HSE 711: Foundations in Data Science
Final Project - Fall 2025

Pritish Ponaka

Table of Contents

List of Figures

1 Introduction

This project utilizes The Cancer Genome Atlas Program (TCGA) Breast Invasive Carcinoma (TCGA-BRCA) cohort data. This dataset comprises samples collected from over 11,000 patients across 33 cancer types, including bulk RNA-seq, DNA methylation, copy number variation, somatic mutations, and clinical annotations across 9 disease types including Adenomas and Adenocarcinomas, Adnexal and Skin Appendage Neoplasms, Basal Cell Neoplasms, Complex Epithelial Neoplasms, Cystic, Mucinous and Serous Neoplasms, Ductal and Lobular Neoplasms, Epithelial Neoplasms NOS, Fibroepithelial Neoplasms, and Squamous Cell Neoplasms.

I chose ENSG00000153002.12 (CPB1, carboxypeptidase B1 gene) for this analysis, focusing on the disease types Fibroepithelial Neoplasms and Squamous Cell Neoplasms. This gene has the highest genomic variance in this dataset. I selected this gene because its expression levels vary significantly across samples due to this high variance. It exhibits rare but dramatic activation when expressed. Most samples have an expression count of zero, but when present, the expression count reaches close to seven million.

2 Methods

This analysis includes six different charts and two summary tables. The scatter plot includes another gene, ENSG00000198804.2 (MT-CO1, mitochondrially encoded cytochrome c oxidase I). The heatmap includes the ten genes with the highest genomic variance.

All plots were developed using the ggplot2 package (version 4.0.0) along with ComplexHeatmap (version 2.22.0) for the heatmap. Most of the data wrangling was performed using the dplyr (version 1.1.4), tidyr (version 1.1.3), stringr (version 1.5.2), and scales (version 1.4.0) libraries.

The covariates for this analysis are ethnicity and pathologic stage. Below are the summary statistics of the selected covariates. The data spread for 'Hispanic or Latino' ethnicity is extremely varied. Similarly, 'Stage IV' of the pathologic stage shows the most variance. From the summary statistics (Figure 2), we can infer that in most cases, the presence of the CPB1 gene in breast cancer predominantly affected samples in Stage 0 of their pathologic sage.

Summary Statistics of Ethnicity:

	Ethnicity	Mean	Median	Standard Deviation	Variance	Max	Min
1	Unknown	173.625	90.5	232.5879	5.409712e+04	671	3
2	hispanic or latino	203001.385	124.0	672690.6268	4.525127e+11	3523763	1
3	not hispanic or latino	97521.310	187.0	495008.2667	2.450332e+11	7032374	0
4	not reported	83204.073	403.0	487602.4897	2.377562e+11	4694659	0

Figure 1: Summary Statistics of Ethnicity: CPB1

Summary Statistics of Pathologic Stage:

	Pathologic Stage	Mean	Median	Standard Deviation	Variance	Max	Min
1	Stage 0	913.250	178.0	1579.497	2.494810e+06	3279	18
2	Stage I	65281.949	245.0	290406.148	8.433573e+10	1904421	0
3	Stage IA	231468.925	229.0	925355.087	8.562820e+11	7032374	0
4	Stage IB	1222882.167	447846.5	1722067.536	2.965517e+12	4281962	18
5	Stage II	4058.000	386.0	9538.578	9.098447e+07	25645	16
6	Stage IIA	74815.650	122.5	423695.426	1.795178e+11	4694659	0
7	Stage IIB	102329.337	195.0	490982.527	2.410638e+11	5484051	0
8	Stage IIIA	55591.703	218.0	343140.474	1.177454e+11	3066263	0
9	Stage IIIB	191808.185	469.0	485250.739	2.354683e+11	1870722	5
10	Stage IIIC	42676.677	248.5	184393.213	3.400086e+10	1071097	2
11	Stage IV	211059.950	188.5	722510.539	5.220215e+11	3235426	1
12	Stage X	3408.286	924.0	7069.720	4.998094e+07	19355	23

Figure 2: Summary Statistics of Pathologic Stage: CPB1

3 Results

3.1 Histogram: CPB1 Expression Count Distribution

The expression distribution is not normal. It exhibits a bimodal distribution. We can also infer from the histogram (Figure 3) that it is right-skewed. This suggests that the CPB1 gene is not universally expressed in breast cancer samples.

