RNA-seq Quality Assessment

Leyla Cufurovic

2023-09-14

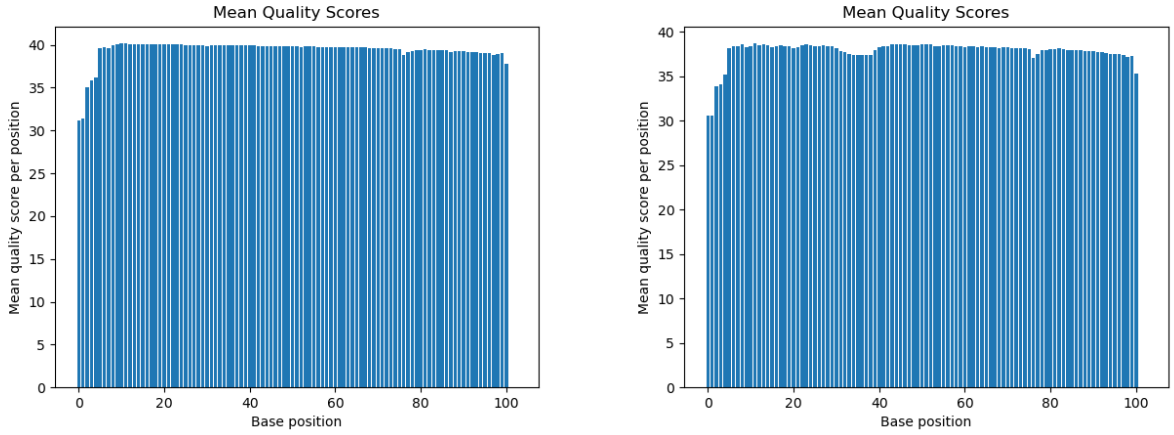
The purpose of this report is to use existing tools for quality assessment and adaptor trimming, compare the quality assessments to those from my own software, and to demonstrate my ability to summarize other important information about this RNA-Seq data set in a high-level report.

The data used for this report can be found here:

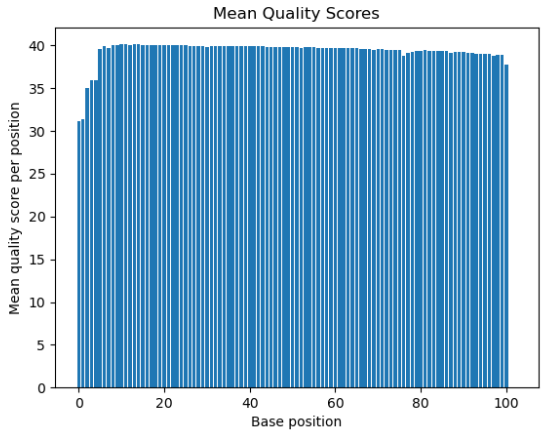
/projects/bgmp/shared/2017_sequencing/demultiplexed/34_4H_both_S24_L008_R1_001.fastq.gz /projects/bgmp/shared/2
/projects/bgmp/shared/2017_sequencing/demultiplexed/6_2D_mbnl_S5_L008_R1_001.fastq.gz /projects/bgmp/shared/20

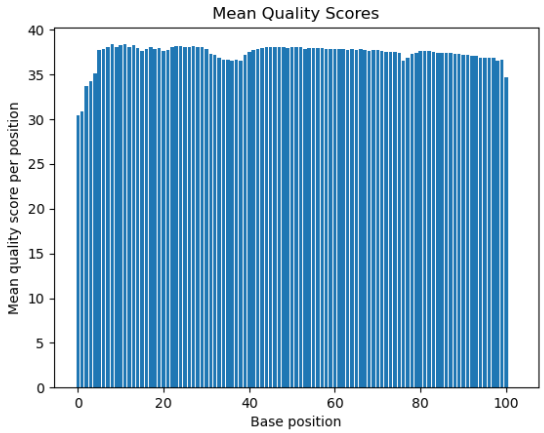
Part 1: Read quality score distributions

