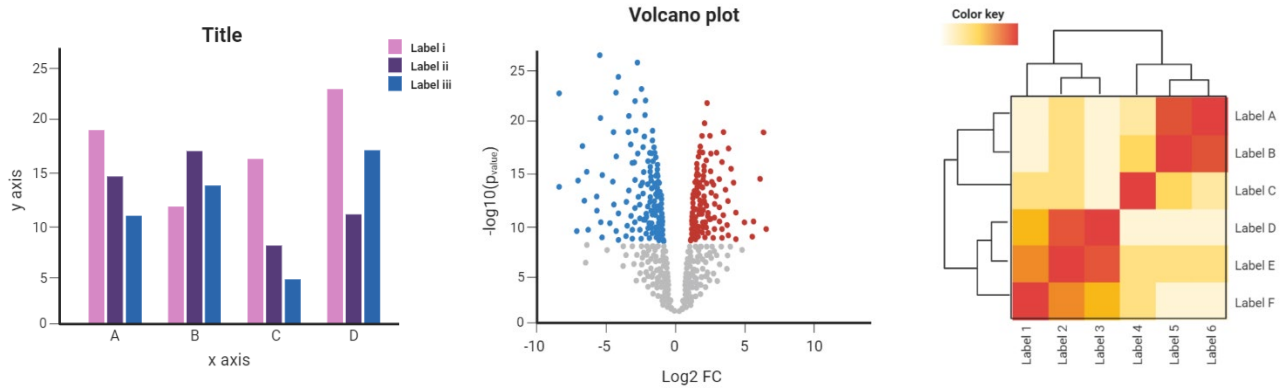
¿Cómo podemos usar datos biológicos abiertos del espacio para aprender sobre visualizaciones de datos?



Contexto

Existen muchas formas de visualización de datos, cada una diseñada para representar información específica y ayudar a identificar tendencias, valores atípicos y patrones. En el análisis -omics, tres gráficos comunes son **mapas de calor**, **gráficos de volcán** y **gráficos de análisis de componentes principales (PCA por sus siglas en inglés)**.

Los **mapas de calor** se utilizan para ilustrar la magnitud a través de un gradiente de intensidad de color mapeado en una matriz bidimensional. En un ejemplo de -omics, las filas típicamente representan un identificador de gen y las columnas indican la muestra de RNAseq. La intensidad del color, en este caso, indica el grado en que el gen se expresa.

Los **gráficos de volcán** son un tipo de gráficos de dispersión que ilustran una relación entre la significancia estadística (también llamada tasa de descubrimiento falso) y el cambio de pliegue (indicativo de la significancia biológica). Estos gráficos se utilizan para identificar los genes más sobreexpresados y subexpresados en los extremos (izquierda o derecha). Los genes más significativos estadísticamente aparecen en la parte superior del gráfico.

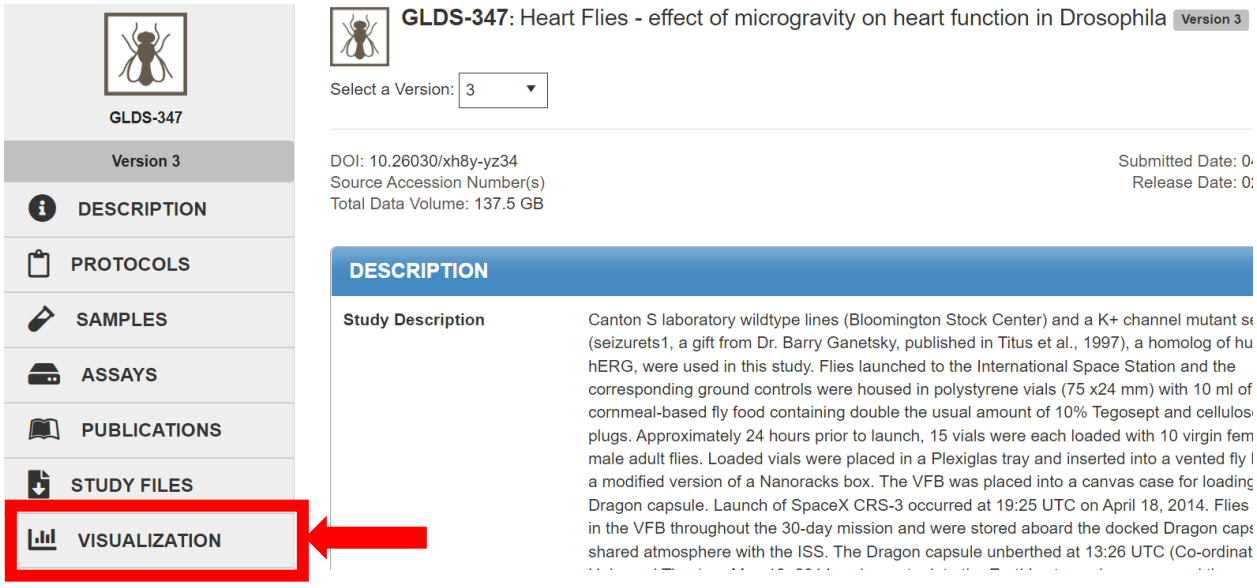
Los **gráficos PCA** agrupan las muestras en función de su similitud, utilizando esta información para representar la variación en el conjunto de datos. Se utilizan típicamente para conjuntos de datos grandes y complejos, como los que se producen en los análisis -omics, que tienen múltiples dimensiones que se mapean en un plano más simple. Por ejemplo, un gráfico PCA se puede utilizar para visualizar si un ratón en una muestra de vuelo es más voluminoso que los otros ratones de vuelo o ratones terrestres.

