RNA-seq Quality Assessment Assignment

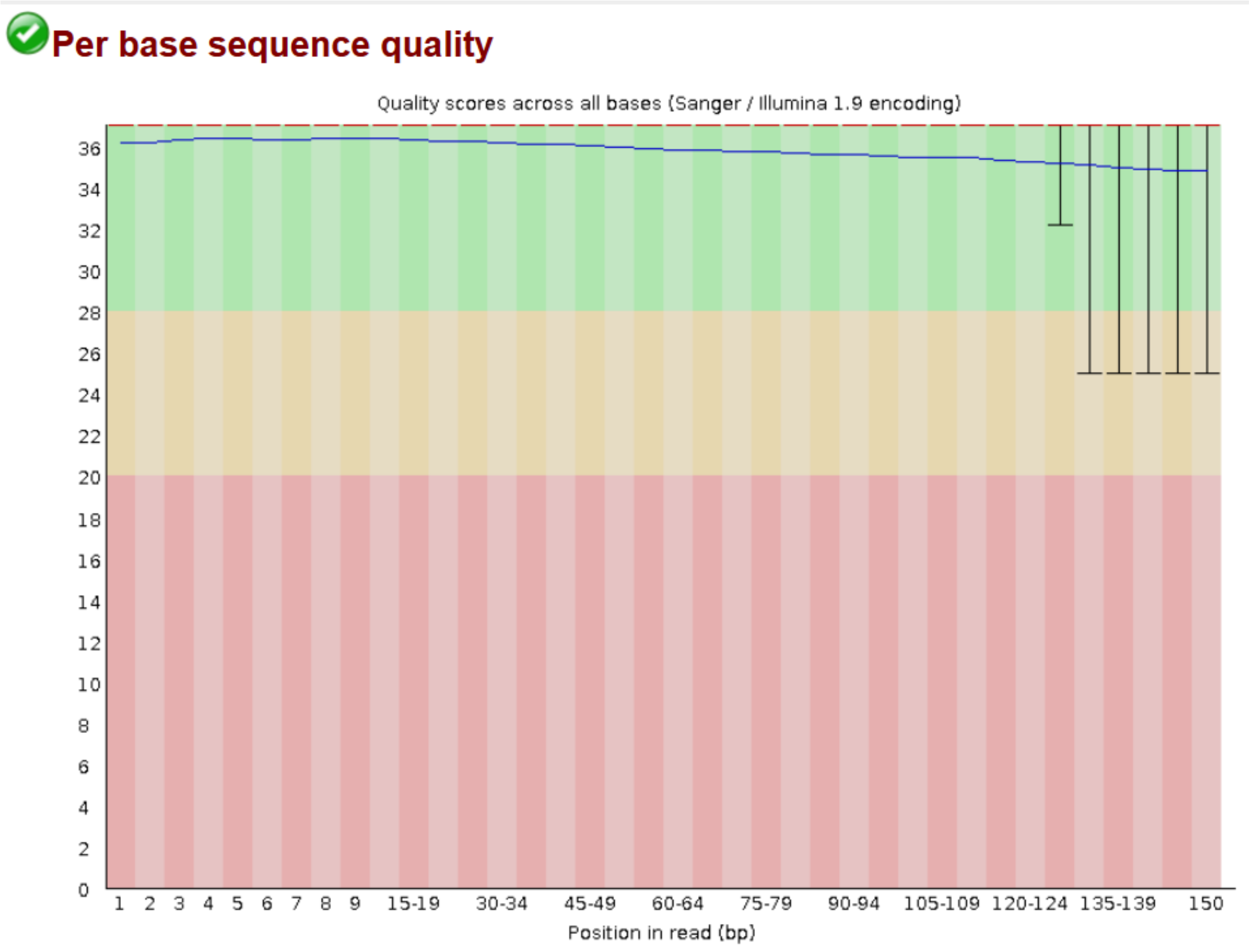
Brandy Corwin

2025-09-07

Part 1

CcoxCrh_comrhy59_EO_6cm_1

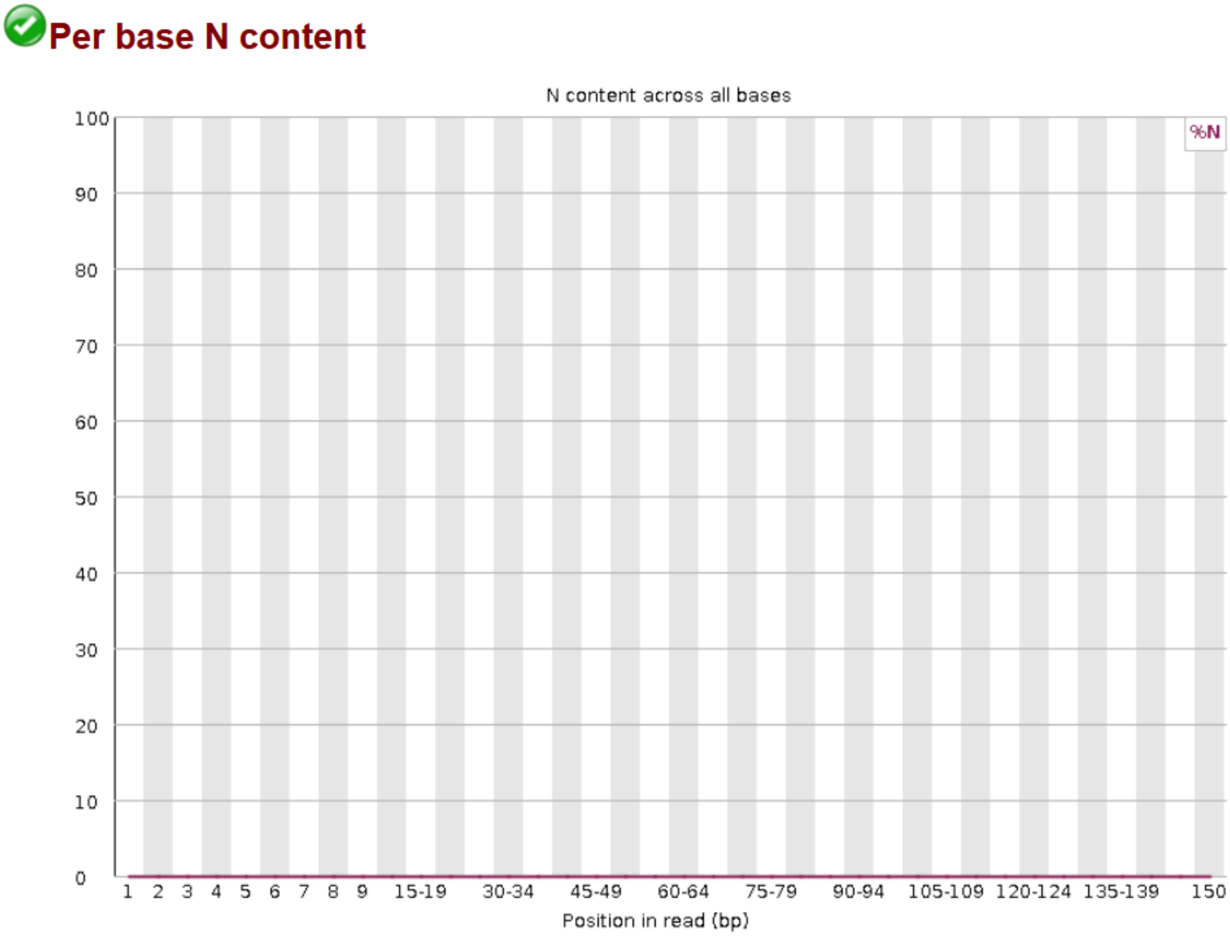
