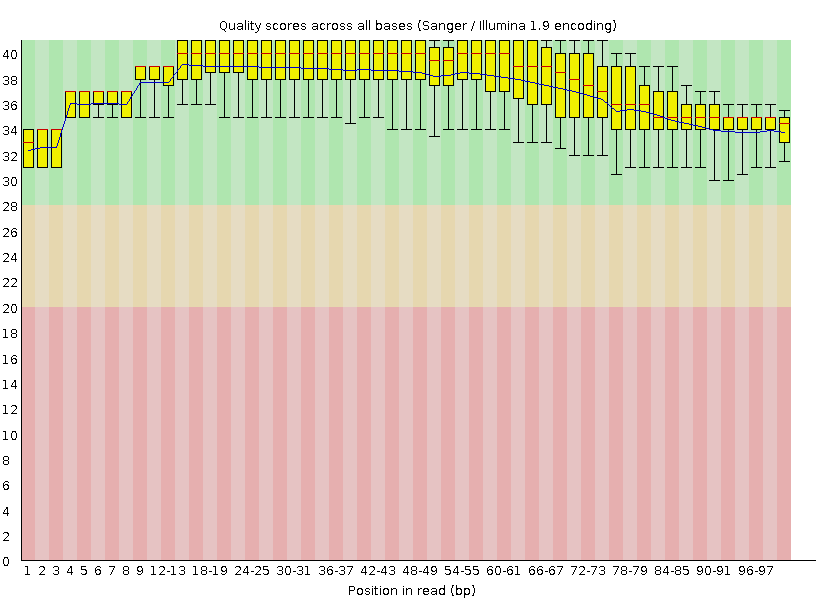
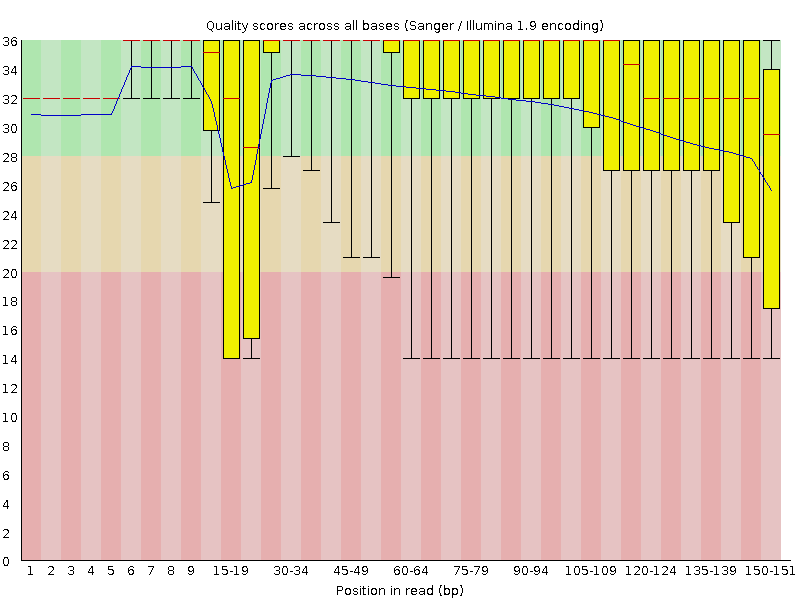
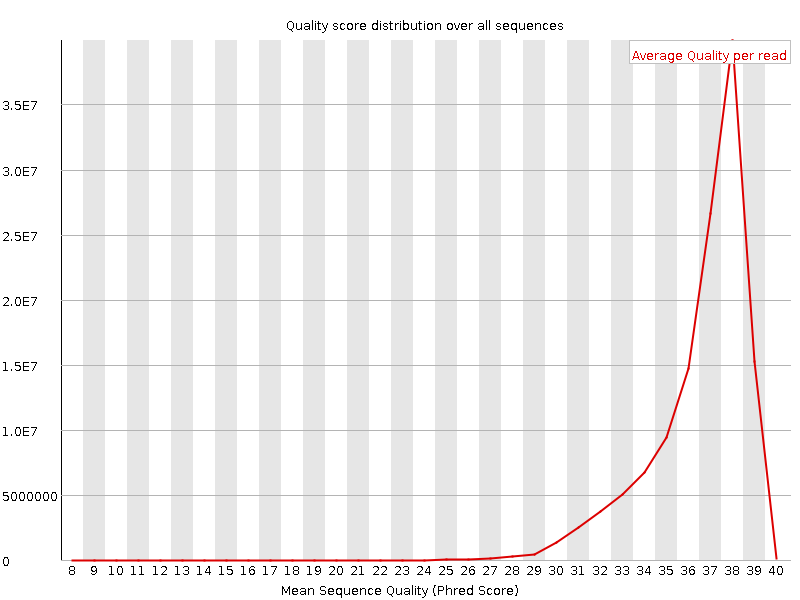
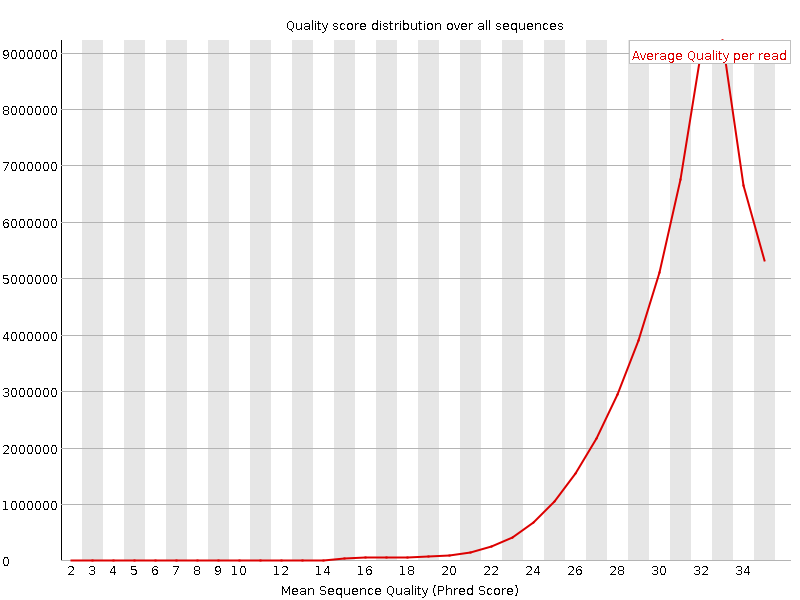
Option 1

