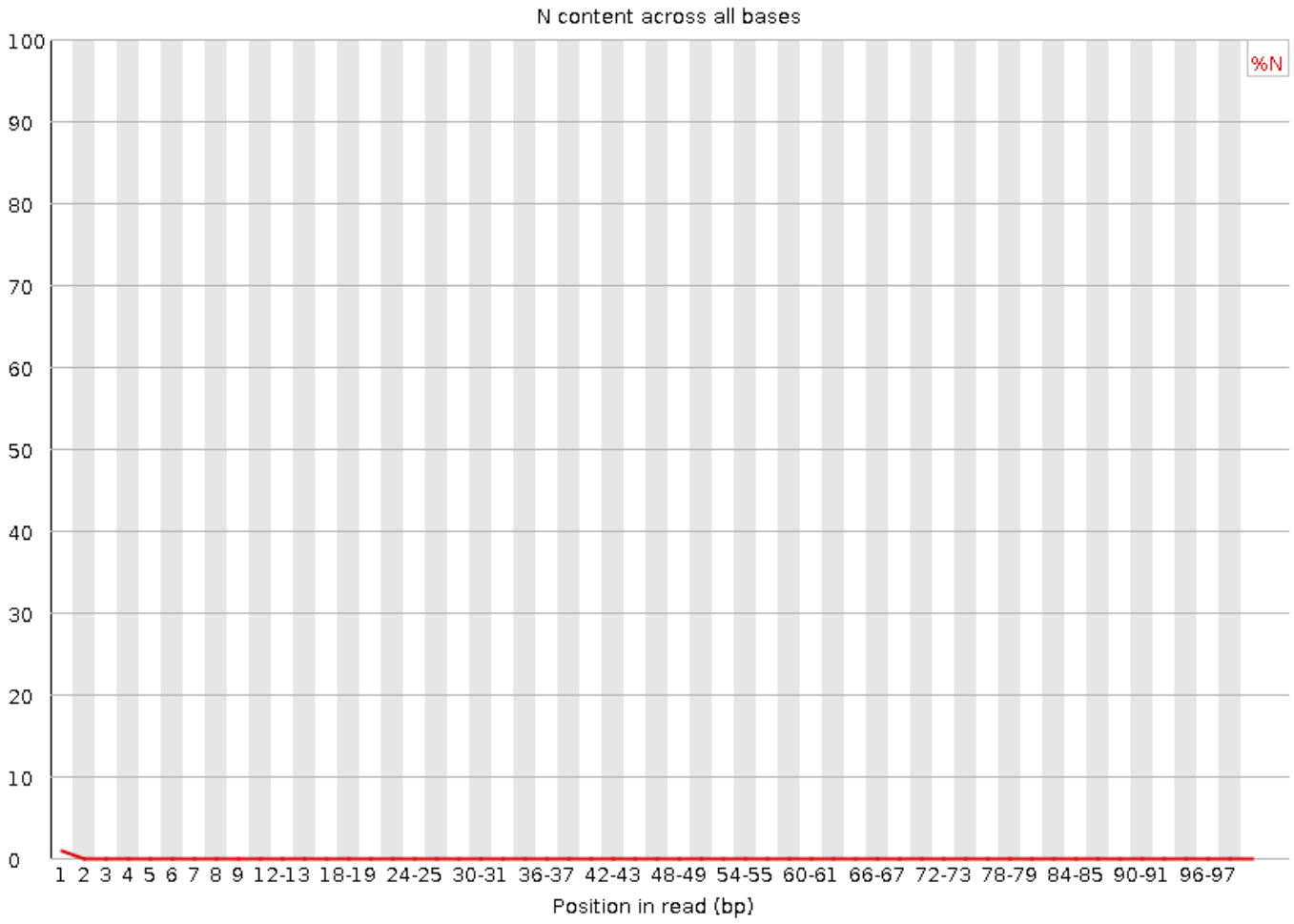
19_3F_R2


N content across all bases
%N
100
90
80
70
60
50
40
30
20
10
0
1 2 3 4 5 6 7 8 9 12-13 18-19 24-25 30-31 36-37 42-43 48-49 54-55 60-61 66-67 72-73 78-79 84-85 90-91 96-97
Position in read (bp)

2_2B_R1


N content across all bases
%N
100
90
80
70
60
50
40
30
20
10
0
1 2 3 4 5 6 7 8 9 12-13 18-19 24-25 30-31 36-37 42-43 48-49 54-55 60-61 66-67 72-73 78-79 84-85 90-91 96-97
Position in read (bp)

2_2B_R2



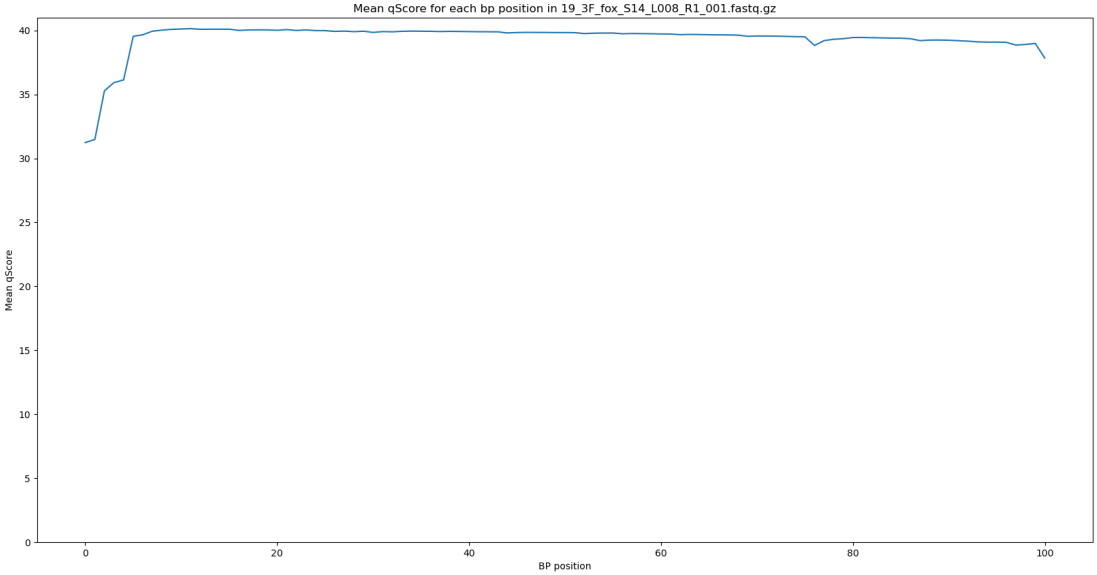
Question 2

Fastqc runs faster, probably because Fastqc is written in Java. Also, the data seems to look pretty similar to the Fastqc graphs.