Per base quality score of read 1



Per base sequence quality

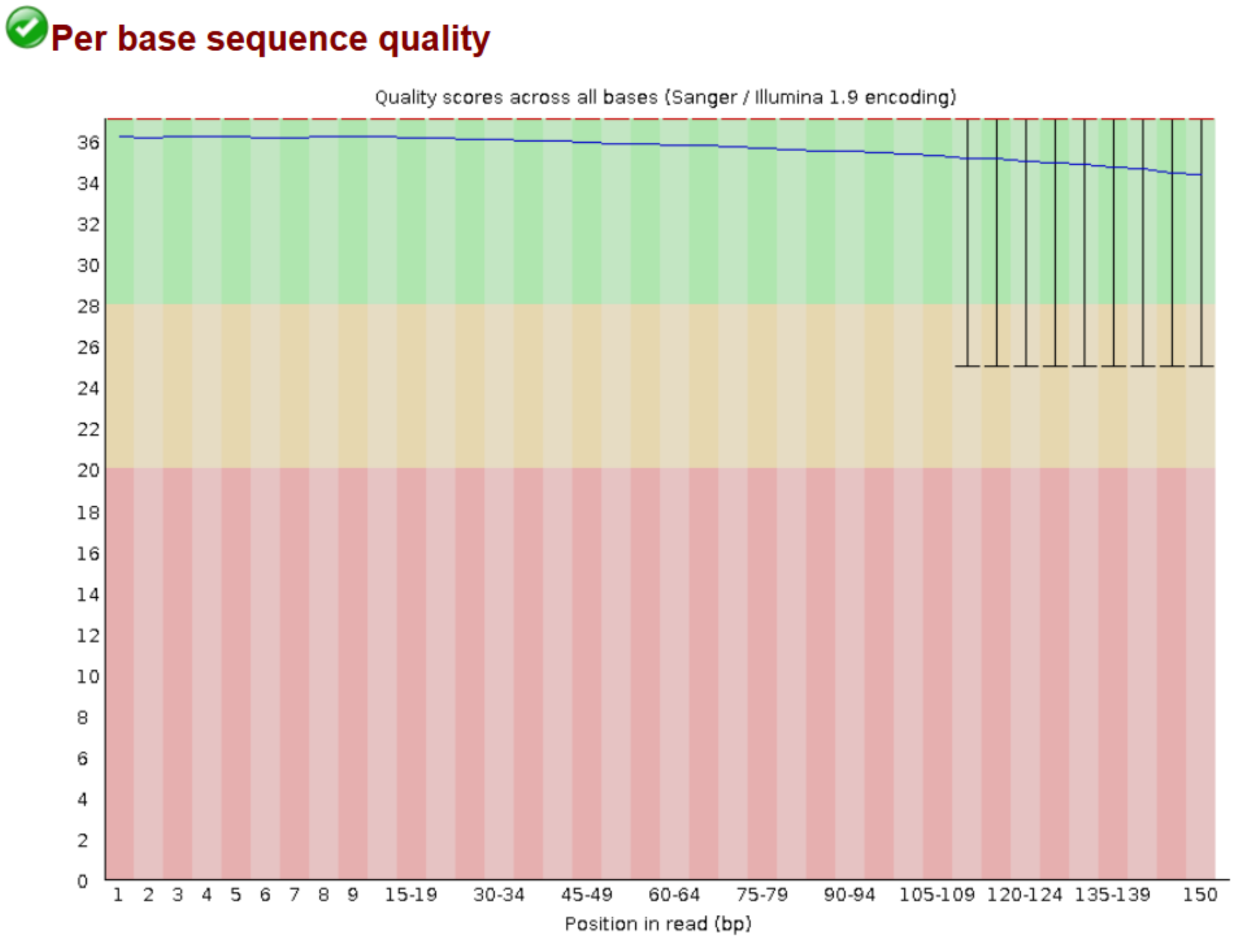
The quality score of each position averaged for each read in the file.

