scientific reports

OPEN

Multi-omic biomarker panel in pancreatic cyst fluid and serum predicts patients at a high risk of pancreatic cancer development

Laura E. Kane[1], Gregory S. Mellotte[2], Eimear Mylod[1], Paul Dowling[3,4], Simone Marcone[1], Caitriona Scaife[5], Elaine M. Kenny[6], Michael Henry[7], Paula Meleady[7], Paul F. Ridgway[8], Finbar MacCarthy[9], Kevin C. Conlon[10], Barbara M. Ryan[2] & Stephen G. Maher[1]✉

Integration of multi-omic data for the purposes of biomarker discovery can provide novel and robust panels across multiple biological compartments. Appropriate analytical methods are key to ensuring accurate and meaningful outputs in the multi-omic setting. Here, we extensively profile the proteome and transcriptome of patient pancreatic cyst fluid (PCF) (n = 32) and serum (n = 68), before integrating matched omic and biofluid data, to identify biomarkers of pancreatic cancer risk. Differential expression analysis, feature reduction, multi-omic data integration, unsupervised hierarchical clustering, principal component analysis, spearman correlations and leave-one-out cross-validation were performed using RStudio and CombiROC software. An 11-feature multi-omic panel in PCF [PIGR, S100A8, REG1A, LGALS3, TCN1, LCN2, PRSS8, MUC6, SNORA66, miR-216a-5p, miR-216b-5p] generated an AUC = 0.806. A 13-feature multi-omic panel in serum [SHROOM3, IGHV3-72, IGJ, IGHA1, PPBP, APOD, SFN, IGHG1, miR-197-5p, miR-6741-5p, miR-3180, miR-3180-3p, miR-6782-5p] produced an AUC = 0.824. Integration of the strongest performing biomarkers generated a 10-feature cross-biofluid multi-omic panel [S100A8, LGALS3, SNORA66, miR-216b-5p, IGHV3-72, IGJ, IGHA1, PPBP, miR-3180, miR-3180-3p] with an AUC = 0.970. Multi-omic profiling provides an abundance of potential biomarkers. Integration of data from different omic compartments, and across biofluids, produced a biomarker panel that performs with high accuracy, showing promise for the risk stratification of patients with pancreatic cystic lesions.

Keywords Pancreatic cancer, Pancreatic cystic lesion, Risk stratification, Biomarker, Multi-omics

Abbreviations

CBF	Cross-biofluid
LOOCV	Leave-one-out cross-validation
PC	Pancreatic cancer
PCA	Principal component analysis
PCF	Pancreatic cyst fluid
PCL	Pancreatic cystic lesion
UHC	Unsupervised hierarchical clustering
VHL	Von-Hippel Lindau

[1]Department of Surgery, Trinity St. James's Cancer Institute, Trinity Translational Medicine Institute, Trinity College Dublin, St. James's Hospital, Dublin 8, Ireland. [2]Department of Gastroenterology, Tallaght University Hospital, Dublin 24, Ireland. [3]Department of Biology, Maynooth University, Maynooth, Ireland. [4]Kathleen Lonsdale Institute for Human Health Research, Maynooth University, Maynooth, Ireland. [5]Mass Spectrometry Facility, Conway Institute of Biomolecular and Biomedical Research, University College Dublin, Dublin 4, Ireland. [6]ELDA Biotech, Newhall, M7 Business Park, Co. Kildare, Ireland. [7]National Institute for Cellular Biotechnology, Dublin City University, Dublin 9, Ireland. [8]Department of Surgery, Centre for Pancreatico-Biliary Diseases, Trinity College Dublin, St. James's Hospital, Dublin 8, Ireland. [9]Department of Clinical Medicine, Trinity Translational Medicine Institute, Trinity College Dublin, St. James's Hospital, Dublin 8, Ireland. [10]Department of Surgery, School of Medicine, Trinity College Dublin, Dublin 2, Ireland. ✉email: maherst@tcd.ie

Multi-omics has been at the forefront of medical research for the last decade, affording scientists and clinicians the opportunity to evaluate multiple compartments of biological data, interrogate differences, make correlations, tease out functional properties, and inform personalized medicine approaches for the management of many diseases[1–3]. However, there are important inherent difficulties to handling the large-scale datasets generated by multi-omics, such as appropriate integration of data from different biological compartments, handling missing data or outliers, scaling of different variables and SI units (International System of Units), inflation of false discovery rates, and choosing the appropriate data handling and analysis pipeline that is suitable to the dataset[4]. In this study, we perform detailed analyses of both single- and multi-omic datasets in a set of matched biological samples for the purposes of biomarker discovery, highlighting the strengths and weaknesses of each approach as we generate a novel and robust multi-omic cross-biofluid (CBF) biomarker panel for pancreatic cancer risk stratification.

Pancreatic cancer (PC) has the worst 5-year survival rate of any cancer as of 2024, at just 13%[5]. Early detection of PC is the primary concern of most PC research, as it has the potential to make a substantial difference to the treatment and survival of these patients. Pancreatic cystic lesions (PCLs) are fluid-filled sacs on or inside the pancreas, that have the potential to become premalignant[6]. While some PCLs are completely benign, others have been shown to have malignant potential and could therefore play a role in the progression to PC[6]. The issue arises in distinguishing which PCLs are benign and which are premalignant and should, as such, be monitored and/or treated accordingly. At present, there are several sets of clinical guidelines worldwide for the stratification of PCLs into risk groups based on their clinical presentation[7,8]. Unfortunately, the presence of several sets of guidelines worldwide indicates the lack of consensus among clinicians as to the cut-offs or defined parameters for these classification factors. As such, the risk stratification of these patients is inaccurate, and could therefore be contributing to the overall problem of early detection.

Importantly, genetic mutations generally occur alongside the development of precursor lesions such as PCLs, and are therefore present at their subsequent progression through increasing histological grades, culminating in invasive carcinoma[9,10]. These genetic mutations can influence the classification of PCL patients into low- and high-risk categories. One such important genetic mutation in this study is von Hippel-Lindau (VHL) syndrome. VHL syndrome is a familial neoplastic condition, caused by a germline mutation to the *VHL* tumour suppressor gene, which can increase a patient's risk of developing PCLs during their lifetime, while also increasing their risk of developing malignant cystic or solid lesions, such as cystic pancreatic neuroendocrine tumours or PDAC[11–13]. As such, while a patient may present with a low-risk PCL, the patient themselves would be regarded as high-risk given their genetic predisposition to PCLs and PC. This highlights yet another potential variable in risk stratification and multi-omic dataset evaluation, that a rigorous analytical pipeline, and subsequently a robust biomarker or biomarker panel, should be able to adequately account for.

The identification of novel, robust biomarkers for the early detection of PC risk is urgently needed for these patients, and could provide a much-needed change to the way in which PCLs are managed, enabling the discovery of high-risk patients at earlier stages of PC development. In this study, the proteome and transcriptome of PCL patient pancreatic cyst fluid (PCF) and serum were profiled in order to identify differentially expressed proteins and miRNA between low- and high-risk patients. The proteome and transcriptome were chosen as target compartments as they have been consistently demonstrated to produce the most promising biomarkers in this setting[14]. Differentially expressed proteins and miRNA were examined both alone, and as part of a multi-omic panel, in both the PCF and the serum, in order to examine their utility as biomarkers of risk stratification. Extensive evaluation of these features was carried out using unsupervised hierarchical clustering (UHC), principal component analysis (PCA) and Spearman correlations, with leave-one-out cross-validation (LOOCV) being used to fit, train and validate predictive linear classification models. Novel and cutting-edge feature selection methodologies were then performed to reduce the PCF- and serum-based panels to the top performing biomarkers in each fluid. The most robust biomarkers identified from each biofluid were then scaled and integrated to create a cross-biofluid (CBF) multi-omic panel. The utility of both CA19-9 and CEA, currently utilised biomarkers in this setting that are imperfect and are frequently dysregulated, were also examined in this cohort, and integrated with the top performing panel in order to assess whether they could improve its performance.

Results

8-protein panel in PCF stratifies patients into risk categories with modest accuracy

Label-free proteomics identified 465 proteins present across PCF samples after data clean-up. Differential expression analysis revealed eight proteins [PIGR, S100A8, REG1A, LGALS3, TCN1, LCN2, PRSS8 and MUC6] to be significantly upregulated in high-risk PCF compared to low-risk (adj-$p < 0.002$, FDR = 0.05, s0 = 0.1) (Fig. 1A and B). These eight proteins were integrated to create an 8-protein biomarker panel. UHC of patients into risk groups using this 8-protein panel was performed with an accuracy of 81.25% (Fig. 1C). The VHL outlier patient was shown to relate more closely to the low-risk classification than any other high-risk patient. In the PCA, modest separation of the low- and high-risk groups can be seen, with the high-risk ellipse being larger in size, indicating more variance in this group (Supplementary Material S1)(Fig. 1D). Spearman correlations showed that expression levels of each of the eight proteins correlated positively with patient risk ($p < 0.01$), with some proteins having significant positive correlations with age, and negative correlations with alcohol consumption ($p < 0.05$)(Fig. 1E).

3-miRNA panel in PCF stratifies patients into risk categories with poor accuracy

Whole transcriptome sequencing identified 2096 miRNAs present across PCF samples after data clean-up. Differential expression analysis revealed three miRNAs [SNORA66, miR-216a-5p and miR-216b-5p] to be significantly upregulated in high-risk PCF compared to low-risk (adj-$p < 0.05$, FDR = 0.05, s0 = 0.1) (Fig. 2A

