Final Project Foundations of Data Science in LaTeX

Author: Donovan Nathaniel King

Fall 2025

1 Summary Table Demonstration

This is a table of summary statistics generated in Base R. It shows the vast variation in the expression of particular genes. Although the assignment stated that one used two genes, I used five to show off.

Table 1: Summary Statistics for Various RNA SEQ Counts

	TSPAN6	DPM1	SCYL3	C1orf112	FGR
Min.	1262	2922	1287	5617	2830
1st Qu.	1262	2922	1287	5617	2830
Median	1262	2922	1287	5617	2830
Mean	1262	2922	1287	5617	2830
3rd Qu.	1262	2922	1287	5617	2830
Max.	1262	2922	1287	5617	2830

2 Histogram of TSPAN6

I improved this histogram by adding color, adjusting the bins, and taking the log10 of the TSPAN6 values. As with other plots, this diminished the skew of extreme outliers. I wanted to overlay several transparent histograms of various genes, but after hours of frustration, I decided to keep it simple so I can turn this in on time.

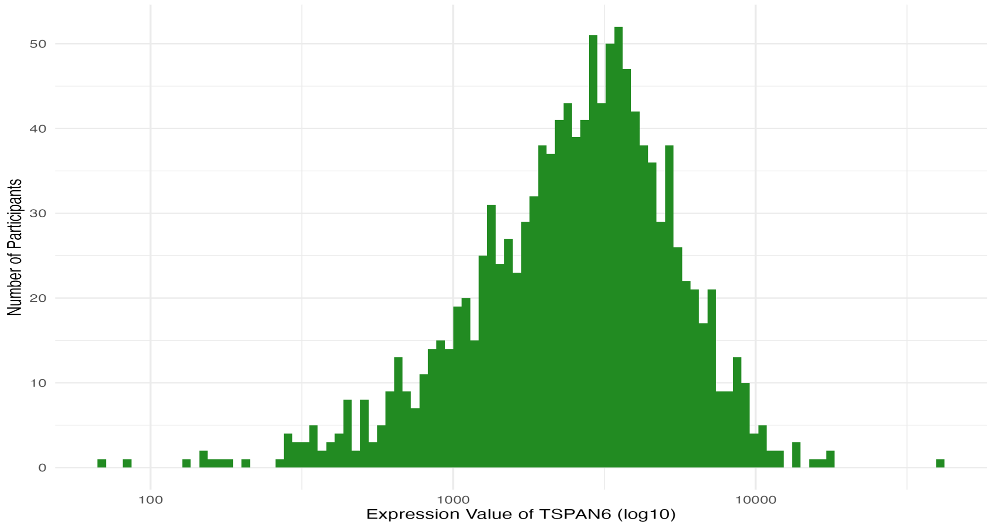

Figure 1: Histogram of log10-transformed TSPAN6 expression counts.

3 Scatter Plot of TSPAN6 and DPM1

This plot really shows the benefit of p-values and trend lines. The blob of points hints at a correlation, but the math proves it. Not the axes are logarithmic. This is, again, to compensate for extreme outliers.

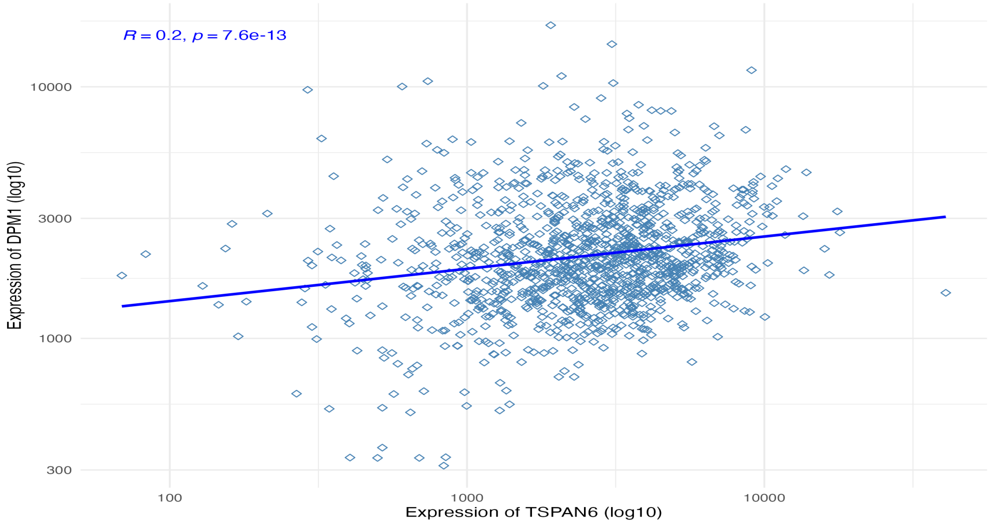

Figure 2: Scatter plot of the expression count for TSPAN6 versus DPM1 (log10).

