ORIGINAL RESEARCH PAPER
Volume - 15 | Issue - 09 | September - 2026 | PRINT ISSN No. 2277 - 8179 | DOI : 10.36106/ijsr
INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH

NETWORK CENTRALITY AND SURVIVAL ASSOCIATIONS OF CONSENSUS HUB GENES IN PANCREATIC DUCTAL ADENOCARCINOMA

Bioinformatics

Payaam Tamboli
Aaditya Faria
Krishna Atreya
Sonalika Ray

ABSTRACT

Pancreatic ductal adenocarcinoma is usually found late, has no reliable screening test, and carries a five-year survival rate that has stayed near 13% for years, so better markers of the disease are still needed. To look for candidate genes, four independent microarray datasets from the Gene Expression Omnibus (GSE15471, GSE71729, GSE183795, GSE28735) were compared between tumor and normal pancreatic tissue, and genes that came up as differentially expressed in at least two of the four datasets were kept as a consensus set. This gave 48 genes, 44 up-regulated and 4 down-regulated in tumor tissue. These genes were tied mainly to extracellular matrix organization, collagen fibril organization, and Wnt signaling. A protein interaction network built from these 48 genes in STRING had 47 nodes and 143 edges, far more connections than expected by chance (PPI enrichment $p < 1.0 \times 10{-}16$), and cytoHubba was used to pull out the top 20 hub genes. When these 20 genes were tested as a combined signature against overall survival in the TCGA-PAAD cohort, the result was not significant (log-rank $p = 0.26$). Tested one at a time, three genes did reach significance: SFRP4 was linked to better survival (HR = 0.743, $p = 0.0115$), while SPP1 (HR = 1.256, $p = 0.0275$) and SFRP2 (HR = 1.504, $p = 0.0267$) were both linked to worse survival. These three genes, identified through both network centrality and survival data, are reasonable candidates for further study as prognostic markers in pancreatic cancer.

KEYWORDS

1INTRODUCTION

Pancreatic cancer remains one of the deadliest cancers in the world [1]. In 2026, an estimated 67,530 people in the United States will be diagnosed with pancreatic cancer, and about 52,740 will die from it [1]. The disease makes up only around 3% of all new cancer cases, yet it is now the third leading cause of cancer death in both men and women, behind only lung and colorectal cancer [1]. The five-year relative survival rate has stayed at 13% for three years in a row, even as survival for most other cancers keeps improving [2]. Only 17% of patients are diagnosed while the tumor is still localized, and even at that early stage, five-year survival is just 44% [1]. Most patients are diagnosed after the cancer has already spread, when treatment options are limited and survival drops sharply [3].

The reason pancreatic cancer is caught so late comes down to how it behaves biologically. Symptoms rarely appear until the tumor has grown large or spread [1]. There is no low-cost, reliable screening test

for the general population the way there is for breast or colon cancer [4]. By the time imaging or blood markers pick it up, the disease has often already reached nearby organs or the bloodstream. Pancreatic tumors are also surrounded by a thick layer of connective tissue called the desmoplastic stroma, which can make up most of the tumor mass and acts as a physical barrier to chemotherapy [5]. Nearly all pancreatic ductal adenocarcinoma (PDAC) cases carry a mutation in the KRAS gene, present in over 90% of tumors [6], which drives constant growth signaling and has proven very hard to target with drugs. On top of this, tumors from different patients look quite different from each other at the molecular level [5], which is part of why a single gene or a single small study rarely holds up as a reliable biomarker once tested in a different group of patients.

This is where a lot of past gene-expression studies on pancreatic cancer run into trouble. Many of them rely on just one microarray or RNA-seq dataset from a single research group, using a single patient cohort and a single microarray platform. A gene that looks significant in one dataset can easily turn out to be a platform-specific quirk or afalse positive that shows up only in that particular group of patients [7]. Findings like this often fail to replicate when someone tries the same analysis on a different cohort. Fewer studies go a step further and combine differential expression results across multiple independent datasets, build a protein interaction network around the genes that hold up, and then check whether those genes actually predict patient survival in real clinical data. Doing all three together, rather than any one of them alone, is what gives more confidence that a candidate gene is not just a statistical artifact of one dataset.

The idea behind combining these methods is fairly simple. If a gene shows up as differentially expressed in several independent patient cohorts profiled on different microarray platforms, that is much stronger evidence than one dataset alone [8]. If that same gene also turns out to be a highly connected hub within the protein interaction network built from the tumor-associated gene set, it is more likely to have a real, central role in tumor biology rather than being a downstream bystander [9]. And if patients with high or low expression of that gene show a measurably different survival outcome in an independent clinical cohort such as The Cancer Genome Atlas, the gene has evidence connecting it to actual patient outcomes and not just to the tumor tissue itself [10]. Each of these three layers on its own has known weaknesses; used together, they tend to cancel out each other's blind spots.

This study set out to apply exactly this kind of layered approach to pancreatic cancer. Four independent, publicly available microarray datasets from the Gene Expression Omnibus were used to identify genes that were consistently differentially expressed between tumor and normal pancreatic tissue across at least two of the four cohorts. These consensus genes were then used to build a protein-protein in-teraction network, from which the most highly connected hub genes were identified. Finally, the prognostic value of these hub genes was tested directly against survival data from the TCGA pancreatic adeno-carcinoma cohort, both as a combined signature and gene by gene, and the genes that came out as significant were checked against previously published pancreatic cancer literature. The aim is to put forward a small set of hub genes for pancreatic cancer that is backed by more than one type of evidence, and that could be worth further validation as prognostic markers rather than simply another list of differentially expressed genes from a single dataset.

2 METHODOLOGY

This study combined four steps to find genes linked to pancreatic cancer that could also be tied to patient survival: gene expression analysis across several public datasets, pathway analysis, a protein interaction network, and a survival check using clinical data. Each step is described below in the order it was carried out.

2.1Dataset Selection and Acquisition

Four microarray datasets were obtained from the Gene Expression Omnibus (GEO), a public repository maintained by the National Center for Biotechnology Information [11, 12]. Each dataset had to include both pancreatic tumor and normal tissue samples, at least 10 samples per group, and a properly annotated microarray platform. Based on these requirements, four datasets were chosen: GSE15471 [13], GSE71729 [5], GSE183795 [14], and GSE28735 [15].

