Gene Expression Analysis of Breast Cancer RNA-Seq Data from the TCGA-BRCA Cohort

Author: Savannah Wallis

Fall 2025

1 Introduction

This report analyzes various gene counts and covariate variables. The gene counts are RNA-Seq Data from breast cancer patients. Their metadeta, such as patient demographics, tumor characteristics, treatment information, survival data, and sample details, also are present in the dataset.

The goal is to better understand these selected genes and variables in relation to breast cancer. This data will provide research projects with a more narrow focus for breast cancer research initiatives. The overarching goal would be to provide information on the genes and factors that link breast cancer, in hopes of finding better treatment pathways.

Gene ENSG00000001460.18, Sperm Tail PG-Rich Repeat Containing 1, was a selected focus. It has been studied in relation to various cancers but no studies have ever been able to directly connect it to cancer.

1.1 Content

1. Gene 1: Summary statistics and count histogram
2. Gene 2: Summary statistics and count histogram
3. Scatter plot to compare Gene 1 and Gene 2
4. Violin plot to compare Gene 1 counts and tumor descriptors

5. Box plot to compare Gene 1 counts and tumor descriptors
6. Beeswarm plot to compare Gene 1 counts and tumor descriptors
7. Heatmap of 10 genes, including Gene 1 and Gene 2
8. Violin plot of Gene 1 counts over laterality
9. Raindrop plot of Gene 1 counts over sex and vital status

2 Methods

The methods involved using resources from Dartmouth Course HSE 711 to create plots and statistical analyses. This was carried out using R/Rstudio and various installed R packages. Additional external resources (stated in the reference section) were used to solidify the coding methods.

2.1 Base Packages

These basic plotting packages were already present on RStudio:
- grid
- stats
- graphics
- grDevices
- utils
- datasets
- methods
- base

2.2 ggplot2

ggplot2_4.0.0
This was installed to utilize its advanced plotting tools.

2.3 dplyr

dplyr_1.1.4
This package was installed for better data manipulation.

2.4 ggpubr

ggpubr_0.6.1
This was used to make the plots publish-ready through annotation and grid arrangement tools.

2.5 gridExtra

gridExtra_2.3
This package was used to create a table image from a data frame.

2.6 ComplexHeatmap

ComplexHeatmap_2.24.1
This heatmap package generates a comprehensive heatmap plot.

2.7 circlize

circlize_0.4.16
This package connects to ComplexHeatmap to help generate and stylize the plot.

2.8 ggbeeswarm

ggbeeswarm_0.7.2
This package was used to generate a beeswarm plot.

2.9 ggdist

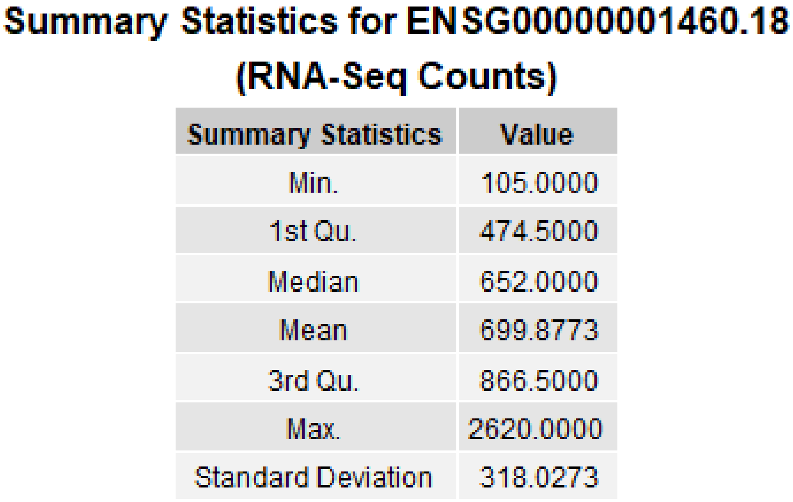
ggdist_3.3.3
This package created the half eye or raindrop plot.

3 Results

3.1 Gene 1: Summary statistics and count histogram

Gene 1 was selected to be ENSG000000014.60.18. The summary statistics for the count data were generated and are shown in Figure 1. You can find the Min and Max, Quartiles, Mean, Median and SD.

