Heliyon 7 (2021) e06000

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Unsupervised subtyping and methylation landscape of pancreatic ductal adenocarcinoma

Shikha Roy [1], Amar Pratap Singh [1], Dinesh Gupta [*]

Translational Bioinformatics Group, International Centre for Genetic Engineering and Biotechnology, New Delhi, India

ARTICLE INFO

Keywords:
Bioinformatics
Multi-omics
Unsupervised classification
Subtyping
Integratomics
DNA Methylation
RNA-Seq
Pancreatic ductal adenocarcinoma

ABSTRACT

Pancreatic Ductal Adenocarcinoma (PDAC) is an aggressive form of pancreatic cancer that typically manifests itself at an advanced stage and does not respond to most treatment modalities. The survival rate of a PDAC patient is less than 5%, with a median survival of just a couple of months. A better understanding of the molecular pathology of PDAC is needed to guide research for the development of better clinical treatment modalities for PDAC patients. Gene expression studies performed to date have identified different subtypes of PDAC with prognostic and clinical relevance. Subtypes identified to date are highly heterogeneous since pancreatic cancer is heterogeneous cancer. Tumor microenvironment and stroma constitute a major chunk of PDAC and contribute to the heterogeneity. Better subtyping methods are need of the hour for better prognosis and classification of PDAC for future personalized treatment. In this work, we have performed an integrated analysis of DNA methylation and gene expression datasets to provide better mechanistic and molecular insights into Pancreatic cancers, especially PDAC. The use of varied and diverse datasets has provided valuable insights into different cancer types and can play an integral role in revealing the complex nature of underlying biological mechanisms. We performed subtyping of TCGA-PAAD gene expression and methylation datasets into different subtypes using state-of-the-art normalization methods and unsupervised clustering methods that reveal latent hidden factors, leading to additional insights for subtyping. Differential expression and differential methylation were performed for each of the subtypes obtained from clustering. Our analysis gave a consensus of five cluster solution with relevant pathways like MAPK, MET. The five subtypes corresponded to the tumor and stromal subtypes. This analysis helps in distinguishing and identifying different subtypes based on enriched putative genes. These results help propose novel experimentally-verifiable PDAC subtyping and demonstrate that using varied data sets and integrated methods can contribute to disease prognostication and precision medicine in PDAC treatment.

1. Introduction

Studies have shown that epigenetic processes are often changed during different stages of cancer, including the initial stage and progression of tumor stages. The changes also include a global change in the DNA methylation profiles concerning normal DNA methylation patterns [1]. Broadly, this change in DNA methylation is characterized by overall genome-wide hypomethylation and DNA hyper-methylation of CpG island promoters [2, 3]. Many studies have been conducted using the TCGA dataset on DNA methylation in different cancers, which has provided new insights into these cancers [4, 5]. PDAC accounts for most exocrine pancreatic cancer cases, with variants being less common and, apart from differences in prognosis, being uninformative for management decisions [6]. Adenosquamous carcinoma is an uncommon variant of PDAC and shares the features of adenocarcinoma and squamous cell carcinoma, showing a mixture of glandular and squamous differentiation [7]. Other carcinomas of the exocrine pancreas with acinar differentiation include pancreatoblastomas, acinar cell carcinomas, and carcinomas with mixed histology and are usually identified by staining for trypsin [8]. Five-year survival of PC is less than 5%, with survival just a couple of months [8].

Pancreatic ductal adenocarcinoma (PDAC) is an aggressive disease that represents itself at an advanced stage and does not respond to most of the available treatment options [9, 10]. Studies have shown that PDAC is predicted to become the second leading cause of cancer mortality by 2030 [11]. Various studies have shed light and have helped decipher and

[*] Corresponding author.
E-mail address: dinesh@icgeb.res.in (D. Gupta).
[1] Equal contribution.

https://doi.org/10.1016/j.heliyon.2021.e06000
Received 7 September 2020; Received in revised form 14 November 2020; Accepted 12 January 2021

characterize the PDAC genetic alterations, provided important insights into the biology of the disease, and laid the foundation for the development of approaches for detection and improved therapies. Initial whole-exome sequencing studies of pancreatic cancer identified several factors, including mutations and changes in somatic copy number alterations (SCNAs) that altered the function of many key oncogenes and tumor suppressor genes, including *KRAS, TP53, SMAD4, and CDKN2.* DNA sequencing of neoplastic cells has demonstrated that most PDACs show complex chromosomal rearrangement patterns, some of which are consistent with PDAC progression [12]. Numerous gene expression studies have identified subtypes of PDAC with prognostic and biological relevance [13, 14, 15, 16]. Genomic analyses have previously revealed heterogeneous landscapes of mutation, copy number variation, structural variation, and gene expression in pancreatic cancer [19]. Better understanding and delineating the molecular pathology of pancreatic ductal adenocarcinoma cancer is an urgent need to achieve advances in clinical treatment for patients. Intra-tumoral heterogeneity makes PDAC a complex disease [17]. Tumor microenvironment, stroma, and immune cell filtrate contribute to the heterogeneity in PDAC. Some have considered Intra-tumor heterogeneity as the Rosetta stone of tumor therapy resistance [18]. Thus, there is a need to identify homogeneous groups from different high throughput datasets, which could be an important step towards better-personalized clinical management of PDAC patients. A plethora of large-scale and valuable experimental data has been generated from extensive experimentations. For example, DNA methylation has been explored in many cancers. It has provided novel and valuable insights into many cancers. Epigenetic mechanisms regulate ontological gene expression networks at different levels, including time and place, giving rise to both normal and disease phenotypes. Recently, multiparametric integrative chromatin immunoprecipitation-sequencing (ChIP-seq) studies on multiple histone modifications, RNA-sequencing (RNA-Seq) and DNA methylation studies have been performed to define the epigenetic landscape of various PDAC subtypes [19]. Integrative analysis of TCGA pancreatic ductal adenocarcinoma, involving different varied and diverse datasets, have revealed a complex molecular landscape of PDAC [20, 21, 22]. Survival analyses based on diverse datasets, including RNA-Seq, DNA methylation, miRNA, long non-coding RNA, representing PDAC patients and normal subjects, have led to the identification of putative gene markers associated with survival and prognosis [23]. In this study, we have performed unsupervised clustering and integrative analysis of PDAC DNA methylation and gene expression data, obtained from The Cancer Genome Atlas (TCGA).

Molecular aberrations and alterations identified in cancers often have multiple synergic interactions as it is a complex disease. Thus, it is important to collect and analyze multiple data types to improve patients' prognosis and response to treatment. A single omics screen cannot fully reveal and decipher the complexity of a biological entity. Therefore, studies involving varied datasets help in a better understanding of the system. Integrating the diverse and rich information from diverse datasets has been an approach to identify latent, hidden factors and putative biomarker identification for cancer and non-communicable studies. Integrative approaches in cancer usually focus on integrating multiple types of omics data such as RNA-Seq, DNA methylation, ChIPseq, etc., rather than using a single omics profile. Varied and diverse data has provided valuable insights into different cancer types and can play an integral role in revealing the underlying complex nature of biological mechanisms in breast cancer, colon cancer, and pancreatic cancer [24, 25, 26]. The earliest example of data integration in omics reported in the literature were studies that involved data analysis from individual omics separately, one by one and the results of these parallel studies were then finally merged [27]. The background behind the integratomics or integrative analysis involving different layers of information is based on emergent property in systems theory. The concept of emergent property has become very popular in the systems biology approach. The emergent properties indicate how some system features are observed only when the system is studied as a whole and not as the sum of its parts [28, 29]. The integrative analysis approach will help in providing better mechanistic and molecular insights into pancreatic ductal adenocarcinoma subtypes. TCGA-PAAD dataset gene expression studies to date include the whole data, which involves PDAC and non-PDAC samples. Our analysis takes this heterogeneity into account and includes only matched PDAC gene expression and DNA methylation samples for better data integration. We have explored the subtyping of PDAC by integrating two different levels of data, using state of the art normalization methods and unsupervised clustering method intNMF, which gives latent factors, thus leading to better subtyping. The current study tries to provide better insight into the understanding of DNA methylation underlying PDAC heterogeneity and identification of epigenetically modified regions in different PDAC subtypes leading to better subtyping, which can serve as potential new markers and therapeutic targets. Our integrated data analysis indicates that gene expression, DNA methylation, and tumor-intrinsic factors, such as the tumor microenvironment and immune cell filtrate, all contribute to the heterogeneous landscape of PDAC and thus, the integrative analysis of these factors lead to better subtyping and a better understanding of the distinct PDAC landscape.

2. Methodology

2.1. Data mining of RNA-Seq and methylation datasets

The study workflow is shown in Figure 1. TCGAbiolinks package was used to obtain pancreatic cancer datasets from TCGA. This package imports and processes molecular profiles from high-throughput experiments such as next-generation sequencing and methylation array and their clinical data for statistical analysis [30]. To date, most of the studies performed on pancreatic cancers have focused on TCGA_PAAD datasets, which is heterogeneous data containing PDAC along with non-PDAC samples. Non-PDAC cancer includes adenosquamous carcinoma, colloid carcinoma, squamous cell carcinoma, and neuroendocrine tumor. Missing and non-matched samples were removed before subtyping and clustering.

2.2. Data pre-processing of methylation datasets

ChAMP or chip analysis methylation pipeline is used for the analysis of methylation datasets such as filtering low-quality probes, adjustment for Infinium I and Infinium II probe design, batch effect correction, detects differentially methylated probes (DMPs), differentially methylated regions (DMRs) and detection of copy number aberrations (CAN) [31]. It also can filter SNPS based on user-specific minor allele frequency in one of four populations as defined by the 1000 genomes project [32]. It uses the algorithm "probe lasso" method for DMR hunting that incorporates annotated genomic features and their corresponding local probe densities and methylation [33].

