Assignment #1: Mapping

16-09-22

This assignment consists of four parts, which will cover the curriculum presented in the first week of the course. The first two parts are pen and paper exercises, and the last two are computer exercises using the ricco server. The data is located here (you will need to copy it into your own folder) `/TEACHING/BIOINF22/assignment1_mapping`.

If you do not have access to the ricco server contact course responsible immediately

0.1 Exact alignment (Rabin-Karp)

Consider the text $ACGTAACGTAACGA$ and pattern $TAACG$

1. What is the length of the strings and what is the size of the alphabet?
2. How many shifts and how many comparisons do we need to do in order to perform an exhaustive naive search?
3. Consider the Rabin-Karp algorithm and define a rolling hash function.
4. Explain why your hash function is a rolling hash function.
5. Compute the hash of the pattern and the hash of each shift.
6. How many operations did the exhaustive search with the Rabin-Karp algorithm require?

0.2 Data Structures

The goal with part of the assignment is to get familiar with the string operations needed to construct the data structures which several modern day alignment algorithms use.

Consider the sequence $T =$ "$DINGELINGDING$" And the query sequence $S =$ "ING"

1. Draw the suffix tree for T.
2. Using tree terminology, describe how you can identify the number of times the query sequence S occurs in sequence T.
3. From sequence T, generate a table consisting of two columns, with the first containing the suffix array and the second the corresponding suffix *(Hint: see slide 22, lecture* `SuffixArraysBWT_2022.pdf`*.*
4. Identify the position for which the query sequence S occurs within the sequence T.
5. Generate the Burrows-Wheeler-Transform of sequence T.

0.3 Mapping Statistics

There are two fastq.gz files in the directory `/TEACHING/BIOINF22/assignment1_mapping`.

One is called `L10.fq.gz` and contains single end reads of length 10, and one is called `L30.fq.gz` and has reads of length 30. You will align both of these files and investigate the effect of read length on the resulting mapped reads. Both of these files are from the organism *Mycobacterium Leprae*. You will also find the reference sequence to map against `Mycobacterium_leprae.fa.gz`.

Due to the simulation software used, the read ID contains the information of the positions for which the sequence originates. Once aligned, there can potentially be differences between the origin of the chromosomal position and the aligned position due to mis-alignments.

e.g. `@T0_RID0_S0_NZ_CP029543.1:2084294-2084323_length:30_R1` is a read (ID) originating from chromosome `NZ_CP029543.1` from position `2084294-2084323`

We can use this fact to investigate whether the reads generated are mapping to the correct position by comparing the position in the read ID to the position in the bam file.

For each question that generates an output, the files generated can be found here `/TEACHING/BIOINF22/assignment1_mapping/output/part3`. These can be used for later questions if you aren't sure about your earlier answers. It is therefore important that for each question, you must **supply the command you used.**

Inspiration and hints can be found in https://github.com/ANGSD/adv_binf_2022_week1/tree/main/day2 and https://github.com/ANGSD/adv_binf_2022_week1/tree/main/day1.

1. How large (in bases) is the bacterial reference genome?
2. Use `bwa aln` (and `bwa samse`) to align the two fastq files to the bacterial reference *(Hint: look at the exercises from day 2)*.
3. Sort and index the resulting files using `samtools`, and make sure they are saved in bam format.
4. Filter out the unaligned reads and create new bam files. Identify which flag to filter out on https://broadinstitute.github.io/picard/explain-flags.html
5. By looking at the reads aligning and the chromosomal positions, we can calculate how many reads are mapped correctly with a 0 nucleotide difference between the origin in the read ID and the mapping coordinate. The script `/TEACHING/BIOINF22/assignment1_mapping/get_stats.sh` takes a bam file and outputs two numbers - firstly the number of reads that map correctly (i.e. those where the start position in the bam file matches the read ID), and then the number that do not map correctly. Use this script to find out how many reads map correctly and incorrectly in the two bam files (e.g. `bash /TEACHING/BIOINF22/assignment1_mapping/get_stats.sh L30.bam`)
6. Now create two new bam files where you filter the reads so we only retain reads with a mapping quality of greater than or equal to 1. *(Hint: look at the day 2 exercises)*
7. Repeat the exercise in question 5 with the two new filtered bam files. How do the numbers differ?
8. Is the proportion of incorrectly mapped reads greater or smaller in the original bams or the quality filtered bams? Why do you think that is?
9. Given the results here, do you think it is relevant to include mapping filters in downstream analysis?

0.4 Ancient Data Analysis

In this part you are given two .fq.gz files in the directory `/TEACHING/BIOINF22/assignment1_mapping`, namely `Hercule.fq.gz` and `Poirot.fq.gz`. One of these files is sequencing data from an ancient bacteria, and one is from a modern bacteria. There were some troubles in the lab, and we have no record of which one is which. Pseudomonas_aeruginosaYour goal is to ascertain which one is which.

For each question that generates an output, the files generated can be found here `/TEACHING/BIOINF22/assignment1_mapping/output/part4`. These can be used for later questions if you aren't sure about your earlier answers. It is therefore important that for each question, you must **supply the command you used.**

Inspiration and hints can be found in https://github.com/ANGSD/adv_binf_2022_week1/tree/main/day2 and https://github.com/ANGSD/adv_binf_2022_week1/tree/main/day1.

1. Briefly describe what characterizes ancient DNA from modern DNA
2. Perform adapter trimming on the provided sample sequence files using fastp. Make sure to discard reads shorter than 30 bp. (-l parameter)
3. How many reads contained adapters in both datasets? *(Hint: Look at the output from fastp)*
4. What is the mean length of the reads before and after trimming? *(Hint: There is a command for this near the bottom of the day 2 exercises)*
5. Perform bwa alignment using `aln` and `samse`. For each sample, sort the sam file, save it as a bam, and index it. Remember, if you are struggling you can find the output for this question in `/TEACHING/BIOINF22/assignment1_mapping/output/part4/q5`. **Note that the reference is not the same as the last part. It is `/TEACHING/BIOINF22/assignment1_mapping/Pseudomonas_aeruginosa_PAO1_107.fna.gz`**
6. For each sample, create new bam files with just the aligned reads. What proportion of reads remain?
7. What is the average depth of each sample (aligned reads only)? Use `samtools depth Hercule.bam|datamash mean 3`
8. Use `mapDamage` to identify the nucleotide misincorporation and fragmentation patterns, and describe (and include) the output plots generated by MapDamage found in the files Fragmisincorporation_plot.pdf and Length_plot.pdf. Make sure to use the command line options `--merge-libraries --no-stats`
9. Based on your results obtained from the previous questions, which of the files looks ancient and which one looks modern? Why?