A histogram was also generated based off of the gene count data across all samples. The histogram can be found in Figure 2. No current analysis of the gene can be performed except that gene expression rates vary greatly from sample to sample.



Summary Statistics for ENSG00000001460.18 (RNA-Seq Counts)

Summary Statistics	Value
Min.	105.0000
1st Qu.	474.5000
Median	652.0000
Mean	699.8773
3rd Qu.	866.5000
Max.	2620.0000
Standard Deviation	318.0273

Figure 1: Summary statistics for Gene 1.

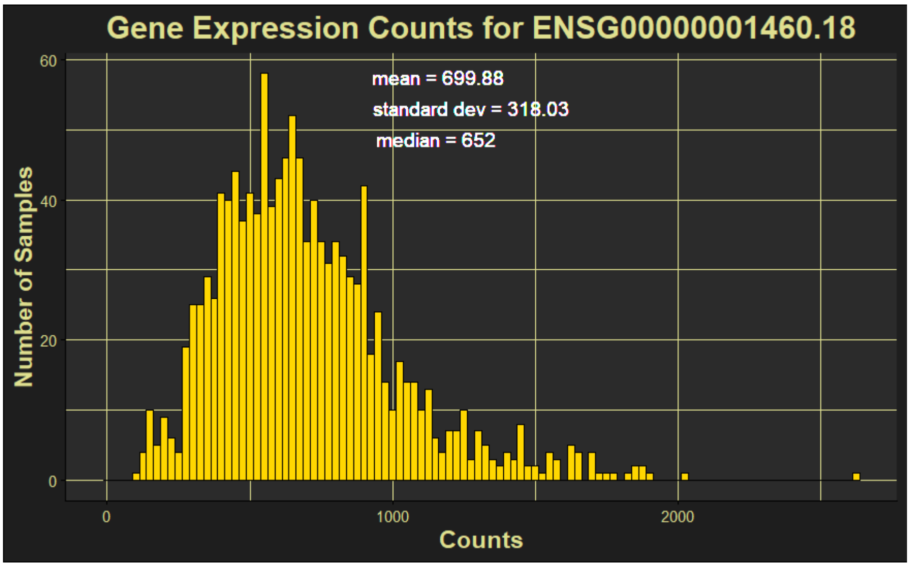
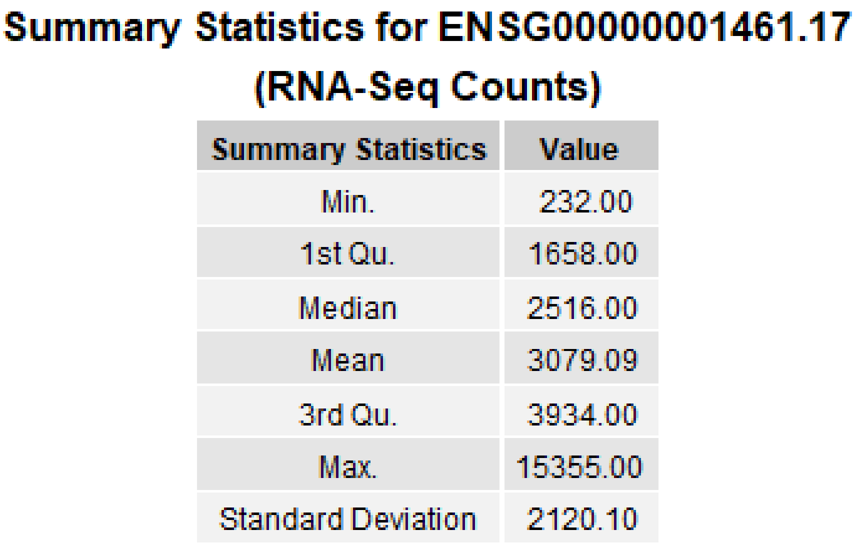


Figure 2: Histogram of counts for Gene 1.

3.2 Gene 2: Summary statistics and count histogram

Gene 2 was selected to be ENSG00000001461.17. Gene 2 is not the main focus in this report. It is being utilized as a way to compare values with Gene 1. The summary statistics for the count data are displayed in Figure 3. You can find the Min and Max, Quartiles, Mean, Median and SD.

