QAA Report

This report presents a comprehensive quality evaluation of demultiplexed sequencing files 1_2A_control_S1_L008 and 24_4A_control_S18_L008 from the 2017 library. It juxtaposes quality assessment from established open-source bioinformatics tools with those generated by custom-developed programs. The analysis includes pre- and post-processing quality assessments, including trimming and filtering operations. Based on these quantitative evaluations, a data utilization recommendation will be formulated to guide subsequent analyses.

Sequence Quality Assessment

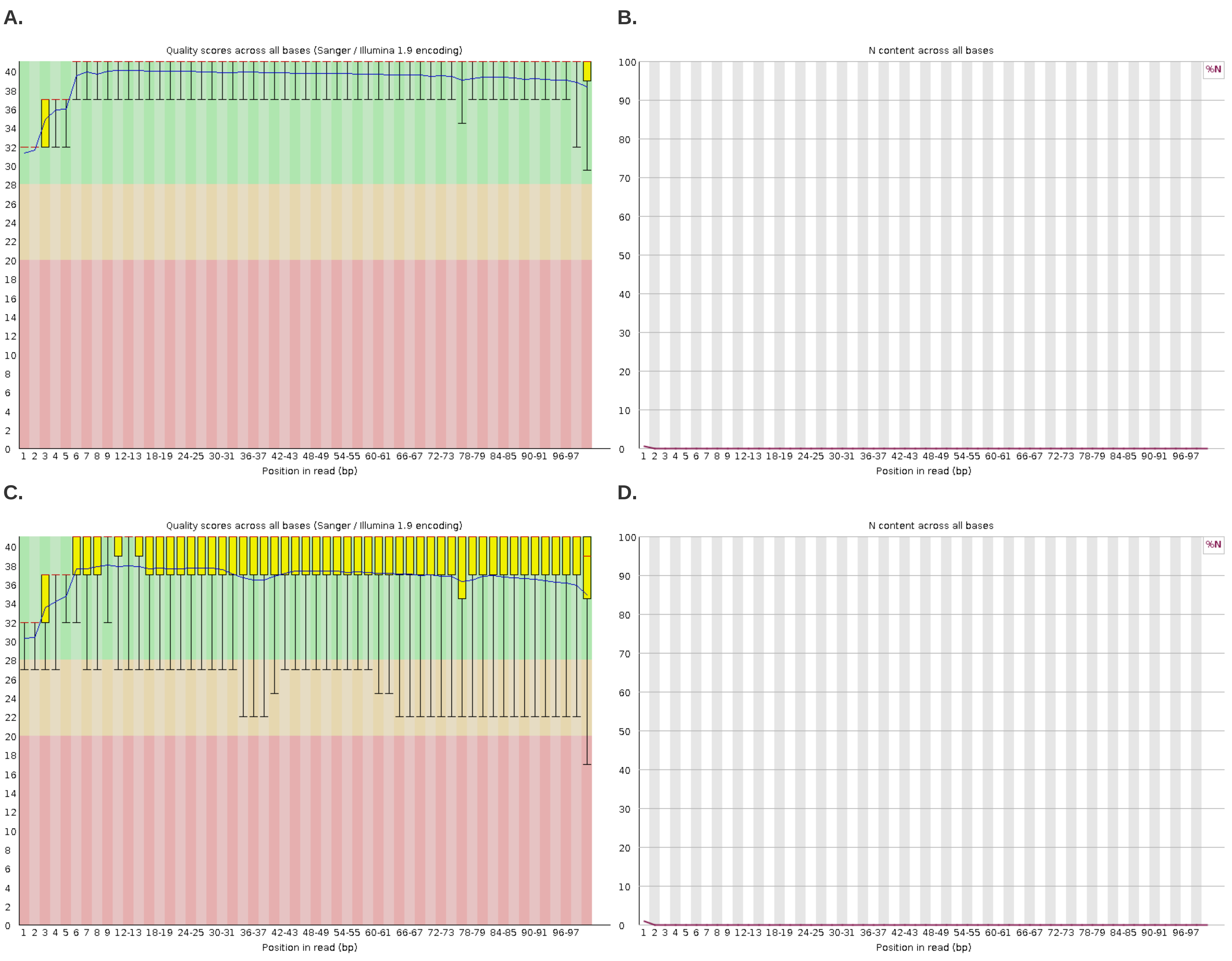

Figure 1: A,B are per base quality score and per base n content for 1_2A_control_S1_L008 read 1 (R1). C,D represent the quality score and n content for read 2 (R2).

Quality analysis shows consistently high Phred scores across all positions in both R1 and R2 reads, with a mean of appoximately 39 per base in R1 and 37 per base in R2. As expected, N content remains minimal throughout both reads, consistent with high Phred score. A slight inverse correlation is observed in the first two positions, where slightly lower quality scores correspond to higher N content in both R1 and R2.

