Worksheet 4.2 - Removing low-quality reads and adapters using Trimmomatic
Authors: Mary Allen & Daniel Ramirez

Trimmomatic webpage: http://www.usadellab.org/cms/?page=trimmomatic

Username: Screenshots show 'daramirez', though you will see your own username!

Sometimes short sequencing reads will contain unwanted adapter sequences. We want to remove these adapters so that the reads will be better aligned or mapped to a reference genome. If the adapter sequences are attached to the informative section of the reads, the mapping program will not know where to align the reads as the adapter part of them will not match to any part of the reference genome.

1. Using an appropriate terminal, log on to the cluster where you will use **fastQC**:
    a. Use ***pwd*** to make sure you know where you are and ***ls*** to make sure you know what is in this directory.

```
[daramirez@ip-172-31-15-245 ~]$ pwd
/Users/daramirez
[daramirez@ip-172-31-15-245 ~]$ ls -ls
total 0
```

    b. Change the working directory (***cd***) to your own scratch directory.

```
[daramirez@ip-172-31-15-245 ~]$ cd /scratch/Users/daramirez/
[daramirez@ip-172-31-15-245 daramirez]$ ls -ls
total 12
4 drwxrwxr-x 2 daramirez daramirez 6144 Jul  9 20:45 eofiles
4 drwxrwxr-x 2 daramirez daramirez 6144 Jul  9 20:45 fastQC
4 drwxrwxr-x 2 daramirez daramirez 6144 Jul  9 20:43 sbatch
```

2. Make 1 new directory/folder (***mkdir***) named trimmomatic.

```
[daramirez@ip-172-31-15-245 daramirez]$ mkdir trimmomatic
[daramirez@ip-172-31-15-245 daramirez]$ ls -ls
total 16
4 drwxrwxr-x 2 daramirez daramirez 6144 Jul  9 20:45 eofiles
4 drwxrwxr-x 2 daramirez daramirez 6144 Jul  9 20:45 fastQC
4 drwxrwxr-x 2 daramirez daramirez 6144 Jul  9 20:43 sbatch
4 drwxrwxr-x 2 daramirez daramirez 6144 Jul  9 21:52 trimmomatic
```

    This directory that will contain the results from trimmomatic. The error and output files generated by your batch scripts job will be stored in "eofiles". The batch script that will be created will live in the "sbatch" directory.

3. Go check the fastq data files in the following public directory using ***cd*** and ***ls***: /scratch/Workshop/SR2019/4_qc/ . In there, there is a folder called fastq. In there, there is a fastq file named "adaptor_dimers.fastq". This is the fastq file you will run trimmomatic on.

```
[daramirez@ip-172-31-15-245 daramirez]$ cd /scratch/Shares/public/sread2019/data_files/
[daramirez@ip-172-31-15-245 data_files]$ ls -lsh
total 28K
4.0K drwxrwxr-x 2 centos centos 6.0K Jul  5 20:36 assesment
4.0K drwxrwxr-x 2 centos centos 6.0K Jul  5 21:04 ATAC-seq
4.0K drwxrwxr-x 3 centos centos 6.0K Jul  5 20:54 ChIP-seq
4.0K drwxrwxr-x 3 centos centos 6.0K Jul  5 20:52 DNAre-seq
4.0K drwxrwxr-x 2 centos centos 6.0K Jul  5 20:33 fastq_for_quality_check
4.0K drwxrwxr-x 3 centos centos 6.0K Jul  5 20:56 RNA-seq
4.0K drwxrwxr-x 5 centos centos 6.0K Jul  5 14:36 videos
[daramirez@ip-172-31-15-245 data_files]$ cd fastq_for_quality_check/
[daramirez@ip-172-31-15-245 fastq_for_quality_check]$ ls -lsh
total 2.0G
5.9M -rwxrwxr-x 1 centos centos 5.9M Jul  5 20:32 adaptor_dimers.fastq
682M -rwxrwxr-x 1 centos centos 682M Jul  5 20:32 Day4HW_R1.fastq
682M -rwxrwxr-x 1 centos centos 682M Jul  5 20:33 Day4HW_R2.fastq
 11M -rwxrwxr-x 1 centos centos  11M Jul  5 20:33 Example_1.fastq.gz
6.5M -rwxrwxr-x 1 centos centos 6.5M Jul  5 20:33 Example_2.fastq.gz
 30M -rwxrwxr-x 1 centos centos  30M Jul  5 20:33 Example_3.fastq.gz
 20M -rwxrwxr-x 1 centos centos  20M Jul  5 20:33 Example_4.fastq.gz
295M -rwxrwxr-x 1 centos centos 295M Jul  5 20:33 Paired_R1.fastq
295M -rwxrwxr-x 1 centos centos 295M Jul  5 20:33 Paired_R2.fastq
```

