IJHSR

ijhighschoolresearch.org

■ RESEARCH ARTICLE

Identification of Pancreatic Cancer Driver Genes: Random Forest Modeling of Principal Components of scRNA-seq

Minnie J. Liang

West Lafayette Jr/Sr High School, 1105 Grant St, West Lafayette, IN, 47906,USA; liangminnie5@gmail.com
Mentor: Dr. Manpreet Katari

ABSTRACT: With the lowest 5-year survival rate among all cancers, pancreatic cancer is considered one of the most lethal diseases in the world. Currently, there are very few reliable genes to target in pancreatic cancer therapies; hence, identifying new biomarkers is crucially needed. This study develops a state-of-the-art machine learning workflow for biomarker discovery to first detect cancerous gene expression patterns through random forest modeling the principal components of the single-cell RNA-sequencing (scRNA-seq) data and, second, to identify the key genes driving these patterns. This method accounts for the effects of complex gene-gene interactions through random forest modeling and, thus, enables the genes driving cancer cell growth to be accurately identified. With a scRNA-seq dataset of pancreatic cancer, our workflow identifies several genes, including KRT17 and PTGS2, which have been validated as potential therapeutic targets in the literature. Our workflow also identifies novel genes that have never been studied, including MXRA5 and NDUFA6, while showing great potential as therapeutic targets. Their overexpression is linked with inferior survival in pancreatic cancer patients. Our workflow potentially accelerates discoveries of therapeutic targets for genetic diseases, and for pancreatic cancer, our newly-discovered genes provide promising directions for advancing its treatments.

KEYWORDS: Bioinformatics; Genomics; Machine Learning; Therapeutic Target; Pancreatic Cancer.

■ Introduction

Pancreatic cancer is known for its lethal and aggressive nature. With a meager 9% five-year survival rate (the lowest among all types of cancer),[1] pancreatic cancer is currently the fourth leading cause of cancer-related deaths in the United States. Unfortunately, its mortality keeps increasing year after year. By 2030, pancreatic cancer is projected to become the second most common cause of cancer-related deaths in the United States, coming second to lung cancer and surpassing both colorectal and breast cancer.[2]

With the development of next-generation sequencing techniques like single-cell RNA-sequencing (scRNA-seq), it has been demonstrated that genetic biomarkers can play an important role in cancer diagnosis and prognosis. Some of these biomarkers can be used as therapeutic targets. However, there are currently very few reliable biomarkers to target in therapies for pancreatic cancer. While for some other types of cancers, therapeutic gene targets have been identified, and the corresponding anti-cancer drugs have been developed. For example, the drug Herceptin has been developed for breast cancer, which targets the gene HER2 and can successfully stop cancerous breast cells from growing and dividing.[3,4] For lung cancer, EGFR and ALK have been identified as the therapeutic gene targets;[5] and for melanoma, BRAF has been identified as the therapeutic gene target.[6] Hence, identifying reliable therapeutic targets for pancreatic cancer is urgently required.

To address this requirement, some progress has been achieved in recent literature, with positive or negative results. For instance, Bouchard *et al.*[7] outlined a bioinformatics workflow for assessing mRNA expression levels, genomic alterations, and survival analysis of putative biomarkers for human cancers. They mainly studied the gene SENP1 but, unfortunately, found it unlikely to be an effective biomarker for pancreatic cancer. Li *et al.*[8] systematically analyzed molecular correlates of pancreatic cancer by bioinformatic analysis based on the TCGA pancreatic cancer dataset and found that two genes, ANLN and MYEOV, have potential value in pancreatic cancer immunotherapy. Cao *et al.*[9] found that some expressed sequence tags (ESTs) may represent diagnostically and therapeutically valuable targets for human cancers. They mainly studied some ESTs that were highly overexpressed in pancreatic adenocarcinomas. Vandenbrouck *et al.*[10] developed a bioinformatic workflow for identifying early cancer diagnosis biomarkers, particularly some candidate biomarkers of pancreatic cancer. It is important to note that these studies identified biomarkers by analyzing variations in expression levels or examining prognosis risk scores. Their analyses were often based on linear models, such as Cox regression, which fail to account for the effects of complex multiple gene-gene interactions, making the studies less effective in identifying valuable biomarkers.

We propose a new workflow for biomarker identification with scRNA-seq data to address this difficulty. The new workflow starts with a search for cancerous gene expression patterns using a machine-learning approach. It then refines the search by finding representative genes for the identified cancerous gene expression patterns. The machine learning approach we employed is a combination of principal component analysis (PCA) and random forest. The former allows us to extract different gene expression patterns from the scRNA-seq data.

DOI: 10.36838/v5i5.16

The latter enables us to identify the cancerous gene expression patterns via complex statistical modeling for cancer cell classification. In particular, the random forest model allows us to account for the effects of complex gene-gene interactions in the biomarker discovery process. With a scRNA-seq dataset of pancreatic cancer, our workflow identifies some known biomarkers of pancreatic cancer and some novel genes that we demonstrate to have high potential as therapeutic targets for pancreatic cancer. We validated the identified biomarkers by conducting a literature review and examining their impact on the survival of patients.

Methods

Dataset:

The dataset[11] used in this study was sequenced through scRNA-seq and consisted of primary pancreatic ductal adenocarcinoma tissues (95% of pancreatic cancer is pancreatic ductal adenocarcinoma[12]) from six different patients. The dataset consists of 19,738 genes and 11,563 cells of 20 different cell types. Two of the cell types are cancerous: Cancer clone A and Cancer clone B. Of the total number of cells, 756 of the cells are Cancer clone A and 1020 of the cells are Cancer clone B. The two cancer clones were merged into a single class in random forest modeling.