1.1 Per base quality score: python plots



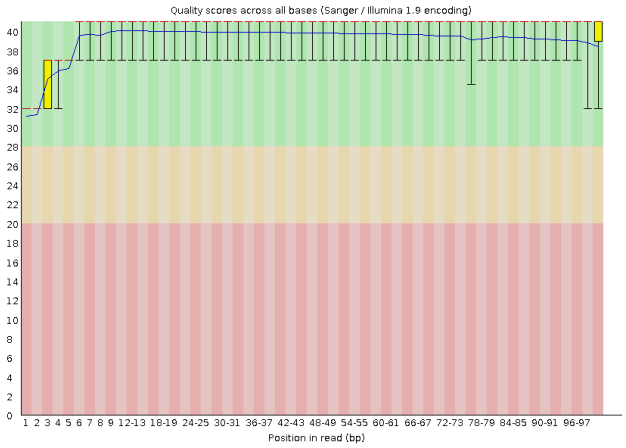
34_4H_both_S24 per base mean quality score python plots. Distribution of average quality scores per sequence plots generated with python script.

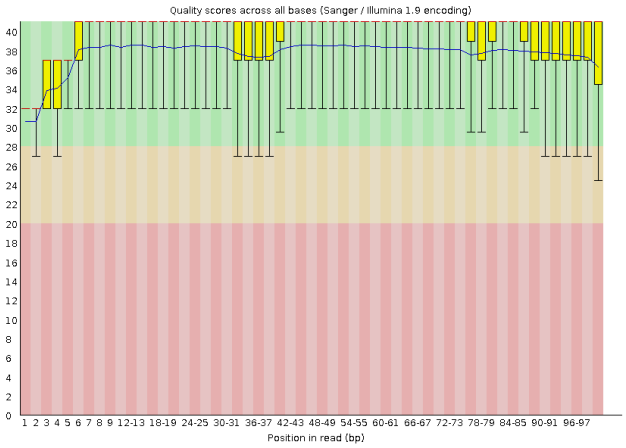




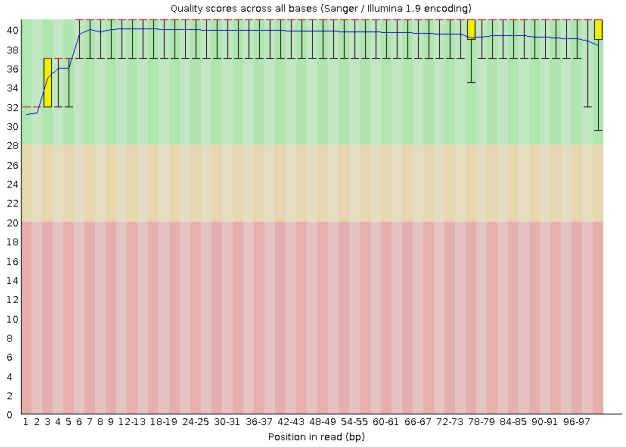
6_2D_mbnl_S5 per base mean quality score python plots. Distribution of average quality scores per sequence plots generated with python. Left: R1; Right: R2.

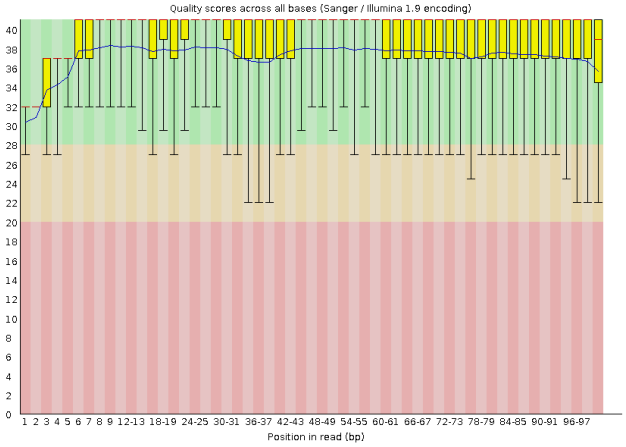
1.1 Per base quality score: FastQC plots





