A new computational framework like Google identifies therapeutic targets based on machine learning unveiling influential genes in pancreatic islet cells

Turki Turki[a*] and Y-h. Taguchi[b*]

[a] King Abdulaziz University, Department of Computer Science, Jeddah 21589, Saudi Arabia. Email: tturki@kau.edu.sa

[b] Chuo University, Department of Physics, Tokyo 112-8551, Japan. Email: tag@granular.com

***Corresponding Authors**.

Abstract

Pancreatic islets comprise a group of cells that produce hormones regulating blood glucose levels. Particularly, the alpha and beta islet cells produce glucagon and insulin to stabilize blood glucose. When beta islet cells are dysfunctional, insulin is not secreted, inducing a glucose metabolic disorder. Identifying effective therapeutic targets against the disease is a complicated task and is yet not conclusive. To close the wide gap between understanding the molecular mechanism of pancreatic islet cells and providing effective therapeutic targets, we present a computational framework to identify potential therapeutic targets against pancreatic disorders. First, we downloaded three transcriptome expression profiling datasets pertaining to pancreatic islet cells (GSE87375, GSE79457, GSE110154) from the Gene Expression Omnibus database. For each dataset, we extracted expression profiles for two cell types. We then provided the extracted expression profile along with the cell types to our proposed constrained optimization problem of a support vector machine and to other existing methods, selecting important genes from the expression profiles. Finally, we performed (1) an evaluation from a classification perspective showing the superiority of our methods against the baseline; and (2) an enrichment analysis showing that our methods achieved better outcomes. Results for the three datasets included 44 unique

genes and 10 unique transcription factors (SP1, HDAC1, EGR1, E2F1, AR, STAT6, RELA, SP3, NFKB1, and ESR1) that have been reported related to pancreatic islet functions, diseases, and therapeutic targets.

Keywords: Machine learning; enrichment analysis; transcriptome expression; pancreatic islet; therapeutic targets; applications in biomedicine

1. Introduction

Pancreatic islets comprise several endocrine cells that play a key role in metabolism [1, 2]. Alpha and beta islet cells are endocrine cells involved in regulating blood glucose levels [3]. As carbohydrates break down into glucose and enter the blood stream, beta cells secrete insulin to absorb glucose into cells for energy. Insulin mainly aims to store excess glucose in the liver [4-6]. When beta cells do not function properly, insulin is not secreted and blood glucose levels are elevated. Left untreated, this condition can lead to the development of diabetic complications, such as blindness, cancer, and kidney damage, among others [7, 8]. Advances in biology and medical research have led to an improved understanding of the pancreatic functions, diseases, and therapeutic targets.

Zheng et al. [9] downloaded transcriptome expression profiling dataset (GSE119794) pertaining to pancreatic adenocarcinoma (PAAD) from the GEO database. Then a bioinformatics approach was utilized, identifying 262 differentially expressed genes between patients and control profiles through the help of limma package in R. Then, an enrichment analysis was applied in which three hub genes (CXCL5, KRT19, and COL11A1) were identified by cytoHubba from the protein-protein interactions (PPI) network. The three hub genes were submitted to GEPIA and TIMER within the TCGA database. Results indicated that COL11A1 was (highly correlated) and expressed in many expression profiles pertaining to PAAD. These results suggest that COL11A1 can act as a biomarker pertaining to PAAD.

Prashanth et al. [10] proposed a bioinformatics approach to identify hub genes that may play a key role in understanding type 1 diabetes (T1D) and can act as targets in the treatment and diagnosis of T1D. First, they downloaded expression-profiling dataset (GSE123658) from the GEO database**.** The dataset consisted of 39 and 43 expression profiles for T1D and healthy individuals, respectively. Totally 284 differentially expressed genes were identified through the limma package in R. An enrichment analysis of the 284 was carried out, to identify 10 hub genes (EGFR, POLR2A, PRKACA, GRIN2B, GABARAP, TLN1, GJA1, CAP2, MIF, and PXN) [10]. Kang et al. [11] aimed to identify genes that can act as a therapeutic targets of T2D patients with periodontitis comorbidity. Five expression profiling datasets were downloaded from the GEO database. A total of 152 commonly DEGs were identified between T2D and periodontitis patients. Then, an enrichment analysis was applied to the 152 genes where four hub genes

(PTPRC, HGF, RAC2, and INPP5D) were identified from the PPI network. These can act as therapeutic targets for T2D patients with periodontitis. Chan et al. identified common driver genes for T2D and cardiovascular disease from genome-wide association studies within women from different ethnicities [12]. Other studies have also aimed to contribute to the understanding of molecular mechanism of pancreas-related disorders and provide therapeutic targets [13-17].

Although current advances in biology and medical research have contributed to a better understanding of pancreatic disorders and diseases, advanced computational frameworks are needed to allow us to uncover the genetic architecture of various metabolic disorders. These frameworks will enable us to identify influential genes that play key roles in pancreatic islet function and identify therapeutic targets. They will advance our understanding of shared molecular mechanisms. The process of drug discovery can be accelerated by computational frameworks because they reduce the search space of potential therapeutic targets.

Contributions. We list our main contributions as follows:

(1) We present a computational framework comprising a novel machine learning formulation of support vector machines (SVM) [18], followed by biological enrichment analysis, to retrieve influential genes that will allow us to (a) understand the molecular mechanisms of the pancreas; (b) improve targeted therapy related to the pancreas; and (c) discover common molecular mechanisms in different cell types.

(2) We compare the proposed framework with other popular methods, including SVM, limma [19], sam, and t test [20] for the identification of genes related to pancreatic islet functions, diseases and therapeutic targets.

(3) We evaluate all methods from two perspectives using three transcription profiling datasets pertaining to pancreatic cells (mainly islet cells) from the GEO database. From the biological perspective, experimental results report the identification of 44 genes and 10 transcription factors. These have been reported related to pancreatic functions, diseases, and therapeutic targets. In terms of the classification perspective, we demonstrate that our new formulation of SVM contributes to improved performance results when incorporated into our computational framework.

Organization. The rest of the study is organized as follows. In Section 2, we give an overview of methods related to our work. In Section 3, we describe transcriptome expression profiling datasets used in this study, followed by our proposed computational framework. Then, we evaluate and report performance results of all methods considered in this study in Section 4 and give a discussion in Section 5. Finally, we provide a conclusion to this work, pointing out future directions.

2. Related Work

2.1 Support Vector Machines

An SVM is a machine learning algorithm that employs training data $\{(x_1,y_1),\ldots,(x_m,y_m)\}$ comprising *m* labeled examples to induce model *h*. The data is used to perform predictions of unseen examples. In Equation 1, we show the traditional constrained optimization problem for SVM in the non-separable case, where the main goal is to learn arguments (i.e., weights, w, and bias, *b*) that minimize the cost function, subject to linear constraints. This method guarantees that a decision boundary with the maximum margin is obtained. Since training data are not linearly separable most of the time, slack variables ξ_i are introduced to allow for learning w and *b* while incurring loss in the training data. It can be seen from Equation 2 that the model induced by SVM is just the dot product of weights and features in addition to the bias (i.e., w.x + *b*). For a given test example x*, prediction is performed by taking the sign of w.x* + *b*. *I*() as an indicator function that maps to 1 if its arguments are true. If its arguments are false, however, it maps to -1. Therefore, the outcome y^* is either +1 or -1, depending on the input class label that maximizes the functional value.

$$\begin{aligned} &\arg\min_{\mathbf{w},b} \ \frac{\|\mathbf{w}\|^2}{2} + c\sum_{i=1}^{m} \xi_i^k \\ \text{subject to} \quad & y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq 1 - \xi_i, \ \ i = 1, 2, ..., m \\ & \xi_i \geq 0 \end{aligned} \tag{1}$$

$$\begin{aligned} & y^* = \arg\max_{y} \ I(h(\mathbf{x}^*) = y), \\ & h(\mathbf{x}^*) = \text{sign}(\mathbf{w}\cdot\mathbf{x}^* + b) \end{aligned} \tag{2}$$

2.1 Huber Regression (HR)

HR is a machine learning algorithm that takes input training data $\{(x_1,y_1),\ldots,(x_m,y_m)\}$ to induce model *h*, which then takes unseen examples to perform predictions [21]. Unlike the constrained optimization problem of SVM (as seen from Equation 1), the optimization problem in HR is unconstrained, aiming to learn weights (also called coefficients) β. Moreover, HR offers two different objective functions, as seen from Equation 3. For a new test example, prediction is performed by taking the dot product of coefficients and features (i.e., x*. β). The outcome is a continuous value.

$$\arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{m} \phi(y_i - \mathbf{x}_i \cdot \boldsymbol{\beta}) \qquad (3)$$

$$\phi(v) = \begin{cases} 2rv - r^2, & \text{if } |v| > r \\ v^2, & \text{otherwise} \end{cases}$$

3. Materials and Methods

3.1 Transcriptome Expression Profiles

We downloaded three single-cell RNA-seq datasets from the NCBI GEO according to the following GEO IDs.

3.1.1 GSE87375: Dataset1

We retrieved this single-cell dataset using the GEO ID GSE87375 [22], considering 913 single-cells and 40916 expression values. Therefore, we have a 913 x 40917 matrix including a column vector for the cell type. The 913 cells are distributed between alpha and beta cells in the pancreatic islets, where the former constitutes 338 out of 913 and the latter constitutes the remaining (i.e., 575 pancreatic beta cells). We refer to the remaining dataset as Dataset1.

3.1.2 GSE79457: Dataset2

This dataset comprised 181 single-cells and 23431 expression values [23]. Thus, we encoded it as a 181 x 23432 matrix including the cell type column vector. We used single-cells pertaining to alpha and

delta cells in the pancreas, where the former constitutes 63 out of the 181 single-cells and the latter constitutes 31 single-cells. There are 47 and 40 remaining cells for beta and other cells, respectively. We refer to this dataset as dataset2. It is noteworthy that although the original dataset had 182 single-cells, one expression profile was excluded because it was defective.