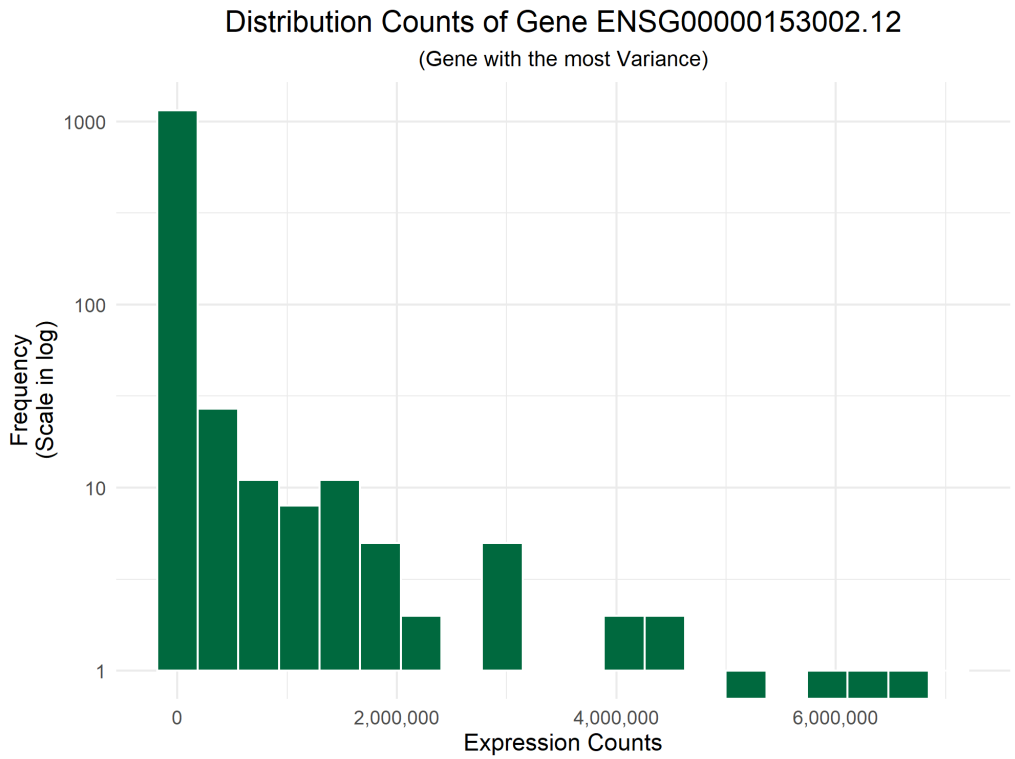

Figure 3: Histogram

3.2 Scatter Plot: CPB1 vs. MT-CO1 Expression by Gender

The scatter plot below (Figure 4) depicts the distribution of CPB1 (the gene with the most genomic variance) versus the MT-CO1 gene. The plot includes gender as a covariate variable. As we can see, for male samples, the MT-CO1 expression ranges between 300,000 and one million. The expression count for CPB1, however, is mostly zero. The regression method applied is linear. The MT-CO1 gene has the highest mean expression in the dataset.

