QAA_report

Amelia Dayton

2024-09-04

Read Quality Score Distributions

FASTQC per-base quality score distributions:

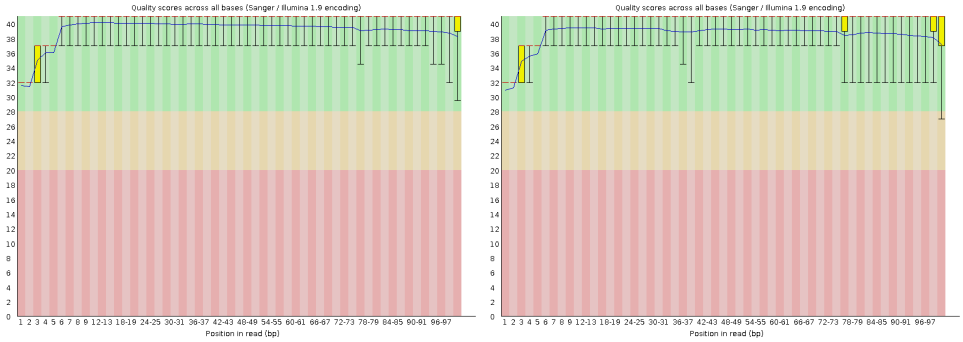

Figure 1: Per base quality score for RNA-seq samples '2 2B control S2' generated with fastqc. (Left) Average quality score per base with standard deviations for R1 of '2 2B control S2'. (Right) Average quality score per base with standard deviations for R2 of '2 2B control S2'.

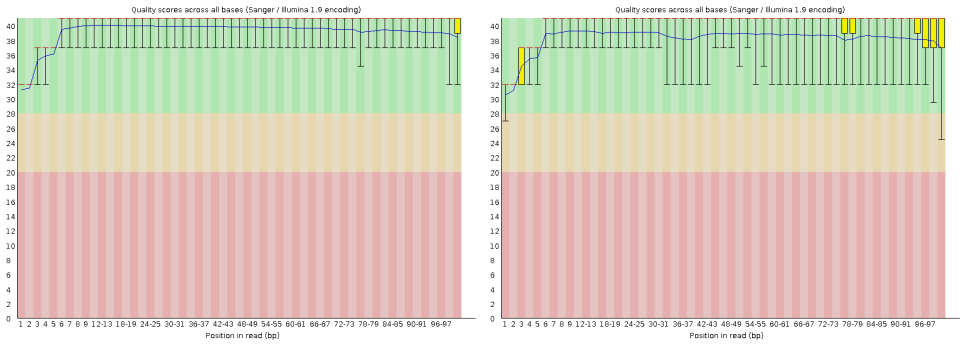

Figure 2: Per base quality score for RNA-seq samples '19 3F fox S14' generated with fastqc. (Left) Average quality score per base with standard deviations for R1 of '19 3F fox S14'. (Right) Average quality score per base with standard deviations for R2 of '19 3F fox S14'.

FASTQC per-base N content:

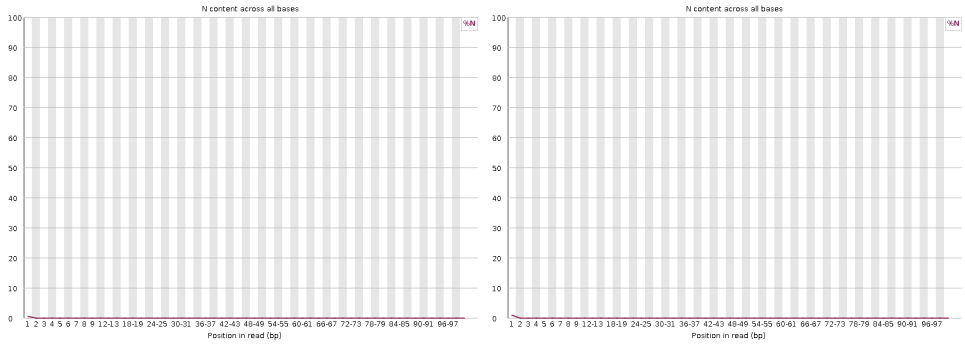

Figure 3: Per base N content for RNA-seq samples '2 2B control S2' generated with fastqc. (Left) Counts of N per base for R1 of '2 2B control S2'. (Right) Counts of N per base for R2 of '2 2B control S2'.

