**Nombre y Apellido: Juan Carlos García Estupiñán**

**Actividad 1.- Manipulación y formateo de archivos: Formato BED**

**Objetivo**

El objetivo de esta actividad es que el estudiante aprenda a manipular y formatear archivos empleando diferentes comandos de Linux aprendidos a lo largo de las sesiones. Específicamente, se trabajará con el formato BED (*Browser Extensible Data*) que se utiliza ampliamente 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.

**Parte I: Obtención de los datos.**

Los datos con los cuales va a trabajar hacen referencia a una serie de regiones de interés detectadas en un tipo de células inmunitarias, las células B, en humanos. Para ello, se han realizado dos réplicas del experimento, obteniendo dos archivos llamados human\_coordinates\_1.bed y human\_coordinates\_2.bed. Estos dos 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

**Acceda a la ruta anterior y descárguese los datos en su entorno de trabajo. Visualice las 5 primeras líneas de cada uno de los archivos. Incluya el código empleado para realizarlo junto a una captura de pantalla (0,5 pts)**

* **Script: 1head.sh**

Texto

Descripción generada automáticamente

* **Stdout: 1head.sh**

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

Seguidamente, responda a cada una de las preguntas que se le indican, adicionando siempre una captura de pantalla con los comandos empleados y la respuesta obtenida por la salida estándar.

* **¿Cuántas líneas presenta cada uno de los archivos descargados? (0,5 pts)**
  + **Script: 2nrow.sh**

Texto

Descripción generada automáticamente

* + **Stdout: 2nrow.sh**

Texto

Descripción generada automáticamente

* **¿Cuántas columnas presenta cada uno de los archivos descargados? (1 pts)**
  + **Script: 3ncol.sh**

Texto

Descripción generada automáticamente

* + **Stdout: 3ncol.sh**

Texto

Descripción generada automáticamente

* **¿Tenemos representación de todos los cromosomas humanos en ambos archivos? (1 pts)**
  + **Script: 4lookchr.sh**

Escala de tiempo

Descripción generada automáticamente con confianza media

* + **Stdout: 4lookchr.sh parte 1**

Texto, Carta

Descripción generada automáticamente

* + **Stdout: 4lookchr.sh parte 2**

Texto, Carta

Descripción generada automáticamente

Al ser réplicas experimentales esperaríamos que ambos archivos fueran idénticos. **Para comprobarlo, primero ordene los dos archivos por el nombre del cromosoma (determinado en la primera columna). Seguidamente, compárelos para mostrar qué regiones son distintas entre ambos. Adjunte una captura de pantalla con los comandos empleados que muestren cuántas y qué regiones son distintas entre ambos archivos (2 pts)**

* **Script: 5diff.sh**

Texto

Descripción generada automáticamente

* **Stdout: 5diff.sh**

Texto, Carta

Descripción generada automáticamente

**Una vez identificadas estas regiones, las debe seleccionar y guardarlas en un archivo nuevo. Ojo solo tiene que guardar las tres columnas, cromosoma, coordenada de inicio y coordenada de fin de cada una de las regiones detectadas. Visualice las primeras líneas de este archivo creado. Incluya una captura de pantalla que muestre el código empleado (1 pts)**

* **Script: 6diff\_processed.sh**

Texto

Descripción generada automáticamente

* **Stdout: 6diff\_processed.sh**

Texto

Descripción generada automáticamente

Ahora va a transformar el formato de estas coordenadas genómicas 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 (1,5 pts)**

