

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.

Detalles sobre la entrega

- La entrega se realizará utilizando este documento como plantilla; adicionando el código y las capturas de pantalla asociados en los espacios determinados a ese uso (puede incrementar o reducir el tamaño de los mismos). Recordar que las actividades a realizar están resaltadas en negrita.
- Será esencial adicionar en las capturas de pantalla su usuario (adicionar el prompt completo) e intentar que la resolución de las mismas sea el máximo posible.
- La entrega se realizará a través del Campus VIU en un archivo único en formato PDF.

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)

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.

los comandos se encuentran en el script: `code/actividad1.sh`

- **¿Cuántas líneas presenta cada uno de los archivos descargados? (0,5 pts)**

INPUT

```
#!/usr/bin/env bash

##### Parte I de la actividad
# Empezaremos con las líneas de cada uno de los archivos
echo ">>> Número de líneas de human_coordinates_1.bed"

wc -l ../data/human_coordinates_1.bed

echo ">>> Número de líneas de human_coordinates_2.bed"

wc -l ../data/human_coordinates_2.bed
```

OUTPUT

```
juancarlos@LAPTOP-GA0CC694:/mnt/c/Users/jcge9/Desktop/VIU/ShellScripting/code$
./actividad1.sh
>>> Número de líneas de human_coordinates_1.bed
1907 ../data/human_coordinates_1.bed
>>> Número de líneas de human_coordinates_2.bed
1909 ../data/human_coordinates_2.bed
```

- **¿Cuántas columnas presenta cada uno de los archivos descargados? (1 pts)**

INPUT

```
#!/usr/bin/env bash

# Seguiremos contando el número de columnas con awk

echo ">>> Número de columnas de human_coordinates_1.bed"

awk -F'\t' '{print NF}' ../data/human_coordinates_1.bed | uniq

echo ">>> Número de columnas de human_coordinates_2.bed"

awk -F'\t' '{print NF}' ../data/human_coordinates_2.bed | uniq
```

OUTPUT

```
juancarlos@LAPTOP-GA0CC694:/mnt/c/Users/jcge9/Desktop/VIU/ShellScripting/code$
./actividad1.sh
>>> Número de columnas de human_coordinates_1.bed
3
>>> Número de columnas de human_coordinates_2.bed
3
```

- **¿Tenemos representación de todos los cromosomas humanos en ambos archivos? (1 pts)**

INPUT

```
#!/usr/bin/env bash

echo ">>> Cmosomas de human_coordinates_1.bed"
cut -f1 ../data/human_coordinates_1.bed | uniq | sort -k1.4 -n
echo ">> Nº de líneas de lo anterior, si es < a 23, no están todos los cormosomas"
cut -f1 ../data/human_coordinates_1.bed | sort | uniq | wc -l

echo ">>> Cmosomas de human_coordinates_2.bed"
cut -f1 ../data/human_coordinates_1.bed | uniq | sort -k1.4 -n
echo ">> Nº de líneas de lo anterior, si es < a 23, no están todos los cormosomas"
cut -f1 ../data/human_coordinates_2.bed | sort | uniq | wc -l
```

OUTPUT

```
juancarlos@LAPTOP-GA0CC694:/mnt/c/Users/jcge9/Desktop/VIU/ShellScripting/code$
./actividad1.sh
>>> Cmosomas de human_coordinates_1.bed
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr20
chr21
chr22
>> Nº de líneas de lo anterior, si es < a 23, no están todos los cormosomas
21
>>> Cmosomas de human_coordinates_2.bed
chr1
chr2
chr3
chr4
chr5
```

```
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr20
chr21
chr22
>> Nº de líneas de lo anterior, si es < a 23, no están todos los cromosomas
21
```

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)**

INPUT

```
#!/usr/bin/env bash

# Ahora vamos a ver si los datos difieren.
echo ">>> Ordenamos ambos datos y guardamos los resultados en un nuevo archivo"
sort -k 1.4 -n ../data/human_coordinates_1.bed >
../data/human_coordinates_1_ordered.bed
sort -k 1.4 -n ../data/human_coordinates_2.bed >
../data/human_coordinates_2_ordered.bed

echo ">> En caso de diferencias ¿Cuántas y que regiones son diferentes?"
diff ../data/human_coordinates_1_ordered.bed
../data/human_coordinates_2_ordered.bed -q ; diff
../data/human_coordinates_1_ordered.bed ../data/human_coordinates_2_ordered.bed
```

OUTPUT

```
juancarlos@LAPTOP-GA0CC694:/mnt/c/Users/jcge9/Desktop/VIU/ShellScripting/code$
./actividad1.sh
>>> Ordenamos ambos datos y guardamos los resultados en un nuevo archivo
>> En caso de diferencias ¿Cuántas y que regiones son diferentes?
Files ../data/human_coordinates_1_ordered.bed and
```

```

../data/human_coordinates_2_ordered.bed differ
232a233
> chr1 204073115      204127743
671a673
> chr6 31164337      31170682
1654a1657
> chr17 42313412     42388540

```

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)

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

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.edu/](http://navegador.genómico.ajogado.por.la.Universidad.de.California.Santa.Cruz.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)

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)**

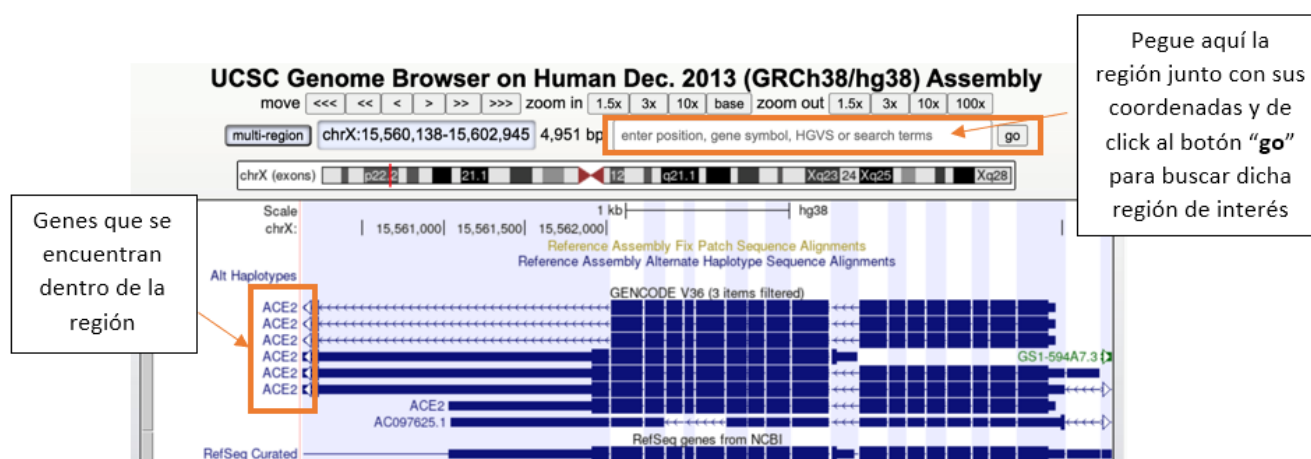


Figura 1. Vista del UCSC Genome Browser.

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).