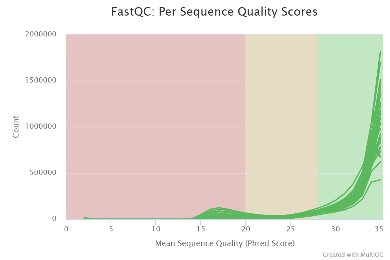
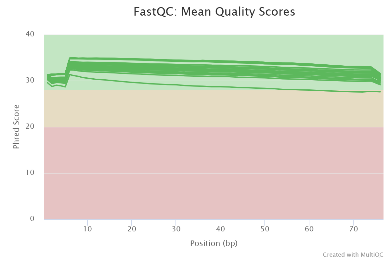
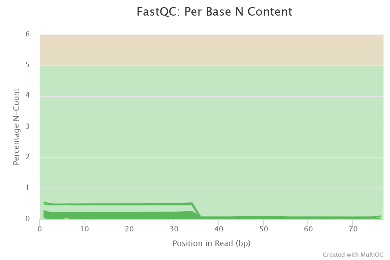
**Paired**

**PRJNA694147 (GSE165352)**

SRR13509767 read quality falls below 30 PhredScore for a considerable amount of positions towards the end suggesting possible issues with this sample due to phasing errors. All other samples showed good average and per base sequence quality. No adapters were detected suggesting these were already removed. There is also a detectable but still low amount of unknown “N” nucleotides in some sequences which should be filtered out following fastp filtering.



Following fastp filtering the minor issues detected in fastqc check were resolved with “N” bases filtered out as well as the low quality bases and reads in sample SRR13509767 which were causing average phred score to fall below 30 towards the end of the reads.

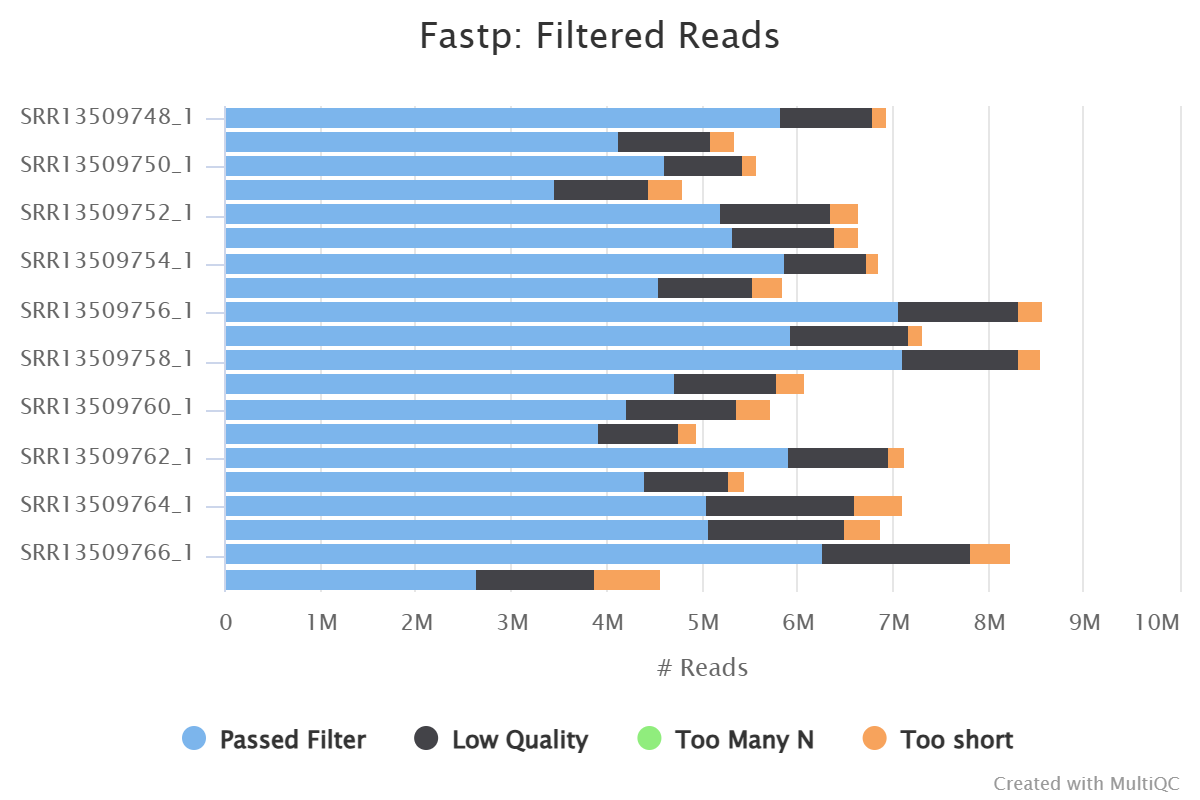
A chart with green and white bars

AI-generated content may be incorrect.A graph with green line

AI-generated content may be incorrect.A graph of a number of scores

AI-generated content may be incorrect.

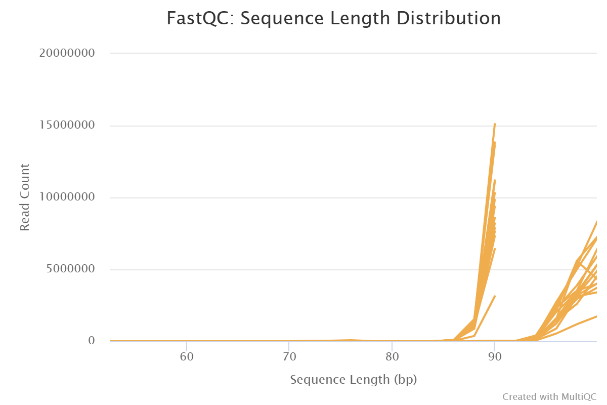
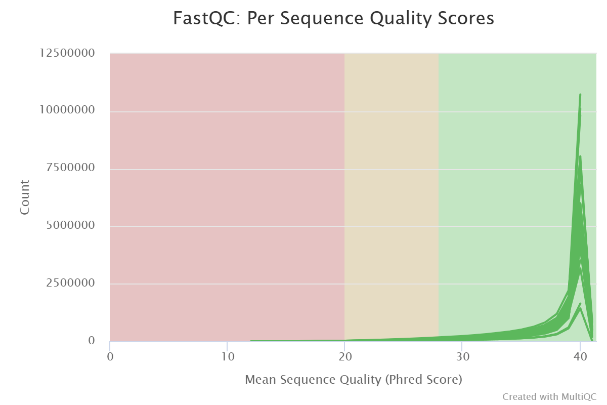
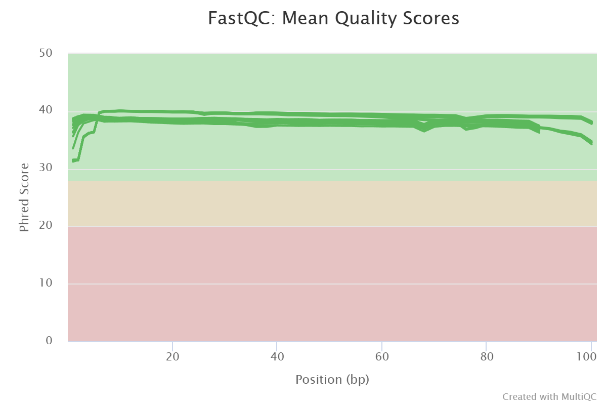
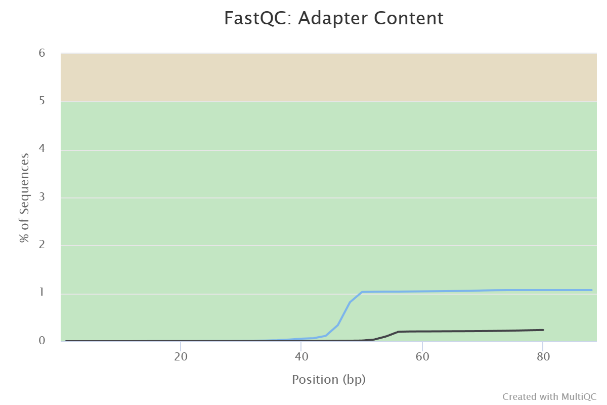
All samples were mapped as no indication of significantly low quality was observed. All samples falling below 90% mapping were excluded from Baerhunter analysis. Interestingly sample with lower quality before trimming also showed the lowest mapping percentage of 65.73% and was excluded from further analysis. This could be due to issues not just with sequencing but also library preparation or contamination.



