LEWIS UNIVERSITY

IDENTIFYING COMMON POTENTIAL DRUG TARGETS AND DIAGNOSTIC BIOMARKERS THROUGH DIFFERENTIALLY EXPRESSED GENES AND RECURSIVE FEATURE ELIMINATION IN PANCREATIC DUCTAL ADENOCARCINOMA AND HEPATIC CELLULAR CARCINOMA

RESEARCH PROJECT SUBMITTED
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE, DATA SCIENCE

BY
AMY NOYES

DIRECTOR: DR. PIOTR SZCZUREK

ROMEOVILLE, IL
DATE: JUNE 2021

ABSTRACT

The tumor microenvironment is complex, different for each cancer type, affects the growth of cancer cells and the prognosis and treatment of cancer patients. New drug therapies are needed to target the components of the tumor microenvironment. Pancreatic ductal adenocarcinoma (PDAC) is the fourth leading cause of cancer death in the United States with 44,915 deaths per year. PDAC patients face a poor prognosis due to many factors including late diagnosis, a complex tumor microenvironment, and lack of treatment options. Hepatic cellular carcinoma (HCC) is the ninth leading cause of cancer death in the United States and has poor prognosis due to late detection leading to limited treatment options. Pancreatic ductal adenocarcinoma and hepatic cellular carcinoma were chosen to identify new diagnostic biomarker and therapeutic targets due to the difficulty in diagnosis and lack of therapeutic treatments.

This study used PDAC and HCC gene expression datasets from the Gene Expression Omnibus (GEO) to identify differentially expressed genes (DEG's). A supervised machine-learning algorithm, support vector machine-recursive feature elimination (SVM-RFE), was used to select the top genes that distinguish between the classes (tumor samples versus non-disease samples). The algorithm ranks the features or genes according to the importance of the gene in separating the cancer samples from the non-disease samples. The features with the lowest ranks are removed until the specified number of features is reached. The SVM-RFE selected results were compared between PDAC and HCC to identify common genes as possible drug targets. Figures were generated with mRNA expression for selected genes across all cancer types to identify future cancers to test the method. Protein-protein interaction networks (PPI) were

generated using the SVM-RFE selected genes to identify interacting genes that are known cancer driver genes and candidate therapeutic targets.

The top 10 downregulated and top 10 upregulated genes were identified for both disease types. Downregulated genes were considered as possible diagnostic biomarkers and upregulated genes as possible therapeutic targets. The collagen triple helix repeat containing 1 (CTHRC1) gene was upregulated in both PDAC and HCC. It has found to be overexpressed in multiple cancer types.

SVM-RFE was performed with 10 features on both PDAC (89% accuracy) and HCC (92% accuracy). The PDAC dataset had three downregulated genes (STAB2, LMX1A-AS2, and UGT3A1) and three upregulated genes (PLK3, GCC2, and CCN4) related to cancer. In addition, the PPI network produced 11 upregulated and seven downregulated interacting proteins.

The HCC dataset had four cancer related downregulated genes (CFP1, NROB2, CACNG4, and RCN3) and three upregulated cancer related genes (SCLY, NOL3, and SHQ1). The PPI network provided three additional drug targets and four diagnostic biomarkers.

This study provides a method to identify new therapeutic targets and diagnostic biomarkers by finding DEGs, selecting genes based on their classification in the tumor subgroup, and identifying additional genes using cancer-specific PPI networks.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Differentially expressed genes	DEGs
Empirical Bayes	eBayes
Gene Expression Omnibus	GEO
Hepatic cellular carcinoma	HCC
Pancreatic ductal adenocarcinoma	PDAC
Protein Interaction Network Analysis	PINA
Protein-protein interaction	PPI
Robust multi-array average expression measure	RMA
Support vector machine-recursive feature elimination	SVM-RFE

CHAPTER I - INTRODUCTION

A. The Tumor Microenvironment

The tumor microenvironment consists of stromal cells including fibroblasts, immune cells, and endothelial cells and extracellular matrix components including fibronectin and collagen. Tumor cells control the function of components in the tumor microenvironment through complex signaling networks of cytokines, chemokines, and growth factors. The tumor microenvironment plays an important role in the growth and invasion of cancer cells and effects cancer progression and treatment (Baghban, Roshangar, Jahanban-Esfahlan, & al., 2020). Changes in the tumor microenvironment can occur through physical cell-to-cell contact, via secreted signals, and via extracellular matrix to cell interactions. New drug therapies can target components of the tumor microenvironment and diagnostic biomarkers can measure differences between disease-free patients and patients with tumors (Privat-Maldonado, Bengtson, Razzokov, Smits, & Bogaerts, 2019).

B. Pancreatic Ductal Adenocarcinoma

Pancreatic ductal adenocarcinoma (PDAC) accounts for more than 90% of pancreatic cancers. It is the fourth leading cause of cancer death in the United States with 44,915 individual deaths in 2018 (Siegel, Miller, Fuchs, & Jemal, 2021). In general, PDAC patients have a poor prognosis due to a lack of symptoms leading to late diagnosis, aggressive growth, resistance to treatment, lack of tumor markers, and complex tumor microenvironments (Kleeff, Korc, Apte, & al., 2016). Treatment of PDAC is mainly limited to surgery, radiation, and chemotherapy with

less than 10% survival rate at five years (Sarantis, Koustas, Papadimitropoulou, Papavassiliou, & Karamouzis, 2020). There is currently only one targeted therapy approved for PDAC, pembrolizumab, an anti-PD1 inhibitor (U.S. Food and Drug Administration, 2021).

C. Hepatic Cellular Carcinoma

Hepatic cellular carcinoma (HCC) is the most common liver cancer and is the ninth leading cause of cancer deaths (Siegel, Miller, Fuchs, & Jemal, 2021). Risk factors for HCC include viral infections, alcohol abuse, autoimmune hepatitis, and obesity. Viral liver injury can affect cell process causing issues with cell growth, survival, and transformation. The mechanisms of carcinogenesis vary depending on diverse factors leading to numerous mechanisms of action making treatment difficult (Singh, Kumar, & Pandey, 2018). Due to the lack of early detection, treatment options are limited. Treatments include surgery, multi-tyrosine kinase inhibitors, and monoclonal antibody therapies targeting PD1, PDL1, and VEGF (Sangro, Sarobe, Hervas-Stubbs, & Melero, 2021).

D. One Drug, Multiple Targets

Ideally, from a drug development standpoint one drug that can treat multiple cancer types would be especially beneficial in terms of reduced development time and reduced cost to patients. An example of an approved drug that treats multiple cancer types is atezolizumab (commercial name: Tecentriq). Atezolizumab treats liver cancer, small cell lung cancer, non-small cell lung cancer, triple-negative breast cancer, melanoma, and bladder cancer. Atezolizumab is a monoclonal antibody against PD-L1, a ligand on tumor cells and tumor-infiltrating immune cells, that restores T-cell activity (Genentech, A Member of the Roche Group, 2021).

E. Objectives of Study

The proposed outcome of this study was to identify common PDAC and HCC genes for drug targets or diagnostic biomarkers. PDAC and HCC were chosen because they effect a large number of people, are difficult to diagnose, and lack treatment options. Cancer datasets from the Gene Expression Omnibus (GEO) were analyzed to identify dysregulated genes. Machine learning was then used to select the most relevant genes. Selected genes were compared between cancer types to identify common genes. Protein-protein interaction networks (PPI) with cancer context were constructed to confirm the selected genes are related to cancer and to identify additional genes. Literature mining was used to identify any previous cancer research on the selected genes. Upregulated genes selected with the SVM-RFE algorithm and upregulated genes with poor survival prognosis from the PPI network were identified as potential therapeutic targets. Downregulated genes selected with the SVM-RFE algorithm and downregulated genes with good and poor survival prognosis from the PPI network were identified as potential diagnostic biomarkers.

CHAPTER II - METHODS

A. Data Collection

The datasets for this study were Affymetrix U133 plus 2.0 microarray data obtained from the Gene Expression Omnibus (GEO) (Table 1) (Edgar, Domrachev, & Lash, 2002). GSE15471 contained 36 matched normal and tumor tissue samples from PDAC patients (Badea, Herlea, Dima, Dumitrascu, & Popescu, 2008). GSE29721 contained 10 matched normal and tumor tissue samples from HCC patients (Bhattacharyya, et al., 2011). Dataset GSE84402 contained 14 matched normal and tissue samples from HCC patients older than 40 years (Wang, et al., 2017). GSE101685 contained eight normal tissue samples and eight HCC samples (Lee, 2021).

B. Identifying Dysregulated Genes (Differentially Expressed Genes)

Referenced R (version 4.1.0) and Python (version 3.8.1) code is available in the Appendix.

The datasets were downloaded as .cel files and processed using the affy() function from the Bioconductor project in R (Gautier, Cope, Bolstad, & Irizarry, 2004). The data was then normalized by the robust multi-array average expression measure (RMA). The RMA function performs quantile normalization to correct for variation between arrays, background correction to correct for spatial variation between arrays, normalization of probes to correct variation within probe sets, and converts the final data into log2 to improve the distribution of the data.

Table 1. Summary of cancer datasets.

Pancreatic Ductal Adenocarcinoma

GEO Series	Sample IDs	Number of Normal Samples	Number of Tumor Samples
GSE15471	GSM388076 - GSM388153	36	36
	Total	**36**	**36**

Hepatic Cellular Carcinoma

GEO Series	Sample IDs	Number of Normal Samples	Number of Tumor Samples
GSE29721	GSM737065 - GSM737084	10	10
GSE84402	GSM2233086 - GSM2233113	14	14
GSE101685	GSM2711996 - GSM2712021	8	8
	Total	**32**	**32**

Limma (Ritchie, et al., 2015) was used to fit a linear model and calculate the empirical Bayes (eBayes) statistic to generate t-statistics, moderated F-statistic, and log-odds for ranking of the genes for differential expression. eBayes uses a Bayesian hierarchical model for gene wise variances with the prior distribution estimated from the marginal distribution of the observed data. Using eBayes improves false discovery rate and statistical power when sample sizes are small (Phipson, Lee, Majewski,, Alexander, & Smyth, 2016).

The limma output was then annotated with the gene symbol and name using the AnnotationDbi package in R (Pagès, Carlson, Falcon, & L, 2021) and the annotation package hgu133plus2.db (Carlson, 2021). Genes with p-values greater than 0.05 (5% probability of finding the observed results when the null hypothesis is true or that there is no difference between tumor and disease-free patients) and log fold changes greater than 0.667 and less than 1.5 were removed from the datasets to identify differentially expressed genes (DEGs) (Lu, Chen, Shan, & Yang, 2016). Fold changes less than 0.667 indicate downregulated genes. Fold changes greater than 1.5 indicate upregulated genes (Lu, Chen, Shan, & Yang, 2016). Data cleaning was performed by removing duplicate genes and removing rows with NA values. The cleaned data was used to generate volcano plots in Python.

C. Identifying Most Relevant Genes (SVM-RFE)

The raw .cel files were also read and processed using the gcrma package from Bioconductor (Wu & Irizarry, 2021). The gcrma package adjusts the background including optical noise and non-specific binding. The output was again annotated with gene information. Data cleaning, including limiting the data to differentially expressed genes was performed.

Recursive feature elimination (SVM-RFE) was performed using scikit-learn in Python (Pedregosa, et al., 2011) to identify the top genes that distinguish the tumor sample subgroup

from the non-disease sample subgroup. SVM-RFE is a supervised machine-learning algorithm that classifies data points by identifying a hyperplane as far as possible from two classes (Huang, et al., 2018). SVM-RFE trains the classifier then iteratively computes the ranking weights for all features and sorts the features according to weight vectors as the classification basis. Then the features with the lowest ranks are removed until the specified number of features are reached. If the model is successful the genes not related to cancer are eliminated (Yan, et al., 2020) (Guyon, Weston, Barnhill, & Vapnik, 2002).

First, SVM-RFE k-fold cross-validation was performed to determine the optimal number of features to select based on the models accuracy. The dataset was randomly shuffled and split into 10 groups or folds. RFE was performed using one group for testing the model and the remaining groups to train the model. This process was repeated until all groups were used exactly once to test the model. The cross-validation was repeated three times and the mean classification accuracy (percentage) and mean standard deviation reported for number of features. The SVM-RFE algorithm was then run using the optimal number of features (highest accuracy and lowest standard deviation) for each disease types.

D. Protein-Protein Interaction Network

Protein-protein interaction networks (PPI) illustrate biological processes and cell function. PPI networks were generated using the Protein Interaction Network Analysis (PINA) v3.0 database (Du, et al., 2021). PINA integrates the human interactome (PPI's from five databases) with transcriptomic profiles (The Cancer Genome Atlas), proteomic profiles (Clinical Proteomic Tumor Analysis Consortium), drug targets (Genomics of Drug Sensitivity in Cancer), and cancer driver genes (The Cancer Genome Atlas) to provide PPI networks with cancer context. PPI networks were generated showing the SVM-RFE selected genes interaction with

tumor type-specific expression. PINA also displays cancer driver genes and candidate therapeutic targets for each PPI network. Kaplan-Meier survival curves and mRNA expression across all cancer types were generated for each selected gene using PINA.

E. Identifying Common Biomarkers and Literature Mining

The selected genes identified by SVM-RFE were compared between the two cancer types to identify common biomarkers. The function of the selected genes from SVM-RFE and the interacting genes in the PPI networks were obtained through GeneCards (Stelzer G, et al., 2016). Literature mining was performed on selected genes through PubMed search (National Center of Biotechnology Information, 2020).

CHAPTER III - RESULTS

A. Top Upregulated and Downregulated Genes

There were 19,866 differentially expressed genes identified in the PDAC dataset with 13,615 downregulated genes and 6,251 upregulated genes. The HCC datasets had 8,426 DEGs with 4,330 downregulated genes and 4096 upregulated genes. The top 10 upregulated and top 10 downregulated genes were identified by their log fold change and are listed in Tables 2 and 3. Volcano plots of each cancer types DEGs are shown in Figures 1 and 2. The gene CTHRC1 was identified as the fifth most upregulated gene in both disease types. Overexpression of CTHRC1 has been found in hepatic cellular carcinoma, melanoma, breast cancer, gastric cancer, and colorectal cancer (Chen, et al., 2019).

B. Top DEGs – Ranked by SVM-RFE

SVM-RFE k-fold cross-validation was used to identify the optimal number of features for each cancer type (Table 4). Twenty features were tested for both PDAC and HCC datasets. Ten features were selected for PDAC with 89% accuracy and a standard deviation of 0.104. Ten features were selected for HCC with 92% accuracy and a standard deviation of 0.079. The top most relevant genes to cancer were selected by SVM-RFE for each cancer type (Tables 5 and 6). There were no common genes between the two datasets generated by SVM-RFE.

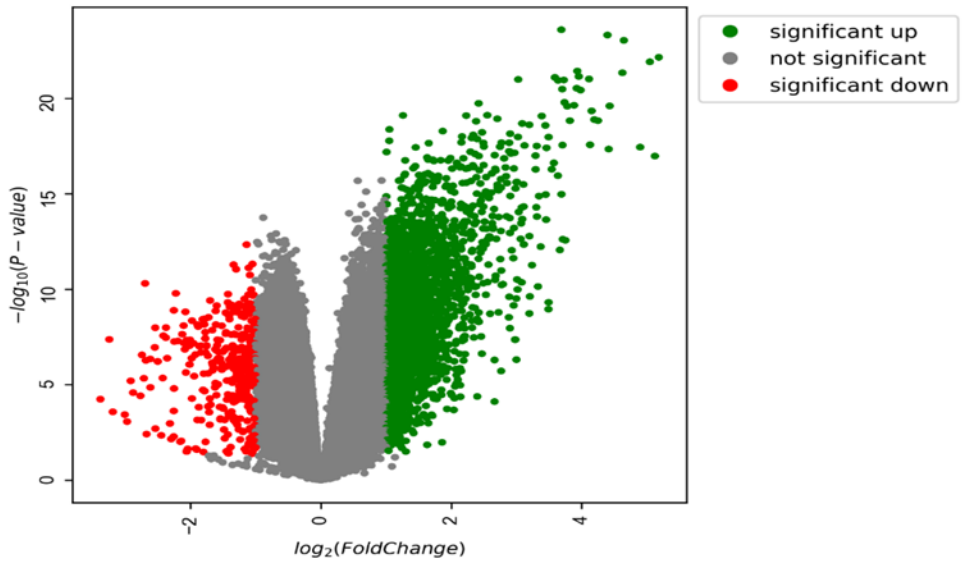
Table 2. Top 20 pancreatic ductal adenocarcinoma differentially expressed genes identified by log fold change.

SYMBOL	GENENAME	logFC	AveExpr	t	P.Value	B	Regulation
CTRL	chymotrypsin like	-3.01	8.99	-3.72	0.000	-0.448	Down
SERPINI2	serpin family I member 2	-2.97	8.11	-3.47	0.001	-1.24	Down
AQP8	aquaporin 8	-2.88	8.24	-4.47	2.601E-05	2.07	Down
TMED6	transmembrane p24 trafficking protein 6	-2.77	7.79	-4.37	3.81E-05	1.70	Down
GNMT	glycine N-methyltransferase	-2.75	6.67	-5.62	2.756E-07	6.45	Down
DNASE1	deoxyribonuclease 1	-2.70	6.67	-7.62	4.934E-11	14.8	Down
SYCN	syncollin	-2.67	9.37	-2.98	0.004	-2.63	Down
GP2	glycoprotein 2	-2.62	6.72	-5.50	4.508E-07	5.97	Down
SLC16A10	solute carrier family 16 member 10	-2.54	5.84	-6.40	1.037E-08	9.63	Down
PNLIPRP1	pancreatic lipase related protein 1	-2.54	7.76	-3.20	0.0020055	-2.02	Down
NTM	neurotrimin	3.96	7.27	1.33E+01	7.11E-22	39.2	Up
THBS2	thrombospondin 2	3.98	9.31	1.29E+01	3.68E-21	37.6	Up
COL1A2	collagen type I alpha 2 chain	4.20	10.7	1.21E+01	1.291E-19	34.1	Up
COL1A1	collagen type I alpha 1 chain	4.25	10.9	1.21E+01	1.464E-19	34.0	Up
COL11A1	collagen type XI alpha 1 chain	4.42	6.78	1.13E+01	4.537E-18	30.7	Up
CTHRC1	**collagen triple helix repeat containing 1**	**4.43**	**9.24**	**1.25E+01**	**2.486E-20**	**35.7**	**Up**
COL8A1	collagen type VIII alpha 1 chain	4.63	7.66	1.34E+01	4.541E-22	39.6	Up
POSTN	periostin	4.90	8.18	1.13E+01	3.596E-18	30.9	Up
COL10A1	collagen type X alpha 1 chain	5.05	7.19	1.38E+01	1.215E-22	40.9	Up
INHBA	inhibin subunit beta A	5.18	7.80	1.39E+01	7.064E-23	41.4	Up

Table 3. Top 20 hepatic cellular carcinoma differentially expressed genes identified by log fold change.