4 Violin Duodectet Plot

This plot shows the density - that is, the most common values expressed - TSPAN6 in 12 disease stages. I have removed the grid lines to allow the shapes to tell the story, and again logged the y axis to account for outliers. The differing shapes show that TSPAN6 has distinct variation in the disease progression.

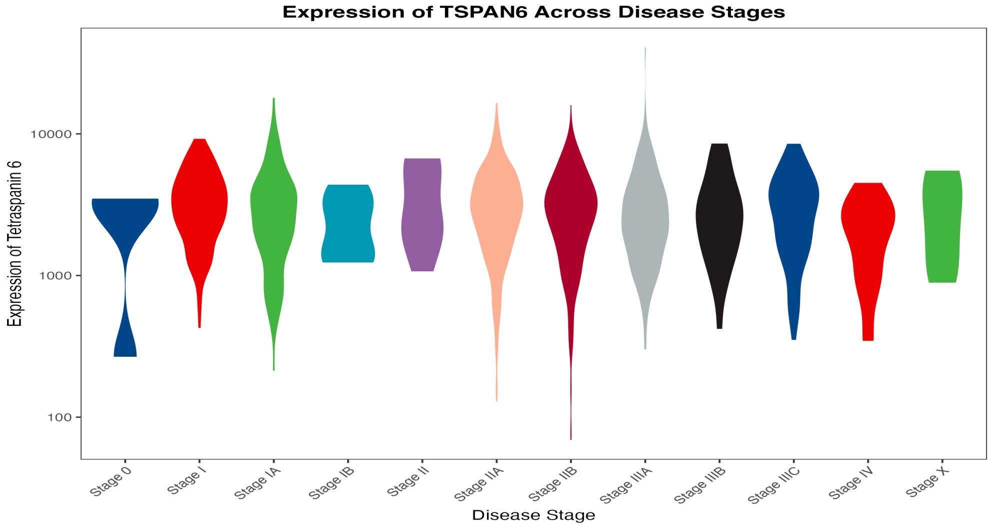

Figure 3: Violin plot mapping the expression of TSPAN6 to disease stages.

5 A Heat-Map of 10 Genes

This heat map shows the values of RNA SEQ counts. Each column represents a gene, and the rows represent the degree to which a participant expresses that gene. Most genes in my list exhibit low values. The plot is stretched to show the higher values within the heat map.