N content of read 1



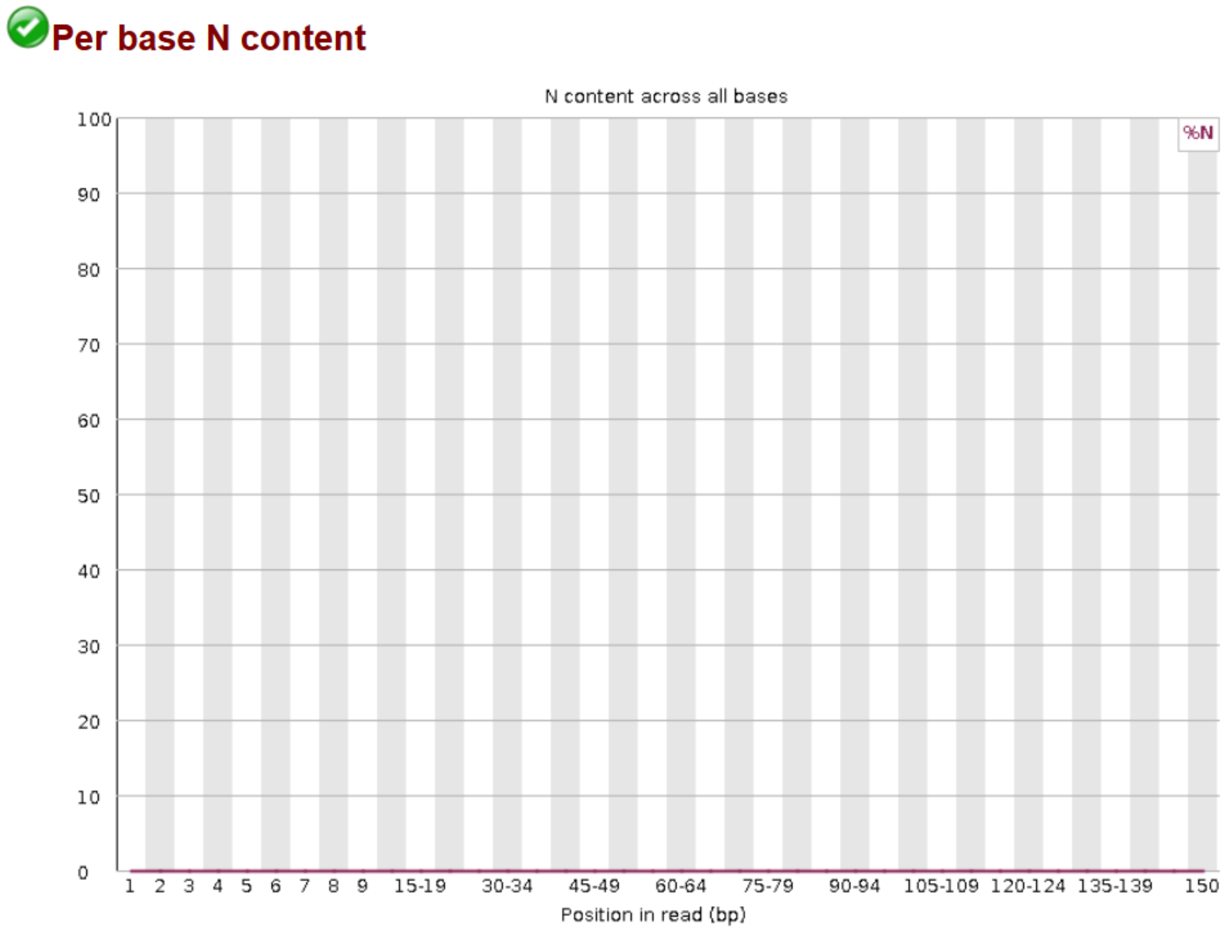
The percent of Ns accross all read in the file at each position in the read.

Per base quality score of read 2



The quality score of each position averaged for each read in the file.

