Received: 28 December 2022 | Revised: 4 May 2023 | Accepted: 11 May 2023

DOI: 10.1111/cts.13563

ARTICLE

Analysis of a large cohort of pancreatic cancer transcriptomic profiles to reveal the strongest prognostic factors

Máté Posta[1,2,3] | **Balázs Győrffy**[4]

[1]Károly Rácz Doctoral School of Clinical Medicine, Semmelweis University, Budapest, Hungary

[2]Oncology Biomarker Research Group, Institute of Enzymology, Research Centre for Natural Sciences, Budapest, Hungary

[3]Systems Biology of Reproduction Research Group, Institute of Enzymology, Research Centre for Natural Sciences, Budapest, Hungary

[4]Department of Bioinformatics and Department of Pediatrics, Semmelweis University, Budapest, Hungary

Correspondence
Balázs Győrffy, Department of Bioinformatics, Semmelweis University, Tűzoltó utca 7-9, Budapest 1094, Hungary.
Email: gyorffy.balazs@med.semmelweis-univ.hu

Abstract

Pancreatic adenocarcinoma remains a leading cause of cancer-related deaths. In order to develop appropriate therapeutic and prognostic tools, a comprehensive mapping of the tumor's molecular abnormalities is essential. Here, our aim was to integrate available transcriptomic data to uncover genes whose elevated expression is simultaneously linked to cancer pathogenesis and inferior survival. A comprehensive search was performed in GEO to identify clinical studies with transcriptome-level gene expression data of pancreatic carcinoma with overall survival data and normal pancreatic tissues. After quantile normalization, the entire database was used to identify genes with altered expression. Cox proportional hazard regression was employed to uncover genes most strongly correlated with survival with a Bonferroni corrected $p < 0.01$. Perturbed biological processes and molecular pathways were identified to enable the understanding of underlying processes. A total of 16 available datasets were combined. The aggregated database comprised data of 1640 samples for 20,443 genes. When comparing with normal pancreatic tissues, a total of 2612 upregulated and 1977 downregulated genes were uncovered in pancreatic carcinoma. Among these, we found 24 genes with higher expression which significantly correlated with overall survival length also. The most significant genes were ANXA8, FAM83A, KRT6A, MET, MUC16, NT5E, and SLC2A1. These genes remained significant after a multivariate analysis also including grade and stage. Here, we assembled a large-scale database of pancreatic carcinoma samples and used this cohort to identify carcinoma-specific genes linked to altered survival outcomes. As our analysis focused on genes with higher expression, these could serve as future therapy targets.

Study highlights

WHAT IS THE CURRENT KNOWLEDGE ON THE TOPIC?

Despite recent advances in cancer treatment for various solid tumors, the prognosis in pancreatic cancer remains very poor. A deeper understanding of the disease will be necessary for further advancement in this field. To achieve this goal, here

we establish a large database for the integrative evaluation of pancreatic tumor tissue gene expression data and the associated survival.

WHAT QUESTION DID THIS STUDY ADDRESS?

We aimed to identify potential therapeutic targets for pancreatic cancer by analyzing the largest available integrated transcriptomic database of pancreatic carcinoma. We planned to determine which genes were downregulated or overexpressed in pancreatic cancer compared with normal pancreatic tissues and filter for those linked to shorter survival. Additionally, we aimed to identify the biological processes linked to pancreatic carcinoma.

WHAT DOES THIS STUDY ADD TO OUR KNOWLEDGE?

We expand the potential therapeutic targets for pancreatic cancer by identifying seven genes whose increased expression is associated with shorter survival. We also found that biological processes linked to the significantly overexpressed genes include metabolic processes, cell cycle and organelle organization, and cellular transport. Additionally, via the provision of a table of significant genes, we support the identification of future biomarker candidates.

HOW MIGHT THIS CHANGE CLINICAL PHARMACOLOGY OR TRANSLATIONAL SCIENCE?

The development of drugs targeting the identified genes could lead to more effective treatments for pancreatic cancer patients. Additionally, the identification of potential biomarker candidates could lead to earlier detection of relapse of pancreatic cancer, which could ultimately improve patient outcomes. Overall, the new therapeutic targets may support clinical pharmacology and translational science by providing new avenues for drug development.

INTRODUCTION

With more than 450,000 deaths every year, pancreatic cancer is the seventh leading cause of death from cancer. The occurrence of pancreatic cancer increased in recent decades in both men and women.[1] Due to extremely poor survival rates, nearly as many people die each year as are diagnosed. More than 95% of malignant tumors of the pancreas arise from the exocrine parts of the gland, and they are histologically in most cases pancreatic ductal adenocarcinomas (PDAC).[2,3] Risk factors include older age, smoking, type 2 diabetes mellitus, and various genetic syndromes.[4] Unfortunately, most patients notice the symptoms only late during the course of the disease preventing the execution of surgery, the only curative treatment. In the last 30 years the prognosis of PDAC has improved significantly, but even now, the 5-year survival is only slightly over 4%.[5] Even after pancreatectomy, the prognosis remains very poor, with a 25% 5-year survival rate.[6,7] Notably, some patients can have rare non-ductal pancreatic tumors including pancreatic neuroendocrine tumor, solid pseudopapillary neoplasm, acinar cell carcinoma, and pancreatoblastoma. The prognosis of these tumors varies, but the expected survival is generally better than for ductal adenocarcinoma.[8]

Currently, apart from the CA-19-9 tumor marker, there is no widely used biomarker for pancreatic carcinoma. However, CA-19-9 cannot be used for tumor screening either, rather it is used to monitor remission after therapy. Carcinoembryonic antigen (CEA), which is also used as a biomarker for other tumors, is frequently increased in PDAC.[9] CA-125 (the product of the MUC16 gene), a marker with potential benefit for ovarian, breast, and lung tumors, is also increased in PDAC, but is less useful as a marker. Nevertheless, the combination of CA-125 and CA-19-9 was able to detect the existence of a pancreatic tumor with higher sensitivity and specificity.[10]

Once systemic therapy is needed, it is very difficult to select drugs for pancreatic cancer. For a long time, the only chemotherapy agent given was gemcitabine. Current recommendations for treatment of patients with metastatic disease include the combination therapy regimen FOLFIRINOX (folinic acid, fluorouracil, irinotecan, and oxaliplatin) and gemcitabine with nab-paclitaxel or gemcitabine alone.[11,12] Response to treatment can be evaluated based on imaging methods, serum markers (such as CA-19-9), and changes in tumor-related symptoms. Among targeted therapies, the epidermal growth factor receptor (EGFR) receptor antagonist erlotinib is the only available option.[13] Blocking

cMET has recently emerged as a potential therapeutic target in pancreatic carcinoma, which can help to prolong overall survival.[14]

Transcriptomic analysis makes it possible to examine the entire genome without a preset hypothesis and reveal pathological processes altered at the molecular level. Two high-throughput techniques are the most widely used for studying gene expression: gene chip and RNA-seq. During a microarray analysis, mRNAs are converted into cDNAs by reverse transcription, labeled, and then these DNA fragments are hybridized to an in situ synthesized and immobilized nucleotide sequence.[15] During RNA-seq, the fragments are identified using a high-throughput sequencing technique, and then the level of expression is quantified based on the number of reads aligned to a selected position in the genome.[16] RNA-seq has some advantages over hybridization-based technologies, as it can provide consistent quantification in a larger dynamic range and without the need of predetermined sequence information. In addition to the quantitative measurement of gene expression, RNA-seq can also determine alternative splicing, single nucleotide polymorphisms (SNPs), and the presence of gene fusions. In contrast, microarrays have other advantages, for example they are cheaper, and their data analysis is more straightforward with established analysis tools and pipelines.[17,18]

In this study, we aimed to establish an integrated database of transcriptome-level gene expression from pancreatic carcinomas by utilizing gene chip and RNA-seq studies of recent years. We set two specific goals to employ this database. First, we aimed to identify transcriptomic changes when comparing tumor and normal tissues to find biomarkers that could indicate the disease presence. Our second goal was to find altered genes linked with disease outcome that can serve as potential therapeutic targets in the future. Finally, we wanted to uncover biological processes that are perturbed in pancreatic carcinoma and could play a pathogenic role in the disease.

