# Final Project — RNA-seq Explorations

Jeremy Childress

October 19, 2025

## 1 Introduction

This project looks at RNA sequencing data from the TCGA-BRCA study, which includes over 1,200 breast cancer samples. The dataset also includes key clinical information like the patient's AJCC pathologic stage and whether they were alive or deceased at the time of data collection.

The main gene I focused on is `ENSG00000000003.15`, better known as **TSPAN6**. It's a cell surface protein that helps with communication between cells and can play a role in how cells grow or move. Because changes in these kinds of proteins are often linked to how cancers spread or respond to treatment, it seemed like an interesting target to explore.

My goal was to see how TSPAN6 expression varies across different stages of breast cancer and between living and deceased patients. I also looked at how it relates to another gene (`ENSG00000000005.6`) and included a broader look at 10 genes to spot any expression patterns in the larger dataset.

## 2 Methods

All analyses were done in R. I used `ggplot2` to make my histogram, scatter, violin, and density plots, and `ComplexHeatmap` to create a heatmap with clinical annotations for each sample. The scatter plot includes a fitted line from a simple linear model to show the trend between the two genes. The violin and density plots show how expression changes across cancer stages, and the heatmap helps visualize how different genes cluster together based on their expression levels.

The data for TSPAN6 and the comparison gene came from the RNA-seq count matrix, and the stage, sex, and vital status information came from the metadata file. I also calculated basic summary statistics (mean, median, standard deviation, etc.) using base R functions. All figures were exported as PNG files and inserted into this LaTeX document. Package versions are listed in my knitted R Markdown file using `sessionInfo()`.

## 3 Results

### 3.1 Summary Statistics

For the gene `ENSG00000000003.15` (TSPAN6), expression counts ranged widely across samples. Most were below 10,000 counts, but a few reached nearly 40,000, showing a strong right skew. The summary table below lists the main statistics.

Table 1: Summary statistics for `ENSG00000000003.15` counts ($n = 1231$).

| Min | 1st Quartile | Median | Mean | 3rd Quartile | Max | SD | N |
|---|---|---|---|---|---|---|---|
| 69 | 1562 | 2700 | 3208 | 4144 | 40780 | 2510.336 | 1231 |

### 3.2 Histogram

This figure shows the distribution of TSPAN6 counts across all samples. Most values are on the lower end, with a small number of very high outliers.