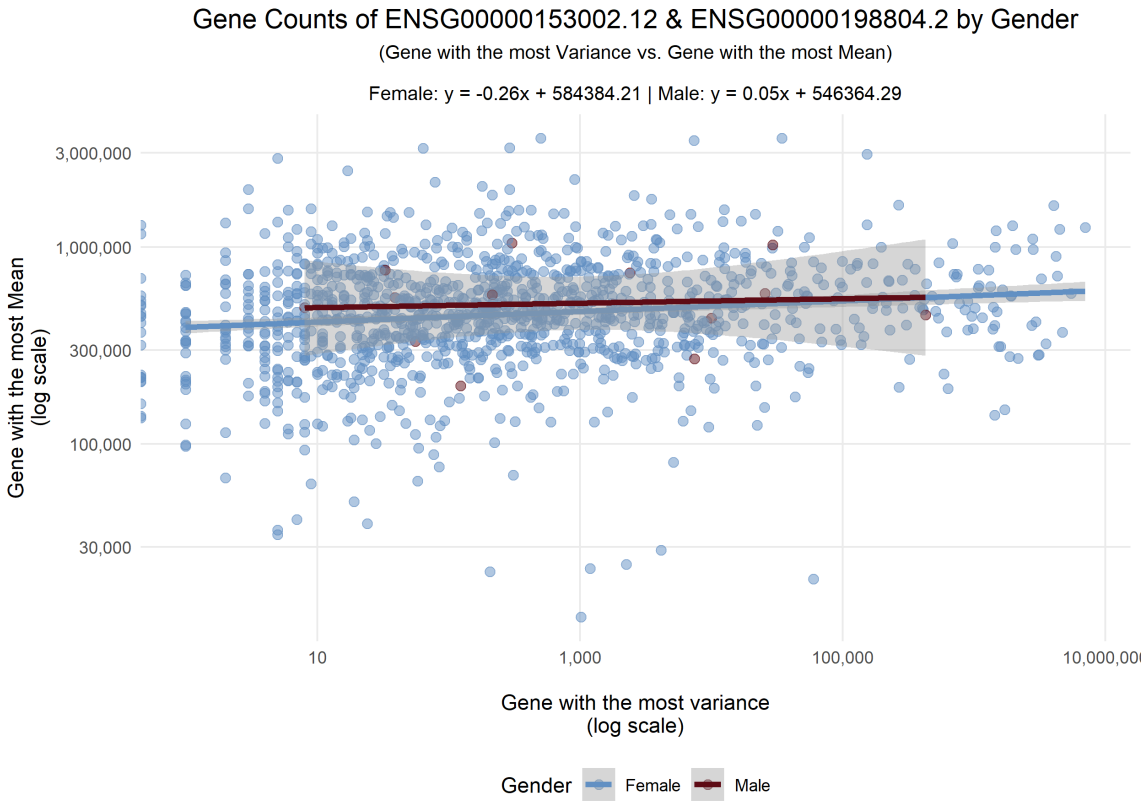

Figure 4: Scatter plot

3.3 Violin Plot: CPB1 Expression Count by Ethnicity

The plot below (Figure 5) shows the CPB1 gene expression count by ethnicity. Interestingly, despite being highly varied, the distributions among the ethnicity subsets are nearly identical, with the exception of 'Unknown'. It is also clear from the plot that when comparing 'Reported' and 'Not Reported' samples, the distributions are inverse.

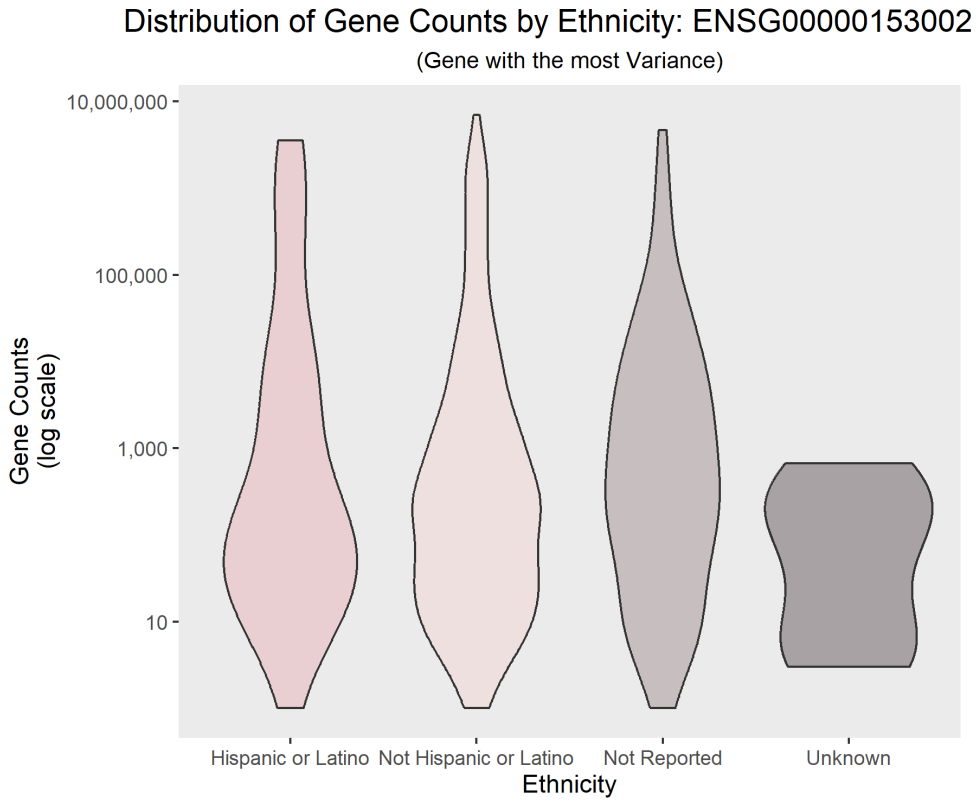

Figure 5: Violin Plot

3.4 Heatmap: Expression Count Distribution of 10 genes with most Genomic Variance

With ethnicity as a covariate, we can infer from the heatmap below (Figure 6) that the CPB1 expression count is higher in the 'Not Hispanic or Latino' subset compared to the 'Hispanic or Latino' subset. The darker the shade of amber, the higher the expression count. We can also observe the clustered genes from the dendrograms on the left side of the plot. Only the first 10 samples are considered for this analysis.

