ORIGINAL RESEARCH
published: 10 September 2020
doi: 10.3389/fgene.2020.572284

A Transcriptomics-Based Meta-Analysis Combined With Machine Learning Identifies a Secretory Biomarker Panel for Diagnosis of Pancreatic Adenocarcinoma

Indu Khatri[1,2] *and Manoj K. Bhasin*[1,3]*

[1] *Division of IMBIO, Department of Medicine, Beth Israel Lahey Health, Harvard Medical School, Boston, MA, United States,* [2] *Department of Immunology and Leiden Computational Biology Center, Leiden University Medical Center, Leiden, Netherlands,* [3] *Department of Pediatrics and Biomedical Informatics, Children's Healthcare of Atlanta, Emory School of Medicine, Atlanta, GA, United States*

OPEN ACCESS

Edited by:
Amit Kumar Yadav,
Translational Health Science and Technology Institute (THSTI), India

Reviewed by:
Deepak Sharma,
Indian Institute of Technology Roorkee, India
Oksana Sorokina,
The University of Edinburgh, United Kingdom

****Correspondence:***
Manoj K. Bhasin
manoj.bhasin@emory.edu

Specialty section:
This article was submitted to Systems Biology, a section of the journal Frontiers in Genetics

Received: *13 June 2020*
Accepted: *12 August 2020*
Published: *10 September 2020*

Citation:
Khatri I and Bhasin MK (2020) A Transcriptomics-Based Meta-Analysis Combined With Machine Learning Identifies a Secretory Biomarker Panel for Diagnosis of Pancreatic Adenocarcinoma.
Front. Genet. 11:572284.
doi: 10.3389/fgene.2020.572284

Pancreatic ductal adenocarcinoma (PDAC) is generally incurable due to the late diagnosis and absence of markers that are concordant with expression in several sample sources (i.e., tissue, blood, plasma) and platforms (i.e., Microarray, sequencing). We optimized meta-analysis of 19 PDAC (tissue and blood) transcriptome studies from multiple platforms. The key biomarkers for PDAC diagnosis with secretory potential were identified and validated in different cohorts. Machine learning approach i.e., support vector machine supported by leave-one-out cross-validation was used to build and test the classifier. We identified a 9-gene panel (IFI27, ITGB5, CTSD, EFNA4, GGH, PLBD1, HTATIP2, IL1R2, CTSA) that achieved ∼0.92 average sensitivity and ∼0.90 average specificity in distinguishing PDAC from healthy samples in five training sets using cross-validation. These markers were also validated in proteomics and single-cell transcriptomics studies suggesting their prognostic role in the diagnosis of PDAC. Our 9-gene classifier can not only clearly discriminate between better and poor survivors but can also precisely discriminate PDAC from chronic pancreatitis (AUC = 0.95), early stages of progression [Stage I and II (AUC = 0.82), IPMA and IPMN (AUC = 1), and IPMC (AUC = 0.81)]. The 9-gene marker outperformed the previously known markers in blood studies particularly (AUC = 0.84). The discrimination of PDAC from early precursor lesions in non-malignant tissue (AUC > 0.81) and peripheral blood (AUC > 0.80) may assist in an early diagnosis of PDAC in blood samples and thus will also facilitate risk stratification upon validation in clinical trials.

Keywords: biomarker, pancreatic cancer, secretory, transcriptome, validation

Abbreviations: AUC, area under the curve; CA 19-9, carbohydrate antigen 19-9; CDF, chip definition file; CP, chronic pancreatitis; DE, differentially expressed; GEO, gene expression omnibus; GGH, γ-glutamyl hydrolase; FDR, false discovery rate; HPA, human protein atlas; IPMA, intraductal papillary-mucinous adenoma; IPMC, intraductal papillary-mucinous carcinoma; IPMN, intraductal papillary mucinous neoplasm; LOOCV, leave-one-out cross-validation; noTM, no transmembrane segments; PanIN, pancreatic intraepithelial neoplasia; PC, pancreatic cancer; PDAC, pancreatic ductal adenocarcinoma; ROC, receiver operating characteristic; SVM, support vector machines; TCGA, tissue cancer genome atlas.

INTRODUCTION

Pancreatic ductal adenocarcinoma (PDAC) is the most common type of pancreatic cancer (PC), which is one of the fatal cancers in the world with 5-year survival rate of <5% due to the lack of early diagnosis (Fesinmeyer et al., 2005). One of the challenges associated with an early diagnosis is distinguishing PDAC from other non-malignant benign gastrointestinal diseases such as chronic pancreatitis (CP) due to the histopathological and imaging limitations (Brand and Matamoros, 1998). Although imaging techniques such as endoscopic ultrasound and FDG-PET have improved the sensitivity of PDAC detection but have failed to distinguish PC from focal mass-forming pancreatitis in >50% cases. Dismal prognosis of PC yields from asymptomatic early stages, speedy metastatic progression, lack of effective treatment protocols, early loco regional recurrence, and absence of clinically useful biomarker(s) that can detect PC in its precursor form(s) (Ballehaninna and Chamberlain, 2012). Studies have indicated a promising 70% 5-year survival for cases where incidental detections happened for stage I pancreatic tumors that were still confined to pancreas (Frena, 2001; Schneider and Schulze, 2003). Therefore, it only seems rational to aggressively screen for early detection of PDAC. CA19-9 is the most common and the only FDA approved blood-based biomarker for diagnosis, prognosis, and management of PC but it has several limitations such as poor specificity, lack of expression in the Lewis negative phenotype, and higher false positive elevation in the presence of obstructive jaundice (Ballehaninna and Chamberlain, 2012). A large number of carbohydrate antigens, cytokeratin, glycoprotein, and Mucinic markers and hepatocarcinoma–intestine–pancreas protein, and PC-associated protein markers have been discovered as a putative biomarkers for management of PC (Ballehaninna and Chamberlain, 2013). However, none of these have demonstrated superiority to CA19-9 in the validation cohorts. Previously, our group discovered a novel five-genes-based tissue biomarker for the diagnosis of PDAC using innovative meta-analysis approach on multiple transcriptome studies. This biomarker panel could distinguish PDAC from healthy controls with 94% sensitivity and 89% specificity and was also able to distinguish PDAC from CP, other cancers, and non-tumor from PDAC precursors at tissue level (Bhasin et al., 2016). The relevance of tissue-based diagnostic markers remains unclear owing to the limitations of obtaining biopsy samples. Additionally, most current studies are based on small sample sizes with limited power to identify robust biomarkers. Provided the erratic nature of PC, the major unmet requirement is to have reliable blood-based biomarkers for early diagnosis of PDAC.

The crucial requisite for better PDAC diagnosis has driven a large number of genome-level studies defining the molecular landscape of PDAC to identify early diagnosis biomarkers and potential therapeutic targets. Despite many genomics studies, we do not have a reliable biomarker that is able to surpass the sensitivity and specificity of CA19-9. The independent studies suffer from inherent statistical limitations where the datasets derived from different batches, techniques and platforms and analytic methods result in the lack of concordance (Ramasamy et al., 2008). The published gene signatures of individual microarray studies are not concordant with comparative analysis and meta-analysis studies when standard approaches are used due to variability in analytical strategies (Ramasamy et al., 2008).

In our work, we have included all the available gene expression datasets for PDAC versus healthy subjects from GEO[1] and ArrayExpress database[2] measured via microarray or sequencing platforms. We have included the datasets derived from blood and tissue sources excluding cell lines in our analysis, which was not included previously. The cell lines were excluded for they do not depict normal cell morphology and do not maintain markers and functions seen *in vivo*.

The approach of combining multiple studies has previously been stated to reveal biological insight by increasing the reproducibility and sensitivity which is generally not evident in the independent original datasets (Wang et al., 2004). Using the uniform pre-processing, normalization and batch correction approaches in the meta-analysis can assist in eliminating false-positive results. Therefore, we used multiple datasets in combinations and further divided them in training, testing and validation sets to identify and validate the markers with secretory signal peptides. We hypothesize that proteins with secretory potential will be secreted out of the tissue into the blood and these markers can be used as prognostic markers in a non-invasive manner. There were no previous studies on identification of marker genes that could be used with least-invasive methods. Also, a set of multiple genes targeting different pathways and biological processes are more reliable and sensitive than single gene-based marker for complex diseases like cancer (Ramasamy et al., 2008). We also corroborated the protein expression of our markers in proteomics datasets obtained from human protein atlas (HPA)[3].

MATERIALS AND METHODS

Dataset Identification

The publicly available microarray repositories i.e., ArrayExpress (see text footnote 2) and GEO (see text footnote 1) were searched for gene expression studies of human pancreatic specimens. The selected datasets were divided into five training sets and fourteen independent validation sets for initial development and validation of biomarkers. To avoid the representation of the datasets only from tissues the few blood studies available were divided across all training and validation phase of this study.

Each training dataset (GSE18670, E-MEXP-950, GSE32676, GSE74629, and GSE49641) included a minimum of four samples of normal pancreas and a minimum of four samples of PDAC. In training set we included minimum two datasets with source pancreatic tissue and peripheral blood. This was done to identify a predictor based on genes that are detectable in both pancreatic tissue and blood. Datasets GSE18670 (Set1: 6 normal, 5 PDAC), GSE32676 (Set6: 6 normal, 24 PDAC) and E-MEXP-950 (Set3: 10 normal, 12 PDAC) was derived from pancreatic tissue, whereas

[1] https://www.ncbi.nlm.nih.gov/geo/

[2] https://www.ebi.ac.uk/arrayexpress/

[3] https://www.proteinatlas.org/

GSE74629 (Set4: 14 normal, 32 PDAC) and GSE49641 (Set5: 18 normal, 18 PDAC) contain transcriptome profile of peripheral blood PDAC patients.

Further, 14 validation sets were also divided into three groups, one "Test sets" (**Table 1A**); second "Validation Sets" (**Table 1A**) and third "Prospective Validation Sets" (**Table 1B**). Five Tissue studies were included: one from microdissected tissue samples (Set6: 6 normal, 6 PDAC) and four from whole tissues (Set7: 45 normal, 40 PDAC; Set8: 6 normal, 6 PDAC; Set9: 8 normal and 12 PDAC and Set10: 15 normal, 33 PDAC). One blood study from peripheral blood was also validated using the biomarker (E-Set11: 14 normal, 12 PDAC).

For Phase I Validation we selected five datasets from different platforms from whole tissues and blood platelets, including comparison of normal versus PDAC samples similar to training and test sets. Four whole tissue datasets (V1: 61 normal, 69 PDAC; V2: 20 normal, 36 PDAC; V3: 9 normal, 45 PDAC; and V4: 12 normal, 118 tumor) and one dataset from blood with samples from blood platelets (V5: 50 normal, 33 PDAC) were included.

In Prospective Validation, the performance of the developed PDAC biomarker panel was tested on four additional independent datasets i.e.,: (i) PDAC versus normal (pancreatic) tissue from TCGA database (PV1: 4 normal, 150 PDAC), (ii) PDAC versus normal pancreatic tissues in early stages [PV2: 61 normal, 69 PDAC (Stage I and II)], (iii) PDAC versus CP (PV3: 9 pancreatitis, 9 PDAC), and (iv) PDAC precursor lesions (IPMA, IPMC, and IPMN) with associated invasive carcinoma [PV4: 6 normal, 15 PDAC precursors (5 IPMA, 5 IPMC, 5 IPMN)] versus normal pancreas tissues (**Table 1B**). Three datasets utilized oligonucleotide- based microarray platforms (two versions of Affymetrix GeneChips and Gene St 1.0 microarrays in one dataset) whereas the cancer genome atlas (TCGA) data is the sequencing data obtained using RNA-sequencing technology.

