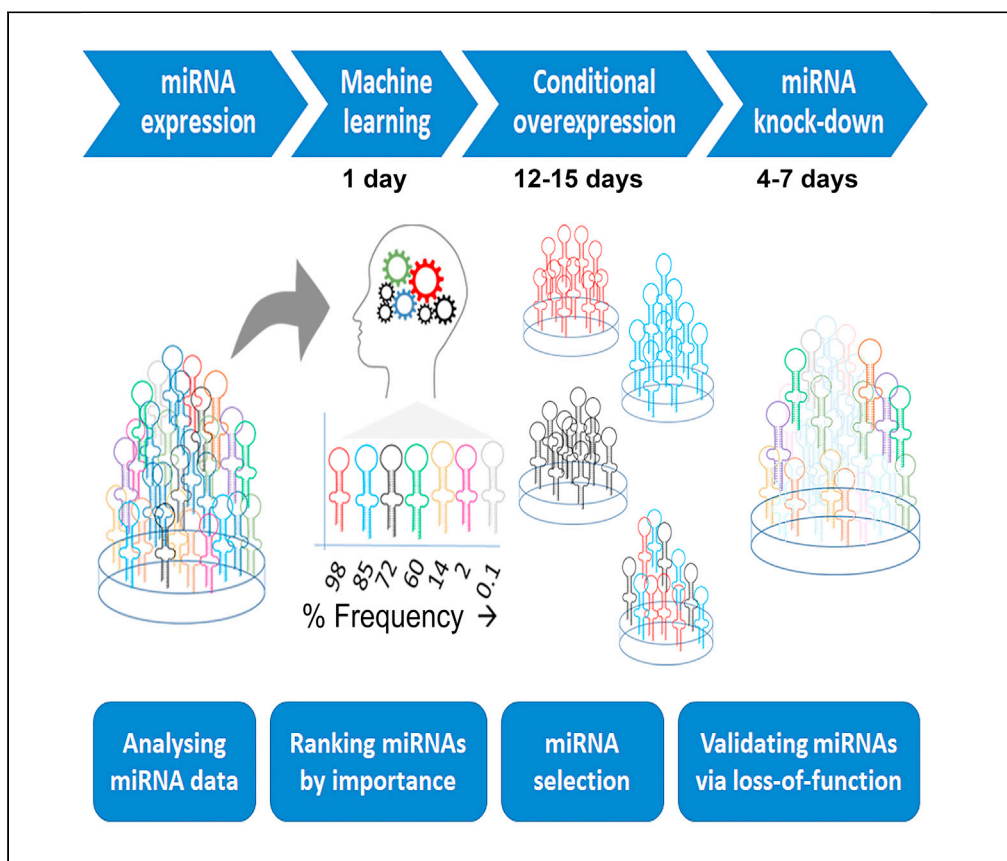


## Protocol

# Manipulating cellular microRNAs and analyzing high-dimensional gene expression data using machine learning workflows



MicroRNAs (miRNAs) are elements of the gene regulatory network and manipulating their abundance is essential toward elucidating their role in patho-physiological conditions. We present a detailed workflow that identifies important miRNAs using machine learning algorithm. We then provide optimized techniques to validate the identified miRNAs through over-expression/loss-of-function studies. Overall, these protocols apply to any field in biology where high-dimensional data are produced.

Vijit Saini, Mugdha V. Joglekar, Wilson K.M. Wong, ..., Ronald C.W. Ma, Louise T. Dalgaard, Anandwardhan A. Hardikar

a.hardikar@westernsydney.edu.au

### Highlights

LASSO and bootstrapping identify important miRNAs associated with gene of interest

Generating puromycin resistant, doxycycline inducible, miRNA overexpressing cells

Transient miRNA knockdown using LNA inhibitors in human primary cells

Optimized reagent concentrations and cell densities for miRNA manipulation

Saini et al., STAR Protocols 2, 100910

December 17, 2021 © 2021

[https://doi.org/10.1016/](https://doi.org/10.1016/j.xpro.2021.100910)

[j.xpro.2021.100910](https://doi.org/10.1016/j.xpro.2021.100910)



## Protocol

## Manipulating cellular microRNAs and analyzing high-dimensional gene expression data using machine learning workflows

Vijit Saini,<sup>1,2,5</sup> Mugdha V. Joglekar,<sup>1,5</sup> Wilson K.M. Wong,<sup>1,5</sup> Guozhi Jiang,<sup>3</sup> Najah T. Nassif,<sup>2</sup> Ann M. Simpson,<sup>2</sup> Ronald C.W. Ma,<sup>3</sup> Louise T. Dalgaard,<sup>4</sup> and Anandwardhan A. Hardikar<sup>1,4,6,7,\*</sup>

<sup>1</sup>Diabetes and Islet Biology Group, School of Medicine, Western Sydney University, Narellan Road & Gilchrist Drive, Campbelltown, NSW 2560, Australia

<sup>2</sup>School of Life Sciences and the Centre for Health Technologies, University of Technology Sydney, 15 Broadway, Ultimo, NSW 2007, Australia

<sup>3</sup>Department of Medicine and Therapeutics, and Hong Kong Institute of Diabetes and Obesity, and Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Prince of Wales Hospital, Hong Kong, Special Administrative Region, China

<sup>4</sup>Department of Science and Environment, Roskilde University, Universitetsvej 1, 4000 Roskilde, Denmark

<sup>5</sup>These authors contributed equally

<sup>6</sup>Technical contact

<sup>7</sup>Lead contact

\*Correspondence: [a.hardikar@westernsydney.edu.au](mailto:a.hardikar@westernsydney.edu.au)  
<https://doi.org/10.1016/j.xpro.2021.100910>

## SUMMARY

**MicroRNAs (miRNAs) are elements of the gene regulatory network and manipulating their abundance is essential toward elucidating their role in patho-physiological conditions. We present a detailed workflow that identifies important miRNAs using a machine learning algorithm. We then provide optimized techniques to validate the identified miRNAs through over-expression/loss-of-function studies. Overall, these protocols apply to any field in biology where high-dimensional data are produced.**

**For complete details on the use and execution of this protocol, please refer to Wong et al. (2021a).**

## BEFORE YOU BEGIN

We describe below steps for using R scripts to identify important miRNAs associated with a specific gene or phenotype of interest. Cell culture-related procedures are written for human islet-derived cells (transient knockdown studies and stable miRNA-overexpression) or for the PANC1 human pancreatic duct cell line (stable miRNA-overexpression). Human islets are obtained following approval from human ethics committee.

