Metanalysis of 3 independent studies exploring gene expression in cancer cells for the identification of significant biomarkers in pancreatic cancer cells

Nathan Ona[1], Ahsan Sarwar[1], Sarthak Sharma [1]

[1] School of Biomedical Engineering, Drexel University, USA

Course: Bmes543 Quantitative Systems Biology

Instructor: Ahmet Sacan

Date: 2022-06-07

Dataset(s) : Datset1: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE42952

Dataset 2: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0148807

Dataset 3: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE32676

ABSTRACT

.

The genetic markers that influence a deadly disease like pancreatic cancer are still being explored in the genomics industry. Because of it's difficult early prognosis and low survival rate, exploring the genetic markers that the disease leaves in patients could lead to breakthrough treatment plans. The study conducted in this report is a review of 3 separate datasets that explored different gene expression levels, biological processes, and drug repositioning tactics to further the understanding of pancreatic cancer. The computational analysis discussed here was conducted in MATLAB, using specific functions from the Bioinformatics toolbox. In this paper, we preform fold change to find the significant genes in each dataset, hypergeometric tests to determine relevant pathways, and use Clue.io to determine potential drug candidates.

1 INTRODUCTION

Pancreatic cancer has one of the lowest survival rates of any cancer type, which is mainly attributed to the difficulty in early detection. It is still the fourth leading cause of cancer-related death, and surgery can only provide a medial survival time frame of 17-23 months[1]. There are variations of the cancer, that make it difficult to provide a feasible treatment for this disease. Specifically, pancreatic ductal adenocarcinoma (PDAC) is one of the most detrimental variations of the disease. It accounts for nearly 90% of all pancreatic cancer cases worldwide, and due to the difficulties of detecting the cancer early, treatments and therapeutics are vastly limited[2] . Current treatment plans for PDAC cases include surgical resection, an uncommon method of treatment as most patients who present with PDAC have locally advanced, non-resectable tumors, and systemic chemotherapy. Chemotherapy treatment is generally not effective, as PDAC tumors have a high degree of radioresistance[3].

The formation of PDACs comes from the transformation of one of the fundamental types of cells in the pancreas, acinar cells, into pancreatic intraepithelial neoplasias (PanINs), which drives uncontrollable cell growth, causing tumor generation. Acinar cells tend to transdifferentiate themselves into more epithelial cells due to the plasticity of the pancreas, a process called acinar-to-ductal metaplasia (ADM), which makes the pancreas more susceptible to mutations in the KRAS gene, which leads to the formation of PanINs. Figure 1 shows a simplified version of this process, starting from normal acinar cells to the formation of the PDAC tumor[3].

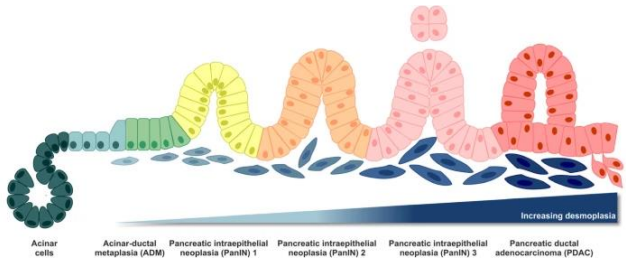

Figure 1- Formation of PDAC Pathway[3]

The main objective of this paper is to analyze and evaluate multiple, pre-existing datasets that explore

gene expression and pathways in relation to PDAC formations. From analyzing the datasets, the goal is to determine significant biomarkers in pancreatic cancer cells and use drug repositioning techniques to determine possible existing drugs that could meaningfully regulate those significant genes. If successful, the results of this study will provide a list of possible drugs that, when repositioned, could be suitable treatments for pancreatic cancer.