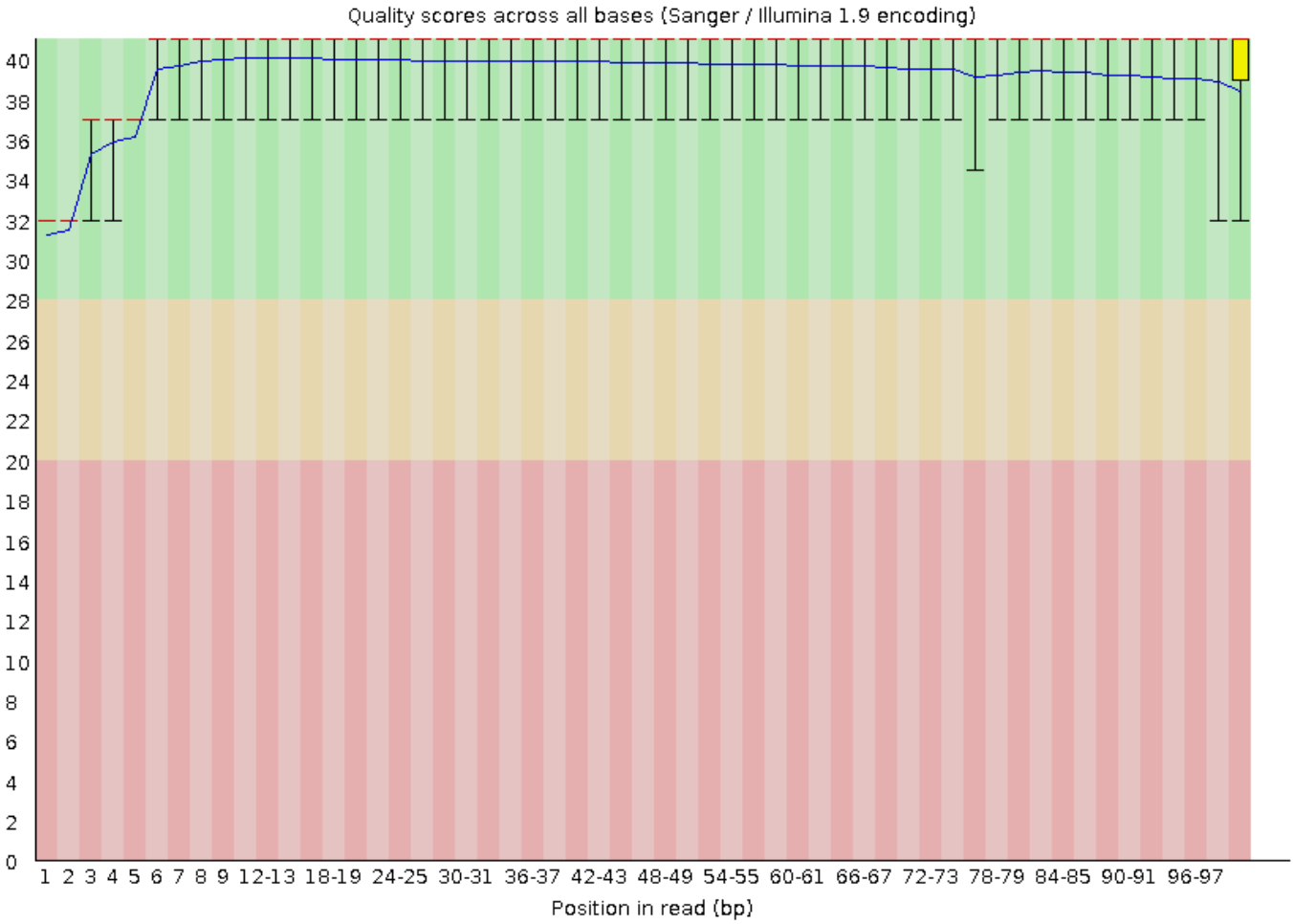
Question 3

The overall state of the data is pretty good. I feel like my graphs and the Fastqc graphs are very similar. They have the same pattern, however, since I did not graph the error bars I don"t know if their error is similar.