## Preparation for miRNA data analysis

⌚ Timing: <1 h

1. Download R (version 3.6.2) and Rstudio software (version 1.3.1093) that are freely available. The following links [https://cran.r-project.org/doc/contrib/Paradis-rdebuts\\_en.pdf](https://cran.r-project.org/doc/contrib/Paradis-rdebuts_en.pdf) and <https://education.rstudio.com/learn/beginner/> are guides for beginners in R.
2. Open Rstudio software.
3. Install R packages using `install.packages('(insert package name of interest here)')` command. Following packages are required: `gdata` (ver. 2.18.0), `hash` (ver. 2.2.6.1), `glmnet` (ver. 4.1),



	A	B	C	D	E	F	G	H
1	Biological Group Name	Sample Name	dv	v4	v5	v6	v7	v8
53	Group A	Sample 52	0	39	25.012	17.835	24.857	19.046
54	Group A	Sample 53	0	35.452	25.143	19.538	26.266	21.794
55	Group A	Sample 54	0	39	25.43	17.82	25.518	19.612
56	Group A	Sample 55	0	39	24.084	17.802	25.047	20.047
57	Group A	Sample 56	0	32.268	24.806	18.317	25.135	39
58	Group A	Sample 57	0	39	25.323	17.587	24.895	19.081
59	Group A	Sample 58	0	39	25.076	18.589	24.883	20.457
60	Group A	Sample 59	0	39	25.392	18.164	26.066	20.269
61	Group A	Sample 60	0	31.289	25.381	18.781	39	39
62	Group A	Sample 61	0	39	23.72	17.344	24.631	19.44
63	Group A	Sample 62	0	39	23.828	17.955	25.548	20.326
64	Group B	Sample 63	1	23.427	15.914	16.094	20.069	18.812
65	Group B	Sample 64	1	24.119	18.77	17.536	20.929	19.975
66	Group B	Sample 65	1	23.538	15.743	15.73	39	18.125
67	Group B	Sample 66	1	22.81	15.315	15.832	19.744	18.372

**Figure 1. Representative image of data arrangement prior to analysis**

An example of how your dataset should be arranged; with the samples in rows and variables in columns. The (Biological) Group Name, Sample Name and dependent variable (dv) are in columns A, B and C respectively. This is followed by the independent variables (miRNAs) (v4...v(n)) starting from column D. Cycle threshold (Ct)-values for each miRNA (v4-v(n)) are presented in this example.

- corrplot (ver. 0.84), penalized (ver. 0.9.51) and readxl (ver. 1.3.1)/or alternatively XLconnect (ver. 1.0.3 which will require ActivePerl software) on Rstudio.
4. Format excel document (i.e., data for analysis) with the samples in rows; sample (biological) group name, sample name, dependent variable (dv) and miRNAs (independent variables) in columns (Figure 1).

**△ CRITICAL: Make sure the excel document dataset does not contain symbols/non-numeric features (such as -).**

**Note:** MiRNA data can be presented and analyzed as normalized cycle threshold (Ct)-values or abundance (fold over detectable). In case of a larger range of data, we recommend to log transform the data using  $\log_2(\text{abundance})$ .

5. Code the miRNA names starting at column D. Label code as the variable (v)4 to v(n) (in this case n=757, for an OpenArray™ discovery panel) in your Excel document. Methodologies related to obtaining such miRNA expression data are presented elsewhere (Wong et al., 2015).

**Note:** Make a separate excel sheet containing details of each code representing the miRNA. Please retain the column headings as shown in Figure 1. For the ease of analysis and to avoid any errors, we recommend to code the miRNA names as (v)4 to v(n). Please remember that R commands are case sensitive. This R script created will detect the independent variables starting with v through the following command:

```
indep_var <- paste("v", 4:(n), sep=" ")
```

Column A is optional and is used to keep the record of which group is considered as 0 and 1 in the analysis.

6. Make sure for logistic regression the dv are binary, thus group A will be 0 while group B will be 1.

**Note:** Sort "dv" column for all the samples from smallest to largest, therefore group A ("0") is before group B ("1").

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
<b>Experimental models: Cell lines</b>		
Human islet cells	Derived from human islets obtained from islet isolation centres	(Gershengorn et al., 2004; Joglekar and Hardikar, 2012; Joglekar et al., 2016)
PANC1 cells	Sigma-Aldrich	87092802
<b>Chemicals, peptides and recombinant proteins</b>		
Potassium chloride (KCl)	VWR	26760.295
Potassium dihydrogen phosphate (KH <sub>2</sub> PO <sub>4</sub> )	VWR	26925.295
Sodium bicarbonate (NaHCO <sub>3</sub> )	Sigma-Aldrich	56014
Sodium chloride (NaCl)	Sigma-Aldrich	59888
Disodium hydrogen phosphate (Na <sub>2</sub> HPO <sub>4</sub> )	VWR	28026.292
D-(+)-Glucose	Sigma-Aldrich	G7021
EDTA	VWR	VWRC20301.290
Trypsin (1:250), powder	Thermo Fisher Scientific	27250018
Medium (CMRL-1066)	Thermo Fisher Scientific	11530037
Medium (DMEM, high glucose)	Thermo Fisher Scientific	11965092
Fetal bovine serum, certified, United States	Thermo Fisher Scientific	16000044
GlutaMAX Supplement	Thermo Fisher Scientific	35050061
Penicillin-Streptomycin (5,000 U/mL)	Thermo Fisher Scientific	15070063
hEGF recombinant, expressed in E. coli, lyophilized powder, suitable for cell culture	Sigma-Aldrich	E9644
Polybrene (Hexadimethrine bromide)	Sigma-Aldrich	H9268
Doxycycline	Sigma-Aldrich	D9891
Puromycin Dihydrochloride	Thermo Fisher Scientific	A1113803
SMARTvector inducible third-generation lentiviral vectors (Non-Target Control)	Dharmacon	VSC6570
SMARTvector inducible third-generation lentiviral vectors (microRNA)	Dharmacon	VSH6906
TRlzol reagent	Thermo Fisher Scientific	15596026
miRCURY LNA microRNA Power Inhibitor	QIAGEN	339131
Cumate-inducible promoters (alternate to doxycycline-inducible SMARTvector™)	System Bioscience	<a href="https://systembio.com/shop/pcdh-cuo-mcs-ires-gfp-ef1%ce%b1-cymr-t2a-puro-all-in-one-inducible-ires-lentivector/#product_info_how_it_works">https://systembio.com/shop/pcdh-cuo-mcs-ires-gfp-ef1%ce%b1-cymr-t2a-puro-all-in-one-inducible-ires-lentivector/#product_info_how_it_works</a>
BLOCK-iT™ Inducible H1 RNAi Entry Vector Kit (Do-It-Yourself shRNA-inducible vector as an alternate to doxycycline-inducible SMARTvector™)	Thermo Fisher Scientific	<a href="https://www.thermofisher.com/au/en/home/references/protocols/rnai-epigenetics-and-gene-regulation/sirna-protocol/inducible-shma-expression-vectors.html">https://www.thermofisher.com/au/en/home/references/protocols/rnai-epigenetics-and-gene-regulation/sirna-protocol/inducible-shma-expression-vectors.html</a>
<b>Software and algorithms</b>		
R software (The R project)	<a href="https://cran.r-project.org">https://cran.r-project.org</a>	ver. 3.6.2
Rstudio software	<a href="https://www.rstudio.com/">https://www.rstudio.com/</a>	Ver 1.3.1093 (used here. Most recent version of Rstudio should also work with R ver 3.6.2).
R script (codes)	<a href="https://github.com/lsletbiology/Penalized-regression">https://github.com/lsletbiology/Penalized-regression</a>	n/a
<b>Other</b>		
24-well tissue culture plates (Falcon)	Falcon	353047
12-well tissue culture plates (Falcon)	Falcon	353043
6-well tissue culture plates (Falcon)	Falcon	353046
6-well suspension plates (Greiner)	Greiner	M9062
T25 tissue culture flasks (Falcon)	Falcon	352097
T75 tissue culture flasks (Falcon)	Falcon	353136

MATERIALS AND EQUIPMENT

**Trypsin:** Once prepared, trypsin should be filtered using 0.2  $\mu\text{m}$  filtration system and then aliquoted as 10–12 mL aliquots into 15 mL conical tubes and stored at  $-20^{\circ}\text{C}$  (for long term storage).

