FastQC and Trimmomatic – Lab 1

**Introduction:**

The field of genomics was revolutionized with the introduction of Next Generation Sequencing (NGS) in the early 2000’s. Another term for NGS is massively parallel sequencing because many versions of a gene can be sequenced simultaneously which has caused an exponential increase in the number of genomes sequenced and overall data accumulated. One of the most popular NGS options is the Illumina sequencing technologies. Illumina addresses some of the sequencing challenges common to genomics. One challenge is the massive size of genomes and massive number of species to sequence. Read errors is another challenge that Illumina has addressed, and compared to other technologies on the market, Illumina is one of the best options for reducing errors in reads. A read error can occur as a result of chemistry error within the polymerase reactions, or as a computational error from CASAVA, the Illumina software responsible for interpreting the base. Interpretation of each base is captured through a colored light emitted from the base once it attaches to the growing chain. One example of chemistry error is referred to as fading, when insufficient strands are being polymerized and the signal is not strong enough to visualize. Another example of misinterpretation is called phasing noise, where a blending of color occurs when strands of varying heights emit their fluorescence. Errors are common, which is why there are so many ways to reduce their effects of these errors. Illumina scores their reads by calculating a quality score for each base pair. The quality score is an estimation or error by calculating the probability an incorrect base is called. The equation Illumina uses is , but a quick estimate can be done by comparing the score to 10,20, or 30. A score of 10 equates to 1 error every 10 bases. Good sequences will be higher than a score of 20, 1 error every 100 bases, or 99% accuracy. High quality sequences score above 30, or 99.9% accuracy. The quality scores can be found in the FASTQ file generated by Illumina with the sequence metadata and sequence reads included as well.

Before you can make meaningful and biological interpretations of your data, it is important to verify the quality of the sequence data. The FASTQ file can be interpreted with a software package called FASTQC. It analyzes the FASTQ file and generates a report focusing on several important features of the sequence data. FastQC provides an average quality score for each sequence, and an average quality score at each position over all sequences. It calculates the distribution of each base on a sequence and the GC content. FastQC also recognizes fragments of sequences that repeat enough to compose over 0.1% of total library. Based on the results of the FastQC report, it may be necessary to modify the data to improve overall quality. Trimmomatic is a software tool that works well with trimming Illumina sequence data. Trimmomatic provides nine different algorithms that can be used to trim or edit sequence reads with high error rate or that are overrepresented in the library. The overall goal of Trimmomatic is to remove read error rates that are high but also retaining the longest possible sequence.

**Methods and Results:**

This lab offered practice with analyzing and trimming sequence data from the bacteria Vibrio Vulnificus. The FASTQ files were collected from NCBI, in the Short Read Archive (SRA). The accession number for the run is SRR1391072. Module load sra-toolkit is used to perform the fastq-dump command. FASTQ was performed on the forward and reverse reads, SRR1391072\_1 and SRR1391072\_2 respectively. After reviewing the reports, the Trimmomatic was used to edit the sequence reads, and then FASTQC was performed again on the trimmed reads. The lab required two combinations of parameters to be completed. After the second trimming combination, a primer sequence was identified by the FASTQC report. I performed one more trimming command to attempt to remove the primer sequence. The results from each step are provided below.

***Retrieving data from NCBI SRA:***

>>module load sra-toolkit

*>>wget -O ./SRR1391072* *https://trace.ncbi.nlm.nih.gov/Traces/sra/?run=SRR1391072*

*>>fastq-dump –outdir /mnt/c/Users/adamh/OneDrive/Desktop/6203\_Genomics/Labs/Lab1 –split-files SRR1391072*

***FASTQ REPORT:***

*>>FastQC SRR1391072\_1*

*>>FastQC SRR1391072\_2*

The basic statistics provide the general information for the FASTQ report. We can see the total sequence length, the length of each sequence, and average %GC content.

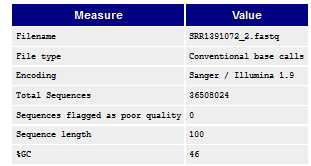
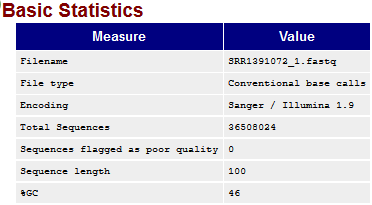


Figure : Basic statistics from the sequence reads of data set.

The per sequence quality score report shows the average quality score among all sequences as a distribution. The graphs below show that most of the sequences have a very high average quality score greater than 36.

