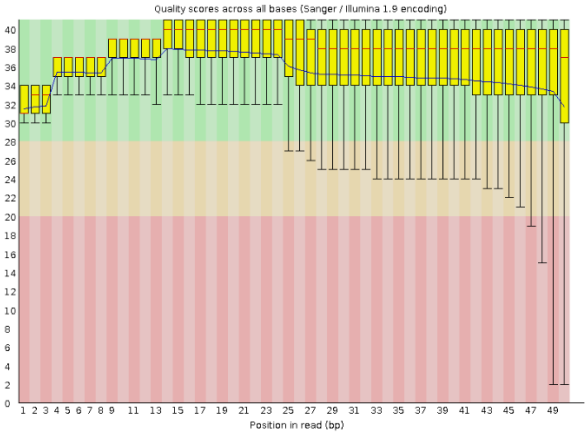
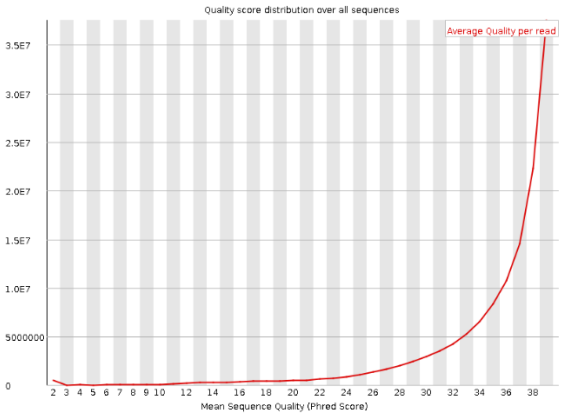
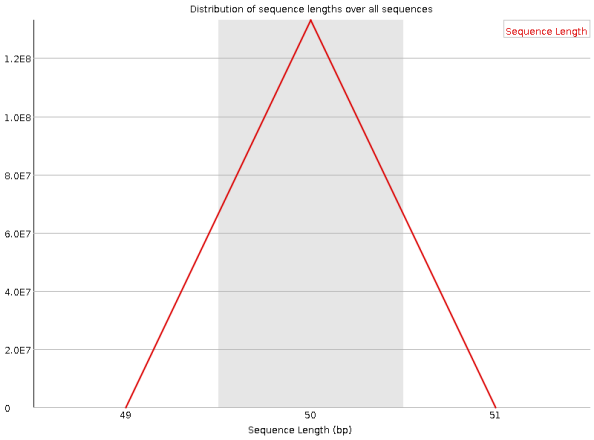
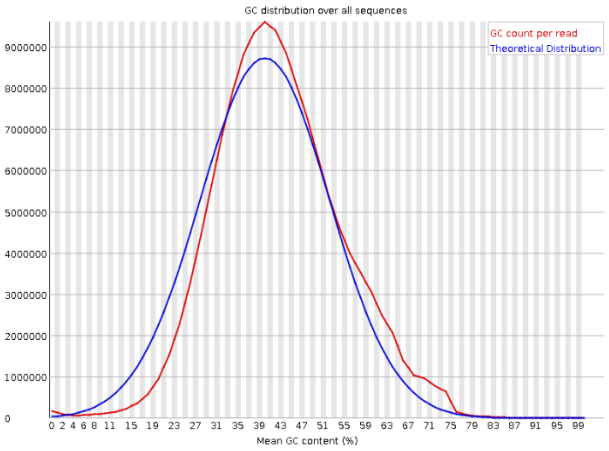
Assignment 2

