Birkbeck, University of London
MSc Bioinformatics
Omics (Spring 2022/2023)
Coursework Assignment
Authors: Dr Igor Ruiz and Dr Irilenia Nobeli

Date of the assignment: 14/04/2023
Date assignment is due: 05/05/2023
Marking: Highest mark possible (for the full coursework): 33

General Guidelines

The aim of the NGS coursework is to give you experience in analyzing genomic data from Illumina next-generation sequencing technologies. Your role as a bioinformatician is not simply to run programs with recommended parameters and pass on the results. Your role is to run and re-run analyses, making informed judgements about the results and improving your methods and pipelines, making use of your experience, shared knowledge and other people's expertise. This coursework aims to train your ability to do just that.

In order to carry out NGS analysis you also need to be familiar with using the command line in unix, understanding shell scripts and creating reports that make your work easy to understand and share and reproducible (we recommend markdown and R markdown files for this).

The coursework below tries to address some of these skills. There are 2 parts to it. Altogether, this coursework counts as **33%** of the mark for the module.

Guidelines for Coursework Part 1

Part 1 of the coursework will contribute **8%** to the total module mark.

Please upload your answers as a PDF file (see below) on Moodle by the deadline shown above.

Description of the coursework (part 1) –
Analysis of whole genome sequencing data using a known reference genome

Your final submission

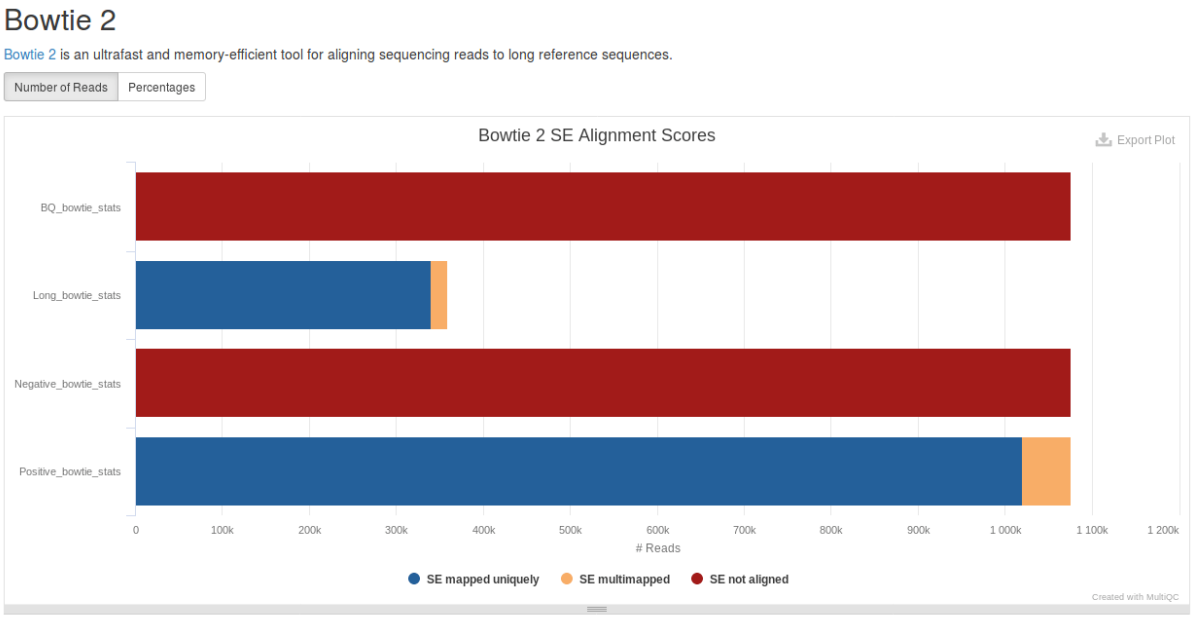
You should submit a PDF or .html file containing all information mentioned at the end of each question. If you submit an .html file, make sure any figures/plots are visible to others opening the file on a web browser.

You can produce this file whichever way you want but you may want to start learning about **markdown** files and produce a PDF, using markdown as your starting point (however, for this exercise your mark will not depend on the use of markdown). You will not be marked on whether you are using markdown or any other type of word processing for this coursework.

> There are plenty of markdown help files on the internet, such as:
> https://www.markdownguide.org/
> We recommend using an IDE that allows you to view the output at the same time as the input (e.g. Dillinger at https://dillinger.io/ or Visual Studio Code with a markdown extension downloaded). Many general editors (like Atom) or specialised ones (like Macdown for Macs) also read and interpret markdown. To include figures/plots in your markdown, take a snapshot and incorporate it following guidelines for including images in markdown files. Save your markdown in your local hard drive as .html, include a locally saved snapshot, open in web browser and then save to pdf.

Question 1 (worth 7 marks)

In the first practical you were given a *fastq* file containing reads from a simulated whole-genome sequencing experiment. When reads were split using the four barcodes (Negative, Positive, Long and BQ), the Negative.fq file gave very poor mapping results (see figure below).



Using good NGS practices, remap Negative.fq so that the mapping statistics improve. You should be able to obtain similar mapping results to those obtained with Positive.fq.

Hint:

a) View the file to work out what could be wrong with the reads! A fastqc report might help you too.

b) Use bowtie2 to align the reads to the reference genome you've been given; you should **try at least two different options** of aligning the reads (playing with different parameters when running bowtie2).

Note: Different options will give you different mapping results (some will be better than others)

c) Run samtools stats and flagstats and use multiQC to summarise your output after mapping.

Your answer for question 1 should include:

- All code used to carry out this exercise.
- The directories containing your input and output so it can be checked (please make these directories readable to "all" by using the unix command:

  `chmod a+r directory_name`
- An explanation of what was wrong with the reads. [2 marks]
- Your mapping trials alongside with your chosen mapping options (with code). [2 marks]
- Final mapping statistics (samtools stats and flagstats, summarised with the help of multiqc). [2 marks]

- A very brief discussion of how your alignment options differ and which one you consider to be better (and why). [2 marks]