Quality Control and Outlier Analysis

Stringent quality control and outlier analysis was performed on all datasets used for training and validation to remove low quality arrays from the analysis. The technical quality of arrays was determined on the basis of background values, percent present calls and scaling factors using various bioconductor packages (Wilson and Miller, 2005; Kauffmann et al., 2009). The arrays with high quality were subjected to outlier analysis using array intensity distribution, principal component analysis, array-to-array correlation and unsupervised clustering. The samples that were identified to be of low quality or identified as outliers were eliminated from the analysis.

Mapping of Platform Specific Identifiers to Universal Identifier

To facilitate the collation of the differentially expressed (DE) genes identified by analysis of individual datasets, the platform specific identifiers associated with each dataset were annotated to corresponding universal gene symbol identifiers. Gene symbols were used in subsequent analyses including comparative analysis of different datasets as well as predictor development. Briefly Affymetrix data was annotated using the custom CDF from brainarray[4]. Affymetrix probe set IDs that could not be mapped to an Entrez gene identifiers were removed from the gene lists. For Agilent- 028004, HumanHT-12 V4.0 and Gene St 1.0 studies the raw matrix was directly retrieved from the GEO interactive web tool, GEO2R[5], which were further processed and normalized. The normalized and annotated genes for TCGA was obtained from Broad GDAC Firehose database[6]. We have removed 29 non-PDAC samples from TCGA during validation as our classifier was trained using PDAC samples (Peran et al., 2018).

Pre-processing and Normalization of Microarray Datasets

Potential bias introduced by the range of methodologies used in the original microarray studies, including various experimental platforms and analytic methods, was controlled by applying a uniform normalization, preprocessing and statistical analysis strategy to each dataset. Raw microarray dataset were normalized using vooma (Law, 2013) algorithm which estimates the mean-variance relationship and use the relationship to compute appropriate gene expression level weights. Similarly, RNA-sequencing datasets were normalized using voom algorithm (Law et al., 2014). The normalized datasets were used for performing meta-analysis as well as predictor development.

Differential Gene Expression Analysis for Generating Meta-Signature

To generate PDAC meta-signature, we performed differential expression analysis on individual datasets from training sets by comparing normal versus cancer samples. To identify DE genes, a linear model was implemented using the linear model microarray analysis software package (LIMMA) (Ritchie et al., 2015). LIMMA estimates the differences between normal and cancer samples by fitting a linear model and using an empirical Bayes method to moderate standard errors of the estimated log-fold changes for expression values from each probe set. In LIMMA, all genes were ranked by t-statistics using a pooled variance, a technique particularly suited to small numbers of samples per phenotype. The DE probes were identified on the basis of absolute fold change and Benjamini and Hochberg corrected *P*-value (Benjamini and Hochberg, 1995). The genes with multiple test corrected *P*-value < 0.05 were considered as DE. Comparative analyses were performed to identify those genes that are significantly DE across multiple PDAC datasets. Genes that are concordantly over or under expressed in three PDAC datasets (two tissues and one blood study) were included in PDAC meta-signature.

Secretory Gene Set Identification

To identify a non-invasive predictor based on genes with secretory potential, we selected genes that had signal peptide for secretory proteins with no transmembrane segments (noTM).

[4] http://brainarray.mbni.med.umich.edu

[5] www.ncbi.nlm.nih.gov/geo/geo2r/

[6] http://gdac.broadinstitute.org

The Biomart package in R (Durinck et al., 2005) with quering the gene symbols to SignalP database facilitated the analysis. The Ensembl Biomart database enables users to retrieve a vast diversity of annotation data for specific organisms. After loading the library, one can connect to either public BioMart databases (Ensembl, COSMIC, Uniprot, HGNC, Gramene, Wormbase and dbSNP mapped to Ensembl) or local installations of these. One set of functions can be used to annotate identifiers such as Affymetrix, RefSeq and Entrez-Gene, with information such as gene symbol, chromosomal coordinates, OMIM and Gene Ontology or vice-versa.

Training and Independent Validation of PDAC Classifier Using Support Vector Machine

The upregulated secretory genes DE from PDAC meta-signature was used for training of PDAC classifier. Classifier was generated by implementing the random forest (RF) using caret{R} and support vector machines (SVM) approach using e1071{R}. Polynomial kernel was used to develop the classifier. RF and SVM was first tuned using 10-fold cross-validation at different costs and the best cost and gamma functions were later used to perform classification on testing and validation sets. Classifiers were trained using normalized, preprocessed gene expression values from each of the five training datasets independently. To independently validate our model in each dataset, performance of classifiers in the training sets was evaluated using internal LOOCV. We assessed the classifier of five to ten genes selected from the set of upregulated genes to identify the biomarker panel that works best in both tissue and blood-based studies. The complete sets of possible combinations of five to ten genes were drawn from the upregulated genes and the accuracy of each classifier was assessed. The performance of classifiers was measured using threshold-dependent (e.g., sensitivity, specificity, accuracy) and threshold-independent ROC analysis. In ROC analysis, the AUC provides a single measure of overall prediction accuracy. The biomarker panel with the highest performance in the training sets (both tissue and blood-based studies) was chosen for assessment of predictive power in six independent test datasets using threshold-dependent and -independent measures i.e., AUC. SVM outperformed the RF models in the training datasets.

Survival Analysis

To determine the association of key genes with survival in PC, we performed survival analysis using the TCGA database[7]. The survival analysis was performed on PDAC mRNA of 150 patients [excluding samples related to normal tissues and non-PDAC tissues (Peran et al., 2018)]. Survival analysis was performed on the basis of individual mRNA expression using the Kaplan-Meier (K-M) approach (Kaplan and Meier, 1958). The normalized expression data for each gene was divided into high and low median groups. The survival analysis was

[7] https://cancergenome.nih.gov/

TABLE 1A | Datasets used for development and validation of secretory genes based PDAC classifier.

Groups	Dataset	Normal	Tumor	Sample type	Platform	Accession
Training Sets	Set 1	6	5	Enriched	U133 Plus 2.0	E-GEOD-18670
	Set 2	6	24	Whole Tissue	U133 Plus 2.0	E-GEOD-32676
	Set 3	10	12	Microdissected	U133A	E-MEXP-950
	Set 4	14	32	Peripheral Blood	HumanHT-12 V4.0	GSE74629
	Set 5	18	18	Peripheral Blood	Gene St 1.0	GSE49641
Test sets	Set 6	6	6	Microdissected	U133A	E-MEXP-1121
	Set 7	45	40	Whole Tissue	Gene St 1.0	GSE28735
	Set 8	6	6	Whole Tissue	Gene St 1.0	GSE41368
	Set 9	8	12	Whole Tissue	U133 Plus 2.0	E-GEOD-71989
	Set 10	15	33	Whole Tissue	U133 Plus 2.0	E-GEOD-16515
	Set 11	14	12	Peripheral Blood	U133 Plus 2.0	E-GEOD-15932
Validation Sets	V1	61	69	Whole Tissue	Gene St 1.0	E-GEOD-62452
	V2	20	36	Whole Tissue	U133 Plus 2.0	E-GEOD-15471
	V3	9	45	Whole Tissue	Agilent-028004	GSE60979
	V4	12	118	Whole Tissue	U219	GSE62165
	V5	50	33	Blood Platelet	HiSeq-2500	GSE68086

TABLE 1B | Datasets used for prospective validation of secretory genes based PDAC classifier.

Group	Dataset	Group	Pancreatic tumor	Sample type	Platform	Accession
Prospective Validation Sets	PV1	4 Normal	150 PDAC	Tissue	RNA-Seq	TCGA
	PV2	61 Normal	69 PDAC (Stage I and II)	Whole Tissue	Gene St 1.0	E-GEOD-62452
	PV3	9 (Pancreatitis)	9 (PDAC)	Whole Tissue	U95Av2	E-EMBL-6
	PV4	7 (Normal)	15 (IPMA, IPMC, IPMN)	Microdissected	U133 Plus 2.0	GSE19650

performed using K-M analysis from survival package in R. The results of the survival analysis were visualized using K-M survival curves with log rank testing. The results were considered significant if the *P*-values from the log rank test were below 0.05. The effects of mRNA on the event were calculated using univariate Cox proportional hazard model without any adjustments.

Pathways Analysis

The biological pathways for the genes was performed using ToppFun software of ToppGene suite (Chen et al., 2009). ToppGene is a one-stop portal for gene list enrichment analysis and candidate gene prioritization based on functional annotations and protein interactions network. ToppFun detects functional enrichment of the provided gene list based on transcriptome, proteome, regulome (TFBS and miRNA), ontologies (GO, Pathway), phenotype (human disease and mouse phenotype), pharmacome (Drug-Gene associations), literature co-citation, and other features. The biological pathways with FDR < 0.05 were considered significantly affected.

RESULTS

PDAC Differential Expression Analysis and Meta-Signature Development

To develop a gene based minimally invasive biomarker for differentiating PDAC from normal/pancreatitis, we identified 19 microarray and RNA sequencing studies containing PDAC and normal samples. These datasets based on their origin i.e., blood or tissue were divided into training sets, independent test sets, validation sets and prospective validation sets (**Figure 1**; Overview of meta-analysis strategy). For classifier training, we performed meta-analysis on 3-tissue and 2-blood-based PDAC studies to identify meta-signature that are DE in blood and tissue during PC. To account for the differences in microarray/sequencing platform used in studies, we processed and normalized studies according to their platforms and selected the genes that are common across multiple studies. The number of DE secretory genes ranged from 480 to 810 genes, totaling 2,010 significantly DE genes in the five training datasets. We identified 74 genes (35 downregulated and 39 upregulated) with concordant directionality in at least two of the three tissue datasets and one of the two blood datasets (**Figure 2A**, shown in red color and **Supplementary Table S1**).

The 74 genes showed consistent expression across the PDAC samples in the selected five datasets (3 tissue and 2 blood datasets) as compared to the normal pancreas (**Figure 2B**), with the extent of over-expression or under-expression denoted by red or green shading, respectively. Pathway analysis of these 74 common PDAC genes depicted significant enrichment (*P*-value < 0.05) in multiple extracellular matrix-associated pathways (e.g., Ensemble of genes encoding extracellular matrix and extracellular matrix-associated proteins, remodeling of the extracellular matrix, structural ECM glycoproteins, Cell adhesion molecules) (**Supplementary Figure S1**). These pathways play important roles in the adhesion of cells that is a key process in progression of PDAC.