METHODS

Database setup

We searched the GEO repository (https://www.ncbi.nlm.nih.gov/geo/) using the terms "pancreas cancer" and "pancreatic adenocarcinoma"; these searches yielded 1097 results. We further filtered the search by defining the species from which the sample originates as *Homo sapiens* and the test type either as "Expression profiling by high throughput sequencing" or "Expression profiling by array" or "Protein profiling by protein array" or "Protein profiling by Mass Spec." In addition, we only kept the datasets that contained at least 30 samples. With these, a total of 160 dataset hits were obtained. We reviewed these series and retained only those that contained overall survival (OS) or disease-free survival (DFS) information in either the GEO database or the supplementary material of the related articles (Figure 1). In addition to the GEO datasets, we have also added projects from the International Cancer Genome Consortium Data Portal on pancreatic carcinomas that contain gene expression data. These include the Pancreatic cancer – Ductal adenocarcinoma – Australia (PACA-AU), Pancreatic cancer – Ductal adenocarcinoma – Canada (PACA-CA), Pancreatic cancer – Adenocarcinoma – United States (PAAD-US), and Pancreatic cancer – Endocrine neoplasms – Australia (PAEN-AU) sets. The aggregated datasets contain gene expression data of tissue samples from pancreatic tumors.

We then uniformized the gene annotation across the various datasets so that each gene expression data was linked to the corresponding gene symbol. Before merging the data tables, the probe with the highest summated expression value across all samples within a given dataset was considered in case a gene was measured by multiple probes. Some samples were measured multiple times or there was more than one sample from the same tumor – in such cases only one measurement result was kept. We also utilized the symbol checker (https://www.genenames.org/tools/multi-symbol-checker/) to identify symbols that had changed and switched to the most recent versions of those symbols. Finally, the individual datasets were combined symbol-wise, thus a concatenated dataset was created.

Clinical data for the samples were collected in a separate table. In this, TNM status was converted to stage using the definition of the American Joint Committee on Cancer,[19] by assigning M1 with any T or N state to stage 4, N2 without metastasis or T4 without metastasis to stage 3, N1 without metastasis (M0) and with a maximum of T3 state to stage 2B, T3N0M0 to stage 2A, T2N0M0 to stage 1B, and T1N0M0 to stage 1A.

Processing of transcriptomic data

We only kept the expression values of the samples for which any clinical data were available. Then, we reviewed the genes and kept only those with expression data in at least 50% of the samples. This was necessary because there were many genes in the processed data that were only measured in a few samples. This step also serves as a filtering stage, since there must be a consensus between the different studies to investigate only genes annotated with a meaningful identifier, thus we eliminated ambiguous artifact transcripts.

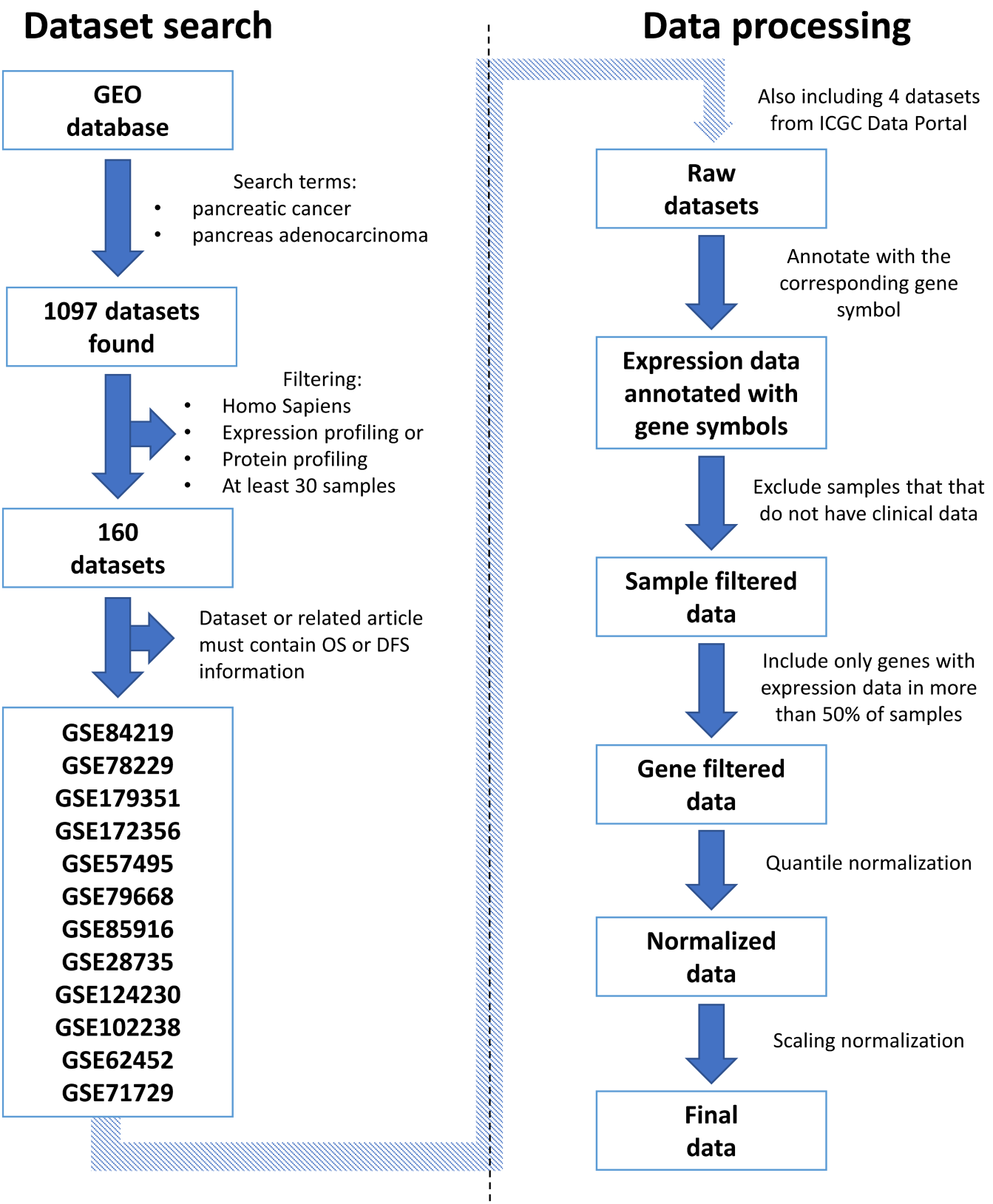

FIGURE 1 Flowchart of dataset search and data processing. DFS, disease-free survival; OS, overall survival.

