Discovering Key Transcriptomic Regulators in Pancreatic Ductal Adenocarcinoma using Dirichlet Process Gaussian Mixture Model

Sk Md Mosaddek Hossain^{1,2,+}, Aanzil Akram Halsana^{1, +}, Lutfunnesa Khatun², Sumanta Ray^{1, *}, and Anirban Mukhopadhyay^{2, *}

¹Aliah University, Computer Science and Engineering, Kolkata, 700160, India

²University of Kalyani, Computer Science and Engineering, Kalyani, 741235, India

*sumantababai86@gmail.com, anirban@klyuniv.ac.in

⁺these authors contributed equally to this work

ABSTRACT

Pancreatic Ductal Adenocarcinoma (PDAC) is the most lethal type of pancreatic cancer (PC), late detection of which leads to its therapeutic failure. This study aims to find out key regulatory genes and their impact on the progression of the disease helping the etiology of the disease which is still largely unknown. We leverage the landmark advantages of time-series gene expression data of this disease, and thereby the identified key regulators capture the characteristics of gene activity patterns in the progression of the cancer. We have identified the key modules and predicted gene functions of top genes from the compiled gene association network (GAN). Here, we have used the natural cubic spline regression model (splineTimeR) to identify differentially expressed genes (DEG) from the PDAC microarray time-series data downloaded from gene expression omnibus (GEO). First, we have identified key transcriptomic regulators (TR) and DNA binding transcription factors (DbTF). Subsequently, the Dirichlet process and Gaussian process (DPGP) mixture model is utilized to identify the key gene modules. A variation of the partial correlation method is utilized to analyze GAN, which is followed by a process of gene function prediction from the network. Finally, a panel of key genes related to PDAC is highlighted from each of the analyses performed.

Please note: Abbreviations should be introduced at the first mention in the main text – no abbreviations lists. Suggested structure of main text (not enforced) is provided below.

Introduction

In genetics, gene expression is one of the elementary constitutional blocks which gives rise to a phenotype from a genotype, i.e., a trait which is observable in all living cells including prokaryotes and eukaryotes. Multiple techniques are available to quantify gene expression and its regulation like DNA microarray, RNA seq, etc. The area of gene expression analysis has undergone several major advancements in biomedical research. With increased efficiency and quality, these measurements have led to improvements in disease sub-classification, gene identification problems, and studying progression characteristics of diseases ¹⁻⁷. Biological mechanisms are dynamic in nature, therefore its activities must be supervised at multiple time points. Time-series gene expression experiments are used widely used to monitor biological processes in a time-series paradigm ⁸. Analyzing these time-series gene expression data helps in identifying transient transcriptional changes, temporal patterns, and causal effects of genes. Time-series gene expression studies can be utilized to predict phenotypic outcomes over a period of time ⁹.

DNA microarrays and RNA-Seq data have been accepted as gold standards for analyzing and measuring gene expressions across different biological circumstances ^{3,5,10}. A gene is considered differentially expressed (DE) if a statistically significant difference in gene expression levels is observed between two experimental conditions. Various statistical distribution models like the Poisson and the Negative Binomial (NB) distribution are used to estimate the patterns of differential gene expression. Gene selection refers to the detection of the most significant differentially expressed genes under various conditions ¹¹. Selection is done based on a combination of score cutoff and expression change threshold, which are commonly generated by the statistical design itself ¹². Popular time course DE analysis tools include edgeR/DESeq2, TimeSeq, and Next maSigPro which are based on NB distribution model. Some DE tools like ImpulseDE2 and splineTimeR, which are based on impulse model and spline regression model between two groups respectively, are used on short time-series data ¹³.

Gene expression is a strongly regulated spatio-temporal process. Genes having identical expression patterns are associated with the same biological function. Clustering genes with similar expression pattern reduces the transcriptional response

complexities by grouping genes responsible for a distinct cellular process^{14,15}. Several statistical clustering techniques have been widely used like k-means, hierarchical clustering^{16,17} and self-organizing maps¹⁸ to produce time-series modules. Gradually, various techniques have been developed especially for clustering time-series data. In¹⁹, Short Time-series Expression Miner (STEM) has been used as a clustering technique that maps genes to their representative expression profile. Cluster Analysis of Gene Expression Dynamics (CAGED), a clustering technique proposed by *Ramoni et al.*²⁰, uses Bayesian method to model gene-expression dynamics using auto-regressive equations. TimeClust²¹ uses temporal gene expression profiles to produce clusters. TMixClust²², Dirichlet process Gaussian process mixture model¹⁴ are some of the significant clustering approaches containing methods with non-parametric models.

The analysis of time-series gene expression modules has helped us unravel major biological complications. It has given us serious insights into disease progression²³, biomarker discovery²⁴, identification of hub genes²⁵, cell cycle progression²⁶, cancer classification²⁷ and several other bio-medically important processes. Moreover, the advancements in information system infrastructure have facilitated creating time-series schemes more feasible method for studying complex psychological phenomena. Multiple tools are available to provide us with an enriched network and pathway analysis of these gene modules. Using such tools enables us for further analysis and deeper observation of the biological mechanisms.

In this article, we have proposed a framework to discover key transcriptomic regulators and key modules containing majority of the key regulators from time-series microarray gene expression data for pancreatic ductal adenocarcinoma. Initially, differentially expressed genes (DEGs) were identified by analyzing empirical Bayes statistics on multivariate time-course gene expression data of pancreatic cancer in R (splineTimeR)²⁸. Also top 100 DEGs at each time point were analyzed using an R/Bioconductor software package called Linear Models for Microarray Data (Limma)²⁹. Clustering on the DEGs to form gene modules based on similar responses was achieved by a non-parametric model, Dirichlet process Gaussian process mixture model¹⁴ in Python. A web tool, REGulator-Gene Association Enrichment (REGGAE)³⁰ was used for the identification of key transcriptional regulators along with the number of targets for each of them from the list of DEGs.

