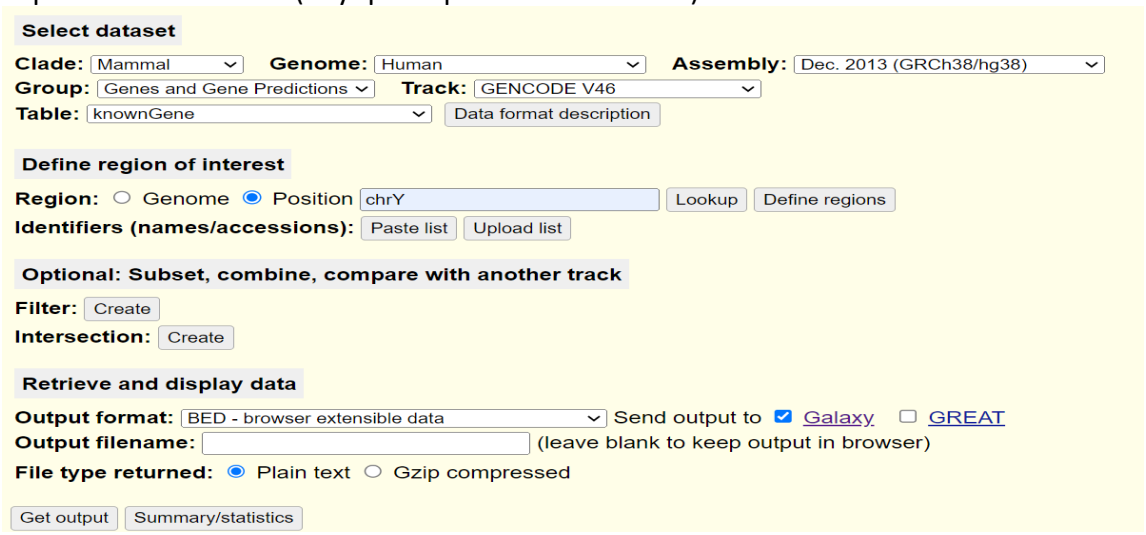
ANÁLISIS DE DATOS GENÓMICOS: Cálculo, cuantificación y localización física de SNPs.

Pretendemos encontrar cuales son los 10 exones de los cromosomas 21, 22 e Y del genoma humano que poseen mayor numero acumulado de SNPs, donde se localizan físicamente en cada cromosoma y a que genes corresponden.

1. Utilizaremos Galaxy (https://usegalaxy.org/) que es una suite abierta de programas para descargar datos genómicos y analizarlos.
2. Lo primero es registrarse en Galaxy y obtener una cuenta. Ir al enlace de 'User' en la parte superior de la interfaz de Galaxy y elegir 'Register' (o 'Login' si ya tienes una cuenta)
3. Cada alumno trabajará con un cromosoma humano que le asignará el profesor.
4. **Obtención de los datos**. Los exones codificantes del cromosoma asignado los obtendremos del servidor UCSC haciendo clic en el menú de la izquierda: "Get Data -> UCSC Main".
   En el panel central de Galaxy se visualiza la interfaz del navegador UCSC.
   Seleccionar el cromosoma asignado siguiendo la figura de abajo, asegurándose de que la configuración es exactamente la misma que se muestra en la pantalla (en particular, el formato de salida debe estar en "BED - browser extensible data" y "Galaxy" debe estar seleccionado como opción de salida 'Send output'). Hacer click '**get output**' para pasar a la pantalla siguiente y seleccionar los exones codificantes. Hacer click en 'Send query to Galaxy' para obtener los datos en el panel de la derecha (hay que esperar un momento):



5. Después de esto se podrá ver el primer proceso en el panel 'History' a la derecha. Irá pasando por el gris (preparación) y amarillo (en ejecución) hasta establecerse en verde cuando se acaba la obtención de los datos.
6. **Obtención de SNPs** (se hace casi exactamente de la misma manera). Hacemos clic de nuevo en "Get Data -> UCSC Main". Ahora en 'group' seleccionamos 'variation' y seleccionamos 'all SNPs 151' y en la siguiente página (después de hacer click en 'get output') seleccionamos 'whole gene' que aquí indica que todas las características de los SNPs son obtenidas. Hacer click en 'Send query to Galaxy' para obtener los nuevos datos en el panel de la derecha.
7. **Edición de los nombres**. Vamos a cambiar el nombre de los dos elementos del historial a "Exones" y "SNPs" haciendo clic en el icono de lápiz junto a cada elemento.

