Data Science in Bioinformatics

Day 5

Let's dig a little more into the topics of the first day of student presentations. You heard a lot of general information about biological sequences and the way they can be presented. In addition to the presentation style, sequences can vary in their quality. Here are a few exercises to improve your understanding of the topic.

The files needed for the exercise can be found here:
https://www.ebi.ac.uk/ena/browser/view/ERR10000000
For downloading:
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR100/000/ERR10000000/ERR10000000_1.fastq.gz
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR100/000/ERR10000000/ERR10000000_2.fastq.gz

1) **Sequences**
    - **a.** What is the difference between short- and long-reads?
    - **b.** What do the extensions "_1" and "_2" in the files stand for?

2) **Quality Control**
    - **a.** What does a bad/good quality sequence mean and how can the quality of a sequence be determined?
    - **b.** Let's put the things to use you learned in this session:
        - **i.** Please download the two sample fastq files to your working directory.
        - **ii.** Create a new Conda environment called "quality" and install FastQC and MultiQC inside the environment.
        - **iii.** Create a FastQC output for the files you downloaded.
        - **iv.** How do FastQC and MultiQC interact?
        - **v.** Create a MultiQC output for the downloaded files and look at it. Please hand in your MultiQC results.
        - **vi.** Which GC content is shown in the first figure of the MultiQC report for the two files?

3) **SARS-CoV-2**
    - **a.** How large is the SARS-CoV-2 genome? What about the genome size of the bacterium E. coli or the human genome? What are general differences between human and SARS-CoV-2 nucleic acids?
    - **b.** Which genes can be found on the SARS-CoV-2 genome? What are their functions?
    - **c.** In this context: what does every sub-component of "S:D614G" mean?
    - **d.** What kind of single-base changes on the nucleic acid level could cause that?
    - **e.** Explain the impact of 'S:D614G'.