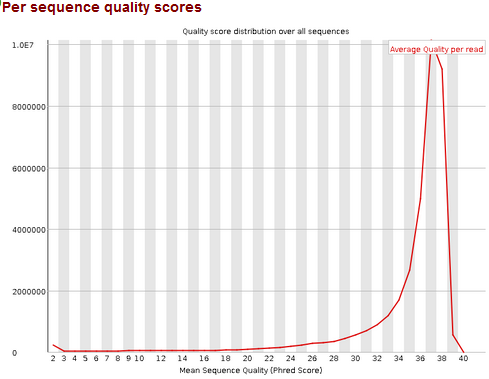
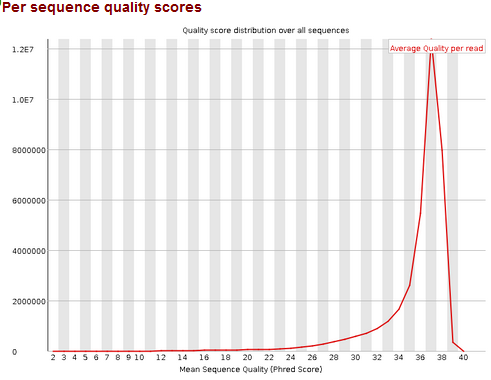


Figure 2: Per sequence quality scores of SRR1391072\_1 (left) and SRR1391072\_2 (right) before trimming.

**CLEANING DATA AND RERUNNING FASTQC REPORT:**

**TRIM 1**: Remove first 10 bases and sliding window of 5, with average quality score of 25 cutoff. The quality score for the initial 10 bases starts lower than most bases downstream. The final 10 bases in SRR1391072\_2 show a much larger distribution of average scores. The sliding window aimed to trim some of the low quality scores downstream.

>>TrimmomaticSE -phred33 SRR1391072\_1.fastq SRR1391072\_1.head.window.fastq HEADCROP:10 SLIDINGWINDOW:5:25

>>FastQC SRR1391072\_1.head.window.fastq

>>TrimmomaticSE -phred33 SRR1391072\_2.fastq SRR1391072\_2.head.window.fastq HEADCROP:10 SLIDINGWINDOW:5:25

>>FastQC SRR1391072\_2.head.window.fastq

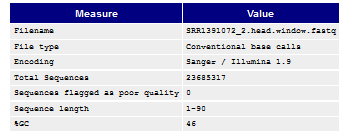
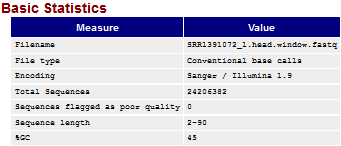


Figure 3: General information for SRR1391072\_1(left) and SRR1391072\_2(right) sequences after trimming.

The figures below show the reads before trimming in the left column and after trimming on the right column. SRR1391072\_1 is represented in the first row, and SRR1391072\_2 in second row. The yellow box represents 25%-75% of the scores that can be found at those averages. The red line is a measure of the median value and blue line is the mean. The black lines provide the boundaries for the 10th and 90th percentiles of scores.

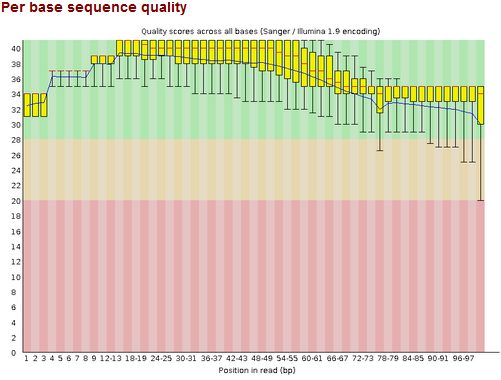
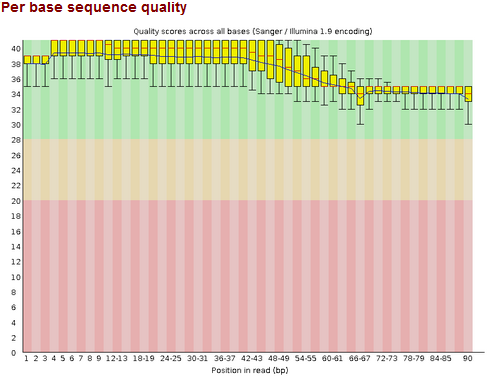
 

Figure 4: SRR1391072\_1 before trimming(left) and after trimming leading 10 bases and 5 base sliding window. (Q-score > 25)

