BINF200 Assignment 1

Tom Michoel

2023-09-14

10 points, 10% of the final grade

1 Introduction

In this assignment you will analyze data from the following paper:

S. Jayaraman *et al.* (2019), Application of long read sequencing to determine expressed antigen diversity in Trypanosoma brucei infections, PLOS Neglected Tropical Diseases 13(4): e0007262.

2 Software

You will need the following ingredients to solve the assignment:

1. **BLAST+**: Install the NCBI command line standalone BLAST+ programs following the online instructions. (Note: there is also an Ubuntu "ncbi-blast+" package not listed on the NCBI page).
2. **GetORF**: Install the getorf tool.
3. **Programming language**: You may use any language: Python, Julia, R, …
4. **Data analysis and visualization, report writing**: Use JupyterLab (or notebooks) and/or an IDE such as Visual Studio Code for data analysis and visualization. Using Quarto you can integrate your notebooks, data visualization, and final report all in a single document.

3 Sequence data

All data for the assignment are available on UiB OneDrive in the **PacBio VSG** folder. You will need to login with your UiB account to have access. You will need:

1. A database of known VSG genes. Download the entire folder **TREU927-v26_VSGTranscripts**.
2. Sequencing data from **one** sample (**one** file) named **PacBio_VSG_filtered_reads_*sample_name*.fa** Twenty such files, for 20 samples (individual mice), are available in total. To find out which one *you* should download:
    1. Go to the **Groups** page on Mitt (https://mitt.uib.no/courses/42444/groups)
    2. Select the **Compulsory Assignment 1 - Sample** tab and find to which group you have been (randomly) assigned.
    3. Each group is named **Compulsory Assignment 1 - Sample *k***, where k is a number from 1 to 20 mapping to the sample names as follows:

Table 1: Mapping of group labels to sample names

k	Sample name	k	Sample name
1	balbc_3_0	11	balbc_10_0
2	balbc_3_1	12	balbc_10_1
3	balbc_3_2	13	balbc_10_2
4	balbc_3_3	14	balbc_10_4
5	balbc_3_4	15	balbc_10_5
6	balbc_6_0	16	balbc_12_1
7	balbc_6_1	17	balbc_12_2
8	balbc_6_2	18	balbc_12_3
9	balbc_6_4	19	balbc_12_4
10	balbc_6_5	20	balbc_12_5

4 Tasks

4.1 Show that you understand the BLAST+ package

The BLAST+ package contains five **core blast search programs**.

1. List all five core blast search programs.
2. Explain the difference between **blastn** and **blastp**.

4.2 Show that you understand the biological experiment and data

Write a short paragraph in your own words to explain the biological experiment that was done to generate the sequence data (see the publication), what sequences are contained in the **PacBio_VSG_filtered_reads_*sample_name*.fasta** files, and what sequences are contained in the **TREU927-v26_VSGTranscripts** database.

4.3 BLAST the sample sequences against the reference VSG database

The relevant BLAST command can be found on the longread-application repository. Adapt this command in the following ways:

1. Change the input file name to *your* assigned sample file (see Table 1).
2. Change the output file name to something containing *your* sample label (see Table 1).
3. What does the paramater `-max_target_seq` do and why was it set to 1?
4. Use a tabular output format *without* comments that includes *only* the following columns:
    - Query sequence id
    - Subject sequence id
    - Raw score
    - Bit score
    - E-value
    - Query sequence length
    - Subject sequence length
    - Alignment length
    - Start of alignment in subject
    - End of alignment in subject
    - Number of identical matches
    - Number of mismatches
    - Total number of gaps
    - Percentage of positive-scoring matches
5. What is the number of sequences in *your* input file? **Hint:** Most popular programming languages have packages that can parse FASTA files automatically.
6. What is the number of sequences in *your* output file? Is it the same as in the input file? **Hint:** Import the blastn output file into a DataFrame.
7. We define the alignment coverage as the percentage of the subject sequence covered by the alignment. Compute the alignment coverage for all sequences from the blastn output. Count and remove alignments with coverage less than 60%.

8. The bit score S' is derived from the raw score S using the formula

$$S' = \frac{\lambda S - \ln K}{\ln 2}$$

Can you find the values of λ and K from your blast results?

4.4 Count VSG expression levels

From the previous BLAST results:

1. Extract the unique VSG ids in *your* sample.
2. For each unique VSG: count the number of sequences aliging to that VSG, and the average number of identical matches, mismatches, and gaps for its alignments. **Hint:** Using the split-apply-combine strategy, these numbers can be computed in one line of code.
3. Identify the 10 most abundant VSGs in *your* sample and visualize their relative expression levels in a pie chart and compare your result against Figure 3 of the paper.
4. Write a paragraph in your report that describes the figures and your interpretation of them.

4.5 Identify open reading frames

The relevant `getorf` command can be found on the longread-application repository.

1. Count the percentage of reads in *your* sample that result in a predicted ORF with a minimum size of 1200 nucleotides.
2. Which explanation was proposed in the paper for this low percentage?