Pre-Processing:

In scRNA-seq, the expression levels of genes from thousands of individual cells are quantified, despite most genes not being expressed in most cells. As a result, we observed that the majority of gene expression levels were very low, with a value between 0 and 0.1. Because this skewed the data distribution toward small numbers, a log2 transformation (with a pseudo-count of 1 added to each expression value) was applied to all the gene expression values to help reduce the skewness of the data distribution.

The scRNA-seq data also suffer from a high dimensionality issue because the expression levels of tens of thousands of genes for each cell were measured. To tackle this issue, gene selection was performed: only 1200 genes with the highest standard deviations across different cells were kept. In this way, many genes, which are likely irrelevant to tumor growth but would possibly complicate the process of biomarker discovery, were eliminated from the analysis. This data processing procedure was conducted using Python packages "NumPy" and "pandas."

Principal Component Analysis:

After gene selection, PCA[13] was then performed to help further combat the problem of high dimensionality. The PCA procedure was conducted via the "scikit-learn" package in Python. Using PCA, we can project our data into a lower dimensional subspace based on the selected principal components (PCs). The first principal component (PC1) accounts for the highest data variability among all PCs, and each successive principal component (PC2, PC3, PC4...) accounts for less and less variability. To help decide which PCs to keep, we constructed and analyzed a scree plot, which displays the amount of variance explained by each PC. A threshold of explained variance of 70% was used as the selection criterion for PCs. We found the number of PCs such that the data's original information was preserved as much as possible while reducing its dimension.

Random Forest:

Our principal components became the features (inputs) of our machine-learning models to classify cancerous cells. We refer to each principal component as a principal feature in this paper.

Random forest[14] is a machine learning model architecture built by assembling multiple decision trees through bagging and random feature selection. The tree structure of random forest represents a conditional model, making it suited to handle nonlinear classification boundaries and account for highly nonlinear interactions, like gene-gene interactions. Numerous studies have identified random forest models to be able to uncover interaction effects.[15-22] In a random forest model, variables on a decision tree that appear together on a transversal path interact. This is because child nodes are based on the conditions of their parent nodes and are a product of other variables' interactions.[23] Other than model-fitting and prediction-making with random forest, we also output an importance measure for each principal feature based on node impurity and the probability of reaching the node. Meanwhile, we compared the results obtained from our random forest model with other machine learning models, including support vector machine (SVM) and logistic regression.

Workflow:

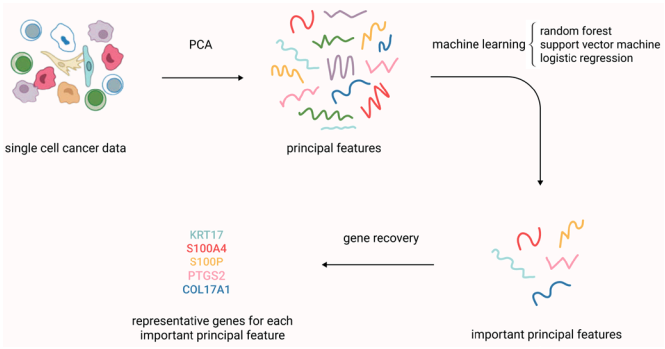

Figure 1: The novel workflow developed in this study for cancer biomarker identification.

As shown in Figure 1, the novel workflow developed in this study takes the scRNA-seq data and reduces its dimensions through PCA. From there, a random forest model is trained that uses those principal features as inputs to predict if a cell is cancerous. The principal features that the model most heavily relies on in cancer cell classification are identified as necessary. Then, gene recovery was performed to find the representative genes for each of the identified principal features.

The idea behind our workflow is that by using a random forest model to analyze the principal features, we can account for the complex nonlinear gene-gene interactions that drive the growth of cancerous cells. Then, the genes causing these effects can be recovered from the critical principal features and identified as the cancer driver genes.

Model Training:

The scRNA-seq data was randomly split into 80% for training (with stratified 5-fold cross-validation) and 20% for testing. Despite the scRNA-seq data containing two types of cancer cells, clones A and B, there was an overwhelming number of noncancerous cells: around 85% of the cells included in the data were noncancerous. The training data was downsampled by deleting some randomly selected noncancerous cells until around 59% of the cells were noncancerous, which prevented the model from overfitting to the noncancerous cell class. The testing data was left untouched.

The machine learning models were developed with the "scikit-learn" package in Python. To train the machine learning models, gridsearch with f1-scoring was used to experiment with all the combinations of hyperparameters and return the hyperparameter settings that resulted in the best performance score. The f1-score is calculated by taking the harmonic mean of precision and recall. The metric f1-score was chosen because it helps combat the problem of imbalanced classes in a dataset.

Gene Recovery:

The representative genes were determined by examining the importance of their contribution to each of the selected principal features, where the importance of each gene is reflected by the magnitude of the corresponding element of the selected principal feature. The higher the magnitude, the more a specific gene contributes to that principal feature and thus, the higher the importance. For each principal feature, we chose the top five genes as the representative genes for at least two reasons. First, we observed that the first few genes often had similar eigenvector values. Second, under the widely accepted sparsity assumption for gene regulatory networks, we believe that choosing "five" would ensure most potentially important representative genes to be included in our analyses, although not all.

Gene Literature Validation:

To validate the genes identified by our workflow, we clustered the scRNA-seq data through Uniform Manifold Approximation and Projection (UMAP).[24] UMAP analysis was conducted using the "umap" package in Python. We examined the expression of the identified genes in cancerous cell clusters compared to noncancerous cell clusters. Additionally, we cross-referenced the identified genes with literature publications that studied their use for pancreatic cancer treatments. By analyzing the breadth and scope of various studies, we show that our discoveries of pancreatic cancer driver genes are backed up by literature.