Reagent	Final concentration (mM)	Amount (for 500 mL)
Potassium chloride (KCl)	5.365	0.2 g
Monopotassium phosphate ( $\text{KH}_2\text{PO}_4$ )	0.441	0.03 g
Sodium bicarbonate ( $\text{NaHCO}_3$ )	4.285	0.18 g
Sodium chloride (NaCl)	136.893	4.0 g
Disodium phosphate ( $\text{Na}_2\text{HPO}_4$ )	0.704	0.05 g
D-glucose	5.551	0.5 g
EDTA	1.300	0.19 g
Trypsin (1:250)	n/a	1.25 g
Autoclave water	n/a	500 mL

**Medium for PANC1 or hIPCs culture:** Once prepared, cell culture media are stored at  $4^{\circ}\text{C}$  (up to three months).

Reagent	Final concentration	Amount
Medium <sup>a</sup>	1 $\times$	450 mL
Fetal bovine serum	10%	50 mL
GlutaMAX <sup>TM</sup>	1%	5 mL
Penicillin-Streptomycin (stock 5,000 U/mL)	1%	5 mL
hEGF (stock 0.1 mg/mL) <sup>b</sup>	10 ng/mL	50 $\mu\text{L}$

<sup>a</sup>High glucose DMEM medium is for PANC1 cells and CMRL (1066) medium is for islet cells.

<sup>b</sup>hEGF is only included in CMRL medium (i.e., for islet cells).

## STEP-BY-STEP METHOD DETAILS

### MiRNA data analysis

⌚ **Timing:** The time to run miRNA data analysis using Least Absolute Shrinkage and Selection Operator (LASSO) penalized regression and bootstrapping depends on the installed memory (RAM) of the computer, dataset size and the number of bootstraps. On a 16.0 GB RAM computer, to run a dataset of  $\sim 90$  samples and  $\sim 750$  independent variables with 1000 bootstraps requires approximately 15–30 mins.

This process of data analysis yields a set of miRNAs that are important in separating the two groups of samples (group A (0) or group B (1)). The miRNAs with higher bootstrap frequencies (ranked higher in importance) are then used for functional analysis using manipulation approaches described later.

1. Ensure that your dataset looks like the example shown in [Figure 1](#).
2. Open or paste our R script used for penalized regression and bootstrap analysis in Rstudio. It can be found through <https://github.com/Isletbiology/Penalized-regression>.

⚠ **CRITICAL:** Make sure to keep the `set.seed()` the same and in the “dv” column all the samples in group A (“0”) are before group B (“1”) to produce a consistent output each time you run the script.

3. Set up directory in `setwd()` function as indicated in the R script. Directory is the location of where your data (excel document) is located.

**Note:** for the directory location entered in R, convert the “\” to “/” at every instance.

4. Read data (excel document) using `read_excel()` function (from `readxl` package).
5. Set the number of independent variables (i.e., miRNAs) as mentioned in the R script to analyze in the following function:

```
indep_var <- paste("v", 4:757, sep="")
```

**Note:** In this example, the independent variables are coded v4 to v757. However, these can be set to any start and end numbers as per your experimental variable numbers and the way data is entered in the excel file. In such case ensure that correct number is reflected in the script.

6. Provide the number of bootstraps (iterations) as indicated in the R script in the following function:

```
boot_num <- 1000.
```

**Note:** In our case bootstrap set at 1000 times was sufficient. It can be increased up to 10,000 bootstraps and would require longer computational times for generating results. Bootstrapping through the `glmnet` R package, involves randomization by eliminating ~37% samples and replacing them with the same number of samples selected randomly at each bootstrap from the remaining samples in the set. This approach is a renowned sampling method (Efron and Tibshirani, 1997).

7. Create an output directory (i.e., the location of where the results will be exported into) and file name (for the results) as presented in the R script in the following function:

```
output <- "../results/name_".
```

**Note:** In this case the "results" folder located in the output directory is the location for your results. The file name(s) will start with "name\_" for the results that will be exported.

8. Run the script.
9. Check the directory folder where the results would be exported to (after the run has completed). The script will produce three results document file:
  - a. A PDF document (name\_\_Lasso\_PENALIZED\_full\_data.pdf), containing the penalized regression graph presenting the coefficient and lambda1 of the selected independent variables.
  - b. A CSV document (name\_select\_var\_using\_entire\_data.csv), containing the independent variables and their coefficient selected through penalized regression analysis.
  - c. A CSV document (name\_Lasso\_IF\_Lambda (min)\_boot\_n=1000.csv), containing the bootstrap results.

**△ CRITICAL:** If the results are not produced, check for errors presented in the run. The errors will be in red text on the Rstudio messages window. Often the results are not produced due to the excel document being in the incorrect format or the function in the script is incorrect (see [Troubleshooting 1, 2, 3, and 4](#)).

10. Open and observe the pdf and csv document containing the penalized regression graph and independent variable coefficients.

**Note:** The L-1 penalized regression approach (Goeman, 2010) is generated using the "penalized" R package. The penalized regression graph presents on the x-axis the lambda (a tuning parameter determined/chosen by cross-validation) and on the y-axis the coefficients for the selected variables (miRNAs). In penalized regression analysis a penalty is applied to each independent variable (miRNA) as presented by the lambda (x-axis) and this value is increased until the coefficient (y-axis) obtained is 0. An optimum lambda is selected as a cutoff for identifying the most important/

	A	B	C	D	E
1	variable	sign	count	frequency	opposite_effect
2	(Intercept)	1	1000	100	FALSE
3	v490	-1	962	96.2	FALSE
4	v542	-1	855	85.5	FALSE
5	v491	-1	833	83.3	FALSE
6	v107	-1	831	83.1	FALSE
7	v5	-1	758	75.8	FALSE
8	v551	-1	736	73.6	FALSE
9	v533	-1	521	52.1	FALSE
10	v132	-1	367	36.7	FALSE
11	v382	1	275	27.5	FALSE
12	v135	-1	244	24.4	FALSE
13	v449	-1	205	20.5	FALSE

**Figure 2. Representative image of the output result**

Figure presents the frequency table produced by our penalized logistic regression with bootstrapping. The frequency table can be used to identify the most important discriminatory miRNAs. For example, v490 appears with >96% frequency, suggesting it is one of the most important miRNAs in discriminating dv "0" and dv "1".

relevant variables. The variables (miRNAs) with coefficients not equal to 0 (right y-axis) at this optimum lambda are considered to be important contributors to the prediction model. Thus, the lambda represents a threshold for identifying the most important variables, the coefficients on y-axis present the weights of the variables selected and together, they inform the user of the contribution of each variable (relative weights) to the prediction model. The coefficients of the intercept and variables shown in csv document can be used to determine if a particular sample is in group A(0) or B(1), using the simple regression formula (Montgomery et al., 2021).

- Open and interpret the csv document containing bootstrap results (an example is shown in [Figure 2](#)).

