# Actividad 2- *From gene counts to DEG and pathways*

## Objetivo

El propósito de esta actividad es que el estudiante demuestre la adquisición de habilidades y competencias necesarias para llevar a cabo un tratamiento estadístico adecuado sobre los datos de conteo de un experimento de RNA-seq. Esto implica determinar los genes expresados de manera diferencial entre diferentes escenarios y extraer los principales términos ontológicos enriquecidos.

## 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 a PDF y se adjuntará a la actividad correspondiente dentro de 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 (eliminar todo aquello que no sea informativo) y los gráficos generados que apoyen sus repuestas.
- No es necesario explicar con detalle los comandos, opciones y/o argumentos empleados.
- Los gráficos o capturas de pantalla deberán tener una calidad y tamaño suficiente para su lectura.

## Preguntas

P1. Lee la matriz de recuentos llamada GSE60450_Lactation-GenewiseCounts.txt, selecciona únicamente las columnas que representan las muestras y reduce el nombre de las muestras tal y como hicimos en clase. Contesta a las siguientes preguntas que te permitirán conocer un poco mejor su contenido y estructura. Recuerda añadir siempre los comandos empleados y el resultado de su ejecución (2 pts).

- ¿Cuántos genes figuran en la matriz de recuentos?
- Genera un gráfico de barras con los valores máximos de lecturas para cada una de las muestras de estudio y comenta los resultados. ¿A qué gen se asocian en cada caso?
- ¿Cuál es el porcentaje de genes que presentan un valor de conteos > 0 en cada una de las muestras? Comenta brevemente los resultados.
- Al realizar la transformación de los identificadores *ENTREZID* de los genes a *SYMBOL* verás que algunos de ellos no se han podido convertir correctamente y la herramienta los ha sustituido por valores NA. ¿Cuántos puedes contabilizar? ¿Qué decisión tomarías respecto a los mismos?

P2. Convierte la matriz de recuentos en un objeto DGEList y elimina todos los genes que presenten un conteo limitado empleado la función filterByExpr (1 pts).

- ¿Qué porcentaje de genes se mantienen en el *dataset* tras el filtrado?
- Realiza un gráfico de cajas (*boxplot*) con los conteos transformados a logCPM antes de computar los factores de normalización y comenta su resultado.

P3. Siguiendo el análisis con la muestra filtrada, computa sus factores de normalización mediante el método TMM. Recuerda añadir siempre los comandos empleados y el resultado de su ejecución y contesta las siguientes preguntas (1,5 pt).

- Realiza un gráfico de cajas (*boxplot*) con los conteos transformados a logCPM tras computar los factores de normalización y comenta su resultado.
- ¿Qué muestras presentan un factor de normalización (norm.factors) menor de 1 y qué puede significar esto?
- Genera un gráfico de MDS en el que cada tipo celular este coloreado de un color distintivo (*basal* vs *luminal*) y las muestras *lactate* se presenten con un círculo lleno, las muestras *pregnant* con un triángulo lleno y las muestras *virgin* con un cuadrado lleno. Comenta brevemente cómo se asocian o diferencian dichas muestras.

P4. Calcula los tres tipos de dispersión en el objeto DGEList creado. Recuerda añadir siempre los comandos empleados y el resultado de su ejecución y contesta las siguientes preguntas (0,5 pt).

- ¿Cuáles son los valores de "trended dispersion", y "tagwise dispersion" para los genes ***Imp4***, ***Xkr4*** y ***Sox17***?
- Después del ajuste de GLM con la función *glmQLFit*, ¿qué valores tienen los *coefficients* de los seis grupos experimentales a estudio para los genes ***Imp4***, ***Xkr4*** y ***Sox17***?

P5. Utiliza la función *glmQLFTest* para realizar la prueba de hipótesis nula en la siguiente comparación:

- *Luminal.lactate vs luminal.pregnant*

Utiliza la función *topTags* sobre la prueba realizada anteriormente y observa cómo se genera una nueva columna llamada FDR. Recuerda añadir siempre los comandos empleados y el resultado de su ejecución y contesta a las siguientes preguntas (2,5 pts).

- ¿Cuántos genes diferencialmente expresados (DEG), aplicando un punto de corte de FDR <= 0.05 y un valor absoluto de logFC >= 1, encuentras?
- Dibuja un gráfico de Volcán con el resultado obtenido y etiqueta los 3 genes con mayor valor absoluto de LogFC. ¿Cuáles son?
- ¿Cuántos genes son únicos y compartidos (diferencia entre sobreexpresados e infraexpresados) puedes encontrar entre esta comparación y aquella que realizamos en clase entre *basal.lactate vs basal.pregnant*?. Para ello realiza un diagrama de Venn o un Upset plot donde muestres los resultados obtenidos. Recuerda emplear los datos obtenidos tras aplicar los puntos de corte especificados en el punto 1 de esta pregunta.

P6. Transforma la matriz de conteos completa a datos normalizados (en logcpm). Recuerda añadir siempre los comandos empleados y el resultado de su ejecución (1,5 pts).

- Filtra la matriz completa tomando únicamente los genes compartidos entre los tipos celulares *basal* y *luminate* que se encuentren sobreexpresados en ambas líneas celulares.
- Genera un mapa de calor empleando la función *pheatmap* y comenta tu resultado. ¿Cómo se organizan las muestras? ¿El resultado (a nivel muestral) es el mismo si aplicas otro método de aglomeramiento o *linkage*?

P7. Finalmente, selecciona el grupo de genes anterior y realiza un análisis de enriquecimiento funcional indicando los 10 primeros resultados que reporta topGO en las categorías BP y MF. Comenta con detalle los resultados obtenidos y añade referencias bibliográficas si es necesario (1 pts).