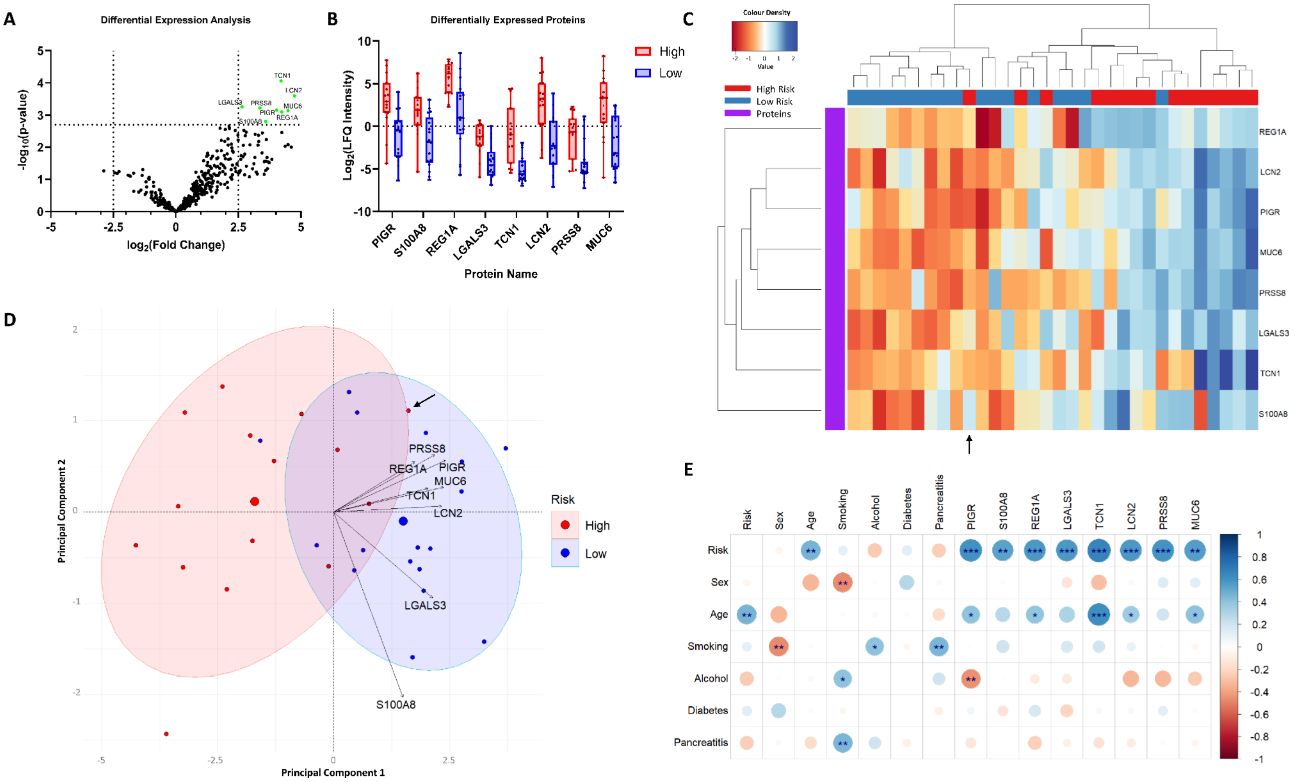

Fig. 1. Proteomic analysis of PCF identifies eight proteins significantly upregulated in high-risk patients compared to low-risk patients. (**A**) Differential expression analysis identified eight proteins that were differentially expressed between low- and high-risk PCF samples (adj-$p < 0.05$, FDR = 0.05, s0 = 0.1). (**B**) Boxplots showing the distribution of patient expression levels for the eight differentially expressed proteins. (**C**) UHC of patients into high- and low-risk groups based on their expression of the eight differentially expressed proteins. Dendrograms show (top) the relatedness of the patients, and (left) the relatedness of the differentially expressed proteins. (**D**) 2D PCA using the eight differentially expressed proteins, with biplot overlayed. Ellipses represent 80% of the data captured within the two risk classifications. Biplot scale is set to zero to ensure vectors (arrows) are scaled to represent their respective loadings. The length of each vector is proportional to the variance of the corresponding protein. (**E**) Spearman correlations between patient clinical data and the eight differentially expressed proteins are given as a corrplot. Colour intensity relates to R value, circle size relates to the p-value (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$). Black arrows show the position of the VHL outlier patient.

and B). These three miRNAs were integrated to create a 3-miRNA biomarker panel. UHC of patients into risk groups using this 3-miRNA panel was performed with an accuracy of 60% (Fig. 2C). In the PCA, the entire low-risk ellipse was captured inside that of the high-risk ellipse, showing the poor separation of the two groups, with the VHL outlier patient being encapsulated within both classifications (Supplementary Material S2) (Fig. 2D). Spearman correlations showed that only miR-216a-5p had a significant positive correlation with patient risk ($p < 0.05$) (Fig. 2E).

11-feature multi-omic panel in PCF stratifies patients into risk categories with high accuracy

The 8-protein and 3-miRNA panels were then scaled and integrated to create an 11-feature multi-omic biomarker panel. UHC of patients into risk groups using this 11-feature multi-omic panel was performed with an accuracy of 95.8%, with only the VHL outlier being grouped incorrectly (Fig. 3A). There were no significant correlations between any of the three miRNAs and eight proteins ($p > 0.05$) (Fig. 3B). S100A8 represented a smaller segment than the other biomarkers and only correlated significantly with REG1A, LGALS3 and TCN1 ($p < 0.05$). In the PCA, the proteins had less variance compared to the miRNAs, with S100A8 having the least variance in the 11-feature panel (Supplementary Material S3) (Fig. 3C). The high-risk ellipse was much larger again, indicating larger variance in the high-risk population. Importantly, the single high-risk datapoint that caused the high-risk ellipse to overlap substantially with the low-risk ellipse represented the VHL outlier patient. Using LOOCV, the 8-protein panel produced an AUC of 0.607 (Sensitivity = 70.6%, Specificity = 60%); the 3-miRNA panel produced an AUC of 0.658 (Sensitivity = 60%, Specificity = 53.3%); and the 11-feature multi-omic panel produced an AUC of 0.806 (Sensitivity = 66.7%, Specificity = 75%) (Fig. 3D). Despite the modest performances of the 8-protein and the 3-miRNA panels alone, the integration of these two panels together improved the overall performance. Importantly, when the VHL outlier patient was reclassified to low-risk, the performance of all panels improved (Fig. 3D).

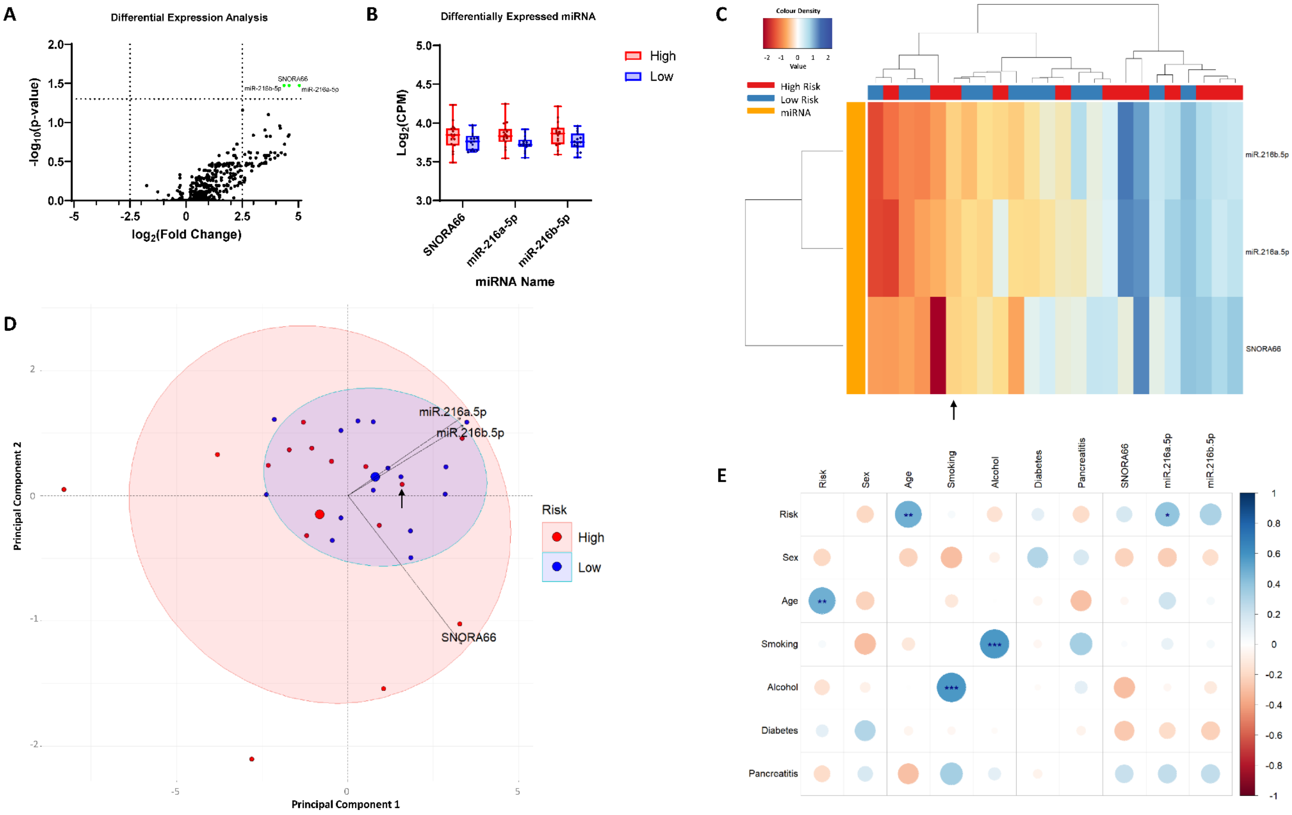

Fig. 2. Transcriptomic analysis of PCF identifies three miRNAs significantly upregulated in high-risk patients compared to low-risk patients. (**A**) Differential expression analysis identified three miRNAs that were differentially expressed between low- and high-risk PCF samples (adj-$p < 0.05$, FDR = 0.05, s0 = 0.1). (**B**) Boxplots showing the distribution of patient expression levels of the three differentially expressed miRNAs. (**C**) UHC of patients into high- and low-risk groups based on their expression of the three differentially expressed miRNAs. Dendrograms show (top) the relatedness of the patients, and (left) the relatedness of the miRNAs. (**D**) 2D PCA using the three differentially expressed miRNAs, with biplot overlayed. Ellipses represent 80% of the data captured within the two risk classifications. Biplot scale is set to zero to ensure vectors (arrows) are scaled to represent their respective loadings. The length of each vector is proportional to the variance of the corresponding miRNA. (**E**) Spearman correlations between patient clinical data and the three differentially expressed miRNAs are given as a corrplot. Colour intensity relates to R value, circle size relates to the p-value (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$). Black arrows show the position of the VHL outlier patient.

8-protein panel in serum stratifies patients into risk categories with modest accuracy

Label-free proteomics identified 145 proteins present across the serum samples after data clean-up. Two proteins [SHROOM3 and IGHV3-72] were significantly downregulated in high-risk serum compared to low-risk ($p < 0.05$) (Fig. 4A). Despite not being significantly differentially expressed, a further six proteins with the lowest p-values were taken forward for biomarker analysis [IGJ, IGHA1, PPBP, APOD, SFN, IGHG1], as panels of multiple biomarkers have been shown to produce better results[14]. A total of eight proteins were examined, as the 8-protein panel in PCF examined above was shown to have good accuracy (Fig. 4B). These eight proteins were integrated to create an 8-protein biomarker panel. UHC of patients into risk groups using this 8-protein panel was performed with an accuracy of 76.5% (Fig. 4C). In the PCA, three components were required to account for 69.5% of the variance (Supplementary Material S4) (Fig. 4D). Separation of the two groups was modest, with some clustering of low- and high-risk samples being seen across the three components. Here, the VHL outlier was on the outer peripheries of both groups, indicating no preferential alignment with either group. Spearman correlations showed that SHROOM3 was the only one of the eight proteins to significantly correlate with patient risk, having a negative correlation ($p < 0.01$) (Fig. 4E). Both SHROOM3 and SFN positively correlated with smoking habits ($p < 0.05$).

5-miRNA panel in serum stratifies patients into risk categories with poor accuracy

Whole transcriptome sequencing identified 2,096 miRNAs present across the serum samples after data clean-up. Differential expression analysis revealed five miRNAs [miR-197-5p, miR-6741-5p, miR-3180, miR-3180-3p and miR-6782-5p] to be significantly upregulated in high-risk serum compared to low-risk (adj-$p < 0.05$, FDR = 0.05, s0 = 0.1) (Fig. 5A and B). These five miRNAs were integrated to create a 5-miRNA biomarker panel. UHC of patients into risk groups using this 5-miRNA panel was performed with an accuracy of 60% (Fig. 5C). The VHL outlier patient did not align with either risk classification. In the PCA, the ellipse of the low-risk classification

