Day 4 Worksheet – Trimmomatic

Author: Jessica Westfall 2021
Edited: Ariel Eraso, Lynn Sanford 2022-23

Introduction: Now that we have evaluated our sequence library initially to determine if the libraries are worth analyzing, we will do some "cleaning up" by trimming unwanted sequences such as adapter sequences. This step is necessary for improved alignment and mapping to the reference genome downstream. Once trimming is completed, we will reevaluate our trimmed files with FastQC for quality to decide if we will move forward with mapping.

❗ Note: The directory and username used in the screenshot will be for my working directory and username and will be different than yours.

Make working directories

Yesterday, we made working directories for running fastQC. Repeat the same process, but this time we will make a directory for trimmomatic.

1. Use command **pwd** to determine what directory you are in and if necessary, **cd** to the directory that you want to place your new trimmomatic directory in.
2. Make a few new directories using the **mkdir** command. Use command **ls -lsh** to confirm the folders are present.

```
[lynn-sanford@ip-172-31-29-36 ~]$  cd /scratch/Users/lynn-sanford/
[lynn-sanford@ip-172-31-29-36 lynn-sanford]$  mkdir day4
[lynn-sanford@ip-172-31-29-36 lynn-sanford]$  mkdir day4/scripts
[lynn-sanford@ip-172-31-29-36 lynn-sanford]$  mkdir day4/eofiles
[lynn-sanford@ip-172-31-29-36 lynn-sanford]$  mkdir day4/trimmomatic
[lynn-sanford@ip-172-31-29-36 lynn-sanford]$  cd day4/
[lynn-sanford@ip-172-31-29-36 day4]$  ls -lsh
total 12K
4.0K drwxrwxr-x 2 lynn-sanford lynn-sanford 6.0K Jul 17 12:19 eofiles
4.0K drwxrwxr-x 2 lynn-sanford lynn-sanford 6.0K Jul 17 12:19 scripts
4.0K drwxrwxr-x 2 lynn-sanford lynn-sanford 6.0K Jul 17 12:19 trimmomatic
```

Trimmomatic

3. Git pull within the sr2023 repo. Then copy (**rsync**) the **d4_trim_qc.sbatch** script from the day04/scripts/ into your script directory. Use **ls -lsh** to confirm the file is present in the directory. You can **ls** with an absolute path as well as relative path.

To copy the script, the command syntax is **rsync <input> <output>**

```
[lynn-sanford@ip-172-31-29-36 day4] $ cd ~/sr2023/
[lynn-sanford@ip-172-31-29-36 sr2023] $ git pull
Already up-to-date.
[lynn-sanford@ip-172-31-29-36 sr2023] $ cd /scratch/Users/lynn-sanford/day4/
[lynn-sanford@ip-172-31-29-36 day4] $ rsync ~/sr2023/day04/scripts/d4_trim_qc.sbatch ./scripts/
[lynn-sanford@ip-172-31-29-36 day4] $ ls scripts/
d4_trim_qc.sbatch
```

4. Edit the sbatch script by using **vim <sbatch>** to open a text editor on your sbatch script. Type **i** to toggle into edit/insert mode. Similar to the previous exercise you will need to change the job name, user email, and the standard output and error log directories. Change the **–job-name=<JOB_NAME>** to a name related to the job you will be running, for example 'trim_qc'. Additionally you will want to change the **–mail user=<YOUR_EMAIL>** to your email, as well as the path to your eofiles directory for the standard output (**--output)** and error log (**--error)**. The **%x** will be replace by your **-job name** and the **%j** will be replace by the job id that will be assigned by the job manager when you run your sbatch script.

```
#!/bin/bash
#SBATCH --job-name=<JOB_NAME>                    # Job name
#SBATCH --mail-type=ALL                          # Mail events (NONE,
#SBATCH --mail-user=<YOUR_EMAIL>                 # Where to send mail
#SBATCH --nodes=1                                # Number of nodes re
#SBATCH --ntasks=8                               # Number of CPUs (pr
#SBATCH --mem=8gb                                # Memory limit
#SBATCH --time=01:30:00                          # Time limit hrs:min
#SBATCH --partition=short                        # Partition/queue re
#SBATCH --output=/scratch/Users/<USERNAME>/day4/eofiles/%x.%j.out
#SBATCH --error=/scratch/Users/<USERNAME>/day4/eofiles/%x.%j.err
```

For this script, I will be change my CPU and nodes for trimmomatic which can use multiple processors per input file. I am going to request 1 node, 8 tasks, 8gb of memory and 90 minutes of wall time.

