QAA

Cameron Kunstadt

2024-09-10

RNA-seq Quality Assessment Assignment - Bi623

Cameron Kunstadt

Part 1

Project involves these 4 files located on talapas, consisting of the R1 and R2 of 2 RNA-seq Datasets.

```
/projects/bgmp/shared/2017_sequencing/demultiplexed/21_3G_both_S15_L008_R1_001.fastq.gz
/projects/bgmp/shared/2017_sequencing/demultiplexed/21_3G_both_S15_L008_R2_001.fastq.gz
/projects/bgmp/shared/2017_sequencing/demultiplexed/16_3D_mbnl_S12_L008_R1_001.fastq.gz
/projects/bgmp/shared/2017_sequencing/demultiplexed/16_3D_mbnl_S12_L008_R2_001.fastq.gz
```

FastQC, as well as a previously developed plotting script was ran on each of the files to get these plots:

Figure 1: 21_3G_both_S15_L008_R1_001.jpg

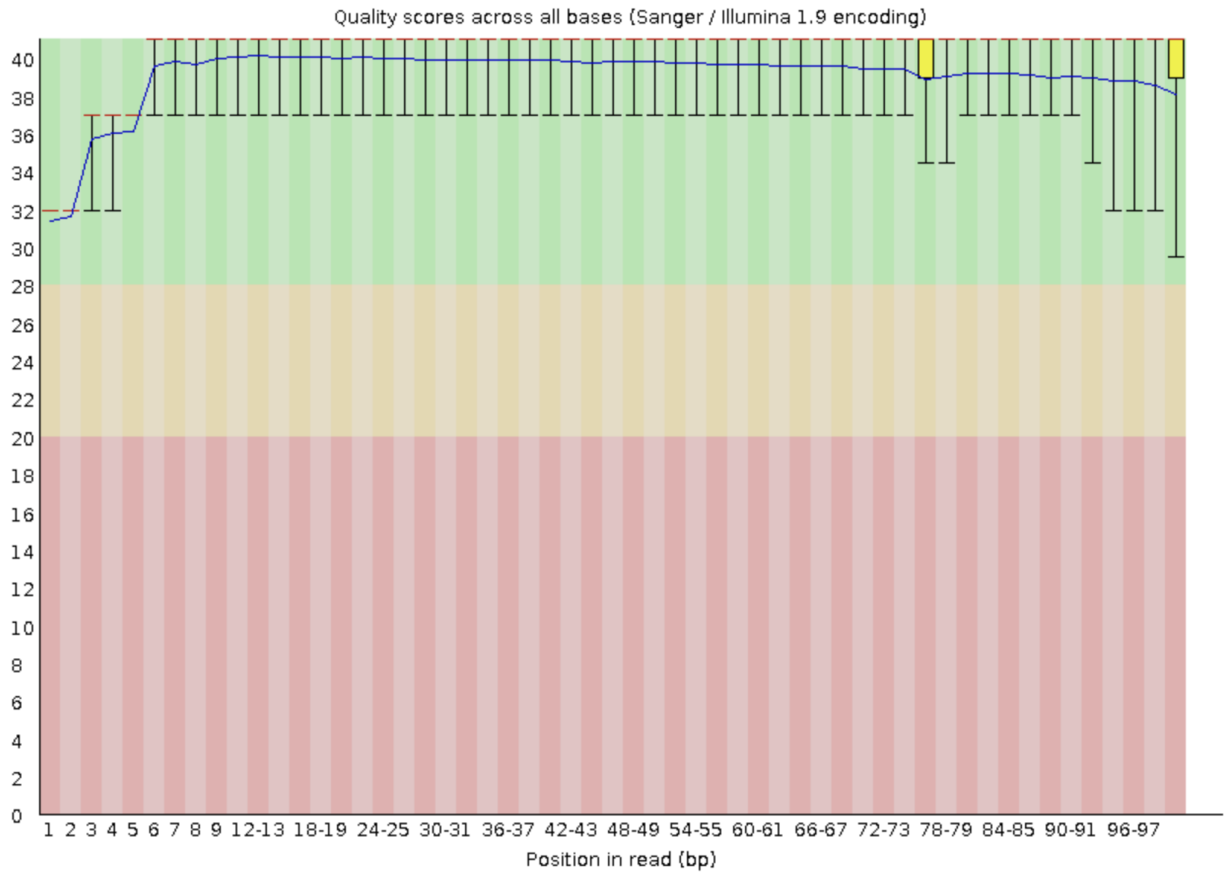

Figure 2: 21_3G_both_S15_L008_R1_001.jpg

Figure 3: 21_3G_both_S15_L008_R2_001.jpg

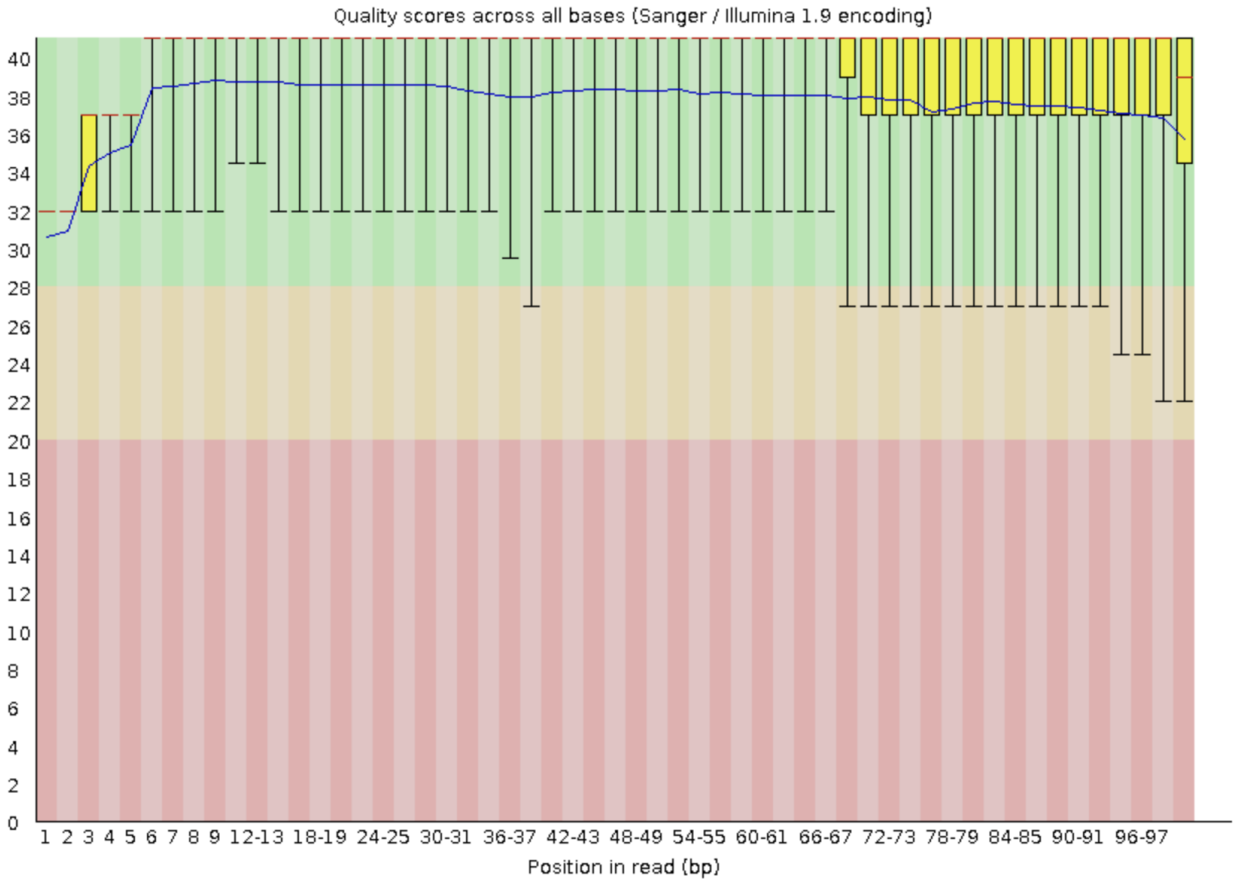

Figure 4: 21_3G_both_S15_L008_R2_001.jpg

Figure 5: 16_3D_mbnl_S12_L008_R1_001

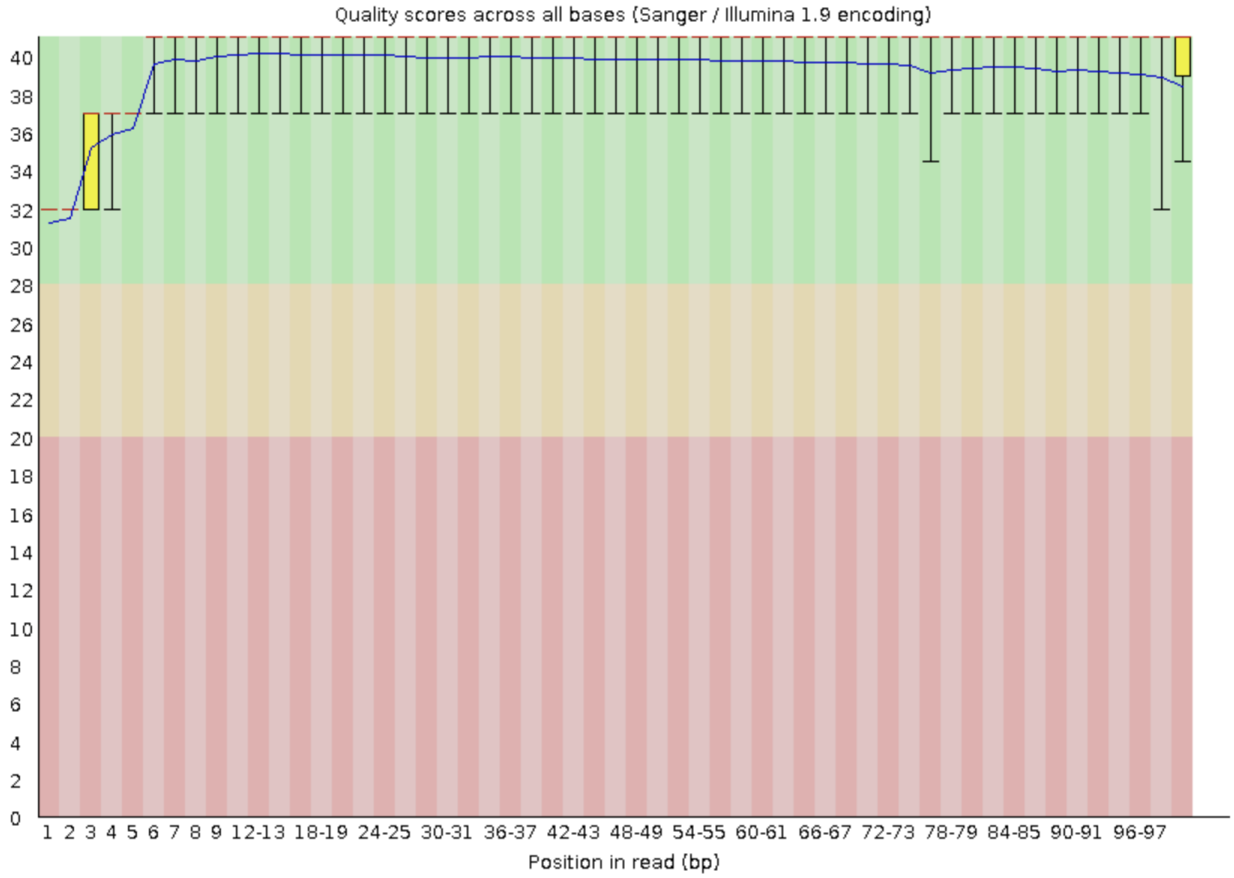

Figure 6: 16_3D_mbnl_S12_L008_R1_001

Figure 7: 16_3D_mbnl_S12_L008_R2_001

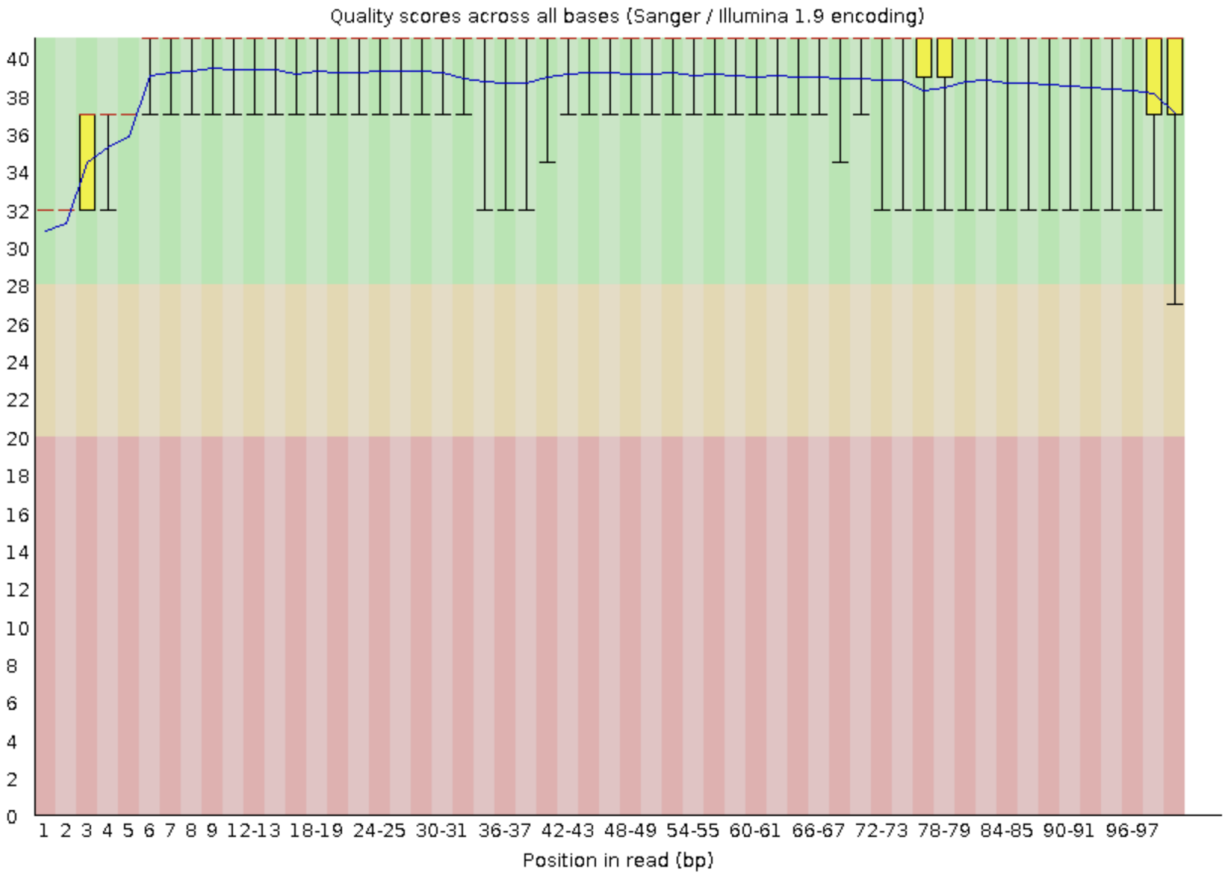

Figure 8: 16_3D_mbnl_S12_L008_R2_001

The Quality score across read position graphs look largely similar between both the in-house plotting script and FastQC. The largest difference of course being the presence of error bars that FastQC calculates, as well as the additional trend line.

The results are largely the same, and the quality accross the data is overall acceptable, staying well over a QC score of 30, and showing a trendline of around 36-38 for most.

Memory usage for both appeared negligable. CPU time as well did not appear to have stark differences between the two methods, FastQC took just slightly longer but created much more data and many more plots than the in house software.

Additionally, all the other statistics that FastQC reports for each file were assessed, and were largely sufficient for each of the files. FastQC only gives warnings to adapter content, which will be filtered out, and per tile sequence quality, which shows only 1 to 2 tiles per file that may contain low quality data.

After assessing the FastQC report, all 4 files are confirmed to be of a high enough quality to continue with the analysis.

Part 2

As mentioned, the adapter sequences showed up in the FastQC report, these adapters are the Illumina Truseq Universal Adapters:

R1: AGATCGGAAGAGCACACGTCTGAACTCCAGTCA

R2: AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT

These sequences were removed from all the data before further analysis, using cutadapt with default parameters.

