Erika Maldonado-Rosado

29 January 2019

Sequence Read QC

BINF 6203

**Introduction**

In this lab we looked at a tomato chloroplast DNA sequence and had to create reads using FastQC with the help of a java based package called Trimmomatic. We used FastQC to look at the sequences and Trimmomatic to trim the reads themselves. We obtained the sequences from NCBI where Dr. Jennifer Weller and University of North Carolina at Charlotte has made available for the public and student needs.

**Method**

In order to obtain the SRR data onto ones computer one needs Linux or IOS system.I used Ubuntu 18.04 LTS and I downloaded Brew, Java, SRA toolkit, and Trimmomatic. Once I got the fastq files I moved them from Ubuntu to my windows in order to view it on FastQC. Java is used to for Trimmomatic. Brew kit is used to get the SRA toolkit, and FastQC can be installed on your window or apple device. Once one gets the data that they want you need to download it using fast dump. Once on there one needs to view the file on FastQC to see what needs to be trimmed. Once you see what needs to be trimmed one uses Trimmomatic and once again views it on FastQC. On Trimmomatic you can do single end trimming and paired end trimming. Single end trimming can be used for one file and paired end can be used for one file that has been split or two files that are connected.

**Results**

With this assignment I noticed that when I increased the minlen increased I lost more base pairs and when I made it small I only got a small amount. When increasing the trail I noticed that I lost more of the base pairs that had better quality score which removed base pairs that I needed. In general one must be wary of the threshold and to only remove what is needed.

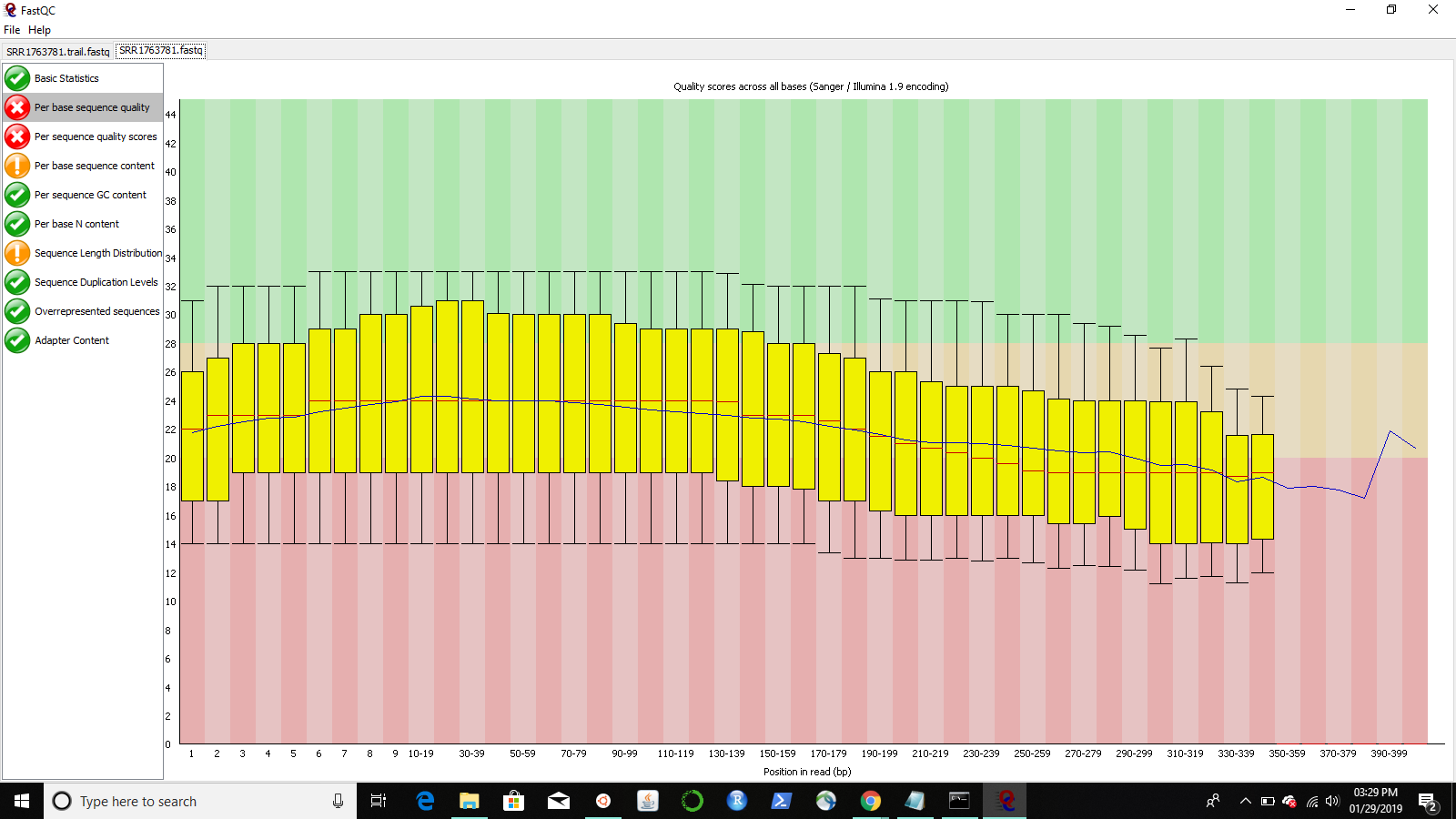
In the first picture is what the original file looked like. The second picture was the trailing at 20 and minlen at 50. The third picture I did a trailing of 20 and a minlen at 50, the image shows a spike in the quality score of the base pairs, I believe that this shows that I trimmed the data too much. Lastly, on the fourth picture I did a leading of 20, trailing of 20, and I used crop 300 to keep the first 300 base pair reads. I Believe that this image is more consistent to what I want my sequences to look like.