Most of the experimental gene expression studies only focus on the determination of DEGs by handling them as independent events and do not investigate the co-action of identified genes. Reconstruction of the possible gene association network (GAN) among DEGs helps us find genes that take part in the interaction network of studied phenotype. Therefore, reconstruction of GAN followed by the identification of top nodes in the network was also performed. Prediction of gene functions of the top genes in the interaction network has also been carried out using GeneMANIA prediction server³¹. We have identified the key gene modules from the list of modules obtained from the clustering. Subsequently, transcriptomic regulatory genes were also detected against a curated database of DNA-binding RNA polymerase II TF (DbTF) using TFcheckpoint³². Furthermore, biological significance like Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway, Gene Ontology (GO) and gene-disease associations of the gene modules were also observed using a gene set enrichment analysis tool, Enrichr³³.

1 Results and Discussion

This section gives us insight about the detailed findings of our present work.

1.1 Evaluation of Differential Expression

We have processed 42412 genes with their normalized expression values from methods discussed in section 2.1. These genes along with their normalized expression values were analyzed for differential expression. Differential gene expression analysis on genes has been executed using splineTimeR as described in section 2.2. We obtained 1397 genes that were differentially expressed by using adjusted P -value ≤ 0.05 using the correction method of Benjamini-Hochberg (BH)³⁴ with 4 as the optimum degree of freedom. Top 20 genes (DEG's) from the pancreatic ductal adenocarcinoma dataset with significant expression value changes were 'Hs.7413', 'CYP26B1', 'NPPB', 'SFRP4', 'Hs.562219', 'C12ORF46', 'SMAD3', 'RN7SK', 'RARRES1', 'ITGA4', 'CLSTN2', 'DKK1', 'CYP26A1', 'EPDR1', 'RARB', 'LOC340598', 'TMEM16C', 'INMT', 'ACTC1', 'C21ORF7'. Figure 8 represents the box and violin plot of the cubic spline normalized gene expressions of samples for each condition at individual time point of the differentially expressed genes. We have also performed DE analysis for all the genes at each time point to distinguish the genes expressed at each time point using Limma package in R. Table 4 shows the top 5 DEGs at each time point, including the top 5 genes of the final DEG list, found using Limma.

We have analyzed the genes in Limma by applying empirical Bayes smoothing on the estimated fold-changes and standard errors from a linear model fitting. We have taken the top 100 genes from each time point using BH adjustment on adj. P -values to determine DEGs. The common differentially expressed genes at each time point have been identified. DEG list of the top 100 genes had the highest similarity of 59% with the top 100 genes of the final list at 168 hours amongst all the time points, followed by 41% at 24 hours. Figure 1 represents the Venn diagram of the top 100 DEGs at each time point with no genes at the intersection of all time points and DEG _{t} stands for the top 100 DEGs at time point t , where $t \in \{30 \text{ Minutes}, 4 \text{ Hours}, 12 \text{ Hours}, 24 \text{ Hours}, 168 \text{ Hours}\}$. DEGs from splineTimeR and top 1397 genes from the final table of Limma were 100% similar.

Figure 1. Venn Diagram of number of DEG's at each time point

1.2 Gene Network Reconstruction and Gene Function Prediction

Figure 2. Reconstructed Gene Association Network using splineNetRecon with top 150 edges based on partial correlation score.

Adopting the methodologies discussed in section 2.3, we have reconstructed the gene association network (GAN). We have obtained two GAN's. We got 45550 significant edges with 1107 nodes for cutoff probability 0.8 and 34080 significant

edges with 1048 nodes for cutoff probability 0.9, which corresponded to 4.67% and 3.5% of possible edges respectively. Figure 2 depicts the reconstructed GAN with a probability cutoff set to 0.9 for the top 150 selected edges based on a higher partial correlation score. Genes with higher degrees in the whole reconstructed GAN are displayed in dark color. From this figure, it can be observed that 'CUEDC1', 'ABCA6', 'MRPL50', 'LYPD3', 'KRT19', 'OLFML3', 'LGALS3', 'Hs.540914', 'LOC285074', 'TBPL1' are the top 10 genes having extremely high connection with others within the GAN. We have also generated reconstructed GAN highlighting the betweenness and closeness centralities for the top 150 selected edges based on higher partial correlation scores which are shown in **Supplementary File 1**. It was discovered that 'CUEDC1', 'ABCA6', 'MRPL50', 'LYPD3', 'KRT19', 'OLFML3', 'TBPL1' were the seven common genes which ranks within top 10 using all of these three centrality measures.

We have also extracted the top 150 genes from the reconstructed GAN with 0.9 as the probability cutoff for prediction of their gene function, chosen based on the aggregation of three centrality measures: degree, betweenness, and closeness centrality. The top 15 genes among them include 'CUEDC1', 'ABCA6', 'MRPL50', 'LYPD3', 'KRT19', 'OLFML3', 'LGALS3', 'TBPL1', 'LOC285074', 'ANKRD26', 'FDXR', 'Hs.540914', 'CCL2', 'CENPJ', and 'CXCL2'. We utilized the list of genes to obtain their gene function using GeneMANIA web server in Cytoscape app. GeneMANIA resulted in the creation of a weighted connected network of query genes including the suggested genes by the algorithm. The resultant network is demonstrated in Figure 3. We have also obtained the predicted gene functions by GeneMANIA. Table 1 provides us with comprehensive details of the top 5 functions of the resultant genes.

