QAA Report

Isis Diaz

2022-09-07

General Objective

The objective of this assignment is to use existing tools for quality assessment and adaptor trimming, compare the quality assessments to those from your own software

Data analyzed

For this assignment I used two different library sequences produced by 2017 cohort on mouse RNA-Seq. It's important to mention that this data was already demultiplexed before performing any analysis.

7_2E_fox_S6_L008

19_3F_fox_S14_L008

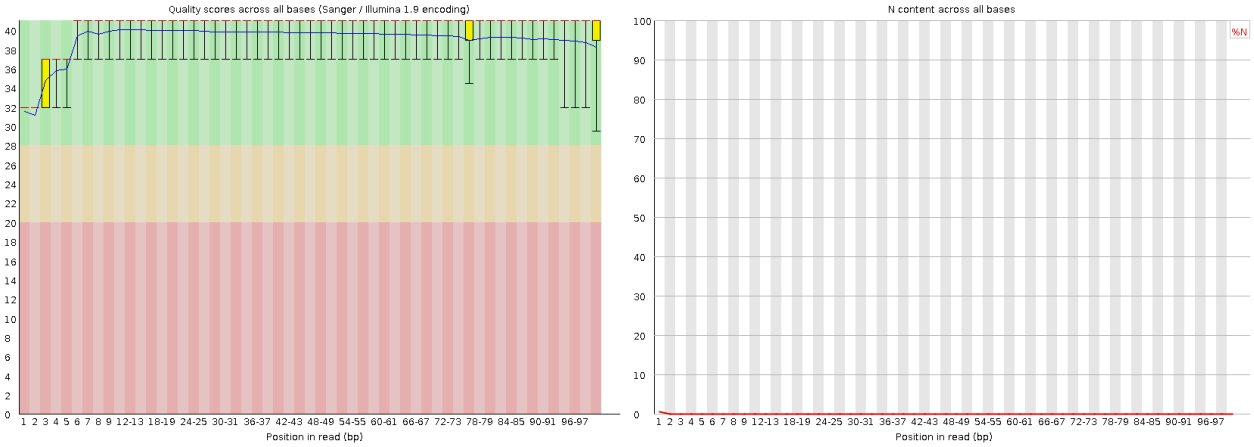
Each file had a forward (R1) and reverse (R2) sequence file, since the sequencing was performed paired end

Part 1

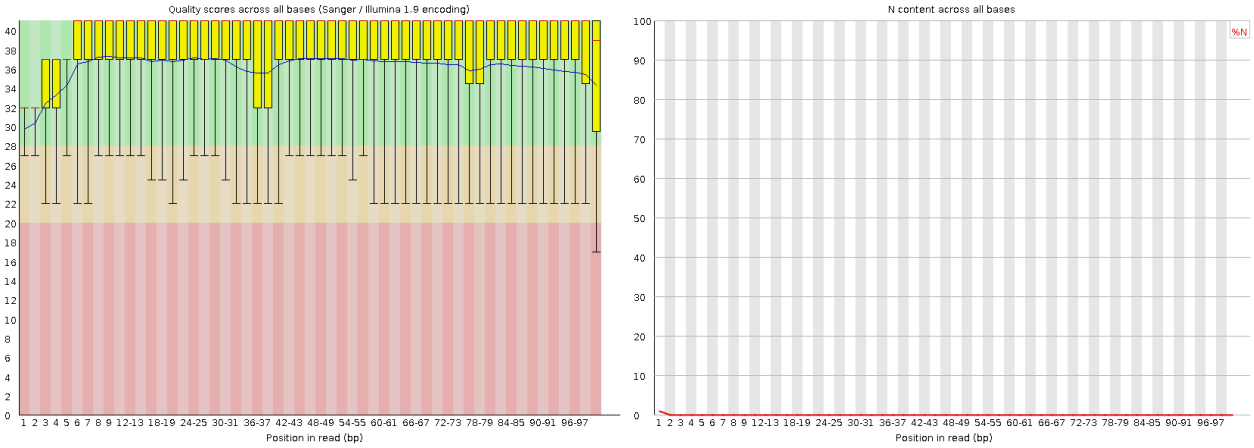
In this part of the assignment we'll compare the quality distribution plots obtained from FastQC and the ones obtained from the distribution code I wrote for Bi622; *Demultiplex/Assignment_The_First*

First I present the graphs obtained from FastQC, which are quality per base position and the ammount of non identified nucleotides (N) in each base position.

