Hallmark Gene Expression Patterns as Predictive Biomarkers in Pan-Cancer Survival

David Hu, Ryan Joyce, Anurag Kumar, Hanna Warren

Carnegie Mellon University

Abstract

"Cancer hallmark genes" are associated with important steps that lead to the spread and development of cancer. In this project, we repeated an earlier study to see how these hallmark genes are linked to patient survival in different types of cancer [7]. We used gene expression data and patient survival information from a large public database called The Cancer Genome Atlas (TCGA). After preparing the data, we used statistical analysis tools to find which genes correlated with better or worse survival outcomes.

Keywords: Survival analysis, Cancer hallmark genes, TCGA, RNA-seq, Prognostic biomarkers

1 Introduction

A pancancer project is a research study that looks at different types of cancer in order to find similarities and differences. Pancancer studies aim to compare genetic and molecular features across dozens of cancers. The goal of our study is to understand how hallmark genes are related to survival in cancer patients.

The Cancer Genome Atlas (TCGA), analyzes data from over 30 different types of cancer to help improve our overall understanding of cancer biology.

Normal cells evolving into cancer cells display "hallmarks of cancer." These hallmarks include sustained loss of growth suppressors, apoptosis resistance, replicative immortality, genomic instability, inflammation, and energy metabolism reprogramming. Hallmark genes are the specific genes that code for these hallmarks of cancer. These genes were grouped into five main hallmark categories. TCGA [3] includes gene expression profiles and clinical information from over 9,700 patients across 26 different cancer types. However, we limited our analysis to 7 different tumor types. By performing survival analysis, we sought to identify genes linked to patient outcomes and determine how these genes vary across different cancers.

In the study, transcriptomic signatures were developed for each hallmark based on the average expression of the genes associated with that hallmark. We examined the prognostic significance of these signature scores in different tumor types. Notably, we observed that gene expression patterns could vary greatly depending on the cancer type, underscoring the complexity of using these hallmark genes as predictive biomarkers.

Our results suggest that different hallmark processes may be more relevant in certain cancer than others. The goal of this study was to rank established cancer hallmark genes based on their relation to the survival of patients. By analyzing which genes are active or inactive in a tumor, doctors can better understand how aggressive the cancer is and which treatments are more likely to work. This could lead to the development of personalized medicine, where treatments are chosen based on a patient's specific genetic profile.

A good example of this is non-small cell lung cancer (NSCLC). In many patients with NSCLC, doctors test for changes in certain genes like EGFR, ALK, or ROS1. If a tumor has a mutation in any of these genes, the doctor may prescribe a targeted drug to block a specific gene's activity. This can be more effective than traditional chemotherapy.[4]

Building on previous research, we utilized the TCGA data set to perform univariate and multivariate survival analyzes using the Cox proportional-hazards regression model. These analyses allow us to assess the relationship between the expression of individual hallmark genes and patient survival, while accounting for other clinical variables such as age, gender, and race. Furthermore, we expand on previous studies by incorporating an updated list of 911 hallmark genes, a significant increase from the 671 genes used in previous analyzes. This broader set of genes provides a more comprehensive understanding of the molecular factors that influence cancer progression.

2 Methods

2.1 Database Setup

We used the R language and statistical environment to perform all data preprocessing and analysis (`https://www.r-project.org/`). The source code is available on GitHub (`https://github.com/ANewRag/Cancer-Hallmark-Gene-Analysis`). RNA-sequencing, clinical, and sample data were taken from The Cancer Genome Atlas (TCGA, `https://portal.gdc.cancer.gov/`).

2.2 Data Preprocessing

All data (which can be found by applying the filters "RNA-Seq", "Gene Expression Quantification", and "open" in the TCGA repository) came in a **star** file. For each sample, this file contained raw counts of sequenced fragments mapping to each gene along with normalized counts using two metrics: Fragments per Kilobase per Million (FPKM) and Transcripts per Million (TPM). These mRNA counts serve as a proxy for gene expression within a tumor cell.