Subsequently, we normalized the data with two algorithms. First, we performed a quantile normalization to bring all datasets to an adjusted metric by using the preprocessCore package in the R environment (https://www.r-project.org/). After that, we performed a scaling normalization, in which the average expression was set

to 1000 for each sample. This makes it easy to determine whether the expression of a given gene exceeds the average expression level across all genes. To make the calculations easier, the obtained expression values were rounded to whole numbers. Note that not every dataset had expression value for each gene, and therefore most genes had to be investigated in slightly different set of samples as we had to exclude samples with missing values from the analysis of a particular gene. The complete data processing is summarized in Figure 1.

Identification of genes showing altered expression between normal and tumor samples

As the datasets contained control pancreatic samples in addition to the tumor samples, we were able to identify expression changes between tumor and non-tumor pancreatic samples across all genes. Differentially expressed genes were determined using the Mann–Whitney U test. To determine significant genes, the following criteria were used: the absolute value of the $\log_2$ transformed fold-change had to be higher than 0.585 ($\log_2 1.5$), the Bonferroni corrected p value of the Mann–Whitney U test had to be less than 0.01, the mean of the expression in at least one of the sample groups (normal or tumor) had to be higher than 500, and finally the mean expression in the other group had to be more than 50. The last requirements ensure that only genes that are meaningfully expressed in at least one of the cohorts and actually expressed in both cohorts are considered.

Survival analysis

To identify genes associated with altered survival, we computed Cox proportional hazards regression[20] for each gene using the overall survival information whenever this was available. We first trichotomized the data, that is, we assigned the samples into three cohorts based on the tertiles of gene expression and then excluded the middle cohort. The sample was assigned into the "upper tertile" cohort in cases where the gene expression levels were higher than the upper tercile and into the "lower tertile" in cases where the gene expression was lower than the lower tercile. It is a common method in the evaluation of clinical parameters to distinguish between a low, a high, and an intermediate group. Although there are other methods for determining cutoff points (e.g., using upper/lower quartile or median, or the computation of the best cutoff with combination of a false discovery rate computation),[21] we aimed to select a method that is similar to Food and Drug Administration (FDA)-approved diagnostic tools where a significant proportion of patients are not classified to increase the robustness of the analysis – see for example tests for breast, lung, and prostate cancer.[22–24] Univariate and multivariate Cox models were carried out in a Python environment using the 'lifelines' package. The expression of a gene was considered significant if the obtained hazard ratio exceeded 1.66 or was less than 0.60, and the Bonferroni corrected p value was less than 0.01.

Determining potential therapeutic targets

Our goal was to uncover target gene products that have the potential for future therapeutic intervention. Since it is only possible to therapeutically influence genes with elevated expression, we only considered genes differentially expressed in the positive direction. Furthermore, it was a second important criterion that overexpression of the gene was significantly linked to reduced survival, so we selected only genes with a hazard ratio over 1.66. Multivariate Cox proportional hazard regression was performed to compare gene expression and the role of sex, grade, and stage in relation to overall survival. The trichotomized lower and upper terciles expression data were used for each gene in the multivariate analysis also.

Network analysis

Gene ontology enrichment analysis was performed using the Database for Annotation, Visualization, and Integrated Discovery tool (DAVID: http://david.abcc.ncifcrf.gov/). Enriched molecular functions and biological processes were identified for the previously identified differentially expressed genes. Gene ontology terms were considered enriched if the Benjamini–Hochberg-based false discovery rate was below 5% and the fold enrichment (FE) was higher than 1.5. The networks of biological processes enriched among genes were created by BINGO[25] and visualized with Cytoscape.[26]

RESULTS

Processing of the datasets

We identified 12 datasets in total during our GEO search that matched our criteria, and added four more datasets from the International Cancer Genome Consortium Data Portal. A complete list of the included cohorts is presented in Table 1.

TABLE 1 Clinical cohorts included in the final integrated database.

Dataset	Project title	Platform	Genes (*n*)	Samples (*n*)
GSE84219	Expression analysis in ductal adenocarcinoma in patients with low and high survival after tumor resection	GPL14951	19,252	30
GSE78229	Microarray gene-expression profiles of 50 pancreatic tumors tissue from patients with pancreatic ductal adenocarcinoma	GPL6244	19,988	50
GSE179351	Radiation therapy enhances immunotherapy response in microsatellite-stable colorectal and pancreatic adenocarcinoma in a phase II trial	GPL18573	20,210	12
GSE172356	Tumor microbiome contributes to an aggressive phenotype in the basal-like subtype of pancreatic cancer	GPL20795	19,240+	53
GSE57495	Microarray analysis of 63 patients with pancreatic cancer tissues resulted in the identification of a 15-gene signature to predict overall survival	GPL15048	19,874	63
GSE79668	RNA-sequencing of human pancreatic adenocarcinoma cancer tissues	GPL11154	19,175	51
GSE85916	Patients with human resected pancreatic cancer	GPL13667	19,322	80
GSE28735	Microarray gene-expression profiles of 45 matching pairs of pancreatic tumor and adjacent non-tumor tissues from 45 patients with pancreatic ductal adenocarcinoma	GPL6244	17,546	90
GSE124230	Pancreatic cancer prognosis is predicted by chromatin accessibility	GPL11154	19,745	49
GSE102238	Gene expression signatures associated with perineural invasion in pancreatic ductal adenocarcinoma	GPL19072	8735	100
GSE62452	Microarray gene-expression profiles of 69 pancreatic tumors and 61 adjacent non-tumor tissue from patients with pancreatic ductal adenocarcinoma	GPL6244	19,988	130
GSE71729	Virtual microdissection of pancreatic ductal adenocarcinoma reveals tumor and stroma subtypes	GPL20769	18,229	252
PACA-AU	Pancreatic cancer – ductal Adenocarcinoma – Australia	GPL10558 & Illumina Hi seq	20,082+	269
PACA-CA	Pancreatic cancer – Ductal adenocarcinoma – Canada	GPL10558	19,757+	234
PAAD-US	Pancreatic cancer – Adenocarcinoma – United States	Illumina Hi seq	19,689	145
PAEN-AU	Pancreatic cancer – Endocrine neoplasms – Australia	GPL10558 & Illumina Hi seq	20,058+	32

The GSE84219 dataset included Illumina probe identifiers and the associated gene symbols are based on the description in the corresponding (GPL14951) platform. The GSE78229, GSE28735, and GSE62452 datasets used the GPL6244 platform, and the gene IDs were paired using the platform description. The GSE57495 and GSE85916 datasets were measured using Affymetrix gene arrays and the corresponding gene symbols were linked according to the GPL15048 and GPL13667 platforms. Only Entrez ID were found without gene symbols in the case of the GSE57495 dataset – here we searched for the corresponding symbols by employing the mygene python package. In the GSE124230 and the GSE79668 datasets the genes were annotated with Ensembl ID and gene read counts values

were available only. We downloaded the GRCh38 (release 106) complete human gene set and calculated the lengths of the genes with the gtftools python package. Then, we performed transcript per million (TPM) normalization by using the median gene length for the calculation. In the GSE102238 dataset the probes were available with Agilent IDs. The associated platform available in GEO (GPL19072) did not contain the gene symbol or any gene ID, only the probe sequences, and we had to determine the Ensembl ID based on data from another platform (GPL26898). In the GSE71729 and PAAD-US the expression data were given with gene symbols so no further action was required. In the PACA-AU, PACA-CA, and PAEN-AU datasets the expression sequencing dataset used Ensembl ID as identifier enabling a direct link with gene symbols. The expression array dataset contained the expression data with Illumina probe IDs. By linking the data with the appropriate platform (GPL10558), we obtained the symbols for these genes as well. In the case of the GPL6244, GPL13667, and GPL26898 platforms, many probes were annotated with multiple gene IDs. As our aim was to avoid missing potentially significant genes, we assigned the measured intensity values to each corresponding gene in these samples.