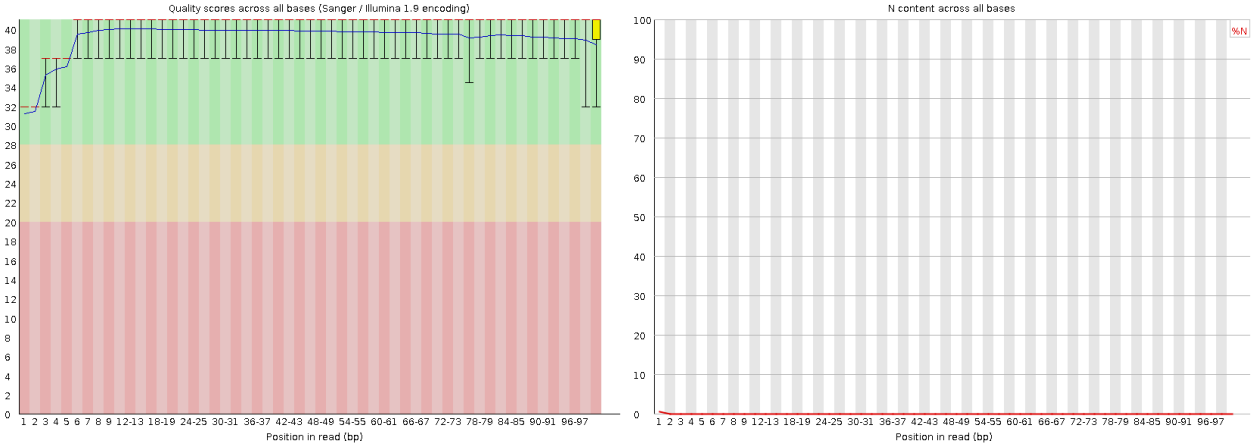
7_2E_fox_S6_L008 R1 (Forward)



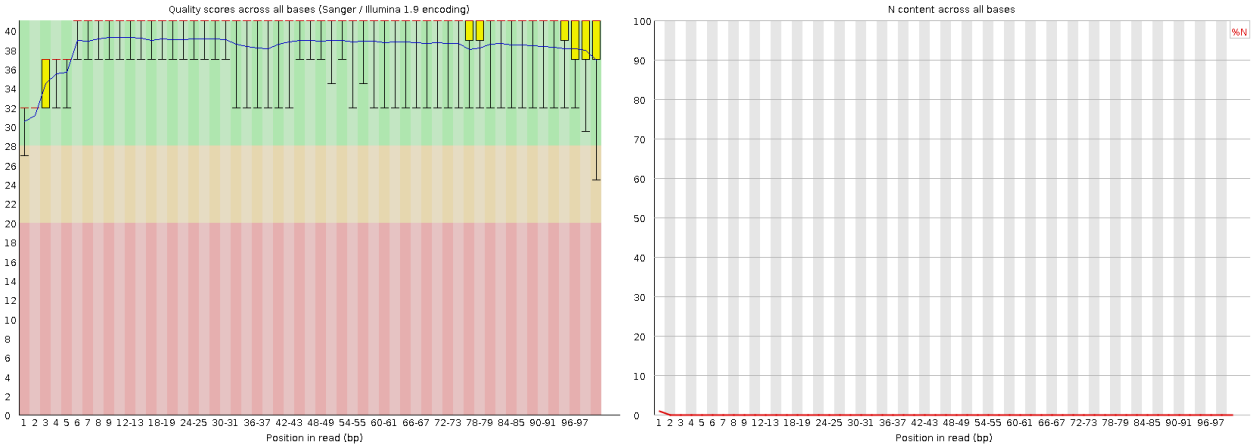
7_2E_fox_S6_L008 R2 (Reverse)



19_3F_fox_S14_L008 R1 (Forward)



>19_3F_fox_S14_L008 R2 (Reverse)

