Identification of early diagnosis markers of pancreatic ductal adenocarcinoma (PDAC) using publicly available transcriptomic tumor and blood sample data

Aristeidis Sionakidis (✉ A.Sionakidis@sms.ed.ac.uk)  
University of Edinburgh

Panagiotis Nikolaos Lalagkas  
University of Massachusetts

Andigoni Malousi  
Aristotle University of Thessaloniki

Ioannis S. Vizirianakis  
Aristotle University of Thessaloniki

Research Article

Keywords: PDAC, gene expression, liquid biopsies, pharmacogenomics, precision medicine

Posted Date: November 7th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-2172022/v2

License: © This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Identification of early diagnosis markers of pancreatic ductal adenocarcinoma (PDAC) using publicly available transcriptomic tumour and blood sample data

Aristeidis Sionakidis1*, Panagiotis Nikolaos Lalagkas2, Andigoni Malousi3, Ioannis S. Vizirianakis4,5*

1Institute of Genetics and Cancer, University of Edinburgh, Scotland, UK
2Department of Biological Sciences, University of Massachusetts, USA
3Laboratory of Biological Chemistry, School of Medicine, Aristotle University of Thessaloniki, Greece
4Laboratory of Pharmacology, School of Pharmacy, Aristotle University of Thessaloniki, Thessaloniki, Greece
5Department of Life and Health Sciences, University of Nicosia, Nicosia, Cyprus

Keywords: PDAC; gene expression; liquid biopsies; pharmacogenomics; precision medicine;

*Corresponding authors: Aristeidis Sionakidis, A.Sionakidis@sms.ed.ac.uk; Ioannis S. Vizirianakis, ivizir@pharm.auth.gr
Abstract
Pancreatic ductal adenocarcinoma (PDAC) is the most frequently diagnosed form of pancreatic cancer worldwide. PDAC is associated with poor survival rate mainly due to the disease being usually diagnosed at late stages. Publicly available gene expression data from 10 studies with tumour tissue (448 samples) and/or blood samples (128 samples) from PDAC patients were pooled together and analysed for the identification of stage-specific (American Joint Committee for Cancer, AJCC staging) and global diagnostic markers. Validation of markers was performed using Cancer Genome Atlas (TCGA) PDAC expression data. Differential gene expression analysis was carried out to compare tumour and normal samples (stage-specific tissue samples vs. normal tissue samples and stage-agnostic blood samples vs. normal blood samples). Active subnetwork search and miRNA enrichment analysis were used to identify enriched gene networks and miRNA interactions. We identified 820 consistently deregulated genes in tissue samples of all stages and blood samples. Active subnetwork analysis revealed enriched ribosome, proteasome, adherens junction and cell cycle pathways across all stages and blood samples suggesting biological plausibility. Stage-specific enriched miRNAs with diagnostic potential were also identified (miR-21, miR-29, miR-124, miR-30, for stages 1-4 respectively). Extensive gene expression deregulation was found in all tumor stages with significant overlap. Additionally, miRNA contribution to PDAC pathology may be important and probably mediated by distinct miRNAs in each stage of PDAC. We therefore present a list of markers and miRNAs that could potentially act as a diagnostic tool for early detection of PDAC onset.
Abbreviations

AJCC: American Joint Committee for Cancer

ATC: Anatomic Therapeutical Chemical

COSMIC: Catalogue Of Somatic Mutations In Cancer

DEG: Differentially Expressed Gene

DGEA: Differential Gene Expression Analysis

DRS: Down-Regulated Score

FDR: False Discovery Rate

GEO: Gene Expression Omnibus

GO-BP: Gene Ontology – Biological Processes

GO-CC: Gene Ontology – Cellular Components

GO-MF: Gene Ontology – Molecular Functions

kNN: k-Nearest Neighbor

KEGG: Kyoto Encyclopedia of Genes and Genomes

miRNA: microRNA

NCBI: National Center for Biotechnology Information

PDAC: Pancreatic Ductal Adenocarcinoma

PIN: Protein Interaction Network
Novelty

In this study we: a) developed an analytic pipeline by gathering data from PDAC studies with clinical gene expression data, normalised them within each study and then joined in a common dataset to increase sample size and the statistical power of the analysis; b) identified a set of gene expression markers that can be found in both tumor tissue samples of all stages and blood samples of PDAC which can distinguish tumors from normal samples; c) validate this set of markers in external data from TCGA and d) report stage-specific diagnostic miRNA markers that can serve as promising candidates particularly for the detection of early-stage disease, as well as pharmacogenomic interactions that could guide personalised therapeutic interventions.
Introduction

Pancreatic ductal adenocarcinoma (PDAC) is one of the most lethal carcinomas with almost as many cases (495,773) as deaths (466,003) in 2020 and a projection to be the second leading cause of cancer-related deaths by 2030\(^1\). It is the most common and aggressive malignancy of pancreas, accounting for at least 95% of the total pancreatic cancers \(^1\). Despite the advances in cancer research and molecular biology, the 5-year overall survival rate remains extremely low (~11\%)\(^2\). This might be due to the often late-stage diagnosis as a result of the largely asymptomatic clinical course and the lack of reliable diagnostic biomarkers\(^3\). The molecular heterogeneity of tumors directly affects treatment outcomes and thus, the application of novel analytical methods to detect molecular changes that favor the growth of cancer cells and tumor progression present promise in the clinical setting\(^4\).