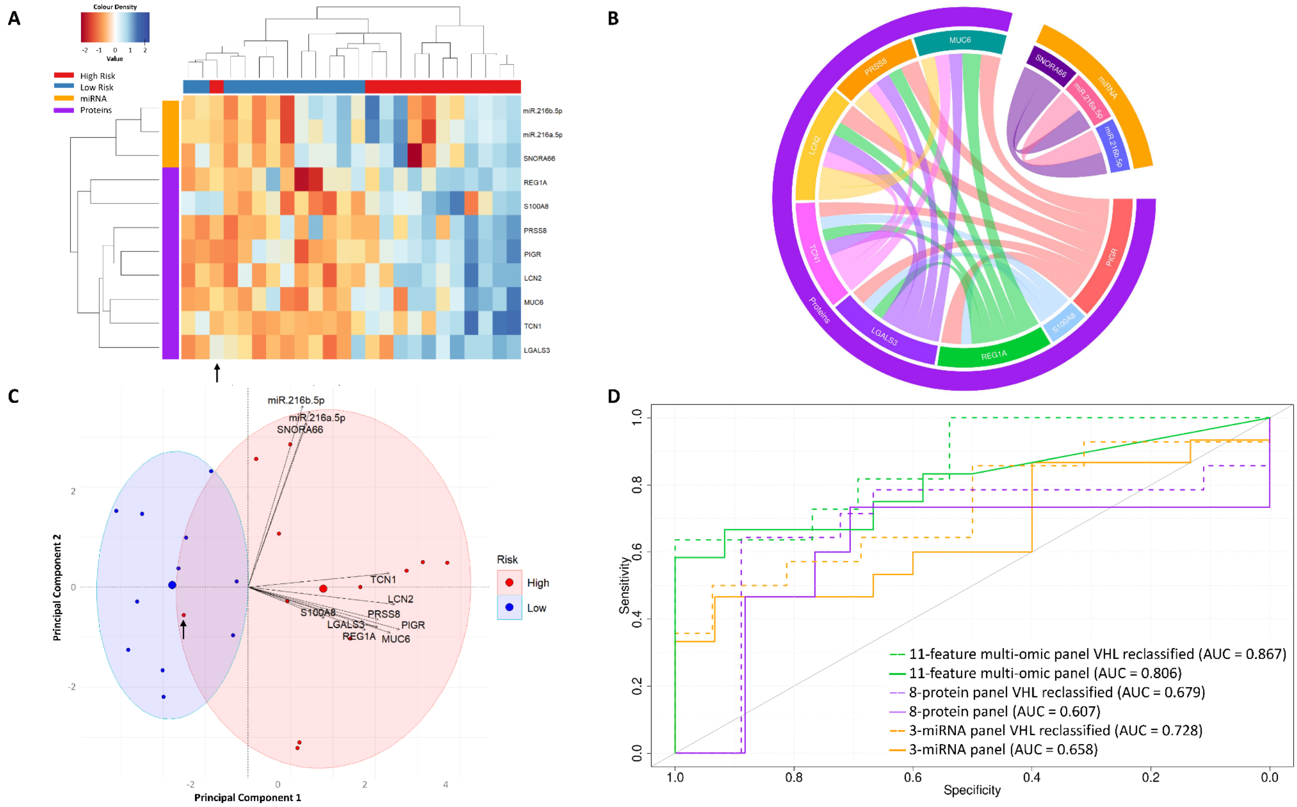

Fig. 3. Integration of the differentially expressed proteins and miRNAs generates a robust 11-feature multi-omic biomarker panel in PCF. (**A**) UHC of patients into high- and low-risk groups based on their expression of the eight proteins and three miRNAs which form an 11-feature multi-omic panel. Dendrograms show (top) the relatedness of the patients, and (left) the relatedness of the differentially expressed proteins and miRNAs. (**B**) Chord diagram showing significant Spearman correlations ($p < 0.05$) between the eight differentially expressed proteins and three differentially expressed miRNAs. Inner chords reflect correlations between the biomarkers. Chord thickness is directly related to the strength of the correlation, with thicker chords indicating stronger correlations. (**C**) 2D PCA using this 11-feature multi-omic panel, with biplot overlayed. Ellipses represent 80% of the data captured within the two risk classifications. Biplot scale is set to zero to ensure vectors (arrows) are scaled to represent their respective loadings. The length of each vector is proportional to the variance of the corresponding protein or miRNA. (**D**) ROC curves generated from LOOCV of the miRNAs alone, the proteins alone, and the 11-feature multi-omic panel, as well as their performances when the VHL outlier patient is reclassified. Black arrows show the position of the VHL outlier patient.

was captured almost entirely inside that of the high-risk classification, indicating poor separation of the two groups (Supplementary Figure S5) (Fig. 5D). Here, the VHL outlier was encapsulated within the high-risk ellipse only. Spearman correlations showed no correlations with patient risk ($p > 0.05$) (Fig. 5E). MiR-6741-5p had a significant positive correlation with increased smoking ($p < 0.05$).

13-feature multi-omic panel in serum stratifies patients into risk categories with high accuracy

The 8-protein and 5-miRNA panels were then scaled and integrated to create a 13-feature multi-omic biomarker panel. UHC of patients into risk groups using this 13-feature panel was performed with an accuracy of 79.3% (Fig. 6A). There were no significant correlations between SFN, PPBP and SHROOM3 and any of the other ten biomarkers ($p > 0.05$) (Fig. 6B). IGHV3-72 had the most correlations within the panel, significantly correlating with IGJ, IGHA1 and IGHG1 ($p < 0.05$). IGHV3-72 had a significant negative correlation with miR-6741-5p, and APOD had a significant negative correlation with miR-6782-5p ($p < 0.05$). The strongest correlations were found between miR-197-5p and both miR-3180 and miR-3180-3p, as indicated by the thicker chords. In the PCA, three components were required to account for 61.4% of the variance (Supplementary Material S6) (Fig. 6C). With three components, the separation of the two groups was modest, though not distinct. Importantly, the VHL outlier patient was grouped within the high-risk classification in this setting. Using LOOCV, the 8-protein panel produced an AUC of 0.608 (Sensitivity = 82.2%, Specificity = 34.8%); the 5-miRNA panel produced an AUC of 0.427 (Sensitivity = 46.7%, Specificity = 40.0%); and the 13-feature multi-omic panel produced an AUC of 0.824 (Sensitivity = 71.4%, Specificity = 80.0%) (Fig. 6D). Despite the poor performance of the 5-miRNA panel, and modest performance of the 8-protein panel, the integration of these two panels together improved the overall performance. Importantly, when the VHL outlier patient was reclassified to low-risk as before, the performance

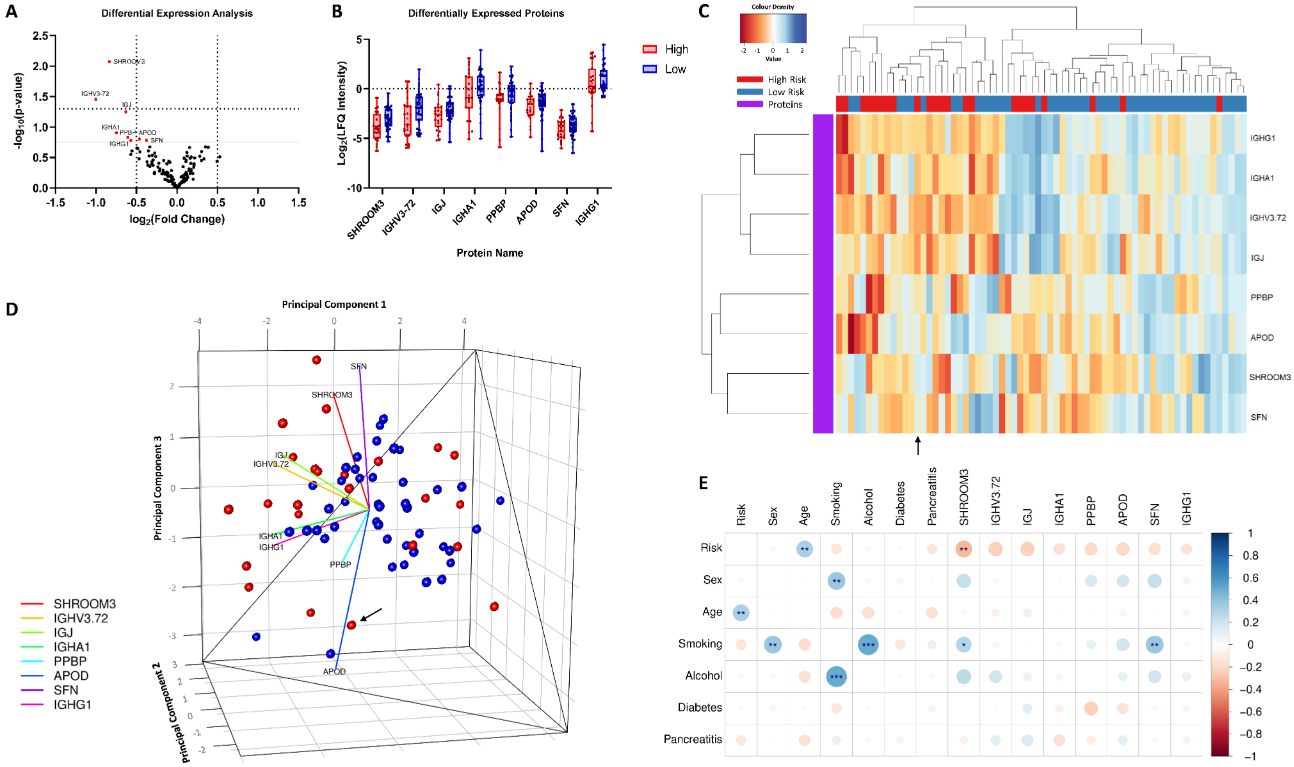

Fig. 4. Proteomic analysis of serum identifies eight proteins downregulated in high-risk patients compared to low-risk patients. (**A**) Differential expression analysis identified eight proteins that were differentially expressed between low- and high-risk serum samples. (**B**) Boxplots showing the distribution of patient expression levels for the eight differentially expressed proteins. (**C**) UHC of patients into high- and low-risk groups based on their expression of the eight differentially expressed proteins. Dendrograms show (top) the relatedness of the patients, and (left) the relatedness of the differentially expressed proteins. (**D**) 3D PCA using the eight differentially expressed proteins, with biplot overlayed. Biplot scale is set to zero to ensure vectors (arrows) are scaled to represent their respective loadings. The length of each vector is proportional to the variance of the corresponding protein. (**E**) Spearman correlations between patient clinical data and the eight differentially expressed proteins are given as a corrplot. Colour intensity relates to R value, circle size relates to the p-value (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$). Black arrows show the position of the VHL outlier patient.

of the 5-miRNA panel and the 13-feature multi-omic panel improved, though the 8-protein panel performance was worse (Fig. 6D).

10-feature multi-omic CBF panel stratifies patients into risk categories with very high accuracy

The performance of every possible combination of the 11 features from the 11-feature multi-omic PCF panel was examined using CombiROC software, with sensitivity and specificity cut-offs of 83% and 25%, respectively (Supplementary Material S7). Twenty combinations of these features were observed to meet these cut-offs, with four being the minimum number of features required to do so. The same analysis was run on the 13-feature multi-omic serum panel, with sensitivity and specificity cut-offs of 93% and 47%, respectively (Supplementary Material S7). Thirty-six combinations were identified that met these criteria, with six being the minimum number of features required to do so. A 4-feature PCF panel consisting of S100A8, LGALS3, SNORA66 and miR-216b-5p, and a 6-feature serum panel consisting of IGHV3-72, IGJ, IGHA1, PPBP, miR-3180 and miR-3180-3p, were identified as the top performing combinations. These two panels were then integrated to create a 10-feature multi-omic CBF panel, and examined in a matched patient cohort. UHC of patients into risk groups using this 10-feature panel was performed with an accuracy of 73.9% (Fig. 7A). While there were no significant correlations between PPBP and any of the other nine biomarkers ($p > 0.05$), S100A8 in the PCF had the most correlations within the panel, significantly correlating with both miR-3180-3p and miR-3180 in the serum, as well as LGALS3 in the PCF ($p < 0.05$) (Fig. 7B). MiR-216b-5p and SNORA66 had the strongest correlation, as indicated by the thicker chords, while IGHA1 and IGHV3-72 had the weakest correlation. PCA was conducted using three components as this was shown to account for 69.6% of the variance (Supplementary Material S8) (Fig. 7C). With three components, the separation of the two groups was modest, with some overlap on the centre of the plot. Importantly, the VHL outlier patient was grouped within the high-risk classification in this setting. Using LOOCV, the 10-feature multi-omic CBF panel produced an AUC of 0.970 (Sensitivity = 91.7%, Specificity = 90.9%) (Fig. 7D). This is the highest reported AUC of any panel examined here thus far. The reduced