5. Assigning path variables will make your scripts easier to read. In addition, this makes it easier to reference to a given path and utilize it in your scripts. For the **INDIR=**change the path to where the data files directories are located and specifically the fastq data. For the **OUTDIR=**, point to the appropriate output file directories for our fastQC and trimmed fastq files. I also use the command **mkdir -p** just in case for my output directories.

```
################ ASSIGNS PATH VARIABLES ############################
## the fastq files will be used as input to fastqc and trimmomatic
## trimmed reads will then be passed on to the mapping step

INDIR=/scratch/Shares/public/sread2023/data_files/day4
FASTQ=${INDIR}/fastq

OUTDIR=/scratch/Users/<USERNAME>/day4
FASTQC=${OUTDIR}/fastqc
TRIM=${OUTDIR}/trimmomatic

FILENAME=chr21Eric_repA

mkdir -p ${OUTDIR}
mkdir -p ${FASTQC}
mkdir -p ${TRIM}
```

6. Load the require modules for running this pipeline. We will be using fastQC and the trimming program trimmomatic. Similar to fastqc, if you are not sure which version of the program is available on the cluster you can use the command **module spider <string>** to find the available versions.

```
-bash-4.2$ module spider trimmomatic

----------------------------------------------------------------------------
  trimmomatic: trimmomatic/0.36
----------------------------------------------------------------------------
    Description:
      No Description Given

    This module can be loaded directly: module load trimmomatic/0.36
```

Now I can add the appropriate versions for the modules I want to load in the pipeline.

```
################ LOAD REQUIRED MODULES ####
module load fastqc/0.11.5
module load trimmomatic/0.36
```

7. For the meat of the script, we will be running 3 steps in the pipeline. (1) To run fastQC on the sample, (2) trim the fastQC and (3) reevaluate the quality of the trimmed fastq with fastQC.

```
##############################################################################################
################## RUN PIPELINE ##############################################################
##1: Run fastqc on the samples (here run on example file ${FILENAME}.RNA.end1.fastq)
fastqc ${FASTQ}/${FILENAME}.RNA.end1.fastq -o ${FASTQC}
fastqc ${FASTQ}/${FILENAME}.RNA.end2.fastq -o ${FASTQC}

##2: Trim FASTQ Files


java -jar /opt/trimmomatic/0.36/trimmomatic-0.36.jar PE \
 -threads 8 \
 -phred33 \
 -trimlog ${TRIM}/trimlog \
 ${FASTQ}/${FILENAME}.RNA.end1.fastq ${FASTQ}/${FILENAME}.RNA.end2.fastq \
 ${TRIM}/${FILENAME}.RNA.end1.trimmed.fastq ${TRIM}/${FILENAME}.RNA.end1.unpaired.fastq \
 ${TRIM}/${FILENAME}.RNA.end2.trimmed.fastq ${TRIM}/${FILENAME}.RNA.end2.unpaired.fastq \
 ILLUMINACLIP:/opt/trimmomatic/0.36/adapters/TruSeq3-PE.fa:2:30:10 \
 CROP:20

##3: Check Post-Trimming QC stats
fastqc ${TRIM}/*.trimmed.fastq -o ${FASTQC}

echo Job finished at `date +"%T %a %d %b %Y"`
```

In this script we are running paired end reads. Trimmomatic can be used on both single end or paired-end reads. When setting your parameters use the appropriate adapters.

Below are the syntaxes needed to run trimmomatic:

Illuminaclip parameter (see below for quick reference to trimming)

ILLUMINACLIP:<path_adapters_fasta>:<seed_mismatches>:
<palindrome_clip_threshold>:<simple_clip_threshold> LEADING:<quality>
TRAILING:<quality> SLIDINGWINDOW:<window_size>:<required_quality>
MINLEN:<length>

For single-end reads

java jar /opt/trimmomatic/0.36/trimmomatic-0.36.jar SE [-threads <n>]
[-phred33 | -phred64] [-trimlog <output_trimlog>] <input_file>
<output_file> ILLUMINACLIP

For pair-end reads

java jar /opt/trimmomatic/0.36/trimmomatic-0.36.jar PE [-threads <n>]
[-phred33 | -phred64] [-trimlog <output_trimlog>] <input_file1>
<input_file2> <output_fileP1> <output_fileU1> <output_fileP2>
<output_fileU2> ILLUMINACLIP

Recall that the '\' at the end is used to break the code up for clarity purpose. We can write this syntax as a single line but it is harder to read. If '\' does not change color as you see above (it may not be yellow, but it should change), you may have an extra space after the '\'. Remove that space or your code will not run properly.