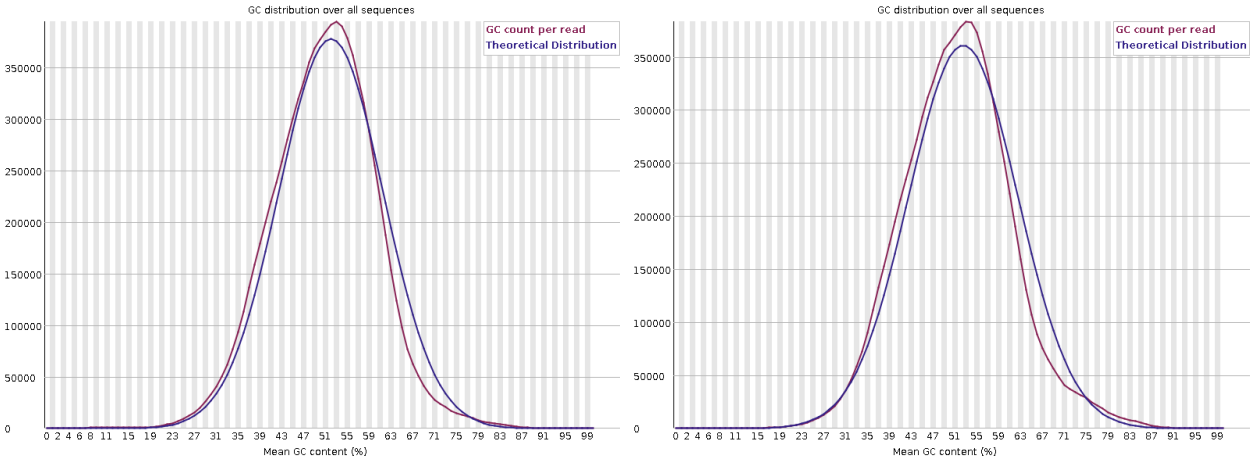
34_4H_both_S24 mean per base quality score FastQC plots. Distribution of average quality scores per sequence plots generated with FastQC. Left: R1; Right: R2.



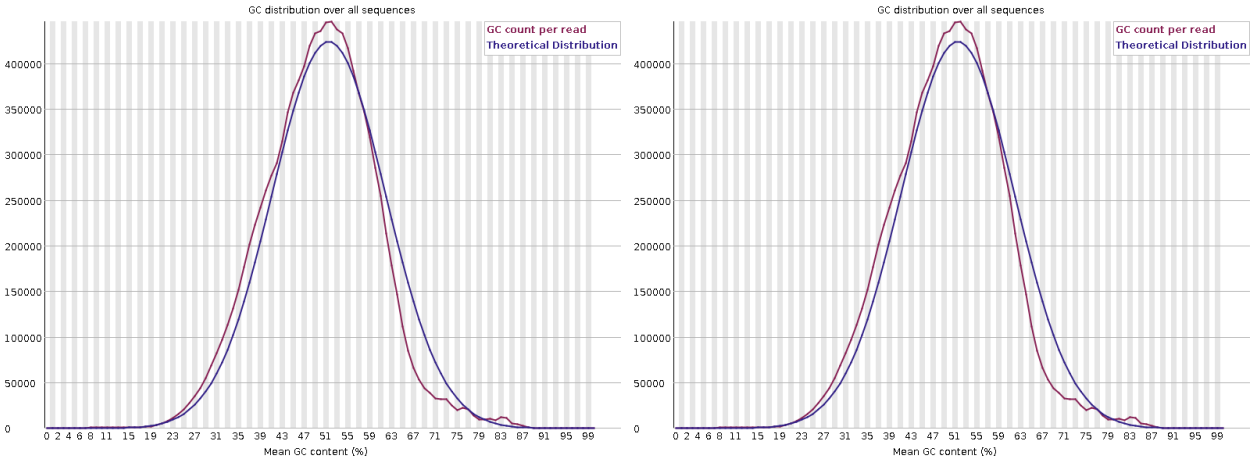


6_2D_mbnl_S5 per base mean quality score FastQC plots. Left: R1; Right: R2.