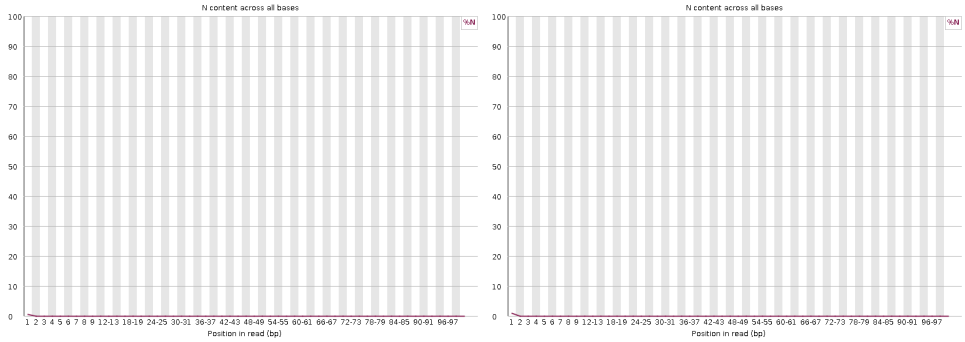

Figure 4: Per base N content for RNA-seq samples '19 3F fox S14' generated with fastqc. (Left) Counts of N per base for R1 of '19 3F fox S14'. (Right) Counts of N per base for R2 of '19 3F fox S14'.

For both conditions the per base N content is consistent with the per base quality score. Both conditions show more Ns in the beginning of the read correlating to a slight dip in quality in the beginning of the read.

Amelia-Written Per-Base Histogram Plots

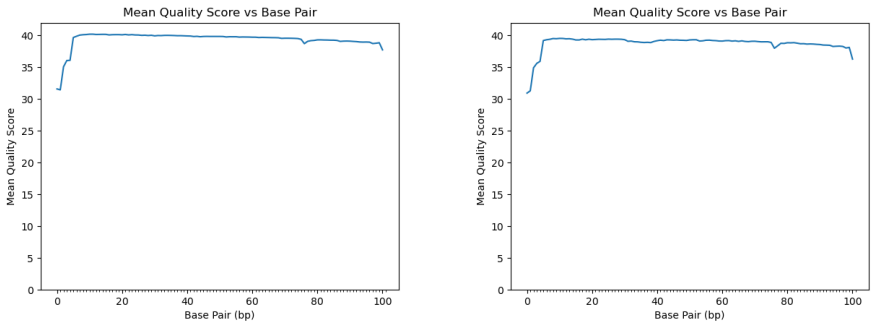

Figure 5: Per base quality Score for RNA-seq samples '2 2B control S2' generated with personal python script. (Left) Average quality score per base with no error bars for R1 of '2 2B control S2'. (Right) Average quality score per base with no error bars for R2 of '2 2B control S2'.

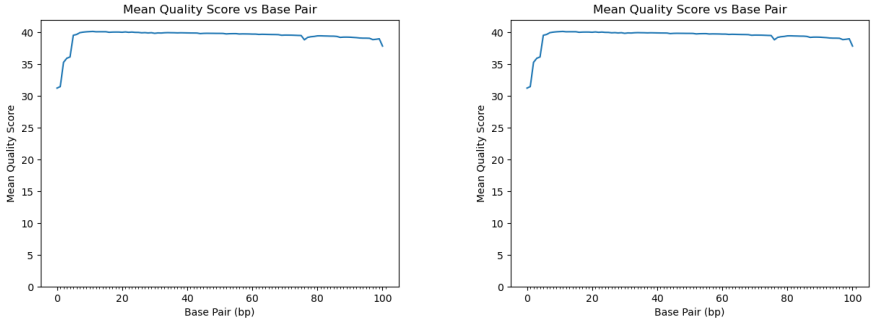

Figure 6: Per base quality Score for RNA-seq samples '19 3F fox S14' generated with personal python script. (Left) Average quality score per base with no error bars for R1 of '19 3F fox S14'. (Right) Average quality score per base with no error bars for R2 of '19 3F fox S14'.

