

Cell free RNA detection of pancreatic cancer in pre diagnostic high risk and symptomatic patients

Received: 18 July 2024

Accepted: 24 July 2025

Published online: 09 August 2025

 Check for updates

Travis W. Moore ^{1,11}, Elias Spiliotopoulos ^{1,11}, Rowan L. Callahan ^{1,2,11}, C. Ward Kirschbaum¹, Conner F. Bailey¹, Hyun Ji Kim ^{1,2}, Breeshey Roskams-Hieter ^{1,3,4}, Florian Goncalves^{1,2}, Dove Keith ⁵, Aaron J. Grossberg ^{1,5,6}, Paul T. Spellman ^{1,7}, Gordon B. Mills ⁸, Rosalie C. Sears ^{5,9}, Terry K. Morgan^{1,2,10} & Thuy T. M. Ngo ^{1,2,8,9} 

Pancreatic ductal adenocarcinomas (PDAC) are among the most fatal cancers, in part due to frequent detection at advanced stages. Endoscopic ultrasound-guided fine-needle aspiration (EUS-FNA), the most sensitive diagnostic method of PDAC in current standard clinical practice, is invasive, costly, with access limited to major healthcare settings. Here, we present a non-invasive evaluation of plasma cell-free RNA (cfRNA) for PDAC detection in pre-diagnostic high-risk and de novo symptomatic patients presenting for EUS-FNA. We develop a cfRNA normalization method to account for preanalytical variation and handling effects and derive 29 potential cfRNA biomarkers for PDAC diagnosis using 153 samples collected prior to the EUS procedure. Biomarkers related to liver function are elevated in PDAC samples, including early-stage patients without liver metastasis. Classification of PDAC using these biomarkers is validated using an independent cohort of 95 samples. Our findings could help to improve diagnostic utility in high-risk and symptomatic individuals.

Pancreatic Ductal Adenocarcinoma (PDAC) is the most common form of pancreatic cancer, and is one of the most lethal malignancies, leading to the 4th most common cancer deaths worldwide¹. Due to a lack of early-stage symptoms, most patients are not diagnosed until the later stages of the disease, where the outcome is very poor². Given the lack of early symptoms and low incidence of PDAC, the United States Preventive Services Task Force does not recommend screening in asymptomatic patients³. Despite the challenges of diagnosing PDAC

at early stages, clinicians frequently opt to use the FDA-approved serum blood test for the Carbohydrate Antigen 19-9 (CA19-9), but it, unfortunately, lacks the necessary sensitivity and positive predictive value (PPV) for diagnostic use⁴. Thus, there remains a need for accurate, reproducible biomarkers that can aid early diagnosis of PDAC in high-risk and clinically indicated populations^{5,6}.

Endoscopic ultrasound-guided (EUS) fine-needle aspiration (FNA) is currently among the most sensitive method in the diagnosis and

¹Cancer Early Detection Advanced Research Center (CEDAR), Knight Cancer Institute, Oregon Health and Science University, Portland, OR, USA. ²Department of Biomedical Engineering, Oregon Health and Science University, Portland, OR, USA. ³Institute of Genetics and Cancer, MRC Human Genetics Unit, Western General Hospital, University of Edinburgh, Edinburgh, UK. ⁴Health Data Research UK, London, UK. ⁵Brenden-Colson Center for Pancreatic Care, Oregon Health and Science University, Portland, OR, USA. ⁶Department of Radiation Medicine, Oregon Health and Science University, Portland, OR, USA. ⁷Department of Medicine and Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, CA, USA. ⁸Division of Oncological Sciences, Knight Cancer Institute, Oregon Health and Science University, Portland, OR, USA. ⁹Department of Molecular and Medical Genetics, Oregon Health and Science University, Portland, OR, USA. ¹⁰Department of Pathology, Oregon Health and Science University, Portland, OR, USA. ¹¹These authors contributed equally: Travis W. Moore, Elias Spiliotopoulos, Rowan L. Callahan. ✉e-mail: ngth@ohsu.edu

management of PDAC^{7–9}. EUS is a nonoperative approach for the detection of pancreatic lesions that warrant further investigation; typical applications are in high-risk patients with a family history of PDAC or hereditary cancer syndromes, as well as symptomatic patients with clinical presentation suspicious for PDAC where findings from other imaging modalities such as computed tomography and magnetic resonance imaging are inconclusive. Additionally, EUS can identify small tumors that are not evident by other imaging modalities⁸. Despite its advantages, EUS-FNA is an expensive and invasive procedure that requires highly skilled operators and is not readily available outside major healthcare centers. Furthermore, in many cases, it may not provide a definite diagnosis. For example, EUS may fail to identify a true pancreatic mass in patients with chronic pancreatitis, a diffusely infiltrating carcinoma, a prominent ventral/dorsal split, or a recent episode of acute pancreatitis⁷. For these reasons, there is an unmet need to have a noninvasive molecular measurement using easily accessible specimens to identify individuals where the EUS-FNA procedure is indicated and improve its diagnostic utility.

Liquid biopsy, the analysis of analytes in easily accessible biofluids, represents an emerging approach for disease identification and monitoring across many disease types^{10,11}. For example, extracellular vesicle surface markers and their cargos, epigenetic profiling of cell-free DNA, and micro and long cRNA all have been shown to discriminate disease states^{12–15}. We and others have previously reported the efficacy of blood-based cRNA biomarkers in obstetrics, transplantation, neurodegeneration, chronic diseases, and oncology^{16–23}. However, prior liquid biopsy studies in PDAC were performed with cross-sectional case-control designs using blood samples collected from PDAC patients with confirmed clinical diagnoses^{12–14,24–33}. The utility of these findings in a pre-diagnostic clinical setting³⁴ remains to be confirmed. Furthermore, to our knowledge, no existing liquid biopsy method can distinguish PDAC from intermediary pathologic conditions that complicate the diagnosis of PDAC, such as biliary stricture, benign pancreatic cysts, pancreatitis, and intraductal papillary mucinous neoplasm (IPMN) in patients with a clinical presentation suspicious of high-risk for PDAC.

In this study, we sequence the cell-free plasma transcriptome of a group of high-risk and clinically indicated patients who had been referred for an EUS-FNA procedure. This pre-diagnosis sampling approach ensures a representative subset of the high-risk patient pool who would most benefit from a pre-diagnostic screening test to determine whether an invasive EUS-FNA is justified. We identify a set of 29 biomarkers predictive of PDAC vs benign and intermediate pathologies of the pancreas and use these biomarkers to train machine learning classifiers to predict PDAC status from sequenced plasma. Our classifier maintains high performance on an external validation set of high-risk patients presenting for EUS, with AUC of 0.896. These results indicate the potential of cf-mRNA biomarkers as an accessible and noninvasive alternative for the prediction and monitoring of PDAC.

Results

Study design and patient characteristics

We recruited patients with clinical indications being referred by physicians as a standard of care procedure to the EUS clinic at Oregon Health and Science University (OHSU). EUS indications included patients previously identified as high risk for pancreatic cancer due to genetic background, family history, physical symptoms associated with a pancreatic disease, pancreatitis, pancreatic cyst, and indefinite pancreatic mass or gastric mass shown by other imaging modalities (Supplementary Data Table 1). Blood samples were drawn before EUS screening, obviating the release of tissue material and immune reactions due to EUS-FNA potentially influencing results. Drawing blood before the procedure also controlled for lifestyle changes, as well as anxiety and stress that can change serum markers due to a cancer diagnosis. Abnormal imaging observations during the EUS procedure

resulted in a FNA to collect specimens for pathology review. Final diagnosis categories included benign pancreas, acute pancreatitis, chronic pancreatitis, IPMN, islet cell tumor, PDAC, or other cancer (Fig. 1). The benign pancreas group included patients with various diseases and benign conditions such as no mass lesion detected, bile duct stones, pancreatic cysts, benign reactive changes, and others.

Two separate cohorts of samples were used for the study (Table 1): 153 samples were collected by Cancer Early Detection Advanced Research Center (CEDAR) between 2019 and 2021 (CEDAR cohort to be used for discovery and model training) and 95 samples were collected by Brenden–Colson Center for Pancreatic Care (BCC) between 2017 and 2019 (BCC cohort to be used for independent validation) at OHSU. The CEDAR discovery cohort included 48 benign pancreas, 59 acute or chronic pancreatitis, 16 IPMN, 3 islet cell tumor, 20 PDAC, and 7 other cancers. The independent BCC validation cohort was composed of 23 benign pancreas, 23 pancreatitis, 11 IPMN, 6 islet cell tumors, 21 PDAC, and 11 other cancers. There were no statistically significant differences between the compositions of the sample sets (Table 1). In the PDAC groups, there were more males than females and the age range was higher than patients with other diagnoses ($P < 0.001$, fisher exact test).

Identification of intrinsic plasma cRNA factors and developing cRNA normalization method

We and others previously found that the composition of extracellular RNA in plasma, including cell-free coding and noncoding RNA, depends on sample handling and processing conditions^{35–39}. Platelet activation and lysis, in particular, is a major source of ex vivo release of RNA into the cRNA compartment³⁵. Therefore, we sought to develop a cell-free RNA-specific method for normalization as an internal control for technical variability. We hypothesize that the level of each gene in a cRNA plasma sample is a combination of the default levels present in the bloodstream at the time of sample collection (the intrinsic factor), and that released from platelets and other blood cellular components due to collection and handling procedures (extrinsic factor). Some genes, such as those expressed by platelets, will be heavily influenced by the extrinsic factor and show a greater variance across sample handling conditions, while others will be more dependent on the intrinsic factor and thus more stable in their apparent level. To correct for handling variations from influencing sequencing read depth normalization, we created a cRNA normalization (cf-normalization) method to balance the expected intrinsic and extrinsic gene contributions in each sample.

