Rayca Precision Assessment Task

Hsiang Yun Lu

July 25, 2024

1 Introduction

The objective of this analysis is to evaluate the impact of a selected gene on survival patterns across different patient clusters. Patients were clustered based on their gene expression profiles. A gene was randomly selected from the top 100 genes, followed by survival analysis within each cluster using Cox Proportional Hazards (CoxPH) models and Kaplan-Meier survival analysis.

2 Method

2.1 Data Preparation

The TCGA-BRCA dataset from cBioPortal was used in this analysis. (`https://www.cbioportal.org/study/summary?id=brca_tcga_pan_can_atlas_2018`)

The clinical data files utilized were *data_clinical_sample.txt* and *data_clinical_patient.txt*. The expression data utilized was *data_mrna_seq_v2_rsem_zscores_ref_all_samples.txt*.

- **Gene Expression Data:**

  Irrelevant data (e.g., HUGO gene symbols) was filtered out. Variance filtering was applied to enhance data quality. The variance of expression levels was calculated for each gene, and genes with variance below a defined threshold were excluded. Duplicated gene names with different z-scores for each sample were identified. To address this, columns with identical gene names were aggregated by calculating their mean z-score values.

- **Clinical Data:**

  Patient ID and sample ID were extracted from *data_clinical_sample.txt* for sample ID mapping. Survival data (*OS_STATUS* and *OS_MONTHS*) and patient ID were extracted from *data_clinical_patient.txt*. Sample IDs and survival data were then used for subsequent analysis.

2.2 Feature Selection

Feature selection was conducted using the Cox proportional hazards model. The top 100 genes were selected based on their hazard ratios.

2.3 Clustering

The K-means clustering algorithm was implemented to identify distinct clusters within the gene expression data. The silhouette score was used to determine the optimal number of clusters. This process was repeated with a random state of 42 to ensure reproducibility. Each patient was assigned to a cluster based on their gene expression profile.

A gene was randomly selected from the top 100 genes for survival analysis. The selected gene for this analysis was 414332 (Entrez_Gene_Id).

2.4 Survival Analysis

To assess the prognostic significance of the identified gene expression clusters, both the Cox Proportional Hazards Model and Kaplan-Meier Survival Analysis were employed. The analysis was conducted separately for each cluster to evaluate the impact of the selected gene's expression on overall survival.

- **Cox Proportional Hazards Model:**

  The Cox Proportional Hazards Model was fitted using data from each cluster for the selected gene. A summary of the fitted model was produced to interpret the association between gene expression and survival within each cluster.

- **Kaplan-Meier Survival Analysis:**

  Within each cluster, patients were divided into high and low expression groups based on the median expression level of the selected gene. Kaplan-Meier survival curves were plotted for both high and low expression groups. The log-rank test was performed to statistically compare the survival distributions of the high and low expression groups.

3 Results

3.1 Cox Proportional Hazards Model (CoxPH)

Model summaries for each cluster:

Cluster	coef	exp(coef)	se(coef)	z	p
0	-0.13	0.88	0.20	-0.66	0.51
1	-0.31	0.73	0.18	-1.77	0.08
2	0.28	1.33	0.26	1.09	0.27
3	-0.15	0.86	0.22	-0.68	0.50

For cluster 0, the hazard ratio (HR) < 1 and a negative coefficient suggest that higher expression of gene 414332 might be associated with a decreased risk, but the p-value > 0.05 indicates this association is not statistically significant.

For cluster 1, the HR < 1 and a negative coefficient suggest that higher expression of gene 414332 might be associated with a decreased risk. The p-value is slightly above 0.05, indicating a trend towards significance but not statistically significant.

For cluster 2, the HR > 1 and a positive coefficient suggest that higher expression of gene 414332 might be associated with an increased risk, but the p-value > 0.05 indicates this association is not statistically significant.

For cluster 3, the HR < 1 and a negative coefficient suggest that higher expression of gene 414332 might be associated with a decreased risk, but the p-value > 0.05 indicates this association is not statistically significant.