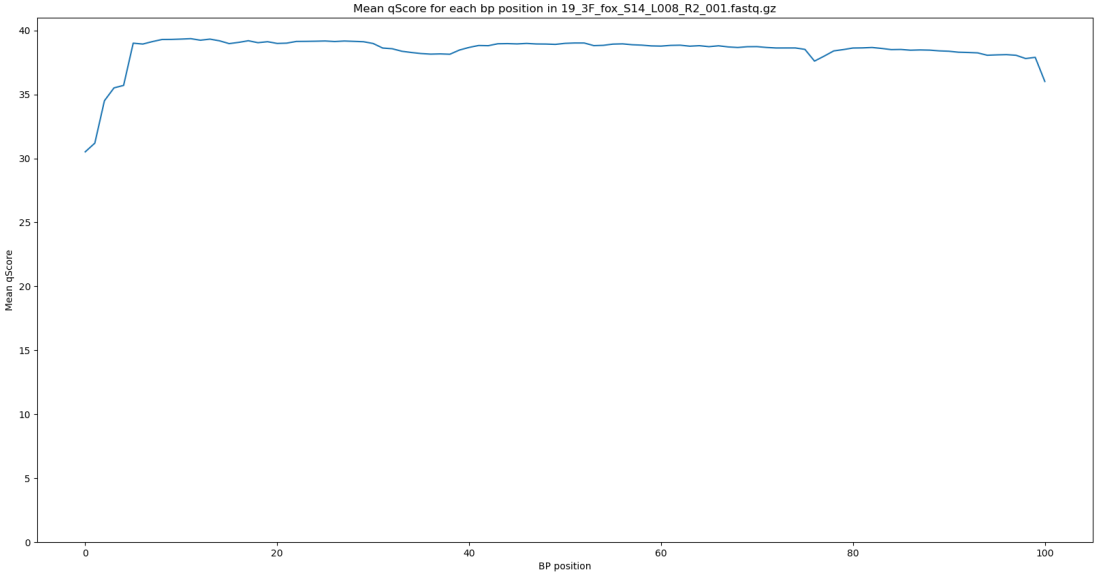
19_3F_fox_S14_L008 R1



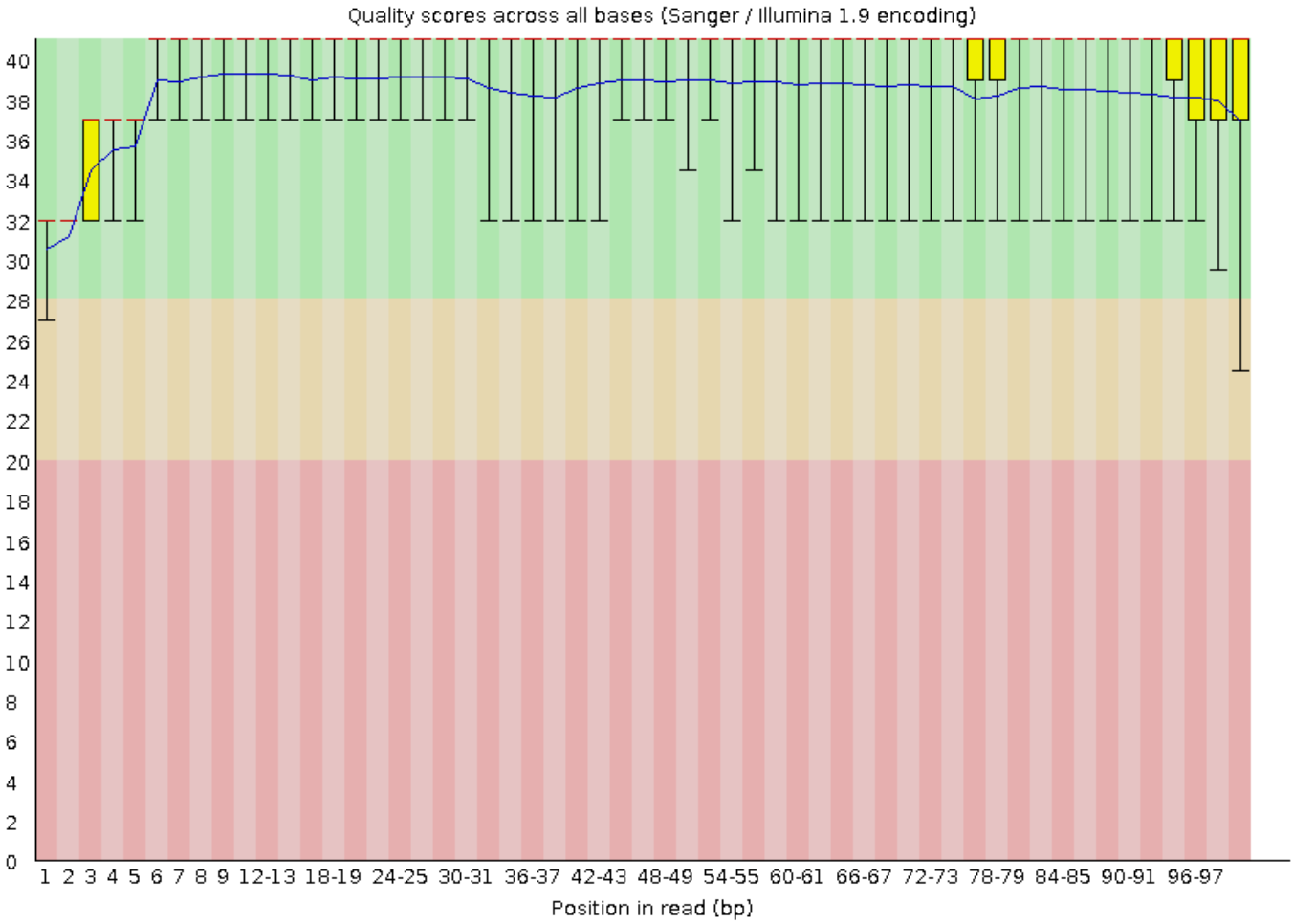
Fastqc R1



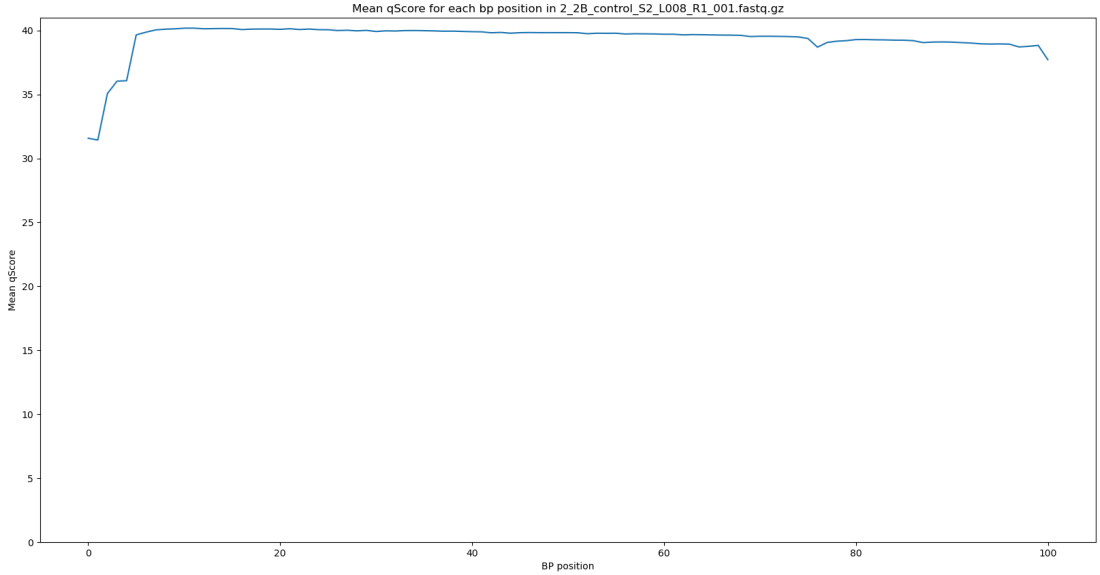
R2


Mean qScore for each bp position in 19_3F_fox_S14_L008_R2_001.fastq.gz
Mean qScore
BP position
0
5
10
15
20
25
30
35
40
0
20
40
60
80
100

