# QAA_report

Anh Vo

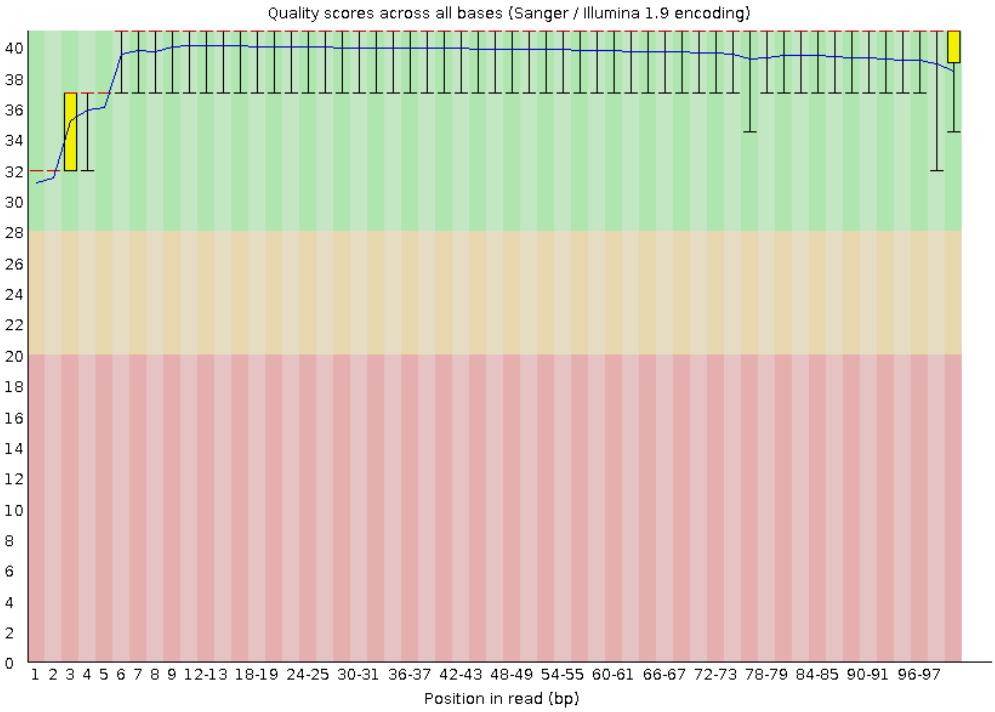
2022-09-07

## PART 1 - Read Quality Score Distributions

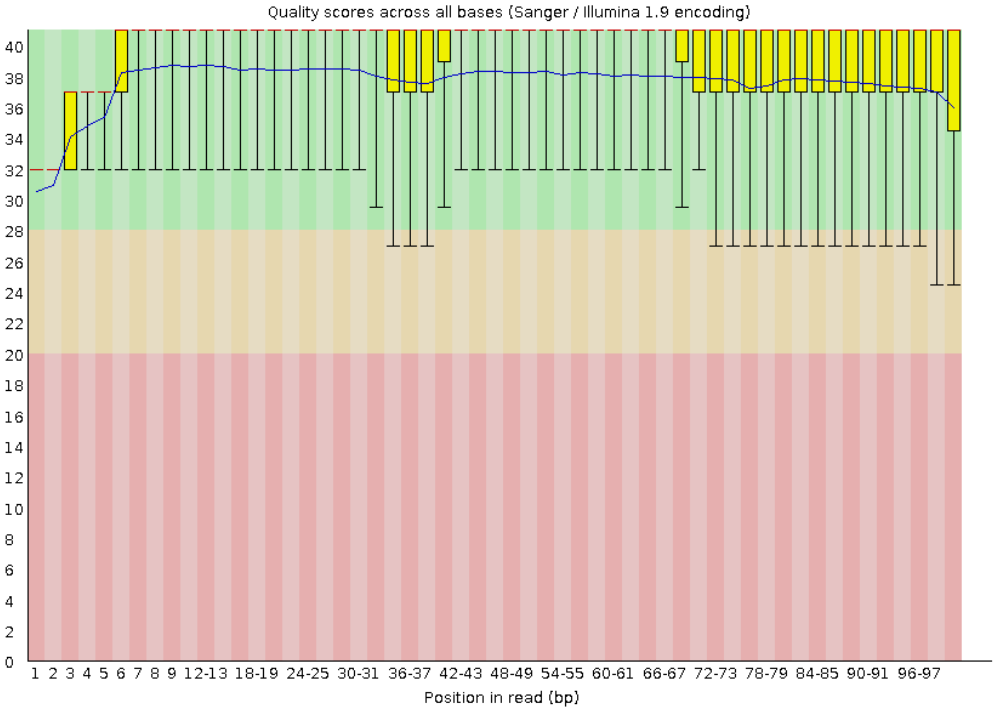
---

### FastQC Plots

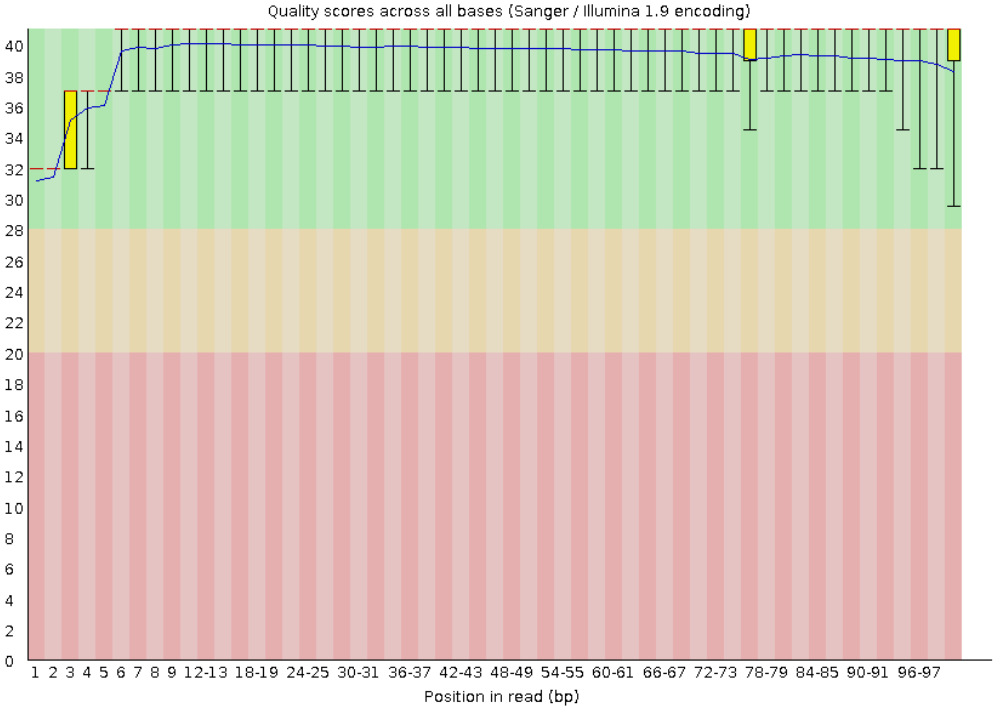