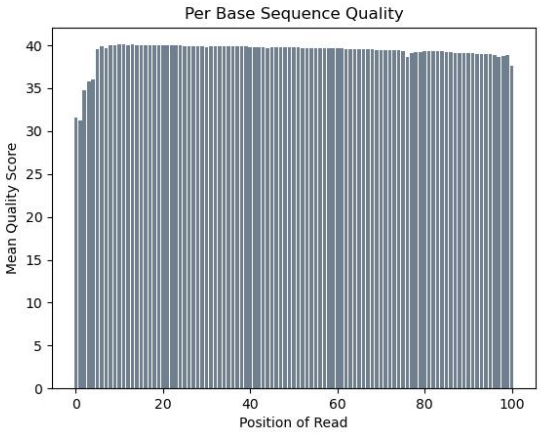


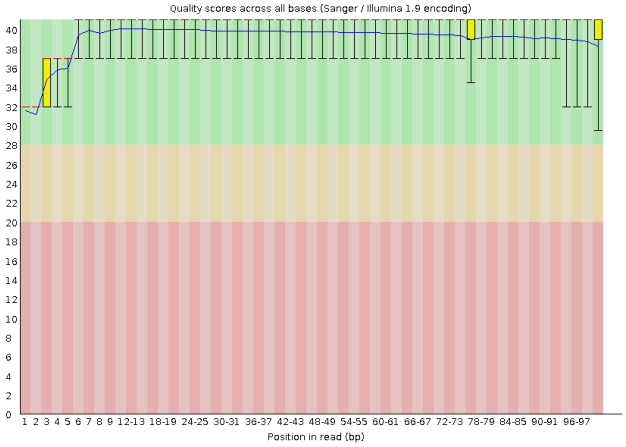
We can see that in all of the graphs we see a low quality in the first base positions of the reads which is consistent with the proportion of N content, since a non identified nucleotide will bring the overall quality score down. The quality at the start of a sequence read for RNA-Seq is known to be low, and this has been

theorized to be caused by different effects like random primer processing, imaging issues near the flow cell, phasing, etc.

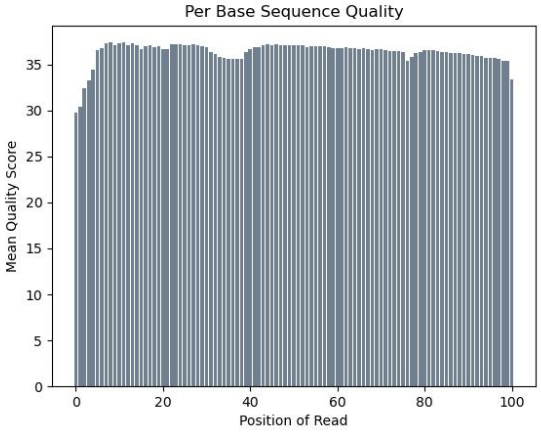
Now I'll present the graphs for quality per base position created from my previously assignment *Demultiplex* compared to the ones created by FastQC

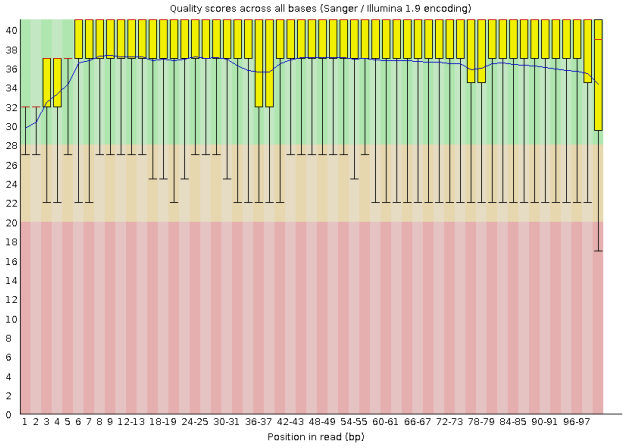
7_2E_fox_S6_L008 R1 (Forward)



