**PREDICTING THE IMPACT OF PATHWAY REGULATIONS ON OVERALL SURVIVAL TIME OF CANCER PATIENT**

**TEAM NAME: LIFELINE ANALYTICS GROUP**

**TEAM NUMBER: 12**

**TEAM MEMBERS**

|  |  |  |  |
| --- | --- | --- | --- |
| **Karan Raj Bagri**  SR No. 23049  *M.Tech CDS*  *karanraj@iisc.ac.in* | **Parvesh Barak**  SR No. 23181  *M.Tech (Res) CDS*  *parveshbarak@iisc.ac.in* | **Sauma Majumdar**  SR No. 23015  *Ph.D CFBR*  *saumasurva@iisc.ac.in* | **Suved Ghanmode**  SR No. 23339  *M.Tech (Res) CDS*  *suvedsanjay@iisc.ac.in* |

**INTRODUCTION**

Cancer is a large group of diseases that can start in almost any organ or tissue when there is an uncontrollable abnormal cell growth in that region of the body. The latter process is called as *metastasizing* and is a major cause of death from cancer. This disease is the second leading cause of death globally, accounting for an estimated 10 million deaths, in 2020 [1]. In cancer, the regulation of various cell signalling pathways is often disrupted, leading to uncontrolled cell growth, survival, and invasion. Upregulation and downregulation of pathways plays a crucial roles in the development and progression of cancer and is directly linked to the survival chances of patients.

We have identified the TCGA Pan Cancer dataset for the transcriptomic data, curated for clinician features. This dataset is comprehensive collection of multi-omics data that provides insights into gene expression patterns across a diverse set of cancer types. [2] This dataset includes a wide range of cancer types (33) which facilitates the identification of shared molecular characteristics and potential therapeutic targets, and additional fields like survival outcomes, tumour stage, etc. In addition to the pan cancer data set, we have used phenotype-curated clinician data which includes endpoints of Overall Survival (OS), Progression Free Interval (PFI), Disease Free Interval (DFI), and Disease Specific Survival (DSS) for each TCGA cancer type [3].

In this project, we have analysed this datasets to access the impact of regulation of certain pathways on the overall survival time of patients. At first, we did an exploratory data analysis to get insights of the dataset and further used statistical methods (discussed in method section) to predict how the pathway regulations are impacting the overall survival time and probability of the cancer patients.

**METHODS**

1. Statistical Methods

We used following statistical tools/methods for our project –

* Single Sample Gene Set Enrichment Analysis (ssGSEA)

This involves the computation of enrichment scores for predefined gene sets in individual gene expression profiles. Higher positive ssGSEA scores indicate increased activity of the corresponding gene sets, offering insights into the biological processes or pathways at play in individual samples. The method facilitates the systematic exploration of gene set enrichment in a single-sample context, contributing to the interpretation of complex gene expression data and the identification of biologically relevant patterns. [4]

* Kaplan-Meier Estimate  
   Kaplan-Meier is a statistical method used in the analysis of time to event data. Time to event means the time from entry into a study until a particular event, for example onset of illness. This method is very useful in survival analysis as it is used by the researchers to determine and/or analyse the patients or participants who lost to follow up or dropped out of the study, those who developed the disease of interest or survived it. It is also used to compare two groups of subjects such as a control group, the one that is given placebo and the other treatment group that is the one given the genuine drug.
* Cox regression model  
   The Cox proportional hazards model is a semiparametric statistical model that is commonly used to analyse survival data. The model assumes that the hazard of death for an individual is proportional to a linear combination of covariates. The model is also known as the Cox regression model or the proportional hazards model. The Cox proportional hazards model can be expressed as follows:  
   h(t; X) = h\_0(t) \* exp(β'X)  
   where:
  + - h(t; X) is the hazard of death for an individual with covariates X at time t
    - h\_0(t) is the baseline hazard function
    - β is a vector of regression coefficients
    - X is a vector of covariate
  + The regression coefficients β can be interpreted as the log-hazard ratios. A positive coefficient indicates that the corresponding covariate increases the hazard of death, while a negative coefficient indicates that the corresponding covariate decreases the hazard of death.
  + The baseline hazard function h\_0(t) is the hazard of death for an individual with all of the covariates set to 0.

1. Workflow

To start with, we first performed exploratory data analysis to get the insights on the data and filtered out top six cancers for further analysis which included BRCA, GBM, OV, UCEC, KIRC and HNSC.

