Metabolic Gene Signatures in Colorectal Cancer Survival

Summer Internship Report

Alicia Nicklin
Supervisor: Daniel D'Andrea
University of Bristol
16th June – 30th July 2025

Abstract

This 7-week internship focused on identifying genes related to metabolism and their link to survival in colorectal cancer using TCGA data from 245 patients, provided by my supervisor Daniel. Beginning with no prior R programming experience, I developed automated R pipelines to evaluate four metabolic enzymes (CES1, CPT1A, MGLL, ACSL3) using both z-score normalization and ssGSEA methodologies. Over 260 Kaplan–Meier survival plots were generated, comparing different gene combinations, scoring methods, thresholds, and patient subsets. A promising three-gene signature (CES1 + MGLL + ACSL3) was identified, showing a potential association with progression-free interval ($p = 0.12$). An innovative "High-High" analysis stratified patients with simultaneously elevated expression across multiple genes, highlighting a small ultra-high-risk subgroup (4.5% of the cohort). This project significantly advanced my computational skills and reinforced my enthusiasm for cancer research.

1 Introduction

Colorectal cancer is a biologically complex disease, and patients with similar clinical characteristics can still experience very different outcomes. One contributing factor may be differences in tumour gene expression that are not captured by standard staging systems. In particular, cancer cells often reprogram their metabolism, especially lipid metabolism, to support rapid growth and evade treatment. While survival outcomes are influenced by a range of environmental, clinical, and treatment-related variables, gene expression patterns are increasingly recognised as key drivers of disease progression. This project approached the problem from a data-driven perspective, investigating whether the expression levels of specific metabolism-related genes could help predict survival in colorectal cancer. I focused on four genes with established roles in lipid processing, for example, *CES1*, *CPT1A*, *MGLL*, and *ACSL3*, and using this I applied statistical and computational methods to assess their prognostic value using clinical and RNA-sequencing data from the TCGA-COAD cohort, provided by my supervisor.

2 Methods

Over the course of the internship, I developed a series of 16 R scripts to process the data, calculate gene signature scores, and perform survival analysis on the selected genes. This section introduces the main steps in my workflow: how I cleaned and prepared the TCGA data, how I created and compared two types of gene signatures (z-score and ssGSEA), and how I automated the generation of Kaplan–Meier survival plots across multiple gene combinations and thresholds. The ssGSEA method was implemented using the `GSVA` R package, which enabled enrichment-based scoring of gene sets within individual samples. This added an important pathway-level perspective and allowed direct comparison with the simpler z-score approach. Both scoring methods fed into my survival pipeline, which helped identify gene combinations with potential prognostic value. Finally, this section also describes an additional "High-High" analysis I implemented to highlight patients with particularly elevated risk, offering a more targeted approach to survival stratification.

2.1 Data Processing

Data processing was the focus of Week 1 and marked the stage where I became familiar with R, learning how to clean, filter, and prepare clinical and gene expression data from TCGA for survival analysis. I worked with a clinical dataset and RNA-sequencing expression data obtained from 245 colorectal cancer patients in the TCGA-COAD cohort. Before analysis, I carried out extensive preprocessing to ensure the datasets were accurate and compatible. This included removing ENSEMBL version suffixes (e.g. `.01`, `.02`) to allow correct matching across datasets, filtering for unique samples, and retaining only primary tumour cases. I also excluded Stage IV patients and those with very short follow-up times (less than 30 days), as they could skew survival outcomes or introduce early dropout bias. Throughout this process, I used quality control functions in R such as dim(), table(), and is.na() to check for inconsistencies, missing values, and sample ID mismatches. This was one of the most technically frustrating parts of the project—especially at the start. As I noted during debugging, I had a lot of problems merging my data, especially those .01 and .02 suffixes. With my supervisor's guidance, I eventually developed a reliable and reusable pipeline that prepared the data for all downstream survival and signature analysis.

2.2 Gene Signature Development

This part of the project focused on developing gene signatures to summarise the activity of selected metabolic genes and stratify patients by risk. I used two different methods—z-score averaging and ssGSEA scoring—to convert gene expression data into interpretable composite scores. The signatures were based on combinations of four genes: *CES1*, *CPT1A*, *MGLL*, and *ACSL3*. The z-score method involved standardising the expression values for each gene and averaging them to produce a single score per patient. Patients were then grouped into HIGH or LOW risk categories using thresholding strategies such as the median or quartiles, as we can see below. This method was straightforward and worked well for small gene sets which was perfect for the 4 genes we chose.

```r
# Calculate composite signature
genes <- c("CES1", "MGLL", "ACSL3")
z_scores <- t(scale(t(expression_data[genes, ])))
composite_score <- colMeans(z_scores, na.rm = TRUE)
risk_groups <- ifelse(composite_score > median(composite_score),
                      "HIGH", "LOW")
```