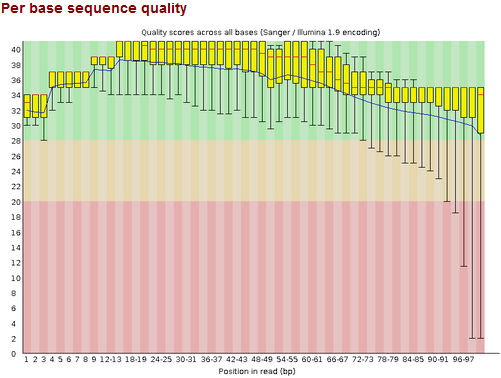
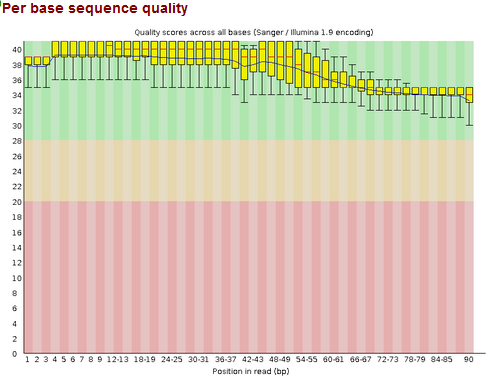
 

Figure 5: SRR1391072\_2 before trimming(left) and after trimming leading 10 bases and 5 base sliding window. (Q-score > 25)

**TRIM 2:** Remove first 8 bases and last 10 bases. More representative in SRR1391072\_2 as the final 10 bases have a quality score drop off shown in the bottom 10% whiskers.

**>>**TrimmomaticSE -phred33 SRR1391072\_1.fastq SRR1391072\_1.head.crop.fastq HEADCROP:8 CROP:82

>>FastQC SRR1391072\_1.head.crop.fastq

**>>**trimmomaticSE -phred33 SRR1391072\_2.fastq SRR1391072\_2.head.crop.fastq HEADCROP:8 CROP:82

>>FastQC SRR1391072\_2.head.crop.fastq

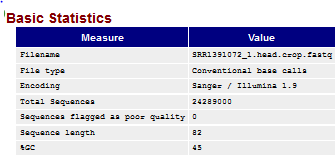
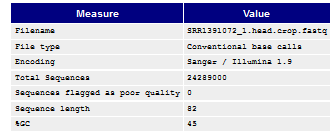
 

Figure 6: General information for SRR1391072\_1(left) and SRR1391072\_2(right) sequences after trimming.

Per base sequence quality shows the average quality score at each base position.

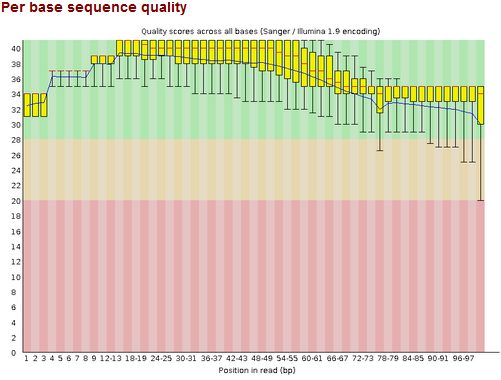
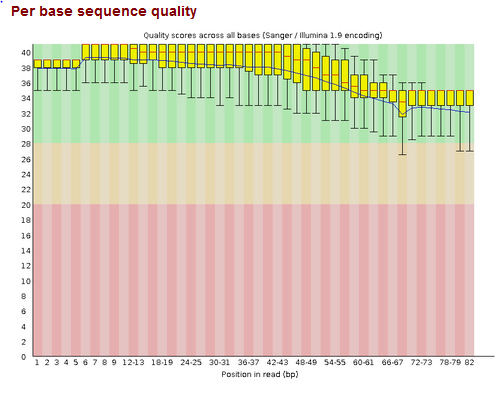
 