|  |  |
| --- | --- |
| SRR13509748 | 96.67 |
| SRR13509749 | 94.88 |
| SRR13509750 | 97.54 |
| SRR13509751 | 94.73 |
| SRR13509752 | 94.21 |
| SRR13509753 | 92.74 |
| SRR13509754 | 98.39 |
| **SRR13509755** | **86.46** |
| SRR13509756 | 94.63 |
| SRR13509757 | 94.92 |
| SRR13509758 | 91.83 |
| **SRR13509759** | **80.16** |
| SRR13509760 | 93.95 |
| SRR13509761 | 94.37 |
| SRR13509762 | 94.63 |
| **SRR13509763** | **81.41** |
| SRR13509764 | 97.84 |
| **SRR13509765** | **84.36** |
| SRR13509766 | 93.22 |
| **SRR13509767** | **65.73** |

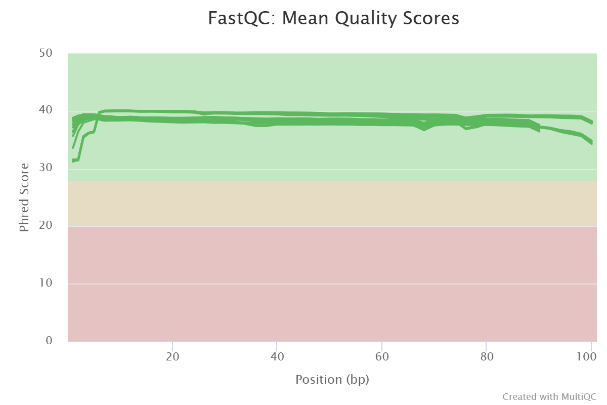
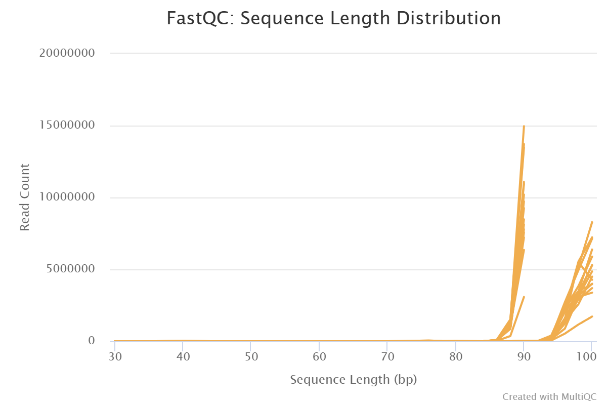
SRR13509759 shows a decrease in read quality towards the end however mean quality is consistently high (in green zone) suggesting few reads are lowering the overall quality and therefore more stringent filtering approaches will be applied during processing with fastp to remove these. Mapping quality has increased from 79.19 to 80.16 following this modification to fastp code from standard approach. This extra step improved the per base quality of the sample but the increase on quality of mapping was insignificant with sample mapping percentage still falling below 90 and therefore excluded from further analysis. Smaples SRR13509767 has the lowest mapping it has also undergone the most stringent filtering and trimming to attempt to combat that. Even with a large proportion of the reads filtered out the mapping quality is not good hence it was excluded from further analysis with no more attempts at recovering mapping percerntage.

**PRJNA886436 (GSE214640)**

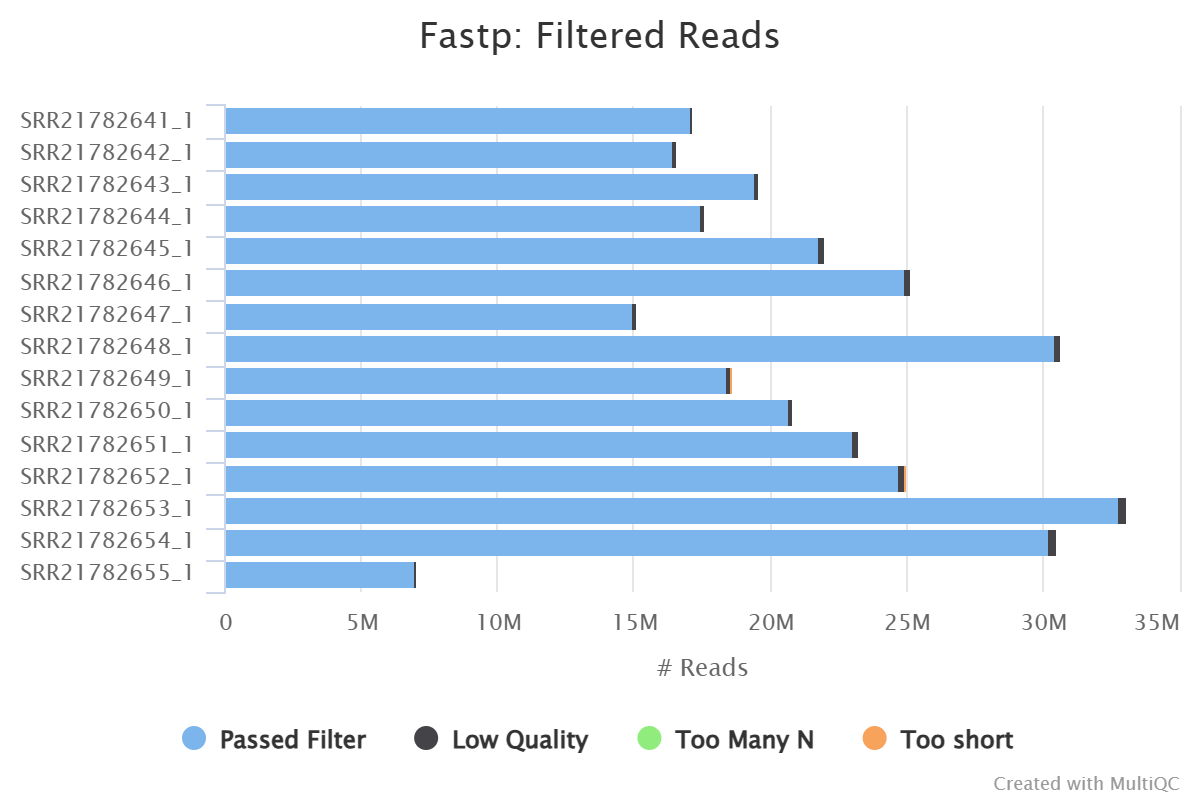
****

All samples were within the acceptable region of per sequence and per base quality with minimal adapter presence. However, two distinct sequence lengths were identified suggesting possible premature truncation during sequencing. In fact all the sequences corresponding to pair 2 are shorter than those in pair 1. This should not cause issues but will be considered during further processing.A graph of a fastqc

AI-generated content may be incorrect.A graph of a number of scores

AI-generated content may be incorrect.

There was no significant improvement in quality of the sequencing following trimming based on multiqc however some low-quality sequences were removed in each sample regardless.