The Series Matrix File for each dataset was downloaded from GEO, and sample information such as tissue type and platform was pulled out

and organized into a pivot table. This made it possible to sort the samples of each dataset into two clear groups, tumor and normal, before running any expression analysis.

2.2 Differential Expression Analysis Using GEO2R

Each of the four datasets was analyzed separately using GEO2R, the online tool built into GEO that runs on the limma package from R/Bioconductor [11, 16]. Samples in every dataset were manually split into a tumor group and a healthy group based on the information given in GEO, and the comparison was set up as tumor versus healthy.

The same cut-offs were used for all four datasets: a fold change of at least 2 on the log2 scale in either direction, a raw p-value for significance, and a Benjamini-Hochberg adjustment to control the false

discovery rate [17]. This correction was used because microarray data involves testing thousands of genes at once, and without it, a large share of the "significant" genes would just be false positives [17].

2.3 Extracting the Gene Lists

For each dataset, GEO2R produced a full table of genes along with their base mean expression, log2 fold change, raw p-value, adjusted p-value, and direction of change (whether the gene went up or down in tumor tissue compared to normal tissue). This table was downloaded for all four datasets and saved as plain text files, giving four separate result sets for GSE15471, GSE71729, GSE183795, and GSE28735.

2.4 Finding Genes Common to Multiple Datasets

Since each dataset comes from a different group of patients and sometimes a different microarray platform, a gene that looks significant in just one dataset could easily be a fluke rather than a real biological signal [7, 8]. To deal with this, the four gene lists were brought to-gether in R [18] using the tidyverse [19], VennDiagram [20], and UpSetR [21] packages, and only genes that showed up consistently across datasets were kept.

The four result files were read into R and their gene symbol, log2FC, and p-value columns were renamed to a common format so they could be handled the same way regardless of which dataset they came from. Within each dataset, genes that appeared more than once because of duplicate probes were reduced to a single entry. Each dataset was then tagged with its own GEO accession number so that, once merged, it would still be clear which dataset each entry originally came from. The four tagged tables were combined into one large table containing every gene from every dataset.

From this combined table, the genes were grouped by their symbol, and for each gene the number of datasets it appeared in was counted, along with its average log2 fold change and average p-value across those datasets. A gene's overall direction, up or down, was decided by whether this average log2 fold change came out positive or negative. Genes that showed up in at least two of the four datasets were treated as consensus genes, since agreement across independent cohorts gives much more confidence than a result from a single dataset [7, 8]. This consensus list was then ordered by how many datasets supported each gene and by the size of its average fold change.

To check that this overlap was reasonable, the raw gene lists from all four datasets were also plotted as a Venn diagram [20] and an UpSet plot [21], both of which show how many genes are shared between any combination of the datasets. The consensus genes were further split into those found in exactly two, exactly three, or all four datasets, and separately into up-regulated and down-regulated groups, so the strength of evidence behind each gene would be easy to see later. All of these tables, the combined list, the consensus list, and the various subsets, were saved as CSV files for the next stage of filtering.

2.5 Filtering the Consensus Genes

The consensus gene list was narrowed down further using the average log2 fold change and average p-value calculated in the previous step. A gene was kept only if its average log2 fold change was above 2 or below

−2, meaning at least a four-fold change in expression, and its average p-value was below 0.05. The resulting filtered gene list was carried forward to the pathway enrichment and network analyses described below.

2.6 Pathway and Functional Enrichment Analysis

The filtered consensus gene list, together with each gene's log2 fold change and p-values, was uploaded to iLINCS [22] as a gene signature. From there, Enrichr [23, 24] was used to check which biological pathways and processes these genes were linked to. Three sources were checked: KEGG pathways [25], Gene Ontology terms covering biological process, cellular component, and molecular function [26, 27], and Reactome pathways [28], all using a cut-off of $p < 0.05$. The enrichment results from each source were downloaded, and iLINCS was also used to look at how the gene signature connected to other known pathway and perturbation signatures in its database.

2.7 Building the Protein Interaction Network

To see how the filtered consensus genes relate to each other at the protein level, a protein-protein interaction network was built using STRING [29]. The gene list was entered into STRING under the multiple-protein search option, restricted to human proteins, using a medium-to-high confidence score for the interactions. The resulting network was then exported and opened in Cytoscape [30] through the stringApp plugin [31] for closer analysis.

2.8 Identifying the Hub Genes

Inside Cytoscape, the cytoHubba plugin [9] was used to score every protein in the network based on how central it was within the network structure. The 20 highest-scoring genes were taken forward as the hub genes, since proteins with the most connections in a network are generally considered more likely to play a real, central role in the disease rather than being minor players. The network was relabeled with gene names for clarity, and the final network image along with the ranked list of 20 hub genes was exported for the next step.

2.9 Checking Survival Outcomes for the Hub Genes

To find out whether these 20 hub genes actually mattered for patient outcomes, and not just for the tumor tissue itself, survival analysis was carried out using GEPIA2 [10]. GEPIA2 pulls together RNA-sequencing and clinical data from the TCGA pancreatic adenocarcinoma cohort along with normal tissue data from GTEx [32,33]. Both overall survival and progression-free interval were looked at.

First, all 20 hub genes were combined into a single signature and submitted to GEPIA2 together. Patients in the TCGA-PAAD cohort were split into a high-expression group and a low-expression group based on the median value of this combined signature. Kaplan-Meier curves were plotted for both groups, and the log-rank test [34] was used to check whether the two curves were actually different. A Cox proportional-hazards model [35] was also run to get the hazard ratio between the high and low groups, with significance set at $p < 0.05$ for both tests.

Next, each of the 20 genes was tested on its own rather than as a combined score, again using a Cox model [35] within GEPIA2, separately for overall survival and progression-free interval. This gave a hazard ratio and a 95% confidence interval for every gene. A gene was treated as individually significant only if its confidence interval did not cross a hazard ratio of 1. Genes with a hazard ratio above 1 were read as risk factors, meaning higher expression pointed to worse survival, while genes with a hazard ratio below 1 were read as protective, meaning higher expression pointed to better survival.

Finally, whichever genes came out significant in this individual survival check were looked up in previously published pancreatic cancer research, to see whether other studies had reported similar findings for the same genes. This was done simply as a sanity check, to see whether the genes picked out by this network and survival pipeline lined up with what has already been reported elsewhere.