I attempted to do option 1, but got stuck because I was not finding the 3

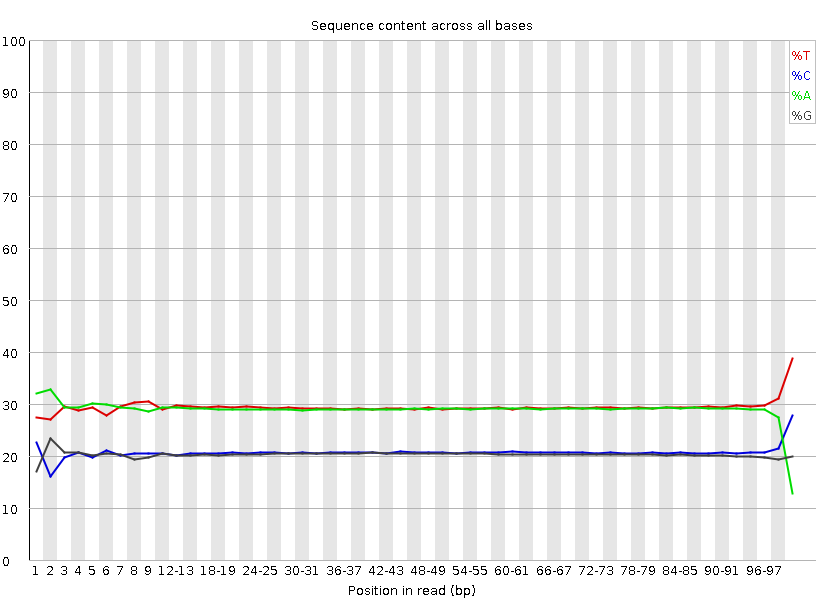
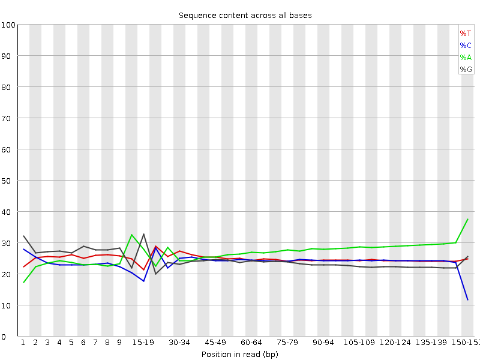
Option 2

