BIOINFORMATIC ANALYSIS TO IDENTIFY AND UNDERSTAND ABERRANT DNA METHYLATION PATTERN ASSOCIATED WITH PANCREATIC CANCER

A Thesis
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Mariam Zamani

In Partial Fulfillment of the Requirements
for the Degree of
MASTER OF SCIENCE

Major Program:
Genomics, Phenomics, and Bioinformatics

September 2021

Fargo, North Dakota

North Dakota State University

Graduate School

Title

BIOINFORMATIC ANALYSIS TO IDENTIFY AND UNDERSTAND ABBERANT DNA METHYLATION PATTERN ASSOCIATED WITH PANCREATIC CANCER

By

Mariam Zamani

The Supervisory Committee certifies that this ***disquisition*** complies with North Dakota State University's regulations and meets the accepted standards for the degree of

MASTER OF SCIENCE

SUPERVISORY COMMITTEE:

Dr. Rick Jansen

Chair

Dr. Phillip E. McClean

Dr. Changhui Yan

Approved:

October 14, 2021 — Date

Dr. Phillip E. McClean — Department Chair

ABSTRACT

In this study, we searched for significant hypo and hyper methylation CpG (*5′-C-phosphate-G-3′*) probes from The Cancer Genome Atlas (TCGA) datasets. First, the relationship between hypo and hypermethylation pattern in significantly expressed genes associated in pancreatic ductal adenocarcinoma (PDAC) was analyzed using computational methodologies in R package. This was done by combining DNA methylation (DM) and gene expression (GE) information, and their corresponding metadata (i.e., clinical data and molecular subtypes) and saved as R files. Next, examination of differentially methylated CpG sites (DMCs) between two groups (normal vs tumor) was identified gene sets. From this analysis, we found nine (09) overexpressed hypomethylated and six (06) under expressed hypermethylated genes near significant CpG probes. Results from this work will shed light on the relationship between CpG methylation and gene expression associated with PDAC.

ACKNOWLEDGEMENTS

This study of pancreatic cancer research was funded by The Center for Diagnostic and Therapeutic Strategies in Pancreatic Cancer (CDTSP) at North Dakota State University (NDSU) and is supported by the National Institute of General Medical Sciences (NIGMS) of the National Institutes of Health (NIH) under award number 1P20GM109024. This project used the resources of the Center for Computationally Assisted Science and Technology (CCAST) at (NDSU), which were made possible in part by NSF MRI Award No. 2019077. Last, the results presented in this thesis was based upon data collected and stored by The Cancer Genome Atlas (TCGA) research network.

The success of this project was made possible by guidance and continual support from many people. First and foremost, I would like to acknowledge my thesis advisor Dr. Rick Jansen, Dept. of Public Health, NDSU, for steering this pancreatic cancer project. Many thanks to team members Amelia Nichols, David Adeleke, Tabassum Tanha, and Nijhum Paul. Special thanks to Dr. Rahul Gomes, Dept. Of Computer Science, University of Wisconsin Eu-Claire (UWEC). Also, I would like to express appreciation to my committee members Dr. Phillip E. McClean, Dept. of Plant Science and Dr. Changhui Yan, Dept of Computer Science, of NDSU.

DEDICATION

This work is dedicated to my family and especially to my daughter Anuva.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

CGIs	Cytosine guanine islands are DNA methylation regions in promoters.
CR	Chromatin remodeling.
CpG	A region of DNA where a cytosine nucleotide is followed by a guanine nucleotide in the same strand.
CTCF	CCCTC binding factor
CNV	Copy number (CNV)
DM	DNA methylation.
DE analysis	Differential expression analysis.
DEGs	Differential expressed genes.
DMC	Different methylated CpG sites
DMR	Different methylated regions
DT	Decide test.
DNMT3L	Non-catalytic homolog of DNMT3A/B is DNMT3L
DNMTs	DNA methyltransferase enzymes
DNMTi	DNA methyltransferase inhibitors
EGFR	Epidermal growth factor
EMT	Epithelial to mesenchymal transition
FDA	Food and Drug Administration
FGF	Fibroblast growth factor
GDC	Genomic Data Commons.
GO	Gene ontology analysis is a technique for annotation of genes.
GO:BP	Go biological process.
GO:CC	Go cellular component

GO:MF	Go molecular function
GO:P	Go process
HA	Histone acetylation
HATs	Histone acetyltransferase
HBF	Heparin-binding growth factors
HDMs	Histone demethylases
HDACs	Histone deacetylases (HDAC1 and HDAC2)
HGF	Hepatocyte growth factor
HMTs	Histone methyltransferase
HKMTs	Histone lysine methyltransferase
IGF	Insulin like growth factor
KEAP1	Kelch like ECH associated protein 1
KEGG	Kyoto Encyclopedia of Genes and Genomes
KEAP1	Kelch like ECH associated protein 1
KRAS	KRAS gene provides instruction for making protein K-ras
IDH	Isocitrate dehydrogenase
ICGC	The International Cancer Genome Consortium (ICGC)
ncRNA	Noncoding RNAs
NT	Normal tissues
MAE	Multi assay experiment
MBDs	methyl-CpG-binding domain proteins
MeCP2	methyl-CpG-binding protein 2 (MeCP2).
RT-PCR	Real time polymerase chain reaction
SAM	S-adenosyl-L-methionine

SE.....Summarized experiment

TCGA.....The Cancer Genome Atlas

TGF.....Transforming growth factor

TEs.....Transposable elements

TET2.....Ten-Eleven Translocation-2

TF.....Transcription factors

TP.....Primary solid tumor

TMM.....Trimmed mean of M-values

TSG.....Tumor suppressor genes

TSS.....Transcription start site

PC.....Pancreatic cancer

PAAD.....Pancreatic ductal adenocarcinoma.

PCA.....Principal component analysis

PDAC.....Pancreatic ductal adenocarcinoma

PNE.....Pancreatic neuroendocrine tumors

UCSC.....The University of California, Santa Cruz

WGS.....Whole genome sequencing

LIST OF SYMBOLS

β .....Beta value

FDR.....False discovery rate

5hmU.....5-hydroxymethyluracil (5hmU)

5caC 5-carboxylcytosine (5caC)

5fC.....5-formylcytosine

5hmC.....5-hydroxymethylcytosine

5mC.....5-methylcytosine

M.....Methylated allele intensity

Pp .....Permutation p-value

Pe.....Empirical p-value

U.....Unmethylated allele intensity

μ .....Mean methylation of group (group 1 vs group 2)

1. BACKGROUND

Pancreatic ductal adenocarcinoma (PDAC) is described as a hard-to-treat malignant disease, with a 5-year survival rate of approximately 10% in the USA, and it is progressively becoming a frequent cause of cancer fatality.[1] The prognosis remains poor with only 20% survival following 5 years post-surgery, even for a small subset of patients who are diagnosed with a localized, resectable tumor.[2] According to the World Health Organization (WHO), a continual rise in PDAC incidence and mortality is expected over the next 20 years, and by the year 2040, it is estimated to be 77.8% and 80%, respectively. [3] [4].

The aim of this thesis is identification of methylated CpG sites by analyzing publicly available datasets from The Cancer Genome Atlas (TCGA) to correlate with specific gene expression patterns of PDAC. [5] [6] This study utilizes samples (178 methylation and 164 from gene expression) downloaded from TCGA for differential and pathway analysis of gene expression and methylation patterns of genes (2 kb) up and downstream of transcription start site (TSS).[7] Kwon et. al. analyzed 23 imprinted genes, and we studied methylation patterns of 178 probes and 164 genes from PDAC datasets.[8] Screening for aberrant DNA methylation (DM) pattern of PDAC datasets with computational tools, such as R statistical software, we can now detect the altered CpG methylation status and specific genes that could support early diagnosis of this fatal disease.[9]

1.1. Epigenetics and gene expression

Gene expression (GE) involves processing DNA sequence information, via transcription and translation that leads to assembly of amino acids based on genetic code encoded in DNA.[10] The cells read the sequence of the gene in group of three bases (codon) and each group of three bases correspond to twenty amino acids that are used to build protein molecules.[11] Epigenetic changes can influence gene expression and affect the proteins that are being translated.[10] Three

major epigenetic modification process of human genome includes DNA methylation (DM), histone modifications (HM) and ribonucleic acid (RNA) associated silencing.[12] Accordingly, DM is a covalent modification in which methyl groups are added to cytosine nucleotides in DNA and are catalyzed by DNA methyltransferase (DNMTs) at CpG sites.[13] Two general methylation states are (1) enhanced DM in the promoter regions, that results in gene silencing; and (2) reduced methylation, which results in gene overexpression.[14] Consequently, DM can also lower GE by reducing the binding of transcription factors (TF), or by increasing the binding of methyl-CpG binding of proteins.[15]

DNA structure is periodically wrapped around histone proteins, forming tightly packed units named nucleosomes.[16] The close association of histone proteins and DNA is, the less accommodating for transcription to occur.[17] Furthermore, RNA's in the form of antisense transcript, noncoding RNAs or RNA interference can provoke a heritable and stable gene silencing.[18] Overall, the three epigenetic modification plays a decisive role in localized control of expression, the regulation of embryonic development, formation, and maintenance of cellular identity in human genomes.[19]

1.2. DNA methylation and epigenetic process

An important epigenetic process is DNA methylation, which involves covalent bonding of a methyl group ($-CH_3$) to a CpG site in mammalian cells.[20] Epigenetic changes are reversible and do not alter DNA sequences but changes how DNA sequence is being read in the human body. [21] Correlation between DM and GE have widely reported, as methylation enhances or silences specific genes.[22] Two dynamic process preserve DM patterns in human genome methylation maintenance and de-novo methylation.[23] Methylation maintenance allows preservation of methylation marks in replication generation. Likewise, de-novo methylation occurs on CpG sites

and increase methylation pattern across cell generation.[24, 25] Moreover, unusual levels of hypo and hyper methylation is well defined characteristics of various cancers. As cancer progresses DM levels can vary and identification of these states can assist in PDAC, long before symptoms become apparent or transform to advanced pathological stages.[26] Furthermore, there are stretches of DNA near promoter regions that are rich in GC- or AT-rich sequences and are found to be unmethylated in normal tissues. [27] These CpG sites become highly methylated during tumorigenesis, causing irregular gene expression or repression. This unusuality is a specific epigenetic mark of carcinogenesis.[28]

DNA methylation can become inactive or obstructed by the action of by Ten-Eleven Translocation-2 (TET2) or isocitrate dehydrogenase (IDH) mutations. IDH, located at codon R132 of IDH1 gene, produces onco-metabolite "2-hydroxyglutarate" that induces epigenetic change, such as DM. Moreover, TET genes, and particularly TET2, catalyze the successive oxidation of 5-methylcytosine (5mC) to 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC), and 5-carboxylcytosine (5caC) and can reverse the DM process.[29] Consistently, the loss of 5hmC observed in tumor cells can be used to identify early-stage malignant disease.[30] TET protein is also known to catalyze the hydroxylation of thymine bases in T:A base pairs to form 5-hydroxymethyluracil (5hmU) by deamination reactions.[31] This thymine base modification of 5hmU has been reported to affect protein-binding to DNA and is a key intermediate in generating site-specific mutations.[27]

Transposable elements (TEs, transposon, or jumping gene) form most repetitive sequences (up to 50% of the human genome) and maintain genomic stability, chromosomal architecture, and transcriptional regulation. Active TEs are considered very mutagenic and linked with multiple steps of cancer development and progression.[32] Mesothelin (MSLN), a tumor-associated antigen,

is found highly expressed in pancreatic ductal adenocarcinoma (PDAC), and a long-terminal repeat (LTR) is its primary promoter. Podocalyxin (PODXL) is a highly glycosylated type I transmembrane protein that can be found in normal tissues as well as in many cancers, including lung, renal, breast, colorectal, and pancreas.[33] In a recent bioinformatics study by Wong et. al, dynamin-2 was identified a binding partner of PODXL in PDAC. The interaction between PODXL with dynamin-2, promotes PC cancer cell migration and metastasis by regulating microtubule and focal adhesion dynamics.[34] Furthermore, overexpression of PODXL in PC cancer cells was linked with advanced clinicopathological stage and poor clinical outcome.[33]

1.3. DNA methylation and chromatin structure

Epigenetics mechanism enables cells to retain memory of preceding cellular environments and disruptions, in nearly all cell types, without changing DNA sequences.[35] Epigenetic changes can communicate with TF and translation process by remodeling chromatin architecture to fine-tune GE patterns.[36] Epigenetic changes include DM and histone modifications (HM), and both are known to inhibit gene transcription.[36] Pertinently, histone protein undergoes post-translational modifications that occur in the N-terminal tails of the core histone, and it also includes acetylation, methylation, phosphorylation, ubiquitination, and sumoylation.[37] As a result, these alteration influences the chromatin structure and can create affinities for chromatin binding proteins, thus regulating expression pattern of genes.[36] Acetylation of histones unwinds their electrostatic interaction with DNA, resulting in unperturbed chromatin states allowing upregulation of transcriptions. Cervoni et al. observed reduced CpG methylation concentrations when a human lymphoma cell line was transfected with HDAC inhibitors. This finding illustrates the suppression of histone deacetylation cause demethylation of DNA.[38]

Histone acetyltransferase (HATs) and deacetylases (HDACs) are responsible for the addition and removal of acetyl groups to/from lysine residues.[39] In oncogenesis, the imbalance of HDACs result in transcriptional inactivation of tumor-suppressor genes (TSG).[40] Moreover, epigenetic gene silencing can occur through unusual docking of HDACs to the gene promoter, resulting in histone hypoacetylation.[41] Importantly, expression of these pathways can resume by inhibition of HDAC activity, such as CpG hypermethylation.[42]

1.4. Growth factor and pancreatic cancer

Growth factor signaling pathways are known to participate in pancreatic tumorigenesis, that include transforming growth factor (TGF), epidermal growth factor (EGF), insulin like growth factor (IGF), hepatocyte growth factor (HGF) and fibroblast growth factor (FGF).[43] Transforming growth factor-β (TGF-β) regulates cell functions and has significant roles in initiation of PDAC development.[44] SMAD4, one of the Smads family of signal transducer from TGF-β, mediates pancreatic cell proliferation and apoptosis.[45] It plays a dual role in tumor initiation and progression (up to 42%) and overexpression of TGF- β is associated with poor prognosis of PDAC. [46] Xia et al. observed a complete blockage of the growth inhibitory effects of TGF- β in PC cell line COLO-357 and in nude mice, following transfection with Smad5 and Smad7.[47] Therefore, PC cells may have multiple mechanisms to evade the tumor suppressive effects of TGF- β, and capacity to express metastasis-promoting genes.[48]

