Foundations in Data Science Final Project

Sarah Christensen

October 2025

1 Introduction

Gene Expression RNA-seq count data and metadata was compiled for patients who were part of The Cancer Genome Atlas (TCGA) Breast Invasive Carcinoma (BRCA) cohort. One of the genes with RNA-seq data is ENSG00000007202.15, which is named BLTP2 or bridge-like lipid transfer protein family member 2. According to a journal article by Banerjee, S., Daetwyler, S., Bai, X. et al, and published in Nature Cell Biology, the expression of the gene for a bridge-like Lipid transport protein (BLTP2) has a correlation with how aggressive a breast cancer is, with potentially connection with cell growth. Another gene that was analyzed, was ENSG00000001631.16, which is named KRIT1. According to a journal article by Orso, F., et al, the gene for KRIT1 might actually be a tumor suppressor, which expression decreases in response to miR-21, an often over-expressed small non-coding RNA in many cancers. This could explain why gene ENSG00000001631.16 RNA-seq count data is significantly lower than the count data for gene ENSG00000007202.15. However, it would be beneficial to know the normal expression levels of gene ENSG00000001631.16 in healthy cells for this comparison.

2 Methods

Multiple functions and packages were used to produce the figures and tables in this report. The Summary Statistics table in Table 1 was created with the group by and summarize functions from the dplyr package. The histogram in Figure 1 was created with the geom histogram from the ggplot2 package. The scatterplot and regression line in Figure 2 was created with the geom point and geom smooth, as well as the lm method from the ggplot2 package. To make the plots cleaner, the columns were capitalized with the str to title function from the stringr package. The violin plot in Figure 3 was created with geom violin and geom jitter from the ggplot2 package as well as the scale fill brewer function to set the palette for the category colors from the RColorBrewer package. The heatmap

in Figure 4 was created with the Heatmap and HeatmapAnnotation functions from the ComplexHeatmap package. The stringr package was also used to capitalize the Gender column for a cleaner plot. The density plot in Figure 5 was created with the geom density from the ggplot2 package. The age data was also cleaned up using the mutate function in dplyr to convert the ages in days to ages in years for a cleaner plot. The packages ggplot2, stringr, and dplyr are also a part of the tidyverse package.

3 Results

3.1 Summary Statistics of Gender

The RNA-seq count data for Gene BLTP2 or ENSG00000007202.15 as labeled in Table 1, was summarized by the gender of each patient in the BRCA cohort. While the expression values for this gene was typically high, compared to other genes, the average for Female patients was slightly lower than the Male patients, at 10712 vs 11806. The Female patients had the lower Minimum Value of 977 to the Male patient minimum value of 1228. The Female patients also had the dramatically higher Maximum Value with 91733 to the Male patient Maximum Value of 26815. The patient that was not identified as Female or Male had the count value of 8927, slightly below the average for the Male and Female patient count data for Gene BLTP2.

Gender	Average	Standard Deviation	Minimum Value	Maximum Value
Female	10712.60	7242.48	977	91733
Male	11806.69	8091.82	1228	26815
NA	8927.00	*NA*	8927	8927

Table 1: Summary Statistics of Gene ENSG00000007202.15 Expression by Gender (from RStudiio df)

Showing 1 to 3 of 3 entries, 5 total columns

3.2 Histogram

The Histogram in Figure 1 displays the frequency, or the number of patients with similar RNA-seq count data or expression of the gene ENSG00000007202.15 (BLTP2). This plot demonstrates that the majority of the patients in the BRCA cohort (Breast Invasive Carcinoma) had expression levels of the BLTP2 gene between 977 (the minimum value) and 25,000.

