frontiers in Endocrinology

PERSPECTIVE
published: 23 March 2022
doi: 10.3389/fendo.2022.853863

Analysis of Half a Billion Datapoints Across Ten Machine-Learning Algorithms Identifies Key Elements Associated With Insulin Transcription in Human Pancreatic Islet Cells

Wilson K. M. Wong[1†], *Vinod Thorat*[2†], *Mugdha V. Joglekar*[1†], *Charlotte X. Dong*[1], *Hugo Lee*[3], *Yi Vee Chew*[4], *Adwait Bhave*[2], *Wayne J. Hawthorne*[4], *Feyza Engin*[3,5], *Aniruddha Pant*[2], *Louise T. Dalgaard*[6], *Sharda Bapat*[2*‡] and *Anandwardhan A. Hardikar*[1,6*‡]

[1] Diabetes and Islet Biology Group, School of Medicine, Western Sydney University, Campbelltown, NSW, Australia, [2] Healthcare Analytics, AlgoAnalytics, Pune, India, [3] Department of Biomolecular Chemistry, School of Medicine and Public Health, University of Wisconsin-Madison, Madison, WI, United States, [4] Centre for Transplant and Renal Research, Westmead Institute for Medical Research, University of Sydney, Westmead, NSW, Australia, [5] Division of Endocrinology, Diabetes & Metabolism, Department of Medicine, School of Medicine and Public Health, University of Wisconsin-Madison, Madison, WI, United States, [6] Department of Science and Environment, Roskilde University, Roskilde, Denmark

Machine learning (ML)-workflows enable unprejudiced/robust evaluation of complex datasets. Here, we analyzed over 490,000,000 data points to compare 10 different ML-workflows in a large (N=11,652) training dataset of human pancreatic single-cell (sc-) transcriptomes to identify genes associated with the presence or absence of insulin transcript(s). Prediction accuracy/sensitivity of each ML-workflow was tested in a separate validation dataset (N=2,913). Ensemble ML-workflows, in particular Random Forest ML-algorithm delivered high predictive power (AUC=0.83) and sensitivity (0.98), compared to other algorithms. The transcripts identified through these analyses also demonstrated significant correlation with insulin in bulk RNA-seq data from human islets. The top-10 features, (including *IAPP, ADCYAP1, LDHA* and *SST*) common to the three Ensemble ML-workflows were significantly dysregulated in scRNA-seq datasets from Ire-1$\alpha^{\beta-/-}$ mice that demonstrate dedifferentiation of pancreatic β-cells in a model of type 1 diabetes (T1D) and in pancreatic single cells from individuals with type 2 Diabetes (T2D). Our findings provide direct comparison of ML-workflows in big data analyses, identify key elements associated with insulin transcription and provide workflows for future analyses.

Keywords: machine-learning (ML) algorithms, insulin, diabetes, beta-cell, single-cell RNA-sequencing (scRNAseq), human islet

Edited by:
Vikash Chandra,
University of Helsinki, Finland

Reviewed by:
OM Prakash Dwivedi,
University of Helsinki, Finland

****Correspondence:***
Sharda Bapat
sbapat@algoanalytics.com
Anandwardhan A. Hardikar
A.Hardikar@westernsydney.edu.au

†These authors have contributed equally to this work and share first authorship

‡These authors have contributed equally to this work and share last authorship

Specialty section:
This article was submitted to Diabetes: Molecular Mechanisms, a section of the journal Frontiers in Endocrinology

Received: *13 January 2022*
Accepted: *22 February 2022*
Published: *23 March 2022*

Citation:
Wong WKM, Thorat V, Joglekar MV, Dong CX, Lee H, Chew YV, Bhave A, Hawthorne WJ, Engin F, Pant A, Dalgaard LT, Bapat S and Hardikar AA (2022) Analysis of Half a Billion Datapoints Across Ten Machine-Learning Algorithms Identifies Key Elements Associated With Insulin Transcription in Human Pancreatic Islet Cells. Front. Endocrinol. 13:853863. doi: 10.3389/fendo.2022.853863

INTRODUCTION

Recent years have witnessed a surge in single-cell transcriptomic technologies; many already generating newer data and insights to address specific biological questions. Machine learning (ML) algorithms offer an unbiased mathematical workflow that facilitates the identification of complex relationships across variables. ML workflows involve an orderly set of instructions using automated, unbiased 'learning' processes usually targeted towards developing (training) a model that can be validated in a separate (test) dataset (1). One goal of ML algorithms is to analyze big data to identify variables that cannot be recognized through conventional biostatistical techniques, and enhance development of predictive algorithms (2, 3).

Currently, several ML algorithms are available to researchers handling big data in omics-based high content analyses. These can be broadly divided into two categories: supervised and unsupervised algorithms (4). Supervised methods (such as decision tree) derive relationships between one dependent and multiple independent variables using a training set and then apply that knowledge in the testing set for predictive/efficacy analysis. Unsupervised methods derive patterns/data clusters amongst all available variables. ML algorithms have been used to unravel patterns/clustering in high-density transcriptome analyses (5) or to build associations (6) or for predictions in several biological processes such as determining DNA methylation states in single cells (7), identifying signatures of lipid or metabolite species (8) or microRNAs (9) in predicting transition from gestational diabetes to type 2 diabetes as well as in genetic studies (10). There are multiple ML algorithms available and it may present a challenge to select the most appropriate method for a particular dataset to answer a specific question. We, therefore, decided to compare different ML methodologies to (i) rank different ML methods for their performance on a large dataset (of 490,855,065 scRNA-sequencing data points) and (ii) understand the most important variables associated with insulin transcription.

Previous studies (11–14) from several laboratories have identified master regulatory transcription factors that regulate the embryonic development of insulin-producing islet β-cells. Although transcription factor-mediated insulin transcription regulation is a well-known mechanism during the development of insulin-producing cells, it is also recognized that active genes localized on different chromosomal regions can dynamically regulate gene transcription in post-natal life (15). One approach to identify genes associated with insulin gene transcription is through single-cell (sc)RNA-seq-based big data analysis.

Here, we examined the performance of 10 different ML algorithms in a curated human pancreatic single-cell sequencing dataset of 490,855,065 data points (N=14,565 single cells and 33,701 expressed gene features). The aims of this study were (i) to provide a comparative account of the predictive potential of 10 different commonly used ML workflows (**Supplementary Table 1**), and (ii) to use existing scRNA-seq datasets in identifying genes (variables) associated with or important for determining insulin transcript-containing cells.

