**Novel putative Methylation Driven Genes identified using network deconvulation approaches in pancreatic cancer.**

Akash Bararia, Agniswar Chakraborty, Paramita Roy, Sudeep Banerjee, Shibajyoti Ghosh, Sumit Gulati, Supriyo Ghatak, Bitan Kumar Chatterjee, Sankhadeep Dutta, Aniruddha Chatterjee, Debabani Ganguly, Indranil Mukhopadhyay, and Nilabja Sikdar

Indian Statistical Institute,

Jadavpur University,

Tata Medical Center

Medical College and Hospital

Apollo Multi superspeciality hospital,

IPGMER & SSKM

Chittarajan Cancer Research Institute

University of Otago.

JIS

Network Deconvulating Methylation Driven Genes In Pancreatic Cancer.

**INTRODUCTION:**

Apprehending the reality behind the fundamental issue of cellular understanding is based on the perception that how different interpretation are depicted by same DNA sequence in varied cell types. Architecture of cellular phenotype is dominated by the genetic and epigenetic machinery as of today’s established determinants. Fact-finding on the epigenetic machinery discloses the importance of DNA methylation, which is a crucial epigenetic mark that fortifies the lifelong architecture of cellular identity.(DNA methylation atlas ) Varied cancer types represents unique methylation patterns which modulates tumorogenesis and associated therapeutic outcomes concurrent with modulated gene expressions.(Prognostic Nomogram/8) Henceforth integrative understanding of Methylation-regulated Differentially Expressed Genes (MeDEGs) might lead the path for better understanding of prognositcation, molecular subtyping and targeted therapeutic approaches.(Prognostic Nomogram)The upsurge in accesibility of cross platform genetic and epigenetic datas has being colloquially supportive in terms of new model and integrative network development. Deconvulating the multiple dataframes and integrating them into suitable models serve as crucial checkpoints for clinical model development such as Nomogram associated patient survival prediction.(Methylation-eQTL)DNA methylation altereed gene expression associated downstream effect on tumor phenotypes and its relevant corelation has provided sustainable apprehension in development of biomarkers for pancreatic cancer (PanCa).(Goggins, 2005) Globally, there is still an inadequate implementation of MeDEGs as therapeutic biomarkers. In course enhanced availability of,high-throughput data has upstream potentiality in utilizing MeDEGs as candidates for future research purposes.(MeDeG)

Current era of digitalized health data implementation has increased in approaches based on artificial intelligence (AI) and machine learning (ML) leading to development of tools concerning precision oncology. The methodologies used circumference around both unsupervised and supervised algorithms with support from funtional bioinformatical analysis for downstream validation of ML methodolgies. (Deep learning\_GM)(ML\_cancer) The implications concerning subseting of cases into high or low risk groups based on cutoff’s typical for particular roles has supported numerous researchers, to evaluate the implication of ML approaches. Alongside, the utility of ML techniques to identify relevant variants from big biological datas validates their importance. Several techniques, including Decision Trees (DTs), Random Forest (RF), K-Neighroust neighbour, K-means and logistic regressions such as Adaptive LASSO (AL) and Support Vector Machines (SVMs) have been vividly implicated in precision oncology for the development of prediction models, repurcession in more sensitive and accurate decision making.In this article, we exhibited varied promising applications of DL in context of PanCa, including, cancer survival, prognostication, clinico-histological inference, role of epigenetic characteristics, and associated gene level expression alterations.

**METHODOLOGY:**

***Study Cohort:***

We assembled multiple compendium datasets from TCGA-PAAD, ICGC-PaCa and GEO (GSE74071 and GSE 49149). The Pancreatic Cancer Dataset from TCGA (TCGA-PAAD) was used to acquire the RNA sequence data and Methylation Array datas. The corresponding clinical information was also acquired from TCGA. The ICGC dataset was used to acquire methylation data of PC patients. The GEO datasets specific for DNA methylation 450K data were acquired for validation purposes only.

***Identifification of Methylation-regulated Differentially Expressed Genes***

The differentially expressed genes (DEGs) and the differentially methylated genes were compared first to identify genes with presence in both datasets. Multiple test corrections was performed using Benjamini & Hochberg’s method and the cutoff values were set at the FDR < 0.01. The |log2FC| > 2 were also observed as secondary screening process as significant methylation alteration doesn’t corresponding imply a |log2FC| > 2. Then, the CpG sites corresponding to these MEDEGs were selected. Pearson correlation analysis was performed to calculate the correlation between the methylation level of DMP and expression level of corresponding DEG. Such DMPs with significant negative correlation and adj. P value < 0.01, which were thought to deeply influence the expression of corresponding DEGs, were selected for subsequent analysis and were referred to as MEDEG.

***Dimensional Reduction and Unsupervised clustering:***

Principal Component Analysis (PCA),a type of unsupervised learning was implemented for dimensionality reduction by assigning correlation between multidimensional information sets. PCA was also used to data set easier to interpret by exclusion of parameter limitation. PCA followed by K means clustering for unsupervised classification of MEDEG across *TCGA-PAAD samples was done to* identify the clustering pattern of the 27 MEDEG’s spread across 69 DMS (47 CpG’s). K means are library tidyverse (data manipulation), librarycluster (clustering algorithms), library factoextra ( clustering algorithms & visualization), and library ggplot2.