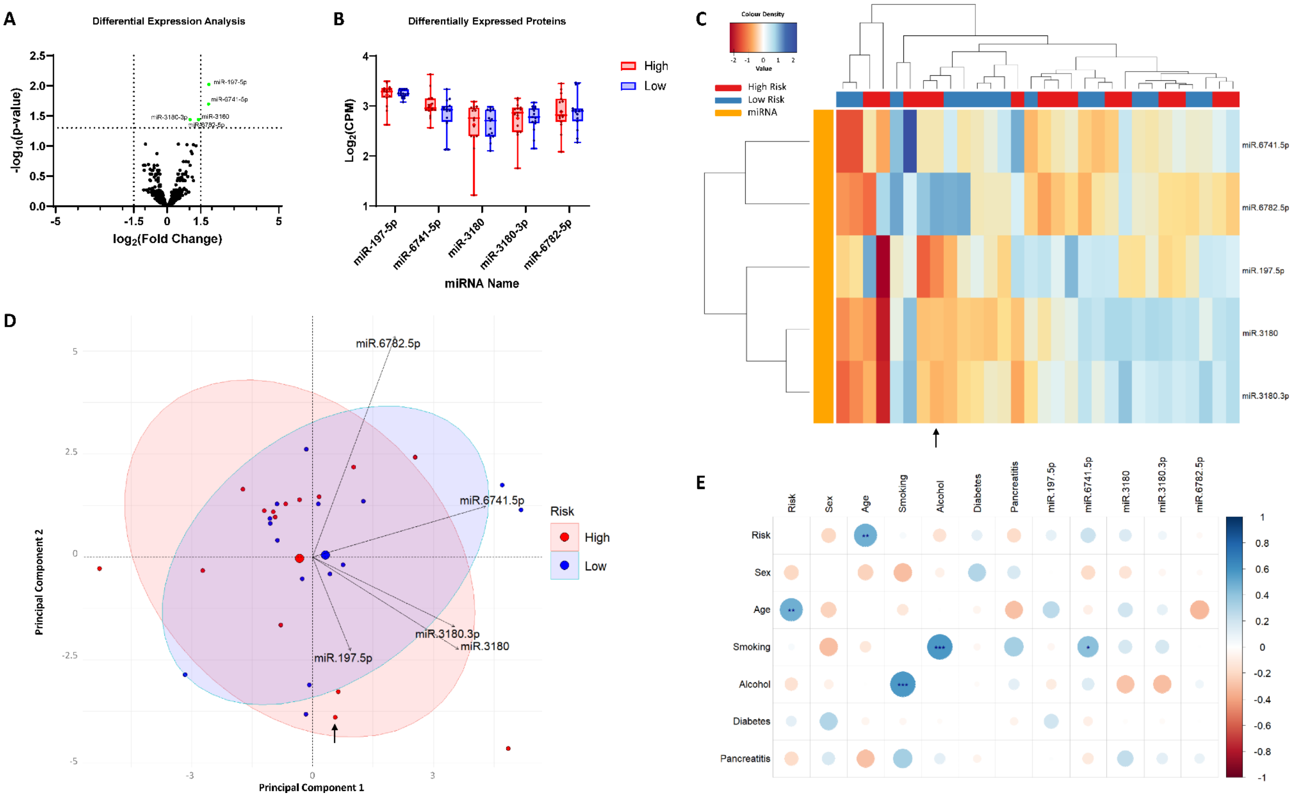

Fig. 5. Transcriptomic analysis of serum identifies five miRNAs significantly upregulated in high-risk patients compared to low-risk patients. (**A**) Differential expression analysis identified five miRNAs that were significantly differentially expressed between low- and high-risk serum samples (adj-$p < 0.05$, FDR = 0.05, s0 = 0.1). (**B**) Boxplots showing the distribution of patient expression levels of the five differentially expressed miRNAs. (**C**) UHC of patients into high- and low-risk groups based on their expression of the five differentially expressed miRNAs. Dendrograms show (top) the relatedness of the patients, and (left) the relatedness of the miRNAs. (**D**) 3D PCA using the five differentially expressed miRNAs, with biplot overlayed. Biplot scale is set to zero to ensure vectors (arrows) are scaled to represent their respective loadings. The length of each vector is proportional to the variance of the corresponding miRNA. (**E**) Spearman correlations between patient clinical data and the differentially expressed miRNA are given as a corrplot. Colour intensity relates to R value, circle size relates to the p-value (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$). Black arrows show the position of the VHL outlier patient.

4-feature PCF panel produced an AUC of 0.927 (Sensitivity = 83.3%, Specificity = 91.7%), and the 6-feature serum panel produced an AUC of 0.686 (Sensitivity = 78.6%, Specificity = 66.7%). Despite the poor performance of the 6-feature serum panel, the integration of this panel with the high-performing 4-feature PCF panel improved the overall performance. Importantly, when the VHL outlier patient was reclassified to low-risk as before, the performance of both the 10-feature multi-omic CBF panel and the 4-feature PCF panel was worsened (Fig. 7D). Reactome pathway analysis of the six proteins within this panel revealed 20 pathways to be significantly enriched (adj-$p < 0.05$, FDR = 0.05), including pathways involved in the immune system, diseases of the immune system, hemostasis, transcription, vesicle-mediated transport and signal transduction (Supplementary Material SE1).

Neither CEA nor CA19-9 improve the performance of the 10-feature multi-omic CBF panel

Finally, the utility of CA19-9 and CEA in this setting were assessed. CA19-9 levels were not significantly different in low- or high-risk patient serum ($p = 0.9826$) (Fig. 8A). CEA levels were significantly increased in the PCF of high-risk patients compared to low-risk ($p < 0.001$) (Fig. 8B). CEA was then integrated into the 10-feature multi-omic CBF panel to investigate whether its addition would improve the performance of the panel. Overall, the performance of the 10-feature multi-omic CBF panel + CEA was worse than the 10-feature multi-omic CBF panel alone, indicating that the addition of CEA to this panel does not improve its performance (Supplementary Material S9) (Fig. 8C–E).

Discussion

At present, there exists no robust single biomarker, or biomarker panel, that can effectively stratify patients with PCLs into low- and high-risk groups. This study aimed to profile the proteome and transcriptome of matched PCL patient PCF and serum to identify promising novel biomarkers for PCL patient risk stratification. We examined the utility of differentially expressed proteins and miRNAs, both as single omic-level panels and as

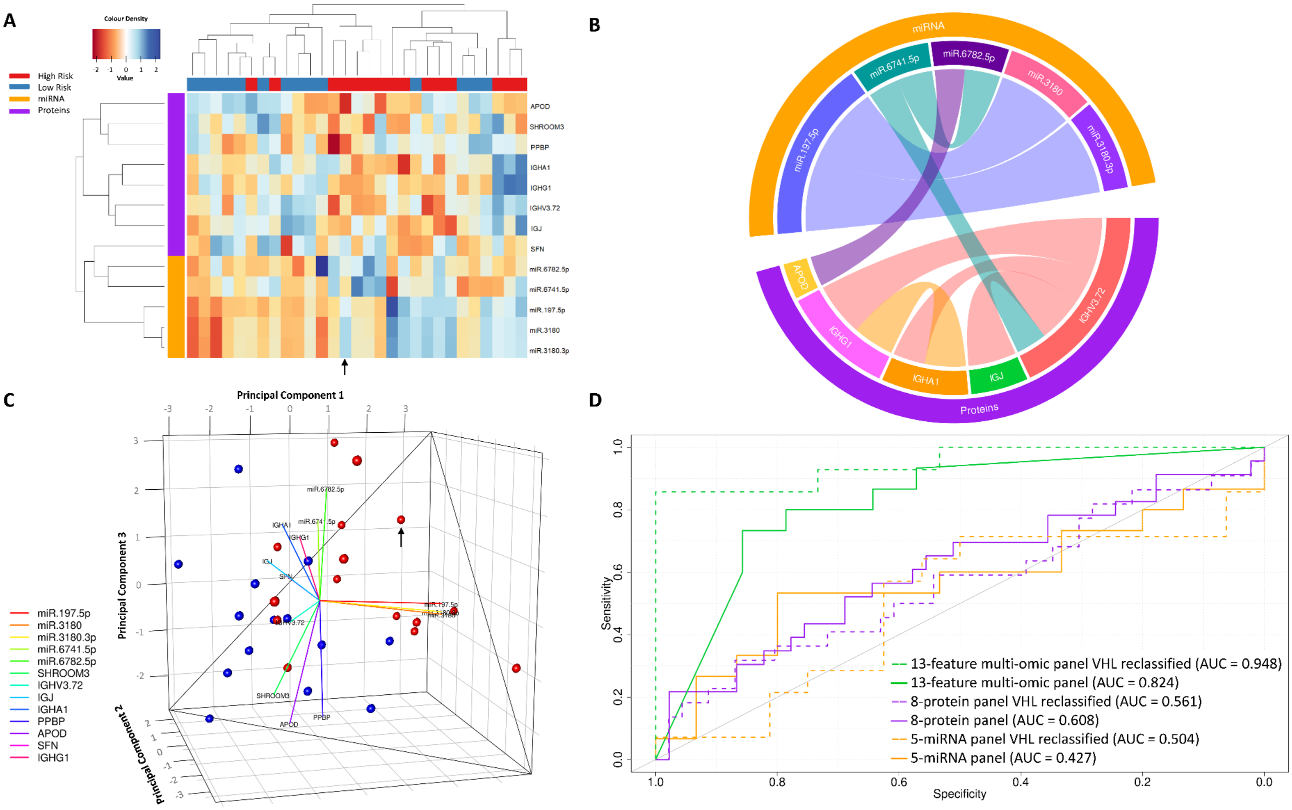

Fig. 6. Integration of the differentially expressed proteins and miRNAs generates a robust 13-feature multi-omic biomarker panel in serum. (**A**) UHC of patients into high- and low-risk groups based on their expression of the eight proteins and the five miRNAs identified as being differentially expressed, which form a 13-feature multi-omic panel. Dendrograms show (top) the relatedness of the patients, and (left) the relatedness of the miRNAs and proteins. (**B**) Chord diagram showing significant Spearman correlations ($p < 0.05$) between the eight differentially expressed proteins and five differentially expressed miRNAs. Inner chords reflect correlations between the biomarkers. Chord thickness is directly related to the strength of the correlation, with thicker chords indicating stronger correlations. (**C**) 3D PCA using the 13-feature multi-omic panel, with biplot overlayed. Biplot scale is set to zero to ensure vectors (arrows) are scaled to represent their respective loadings. The length of each vector is proportional to the variance of the corresponding protein or miRNA. (**D**) ROC curves generated from LOOCV of the miRNAs alone, the proteins alone, and the 13-feature multi-omic panel, as well as their performances when the VHL outlier patient is reclassified. Black arrows show the position of the VHL outlier patient.

multi-omic panels, within the PCF and serum to investigate the value of each biofluid in this setting. Finally, we explored the avenue of multi-omic CBF panels, by integrating the PCF- and serum-based panels, in order to understand whether layering multiple levels of biomarker data could produce improved biomarker efficacy.

Within both the PCF and the serum, the integration of multiple omic compartments to create a multi-omic panel produced the most robust results for patient PC risk classification. Indeed, while the protein and miRNA panels alone demonstrated a modest ability to classify PCL patients based on risk in both biofluids, the performance of these multi-omic panels delivered the best results overall. Interestingly, in the PCF it was shown that the 3-miRNA panel performed better than the 8-protein panel for risk classification by LOOCV, despite only one of the three miRNA significantly correlating with risk status, and demonstrating poor stratification in both the UHC and the PCA. Conversely, in the serum the 8-protein panel performed with better accuracy in the UHC, and with higher AUC in the LOOCV, compared to the 5-miRNA panel. When looking at the correlations of these serum-based biomarkers with clinical factors, one protein (SHROOM3) significantly correlated with risk, while no miRNA correlated with this factor. Given that panels of multiple biomarkers have been demonstrated to produce better results than single biomarkers alone, it would be expected that in both biofluids the 8-protein panels would be superior[14]. These contrasting results highlight the importance of the method of evaluation used for biomarker efficacy. UHC, for example, allows the datapoints to cluster based on patient expression of certain variables, and in this way it groups like-with-like to enable the visualisation of patterns[15]. Here, this analysis was used to investigate whether the patient cohorts would separate into groups based on their expression of these factors. While separation of patient risk groups in the PCA was the worst out of all the panels when using the 3-miRNA panel in the PCF, it is important to note that such methods can be greatly influenced by poor performing variables, especially when there are so few to begin with. Indeed, it was shown that when using this 3-miRNA panel to train and test a LOOCV model, this model performs modestly.