2.3. Data normalization of methylation datasets

There are various methods for the normalization of DNA methylation data such as Noob, Subset-quantile within array normalization (SWAN), Beta-Mixture Quantile (BMIQ), and Functional normalization (FN). SWAN performs scaling of Infinium I and Infinium II probes together within a single array to minimize the differences in beta value distribution [34, 35]. BMIQ normalization uses state-membership probabilities under the beta mixture model to reassign quantile to type2 probe based on type1 probe distribution (Supplementary file I, S1) [36]. Noob capitalizes on the Infinium I probe's unique design to perform with-array normalization of methylation datasets (Supplementary file I, S2) [37]. FN is an unsupervised method that improves replication between

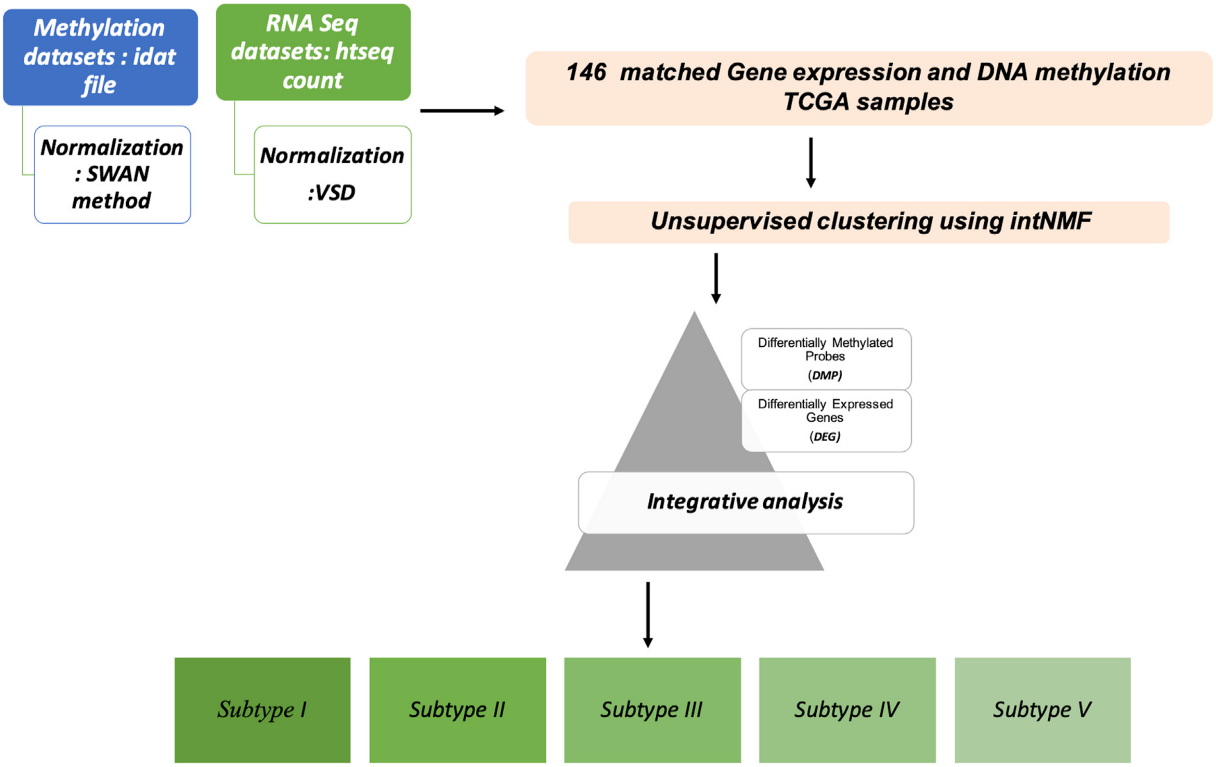
Methylation datasets : idat file
RNA Seq datasets: htseq count
Normalization : SWAN method
Normalization :VSD
146 matched Gene expression and DNA methylation TCGA samples
Unsupervised clustering using intNMF
Differentially Methylated Probes (DMP)
Differentially Expressed Genes (DEG)
Integrative analysis
Subtype I
Subtype II
Subtype III
Subtype IV
Subtype V

Figure 1. Workflow followed in the study to identify major subtypes of pancreatic cancer by integrative clustering analysis of methylation and genomics datasets.

experiments even in batch effect and uses a control probe to act as surrogates variable for unwanted variation [38].

2.4. Data normalization of RNA-Seq datasets

Gene expression datasets of 146 samples were normalized using variance stabilizing normalization (VSN) of the DESeq package. It is used to analyze count datasets generated from high throughput experiments to perform downstream processing such as differential expression analysis [39]. To capture differential signals with high statistical power, its uses a negative binomial generalized linear model where mean and variance are linked to local regression [39]. Data normalization of our datasets was performed using VSN, which uses parametric fit for dispersion using vst function. The vst function calculates variances from fitted dispersion-mean relation and transforms count data into homoscedastic data (https://rdrr.io/bioc/DESeq2/man/varianceStabilizingTransformation.html). Variance stabilized transformed datasets incorporates a correction for size factors or normalization factors.

2.5. Identification of subtypes using clustering

Before intNMF, genes with low variability across samples were removed from gene expression datasets and methylation datasets using Mean absolute deviation (MAD). The basic assumption or premise behind it is that genes with high variability contribute more to the clustering process. intNMF package was used for clustering using highly variable genes and methylation probe. This package utilizes a non-negative matrix factorization method to perform unsupervised integrated clustering of high dimensional datasets [40]. The k represents the number of clusters varied across a suitable range and was repeated 20 times to predict the optimum number of clusters based on cluster predictive index (CPI). Parameters used while running intNMF were n.runs = 30, n.fold cross validation = 5, k.range default value = 2 to 8, st.count = 10, maxiter which is maximum number of iteration = 20 and wt = 1 for each data. It generates CPI that is the measure of the stability of clusters obtained. It signifies the correlation between sample distances obtained for the consensus matrix representing the clusters [41]. It also generates a silhouette index value for interpretation and validation of consistency within the data clusters. The technique provides a graphical representation of how well each object has been classified based on the cluster's tightness and separation [42]. The silhouette value can range from −1 to +1, where a high value representing the object is well matched to its cluster and poorly matched neighboring clusters.

2.6. Differential methylation analysis of obtained subtypes

ChAMP (The Chip Analysis Methylation Pipeline) package was used to perform differential methylation analysis on five subtypes obtained from intNMF. It is used for the various downstream process in methylation datasets such as normalization, detecting differentially methylated regions and copy number aberrations [31]. Differential methylation analysis is performed by limma that uses a linear model for both categorical variables like two phenotypes, like "tumor", "metastasis" or "control," as well as a numeric variable such as age to calculate the p-value for differentially methylated probes. It carries regression analysis to find out covariate-related CpGs for the specified condition. Its output includes some data frames of p-value, t-statistic, difference in mean methylation between two groups (for categorical covariate only), average beta value for sample group, and delta beta value for two comparison groups and annotation for each probe. It also includes the annotation for each probe, the average beta value for the sample group, and the delta beta value for the two groups used in the comparison (https://www.bioconductor.org/packages/release/bioc/vignettes/ChAMP/inst/doc/ChAMP.html). The absolute minimum beta value is 0.2, and for the Benjamini-Hochberg adjustment method, p-value <0.01 is used as a cut-off for DMR analysis.

2.7. Differential gene expression analysis of obtained subtypes

Differential gene expression analysis was performed for 5 subtypes obtained from intNMF, using the edgeR [43]. It uses a negative binomial distribution model for gene count and performs differential expression analysis on RNA-Seq expression profiles [44]. It implements a range of statistical methodologies such as empirical Bayes estimation, which generates gene-specific dispersion estimates, ranking genes that behave consistently across the replicates higher than others [45].

2.8. Correlation analysis of DMR and DEG for obtained subtypes

The in-depth integrative analysis relies on analyzing multiple datasets such as gene expression, methylation, CNV (copy number variations), etc., to extract and identify major biological insights for disease progression. Correlation between methylation and gene expression was carried out to estimate the extent to which methylation influences gene expression in pancreatic ductal adenocarcinoma cancer [46]. The correlation analysis was carried out using TCGAbiolinks starburst function between differentially methylated CpG sites and differentially expressed genes for each subtype individually. This analysis gives clues regarding the epigenetically regulated genes responsible for heterogeneity in pancreatic ductal adenocarcinoma cancer. Starburst plots generate an exponential curve that captures the non-linear relationship between methylation and gene expression utilizing a gene probe that occurs within 20kb windows from each other [47]. Parameters used in Starburst plots are expression p-value cut-off of 0.05 and methylation p-value cut-off of 0.05. DNA methylation platform used is 450K, and genome of reference used to identify nearest probes is hg38.

2.9. Go analysis of obtained subtypes using clusterprofiler to identify gene signatures

Clusterprofiler was used to perform gene ontology analysis on the gene signatures for the predicted subtypes. This package was used for pathway level analysis to obtain a system-level understanding for gene signatures obtained from analysis datasets generated from various platforms such as RNA-Seq, micro-array, etc. [48]. GO was performed on gene signatures returned by the correlation analysis of DMR and DEG to determine major pathways and processes regulated in each of the subtypes using the clusterprofiler.

3. Results

3.1. Data mining of TCGA_PAAD

The standard TCGA dataset for pancreatic cancer TCGA-PAAD was downloaded from TCGAbiolink, including 183 cancer and four normal samples. The curation of TCGA_PAAD samples is important for removing biological and clinical biases from non-PDAC samples [49]. PDAC gene expression RNA-Seq datasets and DNA methylation datasets consisted of 153 and 146 samples, respectively. After looking for seven missing samples for DNA methylation data, 146 matched PDAC samples were selected [49]. These 146 PDAC samples consisted of expression profile for DNA methylation as well as gene expression. Gene expression profile and DNA methylation datasets were normalized by various methods before performing clustering and subsequent integration analysis. There are several methods for preprocessing the 450K array data, adjust for probe-type or color bias, subtract background signals, and eliminate systematic errors [50]. DNA methylation datasets were normalized using various methods such as SWAN, BMIQ, Noob, and Functional normalization (FN). SWAN method gave the best result for Infinium I and Infinium II probe normalizations. In SWAN normalization, type I probe density for type I probe is the same as type II probe density, but in funtonorm type I probe density is 4 but type II probe density is 2.55 (Figure 2a,b). SWAN-normalized datasets were taken further for integrative analysis.