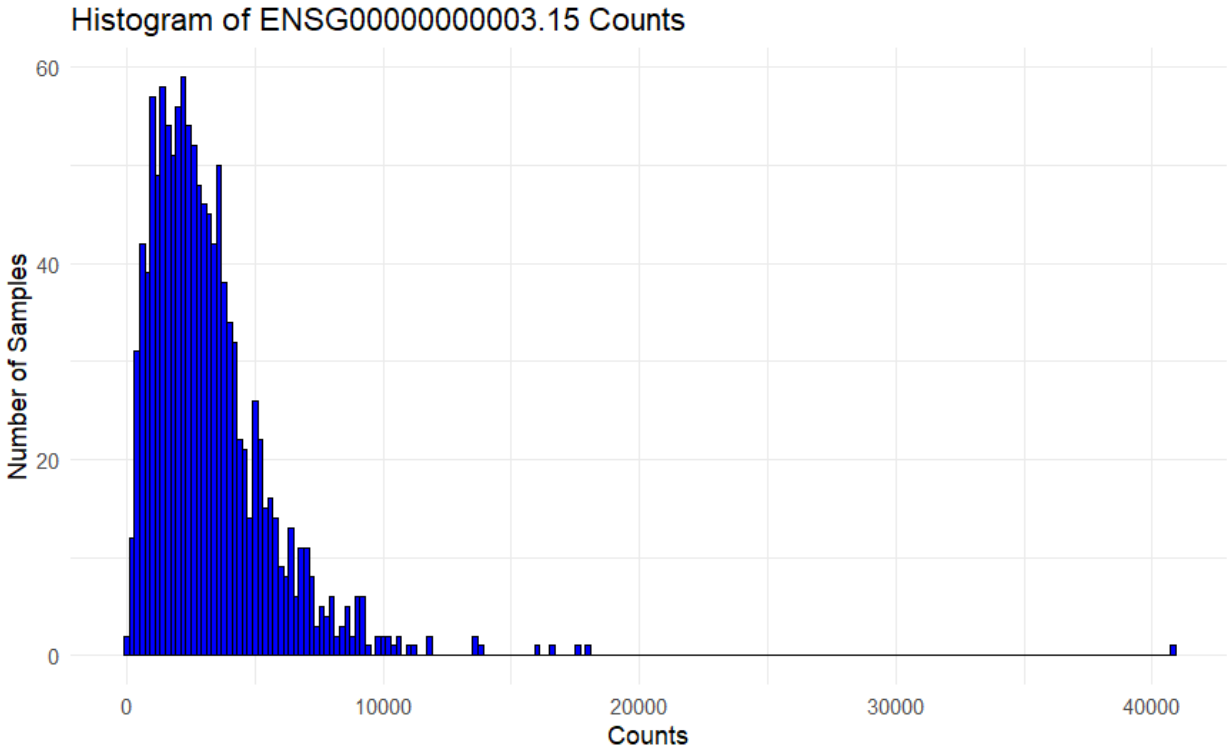


Figure 1: Histogram of TSPAN6 expression counts.

### 3.3 Scatter Plot

Here I compare TSPAN6 with `ENSG00000000005.6`. The fitted line helps show the overall trend, which is a weak positive relationship.