**Note:** The document with the bootstrap results will contain the (independent) variable(s), sign, count, frequency and opposite\_effect in columns A to E respectively. The "variable" column contains the coded variables representing a unique miRNA. The "sign" column identifies which group (A or B) the corresponding miRNA (variable) is higher in Ct- value. For example, miRNA with a sign of "-1" indicates that it has higher Ct-value in group A (i.e. 0). While if the miRNA has a sign of "1", it has higher Ct-value in group B (i.e. 1). The "opposite\_effect" column is used to check if the independent variable (miRNA) appears to be positive ("1" in sign column) in some iterations and negative ("-1" in sign column) in some other bootstrap iterations. When the opposite effect is "TRUE" that means that miRNA is higher (Ct-value) in group "0" for some iterations, while it is lower (Ct-value) in group "0" for other iterations. It will thus appear in the bootstrap table twice, one with a positive sign and another time with a negative sign. If the opposite effect is "FALSE" then that means the miRNA is always higher or lower in all iterations. The "count" column presents the number of times (in this case out of 1000 times(iterations)) the variable is selected to in the model. The "frequency" column reflects the count in percentage. MiRNAs with higher frequencies are of greater importance in our analysis.

- Use output results ([Figure 2](#)) to prioritize the miRNAs based on counts (or % frequency) for over-expression/loss-of-function studies as described below.

**△ CRITICAL: This R script used for penalized regression with bootstrap analysis is not suitable for a dataset with <10 samples.**

**Note:** In instances where unprejudiced workflows for the selection of multiple variables is desired, alternate machine learning approaches, such as Ensemble workflows using Random Forest or similar analytical tools can be used as described elsewhere ([Wong et al., 2021b](#)).

### Derivation and propagation of human islet cells for transient miRNA knockdown

⌚ Timing: prior to transfection

13. Set water bath at 37°C.
14. Clean the biosafety cabinet with 80% v/v ethanol and allow 15 min UV sterilization.
15. Prepare medium (CMRL-1066) supplemented with 1% GlutaMAX™, 10% fetal bovine serum, 1 × of penicillin-streptomycin, and 10 ng/mL epithelial growth factor. Refer to table as described in [materials and equipment](#) section for human islet cells and islet-derived progenitor cells (hIPCs).
16. Add around 3500–5000 islet equivalents (IEQs) per T75 flask in CMRL serum-containing medium (as prepared in step #15). Detailed characterization of these islets to proliferating populations of hIPCs via epithelial-to-mesenchymal transition (EMT) ([Joglekar and Hardikar, 2010](#)) is described ([Gershengorn et al., 2004](#); [Joglekar et al., 2009b, 2016](#); [Joglekar and Hardikar, 2012](#)).
17. We observe that islets attach to the culture flasks in 1–3 days (d), after which the islets flatten out with cells migrating out and then proliferating to form epithelial patches within 7–10 d. We feed these islet-derived cells with fresh medium once they attach and then after every 3–4 d.
18. In this study, we have used islet-derived cells at d8–10 for knockdown experiments. These cells continue to express insulin mRNA, protein and secrete insulin in response to glucose for up to 2 weeks ([Joglekar et al., 2016](#)).

### Knockdown of miRNA expression

⌚ Timing: 1 week

MiRNA manipulation can be performed either by increasing or reducing their abundance, followed by testing their effect on potential targets. In this step, we detail the protocol for using Locked Nucleic Acids (LNA) power inhibitors to reduce miRNA expression in human islet cells. LNA power inhibitors are preferred since they are taken up by the cells without the need for any transfection agents. This significantly reduces adverse effects (including cell death) on primary cells such as islet cells. Transient transfection was performed to understand short term effect of selected miRNA(s).

19. Set water bath at 37°C.
20. Clean the biosafety cabinet with 80% v/v ethanol and allow 15 min UV sterilization.
21. Reconstitute the miRNA LNAs from Qiagen (Qiagen, Hilden, Germany) at a concentration of 50 μM using sterile nuclease-free (NF) water. Aliquot into sterile tubes, each with 5 μL volume to avoid repeated freeze-thaw. Freeze and store the stock at –20°C.
22. Prepare trypsin using the reagents as described in [materials and equipment](#) section.
23. Warm culture media for islet cells (prepared in step #15) and trypsin at 37°C water bath.
24. Ensure human islet cells at d8–10 (in culture) from the time of isolation are available at approximately 90%–95% confluency.
25. Remove medium and add 5 mL trypsin and incubate for 3 min at 37°C in a 5% CO<sub>2</sub> incubator.
26. Ensure the cells are detached by observing them under the microscope. Add 2 mL serum-containing medium to stop further action of trypsin. Collect the cells and transfer them into a 15 mL tube and centrifuge at 500 × g for 3 min. Remove the supernatant.
27. Resuspend the cells in 4 mL of serum-containing medium.
28. Count the cells and add 4 × 10<sup>4</sup> – 5 × 10<sup>4</sup> per well of a tissue-culture treated 24-well plate. Each well should have a final volume of 500 μL.
29. Allow the cells to attach for 16–24 h.



30. Add miRNA LNA power inhibitor to each well to obtain a final concentration of 500 nM/well. See [Troubleshooting 5](#)

**Note:** To inhibit more than one miRNA at the same time, we recommend lowering the concentration of each LNA power inhibitor to keep the combined concentration to 500 nM. For example, in the case of inhibiting five miRNAs together, add 100 nM of each of the miRNA-specific inhibitor to the same well to obtain a total concentration of 500 nM.

31. Leave the cells with miRNA LNA power inhibitor for 3-6d.  
32. Harvest cells for downstream analyses at d3 or d6 post-transfection. Cells can be harvested either as a cell pellet in a Eppendorf (1.7 mL) tube, following steps #25-26 or collected directly in 500  $\mu$ L/well of TRIzol<sup>TM</sup> reagent (ThermoFisher scientific, United States).

**△ CRITICAL:** TRIzol<sup>TM</sup> reagent is harmful if inhaled, swallowed, or if it is in contact with the skin, and can cause severe skin burns and eye damage. Safety measures include wearing PPE and rinsing/washing affected area with water. It is recommended that this is handled in a fume cabinet and safely disposed as it is harmful to aquatic life.

### Preparation for miRNA stable overexpression

⌚ Timing: prior to transduction

MiRNA overexpression can be achieved via transfecting mature miRNAs with the help of transfection agents ([Lahmy et al., 2014](#); [Lopez-Beas et al., 2018](#); [Poudyal et al., 2018](#)). To understand sustained overexpression of a single or combination of miRNAs, we selected an inducible system where miRNA overexpression is regulated by the presence of doxycycline. Other methods based on Cumate-vectors are reported ([Mullick et al., 2006](#); [Poulain et al., 2017](#)). Doxycycline-inducible vectors were preferred since they were readily available. However, Cumate vectors are preferred in scenarios where pre-miRNA overexpression is desired. One has to be mindful of post-processing regulations as some miRNAs are present as pre-miRNAs ([Joglekar et al., 2009a](#)).

33. Set water bath at 37°C.  
34. Clean the biosafety cabinet with 80% (v/v) ethanol and allow 15 min UV sterilization.  
35. Prepare medium (high glucose DMEM) supplemented with 1% GlutaMAX<sup>TM</sup>, 10% fetal bovine serum, 1 × of penicillin and streptomycin cocktail. Refer to table as described in [materials and equipment](#) section for PANC1 cells.  
36. Warm culture media for PANC1 cells (prepared in step #35) at 37°C.  
37. Thaw one vial of third-generation doxycycline-inducible and puromycin-resistant SMARTvector<sup>TM</sup> shRNA (short hairpin RNA) lentiviral vectors (containing mature miRNA sequences and co-expressing either GFP or RFP) on ice with titers of approximately  $1 \times 10^8$  TU/mL.

**△ CRITICAL:** Always work with lentiviral particles in a biosafety cabinet within PC2 facility while wearing recommended PPE. Ensure the protocol is approved by the institutional biosafety committee.

