QAA Report

Varsheni

2024-09-05

Part 1 - Read quality score distributions

Here, an mRNA-seq experiment was carried out using unique dual indexes and sequenced on an Illumina HiSeq 4000. We have four files : `19_3F_fox_S14_L008_R1_001.fastq.gz`, `19_3F_fox_S14_L008_R2_001.fastq.gz`, `7_2E_fox_S6_L008_R1_001.fastq.gz`, and `7_2E_fox_S6_L008_R2_001.fastq.gz` which contain mouse embryonic fibroblasts treated with FOX. R1 is the read one file and R2 is the read 2 file for the 7_2E and the 19_3F samples respectively.

Summary Statistics

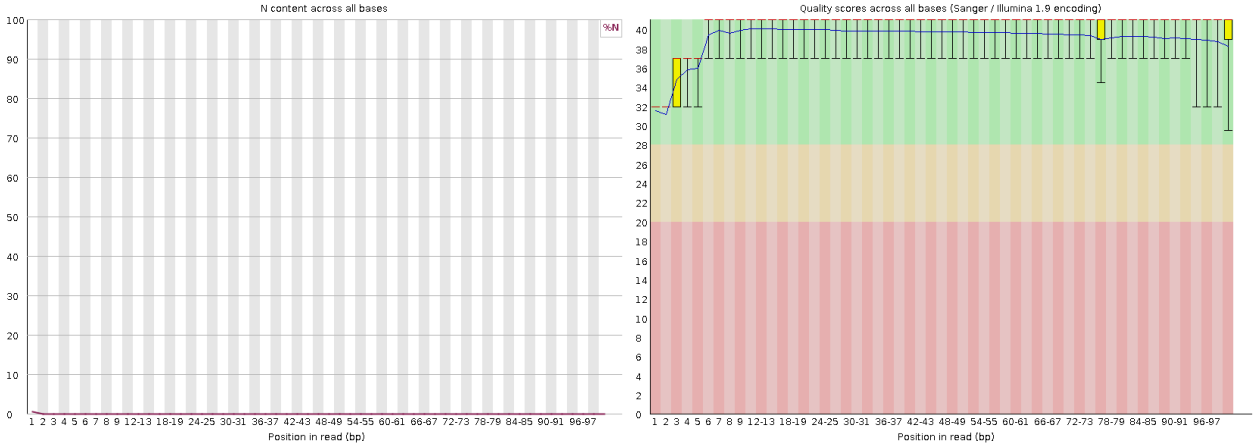
Sample	No of Reads	Length of Reads	Reads with correct matched barcodes
7_2E	5,278,425	101	5,064,906
19_3F	16,348,255	101	15,733,007

Per-base quality score and Per-base N content

We can see that for all the files, the first bases have a low quality score which corresponds with an increased observation of N-content.

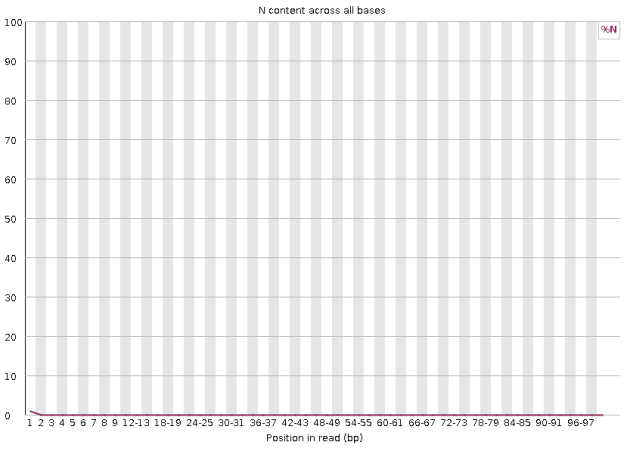
7_2E R1

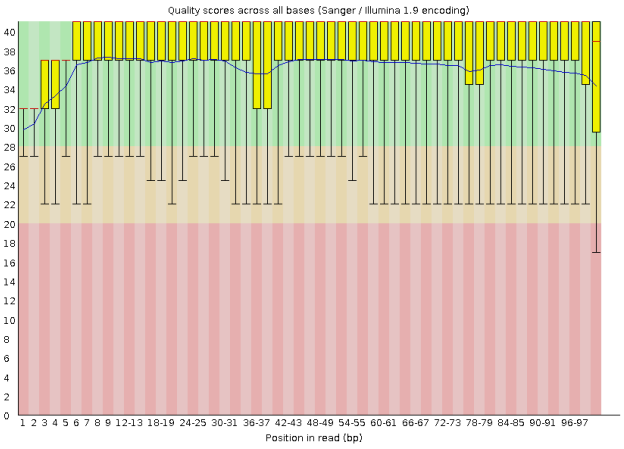
Per-base quality score _____ Per-base N content



7_2E R2

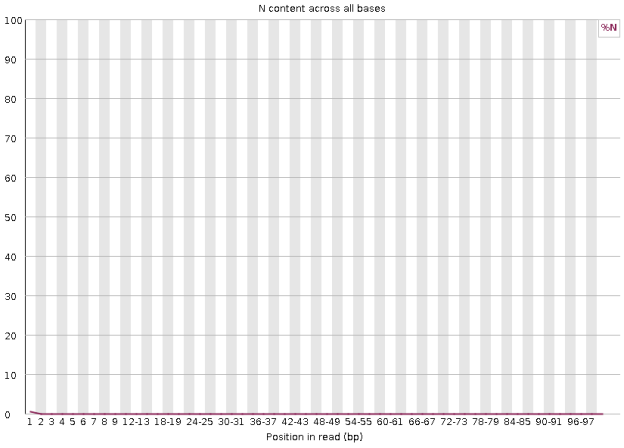
Per-base quality score _____ Per-base N content

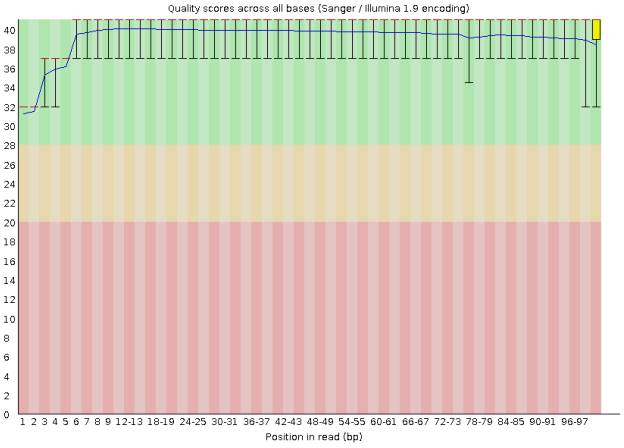




19_3F R1

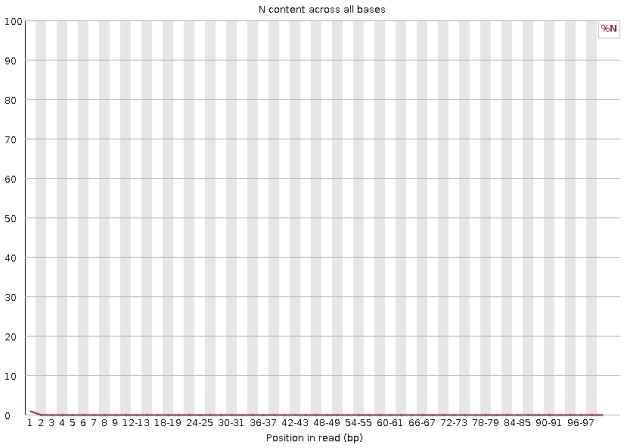
Per-base quality score _____ Per-base N content