Most of the studies conducted involving PDAC development and genetics are completed to find genetic biomarkers, specifically finding certain cell types that were up or down regulated. One study that explores this topic performs in-vitro tests, to determine how the cells of the pancreas react to a stress case scenario. This study found that certain cancer stem-cell-like cells (CSCs) were more prevalent in patients that were diagnosed with pancreatic cancer. One of the significant biomarkers from that study was CD44, which is a biomarker for upregulation of metastasis of cancer cells. [4]

2 DATASET

Dataset 1 - Molecular markers associated with outcome and metastasis in human pancreatic cancer

The first dataset that was analyzed performed a microarray analysis on patients diagnosed with PDACs. The main goal of the study was to determine any possible molecular characteristics that could be impactful to identify early pancreatic cancer development. The experimental groups in this study were divided into patients with good and bad outcomes. Good and bad outcomes were determined using Kaplan-Meier life table. Good outcomes are defined as those patients who had a disease-free survival score (DFS)> 50, and bad outcomes were those with a DFS< 19.5 months. Additionally, there were 4 control samples, which were sourced from the surrounding, non-cancerous areas of pancreatic tissue, 6 bad patient outcomes, and 6 good patient outcomes.

The actual procedure of obtaining the data for this dataset was conducted from extracted PDAC samples from each type of patient (Good, Bad, Control). Trizol and RNeasy were used to extract the RNA from the samples, and the concentration and integrity of each sample was measured using a Bioanalyzer 2100. Once an adequate amount of RNA was extracted for analysis, a GeneChip[R] scanner was used to read the intensities of each signal and store it for computation analysis. The data was normalized using RMA normalization, which corrects for background drift[6]. Once the data was normalized, pathway analysis was completed using Ingenuity Pathway Analysis (IPA), where KEGG pathways on up or down regulated genes was performed. Lastly, there were two major statistical tests performed, hypergeometric distribution and chi-squared test of independence[5].

Dataset 2 - The Metastatic Potential and Chemoresistance of Human Pancreatic Cancer Stem Cells

The main objective of the second dataset was to determine the extent of the chemoresistance and metastatic potential of cancer stem cells, specifically those affiliated with PDACs. The experimenters who conducted the study focused their efforts on evaluating side population (SP) cells to determine if that would be a useful point to further explore PDAC development in its early stages. The dataset itself was not available publicly and was requested by the design team to use for this paper. The received data was normalized, and the corresponding probe IDs and gene symbols were provided. The two main experimental groups were the 56 control samples that were used as a baseline, and the 135 cancer patient samples that were diagnosed with PDACs[7]. The study also had an orthotopic component that was conducted on mice, so for the purposes of our experiment, that data was not used.

The main procedure of the study involves studying the effects of side potential cells of PDACs. There was an SP assay performed, where the cells were first stained with Hoechst 33342 (HOE), which is a popular blue florescent counterstain used in molecular biology[8]. The samples were then incubated for 60 minutes at 37 degrees Celsius, after which, they were suspended in a combination mixture of saline, BSA, and DNase. Once collected, the samples were filtered through a 40μM cell strainer[7]. These strained samples were then used for *in vitro* testing, which found that using SP cells can serve as a viable tool to further evaluate PDAC therapy. Additionally, it was found that verapamil, a calcium channel blocking drug, may be useful in increasing the sensitivity of SP cells[7].

Dataset 3 - Integrative Survival-Based Molecular Profiling of Human Pancreatic Cancer [mRNA]

This dataset is a University of California, Los Angeles (UCLA) study that targets a relatively straightforward goal – to better understand PDACs and their significant pathways as related to genomics. The experiment designers explored the "identification and refinement of prognosis-related genes in PDAC develop-

ment," [9] to further influence the growth of early treatment for pancreatic cancer. The study consisted of two different experimental groups, 7 control samples, and 42 PDAC tumor patient samples. The main form of analysis that was utilized was microarray analysis, to determine different molecular patterns that occur within the cancer samples. The experimenters first conducted *in silico*, or computer-based analysis, and then performed more *in vitro* tests to validate the computational findings.

The gene expression of each of the samples was measured by a computational tool, Affymetrix HGU133 Plus 2 Array. This GeneChip technology lets multiple probe IDs be an input to determine the expression of a single gene throughout an entire sample set[10]. The data obtained from this study was found in publicly available sources and was already RMA normalized and filtered. The significant pathways and gene ontologies were also analyzed using publicly available domains, such as the Fisher exact test[9]. A prediction analysis algorithm was generated to then find the best probe IDs that could be used to predict a patient outcome, by analyzing specific gene expression levels. The results of the survival-based study were that they found consistent results with other studies, in terms of potential biomarkers for early PDAC development. Namely, the core pathway that was found to promote PDAC development was PI3K/AKT. Additionally, the overregulation of the SRC gene, which is a protein tyrosine kinase was present in most PDAC samples. This led to the conclusion that in early prognosis, the SRC gene's downregulation could lead to better treatment plans[9].