N content of read 2

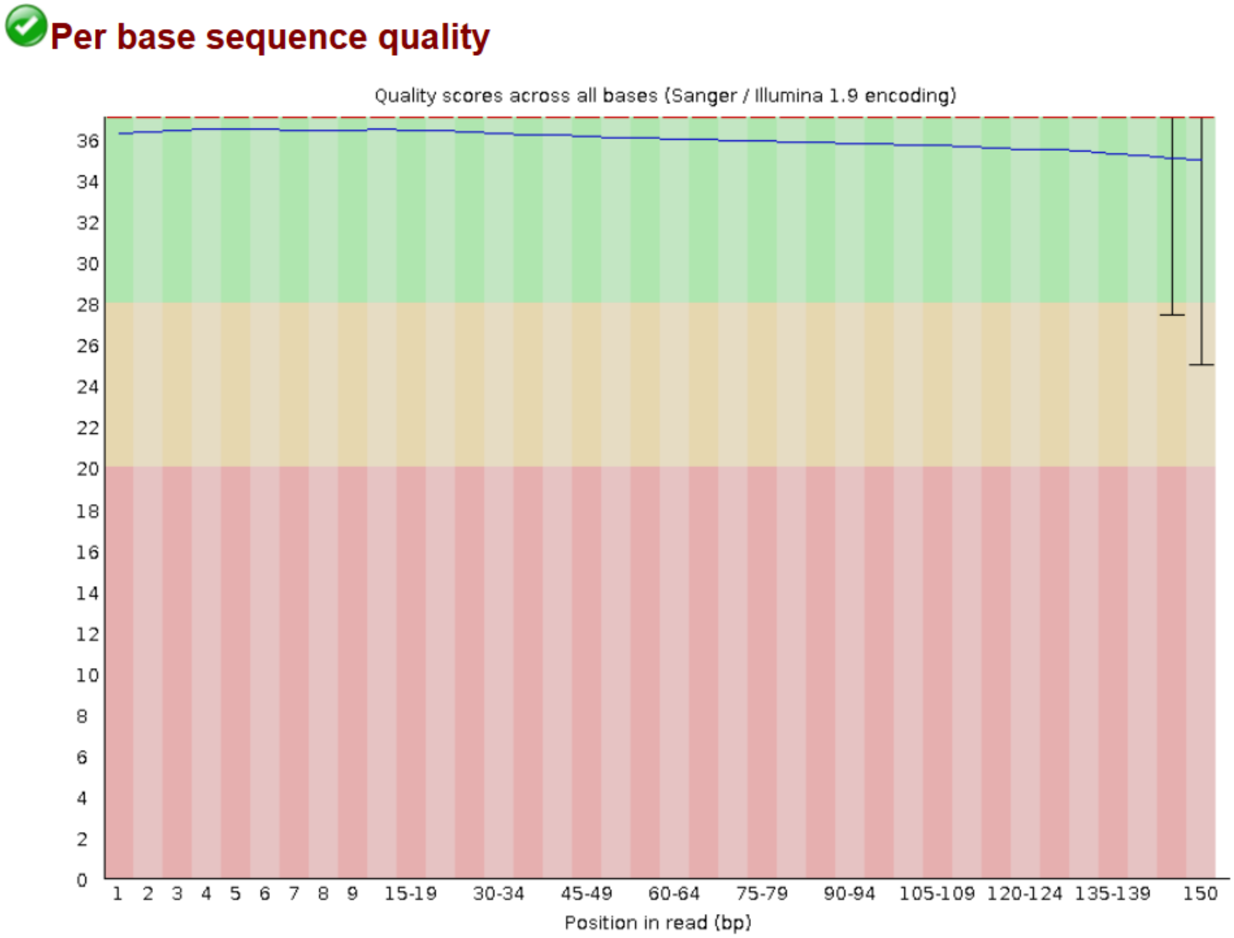


The percent of Ns accross all read in the file at each position in the read.

Both the plots make sense with each other because no/low Ns in the would likely mean that the data is high quality, just like the plot shows.

CcoxCrh_comrhy113_EO_adult_2

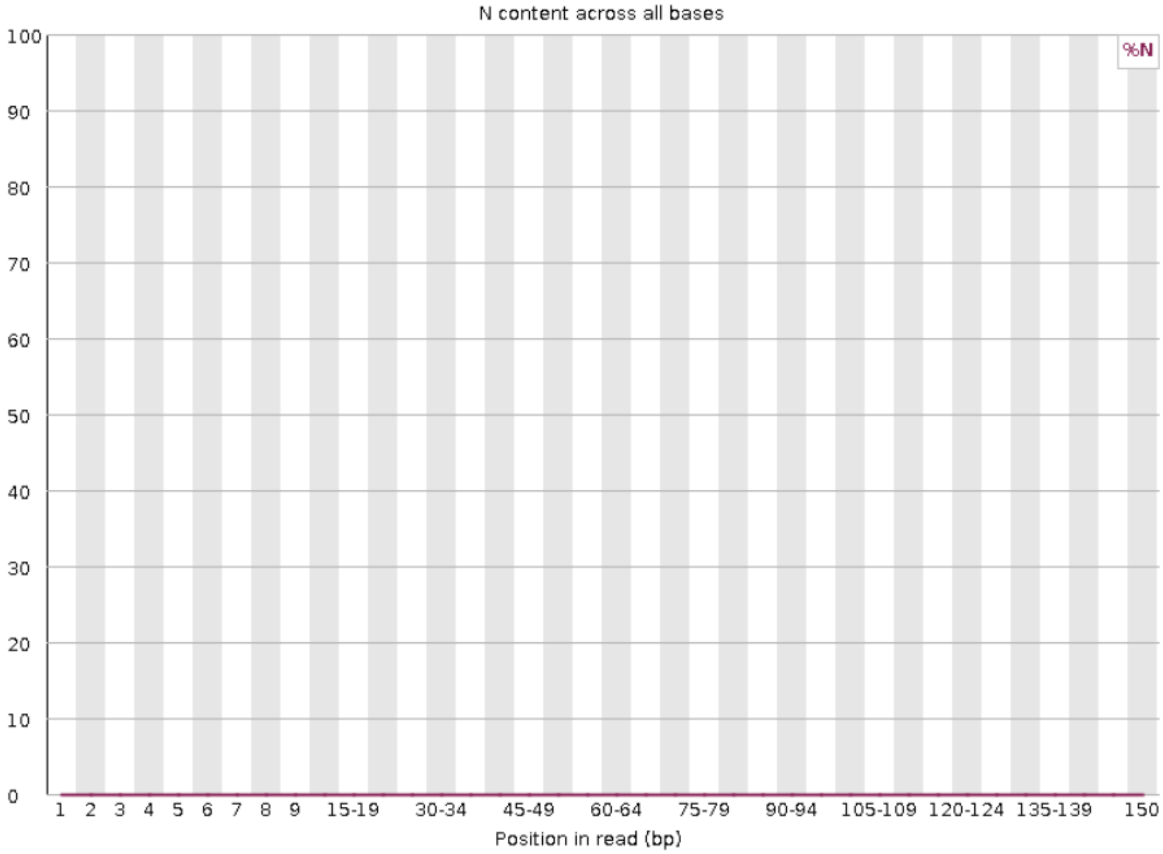
Per base quality score of read 1



The quality score of each position averaged for each read in the file.