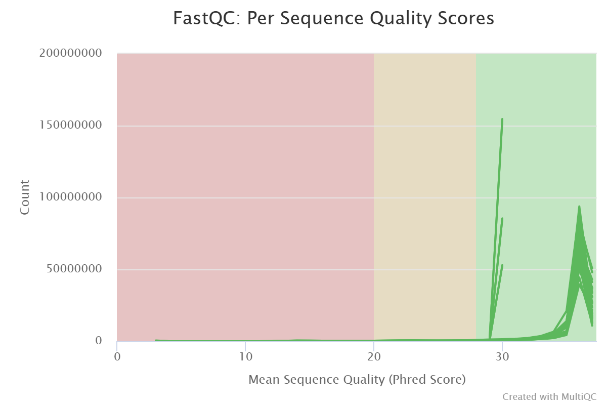
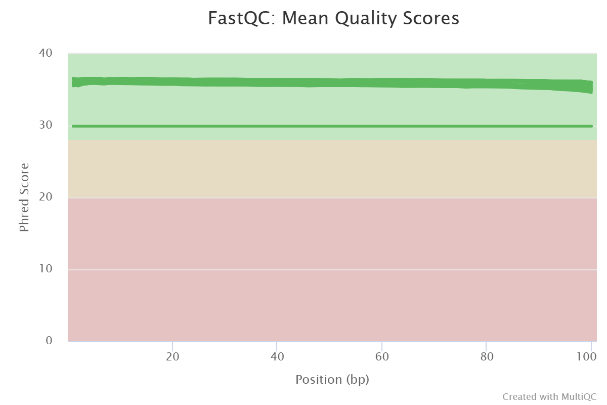
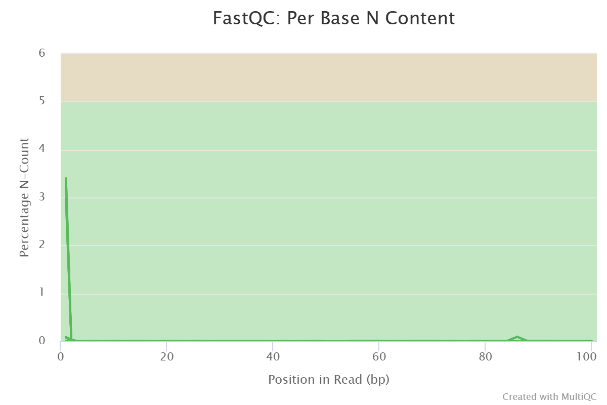
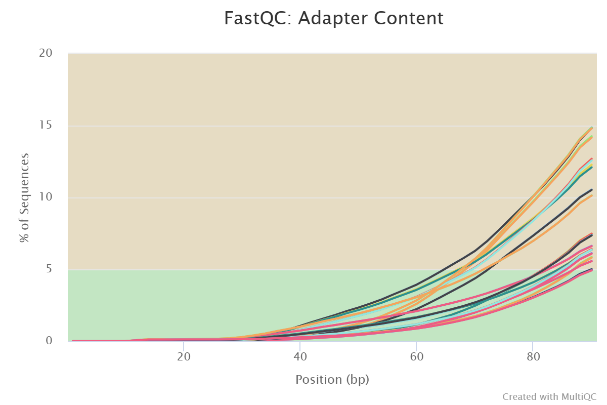


|  |  |
| --- | --- |
| SRR21782641 | 98.91 |
| SRR21782642 | 98.91 |
| SRR21782643 | 98.94 |
| SRR21782644 | 98.59 |
| SRR21782645 | 99.15 |
| SRR21782646 | 99.24 |
| SRR21782647 | 98.23 |
| SRR21782648 | 97.49 |
| SRR21782649 | 98.57 |
| SRR21782650 | 99.11 |
| SRR21782651 | 98.69 |
| SRR21782652 | 97.69 |
| **SRR21782653** | **78.26** |
| **SRR21782654** | **84.95** |
| SRR21782655 | 94.73 |

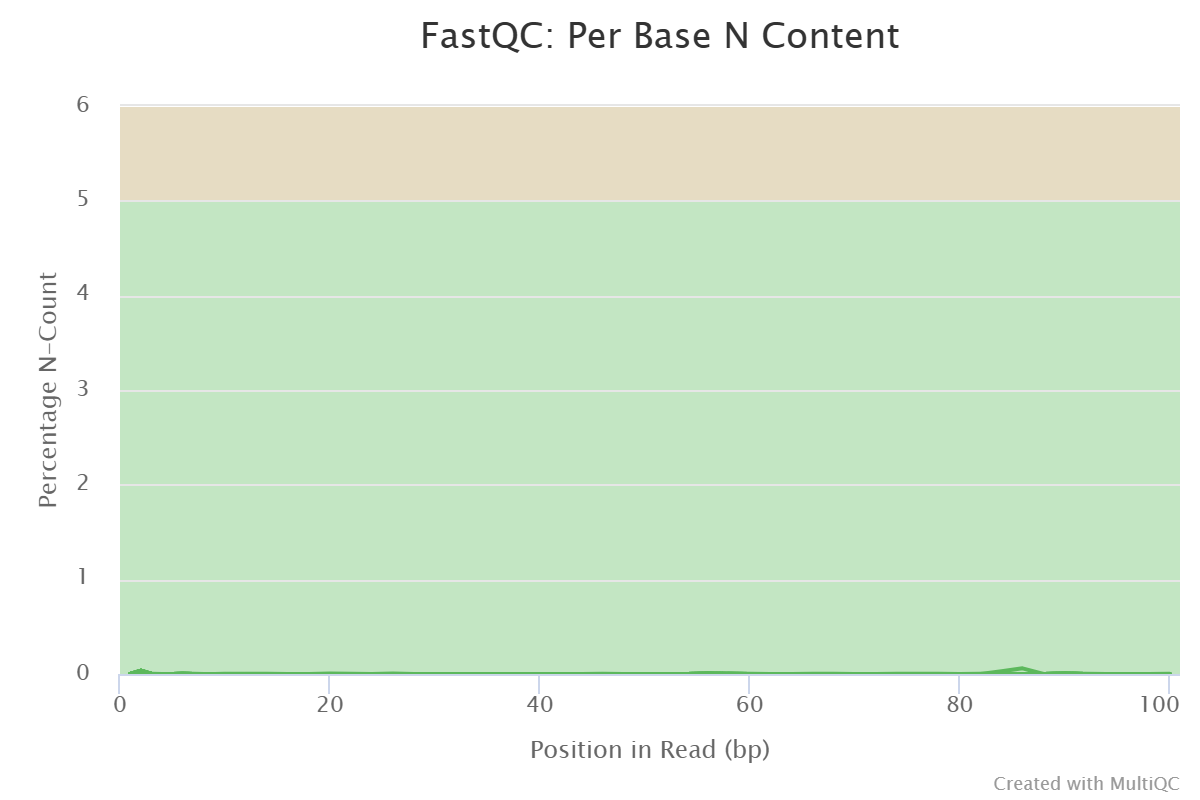
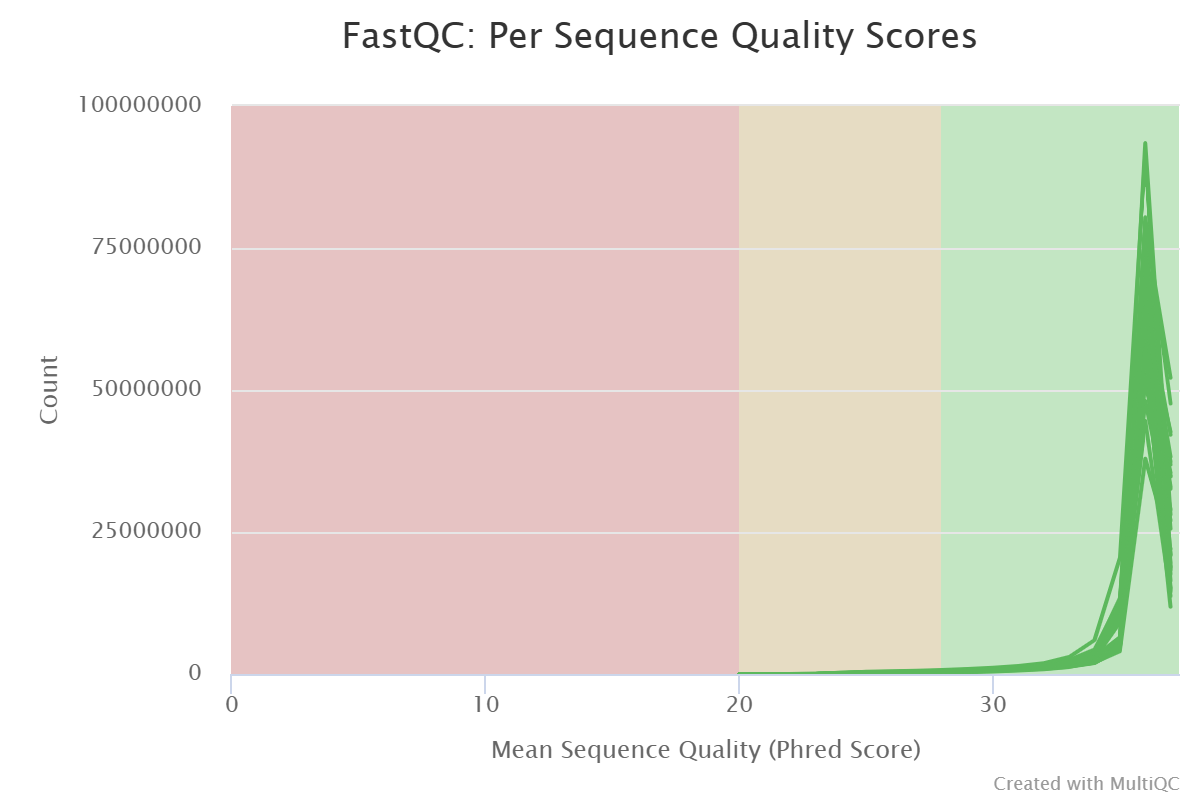
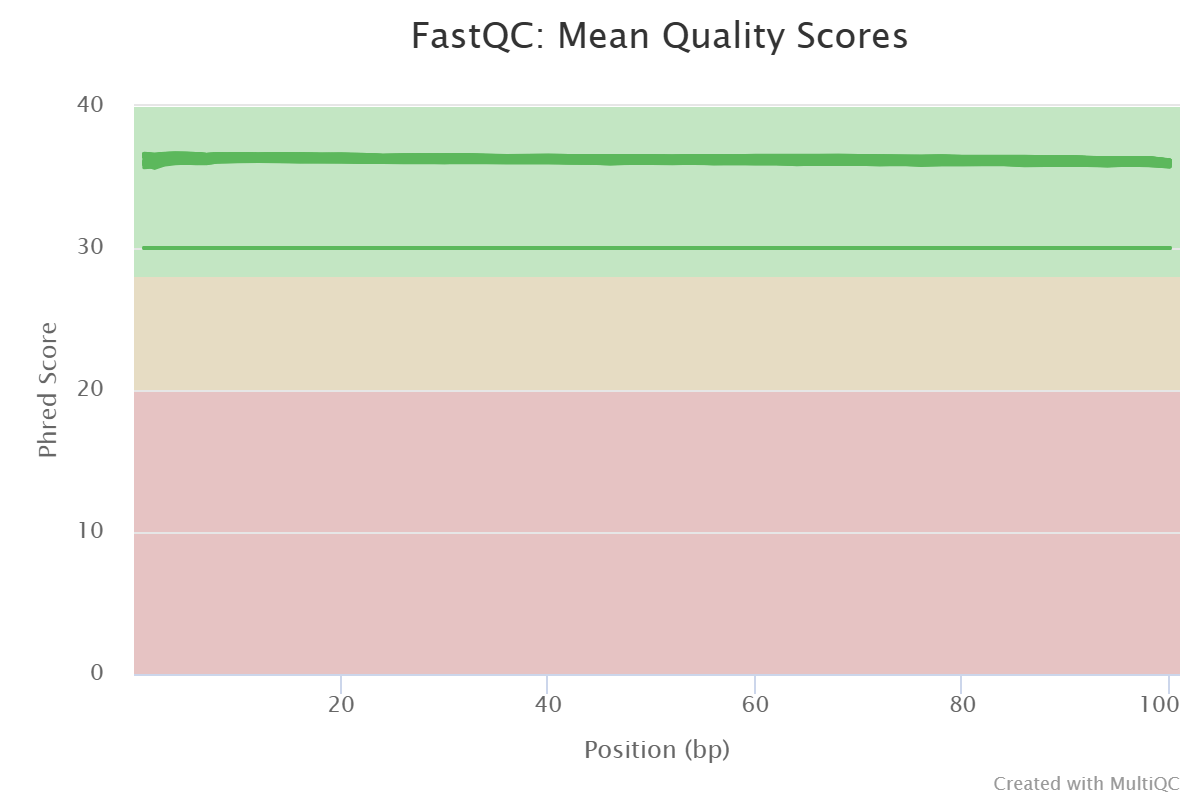
Mapping was good for most samples and the trace number of adapters left in sample SRR21782652 does not look to have caused any issues therefore it is likely that adapters were not true finds. Samples SRR21782653 and SRR21782654 had no observable issues in sequencing quality but were removed from further analysis due to mapping percentage falling below 90%.

