Espinoza_QAA_report

2022-09-08

Title: RNA-seq Quality Assessment Report - Demultiplexed Paired Files 21_3G.fastq and 34_4H.fastq

#Read Quality Score Distributions For Each Demultiplexed FASTQ File # Comparing per base quality scores and per N base content in forward R1 and reverse R2 of 21_3G and 34_4H fastq files. Figure 1 distribution demonstrates a mean quality score of 39 across the majority of base positions for 21_3G R1 fastq file.Overall, every quality score that was examined has a value higher than 30.As expected, the sequence starts out with a few low quality score values (QS of 32 from bp 0-4) and ends with a few more (QS of 30 from bp 98-101). Because read quality deteriorates over time, it is expected to see a decline in quality ratings at these two places.The trend seen in figure 1 is supported by the per N base plot, which show that the base call improves as the quality score increases. As anticipated, a minor peak can be seen between base pairs 0 and 1, but it stays at zero for the rest of the sequence. A general trend of N values of zero indicates high-quality data since sequencers add a N base when they can't confidently recall a base. R2 reverse exhibits a generally significantly lower mean quality score across all base reads than the forward R1 fastq file (Fig. 3). For instance, a mean quality score of 22 is the lowest possible over all reverse reads (Fig.3, bp 98-101). The inner quartile ranges for the 25th to 75th percentile is also more frequently represented in figure 3 by the yellow whisker plots (bp 68-101). The wide ranges shown by the upper and lower whisker plots show that the reverse fastq file has more variation in the quality scores across base readings than the forward fastq file. This observation is also expected as the mean quality scores will always be higher for R1 in contrast to R2 paired end reads. Similar to forward R1 fastq file, the per N base distribution also matches the overall trend seen in the mean per base quality score of figure 3. The mean quality score results in the forward R1 fastq file for 34 4H are comparable to those in 21_3G, with an average quality score of 39 being maintained over most base pairs (Fig. 5,6). It is important to note that in contrast to the reverse R2 fastq file from 21_3G, which has a low-quality score of 22, the reverse R2 file from 34_4H has a slightly higher score of 25. Similar to forward R1 fastq file, reverse R2 has higher variability of quality scores per base (Fig.7,8).

Comparing FastQD quality score distribution software with my own written script.

Overall, the plots produced by my demultiplexed script revealed a mean quality score distribution that was comparable to those generated from fastqc software. It's significant to note that I gave the demultiplexed script a quality score of 30. The reverse R2 file for 34 4H fastq had a quality score of 25, whereas the reverse R2 file for 21_3G fastq had a quality score of 22, therefore the fastqc software must have a different quality score cutoff. The upper and lower quartiles for each base read are among the additional information that Fastqc outputs in comparison to my manual script (very cool!). This information provides further insight as to how the quality scores varies per base (as seen in both reverse fastq files for 21_3G and 34-4H). Fastqc also operates at a significantly faster rate, producing all the graphs in less than two minutes. Additionally, it comes pre-programmed with both a warning and a failure flag. According to its documentation, I discovered that a warning will be issued if the lower quartile of a base is less than a quality score value of 10. Additionally, it will generate an error notice if any base position has a N value higher than 5% or if the total N value is higher than 20% (Fastqc manual). Online Documentation for fastqc: https://dnacore.missouri.edu/PDF/FastQC_Manual.pdf