Transcriptomics have been extensively used in PDAC research. Collisson et al.\(^5\), Moffitt et al.\(^6\) and Bailey et al.\(^7\) identified PDAC subtypes with prognostic and biological relevance, while other studies used transcriptomics to identify differentially expressed genes in PDAC with diagnostic and prognostic potential. Besides protein-coding genes, microRNAs (miRNAs) have also been studied for their role as molecular biomarkers. Interestingly, the most commonly reported miRNA is the miR-21 which has been found to be upregulated in PDAC patients and associated with poor survival\(^8,9\). miR-21 has also been linked to acute pancreatitis\(^10\).

To date, no reliable diagnostic biomarker for early-stage detection of PDAC has been approved for use in routine clinical practice and considering that especially for PDAC, the survival rate is related to disease stage at the time of diagnosis, the discovery of novel
diagnostic biomarkers is an urgent need. Here we describe an analytical methodology using publicly available gene expression data from clinical PDAC samples referring to tumor as well as biopsies from blood samples. We hypothesized that there is a set of genes that are consistently aberrated across PDAC samples and provide a gene signature with diagnostic potential that is consistently deregulated across all tumor stages. We further demonstrate the role of miRNA in PDAC outcomes and progression and provide a summary of pharmacogenomic interactions that could be particularly useful for the development of personalised treatment approaches.

**Methods**

**Database Querying**

We searched the NCBI GEO database for PDAC expression datasets of human samples (see Additional file 8, section 1.1). We filtered out studies where patients (or the corresponding samples) have received pharmacological treatment prior to measurement and studies where human cell lines, cultures or xenografts were used instead of patient samples. We also excluded studies involving samples from very particular categories of patients (e.g. particular mutation carriers) and tissue biopsy studies which did not have any information on the stage of the tumors. Regarding liquid biopsy studies, we used the same exclusion criteria with the exception of tumor stage filtering. 10 studies\(^{11-20}\) satisfied our criteria. Overall, samples from 7 studies\(^{12,14,16-20}\) were used for the tumor tissue analysis and samples from 4 studies\(^{11,13,15,16}\) were used for the blood samples analysis.
**Data aggregation and preprocessing**

Preprocessing and analysis of the gene expression data, was performed in R (version 4.1.1). Probes with more than 25% missing values were excluded from the analysis. The rest of the missing expression values were imputed using k-Nearest Neighbor (kNN, $k = 10$) imputation. For all tumor tissue studies, only samples which were taken from the tumor site and normal adjacent tissue were kept. Tumor stage was determined using the American Joint Committee on Cancer (AJCC) system. We consistently annotated the gene expression matrices with NCBI Entrez identifiers and averaged rows which mapped to the same unique identifier. The expression values in each study's matrix were normalised by conversion to standard scores, gene-wise (see Supplementary Figures 1-6 in Additional file 9). The normalised expression matrices from the tumor tissue studies were then unified into a global tumor tissue expression matrix and, similarly the normalised expression matrices from the blood sample studies into a global blood sample expression matrix. PDAC data from The Cancer Genome Atlas (TCGA) were downloaded for validation purposes: Transcripts Per Million (TPM) expression values were log2-transformed ($\log_2(\text{TPM} + 1)$) and converted to standard scores in a gene-wise manner.

**Differential Gene Expression Analysis**

Differential Gene Expression Analysis (DGEA) was performed by fitting a linear model for the expression values of each gene, adjusted for AJCC classification (normal, stage 1, stage 2, stage 3, stage 4) and study, using the *limma* package in R. We performed DGEA twice: once on the tumor tissue samples and once on the blood samples. The models corresponding to the blood samples were only adjusted for sample type (normal, tumor). Seven comparisons were carried out among tissue samples: 1) stage 1 vs. normal, 2) stage 2 vs. normal, 3) stage 4 vs. normal.
3 vs. normal, 4) stage 4 vs. normal, 5) stages 1 & 2 vs. stages 3 & 4, 6) stage 1 vs. stage 2, and 7) stage 1 vs. stage 4. Blood sample comparisons were made between tumor and normal samples. We used the Benjamini-Hochberg p-value multiple testing adjustment method and set the False Discovery Rate (FDR) guarantee to 0.05. We selected genes with an adjusted p-value below 0.05 for downstream analysis. In typical limma output, the log-fold change statistic reported for each gene for a particular comparison represents the numerical difference between the coefficients estimated for the two contrasts e.g. tumor - normal. In our case, since the data have been normalised prior to being subjected to DGEA, the corresponding statistic represents the numerical difference between the coefficients of the two contrasts measured in standard deviations from the gene’s mean expression across samples (s.d. units).