The overall line of the plots (the means of the scores) is exactly the same, but my plots lack the error bars and color coding. The fastqc plots took 3 minutes and 45 seconds and my histogram script took 10 minutes and 2 seconds. The fastqc plots took less memory (~20 kbytes less) and produced a lot more data! Fastqc is likely much more optimized and written in a language closer to the hardware such as java so it outperforms my script. Essentially, the fastqc plot is faster and prettier, but both methods produce the same data.

The per-base qscore is more than sufficient for these data, and the per-sequence-quality scores are also very high quality. The sequence length distribution is all an even 101 bp. Overall, these data are of very high quality and I feel comfortable moving forward with them.

Adapter Trimming Comparisons

The adapter sequences are:

R1: AGATCGGAAGAGCACACGTCTGAACTCCAGTCA

R2: AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT

By looking through all the reads for the presence of adapter sequences, the R1 adapter is present in abundance in both condition's R1 and not present at all in R2. The R2 adapter is present in both condition's R2 and not present at all in R1. This makes sense!

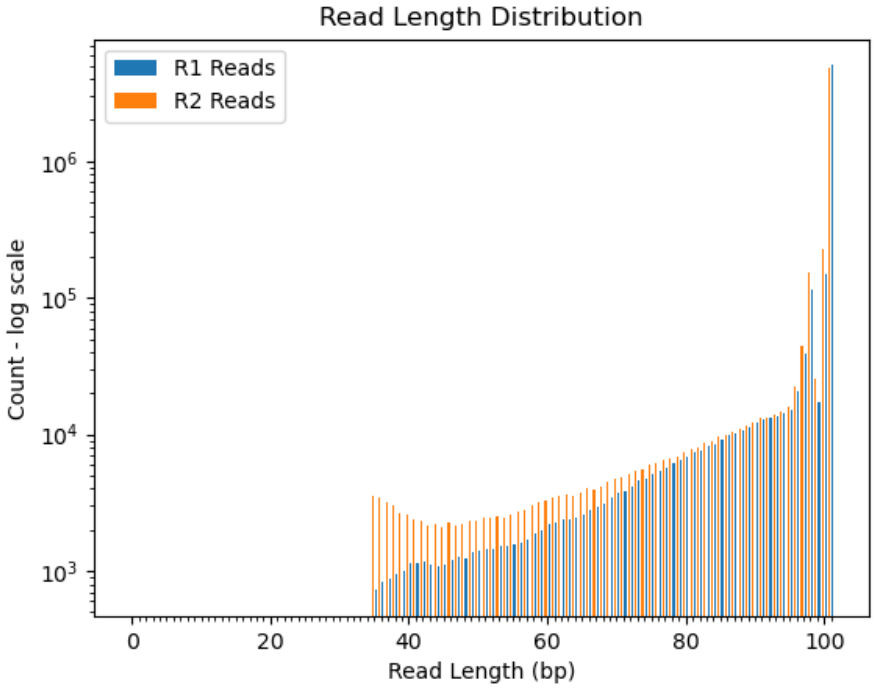

Figure 7: Read length distribution for R1 and R2 of RNA-seq sample '2 2B control S2' post-adapter trimming.

For the '2 2B control S2' sample, 7.3% of reads (423,128) were trimmed from R1 and 8.1% of reads (473,368) were trimmed from R2. For the '19 3F fox S14' sample, 3.3% of reads (546,623) were trimmed from R1 and 4.1% of reads (676,564) were trimmed from R2. In both conditions, all reads passed filters and were written.