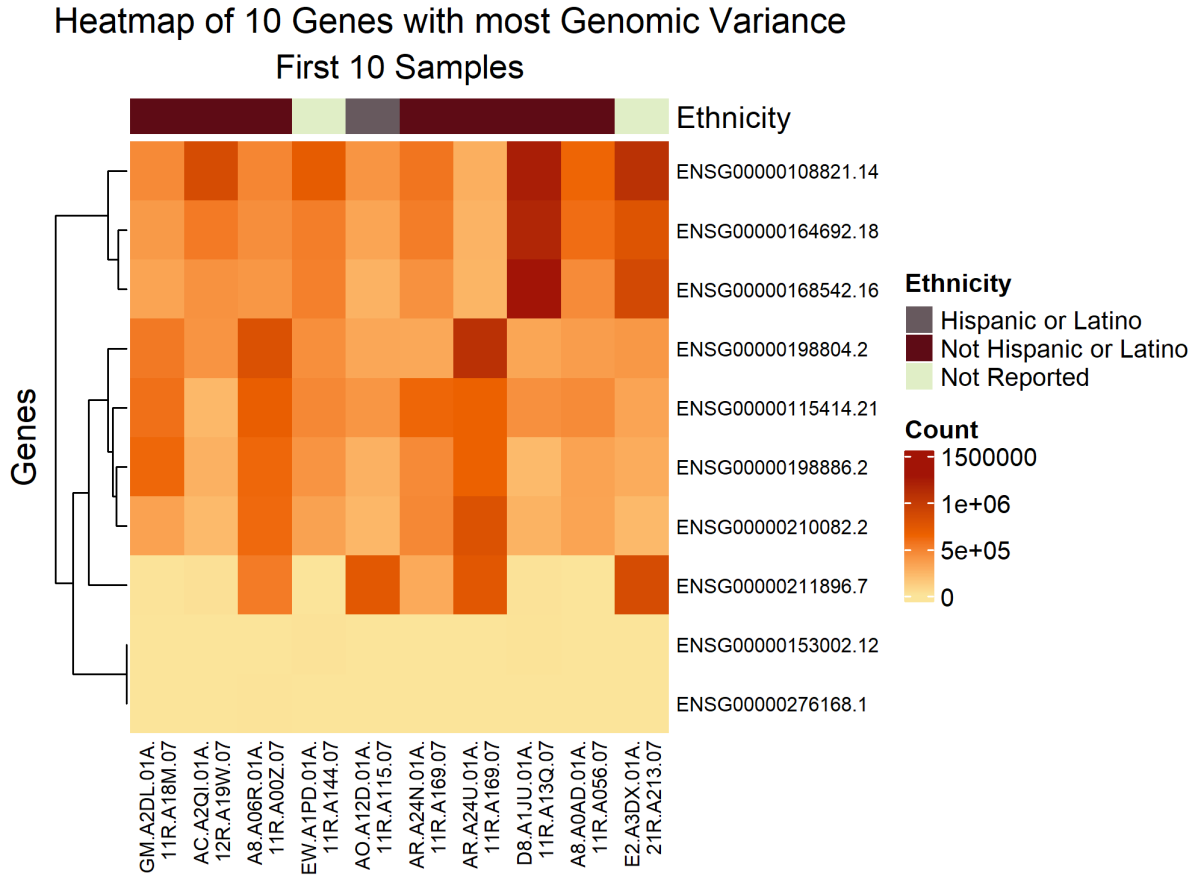

Figure 6: Heatmap (10 Samples)

3.5 Sankey: Gene Expression Across Pathologic Stage

Finally, the sankey plot depicts the 10 genes that I have selected across the pathologic stages, sorted in descending order of expression count magnitude. As we can see from Figure 7, the gene ENSG00000198804.2 has the highest expression count, leading predominantly to Stage IIA of the pathologic stage. We can also infer from the Sankey plot that the highest proportion of samples for these 10 genes are in Stage IIA. The width of the Sankey curves reflects the proportion of expression counts.