**Active Subnetwork and miRNA enrichment Analysis**

For the identification of enriched PDAC-related biological networks we used the pathfindR package. The approach offered by the package incorporates active subnetwork identification and subsequent enrichment analysis using the identified active subnetworks (see Additional file 8, section 1.3). For a given list of significantly differentially expressed genes, an active subnetwork is defined as a group of interconnected genes in a protein-protein interaction network (PIN) that predominantly consists of significantly differentially expressed genes. We selected the BioGRID PIN and performed enrichment analysis within all gene sets available in the package: BioCarta, Gene Ontology - Biological Processes (GO-BP), Gene Ontology - Cellular Components (GO-CC), Gene Ontology - Molecular Functions (GO-MF), Kyoto Encyclopedia of Genes and Genomes (KEGG) and Reactome. Pathways with p.adj<0.05
were selected to produce grouped results and highlight key biological processes using hierarchical clustering.

miRNA enrichment analysis was performed online using Mienturnet\textsuperscript{23}, and setting the database of gene-miRNA interactions to be miRTarBase. The analysis applied on the list of significantly (p.adj<0.05) differentially expressed genes derived from the comparisons of the four different stages to normal tissue samples and the comparison of tumors to normal samples from liquid biopsies (5 lists in total). We removed the miRNAs and miRNA host-genes with p.adj<0.05 before submitting the lists to the tool so that they do not affect the enrichment result. Enriched miRNAs with p.adj<0.05 were considered to be significantly enriched.

**Drug-gene interactions**

Cancer driver status and mutation frequency information on genes was downloaded from COSMIC (Catalogue Of Somatic Mutations In Cancer, accessed: February 22, 2022). Interactions between approved drugs and genes were downloaded from the DrugBank database (accessed February 21, 2022). Drugs were classified into categories using the Anatomic Therapeutic Chemical (ATC) classification system. We mapped our lists of differentially expressed genes from all stages and blood samples to approved drugs with which pharmacological interaction is documented. Circular plots with pharmacogenomic links and additional annotation were then generated using Circos.
Results

Search results

After searching the NCBI GEO database, we identified 132 candidate GEO series for our analysis with PDAC tissue and/or blood samples. Out of them, ten series (GSE18670, GSE21501, GSE42952, GSE49641, GSE62165, GSE62452, GSE74629, GSE84219, GSE102238, and GSE125158) satisfied our criteria. Six out of them (GSE21501, GSE42952, GSE62165, GSE62452, GSE84219, GSE102238) contained 556 tissue samples, while three out of them (GSE49641, GSE74629, GSE125158) contained 116 blood tissue samples of healthy or PDAC patients. Notably, one series (GSE18670) contained 24 samples, including samples from both pancreatic tissue and blood. After preprocessing was completed, 448 tissue samples (318 tumor, 130 normal) from 7 studies and 122 blood samples (77 tumor, 45 normal) from 4 studies were included in our analysis (see Additional file 1).

Gene expression deregulation emerges from early stages and persists in later stages

The global gene expression matrix of tumor tissue samples includes 448 samples (318 tumors, 130 normal) and 25699 genes. Differentially expressed genes (DEGs) in the seven types of comparisons we conducted using DGEA were: 1) stage 1 vs. normal (9,947 DEGs), 2) stage 2 vs. normal (16,328 DEGs), 3) stage 3 vs. normal (10,452 DEGs), 4) stage 4 vs. normal (8,340 DEGs), 5) stages 1 & 2 vs. stages 3 & 4 (one DEG), 6) stage 1 vs. stage 2 (one DEG) and 7) stage 1 vs. stage 4 (no DEGs); see Figure 1 and Additional file 2. The gene overlaps and differences between stage-specific DEG lists were identified. Sets of genes were then created with genes which were not overlapping between lists of different comparisons combined with the overlapping genes which were significantly deregulated towards
different directions (up-regulated/down-regulated in one list and down-regulated/up-regulated in another). We called these genes "dichotomizers" and the results are summarized in Supplementary Table 1 and Additional file 3.

There is significant overlap regarding DEGs across stages. Most of the genes are deregulated (up-regulated or down-regulated) towards the same direction (parallel overlap) when different stages are compared. Overall, there is a 5928-DEG overlap across all four stages (see Figure 2). Especially when stages 1 and 2 are compared, 10,618 out of the 10,891 stage 1 DEGs (97.5%) are also found to be differentially expressed towards the same direction in stage 2 samples (273 dichotomizers). The respective overlaps in later stages are also very high, however, the number of dichotomizers is larger than 2,000 in each comparison.

820 genes are consistently deregulated in tissue and blood sample biopsies

We performed DGEA on 122 blood samples (77 tumors, 45 normal) and 31581 genes. Tumor samples were compared with normal samples and 7145 DEGs (p.adj<0.05, see Additional file 2) were identified. 3708 of these genes are unique for the blood samples when results are compared to those of tissue samples of all stages. The overlap of DEGs across tissue and blood samples amounts to 1330 genes (see Figure 2), of which 820 are consistently up- or down-regulated in all tumor stages and blood samples (see Additional file 4).