- **Hazard Ratios (HR):**

  In most clusters, the HR for the selected gene were less than 1, with the exception of Cluster 2, where HR > 1. This suggests that higher expression of the selected gene might be associated with a decreased risk in Clusters 0, 1, and 3, and an increased risk in Cluster 2. However, these trends are not statistically significant due to the p-values being greater than 0.05.

- **P-values:**

  The p-values from the CoxPH models in all clusters are greater than 0.05, indicating that the associations between the selected gene expression and survival are not statistically significant. This means there is no strong evidence to suggest that the selected gene significantly impacts survival within any of the clusters.

- **Standard Error (se(coef)):**

  The standard error of the coefficients provides an estimate of the variability or precision of the coefficient estimates. Larger standard errors indicate less precise estimates. In this analysis, the standard errors are relatively small, suggesting reasonable precision in the coefficient estimates.

The consistent finding across all clusters, where the HR < 1 in most clusters and p-values > 0.05, suggests a potential trend of higher gene expression being associated with decreased risk in some clusters and increased risk in one cluster. However, due to the lack of statistical significance, we cannot confidently conclude that the selected gene affects survival.

3.2 Kaplan-Meier Survival Analysis

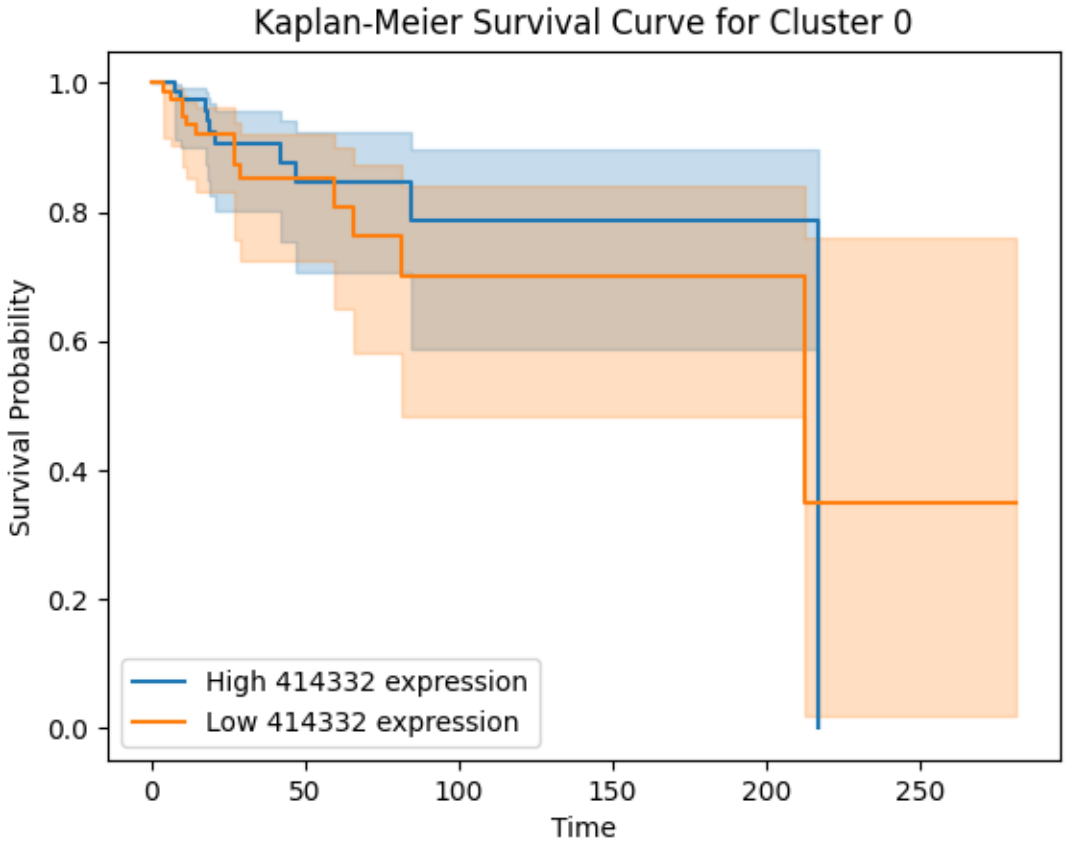
Log-rank test p-value for each cluster:

Cluster	p-value
0	0.4700351667634415
1	0.12049274230290255
2	0.7241615151647266
3	0.32727331457576486

The p-values obtained from the log-rank tests are all greater than 0.05, indicating that there are no statistically significant differences in survival distributions between the high and low expression groups across all clusters.