In cancer cells, the upregulation of EGF and EGF-receptor are common in PDAC with lymph nodes and distant metastasis.[49] An increase in expression of human EGFR is associated with advanced tumor stage and poor survival. Consequently, expression of EGF-receptor and transforming growth factor alpha (TGF-α) or amphiregulin is tied to poor survival prognosis.[50] EGFR family members activate the Ras/MAPK, PI(3)K/Akt pathways. KRAS, an activating

mutation of an isoform of the Ras protein appear during early stages of malignant transformation and are found in almost all PDAC cases.[51] Moreover, DNA methyltransferase enzymes (DNMTs) expression and activity are regulated by Ras and ERK/MAPK signaling pathways, which function downstream of the epidermal growth factor receptor (EGFR) pathway. This means activation of EGFR and its subsequent signaling pathways play a role in DNMT expression and DM.[31, 52] As a result, an increase in DNMTs enzymatic activity are widely linked with cancer progression and poor prognosis.[53]

Cancer cells are known to express insulin and IGF1 receptors, and these receptors are key activators of the Akt and mitogen-activated protein kinase signaling networks in neoplastic tissue.[54] In a recent study, the survivability of a significant subset of latent cancer cells that was dependent on IGF-1R signaling in absence of oncogenic KRAS expression was identified.[55] Glypican-1 (GPC1) is a member of the heparin sulfate proteoglycans family and is the most important coreceptor for heparin-binding growth factors (HBF). Glypican-1 was found overexpressed in majority of PDAC cells and in the fibroblast surrounding the tumor mass. Downregulation of GPC1 decreases the tumorigenicity of pancreatic cancer cells, whereas high levels of GPC1 are associated with poor survival in patients with PC.[56] Therefore, a decrease in GPC-1 expression marks the reduction of sensitivity of PDAC cells to HBF and therefore, makes it a prospective prognostic biomarker in this disease.[57] Fibroblast growth factor (FGF) receptors are found often over-expressed in PDAC cell lines. Fibroblast growth factor (FGF) and its receptors are proposed have a role in tumor angiogenesis, in a study of transgenic mouse model of pancreatic β-cell carcinogenesis.[58]

1.5. Development of novel biomarkers toward pancreatic cancer treatments

Recent advances toward discovery of novel therapeutic agents and to distinguish new biomarkers for existing epigenetic inhibitors is a key research interest. Of late, two group of drugs are currently used in epigenetic therapy: (1) DNA methyltransferase inhibitors (DNMTi) and histone deacetylase inhibitors (HDACi); (2) targeted therapeutic agents to block specific genes that when mutated cause dysregulation of epigenetic markers.[59] Correspondingly, these two classes of inhibitors are under ongoing clinical trial to treat different tumor types, and some have been approved by FDA. Therefore, identification of epigenetic irregularities is vital for the development of new therapies and for discovery of candidate biomarkers. [60] The DNA methylation has many distinctive advantages in individualized cancer therapy.

1.6. DNA methylation and gene expression

The local control of gene expression, regulation of embryonic development, formation, and maintenance of cellular identity of human genomes is significantly influenced by DM. Aberrant DNA methylation patterns contributes to critical signaling pathways involved in pancreatic tumorigenesis.[61] For example, TGF-β promotes epithelial to mesenchymal transition (EMT) of PDAC cells partly by inducing hypermethylation of CpG site in *VAV1* gene body and VAV1 expression.[62] Guo et al. examined 48 pancreatic exocrine and endocrine neoplasms for DM changes of specific gene promoter regions, that includes acinar cell carcinomas, PDAC, and neuroendocrine tumors, and found six most frequently methylated genes (APC 50%, BRCA1 46%, p16INK4a 35%, p15INK4b 35%, RARβ 35%, and p73 33%).[63] Many studies suggest several genes, abnormalities, or mutation are associated with PDAC. Among these, KRAS, TP53, CDKN2A and BRACA have been extensively reported to be the drivers of PDAC [64]. Nones et al.

assessed DM level in 167 untreated resected PDAs and compared them to 29 adjacent non-transformed pancreatic tissue and identified 3522 genes that are differentially methylated.[65 66]

1.7. DNA methyltransferase enzyme

Epigenetic changes occur several ways that can affect genome activity and expression without changing DNA-sequence. Consequently, epigenetic changes or marks occur both on histone and DNA by a network of proteins that includes change by addition of epigenetic marks (writers), changing the state of existing epigenetic marks by their removal (erasers), lastly, react to specific epigenetic marks (readers). Writers of DM are known as DNA methyltransferases (DNMTs), that are essential for the transfer of a methyl group from the universal methyl donor, *S*-adenosyl-L-methionine (SAM), to the 5-position of cytosine residues in DNA. Unusual expression of DNMTs and interruption of DM patterns are linked to many types of cancer. Therefore, DNMTs are prospective therapeutic targets for diagnostic and prognostic markers for PDAC.[67]

DNMT1, a member of the DNMTs family of writers, can be found proliferating cells and copy methylation patterns in the newly created daughter-DNA strands during DNA replication. Although their expression decreases in adult somatic tissues, methylation is evident in the early embryonic stage and is carried out by the writers DNMT3A and DNMT3B on regions of chromosomes where the nucleosome is reduced. DNMT1 can interact with the tumor suppressor gene MEN1 protein, which reversibly regulates PC cell growth. MEN1 gene express Menin protein that are known to be active in all stages of development and are frequently mutated in pancreatic neuroendocrine (PNE) tumors.[68] A quantitative real-time RT-PCR of DNMT expression pattern revealed that the mRNA expression levels of DNMT1, DNMT3A, and DNMT3B increased from normal ducts to PNE and then to PC.[31]

Similarly, a non-catalytic homolog of DNMT3A/B is DNMT3L that support binding to their target DNA. Additionally, proteins such as PCNA, UHRF1, DMAP1, DNMT3l, and HDACs function to steer DNMT proteins to their target regions. These proteins regulate the methyltransferase activity of the genome and epigenetic change that relates to DM. Characteristically, CGIs in normal cells are hypomethylated compared to the rest of the genome. Yet, CGIs in cancer cells of promoter regions adjacent to tumor suppressor genes are hypermethylated by the actions of DNMTs.[69] These enzymes are overexpressed in cancer and enhance tumorigenesis by means of decreasing tumor suppression. Likewise, hypermethylation of CGIs can interfere with binding of CTF protein, leading to loss of vital insulator regions and activation of the oncogene.[70] Hypomethylation of gene bodies and intergenic regions of DNA often correlates with the onset of cancer.[68]

5-methylcytosine (5-mC) is a repressive marker written de-novo by DNMT3A/B and is maintained by the action of DNMT1. Relatedly, 5-hydroxymethylcytosine (5-hmC), an DNA pyrimidine nitrogen base, can potentially switch a gene on or off. 5-hmC is enriched near promoter regions and can interact with methyl-CpG-binding protein 2 (MeCP2). The expression of MeCP2 protein is demonstrated at increasing levels during brain development, and 5-hmC plays a role in its activation. Furthermore, demethylation of CGIs positively regulates gene activation and CTF insulator functions by allowing binding of TFs and CCCTC-binding factor (CTCF), an 11-zinc-finger DNA binding protein. Likewise, 5-mC can be converted to 5-hmC via an enzymatic process involving the Ten-Eleven-Translocation (TET) family of proteins. Recent research identified 5hmC being associated with malignant transformation of *KRAS* in pancreatic cells after deactivation of p53. This process is a frequent clinically observed feature of PC patients[71]

Currently, there are two types of small molecule DNMT inhibitors being reported (i.e., nucleoside and non-nucleoside). Epigenetic inhibitors centered on nucleoside analogs 5-azacytidine and 5-aza-2′-dC are in Phase I-III clinical trials for several human diseases. To this point, DNMT inhibitors, azacytidine and decitabine, have shown promising efficacy and have received FDA approval in the treatment of myelodysplastic syndrome while not in solid tumors.[72] Up to now, studies have shown 5-aza-2′-dC reduced PC cell proliferation and initiated cell cycle arrest in an in vitro model. [73, 74]

The methyl-CpG-binding domain proteins (MBDs) family of proteins can read and identify de-novo-methylation patterns occurring during embryogenic developments. Besides, methyl-CpG binding domain protein-1 (MBD1) is frequently overexpressed in PC in a study by Liu et al.[75] Normally, MeCP2, MBD1, and MBD2/4 readers can create and preserve regions of transcriptionally inactive chromatin by recruiting corepressor proteins such as DNMT1, and histone deacetylases (HDAC1 and HDAC2). [75] Relatedly, Zhang et al. observed that MBD1 gene suppression reduces the antioxidant response and ARE target genes through epigenetic regulation of Kelch like ECH associated protein 1 (KEAP1). [76] Epigenetic events are ubiquitously distributed across normal and cancer cells. Therefore, a better understanding of the specific mechanisms such as DM and its underlying changes of GE in PDAC, is necessary for anticancer treatment.[77]

1.8. DNA methylation analysis techniques

The three most common DNA methylation analysis techniques are (1) digestion of genomic DNA with methylation-sensitive restriction enzymes; (2) affinity-based enrichment of methylated DNA fragment; and (3) bisulfite sequencing (BS) method. BS is a sodium bisulfite conversion technique that provides quantitative DM level with single-base resolution.[78] This treatment of genomic DNA converts unmethylated cytosine to uracil and then uracil becomes

thymidine in subsequent PCR amplification and sequencing. 5mC is resistant to such conversion process and therefore, it can be detected from unmethylated cytosine.[79]

Furthermore, the Infinium methylation 450k microarray is a cost-effective, high-throughput method for detecting DNA methylation in many human samples.[80] It involves bisulfite treatment of genomic DNA and subsequent hybridization to over 450 000 CpG sites throughout the genome. Overall coverage of this platform targets gene regions including promoters, 5′-untranslated regions, the first exons, gene bodies and 3′-untranslated regions, the first exons, gene bodies and 3′-untranslated regions. The TCGA consortium used this platform to profile >7500 samples from over 200 different cancer types.[81] Our analysis included publicly available PC datasets (Infinium methylation 450k microarray) including DM and GE, downloaded from GDC portal using computational methodologies in R software. We aimed at analysis of hypo and hypermethylation mark in significantly expressed genes associated in pancreatic ductal adenocarcinoma (PDAC).[82] This vast wealth of data provided by TCGA network of researchers present us with a unique opportunity to define and understand molecular mechanism associated with differentially methylation and GE pattern associated with PDAC.[83]

2. METHODS AND PROCESSING OF DATA

Publicly available genome datasets were accessed through the NCI Genomic Data Commons (GDC) data portal.[82.] We performed bioinformatic analysis of DM and GE datasets of PDAC by downloading from TCGA database in R statistical software. [81, 84]

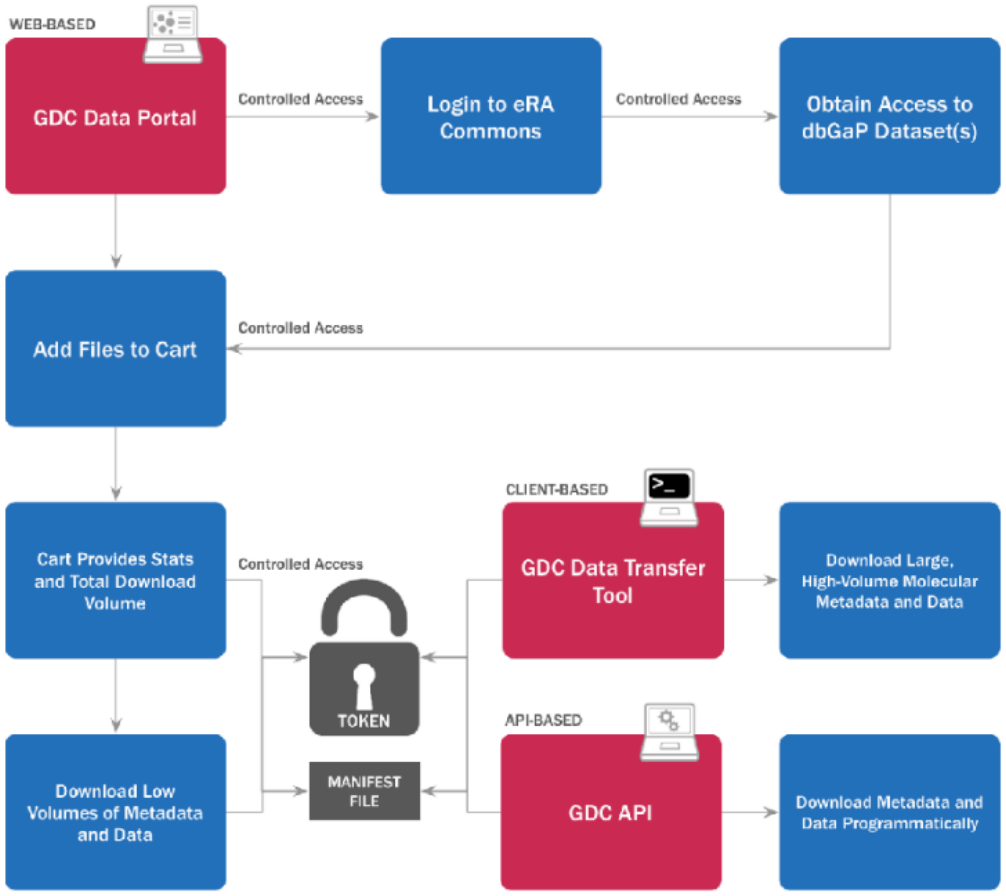

Figure 1: Schema of GDC pipelines and method of programmatic access.[85]

In this study we downloaded PDAC datasets (178 methylation and 164 from gene expression) using TCGAbiolinks package, from the TCGA portal. Next, raw and/or preprocessed DM, gene expression (GE) of pancreatic ductal adenocarcinoma (PDAC) was saved as SummarizedExperiment (SAE) object for analysis.[86] Generally, in many computational genomics analysis workflows a subset of data matrix is applied prior to analysis to exclude poor quality

samples or subset the rows of the matrix to select the most variable features. A subset of data structure was performed with SummarizedExperiment package in R software for storing one or more matrix-like assays along with associated row (e.g., genes, transcripts, exons, CpG probe) and column data (data frame).[87] Furthermore, these objects facilitate the storage and analysis of high-throughput genomic data generated from technologies such as array-based data. Thus, all our data was saved as SAE object for further analysis.

All bioinformatic analysis in this study were performed in the R programming environment (version 4.0.3) and the packages was downloaded from Bioconductor (version: 3.11). The differential methylated CpG (DMC) sites were calculated using the beta-values (methylation values ranging from 0.0 to 1.0) to compare the two means. DEGs were calculated using TCGAbiolinks and Limma package in R. LIMMA is a r package for the analysis of gene expression microarray data, especially the use of linear models for analyzing designed experiments and the assessment of differential expression.[88] Last, a multi-omics experiment of PDAC sample (DM and GE data) was conducted using enhancer linking by methylation/expression relationships (ELMER) package in R programming language. ELMER package detects tumor-specific changes in DNA methylation within distal enhancers, and link these enhancers to target genes in the same sample set.[89]

3. GENE EXPRESSION ANALYSIS

In this experiment, we have analyzed and processed RNA seq (GE) data downloaded from GDC to identify genes associated with methylation sites. Prior to performing DE analysis, we removed genes with low counts and then convert ran-seq data with function of limma. The DE analysis of genes is a method of taking normalized read count data and performing statistical analysis to determine quantitative changes in expression levels between experimental groups and a statistical test is done to decide if, for a given gene, an observed difference in read counts is significant.[90] We downloaded and processed RNA sequence data from GDC from the built-in functions of TCGAbiolinks library, edgeR and limma package. First, we saved the dataset as a single r object (r data single, or rds file). We selected clinical feature of this data to use as class for grouping the samples, normal (NT) versus tumor (TP) with SAE functions. Next, a design matrix was created by functions of limma that indicated the conditions to be compared in DE analysis (NT vs TP).

