Quality Analysis of `8_2F_fox_S7_L008` and `22_3H_both_S16_L008`

Lauren Williams

09-10-2024

Part 1 – Read Quality Score Distributions

The overall data quality of these two libraries are of high enough quality to use for further analysis.

Firstly, all per-base quality scores for reads 1 and 2 in both libraries were above 30, as seen in Figures 1 and 2. While the quality scores are both R2s are lower, this is to be expected as R1 is sequenced first and the DNA may be slightly degraded by the time R2 is sequenced. The `FastQC` quality score distribution plots compare very closely to the quality score distribution plots created using my personal quality score plotting script as seen in Figures 1 and 2. The line trends very similarily, but the `FastQC` plots have more detail about what is classified as "good quality" as well as error bars. `FastQC` also had a runtime that was about half of the runtime of my personal quality score plotting script. This aligns with what I would expect as my personal quality score plotting script was not optimized for runtime.

Additionally, the `FastQC` output reported that neither read 1 or read 2 in the `22_3H_both_S16_L008` library had overrepresetned sequences. For the `8_2F_fox_S7_L008` library, the `FastQC` output reported that read 1 did not have overrepresetned sequences, but read 2 did.

Furthermore, the data in both of these libraries has low per-base N content as seen in Figure 3. On all the graphs in Figure 3, there is a slight uptick at the first base pair, which is consistent with the lower quality seen at the first base pair on the quality score plots.

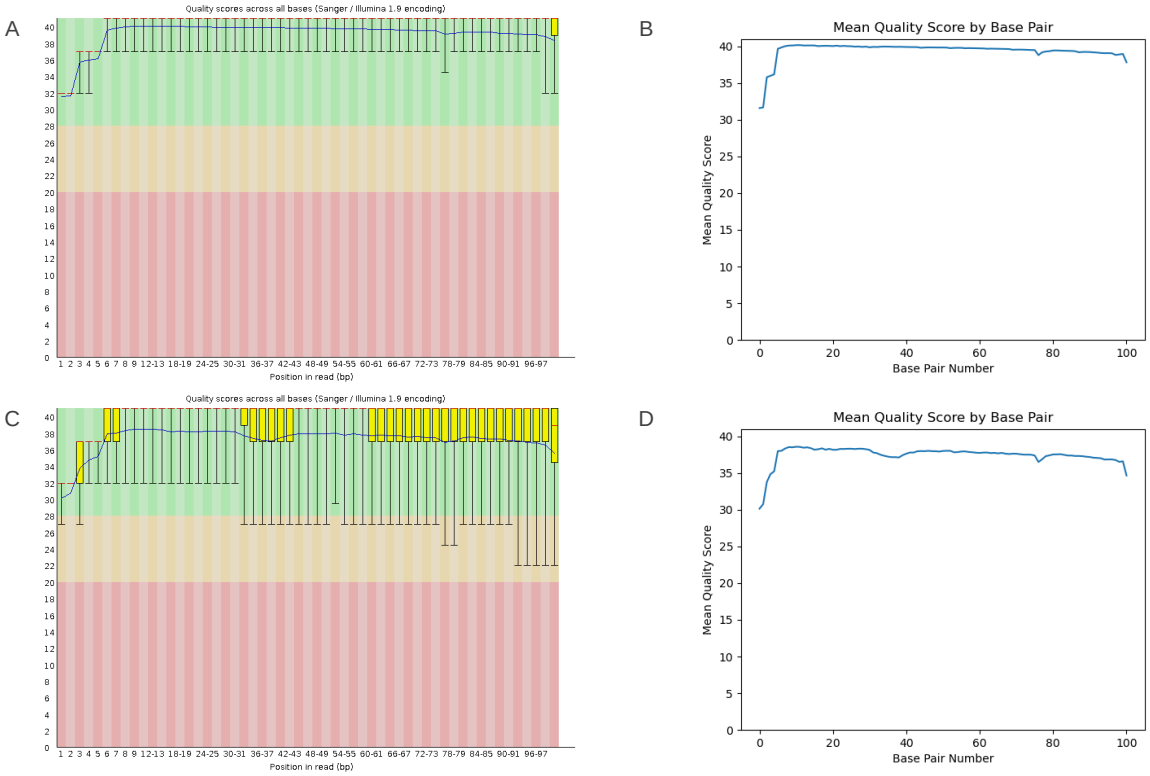

Figure 1: Per-base quality score distributions for R1 and R2 reads of 8_2F_fox_S7_L008. Quality of R1 reads were calculated per base pair and plotted using `FastQC` (**A**) and my personal quality score plotting script (**B**). Quality of R2 reads were calculated per base pair and plotted using `FastQC` (**C**) and my personal quality score plotting script (**D**).