3.2. Filtering of CpG probes using ChAMP

ChAMP performed initial preprocessing on methylation datasets using idat files. If for a probe, a p-value of detection was above 0.01, it was regarded as a failed probe. The selected probes were filtered with a p-value detection value above 0.01, which resulted in the removal of 54800 probes. It also filtered non-CpG probes resulting in the removal of 1616 probes. It filtered the probes associated with SNP resulting in the removal of 48827 probes. It filtered probes with beadcount less than 3, which lead to the removal of 144 probes in at least 5% of the study samples. On applying Multihit stats, 11 probes were removed. It also filtered probes that are mapping to X and Y chromosomes, which resulted in the removal of 8056 probes.

3.3. VSN normalization of RNA-Seq datasets

VSN was performed on RNA-Seq datasets obtained for TCGA-PAAD patients before performing clustering analysis. Before VSN the expression profile had standard deviation elevated in the lower count range (Figure 3a). After VSN the standard deviation was constant across the expression profile (Figure 3b). Normalization resulted in the transformation of data, which was homoscedastic, with constant variance. It removed the influence of technical variation such that the true biological variation can be discovered.

3.4. Identification of subtypes using clustering by intNMF

Before performing clustering, expression profile with low variability across samples was removed for gene expression and methylation datasets. MAD was calculated for each gene and probe, values with less than 0.5 were excluded from our analysis. We have performed MAD (mean absolute deviation) before intNMF as MAD gives highly variable probes. The reason for performing intNMF and no other methods like consensus clustering is that intNMF takes latency and variability into account and thus provides highly significant clusters compared to other methods. It also does not take the data distribution into account, making it highly suitable for analyzing diverse datasets. The PCA analysis has shown no significant variability can be explained by principal components for our

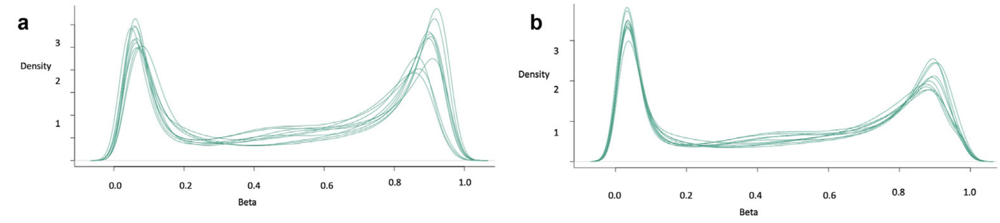

Figure 2. Various normalization methods used for Infinium I and Infinium II probe normalizations. a) SWAN b) Functional normalization. It can be seen from the images that SWAN performs better normalization than functional normalization of Infinium I and Infinium II probe peaks. Hence SWAN normalized datasets were taken further for integrative clustering analysis.

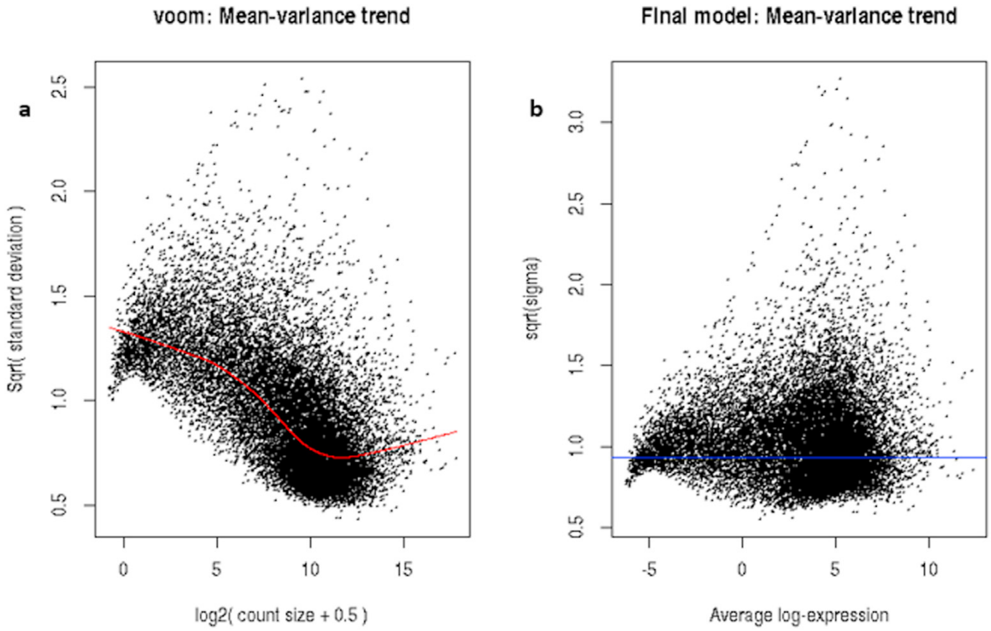

Figure 3. The figure below plots the standard deviation of the transformed data, across samples, against the mean, using the shifted logarithm transformation, the regularized log transformation, and the variance stabilizing transformation. a) The shifted logarithm has elevated standard deviation in the lower count range, and the regularized log to a lesser extent b) In the variance stabilized data the standard deviation is roughly constant along with the whole dynamic range.

multiple datasets (supplementary file I, S3). Subsequently, the top 1000 most variable probes and gene features were used to perform clustering on samples to predict subtypes using intNMF. The intNMF analysis provided optimum cluster solution by k, which was varied from 2 to 10, and the clustering process was repeated 20 times. The value of k corresponding to maximum CPI was chosen as the optimal solution. Our analysis gave an optimum five clusters solution, described in detail below. Since the top 1000 probes have a high CPI of 0.76, these were selected for further analysis (Figure 4). Thus, optimal Five clusters solution was obtained with a high CPI of 0.76 with average silhouette width of 0.74 (Figure 5).

3.5. Differential methylation analysis to identify DMPs

TCGAvisualize_meanMethylation function of TCGAbiolinks was used for visualizing differences in the mean methylation value of patients between comparison groups (Figure 6). It shows differences in the overall methylation expression between the subtypes obtained from the clustering analysis. The distribution of DMRs (hyper-methylation and hypo-methylation) was obtained using the ChAMP. The landscape of methylation include methylation in the promoter region, Transcription Start Site (TSS), intergenic, island, shell, and shore region. In the subtype I, there are 39128 DMPs out of which hypo-methylation in CpG island is higher as compared to that of hyper-methylation. In open sea, hypermethylation is higher as compared to hypomethylation. The shelf region is characterized by low methylation, while in the shore region, there is an equal distribution of hypermethylation and hypomethylation. Exon, 3′UTR, 5′UTR, Intergenic region, and TSS1500 have equal hypomethylation and hypermethylation. Gene body and TSS2000 have more hypermethylation as compared to hypomethylation (Figure 7a). In subtype II, there are 11011 DMPs, out of which there is more hypo-methylation in CpG island as compared to that in the case of hyper-methylation. In the open sea, hypermethylation is higher as compared to hypomethylation. In the shell region, there is low methylation and in the shore region, there is an equal distribution of hyper as well as hypomethylation. The exon and 5′UTR have a low methylation density, where the distribution of hyper and hypomethylation is almost equal. 3′UTR has more hypermethylation as compared to hypomethylation. The Gene body has a high density and

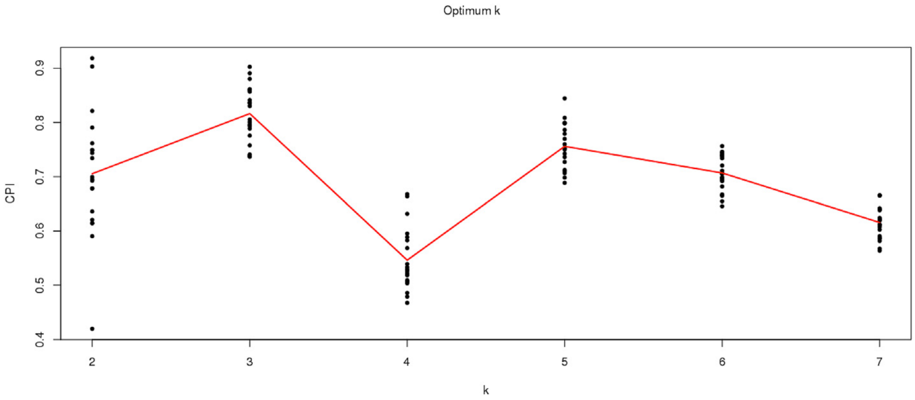

Figure 4. Cophenetic correlation coefficient plot obtained using intNMF. Cophenetic correlation coefficient/Cluster predictive index (CPI) measure of the stability of clusters obtained by intNMF. It signifies the correlation between sample distances obtained for the consensus matrix representing the clusters. We can see from the image that there is high CPI value of 0.76 at five cluster solution.

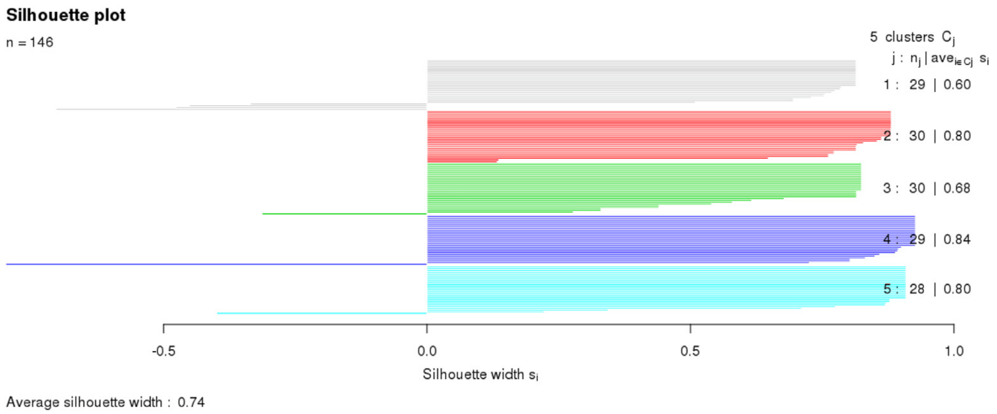

Figure 5. Silhoutte plot obtained for subtypes from clustering analysis by intNMF. It also generates silhouette refers to a method of interpretation and validation of consistency within clusters of data. It also provides a graphical representation of how well each object has been classified based on the tightness and separation of the cluster. Five clusters solution were optimal as they showed the average silhouette width of 0.74.

