A MULTI-OMICS CLASSIFIER ACHIEVES HIGH SENSITIVITY AND SPECIFICITY FOR PANCREATIC DUCTAL ADENOCARCINOMA IN A CASE-CONTROL STUDY OF 146 SUBJECTS

John Blume, Ghristine Bundalian, Jessica Chan, Connie Chao-Shern, Jinlyung Choi, Rea Cuaresma, Kevin Dai, Sara N. Golmaei, Jun Heok Jang, Manoj Khadka, Ehdieh Khaledian, Thidar Khin, Yuya Kodama, Ajinkya Kokate, Joon-Yong Lee, Manway Liu, Hoda Malekpour, Megan Mora, Nithya Mudaliar, Preethi Prasad, Madhuvanthi Ramaiah, Saividya Ramaswamy, Peter Spiro, Kavya Swaminathan, Dijana Vitko, James Yee, Brian Young, Susan Zhang, Chinmay Belthangady, Bruce Wilcox, Brian Koh, Philip Ma

PrognomiQ, Inc., San Mateo, CA

INTRODUCTION

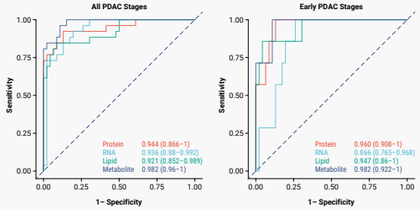
- Pancreatic cancer is currently the third leading cause of cancer-related mortality in the United States,[1] and demographic trends suggest that it will become the second leading cause by 2030[2-4]
- Pancreatic ductal adenocarcinoma (PDAC) and its variants compose more than 90% of pancreatic malignancies,[4] but PDAC is challenging to detect as the onset of clinical symptoms typically coincides with the progression to invasive growth and loss of resective opportunity[5]
- The majority of PDAC cases are not detected until late stage, with 80-85% of initial presentations representing incurable locally advanced or metastatic unresectable disease[6-7]
- Multiple clinical and investigational markers are in use (eg, CA19-9)[8]; however, given the limited performance of those markers, the United States Preventive Services Task Force currently recommends against routine screening for PDAC[9]
- We hypothesize that a plurality of signal inputs, such as those for different types of blood analytes, may be necessary to detect PDAC early enough for interventions that improve patient survival

OBJECTIVE

- To develop and validate a multi-omics PDAC classification model utilizing a combination of orthogonal features (ie, proteins, metabolites, lipids, and RNAs) sampling a wide variety of physiological systems and pathways

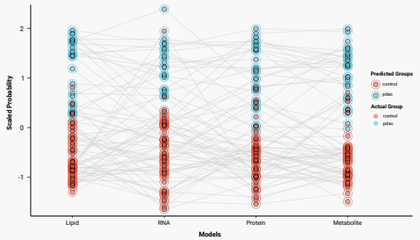
RESULTS

FIGURE 2. Individual omics models demonstrated high classification performance for distinguishing PDAC from non-cancer controls in the validation cohort.



PDAC early-stage (I/II) validation result (subset of all-stage result) comprised 7 PDAC and 46 control subjects.

FIGURE 3. Comparison of individual omics models' predicted class probabilities highlighted omics-specific class-probability ordering.



The scaled, predicted PDAC or control class probabilities are shown for the validation set subjects. Each subject is linked across the models by gray lines.

- Each of the omics models had changes in probability rank order from one model type to another even though overall performance was not significantly different between models (**Figure 3**)
- The difference in model probability rank (**Figure 3**) coupled with the similarity of the models' areas under the ROC curve (AUCs; **Figure 2**) suggest that each omics type may work synergistically to contribute distinct, useful information

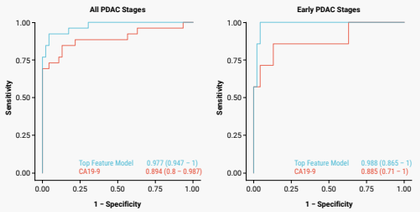
TABLE 2. The top 5 features for each omics type were selected for combined analyte GLMnet model creation.

