Genome Analysis

Trimmomatic: A flexible trimmer for Illumina Sequence Data

Anthony M. Bolger[1,2], Marc Lohse[1] and Bjoern Usadel[2,3,*]

[1]Max Planck Institute of Molecular Plant Physiology, Am Mühlenberg 1, 14476 Golm, Germany.

[2]Institut für Biologie I, RWTH Aachen, Worringer Weg 3, 52074 Aachen, Germany.

[3]Institut of Bio- and Geosciences: Plant Sciences, Forschungszentrum Jülich, Leo-Brandt-Straße, 52425 Jülich, Germany

Trimmomatic Options

- ILLUMINACLIP: Cut adapter and other illumina-specific sequences from the read.
- SLIDINGWINDOW: Perform a sliding window trimming, cutting once the average quality within the window falls below a threshold.
- 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
- HEADCROP: Cut the specified number of bases from the start of the read
- MINLEN: Drop the read if it is below a specified length
- TOPHRED33: Convert quality scores to Phred-33
- TOPHRED64: Convert quality scores to Phred-64

Adaptor Trimming

- ILLUMINACLIP:<fastaWithAdaptersEtc>:<seed mismatches>:<palindrome clip threshold>:<simple clip threshold>
  - **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.
  - **seedMismatches:** specifies the maximum mismatch count which will still allow a full match to be performed
  - **palindromeClipThreshold:** specifies how accurate the match between the two 'adapter ligated' reads must be for PE palindrome read alignment.
  - **simpleClipThreshold:** specifies how accurate the match between any adapter etc. sequence must be against a read.

Trimmomatic trims reads in 2 ways

- ‘Simple’ trimming:
  - Each adapter sequence is tested against the reads, and if a sufficiently accurate match is detected, the read is clipped appropriately
- ‘Palindrome’ trimming:
  - Specifically designed for the case of 'reading through' a short fragment into the adapter sequence on the other end.
  - In this case, the forward read is clipped and the reverse read dropped (since it contains no new data)

Quality Trimming

- SLIDINGWINDOW:<windowSize>:<requiredQuality>
  - windowSize: specifies the number of bases to average across
  - requiredQuality: specifies the average quality required for the window.
- LEADING:<quality>
  - quality: Specifies the minimum quality required to keep a base *at the beginning of the read*.
- TRAILING:<quality>
  - quality: Specifies the minimum quality required to keep a base *at the end of the read*.

Size Trimming

- CROP:<length>
  - length: The number of bases to keep, from the start of the read.
- HEADCROP:<length>
  - length: The number of bases to remove from the start of the read.
- MINLEN:<length>
  - length: Specifies the minimum length of reads to be kept.

Running trimmomatic

- Single-end:
  - java -jar <path to trimmomatic jar> SE [-threads <threads>] [-phred33 | -phred64] [-trimlog <logFile>] <input> <output> <step 1> ...
- Paired-end:
  - java -jar <path to trimmomatic.jar> PE [-threads <threads] [-phred33 | -phred64] [-trimlog <logFile>] <input 1> <input 2> <paired output 1> <unpaired output 1> <paired output 2> <unpaired output 2> <step 1> ...

Quick start settings

java -jar trimmomatic-0.30.jar PE --phred33 input_forward.fq.gz input_reverse.fq.gz output_forward_paired.fq.gz output_forward_unpaired.fq.gz output_reverse_paired.fq.gz output_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:
  - Remove adapters
  - 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 below the 36 bases long