Variables Selection and Class Prediction Analysis in Training Sets

The 39 upregulated genes from the 74 common genes were selected for predictor development. We have specifically targeted upregulated genes for their therapeutics and diagnostic applications. We plotted boxplots of these 39 genes across all the five training sets and removed the genes with opposite direction in any of these five sets. The 27 concordantly upregulated genes (**Supplementary Table S2**) were selected after the boxplot analysis. These combined gene set clearly discriminated between PDAC and normal pancreas samples in all the datasets of training set, as depicted in the heatmap for 27 genes (**Supplementary Figure S2A**) and principal component analysis (PCA) plots (**Supplementary Figure S2B**). The predictors based on 5 to 10 genes were developed and assessed by LOOCV implementing with a polynomial kernel based SVM classifier. All the possible combination of five to ten genes were tested from 27 upregulated genes. The classifiers containing the selected 9 genes i.e., IFI27, ITGB5, CTSD, EFNA4, GGH, PLBD1, HTATIP2, IL1R2, and CTSA performed with highest accuracy. These 9 genes were upregulated in PDAC as compared to the normal pancreas in all the five training sets (**Figures 2C,D**).

We performed LOOCV cross-validation analysis of the 9-gene PDAC classifier across the five training datasets to determine its predictive performance. For each of the five training datasets individually, sensitivity ranged from 0.83 to 1.0 and specificity 0.71 to 1.00 for the predictor (**Supplementary Figure S3A**, **Table 2**). Comparison of the 9-gene PDAC classifier performance in tissues (Set1-Set3) and blood datasets (Set 4 and Set 5) showed approximately 0.94 sensitivity and 0.97 specificity for the tissue datasets, in contrast to 0.88 sensitivity and 0.80 specificity for the blood datasets (**Supplementary Figure S3B**, **Table 2**). AUC for the three tissue datasets ranged from 0.89 to 1.00 with median = 0.96 (**Supplementary Figure S3B**) and for two blood datasets from 0.92 to 0.96 with median = 0.94 (**Table 2**, **Supplementary Figure S3C** and **Figure 2E**), demonstrated threshold independent performance). The average gene expression plots with all the samples combined from the five training sets (**Supplementary Figure S4A**) and the PCA plots of training sets (**Supplementary Figure S4B**) from 9 genes supported the discriminatory power of the marker combinations in identification of PDAC subjects from normal.

Biological Significance of Selected Genes

CTSA and CTSD are involved in extracellular matrix associated proteins; IFI27 and IL1R2 in cytokine signaling in immune system; ITGB5 and HTATIP2 in apoptotic pathway and EFNA4, GGH and PLBD1 are involved in Ephrin signaling, fluoropyrimidine activity and glycerophospholipid biosynthesis, respectively. The genes selected based on the presence of signal

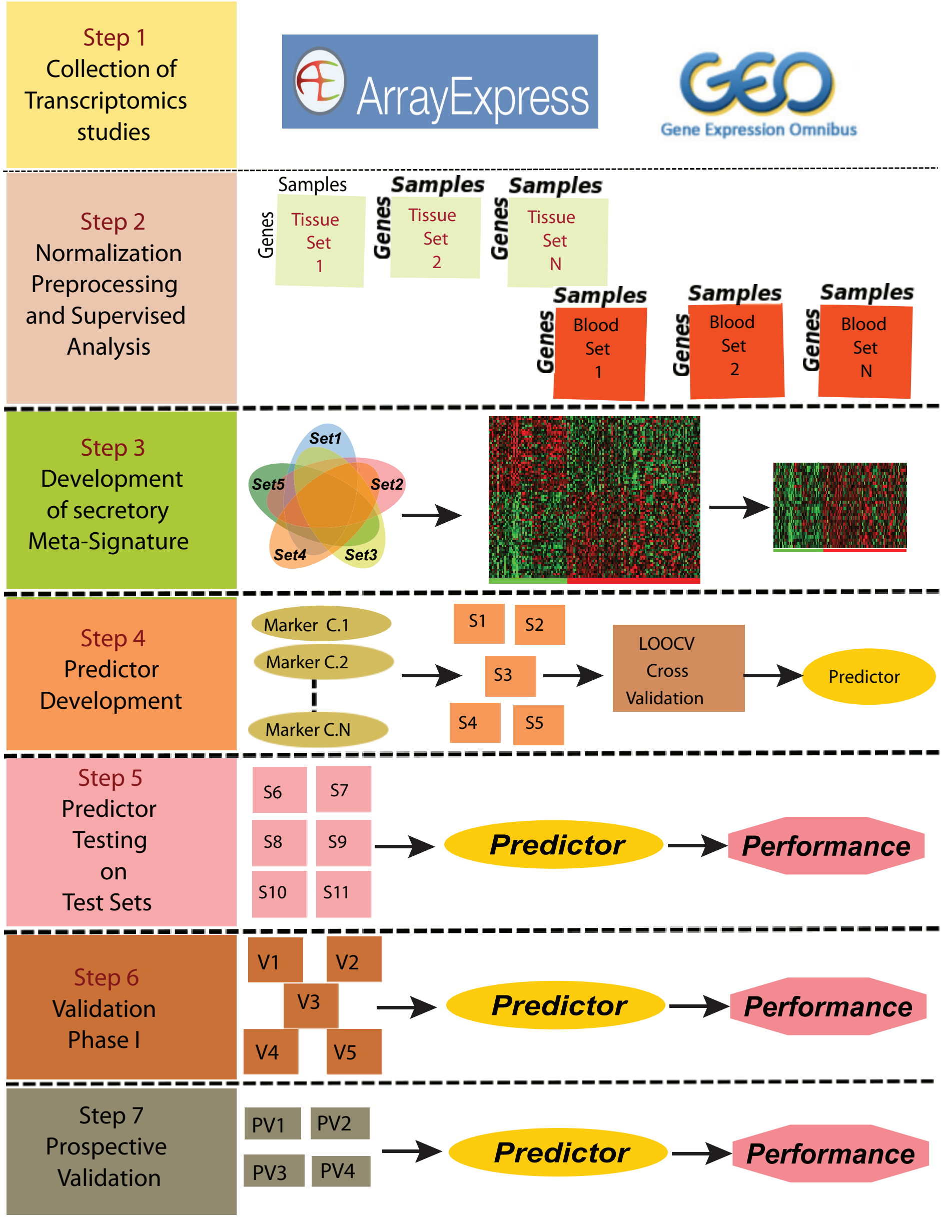


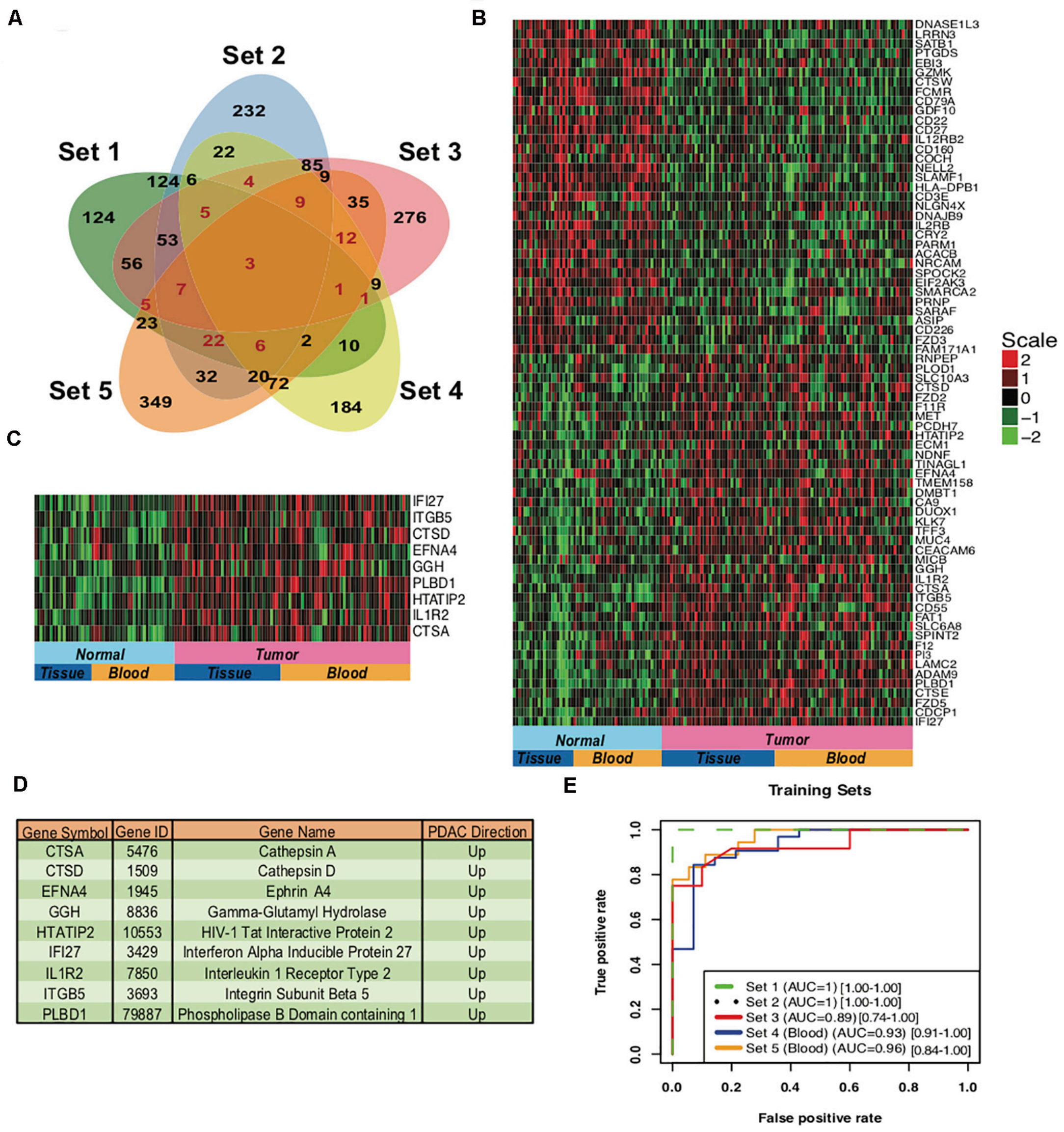
FIGURE 1 | Overview of the meta-analysis approach for development and validation of PDAC biomarker panel. Predictor was developed using the data from Set1-Set5 (S1-S5 in Step 4) and was further tested on Set5-Set10 and validated on V1-V5 and PV1-PV4 datasets.

peptide for secretion are supposed to be secretory; however, the signal peptide is also present in several membrane proteins (Uhlen et al., 2015). In the selected classifier genes, CTSD, EFNA4 and IL1R2 are predicted to be secretory proteins whereas CTSA, GGH, PLBD1, IFI27, ITGB5 and HTATIP2 are predicted to be intracellular or membrane bound proteins in HPA. Furthermore, CTSA and PLBD1 are also localized in Lysosomes and GGH is secretory protein as per UniProtKB[8] predictions. Since our 9 gene markers could be detected with a detectable expression in both tissues and blood samples from PDAC patients, we further validated the performance of these genes for PDAC Diagnosis.