Table 1. GeneMANIA predicted functions

	Function	FDR	Coverage
1.	Cellular response to tumor necrosis factor	0.0015	8/74
2.	Response to tumor necrosis factor	0.0027	8/87
3.	Response to transforming growth factor beta	0.025	9/169
4.	Cellular response to transforming growth factor beta stimulus	0.025	9/169
5.	Cellular response to interleukin-1	0.026	5/36

In table 1, the coverage field defines a ratio of the number of genes in the resultant network to that in the genome, and FDR represents the False Discovery Rate generated from GeneMANIA algorithm. The top 50 genes have been chosen from the resultant ranked gene list provided by GeneMANIA with the assigned score for finding their gene-disease associations. It has been observed that 'CSF2', 'HOXB9', 'ITGA11', 'NUMB' genes are associated with Pancreatic Ductal Adenocarcinoma according to DisGeNET³⁵ web server. The resultant gene list also contained 8 DNA-binding TF's including 'SMAD3', 'STAT1', 'ZNF34', 'ATF3', 'ZNF395', 'JARID2', 'E2F4', 'SOX9'. Results are described in section 1.3.

1.3 Transcriptional Regulators and DNA-binding Transcription Factor Identification

We have identified key transcriptional regulators from the generated DEG list using REGGAE³⁰. We obtained a ranked list of 66 key transcriptional regulators from an RTI collection which contains a list of regulators for each deregulated gene that may influence its expression. Genes were ranked according to the score provided by the REGGAE algorithm. The top 5 genes from the list of key TR's are 'GATA6', 'NFBYB', 'IRF1', 'TRIM22', 'SREBF1'. Among the generated list of key TR's, 5 TR's were not from the genes of the DEG list implying multiple regulators influencing a deregulated gene. Those 5 TR's are 'PEX2', 'FOXO1', 'RREB1', 'HEXIM1', 'ZNF280D'.

Additionally, we have treated the genes in the key TR list with a curated database of specific DbTFs: TFcheckpoint³², to find proteins playing a central role in determining which genes were transcribed. We retrieved 47 DNA-binding RNA polymerase II TFs among 66 key TR's. This 47 DbTFs contained 10 proteins that were directly associated with Pancreatic Ductal Adenocarcinoma (PDAC) disease. The proteins, according to DisGeNET, associated with this disease are 'FOXO1', 'SOX9', 'GATA6', 'SMAD3', 'NFKB1', 'KLF6', 'TBX3', 'SREBF1', 'NR4A2', 'TCF3'. In table 1, the coverage field defines a ratio of the number of genes in the resultant network to that in the genome, and FDR represents the False Discovery Rate generated from the GeneMANIA algorithm. The top 50 genes were chosen from the resultant ranked gene list provided by GeneMANIA with the assigned score for finding their gene-disease associations. It has been observed that 'CSF2', 'HOXB9', 'ITGA11', 'NUMB' genes are associated with Pancreatic Ductal Adenocarcinoma according to DisGeNET³⁵ web server. The resultant gene list also contained 8 DNA-binding TF's including 'SMAD3', 'STAT1', 'ZNF34', 'ATF3', 'ZNF395', 'JARID2', 'E2F4', 'SOX9'. Results are described in section 1.3.

Figure 3. Gene network created from GeneMANIA web server.

1.4 Gene Modules Identification and Determination of Key Gene Modules

After discovering DEGs and DbTFs from the list of DEGs, we have applied the Dirichlet Process Gaussian Process mixture model clustering to obtain robust and accurate gene modules incorporating their expression values at different time points. The optimal number of clusters obtained by using default conditions as discussed in section 2.4 was 32, with 136 and 10 genes being the highest and the lowest number of genes in a module among all the modules, respectively. Figure 4 shows the complex heatmap denoting the posterior distribution of the probability that two genes are being co-clustered.

Subsequently, we have used the modules and the genes inside each module to find the key modules. We have assigned rank to the clusters according to the descending order of the number of regulator genes inside each gene module. The top 4 modules with the highest number of regulators have been considered as the key modules. We have discovered that 'gene module' 2 contains 7 key TRs (now onward will be referred to as '*Key Cluster 1*'), 'gene module' 1 contains 6 key TRs (now onward will be referred to as '*Key Cluster 2*'), 'gene module' 9 contains 6 key TRs (now onward will be referred to as '*Key Cluster 3*') and 'gene module' 26 contains 6 key TRs (now onward will be referred to as '*Key Cluster 4*') are the 4 key modules. The key modules had 78, 136, 50, 64 number of genes in total for each cluster, respectively. Here, we also found out that only 24 gene modules out of 32 modules contained all the key TRs.

Furthermore, we have studied the normalized gene expressions in each of the 4 key modules for 5 time points to understand the magnitude of variation of expression values in the modules. Figure 5 gives us a visual representation of the average fold-change in gene expression along with individual gene trajectory for the 4 key clusters.

1.5 Biological Significance Analysis

We have performed several analysis on the key modules in order to gain more insights about the pathways and biological processes (BP) of the involved genes inside the key modules. Additionally, we have identified key TRs in each of the key module. We have identified KEGG pathways, biological processes (BP) details of the Gene Ontology terms and disease-gene associations using DisGeNet database of Enrichr web server³³. These information along with all the regulator genes and DbTF among the key TRs have been summarized in Table 2, where genes in bold face have been used to represent DbTFs among the TRs. Figure 6 shows the top ranked (based on *p*-value) KEGG pathways (6 A), biological processes (6 B) and drugs/disease association (6 C) enrichment of genes in key cluster 1. These analysis for the other 3 key modules has been attached in the **Supplementary File 1**.