#### Sample 22 Read 1 per base quality



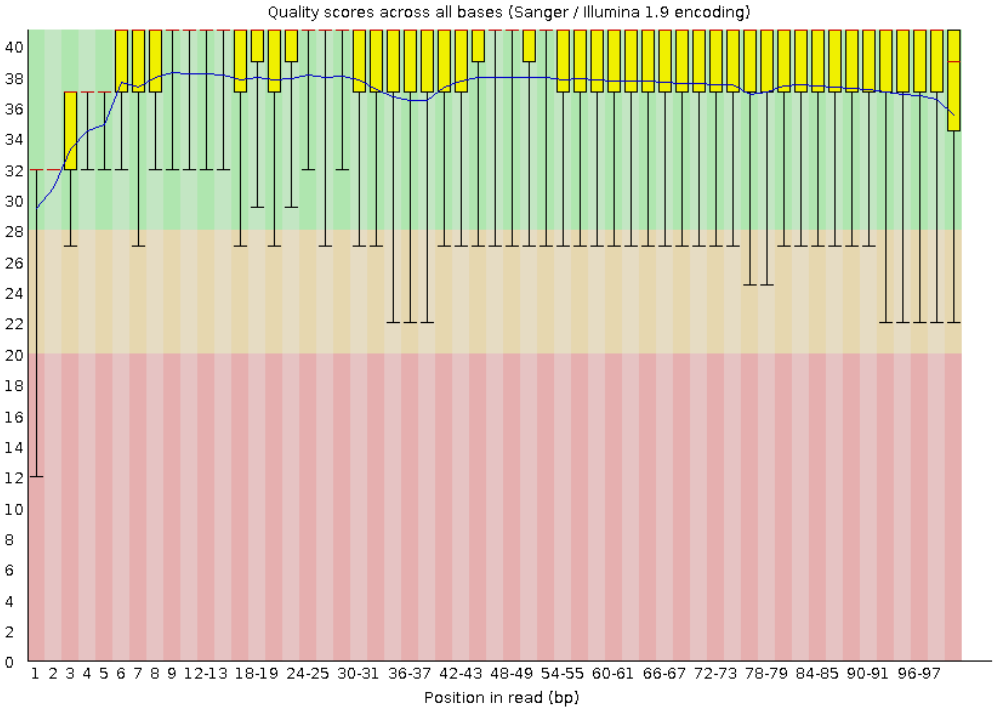
### Sample 22 Read 2 per base quality



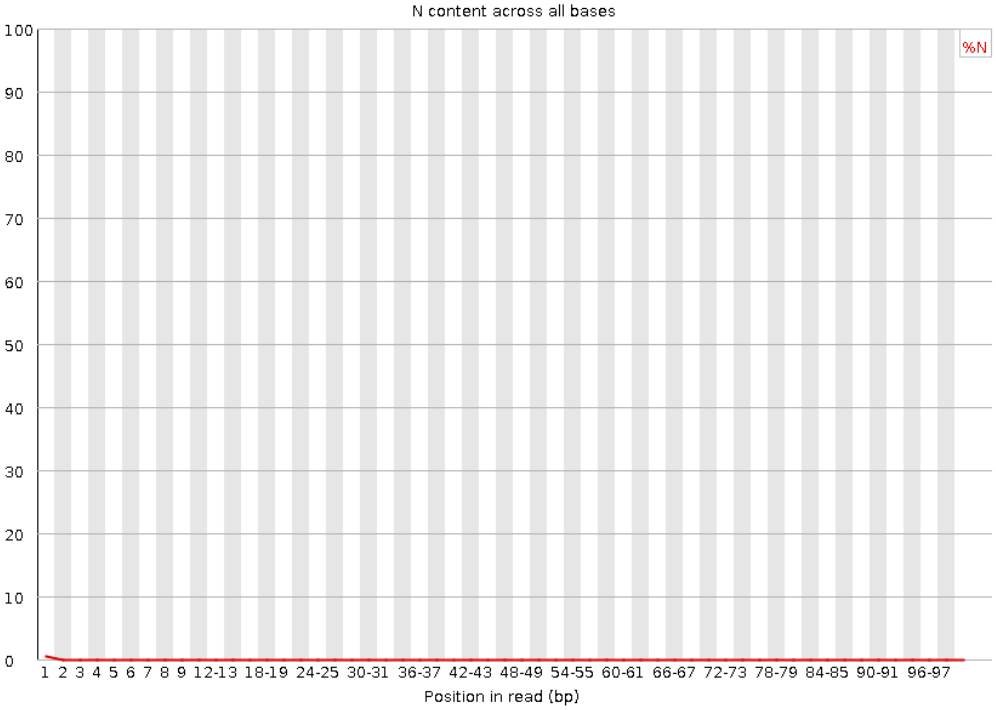

