Leveraging real-world pancreatic cancer datasets to drive drug discovery and patient health

Z. Worman[2], K. Abdilleh[1], A. McNiff[3], D. Shao[3], P. Webster[2] ,R. Beck[2], C. Fisher[2], D. Sain[2], T. Khoyratty[2], J. DiGiovanna[2], G. Aquaah-Mensah[3], L. Matrisian[1], S. Doss[1],B. Davis-Dusenbery[2]

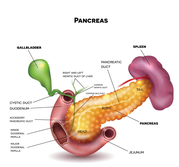
[1]Pancreatic Cancer Action Network, Manhattan Beach, CA, [2]Velsera, Cambridge, MA, [3]Massachusetts College of Pharmacy & Health Sciences

Pancreatic cancer is the third leading cause of cancer-related death in the United States. Current therapeutic options offer a dismal overall survival with the 5-year survival at just ~12%. Analysis of the clinical and molecular underpinnings of pancreatic cancer is critical to developing both early detection methodologies as well as novel therapeutic options. The aggressiveness and deadly nature of this disease warranted the development of a data repository and analysis system dedicated to pancreatic cancer data. The Pancreatic Cancer Action Network's (PanCAN) SPARK platform is a cloud-based data and analytics platform, powered by Velsera, that integrates real-world patient health data from PanCAN research initiatives and accelerates research by making pancreatic cancer data easier to access and use. Encompassing clinical, imaging, and genomics data from over 600 patients with pancreatic cancer within PanCAN's Know Your Tumor ® (KYT) precision medicine service, the SPARK platform connects with petabytes of publicly available cancer data via the Cancer Genomics Cloud (CGC), also powered by Velsera. The CGC is part of NCI's Cancer Research Data Commons (CRDC), a cloud-based data science infrastructure that connects data with analytics tools to allow researchers to share, integrate, analyze, visualize, and drive scientific discovery.
Here, we demonstrate the application of these datasets by providing a case study demonstrating how to combine and enrich data to accelerate pancreatic cancer research. Currently, the genomic and proteomic data available on CRDC amounts to 402 and 304 cases of pancreatic tumor samples, respectively. We will use the capabilities of the SPARK and CGC platforms, which provide ready-to-use tools for multi-omics analysis that require no coding knowledge. Using the KYT and CRDC open-access pancreatic cancer data, we aim to demonstrate how to perform integrated analysis of data from diverse scientific domains, and share with collaborators all in one space, streamlining and increasing the potential for new scientific discoveries. Further expansion of the PanCAN and CGC datasets will undoubtedly provide a more comprehensive understanding of pancreatic cancer tumor biology. SPARK and CGC's cloud-based computation infrastructure, along with numerous available cancer datasets and easy-to-use multi-omics data processing workflows and data analytic tools will be instrumental in this process.

BACKGROUND

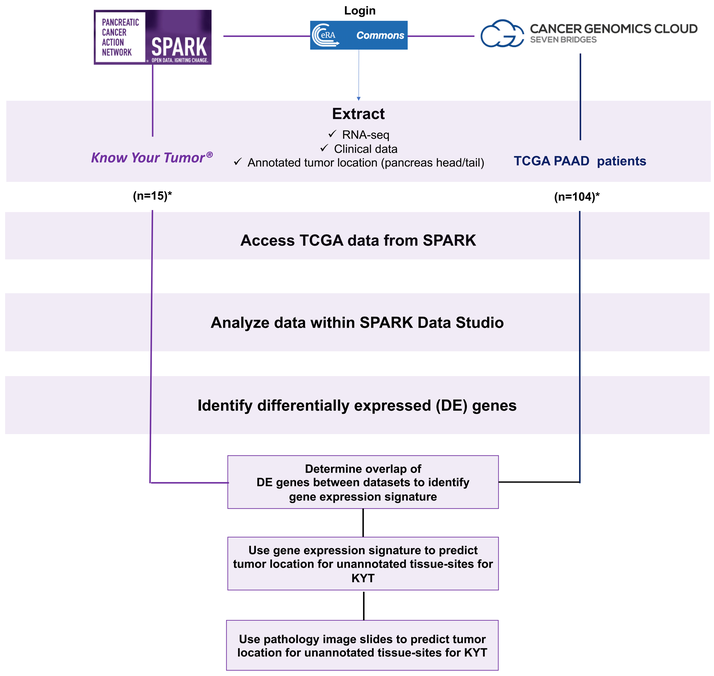
The disease is the ***3rd leading cause of cancer-related deaths*** in the United States, ***with the lowest five-year survival rate at just 10%.***

