**Project #3**

**Objective**

To combine what you’ve learned in High Performance Computing (HPC) and software installation.

**Due**

Friday May 3th at 5pm PST.

**Instructions**

Perform the tasks listed below and when instructed, provide responses in the text boxes provided. When finished, send the completed document to Dr. Ficklin. This is not a group project so please do not work with other students.

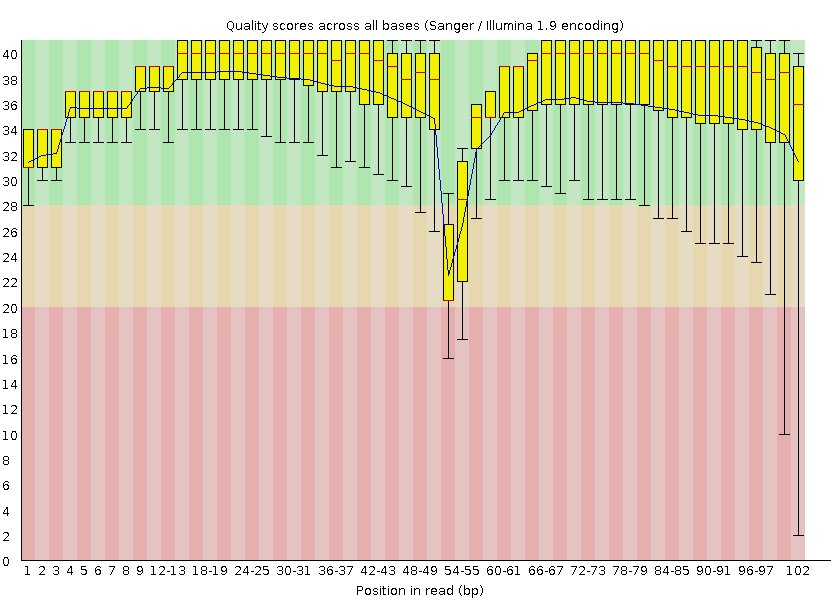
**Background**

Like Project 1, the tool Trimmomatic will remove low quality ends from the beginning and end of DNA sequence reads in FASTQ files. However, it does a lot more. It can use a sliding window to average quality amongst a set of bases to determine where to trim, and it will also remove adaptor sequences from pre-sequencing preparation (library construction) that get added to the ends of the DNA being sequenced. You can learn about it here: <http://www.usadellab.org/cms/?page=trimmomatic>

**Tasks**

**Task A: Get FastQC Results (10 points)**

In class on Thursday, April 27th, together we created a SLURM script that ran the **fastqc** program. The script processed the FASTQ file stored here: /data/hort503/example-data/SRR2931286.fastq. If you did not submit that job, please do so now. Once the job is completed, find the output file created by the job you submitted. It will have a .html file extension. Copy that file to your workstation, open it using a web browser and find the figure titled “Per base sequence quality”. Copy that image from the web page and place it in the box below:



**Task B: Install Trimmomatic (30 points)**

For this task you will download and install trimmomatic from the software’s website: <http://www.usadellab.org/cms/?page=trimmomatic>. Be sure to follow the directory structure discussed in class on Thursday to install the software:

* 1. Store downloaded software in the directory:  
     /data/hort503/$USER/software/src
  2. Install trimmomatic into the directory:   
     /data/hort503/$USER/software/Trimmomatic-0.39

1. Download the 0.39 binary version of Trimmomatic. Cut-and-paste the command-line used to download the software into the box below **(5 points)**:

[pragya.asthana@cn73 src]$ wget 'http://www.usadellab.org/cms/uploads/supplementary/Trimmomatic/Trimmomatic-0.39.zip'

--2019-04-30 10:45:25-- http://www.usadellab.org/cms/uploads/supplementary/Trimmomatic/Trimmomatic-0.39.zip

Resolving www.usadellab.org (www.usadellab.org)... 199.195.142.183

Connecting to www.usadellab.org (www.usadellab.org)|199.195.142.183|:80... connected.