GDC functions:	Description
GDC projects:	TCGA-PAAD
Data category:	Transcriptome profiling, Copy number variation, DNA methylation, Gene expression.
Data type:	DNA methylation, Gene expression quantification, miRA expression quantification.
Workflow type:	DNA methylation, HTSeq-counts, HTSeq-FPKM-UQ.
Platform	Illumina Human Methylation, IlluminaHiseq_RNASeqV2.

Figure 2: GDC functions and data descriptions.

Parsing information from large data requires normalization method to reduce batch effect or technical bias.[91] Normalization allows accurate estimation and detection of differential expression (DE).[92] Therefore, we performed normalization method to remove of genes with low

counts to reduce batch effects and technical variation of trimmed mean of M-values (TMM). Trimmed mean of M-values (TMM) is a normalization method that is based on the hypothesis that most genes are not differentially expressed (DE).[93] For each sample, the TMM factor is calculated when one sample is considered as a reference sample and the other is as a test sample. Besides, for each test sample, TMM is calculated by weighted mean of log ratios between test sample and the reference. Due to the low DE hypothesis, the TMM should be close to 1. If it does not, TMM value will provide an estimate of the correction factor that must be applied to the library sizes. This method of normalization is implemented in the edgeR package as the default normalization method. The data generated in the analysis identified differential expressed genes (DEGs) and was sorted by p-value to assess their significance levels.[94] A heatmap plot was generated with the heatmap.2 package to visualized significant DEGs (figure:4). The samples highlighted are TN samples (red), TP samples (black), and the genes (green) and were narrowed down by the analysis function of limma.

3.1. Enrichment and pathway analysis of differential expressed genes

To understand the biological implication of expression data is enrichment analysis. This approach helps us to identify if the DEG are associated with a certain biological process or underlying molecular function of PC.[88] Gene ontology (GO) analysis is a technique for annotation of genes and gene sets, with biological significance of high-throughput genome or transcriptome data. The Kyoto Encyclopedia of Genes and Genomes (KEGG) is a data-bank for exploration of gene functions and integrated pathways.[95-96] KEGG analysis of PDAC data offers molecular identification of genes, proteins and signaling pathways.

3.2. Results of gene expression analysis

For RNA-sequence data, we performed DE analysis to locate quantitative changes of expression levels between NT vs TP groups. We identified from the tumor samples up to 10179 genes downregulated, 14125 upregulated, and 3721 not significant . Similarly, for normal samples a total of 7705 genes were downregulated, 12822 genes were upregulated, and 7498 were found not significant.[97] We identified the following genes as statistically significant by analysis from limma and plotted as heatmap plot (figure: 4). The gene expression values were extracted from the expression profile for each dataset. A bidirectional hierarchical clustering heatmap was constructed using heatmap.2 package of R language for DEGs in every dataset. Furthermore, the hierarchical clustering was conducted by limiting the analysis up of 36 common DEGs obtained from TCGA datasets. We used heatmap.2 function in gplots package of R to visualize a heat map. For our heatmap.2 function the expression value of gene is in the row and the sample is in the column. After normalizing the value of row, clustering settings are specified via distfun (method = euclidean) and hclustfun (method = complete) function. Last, the samples highlighted are primary solid tumor (black), solid tissues normal tissue (red) and DEGs in clusters are highlighted in green.

To understand biological implication of gene expression data, gene set enrichment (GSEA) analysis based on the functional annotation of the DEGs was conducted. GSEA technique determines if a priori defined set of genes exhibit statistically significant DE between two sample tissues (normal versus tumor) time points or conditions. Furthermore, this information can be obtained through computational interface with public online databases such as GO and KEGG.[98] Gene ontology (GO) analysis is a technique for annotation of genes and gene sets (figure:5). The annotation for gene ontology (GO) includes biological process (GO:BP), cellular component

(GO:CC), molecular function (GO:MF) and pathways (GO:P). Similarly, the Kyoto Encyclopedia of Genes and Genomes (KEGG) is a database for exploration of gene functions and integrated pathways. In summary, we aimed to examine pathway question that could answer important biological function pertaining to PDAC. The canonical pathways represent enrichment by the DEGs (differentially expressed genes) was analyzed with TCGABiolinks function (figure: 5). In more detail, this package allowed us multiple methods for analysis (e.g., differential expression analysis, identifying differentially methylated regions) and methods for visualization (e.g., volcano plot, gene ontology analysis, KEGG interface). In our analysis of DEGs, statistically significant pathways that have been detected in the DEGs list are specified in their p value corrected FDR (-Log) (colored bars) and the ratio of list genes found in each pathway over the total number of genes in this pathway (ratio, dotted red line).

Accordingly, for GO: BP of DEGs we found regulation of cell proliferation (n=106), cell adhesion (n=89), biological adhesion (n=89), cell cycle (n=84) and cell cycle process (n=66) upregulated and sensory perception of chemical stimulus (n=2) and smell (n=2), cognition (n=24), and neurological system process (n=45) downregulated. Next, for GO:CC pathway identification of DEGs we have found, calcium ion binding (n=125), voltage-gated sodium channel activity (n=10), sodium channel activity (n=12) upregulated and zinc ion binding (n=85), transition metal ion binding (n=94), ion binding (n=14), metal ion binding (n=114) downregulated.

ensembl_gene_id	external_gene_name	original_ensembl_gene_id	logFC	AveExpr	t	P.Value	adj.P.Val	B
ENSG00000122566	HNRNPA2B1	ENSG00000122566.19	9.285207	9.286217	511.6718	1.679612e-290	4.707114e-286	649.7812
ENSG00000129351	ILF3	ENSG00000129351.16	7.874172	7.876769	469.3676	1.182169e-283	1.656515e-279	634.9161
ENSG00000165119	HNRNPK	ENSG00000165119.17	8.924601	8.923584	452.9447	7.915346e-281	7.394252e-277	629.1701
ENSG00000136709	WDR33	ENSG00000136709.10	6.051179	6.053545	418.5534	1.455183e-274	8.579737e-271	614.6688
ENSG00000170144	HNRNPA3	ENSG00000170144.17	7.882657	7.882512	418.4375	1.530729e-274	8.579737e-271	615.3925
ENSG00000168066	SF1	ENSG00000168066.19	7.604044	7.607982	415.6481	5.193488e-274	2.425792e-270	614.1661
ENSG00000120948	TARDBP	ENSG00000120948.14	6.661606	6.664441	415.2595	6.161008e-274	2.466604e-270	613.6575
ENSG00000147140	NONO	ENSG00000147140.14	8.216673	8.217049	411.8062	2.831887e-273	9.920453e-270	612.7347
ENSG00000153187	HNRNPU	ENSG00000153187.15	8.386326	8.385176	410.3845	5.325970e-273	1.658448e-269	612.1777
ENSG00000121774	KHDRBS1	ENSG00000121774.16	7.212632	7.214327	393.8179	9.880362e-270	2.768971e-266	604.7739

Figure 3: List of top differentially expressed genes (DEGs) with their significance level and external gene names and id.

To further explore molecular pathways associated with PC, we analyzed GO:MF pathway of DEGs. Interestingly, we have identified extracellular matrix (n=59), extracellular region part (n=200), proteinaceous extracellular matrix (n=70), extracellular space (n=103), basement membrane (n=17) upregulated, plasma membrane part (n=304), intrinsic to plasma membrane(n=179), integral to plasma membrane(n=173), and plasma membrane (n=254) down regulated.

Finally, for GO:P study, we found that noradrenaline and adrenaline degradation (n=16), ethanol degradation-II (n=16), hepatic fibrosis/hepatic stellate activation(n=44), atherosclerosis signaling (n=39), granulocyte adhesion diapedesis (n=49), axonal guidance signaling (n=110) and agranulocyte adhesion and diapedesis (n=55) significantly upregulated. In contrast, we identified EIF2 signaling (n=7), mitotic roles of polo-like kinase (n=24) and LPS/IL mediated inhibition of RXR function (n=56) to be downregulated.

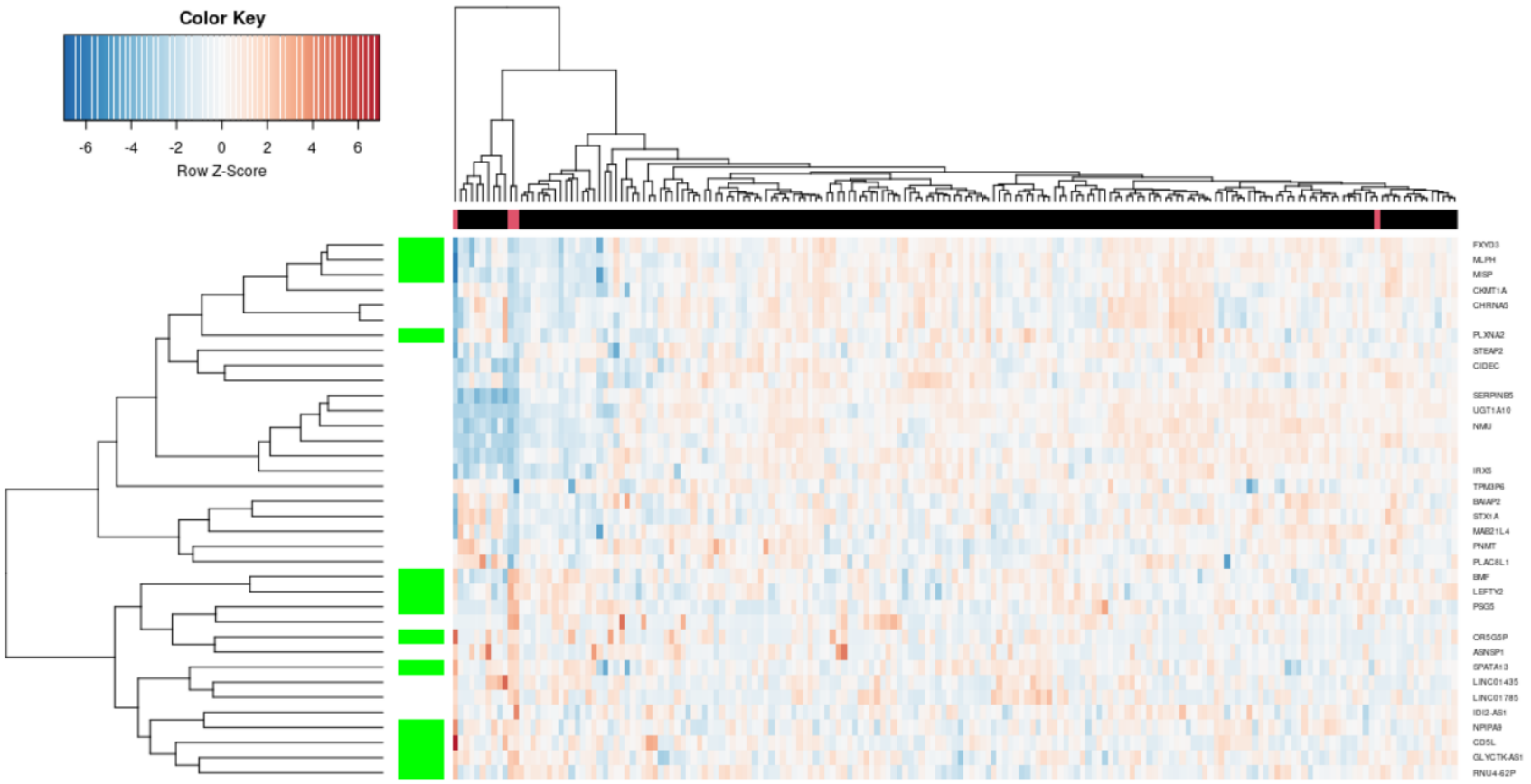

Figure 4: Heatmap of differentially expressed (DE) genes. The samples highlighted are primary solid tumor (black), solid tissues normal tissue (red). The DEGs in clusters are highlighted in green.

Figure 5: Canonical pathways significantly overrepresented (enriched) by the DEGs of PDAC. The most statistically significant canonical pathways identified in the DEGs of PDAC are ranked according to their p-value corrected FDR (-Log10) (colored bars) and the ratio of list genes found in each pathway (ratio, red line) over the 3224 PDAC (NT vs TP) in our pathway analysis with TCGABiolinks enrichment analysis function.

Our study of TCGA datasets (RNA-seq) characterized the DEGs and their corresponding signaling pathways through KEGG pathway analysis. We aimed to identify the most enriched pathways associated with upregulated and downregulated expressed genes of PC. This approach led to the identification of significantly upregulated genes (red) and their pathways. The top KEGG enriched pathways of PC were KRAS in the P13K-Akt, HER2/neu in Jak-STAT, p16 in VEGF, E2F in P53 and lastly, BRCA/Rad51 in TGF-β pathways. The remaining (green) did not show upregulation are PI3K, RacGEF, Rac, NFkB, IKK, Bcl-xl, PKB/Alt, Bad, CASP9, MEK, ERK, JNK, Cdb42Rac, RalBP1, RAL PLD1 genes of P13-Akt signaling pathway. For VEGF signaling pathway the following genes were downregulated are CDK4/6, Rb, E2F and CyclinD1. Similarly, in P53 pathway p21, Bax, p48, Bak, POLK, GADD45 and TGFα, EGF of ErbB pathway. In Jak-STAT signaling pathway, EGFR, PI3K, Jak1,STAT3,STAT1,PKB/Akt1, mTOR, Bcl-xl, NFkB, S6K genes were identified and for TGFβ pathway, TGFβRI, TGFβRII, Smad2/3 genes was identified.

3.3. Discussion of gene expression analysis

For RNA-sequence data, we performed DE analysis to locate quantitative changes of expression levels between NT vs TP groups. As pancreatic cancer progress, more and more genes are differentially expressed. Thus, we identified from the tumor samples up to 10179 genes downregulated, 14125 upregulated, and 3721 not significant from the decision test (DT). Similarly, for normal samples a total of 7705 genes were downregulated, 12822 genes were upregulated, and 7498 were found not significant. Sun et al. identified a total of 2566 DEGs, including 848 upregulated genes and 1718 downregulated ones, (between 178 pancreatic cancer samples and 4 normal samples) from the RNA-seq data deposited in TCGA.[21] Similarly, Tu et al. shown 2060 DEGs associated with PC. The fit function from limma identified the following genes as

statistically significant and plotted them as heatmap (figure: 4) to show significant genes. Hierarchical clusters of DEGs were visualized as heatmap (figure: 4), sorted according to their p value corrected FDR (-Log) (colored bars) and the ratio of list genes found in each pathway over the total number of genes in that pathway (Ratio, red line). A total of 3224 DEGs were examined for pathway enrichment analysis (figure: 5).