Figure 7: SRR1391072\_1 before trimming and after trimming

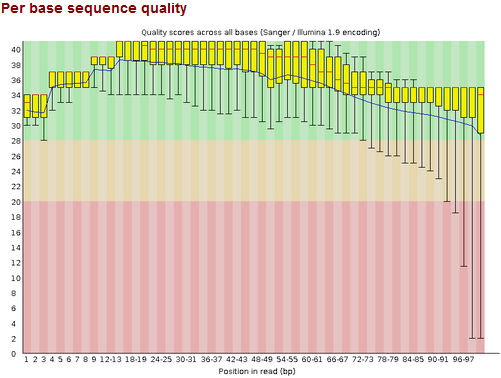
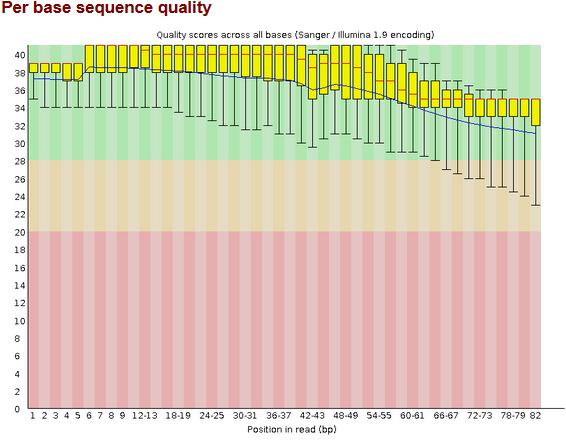
 

Figure 8: SRR1391072\_2 before trimming and after trimming

The FASTQC report had many overrepresented sequences. Only one overrepresented sequence did not have a ‘No-Hit’ classification. Both reads show a possible overrepresented sequence that is identified as a primer used by Illumina. The adapter content shows roughly 5% adapter content at the end of the sequence read. I ran another trim on the trimmed sequence to attempt to remove the adapter sequence.

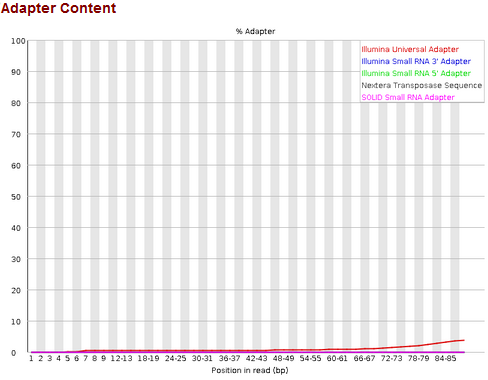
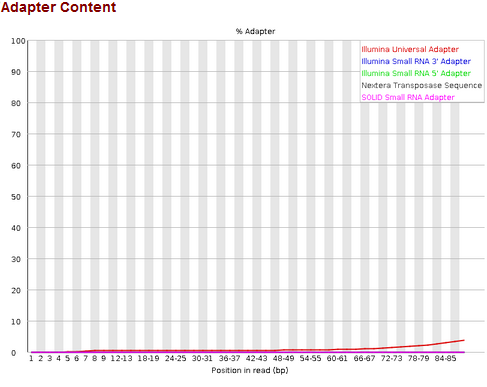
 

Figure 9: Adapter content for SRR1391072\_1(left) and SRR1391072\_2(right)

**TRIM 3: continuation of trim 2**. Remove the adapter sequences.

>>TrimmomaticSE -phred33 ./Labs/Lab1/SRR1391072\_1.head.crop.fastq ./Labs/Lab1/SRR1391072\_1.luminaclip.fastq ILLUMINACLIP:./Trimmomatic-0.39/adapters/TruSeq2-SE.fa:2:30:10 SLIDINGWINDOW:5:30

*Input Reads: 24289000 Surviving: 23868959 (98.27%) Dropped: 420041 (1.73%)*

>>FastQC SRR1391072\_1.luminaclip.fastq

>>TrimmomaticSE -phred33 ./Labs/Lab1/SRR1391072\_2.head.crop.fastq ./Labs/Lab1/SRR1391072\_2.luminaclip.fastq ILLUMINACLIP:./Trimmomatic-0.39/adapters/TruSeq2-SE.fa:2:30:10 SLIDINGWINDOW:5:30