Complete integrated database

In the final table already filtered for genes and samples, we obtained gene expression data for a total of 1640 samples and 20,433 genes. Of the samples, a total of 1435 were pancreatic carcinoma samples, and the remaining 205 samples were controls from healthy (non-tumorous) pancreatic tissues. The controls included not only pancreatic samples from healthy patients, but also parts of the pancreas from cancer patients that were not infiltrated by tumors. From the histopathologic aspect, 94% of the tumor samples were PDAC tumors. Note that for a proportion of patients, histology was not available. To increase the sample number with available gene expression data, all tumors classified as pancreatic cancer were included in the analysis regardless of histological diagnosis. Furthermore, in many cases, the sampling was performed during surgical resection of the tumor, but the sample acquisition is not described for several datasets. Similarly, only a few datasets mention information related to treatment.

Among the tumorous samples, 8.9% were grade 1 (well differentiated), 52.1% grade 2 (moderately differentiated), 37.0% grade 3 (poorly differentiated), and 2.0% grade 4 (undifferentiated). The stage distribution was similar with 12.6% stage 1, 78.5% stage 2, 5.0% stage 3, and 3.9% stage 4. We had survival data for a total of 1245 patients, of which 812 died during the study (mean overall survival time ± SD: 16.9 ± 15.8 months) and 433 were censored (mean time until censoring ± SD: 24.4 ± 22.4 months).

Genes with perturbed expression pattern

When comparing normal tumor tissues and pancreatic carcinoma, we found a total of 4589 genes whose expression was changed, of which 2612 genes were upregulated, while roughly the same number, 1977 genes, were downregulated. Genes with the most significant upregulation were SERPINB5 ($\log_2$FC: 3.39, p: 1.36e-90), MLPH ($\log_2$FC: 3.28, p: 9.76e-90), and TRIM29 ($\log_2$FC: 3.72, p: 4.30e-85). The most significant genes with lower expression were PNRC1 ($\log_2$FC: −1.34, p: 8.17e-81), LONRF2 ($\log_2$FC: −1.78, p: 2.61e-80), and PRKAR2B ($\log_2$FC: −2.57, p: 5.25e-80). The complete list for all genes is provided in Table S1. Note that even with our strict thresholds, more than 22% of the analyzed transcripts were significantly changed.

Genes linked with altered survival outcome

Using the Cox proportional hazards model, we found 24 genes whose expression significantly was linked to altered survival length. The majority of genes ($n = 21$) correlated with shorter survival (top three genes according to hazard ratio [HR]: ANXA8 [HR: 1.99, p: 1.15e-12], KRT6A [HR: 1.94, p: 7.38e-14], and MET [HR: 1.91, p: 4.59e-13]) and only a few genes ($n = 3$) were good prognostic factors (LOC113230 [HR: 0.53, p: 1.97e-7], RGS5 [HR: 0.57, p: 2.37e-10], and RETREG1 [HR: 0.59, p: 7.32e-10]; Table S2).

Potential targets

By using genes with elevated expression in cancerous tissues and linked with increased hazard rate in the survival analysis we identified the nine most robust target genes, ANLN ($\log_2$FC: 2.33; HR: 1.67), ANXA8 ($\log_2$FC: 2.87; HR: 1.99), FAM83A ($\log_2$FC: 3.13; HR: 1.76), KRT6A ($\log_2$FC: 3.49; HR: 1.94), MET ($\log_2$FC: 0.99; HR: 1.91), MUC16 ($\log_2$FC: 3.79; HR: 1.85), NT5E ($\log_2$FC: 1.38; HR: 1.69), SLC2A1 ($\log_2$FC: 1.90; HR: 1.88), and SLTM ($\log_2$FC: 1.28; HR: 1.74). To check the independence of genes in relation to survival, we performed multivariate Cox regression using gene pairs. We found that each gene pair remained significant, with the exception of SLTM and ANLN when they were examined together with the MET gene, so we discarded these genes (Table S3). Their expression difference compared to the controls is depicted

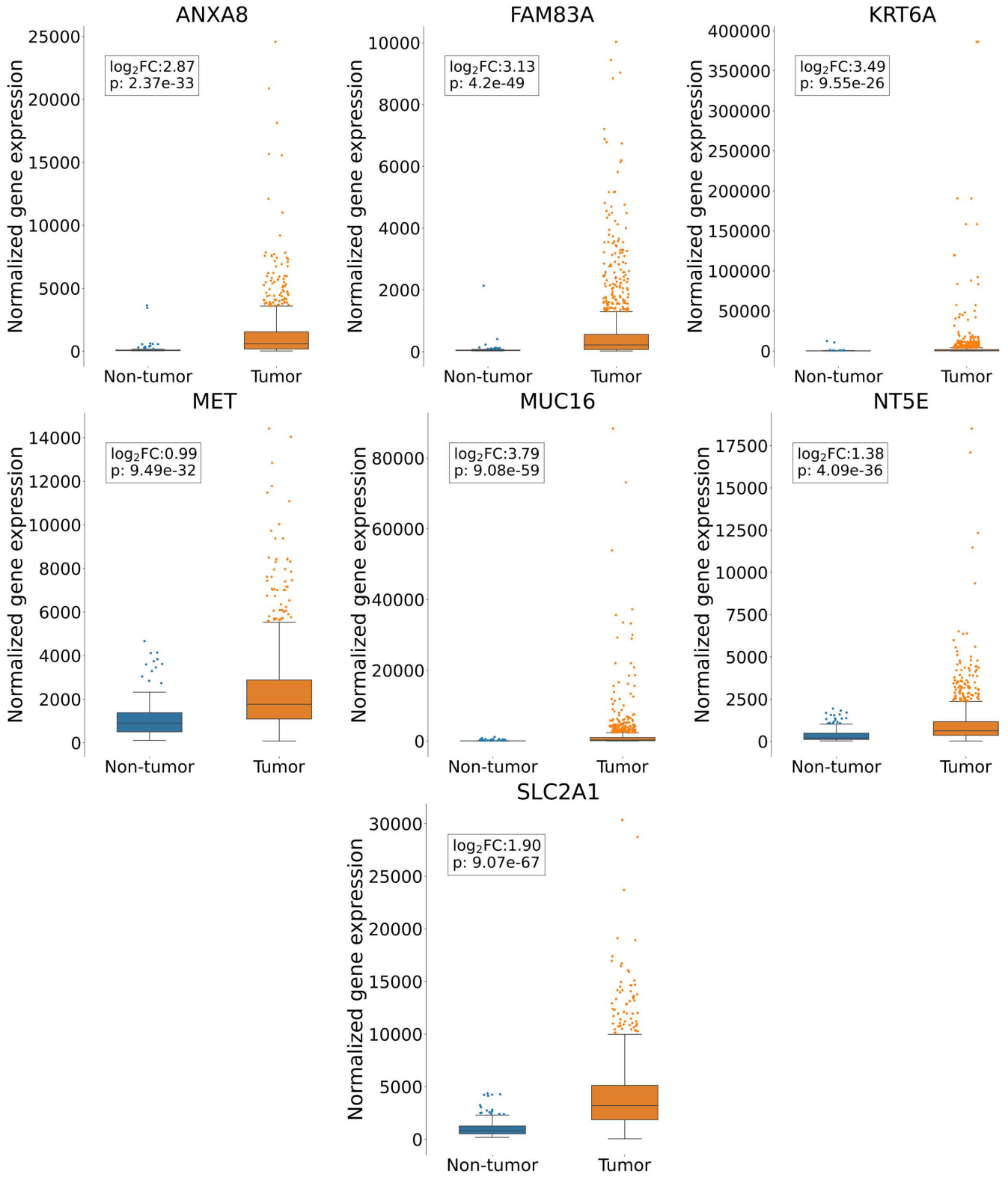

FIGURE 2 Boxplots of the selected candidate therapeutic target genes showing the highest difference in expression between normal and tumorous pancreatic tissues. $\log_2$FC represents the log 2 of the fold-change and *p* represents the significance in the Mann–Whitney U test.

in Figure 2 and the association between expression and survival is depicted in Figure 3. In our multivariate analysis, we found that the role of genes also remained significant while grade and stage were also significantly linked to survival outcome. Gender did not have a significant association with survival (Table 2). These results prove that the selected top genes and the clinical variables capture different, clinically relevant features of the tumors.