### Sample 23 Read 1 per base quality



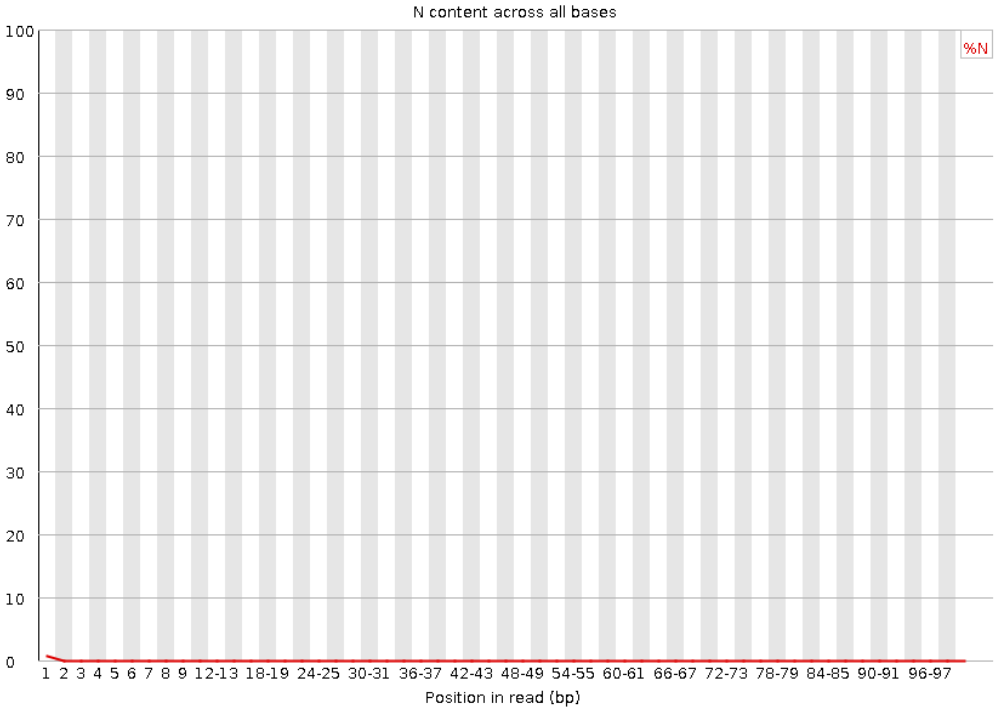
### Sample 23 Read 2 per base quality



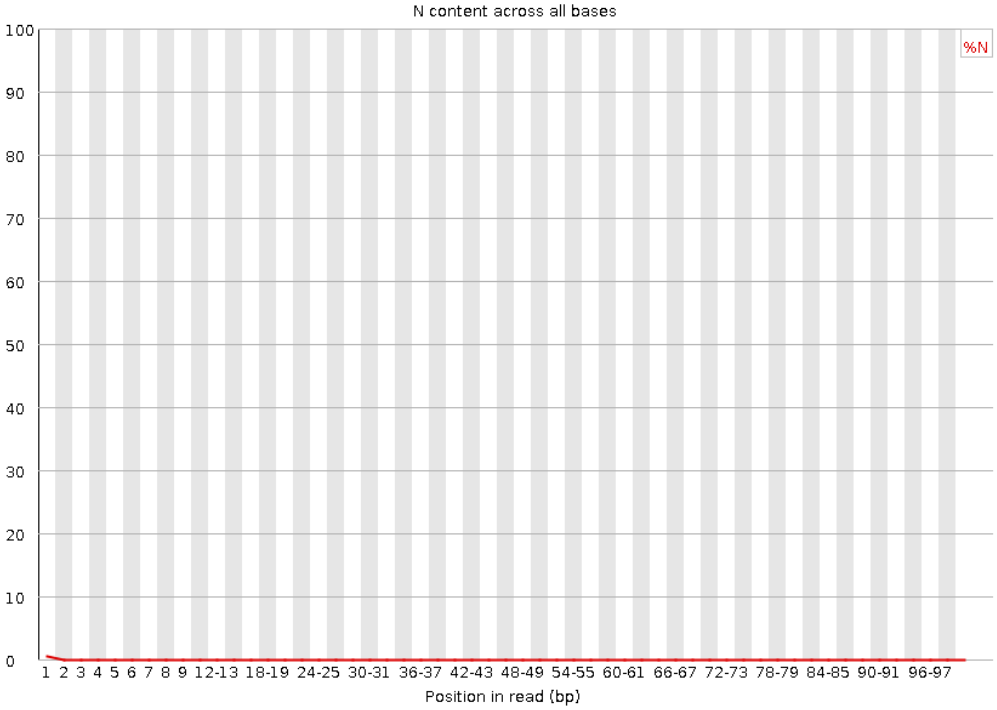

### Sample 22 Read 1 per-base N content