SYMBOL	GENENAME	logFC	AveExpr	t	P.Value	B	Regulation
SLC22A1	solute carrier family 22 member 1	-4.46	8.84	-9.92	1.14E-14	23.1	Down
FCN3	ficolin 3	-4.15	8.56	-13.0	7.01E-20	34.7	Down
MT1M	metallothionein 1M	-4.00	7.68	-9.12	2.81E-13	20.0	Down
GYS2	glycogen synthase 2	-4.00	7.84	-10.9	2.22E-16	26.9	Down
OIT3	oncoprotein induced transcript 3	-3.95	6.86	-14.1	1.85E-21	38.2	Down
CNDP1	carnosine dipeptidase 1	-3.95	6.63	-11.9	4.70E-18	30.7	Down
ADH1B	alcohol dehydrogenase 1B (class I), beta polypeptide	-3.89	8.51	-10.1	6.68E-15	23.6	Down
HAMP	hepcidin antimicrobial peptide	-3.89	9.51	-8.00	2.81E-11	15.5	Down
LINC00844	long intergenic non-protein coding RNA 844	-3.88	7.41	-8.31	7.89E-12	16.7	Down
GLYAT	glycine-N-acyltransferase	-3.87	8.21	-11.7	9.37E-18	30.0	Down
CDKN3	cyclin dependent kinase inhibitor 3	2.97	5.49	1.09E+01	2.04E-16	27.0	Up
CCNB1	cyclin B1	3.07	5.46	1.23E+01	9.36E-19	32.2	Up
CD24	CD24 molecule	3.10	7.17	6.67E+00	6.37E-09	10.18	Up
PEG10	paternally expressed 10	3.17	7.29	6.42E+00	1.75E-08	9.20	Up
TOP2A	DNA topoisomerase II alpha	3.21	5.61	1.23E+01	1.16E-18	32.0	Up
CTHRC1	**collagen triple helix repeat containing 1**	**3.32**	**5.89**	**7.64E+00**	**1.20E-10**	**14.0**	**Up**
ASPM	assembly factor for spindle microtubules	3.34	5.90	1.29E+01	1.30E-19	34.1	Up
SULT1C2	sulfotransferase family 1C member 2	3.40	5.74	9.85E+00	1.48E-14	22.8	Up
GPC3	glypican 3	3.73	7.96	7.08E+00	1.20E-09	11.8	Up
SPINK1	serine peptidase inhibitor Kazal type 1	4.04	8.78	6.24E+00	3.60E-08	8.49	Up

(A) All Data



(B) Adjusted: Fold Change < 0.667 and > 1.5. P-value < 0.05

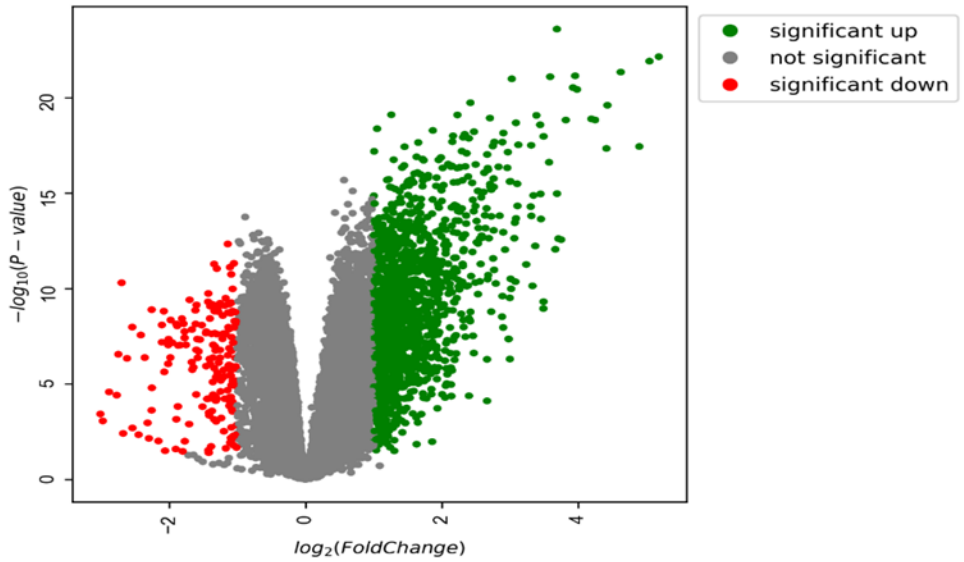
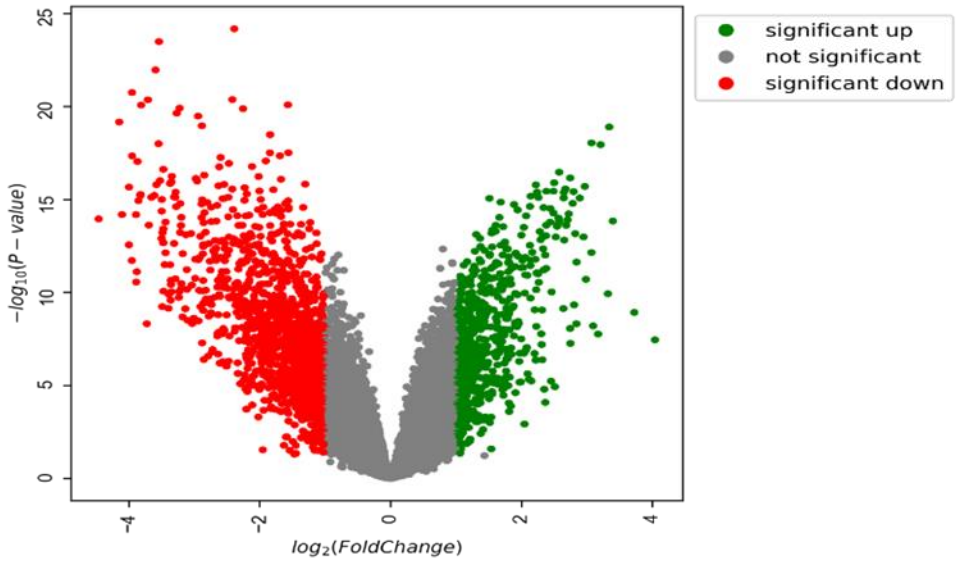


Figure 1. Volcano plot of differentially expressed genes: data from PDAC patients (GEO: GSE15471). (A) 57,193 genes in raw dataset. (B) 22,251 genes after removing duplicates, genes with p-value > 0.05, and fold change between 0.667 and 1.5.

(A) All Data



. (B) Adjusted: Fold Change < 0.667 and > 1.5. P-value < 0.05

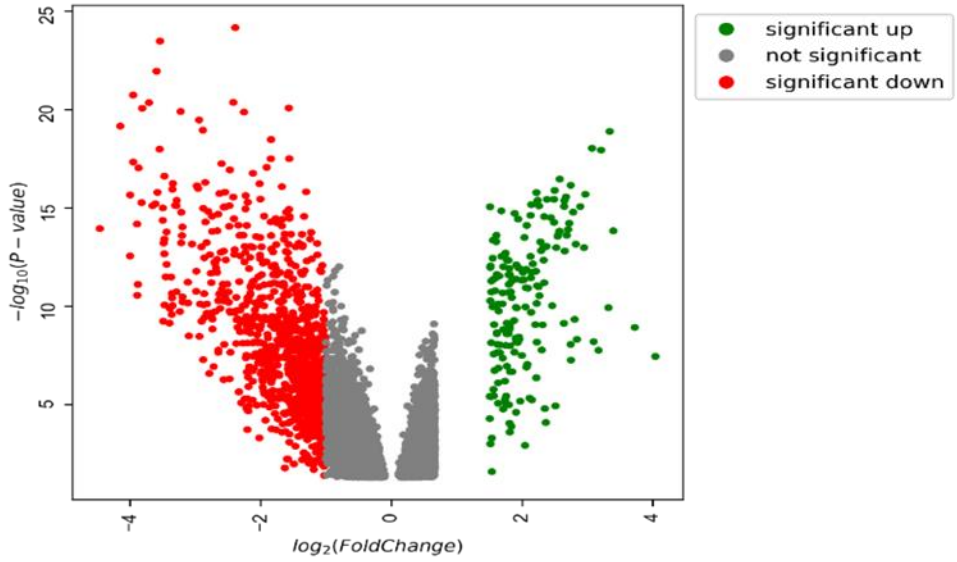


Figure 2. Volcano plot of differentially expressed genes: data from HCC patients (GEO: GSE29721, GSE84402, GSE101685). (A) 57,193 genes in raw dataset. (B) 8,426 genes after removing duplicates, genes with p-value > 0.05, and fold change between 0.667 and 1.5.

Table 4. Optimal number of features for SVM-RFE.

Pancreatic Ductal Adenocarcinoma			Hepatic Cellular Carcinoma		
Features	**Accuracy**	**Standard Deviation**	**Features**	**Accuracy**	**Standard Deviation**
2	87%	0.118	2	89%	0.105
3	88%	0.116	3	92%	0.110
4	86%	0.113	4	90%	0.109
5	87%	0.118	5	92%	0.081
6	87%	0.116	6	91%	0.108
7	88%	0.116	7	90%	0.109
8	86%	0.113	8	92%	0.081
9	88%	0.118	9	89%	0.105
10	**89%**	**0.104**	**10**	**92%**	**0.079**
11	89%	0.104	11	89%	0.105
12	89%	0.104	12	91%	0.108
13	89%	0.104	13	89%	0.105
14	89%	0.104	14	89%	0.105
15	89%	0.104	15	90%	0.109
16	89%	0.104	16	90%	0.109
17	89%	0.104	17	89%	0.105
18	89%	0.104	18	89%	0.105
19	89%	0.104	19	89%	0.105
20	89%	0.104	20	89%	0.105

Table 5. Molecular function of the final selected pancreatic ductal adenocarcinoma genes.

	Symbol	Gene Name	Molecular Function	Associated Diseases	logFc	Regulation
1	STAB2	stabilin 2	May function in angiogenesis, lymphocyte homing, cell adhesion, or receptor scavenging. Receptor that mediates endocytosis of hyaluronic acid	Increased hyaluronic acid in prostate, bladder, lung, breast and other cancers.	-0.618	Down
2	LMX1A-AS2	LMX1A antisense RNA 2	LMX1A-AS2 is an RNA gene affiliated with the lncRNA class.	Deafness, lung cancer	-0.562	Down
3	UGT3A1	UDP glycosyltransferase family 3 member A1	UDP-glucuronosyltransferases catalyzes phase II biotransformation reactions. They are of major importance in the conjugation and subsequent elimination of potentially toxic xenobiotics and endogenous compounds.	Metabolic inactivation of cancer drugs.	-0.483	Down
4	GPR55	G protein-coupled receptor 55	GPR55 is a likely cannabinoid receptor. It may be involved in several physiological and pathological processes by activating a variety of signal transduction pathways.	Colorectal Cancer, Cannabis abuse	-0.457	Down
6	LOC101927093	uncharacterized LOC101927093	Uncharacterized. RNA Gene affiliated with the ncRNA class.	NA	-0.356	Down
8	ZC2HC1B	zinc finger C2HC-type containing 1B	A protein coding gene.	Transient neonatal diabetes mellitus	-0.236	Down
9	LOC105370478	uncharacterized LOC105370478	Uncharacterized. RNA Gene affiliated with the ncRNA class.	NA	-0.190	Down
7	PLK3	polo like kinase 3	Polo-like kinases are important regulators of cell cycle progression. This gene has also been implicated in stress responses and double-strand break repair. Acts as a tumor suppressor.	Increased in non-small lung, head and neck, colorectal, and esophageal cancers.	0.293	Up
5	GCC2	GRIP and coiled-coil domain containing 2	A peripheral membrane protein localized to the trans-Golgi network. It is sensitive to brefeldin A. This encoded protein contains a GRIP domain which is thought to be used in targeting.	Bladder carcinoma and achondrogenesis	0.437	Up
10	CCN4	cellular communication network factor 4	WNT1 is a member of a family of cysteine-rich, glycosylated signaling proteins that mediate diverse developmental processes. This gene is downstream in the WNT1 signaling pathway that is relevant to malignant transformation. It is expressed at a high level in fibroblast cells.	Increased in pancreatic, head and neck, lung, colorectal, breast, brain, prostate, colorectal, esophageal cancers.	3.82	Up

Table 6. Molecular function of the final selected hepatic cellular carcinoma genes.

	Symbol	Gene Name	Molecular Function	Associated Diseases	logFc	Regulation
1	CFP	complement factor properdin	A plasma glycoprotein that positively regulates the alternative complement pathway of the innate immune system. This protein binds to many microbial surfaces and apoptotic cells to form the membrane attack complex and lysis of the target	Downregulated in liver hepatocellular carcinoma and lung adenocarcinoma.	-2.42	Down
2	NR0B2	nuclear receptor subfamily 0 group B member 2	Iinteracts with retinoid and thyroid hormone receptors, inhibiting their ligand-dependent transcriptional activation. Interacts with estrogen receptors leading to inhibition of	Downregulated in liver, kidney, and lung cancer. Upregulated in colon and rectal cancer.	-1.062	Down
3	CACNG4	calcium voltage-gated channel auxiliary subunit	Protein regulates activity of L-type calcium channels that contain CACNA1C as pore-forming subunit. Regulates trafficking and gating properties of AMPA-selective glutamate	Upregulated in breast cancer.	-0.412	Down
4	RCN3	reticulocalbin 3	Probable molecular chaperone assisting protein biosynthesis and transport in the endoplasmic reticulum.	Down-regulated in non-small cell lung cancer.	-0.304	Down
5	CASKIN2	CASK interacting protein 2	Multi-domain scaffolding protein with a role in synaptic transmembrane protein anchoring and ion channel trafficking. Contributes to neural development and regulation of gene expression.	Mental Retardation And Microcephaly.	-0.255	Down
6	SCLY	selenocysteine lyase	Catalyzes the decomposition of L-selenocysteine to L-alanine and elemental selenium.	Upregulated in colon and esophageal cancers.	0.277	Up
7	NOL3	nucleolar protein 3	An apoptosis repressor that blocks multiple modes of cell death.	Upregulated in pancreatic, colorectal, breast, lung, cervical, and prostate cancers.	0.402	Up
8	NPFF	neuropeptide FF-amide peptide precursor	Peptide that modulates morphine-induced analgesia, elevation of arterial blood pressure, and increased somatostatin secretion from the pancreas.	Pain Agnosia	0.415	Up
9	SHQ1	SHQ1, H/ACA ribonucleoprotein assembly factor	Assists in assembly of H/ACA-box ribonucleoproteins that function in the processing of ribosomal RNAs, modification of spliceosomal small nuclear RNAs, and stabilization of	Upregulated in T-acute lymphoblastic leukemia	0.610	Up
10	POLR2J4	RNA polymerase II subunit J4, pseudogene	A Pseudogene.	NA	0.626	Up

i. Top 10 PDAC SVM-RFE Results

Five out of the 10 PDAC SVM-RFE selected genes were related to cancer. Two selected genes are uncharacterized. LOC101927093 and LOC105370478 are RNA genes affiliated with the non-coding RNA class. They were both downregulated in PDAC.

The stabilin 2 (STAB2) gene encodes a transmembrane protein that enhances the engulfment of apoptotic cells. STAB2 is found on the endothelium of liver, lymph nodes, and spleen. It binds to and facilitates the endocytosis of metabolic waste products including hyaluronic acid. Accumulation of hyaluronic acid is implicated in the growth and metastasis of tumor cells (Hirose, et al., 2012). STAB2 was downregulated in PDAC.

The LMX1A antisense RNA 2 (LMX1A-AS2) is an RNA gene that is affiliated with the long non-coding RNA class. LMX1A-AS2 silencing is linked to tumor progression in lung cancer (Wu, et al., 2020). LMX1A-AS2 was downregulated in PDAC.

The UDP glycosyltransferase family 3 member A1 (UGT3A1) encodes an enzyme that is a member of the UDP-glucuronosyltransferase (UGT) family. These enzymes conjugate sugars to amino, carboxyl, or hydroxyl groups of toxic and endogenous compounds for elimination. UGT's can lead to the metabolic inactivation of cancer drugs (Allain, Rouleau, Lévesque, & Guillemette, 2020). UGT3A1 was downregulated in PDAC.

The G protein-coupled receptor 55 (GPR55) gene encodes a receptor that is involved in physiological and pathological processes. The downregulation of GPR55 inhibits activation of an autocrine loop that regulates cell proliferation promoting tumor growth (Pineiro, Maffucci, & Falasca, 2011). In mice, GPR55 promoted pancreatic tumor growth (Ferro, et al., 2018). GPR55 was downregulated in PDAC.

The zinc finger C2HC-type containing 1B (ZC2HC1B) gene encodes a protein that is associated with transient neonatal diabetes mellitus (Stelzer G, et al., 2016). There is no additional information on this protein. ZC2HC1B was downregulated in PDAC.

The polo like kinase 3 (PLK3) gene encodes a protein that is a regulator of cell-cycle progression and plays a role in cellular response to stress. PLK3 expression is decreased in lung cancer, head and neck cancer, and liver cancer. It is increased in ovary and breast cancer (Helmke, Becker, & Strebhardt, 2016). PLK3 was upregulated in PDAC.

The GRIP and coiled-coil domain containing 2 (GCC2) gene encodes a peripheral membrane protein that is required for maintenance of the Golgi structure. It may play a role in transport between the Golgi and recycling endosomes. Associated diseases include bladder carcinoma and achondrogenesis (Stelzer G, et al., 2016). GCC2 was upregulated in PDAC.

The cellular communication network factor 4 (CCN4) gene encodes a cysteine-rich, glycosylated signaling protein that mediates diverse development processes. CCN4 is upregulated in many cancers including pancreatic cancer (Gurbuz & Chiquest-Ehrismann, 2015). CCN4 was upregulated in PDAC.

ii. Top 10 HCC SVM-RFE Results

Seven of the 10 HCC SVM-RFE selected genes were related to cancer.

The complement factor properdin (CFP) gene encodes a protein that is the only positive regulator of the alternative complement pathway. Mutation of CFP in cancer can result in inflammation and tumor progression. It is downregulated in liver hepatocellular carcinoma and lung adenocarcinoma (Mangogna, et al., 2020). CFP was downregulated in HCC.

The nuclear receptor subfamily 0 group B member 2 (NR0B2) gene encodes a protein that is an orphan nuclear receptor, which does not have a known ligand. NR0B2 is a transcriptional repressor that regulates metabolic pathways in the liver, kidney, and pancreas. It is downregulated in most cancers including liver cancer and upregulated in some intestinal cancers (Zhu, et al., 2021). NR0B2 was downregulated in HCC.

The calcium voltage-gated channel auxiliary subunit gamma 4 (CACNG4) gene encodes a protein that helps regulate intracellular calcium levels, cell survival, and homeostasis. Downregulation of CACNG4 results in calcium channels remaining in their open state, a high intracellular calcium level, and decreased tumor growth (Kanwar, et al., 2020). CACNG4 was downregulated in HCC.

The Reticulocalbin 3 (RCN3) gene encodes a protein assisting transport in the endoplasmic reticulum (ER). When RCN3 levels are under-expressed in non-small cell lung cancer patients the ER stress protein GRP78 is upregulated (Hou, et al., 2016). RCN3 was downregulated in HCC.

The CASK interacting protein 2 (CASKIN2) encodes a multi-domain scaffolding protein that contributes to neural development and the regulation of gene expression (Becker, et al., 2020). CASKIN2 was downregulated in HCC.

The selenocysteine lyase (SCLY) gene encodes a protein containing selenium that is over-expressed in colon and esophageal cancers (Jia, Dai, & Zeng, 2020). SCLY was upregulated in HCC.

The nucleolar protein 3 (NOL3) gene is normally highly expressed in heart, brain, and skeletal tissues but deficient elsewhere. It represses apoptosis triggered by hypoxia, hydrogen

peroxide, and the Fas ligand. Wang et al. (2005) found NOL3 in pancreatic, colorectal, breast, lung, cervical, and prostate cancer cell lines. NOL3 was upregulated in HCC.

The neuropeptide FF-amide peptide precursor (NPFF) gene encodes a protein that modulates morphine-induced analgesia. NPFF exerts anti-opioid activity on neurons by an unknown function (Roumy, et al., 2007). NPFF was upregulated in SHH.