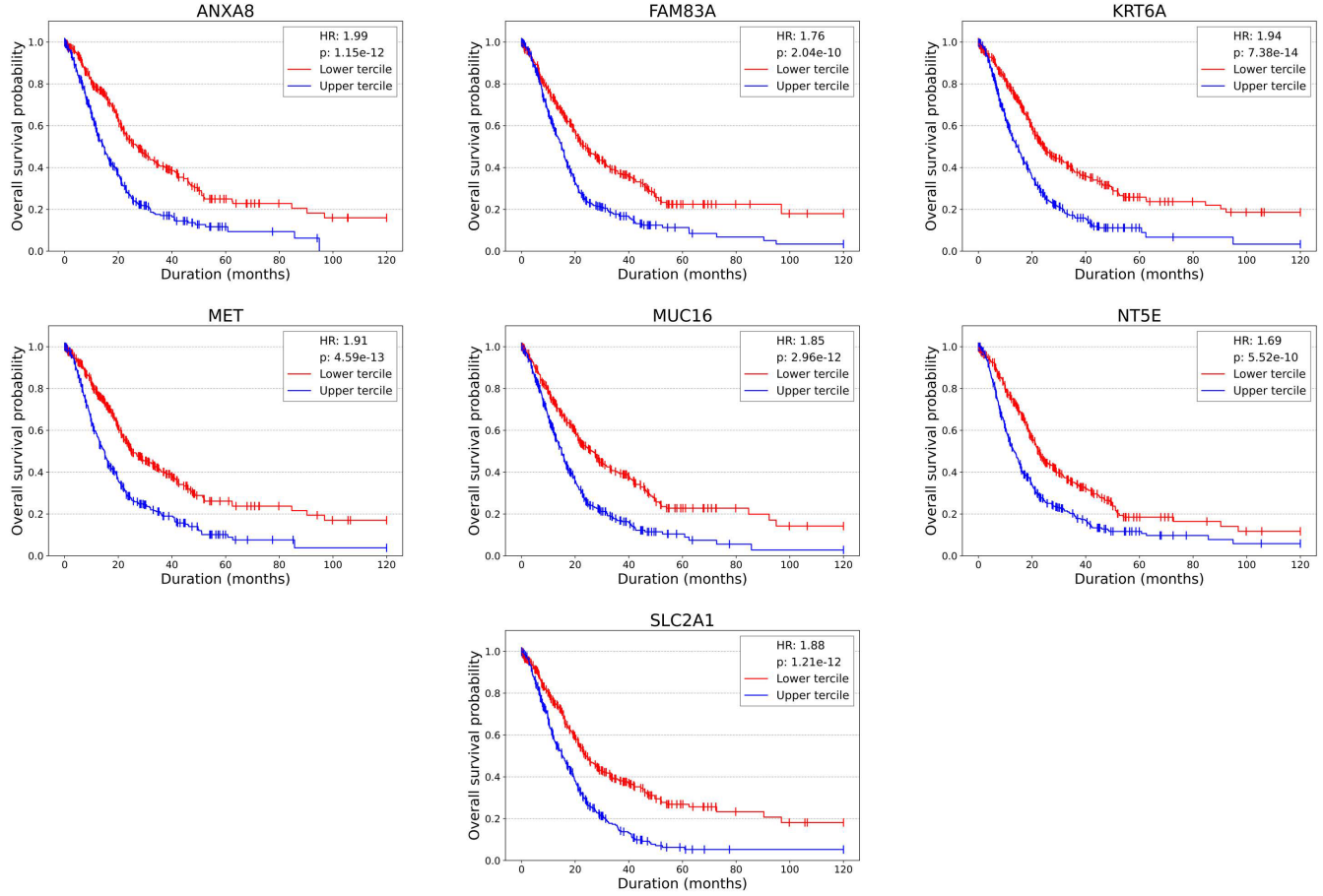

FIGURE 3 Kaplan–Meier survival plots of the selected candidate therapeutic target genes. HR, hazard rate.

TABLE 2 Multivariate analysis results for the selected genes in combination with stage, grade, and gender.

	KRT6A	MET	NT5E	ANXA8	SLC2A1	FAM83A	MUC16
Stage	p_{KRT6A}: 2.6e-8 p_{stage}: 8.3e-3	p_{MET}: 1.9e-5 p_{stage}: 3.3e-3	p_{NT5E}: 3.3e-9 p_{stage}: 5.7e-3	p_{ANXA8}: 3.9e-8 p_{stage}: 1.0e-2	p_{SLC2A1}: 3.4e-8 p_{stage}: 4.5e-3	p_{FAM83A}: 6.0e-6 p_{stage}: 3.6e-3	p_{MUC16}: 4.2e-6 p_{stage}: 8.2e-5
	HR_{KRT6A}: 2.06 HR_{stage}: 1.34	HR_{MET}: 1.78 HR_{stage}: 1.34	HR_{NT5E}: 2.15 HR_{stage}: 1.36	HR_{ANXA8}: 2.11 HR_{stage}: 1.34	HR_{SLC2A1}: 2.06 HR_{stage}: 1.31	HR_{FAM83A}: 1.73 HR_{stage}: 1.35	HR_{MUC16}: 1.81 HR_{stage}: 1.48
Grade	p_{KRT6A}: 3.5e-6 p_{grade}: 2.1e-4	p_{MET}: 1.5e-2 p_{grade}: 2.8e-2	p_{NT5E}: 4.8e-3 p_{grade}: 5.0e-4	p_{ANXA8}: 8.8e-6 p_{grade}: 1.9e-2	p_{SLC2A1}: 6.1e-4 p_{grade}: 2.3e-4	p_{FAM83A}: 2.5e-4 p_{grade}: 4.5e-3	p_{MUC16}: 3.7e-4 p_{grade}: 1.2e-2
	HR_{KRT6A}: 1.91 HR_{grade}: 1.39	HR_{MET}: 1.40 HR_{grade}: 1.21	HR_{NT5E}: 1.49 HR_{grade}: 1.35	HR_{ANXA8}: 1.96 HR_{grade}: 1.24	HR_{SLC2A1}: 1.60 HR_{grade}: 1.37	HR_{FAM83A}: 1.64 HR_{grade}: 1.29	HR_{MUC16}: 1.67 HR_{grade}: 1.27
Sex	p_{KRT6A}: 8.7e-9 p_{sex}: 4.8e-1	p_{MET}: 1.0e-6 p_{sex}: 2.5e-1	p_{NT5E}: 1.2e-6 p_{sex}: 2.7e-2	p_{ANXA8}: 7.8e-10 p_{sex}: 3.3e-1	p_{SLC2A1}: 4.8e-7 p_{sex}: 3.0e-1	p_{FAM83A}: 1.2e-9 p_{sex}: 4.4e-1	p_{MUC16}: 2.7e-7 p_{sex}: 4.5e-1
	HR_{KRT6A}: 2.02 HR_{sex}: 0.92	HR_{MET}: 1.83 HR_{sex}: 0.87	HR_{NT5E}: 1.81 HR_{sex}: 0.77	HR_{ANXA8}: 2.41 HR_{sex}: 0.88	HR_{SLC2A1}: 1.88 HR_{sex}: 0.88	HR_{FAM83A}: 2.23 HR_{sex}: 0.91	HR_{MUC16}: 1.92 HR_{sex}: 0.91

Abbreviation: HR, hazard ratio.