* **Script: 7format\_change.sh**

Escala de tiempo

Descripción generada automáticamente con confianza media

* **Stdout: 7format\_change.sh**

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

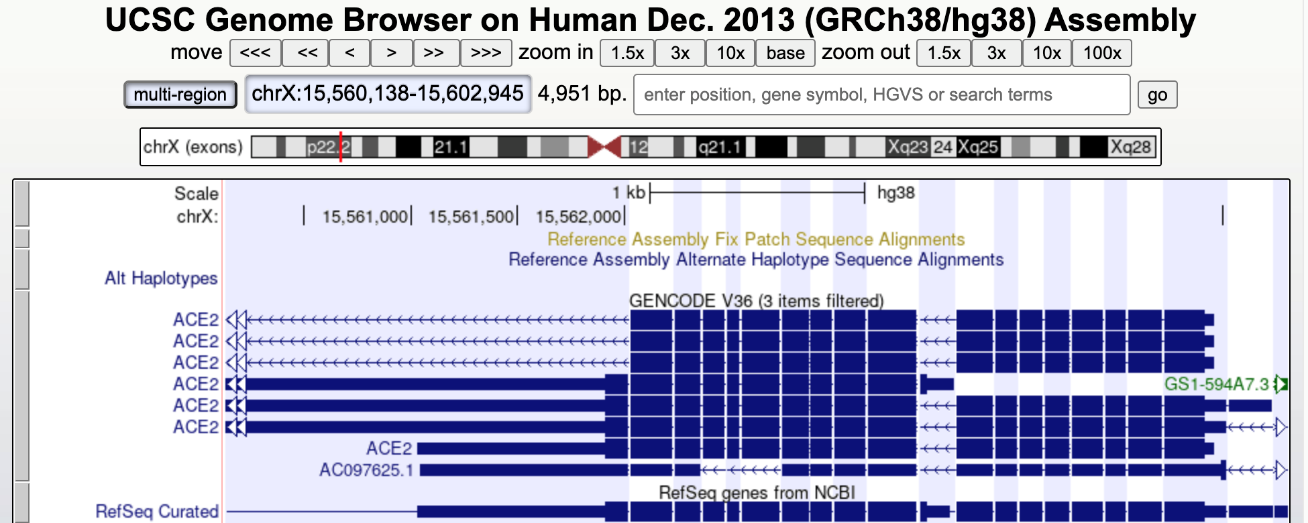
Una vez que tenga las regiones seleccionadas con el formato correcto, las deberá caracterizar e identificar para conocer qué genes alberga en su interior. Para ello, deberá acceder al siguiente navegador genómico alojado por la Universidad de California, Santa Cruz: <https://genome.ucsc.edu/>. Una vez allí, se situará en el menú denominado “**Genomes**” (parte superior derecha) y seleccionará el *assembly* actual y de referencia del genoma del ser humano denominado **Human GRCh38/hg38**.

**¿A qué nos referimos cuando hablamos del *Human Genome Assembly*?. (0,5 pts)**

* El ensamblaje genómico (genome assembly) se refiere al proceso de poner secuencias de nucleótidos en el orden correcto. En el caso de Human Genome Assembly, se referiría al esamblaje genómico humano (<https://www.sciencedirect.com/topics/agricultural-and-biological-sciences/genome-assembly>).

Al dar click 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. **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 (1 pts)**

Pegue aquí la región junto con sus coordenadas y de click al botón “**go**” para buscar dicha región de interés



Genes que se encuentran dentro de la región

Figura . Vista del UCSC Genome Browser.

Imagen que contiene Aplicación

Descripción generada automáticamente

Figura 2. Vista del USC Genome Browser: chr1:204073115-204127743.

Escala de tiempo

Descripción generada automáticamente

Figura 3. Vista del USC Genome Browser: chr6:31164337-31170682.

Interfaz de usuario gráfica, Aplicación, Tabla, Excel

Descripción generada automáticamente

Figura 4. Vista del USC Genome Browser: chr17:42313412-42388540

**Finalmente, cree un archivo final, donde incluya en la primera columna las regiones identificadas previamente con el formato, *cromosoma:inicio-fin* y una segunda columna con el nombre del gen que ha detectado en cada una de ellas. Visualice las primeras líneas del archivo creado. Incluya una captura de pantalla con el código empleado (1pts).**

* **Script: 8coord\_genes.sh**

Texto

Descripción generada automáticamente

* **Stdout: 8coord.\_genes.sh**

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Texto

Descripción generada automáticamente

**En caso de querer ver el código utilizado puede pinchar este enlace:** [**https://github.com/Juankkar/Programacion\_Shell\_Scripting\_VIU/tree/main/actividad1**](https://github.com/Juankkar/Programacion_Shell_Scripting_VIU/tree/main/actividad1)**.**