Actividad 1- *From reads to gene counts*

Objetivo

El propósito de esta actividad es que el estudiante demuestre que ha adquirido las habilidades y competencias necesarias para realizar el **preprocesamiento inicial de datos de RNA-seq.** Esto incluye desde la caracterización y evaluación de calidad de las lecturas crudas hasta su filtrado, alineación y generación de la matriz final de recuentos génicos.

Obtención de los datos

La actividad consiste en el análisis de una muestra real de RNA-seq, que forma parte de un estudio más amplio que examina los perfiles de expresión entre poblaciones de células basales y luminales en glándulas mamarias de ratones en diferentes estadios (vírgenes, gestantes y lactantes).

La publicación de referencia es la siguiente:

Fu, Nai Yang, et al. "EGF-mediated induction of Mcl-1 at the switch to lactation is essential for alveolar cell survival." *Nature Cell Biology* 17.4 (2015): 365-375.

Todo lo necesario para poder realizar esta actividad ha sido previamente abordado en clase y, por tanto, todo el material (incluyendo el entorno de trabajo conda 05MBIF y los archivos necesarios) puede encontrarse en la sección de "Recursos y Materiales/Materiales del profesor/Proyecto_JULIO_2024".

Formato de la entrega

- La entrega se realizará utilizando este documento como plantilla, que se convertirá en PDF y lo adjuntará a la actividad correspondiente dentro del Campus VIU.
- Las respuestas se presentarán de forma clara y concisa, justificando su contenido.
- Se deberá adicionar **SIEMPRE** los comandos empleados, las capturas de pantalla que muestren su ejecución y los gráficos generados que apoyen sus repuestas:
    - Para salidas estándar de larga extensión de algún comando, incluir solo las primeras líneas que muestren su correcta ejecución.
    - No es necesario explicar con detalle los comandos, opciones y/o argumentos empleados.

- Los gráficos deberán tener una calidad suficiente para su lectura y serán acompañados de un pie de figura explicativo con el número de imagen (Figura X: "…").

Preguntas

Para alcanzar la nota máxima, se deberán contestar y justificar cada una de las siguientes preguntas.

P1. Responde brevemente a las siguientes preguntas sobre la caracterización inicial del estudio (1 pts)

- ¿Cuántas réplicas biológicas por grupo se utilizaron en este estudio? ¿Es este un número óptimo para experimentos de RNA-seq?. Justifica tu respuesta.
- Describe el protocolo de secuenciación utilizado en este estudio.

P2. Emplea la herramienta FastQC para evaluar la calidad de las lecturas del archivo SRR1552444.fastq.gz. A continuación, utiliza la herramienta TrimGalore para eliminar adaptadores y extremos de baja calidad. Reporta los comandos utilizados y responde a las siguientes preguntas (3 pts).

- ¿Cuántas lecturas iniciales había en el archivo SRR1552444.fastq.gz y cuántas permanecen tras la utilización de TrimGalore?
- Comenta las diferencias que observas en los reportes de FastQC antes y después del filtrado (recuerda reportar los gráficos resultantes), específicamente en:
    - **Per base sequence quality**
    - **Sequence Length Distribution**
    - **Adapter Content.** ¿Qué secuencia se determina como el adaptador principal? ¿En cuántas lecturas ha encontrado dicho adaptador TrimGalore?

P3. Alinea las lecturas depuradas anteriormente sobre el genoma de referencia indexado empleando la herramienta Hisat2 con la opción (-k = 1). Reporta los comandos empleados y contesta a las siguientes preguntas. (1,5 pts)

- Después del alineamiento con Hisat2, ¿cuál fue el número total de lecturas mapeadas y no mapeadas en la muestra?
- Recuerda que uno de los parámetros de Hisat2 es el valor de -k, que indica el número máximo de alineamientos por lectura a generar. Si empleas un valor de **k = 5**, ¿Cómo afecta esto a los valores finales de alineamiento obtenidos y al tiempo de computación?

P4. Utiliza SAMtools para convertir el archivo SAM (generado con el valor de -k igual a 1) a BAM, ordenarlo por coordenadas e indexarlo. Reporta los comandos utilizados y contesta a las siguientes preguntas. (1,5 pts)

- ¿Cuántas FLAGs distintas se encuentran en el archivo **SRR1552444_hisat2.sam**? Indica cuáles son y sus cantidades.
- ¿Cuántos valores distintos de MAPQ hay en el archivo **SRR1552444_hisat2.sam**? Indícalos y cuenta cuántos son.
- ¿Cuántas letras distintas del alfabeto encontramos en la columna CIGAR del archivo **SRR1552444_hisat2.sam**? Indica cuáles son, sus cantidades y qué información proporcionan.

P5. Como hemos visto en el archivo BAM, hay una ubicación cromosómica para cada lectura asignada. Ahora que ya hemos descubierto de dónde proviene cada lectura en el genoma, necesitamos anotar dicha información. Para ello obtén el archivo de anotaciones en formato GTF y responde a las siguientes cuestiones. (1,5 pts)

- ¿Cuántos y qué programas y bases de datos se emplearon para anotar este archivo? Para responder a esta pregunta, interroga la columna SOURCE.
- ¿Cuáles y cuántas FEATURES podemos encontrar en el archivo de anotaciones?
- ¿Qué porcentaje de GENES en el archivo GTF están ubicados en el cromosoma X?

P6. Finalmente, ejecuta el recuento de los alineamientos sobre el archivo BAM con htseq-count. Reporta los comandos utilizados y computa sobre el archivo tsv final, el porcentaje total de lecturas asignadas, no_feature, ambiguous, too_low_aQual y not_aligned. Una vez computado estos porcentajes, muéstralos con un gráfico (usando cualquier lenguaje de programación) y comenta brevemente su resultado. (1,5 pts)

P7 OPCIONAL. A lo largo del flujo de trabajo, existen etapas que no se han abordado en clase, como la identificación y marcaje de duplicados con MarkDuplicates(Picard), la evaluación de calidad del alineamiento con RseQC, o el uso de herramientas alternativas para el recorte, filtrado, mapeado y recuento de lecturas. En esta parte opcional de la actividad (que puede aportar hasta 0,5 puntos adicionales), se brinda al estudiante la libertad de seleccionar y aplicar alguna de estas etapas adicionales. Se solicita que, además de implementar la etapa seleccionada, se comparta tanto el código desarrollado como las observaciones principales derivadas de su aplicación.