Actividad 1.- Manipulación de Archivos en Formato BED

El objetivo de esta actividad es que el estudiante adquiera habilidades en la manipulación y formateo de archivos usando comandos de Linux, aprendidos en las sesiones teóricas de la asignatura. En particular, se enfocará en el formato **BED** (*Browser Extensible Data*) que se utiliza extensamente en bioinformática para almacenar regiones genómicas, como coordenadas y anotaciones asociadas. Este formato se caracteriza por presentar los datos en forma de columnas separadas por espacios o tabuladores.

Instrucciones de entrega

- La entrega se realizará a través del Campus VIU en un archivo único en formato PDF utilizando este documento como plantilla. Recuerde que las actividades a realizar están resaltadas en negrita.
- Incluya el código empleado, capturas de pantalla con su usuario (agregando el *promp* completo) y resolución máxima.
- Proporcione explicaciones claras y concisas de los comandos utilizados. Si los comandos empleados no se explican brevemente, el valor de la pregunta será penalizado a la mitad.
- Reporte solo una opción o forma para resolver cada una de las preguntas propuestas.

Obtención de los datos.

Los datos con los que va a trabajar se refieren a regiones de interés detectadas en células inmunitarias en humanos. Para ello, se han realizado dos réplicas técnicas del experimento, obteniendo dos archivos llamados human_coordinates_1.bed y human_coordinates_2.bed. Además de estas regiones, se seleccionaron genes candidatos a testear experimentalmente y que pueden encontrarse en el archivo selected_genes.txt. Estos tres archivos están disponibles en la propia actividad propuesta en el campus virtual:

- Actividades/Portafolio de pruebas aplicativas/Prueba aplicativa 1/human_coordinates_1.bed
- Actividades/Portafolio de pruebas aplicativas/Prueba aplicativa 1/human_coordinates_2.bed
- Actividades/Portafolio de pruebas aplicativas/Prueba aplicativa 1/selected_genes.txt

Actividades a realizar

1. Descargue los archivos anteriores en su entorno de trabajo de AWS (no emplee el comando *wget*, realice la descarga mediante la interfaz gráfica de la Universidad) (1,5 pts).

- **Determine cuántas líneas presenta cada archivo descargado.**
- **Determine el número total de líneas vacías en cada archivo (si las hay) y elimínelas. Genere archivos nuevos sin líneas vacías.**

A partir de este punto siempre deberá trabajar con los archivos sin líneas en blanco.

2. Visualice específicamente las líneas 2, 500 y 1500 del archivo human_coordinates_1.bed, adicionando los números de líneas correspondientes, tal y como se muestra en el siguiente ejemplo:

2 chr3 62796234 62796433

Incluya el código empleado para realizarlo junto a una captura de pantalla del resultado (0,5 pts)

3. Calcule el número mínimo y máximo de columnas que encontramos en cada uno de los archivos (1,25 pts)

4. Seleccione el archivo human_coordinates_1.bed para contestar las preguntas siguientes. ¿Cuántas coordenadas únicas se asocian a cada cromosoma? Genere un archivo nuevo con este resultado y ordene los cromosomas de menor a mayor atendiendo a este valor computado. ¿Tenemos representación de todos los cromosomas humanos? ¿Cuál o cuáles faltan? (2 pts)

5. Los archivos human_coordinates_1.bed y human_coordinates_2.bed son réplicas experimentales y, por tanto, esperaríamos que ambos archivos fueran idénticos. **Para comprobarlo, primero ordene los dos archivos por el nombre del cromosoma y las coordenadas de inicio. Seguidamente, compárelos para computar cuántas y qué regiones son distintas entre ambos archivos. Una vez identificadas estas regiones, las debe guardar en un archivo nuevo (solo las tres columnas, cromosoma, coordenada de inicio y coordenada de fin; no emplee ninguna edición manual para realizarlo). Visualice este archivo creado (2 pts)**

6. Ahora va a transformar el formato de estas coordenadas genómicas diferenciales almacenadas. Para ello, debe sustituir el primer tabulador por dos puntos y el segundo por un guion; de forma que las coordenadas presenten la siguiente estructura: chr:inicio-fin. Fíjese en el ejemplo:

- Formato inicial: chr6 20978845 20979044
- Formato final: chr6:20978845-20979044

Incluya una captura de pantalla con el código empleado visualizando el cambio de formato de las regiones (no emplee ningún editor de texto) (1 pts).