3 RESULTS

3.1 Dataset Overview

The four GEO datasets used in this study differed in size and platform. GSE15471 contained 36 tumor and 36 matched normal pancreatic tissue samples, with one tumor-normal pair excluded after quality control. GSE71729 contributed 145 primary PDAC tumor samples and 88 distant-site adjacent normal samples. GSE183795 contributed 139 tumor samples and 102 adjacent non-tumor samples. GSE28735 contributed 45 matched tumor-normal pairs (90 samples total). Differential expression analysis in GEO2R identified 361 differentially expressed probes in GSE15471 (318 up-regulated, 43 down-regulated), 141 in GSE71729 (123 up-regulated, 18 down-regulated), 32 in GSE183795 (17 up-regulated, 15 down-regulated), and 57 in GSE28735 (33 up-regulated, 24 down-regulated).

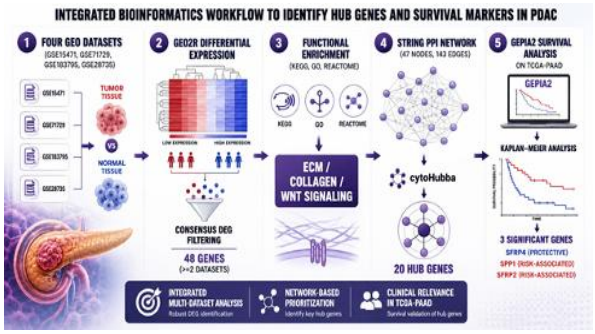

Figure 1: Integrated bioinformatics workflow for identifying consensus DEGs, key pathways, hub genes, and survival-associated markers in pancreatic ductal adenocarcinoma (PDAC)

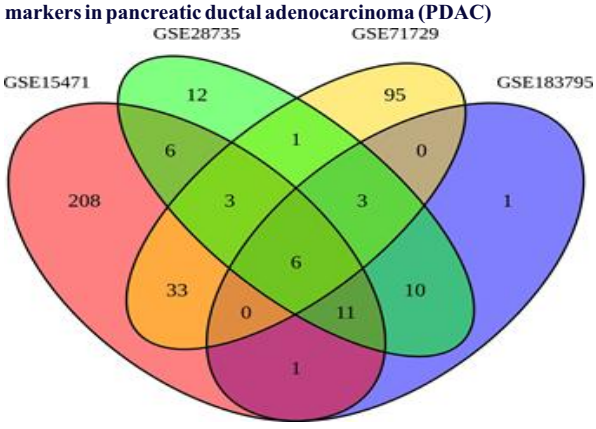

Figure 2: Four-way overlap of differentially expressed genes identified independently in GSE15471, GSE28735, GSE71729, and GSE183795.

3.2 Identification of Consensus DEGs

The overlap among the four per-dataset DEG lists is shown in Figure 2. Most genes identified in each dataset were not shared with any other dataset, reflecting differences in platform, cohort size, and tissue composition across the four series. After merging the four DEG lists and retaining genes identified in two or more datasets, 48 consensus DEGs remained. Of these, 44 were up-regulated and 4 were down-regulated in tumor relative to normal pancreatic tissue. Forty-one of the 48 consensus genes were shared between GSE15471 and GSE71729 alone, six genes (CTRC, CEL, FAM24B-CUZD1///CUZD1, CPA2, CLPS,

CELA2A, PLA2G1B, KRT23, PNLIP///COL1A2) were shared specifically between GSE183795 and GSE28735, and four genes (COL11A1, PNLIPRP2, AGR2, PNLIPRP1) were supported by three or more datasets. The complete consensus DEG list, including mean $\log_2$FC, mean raw *p*-value, regulation direction, and the supporting datasets for each gene, is given in Table 1. The distribution of fold-change and significance across the panel is shown as a volcano plot in Figure 3.

3.3 Functional Enrichment Results

KEGG pathway enrichment of the 48-gene consensus panel returned 11 pathways at adjusted $p < 0.05$. The most significantly enriched pathway was Protein Digestion and Absorption (5/102 genes, adjusted $p = 1.23 \times 10^{-8}$), followed by Cytoskeleton in Muscle Cells (5/230, adjusted $p = 3.70 \times 10^{-7}$), ECM-Receptor Interaction (3/88, adjusted $p = 4.54 \times 10^{-5}$), the AGE-RAGE Signaling Pathway in Diabetic Complications (3/100, adjusted $p = 4.54 \times 10^{-5}$), and Platelet Activation (3/125, adjusted $p = 6.48 \times 10^{-5}$) [25]. Additional enriched pathways included the Relaxin Signaling Pathway, Integrin Signaling, Focal Ad-hesion, Diabetic Cardiomyopathy, the PI3K-AKT Signaling Pathway, and Proteoglycans in Cancer, all driven predominantly by the collagen genes (COL1A1, COL1A2, COL3A1, COL10A1, COL11A1) together with COMP.

Gene Ontology enrichment [26, 27] was run separately for the Biological Process, Cellular Component, and Molecular Function categories through Enrichr [23, 24]. In Biological Process, the top terms were Skeletal System Development (7/169 genes, adjusted $p = 4.43 \times 10^{-11}$), Collagen Fibril Organization (4/54, adjusted $p = 6.84 \times 10^{-7}$), Extracellular Matrix Organization (5/192, adjusted $p = 8.69 \times 10^{-7}$), Skeletal System Morphogenesis (3/51, adjusted $p = 6.43 \times 10^{-5}$), Skin

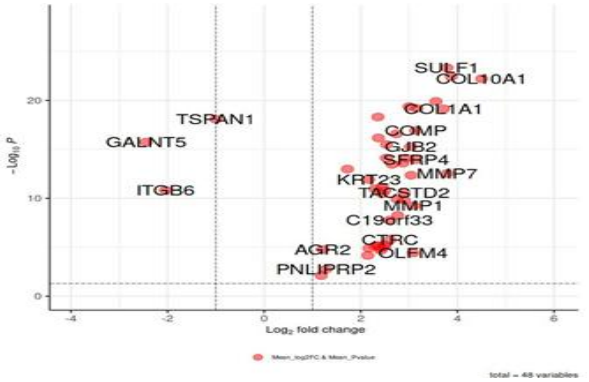

Figure 3: Volcano plot of the 48 consensus DEGs, showing mean $\log_2$ fold change against $-\log_{10}$ mean raw *p*-value across the contributing datasets.