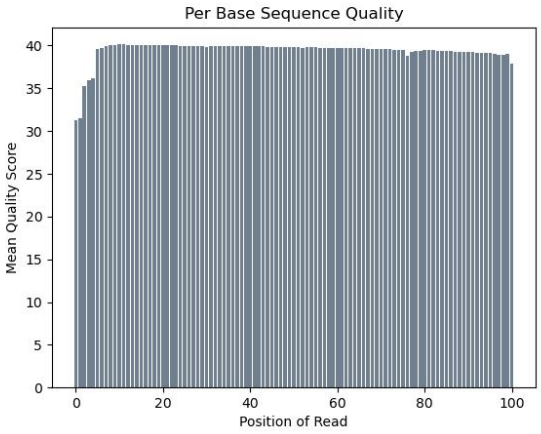


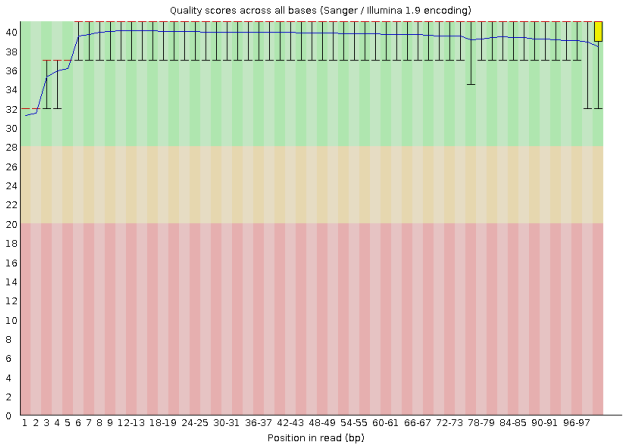
7_2E_fox_S6_L008 R2 (Reverse)



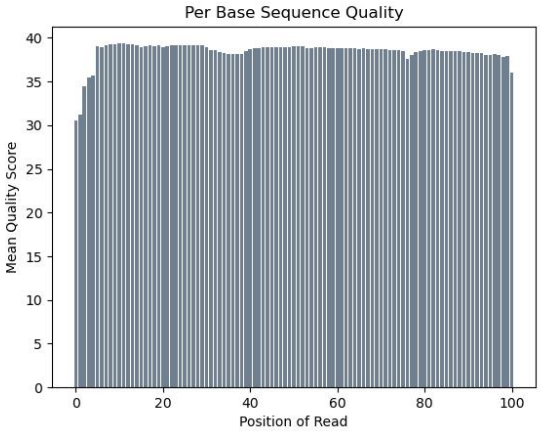


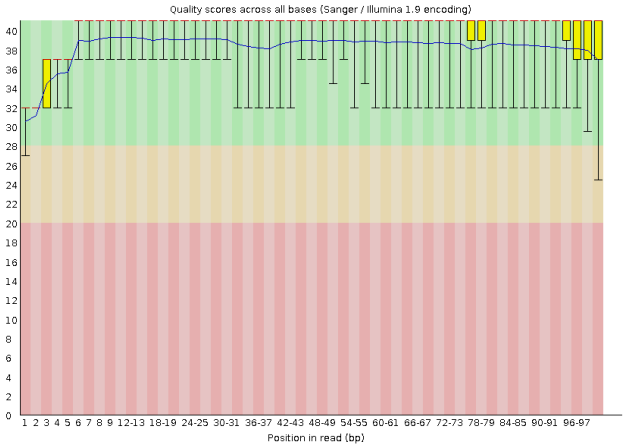
19_3F_fox_S14_L008 R1 (Forward)





19_3F_fox_S14_L008 R2 (Reverse)





Overall Graph Comparison In both graphs we can see the same spikes, the only difference is that FastQC has been developed to be read easier, so its more detailed and with more visual cues for the user.

Runtime Difference FastQC Timing for both files (forward and reverse) was:

- 1:36.37 (one minute with ~36 seconds)

Distribution program for both files (forward and reverse) was:

- 6:16:12 (six minutes with ~16 seconds)
- 18:30 (eighteen minutes with ~30 seconds)

FastQC is far superior in runtime and this could be due to several parameters

- The Distribution program only takes one data set at a time, affecting runtime. FastQC runs all of the data in parallel
- FastQC is multithreaded improving its runtime
- FastQC was developed in C++ with a faster code compilation making it faster than python, which is where the distribution program was developed.

Overall Data Quality The quality data for the libraries is high except for the start positions of the reads, as mentioned prior this is expected due to the Illumina sequencing technical issues.

There is a lower quality for the Reverse Reads (R2 data), but this can be due to the waiting time for processing. R1 will be sequenced first and once it finish sequencing, the strand will be reversed and R2 will be sequenced. Meaning that when starting to sequence R2 the DNA has been in the flow cell for a longer time and could have degraded, decreasing its quality.