3 METHODS

The 3 datasets we took all presented data taken from pancreatic cancer cells, and control cells taken from surrounding tissue. Datasets 1 and 3 were downloaded using a function for GSE download since the datasets had GSE numbers, but our 2nd dataset was in excel format directly from the researchers themselves. The 2nd Dataset already had Gene names associated with every row of the data, but for the other 2 datasets gene names had to be translated from GPL probes after the data was inputted into MATLAB for analysis.

The papers associated with each dataset were checked to see whether normalization was performed on them, and having confirmation that all 3 datasets were normalized, no normalization had to be done by us in MATLAB code. The next step was to segment out cancer vs control data in each dataset, so that appropriate analysis can be performed to make valid conclusions about the expression levels of genes in cancer and control cells. This was done by visually inspecting each dataset, to know where the column cut-offs are, and then writing code that would splice the data into cancer and control groups.

Fold Change

Overall, for the 3 datasets, it was decided to calculate the foldchange which is a measure describing the degree of quantity change between the final and original value. Here the original values would be the gene expression value for control cells, and the final values would be the gene expression values for cancer cells. Taking the difference between both cancer and control average gene expression values gives us a change value, and fold change can be calculated by $2^{-\text{deltadeltachange}}$. If the calculated value is negative inverse is taken, and ultimately the absolute value is taken for all calculations to give a final set of fold change values for the dataset. These values can be sorted in decreasing order, and the top 100 values (fold change value > 2) give us the most significantly up/downregulated genes between cancer and control cells.

Clue.io

These 300 genes were used as queries on a website called Clue.io. Clue.io is an online database that uses connectivity maps to visualize the relationships between diseases, genes, and therapeutics[11]. The functionality of Clue.io used is its drug repurposing database. Using the gene symbols for that query, Clue.io reported back drugs that target any of the queried genes giving a basis for possible drug repurposing for pancreatic cancer.

First Significant Gene Across All Datasets

A bar graph was created in MATLAB to show the average expression level for the first most significant gene across data sets based on the foldchange level.

Hypergeometric Test

A hypergeometric test was conducted on datasets 1 and 3 to see determine the pathways that are most affected by the significant genes (p-value <0.01) between control cells and cancer cells.

The 2nd dataset had gene names generated already, so was put into David Bioinformatics to do essentially what the hypergeometric test did for the 1st and 3rd datasets.

4 EXPERIMENTS AND RESULTS

Fold change of genes was calculated for each dataset. One aspect of that data that was a concern was how many of these genes overlapped between the three datasets. **Error! Reference source not found.Error! Reference source not found.** shows the overlap of significant genes across all 3 datasets.

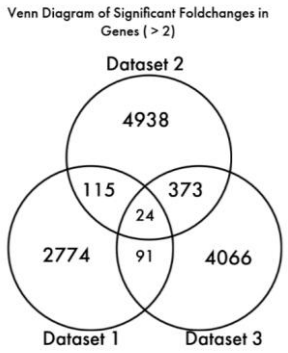

Figure 3 - Venn Diagram of the Significant fold changes (>2) for each dataset.

Having a breakdown of the overlapped genes, an analysis of the significant genes in Clue.io a list of possible drugs to be repurposed was conducted using the top 100 genes. 100 genes were used because Clue.io did not allow genes missing from its database to be used for analysis.

The drug list was sorted by the number of target genes that are significantly regulated in pancreatic cancer. **Table 1** shows the top 5 drugs reported by Clue.io that could be repurposed for treating pancreatic cancer.