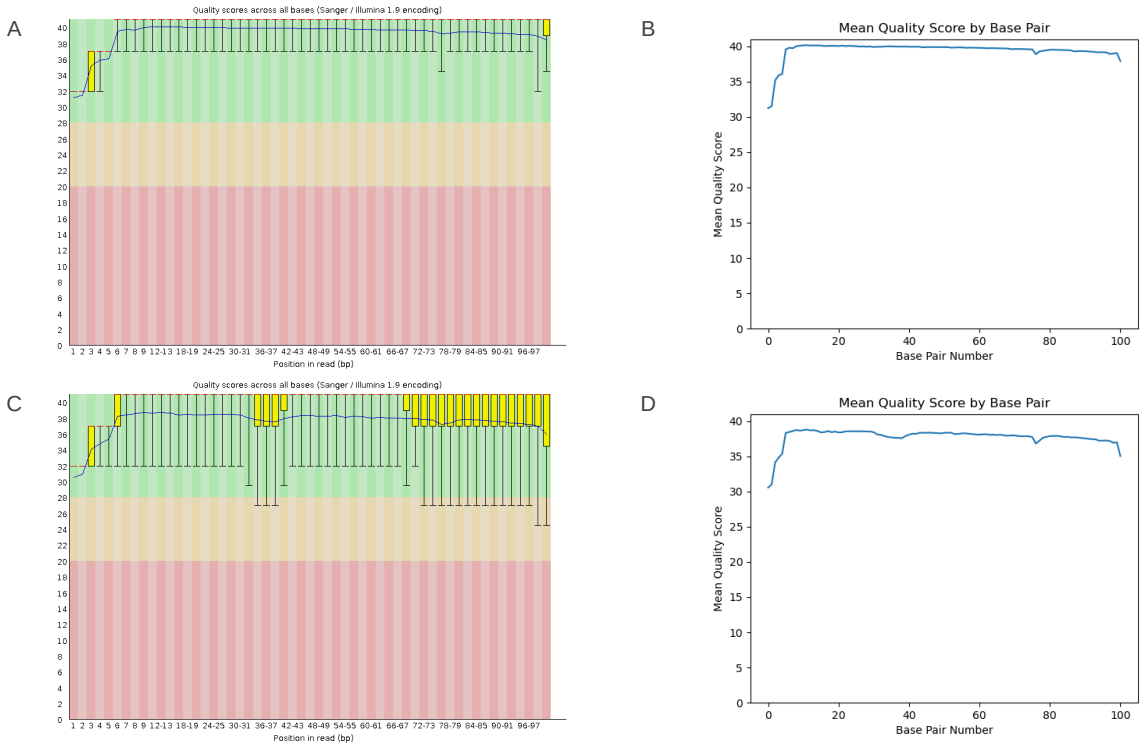

Figure 2: Per-base quality score distributions for R1 and R2 reads of 22_3H_both_S16_L008. Quality of R1 reads were calculated per base pair and plotted using `FastQC` (**A**) and my personal quality score plotting script (**B**). Quality of R2 reads were calculated per base pair and plotted using `FastQC` (**C**) and my personal quality score plotting script (**D**).