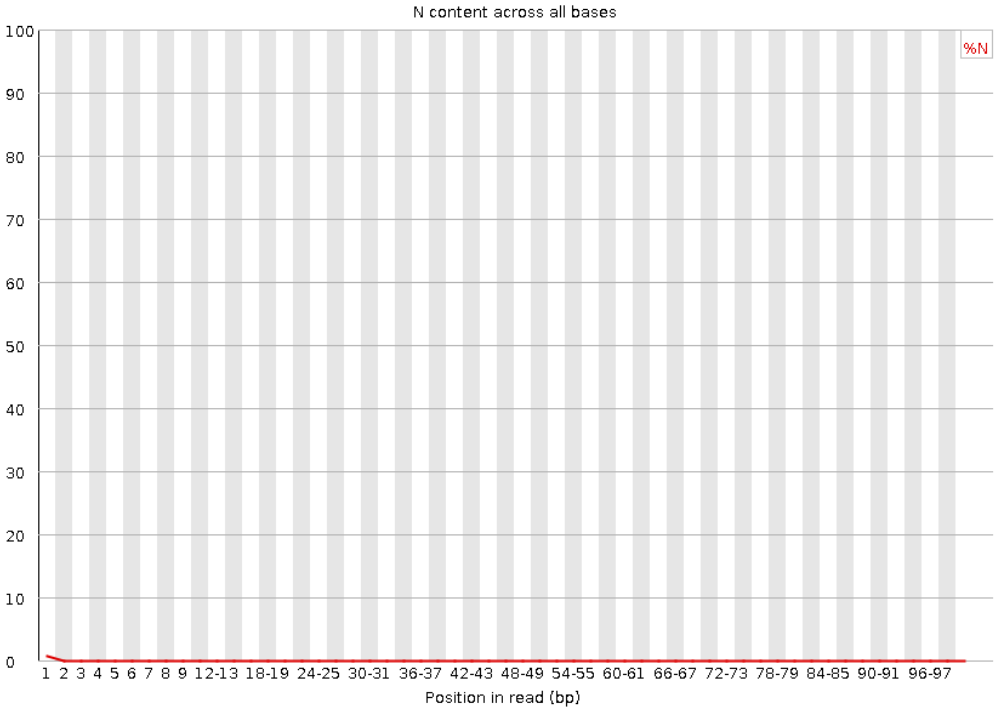
### Sample 22 Read 2 perbase N content




### Sample 22 Read 1 per-base N content



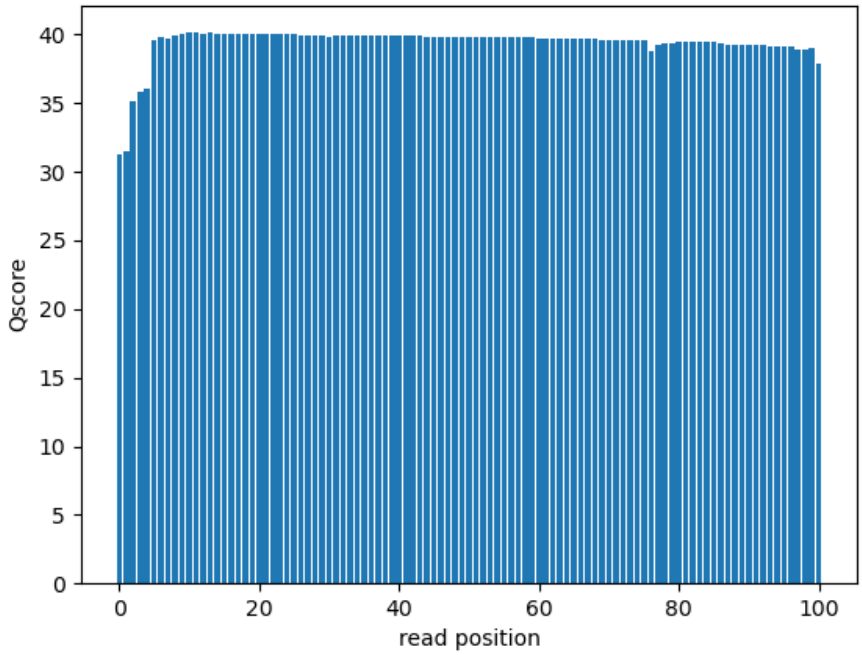
**Sample 22 Read 2 perbase N content**