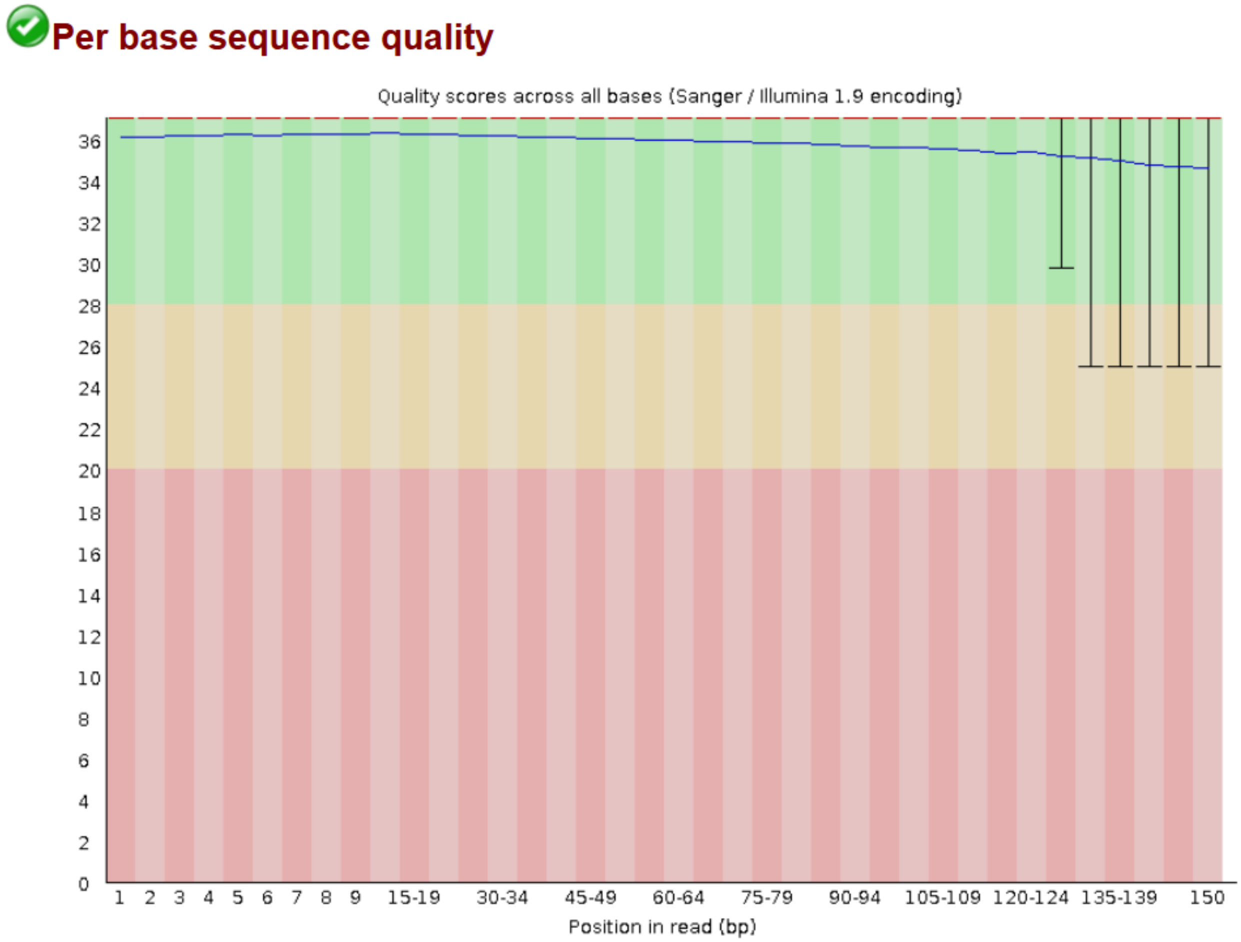
N content of read 1

✓Per base N content



The percent of Ns accross all read in the file at each position in the read.

Per base quality score of read 2

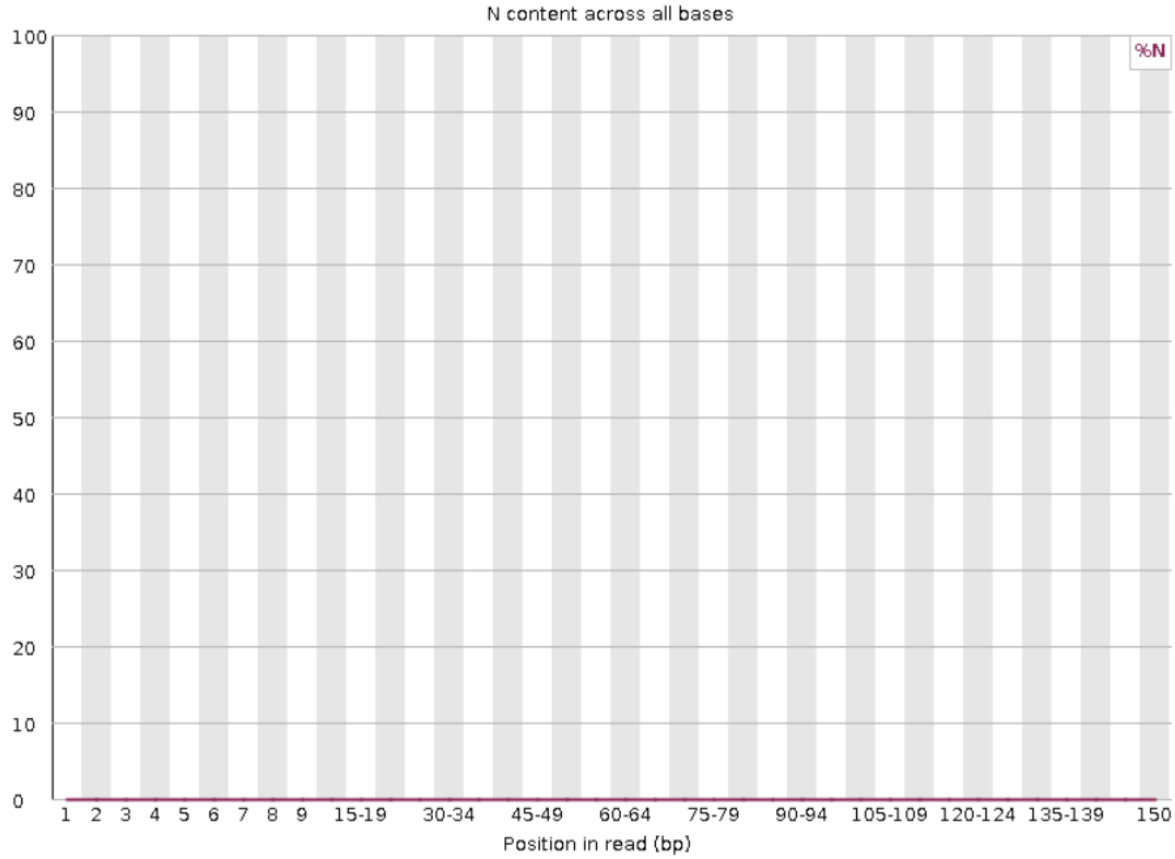


✓ Per base sequence quality

The quality score of each position averaged for each read in the file.

