Bioinformatics Program

Technical University of Munich

Ludwig-Maximilians-Universität München

Master's Thesis in Bioinformatics

Bioinformatic Analysis of the role of MAFA in Stem Cell-Derived Human Pancreatic Islets

Alexander Fastner

Bioinformatics Program

Technical University of Munich

Ludwig-Maximilians-Universität München

Master's Thesis in Bioinformatics

Bioinformatic Analysis of the role of MAFA in Stem Cell-Derived Human Pancreatic Islets

Informatische Analyse von der Funktion von MAFA in aus Stammzellen gewonnenen menschlichen Pankreasinseln

Author:	Alexander Fastner
Supervisors:	PhD Sara Jimènez
	PhD Veronica Cochrane
Advisors:	Prof. Dr. Matthias Hebrok
	Prof. Dr. Fabian Theis
Submitted:	17.06.2024

Declaration of Authorship

I confirm that this Master's thesis is my own work and I have documented all sources and material used.

Date 15.06.2024 Alexander Fastner

Abstract

Type 1 diabetes is a prevalent and costly autoimmune disorder characterized by the destruction of insulin-producing pancreatic β cells, necessitating exogenous insulin administration. While insulin therapy remains the standard treatment, stem cell-based approaches offer a promising solution to restore endogenous insulin production by generating functional β cells to replace those lost. Current stem cell-derived islets exhibit functional immaturity and divergent insulin response compared to primary β cells.
The transcription factor MAFA plays a crucial role in β cell function and insulin production, with expression profiles varying between rodents and humans, necessitating testing on human β cells. In juvenile human β cells, there is residual function but relatively low MAFA activity, while adult cells exhibit increased insulin production and higher MAFA expression. A naturally occurring MAFA mutation, $MAFA^{S64F}$, initially shows high insulin production coupled with elevated MAFA expression, but this heightened activity is short-lived, suggesting MAFA's critical role in β cell maturation and function.
With data obtained from prior Bulk RNA sequencing of both overexpressed $MAFA^{WT}$ and $MAFA^{S64F}$ cell lines and their controls, I seek a better understanding of the complex mechanisms of insulin production. I am looking for what differences can be found between the two cell lines and what might contribute to the instability of the $MAFA^{S64F}$ mutant. The overarching ultimate goal being to find a way to stabilize future stem cell-derived β cells. In this document I describe my process, observations and key findings of this analysis.
Using a variety of bioinformatics methods I first identify genes which are differentially expressed between these samples. Furthermore I discern transcription factors of interest that differ between the samples and change over time. From this I ascertain which pathways are activated or inhibited at different time points in their differentiation.
Leveraging the Bulk RNA dataset including only two time points that were at hand at the time of this work, I developed a solid analysis blueprint to support future research as more comprehensive data becomes available.
I verify that the doxycycline treatment utilized for inducing MAFA overexpression did not elicit significant side effects, thereby allowing any observed effects to be attributed solely to the overexpression itself, rather than confounding factors.
Analysis revealed that the differential gene expression profile of the $MAFA^{MUT}$ cell line exhibited a markedly higher level of differential expression in comparison to other $MAFA^{WT}$ cell lines. Unexpectedly, pathway activity inference uncovered that the Control cell lines demonstrated substantially larger activation scores for pathways than the overexpressed lines, despite exhibiting low differential gene expression.
Furthermore, the final cross-check with the single-cell dataset unveiled an apparent lack of expression of previously identified transcription factors within the β cell population of primary human islets.
While the limited dataset posed challenges in fully elucidating the mechanisms underlying MAFA's role in β cell maturation and insulin production, this analysis lays the groundwork for future investigations by establishing a robust analytical framework and highlighting key areas warranting further exploration.

