Gene Expression and Survival Analysis in Breast Cancer (TCGA BRCA)

Jacob Wheeler

May 2025

Abstract

This report explores the relationship between gene expression and patient survival outcomes in breast cancer, utilizing data from The Cancer Genome Atlas (TCGA) Breast Invasive Carcinoma (BRCA) cohort. The analysis focuses on three key genes: *BRCA1*, *TP53*, and *ESR1*. Through statistical tests, clustering techniques, and survival analysis, we identify patterns and predictive markers associated with survival. The findings contribute to understanding molecular subtypes and may assist in prognosis and treatment stratification for breast cancer patients.

1 Introduction

Breast cancer is one of the most prevalent cancers worldwide and remains a leading cause of cancer-related mortality among women. Advances in genomics have enabled researchers to explore how gene expression impacts disease progression and patient outcomes. The Cancer Genome Atlas (TCGA) provides a rich repository of genomic data across various cancer types. The BRCA cohort includes RNA-Seq-based gene expression profiles and clinical metadata.

2 Data and Methods

2.1 Dataset

The dataset is accessible via the Genomic Data Commons (GDC) portal: `https://portal.gdc.cancer.gov/projects/TCGA-BRCA`. Key variables include expression levels of *BRCA1*, *TP53*, and *ESR1*, as well as survival time, survival status, and tumor stage.

3 Results

3.1 Descriptive Statistics

- BRCA1: 21–8050 counts.

```
> for (gene in genes_of_interest) {
+   cat("\nANOVA for", gene, "by tumor stage:\n")
+   model <- aov(merged_data[[gene]] ~ merged_data$tumor_stage)
+   print(summary(model))
+ }

ANOVA for BRCA1 by tumor stage:
                          Df    Sum Sq Mean Sq F value Pr(>F)
merged_data$tumor_stage   11 2.274e+07 2067604   1.628 0.0852 .
Residuals               1206 1.531e+09 1269780
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
14 observations deleted due to missingness

ANOVA for TP53 by tumor stage:
                          Df    Sum Sq Mean Sq F value Pr(>F)
merged_data$tumor_stage   11 3.087e+07 2806618   0.303  0.985
Residuals               1206 1.116e+10 9256113
14 observations deleted due to missingness

ANOVA for ESR1 by tumor stage:
                          Df    Sum Sq   Mean Sq F value Pr(>F)
merged_data$tumor_stage   11 3.275e+10 2.977e+09   2.046 0.0215 *
Residuals               1206 1.754e+12 1.455e+09
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
14 observations deleted due to missingness
>
```

Figure 1: ANOVA plot of ESR1 expression across tumor stages.

- TP53: 274–25,897 counts.
- ESR1: 14–277,097 counts.

3.2 T-tests and ANOVA

T-tests showed no significant differences between deceased and alive:

- BRCA1: p = 0.60
- TP53: p = 0.21
- ESR1: p = 0.18

ANOVA results:

- BRCA1: p = 0.085
- TP53: p = 0.99
- ESR1: p = 0.0215 (significant)