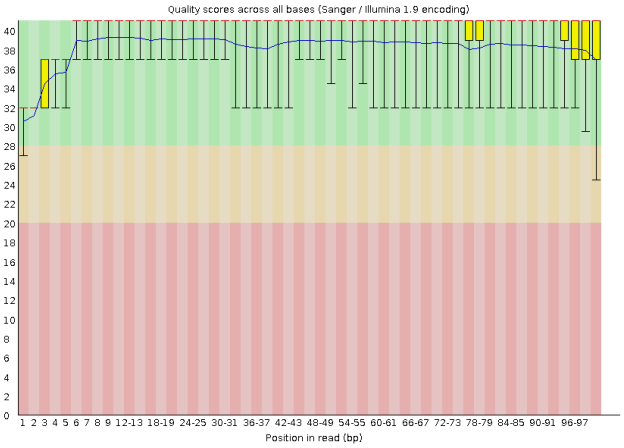




19_3F R2

Per-base quality score _____ Per-base N content





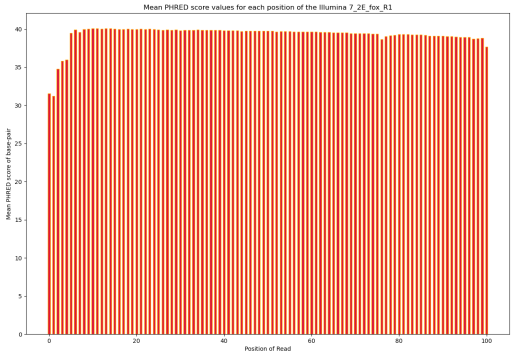
FastQC vs our Demultiplexing code

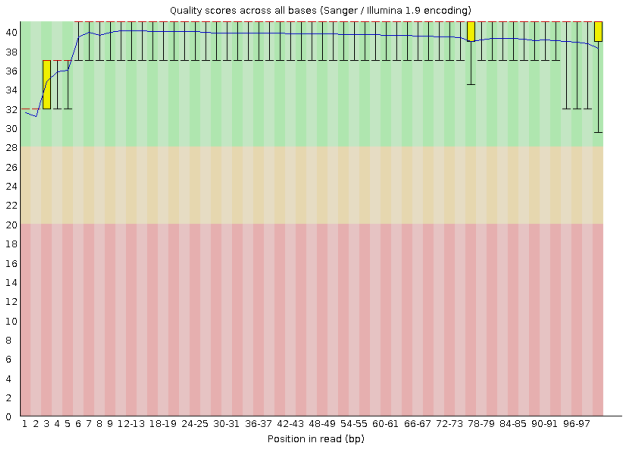
Comparisons of the Plots

The plots look similar but the FastQC one gives us much more information such as the quantile scores for each position, and marks the range for good and bad quality data.

7_2E R1

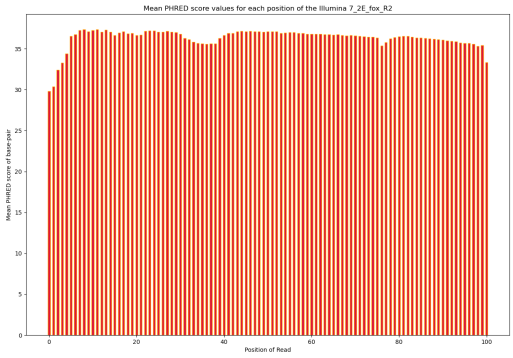
Demultiplex script _____ FastQC output

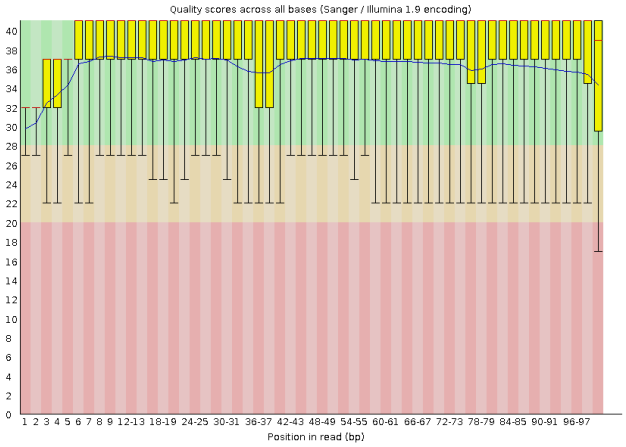




7_2E R2

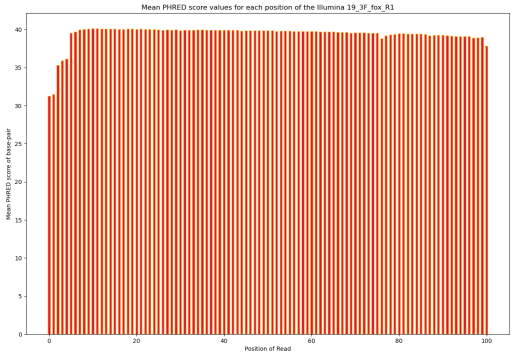
Demultiplex script _____ FastQC output

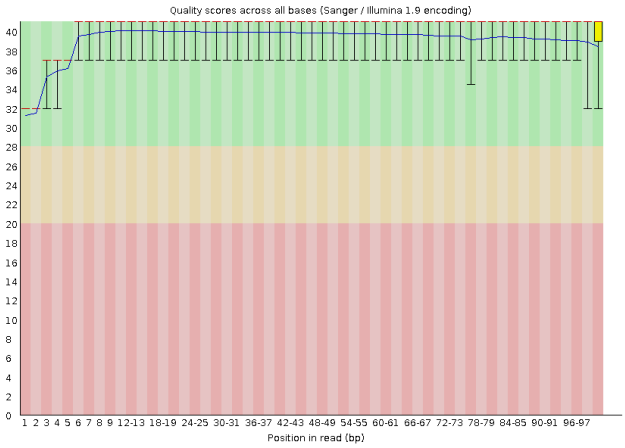




19_3F R1

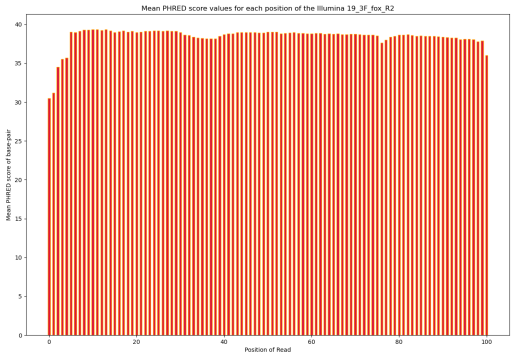
Demultiplex script _____ FastQC output

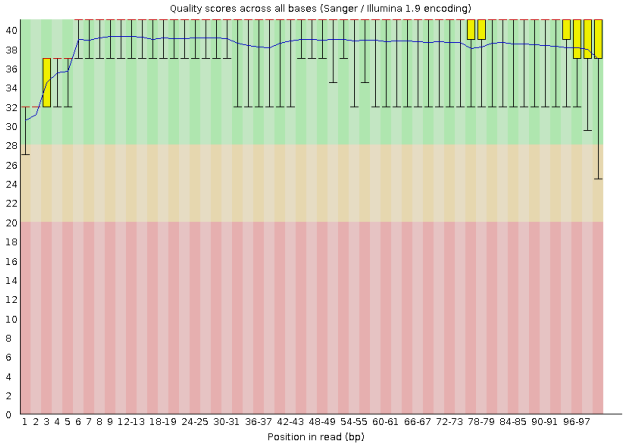




19_3F R2

Demultiplex script _____ FastQC output





Details of our Demultiplexing per base quality check for just one file

```
User time (seconds): 195.60
System time (seconds): 0.39
Percent of CPU this job got: 99%
Elapsed (wall clock) time (h:mm:ss or m:ss): 3:16.04
Maximum resident set size (kbytes): 73136
```

Details of FastQC run for all 4 files

```
User time (seconds): 166.98
System time (seconds): 11.13
Percent of CPU this job got: 251%
```

```
        Elapsed (wall clock) time (h:mm:ss or m:ss): 1:10.89
        Maximum resident set size (kbytes): 2061136
```

FastQC took much less time and much more of the computational resources than our code did. This indicates that FastQC is better equipped to handle large files and use all the computational resources. FastQC has the capability of splitting up a single process to run on multiple cores! To do this, we specified an additional argument -t indicating number of cores. Our script was not able to take advantage of the multiple threads provided to it. FastQC also has much more data analysis done, such as the per sequence quality, the GC content, etc.