Kurzzusammenfassung

Typ-1-Diabetes ist eine weit verbreitete und kostspielige Autoimmunerkrankung, die durch die Zerstörung von insulinproduzierenden β-Zellen der Bauchspeicheldrüse gekennzeichnet ist und eine exogene Insulinverabreichung erforderlich macht. Während die Insulintherapie nach wie vor die Standardbehandlung darstellt, bieten stammzellbasierte Ansätze eine vielversprechende Lösung zur Wiederherstellung der endogenen Insulinproduktion durch die Erzeugung funktionsfähiger β-Zellen als Ersatz für die verlorenen. Aktuelle aus Stammzellen gewonnene Inseln weisen im Vergleich zu primären β-Zellen eine funktionelle Unreife und eine unterschiedliche Insulinreaktion auf.
Der Transkriptionsfaktor MAFA spielt eine entscheidende Rolle bei der Funktion von β-Zellen und der Insulinproduktion, wobei die Expressionsprofile zwischen Nagetieren und Menschen variieren und Tests an menschlichen β-Zellen erforderlich machen. In juvenilen menschlichen β-Zellen besteht eine Restfunktion, aber eine relativ geringe MAFA-Aktivität, während adulte Zellen eine erhöhte Insulinproduktion und eine höhere MAFA-Expression aufweisen. Eine natürlich vorkommende MAFA-Mutation, $MAFA^{S64F}$, zeigt zunächst eine hohe Insulinproduktion in Verbindung mit einer erhöhten MAFA-Expression. Diese erhöhte Aktivität ist jedoch nur von kurzer Dauer, was auf die entscheidende Rolle von MAFA bei der Reifung und Funktion von β-Zellen schließen lässt.
Mit Daten aus früheren Massen-RNA-Sequenzierungen sowohl überexprimierter $MAFA^{WT}$- als auch $MAFA^{S64F}$-Zelllinien und ihrer Kontrollen versuche ich, die komplexen Mechanismen der Insulinproduktion besser zu verstehen. Ich suche nach den Unterschieden zwischen den beiden Zelllinien und was zur Instabilität der $MAFA^{S64F}$-Mutante beitragen könnte. Das übergeordnete Ziel besteht darin, einen Weg zu finden, zukünftige aus Stammzellen gewonnene β-Zellen zu stabilisieren. In diesem Dokument beschreibe ich meinen Prozess, meine Beobachtungen und die wichtigsten Ergebnisse dieser Analyse.
Mithilfe verschiedener bioinformatischer Methoden identifiziere ich zunächst Gene, die in diesen Proben unterschiedlich exprimiert werden. Darüber hinaus identifiziere ich interessante Transkriptionsfaktoren, die sich zwischen den Proben unterscheiden und sich im Laufe der Zeit ändern. Daraus ermittle ich, welche Signalwege zu unterschiedlichen Zeitpunkten ihrer Differenzierung aktiviert oder gehemmt werden. Unter Nutzung des Bulk-RNA-Datensatzes, der nur zwei Zeitpunkte umfasste, die zum Zeitpunkt dieser Arbeit vorlagen, habe ich einen soliden Analyseplan entwickelt, um zukünftige Forschungen zu unterstützen, sobald umfassendere Daten verfügbar werden.
Ich konnte zeigen, dass die zur Induktion der MAFA-Überexpression eingesetzte Doxycyclin-Behandlung keine signifikanten Nebenwirkungen hervorrief, sodass alle beobachteten Effekte ausschließlich auf die Überexpression selbst und nicht auf Störfaktoren zurückzuführen sind. Die Analyse ergab, dass das differentielle Genexpressionsprofil der $MAFA^{MUT}$-Zelllinie im Vergleich zu anderen $MAFA^{WT}$-Zelllinien ein deutlich höheres Maß an differentieller Expression aufwies. Unerwarteterweise ergab die Inferenz der Signalwegaktivität, dass die Kontrollzelllinien trotz geringer differentieller Genexpression wesentlich höhere Aktivierungswerte für Signalwege aufwiesen als die überexprimierten Linien. Darüber hinaus ergab die abschließende Gegenprüfung mit dem Einzelzelldatensatz einen offensichtlichen Mangel an Expression zuvor identifizierter Transkriptionsfaktoren innerhalb der β-Zellpopulation primärer menschlicher Inseln.