Independent Performance of Classifier in Differentiating PDAC From Normal

The biomarker set designed above was further tested in six independent sets with five tissue and one blood based PDAC studies. The classifier genes depicted an upregulation pattern in most of independent validation sets **Supplementary Figure S5**. The boxplot revealed higher expression of all the 9 genes, averaged over test sets, in the tumor samples as compared to the

[8] www.uniprot.org



Gene Symbol	Gene ID	Gene Name	PDAC Direction
CTSA	5476	Cathepsin A	Up
CTSD	1509	Cathepsin D	Up
EFNA4	1945	Ephrin A4	Up
GGH	8836	Gamma-Glutamyl Hydrolase	Up
HTATIP2	10553	HIV-1 Tat Interactive Protein 2	Up
IFI27	3429	Interferon Alpha Inducible Protein 27	Up
IL1R2	7850	Interleukin 1 Receptor Type 2	Up
ITGB5	3693	Integrin Subunit Beta 5	Up
PLBD1	79887	Phospholipase B Domain containing 1	Up

FIGURE 2 | Meta-signature of genes that are consistently DE in multiple datasets and candidate PDAC diagnostic biomarker panel. **(A)** Venn diagram of the five training datasets for the DE genes. 74 genes (marked in red) with concordant directionality are common to at least 2 of the 3 tissue datasets (Set 1 to Set 3) and one of the 2 blood datasets (Set 4 and Set 5). **(B)** Heatmap of the 74 meta-signature genes DE in PDAC from five training datasets. Red = upregulated, Green = downregulated. **(C)** Heatmap of the 9-upregulated marker genes in training sets for PDAC biomarker panel. **(D)** Description of the genes from the 9-gene based PDAC biomarker panels. **(E)** AUC plot [CI: 95%] for 9-gene PDAC classifier across the five training sets using leave one out cross-validation (LOOCV). Set1 and Set 2 are matched normal samples i.e., obtained from same individual. Set 3 normal samples are not matched, Normal samples are obtained from the patients undergoing surgery with other pancreatic diseases. Set 4 and Set 5 are blood sourced studies therefore the normal subjects were matched for gender, age and habits.

healthy (**Figure 3A**). For each of the six datasets individually, sensitivity ranged from 0.75 to 1.00 and specificity from 0.71 to 1.00 for the predictor (**Figure 3B**, **Table 2**). Comparison of the 9-gene PDAC classifier performance in tissue and blood showed an average 0.94 sensitivity and 0.97 specificity for the tissue datasets, in contrast to 0.75 sensitivity and 0.71 specificity for the blood dataset. AUC for the five tissue datasets ranged from 0.94 to 1.00 and for one blood datasets AUC was 0.80 (**Figure 3C**, **Table 2**).

High Accuracy of Our 9-Gene PDAC Classifier in Predicting PDAC in 5 Independent Validation Sets

In five validation sets, the 9-gene PDAC classifier accurately predicted the class of PDAC compared to normal with maximum AUC of 1.00 in the independent validation tissue (V2) set that contained 20 normal and 36 PDAC samples. More than 0.95 AUC was observed in three independent validation tissue sets (V2, V3 and V4) that contained 36, 45 and 118 PDAC

TABLE 2 | The performance matrix of the 9-gene PDAC classifier on the training, testing, validation and prospective validation sets.

Groups	Datasets	Accuracy	Sensitivity	Specificity	AUC
Training Sets	Set 1	1.00	1.00	1.00	1.00
	Set 2	1.00	1.00	1.00	1.00
	Set 3	0.87	0.83	0.90	0.89
	Set 4	0.82	0.93	0.71	0.93
	Set 5	0.86	0.83	0.89	0.97
Test Sets	Set 6	1.00	1.00	1.00	1.00
	Set 7	0.92	0.90	0.93	0.94
	Set 8	1.00	1.00	1.00	1.00
	Set 9	0.95	0.91	1.00	1.00
	Set 10	0.96	0.93	1.00	0.94
	Set 11	0.73	0.75	0.71	0.80
Validation Sets	V1	0.79	0.76	0.83	0.83
	V2	0.98	0.97	1.00	1.00
	V3	0.94	1.00	0.89	0.98
	V4	0.95	1.00	0.91	0.99
	V5	0.83	0.84	0.82	0.89
Prospective Validation Sets	PV1	0.82	0.94	0.72	0.93
	PV2	0.74	0.74	0.75	0.82
	PV3	0.83	0.78	0.89	0.95
	PV4-IPMA	1.00	1.00	1.00	1.00
	PV4-IPMC	0.84	0.83	0.86	0.81
	PV4-IPMN	1.00	1.00	1.00	1.00

and 20, 9 and 12 normal pancreas samples, respectively, (**Figure 4A** and **Table 1B**). The boxplot revealed higher expression of all the 9 genes, averaged over validation sets, in the tumor samples as compared to the healthy samples (**Figure 4B**). In a tissue dataset (V1) containing 61 normal and 69 tumor samples a specificity of 0.83 and sensitivity of 0.76 was determined. In 50 normal and 33 PDAC blood platelet sample (V5) 0.84 sensitivity, 0.82 specificity and 0.88 AUC was achieved. The prediction of the PDAC class in comparison to normal was accurate with a sensitivity ranging 0.76–1.00 and specificity ranging between 0.82 and 1.00 (**Figure 4C** panel II, **Table 2**). **Supplementary Figure S6** presents the heatmap of the nine genes in individual validation datasets and the PCA plots depicting the discrimination of PDAC from normal samples.

Cross-Platform Performance of Classifier on TCGA Pancreatic Samples

We further estimated the cross-platform performance of classifiers on the most widely used PC sample resource namely TCGA. TCGA dataset contains 150 PDAC samples and 4 normal samples and gene expression pattern analysis is not in consistence with other studies (**Supplementary Figure S7C**). The cross-platform validation of classifier on TCGA data also achieved high sensitivity (0.94) and specificity (0.72) indicating the stability of the classifier in handling the cross-platform variation in absolute gene expression signal (**Figure 5** PV1). The classifier achieved an excellent AUC of 0.93 (**Table 2**). The lower specificity of TCGA datasets might be due to the limited number of normal samples in the dataset. Heatmap of the 9 genes and PCA plots depicts the discrimination of two classes with the nine genes in the TCGA samples (**Supplementary Figure S7** PV1).

The markers did not show concordance in the TCGA dataset; however, the significance of these genes in the survival analysis can be very well established using the TCGA database. The samples were partitioned at median for selected nine-genes and survival analysis was performed on two clusters (**Supplementary Figure S8**). The results showed the combined survival of genes was able to clearly discriminate between better and poor survivors (*P*-value significance of 0.05 and hazard Ratio of 0.85), indicating their prognostic role in PDAC. High CTSD, EFNA4, HTATIP2, IFI27, ITGB5 and PLBD1 expression is associated with shortened survival time. Also, the survival analysis of these genes with a hazard ratio of >1 at significant *P*-value indicates their prognostic importance.

Performance of Classifier in Identifying Early Stage PDAC

As it is well established in literature that lack of established strategies for early detection of PDAC result in poor prognosis and mortality, we therefore tested performance of our classifiers on stage I and II PDAC. The predictor could distinguish stage I and II PDACs from normals with 0.74 sensitivity and 0.75 specificity and an AUC 0.82 (**Figure 5** PV2, **Table 2**). Heatmap of the nine genes and PCA plots depicts the discrimination of two classes with the nine genes in early stages PDAC samples (**Supplementary Figure S7** PV2).

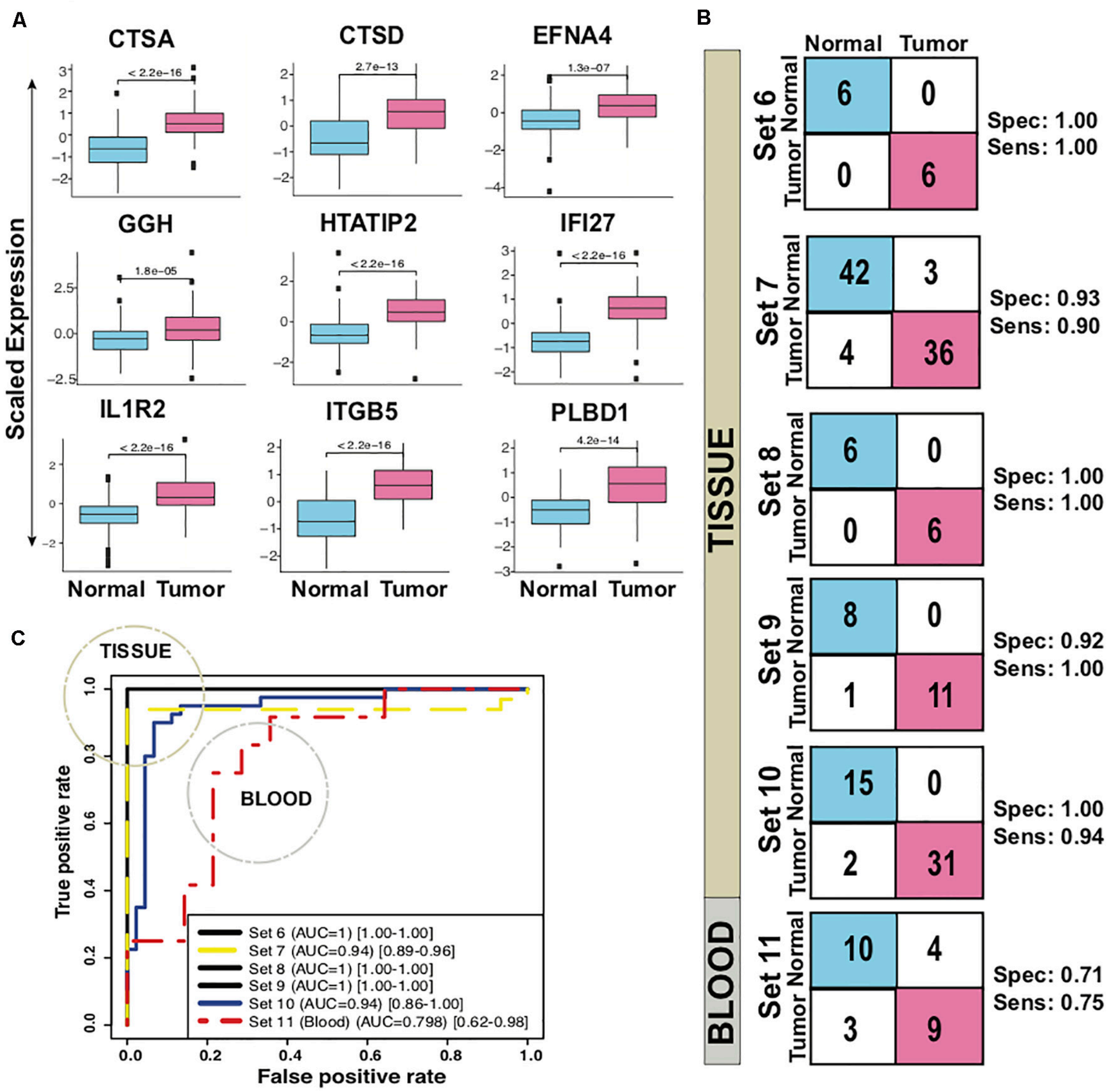


FIGURE 3 | Performance of 9-gene PDAC Classifier on test sets using leave one out cross-validation (LOOCV). **(A)** The boxplot of the averaged expression of the genes across all the six test datasets. The *P*-values as calculated by *t*-test between the groups are on the individual genes. **(B)** Diagnostic performance of the 9-gene PDAC classifier on the six test sets of PDAC vs. normal pancreas. Sensitivity (Sens.) and specificity (Spec.) indicated besides each set. **(C)** AUC plot for 9-gene [CI: 0.95–0.99] PDAC classifier across the six test datasets.