MATERIALS AND METHODS

Pancreatic Single-Cell (sc)RNA Sequencing Datasets and Analyses

Human Pancreatic Single-Cell Sequencing Datasets

The pancreatic single-cell sequencing dataset (N=14,890) was extracted using the Panc8 data (16) containing multiple publicly available scRNA-seq transcriptomes (GSE84133, GSE85241, E-MTAB-5061,GSE81076, GSE86469). The original publications citing the listed GEO datasets (GSE84133, GSE85241, E-MTAB-5061, GSE81076, GSE86469) add up to a total of 31 pancreas samples across all the studies. Clinical and/or donor details are available for 26 of these samples; seven of which were indicated to be from donors with type 2 diabetes. The number of cell types in the combined single cell dataset (Panc8) can be found in the metadata of the SeuratData (version 0.2.1), using the command "panc8@meta.data" (in R studio version 1.2.5033). This panc8 dataset contains scRNA-seq data from acinar (n=1864), activated stellate (n=474), alpha-(n=4615), beta-(n=3679), delta-(n=1013), ductal (n=1954), endothelial (n=296), epsilon (n=30), gamma (n=625), macrophage (n=79), mast (n=56), quiescent stellate (n=180) and schwann (n=25) cell transcriptomes. Analysis was carried out by using R studio version 1.2.5033 as detailed in SOM.

Ire1α$^{β-/-}$ Mouse Pancreatic Single-Cell Dataset

Single-cell RNA-seq dataset from pancreatic islets of Ire1α$^{fl/fl}$ (N=1,163 single-cell transcriptomes from one mouse) and Ire1α$^{β−/−}$ (N=1,683 single-cell transcriptomes from two mice) were obtained through GSE144471 (17). The β-cells (Ire1α$^{fl/fl}$: 830 cells; Ire1α$^{β−/−}$: 816 cells) were separated from the dataset and the expression values of selected genes were evaluated in the β-cell population.

T2D Pancreatic Single-Cell (sc)RNA Sequencing Dataset

Pancreatic single cell normalized read dataset of adult ND (with no diabetes; N=4) and T2D (N=10) donors were obtained from GSE154126 (18). The adult ND (N=296) and T2D (N=505) insulin transcribing cells were compared and used for validation. In this dataset, insulin-transcribing cells were identified and defined as any single cell that contained (non-zero) *INS* transcript.

Human Pancreatic Single-Cell Sequencing Classification and Analyses

Deidentified datasets were shared with data scientists. A random number generator function was used to allocate 80% of samples to a training set. Analyses were carried using Python (Ver:3.4), wherein the data was imported, transposed, edited to delete INS and INS-IGF2 columns from the data frame and labeled (label=0 where INS=0 and label=1 where INS>0). Classifiers were initialized and model trained using the discovery (80%) data set. Predictive analyses were then carried out on the validation (20%) set and the resulting accuracy metrics were saved to compare the feature importance. Selected classifiers (Random

Forest, Gradient Boosting, Decision Tree Classifier, Logistic Regression, Multinomial Naive Bayes Classifier, ADA Boost Classifier, Linear Discriminant Analysis, Ridge Classifier, KNeighbors Classifier and Linear Support Vector Classifier) were used on the same set.

Pancreatic Islet RNA Sequencing Dataset and Analysis

Human Pancreatic Islet Bulk RNA Sequencing Dataset

Human pancreatic islet RNA-seq dataset was obtained from GSE152111 (6). RNA-seq dataset contains n=66 human islet samples, across 65 different donors with no diabetes. Two of the 66 RNA-seq samples were duplicates from the same donor and their RNA-seq profile highly correlative (Pearson r=0.99) to each other. The average of the duplicates of this donor was calculated prior to analysis. Data was analyzed in DEseq values. DEseq values are the normalized RNA-seq. DEseq compares the different read depths between samples by estimating the effective library size (using the estimate size factors function). The size factor for each sample is the median raw count of a gene's geometric mean across all samples. DEseq normalization involves dividing the raw count of a gene in a sample by the size factor. The implementation of DEseq have been described previously (19).

Pathway Analysis

To analyze enrichment for β-cell pathways, lists of pancreatic single-cell features generated/predicted by ML algorithms (Random forest, Gradient boosting, Decision tree classifier and ADA Boost classifier) were compared with β cell-expressed genes (E-GEOD-20966) using Gene Ontology (GO) over-representation analysis on Pantherdb.org (20). Preanalytical workflows included cleaning up entries not mapping to protein-coding gene symbols in E-GEOD-20966. Gene lists for each ML algorithm consisted of up to the top 100 genes as predictors of insulin expression, which were compared against the data set of β-cell expressed genes (N=13,165 from E-GEOD-20966). Overrepresentation analysis using GO categories for biological processes (GO: BP) was performed using binomial testing using false-detection-rate to correct for multiple testing. Lists of significantly enriched pathways associated with each ML algorithm were compared using Venn diagrams (21).

Statistical Analysis

The R software (ver. 3.6.1; R Foundation for Statistical Computing, Vienna, Austria) was used to create the categorical bubble plot using the packages ggplot2 (3.3.3), ggpubr (0.4.0) and proto (1.0.0). Spearman correlation matrix analysis was generated through using R packages corrplot (0.90), Hmisc (4.6.0), dplyr (1.0.7) and readxl (1.3.1) in R and Rstudio software. Statistical software, Microsoft Excel (ver. 2016; Microsoft, Redmond, WA, USA), the R software and/or GraphPad Prism (ver. 8.4.1; GraphPad Software, San Diego, CA, USA) were used for univariate test comparisons and Benjamini-Hochberg method for multiple testing.

RESULTS

Machine Learning (ML) Algorithms Yield Varying Performance Outputs

The scRNA-seq data were obtained from public databanks (GSE84133, GSE85241, E-MTAB-5061, GSE81076, GSE86469) of human pancreatic single-cell transcriptomes. We first randomized this available pancreatic scRNA-seq transcriptomic data and allocated 80% of samples to a discovery/training set (Training; N=11,652 samples) and remaining into a validation/testing set (Test; N=2,913 samples). With the availability of several ML algorithms (**Supplementary Table 1**), we probed the discovery dataset using 10 different ML workflows (**Figure 1A**) to identify features highly associated with the presence of insulin transcripts in a single cell. Genes (features) identified as the most important/predictive variables for each of these ML workflows were used to identify insulin transcript-containing cells from the validation set (remainder 20% of the samples). Validation results of the identified gene features from each of the 10 ML workflows are presented in the form of receiver operator characteristic (ROC) curves (**Figure 1B**). The top three ML algorithms; Gradient boosting, Random Forest and ADA boost (all Ensemble workflows), demonstrated similar performance returning an Area Under Curve (AUC) of between 0.83 – 0.86. A confusion matrix is presented below each ROC curve dataset (**Figure 1B**) to demonstrate the false-positive and false-negative predictions within every workflow. These analyses show that although Ensemble machine learning workflows are the best in predicting insulin-transcribing cells, other workflows, such as logistic regression, also perform closely to the Ensemble methods.

