QAA

Isabel Quesada

2023-09-08

Contents

Part 1 – Read quality score distributions 1
Part 2 – Adaptor trimming comparison 4
Part 3 – Alignment and strand-specificity 5

Part 1 – Read quality score distributions

- Library Assignment:
  - 6_2D_mbnl_S5_L008
  - 16_3D_mbnl_S12_L008

FastQC

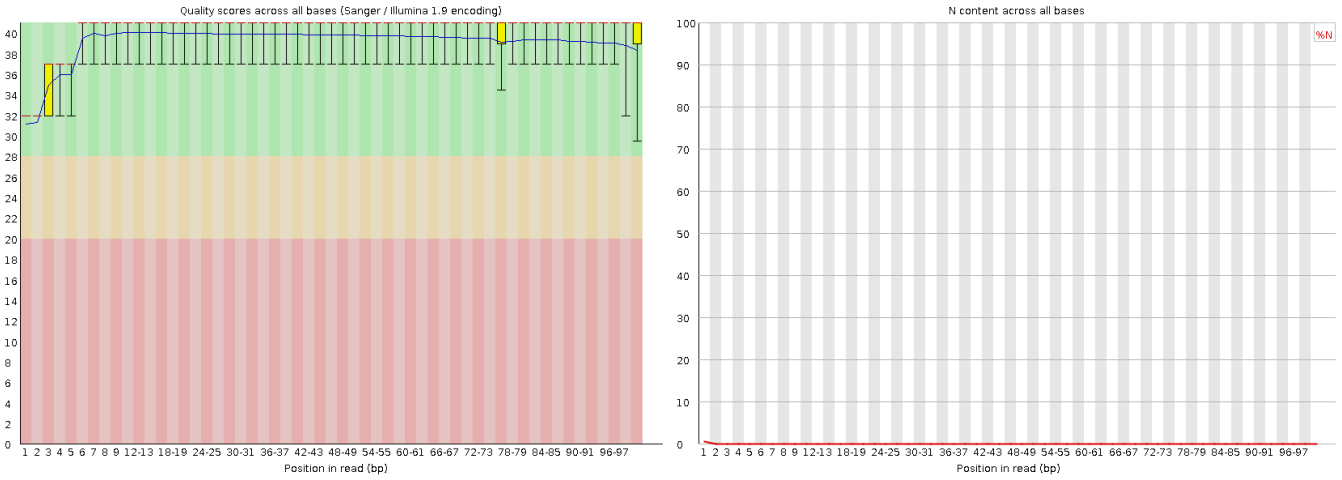

Figure 1: FastQC Generated Plots: 6-2D-mbnl-S5-L008-R1

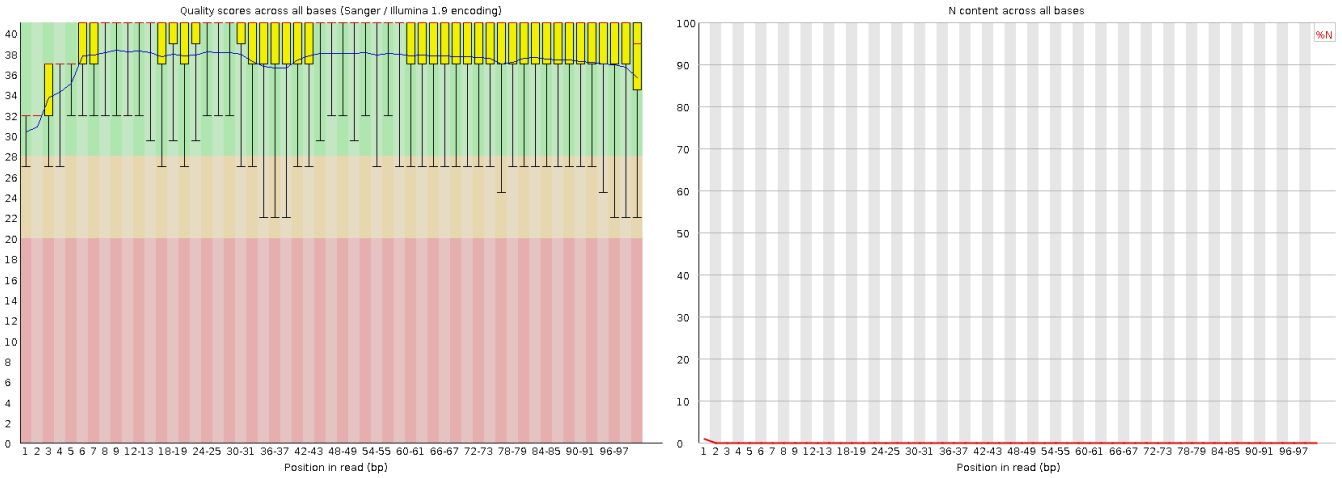

Figure 2: FastQC Generated Plots: 6-2D-mbnl-S5-L008-R2