8. **Localización de los exones con el mayor número de SNPs**. Para encontrar los 10 exones que contienen la mayoría de los SNPs lo primero que tenemos que hacer es unir los datos de exones y de SNPs en una sola matriz. Esto se hace utilizando la herramienta: "Operate on Genomics Intervals -> Join", añadiendo los exones como primer item y los SNPs como segundo item. Hacer clic en 'Run Tool' para general el tercer elemento del historial.

9. **Visualizamos el conjunto de datos**. Las primeras seis columnas corresponden a los exones. Las últimas seis corresponden a SNPs. Algunos exones se repiten varias veces en la matriz de datos, ya que hay más de un SNP en ese exón y por tanto se pueden calcular fácilmente el número de SNPs por cada exón simplemente contando el número de repeticiones de nombre para cada exón. Esto se hace con la herramienta "Join, Subtract, and Group" -> "Group". Elegir en esta pantalla la columna 4 seleccionando "Column: 4" en el Grupo de columna. Luego haga clic en 'Insert Operation' y asegurarse de que el tipo es 'Count' en la "columna 4". Hacer click en 'Run Tool'.

10. **Ordenación de los datos.** Si visualizamos el resultado de la agrupación veremos que ahora contiene solamente dos columnas; la primera con el nombre del exón y la segunda indicando el número de veces que se ha repetido este nombre en el conjunto de datos anteriores. Para ordenar los exones por número de SNPs y localizar así los 10 exones que tiene el mayor número de SNPs, simplemente tenemos que ordenar descendentemente el conjunto de datos obtenidos a partir de la segunda columna. Esto se hace con "Sort" localizado dentro de "Text Manipulation", seleccionando nuestro grupo de datos, la segunda columna y "descending" order.

11. **Selección de los exones.** Ahora necesitamos identificar a que genes corresponden esos exones para seleccionar los diez exones con el mayor número de SNPs. Vamos a utilizar la herramienta "Text Manipulation -> Select First lines from a datashet". Es posible que algunos exones se repitan en parte debido a la nomenclatura diferente de algunos exones en secuencias alternativas. Para eliminar repeticiones y elegir los 10 exones diferentes es conveniente seleccionar un mayor numero de líneas (e.g. 30). Al hacer clic en 'Run Tool' se generará el sexto elemento del historial que contienen sólo las 30 líneas correspondiente a cada exón con mayor número de SNPs.

12. **Recuperación de coordenadas.** Una vez tenemos los 10 exones con mayor número SNPs, debemos localizar el gen y el exón en el que se encuentran y para ello necesitamos recuperar la información de posición (coordenadas) de estos exones (información que se ha perdido en la etapa de agrupación). Para obtener las coordenadas de nuevo vamos a hacer coincidir los nombres de los exones en el conjunto de datos recién generado (columna 1) con los nombres de los exones en el conjunto de datos originales obtenidos en primer lugar (columna 4 de "Exones"): herramienta "Join, Subtract and Group" -> "Join two Datasets". Unir datos 'Exones' usando la columna 'c4' frente a 'Select first on data' y la columna 'c1'. Dar a 'Run Tool' para obtener el séptimo elemento de la historia.

13. **Ordenación de los datos**. Tras este último paso, los exones se han vuelto a desordenar. Podemos ordenarlos de nuevo <usando la herramienta "Sort" como hicimos en el paso 10 (ten en cuenta que ahora la columna para ordenarlos de manera descendente no es la misma).

14. **Localización de exones.** La mejor manera de saber acerca de estos exones es mirar en su entorno genómico y para ello vamos a utilizar ENSEMBL, un navegador genómico. Abrir los datos en el navegador, utilizar las coordenadas obtenidas para localizar los genes y rellenar la tabla del ANEXO I a partir de los resultados obtenidos.

ANEXO I.

APELLIDOS Y NOMBRE:

Cromosoma humano asignado:	

Gen	**Exón**	**No. SNPs**	**Localización (banda)**	**Función**