Ensemble ML Workflows to Identify Genes Associated With Insulin Transcription

The scRNA-seq datasets obtained from public databanks of human pancreatic single-cell transcriptomes were classified as insulin-transcribing (1) or those with no insulin (0) (**Figure 2A**). As described earlier, all the three Ensemble ML workflows presented with an AUC that was better than any of the other ML workflows tested in our ROC curve analysis. Ensemble workflows also presented with high accuracy (≥87%), precision (≥0.89), and sensitivity (≥0.95), which was comparable to other popular workflows such as logistic regression (**Figure 2B**). As Ensemble ML workflows such as Random Forest use a collection of decision trees (forest), we decided to compare the performance of the top three (Ensemble) workflows to a single (Decision tree) algorithm. The relative contribution of the top 10 features (genes) from each of these ML workflows are presented as radar plots (**Figure 2C**), whilst the longer list of genes ranked by their importance is presented in **Supplementary Table 2**. *IAPP*, *ADCYAP1*, *LDHA* and *SST* were common to all three Ensemble workflows. We then examined the pathways targeted by these features (genes) identified through each of the Ensemble and Decision Tree classifier by comparing them to a separate islet β-cell dataset (**Figure 2D**). Number of GO terms enriched across all four ML workflows (**Figure 2D**) suggests several common pathways (including insulin secretion) targeted by the features identified through these analysis

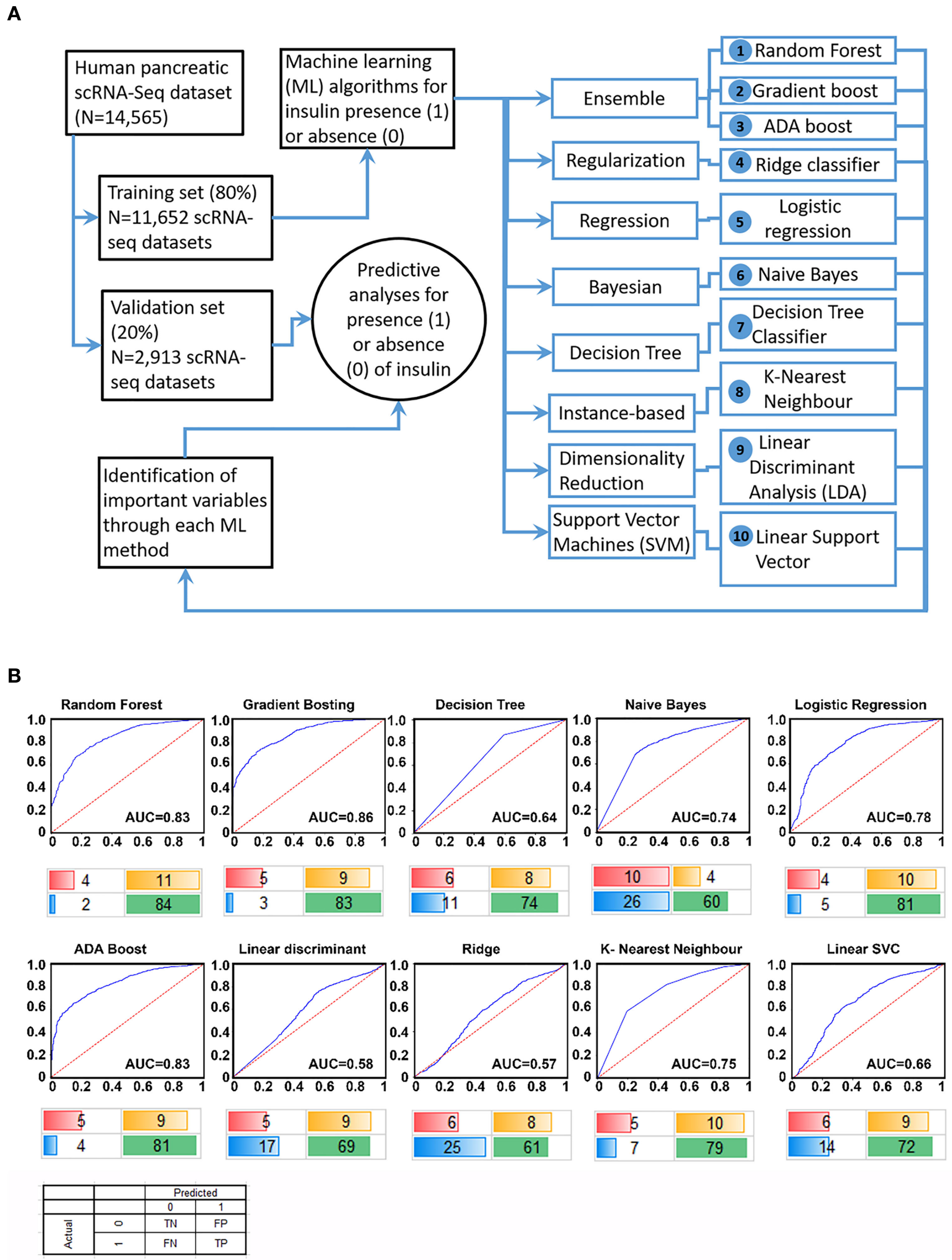
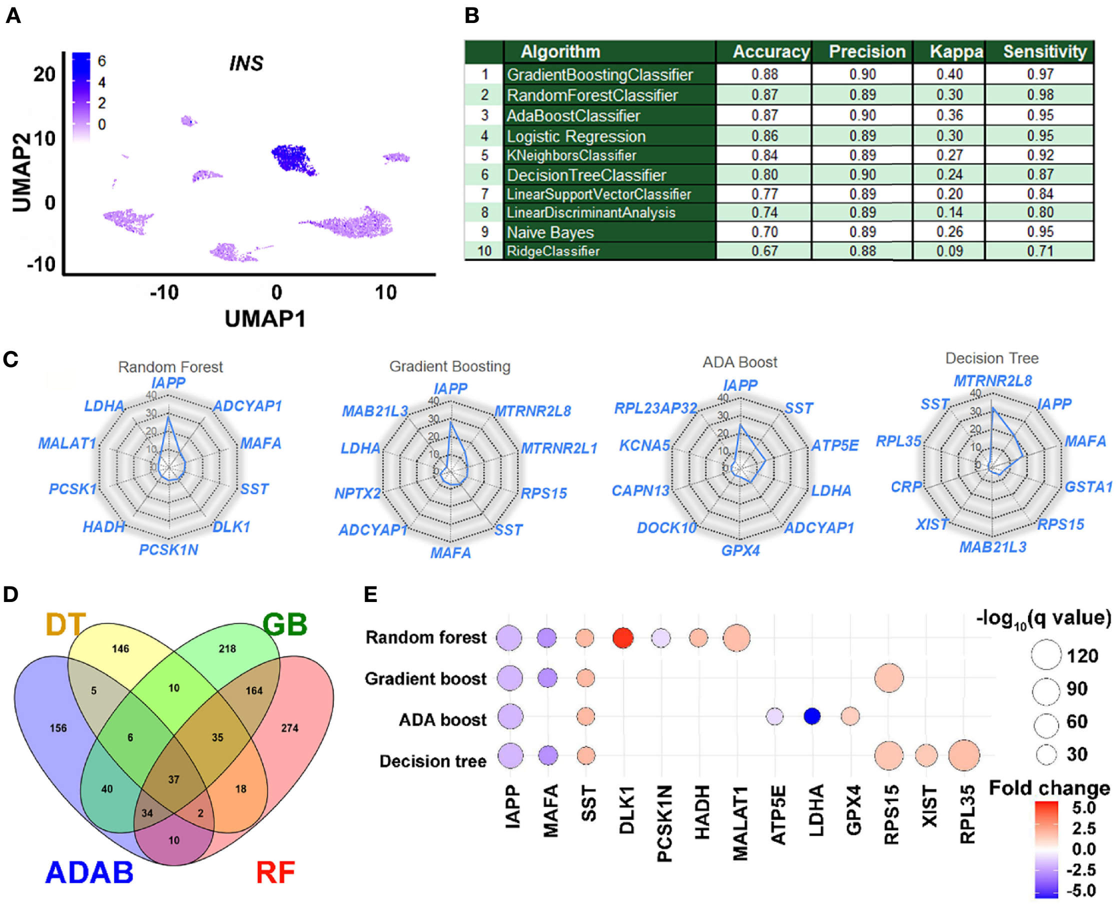