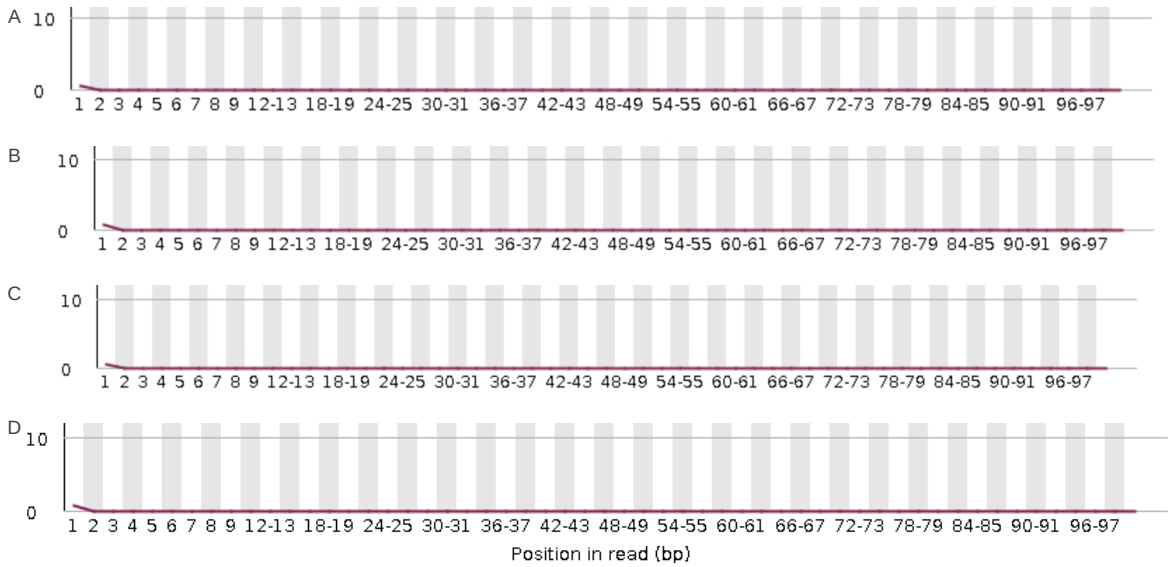

Figure 3: Per-base N content for R1 and R2 reads of 8_2F_fox_S7_L008 and 22_3H_both_S16_L008. (**A**) R1 and (**B**) R2 reads of 8_2F_fox_S7_L008. (**C**) R1 and (**D**) R2 reads of 22_3H_both_S16_L008. All plots were calculated and plotted using `FastQC`.

Part 2 – Adaptor Trimming Comparison

Before preforming adapter trimming, I verified that the R1 files contained the R1 adapter, but not the R2 adapter. I also verified that the R2 files contained the R2 adapter, but not the R2 adapter. To do this I used the following unix commands: `zcat <file_name> | grep "AGATCGGAAGAGCACACGTCTGAACTCCAGTCA" | wc -l` and `zcat <file_name> | grep "AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT" | wc -l`. I also used these commands to verify that the files no longer contained their respective adapter AFTER using the `cutadapt` tool.

While only a small portion of of R1 and R2 reads were trimmed, read 2 had a high percentage of trimmed reads for both libraries (Table 1). Furthermore, Figure 4 shows that R2s are trimmed more extensively than R1s for both libraries. One would expect R1s and R2s to be adapter-trimmed at different rates because the quality of R1 and R2 data is different by nature of sequencing.

Table 1. Proportion of trimmed reads.

	8__2F__fox__S7__L008	**22__3H__both__S16__L008**
Total read pairs processed:	36,482,601	4,050,899
Read 1 with adapter:	2,145,600 (5.9%)	153,089 (3.8%)
Read 2 with adapter:	2,403,490 (6.6%)	186,534 (4.6%)