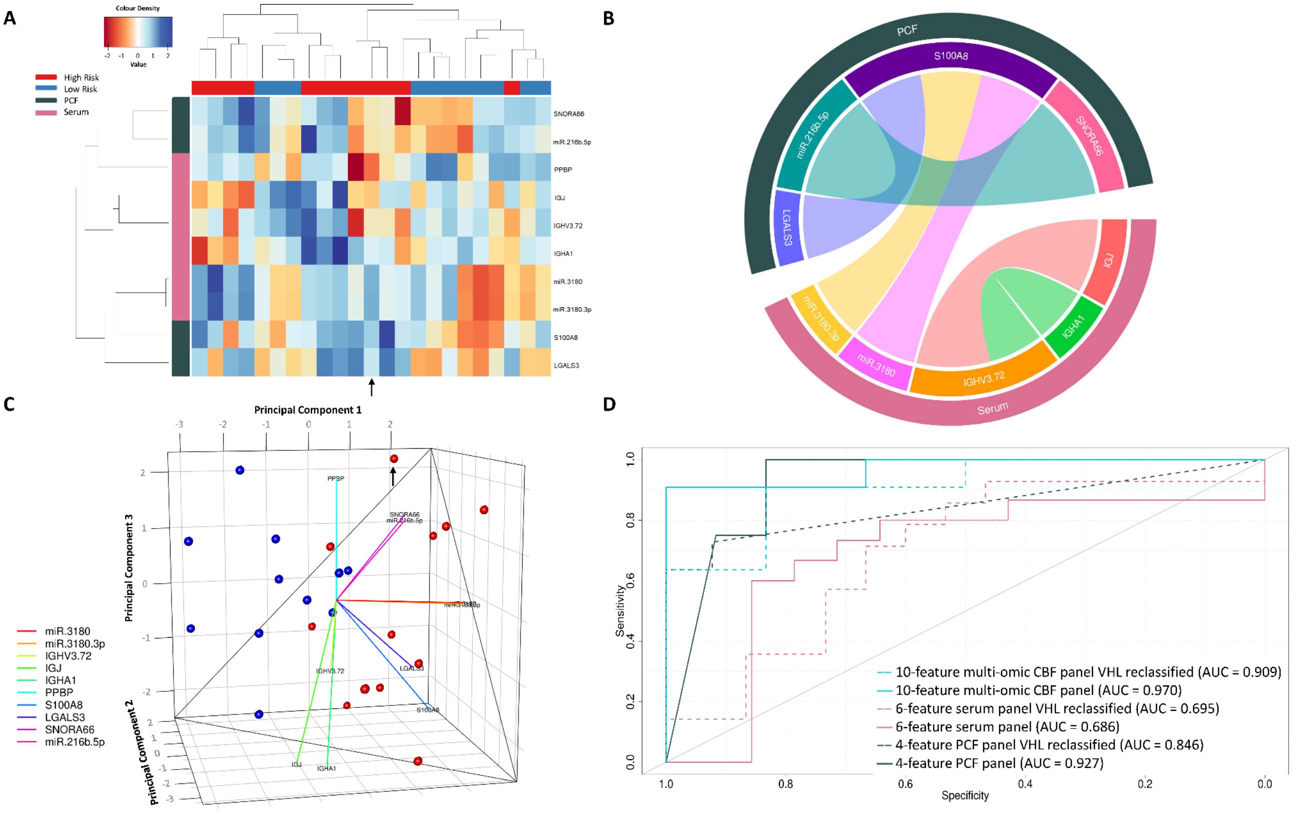

Fig. 7. Integration of the reduced PCF and serum panels generates a robust 10-feature multi-omic CBF biomarker panel. (**A**) UHC of patients into high- and low-risk groups based on their expression of the reduced serum panel and the reduced PCF panel. Dendrograms show (top) the relatedness of the patients, and (left) the relatedness of the biomarkers. (**B**) Chord diagram showing significant Spearman correlations ($p < 0.05$) between the serum biomarkers and PCF biomarkers. Inner chords reflect correlations between the biomarkers. Chord thickness is directly related to the strength of the correlation, with thicker chords indicating stronger correlations. (**C**) 3D PCA using this 10-feature multi-omic CBF panel, with biplot overlayed. Biplot scale is set to zero to ensure vectors (arrows) are scaled to represent their respective loadings. The length of each vector is proportional to the variance of the corresponding biomarker. (**D**) ROC curves generated from LOOCV of the reduced PCF panel alone, the reduced serum panel alone, and the 10-feature multi-omic CBF panel, as well as their performances when the VHL outlier patient is reclassified. Black arrows show the position of the VHL outlier patient.

PCA analysis allows the visualisation of large datasets in smaller components or dimensions via dimensionality reduction, clustering similar samples together and aligning highly correlated variables[16]. In this way, PCA finds the majority of its utility in ‘long’ datasets with dimensionality issues, where there are more variables than the number of samples and as such, it can be difficult to discern the individual effects of a single variable[17]. Here, the 3-miRNA panel from the PCF was dimensionally reduced to two components using PCA, which is not a common practice for such small datasets. However, this approach highlighted the outlier among the three biomarkers, SNORA66, and illustrated that the majority of variance could be accounted for by miR-216a-5p and miR-216b-5p, suggesting that SNORA66 is perhaps the least important component of this 3-miRNA panel. Conversely, the 8-protein panels in both the PCF and the serum performed well in all analyses, despite the fact that while all eight proteins in the PCF significantly correlated with risk, just one of the proteins in the serum had a significant correlation with risk. This is likely due to the number of variables in these panels, with eight variables helping to distinguish and restructure the data according to their expression. In this way, larger panels allow for better handling of outlier patients, as when one variable becomes dysregulated, the others within the panel can compensate for this. However, the utility of both the miRNA- and protein-based panels alone are limited, with modest separation being seen in the UHC and PCA, and modest AUC values being obtained in the LOOCV. The integration of these omic compartments to form multi-omic biomarker panels is where the true potential of these biomarkers can be seen. Indeed, current trends in biomarker identification lean towards the creation of multi-omic panels that can better control for the complexity of the disease[4]. By encompassing factors from multiple biological levels, multi-omic biomarker panels can better compensate for the dysregulation of individual biological compartments. In fact, substantial improvements in results from either omic level alone can be seen through the use of the 11-feature multi-omic panel in the PCF and the 13-feature multi-omic panel in the serum. Importantly, the identification of one outlier patient highlights the limitation of these panels. Indeed, in the initial models with this patient classified as high-risk, this datapoint can be frequently seen as an outlier.

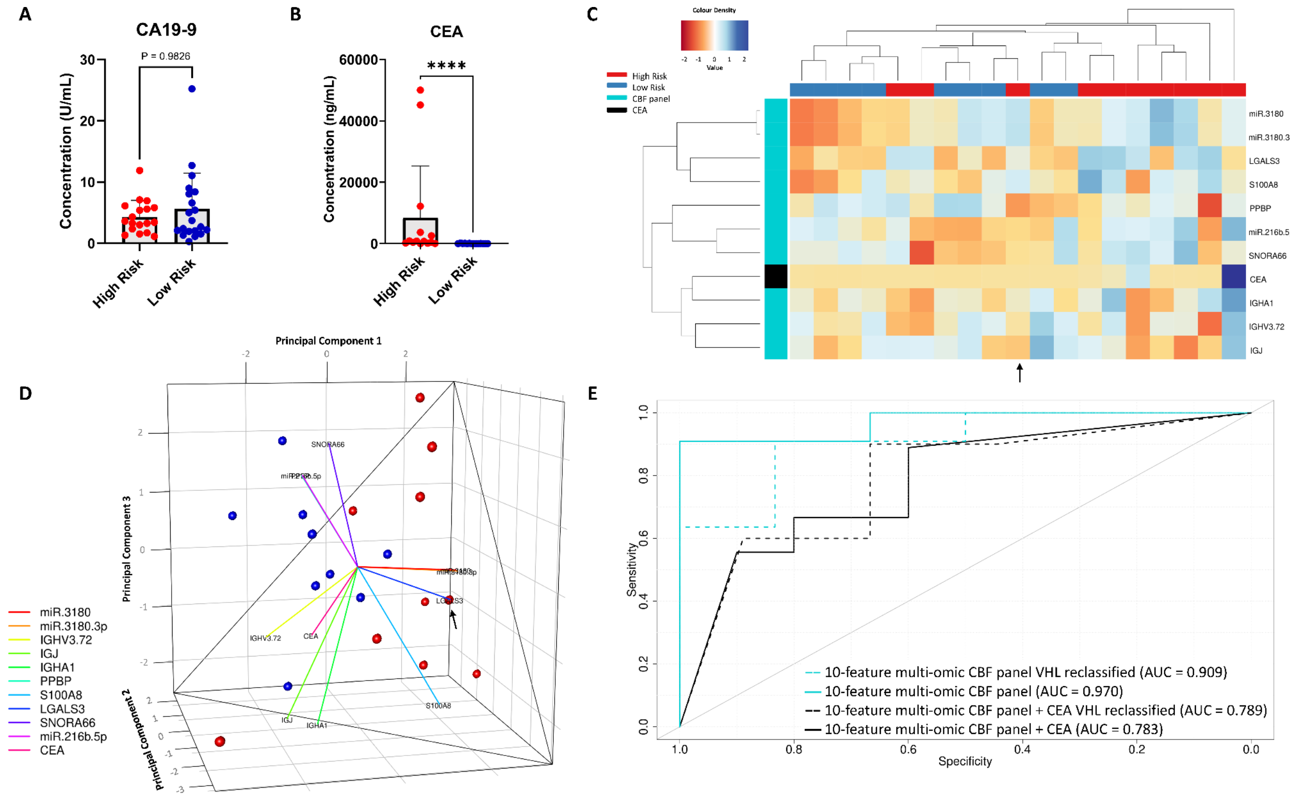

Fig. 8. Neither CA19-9 nor CEA improve the performance of the 10-feature multi-omic CBF panel. (**A**) Serum concentration of CA19-9 (U/mL) in high-risk (red) and low-risk (blue) patients. Mann–Whitney test. Data are presented as mean ± SEM. (**B**) PCF concentration of CEA (ng/mL) in high-risk (red) and low-risk (blue) patients. Mann–Whitney test. Data are presented as mean ± SEM, ****$p < 0.0001$. (**C**) UHC of patients into high-risk (red) and low-risk (blue) groups based on their expression of the 10-feature multi-omic CBF panel and CEA. Dendrograms show (top) the relatedness of the patients, and (left) the relatedness of the biomarkers. (**D**) 3D PCA using the 10-feature multi-omic CBF panel + CEA, with biplot overlayed. Biplot scale is set to zero to ensure vectors (arrows) are scaled to represent their respective loadings. The length of each vector is proportional to the variance of the corresponding biomarker. (**E**) ROC curves generated from LOOCV of the 10-feature multi-omic CBF panel (light blue lines), and the 10-feature multi-omic CBF panel + CEA (black), as well as their performances when the VHL outlier patient is reclassified (dashed lines). Black arrows show the position of the VHL outlier patient.

In the UHC for the 11-feature multi-omic panel in the PCF specifically, it is the only datapoint that is incorrectly clustered. When reclassified, the model performance improves for all analyses in the PCF, with improved separation of the two groups in the PCA, and increased AUC for the LOOCV. In the serum, the performance of the 5-miRNA panel and the 13-feature multi-omic panel is improved when the VHL patient is reclassified, while the 8-protein panel performance worsened. Here, it is important to highlight the substantial alteration to model performance that just one patient could make. Furthermore, while in this case the presence of a *VHL* mutation stood out clinically as a potential confounding factor, it may not be appropriate to reclassify or remove this patient from this analysis as their original classification as high-risk was based on the same guidelines as all other patients. These data emphasise the need for validation of these results in a larger, independent patient cohort where longitudinal progression data can confirm whether high-risk patients progressed to PC.