The H/ACA ribonucleoprotein assembly factor (SHQ1) gene encodes a protein that is an assembly chaperone of the H/ACA-box ribonucleoproteins that processes ribosomal RNA, modifies spliceosomal small nuclear RNA, and stabilizes telomerase. Increased SHQI expression is essential for T-acute lymphoblastic leukemia cell growth (Su, et al., 2018). SHQ1 was upregulated in HCC.

RNA polymerase II subunit J4, pseudogene (POLR2J4) is a pseudogene that has ubiquitous expression in many tissues (NCBI, 2021). POLR2J4 was upregulated in HCC.

C. PPI Network's

The PPI network displays the SVM-RFE selected gene, interacting proteins, and disease-specific cancer context. Cancer context includes cancer driver genes, drug targets, protein expression, and survival curves. Potential drug targets were identified by over-expression or unknown expression and diagnostic biomarkers by under-expression.

i. PDAC PPI Network

Six of the top 10 PDAC SVM-RFE selected genes had PPI interactions (Figures 3-7). Table 7 shows the interacting proteins from the PPI network with PDAC- specific cancer context. Possible drug targets specific to PDAC were identified (Table 8). The correlation of expression shows the interacting proteins expression for the specific cancer type.

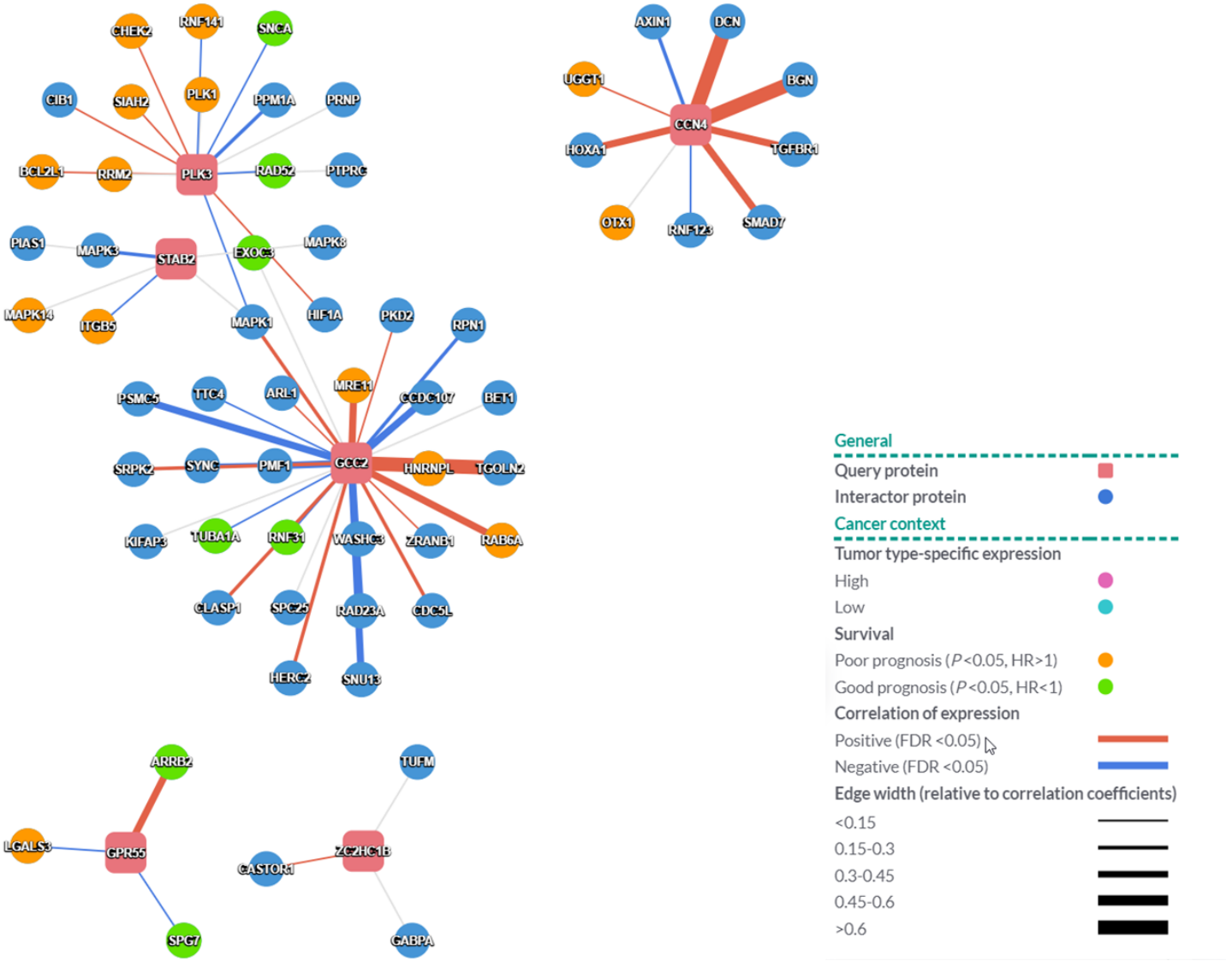


Figure 3. Protein-protein interaction (PPI) network of PDAC top ten dysregulated genes.

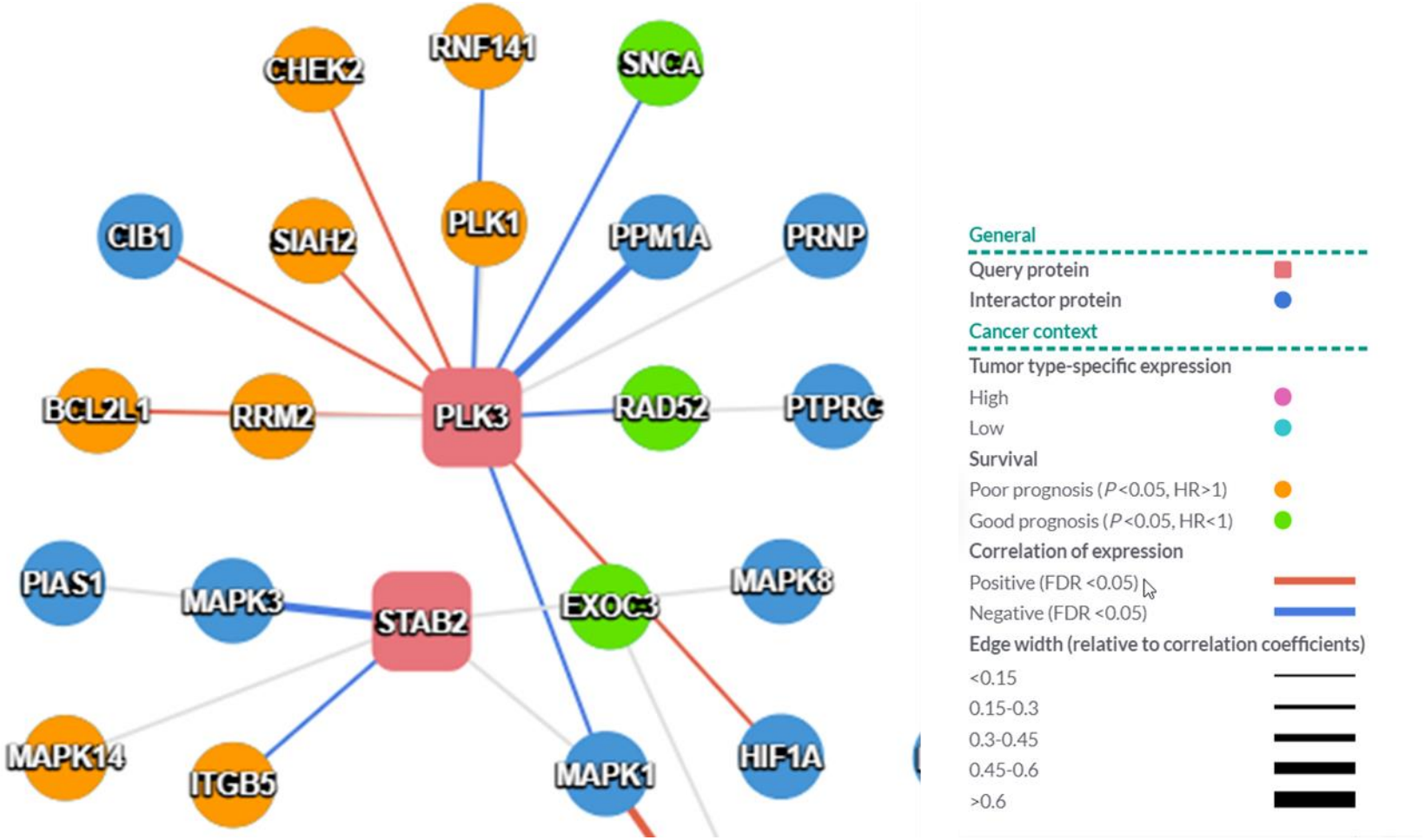


Figure 4. Upper left view of protein-protein interaction (PPI) network of PDAC top ten dysregulated genes.

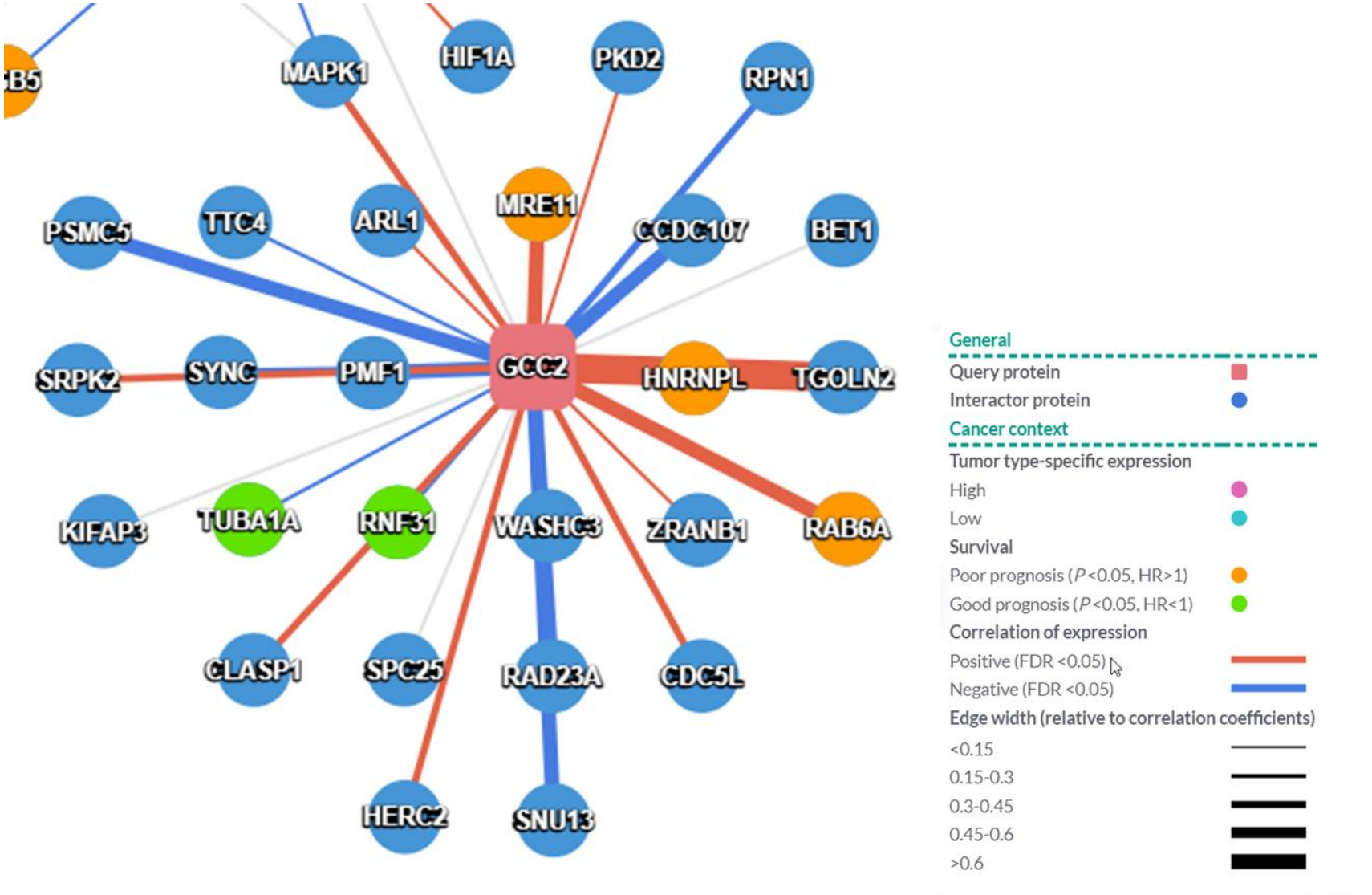


Figure 5. Middle view of protein-protein interaction (PPI) network of PDAC top ten dysregulated genes.

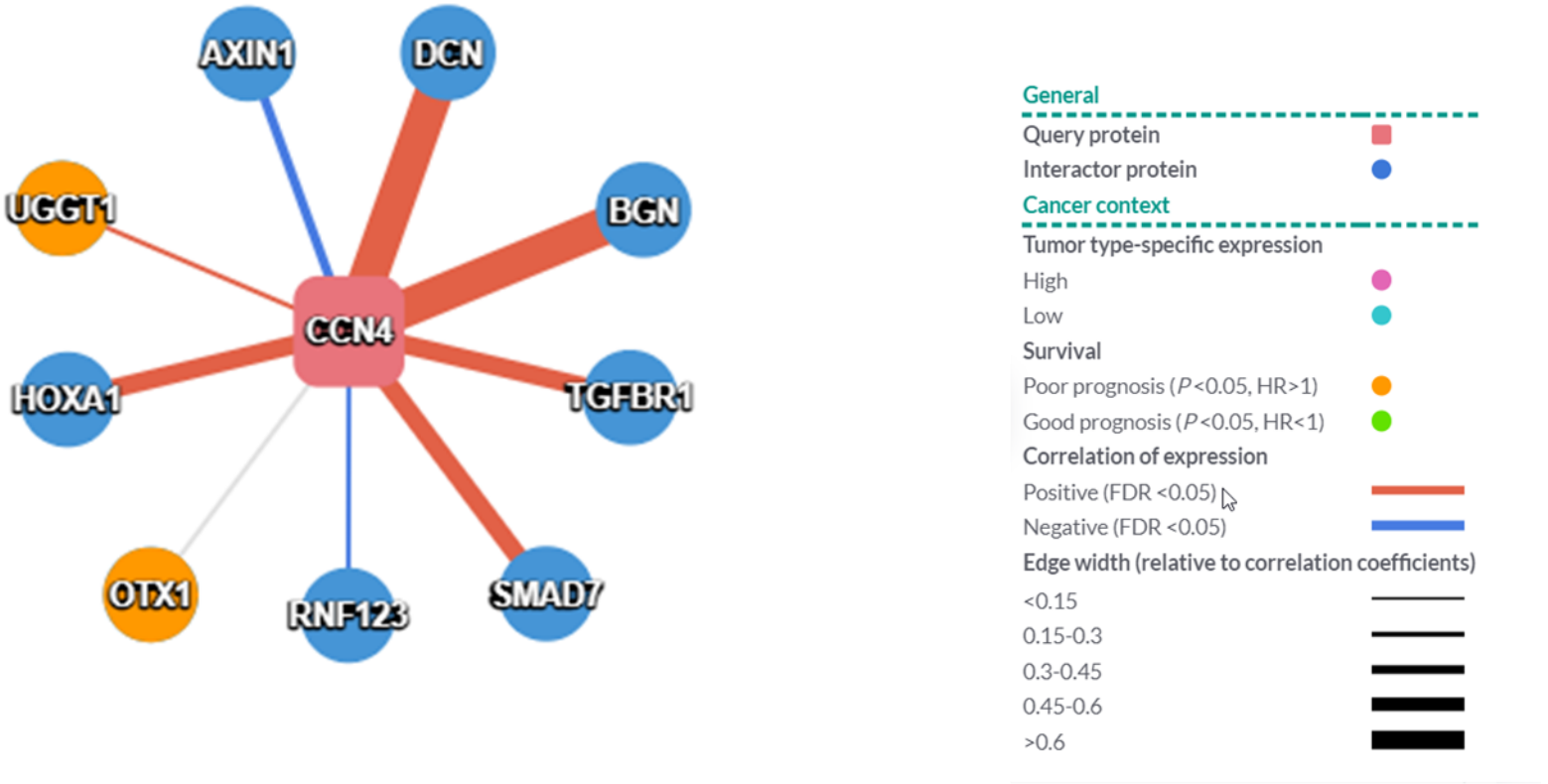


Figure 6. Upper right view of protein-protein interaction (PPI) network of PDAC top ten dysregulated genes.

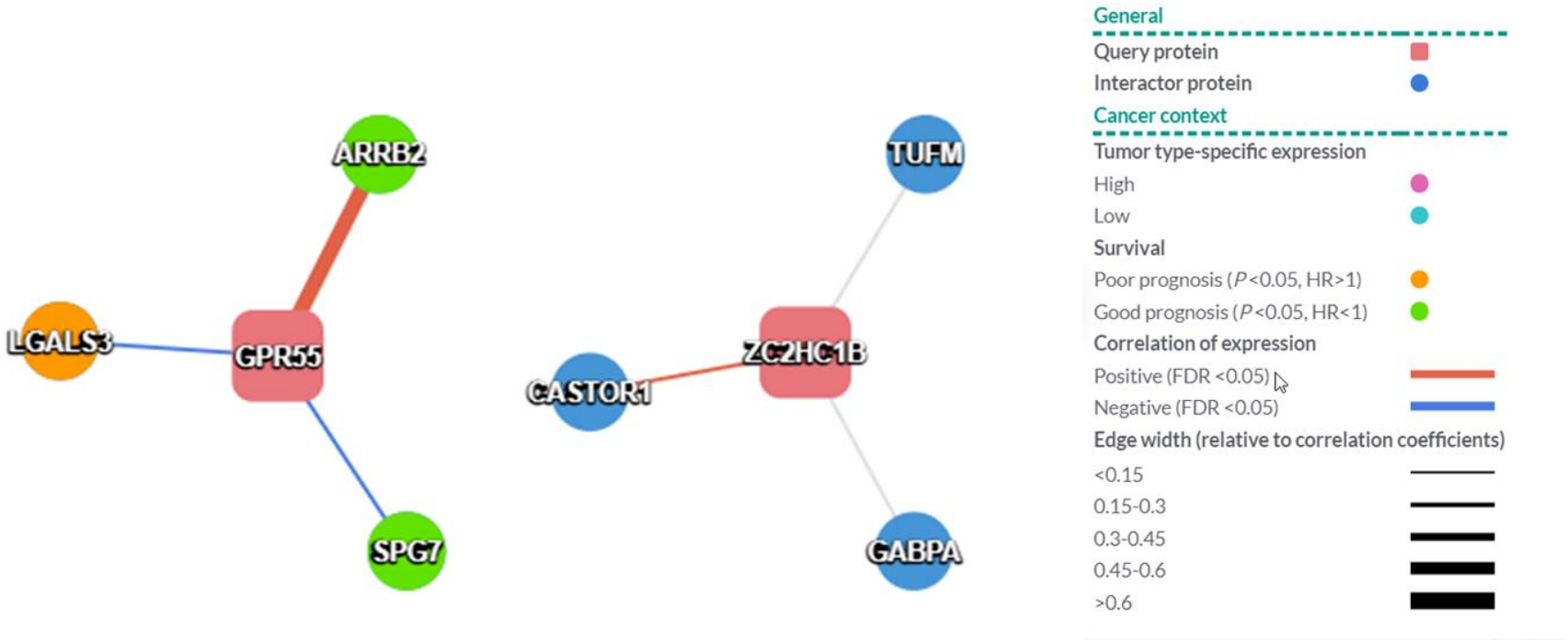


Figure 7. Lower view of protein-protein interaction (PPI) network of PDAC top ten dysregulated genes.

