Molecular signatures of tumor progression in pancreatic adenocarcinoma identified by energy metabolism characteristics

Xin Wang

Department of Pathology, Fudan University Shanghai Cancer Center

Cong Tan

Department of Pathology, Fudan University Shanghai Cancer Center

Weiwei Weng

Department of Pathology, Fudan University Shanghai Cancer Center

Shujuan Ni

Department of Pathology, Fudan University Shanghai Cancer Center

Meng Zhang

Department of Pathology, Fudan University Shanghai Cancer Center

Hesheng Jiang

Department of Oncology, Shanghai Medical College, Fudan University

Lei Wang

Department of Pathology, Fudan University Shanghai Cancer Center

Dan Huang

Department of Pathology, Fudan University Shanghai Cancer Center

Weiqi Sheng

Department of Pathology, Fudan University Shanghai Cancer Center

Midie Xu (✉ xumd27202003@sina.com)

Department of Pathology, Fudan University Shanghai Cancer Center

Research Article

Keywords: Pancreatic adenocarcinoma, molecular subtype, energy-metabolism-related genes, prognosis signature

Posted Date: May 20th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-478202/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background: In this study, we aimed to describe a molecular evaluation of primary pancreatic adenocarcinoma (PAAD) based on comprehensive analysis of energy-metabolism-related gene (EMRG) expression profiles.

Methods: Molecular subtypes were identified by non-negative matrix clustering algorithm clustering on 565 EMRGs. The overall survival (OS) predictive gene signature was developed, internally and externally validated based on three online PAAD datasets. Hub genes were identified in molecular subtypes by weighted gene correlation network analysis (WGCNA) co-expression algorithm analysis, and then enrolled for determination of prognostic genes. Univariate, LASSO and multivariate Cox regression analyses were performed to assess prognostic genes and construct the prognostic gene signature. Time-dependent receiver operating characteristic (ROC) curve, Kaplan-Meier curve and nomogram were used to assess the performance of the gene signature.

Results: On the basis of EMRGs expression profile, we propose a molecular classification dividing PAAD into two subtypes: Cluster 1, which display more immune and stromal cell components in tumor microenvironment and higher tumor purity; and Cluster 2, which display worse OS. Moreover, by using a three-phase training, test and validation process, we construct a 4-gene signature that can constantly classify the prognostic risk of patients in all three datasets, and which present higher robustness and clinical usability compared with four previous reported prognostic gene signatures. In addition, a novel nomogram constructed by combining clinical features and the 4-gene signature showed confident clinical utility in PAAD. According to gene set enrichment analysis (GSEA), gene sets related to the high-risk group were participated in the neuroactive ligand receptor interaction pathway.

Conclusions: In summary, the EMRG-based molecular subtypes and prognostic gene model provides a roadmap for patient stratification and trials of targeted therapies.

Introduction

Pancreatic adenocarcinoma (PAAD) is one of the most lethal malignancies, which caused 459,000 incidence and 432,000 death in the world, according to GLOBOCAN 2018(1). Our current understanding of the complicated genetic, epigenetic alteration and its interaction with microenvironment has not resulted in a leap in patients' survival(2). Challenges including exploration the pathogenesis and progression of disease, identification the early detection and risk evaluation biomarker which led to diverse treatment option still need a large amount of efforts.

The reprogramming of cellular metabolism plays an indispensable role as both direct and indirect outcomes of oncogenic alteration in tumorigenesis. It enables tumor cells to produce ATP to maintain the reduction-oxidation balance and macromolecular biosynthesis processes required for cell growth, proliferation, and migration. For a long time, it has been believed that malignancies mainly restrict their energy metabolism to glycolysis, even in the presence of oxygen, which is called the Warburg effect (3).

However, increasingly researches acknowledged the heterogeneous metabolic phenotype of cancer cells(4). Although several highly conserved pathways including Kras, p53, c-Myc, and Lkb1 signaling are pivotal for maintaining uncontrolled proliferation of cancer cells and thus involved in tumorigenesis and cancer development(5), almost nothing is known about the precise role and underlying mechanism of energy-metabolism-related gene (EMRG) and their gene expression profiles in primary PAAD, not yet anything known related to the prognostic distinctions of PAAD. A deep understanding of EMRGs in tumors might provide an important step for the development of new therapies.

In this study, we constructed an energy metabolism molecular subtypes of PAAD by using expression data of EMRGs from public databases, including TCGA, GEO, and ICGC. Furthermore, we assess its relationship with prognosis and identification differences on clinical and immune characteristics. The prognostic risk model constructed by differentially expressed genes between PAAD molecular subtypes can better evaluate the prognosis of PAAD samples. We further use the gene expression data sets from GEO and ICGC database to verify the performance of the prognostic risk model.