3.1.3 GSE110154: Dataset3

The third dataset comprised 1860 single-cells and 22762 expression values [24]. In this dataset, we obtained a total of 1860 cells, of which 376 were acinar and 42 were delta pancreatic cells . There were 1442 remaining cells for other types. for We refer to this dataset in the remaining sections as dataset3.

3.2 Computational Framework

In Figure 1, we show the computational steps carried out by the proposed computational framework. Suppose that we have a single-cell dataset as $\{(x_1,y_1),\ldots,(x_m,y_m)\}$, where x_i represents the i*th* cell and y_i is the corresponding cell type. In this study, we consider y_i in the binary case. We encode all cells x_i (for i = 1..m) in this dataset as an $m \times n$ matrix, where m and n are the number of cells and genes, respectively. Similarly, we encode all cell types y_i (for i = 1..m) as a $1 \times m$ column vector. We seek to identify p influential genes out of the n genes, such that these p genes are mainly expressed in a given cell type. As a result, we propose a robust variant of the SVM in Section 2.1 based on the HR in Section 2.2, working as follows. In Equation 4, the constrained optimization problem finds the weights $w = [w_1 \ldots w_n]$ (a vector consisting of n elements) and bias $b \in R$ (consisting of one value) such that the cost function is minimized. As each cell is encoded by n genes, the weights w correspond to the importance of these genes among all cells. The higher the weight w_i is, the more important the gene i is.

$$\begin{aligned} &\arg\min_{\mathbf{w},b} \; u\sum_{i=1}^{m}\phi\left(\frac{\|\mathbf{w}\|^2}{2}\right) + c\sum_{i=1}^{m}\xi_i^k \\ &\phi(v) = \begin{cases} 2rv - r^2, & \text{if } |v| > r \\ v^2, & \text{otherwise} \end{cases} \\ \text{subject to} \quad & y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq 1 - \xi_i, \; i = 1, 2, ..., m \\ & \xi_i \geq 0 \end{aligned} \tag{4}$$

According to Equation 4, the constrained optimization problem becomes as shown from Equation 5 when $|v| > r$. Otherwise, the optimization problem becomes as in Equation 6. Depending on the dataset, this leads to different cost functions and different weights as a result, when compared to the original SVM in Section 2.1. To solve both constrained optimization problems in Equations 5 and 6 and obtain weights w and bias b, we utilize the CVXR solver in R [25].

$$\begin{aligned} &\arg\min_{\mathbf{w},b} \; u\sum_{i=1}^{m}(2r\frac{\|\mathbf{w}\|^2}{2} - r^2) + c\sum_{i=1}^{m}\xi_i^k \\ &\text{subject to} \quad y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq 1 - \xi_i, \; i = 1, 2, ..., m \\ &\qquad\qquad \xi_i \geq 0 \end{aligned} \tag{5}$$

$$\begin{aligned} &\arg\min_{\mathbf{w},b} \; u\sum_{i=1}^{m}(\frac{\|\mathbf{w}\|^2}{2})^2 + c\sum_{i=1}^{m}\xi_i^k \\ &\text{subject to} \quad y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq 1 - \xi_i, \; i = 1, 2, ..., m \\ &\qquad\qquad \xi_i \geq 0 \end{aligned} \tag{6}$$

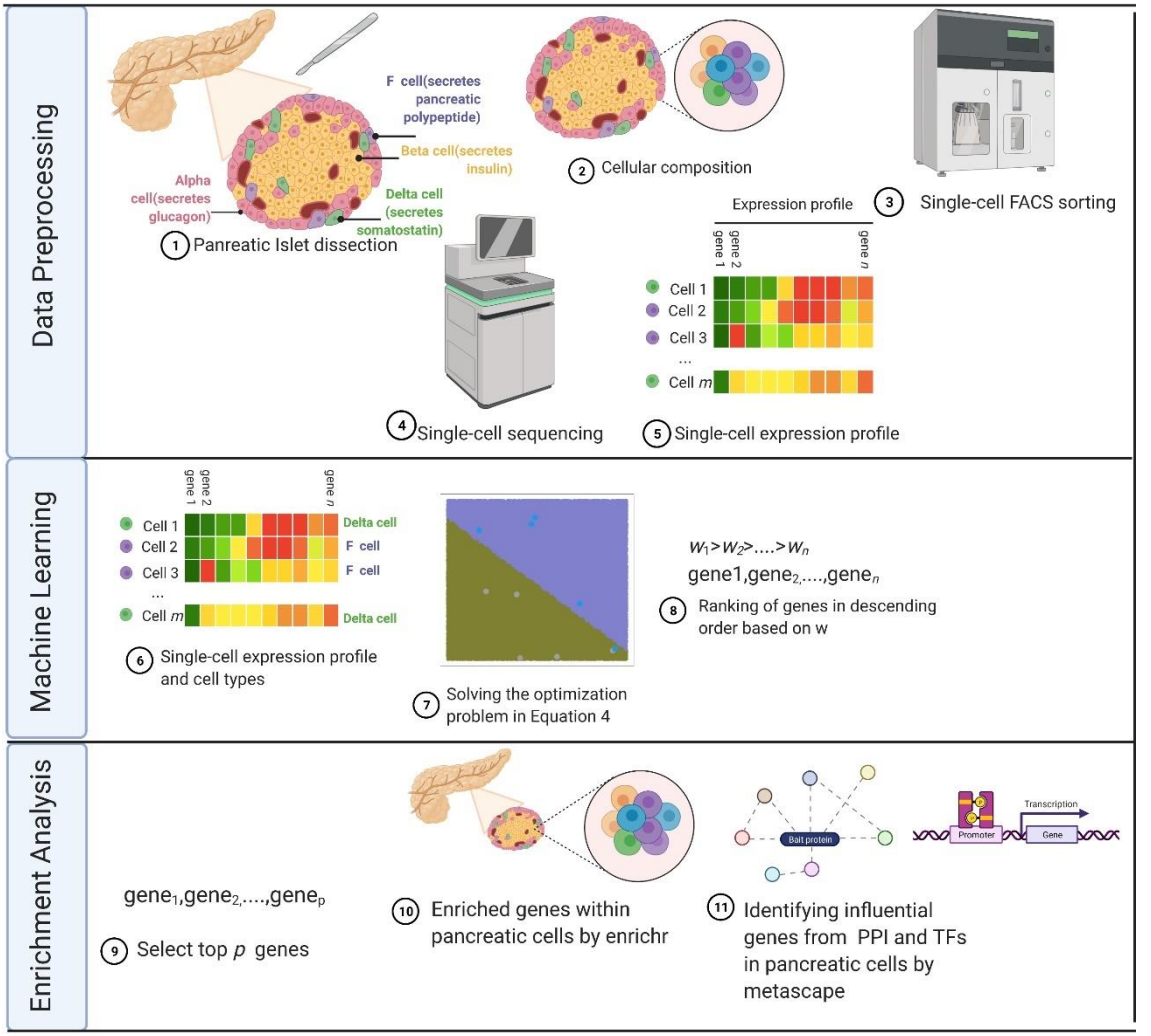


Figure 1: Flowchart of the proposed computational framework to identify influential genes expressed in pancreatic islet cells. Figure created with BioRender.com.

4 Experiments and Results

4.1 Experimental Methodology

We compare our proposed methods (i.e., scsvm1 and scsvm2) against the baseline and other well-known computational methods, such as limma [19], sam, and t test. [20]. All the approaches work under the supervised learning approach, where each method takes labeled examples to induce the model. For methods using SVM (i.e., scsvm1, scsvm2 and baseline), because the SVM model is expressed as

weighted features plus the bias (i.e., w.x + *b*), we select the top *p* genes based on the top *p* values in w, as those correspond to genes with high importance. For the remaining methods, we select significant genes according to adjusted *p*-values < 0.01. In Table 1, we give a summary of the methods used in this study.

Table 1: Summary of studied Computational Methods.

Abbreviation	Prediction Algorithm
scsvm1	The computational framework with *k* =1 in Equation 4
scsvm2	The computational framework with *k* =2 in Equation 4
baseline	The baseline SVM in Section 2.1
limma	linear models for microarray data
sam	significance analysis of microarrays
ttest	Statistical test

To evaluate the performance results, we performed biological enrichment analysis by uploading genes of each method to Enrichr (https://maayanlab.cloud/Enrichr/) [26] and Metascape (https://metascape.org/gp/index.html) [27]. Assuming the term is a cell, the greater the expressed genes in a cell, the better the performance is when the cell is related to the studied cell type. The same holds when a term is a tissue (i.e., consisting of cells performing the same function). Moreover, the lower the ranking of the term (e.g., cell) is, the better the performance is. We also evaluate the models from a classification perspective, using area under curve (AUC) performance measures, where the higher the AUC, the better the performance is. We performed experiments in this study using R [28]. Specifically, we used the CVXR package in R to solve optimization problems [25], sam function in siggenes package in R to run sam [29], lmFit and eBayes functions in limma package in R to run limma [19] and the t test function in stats package in R to run the t test [28].

4.2 Results

4.2.1 Dataset1

In Table 2, we show terms retrieved by each method when uploading 95 genes to Enrichr. The lower that ranking is with more expressed genes in cells, the better the result is. It can be shown from Table 2 that the top three retrieved terms (when scsvm1 is utilized) are pancreatic islets, beta cells, and alpha cells

(written in bold in the overlap column). Thirty-nine out of 95 genes were expressed in the pancreatic islets and distributed into its cells. Twenty-nine genes out of the 95 enriched genes were expressed in pancreatic beta and alpha cells. Twenty-seven genes (PCSK2, SLC7A14, SLC7A2, TMEM27, PCSK1N, TTR, EPCAM, PEG10, NPY, TSPAN7, PPY, CHGB, CHGA, PEG3, GPX3, ETV1, KCNK16, PAX6, TM4SF4, SMARCA1, ISL1, PYY, G6PC2, SCG5, CPE, IAPP, GHRL) out of the 29 were common between alpha and beta islet cells. The *p*-values indicate that these outcomes are significant. The second-best method is scsvm2, where the top two retrieved items are pertaining to pancreatic islets and beta cells. The term alpha cell ranked sixth. These results demonstrate the superiority of scsvm1 and scsvm2, when compared to all other methods where all cells pertaining to the pancreas are not in the top three items with more expressed genes. Regarding Dataset 1, we provide a list of genes obtained via each method in Supplementary Data Sheet 1. Moreover, we provide a full list of enrichment analysis results of each method in Supplementary Table 1.

