trimmomatic: paired end mode

Slurm:
this will load the trimmomatic software
you will have two new files, Truseq3-SE.fa and seq.fq, which are necessary to run trimmomatic

```
module load trimmomatic/0.33
wget
https://public.s3.msi.umn.edu/reframe/sw/trimmomatic/Truseq3-SE.fa
wget https://public.s3.msi.umn.edu/reframe/sw/trimmomatic/seq.fq
```

here is the code to run the trimmomatic program
this is specific to PAIRED END data
some notes:
'PE' = 'paired-end'
-phred33 is the standard quality check
the output does not need to exist; trimmomatic will create it for you
make sure you are in the correct directory using pwd (print working directory) and cd (change directory) commands
if TruSeq3-SE.fa is in a different directory, make sure to provide the proper path
i.e. /home/user/TruSeq3-SE.fa

```
trimmomatic PE -phred33 [INPUT FILE 1 .fastq]  [OUTPUT FILE 1 .fq]
[INPUT FILE 2 .fastq]  [OUTPUT FILE 2 .fq] \
ILLUMINACLIP:TruSeq3-SE.fa:2:30:10 LEADING:3 TRAILING:3
SLIDINGWINDOW:4:15 MINLEN:36
```

Getting started:
the following code will remove adapters, remove leading low quality or N bases (`LEADING:3`), remove trailing low quality or N bases (`TRAILING:3`), scan the read with a 4-base wide sliding window, cutting when the average quality per base drops below 15 (`SLIDINGWINDOW:4:15`), and drops reads below the 36 bases long (`MINLEN:36`)

Current trimming steps:

1. ILLUMINACLIP: clip adapter and other illumina-specific sequences from the read
    a. fastaWithAdaptersEtc: specifies the path to a fasta file containing all the adapters, PCR sequences etc. The naming of the various sequences within this file determines how they are used. See below.
2. SLIDINGWINDOW: perform a sliding window trimming, cutting once the average quality falls below a threshold
3. LEADING: Cut bases off the start of a read, if below a threshold quality
4. TRAILING: Cut bases off the end of a read, if below a threshold quality
5. CROP: Cut the read to a specified length
6. HEADCROP: Cut the specified number of bases from the start of the read
7. MINLEN: Drop the read if it is below a specified length

8. TOPHRED33: Convert quality scores to Phred-33
9. TOPHRED64: Convert quality scores to Phred-64

- Works with FASTQ, either uncompressed or gzipp'ed

Running Trimmomatic
if no quality score is specified, phred-64 is the default

Specifying a trimlog file creates a log of all read trimmings, including:

- The read name
- Surviving sequence length
- Location of the first surviving base aka amount trimmed from start
- Location of last surviving base in the original read
- Amount trimmed from the end

For paired-end data, two input files, and 4 output files are specified, 2 for the 'paired' output where both reads survived the processing, and 2 for corresponding 'unpaired' output where a read survived, but the partner read did not.

- For input files, either can be used
    - Explicitly naming the 2 input files
    - Naming the forward file using the -basein flag, where the reverse file can be determined automatically. The second file is determined by looking for common patterns of file naming, and changing the appropriate character to reference the reverse file. Examples which should be correctly handled include:
        - Sample_Name_R1_001.fq.gh -> Sample_Name_R2_001.fq.gz
        - Sample_Name.f.fastq -> Sample_Name.r.fastq
        - Sample_Name.1.sequence.txt -> Sample_Name.2.sequence.txt
- For output files, " "
    - Explicitly naming the 4 output files
    - Providing a base file name using the –baseout flag, from which the 4 output files can be derived. If the name "mySampleFiltered.fq.gz" is provided, the following 4 file names will be used:
        - mySampleFiltered_1P.fq.gz - for paired forward reads
        - mySampleFiltered_1U.fq.gz - for unpaired forward reads
        - mySampleFiltered_2P.fq.gz - for paired reverse reads
        - mySampleFiltered_2U.fq.gz - for unpaired reverse reads

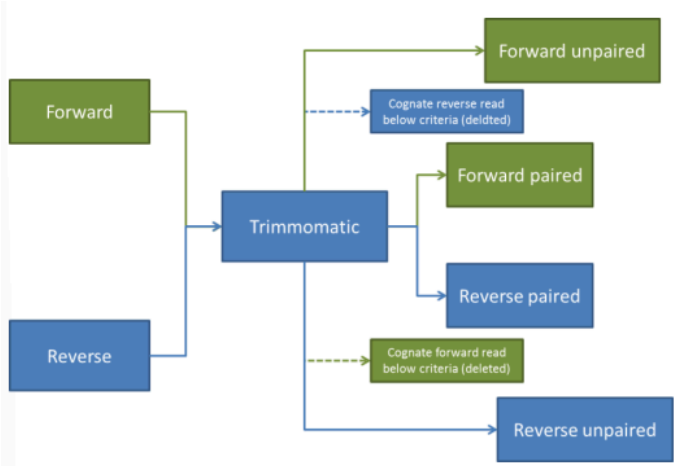

Figure 1: Flow of reads in Trimmomatic Paired End mode

EXAMPLE:

```
java -jar trimmomatic-0.30.jar PE s_1_1_sequence.txt.gz
s_1_2_sequence.txt.gz
lane1_forward_paired.fq.gz lane1_forward_unpaired.fq.gz
lane1_reverse_paired.fq.gz lane1_reverse_unpaired.fq.gz
ILLUMINACLIP:TruSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3
SLIDINGWINDOW:4:15 MINLEN:36
```