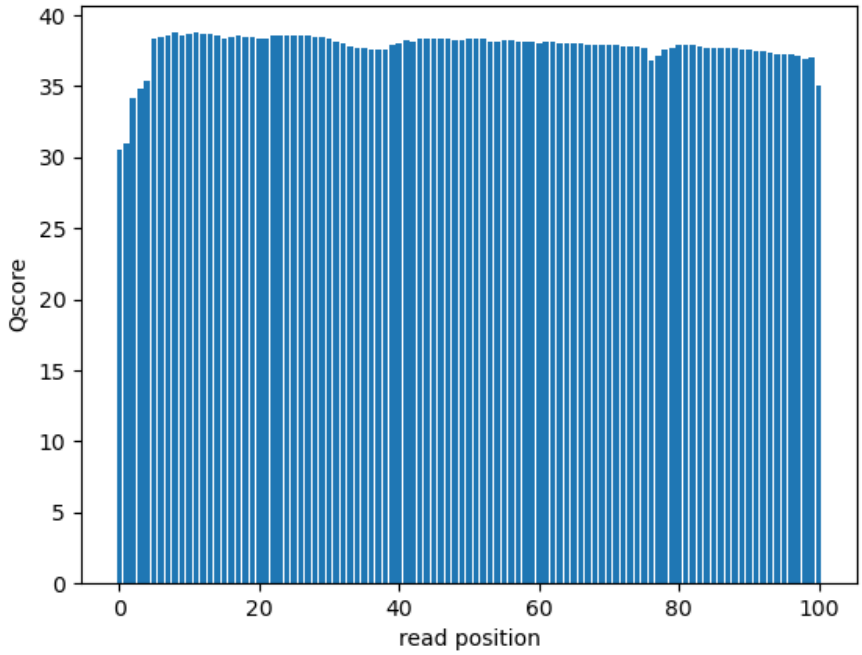

The quality scores at read position one are approximately 32 for each sample, which is consistent with the N content plot where position 1 has an unknown base.

