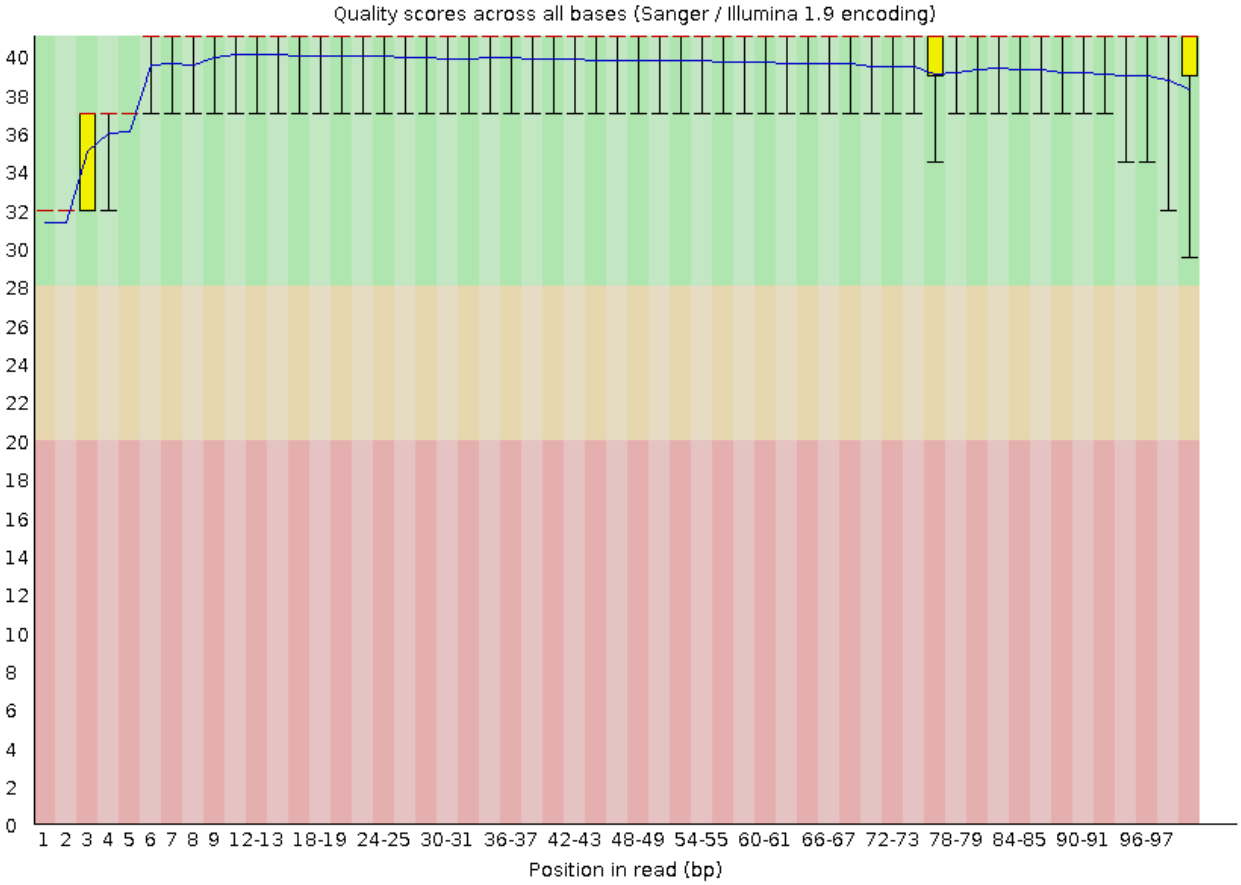
QAA

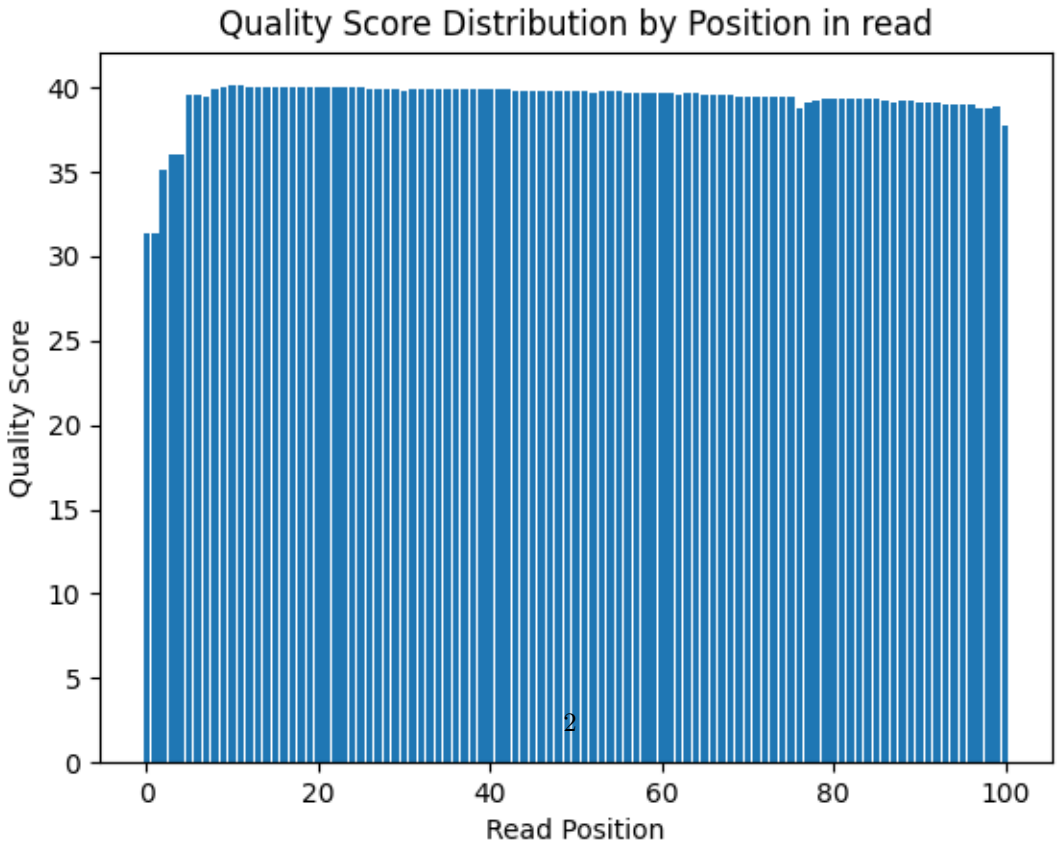
Logan Wallace

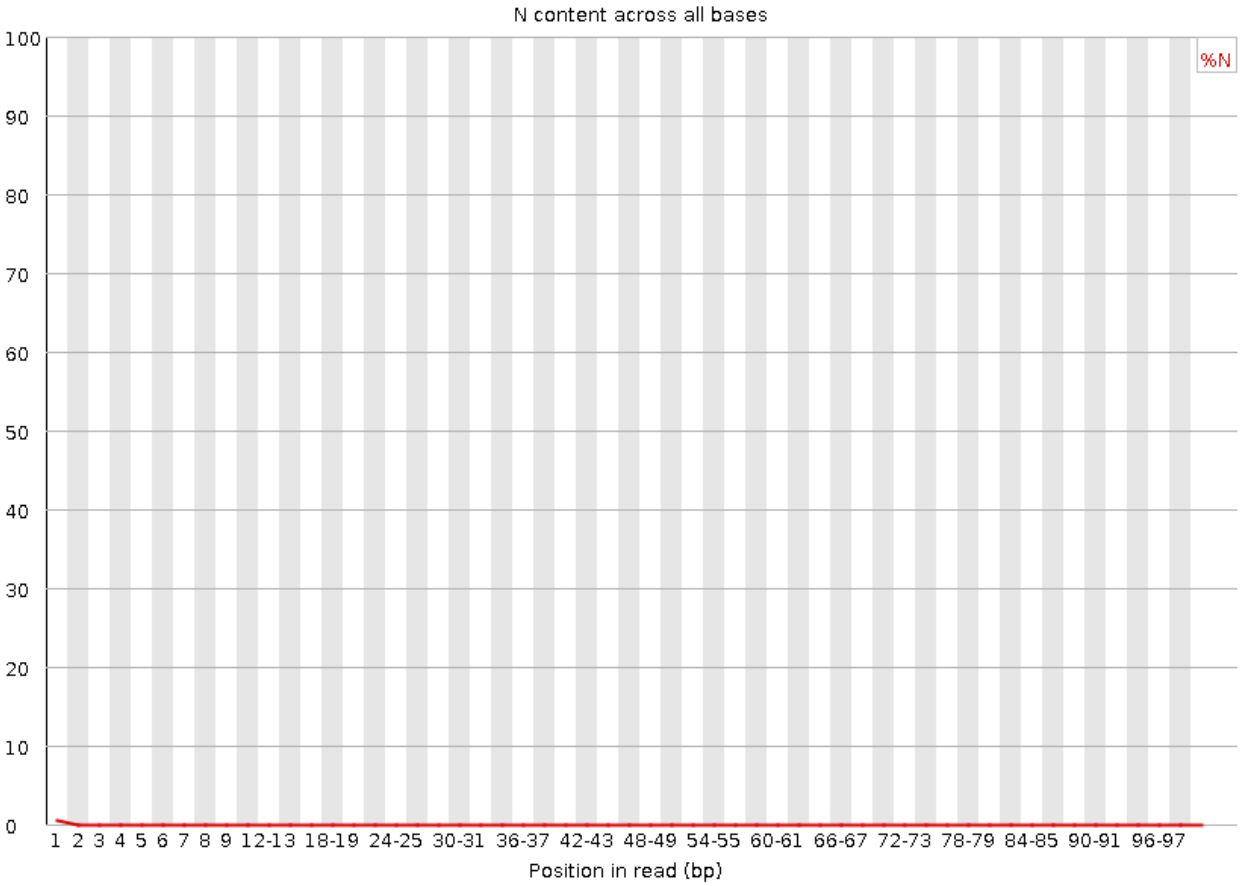
2022-08-31

Part 1 – Read quality and score distributions

3_2B_control_S3_L008_R1_001






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)

