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

Jacob Wheeler

May 2025

# Outline

# Breast Cancer Overview

Breast cancer is a leading cause of cancer mortality. Advances in genomics allow detailed exploration of gene expression and survival outcomes.

# Project Objectives

- Assess gene expression and survival status.
- Examine differences across tumor stages.
- Identify molecular subgroups.
- Build predictive models of survival.

# Dataset

TCGA-BRCA dataset from the Genomic Data Commons.

- RNA-Seq gene expression data.
- Clinical metadata (tumor stage, survival time, status).

# Genes of Interest

- **BRCA1**: DNA repair gene.
- **TP53**: Tumor suppressor gene.
- **ESR1**: Estrogen receptor gene.

# Tumor Stages

Patients' tumor stages range from I to IV, included as a key clinical variable.

# T-test Results

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

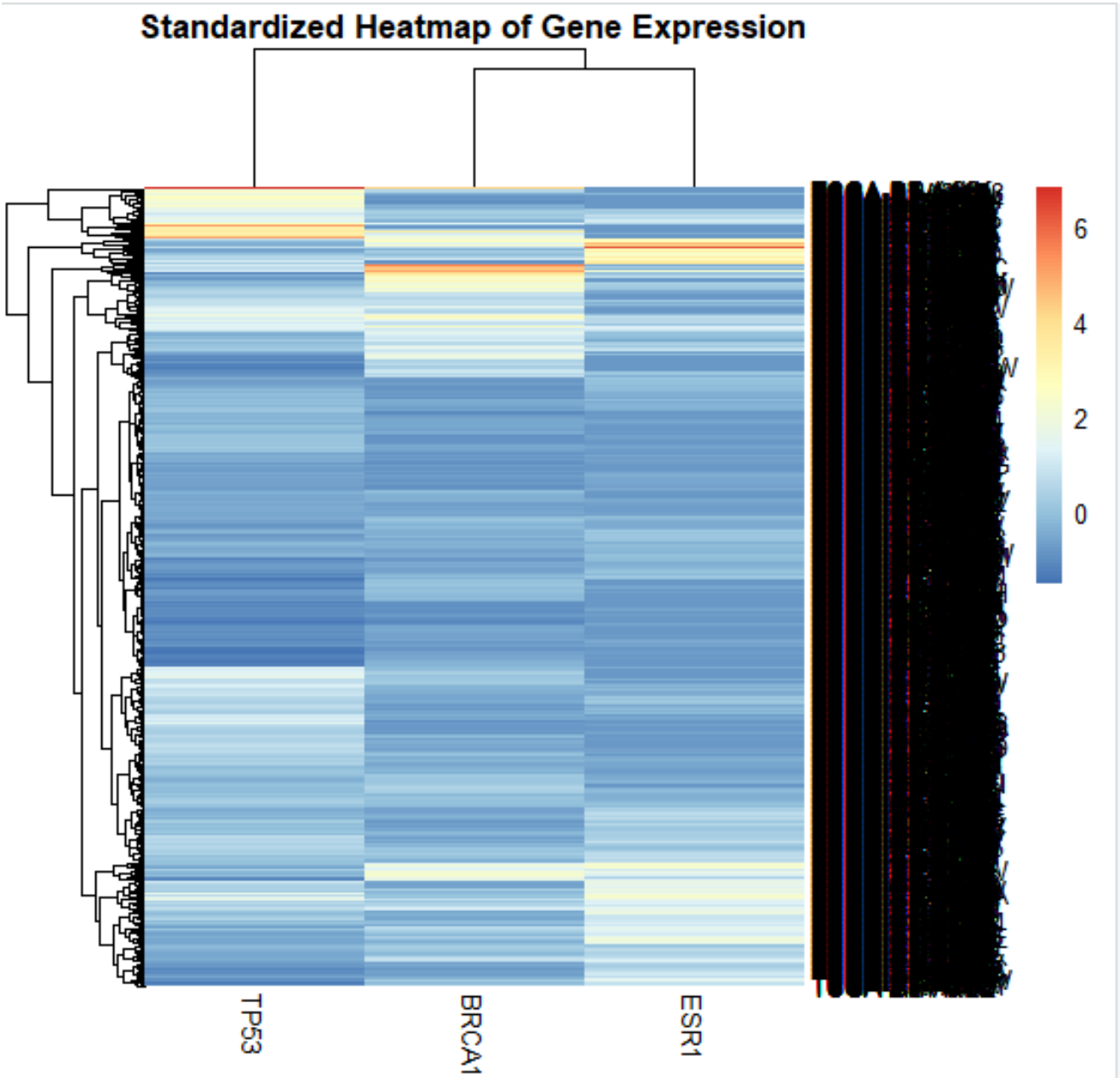
# ANOVA Results

```
> 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
>
```

# Clustering Approach

K-means clustering (k=2) on standardized gene expression.