Molecular functions and biological processes perturbed in pancreatic carcinoma

We separated the significantly changed genes according to whether they were upregulated or downregulated compared to the controls and searched for the gene ontology molecular function and biological processes in DAVID. When using the upregulated genes, we found 45 enriched molecular functions, the top three including cell adhesion molecule binding (FE: 2.38, *p*: 1.88e-28), cadherin binding (FE: 2.72, *p*: 9.58e-26), and macromolecular complex binding (FE: 1.57, *p*: 1.26e-15). The most significant enriched biological process terms were regulation of cellular

component organization (FE: 1.51, *p*: 3.50e-27), cellular protein localization (FE: 1.52, *p*: 1.15e-18), and cellular macromolecule localization (FE: 1.51, *p*: 2.17e-18). The complete set of biological processes influenced by the upregulated genes is provided in Figure 4 (high resolution in Figure S1). In the case of downregulated genes, we found 18 enriched biological processes but no significantly altered molecular functions. The three most significant biological processes were detoxification (FE: 2.68, *p*: 5.53e-6), muscle cell proliferation (FE: 2.01, *p*: 1.37e-5), and detoxification of copper ion (FE: 6.77, *p*: 1.56e-5).

DISCUSSION

The identification of differentially expressed genes in pancreatic cancer is critical for developing new markers for detection and treatment in this highly lethal tumor. Our findings confirm that several previously identified highly expressed genes, many of which offer potential clinical targets, are overexpressed in pancreatic cancer. These differentially expressed genes can form the basis for screening methods to detect the disease at an earlier, potentially curable stage, or serve as new targets for drug development or imaging.

In the course of our work, we combined 16 datasets, and used this integrated database for correlating gene expression and survival. Genes linked with the poorest prognosis were ANXA8, KRT6A, and MET whereas the best prognostic genes were LOC113230, RGS5, and RETREG1. During the data analysis, we identified seven genes, the expression of which was higher in pancreatic carcinoma, and this higher level is also linked to shorter survival, including ANXA8, FAM83A, KRT6A, MET, MUC16, NT5E, and SLC2A1. We also found that these genes, with the exception of SLTM and ANLN, independently affect overall survival. The results of the multivariate analysis show that there is a correlation between the MET gene and the ANLN and SLTM genes in terms of survival, so the prognostic role of the genes is not independent of each other.

ANXA8 (annexin A8) and its paralogous gene, ANXA8L1 (annexin A8 like 1) had the highest hazard ratio in the Cox regression. The overexpression of these genes has been previously described in several tumors,[27,28] and they also have an important function in promoting

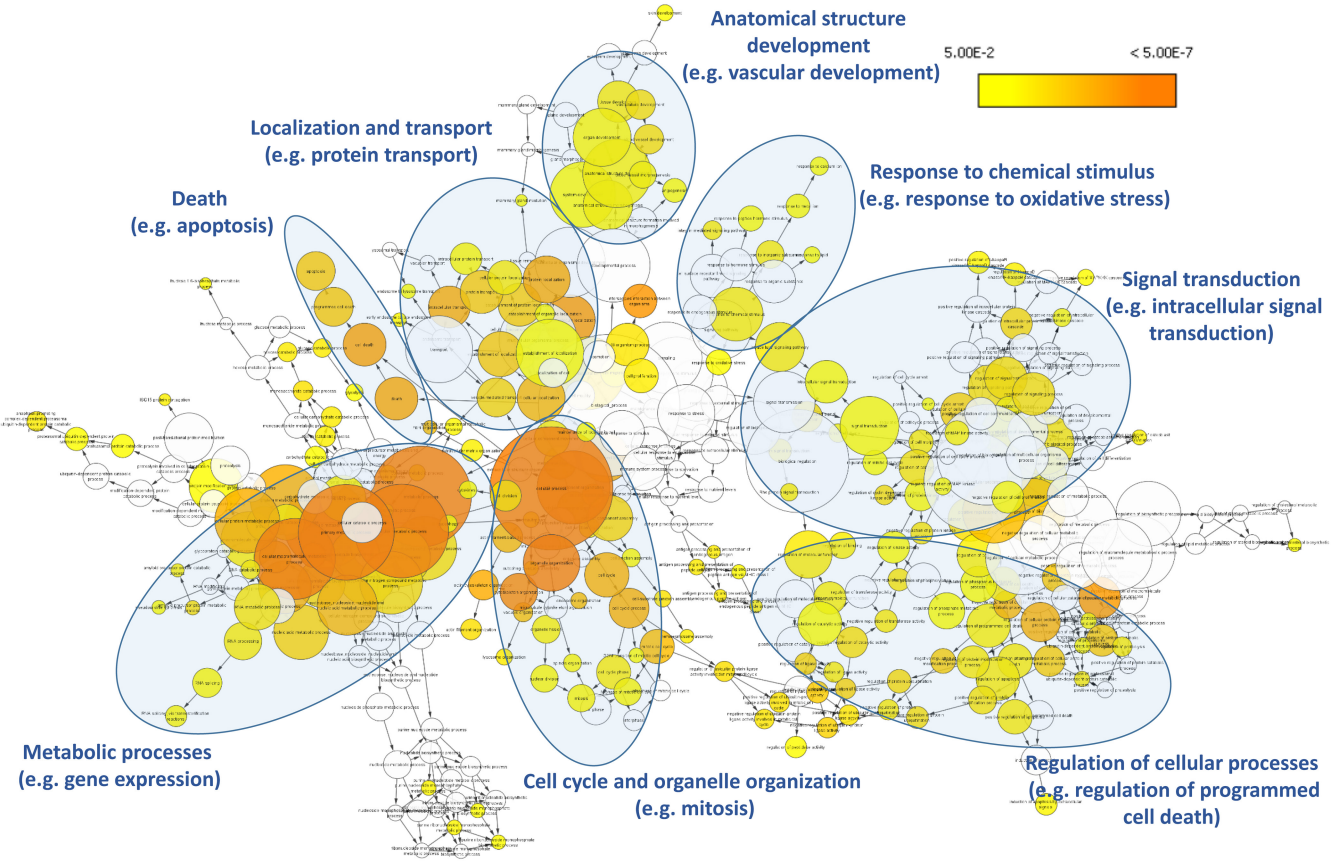

FIGURE 4 Perturbed biological processes for genes upregulated in pancreatic carcinoma. Circle sizes relate to the number of genes involved in the biological processes and colors refer to *p* values according to the color code. White bubbles represent non-significant categories connecting other biological processes. High resolution image is provided as Figure S1.

invasion and proliferation.[28,29] Keratin 6A (KRT6A) belongs to the keratin protein family, which is an essential component of the cytoskeleton. It may affect the pentose phosphate pathway, that was described to promote tumor growth and metastasis in lung cancer.[30] Although the molecular mechanism by which KRT6A promotes tumor progression is not fully understood, a high level of KRT6A is also associated with poor clinical outcome in pancreatic cancer.[31] SLC2A1 (solute carrier family 2 member 1), also known as GLUT-1 (glucose transporter protein type 1), was previously linked with the presence and prognosis of pancreatic cancer in multiple studies.[32–34] NT5E (5′-nucleotidase ecto, also called CD73) has immunosuppressive properties, and NT5E inhibitors are currently under development.[35] As here we show higher NT5E expression in pancreatic cancer, our results suggest that NT5E inhibitors might also be potentially effective in PDAC.