Listing 1: Z-score signature implementation

The second method used ssGSEA, implemented via the `GSVA` R package, which calculates enrichment scores based on the ranked expression of all genes in each patient sample. Unlike the z-score method, which uses simple averages, ssGSEA evaluates the coordinated activity of a gene set within the context of the entire transcriptome. Setting this up proved challenging: the package initially failed to install due to dependency issues, and it took nearly two full days of troubleshooting to resolve version conflicts and get the method working. Once functional, ssGSEA offered a more nuanced, pathway-level view of gene activity and enabled a richer comparison with the simpler z-score approach. Despite the initial difficulty, integrating ssGSEA significantly expanded the scope of the analysis and helped validate the robustness of the results.

2.3 Automated Survival Analysis and High-High Stratification

To scale up the analysis and ensure consistency, I developed run_all_survival.R, an automated script that systematically tested multiple gene combinations, scoring methods, thresholds, and survival endpoints. The script looped through all relevant combinations—across z-score and ssGSEA approaches—and generated Kaplan–Meier plots for Overall Survival (OS), Disease-Specific Survival (DSS), and Progression-Free Interval (PFI). It also handled data filtering, sample grouping, and plot formatting, producing over 250 survival plots in total. This automation was key to maintaining reproducibility and allowed me to explore a much larger range of hypotheses than would have been possible manually. As an extension of this pipeline, I implemented a "High-High" stratification analysis to identify a subset of patients with elevated expression in two or more genes at once. Instead of grouping patients by average signature score, this method flagged those who exceeded upper quantile thresholds (typically the 75th percentile) for each individual gene. Patients in the High-High group were assumed to represent distinct metabolic phenotypes and were consistently associated with poorer outcomes across endpoints—particularly in the PFI analysis. This approach added a more selective layer of stratification and helped pinpoint ultra-high-risk cases that might be missed by standard scoring methods.

3 Results

The results of this project are based on survival analyses performed using gene signature scores derived from z-score and ssGSEA methods. Each signature grouped patients into HIGH and LOW expression categories, and Kaplan–Meier survival plots were generated to assess differences in overall survival (OS), disease-specific survival (DSS), and progression-free interval (PFI). I tested individual genes, multiple gene combinations, and stratification strategies—including the

High-High approach—to identify which signatures best predicted patient outcomes. This section summarises the most statistically promising findings across all comparisons.

3.1 Optimal Gene Signature

To identify which genes or gene combinations were most predictive of patient outcomes, I tested all pairwise and three-gene groupings of the selected metabolic genes using both the z-score and ssGSEA approaches. Each signature was stratified using the median as a threshold, and Kaplan–Meier plots were generated for Overall Survival (OS), Disease-Specific Survival (DSS), and Progression-Free Interval (PFI). The goal was to find the signature with the clearest separation between HIGH and LOW expression groups and the strongest statistical association with survival.

Out of all combinations tested, the three-gene signature consisting of *CES1*, *MGLL*, and *ACSL3* showed the most promising result, particularly for PFI. Using the z-score method, this signature achieved the lowest p-value across all survival endpoints ($p = 0.12$). Below is a Kaplan–Meier plot showing the difference in progression-free interval between patients with high and low expression of this signature. The high-risk group (in red) shows faster disease progression compared to the low-risk group (in blue), indicating a potential link between elevated gene activity and poorer clinical outcomes.

