Prediction of pancreatic cancer risk pathways based on multiple omics

Hongyu Zhang[*]*, Xiao Liu, Haotian Guo*

Harbin Medical University, Heilongjiang, China

ABSTRACT

As a common and dangerous malignant tumor, the pathogenesis of pancreatic cancer is still not completely clear. This study aims to reveal the potential molecular mechanisms and therapeutic targets of pancreatic cancer by comprehensively utilizing the protein and gene expression data in TCGA database, combining survival information and protein interaction network analysis, and provide new theoretical support for clinical diagnosis and treatment.

First, we obtained protein and gene expression data from pancreatic cancer patients from the TCGA database, covering protein expression for 328 samples and gene expression information for 493 samples. By incorporating survival information, we obtained a complete expression profile, which lays the foundation for subsequent analysis. Second, 23 differential proteins were identified associated with pancreatic cancer survival that may play an important role in the development and progression of pancreatic cancer. Further, we constructed a protein interaction network for pancreatic cancer, and combined the String database and GN algorithm to identify the proteins with the most significant impact on pancreatic cancer and their associated primary and secondary proteins, providing important clues for further research. Subsequently, we analyzed the gene expression data using the Lasso regression model, identified 6 genes with significant differences (TMEM176A, ANLN, IGFBP2, MROH 9, OLFM 3, TRIM67), and drew their Lasso coefficient pathway maps and cross-validation curves. The random walk algorithm and the perturbation method were further used to identify the two most important pathway genes, SYNE 1 and STARD8, which provided a new perspective on the pathogenesis of pancreatic cancer. Finally, we input SYNE 1 and STARD8 into the KEGG database, and constructed a pathway network of pancreatic cancer, revealing the potential mechanism of action of these two genes in the occurrence and development of pancreatic cancer. These results provide an important theoretical basis for the early diagnosis of pancreatic cancer, the discovery of therapeutic targets, and the development of individualized treatment strategies, and are expected to bring about a significant improvement in the survival rate and quality of life of pancreatic cancer patients.

Keywords: Pancreatic Cancer; Cox Regression; Protein Interaction; Network Biological Pathway; Lasso Regression

1 INTRODUCTION

Pancreatic cancer is the seventh leading cause of cancer-related death in the world. It is also one of the cancer types with the highest mortality rate among all solid tumors. It has the title of "King of cancer" and seriously endangers human life and health. According to the latest data released by the National Cancer Center of the Chinese Academy of Medical Sciences, 134,374 new pancreatic cancer patients and 131,203 pancreatic cancer deaths are

expected in China in 2022, ranking the eighth in the total population incidence rate and the sixth in the mortality rate. Among them, pancreatic ductal adenocarcinoma (PDAC), which accounts for 90% of pancreatic cancer, has a very poor prognosis due to the difficult early detection, and the five-year survival rate of patients is less than 5-10%. About 80% of PDAC patients are already in the middle and late stage of the first diagnosis, and are almost "death" (that is, the presence of unresectable or metastatic lesions and the loss of radical surgery). However, if it can be timely detected and treated in phase I, the five-year survival rate will significantly increase to 73.3%, so it is particularly important to accurately screen for early pancreatic cancer detection methods.

The pathogenesis of pancreatic cancer is very complex, involving various genetic variations, abnormal signaling pathways and micro environment changes. By studying the interactions and regulatory mechanisms between these factors, a deeper understanding of the pathogenesis of pancreatic cancer can provide a theoretical basis for prevention and treatment. The differential protein-pathway interaction network constructed in this paper enhances the interactions between proteins and the associations between pathways. During the random walk, we consider the differential expression of genes, make up for the shortcomings of the traditional method, and effectively identify the risk pathways with significant relationship with the disease, which has important guidance for the mechanism of disease development in the future.

2 MODEL ESTABLISHMENT AND SOLUTION

2.1 A Cox regression analysis

The Cox regression model is a semi-parametric regression model. This model takes the survival outcome and survival time as the dependent variables, which can simultaneously analyze the influence of many factors on the survival time, can analyze the data with censored survival time, and does not require to estimate the survival distribution type of the data.

The main purpose of the survival analysis is to study the relationship between the variable X and the observation, namely the survival function S (t, X). When S (t, X) is affected by many factors, namely X= (X1,..., X n) as the vector, the traditional method is to consider the regression equation namely the variables Xi influence on S (t, X), and Cox proportional risk regression model, it is not directly investigate the relationship between S (t, X) and X, but with h (t, X) as the dependent variable, the basic form of the model is:

$$h(t,X) = h_0(t)exp(\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_m X_m) \quad (1)$$

2.2 GN algorithm

The GN algorithm is a classical community discovery algorithm, which belongs to the split hierarchical clustering algorithm. Initially, as proposed by Michelle Girvan and Mark Newman.