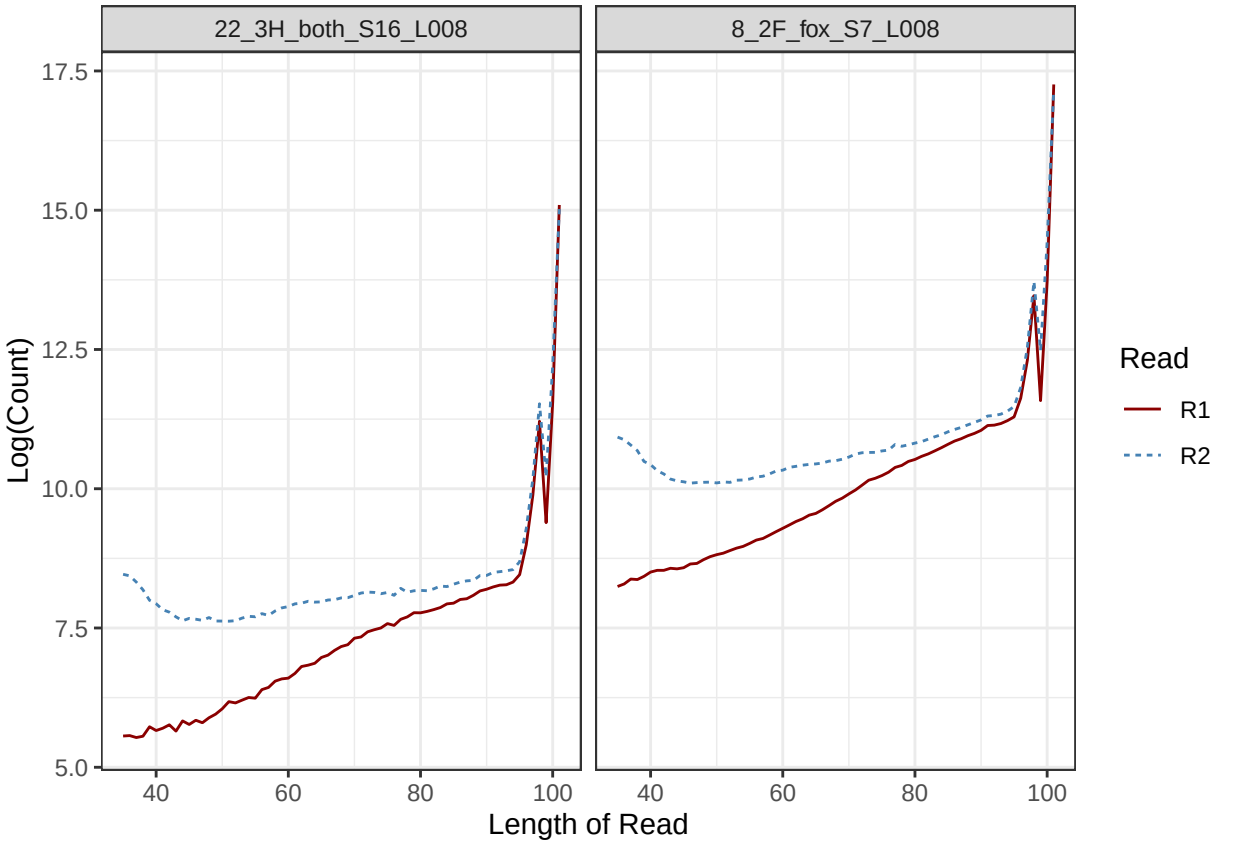

Figure 4: Trimmed read length distributions. Data from both libaries were trimmed using `cutadapt` and `trimmomatic` and then plotted using R.

Part 3 – Alignment and Strand-specificity

I propose that the `8_2F_fox_S7_L008` data are strand-specific, because 80.59% of the reads are mapped stranded reverse, as opposed to 3.46% mapped stranded forwards. Furthermore, I propose that the `22_3H_both_S16_L008`data are also strand-specific, because 87.00% of the reads are mapped reverse, as opposed to 3.71% mapped forwards. Because these values are different, we know that these data are strand-specific. If these values were the same, strandedness would not matter and it would map the same number of reads forwards and backwards.

The difference in percent of reads mapped from my parse script vs the htseq-count script is due to the additional classification of reads as no_feature, ambiguous, too_low_aQual, not_aligned, alignment_not_unique. However, total number of reads from the HTSEQ count are exactly half of the total number of reads from my personal mapping script.

Table 2. Number of Reads for `8_2F_fox_S7_L008` Library.

	My personal mapping script	HTSEQ - stranded	HTSEQ - reverse stranded
Mapped Reads:	67,070,894	1,205,384	28,037,914
Unmapped Reads:	2,511,420	33,585,773	6,753,243
Total Number of Reads:	69,582,314	34,791,157	34,791,157
Percent of reads mapped:	**96.39%**	**3.46%**	**80.59%**

Table 3. Number of Reads for `22_3H_both_S16_L008` Library.

