QAA_report

Zach Girard

2024-09-08

Part 1

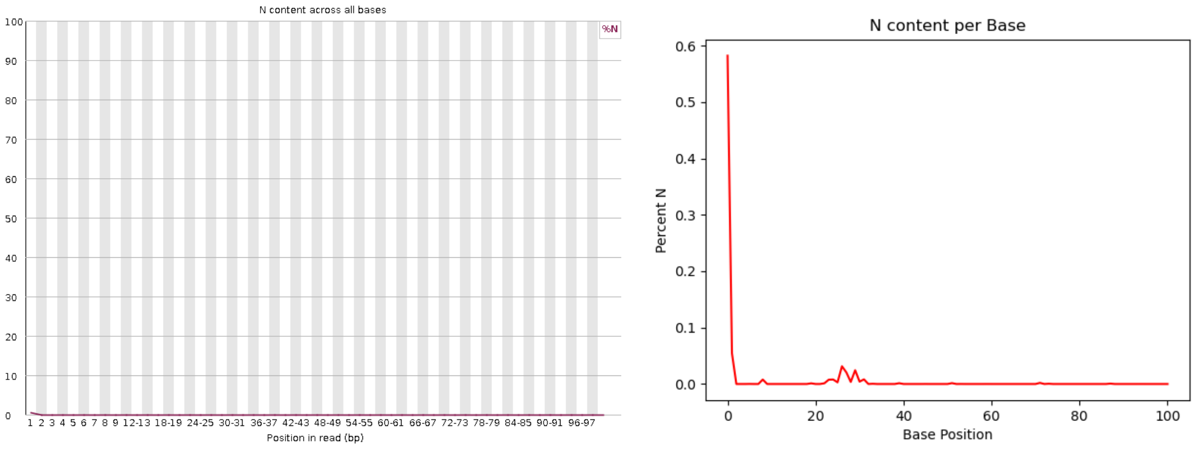
Per base N-content

Step 2

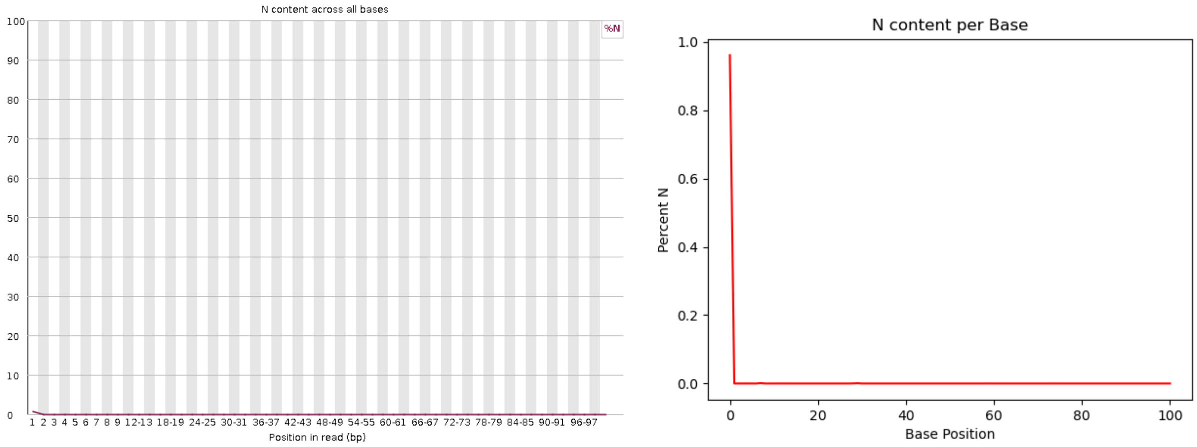
Using FastQC via the command line on Talapas, produce plots of the per-base quality score distributions for R1 and R2 reads. Also, produce plots of the per-base N content, and comment on whether or not they are consistent with the quality score plots.

Answer: My personally produced Per Base N-Content plots show the same trend as those produced by fastqc. However, my plots have their y-axis scaled by the max value. The fastqc plots are out of 100%. This allows me to see much smaller trends that are not visible in the fastqc plots. It appears as though R2 reads have extremely small non-zero counts of N's after the first base, making them visibly indistinguishable from zero even at a small scale.