Drug Name	Mechanism of Action	Target Gene(s)
marimastat	matrix metalloprotease inhibitor	MMP1, MMP10, MMP14, MMP7
niflumic-acid	cyclooxygenase inhibitor	KCNQ1, PLA2G1B, UGT1A9
bardoxolone-methyl	Nuclear factor erythroid-derived, like (NRF2) activator	PPARG, STAT3
cholic-acid	bile acid	COX7A1, PLA2G1B
dexfosfoserine	membrane integrity inhibitor	CFTR, REG1A

Table 1. Top-5 potential drugs for repurposing for pancreatic cancer Using Clue.io, we were able to find drug candidates, based on the target genes from our analysis of the three datasets.

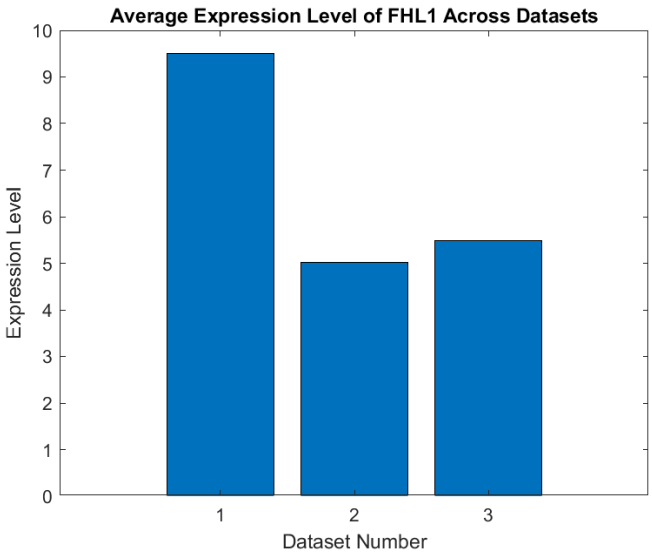

Figure 2 - Bar graph of first most significant gene across all 3 datasets

There were 24 genes that were shared between the three datasets, so a search was conducted to find the first most regulated gene among all datasets. Figure 3 shows the average expression of that gene, which is FHL1, a gene that provides instructors for the creation of skeletal and cardiac muscle.

A hypergeometric test was conducted using datasets 1 and 3 to see what biological processes are most significant in pancreatic cancer given the significant genes found between control and patient pancreatic cells. **Table 2** shows the top 10 pathways that these genes are likely to affect. Most of the pathways presented affect the function of individual cells from the formation of the cell to how the cells communicate and proliferate (mRNA).

For Dataset 2, DAVID Bioinformatic was used in place of a hypergeometric test. However, the genes from that study were from *homo sapiens*, but DAVID registered them as other species, so no conclusive biological pathway was recovered.

Pathway	P-Value
gene expression	1.1E-06
response to stimulus	5.8E-05
mRNA metabolic process	1.1E-05
viral life cycle	1.3E-05
nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	2.0E-05
SRP-dependent co-translational protein targeting to membrane	3.7E-05
pattern specification process	2.0E-04
cell morphogenesis	2.4E-04
nucleic acid phosphodiester bond hydrolysis	4.2E-04
detection of chemical stimulus involved in sensory perception of smell electronic annotation'	7.3E-04

Table 22. Top-10 significant pathways for genes in pancreatic cancer. This data was constructed from functions in MATLAB, and provides context on relevant biological pathways

5 DISCUSSION

Our experimental design conducted a meta-analysis of the three datasets discussed, to find potential biomarkers of pancreatic cancer, specifically PDACs. However, because our methods were different from the datasets, and approached the analysis more holistically, the results from our study do not universally align with those we researched. For example, Dataset 3 reported that the SRC gene was one of the most promising genes to affect pancreatic cancer, however, our data did not report it as one. Additionally, dataset 2 found that verapamil could potentially be a drug candidate for drug repositioning, which could be a helpful tool in treating pancreatic cancer. However, our study found that a different drug, Marimastat, which is a metalloproteinase inhibitor, would be the most probable choice, as it targets multiple significant genes found in our study and is essential for the suppression of tumor metastasis.

The data we extracted from each dataset for the meta-analysis was most relevant to identifying biomarkers for PDACs in humans. As an example, dataset 2 conducted part of their study by injecting human PDAC cells in mice but we used none of that data as it wasn't relevant to the analysis we conducted. Our approach for the metanalysis was to be able to find significantly expressed genes in pancreatic cancer and controls, significantly differentially expressed genes between cancer samples and controls, and identify potential drugs for repotting for treatment.