Mary Silva









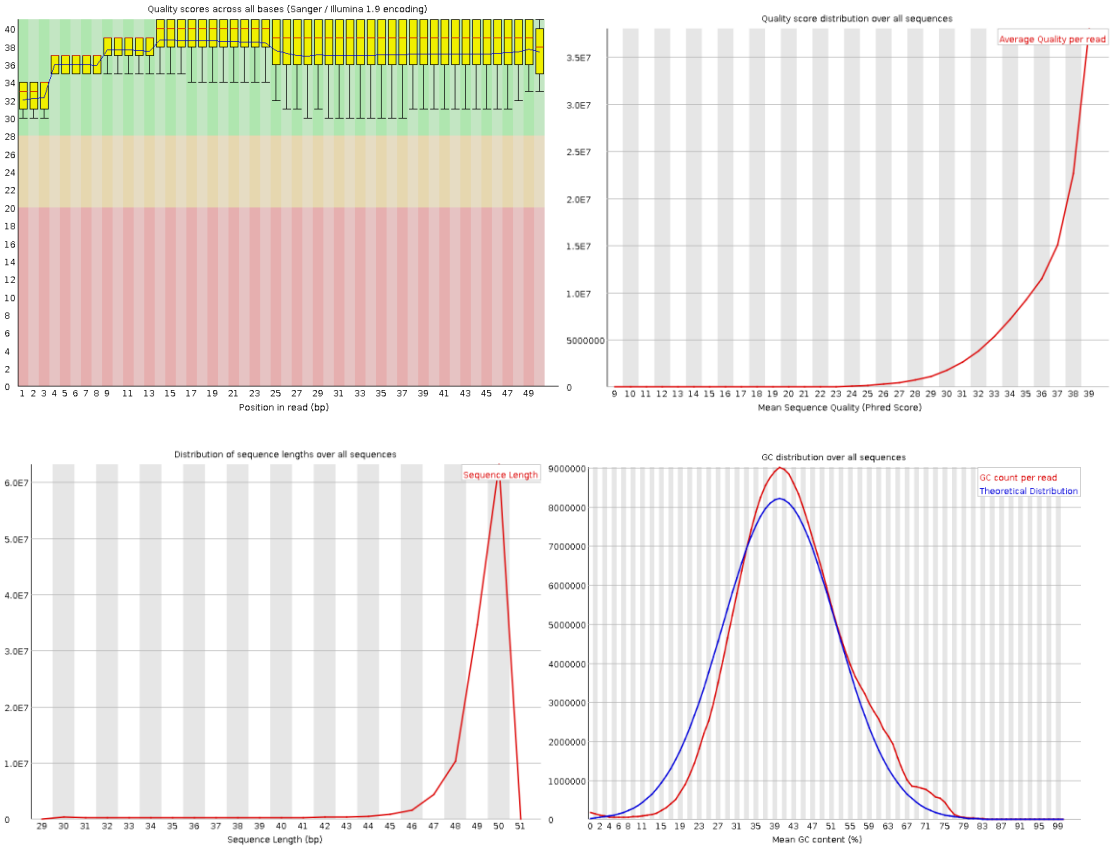
2. The Phred score across all bases appears to be within an acceptable range. There are some outliers in the rightmost positions, which could indicate that trimming would be helpful. The per sequence quality scores show a peak converging over 30, which correlates to a probability of an incorrect base call 1 in 1000 times. The sequence length distribution shows that the read lengths peak at 50 base pairs. And the per sequence GC distribution is very close to the expected theoretical distribution.
3. After using Trim Galore, the summary is shown below

```
=== Summary ===

Total reads processed:             133,020,500
Reads with adapters:                47,849,621 (36.0%)
Reads written (passing filters):   133,020,500 (100.0%)

Total basepairs processed: 6,651,025,000 bp
Quality-trimmed:             509,904,004 bp (7.7%)
Total written (filtered):  6,070,045,697 bp (91.3%)
```

Which shows that 91.3% passed the filter, and 7.7% was quality-trimmed down under the condition of a Phred score cutoff of 30.



4. After trimming, we see the distributions of quality scores across all bases are within a good range, without any outliers. The Phred scores converge more quickly on the top right plot and have a higher average, almost to a Phred score of 40. The peak of the sequence lengths are still 50 base pairs and the mean GC count per read didn't change much either. This means that the only difference trimming the 7.7% of reads mainly affected the Phred score.