8. Save your sbatch script. Press **esc** to exit out of edit mode, then type **:wq**. This will write/save (w) and quit (q) the script.

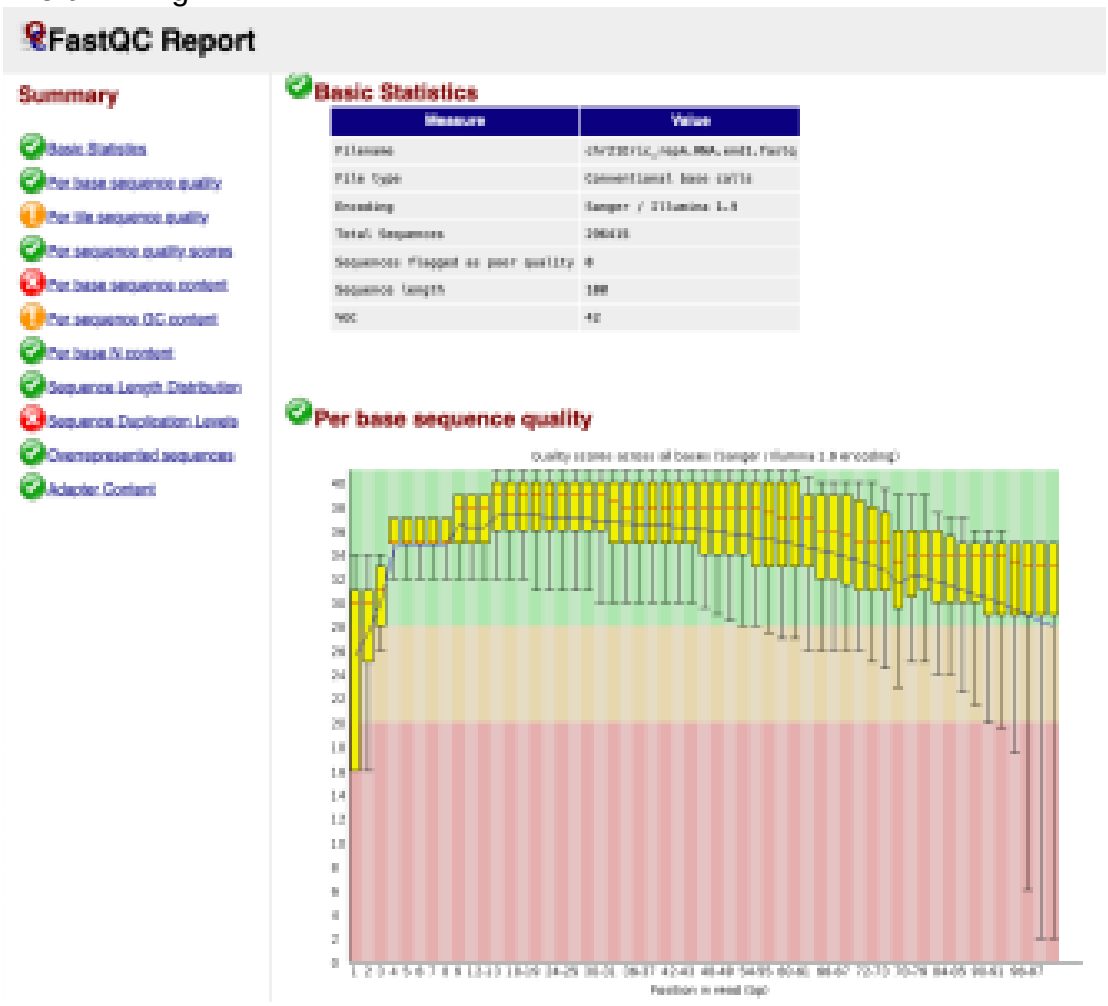
9. Let's run the script. Submit the job to the job manager SLURM using the command **sbatch <sbatch_file>**. The job manager will assign a job id to your run. 12. This pipeline has more tasks than the previous worksheet, so you will want to check the status of your job using the command **squeue -u <username>** to see if the job is running (R) or completed (C). If there are any errors, often time these are just typos in your scripts, you will want to access your error log to make necessary corrections. I will **ls -lahtr /path/to/eofiles** to get the name of the error log for the job id so that I can view it using **more**, **less**, or **cat**. I use **-tr** with the **ls** command to get order my files based on time so I can quickly find the latest error log.