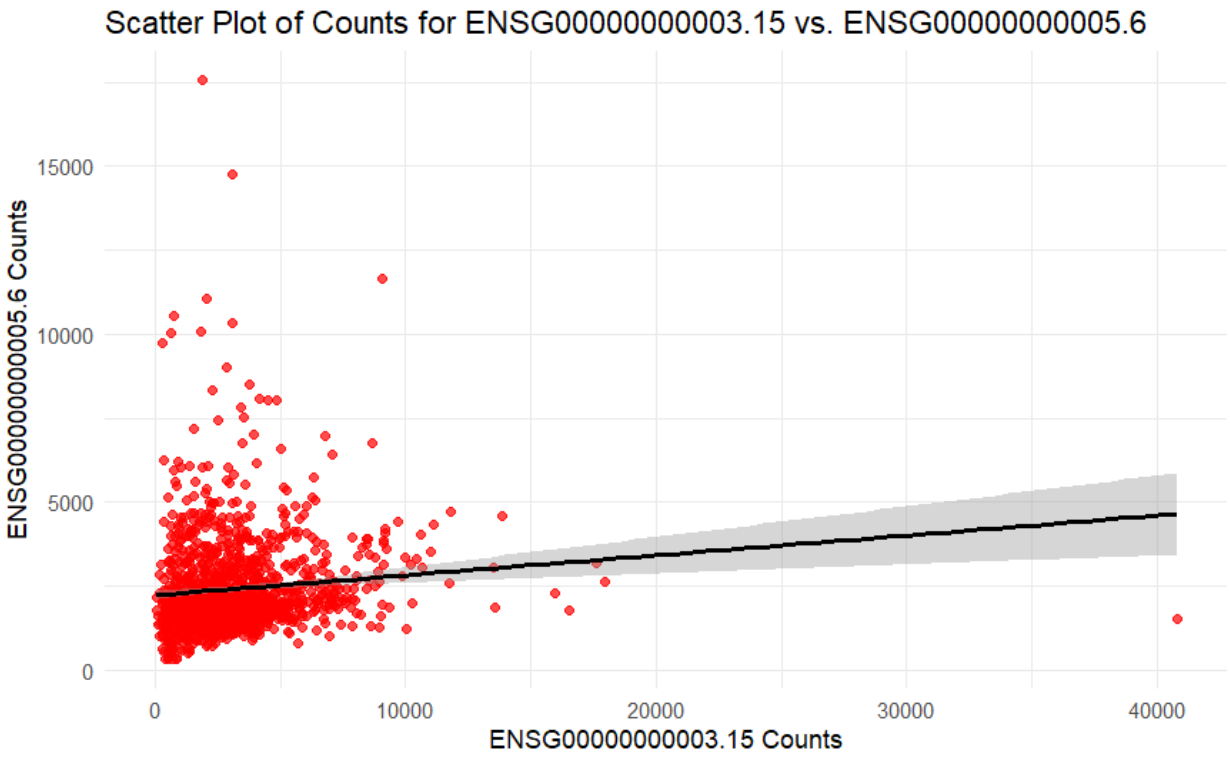


Figure 2: Scatter plot of TSPAN6 vs. `ENSG00000000005.6` with linear regression line.

### 3.4 Violin Plot

This plot shows how TSPAN6 expression varies by AJCC pathologic stage. Medians are similar across stages, with wider upper tails in later stages.

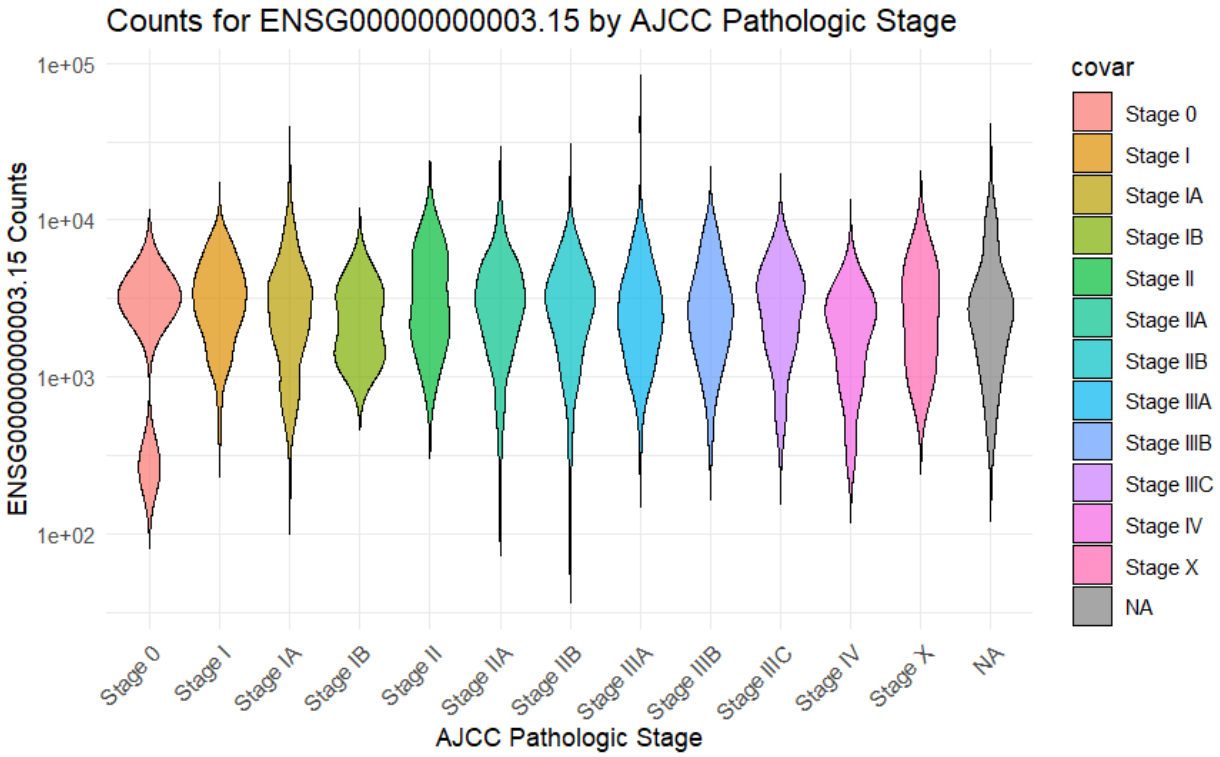


Figure 3: Violin plot of TSPAN6 expression across AJCC pathologic stages.