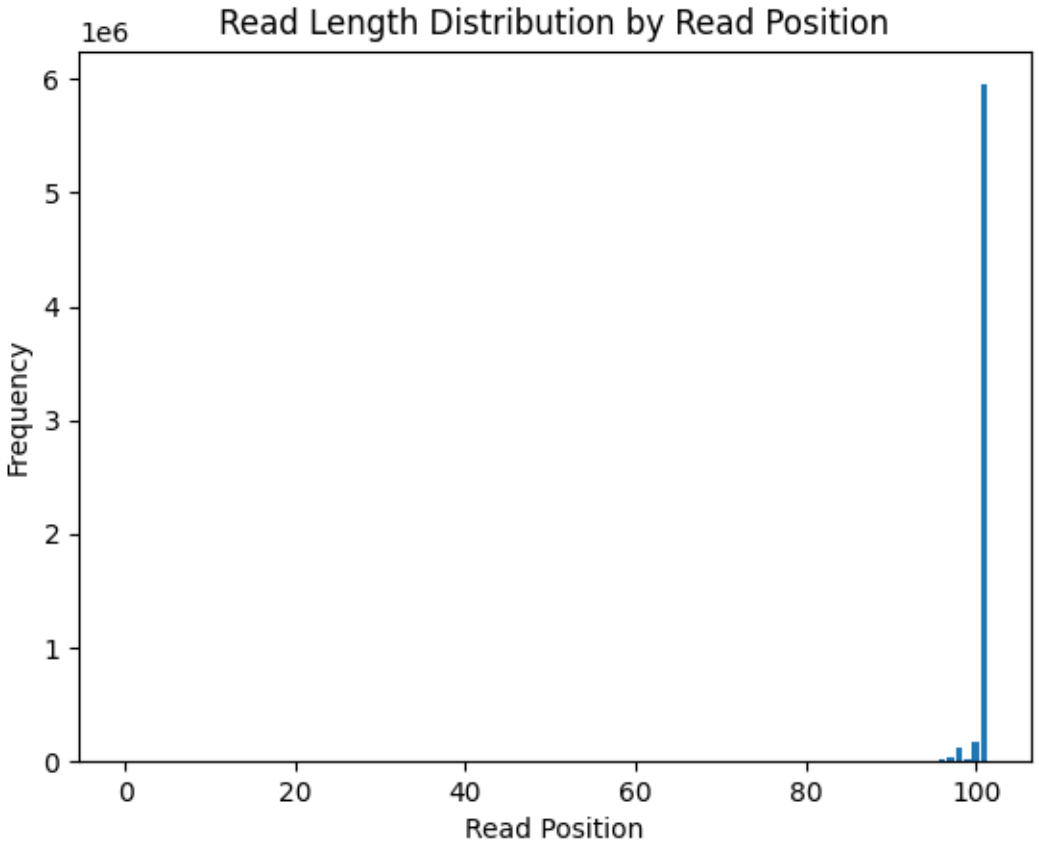
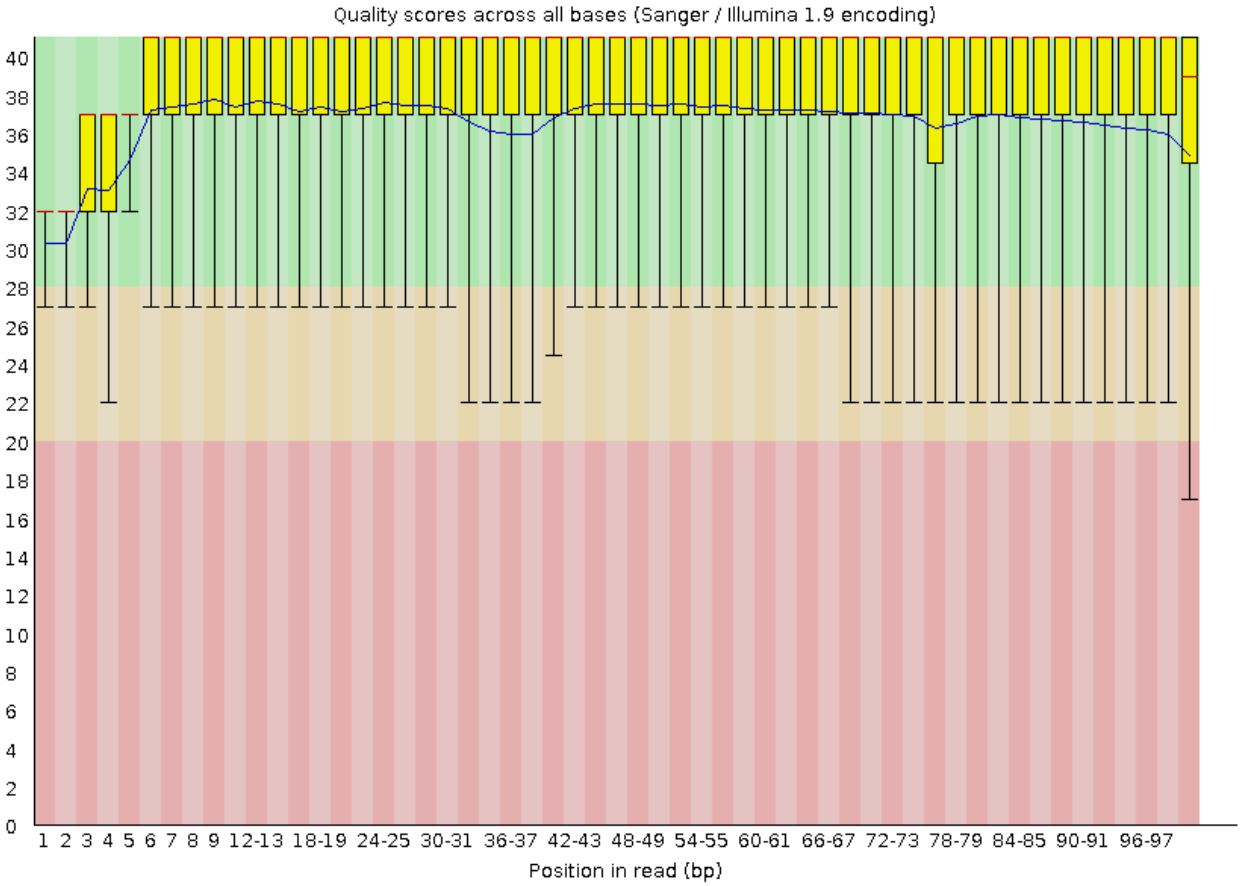
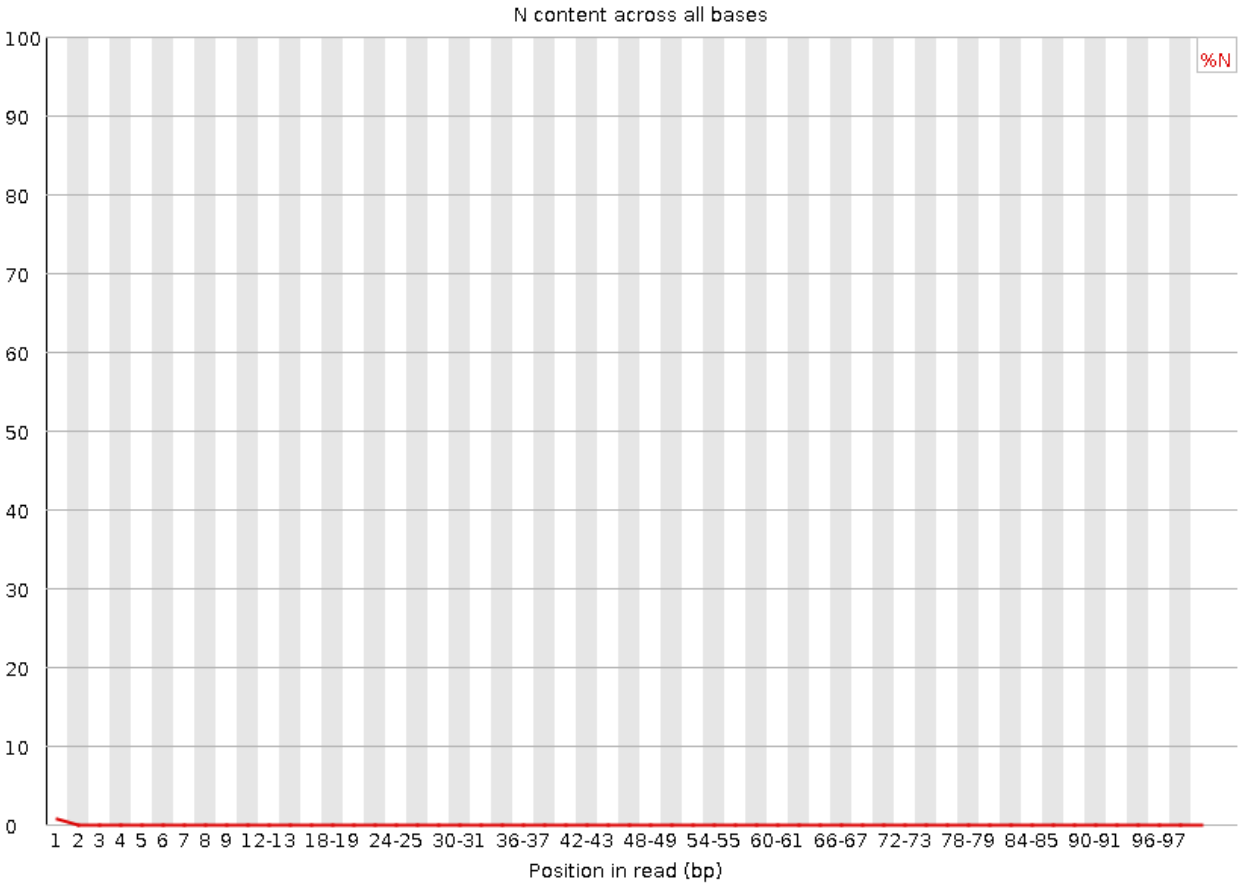


