University of TEHRAN

Introduction to Bioinformatics

Project 1- Read mapping and genome assembly

Fall 2023

Due date: Saturday 02 December 2023, 11 Azar 1402

Objective

In this project, you will learn how to analyze short-read sequencing data, perform quality control, create genome assemblies, and conduct read mapping.

Part A - Downloading E. Coli WGS data, preliminary analyses, and quality controls

First, download the data according to the following steps and then proceed with the subsequent analyses.

1. Download SRR8185316 (short-read WGS of E.coli) from SRA using the SRA Toolkit in Linux or Windows. (Hint: you may use fastq-dump [options] <accession>).
2. Answer the following questions about the short reads fastq file (you can use existing R packages such as ShortRead, or similar python packages):

    I. How many reads are in the fastq file?

    II. Print the identifier, quality, and sequence of the first read of the fastq file.

    III. How many times does the TTAAATGGAA subsequence appear in the file?

    IV. Extract the first 1000 sequences of the fastq files (4000 lines).

    V. Plot the quality of the reads in the fastq file using a box plot.

    VI. Show the distribution of read lengths using a density plot.

3. Perform quality control of the reads using FastQC and interpret the results.

Part B - De novo genome assembly

Install SPAdes, Quast, BWA, Samtools, or any other alternative software of your choice. Afterwards, follow the steps provided below.

1. Run SPAdes to generate draft genome assemblies from short reads.
2. Assess the quality of the draft genome assembly using Quast and compare it to the reference genome (Download the reference genome from here)

Part C - Read mapping

1. Map the Illumina short-read data to the reference genome using BWA.
2. Print the head of the obtained SAM file from previous question. Explain what you see for the first hit (you can do this step either in Linux or R).
3. Convert the SAM file to an indexed BAM file. Hint: use samtools view, samtools sort, samtools index.
4. Use the Integrative Genomics Viewer (IGV) to visualize the mapped reads in a 200-b genomic region of your choice. Select the reference genome fasta and GTF file (GTF is optional).
5. Determine the percentage of short reads that are mapped to the reference genome. Hint: use samtools flagstat.
6. Get the read depth for the sorted BAM file at all positions of the reference genome and report the mean of all reads. Hint: use samtools depth.
7. Make yourself familiar with the CIGAR format. How do you interpret the following expressions?

   “29S21=1X25=”

   “20M2I1M1D10M”

   “5M10N25M”

Important note: please ensure that all results are organized within a directory named Project1. This directory should contain subdirectories for the input data, QC results, the output of mapping to the reference genome, de novo assembly outputs, and the alignment results of the assembled genome with the reference genome.