Development (3/72, adjusted $p = 1.47 \times 10^{-4}$), and Positive Regula-tion of the Wnt Signaling Pathway (3/121, adjusted $p = 4.99 \times 10^{-4}$). The full distribution of enriched Biological Process terms is shown as a treemap in Figure 4, grouped by relative term size and significance. In Cellular Component, the panel was enriched for the Endoplasmic Reticulum Lumen (5/306, adjusted $p = 1.73 \times 10^{-6}$) and the Intracellular Organelle Lumen (5/955, adjusted $p = 2.28 \times 10^{-4}$). In Molecular Function, the strongest terms were Platelet-Derived Growth Factor Binding (3/11, adjusted $p = 1.04 \times 10^{-7}$) and Protease Binding (4/121, adjusted $p = 9.11 \times 10^{-7}$), along with Heparan Sulfate Proteoglycan Binding and Arylsulfatase Activity, both contributed by SULF1 and COMP.

Reactome pathway enrichment [28] returned seven pathways at FDR < 0.05, all supported by single genes. COL1A1 mapped to RUNX2-related pathways, including RUNX2 Regulates Osteoblast Differentiation (FDR = 0.042) and Transcriptional Regulation by RUNX2 (FDR = 0.042). COL1A2 mapped to TGF-beta and SMAD-related pathways, including Signaling by TGF-beta Receptor Complex, Signaling by TGFB Family Members, and the SMAD2/SMAD3:SMAD4 transcriptional complex (all FDR = 0.042).

3.4 PPI Network and Hub Gene Results

Of the 48 consensus DEGs submitted to STRING, 47 mapped to nodes in the resulting protein-protein interaction network [29]. The network contained 143 edges, an average node degree of 6.09, and an average local clustering coefficient of 0.662, against an expected 7 edges for a random gene set of the same size (PPI enrichment $p < 1.0 \times 10^{-16}$), indicating substantially more interaction than expected by chance. This network was imported into Cytoscape [30] and ranked using cytoHubba[9]

to identify the top 20 hub genes (Table 2, Figure 5). Collagen and extracellular-matrix-associated genes (COL1A1, COL1A2, COL3A1, COL10A1, COL11A1, COMP, FNDC1, MXRA5) made up the largest functional subgroup among the top 20 hub genes, alongside two Wnt pathway modulators (SFRP2, SFRP4), the proteoglycan-modifying enzyme SULF1, and a cluster of digestive-enzyme genes (CEL, CLPS, CPA2, PLA2G1B, PNLIP) contributed mainly by GSE183795 and GSE28735.

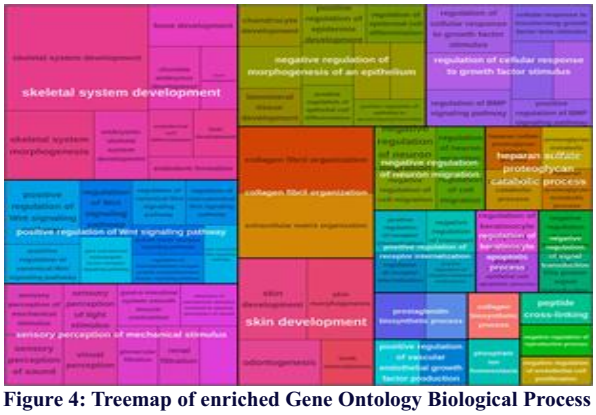

Figure 4: Treemap of enriched Gene Ontology Biological Process terms for the 48-gene consensus DEG panel. Tile size reflects relative term size within the enrichment result.

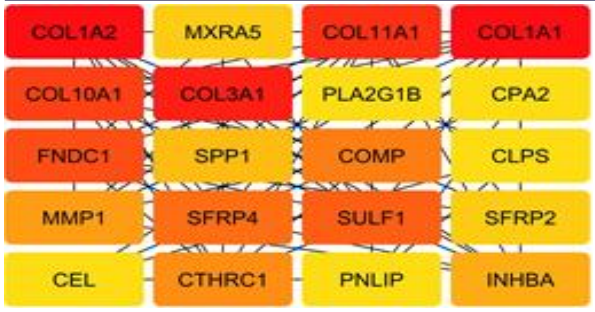

Figure 5: Top 20 hub genes identified by cytoHubba, colored by relative log_2FC. Edges represent STRING protein-protein interactions.

3.5 Survival Analysis Results

The 20 hub genes were first evaluated as a combined signature in GEPIA2 [10], stratifying the TCGA-PAAD cohort (n = 178; 89 high, 89 low) by median expression of the signature score. The combined 20-gene signature was not significantly associated with overall survival (log-rank p = 0.26; HR = 1.3, p(HR) = 0.25; Figure 6) [34, 35].

Each of the 20 hub genes was then evaluated individually using a univariate Cox proportional-hazards model for overall survival (OS) (Table 3, Figure 7). Three genes reached significance. SFRP4 was associated with a favorable prognosis (HR = 0.743, 95% CI 0.590-0.935, p = 0.0115). SPP1 (HR = 1.256, 95% CI 1.026-1.537, p =0.0275) and SFRP2 (HR = 1.504, 95% CI 1.048-2.159, p = 0.0267)

were both associated with an unfavorable prognosis. The remaining 17 hub genes did not reach significance.

None of the 20 individually significant or non-significant hub gene results were adjusted for multiple comparisons; they are reported here as raw univariate Cox model outputs from GEPIA2.

Across the four GEO datasets, 48 genes were consistently differen-tially expressed in PDAC tumor tissue relative to normal pancreatic tissue, 44 of them up-regulated. These genes were enriched for ex-tracellular matrix organization, collagen fibril organization, skeletal system development, and Wnt and TGF-beta signaling pathways. A PPI network built from the 48 genes showed significantly more interac-tion than expected by chance, and cytoHubba identified 20 hub genes dominated by collagen and ECM-associated proteins. The combined 20-gene signature was not significantly associated with overall survival in TCGA-PAAD. Individually, SFRP4 was associated with favorable

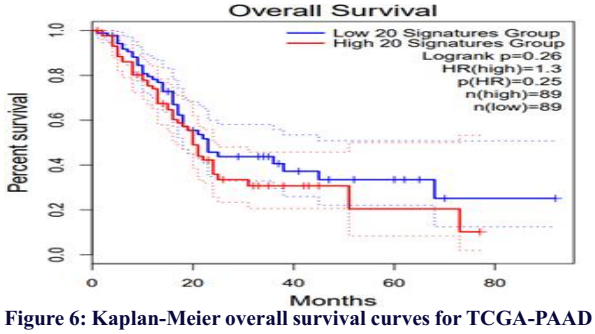

Figure 6: Kaplan-Meier overall survival curves for TCGA-PAAD patients stratified by high versus low expression of the combined 20-gene hub signature.