Performance of Classifier in Discriminating PDAC From Pancreatitis

Identification of CP and discriminating it from PDAC is a key challenge. As our 9-gene PDAC classifier accurately established the differences between PDAC and CP, it is important to include further validation steps for the biomarker panel. The array U95Av2 have the recorded signal intensity values for all the genes except PLBD1, hence only 8 genes were tested as a classifier for the discrimination of CP from PDAC. We tested the biomarker on the PV3 dataset wherein there were nine samples each for CP and PDAC. The classifier genes on PV3 dataset depicted significantly altered expression pattern between PDAC from CP (**Supplementary Figure S7** PV3). The classifier achieved a specificity of 0.89 and sensitivity of 0.78 with an overall accuracy of 0.83 and an AUC of 0.95 in discriminating PDAC from CP (**Figure 5** PV3, **Table 2**).

Classifier Discriminated Pre-cancerous Lesions From Normal Pancreas With Good Accuracy

To estimate the ability of the biomarker panel in discriminating precancerous lesions from a normal pancreas, we tested its performance on independent dataset containing normal main pancreatic duct epithelial cells microdissected by lasers and neoplastic epithelial cells from potential PDAC precursor lesions i.e., IPMA, IPMC and IPMN [15]. Classifier genes

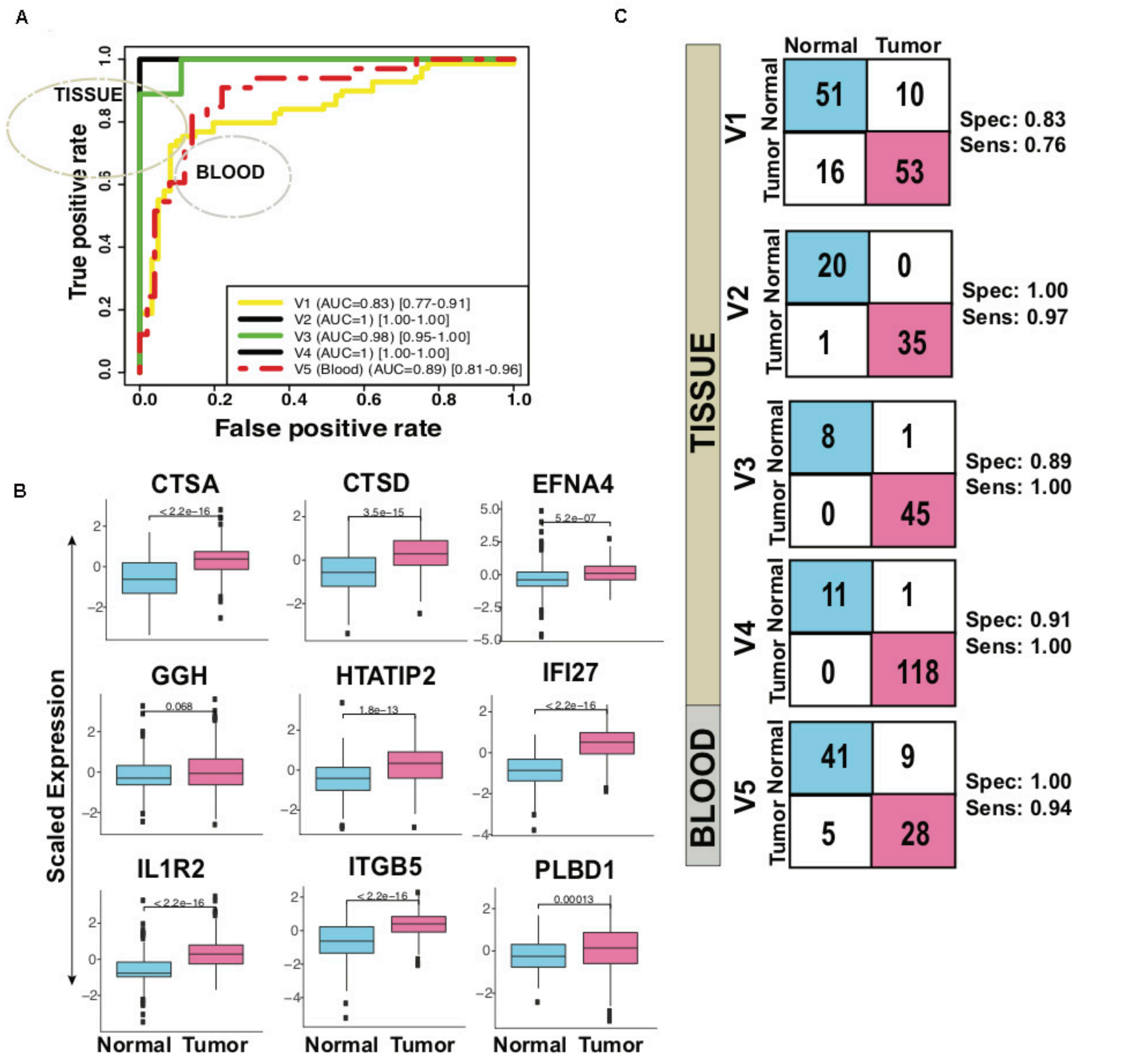


FIGURE 4 | Performance of 9-gene PDAC Classifier on validation sets using leave one out cross-validation (LOOCV). **(A)** The boxplot of the averaged expression of the genes across all the five validation datasets. The *P*-values as calculated by *t*-test between the groups are mentioned on the individual genes. **(B)** Diagnostic performance of the 9-gene PDAC classifier on the five validation sets of PDAC vs. normal pancreas. Sensitivity (Sens.) and specificity (Spec.) indicated besides each set. **(C)** AUC plot [CI: 0.95–0.99] for 9-gene PDAC classifier across the five validation datasets.

were consistently overexpressed in the PDAC samples, GGH was under-expressed in IPMA samples whereas it was overexpressed across the other PDAC precursors, IPMC and IPMN (**Supplementary Figure S9**). The 9-gene PDAC classifier separated all potential PDAC precursor (IPMA, IPMC, IPMN) samples from the normal pancreatic duct samples except for one normal sample and one IPMC sample (**Figure 5** PV4). The biomarker panel differed IPMA and IPMN from normal pancreas with 1.00 sensitivity and 1.00 specificity, achieving an AUC of 1.00 (**Figure 5** PV4). The predictor separated IPMC from healthy pancreas with 0.83 sensitivity and 0.86 specificity, achieving an AUC of 0.81 (**Table 2**).

Classifier Performed Better Than Previously Known Markers

To estimate the performance of our current marker as compared to the previously established markers we compared the performance of our marker with each study [Bhasin et al. (2016), Balasenthil et al. (2017), Kisiel et al. (2015), and Immunovia (Mellby et al., 2018)]. We used polynomial

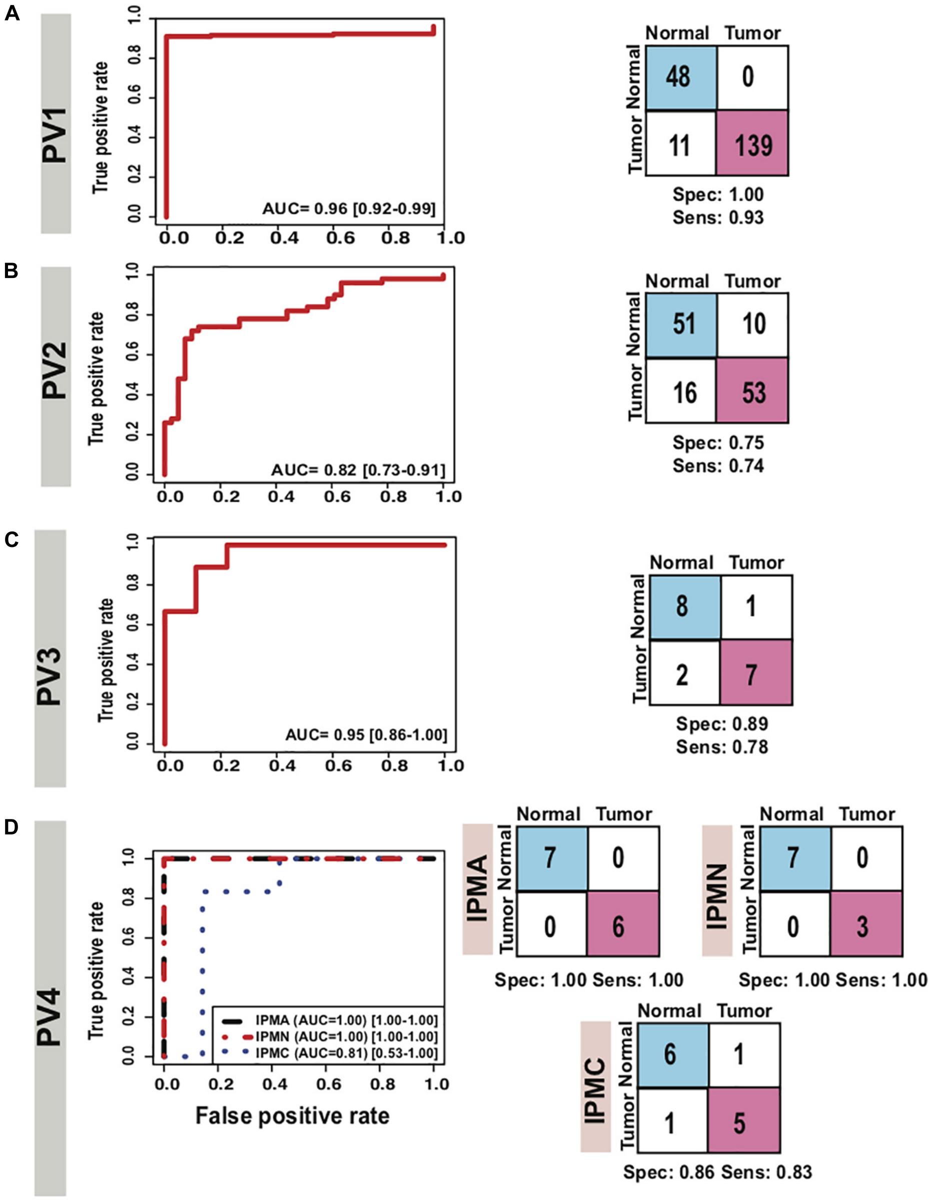


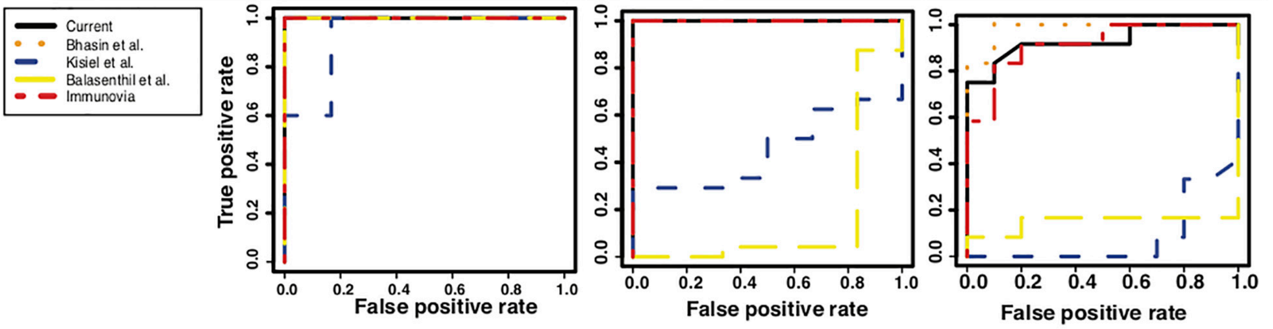
FIGURE 5 | Performance of 9-gene PDAC Classifier on prospective validation sets using leave one out cross-validation (LOOCV). AUC plot [CI: 0.95–0.99] for 9-gene PDAC classifier and the diagnostic performance of **(A)** the classifier for PV1 dataset, **(B)** the classifier for PV2 dataset. **(C)** The classifier for IPMA, IPMC and IPMN subjects in PV4 dataset and **(D)** the classifier for PV3 dataset.

kernel for each set of markers and selected best model to record the performance on all the training, test and validation datasets (**Supplementary Figure S10** and **Supplementary Table S3**). We found that all the methods performed well in tissue biopsies samples whereas when applied to the blood studies the performance of our marker set is the best (**Figure 6**). Our set of markers has performed well in tissues as well as blood studies and will be an ideal minimally invasive biomarker for studying in future studies and clinical trials.