Here's what the head of one of these files looks like (the unstranded label refers to the fact that we are looking at fragments from both strands of the DNA):

gene_name	**unstranded**	**tpm_unstranded**	**fpkm_unstranded**
TSPAN6	609	10.6344	5.0654
TNMD	2	0.1073	0.0511
DPM1	885	58.0771	27.6634
SCYL3	699	8.0439	3.8315
C1orf112	303	4.0201	1.9149
FGR	2011	47.0985	22.4341
CFH	6541	64.9247	30.9251

Unfortunately, because neither FPKM nor TPM model the discrete count nature of RNA-seq data or accounts for compositional biases, they are not ideal for formal differential-expression testing. Therefore, we re-normalized the raw count matrix using the DESeq2 package (`https://bioconductor.org/packages/release/bioc/html/DESeq2.html`) [1]. DESeq2 estimates a size factor for each sample by computing, for each gene, the ratio of its count to the geometric mean of that gene's counts across all samples, and then taking the median of those ratios. These size factors correct for differences in the size and composition of different samples, which is important as TCGA data comes from many different experiments conducted on many different patients. This enables accurate comparison of expression levels in downstream negative-binomial–based differential-expression analysis. While the original paper used **htseq** files (which don't contain normalized counts) and used the original DESeq library, the differences are minute.

Since each RNA-Seq count file only contains data from a single sample in a single study, these normalized counts were arranged into a table with rows representing the name of each gene represented in the count files and columns representing the sample from which they came.

2.3 Cancer Hallmark Genes

A list of 911 cancer hallmark genes were generated using the Bioconductor package (`https://www.bioconductor.org/install/`). These represent genes that have been identified to be associated with cancer in humans. This contrasts the previous study – which only used 671 hallmark genes – as the list of known hallmark genes has grown since its publication.

These genes were then further divided into five categories:

- Apoptosis - genes that regulate the process of programmed cell death,
- DNA Repair - genes that regulate the repair of damaged DNA,
- G2M Checkpoint - genes that control the transition from the G2 phase to mitosis,
- E2F Targets - genes activated by E2F transcription factors that drive DNA replication and the cell cycle,

- p53 Pathway - genes regulated by the tumor suppressor p53, involved in DNA repair, cell cycle arrest, and apoptosis

Note that a gene may fall into multiple of these categories [5, 6]. While all statistical analysis was done for each gene individually, plots were generated with respect to these categories.

2.4 Univariate Survival Analysis

We replicated the univariate survival analysis conducted by the original study. To analyze the relationship between the presence of different cancer hallmark genes and the survival time among samples from a given tumor type, we used a Cox proportional-hazards regression analysis. Model fitting and result extraction was done using the survival package (`https://cran.r-project.org/web/packages/survival/index.html`).

Files containing mRNA expression data, clinical metadata, and the list of cancer hallmark genes were loaded into the R statistical environment. These data were filtered to only include patients for whom clinical data was available. We then selected a tumor type for analysis. Survival time and event status (whether the patient was alive or dead) were extracted from the clinical dataset.

Each patient was then sorted into a ”high” or ”low” expression group for each gene by identifying an optimal cutoff point. This was done by choosing candidate cut-points from the 25th to the 75th percentile (in 1% increments), computing a univariate Cox model for each candidate, and selecting the one that yielded the lowest p-value (i.e., the most statistically significant split).