FIGURE 1 | Study design and performance of different ML workflows. A flowchart of our analytical plan is presented in **(A)**. Previously published datasets of single-cell RNA-sequencing analyses from pancreatic islet cell preparations were randomly divided into a training (N = 11,652) and a validation (N = 2,913) set. The learning phase (Training) involved identifying features (genes) and their associated weights/coefficients in each of the 10 machine learning (ML) methods (listed 1-10). Weighted features were used in the prediction of insulin transcription (across 10 ML algorithms) to test the performance of these models in an independent validation set of samples (N = 2,913). ROC curve plots for each ML algorithm using validation set data are presented in **(B)**. The area under the curve (AUC) for the tested workflows are presented along with a confusion matrix below the plot. Percent values are rounded off to the nearest integer (and hence may not sum up to an absolute 100%) and represent true negative (red), true positive (green), false positive (yellow) and false negative (blue) samples identified in the validation set.



	Algorithm	Accuracy	Precision	Kappa	Sensitivity
1	GradientBoostingClassifier	0.88	0.90	0.40	0.97
2	RandomForestClassifier	0.87	0.89	0.30	0.98
3	AdaBoostClassifier	0.87	0.90	0.36	0.95
4	Logistic Regression	0.86	0.89	0.30	0.95
5	KNeighborsClassifier	0.84	0.89	0.27	0.92
6	DecisionTreeClassifier	0.80	0.90	0.24	0.87
7	LinearSupportVectorClassifier	0.77	0.89	0.20	0.84
8	LinearDiscriminantAnalysis	0.74	0.89	0.14	0.80
9	Naive Bayes	0.70	0.89	0.26	0.95
10	RidgeClassifier	0.67	0.88	0.09	0.71

FIGURE 2 | Performance and application of learned features in understanding insulin gene transcription. **(A)** A 2D clustering of pancreatic single cells assessed in this study using UMAP (Uniform Manifold Approximation and Projection plot). Cellular subtypes based on the UMAP clustering algorithm are labeled and graded (scale, inset) as per the level of insulin gene transcripts. **(B)** The performance of learning models on accurately identifying insulin-positive (1) and insulin-negative (0) single cells from the validation dataset are presented. **(C)** Relative weighted rank contributions of the top 10 genes in each of the four listed ML algorithms are presented as spider plots plotted in the order of importance (starting clockwise at 12-O'clock position). Percent representation of each of the genes indicates their relative contribution in the set on the spider plot with a logarithmic scale (center=1% and outer circle=100%). A comparison of the gene features identified by the top three ensemble workflows is presented along with those identified by the Decision Tree classifier. **(D)** Pathways targeted by up to the top 100 features (**Supplementary Table 2**) from each of the four selected ML methods (RF, Random Forest; GB, Gradient Boosting; ADAB, ADA Boost; DT, Decision Tree) identified using gene ontology (GO) function analysis are presented in the Venn diagram. Number of GO terms enriched and common for top features (genes) in each ML method are plotted. **(E)** All significantly dysregulated genes identified from and common to the four ML algorithms **(C)** presented herein were assessed in the scRNA-seq dataset from Ire1$\alpha^{\beta-/-}$ mice. Bubble plot presenting fold-change and statistical significance (q-value) for each of the genes in Ire1$\alpha^{fl/fl}$ and Ire1$\alpha^{\beta-/-}$ mice are shown. Blue color represents downregulation while red color indicates increased abundance of transcripts in Ire1$\alpha^{\beta-/-}$ mice compared to control.

(**Figure 2D** and **Supplementary Table 3**). These genes were also validated in a bulk RNA-seq dataset (GSE152111, n=66) of human islet samples (**Supplementary Figure 1**). In this analysis (**Supplementary Figure 1**), most of these gene transcripts had significant positive correlation with insulin transcript. While some of the gene transcripts such as *LDHA*, *CRP*, *RPS15* and *RPL35* negatively correlated with insulin transcript in human islets.

Insulin-Associated Genes Are Dysregulated During β-Cell Dedifferentiation

Dedifferentiation of β-cells, characterized by the loss of expression of key β-cell maturation marker genes with an accompanying reduction in insulin secretion, has been observed in mouse models of type 1 (T1D) and type 2 (T2D) diabetes, as well as in individuals with diabetes (22–25). We questioned if the expression of gene variables identified and validated (*in silico*) as being associated with insulin gene transcription (**Figure 2C**) are dysregulated in a mouse model of T1D with evidence of islet dedifferentiation. Transient dedifferentiation of islet β-cells was recently reported in an established T1D preclinical mouse model upon β-cell-specific deletion of a key stress response gene, *Ire1α*, (Ire1$\alpha^{\beta-/-}$) (17). These mice also demonstrated reduced β-cell number as well as diminished expression of insulin transcripts in β-cells compared to control (Ire-1$\alpha^{fl/fl}$) mice. Therefore, we evaluated the expression of the total 25 gene transcripts that made up the top 10 features across the four different ML workflows (**Figure 2C**) in the single cell datasets generated from these (Ire1α$^{β-/-}$ and Ire1α$^{fl/fl}$) islets. Twelve of these features were not significantly different between Ire1α$^{β-/-}$ and Ire1α$^{fl/fl}$ islets. However, the remaining thirteen features were significantly dysregulated in β-cells of Ire-1α$^{β-/-}$ mice that were undergoing dedifferentiation (**Figure 2E**). Dedifferentiating β-cells showed significant downregulation of five key genes; *Iapp*, *MafA*, *Pcsk1n*, *Atp5e* and *Ldha*, whilst all other insulin-associated gene transcripts showed significantly higher levels (**Figure 2E**).