	My personal mapping script	HTSEQ - stranded	HTSEQ - reverse stranded
Mapped reads:	7,677,917	144,728	3,393,906
Unmapped reads:	124,337	3,756,399	507,221
Total number of reads:	7,802,254	3,901,127	3,901,127
Percent of reads mapped:	**98.41%**	**3.71%**	**87.00%**

CHALLENGE: Difference Between Trimmed and Untrimmed Data

While it maybe difficult to observe the slight increase in quality score for the R1 plots (Figure 5 and 7, A and B), there is a more noticeable increase in quality score for both R2 plots (Figure 6 and 8, A and B). The R2 reads were previously the lower quality reads and R2s were trimmed more extensively than R1s for both libraries, so increasing the quality of these reads by trimming the data really highlights the importance of this step in data cleaning. Furthermore, the slight uptick at the first base pair in the untrimmed N content plots, which is consistent with the lower quality seen at the first base pair on the untrimmed quality score plots, does not occur at the first base pair in the trimmed N content plots (Figures 5-8, C and D). This aligns with idea that trimming your data makes it higher quality.

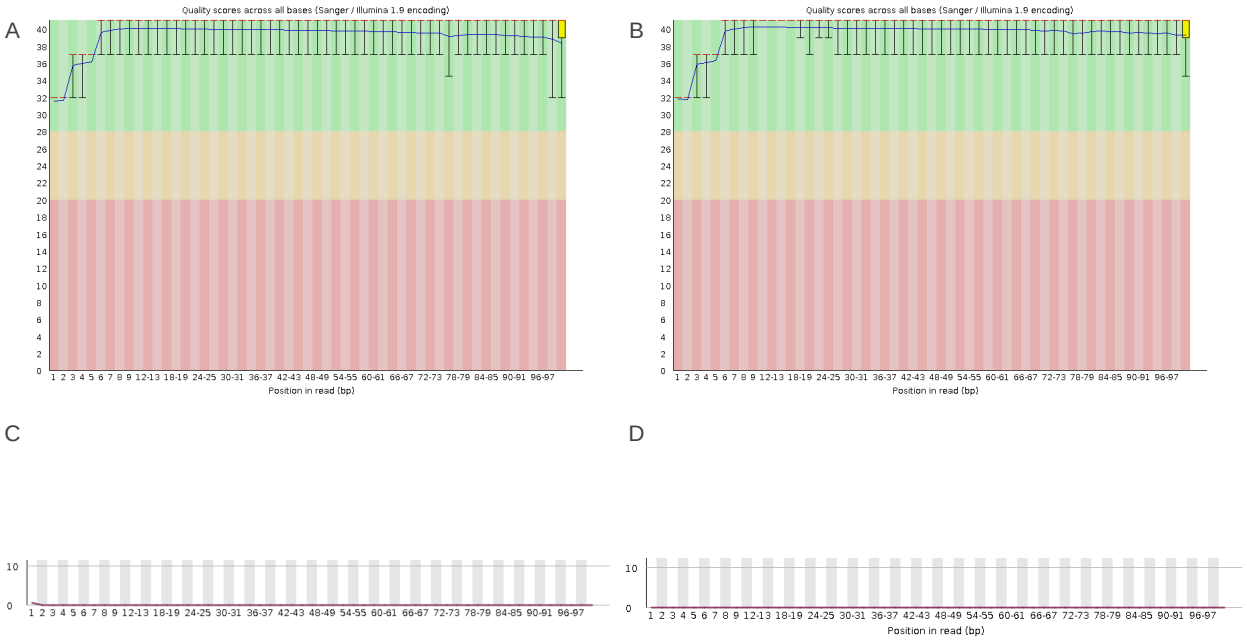

Figure 5: Per-base quality score and n-content distributions for trimmed and untrimmed R2 reads of 8_2F_fox_S7_L008 Quality of untrimmed R1 reads (**A**) and trimmed R1 reads (**B**) as well as N content of untrimmed R1 reads (**C**) and trimmed R1 reads (**D**) were calculated per base pair and plotted using `FastQC`