To inform the creation of the normalization procedure, we generated cRNA sequencing data for plasma samples from ten healthy individuals. Each blood draw was processed with seven different processing conditions by varying tube types, storage time, and centrifuge settings with single or double spins (Supplementary Table S1 and Supplementary Fig. S1). We then fitted the resulting count table to a nonnegative matrix factorization (NMF) model to identify the intrinsic and extrinsic contributions from this data set (Fig. 2a). NMF allows us to decompose the gene by sample matrix into a composition of factors across those genes, in this case representing the cRNA amount of each sample as a weighted mixture of the true (intrinsic) signal and the processing (extrinsic) based signal. As expected, unsupervised NMF modeling resulted in a high extrinsic coefficient for platelet genes (Supplementary Table S2). We use these factors learned in healthy plasma data to decompose our training and validation data into the expected intrinsic and extrinsic contributions from each gene in each sample. The intrinsic and extrinsic count tables are then normalized for total sample counts separately, before being recombined into the normalized cRNA (cf-normalized) count table. This not only normalizes the library size for each sample, but also sets the predicted total contribution to library size by intrinsic and extrinsic contributions to be the same for each sample (Fig. 2b). We include an additional normalization using the trimmed mean of M values (TMM) to further

reduce the effects of gene specific read depth variation from sequencing⁴⁰.

Identification and biological significance of cfRNA biomarkers distinguishing PDAC from intermediary pathologic conditions

We isolated and sequenced cfRNA from plasma specimens to identify potential cfRNA biomarkers and build classification models. Both the discovery CEDAR cohort and the validation BCC cohort were sequenced to saturation (Supplementary Fig. S2). After selecting reads that mapped uniquely to the human genome and deduplication, the cfRNA libraries had an average read depth of 1,069,874. A total of 40,279 annotated features were detected with at least 1 mapped read across all samples. The majority of detected RNAs were protein-coding with a mean fraction of 91.2% (Supplementary Table S3 and Supplementary Fig. S3).

After normalization using the method detailed above, we used the CEDAR cohort for biomarker discovery. We used a Wilcoxon rank sum test to identify differentially expressed (DE) genes in PDAC samples, scored by False Discovery Rate (FDR) corrected *P* value. A cutoff of 0.05 produced 29 cfRNA biomarker candidates that were DE in PDAC compared to all other diagnoses (Supplementary Figs. S4–6).

Comparing the relative expression across tissue types in the Human Protein Atlas, we found that our cfRNA biomarker candidates were highly expressed in the liver, lymph node, pancreas, and other immune-related tissues such as bone marrow, tonsil, and spleen (Fig. 3a). Of note, the most common sites of metastasis for PDAC patients in the CEDAR dataset were the liver and lymph nodes (6 and 4

samples, respectively, out of 20). Liver expression, while also being the highest overall, contained a unique combination of the DE genes without enrichment in other highly expressed tissues, notably *SERPINI1*, *ACOT12*, *CRP*, and *AFMID*. All four of these genes were upregulated in PDAC samples within both the discovery cohort and the validation cohort (Supplementary Figs. S5 and 6). Conversely, the genes associated with the top immune-related tissues (*RPS4X*, *RPS13*, *RPS14*, *CD37*, *MS4A1*, *TNFRSF13C*, *PAX5*, *FAM117A*) were all downregulated in PDAC samples for both datasets. This corresponds with previous observations that PDAC has very low levels of infiltrating immune cells^{41–43}.

To further dissect the sources of the potential PDAC cfRNA biomarkers, we performed deconvolution of the DE genes using nu-support vector regression, with a signature matrix constructed from the Human Protein Atlas tissue RNA values. This deconvolution estimates the DE gene level of each sample as a combination of tissue types from the atlas, identifying which tissues are the most likely sources of the observed gene level. Samples from both data sets were deconvolved over the 256 tissue types in the atlas. Based on the average relative tissue proportions predicted by deconvolution, we saw that the liver had over 3 times the contribution in PDAC patients as in all other samples, consistently marking it as a likely contributor of potential up-regulated biomarkers (Fig. 3b). A similar analysis on a subset of PDAC patients at early stages, and on PDAC patients who did not have clinically detected liver metastasis showed a similarly high presence of liver contribution to cfRNA composition (Fig. 3b). Deconvolution of the two cohorts sepa-

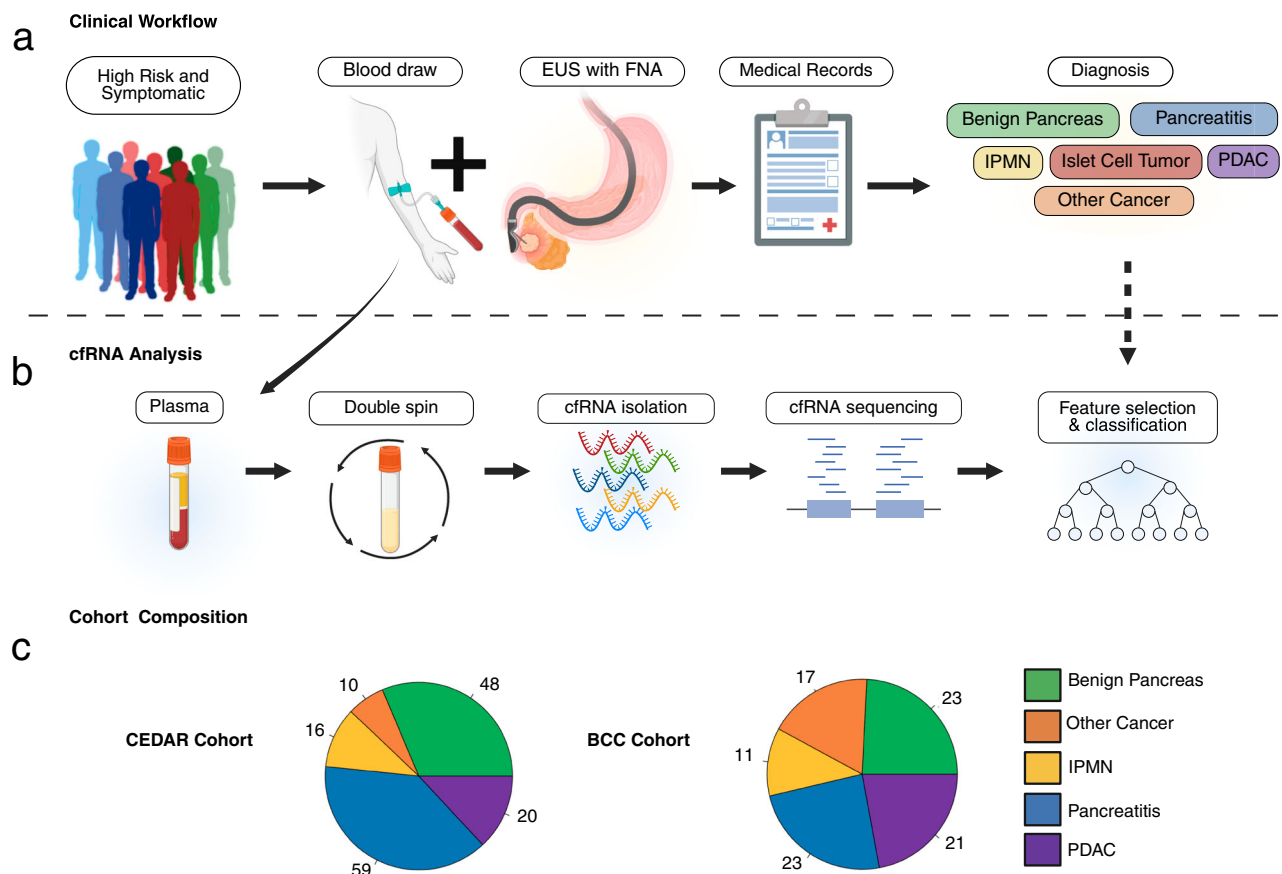


Fig. 1 | Overview of the study design. a High-risk and symptomatic patients who were referred to an EUS procedure were recruited. Blood draws were performed before the procedure. The final annotation by a board-certified physician on the team, based on a full charge review after more than 1 year from blood collection, categorizes patients into 5 groups: benign pancreas, pancreatitis, IPMN, PDAC, and

other cancers. **b** cfRNA sequencing preparations were performed with randomized samples in a blinded manner at the time of sample processing. **c** Patient composition of the CEDAR (discovery) and BCC (validation) cohorts. Figure created using BioRender (<https://biorender.com>).

Table 1 | Patients’ characteristics

CEDAR discovery and training cohort (n = 153)						BCC validation cohort (n = 95)				
	Benign pancreas	Pancreatitis	IPMN	PDAC	Cancer other	Benign pancreas	Pancreatitis	IPMN	PDAC	Cancer other
N	48	59	16	20	10	23	23	11	21	17
Sex $p < 0.001$										
Female	23	22	11	1	7	15	6	8	4	6
Male	25	37	5	19	3	8	17	3	17	11
Race $p = 0.275$										
White, non hispanic	45	52	16	18	6	22	22	11	20	17
Non white OR hispanic	3	7	0	2	4	1	1	0	1	0
Age, years $p < 0.001$										
Median	57.5	58	67	70	55.5	58	55	67	72	75
Range	25–82	25–77	32–83	50–83	27–73	30–81	34–78	42–83	58–85	39–86
PDAC stage										
1				5					4	
2				2					3	
3				5					9	
4				8					5	

