1 Methology

Algorithm 1 Study Cohort

1: Assemble multiple compendium datasets from TCGA-PAAD, ICGC-PaCa and GEO (GSE74071 and GSE 49149).

2: Acquire RNA sequence data and Methylation Array data from the Pancreatic Cancer Dataset of TCGA (TCGA-PAAD).

3: Acquire corresponding clinical information from TCGA.

4: Acquire methylation data of PC patients from the ICGC dataset.

5: Acquire GEO datasets specific for DNA methylation 450K data for validation purposes only. =0

Algorithm 2 Identification of Methylation-regulated Differentially Expressed Genes

0: **procedure** IDENTIFICATION($RNAseq, Methylation, clinical$)

1: $DEGs \leftarrow$ Differentially expressed genes from RNAseq data

2: $DMPs \leftarrow$ Differentially methylated genes from Methylation data

3: $overlap \leftarrow$ Genes with presence in both datasets

4: $pvalues \leftarrow$ Multiple test corrections using Benjamini and Hochberg's method

5: $DEGs \leftarrow$ DEGs with $FDR < 0.01$ and $|log_2FC| > 2$

6: $MEDEGs \leftarrow$ Genes with significant negative correlation between methylation and expression

7: **return** $MEDEGs$

7: **end procedure**=0

Algorithm 3 Dimensional Reduction and Unsupervised Clustering

1: Principal Component Analysis (PCA) was implemented for dimensionality reduction by assigning correlation between multidimensional information sets.

2: PCA was also used to make the dataset easier to interpret by excluding parameter limitations.

3: PCA followed by K-means clustering was done to identify the clustering pattern of the 27 MEDEGs spread across 69 DMS (47 CpGs).

4: The following packages were used for K-means clustering: `tidyverse` (data manipulation), `cluster` (clustering algorithms), `factoextra` (clustering algorithms and visualization), and `ggplot2`. =0

Algorithm 4 Supervised Algorithm-based Machine Learning Models

1: Multiple predictive models such as the K-nearest neighbor (kNN) classifier were used. kNN is a non-parametric supervised machine learning algorithm that is distance-based and is classically suitable for smaller datasets like our 47 CpG probe dataset.

2: For handling bigger datasets and also in terms of enhanced non-parametric approach, a random forest-based (RF) classifier was built for prediction model development.

3: RF works on the Breiman and Cutler algorithm.

4: The following standard Python packages were used for RF classification: `scikit-learn`, `pandas`, `numpy`, `matplotlib`, and `seaborn` for visualization. =0

Algorithm 5 Clinical Relevance of Methylation-regulated Differentially Expressed Genes

1: Construct MEDEGs-convoluted prognostic signature

2: Perform univariate Cox regression analysis to identify associations between methylation level of each MEDEG and patient's overall survival (OS) in scaled cohort

3: Identify prognosis-related DMSs with P-values less than 0.05

4: Implement Adaptive LASSO Regression method to identify prognosis associated MEDEGs and obtain a prime model

5: Label MEDEGs with coefficient, C-index $\neq 0$ as significant variables

6: Establish risk scoring model using combination of weighted methylation values

7: Calculate risk scores using the following equation:

$$Risk\ Score = \sum_{i=1}^{n} \beta_i \times X_i$$

8: Use mean risk score value as cutoff score to split patients into low-risk (risk score below mean value) or high-risk (risk score above mean value) group

9: Construct clinical nomogram including risk score and clinicopathological parameters of TCGA-PAAD patients evaluated by multivariate Cox proportional-hazards regression

10: Use nomogram to evaluate independent prognostic value of signature after adjusting for age, sex, and stage alongside predicting overall survival (1-, 2-, and 3-years) (OS) in TCGA-PAAD cohort

11: Evaluate discriminatory ability of nomogram by calculating concordance index (C-index)

12: Plot calibration plots to compare observed and predicted probabilities for nomogram =0

Algorithm 6 Analysis of Methylation-regulated Differentially Expressed Genes

Tissue of Origin

Compare selected MEDEGs with normal pancreas tissue and other tissue types using GTEx V8 dataset

Develop PanCancer profile to highlight role of selected MEDEGs in pancreas tissue using GEPIA

Obtain protein level data from CPTAC database to highlight downstream protein level signature role in pancreas tissue using UALCAN

Compare profile of selected CpGs with profile obtained from two other PanCa databases taken from GEO database for global profiling purpose irrespective of demography and ethnicity

Functional Relevance

Obtain enriched gene ontology (GO) and associated pathways under biological processes (BP) ontology using DEGs in PanCa based on BP Ontology

Use MetaScape to obtain enrichment network with nodes colored based on p-value

Use ENRICHR to obtain functional characterization of MEDEGs in context of their role in other associated molecular processes and cellular functions ($\mathrm{DEG}_D MG_N omo$)

Immune and Cell Ecosystem Profile

Obtain immune and stromal cell infiltration data with selected MEDEGs using ESTIMATE and identify significance using Kruskal-Wallis

Obtain immune profile and associated ecosystem data using deconvolution-based immune classifiers such as CIBERSORT and EcoTyper

Use gene marker-associated immune classifier via xCell to infer 64 immune and stromal cell types

Pharmacogenomic Screening

Select top three genes for molecular dynamics lead simulation-based visualization of interaction among molecules

Evaluate interaction between genes and drug molecules specific for pancreatic cancer tissue for suitable drug target identification of genes of interaction

Consider drugs undergoing both clinical trial and in vitro development of small molecules =0