Figure 1. *Control Read 1. A. Quality score distribution. B. Quality score distribution C. Number of 'N's called across the read. D. Read length distribution.*

3__2B__control__S3__L008__R2__001




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)

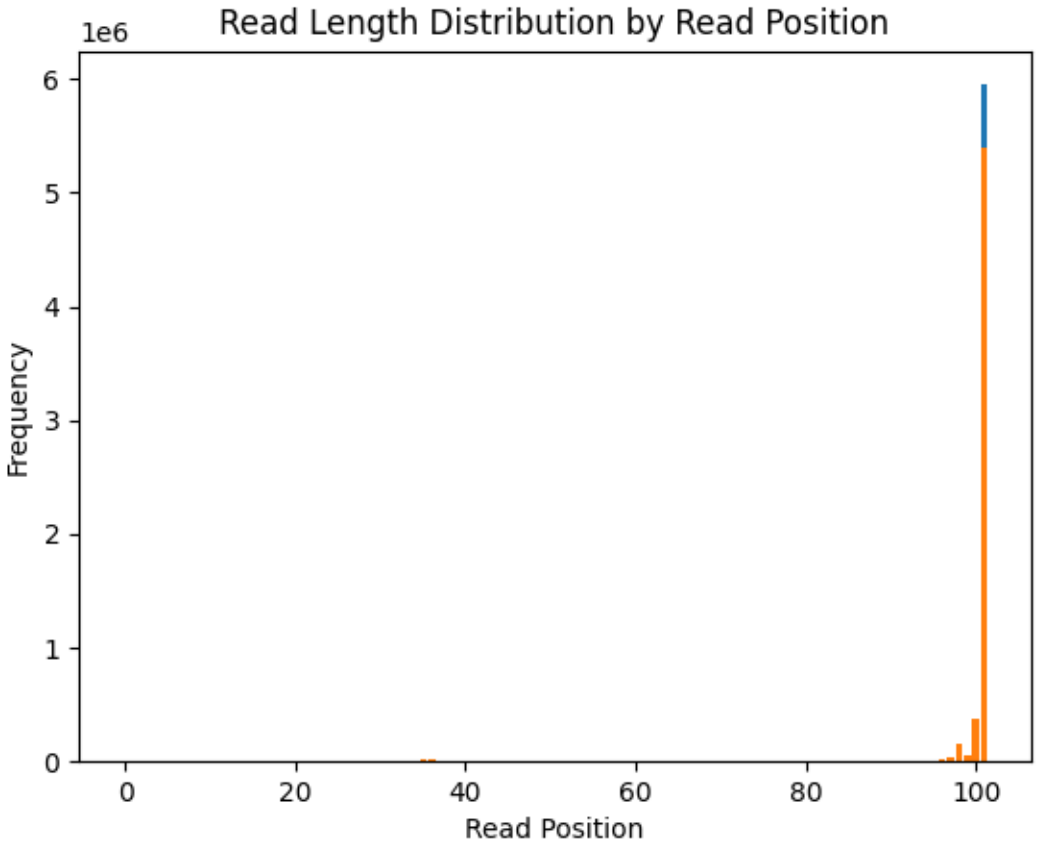
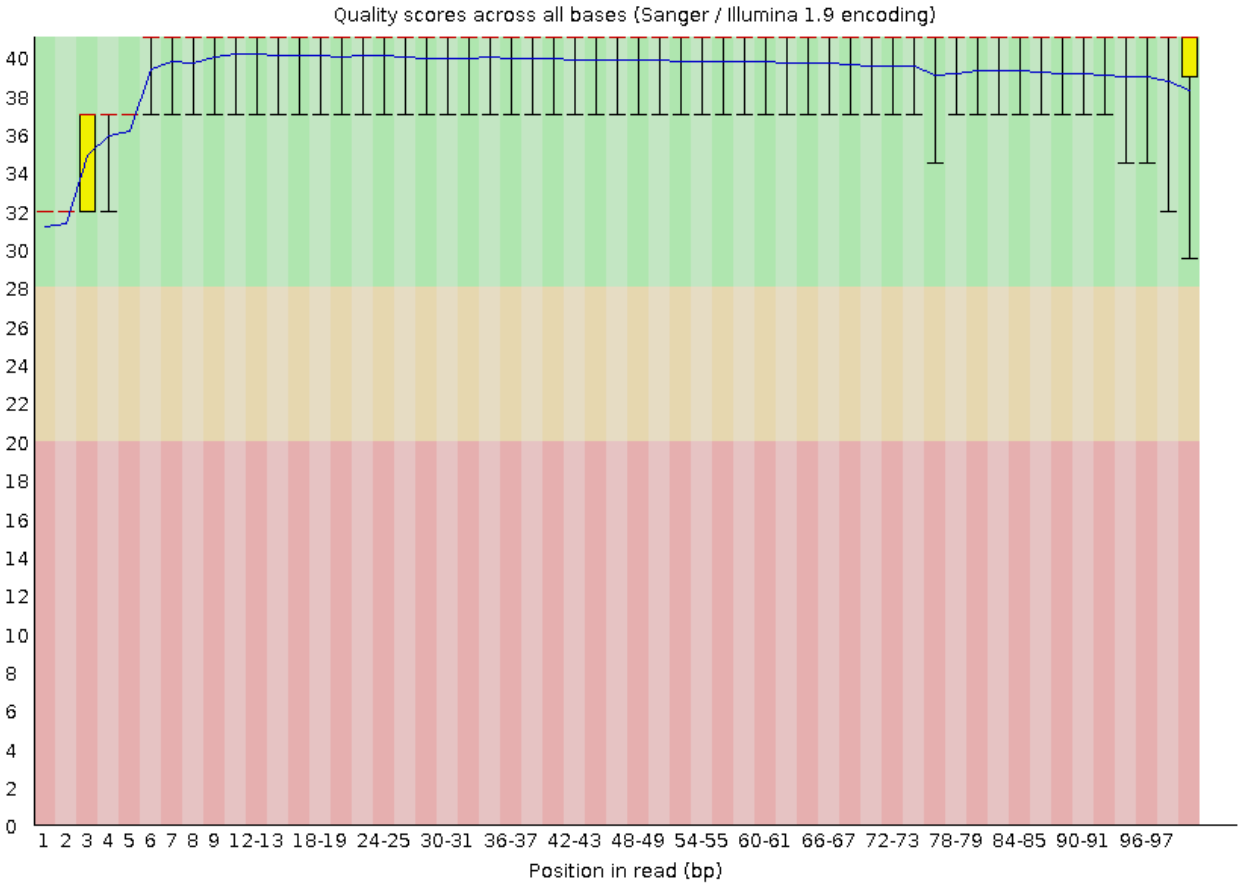
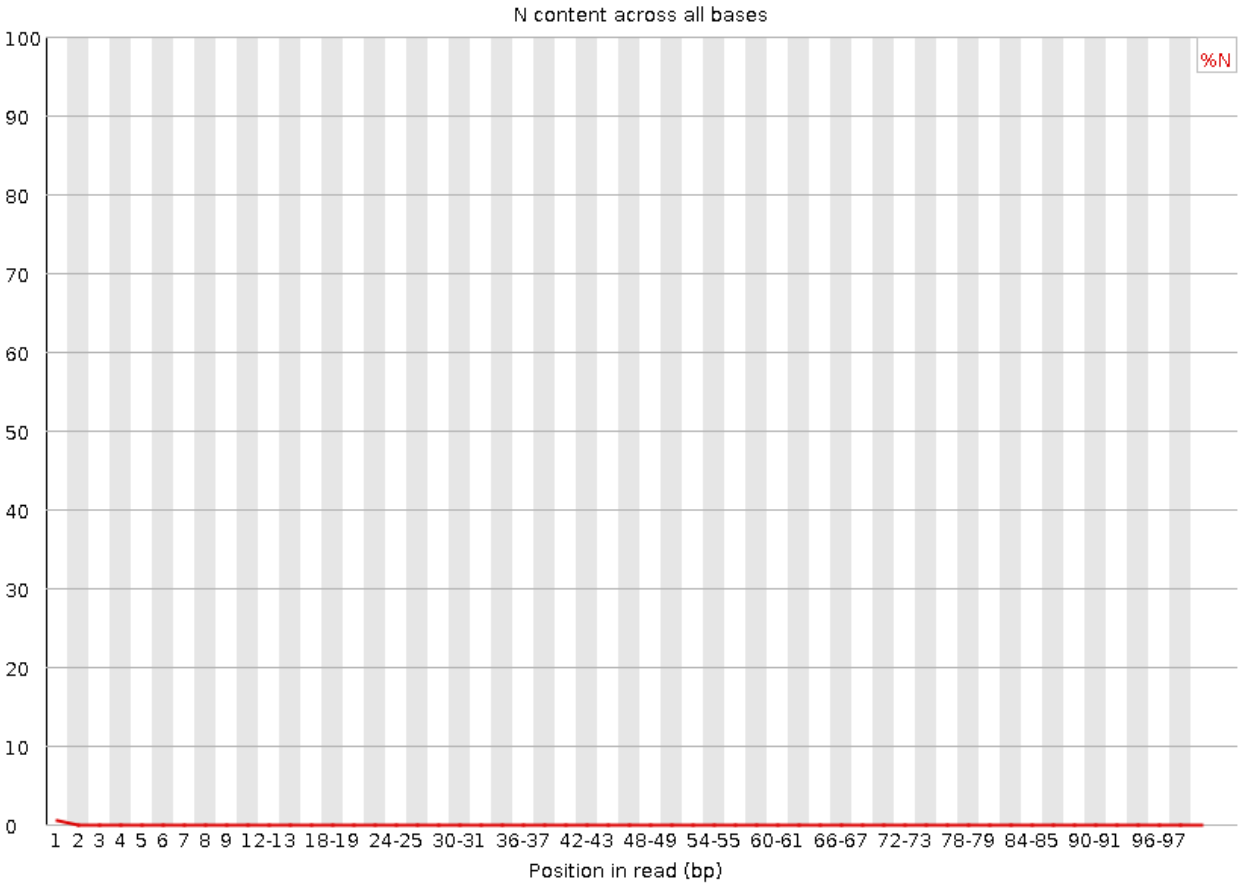