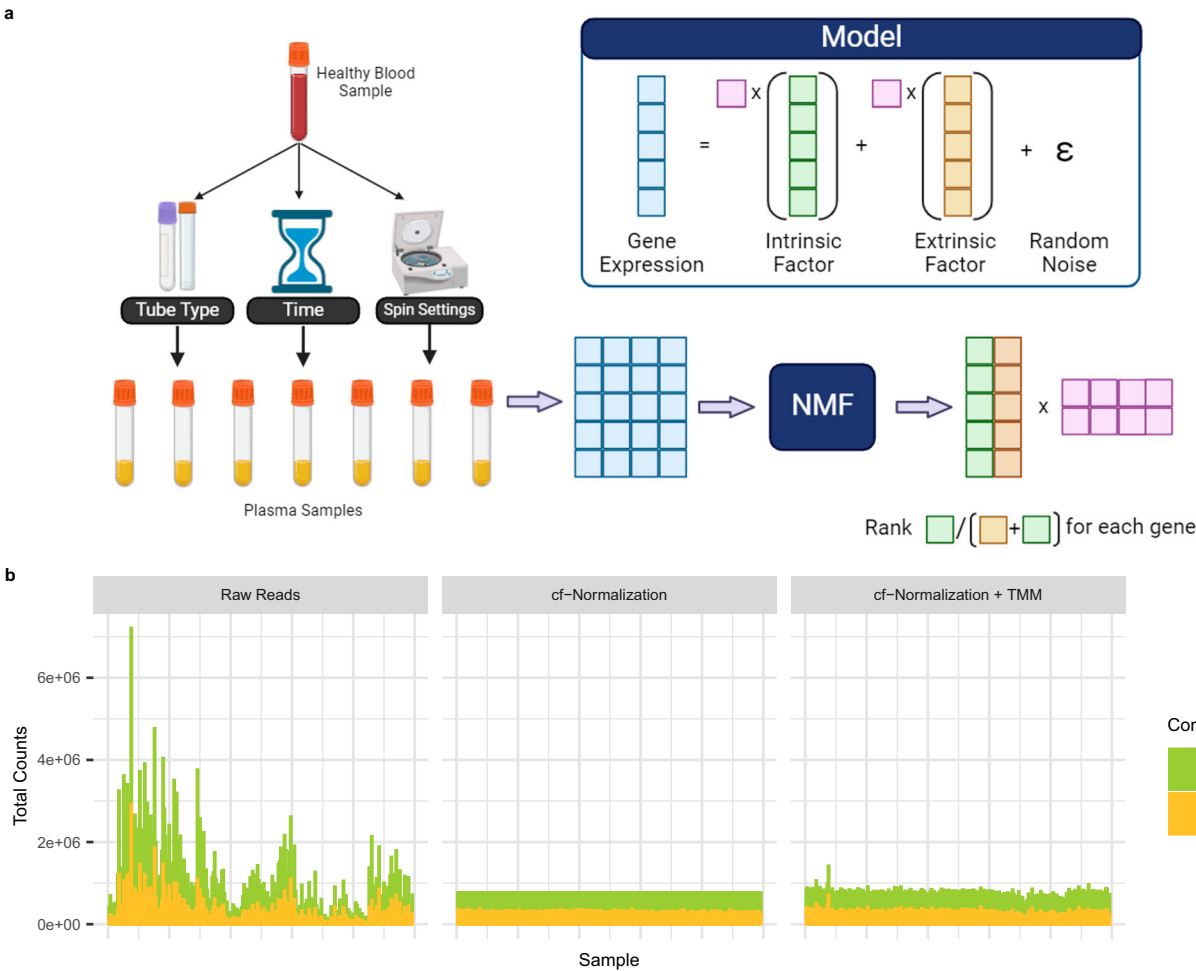


Fig. 2 | Identification of intrinsic and extrinsic cfRNA factors using nonnegative matrix factorization. **a** Representation of the identification of gene factors and underlying gene count model. A cfRNA dataset of 70 samples from 10 healthy donors was collected, with samples from each donor undergoing seven different handling conditions involving tube type, storage time, and centrifuge setting variations before plasma extraction and sequencing. Nonnegative matrix factorization was performed on the gene count matrix of the dataset to identify the intrinsic and

extrinsic factors for each gene. **b** Expected intrinsic and extrinsic contributions of total gene counts for CEDAR and BCC samples, based on NMF factors. cfRNA normalization balances these contributions across samples (middle panel), while TMM normalization (right panel) adjusts for sample and gene outliers. **a** created using BioRender (<https://biorender.com>).

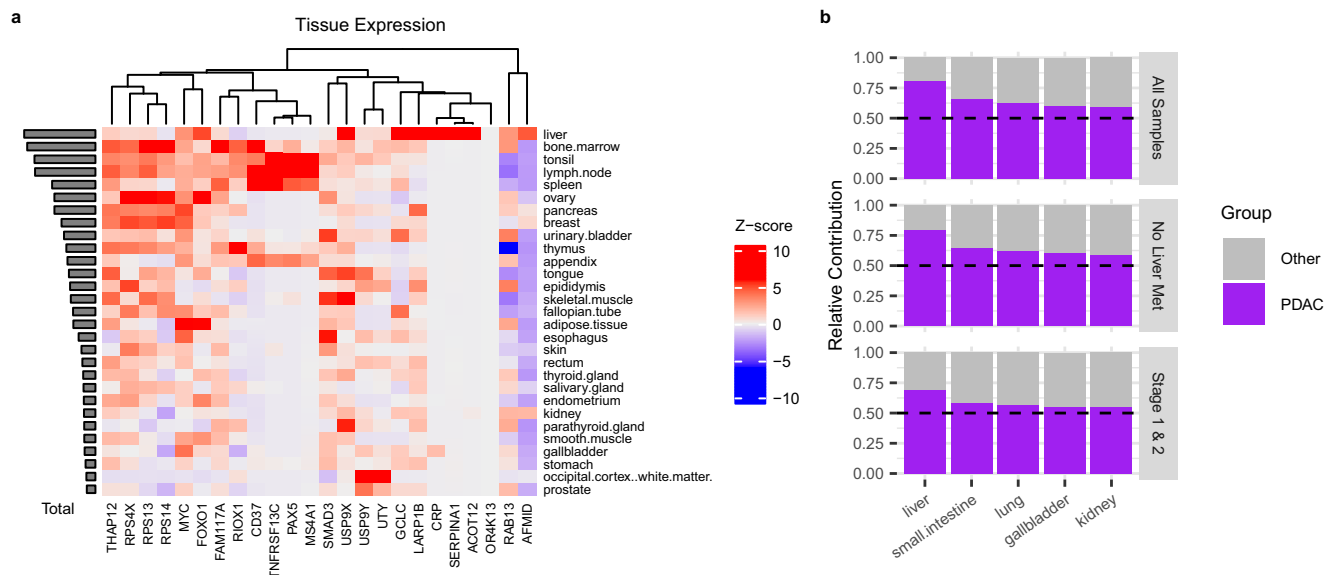


Fig. 3 | Deconvolution of tissue contribution to PDAC cfRNA biomarker levels in plasma. **a** Z-score normalized expression for DE genes in top tissues, sorted by total count. Average tissue RNA count values were taken from the Human Protein Atlas version 22.0. DE genes not present in tissue atlas data are omitted (5 total). **b** Relative comparison of tissue proportions predicted from nu-SVR deconvolution of DE genes in PDAC and all other samples. Tissue proportions from deconvolution

are averaged over PDAC and other groups and normalized to sum to 1. Three deconvolutions were performed: using all PDAC samples, PDAC samples without liver metastasis, and stage 1 and 2 PDAC samples. Relative contributions above the 0.50 line indicate elevated tissue proportions in PDAC samples compared to all other diagnoses.

rately produced largely the same results, except for a lower liver signature in the seven early stage PDAC patients in the BCC cohort (Supplementary Fig. S7). This observation indicates that the liver may be involved early during PDAC carcinogenesis and progression, potentially by circulating factors from the PDAC^{44–46}.

Distinguishing PDAC from benign and intermediary pathologies of the pancreas and other cancers by cfRNA PDAC score

We trained a machine learning model on the normalized CEDAR data set to classify patients with PDAC diagnosis. The workflow of our machine learning pipeline and tuning results are detailed in the supplemental material (Supplementary Fig. S8). The model was trained only on the discovered DE gene set and our classifier was tasked with identifying a binary classification of “PDAC” and “Not PDAC”. To address the large imbalance between these two classes in the data, we used the SMOTE interpolation method⁴⁷ to create 100 synthetic PDAC samples for training. We then used 10-fold cross-validation of the CEDAR cohort to select the best-performing classifier, leaving 1-fold (10% of samples) out for validation each iteration while training on the other 9. Random forest (RF) demonstrated better results than support vector machine (SVM) and linear discriminant analysis (LDA) (Supplementary Table S4). A second round of cross-validation was then used to tune the parameters of the RF classifier, with the best parameter settings used to evaluate the performance in the training and independent validation data (Supplementary Fig. S9).

Next, we used the trained classifier to predict the probability of a sample having a diagnosis of PDAC, which we called the PDAC score. Evaluation of performance in the training data was assessed using 10-fold cross-validation, allowing each sample to have a PDAC score assigned while in the validation set. This setup was repeated for three different experiments. In the first experiment, we only included samples with PDAC and benign diagnosis. The second experiment included pancreatitis and IPMN samples, while the third used all available samples. We also assessed cross-validation performance on different sample populations for each experiment, separating performance based on sex and age. For age comparison, we evaluated a binary split

of patients younger or older than 66. This cutoff value provided the optimal age balance between PDAC and non-PDAC diagnoses, ensuring each age split contained at least 30% of the samples for both categories. We computed the PDAC score for each sample in these three experiments, which was used to calculate the ROC curve for predicting PDAC vs. Not PDAC (Fig. 4).

In each setting, the classifier was able to learn an accurate differentiation of PDAC samples from all others, even when intermediate pancreas pathologies and other cancer types were included. Differentiating PDAC from benign, other intermediary pathologies of the pancreas, and all other diagnoses yielded an area under the curve (AUC) of 0.945, 0.965, and 0.956, respectively, without notable bias towards age or sex. These results demonstrate the predictive potential of cfRNA biomarkers in this setting and warrant further investigation into their performance on other data sets.

