Exploring the Genomic Landscape of Pancreatic Cancer through Mutational Patterns and Survival Outcome Analysis

Tianyi Chu and Xinyue Qie

Introduction

Pancreatic ductal adenocarcinoma (PDAC) is a formidable challenge in oncology, often characterized by low survival rates, late diagnosis at advanced stages, and very limited treatment options. The majority of patients have unresectable, locally advanced, or metastatic disease at the time of diagnosis. Moreover, traditional treatments such as chemotherapy, surgery, and radiation have not been shown to significantly improve survival. With its aggressive nature and tendency to metastasize early, the five-year survival rate for pancreatic cancer remains dismal at around 10% (Oberstein and Olive 2013; Cancer Genome Atlas Research Network 2017; Sarantis et al. 2020.)

The etiology of PDAC is multifactorial, with risk factors including smoking, obesity, chronic pancreatitis, and family history of the disease. However, the exact mechanisms driving pancreatic tumorigenesis are not fully understood. Recent genomic studies have shed light on the molecular landscape of pancreatic cancer, revealing a complex interplay of genetic alterations that drive tumor initiation, progression, and metastasis.

Key mutations in genes such as KRAS and TP53, and those involved in DNA repair pathways have been identified as common drivers of pancreatic cancer (Saiki et al. 2021.) These mutations disrupt crucial cellular processes, including cell cycle regulation, DNA repair mechanisms, and chromatin remodeling, leading to uncontrolled cell proliferation and tumor growth. Targeted therapies directed against these specific mutations, such as inhibitors of the KRAS pathway, hold promise for improving treatment efficacy and patient survival. Therefore, understanding the genetic mutations driving pancreatic cancer is critical for developing personalized treatment approaches and prognostic stratification for improving patient outcomes, and integrating genomic data with clinical parameters can enhance risk assessment and guide therapeutic decisions, ultimately improving patient care and survival rates.

In this study, we aim to explore the genomic landscape of pancreatic cancer, delineate mutational signatures associated with disease progression, and investigate their implications for patient survival. By elucidating molecular mechanisms and identifying prognostic biomarkers, we hope to contribute to enhancing the diagnosis, treatment, and management of pancreatic cancer.

Research Strategy

Significance

In the context of PDAC prognosis, critical mutations in specific genes, notably KRAS, CDKN2A, TP53, and SMAD4, along with the subsequent activation of related signaling pathways, play an essential role in treatment resistance. Among these genes, KRAS stands out for its exceptionally high mutation rate; a cohort study involving whole-exome sequencing of 109 PDAC cases found KRAS mutations in over 90% of the samples. However, it has been established that KRAS mutations alone are insufficient to initiate cancer. The precise mechanisms through which secondary mutations influence PDAC outcomes remain unclear, underscoring the need for further investigation into how these additional genetic alterations impact prognosis.

Past research has demonstrated that mutations in the TP53 gene are present in 50–75% of human pancreatic ductal adenocarcinomas (PDAC), typically following an initial activating mutation in the KRAS gene (Morton et al. 2010.) Efforts to develop treatments have focused on targeting the mutant P53 protein. By examining the Pancreatic Adenocarcinoma TCGA PanCancer dataset, our objective is to identify additional genes that are prone to secondary mutations, similar to TP53. This approach aims to uncover new avenues for therapeutic intervention.

Furthermore, we propose to conduct survival analysis utilizing various models such as the Cox hazard proportional model, random survival forest, and deep learning-based models. This approach aims to illuminate the survival outcomes among patients with differing gene mutations. Notably, through the utilization of the deep learning-based models, we aim to leverage the advantage of feature selection. This capability is crucial as it allows us to distill a concise subset of features from extensive RNA seq data, which comprises over 20,000 genes.

This targeted selection can offer clinicians valuable insights into the etiology of PDAC. Given that gene information often constitutes weak signals, leveraging these feature selection models on clinical data could further empower clinicians to pinpoint the most relevant features, potentially enhancing the prognosis and diagnosis of PDAC.

Innovation

Prior research has explored the genes frequently mutated in PDAC through both in-vivo and in-vitro studies. In this study, leveraging the TCGA PanCancer dataset, we intend to delineate the genomic landscape of PDAC using a data-driven methodology. Our analysis will particularly concentrate on identifying mutations that occur independently of TP53 mutations, noting that about 30% of secondary mutations remain uncharted in their specific locations.