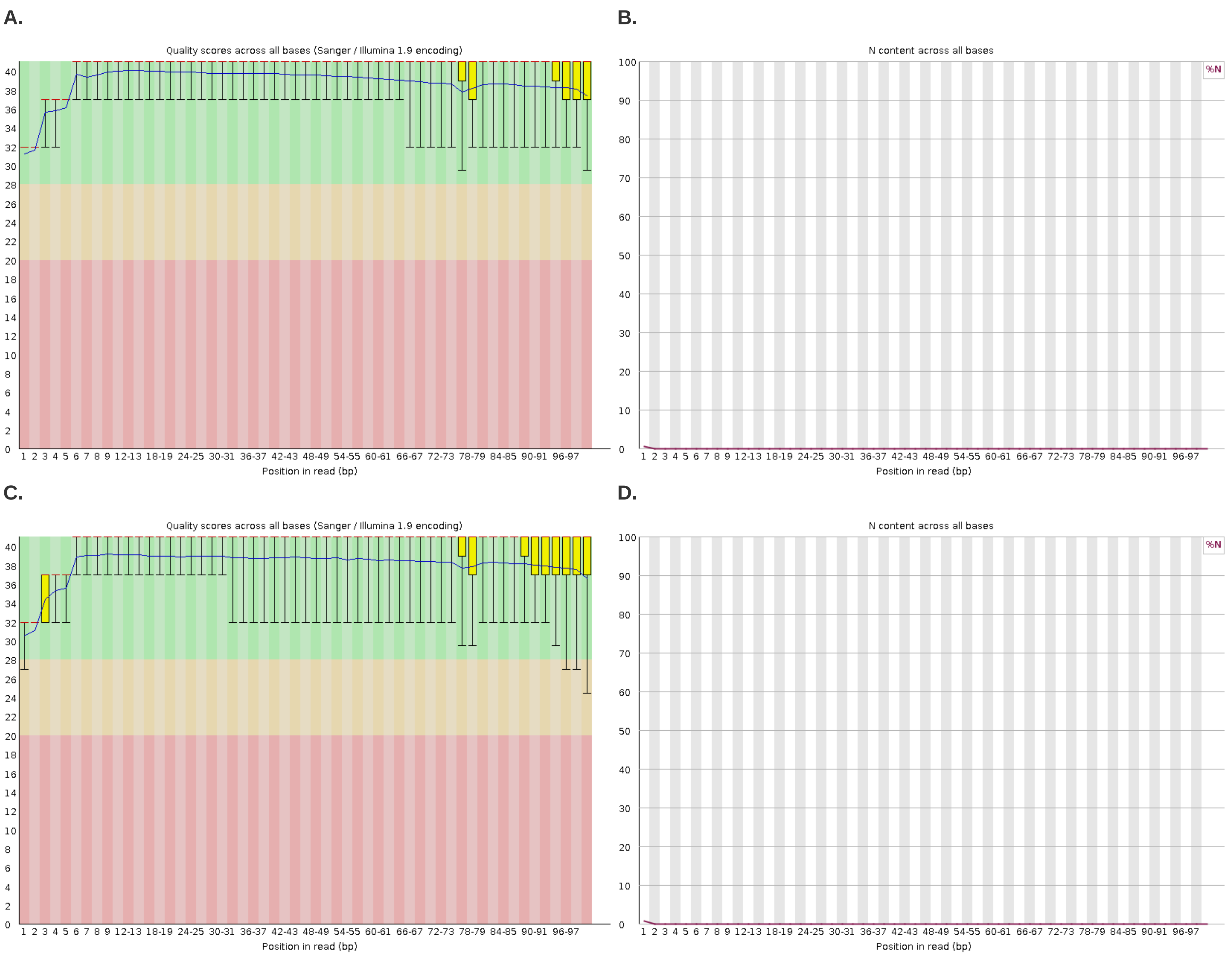

Figure 2: A,B are per base quality score and per base n content for 24_4A_control_S18_L008 read 1 (R1). C,D represent the quality score and n content for read 2 (R2).

The high-quality score consistency observed in Fig.1(A,C) is replicated in these reads, which aligns with expectations for control data, which is designed to to be of superior quality.