Validation of PDAC cfRNA biomarkers in an independent cohort

To validate the performance of our cfRNA biomarker set, we used the BCC cohort, which was independently collected with a similar patient recruitment and specimen collection workflow as the CEDAR cohort. Using the PDAC classifier, which was trained separately on the entire CEDAR cohort, we evaluated three model settings of PDAC vs. Benign, PDAC vs. non-cancer samples, and PDAC vs all other diagnoses, filtering samples in each dataset to match the model groups. Both training and validation cohorts were normalized using our cf-normalization procedure and filtered to the DE gene set identified in the CEDAR cohort. Consistent with the previous finding, PDAC samples strongly separated from other diagnoses in all 3 model settings for the independent BCC validation cohort (Fig. 5). The cfRNA classifier was validated in the BCC cohort with an AUC of 0.904, 0.937, and 0.896 when differentiating PDAC from benign, other intermediary pathologies of the pancreas, and all other diagnoses, respectively. This performance was largely stable when splitting the samples by sex and age (Fig. 5b, c).

The consistently strong classifier performance on the external validation cohort is due in large part to our normalization procedure

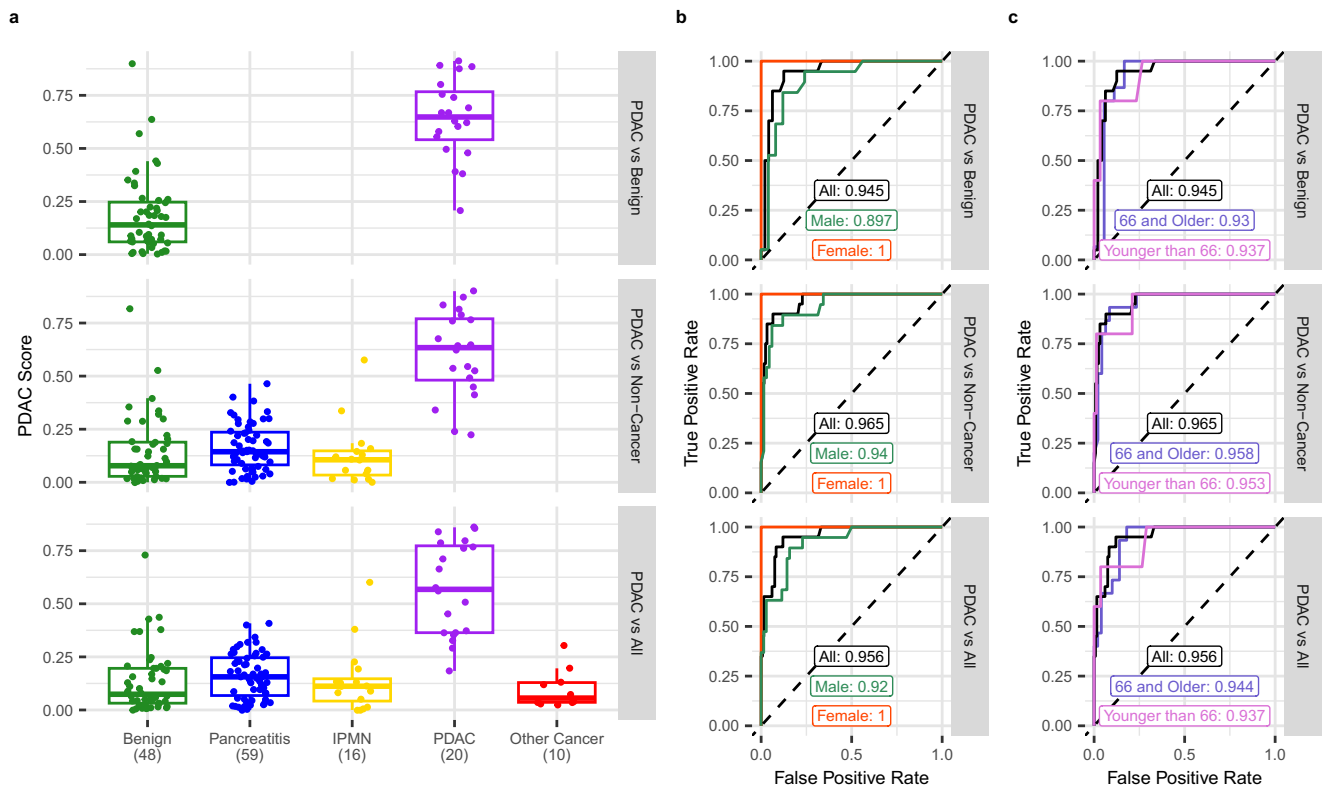


Fig. 4 | Classification model differentiating PDAC from other diagnoses cross-validated in CEDAR cohort. **a** PDAC score assigned to each sample in cross-validation, separated by diagnosis group (sample size noted underneath each group). Each row is a different experiment, with the first comparing only PDAC and Benign, PDAC and intermediate pathologies in the second row, and all conditions in the

bottom row. Boxplots display the 25th, 50th, and 75th percentiles of groups, with whiskers extending up to 1.5 times beyond inter-quartile range. **b, c** ROC plots based on PDAC score, and associated AUC values, for each analysis condition. Additional ROC plots are shown for stratification of patients by sex (**b**) and age (**c**).

and adding synthetic SMOTE samples (Supplementary Figs. S10 and 11). Replicating this experimental workflow using standard TPM normalization resulted in lower training set performance (0.916 AUC) and a sharp drop off in predictive power when applied to the validation cohort (0.682 AUC) (Supplementary Figs. S12 and 13), as did using TMM normalization without our cf-normalization procedure (Supplementary Figs. S14 and 15). Our classification results not only indicate the potential power of cfRNA biomarkers as a tool to aid in diagnosis but also the importance of adjusting for processing effects in cell-free samples to enhance their predictive power. In summary, these results confirm the diagnostic potential of 29 cfRNA biomarkers with cf-normalization to detect PDAC in clinically indicated patients suspicious of PDAC before referral to the EUS-FNA procedure.

We additionally performed the same classification analysis after removing PDAC samples with liver metastasis from both the training and validation cohorts (Supplementary Figs. S16 and 17). In this setting, our cfRNA biomarkers were able to accurately separate PDAC without liver metastasis from other diagnoses with an AUC of 0.840 or better. Even though many of our biomarkers show significant enrichment in genes associated with the liver, they can still be an accurate identification of PDAC for patients without clinically diagnosed liver metastasis.

The RNA sequencing reads in both cohorts showed a variability of exon fraction. We performed an additional version of our classification pipeline after filtering any samples with an exon fraction lower than 0.6 (Supplementary Fig. S18 and Supplementary Table S5). While this analysis version with an exon filter produced a larger number of DE genes, 41 compared to the original 29, the majority of these (25) were the same, and there was no significant change in the classification AUC. For this reason, we do not filter samples by exon fraction in the remainder of the analysis.

cfRNA classifier outperformed the CA19-9 biomarker

Clinical measurements of serum CA19-9 levels were available for 126 of 153 samples in the CEDAR set and 86 of 95 samples in the BCC set. We assessed the accuracy of our cfRNA PDAC classifier compared to CA19-9 by itself, and with CA19-9 included as an additional covariate in the model. We show here the CA19-9 levels by diagnosis in each cohort, with a dashed line indicating the clinical cutoff value of 37 (Fig. 6a, b). Clinical measurements of CA19-9 showed an encouraging negative predictive value (NPV) of 0.942 and 0.935 but a poor PPV of 0.364 and 0.417 in CEDAR and BCC sample cohorts, respectively. By comparison, our predicted PDAC score had PPV values of 0.857 and 0.531 on the CEDAR and BCC cohorts, when NPV is set to the same levels as in CA19-9 (Supplementary Figs. S19 and 20). We additionally compared the performance of our classifier using feature sets of just CA19-9, our discovered cfRNA biomarker genes, and a combination of cfRNA biomarker genes with CA19-9. In this setting, we did not use the clinical cutoff for CA19-9, instead allowing the classifier to learn a cutoff from the training data. Our cfRNA classifier outperformed CA19-9 with both cross-validation in the CEDAR sample cohort and independent validation in the BCC cohort (Fig. 6c, d). This difference was statistically significant in experiments that included intermediate pathologies and other cancers ($P < 0.05$, Delong's AUC comparison test). The combination of CA19-9 with cfRNA did not improve the classification accuracy of cfRNA significantly (Fig. 6c, d). This may be due to several of the DE genes being correlated with CA19-9 levels in both cohorts (*SERPINA1*, *RAB13*, and *CRP*, Spearman correlations of 0.462, 0.363, and 0.306, respectively).