Omics Type	Feature ID[a]	Protein[b] *(Gene)*
Proteins[c]	NP4_GAGGQSMSEAPTGDHAPAPTR	Q14767-LTBP2 *(LTBP2)*
	NP3_TFVIIPELVLPNR	Q9BYE9-CDHR2 *(CDHR2)*
	NP2_TFVIIPELVLPNR	Q9BYE9-CDHR2 *(CDHR2)*
	NP3_DSC(UniMod:4) TMRPSSLGQGAGEVWLR	Q9BYJ0-FGFBP2 *(FGFBP2)*
	NP3_DNC(UniMod:4) PHLPNSGQEDFDK	P35442-TSP2 *(THBS2)*
Metabolites[d]	NEG_AICAR	
	POS_Cystine	
	NEG_CMP	
	NEG_Gentisate	
	POS_Creatine	
Lipids[d]	NEG_PC(18:2_20:5)+AcO	
	POS_DAG(18:1_20:0)+NH4	
	NEG_PE(O-16:0_22:6)-H	
	NEG_PC(18:2_20:3)+AcO	
	POS_CER(d18:1/18:0)+H	
RNAs	ENST00000483727.5	NA[e] (ENSG00000078747)
	ENST00000531734.6	Q01433-AMPD2 *(AMPD2)*
	ENST00000437154.6	Q9UKQ2-ADA28 *(ADAM28)*
	ENST00000531997.1	NA[e] *(GLYATL1 pseudogene)*
	ENST00000424185.7	Q8IW45-NNRD *(NAXD)*

[a]Top features were selected from the ranked feature importance list for each individual omics type XGBoost model from the first 10x10 RCV.
[b]UniProt identifier-UniProt protein name.
[c]Nanoparticles from the Proteograph-based proteomics sample preparation platform are identified as NPx.
[d]NEG and POS refer to MS ionization mode during data acquisition.
[e]No associated protein identifiers.

- Inspection of the ranked importance plots suggested that the top 5 features provide most of the discrimination performance (**Table 2**)

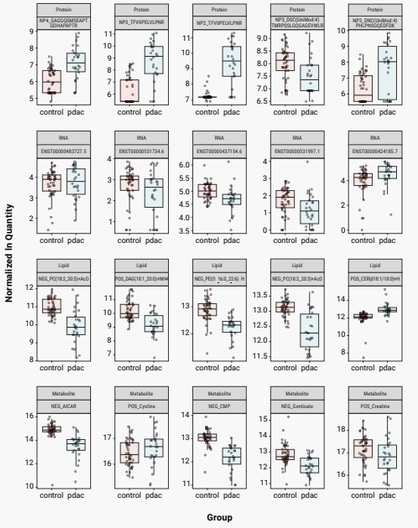
FIGURE 4. The combined top feature, multi-omics model achieved high sensitivity and specificity for PDAC in the validation cohort.



For the multi-omics (blue) and CA19-9 (red) models, performance in the validation subjects was plotted as a ROC curve. AUCs with 95% confidence intervals are annotated.

- At 99% specificity, the observed sensitivity for the multi-omics model for all- and early-stage PDAC was 80.8% and 71.4%, respectively. The CA19-9 model had 69.2% and 57.1% sensitivity for all- and early-stage PDAC, respectively (**Figure 4**)
- Although this preliminary study was neither intended nor powered to demonstrate superiority to CA19-9, the results are suggestive of such and will factor into future study designs
- Differences in plasma levels of the 20 features between PDAC and non-cancer subjects suggest clinical development feasibility of these analytes (**Figure 5**)
- Although some of the features of the final classifier have known links to PDAC, others do not yet have an established connection, which is expected from an unbiased, untargeted approach to multi-omics data generation and evaluation

FIGURE 5. Plasma levels of the 20 features from the multi-omics model differ between PDAC and non-cancer subjects from the validation cohort.



Columns are ordered by feature importance (reducing importance from left to right). The second and third most important protein features comprise the same peptide, TFVIIPELVLPNR, as selected on 2 distinct nanoparticles.