Table 7. Interacting proteins from PPI with PDAC-specific cancer context.

Query Gene: PLK3		**Regulation: Up**			
Interactor Protein	**Gene Name**	**Molecular Function**	**Expression Level**	**Survival Prognosis**	**Correlation of Expression**
CHEK2	Checkpoint Kinase 2	Cell cycle checkpoint regulator and putative tumor suppressor.	NA	Poor	NA
RNF141	Ring Finger Protein 141	Involved in protein-DNA and protein-protein interactions.	NA	Poor	Negative
SIAH2	Siah E3 Ubiquitin Protein Ligase 2	E3 ligase involved in ubiquitination and proteasome-mediated degradation of specific proteins. Implicated in regulating cellular response to hypoxia.	NA	Poor	NA
PLK1	Polo Like Kinase 1	Highly expressed during mitosis. Elevated levels are found in many different types of cancer. Depletion inhibited cell proliferation and induced apoptosis.	NA	Poor	Negative
BCL2L1	BCL2 Like 1	Controls the production of reactive oxygen species and release of cytochrome C, which are potent inducers of cell apoptosis.	NA	Poor	Positive
RRM2	Ribonucleotide Reductase Regulatory Subunit M2	Catalyzes the formation of deoxyribonucleotides from ribonucleotides.	NA	Poor	Positive
SNCA	Synuclein Alpha	May serve to integrate presynaptic signaling and membrane trafficking.	NA	Good	Negative
RAD52	RAD52 Homolog, DNA Repair Protein	Binds ssDNA ends and mediates the DNA-DNA interaction necessary for the annealing of complementary DNA strands.	NA	Good	Negative

Table 7, continued. Interacting proteins from PPI with PDAC-specific cancer context.

Query Gene: STAB2		**Regulation: Down**			
Interactor Protein	**Gene Name**	**Molecular Function**	**Expression Level**	**Survival Prognosis**	**Correlation of Expression**
MAPK14	Mitogen-Activated Protein Kinase 14	Role in stress related transcription, cell cycle regulation, and genotoxic stress response.	NA	Poor	NA
ITGB5	Integrin Subunit Beta 5	Integrins are integral cell-surface receptors that participate in cell adhesion as well as cell-surface mediated signaling.	NA	Poor	Negative
EXOC3	Exocyst Complex Component 3	Component of the exocyst complex, essential for targeting exocytic vesicles to specific docking sites on the plasma membrane.	NA	Good	NA

Query Gene: CCN4		**Regulation: Up**			
Interactor Protein	**Gene Name**	**Molecular Function**	**Expression Level**	**Survival Prognosis**	**Correlation of Expression**
UGGT1	UDP-Glucose Glycoprotein Glucosyltransferase 1	Soluble protein of the ER that selectively reglucosylates unfolded glycoproteins, providing quality control for protein transport out of the ER.	NA	Poor	Positive
OTX1	Orthodenticle Homeobox 1	Acts as a transcription factor and may play a role in brain and sensory organ development.	NA	Poor	NA

Table 7, continued. Interacting proteins from PPI with PDAC-specific cancer context.

Query Gene: GCC2		**Regulation: Up**			
Interactor Protein	**Gene Name**	**Molecular Function**	**Expression Level**	**Survival Prognosis**	**Correlation of Expression**
MRE11	MRE11 Homolog, Double Strand Break Repair Nuclease	A nuclear protein involved in homologous recombination, telomere length maintenance, and DNA double-strand break repair.	NA	Poor	Positive
HNRNPL	Heterogeneous Nuclear Ribonucleoprotein L	Associated with hnRNP complexes. Likely to play a major role in the formation, packaging, processing, and function of mRNA.	NA	Poor	Positive
RAB6A	RAB6A, Member RAS Oncogene Family	Regulates traffic from early endosomes and Golgi to endoplasmic reticulum and from Golgi to the plasma membrane.	NA	Poor	Positive
TUBA1A	Tubulin Alpha 1a	Encoding these microtubule constituents belong to the tubulin superfamily.	NA	Good	Negative
RNF31	Ring Finger Protein 31	E3 ubiquitin-protein ligase component of the linear ubiquitin chain assembly complex.	NA	Good	Positive

Table 7, continued. Interacting proteins from PPI with PDAC-specific cancer context.

Query Gene: GPR55		**Regulation: Down**			
Interactor Protein	**Gene Name**	**Molecular Function**	**Expression Level**	**Survival Prognosis**	**Correlation of Expression**
LGALS3	Galectin 3	Plays a role in numerous cellular functions including apoptosis, innate immunity, cell adhesion, and T-cell regulation.	NA	Poor	Negative
ARRB2	Arrestin Beta 2	Inhibits beta-adrenergic receptor function in vitro. Expressed at high levels in the central nervous system and may play a role in the regulation of synaptic receptors.	NA	Good	Positive
SPG7	SPG7 Matrix AAA Peptidase Subunit, Paraplegin	Members of this protein family have roles in diverse cellular processes including membrane trafficking, intracellular motility, organelle biogenesis, protein folding, and proteolysis.	NA	Good	Negative

Table 8. Possible PDAC drug targets and diagnostic biomarkers derived from RFE selected genes and PPI interactions .

Symbol	Gene Name	Origin	Drug Target/Diagnostic Biomarker
CHEK2	Checkpoint Kinase 2	PPI	Drug Target
SIAH2	Siah E3 Ubiquitin Protein Ligase 2	PPI	Drug Target
BCL2L1	BCL2 Like 1	PPI	Drug Target
RRM2	Ribonucleotide Reductase Regulatory Subunit M2	PPI	Drug Target
MAPK14	Mitogen-Activated Protein Kinase 14	PPI	Drug Target
ITGB5	Integrin Subunit Beta 5	PPI	Drug Target
OTX1	Orthodenticle Homeobox 1	PPI	Drug Target
MRE11	MRE11 Homolog, Double Strand Break Repair Nuclease	PPI	Drug Target
HNRNPL	Heterogeneous Nuclear Ribonucleoprotein L	PPI	Drug Target
RAB64	RAB6A, Member RAS Oncogene Family	PPI	Drug Target
LGALS3	Galectin 3	PPI	Drug Target
PLK3	Polo like kinase 3	RFE	Drug Target
CCN4	Cellular communication network factor 4	RFE	Drug Target
GCC2	GRIP and coiled-coil domain containing 2	RFE	Drug Target
RNF141	Ring Finger Protein 141	PPI	Diagnostic Biomarker
PLK1	Polo Like Kinase 1	PPI	Diagnostic Biomarker
SNCA	Synuclein Alpha	PPI	Diagnostic Biomarker
RAD52	RAD52 Homolog, DNA Repair Protein	PPI	Diagnostic Biomarker
UGGT1	UDP-Glucose Glycoprotein Glucosyltransferase 1	PPI	Diagnostic Biomarker
TUBA1A	Encoding these microtubule constituents belong to the tubulin superfamily.	PPI	Diagnostic Biomarker
SPG7	SPG7 Matrix AAA Peptidase Subunit, Paraplegin	PPI	Diagnostic Biomarker
LMX1A-AS2	LMX1A antisense RNA 2	RFE	Diagnostic Biomarker
UGT3A1	UDP glycosyltransferase family 3 member A1	RFE	Diagnostic Biomarker
LOC101927093	uncharacterized LOC101927093	RFE	Diagnostic Biomarker
ZC2HC1B	Zinc finger C2HC-type containing 1B	RFE	Diagnostic Biomarker
LOC105370478	uncharacterized LOC105370478	RFE	Diagnostic Biomarker
STAB2	Stabilin 2	RFE	Diagnostic Biomarker
GPR55	G protein-coupled receptor 55	RFE	Diagnostic Biomarker

The survival curves displayed in Figure 8 show how the probability of patient survival changed depending on high or low expression of each selected gene. Patients with low gene expression of PLK3, GCC2, and CCN4 have a higher probability of survival after 500 days. All three of the genes were upregulated in the DEG analysis.

Figure 9 shows the mRNA expression of PDAC SVM-RFE selected genes across cancer types. GCC2 has high expression in PDAC (PAAD), stomach adenocarcinoma (STAD), prostate adenocarcinoma (PRAD), and esophageal carcinoma (ESCA). PLK3 has high expression in PDAC (PAAD), thyroid carcinoma (THCA), bladder urothelial carcinoma (BLCA), and cervical squamous cell carcinoma/endocervical adenocarcinoma (CESC). CCN4 has high expression in PDAC (PAAD), sarcoma (SARC), head and neck squamous cell carcinoma (HNSC), breast invasive carcinoma (BRCA), and mesothelioma (MESO).

ii. HCC PPI Network

Five of the top ten HCC SVM-RFE selected genes had PPI interactions (Figures 10-13). Table 9 shows the interacting proteins from the PPI network with HCC-specific cancer context. Possible drug targets and diagnostic biomarkers identified by SVM-RFE and the PPI network are shown in Table 10.

Survival curves show that patients with high gene expression of CFP1, CASKIN2, and SHQ1 have a higher probability of survival after 1500 days (Figure 14). CFP1 and CASKIN2 were both downregulated in the HCC dataset. SHQI was upregulated in the HCC dataset.

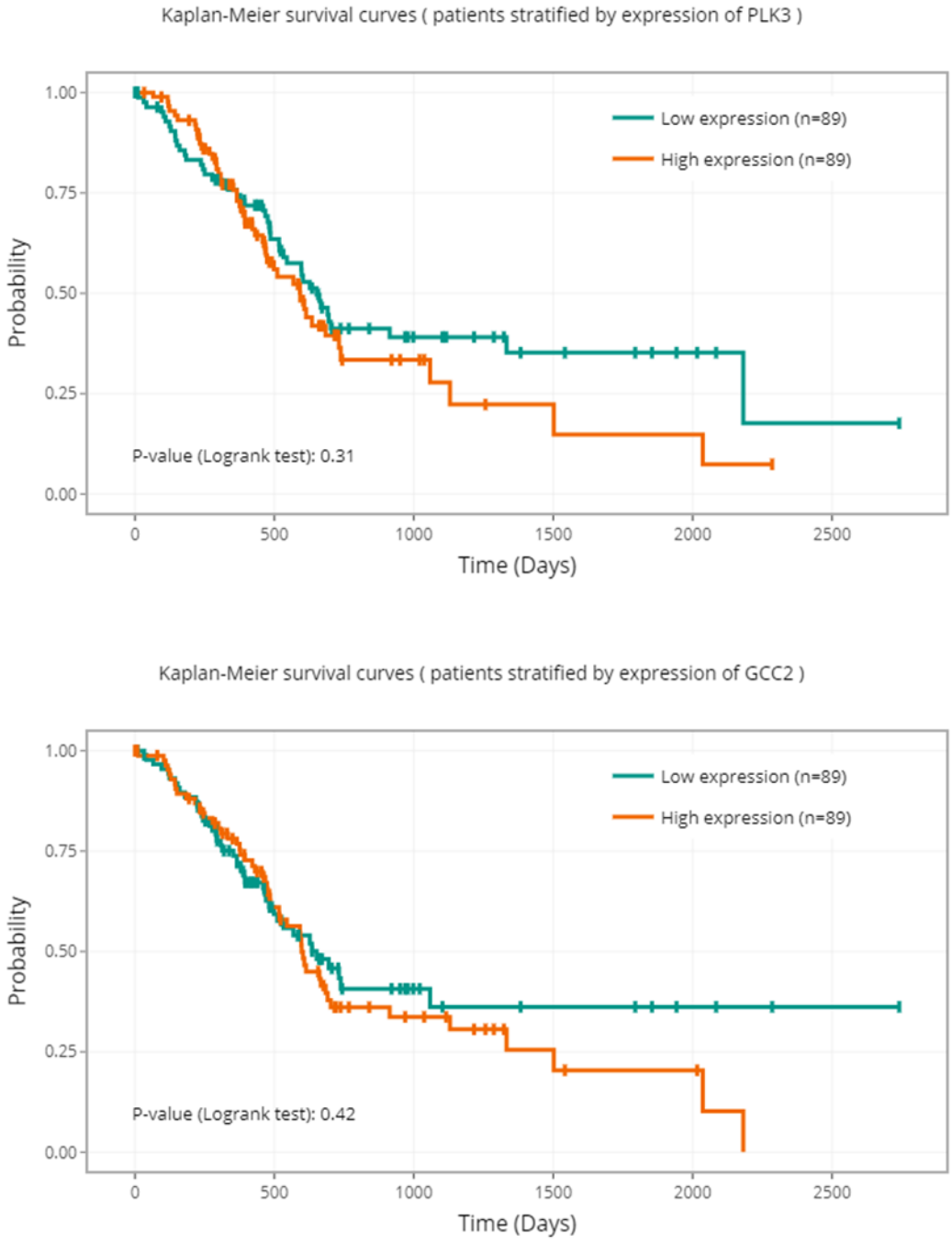


Figure 8. Cancer survivor curves for selected dysregulated genes in PDAC dataset

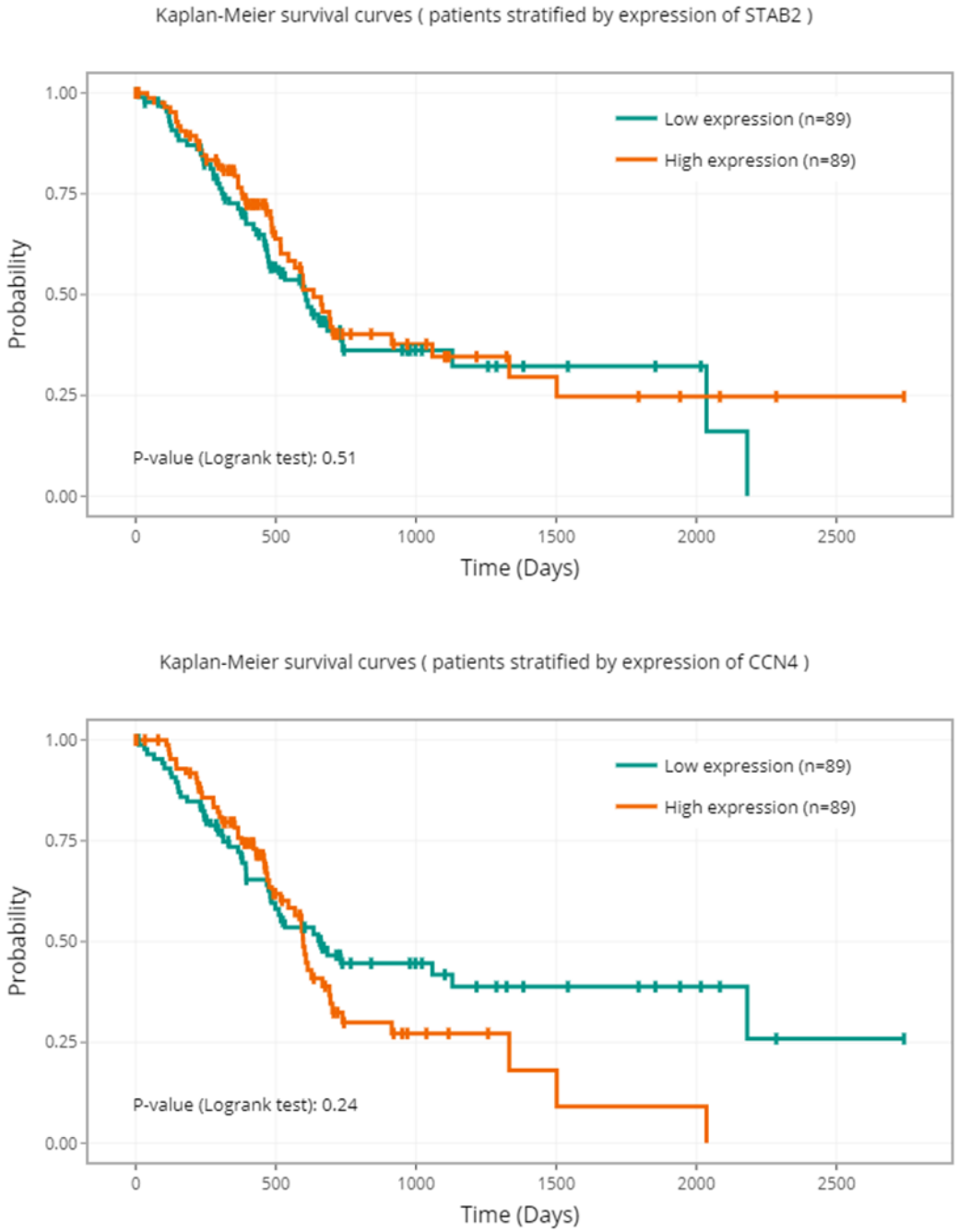


Figure 8, continued. Cancer survivor curves for selected dysregulated genes in PDAC dataset.

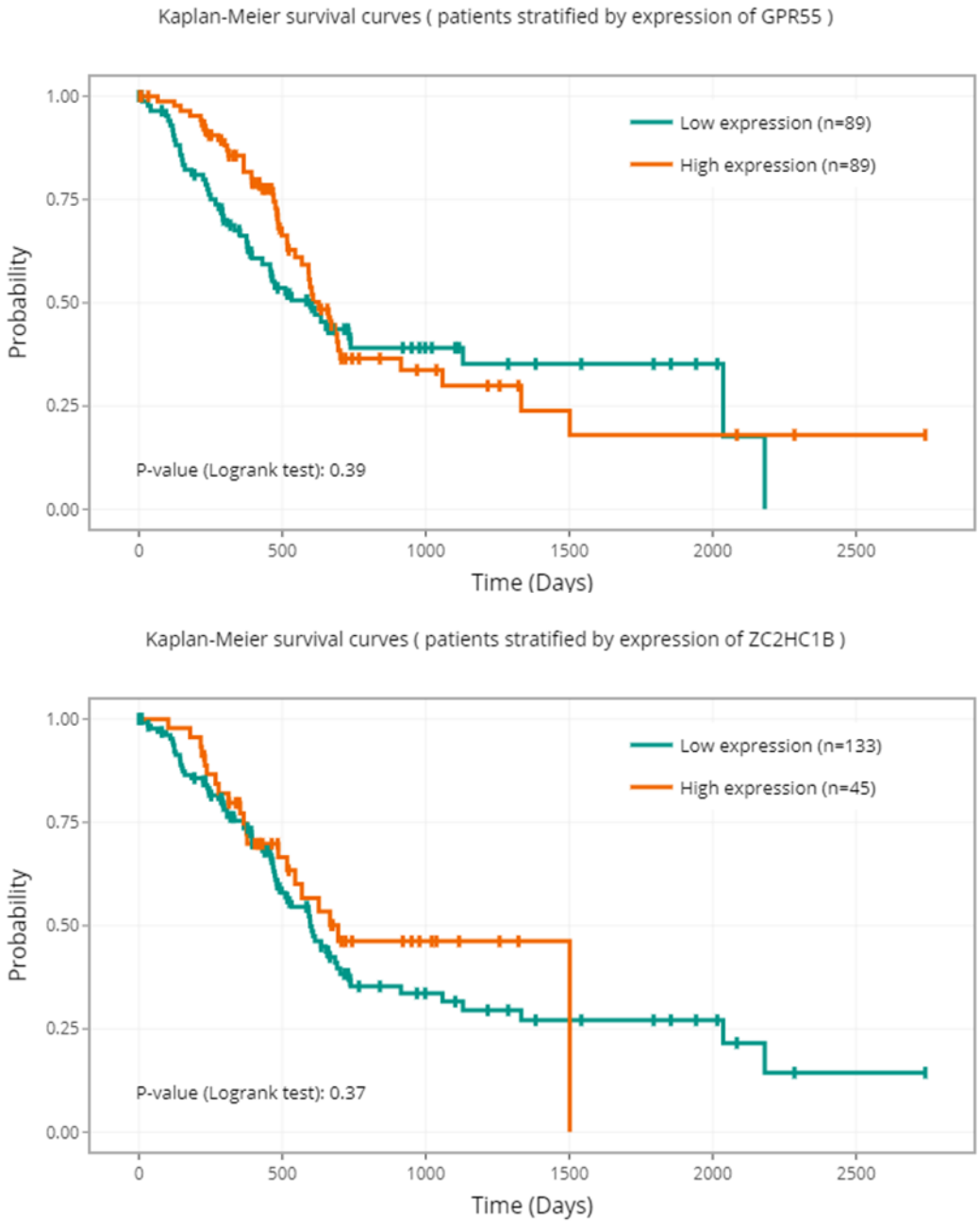


Figure 8, continued. Cancer survivor curves for selected dysregulated genes in PDAC dataset