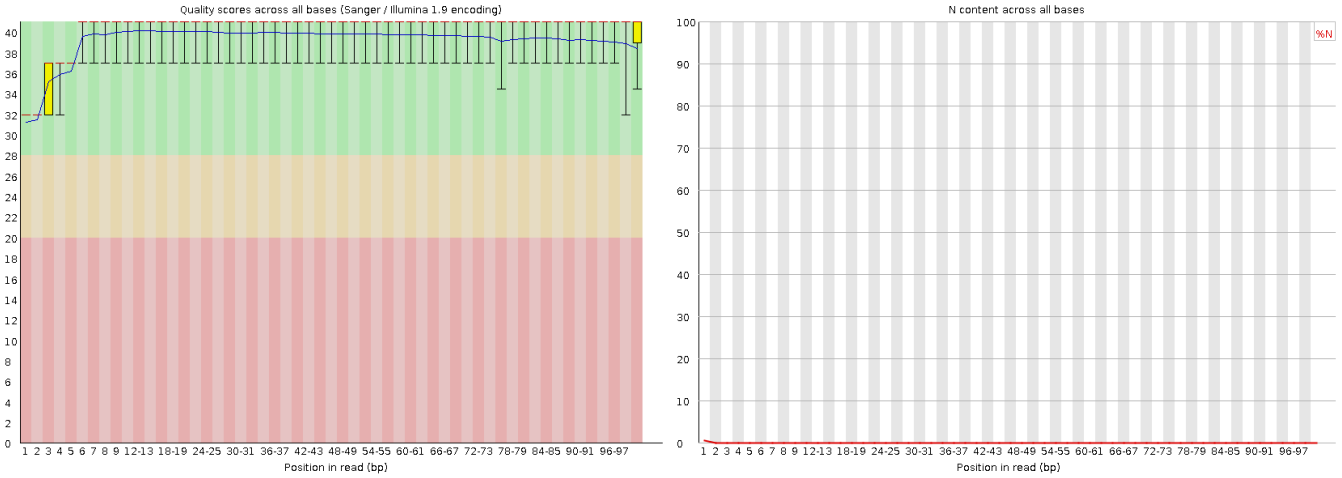

Figure 3: FastQC Generated Plots: 16-3D-mbnl-S12-L008-R1

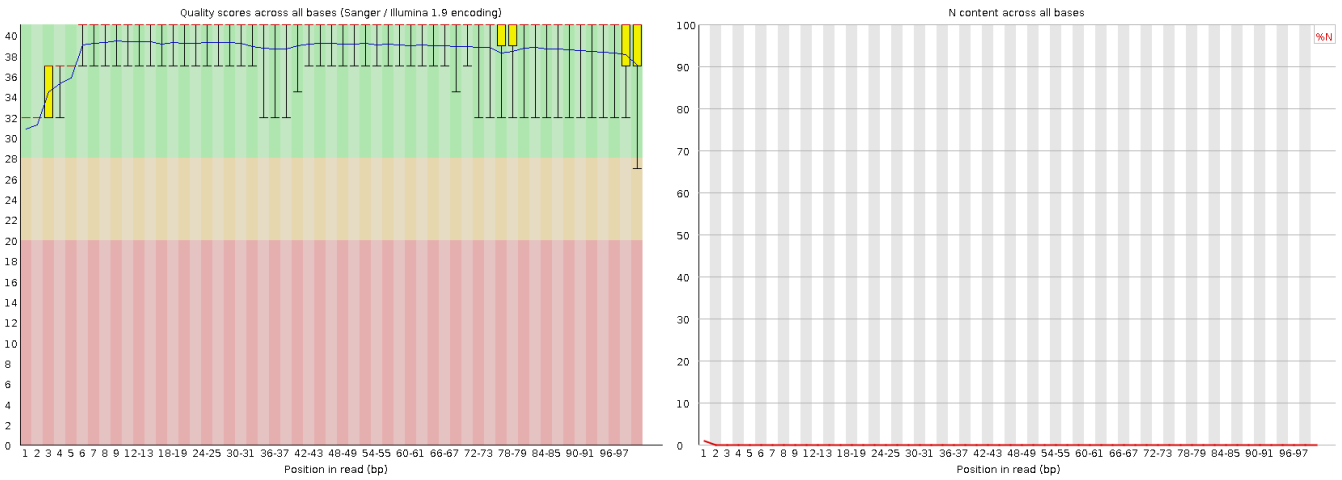

Figure 4: FastQC Generated Plots: 16-3D-mbnl-S12-L008-R2

The per-base N content plots are consistent with the quality score plots across all reads. All the plots displayed a small increase in N content at the start of the read which corresponded to a decrease in quality score at that base position.