Validation of the Markers in Single-Cell Transcriptomics Studies

Furthermore, as the markers are derived from bulk sequencing protocols it is important to know if the markers discovery is not influenced by different cell-types in normal and cancerous pancreas. Therefore, we used single-cell RNA-sequencing data published by Peng et al. (2019) suggesting heterogeneity in PDAC tumor to plot expression of our markers on different cell-types. Using standard Seurat single-cell analysis methodology (Butler et al., 2018; Stuart et al., 2019), we identified that our

TISSUE	Set 1				Set 2				Set 3			
	Acc	Sens	Spec	AUC	Acc	Sens	Spec	AUC	Acc	Sens	Spec	AUC
Current	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.87	0.83	0.90	0.89
Bhasin et al	0.91	1.00	0.83	1.00	1.00	1.00	1.00	1.00	0.95	1.00	0.90	0.98
Balasenthil et al	1.00	1.00	1.00	1.00	0.50	1.00	0.00	0.57	0.47	0.83	0.1	0.16
Kisiel et al	0.63	0.60	0.67	0.73	0.50	1.00	0.00	0.07	0.20	0.41	0.00	0.15
Immunovia	0.80	0.60	1.00	1.00	1.00	1.00	1.00	1.00	0.78	0.67	0.90	0.92



BLOOD	Set 4				Set 5				Set 11			
	Acc	Sens	Spec	AUC	Acc	Sens	Spec	AUC	Acc	Sens	Spec	AUC
Current	0.82	0.93	0.71	0.93	0.89	0.89	0.89	0.93	0.73	0.75	0.71	0.80
Bhasin et al	0.49	0.97	0.00	0.12	0.50	0.45	0.56	0.53	0.47	0.00	0.93	0.22
Balasenthil et al	0.50	1.00	0.00	0.01	0.78	0.78	0.78	0.81	0.71	0.58	0.85	0.65
Kisiel et al	0.50	1.00	0.00	0.35	0.47	0.44	0.50	0.49	0.51	0.16	0.85	0.23
Immunovia	0.72	0.88	0.57	0.77	0.59	0.56	0.62	0.64	0.29	0.09	0.50	0.18

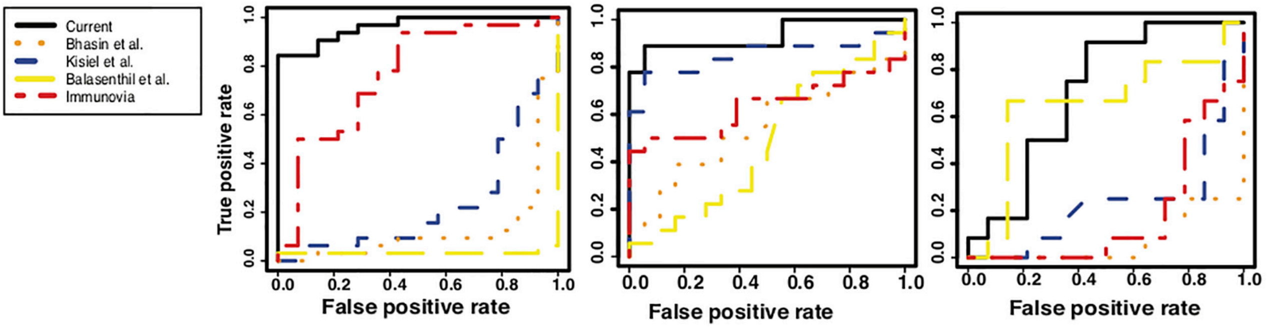


FIGURE 6 | Comparative performance of 9-gene PDAC Classifier with different previously established biomarkers. AUC plot [CI: 0.95–0.99] for 9-gene PDAC classifier across the three tissue and three blood datasets. The boxes colored in mustard color have greater than 0.80 AUC.

markers are not associated with any cell-types and are expressed across major cell types in pancreatic cancer (**Supplementary Figure S11**). All our markers depicted upregulation in various tumor microenvironment cells including immune cells and endothelial cells.

Validation of Markers in Blood-Based Proteomics Study

The nine-gene markers in the classifier were discovered and validated from the transcriptomics studies, hence the validation of their expression at the protein level is necessary. Therefore, we confirmed the expression of the nine genes at the protein level in publicly available proteomics studies and HPA. The immunolabeling of the proteins of the respective genes in HPA (**Supplementary Figure S12**) suggest higher staining of the proteins in tumors as compared to the normal samples except IFI27 where the expression of the protein cannot be detected. To further validate the protein expression of our markers we searched for the corresponding proteins in multiple pancreatic cancer proteomics studies (Chen et al., 2005; Crnogorac-Jurcevic et al., 2005; Cui et al., 2009; McKinney et al., 2011; Kosanam et al., 2013; Wang et al., 2013; Iuga et al., 2014). CTSD, a cathepsin family protein, and Ephrin and Interferon gamma family markers are found to be highly expressed in multiple proteomics studies (Chen et al., 2005; Cui et al., 2009; McKinney et al., 2011).

DISCUSSION

We applied a data mining approach to a large number of publicly available transcriptome datasets derived from pancreatic cancer and healthy blood and tissues, followed by class prediction analysis using machine learning and validation of the classifier in the independent datasets to discover candidate PDAC biomarkers (Harsha et al., 2009; Ranganathan et al., 2009). We explored the genes with secretory peptide DE in the PDAC as compared to normal pancreas/blood, for the first time to investigate an accurate secretory/non-invasive biomarker panel for the PDAC diagnosis. We report here a 9-gene PDAC classifier that differentiates PDAC as well as the precursor lesions from the normal with high accuracy. This 9-gene PDAC classifier was validated independently in 12 different blood and tissue studies. The 9-gene PDAC classifier encodes proteins with secretory potential in pancreas and few other tissues.

Approximately 2500 candidate biomarkers have been associated with PDAC previously while some of these biomarkers are in various evaluation stages only CA19-9 is approved by FDA (Koprowski et al., 1979, 1981; Hyöty et al., 1992). However, accuracy of CA19-9 is not accurate enough for screening, especially for an early detection of PDAC. Presently, the extensive validation of diagnostic or predictive gene/protein expression biomarkers for accurate discrimination between healthy patients, benign, premalignant and malignant disease are still lacking. Therefore, we aimed to identify a biomarker panel with greater sensitivity and specificity and identified a 9-gene marker panel that performs with high accuracy in discriminating PDAC with normal pancreas across multiple platforms, using either whole/microdissected tissue or peripheral blood.

To determine whether the genes in our classifier reflect key pathophysiological pathways associated with the development of PDAC, we reviewed available information for the role of these genes. Most of our 9-gene classifier genes have been linked to tumorigenesis, indicating a causal role in the development and progression of PDAC. HTATIP2 is involved in apoptosis function in liver metastasis related genes (Shi et al., 2009), gastric cancer (Xu et al., 2010) and pancreatic cancer (Ouyang et al., 2014). IFI27, functioning in immune system, has been suggested as a marker of epithelial proliferation and cancer (Grutzmann et al., 2003; López-Casas and López-Fernández, 2010). ITGB5 involved in integrin signaling have been found to be upregulated in several analysis studies (Van den Broeck et al., 2012). The Integrin and ephrin pathways have been proposed to play a crucial role in pancreatic carcinogenesis and progression, including *ITGB1*, a paralog of *ITGB5*, and EPHA2 as most important regulators (Van den Broeck et al., 2012). EPHA2 belongs to ephrin receptor subfamily and is involved in developmental events, especially in the nervous system and in erythropoiesis. To this family belongs one of our genes EFNA4 which activates another ephrin receptor EPHA5. IL1R2 was identified as possible candidate gene in PDAC that can lead to defects of the apoptosis pathway (Rückert et al., 2010). Moreover, Il1, the ligand of IL1R2, is secreted by the pancreatic cells (Arlt et al., 2002) and has an important function in inflammation and proliferation that can also trigger the apoptosis (Dupraz et al., 2000; Ruckdeschel et al., 2002; Yoshida et al., 2004). CTSD have been shown to be upregulated in the PDAC cancer (Iacobuzio-Donahue et al., 2003). AGR2, a surface antigen, has been shown to promote the progression of PDAC cells through regulation of Cathepsins B and D genes (Dumartin et al., 2011). CTSA was identified as one of the 76 deregulated genes in a study aiming for the development of early diagnostic markers as well as potential novel therapeutic targets for both familial and sporadic PDAC (Crnogorac-Jurcevic et al., 2013). PLBD1 has been found to be upregulated in various studies with five-fold increase in cell lines (Makawita et al., 2011) and in study where the effect of pancreatic β-cells inducing immune-mediated diabetes was studied (Salem et al., 2014). Metabolism-related GGH has been found to be relevant and upregulated in gallbladder carcinomas (Washiro et al., 2008).

Most of the genes in the 9-gene classifier (ITGB1, EPHA2, IL1R2) are involved in the migration, immune pathways, adhesion and metastasis of PDAC or other cancers, that are specifically associated with the developmental events and signaling in the progression of cancer. To corroborate the involvement of these genes in PDAC progression and early stages of PDAC development, we evaluated the expression levels of these genes in the early lesions of PDAC precursors i.e., LIGD-IPMN, HGD-IPMN and InvCa-IPMN (**Figure 5**) [15]. Eight genes except GGH are upregulated in IPMA, IPMN, and IPMC as well as in PanINs, as compared to a normal pancreas, demonstrating their enhanced expression is linked with the progression of PDAC that occurs early during development of malignancy. The outcomes of our study clearly show that our 9-gene classifier reflect drivers of early defects during progression and development of PDAC. This argument is further strengthened by the survival analysis of the genes where five of the nine genes (CTSA, CTSD, EFNA4, IFI27 and IL1R2) are strongly related to discriminating better and poor survivors.

