Module 4.1: Learn-Biology

Overlap Analysis

Overlap analysis refers to the simple question of how many items are overlapping between two lists. In gene expression analysis, we are generally asking the question of how many genes are expressed in common between two tissues, two conditions, two cell lines, etc. When the sources of the lists we are analyzing are independent from one another, we can use a statistic called the hypergeometric p-value to ask how the number of overlapping genes compares to what we would get by chance with lists of given lengths, but in our case we will be comparing differentially expressed gene lists derived from different analysis of the same original data (so definitely not independent sources).

One of those most common and intuitive ways to visualize overlap is using a Venn diagram. This is a figure that shows overlapping shapes to represent logical relationships between two groups. A simple example using Dr. Wilson's favorite animal, the platypus, is shown below:

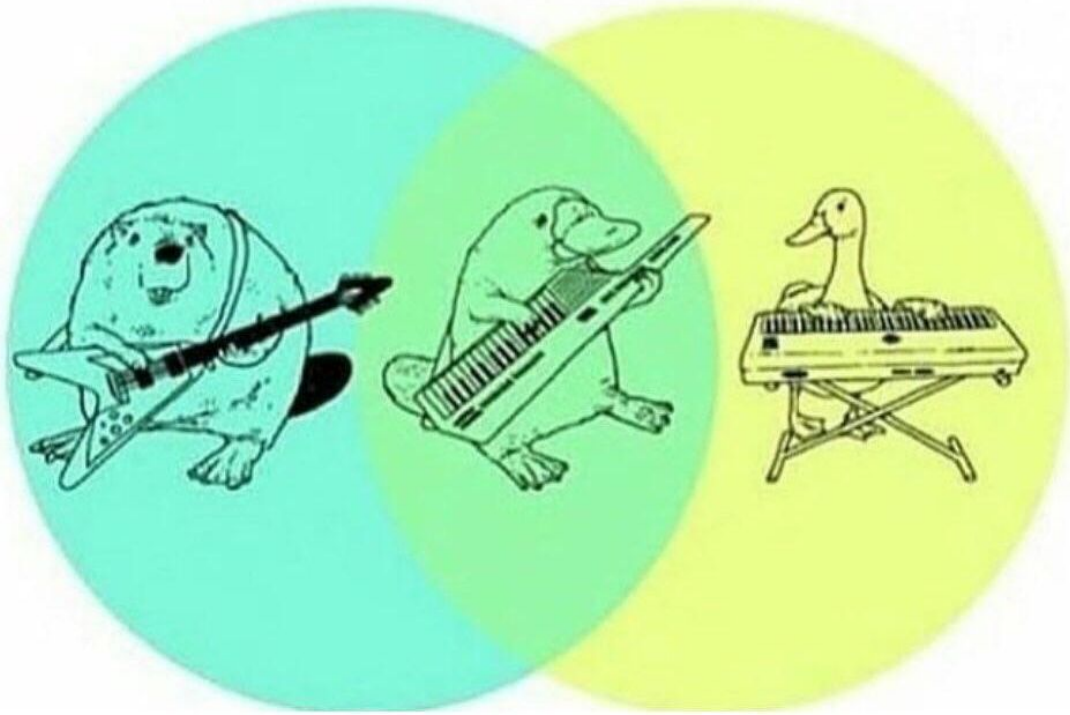


Figure. Platypus diagram.

Source. (https://imgflip.com/memetemplate/160845248/Venn-Diagram)

The shapes are generally circles and limited to 2 or three groups being compared, but you can see in a fancier e
can go higher if you are creative about making sure the numbers can be read and interpreted easily. In genomics
representing the number of biomolecules (genes, proteins, etc) that are in the intersection of the groups being co