Survival Analysis:

Another method of validating the discovered genes by our workflow was through survival curves. These survival curves were carried out using the Kaplan-Meier Plotter.[25] The Kaplan-Meier survival analysis with log-rank and Cox proportional hazard tests was used to compare survival times among patients with pancreatic cancer, depending on their gene expression levels. The cutoffs to differentiate the gene expression levels between high and low were given in Z-scores. Specifically, we first converted the gene expression levels to the Z-scale, and then regarded a gene expression level as high if its Z-score is greater than 1, and as low if its Z-score is less than -1.

Results and Discussion

Principal Component Selection:

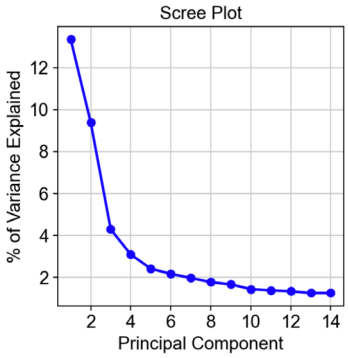

Figure 2: A scree plot of the first few principal components of our data. The x-axis displays the principal component number, and the y-axis displays the percentage of variance of the data that each principal component explains.

Examining the scree plot of the principal components in Figure 2, the curve begins flattening out at around principal component 4. However, the cumulation of principal components 1 to 4 only accounted for 30.10% of the total variance of the data. It is common practice to select principal components to be used that meet a threshold of explained variance of at least 70%.[26-29] For our dataset, the cumulation of principal components 1 to 34, accounting for 70.23% of the total variance, achieved this threshold. Furthermore, research has been conducted, like by Chang,[30] that showed theoretically the first few principal components do not necessarily contain more important information compared to later components. Combining Chang's findings with the 70% threshold criterion, principal components 1 to 34 were chosen to be kept.

Random Forest Performance:

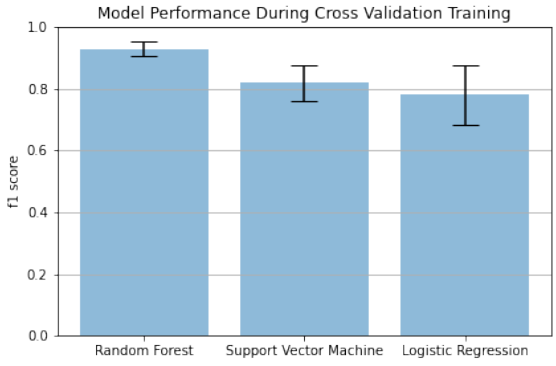

Figure 3: Bar plot with the standard error of the mean (SEM) f1-scores produced by random forest, support vector machine, and logistic regression in a 5-fold cross-validation assessment for their prediction performance.

For the three machine learning model architectures, hyperparameter tuning was performed during cross-validation training to develop each model. The f1-scores obtained by the three models during cross-validation training are shown in Figure 3. The random forest model yielded the highest f1-score, with statistically significant results as its SEM error bars

DOI: 10.36838/v5i5.16

do not overlap with the other two models. Then, the three developed models were tested by making predictions on the testing data, where each model's performance is summarized by a single f1-score, and it is not part of the cross-validation procedure. The f1-scores on the test dataset are displayed in Table 1. Out of the three machine learning model architectures, the best-performing model was the random forest. The random forest model achieved an f1- score of 0.85 out of 1.0 on the testing data.

Table 1: The f1-scores produced by three machine learning models on the testing data.

Model	f1 Score
Random Forest	0.85
Support Vector Machine	0.74
Logistic Regression	0.65

The random forest model performed best for the problem because it works on dimension-reduced data and it takes into account complex, nonlinear interactions amongst different features. In comparison, we used logistic regression in the standard way, which does not include feature interaction terms. Also it should be noted that the high-order feature-feature interaction terms would be hard for the logistic regression to handle, as there would be C(34,2)=561 two-way interaction terms, C(34,3)=5984 three-way interaction terms, and C(34,4)=46,376 four-way interaction terms. For the SVM model, we used the RBF kernel, which is a non-linear kernel. After trying all other kernels, including the linear one, we found that RBF performs the best. Compared to logistic regression, SVM, as a nonlinear model, is also able to account for feature-feature interactions. However, as explained below, working on the dimension-reduced data might deteriorate its performance in classification. Also, feature selection, the goal of the paper, seems not the strength of SVM.

In the literature, many authors have compared the performance of random forest and SVM for classification tasks, see e.g., Diaz-Uriarte[31] and Statnikov,[32] with diverse results presented. This indicates the performance of the two methods may be dataset dependent. In general, SVM works well for high-dimensional data due to its distance metric-dependent nature. In contrast, the random forest works well for the data with a small number of useful features. Therefore, it is not surprising that the random forest outperforms SVM on our dimension-reduced data. We expect that the results of model performances can be further improved with iterative random forests,[33] which can help to discover predictive and stable high-order interactions.

Important Principal Features:

The best-performing model, random forest, was then analyzed to determine the most critical principal features. As seen in Figure 4, the main features that are the most important to the random forest model are PC1, PC2, and PC3, as well as PC11 and PC12, according to the feature importance scores produced by the model.

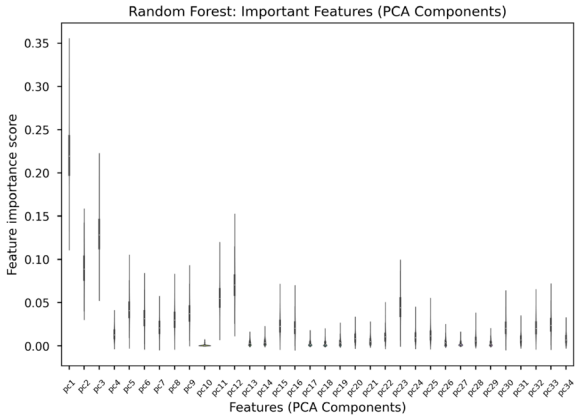

Figure 4: Feature importance scores produced by the random forest model, where PC1, PC2, PC3, PC11, and PC12 have the top five highest feature importance scores.