HTTP request sent, awaiting response... 200 OK

Length: 133596 (130K) [application/zip]

Saving to: ‘Trimmomatic-0.39.zip’

100%[=====================================>] 133,596 399KB/s in 0.3s

2019-04-30 10:45:26 (399 KB/s) - ‘Trimmomatic-0.39.zip’ saved [133596/133596]

[pragya.asthana@cn73 src]$ ls

Trimmomatic-0.39.zip

1. Install Trimmomatic. Trimmomatic is compressed as a ZIP file rather than with TAR + Gzip fire. You can decompress it using this commend:  
     
   unzip Trimmomatic-0.39.zip   
     
   Cut-and-paste the commands you used to install the software in the box below **(15 points)**:

[pragya.asthana@sn13 ~]$ cd /data/hort503/pragya/software/Trimmomantic

[pragya.asthana@sn13 Trimmomantic]$ unzip ../src/Trimmomatic-0.39.zip

1. Demonstrate that you have the software install in the correct directory by running these commands **(5 points)**.   
     
   cd /data/hort503/$USER/software  
   tree -L 2  
     
   Enter the results of the command above into the box below:

├── 0.39

├── bin

├── include

│   └── impl\_sse.h

├── lib

├── share

│   ├── doc

│   └── man

├── src

│   └── Trimmomatic-0.39.zip

└── Trimmomantic

└── Trimmomatic-0.39

10 directories, 2 files

1. Test running the program by executing this command in the Trimmomatic directory **(5 points)**:  
     
   java -jar trimmomatic-0.39.jar -version

What do you think the above command does?

It would check the version of the Trimmomatic program just downloaded.

Enter the results of the command above into the box below:

[pragya.asthana@sn13 Trimmomatic-0.39]$ java -jar trimmomatic-0.39.jar -version

0.39

**Task C: Run Trimmomatic (30 points)**

It was clear from the FastQC report generated in Task A that the original sequence data (named SRR2931286.fastq) had some low-quality regions. We can try to clean the data by using Trimmomatic. To do this, create a SLURM script that uses the new version of Trimmomatic that you just installed. Because Trimmomatic may be new to you the exact command to execute is provided below.

export CLASSPATH=$CLASSPATH:/data/hort503/$USER/software/Trimmomatic-0.39/trimmomatic-0.39.jar

java -Xmx512m org.usadellab.trimmomatic.Trimmomatic SE -threads 10 -phred33 /data/hort503/example-data/SRR2931286.fastq SRR2931286\_1.trim.fastq ILLUMINACLIP:/data/hort503/$USER/software/Trimmomatic-0.39/adapters/TruSeq3-PE.fa:2:40:15 LEADING:3 TRAILING:6 SLIDINGWINDOW:4:15 MINLEN:50

The first command alters a JAVA specific environment variable. The CLASSPATH variable is used by JAVA to find **.jar** files. This way we don’t have to be in the Trimmomatic installation directory to run it. The second command will cause Trimmomatic to read the SRR2931286\_1.fastq file, clean up low-quality ends and create a new file named SRR2931286\_1.trim.fastq. This new file will be created in the directory where you run Trimmomatic. Write a SLURM script that runs the commands above. Be sure to request the following resources:

1. 10 threads
2. 6 GB of RAM
3. 1 hour for execution
4. Set the submission script to email you when the job starts/completes/fails.

Cut-and-paste your SLURM script below **(20 points)**:

[pragya.asthana@login-p1n02 pragya]$ cat slurm\_project3

#!/bin/bash

#SBATCH --partition=hort503

#SBATCH --account=hort503

#SBATCH --job-name=trimmomatic\_trim

#SBATCH --time=00:01:00

#SBATCH --nodes=1

#SBATCH --ntasks-per-node=1

#SBATCH --cpus-per-task=10

#SBATCH --output=trimmomatic\_trim.out

#SBATCH --mem=6000

#SBATCH --mail-user=pragya.asthana@wsu.edu

export CLASSPATH=$CLASSPATH:/data/hort503/pragya/software/Trimmomantic/Trimmomatic-0.39/trimmomatic-0.39.jar

java -Xmx512m org.usadellab.trimmomatic.Trimmomatic SE -threads 10 -phred33 /data/hort503/example-data/SRR2931286.fastq SRR2931286\_1.trim.fastq ILLUMINACLIP:/data/hort503/$USER/software/Trimmomatic-0.39/adapters/TruSeq3-PE.fa:2:40:15 LEADING:3 TRAILING:6 SLIDINGWINDOW:4:15 MINLEN:50