A histogram was also generated for the gene count data across all samples. The histogram can be found in Figure 4. No current analysis of the gene can be deduced except that the gene expression rates are, on average, much higher than the expression counts of Gene 1.



Summary Statistics for ENSG00000001461.17 (RNA-Seq Counts)

Summary Statistics	Value
Min.	232.00
1st Qu.	1658.00
Median	2516.00
Mean	3079.09
3rd Qu.	3934.00
Max.	15355.00
Standard Deviation	2120.10

Figure 3: Summary statistics for Gene 2.

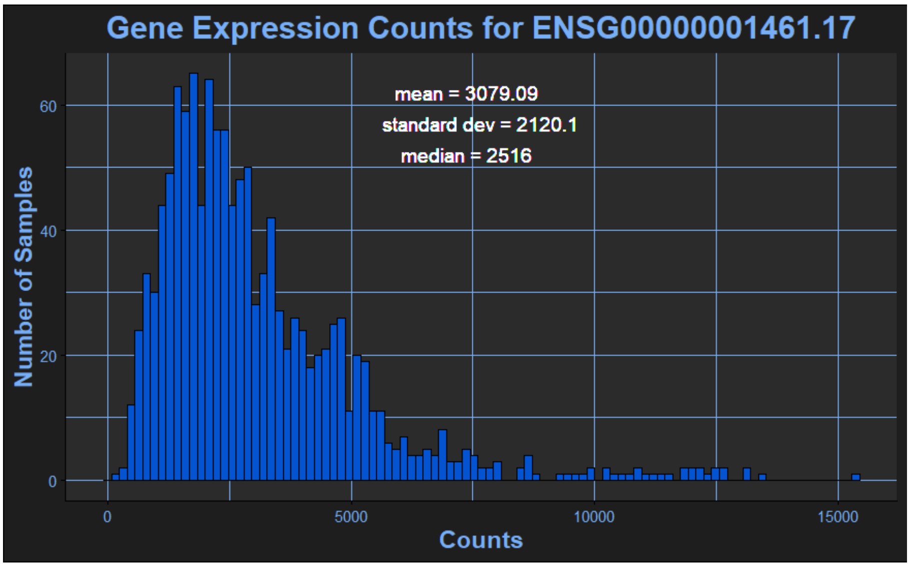


Figure 4: Histogram of counts for Gene 2.

3.3 Scatter plot to compare Gene 1 and Gene 2

A scatter plot was generated to compare count values on a linear regression between both genes. Gene 2 has a higher expression count on average. The equation of the regression resulted as y = 4.06x + 240.78 and the R squared was 0.37. From these numbers we can deduce that there is very little linear trend between the genes. See Figure 5.

$$y = 4.06x + 240.78 \quad (R^2 = 0.37)$$

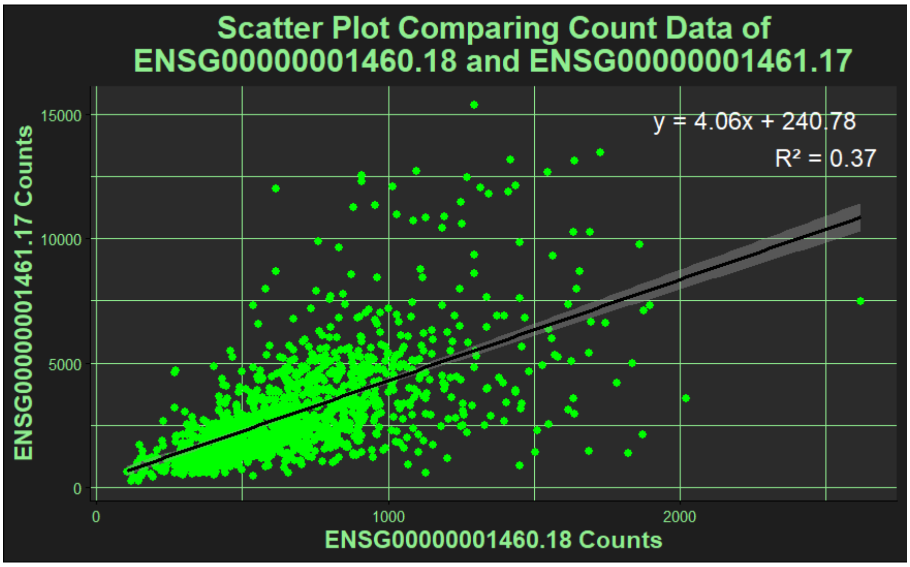


Figure 5: Scatter Plot of Gene 1 and Gene 2.

