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.

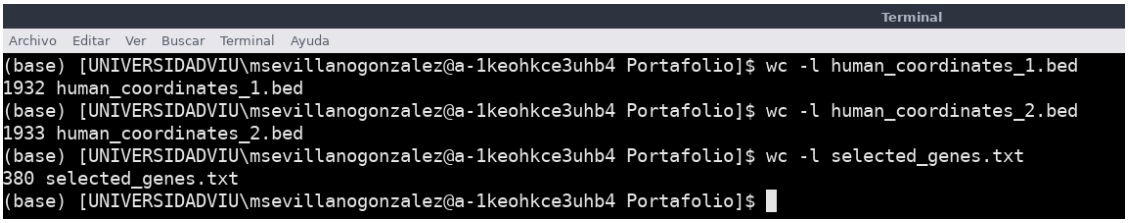
Los datos con los que va a trabajar se refieren a regiones de interés detectadas en células inmunitarias en humanos. Para ello, se han realizado dos réplicas técnicas del experimento, obteniendo dos archivos llamados human_coordinates_1.bed y human_coordinates_2.bed. Además de estas regiones, se seleccionaron 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

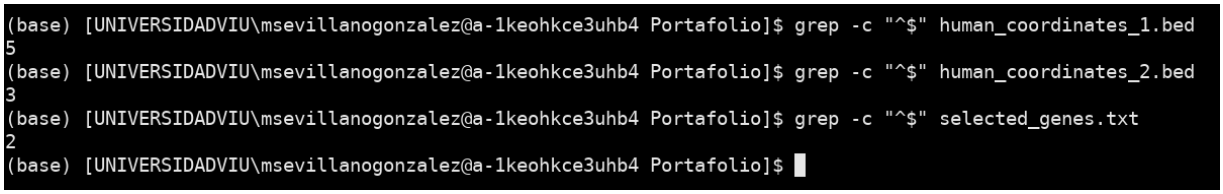
Actividades a realizar

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.**

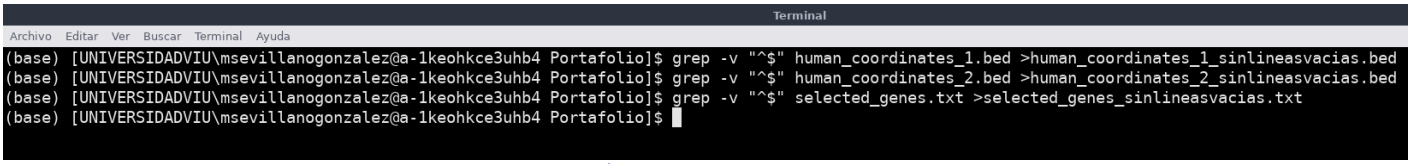


El comando wc con la opción -l permite contar todas las líneas de un archivo (que tienen saltos de línea).



El comando grep con la opción -c cuenta el número de veces que aparece un patrón. En este caso el patrón indicado es "^$", que permite encontrar líneas vacías.

Por tanto, con el comando grep -c "^$", se determina el número total de líneas vacías de cada archivo.



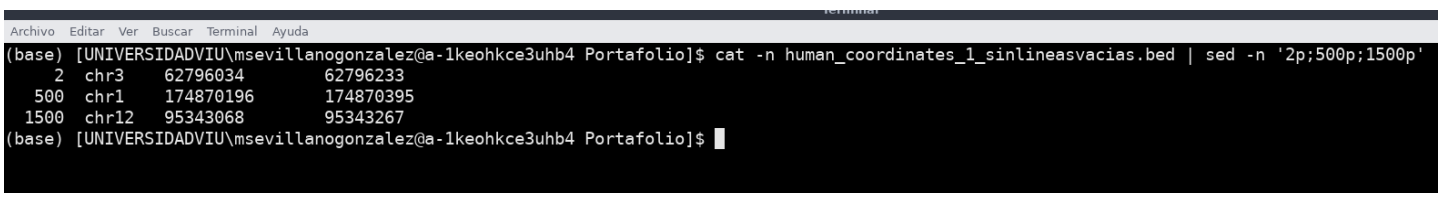
La opción -v del comando grep seguida de "^$", nos permite eliminar esas líneas vacías.

