**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.

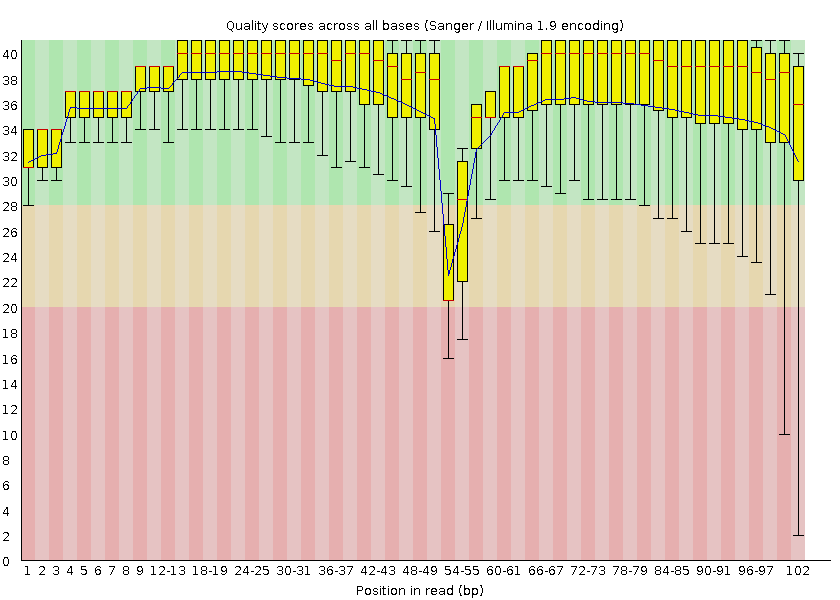
**Background**

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)**:

[richard.manasseh@login-p1n01 src]$ wget "http://www.usadellab.org/cms/uploads/supplementary/Trimmomatic/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)**:

[richard.manasseh@login-p1n01 src]$ unzip Trimmomatic-0.39.zip

[richard.manasseh@login-p1n01 src]$ mv Trimmomatic-0.39 ../Trimmonatic-0.39

1. Demonstrate that you have the software install in the correct directory by running these commands **(5 points)**.   
     
   cd /data/hort503/$USER/software  
   tree -L 2  
   [richard.manasseh@login-p1n01 software]$ tree -L 2:
2. Enter the results of the command above into the box below:

├── bin

│   ├── alimask

│   ├── hmmalign

│   ├── hmmbuild

│   ├── hmmconvert

│   ├── hmmemit

│   ├── hmmfetch

│   ├── hmmlogo

│   ├── hmmpgmd

│   ├── hmmpress

│   ├── hmmscan

│   ├── hmmsearch

│   ├── hmmsim

│   ├── hmmstat

│   ├── jackhmmer

│   ├── makehmmerdb

│   ├── muscle

│   ├── nhmmer

│   ├── nhmmscan

│   └── phmmer

├── class\_slurm

├── hmmer

├── hmmer-3.2

│   ├── config.guess

│   ├── config.log

│   ├── config.status

│   ├── config.sub

│   ├── configure

│   ├── configure.ac

│   ├── documentation

│   ├── easel

│   ├── INSTALL

│   ├── install-sh

│   ├── libdivsufsort

│   ├── LICENSE

│   ├── Makefile

│   ├── Makefile.in

│   ├── makeTAGS.sh

│   ├── profmark

│   ├── README.md

│   ├── RELEASE-3.2

│   ├── src

│   ├── test-speed

│   ├── testsuite

│   ├── tutorial

│   └── Userguide.pdf

├── muscle

├── muscle3.8.31\_i86darwin64

├── muscle3.8.31\_i86darwin64.tar

├── muscle3.8.31\_i86linux64

├── muscle3.8.31\_i86linux64.tar

├── share

│   └── man

├── sickle

│   ├── LICENSE

│   ├── Makefile

│   ├── print\_record.o

│   ├── README.md

│   ├── sickle

│   ├── sickle.o

│   ├── sickle.xml

│   ├── sliding.o

│   ├── src

│   ├── test

│   ├── trim\_paired.o

│   └── trim\_single.o

├── sickle-1.33

│   └── sickle

├── src

│   ├── sickle

│   └── Trimmomatic-0.39.zip

└── Trimmonatic-0.39

├── adapters

├── LICENSE

└── trimmomatic-0.39.jar

20 directories, 55 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?

java indicates we’re running a Java program.