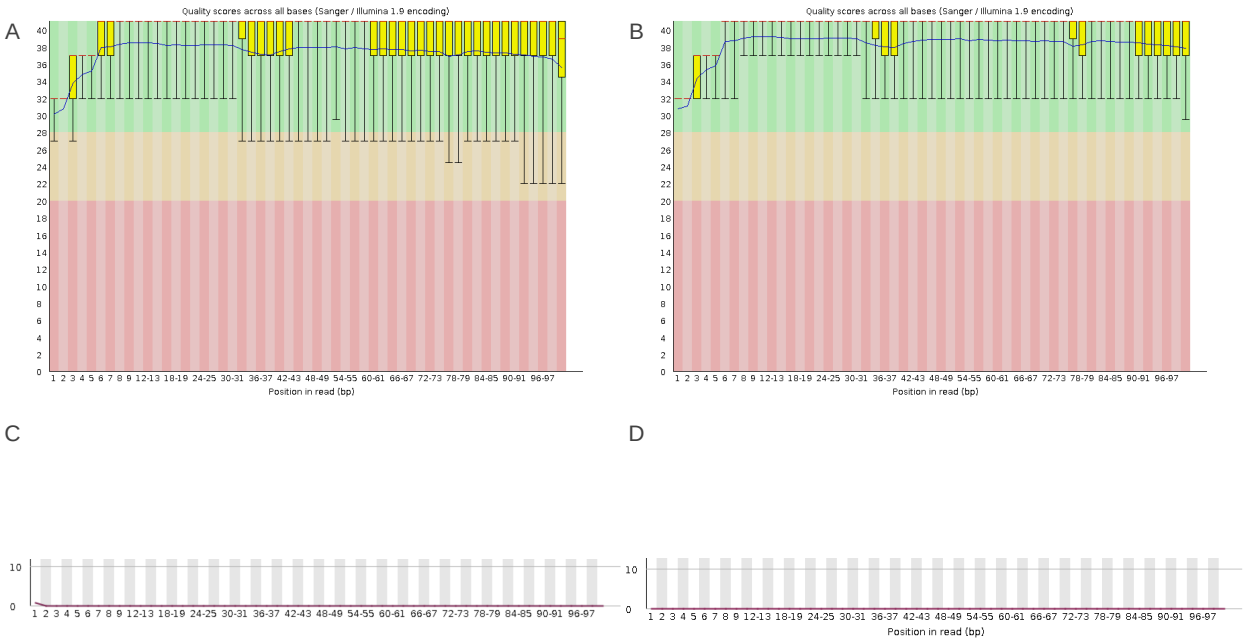

Figure 6: Per-base quality score and n-content distributions for trimmed and untrimmed R2 reads of 8_2F_fox_S7_L008 Quality of untrimmed R2 reads (**A**) and trimmed R2 reads (**B**) as well as N content of untrimmed R2 reads (**C**) and trimmed R2 reads (**D**) were calculated per base pair and plotted using `FastQC`