- **Kaplan-Meier Plot: Cluster 0**

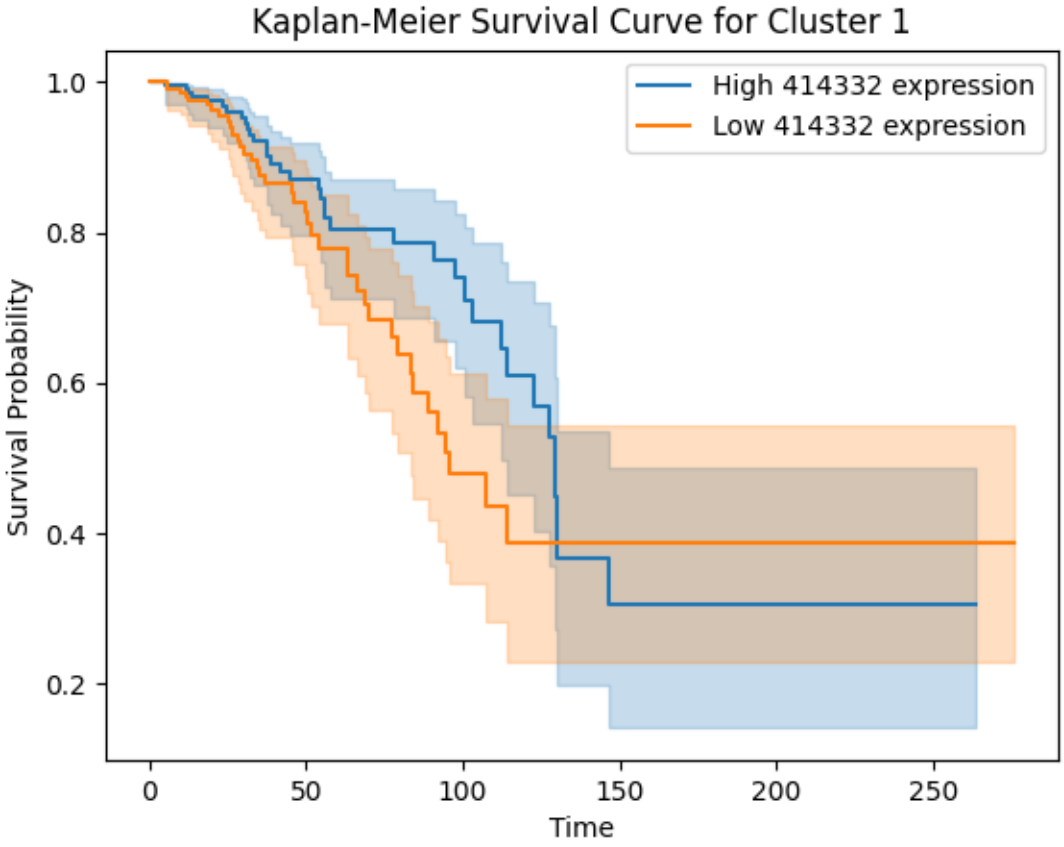
  Initially, both high and low expression groups exhibit similar survival probabilities. Over time, the high expression group (blue line) maintains a slightly better survival probability compared to the low expression group (orange line). The divergence between the survival curves becomes more pronounced after approximately 50 months, with the high expression group showing a higher survival probability. Around the 200-month mark, the high expression group experiences a noticeable drop in survival probability, while the low expression group continues to decline more gradually. Initially, both groups have overlapping confidence intervals, indicating no significant difference in survival probability at the outset. As time progresses, the confidence intervals widen, reflecting increased uncertainty in the survival estimates. Despite this, the overall trend suggests better survival for the high expression group.