PC1 having the highest feature importance score can be rationalized based on the theory behind PCA, as PC1 always accounts for the highest variability among all PCs of the data. While it has ordinarily been the practice to prioritize the first few principal components as they contain most data variations, Chang[30] showed theoretically that the first few principal components only sometimes contain more information. It is essential to consider these latter principal components. In our case, the top five highest feature importance scores include PC1, PC2, PC3, PC11, and PC12.

Gene Recovery:

Table 2: Top contributing genes for each of the identified important principal features.

PC	Top 1 Contributing Gene
PC1	KRT17
PC2	S100A4
PC3	S100P
PC4	PTGS2
PC5	COL17A1

Cancer driver genes were then identified (see Table 2) by finding the top contributing gene to each of these five notable principal features: PC1, PC2, PC3, PC11, and PC12. To see if the identified genes are related to pancreatic cancer, we clustered the scRNA-seq data by cell type and further colored them by the expression levels of these genes. Examining the colored gene expressions in Figure 5, all the identified genes have very high expression levels in the Cancer Clone A and B clusters. This affirms that our workflow indeed identified driver genes of pancreatic cancer.

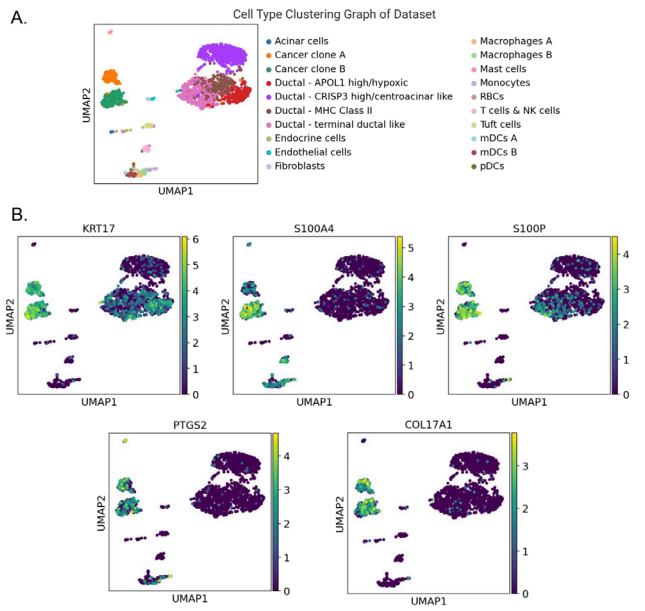

Figure 5: A) Clustering cell types of the dataset with UMAP. B) The same cell clustering map from part A but now colored by gene expression. The genes used to color were the top contributing genes for the five important principal features: PC1, PC2, PC3, PC11, and PC12.

It is interesting to note that although PC11 and PC12 represented relatively low portions of the variability of the original data, they still represented key cancerous gene expression patterns and led to the identification of PTGS2 and COL17A1. Our findings corroborate Chang's finding[30] that the later PCs also contain important information about the data. Additionally, genes KRT17 and S100P have notable expression in the ductal cell clusters, which can be explained as our dataset was sequenced from pancreatic ductal adenocarcinoma tumors.

Literature Validation:

Next, we identified each principal feature's top five contributing genes (see Table 3). The association between the model's top contributing genes and pancreatic cancer was validated through the literature on PubMed at https://www.ncbi.nlm.nih.gov/pmc/. As shown in Table 3, our developed workflow identified several genes color-coded in red and orange whose use for pancreatic cancer therapeutics has been substantiated by hundreds of publications in the literature.

Table 3: Top five contributing genes for each of the identified important principal features, where different colors indicate the number of relevant PubMed publications on the gene and pancreatic cancer: 0, 1-15, 15-150, 150+.

PC	Top 5 Contributing Genes
PC1	KRT17 , KRT16 , SLC20A1, COL6A1 , THBS1
PC2	S100A4 , VIM , CTSE , ITPRID2 , CXCL14
PC3	S100P , IGFBP7 , PITX1 , TRIM29 , NDUFA6
PC4	PTGS2 , PLAC8 , SERPINB5 , MXRA5 , AHNAK2
PC5	COL17A1 , TPX2 , TNK2 , PADI1 , MROH6

Let us examine some genes in further depth. For instance, KRT17 is known as a protein-coding gene expressed in hair follicles and nail beds. On the surface, this gene seems unrelated to pancreatic cancer. However, multiple publications[34,35] have reported that KRT17 regulates the proliferation, migration, and invasion of pancreatic cancer cells. Chen *et al.*[36] found that the overexpression of this gene is linked with low survival rates in patients with pancreatic cancer. Additionally, Pan *et al.*[37] identified KRT17 to have high potential as a novel target for making biomarker-based personalized treatment for pancreatic cancer.

In the literature,[38] the gene S100A4 is known to be associated with poor clinical outcomes for cancer patients. Che *et al.*[39] found that silencing this gene decreases pancreatic cancer cell invasion. Both Jia *et al.*[40] and Tsukamoto *et al.*[41] reported that S100A4 promotes tumorigenic phenotypes of pancreatic cancer *in-vivo* by regulating cell migration and invasion. Numerous studies[41-43] also showed that the gene is an attractive candidate as a prognostic marker and a therapeutic target in pancreatic cancer.

For the gene S100P, Deng *et al.*[44] reported that this gene plays a vital role in detecting pancreatic cancer with high sensitivity and specificity, and it might be a promising diagnostic marker for pancreatic cancer. Several studies[45-47] also found this gene to play a significant role in the aggressiveness of pancreatic cancer, and interference with S100P showed high promise to provide a novel approach for treatment. Camara *et al.*[49] showed the potential of using S100P's small molecule inhibitors for pancreatic cancer drug candidate development.