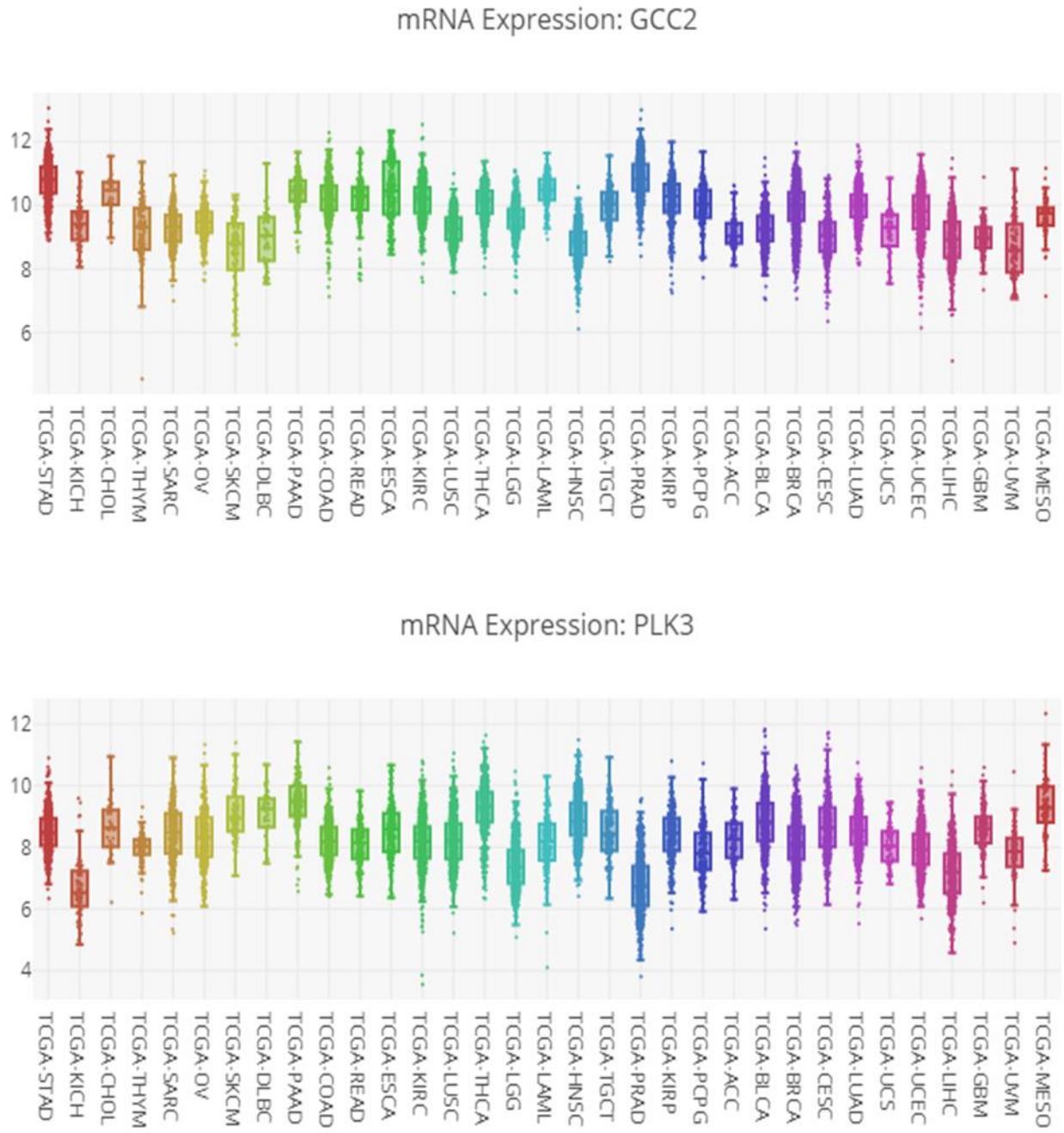


Figure 9. . mRNA expression of RFE selected genes across cancer types. PAAD is pancreatic adenocarcinoma.

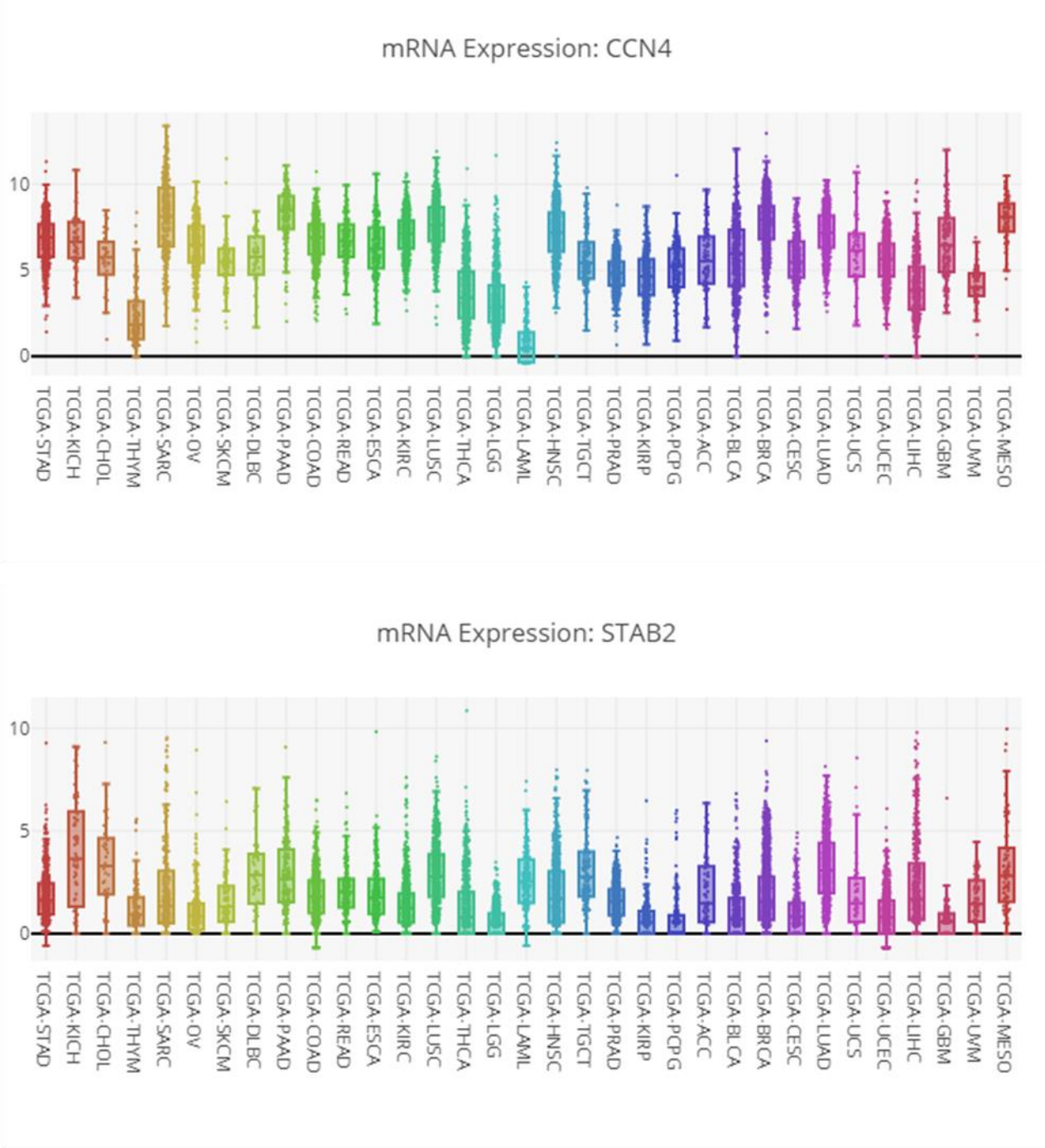


Figure 9, continued. mRNA expression of RFE selected genes across cancer types. PAAD is pancreatic adenocarcinoma.

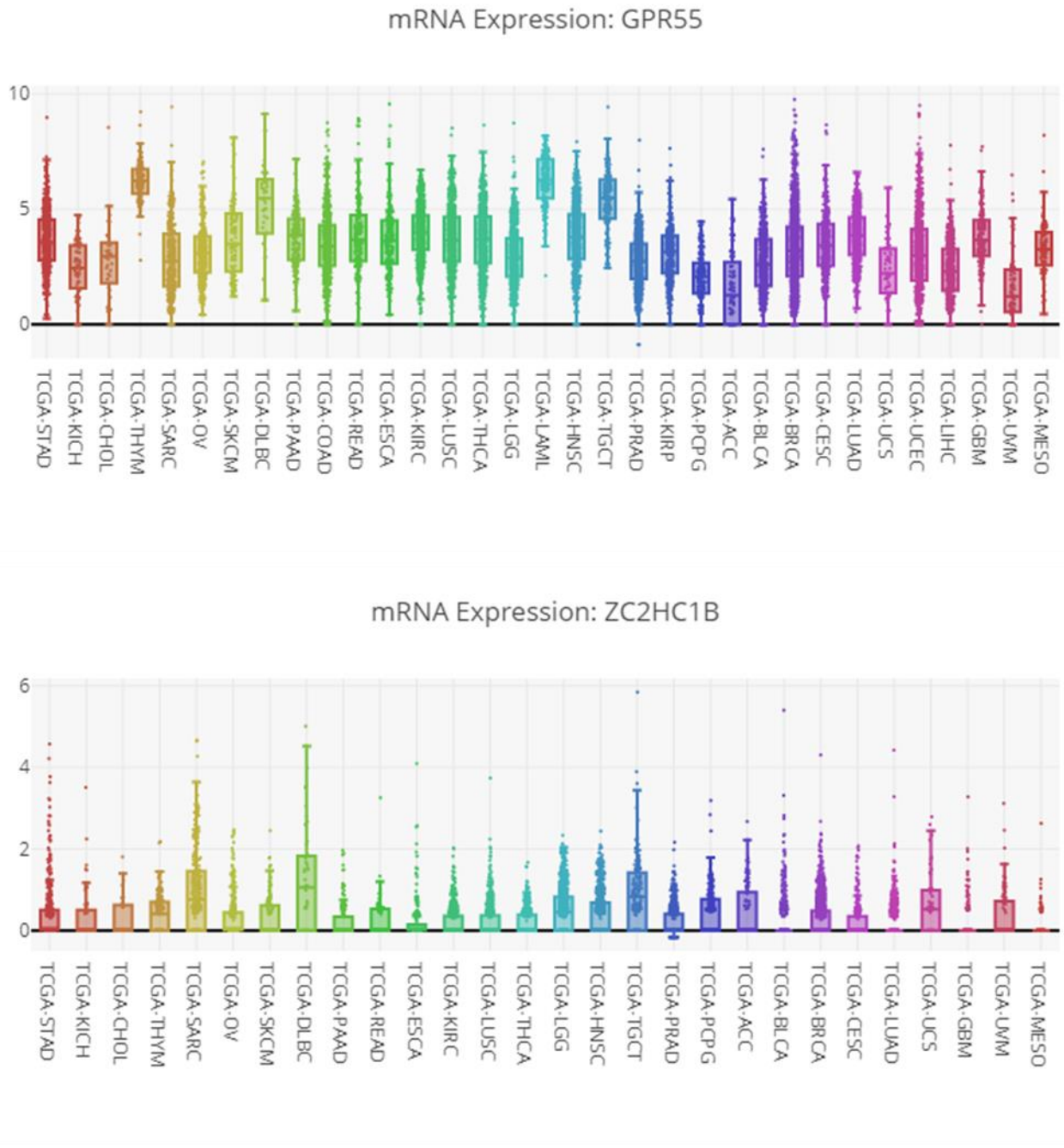


Figure 9, continued. mRNA expression of RFE selected genes across cancer types. PAAD is pancreatic adenocarcinoma.

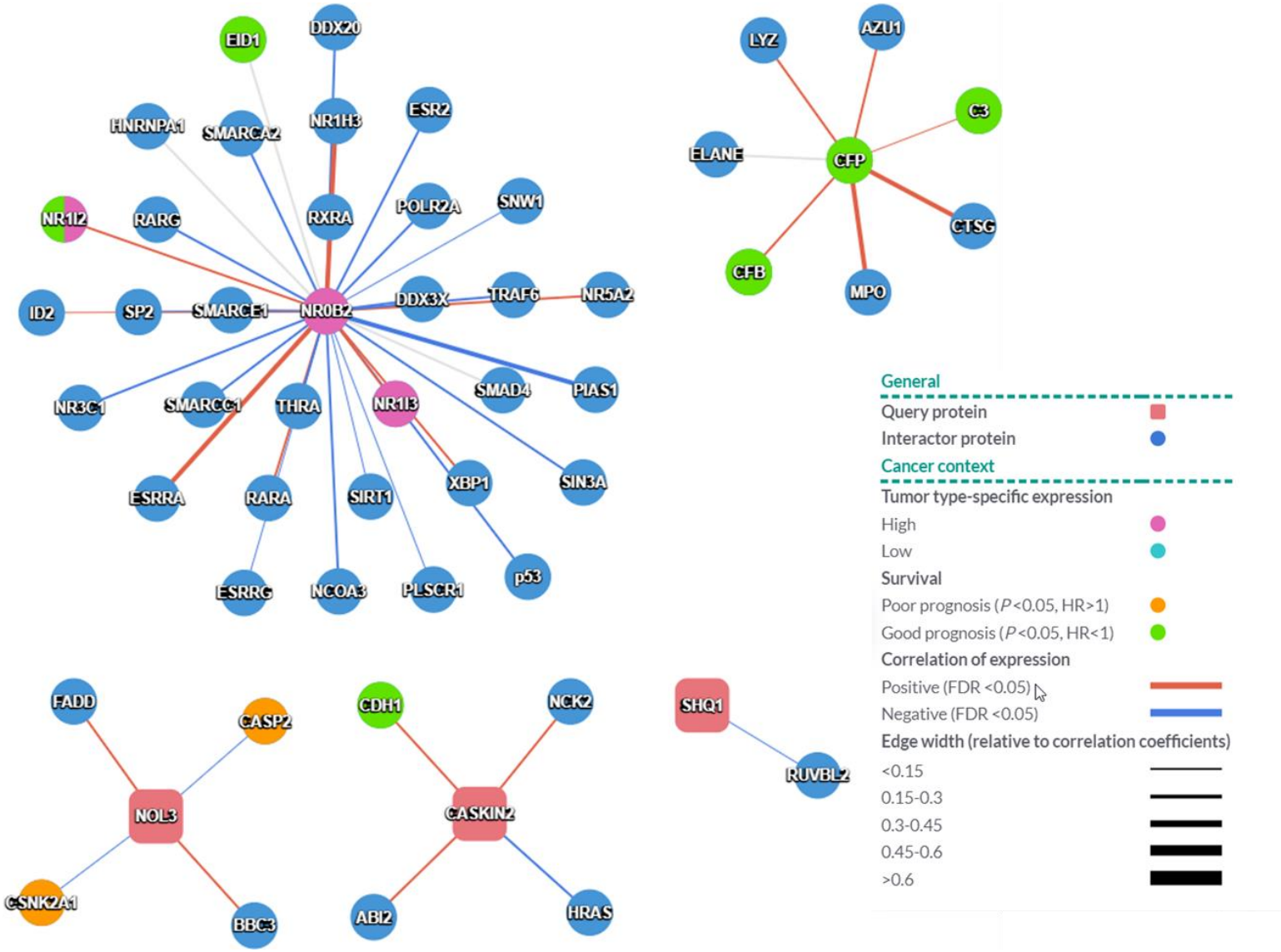


Figure 10. Protein-protein interaction (PPI) network of HCC top ten dysregulated genes.

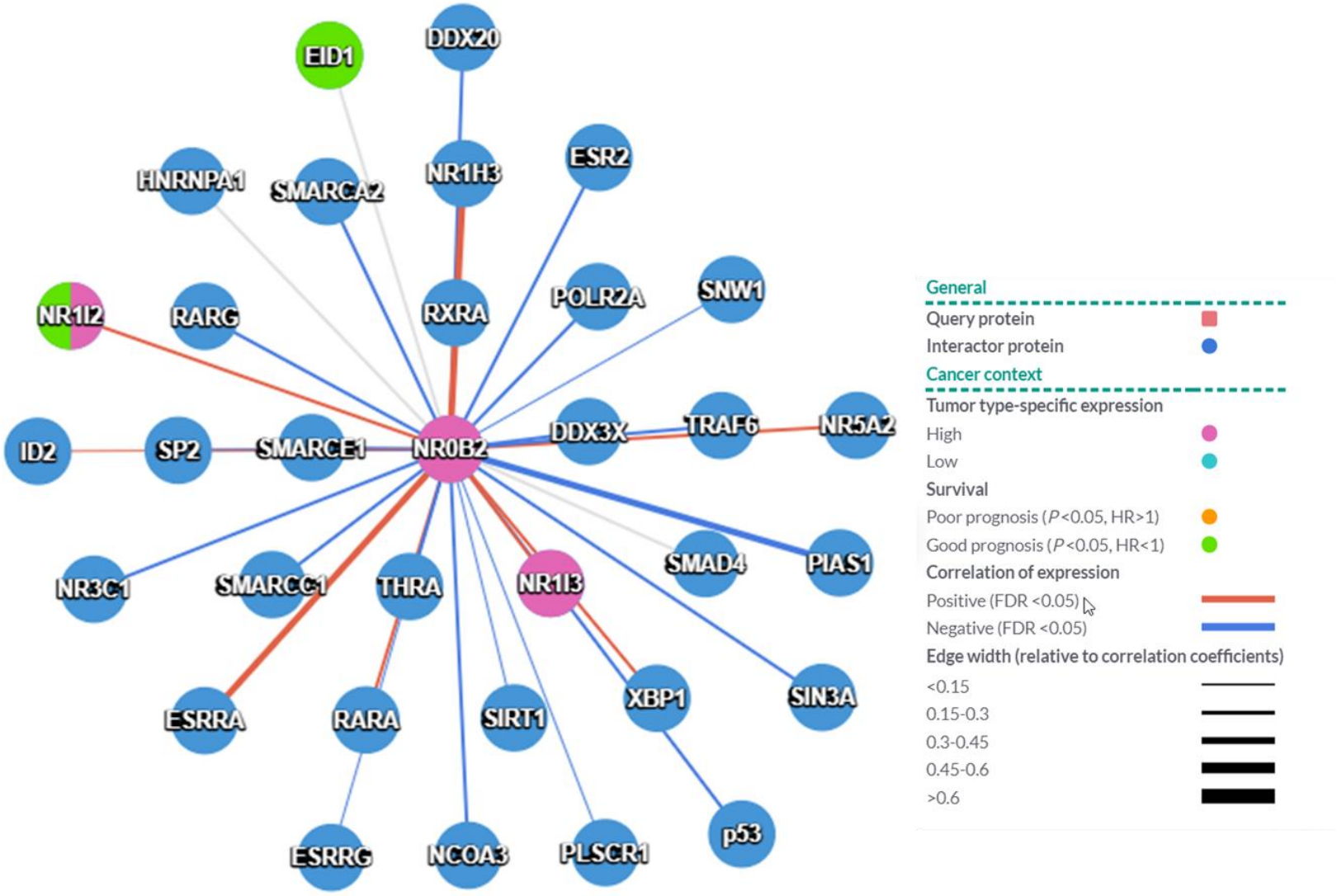


Figure 11. Upper left view of protein-protein interaction (PPI) network of HCC top ten dysregulated genes.