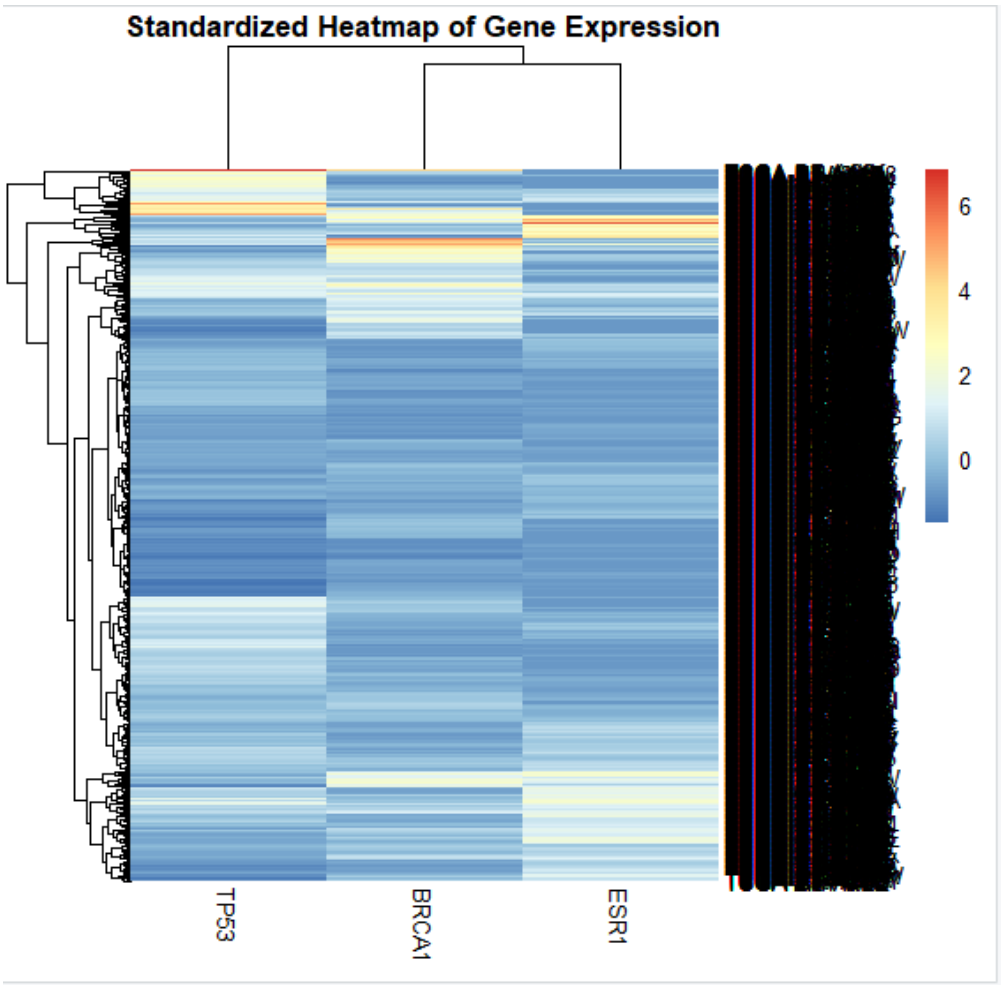

Figure 2: Heatmap of standardized gene expression clustered by K-means.

3.3 Clustering

K-means clustering (k=2) resulted in two clusters:

- Cluster 1: 864 patients
- Cluster 2: 231 patients

3.4 Survival Analysis

Kaplan-Meier survival curves showed no significant difference between clusters (p = 0.46).

3.5 Regression Modeling

Tumor stage was a significant survival predictor (p ¡ 0.001), while ESR1 was marginal (p 0.06). BRCA1 and TP53 were not significant.

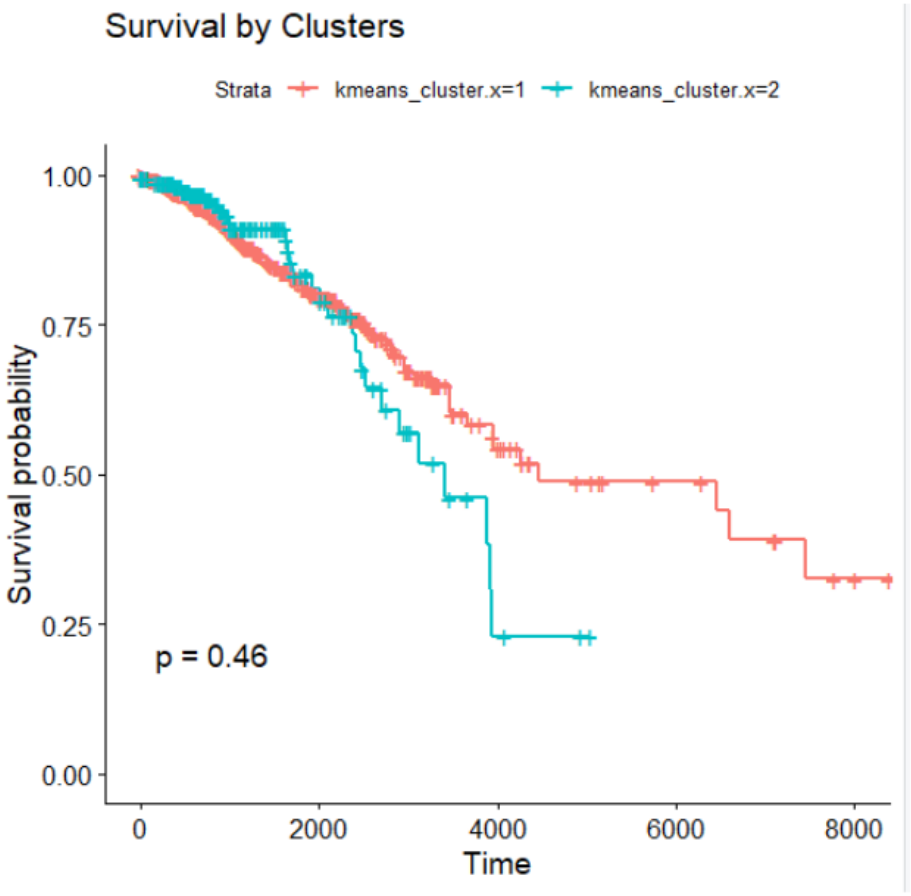

Figure 3: Kaplan-Meier survival curves for the two clusters.

4 Discussion

The results confirm tumor stage as the strongest predictor of survival. ESR1 showed stage-related variation and marginal predictive power. BRCA1 and TP53 did not show significant associations with survival outcomes. Clustering based on these genes did not yield meaningful survival stratification.

5 Conclusion

This analysis demonstrates the importance of clinical variables like tumor stage in breast cancer prognosis. ESR1 may provide additional insight, but larger gene panels and integrated data types will be key to improving predictive accuracy.