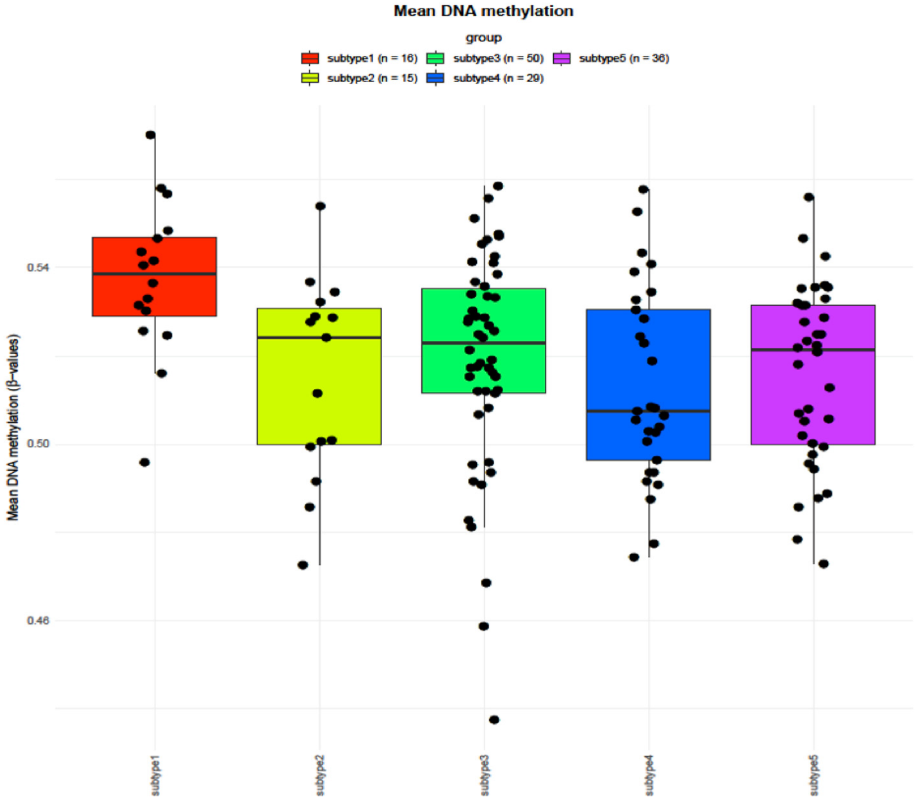

Figure 6. Visualizing differences in mean methylation pattern between comparison groups obtained by subtyping using TCGAbiolink package. We can see from the images that means methylation value between the obtained subtypes shows differences. Thus showing us the significance of obtained subtypes I terms of methylation.

equal distribution of hyper and hypomethylation. The intergenic region has a high density and hypermethylation is more as compared to hypomethylation. Both TSS1500 and TSS2000 have more hypermethylation as compared to hypomethylation (Figure 7b). In the subtype III, there are 7721 DMPs, out of which there is more hypomethylation than hypermethylation in CpG islands. In the open sea, hypermethylation is more as compared to hypomethylation. In the shelf region, hypermethylation is more as compared to hypomethylation and in the shore region, hypermethylation, as well as hypomethylation, are equally abundant. In the subtype III, exon region, 5′UTR, TSS1500 and TSS2000 have more hypomethylation as compared to hypermethylation. 3′UTR gene body and intergenic region have more hypermethylation as compared to hypomethylation (supplementary file I-Figure S4). In subtype IV, there are 6731 DMPs, out of which there is more hypermethylation than hypomethylation in the CpG island region. In the open sea region, hypomethylation is greater than hyper-methylation whereas, in the shelf and shore region, hypo-methylation is more abundant as compared to hypermethylation. In subtype IV, the exon, 3′UTR, 5′UTR, gene body, intergenic region, and TSS1500 have low hypermethylation probes as compared to high hypermethylation probes whereas only TSS2000 has more hypermethylation as compared to hypomethylation. In subtype V, there are 6728 DMPs, out of which there is more hypermethylation as compared to hypomethylation in CpG. In the open sea region, hypermethylation is more as compared to hypo-methylation. In the shell region, hypo-methylation is more as compared to hyper-methylation while in the shore region, distribution is equal (supplementary file I-Figure S5).

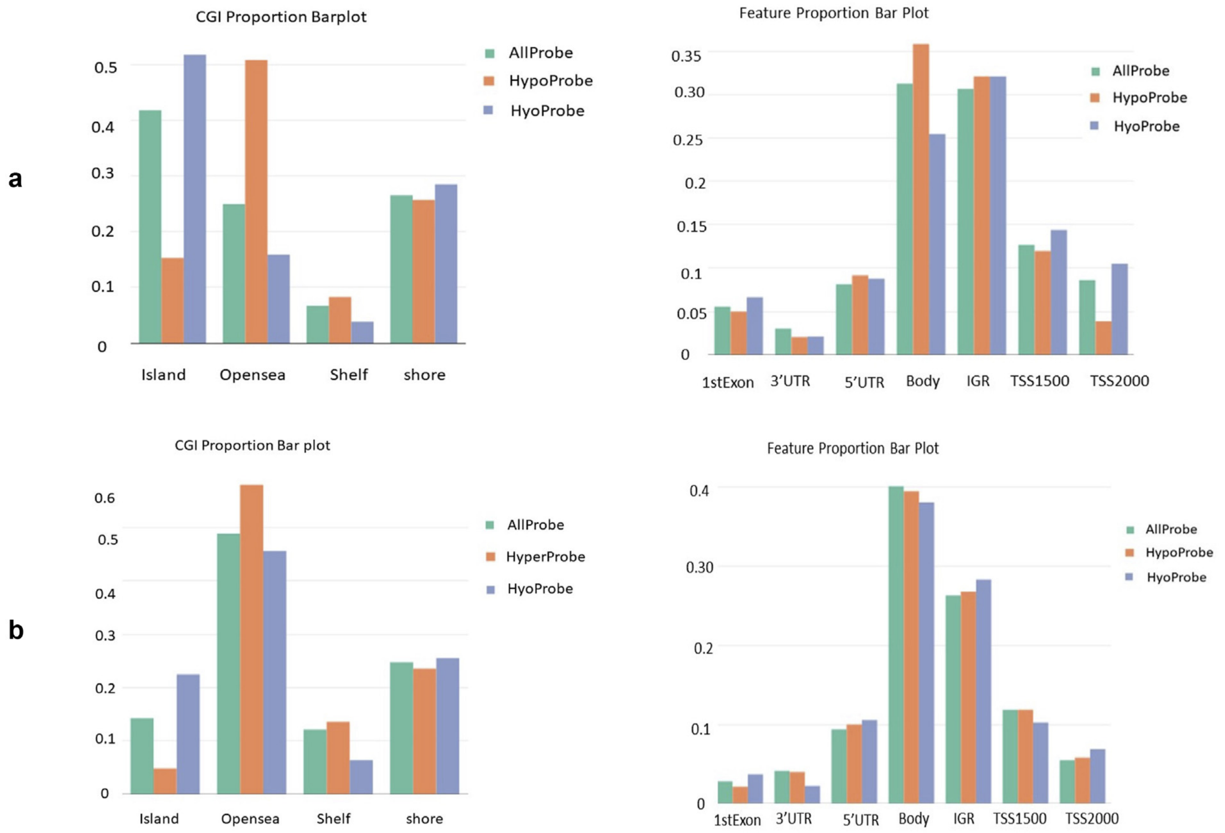

Figure 7. Differential methylation analysis between subtypes obtained by clustering using ChAMP package shows differences in methylation pattern in different regions of the genome for: a) Subtype I. b.) Subtype II.

In the subtype V, the exon, 5′UTR, TSS1500, and TSS2000 have more hypermethylation as compared to hypomethylation, whereas 3′UTR, gene body, and intergenic region have lower hypermethylation sites as compared to hypomethylation (supplementary file I-Figure S6). Overall hypo-methylation expression landscape between all subtypes shows that subtype III has most hypomethylation in island region whereas subtype II and subtype III have in open sea region (Supplementary file I, S7). Overall hyper-methylation expression landscape between all subtypes shows that subtype II has hyper methylation open sea region whereas subtype IV and subtype V in island region (Supplementary file I, S8). There is no difference in hypomethylation expression in shore region. Thus, there are difference in methylation landscape in different subtypes obtained by our analysis.

3.6. *Correlation analysis*

Correlation analysis was performed between DNA methylation and gene expression to determine the extent to which DNA methylation influenced gene expression in the subtypes obtained in our analysis. Differentially methylated CpGs and the differentially methylated genes were used for correlation analysis for each of the obtained subtypes using starburst function. Correlation is observed for each of the subtype (I, II, III, IV, V) individually, based on its DMR and DEG profile. Therefor obtained gene signatures and pathway regulated by them has been used to characterize that particular subtypes. The starburst performs a correlation analysis of DMR and DEG to find out the genes that have a significant correlation with the expression pattern of DMR. In the subtype I there are total 768 gene signature that have a significant correlation with methylation expression pattern, out of which 36 genes show hypermethylation and 732 genes show hypomethylation (supplementary file I-Figure S9, supplementary file II-subtype I). There is a gene signature with 254 genes in subtype II that has a significant correlation with methylation expression pattern, out of which 204 genes show hypermethylation and 50 genes show hypomethylation (supplementary file I-Figure S10, supplementary file II-subtype II). In subtype III, there is gene signature with 76 genes with significant correlation with methylation expression pattern, out of which 26 genes show hypermethylation and another set of 50 genes show hypomethylation (supplementary file I-Figure S11, supplementary file II-subtype III).