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:

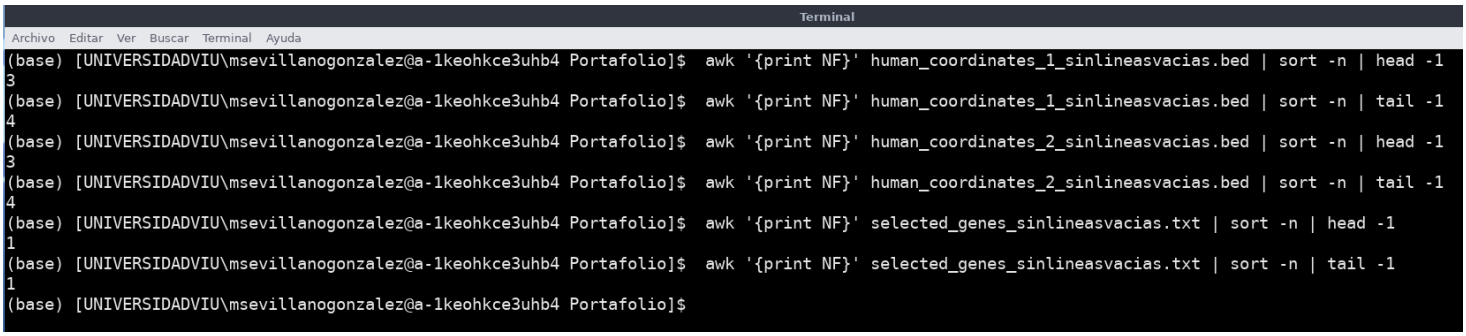
2 chr3 62796234 6279643

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



El comando cat con la opción -n permite imprimir el contenido completo del archivo y numerar todas las líneas. Con la tubería o pipe (|) el resultado de este comando es transferido al comando sed, que con la opción -n va a suprimir la impresión automática de todas las líneas (que tiene por defecto), permitiendo con la orden p imprimir los números específicos de línea indicados.

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



El comando awk nos permite realizar una acción que definimos entre corchetes, {print NF}, que nos permite imprimir el número de campos/columnas (NF).

El resultado de este comando se transfiere mediante pipe al comando sort con la opción -n que permite ordenar por número, en este caso el número de columnas por fila resultado del comando awk, que de manera predeterminada lo ordena de menor a mayor.

El comando head -1 enlazado con otra tubería, nos permite obtener el primer resultado, que será por tanto el número mínimo de columna al estar ordenador de menor a mayor. El comando tail -1, nos permite obtener el último resultado, que será entonces el número máximo de columnas.

Por tanto, para el archivo human_coordinates_1.bed el número mínimo de columnas es 3 y el número máximo de columnas es 4, lo mismo que para el archivo de human_coordinates_2.bed. El archivo selected_genes.txt solo tiene una columna.

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)

El comando sort permite ordenar las líneas de un archivo, y con la opción -V trabaja tanto con números como con texto. La opción -u permite además eliminar duplicados. Por tanto, se ordenará el archivo por cromosoma y número, y por coordenadas.
A través de > redirigimos la salida a un nuevo archivo.
El comando cat permite imprimir el contenido completo del archivo.

```
Terminal
Archivo  Editar  Ver  Buscar  Terminal  Ayuda
(base) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 Portafolio]$ sort -V -u human_coordinates_1_sinlineasvacias.bed > coordenadas_unicas_1.bed
(base) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 Portafolio]$ cat coordenadas_unicas_1.bed
chr1    75339199        75339398
chr1    75339399        75339598
chr1    75339599        75339798
chr1    75339799        75339998
chr1    92832599        92832798
chr1    92832799        92832998
chr1    92832999        92833198
chr1    92960399        92960598
chr1    92960599        92960798
chr1    92960799        92960998
```

```
(base) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 Portafolio]$ awk '{print $1}' coordenadas_unicas_1.bed | uniq -c
    322 chr1
    134 chr2
     89 chr3
     18 chr4
     44 chr5
     96 chr6
    129 chr7
     28 chr8
     12 chr9
     44 chr10
     85 chr11
    506 chr12
     28 chr13
     47 chr14
     38 chr15
     30 chr16
     42 chr17
      9 chr18
     25 chr20
    162 chr21
     15 chr22
      1 chrX
(base) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 Portafolio]$
```

Con el archivo ya ordenado y eliminadas las líneas duplicadas, quedarán por tanto los cromosomas con coordenadas únicas. Para obtener este resultado, se imprime con awk la columna 1 mediante la acción {print $1}, los nombres de los cromosomas, y este resultado se transfiere mediante tubería al comando uniq con la opción -c que imprime el número de veces que se repite una línea.