FastQC summary

Quality Check	7_2E R1	7_2E R2	19_3F R1	19_3F R2
Basic Statistics	PASS	PASS	PASS	PASS
Per base sequence quality	PASS	PASS	PASS	PASS
Per tile sequence quality	**FAIL**	**WARN**	**FAIL**	**FAIL**
Per sequence quality scores	PASS	PASS	PASS	PASS
Per base sequence content	**FAIL**	**WARN**	**WARN**	**WARN**
Per sequence GC content	PASS	PASS	PASS	PASS
Per base N content	PASS	PASS	PASS	PASS
Sequence Length Distribution	PASS	PASS	PASS	PASS
Sequence Duplication Levels	**WARN**	PASS	**FAIL**	**WARN**
Overrepresented sequences	PASS	PASS	PASS	PASS
Adapter Content	PASS	PASS	PASS	PASS

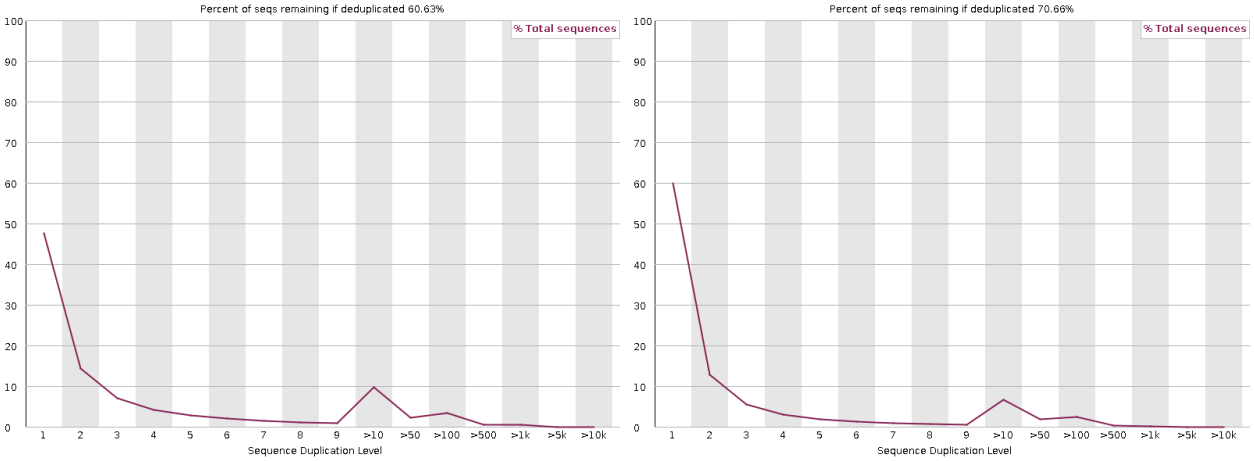
Overall Data Quality

The data seems to be of good quality.

- All sequence lengths are 101 base pairs long.
- Majority of the sequences have quality scores greater than 39.
- There is low duplication observed, there is a peak at 10x for all files but thats okay because it could be differential expression; we're looking at RNA-seq data. However, for the 19_3F data, the file failed the FastQC parameters, because non-unique sequences make up more than 50% of the total. According to the metadata, 19_3F contains significant adapter and adapter dimer peaks, which could be the reason why it contains so much duplication. However, the metadata also says that there was size selection for 300-400 bp so it probably is only represented by the differentially expressed transcripts.

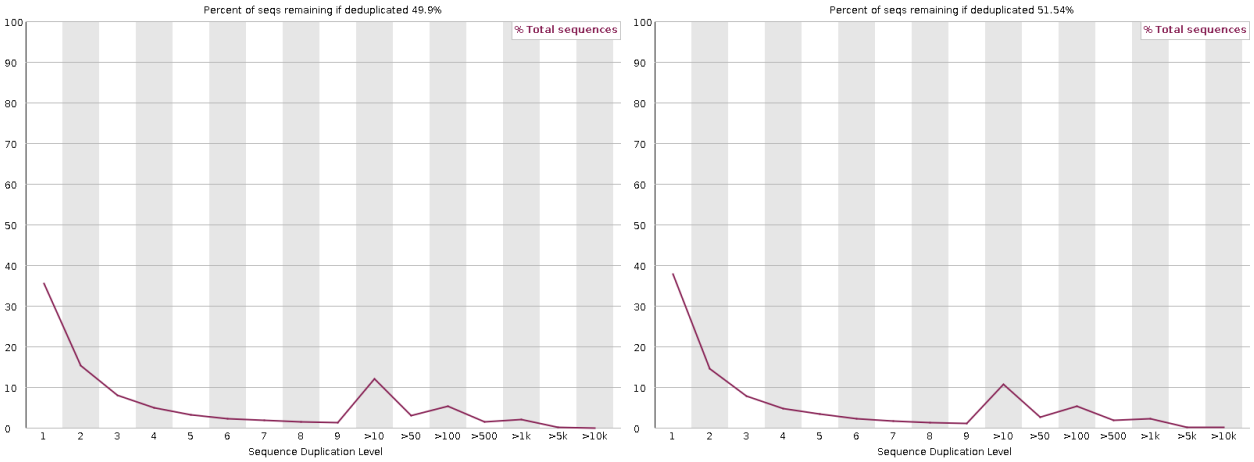
7_2E duplication levels

R1 _____ R2



19_3F duplication levels

R1 _____ R2

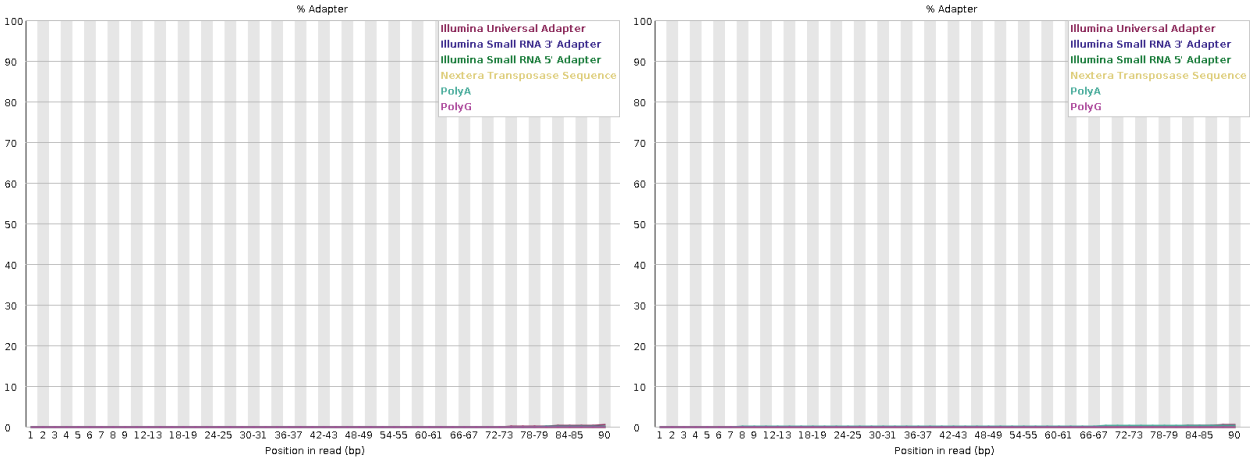


- There is a small bump of nearly ~1% of Ns for all the files for the first position of the reads, but for the other positions 2-101, there is close to no Ns. This corresponds with the first bases having lower quality scores as shown previously.

- Adapter content also seems to be super low (shown by cutadapt and trimmomatic results later as well). However they are present, near the 3‘ ends of the sequences. This further indicates that the data from 19_3F could just indicate increased expression of certain genes.