The Cox proportional-hazards model is a regression technique for time-to-event data (in our case, survival time) that relates predictors (“high” vs. “low” gene expression) to some hazard (death). It computes a hazard rate, $h(t)$ for a patient at a given time step t using the following formula:

$$h(t) = h(t)_0 \cdot \exp(\beta_{\text{exp}} \cdot X_{\text{exp}})$$

where:

- $h_0(t)$ is some “baseline” hazard common to all subjects.
- X_{exp} is 1 if the patient is “high” expression or 0 if the patient is “low” expression.
- β_{exp} is the ”partial likelihood” of hazard that only depends on the regression coefficient β_{exp} (in the case of univariate analysis, there is only one β, but in multivariate analysis this makes more sense). The value of $\exp(\beta_{\text{exp}})$ is the hazard ratio for ”high” vs. ”low” expressers (i.e., the hazard rate of ”high” expressers of the gene divide by that of ”low” expressers of the gene).

2.5 Multivariate Survival Analysis

Multivariate survival analysis was nearly identical to univariate survival analysis. The main difference was that the Cox regression model was fit with the following covariates:

- Gene expression group (high vs. low),
- Gender,
- Race,
- Age.

In full, the new formula for the hazard rate in the Cox model is

$$h_i(t) = h_0(t)\,\exp\Big(\sum_{j=1}^{7} \beta_j\, X_{ij}\Big).$$

Here, the covariates are

$$(X_{i1}, \dots, X_{i7}) =$$
$$\big(X_{i,\text{expr}},\ X_{i,\text{gender}},\ X_{i,\text{race,Asian}},\ X_{i,\text{race,Black}},\ X_{i,\text{age}},\ X_{i,\text{stage}},\ X_{i,\text{grade}}\big),$$

corresponding to high-vs-low gene expression, gender ($0 = $ female, $1 = $ male), race indicators vs. White, age in years, tumor stage, and tumor grade, respectively.

The original study also used tumor stage and tumor grade as covariates, but they were omitted from our analysis because a large fraction of the TCGA database did not have these data.

2.6 *p*-Value Correction

We calculated hazard ratios and p-values for each covariate and stored the results in a table. Then, p-values were adjusted using a Bonferroni correction where the number of "trials" was the number of genes (911). So, all p-values were multiplied by 911 and only considered significant if this product exceeded the threshold of 0.05.

3 Results

3.1 Transcriptomic database

The complete data set of RNA-seq samples that we analyzed consisted of 7 different tumor types. In the data set, gliomas was the largest (n=505) and thymoma was the smallest set (n=121). Most of the different tumor datasets has both overall survival (OS) and relapse-free survival (RFS) data, with the exception of AML and thymoma, which only had RFS data. In addtion to this, glioma and liver cancer patients had the longest and shortest median RFS at 23.8 and 6.7 months, respectively.

Clinico-pathological characteristics of patients, including, sex, race, ethnicity, and age were available for all 1915 of the cases we looked at in this paper. This was mainly due to the fact that the implementation of our code and analysis required these fields to be filled for proper analysis to be conducted.

3.2 Prognostic significance of hallmark gene types

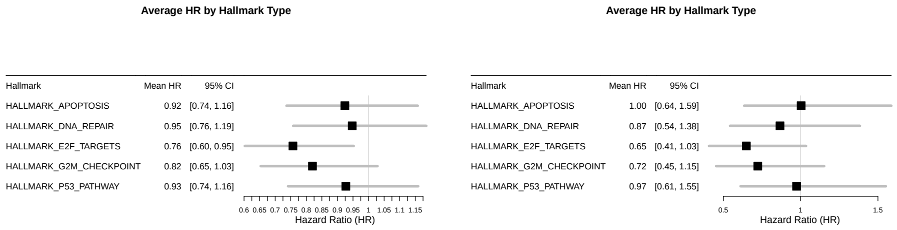

Figure 1: The hazard ratio distribution of each hallmark gene types in gliomas(left) compared to pacreatic adenocarcinoma(right)