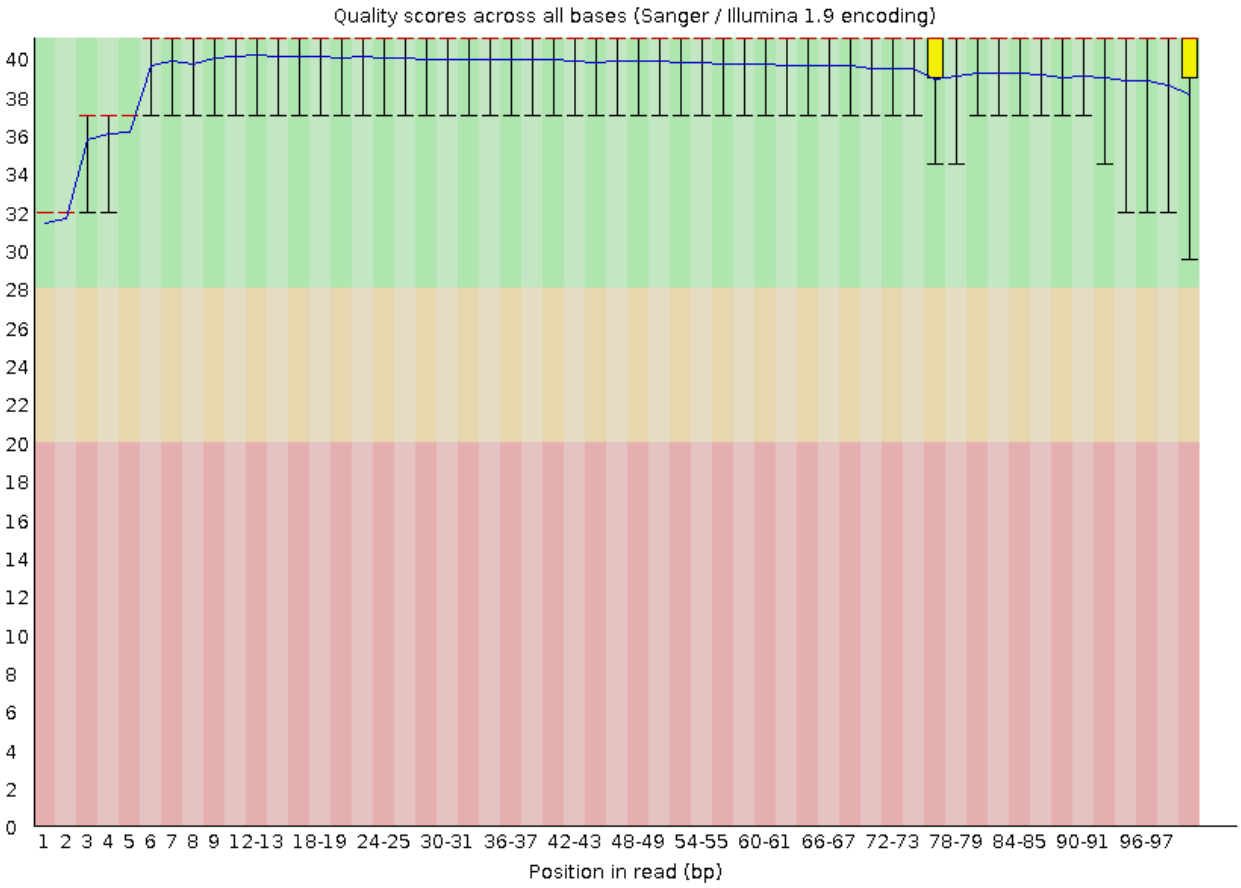


Figure 1: Figure 1: 21_3G Forward R1 Per Base Quality Distribution

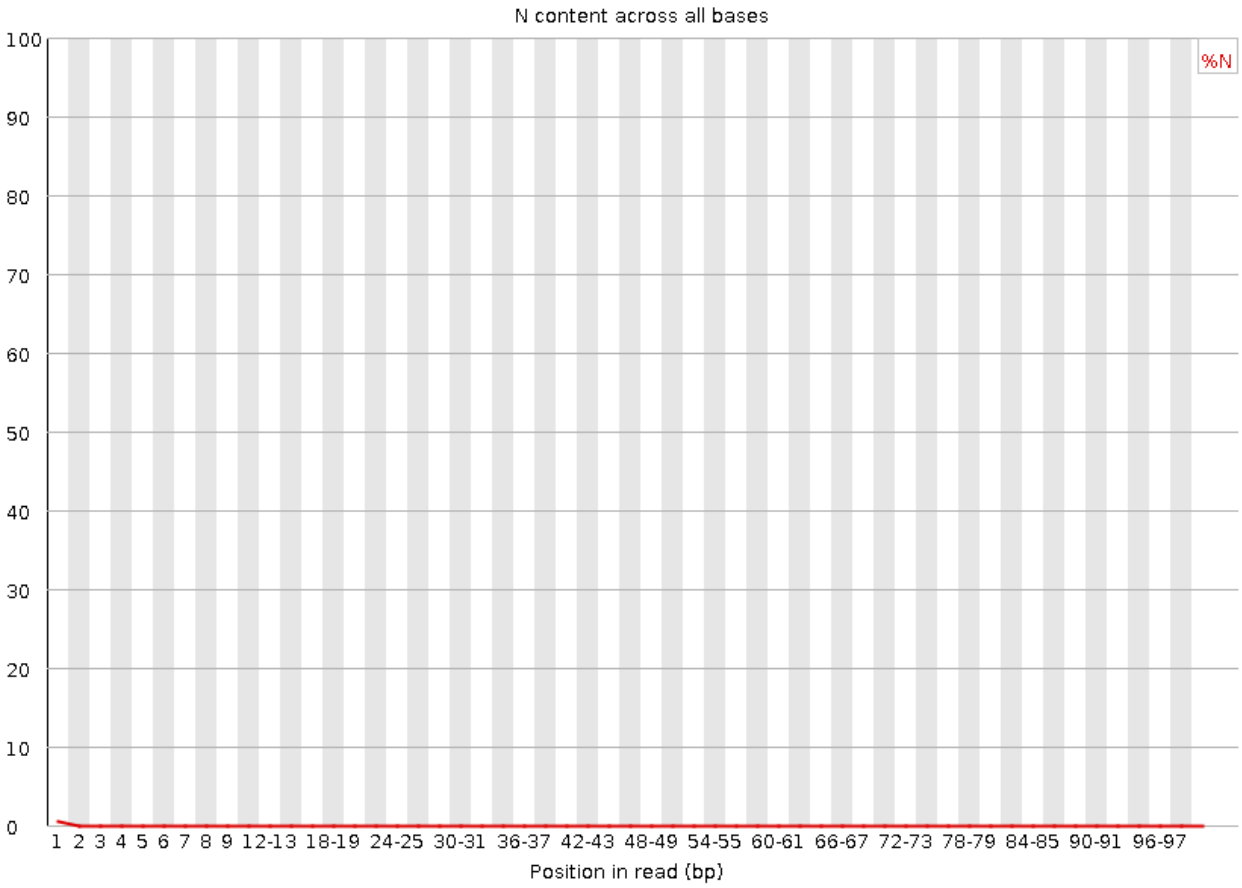


Figure 2: Figure 2: 21_3G Forward R1 Per N Content

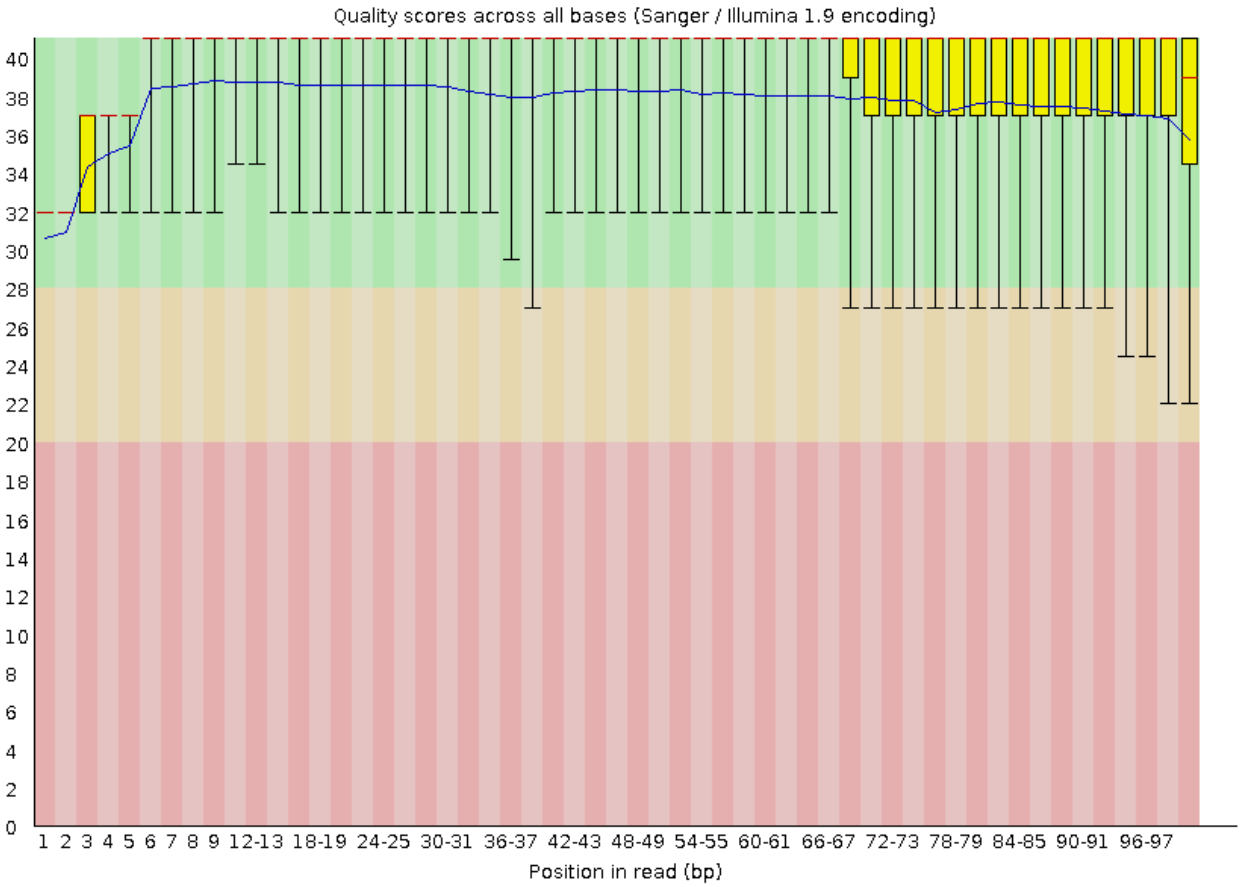


Figure 3: Figure 3: 21_3G Reverse R2 Per Base Quality Distribution