The gene PTGS2 is known to be a key factor in inflammation, and it is upregulated in pancreatic cancer.[50] Multiple studies[51,52] reported that high levels of PTGS2 are associated with poor prognosis of cancer. Markosyan *et al.*[53] also showed that pharmacological inhibition of PTGS2 sensitizes pancreatic tumors to immunotherapy. Additionally, Yip-Schnieder *et al.*[54] found the ability of PTGS2 inhibitors to inhibit tumor cell growth in-vitro, showing high promise as a therapeutic target for pancreatic cancer.

In the literature, high expression of COL17A1 is associated with poorer overall survival in patients with pancreatic cancer. Mao *et al.*[55] also found the gene to have significant prognostic value and potential as a therapeutic target. When analyzing survival in patients with pancreatic cancer, both Zhang *et al.*[56] and Wu *et al.*[57] reported COL17A1 as one of the critical gene expression signatures.

SLC20A1 and PADI, identified genes that have fewer literature publications on them, were established using R software by Demirkol *et al.*[58] and Chen *et al.*[59], respectively, to be part of novel gene signatures that define not only prognostic and biological sub-groups but can predict response to targeted therapy for pancreatic cancer as well. However, this was only studied through bioinformatic tools. Further experimental research is needed to explore the high potential of these two genes in clinical use.

ITPRID2, NDUFA6, MXRA5, and MROH6 are genes identified by our developed workflow that have few to no studies on their relation with pancreatic cancer. Further research of these genes in the experimental and clinical setting

DOI: 10.36838/v5i5.16

may provide new insights into advancing pancreatic cancer prognosis and treatment.

Survival Analysis:

Kaplan-Meier survival curves were constructed for the barely-researched genes or never researched before: SLC20A1, PADI1, ITPRID2, MXRA5, NDUFA6, and MROH6.

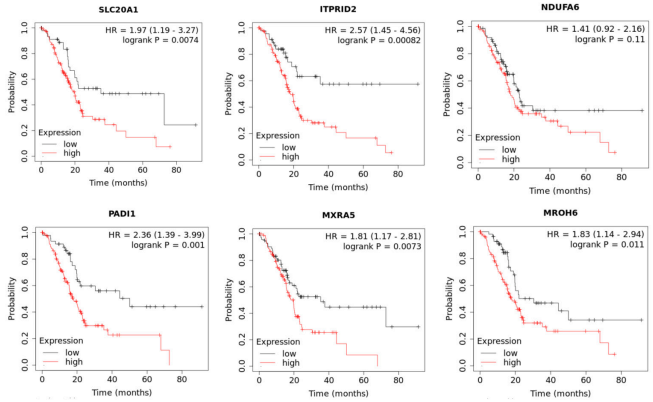

Figure 6: Kaplan-Meier survival curves constructed for barely-researched and novel genes (genes in green and blue in Table 3). All genes identified had their overexpression linked with inferior survival.

It is noted that the survival curve comparison between high and low expressions of NDUFA6 produced a p-value of 0.11, which is not statistically significant at the significance level of 0.05. That is, NDUFA6 might be a false positive finding. However, as seen in Figure 6, the comparisons for all the other five genes show significant results; their p-values are smaller or much smaller than 0.05. These comparisons suggest that developing a new therapy or drug targeting these genes—SLC20A1, PADI1, ITPRID2, MXRA5, and MROH6—may lead to life-saving treatment for pancreatic cancer.

However, there are some limitations to this study. First, a further improvement of the random forest model is possible, for example, deploying an iterative random forest architecture, which will be able to more efficiently detect high-order interactions among genes. Second, our findings are limited by the retrospective nature of our analysis. Our dataset was obtained from the GEO database and may contain potential sampling bias. Most of the ethnicities of the patients in this database are Whites, Latinos, or Africans. Caution must be used when generalizing this study's results to other patients from other ethnicities. Expanding our machine learning models using external datasets with a wider spectrum of patient composition will further solidify our findings. Thirdly, further experimental studies should be conducted to verify the identified cancer driver genes. This can include conducting a gene knockdown experiment in a cancer cell culture and measuring the growth of the cancer cells over time. Additionally, further research is needed to clarify the molecular mechanisms behind these genes for the progression of pancreatic cancer.

Conclusion

In summary, we have developed a novel machine-learning method for identifying cancer driver genes with scRNA-seq data through principal component analysis and random forest modeling combined. Compared to the existing methods, our process accounts for the effects of complex multiple gene-gene interactions on cancer cell growth. It thus enables cancer driver genes to be accurately identified. Our approach has been applied to a scRNA-seq pancreatic cancer dataset and achieved promising results. The identified cancer driver genes have been validated through gene expression clustering graphs, literature review, and survival curve comparisons. Not only did our method identify genes supported by numerous existing publications in the literature, but it also identified genes with little to no studies on their use for pancreatic cancer treatment. These newly identified genes provide potential targets for further experimental study to advance pancreatic cancer therapies and drugs.

Acknowledgments

I want to thank my mentor, Dr. Manpreet Katari, for his support and advice in this project.