Several studies have revealed the loss of appropriate cell cycle regulation leads to genomic instability and play a role in the etiology of spontaneous cancers.[99] We identified regulation of cell proliferation cell adhesion, biological adhesion, cell cycle and cell cycle process upregulated in PDAC samples. He et al. observed, upregulation of DEGs enriched in digestion, lipid digestion and proteolysis. Similarly, in our analysis of DEGs we identified downregulated BP in sensory perception of chemical stimulus and smell, cognition, and neurological system process. Atay et al. identified extracellular structure constituents, collagen binding and integrin binding in their analysis.[100] In addition, GO:BP analysis DEGs of PC was found to be enriched in immune response, cell growth and maintenance, protein metabolism in a study by Tu et al.[101] Significantly enriched GO:BP was identified by Wu et al. include extracellular matrix organization, cell adhesion, collagen catabolic process, extracellular matrix disassembly, hemidesmosome assembly, proteolysis, and cell migration.[102]

Next, for GO:CC upregulated pathway identification, we found calcium ion binding, voltage-gated sodium channel activity, sodium channel activity and Tu et al. identified significantly DEGs of PC, enriched in extracellular matrix/region/space, exosomes, and plasma membrane.[101] Likewise, we found in GO:CC the following downregulated DEGs pathways, zinc ion binding, transition metal ion binding, ion binding, and metal ion binding.

For GO: Molecular Function (MF) pathway we identified the following: extracellular matrix, extracellular region part, proteinaceous extracellular matrix, extracellular space, basement membrane upregulated. Additionally, plasma membrane part, intrinsic to plasma membrane, integral to plasma membrane, plasma membrane down regulated. Mishra at al. identified receptor activity, transmembrane receptor signaling activity, signal transducer activity significant in GO: MF analysis.[103]

Our study detected upregulated genes in noradrenaline and adrenaline degradation, ethanol degradation II, hepatic fibrosis/hepatic stellate activation, atherosclerosis signaling, granulocyte adhesion and diapedesis, axonal guidance signaling, mitotic roles of polo-like kinase, agranulocyte adhesion and diapedesis in GO: Pathway analysis. He et al. showed upregulated DEGs were enriched in pancreatic secretion, protein digestion and absorption. Last, we also found downregulated DEGs in GO:P analysis is EIF2 signaling, and LPS/IL-1 mediated inhibition of RXR functions and Khan et al. identified cellular senescence, chronic myeloid leukemia, focal adhesion and fluid shear stress and atherosclerosis associated with PC.[104]

In our study we examined DEGs and their corresponding signaling paths through a programmatic interface with the KEGG database (figure: 6). Different signaling pathways contributing to various biological processes during tumorigenesis progression of pancreas have been identified.[105] We have analyzed enhanced pathways associated with upregulated (red) and downregulated (green) DEGs of tumorigenesis of pancreas. Accordingly, significantly upregulated genes and their pathways detected are KRAS in the P13K-Akt, HER2/neu in Jak-STAT, p16 in VEGF, E2F in P53 and lastly BRACA/Rad51 in TGF-β pathways.The remaining DEGs not showing upregulation are PI3K, RacGEF, Rac, NFkB, IKK, Bcl-xl, PKB/Alt, Bad, CASP9, MEK, ERK, JNK, Cdb42Rac, and RalBP1, RAL, PLD1 genes of P13-Akt signaling pathway. For VEGF

signaling pathway the following genes were discovered downregulated are CDK4/6, Rb, E2F and CyclinD1.

Correspondingly, in P53 pathway p21, Bax, p48, Bak, POLK, GADD45 and TGFα, EGF in ErbB signaling pathway and TGFβ, TGFβRI, TGFβRII, Smad2/3 genes of TGFβ pathways was downregulated. Last, for Jakt-STAT signaling pathway EGFR, PI3K, Jak1, STAT3, STAT1, PKB/Akt1, mTOR, Bcl-xl, NFkB, and S6K genes were identified not significantly upregulated. KEGG pathway analysis by Wu et al. showed that DEGs participated in PI3K-Akt signaling pathway, pathways in cancer and pathways related to cellular dissociation from *in situ*, including ECM-receptor interaction, focal adhesion, and protein digestion and absorption.[102]

Additionally, Grimont et al. found participation of ErbB signaling pathway and Javle et al., detected TGF-beta signaling pathway in transmembrane signal transduction of PDAC. Furthermore, both of these studies have demonstrated the transmembrane signaling pathways transfer their signals to intracellular pathways (such as MAPK signaling pathway, PI3K-Akt signaling pathway, p53 signaling pathway, and VEGF signaling pathway).[106] Based on KEGG database, we were able to integrate pathway map which identified functional pathway based on the biological network of PDAC cells. Similarly, bioinformatic features of GO terms allowed us to determine genes and cell processes distributed across PDAC disease. To that end, enrichment methods can provide accurate depiction of underlying biological processes involved in tumorigenesis progression of pancreas. Decoding how these pathways interact with each other in PC can reveal unique features toward precision medicine and biomarker discovery of this disease.

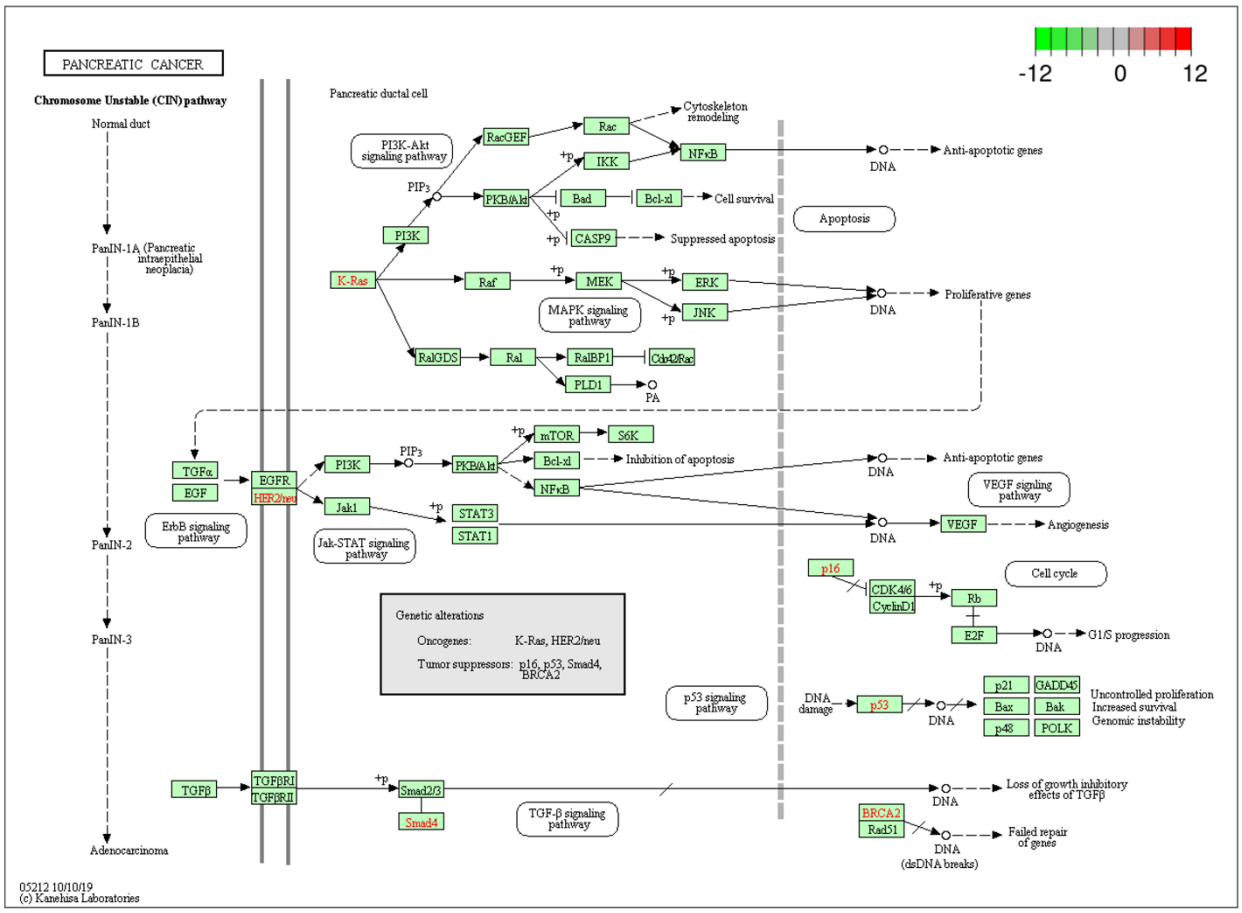

Figure 6: KEGG pathway analysis of DEGs of PDAC and their pathways are highlighted in red.

4. DNA METHYLATION ANALYSIS

Methylation data of PDAC was downloaded and processed by TCGAbiolinks packages. First, a preprocessing step was done to exclude NA values and then we tested mean of DM of each patient in groups (TP vs NT). A Summarized Experiment (SE) object was created with DM data, and the functions of group column and subgroup column, having two columns from the sample data matrix of the SE object was available from column data function. The nucleotide base or regions with different methylated proportion across samples are defined as different methylated CpG sites (DMCs) and different methylated regions (DMRs). Once we had pre-processed our data, we studied the mean DM of each patient in a group normal (TP) vs tumor (NT) with mean Methylation function of TCGABiolinks in R.[82]

Next, we examined DMCs between groups (NT vs TP) and analyzed them using the TCGAanalyze-DMC function. The DM data (level 3) was processed by taking beta-values in a scale ranging from 0.0 (unmethylated probe) up to1.0 (methylated probe). Accordingly, for searching DMRs, a mean beta-values of each group and probe were calculated. Next, differential expression between groups (NT vs TP) were calculated with minimum absolute beta-values difference of 0.15 and an adjusted p-value of less than 0.05. After this analysis, a volcano plot (x-axis: difference of mean DNA methylation, y-axis: statistical significance) was created (figure: 8) to identify hypo and hyper methylated CpG sites.

At this step, we identified genomic annotation of methylation datasets by accessing Ensemble and UCSC database functions from ChiPeakAnno and annotatr packages in R.[107, 108] Genomic annotations include 1-5Kb upstream of the TSS, the promoter (< 1Kb upstream of the TSS), 5’UTR, first exons, exons, introns, CDS, 3’UTR, and intergenic regions. The proportion of

binding sites were calculated by function from Genomic Features and data from TxDb. and org.eg.db packages. [107]

4.1. Differential methylated CpG (DMCs) analysis

A multi assay experiment (MAE) object was created from the features of DM and GE datasets, with input function of ELMER packages (r script utilized is listed in appendices section). MAE is a data structure method for integrating and manipulation of multi assay genomic datasets (DM and GE). The three general components of MAE are: (a) colData, that provides data about the patients, cell lines, or other biological units, with one row per unit and one column per variable, (b) experiments, that are a list of assay datasets with one column per observation and (c) sampleMap, which links a single table of patient data (colData) to a list of experiments.[110] First, probe data stored as GRanges object returned the coordinates, clinical data, molecular subtype information, names of each probe and DM array with MAE functions. [111] Similarly, gene information containing the coordinates of each gene, gene id, gene symbol and gene isoform, from annotations were retrieved from the biomaRt package in R. Last, for analysis of DM and GE patterns of PDAC, we combined the information above and created a MAE object in R with ELMER package.

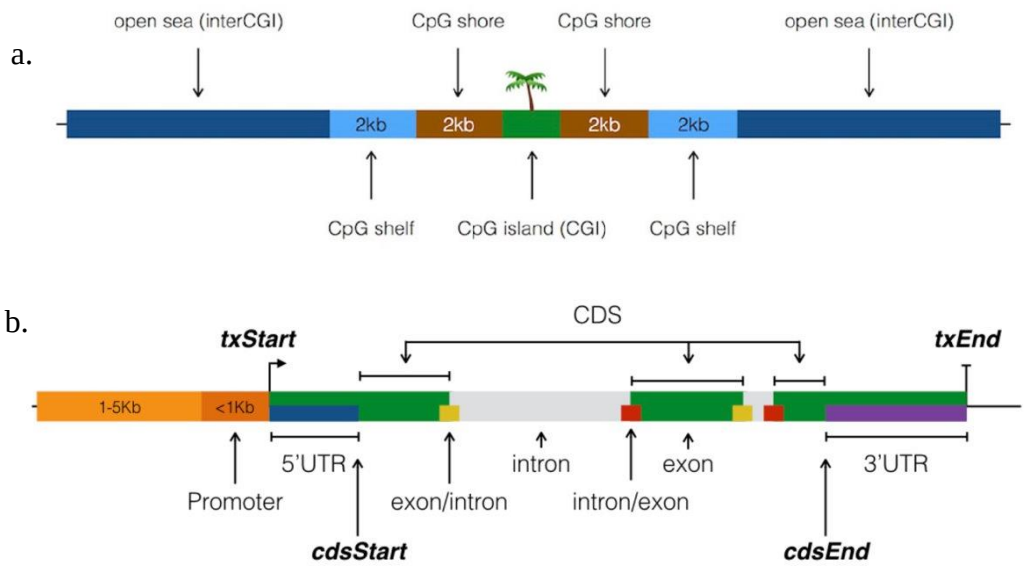

Figure 7: (a) Schematic representation of CpG islands (b) Schematic representation of genomic annotation.[109]

We analyzed methylation status and expression of contiguous genes for identification of hypo and hyper methylated targets.[89] The processed DM data were calculated as M/(M+U)=beta values, where M signifies the methylated allele intensity and U the unmethylated allele intensity. Beta values range from 0 to 1 and signify the fraction of methylated alleles at each CpG site for tumors. Correspondingly, beta values close to 0 indicates low level of DM, and beta values near to 1 indicates higher level of DM.[80]

DMCs sites were evaluated to recognize the differences of DNA methylation level for each probe and their significance values. To compare DM level, the samples of each group (group 1 and group 2) are ranked by their DNA methylation beta values for a given probe. The samples in the lower quantile (20% samples with lowest methylation levels) of each group were used to identify if the probe is hypomethylated in group 1 compared to group 2. Likewise, for identification of hypermethylated probes, we used upper quantile (20% samples with high methylation levels) of each group to detect hypermethylated probes.

Samples of each group (NT and TP) were ranked by their DM beta values. In addition, samples in the lower quintile of each group were used to identify hypomethylated probes in group 1 (NT) matched to group 2 (TP). Similarly, samples in upper quantile were used to identify hypermethylated probes in group 1 (NT) matched to group 2 (TP). The selection criteria for DMC probes was set to p-value less than 0.05.[112]

Furthermore, DMCs were examined for putative target genes (closest 10 upstream genes and the closest 10 downstream genes) by inverse correlation between methylation of the probe and expression pattern of the gene. Next, a Mann-Whitney U test was conducted to determine that overall gene 1 expression (i.e. group M is greater than or equal to that in group U) for each candidate methylated probe-gene pair.[113]Furthermore, the raw p-value (Pr) was corrected for multiple hypothesis testing using a permutation approach. The gene in the pair remained constant, and then random methylation probes were chosen to perform the "one-tailed U test", to identify permutation p-values (Pp). Moreover, probes were being identified by null-model from the same set as they were being tested, and an empirical p-value (Pe) was calculated by a pseudo-count of 1 to obtain significant results. [114]

4.2. Results of DNA methylation analysis

The nucleotide base or regions with different methylated proportions across groups are defined as DMCs and DMRs.[115] Due to their importance in regulation of gene expression and onset of diseases, DM analysis is highly informative and are considered biomarkers. In our analysis of DMC (figure: 4) we observed 59 hyper-methylated markers and 55 hypo-methylated CpG sites when comparing tumor (TP) to normal tissue (NT) using the TCGAbiolinks-DMC function.