De esta forma, al estar ordenado el archivo y eliminados duplicados, contando las veces que se repite cada cromosoma se obtiene el número de coordenadas únicas por cada cromosoma.

No hay representación de todos los cromosomas, faltarían el cromosoma 19 y el Y.

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)**

```
(base) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 Portafolio]$ sort -k1,2 -V -u  human_coordinates_1_sinlineasvacias.bed -o cromosomas_ordenados_1.bed
(base) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 Portafolio]$ sort -k1,2 -V -u  human_coordinates_2_sinlineasvacias.bed -o cromosomas_ordenados_2.bed
(base) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 Portafolio]$ cat cromosomas_ordenados_1.bed
chr1    75339199        75339398
chr1    75339399        75339598
chr1    75339599        75339798
chr1    75339799        75339998
chr1    92832599        92832798
chr1    92832799        92832998
chr1    92832999        92833198
chr1    92960399        92960598
chr1    92960599        92960798
```

Con el comando sort -k1,2 le decimos que ordene por la columna 1 y la columna dos, es decir, el nombre del cromosoma y las coordenadas de inicio. Con -V para que trabaje con texto y número y -u para que elimine duplicados. La opción -o permite cambiar la salida a otro archivo.

```
(base) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 Portafolio]$ diff --suppress-common-lines cromosomas_ordenados_1.bed cromosomas_ordenados_2.bed
315a316
> chr1  204073115       204127743
618a620
> chr6  31164337        31170682
1651a1654
> chr17 42313412        42388540
```

El comando diff compara línea por línea dos archivos, y con la opción –supress-common-lines elimina las líneas comunes, dejando solo las diferencias entre los dos archivos. Nos indica en este caso, que en el archivo 1 faltan las 3 regiones que imprime por pantalla del archivo 2.

```
(base) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 Portafolio]$ awk '{OFS="\t"} NR==316 || NR==620 || NR==1654 {print $1, $2, $3}' cromosomas_ordenados_2.bed > region
es_distintas.bed
(base) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 Portafolio]$ cat regiones_distintas.bed
chr1    204073115       204127743
chr6    31164337        31170682
chr17   42313412        42388540
```

Sabiendo los números de línea concretos del archivo 2 para esas regiones, con el comando awk podemos obtener las líneas concretas mediante su variable interna predefinida NR (número de registros/líneas), concatenando las instrucciones con ||. El número de columna específico se le indica con $. Y la variable predefinida de awk OFS nos permite definir el separador de la salida, en este caso, por tabulador (\t).
Redirigimos el resultad a un nuevo archivo con >.
El comando cat permite imprimir el contenido completo del archivo para poder visualizarlo por pantalla.

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).

```
(base) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 Portafolio]$ awk '{print $1 ":" $2 "-" $3}' regiones_distintas.bed > coordenadas_cromosomicas.bed
(base) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 Portafolio]$ cat coordenadas_cromosomicas.bed
chr1:204073115-204127743
chr6:31164337-31170682
chr17:42313412-42388540
(base) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 Portafolio]$
```

El comando awk nos permite definir la acción a realizar mediante corchetes, permitiendo imprimir las columnas con $ y el tipo de separador entre columnas, establecido entre dobles comillas “ “. Redirigimos el resultado a otro archivo nuevo (>).