Cut-and-paste the command you used to launch the job **(5 points)**

sbatch slurm\_project3

Cut-and-paste the contents of the output file from the job in the box below **(5 points)**:

TrimmomaticSE: Started with arguments:

-threads 10 -phred33 /data/hort503/example-data/SRR2931286.fastq SRR2931286\_1.trim.fastq ILLUMINACLIP:/data/hort503/pragya/software/Trimmomatic-0.39/adapters/TruSeq3-PE.fa:2:40:15 LEADING:3 TRAILING:6 SLIDINGWINDOW:4:15 MINLEN:50

Using PrefixPair: 'TACACTCTTTCCCTACACGACGCTCTTCCGATCT' and 'GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT'

ILLUMINACLIP: Using 1 prefix pairs, 0 forward/reverse sequences, 0 forward only sequences, 0 reverse only sequences

Input Reads: 45844996 Surviving: 42788357 (93.33%) Dropped: 3056639 (6.67%)

TrimmomaticSE: Completed successfully

**Task D: Run FASTQC again (30 points)**

Now that Trimmomatic has completed, we want to determine if trimming improved the quality of the dataset. To do this. Create a new SLURM script that runs the FastQC program on the newly created SRR2931286\_1.trim.fastq file. Cut-and-paste your SLURM script into the box below **(20 points)**:

#!/bin/bash

#SBATCH --partition=hort503

#SBATCH --account=hort503

#SBATCH --job-name=trimmed\_fastqc

#SBATCH --time=01:00:00

#SBATCH --nodes=1

#SBATCH --ntasks-per-node=1

#SBATCH --cpus-per-task=10

#SBATCH --output=trimmed\_fastqc.out

#SBATCH --mem=6000

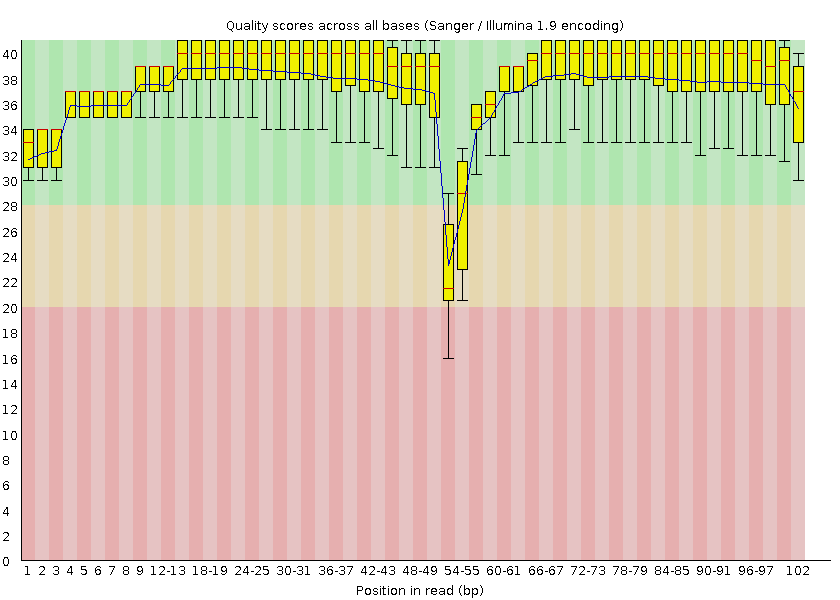
#SBATCH --mail-user=pragya.asthana@wsu.edu

module use /data/ficklin/modulefiles

module load fastqc/0.11.8

fastqc -o ./ -t 5 /data/hort503/pragya/SRR2931286\_1.trim.fastqqq

Finally, copy the resulting output file with the **.html** extension to your workstation. Open the file with a web browser, find the figure titled “Per base sequence quality” and copy that image from the web page and place it in the box below **(10 points)**:



Do you see any improvement?

Yes, I do see a good improvement in the quality of the trimmed reads. By looking at the new boxplot we could figure out that the range of the standard error bars have reduced significantly along with reduction in the size of the boxes as well. All the box plots have shifted upwards in the green belt.

For the 54-55 base pairs, even after trimming the quality did not improve much probably because there was a lot of mess picked up while sequencing for this position which did not allow it to improve a lot after trimming as well.

All these observations signify that the trimmed reads, overall, are all of better quality than the original ones.