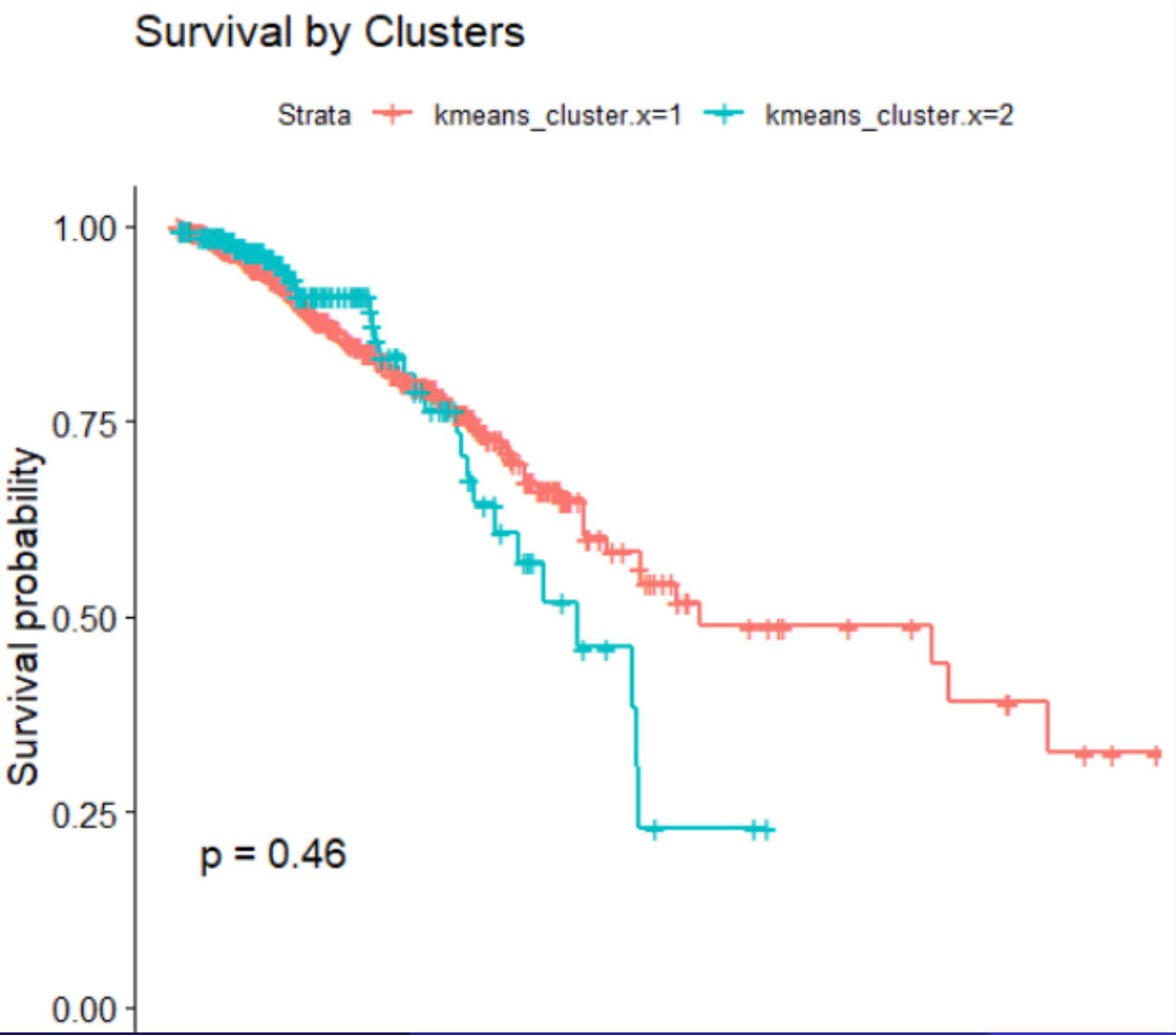
# Heatmap


Standardized Heatmap of Gene Expression
6
4
2
0
TP53
BRCA1
ESR1

# Cluster Sizes

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

# Kaplan-Meier Curves



# Survival Results

No significant survival difference between clusters ($p = 0.46$).

# Logistic Regression

Predictors:

- BRCA1
- TP53
- ESR1
- Tumor Stage

# Regression Results

- Tumor Stage IV: p ¡ 0.001 (strong predictor)
- ESR1: p $\tilde{0}$.06 (marginal)
- BRCA1 and TP53: not significant

# Key Findings

- Tumor stage is the strongest survival predictor.
- ESR1 shows stage-based variation.
- Clustering alone did not stratify survival.

# Limitations

- Small gene panel.
- Missing data reduced sample size.
- Cross-sectional dataset.

# Future Work

- Expand gene panels.
- Integrate multi-omics data.
- Apply machine learning models.

# Conclusion

Tumor stage remains a dominant factor in survival prediction. ESR1 may have added value, but broader data integration is needed for improved models.

# Acknowledgments

Thank you Dr. Kang.

# Questions

Questions?