1.2 GC Content

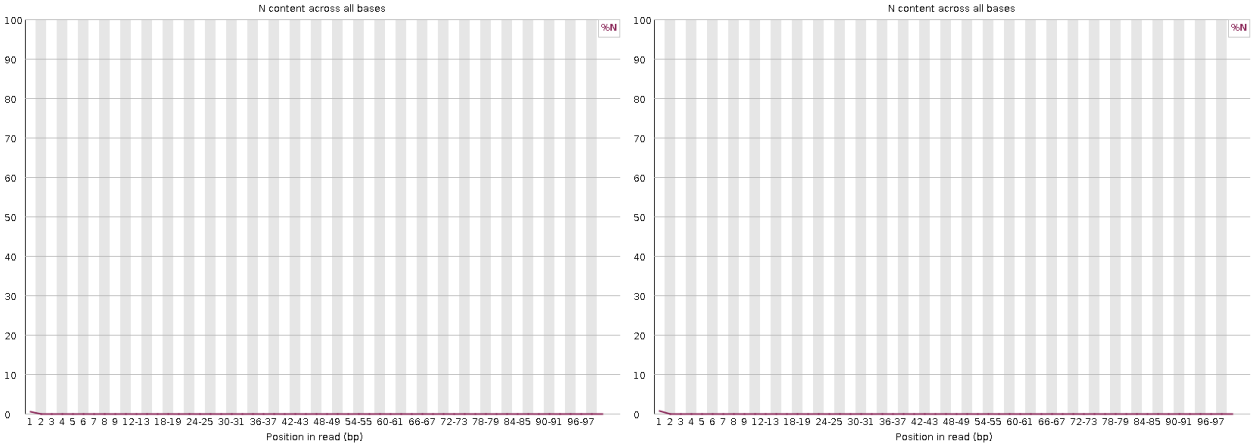


34_4H_both_S24 per sequence gc content for R1 (left) and R2 (right)

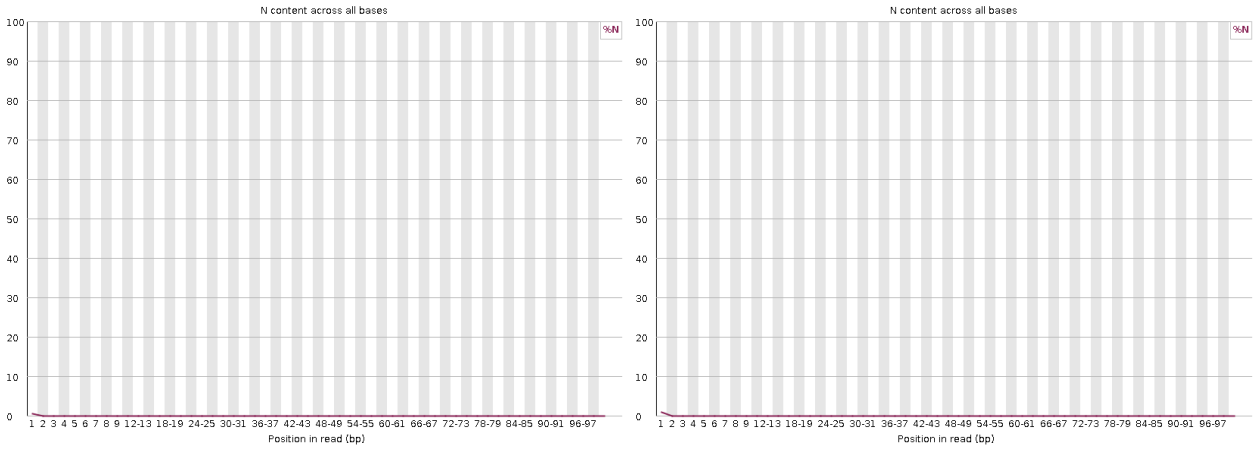


6_2D_mbnl_S5 per sequence gc content for R1 (left) and R2 (right)

1.3 Per base N content



34_4H_both_S24 per base N content for R1 (left) and R2 (right).



6_2D_mbnl_S per base N content for R1 (left) and R2 (right)

1.4 FastQC vs Python script

Library	FastQC times (m:s.ms)	Python times (m:s.ms)
R1 both	0:53.49	11:33.67
R2 both	0:56.25	11:34.58
R1 mbnl	1:00.63	12:49.02
R2 mbnl	1:00.72	11:52.32

FastQC was able to run significantly faster than my python script. One reason might be because FastQC is optimized to be faster.

Discussion of plots:

The per base sequence quality plots indicate that the data all have relatively high quality scores. While there is a lower quality score in the first several bases (although the lowest qscore is about 31, which is still a high score), the majority of the data has a quality score of about 40. It's important to note that while lower quality scores in the initial base positions are common, they don't necessarily indicate low data quality overall. The plots from the FastQC outputs as well as the plots generated from my python script followed the same trend lines for each plot, indicating that the quality scores are accurate.