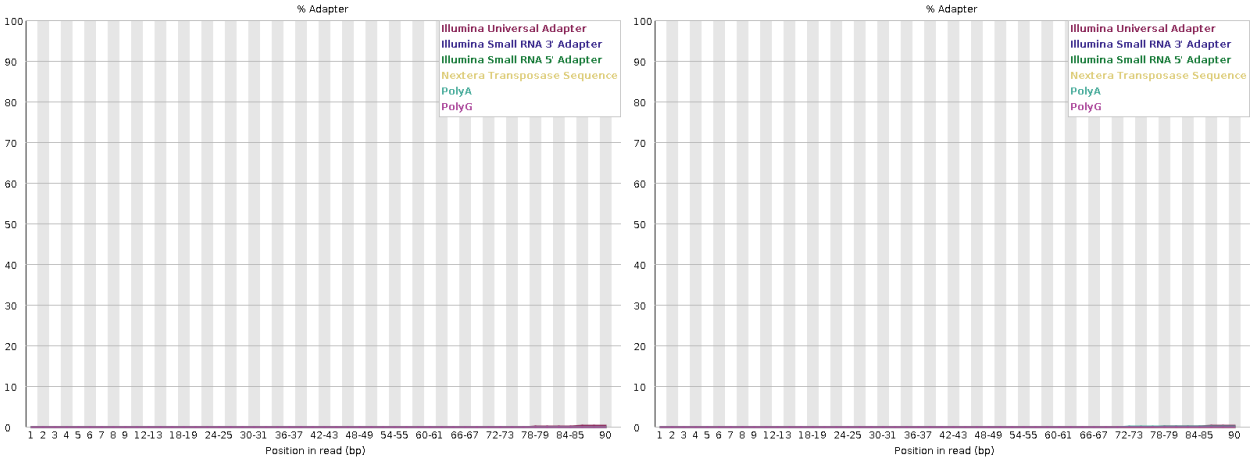
7_2E Adapter Content

R1 _____ R2



19_3F Adapter Content

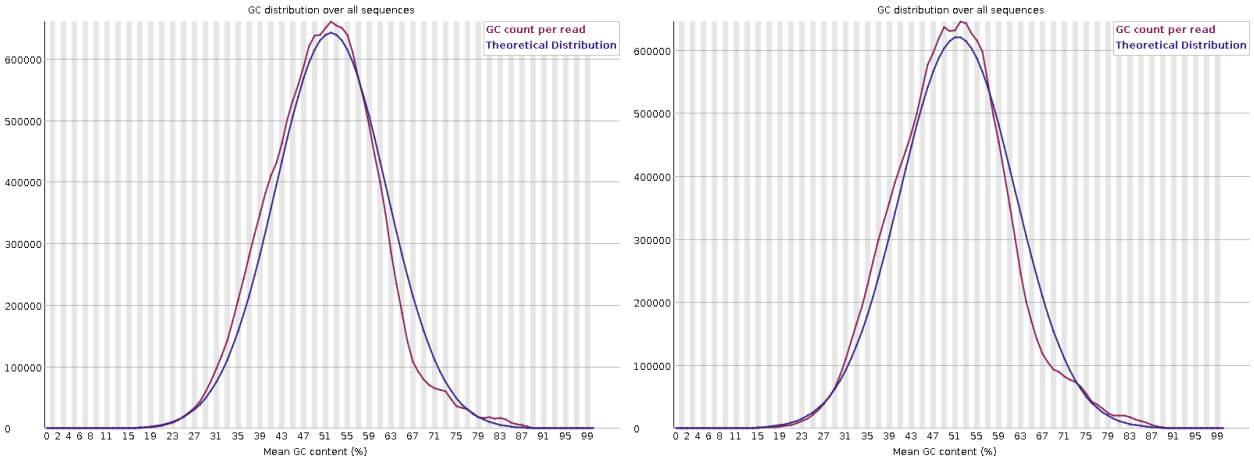
R1 _____ R2



- The GC content differs a bit from the theoretical distribution not significantly enough though.

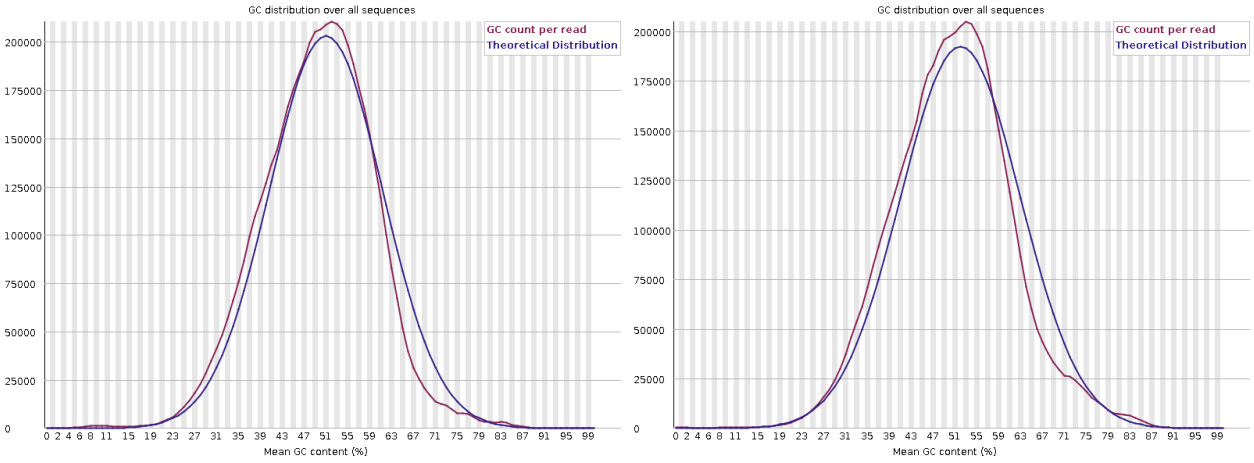
7_2E GC Content

R1 _____ R2



19_3F GC Content

R1 _____ R2

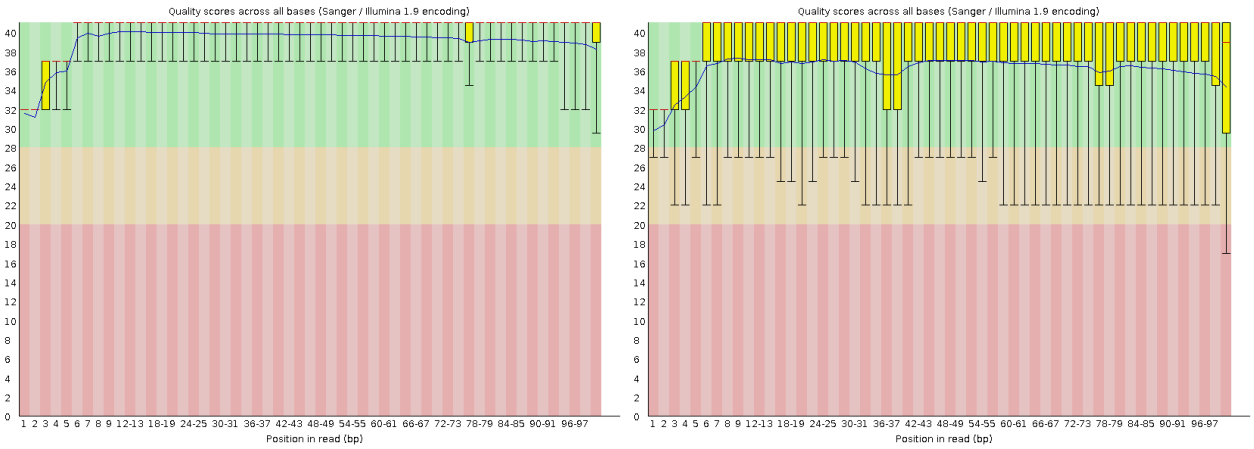


- Per base average quality for each position is greater than 32 for all files, which is good.

- Reverse reads (R2) have lower quality than Forward reads (R1) both sequence and basepair wise.

