Bioinformática y análisis de datos ómicos

EJERCICIO TEMA 9

Alumno:

Realiza las distintas actividades de abajo copiando debajo de cada actividad los comandos necesarios. Para ello descarga la carpeta para realizar los ejercicios del tema 9. Finalmente, convierte este documento a PDF y entrégalo a través del moodle.

1.- Utiliza el comando gzip para descomprimir los archivos Homo_sapients.GRCh38.109.5.gtf.gz y Homo_sapiens.GRCh38.dna.chromosome.5.fa.gz.

2.- Utiliza el comando hisat2-build para generar los índices correspondientes al genoma Homo_sapiens.GRCh38.dna.chromosome.5.fa.gz

3.- Utiliza la herramienta hisat2 para alinear los archivos fastq de la muestra H460_shARID2 frente al genoma que has indexado en el ejercicio 2. Convierte el archivo SAM al formato BAM, ordénalo de acuerdo a las coordenadas genómicas e indéxalo con SAMTOOLS.

4.- Utiliza la herramienta htseq-count para generar los contajes de lecturas correspondientes a cada gen utilizando el archivo Homo_sapiens.GRCh38.109.5.gtf como modelo de genes. Utiliza la opción para decirle a htseq-count que no tenga en cuenta la hebra al contar las lecturas (-s).

5.- Utilizando R, modifica el script Tema9.R para generar un dataframe a partir de los archivos con extensión .counts que contienen información de tres replicados biológicos de células deficientes en ARID2 (H661_ARID2) y de células control (H661_EMPTY) generados a partir de experimentos de RNA-seq. El dataframe debe contener los genes en filas y las distintas muestras en columnas.

6.- Genera un segundo dataframe con el nombre de las muestras en filas y una sola columna que contenga, en forma de factor, a qué grupo de estudio pertenece cada muestra.

7.- Carga la librería "DESeq2" para analizar los datos y genera un objetivo tipo Deseq utilizando la función "DESeqDataSetFromMatrix"

8.- Corre el análisis de Deseq utilizando la función "DESeq" sobre el objetivo generado en el ejercicio 7.

9.- Extrae los resultados del objeto deseq usando la función "results" y los contajes normalizados usando la función "counts".

10.- Haz una representación de puntos tipo (Volcano plot) en donde se representa el -log10 del padj en el eje Y y el log2FoldChange en el eje X.

11.- Carga la librería "pheatmap" y dibuja un diagrama heatmap con los 20 genes con mayor foldchange en ambas direcciones y con un padj < 0.05.

12.- Calcula los componentes principales con la función prcomp a partir de los datos contenidos en el archivo "Breast_normalized_counts.txt ".

13.- Clasifica las muestras utilizando las primeras 10 dimensiones del PCA con la función kmeans.

14.- Dibuja un diagrama de puntos en donde representes el PC1 en el eje Y, el PC2 en el eje X y colorea los puntos de acuerdo a las agrupaciones (clusters) predichas por kmeans.