Pancreatic cancer is usually not diagnosed until advanced stages due to the location of the pancreas. Moreover, studies have shown that carcinogenesis in pancreatic tissue may differ depending on tumor location – i.e. head of pancreas vs body/tail of pancreas. Clinicopathology studies have shown that Pancreatic Ductal AdenoCarcinoma (PDAC) PDAC head and body/tail tumors have different physiological and clinical presentations as well as different survival outcomes.



The aim of this study is to characterize unique gene expression signatures that differentiate between head and tail PDAC cancers. We coupled data from PanCAN's Know Your Tumor® (KYT) program from PanCAN's SPARK platform with TCGA-PAAD, phs000178, Pancreatic Ductal AdenoCarcinoma. This study was retrieved from the NCI Cloud Resource, Cancer Genomics Cloud. Using multi-modal data from both datasets, we couple transcriptomic and ML methods to identify a putative anatomical-site based prognostic signature of PDAC.

METHODS



* These numbers correspond to patients with annotated anatomical site tumor location as well as RNAseq data.

RESULTS

Table 1: 11 genes differentially expressed between head and tail pancreas tumors in both the KYT & TCGA datasets (padj < 0.1)

Genes with expression downregulated in the head relative to the tail of pancreas	
PGA3	Pepsinogen A3
PGA4	Pepsinogen A4
PGA5	Pepsinogen A5

Genes with expression upregulated in the head relative to the tail of pancreas	
APOA4	Apolipoprotein A4
APOB	Apolipoprotein B
CHP2	Calcineurin Like EF-Hand Protein 2
DES	Desmin
LCT	Lactase
PCK1	Phosphoenolpyruvate Carboxykinase 1
SLC26A3	Solute Carrier Family 26 Member 3
SI	Sucrase-Isomaltase

Machine learning using gene expression signature

- The 11 DE genes were used as input in ML models (-J48 Decision Tree in WeKa with 10-fold cross validation)
  to classify between head and tail pancreatic cancer tumors
- For KYT data, APOA4 was an important gene in predicting head vs tail tumors (Fig 1A)
- For TCGA, the pepsinogen genes (PGA4, PGA5) were important in predicting head vs tail tumors (Fig 1B)
- Differences observed may be attributable to the different sampling protocols between the two datasets

Figure 1A

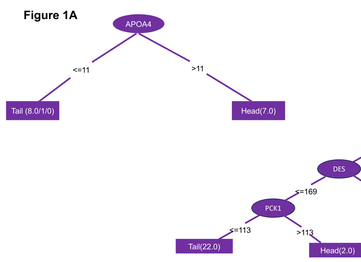


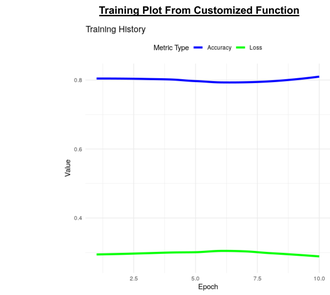
Figure 1B

Figure 1A-B: Decision trees were generated by training the J48 algorithm on the gene expression data from the 11 genes. The top ellipse is the node root of tree and represents the most important condition for discriminating head from tail tumors. For the KYT data, the APOA4 gene expression signature is the most discriminating feature between head and tail tumors based on the J48 model. For the TCGA data, several of the 11 DE genes (including the pepsinogen genes) represent conditions for considering head vs tail tumors. The rectangles are the leaf nodes that represent the final classification (head vs tail). In the round brackets inside rectangles, the number before the slash indicates the total number of instances of head/tail and the number after the slash indicates how many head/tail tumors were incorrectly predicted.