Part 2

In this part of the assignment we'll remove the adapters and trim low quality reads and create plots to view the distribution of the length of remaining reads.

The first processing step for this part was to determine the adapter sequences from the reads, to determine this I looked at the known Illumina sequence adapters and compare them to the sequences in our data.

The corresponding adapters are from IDT for Illumina-TruSeq DNA and RNA UD Indexes kit:

Read1 (Forward)

- AGATCGGAAGAGCACACGTCTGAACTCCAGTCA

Read2 (Reverse)

- AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT

Cutadapt Cutadapt is a program to remove adapters, in this case we removed the adapters considering a pair-end read.

=== Summary 7_2E_fox_S6_L008===

- Total read pairs processed: 5,278,425
  - Read 1 with adapter: 56,771 (1.1%)
  - Read 2 with adapter: 56,771 (1.1%)
- Total basepairs processed: 1,066,241,850 bp
  - Read 1: 533,120,925 bp
  - Read 2: 533,120,925 bp
- Total written (filtered): 1,064,222,649 bp (99.8%)
  - Read 1: 532,111,677 bp
  - Read 2: 532,110,972 bp

Cutadapt summary reveals that for reads 7_2E_fox_S6_L008, it removed 2,019,201 of basepairs (not written) which accounts for 0.1893755% of total basepairs processed

=== Summary 19_3F_fox_S14_L008===

- Total read pairs processed: 16,348,255
  - Read 1 with adapter: 193,497 (1.2%)
  - Read 2 with adapter: 193,497 (1.2%)

- Total basepairs processed: 3,302,347,510 bp
  - Read 1: 1,651,173,755 bp
  - Read 2: 1,651,173,755 bp
- Total written (filtered): 3,296,944,816 bp (99.8%)
  - Read 1: 1,648,472,709 bp
  - Read 2: 1,648,472,107 bp

Cutadapt summary reveals that for reads 19_3F_fox_S14_L008, it removed 5,402,694 of base pairs (not written) which accounts for 0.1636016% of total base pairs processed.

Overall 7_2E_fox_S6_L008 had more of a percentage filtered.

Trimmomatic After removing the adapters we can run trimmomatic. Trimmomatic-0.39 is a tool to trim reads, the triming type will be determined by the arguments used for the data. Bellow I also added the arguments that are going to be used according with the instructions:

- LEADING: quality of 3
- TRAILING: quality of 3
- SLIDING WINDOW: window size of 5 and required quality of 15
- MINLENGTH: 35 bases

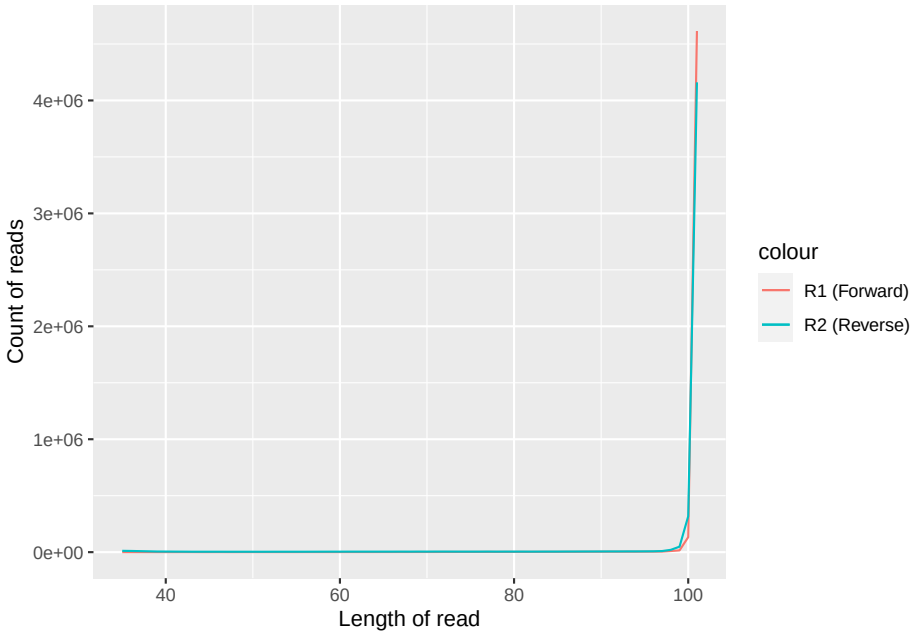
What this arguments do is specified bellow:

- Remove leading low quality or N bases (below quality 3)
- Remove trailing low quality or N bases (below quality 3)
- Scan the read with a 5-base wide sliding window, cutting when the average quality per base drops below 15
- Drop reads below the 35 bases long

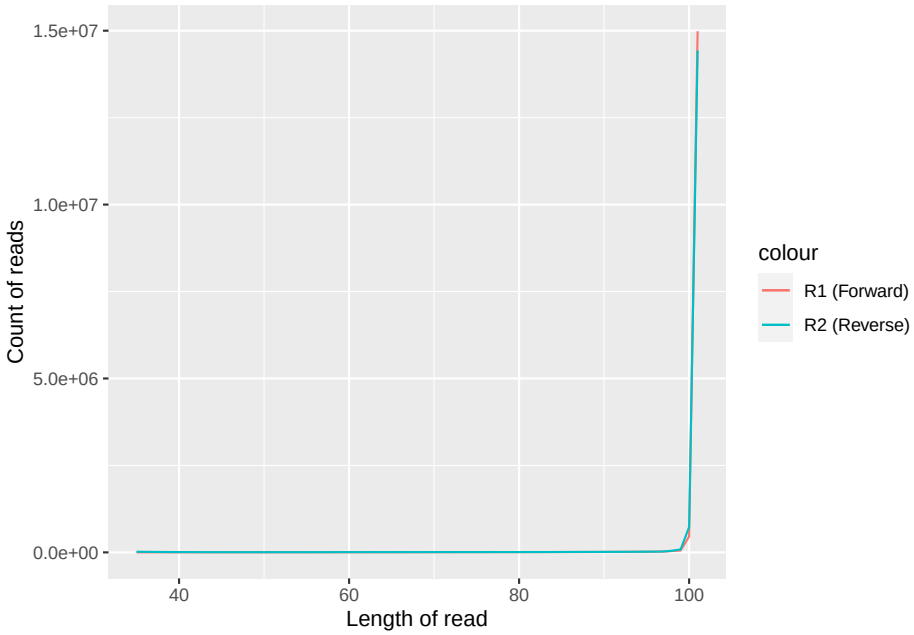
After processing our reads I created Distribution graphs for both files (Forward and Reverse in same graph)