Additionally, the Fastqc output GC content plots that show that the GC content of the libraries is around 50%. Eukaryotic genomes tend to have a more moderate GC content, often falling in the range of 30% to 65%, but a range of 50 to 55 percent is ideal. An unusually shaped distribution could indicate a contaminated library or some other kinds of biased subset. A normal distribution which is shifted indicates some systematic bias which is independent of base position. If there is a systematic bias which creates a shifted normal distribution then this won't be flagged as an error by the module since it doesn't know what your genome's GC content should be. In this case, the plots generally follow a normal distribution.

The per base N content of the libraries is generally around 0, except for a small increase around the first few bases. This is to be expected because the ends of the extracted RNA are prone to degradation. Overall, the N content stays at 0 and indicates that every base position had a base call.

Overall, I would suggest that the data is of high enough quality to use for further downstream analysis.

Part 2: Adaptor trimming comparison

2.1 Adapter and quality trimming

Adapter sequences:

```
R1: AGATCGGAAGAGCACACGTCTGAACTCCAGTCA
R2: AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
```

Command to search for adapter sequence in dataset:

```
R1 34_4H_both_S24 count: 138880
zcat /projects/bgmp/shared/2017_sequencing/demultiplexed/34_4H_both_S24_L008_R1_001.fastq.gz |
grep "AGATCGGAAGAGCACACGTCTGAACTCCAGTCA" | wc -l


R2 34_4H_both_S24 count: 137879
zcat /projects/bgmp/shared/2017_sequencing/demultiplexed/34_4H_both_S24_L008_R2_001.fastq.gz |
grep "AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT" | wc -l


R1 6_2D_mbnl_S5 count: 12870
zcat /projects/bgmp/shared/2017_sequencing/demultiplexed/6_2D_mbnl_S5_L008_R1_001.fastq.gz |
grep "AGATCGGAAGAGCACACGTCTGAACTCCAGTCA" | wc -l


R2 6_2D_mbnl_S5 count: 13562
zcat /projects/bgmp/shared/2017_sequencing/demultiplexed/6_2D_mbnl_S5_L008_R2_001.fastq.gz |
grep "AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT" | wc -l
```

Cutadapt output

Library	Read-pairs	Read 1 Trimmed	% Read 1 Trimmed	# Read 2 Trimmed	% Read 2 Trimmed
34_4H_both_S24	9040597	819166	9.1%	886595	9.8%
6_2D_mbnl_S5	11028244	416045	3.8%	502045	4.6%

Trimmomatic output

Library	Total Surviving	% Total surviving	Total dropped	% Total dropped
34_4H_both_S24	8671861	95.92%	3133	0.03%
6_2D_mbnl_S5	10463870	94.88%	4084	0.04%

Trimmomatic version: 0.39

cutadapt version: 4.4

In this section, Cutadapt was performed to remove adapter sequences and Trimmomatic was used to quality trim the data. Cutadapt trimmed 9.1% and 9.8% for read 1 and read 2, respectively, for the 34_4H_both_S24 library; 3.8% and 4.6% of reads were trimmed for read 1 and read 2, respectively, for the 6_2D_mbnl_S5 library. The trimmomatic plots show that the reverse reads have consistently lower read lengths than read 1. This can be attributed to sample degradation due to the wait time on the sequencer for read 2 to be processed.