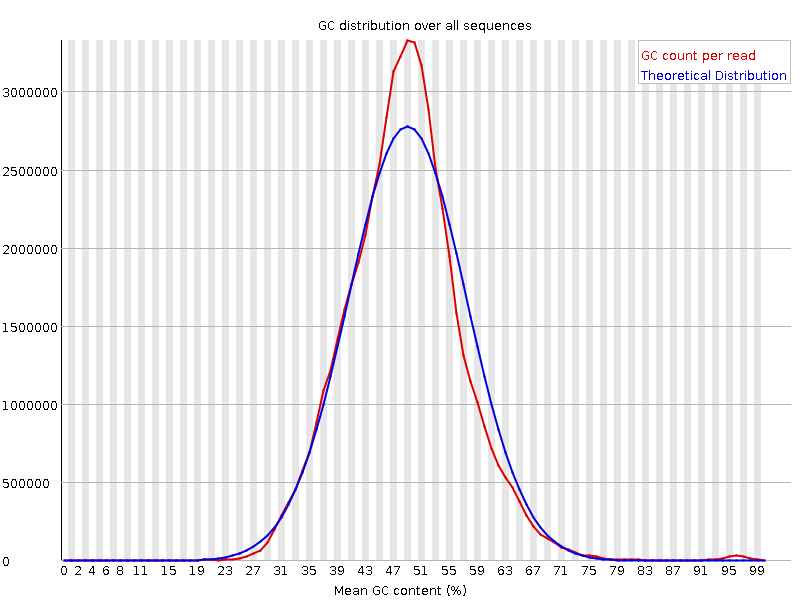


There is a significant difference between the FastQC report without trimming on the left and with trimming and Phred score on the right. Positions 15-19, and 105+ had a larger spread of quality scores that was cleaned with the Trim tool.



Additionally, the quality score distribution improved to an average of 38. Some other significant changes:



Chart, line chart

Description automatically generated

The FastQC report seemed to flag the per base sequence content for the trimmed reads, but further reading from Biostars examples state that this is common and usually doesn’t affect downstream results. It’s possibly a result of some small bias introduced in the library prep. The mean GC content appears greater than the normal distribution in both cases, and for the trimmed reads, there the left side of the curve has a lower than theoretically expected GC while the right has a higher than theoretically expected GC. This still doesn’t seem bad compared to some examples.

Next, I ran the markDuplicates software on the trimmed reads. First, I used Bowtie2 to align the hg38 canonical reference genome because I wasn’t sure what the difference was. It does seem like BWA-MEM software was faster. Next, I used freeBayes and VCFfilter as specified.