Una vez que tenga las regiones seleccionadas con el formato correcto, las deberá caracterizar e identificar para conocer qué genes alberga en su interior. Para ello, deberá acceder al siguiente navegador genómico alojado por la Universidad de California, Santa Cruz: https://genome.ucsc.edu/. Una vez allí, se situará en el menú denominado "**Genomes**" (parte superior derecha) y seleccionará el *assembly* actual y de referencia del genoma humano denominado **Human GRCh38/hg38**. Al dar clic en él, se abrirá un sitio web interactivo donde podrá pegar cada una de las regiones detectadas para identificar qué genes se encuentran en dichas coordenadas genómicas.

7. Adjunte una captura de pantalla (como la que se muestra a continuación) para cada una de las regiones encontradas previamente donde se visualice la región y el o los genes que se encuentran en ella (0,75 pts)

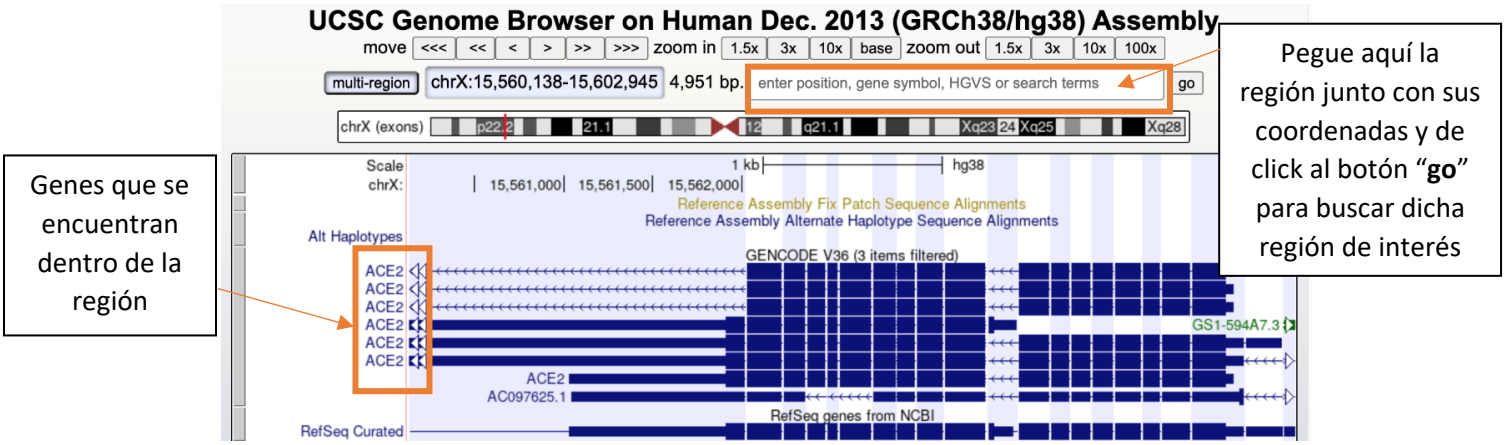

Figura 1. Vista del UCSC Genome Browser.

8. Finalmente, del archivo llamado selected_genes.txt, deberá seleccionar aquellos genes que ha obtenido en cada una de las búsquedas realizadas y añadirlos a un archivo final; donde incluya en la primera columna las regiones identificadas previamente con el formato, *cromosoma:inicio-fin,* una segunda columna con el nombre del gen que ha detectado en cada una de ellas y una tercera columna donde indique el número de línea donde ha encontrado el gen en el archivo selected_genes.txt. Incluya una captura de pantalla con el código empleado y el archivo final generado (no emplee ninguna edición de texto manual para realizarlo) (1 pts).