Adapter was trimmed at a similar rate in both sets with R2 being trimmed slightly more extensively. This makes sense as the reads should be the same length, ????

These plots agree that R1 is trimmed more extensively than R2 as R1 reads are shorter than R2.

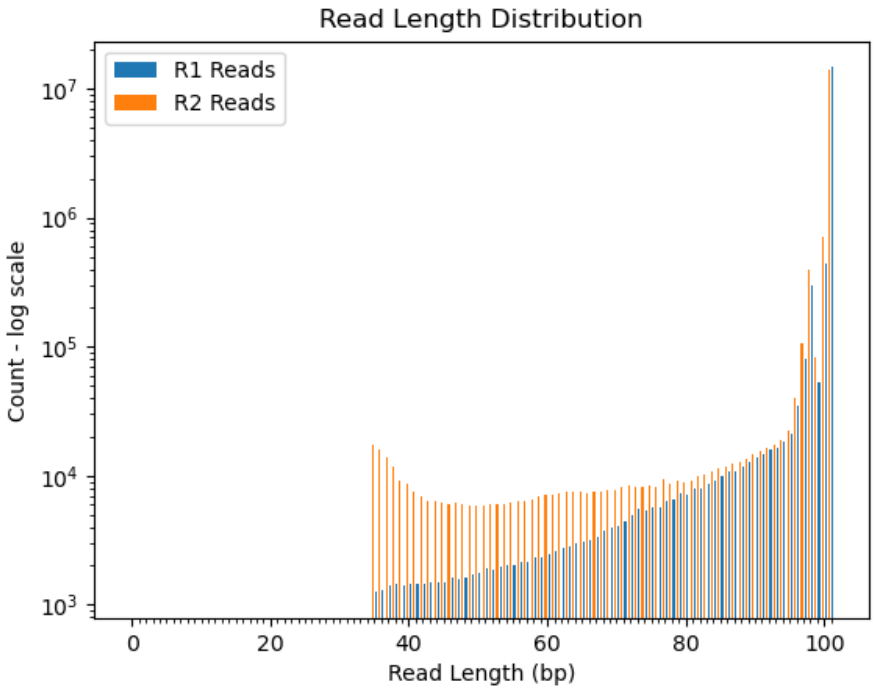

Figure 8: Read Length Distribution for R1 and R2 of RNA-seq samples '19 3F fox S14' post-adapter trimming.

Alignment and Strand-Specificity

Table 1: Table 1. Aligned '2 2B control S2' reads that map or do not map to reference mouse genome. 'Mapped to Forward' and 'Mapped to Reverse' columns found with htseq-count package and –stranded parameter.

Mapped	Unmapped	Mapped.to.Forward	Mapped.to.Reverse	TotalReads
200241	11388657	4598	54800	11588898

Table 2: Table 2. Percentage of aligned '2 2B control S2' reads that map to forward or reverse strand.

Percentage.Mapped.to.Forward	Percentage.Mapped.to.Reverse
0.0793518	0.945733

Table 3: Table 3. Aligned '19 3F fox S14' reads that map or do not map to reference mouse genome. 'Mapped to Forward' and 'Mapped to Reverse' columns found with htseq-count package and –stranded paramater.

Mapped	Unmapped	Mapped.to.Forward	Mapped.to.Reverse	TotalReads
732750	31963176	11853	282927	32695926

Table 4: Table 2. Percentage of aligned '2 2B control S2' reads that map to forward or reverse strand.

Percentage.Mapped.to.Forward	Percentage.Mapped.to.Reverse
0.0725044	1.73066

I propose that these data are strand-specific to the reverse strand, because 0.8674812% in 2_2B_control_S2 and 1.658156% in 19_3F_fox_S14 more reads mapped to the reverse strand. Reads preferentially mapping to any strand indicates that this is a strand-specific library. If the library was not strand specific, there would be an approximately equal distribution of reads between the two strands.