Materials And Methods

1. Data collection and processing

Raw gene expression data and corresponding clinical information of patients with PAAD were obtained from the Cancer Genome Atlas website (TCGA), Gene Expression Omnibus (GEO), and the International Cancer Genome Consortium (ICGC). The RNA-seq expression data, RNA-seq count data and clinical follow-up information of 177 patients diagnosed as PAAD was downloaded through the TCGA GDC API, among them 171 patients (90%) were randomly selected as the training set for model construction (**Table 1**). Subsequently, in order to verify the robustness of the model over different sequencing platforms, all PAAD samples in *TCGA* database were used as internal verification set. Further, A GEO dataset, GSE57495, containing transcriptome and clinical data of 63 patients, and a series of RNA-seq profile of 269 samples obtained from the ICGC database, were downloaded as validation datasets (**Table 1**). Eleven annotated metabolism related pathways referred to the Molecular Signature Database v7.0 (MSigDB), which including 594 EMRGs, were downloaded from Reactome database (<https://reactome.org/>, **Supplementary Table 1**). We matched the candidate gene with the TCGA transcriptome matrix and retained genes with detectable signals of more than half of the tissues, and finally obtained 565 genes for subsequent analysis. The workflow was exhibited in **Supplementary Figure 1**. Patient informed consent was existing in these three public datasets, and this study was approved by the institutional review board (IRB) of Fudan University Shanghai Cancer Center (FUSCC) and conducted in accordance with the Helsinki Declaration.

2. Identification of energy metabolism molecular subtypes

Among all TCGA and ICGC PAAD samples, 565 EMRGs were extracted. The non-negative matrix factorization (NMF) (6) was utilized to cluster all PAAD samples, and the optimal numbers of cluster were

determined according to indicators including cophenetic correlation(6), silhouette coefficient(7), and residual sum of squares (RSS) (8)

3. Analysis of immune scores between molecular subtypes

The fragments per kilobase of exon model per million mapped reads (FPKM) data of gene in *TCGA* PAAD dataset was input into the TIMER (tumor immune estimation resource) tool (9) and the R software package estimate for calculation of the immune score. Next, difference on the Immune Score and Stromal Score, which represent the relative proportion of immune cells and stromal cells in tumor tissues, was calculated by using R package Estimation of Stromal and Immune cells in Malignant Tumors using Expression data (ESTIMATE)(10). The Estimate Score, which refers to the purity of tumor tissues, is the sum of Immune Score and Stromal Score. Then the differences in the immune score of the samples between the two subtypes were compared.

4. Identify differentially co-expressed genes between molecular subtypes

To identify the differential co-expression genes between each subtype, the R software package DESeq2 was used to calculate the differentially expressed genes (DEGs) between the two subtypes, and the thresholds were set to $FDR < 0.05$ and $|\log_2FC| > 1$. Weighted gene correlation network analysis (WGCNA) co-expression algorithm was used for detecting co-expressed genes and modules by the R package WGCNA(11). To improve the accuracy of network construction, TPM profile of genes were input into hierarchical cluster analysis to remove the outlier samples. Second, distance between each gene were calculated using *Pearson* correlation coefficient; a weighted co-expression network was constructed using the R package WGCNA, and co-expression modules were screened by set the soft threshold power β as 10. Third, the topology overlap matrix (TOM) was then constructed from the adjacency matrix to avoid the influence of noise and spurious associations. On the basis of TOM, average-linkage hierarchical clustering using the dynamic shear tree method was subsequently conducted to define co-expression modules, and the minimum gene size of each module was set as 30. The feature vector values (eigengenes) of each module was calculated in turn to explore the relationship among modules, and then modules with highly correlated eigengenes were merged into new module by performed cluster analysis with threshold as follows: height = 0.25, DeepSplit = 2, and minModuleSize = 30. In order to identify the modules of interest, the correlation between each co-expression module and patients' clinical features as well as cluster subtypes was further evaluated. Modules with significant correlation with the energy-metabolism subtypes were defined as key modules for the subsequent selection of hub genes (Spearman correlation coefficient > 0.4 , $P < 0.05$). Finally, pathway enrichment analysis of differentially co-expressed genes was performed through the R package WebGestaltR (the threshold $FDR < 0.05$).

5. Establishment of prognosis prediction modal

The R package survival coxph function was used for analysis of univariate Cox proportional hazards regression model, and log rank $p < 0.01$ was selected as the threshold. To narrow the gene range and maximize the accuracy, Least Absolute Shrinkage and Selection Operator (LASSO) Cox regression

analysis (12), a method screening signatures with generally effective prognostication performance by performing automatic feature selection, was performed by using the glmnet package of R to identify the prognostic gene. And optimal genes were evaluated by 10-fold cross validation. And genes obtained by LASSO analysis were subjected to multivariate Cox survival analysis to construct a final prognostic risk model. Time-dependent Receiver operating characteristic (ROC) curve analysis was conducted to assess the prognostic value of the identified model using the R package timeROC (13). The risk scores of patients in the internal verification set and the external verification set were analyzed using the same model coefficients as the training set for verifying the robustness of the gene signature. Kaplan-Meier curve was used to evaluate the difference in survival time between groups, and then univariate and multivariate Cox regression analysis was performed to evaluate independent prognostic factors. P value < 0.05 was considered significant statistically. Decision curve analysis (DCA), which can evaluate predictive models from the perspective of clinical consequences (14), was performed in the entire cohort to test the clinical usefulness of the nomogram in comparison with the gene signature and clinicopathological parameters. Restricted mean survival time (RMST) curve was drawn to construct the comparison with R package rms.

6. Bioinformatic analysis

The data processing and symbol remapping was conducted using the R-4.0.1 software. P value < 0.05 was considered significant statistically. Single-sample Gene set enrichment analysis (ssGSEA) was applied for identifying relationship between the risk scores of different samples and biological functions using the R package GSVA. The classical gene sets of Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways (c2.cp.kegg.v7.0.symbols) were considered to decipher the phenotype. For each analytical pathway, the enrichment score (ES) and the significance of ES were calculated, and the normalized enrichment score (NES) and false discovery rate (FDR) were further calculated to examine functional enrichment results.