Table 2: Enriched terms in "ARCHS4 Tissues" using Enrichr based on genes obtained from Dataset1, consisting of alpha and beta pancreatic cells. Rank column shows the order of terms when retrieved. Overlap displays the number of common genes between the uploaded gene and those in the term divided by the number of genes in that term. Best results of a method is in bold.

Method	**Rank**	**Term**	**Overlap**	***P*-value**	**Adjusted *P*-value**
scsvm1	1	PANCREATIC ISLET	**39**/2316	2.23×10^{-13}	2.41×10^{-11}
	3	BETA CELL	**29**/2316	5.79×10^{-7}	2.08×10^{-5}
	2	ALPHA CELL	**29**/2316	5.79×10^{-7}	2.08×10^{-5}
scsvm2	1	PANCREATIC ISLET	34/2316	6.31×10^{-10}	6.56×10^{-8}
	2	BETA CELL	24/2316	1.60×10^{-4}	8.30×10^{-3}
	6	ALPHA CELL	20/2316	5.60×10^{-3}	8.33×10^{-2}
baseline	1	PANCREATIC ISLET	33/2316	2.71×10^{-9}	2.81×10^{-7}
	5	BETA CELL	22/2316	1.06×10^{-3}	1.83×10^{-2}
	8	ALPHA CELL	19/2316	1.19×10^{-2}	1.23×10^{-1}
limma	101	PANCREATIC ISLET	4/2316	9.97×10^{-1}	1.00×10^{0}
	46	BETA CELL	11/2316	5.48×10^{-1}	1.00×10^{0}
	83	ALPHA CELL	7/2316	9.34×10^{-1}	1.00×10^{0}
sam	3	PANCREATIC ISLET	16/2316	7.91×10^{-2}	1.00×10^{0}

	1	BETA CELL	16/2316	7.91×10^{-2}	1.00×10^{0}
	10	ALPHA CELL	11/2316	5.48×10^{-1}	1.00×10^{0}
ttest	50	PANCREATIC ISLET	8/2316	8.73×10^{-1}	1.00×10^{0}
	2	BETA CELL	14/2316	2.07×10^{-1}	1.00×10^{0}
	28	ALPHA CELL	9/2316	7.85×10^{-1}	1.00×10^{0}

Because scsvm1 exhibited good results in terms of retrieving the cells related to studying trancriptome expression profiles in dataset1, we uploaded 95 genes of scsvm1 to Metascape. Figure 2 shows the protein-protein interaction (PPI) network obtained from Metascape. Figure 3 shows the highly connected subgraphs obtained from the PPI network, which include the following 16 genes: PPY, PYY, NPY, GNG12, BSG, HSP90AA1, HSPA8, RBM39, H1-2, HSP90AB1, RRM2, TUBBB, ATP5F1B, ACTB, SLC25A5, and EEF1A1.

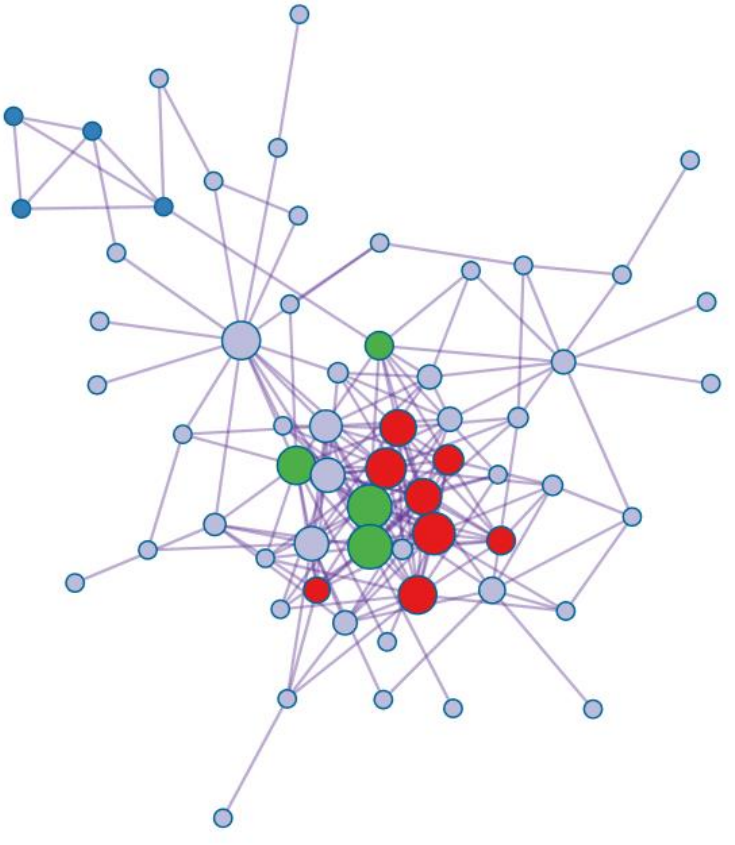


Figure 2: Protein-Protein interaction network provided by metascape for enriched genes of scsvm1

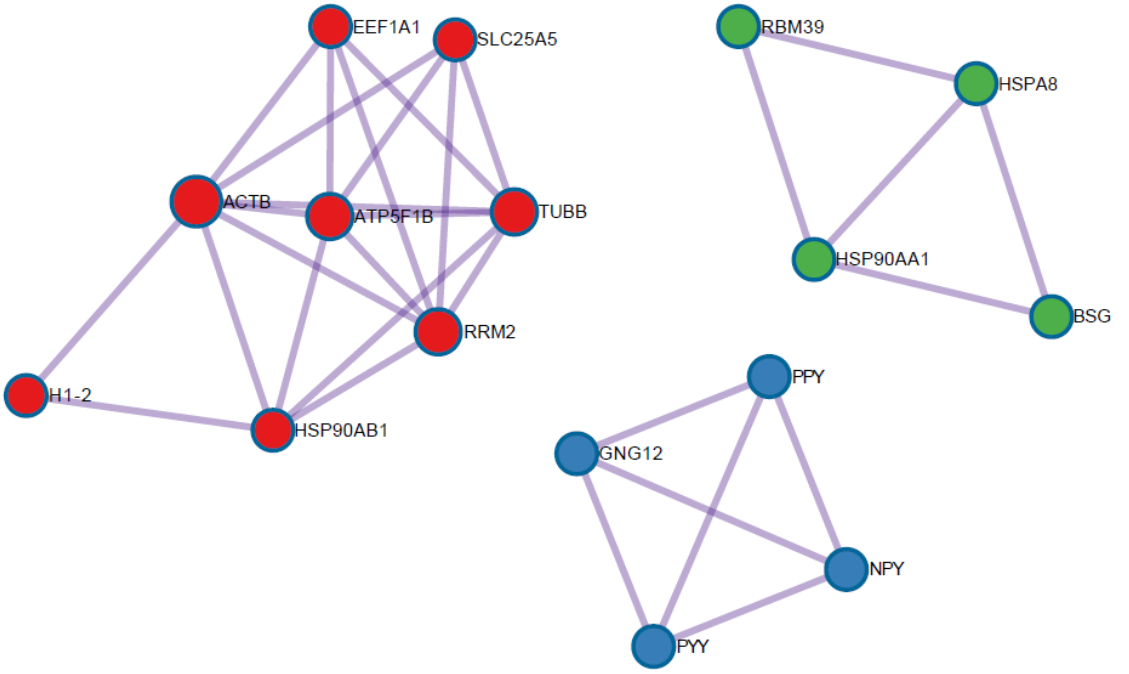


Figure 3: Three highly interconnected regions (clusters) in a protein-protein interaction enrichment analysis according to 95 genes of scsvm1, submitted to metascape.

Some of these identified genes have played a key role in understanding the working mechanism of pancreatic cells [30, 31]. In Figure 4, we identified four transcription factors- SP1, HDAC1, EGR1 and E2F1. SP1 has been reported as a biomarker for identifying the most common type of pancreatic cancer- pancreatic ductal adenocarcinoma [32]. HDAC1 has also been reported to control the expression of the p53 gene in pancreatic cancer. Additionally, EGR1 and E2F1 control the expression of cancer cells [33, 34]. These 16 identified genes and their 4 transcription factors can aid as therapeutic targets pertaining to common disorders in the pancreas.

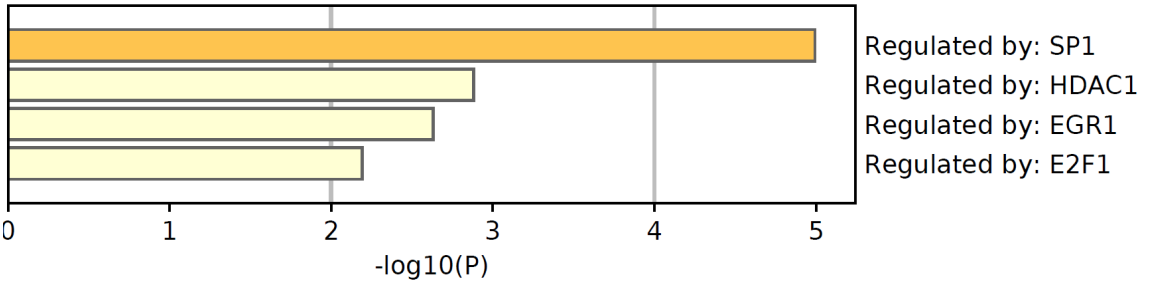


Figure 4: Four transcription factors regulating the submitted genes by scsvm1 to metascape.

