Birkbeck, University of London
MSc Bioinformatics
Sequence Analysis and Omics (Spring 2023/2024)
Coursework Assignment 1
Authors: Dr Irilenia Nobeli and Dr Igor Ruiz

Date of the assignment:	10/03/2024
Date assignment is due:	17/03/2024
Weight:	20% of the total mark for the SAO module.
Max mark:	40

Guidelines

The aim of this 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).

There are two parts to this coursework:
a) The first part asks you to **apply your command line skills and knowledge of NGS-relevant software to manipulate files with artificially produced NGS reads**. You will need to carry out this work on our server thoth OR on a local installation on your own computer.
b) The second part **gives you the opportunity to apply your skills in using the BV-BRC server to answer a series of questions focused on genome and sequence analysis of the SARS-Cov2 virus**.

Queries regarding the content of the task can be addressed by e-mail to Irilenia Nobeli– i.nobeli@bbk.ac.uk.

Notes on College regulations

Plagiarism: In preparing their answers, students should take note of the School's policy on plagiarism. If you have any doubt about whether part of your coursework may be considered plagiarism, please discuss it with the module organiser prior to submission. In submitting your coursework, you declare that you have read and understood the College's policy on plagiarism and that all the submitted work is your own.

The electronic copy (PDF format) of the coursework should be submitted on Moodle.

Do not add links to temporary results web pages that you have generated while carrying out work for the report – these may not work for someone else trying to open them later! Put any information that you wish to be considered for marks in the report or an appendix.

In the interests of fairness, and following Birkbeck rules the deadline will be strictly enforced. In the absence of mitigating circumstances, if you hand your coursework in after the agreed deadline, your mark will be reduced by 10% for the first week late and capped at 50% for the second week late. Should you have unexpected difficulties in submitting on time you can apply for these penalties to be waived/removed via the College's "mitigating circumstances" procedures. The form is available on MyBirkbeck.

Part A [15 marks]

It is recommended that you carry out this part on our Birkbeck server thoth.

In this part of the coursework, you are provided with a file of NGS genomic reads (`/d/in4/u/ubcg71a/teaching/sao/data/final_merge_synthetic_reads.fq`). This file is in FASTQ format and was created by Dr Igor Ruiz. The reads are not totally artificial, in the sense that they originate from a real *E. coli* genome but Igor has manipulated the reads for didactic purposes. The aim is to examine and manipulate this file so that the reads can be aligned to a reference genome (the reference is also provided by Igor).

a) Run FastQC on final_merge_synthetic_reads.fq and summarise your findings in no more than a few sentences. Include relevant FastQC plots as evidence to back up your conclusion.

b) We are told that the original reads file contains three samples mixed together. To identify the samples, the centre used barcoding, i.e. added a specific sequence in front of each read to keep track of its origin. Split the original file into four fastq files using the software `cutadapt` and the information that the barcode "GATACA" identifies reads originating from a sample called "positive", and barcodes "AGTAGT", "CACACA" and "AAACCC" identify reads originating respectively from samples: "negative", "bq", and long. Make sure the barcodes are trimmed by cutadapt.

c) Map both positive and negative fastq files with bowtie2 (end-to-end alignment) to the reference genome provided under:
`/d/in4/u/ubcg71a/teaching/sao/genomes/AFPN02.1/AFPN02.1_merge.fasta`
Compare the output of bowtie2 for the two fastq files mapped using the bowtie statistics for each case (bowtie produces these stats at the end of the run).

d) 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:

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

ii) 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 – read the documentation of this software to help you decide what to do!).

iii) Run samtools stats and flagstats and use multiQC to summarise your outputs after mapping.

Your answer for part A should include:

- All code used to carry out this exercise (you can simply include your commands in your report).

[1 mark]

- 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`
- A summary of your FASTQC run for part (a); you should include some plots and put emphasis on modules that have failed or look worrying.
  [2 marks]
- The cutadapt command for separating the original file into four fastq files, each containing reads from one sample.
  [1 marks]
- An explanation of what was wrong with the “negative” sample reads.
  [1 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). [4 marks]

- Two marks will be reserved for neat presentation of results and a high-quality report.
  [2 marks]

Part B [25 marks]

Note: the marks in this part add up to 23. Two additional marks are reserved for a neat and accurate presentation of results.

Question 1. [12 marks]

For this question, you will need to use the BV-BRC server we used in practical 1 (https://www.bv-brc.org) . Note that if you haven't already created an account during that practical, you will need to do so now, in order to run your own jobs on the server.

The accession number SRR23290383 in the European Nucleotide Archive (www.ebi.ac.uk/ena) corresponds to the Illumina paired-end genome sequencing results of a clinical sample obtained from a patient with COVID-19 in Texas, USA in July 2023. We will use this sample for this part of the coursework.

Note: You do not need to download and re-upload the raw fastq files to the BV-BRC server. The server can use the ENA codes to access the data directly (we did this in the practical too).

a) Start by using FastQC (Tools and Services -> Utilities -> Fastq utilities) on the data from accession number SRR23290383 to carry out some basic quality control. Use the html report you get back from FastQC to answer the following questions:

1. What is the length of the reads in this dataset? Are all reads in the fastq files of this dataset the same length? **[1 mark]**

2. Is the quality of the reads good? Justify your answer. (*Hint: check the "per base sequence quality" module*). **[1 mark]**

3. FastQC reports a failure of the "per base sequence content" module for both forward and reverse reads. What do you think is the most likely origin of that failure? *Hint: Examine the rest of the fastqc report and this should give you a clue about the origin of the problem.* **[2 marks]**

4. Based on the report from FastQC, what do you expect the %GC content of this virus to be? **[1 mark]**

b) Use the SARS-CoV-2 Genome Analysis pipeline on the BV-BRC server under Tools and Services) to assemble the viral genome from this sample (SRR23290383). In the practical session, we used the Comprehensive Genome Analysis, but this is set up for bacteria only and should not be used for data from viruses.