In our experiment, we annotated our dataset by interfacing with AnnotationHub and annotator package. We annotated genomic regions (figure: 9), that include 1-5Kb upstream of the

TSS, the promoter (< 1Kb upstream of the TSS), 5'UTR, first exons, exons, introns, 3'UTR, and intergenic regions. We identified 0.42% at 1-2kb, 0.72% in region <=1kb, 5.6% in regions 2-3kb away of promoters, 0.18% at 5'UTR, 5.27% at 3'UTR, 0.15% at 1st exon, another exon's having 3.32%, 13.94% at 1st intron, 28.45% at other intron, 1.29% at downstream <=300 and finally, 40.65% in distal intergenic regions.

We observed significant probes associated with differentially methylated genes. For methylated DMCs, nearest 10 upstream genes and 10 downstream genes have been tested for inverse correlation between methylation of the probe and expression of the gene by number of flanking genes constraint. A probe-gene spacing has been defined as the distance between probe to the transcription start site specified by gene level annotations. A 20% minimum subgroup fraction was set in the get.pair function of ELMER package in R. Accordingly, each probe-gene pair the samples from both groups were separated into two groups upper and lower methylation levels. Thus, the M group, which is comprised of upper methylation quintile (the 20% of samples with the highest methylation at enhancer probe), and the U group, with having lowest methylation quintile (the 20% of samples with the lowest methylation at enhancer probe).

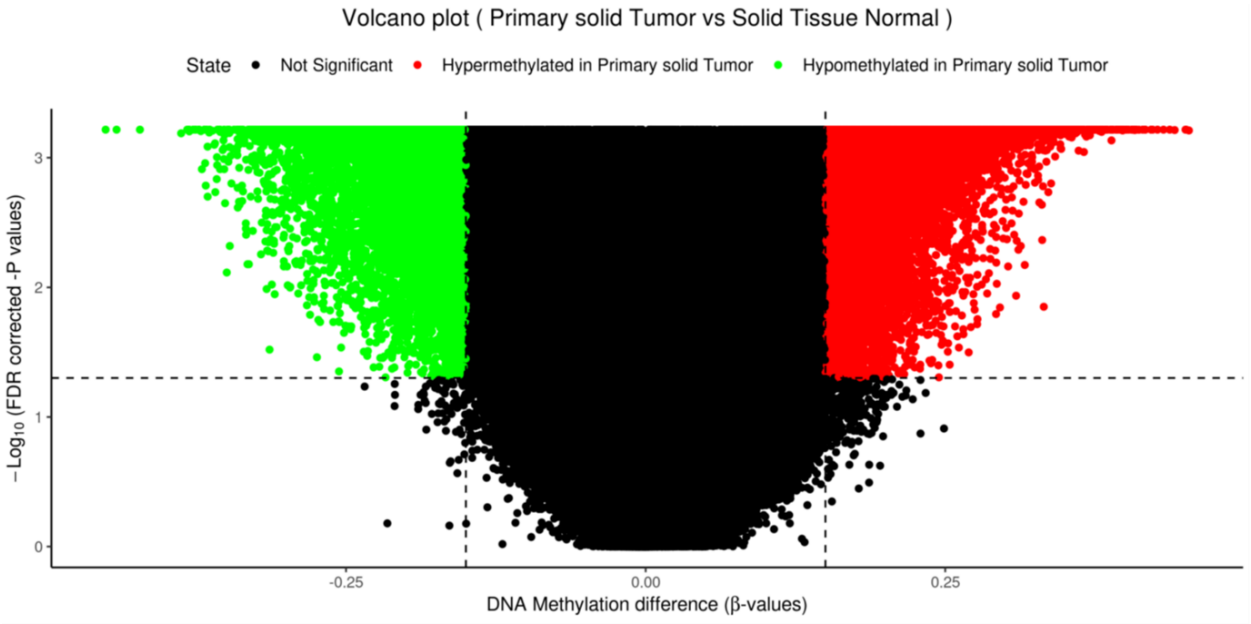

Figure 8: Volcano plot of DNA methylation in tumor tissues compared to normal tissue. Significance associations are indicated in hypomethylated (green) and hypermethylated (red).

In our hyper analysis (figure: 10) AGPAT6, ANK1, KAT6A, AP3M2, PLAT, IKBKB, POLB, SLC20A2, VDAC3m, SMIM19, THAP1, RNF170, HOOK3, FNTA and HGSNAT genes were in equal measure methylated and over-expressed. Likewise, genes that were under-expressed but hyper methylated are as follows NKx6-3, CHRNB3, CHRNB3, CHRNA6, POMK and DKK4 genes. Similarly, for our hypo methylated DMCs analysis (figure: 11) we identified PLAT, VDAC3, and IKBKB, gene equally methylated and over expressed. In summary, we observed that methylation and gene expression levels varied among DMCs that were identified. In our list, low expressed genes were RP11-231D20.2, RP11-589C21.6, RPL5P23, DKK4, RP11-503E24.2, R1007J8.1, RP11-412B14.1 and 2, CHRNA6. Conversely, highly expressed genes with low methylation levels were AP3M2, POLB, SLC20A2, SMIM19 and RNF170 during tumorigenesis.

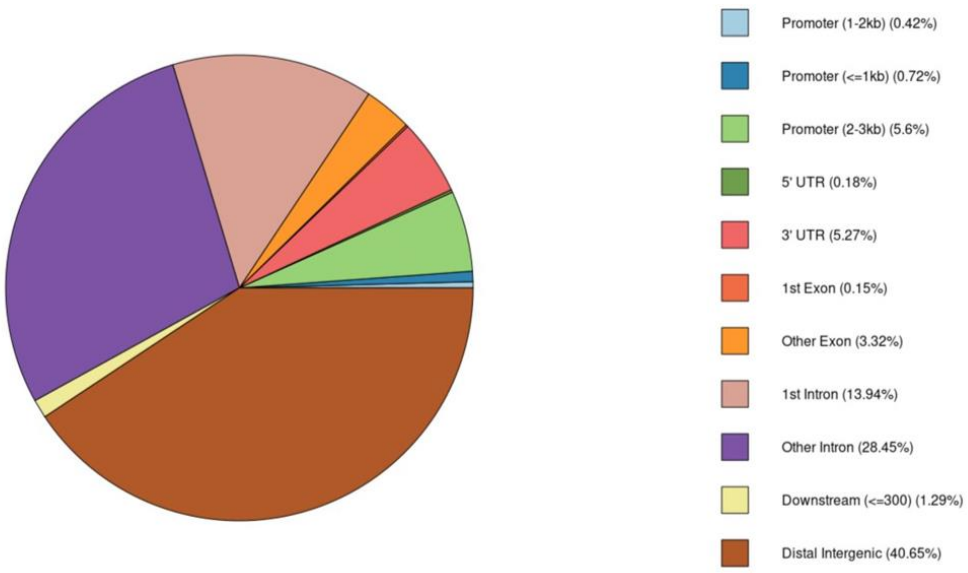

Figure 9: Pie chart result of our genomic annotation of DNA methylation datasets.

4.3. Discussion of DNA methylation analysis

The nucleotide base or regions with different methylated proportions across samples are defined as DMCs. Accordingly, DM analysis is considered very informative and is considered biomarkers given their importance in regulation of gene expression and inception of pancreatic tumorigenesis. In our analysis of DMCs, we observed from 400 samples, 59 hyper-methylated markers and 55 hypo-methylated CpG sites when comparing tumor (TP) to normal tissue (NT) using the TCGAbiolinks-DMC function.

Our tests show prevalence of hypermethylation compared to hypomethylation sites (figure: 8). We annotated genomic regions (figure: 9), that include 1-5Kb upstream of the TSS, the promoter (< 1Kb upstream of the TSS), 5'UTR, first exons, exons, introns, 3'UTR, and intergenic regions. We identified .0.42% at 1-2kb, 0.72% in region <=1kb, 5.6% in regions 2-3kb away of promoters, 0.18% at 5'UTR, 5.27% at 3'UTR, 0.15% at 1st exon, another exon's having 3.32%, 13.94% at 1st intron, 28.45% at other intron, 1.29% at downstream <=300 and finally, 40.65% in distal intergenic regions.

We observed significant differentially methylation level between genes. For each DMCs, nearest 10 upstream genes and 10 downstream genes have been tested for inverse correlation between methylation of the probe and expression of the gene by numFlankingGenes constraint. [116] A probe-gene spacing has been defined as the distance between probe to the transcription start site specified by gene level annotations. Accordingly, each probe-gene pair the samples from both groups were separated into two groups upper and lower methylation levels. Thus, the M group, which is comprised of upper methylation quintile (the 20% of samples with the highest methylation at enhancer probe), and the U group, with having lowest methylation quintile (the 20% of samples with the lowest methylation at enhancer probe).

For our hyper analysis AGPAT6, ANK1, KAT6A, AP3M2, PLAT, IKBKB, POLB, SLC20A2, VDAC3m SMIM19, THAP1, RNF170, HOOK3, FNTA and HGSNAT genes were in equal measure hyper methylated and over-expressed.[117] Likewise, genes that were under-expressed but hyper methylated are as follows NKx6-3CHRNB3, CHRNB3, CHRNA6, POMK and DKK4 genes. Correspondingly, in a study by Sun et al. reported five (05) hypomethylated/overexpressed and twenty six (26) hypermethylated/under expressed differential methylated expressed genes.[21]

In our hypo methylated DMCs analysis we identified PLAT gene equally methylated and over expressed.[118] The methylation and gene expression levels varied among DMCs that were identified. In our list, hypo methylated and the low expressed genes were RP11-231D20.2, RP11-589C21.6, RPL5P23, DKK4, RP11-503E24.2, R1007J8.1, RP11-412B14.1 and 2, CHRNA6.[119] Similarly, the following genes had high expression level and reduced methylation state are AP3M2, VDAC3, POLB, IKBKB, SLC20A2, SMIM19 and RNF170 genes.[120]

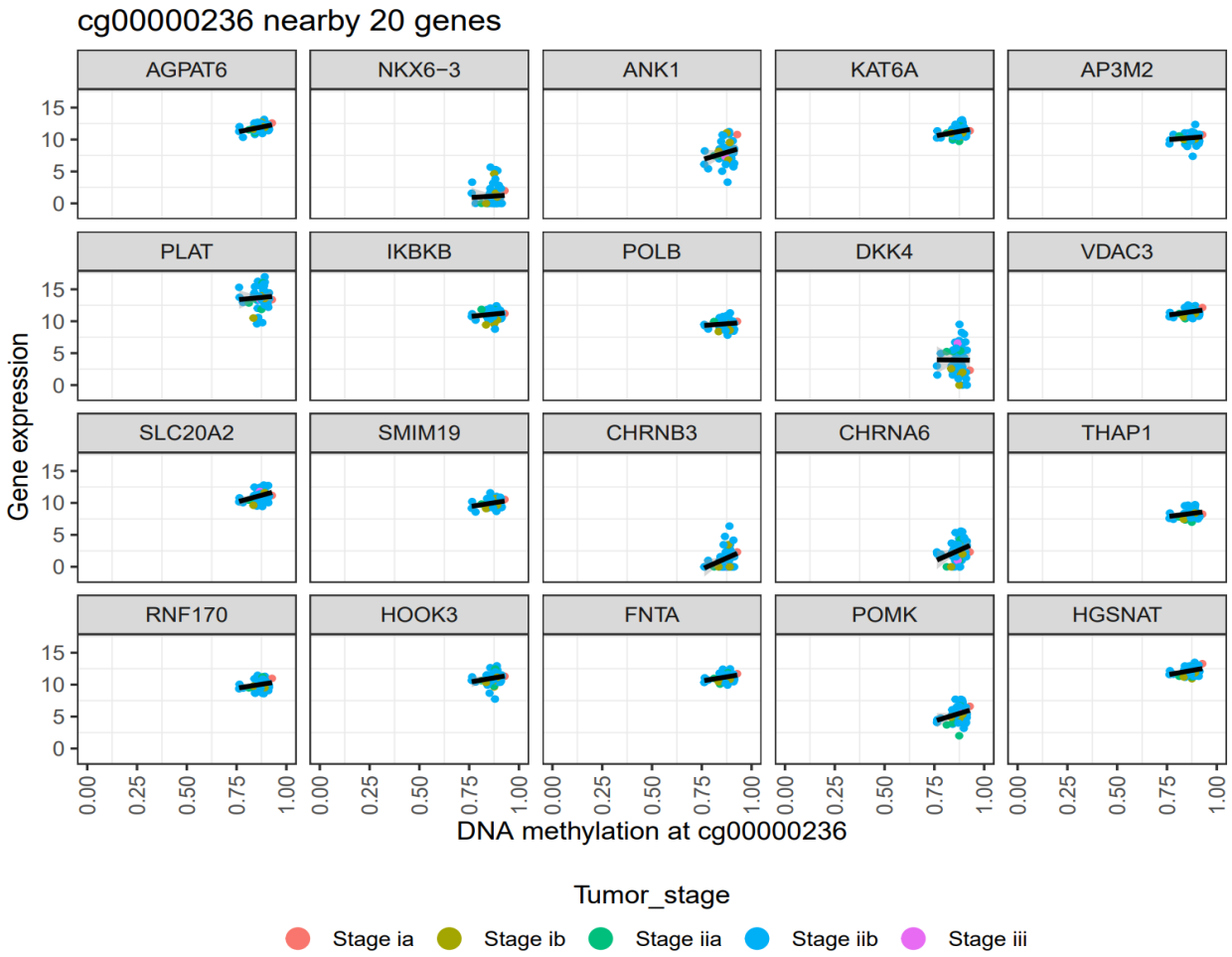

Figure 10: Scatter plot of hyper-methylation DMC analysis reveals the methylation and gene expression level of significant probe cg00000236 plotted against nearby 20 genes.