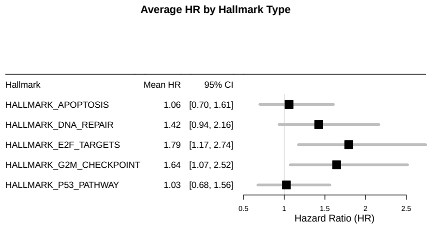

Figure 2: The hazard ratio distribution of each hallmark gene types in Thymoma

The study performed cox regression analysis on the RNA-seq expression of 911 hallmark genes. The study generate forest plot based on the results of univariate survival analysis across 7 tumor types, with genes that have p-value lower than 0.05 after Bonferroni correction. In both the gliomas and pancreatic adenocarcinoma cohort, there is a significant correlation between downregulation of E2F Targets pathway and G2M Checkpoint pathway with lower survival rates of patients with 95% confidence (See Fig.1). Conversely in the Thymoma cohort , there is a strong correlation between the upregulation of E2F Targets pathway, G2M Checkpoint pathway, and the overall patient survival rate with 95 % confidence (See Fig.2). Together, these findings show a divergent impact of hallmark gene types on the cancer prognosis based on the tumor type context.

3.3 Hallmark signatures and survival in different tumor types

The "survplot" R package (`http://www.cbs.dtu.dk/~eklund/survplot/`) was used to generate Kaplan-Meier plots that mapped survival probability over time across different hallmarks and levels of gene expression of these said hallmarks. In Figures 3 and 4, E2F target activation and DNA repair were selected as a few examples of how these survplots can be used to potentially learn more about different types of tumors.

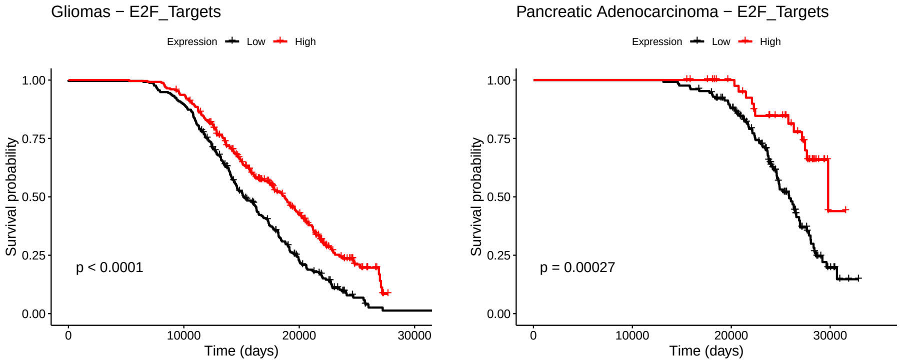

Figure 3: Kaplan–Meier survival curves showing overall survival arranged by E2F target gene expression levels (high, red; low, black) in gliomas (left) and pancreatic adenocarcinoma (right).

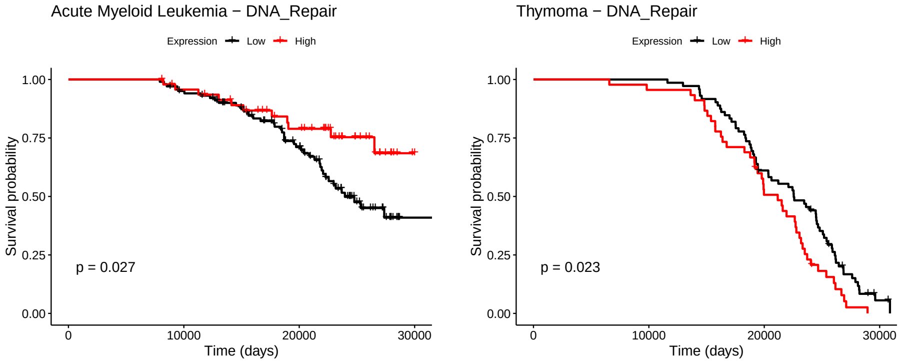

Figure 4: Kaplan–Meier survival curves for patients with acute myeloid leukemia (left) and thymoma (right), stratified by DNA repair gene expression levels (high, red; low, black).