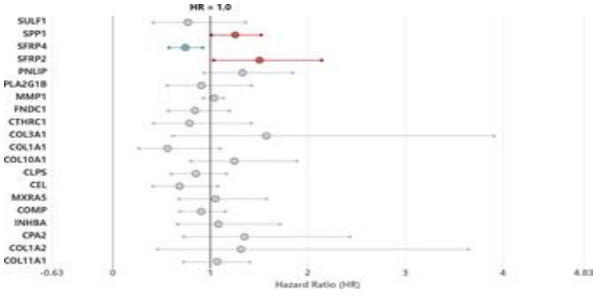

Figure 7: Forest plot of hazard ratios for overall survival across the 20 hub genes, TCGA-PAAD cohort.

overall survival, while SPP1 and SFRP2 were both associated with unfavorable overall survival.

4 DISCUSSION

The genes that came out on top of this analysis, both in the con-sensus DEG list and in the hub gene ranking, point mostly toward one thing: the extracellular matrix that surrounds pancreatic tumors. COL1A1, COL1A2, COL3A1, COL10A1, COL11A1, COMP, FNDC1, and MXRA5 were all up-regulated and all showed up among the top hub genes, and the enrichment results (collagen fibril organization, extracellular matrix organization, ECM-receptor interaction) back this up from a different angle. This lines up with what is already known about PDAC, where tumors are wrapped in a thick layer of dense connective tissue, the desmoplastic stroma, that can make up most of the tumor mass and gets in the way of chemotherapy reaching the cancer cells [5]. The genes identified here look less like drivers of the can-cer cells themselves and more like a transcriptional signature of that stromal wall. This matters for how the result should be read: a gene showing up as a hub gene in this kind of network does not mean it is doing something inside the tumor cell, it may just mean the surrounding tissue has changed shape.

SPP1, which codes for the protein osteopontin, is one of the few hub genes here with a real track record in pancreatic cancer research specifically. A recent single-cell RNA-seq study of PDAC combined with TCGA data found that SPP1 was tied to worse survival and was expressed mainly by malignant ductal cells and by macrophages inside the tumor [36]. The result here, SPP1 coming out as both a network hub and a risk gene for overall survival (HR = 1.256, p = 0.0275), agrees with that. It is one of the more consistent findings across studies of this disease.

SFRP4 and SFRP2 are harder to make sense of together. Both are Wnt pathway antagonists from the same protein family, both showed up as consensus DEGs and hub genes here, and both were significant for survival, but in opposite directions. SFRP4 was protective (HR= 0.743), while SFRP2 was a risk factor (HR = 1.504). Looking at prior work does not settle this cleanly. A study specifically in PDAC found that SFRP4 is lost in tumor tissue compared to normal pancreas, and that patients with lower SFRP4 had worse outcomes, which points the same direction as the protective result found here [37]. But a large pan-cancer analysis covering close to 8,000 tumor samples across 29 cancer types found that SFRP2 and SFRP4 usually go together, and that both are more often linked to worse prognosis and to genes involved in epithelial-to-mesenchymal transition, with both proteins mostly coming from the tumor stroma rather than the cancer cells themselves [38]. That pan-cancer picture does not match the SFRP4 direction found in the present PDAC data. This is left as an open question rather than something this study can resolve. It is possible that SFRP4 behaves differently in pancreatic cancer than it does across cancer types in general, or that the stromal versus tumor-cell source of the two genes plays a bigger role than a bulk-tissue gene expression analysis can capture.

One additional enrichment result is worth noting cautiously. SFRP4 alone maps to the Gene Ontology term positive regulation of receptor-mediated endocytosis (GO:0048260, adjusted p = 0.043). To our knowledge, this specific association has not been examined in the pancreatic cancer literature. However, this term is supported by a single gene in the panel, and GO annotations can reflect inferred or cross-species evidence rather than PDAC-specific experimental data, so this is presented here as a narrow, hypothesis-generating obser-vation rather than a validated pathway link.

The combined 20-gene signature not reaching significance, despite three individual genes doing so, is not really a contradiction. A com-bined signature score works by averaging expression across all 20 genes at once. If SFRP4 pulls survival in one direction and SPP1 and SFRP2 pull it in the other, and the other 17 genes contribute close to nothing on their own, the combined score ends up diluted rather than sharpened. A handful of individually meaningful genes can therefore fail to add up to a meaningful group score once mixed in with a larger set of weaker or opposing signals.

This study has several limitations that are worth stating plainly. The expression data used throughout come from bulk tissue at a single time point, so there is no way to tell from this data alone how much of the signal comes from tumor cells versus the surrounding stroma or immune cells. DEGs were selected using a raw p-value cutoff along with a fold-change threshold rather than an adjusted p-value, which is a

looser filter and could let some false-positive genes into the consensus list. Two of the four datasets, GSE15471 and GSE71729, accounted for the large majority of the consensus genes, so the panel leans more heavily on those two cohorts than on all four equally. Being a hub gene in a STRING network also does not mean a gene causes any-thing; STRING scores are built partly from existing literature and co-expression data, so genes that are already well studied tend to end up more connected regardless of their true biological role. For example, several of SFRP4's STRING co-expression edges, with collagen genes such as COL1A1, COL1A2, COL3A1, and COMP, are drawn from breast cancer expression compendia (e.g., Perou-Botstein-2000, Ross-Perou-2001, Bild-Nevins-2006) rather than from pancreatic-specific data. This illustrates that a high STRING confidence score can reflectbroad, cross-tissue stromal co-expression patterns rather than an inter-action specific to pancreatic cancer. Finally, the survival analysis used TCGA-PAAD alone, so none of the individual gene results have been checked against a second, independent cohort.

Given these limitations, the most useful next step would be testing whether SFRP4, SPP1, and SFRP2 hold up as prognostic markers in an independent PDAC cohort outside of TCGA. Single-cell or spatial transcriptomic data would also help separate how much of the collagen and ECM signal is coming from tumor cells compared to stromal fibroblasts, which bulk expression data cannot resolve on its own.

Table 1: Consensus differentially expressed genes (present in ≥ 2 of the four datasets), ranked by mean log_2FC.