/

Result

1. Construction of energy metabolism related molecular subtypes

By using the NMF analysis based on the expression of the 565 EMRGs (**Supplementary Figure 2A**), we identified two distinct subtypes (Cluster1[n=74], Cluster2[n=97]) between the 171 patients in the *TCGA* PAAD dataset (**Figure 1A-B**). Clinically, patients in Cluster 1 showed significant higher tumor grade than that in Cluster 2 (**Supplementary Figure 2C**). Moreover, we assessed the potential difference on prognosis between the two subtypes, which demonstrated that patients in Cluster 1 had significant better OS than that in Cluster 2 (p = 0.017, HR = 0.597 95%CI 0.383-0.914, **Figure 1B**). Similarity, the expression profiles of these 565 EMRGs can also divided 257 patients into two molecular subtypes in the ICGC PAAD dataset (**Figure 1C-D, Supplementary Figure 2B**), and also patients in Cluster 1 showed significant better OS than

that in Cluster 2 ($p = 0.003$ HR = 0.610 95%CI 0.441-0.844, **Figure 1D**). These data showed the consistent existing of the molecular types in PAAD.

Then, we calculate the immune scores of six cells (B cell, CD4 T cell, CD8 T cell, Neutrophil, Macrophage, and Dendritic) in each PAAD sample and analyzed the potential difference between Cluster 1 and Cluster 2. The results showed that except for the B cell immune score, Cluster 1 showed higher immune score than Cluster 2 (**Figure 1E**). We further observed that the scores of immunity, stroma and tumor purity in Cluster 1 were also significantly higher than that in Cluster 2 (**Figure 1F**). These results indicate that lower immune cells infiltration in tumor environments (TME) may confers worse prognosis in patients with PAAD.

2. Identification of differential co-expression genes between subtypes

We extracted the expression profile of protein coding genes from *TCGA* PAAD dataset, and clustered all samples through hierarchical clustering (**Supplementary Figure 3A**), from which we confirmed that there is no outlier sample. To ensure that the network constructed by WGCNA is scale-free, β was set as 10 (**Supplementary Figure 3B**). Then we run cluster analysis and obtained 14 modules, among which grey module represent gene sets that cannot be aggregated to other modules (**Figure 2A**). Moreover, by analyzing the correlation of the module and genes in the module with phenotypes (**Supplementary Table 2**), we found that the blue module (contains 1692 co-expression genes) is significant correlated with Cluster 1, and the yellow module (contains 645 co-expression genes) is significant correlated with Cluster 2, respectively (**Figure 2B-D**). In addition, by analyzing the differential expression genes (DEGs) between Cluster 1 and Cluster 2, we obtained 2411 DEGs, comprised of 1641 up-regulated DEGs and 770 down-regulated DEGs between the two subtypes (**Figure 2E-F**, **Supplementary Table 3**). We further analyzed these 2411 DEGs and those co-expression genes in blue and yellow modules, and obtained 743 overlapping genes (**Supplementary Table 4**). These 743 co-expression DEGs were analyzed by GO function and KEGG pathway enrichment (**Supplementary Table 5**), and 38 KEGG pathways, 52 GO cellular component (CC), 126 GO molecular function (MF) and 977 GO biological process (BP) were enriched. The top enriched pathways include cell adhesion molecules (CAMs), transcriptional mis-regulation in cancer, immunological synapse and T cell differentiation (**Supplementary Figure 3C-F**), suggesting that these co-expression DEGs may involve in PAAD molecular regulative network by exhibiting pivotal function through these pathways.

3. Development of prognostic risk modal based on co-expression DEGs

By analyzed the expression profiles of 743 co-expression DEGs and corresponding survival of training set for using univariate Cox proportional hazard regression model, we obtained sixty-seven prognostic co-expression DEGs ($P < 0.01$, **Supplementary Table 6**). After Lasso Cox regression analysis and 10-fold cross validation, we selected four genes ($\lambda = 0.1042$) as the candidate genes for construction of prognostic risk modal (**Supplementary Figure 4A-B**). And we then established a gene-based prognostic model by using univariate Cox regression analysis (**Table 2**). The high expression level of GJB5, MET and TMEM139 were identified as risk factors, while AFF3 as protective factors. The final 4-gene signature

formula is as follows: $\text{RiskScore} = -0.1513 \times \exp^{\text{AFF3}} + 0.0156 \times \exp^{\text{GJB5}} + 0.0045 \times \exp^{\text{MET}} + 0.0164 \times \exp^{\text{TMEM139}}$.

We calculated the risk score of each sample according to the established model, and plotted the risk score distribution, which showed that the survival time of the samples with high risk score is significantly shorter than that with low risk score (**Figure 3A**). In addition, the AUCs of 1-, 3-, and 5-year ROC curves for the 4-gene signature to predict PAAD survival were all above 0.70 (**Figure 3B**). Finally, we carry out Z-score normalized on RiskScore, which classified samples with risk score greater than zero into high-risk group, and samples with risk score less than zero into low-risk group. Kaplan-Meier survival analysis demonstrated there were significant differences between the high- and low- risk group (log rank $P < 0.001$, HR = 2.413, **Figure 3C**).