>>FastQC SRR1391072\_2.luminaclip.fastq

*Input Reads: 24289000 Surviving: 23018834 (94.77%) Dropped: 1270166 (5.23%)*

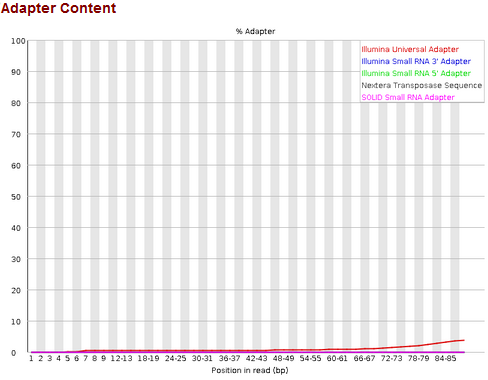


Figure 10: Adapter content for SRR1391072\_1 before Illumina clip

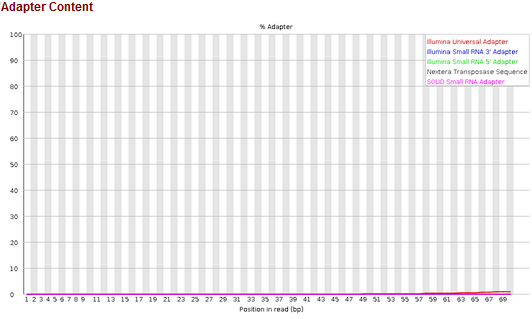


Figure 11: Adapter content for SRR1391072\_1 after Illumina clip

Adapter content of SRR1391072\_2 before the Illumina adapter trim (top) and after clipping the Illumina adapter sequence (bottom).

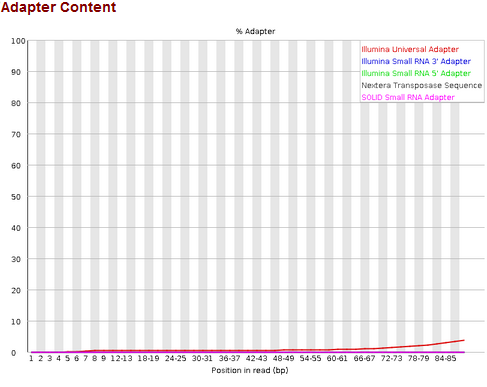


Figure 12: Adapter content for SRR1391072\_2 before Illumina clip

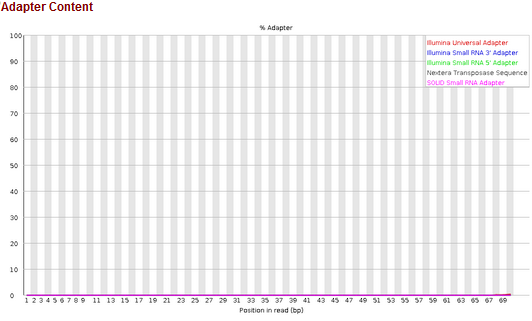
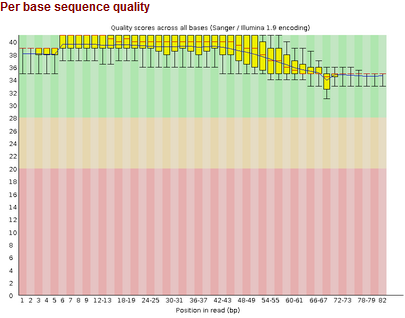
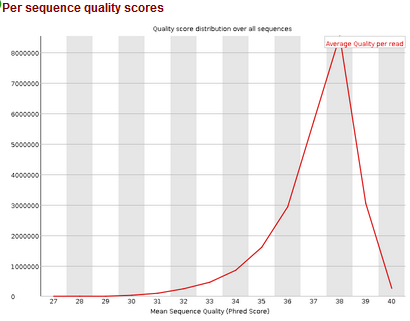


Figure 13: Adapter content of 1391072\_2 after Illumina clip

The FASTQC report for Trim 3. SRR1391072\_1 on the top row and SRR1391072\_2 on bottom row.

2

3 4

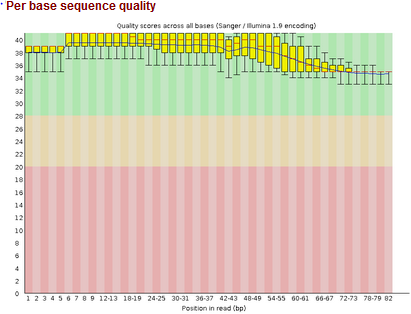
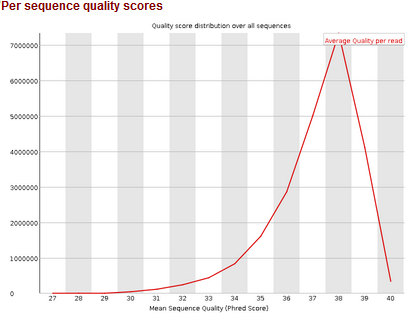
 

Table 1: (1) and (2) represents the reports from SRR1391072\_1. (3) and (4) are the reports from SRR1391072\_2 after Illumina clip.