The CBF integration of both multi-omic panels from PCF and serum produced the highest classification accuracy of any panel examined, without the need to reclassify the outlier VHL patient. Indeed, while the integration of proteomic and transcriptomic biomarkers to create a multi-omic panel in both the PCF and the serum separately produced substantially better risk stratification than either omic level alone, it is clear from these results that the layering of data from multiple biological compartments could be the key to the generation of more robust biomarkers. Here, CombiROC software was used to interrogate both the 11-feature multi-omic panel in PCF and the 13-feature multi-omic panel in the serum. Using appropriate cut-offs, these panels were reduced down to the least number of biomarkers that would still produce highly sensitive results in order to allow the integration of these biomarkers to create a CBF panel. In the same way that multi-omic panels have the potential to encapsulate the complexity of disease, and have the unique ability to control for the dysregulation of one omic compartment or factor via compensation of other biomarkers within the panel[4], CBF panels present a new and exciting progression from this. By selectively reducing the PCF and serum panels, and generating a new panel that consists of two omic compartments, as well as two distinct biofluids from the same patient, high sensitivity, specificity and AUC metrics were achieved. Furthermore, the performance of this panel is not improved by the reclassification of the VHL patient, further emphasising its utility in this setting. Importantly, while the sensitivity cut-off was higher in the serum panel (93%) compared to the PCF panel (83%), more features were required to achieve this high performance in the serum than in the PCF, perhaps indicating that PCF-based biomarkers perform better than serum-based biomarkers in this cohort, which would not be unexpected given the direct proximity of PCF to the PCL in question. Also of note, is the ratio of proteins to miRNAs in these two reduced panels, with two proteins (S100A8 and LGALS3) and two miRNAs (SNORA66 and miR-216b-5p) making up the PCF panel, while four proteins (IGHV3-72, IGJ, IGHA1 and PPBP) and two miRNAs (miR-3180 and miR-3180-3p) make up the serum panel. Interestingly, SHROOM3, the only significant protein in the serum that also significantly correlated with patient risk, was not included in the reduced panel. Importantly, CA19-9 levels were shown to have no significant difference in the serum of low- and high-risk patients, indicating a limited utility in the risk stratification setting for this FDA-approved diagnostic biomarker for PC. While CEA was significantly differentially expressed in the PCF of low- and high-risk patients, the addition of CEA to the 10-feature multi-omic CBF panel worsened the LOOCV performance of the panel.

Among the 10 biomarkers within the final CBF panel, several of these factors have been previously identified as dysregulated in the pancreatic setting. Within the transcriptomic compartment, miR-216-5p has been studied extensively in the context of PC, as it is a pancreas-specific miRNA[18]. MiR-216b-5p has been demonstrated to function as a tumour suppressive miRNA in pancreatic tissues by repressing PC cell proliferation, inducing apoptosis and cell cycle arrest, and supressing invasive and migratory capabilities[19,20]. As such, its expression is generally reduced in PC tissues, and this is associated with poor prognosis[19,21]. Interestingly, despite the downregulation of this miRNA in PC, miR-216b-5p has previously been shown to be increased in high-risk IPMNs compared to low-risk, which aligns with the elevated levels identified in the PCF of high-risk patients in this study[22]. SNORA66, mir-3180 and miR-3180-3p, unfortunately, remain largely unstudied in PC or PCLs. MiR-3180 has been suggested as a potential biomarker of hepatitis B virus infection persistence[23], while miR-3180-3p was shown to be significantly upregulated in the serum of chemotherapy (cisplatin) resistant gastric cancer patients compared to chemotherapy sensitive patients, and significantly correlated with high TNM stage[24]. Within the proteomic compartment, no studies to date have reported on IGHV3-72 in pancreatic disease, however, the levels of this protein in plasma exosomes have been shown to have utility in distinguishing lung adenocarcinoma from lung squamous cell carcinoma[25]. Importantly, increased LGALS3 expression has been observed as an early PC event, with LGALS3 expression shown to be 1.5-fold higher in chronic pancreatitis tissues compared to healthy controls, but up to 6.5-fold higher in PC tissue, increasing incrementally as the disease progresses[26,27]. Indeed, both LGALS3 mRNA and protein levels have been shown to be significantly increased in PC tissue compared to healthy controls[28]. Interestingly, LGALS3, along with LCN2, REG1A and S100A8, has been previously detected in patient PCF, with S100A8 and LCN2 also being detected in PCF cell pellets, though no indication as to the level of expression or differential expression between controls were reported[6]. High S100A8 expression in pancreatic ductal fluid has been demonstrated to predict worse disease-free and overall survival in late-stage PC patients[29], with the current study providing evidence that S100A8 is elevated in the PCF of high-risk patients. Furthermore, S100A8 was shown to be overexpressed in PC tumours compared to normal and pancreatitis tissues[30]. IGHA1 plays a key role in immunoglobulin receptor binding activity, and has been measured previously in the PCF of both chronic pancreatitis and non-pancreatitis patients[31]. Importantly, IGHA1 has also been measured in normal pancreatic FFPE tissue samples, but was not detected in chronic pancreatitis or PC FFPE tissue specimens[32]. This study demonstrated lower levels of IGHA1 in high-risk PCF, suggesting that expression may be lost during disease progression. IGJ has previously been identified, via proteomic evaluation of pancreatic patient plasma, as being upregulated in PDAC plasma compared to healthy controls[33]. While this does not align with the results reported in this study, discrepancies in the concentrations of numerous proteins in serum versus plasma have been reported, and the two fluids are therefore not directly comparable[34]. PPBP, also known as CXCL7, is a neutrophil chemoattractant that has been demonstrated by Matsubara et al.[35] to be significantly decreased in the plasma of PC patients compared to healthy controls. However, Pan et al.[33] found PPBP levels to be elevated in chronic pancreatitis and PDAC plasma compared to healthy controls. In a 2021 study, Kim et al.[36] generated a plasma-based multi-biomarker panel consisting of 14 proteins, including PPBP, that could distinguish PDAC from controls with AUC values of up to 0.977. Here, Kim et al.[36] found PPBP levels to be increased in PDAC patients compared to controls. These three studies highlight the importance of validation of biomarkers across independent patient cohorts, and the vast differences that can be seen in biomarker expression profiles across different patient cohorts. Overall, six out of the ten biomarkers have previously been studied in the pancreatic setting, with the remaining four being reported in this study for the first time. Interestingly, pathway enrichment of the proteins within this panel also revealed significant associations with the immune system, providing evidence of the potential involvement of these features in processes such as neutrophil degranulation, regulation of the TLR signalling cascade, and diseases of the immune system.

While the various biomarker panels generated in this study have shown promise in the risk stratification setting for patients with PCLs, it is important to note that this study is not without its limitations. While the patients included in this analysis were stratified as low- or high-risk based on the European evidence-based guidelines for PCL management, these guidelines are one of several that exist globally. Indeed, there is a lack of consensus among clinicians as to the characteristics that constitute a low- or high-risk of PC development. As such, while these data were generated using the European guidelines, they may not align with other guidelines globally, and this is a major issue with current PCL research. Further to this, another limitation of this study is the lack of a validation cohort to verify these results. While validation is an important part of any biomarker research, it was not possible to generate a validation cohort over the course of this study. Indeed, in order to validate these results another dataset consisting of proteomic and transcriptomic data for both PCF and serum of PCL patients would be needed. As such, while the results presented are very promising, these data remain to be validated in an independent patient cohort. This lack of readily available matched datasets remains one of the major caveats with the use of multi-omic data at present, hampering validation of results through standard data mining efforts.

Overall, the various approaches used to analyse these data highlight the strengths and weaknesses of the panels identified in this study, and demonstrate that while UHC and PCA are useful for interrogating datasets, training and testing models that are developed for the examination of biomarkers, such as LOOCV, gives the best sense of biomarker performance. Metrics such as AUC value, sensitivity and specificity are the most important in this context, and should therefore be given the most weight. The results reported here not only describe the dysregulation of proteins and miRNAs in pancreatic disease that have not previously been seen, but also demonstrate their potential utility as biomarkers of patient PC risk in this PCL cohort. Promising multi-omic panels have been identified in both the PCF and the serum that have the potential to classify patients based on their risk of PC with high accuracy. Using novel CombiROC software these two multi-omic panels were reduced and integrated to create a CBF multi-omic panel that could stratify patients with improved accuracy compared to either multi-omic panel alone. This research not only highlights promising novel biomarkers of patient PC risk stratification, but provides a unique methodology for the generation of biomarker panels across biological samples. Importantly, these data also highlight potential caveats to biomarker panel design and analysis, and as such demonstrate the importance of careful and extensive validation of results in novel patient cohorts. While these data remain to be further validated in an independent patient cohort, the outputs reported here give hope not just for the establishment of robust biomarkers in pancreatic disease, but for biomarker research as a whole. Lastly, this work showcases the vast diversity of dysregulated components to be found within the PCF and serum of PCL patients. Given the expansive research conducted to date demonstrating the various factors within the PCF, it is important to understand how these factors become dysregulated and what role they may have in the progression of PCLs to PC.

Materials and methods

Patient sample collection

PCF and peripheral blood serum were collected prospectively from patients presenting with a PCL in one of three tertiary hospitals in Dublin, Ireland (Tallaght University Hospital, St. James's Hospital and St. Vincent's University Hospital) from July 2019 to July 2022. PCF samples (n = 32) were collected via endoscopic ultrasound-guided fine-needle aspiration, with serum samples (n = 68) being collected prior to cyst puncture. As serum samples were taken prior to cyst puncture, matched serum and PCF were only obtained for n = 32 patients due to a low volume of PCF present, difficulty puncturing the cyst, or complications during endoscopy. Demographic information for all cohorts are provided in Supplementary Materials S10, S11, and S12. PCF CEA levels were assessed clinically for all patients as part of routine cytology. Patients were stratified into low- and high-risk groups for PC development by the clinical team using the 2018 European evidence-based guidelines on pancreatic cystic neoplasms[37]. Pathology for these patients was not available for a definitive classification as tissue is not taken as part of the routine clinical workup for these patients. Importantly, one patient within the cohort possesses a *VHL* mutation, and as such this patient was classified as high-risk due to their genetic predisposition to PCLs and PC, despite their PCL receiving a low-risk classification. As such, the data presented here were examined, where appropriate, with this patient classified as both low- and high-risk, and changes in panel performance were discussed. A detailed illustration of the methods can be found in Supplementary Material S13.

Materials

All chemicals and reagents used were purchased from Sigma-Aldrich (Wicklow, Ireland), unless otherwise stated. Triton-X100 (Product Code 306324N) was purchased from British Drug Houses Ltd (London, UK).

HTG EdgeSeq miRNA whole transcriptome sequencing of PCF and serum

PCF and serum samples were processed in accordance with OP-00034, HTG EdgeSeq processing. Serum samples were processed with a modified protocol, using a 1:2 dilution with Plasma Lysis Buffer in place of the Biofluid Lysis Buffer to overcome the presence of inhibitors in the samples. Target capture was done by HTG EdgeSeq chemistry. The library was prepared in accordance with OP-00035, HTG EdgeSeq PCR processing. Clean-up procedures were performed according to OP-00037, HTG EdgeSeq AMPure clean-up of Illumina Sequencing Libraries. The library was quantified in accordance with OP-00079, HTG EdgeSeq KAPA Library Quantification for Illumina Sequencing. All samples and controls were quantified in triplicate.