In type 2 diabetes (T2D), it is known that *INS* transcript expression is reduced. Therefore we validated the top, common gene features (*IAPP*, *SST*, *MAFA*, *ADCYAP1* and *LDHA*) from the three ML workflows using a separate publicly available single-cell RNA-seq dataset from non-diabetic (ND) vs T2D adult human pancreas (GSE154126 (18)). Four of the five genes (*IAPP*, *SST*, *MAFA*, *ADCYAP1*), were significantly lower in T2D insulin-transcribing cells compared to ND insulin-transcribing cells (**Supplementary Table 4**).

DISCUSSION

In this study, we compared the performance characteristics of 10 different ML algorithms, (**Supplementary Table 1**) that are currently used in big data analyses. We analyzed a scRNA-seq dataset that was randomly split to a larger (80%; 392,684,052 data points) training set involving model learning, and then a smaller (20%; 98,171,013 data points) validation set. All algorithms identified a set of genes (features) that associate with insulin-production (1) defined as the presence of one or more transcripts of insulin in a sample, or no insulin production (0) from the 11,652 single cells analyzed in the training test. We validated the predictive features identified through each ML workflow in the validation/test set of 2,913 single cell transcriptomes. ML workflows that returned high performance (based on AUC, sensitivity/specificity) were selected and the top 10 genes (ranked by their importance) in each of those ML methods were re-validated in discrete mouse and human datasets that model beta cell dedifferentiation (summarized in **Figure 3**).

Our analysis provides two major outcomes that are of interest to a broad range of data analysts and biologists. First, a comparison of the ML algorithms identified Ensemble-based ML methods as the best performing algorithms in our analyses. Logistic regression performed closest to Ensemble methods, in line with previous reports in clinical datasets (26). We then compared Ensemble methodologies to the Decision tree algorithm. Decision tree offers the often-desired simplistic model generation method as compared to Ensemble methods such as Random Forest. The latter builds multiple decision trees independently and offers an overall learning model that is closest to the best possible prediction. Indeed, Decision tree was determined to be a weaker predictor than the Random Forest as the latter reduces variance using different sample sets (bootstrap) in training, randomizing feature subsets, and combining the predictive learning by building multiple decision trees. Random Forest prediction outcomes were similar to gradient boosting, which also builds a set of decision trees, but one tree at a time. The bagging and boosting approach used in ADA/Gradient boosting methods seems to have offered better accuracy and performance in insulin prediction analysis than those observed using Random Forest, whereas the Random Forest algorithm offered the highest sensitivity (**Figure 2B**) amongst all methodologies tested.

The other outcome from this analysis is the identification of genes that are associated with and predictive of insulin gene transcription in single cells. Since bulk RNA-sequencing studies do not offer the desired single-cell resolution to identify transcriptional regulation at a cellular level, our analyses provide a firsthand view of insulin gene transcriptional determinants identified through an unbiased, big data machine learning approach. The top three methodologies (based on high AUC values) belonged to Ensemble machine learning workflow. Weighted relative importance of the top-10 most important features are compared (**Figure 2C**). Interestingly, five genes were common to the top 10 features from all the algorithms compared – *IAPP, ADCYAP1, MAFA, SST* and *LDHA*. The top-ranked gene associated with insulin gene transcription across all the Ensemble workflows was *IAPP*. Islet amyloid polypeptide (*IAPP*) and insulin are known to be expressed in pancreatic islet β-cells and co-secreted in response to changes in glucose concentration (27, 28). Their mRNA levels are also regulated by glucose. The promoters of both these genes share similar cis-acting sequence elements, and both bind the master regulatory transcription factor *PDX1* (27). *FoxA2* (*HNF-3β*) negatively regulates *IAPP* promoter activity (29) and has also been shown to suppress insulin gene expression (30). Although insulin gene expression is known to be regulated by several islet-enriched transcription factors, *MafA* is the most well recognized β-cell-specific activator of insulin gene expression (31). The selection of *MAFA* as a key feature by three of the compared ML approaches tested through this analysis is therefore not surprising. The inclusion of *SST* in the top three gene features is intriguing. Somatostatin expression is known to be important in control of insulin release and ablation of somatostatin-expressing delta cells impairs pancreatic islet function and cause neonatal death in rodents (32). SST analogs were shown to inhibit the release of insulin *via* the activation of both ATP sensitive K+ channels and G protein-coupled inward rectifier K+ channels (33). Another candidate that was identified through these analyses is *MTRNR2L8*, a neuroprotective and antiapoptotic peptide derived from a portion of the mitochondrial *MT-RNR2* gene and reported in fetal as well as adult beta cells (34). *ADCYAP1* stimulates insulin secretion in a glucose-dependent manner (35) and genetic screening in T2D Caucasians indicated the presence of two SNPs in exons 3 and 5 of this gene to be associated with T2D (36). Finally, *LDHA*, which was also selected through these unbiased analyses across the top-three ML workflows is a pancreatic β-cell disallowed gene (37–39) and human *LDHA* levels are predictive of insulin transcription (40). Consistent with these previous reports, our validation analysis in human islets