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)

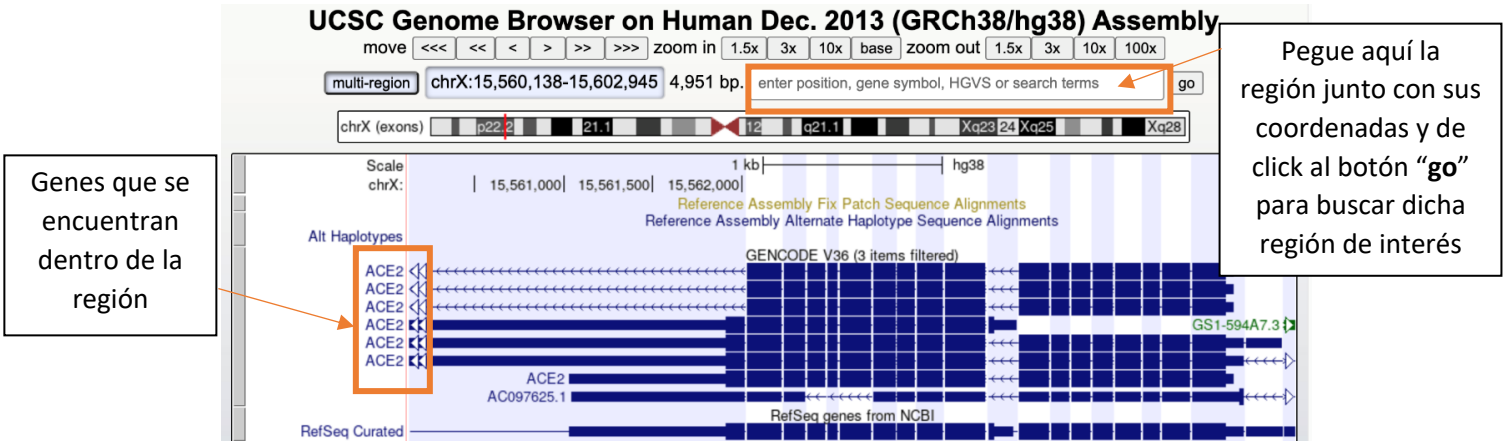
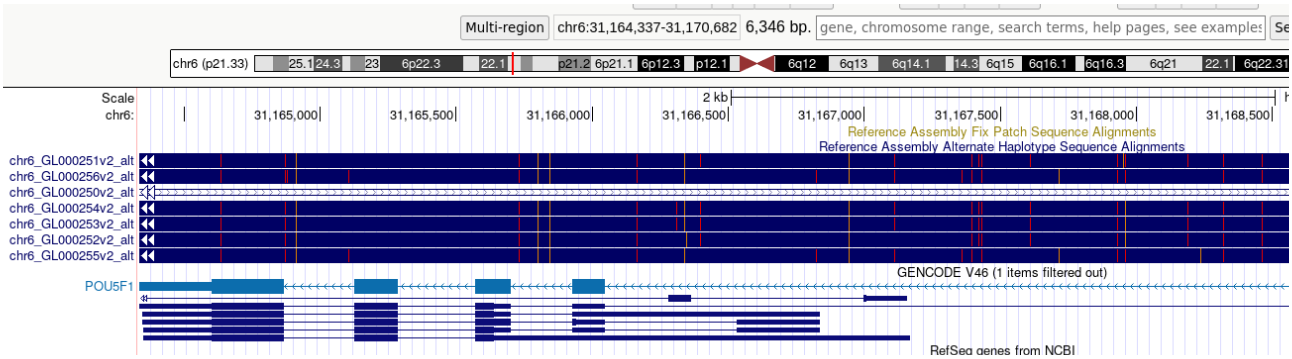
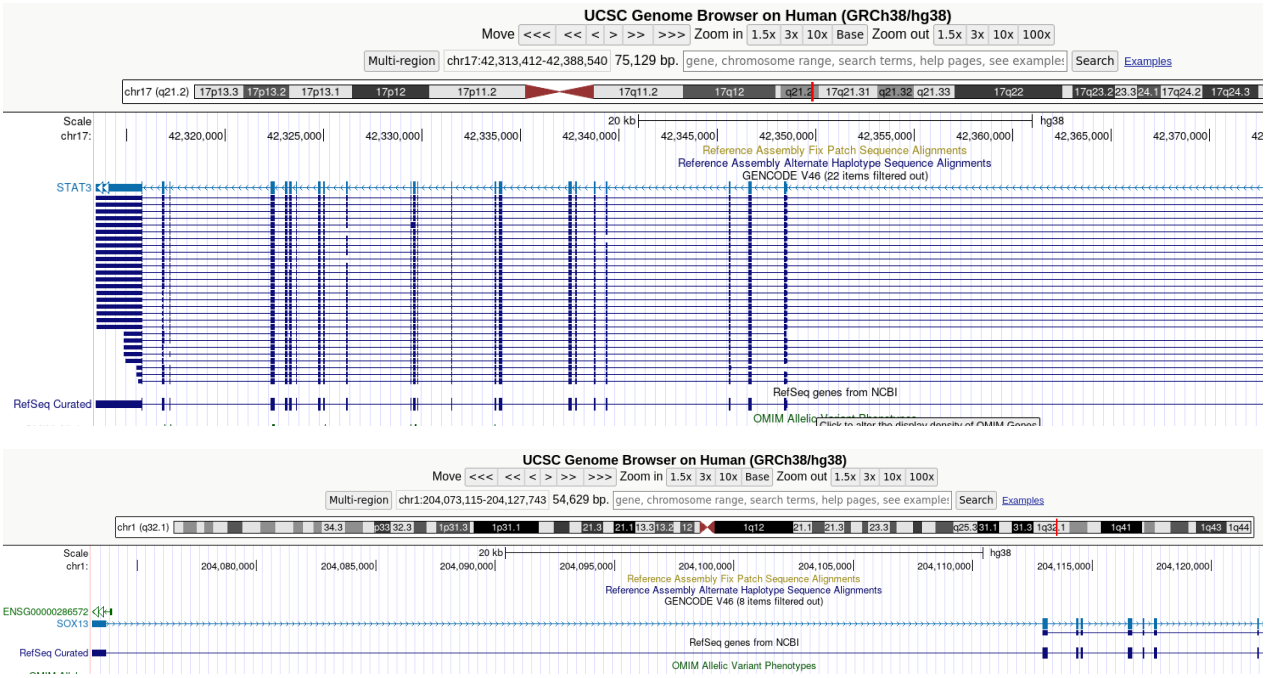