38. Calculate the volume of the vector for a multiplicity of infection (MOI) of 0.2 for  $1.25 \times 10^5$  number of cells. The following formula can be used

$V = (MOI \times N) / T$ , where V=required volume of viral vectors in mL; MOI=multiplicity of infection; N=number of cells and T=titer of the viral vectors per mL.

39. PANC1 cells ([Hardikar et al., 2003](#)), obtained from Sigma Aldrich (or similar/ATCC) are recommended to be used at lower passages (up to 10 passages).

**Note:** Alternatively, we have also applied methods described below on human islet-derived progenitor cells (hIPCs). For further details on hIPCs, please refer to our publications (Gershengorn et al., 2004; Joglekar et al., 2009b, 2016; Joglekar and Hardikar, 2012).

40. Ensure PANC1 cells or hIPCs are available at desired density for use.

## Generating miRNA overexpressing PANC1 cell lines

⌚ Timing: 4 weeks

41. Puromycin dose optimization (week 1, d1-7)

- Grow PANC1 or hIPCs in the respective serum-containing medium in a T75 flask until they are approximately 90%–95% confluent.
- Remove medium and add 5 mL trypsin and incubate for 3 min at 37°C in 5% CO<sub>2</sub> incubator.
- Ensure the cells are detached by observing them under the microscope. Add 2 mL serum-containing medium to stop further action of trypsin. Collect the cells and transfer into a 15 mL tube and centrifuge at 500 × g for 3 min.
- Remove the supernatant and resuspend the cells in 4 mL of serum-containing medium.
- Count the cells and add  $4 \times 10^4$  –  $5 \times 10^4$  per well of a tissue-culture treated 24-well plate. Each well should have final 500 μL volume.
- Allow the cells to attach for 16–24 h. Ensure that the wells are just 50%–60% confluent. See [Troubleshooting 6](#)

⚠ **CRITICAL:** It is important to have just enough cells (50%–60%) at the time of puromycin addition. Higher cell density may lead to inappropriate puromycin dose calculation. A wrong dosage of puromycin in later stages of this experiment would kill untransduced as well as transduced cells thereby reducing the transduction efficiency.

- Add different doses of puromycin per well, ranging from 0–10 μg/mL ([Figure 3](#)). Change the medium in the wells with the same dose of puromycin after d3.

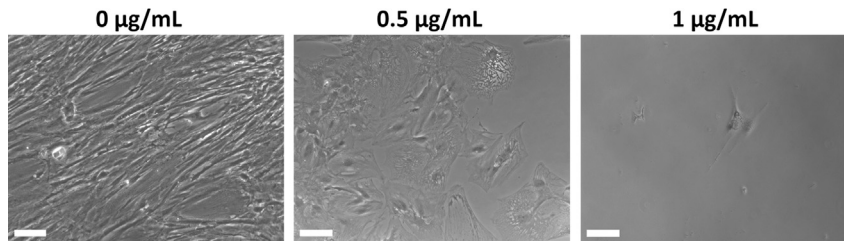
⚠ **CRITICAL:** Puromycin has acute oral toxicity and is harmful if swallowed. Safety measures include washing hands thoroughly and rinsing the mouth with water if consumed.

- Visually inspect all wells at the end of d7 to determine the lowest concentration of puromycin that kills all the non-modified/untransduced cells (100%). In our experience, 8 μg/mL of puromycin was sufficient to kill all PANC1 cells or 2 μg/mL in hIPCs.

42. Transduction of lentiviral vectors carrying miRNAs (week 2, d1-5)

- Culture and trypsinize PANC1 or hIPCs as described above for step 41 (a-d).
- Count the cells and add  $1.25 \times 10^5$  per well of a tissue-culture treated 12-well plate. Each well should have final 1 mL volume.
- Allow the cells to attach for 16–24 h. Ensure that the wells are just 60%–70% confluent. [Troubleshooting 6](#)

⚠ **CRITICAL:** It is important to know the cells you are working with, to understand their doubling time as well as cell size. These factors decide the number of cells to be added per well. Ideally, the cell density at this point should be 60%–70%, which allows these cells to grow and reach confluency in d3 post-transduction. The number of cells is also crucial to calculate the MOI. Each cell type has a different growth rate as well as the number/cm<sup>2</sup> of a particular tissue culture surface area may differ.



**Figure 3. Optimization for puromycin concentration**

Untransduced hIPCs with increasing puromycin concentrations (0–10 µg/mL) in the serum-containing medium. Cells were observed every day and imaged on d7. Images of cells with concentrations of >1 µg/mL are not shown since all cells were killed off. Scale bar 100µm.

- d. Add third-generation doxycycline-inducible and puromycin-resistant SMARTvector™ shRNA lentiviral particles (containing mature miRNA sequences and co-expressing either GFP or RFP) at a MOI of 0.2 along with 8 µg/mL polybrene to fresh culture medium. Replace the media in the wells with this freshly prepared medium pre-mixed with lentivirus and polybrene.

**Note:** We did not observe higher cell death after transduction in a medium containing antibiotics in comparison to performing transduction in an antibiotic-free medium.

- e. Incubate the cells with viral vectors for 24 h and then change the culture medium with fresh serum-containing medium.
- f. After d3 post-transduction, trypsinize the cells and transfer them to 6-well plate. This ensures the desired lower density of cells before puromycin addition, as detailed in step #41.

#### 43. Puromycin selection of transduced cells (week 3, d1-7)

- a. Add optimal concentration of puromycin as determined in step #41, for 7 days to kill off any untransduced cells.

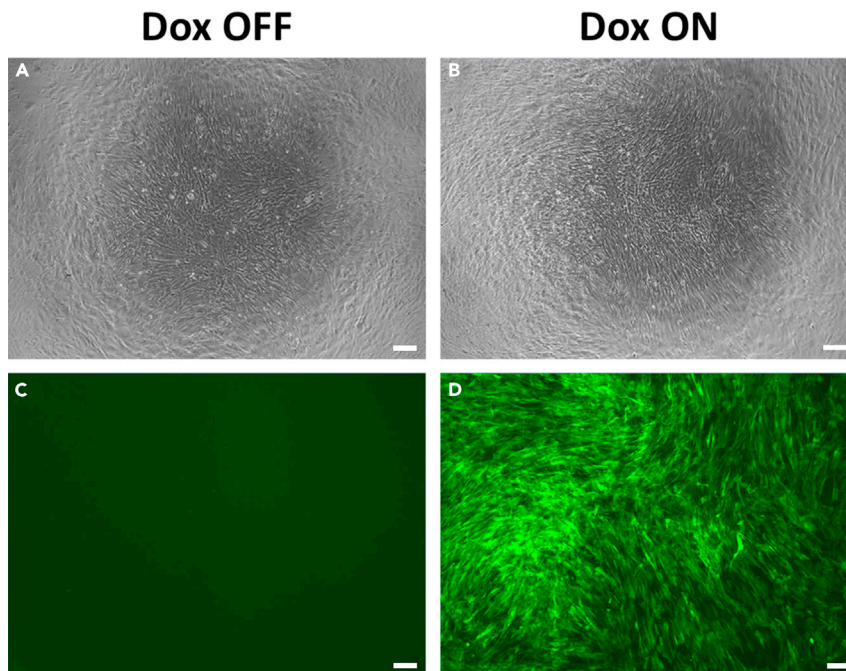
**Note:** After the d7, only transduced cells will survive since they harbor the puromycin-resistance gene. Afterwards, do not use puromycin for culturing/growing transduced cells.