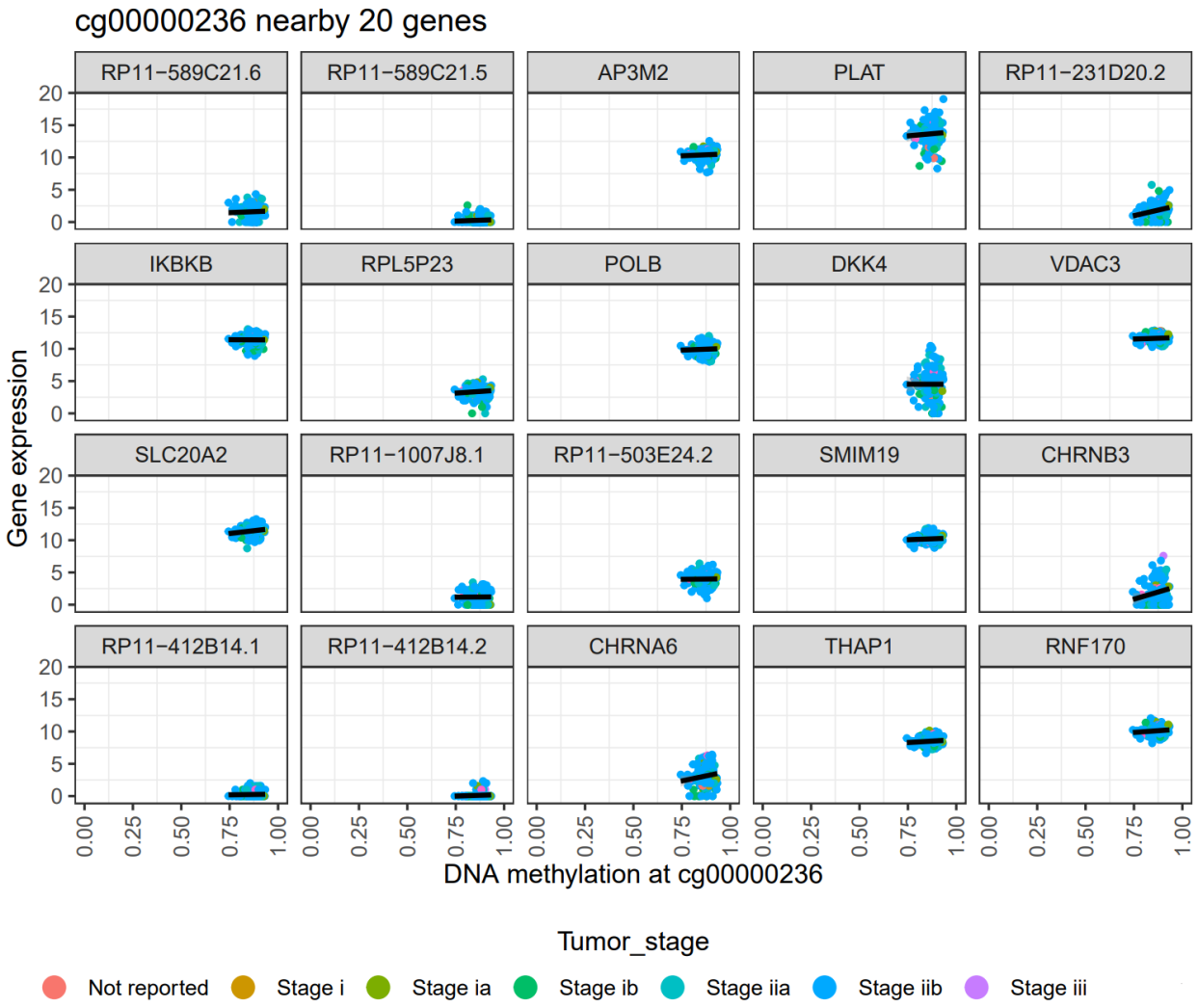

Figure 11: Scatter plot of hypo-methylation DMC analysis indicates the methylation and gene expression level of significant probe cg00000236 plotted against nearby 20 genes.

5. CONCLUSION

Cancer incidence and mortality trend of pancreatic carcinoma remains grim at present. In this capacity, early detection is a key factor in overall disease management. In-silico analysis of publicly available high-throughput genomic datasets can aid in decoding molecular basis of PDAC.[5, 121] Furthermore, examination of methylation markers and a systemic enquiry of epigenetic influence of DNA methylation (DM) on gene expression will provide a unique angle toward precision medicine.[79] The process of DM can be reversed and therefore could serve as a potential biomarker. This study will provide crucial insights toward identification of hypo and hyper methylated gene expression. Therefore, understanding methylation pattern and the complex process of tumorigenesis will offer new methods for detection, classification, and clinical insight into this disease.

REFERENCES

1. Rawla, P.; Sunkara, T.; Gaduputi, V., Epidemiology of Pancreatic Cancer: Global Trends, Etiology and Risk Factors. *World journal of oncology* **2019,** *10* (1), 10-27.

2. Grossberg, A. J.; Chu, L. C.; Deig, C. R.; Fishman, E. K.; Hwang, W. L.; Maitra, A.; Marks, D. L.; Mehta, A.; Nabavizadeh, N.; Simeone, D. M.; Weekes, C. D.; Thomas Jr, C. R., Multidisciplinary standards of care and recent progress in pancreatic ductal adenocarcinoma. *CA: A Cancer Journal for Clinicians* **2020,** *70* (5), 375-403.

3. Bray, F.; Ferlay, J.; Soerjomataram, I.; Siegel, R. L.; Torre, L. A.; Jemal, A., Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians* **2018,** *68* (6), 394-424.

4. Adamska, A.; Domenichini, A.; Falasca, M., Pancreatic Ductal Adenocarcinoma: Current and Evolving Therapies. *International journal of molecular sciences* **2017,** *18* (7), 1338.

5. Kourou, K.; Exarchos, T. P.; Exarchos, K. P.; Karamouzis, M. V.; Fotiadis, D. I., Machine learning applications in cancer prognosis and prediction. *Comput Struct Biotechnol J* **2015,** *13*, 8-17.

6. Slatko, B. E.; Gardner, A. F.; Ausubel, F. M., Overview of Next-Generation Sequencing Technologies. *Curr Protoc Mol Biol* **2018,** *122* (1), e59.

7. Chen, H.; Li, C.; Peng, X.; Zhou, Z.; Weinstein, J. N.; Liang, H., A Pan-Cancer Analysis of Enhancer Expression in Nearly 9000 Patient Samples. *Cell* **2018,** *173* (2), 386-399.e12.

8. Kwon, M. S.; Kim, Y.; Lee, S.; Namkung, J.; Yun, T.; Yi, S. G.; Han, S.; Kang, M.; Kim, S. W.; Jang, J. Y.; Park, T., Integrative analysis of multi-omics data for identifying multi-markers for diagnosing pancreatic cancer. *BMC Genomics* **2015,** *16 Suppl 9* (Suppl 9), S4.

9. Kwon, M. S.; Kim, Y.; Lee, S.; Namkung, J.; Yun, T.; Yi, S. G.; Han, S.; Kang, M.; Kim, S. W.; Jang, J. Y.; Park, T., Erratum to: Integrative analysis of multi-omics data for identifying multi-markers for diagnosing pancreatic cancer. *BMC Genomics* **2017,** *18* (1), 88.

10. Gibney, E. R.; Nolan, C. M., Epigenetics and gene expression. *Heredity* **2010,** *105* (1), 4-13.

11. Pope, S. D.; Medzhitov, R., Emerging Principles of Gene Expression Programs and Their Regulation. (1097-4164 (Electronic)).

12. Kawasaki, F.; Beraldi, D.; Hardisty, R. E.; McInroy, G. R.; van Delft, P.; Balasubramanian, S., Genome-wide mapping of 5-hydroxymethyluracil in the eukaryote parasite Leishmania. *Genome Biology* **2017,** *18* (1), 23.

13. Jin, B.; Li, Y.; Robertson, K. D., DNA methylation: superior or subordinate in the epigenetic hierarchy? *Genes & cancer* **2011,** *2* (6), 607-617.

14. Shi, J.; Xu, J.; Chen, Y. E.; Li, J. S.; Cui, Y.; Shen, L.; Li, J. J.; Li, W., The concurrence of DNA methylation and demethylation is associated with transcription regulation. *Nature Communications* **2021,** *12* (1), 5285.

15. Mohn, F.; Weber, M.; Rebhan, M.; Roloff, T. C.; Richter, J.; Stadler, M. B.; Bibel, M.; Schübeler, D., Lineage-specific polycomb targets and de novo DNA methylation define restriction and potential of neuronal progenitors. *Mol Cell* **2008,** *30* (6), 755-66.

16. Luger, K.; Dechassa, M. L.; Tremethick, D. J., New insights into nucleosome and chromatin structure: an ordered state or a disordered affair? *Nature Reviews Molecular Cell Biology* **2012,** *13* (7), 436-447.

17. Ramazi, S.; Allahverdi, A.; Zahiri, J., Evaluation of post-translational modifications in histone proteins: A review on histone modification defects in developmental and neurological disorders. *Journal of Biosciences* **2020,** *45* (1), 135.

18. Jones, P. A.; Baylin, S. B., The fundamental role of epigenetic events in cancer. *Nature Reviews Genetics* **2002,** *3* (6), 415-428.

19. van der Knaap, J. A.; Verrijzer, C. P., Undercover: gene control by metabolites and metabolic enzymes. *Genes Dev* **2016,** *30* (21), 2345-2369.

20. Wade, P. A., Methyl CpG-binding proteins and transcriptional repression. *Bioessays* **2001,** *23* (12), 1131-7.

21. Sun, H.; Xin, R.; Zheng, C.; Huang, G., Aberrantly DNA Methylated-Differentially Expressed Genes in Pancreatic Cancer Through an Integrated Bioinformatics Approach. *Frontiers in genetics* **2021,** *12*, 583568-583568.

22. Sharif, J.; Koseki, H., Hemimethylation: DNA's lasting odd couple. *Science* **2018,** *359* (6380), 1102-1103.

23. Li, E.; Zhang, Y., DNA methylation in mammals. *Cold Spring Harb Perspect Biol* **2014,** *6* (5), a019133.

24. Ehrlich, M., DNA hypomethylation in cancer cells. *Epigenomics* **2009,** *1* (2), 239-259.

25. Sun, S.; Zane, A.; Fulton, C.; Philipoom, J., Statistical and bioinformatic analysis of hemimethylation patterns in non-small cell lung cancer. *BMC Cancer* **2021,** *21* (1), 268.

26. Kenner, B.; Chari, S. T.; Kelsen, D.; Klimstra, D. S.; Pandol, S. J.; Rosenthal, M.; Rustgi, A. K.; Taylor, J. A.; Yala, A.; Abul-Husn, N.; Andersen, D. K.; Bernstein, D.; Brunak, S.; Canto, M. I.; Eldar, Y. C.; Fishman, E. K.; Fleshman, J.; Go, V. L. W.; Holt, J. M.; Field, B.; Goldberg, A.; Hoos, W.; Iacobuzio-Donahue, C.; Li, D.; Lidgard, G.; Maitra, A.; Matrisian, L. M.; Poblete, S.; Rothschild, L.; Sander, C.; Schwartz, L. H.; Shalit, U.; Srivastava, S.; Wolpin, B., Artificial Intelligence and Early Detection of Pancreatic Cancer: 2020 Summative Review. *Pancreas* **2021,** *50* (3), 251-279.

27. Moore, L. D.; Le, T.; Fan, G., DNA methylation and its basic function. *Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology* **2013,** *38* (1), 23-38.

28. Egger, G.; Liang, G.; Aparicio, A.; Jones, P. A., Epigenetics in human disease and prospects for epigenetic therapy. *Nature* **2004,** *429* (6990), 457-63.

29. Nissim, S.; Leshchiner, I.; Mancias, J. D.; Greenblatt, M. B.; Maertens, O.; Cassa, C. A.; Rosenfeld, J. A.; Cox, A. G.; Hedgepeth, J.; Wucherpfennig, J. I.; Kim, A. J.; Henderson, J. E.; Gonyo, P.; Brandt, A.; Lorimer, E.; Unger, B.; Prokop, J. W.; Heidel, J. R.; Wang, X. X.; Ukaegbu, C. I.; Jennings, B. C.; Paulo, J. A.; Gableske, S.; Fierke, C. A.; Getz, G.; Sunyaev, S. R.; Wade Harper, J.; Cichowski, K.; Kimmelman, A. C.; Houvras, Y.; Syngal, S.; Williams, C.; Goessling, W., Mutations in RABL3 alter KRAS prenylation and are associated with hereditary pancreatic cancer. *Nat Genet* **2019,** *51* (9), 1308-1314.

30. Pfeifer, G. P.; Kadam S Fau - Jin, S.-G.; Jin, S. G., 5-hydroxymethylcytosine and its potential roles in development and cancer. (1756-8935 (Print)).

31. Zhang, J. J.; Zhu, Y.; Wu, J. L.; Liang, W. B.; Zhu, R.; Xu, Z. K.; Du, Q.; Miao, Y., Association of increased DNA methyltransferase expression with carcinogenesis and poor prognosis in pancreatic ductal adenocarcinoma. *Clin Transl Oncol* **2012,** *14* (2), 116-24.

32. Villar, D.; Flicek, P.; Odom, D. T., Evolution of transcription factor binding in metazoans - mechanisms and functional implications. *Nature reviews. Genetics* **2014,** *15* (4), 221-233.

33. Dallas, M. R.; Chen, S. H.; Streppel, M. M.; Sharma, S.; Maitra, A.; Konstantopoulos, K., Sialofucosylated podocalyxin is a functional E- and L-selectin ligand expressed by metastatic pancreatic cancer cells. *Am J Physiol Cell Physiol* **2012,** *303* (6), C616-24.

34. Wong, B. S.; Shea, D. J.; Mistriotis, P.; Tuntithavornwat, S.; Law, R. A.; Bieber, J. M.; Zheng, L.; Konstantopoulos, K., A Direct Podocalyxin-Dynamin-2 Interaction Regulates Cytoskeletal Dynamics to Promote Migration and Metastasis in Pancreatic Cancer Cells. *Cancer research* **2019,** *79* (11), 2878-2891.

35. Rodenhiser, D.; Mann, M., Epigenetics and human disease: translating basic biology into clinical applications. *Canadian Medical Association Journal* **2006,** *174* (3), 341.

36. Corces, M. R.; Granja, J. M.; Shams, S.; Louie, B. H.; Seoane, J. A.; Zhou, W.; Silva, T. C.; Groeneveld, C.; Wong, C. K.; Cho, S. W.; Satpathy, A. T.; Mumbach, M. R.; Hoadley, K. A.; Robertson, A. G.; Sheffield, N. C.; Felau, I.; Castro, M. A. A.; Berman, B. P.; Staudt, L. M.; Zenklusen, J. C.; Laird, P. W.; Curtis, C.; Greenleaf, W. J.; Chang, H. Y., The chromatin accessibility landscape of primary human cancers. *Science* **2018,** *362* (6413).

37. Yen, C.-Y.; Huang, H.-W.; Shu, C.-W.; Hou, M.-F.; Yuan, S.-S. F.; Wang, H.-R.; Chang, Y.-T.; Farooqi, A. A.; Tang, J.-Y.; Chang, H.-W., DNA methylation, histone acetylation and methylation of epigenetic modifications as a therapeutic approach for cancers. *Cancer Letters* **2016,** *373* (2), 185-192.

38. Cervoni, N.; Szyf, M., Demethylase activity is directed by histone acetylation. *J Biol Chem* **2001,** *276* (44), 40778-87.