Before cutadapt was run, % of reads containing an adapter was recorded for each of the 4 files:

```
21_3G_both_S15_L008_R1_001.fastq.gz

(66732 adapter containing reads / 36949196 reads) * 100 = 0.18% adapter content

21_3G_both_S15_L008_R1_002.fastq.gz

(67707 / 36949196) * 100 = 0.18% adapter content

16_3D_mbnl_S12_L008_R2_001.fastq.gz

(115556 / 32940788) * 100 = 0.35% adapter content

16_3D_mbnl_S12_L008_R2_002.fastq.gz

(115921 / 32940788) * 100 = 0.35% adapter content
```

Sequences were cut using cut adapt, list parameters and adapters

Following cutadapt, sequences were trimmed using trimmomatic, with parameters

LEADING:3

TRAILING:3

SLIDINGWINDOW:5:15

MINLEN:35

The read length distributions of these newly tripped reads were plotted using MatPlotLib

From visual inspection, it can be concluded that both of the R1 files in the two RNA-seq datasets are trimmed more aggresively than R2. It is inconclusive as to exactly why, but the R2 files contain much higher frequencies of low read lengths than the R1 files.

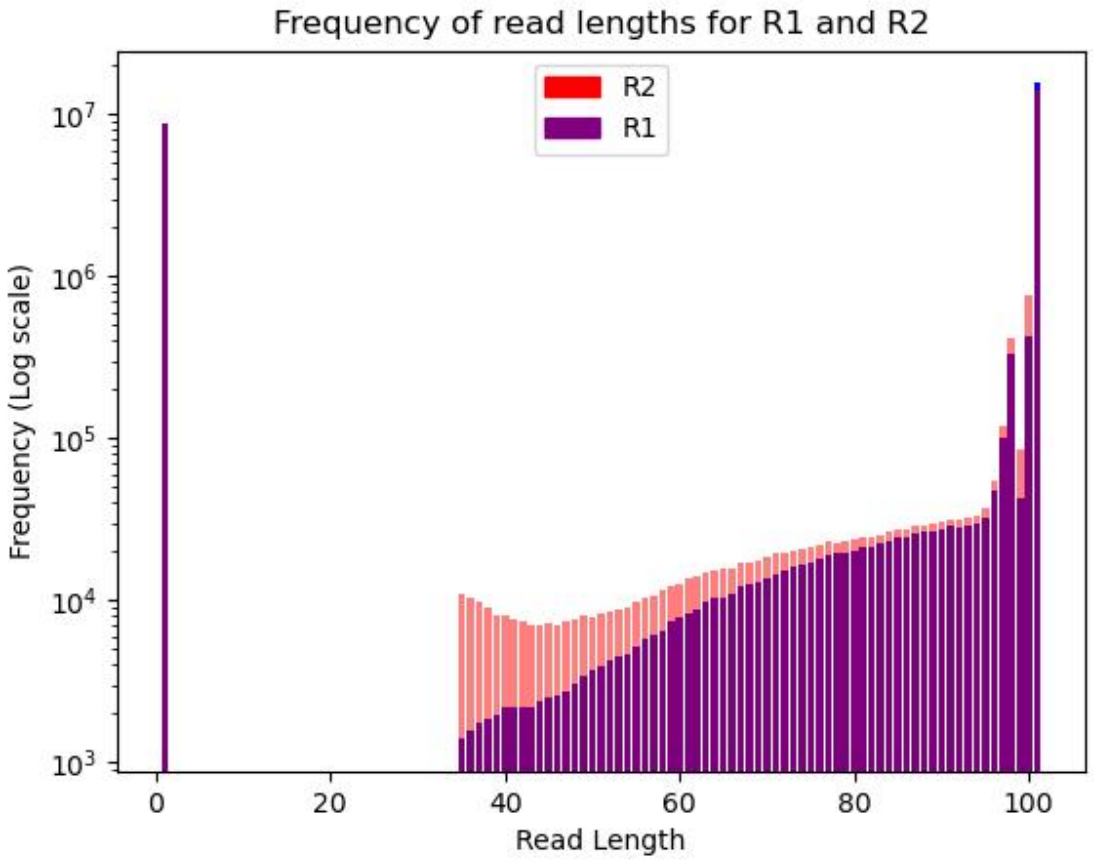

Figure 9: 21_3G_both_S15_L008

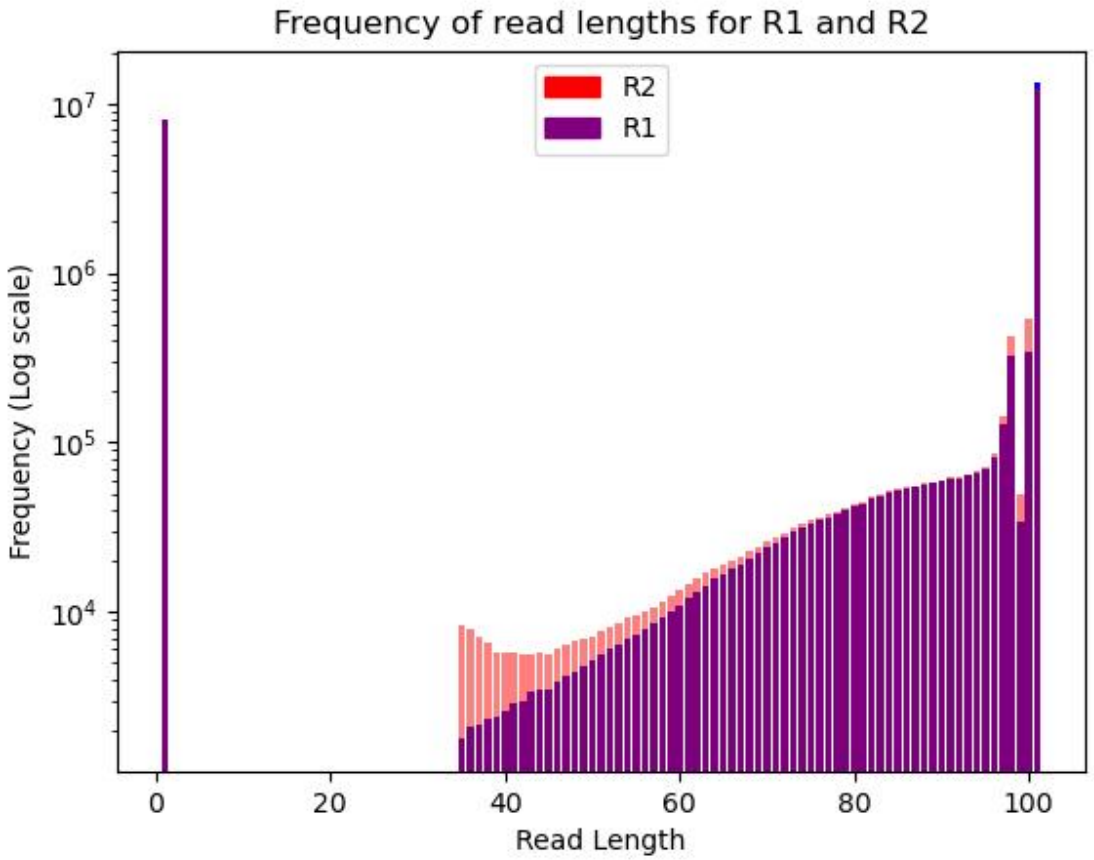

Figure 10: 16_3D_mbnl_S12_L008

Table 1: Mapped and Unmapped Reads from SAM Files

Datatset	Mapped_Reads	Unmapped_Reads
16_3D_mbnl_S12_L008	15662583	365733
21_3G_both_S15_L008	17061180	645444

Part 3