- b. Continue culturing the puromycin-selected miRNA-vector transduced PANC1 cells in serum-containing DMEM medium for at least another week and transfer to larger culture flasks when they become confluent.

#### 44. Optimization of doxycycline concentration (week 4, d1-4)

- a. Grow transduced cells in the respective serum-containing medium until they are approximately 90%–95% confluent.
- b. Trypsinize the cells and add to tissue-culture treated 24-well plate at 60%–70% confluency.
- c. Add doxycycline at concentrations of 0, 50, 200, 500 or 1000 ng/mL for 3 days to determine the optimal presence of fluorescent protein (GFP or RFP, inserted along with the miRNAs in the lentiviral vector backbone (<https://horizondiscovery.com/-/media/Files/Horizon/resources/Technical-manuals/smartvector-inducible-lentiviral-shRNA-manual.pdf>) via fluorescent microscopy (Figure 4).

**△ CRITICAL:** Doxycycline is harmful if inhaled, swallowed, or if it is in contact with skin. Safety measures include wearing personal protective equipment (PPE) and rinsing/washing affected area with water.



**Figure 4. Transduced human islet-derived progenitor cells**

hiPCs transduced with doxycycline-inducible lentiviral vectors expressing green fluorescence protein (TurboGFP) in serum-containing medium with either no doxycycline (A and C) or 200 ng/mL doxycycline (B and D). Scale bar 200 $\mu$ m.

- d. We observe a dose of 1000 ng/mL doxycycline to induce fluorescence (an indirect measure of miRNA overexpression) in all PANC1 cells with the highest intensity and lowest cell death (or 200 ng/mL doxycycline in hiPCs).

**Note:** MiRNA expression can be determined using real-time PCR (Wong et al., 2015) and compared between untransduced and transduced cells (Hardikar et al., 2014). To maintain miRNA expression in these cells, doxycycline should be added at every medium change. Considering the half-life of exogenously added doxycycline, we recommend media change every 3-days.

- e. This puromycin-selected miRNA-vector transduced PANC1 or hiPCs can be continued to grow in serum-containing medium or can be cryofrozen based on experimental need.

## EXPECTED OUTCOMES

A typical datasheet containing miRNA profiles for various samples is shown as an example in Figure 1. Using our script and the above step-by-step protocol, we obtain a frequency table that can be used to identify the most important discriminatory miRNAs with highest frequencies (Figure 2). In our recent work (Joglekar et al., 2021; Wong et al., 2019, 2021a), we have selected either some of the top ranked non-coding RNAs based on their bootstrap frequencies, for validation or for overexpression/knockdown studies. Using our miRNA overexpression strategies with lentiviral vectors, we observe GFP-expressing cells (Figure 4) as a surrogate marker for miRNA overexpression. Final confirmation on miRNA overexpression or knockdown is obtained using real-time quantitative PCR methods.

A strength of the protocol is that it presents a one-stop solution to identify the most important variables using machine-learning algorithms that facilitate an unprejudiced selection of key variables from multi-dimensional data. The transfection and transduction methodologies presented here detail specific steps that are gained through a first-hand experience of commercially available

methods, enabling users to adopt this workflow without the need for extensive optimization. Although commercial protocols are generalized for multiple cell types and need optimization for cell density, concentrations of polybrene, puromycin, doxycycline and the MOIs, our study presents an optimized workflow for human pancreatic duct and islet cell research.

## LIMITATIONS

We have used the miRNA analysis method on several datasets to identify important miRNAs. This script is not suitable for less than 5 samples per group or total samples <10, which could be a limitation where sample number/availability is a constraint. When dealing with low sample numbers for high dimension data, it is recommended to use univariate approaches as described elsewhere (Barracough et al., 2020; Shihana et al., 2021).

Our method is most successful in the case of single miRNA overexpression. We have performed overexpression of up to eight different miRNAs using co-transduction with eight different lentiviral vectors at a higher MOI of 5. It is likely that with the co-transduction approach, every cell may not receive an equal ratio of multiple miRNA sequences. This may not be the best approach where all miRNA overexpression is desired in every transduced cell. A polycistronic vector that contains all desired miRNA sequences in a single open reading frame under a single promoter (Jin et al., 2019) may be used to achieve equimolar ratios of miRNA overexpression in every transduced cell.

Doxycycline inducible promoters are susceptible to be “leaky”, which means that the miRNA expression may occur even in absence of exogenously added doxycycline. This usually happens due to trace amounts of doxycycline present in media reagents such as fetal bovine serum. One option is to purchase doxycycline-free reagents to mitigate the risk of unwanted gene expression. Another option is to use non-doxycycline inducible promoters, such as the cumate-inducible promoters ([https://systembio.com/shop/pcdh-cuo-mcs-ires-gfp-ef1%ce%b1-cymr-t2a-puro-all-in-one-inducible-ires-lentivector/#product\\_info\\_how\\_it\\_works](https://systembio.com/shop/pcdh-cuo-mcs-ires-gfp-ef1%ce%b1-cymr-t2a-puro-all-in-one-inducible-ires-lentivector/#product_info_how_it_works)) or the BLOCK-iT™ Inducible H1 RNAi Entry Vector Kit (<https://www.thermofisher.com/au/en/home/references/protocols/rnai-epigenetics-and-gene-regulation/sirna-protocol/inducible-shrna-expression-vectors.html>, ThermoFisher, MA, USA). Cumate may be less cytotoxic than doxycycline; however, these vectors were not available as pre-made viral stocks.

Human islet-derived progenitor cells (hIPCs) generated from different donor islet tissues may have a different response to vector overexpression or knockdown based on purity, viability, and cell composition. Being primary cells, these cells have limited proliferation capability in vitro compared to any other established cell lines. We also observe slightly different doubling rates and cell size for different biological preparations of hIPCs, which may need adjustment while adding cells to culture well for a desired 50%–60% density. However, these limitations can be observed with many other human tissue-derived primary cells.

## TROUBLESHOOTING

### Problem 1

(Step-by-step methods section; step 9)

Since bootstrapping is a statistical technique that involves random resampling for each iteration (performed here in this example for 1000 iterations), a slight (miniscule) variation in the bootstrapping results may be observed in different analysis runs (as per Figure 5 showing two results) on the same dataset analyzed using the same R script.

### Potential solution

The R function `set.seed()` at the start of the script is used to secure the same result each time when you run the script for analyzing the same dataset. Select the same number (any number) to place in the brackets for the `set.seed()` function at every time the script is used on the same dataset in order to obtain the same results.

	A	B	C	D	E
1	variable	sign	count	frequency	opposite_effect
2	(Intercept)	1	1000	100	FALSE
3	v490	-1	962	96.2	FALSE
4	v542	-1	855	85.5	FALSE
5	v491	-1	833	83.3	FALSE
6	v107	-1	831	83.1	FALSE
7	v5	-1	758	75.8	FALSE
8	v551	-1	736	73.6	FALSE
9	v533	-1	521	52.1	FALSE
10	v132	-1	367	36.7	FALSE
11	v382	1	275	27.5	FALSE
12	v135	-1	244	24.4	FALSE
13	v449	-1	205	20.5	FALSE

	A	B	C	D	E
1	variable	sign	count	frequency	opposite_effect
2	(Intercept)	1	1000	100	FALSE
3	v490	-1	969	96.9	FALSE
4	v542	-1	836	83.6	FALSE
5	v107	-1	835	83.5	FALSE
6	v491	-1	811	81.1	FALSE
7	v551	-1	755	75.5	FALSE
8	v5	-1	741	74.1	FALSE
9	v533	-1	534	53.4	FALSE
10	v132	-1	343	34.3	FALSE
11	v382	1	283	28.3	FALSE
12	v135	-1	237	23.7	FALSE
13	v449	-1	196	19.6	FALSE

Figure 5. Images of two output results produced on the same dataset without the same number in set.seed()

### Problem 2

(Step-by-step methods section; step 9)

An error will occur when the incorrect directory or file name is set (as per Figure 6).