39. Kang, X.; Lin, Z.; Xu, M.; Pan, J.; Wang, Z.-W., Deciphering role of FGFR signalling pathway in pancreatic cancer. *Cell proliferation* **2019,** *52* (3), e12605-e12605.

40. Akhavan-Niaki, H.; Samadani, A. A., DNA Methylation and Cancer Development: Molecular Mechanism. *Cell Biochemistry and Biophysics* **2013,** *67* (2), 501-513.

41. Seligson, D. B.; Horvath, S.; Shi, T.; Yu, H.; Tze, S.; Grunstein, M.; Kurdistani, S. K., Global histone modification patterns predict risk of prostate cancer recurrence. *Nature* **2005,** *435* (7046), 1262-6.

42. Raynal, N. J. M.; Si, J.; Taby, R. F.; Gharibyan, V.; Ahmed, S.; Jelinek, J.; Estécio, M. R. H.; Issa, J.-P. J., DNA Methylation Does Not Stably Lock Gene Expression but Instead Serves as a Molecular Mark for Gene Silencing Memory. *Cancer Research* **2012,** *72* (5), 1170.

43. Goodrich, J. M.; Hector, E. C.; Tang, L.; LaBarre, J. L.; Dolinoy, D. C.; Mercado-Garcia, A.; Cantoral, A.; Song, P. X. K.; Téllez-Rojo, M. M.; Peterson, K. E., Integrative Analysis of Gene-Specific DNA Methylation and Untargeted Metabolomics Data from the ELEMENT Cohort. *Epigenetics insights* **2020,** *13*, 2516865720977888.

44. Wang, Y.-Q.; Li, Y.-M.; Li, X.; Liu, T.; Liu, X.-K.; Zhang, J.-Q.; Guo, J.-W.; Guo, L.-Y.; Qiao, L., Hypermethylation of TGF-β1 gene promoter in gastric cancer. *World journal of gastroenterology* **2013,** *19* (33), 5557-5564.

45. Shugang, X.; Hongfa, Y.; Jianpeng, L.; Xu, Z.; Jingqi, F.; Xiangxiang, L.; Wei, L., Prognostic Value of SMAD4 in Pancreatic Cancer: A Meta-Analysis. *Translational oncology* **2016,** *9* (1), 1-7.

46. Ahmed, S.; Bradshaw, A.-D.; Gera, S.; Dewan, M. Z.; Xu, R., The TGF-β/Smad4 Signaling Pathway in Pancreatic Carcinogenesis and Its Clinical Significance. *Journal of clinical medicine* **2017,** *6* (1), 5.

47. Xia, X.; Wu, W.; Huang, C.; Cen, G.; Jiang, T.; Cao, J.; Huang, K.; Qiu, Z., SMAD4 and its role in pancreatic cancer. *Tumour Biol* **2015,** *36* (1), 111-9.

48. Xia, X.; Wu W Fau - Huang, C.; Huang C Fau - Cen, G.; Cen G Fau - Jiang, T.; Jiang T Fau - Cao, J.; Cao J Fau - Huang, K.; Huang K Fau - Qiu, Z.; Qiu, Z., SMAD4 and its role in pancreatic cancer. (1423-0380 (Electronic)).

49. Sigismund, S.; Avanzato, D.; Lanzetti, L., Emerging functions of the EGFR in cancer. *Molecular oncology* **2018,** *12* (1), 3-20.

50. Juhász, M.; Nitsche B Fau - Malfertheiner, P.; Malfertheiner P Fau - Ebert, M. P. A.; Ebert, M. P., Implications of growth factor alterations in the treatment of pancreatic cancer. (1476-4598 (Electronic)).

51. Lee, J.; Snyder, E. R.; Liu, Y.; Gu, X.; Wang, J.; Flowers, B. M.; Kim, Y. J.; Park, S.; Szot, G. L.; Hruban, R. H.; Longacre, T. A.; Kim, S. K., Reconstituting development of pancreatic intraepithelial neoplasia from primary human pancreas duct cells. *Nature Communications* **2017,** *8* (1), 14686.

52. Zheng, X.; Zhang, N.; Wu, H.-J.; Wu, H., Estimating and accounting for tumor purity in the analysis of DNA methylation data from cancer studies. *Genome biology* **2017,** *18* (1), 17-17.

53. Zhang, W.; Xu, J., DNA methyltransferases and their roles in tumorigenesis. *Biomarker Research* **2017,** *5* (1), 1.

54. Gallagher, E. J.; LeRoith, D., The proliferating role of insulin and insulin-like growth factors in cancer. *Trends Endocrinol Metab* **2010,** *21* (10), 610-8.

55. Rajbhandari, N.; Lin, W.-C.; Wehde, B. L.; Triplett, A. A.; Wagner, K.-U., Autocrine IGF1 Signaling Mediates Pancreatic Tumor Cell Dormancy in the Absence of Oncogenic Drivers. *Cell reports* **2017,** *18* (9), 2243-2255.

56. Qian, J. Y.; Tan, Y. L.; Zhang, Y.; Yang, Y. F.; Li, X. Q., Prognostic value of glypican-1 for patients with advanced pancreatic cancer following regional intra-arterial chemotherapy. *Oncol Lett* **2018,** *16* (1), 1253-1258.

57. Frampton, A. E.; Prado, M. M.; López-Jiménez, E.; Fajardo-Puerta, A. B.; Jawad, Z. A. R.; Lawton, P.; Giovannetti, E.; Habib, N. A.; Castellano, L.; Stebbing, J.; Krell, J.; Jiao, L. R., Glypican-1 is enriched in circulating-exosomes in pancreatic cancer and correlates with tumor burden. *Oncotarget* **2018,** *9* (27), 19006-19013.

58. Compagni, A.; Wilgenbus, P.; Impagnatiello, M. A.; Cotten, M.; Christofori, G., Fibroblast growth factors are required for efficient tumor angiogenesis. *Cancer Res* **2000,** *60* (24), 7163-9.

59. Xu, P.; Hu, G.; Luo, C.; Liang, Z., DNA methyltransferase inhibitors: an updated patent review (2012-2015). *Expert Opin Ther Pat* **2016,** *26* (9), 1017-30.

60. Jones, P. A.; Issa, J. P.; Baylin, S., Targeting the cancer epigenome for therapy. (1471-0064 (Electronic)).

61. Cheng, Y.; He, C.; Wang, M.; Ma, X.; Mo, F.; Yang, S.; Han, J.; Wei, X., Targeting epigenetic regulators for cancer therapy: mechanisms and advances in clinical trials. *Signal Transduction and Targeted Therapy* **2019,** *4* (1), 62.

62. Silverman, B. R.; Shi, J., Alterations of Epigenetic Regulators in Pancreatic Cancer and Their Clinical Implications. *International journal of molecular sciences* **2016,** *17* (12), 2138.

63. Guo, M.; Jia, Y.; Yu, Z.; House, M. G.; Esteller, M.; Brock, M. V.; Herman, J. G., Epigenetic changes associated with neoplasms of the exocrine and endocrine pancreas. *Discovery medicine* **2014,** *17* (92), 67-73.

64. Díaz-Talavera, A.; Calvo, P. A.; González-Acosta, D.; Díaz, M.; Sastre-Moreno, G.; Blanco-Franco, L.; Guerra, S.; Martínez-Jiménez, M. I.; Méndez, J.; Blanco, L., A cancer-associated point mutation disables the steric gate of human PrimPol. *Sci Rep* **2019,** *9* (1), 1121.

65. Nones, K.; Waddell, N.; Song, S.; Patch, A. M.; Miller, D.; Johns, A.; Wu, J.; Kassahn, K. S.; Wood, D.; Bailey, P.; Fink, L.; Manning, S.; Christ, A. N.; Nourse, C.; Kazakoff, S.; Taylor, D.; Leonard, C.; Chang, D. K.; Jones, M. D.; Thomas, M.; Watson, C.; Pinese, M.; Cowley, M.; Rooman, I.; Pajic, M.; Butturini, G.; Malpaga, A.; Corbo, V.; Crippa, S.; Falconi, M.; Zamboni, G.; Castelli, P.; Lawlor, R. T.; Gill, A. J.; Scarpa, A.; Pearson, J. V.; Biankin, A. V.; Grimmond, S. M., Genome-wide DNA methylation patterns in pancreatic ductal adenocarcinoma reveal epigenetic deregulation of SLIT-ROBO, ITGA2 and MET signaling. *Int J Cancer* **2014,** *135* (5), 1110-8.

66. Wasif, N.; Ko, C. Y.; Farrell, J.; Wainberg, Z.; Hines, O. J.; Reber, H.; Tomlinson, J. S., Impact of tumor grade on prognosis in pancreatic cancer: should we include grade in AJCC staging? *Ann Surg Oncol* **2010,** *17* (9), 2312-20.

67. Ehrlich, M.; Gama-Sosa, M. A.; Huang, L. H.; Midgett, R. M.; Kuo, K. C.; McCune, R. A.; Gehrke, C., Amount and distribution of 5-methylcytosine in human DNA from different types of tissues of cells. *Nucleic acids research* **1982,** *10* (8), 2709-2721.

68. Cheng, P.; Wang, Y. F.; Li, G.; Yang, S. S.; Liu, C.; Hu, H.; Jin, G.; Hu, X. G., Interplay between menin and Dnmt1 reversibly regulates pancreatic cancer cell growth downstream of the Hedgehog signaling pathway. *Cancer Lett* **2016,** *370* (1), 136-44.

69. Sproul, D.; Meehan, R. R., Genomic insights into cancer-associated aberrant CpG island hypermethylation. *Briefings in functional genomics* **2013,** *12* (3), 174-190.

70. Jeziorska, D. M.; Murray, R. J. S.; De Gobbi, M.; Gaentzsch, R.; Garrick, D.; Ayyub, H.; Chen, T.; Li, E.; Telenius, J.; Lynch, M.; Graham, B.; Smith, A. J. H.; Lund, J. N.; Hughes, J. R.; Higgs, D. R.; Tufarelli, C., DNA methylation of intragenic CpG islands depends on their transcriptional activity during differentiation and disease. *Proceedings of the National Academy of Sciences* **2017,** *114* (36), E7526.

71. Morgan, M. A.; Shilatifard, A., Chromatin signatures of cancer. *Genes Dev* **2015,** *29* (3), 238-49.

72. Xu, F.; Liu, J.; Na, L.; Chen, L., Roles of Epigenetic Modifications in the Differentiation and Function of Pancreatic β-Cells. *Frontiers in Cell and Developmental Biology* **2020,** *8* (748).

73. Susanto, J. M.; Colvin, E. K.; Pinese, M.; Chang, D. K.; Pajic, M.; Mawson, A.; Caldon, C. E.; Musgrove, E. A.; Henshall, S. M.; Sutherland, R. L.; Biankin, A. V.; Scarlett, C. J.,

The epigenetic agents suberoylanilide hydroxamic acid and 5-AZA-2' deoxycytidine decrease cell proliferation, induce cell death and delay the growth of MiaPaCa2 pancreatic cancer cells in vivo. *Int J Oncol* **2015,** *46* (5), 2223-30.

74. Wang, X.; Wang, H.; Jiang, N.; Lu, W.; Zhang, X. F.; Fang, J. Y., Effect of inhibition of MEK pathway on 5-aza-deoxycytidine-suppressed pancreatic cancer cell proliferation. *Genet Mol Res* **2013,** *12* (4), 5560-73.

75. Liu, C.; Chen, Y.; Yu, X.; Jin, C.; Xu, J.; Long, J.; Ni, Q.; Fu, D.; Jin, H.; Bai, C., Proteomic analysis of differential proteins in pancreatic carcinomas: Effects of MBD1 knock-down by stable RNA interference. *BMC Cancer* **2008,** *8* (1), 121.

76. Zhang B Fau - Xu, J.; Xu J Fau - Li, C.; Li C Fau - Shi, S.; Shi S Fau - Ji, S.; Ji S Fau - Xu, W.; Xu W Fau - Liu, J.; Liu J Fau - Jin, K.; Jin K Fau - Liang, D.; Liang D Fau - Liang, C.; Liang C Fau - Liu, L.; Liu L Fau - Liu, C.; Liu C Fau - Qin, Y.; Qin, Y.; Yu, X., MBD1 is an Epigenetic Regulator of KEAP1 in Pancreatic Cancer. (1875-5666 (Electronic)).

77. Cheng, X.; Hashimoto, H.; Horton, J. R.; Zhang, X., Chapter 2 - Mechanisms of DNA Methylation, Methyl-CpG Recognition, and Demethylation in Mammals. In *Handbook of Epigenetics*, Tollefsbol, T., Ed. Academic Press: San Diego, **2011**; pp 9-24.

78. Meissner, A.; Gnirke, A.; Bell, G. W.; Ramsahoye, B.; Lander, E. S.; Jaenisch, R., Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res* **2005,** *33* (18), 5868-77.

79. Kim, M.; Costello, J., DNA methylation: an epigenetic mark of cellular memory. *Experimental & Molecular Medicine* **2017,** *49* (4), e322-e322.

80. Zhou, W.; Laird, P. W.; Shen, H., Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. (1362-4962 (Electronic)).

81. Tomczak, K.; Czerwińska, P.; Wiznerowicz, M., The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp Oncol (Pozn)* **2015,** *19* (1a), A68-77.

82. Mounir, M.; Lucchetta, M.; Silva, T. C.; Olsen, C.; Bontempi, G.; Chen, X.; Noushmehr, H.; Colaprico, A.; Papaleo, E., New functionalities in the TCGAbiolinks package for the study and integration of cancer data from GDC and GTEx. *PLoS Comput Biol* **2019,** *15* (3), e1006701.

83. Sun, H.; Xin, R.; Zheng, C.; Huang, G., Aberrantly DNA Methylated-Differentially Expressed Genes in Pancreatic Cancer Through an Integrated Bioinformatics Approach. *Frontiers in Genetics* **2021,** *12* (29).

84. Lawrence, M.; Huber, W.; Pagès, H.; Aboyoun, P.; Carlson, M.; Gentleman, R.; Morgan, M. T.; Carey, V. J., Software for computing and annotating genomic ranges. *PLoS Comput Biol* **2013,** *9* (8), e1003118.

85. Grossman, R. L.; Heath, A. P.; Ferretti, V.; Varmus, H. E.; Lowy, D. R.; Kibbe, W. A.; Staudt, L. M., Toward a Shared Vision for Cancer Genomic Data. *N Engl J Med* **2016,** *375* (12), 1109-12.

86. Lee, S.; Cook, D.; Lawrence, M., plyranges: a grammar of genomic data transformation. *Genome Biology* **2019,** *20* (1), 4.

87. Sepulveda, J. L., Using R and Bioconductor in Clinical Genomics and Transcriptomics. *The Journal of Molecular Diagnostics* **2020,** *22* (1), 3-20.

88. Ritchie, M. E.; Phipson, B.; Wu, D.; Hu, Y.; Law, C. W.; Shi, W.; Smyth, G. K., limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **2015,** *43* (7), e47.