## Mean Quality Plots from Demultiplex Script

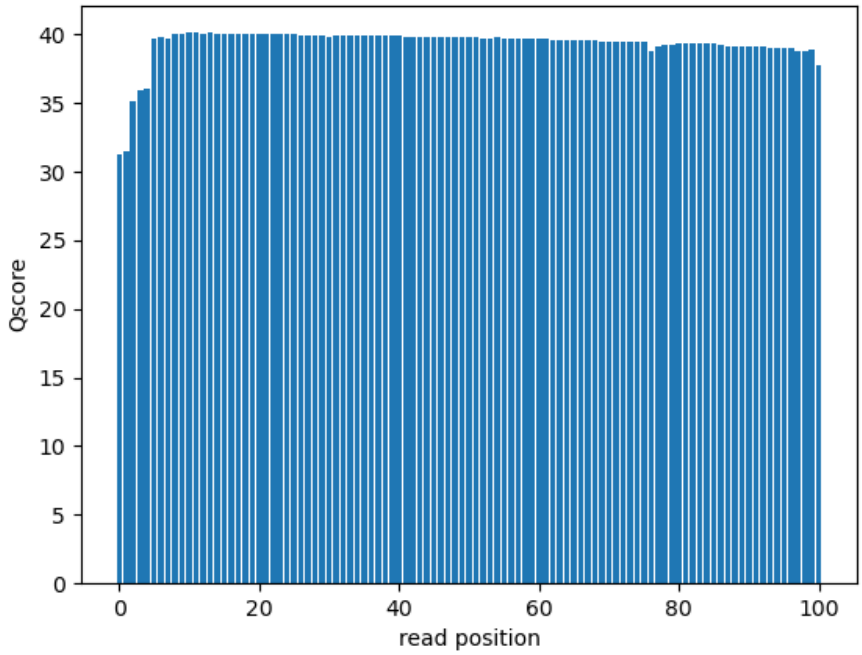
### Sample 22 Read 1



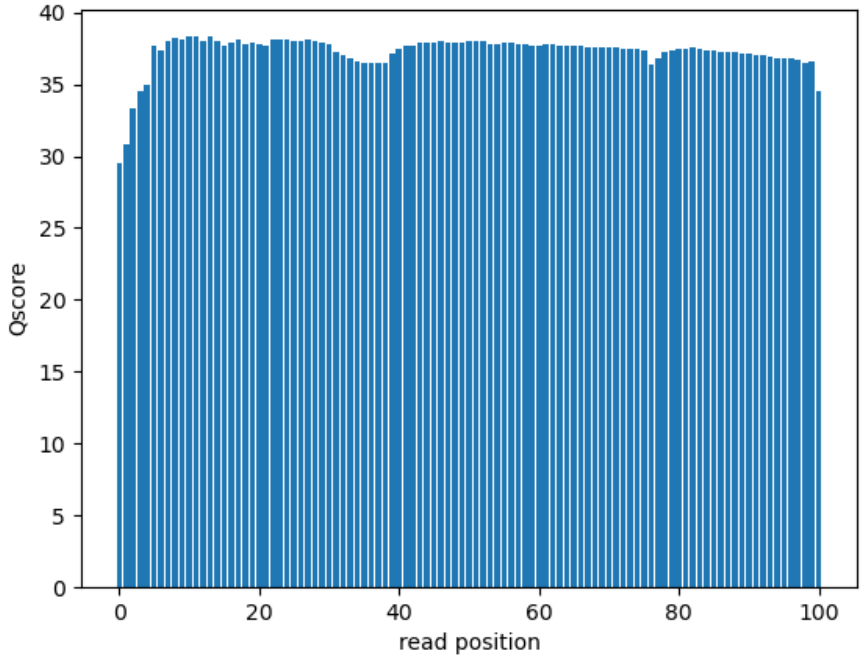
**Sample 22 Read 2**




**Sample 23 Read 1**



**Sample 23 Read 2**




The FastQC quality plots and the demultiplex quality plots are very similar in quality scores at each base position.

The FastQC runtime was 7.5x faster than the demultiplex script.

The overall mean quality scores for sample 22 and 23 are above 35, indicating that these are high quality reads.