The PI3K-AKT signaling plays a major role in the pathophysiology of pancreatic cancer, which is also supported by our discovery of multiple genes linked to this pathway. The AKT kinase promotes cell survival, initiation of cell division, increased metabolism, growth, angiogenesis, and DNA repair through the phosphorylation of many other proteins.[36] The PI3K-AKT pathway is affected by MUC16,[37] FAM83A,[38] and NT5E,[39] which all have been previously identified as poor prognostic factors in pancreatic cancer[40–44] and as potential therapeutic targets.[45–49] However, MET is currently the most significant gene in this regard, according to our clinical knowledge. The MET proto-oncogene encodes the protein MET (c-MET), a membrane tyrosine kinase receptor. First, MET binds to its ligand, hepatocyte growth factor (HGF). Then, HGF is released by stromal cells, dimerizes, and by this it gets into an activate state where it can activate the PI3K/AKT pathway.[50] There are MET inhibitors already in use or under clinical development, including small molecule MET receptor inhibitors (e.g., crizotinib, savolitinib, tepotinib, cabozantinib), MET receptor monoclonal antibodies (e.g., onartuzumab), and antibodies against the HGF ligand (e.g., ficlatuzumab).[51] It has been known for a long time that the c-MET gene expression and serum HGF levels are significantly increased in PDAC.[52,53] However, cMET upregulation worsens patient survival, as it increases resistance to gemcitabine, promotes tumor cell motility, and the secretion of angiogenic factors, which ultimately affect disease progression. The preclinical data indicate that the most effective results can be achieved by simultaneously blocking the ligand and the receptor in combination with chemotherapy.

Our further goal was to explore the changes in biological processes in pancreatic carcinoma. According to our analysis, altered metabolic processes, cell cycle alterations, and growth became very pronounced. However, other processes that are also crucial for the functioning of the tumor have also changed, such as apoptosis, the response to stimuli, or alteration in regulatory mechanisms.

In our study we identified seven genes that could serve as potential therapeutic targets for pancreatic carcinoma. The fact that these genes share excessive similarity with previous findings validates the strength of our discovery. The identified biological processes are well-known characteristics of tumor and provide further evidence of the importance of our identified genes in the disease.

The advantage of our analysis is that we integrated the results of multiple previous expression cohorts to examine the expression changes observed in pancreatic carcinoma, which was not possible in preceding studies. Nevertheless, we must mention a few limitations of our study. Since we combined different studies, not all platforms had expression values for all genes. Thus, we had to omit analysis of some genes which might have a pathogenic role in the development of pancreatic carcinoma. A second limitation is that although we collected both overall survival and relapse-free survival (RFS) data, only a fraction of tumors had RFS data – for this reason, our analysis was restricted to overall survival.

In conclusion, the expression level of a significant proportion of genes varies considerably in pancreatic cancer and these genes impact a wide range of biological processes. A number of genes previously recognized as highly expressed were confirmed by the thorough characterization of the most overexpressed genes in ductal pancreatic adenocarcinomas, and it is now clear which genes have the greatest promise for further study. We found that the PI3K-AKT signaling pathway plays a central role, including the MET gene, therapeutic targeting of which may have clinical benefits. We analyzed a patient cohort that had a sufficient size to enable a robust comparison and ranking of all potentially relevant genes.

AUTHOR CONTRIBUTIONS

Both authors wrote the manuscript. B.G. designed the research. M.P. performed the research. Both authors analyzed the data.

ACKNOWLEDGMENTS

The authors acknowledge the support of ELIXIR Hungary (www.bioinformatics.hu). The authors thank Ms. Viktoria Lakatos for the careful English editing of the manuscript.

FUNDING INFORMATION

This project was supported by the National Research, Development and Innovation Office (PharmaLab, RRF-2.3.1-21-2022-00015 and TKP2021-NVA-15).

CONFLICT OF INTEREST STATEMENT

The authors declared no competing interests for this work.

ETHICS APPROVAL

No ethics approval needed.