Following qPCR quantification, the HTG EdgeSeq RUO Library Calculator (v3.2) was used to ensure there was a sufficient concentration of sample for library pooling and to determine the appropriate dilutions for building the library pool. Use of the HTG EdgeSeq RUO Library Calculator was guided by the HTG EdgeSeq RUO Library Calculator Instructions for Use (P/N 10290200). All samples processed within this study had sufficient PCR product to be pooled for sequencing. The HTG EdgeSeq RUO library calculator was also used to determine the volume and specific type of denaturation reagents to be used for the library.

The sequencing was performed on the Illumina NextSeq sequencer in accordance with OP-00093, HTG EdgeSeq Illumina NextSeq sequencing. The sequencing data on miRNA expression of target genes were imported into HTG EdgeSeq Parser software (v5.3.0.7184). The HTG EdgeSeq Reveal Application (v3.1.0) was utilized to quality check and normalize data. Post-sequencing quality control (QC) metrics were used to detect sample failure modes. Data were returned from the sequencer in the form of demultiplexed FASTQ files, with four files per original well of the assay. The HTG EdgeSeq Parser was used to align the FASTQ files to the probe list to collate the data. All multi-tissue control correlations passed Pearson and Spearman correlation acceptance criteria of ≥ 0.85, and all samples passed the HTG EdgeSeq post-sequencing metrics. Transcriptomic data have been uploaded to the Gene Expression Omnibus (GSE280768 and GSE280772).

Preparation of PCF samples for LC–MS

The protein concentration of PCF samples was assessed by Pierce™ BCA Protein Assay (Cat. No. 23225) (ThermoFisher, UK) as per the manufacturer's instructions. Hydrophobic and hydrophilic Speedbead Magnetic Carboxylate Modified Particles (GE Healthcare, Cat. No. 45152105050250 and 65152105050250) (Cytiva, MS, USA) were combined in a ratio of 1:1 (v/v), rinsed on the DynaMagTM (Cat No. 12321D) (ThermoFisher, London, UK), and reconstituted in the starting volume of LC–MS grade water. Volumes representing 50 µg of protein from each sample were combined with an equal volume of 2X sample buffer [300 mM NaCl, 100 mM tris (pH 8.0), 3 mM MgCl2, 2% Triton-X100 and 1 tablet of Complete Mini Protease Inhibitor in LC–MS grade water]. After incubating at 4 °C with intermittent agitation for 20 min, samples were centrifuged before adding lysis buffer [6 M urea, 2 M thiourea and 50 mM MOPS in LC–MS grade water] and 0.2 M dithiothreitol. Samples were incubated on a thermoshaker at 700 RPM for 15 min at 30 °C, then cooled to RT before adding 0.4 M iodoacetamide and incubating on a thermoshaker at 700 RPM for 15 min at room temperature (RT) in the dark. After incubating, 100% acetonitrile was added to each sample, followed by magnetic bead mix, and samples were placed on a rotation mixer at RT for 1 h. Samples were then rinsed on the DynaMagTM stand. Samples were processed in this way by rinsing with both 70% (v/v) ethanol, and 100% acetonitrile, consecutively. After the final rinse, 50 mM ammonium bicarbonate was added to each tube, followed by Promega Sequencing Grade Modified Trypsin (Product Code V5111) (MyBio Ltd, Kilkenny, Ireland). Samples were then incubated overnight on a thermoshaker at 500 RPM and 37 °C. Samples were then quick spun and resuspended with more fresh magnetic bead mix, along with 100% acetonitrile. Tubes were then incubated on the rotation mixer for 18 min at RT, before being rinsed on the DynaMagTM stand. Samples were then rinsed with 100% acetonitrile before being removed from the stand and LC–MS grade water added to elute the peptides from the beads. The magnetic beads were vortexed intermittently for 5 min at RT before being placed on the DynaMagTM stand a final time for 5 min. The eluted peptide supernatant was then carefully transferred to a fresh tube. Peptide concentrations of the elutants were assessed by PierceTM Quantitative Colorimetric Peptide Assay (Cat. No. 23275) (ThermoFisher, UK) as per the manufacturer's instructions. Peptide sample dilutions of 100 ng/mL were prepared in 0.1% (v/v) formic acid in mass-spec vials.

Label-free LC–MS/MS analysis of PCF

Samples were run on a Thermo Scientific Q Exactive mass spectrometer coupled to a Dionex Ultimate 3000 (RSLCnano) chromatography system to perform the LC–MS/MS analysis of PCF samples in the Mass Spectrometry Facility, Conway Institute of Biomolecular and Biomedical Research, University College Dublin. The tryptic peptides were separated on a reversed-phase C18 column packed in-house (8 cm × 75 µm ID; C 18, 3.0 µm (ReproSil-Pur 120 Dr Maitsch GmbH.)) and separated at a constant flow rate of 250 nL/min by an increasing acetonitrile gradient. Mobile phases were 0.5% (v/v) acetic acid, 2% (v/v) acetonitrile, 97.5% (v/v) water (phase A), and 0.5% (v/v) acetic acid, 2% (v/v) water, 97.5% (v/v) acetonitrile (phase B). The peptides were separated by a gradient starting from 1% of mobile phase B and increased linearly to 30% for 58 min at a flow rate of 250 nL/min. The mass spectrometer was operated in data dependent TopN 12 mode, with the following settings: mass range 320-1600 Th; resolution for MS1 scan 70,000; AGC target 3e6; resolution for MS2 scan 17,500; AGC target 5e4.

Preparation of serum samples for LC–MS

Immunodepletion of serum samples was carried out using the Proteome Purify 12 Human Serum Protein Immunodepletion Resin kit (Cat. No. IDR012) (R&D Systems, MN, USA) as per the manufacturer's instructions. Following this, sample suspension was placed into the upper chamber of a Corning™ Costar™ Spin-X™ Centrifuge Tube Filter (Product Code 10310361) (Fisher Scientific, Dublin, Ireland) and centrifuged for 2 min at 2000 × *g*. The immunodepleted elutants were moved to a fresh tube and combined with − 20 °C 100% acetone and stored at − 20 °C overnight. After this time, samples were centrifuged and the supernatant discarded to waste, before adding − 20 °C 50% (v/v) acetone. Samples were centrifuged again and the supernatant was discarded to waste before adding more − 20 °C 50% (v/v) acetone and centrifuging as before. The pellet was allowed to air dry for 24 h at RT. Once dry, the protein pellets were immediately processed using the PreOmics iST 96 × kit (Product code P.O.00027) (PreOmics GmbH, Munich, Germany) as per the manufacturer's instructions.

Label-free LC–MS/MS analysis of serum

An UltiMate 3000 nano RSLC (ThermoFisher, UK) system interfaced with an Orbitrap Fusion Tribrid Mass Spectrometer (ThermoFisher, UK) was used to perform the LC–MS/MS analysis of serum samples in the Proteomics Facility of the National Institute for Cellular Biotechnology, Dublin City University. A volume of 2 µL from each sample was loaded onto a PepMap100, C18, 300 µm × 5 mm trapping column using a flow rate of 25 µL/min with 2% (v/v) acetonitrile and 0.1% (v/v) trifluoroacetic acid in LC–MS grade water for 3 min. Each sample was then resolved onto an Acclaim PepMap 100, 75 µm × 50 cm, 3 µm analytical column. A binary gradient of: solvent A (0.1% (v/v) formic acid in LC–MS grade water) and solvent B (80% (v/v) acetonitrile, 0.08% (v/v) formic acid in LC–MS grade water), using 2–32% B for 50 min, 32–90% B for 5 min, and holding at 90% for 5 min at a flow rate of 300 nL/min was used to elute peptides. A column temperature of 47 °C and a voltage of 2.0 kV was used for peptide ionization. Data-dependent acquisition was performed using a full scan range of 380–1500 m/z. The Orbitrap mass analyser with a resolution of 120,000 (at m/z 200), a maximum injection time of 50 ms and an automatic gain control (AGC) value of 4.0 × E5 was used to perform scans. A top-speed acquisition algorithm was used to determine the number of selected precursor ions for fragmentation.

Selected precursor ions were isolated in the quadrupole using an isolation width of 1.6 Da. A dynamic exclusion was applied to analysed peptides after 60 s and only peptides with a charge state between 2 + and 7 + were analysed. Precursor ions were fragmented using higher energy collision-induced dissociation with a normalized collision energy of 28%. The resulting MS/MS ions were measured in the Orbitrap analyser with a resolution of 30,000 (at m/z 200). MS/MS scan conditions were typically the following: a targeted AGC value of 5 × E4 and a maximum fill time of 300 ms.

Protein identification from LC–MS/MS

Data from both PCF and serum LC–MS/MS analysis were searched against the Human Reference Proteome (reviewed entries) downloaded from Uniprot.org (21-05-2021), using MaxQuant (v1.6.17.0). Label-Free Quantitation was selected as was the Match between Runs option. The following parameters were selected for the search—Fixed Mod: carbamidomethylation; Variable Mods: methionine oxidation, acetyl (protein N-term); Trypsin/P digest enzyme; Precursor mass tolerances 4.5 ppm; Fragment ion mass tolerances 20 ppm; Peptide FDR 1%; Protein FDR 1%. Proteomic data have been uploaded to the PRIDE database (PXD057661 and PXD057299).

Quantification of serum CA19-9 concentrations via sandwich ELISA

Soluble CA19-9 concentrations in patient serum were measured using the Human CA19-9 PharmaGenie ELISA kit (Cat. Code: SBRS0338) from Assay Genie Ltd (Dublin, Ireland) as per the manufacturer's instructions. The absorbance at 450 nm was measured using the GloMax Explorer microplate reader (Promega, WI, USA).

Bioinformatic analysis of omics data

Both PCF and serum proteomic data obtained from LC–MS were cleaned and normalised using Perseus software (v1.6.15.0)[38]. Briefly, label-free quantification (LFQ) intensity data were filtered to remove reverse sequences, potential contaminants and proteins that were only identified by peptides carrying one or more modified amino acids ("only identified by site"). The data were then filtered based on valid values, and proteins with zero values in more than 30% of samples were removed. The data were then log2 transformed, normalised using a linear transformation, and imputed to replace missing values from the normal distribution. Differential expression analysis with Benjamini-Hochberg corrections was conducted in Perseus using the built-in 'edgeR' package from RStudio.

PCF and serum transcriptomic data were obtained from HTG Molecular in CPM normalised form. Normalised transcriptomic data were loaded into RStudio (v21.09.0) and differential expression analysis using Empirical Bayes Statistics for Differential Expression was conducted using packages 'readxl' (v1.4.1), 'edgeR' (v3.32.1) and 'DESeq' (v1.30.1). Multiple comparisons for differential expression analysis was corrected using Benjamini-Hochberg corrections.

Power calculations suggest that a minimum of 15 patients per cohort is required for sufficient statistical power ($z = 0.8$, $\alpha = 0.05$, $\beta = 0.2$, $k = 1$)[39]. Box plots and volcano plots of significantly differentially expressed factors were created in GraphPad Prism (v9.5.0). Processed proteomic and transcriptomic data for both PCF and serum were scaled individually before being integrated to create a single data matrix. UHC with supporting heatmap and dendrograms were generated in RStudio using packages 'edgeR' (v3.32.1), 'cluster' (v2.1.4), 'purrr' (v0.3.4), 'dendextend' (v1.15.2), 'dplyr' (v1.0.9), 'ggplot2' (v3.3.5), 'ComplexHeatmap' (v2.6.2), 'RColorBrewer' (v1.1-3), 'gplots' (v3.1.1), 'pheatmap' (v1.0.12) and 'factoextra' (v1.0.7). Corrplots illustrating the correlations between patient clinical data and omic factors were created in RStudio using packages 'Hmisc' (v4.7-2), 'heatmap3' (v1.1.9), 'pheatmap' (v1.0.12), 'plot.matrix' (v1.6.2), 'RColorBrewer' (v1.1-3), 'gplots' (v3.1.1), 'corrplot' (v0.90) and 'ggcorrplot' (v0.1.3). Clinical data were converted to binary code where appropriate using the key in Supplementary Material S14. PCA was conducted in RStudio using packages 'tidyverse' (v1.3.1), 'ggplot2' (v3.3.5), 'factoextra' (v1.0.7), 'rgl' (v0.108.3) and 'plot3D' (v1.4). LOOCV and corresponding ROC plots were created in RStudio using packages 'tidyverse' (v1.3.1), 'dplyr' (v1.0.9), 'plyr' (v1.8.7), 'klaR' (v1.7-1) and 'caret' (v6.0-93). Predictive linear classification models were used in the LOOCV. All linear models (used in the LOOCV, differential expression analysis, and in the Spearman correlations) report on the linear relationships between variables, and how changes in one variable can impact the other. *P*-values report on whether these relationships are statistically significant, and whether there is a significant association between the variables. For example, where increases in certain protein levels are significantly associated with high-risk PCL patients.

