Integrated Bioinformatics Analysis of Gene Expression Profiles for Potential Biomarker Identification Towards Early Therapeutic Intervention in Pancreatitis and Pancreatic ductal adenocarcinoma

Manoj M Wagle
Department of Bioinformatics, Manipal School of Life Sciences, Manipal Academy of Higher Education, Manipal, India - 576104
https://orcid.org/0000-0001-5339-1453

Ananya Rao Kedige
Department of Bioinformatics, Manipal School of Life Sciences, Manipal Academy of Higher Education, Manipal, India - 576104
https://orcid.org/0000-0002-5337-2209

Shama P Kabekkodu
Department of Cell and Molecular Biology, Manipal School of Life Sciences, Manipal Academy of Higher Education, Manipal, India - 576104
https://orcid.org/0000-0002-4158-3893

Sandeep Mallya (✉ sandeep.mallya@manipal.edu)
Department of Bioinformatics, Manipal School of Life Sciences, Manipal Academy of Higher Education, Manipal, India - 576104
https://orcid.org/0000-0001-6318-7165

Research Article

Keywords: Pancreatic Ductal Adenocarcinoma, Pancreatitis, Inflammation, Pancreatic cancer, Microarray analysis, Differentially Expressed Genes

Posted Date: March 30th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-879048/v2

License: ☑️ ☑️ This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Abstract

Pancreatic ductal adenocarcinoma (PDAC) is a highly aggressive malignancy associated with rapid progression and an abysmal prognosis. It is reported that chronic pancreatitis can increase the risk of developing PDAC by 16-fold. The overarching hypothesis is that some of the biological processes disrupted during the inflammatory stage show significant dysregulation, even in cancer. This might explain why chronic inflammation increases the risk of carcinogenesis and uncontrolled proliferation. Here, we try to pinpoint such processes by comparing the expression profiles of pancreatitis and PDAC tissues. The gene expression datasets were retrieved from the EMBL-EBI ArrayExpress and NCBI GEO database. A total of 172 samples of normal pancreatic tissue, 68 samples of pancreatitis, and 306 samples of PDAC were used in this study. The differentially expressed genes (DEGs) identified were used to perform downstream analysis for ontology, interaction, and associated pathways.

Further, we validated hub gene expression, potential druggability, and the associated prognostic value. We also checked for promoter methylation and performed expression analysis based on different factors (such as gender, patient's drinking habit, race, and pancreatitis status). Our study identified a total of 45 genes found to have altered expression levels in both PDAC and pancreatitis. Over-representation analysis revealed that protein digestion and absorption, ECM-receptor interaction, PI3k-Akt signaling, and proteoglycans in cancer pathways as significantly enriched. Module analysis revealed 15 hub genes, of which 14 were found to be in the druggable genome category. Through bioinformatics analysis, we have identified crucial genes and biochemical processes disrupted at a molecular level. These results can provide valuable insights into PDAC treatment, and intervening therapeutically at an earlier stage by targeting these specific processes can be an effective strategy to decrease the incidence and severity of PDAC.

1. Introduction

Pancreatic ductal adenocarcinoma (PDAC) is one of the most prevalent forms of pancreatic cancer worldwide, with a bad prognosis accounting for approximately 90% of neoplastic diseases of the pancreas [1]. Patients with the malignancy rarely present the symptoms resulting in a poor diagnosis and increased mortality rates. Despite advancements in the treatment options, the five-year overall survival rate is roughly around 8% making it the 4th common cause of cancer-related deaths [2]. Although lifestyle factors such as age, alcohol consumption, tobacco use, and obesity play a vital role in the disease, family history and genetic susceptibility also account for ~10% of pancreatic cancers [3, 4]. The incidence of PDAC in both males and females is higher in developed countries than in developing countries. Some studies have estimated that PDAC will become the second most common cause of cancer-related mortality by 2030 [5, 6]. With a poorly understood etiology, potential treatment options for the management of PDAC include surgical resection (such as pancreaticoduodenectomy or Whipple procedure, total pancreatectomy), adjuvant chemotherapy, and radiation therapy [7, 8].