From Figure 3, we can see that in gliomas and pancreatic cancer, higher expression of E2F targets was associated with improved overall survival. Similarly, from Figure 4, we can see that in acute myeloid leukemia higher DNA repair expression corresponded to better outcomes, whereas in thymoma, higher DNA repair expression was actually associated with a lower survival probability.

3.4 Genes with the greatest prognostic power in multiple tumor types

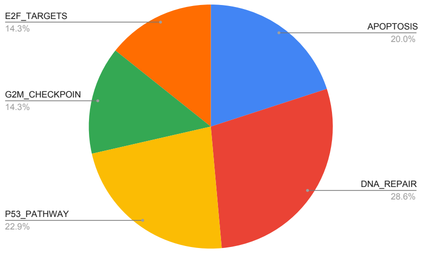

Figure 5: The top 5 hallmark gene types with highest hazard ratio distribution across 7 tumor types

We summarized the distribution of hallmark pathways among the 35 highest-risk genes (five per tumor type) across seven cancers using a pie chart. Each slice represents the percentage of hallmark gene types among all 35 genes chosen. The result of pie chart reveals that the DNA Repair pathway has the highest percentage among all hallmark gene types and shows its pan-cancer driven potential in driving poor prognosis in the 7 tumor types studied.

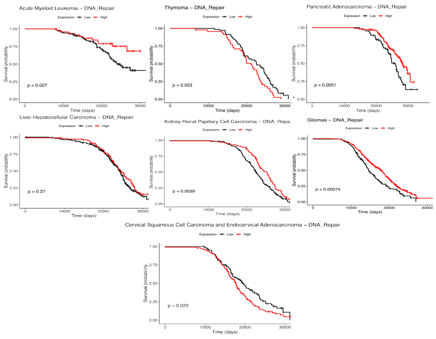

Figure 6: The Comparison of Survival plot of DNA Repair Pathway across 7 tumors

Across the 7 tumor types, the study examines the relation between DNA Repair pathway expression level and patient survival rate. Except Liver Hepatocellular Carcinoma, there is a significant survival rate difference between the patient group that has high expression and low expression (See Fig.6). In Acute Myeloid Leukemia, Pancreatic Adenocarcinoma, Kidney Renal Papillary Cell Carcinoma, and Gliomas, patients have a higher survival rate if the DNA Repair related hallmark genes have high expression. Conversely, for Thymoma and Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma, lower expression of DNA Repair hallmark genes produce higher survival rates.

4 Discussion

4.1 Hallmark Expression Patterns and Clinical/Experimental Outcomes

The survival plots shown in Figures 3 and 4 represent a small subset of the total number of plots generated by our computational pipeline and highlight key trends observed across the different tumor types that we covered in our datasets. In Figure 3, higher expression of E2F target genes was associated with improved survival in both gliomas and pancreatic adenocarcinoma. This finding suggests that therapuetic strategies targeting the E2F related pathways may have cross-cancer applicability, specifically regarding the gliomas and pancreatic cancer. In addition to this, the shared surivival benefit associated with E2F activation raises the question of whether common molecular features underlies this response in the otherwise different cancers. This data also potentially warrants future laboratory and clinical studies about whether these observations reflect shared biological features or just represnt independent, tumor-specific phenomena.

In Figure 4, there was no clear correlation between the expression of the DNA repair hallmark and survival rate between thymoma cancer and leukemia. In acute myeloid leukemia, higher expression of the DNA repair hallmark clearly led to increased survial rates among those affected. On the other hand, in thymoma, lower expression of DNA repair hallmark genes actually resulted in a higher survival rate, but there was overlap at points on the graph. First, this illustrates the important fact that not all cancers are the same and respond to the expression hallmark genes in various ways. In this specific case for thymoma, the data might suggest that due to the slight overlap in the graph, where high expression overtook low at one point, there might be certain DNA repair pathways within the general DNA repair hallmark that when expressed at a higher level, have a positive effect on survival rate, which could warrant further inspection or clinical observation.