Assessment of optimal biomarker combinations was conducted using CombiROC software (v1.2)[40]. Data were scaled in RStudio and processed using a linear transformation to ensure no negative values were present before being brought into the CombiROC software. Using the graphics function, the minimum number of biomarker features was set to 1 in order to evaluate the number of biomarkers that produced the best results. The test signal cut-off was calculated as the mean of the control group plus the standard deviation, rounded to the nearest whole number to be compatible with the software. For PCF the cut-off was set to 3, for serum this was set to 4. PCF sensitivity and specificity limitations were set at 83% and 25%, respectively. Serum sensitivity and specificity limitations were set at 93% and 47%, respectively.

Pathway mapping

Proteins in the final CBF panel were entered into the Reactome Pathway Browser (v 3.7) analysis tool to evaluate potential pathways that these features may be involved in[41]. Significantly enriched pathways ($p\text{-value} < 0.05$, $FDR = 0.05$) were sorted under the platform's hierarchy for ease of interpretation.

Data availability

Study data have been uploaded to online repositories. Transcriptomic data have been uploaded to the Gene Expression Omnibus (GSE280768 and GSE280772). Proteomic data have been uploaded to the PRIDE database (PXD057661 and PXD057299).

Received: 7 October 2024; Accepted: 17 December 2024
Published online: 02 January 2025

References

1. Chen, C. et al. Applications of multi-omics analysis in human diseases. *MedComm* **4**(4), e315 (2023).
2. Vandereyken, K., Sifrim, A., Thienpont, B. & Voet, T. Methods and applications for single-cell and spatial multi-omics. *Nat. Rev. Genet.* **24**(8), 494–515 (2023).
3. Babu, M. & Snyder, M. Multi-omics profiling for health. *Mol. Cell. Proteom.* **22**(6), 100561 (2023).
4. Kane, L. E., Mellotte, G. S., Conlon, K. C., Ryan, B. M. & Maher, S. G. Multi-Omic biomarkers as potential tools for the characterisation of pancreatic cystic lesions and cancer: Innovative patient data integration. *Cancers* **13**(4), 769 (2021).
5. Siegel, R. L., Giaquinto, A. N. & Jemal, A. Cancer statistics, 2024. *CA Cancer J. Clin.* **74**(1), 12–49 (2024).
6. Park, J. et al. Proteome characterization of human pancreatic cyst fluid from intraductal papillary mucinous neoplasm by liquid chromatography/tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* **31**(20), 1761–1772 (2017).
7. Hasan, A., Visrodia, K., Farrell, J. J. & Gonda, T. A. Overview and comparison of guidelines for management of pancreatic cystic neoplasms. *World J. Gastroenterol.* **25**(31), 4405 (2019).
8. Marchegiani, G. et al. Guidelines on pancreatic cystic neoplasms: major inconsistencies with available evidence and clinical practice—Results from an international survey. *Gastroenterology* **160**(7), 2234–2238 (2021).
9. Ho, W. J., Jaffee, E. M. & Zheng, L. The tumour microenvironment in pancreatic cancer—Clinical challenges and opportunities. *Nat. Rev. Clin. Oncol.* **17**(9), 527–540 (2020).
10. Maitra, A. & Hruban, R. H. Pancreatic cancer. *Ann. Rev. Pathol. Mech. Dis.* **3**, 157–188 (2008).
11. Varshney, N. et al. A review of Von Hippel-Lindau syndrome. *J. Kidney Cancer VHL* **4**(3), 20 (2017).
12. Tirosh, A. et al. Association of VHL genotype with pancreatic neuroendocrine tumor phenotype in patients with von Hippel-Lindau disease. *JAMA Oncol.* **4**(1), 124–126 (2018).
13. Ayloo, S. & Molinari, M. Pancreatic manifestations in von Hippel-Lindau disease: A case report. *Int. J. Surg. Case Rep.* **21**, 70–72 (2016).
14. Kane, L. E. et al. Diagnostic accuracy of blood-based biomarkers for pancreatic cancer: A systematic review and meta-analysis. *Cancer Res. Commun.* **2**(10), 1229–1243 (2022).
15. Alashwal, H., El Halaby, M., Crouse, J. J., Abdalla, A. & Moustafa, A. A. The application of unsupervised clustering methods to Alzheimer's disease. *Front. Comput. Neurosci.* **13**, 31 (2019).
16. Ringnér, M. What is principal component analysis?. *Nat. Biotechnol.* **26**(3), 303–304 (2008).
17. Yao, F., Coquery, J. & Lê Cao, K.-A. Independent principal component analysis for biologically meaningful dimension reduction of large biological data sets. *BMC Bioinform.* **13**, 1–15 (2012).
18. Endo, K., Weng, H., Kito, N., Fukushima, Y. & Iwai, N. MiR-216a and miR-216b as markers for acute phased pancreatic injury. *Biomed. Res.* **34**(4), 179–188 (2013).
19. You, Y. et al. MicroRNA-216b-5p functions as a tumor-suppressive RNA by targeting TPT1 in pancreatic cancer cells. *J. Cancer* **8**(14), 2854 (2017).
20. Wu, X. et al. MiR-216b inhibits pancreatic cancer cell progression and promotes apoptosis by down-regulating KRAS. *Arch. Med. Sci.* **14**(6), 1321–1332 (2018).
21. Felix, T. F. et al. MicroRNA modulated networks of adaptive and innate immune response in pancreatic ductal adenocarcinoma. *PloS one* **14**(5), e0217421 (2019).
22. Saha, B., Chhatriya, B., Pramanick, S. & Goswami, S. Bioinformatic analysis and integration of transcriptome and proteome results identify key coding and noncoding genes predicting malignancy in intraductal papillary mucinous neoplasms of the pancreas. *BioMed Res. Int.* **2021**, 1–11 (2021).
23. Chen, J. et al. Circulating microRNAs as potential biomarkers of HBV infection persistence. *Infect. Genet. Evol.* **54**, 152–157 (2017).
24. Jin, L. & Zhang, Z. Serum miR-3180–3p and miR-124–3p may function as noninvasive biomarkers of cisplatin resistance in gastric cancer. *Clin. Lab.* **66**(12) (2020).
25. Bao, M. et al. Proteomic analysis of plasma exosomes in patients with non-small cell lung cancer. *Trans. Lung Cancer Res.* **11**(7), 1434 (2022).
26. Chen, R. et al. Pilot study of blood biomarker candidates for detection of pancreatic cancer. *Pancreas* **39**(7), 981 (2010).
27. Pan, S. et al. Quantitative glycoproteomics analysis reveals changes in N-glycosylation level associated with pancreatic ductal adenocarcinoma. *J. Proteome Res.* **13**(3), 1293–1306 (2014).
28. Berberat, P. O. et al. Comparative analysis of galectins in primary tumors and tumor metastasis in human pancreatic cancer. *J. Histochem. Cytochem.* **49**(4), 539–549 (2001).
29. Chen, K. T. et al. Potential prognostic biomarkers of pancreatic cancer. *Pancreas* **43**(1), 22–27 (2014).
30. Shen, J., Person, M. D., Zhu, J., Abbruzzese, J. L. & Li, D. Protein expression profiles in pancreatic adenocarcinoma compared with normal pancreatic tissue and tissue affected by pancreatitis as detected by two-dimensional gel electrophoresis and mass spectrometry. *Cancer Res.* **64**(24), 9018–9026 (2004).
31. Paulo, J. A. et al. Proteomic analysis (GeLC–MS/MS) of ePFT-collected pancreatic fluid in chronic pancreatitis. *J. Proteome Res.* **11**(3), 1897–1912 (2012).
32. Paulo, J. A., Lee, L. S., Banks, P. A., Steen, H. & Conwell, D. L. Proteomic analysis of formalin-fixed paraffin-embedded pancreatic tissue using liquid chromatography tandem mass spectrometry (LC-MS/MS). *Pancreas* **41**(2), 175 (2012).
33. Pan, S. et al. Protein alterations associated with pancreatic cancer and chronic pancreatitis found in human plasma using global quantitative proteomics profiling. *J. Proteome Res.* **10**(5), 2359–2376 (2011).
34. Plebani, M. et al. Serum or plasma? An old question looking for new answers. *Clin. Chem. Lab. Med.* **58**(2), 178–187 (2020).
35. Matsubara, J. et al. Reduced plasma level of CXC chemokine ligand 7 in patients with pancreatic cancer. *Cancer Epidemiol. Biomark. Prev.* **20**(1), 160–171 (2011).
36. Kim, Y. et al. Development and multiple validation of the protein multi-marker panel for diagnosis of pancreatic Cancer. *Clin. Cancer Res.* **27**(8), 2236–2245 (2021).
37. European Study Group on Cystic Tumours of the Pancreas. European evidence-based guidelines on pancreatic cystic neoplasms. *Gut* **67**(5), 789–804 (2018).
38. Tyanova, S. et al. The Perseus computational platform for comprehensive analysis of (prote) omics data. *Nat. Methods* **13**(9), 731–740 (2016).
39. Hickey, G. L., Grant, S. W., Dunning, J. & Siepe, M. Statistical primer: sample size and power calculations—Why, when and how?†. *Eur. J. Cardio Thoracic Surg.* **54**(1), 4–9 (2018).

40. Mazzara, S. et al. CombiROC: An interactive web tool for selecting accurate marker combinations of omics data. *Sci. Rep.* **7**(1), 1–11 (2017).
41. Fabregat, A. et al. Reactome pathway analysis: A high-performance in-memory approach. *BMC Bioinform.* **18**, 1–9 (2017).

Acknowledgements

We are deeply grateful to the patients who kindly gave their samples to this research. We would also like to thank the Meath Foundation and Viatris Ireland for funding this research.

Author contributions

L.E.K., G.S.M., F.M, K.C.C., B.M.R. and S.G.M. conceptualized the study. G.S.M., P.F.R., F.M., K.C.C. and B.M.R. recruited patients to the study and processed clinical samples. L.E.K., G.S.M. and E.M. prepared and processed samples for LC–MS, whole transcriptome sequencing and ELISA. L.E.K., S.G.M., P.D., S.M. and B.M.R. were responsible for experimental design. C.S., M.H., and P.M. were responsible for running LC–MS and pre-processing of data. E.M.K. was responsible for sequencing of samples and pre-processing data. L.E.K. performed the data analysis and generated the figures. L.E.K, S.G.M. and B.M.R. wrote the original draft of the manuscript. All authors reviewed the manuscript. S.G.M. and B.M.R. acquired the funding to conduct this research.

Funding

This research was funded by the Meath Foundation (grant number RG105/2018) and Viatris Ireland.

Declarations

Competing interests

The authors declare no competing interests.

Ethics approval and consent to participate

This work was performed in accordance with the Code of Ethics of the World Medical Association (Declaration of Helsinki) for experiments involving humans. Patients provided informed consent for sample and data acquisition, and the study received full ethical approval from Tallaght University Hospital Joint Research Ethics Committee Review Board (ID: 0319-264).

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-83742-4.

Correspondence and requests for materials should be addressed to S.G.M.

Reprints and permissions information is available at www.nature.com/reprints.

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