All downloads from `http://www.mimuw.edu.pl/~szczurek/TSG2/01_lab/`

Exercise #1

Quality control with FastQC.

1. Download example sequence file test1.fastq
2. Download the FastQC software fastqc_v0.11.3.zip
   - Unzip
   - chmod 755 fastqc
   - ./fastqc
3. Open and analyze test1.fastq
4. What is the number of sequences in this file?
5. Save report 1

Exercise #2

Quality filtering with Trimmomatic.

1. Download the Trimmomatic software Trimmomatic-0.33.zip
2. java -jar trimmomatic-0.33.jar SE -phred33 test1.fastq output.fq ILLUMINACLIP:TruSeq3-SE:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36
3. Run fastqc on the output file and compare with the input.
   - What is the number of remaining reads?
   - Which of the problems disappeared entirely?

Exercise #2 continued on next page. . .

Meaning of the Trimmomatic parameters:
ILLUMINACLIP:

- Trimmomatic uses a two-step approach to find matches between the adapters and reads. First, short sections of each adapter (maximum 16 bp) are tested in each possible position within the reads. If this short alignment, known as the seed is a perfect or sufficiently close match, determined by the seedMismatch parameter (see below), the entire alignment between the read and adapter is scored. In the full alignment, a perfect match of a 12 base sequence will score just over 7, while 25 bases are needed to score 15. As such we recommend values of between 7 -15 as the threshold value for simple alignment mode.
- ILLUMINACLIP: fastaWithAdaptersEtc : seed mismatches: palindrome clip threshold : simple clip threshold
- more strict setting than the default: ILLUMINACLIP:TruSeq3-SE:1:30:15

SLIDINGWINDOW:

- Perform a sliding window trimming, cutting once the average quality within the window falls below a threshold. By considering multiple bases, a single poor quality base will not cause the removal of high quality data later in the read.
- SLIDINGWINDOW: windowSize: requiredQuality
- less strict setting than the default: SLIDINGWINDOW:4:2

LEADING

- Remove low quality bases from the beginning. As long as a base has a value below this threshold the base is removed and the next base will be investigated.
- LEADING: quality
- quality: Specifies the minimum quality required to keep a base.

TRAILING

- Remove low quality bases from the end. As long as a base has a value below this threshold the base is removed and the next base (which as trimmomatic is starting from the 3 prime end would be base preceding the just removed base) will be investigated.
- TRAILING: quality
- quality: Specifies the minimum quality required to keep a base.

MINLEN

- This module removes reads that fall bel ow the specified minimal length. If required, it should normally be after all other processing steps. Reads removed by this step will be counted and included in the dropped reads count presented in the trimmomatic summary.
- MINLEN: length
- length: Specifies the minimum length of reads to be kept

Exercise #3

Best practice in quality filtering. Work in pairs.

Exercise #3: (a)

Defining the quality objective.

Within each pair,

- Discuss what are the minimum characteristics (in terms of FastQC checks) of a dataset with acceptable quality.
- Prepare one person in the pair to present your definition of acceptable quality.

Exercise #3: (b)

CONTEST: Optimizing the quality objective.

Using trimmomatic, reach the (consensus) acceptable quality while keeping the maximum number of reads.

Exercise #4

Quality control and filtering with the Galaxy server

1. http://centromere:8080
2. login: your email
3. password: your last name (first letter is large)
4. upload data SRR020192.fastq
5. what is the number of reads in this file
6. what is their length?
7. run FastQC
8. optimize quality with

CLIP the adapters

TRIM the reads based on their quality

FILTER low quality reads

Homework in quality control #5

Working with real size problem.

- Run quality control on test2 with fastQC.
- Run quality filtering on test2 with your tool of choice.
- Write which tool you chose and which parameters and why.
- Report fastQC results on the filtered data.

Homework in R #6

Programming exercises

Drawn at random per each person. Report both code and results.