4. Internal and external validation of the prognostic risk model

In order to determine the robustness of the model, we submit patients in the entire TCGA dataset into the formula aforementioned. The risk score distribution of all samples (**Figure 4A**), corresponding ROC curve (**Figure 4B**), and Kaplan-Meier survival curves (**Figure 4C**) showed that the AUCs of the signature remained high, and the high-risk groups had consistently shorter OS than the low-risk groups.

We further verified the robustness of the 4-gene prognosis signature by external analyzed in the GSE57495 dataset (**Figure 5A-C**) and ICGC PAAD dataset (**Figure 5D-F**) using the same coefficients aforementioned. Excellent performance was overserved in the prognostic risk indication.

5. Independence of the 4-gene prognosis signature

In order to identify the independence of 4-gene signature in clinical application, we conducted univariate and multivariate Cox regression in *TCGA* PAAD dataset. We systematically analyzed the clinical data of patients, including age, gender, pathologic T stage, pathologic N stage, pathologic M stage, tumor stage, tumor grade, and the 4-gene signature. Univariate Cox regression analysis showed that gender, tumor grade, pathologic T stage, pathologic M stage, tumor stage and the 4-gene signature were significantly associated with survival ($P < 0.05$, **Figure 6A**). However, multivariate Cox regression analysis showed that age, pathologic N stage and the 4-gene signature (**Figure 6B**) were independent prognostic indicators in PAAD. The above conditions indicated that the 4-gene signature has good predictive performance in clinical application.

Furthermore, we combined clinical features and the 4-gene signature and constructed a nomogram using the entire TCGA PAAD dataset (**Figure 6C**). Nomogram suggested the 4-gene signature has the greatest impact on the survival rate prediction. We calibrated the performance of 1, 2, and 3-year nomography data for visualization of nomogram, which further verified the consistency between predicted and actual survival probability (**Figure 6D**).

6. Comparison with previous prognostic models.

Previous studies had identified several prognostic models for survival prediction of PAAD patients. The predictive performance of the present 4-gene signature was further compared with four previous models (a 15-gene signature proposed by Chen et al.(15), a 7-gene signature proposed by Cheng et al. (16), a 5-gene signature proposed by Raman et al. (17), and a 7-gene signature proposed by Magouliotis et al. (18). We calculated the risk score of each PAAD sample in *TCGA* PAAD dataset based on the corresponding coefficients provided by each model, evaluated the ROC of each model, and divided the samples into high-risk and low-risk groups based on the median risk score of each signature. All of the four models could divide the patients into high-risk group and low-risk group (**Supplementary Figure 5**). Kaplan-Meier curves showed that except for Li model ($P = 0.076$), there are significant differences between high-risk group and low-risk group in Chen model, Cheng model and Raman model ($P < 0.05$, **Supplementary Figure 5A-D**). Among the four models, the AUC of Chen model and Raman model are greater than 0.70, but generally the prediction effect of the four models is worse than that of our four gene models (**Supplementary Figure 5E-H**). Furthermore, RMST curve (**Figure 7A**) and DCA curve (**Figure 7B**) were used to evaluate the predictive effect of our 4-gene signature and the four published models on the prognosis of PAAD patients, both demonstrated that the performance of our four gene models was significantly better than those of the four models.

7. GSEA analysis of enriched pathway based on risk score

To investigate the relationship between the risk score and biological function of different samples, we conducted single sample GSEA (ssGSEA) analysis, and calculated ssGSEA score of each sample on different biological functions. The correlation between these functions and RiskScore with the coefficient cutoff of 0.4 showed that most of the functional pathways were negatively correlated with the RiskScore of samples (**Figure 8A**). Moreover, we divided the training set into high-risk group and low-risk group referring to the RiskScore. GSEA was used to analyze the significantly enriched pathways in the two groups (**Supplementary Table 7**). The result showed that pathways including bladder cancer, pentose phosphate pathway, p53 signaling pathway and thyroid cancer were significant negatively correlated with the low-risk group, whilst neuroactive ligand receptor interaction pathway was negatively correlated with the high-risk group ($P < 0.01$, **Figure 8B**).

Discussion

Cumulative evidence has revealed that metabolic reprogramming in cancer was extensive linked to oncogenesis and immune disorder(19, 20). In PAAD, previous studies suggested that the metabolic alteration in PAAD was typically characterized by over-expression of glycolytic enzymes and lactate dehydrogenase for glucose, amino acids, and lipids metabolism (21, 22). Moreover, there is complex crosstalk among these reprogrammed metabolism approach within the tumor microenvironment, which contribute to the extraordinary growth advantages of tumor cells and unlimited development of PAAD(22).