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. Hackeng WM, Hruban RH, Offerhaus GJA, Brosens LAA. Surgical and molecular pathology of pancreatic neoplasms. *Diagn Pathol*. 2016;11(1):47.
3. Vincent A, Herman J, Schulick R, Hruban RH, Goggins M. Pancreatic cancer. *Lancet*. 2011;378(9791):607-620.
4. Mario C, Marilisa F, Kryssia IRC, et al. Epidemiology and risk factors of pancreatic cancer. *Acta Biomed*. 2018;89(Suppl. 9):141-146.
5. Bengtsson A, Andersson R, Ansari D. The actual 5-year survivors of pancreatic ductal adenocarcinoma based on real-world data. *Sci Rep*. 2020;10(1):16425.
6. Distler M, Rückert F, Hunger M, et al. Evaluation of survival in patients after pancreatic head resection for ductal adenocarcinoma. *BMC Surg*. 2013;13(1):12.
7. Vareedayah AA, Alkaade S, Taylor JR. Pancreatic adenocarcinoma. *Mo Med*. 2018;115(3):230-235.
8. Klimstra DS. Nonductal neoplasms of the pancreas. *Mod Pathol*. 2007;20(1):S94-S112.
9. Meng Q, Shi S, Liang C, et al. Diagnostic and prognostic value of carcinoembryonic antigen in pancreatic cancer: a systematic review and meta-analysis. *OncoTargets Ther*. 2017;15(10):4591-4598.
10. Khomiak A, Brunner M, Kordes M, et al. Recent discoveries of diagnostic, prognostic and predictive biomarkers for pancreatic cancer. *Cancer*. 2020;12(11):E3234.
11. Conroy T, Desseigne F, Ychou M, et al. FOLFIRINOX versus gemcitabine for metastatic pancreatic cancer. *N Engl J Med*. 2011;364(19):1817-1825.
12. Orth M, Metzger P, Gerum S, et al. Pancreatic ductal adenocarcinoma: biological hallmarks, current status, and future perspectives of combined modality treatment approaches. *Radiat Oncol*. 2019;14(1):141.
13. Kelley RK, Ko AH. Erlotinib in the treatment of advanced pancreatic cancer. *Biol Targets Ther*. 2008;2(1):83-95.
14. Takiguchi S, Inoue K, Matsusue K, Furukawa M, Teramoto N, Iguchi H. Crizotinib, a MET inhibitor, prevents peritoneal dissemination in pancreatic cancer. *Int J Oncol*. 2017;51(1):184-192.
15. Govindarajan R, Duraiyan J, Kaliyappan K, Palanisamy M. Microarray and its applications. *J Pharm Bioallied Sci*. 2012;4(Suppl. 2):S310-S312.
16. Kukurba KR, Montgomery SB. RNA sequencing and analysis. *Cold Spring Harb Protoc*. 2015;2015(11):951-969.
17. Trevino V, Falciani F, Barrera-Saldaña HA. DNA microarrays: a powerful genomic tool for biomedical and clinical research. *Mol Med*. 2007;13(9–10):527-541.
18. Stark R, Grzelak M, Hadfield J. RNA sequencing: the teenage years. *Nat Rev Genet*. 2019;20(11):631-656.
19. Amin MB, Greene FL, Edge SB, et al. The Eighth Edition AJCC Cancer Staging Manual: continuing to build a bridge from a population-based to a more 'personalized' approach to cancer staging. *CA Cancer J Clin*. 2017;67(2):93-99.
20. Cox DR. Regression models and life-tables. *J R Stat Soc Ser B Methodol*. 1972;34(2):187-220.
21. Lánczky A, Győrffy B. Web-based survival analysis tool tailored for medical research (KMplot): development and implementation. *J Med Internet Res*. 2021;23(7):e27633.
22. McVeigh TP, Kerin MJ. Clinical use of the oncotype DX genomic test to guide treatment decisions for patients with invasive breast cancer. *Breast Cancer (Dove Med Press)*. 2017;9:393-400.
23. Loeb S, Catalona WJ. The prostate health index: a new test for the detection of prostate cancer. *Ther Adv Urol*. 2014;6(2):74-77.
24. Mezquita L, Auclin E, Ferrara R, et al. Association of the Lung Immune Prognostic Index with immune checkpoint inhibitor outcomes in patients with advanced non-small cell lung cancer. *JAMA Oncol*. 2018;4(3):351-357.
25. Maere S, Heymans K, Kuiper M. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*. 2005;21(16):3448-3449.
26. Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*. 2003;13(11):2498-2504.
27. Ma F, Li X, Fang H, Jin Y, Sun Q, Li X. Prognostic value of ANXA8 in gastric carcinoma. *J Cancer*. 2020;11(12):3551-3558.
28. Gou R, Zhu L, Zheng M, et al. Annexin A8 can serve as potential prognostic biomarker and therapeutic target for ovarian cancer: based on the comprehensive analysis of annexins. *J Transl Med*. 2019;17(1):275.
29. Hata H, Tatemichi M, Nakadate T. Involvement of annexin A8 in the properties of pancreatic cancer. *Mol Carcinog*. 2014;53(3):181-191.
30. Che D, Wang M, Sun J, et al. KRT6A promotes lung cancer cell growth and invasion through MYC-regulated pentose phosphate pathway. *Front Cell Dev Biol*. 2021;9:694071.
31. Raman P, Maddipati R, Lim KH, Tozeren A. Pancreatic cancer survival analysis defines a signature that predicts outcome. *PloS One*. 2018;13(8):e0201751.
32. Sharen G, Peng Y, Cheng H, Liu Y, Shi Y, Zhao J. Prognostic value of GLUT-1 expression in pancreatic cancer: results from 538 patients. *Oncotarget*. 2017;8(12):19760-19767.
33. Pizzi S, Porzionato A, Pasquali C, et al. Glucose transporter-1 expression and prognostic significance in pancreatic carcinogenesis. *Histol Histopathol*. 2009;24(2):175-185.
34. Achalandabaso Boira M, Di Martino M, Gordillo C, Adrados M, Martín-Pérez E. GLUT-1 as a predictor of worse prognosis in pancreatic adenocarcinoma: immunohistochemistry study showing the correlation between expression and survival. *BMC Cancer*. 2020;20(1):909.
35. Nocentini A, Capasso C, Supuran CT. Small-molecule CD73 inhibitors for the immunotherapy of cancer: a patent and literature review (2017–present). *Expert Opin Ther Pat*. 2021;31(10):867-876.
36. Hemmings BA, Restuccia DF. PI3K-PKB/Akt pathway. *Cold Spring Harb Perspect Biol*. 2012;4(9):a011189.
37. Ma X, Thapi D, Yan X, She QB, Rosales N, Spriggs D. Muc16 carboxyl portion expression alternates PI3K/Akt and EGFR/HER2/ERK signal pathways in human ovarian cells. *Cancer Res*. 2008;68(9_Suppl):3435.

38. Hu H, Wang F, Wang M, et al. FAM83A is amplified and promotes tumorigenicity in non-small cell lung cancer via ERK and PI3K/Akt/mTOR pathways. *Int J Med Sci*. 2020;17(6):807-814.
39. Zhou L, Jia S, Chen Y, et al. The distinct role of CD73 in the progression of pancreatic cancer. *J Mol Med (Berl)*. 2019;97(6):803-815.
40. Chen S, Huang J, Liu Z, Liang Q, Zhang N, Jin Y. FAM83A is amplified and promotes cancer stem cell-like traits and chemoresistance in pancreatic cancer. *Oncogenesis*. 2017;6(3):e300.
41. Liu L, Xu HX, Wang WQ, et al. Serum CA125 is a novel predictive marker for pancreatic cancer metastasis and correlates with the metastasis-associated burden. *Oncotarget*. 2016;7(5):5943-5956.
42. Haridas D, Chakraborty S, Ponnusamy MP, et al. Pathobiological implications of MUC16 expression in pancreatic cancer. *PloS One*. 2011;6(10):e26839.
43. Chen Q, Pu N, Yin H, et al. CD73 acts as a prognostic biomarker and promotes progression and immune escape in pancreatic cancer. *J Cell Mol Med*. 2020;24(15):8674-8686.
44. Nevedomskaya E, Perryman R, Solanki S, Syed N, Mayboroda OA, Keun HC. A systems oncology approach identifies NT5E as a key metabolic regulator in tumor cells and modulator of platinum sensitivity. *J Proteome Res*. 2016;15(1):280-290.
45. Tuan NM, Lee CH. Role of anillin in tumour: from a prognostic biomarker to a novel target. *Cancer*. 2020;12(6):E1600.
46. Ma Z, Zhou Z, Zhuang H, et al. Identification of prognostic and therapeutic biomarkers among FAM83 family members for pancreatic ductal adenocarcinoma. *Dis Markers*. 2021;2021:6682697.
47. Grant S. FAM83A and FAM83B: candidate oncogenes and TKI resistance mediators. *J Clin Invest*. 2012;122(9):3048-3051.
48. Aithal A, Rauth S, Kshirsagar P, et al. MUC16 as a novel target for cancer therapy. *Expert Opin Ther Targets*. 2018;22(8):675-686.
49. Ghalamfarsa G, Kazemi MH, Raoofi Mohseni S, et al. CD73 as a potential opportunity for cancer immunotherapy. *Expert Opin Ther Targets*. 2019;23(2):127-142.
50. Organ SL, Tsao MS. An overview of the c-MET signaling pathway. *Ther Adv Med Oncol*. 2011;3(1 Suppl):S7-S19.
51. Pothula SP, Xu Z, Goldstein D, Pirola RC, Wilson JS, Apte MV. Targeting HGF/c-MET axis in pancreatic cancer. *Int J Mol Sci*. 2020;21(23):E9170.
52. Pothula SP, Xu Z, Goldstein D, et al. Hepatocyte growth factor inhibition: a novel therapeutic approach in pancreatic cancer. *Br J Cancer*. 2016;114(3):269-280.
53. Zhu GH, Huang C, Qiu ZJ, et al. Expression and prognostic significance of CD151, c-met, and integrin alpha3/alpha6 in pancreatic ductal adenocarcinoma. *Dig Dis Sci*. 2011;56(4):1090-1098.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Posta M, Győrffy B. Analysis of a large cohort of pancreatic cancer transcriptomic profiles to reveal the strongest prognostic factors. *Clin Transl Sci*. 2023;16:1479-1491. doi:10.1111/cts.13563