This will perform the following in this order

- Remove Illumina adapters provided in the TruSeq3-PE.fa file (provided). Initially Trimmomatic will look for seed matches (16 bases) allowing maximally 2 mismatches. These seeds will be extended and clipped if in the case of paired end reads a score of 30 is reached (about 50 bases), or in the case of single ended reads a score of 10, (about 17 bases).
- Remove leading low quality or N bases (below quality 3)
- Remove trailing low quality or N bases (below quality 3)
- Scan the read with a 4-base wide sliding window, cutting when the average quality per base drops below 15
- Drop reads which are less than 36 bases long after these steps

trimmomatic attempt:

template:
trimmomatic PE -phred33 <input_forward_file> <input_reverse_file>
<output_forward_paired_file> <output_forward_unpaired_file> <output_reverse_paired_file>
<output_reverse_unpaired_file>

trimmomatic PE -phred33 2408-39-Peony_1.fastq.gz 2408-39-Peony_2.fastq.gz
reverse_Peony1.fq reverse_Peony1_unpaired.fq
reverse_Peony2.fq reverse_Peony2_unpaired.fq TRAILING:20 LEADING:20 MINLEN:36

TrimmomaticPE: Started with arguments:
-phred33 2408-39-Peony_1.fastq.gz 2408-39-Peony_2.fastq.gz reverse_Peony1.fq
reverse_Peony1_unpaired.fq reverse_Peony2.fq reverse_Peony2_unpaired.fq TRAILING:20
LEADING:20 MINLEN:36
Multiple cores found: Using 4 threads
Input Read Pairs: 24742524 Both Surviving: 24742524 (100.00%) Forward Only Surviving: 0
(0.00%) Reverse Only Surviving: 0 (0.00%) Dropped: 0 (0.00%)
TrimmomaticPE: Completed successfully

abels053@acn23 [~/Plant_Virology/Plant_Virology/Peony_RawData] % trimmomatic PE
-phred33 2408-39-Peony_1.fastq.gz 2408-39-Peony_2.fastq.gz reverse_Peony1.fq
reverse_Peony1_unpaired.fq
reverse_Peony2.fq reverse_Peony2_unpaired.fq TRAILING:30 LEADING:30 MINLEN:36
TrimmomaticPE: Started with arguments:
-phred33 2408-39-Peony_1.fastq.gz 2408-39-Peony_2.fastq.gz reverse_Peony1.fq
reverse_Peony1_unpaired.fq reverse_Peony2.fq reverse_Peony2_unpaired.fq TRAILING:30
LEADING:30 MINLEN:36
Multiple cores found: Using 4 threads
Input Read Pairs: 24742524 Both Surviving: 24738134 (99.98%) Forward Only Surviving: 2860
(0.01%) Reverse Only Surviving: 1494 (0.01%) Dropped: 36 (0.00%)
TrimmomaticPE: Completed successfully

abels053@acn05 [~/Plant_Virology/Plant_Virology/Barley_RawData] % trimmomatic PE -phred33 2311-10-BY-1E1_1.fastq.gz 2311-10-BY-1E1_2.fastq.gz E1_paired.fq E1_unpaired.fq E2_paired.fq
E2_unpaired.fq TRAILING:30 LEADING:30 MINLEN:36
TrimmomaticPE: Started with arguments:
 -phred33 2311-10-BY-1E1_1.fastq.gz 2311-10-BY-1E1_2.fastq.gz E1_paired.fq E1_unpaired.fq E2_paired.fq E2_unpaired.fq TRAILING:30 LEADING:30 MINLEN:36
Multiple cores found: Using 4 threads
Input Read Pairs: 33153647 Both Surviving: 33129491 (99.93%) Forward Only Surviving: 16281 (0.05%) Reverse Only Surviving: 7501 (0.02%) Dropped: 374 (0.00%)
TrimmomaticPE: Completed successfully

abels053@acn05 [~/Plant_Virology/Plant_Virology/Buckthorn_RawData] % trimmomatic PE -phred33 2401-16-BTH_1.fastq.gz 2401-16-BTH_2.fastq.gz BTH_1_paired.fq BTH_1_unpaired.fq BTH_2_pair
ed.fq BTH_2_unpaired.fq TRAILING:30 LEADING:30 MINLEN:36
TrimmomaticPE: Started with arguments:
 -phred33 2401-16-BTH_1.fastq.gz 2401-16-BTH_2.fastq.gz BTH_1_paired.fq BTH_1_unpaired.fq BTH_2_paired.fq BTH_2_unpaired.fq TRAILING:30 LEADING:30 MINLEN:36
Multiple cores found: Using 4 threads
Input Read Pairs: 21801579 Both Surviving: 21791559 (99.95%) Forward Only Surviving: 8633 (0.04%) Reverse Only Surviving: 1352 (0.01%) Dropped: 35 (0.00%)
TrimmomaticPE: Completed successfully