3.4 Violin plot to compare Gene 1 counts and tumor descriptors

The covariate from the metadata was chosen as the tumor descriptor for analysis. The categories were Primary, Metastatic, and Not Applicable. Not Applicable were treated as NA values simply because it does not add anything to the discovery. A violin plot with box plot analysis and jitters was formed from this data. The same Gene 1 expression values were utilized. From Figure 6, it is clear that primary tumors are much more common than metastatic. The expression for primary is also much more spread out, while metastatic is condensed to just a few data points around the mean. Therefore, Gene 1 is more applicable to primary breast cancer tumors.

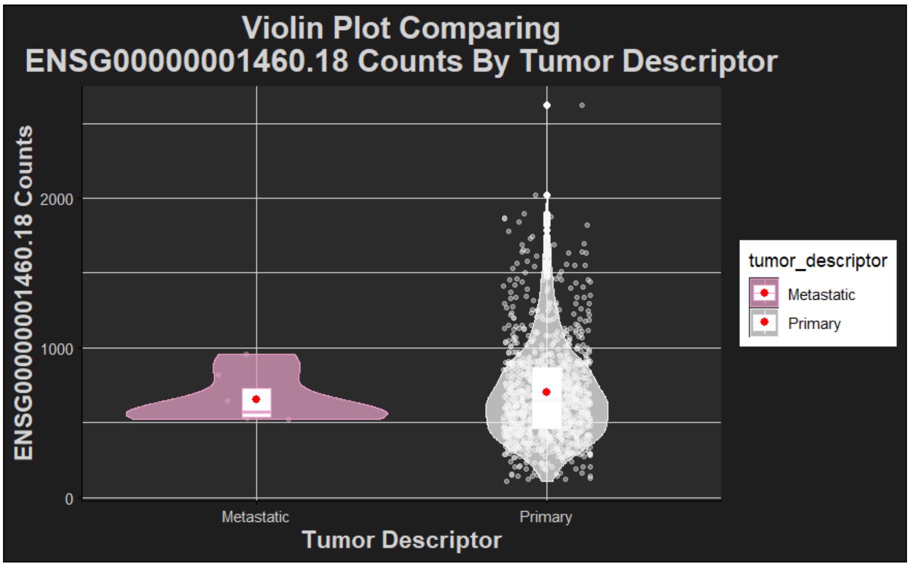


Figure 6: Violin Plot of counts and tumor descriptors

3.5 Box plot to compare Gene 1 counts and tumor descriptors

Again, we are using the same covariate of tumor descriptors and Gene 1 expression values. However, a box plot is utilized alone for easy viewing along with the addition of more annotations. See Figure 7. Upon this analysis, we actually see that the means of both categories are similar. However, the density of data for the primary tumor is significantly more dense and spread out. It is still safe to assume that expression for this gene in breast cancer patients hovers around 650-700 counts, yet is much more common in primary tumors.