4. Find and explore the contents (e.g. ***vim <file>***) of the script batch template "template.sbatch" in the directory: /scratch/Workshop/SR2019/scripts/
   You cannot edit, only look. The top of the file has information for the queue. The middle section contains job specific documentation. We will change this file so that it can be used for trimmomatic. This is your template. When you are done looking use ***:q!*** then press enter to exit the file.

```
#!/bin/bash
#SBATCH --job-name=<JOB-NAME>                       # Job name
#SBATCH --mail-type=ALL                             # Mail events (NONE, BEGIN, END, FAIL, ALL)
#SBATCH --mail-user=<EMAIL>                         # Where to send mail
#SBATCH --nodes=<n>                                 # Number of nodes requested
#SBATCH --ntasks=<n>                                # Number of CPUs (processor cores/tasks)
#SBATCH --mem=<n>gb                                 # Memory limit
#SBATCH --time=<00:00:00>                           # Time limit hrs:min:sec
#SBATCH --partition=compute                         # Partition/queue requested on server
#SBATCH --output=/scratch/Users/<USERNAME>/eofiles/%x.%j.out       # Standard output
#SBATCH --error=/scratch/Users/<USERNAME>/eofiles/%x.%j.err        # Standard error log

### Displays the job context
echo Job: $SLURM_JOB_NAME with ID $SLURM_JOB_ID
echo Running on host `hostname`
echo Job started at `date +"%T %a %d %b %Y"`
echo Directory is `pwd`
echo Using $SLURM_NTASKS processors across $SLURM_NNODES nodes

### Assigns path variables
INPUT_DIRECTORY=<PATH_TO_FILES>

### Loads modules
<MODULES_TO_LOAD>

### <SOFTWARE SPECIFICS>

echo Job finished at `date +"%T %a %d %b %Y"`
~
~
~
"template.sbatch" 28L, 1221C
```

5. Copy the script batch "template.sbatch" that you just looked at to your sbatch directory "/scratch/Users/<username>/sbatch/" using the new name "trimmomatic.sbatch" (***cp***

<***input***> <***output***>). Check that copying worked by moving to the sbatch directory and listing its contents (hint: ***cd*** & ***ls***).

```
[daramirez@ip-172-31-15-245 scripts]$ pwd
/scratch/Shares/public/sread2018/scripts
[daramirez@ip-172-31-15-245 scripts]$ ls -ls
total 8
4 drwxrwxr-x 2 centos    centos   6144 Jul  2 16:23 cookingshow
4 -rw-r--r-- 1 daramirez workshop 1281 Jul  9 20:42 template.sbatch
[daramirez@ip-172-31-15-245 scripts]$ cp template.sbatch /scratch/Users/daramirez/sbatch/trimmomatic.sbatch
[daramirez@ip-172-31-15-245 scripts]$ cd /scratch/Users/daramirez/sbatch/
[daramirez@ip-172-31-15-245 sbatch]$ ls -ls
total 4
4 -rw-r--r-- 1 daramirez daramirez 1281 Jul  9 22:26 trimmomatic.sbatch
```

6. Complete the new "trimmomatic.sbatch" file with the right content to run trimmomatic. (hint: transition to insert mode by pressing ***i*** if using vim.)