### 3.5 Ridgeline Densities

These ridgeline densities show the distribution of TSPAN6 within each stage. Most peaks are under 10,000 counts; later stages have slightly higher tails.

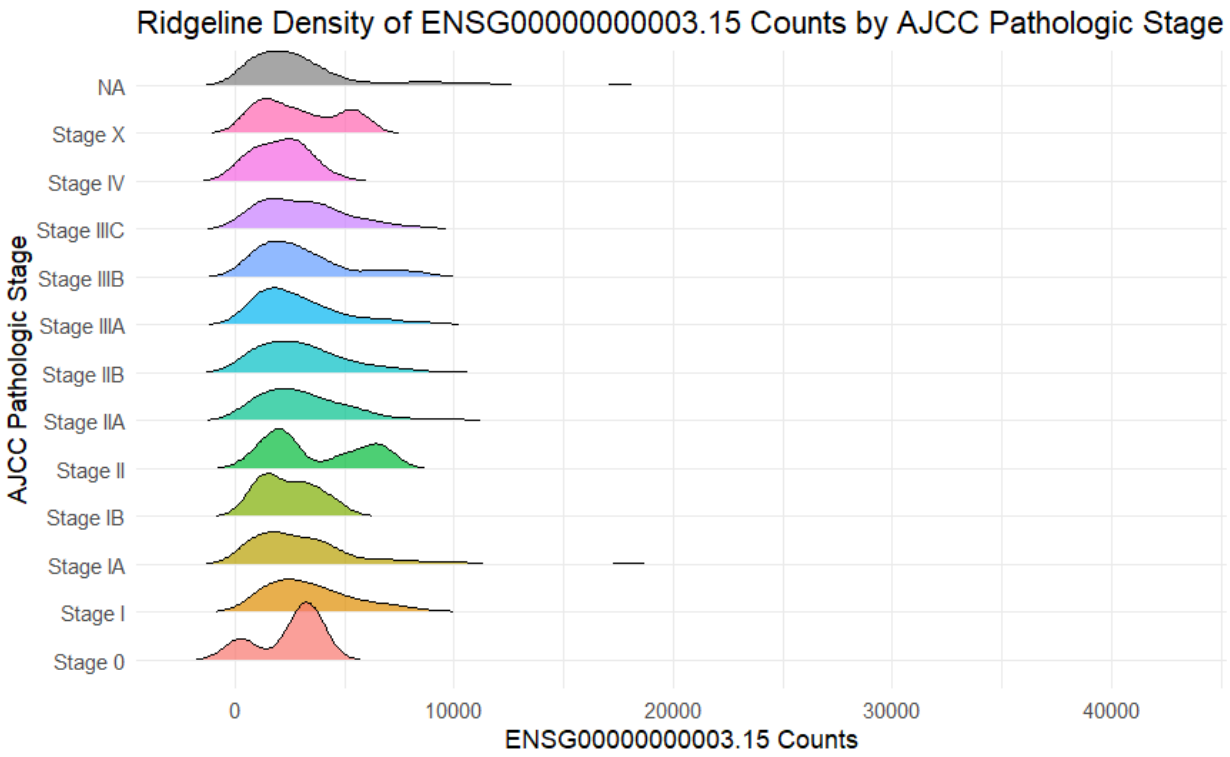


Figure 4: Ridgeline density of TSPAN6 expression by AJCC pathologic stage.

### 3.6 Heatmap

The heatmap shows ten genes across samples with two annotations (Vital Status and Sex). Some clustering lines up with Vital Status.