There is a gene signature with 390 genes in Subtype IV with a significant correlation with methylation expression pattern, out of which 364 genes show hypermethylation and 26 genes show hypomethylation (supplementary file I-Figure S12, supplementary file II-subtype IV). In subtype V there are total 148 gene signature that have significant correlation with methylation expression pattern, out of which 122 genes show hypermethylation and 26 genes show hypomethylation (supplementary file I-Figure S13, supplementary file II-subtype V). Our DMP and correlation analysis show there is direct relation between proportion of differentially methylated probe to gene signatures obtained by correlation analysis for individual subtypes (supplementary file I, S14).

3.7. *Functional annotation of PDAC subtypes using clusterprofiler*

Correlation analysis of the obtained subtypes resulted in a distinct pattern of gene expression and methylation. Our analysis results in subtypes with considerable overlap and correlation with the previously reported pancreatic ductal adenocarcinoma subtypes with their characteristic pathways. A novel gene signature was reported for each of the obtained subtypes. This analysis has tried to characterize the obtained subtypes based on the available literature regarding the gene signatures.

3.8. Subtype I – ADEX subtype genes

Pancreatic progenitor displays a transcriptional network of early pancreatic development (FOXA2/3 and PDX1) [13]. Subtype I display upregulation of genes involved in the latter stages of pancreatic development, differentiation and endocrine differentiation (NEUROG1 and NKX2-2) similar to ADEX subtypes. ADEX subtype includes genes responsible for endocrine/exocrine differentiation of pancreas [13]. The key genes identified in subtype I include HOXA3. The HOXA3 family genes are involved in pancreas development and upregulated in pancreatic cancer. The human protein atlas also shows the HOX gene family's oncogenic role with reduced survival (Figure 8a) [51].

Another signature gene in the subtype is CDH3, a classic cadherin protein, a member of a single-span transmembrane domain glycoprotein, involved in cell-cell adhesion. It is hypomethylated in the promotor region (Figure 8a) [52].

LIMK1 is a serine/threonine kinase that regulates actin filament dynamics. It phosphorylates and deactivates de-polymerization factors such as CFL1 and CFL2 resulting in the stabilization of actin filament (Figure 8a) [53]. It is involved in metastasis and tumor-cell induced angiogenesis in pancreatic cancer [54].

SLC17A7, SLC25A5, SLC35A2 are series of transporters expressed in organ and tissue of the digestive tract involved in the uptake of a small molecule (Figure 8a) [55, 56, 57].

NRP1 is a prognostic marker in stomach cancer, cervical cancer, renal cancer and glioma (Figure 8a) [58]. NRP1 is a prognostic marker, hypomethylated and co-expressed with PDGFRB resulting in reduced gastric cancer survival [59].

3.9. Subtype II genes

Subtype II is similar to the classical/pancreatic progenitor subtype. The classical/progenitor subtype includes transcription factors that determine the pancreas endoderm fate [13]. The key genes identified in subtype II are: FGFR2, which acts as a cell surface receptor for fibroblast growth factors regulating cell differentiation, proliferation, migration and embryonic development (Figure 8b) [60]. It is responsible for activating MAPK and AKT1 signaling pathway by phosphorylation of FRS2 that activates RAS, MAPK and ERK [60].

Other important genes in the subtype include PDIA2, a member of endoplasmic reticulum family disulphide isomerase that catalyzes protein folding by thiol-disulphide interaction changes specific to the pancreas (Figure 8b) [61]. It is involved in various tumors and is specific to the pancreas. PDIA2 is engaged in multiple tumors, according to recent research [62]. Subtype II is associated with the expression of genes related to digestive enzymes, characteristic of the exocrine pancreatic function such as CYPA1, CYPB.

3.10. Subtype III genes

Subtype III is similar to the immunogenic subtype. The subtype is enriched in key immunological genes. The key subtype III genes are:

BTK encodes Bruton's Tyrosine Kinase that regulates cytokine signaling by PLCG phosphorylation in close cooperation with B cell linker protein BLNK resulting in B lymphocyte development, differentiation and signaling (Figure 8c) [63]. The therapeutic role of BTK inhibition has been reported in PDAC [64]. Mice model studies have shown the therapeutic role of BTK inhibition in PDAC [63].

Another key gene in the signature includes IRAK1, a serine-threonine kinase involved in toll-like receptor and IL-1R signaling that initiates innate immune response against foreign pathogens. TLR activation helps in recruiting MYD88 that phosphorylates IRAK1, which brings together IRAK4, MYD88 and tollip and leads to NK-KB activation (Figure 8c) [65].

DOCK 11 is a guanine nucleotide exchange factor and is important for B cell development and plays a role in the development of B cell in the marginal zone (Figure 8c) [66, 67]. Dock180 contributes to ovarian carcinogenesis and since its overexpression is correlated with poor patient survival, it can be a potential prognostic marker and therapeutic target [68]. High DOCK2 expression is involved in better prognosis in AML [69, 70].

3.11. Subtype IV – Stroma (Microenvironment of tumor) genes

One of the important features of PDAC is non-tumor cells collectively known as stroma responsible for its progression. This feature contributes to the heterogeneity associated with PDAC resulting in less established patients. Therefore, it is important to recognize the key molecular features and biological processes responsible for the heterogeneity and PDAC progression [71]. The key genes identified in subtype IV are:

TAB3 gene forms a ternary complex with protein kinase MAPK3K7/TAK1 leading to the stimulation of pro-inflammatory cytokine and NF-kappa signaling activation (Figure 8d). TAB2 gene exhibits polymorphism and is associated with ovarian cancer susceptibility [72]. TAB3 gene overexpression is associated with poor survival in human esophageal squamous cell carcinoma [73].

SMAD3 is a potential biomarker in PDAC, which promotes cancer's malignant potential through EMT induction in malignant cells [74]. Hepatocyte growth factor promotes pancreatic cancer's growth and behavior by promoting the ductal phenotype (Figure 8d) [75].

Another important gene in the subtype, NSDHL, encodes a gene localized in the endoplasmic reticulum involved in cholesterol biosynthesis (Figure 8d) [76].

3.12. Subtype V genes

Subtype V shows resemblance to the squamous subtype enriched in major pathways like MAPK, Ras protein signaling and chromatin modification. The key genes identified in subtype V are:

MAP3K15 gene is a member of the mitogen-activated protein kinase that is involved in the protein kinase signal transduction pathway (Figure 8e). MKK3/6-p38 MAPK-caspase signaling pathway activation results in the induction of apoptosis induced by Gemcitabine in human pancreatic cancer, serving as a novel marker [77].

PAK3 is a serine-threonine protein kinase that regulates various signaling pathways such as cell migration, cytoskeleton regulation and cell cycle regulation (Figure 8e). PAK3 acts on Ser473-Akt kinase regulating the Akt-GSK3β-β-catenin signaling in several pancreatic cancer cell lines [78].

KDM5D gene is a histone demethylase that plays a major role in histone modification by demethylation of lysine of histone H3 (Figure 8e) [79]. It is found to promote pancreatic cancer by modification of the epigenetic landscape [80].

3.13. Survival curve of subtypes

Survival analysis of the five subtypes was obtained using Kaplan-Meier analysis. Subtype V has the worst clinical outcome as compared to the other subtypes in the survival analysis. No significant differences were observed for survival analysis between samples classified into subtype III and subtype IV (supplementary file II-Figures S17-S18). Subtype II has the best clinical outcomes compared to the other subtypes (supplementary file II-Figure S16). Subtype I has the worst clinical outcomes than that of subtype II, III, and IV. Subtype I show a drastic decline in survival around 1000 days (supplementary file II-Figure S15).

4. Discussion

Pancreatic cancer treatment is faced with the significant challenge of heterogeneity in the genomic profile of patients. However, the advancement of molecular profiling techniques has led to a better understanding of heterogeneity in pancreatic cancer.