BRCA - Breast Invasive Carcinoma   
GBM - Glioblastoma Multiforme (Cancer of the central nervous system)   
OV - Ovarian Cancer   
UCEC - Uterine Corpus Endometrial Carcinoma   
KIRC - Kidney Renal clear cell Carcinoma   
HNSC - Head and Neck Squamous cell Carcinoma

A black and white rectangular sign

Description automatically generated

Fig. Modelling workflow

* We used ‘ssgsea’ python module [5] to get the gene set enrichment score from gene expression file. We calculated the pathway enrichment scores for each sample for 1387 pathways taken from PARADIGM [6]. The scores were analysed, and we calculated the top 10 upregulated and downregulated pathways across the 6 cancer types.
* Grouping of patients according to pathway regulation  
  Numerous genes are linked to the signalling pathways, which have been repeatedly found to be dysregulated in various cancer types as a result of mutations or changed function of their products. There have been previous studies showing these effects can be traced down to the fundamental survival of the cells with increased proliferation and decreased apoptosis and directly affect the overall survival outcome of the patient. Here, we have studied the effect of up/downregulation of the top 3 gene sets (in terms of enrichment scores of the pathways), subset from the top 10 pathways as explained in the previous section, on the overall survival of the patients, by grouping them according to pathway enrichment levels.   
  The patients are initially categorized into three groups:
  + *No Regulation Group*   
    This group consists of patients where the expression of the selected gene sets is neither upregulated nor downregulated.
  + *Upregulated Gene Sets Groups*   
    For each of the top three upregulated gene sets, patients with gene expression levels above zero are grouped together.
  + *Downregulated Gene Sets Groups*   
    Similarly, for each of the top three downregulated gene sets, patients with pathway enrichment levels below zero are grouped together.
* We calculated the cumulative hazard H(t) at a given time t, which is defined as the integral of the hazard function h(t) from 0 to t. The cumulative hazard provides a measure of the cumulative risk of an event (e.g., death in this case) up to time t. To calculate the cumulative hazard at the median survival time, the Kaplan-Meier survival curves were fitted for each group of patients associated with a specific gene set. The survival function S(t) from the Kaplan-Meier analysis was used to derive the hazard function h(t). The cumulative hazard at the median survival time was then determined by evaluating H(t) at this specific time point. The pathway with the highest cumulative hazard at the median survival time was identified by comparing the calculated cumulative hazards across all gene sets, providing a quantitative measure of the adverse impact of each gene set on patient survival. The pathway associated with the maximum cumulative hazard represents the worst outcome on survival at the median time.
* Further, we selected the most hazardous pathway and performed Survival analysis using the Cox regression model to figure out how that pathway’s enrichment scores affected the survival probability over time. We fit the cox regression model to the data (whether the patient dies or not, observed time to death and the enrichment score calculated for that patient’s sample) and obtain p-value (null hypothesis being that the enrichment score does not influence the survival probability over time) and form a single plot containing survival curves for the min, mean and max enrichment scores amongst these samples.
* The Cox proportional hazards model is a semiparametric statistical model that is commonly used to analyse survival data. The model assumes that the hazard of death for an individual is proportional to a linear combination of covariates. The model is also known as the Cox regression model or the proportional hazards model.

**RESULTS**

1. Exploratory Data Analysis  
    **A graph of different colored rectangular shapes

   Description automatically generated with medium confidence A graph of cancer type

   Description automatically generated**

A graph showing different colored bars

Description automatically generated A graph of a number of bars

Description automatically generated with medium confidence

A graph with different colored bars

Description automatically generated A graph of different colored bars

Description automatically generated

Fig 2. Data insights from EDA

1. Pathway Analysis  
   We interestingly found that Ovarian cancer (OV) and uterine carcinosarcoma [UCE]) had very similar pathway upregulation and downregulation trends, which was unlike other cancer types.

A close-up of a graph

Description automatically generated

Fig 3. Heatmap showing common pathways upregulated and downregulated across the cancer types

1. Survival Analysis  
   Below plots shows KM survival curves for each cancer and the effect of most hazardous pathway on the survival

A graph showing a number of different colored lines

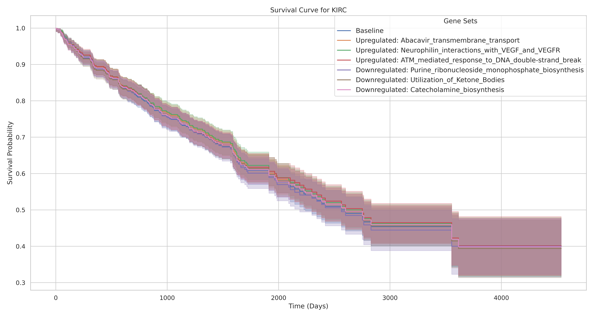
Description automatically generated A graph showing the number of patients

Description automatically generated

A graph showing a number of patients

Description automatically generated with medium confidence A graph showing a number of patients

Description automatically generated

 A graph showing the number of patients

Description automatically generated

Fig 4. Figures showing the effect of upregulation, downregulation and baseline expression of the selected pathways on the overall survival of patients.

|  |  |  |
| --- | --- | --- |
| **Cancer Type** | **Hazardous Pathway** | **Cumulative Hazard** |
| BRCA | Upregulated: Mtb\_iron\_assimilation\_by\_chelation | 0.06375662797876425 |
| UCEC | Upregulated: p53-Independent\_DNA\_Damage\_Response | 0.13397947146229183 |
| OV | Downregulated: Fructose\_catabolism | 0.30975623926977625 |
| KIRC | Baseline | 0.2569762813702292 |
| GBM | Downregulated: Vitamin\_B2\_(riboflavin)\_metabolism | 0.403731163717974 |
| HNSC | Downregulated: Beta\_oxidation\_of\_hexanoyl-CoA\_to\_butanoyl-CoA | 0.32253908004216814 |