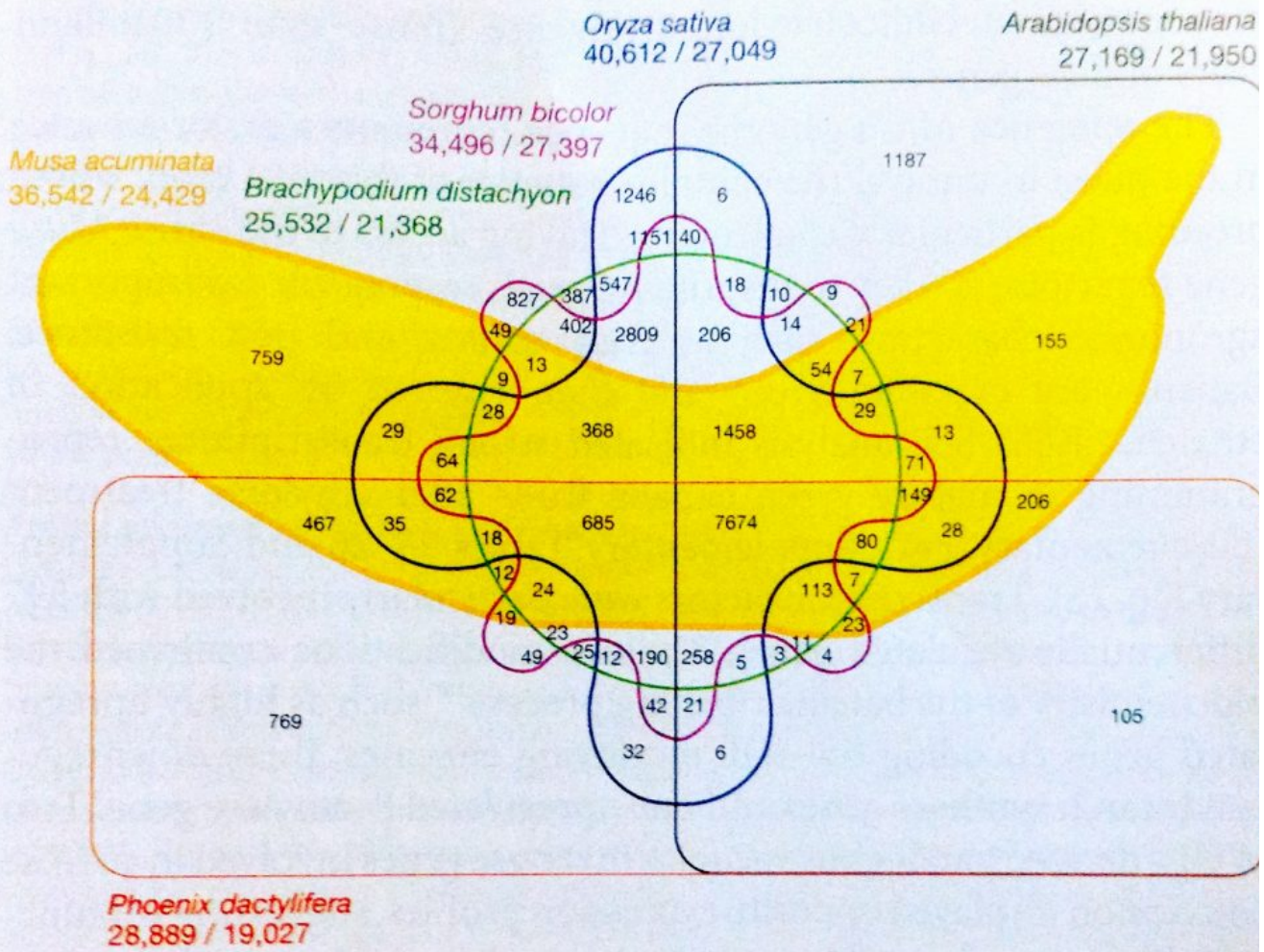


Figure 4 | Six-way Venn diagram showing the distribution of shared gene families (sequence clusters) among *M. acuminata*, *P. dactylifera*, *Arabidopsis thaliana*, *Oryza sativa*, *Sorghum bicolor* and *Brachypodium distachyon* genomes. Numbers of clusters are provided in the intersections. The total number of sequences for each species is provided under the species name (total number of sequences/total number of clustered sequences).

Figure. Venn diagram comparing plant genomes.