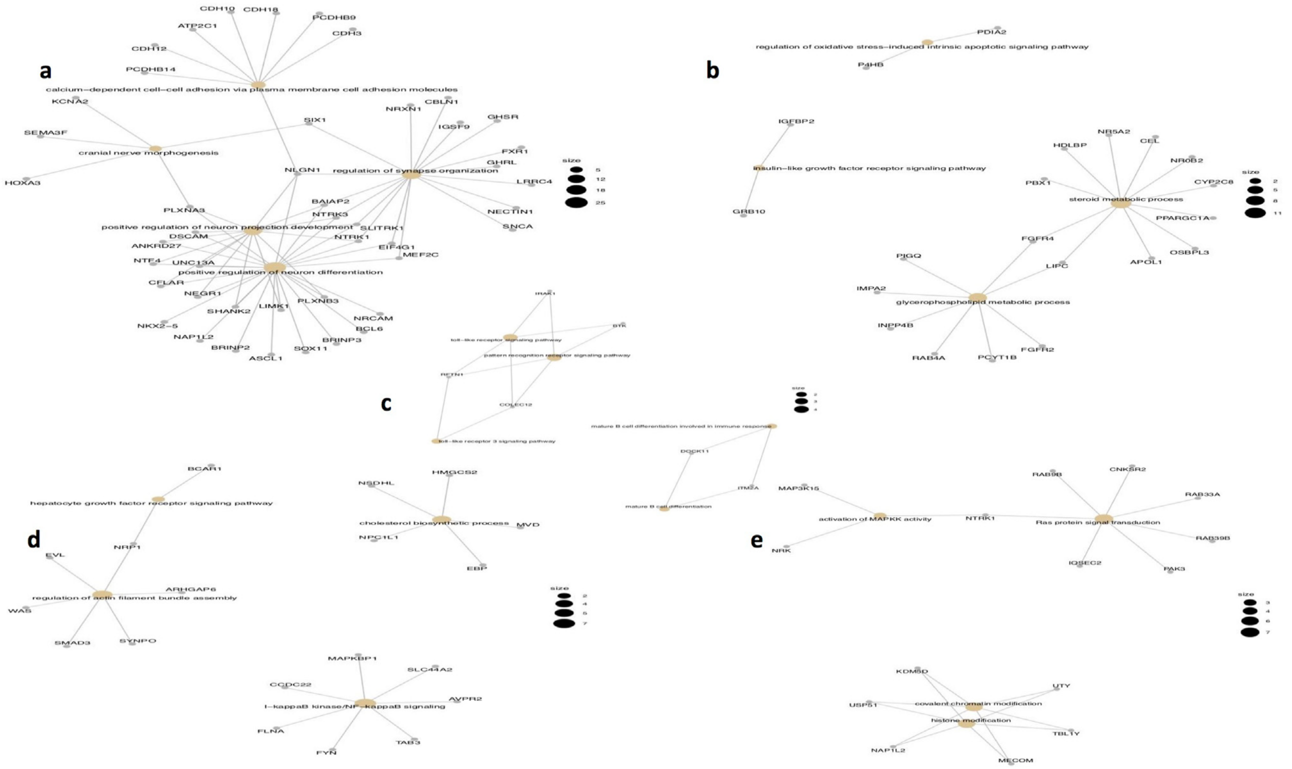
a
CDH10
CDH18
PCDHB9
ATP2C1
CDH3
CDH12
PCDHB14
calcium−dependent cell−cell adhesion via plasma membrane cell adhesion molecules
KCNA2
NRXN1
CBLN1
SIX1
GHSR
SEMA3F
IGSF9
FXR1
cranial nerve morphogenesis
GHRL
NLGN1
regulation of synapse organization
LRRC4
HOXA3
size
5
12
18
25
BAIAP2
PLXNA3
NTRK3
NECTIN1
positive regulation of neuron projection development
SLITRK1
SNCA
DSCAM
NTRK1
ANKRD27
EIF4G1
MEF2C
NTF4
UNC13A
positive regulation of neuron differentiation
CFLAR
NEGR1
PLXNB3
LIMK1
SHANK2
NRCAM
NKX2−5
BCL6
NAP1L2
BRINP3
BRINP2
SOX11
ASCL1
b
PDIA2
regulation of oxidative stress−induced intrinsic apoptotic signaling pathway
P4HB
IGFBP2
NR5A2
HDLBP
CEL
NR0B2
insulin−like growth factor receptor signaling pathway
PBX1
CYP2C8
size
2
5
8
11
steroid metabolic process
GRB10
PPARGC1A
FGFR4
OSBPL3
PIGQ
LIPC
APOL1
IMPA2
glycerophospholipid metabolic process
INPP4B
FGFR2
RAB4A
PCYT1B
c
IRAK1
toll−like receptor signaling pathway
BTK
pattern recognition receptor signaling pathway
RFTN1
COLEC12
size
2
3
4
toll−like receptor 3 signaling pathway
mature B cell differentiation involved in immune response
DOCK11
ITM2A
mature B cell differentiation
d
BCAR1
hepatocyte growth factor receptor signaling pathway
NRP1
EVL
ARHGAP6
regulation of actin filament bundle assembly
WAS
SMAD3
SYNPO
NSDHL
HMGCS2
cholesterol biosynthetic process
MVD
NPC1L1
EBP
size
2
4
5
7
MAPKBP1
SLC44A2
CCDC22
AVPR2
I−kappaB kinase/NF−kappaB signaling
FLNA
TAB3
FYN
e
MAP3K15
activation of MAPKK activity
NTRK1
NRK
RAB6B
CNKSR2
RAB33A
Ras protein signal transduction
RAB39B
IQSEC2
PAK3
size
3
4
6
7
KDM5D
UTY
USP51
covalent chromatin modification
histone modification
TBL1Y
NAP1L2
MECOM

Figure 8. a) Gene ontology analysis of the gene set obtained from correlation analysis using clusterprofiler package shows subtype I having similarity to major pathways of ADEX subtype. b) Gene ontology analysis of the gene set obtained from correlation analysis using clusterprofiler package shows subtypes II having similarity to the major pathways of classical/pancreatic progenitor subtype. c) Gene ontology analysis of the gene set obtained from correlation analysis using clusterprofiler package show subtypes III having similarity to the major pathways of Immunogenic subtype. d) Gene ontology analysis of the gene set obtained from correlation analysis using clusterprofiler package shows subtypes IV having similarity to the major pathways of stroma subtype. e) Gene ontology analysis of the gene set obtained from correlation analysis using clusterprofiler package for shows subtypes V having similarity to the major pathways of squamous subtype.

Compared to the traditional classification technique based on staining and histochemical studies, these methods classify samples into distinct subgroups based on molecular characteristics having clinical implications. However, heterogeneous classification results may be obtained by varying patient cohorts, gene expression platforms, and clustering methods. Different methodologies and platforms have resulted in different classifications of PDAC and it has been classified into two to six subtypes by other groups. But they have their limitations and inconsistency. This prompts a better classification of PDAC, where the role of data integration comes into play. Data integration helps in identifying latent factors hidden across different levels of data and thus helps in better identification of the heterogeneity in data. PDAC has a characteristic feature of abundant stroma that constitutes a major percentage of the tumor mass. The presence of microenvironment, stroma and tumor cell infiltrate make PDAC highly heterogeneous. Besides, PDAC also has infiltrative natures having normal pancreatic components along with the tumor. Tumor microenvironment cells molecular profiling may help define molecular subgroups and identify carcinogenic mechanisms based on mRNA expression and their epigenetic regulation. Studies involving various other datasets involving gene expression, DNA methylation, miRNA and long non-coding RNA have shown putative markers important for survival in PDAC. Genome-wide methylation studies have been performed in TCGA pancreatic cancer datasets involving all pancreatic cancer dataset and normal samples providing three cluster solution with significant insights showing stage-specific subtyping like histologic grade G1 and T3 stage subtype and have shown important gene and methylation on histone modifying core genes like histone reader, editor and eraser genes [81]. In comparison, we have removed the heterogeneity in the TCGA_PAAD data set by removing the non-PDAC samples from PDAC samples. All studies were carried out on matched samples of PDAC. We have applied intNMF to perform an integrative study of varied datasets to perform unsupervised classification of PDAC, which involves two levels of data, namely RNA-Seq and DNA methylation data. We have used all state-of-the-art normalization processes for the DNA methylation data namely SWAN, BMIQ, Noob, and Functional normalization for normalization of DNA methylation data. Amongst these four methods, we showed that SWAN performed best and then we performed integratomics and further downstream studies, which included clustering using intNMF, DEG, DMR, and the correlation study. We propose five molecular and clinical distinct PDAC subtypes and studied survival analysis of these subtypes, based on the integrative clustering approach. Our study improves our understanding of PDAC heterogeneity and further helps decipher the molecular and clinical significance of different subtypes.

The five subtypes emerging out of this analysis correlate properly with the already identified subtypes of PDAC based on Bailey's and Moffitt's classification. This study has shown the methylation landscape in different subtypes of PDAC obtained after clustering and correlation studies.

This study will strengthen our understanding of the impact of methylation landscape of hyper-methylation and hypo-methylation on different gene regions in gene expression profiles of obtained PDAC subtypes. Our study focuses on the role of DNA methylation on gene expression at different loci in the different genes in the heterogeneous PDAC landscape. It will help improve and help in a better understanding of epigenetic regulation on the gene expression in PDAC, using unsupervised classification that will lead to better subtyping, prognosis and personalized medical treatment.

5. Conclusion

Our integrative study proposes five biological subtypes of PDAC, with their distinct molecular features and clinical outcomes. The proposed study will lead to better identification of PDAC and help in a better prognosis, personalized treatment and help in delineating the heterogenetic landscape of PDAC. The obtained subtype-specific genes by our analysis have the potential to drive personalized therapies and risk prediction for PDAC patients [82].

Declarations

Author contribution statement

Shikha Roy: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Amar Pratap Singh: Conceived and designed the experiments; Analyzed and interpreted the data; Wrote the paper.

Dinesh Gupta: Conceived and designed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Funding statement

D. Gupta was supported by the Bioinformatics Infrastructure Facility (BIF) at ICGEB, by The Department of Biotechnology, Ministry of Science and Technology (BT/BI/25/066/2012). S. Roy was supported by the Council for Scientific and Industrial Research, India (IN (09/0512(0212)/2016/EMR-1). A. P. Singh was supported by the Council of Scientific and Industrial Research, India (IN) (09/0512(0209)/2016/EMR-1).

Data availability statement

Data associated with this study is available online in the public domain in the TCGA resource.

Declaration of interests statement

The authors declare no conflict of interest.

Additional information

Supplementary content related to this article has been published online at https://doi.org/10.1016/j.heliyon.2021.e06000.