4.2.2 Dataset2

Table 3 reports the retrieved terms for the genes uploaded by each method to Enrichr. It can be shown that our method scsvm2 outperforms all other methods. Specifically, pancreatic islet is ranked at the top, with 20 expressed genes out of 2316. Thirteen and twelve genes were expressed in beta and alpha cells, respectively. Eleven genes (PCSK2, MPP2, PEG3, ABCC8, GCG, BAIAP3, EML5, NR4A1, ARC, GNG4, and CHGB) were commonly expressed in beta and alpha cells. In terms of the number of genes in sam and t test, it is not likely that these will be expressed by scsvm2 and scsvm1. These results reflect that scsvm2 has more expressed genes in these terms than all other methods (see results in bold in the overlap column). Results obtained by scsvm1 and baseline are almost the same.

Table 3: Enriched terms in "ARCHS4 Tissues" using Enrichr based on genes obtained from Dataset2, consisting of alpha and delta pancreatic cells. Rank column shows the order of terms when retrieved. Overlap displays the number of common genes between the uploaded gene and those in the term divided by the number of genes in that term. Best results of a method is in bold.

Method	Rank	Term	Overlap	P-value	Adjusted P-value
scsvm1	1	PANCREATIC ISLET	18/2316	2.37×10^{-2}	1.00×10^{0}
	5	BETA CELL	12/2316	4.21×10^{-1}	1.00×10^{0}
	9	ALPHA CELL	11/2316	5.48×10^{-1}	1.00×10^{0}
scsvm2	1	PANCREATIC ISLET	**20**/2316	5.61×10^{-3}	5.94×10^{-1}
	4	BETA CELL	**13**/2316	3.04×10^{-1}	1.00×10^{0}
	6	ALPHA CELL	**12**/2316	4.21×10^{-1}	1.00×10^{0}

baseline	1	PANCREATIC ISLET	18/2316	2.37×10^{-2}	1.00×10^{0}
	5	BETA CELL	12/2316	4.21×10^{-1}	1.00×10^{0}
	8	ALPHA CELL	11/2316	5.48×10^{-1}	1.00×10^{0}
limma	90	PANCREATIC ISLET	2/2316	1.00×10^{0}	1.00×10^{0}
	96	BETA CELL	1/2316	1.00×10^{0}	1.00×10^{0}
	61	ALPHA CELL	3/2316	9.99×10^{-1}	1.00×10^{0}
sam	3	PANCREATIC ISLET	4/2316	8.14×10^{-4}	1.41×10^{-2}
	2	BETA CELL	4/2316	8.14×10^{-4}	1.41×10^{-2}
	1	ALPHA CELL	4/2316	8.14×10^{-4}	1.41×10^{-2}
ttest	3	PANCREATIC ISLET	3/2316	1.55×10^{-3}	2.17×10^{-2}
	2	BETA CELL	3/2316	1.55×10^{-3}	2.17×10^{-2}
	1	ALPHA CELL	3/2316	1.55×10^{-3}	2.17×10^{-2}

Because scsvm2 had more expressed genes in pancreatic cells, we uploaded these 95 genes by scsvm2 to Metascape. In Figure 5, we show the PPI network provided by Metascape. Two highly connected subgraphs are obtained as shown in Figure 6, which include the following 14 genes: ITG1, HSPA5, ACTB, ATP5F1B, CCT5, GAPDH, ARPC1A, GNA12, ARFGAP1, INS, CDC42, BSG, ARFGAP3, and ARF1. Among these 14 genes, it has recently been reported that ARFGAP1 is a predictive factor for pancreatic cancer [35], and GAPDH has been used as a reference gene in fetal rat pancreas analysis [36]. In Figure 7, only the androgen receptor (AR) transcription factor was reported by Metascape. It has been reported recently that AR plays a key role in the progression of pancreatic cancer [37]. These results demonstrate that the inhibition of AR and the understanding of these 14 genes can act as a therapeutic target in pancreatic cancer. In terms of dataset 2, a list of genes obtained via each method is available in Supplementary Data Sheet 2. Additionally, a full list of enrichment analysis results of each method is available in Supplementary Table 2.

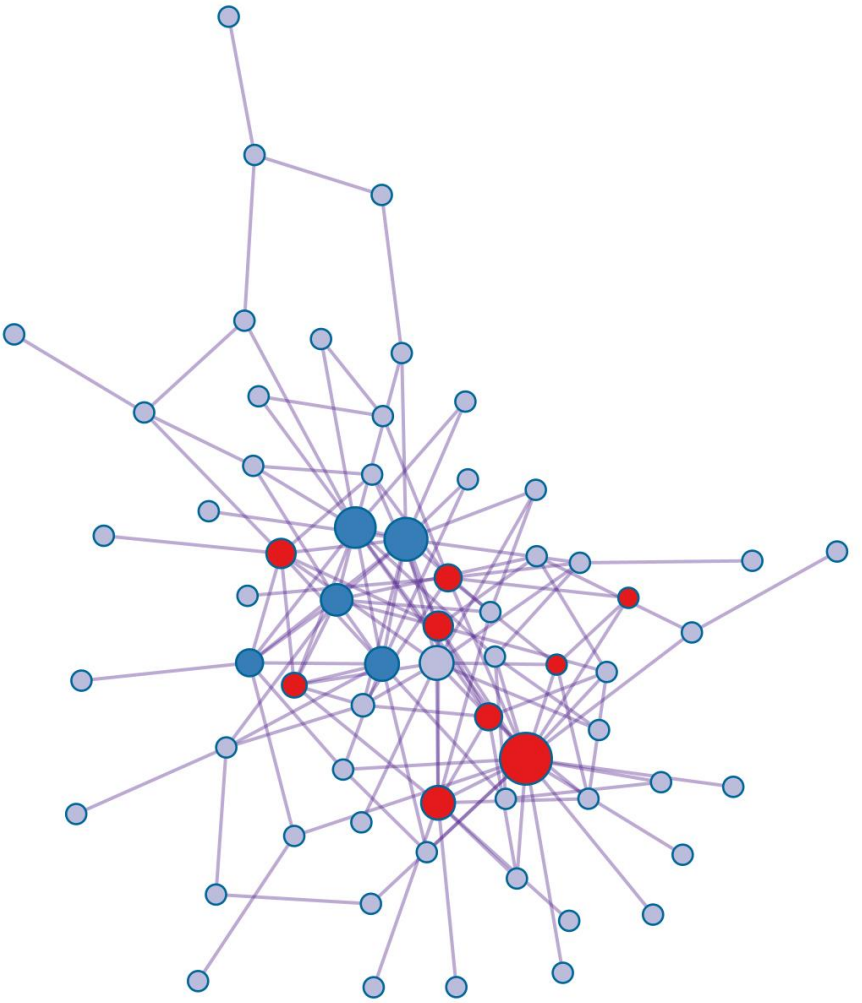


Figure 5: Protein-Protein interaction network provided by metascape for enriched genes of scsvm2

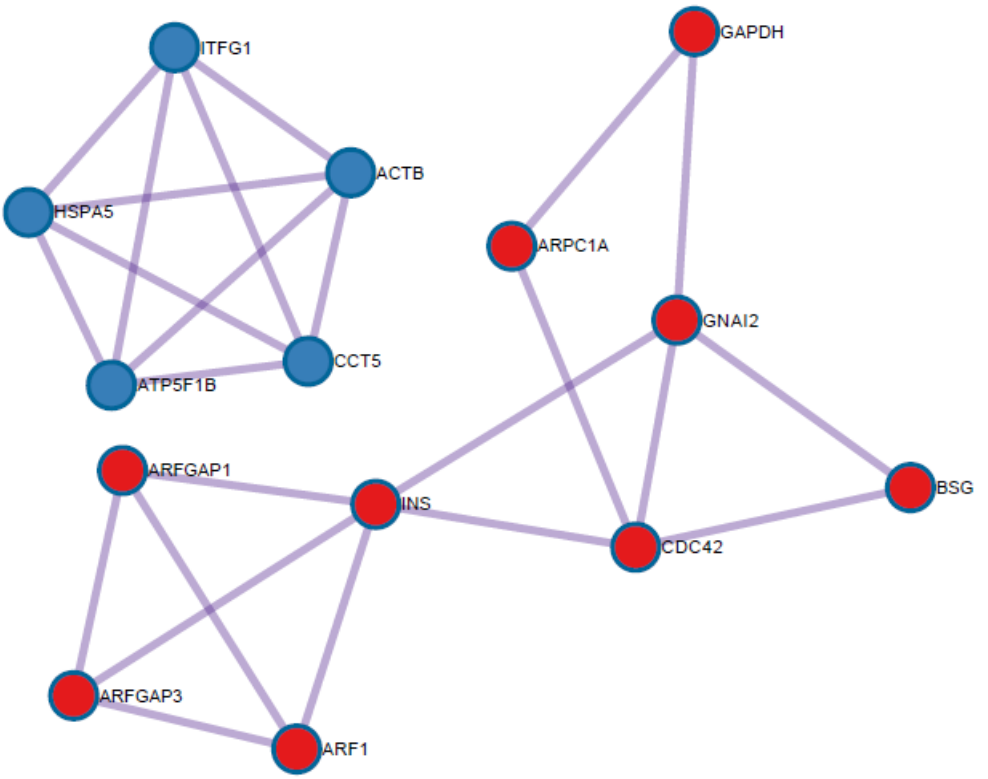


Figure 6: Two highly interconnected regions (clusters) in a protein-protein interaction enrichment analysis according to 95 genes of scsvm2, submitted to metascape.