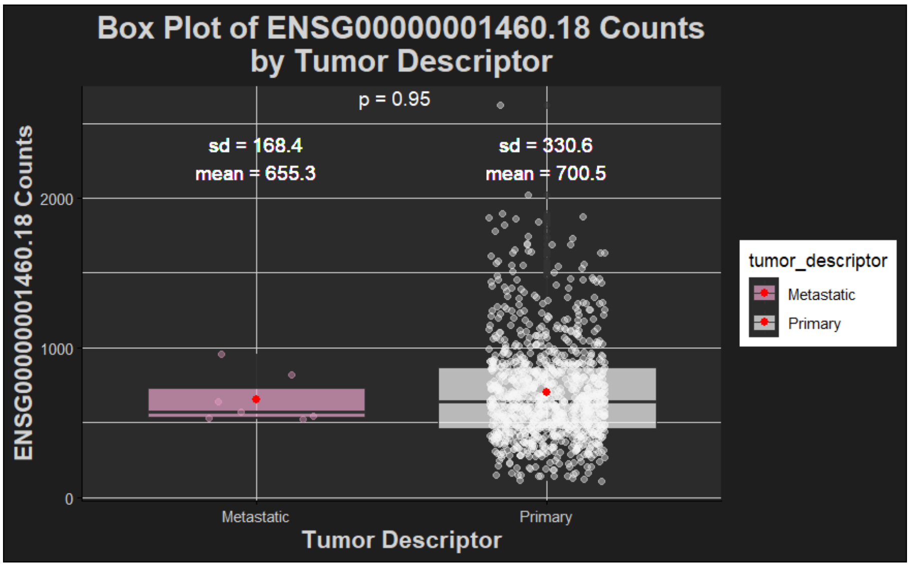


Figure 7: Box plot of counts and tumor descriptors

3.6 Beeswarm plot to compare Gene 1 counts and tumor descriptors

A beeswarm plot is utilized to see just how dense our data is for the tumor descriptor category. It is quite clear from Figure 8 that there is a difference in density. The lack of outliers is interesting as well.

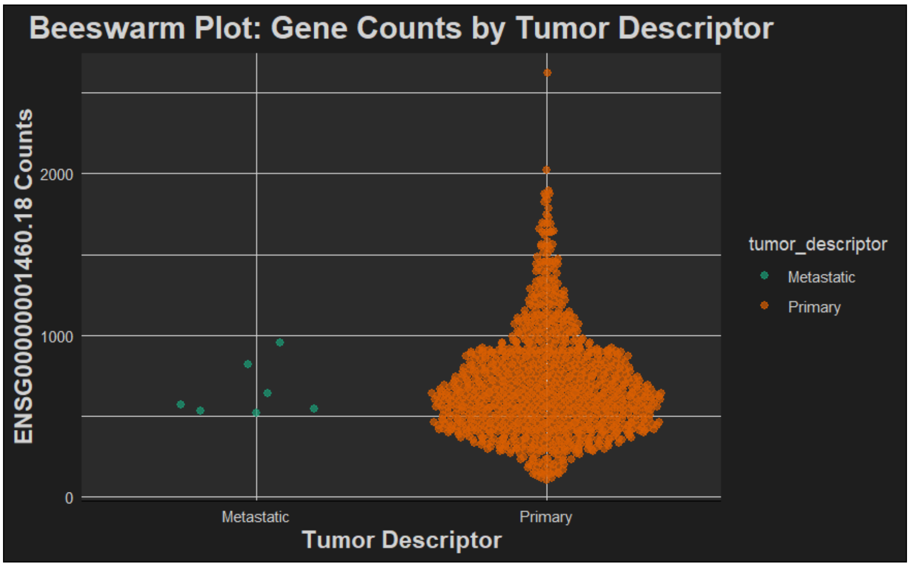


Figure 8: Beeswarm plot of counts and tumor descriptors

3.7 Heatmap of 10 genes, including Gene 1 and Gene 2

A heatmap of 10 random genes, including our two genes of interest (Gene 1 and Gene 2), was generated. From the heatmap (Figure 9), it is clear that the count density for Gene 1 is significantly lower than Gene 2. Both Genes are much lower in counts than other randomly selected genes. It also appears that out of the few metastatic samples, the expression of all of these genes is still low. However, after seeing this heatmap, ENSG000000016.17.12 might be a gene of interest -possibly linked to breast cancer. Please see the attached png in the GitHub Repository to explore the heatmap up close.