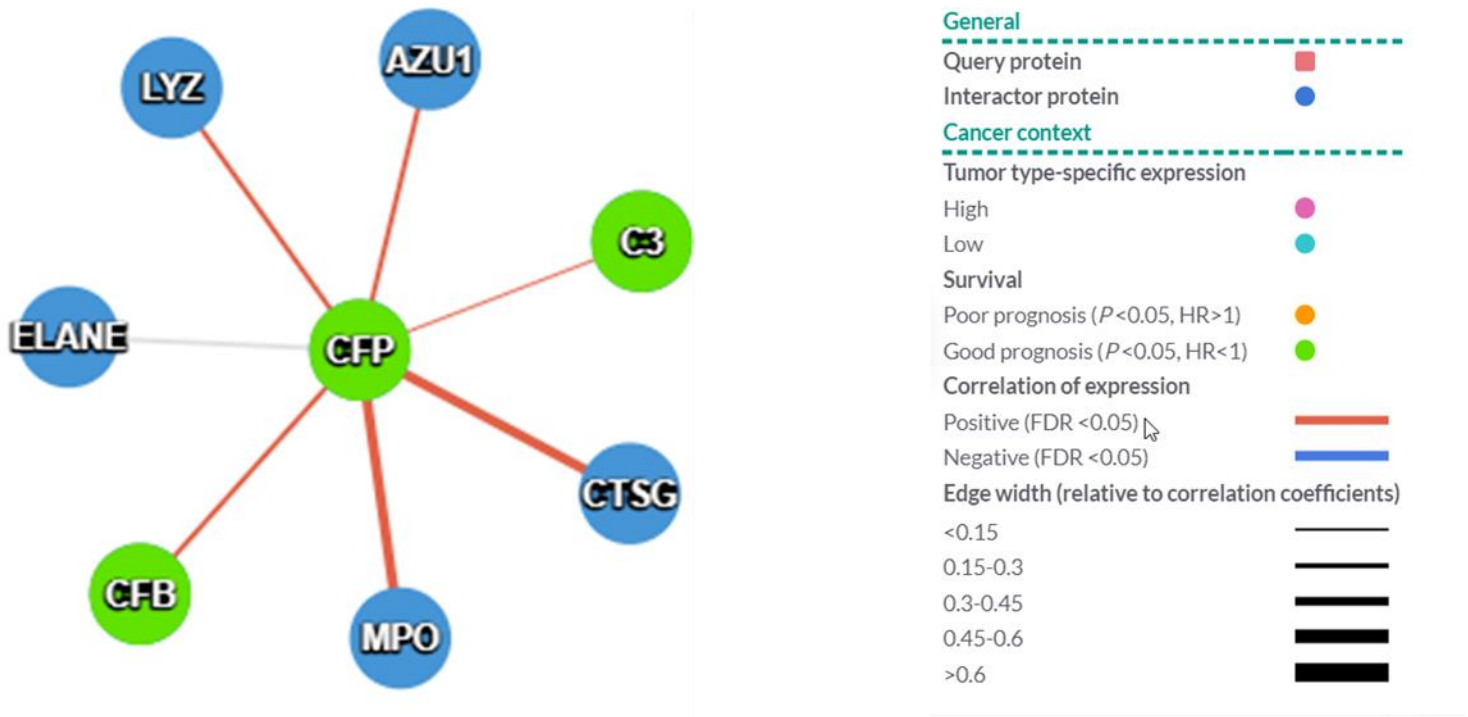


Figure 12. Upper right view of protein-protein interaction (PPI) network of HCC top ten dysregulated genes.

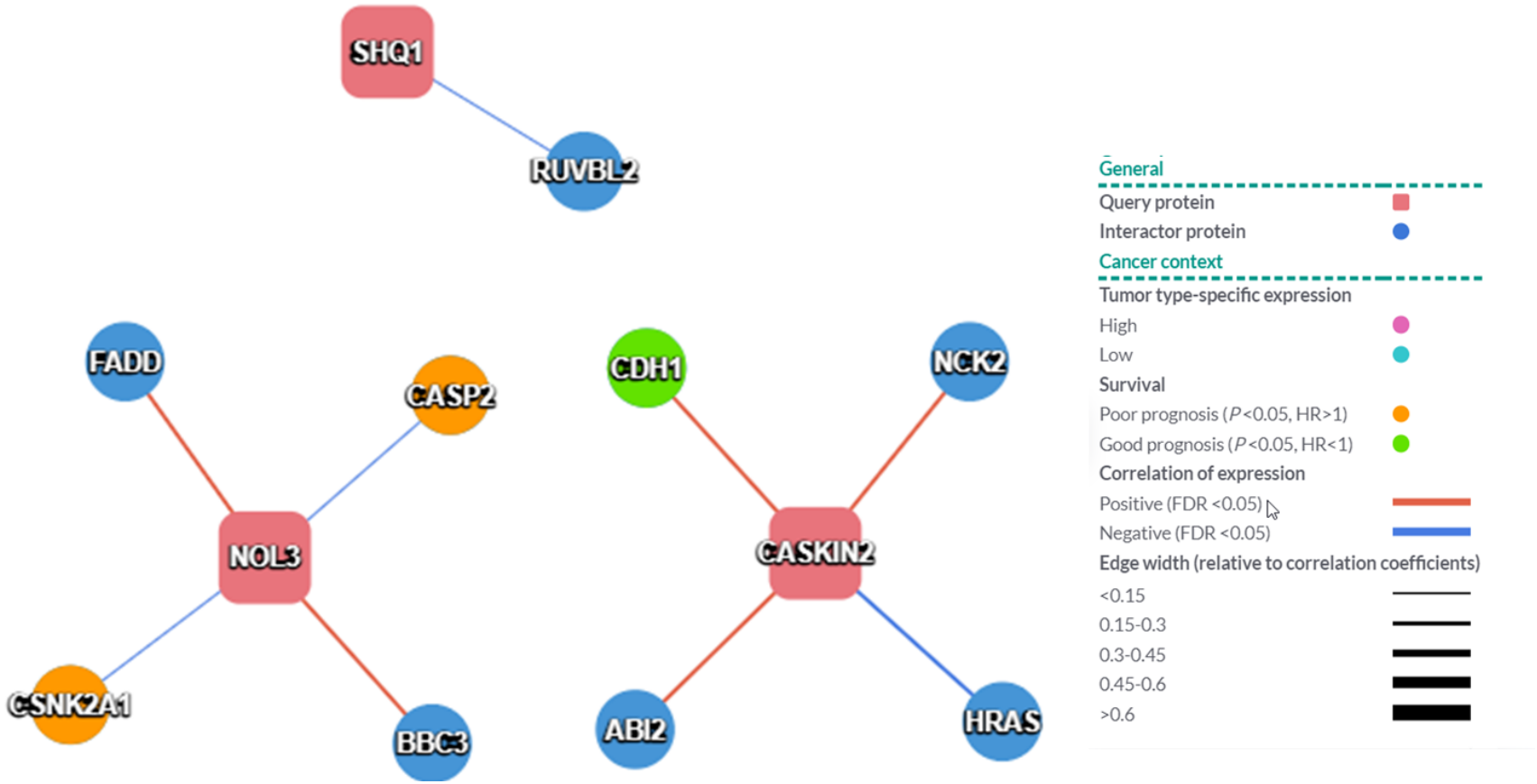


Figure 13. Lower view of protein-protein interaction (PPI) network of HCC top ten dysregulated genes.

Table 9. Interacting proteins from PPI with HCC-specific cancer context.

Query Gene: NROB2		**Regulation: Down**			
Interactor Protein	**Gene Name**	**Molecular Function**	**Expression Level**	**Survival Prognosis**	**Correlation of Expression**
NR1I3	Nuclear Receptor Subfamily 1 Group I Member 3	Key regulator of xenobiotic and endobiotic metabolism.	High	NA	Positive
NR1I2	Nuclear Receptor Subfamily 1 Group I Member 2	Transcriptional regulator of the cytochrome P450 gene CYP3A4 (enzyme, found in the liver and in the intestine, that oxidizes small foreign organic molecules).	High	Good	Positive
EID1	EP300 Interacting Inhibitor Of Differentiation 1	Repressor of MYOD1 transactivation. Inhibits EP300 and CBP histone acetyltransferase activity. May be involved in coupling cell cycle exit to the transcriptional activation of genes required for cellular differentiation.	NA	Good	NA

Table 9, continued. Interacting proteins from PPI with HCC-specific cancer context.

Query Gene: NOL3		**Regulation: Up**			
Interactor Protein	**Gene Name**	**Molecular Function**	**Expression Level**	**Survival Prognosis**	**Correlation of Expression**
CASP2	Caspase 2	Mediates cellular apoptosis through the proteolytic cleavage of specific protein substrates. May function in stress-induced cell death pathways, cell cycle maintenance, and the suppression of tumorigenesis.	NA	Poor	Negative
CSNK2A1	Casein Kinase 2 Alpha 1	Phosphorylates acidic proteins such as casein. It is involved in various cellular processes, including cell cycle control, apoptosis, and circadian rhythm.	NA	Poor	Negative

Query Gene: CASKIN2		**Regulation: Down**			
Interactor Protein	**Gene Name**	**Molecular Function**	**Expression Level**	**Survival Prognosis**	**Correlation of Expression**
CDH1	Cadherin 1	Calcium-dependent cell adhesion protein. Mutations correlated with gastric, breast, colorectal, thyroid and ovarian cancer. Loss of function thought to contribute to cancer progression by increasing proliferation, invasion,	NA	Good	Positive

Table 10. Possible HCC drug targets and diagnostic biomarkers derived from RFE selected genes and PPI interactions .

Symbol	Gene Name	Origin	Drug Target/Diagnostic Biomarker
NR1I3	Nuclear Receptor Subfamily 1 Group I Memb	PPI	Drug Target
CASP2	Caspase 2	PPI	Drug Target
CSNK2A1	Casein Kinase 2 Alpha 1	PPI	Drug Target
NOL3	nucleolar protein 3	RFE	Drug Target
SCLY	selenocysteine lyase	RFE	Drug Target
NPFF	neuropeptide FF-amide peptide precursor	RFE	Drug Target
SHQ1	SHQ1, H/ACA ribonucleoprotein assembly fa	RFE	Drug Target
POLR2J4	RNA polymerase II subunit J4, pseudogene	RFE	Drug Target
NR1I2	Nuclear Receptor Subfamily 1 Group I Memb	PPI	Diagnostic Biomarker
EID1	EP300 Interacting Inhibitor Of Differentiation	PPI	Diagnostic Biomarker
C3	Complement C3	PPI	Diagnostic Biomarker
CFB	Complement Factor B	PPI	Diagnostic Biomarker
NROB2	nuclear receptor subfamily 0 group B member	RFE	Diagnostic Biomarker
CFP	complement factor properdin	RFE	Diagnostic Biomarker
CASKIN2	CASK interacting protein 2	RFE	Diagnostic Biomarker
CDH1	Cadherin 1	RFE	Diagnostic Biomarker
CACNG4	calcium voltage-gated channel auxiliary subun	RFE	Diagnostic Biomarker
RCN3	reticulocalbin 3	RFE	Diagnostic Biomarker

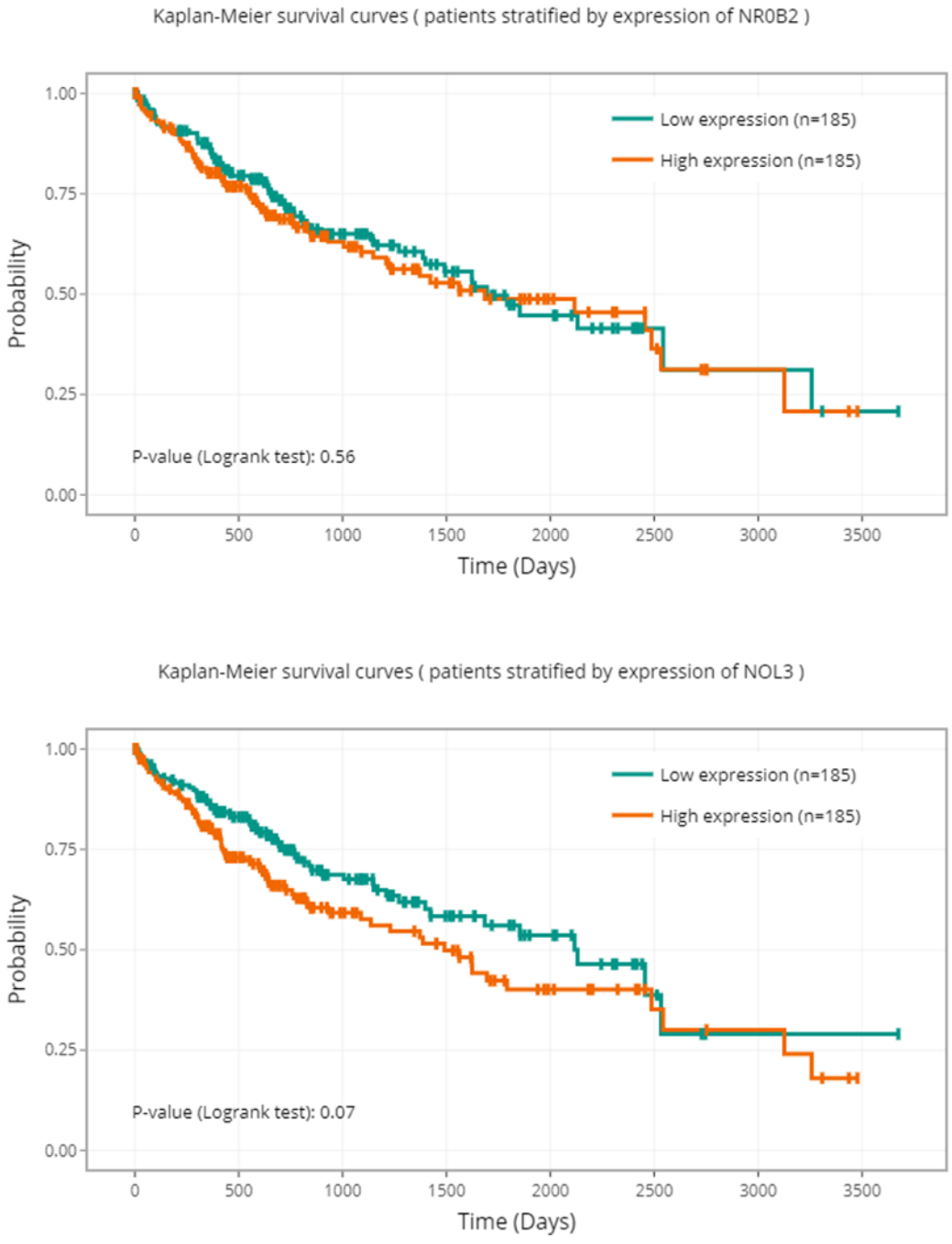


Figure 14. Cancer survivor curves for selected dysregulated genes in HCC dataset.

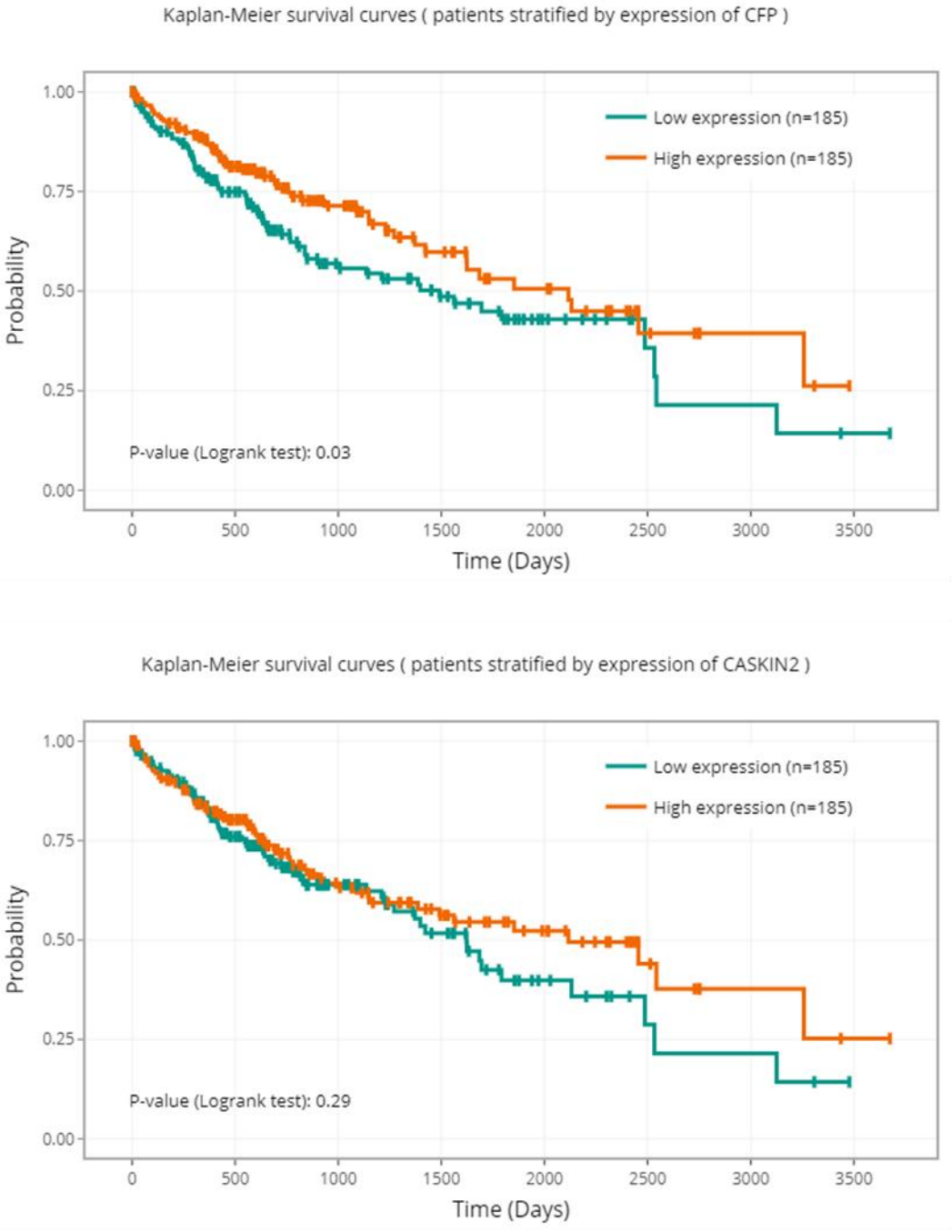


Figure 14, continued. Cancer survivor curves for selected dysregulated genes in HCC dataset

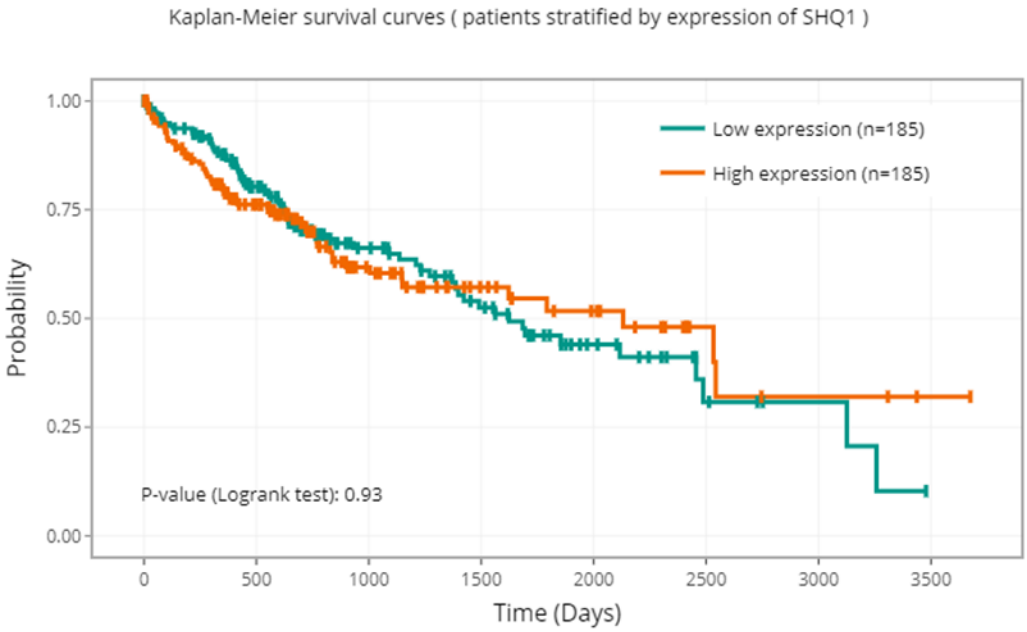


Figure 14, continued. Cancer survivor curves for selected dysregulated genes in HCC dataset

Figure 15 shows the mRNA expression of HCC SVM-RFE selected genes across cancer types. NROB2 has high expression in liver hepatocellular carcinoma (LIHC), adrenocortical carcinoma (ACC), esophageal carcinoma (ESCA), PDAC (PAAD), colon adenocarcinoma (COAD), rectum adenocarcinoma (READ), and lung adenocarcinoma (LUAD). CFP has high expression in all cancer types with the highest in thymoma (THYM), lymphoid neoplasm diffuse large B-cell lymphoma (DLBC), and acute myeloid leukemia (LAML). NOL3 has high expression in many cancers with the highest being kidney renal clear cell carcinoma (KIRC). CASKIN2 has high expression in all cancer types except for acute myeloid leukemia (LAML) and sarcoma (SARC). SHQ1 has high expression in all cancer types.