The aggressive biology and the complicated tumor microenvironment often promote metastasis microscopically, making it challenging to treat. In addition, gene instability, pre-existing cancer stem cells, and alterations in multiple signaling pathways result in an intrinsic chemoresistance that hinders the therapeutic drug delivery [9–11]. Dysregulation of molecular pathways such as K-Ras occurs in 75-90% of pancreatic carcinomas, further stimulating downstream signaling cascades [12, 13]. Moreover, mutations in the transcription factor P53 (TP53) gene can be seen in more than ~60% of pancreatic cancers [14]. Recent studies have shown how inflammation and an elevation in inflammatory cytokines often play a role in developing various cancers. PDAC is associated with significant peri and intra-tumoral inflammation and epithelial-mesenchymal transition (EMT) induction that serves as key mediators contributing to tumor initiation and its rapid progression. This is especially true in cases of chronic pancreatitis (an inflammatory pathophysiological disease of the pancreas), with the predisposing genes associated with a higher risk of developing pancreatic cancer. Hence, it is crucial to identify the underlying mechanisms, cellular processes, and inflammatory pathways, which can further help us design drugs targeting these biomarkers [15, 16].

The present study analyzes the microarray data of PDAC and pancreatitis tissues from publicly available datasets to derive the biological meaning of differentially expressed genes (DEGs) using different bioinformatics methods. The results of this study provide valuable insights which could be further explored to identify novel therapeutic targets.

2. Material And Methods

2.1 Retrieval of Gene expression datasets

The microarray data for normal, PDAC, and pancreatitis tissues were obtained from NCBI Gene Expression Omnibus (GEO) and EMBL-EBI ArrayExpress. A total of 6 datasets (GSE15471, GSE32676, GSE46234, E-MTAB-1791, E-GEOD-71989, and E-MEXP-2780) were analyzed in
this study [17, 18]. Since data was generated using different platforms, all the datasets belonging to the respective platforms (Affymetrix GPL570 [HG-U133_Plus_2] or Illumina human WG6 BeadChip v3) were processed and analyzed independently. The results obtained were later pooled for a more comprehensive analysis. The detailed description of the methodology followed in the study is represented in Fig. 1.

2.2 Data pre-processing and DEG screening

The datasets were pre-processed, normalized, and analyzed for differential expression using BRB-Array tool 4.6.1 (Stable Version) [19]. The DEGs screened conformed to the following cutoff criteria: \(|\log FC| > 2\) and a high significance threshold of 0.001 of univariate tests. Overlapping DEGs between pancreatitis and the PDAC samples was identified using the Funrich software [20].

2.3 Ontology and Pathway enrichment analysis

Gene ontology (GO) terms describe non-overlapping information on biological process (BP), cellular component (CC), and molecular function (MF) of individual gene products [21]. In contrast, the ontologies and comprehensive information on human diseases are described in the Disease Ontology (DO) [22]. KEGG is a database resource encompassing the functional meaning of a biological system derived mainly from high-throughput experiments [23]. We used the R Bioconductor package, clusterProfiler, which integrates the data from the above resources to perform ontology and enrichment analysis [24–28].

2.4 Protein-Protein Interaction (PPI) and Module analysis

PPIs are crucial for several biological functions in the body, and any dysregulation can often indicate diseases [29]. In this study, we used the Search Tool for the Retrieval of Interacting Genes (STRING) database and the Cytoscape software (version 3.8.2) for the construction and visualization of interaction networks [30, 31]. MCODE (Molecular Complex Detection) was used to identify the densely interconnected regions in the network with the following analysis parameters: node score cutoff = 0.2, k-score = 2, degree cutoff = 2, node density cutoff = 0.1, and max depth = 100 [32].

2.5 Pathway reanalysis, Potential druggability, and Gene expression analysis

The hub genes obtained were then reanalyzed to identify core genes. The potential druggability was determined using the Drug-Gene Interaction Database (DGIdb). DGIdb is a web-based resource providing information about druggable candidate genes and potential drug-gene interactions [33]. The expression of hub genes was evaluated using the GPEIA2 tool. GEPIA2 performs gene expression analysis using the data from The Cancer Genome Atlas (TCGA) and Genotype-Tissue Expression (GTex) [34], which helps us to validate and correlate the expression profiles between normal and PDAC tissues. The results of the KEGG pathway reanalysis were visualized using the R package, circlize [35].

2.6 Survival Analysis, Tumor subgroup expression analysis, and Promoter methylation

Survival analysis of the hub genes was performed using the Kaplan-Meier (KM) plotter [36]. We also performed hub gene expression analysis based on factors such as gender, patient's drinking habit, race, and pancreatitis status using the UALCAN tool. Further, the promoter methylation profile of individual hub genes was compared between normal and PDAC tissues [37].

