Trabajo Transcriptómica Curso 2021 – 2022

El objetivo de este trabajo es que el alumno demuestre los conocimientos obtenidos acerca del análisis de RNA-seq. Para ello deberá **redactar un informe** en el que se expliquen los datos de partida y se extraiga una **conclusión de los resultados.** Así mismo, el alumno debe **detallar el proceso de análisis** indicando el *software* (incluída la versión) empleado, así como los parámetros utilizados en cada uno de los pasos. En caso de que hubiera que eliminar muestras por motivos técnicos o biológicos, el alumno debe indicar y justificar el por qué en cada caso. Se pueden introducir *code chunks*, e imágenes para apoyar el informe. De manera alternativa, puede aportarse todo el código en forma de repositorio público. Se han planteado **5 preguntas (2 puntos cada una)** para **guiar** la redacción del informe.

El trabajo consta de dos apartados en los que se utilizarán *datasets* diferentes. El **primer apartado (3 preguntas)** abarca los pasos de control de calidad y de fuentes de contaminación, *trimming*, alineamiento y cuantificación para obtener cuentas crudas y normalizadas a partir de ficheros fastq. El **segundo apartado (2 preguntas)** parte de una matriz de cuentas crudas de un experimento distinto y está enfocado a realizar un control de calidad biológico, detectar los genes diferencialmente expresados entre condiciones y los *pathways* enriquecidos en cada una de ellas.

Apartado 1 (6 puntos)

Se ha generado un *dataset* de tamaño reducido (similar al utilizado en clase), con el fin de poder abordar las cuestiones planteadas a continuación de manera satisfactoria en vuestros ordenadores portátiles.

Dicho dataset consta de los siguientes ficheros:

(a) Cuatro ficheros fastq correspondientes a cuatro muestras diferentes.
(b) Un fichero fasta con la secuencia de la referencia genómica, en este caso correspondiente al cromosoma 21 humano (ensamblaje GRCh38).
(c) Un fichero GTF con la anotación génica para los genes del cromosoma 21 (GRCh38.gencode.v38).

Se pide realizar un análisis de dichas muestras similar al realizado en clase, considerando los siguientes puntos:

Pregunta 1: Control de calidad de secuenciación. Incluir en el informe las figuras de calidad media por nucleótido de las lecturas (*per base quality*) y la figura de niveles de duplicación (*duplication levels*), comentando ambas figuras.

Pregunta 2: Indexación de la referencia genómica, alineamiento de las diferentes muestras contra la referencia genómica ya indexada y elaboración de los ficheros necesarios para el posterior contaje de lecturas. Razonar los parámetros usados en el alineamiento con Hisat2, y apuntar porcentajes de alineamiento.

Pregunta 3: Contaje de lecturas para la obtención de un fichero con cuentas crudas (*raw*) y un fichero con cuentas normalizadas. Comentar los parámetros usados con HTSeq-count.

Apartado 2 (4 puntos)

Disponemos de 24 cultivos primarios de tumores paratiroideos negativos para receptores de estrógenos alfa (ERα). Las muestras, procedentes de 4 pacientes diferentes, se han tratado con dos fármacos diferentes: diarilpropionitrilo (DPN) o 4-hidroxitamoxifeno (OHT) a 24h o 48h. El DPN es un agonista del ERα mientras que el OHT es un inhibidor competitivo de los receptores de estrógenos.

Dicho dataset consta de los siguientes ficheros:

(a) Matriz de cuentas crudas.
(b) *Data frame* con los metadatos asociados al experimento.

Pregunta 1: ¿Qué genes se encuentran diferencialmente expresados entre las muestras pertenecientes al grupo tratado con OHT con respecto al control tras 24h? ¿Y en el caso de las muestras tratadas con DPN, también tras 24h? Como parte de la respuesta a esta pregunta, podéis entregar una o más tablas adjuntas donde se incluyan los genes diferencialmente expresados, indicando el criterio o los criterios que habéis seguido para filtrar los resultados, así como cualquier otro gráfico o gráficos que hayáis elaborado durante el análisis.

Pregunta 2: Genera dos firmas con los 100 genes más up-/down-regulados en las células tratadas con DPN durante 48h. ¿Están estas firmas enriquecidas en las células tratadas con DPN durante 24h? ¿Qué conclusiones extraes de este resultado? Incluir una tabla con los resultados del análisis, destacando las columnas con los valores utilizados para extraer vuestras conclusiones. También incluir los gráficos característicos de este tipo de análisis.