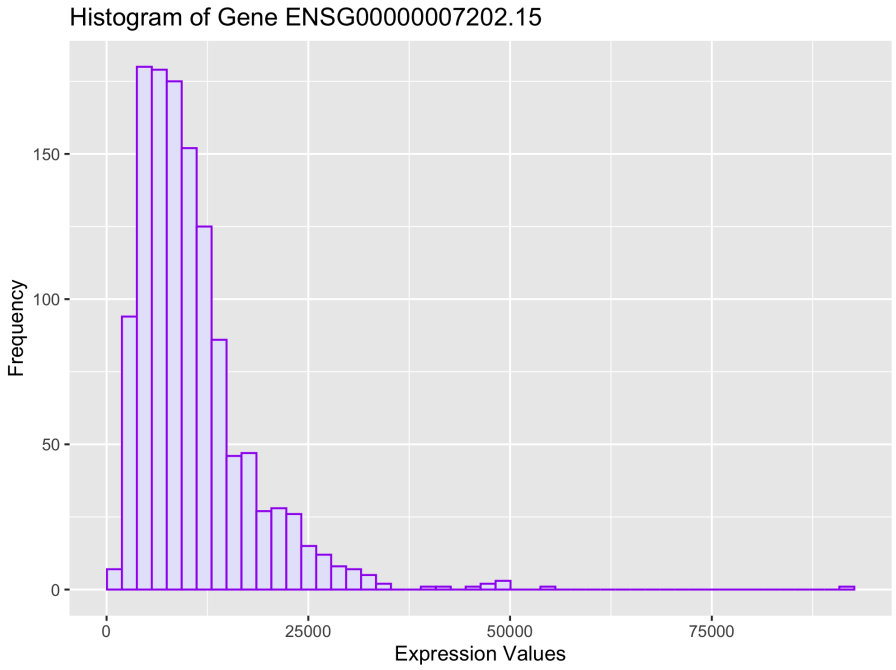

Figure 1: Histogram of Gene ENSG00000007202.15 Expression

3.3 Scatterplot

When comparing the RNA-seq count data of gene ENSG00000007202.15 (BLTP2) and gene ENSG00000001631.16 (KRIT1) via a scatterplot, as seen in Figure 2, there appears to be a slight positive relationship between the expression of the two genes, with the ENSG00000007202.15 (BLTP2) is expressed at much higher levels than the KRIT1 gene.

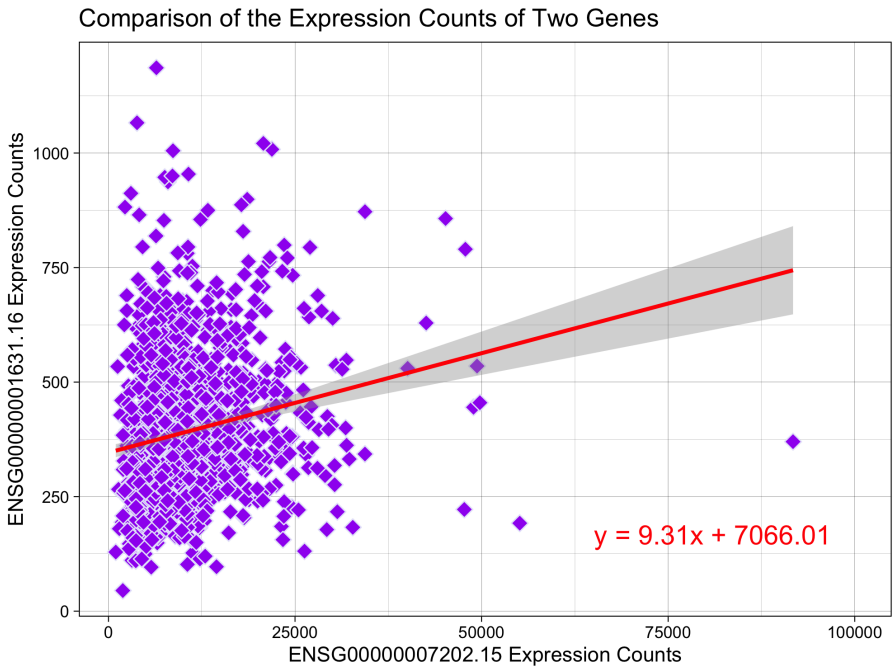

Figure 2: Relationship between the Gene Expression of ENSG00000007202.15 and ENSG00000001631.16

3.4 Violin Plot

When analyzing the RNA-seq counts for gene ENSG00000007202.15 by Gender in a Violin Plot, as seen in Figure 3, some of the female patients did have expression levels much higher than the male patients, but the sample size for the male patients is significantly smaller than the sample size for female patients. This is reinforced by the widest parts of the violin plots remaining below 25,000, or even 12,500, for both Male and Female patients. The one patient that was not identified as Male or Female also had a count below 12,500.