The detection of aberrant metabolomics also contributes to the identification of novel biomarkers for diagnosis and prognostication, and the discovery of potential therapeutic targets for PAAD. For example, there are significant differences in metabolic profiles not only between PAAD patients and normal controls but also among different pathological PAAD subtypes(23, 24), and the metabolic alterations have helped identify several promising metabolomics-based diagnostic biomarker such as single serum metabolite (25), or even metabolomics-based biomarker signature in blood(26). Oliver F. Bathe et al. proposed the potential utility of serum metabolomic profile in discriminate PAAD patients from healthy controls(27). PAAD patients with higher levels of PE in the serum exosome might have a worse prognosis according to a population-based study(28). Taken together, the distinct characteristics of energy metabolism in PAAD are worth exploring and may shed new light on development of novel biomarkers related to metabolism. However, the accurate detection of metabolites in biological samples is still hampered by some technical defects such as lack of optimized study methods, limited coverage in metabolomics fingerprints and interference caused by unwanted sources(29). Moreover, the abundance of some metabolites can be quite low even less than the detection limit (30). Gene expression profiling, with the advantage of being convenient and precise, can give a whole picture of tumor properties based on quantitative data(31). By analyzing the expression levels of EMRGs in PAAD tumor tissue, the metabolic characteristics of PAAD can be comprehensively interpreted from another dimension.

In the present study, a total of 565 EMRGs were selected from Reactome database. These genes mainly participant in the key pathways of carbohydrate, fatty Acid, and glycogen metabolism. Based on the expression data of the *TCGA*-PAAD dataset, pancreatic cancer patients were divided into two metabolic subtypes using the NMF algorithm. Significant difference was observed in patients' immune cell infiltration and survival status between the two subtypes. Moreover, the proportion of almost all the immune cells and the fraction of immune components were significantly higher in the subtype with significantly better clinical outcomes, which strongly indicated the close relationship between tumor energy metabolism and immunology in PAAD. Previous evidence has proved that metabolic interventions can impact the immune functions of immune cells upon activation (32, 33). This phenomenon revealed the potential influence of the cross-talk between energy metabolism and immune microenvironment on the development and long-term survival of PAAD.

In order to select the hub genes that may significantly modulate cancer metabolism in PAAD, WGCNA co-expression analysis was firstly conducted and a total of 743 genes that strongly correlated with the two metabolic subtypes and differentially expressed between the two subtypes were screened out and considered as candidates for the construction of prognostic model. Using the Lasso regression analysis, a four-gene (AFF3, GJB5, MET and TMEM139) signature was identified after the verification of the training, internal validation sets, and external validation sets which included a total of 491 patients from the *TCGA*, ICGC and GEO PAAD datasets. The model interpreted the information of gene expression into risk score for the accurate estimation of prognosis in PAAD. Notably, the 3-year AUCs for the signature in all datasets were solid (higher than 0.70). When clinicopathologic parameters were taken into consideration, the constructed risk-score system could still independently predict the prognosis of PAAD

patients. A nomogram integrating the calculated risk score and clinical information constructed for the accurate prediction of survival probability of PAAD patients also showed confident clinical utility in PAAD.

Among the four genes, GJB5, MET and TMEM139 were risk factors while AFF3 was a protective factor for clinical outcomes in PAAD. The prognostic value of MET in PAAD have been reported in previous studies(34, 35), MET is a well-recognized regulator in the progression of PAAD, and MET inhibitor has shown promising results in preclinical studies (36, 37); while the risky or protective value of the other three genes in PAAD was rarely identified. Functional enrichment analysis revealed that this metabolism-related signature was significantly involved in some classical cancer-related pathways. The interaction between the four genes and tumor metabolism and progression in PAAD deserves further investigation.

Several previous studies have also identified specific prognostic models for the risk prediction of PAAD. For example, Chen et al. proposed a 15-gene signature which contained C6orf15, CAPN8, HIST1H3H, IGF2BP3, KIF14, KRT6A, PMAIP1, PPBP, RTKN2, SCEL, SERPINB5, SLC2A1, SLC45A3, TMPRSS3 and UCA1 (15). Cheng et al. identified a biomarker consisting of 7 genes such as SCEL, SLC2A1 and SERPINB5, et al, and which shared three same genes with Chen's gene-signature (16). Raman et al. discovered another 5-gene signature by enrolled the gene expression level of ADM, ASPM, DCBLD2, E2F7, and KRT6A (17), that was totally different from the previous two module. Magouliotis et al. discovered another gene signature containing 3 protein-coding RNA and 4 microRNAs that was totally different from that of Jiang et al(18).The prognostic performance of the present model was further compared with that of the four previous models. Among the four different signatures, our four-gene biomarker had the highest AUC and C-index values. It could be concluded that this EMRGs outperforms some previous biomarkers in the survival prediction of PAAD patients, and has great potential to be used in clinical application in the future.

However, there are still some limitations of this study. For example, the analysis was based on just retrospective data and needs to be verified in a prospective cohort containing samples from multi-centers before clinical application. Deeper mechanism research was also in need to elucidate the exact functions of the identified signature in PAAD.

Conclusion

In summary, by analyzing the expression levels of EMRGs in PAAD tumor tissues, two different clusters with varied overall survival and immune status were identified in the TCGA-PAAD dataset. A 4-gene prognostic signature containing and a novel nomogram was identified for the accurate risk prediction of PAAD patients.