Figure 2. *Control Read 2. A. Quality score distribution. B. Quality score distribution C. Number of 'N's called across the read. D. Read length distribution.*

17__3E__fox__S13__L008__R1__001




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)

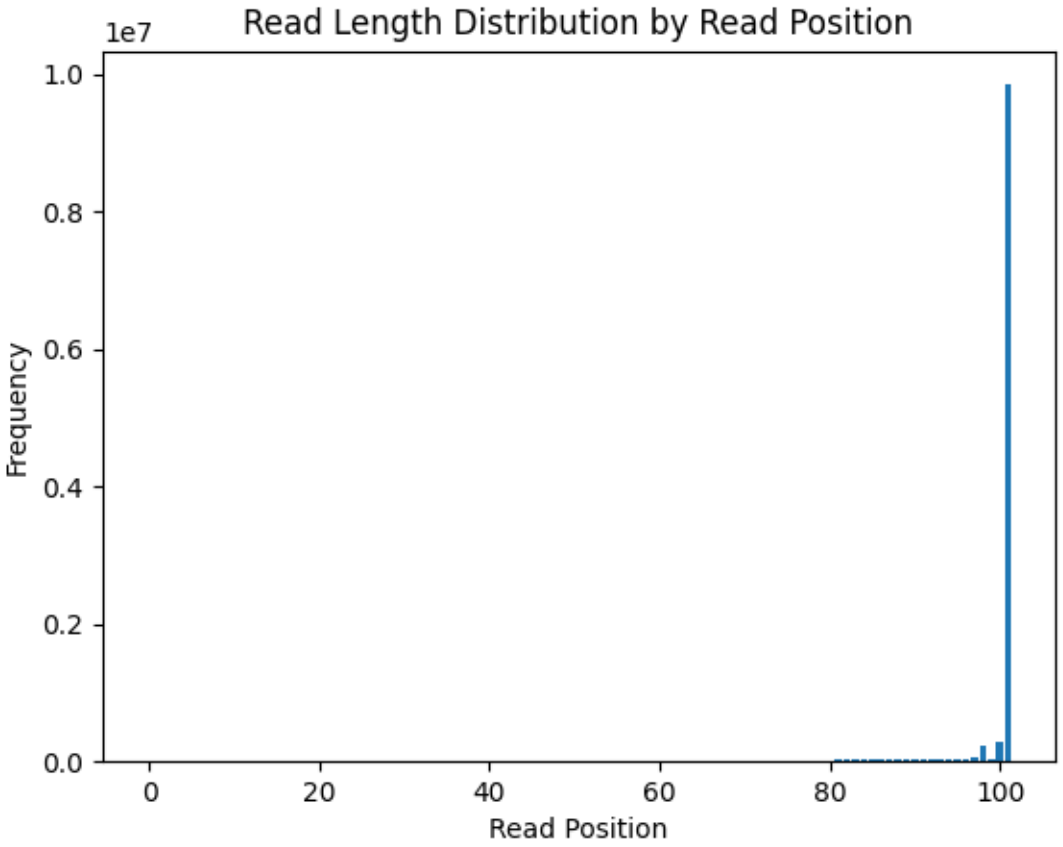