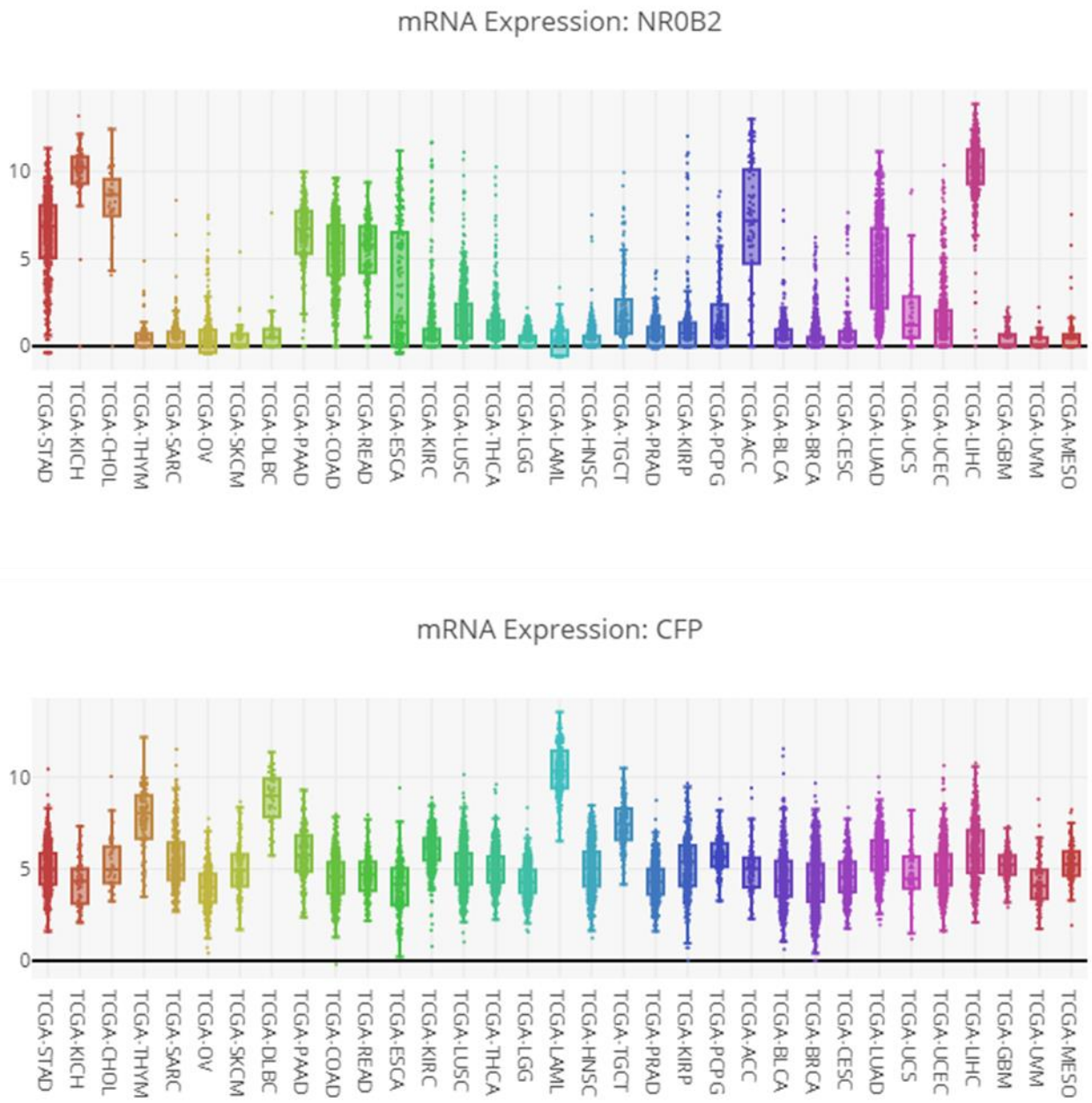


Figure 15. mRNA expression of RFE selected genes across cancer types. LIHC is liver hepatocellular carcinoma.

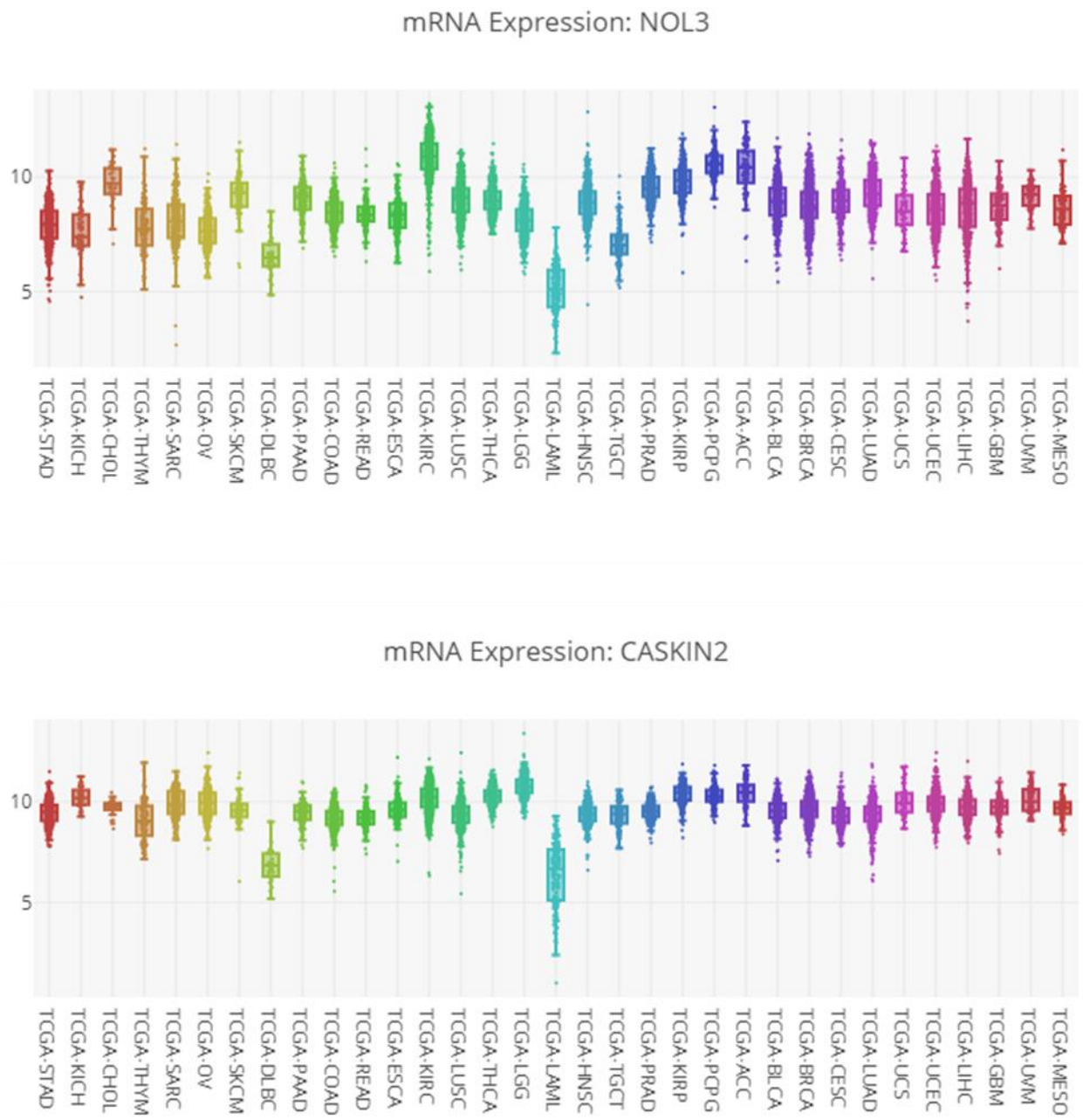


Figure 15, continued. mRNA expression of RFE selected genes across cancer types. LIHC is liver hepatocellular carcinoma.

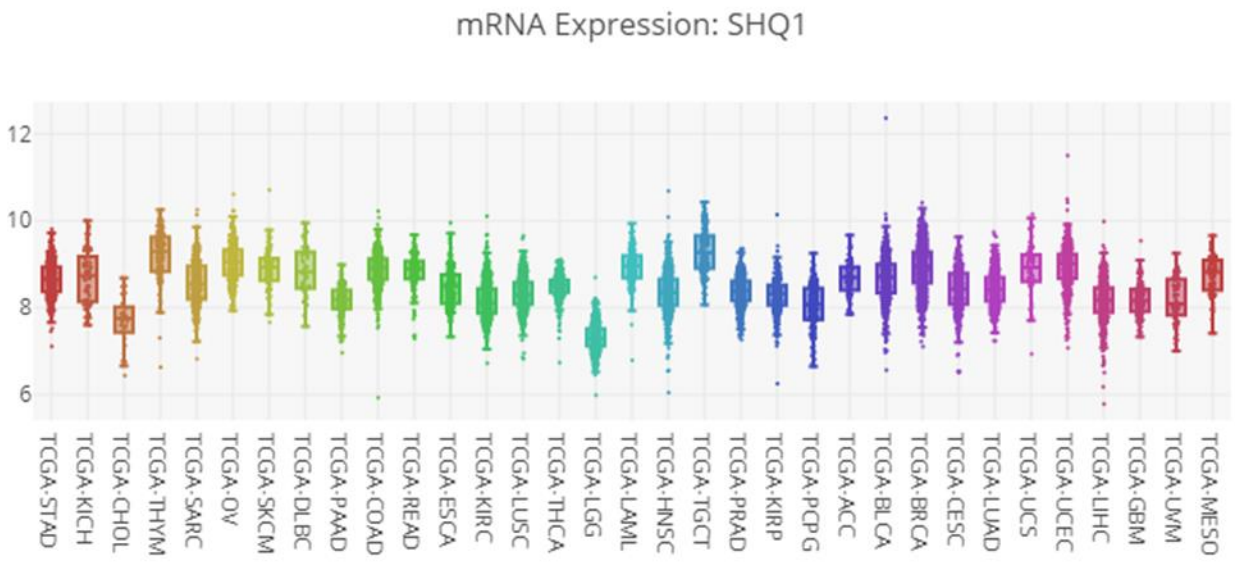


Figure 15, continued. mRNA expression of RFE selected genes across cancer types. LIHC is liver hepatocellular carcinoma.

CHAPTER IV - CONCLUSIONS

A. Summary

This study outlined a method that identified DEGs in PDAC and HCC cancer types. PDAC and HCC were selected based on number of deaths, poor prognosis, and lack of available treatments. The DEGs were analyzed by the SVM-RFE algorithm to select the top genes that differentiated between tumor and disease-free patient samples. PPI networks with cancer specific context were created using the SVM-RFE selected genes to identify additional genes. Upregulated genes from both SVM-RFE and PPI interactions are possible therapeutic targets and downregulated genes are possible diagnostic biomarker targets.

B. Conclusions

The top 10 downregulated and top 10 upregulated genes were identified for both PDAC and HCC datasets using their log fold change and p-values. The collagen triple helix repeat containing 1 (CTHRC1) gene was the fifth most upregulated gene in both cancer types and is upregulated in many cancer types. It is a potential drug target or co-drug target.

SVM-RFE was used to select the most important genes that differentiate between tumor and disease-free cells. The SVM-RFE algorithm had the best accuracy for both PDAC and HCC with 10 features or genes selected. There were no common genes between cancer types in the SVM-RFE analysis.

The PDAC dataset had seven SVM-RFE selected genes related to cancer. Cancer related genes STAB2, LMX1A-AS2, and UGT3A1 were downregulated and are therefore candidates for diagnostic biomarkers. GPR55 was downregulated in the PDAC dataset however; studies show that highly expressed GPR55 promotes tumor growth. It is possible that the model mischaracterized this gene.

The PDAC cancer genes PLK3, GCC2, and CCN4 were upregulated and are therefore potential drug targets. Survival curves show that low expression of PLK3, GCC2, and CCN4 have a higher probability of survival, which indicates that targeted therapies against them would be successful. There were also two uncharacterized gene loci affiliated with the non-coding RNA class. These should be investigated to confirm that they are actually non-coding. The remaining non-cancer related selected gene, ZC2HC1B, is a possible diagnostic biomarker as it was selected by the SVM-RFE algorithm as important.

The PDAC PPI network provided additional therapeutic targets and diagnostic biomarkers. There were 11 upregulated proteins (CHEK2, SIAH2, BCL2L1, RRM2, MAPK14, ITGB5, OTX1, MRE11, HNRNPL, RAB64, and LGALS3) with poor survival prognosis that are possible drug targets. There were also seven downregulated proteins (RNF141, PLK1, SNCA, RAD52, UGGT1, TUBA1A, and SPG7) with both good and poor survival that are possible diagnostic biomarkers.

The HCC dataset had seven SVM-RFE selected genes related to cancer. The cancer related genes CFP1, NROB2, CACNG4, and RCN3 were downregulated and are therefore potential diagnostic biomarkers. The downregulated gene CASKIN2 is not directly associated with cancer however; it is important in regulation of gene expression and is a possible diagnostic

biomarker. In addition, survival curves show that high expression of CFP1 and CASKIN2 have a higher probability of survival.

Upregulated HCC cancer related genes SCLY, NOL3, and SHQ1 are potential targets for therapeutics. NPFF is involved in morphine-induced analgesia and is elevated in patients receiving morphine for pain (Gibula-Tarlowska & Kotlinska, 2020). The upregulated gene POLR2J4 is identified as a pseudogene however many pseudogenes have biological functions (Cheetham, Faulkner, & Dinger, 2020). It is therefore a possible therapeutic target.

The HCC PPI network provided three additional drug target and four additional diagnostic biomarkers. CSNK2A1, NR1I3, and CASP2 are upregulated in HCC and are possible drug targets. NR1I2, EID1, C3, and CFB are downregulated in HCC and are therefore possible diagnostic biomarkers.

As common biomarkers were not identified between the PDAC and HCC datasets using SVM-RFE, figures with mRNA expression across cancer types for each selected gene were generated. These figures was used to identify additional cancer types to use the method on to find common genes.

The following upregulated SVM-RFE selected genes also have high expression in other cancer types. GCC2 is upregulated in PDAC, stomach adenocarcinoma, prostate adenocarcinoma, and esophageal carcinoma. PLK3 is upregulated in PDAC, thyroid carcinoma, bladder urothelial carcinoma, cervical squamous cell carcinoma, and endocervical adenocarcinoma. CCN4 is upregulated in PDAC, sarcoma, head and neck squamous cell carcinoma, breast invasive carcinoma, and mesothelioma. NOL3 is upregulated in many cancers

including HCC with the highest being kidney renal clear cell carcinoma. SHQ1 had high expression in HCC and across all cancer types.

C. Limitation of the Study

This study identified promising therapeutic targets and diagnostic biomarkers for both PDAC and HCC. However, the method had some limitations. First, although the accuracy of the SVM-RFE algorithms were good the sample size was relatively small. Secondly, the data was obtained from public databases so the quality of the data cannot be evaluated. Finally, metadata including gender, age, location, et cetera were not taken into account.

D. Next Steps

Future research should include testing the method in other cancer types to attempt to identify common biomarkers, using more samples per cancer type with a goal of 95% or better accuracy, testing different patient populations, and investigating the possible drug targets and diagnostic biomarkers identified.

LITERATURE CITED

Allain, E., Rouleau, M., Lévesque, E., & Guillemette, C. (2020). Emerging roles for UDP-glucuronosyltransferases in drug resistance and cancer progression. *British Journal of Cancer, 122*, 1277-1287. doi:https://doi.org/10.1038/s41416-019-0722-0

Badea, L., Herlea, V., Dima, S., Dumitrascu, T., & Popescu, I. (2008). Combined gene expression analysis of whole-tissue and microdissected pancreatic ductal adenocarcinoma identifies genes specifically overexpressed in tumor epithelia. *Hepatogastroenterology, 88*, 2016-2027.

Baghban, R., Roshangar, L., Jahanban-Esfahlan, R., & al., e. (2020). Tumor microenvironment complexity and therapeutic implications at a glance. *Cell Communication and Signalling, 18*(59), 59. doi:https://doi.org/10.1186/s12964-020-0530-4

Becker, M., Mastropasqua, F., Reising, J., Maier, S., Ho, M. L., Rabkina, L., . . . Tammimies. (2020). Presynaptic dysfunction in CASK-related neurodevelopmental disorders. *Translational Psychiatry, 10*, 312. doi:https://doi.org/10.1038/s41398-020-00994-0

Bhattacharyya, B., Suderman, M., Szyf, M., Huang, J., Han, Z., & Hallett, M. (2011). Definition of the landscape of promoter DNA hypomethylation in liver cancer. *Cancer Research, 17*, 5891-5903.

Carlson, M. (2021, 06 21). *Bioconductor*. Retrieved from hgu133plus2.db: Affymetrix Human Genome U133 Plus 2.0 Array annotation data (chip hgu133plus2).: https://bioconductor.org/packages/release/data/annotation/html/hgu133plus2.db.html

Cheetham, S., Faulkner, G., & Dinger, M. (2020). Overcoming challenges and dogmas to understand the functions of pseudogenes. *National Reviews Genetics, 21*, 191–201. doi:https://doi.org/10.1038/s41576-019-0196-1

Chen, Y., Sun, Y., Cui, Y., Lei, Y., Jiang, N., Jiang, W., . . . Ke, Z. (2019). High CTHRC1 expression may be closely associated with angiogenesis and indicates poor prognosis in lung adenocarcinoma patients. *Cancer Cell International, 19*, 318. doi:https://doi.org/10.1186/s12935-019-1041-5

Du, Y., Cai, M., Xing, X., Ji, J., Yang, E., & Wu, J. (2021). PINA 3.0: mining cancer interactome. *Nucleic Acids Research, 49*, D1351-D1357.