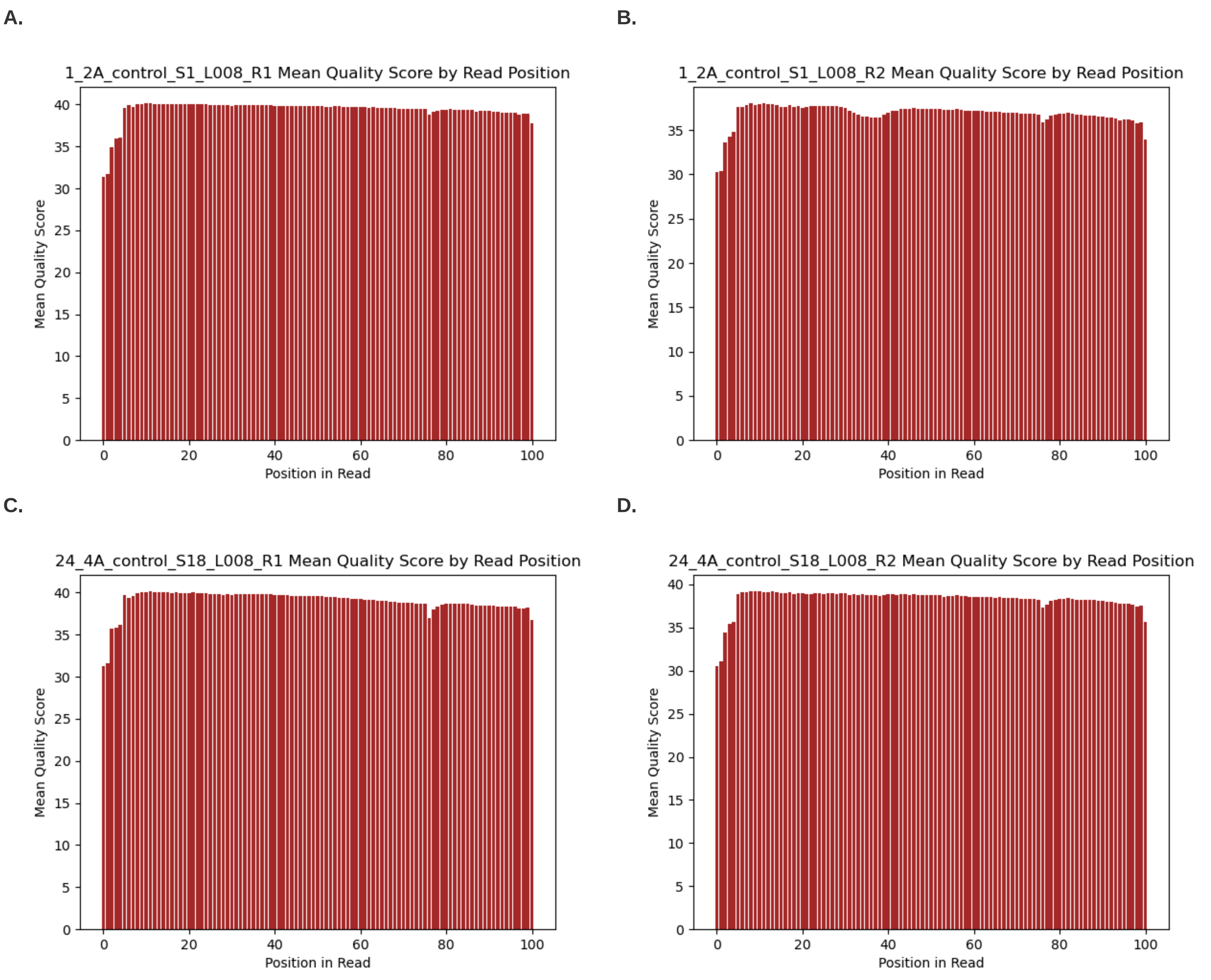

Figure 3: A,B. 1_2A_control_S1_L008 mean quality score distributions for R1 and R2 respectively generated using a program custom developed for this task. 24_4A_control_S18_L008 mean quality score distributions for R1 (C) and R2 (D) made with the same program.

The mean quality score distributions generated with each method produce the same numerical results; however, more in-depth plots are produced with fastqc, which displays standard deviation by position as well. The fastqc program runs more quickly than the custom program, it takes only 1 minute as opposed to 3. Both methods use similar computational resources (fastqc 97% CPU, custom 99% CPU), yet with those resources much more information is generated such as the per base n content distribution, and a summary of per base GC content.

Because both 1_2A_control_S1_L008 and 24_4A_control_S18_L008 are controls for the 2017 sequence dataset, overall quality fo these sequences is high. Based on mean quality score, per base N content, and overall sequence quality, it is recommended both 1_2A_control_S1_L008 and 24_4A_control_S18_L008 be used in further analyses.