References

1. Rawla, P.; Sunkara, T.; Gaduputi, V. Epidemiology of Pancreatic Cancer: Global Trends, Etiology and Risk Factors. World Journal of Oncology 2019, 10 (1), 10–27.
2. Rahib, L.; Wehner, M. R.; Matrisian, L. M.; Nead, K. T. Estimated Projection of Us Cancer Incidence and Death to 2040. JAMA Network Open 2021, 4 (4).
3. Tokumaru, Y.; Joyce, D.; Takabe, K. Current Status and Limitations of Immunotherapy for Breast Cancer. Surgery 2020, 167 (3), 628–630.
4. Nahta, R.; Yu, D.; Hung, M.-C.; Hortobagyi, G. N.; Esteva, F. J. Mechanisms of Disease: Understanding Resistance to HER2-Targeted Therapy in Human Breast Cancer. Nature Clinical Practice Oncology 2006, 3 (5), 269–280.
5. Lindeman, N. I.; Cagle, P. T.; Beasley, M. B.; Chitale, D. A.; Dacic, S.; Giaccone, G.; Jenkins, R. B.; Kwiatkowski, D. J.; Saldivar, J.-S.; Squire, J.; Thunnissen, E.; Ladanyi, M. Molecular Testing Guideline for Selection of Lung Cancer Patients for EGFR and ALK Tyrosine Kinase Inhibitors: Guideline from the College of American Pathologists, International Association for the Study of Lung Cancer, and Association for Molecular Pathology. Journal of Thoracic Oncology 2013, 8 (7), 823–859.
6. Flaherty, K. T.; Puzanov, I.; Kim, K. B.; Ribas, A.; McArthur, G. A.; Sosman, J. A.; O'Dwyer, P. J.; Lee, R. J.; Grippo, J. F.; Nolop, K.; Chapman, P. B. Inhibition of Mutated, Activated BRAF in Metastatic Melanoma. New England Journal of Medicine 2010, 363 (9), 809–819.
7. Bouchard, D. M.; Matunis, M. J. A Cellular and Bioinformatics Analysis of the SENP1 Sumo Isopeptidase in Pancreatic Cancer. Journal of Gastrointestinal Oncology 2019, 10 (5), 821–830.
8. Li, Z.; Hu, C.; Yang, Z.; Yang, M.; Fang, J.; Zhou, X. Bioinformatic Analysis of Prognostic and Immune-Related Genes in Pancreatic Cancer. Computational and Mathematical Methods in Medicine 2021, 2021, 1–23.
9. Cao, D.; Hustinx, S. R.; Sui, G.; Bala, P.; Sato, N.; Martin, S.; Maitra, A.; Murphy, K. M.; Cameron, J. L.; Yeo, C. J.; Kern, S. E.; Goggins, M.; Pandey, A.; Hruban, R. H. Identification of Novel Highly Expressed Genes in Pancreatic Ductal Adenocarcinomas through a Bioinformatics Analysis of Expressed Sequence Tags. Cancer Biology & Therapy 2004, 3 (11), 1081–1089.
10. Vandenbrouck, Y.; Christiany, D.; Combes, F.; Loux, V.; Brun, V. Bioinformatics Tools and Workflow to Select Blood Biomarkers for Early Cancer Diagnosis: An Application to Pancreatic Cancer. PROTEOMICS 2019, 19 (21-22), 1800489.
11. Moncada, R.; Barkley, D.; Wagner, F.; Chiodin, M.; Devlin, J. C.; Baron, M.; Hajdu, C. H.; Simeone, D. M.; Yanai, I. Integrating Mi