Während der begrenzte Datensatz eine Herausforderung bei der vollständigen Aufklärung der Mechanismen darstellte, die der Rolle von MAFA bei der Reifung von β-Zellen und der Insulinproduktion zugrunde liegen, legt diese Analyse den Grundstein für zukünftige Untersuchungen, indem sie einen robusten analytischen Rahmen schafft und Schlüsselbereiche hervorhebt, die eine weitere Erforschung erfordern.

Acknowledgments

I would like to thank Prof. Dr. Matthias Hebrok for offering me the opportunity to contribute to such an interesting project.

My sincere appreciation extends to my supervisor Sara Jimènez for her invaluable assistance in pivoting this project as circumstances changed as well as for her assistance in manual annotation of a comparison dataset.

Furthermore, I wish to express my heartfelt gratitude to Veronica Cochrane for her guidance and for keeping me focused on and grounded in the biological relevance of this project.

I am also grateful to my parents whose unwavering support has made my educational journey up to this point possible.

Table of Contents

Abstract **3**
Kurzzusammenfassung **5**
Acknowledgments **7**
Introduction **11**
General Background 11
Specific Background 11
Experiment Design 12
Methods and Materials **15**
Input Data 16
Bulk RNA 16
OmniPath 16
Human Single Cell RNAseq Data for β Cells 16
Overview of Technical Implementation 17
Differential Expression Analysis 18
Quality Control and Filtering 18
Analysis 18
Transcription Factor Inference 20
Functional Annotation Clustering 22
Data preparation 22
Clustering 22
Pathway Inference 23
Verify Against Millman Human Cell Data 24
Results **25**
The variance in differential expression of the Controls is in line with expectations 25
The Doxycycline treatment does not induce detectable differential expression 27
$MAFA^{MUT}$ cell line displayed significantly higher Differential Gene Expression 29
The Multivariate Linear Model is more useful for Transcription Factor inference with higher number of target genes 31
The Control Problem for Pathway Inference Analysis 32
JAK-STAT Pathway has higher gene count in $MAFA^{MUT}$ 35
Transcription Factors Identified prior do not match up with expression in Primary Human Islets 36
Discussion **39**
Conclusion **41**
Code Availability **42**
List of Figures and Tables **42**
Bibliography **44**

Introduction

General Background

Type 1 diabetes is a complex, chronic autoimmune disorder that affects millions of people worldwide, and its prevalence continues to rise. It is one of the most common and costly diseases, and is expected to increase in prevalence in the future.[1] In this condition, the body's immune system mistakenly attacks and destroys the pancreatic β cells, which are responsible for producing insulin, a vital hormone that regulates blood sugar levels. As Insulin is a protein hormone, it cannot be taken orally as it would be denatured by stomach acid and thus must be administered by injection.[2] While insulin therapy remains the cornerstone of type 1 diabetes management since its discovery by Frederick Banting and Charles Best, researchers are continuously exploring new and innovative approaches to improve the quality of life for those living with this chronic condition. [3]

Stem cells, with their unique ability to differentiate into various cell types, offer a potential solution to the loss of pancreatic β cells in type 1 diabetes. By harnessing the regenerative power of stem cells, researchers are exploring ways to generate functional, insulin-producing β cells that could be used to replace the damaged or destroyed cells in individuals with type 1 diabetes.[4]

Current state-of-the-art stem cell-derived islets are functionally immature and respond differently to signals for insulin production than primary β cells. Left unprotected they are also prone to being destroyed by the immune system. One of the primary hurdles is the risk of immune rejection, as the transplanted cells may be recognized as foreign by the recipient's immune system. While the use of immunosuppressant drugs can mitigate this risk, the adverse effects of long-term immunosuppression often outweigh the potential benefits.