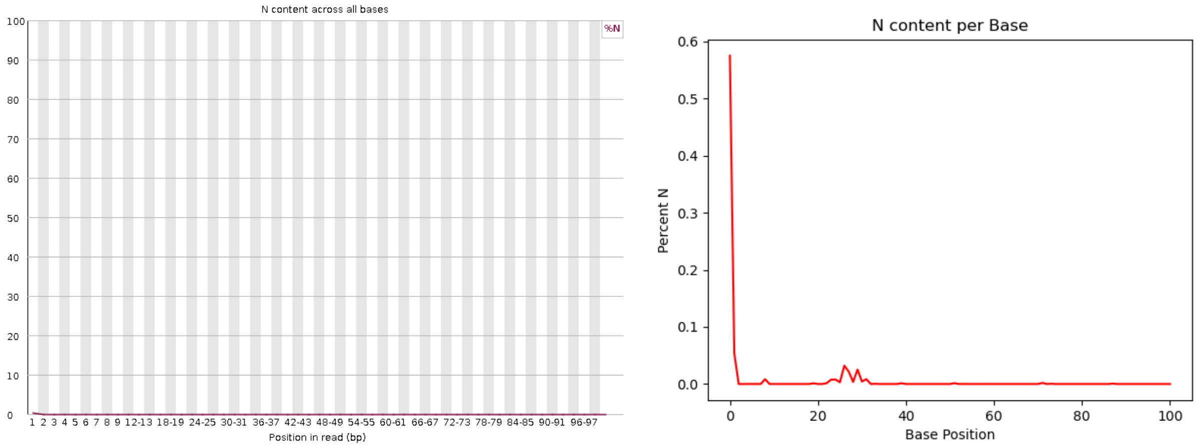
Per Base N−Content for 3_2B_control_S3_L008_R1_001



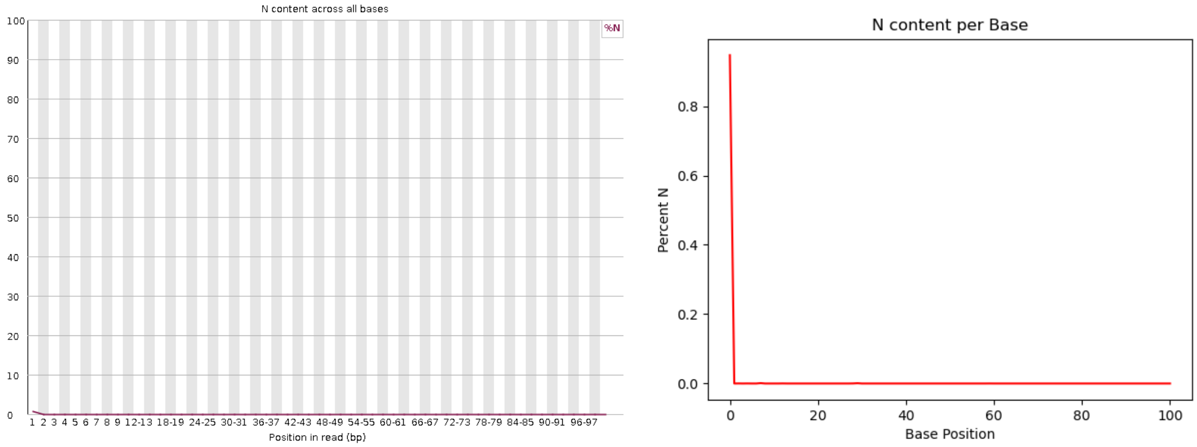
Per Base N−Content for 3_2B_control_S3_L008_R2_001