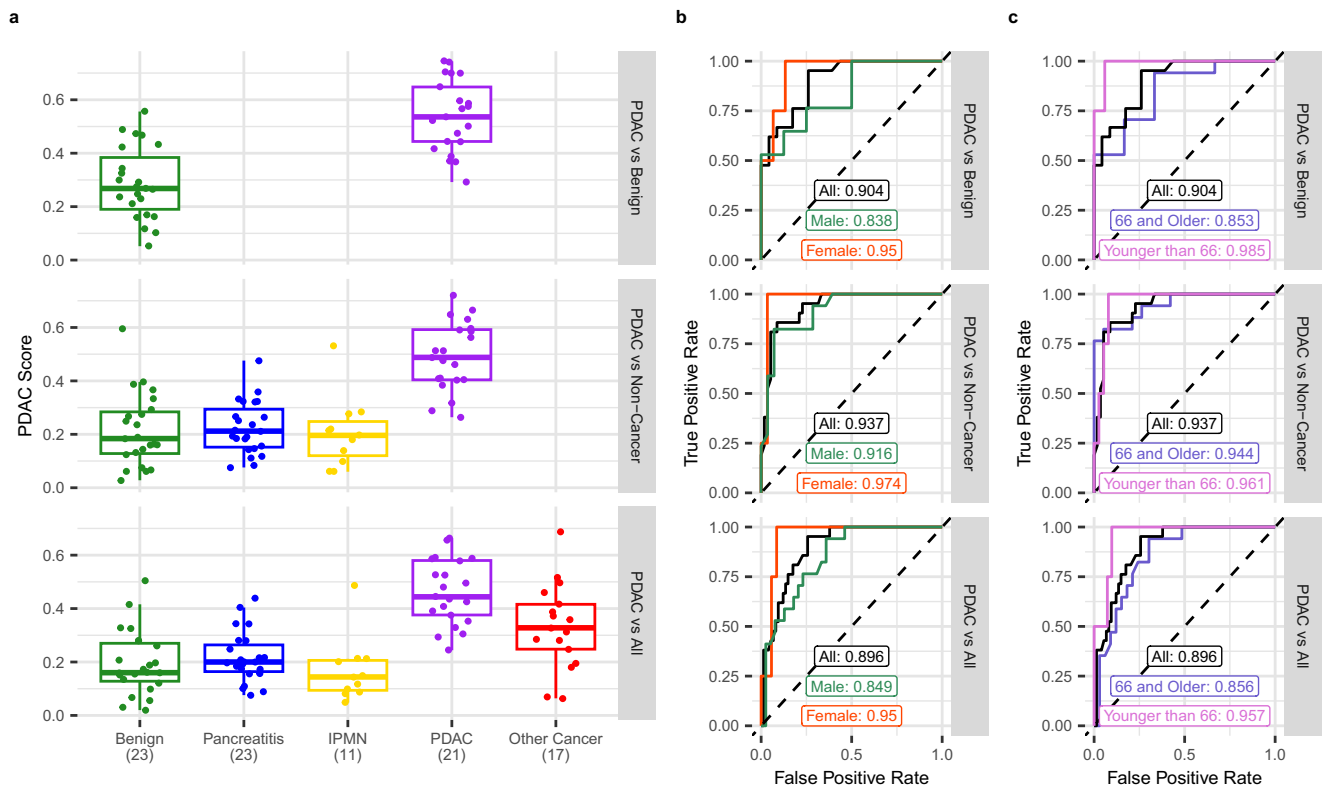


Fig. 5 | Independent validation of the PDAC cfRNA classifier in the separate BCC cohort. **a** PDAC score assigned to each sample in the validation set, separated by diagnosis group (sample size noted underneath each group). Each row is a different experiment, with the first comparing only PDAC and Benign, PDAC and intermediate pathologies in the second row, and all conditions in the bottom row.

Boxplots display the 25th, 50th, and 75th percentiles of groups, with whiskers extending up to 1.5 times beyond inter-quartile range. **b**, **c** ROC plots based on PDAC score, and associated AUC values, for each analysis condition. Additional ROC plots are shown for stratification of patients by sex (**b**) and age (**c**).

Stage dependence of cfRNA PDAC score and survival correlation

We obtained follow-up data for the PDAC patients in both cohorts, indicating their survival and disease state at the time of follow-up. Follow-up dates were inconsistent between patients and limited for some, but provided an indication of the short-term outcome for each patient.

We computed Kaplan-Meier curves for each dataset, using our predicted PDAC score to separate patients into high and low-risk groups. We used the 50th percentile of scores for PDAC patients in each dataset as the cutoff mark for these two groups, creating group sizes of 10 and 10 for CEDAR patients, and 10 and 11 for BCC patients. While the differences in these curves were not significant for either cohort, unsurprising given the small sample sizes, we did observe a notable correlation of PDAC score with survival in the CEDAR cohort, and a correlation of the PDAC score with disease stage in the BCC cohort (Supplementary Fig. S21).

We additionally computed Cox proportional hazard models for each analysis group, combining the two cohorts to improve power. In these models we used PDAC score (as computed in each analysis setting), sex, age, and PDAC stage as variables for the prediction of survival time in PDAC patients. Notably, in these models, we found that PDAC score is a stronger prediction of outcome than patient age (Supplementary Table S6), though disease stage is the most significant predictor.

Discussion

Liquid biopsy has emerged as a promising approach in cancer screening and diagnosis in the last decade⁴⁸. However, most previous studies into the potential of blood-based methods for the detection of PDAC used a case-control design with sample collection from patients after a confirmed diagnosis^{12–14,24–32}. The invasive diagnostic procedures with needle biopsy may induce the dissemination of potential

biomarkers into the blood, which may not be present or at low levels before diagnosis^{49,50}. In addition, lifestyle changes, as well as stress from the diagnosis of cancer in patients with a different confirmed final diagnosis, may alter the composition of blood analytes^{51,52}. This may partly explain the significant drop in the performance of blood tests in clinical trials compared to the discovery phase in previous studies. In this study, blood was drawn before the EUS screening in high-risk and clinically indicated patients whose diagnoses were not confirmed. This presents a realistic sampling approach where patients are more likely to have shared background conditions or reasons for undergoing screening. Further, it is the most appropriate patient set to identify patients where EUS is most likely to demonstrate utility.

To date, the validity of blood biomarkers in a clinical setting in which high-risk and de novo symptomatic patients are indicated for screening and definitive diagnosis of pancreatic diseases is under-developed. Previously established biomarkers, such as CA19-9, lack sufficient power to reliably distinguish cancers from intermediate pathologies^{53,54}. Our newly identified cfRNA biomarkers can differentiate PDAC from intermediary pathologies of the pancreas, including all diverse diagnoses such as benign conditions of the pancreas, chronic and acute pancreatitis, IPMN, and other cancer types. These biomarkers offer a significant improvement in identifying PDAC compared to CA19-9 at the clinical cutoff level.

The 29 cfRNA biomarkers we identified in this study are highly expressed in the liver, lymph node, pancreas, and other immune-related tissues (bone marrow, tonsil, spleen). The level of liver cfRNA biomarkers is elevated while immune-related genes are lower in PDAC. Deconvolution of the global cell-free transcriptome to tissue contribution also confirms a three-fold increase in liver contribution in PDAC patients compared to all others. Note that only a fraction of PDAC

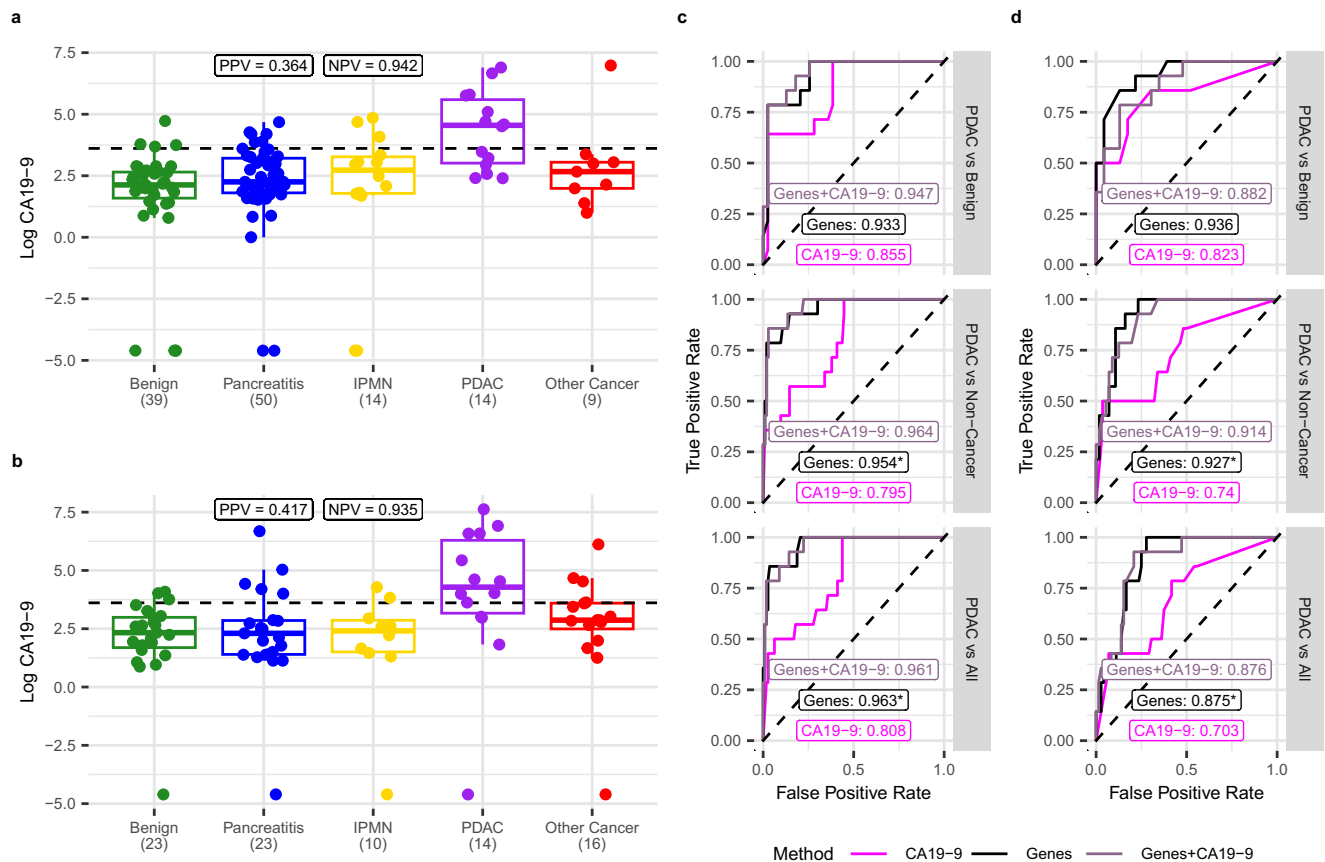


Fig. 6 | Comparison of cfRNA classifier with CA19-9 measurement. Measured CA19-9 levels in CEDAR (**a**) and BCC (**b**) cohorts. The clinical cutoff value of 37 shown as a dotted line, is used to calculate PPV and NPV for the detection of PDAC. Number of available samples with CA19-9 measurements is shown in parentheses. Boxplots display the 25th, 50th, and 75th percentiles of groups, with whiskers extending up to 1.5 times beyond inter-quartile range. Classifier results

using CA19-9, discovered gene set, and discovered gene set plus CA19-9 in CEDAR cross-validation (**c**) and BCC independent validation (**d**). *indicates when AUC using DE genes is significantly better than just CA19-9 ($P < 0.05$, two-sided Delong's hypothesis test). P values for **c** (top to bottom) are 0.175, 0.0047, 0.0039. P values for **d** (top to bottom) are 0.0895, 0.0177, 0.0424.

patients had liver metastasis at the time of blood draw, and the relative liver contribution of PDAC samples remains high when these patients are omitted. The involvement of the liver as one of the top tissues contributing to cfRNA biomarkers in PDAC may indicate the early remodeling of the liver in the early progression of PDAC, as indicated in previous mouse studies^{44–46}. This finding prompts the need for further investigation into the mechanism of pancreas-liver communication and the modulation of biomarker secretion from these two organs and other tissues in PDAC without detectable liver metastasis.