Conjuntos de datos

El conjunto de datos utilizado en esta actividad es GLDS-347: Heart Flies – efecto de la microgravedad en la función cardíaca en Drosophila. Las *Drosophila* se utilizan a menudo como organismos modelo al estudiar trastornos cardíacos y la genética asociada con el desarrollo de órganos y tejidos. Su genoma es aproximadamente un 60% homólogo al de los humanos, y alrededor del 75% de las enfermedades humanas tienen homólogos genéticos en Drosophila.

Activity

1) Navega al Repositorio de Datos GeneLab (https://genelab.nasa.gov). Haz clic en el botón de *Data Repository* (Repositorio de Datos).
2) Usando la barra de búsqueda, navega a GLDS-347 o utiliza el hipervínculo en la sección de Conjuntos de Datos anterior.
3) En el panel de navegación a la izquierda, haz clic en la pestaña *Visualization* (Visualización). (Pueda que sea lento al cargar inicialmente.)



4) La página se poblará con varias visualizaciones, incluyendo un gráfico PCA, un mapa de calor y un gráfico de volcán. Este proceso puede tardar un poco más en cargar dependiendo de tu conexión a internet, por favor sé paciente.
5) Explora los gráficos, comenzando con una evaluación preliminar de lo que se presenta en cada gráfico.
    c. Intenta hacer clic en la opción "**2D**" para el gráfico PCA, luego haz clic en *Update* (Actualizar).

i. Los ejes indican una relación entre múltiples muestras y múltiples variables.
ii. Los gráficos PCA típicamente representan agrupamientos. En este caso, las diferencias entre *wildtype* (también llamado 'salvaje' o 'silvestre') y vuelo espacial están más estrechamente agrupadas, lo que significa que tienen menos variabilidad interna, en contraste con los datos de ambos mutantes *seizure channel mutants*.
iii. ¿Cuáles podrían ser algunas razones para que esto ocurra? (Las respuestas pueden variar, por ejemplo, podrían abordar variaciones en los organismos individuales, como el tamaño, etc.)

b. Intenta hacer clic en el mapa de calor para una vista ampliada.
   i. Esta visualización muestra que hubo tanto mutantes *seizure channel mutants* como moscas de tipo *wildtype* en condiciones de vuelo espacial y de Tierra.
   ii. El gen Nplp3 está más fuertemente expresado en el vuelo espacial que en el control de Tierra tanto para el *wildtype* como para los mutantes *seizure channel mutants*. Esto se indica por un color azul más oscuro en las muestras de vuelo espacial comparado con las muestras de control de Tierra.

d. Intenta pasar el cursor sobre los puntos rojos y azules en el gráfico de volcán.
   i. El punto rojo más a la derecha (CG9684 X: 12.92, Y: 16.30) es el gen más sobreexpresado. El punto rojo que aparece más alto en el eje vertical es el gen sobreexpresado más estadísticamente significativo (CG14352 X:11.59, Y: 28.36).
   ii. El punto azul más a la izquierda es el gen más sobreexpresado. ¿Cuál es su identificador? (Respuesta: Alp9). El punto azul más alto en el eje vertical es Mal-A1. ¿Qué significa la posición de este gen en este gráfico? (Respuesta: Es el gen subexpresado más estadísticamente significativo.)

¿Por qué es esto importante?

Evaluar rápidamente un conjunto de datos es útil para un investigador, y las visualizaciones facilitan esto, especialmente con grandes cantidades de información. Un investigador, por ejemplo, puede ver rápidamente qué genes están más sobreexpresados o subexpresados en un conjunto.

a. Navega a FlyBase.org, una base de datos de genes y genomas de Drosophila. En la barra de búsqueda, ingresa el nombre de uno de los genes sobreexpresados del gráfico de volcán GLDS-347, como CG9684. ¿Qué información se incluye en el resumen del gen? (CG9684 codifica una proteína que pertenece a la familia TDRD1 que se predice que tiene un papel en la vía piRNA.)
b. Explora los resúmenes de otros genes sobreexpresados. ¿Tienen algo en común?¿Qué tienen en común? (Las respuestas pueden variar.)

NGSS Standards

Áreas: HS-LS1-2; HS-LS1-3; HS-LS4-1

Prácticas: Desarrollo y uso de modelos; Hacer preguntas y definir problemas; Análisis e interpretación de datos

Conceptos transversales: Interdependencia de la ciencia, la ingeniería y la tecnología; Influencia de la ingeniería, la tecnología y la ciencia en la sociedad y el mundo natural