Abbreviations

PAAD: primary pancreatic adenocarcinoma

EMRG: energy-metabolism-related gene

OS: survival

WGCNA: weighted gene correlation network analysis

ROC: receiver operating characteristic

GSEA: gene set enrichment analysis

TCGA: the Cancer Genome Atlas

GEO: Gene Expression Omnibus

ICGC: International Cancer Genome Consortium

MSigDB: Molecular Signature Database

IRB: institutional review board

FUSCC: Fudan University Shanghai Cancer Center

NMF: non-negative matrix factorization

RSS: residual sum of squares

FPKM: fragments per kilobase per million

TOM: topology overlap matrix

LASSO: Least Absolute Shrinkage and Selection Operator

DCA: Decision curve analysis

RMST: Restricted mean survival time

ES: enrichment score

NES: normalized enrichment score

FDR: false discovery rate

Declarations

Availability of data and materials

The datasets generated and analyzed during the current study are available in the TCGA repository (<https://portal.gdc.cancer.gov/>), ICGC database (<https://dcc.icgc.org/repositories>) and the GEO repository (GSE57495, <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE57495>).

Acknowledgements

The authors would like to thank all researchers contributed to the TCGA, GEO and ICGC data sets included.

Funding

This work was supported by National Natural Science Foundation of China (81972249, 81602078, 81802367, 81802361), Shanghai Clinical Research Plan of SHDC (SHDC2020CR4068), Shanghai Clinical science and technology innovation project of municipal hospital (SHDC12020102), Fudan University's 2019 "Double First-class" Original Research Personalized Support Project (XM03190634), Shanghai Science and Technology Development Fund (18ZR1408000, 17ZR1406500), Shanghai Science and technology development fund (19MC1911000), Clinical Research Project of Shanghai Municipal Health Committee (20194Y0348), Shanghai Anticancer Association EYAS project (SACA-CY19B10) and Hospital Foundation of Fudan University Shanghai Cancer Center (YJMS201907, YJQN201906, YJ201704).

Competing interests

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Ethics approval and consent to participate

Not applicable.

Patient consent for publication

Not applicable.

Authors' contributions

WS and MX designed the study. XW conducted the data process, modal establishment and visualization of analysis. MX, CT, MY and WW did the data analysis and interpretation. SN and MZ performed statistical analysis. CT and LW, WS and MX planned and supervised the project, performed data analysis and wrote the manuscript. CT and DH revised the manuscript. All authors have read and approved the final manuscript.

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* 2018;68(6):394-424.

2. Storz P, Crawford HC. Carcinogenesis of Pancreatic Ductal Adenocarcinoma. *Gastroenterology*. 2020;158(8):2072-81.
3. Liberti MV, Locasale JW. The Warburg Effect: How Does it Benefit Cancer Cells? *Trends Biochem Sci*. 2016;41(3):211-8.
4. Nayak AP, Kapur A, Barroilhet L, Patankar MS. Oxidative Phosphorylation: A Target for Novel Therapeutic Strategies Against Ovarian Cancer. *Cancers*. 2018;10(9).
5. Regel I, Kong B, Raulefs S, Erkan M, Michalski CW, Hartel M, et al. Energy metabolism and proliferation in pancreatic carcinogenesis. *Langenbecks Arch Surg*. 2012;397(4):507-12.
6. Brunet JP, Tamayo P, Golub TR, Mesirov JP. Metagenes and molecular pattern discovery using matrix factorization. *Proc Natl Acad Sci U S A*. 2004;101(12):4164-9.
7. Lovmar L, Ahlford A, Jonsson M, Syvänen AC. Silhouette scores for assessment of SNP genotype clusters. *BMC Genomics*. 2005;6:35.
8. Glen S. Sum of Squares: Residual Sum, Total Sum, Explained Sum: StatisticsHowTo.com: Elementary Statistics for the rest of us! ; [Available from: <https://www.statisticshowto.com/residual-sum-squares/>].
9. Li B, Severson E, Pignon JC, Zhao H, Li T, Novak J, et al. Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. *Genome biology*. 2016;17(1):174.
10. Yoshihara K, Shahmoradgoli M, Martínez E, Vegesna R, Kim H, Torres-Garcia W, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nature communications*. 2013;4:2612.
11. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics*. 2008;9:559.
12. Hughey JJ, Butte AJ. Robust meta-analysis of gene expression using the elastic net. *Nucleic Acids Res*. 2015;43(12):e79.
13. Blanche P, Dartigues JF, Jacqmin-Gadda H. Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Statistics in medicine*. 2013;32(30):5381-97.
14. Kerr KF, Brown MD, Zhu K, Janes H. Assessing the Clinical Impact of Risk Prediction Models With Decision Curves: Guidance for Correct Interpretation and Appropriate Use. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*. 2016;34(21):2534-40.
15. Chen DT, Davis-Yadley AH, Huang PY, Husain K, Centeno BA, Permuth-Wey J, et al. Prognostic Fifteen-Gene Signature for Early Stage Pancreatic Ductal Adenocarcinoma. *PLoS One*. 2015;10(8):e0133562.
16. Cheng Y, Wang K, Geng L, Sun J, Xu W, Liu D, et al. Identification of candidate diagnostic and prognostic biomarkers for pancreatic carcinoma. *EBioMedicine*. 2019;40:382-93.
17. Raman P, Maddipati R, Lim KH, Tozeren A. Pancreatic cancer survival analysis defines a signature that predicts outcome. *PLoS One*. 2018;13(8):e0201751.