Edgar, R., Domrachev, M., & Lash, A. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research, 30*(1), 207-210. doi:https://doi.org/10.1093/nar/30.1.207

Ferro, R., Adamska, A., Lattanzio, R., Mavrommati, I., Edling, C. E., Arifin, S. A., . . . Falasca, M. (2018). GPR55 signalling promotes proliferation of pancreatic cancer cells and tumour growth in mice, and its inhibition increases effects of gemcitabine. *Oncogene, 37*(49), 6368–6382. doi:https://doi.org/10.1038/s41388-018-0390-1

Gautier, L., Cope, L., Bolstad, B., & Irizarry, R. (2004). affy---analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics, 20*(3), 307-315.

Genentech, A Member of the Roche Group. (2021, 02 26). *TECENTRIQ Proposed Mechanism of Action*. Retrieved from Tecentriq (atezolizumab): https://www.tecentriq-hcp.com/urothelial-carcinoma/tecentriq-moa.html

Gibula-Tarlowska, E., & Kotlinska, J. H. (2020). Crosstalk between Opioid and Anti-Opioid Systems: An Overview and Its Possible Therapeutic Significance. *Biomolecules, 10*, 1376. doi:doi:10.3390/biom10101376

Gurbuz, I., & Chiquest-Ehrismann, R. (2015). CCN4/WISP1 (WNT1 inducible signaling pathway protein 1): A focus on its role in cancer. *The International Journal of Biochemistry & Cell Biology, 62*, 142-146. doi:https://doi.org/10.1016/j.biocel.2015.03.007

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning, 46*, 389-422. doi:https://doi.org/10.1023/A:1012487302797

Helmke, C., Becker, S., & Strebhardt, K. (2016). The role of Plk3 in oncogenesis. *Oncogene, 35*, 135–147. doi:https://doi.org/10.1038/onc.2015.105

Hirose, Y., Saijou, E., Sugano, Y., Takeshita, F., Nishimura, S., Nonaka, H., . . . Miyajima, A. (2012). Inhibition of Stabilin-2 elevates circulating hyaluronic acid levels and prevents tumor metastasis. *Proceedings of the National Academy of Sciences of the United States of America, 109*(11), 4263–4268. doi:https://doi.org/10.1073/pnas.1117560109

Hou, Y., Li, Y., Gong, F., Jin, J., Huang, A., Fang, Q., & Ma, R. Z. (2016). A Preliminary Study on RCN3 Protein Expression in Non-small Cell Lung Cancer. *Clinical Laboratory, 62*(3), 293-300. doi:doi: 10.7754/clin.lab.2015.150411

Huang, S., Cai, N., Pacheco, P. P., Narrandes, S., Wang, Y., & Xu, W. (2018). Applications of Support Vector Machine (SVM) Learning in Cancer Genomics. *Cancer genomics & proteomics, 15*(1), 41–51. doi:https://doi.org/10.21873/cgp.20063

Jia, Y., Dai, J., & Zeng, Z. (2020). Potential relationship between the selenoproteome and cancer. *Molecular and Clinical Oncology, 13*(6), 83. doi:https://dx.doi.org/10.3892%2Fmco.2020.2153

Kanwar, N., Carmine-Simmen, K., Nair, R., Wang, C., Moghadas-Jafari, S., Blaser, H., . . . Done, S. J. (2020). Amplification of a calcium channel subunit CACNG4 increases breast cancer metastasis. *EBioMedicine, 52*, 102646. doi:https://dx.doi.org/10.1016%2Fj.ebiom.2020.102646

Kleeff, J., Korc, M., Apte, M., & al., e. (2016). Pancreatic cancer. *Nature Reviews Disease Primers, 2*, 16022. doi:https://doi.org/10.1038/nrdp.2016.22

Lee, Y. S. (2021, 06 27). *Series GSE101685 (unpublished)*. Retrieved from Gene Expression Omnibus (GEO): https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi

Lu, C., Chen, H., Shan, Z., & Yang, L. (2016). Identification of differentially expressed genes between lung adenocarcinoma and lung squamous cell carcinoma by gene expression profiling. *Molecular Medicine Reports, 14*(2), 1483-1490. doi:https://dx.doi.org/10.3892%2Fmmr.2016.5420

Mangogna, A., Varghese, P. M., Agostinis, C., Alrokayan, S. H., Khan, H. A., Stover, C. M., . . . Kishore, U. (2020). Prognostic Value of Complement Properdin in Cancer. *Frontiers in Immunology, 11*, 614980. doi:https://doi.org/10.3389/fimmu.2020.614980

National Center of Biotechnology Information. (2020, 02 27). *PubMed*. Retrieved from National Library of Medicine, National Center of Biotechnology Information: https://pubmed.ncbi.nlm.nih.gov/

NCBI. (2021, 06 30). Retrieved from POLR2J4 RNA polymerase II subunit J4, pseudogene [Homo sapiens (human)]: https://www.ncbi.nlm.nih.gov/gene/84820

Nivison, M., & Meier, K. (2018). The role of CCN4/WISP-1 in the cancerous phenotype. *Cancer Management and Research, 10*, 2893–2903. doi:https://dx.doi.org/10.2147%2FCMAR.S133915

Pagès, H., Carlson, M., Falcon, S., & L, N. (2021, 06 13). *AnnotationDbi: Manipulation of SQLite-based annotations in Bioconductor*. Retrieved from Bioconductor: https://bioconductor.org/packages/AnnotationDbi

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Duche, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12*(85), 2825-2830.

Phipson, B., Lee, S., M. I., Alexander, W. S., & Smyth, G. K. (2016). ROBUST HYPERPARAMETER ESTIMATION PROTECTS AGAINST HYPERVARIABLE GENES AND IMPROVES POWER TO DETECT DIFFERENTIAL EXPRESSION. *The annals of applied statistics, 10*(2), 946-963. doi: doi:10.1214/16-AOAS920

Pineiro, R., Maffucci, T., & Falasca, M. (2011). The putative cannabinoid receptor GPR55 defines a novel autocrine loop in cancer cell proliferation. *Oncogene, 30*, 142-152. doi:https://doi.org/10.1038/onc.2010.417

Privat-Maldonado, A., Bengtson, C., Razzokov, J., Smits, E., & Bogaerts, A. (2019). Modifying the Tumour Microenvironment: Challenges and Future Perspectives for Anticancer Plasma Treatments. *Cancers (Basel), 11*(12), 1920. doi:https://dx.doi.org/10.3390%2Fcancers11121920

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research, 43*(7), e47. doi:https://doi.org/10.1093/nar/gkv007

Roumy, M., Lorenzo, C., Mazères, S., Bouchet, S., Zajac, J. M., & Mollereau, C. (2007). Physical association between neuropeptide FF and micro-opioid receptors as a possible molecular basis for anti-opioid activity. *Journal of Biological Chemistry, 282*(11), 8332-8342. doi:https://doi.org/10.1074/jbc.M606946200

Sangro, B., Sarobe, P., Hervas-Stubbs, S., & Melero, I. (2021). Advances in immunotherapy for hepatocellular carcinoma. *Nature Reviews Gastroenterology Hepatology*, 1-19. doi:https://doi.org/10.1038/s41575-021-00438-0

Sarantis, P., Koustas, E., Papadimitropoulou, A., Papavassiliou, A., & Karamouzis, M. (2020). Pancreatic ductal adenocarcinoma: Treatment hurdles, tumor microenvironment and immunotherapy. *World Journal of Gastrointestinal Oncology, 12*(2), 173-181. doi:https://dx.doi.org/10.4251%2Fwjgo.v12.i2.173

Siegel, R., Miller, K., Fuchs, H., & Jemal, A. (2021). Cancer Statistics, 2021. *CA: A Cancer Journal for Clinicians, 71*(1), 7-33. doi:https://doi.org/10.3322/caac.21654

Singh, A. K., Kumar, R., & Pandey, A. K. (2018). Hepatocellular Carcinoma: Causes, Mechanism of Progression and Biomarkers. *CurrentCchemical Genomics and Translational Medicine, 12*, 9-26. doi:https://dx.doi.org/10.2174%2F2213988501812010009

Stelzer G, R. R., Fishilevich, S., Iny, S., Nudel, R., Lieder, I., Mazor, Y., . . . Lancet, D. (2016). The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analysis. *Current Protocols in Bioinformatics, 54*, 1.30.1–1.30.33. doi:https://doi.org/10.1002/cpbi.5

Su, H., Hu, J., Huang, L., Yang, Y., Thenoz, M., Kuchimy, A., . . . Liu, H. (2018). SHQ1 regulation of RNA splicing is required for T-lymphoblastic leukemia cell survival. *Nature Communications, 9*, 4281. doi:https://doi.org/10.1038/s41467-018-06523-4

U.S. Food and Drug Administration. (2021, 06 26). *FDA grants accelerated approval to pembrolizumab for first tissue/site agnostic indication*. Retrieved from U.S. Food and Drug Administration: https://www.fda.gov/drugs/resources-information-approved-drugs/fda-grants-accelerated-approval-pembrolizumab-first-tissuesite-agnostic-indication

Wang, H., Huo, X., Yang, X. R., He, J., Cheng, L., Wang, N., . . . Qin, W. (2017). TAT3-mediated upregulation of lncRNA HOXD-AS1 as a ceRNA facilitates liver cancer metastasis by regulating SOX4. *Molecular Cancer, 16*(1), 136. doi:10.1186/s12943-017-0680-1

Wang, W., Qanungo, S., Crow, M. T., Watanabe, M., & Nieminen, A. L. (2005). Apoptosis repressor with caspase recruitment domain (ARC) is expressed in cancer cells and localizes to nuclei. *FEBS Letters, 579*(11), 2411-2415. doi:https://doi.org/10.1016/j.febslet.2005.03.040

Wu, J., & Irizarry, R. (2021, 06 13). *gcrma: Background Adjustment Using Sequence Information*. Retrieved from Bioconductor: https://www.bioconductor.org/packages/release/bioc/html/gcrma.html

Wu, T. H., Chang, S. Y., Shih, Y. L., Chian, C. F., Chang, H., & Lin, Y. W. (2020). Epigenetic Silencing of LMX1A Contributes to Cancer Progression in Lung Cancer Cells. *International journal of molecular sciences, 21*(15), 5425. doi:https://doi.org/10.3390/ijms21155425

Yan, W., Liu, X., Wang, Y., Han, S., Wang, F., Liu, X., . . . Hu, G. (2020). Identifying Drug Targets in Pancreatic Ductal Adenocarcinoma Through Machine Learning, Analyzing Biomolecular Networks, and Structural Modeling. *frontiers in Pharmacology, 11*, 534. doi:https://dx.doi.org/10.3389%2Ffphar.2020.00534

Zhu, R., Tu, Y., Chang, J., Xu, H., Li, J. C., Liu, W., . . . Li, B. (2021). The Orphan Nuclear Receptor Gene NR0B2 Is a Favorite Prognosis Factor Modulated by Multiple Cellular Signal Pathways in Human Liver Cancers. *Frontiers in oncology, 11*, 691199. doi:https://doi.org/10.3389/fonc.2021.691199

APPENDIX

A. Code

Code is in order of execution with programming language used.

R Program:

1.

```
# Pancreas_limma.R
# read .cel files, perform RMA, fit linear model, perform eBayes, annotate with gene
information.

library(affy)
library(limma)
library(hgu133plus2.db)
library(AnnotationDbi)
library(dplyr)

# read .CEL files from working directory
Data <- ReadAffy()
# boxplot of raw data
boxplot(Data)

# perform RMA. Background correction, log2 transformation, quantile normalization
DataRMA <- rma(Data)
# boxplot of after RMA algorithm
boxplot(DataRMA)
head(DataRMA)
write.table(DataRMA,"RMA_pancreas.txt",sep="\t", col.names = NA, row.names =
TRUE)

# create design matrix to compare between tumor and normal. Update for # of samples of
each type.
group=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,1,1,1,
1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1)
design <- model.matrix(~-1 + factor(group))
design
colnames(design) <- c("tumor","normal")
contrast <- makeContrasts(tumor - normal, levels = design)

# fits linear model using expression data and design matrix
fit <- lmFit(DataRMA, design)
fit2 <- contrasts.fit(fit, contrast)
```

```
# calculates logFc, t-statistics
fit2 <- eBayes(fit2)

# results table
res <- topTable(fit2, number=Inf, adjust.method="none")

# save differential expression data
write.table(res,"dif_exp.txt",sep="\t", col.names = NA, row.names = TRUE)

# add in PROBEID to column A header for merge
diff <- read.delim("dif_exp.txt", header=TRUE)
names(diff)[1] <- "PROBEID"
head(diff)

vector_diff = diff[['PROBEID']]
head(vector_diff)

geneID <- AnnotationDbi::select(hgu133plus2.db,
                                keys = (vector_diff),
                                columns = c("SYMBOL", "GENENAME"),
                                keytype = "PROBEID")

write.csv(geneID, file = "gene_ID.csv", row.names = FALSE)

# join expression data and gene annotation
joined <- left_join(diff, geneID,
                    by = c("PROBEID"))

write.csv(joined, file = "pancreas_final.csv", row.names = FALSE)
```

Python:

2.

```
# drop_duplicate_genes.py
# removes rows with NA, sorts by SYMBOL and then P.Value, drops duplicate gene
with lowest p-value
# data from Pancreas_limma.py output

import pandas as pd

df = pd.read_csv("pancreas_final.csv")

# removes rows with NA
df = df.dropna()
```

```
# Sorts by SYMBOL, then P.Value.
df = df.sort_values(["SYMBOL", "P.Value"], ascending = (True, True))

# Removes SYMBOL duplicates. Keeps gene with lowest p-value
df.drop_duplicates(subset=['SYMBOL'], keep='first', inplace = True)

# save to csv
df.to_csv(pancreas_final_drop.csv', index = False)
```

3.

```
# volcano_plot.py
# creates volcano plot from RMA_pancreas.csv (generated using
drop_duplicate_genes.py)

from bioinfokit import analys, visuz
import pandas as pd

df = pd.read_csv('pancrease_final_drop.csv')
df.dropna()

visuz.gene_exp.volcano(df=df, lfc='logFC', pv='P.Value', geneid="SYMBOL",
plotlegend=True, legendpos='upper right',
   legendanchor=(1.46,1))
```

R Program:

4.

```
# mas5.R
# read .cel files, perform mas5 for RFE.

library(affy)
library(limma)
library(dplyr)
library(gcrma)

# read .CEL files
Data <- ReadAffy()

eset <- mas5(Data)

write.csv(eset, file = "mas5_pancreas.csv", row.names = TRUE)
```

Python:

5.

```
# transpose.py
# data from mas5.R
# transposes large dataframes

import pandas as pd

df = pd.read_csv("mas5_pancreas.csv")

# transpose
df = df.T

df.to_csv(' mas5_pancreas_transposed.csv', index = True, header = False)
```

6.

```
# merge_mas5_geneID.py
# data from transpose.py
# Data cleaning, join mas5 and gene ID data

import pandas as pd

# mas5 data from mas5.R, transpose.py
mas5 = pd.read_csv("RMA_pancreas_transposed.csv")

# remove leading X from PROBEID
mas5['PROBEID'] = mas5['PROBEID'].str.replace('^X','')

# gene ID data from Pancreas_limma.R
gene = pd.read_csv("gene_ID.csv")

# inner join. removes mas5 data not found in gene_ID
merged = pd.merge(mas5, gene)

merged.to_csv(mas5 _gene_pancreas.csv', index = False)
```

7.

```
# merge_ebayes_mas5_annotation.py
# data from merge_mas5_geneID.py (mas5_gene_pancreas.csv) and
drop_duplicate.gene.py (pancreas_final_drop.csv)
# Data cleaning, join mas5 and gene ID data

import pandas as pd

# data from merge_ mas5_geneID.py
```

```
mas5= pd.read_csv("rma_gene_pancreas.csv")

# remove leading X from PROBEID to match gene df
mas5 ['PROBEID'] = mas5 ['PROBEID'].str.replace('^X','')

# data from drop_duplicate.gene.py
gene = pd.read_csv("pancreas_final_drop.csv")

# inner join. removes mas5 data not found in gene
merged = pd.merge(mas5, gene)

merged.to_csv('final_ mas5_drop_pancreas.csv', index = False)
```

8.

final_ mas5_drop_pancreas.csv is transposed using transpose.py

9.

```
# remove_rows.py
# makes SYMBOL row header row
# removes non-mas5 data rows

import pandas as pd

df = pd.read_csv("final_mas5_drop_pancreas_transposed.csv")

# make SYMBOL row the header
header_row = 78
df.columns = df.iloc[header_row]

# drop rows without mas5 data
df.drop(df.tail(8).index,
       inplace = True)

# convert all columns except PROBEID to numeric
cols=[i for i in df.columns if i not in ["SYMBOL"]]
for col in cols:
   df[col]=pd.to_numeric(df[col])
print(df.dtypes)

# change rows names from sample IDs to numeric
df.index = ['1', '2', '3', '4', '5', '6', '7', '8', '9', '10', '11', '12', '13', '14', '15', '16', '17', '18', '19',
'20', ]
df.drop(df.columns[[0]], axis = 1, inplace = True)

df.to_csv('final_mas5_drop_pancreas_transposed_final.csv', index = False, header =
True)
```

10.
```
# RFE_number_features.py
# evaluate RFE for classification
import pandas as pd

from numpy import mean
from numpy import std
from sklearn.datasets import make_classification
from sklearn.model_selection import cross_val_score
from sklearn.model_selection import RepeatedStratifiedKFold
from sklearn.feature_selection import RFE
from sklearn.tree import DecisionTreeClassifier
from sklearn.pipeline import Pipeline

df = pd.read_csv("final_mas5_drop_pancreas_transposed_final.csv")

# define dataset
X = df.iloc[:,:-1].values
y = df.iloc[:,-1].values

# create pipeline
rfe = RFE(estimator=DecisionTreeClassifier(), n_features_to_select=20)
model = DecisionTreeClassifier()
pipeline = Pipeline(steps=[('s',rfe),('m',model)])

# evaluate model
cv = RepeatedStratifiedKFold(n_splits=10, n_repeats=3, random_state=1)
n_scores = cross_val_score(pipeline, X, y, scoring='accuracy', cv=cv, n_jobs=-1,
error_score='raise')

# report performance
print('Accuracy: %.3f (%.3f)' % (mean(n_scores), std(n_scores)))
```

11.
```
# rfe.py
# performs recursive feature elimination (RFE)
import pandas as pd
import sys
from sklearn.feature_selection import RFE
from sklearn.tree import DecisionTreeClassifier

df = pd.read_csv("final_mas5_drop_pancreas_transposed_final.csv")
# create new df with gene names
names = df.iloc[[0],:]
names.to_csv('names_liver.csv', index = False, header = True)

# separate features (X) and labels(y)
```

```
X = df.iloc[:,:-1].values
y = df.iloc[:,-1].values

# define RFE
rfe = RFE(estimator=DecisionTreeClassifier(), n_features_to_select=10)

# fit RFE
rfe.fit(X, y)

# save print to console as file
sys.stdout = open("RFE_pancreas_final.csv", "w")

# summarize all features
for i in range(X.shape[1]):
        print('Column: %d, Selected %s, Rank: %.3f' % (i, rfe.support_[i],
rfe.ranking_[i]))

sys.stdout.close()
```