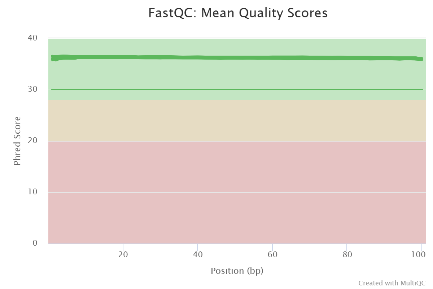
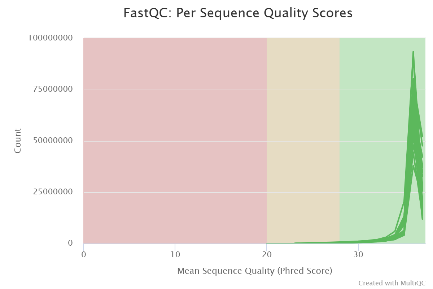
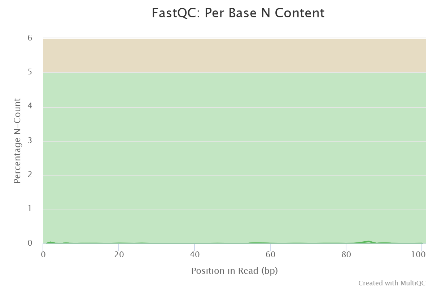
**PRJNA917642 (GSE222081)**

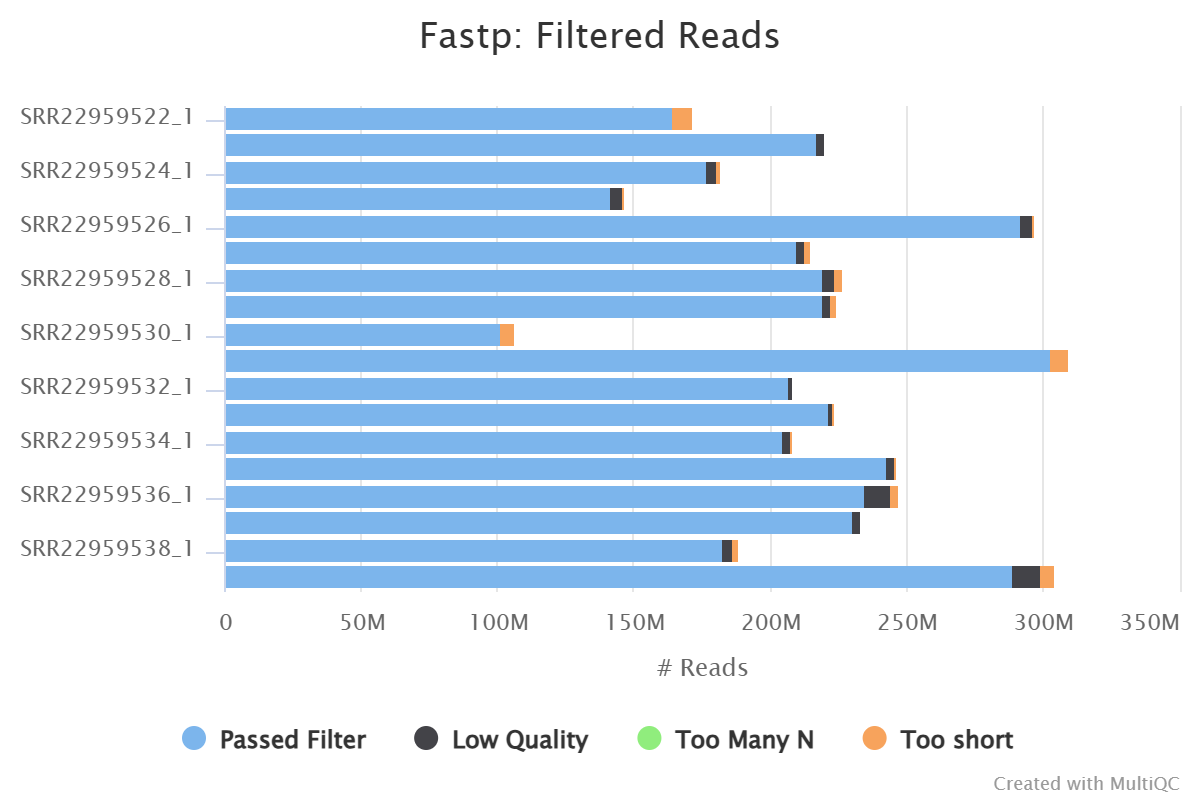
All samples are within acceptable for per base and average sequence quality however sample SRR22959522 has overall decreased quality across all positions though still within acceptable range. Similarly sample SRR22959530 and SRR22959531 has an increased count of sequences below Phred score of 30 paired with a higher percentage of “N”/unknown nucleotides towards the start of the reads. This could be due to an issue with sequencing for those samples. However, as they still pass all the checks they will be further processed in line with the other samples. There was a good amount of adapters still present but they should be taken care of during filtering and trimming with fastp.