The basic idea is to constantly delete edges in the network with the largest number of edges relative to all source nodes, and then recompute the number of the remaining edges in the network relative to all source nodes, repeating the process until all edges are deleted in the network.

Modularity Q, also known as a modularity measure, is an indicator to evaluate the structural strength of a community. The larger the index result, the better the community division effect.

The calculation formula is as follows:

$$Q = \sum_i (e_{ii} - a_i^2) = Tre - ||e^2|| \quad (2)$$

2.3 random walk (RWR)

Based on network randomwalk (randomwalk, RW) algorithm used to represent an irregular form of change, some interested nodes in the network is given weights as a seed node, from the seed nodes to other neighbors along the network structure, at the same time in the process of wandering weight distribution of other nodes, eventually makes the closely associated with the seed node nodes tend to have higher weight. The restart-type random walk (random walk with restart, RWR) is used here, and the formula is defined as:

$$Pt + 1 = \quad (1 - r)\ Wpt + rp0 \quad (3)$$

Where r represents the probability that a node assigns its weights to its neighbor node in each walk, where the default value of 0.7 is used, W represents the standardized adjacency matrix of the network, p0 represents the initial weight vector of the node, and p t represents the new weight of the node after the network t walk.

The principle is shown as follows:

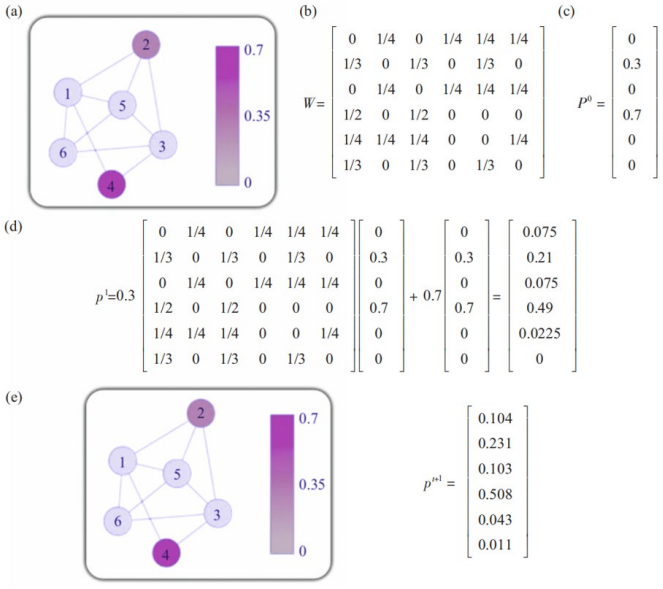

Fig. 1 Interpretation of the random walk principle

2.4 Perturbation and disturbance network evaluation model

We build a high degree in the protein-pathway interaction network, so we tend to get higher scores during the network random walk, which may lead to some bias in the identification results. In order to eliminate such bias, we do randomization disturbance,

disturb the distribution of seed nodes in the network, the network specific disturbance method is randomly selected from the real network and identify the number of disease differential expressed genes equal nodes as seed nodes, and then the fold change value of the differential genes randomly assigned to the seed nodes as weights, random walk, calculate the score of each pathway. According to the above method, after 1000 disturbances of the network, the significance of the candidate pathway, P=N/1000. For those pathway nodes with high degree in the protein-pathway interaction network, there are still a large number of closely connected nodes to assign weights, so that the corresponding P value will increase. Finally, passing the threshold P <0.01, we were able to correct the effect of pathway size on the outcome and identify pathways significantly associated with disease. Finally, P-values were used to evaluate the significant relationship between the risk pathway and the disease.

2.5 The Lasso regression analysis

The Lasso regression method is a commonly used statistical analysis tool with extensive applications in the field of data mining and machine learning. It performs feature selection by introducing the L1 regularization term, which can effectively screen the features that have a significant impact on the target variable in high-dimensional data to improve the predictive power and interpretability of the model.

The core idea of Lasso regression method is to introduce the L1 regularization term on the basis of the least squares method and solve the parameters of the model by minimizing the objective function. The L1 regularization term has sparsity and is able to compress a part of the coefficients to zero for feature selection. Compared with the ridge regression method, the Lasso regression method enables more accurate feature selection, which is applicable to the problems with explanatory requirements on the model.