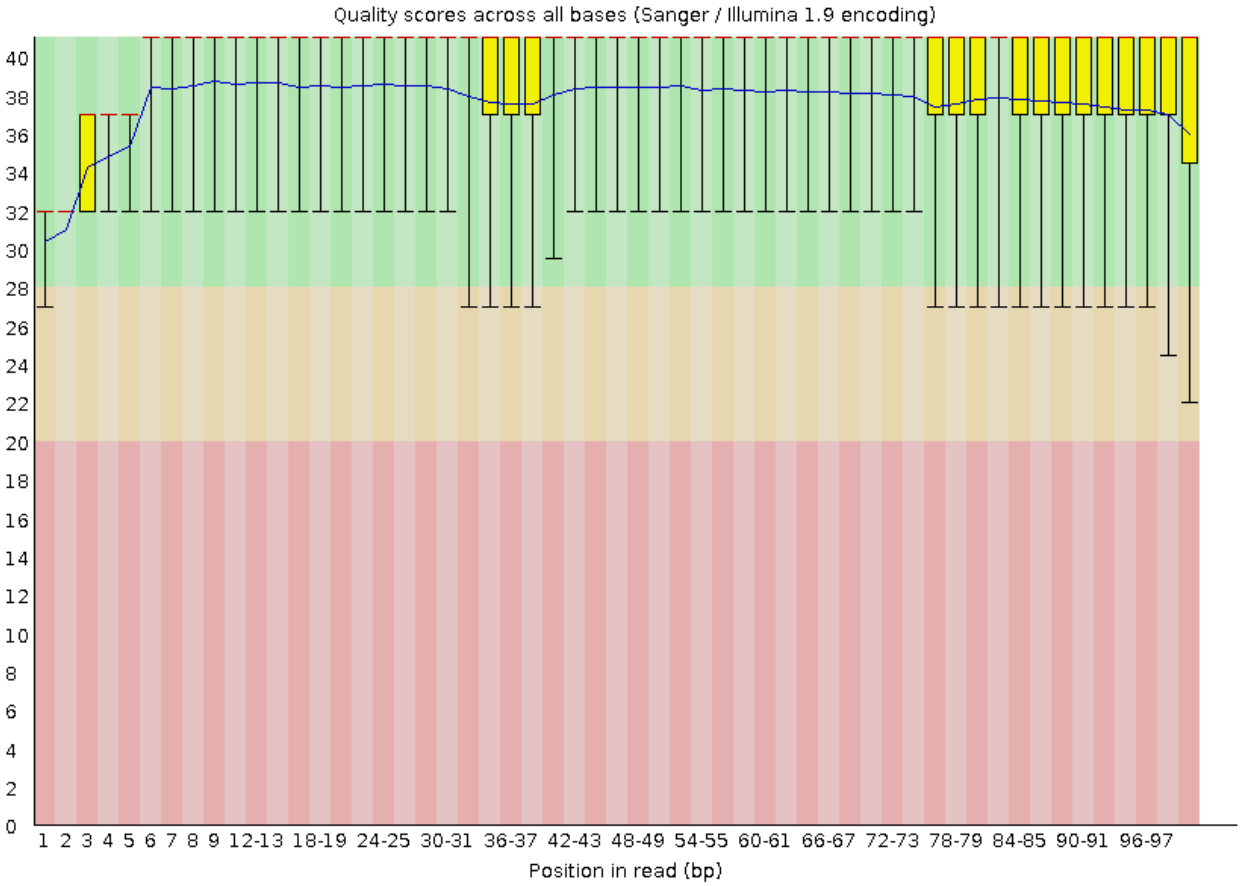
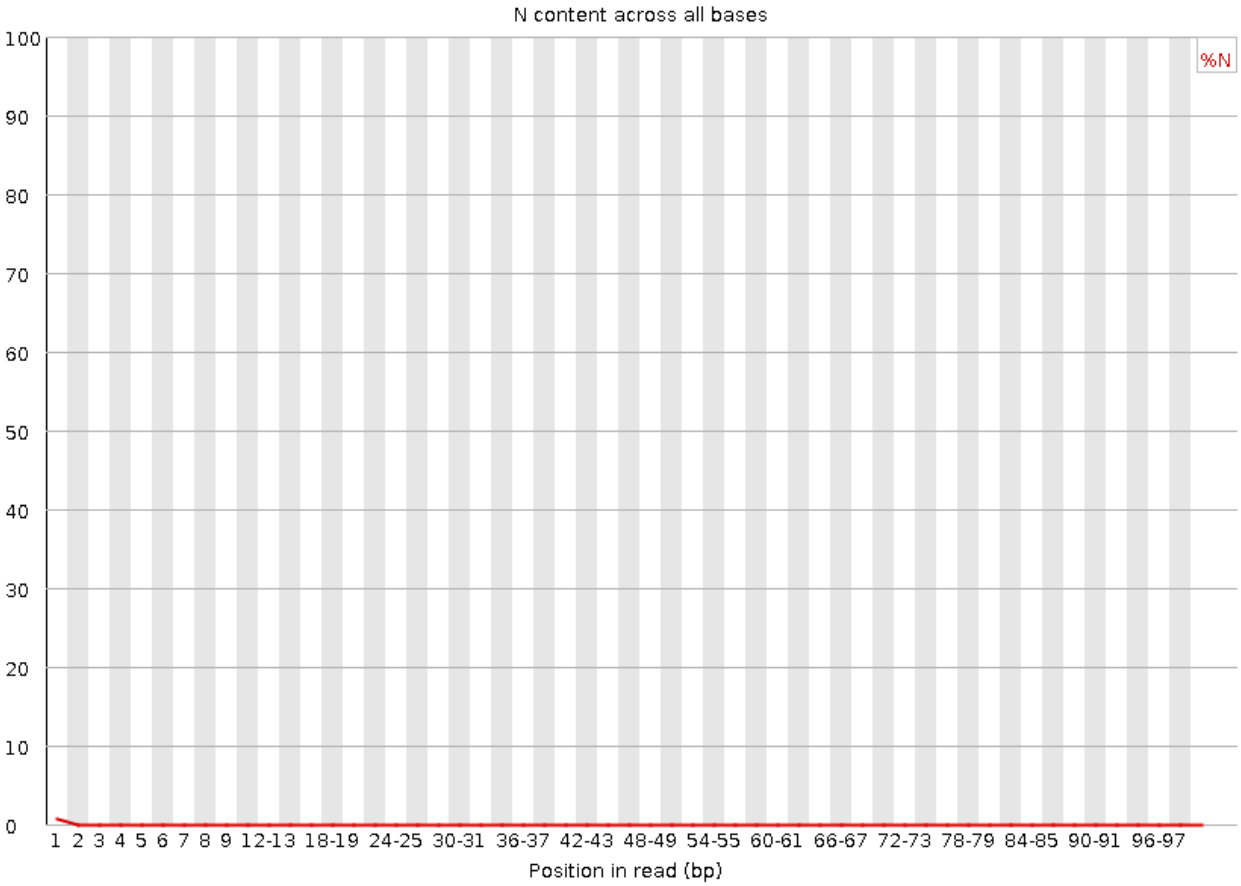


Figure 3. *Fox Read 1. A. Quality score distribution. B. Quality score distribution C. Number of 'N's called across the read. D. Read length distribution.*

17__3E__fox__S13__L008__R2__001




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)