3. Results

3.1 Identification of Differentially Expressed Genes

A total of 6 datasets were retrieved that included 172 samples of normal pancreatic tissue, 68 samples of pancreatitis, and 306 samples of PDAC. The microarray analysis was performed to identify the differentially expressed genes in PDAC and pancreatitis. The results of these studies were then combined to identify 45 genes that were differentially expressed in both PDAC and pancreatitis.
3.2 Enrichment analysis and KEGG pathway analysis

Ontology analysis and KEGG pathway enrichment analysis for the 45 DEGs were conducted using the R Bioconductor package, clusterProfiler, with the criterion set at p < 0.05. Gene Ontology analysis showed that (i) The most enriched biological processes were extracellular matrix (ECM) organization, extracellular structure organization, ossification, cell-substrate adhesion, and collagen fibril organization (Fig. 2a). (ii) In the molecular function group, the DEGs were mainly enriched in collagen binding, growth factor binding, EMSC conferring tensile strength, and glycosaminoglycan binding (Fig. 2b). (iii) In the cellular component group, the DEGs were significantly connected with the collagen-containing ECM, collagen trimer, and its complex and endoplasmic reticulum lumen. (Fig. 2c).

As shown in Fig. 3a, Disease Ontology analysis indicated that the DEGs were significantly associated with lung disease, cell type benign neoplasm, and pancreatic cancer. As for the KEGG pathway analysis, protein digestion and absorption pathway, PI3k-Akt signaling pathway, ECM-receptor interaction pathway, and proteoglycans in cancer pathways were significantly enriched (Fig. 3b).

3.3 PPI network and Identification of hub genes

The Protein-protein interaction network (PPI) constructed for the overlapping DEGs using the STRING database with a combined score > 0.4 (default threshold) showed 45 nodes with a total of 152 edges, representing a densely interconnected network (Fig. 4a). Module analysis using MCODE revealed 15 hub genes (with 92 edges), including - COL6A3, COL1A1, FBLN1, COL8A1, THBS2, CDH11, COL5A2, SPARC, COL3A1, THBS1, COL6A1, LUM, COL1A2, COL6A2, and COL5A1 (Fig. 4b).

3.4 Pathway reanalysis, Druggability, and Gene expression analysis

The 15 hub genes identified were then reanalyzed for KEGG pathways, and the following five core genes were identified - COL1A1, THBS1, COL1A2, THBS2, and COL3A1 (Fig. 5). Among these, COL1A1 and COL1A2 were associated with 11 different pathways each. Expression analysis using the GEPIA2 tool between normal and PDAC tissues showed that all 15 hub genes were found to be significantly expressed (P-value < 0.001 and Log2FC > 2) (Fig. 6). The potential druggability of the hub genes was determined using the DGIdb database. It was found that 14 out of 15 genes were in the druggable genome category, suggesting that they could be modulated and interact with small molecules. The complete list of genes and their corresponding druggable gene category is shown in Table 1.

<table>
<thead>
<tr>
<th>Druggable gene category</th>
<th>Gene Count</th>
<th>Gene(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Druggable genome</td>
<td>14</td>
<td>COL6A3, COL1A1, FBLN1, COL8A1, THBS2, COL5A2, SPARC, COL3A1, THBS1, COL6A1, LUM, COL1A2, COL6A2, COL5A1</td>
</tr>
<tr>
<td>Cell surface</td>
<td>2</td>
<td>SPARC, THBS1</td>
</tr>
<tr>
<td>Clinically actionable</td>
<td>2</td>
<td>COL1A1, CDH11</td>
</tr>
<tr>
<td>Drug resistance</td>
<td>2</td>
<td>COL1A1, THBS1</td>
</tr>
<tr>
<td>External side of plasma membrane</td>
<td>1</td>
<td>THBS1</td>
</tr>
<tr>
<td>Protease inhibitor</td>
<td>1</td>
<td>COL6A3</td>
</tr>
<tr>
<td>Transcription factor</td>
<td>1</td>
<td>THBS1</td>
</tr>
</tbody>
</table>

Table 1 Of the 15 hub genes analyzed for druggability using the DGIdb database, a total of 14 were found to be in the druggable genome category indicating their use as potential drug targets.
3.5 KM Survival Analysis, Tumor subgroup & Promoter methylation