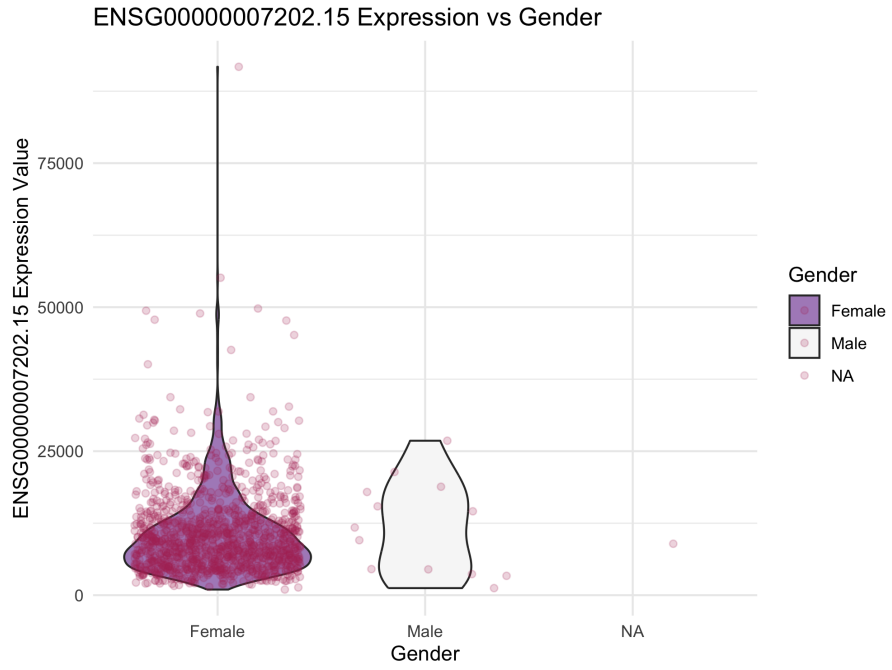

Figure 3: Expression of Gene ENSG00000007202.15 by Gender

3.5 Heatmap

When looking at the expression levels of 10 genes from the RNA-seq count data in the Heatmap in Figure 4 below, gene ENSG00000000971.16 had the highest expression levels (of the 10 genes analyzed) and gene ENSG00000000005.6 with the lowest expression levels (of the 10 genes analyzed) among the patients in the BRCA cohort.

3.6 Density Plot

The Density Plot in Figure 5 displays the age at which each patient was diagnosed with breast cancer. The density of ages at diagnosis were split by Gender, demonstrating that the peak ages for both Male and Female genders for diagnosis was around 60 years of

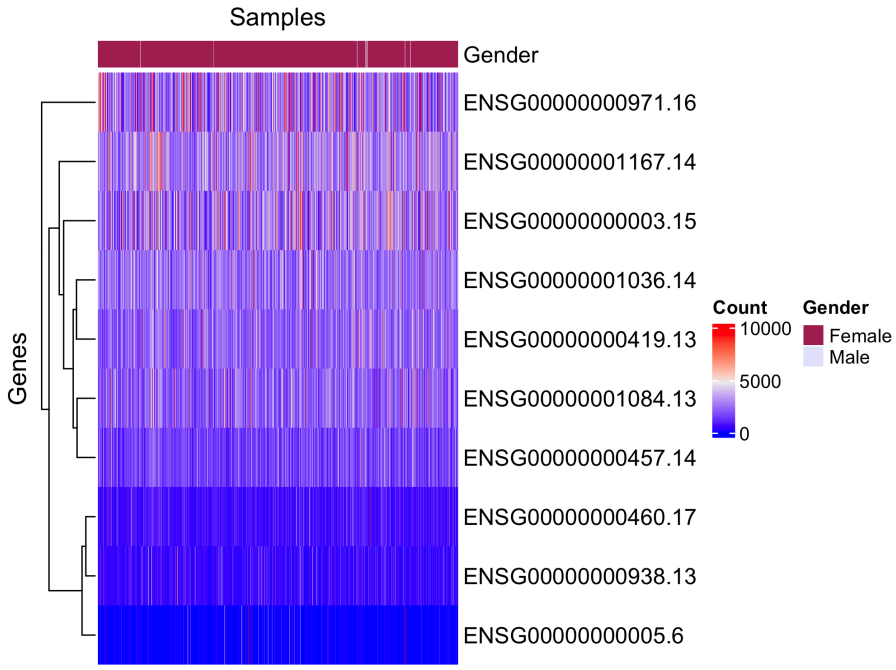

Figure 4: Heatmap of Gene Expression of by Gender

age, with some females being diagnosed sooner and some males had another smaller peak of diagnoses around 80 years of age.