In the medical field, the Lasso regression method can be applied to disease diagnosis, biomarker discovery, etc.

3 CONCLUSION

Expression profile synthesis

Protein data: the expression profile synthesized by the data downloaded by TCGA, including the sample number and the measured protein expression amount of each sample, and then combined with the survival information (including survival status and survival time) to obtain the expression profile required for the final experiment.

Gene data: the expression profile synthesized by the data downloaded from TCGA, including the sample number and the amount of measured gene expression of each sample, and then combined with the survival information (including survival status and survival time) to obtain the expression profile required for the final experiment.

Cox regression analysis identified the differential proteins

We used expression profiles for Cox regression, set p 0.05 representative with significant difference, and finally selected the differential proteins of pancreatic cancer, resulting in the following 23 proteins.

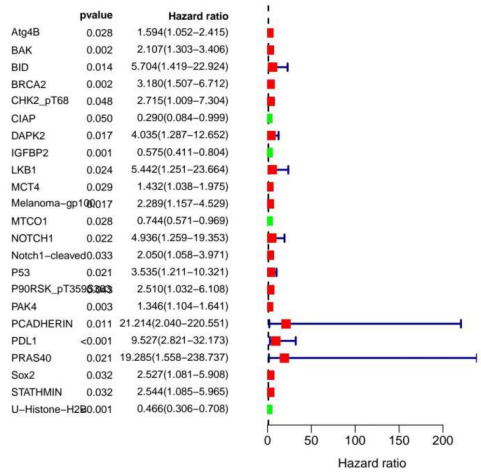

Fig. 2 Differential protein recognition results in pancreatic cancer

Construction of a protein interaction network (ppi)

We constructed the pancreatic cancer protein interaction network based on the related proteins obtained by COX, combined the protein interaction pairs of String database and applied the GN algorithm to get the largest influence on pancreatic cancer and the associated primary and secondary proteins.

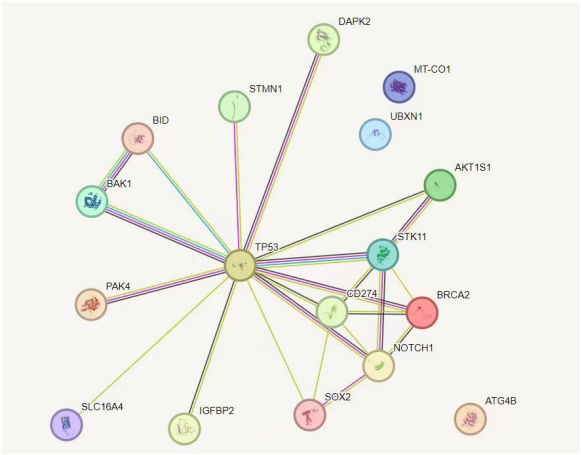

Fig. 3 Protein interaction network for pancreatic cancer

LASSO regression to identify specific genes

For the gene-survival information combined expression profile of TCGA data, we used the Lasso regression model to obtain the Lasso coefficient path diagram of 6 specific genes (TMEM176A, ANLN, IGFBP2, MROH 9, OLFM 3, TRIM67) and their cross-validation curves,

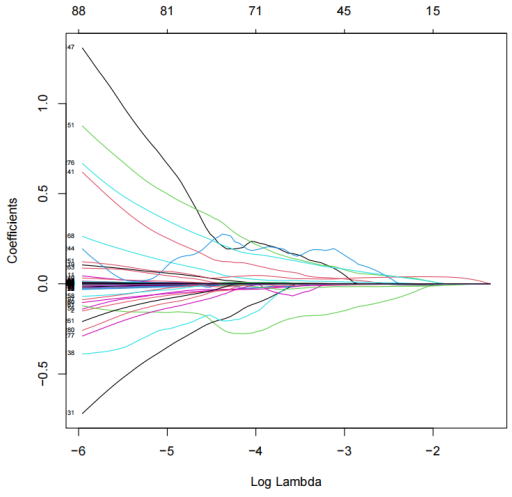

Fig. 4 Path of specific gene coefficients

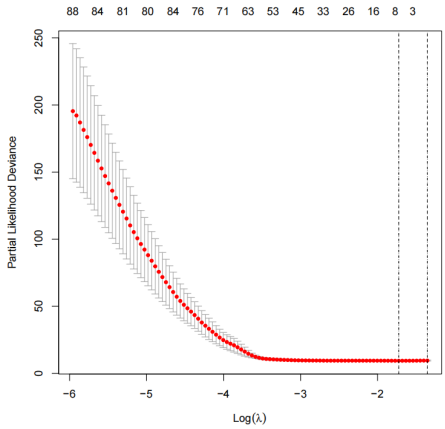

Fig. 5 Cross-validation curve

Random walk and perturbation method to find gene targets

The differential genes analyzed by our Lasso regression model are input to the random walk algorithm, divide particle populations, and then apply perturbation to score, the two most important pathway genes are SYNE 1 and STARD8.