Per Base N−Content for 28_4D_mbnl_S20_L008_R1_001



Per Base N−Content for 28_4D_mbnl_S20_L008_R2_001

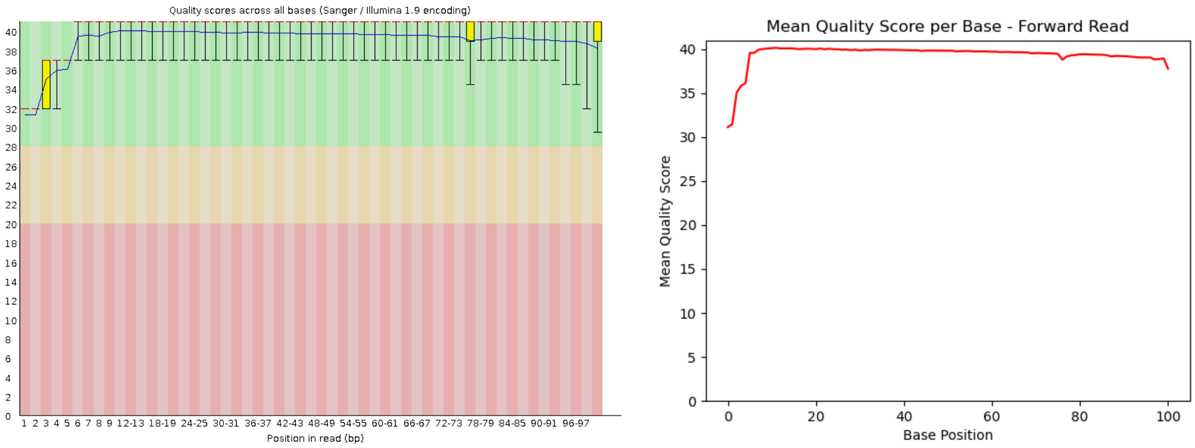


Step 3

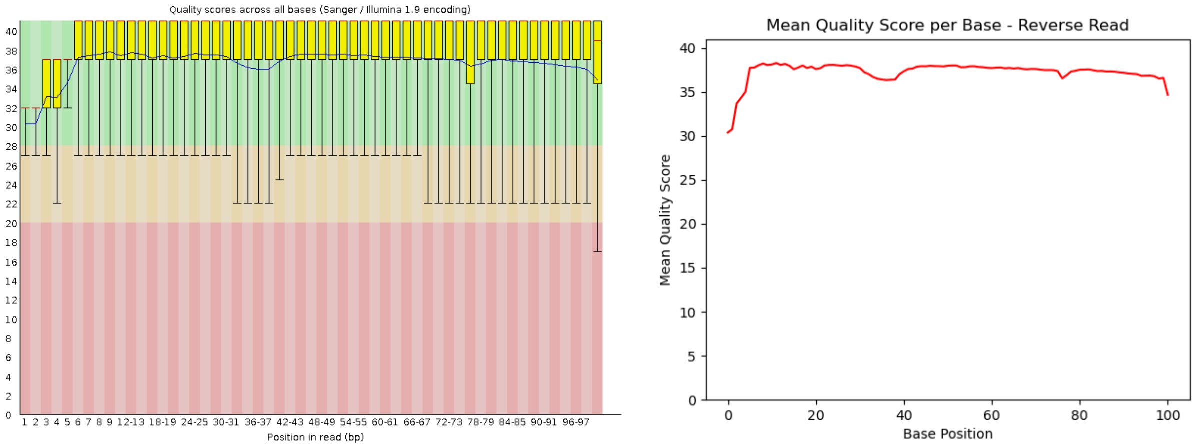
Run your quality score plotting script from your Demultiplexing assignment in Bi622. (Make sure you're using the "running sum" strategy!!) Describe how the FastQC quality score distribution plots compare to your own. If different, propose an explanation. Also, does the runtime differ? Mem/CPU usage? If so, why?

Answer: The FastQC quality score distribution plots reflect the same trend shown in my produced plots. The fastqc plots also show the interquartile range for each base. Fastqc produced several plots for each file within 1 minute, using 300% of CPU given. Just producing 1 plot for each file using my Demultiplexing script took 7 minutes, using 98% of the CPU given(same as fastqc). The large difference in runtime is due to the level of coding using, the language used, and programs used to produce the plots.

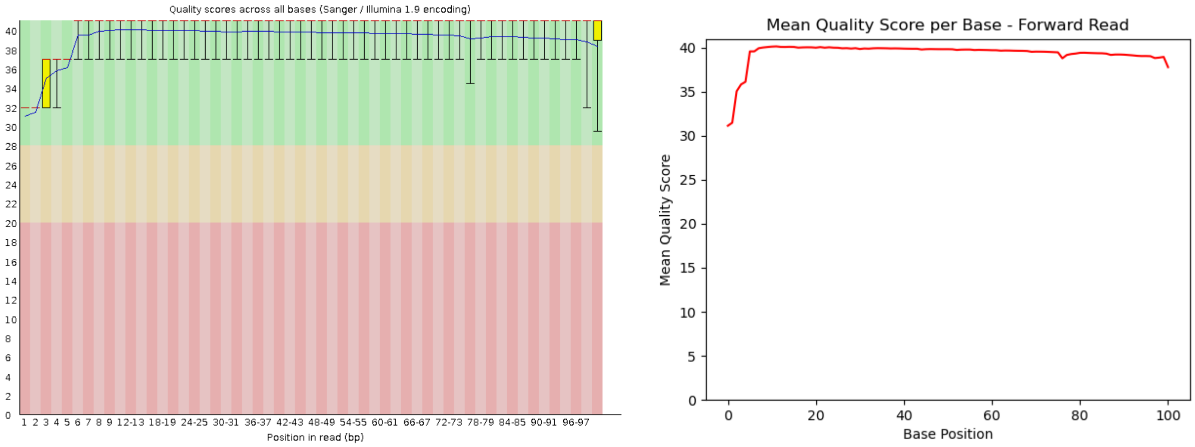
Per Base Quality Score for 3_2B_control_S3_L008_R1_001