***Supervised Algorithm based Machine Learning Models***

Multiple predictive models such the K-nearest neighbor (kNN) classifier is a non-parametric supervised machine learning algorithm which is adistance-based classifier**.** It is classically suitable for smaller datasets like our 47 CpG probe dataset. For handling bigger data sets and also in terms of enhanced non-parametric approach a random forest-based (RF) classifier was built for prediction model development. RF works on the Breiman and Cutler algorithm. Standard python packages were used which includes SciKit Learn, Pandas, Numpy, matploit.lib and Seaborne for visualization.

***Clinical Relevance of Methylation-regulated Differentially Expressed Genes***

The Construction of MEDEGs-convulated prognostic signature was done followed by the evolution of MEDEG’s based competiting risk-Nomogram. The process initated with univariate Cox regression analysis that identifued association between the methylation level of each MEDEG’s and patient’s overall survival (OS) in our scaled cohort. Those sites with P-values less than 0.05 were identified as prognosis-related DMSs. Then, Adaptive LASSO Regression method was implemented to identify the prognosis associated MEDEG’s and concurently an prime model was obtained. The MEDEG’s with coefficient, C-index ≠ 0 were labeled as significant variables which lead to risk scoring model establishment using the combination of weighted methylation values. The risk scores were calculated as shown in the following equation:

The mean risk score value was used as the cuttoff score for spliting, patients into low-risk (risk score below the mean value) or high-risk (risk score above the mean value) group correspondingly. To translate the prognostic score of MEDEG’s into clinical application, we then constructed a clinical nomogram, which included the risk score and the clinicopathological parameters of TCGA-PAAD patients evaluated by multivariate Cox proportional-hazards regression. It was used to evaluate the independent prognostic value of the signature after adjusting for age, sex and stage alongside predicting the overall survial (1-,2- and 3-years) (OS) in TCGA-PAAD cohort.

The discriminatory ability of the nomogram was evaluated by calculating the concordance index (C-index), which is a measure of discrimination. Calibration plots were plotted to compare the observed and predicted probabilities for the nomogram.

***Tissue of Origin of Methylation-regulated Differentially Expressed Genes***

Tissue specificity of selected ***MEDEG’s*** were done by comparing with normal pancreas tissue and other tissue types . The data were taken from GTEx V8 dataset. The PanCancer profile was developed to highlight role of selected MEDEG’s in pancreas tissue by use of GEPIA. Protein level data was also obtained from CPTAC database for highlighting downstream protein level signature role in tissue of pancreas origin by use of UALCAN. Finally the profile of selected CpG were compared with profile obtained from two other PanCa databases taken from GEO database for global profiling purpose irrespective of demography and ethinicity.

***Functional Relevance of Methylation-regulated Differentially Expressed Genes***

Functional characterization of MEDEG’s were done to obtaine enriched gene ontology (GO) and its associated pathways Under the category, biological processes (BP) ontology that showcased the pathway enrichment which is an crucial resource for a intimate understanding of biological process from large-scale molecular cohorts produced by highthroughput experimental technology. The criterion for significant enrichment was based on Benjamini-Hochberg adjustment (p < 0.05). Functional characterization leading to enriched GO were also obtained using DEG in PanCa based on BP Ontology. Enrichment network was also obtained by MetaScape which has its nodes colored based on p-value. ENRICH R was used to obtain functional characterisation of the MEDEG’s in context of their role in other associated molecular processes and cellular functions. (DEG\_DMG\_Nomo)

***Immune and Cell Ecosystem profile of Methylation-regulated Differentially Expressed Genes***

Immune and stromal cell infiltration data concurent with selected MEDEG’s were obtained via ESTIMATE and significance was identified using Kruskar Wallis***.*** Immune profile and its associated ecosystem associated data was obtained using deconvulation based immune classifier such as CIBERSORT and ECOTYPER. Gene marker associated immune classifier was used alongside via Xcell. xCell, a novel gene signature-based method, was used to infer 64 immune and stromal cell types. EcoTyper, an upgrade of CIBERSORT X (machine learning framework) was used for large-scale identification and validation of cell states and multicellular communities from bulk gene expression data. Applied to 12 major cell lineages across 16 types of human carcinoma.

***Pharmacogenomic screening of Methylation-regulated Differentially Expressed Genes***

Top three genes were selected for molecular dynamics lead simulation based visualization of interaction amongst molecules. The interaction between the genes and drug molecules specific for pancreatic cancer tissue will also be evaluated for suitable drug targetibility identification of the genes of interaction. Drugs undergoing both clinical trail and in vitro development of small molecules will be considered.

**RESULTS:**

Conclusion and Future Prosepective:

Even though it is evident that the use of ML methods can improve our understanding of cancer progression, an appropriate level of validation is needed in order for these methods to be considered in the everyday clinical practice. (ML\_cancer)