89. Silva, T. C.; Coetzee, S. G.; Gull, N.; Yao, L.; Hazelett, D. J.; Noushmehr, H.; Lin, D.-C.; Berman, B. P., ELMER v.2: an R/Bioconductor package to reconstruct gene regulatory networks from DNA methylation and transcriptome profiles. *Bioinformatics* **2019,** *35* (11), 1974-1977.

90. Olivier, M.; Asmis, R.; Hawkins, G. A.; Howard, T. D.; Cox, L. A., The Need for Multi-Omics Biomarker Signatures in Precision Medicine. *Int J Mol Sci* **2019,** *20* (19).

91. Robinson, M. D.; Oshlack, A., A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology* **2010,** *11* (3), R25.

92. Bolstad, B. M.; Irizarry Ra Fau - Astrand, M.; Astrand M Fau - Speed, T. P.; Speed, T. P., A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. (1367-4803 (Print)).

93. Abbas-Aghababazadeh, F.; Li, Q.; Fridley, B. L., Comparison of normalization approaches for gene expression studies completed with high-throughput sequencing. *PLoS One* **2018,** *13* (10), e0206312.

94. Aryee, M. J.; Gutiérrez-Pabello, J. A.; Kramnik, I.; Maiti, T.; Quackenbush, J., An improved empirical bayes approach to estimating differential gene expression in microarray time-course data: BETR (Bayesian Estimation of Temporal Regulation). *BMC Bioinformatics* **2009,** *10* (1), 409.

95. Kanehisa, M.; Furumichi, M.; Tanabe, M.; Sato, Y.; Morishima, K., KEGG: new perspectives on genomes, pathways, diseases and drugs. (1362-4962 (Electronic)).

96. Huang da, W.; Sherman Bt Fau - Lempicki, R. A.; Lempicki, R. A., Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. (1362-4962 (Electronic)).

97. Chrominski, K.; Tkacz, M., Comparison of High-Level Microarray Analysis Methods in the Context of Result Consistency. *PloS one* **2015,** *10* (6), e0128845-e0128845.

98. Subramanian, A.; Tamayo, P.; Mootha, V. K.; Mukherjee, S.; Ebert, B. L.; Gillette, M. A.; Paulovich, A.; Pomeroy, S. L.; Golub, T. R.; Lander, E. S.; Mesirov, J. P., Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **2005,** *102* (43), 15545-50.

99. Hartwell, L. H.; Kastan, M. B., Cell cycle control and cancer. *Science* **1994,** *266* (5192), 1821-8.

100. Atay, S., Integrated transcriptome meta-analysis of pancreatic ductal adenocarcinoma and matched adjacent pancreatic tissues. *PeerJ* **2020,** *8*, e10141-e10141.

101. Tu, J.; Huang, Z.; Wang, Y.; Wang, M.; Yin, Z.; Mei, X.; Li, M.; Li, L., Transcriptome analysis of the procession from chronic pancreatitis to pancreatic cancer and metastatic pancreatic cancer. *Scientific Reports* **2021,** *11* (1), 3409.

102. Wu, M.; Li, X.; Zhang, T.; Liu, Z.; Zhao, Y., Identification of a Nine-Gene Signature and Establishment of a Prognostic Nomogram Predicting Overall Survival of Pancreatic Cancer. *Frontiers in Oncology* **2019,** *9* (996).

103. Mishra, N. K.; Guda, C., Genome-wide DNA methylation analysis reveals molecular subtypes of pancreatic cancer. *Oncotarget* **2017,** *8* (17), 28990-29012.

104. Khan, T.; Paul, B. K.; Hasan, M. T.; Islam, M. R.; Arefin, M. A.; Ahmed, K.; Islam, M. K.; Moni, M. A., Significant pathway and biomarker identification of pancreatic cancer associated lung cancer. *Informatics in Medicine Unlocked* **2021,** *25*, 100637.

105. Perera, R. M.; Bardeesy, N., Pancreatic Cancer Metabolism: Breaking It Down to Build It Back Up. (2159-8290 (Electronic)).

106. Li, R.; Yang, Y.-E.; Yin, Y.-H.; Zhang, M.-Y.; Li, H.; Qu, Y.-Q., Methylation and transcriptome analysis reveal lung adenocarcinoma-specific diagnostic biomarkers. *Journal of translational medicine* **2019,** *17* (1), 324-324.

107. Cavalcante, R., annotatr: genomic regions in context. B*ioinformatics.* Sartor, M., Ed. Bioconductor: R package version 1.18.1, **2017**.

108. Nakato, R.; Sakata, T., Methods for ChIP-seq analysis: A practical workflow and advanced applications. *Methods* **2021,** *187*, 44-53.

109. Cavalcante, R. G.; Sartor, M. A., annotatr: genomic regions in context. *Bioinformatics (Oxford, England)* **2017,** *33* (15), 2381-2383.

110. Ramos, M.; Schiffer, L.; Re, A.; Azhar, R.; Basunia, A.; Rodriguez, C.; Chan, T.; Chapman, P.; Davis, S. R.; Gomez-Cabrero, D.; Culhane, A. C.; Haibe-Kains, B.; Hansen, K. D.; Kodali, H.; Louis, M. S.; Mer, A. S.; Riester, M.; Morgan, M.; Carey, V.; Waldron, L., Software for the Integration of Multiomics Experiments in Bioconductor. *Cancer research* **2017,** *77* (21), e39-e42.

111. Durinck, S.; Spellman, P. T.; Birney, E.; Huber, W., Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature protocols* **2009,** *4* (8), 1184-1191.

112. Yao, L.; Shen, H.; Laird, P. W.; Farnham, P. J.; Berman, B. P., Inferring regulatory element landscapes and transcription factor networks from cancer methylomes. *Genome Biology* **2015,** *16* (1), 105.

113. Purcell, S.; Cherny, S. S.; Sham, P. C., Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics* **2003,** *19* (1), 149-50.

114. Hsiao, C. L.; Hsieh, A. R.; Lian Ie, B.; Lin, Y. C.; Wang, H. M.; Fann, C. S., A novel method for identification and quantification of consistently differentially methylated regions. *PLoS One* **2014,** *9* (5), e97513.

115. Ouyang, Y.; Pan, J.; Tai, Q.; Ju, J.; Wang, H., Transcriptomic changes associated with DKK4 overexpression in pancreatic cancer cells detected by RNA-Seq. *Tumour Biol* **2016,** *37* (8), 10827-38.

116. Lu Y, Identification of Critical Pathways and Potential Key Genes in Poorly Differentiated Pancreatic Adenocarcinoma. Li D, L. G., Xiao E, Mu S, Pan Y, Qin F, Zhai Y, Duan S, Li D, Yan G, Ed. Dove Press: *Onco Targets Ther*. **2021**; Vol. 14:711-723.

117. Sato, N.; Fukushima, N.; Matsubayashi, H.; Goggins, M., Identification of maspin and S100P as novel hypomethylation targets in pancreatic cancer using global gene expression profiling. *Oncogene* **2004,** *23* (8), 1531-1538.

118. Maehata, T.; Taniguchi, H.; Yamamoto, H.; Nosho, K.; Adachi, Y.; Miyamoto, N.; Miyamoto, C.; Akutsu, N.; Yamaoka, S.; Itoh, F., Transcriptional silencing of Dickkopf gene family by CpG island hypermethylation in human gastrointestinal cancer. *World J Gastroenterol* **2008,** *14* (17), 2702-14.

121. Cai, X.; Yao, Z.; Li, L.; Huang, J., Role of DKK4 in Tumorigenesis and Tumor Progression. *International journal of biological sciences* **2018,** *14* (6), 616-621.

122. González-Borja, I.; Viúdez, A.; Goñi, S.; Santamaria, E.; Carrasco-García, E.; Pérez-Sanz, J.; Hernández-García, I.; Sala-Elarre, P.; Arrazubi, V.; Oyaga-Iriarte, E.; Zárate, R.; Arévalo, S.; Sayar, O.; Vera, R.; Fernández-Irigoyen, J., Omics Approaches in Pancreatic Adenocarcinoma. *Cancers (Basel)* **2019,** *11* (8).

APPENDIX A. SUMMARIZED R SCRIPT

```
library(SummarizedExperiment)
library(TCGAbiolinks)
# Download pancreatic cancer datasets from GDC #
query.exp <- GDCquery(project = "TCGA-PAAD",
legacy = TRUE,
data.category = "Gene expression",
data.type = "Gene expression quantification",
platform = "Illumina HiSeq",
file.type = "results",
experimental.strategy = "RNA-Seq",
sample.type = c("Primary solid Tumor","Solid Tissue Normal"))
GDCdownload(query.exp)
pc.exp <- GDCprepare(query = query.exp, save = TRUE, save.filename = "PDAC.Exp.rda")

# Get subtype information
dataSubt <- TCGAquery_subtype(tumor = "PAAD")

# Get clinical data
dataClin <- GDCquery_clinic(project = "TCGA-PAAD","clinical")

# Which samples are primary solid tumor
dataSmTP <- TCGAquery_SampleTypes(getResults(query.exp,cols="cases"),"TP")
# Which samples are solid tissue normal
dataSmNT <- TCGAquery_SampleTypes(getResults(query.exp,cols="cases"),"NT")

dataPrep <- TCGAanalyze_Preprocessing(object = pdac.exp, cor.cut = 0.5)

dataNorm <- TCGAanalyze_Normalization(tabDF = dataPrep,
        geneInfo = geneInfo,
        method = "gcContent")

dataFilt <- TCGAanalyze_Filtering(tabDF = dataNorm,
       method = "quantile",
       qnt.cut = 0.25)

dataDEGs <- TCGAanalyze_DEA(mat1 = dataFilt[,dataSmNT],
     mat2 = dataFilt[,dataSmTP],
     Cond1type = "Normal",
     Cond2type = "Tumor",
     fdr.cut = 0.01 ,
     logFC.cut = 1,
     method = "glmLRT")
# DEGs table with expression values in normal and tumor samples
dataDEGsFiltLevel <- TCGAanalyze_LevelTab(dataDEGs,"Tumor","Normal",
```

```
            dataFilt[,samplesTP],dataFilt[,samplesNT])
```

#TCGAbiolinks outputs bar chart with the number of genes for the main categories of three ontologies (GO:biological process, GO:cellular component, and GO:molecular function, respectively).

```
ansEA <- TCGAanalyze_EAcomplete(TFname="DEA genes Normal Vs Tumor",
      RegulonList = rownames(dataDEGs))

TCGAvisualize_EAbarplot(tf = rownames(ansEA$ResBP),
    GOBPTab = ansEA$ResBP,
    GOCCTab = ansEA$ResCC,
    GOMFTab = ansEA$ResMF,
    PathTab = ansEA$ResPat,
    nRGTab = rownames(dataDEGs),
    nBar = 20)

# ELMER pipeline for differential methylated CpG site analysis.
library(TCGAbiolinks)
library(SummarizedExperiment)
library(ELMER)
library(parallel)
dir.create("dmc_analysis")
setwd("dmc_analysis")


#----------------------------------
# STEP 1: Search, download, prepare |
#----------------------------------
# 1.1 - DNA methylation
# ----------------------------------
query.met <- GDCquery(project = "TCGA-PAAD",
    data.category = "DNA Methylation",
    platform = "Illumina Human Methylation 450")
GDCdownload(query.met)
pc.met <- GDCprepare(query = query.met,
    save = TRUE,
    save.filename = "pc.met.rda",
    summarizedExperiment = TRUE)
#For gene expression we will use Gene Expression Quantification.

# Step 1.2 download expression data # Note this step was carried out previously#
#----------------------------------
# 1.2 - RNA expression
# ----------------------------------
query.exp <- GDCquery(project = "TCGA-PAAD",
```

```
    data.category = "Transcriptome Profiling",
    data.type = "Gene Expression Quantification",
    workflow.type = "HTSeq - FPKM-UQ")
GDCdownload(query.exp)
pc.exp <- GDCprepare(query = query.exp,
    save = TRUE,
    save.filename = "pcExp.rda")
pc.exp <- pc.exp
```

#A MultiAssayExperiment object from the r BiocStyle::Biocpkg("MultiAssayExperiment") package is the input for multiple main functions of r BiocStyle::Biocpkg("ELMER").

#We will first need to get distal probes (2 KB away from TSS).

```
distal.probes <- get.feature.probe(genome = "hg38", met.platform = "450K")
```

#To create it you can use the createMAE function. This function will keep only samples that have both DNA methylation and gene expression.

```
library(MultiAssayExperiment)
mae <- createMAE(exp = pc.exp,
    met = pc.met,
    save = TRUE,
    linearize.exp = TRUE,
    filter.probes = distal.probes,
    save.filename = "mae_kirc.rda",
    met.platform = "450K",
    genome = "hg38",
    TCGA = TRUE)
# Remove FFPE samples
mae <- mae[,!mae$is_ffpe]
```

We will execute ELMER to identify probes that are hypomethylated in tumor samples compared to the normal samples.
We will also repeat this step for hyper methylation analysis to identify probes that are hypermethylated in tumor samples compared to the normal samples.

```
group.col <- "definition"
group1 <- "Primary solid Tumor"
group2 <- "Solid Tissue Normal"
direction <- "hypo" # for hyper analysis replace hypo with "hyper"
dir.out <- file.path("pc",direction)
dir.create(dir.out, recursive = TRUE)
#-------------------------------------
# STEP 3: Analysis      |
#-------------------------------------
# Step 3.1: Get diff methylated probes |
```

```
#-------------------------------------

sig.diff <- get.diff.meth(data = mae,
      group.col = group.col,
      group1 = group1,
      group2 = group2,
      minSubgroupFrac = 0.2,
      sig.dif = 0.3,
      diff.dir = direction, # Search for hypomethylated probes in group 1
      cores = 1,
      dir.out = dir.out,
      pvalue = 0.01)

#-------------------------------------------------------------
# Step 3.2: Identify significant probe-gene pairs   |
#-------------------------------------------------------------
# Collect nearby 20 genes for Sig.probes
nearGenes <- GetNearGenes(data = mae,
      probes = sig.diff$probe,
      numFlankingGenes = 20, # 10 upstream and 10 downstream genes
      cores = 1)

pair <- get.pair(data = mae,
    group.col = group.col,
    group1 = group1,
    group2 = group2,
    nearGenes = nearGenes,
    minSubgroupFrac = 0.4, # % of samples to use in to create groups U/M
    permu.dir = file.path(dir.out,"permu"),
    permu.size = 100, # Please set to 100000 to get significant results
    raw.pvalue = 0.05,
    Pe = 0.01, # Please set to 0.001 to get significant results
    filter.probes = TRUE, # See preAssociationProbeFiltering function
    filter.percentage = 0.05,
    filter.portion = 0.3,
    dir.out = dir.out,
    cores = 1,
    label = direction)

# From this analysis the relationship between nearby 20 gene expression vs DNA methylation at
can be verified. The result of this is shown by ELMER scatter plot function.

scatter.plot(data = mae,
   byProbe = list(probe = sig.diff$probe[1], numFlankingGenes = 20),
   category = "definition",
   dir.out = "plots",
```

```
    lm = TRUE, # Draw linear regression curve
    save = TRUE)
# End not run #
# Setwd(mariam.zamani/analysis/files/ccast)
```