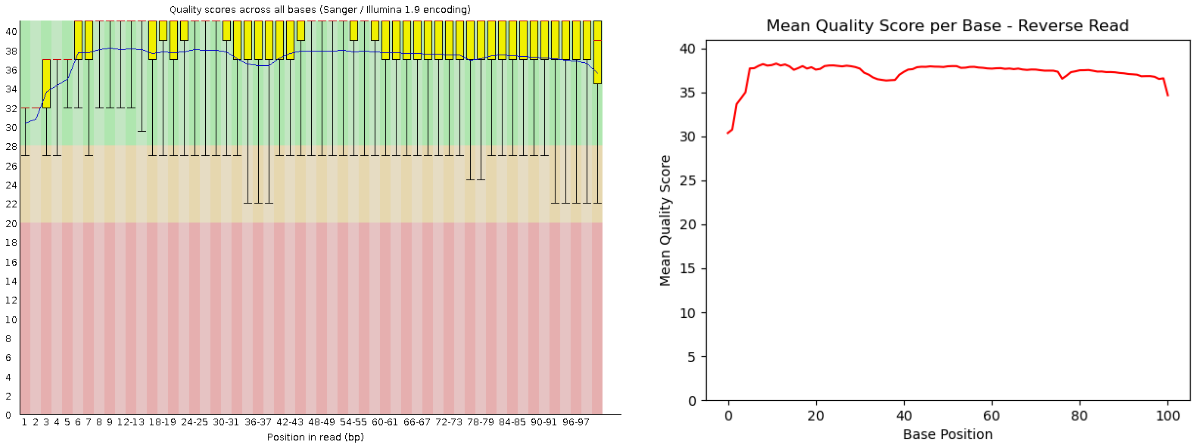
Per Base Quality Score for 3_2B_control_S3_L008_R2_001



Per Base Quality Score for 28_4D_mbnl_S20_L008_R1_001



Per Base Quality Score for 28_4D_mbnl_S20_L008_R2_001



Step 4

Comment on the overall data quality of your two libraries. Go beyond per-base qscore distributions. Make and justify a recommendation on whether these data are of high enough quality to use for further analysis.

Answer: I believe these data are high enough quality to use for further analysis. Looking beyond per-base qscore distributions, I can consider other statistics such as Sequences flagged as poor quality. All had 0 flagged. I can then look at sequence duplication and overrepresented sequences, none of which give me a "red x" evaluation. Per sequence GC content is also "green level" for all reads.