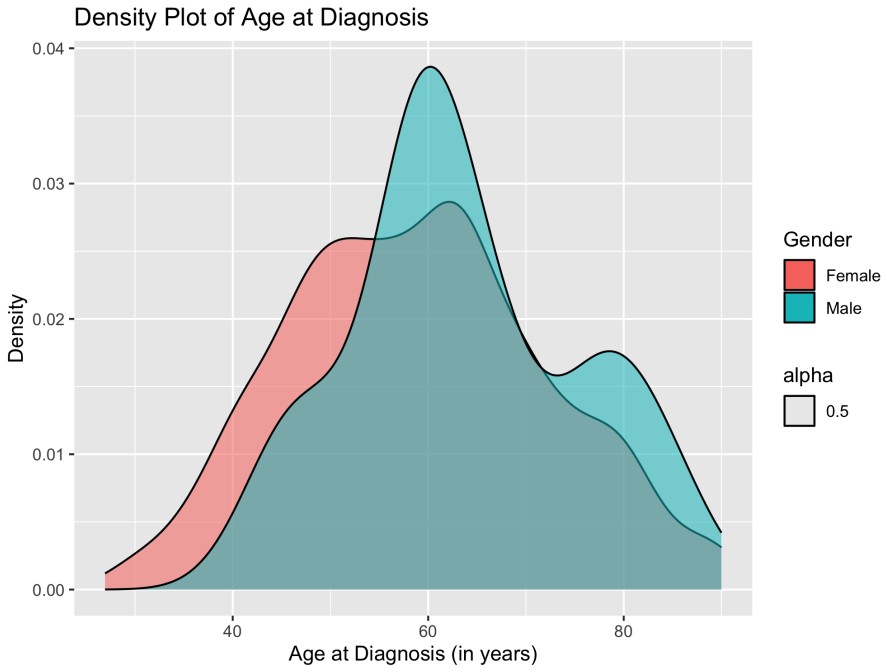

Figure 5: Density of Age at Diagnosis (in years) by Gender

4 References

4.1 Ensembl Gene Info

Specific gene information for ENSG00000007202.15 (BLTP2) and ENSG00000001631.16 (KRIT1) came from useast.ensembl.org.

4.2 How Gene BLTP2 relates to breast cancer

Banerjee, S., Daetwyler, S., Bai, X. et al. The Vps13-like protein BLTP2 regulates phosphatidylethanolamine levels to maintain plasma membrane fluidity and breast cancer aggressiveness. Nat Cell Biol 27, 1125–1135 (2025). https://doi.org/10.1038/s41556-025-01672-3

4.3 How Gene KRIT1 relates to breast cancer

Orso F, Balzac F, Marino M, Lembo A, Retta SF, Taverna D. miR-21 coordinates tumor growth and modulates KRIT1 levels. Biochem Biophys Res Commun. 2013 Aug 16;438(1):90-6. doi: 10.1016/j.bbrc.2013.07.031. Epub 2013 Jul 18. PMID: 23872064; PMCID: PMC3750217.

4.4 Raw Data

Raw data from The Cancer Genome Atlas (TCGA) Breast Invasive Carcinoma (TCGA-BRCA) cohort. (count.csv and metadata.csv).

https://portal.gdc.cancer.gov/projects/TCGA-BRCA (dbGaP Study Accession phs000178)

Heath, A.P., Ferretti, V., Agrawal, S. et al. The NCI Genomic Data Commons. Nat Genet 53, 257-262 (2021). https://doi.org/10.1038/s41588-021-00791-5 Study Attribution: The Cancer Genome Atlas Research Network. Acknowlegment Statement from the dbGaP Study Accession phs000178.v11.p8 website: "The results published here are in whole or part based upon data generated by The Cancer Genome Atlas managed by the NCI and NHGRI. Information about TCGA can be found at http://cancergenome.nih.gov."