Figure 4. Heatmap of posterior similarity matrix from clustering.

Table 2. Biological significance analysis of key regulators in each key module.

Regulator genes and DbTF among them	Key Cluster	Gene Ontology (Biological Process 2018)	Biological Pathway (KEGG Human 2019)
GATA6, NFYB, IRF1, TRIM22, SREBF1, RFX5, PEX2, HOXA4, SMAD3, ID3, STAT1, FOXO1, ZNF34, RING1, LYL1, TOP2B, ATF3, NR1H4, EGR2, HOXA5, HOXA2, RARB, HOXC13, DACH1, ARNT2, SMAD6, TAF15, FOXF1, RELB, ZNF395, FOSB, ZNF281, ID1, LIN54, CREB3L2, NKX3-1, MSX1, JARID2, TCF3, NR4A2, RREB1, HES4, HIVEP1, NFKB1, NCAPG, ZBTB20, NOTCH1, E2F4, KLF6, RBPJ, LMNA, EGR1, DNMT1, MYB, HEXIM1, ZNF280D, ELL2, PBX3, E2F7, POU2F2, KLF13, SOX9, TBX3, FOSL1, PRPF4, ZNF792	Key Cluster 1 : GATA6, IRF1, SREBF1, SMAD3, TOP2B, HOXA5, HOXA2	Regulation of transforming growth factor beta2 production (GO:0032909); Regulation of angiogenesis (GO:0045765)	Adherens junction; Pancreatic cancer
	Key Cluster 2 : NFYB, HOXA4, NR1H4, RARB, DACH1, TCF3	Muscle contraction (GO:0006936); Regulation of cellular biosynthetic process (GO:0031326)	Leukocyte transendothelial migration; Hypertrophic cardiomyopathy (HCM)
	Key Cluster 3 : STAT1, ZNF34, ATF3, EGR2, ZNF281, PBX3	Purine nucleobase metabolic process (GO:0006144); Regulation of JNK cascade (GO:0046328)	Hepatocellular carcinoma; Human papillomavirus infection
	Key Cluster 4 : FOXF1, MSX1, JARID2, HES4, POU2F2, TBX3	Embryonic digit morphogenesis (GO:0042733); Artery morphogenesis (GO:0048844)	TNF signaling pathway; IL-17 signaling pathway

Moreover, we have observed the top 5 KEGG pathways and GO terms (BP) of each of the key modules to understand their relevance and significance in PDAC. We have found several important factors essential for PDAC disease among the observed results. We observed that key module 1 regulates Transforming Growth Factor (TGF) beta2 production which leads to high damage for over-expressing TGF- β in PDAC³⁶. In key module 2, 'MYL7', 'SHC3', 'ITGA4', 'COL4A4', 'ITGA1', 'FLNA', 'MYL9' genes are significant in focal adhesion pathway which plays a key role being over-expressed and highly active in PDAC³⁷. In a publication, Gordon et. al. demonstrated that several bone morphogenesis signaling component targets has been up-regulated in pancreatic cancer tissues including genes like 'SFRP4', 'CYP26B1', 'RARB'³⁸. Key module 3 is responsible for the regulation of c-Jun N-terminal kinase cascade regulation for the presence of 'DUSP10', 'GADD45A', 'ANKRD6'

Figure 5. Statistics of gene expression in top 4 key modules (Key Clusters).

genes which is one of the pathways activated in PDAC³⁹. Regulation of the p38MAPK cascade pathway in key module 3 is involved in Interleukin 1 alpha (IL-1 α) expression by PDAC⁴⁰. We found that Interleukin-17 signaling pathway promoted changes from chronic pancreatitis to pancreatic cancer accompanied by ‘CSF2’, ‘CCL20’, ‘TNFAIP3’ genes in key module 4⁴¹. Importance of tumor necrosis factor, notch signaling pathway are also significant for promoting PDAC^{42,43}. Finally, we found that ‘TGFB2’, ‘SMAD3’, ‘RAC2’ genes in key module 1 were responsible for pancreatic cancer pathways.

Figure 7 shows the top 15 gene ontology terms (biological processes) performed by the genes in the ‘key module 1’ with their proportion of counts in the module. These GO terms have been selected according to the lowest p -values. This analysis for the other 3 key modules has been attached in the **Supplementary File 1**.

Additionally, we have analyzed genes in the top 4 key modules for their direct association with PDAC. We have observed the gene list from each key module against DisGeNet web server. We have found that 6, 8, 3, 7 genes were directly associated with PDAC in the key modules 1, 2, 3, 4, respectively. The names of the genes have been tabulated in Table 3.

Table 3. Genes related to PDAC in each key module.

Key Module 1	Key Module 2	Key Module 3	Key Module 4
GATA6, IGFBP3, SMAD3, SREBF1, TIMP3, GPC5A	IGF2BP3, CNN2, IDH1, CXCL12, MMP11, SFRP4, TCF3, NCOA3	ROR2, CLIC3, ITGA11	LGALS3, CSF2, ISG20, TBX3, NUMB, MTSS1, KHDRBS1

Figure 6. A. KEGG Pathway of *Key Module 1*, B. Biological Process of *Key Module 1*, C. DisGeNet analysis of *Key Module 1*

2 Methods

The present section provides the overview of our systematic approach on data collection, data preprocessing and the overall framework for the different methodologies used in our present analysis.