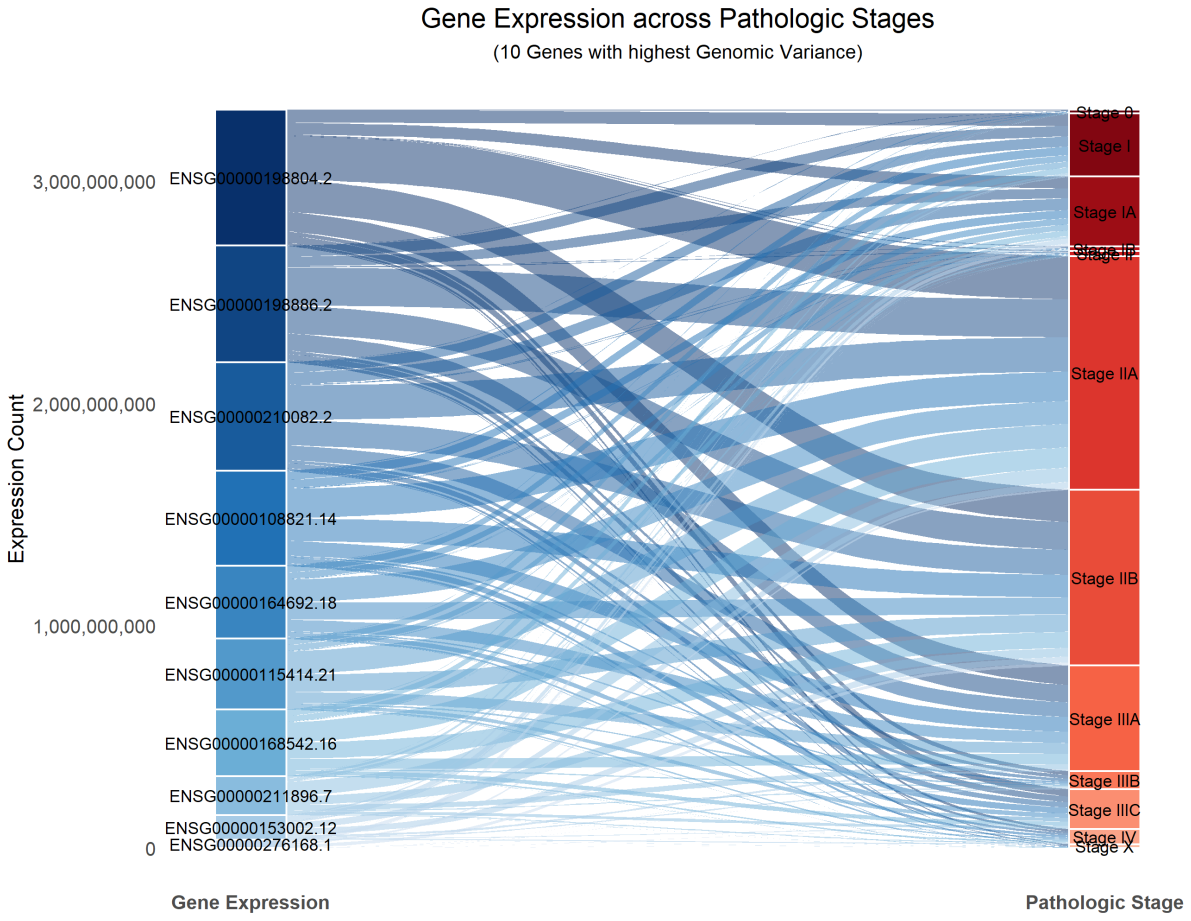

Figure 7: Sankey: Gene Expression Across Pathologic Stage

4 References

1. National Cancer Institute. (n.d.). TCGA-BRCA project. Genomic Data Commons. https://portal.gdc.cancer.gov/projects/TCGA-BRCA [portal.gdc...cancer.gov]

2. Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., ... Pontén, F. (n.d.). CPB1 - Carboxypeptidase B1 - pathology summary. The Human Protein Atlas. https://v21.proteinatlas.org/ENSG00000153002-CPB1/pathology

3. National Center for Biotechnology Information (NCBI). (n.d.). Gene: CPB1 carboxypeptidase B1 [Homo sapiens]. National Library of Medicine (US). Retrieved October 16, 2025, from https://www.ncbi.nlm.nih.gov/gene/1360

4. Holtz, Y., Healy, C. (n.d.). Sankey diagram. Data to Viz. https://www.data-to-viz.com/graph/sankey.html

5. Dartmouth College. (n.d.). Dartmouth chat service. Retrieved December 15, 2024, from https://chat.dartmouth.edu