Demultiplex

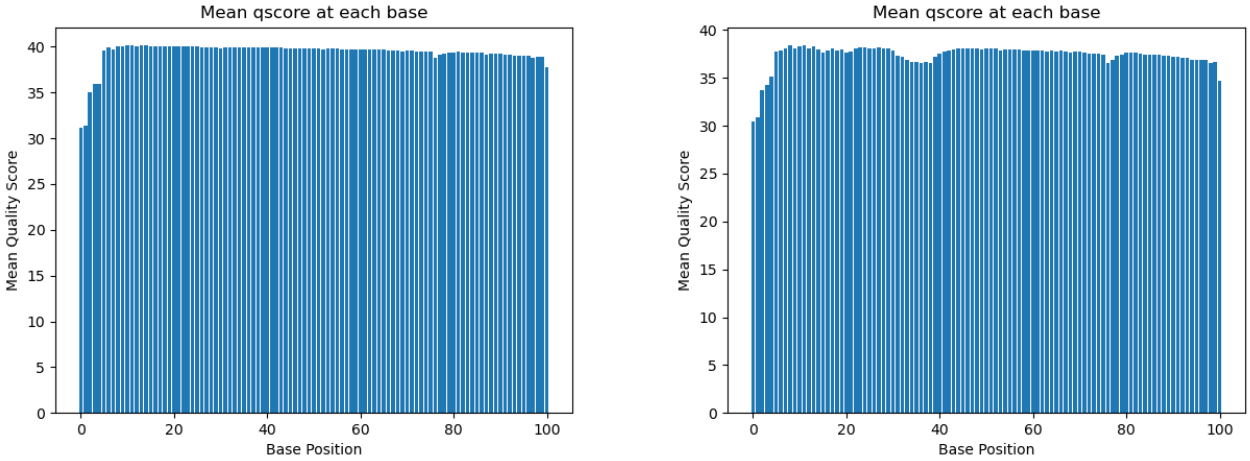

Figure 5: Demultiplex Generated Plots: 6-2D-mbnl-S5-L008-R1(left), 6-2D-mbnl-S5-L008-R2(right)

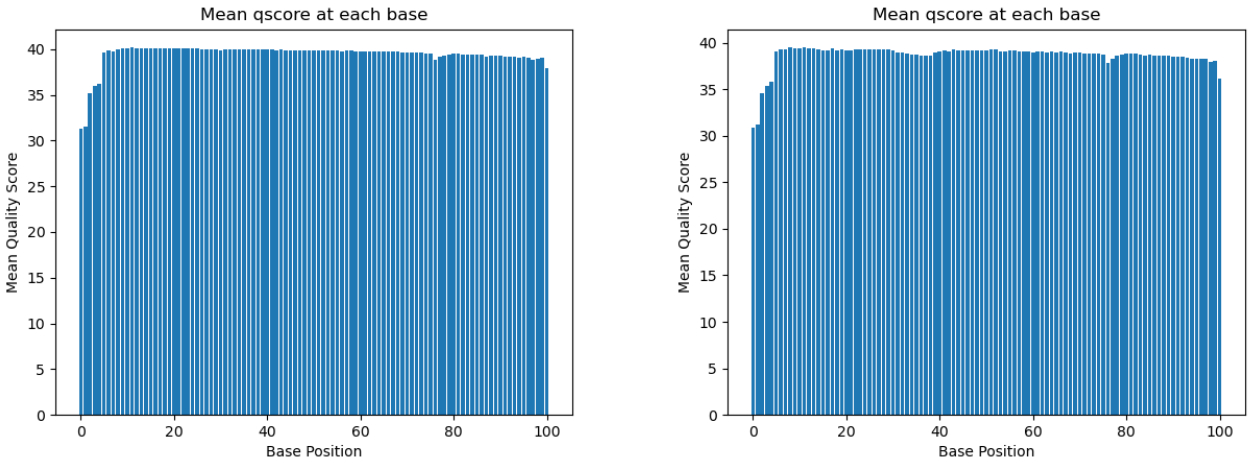

Figure 6: Demultiplex Generated Plots: 16-3D-mbnl-S12-L008-R1(left), 16-3D-mbnl-S12-L008-R2(right)

The FastQC quality score distribution plots and the quality score distribuiton plots I generated, using a script I made, showed the same distribution across bases. Regardless of the method used the distributions were the same when compared to each other. There was a significant difference in run time, the script I made took 3.5 times longer to run than FastQC which could have been due to FastQC having optimized code.

Both of the libraries analyzed have good quality data. After initial data exploration using UNIX I was able to confirm that all the four fastq files were formatted properly and had the proper read length of 101 bp. Using the plots generated by FastQC I was able to check the GC content of the reads, mean quality score per read, mean quality score per base pair, and the N content in each read. All of these checks passed and no abnormal results were present in the data. Since all four data files passed the preliminary screening they are high enough quality to use for further analysis.