2.1 Data Preparation

Gene Expression Omnibus (GEO)⁴⁴ has been used to collect Pancreatic ductal adenocarcinoma microarray dataset (GSE14426), which contains treated pancreatic stellate cell line raw counts. The dataset has 48701 genes and 30 samples of each gene along with their periodic gene expression changes. Samples include cubic spline-normalized intensity values of two conditions, viz., pancreatic stellate cell line before and after being treated with 1-molecular concentration of all-trans retinoic acid (ATRA) on plastic. The gene expression values were recorded at 5 time points (30 min, 4 hours, 12 hours, 24 hours and 168 hours) with 3 replicates at each time point. All expression values were $\log_2$ -transformed for better performance in finding DEGs.

The R package *org.Hs.eg.db*, genome-wide annotation database for Homo Sapiens⁴⁵ was used to map the Illumina gene identifiers from the raw data file to their official gene symbols along with the Illumina human platform information (GPL6102) from GEO. We have obtained the final set of genes after mapping with official gene symbols, considering all gene names found from 1:1 as well as 1:many mapping and aggregating expression values of duplicate genes from the total number of genes.

2.2 Differential Expression Analysis

In order to analyze and extract the genes with significant gene expression changes, differential gene expression analysis has been performed using the R/Bioconductor *splineTimeR* package. SplineTimeR operates on the values obtained from the parameters of a fitted natural cubic spline regression model. The regression model is fitted with time-dependent gene expression data for two expression groups (e.g. control and treated)²⁸. The differential expression of a gene has been discovered by applying empirical Bayes moderate F-statistics on the coefficient differences of the spline regression model. The mathematical natural

Figure 7. Lollipop plot describing percentage of gene in *key module 1* contributing to top 15 Gene Ontology terms

cubic spline regression model is defined as:

$$f = f(t, z) = s_0 + \sum_{i=1}^n s_i B_i(t - t_0) + z(d_0 + \sum_{i=1}^n d_i B_i(t - t_0)) \quad (1)$$

where s_i and d_i is the reference group's spline coefficient and compared group's differential spline coefficient, respectively and $B_i(t - t_0)$ is spline base function $\forall i = 1, 2, \dots, n$ with t_0 being the first measurement time.

$$f = \begin{cases} f_{\text{reference}}, & \text{if } z = 0 \\ f_{\text{compared}}, & \text{if } z = 1 \end{cases} \quad (2)$$

The detection of DEGs using splineTimeR has been carried out by setting the Benjamini-Hochberg³⁴ adjusted P -value threshold to 0.05 and with a degree of freedom as 4 for all genes. Top regulated genes have been identified as differentially expressed and used for further analysis. Figure 8 represents the box and violin plot of the cubic spline normalized gene expressions of samples for each condition at the individual time point of the differentially expressed genes.

Additionally, the Limma²⁹ software package has been utilized to find top DEG's at each time point considering the replicates of that particular time point. Limma fits linear models to gene expression value matrix to assess differential expression by taking care of the differences among gene sample conditions. A linear model can be defined as:

$$\beta[e_g] = D \cdot \Gamma_g, \quad (3)$$

where e_g represents a vector of expression values and D represents the design matrix which maps these vector to some coefficient of interest Γ_g for each gene g . A contrast matrix was also used for computing $\log_2$ -fold-changes and t -statistics by allowing

Figure 8. Box and Violin plots of cubic spline normalized gene expression of the differentially expressed genes.

all possible pairwise comparisons between gene samples. Furthermore, test-statistics are obtained to perform gene ranking. Normalized expression values for 3 replicates at each of the 5 time points were used for differential expression analysis.

Table 4 shows top 5 DEGs at each time point, including the top 5 genes of the final DEG list, found using Limma.

Table 4. Differentially Expressed Genes using Limma

30 Minutes	4 Hours	12 Hours	24 Hours	168 Hours	Final DEG's
SNAR-A1	CXCL2	FKBP11	CYP26B1	CYP26B1	Hs.7413
HMOX1	CCL20	MYH10	SYNPO2L	FHOD3	CYP26B1
LFNG	DHRS3	CRLF1	CXCL1	NPPB	NPPB
BCYRN1	IL8	SLC12A8	NPPB	ACTC1	SFRP4
LOC255783	SMAD3	Hs.559673	LOC340598	SFRP4	Hs.562219

2.3 Gene Association Network Reconstruction and Prediction of Gene Function

In this work, we have performed a gene association network (GAN) reconstruction from time-course data of the prior selected DEGs using a regularized dynamic partial correlation method. GeneNet⁴⁶ R-package has been used to analyze covariance matrices with selected shrinkage method as dynamic. Analyses have been performed with a posterior probability of 0.8, and

0.9. GAN reconstruction has been implemented using splineNetRecon function from the splineTimeR package to identify regulatory association between genes under a specific condition (Treatment)²⁸. Top significant edges have been identified based on their cutoff posterior probability.

Important genes have been identified by analyzing the overall graph topology in the resultant GAN using the commonly used centrality measures: degree, betweenness and closeness centrality. These centrality measures were combined for selecting the top 150 ranked genes by taking the average of the quantile normalized values of these measures to perform gene function prediction.

GeneMANIA³¹ provides us with a very adjustable interface to obtain possible connections between the query genes by searching several prominent networks containing gene function, proteomic data. GeneMANIA operates on a ridge regression derived fast heuristic algorithm to integrate multiple functional gene association networks using label propagation. GeneMANIA weights data sources based on the connectivity strength of all genes with each other in the query list and suggests relatively similar non-queried genes and their connection types. It returns an interactive functional gene association network of all resultant genes and their relationship. It also returns functions of interest based on non-negative weights of gene sources (i.e. association of two genes) by considering the problem of prediction of gene function as a binary classification problem. GeneMANIA further returns a ranked gene list presumably sharing phenotypes with query list genes based on its large and diverse data. All 150 genes from the ranked list of GAN were analyzed for gene function using the GeneMANIA Cytoscape app.