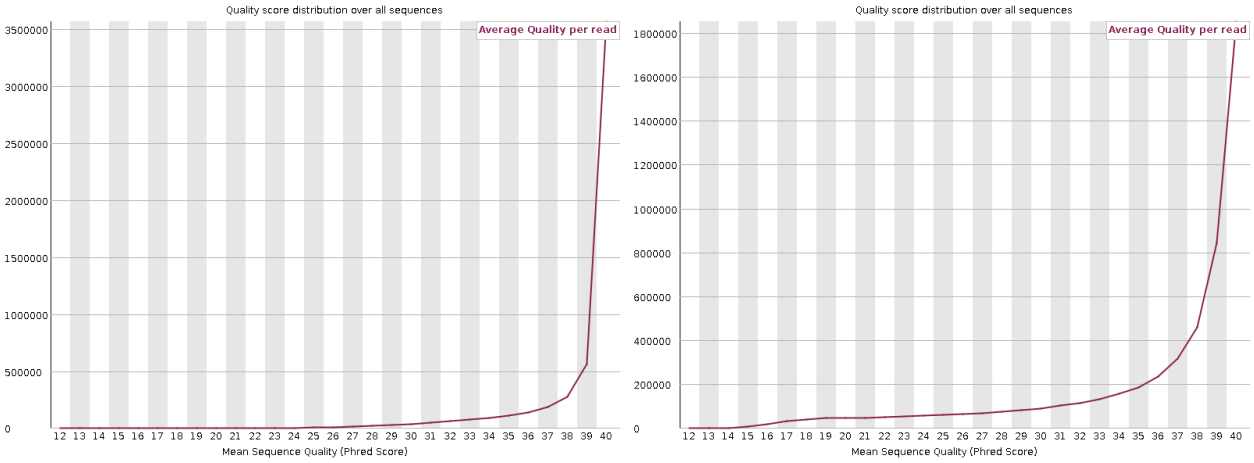
7_2E per basepair quality

R1 _____ R2



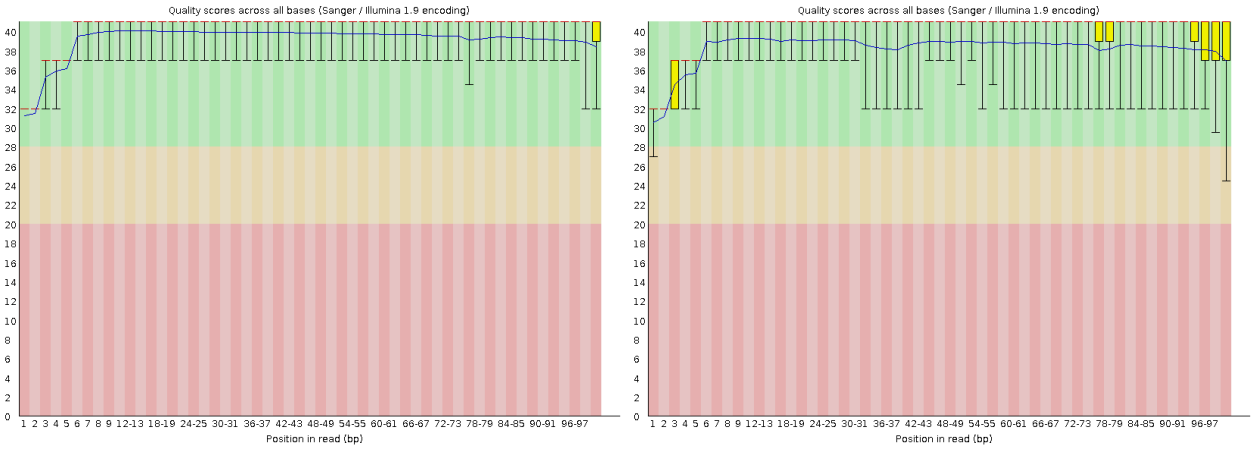
7_2E per sequence quality

R1 _____ R2



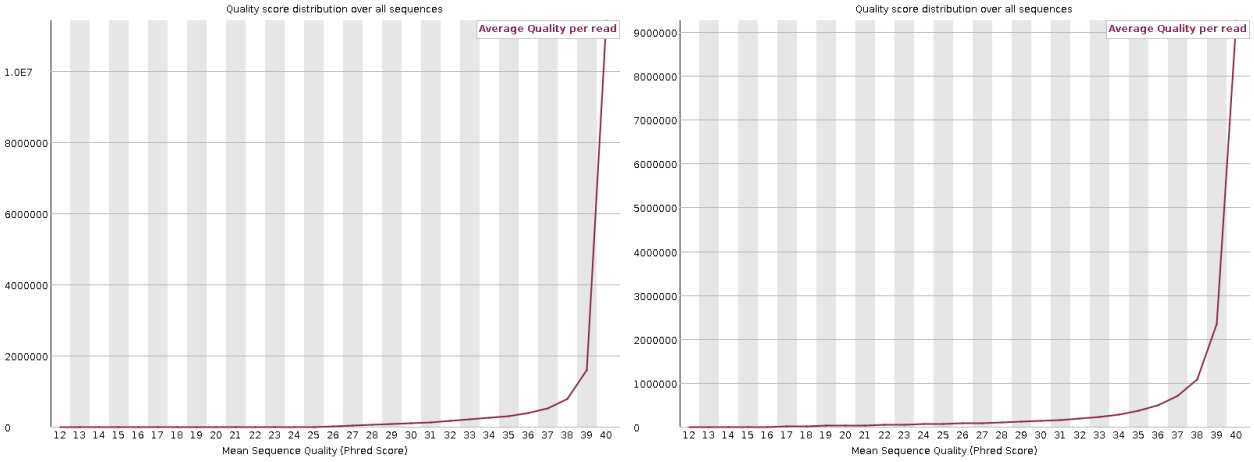
19_3F per basepair quality

R1 _____ R2



19_3F per sequence quality

R1 _____ R2



- 5 million of the 5.2 million reads have a perfect match of Barcodes for 7_2E. According to the metadata, the barcodes should be CGGTAATC+GATTACCG. 15.7 million of the 16.3 million reads have a perfect match of Barcodes for 19_3F. According to the metadata, the barcodes should be TGTTCCGT+ACGGAACA.

Overall, this data seems to be of decent enough quality for downstream analyses. There do seem to be some things to look at in case the data leads to bad output, in particular the high level of duplication in 19_3F. This can be clarified after looking at the data after adaptor trimming.

Part 2 – Adaptor trimming comparison

Proportion of reads (both R1 and R2) trimmed using Cutadapt

About the same percentage of reads were trimmed in both samples. Cutadapt didn't filter out low quality reads it just trimmed them. The number of total reads in both files remained the same.

19_3F	
Total read pairs processed:	16,348,255
Read 1 with adapter:	546,623 (3.3%)
Read 2 with adapter:	676,564 (4.1%)
Pairs written (passing filters):	16,348,255 (100.0%)

7_2E	
Total read pairs processed:	5,278,425
Read 1 with adapter:	173,473 (3.3%)
Read 2 with adapter:	212,512 (4.0%)
Pairs written (passing filters):	5,278,425 (100.0%)

Checking Adapter Sequence orientations

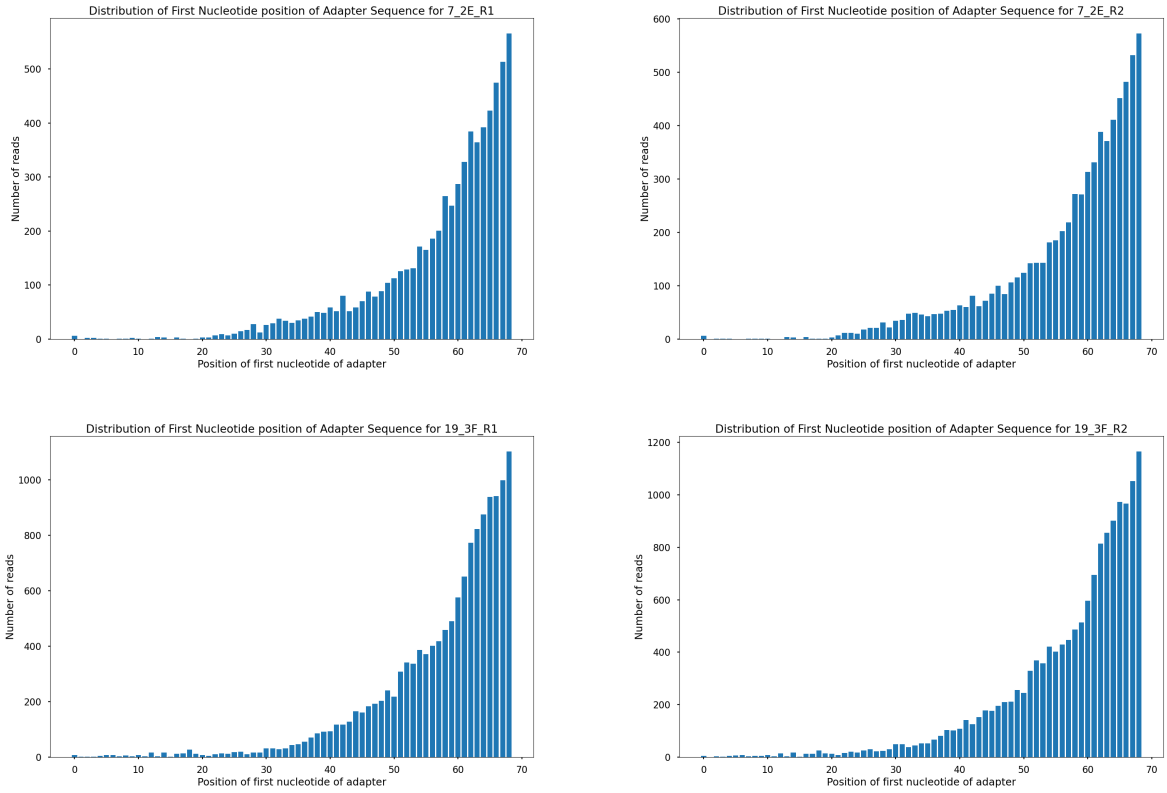
Example Command used:

```
zcat <file_name(R1/R2)> | grep <AdapterSequence(R1/R2)> |
sed -E -r 's/(.*)(<AdapterSequence(R1/R2)>)(.*)/\1/' |
awk '{print length($0)}' | sort | uniq -c | sort -nr | head -10
```

This code tells us where the adapters are present in the sequences, and lets us know the top 10 most common first nucleotide positions of the adapter sequence(n) with their frequency(f). Each column is written in the form (f n).

