Actividad 2: Manipulación y formateo de archivos: Formato FASTQ y FASTA

Objetivo

El propósito de esta actividad es desarrollar un flujo de trabajo bioinformático integral, conocido como *pipeline*, para procesar datos biológicos. La intención es que el estudiante adquiera destrezas para interactuar con el sistema operativo mediante la línea de comandos y pueda desarrollar scripts en Shell para abordar diferentes desafíos bioinformáticos, centrándose en dos formatos de texto específicos: **FASTQ y FASTA.**

Detalles sobre la entrega

- La entrega se realizará utilizando este documento como plantilla; adicionando capturas de pantalla que ilustren el código empleado y su ejecución. Será esencial incluir el nombre de usuario completo (*prompt*) en las capturas de pantalla y se recomienda ajustar la resolución al máximo posible.
- Para cada comando empleado, deberá realizar una **breve explicación** donde indique su función y describa las opciones o parámetros utilizados. Para facilitar la lectura, emplee un color o formato de letra distinto al del enunciado propuesto.
- La entrega se realizará a través del Campus VIU, en un archivo **PDF** único descomprimido.

Parte I: Creación del directorio de trabajo.

Para inicializar este ejercicio, deberá crear y organizar un nuevo directorio de trabajo que contenga los elementos clave que se explorarán. La estructura de organización busca que cualquier persona no familiarizada con el proyecto pueda examinar los archivos en la computadora y entender en detalle lo realizado y por qué.

Utilizando comandos de Linux en la terminal, deberá crear la siguiente estructura de directorios para el proyecto, identificándolo de la siguiente manera:

User_proyect_year_publication donde:

- **User**: Corresponde al apellido del estudiante.
- **Project:** Es el nombre del proyecto hipotético en el que está trabajando.
- **Year**: Representa el año de la publicación o investigación en curso.
- **Publication**: Es el nombre de la revista donde planea publicar sus resultados.

Este directorio, en mi caso, ***Soler_humancells_2024_Nature*** será creado dentro del directorio ***Documents,*** que a su vez contendrá tres directorios ***data, code*** y ***submission***. Además, el directorio ***data*** tendrá a su vez dos subdirectorios llamados ***raw*** y **processed.**

Parte II: Obtención de datos

Ahora procederá a obtener el conjunto de datos con el cual trabajará. El archivo en cuestión puede encontrarse en la siguiente ruta del campus virtual y

Actividades/Portafolio de pruebas aplicativas/Prueba aplicativa 2/SRR1984406_1.fastq

Seguidamente, guarde este archivo en el directorio ***data/raw*** y responda a las preguntas indicadas, adjuntando los comandos empleados y capturas de pantalla.

- **P1.** ¿Cuántas secuencias podemos encontrar en el archivo? **(0,5 pts)**

- **P2.** ¿Cuál es la longitud mínima y máxima de las secuencias incluidas en el archivo? ¿Cuántas secuencias tienen una longitud igual al valor de longitud mínima previamente computada? **(1,5 pts)**

- **P3.** ¿Cuántas veces aparece el patrón "ATGATG" en las secuencias del archivo descargado? **(0,5 pts)**

- **P4.** Con el propósito de transformar el archivo de formato FASTQ a FASTA, deberá seleccionar las líneas correspondientes que representan el encabezado (header) y la secuencia de cada lectura. El resultado se almacenará en un archivo llamado *all_sequences.fasta* dentro del directorio *data/processed*. Incluya una captura de pantalla con el código empleado para realizar esta conversión de formato y visualice las 4 primeras líneas del archivo *all_sequences.fasta* en el directorio determinado. **(1,5 pts)**

  FALTA HACERLO EN FORMA DE SCRIPT

- **P5.** Posterior a la conversión, seleccione las primeras 5 lecturas del archivo all_sequences.fasta (con su cabecera y su secuencia asociada) y cree archivos individuales denominados secuencia1.fasta, secuencia2.fasta, y así sucesivamente. En cada uno de ellos, deberá modificar la cabecera, por Secuencia 1, Secuencia 2 y así, sucesivamente (no emplee ningún editor de texto manual para realizarlo).

  Como ejemplo, cada uno de ellos contendrá lo siguiente:

  >Secuencia 1
  GACGACTGCCATCTGAACGTGTGGAATCAACGGAGCCACATCTGACTTCCAGTATCCATCCGAAGTTCTC
  CATTCAATAGTGAGGAATCTGACGACTGCCATCTGAACGTGTGGAATCAACGGAGCCACATCTG

  Incluya una captura de pantalla con el código empleado y su ejecución, visualice el contenido de los archivos creados y con el comando *ls* muestre la creación de estos en el directorio *data/processed* **(1,25 pts)**

- **P6**. Ahora deberá crear un script llamado *analyze_sequences.sh* en el directorio *code.* Este script deberá leer cada uno de los archivos, específicamente su secuencia, y reportar por la salida estándar estadísticas detalladas de las secuencias:
    - Longitud de cada secuencia.
    - Identificación del nucleótido inicial y final de cada secuencia.
    - Porcentaje de contenido de GC para cada secuencia. Redondee el resultado a dos decimales.
    - Determine con una estructura condicional si en alguna de las secuencias se encuentra el patrón "GGGG". Si no lo encuentra reporte un mensaje de negación al usuario y si lo encuentra, reporte las secuencias donde lo ha encontrado y sustituya este patrón por la palabra "Mutation".

  Adicione una captura de pantalla que muestre la estructura del script, su ejecución y los resultados obtenidos **(2,5 pts)**

Pero se refiere a que cada uno se debe guardar en varios archivos distintos nuevos o todas las secuencias en un nuevo archivo? y hay que guardarlo con la cabecera?

- **P7**. Implemente otro script llamado *Metamorphosis.sh* en el directorio *code.* Este script deberá leer cada uno de los cinco archivos e invertir el orden de las secuencias de nucleótidos en cada registro fasta. Después, deberá convertir todas las secuencias a minúsculas y escribir las secuencias procesadas en un nuevo archivo fasta con un nombre que indique que se ha invertido y convertido a minúscula. Adicione una captura de pantalla que muestre la estructura del script, su ejecución y el contenido de los resultados obtenidos. **(2 pts)**

- **P8.** Para concluir el proyecto, genere una copia de esta actividad cumplimentada en formato PDF y trasládela al directorio submission. Además, incluya una captura de pantalla con la estructura de directorios actual usando el comando *tree* desde el terminal **(0,25 pts).**