2.4 Finding Gene Modules from Time-Series data

Piece-wise Aggregate Approximation (PAA) has been used from the *TSrepr*⁴⁷ package in R for dimensionality reduction of multivariate time-series data. The DEGs obtained from the dataset contained 3 replicate values at each time point, which have been converted into univariate expression values for each of the 5 time points using the mean function of PAA. These expression values were clustered to form gene modules.

We have used a model-based approach, the Dirichlet Process Gaussian Process mixture model (DPGP), presented by McDowell *et al.* to find gene modules from univariate time-course expression data¹⁴. It simultaneously models cluster number with a Dirichlet process and also uses temporal dependencies as a Gaussian Process. DPGP uses a Bayesian non-parametric model for time course paths $P \in \mathbb{R}^{N \times T}$, where N is the number of genes and T is the number of time points. A generative Dirichlet process (DP) mixture model is defined as

$$V_i \sim DP(\delta, G) \quad (4)$$

Here, DP represents a draw which generates latent variables V_i for a cluster i with δ being the concentration parameter and G being the base distribution, and the observation distribution $M_j \sim p(\cdot | V_i)$ is specified with a Gaussian Process (GP) for gene $j \in \{1, 2, \dots, N\}$. Cluster-specific parameter values V_i is obtained from conditional probability $p(V_i | V_{-i})$ on all other variables except V_i by integrating draws from DP . The within-cluster parameters of the trajectories for a cluster i , i.e. $V_i = \{\mu_i, \ell_i, \zeta_i, \sigma_i^2\}$ is described as $\mu_i \sim GP(\mu_0, L)$ with $\ell_i, \zeta_i \sim \ln \mathcal{N}(0, 1)$ and σ_i^2 being an inverse gamma function of shape and rate of the trajectory. DPGP uses a cluster-specific positive definite Gram matrix L quantifying similarity between each and every time point $T_x, T_{x'}$ for a cluster i which is defined as

$$L_i(T_x, T_{x'}) = \zeta_i^2 \exp \left\{ -\frac{\|T_x - T_{x'}\|_2}{2\ell_i^2} \right\} \quad (5)$$

The hyperparameter ℓ_i, ζ_i corresponds to input space distance among data points having correlated output (*characteristic length scale*), gene expression path variance over time (*signal variance*), respectively. Thus, the probability distribution of each observation M_j corresponding GP for a specific cluster including unique marginal variance is characterized by

$$M_j \sim \mathcal{N}(\mu_i, L_i + \sigma_i^2 I) \quad (6)$$

DPGP uses Markov Chain Monte Carlo Method (MCMC)⁴⁸ to assess the posterior distribution of the parameters used in the model. In MCMC, "Algorithm 8" of Neal's Gibbs Sampling⁴⁹ has been implemented to estimate the posterior distribution of the probability that a cluster i containing trajectory of a gene j corresponding to DP. We ran the MCMC for 1000 iterations which is also the default number of iterations, among which the first 48% are divided into two burn-in phases equally. Initially, every gene is assigned to the clusters of their own. In burn-in phases, a cluster is chosen at each iteration for a gene based on its likelihood to fit a GP having cluster mean and covariance as initial parameters. After the second burn-in phase posterior probability of the kernel hyperparameters was computed. MCMC generates a state sequence from Gibbs' sampler in which each state contains a section of genes into each cluster. Maximum a *posteriori* clustering or partition with the maximum value of posterior probability has been chosen to summarize results. We have used a Python implementation of DPGP from the command line interface and resultant clusters of genes have been used for further analysis.

2.5 Key module Identification and Biological Significance Analysis

This subsection provides information about approaches that have been used to find biological significance of our obtained results.

2.5.1 Finding Transcriptional Regulators

We have analyzed DEGs along with their expression values between two sample points to identify key transcriptional regulators (TR) using REGGAE³⁰. It operates by using Kolmogorov-Smirnov-like test statistics along with an implicit combination of Regulator Target Interactions (RTI's) for the prioritizing influence of TR's. It returns a ranked list of key TRs with p-value aggregations.

2.5.2 DNA-binding Transcription Factor Identification

Key TRs have been observed against a cumulative and high-quality knowledge source of genome-scale information referring their biological function potentiality as DNA Binding Transcription Factors (DbTF) using TFcheckpoint³².

2.5.3 Identifying Key Modules

Gene modules have been studied and ranked for having the highest number of key TRs. Gene modules having the highest number of key TRs were identified as the key modules. Further analysis of the top 3 modules was carried out.

2.5.4 Gene Set Enrichment Analysis

Top 3 key modules have been analyzed to discover biological processes and pathway analysis for the involved genes inside each module. Biological aspects like gene ontology, disease-gene associations and pathway analysis were determined using Enrichr³³.