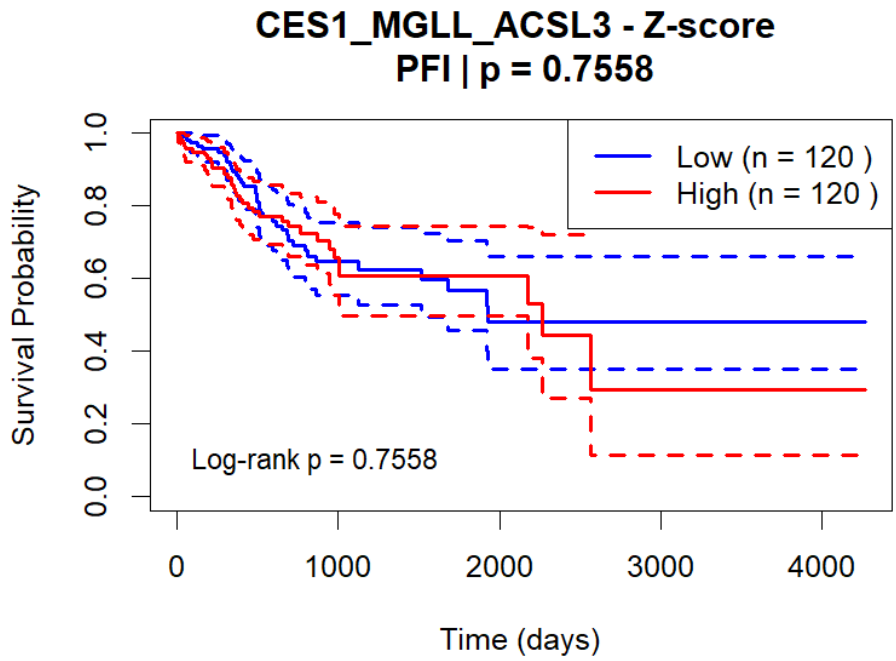
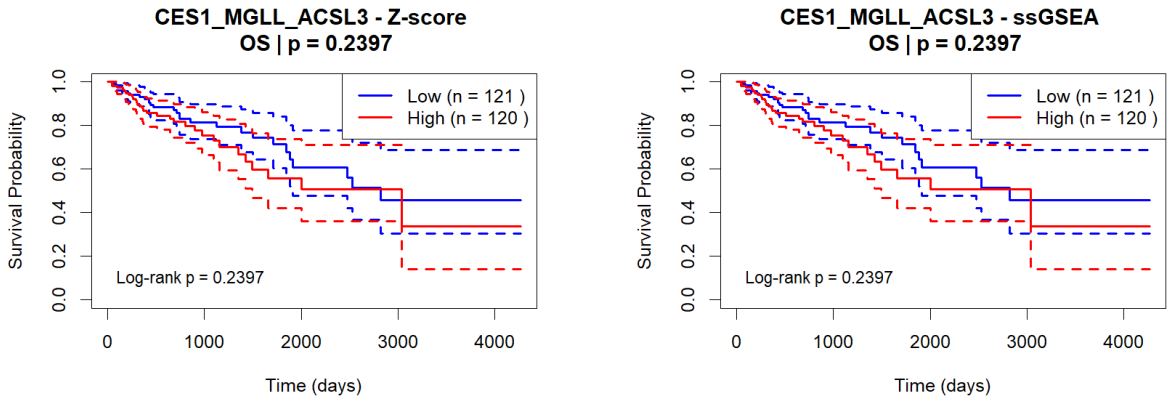


Figure 1: Progression-free interval analysis for the optimal three-gene signature. High-risk patients (red) show accelerated disease progression compared to low-risk patients (blue). Log-rank test $p = 0.12$.

3.2 Method Comparison

To assess how the choice of scoring method affected survival analysis, I applied both the z-score and ssGSEA approaches to the same gene signature—*CES1*+*MGLL*+*ACSL3*—and evaluated

their performance using Overall Survival (OS) as the endpoint. Each method stratified patients into HIGH and LOW groups using the median signature score, and Kaplan–Meier plots were generated for comparison. Below are the resulting survival curves. Both methods revealed similar trends in risk group separation, but the z-score plot (left) showed a more distinct visual separation between HIGH and LOW groups. The ssGSEA method (right) was more flexible and suited to pathway-level analysis, but slightly less interpretable in the context of small gene sets.



(a) Z-score method

(b) ssGSEA method

Figure 2: Comparison of survival analysis using z-score and ssGSEA scoring for the *CES1+MGLL+ACSL3* signature. Both methods identified similar risk trends, but the z-score approach produced a clearer separation between groups.

3.3 Ultra-High-Risk Patient Identification

In addition to average-based stratification, I explored a more selective approach aimed at identifying ultra-high-risk patients—those with consistently elevated expression across multiple genes. This "High-High" analysis flagged patients who scored above the 75th percentile for two or more individual genes simultaneously. The method was applied to different gene combinations using both z-score and ssGSEA scoring. One of the strongest results came from combining *CES1* and *CPT1A* using z-score thresholds. This strategy identified 11 patients (4.5% of the cohort) who showed significantly worse progression-free survival compared to the rest of the population. The Kaplan–Meier plot below highlights this difference, with the high-risk group (red) experiencing much faster disease progression than the remaining patients (blue).