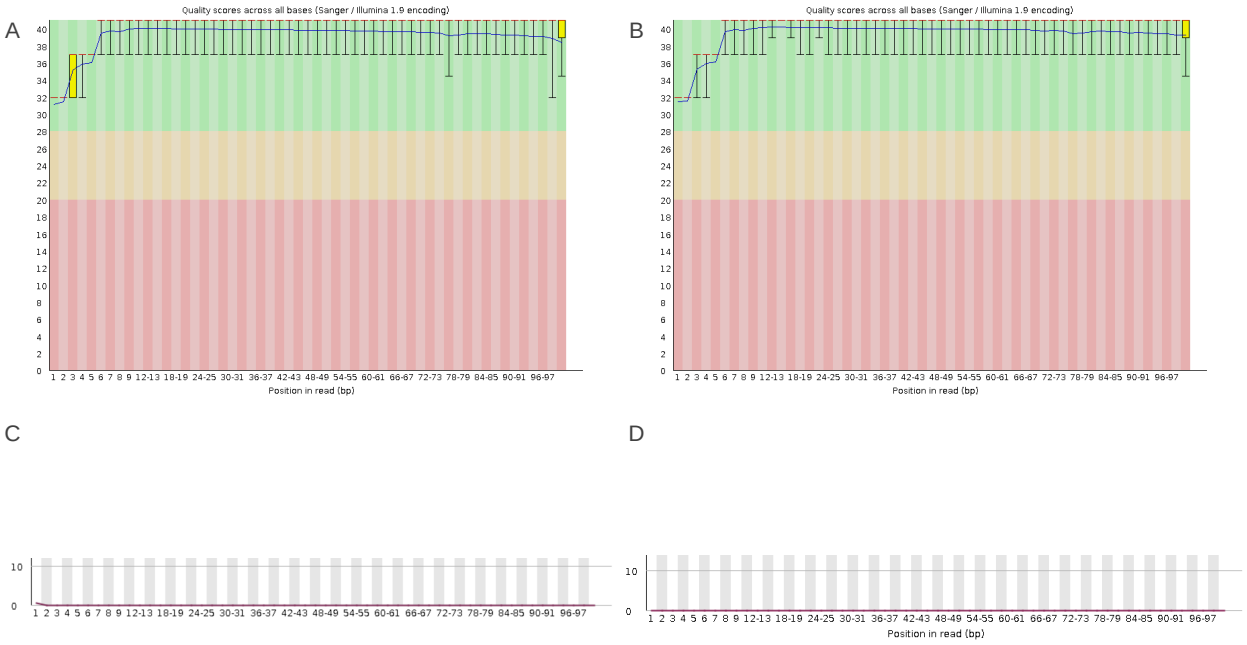

Figure 7: Per-base quality score and n-content distributions for trimmed and untrimmed R1 reads of 22_3H_both_S16_L008. Quality of untrimmed R1 reads (**A**) and trimmed R1 reads (**B**) as well as N content of untrimmed R1 reads (**C**) and trimmed R1 reads (**D**) were calculated per base pair and plotted using `FastQC`

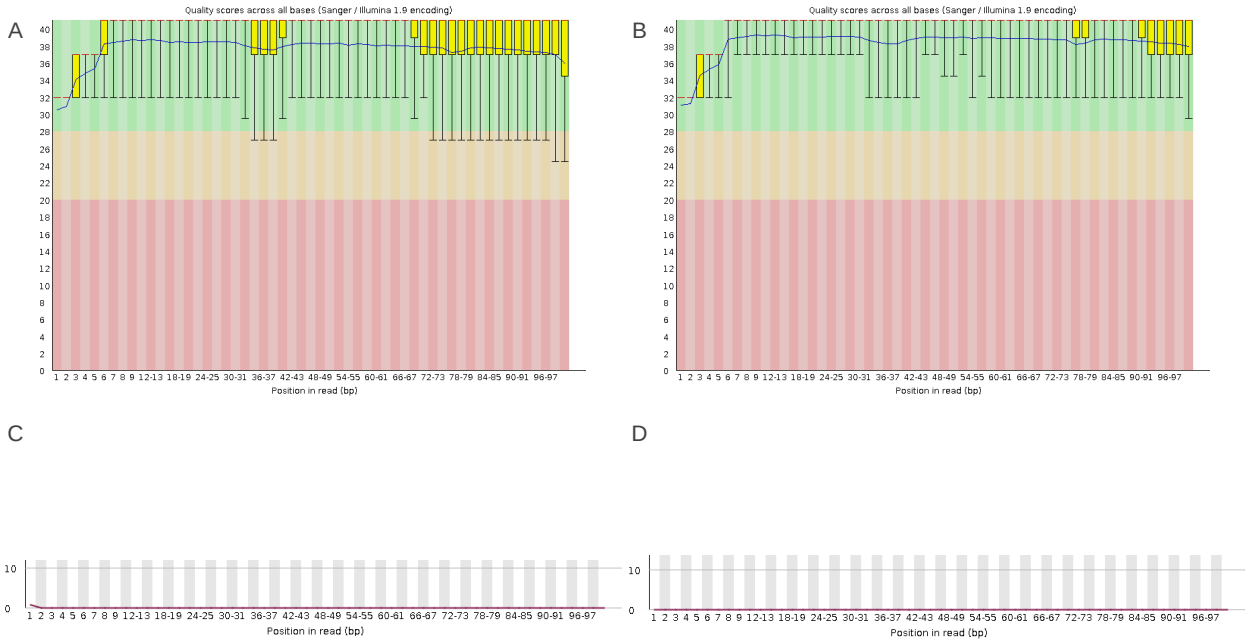

Figure 8: Per-base quality score and n-content distributions for trimmed and untrimmed R2 reads of 22_3H_both_S16_L008. Quality of untrimmed R2 reads (**A**) and trimmed R2 reads (**B**) as well as N content of untrimmed R2 reads (**C**) and trimmed R2 reads (**D**) were calculated per base pair and plotted using `FastQC`