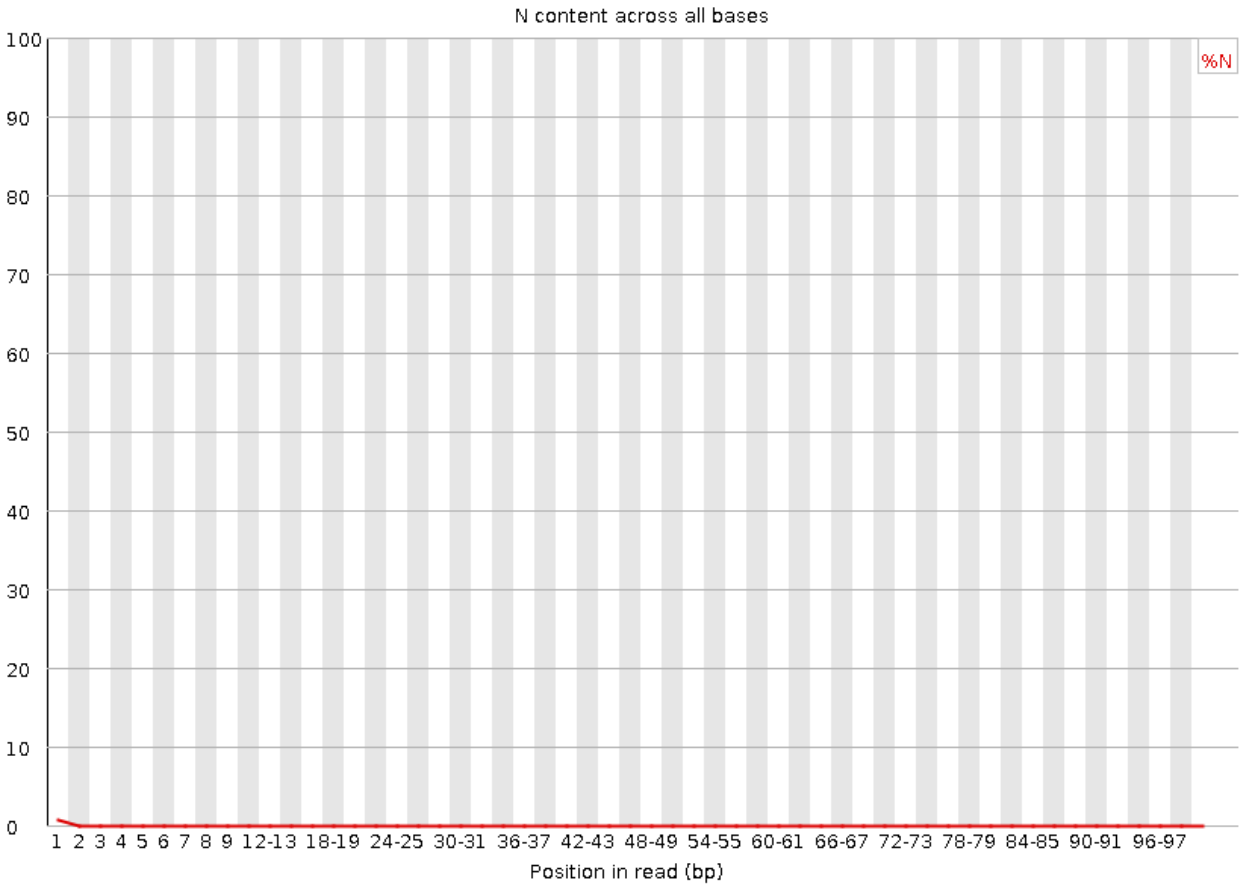


Figure 4: Figure 4: 21_3G Reverse R2 Per N Content

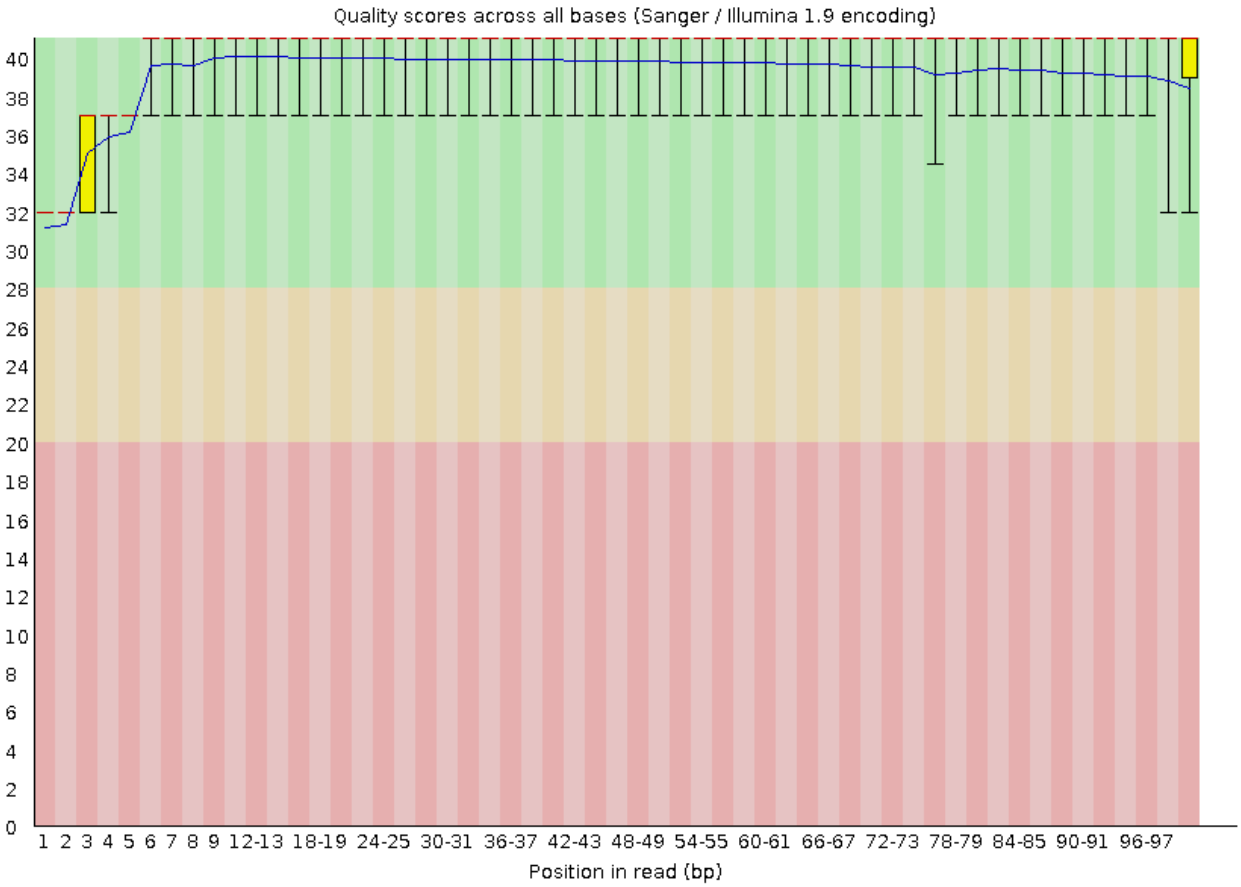


Figure 5: Figure 5: 34_4H Forward R1 Per Base Quality Distribution

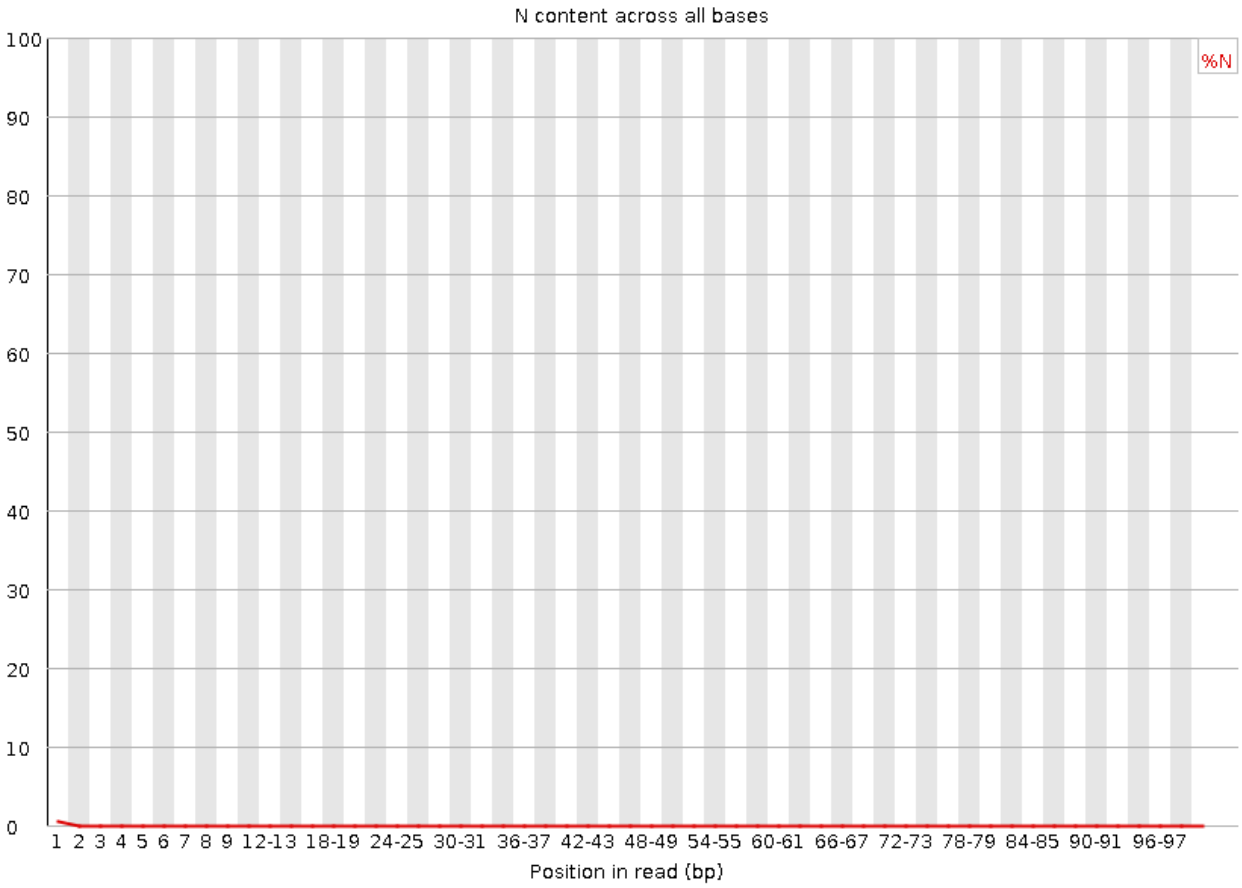


Figure 6: Figure 6: 34_4H Forward R1 Per N Content