Following trimming mean sequence quality improved for some samples to match the rest N content was noticeably decreased and there were “No samples found with any adapter contamination > 0.1%”. Additional parameters were introduced to remove adapters (--detect\_adapter\_for\_pe) and trim the N bases at the start of the reads (-5).





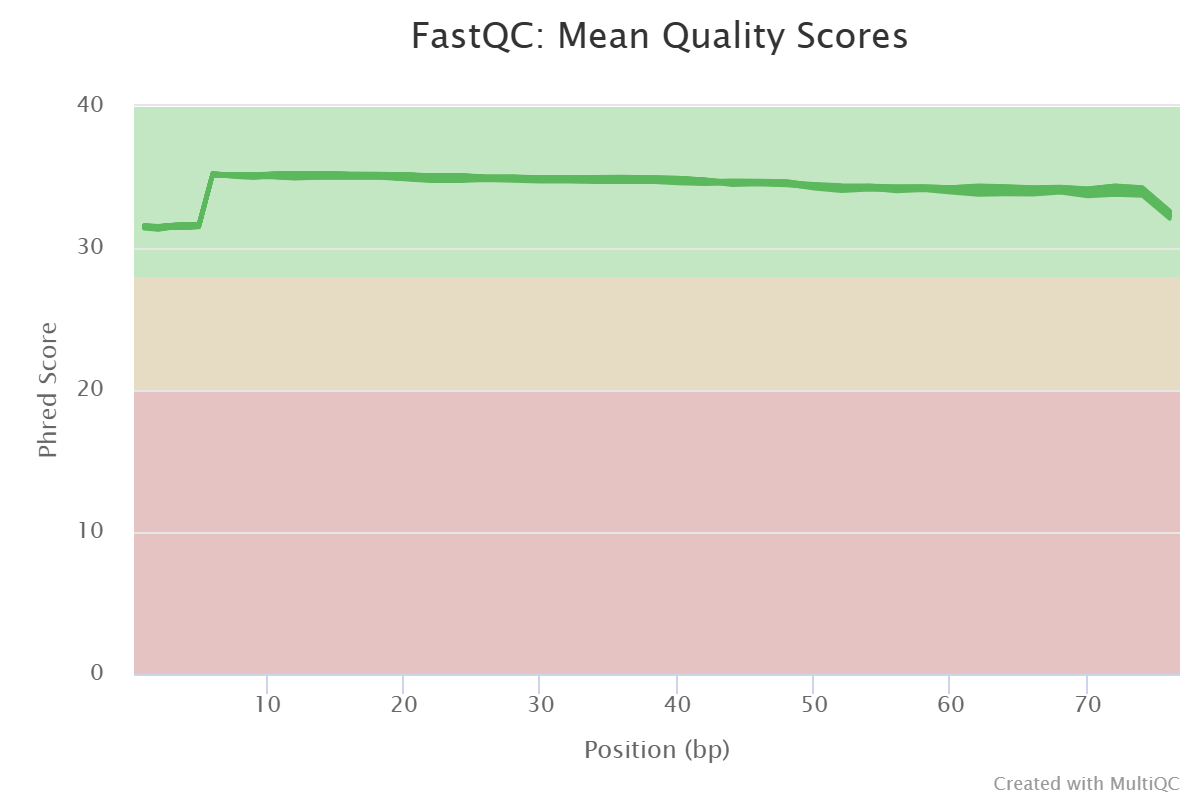
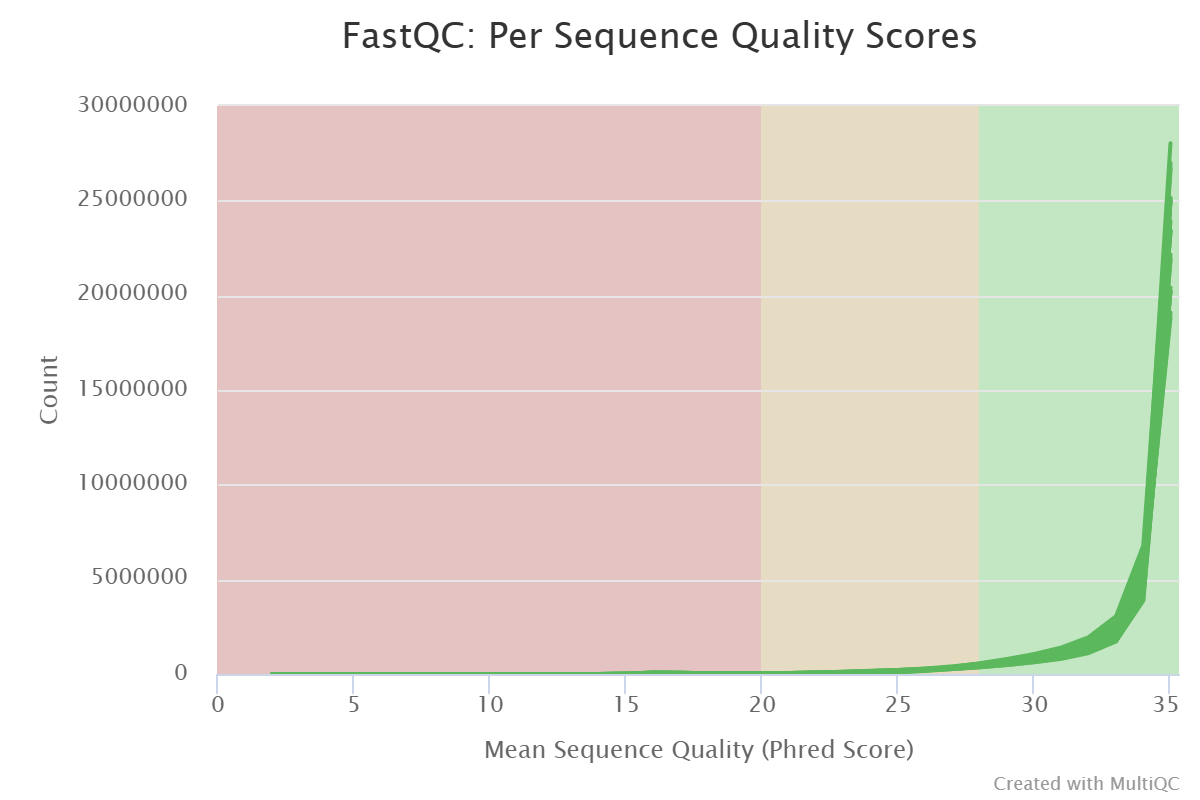
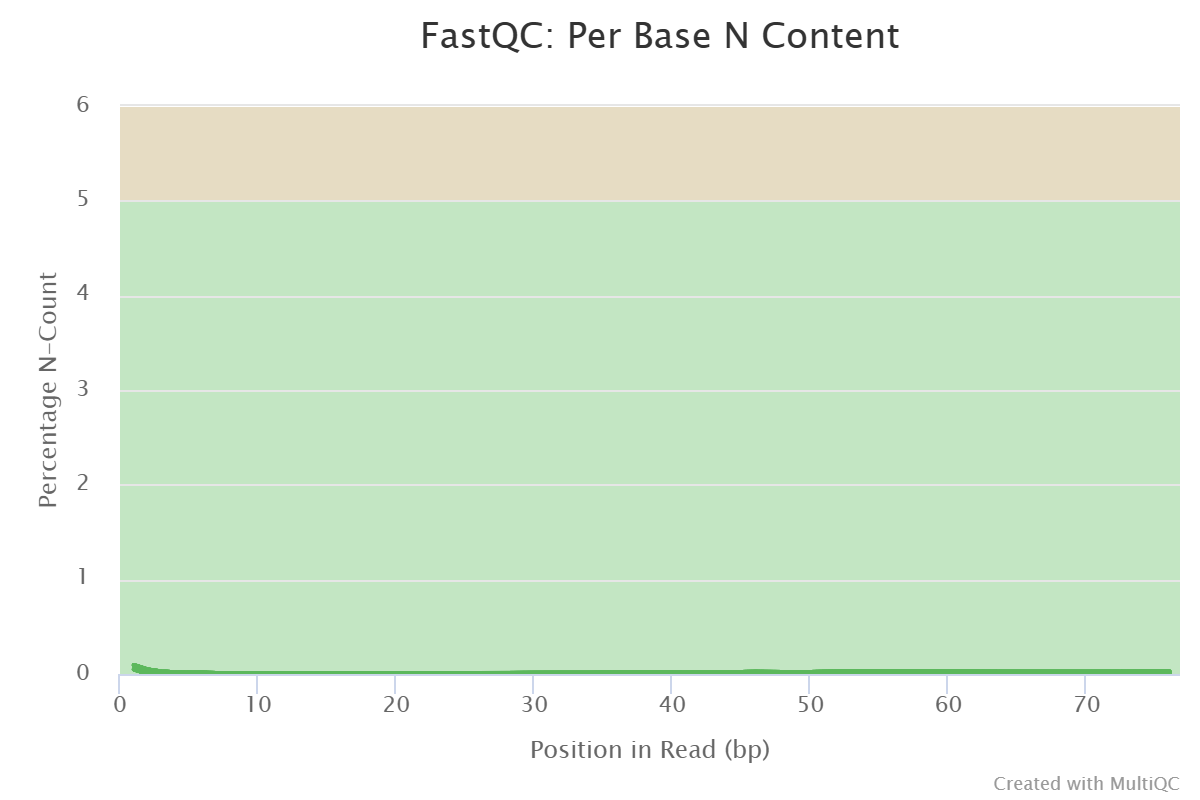
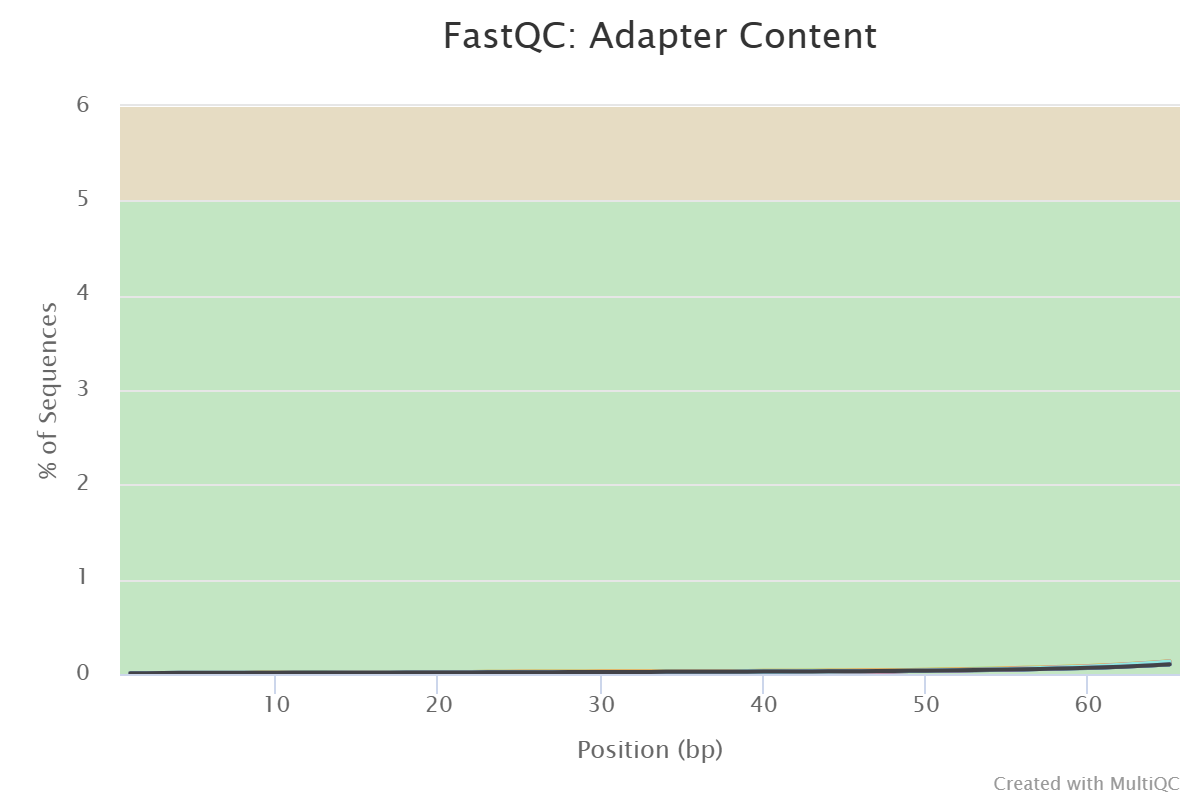