Out 2nd dataset did not contain GPL Probe information which would be used to conduct a hypergeometric test, since it had GPL information already converted into gene names. Initial thought was that since we have gene names we can use David Bioinformatics for the test, but David Bioinformatics suggested that we had information from multiple species. We were unable to obtain any information from DAVID for this reason and were confident that there is an issue with the program since all our papers focused on human PDAC. Not being able to run the hypergeometric test on our 2nd dataset was a big limitation, as was that all the 3 studies discussed used different methods. One thing that could help improve our metanalysis would be to find studies conducted around the world that used the same methods, and perhaps the same GPL probes to better support a metanalysis done on them.

6 REFERENCES

1. I. Aier, R. Semwal, A. Sharma, and P. K. Varadwaj, "A systematic assessment of statistics, risk factors, and underlying features involved in pancreatic cancer," *Cancer Epidemiology*, 08-Dec-2018. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S1877782118305101. [Accessed: 29-May-2022].

2. P. Sarantis, E. Koustas, A. Papadimitropoulou, A. G. Papavassiliou, and M. V. Karamouzis, "Pancreatic ductal adenocarcinoma: Treatment hurdles, tumor microenvironment and immunotherapy," *World journal of gastrointestinal oncology*, 15-Feb-2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7031151/#:~:text=Pancreatic%20ductal%20adenocarcinoma%20(PDAC)%20is%20a%20highly%20aggressive%20lethal%20malignancy,90%25%20of%20pancreatic%20cancer%20cases. [Accessed: 29-May-2022].

3. M. Orth, P. Metzger, S. Gerum, J. Mayerle, G. Schneider, C. Belka, M. Schnurr, and K. Lau-

ber, “Pancreatic ductal adenocarcinoma: Biological hallmarks, current status, and future perspectives of combined modality treatment approaches - radiation oncology,” *BioMed Central*, 08-Aug-2019. [Online]. Available: https://ro-journal.biomedcentral.com/articles/10.1186/s13014-019-1345-6. [Accessed: 29-May-2022].

4. L. T. Senbanjo and M. A. Chellaiah, “CD44: A multifunctional cell surface adhesion receptor is a regulator of progression and metastasis of cancer cells,” *Frontiers in cell and developmental biology*, 07-Mar-2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5339222/. [Accessed: 04-Jun-2022].

6. “Molecular organisation and assembly in cells,” *RMA and GC-RMA Normalisation*. [Online]. Available: https://warwick.ac.uk/fac/sci/moac/people/students/2003/sam_robson/usergroups/rmavsmas5/. [Accessed: 04-Jun-2022].

7. V. J. Bhagwandin, J. M. Bishop, W. E. Wright, and J. W. Shay, “The metastatic potential and chemoresistance of human pancreatic cancer stem cells,” *PLOS ONE*. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0148807#sec002. [Accessed: 05-Jun-2022].

8. “Hoechst 33342, trihydrochloride, trihydrate - 10 mg/ML solution in water,” *Thermo Fisher Scientific - US*. [Online]. Available: https://www.thermofisher.com/order/catalog/product/H3570#:~:text=Hoechst%2033342%20nucleic%20acid%20stain,studies%20in%20combination%20with%20BrdU. [Accessed: 05-Jun-2022].

9. T. R. Donahue, L. M. Tran, R. Hill, Y. Li, A. Kovochich, J. H. Calvopina, S. G. Patel, N. Wu, A. Hindoyan, J. J. Farrell, X. Li, D. W. Dawson, and H. Wu, “Integrative survival-based molecular profiling of human pancreatic cancer,” *American Association for Cancer Research*, 01-Mar-2012. [Online]. Available: https://aacrjournals.org/clincancerres/article/18/5/1352/77508/Integrative-Survival-Based-Molecular-Profiling-of. [Accessed: 06-Jun-2022].

10. “Data Sheet, GeneChip® human genome arrays - affymetrix.” [Online]. Available: https://www.affymetrix.com/support/technical/datasheets/human_datasheet.pdf. [Accessed: 06-Jun-2022].