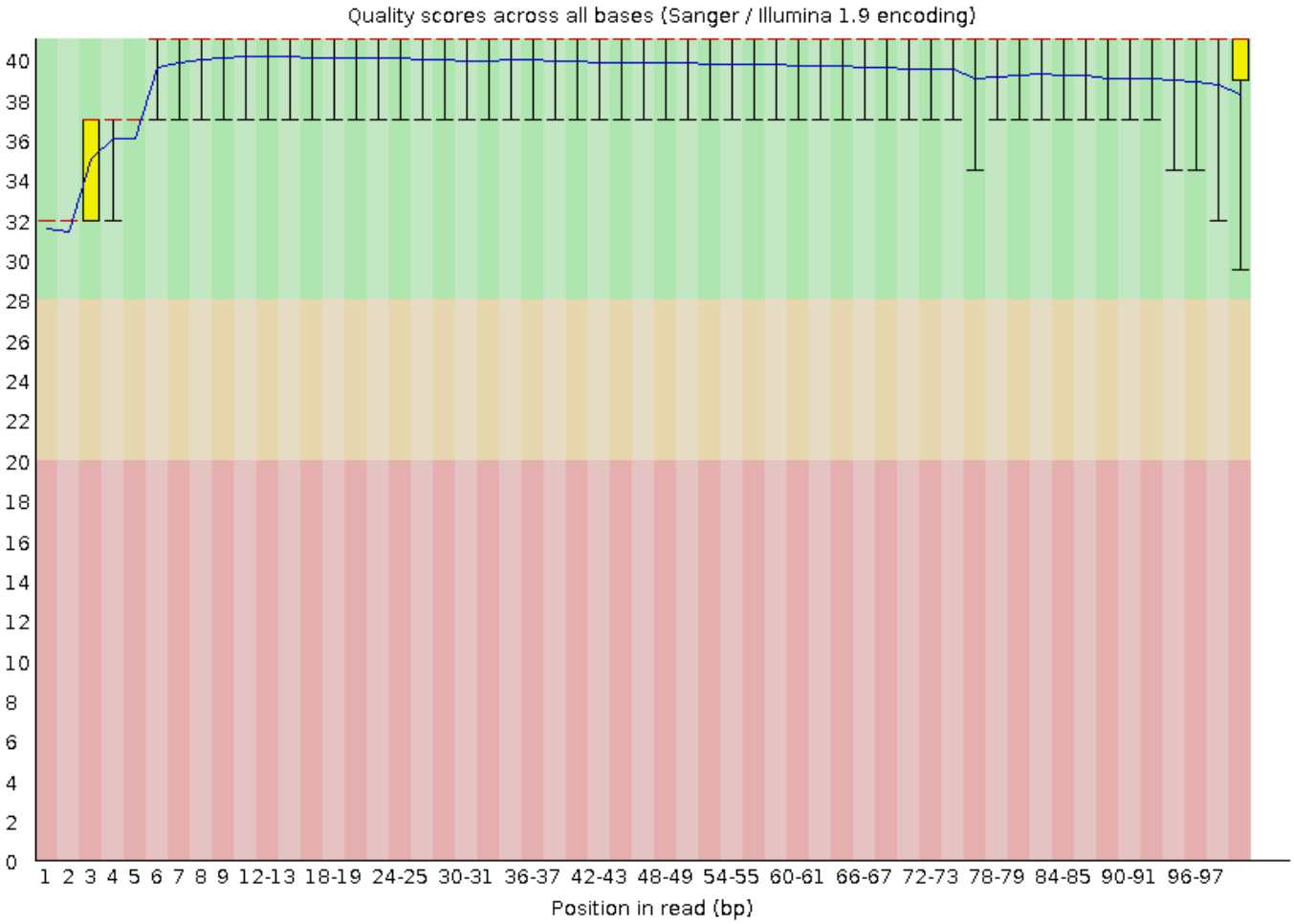
Fastqc R2


Quality scores across all bases (Sanger / Illumina 1.9 encoding)
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36
38
40
1 2 3 4 5 6 7 8 9 12-13 18-19 24-25 30-31 36-37 42-43 48-49 54-55 60-61 66-67 72-73 78-79 84-85 90-91 96-97
Position in read (bp)