N content of read 2

✓ Per base N content

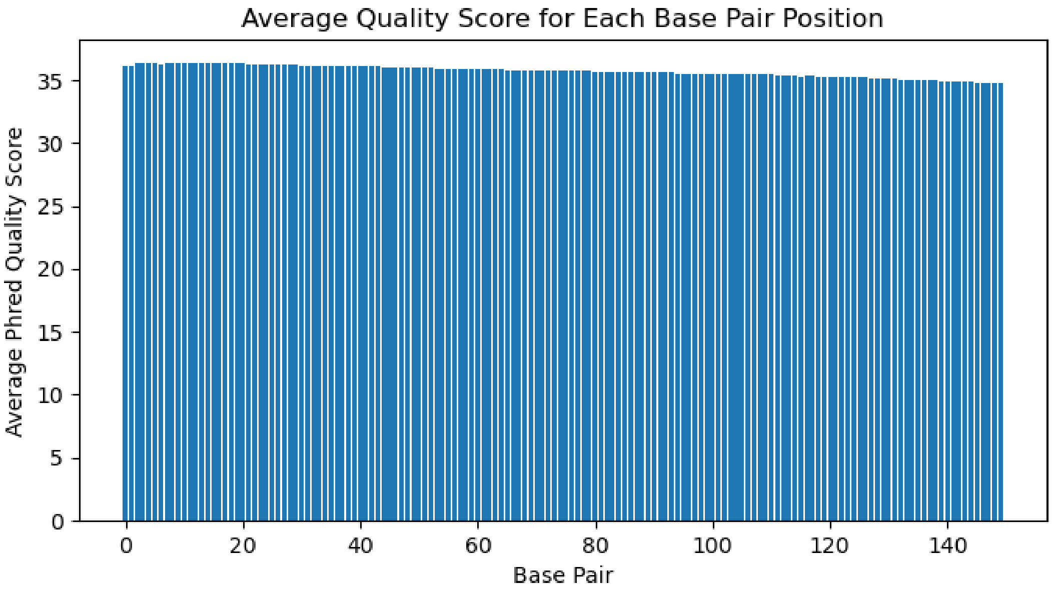


The percent of Ns accross all read in the file at each position in the read.

All of these plots also had a high quality score with little to no Ns.