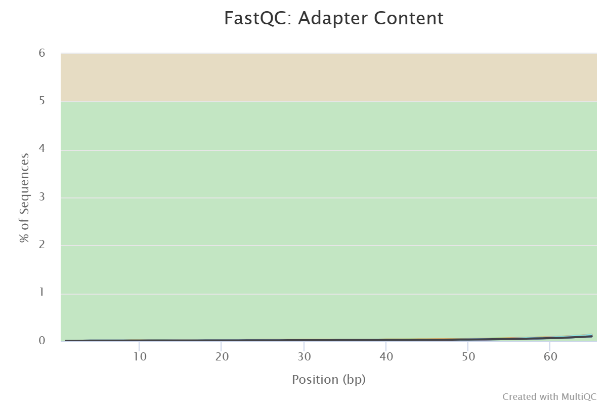
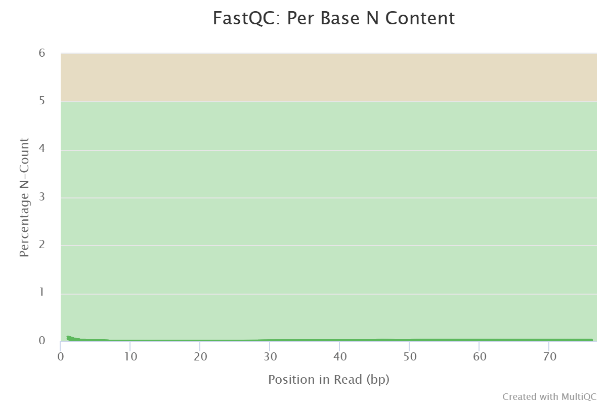
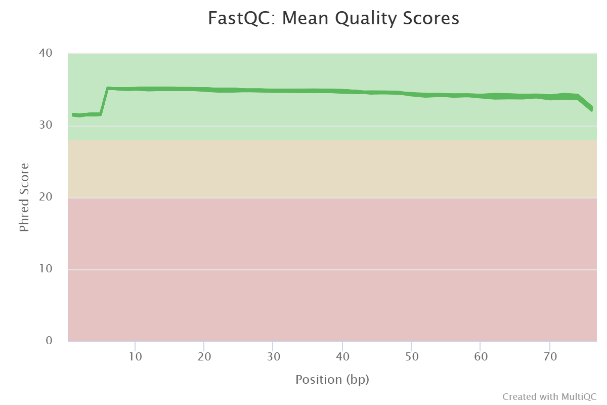
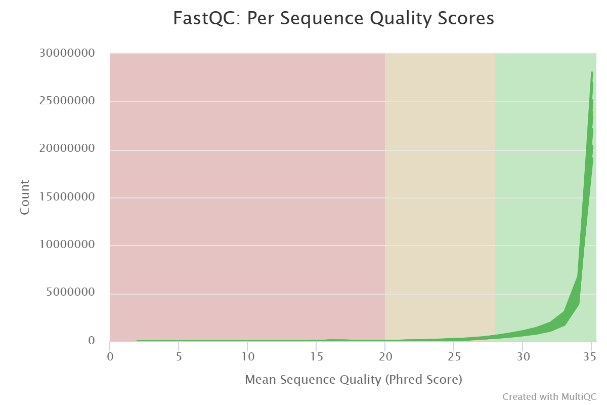
Following trimming the few sequences which had lower quality than the rest have been removed but the sample with lower quality remained slightly lower than the rest. There was no detectable adapters and the N content at the beginning of some reads has been removed successfully.