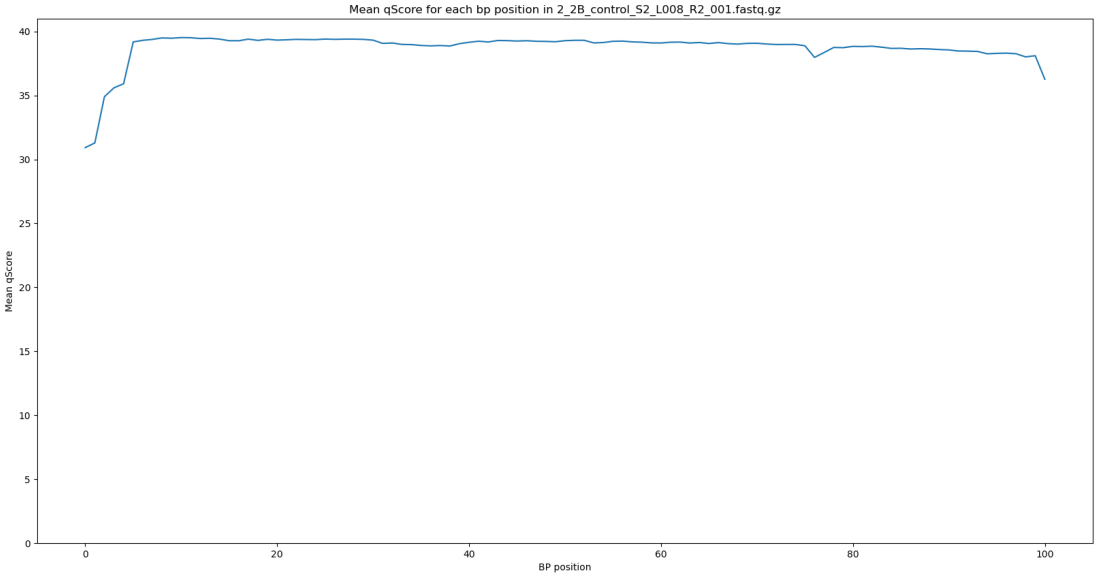
2_2B_control_S2_L008 R1



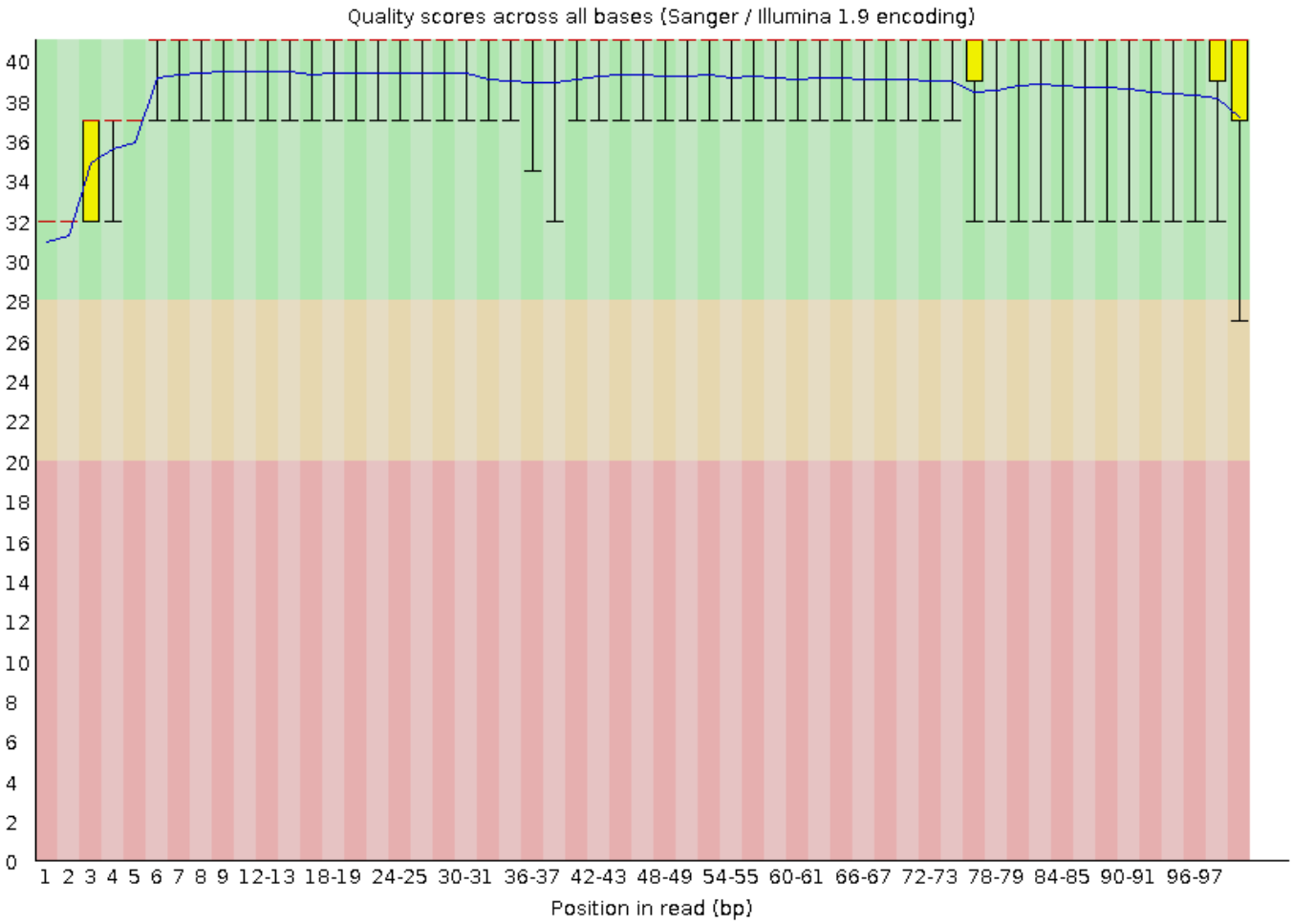
Fastqc R1



R2



Fastqc R2



Part 2

Question 5

You can see the adapters outlined in red if you run the command. Unfortunately it doesn't translate well to a markdown format.

```
zcat
/projects/bgmp/shared/2017_sequencing/demultiplexed/2_2B_control_S2_L008_R1
_001.fastq.gz | sed -n '2~4p' | grep --color=always
'AGATCGGAAGAGCACACGTCTGAACTCCAGTCA' | awk '{print $0}' | head -n 20
ANGCCCCAAACCCAAACAACACACACACACACACACACACACACACACACACACACACAAGATCGGAAGAGCACACG
TCTGAACTCCAGTCACCGATCGATAT
TGACTAGTGACTGACCGGCCTCTAGGCCATTAATGCCCGCCAGGATGTTGATGGCATTAGATCGGAAGAGCACAC
GTCTGAACTCCAGTCACCGATCGATA
GGATGATCAGCCCATCCTTGATCAGCTTCCTGATCTGCTGACGGGAGTTGGCATTGGCGATTTCATTAGATCGGA
AGAGCACACGTCTGAACTCCAGTCAC
GTCCGGATAGCCGTGGCCTGTCACGTTGCGGAGATCGGAAGAGCACACGTCTGAACTCCAGTCACCGATCGATAT
CTCGTATGCCGTCTTCTGCTTGAAAA
TGCTTAAAGTCAGTTCGGACACGCCAGCCTTTATCATAAGCCAATGTGTGCCGGTCTTTCTGGAAGAGAGATCGG
AAGAGCACACGTCTGAACTCCAGTCA
CCGCCATGTGGTTGCTGGGATTTGAACTCCGGACCTTCGGAAGAGCAGTCGGGTGCTCTTACCCACTAGATCGGA
AGAGCACACGTCTGAACTCCAGTCAC
CGGCACTTCTGTCTCTGTTTCAGGTTAATGGCCAGCCACAGGGGCACGTCCACGGAGATCGGAAGAGCACACGTC
TGAACTCCAGTCACCGATCGATATCT
GTAGAGTGTAATGATCCCCCTGTGCTTGGTGACAGTATGGTGCATTCCATCCTTTTCTGAGAGATCGGAAGAGCA
CACGTCTGAACTCCAGTCACCGATCG
GCACCGATTTAACAACAGTTTTCGAAAATTCACAGTTACTGTTGGCTTTTCTGTAGTGGAGATCGGAAGAGCACA
CGTCTGAACTCCAGTCACCGATCGAT
CGTGATCATTTCAAAATCATTCCCGTTCCCGCTACTGTGTTGCGAGCGGTCGAGATCGGAAGAGCACACGTCTGA
ACTCCAGTCACCGATCGATATCTCGT
CCCCGAGGAGCTCCTCACCCCACAGCTTCTTCCAGCTTTATTGGTGTCTGATGGCCTTGGGAGATCGGAAGAGCA
CACGTCTGAACTCCAGTCACCGATCG
TGCTTGTTGACAATGATGCCCACGGCATGCTGGGTGACATTGTAGACTCTTCCGGTTTTGCAGATCGGAAGAGCA
CACGTCTGAACTCCAGTCACCGATCG
TGGGGTTGGAGTTTCCCTCAGCTTACACCATTTGTTTGGGGCAAGCAGATCTGAGAGTTCCAGATCGGAAGAGCA
CACGTCTGAACTCCAGTCACCGATCG
CACCAACTTACGAGCCACCTCTTCATACTTCCTATCGGCCTCTTCTGCAATGTGCAGATCGGAAGAGCACACGTC
TGAACTCCAGTCACCGATCGATATCT
ATTCTTGTCAGTAGCCAGTTTGTGCAGTTCCAGTAGTGACTGATTCACACAGATCGGAAGAGCACACGTCTGAAC
TCCAGTCACCGATCGATATCTCGTAT
CGCTCCTCCTTGCACTGTTTCGTCTGTTCCACACCAGGACCCAGTCCTGCAGCTGTCATCTGTGAGATCGGAAGA
GCACACGTCTGAACTCCAGTCACCGA
CCTGCACTATAGGCACCAGACCTCATCCTATCCAGACCTGCCTGTTAGATCGGAAGAGCACACGTCTGAACTCCA
GTCACCGATCGATATCTCGTATGCCG
GCAGCATATATGTTGGATTTTTAGGAAAGACCAATTCACAGCCCTCATGTGGGCTATAATTTTTAGATCGGAAGA
GCACACGTCTGAACTCCAGTCACCGA
CTGGGCTGGAACCCTGCGGTCTACTTGAGCAGGTTCTGCAGCATGGCCGTAGCATAGGAGATCGGAAGAGCACAC
GTCTGAACTCCAGTCACCGATCGATA