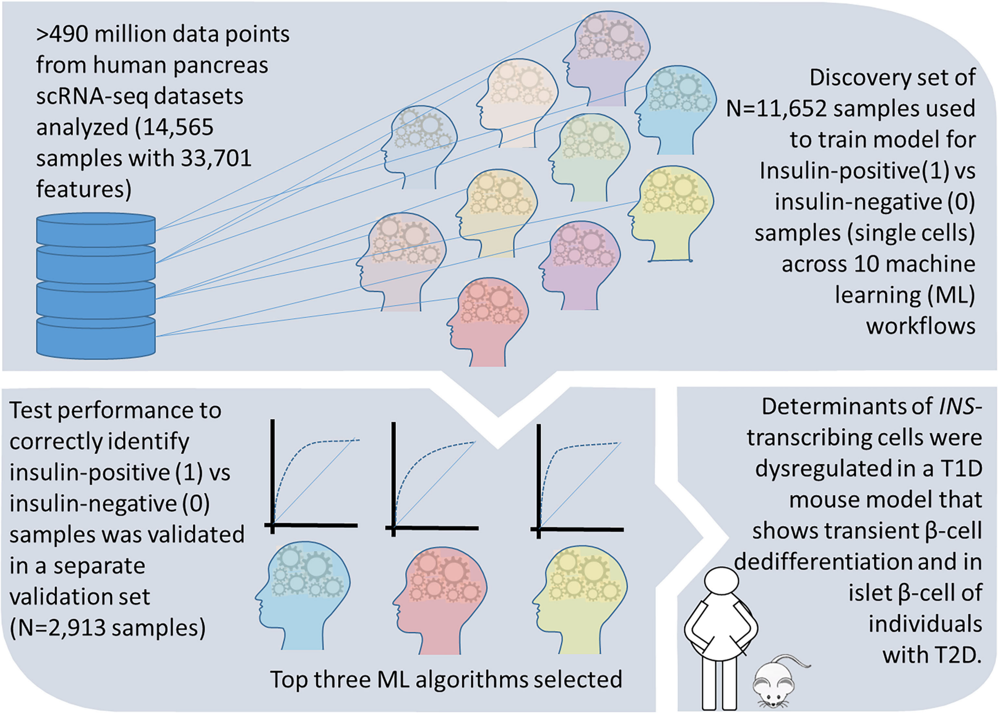

>490 million data points from human pancreas scRNA-seq datasets analyzed (14,565 samples with 33,701 features)
Discovery set of N=11,652 samples used to train model for Insulin-positive(1) vs insulin-negative (0) samples (single cells) across 10 machine learning (ML) workflows
Test performance to correctly identify insulin-positive (1) vs insulin-negative (0) samples was validated in a separate validation set (N=2,913 samples)
Top three ML algorithms selected
Determinants of *INS*-transcribing cells were dysregulated in a T1D mouse model that shows transient β-cell dedifferentiation and in islet β-cell of individuals with T2D.

FIGURE 3 | A summary of study design and results. Workflow and findings of our study are presented in this schematic, which illustrates the steps in discovery and validation of the important gene features associated with insulin transcript in pancreatic single cell transcriptomes. We further confirm these features to be dysregulated during β-cell dedifferentiation in a T1D mouse model and in individuals with T2D.

RNA-seq data, demonstrated negative correlation of *LDHA* and positive correlation of *ADCYAP1*, *MAFA* and *SST* transcripts with insulin (**Supplementary Figure 1**). Together, these algorithms help in identifying a set of genes expressed in or disallowed from insulin-producing pancreatic β-cells.

Mouse models often provide the validation to understand mechanisms that cannot be tested in human studies. The Ire1α$^{β-/-}$ mouse offers a unique model, wherein pancreatic β-cells transiently dedifferentiate during early post-natal life, allowing these knockout mice to escape immune-mediated β-cell destruction and T1D in later life (17). Analysis of islet single cell sequencing data from this model identified genes that were significantly dysregulated in β-cells of Ire1αβ-/- mice when compared to control (Ire1α$^{fl/fl}$) mice. Eight of thirteen features (from the top 10 features in each of the four ML workflows, **Figure 2C**), which showed significant dysregulation between Ire1α$^{β-/-}$ and Ire1α$^{βfl/fl}$ mice are upregulated in β-cells of Ire1α$^{β-/-}$ mice (**Figure 2E**). Analysis of T2D islet single cell data also revealed down-regulation of four common gene features (*IAPP, SST, MAFA* and *ADCYAP1* identified across our three top ML workflows) in T2D compared to ND insulin transcribing cells (**Supplementary Table 4**). Interestingly, Delta Like Non-Canonical Notch Ligand 1 (*DLK1*) was also significantly downregulated in T2D compared to ND insulin transcribing cells (Mann-Whitney test P-value=0.0017). The imprinted region of chromosome 14q32.2, contains microRNA cluster of *DLK1-MEG3* which are highly expressed and more specific in human β-cells compared to α-cells. Previous study had also shown that in T2D human islets, the *MEG3*-microRNA locus expression levels are significantly lower (41). The 14q32 locus of microRNAs (such as co-expression of miR-376a and miR-432) also have been shown to target and suppress the expression of *IAPP* (41), that was one of the top features in our analyses.

Strength and Limitation

This is a first demonstration comparing multiple ML algorithms to identify key genes associated with insulin transcription using a big dataset of over 490 million data points. As anticipated, Ensemble methods perform better than most other workflows and identified a set of genes that corroborate with previous reports of transcriptional regulation of insulin in mouse and human β-cells. These findings indicate that unbiased ML workflows for big data analyses can generate biologically meaningful results, when applied to large training datasets. Our study provides the codes/scripts for other researchers to use in existing as well as emerging datasets for identification of gene candidates associated with other genetic pathways (e.g., related to *GCG* or *GCK*) in future or to genes recognized to be associated with T2D GWAS datasets. We recognize that there are several limitations: we are unsure as to why some other well known candidates (such as *PDX1*, and *NEUROD*) were not selected by our top predictive models. An explanation is that we used a whole pancreatic single cell dataset and that the predictive models generated through filtering out β-cells may be more enriched for known pro-endocrine gene regulators such as *PDX1*. The other explanation is that although *PDX1* is a key regulator, the transcript levels in these datasets using multiple scRNA-seq technologies may not be sufficient considering the sequencing depth offered by some of these scRNA-seq workflows. It would be of interest to explore β-cell factors associated with insulin transcript levels through subset analyses in β-cell types. This is becoming increasingly important to the islet community as differences in insulin transcripts across islet β-cells [i.e., β-cell heterogeneity (42)] may drive optimal β-cell function (43) as well as diabetes progression.

We recognize that exhaustive [e.g., LOOCV (44)] as well as non-exhaustive cross-validation approaches [such as K-fold cross-validation (45)] were not performed here. Such cross-validation approaches, although useful in assessing how results will generalize to an independent dataset, are mostly used in the validation of much smaller datasets. In big data analyses, the use of such cross-validation methodologies would limit the analyses to only those with an access to high-end cluster computing. The 10 different ML scripts used in these analyses are designed to work on a high-end personal computing device (i7 processor with 4 cores and 32GB RAM or better). We believe that the application of such ML algorithms to the expanding scRNA-seq datasets would lead to the confirmation/validation of current as well as identification of determinants of gene transcription, thereby accelerating innovation in discovery of gene targets in biology and medicine.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: NCBI, GEO, GSE152111 GSE144471, Panc8 (GSE84133, GSE85241, E-MTAB-5061, GSE81076, GSE86469). The data codes/scripts are available through https://github.com/Isletbiology/ML.

AUTHOR CONTRIBUTIONS

Conceptualization: AH. Methodology: AH, AP, SB, and LD. Software: VT, MJ, WW, CD, HL, FE, and LD. Validation: WW, MJ, VT, CD, and AB. Data curation: WW, MJ, VT, CD, HL, YC, WH, FE, LD, and AH. Writing—original draft: AH, MJ, WW, and LD. Review and editing: all authors. Visualization: AH. Supervision: AH. Project administration: AH. Funding acquisition: AH. All authors contributed to the article and approved the submitted version.