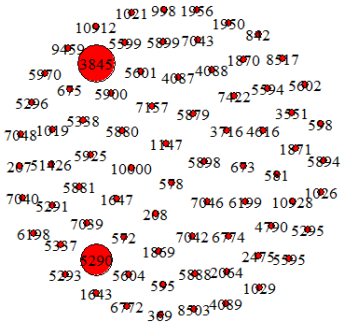

Fig. 6 Pathway gene identification results

Construction of the gene pathways

In the previous step, we concluded that the two most important pathway genes in pancreatic cancer are SYNE 1 and STARD8, which are now entered into the KEGG database to read the pathway information and construct the network.

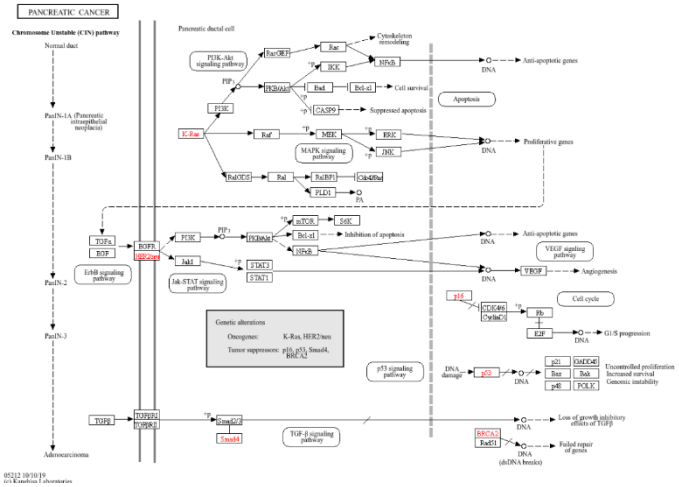

Fig. 7 Pancreatic cancer pathway network construction

4 DISCUSSION

Pancreatic cancer is a highly lethal malignancy whose pathogenesis involves the dysregulation of multiple genes and pathways. Through high-throughput sequencing technology, the genetic information of gene expression, mutation and copy number variation in pancreatic cancer tissues, which reveals the molecular mechanism of the occurrence and development of pancreatic cancer and provides an important basis for individualized treatment.

In performing gene full pathway testing in pancreatic cancer, we used several methods to analyze and interpret the data. We used Cox regression analysis to determine gene variation, pathway activity and other factors related to patient survival, use random walk model to explore the interaction network between genes, identify key signaling pathways and regulators, and provide theoretical basis for the analysis of disease pathogenesis. We applied the GN algorithm to discover potential patterns and subtypes in gene expression data, thereby identifying molecular features and therapeutic targets for different subtypes. Lasso regression was used to screen out key genes and pathways associated with tumor development and development, reduce data dimension and noise interference, and improve the predictive performance and interpretability of the model. Using perturbation algorithm to simulate changes and noise in gene expression data, evaluate the stability of the analysis results, and optimize the data processing and modeling strategies.

In the future, there are still many developments and challenges for gene full pathway detection in the field of pancreatic cancer research. With the development of single-cell sequencing technology, the whole-gene pathway detection will be more refined, which can reveal the genetic heterogeneity between different subtypes and provide a more accurate basis for the individualized treatment of pancreatic cancer. Comprehensive analysis of multi-omics data will develop as a trend in the future. Combining genomic data, transcriptome and proteome, we can provide a more comprehensive understanding of the pathogenesis of pancreatic cancer and discover new therapeutic targets and biomarkers. In addition, the application of artificial intelligence and machine learning algorithms will accelerate the analysis and interpretation of the whole-pathway detection results of genes. By establishing an accurate prediction model, the personalized treatment plan for pancreatic cancer patients can be realized, and the treatment effect and survival rate can be improved.

In conclusion, gene whole-wide pathway detection is an important tool in the field of pancreatic cancer research, which is important in revealing disease mechanisms, guiding treatment and prognosis evaluation. The future development will pay more attention to the comprehensive analysis of multi-omics data and the realization of individualized treatment strategies, and bring new breakthroughs and opportunities for the precision medicine of pancreatic cancer.

