

Practica 4.

Análisis de expresión diferencial utilizando datos de secuenciación masiva de ARN (RNA-seq)

Genómica Computacional 2019-2
Licenciatura en Ciencias de la Computación
Facultad de Ciencias
Universidad Nacional Autónoma de México

El reporte correspondiente a esta practica deberá ser entregado el día 24 de mayo antes de las 23:59 hrs.

1. Obtención de los datos. Utilizando el número de acceso GSE108912, descarga desde NCBI los datos a utilizar. Este número de acceso está ligado a un GEO Data Set (la primera opción). Haciendo clic en la primera opción, en la tabla "Supplementary file" en la parte inferior de la página, haz click en en Download > http para descargar los datos.

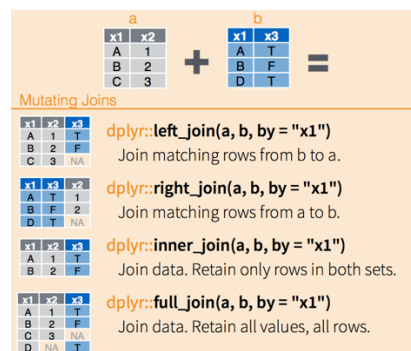
Supplementary file	Size	Download	File type/resource
GSE108912_RAW.tar	1.8 Mb	(http)(custom)	TAR (of TXT)

Raw data are available in SRA

Processed data provided as supplementary file

2. Para trabajar con los archivos deberás descomprimir el archivo descargado, el cual resultará en una carpeta que contiene seis archivos que a su vez deben ser descomprimidos. Cada uno de estos archivos contiene una tabla en la que se enlista para cada gen el número de lecturas que fueron mapeadas a ese gen. Este número es utilizado como una medida de la expresión génica del gen en cada una de las muestras.

Para trabajar con los datos primero deben combinarse en una sola matriz para lo cual utilizaremos la función `inner_join()` del paquete `dplyr`.



El script que se provee permite hacer un análisis de expresión génica utilizando este tipo de datos. Al igual que con los microarreglos algunos pasos estándar son:

- i. Visualizar la distribución de las muestras a través de cladogramas y PCAs.
- ii. Ajustar los datos a un modelo lineal y realizar pruebas de hipótesis para determinar los genes diferencialmente expresados.
- iii. Visualizar la expresión de los genes diferencialmente expresados a través de las muestras con un mapa de calor (heatmap).
- iv. Realizar un análisis de enriquecimiento de los términos GO.

Para esta practica se hará un reporte siguiendo como guía las siguientes preguntas.

- a. Realiza una breve descripción de qué es *Arabidopsis thaliana* y la hormona vegetal citocinina (cytokinin en inglés).
- b. En las muestras que se analizan, las muestras “tratadas” fueron expuestas al compuesto 6-benzylaminopurine (BA). ¿Qué es este compuesto? ¿Y cuál es su relación respecto a las citocininas?
- c. A partir del agrupamiento y del PCA, ¿Dirías que las muestras se agrupan adecuadamente?
- d. ¿Cuántos genes se encuentran diferencialmente expresados entre las muestras tratadas en relación con el grupo control?
- e. Dado el efecto esperado de BA sobre las células, ¿hay procesos asociados a las citocininas que se encuentren enriquecidos en las muestras?
- f. ¿Qué otros procesos se ven enriquecidos?
- g. Añade cualquier observación que consideres pertinente.