References

[1] P.A. Jones, S.B. Baylin, The epigenomics of cancer, Cell 128 (4) (2007) 683–692.
[2] S.B. Baylin, P.A. Jones, A decade of exploring the cancer epigenome - biological and translational implications, Nat. Rev. Cancer 11 (10) (2011) 726–734.
[3] P.A. Jones, S.B. Baylin, The fundamental role of epigenetic events in cancer, Nat. Rev. Genet. 3 (6) (2002) 415–428.
[4] M. Aine, G. Sjodahl, P. Eriksson, S. Veerla, D. Lindgren, M. Ringner, M. Hoglund, Integrative epigenomic analysis of differential DNA methylation in urothelial carcinoma, Genome Med. 7 (1) (2015) 23.
[5] H. Noushmehr, D.J. Weisenberger, K. Diefes, H.S. Phillips, K. Pujara, B.P. Berman, F. Pan, C.E. Pelloski, E.P. Sulman, K.P. Bhat, et al., Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma, Canc. Cell 17 (5) (2010) 510–522.
[6] G. Kloppel, J. Luttges, WHO-classification 2000: exocrine pancreatic tumors, Verhandlungen der Deutschen Gesellschaft fur Pathologie 85 (2001) 219–228.
[7] A. Rahemtullah, J. Misdraji, M.B. Pitman, Adenosquamous carcinoma of the pancreas: cytologic features in 14 cases, Cancer 99 (6) (2003) 372–378.
[8] O. Basturk, L. Tang, R.H. Hruban, V. Adsay, Z. Yang, A.M. Krasinskas, E. Vakiani, S. La Rosa, K.T. Jang, W.L. Frankel, et al., Poorly differentiated neuroendocrine carcinomas of the pancreas: a clinicopathologic analysis of 44 cases, Am. J. Surg. Pathol. 38 (4) (2014) 437–447.
[9] O.W. Ryan, J.M. Skerker, M.J. Maurer, X. Li, J.C. Tsai, S. Poddar, M.E. Lee, W. DeLoache, J.E. Dueber, A.P. Arkin, et al., Selection of chromosomal DNA libraries using a multiplex CRISPR system, eLife 3 (2014).
[10] C.L. Wolfgang, J.M. Herman, D.A. Laheru, A.P. Klein, M.A. Erdek, E.K. Fishman, R.H. Hruban, Recent progress in pancreatic cancer, CA: Cancer J Clin 63 (5) (2013) 318–348.
[11] L. Rahib, B.D. Smith, R. Aizenberg, A.B. Rosenzweig, J.M. Fleshman, L.M. Matrisian, Projecting cancer incidence and deaths to 2030: the unexpected burden of thyroid, liver, and pancreas cancers in the United States, Cancer Res. 74 (11) (2014) 2913–2921.
[12] F. Notta, M. Chan-Seng-Yue, M. Lemire, Y. Li, G.W. Wilson, A.A. Connor, R.E. Denroche, S.B. Liang, A.M. Brown, J.C. Kim, et al., A renewed model of pancreatic cancer evolution based on genomic rearrangement patterns, Nature 538 (7625) (2016) 378–382.
[13] P. Bailey, D.K. Chang, K. Nones, A.L. Johns, A.M. Patch, M.C. Gingras, D.K. Miller, A.N. Christ, T.J. Bruxner, M.C. Quinn, et al., Genomic analyses identify molecular subtypes of pancreatic cancer, Nature 531 (7592) (2016) 47–52.
[14] E.A. Collisson, A. Sadanandam, P. Olson, W.J. Gibb, M. Truitt, S. Gu, J. Cooc, J. Weinkle, G.E. Kim, L. Jakkula, et al., Subtypes of pancreatic ductal adenocarcinoma and their differing responses to therapy, Nat. Med. 17 (4) (2011) 500–503.
[15] R.A. Moffitt, R. Marayati, E.L. Flate, K.E. Volmar, S.G. Loeza, K.A. Hoadley, N.U. Rashid, L.A. Williams, S.C. Eaton, A.H. Chung, et al., Virtual microdissection identifies distinct tumor- and stroma-specific subtypes of pancreatic ductal adenocarcinoma, Nat. Genet. 47 (10) (2015) 1168–1178.
[16] L. Zhao, H. Zhao, H. Yan, Gene expression profiling of 1200 pancreatic ductal adenocarcinoma reveals novel subtypes, BMC Cancer 18 (1) (2018) 603.
[17] A.V. Biankin, A. Maitra, Subtyping pancreatic cancer, Cancer Cell 28 (4) (2015) 411–413.
[18] A. Marusyk, M. Janiszewska, K. Polyak, Intratumor heterogeneity: the Rosetta stone of therapy resistance, Cancer Cell 37 (4) (2020) 471–484.
[19] G. Lomberk, Y. Blum, R. Nicolle, A. Nair, K.S. Gaonkar, L. Marisa, A. Mathison, Z. Sun, H. Yan, N. Elarouci, et al., Distinct epigenetic landscapes underlie the pathobiology of pancreatic cancer subtypes, Nat. Commun. 9 (1) (2018) 1978.
[20] Cancer Genome Atlas Research Network, Electronic address aadhe, cancer genome atlas research N: integrated genomic characterization of pancreatic ductal adenocarcinoma, Cancer Cell 32 (2) (2017) 185–203, e113.
[21] N.A. Juiz, J. Iovanna, N. Dusetti, Pancreatic cancer heterogeneity can Be explained beyond the genome, Front. Oncol. 9 (2019) 246.
[22] M.J. Pishvaian, E.M. Blais, J.R. Brody, E. Lyons, P. DeArbeloa, A. Hendifar, S. Mikhail, V. Chung, V. Sahai, D.P.S. Sohal, et al., Overall survival in patients with pancreatic cancer receiving matched therapies following molecular profiling: a retrospective analysis of the Know Your Tumor registry trial, Lancet Oncol. 21 (4) (2020) 508–518.
[23] N.K. Mishra, S. Southekal, C. Guda, Survival analysis of multi-omics data identifies potential prognostic markers of pancreatic ductal adenocarcinoma, Front. Genet. 10 (2019) 624.
[24] Y. Xiao, D. Ma, S. Zhao, C. Suo, J. Shi, M.Z. Xue, M. Ruan, H. Wang, J. Zhao, Q. Li, et al., Multi-omics profiling reveals distinct microenvironment characterization and suggests immune escape mechanisms of triple-negative breast cancer, Clin. Cancer Res. – Offic. J. Am. Assoc. Cancer Res. 25 (16) (2019) 5002–5014.
[25] W. Hu, Y. Yang, X. Li, M. Huang, F. Xu, W. Ge, S. Zhang, S. Zheng, Multi-omics approach reveals distinct differences in left- and right-sided colon cancer, Mol. Cancer Res. : MCR 16 (3) (2018) 476–485.
[26] M.S. Kwon, Y. Kim, S. Lee, J. Namkung, T. Yun, S.G. Yi, S. Han, M. Kang, S.W. Kim, J.Y. Jang, et al., Integrative analysis of multi-omics data for identifying multimarkers for diagnosing pancreatic cancer, BMC Genom. 16 (Suppl 9) (2015) S4.
[27] U.T. Shankavaram, W.C. Reinhold, S. Nishizuka, S. Major, D. Morita, K.K. Chary, M.A. Reimers, U. Scherf, A. Kahn, D. Dolginow, et al., Transcript and protein expression profiles of the NCI-60 cancer cell panel: an integromic microarray study, Mol. Cancer Therapeut. 6 (3) (2007) 820–832.
[28] A.C. Ahn, M. Tewari, C.S. Poon, R.S. Phillips, The limits of reductionism in medicine: could systems biology offer an alternative? PLoS Med. 3 (6) (2006) e208.
[29] Y. Liu, V. Devescovi, S. Chen, C. Nardini, Multilevel omic data integration in cancer cell lines: advanced annotation and emergent properties, BMC Syst. Biol. 7 (2013) 14.
[30] A. Colaprico, T.C. Silva, C. Olsen, L. Garofano, C. Cava, D. Garolini, T.S. Sabedot, T.M. Malta, S.M. Pagnotta, I. Castiglioni, et al., TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data, Nucleic Acids Res. 44 (8) (2016) e71.
[31] T.J. Morris, L.M. Butcher, A. Feber, A.E. Teschendorff, A.R. Chakravarthy, T.K. Wojdacz, S. Beck, ChAMP: 450k chip analysis methylation pipeline, Bioinformatics 30 (3) (2014) 428–430.
[32] J. Nordlund, C.L. Backlin, P. Wahlberg, S. Busche, E.C. Berglund, M.L. Eloranta, T. Flaegstad, E. Forestier, B.M. Frost, A. Harila-Saari, et al., Genome-wide signatures of differential DNA methylation in pediatric acute lymphoblastic leukemia, Genome Biol. 14 (9) (2013) r105.
[33] L.M. Butcher, S. Beck, Probe Lasso: a novel method to rope in differentially methylated regions with 450K DNA methylation data, Methods 72 (2015) 21–28.
[34] J. Maksimovic, L. Gordon, A. Oshlack, SWAN: Subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips, Genome Biol. 13 (6) (2012) R44.
[35] S. Dedeurwaerder, M. Defrance, E. Calonne, H. Denis, C. Sotiriou, F. Fuks, Evaluation of the infinium methylation 450K technology, Epigenomics 3 (6) (2011) 771–784.
[36] A.E. Teschendorff, F. Marabita, M. Lechner, T. Bartlett, J. Tegner, D. Gomez-Cabrero, S. Beck, A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data, Bioinformatics 29 (2) (2013) 189–196.
[37] T.J. Triche Jr., D.J. Weisenberger, D. Van den Berg, P.W. Laird, K.D. Siegmund, Low-level processing of illumina infinium DNA methylation BeadArrays, Nucleic Acids Res. 41 (7) (2013) e90.

[38] J.P. Fortin, A. Labbe, M. Lemire, B.W. Zanke, T.J. Hudson, E.J. Fertig, C.M. Greenwood, K.D. Hansen, Functional normalization of 450k methylation array data improves replication in large cancer studies, Genome Biol. 15 (12) (2014) 503.

[39] S. Anders, W. Huber, Differential expression analysis for sequence count data, Genome Biol. 11 (10) (2010) R106.

[40] P. Chalise, B.L. Fridley, Integrative clustering of multi-level 'omic data based on non-negative matrix factorization algorithm, PloS One 12 (5) (2017), e0176278.

[41] R. Gaujoux, C. Seoighe, A flexible R package for nonnegative matrix factorization, BMC Bioinf. 11 (2010) 367.

[42] S. Zhao, J. Sun, K. Shimizu, K. Kadota, Silhouette scores for arbitrary defined groups in gene expression data and insights into differential expression results, Biol. Proced. Online 20 (2018) 5.

[43] M.D. Robinson, D.J. McCarthy, G.K. Smyth, edgeR: a Bioconductor package for differential expression analysis of digital gene expression data, Bioinformatics 26 (1) (2010) 139–140.

[44] M.D. Robinson, G.K. Smyth, Small-sample estimation of negative binomial dispersion, with applications to SAGE data, Biostatistics 9 (2) (2008) 321–332.

[45] M.D. Robinson, G.K. Smyth, Moderated statistical tests for assessing differences in tag abundance, Bioinformatics 23 (21) (2007) 2881–2887.

[46] T.C. Silva, A. Colaprico, C. Olsen, F. D'Angelo, G. Bontempi, M. Ceccarelli, H. Noushmehr, TCGA Workflow: analyze cancer genomics and epigenomics data using Bioconductor packages, F1000Research 5 (2016) 1542.

[47] K.D. Siegmund, Statistical approaches for the analysis of DNA methylation microarray data, Hum. Genet. 129 (6) (2011) 585–595.

[48] G. Yu, L.G. Wang, Y. Han, Q.Y. He, clusterProfiler: an R package for comparing biological themes among gene clusters, OMICS A J. Integr. Biol. 16 (5) (2012) 284–287.