4.2 Relation of DNA Repair Hallmark Genes with Survival Rates

DNA Repair plays important role in maintaining the cell stability from any harmful mutation. The DNA Repair enzymes ”correct damaged nucleotide residues generated by exposure to carcinogens and cytotoxic compounds”[8]. In our analysis across 7 tumor types, DNA Repair hallmark genes exhibit the highest hazard ratio. The survival plots also indicate that the expression level of DNA repair hallmark genes is prognostic in 6 out of 7 tumor types. The study observes in 2 of 6 tumor types that DNA Repair genes show prognostic effect, the higher expression of DNA Repair genes leads to lower survival rate of patients. This contradicts with DNA Repair genes' function to repair damaged chromosome. However, Bader and Bushell's article explains: ”association with high mutational burden and poor cancer prognosis indicates expression of these [DNet-genes] facilitates mutational burden via the repair of high levels of DNA damage. . . thus reducing patient survival” [2]. Accordingly, the study interprets the negative prognostic impact of elevated DNA repair gene expression as a surrogate for an underlying high mutational load. In patients with extensive genomic damage, DNA Repair pathways are regulated in response to the widespread mutation.

4.3 Limitation and Future Directions

In the study, due to the time limit, there are only 7 tumor types analyzed for the correlation between hallmark genes expression and the patient survival rates. Another limitation of the study is that all multivariate analysis results are removed due to the hazard ratio being greater than 1 in all genes in multiple factor categories. This is possibly due to the omission of cancer stage, which is a confounding clinical variable. Since cancer stage is positively correlated with hazard ratio, omitting this factor can cause skewing the hazard ratio upward.

For the next steps of the study, more tumors types should be included to find novel hallmark genes that have prognostic effect on multiple common tumor types. The data about cancer stage should also be included to negate its skewing effect. Moreover, Gene set enrichment analysis will also be performed to confirm the statistical significance of hallmark gene types' prognostic effects on patients survival rates.

Conflicts of Interest. The authors declare no conflicts of interest.

Acknowledgment. We would like to thank Dr. Phillip Compeau, Sarah Fisher, and Stephen Wu for their support throughout this project.

References

[1] W. Anders, S. Huber, Differential expression analysis for sequence count data, *Genome Biol*, **10**(R106), 2011, `https://doi.org/10.1186/gb-2010-11-10-r106`.

[2] Aldo S Bader, Martin Bushell, Damage-Net: A program for DNA repair meta-analysis identifies a network of novel repair genes that facilitate cancer evolution, *DNA Repair (Amst.)*, **105**(103158):103158, September 2021.

[3] L.A. Cooper, Pancancer insights from the cancer genome atlas: The pathologist's perspective, *The Journal of Pathology*, **244**:512–524, 2018, `https://doi.org/10.1002/path.5028`.

[4] J. H. Doroshow, Targeting egfr in non–small-cell lung cancer, *New England Journal of Medicine*, **353**:116–118, 2005, `https://doi.org/10.1056/NEJMe058113`.

[5] R. A. Hanahan, D. Weinberg, Hallmarks of cancer: The next generation, *Cell*, **144**:646–674, 2011, `https://doi.org/10.1016/j.cell.2011.02.013`.

[6] O. et al. Menyhart, Guidelines for the selection of functional assays to evaluate the hallmarks of cancer, *Biochem. Biophys. Acta.*, p. 300–319, 2016, `https://doi.org/10.1016/j.bbcan.2016.10.002`.

[7] Munkácsy G. Győrffy B. Nagy, Á., Pancancer survival analysis of cancer hallmark genes., *Sci Rep*, **11**, 2021, `https://doi.org/10.1038/s41598-021-84787-5`.

[8] D. Wood et al. Richard, Human dna repair genes, *Science*, **291**:1284–1289, 2001, `https://doi.org/10.1126/science.1056154`.