Figura 1. Vista del UCSC Genome Browser.

Los genes encontrados con estas coordenadas cromosómicas son: POU5F1, STAT3 y SOX13.





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).

```
(base) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 Portafolio]$ awk 'NR==345 || NR==366 || NR==378 {print
 $2,$1, NR}' selected_genes_sinlineasvacias.txt > genes_seleccionados.txt
(base) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 Portafolio]$ cat genes_seleccionados.txt
 Sox13 345
 Stat3 366
 Pou5f1 378
(base) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 Portafolio]$
```

Del archivo selected_genes con cat -n obtenemos los números de línea y buscamos los genes en cuestión.

El comando awk nos permite imprimir las líneas específicas con NR indicándoselo mediante ==, y las columnas en el orden que queremos: $2,$1, siendo la columna 2 el nombre de los genes (que ahora será la primera columna) y la columna 1 el número de línea (que ahora será la segunda columna), que imprimimos gracias a la variable predefinida NR.

Todo esto lo redirigimos a un nuevo archivo (genes_seleccionados) que contendrá en la primera columna el nombre del gen y en la segunda columna el número de línea correspondiente del archivo original.

```
(base) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 Portafolio]$ paste coordenadas_cromosomicas.bed genes_
seleccionados.txt > genes_definitivo.txt
(base) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 Portafolio]$ cat genes_definitivo.txt
chr1:204073115-204127743        Sox13 345
chr6:31164337-31170682   Stat3 366
chr17:42313412-42388540  Pou5f1 378
(base) [UNIVERSIDADVIU\msevillanogonzalez@a-1keohkce3uhb4 Portafolio]$
```

El comando paste permite fusionar líneas de archivos, línea a línea. Por tanto, con paste se fusiona el archivo que contiene las coordenadas cromosómicas seleccionadas previamente (ya con el formato cromosoma:inicio-fin que corresponden a los genes en cuestión), y el archivo anteriormente creado con el gen y el número de línea, en el orden ya establecido, redirigiéndolo a un nuevo archivo.

El comando cat permite visualizar el contenido completo del archivo.