Part 2 – Adaptor trimming comparison

Adapter Sequences Used:

-R1:AGATCGGAAGAGCACACGTCTGAACTCCAGTCA

-R2:AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT

These adapter sequences were verified by using UNIX commands on each of the four fastq files. I isolated the sequence line and used grep to find the respective adapter sequence for R1 or R2. The adapter sequences were found on the 3' end which helped with confirming that the adapters had the proper orientation.

UNIX commands used for adapter confirmation:

- zcat 6_2D_mbnl_S5_L008_R1_001.fastq.gz | grep -A 1 ^"@" | grep -v ^"@" | grep -v ^"–" | grep -E –color="always" "AGATCGGAAGAGCACACGTCTGAACTCCAGTCA" | head
- zcat 6_2D_mbnl_S5_L008_R2_001.fastq.gz | grep -A 1 ^"@" | grep -v ^"@" | grep -v ^"–" | grep -E –color="always" "AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT" | head
- zcat 16_3D_mbnl_S12_L008_R1_001.fastq.gz | grep -A 1 ^"@" | grep -v ^"@" | grep -v ^"–" | grep -E –color="always" "AGATCGGAAGAGCACACGTCTGAACTCCAGTCA" | head
- zcat 16_3D_mbnl_S12_L008_R2_001.fastq.gz | grep -A 1 ^"@" | grep -v ^"@" | grep -v ^"–" | grep -E –color="always" "AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT" | head

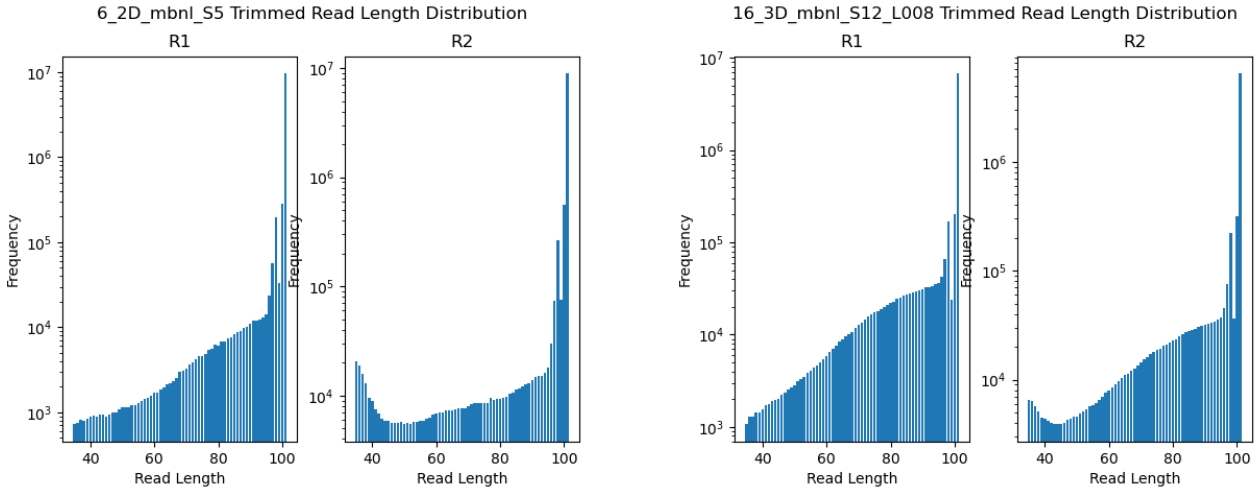

Figure 7: Trimmed Read Lengths Distributions: 6-2D-mbnl-S5(left), 16-3D-mbnl-S12(right)

I would expect R1 and R2 to be trimmed at different rates which is what is observed in figure 7. Overall, R2 tends to have lower quality sequences due to the sequencing process. R2 is the last thing to be sequenced on an Illumina instrument which means the sample has already been sitting on the instrument for a while before it gets sequenced. So when we trim the reads, using Trimmomatic, more data will be removed from R2 than R1.

Part 3 – Alignment and strand-specificity

mapped/unmapped read counts

```
##               library   mapped unmapped
## 1 16_3D_mbnl_S12_L008 15662605   365711
## 2   6_2D_mbnl_S5_L008 20186295   736311
```

I propose that these data are not strand-specific, because 50% of the reads are mapped to each strand (fw and rv).

I used the following UNIX command on each output sam file from running htseq.sh:

- cat 6_2D_mbnl_S5_L008_fw_genecount.txt | awk '{sum+=$2} END{print sum}'

This provided the number of reads mapped to each strand (fw and rv) so I can determine if the data are/are not strand-specific.