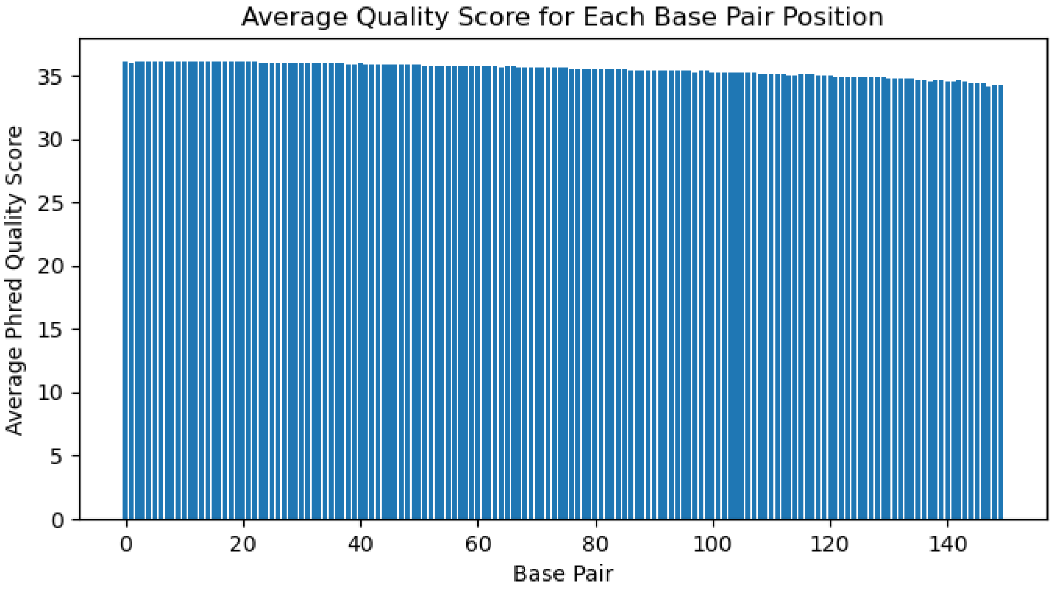
My per base quality score graphs

CcoxCrh_comrhy59_EO_6cm_1 read 1



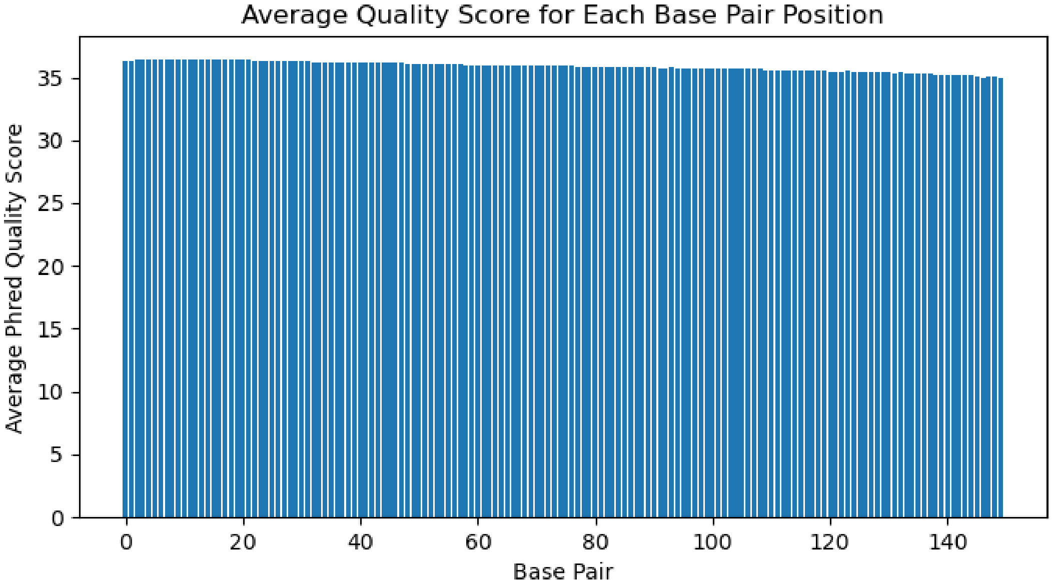
The quality score of each position averaged for each read in the file.

CcoxCrh_comrhy59_EO_6cm_1 read 2



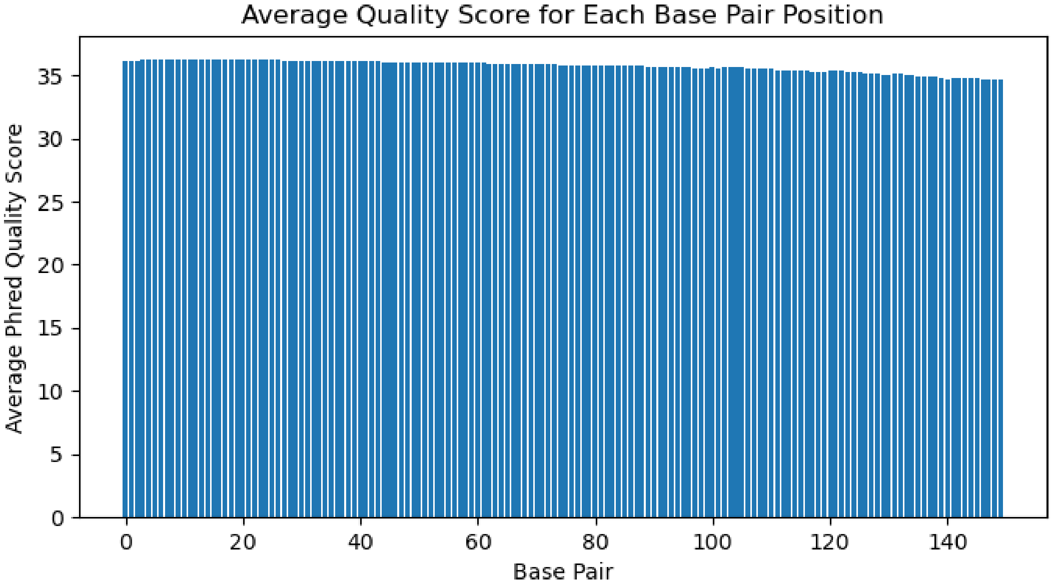
The quality score of each position averaged for each read in the file.

CcoxCrh__comrhy133__EO__adult__2 read 1



The quality score of each position averaged for each read in the file.

CcoxCrh__comrhy133__EO__adult__2 read 2



The quality score of each position averaged for each read in the file.

My per base quality score graphs show about the same quality scores as the ones from fastqc, but they took much longer to run. I only gave 1 cpu to both types of runs so they both used 99% cpu, but my script took 3 times longer than fastqc and I was only doing the quality distribution, while fastqc also did many others. This is likely because there code is optimized much better than mine is.

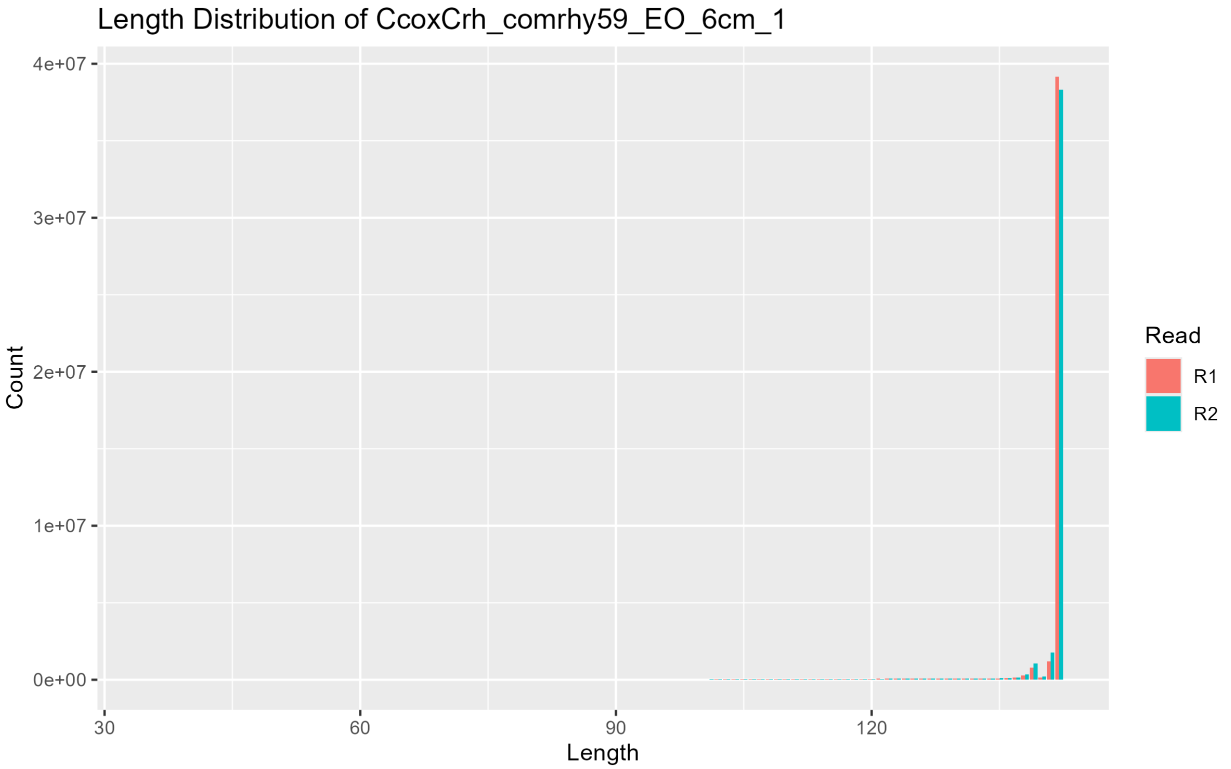
Overall, the sequence quality looks good for all four files. The first 10 bases of the reads look bad due to barcodes and there are some adapters still in my sequences. Those can easily be fixed by trimming the data.

