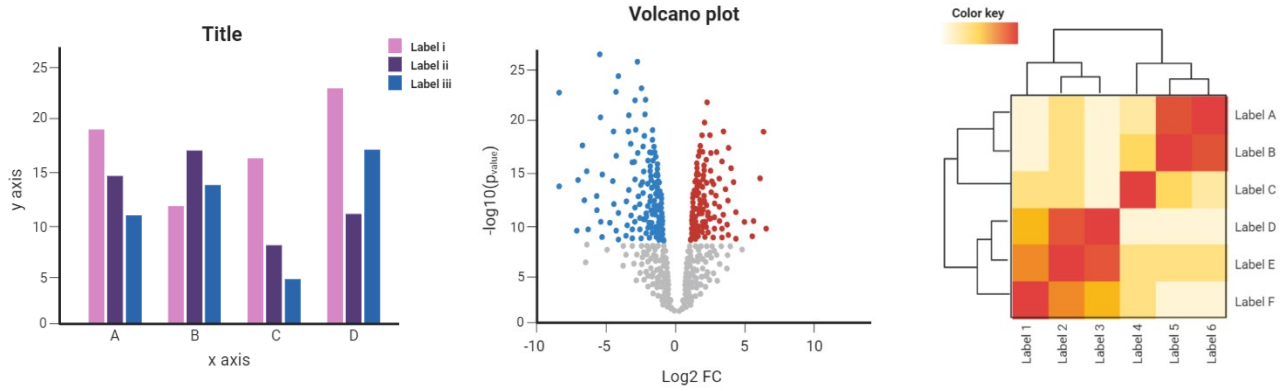
How can we use open space biology data to learn about data visualizations?



Background

Many forms of data visualizations exist and each serves a specific method of representing information to help identify trends, outliers, and patterns. In -omics analysis, three common plots are **heat maps**, **volcano plots**, and **principal component analysis (PCA) plots**.

Heat maps are used to demonstrate magnitude through a color intensity gradient mapped in a two dimensional matrix. In an omics example, rows typically represent a gene identifier and columns indicate the RNAseq sample. The color intensity, in this case, consequently reports the degree to which the gene is expressed.

Volcano plots are a subcategory of scatter plots that illustrates a relationship between statistical significance (also called the false discovery rate) and fold change (indicative of biological significance). These plots are used to identify the most highly up- or down-regulated genes on either edge (left or right) and the most statistically significant genes appear at the top of the plot.

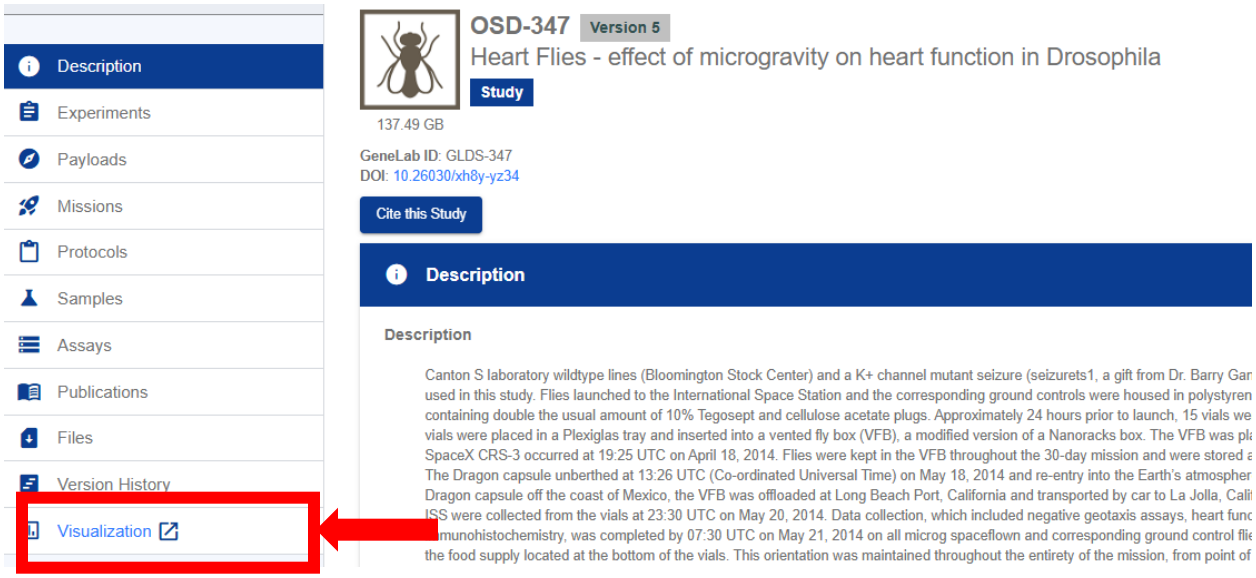
PCA plots will cluster samples based on similarity and uses this information to represent variation in the dataset. It is typically used for large and complex datasets, such as the type that occurs from -omics analyses, that have multiple dimensions that become mapped onto a simpler plane. For example, a PCA plot can be used to visualize whether one mouse in a flight sample is chunkier than the other flight mice or ground mice.