**Discussion:**

FASTQC is a very useful tool for quickly assessing the quality of the sequence reads and analyzing the composition of the sequence. It does not guarantee the quality of the reads, but it can alert you to take a closer look at areas that could be problematic. It is important to know what kind of data you are looking at as well. For example, the sequences we examined are overall high quality and did not require much additional trimming. However, there are a considerable amount of overrepresented sequences that could cause one to excessively trim their data. This could be representative of adapter or primer technical sequences being duplicated, or repeating elements in the sequence that could provide error. We are using RNA-seq data, so gene expression can play a major role in visualizing overrepresented sequences, which is normal from RNA sequence data.

I used the per base sequence quality report as my primary source for information regarding the sequence quality. It provides the most intuitive view of sequence quality, and I used those scores to develop a trimming strategy. Some error is per base quality is expected, especially at the beginning of the reads. When there is a small sample, one error can have a high significance before regressing to the norm. Chemistry errors, like the technical sequences of adapters and primers can be found in the beginning of sequences if they are not adequately trimmed from the Illumina software. These factors were considered when making the trim commands.

The first trim benefitted SRR1391072\_2 more than the forward read. The range of average scores in the bottom 10 percentile of the reverse read increases from 2 to 24 compared to an increase from 21 to 31 in the forward read. Another positive result is the blue line, which represents the mean average score, is closer to the median score, denoted by the red bar, on both reads after trimming. This indicates a more even distribution of quality scores at each base position. I thought the beginning 10 positions had an interesting pattern, where it increases from around 30 to 40, before gradually decreasing downstream. I trimmed the beginning of the sequence off just in case it was residual bases from the adapters. I used the sliding window with a small window from of 5 bases and quality score cutoff of 25. I believed this could remove drastically low-quality reads that cause the small window average to drop below 25.

After observing the beginning sequences and the ending of the sequences I tried a different approach to improving the quality by focusing on sequence length. I removed the initial 8 bases similarly to the first trim, but then cropped the reads after position 90. This essentially only scores the reads from position 8 to position 90 and produces a read length of 82, which is found in the Basic Statistics figure 6. We see a much more noticeable change in the per base sequence quality report on the reverse read compared to forward read, but both have an overall high quality. There are many unrecognized overrepresented sequences, but the FASTQ report did identify a possible source, or hit, for one overrepresented sequence. SRR1391072\_1 reported the following overrepresented sequence:

*ATCGGAAGAGCACACGTCTGAACTCCAGTCACACAGTGATCTCGTATGCCG*

The sequence is identified 46,471 times and composed 0.169% of the total sequence reads. The possible source is identified as TruSeq Adapter Index 5. The TruSeq Adapter Index 5 was located again 25,675 times but with the leading A removed. Additionally, SRR1391072\_2 reported the following overrepresented sequence:

*ATCGGAAGAGCACACGTCTGAACTCCAGTCACACAGTGATCTCGTATGCC.*

The sequence is identified 41,187 times and composed 0.169% of the total sequence reads. The possible source is identified as Illumina Single End PCR Primer 1. From the trimmed sequence data, I ran a third trim step, using the Illumina clip command to attempt to the remove the possible adapter. Figure 10 and 11 shows the adapter content on the forward read before the Illumina clip step and after clipping the adapter sequence respectively. Figures 12 and 13 show the before and after on the reverse read. You can see that after clipping, the adapter content for both reads reduces to nearly 0%. The FASTQC report from trim #3 had an interesting finding in SRR1391072\_1. It reported the following overrepresented sequence:

*ATCGGAAGAGCACACGTCTGAACTCCAG.*

The sequence is identified 39,929 times and composed 0.167% of the total sequence reads. The possible source is identified as Illumina Multiplexing PCR Primer 2.01. The sequence is very similar to the overrepresented sequences found in previous step, but just a shortened version. SRR1391072\_2 had no reported technical sequences recognized as an overrepresented sequence but did have many overrepresented sequences that are labeled as ‘No Hit’. Without more extensive knowledge of the sequence, I would presume the overrepresented sequences are genes being highly expressed. In table 1, (2) and (4) show the pers equence quality scores of the forward and reverse reads respectively. After trimming the sequence and removing the potential adapter sequence, we can see the distribution of quality scores is wider than the initial FASTQC report before trimming. While the quality scores are still quite high, the average scores vary more than before doing any trimming.

References

BabrahamBioinf. Sept. 22, 2010. Using FastQC to check the quality of high throughput sequence. https://www.youtube.com/watch?v=bz93ReOv87Y

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)*, *30*(15), 2114–2120. doi:10.1093/bioinformatics/btu170