3 Conclusion

In this article, we have designed a framework for analyzing top regulated genes from multivariate time-series Pancreatic Ductal Adenocarcinoma (PDAC) microarray data to identify key gene modules, DNA binding Transcription Factors (DbTF) among key Transcriptomic Regulators (TR), and gene function prediction of potential crucial genes in construction of Gene Association Networks (GAN). Top differentially expressed genes (DEG) were identified with a cubic spline regression model. Gene modules from normalized DEG expressions have been determined using a Dirichlet Process Gaussian Process (DPGP) mixture model. DEG list have been used for the detection of key TRs using REGulator-Gene Association Enrichment (REGGAE) and DbTFs among the TRs using TFcheckpoint databases. Thereafter, we used key TRs' information in each of the gene modules to find top clusters by ranking with the number of key TRs inside them. We have labeled the top modules as "key modules". We have also reconstructed GAN of DEGs with the help of a regularized dynamic partial correlation approach to find top genes in the GAN by ranking them according to their centrality measures in the network. We, then, predicted gene functions of the top genes using the GeneMANIA web server and analyzed obtained genes for their association with PDAC. We identified 8 TR's and 10 genes in the reconstructed GAN to be directly associated with PDAC. We have detected 24 directly associated genes with PDAC in the top 4 key modules. Our analysis also reveals that only 24 out of 32 key gene modules contained all the key TRs. KEGG pathway, biological processes GO terms and DisGeNet analysis of top key modules were also determined.

Additionally, our analysis can be further extended by combining several biologically significant data sources for this disease by data integration techniques. Common symptoms of PDAC include weight loss, indigestion, abdominal and back pain. Thus, studying network pathways of the key gene modules may help unravel the serious complexities of this disease.

Moreover, the genes in the key modules may be further validated using *in vitro* experiments to reveal some important findings in pancreatic ductal adenocarcinoma and its pathogenesis. One may also verify the role of key regulators in the modules being potential biomarkers. Survival analysis using key transcriptional regulators and Protein-Protein interaction networks (PPIN) may enlighten us with greater insights about this disease.

References

1. Douglas, Y. The next decade of gene expression profiling. *Drug Discov.* 55 (2005).
2. Robertson, S. Gene expression techniques (2019).
3. Hossain, S. M. M., Ray, S. & Mukhopadhyay, A. Preservation affinity in consensus modules among stages of HIV-1 progression. *BMC Bioinforma.* 18, 181 (2017).
4. Hossain, S. M. M., Ray, S., Tannee, T. S. & Mukhopadhyay, A. Analyzing prognosis characteristics of Hepatitis C using a biclustering based approach. *Procedia Comput. Sci.* 115, 282 – 289 (2017).
5. Ray, S., Hossain, S. M. M., Khatun, L. & Mukhopadhyay, A. A comprehensive analysis on preservation patterns of gene co-expression networks during Alzheimer's disease progression. *BMC Bioinforma.* 18, 579 (2017).

6. Ray, S., Hossain, S. M. M. & Khatun, L. Discovering preservation pattern from co-expression modules in progression of HIV-1 disease: An eigengene based approach. In *2016 IEEE International Conference on Advances in Computing, Communications and Informatics, ICACCI 2016, Jaipur, India, September 21-24, 2016*, 814–820 (IEEE, USA, 2016).
7. Hossain, S. M. M., Ray, S. & Mukhopadhyay, A. Identification of hub genes and key modules in stomach adenocarcinoma using nsnmf-based data integration technique. In *IEEE 2019 International Conference on Information Technology (ICIT)*, 331–336 (2019).
8. Bar-Joseph, Z., Gitter, A. & Simon, I. Studying and modelling dynamic biological processes using time-series gene expression data. *Nat. Rev. Genet.* **13**, 552–564 (2012).
9. Ernst, J., Nau, G. J. & Bar-Joseph, Z. Clustering short time series gene expression data. *Bioinformatics* **21**, i159–i168 (2005).
10. Leek, J. T., Monsen, E., Dabney, A. R. & Storey, J. D. Edge: extraction and analysis of differential gene expression. *Bioinformatics* **22**, 507–508 (2006).
11. Yang, K., Cai, Z., Li, J. & Lin, G. A stable gene selection in microarray data analysis. *BMC bioinformatics* **7**, 228 (2006).
12. Anjum, A. *et al.* Identification of differentially expressed genes in rna-seq data of arabidopsis thaliana: A compound distribution approach. *J. Comput. Biol.* **23**, 239–247 (2016).
13. Spies, D., Renz, P. F., Beyer, T. A. & Ciaudo, C. Comparative analysis of differential gene expression tools for rna sequencing time course data. *Briefings bioinformatics* **20**, 288–298 (2019).
14. McDowell, I. C. *et al.* Clustering gene expression time series data using an infinite gaussian process mixture model. *PLoS computational biology* **14**, e1005896 (2018).
15. Fujita, A. *et al.* Functional clustering of time series gene expression data by granger causality. *BMC systems biology* **6**, 137 (2012).
16. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci.* **95**, 14863–14868 (1998).
17. Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J. & Church, G. M. Systematic determination of genetic network architecture. *Nat. genetics* **22**, 281–285 (1999).
18. Tamayo, P. *et al.* Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci.* **96**, 2907–2912 (1999).
19. Ernst, J. & Bar-Joseph, Z. Stem: a tool for the analysis of short time series gene expression data. *BMC bioinformatics* **7**, 191 (2006).
20. Ramoni, M. F., Sebastiani, P. & Kohane, I. S. Cluster analysis of gene expression dynamics. *Proc. Natl. Acad. Sci.* **99**, 9121–9126 (2002).
21. Magni, P., Ferrazzi, F., Sacchi, L. & Bellazzi, R. Timeclust: a clustering tool for gene expression time series. *Bioinformatics* **24**, 430–432 (2008).
22. Golumbeanu, M. & Beerenwinkel, N. Clustering time series gene expression data with tmixclust (2018).
23. Luong, D. T. A. & Chandola, V. A k-means approach to clustering disease progressions. In *2017 IEEE International conference on healthcare informatics (ICHI)*, 268–274 (IEEE, 2017).
24. Yang, Y., Huang, N., Hao, L. & Kong, W. A clustering-based approach for efficient identification of microrna combinatorial biomarkers. *BMC genomics* **18**, 210 (2017).
25. Li, X. *et al.* Identification of hub genes and key pathways associated with angioimmunoblastic t-cell lymphoma using weighted gene co-expression network analysis. *Cancer management research* **11**, 5209 (2019).
26. Cho, R. J. *et al.* A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. cell* **2**, 65–73 (1998).
27. Kourou, K., Rigas, G., Papaloukas, C., Mitsis, M. & Fotiadis, D. I. Cancer classification from time series microarray data through regulatory dynamic bayesian networks. *Comput. Biol. Medicine* **116**, 103577 (2020).
28. Michna, A. *et al.* Natural cubic spline regression modeling followed by dynamic network reconstruction for the identification of radiation-sensitivity gene association networks from time-course transcriptome data. *PLoS one* **11** (2016).
29. Ritchie, M. E. *et al.* limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research* **43**, e47–e47 (2015).