Trimming Comparison

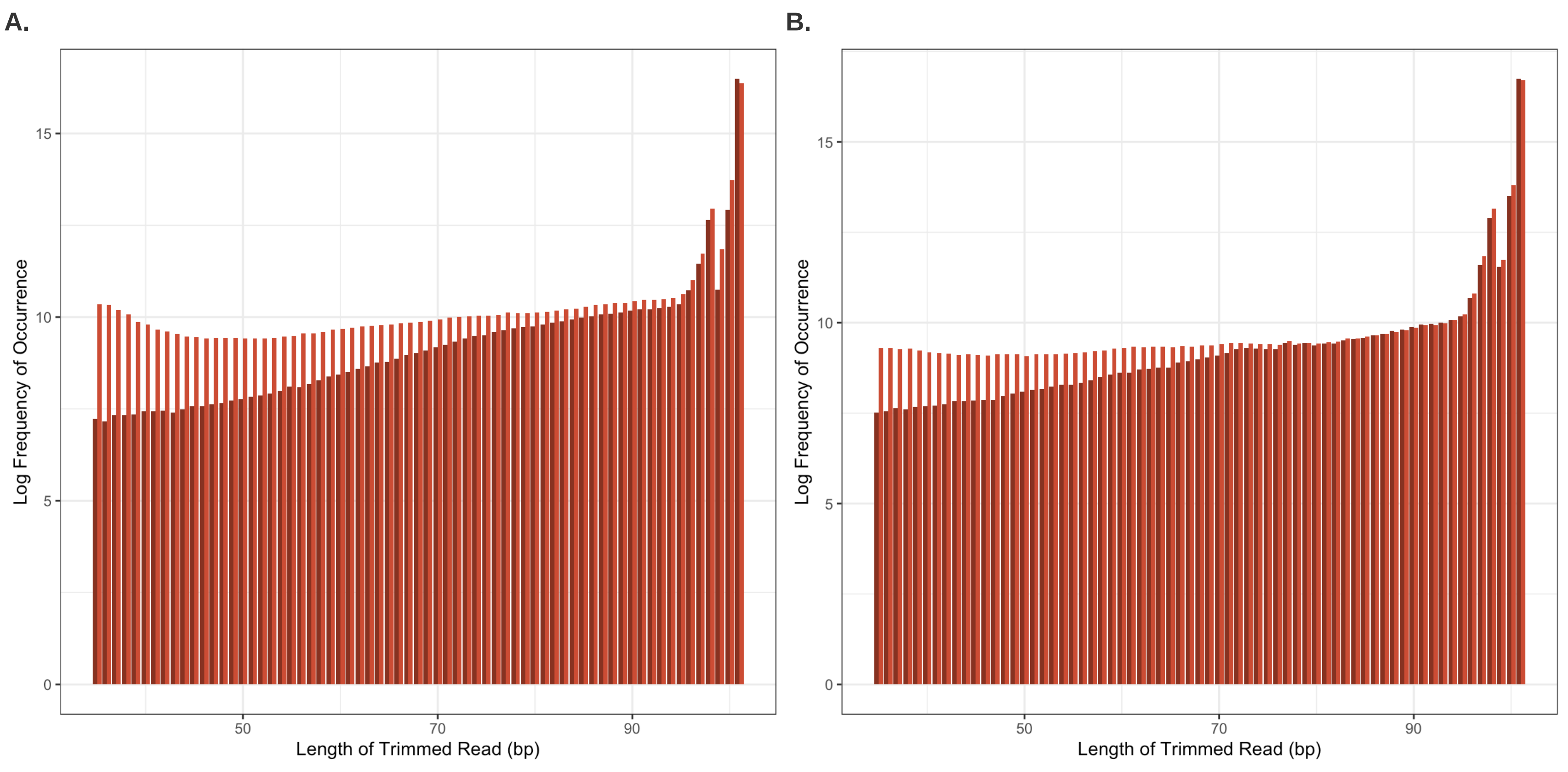

Figure 4: A. 1_2A_control_S1_L008 distribution of trimmed read lengths and the frequency of their occurrences, logged. Dark red is R1 and Light red is R2. B. 24_4A_control_S18_L008 distribution of trimmed read lengths and the frequency of their occurrences, logged. Dark red is R1 and Light red is R2.

Due to the functionality of Illumina sequencing, it is expected that R1 will not get trimmed as frequently as R2. Portions of adaptor sequence are more likely to be found in R1 and be trimmed out and thus R1 will have shorter reads more frequently. This is seen in Fig.4.

Alignment and Strand-Specificity

According to the current htseq parameters, strand-specific datasets will produce different numbers of mapped reads per count depending on the setting stranded=yes or stranded=reverse. Based off the htseq counts, it appears for both 24_4A_control_S18_L008 and 1_2A_control_S1_L008, different numbers of mapped and unmapped reads were produced depending on the “strandedness” setting (7158427 map to no feature when stranded=yes for 1_2A_control_S1_L008, and 447091 when reverse. 873868 mapping to no feature when reverse and 9072871 when yes for 24_4A_control_S18_L008). This indicates that the datasets are strand-specific.

	Aligned	Secondary+ Alignment	Unaligned	Total	Percent Aligned
1_2A_control_S1_L008	15627437	1080021	0	16707458	93.54%
24_4A_control_S18_L008	19780624	1804084	0	21584708	91.64%

Table 1: Percentage of reads aligned to a feature for 1_2A_control_S1_L008 and 24_4A_control_S18_L008. Reads that map to features more than once are not counted multiple times in percentage.