The list of 820 genes was compared with the lists published by Collisson et al. (62 genes)⁵, Moffitt et al. (50 genes)⁶ and Bailey et al. (614 genes)⁷. There was no overlap with the lists by Bailey et al., and Collisson et al., but there was a 6-gene overlap with the list by Moffitt et al.: AIM2, HK2, HMMR, S100P, PMAIP1, and CEACAM6. Additionally, when our list is compared...
Figure 1. Volcano plots for the comparisons of samples of different stages against normal samples (tissue samples). The negative logarithm of the adjusted p-value of each gene is plotted on the y-axis and the expression change (tumor/normal) of each gene (measured in standard deviations) is plotted on the x-axis. Vertical and horizontal dashed lines define expression change and p-value thresholds to indicate the most notable results. Expression change threshold was set to 1 and adjusted p-value threshold was set to 0.05. Depending on the partition of the plot each gene falls into, the color of the dot changes as indicated by the legend at the top. Gray: NS (not significant), pink: passes expression change threshold, purple: passes p-value threshold, red: passes both thresholds.
Figure 2. Venn diagram of significantly differentially expressed genes. 4598 overlapping genes are found across all stages of tumor tissue samples (and are not found in liquid biopsies). When blood samples are also considered, the overlap across all samples (regardless of origin) consists of 1330 genes.

to the PDAC prognostic signature (clinical outcome prognosis, 36 genes) published by Haider et al.\textsuperscript{24}, we find an overlap of two genes: \textit{CNNM3}, \textit{QDPR}.

Validation of our signature

We broke our signature (820 genes) down to the down-regulated genes and up-regulated genes and derived two metascores: 1) \textbf{DRS (Down-Regulated Score)}: the mean expression of the down-regulated genes in a given sample and 2) \textbf{URS (Up-Regulated Score)}: the mean expression for up-regulated genes. We then assessed how these two scores are associated with tumor stage and/or sample type in our data (see Figure 3). Independent sample t-tests showed that both scores are significantly different in samples of any stage when compared to normal samples and the effect size is large as estimated by Cohen's d. The association of
both scores with vital status (binary variable: dead/alive) in TCGA expression data of 183 PDAC samples was also statistically

Figure 3. Boxplots of metascores in our data and TCGA data. Top: Boxplots of the mean metascores across sample types and tumor stages in our data (down-regulated genes to the left, up-regulated genes to the right). Bottom: Mean metascore comparisons between "dead" and "alive" subjects in TCGA data (down-regulated genes to the left, up-regulated genes to the right).
significant (**DRS**: $t = 2.2$, $p = 0.03$, **URS**: $t = 2.9$, $p = 0.004$) with a small effect size as estimated by Cohen’s d (**DRS**: $d = -0.33$, **URS**: $d = 0.43$, $|d| < 0.5$).

**A miRNA-guided map to disease progression**

We investigated miRNA involvement in the disease in two ways: a) by identifying differentially expressed miRNAs in tissue and blood samples (see Additional file 5 and Supplementary Table 2) and b) by running miRNA enrichment analysis on the different lists of DEGs (excluding miRNAs and miRNA host genes) using an online tool (Mienturnet\textsuperscript{23}). Unlike the DEGs identified by different DGEA comparisons, the lists of differentially expressed miRNAs do not share significant overlap across stages and type of samples (see Supplementary Figure 8; Additional file 9). 35 out of the 67 stage 2 miRNAs are unique for stage 2. Similarly, 45 out of the 53 blood sample miRNAs are uniquely found to be differentially expressed in blood samples. Interestingly, there is only one miRNA that is a unique overlap between stage 1 samples and blood samples: miR-21. Our results revealed a higher expression in stage 1 tumors versus normal samples (by 0.79 standard deviations from the mean, $p\text{adj}=0.048$) and a higher expression in PDAC blood samples versus normal blood samples (by 0.82 standard deviations from the mean, $p\text{adj}=0.0001$).

According to Mienturnet results (see Additional file 6) miR-192-5p, miR-200b-3p, miR-200c-3p and miR-429 are consistently found to be significantly enriched across all four stages (Figure 4). Among these four, miR-192-5p interacts with most DEGs from the respective stage DEG lists (around 500 DEGs in each stage), followed by miR-200c-3p and miR-200b-3p (just above 100 DEGs) and miR-429 (around 100 DEGs). miR-375 is enriched in stages 2, 3 and 4, thus indicating a potential role in the progression of the disease.
Interestingly, miR29b-3p is enriched in stage 2 (p.adj=0.0036) and 200 interactions with our stage 2 DEGs are predicted. Furthermore, the corresponding gene (MIR29B2) and its host gene (MIR29B2CHG) are both found to be differentially expressed (MIR29B2: -0.70 standard deviations from the mean, p.adj=0.0005, MIR29B2CHG: -1.01 standard deviations from the mean, p.adj=2.2⋅10^{-7}). Other notable results include miR-124-3p in stage 3 which is the top enriched miRNA (p.adj=0.0001) with more than 700 interactions in the list of DEGs for stage 3 tumors. This miRNA is also found to be differentially expressed in this stage (-0.88 standard deviations from the mean, p.adj=0.002), which suggests that its down-regulation may contribute to progression in later stages. Similarly, miR-30a-5p is enriched in stage 4 (p.adj=0.001) with more than 300 interactions and is also found to be differentially expressed at stage 4 (-1.11 standard deviations from the mean, p.adj=0.049). miR-124-3p is also enriched in stage 4 (p.adj=0.0006). Regarding blood sample miRNA enrichment results