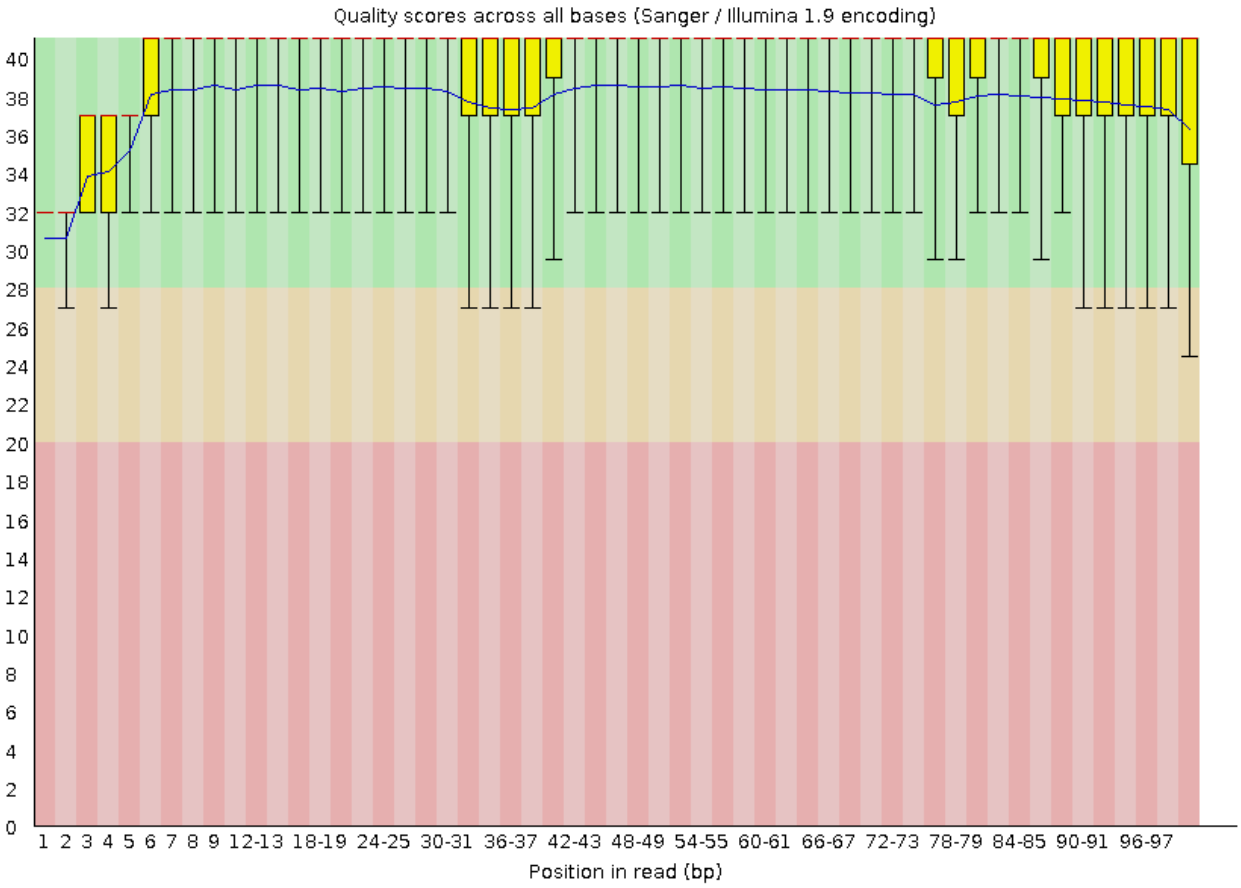


Figure 7: Figure 7: 34_4H Reverse R2 Per Base Quality Distribution

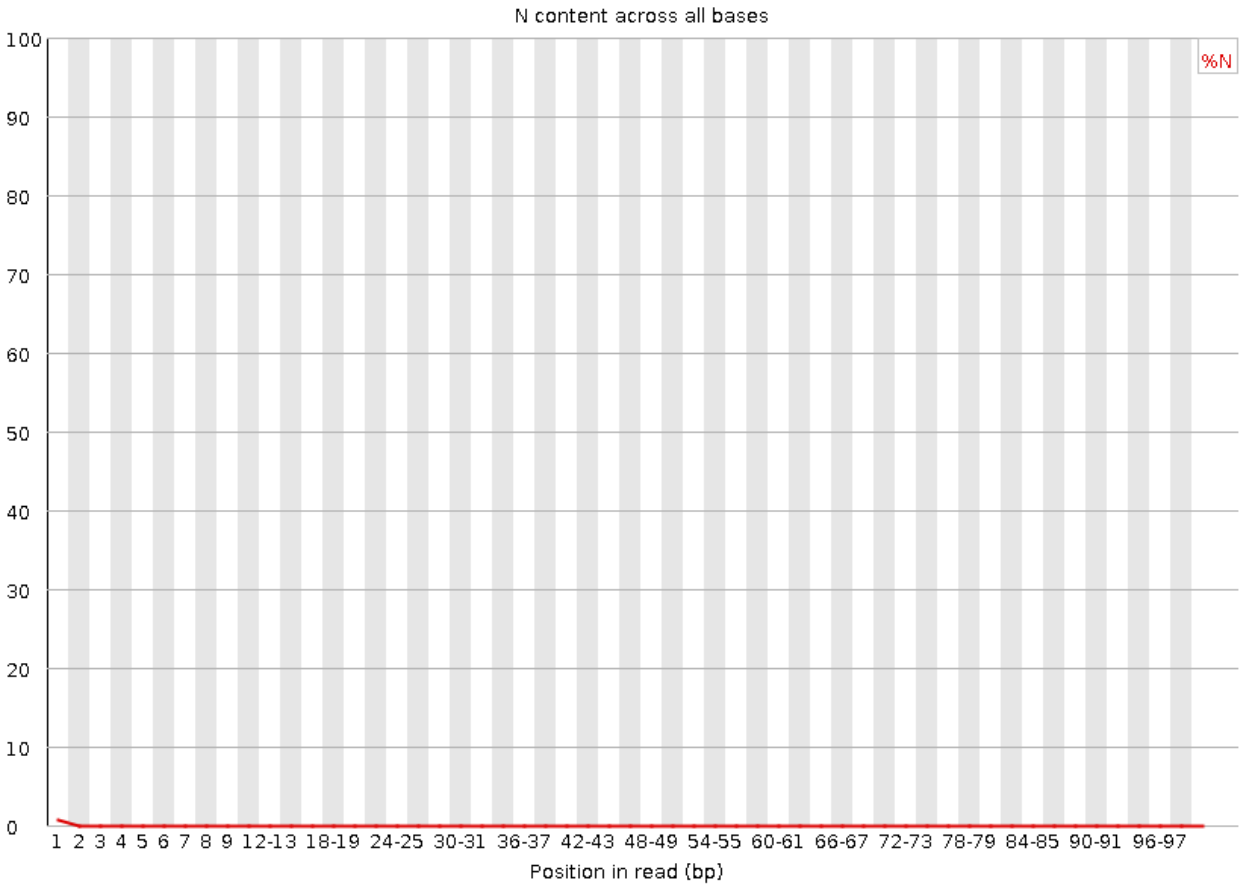


Figure 8: Figure 8: 34_4H Reverse R2 Per Per N Content

Using cutadapt and Trimmomatic softwares to trim adapter sequences and loq quality reads from both paired fastq files

For 21_3G fastq file, the forward R1 reads had a total of 6.6 % of reads trimmed, in comparison to 7.4 % in the reverse R2 fastq file (Summary table 1). While for 34_4H fastq file, the forward R1 file had 9.1% of its reads trimmed, with 9.8 % for the reverse R2 file (Summary table 2).