Furthermore, we plan to conduct exploratory analyses on samples with the identified secondary mutations. Through survival analysis, we aim to assess the impact of these mutations on patient survival, comparing the outcomes of patients with these genetic alterations against those without, and against patients with TP53 mutations. Our findings will elucidate the influence of gene mutations on survival prospects, highlighting how secondary mutations might differentiate survival outcomes in PDAC patients.

Concurrently, we will examine the nature of mutations within these frequently mutated genes. By integrating mutation data from the TCGA dataset with mRNA sequencing data, we will analyze patient samples based on gene expression levels of these mutated genes. This will allow us to classify patients with secondary mutations and examine how different mutations within the same gene affect patient survival. This comprehensive analysis will provide insights into whether specific types of mutations in the same genes influence the survival rates of PDAC patients differently.

Research Plan

In our project, we will utilize the Pancreatic Adenocarcinoma dataset from The Cancer Genome Atlas (TCGA) PanCancer Atlas project. This dataset provides comprehensive genomic and clinical information on pancreatic adenocarcinoma, a particularly aggressive and challenging cancer type. The TCGA PanCancer data offers a wealth of molecular profiling data,

including somatic mutations, gene expression patterns, DNA copy number alterations, and clinical outcomes, collected from a large cohort of pancreatic cancer patients.

Specifically, we will focus on the mutation dataset, the mRNA sequencing z-score dataset, the survival status of patients, and the clinical record of the patients which contains demographic information. Our variables of interest will be time-to-event (death), mutation locations, gene expression and the demographics of patients.

By leveraging this rich dataset, we aim to gain deeper insights into the molecular mechanisms driving pancreatic adenocarcinoma pathogenesis, identify key genetic drivers associated with disease progression and treatment response, and explore potential biomarkers for prognostication and personalized therapy. In particular, we aim to harness this information to analyze mutations and survival outcomes for patients using their clinical features and bulk RNA sequencing data.

Specific Aims

Hypothesis

We would like to divide the PDAC patient data into 4 groups based on their mutation data: (1) patients with no mutations in KRAS, (2) patients with only first-hit mutations in KRAS, (3) patients with both mutations in KRAS and TP53, and (4) patients with mutations in KRAS but not in TP53.

We believe that PDAC patients with mutations in KRAS but not TP53 (Group 4 patients) are likely to have mutations in other certain genes, which we denote as frequent genes for secondary mutations. We hypothesize that patients having mutations in these frequent genes have less optimal survival when compared to patients in group 1 and 2. We also hypothesize that patients in group 4 have different survival when compared to those in group 3.

Specifically, our hypothesis posits that the presence of specific mutations in key genes is correlated with lower survival outcomes in pancreatic cancer patients. We expect that individuals harboring mutations in these genes will exhibit reduced overall survival rates compared to those without such mutations.

Rationale

We intend to perform exploratory analyses on PDAC patient samples and identify frequent locations for secondary mutations other than TP53. By employing survival analysis, our goal is to evaluate how these mutations affect patient survival. We will compare the survival rates of patients harboring these genetic alterations with those who do not possess them, as well as with patients who have TP53 mutations. Our research will shed light on the role of gene mutations in influencing survival rates, emphasizing the distinct impact that secondary mutations have on the survival outcomes of PDAC patients.

Our rationale for this hypothesis is grounded in extensive literature documenting the critical roles of KRAS, CDKN2A, TP53, and SMAD4 in pancreatic tumorigenesis and disease progression. These genes are frequently mutated in pancreatic cancer and are known to exert significant influence over crucial cellular processes such as cell cycle regulation, DNA repair mechanisms, and signaling pathways involved in cancer development and metastasis.

Studies have demonstrated that mutations in KRAS, for instance, are among the earliest and most common genetic alterations in pancreatic cancer, contributing to tumor initiation and progression. Similarly, alterations in CDKN2A, TP53, and SMAD4 have been implicated in the development of aggressive tumor phenotypes, treatment resistance, and metastatic dissemination.

Given the pivotal roles of these genes in pancreatic cancer pathogenesis, it is plausible to hypothesize that the presence of mutations in KRAS, CDKN2A, TP53, and SMAD4 may confer a survival disadvantage to patients by promoting tumor aggressiveness, treatment resistance, and disease recurrence.