### Potential solution

To set the correct directory and file name, copy the directory (of where the data file is located) and paste it into the blank space in:

```
setwd(" ____ ")
```

of script. Then copy the file name (of the data file) and paste file name into blank space in:

```
ori_dat <- read_excel(" ____ ")
```

of the script. Convert the "\ " to "/" in the script.

### Problem 3

(Step-by-step methods section; step 9)

Errors will occur when the incorrect directory location for the (output) export data is set (as per Figure 7).

### Potential solution

To set the correct directory for the output location, copy the folder location (of where the output document files will be exported to) and paste this directory blank space in:

```
output <- "../ ____ "
```

of the script. Convert the "\ " to "/" in the script.

### Problem 4

(Step-by-step methods section; step 9)

An error will occur when the non-numeric value(s) is detected in your dataset used for analysis (as per Figure 8).

### Potential solution

Open the data file and check the dataset to make sure all values are numeric.

```
> #working directory HAS to have subfolders /pro, /data and /result
> setwd("C:/Users/Dell11/Documents/R/data/")
>
> #Note XLconnect requires ActivePerl. Alternate package to XLConnect is "readxl" package.
> #Readxl does not require ActivePerl and sep=, stringsAsFactors, check.names=. Readxl uses read_excel command
>
> ori_dat <- read_excel("C:/Users/Documents/R/data/sample.xlsx")
Error: 'path' does not exist: 'C:/Users/Documents/R/data/sample.xlsx'
```

**Figure 6. Error message if the directory/file name is incorrect**

### Problem 5

([Step-by-step methods](#) section; step 30)

Cell death observed after adding 500 nM LNA power inhibitors.

### Potential solution

As per the manual of LNA power inhibitors, cell death is not expected (between the recommended range 100 nM–5  $\mu$ M). However, if cell death (more than that in normal cultures/untransfected controls) is observed, one can lower the concentration to 100 nM. In our hands, 500 nM was optimal.

### Problem 6

([Step-by-step methods](#) section; step 42)

Wells are more confluent than desired density after seeding the above-mentioned number of cells.

### Potential solution

The cell number mentioned for these steps is based on our experience in using PANC1 cells/hIPCs to achieve 50%–60% confluency in different plates (24-well or 12-well). One must determine the optimal cell number to reach desired density based on the cells of interest. Different cell types may need a different number to reach confluency based on their shape and size. If the cell density is found more than desired, we recommend to trypsinize and reset the experiment.

## RESOURCE AVAILABILITY

### Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Anandwardhan A. Hardikar, PhD ([A.Hardikar@westernsydney.edu.au](mailto:A.Hardikar@westernsydney.edu.au)).

### Materials availability

MiRNA overexpressing PANC1 cells can be obtained with a reasonable request to the lead contact.

```
> #opt-variable-table-export
> #opt-variable-table-export
> #opt-variable-table-export
> opt_coef_result <- data.frame(opt_coef)
> opt_coef_result <- cbind(row.names(opt_coef_result), opt_coef_result)
> names(opt_coef_result) <- c("selected_var", "coefficients")
> write.table(opt_coef_result, paste(output, "select_var_using_entire_data.csv", sep=""), sep=";", row.names=F)
Error in file(file, ifelse(append, "a", "w")) :
  cannot open the connection
In addition: warning message:
In file(file, ifelse(append, "a", "w")) :
  cannot open file './result/name_select_var_using_entire_data.csv': No such file or directory
>
> #plot variable selection using PENALIZED package (coefficients vs. shrinkage parameter lambda).
> #plot variable selection using PENALIZED package (coefficients vs. shrinkage parameter lambda).
> pdf(paste(output, "Lasso_PENALIZED_full_data.pdf"))
Error in pdf(paste(output, "Lasso_PENALIZED_full_data.pdf")) :
  cannot open file './result/name_Lasso_PENALIZED_full_data.pdf'
> full_fit <- penalized(cc_dat[, dep_var], cc_dat[, indep_var], lambda1=opt_lambda, steps=5, trace=F)
> plotpath(full_fit, lwd=3, font=2, font.lab=2, log="x")
dev.off()
null device
1

> #
> #
> write.table(stat_nonzero, paste(output, paste("Lasso_IF_", "Lambda (" ,lambda_label, ")_boot_nm", boot_num, ".csv", sep=""), sep=""),
+ quote=F, sep=";", row.names=F)
Error in file(file, ifelse(append, "a", "w")) :
  cannot open the connection
In addition: warning message:
In file(file, ifelse(append, "a", "w")) :
  cannot open file './result/name_Lasso_IF_Lambda (min)_boot_nm=2000.csv': No such file or directory
> #
```

**Figure 7. Error message if the output directory location is incorrect**

```
>
> #LASSO using the full CC data
> (cvfit <- optL1(cc_dat[, dep_var], cc_dat[, indep_var], fold=n_fold, trace=F))
Error in weights * y : non-numeric argument to binary operator
> (opt_lambda <- cvfit$lambda)
Error: object 'cvfit' not found
> (opt_coef <- coefficients(cvfit$fullfit))
Error in coefficients(cvfit$fullfit) : object 'cvfit' not found
> (select_var <- names(opt_coef)[-1])
Error: object 'opt_coef' not found
>
```

**Figure 8. Error message if dataset contains non-numeric values**

### Data and code availability

Our script used for penalized regression and bootstrap analyses carried herein is available through <https://github.com/Isletbiology/Penalized-regression>

### ACKNOWLEDGMENTS

This study was supported by the Australian Research Council Future Fellowship (FT110100254), the Juvenile Diabetes Research Foundation (JDRF) Australia T1D Clinical Research Network (JDRF/4-CDA2016-228-MB), a special initiative of the Australian Research Council and the University of Sydney CDIP grant to A.A.H. Collaborative exchange between A.A.H. and L.T.D. are supported through visiting professorships from the Danish Diabetes Academy, funded by the Novo Nordisk Foundation, grant number NNF17SA0031406 (2016-18 and 2019-22). W.K.M.W. acknowledges previous support from the Australian Postgraduate Award, University of Sydney, JDRF Australia PhD top-up award and current support through the Leona M. and Harry B. Helmsley Charitable Trust (Grant 2018PG-T1D009 to A.A.H.) in collaboration with the JDRF Australian Type 1 Diabetes Clinical Research Network (Grant 3-SRA-2019-694-M-B). M.V.J. is supported through a transition award from JDRF, USA (1-FAC-2021-1063-A-N). R.C.W.M. acknowledges support from the RGC Theme-based Research Scheme (T12-402/13N) and Research Impact Fund (R4012-18), the Focused Innovation Scheme, and Faculty Postdoctoral Scheme of the Chinese University of Hong Kong. We acknowledge the infrastructure support from the School of Medicine, Western Sydney University; Ingham Institute; University of Technology Sydney; and the Rebecca L. Cooper Medical Research Foundation.