	7_2E	19_3F		7_2E	19_3F
R1	566 68	1103 68	R2	573 68	1166 68
	513 67	999 67		532 67	1053 67
	475 66	941 66		482 66	973 65
	423 65	939 65		452 65	967 66
	392 64	876 64		411 64	901 64
	384 62	823 63		388 62	856 63
	364 63	773 62		371 63	814 62
	328 61	651 61		331 61	695 61
	287 60	576 60		313 60	597 60
	265 58	490 59		272 58	514 59

To visualize it more clearly

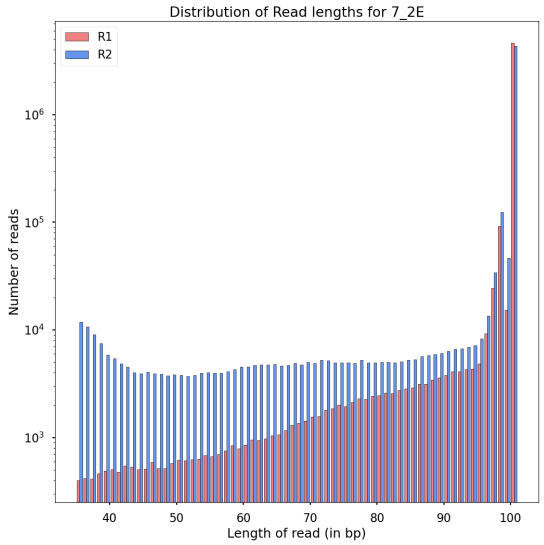


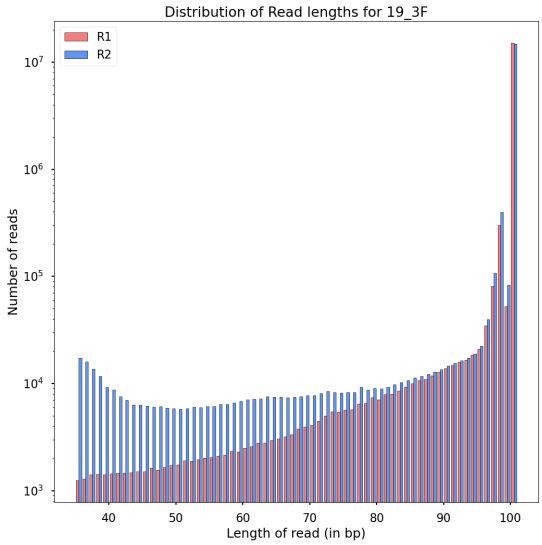
From the table and the figures above, the adapter sequence always tends to be at the end of the sequence for all 4 files

- The adapter sequence is 33 bases long
- Most of the adapter sequences begin at the 68^{th} bp position meaning that for most of the reads, the adapters are at the very end.
- In other words, most of the reads with adapters have them at the 3' end. This coincides with the results from FASTQC.

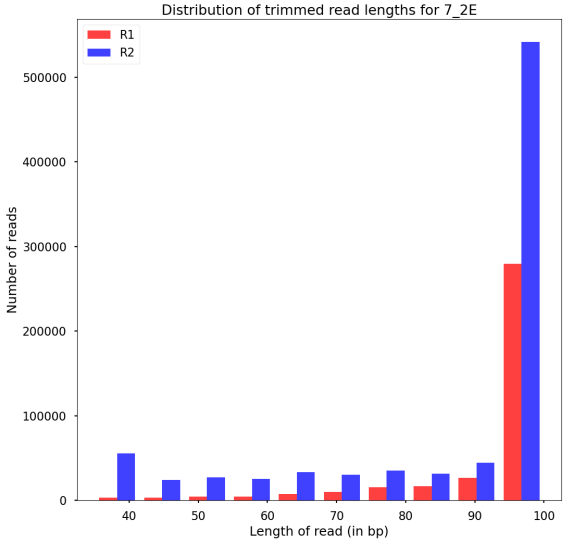
Trimmed Read Length Distributions

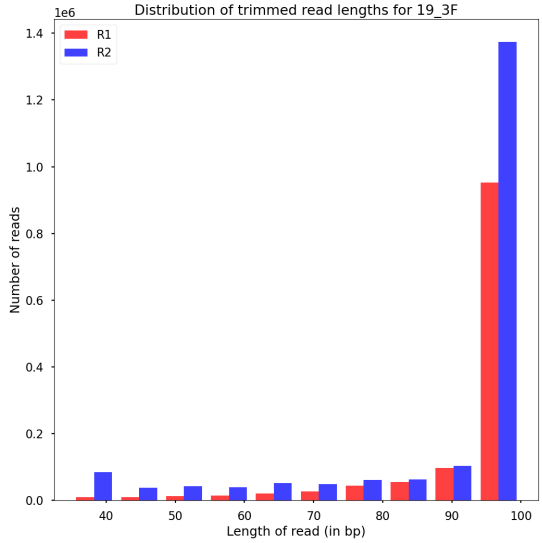
The distribution of read lengths including untrimmed reads





The distribution of only trimmed read lengths





We can see that for both samples, R2 has more trimmed reads than R1. This could be due to the fact that Reverse reads tend to have slight lower quality than forward reads, so quality trimming would affect R2 more.

R2 generally has lower quality, as R2 occurs towards the end of sequencing and reagents slightly degrade from the run time. The signal that comes from each cluster on the flowcell also gets worse over time due to more mutations during DNA synthesis.

From Trimmomatic

After running through Trimmomatic, the low quality reads were filtered out, along with quality trimming. Paired reads and unpaired reads were put into separate files for each of R1 and R2. Only the paired reads will be used for downstream analysis. The unpaired reads basically have the reads whose mate was low quality and trimmed out.

```
Percentage of reads left after Trimmomatic
-------------------------------------------
              7_2E     92.5%   (4,882,703)
              19_3F    97.2%   (15,899,268)
```

A high proportion of the reads seem to be present after filtering from both the samples.

FastQC Summary on trimmed data

Quality Check	7_2E R1	7_2E R2	19_3F R1	19_3F R2
Basic Statistics	PASS	PASS	PASS	PASS
Per base sequence quality	PASS	PASS	PASS	PASS
Per tile sequence quality	**FAIL*	* **WARN**	**FAIL**	**FAIL**
Per sequence quality scores	PASS	PASS	PASS	PASS
Per base sequence content	**FAIL*	* **FAIL**	**WARN**	**WARN**
Per sequence GC content	PASS	PASS	PASS	PASS
Per base N content	PASS	PASS	PASS	PASS
Sequence Length Distribution	**WARN**	**WARN**	**WARN**	**WARN**
Sequence Duplication Levels	**WARN**	**WARN**	**FAIL**	**WARN**
Overrepresented sequences	PASS	**WARN**	PASS	PASS
Adapter Content	PASS	PASS	PASS	PASS

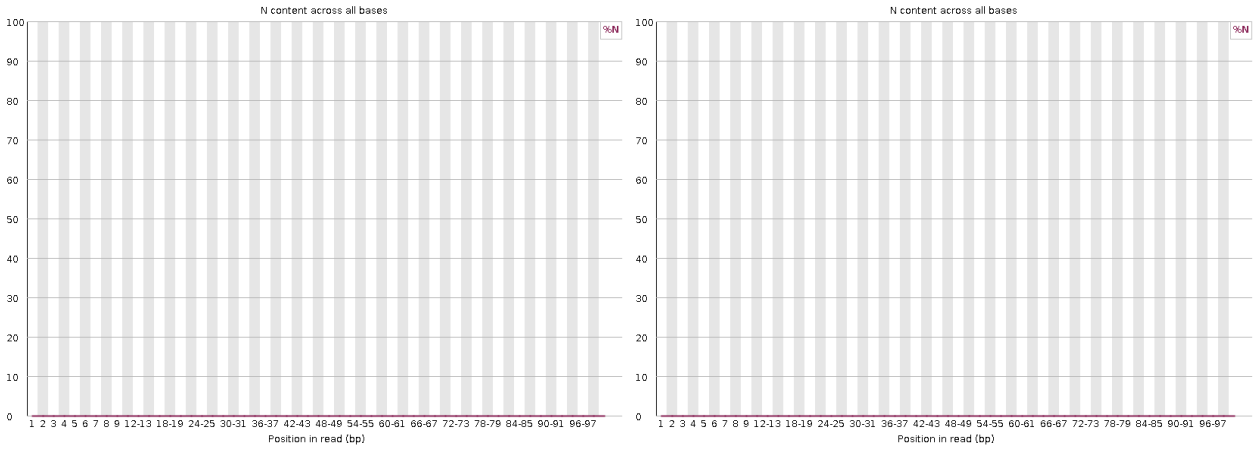
FastQC results on trimmed data: Overall Quality report

In the trimmed reads vs the normal reads

- The per-base-N content at the first positions have reduced to nearly 0% at the first position as compared to before trimming.