- **Kaplan-Meier Plot: Cluster 1**

  Patients with high expression of the selected gene exhibit a gradual decline in survival probability over time. This decline becomes more pronounced between the 50-month

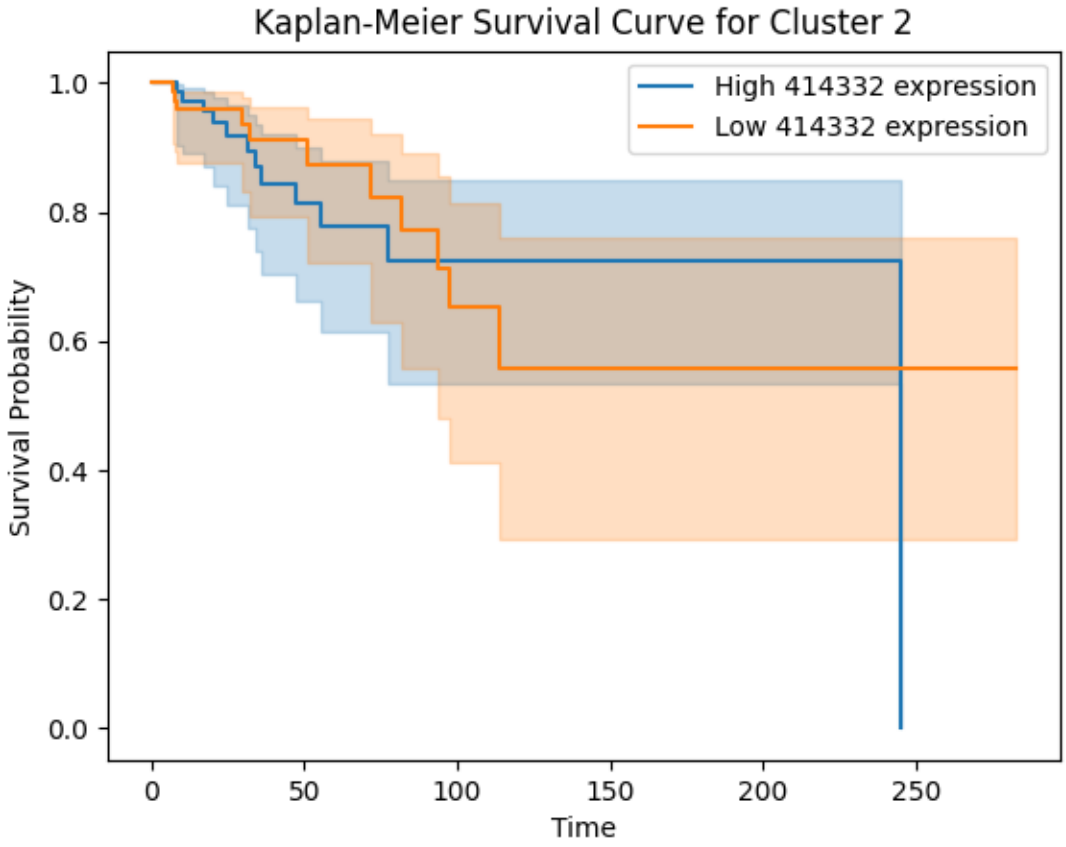
and 150-month marks, after which the rate of decline slows. In contrast, patients with low expression also experience a gradual decline in survival probability, but this decline is steeper compared to the high expression group. The disparity in survival probability between the high and low expression groups becomes more evident around the 100-month mark, suggesting a higher event (relapse) rate among the low expression group during this period. However, the confidence intervals overlap significantly between the two groups, indicating that while a trend is observed, the difference may not be statistically significant.