Experimental Approach

We aim to compare the impact of various mutations on survival outcomes by evaluating different computational survival methodologies and models. This encompasses, among others, the Cox proportional hazard model (CPH), the random survival forest model (RSF), and deep learning-based Cox regression models such as DeepSurv (Katzman et al. 2018), Stochastic Gates (STG) (Yamada et al. 2020), and Locally Sparse Interpretable Network (LSPIN) (Yang, Lindenbaum, and Kluger 2022). Specifically, DeepSurv is a simple multi-layer perceptron

version of the Cox proportional hazard model, STG offers global feature selection, and LSPIN provides instance-wise feature selection.

Moreover, we will investigate models incorporating different regularization methods to potentially explore dimensionality reduction and feature selection techniques. These methods aim to identify features that are related to or significantly contribute to survival outcomes. By employing a diverse range of computational approaches, our objective is to comprehensively analyze the impact of mutations on patient survival and uncover novel insights into the prognostic factors associated with pancreatic adenocarcinoma.

Results

The TCGA mutation dataset contains around 170 samples from 165 patients, with more than 15,000 mutated genes recorded. By grouping the TCGA mutation data, we found TP53 to be the gene that carries the most mutations outside KRAS, hosting approximately 30% of the identified mutations. This aligns with established results that TP53 is the most common second-hit mutation spot. To find other potential second-hit mutation markers, we filtered out patients with mutations in TP53 and examined the distribution of mutations. Since this group of patients don't have mutations in TP53, their second-hit mutation will be of interest since we are looking for mutually exclusive mutation hotspots. In similar manners, TTN and CDKN2A were found to be the next two frequent mutation carriers. Our exploratory analysis suggests that TP53, TTN, and CDKN2A can potentially be the mutation hotspots for PDAC.

Patients with mutations in these identified mutation hotspots are then grouped together, and their clinical outcomes were compared to patients without mutations in the hotspots. Using the clinical outcome dataset from TCGA, we examined the survival outcome of the two groups. Overall, the ratio between the surviving and deceased patients is 1:1, whereas the ratio falls in the group of patients who have mutations in the hotspot. In patients with mutations in TP53 and CDKN2A, the ratios between the survival and deceased patients fall as low as 1:2. This suggests that having mutations in these identified mutation hotspots likely affects the survival of patients and invites detailed analysis on survival outcome.

In the study, we explored the predictive performance, as well as the potential for feature selection, of various models on two distinct types of datasets: one focusing on clinical features and the other on gene mutation hotspots. For both constructed datasets, standard survival

models such as the Cox Proportional Hazard Model (CPH) (Katzman et al. 2018) and Random Survival Forest (RSF) (Ishwaran et al. 2008), alongside deep-learning-based models like DeepSurv (MLP), Stochastic Gates (STG), and Locally Sparse Interpretable Network (LSPIN), were trained and evaluated to analyze performance.

Given our interest in feature selection analysis, particularly concerning the Cox Proportional Hazard Model, we included and compared four versions with different penalization choices: Cox Vanilla (without any penalization), Cox Lasso (L1 norm), Cox Ridge (L2 norm), and Cox Elastic Net (a combination of L1 and L2 norms). To evaluate these models, we used the concordance index (CI) to assess survival predictive accuracy, and we assessed feature selection accuracy by determining the probability of including each feature in the dataset.

For the dataset centered around clinical features, we incorporated four primary clinical attributes: age, gender, cancer tumor stage, and whether the patient received radiation therapy. In order to evaluate the models' capability for feature selection, we augmented these features with 40 additional non-informative features sampled from Gaussian noise.

The benchmarked models showcased diverse levels of predictive accuracy. As illustrated in Table 1, LSPIN and DeepSurv demonstrated superior performance, achieving concordance indices of 0.707 and 0.679, respectively. Interestingly, there was little distinction in performance among the CPH models, RSF, and STG, as their concordance indices hovered around 0.5, indicating performance akin to random chance. However, the STG model exhibited strong performance in feature selection, as evidenced by Figure 1.

	Clinical Features + Gaussian Noise	Mutation Hotspots + Random Genes
Cox vanilla	0.429	0.319
Cox Lasso	0.556	0.622
Cox Ridge	0.435	0.681
Cox Elastic Net	0.546	0.62
RSF	0.505	0.638

STG	0.522	0.633
LSPIN	**0.707**	0.809
DeepSurv	0.679	**0.830**

Table 1: Model Performance Comparison on Clinical Features and Gene Mutation Hotspots Evaluated based on Concordance Index.