Distributions

Plot for 7_2E_fox_S6_R2_distribution.txt



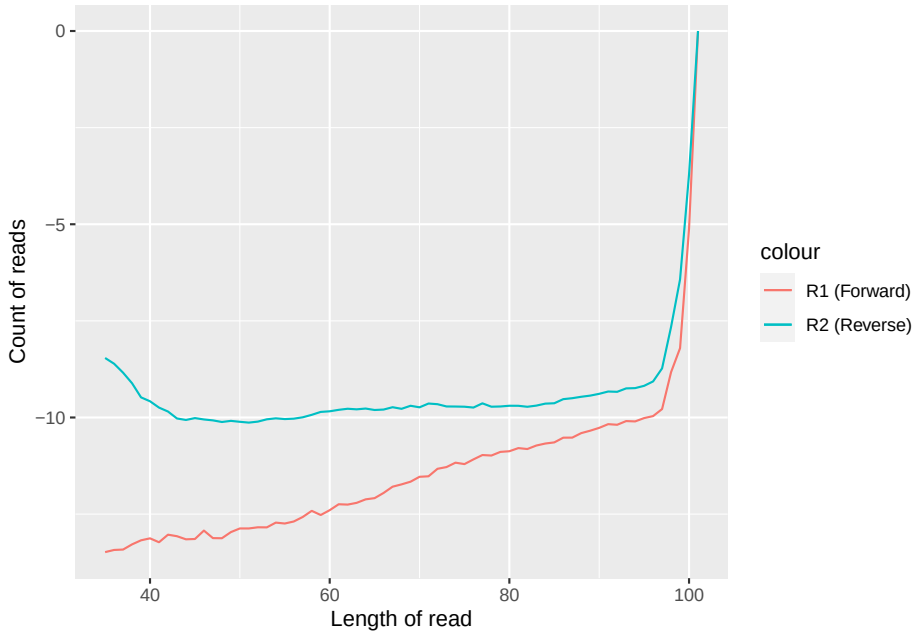
Plot for 19_3F_fox_S14_distribution.txt



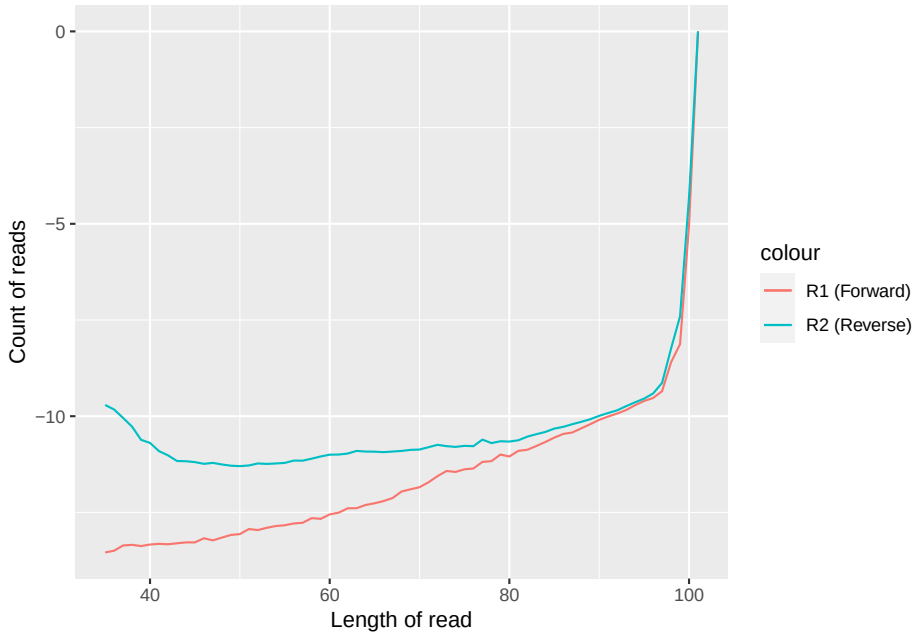
Log Transform Distributions

This graphs are created to be able to visualize the data better than the normal distribution of information.

Plot for distribution of 7_2E_fox_S6



Plot for distribution of 19_3F_fox_S14



As Expected we see a bigger length reduction in R2 reads and that is because of the trimming of low quality reads, which as shown in Part1 R2 had a lower quality than R1 reads. But the length distribution of read is still mostly concentrated at 101 bp which is more desirable.

Part 3

The last part of this assignment is to align our sequence read to the genome and determine if the sequencing was strand-specific.The sequence reads we are working with are from Mouse RNA, so we would need to align to the current Mouse Genome (Ensemble release 107).

In this case I use the STAR software (STAR-2.7.10a) to create a Database and then align the sequences to the genome. We obtained on sam file per sequence (forward and reverse combined and identified through bit flags). I used a previously developed code for Bi621; *PS8* which will identified the amount of reads that align to the genome and those that where not mapped:

Condition	7_2E_Mus_musculus	19_3F_Mus_musculus
Properly Mapped	9424344	30509928
Not Mapped	341296	1289070

But whether they map or not doesn't help us determine accurately if the sequencing was performed in a strand-specific fashion, to aid us with this determination we could use the htseq software which determines to what it aligns and whether is related to a gene. I ran htseq two times per file to account for forward strandedness and reverse strandedness. The resulting counts for alignment to genes is the one shown below:

Sequence	Mapped to Gene
7_2E_Mus_musculus (Forward)	182741
7_2E_Mus_musculus (Reverse)	4027416
19_3F_Mus_musculus (Forward)	555542
19_3F_Mus_musculus (Reverse)	12936768

Now we can determine that the sequencing was strand-specific and we can imply this by the vast difference between the gene alignment done in the files for forward and reverse strandedness.The reverse for 7_2E_Mus_musculus has 22.0389294 times the alignment to genes than it's forward strand, and the reverse for 19_3F_Mus_musculus as 22.0389294 times the alignment to genes than it's forward strand. So we can confidently say that R2 is the specific strand of which the RNA-Seq was based on for this experiment.