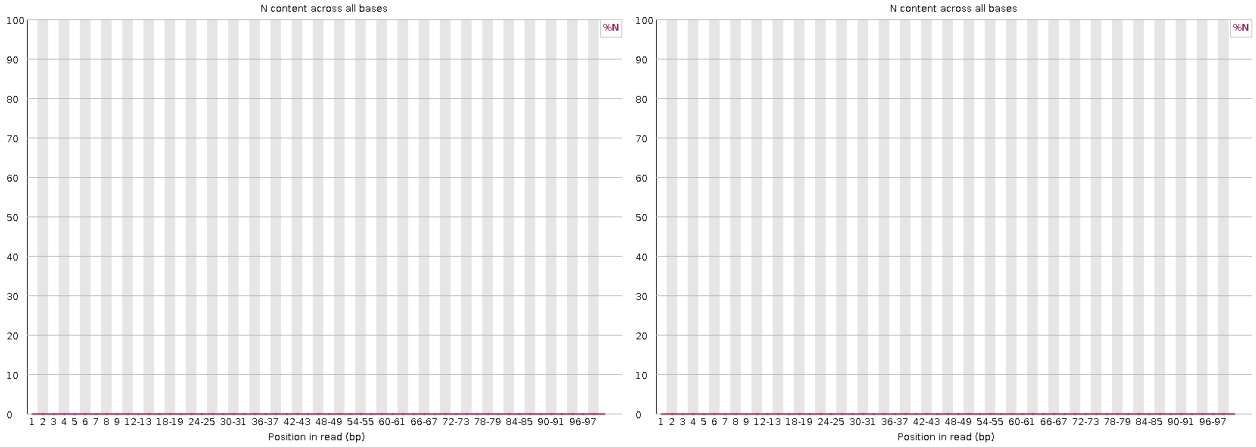
7_2E per-base-N content

R1 _____ R2



19_3F per-base-N content

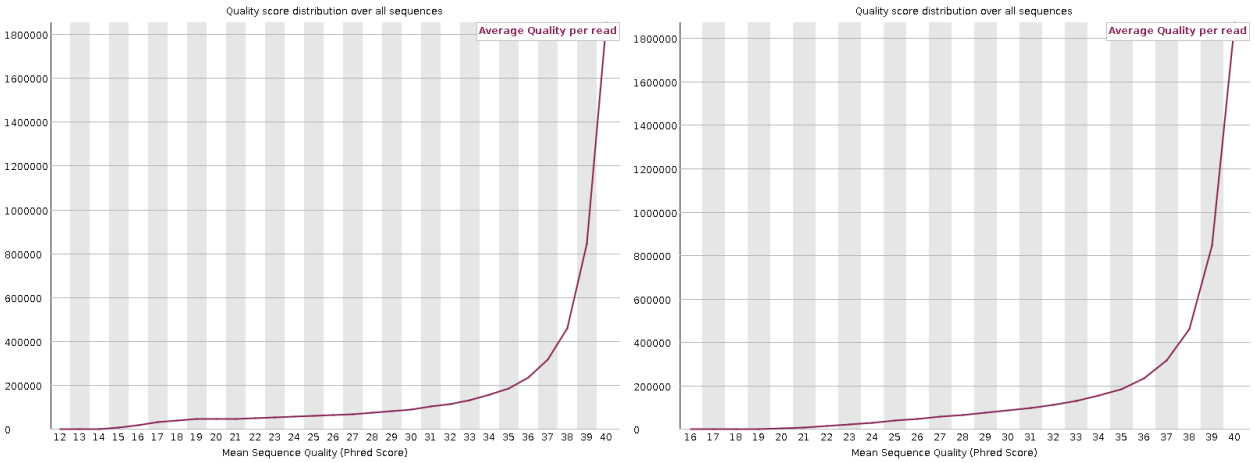
R1 _____ R2



- More sequences have high quality scores of 39 and 40, especially noticable in R2 reads.

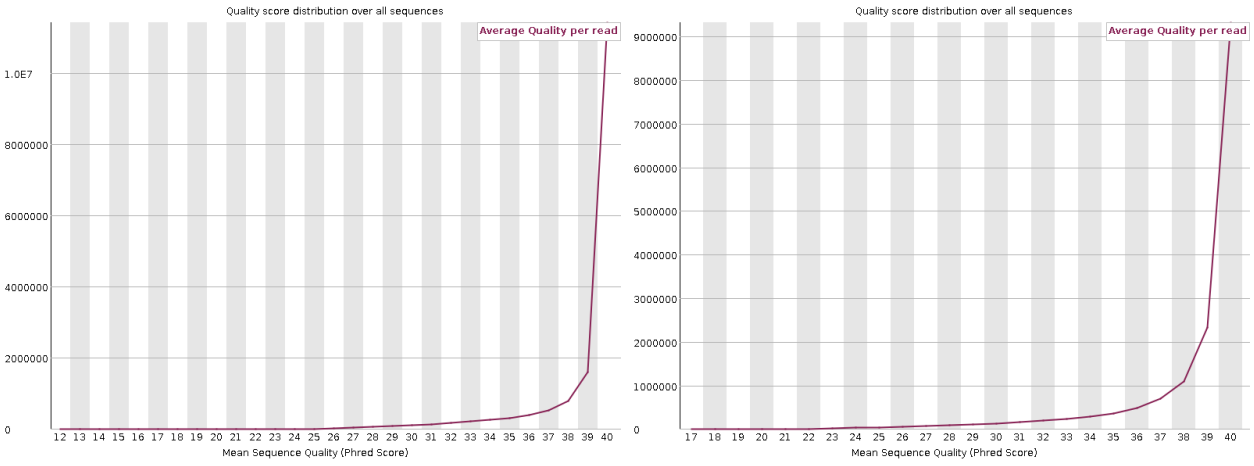
7_2E R2 per-sequence-quality

Before Trimming _____ After Trimming



19_3F R2 per-sequence-quality

Before Trimming _____ After Trimming



- More different sequence lengths in the distribution due to the trimming.
- The GC-content distribution seems to remain similar to how it was before trimming. As it passes the FastQC parameter, we can assume that the content is as expected.
- The per-base sequence quality is also very similar, if not the exact same to how it was before trimming. This could be due to the differential expression of transcripts, and hence, the trimming could have had no effect on this feature.

- The data shows decrease of reads with low sequence duplication levels in R2 reads. The number of reads which were duplicated seemed to hence make up a slightly bigger proportion of the total reads then before in the R2 reads, but not much change in the R1 reads.