Figure 4. Top 10 Mienturnet miRNA enrichment results across tumor stages. Each bar chart illustrates the number of interactions, with genes of our stage-specific DEG lists, that available in miRTarBase for a predicted miRNA. miRNAs that are consistently significantly enriched and/or differentially expressed and/or whose host genes are differentially expressed are annotated as shown at the legend on the top. miRNAs which are not statistically significantly enriched (p.adj >0.05) are underlined with a gray line.
(see Supplementary Figure 9), no overlap of enriched miRNAs was found with tumor tissue results. Among the enriched miRNAs, miR-193B-3p stands out (around 300 interactions, \( p\text{-adj}=1.26 \times 10^{-5} \)), because its host gene \((MIR193BHG)\) is down-regulated in tumor blood samples (-0.53 standard deviations from the mean, \( p\text{-adj}=0.045 \)). When we provided the tool with the list of 820 genes we identified at the previous step, only four miRNAs were statistically significantly enriched: miR-375 (\( p\text{-adj}=1.2 \times 10^{-5} \)), miR-146a-5p (\( p\text{-adj}=2.6 \times 10^{-5} \)), miR-192-5p (\( p\text{-adj}=0.0008 \)) and miR-215-5p (\( p\text{-adj}=0.001 \)); see Supplementary Figure 9.

**Molecular pathway homogeneity across stages reveals global tumor progression driving processes**

Results from active subnetwork enrichment analysis revealed major pathway overlaps across stages and types of samples, regardless of the gene set that was used. When focusing on results from the KEGG set, we find processes related with adherens junction, cell cycle, proteasome, ribosome, spliceosome, proteoglycans in cancer and glycolysis/gluconeogenesis to be globally significantly enriched (see Additional file 7). When the top five common KEGG terms from stage 1 and stage 2 tumors are compared, four sugar metabolism pathways (pentose phosphate, fructose and mannose phosphate, galactose metabolism and glycolysis/gluconeogenesis pathways) and the citrate cycle appear to be identically deregulated (Figure 5; see also Supplementary Figure 9). In terms of cellular components, enrichment results are again very similar between the stages; the ribosome, proteasome, spliceosome and adherens junction components stand out again. Most notably, a great number of ribosomal pathway components are down-regulated, while proteasomal pathway components are up-regulated.
In terms of molecular functions in higher resolution, results from the Reactome gene set revealed that the ROBO-SLIT-mediated molecular pathways appear to be considerably enriched in all stages (particularly stages 1, 2 and 3; see Supplementary Figures 10, 11). The wide ribosomal pathway repression is notable at higher resolution here, as processes like formation of free 40S subunits, translation initiation and elongation are heavily down-regulated (see Supplementary Figure 11 and Additional file 7). Other similar down-regulated processes include rRNA processing and nonsense-mediated decay. In contrast with the aforementioned results, processes such as proteasome-mediated degradation and
deubiquitination are consistently found to be up-regulated in all stages of PDAC. Furthermore, mitosis promoting pathways and cell cycle regulation pathways were associated with gradually increasing fold-enrichment scores, indicating that tumors demonstrate more aggressive proliferation patterns in later stages.

**Genes in driver pathways represent promising targets for pharmacogenomics-based strategies**

Figure 6 illustrates the complex pharmacogenomic and miRNA interactions between genes from two consistently enriched KEGG pathways in our samples: "Cell cycle" and "Proteoglycans in cancer". Both pathways could serve as pools of targets for pharmacological interventions because of their contribution to PDAC pathology. The suggested interactions can be further evaluated to test whether they can be harnessed for the treatment of unresectable PDAC tumors or as potential neoadjuvant treatment options to reduce tumor size prior to resection.