When your genome is assembled, use the report returned by BV-BRC to answer the following questions:

1. The BV-BRC server sends the assembled genome to the Pangolin server to get a prediction of the SARS-CoV-2 lineage to which the genome belongs. Report the lineage and variant (Greek letters are assigned to variants e.g. alpha, delta etc) predicted for the genome you just assembled. **[2 marks]**

2. Which protein or open reading frame is predicted to contain the largest number of UNIQUE non-synonymous SNPs in this variant compared with the reference sequence for SARS-CoV-2? **[2 marks]**

3. The report returned by BV-BRC describes insertions as well as SNPs in the genomic assembly of SRR23290383 compared with the reference. One of the reported insertions is at position 11287. Which gene/ORF/protein does this correspond to? **[1 mark]**

   (*Hint: you can use the annotation of features to answer this question*).

4. Download the fasta file for the assembly (you should have a **.**fasta file in the directory where the results of the assembly were saved) and use that file together with the reference sequence (NC_045512.2; you can get this from the NCBI Nucleotide database as we did before in the practical) to obtain a global pairwise alignment with the EMBOSS tool "stretcher" (https://www.ebi.ac.uk/Tools/psa/emboss_stretcher/).

   Upload a screenshot of:

   i) the header information of the alignment where the sequence names, gap/extend penalty and results (length, identity, similarity, gaps and score) are reported **[1 mark]**

   ii) a screenshot near position 11287 to show that the insertion reported by BV-BRC is indeed there. **[1 mark]**

Question 2. [3 marks]

In the data folder uploaded on Moodle, you are provided with three sequences from three major variants of SARS-CoV-2: an alpha, a delta and an omicron variant. You are also provided with four "mystery" genomes that were assembled from samples collected at different times and places since the start of the pandemic.

Use the Viral Genome Tree service of the BV-BRC server to obtain a phylogenetic tree of the seven genomes listed above (three known and four mystery genomes). Leave all parameters to default and upload the fasta files from your computer selecting the "unaligned fasta" option.

a) Upload a screenshot of your phylogenetic tree (labels should be clearly visible on the branches). **[2 marks]**

b) Mystery genome 3 does not belong to any of the three known variants I provided you example sequences for. Use the phylogenetic tree to predict which of the three variants (alpha, delta or omicron) is closest to mystery genome 3. Check this result against the pairwise sequence identity matrix. **[1 mark]**

Question 3. [5 marks]

In question 1b, you assembled a SARS-CoV-2 genome with the help of the BV-BRC server. In the report accompanying the assembly, the server listed a number of non-synonymous SNPs for all predicted features (genes) in the genome. First, find and report all SNPs predicted for the spike (S) protein in this assembly. Next, open the PDB structure 6lzg in Chimera (this is a structure of the spike receptor-binding domain complexed with human ace2) and highlight all residues that are predicted to have mutations in the genome you assembled. Note that this PDB structure contains only part of the spike protein sequence so you may not be able to find all mutated residues in this structure.

Create and upload a clear, high-quality image of the structure with all mutated residues highlighted (you do not need to change the original residue – simply highlight it using, for example, a different colour and/or representation). You are free to use any representation you want but the image should be of publication quality and residues carrying mutations should be clearly visible. Accompany your image with a suitable figure caption. The caption should include the list of amino acids highlighted as carrying mutations.

Question 4. [3 marks]

Obtain the **spike protein** (amino acid) FASTA sequence for the SARS-Cov-2 reference genome (NC_045512.2) from the NCBI Nucleotide database.

Hint: find the spike entry in the Genbank record and download the sequence; alternatively, you can get the sequence from the corresponding Uniprot entry.

Go to NCBI's BLAST server, exclude all SARS-CoV-2 (taxid:2697049) sequences and search the nr database with the sequence of the spike protein from the SARS-CoV-2 reference genome. Leave all other parameters to default. Answer the following questions:

a) Surprisingly, among the results returned there are hits to Homo sapiens (human) sequences. Examine the hits and explain why that is.
[1 mark]

b) Among the hits in the database, there are several hits to pangolin coronavirus sequences. Choose “Descriptions” in the results page, order the descriptions by “scientific name” in order to easily select the pangolin hits and then click on the MSA viewer at the top to obtain a multiple alignment of all sequences from pangolin. On the page showing the multiple alignment, choose Coloring -> Show Differences to get a clearer view of the positions that are most commonly mutated in the pangolin coronavirus, relative to the SARS-CoV-2 reference spike protein. There is a region of about 300 nucleotides in the pangolin sequences that has almost no mutations compared with the SARS-CoV-2 sequence.

i) Identify the region of SARS-CoV-2 that is highly conserved in the pangolin sequences and upload a screenshot of the alignment to justify your answer (approximately, e.g. your answer could be “from approximately residue 200 to approximately residue 500”) **[1 mark]**

ii) highlight this region in the EM-determined structure of the near-complete sequence of SARS-CoV-2, PDB id: 6zb5 (submit as your answer a high-quality image produced in Chimera, highlighting this conserved part of the sequence). **[1 mark]**