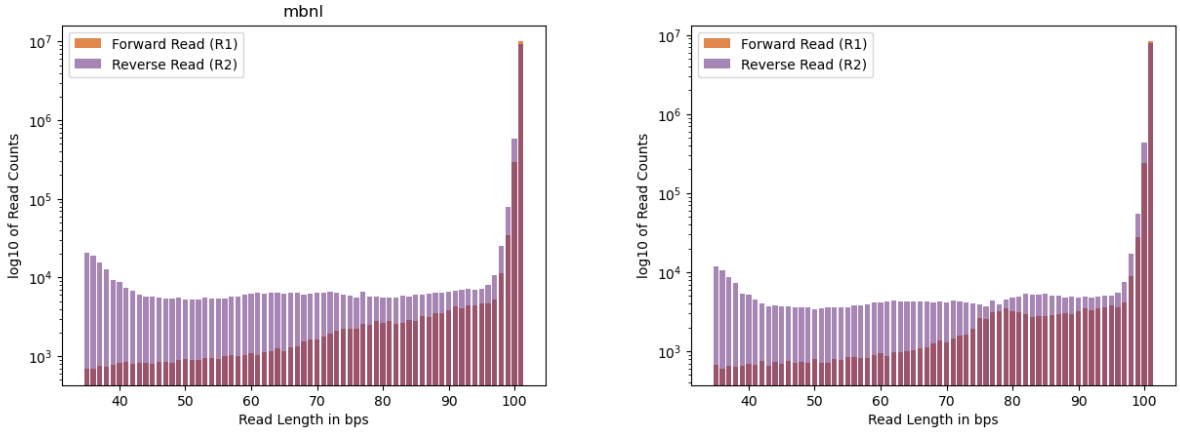


Figure 1: Trimmed read length distributions for both R1 (left) and R2 reads (right)

Part 3 Alignment and strand-specificity

```
STAR version 2.7.10b
htSeq version 2.0.3
matplotlib version 3.7.2
numpy version 1.25.2

Fasta file used for STAR:     Mus_musculus.GRCm39.dna.primary_assembly.fa (Ensemble release 110)
  GTF file used for STAR:     Mus_musculus.GRCm39.110.gtf (Ensemble release 110)
```

3.1 STAR results: mapped vs unmapped reads

Sample	Mapped Reads	Unmapped Reads
34_4H_both_S24	16213347	20049202
6_2D_mbnl_S5	1130375	878538

3.2 htSeq results: stranded vs unstranded

Sbatch script for running htSeq

conda activate bgmp-QAA /usr/bin/time -v htseq-count –stranded=reverse /projects/bgmp/leylacuf/bioinfo/Bi623/QAA/mus_musc_files/STAR_output/both_trimmed_Aligned.out.sam /projects/bgmp/leylacuf/bioinfo/Bi623/QAA/mus_musc_files/Mus_musculus.GRCm39.110.gtf > htseq_both_reverse.txt

conda activate bgmp-QAA /usr/bin/time -v htseq-count –stranded=yes /projects/bgmp/leylacuf/bioinfo/Bi623/QAA/mus_musc_files/STAR_output/both_trimmed_Aligned.out.sam /projects/bgmp/leylacuf/bioinfo/Bi623/QAA/mus_musc_files/Mus_musculus.GRCm39.110.gtf > htseq_both_reverse.txt

Commands for parsing htseq outputs

```
cat htseq_both_reverse.txt | grep -v "^__" | awk '{sum+=$2} END {print sum}'
cat htseq_both_stranded.txt | grep -v "^__" | awk '{sum+=$2} END {print sum}'

cat htseq_mbnl_reverse.txt | grep -v "^__" | awk '{sum+=$2} END {print sum}'
cat htseq_mbnl_stranded.txt | grep -v "^__" | awk '{sum+=$2} END {print sum}'
```

Library	Total Reads	Stranded Reads	% Stranded Reads
34_4H_both_S24	8671861	432793	4.99%
6_2D_mbnl_S5	10463870	382583	3.66%

Library	Total Reads	Reverse Reads	% Reverse Reads
34_4H_both_S24	8671861	6986953	80.57%
6_2D_mbnl_S5	10463870	8553965	81.74%

Strand-Specific Reads: A significant proportion of reads in both libraries correspond to strand-specific alignments. For the "34_4H_both_S24" library, approximately 4.99% of the reads are stranded, while for the "6_2D_mbnl_S5" library, about 3.66% of the reads are stranded.

Reverse Reads: The majority of reads in both libraries align in the reverse orientation. Approximately 80.57% of the reads in the "34_4H_both_S24" library and 81.74% of the reads in the "6_2D_mbnl_S5" library align in the reverse direction.

These quantitative statements suggest that a substantial portion of the reads in the RNA-Seq data corresponds to strand-specific alignments, indicating that the data are derived from strand-specific RNA-Seq libraries. The high percentage of reverse reads further supports that the majority of alignments are on the reverse strand.