Gene	Datasets present	Mean log_2FC	Mean p-value	Direction	Supporting datasets
COL10A1	2	4.504	6.05E-23	Up	GSE15471, GSE71729
INHBA	2	3.873	3.53E-23	Up	GSE15471, GSE71729
MMP7	2	3.787	2.93E-13	Up	GSE15471, GSE71729
SULF1	2	3.784	4.55E-24	Up	GSE15471, GSE71729
COL1A1	2	3.717	7.30E-20	Up	GSE15471, GSE71729
CTHRC1	2	3.556	1.25E-20	Up	GSE15471, GSE71729
COL1A2	2	3.160	6.45E-20	Up	GSE15471, GSE71729
SFRP4	2	3.143	1.13E-14	Up	GSE15471, GSE71729
COMP	2	3.139	1.20E-17	Up	GSE15471, GSE71729
MMP1	2	3.099	5.55E-10	Up	GSE15471, GSE71729
OLFM4	2	3.083	3.83E-05	Up	GSE15471, GSE71729
S100P	2	3.044	4.37E-13	Up	GSE15471, GSE71729
GJB2	2	3.032	5.30E-16	Up	GSE15471, GSE71729
FAP	2	3.001	4.21E-20	Up	GSE15471, GSE71729
SFRP2	2	2.929	7.80E-15	Up	GSE15471, GSE71729
TMC5	2	2.901	1.54E-10	Up	GSE15471, GSE71729
TACSTD	2	2.893	2.71E-11	Up	GSE15471, GSE71729
MMP11	2	2.872	2.54E-14	Up	GSE15471, GSE71729
TCN1	2	2.766	5.45E-09	Up	GSE15471, GSE71729
LCN2	2	2.753	1.24E-10	Up	GSE15471, GSE71729
COL3A1	2	2.733	2.63E-17	Up	GSE15471, GSE71729
PLAT	2	2.651	3.38E-14	Up	GSE15471, GSE71729
CTRC	2	2.605	1.74E-06	Up	GSE183795, GSE28735
C19orf33	2	2.596	1.67E-08	Up	GSE15471, GSE71729
FNDC1	2	2.536	3.29E-16	Up	GSE15471, GSE71729
PHLDA2	2	2.530	7.60E-15	Up	GSE15471, GSE71729
SPP1	2	2.498	1.60E-11	Up	GSE15471, GSE71729
CEL	2	2.494	5.70E-06	Up	GSE183795, GSE28735
FAM24B-CUZD1///CUZD1	2	2.493	9.20E-06	Up	GSE183795, GSE28735
KLK10	2	2.442	6.40E-12	Up	GSE15471, GSE71729
C15orf48	2	2.440	3.48E-11	Up	GSE15471, GSE71729
TFF1	2	2.417	2.07E-05	Up	GSE15471, GSE71729
CPA2	2	2.410	1.19E-05	Up	GSE183795, GSE28735
CLPS	2	2.369	6.45E-06	Up	GSE183795, GSE28735
MXRA5	2	2.365	6.75E-17	Up	GSE15471, GSE71729
PMEPA1	2	2.356	4.96E-19	Up	GSE15471, GSE71729
CELA2A	2	2.295	7.15E-06	Up	GSE183795, GSE28735
KRT6B	2	2.294	1.11E-11	Up	GSE15471, GSE71729
PLA2G1B	2	2.169	1.26E-05	Up	GSE183795, GSE28735
KRT23	2	2.166	1.21E-12	Up	GSE15471, GSE71729
PNLIP///COL1A2	2	2.146	6.70E-05	Up	GSE183795, GSE28735
COL11A1	3	1.723	1.02E-13	Up	GSE15471, GSE28735, GSE71729
PNLIPRP2	4	1.279	1.77E-03	Up	GSE15471, GSE183795, GSE28735, GSE71729
AGR2	3	1.210	1.71E-05	Up	GSE15471, GSE28735, GSE71729
PNLIPRP1	3	1.182	7.73E-03	Up	GSE15471, GSE183795, GSE28735
TSPAN1	3	-1.000	8.30E-19	Down	GSE183795, GSE28735, GSE71729
ITGB6///ITGB6	2	-2.042	1.49E-11	Down	GSE183795, GSE28735
Galnt5///galnt5	2	-2.445	1.81E-16	Down	GSE183795, GSE28735

Table 2: Top 20 hub genes ranked by cytoHubba, with mean log_2FC from the consensus DEG analysis.

Gene	Mean log_2FC
COL10A1	4.504
INHBA	3.873
SULF1	3.784
COL1A1	3.717
CTHRC1	3.556
COL1A2	3.160
SFRP4	3.143
COMP	3.139
MMP1	3.099
SFRP2	2.929
COL3A1	2.733
FNDC1	2.536
SPP1	2.498
CEL	2.494
CPA2	2.410
CLPS	2.369
MXRA5	2.365
PLA2G1B	2.169
PNLIP	2.146
COL11A1	1.723

Table 3: Univariate Cox proportional-hazards results for overall survival (OS) across the 20 hub genes, TCGA-PAAD cohort. Bold rows reached $p < 0.05$.

Gene	HR (95% CI)	p
CEL	0.683 (0.428-1.091)	0.111
CLPS	0.851 (0.614-1.179)	0.332
COL10A1	1.246 (0.817-1.901)	0.306
COL11A1	1.027 (0.743-1.419)	0.872
COL1A1	0.560 (0.282-1.114)	0.0985
COL1A2	1.313 (0.471-3.662)	0.602
COL3A1	1.574 (0.632-3.924)	0.330
COMP	0.906 (0.702-1.168)	0.445
CPA2	1.350 (0.745-2.447)	0.323
CTHRC1	0.786 (0.431-1.435)	0.433
FNDC1	0.842 (0.585-1.211)	0.353
INHBA	1.081 (0.676-1.729)	0.744
MMP1	1.039 (0.942-1.145)	0.445
MXRA5	1.051 (0.693-1.593)	0.815
PLA2G1B	0.908 (0.573-1.437)	0.680
PNLIP	1.331 (0.952-1.860)	0.0947
SFRP2	1.504 (1.048-2.159)	0.0267
SFRP4	0.743 (0.590-0.935)	0.0115
SPP1	1.256 (1.026-1.537)	0.0275
SULF1	0.769 (0.430-1.374)	0.375