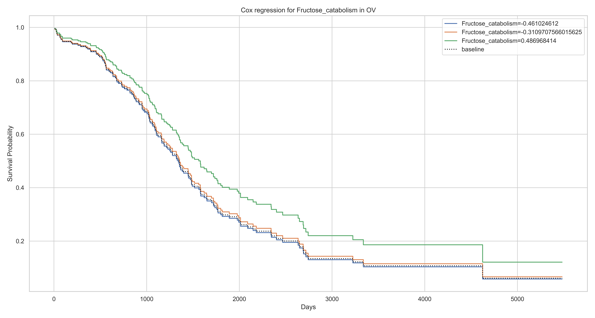
Table 1: The pathway with the worst outcome on survival at the median time for each cancer type

A graph showing different colored lines

Description automatically generated A graph showing the amount of vitamins

Description automatically generated

A graph showing a number of patients

Description automatically generated 

A graph showing a number of data

Description automatically generated with medium confidence

Fig 5. Effects of the minimum, maximum and average enrichment of the most hazardous pathway for each type of cancer

Despite the presence of similar upregulated and downregulated pathways in both OV and UCEC cancers, the most hazardous pathways differed markedly in terms of functionality. Ovarian cancer exhibited heightened mortality risk associated with a metabolic pathway, while uterine cancer linked to a DNA repair pathway. This disparity highlights the nuanced impact of pathway irregularities on patient survival, even when similar pathways are dysregulated across distinct cancer types.

**DISCUSSION**

Contrary to expectations, our analysis of KIRC did not pinpoint a specific pathway amongst the 6 identified i.e. 3 upregulated and 3 downregulated, irregularity associated with reduced survival. Instead, baseline expressions of the top differentially regulated pathways emerged as the most hazardous, underscoring the complexity of KIRC pathogenesis.

In contrast, the comparison between ovarian and uterine cancers revealed a noteworthy observation. Despite the presence of similar upregulated and downregulated pathways in both cancers, the most hazardous pathways differed markedly in terms of functionality. Ovarian cancer exhibited heightened mortality risk associated with a metabolic pathway, while uterine cancer linked to a DNA repair pathway. This disparity highlights the nuanced impact of pathway irregularities on patient survival, even when similar pathways are dysregulated across distinct cancer types.

The scale of our dataset, comprising around 500 patient samples for each cancer type analysed, enhances the robustness and significance of our findings. The diversity and size of the cohort allow for a more nuanced understanding of the intricate relationship between pathway irregularities and patient outcomes.

**CONCLUSION**

This study underscores the importance of considering cancer-specific nuances in pathway irregularities when developing targeted therapies. The fact that similar pathways can yield different impacts on patient survival across cancer types emphasizes the need for precision oncology approaches. As we delve deeper into the complexities of cellular signalling pathways, the integration of this knowledge into clinical practice holds great promise for improving patient outcomes. Our findings contribute valuable insights to the ongoing efforts in advancing precision medicine and underscore the necessity of a nuanced, cancer-type-specific approach in the quest to enhance patient survival.

# REFERENCES

|  |  |
| --- | --- |
| [1] | "Cancer," [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cancer. |
| [2] | "The Cancer Genome Atlas Program (TCGA)," [Online]. Available: https://www.cancer.gov/ccg/research/genome-sequencing/tcga. |
| [3] | "Assessment and Recommended Use of the Endpoints of OS, PFI, DFI, and DSS," [Online]. Available: https://www.cell.com/action/showFullTableImage?isHtml=true&tableId=tbl3&pii=S0092867418302290. |
| [4] | "ssGSEAProjection (v4)," [Online]. Available: https://www.genepattern.org/modules/docs/ssGSEAProjection/4. |
| [5] | "ssGSEA Python Module," [Online]. Available: https://gseapy.readthedocs.io/. |
| [6] | L. T. H. K. P. L. L. A. C. A. K. A. B. C. L. D. L. A. O. L. W. D. S. C. T. V. Liu J, C. G. A. R. Network and H. H., "An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics," in *Cell. 2018 Apr 5*, 2018. |
| [7] | X. S. D. &. Z. H. Huang, "Transcriptional Profiles from Paired Normal Samples Offer Complementary Information on Cancer Patient Survival – Evidence from TCGA Pan-Cancer Data," in *Sci Rep 6, 20567*, 2016. |
| [8] | D. C. A. F. H. S. C. L. M. R. E. B. R. G. E. A. R. B. R. B. D. C. T. G. R. G. A. H. S. H. A. H. E. H. J. H. T. K. R. K. B. K. D. L. W. M. M. M. Ricketts CJ, "The Cancer Genome Atlas Comprehensive Molecular Characterization of Renal Cell Carcinoma.," in *Cell Rep*, 2018. |
| [9] | "World Health Organization," [Online]. Available: https://www.who.int/health-topics/cancer. |