- **Kaplan-Meier Plot: Cluster 2**

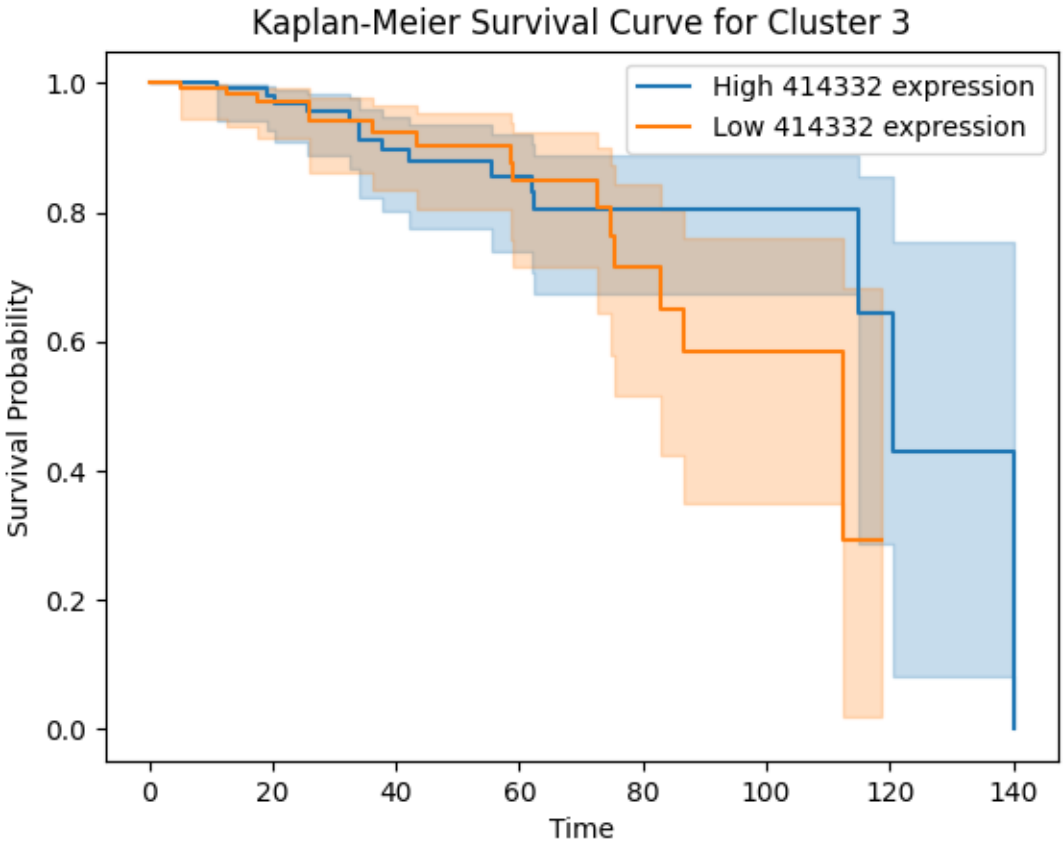
  Both the high and low expression groups exhibit a relatively steep survival curve until the 100-month mark, where the high expression group shows a stronger association with decreased survival. There is a crossover between the two curves, where the high expression group then shows a stronger association with increased survival compared to the low expression group. A noticeable decline occurs at around the 250-month mark. Between the 100 and 250-month marks, the high expression group maintains a higher survival probability of 0.75 compared to the low expression group's 0.55, indicating a higher event (relapse) rate in the low expression group during this period. However, the confidence intervals for both groups are wide and overlap significantly, suggesting substantial uncertainty in the survival estimates and indicating that the

observed differences between the groups may not be statistically significant.



- **Kaplan-Meier Plot: Cluster 3**

  Similar to Cluster 0, both high and low expression groups in this cluster start with similar survival probabilities. Over time, the high expression group shows a consistently higher survival probability compared to the low expression group. The divergence between the survival curves is more pronounced than in Cluster 0, indicating a stronger association between high gene expression and better survival in this cluster. There is a steady decline in survival probability for both groups, but the high expression group maintains a higher probability throughout the observed period. The confidence intervals for the high expression group remain relatively narrow, suggesting more reliable survival estimates. In contrast, the low expression group's confidence intervals widen significantly after the 80-month mark, indicating increased uncertainty and lower survival probability.



The survival curves of Clusters 0 and 3 exhibit a similar pattern, starting less steeply and overlapping. In Cluster 3, the high expression group shows a significantly better survival probability compared to the low expression group. The Kaplan-Meier survival curves suggest that higher expression of the selected gene is associated with better survival outcomes in both clusters, with a stronger effect observed in Cluster 3.

Cluster 2's survival curves resemble those of Clusters 0 and 3, but with notable variations: before the 100-month mark, the high expression group shows a stronger association with decreased survival, whereas after the 100-month mark, an opposite pattern is observed.