18. Magouliotis DE, Sakellaridis N, Dimas K, Tasiopoulou VS, Svokos KA, Svokos AA, et al. In Silico Transcriptomic Analysis of the Chloride Intracellular Channels (CLIC) Interactome Identifies a Molecular Panel of Seven Prognostic Markers in Patients with Pancreatic Ductal Adenocarcinoma. *Curr Genomics*. 2020;21(2):119-27.
19. Andrejeva G, Rathmell JC. Similarities and Distinctions of Cancer and Immune Metabolism in Inflammation and Tumors. *Cell Metab*. 2017;26(1):49-70.
20. Hirschey MD, DeBerardinis RJ, Diehl AME, Drew JE, Frezza C, Green MF, et al. Dysregulated metabolism contributes to oncogenesis. *Semin Cancer Biol*. 2015;35 Suppl:S129-S50.
21. Chan AK, Bruce JI, Siriwardena AK. Glucose metabolic phenotype of pancreatic cancer. *World journal of gastroenterology : WJG*. 2016;22(12):3471-85.
22. Qin C, Yang G, Yang J, Ren B, Wang H, Chen G, et al. Metabolism of pancreatic cancer: paving the way to better anticancer strategies. *Molecular Cancer*. 2020;19(1):50.
23. Liang C, Qin Y, Zhang B, Ji S, Si S, Xu W, et al. Energy sources identify metabolic phenotypes in pancreatic cancer. *Acta biochimica et biophysica Sinica*. 2016;48.
24. Follia L, Ferrero G, Mandili G, Beccuti M, Giordano D, Spadi R, et al. Integrative Analysis of Novel Metabolic Subtypes in Pancreatic Cancer Fosters New Prognostic Biomarkers. *Frontiers in oncology*. 2019;9(115).
25. Akita H, Ritchie SA, Takemasa I, Eguchi H, Pastural E, Jin W, et al. Serum Metabolite Profiling for the Detection of Pancreatic Cancer: Results of a Large Independent Validation Study. *Pancreas*. 2016;45(10):1418-23.
26. Mayerle J, Kalthoff H, Reszka R, Kamlage B, Peter E, Schniewind B, et al. Metabolic biomarker signature to differentiate pancreatic ductal adenocarcinoma from chronic pancreatitis. *Gut*. 2018;67(1):128-37.
27. Bathe OF, Shaykhutdinov R, Kopciuk K, Weljie AM, McKay A, Sutherland FR, et al. Feasibility of Identifying Pancreatic Cancer Based on Serum Metabolomics. *Cancer Epidemiology Biomarkers & Prevention*. 2011;20(1):140-7.
28. Tao L, Zhou J, Yuan C, Zhang L, Li D, Si D, et al. Metabolomics identifies serum and exosomes metabolite markers of pancreatic cancer. *Metabolomics*. 2019;15(6):86.
29. Scalbert A, Brennan L, Fiehn O, Hankemeier T, Kristal BS, van Ommen B, et al. Mass-spectrometry-based metabolomics: limitations and recommendations for future progress with particular focus on nutrition research. *Metabolomics*. 2009;5(4):435-58.
30. Kang YP, Ward NP, DeNicola GM. Recent advances in cancer metabolism: a technological perspective. *Exp Mol Med*. 2018;50(4):31.
31. Nevins JR, Potti A. Mining gene expression profiles: expression signatures as cancer phenotypes. *Nat Rev Genet*. 2007;8(8):601-9.
32. O'Sullivan D, Sanin DE, Pearce EJ, Pearce EL. Metabolic interventions in the immune response to cancer. *Nat Rev Immunol*. 2019;19(5):324-35.

33. Biswas SK. Metabolic Reprogramming of Immune Cells in Cancer Progression. Immunity. 2015;43(3):435-49.

34. Lux A, Kahlert C, Grützmann R, Pilarsky C. c-Met and PD-L1 on Circulating Exosomes as Diagnostic and Prognostic Markers for Pancreatic Cancer. International journal of molecular sciences. 2019;20(13).

35. Zhu GH, Huang C, Qiu ZJ, Liu J, Zhang ZH, Zhao N, et al. Expression and prognostic significance of CD151, c-Met, and integrin alpha3/alpha6 in pancreatic ductal adenocarcinoma. Dig Dis Sci. 2011;56(4):1090-8.

36. Qian LW, Mizumoto K, Inadome N, Nagai E, Sato N, Matsumoto K, et al. Radiation stimulates HGF receptor/c-Met expression that leads to amplifying cellular response to HGF stimulation via upregulated receptor tyrosine phosphorylation and MAP kinase activity in pancreatic cancer cells. International journal of cancer Journal international du cancer. 2003;104(5):542-9.

37. Rucki AA, Xiao Q, Muth S, Chen J, Che X, Kleponis J, et al. Dual Inhibition of Hedgehog and c-Met Pathways for Pancreatic Cancer Treatment. Molecular cancer therapeutics. 2017;16(11):2399-409.