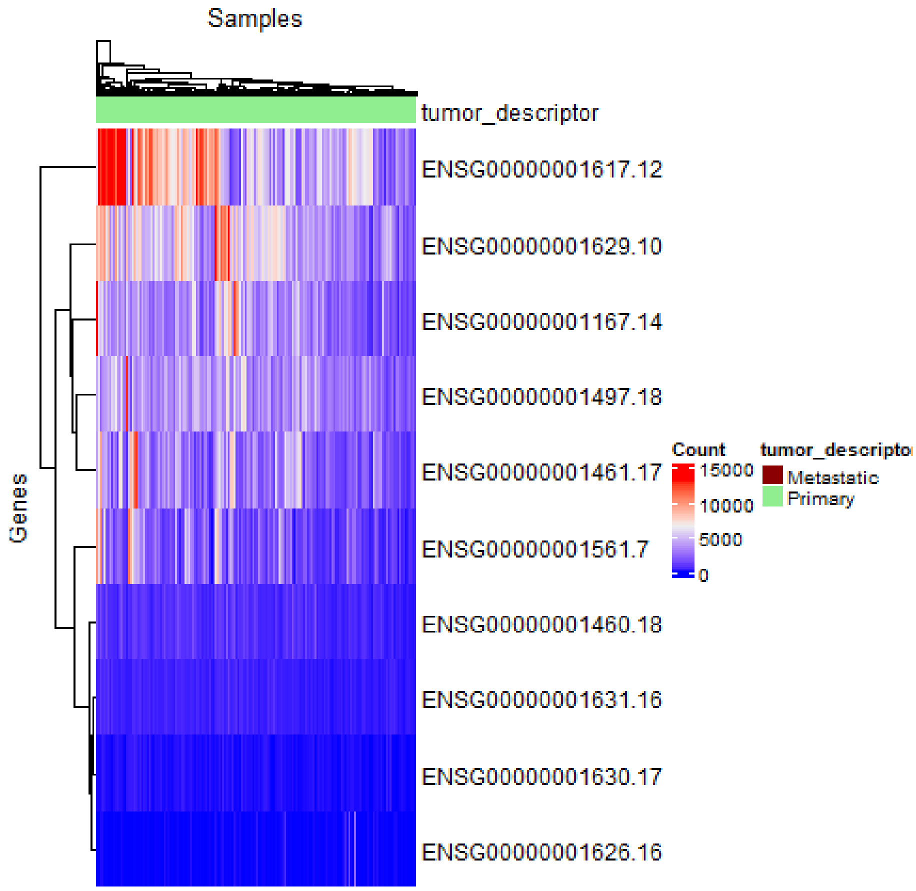


Figure 9: Heatmap of expression and tumor descriptors - no sample names

3.8 Violin plot of Gene 1 counts over laterality

To introduce a new covariate, laterality has been explored in connection to gene expression. Laterality explains which side of the body (or breast) the tumor sample was extracted from. Figure 10 shows almost exact symmetry between both the right and left side of the body. Therefore, there is no significant difference of Gene 1 expression counts in relation to laterality.