### AUTHOR CONTRIBUTIONS

V.S., M.V.J., and W.K.W.M. optimized protocols, performed experiments, and wrote the first draft of the manuscript; G.J. wrote the PLR and bootstrapping script; N.T.N., A.M.S., R.C.W.M., L.T.D., and A.A.H. supervised the work, troubleshooting, and provided resources; A.A.H. conceptualized and finalized the draft; all authors reviewed, commented, contributed to revision, and agreed on the final version for submission.

### DECLARATION OF INTERESTS

The authors declare no competing interests.

### REFERENCES

- Barraclough, J.Y., Joglekar, M.V., Januszewski, A.S., Martinez, G., Celermajer, D.S., Keech, A.C., Hardikar, A.A., and Patel, S. (2020). A MicroRNA signature in acute coronary syndrome patients and modulation by colchicine. *J. Cardiovasc. Pharmacol. Ther.* 25, 444–455.
- Efron, B., and Tibshirani, R. (1997). Improvements on cross-validation: the .632+ bootstrap method. *J. Am. Stat. Assoc.* 92, 548–560.
- Gershengorn, M.C., Hardikar, A.A., Wei, C., Geras-Raaka, E., Marcus-Samuels, B., and Raaka, B.M. (2004). Epithelial-to-mesenchymal transition generates proliferative human islet precursor cells. *Science* 306, 2261–2264.
- Goeman, J.J. (2010). L1 penalized estimation in the Cox proportional hazards model. *Biom. J.* 52, 70–84.
- Hardikar, A.A., Farr, R.J., and Joglekar, M.V. (2014). Circulating microRNAs: understanding the limits for quantitative measurement by real-time PCR. *J. Am. Heart Assoc.* 3, e000792.
- Hardikar, A.A., Marcus-Samuels, B., Geras-Raaka, E., Raaka, B.M., and Gershengorn, M.C. (2003). Human pancreatic precursor cells secrete FGF2 to stimulate clustering into hormone-expressing islet-like cell aggregates. *Proc. Natl. Acad. Sci. U S A* 100, 7117–7122.
- Jin, W., Mulas, F., Gaertner, B., Sui, Y., Wang, J., Matta, I., Zeng, C., Vinckier, N., Wang, A., Nguyen-Ngoc, K.V., et al. (2019). A network of microRNAs acts to promote cell cycle exit and differentiation of human pancreatic endocrine cells. *iScience* 21, 681–694.
- Joglekar, M.V., and Hardikar, A.A. (2010). Epithelial-to-mesenchymal transition in pancreatic islet beta cells. *Cell Cycle* 9, 4077–4079.



- Joglekar, M.V., and Hardikar, A.A. (2012). Isolation, expansion, and characterization of human islet-derived progenitor cells. *Methods Mol. Biol.* *879*, 351–366.
- Joglekar, M.V., Joglekar, V.M., and Hardikar, A.A. (2009a). Expression of islet-specific microRNAs during human pancreatic development. *Gene Expr. Patterns* *9*, 109–113.
- Joglekar, M.V., Patil, D., Joglekar, V.M., Rao, G.V., Reddy, D.N., Mitnala, S., Shouche, Y., and Hardikar, A.A. (2009b). The miR-30 family microRNAs confer epithelial phenotype to human pancreatic cells. *Islets* *1*, 137–147.
- Joglekar, M.V., Trivedi, P.M., Kay, T.W., Hawthorne, W.J., O'connell, P.J., Jenkins, A.J., Hardikar, A.A., and Thomas, H.E. (2016). Human islet cells are killed by BID-independent mechanisms in response to FAS ligand. *Apoptosis* *21*, 379–389.
- Joglekar, M.V., Wong, W.K.M., Ema, F.K., Georgiou, H.M., Shub, A., Hardikar, A.A., and Lappas, M. (2021). Postpartum circulating microRNA enhances prediction of future type 2 diabetes in women with previous gestational diabetes. *Diabetologia* *64*, 1516–1526.
- Lahmy, R., Soleimani, M., Sanati, M.H., Behmanesh, M., Kouhkan, F., and Mobarra, N. (2014). MiRNA-375 promotes beta pancreatic differentiation in human induced pluripotent stem (hiPS) cells. *Mol. Biol. Rep.* *41*, 2055–2066.
- Lopez-Beas, J., Capilla-Gonzalez, V., Aguilera, Y., Mellado, N., Lachaud, C.C., Martin, F., Smani, T., Soria, B., and Hmadcha, A. (2018). miR-7 modulates hESC differentiation into insulin-producing beta-like cells and contributes to cell maturation. *Mol. Ther. Nucleic Acids* *12*, 463–477.
- Montgomery, D.C., Peck, E.A., and Vining, G.G. (2021). *Introduction to Linear Regression Analysis* (John Wiley and Sons, Inc).
- Mullick, A., Xu, Y., Warren, R., Koutroumanis, M., Guilbault, C., Broussau, S., Malenfant, F., Bourget, L., Lamoureux, L., Lo, R., et al. (2006). The cumate gene-switch: a system for regulated expression in mammalian cells. *BMC Biotechnol.* *6*, 43.
- Poudyal, D., Herman, A., Adelsberger, J.W., Yang, J., Hu, X., Chen, Q., Bosche, M., Sherman, B.T., and Imamichi, T. (2018). A novel microRNA, hsa-miR-6852 differentially regulated by Interleukin-27 induces necrosis in cervical cancer cells by downregulating the FoxM1 expression. *Sci. Rep.* *8*, 900.
- Poulain, A., Perret, S., Malenfant, F., Mullick, A., Massie, B., and Durocher, Y. (2017). Rapid protein production from stable CHO cell pools using plasmid vector and the cumate gene-switch. *J. Biotechnol.* *255*, 16s–27.
- Shihana, F., Wong, W.K.M., Joglekar, M.V., Mohamed, F., Gawarammana, I.B., Isbister, G.K., Hardikar, A.A., Seth, D., and Buckley, N.A. (2021). Urinary microRNAs as non-invasive biomarkers for toxic acute kidney injury in humans. *Sci. Rep.* *11*, 9165.
- Wong, W., Farr, R., Joglekar, M., Januszewski, A., and Hardikar, A. (2015). Probe-based real-time PCR approaches for quantitative measurement of microRNAs. *J. Vis. Exp.* *98*, 52586. <https://doi.org/10.3791/52586>.
- Wong, W.K., Jiang, G., Sorensen, A.E., Chew, Y.V., Lee-Maynard, C., Liuwantara, D., Williams, L., O'connell, P.J., Dalgaard, L.T., Ma, R.C., et al. (2019). The long noncoding RNA MALAT1 predicts human pancreatic islet isolation quality. *JCI Insight* *5*, e129299.
- Wong, W.K.M., Joglekar, M.V., Saini, V., Jiang, G., Dong, C.X., Chaitarvornkit, A., Maciag, G.J., Gerace, D., Farr, R.J., Satoor, S.N., et al. (2021a). Machine learning workflows identify a microRNA signature of insulin transcription in human tissues. *iScience* *24*, 102379.
- Wong, W. K. M., Thorat, V., Joglekar, M. V., Dong, C. X., Lee, H., Bhave, A., Engin, F., Pant, A., Dalgaard, L. T., Bapat, S. & Hardikar, A. A. 2021b. Machine learning algorithms in big data analyses identify determinants of insulin gene transcription. <https://doi.org/10.1101/2021.03.27.437353>.