Part 2

After using cutadapt on my CcoxCrh_comrhy59_EO_6cm_1 files, 6.4% from read 1 and , 7.2% from read 2 were trimmed.

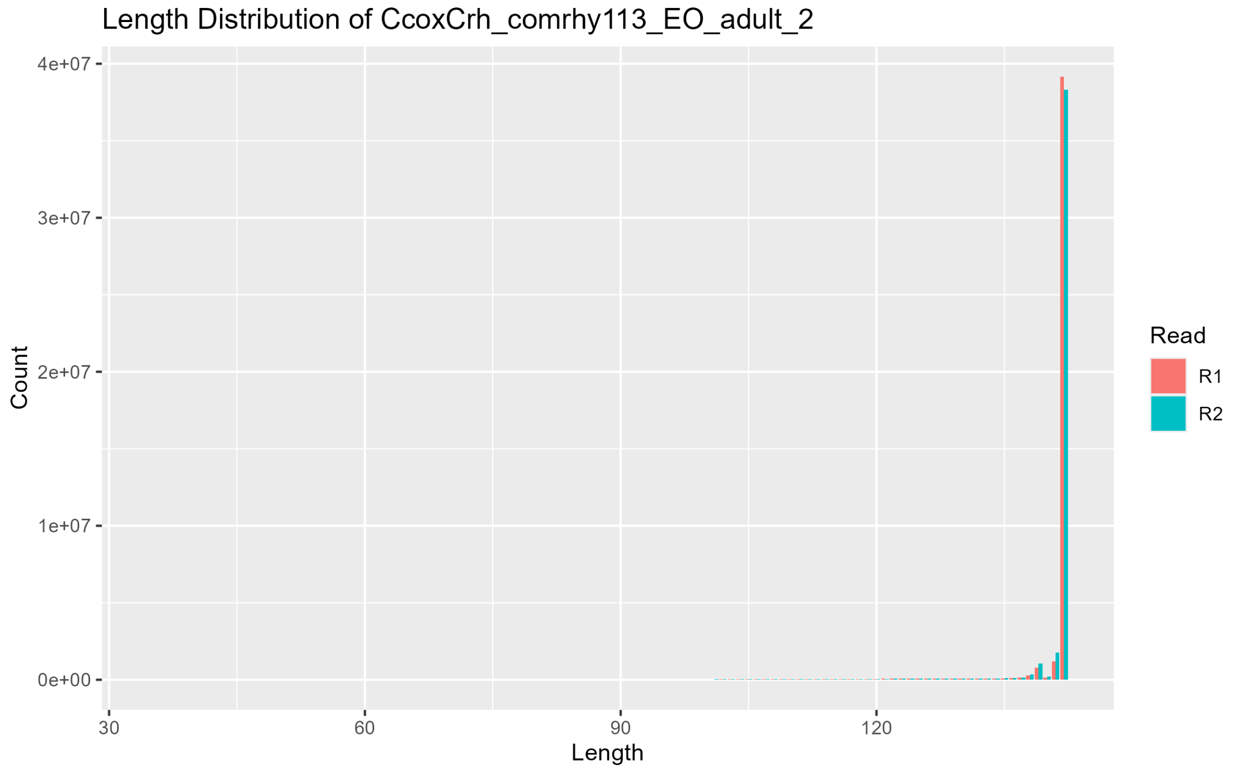
After using cutadapt on my CcoxCrh_comrhy133_EO_adult_2 files, 11.1% from read 1 and , 11.7% from read 2 were trimmed.

CcoxCrh_comrhy59_EO_6cm_1 length distribution



The counts of reads of each length size in both read 1 and read 2 files after trimming.

CcoxCrh__comrhy113__EO__adult__2 length distribution



The counts of reads of each length size in both read 1 and read 2 files after trimming.

Most reads from both files were only trimmed by 8 bases. Both read 2s were trimmed slightly more than their read 1 counterparts after using cutadapt, but not by a drastic percentage. This is likely because the different reads can have different adapter content. Since their adapters are in different places and sometimes get missed, one read could have the adapters be missed more than another.

Part 3

CcoxCrh_comrhy59_EO_6cm_1

mapped: 29749319
unmapped: 9402878

CcoxCrh_comrhy113_EO_adult_2

mapped: 25900086
unmapped: 10680808

Based on both my bash commands listed in strandedness_test/ the reads are reverse strand specific. For both samples, the stranded=yes had about 2% of reads mapped while the stranded=reverse had about 30% of reads mapped.