5 DATA SOURCES

We downloaded the protein expression data and survival information of pancreatic cancer disease from TCGA database, which determined the expression level of 230 proteins from 328 samples and synthesized an expression profile; we downloaded the gene expression data and survival information of pancreatic cancer disease from TCGA database, which determined the expression level of 1626 genes in 493 samples and synthesized an expression profile.

After the differential proteins were extracted by using the TCGA data expression profiles, the differential proteins were input from the String database to obtain the interacting primary and secondary proteins.

We downloaded pathway data for pancreatic cancer from the KEGG database and stored basic pathway information in the database.

6 ACKNOWLEDGEMENTS

During the completion of this study, we express our utmost gratitude to those who have always supported and encouraged us.

First of all, we would like to sincerely thank every member of our team. The hard work, expertise and cooperative spirit of everyone have made an indispensable contribution to the smooth progress of this study. The spirit of exploration and enthusiasm for knowledge not only inspire the growth of our team, but also provide a solid foundation for the completion of this article.

In particular, we should express our sincere gratitude to our instructor Lu Pengju. Mr.Lu gave us great guidance and help in the topic selection, conception and writing of this paper, and provided many valuable suggestions in the process of revising the paper. His repeated careful review and patient guidance enabled our paper to continuously improve and finalized the final draft. Mr.Lus careful guidance made every member of our team benefit a lot from the research, which further strengthened our research direction and goals.

Finally, we would like to thank all individuals and organizations who provided support and assistance for this study. Their encouragement, support and contribution provide an important guarantee for the smooth progress of our research, and without their support, we will not be able to complete this research work.

We sincerely thank you for your support and encouragement, and thank you for your company and support on our research road.

REFERENCES

[1] Fraunhoffer, N. A., Abuelafia, A. M., Bigonnet, M., Gayet, O., Roques, J., Nicolle, R., ... & Iovanna, J. (2022). Multi-omics data integration and modeling unravels new mechanisms for pancreatic cancer and improves prognostic prediction. *NPJ precision oncology*, *6*(1), 57.

[2] Su, Y., Wang, F., Lei, Z., Li, J., Ma, M., Yan, Y., ... & Hu, T. (2023). An Integrated Multi-Omics Analysis Identifying Immune Subtypes of Pancreatic Cancer. *International Journal of Molecular Sciences*, *25*(1), 142.

[3] Turanli, B., Yildirim, E., Gulfidan, G., Arga, K. Y., & Sinha, R. (2021). Current state of "omics" biomarkers in pancreatic cancer. *Journal of Personalized Medicine*, *11*(2), 127.

[4] Nicoletti, A., Paratore, M., Vitale, F., Negri, M., Quero, G., Esposto, G., ... & Zileri Dal Verme, L. (2024). Understanding the conundrum of pancreatic cancer in the omics sciences era. *International Journal of Molecular Sciences*, *25*(14), 7623.

[5] Xu, D., Wang, Y., Liu, X., Zhou, K., Wu, J., Chen, J., ... & Zheng, J. (2021). Development and clinical validation of a novel 9-gene prognostic model based on multi-omics in pancreatic adenocarcinoma. *Pharmacological Research*, *164*, 105370.

[6] Bagante, F., Spolverato, G., Ruzzenente, A., Luchini, C., Tsilimigras, D. I., Campagnaro, T., ... & Pawlik, T. M. (2021). Artificial neural networks for multi-omics classifications of hepato-pancreato-biliary cancers: towards the clinical application of genetic data. *European Journal of Cancer*, *148*, 348-358.

[7] Gao, Y., Zhang, E., Fei, X., Kong, L., Liu, P., & Tan, X. (2021). Identification of novel metabolism-associated subtypes for pancreatic cancer to establish an eighteen-gene risk

prediction model. *Frontiers in Cell and Developmental Biology*, *9*, 691161.

[8] Wang, L., Liu, Z., Zhu, R., Liang, R., Wang, W., Li, J., ... & Sun, Y. (2022). Multi-omics landscape and clinical significance of a SMAD4-driven immune signature: Implications for risk stratification and frontline therapies in pancreatic cancer. *Computational and Structural Biotechnology Journal*, *20*, 1154-1167.

[9] Pettini, F., Visibelli, A., Cicaloni, V., Iovinelli, D., & Spiga, O. (2021). Multi-omics model applied to cancer genetics. *International journal of molecular sciences*, *22*(11), 5751.

[10] Connor, A. A., & Gallinger, S. (2022). Pancreatic cancer evolution and heterogeneity: integrating omics and clinical data. *Nature Reviews Cancer*, *22*(3), 131-142.