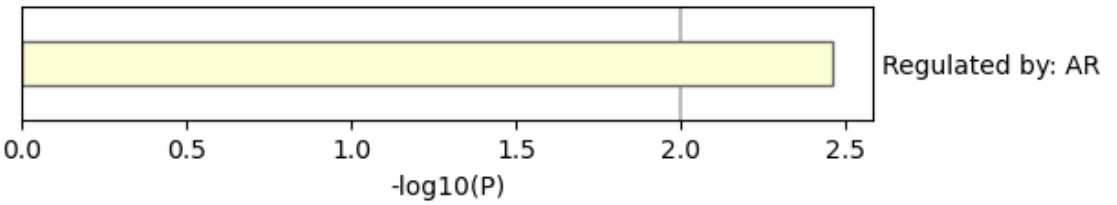


Figure 7: A Transcription factor regulating the submitted genes by scsvm2 to metascape.

4.2.3 Dataset3

Table 3 reports the top retrieved terms for each method by Enrichr. It is clear that scsvm2 has more expressed genes (and good ranking results) in pancreatic islet and alpha cell terms when compared to all other methods. In particular, 27 out of 2316 genes were expressed in pancreatic islets. Totally 20 and 21

genes were expressed in beta and alpha cells, respectively. Seventeen genes (RBPJL, CELA3A, EGR1, PRSS1, CPA1, CELA3B, JUN, SPINK1, CTRB2, CTRB1, GP2, GCG, MTRNR2L8, DLK1, CLPS, ID4, and FOSB) were commonly expressed in beta and alpha cells. In terms of beta cell, sam had 22 expressed genes (KLF10, CELA3A, PRSS1, CDKN1A, SPINK1, KIRREL2, DTNA, KLK1, REG1A, CTRB1, SLC16A12, CEL, SLC39A14, TMC4, SERPINA4, DLK1, NR4A1, SCG5, GNAS, FAT1, CELP, and KCTD16). However, the sam ranking was worse than that of scsvm2 and scsvm1. These results demonstrate the superiority of scsvm2 in identifying pancreatic cell types within the pancreatic islet.

Table 4: Enriched terms in "ARCHS4 Tissues" using Enrichr based on genes obtained from Dataset3, consisting of acinar and delta pancreatic cells. Rank column shows the order of terms when retrieved. Overlap displays the number of common genes between the uploaded gene and those in the term divided by the number of genes in that term. Best results of a method is in bold.

Method	Rank	Term	Overlap	P-value	Adjusted P-value
scsvm1	1	PANCREATIC ISLET	25/2316	5.74×10^{-5}	5.75×10^{-3}
	10	BETA CELL	19/2316	1.19×10^{-2}	1.17×10^{-1}
	8	ALPHA CELL	20/2316	5.61×10^{-3}	6.73×10^{-2}
scsvm2	1	PANCREATIC ISLET	**27**/2316	6.36×10^{-6}	6.87×10^{-4}
	8	BETA CELL	20/2316	5.61×10^{-3}	6.73×10^{-2}
	6	ALPHA CELL	**21**/2316	2.50×10^{-3}	3.86×10^{-2}
baseline	1	PANCREATIC ISLET	24/2316	1.60×10^{-4}	8.62×10^{-3}
	21	BETA CELL	15/2316	1.32×10^{-1}	5.70×10^{-1}
	11	ALPHA CELL	18/2316	2.37×10^{-2}	2.13×10^{-1}
limma	18	PANCREATIC ISLET	10/2316	6.74×10^{-1}	9.99×10^{-1}
	24	BETA CELL	9/2316	7.85×10^{-1}	9.99×10^{-1}
	72	ALPHA CELL	6/2316	9.70×10^{-1}	9.99×10^{-1}
sam	25	PANCREATIC ISLET	19/2316	1.19×10^{-2}	4.93×10^{-2}
	6	BETA CELL	**22**/2316	1.06×10^{-3}	1.90×10^{-2}
	12	ALPHA CELL	20/2316	5.61×10^{-3}	3.36×10^{-2}
ttest	34	PANCREATIC ISLET	13/2316	3.04×10^{-1}	8.42×10^{-1}

77	BETA CELL	8/2316	8.73×10^{-1}	1.00×10^{0}
76	ALPHA CELL	8/2316	8.73×10^{-1}	1.00×10^{0}

Because scsvm2 achieved good performance results, we uploaded the 95 genes obtained from scsvm2 to Metascape. Figure 8 shows the PPI network obtained by Metascape. In Figure 9, four highly connected subgraphs were extracted from the PPI network in Figure 8. These four subgraphs include the following 18 genes: COL1A2, COL1A1, SPARC, COL3A1, SEC24B, CD59, SERPINA1, TMED10, LAMB2, SPP1, HSP90B1, RPS19, EEF1A1, RPL3, RPLP2, RPS12, UBC, and RPS27. Among these genes, Lin et al. [38] suggested that COL1A1 can act as a biomarker and therapeutic target for type 2 diabetes. RPS12 and RPS27 have played a key role in the differentiation and proliferation of pancreatic cancer cells [39].

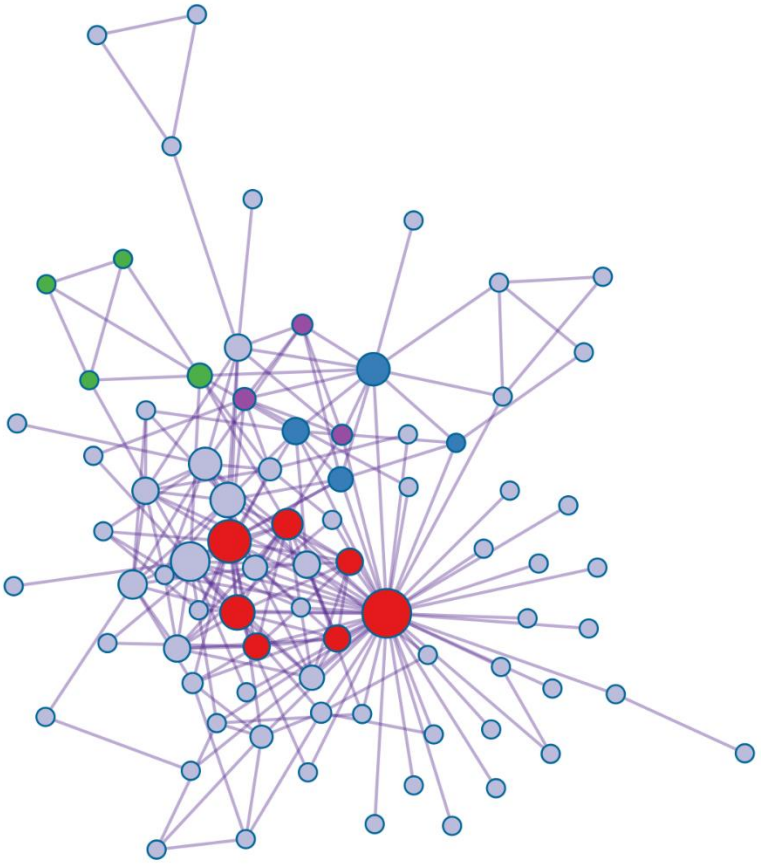


Figure 8: Protein-Protein interaction network provided by metascape for enriched genes of scsvm2.

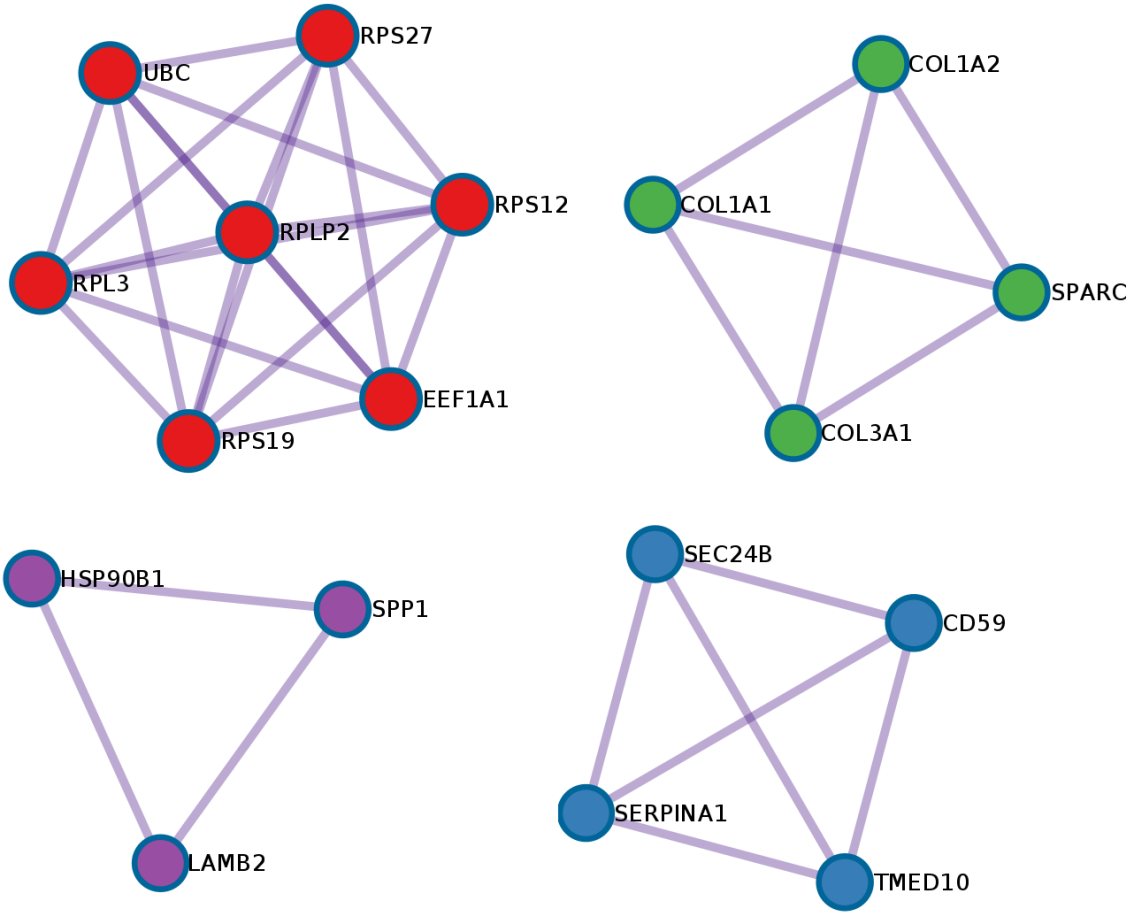


Figure 9: Four highly interconnected regions (clusters) in a protein-protein interaction enrichment analysis according to 95 genes of scsvm2, submitted to metascape.