Part 2

Trim Reads

Step 6

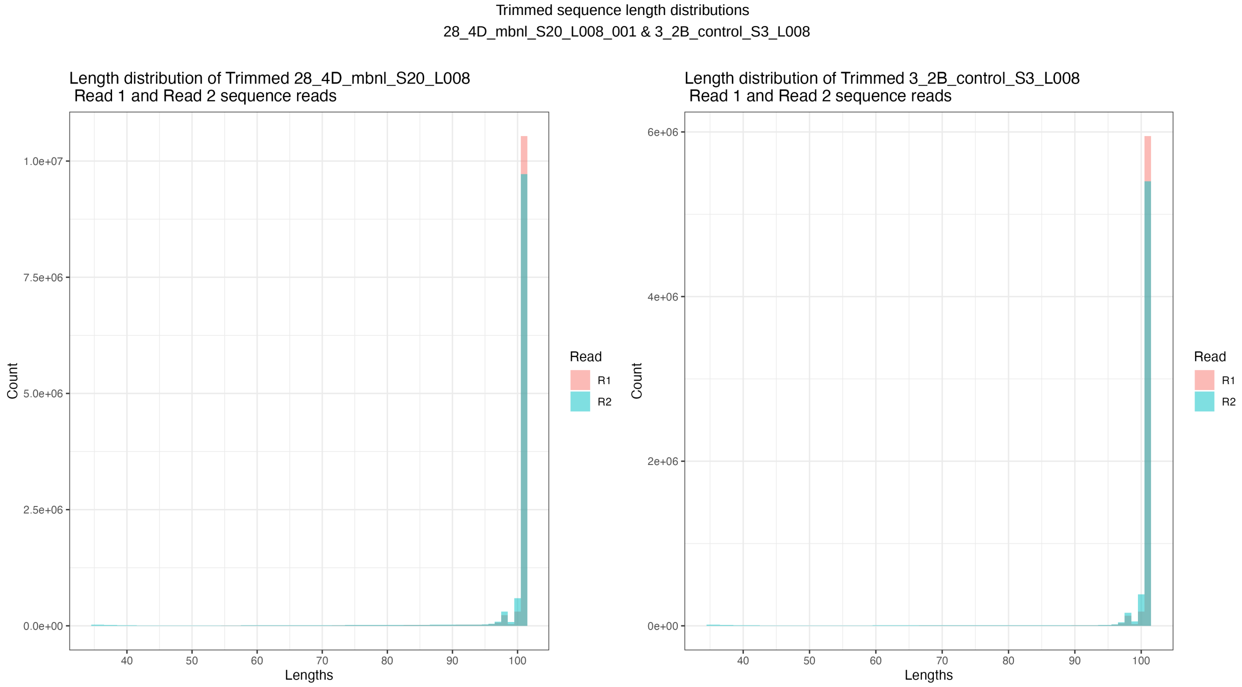
Using cutadapt, properly trim adapter sequences from your assigned files. Be sure to read how to use cutadapt. Use default settings. What proportion of reads (both R1 and R2) were trimmed?

Answer: 3_2B_control_S3_L008_R1_001.fastq.gz had 3.2% trimmed.
3_2B_control_S3_L008_R2_001.fastq.gz had 3.9% trimmed.
28_4D_mbnl_S20_L008_R1_001.fastq.gz had 6.0% trimmed.
28_4D_mbnl_S20_L008_R2_001.fastq.gz had 6.8% trimmed.

Step 8

Plot the trimmed read length distributions for both R1 and R2 reads (on the same plot - yes, you will have to use Python or R to plot this. See ICA4 from Bi621). You can produce 2 different plots for your 2 different RNA-seq samples. There are a number of ways you could possibly do this. One useful thing your plot should show, for example, is whether R1s are trimmed more extensively than R2s, or vice versa. Comment on whether you expect R1s and R2s to be adapter-trimmed at different rates and why.

Answer: I would expect R2 reads to be trimmed more extensively due to the lower quality of R2 reads. R2 reads also have a higher N-content. R2 reads have also been on the sequencer long and have thus experienced higher rates of degradation.



Part 3

Step 12

Using your script from PS8 in Bi621, report the number of mapped and unmapped reads from each of your 2 sam files. Make sure that your script is looking at the bitwise flag to determine if reads are primary or secondary mapping (update/fix your script if necessary).

28_4D_mbnl_S20_L008.Aligned.out.sam
Mapped: 22657642
Unmapped: 793158

3_2B_control_S3_L008.Aligned.out.sam
Mapped: 12359963
Unmapped: 496075

Step 14

Demonstrate convincingly whether or not the data are from "strand-specific" RNA-Seq libraries. Include any comands/scripts used. Briefly describe your evidence, using quantitative statements (e.g. "I propose that these data are/are not strand-specific, because X% of the reads are y, as opposed to z.").

Answer: I propose the library that these data are not strand-specific, because 42.6% and 82.7% of reads in the stranded=reverse htseq-count data is mapped to genes, as opposed to 1.79% and 3.5% in the stranded=yes htseq-count data. Per the htseq-count documentation, -stranded=yes is for strand-specific libraries.

htseq_stranded_3_2B_control_S3_L008
Mapped: 221658
Total: 12359963
Proportion: 0.0179

htseq_reverse_3_2B_control_S3_L008
Mapped: 5260168
Total: 12359963
Proportion: 0.426

htseq_stranded_28_4D_mbnl_S20_L008
Mapped:411314
Total: 11725400
Proportion: 0.035

htseq_reverse_28_4D_mbnl_S20_L008
Mapped: 9700357
Total: 11725400
Proportion: 0.827