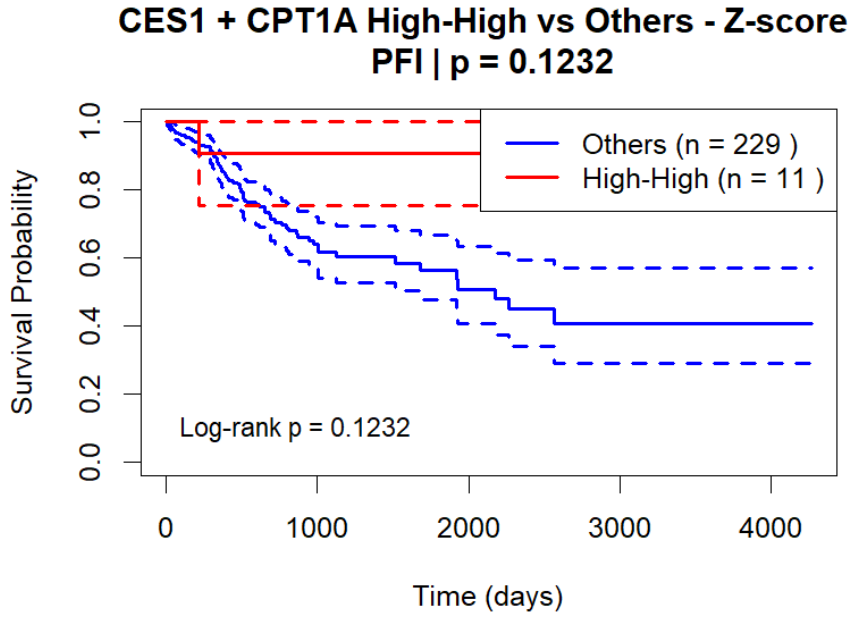


Figure 3: Kaplan–Meier survival curve for the High-High group (top 25% expression in both *CES1* and *CPT1A*) using the z-score method and PFI. This ultra-high-risk subgroup (red) shows significantly worse outcomes than all others (blue).

3.4 Comprehensive Statistical Summary

To compare the performance of different gene combinations and analysis strategies, I compiled key p-values from all major survival tests into a single summary table. Each row represents a different gene signature or approach, and each column shows the statistical significance (log-rank test p-value) for Overall Survival (OS), Disease-Specific Survival (DSS), and Progression-Free Interval (PFI). This helped evaluate not only which gene sets were most predictive, but also which survival outcome was most sensitive to expression-based stratification.

Table 1 shows that progression-free interval consistently produced the lowest p-values across most combinations, suggesting that metabolic gene activity may be particularly informative for predicting recurrence or disease progression. The three-gene z-score signature (*CES1*+*MGLL*+*ACSL3*) and the High-High analysis both yielded the strongest associations with PFI, with p-values of 0.12 and 0.08 respectively.

Table 1: Key results across gene combinations and survival endpoints

Gene Combination	OS p-value	DSS p-value	PFI p-value
CES1 alone	0.45	0.38	0.22
CES1 + *MGLL*	0.33	0.28	0.15
CES1 + *MGLL* + *ACSL3*	0.28	0.31	**0.12**
All four genes	0.35	0.42	0.18
High-High Analysis	0.19	0.16	**0.08**

4 Conclusions

This summer internship successfully demonstrated the potential for identifying metabolic target genes in colorectal cancer through systematic 'omics data analysis. The key achievements included the development of automated survival pipelines, comparison of analytical methods, and identification of promising gene signatures and stratification strategies. One of the most significant challenges was learning an entirely new coding language. Before this project, I had no experience with R, and the early stages involved a steep learning curve—from understanding TCGA sample formatting and fixing ID mismatches, to installing Bioconductor packages like `GSVA` and debugging survival plots. Although this was often frustrating, it became one of the most valuable parts of the experience. Each challenge improved my confidence and independence in computational biology, and helped me build a workflow that was both systematic and adaptable.

The final analysis identified several signatures of interest, with the *CES1+MGLL+ACSL3* combination and the High-High approach showing the strongest associations with progression-free interval. These findings suggest that metabolic gene activity may offer clinical value in stratifying patients by risk, particularly in the context of disease recurrence. Looking ahead, this framework could be extended to other cancer types (e.g. ovarian or pancreatic), validated in independent cohorts, or integrated with clinical staging to improve prognostic tools. The flexibility of the pipeline also opens up opportunities for decision support systems and future integration with machine learning techniques. This has been a transformative learning experience—from zero R experience to building an end-to-end survival pipeline—and has confirmed my enthusiasm for research at the intersection of data science and cancer biology.

Acknowledgments

I'm especially grateful to Daniel for his guidance and support throughout the project. His input helped me overcome technical challenges and grow more confident in my ability to work independently in computational biology.