In Figure 10, we identified seven TFs, including HDAC1, STAT6, RELA, SP3, NFKB1, SP1, and ESR1. Two of these seven TFs were reported in Figure 4, namely HDAC1 and SP1. Furthermore, SP3 and NFKB1 have been reported strongly correlated with RAP1 expression, which plays a key role in multiple functions pertaining to pancreatic islet cells [40-42]. For Dataset 3, a list of genes obtained by each method is provided in Supplementary Data Sheet 3. Moreover, a full list of enrichment analysis results of each method is deposited in Supplementary Table 3.

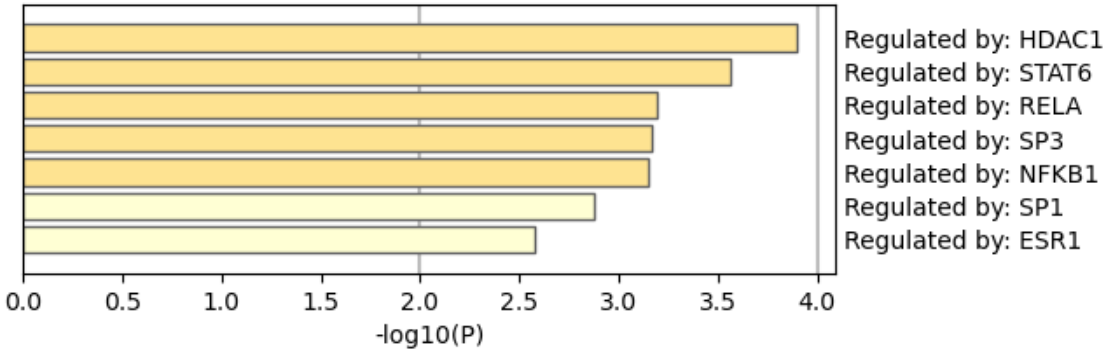


Figure 10: Seven Transcription factor regulating the submitted genes by scsvm2 to metascape.

4.2.4 Demonstrating the Discriminative Power

Consider that we are given a dataset pertaining to two cell types (i.e., alpha and beta cells) as shown in Figure 11 (a). Our task is to provide this dataset to a machine-learning algorithm to induce a model, which is then utilized for predicting the types of unseen cells. It can be shown from Figure 11 (c, d) that scsvm1 and scsvm2 outperform the baseline (Figure 11 b) by learning models that make no errors. Furthermore, scsvm1 (p-value = 2.17×10^{-3}) and scsvm2 (p-value = 5.50×10^{-3}) generate statistically significant results (Figure 12 b, c). Because the baseline in Figure 11 (b) has one correctly predicted alpha cell and another incorrectly classified beta cell, no p-value was reported, as there were not enough examples in the category. In terms of measuring the classification performance, using the area under curve (AUC) performance measure, it can be shown from Figure 12 (d) that scsvm1 and scsvm2 achieved the highest results, 100%, when compared to the baseline, which generated 75%. These results demonstrate the superiority of the proposed learning algorithms, scsvm1 and scsvm2.

(a) (b) (c) (d)

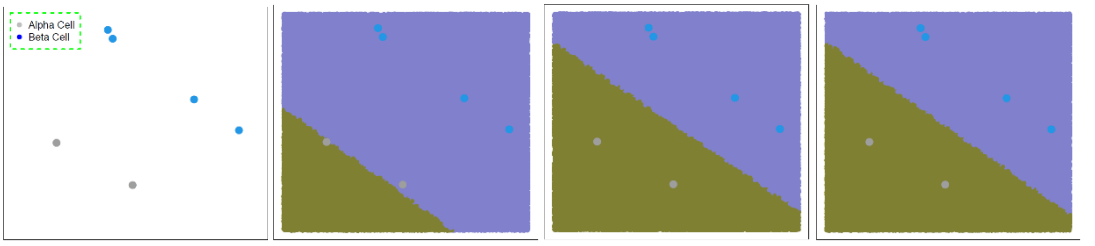
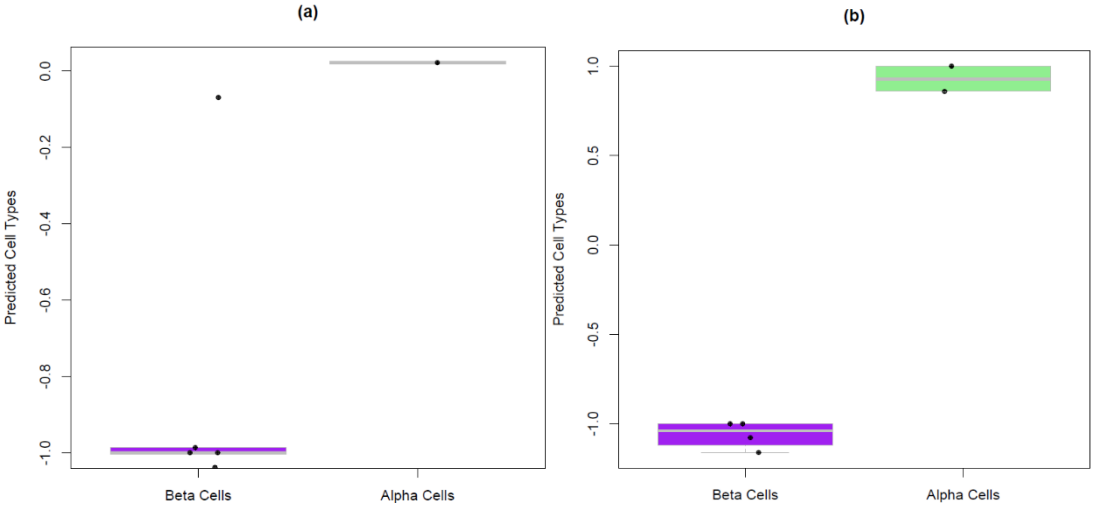


Figure 11: (a) Dataset, (b) decision boundary obtained by the baseline, (c-d) decision boundaries obtained by the scsvm1 and scsvm2.



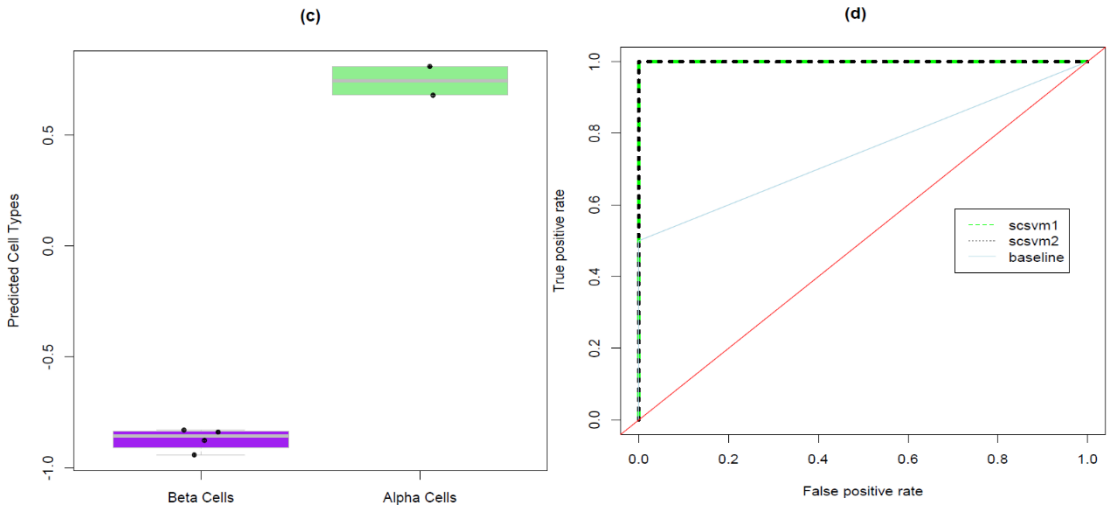


Figure 12: Prediction results pertaining to alpha and beta cell types using dataset in Figure 11(a). Boxplot and strip charts (a-c) show the differences predicted cell types, beta or alpha, for the baseline, scsvm1, and scsvm2, respectively. (d) shows the ROC curves for the baseline, scsvm1 and scsvm2 by displaying the true positive rate against the false positive rate of each model.

Additionally, assume that we have another dataset as shown in Figure 13, where the task is to provide this dataset to a machine learning algorithm, to generate a model, which then takes new examples to predict the cell type. It can be seen from Figure 13 (d) that the induced model scsvm2 classifies all examples correctly. The second-best model is scsvm1, which incorrectly classifies one out of five beta cells as an alpha cell. In contrast, the baseline was the worst among the three, and incorrectly classified two out of five beta cells as alpha cells. Results obtained by scsvm1 and scsvm2 were statistically significant, where the former had the *p*-value = 2.846×10^{-3}, and the latter had the *p*-value = 2.842×10^{-3} (see Figures 14 b, c). Although the baseline generated highly statistically significant results (*p*-value $=1.865\times10^{-5}$), the performance result was the worst as it achieved an AUC of 80% which was quite low compared to scsvm1 and scsvm2 which generated an AUC of 90 and 100%, respectively (See Figure 14 d).