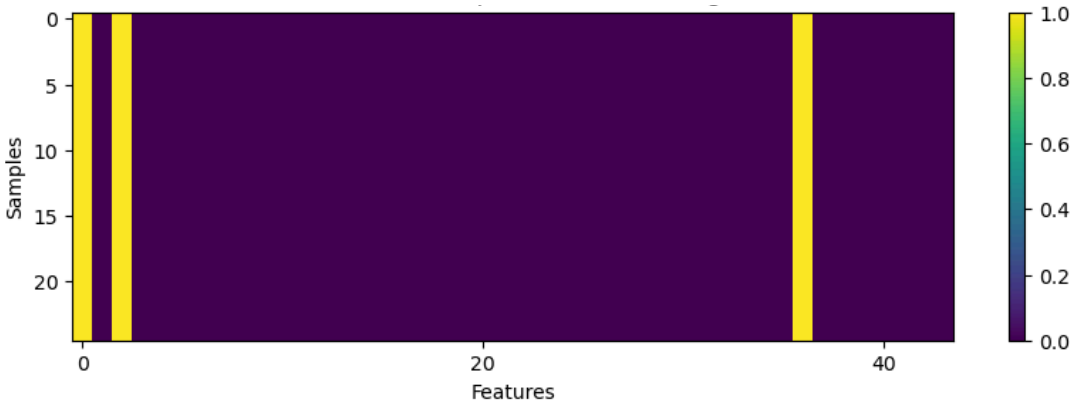

Figure 1: Feature Selection Heatmap for Test Data using STG Model. The features selected by the model include Age, Radiation Therapy Status, and a Gaussian noise feature.

In terms of the dataset concerning gene mutation hotspots, we included six mutational hotspots as discussed above: TTN, TP53, CDKN2A, RBM10, SF3B1, and ATM, which are considered informative features. Additionally, similar to the clinical feature dataset, we incorporated 194 random genes from the remaining genes to constitute the dataset's non-informative features, resulting in a total of 200 gene features. As before, all models were trained and evaluated using this gene dataset.

As shown in Table 1, regarding model prediction performance, LSPIN and DeepSurv continued to exhibit strong predictive power, achieving notably high concordance indices of 0.809 and 0.830, respectively. Except for the CPH model without any penalization, the remaining models demonstrated similar performance, with concordance indices between 0.6 and 0.7, while Cox-Ridge slightly outperformed them with a CI of 0.681. However, despite LSPIN and DeepSurv demonstrating good prediction accuracy, their performance in terms of feature selection was not satisfactory. The models achieving the highest concordance index

through optimization were achieved with a small penalization parameter λ, indicating minimal feature selection. However, Cox-Lasso provides feature selection results that can offer useful insights into informative genes, as the top features associated with large magnitudes of coefficients include two of the six mutation hotspots, CDKN2A and TP53 (see Figure 2).

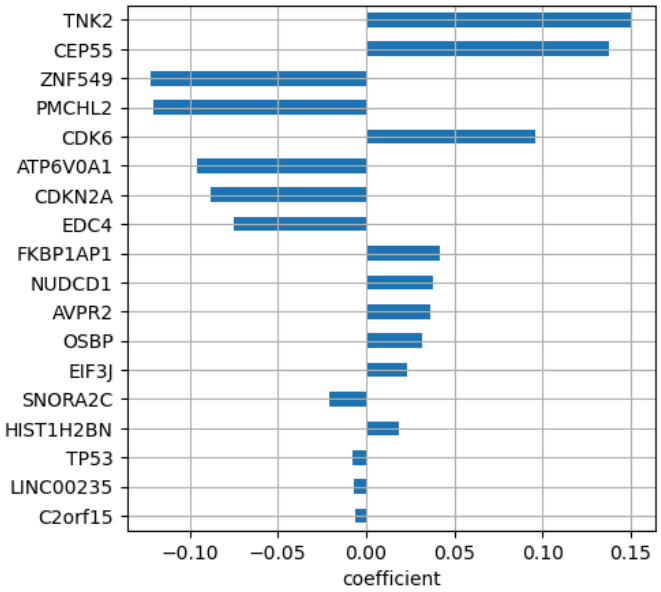

Figure 2: Feature Selection Results for the CPH Model with Lasso Penalization. Genes selected and their associated estimated coefficients are ranked by magnitude of coefficients in descending order.

Discussion

Our analysis reveals a significant association between mutations in KRAS, CDKN2A, TP53, and TTN and reduced survival outcomes in pancreatic cancer patients. The identification of specific genetic markers associated with poor prognosis could inform personalized treatment strategies and facilitate risk stratification for improved patient management.

After identifying the frequent locations for secondary mutations, we examined and compared the survival status of patients with or without mutations in these genes. The two patient groups have different survival outcomes, and patients without these secondary mutations tend to have longer time-to-death or have survived.