    a. Change the name of the script batch from <JOB-NAME> to something more useful, such as "trimmomatic".

```
#SBATCH --job-name=<JOB-NAME>                          # Job name
```

    b. Replace <EMAIL> with your own email address to which you want to receive any notifications.

```
#SBATCH --mail-user=<EMAIL>                          # Where to send mail
```

    c. Replace <USERNAME> with your own username to complete the path directory to where to store the error and output files.

```
#SBATCH --output=/scratch/Users/<USERNAME>/eofiles/%x.%j.out       # Standard output
#SBATCH --error=/scratch/Users/<USERNAME>/eofiles/%x.%j.err        # Standard error log
```

    d. Complete the following fields: nnodes, ntasks, mem and time. Trimmomatic can multi-thread, or use multiple processors per input file. So 1 node, 8 tasks or processors, 10gb for memory and 1 hour for wall-time should be enough.

    e. Specify first the path of the adaptor_dimmers.fastq file as the value of the variable "INPUT_DIRECTORY", and second the path that leads to the trimmomatic directory you created earlier in your scratch directory as the value of the variable "OUTPUT_DIRECTORY".

    Such that you have:
    "/scratch/Shares/public/sread2019/data_files/fastq_for_quality_check" to the INPUT_DIRECTORY variable, and
    "/scratch/Users/daramirez/trimmomatic/"
    to the OUTPUT_DIRECTORY variable.

```
### Assigns path variables
INPUT_DIRECTORY=<PATH_TO_INPUT_FILE>
OUTPUT_DIRECTORY=<PATH_TO_OUTPUT_FILE>
```

f. Assign the required modules necessary to run this trimmomatic job. To do this, exit vim by saving all changes (press ***ESC*** and type ***:wq!***). To look for the correct trimmomatic module, list all available modules on the computer cluster that contain the word "trimmomatic" in them. Type the following command ***module spider <string>*** and look for the one for trimmomatic.

```
[daramirez@ip-172-31-15-245 sbatch]$ module spider trimmomatic

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

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

Copy "module load trimmomatic/0.36". Re-open "trimmomatic.sbatch" using vim and replace "MODULES_TO_LOAD" with what you just copied.

```
### Loads modules
<MODULES_TO_LOAD>
```

g. The last edit you need to do is the actual text that runs trimmomatic!

The syntax to use trimmomatic for single-end reads is as follows:
java jar /opt/trimmomatic/0.36/trimmomatic-0.36.jar SE [-threads <n>] [-phred33 | -phred64] [-trimlog <output_trimlog>] <input_file> <output_file> ILLUMINACLIP ...

The syntax to use trimmomatic for paired-end reads is as follows:
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 ...

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

In our example, adaptor_dimmers.fastq is single-end reads, so choose:
<threads> is 8 (processors/threads/CPUs)
-phred33 is the base quality encoded in the fastq file.
<output_trimlog> is $OUTPUT_DIRECTORY/adaptor_dimers.trimlog
<input_file> is $INPUT_DIRECTORY/adaptor_dimers.fastq
<output_file> is $OUTPUT_DIRECTORY
ILLUMINACLIP:/opt/trimmomatic/0.36/adapters/TruSeq3-SE.fa:2:30:10
LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36
So we can go from having in the template:

```
### <SOFTWARE SPECIFICS>
```

To having a complete trimmomatic command:

```
### Running trimmomatic
java -jar /opt/trimmomatic/0.36/trimmomatic-0.36.jar SE -threads 8 -phred33 \
-trimlog $OUTPUT_DIRECTORY/adaptor_dimers.trimlog \
$INPUT_DIRECTORY/adaptor_dimers.fastq \
$OUTPUT_DIRECTORY/adaptor_dimers.trimmed.fastq \
ILLUMINACLIP:/opt/trimmomatic/0.36/adapters/TruSeq3-SE.fa:2:30:10 \
LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36
```

The \ at the end of every line in the trimmomatic section is used to break up what would be a long and confusing single line command into pieces corresponding to every part of the command, just for clarity purposes.