The RNA-seq datasets were then aligned to a publically available mouse genome from ensemble using the splice aware aligner, STAR.

The number of mapped and unmapped reads was recorded for each file and presented in Table 1:

Reads that map to features were computed using HTseq, and are shown here:

```
cat 21_3G_both_S15_L008.strand.txt

__no_feature 7,811,301
__ambiguous 7301
__too_low_aQual 12685
__not_aligned 315910
__alignment_not_unique 383496

cat 21_3G_both_S15_L008.rev.txt

__no_feature 819,790
__ambiguous 141100
__too_low_aQual 12685
__not_aligned 315910
__alignment_not_unique 383496

cat 16_3D_mbnl_S12_L008.strand.txt

__no_feature 7,107,777
__ambiguous 6384
__too_low_aQual 7398
__not_aligned 178950
__alignment_not_unique 388074

cat 16_3D_mbnl_S12_L008.rev.txt

__no_feature 429,966
__ambiguous 136,863
__too_low_aQual 7398
__not_aligned 178950
__alignment_not_unique 388074
```

Based on the fact that the reverse strands show a much lower rate of no-features, it can be hypothesized that this particular RNA-seq protocol favors the reverse strand for the actual purposes of feature identification. Upon research of documentation, this is supported by Illumina documentation on their Truseq platform.

It can therefore be concluded that these RNA reads are in fact strand specific due to this descrepancy, which is supported by the protocol documentation.