Genes that represent potential treatment target candidates (Figure 6) are: *SRC, MET* and *MAP2K1*. All three of these genes are cancer drivers (not pancreas-specific). *MET* is a receptor tyrosine kinase and a popular antineoplastic drug target, targeted by crizotinib and cabozantinib (shown here), and amivantamab, capmatinib, tepotinib (see Supplementary Figure 13). *SRC* is a proto-oncogene for a non-receptor tyrosine kinase which is heavily involved in several intracellular processes and is targeted by dasatinib (shown here) and tirbanibulin, a dermatological drug (see Supplementary Figure 13). Cabozantinib was successfully combined with gemcitabine in a Phase I trial and is currently being studied at more advanced clinical settings (Phase II study in combination with pembrolizumab: NCT05052723). On the other hand, dasatinib has failed in Phase II trials in...
Figure 6. Circos plot of antineoplastic and immuno-modulating drugs and miRNAs vs. genes from cell cycle and proteoglycans in cancer pathways (Stage 1). Drug-gene and miRNA-gene interactions from the DrugBank and miRTarBase databases respectively. A) Labels. 1. Drugs: Antineoplastic and immunomodulating drugs are drawn with black labels at the top right arc. 2. Genes: Genes from the KEGG Cell cycle pathway are drawn in purple at the bottom right arc (smaller gene track) and genes from the KEGG proteoglycans in cancer pathway are drawn in purple at the left side (larger gene track). 3. miRNA: Black labels on a small arc at the top of the plot represent miRNAs that interact with these genes. B) Plots. 1. Links: Red ribbons represent pharmacogenetic links between drugs and up-regulated genes. Blue ribbons indicate pharmacogenetic links between drugs and down-regulated genes. Links for genes with lower DGEA adjusted p-values and larger absolute expression changes are plotted on top of others and given darker colors. miRNA-gene interactions are drawn in purple ribbons. Darker colors and overlaid ribbons indicate lower DGEA adjusted p-values and higher absolute expression shifts in genes, and lower miRNA enrichment adjusted p-values for miRNAs. 2. Inner circle histogram: The inwards-oriented histogram (gray) illustrates the DGEA adjusted p-value for each gene. 3. Cancer drivers glyph track: Colored dots indicate global cancer drivers (orange) and pancreas-specific cancer drivers (green). 4. Connector track: Lines which connect gene labels to the corresponding glyphs. 5. Mutation frequency: The orange histogram in the outer layer gives an estimate of the COSMIC mutation frequency of each gene in pancreatic cancer samples. 6. miRNA degree histogram: A histogram that displays the number of genes (from our full list of genes) each miRNA interacts with.
combination with established protocols (FOLFOX or gemcitabine treatment) to induce remarkable improvements in patient survival. Finally, MAP2K1 is a dual-specificity kinase in the MAP signal transduction pathway, which is involved in growth, adhesion, differentiation among other processes, and is targeted by cobimetinib and trametinib (shown here), and solumetinib (see Supplementary Figure 13). Trametinib was not as successful in combination with gemcitabine and other drugs before, but more recent results of trametinib in combination with radiation and pembrolizumab adjuvant treatment showed a borderline improvement in median overall survival.

Additional genes of interest include up-regulated pancreatic cancer drivers KRAS, HIF1A and EP300. KRAS is a well-known proto-oncogene targeted by sotorasib (see Supplementary Figure 13), the first FDA-approved KRAS inhibitor which was recently introduced. Sotorasib is currently approved for KRAS mutation carriers with non-small cell lung carcinoma, however, KRAS is the most frequently mutated gene in PDAC samples too and therefore, sotorasib could be a candidate for neoadjuvant or adjuvant PDAC treatment. Genes that may not pharmacologically interact with approved drugs, such as EP300 and HIF1A, may be promising candidates for RNA interference approaches. Both these genes are up-regulated and are predicted targets of miR-429 (shown here) and therefore, enhanced miR-429 function could perhaps bring their expression levels back to normal.

**Discussion**

Progress in the early diagnosis of PDAC has been slow and incidence of the disease is gradually increasing, while outcomes are among the poorest in all cancers. There is therefore, a dire need for robust diagnostic biomarkers for early-stage diagnosis. We present
here a comprehensive gene expression analysis of 570 human PDAC samples (448 tissue biopsies, 122 liquid biopsies). We report lists of genes that are significantly deregulated in tumors of specific stages in comparison with normal samples and derive an 820-gene signature that is consistently and homodirectionally deregulated in tissue and blood sample biopsies. Regarding miRNAs, miR-21 has been previously reported to be up-regulated in PDAC and associated with decreased survival. It has also been linked to PDAC development and has been argued to be a promising PDAC biomarker. This result is reproduced in our analysis where miR-21 is found to be up-regulated in stage 1 tissue samples and blood samples, a property that makes it a promising early diagnostic biomarker.