croarray-Based Spatial Transcriptomics and Single-Cell RNA-Seq Reveals Tissue Architecture in Pancreatic Ductal Adenocarcinomas. Nature Biotechnology 2020, 38 (3), 333–342.
12. Becker, A. E. Pancreatic Ductal Adenocarcinoma: Risk Factors, Screening, and Early Detection. World Journal of Gastroenterology 2014, 20 (32), 11182.
13. Bro, R.; Smilde, A. K. Principal Component Analysis. Anal. Methods 2014, 6 (9), 2812–2831.
14. Breiman, L. Random Forests. Machine Learning 2001, 45, 5–32.
15. McKinney, B. A.; Reif, D. M.; Ritchie, M. D.; Moore, J. H. Machine learning for detecting gene-gene interactions: a review. Appl Bioinforma 2006, 5(2), 77–88.
16. Hastie, T.; Tibshirani, R.; Friedman, J. The Elements of Statistical Learning 2009.
17. Goldstein, B. A.; Polley, E. C.; Briggs, F. Random forests for genetic association studies. Stat Appl Genet Mol Biol 2011, 10(1), 32.
18. Liu, C.; Ackerman, H. H.; Carulli J. P. A genome-wide screen of gene–gene interactions for rheumatoid arthritis susceptibility. Hum Genet 2011, 129(5), 473–85.
19. Grömping, U. Variable importance assessment in regression: linear regression versus random forest. Am Stat 2009, 63(4), 308–19.
20. Yang, P.; Hwa Yang, Y.; Zhou, B. B.; Zomaya, A. Y. A review of ensemble methods in bioinformatics. Curr Bioinform 2010, 5(4), ,296–308.
21. Moore, J. H.; Asselbergs, F. W.; Williams, S. M. Bioinformatics challenges for genome-wide association studies. Bioinformatics 2010, 26(4), 445–55.
22. Touw, W. G.; Bayjanov, J. R.; Overmars, L.; Backus, L.; Boekhorst, J.; Wels, M.; van Hijum, S. Data mining in the life sciences with random forest: a walk in the park or lost in the jungle? Brief Bioinform 2013, 14(3), 315–26.
23. Kingsford, C.; Salzberg, S. What are decision trees?. Nat Biotechnology 2008, 26, 1011–1013.
24. McInnes, L.; Healy, J.; Saul, N.; Großberger, L. UMAP: Uniform Manifold Approximation and Projection. Journal of Open Source Software 2018, 3 (29), 861.
25. Lanczky, A.; Gyorffy, B. Web-Based Survival Analysis Tool Tailored for Medical Research (KMplot): Development and Implementation. J Med Internet Res 2021, 23(7), e27633.
26. Jolliffe, I. T.; Cadima, J. Principal component analysis: a review and recent developments. Philos Trans A Math Phys Eng Sci 2016, 374(2065), 20150202.
27. Shen, H.; Huang, J. Z. Sparse principal component analysis via regularized low rank matrix approximation. Journal of multivariate analysis 2008, 99(6), pp.1015-1034.
28. Suhr, D. D. Principal component analysis vs. exploratory factor analysis. SUGI 30 proceedings 2005, 230.
29. Yoshioka, Y.; Iwata, H.; Ohsawa, R. Y. O.; Ninomiya, S. Analysis of petal shape variation of Primula sieboldii by elliptic Fourier descriptors and principal component analysis. Annals of Botany 2004, 94(5), 657-664.
30. Chang, W. C. On Using Principal Components before Separating a Mixture of Two Multivariate Normal Distributions. Applied Statistics 1983, 32 (3), 267.
31. Díaz-Uriarte, R.; Alvarez de Andrés, S. Gene selection and classification of microarray data using random forest. BMC Bioinformatics 2007, 7, 3.
32. Statnikov, A.; Wang, L.; Aliferis, C. F. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. BMC Bioinformatics 2008, 9, 319.
33. Basu, S.; Kumbier, K.; Brown, J. B.; Yu, B. Iterative random forests to discover predictive and stable high-order interactions. Proc Natl Acad Sci 2008, 115, 1943–1948.
34. Li, D.; Ni, X. F.; Tang, H.; Zhang, J.; Zheng, C.; Lin, J.; Wang, C. ; Sun, L.; Chen, B. KRT17 Functions as a Tumor Promoter and Regulates Proliferation, Migration and Invasion in Pancreatic Cancer via MTOR/S6K1 Pathway. Cancer Management and Research 2020, Volume 12, 2087–2095.
35. Zeng, Y.; Zou, M.; Liu, Y.; Que, K.; Wang, Y.; Liu, C.; Gong, J.: You, Y. Keratin 17 Suppresses Cell Proliferation and Epithelial-Mesenchymal Transition in Pancreatic Cancer. Frontiers in Medicine 2020, 7.
36. Chen, P.; Shen, Z.; Fang, X.; Wang, G.; Wang, X.; Wang, J.; Xi, S. Silencing of Keratin 17 by Lentivirus-Mediated Short Hair pin RNA Inhibits the Proliferation of PANC-1 Human Pancreatic Cancer Cells. Oncology Letters 2020.
37. Pan, C. H.; Otsuka, Y.; Sridharan, B. P.; Woo, M.; Leiton, C. V.; Babu, S.; Torrente Gonçalves, M.; Kawalerski, R. R.; K. Bai, J. D.; Chang, D. K.; Biankin, A. V.; Scampavia, L.; Spicer, T.; Escobar-Hoyos, L. F.; Shroyer, K. R. An Unbiased High-Through put Drug Screen Reveals a Potential Therapeutic Vulnerability in the Most Lethal Molecular Subtype of Pancreatic Cancer. Molecular Oncology 2020, 14 (8), 1800–1816.
38. Kozono, S.; Ohuchida, K.; Ohtsuka, T.; Cui, L.; Eguchi, D.; Fujiwara, K.; Zhao, M.; Mizumoto, K.; Tanaka, M. S100A4 Mrna Expression Level Is a Predictor of Radioresistance of Pancreatic Cancer Cells. Oncology Reports 2013, 30 (4), 1601–1608.
39. Che, P.; Yang, Y.; Han, X.; Hu, M.; Sellers, J. C.; Londono-Joshi, A. I.; Cai, G.-Q.; Buchsbaum, D. J.; Christein, J. D.; Tang, Q.; Chen, D.; Li, Q.; Grizzle, W. E.; Lu, Y. Y.; Ding, Q. S100A4 Promotes Pancreatic Cancer Progression through a Dual Signaling Pathway Mediated by SRC and Focal Adhesion Kinase. Scientific Reports 2015, 5 (1).
40. Jia, F.; Liu, M.; Li, X.; Zhang, F.; Yue, S.; Liu, J. Relationship between S100A4 Protein Expression and Pre-Operative Serum CA19.9 Levels in Pancreatic Carcinoma and Its Prognostic Signifiance. World Journal of Surgical Oncology 2019, 17 (1).
41. Tsukamoto, N.; Egawa, S.; Akada, M.; Abe, K.; Saiki, Y.; Kaneko, N.; Yokoyama, S.; Shima, K.; Yamamura, A.; Motoi, F.; Abe, H.; Hayashi, H.; Ishida, K.; Moriya, T.; Tabata, T.; Kondo, E.; Kanai, N.; Gu, Z.; Sunamura, M.; Unno, M.; Horii, A. The Expression of S100A4 in Human Pancreatic Cancer Is Associated with Invasion. Pancreas 2013, 42 (6), 1027–1033.
42. Treese, C.; Hartl, K.; Pötzsch, M.; Dahlmann, M.; von Winter feld, M.; Berg, E.; Hummel, M.; Timm, L.; Rau, B.; Walther, W.; Daum, S.; Kobelt, D.; Stein, U. S100A4 Is a Strong Negative Prognostic Marker and Potential Therapeutic Target in Adenocarcinoma of the Stomach and Esophagus. Cells 2022, 11 (6), 1056.
43. Huang, S.; Zheng, J.; Huang, Y.; Song, L.; Yin, Y.; Ou, D.; He, S. ; Chen, X.; Ouyang, X. Impact of S100A4 Expression on Clinicopathological Characteristics and Prognosis in Pancreatic Cancer: A Meta-Analysis. Disease Markers 2016, 2016, 1–9.
44. Deng, H.; Shi, J.; Wilkerson, M.; Meschter, S.; Dupree, W.; Lin, F. Usefulness of S100P in Diagnosis of Adenocarcinoma of Pancreas on Fine-Needle Aspiration Biopsy Specimens. American Journal of Clinical Pathology 2008, 129 (1), 81–88.
45. Chiba, M.; Imazu, H.; Kato, M.; Ikeda, K.; Arakawa, H.; Kato, T. ; Sumiyama, K.; Homma, S. Novel Quantitative Analysis of the S100P Protein Combined with Endoscopic Ultrasound-Guided Fiedle Aspiration Cytology in the Diagnosis of Pancreatic Neatic Adenocarcinoma. Oncology Reports 2017, 37 (4), 1943–1952.
46. Arumugam, T.; Simeone, D. M.; Van Golen, K.; Logsdon, C. D. S100P Promotes Pancreatic Cancer Growth, Survival, and Invasion. Clinical Cancer Research 2005, 11 (15), 5356–5364.
47. Arumugam, T.; Logsdon, C. D. S100P: A Novel Therapeutic Tar