Tables

Table 1 Clinical characteristic of training and validation datasets

Characteristic		TCGA Set	Training Set	GSE57495 Set	ICGC Set
Age(years)	<65	78	71	-	103
	>=65	93	83	-	154
Survival state	Alive	80	74	21	151
	Dead	91	80	42	106
Gender	female	78	71	-	120
	male	93	83	-	137
Pathologic T	T1	7	6	-	-
	T2	21	20	-	-
	T3	138	123	-	-
	T4/Tx	4	4	-	-
Pathologic N	N1	119	107	-	-
	N /Nx	51	46	-	-
Pathologic M	Mx	90	81	-	-
	M0/ M1	81	72	-	-
Tumor Stage	Stage I	19	17	-	-
	Stage II	142	128	-	-
	Stage III	3	3	-	-
	Stage IV	3	3	-	-
Grade	G1	28	24	-	-
	G2	92	82	-	-
	G3	47	40	-	-
	G4/Gx	4	4	-	-
Total		171	154	63	257

Table 2 Univariant Cox regression of the 4-gene signature

Symbol	coefficient	Hazard ration	Z-score	P value	Low 95%CI	High 95%CI
AFF3	-0.1513	0.8595	-1.7450	0.0809	0.7252	1.0190
GJB5	0.0156	1.0157	3.4580	0.0005	1.0068	1.0250
MET	0.0045	1.0045	2.3600	0.0183	1.0008	1.0080
TMEM139	0.0164	1.0165	1.8980	0.0577	0.9995	1.0340

Figures

Figure 1

Identification of molecular subtypes in PAAD A: Consensus map of NMF clustering in TCGA PAAD dataset; B: Kaplan-Meier curves showed the overall survival (OS) curve of the two subtypes in TCGA PAAD dataset; C: Consensus map of NMF clustering in ICGC PAAD dataset; D: Kaplan-Meier curves showed the overall survival (OS) curve of the two subtypes in ICGC PAAD dataset; E: The proportions of B

cell, CD4+T cell, CD8+T cell, Neutrophil, Macrophage, and Dendritic cell (DC) between the two subtypes; F: Distribution of the ImmuneScore, StromalScore, and ESTIMATEScore between the two subtypes.

Figure 2

WGCNA co-expression analysis A: Gene dendrogram and module colors; B: Relationship between the 29 modules and the clinical phenotypes and molecular subtypes. C-D: The correlation of blue module with Cluster 1 (C) and yellow module with Cluster 2 (D) in the TCGA PAAD dataset; E: Volcano map of

differentially expressed genes (DEG) between the two subtypes; F: Heat map of the expression profile of 2411 DEGs and the distribution of clinicopathological parameters in the two subtypes.

Figure 3

Construction of the 4-gene signature in the training dataset A: Risk score, survival time, survival status and expression of the 4-gene signature in the training set; B: ROC curve of the 11-gene signature for 1-year, 3-year and 5-year survival in the training set. C: Kaplan-Meier survival analysis of overall survival for high-risk or low-risk group patients in the training set. ROC, receiver operating characteristic; AUC, area under the curve; HR, hazard ratio; CI, confidence interval.

Figure 4

Internal validation of the 4-gene signature's robustness in the entire TCGA cohort A: Risk score, survival time, survival status and expression of the 4-gene signature in the training set; B: ROC curve of the 11-gene signature for 1-year, 3-year and 5-year survival in the entire TCGA cohort; C: Kaplan-Meier survival analysis of overall survival for high-risk or low-risk group patients in the entire TCGA set. ROC, receiver operating characteristic; AUC, area under the curve; HR, hazard ratio; CI, confidence interval.

Figure 5

External validation of the 4-gene signature's robustness in the GSE57495 and ICGC PAAD cohorts A: Risk score, survival time, survival status and expression of the 4-gene signature in the GSE57495 dataset; B: ROC curve of the 4-gene signature for 1-year, 3-year and 5-year survival in the GSE57495 cohort; C: Kaplan-Meier survival curve based on the 4-gene signature in the GSE57495 cohorts. D: Risk score, survival time, survival status and expression of the 4-gene signature in the ICGC PAAD dataset; E: ROC

curve of the 4-gene signature for 1-year, 3-year and 5-year survival in the ICGC PAAD cohort; (D) Kaplan-Meier survival curve based on the 4-gene signature in the ICGC PAAD cohorts. ROC, receiver operating characteristic; AUC, area under the curve; HR, hazard ratio; CI, confidence interval.

Figure 6

Independence of the 4-gene prognosis signature A-B: Forest plot of the univariate (A) and multivariate (B) Cox regression analyses in TCGA PAAD dataset; C: A nomogram was developed by integrating the signature risk-score with the clinicopathologic features in the TCGA PAAD dataset. D: Calibration curves of nomogram for predicting OS at 1-year, 3-year and 5-year in the TCGA PAAD dataset.

Figure 7

The performance of the 6-gene signature in comparison to previous signatures in TCGA STAD dataset A: Restricted mean survival time (RMST) curve developed by integrating the indicated 5 signatures; B: DCA plots developed by integrating the indicated 5 signatures.

Figure 8

ssGSEA result according to the risk-score of PAAD samples in TCGA dataset A: Clustering of KEGG pathways correlated with RiskScore, with correlation coefficients greater than 0.40; B: Enrichment pathways that were significantly correlated in the high-risk and low-risk groups.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryFigure1.pdf](#)
- [SupplementaryFigure2.pdf](#)
- [SupplementaryFigure3.pdf](#)
- [SupplementaryFigure4.pdf](#)
- [SupplementaryFigure5.pdf](#)
- [SupplementaryTable1.pdf](#)
- [SupplementaryTable2.pdf](#)
- [SupplementaryTable3.pdf](#)
- [SupplementaryTable4.pdf](#)
- [SupplementaryTable5.pdf](#)
- [SupplementaryTable6.pdf](#)
- [SupplementaryTable7.pdf](#)