# PART 2 - Adapter Trimming Comparison

---

## Adapter trimmed using Cutadapt

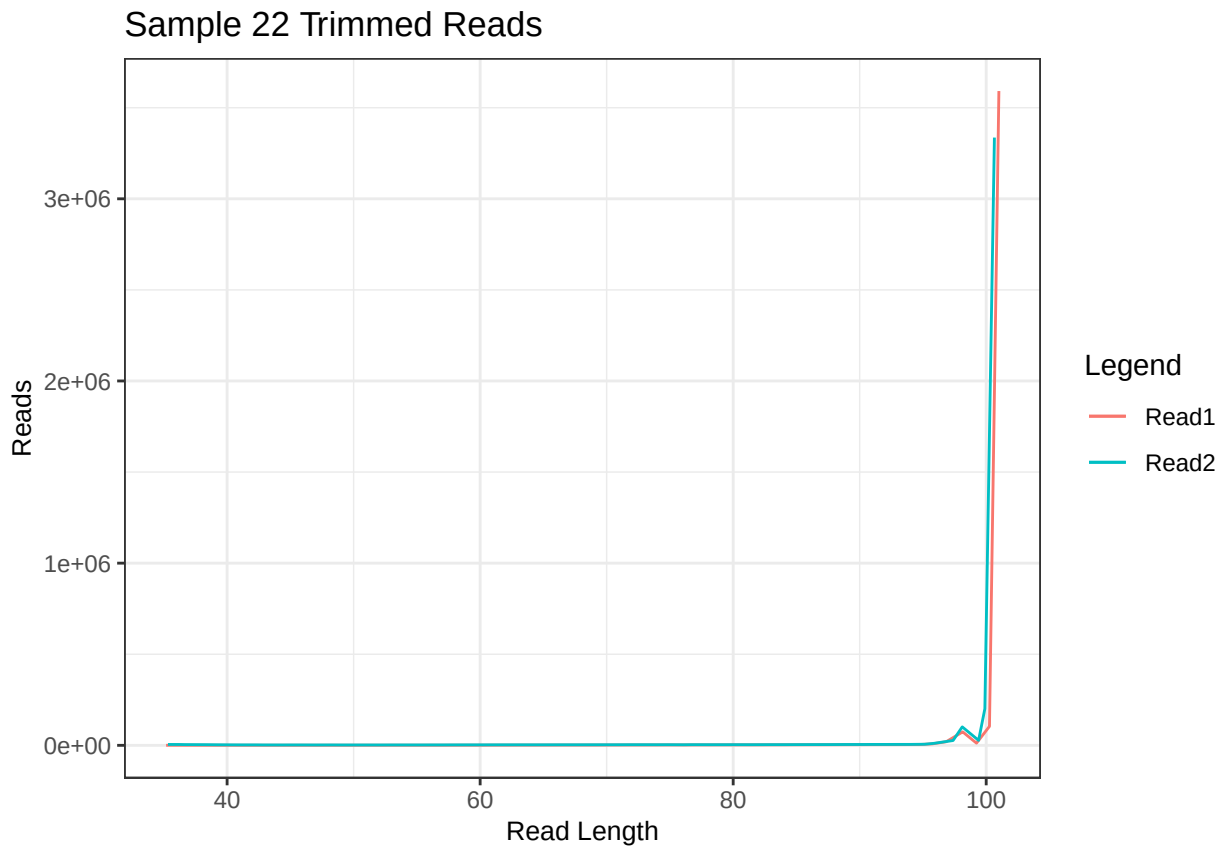
Adapters were trimmed using cutadapt and sanity checked using `grep` to confirm the adapters were removed:

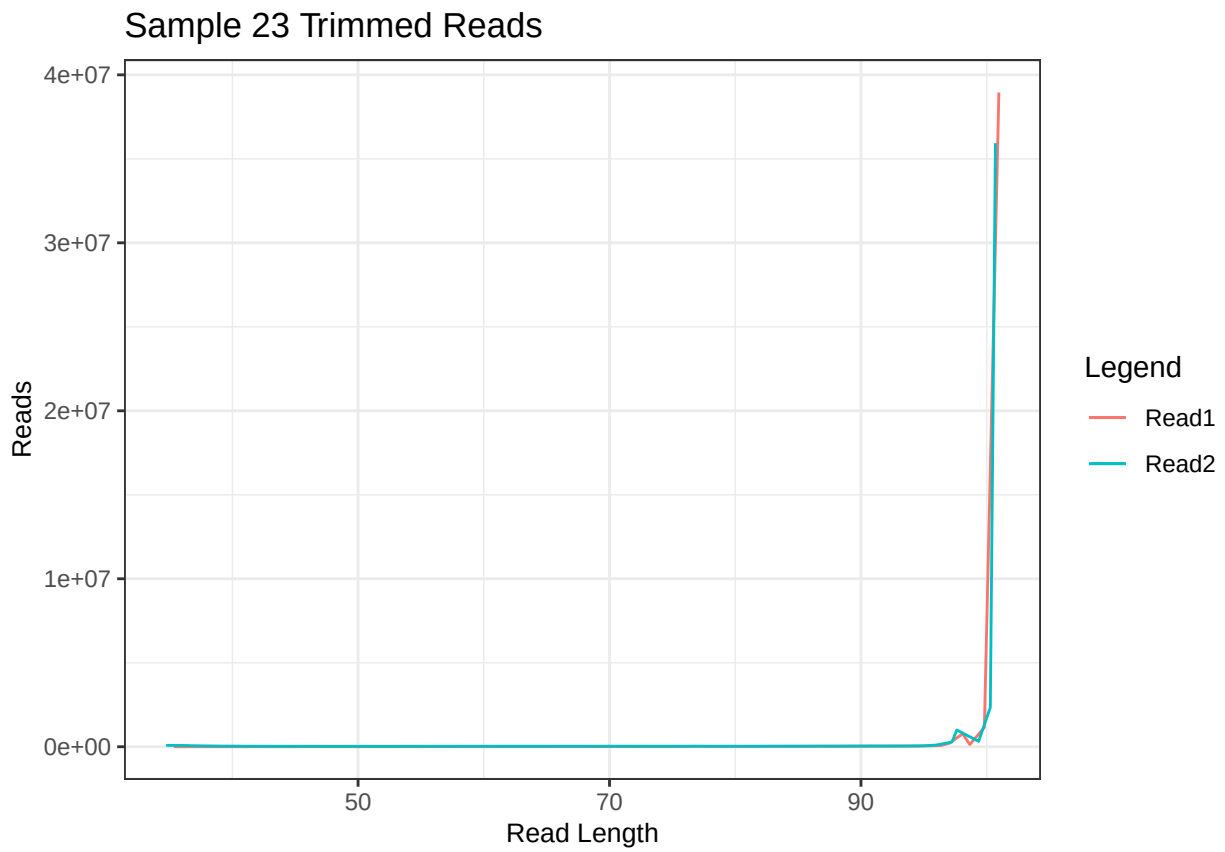
```
cat 22_3H_both_S16_L008_R1_001.fastq.gz.trimmed | grep "AGATCGGAAGAGCACACGTCTGAACTCCAGTCA" | wc -l
cat 23_4A_control_S17_L008_R1_001.fastq.gz.trimmed | grep "AGATCGGAAGAGCACACGTCTGAACTCCAGTCA" | wc -l
cat 22_3H_both_S16_L008_R2_001.fastq.gz.trimmed | grep "AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT" | wc -l
cat 22_3H_both_S16_L008_R2_001.fastq.gz.trimmed | grep "AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT" | wc -l
```

Sample 22:
3.8% of R1 reads were adapter-trimmed.
4.6% of R2 read were adapter-trimmed.

Sample 23:
3.1% of R1 reads were adapter-trimmed.
3.7% of R2 reads were adapter-trimmed.

### Quality trimmed reads using Trimmomatic






Read 2 is expected to be adapter-trimmed at a slightly higher rate than read 1 since the samples sit on the sequencer for a longer period of time. Therefore, read 2 will be degraded at a higher rate and result in lower number of reads at 101 bps.

# PART 3 - Alignment and Strand-specificity

---

## HtSeq Mappped Gene Features

### Sample 22 Mapped Reads

| | Mapped | %Mapped | Unmapped | %Unmapped |
|---|---|---|---|---|
| **Stranded** | 141326 | 3.62% | 37598016 | 96.38% |
| **Rev Stranded** | 3228275 | 82.75% | 672852 | 17.25% |

### Sample 23 Mapped Reads

| | Mapped | %Mapped | Unmapped | %Unmapped |
|---|---|---|---|---|
| **Stranded** | 1298933 | 3.09% | 40757643 | 96.91% |

| | Mapped | %Mapped | Unmapped | %Unmapped |
|---|---|---|---|---|
| **Rev Stranded** | 28696716 | 68.23% | 13359860 | 31.77% |

The data from these RNA-Seq libraries are from strand-specific libraries because the majority of the samples are "reversed-stranded" mapped in each sample. In Sample 22, 82.75% of the reads are mapped when performed under a "reverse-stranded" analysis compared to only 3.62% mapped when doing a "stranded" analysis. The trend is the same in Sample 23 with 68.23% of the reads also "reverse-stranded" mapped.

The libraries would be unstranded if there were an even distribution of reads mapped and unmapped in in both stranded and reverse-stranded analysis.