10. Check the error log to find information about the fastqc and trimming job.

```
Approx 95% complete for chr21Eric_repA.RNA.end2.fastq
TrimmomaticPE: Started with arguments:
 -threads 8 -phred33 -trimlog /scratch/Users/jewe1055/sread//trimmomatic/trimlog /scratch/Shares/
dowell/sread/data_files/day4/fastq/chr21Eric_repA.RNA.end1.fastq /scratch/Shares/dowell/sread/dat
a_files/day4/fastq/chr21Eric_repA.RNA.end2.fastq /scratch/Users/jewe1055/sread//trimmomatic/chr21
Eric_repA.RNA.end1.trimmed.fastq /scratch/Users/jewe1055/sread//trimmomatic/chr21Eric_repA.RNA.en
d1.unpaired.fastq /scratch/Users/jewe1055/sread//trimmomatic/chr21Eric_repA.RNA.end2.trimmed.fast
q /scratch/Users/jewe1055/sread//trimmomatic/chr21Eric_repA.RNA.end2.unpaired.fastq ILLUMINACLIP:
/opt/trimmomatic/0.36/adapters/TruSeq3-PE.fa:2:30:10 CROP:20
Using PrefixPair: 'TACACTCTTTCCCTACACGACGCTCTTCCGATCT' and 'GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT'
ILLUMINACLIP: Using 1 prefix pairs, 0 forward/reverse sequences, 0 forward only sequences, 0 reve
rse only sequences
Input Read Pairs: 296415 Both Surviving: 296399 (99.99%) Forward Only Surviving: 16 (0.01%) Rever
se Only Surviving: 0 (0.00%) Dropped: 0 (0.00%)
TrimmomaticPE: Completed successfully
Started analysis of chr21Eric_repA.RNA.end1.trimmed.fastq
Approx 5% complete for chr21Eric_repA.RNA.end1.trimmed.fastq
```

Pre- and post-trim fastQC

Pre-trimming



Post-trimming

FastQC Report

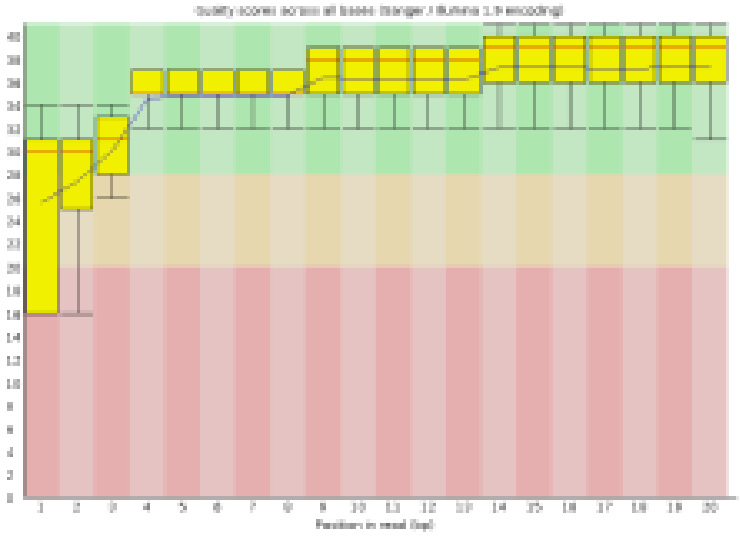
Summary

- Basic Statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

Basic Statistics

Measure	Value
Filename	ch211r1c_rep4_MMA_ecol_trimmed.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	298366
Sequences flagged as poor quality	0
Sequence length	20
%GC	43

Per base sequence quality



Implemented trimming steps (Quick reference)

Trimmomatic performs a variety of useful trimming tasks for illumina paired-end and single ended data. The selection of trimming steps and their associated parameters are supplied on the command line.

The current trimming steps are:

- ILLUMINACLIP: Cut adapter and other illumina-specific sequences from the read.
- SLIDINGWINDOW: Performs a sliding window trimming approach. It starts scanning at the 5' end and clips the read once the average quality within the window falls below a threshold.
- MAXINFO: An adaptive quality trimmer which balances read length and error rate to maximise the value of each read
- LEADING: Cut bases off the start of a read, if below a threshold quality
- TRAILING: Cut bases off the end of a read, if below a threshold quality
- CROP: Cut the read to a specified length by removing bases from the end
- HEADCROP: Cut the specified number of bases from the start of the read
- MINLEN: Drop the read if it is below a specified length
- AVGQUAL: Drop the read if the average quality is below the specified level
- TOPHRED33: Convert quality scores to Phred-33
- TOPHRED64: Convert quality scores to Phred-64