Cluster 1 displays a more pronounced decline in survival curves compared to all other clusters, indicating that the impact of the selected gene's expression on survival may vary between clusters.

Overall, there is a general trend across all clusters where patients with high expression of the selected gene tend to have slightly lower survival probabilities compared to those with low expression. However, these differences are subtle and may not be statistically significant due to the overlapping confidence intervals.

4 Discussion

4.1 Statistical Significance

The CoxPH models show that the hazard ratios (HRs) for the selected gene across all clusters are not statistically significant (p-values > 0.05). This suggests that, based on the available data, there is no strong evidence to conclude that the selected gene's expression significantly impacts survival in any of the clusters. The Kaplan-Meier survival analysis also indicates no statistically significant differences in survival distributions between high and low expression groups within each cluster (log-rank test p-values > 0.05). Despite the lack of statistical significance, the observed trends in HRs and survival probabilities provide insights into potential patterns and biological relevance that might warrant further investigation with larger sample sizes or additional data.

4.2 Cluster-Specific Interpretation

- **Cluster 0:**

  The Cox Proportional Hazards (CoxPH) model indicates a hazard ratio (HR) of less than 1, suggesting a potential association between higher gene expression and decreased risk of events, though this finding is not statistically significant. The Kaplan-Meier plot supports this trend, showing that the high expression group exhibits better survival probability over time, with a noticeable divergence occurring after approximately the 50-month mark.

- **Cluster 1:**

  The Cox Proportional Hazards (CoxPH) model indicates a hazard ratio (HR) of less than 1, suggesting a trend towards decreased risk with higher gene expression, although this finding is not statistically significant. The Kaplan-Meier plot aligns with this trend, showing that the high expression group experiences a slower decline in survival probability compared to the low expression group, with more pronounced differences around 100 months.

- **Cluster 2:**

  The Cox Proportional Hazards (CoxPH) model shows a hazard ratio (HR) greater than 1, indicating a potential association between higher gene expression and an increased risk of events, though this is not statistically significant. The Kaplan-Meier plot reveals that both high and low expression groups experience relatively steep initial declines, with the high expression group exhibiting better survival probability only between 100 and 250 months. The inconsistency in the survival patterns over the entire duration suggests that the increased risk indicated by the CoxPH model may only be relevant for certain time periods, and the wide confidence intervals reflect substantial uncertainty.

- **Cluster 3:**

  The Cox Proportional Hazards (CoxPH) model shows a hazard ratio (HR) less than 1, suggesting a potential decreased risk with higher gene expression, although this finding

is not statistically significant. The Kaplan-Meier plot indicates that the high expression group consistently shows better survival probabilities over time, aligning well with the trend observed in the CoxPH model. Additionally, the narrow confidence intervals for the high expression group indicate more reliable survival estimates compared to the low expression group.

While the results are not statistically significant, the consistent trends across multiple clusters (especially in Clusters 0, 1, and 3) suggest that higher expression of the selected gene might be associated with better survival outcomes. This could point towards a potential biological role of the gene in influencing survival, warranting further investigation.

The discrepancy in Cluster 2, where the CoxPH model indicates increased risk but the Kaplan-Meier plot shows improved survival for a certain period, suggests that the relationship between gene expression and survival might be more complex and time-dependent. This could be due to varying biological mechanisms at different stages of the disease or treatment responses.

The lack of statistical significance might be due to insufficient sample size or power. Future studies with larger cohorts or more refined clustering methods might provide clearer insights and potentially reveal significant associations.

The standard errors in the CoxPH models are relatively small, indicating reasonable precision in the coefficient estimates. However, the wide confidence intervals in the Kaplan-Meier plots for some clusters highlight the uncertainty in survival estimates, especially for the low expression groups. This underscores the need for cautious interpretation of the trends.

5 Conclusion

The combined analysis of CoxPH models and Kaplan-Meier survival curves suggests potential associations between higher gene expression and better survival outcomes in several clusters, though these associations are not statistically significant. The trends observed provide a foundation for future research, emphasizing the need for larger studies to confirm these findings and explore the biological mechanisms underlying these patterns.