Summary Table 1: Proportion of Trimmed Reads-21_3G.fastq.gz

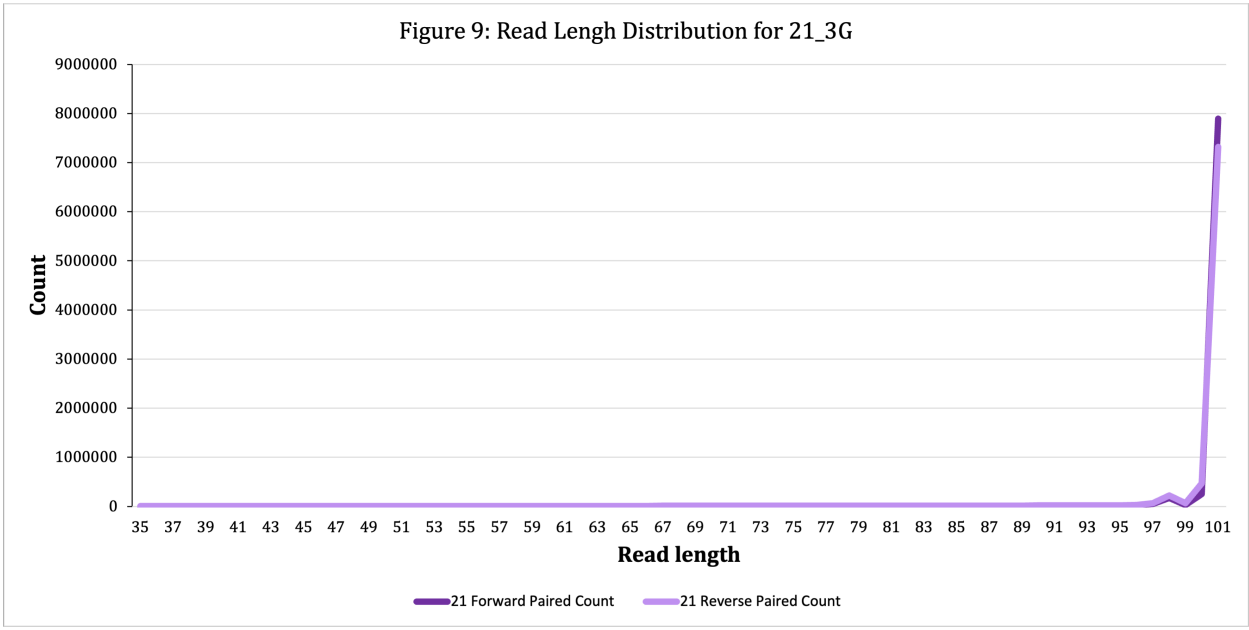
	Number of Reads	Percentage
Read 1 with adapter	613,874	6.6
Read 2 with adapter	679,275	7.4
Total read pairs processed	9,237,299	

Summary Table 2: Proportion of Trimmed Reads-34_4H.fastq.gz

	Number of Reads	Percentage
Read 1 with adapter	819,166	9.1
Read 2 with adapter	886,595	9.8
Total read pairs processed	9,040,597	

Comparing Read Length Distribution Plots after completing Trimming quality procedures for both paired fastq files: R1 Forward vs R2 Reverse

Figure 9 displays the forward R1 and reverse R2 read length distributions after trimming low quality reads from 21_3G fastq file. In comparison to R1 Reverse Paired file, R1 Forward Paired file has more reads with lengths 99 to 101. Given the disparity at read lengths 99–101, this further supports the idea that R2 was cut more severely than R1. Figure 10 displays the forward R1 and reverse R2 read length distributions after trimming loq quality reads from 34_4H fastq file. Overall, R1 Forward paired file has higher reads with length of 99-101 when compared to R2 Reverse Paired file. Given the disparity at read lengths 99–101, this further supports the idea that R2 was trimmed more severely than R1. In contrast to the R1 and R2 file from the 21_3G fastq, the R1 and R2 file from 34_4H fastq file exhibits a smaller trimmed difference between the two paired files. This is further supported as R2 was trimmed 0.8% more than R1 in 21_3G fastq, while R2 was trimmed 0.7% more than R1 in 34_4H fastq (Summary Table 1, 2).



Comparing Read Quality Score Distributions after Trimming adapter sequences and low quality reads

It is typically normal for forward R1 and reverse R2 to be trimmed at different rates. The R2 file displayed worse quality scores across base reads than the forward R1 file when fastqc was applied to the raw data. Because of this, it is anticipated that the bioinformatic tools cuatadpt and trimmomatic will eliminate more flagged low-quality reads from the R2 dataset than the R1 file. The percentages of clipped readings listed in Summary tables 1 and 2 provide more evidence for this. For example, in 21_3G fastq file the forward R1 had a 6.6% of reads trimmed, while reverse R2 had a 7.4%, of reads trimmed; a difference of 0.8% (Figs. 11-14). For 34_4H fastq file, forward R1 had 9.1% of the reads removed compared to 9.8% for R2, with a difference of 0.7% (Figs. 15-18). As observed in fastqc raw files, the 34_4H fastq showed an overall lower quality values with greater variability, which is why it has a higher trimmed percentage when compared to 21_3G fastq.

Determining the number of mapped reads to reference genome for each paired fastq file

Summary tables 3 and 4 provide the number of mapped, un-mapped, and total reads for each SAM file generated from the 21_3G and 34_4H fastq. Aligned 21 Mus musculus.GRCm39.dna SAM file had 91.8% of reads mapped, with 17,061,173 reads aligning to the reference genome out of 18,586,906 reads. In contrast, the aligned 34 Mus musculus.GRCm39.dna SAM file had 92.6% of reads mapped, with 16,822,704 reads aligned to the reference genome out of a total of 18,158,914 reads.