Datasets

The dataset used in this activity is OSD-347: Heart Flies – effect of microgravity on heart function in Drosophila. *Drosophila* are often used as model organisms when studying heart disorders and genetics associated with organ and tissue development. Their genome is approximately 60% homologous to humans and approximately 75% of human diseases have homologues of genes responsible for human diseases.

Activity

1) Navigate to the Open Science Data Repository (from https://nasa.gov/osdr/, scroll down and click on the *Explore the Data Repository* button).
2) Using the search bar, navigate to **OSD-347** or use the hyperlink in the Datasets section above.
3) In the left navigation panel, click on the tab for "Visualization". (Note it might be slow at initial loading.)



4) A page will populate with several visualizations, including a PCA plot, heat map, and volcano plot. This might take a little bit of extra processing time depending on your internet connection, so please be patient.
5) Explore the plots, beginning with a cursory evaluation of what is presented in each plot
    a. Try clicking on the option for "**2D**" for the PCA plot, then click on **Update**.
        i. The axes tell you a relationship between multiple samples and multiple variables
        ii. PCA plots typically represent clustering. In this case, differences between the wildtype and spaceflight are more closely clustered, meaning they have less internal variability, in contrast to data from both of the seizure channel mutants.
        iii. What might be some reasons that this happens? (Answers will vary, but could address variations in individual organisms such as size, etc)
    b. Try clicking on the heat map for an expanded view.
        i. This visualization shows that there were both seizure channel mutants and wild type flies in both spaceflight and ground conditions.
        ii. Gene Nplp3 is more strongly expressed in spaceflight than in the ground control for both the wildtype and the seizure channel mutants. This is indicated by the blue color coding that is darker in the spaceflight samples than the ground control samples.
    c. Try hovering over the red and blue points in the volcano plot.

i. The red dot furthest to the right (CG9684 X: 12.92, Y: 16.30) is your most upregulated gene. The red dot that appears highest on the vertical axis is your most statistically significant upregulated gene (CG14352 X:11.59, Y: 28.36).
ii. The blue dot furthest to the left is your most down regulated gene. What is its identifier? (Answer: Alp9). The blue dot highest on the vertical axis is Mal-A1. What does the position of this gene on this plot mean? (Answer: It is the most statistically significant downregulated gene.)

Why Does This Matter?

Being able to evaluate a dataset quickly can be helpful to a researcher, and visualizations allow that to happen, especially when large quantities of information are involved. A researcher, for instance, can quickly see which genes are most up or down regulated in a set.

a. Navigate to FlyBase.org, a database of *Drosophila* genes and genomes. In the search bar, enter the name of one of the upregulated genes from the GLDS-347 volcano plot such as CG9684. What is reported in the gene summary? (CG9684 encodes a protein belonging to the TDRD1 family that is predicted to have a role in the piRNA pathway.)
b. Explore the summaries of other upregulated genes. What do they have in common, if anything? (Answers will vary)

NGSS Standards

Strands: HS-LS1-2; HS-LS1-3; HS-LS4-1
Practices: Developing and Using Models; Asking Questions and Defining Problems; Analyzing and Interpreting Data
Crosscutting Concepts: Interdependence of Science, Engineering, and Technology; Influence of Engineering, Technology, and Science on Society and the Natural World