Quality Control Tutorial

Within this tutorial we will

- run fastqc of some NGS files to learn to understand the output
- run trimmomatics to test some parameter and prepare the NGS file for the next exercises

FASTQC

https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

As mentioned during the lecture, fastqc is a simple tool to monitor the quality and the properties of a NGS sequencing file in fastq.

We will check the quality of three NGS experiments, one from DNA sequencing, one from RNA sequencing and one from a small RNA sequencing.

Since we will create a lot of output, some we will use in a downstream analysis, I would recommend to get very well organized with a clear system of directories.

Create a working directory for example called NGSTutorial:

mkdir QCTutorial

within this directory create a directory rawData:

mkdir QCTutorial /rawData

Copy files form /data/ with the following files into the rawData directory:

- TMEB117_R1_frac.fastq
- TMEB117_R2_frac.fastq
- TMEB419RNA_frag_R1.fastq
- TMEB419RNA_frag_R2.fastq

Be sure that you are within the working directory:

cd ~/ QCTutorial

Then create a directory for the quality control:

mkdir qc

For each fastq file we run the the quality control with the following command:

- *fastqc -o qc rawData/TMEB117_R1_frac.fastq*
- *fastqc -o qc rawData/TMEB117_R2_frac.fastq*
- *fastqc -o qc rawData/TMEB419RNA_frag_R1.fastq*
- *fastqc -o qc rawData/TMEB419RNA_frag_R2.fastq*

You will get for each fastq file two output files:

- TMEB117_R1_frac_fastqc.zip (report, data files and graphs)
- TMEB117_R1_frac_fastqc.html (report in html)
- etc.

We download both files to the local computer for consulting.

scp-r mlandi@13.36.156.214:/home/mlandi/QCTutorial/qc ./

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

Dependent on what analysis you need to do with the NGS data it is wise to process the data according to the quality control and remove low score sequences and/or low score 5' and 3' fragments. Especially if you plan to run a sequence assembly trimming low quality fragments will improve the assembly accuracy.

In this tutorial we will run Trimmomatic for the two sequencing experiments we were checking the quality.

Let's get the output into a different directory:

mkdir trim

We process the reads according to the following parameters:

- clip the first 10 nucleotides
- Remove adapters (if any)
- Remove trailing low quality or N bases (below quality 10)
- Scan the read with a 4-base wide sliding window, cutting when the average quality per base drops below 15
- Drop reads below the 70 bases long

Here the example command for the first experiment:

java -jar /usr/local/src/Trimmomatic-0.39/trimmomatic-0.39.jar *PE -phred33 rawData/TMEB117_R1_frac.fastq rawData/TMEB117_R2_frac.fastq trim/TMEB117_forward_paired.fq.gz trim/TMEB117_forward_unpaired.fq.gz trim/TMEB117_reverse_paired.fq.gz trim/TMEB117_reverse_unpaired.fq.gz ILLUMINACLIP:TruSeq3-PE.fa:2:30:10 HEADCROP:10 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:70*

Once you get the cleaned sequences run again fastqc to check the result.

Run the other experiments accordingly.