The prognostic value associated with hub genes was analyzed using the KM plotter at a P-value threshold of < 0.05 (Fig. 7). The results showed that the genes - COL6A1, COL6A3, COL8A1, LUM & THBS2 caused a significant reduction in the overall survival rate of PDAC patients, with COL6A1 being the most statistically significant (log-rank P = 0.0061). Next, the tumor subgroup analysis of the hub genes was performed. The analysis based on gender did not reveal any notable differences in the gene expression between male and female PDAC patients (Suppl. Fig. 1). Similarly, the expression values did not vary much with or without the presence of chronic pancreatitis for most of the hub genes. However, higher transcript per million (TPM) values were observed for the gene - COL6A1 in patients with pancreatitis than in normal and non-pancreatitis patients (Suppl. Fig. 2). Although not statistically significant, samples from ‘occasional drinkers’ showed higher TPM values compared to other groups. But this can be overlooked based on the observation that the data from occasional drinkers cover a greater range indicating highly probable values making it less reliable to conclude (Suppl. Fig. 3). Expression analysis based on different races showed that African Americans exhibited higher median TPM values, with gene - FBLN1 being most statistically significant compared to Asians, Caucasians, and normal samples. The observations and comparisons between the races are biased due to differences in sample count and thus have low statistical significance (Suppl. Fig. 4). Evaluation of regulation of gene expression by promoter methylation revealed no significant change in the methylation profiles for most of the hub genes, except gene - COL3A1, which showed a deviation compared to normal samples. No methylation profile data was available for the gene - LUM (Suppl. Fig. 5).

4. Discussion

Although there has been a massive improvement in the treatment for PDAC, mortality and incidence rates are still increasing at an unprecedented rate. Several studies have been carried out to uncover the molecular mechanisms involved in the onset, growth, and progression of PDAC. It has also been reported that chronic cases of pancreatitis can increase the risk of developing PDAC by 16-fold [38, 39]. Our study aims to identify essential genes, pathways, and interactions mediating both pancreatitis and PDAC. Analysis of the datasets revealed that the 45 overlapping DEGs were mainly enriched in collagen and growth factor binding, extracellular environment, and cell adhesion. Collagen is a crucial component of the extracellular matrix (ECM), and specific orientation and arrangements of ECM in a microscopic environment are thought to play essential roles in tumor progression [40, 41]. This disruption in the ECM homeostasis can be caused by degradation and even deposition of collagen. Since tumor cells continuously interact with ECM, an increased disruption can accelerate tumor progression by negatively interfering with cell adhesion [42–45].

KEGG pathway analysis showed that the protein digestion and absorption pathway, ECM-receptor interaction pathway, PI3k-Akt signaling pathway, and proteoglycans in cancer pathways might play essential roles in the progression of PDAC. Aberration of the PI3k-Akt signaling pathway can be seen in many different cancers. Moreover, an increase in Akt activity is regularly seen in PDAC (~60% of cases) due to the loss of key regulators or mutations. K-Ras is an essential gene of the RAS/MAPK pathway (required for proliferation and maturation of cells), and activating mutations in this gene can be seen in ~95% of pancreatic cancers, which further activates PI3K signaling [46–48]. These are the major reasons why targeting the PI3k-Akt pathway has been a significant interest in cancer drug discovery. Proteoglycans are another important biomolecule of interest, having multiple functions in angiogenesis and cancer. They often influence cell growth through their interaction with growth factors and can sometimes cause deregulation of cell proliferation [49, 50]. Thus, their integration in tumor cell diagnostics can facilitate early diagnosis, as demonstrated in a few studies [51, 52].

We further constructed a protein-protein interaction network and performed a module analysis. The module consisted of 15 nodes with 92 edges. Expression analysis using the GEPIA2 tool revealed all 15 hub genes to be significantly expressed in PDAC. Further, the expression of most of the hub genes was independent of factors such as gender, drinking habits, race, and pancreatitis status, which suggests that these genes can be used as biomarkers on a global scale for advancing PDAC treatment. Potential druggability determined using the DGIdb database showed that 14 out of 15 genes are in the druggable genome category and thus have a potential value for developing targeted drugs. Next, to understand which genes were significantly involved in the pathways analyzed before, we performed KEGG pathway reanalysis for the 15 hub genes. Based on this, we identified five core genes - COL1A1, COL1A2, THBS1, THBS2, and COL3A1. Among these, three were protein-coding collagen genes. Recently, a few studies have demonstrated how various collagen genes can play a role in tumorigenesis leading to poor clinical outcomes [53, 54]. Further, differential expression of genes COL1A1, COL1A2, and THBS1 has been reported in several cancers, including colorectal cancer, hepatocellular carcinoma, and melanoma [55–57]. Some researchers have demonstrated the utilization
of Thrombospondin-2 (THBS2) as a biomarker for risk prediction and early detection of PDAC and as a robust prognostic indicator in colorectal cancer [58, 59].

Our study presents useful findings elucidating the role of inflammation in carcinogenesis and its progression. The clinical significance of this study can be further validated by in-depth experimental research. However, there are a few limitations to this study. Firstly, this study compared pancreatitis and PDAC samples but did not consider the stage of individual samples. Secondly, our results are limited to bioinformatics findings, and we did not analyze the clinical data of samples due to inaccessibility.