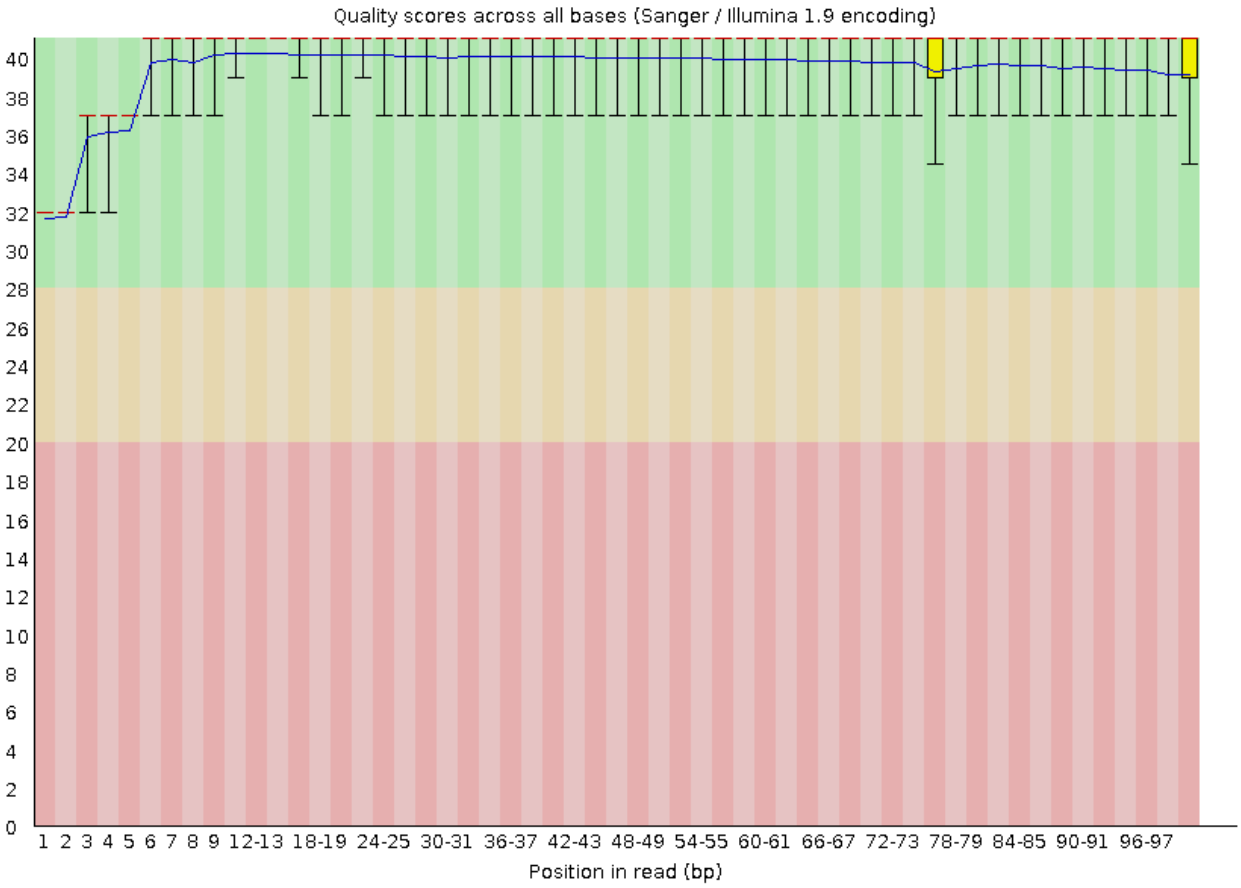


Figure 9: Figure 11: Trimmed 21_3G Forward R1 Per Base Quality Distribution

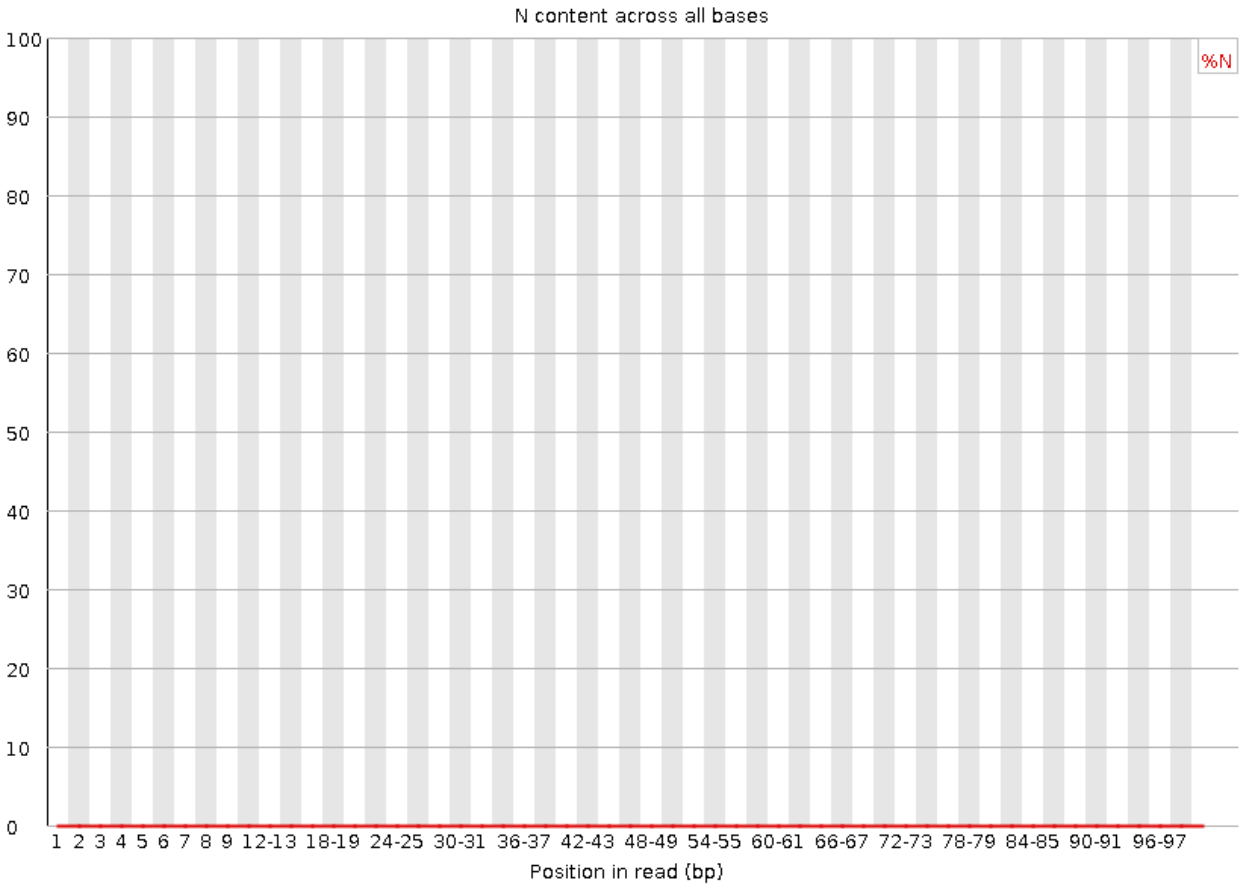


Figure 10: Figure 12: Trimmed 21_3G Forward R1 Per N Content

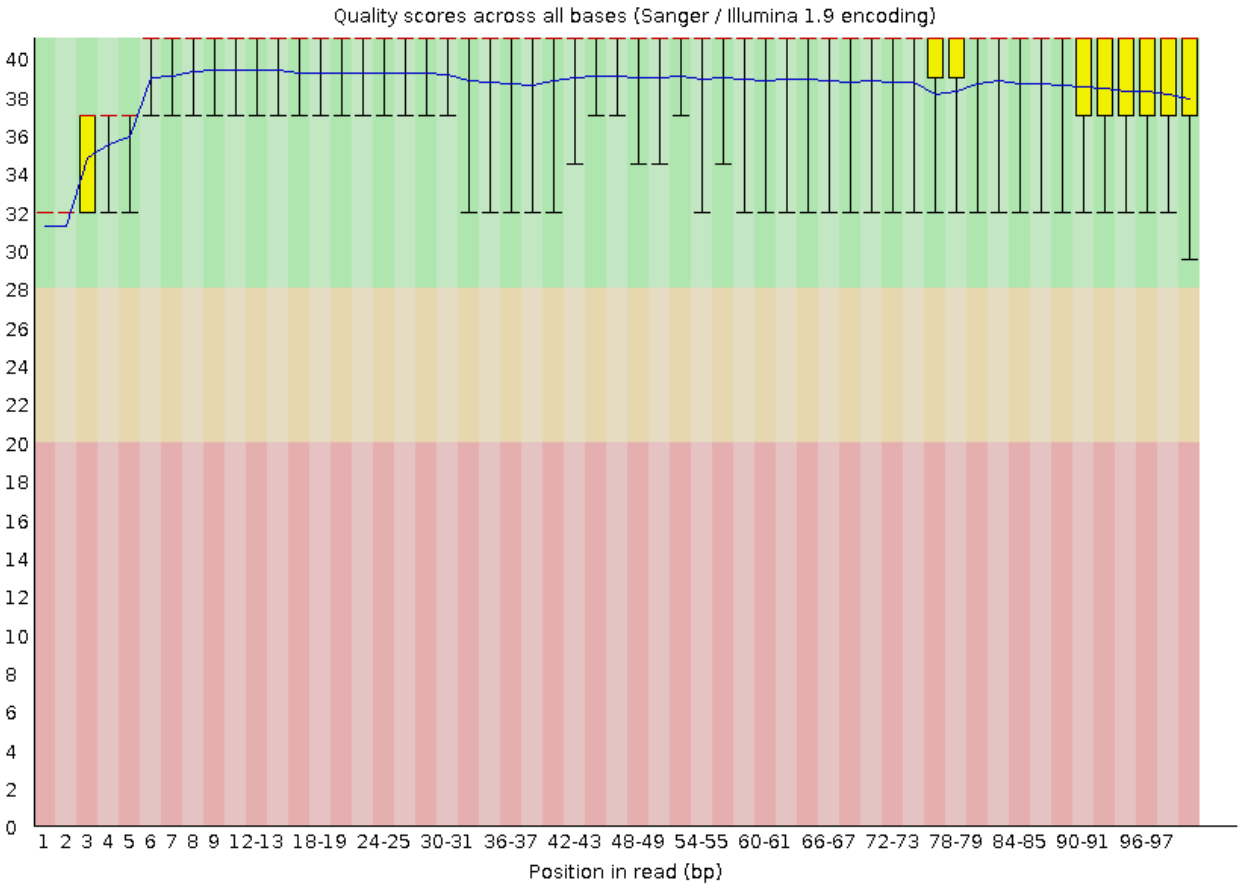


Figure 11: Figure 13: Trimmed 21_3G Reverse R2 Per Base Quality Distribution

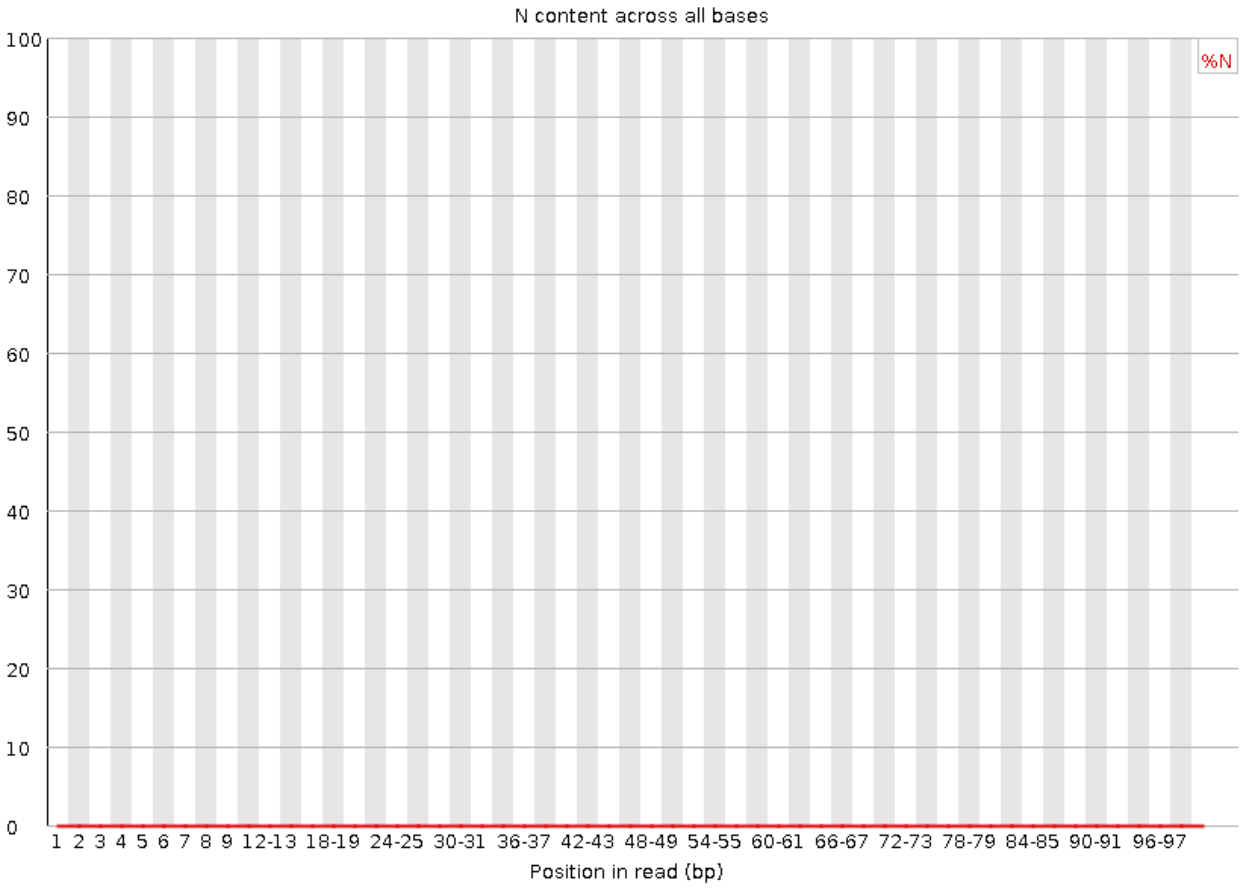


Figure 12: Figure 14: Trimmed 21_3G Reverse R2 Per N Content

Summary Table 3: Number of Mapped and Unmapped Reads in aligned 21_Mus_musculus.GRCm39.dna SAM File

	Number of reads	Percentage
Mapped	17061173	91.8
Un-Mapped	645451	8.2
Total	18586906	

Summary Table 4: Number of Mapped and Unmapped Reads in aligned 34_Mus_musculus.GRCm39.dna SAM File

	Number of reads	Percentage
Mapped	16822704	92.6
Un-Mapped	483578	7.4
Total	18158914	

Demonstrating how the data analyzed are "strand-specific" RNA-Seq libraries.

The fact that more than half of my sequences read in one direction and the other half in the opposite direction leads me to believe that my data is stranded specific. Using the information reported below, for 21_3G fastq, 3.8% were mapped in the forward direction and 81.2% in the reverse direction. For 34_4H fastq, 5.6% were mapped in the forward direction while 83.4% in the reverse direction (Summary table 5,6). These stated figures are consistent with the hypothesis that stranded specific RNA libraries exist when more than 50% of reads for a given value are read in one way over the other.

Summary Table 5: Strand Specific RNA Libraries 21_3G Fastq

Direction of Reads	Number of Reads	Percentage %	Total Number of Reads
Forward	340380	3.8	
Reverse	7181307	81.2	
Other	1331625	15	
			8853312

Summary Table 6: Strand Specific RNA Libraries 34_4H Fastq

Direction of Reads	Number of Reads	Percentage %	Total Number of Reads
Forward	482525	5.6	
Reverse	7214830	83.4	
Other	955786	11	
			8653141