```
#!/bin/bash
#SBATCH --job-name=trimmomatic                    # Job name
#SBATCH --mail-type=ALL                           # Mail events (NONE, BEGIN, END, FAIL, ALL)
#SBATCH --mail-user=dara6367@colorado.edu         # Where to send mail
#SBATCH --nodes=1                                 # Number of nodes requested
#SBATCH --ntasks=8                                # Number of CPUs (processor cores/tasks)
#SBATCH --mem=10gb                                # Memory limit
#SBATCH --time=01:00:00                           # Time limit hrs:min:sec
#SBATCH --partition=compute                       # Partition/queue requested on server
#SBATCH --output=/scratch/Users/daramirez/eofiles/trimmomatic.%j.out       # Standard output
#SBATCH --error=/scratch/Users/daramirez/eofiles/trimmomatic.%j.err        # Standard error log

### Displays the job context
echo Job: $SLURM_JOB_NAME with ID $SLURM_JOB_ID
echo Running on host `hostname`
echo Job started at `date +"%T %a %d %b %Y"`
echo Directory is `pwd`
echo Using $SLURM_NTASKS processors across $SLURM_NNODES nodes

### Assigns path variables
INPUT_DIRECTORY=/scratch/Shares/public/sread2019/data_files/fastq_for_quality_check
OUTPUT_DIRECTORY=/scratch/Users/daramirez/trimmomatic

### Loads modules
module load trimmomatic/0.36

### Running trimmomatic
java -jar /opt/trimmomatic/0.36/trimmomatic-0.36.jar SE -threads 8 -phred33 \
-trimlog $OUTPUT_DIRECTORY/adaptor_dimers.trimlog \
$INPUT_DIRECTORY/adaptor_dimers.fastq \
$OUTPUT_DIRECTORY/adaptor_dimers.trimmed.fastq \
ILLUMINACLIP:/opt/trimmomatic/0.36/adapters/TruSeq3-SE.fa:2:30:10 \
LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36

###  Final time stamp
echo Job finished at `date +"%T %a %d %b %Y"`
~
"trimmomatic.sbatch" 36L, 1712C
```

Save all changes to “trimmomatic.sbatch” and exit vim.

7. Now that the batch script is ready, submit it to the job manager SLURM to begin processing the data. In the terminal, while located in the “sbatch” directory where “trimmomatic.sbatch” lives, type ***sbatch <sbatch file>***. The job manager will give you a job number. Once submitted, you can check on the status of jobs by typing ***squeue -u username***.

```
[daramirez@ip-172-31-15-245 sbatch]$ sbatch trimmomatic.sbatch
Submitted batch job 39
```

8. Move to the “eofiles” directory and open the error and output files.

```
[daramirez@ip-172-31-15-245 sbatch]$ cd ../eofiles/
[daramirez@ip-172-31-15-245 eofiles]$ ls -ls
total 16
4 -rw-rw-r-- 1 daramirez daramirez 855 Jul  9 20:45 fastQC.36.err
4 -rw-rw-r-- 1 daramirez daramirez 255 Jul  9 20:45 fastQC.36.out
4 -rw-rw-r-- 1 daramirez daramirez 755 Jul 10 00:03 trimmomatic.39.err
4 -rw-rw-r-- 1 daramirez daramirez 219 Jul 10 00:03 trimmomatic.39.out
```

```
[daramirez@ip-172-31-15-245 eofiles]$ more trimmomatic.39.out
Job: trimmomatic with ID 39
Running on host ip-172-31-5-22
Job started at 00:03:29 Tue 10 Jul 2018
Directory is /scratch/Users/daramirez/sbatch
Using 8 processors across 1 nodes
Job finished at 00:03:30 Tue 10 Jul 2018
```

```
[daramirez@ip-172-31-15-245 eofiles]$ more trimmomatic.39.err
TrimmomaticSE: Started with arguments:
 -threads 8 -phred33 -trimlog /scratch/Users/daramirez/trimmomatic/adaptor_dimers.trim
log /scratch/Shares/public/sread2018/data_files/fastq_for_quality_check/adaptor_dimers
.fastq /scratch/Users/daramirez/trimmomatic/adaptor_dimers.trimmed.fastq ILLUMINACLIP:
/opt/trimmomatic/0.36/adapters/TruSeq3-SE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDO
W:4:15 MINLEN:36
Using Long Clipping Sequence: 'AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTA'
Using Long Clipping Sequence: 'AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC'
ILLUMINACLIP: Using 0 prefix pairs, 2 forward/reverse sequences, 0 forward only sequen
ces, 0 reverse only sequences
Input Reads: 37500 Surviving: 36295 (96.79%) Dropped: 1205 (3.21%)
TrimmomaticSE: Completed successfully
```

9. Check the “trimmomatic” directory. There will be two files: a new fastq file containing the trimmed version of the adaptor_dimers file and its corresponding trimlog file.