5. Conclusion

Here, we present a comprehensive analysis of the gene expression profiles using an integrated bioinformatics approach to identify critical genes, biochemical processes, and various pathways disrupted in both pancreatitis and PDAC. Literature studies indicate how inflammation can serve as a risk factor for developing various cancers. We majorly identified 15 hub genes and associated biological processes that might play crucial roles in events leading to cancer initiation. Our study suggests that targeting these specific processes at an earlier stage can be used as one of the effective strategies to decrease the incidence and severity of PDAC. These personalized therapies can significantly reduce the sequelae of cancer treatment while improving patients’ quality of life and overall survival rate.

Declarations

Availability of data and material

The authors declare that the data analyzed during the study are available in the NCBI Gene Expression Omnibus [GSE15471, GSE32676, GSE46234] and EMBL-EBI ArrayExpress [E-MTAB-1791, E-GEOD-71989, and E-MEXP-2780].

Credit authorship contribution statement

MMW - Conceptualization, Methodology, Formal analysis, Investigation, Writing - original draft, review & editing; ARK - Methodology, Formal analysis, Investigation; SPK - Conceptualization, Supervision, Validation, Project administration, Resources; SM - Conceptualization, Methodology, Supervision, Validation, Project administration, Resources, Writing - review & editing. All authors approved the final version of the manuscript.

Ethics approval:

Not applicable

Competing Interest:

None

Consent to participate:

Not applicable

Consent for publication:

Not applicable

Funding:
This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Acknowledgment

The authors wish to thank DST-FIST, Government of India, TIFAC-CORE in pharmacogenomics, and the Manipal Academy of Higher Education, Manipal, India, for providing the necessary facilities and infrastructure for the study. We would also like to thank Dr. K. Satyamoorthy (Director, Manipal School of Life Sciences, MAHE, Manipal) and Dr. Bobby Paul (Assistant Professor and Head, Department of Bioinformatics, Manipal School of Life Sciences, MAHE, Manipal) for their support and suggestions towards valuable additions to the manuscript.

References


Figures

**Figure 1**

Flowchart describing the steps to obtain DEGs and their downstream analysis. This involved collection of raw data and preprocessing, screening and identification of overlapping DEGs, ontology and pathway enrichment analysis, protein-protein interaction, and analysis of hub genes.
Figure 2

**Fig. 2a** Dot plot representing GO - biological process, where the DEGs were mainly enriched in ECM organization, extracellular structure organization, ossification, cell-substrate adhesion, and collagen fibril organization. **Fig. 2b** Dot plot representing GO - molecular function. The DEGs were mainly enriched in collagen binding, growth factor binding, extracellular matrix structural constituent (EMSC), EMSC conferring tensile strength, and glycosaminoglycan binding. **Fig. 2c** Dot plot representing GO - cellular process, where the DEGs were significantly associated with the collagen-containing ECM, collagen trimer and its complex, and endoplasmic reticulum lumen.
**Figure 3**

**Fig. 3a** Bar plot representing disease ontology analysis for the 45 DEGs common to both pancreatitis and PDAC. The DEGs were significantly associated with lung disease, cell type benign neoplasm, and pancreatic cancer. **Fig. 3b** Bar plot representing KEGG Pathway Analysis for the 45 DEGs common to both pancreatitis and PDAC. The DEGs were mainly enriched in protein digestion and absorption pathway, PI3k-Akt signaling pathway, ECM-receptor interaction pathway, proteoglycans in cancer pathways.

**Figure 4**

**Fig. 4a** Represents the protein-protein interaction (PPI) network constructed using the STRING database for the 45 DEGs common to both pancreatitis and PDAC. **Fig. 4b** Represents the module analysis of the PPI network constructed using the Cytoscape app, MCODE showing 15 nodes (hub genes) with 92 edges.
Figure 5

Chord diagram representing the KEGG pathway reanalysis for the 15 hub genes. A total of five core genes were identified - COL1A1, THBS1, COL1A2, THBS2, and COL3A1. Among these, the genes COL1A1 and COL1A2 alone were significantly associated with 11 different pathways each.
Figure 6

Box plots comparing the gene expression levels of PDAC and normal tissues for 15 hub genes P-value < 0.001 and Log2FC > 2. The samples from normal tissues are shown in grey and PDAC tissues in red. Gene expression changes between groups in all hub genes were found to be significant (marked with an *).
Figure 7

Kaplan-Meier plots representing the survival analysis for 15 hub genes with respect to low expression (black color) and high expression (red) in PDAC tissue samples. Among these, genes - COL6A1, COL6A3, COL8A1, LUM, and THBS2 (marked with an *) are statistically significant P-Value < 0.05.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterial.pdf