CCCCACTTTGTTCTCCACAGGCTCTGTGCTCTCCTGCCCATCGCCGCTCTGCCTCTCCCCAGAGATCGGAAGAGC
ACACGTCTGAACTCCAGTCACCGATC
(fastqc) [apowers4@n226:/projects/bgmp/apowers4/Bi623/PS/ps3-QAA]
└─╼ zcat
/projects/bgmp/shared/2017_sequencing/demultiplexed/2_2B_control_S2_L008_R2
_001.fastq.gz | sed -n '2~4p' | grep --color=always
'AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT' | awk '{print $0}' | head -n 20
NGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTTGTTTGGGTTTGGGGCTTAGATCGGAAGAGCGTCGT
GTAGGGAAAGAGTGTATCGATCGGTG
NATGCCATCAACATCCTGGCGGGCATTAATGGCCTAGAGGCCGGTCAGTCACTAGTCAAGATCGGAAGAGCGTCG
TGTAGGGAAAGAGTGTATCGATCGGT
```

```
NATGAAATCGCCAATGCCAACTCCCGTCAGCAGATCAGGAAGCTGATCAAGGATGGGCTGATCATCCAGATCGGA
AGAGCGTCGTGTAGGGAAAGAGTGTA
NCGCAACGTGACAGGCCACGGCTATCCGGACAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTATCGATCGGTG
TAGATCTCGGTGGTCGCCGTATCATT
NTCTTCCAGAAAGACCGGCACACATTGGCTTATGATAAAGGCTGGCGTGTCCGAACTGACTTTAAGCAAGATCGG
AAGAGCGTCGTGTAGGGAAAGAGTGT
NGTGGGTAAGAGCACCCGACTGCTCTTCCGAAGGTCCGGAGTTCAAATCCCAGCAACCACATGGCGGAGATCGGA
AGAGCGTCGTGTAGGGAAAGAGTGTA
NCGTGGACGTGCCCCTGTGGCTGGCCATTAACCTGAAACAGAGACAGAAGTGCCGAGATCGGAAGAGCGTCGTGT
AGGGAAAGAGTGTATCGATCGGTGTA
CTCAGAAAAGGATGGAATGCACCATACTGTCACCAAGCACAGGGGGATCATTACACTCTACAGATCGGAAGAGCG
TCGTGTAGGGAAAGAGTGTATCGATC
CCACTACAGAAAAGCCAACAGTAACTGTGAATTTTCGAAAACTGTTGTTAAATCGGTGCAGATCGGAAGAGCGTC
GTGTAGGGAAAGAGTGTATCGATCGG
CGACCGCTCGCAACACAGTAGCGGGAACGGGAATGATTTTGAAATGATCACGAGATCGGAAGAGCGTCGTGTAGG
GAAAGAGTGTATCGATCGGTGTAGAT
CCCAAGGCCATCAGACACCAATAAAGCTGGAAGAAGCTGTGGGGTGAGGAGCTCCTCGGGGAGATCGGAAGAGCG
TCGTGTAGGGAAAGAGTGTATCGATC
GACACCACAGCGACCTCAGAGAACAAGAGCGGCTTCAACTTTGGAACCCTAGACACAAAGAGTGTCAGATCGGAA
GAGCGTCGTGTAGGGAAAGAGTGTAT
GTAAGAACGTGATGCCAAAAGAGGAGACGCCTGCTGAGGATGAAAGTGAAAAGATCGGAAGAGCGTCGTGTAGGG
AAAGAGTGTATCGATCGGTGTAGATC
GCAAAACCGGAAGAGTCTACAATGTCACCCAGCATGCCGTGGGCATCATTGTCAACAAGCAAGATCGGAAGAGCG
TCGTGTAGGGAAAGAGTGTATCGATC
GGAACTCTCAGATCTGCTTGCCCCAAACAAATGGTGTAAGCTGAGGGAAACTCCAACCCCAAGATCGGAAGAGCG
TCGTGTAGGGAAAGAGTGTATCGATC
GCACATTGCAGAAGAGGCCGATAGGAAGTATGAAGAGGTGGCTCGTAAGTTGGTGAGATCGGAAGAGCGTCGTGT
AGGGAAAGAGTGTATCGATCGGTGTA
GGCCCATTCAATCATCTGCTTGTTCTGCACTTCCACAGCCTTGCCACTGTCACTTTCATCACTGTAGATCGGAAG
AGCGTCGTGTAGGGAAAGAGTGTATC
GTGTGAATCAGTCACTACTGGAACTGCACAAACTGGCTACTGACAAGAATAGATCGGAAGAGCGTCGTGTAGGGA
AAGAGTGTATCGATCGGTGTAGATCT
CACAGATGACAGCTGCAGGACTGGGTCCTGGTGTGGAACAGACGAAACAGTGCAAGGAGGAGCGAGATCGGAAGA
GCGTCGTGTAGGGAAAGAGTGTATCG
AACAGGCAGGTCTGGATAGGATGAGGTCTGGTGCCTATAGTGCAGGAGATCGGAAGAGCGTCGTGTAGGGAAAGA
GTGTATCGATCGGTGTAGATCTCGGT
```