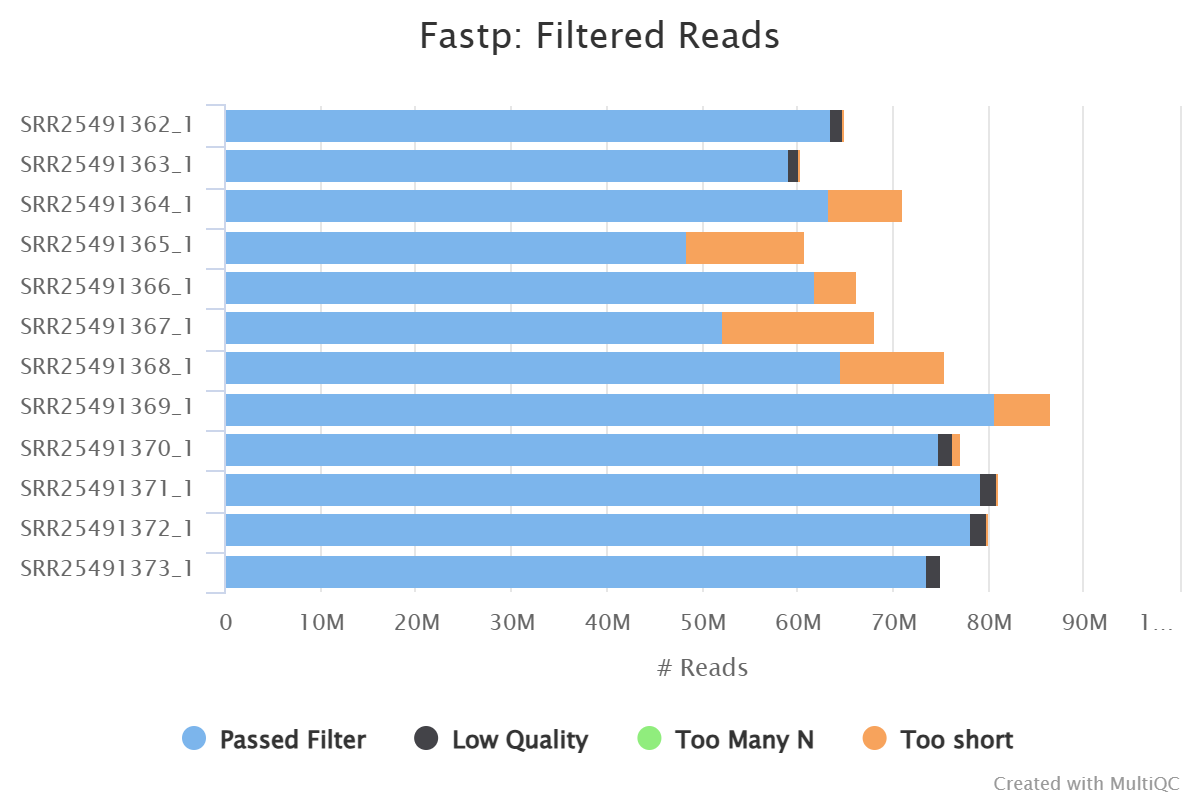
|  |  |
| --- | --- |
| SRR22959522 | 96.13 |
| SRR22959523 | 99.21 |
| SRR22959524 | 98.14 |
| SRR22959525 | 97.84 |
| SRR22959526 | 99.32 |
| SRR22959527 | 99.21 |
| SRR22959528 | 99.17 |
| SRR22959529 | 99.52 |
| SRR22959530 | 96.27 |
| SRR22959531 | 98.19 |
| SRR22959532 | 99.57 |
| SRR22959533 | 99.31 |
| SRR22959534 | 99.01 |
| SRR22959535 | 99.27 |
| SRR22959536 | 99.03 |
| SRR22959537 | 98.87 |
| SRR22959538 | 97.71 |
| SRR22959539 | 97.96 |

There were no samples with mapping percentage lower than 90% and therefore none were excluded from further analysis.

**PRJNA1001307(GSE239869)**

From the start there is nothing concerning to be found in this dataset in terms of quality, adapters have already been removed and N content is low per sequence and mean quality scores are both well within the good range. There is a slight decrease in quality towards the start and end on the reads so “-5” and “–3” parameters will be added to the trimming script to combat these.





There was not much of a difference that could be observed following filtering. However, some low quality or too short samples were removed. Majority of these were removed due to being too short and some due to poor quality.

|  |  |
| --- | --- |
| SRR25491362 | 97.19 |
| SRR25491363 | 98.24 |
| SRR25491364 | 97.55 |
| SRR25491365 | 96.7 |
| SRR25491366 | 98.08 |
| SRR25491367 | 96.49 |
| SRR25491368 | 96.62 |
| SRR25491369 | 95.91 |
| SRR25491370 | 96.27 |
| SRR25491371 | 96.79 |
| SRR25491372 | 96.71 |
| SRR25491373 | 97.34 |

All samples mapped at a percentage greater than 90% so none were excluded from further analysis.

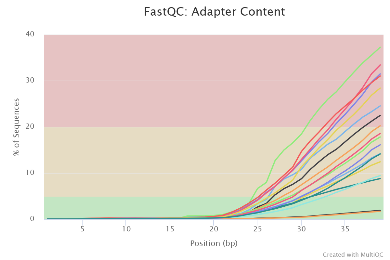
**Single**

**PRJNA313774(GSE78787)**

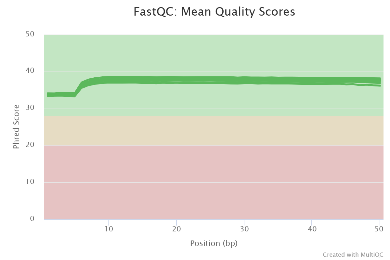
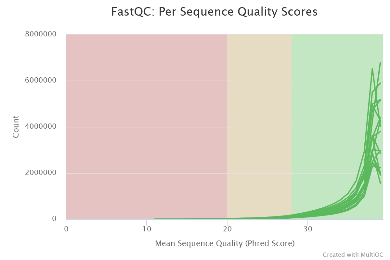
Following initial quality check utilising FastQC all samples show good quality both in terms of mean quality score and per base quality score however show a substantial amount of adapter content which will have to be removed during the trimming process.

A graph of a number of scores

AI-generated content may be incorrect.A green line on a pink and green line

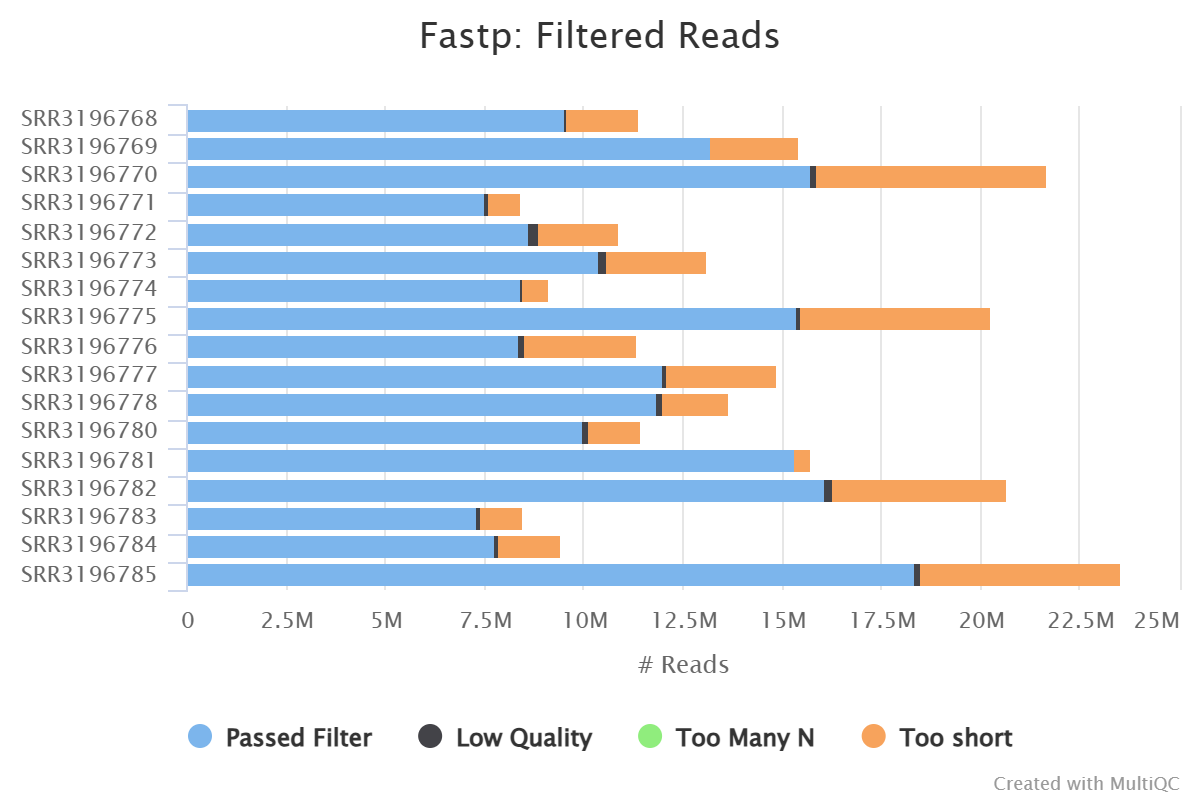
AI-generated content may be incorrect.