[49] R. Nicolle, J. Raffenne, V. Paradis, A. Couvelard, A. de Reynies, Y. Blum, J. Cros, Prognostic biomarkers in pancreatic cancer: avoiding errata when using the TCGA dataset, Cancers 11 (1) (2019).

[50] T. Wang, W. Guan, J. Lin, N. Boutaoui, G. Canino, J. Luo, J.C. Celedon, W. Chen, A systematic study of normalization methods for Infinium 450K methylation data using whole-genome bisulfite sequencing data, Epigenetics 10 (7) (2015) 662–669.

[51] T.L. Kuo, K.H. Cheng, L.T. Chen, W.C. Hung, Deciphering the potential role of hox genes in pancreatic cancer, Cancers 11 (5) (2019).

[52] K. Sakamoto, K. Imai, T. Higashi, K. Taki, S. Nakagawa, H. Okabe, H. Nitta, H. Hayashi, A. Chikamoto, T. Ishiko, et al., Significance of P-cadherin overexpression and possible mechanism of its regulation in intrahepatic cholangiocarcinoma and pancreatic cancer, Cancer Sci. 106 (9) (2015) 1153–1162.

[53] T. Sumi, K. Matsumoto, Y. Takai, T. Nakamura, Cofilin phosphorylation and actin cytoskeletal dynamics regulated by rho- and Cdc42-activated LIM-kinase 2, J. Cell Biol. 147 (7) (1999) 1519–1532.

[54] D.H. Vlecken, C.P. Bagowski, LIMK1 and LIMK2 are important for metastatic behavior and tumor cell-induced angiogenesis of pancreatic cancer cells, Zebrafish 6 (4) (2009) 433–439.

[55] J. Xie, C.S. Cheng, X.Y. Zhu, Y.H. Shen, L.B. Song, H. Chen, Z. Chen, L.M. Liu, Z.Q. Meng, Magnesium transporter protein solute carrier family 41 member 1 suppresses human pancreatic ductal adenocarcinoma through magnesium-dependent Akt/mTOR inhibition and bax-associated mitochondrial apoptosis, Aging 11 (9) (2019) 2681–2698.

[56] S. Panda, N. Banerjee, S. Chatterjee, Solute carrier proteins and c-Myc: a strong connection in cancer progression, Drug Discov. Today (2020).

[57] B. Mohelnikova-Duchonova, V. Brynychova, V. Hlavac, M. Kocik, M. Oliverius, J. Hlavsa, E. Honsova, J. Mazanec, Z. Kala, B. Melichar, et al., The association between the expression of solute carrier transporters and the prognosis of pancreatic cancer, Canc. Chemother. Pharmacol. 72 (3) (2013) 669–682.

[58] G. Zhang, L. Chen, A.A. Khan, B. Li, B. Gu, F. Lin, X. Su, J. Yan, miRNA-124-3p/neuropilin-1(NRP-1) axis plays an important role in mediating glioblastoma growth and angiogenesis, Int. J. Cancer 143 (3) (2018) 635–644.

[59] G. Wang, B. Shi, Y. Fu, S. Zhao, K. Qu, Q. Guo, K. Li, J. She, Hypomethylated gene NRP1 is co-expressed with PDGFRB and associated with poor overall survival in gastric cancer patients, Biomed. Pharmacother. – Biomedecine & pharmacotherapie 111 (2019) 1334–1341.

[60] D.M. Ornitz, N. Itoh, The fibroblast growth factor signaling pathway, Wiley Interdiscipl. Rev. Dev. Biol. 4 (3) (2015) 215–266.

[61] E. Lee, D.H. Lee, Emerging roles of protein disulfide isomerase in cancer, BMB Rep. 50 (8) (2017) 401–410.

[62] Z. Wang, H. Zhang, Q. Cheng, PDIA4: the basic characteristics, functions and its potential connection with cancer, Biomed. Pharmacother. – Biomedecine & pharmacotherapie 122 (2020) 109688.

[63] A.J. Gunderson, M.M. Kaneda, T. Tsujikawa, A.V. Nguyen, N.I. Affara, B. Ruffell, S. Gorjestani, S.M. Liudahl, M. Truitt, P. Olson, et al., Bruton tyrosine kinase-dependent immune cell cross-talk drives pancreas cancer, Cancer Discov. 6 (3) (2016) 270–285.

[64] M. Overman, M. Javle, R.E. Davis, P. Vats, C. Kumar-Sinha, L. Xiao, N.B. Mettu, E.R. Parra, A.B. Benson, C.D. Lopez, et al., Randomized phase II study of the Bruton tyrosine kinase inhibitor acalabrutinib, alone or with pembrolizumab in patients with advanced pancreatic cancer, J. Immunother. Cancer 8 (1) (2020).

[65] D. Zhang, L. Li, H. Jiang, B.L. Knolhoff, A.C. Lockhart, A. Wang-Gillam, D.G. DeNardo, M.B. Ruzinova, K.H. Lim, Constitutive IRAK4 activation underlies poor prognosis and chemoresistance in pancreatic ductal adenocarcinoma, Clin. Cancer Res. Offic. J. Am. Assoc. Cancer Res. 23 (7) (2017) 1748–1759.

[66] J.F. Cote, K. Vuori, Identification of an evolutionarily conserved superfamily of DOCK180-related proteins with guanine nucleotide exchange activity, J. Cell Sci. 115 (Pt 24) (2002) 4901–4913.

[67] Q. Lin, W. Yang, D. Baird, Q. Feng, R.A. Cerione, Identification of a DOCK180-related guanine nucleotide exchange factor that is capable of mediating a positive feedback activation of Cdc42, J. Biol. Chem. 281 (46) (2006) 35253–35262.

[68] F. Zhao, M.K. Siu, L. Jiang, K.F. Tam, H.Y. Ngan, X.F. Le, O.G. Wong, E.S. Wong, H.Y. Chan, A.N. Cheung, Overexpression of dedicator of cytokinesis I (Dock180) in ovarian cancer correlated with aggressive phenotype and poor patient survival, Histopathology 59 (6) (2011) 1163–1172.

[69] N. Hu, Y. Pang, H. Zhao, C. Si, H. Ding, L. Chen, C. Wang, T. Qin, Q. Li, Y. Han, et al., High expression of DOCK2 indicates good prognosis in acute myeloid leukemia, J. Cancer 10 (24) (2019) 6088–6094.

[70] E.A. Collisson, P. Bailey, D.K. Chang, A.V. Biankin, Molecular subtypes of pancreatic cancer, Nat. Rev. Gastroenterol. Hepatol. 16 (4) (2019) 207–220.

[71] V.L. Veenstra, A. Garcia-Garijo, H.W. van Laarhoven, M.F. Bijlsma, Extracellular influences: molecular subclasses and the microenvironment in pancreatic cancer, Cancers 10 (2) (2018).

[72] X. Huang, C. Shen, Y. Zhang, Q. Li, K. Li, Y. Wang, Y. Song, M. Su, B. Zhou, W. Wang, Associations between TAB2 gene polymorphisms and epithelial ovarian cancer in a Chinese population, Dis. Markers 2019 (2019) 8012979.

[73] J. Zhao, L. Gai, Y. Gao, W. Xia, D. Shen, Q. Lin, W. Mao, F. Wang, P. Liu, J. Chen, TAB3 promotes human esophageal squamous cell carcinoma proliferation and invasion via the NFkappaB pathway, Oncol. Rep. 40 (5) (2018) 2876–2885.

[74] K. Yamazaki, Y. Masugi, K. Effendi, H. Tsujikawa, N. Hiraoka, M. Kitago, M. Shinoda, O. Itano, M. Tanabe, Y. Kitagawa, et al., Upregulated SMAD3 promotes epithelial-mesenchymal transition and predicts poor prognosis in pancreatic ductal adenocarcinoma, Lab. Invest. J. Tech. Methods Pathol. 94 (6) (2014) 683–691.

[75] M.F. Di Renzo, R. Poulsom, M. Olivero, P.M. Comoglio, N.R. Lemoine, Expression of the Met/hepatocyte growth factor receptor in human pancreatic cancer, Cancer Res. 55 (5) (1995) 1129–1138.

[76] H. Caldas, G.E. Herman, NSDHL, an enzyme involved in cholesterol biosynthesis, traffics through the Golgi and accumulates on ER membranes and on the surface of lipid droplets, Hum. Mol. Genet. 12 (22) (2003) 2981–2991.

[77] A. Habiro, S. Tanno, K. Koizumi, T. Izawa, Y. Nakano, M. Osanai, Y. Mizukami, T. Okumura, Y. Kohgo, Involvement of p38 mitogen-activated protein kinase in gemcitabine-induced apoptosis in human pancreatic cancer cells, Biochem. Biophys. Res. Commun. 316 (1) (2004) 71–77.

[78] H.Y. Wu, M.C. Yang, L.Y. Ding, C.S. Chen, P.C. Chu, p21-Activated kinase 3 promotes cancer stem cell phenotypes through activating the Akt-GSK3beta-beta-catenin signaling pathway in pancreatic cancer cells, Cancer Lett. 456 (2019) 13–22.

[79] J. Cui, M. Quan, D. Xie, Y. Gao, S. Guha, M.B. Fallon, J. Chen, K. Xie, A novel KDM5A/MPC-1 signaling pathway promotes pancreatic cancer progression via redirecting mitochondrial pyruvate metabolism, Oncogene 39 (5) (2020) 1140–1151.

[80] A. D'Oto, Q.W. Tian, A.M. Davidoff, J. Yang, Histone demethylases and their roles in cancer epigenetics, J. Med. Oncol. Therapeut. 1 (2) (2016) 34–40.

[81] N.K. Mishra, C. Guda, Genome-wide DNA methylation analysis reveals molecular subtypes of pancreatic cancer, Oncotarget 8 (17) (2017) 28990–29012.

[82] J.S. Parker, M. Mullins, M.C. Cheang, S. Leung, D. Voduc, T. Vickery, S. Davies, C. Fauron, X. He, Z. Hu, et al., Supervised risk predictor of breast cancer based on intrinsic subtypes, J. Clin. Oncol. Offic. J. Am. Soc. Clin. Oncol. 27 (8) (2009) 1160–1167.