FUNDING

The research presented herein has been funded through grants from the Australian Research Council Future Fellowship (FT110100254) the Juvenile Diabetes Research Foundation (JDRF) Australia T1D Clinical Research Network (JDRF/4-CDA2016-228-MB) and the Visiting Professorships (2016-18 and 2019-22) from the Danish Diabetes Academy, funded by the Novo Nordisk Foundation, grant number NNF17SA0031406 to AH. WW is supported through the Leona M. and Harry B. Helmsley Charitable Trust (Grant 2018PG-T1D009) in collaboration with the JDRF Australian Type 1 Diabetes Clinical Research Network funding (Grant 3-SRA-2019-694-M-B) to AH. MJ was supported through a JDRF International Advanced Post-doctoral award (3-APF-2016-178-A-N) and currently a transition award from JDRFI (1-FAC-2021-1063-A-N). AH acknowledges the support through the Ainsworth Medical Research Fund, Western Sydney University School of Medicine, Australia. FE was supported by a grant from the JDRF-5-CDA-2014-184-A-N and currently through grants from NIH NIDDK (1R01DK130919-01 and 1R56DK128136-01). HL is supported by NIH National Research Service Award T32 GM007215.

ACKNOWLEDGMENTS

AH acknowledges the initial interactions with Tune Pers, University of Copenhagen, Denmark.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fendo.2022.853863/full#supplementary-material

REFERENCES

1. Zou J, Huss M, Abid A, Mohammadi P, Torkamani A, Telenti A. A Primer on Deep Learning in Genomics. *Nat Genet* (2019) 51:12–8. doi: 10.1038/s41588-018-0295-5
2. Culos A, Tsai AS, Stanley N, Becker M, Ghaemi MS, McIlwain DR, et al. Integration of Mechanistic Immunological Knowledge Into a Machine Learning Pipeline Improves Predictions. *Nat Mach Intell* (2020) 2:619–28. doi: 10.1038/s42256-020-00232-8
3. Zheng A, Lamkin M, Zhao H, Wu C, Su H, Gymrek M. Deep Neural Networks Identify Sequence Context Features Predictive of Transcription Factor Binding. *Nat Mach Intell* (2021) 3:172–80. doi: 10.1038/s42256-020-00282-y
4. Xu C, Jackson SA. Machine Learning and Complex Biological Data. *Genome Biol* (2019) 20:76. doi: 10.1186/s13059-019-1689-0
5. Hu J, Li X, Hu G, Lyu Y, Susztak K, Li M. Iterative Transfer Learning With Neural Network for Clustering and Cell Type Classification in Single-Cell RNA-Seq Analysis. *Nat Mach Intell* (2020) 2:607–18. doi: 10.1038/s42256-020-00233-7
6. Wong WKM, Joglekar MV, Saini V, Jiang G, Dong CX, Chaitarvornkit A, et al. Machine Learning Workflows Identify a microRNA Signature of Insulin Transcription in Human Tissues. *iScience* (2021) 24:102379. doi: 10.1016/j.isci.2021.102379
7. Angermueller C, Lee HJ, Reik W, Stegle O. DeepCpG: Accurate Prediction of Single-Cell DNA Methylation States Using Deep Learning. *Genome Biol* (2017) 18:67. doi: 10.1186/s13059-017-1189-z