Having internal controls to account for technical variability is crucial for measurement reproducibility and eliminating batch effects. Previous efforts in normalization and batch correction for RNA sequencing for cellular RNA⁴⁰ are unlikely to be effective for cfRNA in the blood due to the variation in ex vivo generation of cfRNA from platelets and other blood cellular components³⁵. Here, we developed a normalization method specifically for cfRNA that can be used to account for technical variations throughout the whole process from blood handling, storage, processing, and sequencing. Please note that these various processing conditions were only used to identify the intrinsic and extrinsic factors for developing the normalization method. The processing condition for actual patient blood collection was uniform without extended delay as described in the method section. The consistently strong classifier performance on the external validation set is due in large part to our cf-normalization. The method

is applicable to general analyses of extracellular RNA in the blood for not only PDAC but also for other disease conditions.

Predictive machine learning models are an increasingly common tool for analyzing biomarker potential, but the rarity of many pathologies means that class imbalance is a critical problem to address in their use in diseases such as pancreatic cancer. In settings where disease samples only account for a small percentage of the total data, optimization methods may ignore the disease class altogether in order to achieve a high overall accuracy. General solutions to this issue include adjusting the misclassification cost used to train specific algorithms⁵⁵, under-sampling majority classes^{56,57}, or creating synthetic/repeated samples of minority classes. We opted to use the latter approach in order to maintain usability on more general classifier types and to not lose potential signals by excluding samples. The SMOTE method of creating synthetic minority samples randomly generates new samples such that they lie between two neighboring minority samples in the feature space. This ensures that (1) synthetic samples are bounded by real samples in terms of their features, and (2) the relative shape of the minority class in feature space is preserved when including the synthetic samples. These properties help to ensure that the learned classifier is driven by the properties of the real biological samples, rather than random artifacts introduced from synthetic samples. Using SMOTE helped to improve the separation of PDAC samples from all others in our classification, improving the AUC.

The robustness of the approach can be seen in the model performance when translated to the external validation cohort.

It is important to note a significant limitation of our study regarding the small sample size. The number of cases in the discovery and validation sets were 20 and 21, respectively. Due to the nature of disease heterogeneity, the control group included a small patient number with various diagnostic outcomes. For this study, we focused only on pre-diagnosis blood draws only, limiting the study population to patients referred for the EUS procedure. This study design may introduce bias toward patients with EUS indications. In clinical practice, many patients with pancreatitis and IPMN may not require EUS. Whether the cfRNA approached reported in the current study is applicable for this patient population remains to be explored in future studies.

Further development is needed before fully evaluating the utility of cfRNA to combine with EUS-FNA screening for PDAC diagnosis. Our result demonstrated diagnostic performance of the cfRNA classifier in differentiating PDAC from various groups of controls with intermediary pathologic and benign conditions of the pancreas. However, the cfRNA classifier presented here needs further development to be used as a method to determine the need for either EUS or EUS-FNA. Technically, a targeted assay to capture the biomarkers with high efficiency may give better sensitivity as well as decreased cost and time compared to whole transcriptome sequencing. The sensitivity for detecting PDAC at early stages and false negative calls needs to be established in large-scale clinical settings. Combining plasma cell-free biomarkers with other blood analytes, such as serum protein biomarkers, nucleic-acid-based biomarkers, protease assays, and extracellular vesicle cargoes, should be considered to improve sensitivity and specificity as well as the utility of the approach.

Methods

Patient recruitment and specimen collection

Our research complies with all relevant ethical regulations approved by the Institutional Review Board at OHSU with the IRB study number 3609. Informed consent was obtained from all subjects. All methods were carried out in accordance with relevant guidelines and regulations.

We recruited patients who were referred to the EUS procedure at OHSU for this study. Inclusion criteria include patients with hereditary high risk, undergoing surveillance for pancreatic cancer, OR patients with a history of chronic pancreatitis or other gastric conditions, OR patients with obstructive pancreatic or biliary symptoms. Exclusion criteria include patients unable to provide consent. Blood samples were collected after consenting and before any imaging and diagnostic procedure. Blood specimens were collected in EDTA tube before the clinical procedure. We do not control for time of blood draw and fasting status other than what was required by the standard clinical visit. The CEDAR cohort includes patients recruited by CEDAR between 2019 and 2021. BCC cohort includes patients recruited by the BCC between 2017 and 2019. All the samples were collected and banked by BCC and CEDAR. Banked samples were released to multiple studies. Upon our request, samples with sufficient plasma volume were released to perform this study.

Sample preparation

Patient plasma (2 mL) was processed from whole blood by centrifugation twice; first at $1000 \times g$ for 10 min, then the supernatant plasma was removed and spun at $2500 \times g$ for 10 min at room temperature. Plasma aliquots were frozen and stored at -80°C until we began RNA processing, which took place in batches of eight in a randomized order for each cohort. We purified RNA from these plasma samples using a Plasma/Serum RNA purification kit available commercially from Norgen Biotek (ref. 42800). The RNA elution volume is 100 μL .

Genomic DNA was removed from the resulting RNA with Lucigen Baseline-ZERO Dnase I (ref. DB0715K) in a 20 min incubation. The RNA was then further purified using a commercially available RNA Clean

and Concentrator kit from Zymo Research (ref. R1016), and aliquoted to be stored at -80°C . The final RNA elution volume is 20 μL .

Library preparation and sequencing

We prepared sequencing libraries from 5 μL purified RNA for each sample using the Takara SMARTer Stranded Total RNA-seq kit v2- Pico Input Mammalian formulation (ref. 634413) according to the manufacturer's instructions. We used option 2 during the library processing, without fragmentation. During the adapter addition step, we used the Takara SMARTer RNA Unique Dual Index kit -96U (ref. 634452) to append a unique index to each 5' and 3' read to aid in distinguishing samples. The final PCR amplification of the RNA-seq libraries used 16 cycles, and libraries were eluted into 20 μL Tris buffer and stored at -20°C before sequencing.

We randomized all samples for RNA extraction, library preparation and sequencing. All samples have a unique index and were pooled and randomly assigned to the same sequencing lanes. We control and measure the concentration of the libraries by bioanalyzer and qbit. Samples were again controlled by the sequencing services using Tape Station and qPCR before loading to the sequencer according to the standard operating procedure at the sequencing facility. Samples were sequenced at 100 million 75 or 150 bp paired-end read depth using a Novaseq 6000 sequencer, and processed on a computation cluster. We trimmed adapter sequences using sickle (ver 1.33) and ran an initial quality check with FastQC (ver 0.11.7). We then aligned each sample to hg38 using STAR (ver 2.5.3) and removed duplicate reads using Picard (ver 1.119), followed by generating read counts for each gene using htseq-count (ver 0.11.2).

Intrinsic and extrinsic factor identification

Let $g_{i,j}$ be the observed level for gene i in sample j , $\mathbf{I}$ be the intrinsic factor for all genes, and $\mathbf{E}$ be the extrinsic factor. We use the model $g_{i,j} = \alpha_j I_i + (1 - \alpha_j) E_i$ where α_j controls the percentage of the total level due to each factor in each sample. In matrix notation, for a collection of n samples over p genes, we can write $\mathbf{G}_{[p \times n]} = \mathbf{F}_{[p \times 2]} \mathbf{A}_{[2 \times n]}$ where $\mathbf{G}$ is the gene level, $\mathbf{F}$ the intrinsic and extrinsic factors, and $\mathbf{A}$ the concatenation of α_j and $(1 - \alpha_j)$. In this form, we can estimate the factor level and alpha values using NMF, with the added constraint that columns of $\mathbf{A}$ sum to 1. We use the platelet marker genes *PF4*, *PPBP*, *TAGLN2*, *TM2D2*, and *IFRD1* to identify which column of $\mathbf{F}$ is the extrinsic factor and which is the intrinsic factor.

Factor estimation was done on a separate dataset of 70 cfRNA samples collected from 10 healthy donors. Plasma from each donor was processed using seven different handling conditions before sequencing (Supplementary Table S2). Constrained NMF was used on the CPM normalized samples to estimate the intrinsic and extrinsic factors for each gene.

Cf-normalization

The intrinsic and extrinsic factors learned on our healthy reference data are estimates of the contribution level of intrinsic and extrinsic processes for each gene when both of these contributions are balanced in the sample. We use the normalized ratio of intrinsic and extrinsic factors for each gene to split the cfRNA count table into an intrinsic and extrinsic table, such that $g_{i,j} = I_i^n g_{i,j} + E_i^n g_{i,j}$ for each sample j and gene i , where $I_i^n = I_i / (I_i + E_i)$ and $E_i^n = E_i / (I_i + E_i)$. Both $I_i^n g_{i,j}$ and $E_i^n g_{i,j}$ undergo total sum normalization (abbreviated as TSN), where each sample is divided by its total counts and multiplied by the median sample total. These values are then combined back into the cf-normalized gene table $g_{i,j}^n = \text{TSN}(I_i^n g_{i,j}) + \text{TSN}(E_i^n g_{i,j})$. Performing TSN on these contributions separately ensures that the total expected contributions across all genes is normalized across samples, assuming the per-gene contribution ratios learned in the reference data, though there may still be some deviations in individual genes for different samples.