30. Kehl, T. *et al.* Reggae: a novel approach for the identification of key transcriptional regulators. *Bioinformatics* **34**, 3503–3510 (2018).
31. Warde-Farley, D. *et al.* The genemania prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic acids research* **38**, W214–W220 (2010).
32. Chawla, K., Tripathi, S., Thommesen, L., Lægreid, A. & Kuiper, M. Tfcheckpoint: a curated compendium of specific dna-binding rna polymerase ii transcription factors. *Bioinformatics* **29**, 2519–2520 (2013).
33. Chen, E. Y. *et al.* Enrichr: interactive and collaborative html5 gene list enrichment analysis tool. *BMC bioinformatics* **14**, 128 (2013).
34. Thissen, D., Steinberg, L. & Kuang, D. Quick and easy implementation of the benjamini-hochberg procedure for controlling the false positive rate in multiple comparisons. *J. educational behavioral statistics* **27**, 77–83 (2002).
35. Piñero, J. *et al.* Disgenet: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic acids research* gkw943 (2016).
36. Nagaraj, N. S. & Datta, P. K. Targeting the transforming growth factor- β signaling pathway in human cancer. *Expert. opinion on investigational drugs* **19**, 77–91 (2010).
37. Gagliano, N. *et al.* Epithelial-to-mesenchymal transition in pancreatic ductal adenocarcinoma: Characterization in a 3d-cell culture model. *World J. Gastroenterol.* **22**, 4466 (2016).
38. Gordon, K. J., Kirkbride, K. C., How, T. & Blobe, G. C. Bone morphogenetic proteins induce pancreatic cancer cell invasiveness through a smad1-dependent mechanism that involves matrix metalloproteinase-2. *Carcinogenesis* **30**, 238–248 (2009).
39. Recio-Boiles, A. *et al.* Jnk pathway inhibition selectively primes pancreatic cancer stem cells to trail-induced apoptosis without affecting the physiology of normal tissue resident stem cells. *Oncotarget* **7**, 9890 (2016).
40. Tjomsland, V. *et al.* Il-1 α expression in pancreatic ductal adenocarcinoma affects the tumor cell migration and is regulated by the p38mapk signaling pathway. *PLoS One* **8**, e70874 (2013).
41. Loncle, C. *et al.* Il17 functions through the novel reg3 β -jak2-stat3 inflammatory pathway to promote the transition from chronic pancreatitis to pancreatic cancer. *Cancer research* **75**, 4852–4862 (2015).
42. Chopra, M. *et al.* Tumor necrosis factor induces tumor promoting and anti-tumoral effects on pancreatic cancer via tnfr1. *PLoS One* **8**, e75737 (2013).
43. Gao, J., Long, B. & Wang, Z. Role of notch signaling pathway in pancreatic cancer. *Am. journal cancer research* **7**, 173 (2017).
44. Barrett, T. *et al.* Ncbi geo: archive for functional genomics data sets—update. *Nucleic acids research* **41**, D991–D995 (2012).
45. Carlson, M. org.hs.eg.db: Genome wide annotation for human. r package version 3.8.2, DOI: [10.18129/B9.BIOC.ORG.HS.EG.DB](https://doi.org/10.18129/B9.BIOC.ORG.HS.EG.DB) (2019).
46. Opgen-Rhein, R. & Strimmer, K. From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. *BMC systems biology* **1**, 37 (2007).
47. Laurinec, P. Tsrepr r package: Time series representations. *J. Open Source Softw.* DOI: [10.21105/joss.00577](https://doi.org/10.21105/joss.00577) (2018).
48. Medvedovic, M. & Sivaganesan, S. Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics* **18**, 1194–1206 (2002).
49. Neal, R. M. Markov chain sampling methods for dirichlet process mixture models. *J. computational graphical statistics* **9**, 249–265 (2000).

Acknowledgements (not compulsory)

Acknowledgements should be brief, and should not include thanks to anonymous referees and editors, or effusive comments. Grant or contribution numbers may be acknowledged.

Author contributions statement

Must include all authors, identified by initials, for example: A.A. conceived the experiment(s), A.A. and B.A. conducted the experiment(s), C.A. and D.A. analysed the results. All authors reviewed the manuscript.

Additional information

To include, in this order: **Accession codes** (where applicable); **Competing interests** (mandatory statement).

The corresponding author is responsible for submitting a [competing interests statement](#) on behalf of all authors of the paper. This statement must be included in the submitted article file.