	% correctly predicted		% unannotated samples predicted as head or tail	
	J48	Random Forest	J48	Random Forest
Head	7/8 (88%)	7/8 (88%)	206/215 (96%)	205/215 (95%)
Tail	1/7 (14%)	1/7 (14%)	9/215 (4%)	10/215(5%)

Table 2: 95% of unannotated KYT samples were predicted to be tumors deriving from the head of pancreas using the gene expression signature for both the J48 and Random Forest classification schemes. This is in line with observations that tumors in head of pancreas are more common than tumor in body/tail of pancreas (75-80% vs 20-25%)

RESULTS



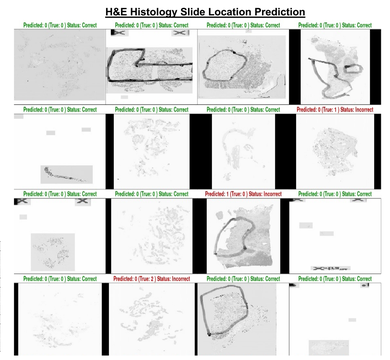

```
                Overall  Training Statistics
Accuracy : 0.7971
95% CI : (0.7507, 0.8383)
No Information Rate : 0.9101
P-Value [Acc > NIR] : 1.00
 Kappa : -0.0133

 Mcnemar's Test P-Value : 0.27
Statistics by Class:
```

	Class: body	Class: head	Class: tail
Sensitivity	0.86943	0.0625	0.066667
Specificity	0.09677	0.948328	0.924242
Pos Pred Value	0.90698	0.055556	0.038462
Neg Pred Value	0.06818	0.954128	0.956113
Prevalence	0.91014	0.046377	0.043478
Detection Rate	0.7913	0.002899	0.002899
Detection Prevalence	0.87246	0.052174	0.075362
Balanced Accuracy	0.4831	0.505414	0.495455

model_summary	Total params: 15764675 (60.14 MB)	Trainable params: 15764675 (60.14 MB)
Model: "pancan_model"	Non-trainable params: 0 (0.00 Byte)	
Layer (type)	Output Shape	Param #
conv2d_5 (Conv2D)	(None, 254, 254, 32)	896
max_pooling2d_5 (MaxPooling2D)	(None, 127, 127, 32)	0
conv2d_4 (Conv2D)	(None, 125, 125, 64)	18496
max_pooling2d_4 (MaxPooling2D)	(None, 62, 62, 64)	0
flatten_2 (Flatten)	(None, 246016)	0
dense_3 (Dense)	(None, 64)	15745088
dense_4 (Dense)	(None, 3)	195

Fig 2A. We performed data augmentation on 192 H&E images from pancreatic cancer biopsies taken from either the head, tail, or body of the pancreas. we created a custom training and plotting function. Below the plot is the summary of the pancan_model and to the right are the overall training statistics for each tumor location.

H&E Histology Slide Location Prediction



Class	Sensitivity	Specificity
body	87%	10%
head	6%	95%
tail	7%	92%
Class	**Neg.Pred.Value**	**Prevalence**
body	7%	91%
head	95%	5%
tail	96%	4%
Class	**Pos.Pred.Value**	**Detection.Rate**
body	91%	79%
head	6%	0%
tail	4%	0%

Fig 2B. As a visual representation for the effectiveness of our training model we generated the images and prediction of the first 25 images within the data array we used to hold imaging data. Green text represents a correct prediction and Red indicated incorrect with the actual location encoded as Body (0), Head (1), and Tail (2). The associated table above represents the prediction statistics for each location.

CONCLUSIONS

In this study, we demonstrate how to:

- **Leverage real-world datasets** like the Know Your Tumor data
- **Access and analyze TCGA** within the PanCAN SPARK platform
- Conduct **multi-modal data analysis** capabilities of the Velsera platform on combined datasets (i.e. analysis of clinical, RNAseq, and whole-slide pathology imaging data from KYT and TCGA)
- **Identify putative prognostic signatures from real-world data and use TCGA as a validation dataset**

Scan to visit
www.pancan.org/spark

Scan to visit
www.cancergenomicscloud.org

Sign up for the April webinar on the CGC website to learn more about SPARK!