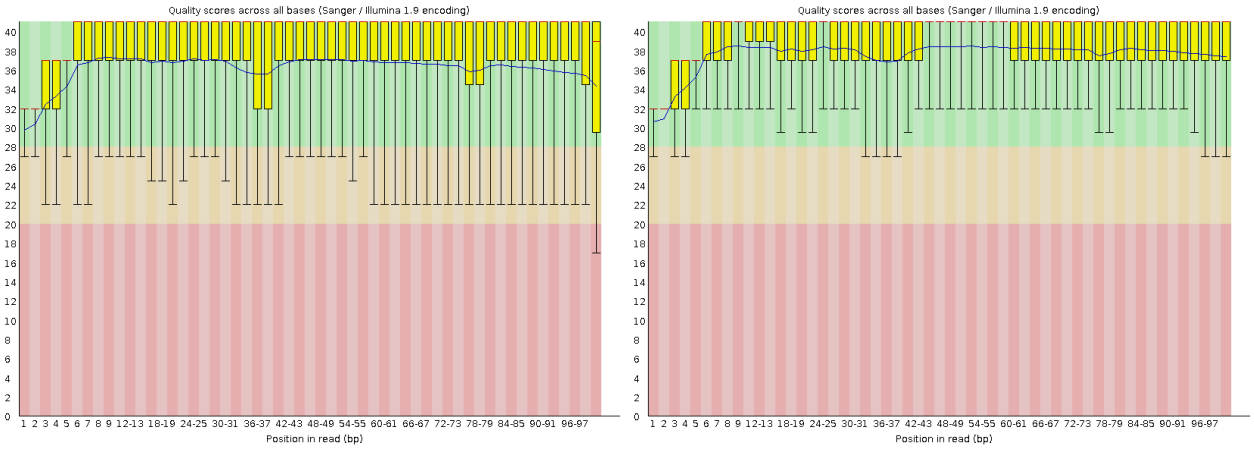
Percentage of sequences remaining if deduplicated

	Before Trimming	After Trimming
7_2E R1	60.63	61.07
7_2E R2	70.66	68.3
19_3F R1	49.9	49.84
19_3F R2	51.54	50.01

- Adapter content also improves in the trimmed reads, especially at the ends, with nearly 0% observed compared to the untrimmed data
- The per-base average quality for each position improves, reduced variance of quality score values for each base pair position.

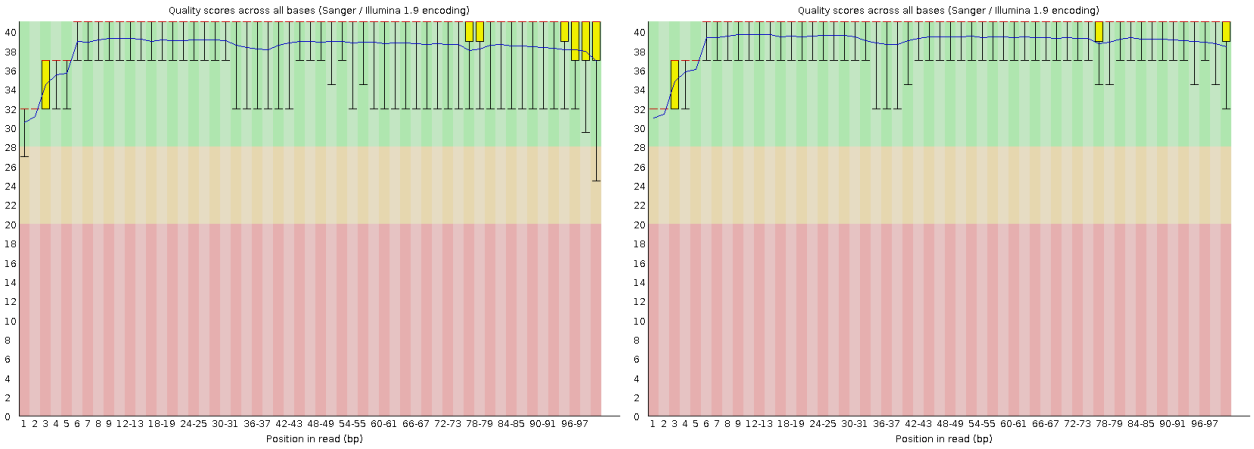
7_2E R2 per-base-quality

Before Trimming _____ After Trimming



19_3F R2 per-base-quality

Before Trimming _____ After Trimming



Overall the data quality seems to have improved after trimming.

Part 3 – Alignment and strand-specificity

Output from STAR

We used STAR, the splice aware aligner to align the reads from the 2 samples to a generated mouse genomic database (Ensemble release 112).

7_2E

Number of input reads	4882703
Average input read length	198
UNIQUE READS:	
Uniquely mapped reads number	4508671
Uniquely mapped reads %	92.34%
Average mapped length	198.34
Number of splices: Total	3154694
Number of splices: Annotated (sjdb)	3126019
Number of splices: GT/AG	3121586
Number of splices: GC/AG	25565
Number of splices: AT/AC	3429
Number of splices: Non-canonical	4114
Mismatch rate per base, %	0.27%
Deletion rate per base	0.01%
Deletion average length	3.16
Insertion rate per base	0.01%
Insertion average length	1.72
MULTI-MAPPING READS:	
Number of reads mapped to multiple loci	210695
% of reads mapped to multiple loci	4.32%
Number of reads mapped to too many loci	32788
% of reads mapped to too many loci	0.67%

19_3F

Number of input reads	15899268
Average input read length	200
UNIQUE READS:	
Uniquely mapped reads number	14500960
Uniquely mapped reads %	91.21%
Average mapped length	199.20
Number of splices: Total	10204170
Number of splices: Annotated (sjdb)	10115095
Number of splices: GT/AG	10092385
Number of splices: GC/AG	86993
Number of splices: AT/AC	11496
Number of splices: Non-canonical	13296
Mismatch rate per base, %	0.25%
Deletion rate per base	0.02%
Deletion average length	3.90
Insertion rate per base	0.01%

```
                Insertion average length |   1.82
                      MULTI-MAPPING READS:
Number of reads mapped to multiple loci |   764498
     % of reads mapped to multiple loci |   4.81%
Number of reads mapped to too many loci |   110609
     % of reads mapped to too many loci |   0.70%
```

Output from the `check_if_read_mapped.py` script

Our script counts the R1 and R2 reads as separate alignments, which is why our total number of reads according to this is double of the number outputted from Trimmomatic and after the STAR alignment.

	7_2E	19_3F
Number of mapped reads	9,424,733	30,512,167
Number of unmapped reads	340,673	1,286,369

Checking output from htseq-count

htseq-count tells us how many reads map to features. The reads which map ambiguously are not counted.

Stranded = yes

	7_2E	19_3F
total reads	4,882,703	15,899,268
mapped reads	171,207	500,167
% of mapped reads	3.5064	3.14585

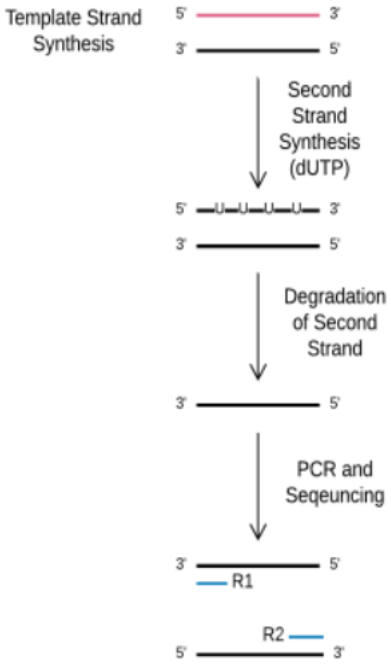
Stranded = reverse

	7_2E	19_3F
total reads	4,882,703	15,899,268
mapped reads	4,026,702	12,934,731
% of mapped reads	82.4687	81.3543

The primary intended use case for htseq-count is differential expression analysis, where one compares the expression of the same gene across samples. HTSeq-count is counting the reads, which align to the given exons. If we use the stranded option "yes", it checks whether the reads are in the same orientation as the transcript. Illumina's TruSeq Stranded protocol produces libraries, which are in reverse orientation to the transcripts' one.

> Second strand cDNA libraries were more common when htseq-count was designed than now, resulting in the default 'Yes' option.[1]

First strand library synthesis steps



Hence, data is strand specific (reverse to the "biological" reading direction) , as shown by our code and the fact that the output reads from TruSeq library preparation are known to be reverse to the "biological" reading direction (5' to 3') of the transcripts. They follow the reading direction of the DNA template strand.

Conclusions from Metadata

19_3F according to the metadata has adapter dimers and dimer peaks. However, size selection was also conducted for segments of size 300-400 bp. The sample doesn't seem to have too high of a concentration of adapters according to FastQC and the downstream analysis also looks similar for both these samples. Hence, it looks like the significant adapter and adapter dimer peak content of 19_3F doesn't cause any change in data quality as compared to 7_2E. We can hence use the 19_3F sample for the downstream analyses along with the 7_2E one.

References

[1] Srinivasan, Krishna A., Suman K. Virdee, and Andrew G. McArthur. "Strandedness during cDNA synthesis, the stranded parameter in htseq-count and analysis of RNA-Seq data." *Briefings in Functional Genomics* 19.5-6 (2020): 339-342.