After trimming with default settings sample SRR3196779 and SRR3196781 still showed some adapter content although within the acceptable range.

A graph of a graph

AI-generated content may be incorrect.

Following Fastp processing a large amount of filtered out reads were too short with some low quality reads filtered out as well.

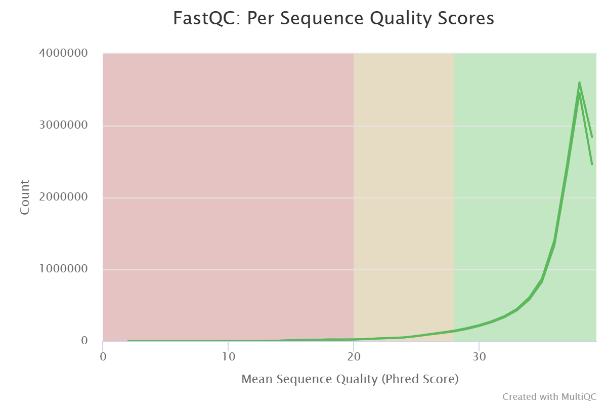
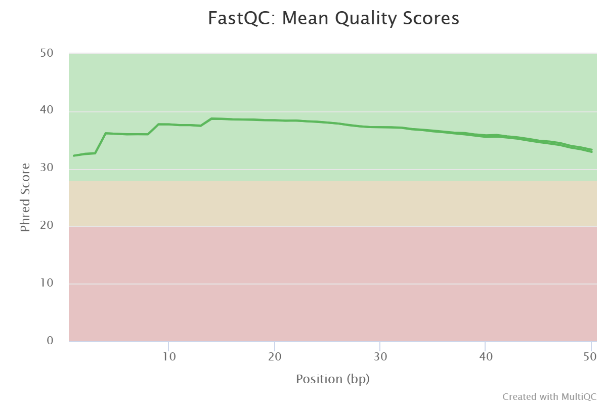


All samples in PRJNA313774 were mapped and below is a summary of mapping percentages. Those in bold have been excluded from further analysis due to insufficient mapping quality. As the mapping quality of single end data normally has a lower mapping percentage compared to paired and as this has been observed in this dataset lower acceptance threshold of at least 80 % mapping has been observed. Considering the quality observed when analysing these samples with FastQC low mapping percentage must be attributed to factors outside of sequencing quality such as contamination or experimental errors in library preparation. Particularly the first 3 samples show very low mapping and higher duplication some more significant error. These are likely due to large amounts of contamination as a homology search of unmapped reads seem to align well to other bacterial species thus suggesting sample preparation error. After excluding all samples with low mapping percentage only 8 out of 18 samples were used to predict ncRNAs using Baerhunter. Surprisingly the best mapped were SRR3196779 and SRR3196781 which did not stand out from the rest of the rest of the samples other than having a larger detected number of adapters following fastp trimming and filtering.

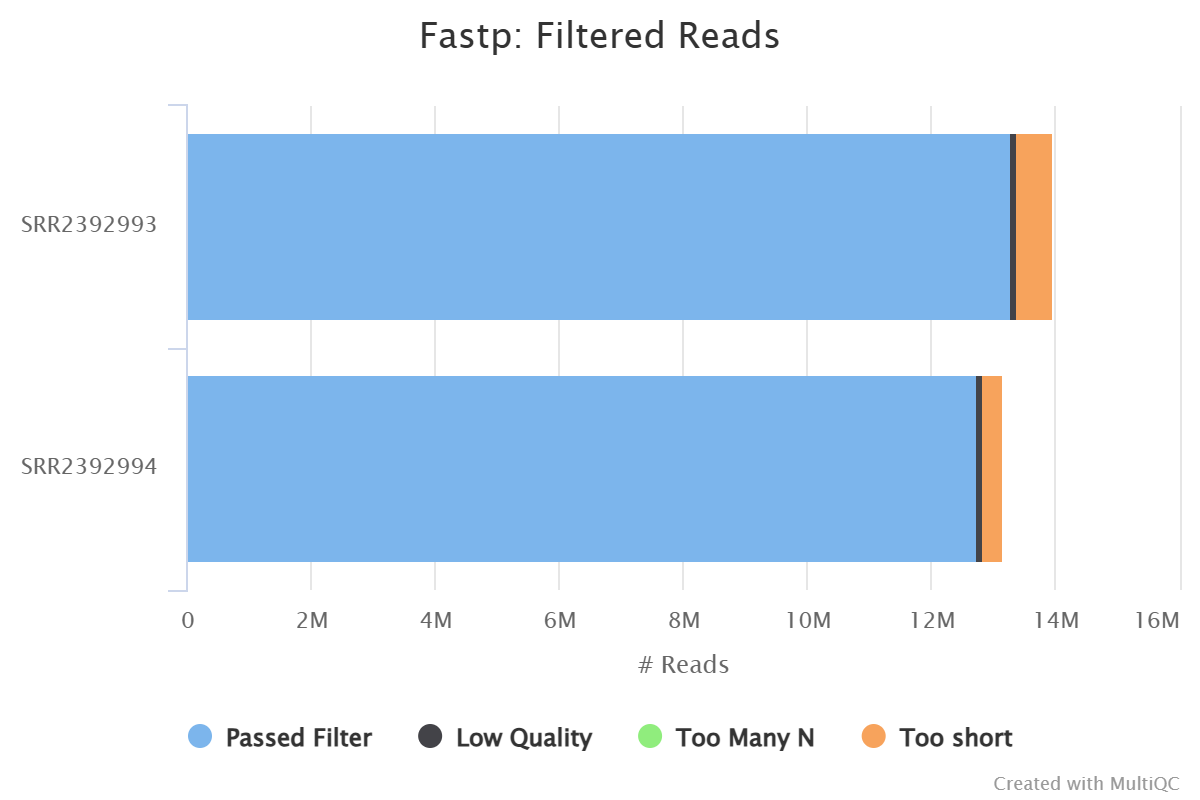
|  |  |
| --- | --- |
| accession code | mapping % |
| **SRR3196768** | **4.56** |
| **SRR3196769** | **12.92** |
| **SRR3196770** | **7.11** |
| **SRR3196771** | **73.75** |
| **SRR3196772** | **58.35** |
| **SRR3196773** | **56.96** |
| SRR3196774 | 85.23 |
| SRR3196775 | 89.24 |
| **SRR3196776** | **65.56** |
| SRR3196777 | 88.27 |
| SRR3196778 | 82.16 |
| SRR3196779 | 94.78 |
| **SRR3196780** | **77.62** |
| SRR3196781 | 96.14 |
| SRR3196782 | 81.10 |
| **SRR3196783** | **78.60** |
| **SRR3196784** | **66.43** |
| SRR3196785 | 80.55 |

There does not seem to be any relationship between quality of samples, number or length of reads that corresponds to the poor mapping quality of some of the sample.

**PRJNA295556(GSE72996)**

Both files in this dataset appeared to be of good quality with adapters already removed nevertheless fastp was still implemented with default parameters to remove lower quality sequences and allow for correct mapping. 

Following trimming with Fastp majority of reads which were removed were due to being too short as well as a small proportion of low-quality reads.



As expected, these samples mapped with high mapping percentage showed in the table below.

|  |  |
| --- | --- |
| SRR2392993 | 96.77 |
| SRR2392994 | 96.48 |