8. Lai M, Liu Y, Ronnett GV, Wu A, Cox BJ, Dai FF, et al. Amino Acid and Lipid Metabolism in Post-Gestational Diabetes and Progression to Type 2 Diabetes: A Metabolic Profiling Study. *PloS Med* (2020) 17:e1003112. doi: 10.1371/journal.pmed.1003112
9. Joglekar MV, Wong WKM, Ema FK, Georgiou HM, Shub A, Hardikar AA, et al. Postpartum Circulating microRNA Enhances Prediction of Future Type 2 Diabetes in Women With Previous Gestational Diabetes. *Diabetologia* (2021) 64:1516–26. doi: 10.1007/s00125-021-05429-z
10. Schrider DR, Ayroles J, Matute DR, Kern AD. Supervised Machine Learning Reveals Introgressed Loci in the Genomes of Drosophila Simulans and D. *Sechellia PloS Genet* (2018) 14:e1007341. doi: 10.1371/journal.pgen.1007341
11. Stoffers DA, Zinkin NT, Stanojevic V, Clarke WL, Habener JF. Pancreatic Agenesis Attributable to a Single Nucleotide Deletion in the Human IPF1 Gene Coding Sequence. *Nat Genet* (1997) 15:106–10. doi: 10.1038/ng0197-106
12. Harrison KA, Thaler J, Pfaff SL, Gu H, Kehrl JH. Pancreas Dorsal Lobe Agenesis and Abnormal Islets of Langerhans in Hlxb9-Deficient Mice. *Nat Genet* (1999) 23:71–5. doi: 10.1038/12674
13. Oliver-Krasinski JM, Stoffers DA. On the Origin of the Beta Cell. *Genes Dev* (2008) 22:1998–2021. doi: 10.1101/gad.1670808
14. Doyle MJ, Sussel L. Nkx2.2 Regulates Beta-Cell Function in the Mature Islet. *Diabetes* (2007) 56:1999–2007. doi: 10.2337/db06-1766
15. Osborne CS, Chakalova L, Brown KE, Carter D, Horton A, Debrand E, et al. Active Genes Dynamically Colocalize to Shared Sites of Ongoing Transcription. *Nat Genet* (2004) 36:1065–71. doi: 10.1038/ng1423
16. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, et al. Comprehensive Integration of Single-Cell Data. *Cell* (2019) 177:1888–902.e21. doi: 10.1016/j.cell.2019.05.031
17. Lee H, Lee YS, Harenda Q, Pietrzak S, Oktay HZ, Schreiber S, et al. Beta Cell Dedifferentiation Induced by IRE1alpha Deletion Prevents Type 1 Diabetes. *Cell Metab* (2020) 31:822–36.e5. doi: 10.1016/j.cmet.2020.03.002
18. Avrahami D, Wang YJ, Schug J, Feleke E, Gao L, Liu C, et al. Single-Cell Transcriptomics of Human Islet Ontogeny Defines the Molecular Basis of Beta-Cell Dedifferentiation in T2D. *Mol Metab* (2020) 42:101057. doi: 10.1016/j.molmet.2020.101057
19. Anders S, Huber W. Differential Expression Analysis for Sequence Count Data. *Genome Biol* (2010) 11:R106. doi: 10.1186/gb-2010-11-10-r106
20. Mi H, Muruganujan A, Huang X, Ebert D, Mills C, Guo X, et al. Protocol Update for Large-Scale Genome and Gene Function Analysis With the PANTHER Classification System (V. *14.0) Nat Protoc* (2019) 14:703–21. doi: 10.1038/s41596-019-0128-8
21. Oliveros JC. *Venny: An Interactive Tool for Comparing Lists With Venn's Diagrams* . Available at: https://bioinfogp.cnb.csic.es/tools/venny/index.html.
22. Wang YJ, Schug J, Won KJ, Liu C, Naji A, Avrahami D, et al. Single-Cell Transcriptomics of the Human Endocrine Pancreas. *Diabetes* (2016) 65:3028–38. doi: 10.2337/db16-0405
23. Cinti F, Bouchi R, Kim-Muller JY, Ohmura Y, Sandoval PR, Masini M, et al. Evidence of Beta-Cell Dedifferentiation in Human Type 2 Diabetes. *J Clin Endocrinol Metab* (2016) 101:1044–54. doi: 10.1210/jc.2015-2860
24. Guo S, Dai C, Guo M, Taylor B, Harmon JS, Sander M, et al. Inactivation of Specific Beta Cell Transcription Factors in Type 2 Diabetes. *J Clin Invest* (2013) 123:3305–16. doi: 10.1172/JCI65390
25. Weir GC, Bonner-Weir S. Five Stages of Evolving Beta-Cell Dysfunction During Progression to Diabetes. *Diabetes* (2004) 53 Suppl 3:S16–21. doi: 10.2337/diabetes.53.suppl_3.S16
26. Lynam AL, Dennis JM, Owen KR, Oram RA, Jones AG, Shields BM, et al. Logistic Regression has Similar Performance to Optimised Machine Learning Algorithms in a Clinical Setting: Application to the Discrimination Between Type 1 and Type 2 Diabetes in Young Adults. *Diagn Progn Res* (2020) 4:6. doi: 10.1186/s41512-020-00075-2
27. Macfarlane WM, Campbell SC, Elrick LJ, Oates V, Bermano G, Lindley KJ, et al. Glucose Regulates Islet Amyloid Polypeptide Gene Transcription in a PDX1- and Calcium-Dependent Manner. *J Biol Chem* (2000) 275:15330–5. doi: 10.1074/jbc.M908045199
28. Mulder H, Ahren B, Sundler F. Islet Amyloid Polypeptide and Insulin Gene Expression Are Regulated in Parallel by Glucose *In Vivo* in Rats. *Am J Physiol* (1996) 271:E1008–14. doi: 10.1152/ajpendo.1996.271.6.E1008
29. Shepherd LM, Campbell SC, Macfarlane WM. Transcriptional Regulation of the IAPP Gene in Pancreatic Beta-Cells. *Biochim Biophys Acta* (2004) 1681:28–37. doi: 10.1016/j.bbaexp.2004.09.009
30. Wang H, Gauthier BR, Hagenfeldt-Johansson KA, Iezzi M, Wollheim CB. Foxa2 (HNF3beta) Controls Multiple Genes Implicated in Metabolism-Secretion Coupling of Glucose-Induced Insulin Release. *J Biol Chem* (2002) 277:17564–70. doi: 10.1074/jbc.M111037200
31. Matsuoka TA, Artner I, Henderson E, Means A, Sander M, Stein R. The MafA Transcription Factor Appears to be Responsible for Tissue-Specific Expression of Insulin. *Proc Natl Acad Sci USA* (2004) 101:2930–3. doi: 10.1073/pnas.0306233101
32. Badi I, Mancinelli L, Polizzotto A, Ferri D, Zeni F, Burba I, et al. miR-34a Promotes Vascular Smooth Muscle Cell Calcification by Downregulating SIRT1 (Sirtuin 1) and Axl (AXL Receptor Tyrosine Kinase). *Arterioscler Thromb Vasc Biol* (2018) 38:2079–90. doi: 10.1161/ATVBAHA.118.311298
33. Smith PA, Sellers LA, Humphrey PP. Somatostatin Activates Two Types of Inwardly Rectifying K+ Channels in MIN-6 Cells. *J Physiol* (2001) 532:127–42. doi: 10.1111/j.1469-7793.2001.0127g.x
34. Blodgett DM, Nowosielska A, Afik S, Pechhold S, Cura AJ, Kennedy NJ, et al. Novel Observations From Next-Generation RNA Sequencing of Highly Purified Human Adult and Fetal Islet Cell Subsets. *Diabetes* (2015) 64:3172–81. doi: 10.2337/db15-0039
35. Filipsson K, Kvist-Reimer M, Ahren B. The Neuropeptide Pituitary Adenylate Cyclase-Activating Polypeptide and Islet Function. *Diabetes* (2001) 50:1959–69. doi: 10.2337/diabetes.50.9.1959
36. Gu HF. Genetic Variation Screening and Association Studies of the Adenylate Cyclase Activating Polypeptide 1 (ADCYAP1) Gene in Patients With Type 2 Diabetes. *Hum Mutat* (2002) 19:572–3. doi: 10.1002/humu.9034
37. Rutter GA, Pullen TJ, Hodson DJ, Martinez-Sanchez A. Pancreatic Beta-Cell Identity, Glucose Sensing and the Control of Insulin Secretion. *Biochem J* (2015) 466:203–18. doi: 10.1042/BJ20141384
38. Rutter GA, Pullen TJ. Comment on: Schuit Et Al. Beta-Cell-Specific Gene Repression: A Mechanism to Protect Against Inappropriate or Maladjusted Insulin Secretion? *Diabetes* (2012) 61:969–75. doi: 10.2337/db12-0775
39. Schuit F, Van Lommel L, Granvik M, Goyvaerts L, de Faudeur G, Schraenen A, et al. Beta-Cell-Specific Gene Repression: A Mechanism to Protect Against Inappropriate or Maladjusted Insulin Secretion? *Diabetes* (2012) 61:969–75. doi: 10.2337/db11-1564
40. Cantley J, Walters SN, Jung MH, Weinberg A, Cowley MJ, Whitworth TP, et al. A Preexistent Hypoxic Gene Signature Predicts Impaired Islet Graft Function and Glucose Homeostasis. *Cell Transplant* (2013) 22:2147–59. doi: 10.3727/096368912X658728
41. Kameswaran V, Bramswig NC, McKenna LB, Penn M, Schug J, Hand NJ, et al. Epigenetic Regulation of the DLK1-MEG3 microRNA Cluster in Human Type 2 Diabetic Islets. *Cell Metab* (2014) 19:135–45. doi: 10.1016/j.cmet.2013.11.016
42. Joglekar MV, Dong CX, Wong WKM, Dalgaard LT, Hardikar AA. A Bird's Eye View of the Dynamics of Pancreatic Beta-Cell Heterogeneity. *Acta Physiol (Oxf)* (2021) 233:e13664. doi: 10.1111/apha.13664
43. Benninger RKP, Hodson DJ. New Understanding of Beta-Cell Heterogeneity and *In Situ* Islet Function. *Diabetes* (2018) 67:537–47. doi: 10.2337/dbi17-0040
44. Zou M, Zhang PJ, Wen XY, Chen L, Tian YP, Wang Y. A Novel Mixed Integer Programming for Multi-Biomarker Panel Identification by Distinguishing Malignant From Benign Colorectal Tumors. *Methods* (2015) 83:3–17. doi: 10.1016/j.ymeth.2015.05.011
45. Dankers F, Traverso A, Wee L, van Kuijk SMJ. Prediction Modeling Methodology. In: Kubben P, Dumontier M, Dekker A, editors. *Fundamentals of Clinical Data Science*. Cham (CH) (2019). p. 101–20.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.