As outlined above, our cf-normalization can only be used on genes that are present in the reference cohort. For genes present in the PDAC data and not present in the reference cohort, we estimated their intrinsic and extrinsic contributions by re-fitting the same NMF model on the training and validation data, but holding the factor values for reference cohort genes constant. This learns the factor contributions on new genes while using the previously found gene factors to guide the factorization. We note that this method is an approximation that can be used when reference values of the genes are unavailable. One of our discovered biomarkers, *AC139099.4*, was added in this way, as it had no observed reads in the reference cohort.

DE gene discovery

DE genes were discovered on the CEDAR cohort. We first removed all genes with sample variance <1 , as computed on the normalized count matrices. For each gene, a Wilcoxon rank sum test is performed, comparing PDAC samples against all other diagnoses. Any gene with an FDR-corrected $p \leq 0.05$ is designated as DE and included in the model.

Tissue enrichment analysis

Tissue enrichment was performed in reference to the “RNA HPA tissue gene data” dataset from Human Protein Atlas version 22.0⁵⁸. The TPM levels of each DE gene were normalized across tissues using Z-score, which involves subtracting the mean for each gene across tissues and dividing by the standard deviation of that gene. Tissues are then ranked by total normalized count for all DE genes. For heatmap generation, Z-scores were capped at 5 to improve visualization.

Tissue deconvolution

The signature matrix for tissue deconvolution was created using the Human Protein Atlas tissue gene dataset TPM values, filtered for our DE gene list and with a minimum count capped at 1. We used support-vector regression to perform deconvolution, similar to CIBERSORT⁵⁹ and Vorperian et al.⁶⁰. The combination of CEDAR and BCC cohorts, filtered for DE genes, was used as the input matrix for deconvolution.

Both input matrix and signature matrix columns (i.e., samples/tissues) were zero-centered and scaled by standard deviation. Support vector nu-regression was performed on each sample individually, using the sample as the target vector and the signature matrix as the input variables. We used a linear kernel function, with a parameter search over the value of nu, selecting the value with the lowest sum of squares error. To find the tissue proportions for each sample, we took the learned coefficients for the model on that sample, set negative values to 0, and then normalized them to sum to one.

Figure 3b was created using tissue proportions learned from deconvolution. We averaged the predicted tissue proportions for PDAC samples and for non-PDAC samples over both datasets. The averages were normalized, per tissue, to sum to one for the bar plot.

Machine learning experiments

Our machine learning pipeline takes as input a training set of labels and features, a validation set of features, a classifier model to use, and parameter settings for the classifier. We perform a $\log(x+1)$ transform on all input feature tables to reduce the effects of outliers. Labels are in the binary categories of PDAC and not-PDAC. We use SMOTE⁴⁷ to create additional synthetic PDAC training samples, so that the number of samples for each class is balanced in the training cohort. The modified training data is used to train the given classifier with the given parameter settings, and the trained model predicts the probability of each validation sample being PDAC, which we call the PDAC score. The predicted PDAC score and the true validation set labels are used to compute ROC and AUC as evaluation measures.

We performed two sets of experiments to evaluate the best choice of classifier and the optimal parameter settings. For both experiments, we used only the PDAC and benign samples of the CEDAR cohort and

evaluated performance using 10-fold cross-validation. Since there are only 20 total PDAC samples in the CEDAR cohort, we ensured that each fold of data contained exactly two random PDAC samples. For classifier evaluation, we compared LDA, SVM, and RF, implemented in the R packages “MASS”, “e1071”, and “randomForest”, respectively. Each classifier was evaluated using the default choices of parameters, with RF showing the best performance.

For parameter evaluation of RF, we looked at the number of variables to use in each split (1, 5, or 10), the size of the leaf nodes (1, 5, or 10), the total number of trees created (250, 500, or 1000), and whether sampling was performed with or without replacement. This equated to 54 different parameter settings that were evaluated. Instead of using the highest single-performing parameter combination, we looked at the box plots of performance for each parameter and selected the option for each parameter with the highest median value (Supplementary Fig. S9). This corresponded to one variable used in each split and sampling with replacement. For the other parameters where there was no obvious best choice, we used the lowest values tested: leaf node size of 1 and 250 trees.

To generate the results for cross-validation and external validation, we use the parameter setting above in three different experiment settings. In the first setting, we compare PDAC samples to benign samples, the second setting excludes all non-PDAC cancer samples, and the final setting includes all available samples. These experiment settings show the performance of the PDAC score with increasingly heterogeneous diagnoses to compare against.

Survival analysis

For Kaplan–Meyer curves, we use the predicted PDAC score of PDAC samples to create low and high-risk groups for survival analysis. We use the 50th percentile of the scores, calculated for each dataset individually, as the cutoff between the two groups. We use the R package “survival” to compute the p value of a chi-squared test of equality between the two groups, modeling survival time with PDAC score as the only predictor.

As a separate model for survival analysis, we fit a Cox proportional hazards model to PDAC samples in each cohort, using PDAC score, sex, and age to predict survival time. This was done using the “coxph” function in the R “survival” package.

CA19-9 measurement

The CA19-9 levels in the samples were measured by the OHSU Pathology Lab Services Core lab using the Beckman Coulter Access GI Monitor Cancer Antigen 19-9 (ref. 387687) immunoassay. All samples with serum aliquots of $>200\ \mu\text{l}$ were run through the assay, which produced results for 126 samples in the CEDAR cohort and 86 samples in the BCC cohort. The Core lab determined the reference range of 0–37 U/ml.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Sequencing data associated with patient samples is publicly available in the Sequence Read Archive with the accession number [PRJNA1135202](https://www.ncbi.nlm.nih.gov/sra/PRJNA1135202). Sequencing data associated with the normalization samples is also publicly available in the Sequence Read Archive with the accession number [PRJNA1199029](https://www.ncbi.nlm.nih.gov/sra/PRJNA1199029). Processed data is available using the github link provided.

Code availability

In-house scripts used in this manuscript, along with count tables, metadata, and software versions, are publicly available on Github: https://github.com/ohsu-cedar-comp-hub/2024_ngolab_cfRNA. Code

and scripts used to produce this manuscript are citable via <https://doi.org/10.5281/zenodo.15678348>.