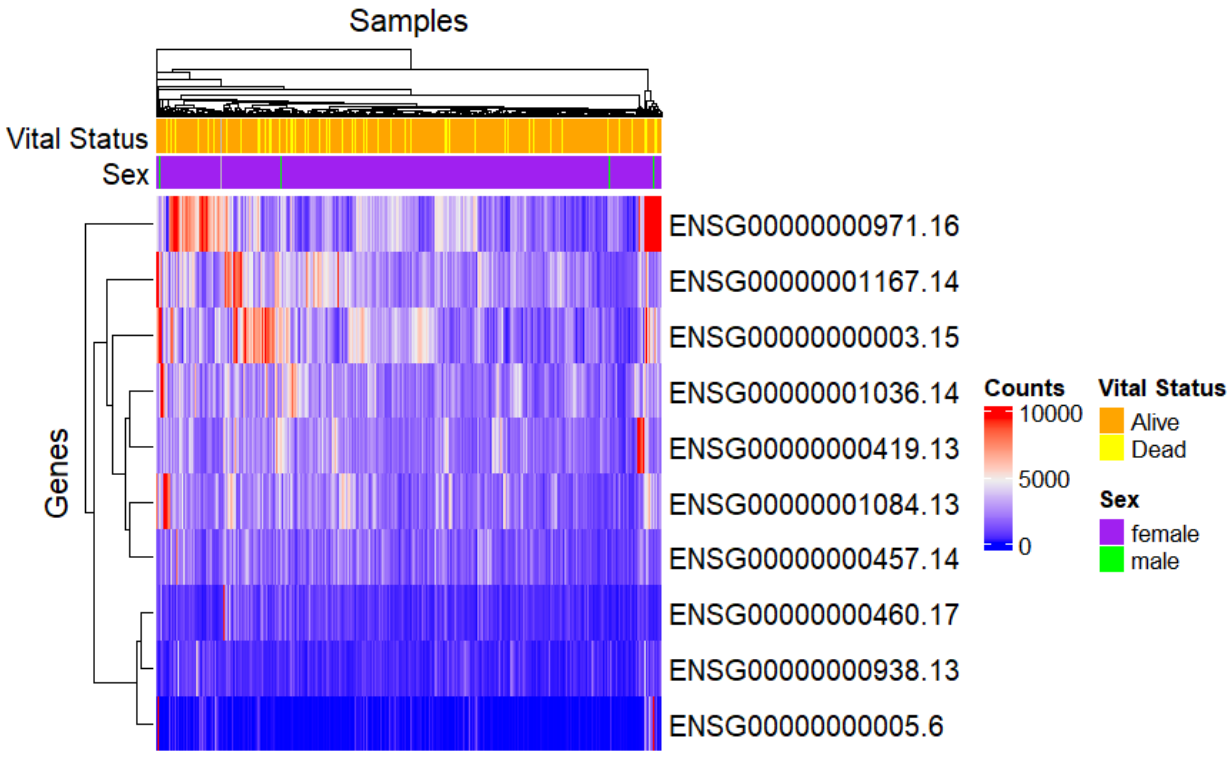


Figure 5: Heatmap of 10 genes annotated by Vital Status (Alive/Dead) and Sex (Female/Male).

## 4 References

The Cancer Genome Atlas (TCGA) Research Network. *TCGA-BRCA: Breast Invasive Carcinoma RNA-Seq Counts and Clinical Data*. Genomic Data Commons (GDC), National Cancer Institute. Available at: `https://portal.gdc.cancer.gov/projects/TCGA-BRCA`.

GeneCards. *TSPAN6 Gene (ENSG00000000003) – Tetraspanin 6*. Weizmann Institute of Science, GeneCards Human Gene Database. Available at: `https://www.genecards.org/cgi-bin/carddisp.pl?gene=TSPAN6`.

OpenAI. ChatGPT (GPT-5). Assisted in editing and formatting the LaTeX document and summarizing results. Available at: `https://chat.openai.com/`.