BIOINFORMATICS

PROJECT: Quality control and filtering of NGS data

Quality control and filtering of NGS data is very important and can be done by the Galaxy web interface https:// usegalaxy.org. This website provides a variety of tools needed to manipulate, check and visualise the sequence data.

1. Go to the https:// usegalaxy.org website and create a user account (User --> Register). Login under the "User" tab (User --> Login).

2. Upload the data (SeqA.fastq and SeqB.fastq) to your Galaxy account (Get Data --> Upload File). You will see both datasets in right hand side panel.

3. We have to "groom" the fastq file. Because, different quality formats exist (different symbols for same phred score), this step converting all formats to Sanger format is required. (FASTQ Groomer --> select the fastq file --> select "Sanger" --> Excecute). Do this for both fastq files. As always, the newly created files are shown in the right panel.

4. In order to control the quality of data, run the FastQC programme. In the left panel, click on "NGS: QC and Manipulation" --> "FastQC Read Quality reports"). Based on "Per base sequence quality" and "Per sequence quality scores" plots answer the questions below:
   a. what is the most frequent average quality of reads in both files?
   b. how many reads represents the highest average quality?
   c. what is the average quality per base in the position 36?

5. Which file do you think is higher quality?

6. Filter out poor quality sequences by tools "NGS: QC and manipulation" --> "Filter FASTQ". Set the minimum base quality at 20. Rerun FastQC on the filtered files. Have they improved?

7. Try at least one different quality threshold and rerun FastQC.

8. Try one other method of reads filtering.

Based on obtained results prepare the presentation about data filtering. Describe used methods and explain how they work. Interpret the results focusing on quality of reads.