CONCLUSIONS

- In this proof-of-concept study, we leveraged a broad, multi-omics profiling platform to identify novel combinations of analytes with both known and unknown relation to PDAC into a high-performance biomarker panel for detection and discrimination of PDAC from non-cancer controls
- The manageable number of easily assayed analytes collected from blood draws makes this panel ideal for rapid development and clinical study
- Results support the feasibility of this approach as a clinically useful test for PDAC, with potential for earlier detection

REFERENCES

1. Siegel RL, et al. *CA Cancer J Clin*. 2022;72(1):7-33.
2. Smith BD, et al. *J Clin Oncol*. 2009;27(17):2758-65.
3. Rahib L, et al. *Cancer Res*. 2014;74(11):2913-21.
4. Park W, et al. *JAMA*. 2021;326(9):851-862.
5. Chari ST, et al. *Pancreas*. 2015;44(5):693-712
6. Earle CC, et al. *Can J Gastroenterol*. 2003;17(3):161-7.
7. De Dosso S, et al. *Cancer Treat Rev*. 2021;96:102180.
8. Ge L, et al. *BMJ Open*. 2017;7(12):e018175.
9. US Preventive Services Task Force. *JAMA*. 2019;322(5):438-444.
10. Schwenk JM, et al. *J Proteome Res*. 2017;16(12):4299-4310.

DISCLOSURES

Study funded by PrognomiQ, Inc. All authors are current or former employees of PrognomiQ, Inc.

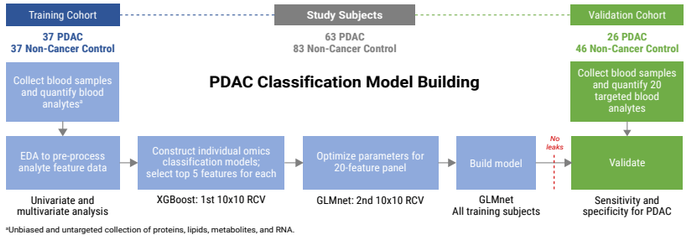
ACKNOWLEDGEMENTS

Funded by PrognomiQ, Inc (San Mateo, CA). Editorial and graphical assistance provided by Prescott Medical Communications Group (Chicago, IL).

METHODS

- This case-controlled study was comprised of 63 diagnosis-aware, but treatment-naïve PDAC subjects (12 stage I, 8 stage II, 4 stage III, 36 stage IV, and 3 stage unknown) and 83 age- and sex-matched non-cancer control subjects enrolled across 16 clinical sites as part of an ongoing, IRB-approved, observational study
- Unbiased liquid chromatography mass spectrometry (LCMS), multiplexed and targeted multiple reaction monitoring-LCMS, RNA-Seq, and ELISA assays were used to detect and quantify proteins, metabolites and lipids, RNAs, and CA19-9, respectively, from subject blood samples
- Distinct cohorts of subjects were created for machine learning-based classification model training (ie, repeated cross-validation [RCV] and final model construction) and validation (**Figure 1**). The proportionality of PDAC cancer stages was maintained across cohorts
- Univariate and multivariate exploratory data analysis (EDA) was performed for initial evaluation, and the significant differences included several features related to acute and inflammatory response. To improve classification specificity to PDAC, proteins were additionally filtered to remove features associated with high-abundance proteins (based on Human Plasma Proteome Project top 25%),[10] and both proteins and RNAs were filtered to remove any features associated with acute immune/inflammatory response-related Gene Ontology Biological Process terms (**Table 1**)
- We used a robust machine-learning modeling engine XGBoost, a gradient-boosted, ensemble-tree method, to construct individual classification models for each of the 4 omics types using 10 repeats of 10-fold cross-validation (10x10 RCV; **Figure 1**)
- The 5 features that contributed most to each individual omic model were identified and used as input for a new RCV with a re-shuffling of the training subjects and a grid of 50 hyperparameter combinations to create a GLMnet regularized logistic regression model. This final, 20-feature, multi-omics model was applied to the validation cohort to evaluate PDAC classification performance (**Figure 1**)

FIGURE 1. The classification modeling algorithm used a 2-tiered approach with a training cohort for analyte feature selection and model optimization and a validation cohort for performance assessment.



*Unbiased and untargeted collection of proteins, lipids, metabolites, and RNA.

TABLE 1. Pre-processing of omics data refined the number of features for each omics type.

Omics Type	Feature Group	Feature Count
Proteins	Original	151,461
	EDA	54,108
	Classification	21,176
Metabolites	Original	377
	EDA	373
	Classification	372
Lipids	Original	898
	EDA	898
	Classification	879
RNAs	Original	202,125
	EDA	110,734
	Classification	107,631

Original data counts represent raw omics data. EDA data counts reflect features filtered for presence in ≥25% of all 146 subject samples. Classification feature counts reflect features filtered for presence in ≥50% of at least 1 of the classes (PDAC or control) in the training subjects.