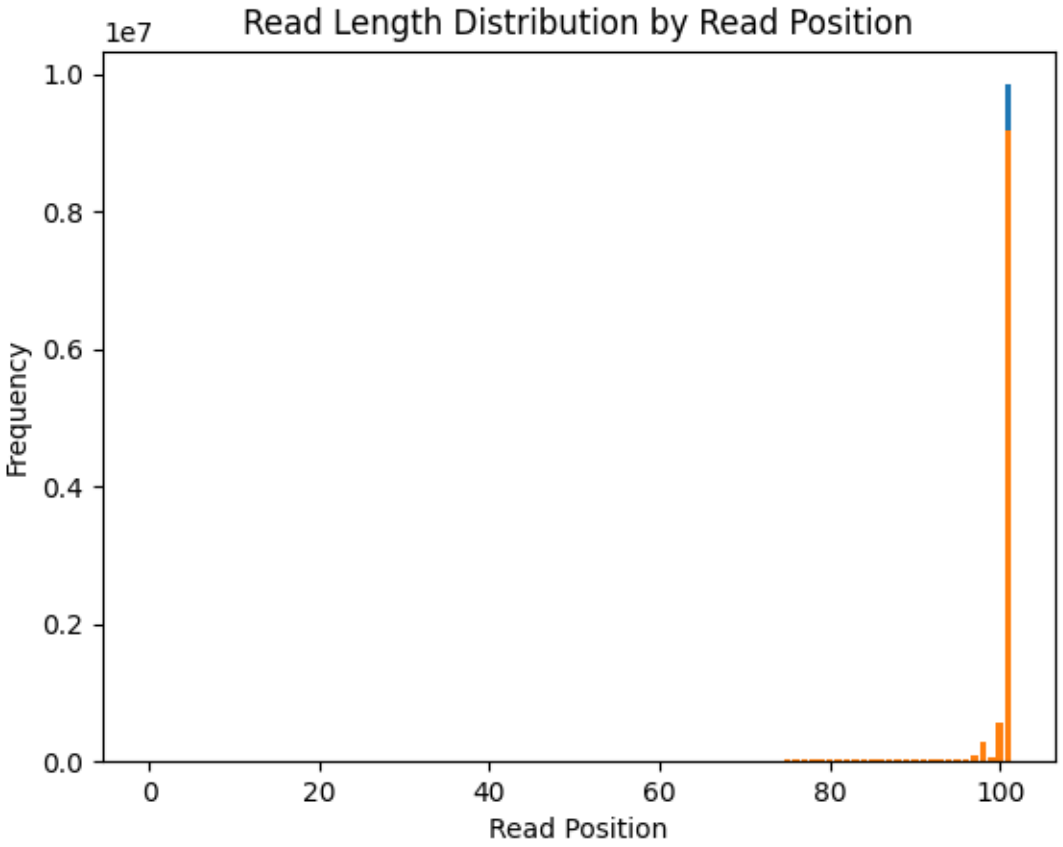


Figure 4. *Fox Read 2. A. Quality score distribution. B. Quality score distribution C. Number of 'N's called across the read. D. Read length distribution.*

Are the plots of N's per base consistent with they other data?

Yes, we see that the highest frequency of N's called are all within the first several bases of the run and this is consistent with the lowest q-scores for each run.

Comparison of output from our script and fastQC

First, when comparing the results between my output and the output from fastQC they are very similar! We see drops in the Q-score in identical base pair positions and the scores themselves are near identical. Second, the runtime did significantly differ. My script took several minutes to run and the fastQC took only moments. I don't have any insight into how the actual code differs, however fastQC was written by a team of experienced authors and my code was written by a masters student with minimal experience under a time constraint.

How is the overall quality of the two libraries

Overall, the quality of the libraries seem quite good with average q-scores well above 30. 3_2B_control_S3_L008_R2_001, seems to have the lowest average quality scores.

Part 2 – Adaptor Trimming Comparison

5. What proportion of reads (both R1 and R2) were trimmed?

Reads Trimmed

stranded arg	3_2B_control	17_3E_fox_S13
Total Pairs	6,873,509	11,784,410
Read 1	219,477 (3.2%)	1,024,588 (8.7%)
Read 2	268,119 (3.9%)	1,104,503 (9.4%)

Table 1. *Table showing data from trimmomatic. Number of reads trimmed total and from each of read one and read two.*

UNIX SKILLZ $ zcat 3_2B_control_S3_L008_R1_001.fastq.gz | grep 'AGATCGGAAGAGCACACGTCT-GAACTCCAGTCA' | wc -l 7659 $ zcat 3_2B_control_S3_L008_R1_001.fastq.gz | grep 'AGATCGGAA-GAGCGTCGTGTAGGGAAAGAGTGT' | wc -l 0 Explaining the above, I ran a grep for the adapter sequences in question for each of the files (Read1 and Read2) for both the 'control' and 'fox' datasets. Above you'll see, as expected, we only observed the R1 adapters within the R1 reads vice versa for R2.

7. Read length distribution for R1 and R2.

I do not anticipate that trimming should occur differently between reads one and two. This is upheld by the agreement between our plots for the read length distribution (Figures 1-4(d)). We see that a very high fraction of the reads are at 101 for both read one and read two.

Part 3 – Alignment and Strand Specificity

11. Demonstrate convincingly whether or not the data are from stranded libraries.

More Unix skillz $ cat htseq_count_control_rev | grep ^'ENSMUSG' | cut -f 2 | paste -sd+ | bc

HTSEQ - Read Counts

stranded arg	3_2B_control	17_3E_fox_S13
YES	245,058	5,260,739
REVERSE	443,307	8,952,669

Table 2. *The above table shows the read counts mapped for each dataset as counted by htseq-count using the -s argument. Does the read map to the same or a different strand?*

Using the above bash command, I summed the total reads mapped to a feature and derived a percentage of reads mapped in either direction. We can infer that the library was stranded because of the discrepancy in percent of reads mapped to features on either strand. For example, we observed approximately 20x more reads mapped to features when using -reverse (8,952,669) than when using -yes (443,307) for our Fox set of RNA-seq data.