cutadapt commands run:

```
#19_3F
cutadapt -a AGATCGGAAGAGCACACGTCTGAACTCCAGTCA -A
AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT -o trimmed_19_3F_1R.fastq.gz -p
trimmed_19_3F_2R.fastq.gz 19_3F_fox_S14_L008_R1_001_fastqc.zip
19_3F_fox_S14_L008_R2_001_fastqc.zip

#2_2B
cutadapt -a AGATCGGAAGAGCACACGTCTGAACTCCAGTCA -A
AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT -o trimmed_2_2B_1R.fastq.gz -p
trimmed_2_2B_2R.fastq.gz 2_2B_control_S2_L008_R1_001_fastqc.zip
2_2B_control_S2_L008_R2_001_fastqc.zip
```

Question 6

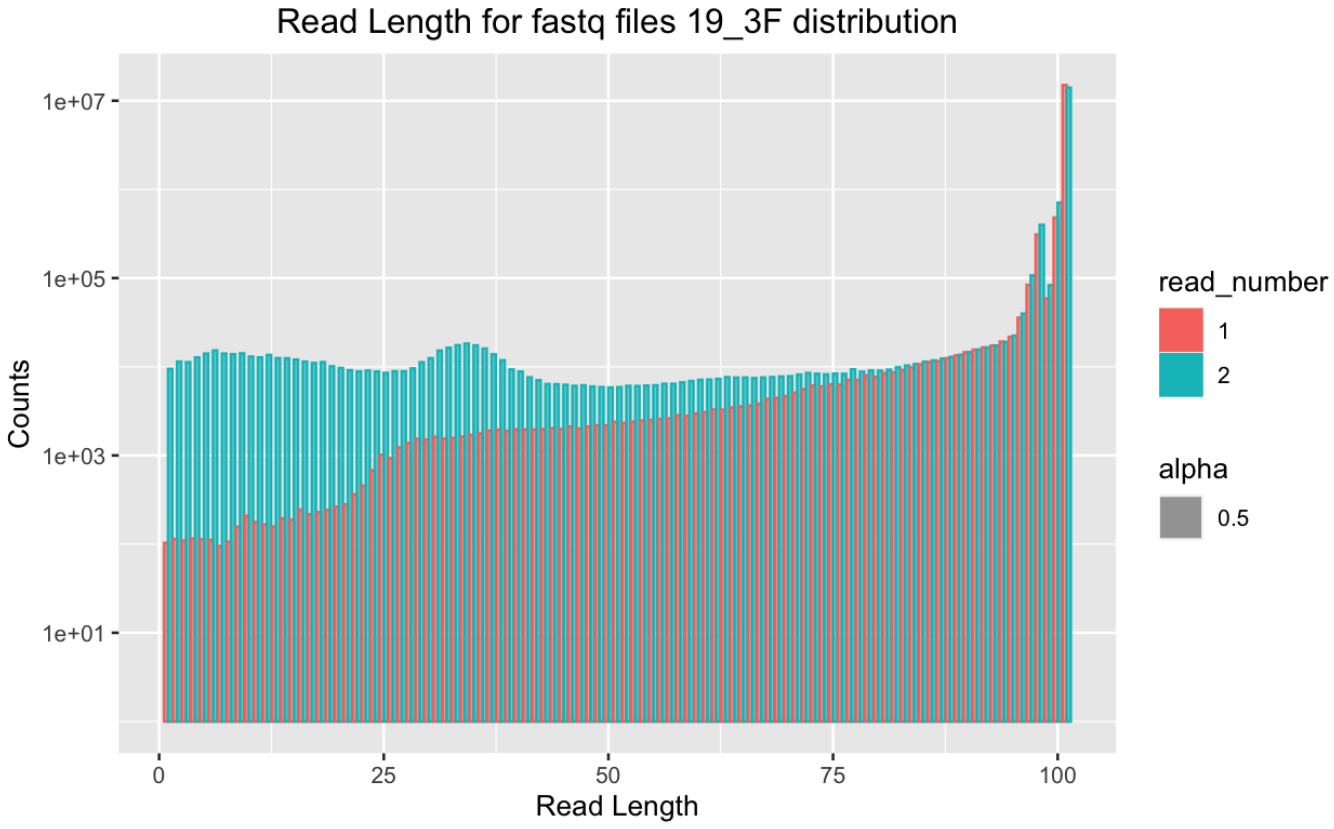
Trimmomatic commands run

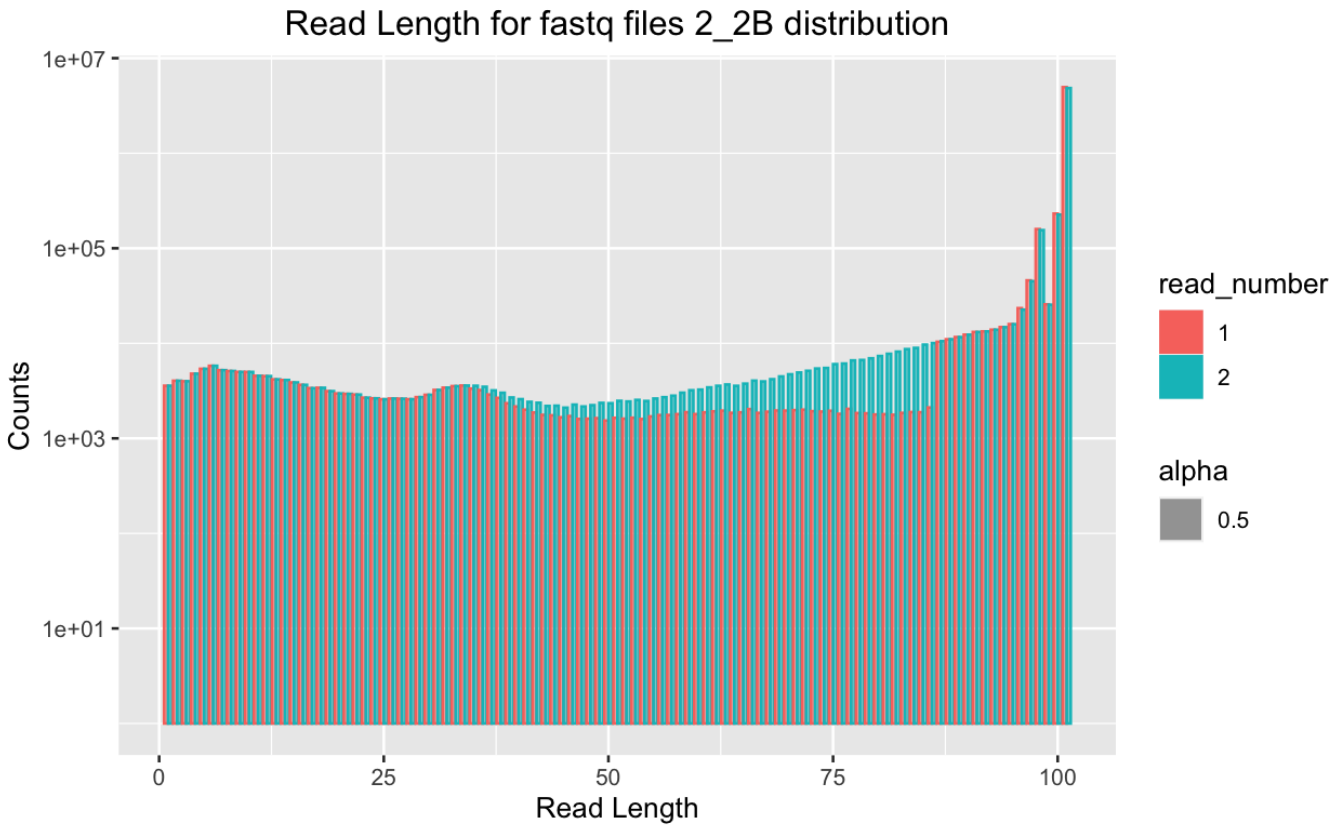
```
# 2_2B
trimmomatic PE output_2_2B/trimmed_2_2B_1R.fastq.gz
output_2_2B/trimmed_2_2B_2R.fastq.gz
trimmomatic_2_2B/trimmed_2_2B_1P.fastq.gz
trimmomatic_2_2B/trimmed_2_2B_1u.fastq.gz
trimmomatic_2_2B/trimmed_2_2B_2P.fastq.gz
trimmomatic_2_2B/trimmed_2_2B_2u.fastq.gz LEADING:3 TRAILING:3 MINLEN:35
SLIDINGWINDOW:5:15

# 19_3F
trimmomatic PE output_19_3F/trimmed_19_3F_1R.fastq.gz
output_19_3F/trimmed_19_3F_2R.fastq.gz
trimmomatic_19_3F/trimmed_19_3F_1P.fastq.gz
trimmomatic_19_3F/trimmed_19_3F_1u.fastq.gz
trimmomatic_19_3F/trimmed_19_3F_2P.fastq.gz
trimmomatic_19_3F/trimmed_19_3F_2u.fastq.gz LEADING:3 TRAILING:3 MINLEN:35
SLIDINGWINDOW:5:15
```

Question 7

R made graphs:





Part 3

Question 10

For 19_3F reads: Sequences Mapped: 31075050 Sequences Unmapped: 1569764

For 2_2B reads: Sequences Mapped: 58690 Sequences Unmapped: 11508686

Question 12

I propose that the data are strand specific, because __no_feature in 19_3F_unstranded.txt = 1256405 out of 16322407 which is 7.69% __no_feature in 19_3F_sam_stranded.txt = 14086570 out of 16322405 which is 86.30%

More then 3/4 of the data has no features if it is given the stranded comand which would lead me to believe that the rest of the data would bind to the reverse strand. Since there are so many binding when made unstranded. Meaning that the library is stranded. Since ~79% would seem to bind to the reverse strand.