DOI: 10.36838/v5i5.16

get for Cancer. Amino Acids 2010, 41 (4), 893–899.
48. Wu, Y.; Zhou, Q.; Guo, F.; Chen, M.; Tao, X.; Dong, D. S100 Proteins in Pancreatic Cancer: Current Knowledge and Future Perspectives. Frontiers in Oncology 2021, 11.
49. Camara, R.; Ogbeni, D.; Gerstmann, L.; Ostovar, M.; Hurer, E.; Scott, M.; Mahmoud, N. G.; Radon, T.; Crnogorac-Jurcevic, T.; Patel, P.; Mackenzie, L. S.; Chau, D. Y. S.; Kirton, S. B.; Rossiter, S. Discovery of Novel Small Molecule Inhibitors of S100P with In Vitro Anti-Metastatic Effects on Pancreatic Cancer Cells. European Journal of Medicinal Chemistry 2020, 203, 112621.
50. Wei, D.; Wang, L.; He, Y.; Xiong, H. Q.; Abbruzzese, J. L.; Xie, K. Celecoxib Inhibits Vascular Endothelial Growth Factor Expression in and Reduces Angiogenesis and Metastasis of Human Pancreatic Cancer via Suppression of SP1 Transcription Factor Activity. Cancer Research 2004, 64 (6), 2030–2038.
51. Hill, R.; Li, Y.; Tran, L. M.; Dry, S.; Calvopina, J. H.; Garcia, A.; Kim, C.; Wang, Y.; Donahue, T. R.; Herschman, H. R.; Wu, H. Cell Intrinsic Role of Cox-2 in Pancreatic Cancer Development. Molecular Cancer Therapeutics 2012, 11 (10), 2127–2137.
52. Matsubayashi, H.; Infante, J. R.; Winter, J. M.; Klein, A. P.; Schulick, R.; Hruban, R.; Visvanathan, K.; Visvanathan, K.; Goggins, M. Tumor Cox-2 Expression and Prognosis of Patients with Resectable Pancreatic Cancer. Cancer Biology & Therapy 2007, 6 (10), 1569–1575.
53. Markosyan, N.; Li, J.; Sun, Y. H.; Richman, L. P.; Lin, J. H.; Yan, F.; Quinones, L.; Sela, Y.; Yamazoe, T.; Gordon, N.; Tobias, J. W.; Byrne, K. T.; Rech, A. J.; FitzGerald, G. A.; Stanger, B. Z.; Vonderheide, R. H. Tumor Cell–Intrinsic EPHA2 Suppresses Antitumor Immunity by Regulating PTGS2 (COX-2). Journal of Clinical Investigation 2019, 129 (9), 3594–3609.
54. Yip-Schneider, M. T. Cyclooxygenase-2 Expression in Human Pancreatic Adenocarcinomas. Carcinogenesis 2000, 21 (2), 139–146.
55. Mao, F.; Li, D.; Xin, Z.; Du, Y.; Wang, X.; Xu, P.; Li, Z.; Qian, J.; Yao, J. High Expression of COL17A1 Predicts Poor Prognosis and Promotes the Tumor Progression via NF-KB Pathway in Pancreatic Adenocarcinoma. Journal of Oncology 2020, 1–12.
56. Wu, M.; Li, X.; Zhang, T.; Liu, Z.; Zhao, Y. Identification of a Nine-Gene Signature and Establishment of a PrognosticNomogram Predicting Overall Survival of Pancreatic Cancer. Frontiers in Oncology 2019, 9.
57. Zhang, Z.; Gu, J.; Yin, M.; Wang, D.; Ma, C.; Du, J.; Lin, Z.; Hu, S.; Wang, X.; Li, Y.; Tan, G.; Luo, H.; Wei, G. Establishment and Investigation of a Multiple Gene Expression Signature to Predict Long-Term Survival in Pancreatic Cancer. BioMed Research International 2020, 2020, 1–20.
58. Demirkol Canlı, S.; Dedeoğlu, E.; Akbar, M. W.; Küçükkaradumban, B.; İşbilen, M.; Erdoğan, Ö. Ş.; Erciyas, S. K.; Yazıcı, H.; Vural, B.; Güre, A. O. A Novel 20-Gene Prognostic Score in Pancreatic Adenocarcinoma. PLOS ONE 2020, 15 (4).
59. Chen, Y.; Xu, R.; Ruze, R.; Yang, J.; Wang, H.; Song, J.; You, L.: Wang, C.; Zhao, Y. Construction of a Prognostic Model with Histone Modification-Related Genes and Identification of Potential Drugs in Pancreatic Cancer. Cancer Cell International 2021, 21 (1).

Author

Minnie Liang is a senior at West Lafayette Jr/Sr High School in Indiana. She is passionate about computer science and biology, so she wants to pursue an MD-PhD in computational biology. By translating her findings into the clinic, she aims to advance medicine to create innovative diagnostics and therapeutics.