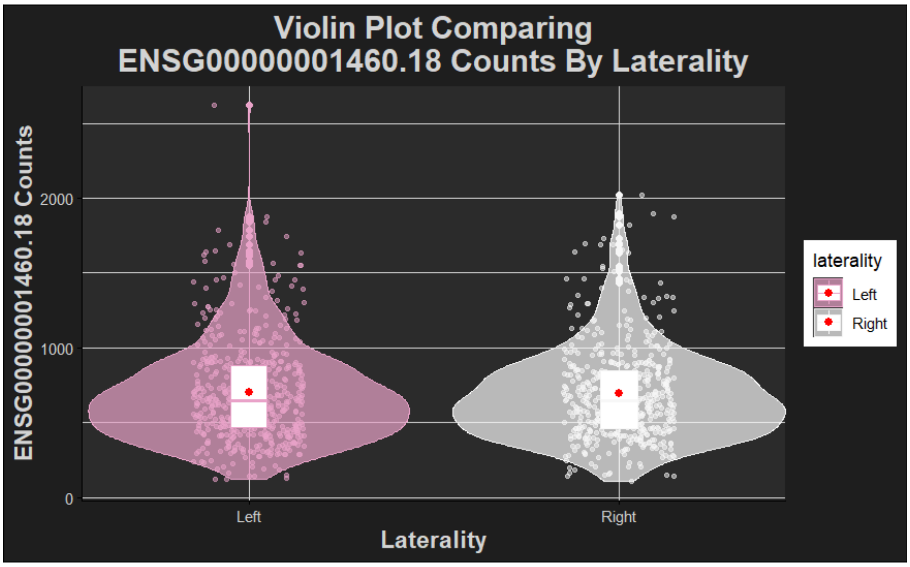


Figure 10: Violin plot of expression over laterality

3.9 Raindrop plot of Gene 1 counts over sex and vital status

Lastly, Figure 11 is a raindrop plot displaying the gene expression counts in connection to gender and vital status (alive or deceased). The plot shows a significant difference between male and female vitality. It appears that the female deceased and alive expression rates are very similar spreads. The male death rate is extremely low and almost non-existent compared to the number of living males. If further analysis were to be conducted, the lack of data for male breast cancer would likely need to be taken into account.

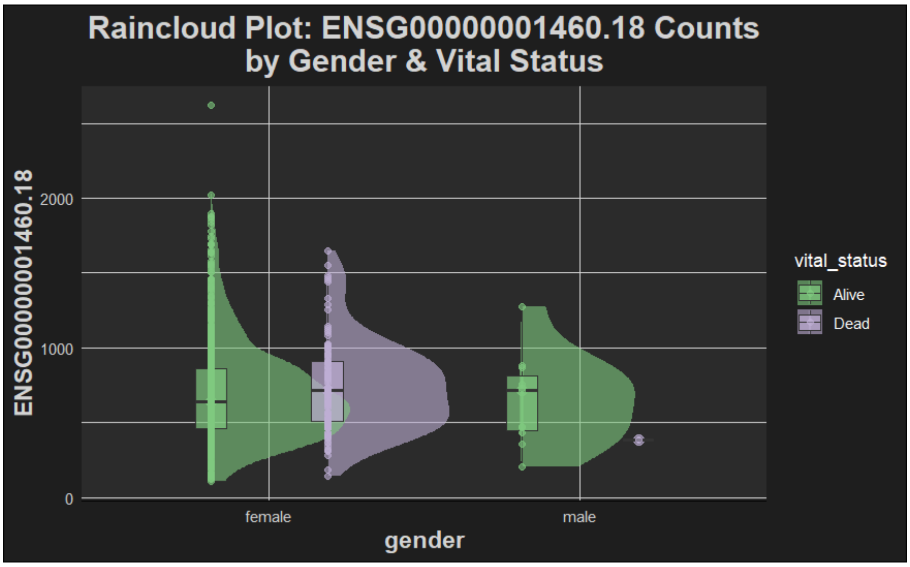


Figure 11: Raindrop plot of expression over gender and vital status

4 References

4.1 Dataset

The Cancer Genome Atlas Research Network. (2025). *TCGA Breast Invasive Carcinoma (TCGA-BRCA) [dataset]*. National Cancer Institute & National Human Genome Research Institute. Retrieved October 18, 2025, from https://portal.gdc.cancer.gov/projects/TC
BRCA

Atlas of Genetics and Cytogenetics in Oncology and Haematology. (2014, November 1). STPG1 (sperm tail PG-rich repeat containing 1). http://atlasgeneticsoncology.org/gene/74322/stpg1-

4.2 Main Coding

Dartmouth HSE 711, Dr. Noelle Kosarek. (2025). Course materials and coding guidance. Dartmouth College.

4.3 Additional Coding

Auguie, B. (2017, September 9). *Displaying tables as grid graphics*. R Project. Retrieved from https://cran.r-project.org/web/packages/grid

STHDA. (2025). *ggplot2 themes and background colors: The*

3 elements. Easy Guides Wiki. Retrieved from https://www.sthda.com/engl themes-and-background-colors-the-3-elements

Stack Overflow. (n.d.). *How to adjust the font size for axis labels in Complex Heatmap?* Retrieved from https://stackoverflow.com/quest to-adjust-the-font-size-for-axis-labels-in-complex-heatmap

Stack Overflow. (n.d.). *How to format ggplot geom_text with formula, getting unwanted "c(...)."* Retrieved from https://stackoverflow.co to-format-ggplot-geom-text-wit