You can't really draw any conclusions from the 2_2B reads as they do not seem to align to the Mouse genome.

Slurm Scripts

Genome index

```
#!/bin/bash
#SBATCH --account=bgmp
```

```
#SBATCH --partition=bgmp
#SBATCH --job-name=star_genomeindex
#SBATCH --output=star-%j-sbatch.out
#SBATCH --nodes=1
#SBATCH --ntasks-per-node=1
#SBATCH --cpus-per-task=8
#SBATCH --time=2:00:00

# Activate the environment
conda activate fastqc

#Unzip the files
gunzip /projects/bgmp/apowers4/Bi623/PS/ps3-
QAA/mouse_genome/Mus_musculus.GRCm39.dna.primary_assembly.fa.gz
gunzip /projects/bgmp/apowers4/Bi623/PS/ps3-
QAA/mouse_genome/Mus_musculus.GRCm39.104.gtf.gz

# Run the actual database indexer STAR
/usr/bin/time -v STAR --runThreadN 8 \
--runMode genomeGenerate \
--genomeDir /projects/bgmp/apowers4/Bi623/PS/ps3-
QAA/mouse_genome/genome_mous_star_2.7.1a \
--genomeFastaFiles /projects/bgmp/apowers4/Bi623/PS/ps3-
QAA/mouse_genome/Mus_musculus.GRCm39.dna.primary_assembly.fa \
--sjdbGTFfile /projects/bgmp/apowers4/Bi623/PS/ps3-
QAA/mouse_genome/Mus_musculus.GRCm39.104.gtf \

# Rezip the files
zip /projects/bgmp/apowers4/Bi623/PS/ps3-
QAA/mouse_genome/Mus_musculus.GRCm39.dna.primary_assembly.fa
zip /projects/bgmp/apowers4/Bi623/PS/ps3-
QAA/mouse_genome/Mus_musculus.GRCm39.104.gtf
```

Star Align

Star Align 2_2B

```
#!/bin/bash
#SBATCH --account=bgmp
#SBATCH --partition=bgmp
#SBATCH --job-name=star_align2
#SBATCH --output=staralign2-%j-sbatch.out
#SBATCH --nodes=1
#SBATCH --ntasks-per-node=1
#SBATCH --cpus-per-task=8
#SBATCH --time=10:00:00

# Activate the environment
conda activate fastqc

# Run the actual database indexer STAR
```

```
/usr/bin/time -v STAR --runThreadN 8  --runMode alignReads \
--outFilterMultimapNmax 3 \
--outSAMunmapped Within KeepPairs \
--alignIntronMax 1000000 --alignMatesGapMax 1000000 \
--readFilesCommand zcat \
--readFilesIn /projects/bgmp/apowers4/Bi623/PS/ps3-
QAA/trimmomatic_2_2B/trimmed_2_2B_1P.fastq.gz
/projects/bgmp/apowers4/Bi623/PS/ps3-
QAA/trimmomatic_2_2B/trimmed_2_2B_2P.fastq.gz \
--genomeDir /projects/bgmp/apowers4/Bi623/PS/ps3-
QAA/mouse_genome/genome_mous_star_2.7.1a \
--outFileNamePrefix aligned_star_sam_2_2B
```

Star Align 19_3F

```
#!/bin/bash
#SBATCH --account=bgmp
#SBATCH --partition=bgmp
#SBATCH --job-name=star_align19
#SBATCH --output=staralign19-%j-sbatch.out
#SBATCH --nodes=1
#SBATCH --ntasks-per-node=1
#SBATCH --cpus-per-task=8
#SBATCH --time=10:00:00

# Activate the environment
conda activate fastqc

# Run the actual database indexer STAR
/usr/bin/time -v STAR --runThreadN 8  --runMode alignReads \
--outFilterMultimapNmax 3 \
--outSAMunmapped Within KeepPairs \
--alignIntronMax 1000000 --alignMatesGapMax 1000000 \
--readFilesCommand zcat \
--readFilesIn /projects/bgmp/apowers4/Bi623/PS/ps3-
QAA/trimmomatic_19_3F/trimmed_19_3F_1P.fastq.gz
/projects/bgmp/apowers4/Bi623/PS/ps3-
QAA/trimmomatic_19_3F/trimmed_19_3F_2P.fastq.gz \
--genomeDir /projects/bgmp/apowers4/Bi623/PS/ps3-
QAA/mouse_genome/genome_mous_star_2.7.1a \
--outFileNamePrefix aligned_star_sam_19_3F
```

HTSeq-count

2_2B

```
# Unstranded
#!/bin/bash
#SBATCH --account=bgmp
```

```
#SBATCH --partition=bgmp
#SBATCH --job-name=count_2B_unstranded
#SBATCH --time=2:00:00

htseq-count --stranded=no aligned_star_sam_2_2BAligned.out.sam
Mus_musculus.GRCm39.104.gtf > htseq_outputs/2_2B_unstranded.txt

# Stranded
#!/bin/bash
#SBATCH --account=bgmp
#SBATCH --partition=bgmp
#SBATCH --job-name=count_2B_stranded
#SBATCH --time=2:00:00

htseq-count --stranded=yes aligned_star_sam_2_2BAligned.out.sam
Mus_musculus.GRCm39.104.gtf > htseq_outputs/2_2B_stranded.txt
```

19_3F

```
# Unstranded
#!/bin/bash
#SBATCH --account=bgmp
#SBATCH --partition=bgmp
#SBATCH --job-name=count_19_unstranded
#SBATCH --time=2:00:00

htseq-count --stranded=no aligned_star_sam_19_3FAligned.out.sam
Mus_musculus.GRCm39.104.gtf > htseq_outputs/19_3F_unstranded.txt

# Stranded
#!/bin/bash
#SBATCH --account=bgmp
#SBATCH --partition=bgmp
#SBATCH --job-name=count_19_stranded
#SBATCH --time=2:00:00

htseq-count --stranded=yes aligned_star_sam_19_3FAligned.out.sam
Mus_musculus.GRCm39.104.gtf > htseq_outputs/19_3F_sam_stranded.txt
```