Source. (https://adamnorwood.com/notes/six-way-banana-venn-diagram/)

Correlation

In order to do overlap analysis with gene expression data, we need to apply a threshold for differential expressio of genes in each condition we are analyzing. Another question we can ask is how does the enter profile of gene between two conditions. That is, for example, do genes that have higher expression in females in one data set te expression in females in another data set? One measure we can use to answer this question is correlation. Give as average expression of a gene in two groups of samples, correlation refers to how well we can predict the valu other. There are many ways that we can assess this question, but a common statistic used is a Pearson correlati calculated by finding a best fit line through a map of 2 dimensional data and then adjusting based on how much c a whole have to that line. The correlation coefficient ranges from +1 as an exact positive correlation (when one v other increases) to 0 (the behavior of one variable has no effect on the other) to -1 (the two variables have oppos figure below shows this range for linear relationships on the top row. The correlation coefficient does not accoun best fit line (by how much one variable changes versus the other) as shown in the second row of the figure below to have nonlinear relationships (bottom row of figure below), so it is always best to visualize the scatter plot when correlation in case there is another statistic that is more appropriate.

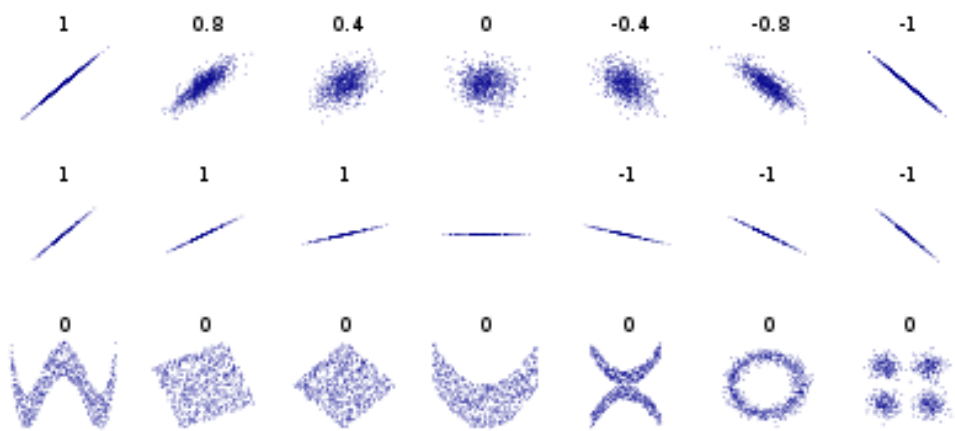
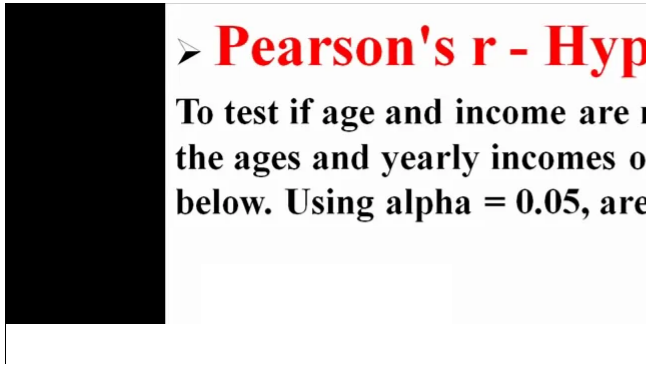


Figure. Correlation coefficient of x to y in various (x,y) data sets.

Source. (https://en.wikipedia.org/wiki/Correlation)

Pearson correlation coefficient, typically expressed as *r* or *R*, can be used to do hypothesis testing in order to ask correlation we are seeing between two variables is enough to determine whether those two variables are truly rel This video explains how the correlation coefficient is associated with a p-value so that we can do hypothesis testi learned about with t-test p-values in the previous module.



Video. Hypothesis testing with Pearson's r.

In this video, hypothesis testing with Pearson's r is used to test the relation of two datasets.

View Transcript. (https://canvas.asu.edu/courses/122165/files/54792233?wrap=1)

(https://canvas.asu.edu/courses/122165/files/54792233/download?download_frd=1)

We will be using R code in this module that does a Pearson correlation analysis to determine if fold change in fer using untrimmed data compares to the same using trimmed data. We would expect a fairly high positive correlat identify regions that have lower correlation than others and we can view the scatter plot to determine the slope of

Module 4.1 Additional Resources

- **StatQuest video on correlations** (https://www.youtube.com/watch?v=xZ_z8KWkhXE)
- Examples showing how the correlation coefficient is calculated and how is adjusted with the noise level
  - **Correlation - The Basic Idea Explained** (https://www.youtube.com/watch?v=qC9_mohleao)
  - **The Correlation Coefficient - Explained in Three Steps** (https://www.youtube.com/watch?v=ugd4k3d