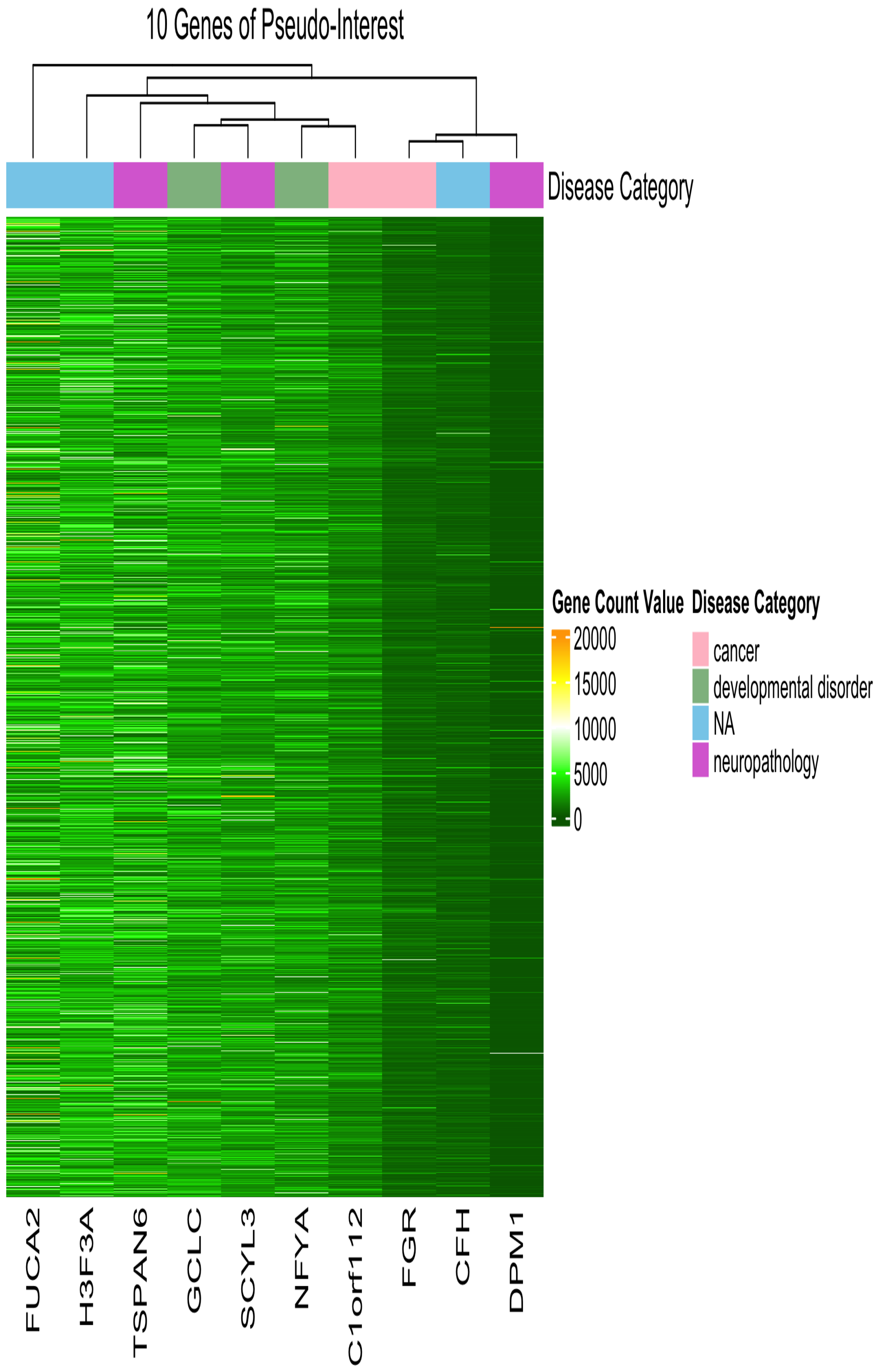

Figure 4: A heat map showing the similarity among participants of several genes.

6 My Chosen Plot

I chose a line plot, the classic "over time" plot. I looked to see if there was a variation in TSPAN6 expression and age of diagnosis. There is quite a bit of variation, but no trend showing a relationship.

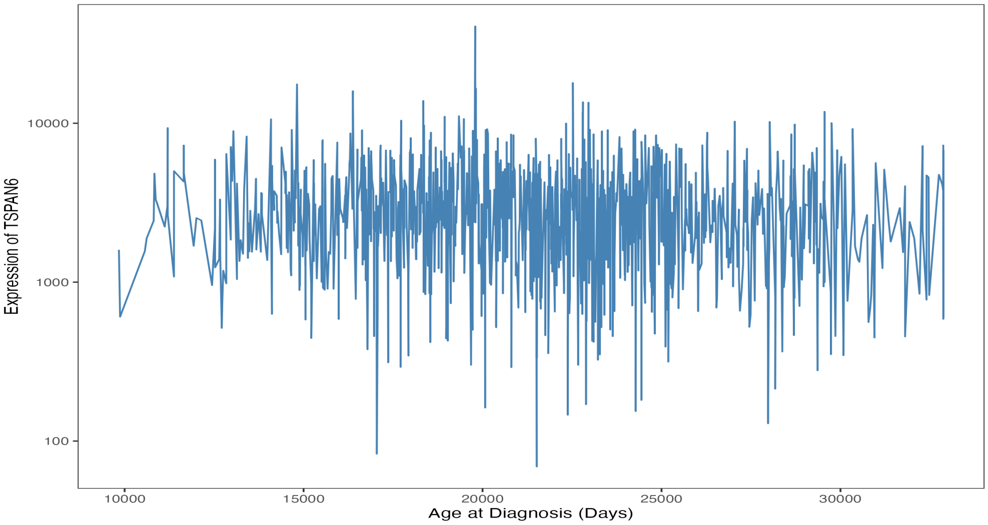

Figure 5: A line plot seeking a correlation with the expression of TSPAN6 and the age in which the patient was diagnosed.