Pooling together the results from our miRNA analytical pipeline, we propose a miRNA-driven map to PDAC progression with key miRNAs that can potentially serve as informative clinical biomarkers. miR-192-5p and miR-375 can be used as disease-specific all-stage biomarkers. It has been previously reported that miR-192-5p has a tumor-suppressing-role in PDAC and its down-regulation is associated with disease progression. This is further verified by the fact that miR-192-5p has 77 targets in our 820-gene signature, 52 of which (67.53%) are consistently up-regulated in all-stage tissue samples and blood samples. miR-375 over-expression and also down-regulation have both been previously linked to PDAC, but more recent results indicate that it is more consistently down-regulated in PDAC. Interestingly, 36 out of its 51 predicted targets in our 820-gene signature are consistently up-regulated. Overall, according to our results, it can be inferred that expression of both miR-192-5p and miR-375 is generally repressed in PDAC samples, regardless of stage and sample type.
Regarding stage-specific results, we suggest four tissue biopsy miRNA biomarkers: 1) miR-21 up-regulation could serve as a stage 1 biomarker, 2) miR-29b-3p down-regulation as a stage 2 biomarker, 3) miR-124-3p down-regulation as a stage 3 biomarker and 4) miR-30a-5p down-regulation as a stage 4 biomarker. We validated the potential of miR-21 as discussed previously, however the rest of the miRNAs were not found to be differentially expressed in blood samples, but that can also be due to the limited miRNAs in our DGEA input gene lists.

Additional blood biomarker candidates for stage 2 are miR-29c, miR-148b, miR-222 and miR-331 which are in parallel differentially expressed in stage 2 and blood samples (all down-regulated). miR-29c repression has been previously associated with PDAC progression. Similarly, miR-148b has been shown to act as a tumor suppressor in PDAC cell lines. In contrast, both miR-222 and miR-331 up-regulation has been linked to PDAC tumor progression and oncogenicity. For stage 3, miR-29c can again serve as a potential biomarker (down-regulated in blood and tissue samples) as well as miR-376c (down-regulated in blood and tissue samples), high expression of which has been linked to increased survival rates. No unique miRNA overlap was found for blood samples and stage 4 tissue samples.

Finally, we demonstrate that PDAC pathology is generally consistent across stages. Molecular pathways and cellular components that are found to be aberrant in early stages are also aberrant in later stages with only few differences. The pathway analysis highlights adherens junction, proteoglycans, the proteasome, the ribosome, and the spliceosome as some of the most deregulated processes. Proteasomal complexes have been previously reported to be highly heterogeneous between different tumors and cell lines, and partially
drive the tumor’s sensitivity to selective drugs such as bortezomib\textsuperscript{42}. The adherens junction has also been found to be enriched in PDAC tumors in other studies\textsuperscript{39} and alternative splicing diversity has been reported to be significantly reduced in pancreatic cancer in comparison to normal pancreas, particularly when tumors are of a later stage\textsuperscript{43}. The results indicate a consistent up-regulation of the catalytic step 2 spliceosome pathways and the stage 1-fold enrichment is 0.8 while the stage 4-fold enrichment rises to 1.1. Certain genes from these pathways represent targets of approved and more novel pharmaceutics as well as predicted targets of miRNAs and therefore, they could serve as the basis for the development of therapeutic strategies, either for neoadjuvant protocols or protocols for cases of unresectable PDAC.

There are certain limitations to our approach and the analytical methods we used. There is no gold-standard method to combine results from gene expression studies. Although various methods have been suggested, each carries its own limitations\textsuperscript{44}. Here, we normalised the samples of each study prior to joining them together, a transformation that is not guaranteed to retain the full original biological in the data\textsuperscript{44}. When conducting DGEA we adjusted the results for tumor stage, sample type (normal/tumor) and study of origin to account for batch effects. It is hoped that with increasing usage of RNA-Seq technologies and the development of more sophisticated batch correction methods, such limitations of comprehensive analysis of expression data will be less eminent. When a gene symbol was not available in a study’s expression matrix, all samples from that study end up having missing values in the final expression matrix with respect to that symbol. We did not remove these genes from our analysis, since the \textit{limma} package ignores the missing values during the calculation of the coefficients of each model. However, this inevitably results in coefficients that were derived
from different sample sizes across genes. Finally, comparing results from the separate analysis of tumor tissue samples and blood samples also requires caution when interpreting the overlaps and making inference about the translational potential of markers.

**Conclusion**

PDAC is a malignancy characterized by widespread molecular deregulation in terms of gene expression. We demonstrate here that the gene expression aberrations found in tumor samples when compared to normal samples are similar across stages of the disease. Furthermore, we show how particular tumor markers found in tissue biopsies can also be identified in blood samples (liquid biopsies) and derive a consistent signature of 820 PDAC-specific genes. Such markers are strong candidates for early diagnostic interventions with the potential of detecting the disease at early stages where tumors are operable and neoadjuvant therapeutic interventions may be beneficial for patients. We also describe here the potential role of miRNAs in the pathology and the progression of the disease and discuss their stage-specific diagnostic potential. Finally, we present interactions between our PDAC signature and identified miRNAs as well as clinically approved drugs and suggest plausible therapeutic interventions that could be of substantial benefit for PDAC patients. Evidently, the complexity of these relationships combined with the different clinical manifestation of PDAC across tumor stages will require careful planning regarding the therapeutic strategies that could be beneficial for PDAC patients. Overall, we studied the molecular complexity of PDAC, explored the gene expression and miRNA deregulation of the disease and provide a set of markers that are strong candidates for diagnostic and therapeutic interventions that could improve patient outcomes and increase survival.
Availability of data and materials