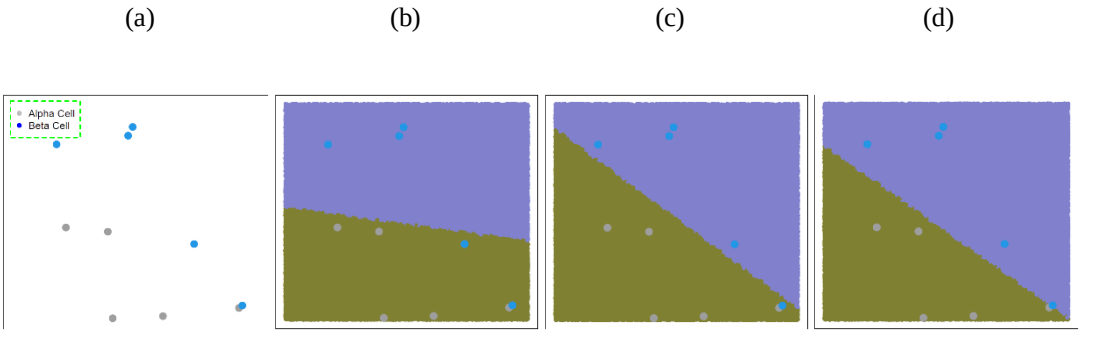
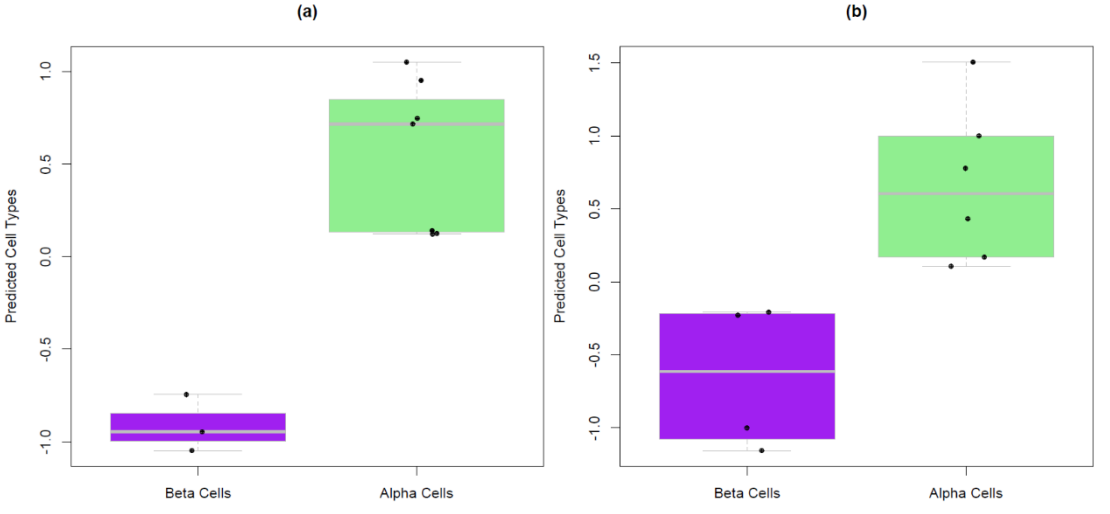


Figure 13: (a) Dataset, (b) decision boundary obtained by the baseline, (c-d) decision boundaries obtained by the scsvm1 and scsvm2, respectively.



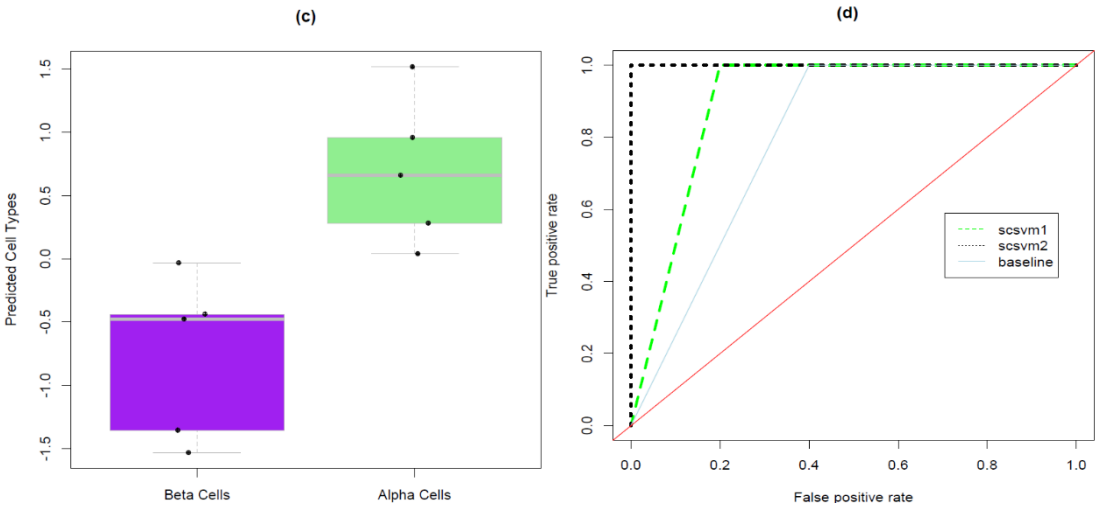


Figure 14: Prediction results pertaining to alpha and beta cell types using dataset in Figure 13(a). Boxplot and strip charts (a, c) show the differences predicted cell types, beta or alpha, for the baseline, scsvm1, and scsvm2, respectively. (d) shows the ROC curves for the baseline, scsvm1 and scsvm2 by displaying the true positive rate against the false positive rate of each model.

5. Discussion

Therapeutic targets pertaining to the pancreas provide a deep biological insight into its complex molecular mechanisms. Therefore, advanced computational methods can aid in speeding up the drug discovery process by providing a basis for the development of drugs for pancreatic disorders. Therefore, we proposed a computational framework to identify influential genes from three expression-profiling datasets. We introduced a new constrained optimization version of SVM and solved an optimization problem using the CVXR package in R, to get weights w, which correspond to the importance of genes in the dataset. The higher the weight, the more important the gene. Then, we selected the top 95 genes and performed enrichment analysis. In terms of cell types, the top retrieved terms by our methods (scsvm1 and scsvm2) consisted of more expressed genes in the pancreatic islet cells, including beta and alpha islet cells. Moreover, from the highly connected subgraphs within the PPI networks and TFs, we identified important genes that have been reported to be related to pancreatic islet functions, diseases, and treatments.

In terms of the results (See Tables 2-4), our computational framework (when coupled with enrichment analysis) acts as a search engine which retrieves the most relevant terms (e.g., pancreatic islet cells) out of many terms. In the context of a search engine, genes and terms can be thought of as contents and pages,

respectively. The retrieved terms are related to the selected genes. If the genes in question are not related to pancreatic islet cells, then the retrieved terms are not related to the pancreas. It can be seen from the results that the lower the ranking with more expressed genes, the better the performance. Therefore, the results obtained by our computational framework are better than those obtained from other methods are.

For the constrained optimization problem in Equation 4, we set $u = 0.5$, $m = 0.1$, and $c = 1$. The only difference is k, which is set to 1 in the scsvm1, while it is set to 2 in scsvm2. In terms of solving the problem, we defined the objective function and constraints, then used the 'Problem' and 'Solve' functions in the CVXR package, to define and solve the optimization problem [25]. In Dataset 2, It is worth pointing out that sam and t test provide five and three genes, respectively, for further enrichment analysis. This is because the selection of genes in these methods (including limma) is based on p-values < 0.01 (See list of genes in Supplementary Data Sheet 2). Therefore, fewer genes were selected, which would meet this requirement [43, 44].

The strength of the proposed computational framework relies heavily on the new variant of the constrained optimization problem for SVM. As can be seen in Figures 11-14, our methods generate better performance results when compared to the baseline. In addition, our method is more robust as it finds the maximum marginal hyperplane and conceals the effects of slack variables, which allow for some training errors in the baseline (See Figures 11 and 13). If slack variables are not introduced, hard margins are created in SVM, producing no solution when training data is not linearly separable. Moreover, for Datasets 1-3, the probabilities of any method coinciding with other methods in terms of the 95 selected genes are $\frac{1}{\binom{40916}{95}}$, $\frac{1}{\binom{23431}{95}}$, and $\frac{1}{\binom{22762}{95}}$, respectively, where $\binom{n}{k}$ is the binomial coefficient, and n and k are the number of genes and selected genes, respectively. Therefore, the reproduction of the exact same results as reported in our study is highly unlikely.

6. Conclusion and Future Work

We present a computational framework for identifying influential genes related to pancreatic islet functions and diseases. The framework works by processing expression profiles pertaining to pancreatic islet cells, which are then provided to a new constrained optimization of SVM, to quantify the importance of genes in these transcriptomic expression profiles. Three transcriptome expression profile datasets (GSE87375, GSE79457, and GSE110154) were downloaded from the GEO database. Then, an enrichment analysis was performed on the top p selected genes, comparing proposed methods against existing methods, which include the baseline. Results indicate that our methods outperformed all other methods by retrieving the top terms with more expressed genes in pancreatic islet cells. Furthermore, we

identified 44 genes and 10 TFs (SP1, HDAC1, EGR1, E2F1, AR, STAT6, RELA, SP3, NFKB1, and ESR1) which have been related to pancreatic functions, diseases and therapeutic targets. In terms of the assessment from a classification perspective, the proposed methods utilizing the new variant also performed better than the baseline, achieving significant results when compared to the baseline.

We suggest that similar strategies should be implemented to (1) utilize the computational framework to unveil molecular mechanisms and find therapeutic targets in various cells, including the liver, heart, and lung; (2) collaborate with hospitals to improve the poor therapeutic outcomes for diseases; and (3) derive gene signatures from expression-profiling data using our new SVM variant, for assessment in various biological tasks.

Acknowledgment

This project was funded by the Deanship of Scientific Research (DSR) at King Abdulaziz University, Jeddah, under grant no. G: 028-611-1442. The authors, therefore, acknowledge with thanks DSR for technical and financial support.

Conflict of interest

None declared.