References

1. Siegel, R., Miller, K., Fuchs, H. & Jemal, A. Cancer statistics, 2022. *CA Cancer J. Clin.* **72**, 7–33 (2022).
2. Bengtsson A., Andersson R. & Ansari D. The actual 5-year survivors of pancreatic ductal adenocarcinoma based on real-world data. *Sci. Rep.* **10**, 16425 (2020).
3. Henrikson, N. et al. Screening for pancreatic cancer: updated evidence report and systematic review for the us preventive services task force. *JAMA* **322**, 445–454 (2019).
4. Poruk, K. E. et al. The clinical utility of CA 19-9 in pancreatic adenocarcinoma: diagnostic and prognostic updates. *Curr. Mol. Med.* **13**, 340–351 (2013).
5. Trikudanathan, G., Lou, E., Maitra, A. & Majumder, S. Early detection of pancreatic cancer: current state and future opportunities. *Curr. Opin. Gastroenterol.* **37**, 532–538 (2021).
6. Overbeek, K. A. et al. Timeline of development of pancreatic cancer and implications for successful early detection in high-risk individuals. *Gastroenterology* **162**, 772–785 (2022).
7. Luz, L. P., Al-Haddad, M. A., Sey, M. S. & DeWitt, J. M. Applications of endoscopic ultrasound in pancreatic cancer. *World J. Gastroenterol.* **20**, 7808–7818 (2014).
8. Kitano, M. et al. Impact of endoscopic ultrasonography on diagnosis of pancreatic cancer. *J. Gastroenterol.* **54**, 19–32 (2019).
9. Blackford, A. L. et al. Pancreatic cancer surveillance and survival of high-risk individuals. *JAMA Oncol.* **10**, 1087–1096 (2024).
10. Heitzer, E., Perakis, S., Geigl, J. & Speicher, M. The potential of liquid biopsies for the early detection of cancer. *NPJ Precis. Oncol.* **1**, 36 (2017).
11. Batool, S. M. et al. The liquid biopsy consortium: challenges and opportunities for early cancer detection and monitoring. *Cell Rep. Med.* **4**, 101198 (2023).
12. Cao, F. et al. Integrated epigenetic biomarkers in circulating cell-free DNA as a robust classifier for pancreatic cancer. *Clin. Epigenetics* **12**, 112 (2020).
13. Khan, I. A. et al. Panel of serum miRNAs as potential non-invasive biomarkers for pancreatic ductal adenocarcinoma. *Sci. Rep.* **11**, 2824 (2021).
14. Yu, S. et al. Plasma extracellular vesicle long RNA profiling identifies a diagnostic signature for the detection of pancreatic ductal adenocarcinoma. *Gut* **69**, 540–550 (2020).
15. Yekula, A. et al. Longitudinal analysis of serum-derived extracellular vesicle RNA to monitor dacomitinib treatment response in EGFR-amplified recurrent glioblastoma patients. *Neurooncol. Adv.* **5**, vdad104 (2023).
16. Roskams-Hieter, B. et al. Plasma cell-free RNA profiling distinguishes cancers from pre-malignant conditions in solid and hematologic malignancies. *NPJ Precis. Oncol.* **6**, 28 (2022).
17. Ngo, T. T. et al. Noninvasive blood tests for fetal development predict gestational age and preterm delivery. *Science* **360**, 1133–1136 (2018).
18. Vermeirssen, V. et al. Whole transcriptome profiling of liquid biopsies from tumour xenografted mouse models enables specific monitoring of tumour-derived extracellular RNA. *NAR Cancer* **4**, zcac037 (2022).
19. Reggiardo, R. E. et al. Profiling of repetitive RNA sequences in the blood plasma of patients with cancer. *Nat. Biomed. Eng.* **7**, 1627–1635 (2023).
20. Ning, C. et al. A comprehensive evaluation of full-spectrum cell-free RNAs highlights cell-free RNA fragments for early-stage hepatocellular carcinoma detection. *EBioMedicine* **93**, 104645 (2023).
21. Chen, S. et al. Cancer type classification using plasma cell-free RNAs derived from human and microbes. *Elife* **11**, e75181 (2022).
22. Larson, M. H. et al. A comprehensive characterization of the cell-free transcriptome reveals tissue- and subtype-specific biomarkers for cancer detection. *Nat. Commun.* **12**, 2357 (2021).
23. Hulstaert, E., Morlion, A., Levanon, K., Vandesompele, J. & Mestdagh, P. Candidate RNA biomarkers in biofluids for early diagnosis of ovarian cancer: a systematic review. *Gynecol. Oncol.* **160**, 633–642 (2021).
24. Majumder, S. et al. High detection rates of pancreatic cancer across stages by plasma assay of novel methylated DNA markers and CA19-9. *Clin. Cancer Res.* **27**, 2523–2532 (2021).
25. Guler, G. D. et al. Detection of early stage pancreatic cancer using 5-hydroxymethylcytosine signatures in circulating cell free DNA. *Nat. Commun.* **11**, 5270 (2020).
26. García-Ortiz, M. V., Cano-Ramírez, P., Toledano-Fonseca, M., Aranda, E. & Rodríguez-Ariza, A. Diagnosing and monitoring pancreatic cancer through cell-free DNA methylation: progress and prospects. *Biomark. Res.* **11**, 88 (2023).
27. Ben-Ami, R. et al. Protein biomarkers and alternatively methylated cell-free DNA detect early stage pancreatic cancer. *Gut* **73**, 639–648 (2024).
28. Haan, D. et al. Epigenomic blood-based early detection of pancreatic cancer employing cell-free DNA. *Clin. Gastroenterol. Hepatol.* **21**, 1802–1809 (2023).
29. Mohan, S. et al. Analysis of circulating cell-free DNA identifies KRAS copy number gain and mutation as a novel prognostic marker in Pancreatic cancer. *Sci. Rep.* **9**, 11610 (2019).
30. Sickels, A., Jain, T. & Dudeja, V. Peritoneal cell-free DNA: a novel biomarker for recurrence in pancreatic cancer. *Ann. Surg. Oncol.* **30**, 6308–6310 (2023).
31. Bryce, A. H. et al. Performance of a cell-free dna-based multi-cancer detection test in individuals presenting with symptoms suspicious for cancers. *JCO Precis. Oncol.* **7**, e2200679 (2023).
32. Botrus, G. et al. Serial cell-free DNA (cfDNA) sampling in advanced pancreatic ductal adenocarcinoma (PDAC) patients may predict therapeutic outcome. *J. Clin. Oncol.* **39**, 423–423 (2021).
33. Macgregor-Das, A. et al. Detection of circulating tumor DNA in patients with pancreatic cancer using digital next-generation sequencing. *J. Mol. Diagn.* **22**, 748–756 (2020).
34. Pepe, M. S., Feng, Z., Janes, H., Bossuyt, P. M. & Potter, J. D. Pivotal evaluation of the accuracy of a biomarker used for classification or prediction: standards for study design. *J. Natl. Cancer Inst.* **100**, 1432–1438 (2008).
35. Kim, H. J. et al. Irreversible alteration of extracellular vesicle and cell-free messenger RNA profiles in human plasma associated with ex vivo platelet activation. *Sci. Rep.* **12**, 2099 (2022).
36. Suzuki, K. et al. Establishment of preanalytical conditions for microRNA profile analysis of clinical plasma samples. *PLoS ONE* **17**, e0278927 (2022).
37. Haberberger, A. et al. Changes in the microRNA expression profile during blood storage. *BMJ Open Sport Exerc. Med.* **4**, e000354 (2018).
38. Cheng, H. H. et al. Plasma processing conditions substantially influence circulating microRNA biomarker levels. *PLoS ONE* **8**, e64795 (2013).
39. Mompeón, A. et al. Disparate miRNA expression in serum and plasma of patients with acute myocardial infarction: a systematic and paired comparative analysis. *Sci. Rep.* **10**, 5373 (2020).
40. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, 1–9 (2010).
41. Ke-Yu, L. et al. Pancreatic ductal adenocarcinoma immune micro-environment and immunotherapy prospects. *Chronic Dis. Transl. Med.* **6**, 6–17 (2020).

42. Yarchoan, M., Hopkins, A. & Jaffee, E. M. Tumor mutational burden and response rate to PD-1 inhibition. *N. Engl. J. Med.* **377**, 2500–2501 (2017).
43. Laklai, H. et al. Genotype tunes pancreatic ductal adenocarcinoma tissue tension to induce matricellular fibrosis and tumor progression. *Nat. Med.* **22**, 497–505 (2016).
44. Costa-Silva, B. et al. Pancreatic cancer exosomes initiate pre-metastatic niche formation in the liver. *Nat. Cell Biol.* **17**, 816–826 (2015).
45. Wang, G. et al. Tumour extracellular vesicles and particles induce liver metabolic dysfunction. *Nature* **618**, 374–382 (2023).
46. Lee, J. W. et al. Hepatocytes direct the formation of a pro-metastatic niche in the liver. *Nature* **567**, 249–252 (2019).
47. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, P. W. SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002).
48. Alix-Panabières, C., Marchetti, D. & Lang, J. E. Liquid biopsy: from concept to clinical application. *Sci. Rep.* **13**, 21685 (2023).
49. Holmes, D. R. Reducing the risk of needle tract seeding or tumor cell dissemination during needle biopsy procedures. *Cancers* **16**, 317 (2024).
50. Bai, R. Y., Staedtke, V., Xia, X. & Riggins, G. J. Prevention of tumor seeding during needle biopsy by chemotherapeutic-releasing gelatin sticks. *Oncotarget* **8**, 25955–25962 (2017).
51. Thomas, R. & Davies, N. Lifestyle during and after cancer treatment. *Clin. Oncol.* **19**, 616–627 (2007).
52. Stamatakis, E. et al. Vigorous intermittent lifestyle physical activity and cancer incidence among nonexercising adults: the UK biobank accelerometry study. *JAMA Oncol.* **9**, 1255–1259 (2023).
53. Balleshannina, U. K. & Chamberlain, R. S. 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 (2012).
54. Gold, G., Goh, S. K., Christophi, C. & Muralidharan, V. Dilemmas and limitations interpreting carbohydrate antigen 19-9 elevation after curative pancreatic surgery: a case report. *Int. J. Surg. Case Rep.* **54**, 20–22 (2019).
55. Pazzani, M. et al. Reducing misclassification costs. In *Proc. 11th International Conference on Machine Learning*. 217–225 (Morgan Kaufmann Publishers Inc., USA, 1994).
56. Kubat, M. & Matwin, S. Addressing the curse of imbalanced training sets: one-sided selection. In *Proc. 14th International Conference on Machine Learning*. 179–186 (Morgan Kaufmann Publishers Inc., USA, 1997).
57. Lewis, D. D. & Catlett, J. Heterogeneous uncertainty sampling for supervised learning. In *Proc. 11th International Conference on Machine Learning*. 148–156 (Morgan Kaufmann Publishers Inc., USA, 1994).
58. Karlsson, M. et al. A single-cell type transcriptomics map of human tissues. *Sci. Adv.* **7**, eab2169 (2021).
59. Newman, A. M. et al. Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12**, 453–457 (2015).
60. Vorperian, S. K., Moufarrej, M. N. & Quake, S. R. Cell types of origin of the cell-free transcriptome. *Nat. Biotechnol.* **40**, 855–861 (2022).

Acknowledgements

Research in Ngo lab was supported by the CEDAR at OHSU Knight Cancer Institute, Cancer Research UK/OHSU Project Award (C63763/A27122 to T.T.M.N.), the Department of Defense (W81XWH2110853 to T.T.M.N.), the Susan G. Komen Foundation (CCR21663959 to T.T.M.N.), and the Kuni Foundation. We would like to acknowledge the CEDAR repository and the BCC, and the Oregon Pancreas Tissue Registry for

their support in providing samples and clinical data. The research reported in this publication used computational infrastructure supported by the Office of Research Infrastructure Programs, Office of the Director, of the National Institutes of Health under Award Number S10OD034224. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Parts of Figures 1 and 2 were created with BioRender.com under licenses agreement numbers LZ28CN1XFT, WI28DNVJJW, and WY28DRWMVO.

Author contributions

T.W.M.: algorithm, analysis, illustration, & writing; E.S.: methodology, samples processing, data processing, & editing; R.L.C.: data processing, analysis, illustration & editing; C.W.K.: methodology, samples processing; C.F.B.: performing experiments; H.J.K.: methodology; B.R.: methodology, F.G.: performing experiments; D.K.: sample acquisition; A.G.: involving in study design; P.T.S.: sample acquisition; G.B.M.: involving in study design, funding acquisition & editing; R.C.S.: sample acquisition, T.K.M.: involving in study design, clinical annotation & sample acquisition; T.T.M.N.: study design, funding acquisition, methodology, analysis strategy, writing, and supervised all aspects of the study. All authors approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-025-62685-y>.

Correspondence and requests for materials should be addressed to Thuy T. M. Ngo.

Peer review information *Nature Communications* Jo Vandesompele, Jun Zhong and the other anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025