5 CONCLUSION

Pancreatic ductal adenocarcinoma is usually found late, has no reliable screening test, and carries a five-year survival rate that has stayed near 13% for years, so better markers of the disease are still needed. To look for candidate genes, four independent microarray datasets from the Gene Expression Omnibus (GSE15471, GSE71729, GSE183795, GSE28735) were compared between tumor and normal pancreatic tissue, and genes that came up as differentially expressed in at least two of the four datasets were kept as a consensus set. This gave 48 genes, 44 up-regulated and 4 down-regulated in tumor tissue. These genes were tied mainly to extracellular matrix organization, collagen fibril organization, and Wnt signaling. A protein interaction network built from these 48 genes in STRING had 47 nodes and 143 edges, far more connections than expected by chance (PPI enrichment p < 1e-16), and

cytoHubba was used to pull out the top 20 hub genes. When these 20 genes were tested as a combined signature against overall survival in the TCGA-PAAD cohort, the result was not significant (log-rank p = 0.26). Tested one at a time, three genes did reach significance: SFRP4 was linked to better survival (HR = 0.743, p = 0.0115), while SPP1 (HR = 1.256, p = 0.0275) and SFRP2 (HR = 1.504, p = 0.0267) were both linked to worse survival. These three genes, identified through both network centrality and survival data, are reasonable candidates for further study as prognostic markers in pancreatic cancer.

Supplemental Materials

Two supplementary spreadsheet files accompany this paper.

PC_DEGs_All.xlsx contains the complete, unfiltered GEO2R out-put for each of the four datasets, with one sheet per dataset (GSE15471, GSE71729, GSE183795, GSE28735). Each sheet lists every probe or gene tested in that dataset along with its adjusted *p*-value, raw *p*-value, *t*-statistic, *B*-statistic, $\log_2$ fold change, gene symbol, and gene title, before any cross-dataset filtering was applied.

Supplementary_File.xlsx contains the processed results refer-enced throughout the Methodology and Results sections, split across the following sheets:

- **DEGs:** the full consensus DEG list (genes present in two or more of the four datasets), matching Table 1 and including all 48 genes with their dataset counts, mean $\log_2$FC, mean raw *p*-value, direction, and supporting datasets.
- **Hub Genes:** the top 20 hub genes ranked by cytoHubba, matching Table 2.
- **Pathways:** the full KEGG pathway enrichment results for the 48-gene consensus panel.
- **Gene Interactions:** the STRING protein-protein interaction edge list used to build the network described in Section 3.4.
- **Reactome Results:** the full Reactome pathway enrichment out-put.
- **Biological Processes, Cellular Components, and Molecular Functions:** the complete Gene Ontology enrichment results for each of the three GO categories, beyond the top terms discussed in Section 3.3.

REFERENCES

[1] American Cancer Society. Cancer facts & figures 2026. Technical report, American Cancer Society, Atlanta, GA, 2026. Available at:https://www.cancer.org/research /cancer-facts-statistics/ all-cancer-facts-figures.html. 1

[2] OncoDaily. The 5-year survival rate for pancreatic cancer remains at 13%, 2026. Reporting on American Cancer Society Cancer Statistics, 2026. Available at: https://oncodaily.com/voices/ survival-rate-448909. 1

[3] Seena Magowitz Foundation. Pancreatic cancer incidence and survival rates by stage of pancreatic cancer, 2026. Avail-able at: https://seenamagowitzfoundation.org/resource/pancreatic-cancer-statistics/. 1

[4] Pancreatic Cancer Action Network. Pancreatic cancer survival rate, 2026. Available at: https://pancan.org/facing-pancreatic-cancer/ about-pancreatic-cancer/survival-rate/. 1

[5] Richard A Moffitt, Raoud Marayati, Elizabeth L Flate, Keith E Volmar, S Gabriela Herrera Loeza, Katherine A Hoadley, Naim U Rashid, Lind-say A Williams, Samuel C Eaton, Alexander H Chung, et al. Virtual microdissection identifies distinct tumor-and stroma-specific subtypes of pancreatic ductal adenocarcinoma. Nature genetics, 47(10): 1168–1178, 2015. 1, 2, 5

[6] Dongwoo Cho, Kabsoo Shin, Tae Ho Hong, Sung Hak Lee, Younghoon Kim, In-Ho Kim, Sook-Hee Hong, MyungAh Lee, and Se Jun Park. Clini-cal and molecular characteristics of kras codon-specific mutations in ad-vanced pancreatic ductal adenocarcinoma with prognostic and therapeutic implications. International Journal of Molecular Sciences, 26(22):10908, 2025. 1

[7] Adaikalavan Ramasamy, Adrian Mondry, Chris C. Holmes, and Douglas G. Altman. Key issues in conducting a meta-analysis of gene expression microarray datasets. PLoS Medicine, 5(9):e184, 2008. 2

[8] Daniel R. Rhodes, Terrence R. Barrette, Mark A. Rubin, Debashis Ghosh, and Arul M. Chinnaiyan. Meta-analysis of microarrays: interstudy valida-tion of gene expression profiles reveals pathway dysregulation in prostate cancer. Cancer Research, 62(15):4427–4433, 2002. 2

[9] Chia-Hao Chin, Shu-Hwa Chen, Hsin-Hung Wu, Chin-Wen Ho, Ming-Tat Ko, and Chung-Yen Lin. cytohubba: identifying hub objects and sub-networks from complex interactome. BMC Systems Biology, 8(Suppl 4):S11, 2014. 2, 3, 4

[10] Zefang Tang, Boxi Kang, Chenwei Li, Tianxiang Chen, and Zemin Zhang. Gepia2: an enhanced web server for large-scale expression profiling and interactive analysis. Nucleic Acids Research, 47(W1):W556–W560, 2019. 2, 3, 4

[11] Tanya Barrett, Stephen E. Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F. Kim, Maxim Tomashevsky, Kimberly A. Marshall, Kather-ine H. Phillippy, Patti M. Sherman, Michelle Holko, et al. Ncbi geo: archive for functional genomics data sets–update. Nucleic Acids Research, 41(D1):D991–D995, 2013. 2

[12] Ron Edgar, Michael Domrachev, and Alex E. Lash. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. Nucleic Acids Research, 30(1): 207–210, 2002. 2

[13] Liviu Badea, Vlad Herlea, Simona Olimpia Dima, Traian Dumitrascu, Irinel Popescu, et al. Combined gene expression analysis of whole-tissue and microdissected pancreatic ductal adenocarcinoma identifies genes specifically overexpressed in tumor epithelia-the authors reported a com-bined gene expression analysis of whole-tissue and microdissected pancre-atic ductal adenocarcinoma identifies genes specifically overexpressed in tumor epithelia. Hepato-gastroenterology, 55(88):2016, 2008. 2