Since individuals with CP are at increased risk of developing PDAC and pathological discrimination is challenging between CP and PDAC which makes it important for a classifier to discriminate between these two disease stages. While other studies have performed meta-analysis of transcriptome data for PDAC to identify the genes that are overexpressed in PDAC (Iacobuzio-Donahue et al., 2003; López-Casas and López-Fernández, 2010; Munding et al., 2012), they are irrelevant in identifying the markers for prognosis of PDAC. Our 9-gene biomarker classifier accurately distinguished premalignant and malignant pancreatic lesions such as PanIN, IPMA, IPMN and IPMC from healthy pancreas. As all 9 genes of our classifier are upregulated in PanIN (as compare to normal pancreas) already, it indicates that these 9 genes are dysregulated in early lesions during the process of PDAC development and therefore could assist in an early detection of PDAC.

Further, to analyze the potential of the 9-gene biomarker in accurate classification of PDAC subjects versus healthy subjects we compared our biomarker combination with previously known and established biomarkers. Our analysis also indicates that the 9-gene biomarker panel including multiple genes, rather than a single biomarker, is more powerful and had possibility to improve the specificity and selectivity for an accurate detection of PDAC. The idea behind generation of biomarker panel with the better identification in blood sample, in corroboration with the tissue studies, is fulfilled here. The previously established markers worked well in the tissue studies but could not show their similar potential in blood studies.

Further, the protein expression of selected biomarker genes was also examined to determine their association with PDAC at protein levels. The analysis depicted that multiple gene product/proteins corresponding to biomarkers genes depicted higher expression in pancreatic cancer tissues. Interestingly some marker (e.g., EFNA4, GGH) also depicted over-expression in other cancers indicating their association with tumor development and progression related hallmark processes. In recent years, multiple proteomics studies were performed to understand the proteome landscape of the PDAC but still lack in generating comprehensive picture due to technological limitations. Most of the proteomics technique can measure the expression of 2,000-3,000 proteins that is far from generating the global overview of proteome. High expression of Cathepsin family proteins specifically CTSD is noted in several proteomics studies which was also the case for Ephrin and Interferon gamma family markers (Chen et al., 2005; Cui et al., 2009; McKinney et al., 2011). Also, the expression of these genes is not found to be related to a particular cell-type in pancreatic cancer cell lineage. However, the fact that the overall study is based on bulk sequencing data cannot be overlooked and these cells may comprise of multiple cell-types which may or may not influence the overall methodology of marker selection. Overall, the protein-expression of the selected genes and their expression in multiple cell-types of pancreatic cancer is established. However, the aforementioned limitations have to be challenged before designing the diagnostic panel. The 9-gene markers identified here still needs validation in a bigger cohort for its potential in identifying accurately the early stages but this marker combination potentially has shown its discriminatory power across various blood and tissue datasets obtained from different sources and different platforms.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

IK performed the bioinformatics analysis and wrote the manuscript. MB supervised the bioinformatics analysis and edited the manuscript. Both the authors read and approved the final manuscript.

FUNDING

This study was supported through BIDMC CAO Innovation grant.

ACKNOWLEDGMENTS

This manuscript has been released as a pre-print at medRxiv, Khatri and Bhasin (2020).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.572284/full#supplementary-material

FIGURE S1 | Pathway enrichment analysis of the 74 PDAC-specific secretory genes.

FIGURE S2 | Upregulated Secretory genes in training datasets. **(A)** Heatmap of 27 upregulated secretory genes in PDAC for two of the three tissues and one of the two blood datasets. **(B)** PCA plots for each training datasets using 27 upregulated secretory genes.

FIGURE S3 | Performance of 9-gene PDAC classifier on training sets using leave one out cross-validation (LOOCV). **(A)** Diagnostic performance of the 9-gene PDAC classifier on the five training sets. Sensitivity (Sens) and Specificity (Spec) are indicated for each dataset. **(B)** AUC plot for 9-gene PDAC classifier on the three tissue training datasets. **(C)** AUC plot for 9-gene PDAC classifier on the two blood training datasets.

FIGURE S4 | The metrics for training datasets using the 9-biomarker panel genes. **(A)** Boxplot of the averaged expression of the genes across all the five training datasets. **(B)** PCA plots for each training datasets using the 9-biomarker panel genes.

FIGURE S5 | The assessment metrics for testing datasets using the 9-biomarker panel genes. **(A)** Heatmap of the 9 PDAC-upregulated marker genes. **(B)** PCA plots in six independent testing datasets.

FIGURE S6 | The assessment metrics for validation datasets using the 9-biomarker panel genes. Heatmaps **(A)** and PCA plots **(B)** based on biomarker panel genes in validation sets.

FIGURE S7 | The assessment metrics for PV1-3 dataset using the 9-biomarker panel genes. **(A)** PCA plots of three different prospective validation datasets. **(B)** Heatmaps of the 9-marker genes panel. **(C)** Boxplots of the expression of the genes.

FIGURE S8 | Survival curve of 9-gene-based PDAC classifier and combined genes.

FIGURE S9 | The assessment metrics for PV4 dataset using the 9-biomarker panel genes. **(A)** PCA plots for precursor lesions in three stages IPMA, IPMN and IPMC. **(B)** Heatmaps of the 9-marker genes panel. **(C)** Boxplots of the expression of the genes in precursor lesions.

FIGURE S10 | Comparative performance of 9-gene-based PDAC classifier with different previously established biomarkers. AUC plot for 9-gene-based PDAC classifier across the training and validation datasets. The measures of performances e.g., accuracy, sensitivity, specificity and AUC are mentioned in **Supplementary Table S3**.

FIGURE S11 | Expression of 9-gene markers in different pancreas cell-types in both healthy and tumor states. The expression of these genes is high in tumor state (CTSA, CTSD, EFNA4, GGH, HTATIP2, IFI27, and ITGB5) or they are not expressed at all in healthy state (IL1R2 and PLBD1). This is also consistent with protein expression of the genes as measured by antibody staining experiments by HPA.

FIGURE S12 | Immunolabeling of protein expression of nine genes selected for the classifier in pancreatic cancer. Light blue is low staining; blue is moderate staining and brown is high.

TABLE S1 | Log2 fold change of the significantly DE genes identified from different training datasets.

TABLE S2 | Direction of differentially upregulated genes validated via boxplot analysis. Upregulated are shown with green background and ones with opposite direction are colored black.

TABLE S3 | Comparative performance of 9-gene PDAC Classifier with different previously established biomarkers in training, test and validation datasets. Sets with green background are datasets derived from blood. All mustard colored cells have AUC > 0.80 whereas light blue cells indicate low specificity or sensitivity despite of high AUC. For black shaded cells all the genes corresponding to the mentioned studies cannot be identified.

REFERENCES

Arlt, A., Vorndamm, J., Müerköster, S., Yu, H., Schmidt, W. E., Fölsch, U. R., et al. (2002). Autocrine production of interleukin 1beta confers constitutive nuclear factor kappaB activity and chemoresistance in pancreatic carcinoma cell lines. *Cancer Res.* 62, 910–916.

Balasenthil, S., Huang, Y., Liu, S., Marsh, T., Chen, J., Stass, S. A., et al. (2017). A plasma biomarker panel to identify surgically resectable early-stage pancreatic cancer. *JNCI J. Natl. Cancer Inst.* 109:djw341. doi: 10.1093/jnci/djw341

Ballehaninna, U. K., and Chamberlain, R. S. (2012). The clinical utility of serum CA 19-9 in the diagnosis, prognosis and management of pancreatic adenocarcinoma: an evidence based appraisal. *J. Gastrointest. Oncol.* 3, 105–119. doi: 10.3978/j.issn.2078-6891.2011.021

Ballehaninna, U. K., and Chamberlain, R. S. (2013). Biomarkers for pancreatic cancer: promising new markers and options beyond CA 19-9. *Tumor Biol.* 34, 3279–3292. doi: 10.1007/s13277-013-1033-3

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* 57, 289–300. doi: 10.2307/2346101

Bhasin, M. K., Ndebele, K., Bucur, O., Yee, E. U., Otu, H. H., Plati, J., et al. (2016). Meta-analysis of transcriptome data identifies a novel 5-gene pancreatic adenocarcinoma classifier. *Oncotarget* 7, 23263–23281. doi: 10.18632/oncotarget.8139

Brand, R. E., and Matamoros, A. (1998). Imaging techniques in the evaluation of adenocarcinoma of the pancreas. *Dig. Dis.* 16, 242–252. doi: 10.1159/000016872

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* 36, 411–420. doi: 10.1038/nbt.4096

Chen, J., Bardes, E. E., Aronow, B. J., and Jegga, A. G. (2009). ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res.* 37, W305–W311. doi: 10.1093/nar/gkp427

Chen, R., Yi, E. C., Donohoe, S., Pan, S., Eng, J., Cooke, K., et al. (2005). Pancreatic cancer proteome: the proteins that underlie invasion, metastasis, and immunologic escape. *Gastroenterology* 129, 1187–1197. doi: 10.1053/j.gastro.2005.08.001

Crnogorac-Jurcevic, T., Chelala, C., Barry, S., Harada, T., Bhakta, V., Lattimore, S., et al. (2013). Molecular analysis of precursor lesions in familial pancreatic cancer. *PLoS One* 8:e54830. doi: 10.1371/JOURNAL.PONE.0054830

Crnogorac-Jurcevic, T., Gangeswaran, R., Bhakta, V., Capurso, G., Lattimore, S., Akada, M., et al. (2005). Proteomic analysis of chronic pancreatitis and pancreatic adenocarcinoma. *Gastroenterology* 129, 1454–1463. doi: 10.1053/j.gastro.2005.08.012

Cui, Y., Tian, M., Zong, M., Teng, M., Chen, Y., Lu, J., et al. (2009). Proteomic analysis of pancreatic ductal adenocarcinoma compared with normal adjacent pancreatic tissue and pancreatic benign cystadenoma. *Pancreatology* 9, 89–98. doi: 10.1159/000178879

Dumartin, L., Whiteman, H. J., Weeks, M. E., Hariharan, D., Dmitrovic, B., Iacobuzio-Donahue, C. A., et al. (2011). AGR2 is a novel surface antigen that promotes the dissemination of pancreatic cancer cells through regulation of cathepsins B and D. *Cancer Res.* 71, 7091–7102. doi: 10.1158/0008-5472.can-11-1367

Dupraz, P., Cottet, S., Hamburger, F., Dolci, W., Felley-Bosco, E., and Thorens, B. (2000). Dominant negative MyD88 proteins inhibit interleukin-1beta /interferon-gamma -mediated induction of nuclear factor kappa B-dependent nitrite production and apoptosis in beta cells. *J. Biol. Chem.* 275, 37672–37678. doi: 10.1074/jbc.M005150200

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., et al. (2005). BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* 21, 3439–3440. doi: 10.1093/bioinformatics/bti525

Fesinmeyer, M. D., Austin, M. A., Li, C. I., De Roos, A. J., and Bowen, D. J. (2005). Differences in survival by histologic type of pancreatic cancer. *Cancer Epidemiol. Biomarkers Prev.* 14, 1766–1773. doi: 10.1158/1055-9965.EPI-05-0120

Frena, A. (2001). SPan-1 and exocrine pancreatic carcinoma. The clinical role of a new tumor marker. *Int. J. Biol. Markers* 16, 189–197. doi: 10.1177/172460080101600306

Grutzmann, R., Foerder, M., Alldinger, I., Staub, E., Brummendorf, T., Ropcke, S., et al. (2003). Gene expression profiles of microdissected pancreatic ductal adenocarcinoma. *Virchows Arch.* 443, 508–517. doi: 10.1007/s00428-003-0884-1

Harsha, H. C., Kandasamy, K., Ranganathan, P., Rani, S., Ramabadran, S., Gollapudi, S., et al. (2009). A compendium of potential biomarkers of pancreatic cancer. *PLoS Med.* 6:e1000046. doi: 10.1371/journal.pmed.1000046

Hyöty, M., Hyöty, H., Aaran, R. K., Airo, I., and Nordback, I. (1992). Tumour antigens CA 195 and CA 19-9 in pancreatic juice and serum for the diagnosis of pancreatic carcinoma. *Eur. J. Surg.* 158, 173–179.