Preprocessing and analysis of the gene expression data, was performed in R (http://www.r-project.org – version 4.1.1). We used packages from both CRAN (http://cran.r-project.org/) and Bioconductor (version 3.13, http://www.bioconductor.org/). All code to reproduce the analysis is available at a GitHub repository (https://github.com/lalagkaspn/pdac_omics). Code for the preprocessing and analysis of the tumor tissue biopsy data is included in the "GSE_tumor_stage.R" script. Code for blood biopsy studies is included in the "GSE_blood_samples.R". The "Pathway_ontology_driver_genes.R" script contains code for the active subnetwork analysis of the differentially expressed genes and finally the "Circos.R" script contains the code for the preparation of all text files that required by the Circos software to produce pharmacogenomic plots. For every analytical process which included a random component (e.g., enrichment analysis), we used the random number generator version (RNGversion) "4.0.2" and random seed (123), for consistency and reproducibility. miRNA enrichment analysis was performed using Mienturnet, an online tool (http://userver.bio.uniroma1.it/apps/mienturnet/). The input lists for Mienturnet are available in the repository. Interactions between approved drugs and genes were downloaded from the DrugBank database (https://go.drugbank.com/). Drugs were classified into categories using the Anatomic Therapeutic Chemical (ATC) classification system (https://www.who.int/tools/atc-ddd-toolkit/atc-classification, August 2020 version). All outputs are available in the GitHub repository, apart from the pathway-specific pathfindR plots due to size.
References


Acknowledgements

This work was presented as poster in the recent EACR 2022 Congress – “Innovative Cancer Science: Translating Biology to Medicine”, 20-23 June 2022, Seville, Spain.

Funding

This work was supported by the Medical Research Council DTP in Precision Medicine (to A.S.), the University of Edinburgh (A.S.), Bodossaki Foundation and Onassis Foundation (A.S.), as well as interdepartmental public resources of Aristotle University of Thessaloniki, Greece.
Author information

Authors and Affiliations

Institute of Genetics and Cancer, University of Edinburgh, Scotland, UK, EH4 2XU

Aristeidis Sionakidis

Department of Biological Sciences, University of Massachusetts, Lowell, Boston, USA

Panagiotis-Nikolaos Lalagkas

Laboratory of Biological Chemistry, School of Medicine, Aristotle University of Thessaloniki, Thessaloniki, Greece

Andigoni Malousi

Laboratory of Pharmacology, School of Pharmacy, Aristotle University of Thessaloniki, Thessaloniki, Greece

Ioannis S. Vizirianakis

Department of Life and Health Sciences, University of Nicosia, Nicosia, Cyprus

Ioannis S. Vizirianakis

Contributions

All authors jointly designed the analysis plan. A.S. and P.N.L. performed the bioinformatic and statistical analysis. All authors interpreted the data together. A.S and P.N.L wrote the paper and all other authors contributed to revisions. I.S.V. coordinated the study. All authors read and approved the manuscript.

Corresponding author

Correspondence to Aristeidis Sionakidis, Ioannis S. Vizirianakis.
Ethics declaration

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors report no competing interests.

Supplementary Information
Additional file 1.xlsx. Table of study characteristics. A summary of the studies that were included in the analysis.

Additional file 2.xlsx. Results of Differential Gene Expression Analysis. 1. Results summary for all comparisons. 2. Stage 1 vs. normal samples. 3. Stage 2 vs. normal samples. 4. Stage 3 vs. normal samples. 5. Stage 4 vs. normal samples. 6. Stages 1 & 2 vs. Stages 3 & 4. 7. Stage 1 vs. Stage 4. 8. Stage 1 vs. Stage 2. 9. Liquid biopsy tumors vs. liquid biopsy normal.


Additional file 5.xlsx. Differentially expressed miRNAs and host genes. 1. Stage 1 miRNAs and host genes. 2. Stage 2 miRNAs and host genes. 3. Stage 3 miRNAs and host genes. 4. Stage 4 miRNAs and host genes.

Additional file 6.xlsx. miRNA enrichment analysis results. 1. Stage 1 miRNA enrichment results. 2. Stage 2 miRNA enrichment results. 3. Stage 3 miRNA enrichment results. 4. Stage 4 miRNA enrichment results. 5. Liquid biopsy miRNA enrichment results. 6. 820-gene signature miRNA enrichment results.

Additional file 7.zip. Active Subnetwork Enrichment Analysis results. Results from the analysis conducted using the pathfindR package.

Additional file 8.pdf Supplementary Methods. An extensive step-by-step description of the methodology that was used in the analysis.

Additional file 9.pdf Supplementary Figures and Tables. Additional figures and tables that support the results of this analysis.
Supplementary Files

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

- Additionalfile1.xlsx
- Additionalfile2.xlsx
- Additionalfile3.xlsx
- Additionalfile4.xlsx
- Additionalfile5.xlsx
- Additionalfile6.xlsx
- Additionalfile7.zip
- Additionalfile8.pdf
- Additionalfile9.pdf