1. I was actually able to select the UCSC genome browser directly from Galaxy without having to download. This was done with the following steps:

Graphical user interface, text, application

Description automatically generatedGraphical user interface, text, application

Description automatically generated

Timeline

Description automatically generated

After running the R script, I did not find the variants rs104886003, rs121434569, or rs80359031. The first columns appeared as follows:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| chrom | pos | ref | alt | chrom | chromStart | chromEnd | name | refUCSC | func |
| chr1 | 382523 | G | A | chr3 | 382522 | 382523 | rs142697664 | G | coding-synon,intron |
| chr1 | 123487610 | G | T | chr6 | 123487609 | 123487610 | rs207467421 | A | intron |
| chr1 | 124559883 | T | A | chr7 | 124559882 | 124559883 | rs207468543 | C | unknown |
| chr1 | 182382492 | TCTTAT | ACTTAAATGAAT | chr1 | 182382491 | 182382495 | rs886045605 | TCTT | near-gene-3,untranslated-3 |
| chr2 | 33989353 | A | G | chr5 | 33989352 | 33989353 | rs886060530 | C | ncRNA,missense,untranslated-3 |
| chr3 | 75739979 | T | C | chr1 | 75739978 | 75739979 | rs875989869 | G | splice-3 |
| chr4 | 6003804 | G | A | chr7 | 6003803 | 6003804 | rs562762722 | T | intron |
| chr4 | 9362076 | T | C | chr6 | 9362075 | 9362076 | rs207466697 | G | unknown |
| chr4 | 38550791 | G | A | chr3 | 38550790 | 38550791 | rs199473636 | C | missense |
| chr4 | 49104455 | G | A | chr3 | 49104454 | 49104455 | rs587777331 | C | ncRNA,missense |
| chr4 | 49117775 | A | C | chr3 | 49117774 | 49117775 | rs267599866 | G | coding-synon,near-gene-5 |
| chr4 | 49133013 | T | C | chr3 | 49133012 | 49133013 | rs886058678 | C | untranslated-5 |
| chr4 | 49636656 | T | G | chr6 | 49636655 | 49636659 | rs387906519 | GAGG | frameshift |
| chr4 | 189896450 | G | C | chr3 | 189896449 | 189896449 | rs886058235 | - | untranslated-3 |
| chr5 | 47403309 | T | A | chr2 | 47403308 | 47403309 | rs63751260 | G | intron,ncRNA,missense |
| chr5 | 47463034 | A | G | chr2 | 47463033 | 47463034 | rs587779088 | G | ncRNA,frameshift |
| chr5 | 47476525 | CTTT | TTTC | chr2 | 47476524 | 47476525 | rs587781996 | G | ncRNA,missense |
| chr5 | 47806863 | G | C | chr2 | 47806862 | 47806863 | rs786203308 | T | near-gene-3,untranslated-3 |
| chr5 | 48499608 | T | C | chr4 | 48499607 | 48499608 | rs267600177 | G | coding-synon |
| chr5 | 48587187 | GTTTTTAAATTC | GTTTTTTAAATTC | chr7 | 48587186 | 48587187 | rs267601532 | G | ncRNA,missense |
| chr5 | 48963778 | A | T | chr2 | 48963777 | 48963778 | rs386833510 | G | missense |
| chr5 | 49126157 | T | G | chr3 | 49126156 | 49126157 | rs886058675 | C | coding-synon |
| chr5 | 49133067 | G | C | chr3 | 49133066 | 49133067 | rs886058679 | G | near-gene-5,untranslated-5 |
| chr5 | 49420216 | GT | AC | chr3 | 49420215 | 49420230 | rs386833683 | TGAGGGCCAAATCTT | intron,ncRNA,cds-indel |
| chr5 | 50136285 | G | C | chr8 | 50136284 | 50136285 | rs207469145 | T | intron |
| chr6 | 5973467 | T | A | chr7 | 5973466 | 5973467 | rs886039646 | A | ncRNA,frameshift |
| chr6 | 31355398 | C | T | chr6 | 31355397 | 31355398 | rs41558116 | C | near-gene-5,ncRNA,missense |
| chr7 | 58078779 | A | G | chr3 | 58078778 | 58078779 | rs121908895 | A | ncRNA,missense |
| chr7 | 58094901 | C | G | chr3 | 58094900 | 58094901 | rs886058760 | G | ncRNA,missense |
| chr7 | 61048558 | C | T | chr2 | 61048557 | 61048558 | rs149581494 | C | missense |

After this, I moved on to option 1.