[14] Shouhui Yang, Wei Tang, Azadeh Azizian, Jochen Gaedcke, Philipp Ströbel, Limin Wang, Helen Cawley, Yuuki Ohara, Paloma Valenzuela, Lin Zhang, et al. Dysregulation of hnf1b/clusterin axis enhances disease progression in a highly aggressive subset of pancreatic cancer patients.Carcinogenesis, 43(12):1198–1210, 2022. 2

[15] Geng Zhang, Aaron Schetter, Peijun He, Naotake Funamizu, Jochen Gaed-cke, B Michael Ghadimi, Thomas Ried, Raffit Hassan, Harris G Yfantis, Dong H Lee, et al. Dpep1 inhibits tumor cell invasiveness, enhances chemosensitivity and predicts clinical outcome in pancreatic ductal adeno-carcinoma. PloS one, 7(2):e31507, 2012. 2

[16] Matthew E. Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. limma powers differential expression anal-yses for rna-sequencing and microarray studies. Nucleic Acids Research, 43(7):e47, 2015. 2

[17] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B (Methodological), 57(1):289–300, 1995. 2

[18] R Core Team. R: A Language and Environment for Statistical Computing.R Foundation for Statistical Computing, Vienna, Austria, 2023. 2

[19] Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, et al. Welcome to the tidyverse. Journal of Open Source Software, 4(43):1686, 2019. 2

[20] Hanbo Chen and Paul C. Boutros. Venndiagram: a package for the generation of highly-customizable venn and euler diagrams in r. BMC Bioinformatics, 12:35, 2011. 2

[21] Jake R. Conway, Alexander Lex, and Nils Gehlenborg. Upsetr: an r package for the visualization of intersecting sets and their properties. Bioinformatics, 33(18):2938–2940, 2017. 2

[22] Marcin Pilarczyk, Mehdi Fazel-Najafabadi, Michal Kouril, Behrouz Sham-saei, Juozas Vasiliauskas, Wen Niu, Naim Mahi, Lixia Zhang, Nicholas A. Clark, Yan Ren, et al. Connecting omics signatures and revealing bio-logical mechanisms with ilincs. Nature Communications, 12:1670, 2021. 2

[23] Edward Y. Chen, Christopher M. Tan, Yan Kou, Qiaonan Duan, Zichen Wang, Gabriela Vaz Meirelles, Neil R. Clark, and Avi Ma'ayan. Enrichr: interactive and collaborative html5 gene list enrichment analysis tool.BMC Bioinformatics, 14:128, 2013. 2, 3

[24] Maxim V. Kuleshov, Matthew R. Jones, Andrew D. Rouillard, Nicolas F. Fernandez, Qiaonan Duan, Zichen Wang, Simon Koplev, Sherry L. Jenkins, Kathleen M. Jagodnik, Alexander Lachmann, et al. Enrichr: a comprehen-sive gene set enrichment analysis web server 2016 update. Nucleic Acids Research, 44(W1):W90–W97, 2016. 2, 3

[25] Minoru Kanehisa and Susumu Goto. Kegg: Kyoto encyclopedia of genes and genomes. Nucleic Acids Research, 28(1):27–30, 2000. 2, 3

[26] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, et al. Gene ontology: tool for the unification of biology. Nature Genetics, 25(1):25–29, 2000. 2, 3

[27] The Gene Ontology Consortium. The gene ontology resource: enriching a gold mine. Nucleic Acids Research, 49(D1):D325–D334, 2021. 2, 3

[28] Bijay Jassal, Lisa Matthews, Guilherme Viteri, Chuqiao Gong, Pascual Lorente, Antonio Fabregat, Konstantinos Sidiropoulos, Justin Cook, Marc Gillespie, Robin Haw, et al. The reactome pathway knowledgebase. Nu-cleic Acids Research, 48(D1):D498–D503, 2020. 2, 4

[29] Damian Szklarczyk, Annika L. Gable, Katerina C. Nastou, David Lyon, Rebecca Kirsch, Sampo Pyysalo, Nadezhda T. Doncheva, Marc Legeay, Tao Fang, Peer Bork, et al. The string database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Research, 49(D1):D605–D612, 2021. 3, 4

[30] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S. Baliga, Jonathan T. Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Research, 13(11):2498–2504, 2003. 3, 4

[31] Nadezhda T. Doncheva, John H. Morris, Jan Gorodkin, and Lars J. Jensen. Cytoscape stringapp: Network analysis and visualization of proteomics data. Journal of Proteome Research, 18(2):623–632, 2019. 3

[32] John N. Weinstein, Eric A. Collisson, Gordon B. Mills, Kenna R. Mills Shaw, Brad A. Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M. Stuart. The cancer genome atlas pan-cancer analysis project.Nature Genetics, 45(10):1113–1120, 2013. 3

[33] John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Ed-mund Lo, Saboor Shad, Richard Hasz, Gary Walters, Fernando Garcia, Nancy Young, et al. The genotype-tissue expression (gtex) project. Nature Genetics, 45(6):580–585, 2013. 3

[34] J. Martin Bland and Douglas G. Altman. The logrank test. BMJ, 328(7447):1073, 2004. 3, 4

[35] David R. Cox. Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological), 34(2):187–202, 1972. 3, 4

[36] Kai Chen, Qi Wang, Xinxin Liu, Feng Wang, Yongsu Ma, Shupeng Zhang, Zhijiang Shao, Yinmo Yang, and Xiaodong Tian. Single cell rna-seq identifies immune-related prognostic model and key signature-spp1 in pancreatic ductal adenocarcinoma. Genes, 13(10):1760, 2022. 5

[37] Xu Han, Hexige Saiyin, Junjie Zhao, Yuan Fang, Yefei Rong, Chenye Shi, Wenhui Lou, and Tiantao Kuang. Overexpression of mir-135b-5p promotes unfavorable clinical characteristics and poor prognosis via the repression of sfrp4 in pancreatic cancer. Oncotarget, 8(37):62195–62207, 2017. 5

[38] Krista Marie Vincent and Lynne-Marie Postovit. A pan-cancer analysis of secreted frizzled-related proteins: re-examining their proposed tumour suppressive function. Scientific Reports, 7:42719, 2017. 5