References

1. Benninger, R.K. and V. Kravets, *The physiological role of β-cell heterogeneity in pancreatic islet function.* Nature Reviews Endocrinology, 2022. **18**(1): p. 9-22.
2. Yu, X.-X., et al., *Sequential progenitor states mark the generation of pancreatic endocrine lineages in mice and humans.* Cell Research, 2021. **31**(8): p. 886-903.
3. Bahl, V., et al., *Genetic activation of α-cell glucokinase in mice causes enhanced glucose-suppression of glucagon secretion during normal and diabetic states.* Molecular metabolism, 2021. **49**: p. 101193.
4. Wardman, J.F., et al., *Carbohydrate-active enzymes (CAZymes) in the gut microbiome.* Nature Reviews Microbiology, 2022: p. 1-15.
5. James, D.E., J. Stöckli, and M.J. Birnbaum, *The aetiology and molecular landscape of insulin resistance.* Nature Reviews Molecular Cell Biology, 2021. **22**(11): p. 751-771.
6. Xia, H., et al., *Insulin action and resistance are dependent on a GSK3β-FBXW7-ERRα transcriptional axis.* Nature Communications, 2022. **13**(1): p. 1-19.
7. Kleiber, T., et al., *Single cell transcriptomics reveal trans-differentiation of pancreatic beta cells following inactivation of the TFIID subunit Taf4.* Cell death & disease, 2021. **12**(8): p. 1-12.

8. Wagner, R., et al., *Pathophysiology-based subphenotyping of individuals at elevated risk for type 2 diabetes.* Nature medicine, 2021. **27**(1): p. 49-57.
9. Zheng, X., et al., *Integrated bioinformatics analysis identified COL11A1 as an immune infiltrates correlated prognosticator in pancreatic adenocarcinoma.* International Immunopharmacology, 2021. **90**: p. 106982.
10. Prashanth, G., et al., *Identification of hub genes related to the progression of type 1 diabetes by computational analysis.* BMC Endocrine Disorders, 2021. **21**(1): p. 1-65.
11. Kang, J., et al., *Identification of Shared Genes and Pathways in Periodontitis and Type 2 Diabetes by Bioinformatics Analysis.* Frontiers in endocrinology, 2021. **12**.
12. Chan, K.H.K., et al., *Shared molecular pathways and gene networks for cardiovascular disease and type 2 diabetes mellitus in women across diverse ethnicities.* Circulation: Cardiovascular Genetics, 2014. **7**(6): p. 911-919.
13. Islam, S., et al., *ITGA2, LAMB3, and LAMC2 may be the potential therapeutic targets in pancreatic ductal adenocarcinoma: an integrated bioinformatics analysis.* Scientific reports, 2021. **11**(1): p. 1-16.
14. Shi, T. and G. Gao, *Identify potential prognostic indicators and tumor-infiltrating immune cells in pancreatic adenocarcinoma.* Bioscience Reports, 2022. **42**(2): p. BSR20212523.
15. Wu, C.-C., et al., *Prediction of biomarkers and therapeutic combinations for anti-PD-1 immunotherapy using the global gene network association.* Nature communications, 2022. **13**(1): p. 1-14.
16. Purfield, D., R. Evans, and D. Berry, *Reaffirmation of known major genes and the identification of novel candidate genes associated with carcass-related metrics based on whole genome sequence within a large multi-breed cattle population.* BMC genomics, 2019. **20**(1): p. 1-17.
17. Kim, K.-S., et al., *Identification of novel biomarker for early detection of diabetic nephropathy.* Biomedicines, 2021. **9**(5): p. 457.
18. Scholkopf, B. and A.J. Smola, *Learning with kernels: support vector machines, regularization, optimization, and beyond*. 2018: MIT press.
19. Ritchie, M.E., et al., *limma powers differential expression analyses for RNA-sequencing and microarray studies.* Nucleic acids research, 2015. **43**(7): p. e47-e47.
20. Tusher, V.G., R. Tibshirani, and G. Chu, *Significance analysis of microarrays applied to the ionizing radiation response.* Proceedings of the National Academy of Sciences, 2001. **98**(9): p. 5116-5121.
21. Huber, P.J., *Robust estimation of a location parameter*, in *Breakthroughs in statistics*. 1992, Springer. p. 492-518.
22. Qiu, W.-L., et al., *Deciphering pancreatic islet β cell and α cell maturation pathways and characteristic features at the single-cell level.* Cell metabolism, 2017. **25**(5): p. 1194-1205. e4.
23. Chakravarthy, H., et al., *Converting adult pancreatic islet α cells into β cells by targeting both Dnmt1 and Arx.* Cell metabolism, 2017. **25**(3): p. 622-634.
24. Barrett, T., et al., *NCBI GEO: archive for functional genomics data sets—update.* Nucleic Acids Research, 2012. **41**(D1): p. D991-D995.
25. Fu, A., B. Narasimhan, and S. Boyd, *CVXR: An R Package for Disciplined Convex Optimization.* Journal of Statistical Software, 2020. **94**(14): p. 1 - 34.
26. Kuleshov, M.V., et al., *Enrichr: a comprehensive gene set enrichment analysis web server 2016 update.* Nucleic acids research, 2016. **44**(W1): p. W90-W97.
27. Zhou, Y., et al., *Metascape provides a biologist-oriented resource for the analysis of systems-level datasets.* Nature communications, 2019. **10**(1): p. 1-10.
28. Team, R.C., *R: A Language and Environment for Statistical Computing.* Journal of Statistical Software, 2008. **25**(1).

29. Schwender, H., *siggenes: Multiple testing using SAM and Efron's empirical Bayes approaches.* R package version, 2012. **1**(0): p. 1-70.
30. Yang, C.-H., et al., *Neuropeptide Y1 receptor antagonism protects β-cells and improves glycemic control in type 2 diabetes.* Molecular metabolism, 2022. **55**: p. 101413.
31. de Souza Marinho, T., et al., *Pancreatic islet cells disarray, apoptosis, and proliferation in obese mice. The role of Semaglutide treatment.* Biochimie, 2022. **193**: p. 126-136.
32. Jiang, N.Y., et al., *Sp1, a new biomarker that identifies a subset of aggressive pancreatic ductal adenocarcinoma.* Cancer Epidemiology and Prevention Biomarkers, 2008. **17**(7): p. 1648-1652.
33. Denechaud, P.-D., L. Fajas, and A. Giralt, *E2F1, a novel regulator of metabolism.* Frontiers in endocrinology, 2017. **8**: p. 311.
34. Wang, B., et al., *The role of the transcription factor EGR1 in cancer.* Frontiers in Oncology, 2021. **11**: p. 775.
35. Meng, D., et al., *ArfGAP1 inhibits mTORC1 lysosomal localization and activation.* The EMBO journal, 2021. **40**(12): p. e106412.
36. Dai, Y., et al., *Identification and validation of reference genes for RT-qPCR analysis in fetal rat pancreas.* Reproductive Toxicology, 2021. **105**: p. 211-220.
37. Fan, Y.-C., et al., *Investigation of anti-tumor effects of an MLK1 inhibitor in prostate and pancreatic cancers.* Biology, 2021. **10**(8): p. 742.
38. Lin, G., et al., *COL1A1 as a potential new biomarker and therapeutic target for type 2 diabetes.* Pharmacological Research, 2021. **165**: p. 105436.
39. Reza, A.M.M.T. and Y.-G. Yuan, *microRNAs Mediated Regulation of the Ribosomal Proteins and its Consequences on the Global Translation of Proteins.* Cells, 2021. **10**(1): p. 110.
40. Kelly, P., et al., *Rap1 promotes multiple pancreatic islet cell functions and signals through mammalian target of rapamycin complex 1 to enhance proliferation.* Journal of Biological Chemistry, 2010. **285**(21): p. 15777-15785.
41. Kurashima, Y., et al., *Pancreatic glycoprotein 2 is a first line of defense for mucosal protection in intestinal inflammation.* Nature communications, 2021. **12**(1): p. 1-17.
42. Hossain, S.M.M., et al., *Discovering key transcriptomic regulators in pancreatic ductal adenocarcinoma using Dirichlet process Gaussian mixture model.* Scientific reports, 2021. **11**(1): p. 1-15.
43. Taguchi, Y. and T. Turki, *A new advanced in silico drug discovery method for novel coronavirus (SARS-CoV-2) with tensor decomposition-based unsupervised feature extraction.* PloS one, 2020. **15**(9): p. e0238907.
44. Taguchi, Y. and T. Turki, *Novel feature selection method via kernel tensor decomposition for improved multi-omics data analysis.* BMC medical genomics, 2022. **15**(1): p. 1-12.

Turki Turki received a B.S. in computer science from King Abdulaziz University, an M.S. in computer science from NYU.POLY, and a Ph.D. in computer science from the New Jersey Institute of Technology. He is currently an associate professor with the Department of Computer Science, King Abdulaziz University, Saudi Arabia. His research interests include machine learning, data science and bioinformatics. His research has been published in journals such as *Expert Systems with Applications*, *IEEE Journal on Selected Topics in Signal Processing*, *BMC Medical Genomics and Computers in Biology and Medicine*. He is an editorial board member of *Sustainable Computing: Informatics and Systems*, *BMC Medical Genomics*, *PLOS ONE*, and *Computers in Biology and Medicine*.

Y-h. Taguchi received a B.S. degree in physics from the Tokyo Institute of Technology and a Ph.D. degree in physics from the Tokyo Institute of Technology. He is currently a full professor with the Department of Physics, Chuo University, Japan. His works have been published in leading journals such as *Physical Review Letters*, *Bioinformatics*, and *Scientific Reports*. His research interests include bioinformatics, machine learning, and nonlinear physics. He is also an editorial board member of *PloS ONE*, *BMC Medical Genomics*, *Medicine* (Lippincott Williams & Wilkins journal), *BMC Research Notes*, and *IPSJ Transaction on Bioinformatics*.