Iacobuzio-Donahue, C. A., Maitra, A., Olsen, M., Lowe, A. W., van Heek, N. T., and Rosty, C. (2003). Exploration of global gene expression patterns in pancreatic adenocarcinoma using cDNA microarrays. *Am. J. Pathol.* 162, 1151–1162. doi: 10.1016/S0002-9440(10)63911-9

Iuga, C., Seicean, A., Iancu, C., Buiga, R., Sappa, P. K., Völker, U., et al. (2014). Proteomic identification of potential prognostic biomarkers in resectable pancreatic ductal adenocarcinoma. *Proteomics* 14, 945–955. doi: 10.1002/pmic.201300402

Kaplan, E. L., and Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.* 53, 457–481. doi: 10.1080/01621459.1958.10501452

Kauffmann, A., Gentleman, R., and Huber, W. (2009). arrayQualityMetrics–a bioconductor package for quality assessment of microarray data. *Bioinformatics* 25, 415–416. doi: 10.1093/bioinformatics/btn647

Khatri, I., and Bhasin, M. K. (2020). A transcriptomics-based meta-analysis combined with machine learning approach identifies a secretory biomarker panel for diagnosis of pancreatic adenocarcinoma. *medRxiv* [Preprint]. doi: 10.1101/2020.04.16.20061515

Kisiel, J. B., Raimondo, M., Taylor, W. R., Yab, T. C., Mahoney, D. W., Sun, Z., et al. (2015). New DNA methylation markers for pancreatic cancer: discovery, tissue validation, and pilot testing in pancreatic juice. *Clin. Cancer Res.* 21, 4473–4481. doi: 10.1158/1078-0432.CCR-14-2469

Koprowski, H., Herlyn, M., Steplewski, Z., and Sears, H. F. (1981). Specific antigen in serum of patients with colon carcinoma. *Science* 212, 53–55. doi: 10.1126/science.6163212

Koprowski, H., Steplewski, Z., Mitchell, K., Herlyn, M., Herlyn, D., and Fuhrer, P. (1979). Colorectal carcinoma antigens detected by hybridoma antibodies. *Somatic Cell Genet.* 5, 957–971. doi: 10.1007/bf01542654

Kosanam, H., Prassas, I., Chrystoja, C. C., Soleas, I., Chan, A., Dimitromanolakis, A., et al. (2013). Laminin, gamma 2 (LAMC2): A promising new putative pancreatic cancer biomarker identified by proteomic analysis of pancreatic adenocarcinoma tissues. *Mol. Cell. Proteomics* 12, 2820–2832. doi: 10.1074/mcp.M112.023507

Law, C. W., Chen, Y., Shi, W., and Smyth, G. K. (2014). voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* 15:R29. doi: 10.1186/gb-2014-15-2-r29

Law, C. W. M. (2013). *Precision Weights for Gene Expression Analysis*. Available online at: https://minerva-access.unimelb.edu.au/handle/11343/38150 (accessed October 11, 2017).

López-Casas, P. P., and López-Fernández, L. A. (2010). Gene-expression profiling in pancreatic cancer. *Expert Rev. Mol. Diagn.* 10, 591–601. doi: 10.1586/erm.10.43

Makawita, S., Smith, C., Batruch, I., Zhengţ, Y., Rü, F., Grü, R., et al. (2011). Integrated proteomic profiling of cell line conditioned media, and pancreatic juice for the identification of pancreatic cancer biomarkers. *Mol Cell Proteomics* 10:M111.008599.

McKinney, K. Q., Lee, Y. Y., Choi, H. S., Groseclose, G., Iannitti, D. A., Martinie, J. B., et al. (2011). Discovery of putative pancreatic cancer biomarkers using subcellular proteomics. *J. Proteomics* 74, 79–88. doi: 10.1016/j.jprot.2010.08.006

Mellby, L. D., Nyberg, A. P., Johansen, J. S., Wingren, C., Nordestgaard, B. G., Bojesen, S. E., et al. (2018). Serum biomarker signature-based liquid biopsy for diagnosis of early-stage pancreatic cancer. *J. Clin. Oncol.* 36, 2887–2894. doi: 10.1200/JCO.2017.77.6658

Munding, J. B., Adai, A. T., Maghnouj, A., Urbanik, A., Zöllner, H., Liffers, S. T., et al. (2012). Global microRNA expression profiling of microdissected tissues identifies miR-135b as a novel biomarker for pancreatic ductal adenocarcinoma. *Int. J. Cancer* 131, E86–E95. doi: 10.1002/ijc.26466

Ouyang, H., Gore, J., Deitz, S., and Korc, M. (2014). microRNA-10b enhances pancreatic cancer cell invasion by suppressing TIP30 expression and promoting EGF and TGF-β actions. *Oncogene* 33, 4664–4674. doi: 10.1038/onc.2013.405

Peng, J., Sun, B.-F., Chen, C.-Y., Zhou, J.-Y., Chen, Y.-S., Chen, H., et al. (2019). Single-cell RNA-seq highlights intra-tumoral heterogeneity and malignant progression in pancreatic ductal adenocarcinoma. *Cell Res.* 29, 725–738. doi: 10.1038/s41422-019-0195-y

Peran, I., Madhavan, S., Byers, S. W., and McCoy, M. D. (2018). Curation of the pancreatic ductal adenocarcinoma subset of the cancer genome atlas is essential for accurate conclusions about survival-related molecular mechanisms. *Clin. Cancer Res.* 24, 3813–3819. doi: 10.1158/1078-0432.CCR-18-0290

Ramasamy, A., Mondry, A., Holmes, C. C., and Altman, D. G. (2008). Key issues in conducting a meta-analysis of gene expression microarray datasets. *PLoS Med.* 5:e184. doi: 10.1371/journal.pmed.0050184

Ranganathan, P., Harsha, H. C., and Pandey, A. (2009). Molecular alterations in exocrine neoplasms of the pancreas. *Arch. Pathol. Lab. Med.* 133, 405–412. doi: 10.1043/1543-2165-133.3.405

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

Ruckdeschel, K., Mannel, O., and Schröttner, P. (2002). Divergence of apoptosis-inducing and preventing signals in bacteria-faced macrophages through myeloid differentiation factor 88 and IL-1 receptor-associated kinase members. *J. Immunol.* 168, 4601–4611. doi: 10.4049/jimmunol.168.9.4601

Rückert, F., Dawelbait, G., Winter, C., Hartmann, A., Denz, A., Ammerpohl, O., et al. (2010). Examination of apoptosis signaling in pancreatic cancer by computational signal transduction analysis. *PLoS One* 5:e12243. doi: 10.1371/journal.pone.0012243

Salem, H. H., Trojanowski, B., Fiedler, K., Maier, H. J., Schirmbeck, R., Wagner, M., et al. (2014). Long-term IKK2/NF- B signaling in pancreatic -cells induces immune-mediated diabetes. *Diabetes* 63, 960–975. doi: 10.2337/db13-1037

Schneider, J., and Schulze, G. (2003). Comparison of tumor M2-pyruvate kinase (tumor M2-PK), carcinoembryonic antigen (CEA), carbohydrate antigens CA 19-9 and CA 72-4 in the diagnosis of gastrointestinal cancer. *Anticancer Res.* 23, 5089–5093.

Shi, W.-D., Zhi, Q. M., Chen, Z., Lin, J.-H., Zhou, Z.-H., and Liu, L.-M. (2009). Identification of liver metastasis-related genes in a novel human pancreatic carcinoma cell model by microarray analysis. *Cancer Lett.* 283, 84–91. doi: 10.1016/J.CANLET.2009.03.030

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., et al. (2019). Comprehensive integration of single-cell data. *Cell* 177, 1888.e21–1902.e21. doi: 10.1016/j.cell.2019.05.031

Uhlen, M., Fagerberg, L., Hallstrom, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., et al. (2015). Tissue-based map of the human proteome. *Science* 347:1260419. doi: 10.1126/science.1260419

Van den Broeck, A., Vankelecom, H., Van Eijsden, R., Govaere, O., and Topal, B. (2012). Molecular markers associated with outcome and metastasis in human pancreatic cancer. *J. Exp. Clin. Cancer Res.* 31:68. doi: 10.1186/1756-9966-31-68

Wang, J., Coombes, K. R., Highsmith, W. E., Keating, M. J., and Abruzzo, L. V. (2004). Differences in gene expression between B-cell chronic lymphocytic leukemia and normal B cells: a meta-analysis of three microarray studies. *Bioinformatics* 20, 3166–3178. doi: 10.1093/bioinformatics/bth381

Wang, W. S., Liu, X. H., Liu, L. X., Lou, W. H., Jin, D. Y., Yang, P. Y., et al. (2013). ITRAQ-based quantitative proteomics reveals myoferlin as a novel prognostic predictor in pancreatic adenocarcinoma. *J. Proteomics* 91, 453–465. doi: 10.1016/j.jprot.2013.06.032

Washiro, M., Ohtsuka, M., Kimura, F., Shimizu, H., Yoshidome, H., Sugimoto, T., et al. (2008). Upregulation of topoisomerase IIα expression in advanced gallbladder carcinoma: a potential chemotherapeutic target. *J. Cancer Res. Clin. Oncol.* 134, 793–801. doi: 10.1007/s00432-007-0348-0

Wilson, C. L., and Miller, C. J. (2005). Simpleaffy: a bioconductor package for affymetrix quality control and data analysis. *Bioinformatics* 21, 3683–3685. doi: 10.1093/bioinformatics/bti605

Xu, Z.-Y., Chen, J.-S., and Shu, Y.-Q. (2010). Gene expression profile towards the prediction of patient survival of gastric cancer. *Biomed. Pharmacother.* 64, 133–139. doi: 10.1016/J.BIOPHA.2009.06.021

Yoshida, Y., Kumar, A., Koyama, Y., Peng, H., Arman, A., Boch, J. A., et al. (2004). Interleukin 1 activates STAT3/nuclear factor-kappaB cross-talk via a unique TRAF6- and p65-dependent mechanism. *J. Biol. Chem.* 279, 1768–1776. doi: 10.1074/jbc.M311498200

Conflict of Interest: BIDMC will be filling patent on behalf of MB and IK on the use of biomarker panel for early PDAC diagnosis. MB is an equity holder at BiomaRx and Canomiks.