**Code**

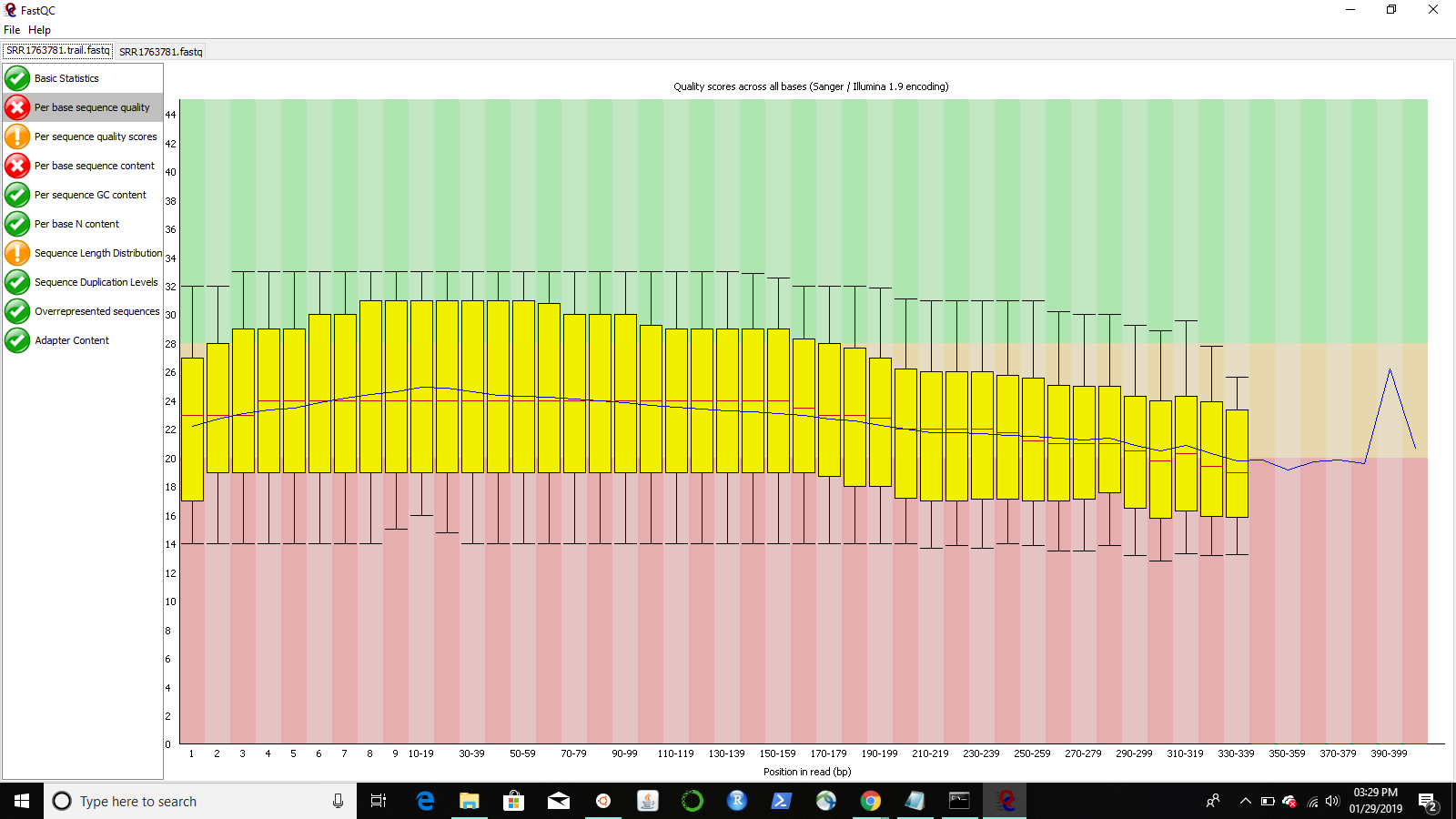
fastq-dump -A SRR1763781

fastq-dump -X 5 -Z SRR1763781

cp SRR1763781.fastq /mnt/c/Users/erika/Onedrive/Desktop



java -jar ~/Trimmomatic-0.36/trimmomatic-0.36.jar SE -threads 3 SRRCrop.fastq SRR1763781.trail.fastq TRAILING:20 MINLEN:50

cp SRR1763781.trail.fastq /mnt/c/Users/erika/Onedrive/Desktop

java -jar ~/Trimmomatic-0.36/trimmomatic-0.36.jar SE -threads 3 SRRCrop.fastq SRR.t.fastq TRAILING:20 MINLEN:150

cp SRR.t.fastq /mnt/c/Users/erika/OneDrive/Desktop/



java -jar ~/Trimmomatic-0.36/trimmomatic-0.36.jar SE -threads 3 SRRCrop.fastq SRR.3.fastq LEADING:20 TRAILING:20 CROP:300

cp SRR.3.fastq /mnt/c/Users/erika/OneDrive/Desktop/