-jar is an option for specifying the Java program we want to run, which is trimmomatic-0.39.jar

-version returns the version of the Java program (trimmomatic-0.39.jar)

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

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

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

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)**:

#!/bin/bash

#SBATCH --mail-user=richard.manasseh@wsu.edu

#SBATCH --mail-type=ALL

#SBATCH --partition=hort503

#SBATCH --account=hort503

#SBATCH --job-name=Trimmomatic

#SBATCH --time=00:60:00

#SBATCH --mem=6000

#SBATCH --nodes=1

#SBATCH --ntasks-per-node=1

#SBATCH --cpus-per-task=5

#SBATCH --output="trim.out"

export CLASSPATH=$CLASSPATH:/data/hort503/richard.manasseh/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/richard.manasseh/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)**

[richard.manasseh@login-p1n02 software]$ sbatch trimmomatic.sh

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

[richard.manasseh@login-p1n02 software]$ cat trim.out

TrimmomaticSE: Started with arguments:

-threads 10 -phred33 /data/hort503/example-data/SRR2931286.fastq SRR2931286\_1.trim.fastq ILLUMINACLIP:/data/hort503/richard.manasseh/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 --mail-user=richard.manasseh@wsu.edu

#SBATCH --mail-type=ALL

#SBATCH --partition=hort503

#SBATCH --account=hort503

#SBATCH --job-name=fastaqc

#SBATCH --time=00:60:00

#SBATCH --mem=4000

#SBATCH --nodes=1

#SBATCH --ntasks-per-node=1

#SBATCH --cpus-per-task=5

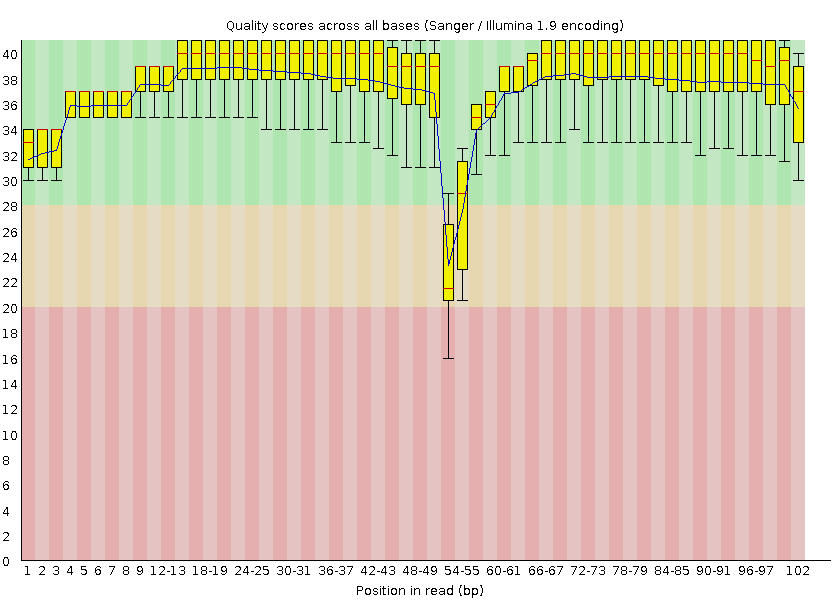
#SBATCH --output="fastqc.out"

module use /data/ficklin/modulefiles

module load FastQC/0.11.8

fastqc -o /data/hort503/richard.manasseh -t 5 /data/hort503/richard.manasseh/software/SRR2931286\_1.trim.fastq

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?

|  |  |
| --- | --- |
| C:\Users\rmann\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\B0CE60AC.tmp | C:\Users\rmann\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\3442BED3.tmp |

Yes, there was improvement. The mean quality scores (blue lines) and the 90th percentile scores (represented by the lower whiskers) rose dramatically for bins from 78-79 to position 102.