```
[daramirez@ip-172-31-15-245 eofiles]$ cd ../trimmomatic/
[daramirez@ip-172-31-15-245 trimmomatic]$ ls -ls
total 8128
2392 -rw-rw-r-- 1 daramirez daramirez 2449253 Jul 10 00:03 adaptor_dimers.trimlog
5736 -rw-rw-r-- 1 daramirez daramirez 5871674 Jul 10 00:03 adaptor_dimers.trimmed.fastq
[daramirez@ip-172-31-15-245 trimmomatic]$ head adaptor_dimers.trimlog
HWI-ST753:239:C6YUTACXX:6:1101:1079:2089 1:N:0:CTTGTA 51 0 51 0
HWI-ST753:239:C6YUTACXX:6:1101:1048:2124 1:N:0:CTTGTA 51 0 51 0
HWI-ST753:239:C6YUTACXX:6:1101:1462:2189 1:N:0:CTTGTA 50 0 50 1
HWI-ST753:239:C6YUTACXX:6:1101:1550:2084 1:N:0:CTTGTA 51 0 51 0
HWI-ST753:239:C6YUTACXX:6:1101:1605:2187 1:N:0:CTTGTA 51 0 51 0
HWI-ST753:239:C6YUTACXX:6:1101:2384:2060 1:N:0:CTTGTA 51 0 51 0
HWI-ST753:239:C6YUTACXX:6:1101:2458:2086 1:N:0:CTTGTA 51 0 51 0
HWI-ST753:239:C6YUTACXX:6:1101:2494:2121 1:N:0:CTTGTA 51 0 51 0
HWI-ST753:239:C6YUTACXX:6:1101:2680:2060 1:N:0:CTTGTA 51 0 51 0
HWI-ST753:239:C6YUTACXX:6:1101:2898:2184 1:N:0:CTTGTA 51 0 51 0
```

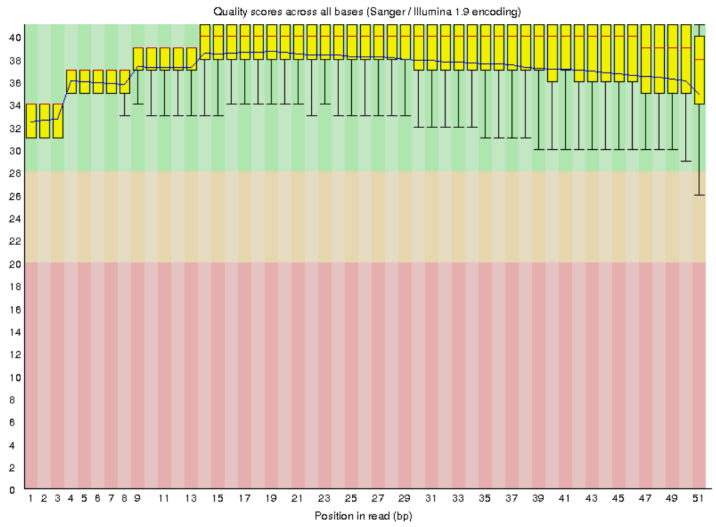
10. Now run fastQC on both the original adaptor_dimers fastq and the trimmed adaptor_dimers fastq, download the resulting html files to your computer and open them using a web browser, as shown in the worksheet 4.1.

FastQC Report

Tue 10 Jul 2018
adaptor_dimers.fastq

Summary

Per base sequence quality



FastQC Report

Tue 10 Jul 2018
adaptor_dimers.trimmed.fastq

Summary

Per base sequence quality