In terms of survival analysis, our findings suggest that clinical features generally serve as stronger predictors of survival compared to gene information. This superiority is attributed to the better feature selection ability observed in clinical features. While deep learning-based models exhibit higher survival prediction accuracy on the gene dataset, this could be attributed to the dataset's larger feature set, making it easier for the model to achieve higher concordance indices compared to datasets with fewer features. However, feature selection for deep learning-based models on the gene dataset is suboptimal. This may be because individual genes are weak predictors, and the RNA expression levels used in our analysis may not adequately represent mutations occurring in these hotspot genes. Additionally, results for deep learning-based models are generally unsatisfactory, likely due to the small sample size of the data. Nonetheless, simpler models such as the Cox proportional hazard model with appropriate penalization methods offer promising and insightful results.

This result provides insights into clinical research and prognosis because clinicians can look for mutations in these genes as a first step in diagnosis, and treatment specifically targeting these genes may be more efficient. Our result also provides insight for both in-vivo and in-vitro experiments. Using mouse models, researchers can knock out specific genes (for example, the identified mutation hotspots), and compare the tumor development with mice that have wild-type genotype. Findings from the mouse models can then be mapped back to human PDO (patient developed organoids) samples, as there exists homologous relationships between mouse and human chromosomes. Results can then be used to develop specific anti-splicing treatments.

Challenges may arise in the identification and validation of mutations, sample size limitations, and confounding factors affecting survival outcomes. Alternative approaches may involve integrating multi-omics data, employing machine learning algorithms for predictive modeling, and validating findings in independent cohorts to enhance the robustness and generalizability of results.

References

Cancer Genome Atlas Research Network. 2017. “Integrated Genomic Characterization of Pancreatic Ductal Adenocarcinoma.” Cancer Cell 32 (2): 185–203.e13. https://doi.org/10.1016/j.ccell.2017.07.007.

Ishwaran, Hemant, Udaya B. Kogalur, Eugene H. Blackstone, and Michael S. Lauer. 2008. “Random Survival Forests.” The Annals of Applied Statistics 2 (3). https://doi.org/10.1214/08-AOAS169.

Katzman, Jared L., Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. 2018. “DeepSurv: Personalized Treatment Recommender System Using a Cox Proportional Hazards Deep Neural Network.” BMC Medical Research Methodology 18 (1): 24. https://doi.org/10.1186/s12874-018- 0482-1.

Morton, Jennifer P., Paul Timpson, Saadia A. Karim, Rachel A. Ridgway, Dimitris Athineos, Brendan Doyle, Nigel B. Jamieson, et al. 2010. “Mutant P53 Drives Metastasis and Overcomes Growth Arrest/Senescence in Pancreatic Cancer.” Proceedings of the National Academy of Sciences 107 (1): 246–51. https: //doi.org/10.1073/pnas.0908428107.

Oberstein, Paul E., and Kenneth P. Olive. 2013. “Pancreatic Cancer: Why Is It so Hard to Treat?” Therapeutic Advances in Gastroenterology 6 (4): 321–37. https://doi.org/10.1177/1756283X13478680. Saiki, Yuriko, Can Jiang, Masaki Ohmuraya, and Toru Furukawa. 2021. “Genetic Mutations of Pancreatic

Cancer and Genetically Engineered Mouse Models.” Cancers 14 (1): 71. https://doi.org/10.3390/cancers1

4010071.

Sarantis, Panagiotis, Evangelos Koustas, Adriana Papadimitropoulou, Athanasios G Papavassiliou, and

Michalis V Karamouzis. 2020. “Pancreatic Ductal Adenocarcinoma: Treatment Hurdles, Tumor Mi- croenvironment and Immunotherapy.” World Journal of Gastrointestinal Oncology 12 (2): 173–81. https://doi.org/10.4251/wjgo.v12.i2.173.

Yamada, Yutaro, Ofir Lindenbaum, Sahand Negahban, and Yuval Kluger. 2020. “Feature Selection Using Stochastic Gates.” In International Conference on Machine Learning, 10648–59. PMLR. https://proceedi ngs.mlr.press/v119/yamada20a.html.

Yang, Junchen, Ofir Lindenbaum, and Yuval Kluger. 2022. “Locally Sparse Neural Networks for Tabular Biomedical Data.” In International Conference on Machine Learning, 25123–53. PMLR. https://proceedi ngs.mlr.press/v162/yang22i.html.