Researchers are therefore exploring alternative strategies, such as shielding or disguising the transplanted cells, to evade the host's immune response without the need for lifelong medication. Additionally, ensuring the long-term viability and function of the transplanted cells remains an ongoing challenge, as current data suggests that after 5 years, 50-70% of transplant recipients still require insulin supplementation.[5]

The goal of this thesis is to provide the means to better understand the complex biological systems in stem cell-derived β cells eventually leading to more stable insulin production for diabetes patients, and improving the stability of stem cells in general.

Ultimately, the goal of stem cell-derived islet transplantation is to achieve insulin independence, where the transplanted cells can fully restore the body's natural insulin production and regulation.

Specific Background

The transcription factor MAFA plays a crucial role in the function of pancreatic β cells, which are responsible for insulin production. While MAFA is crucial for both rodent and human β functionality the expression profile diverges making testing on human β cells essential. In human cells, the expression and activity of $MAFA^{WT}$

varies with the age and maturation of β cells. In juvenile β cells (under 9 years old), there is some residual β cell function but relatively low $MAFA^{WT}$ activity, while adult β cells exhibit both increased insulin production and higher levels of $MAFA^{WT}$ expression.
Researchers have identified a naturally occurring mutation in the MAFA gene, known as $MAFA^{S64F}$, which initially shows high levels of insulin production coupled with elevated $MAFA^{WT}$ expression, but this heightened activity is short-lived, as the insulin production eventually crashes despite the continued high expression of the mutant $MAFA^{S64F}$. These findings suggest that MAFA is a critical regulator of β cell maturation and function, and further research into the mechanisms underlying MAFA's influence may lead to the development of targeted therapies or strategies to enhance β cell health and insulin secretion, ultimately benefiting individuals with diabetes.

Experiment Design

The general strategy for generating β cells from human pluripotent stem cells (hPSCs) is to closely recapitulate the path that hPSCs take during embryogenesis. They are directed from the initial starting hPSCs through various intermediary stages to finally become pancreatic islet β cells.

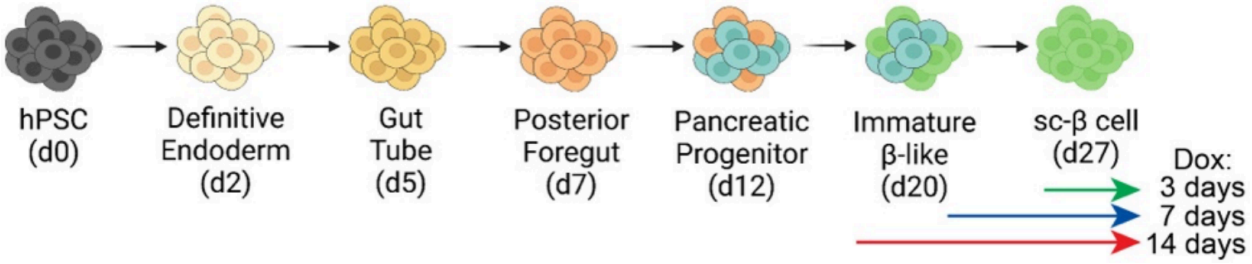

Figure 1| The differentiation process and the timeline for Dox treatment. To ensure the same maturity of the cells all samples were taken on the same end date, having undergone the treatment for differing durations.

Despite recent advances stem cell derived β cells generated *in vitro* do not bear the same characteristics as endogenous adult cells. The inability to express MAFA, whose expression coincides with human β cell maturation likely plays a role in this. To test the effect of MAFA regulation on stem cell-derived β cells, an overexpression model was developed.[5] Using a TALEN-based strategy to insert a Doxycycline (Dox) inducible cassette into the AAVS1 locus of the MEL1-INS-GFP hPSC line. After stimulation with Dox, MAFA expression is detectable. The stem cell lines were then differentiated towards β cells and treated with Dox starting at -3, -7, and -14 days shown in (Fig. 1).
The samples were then assayed for glucose-stimulated insulin secretion (GSIS) functionality after 7 and 14 days of Dox treatment. This process measures the insulin response of cells when given a specified amount of glucose. The data from the 3 day Dox treatment was not included in the data used for this project limiting the time series analysis performed in my work.

The originating question for my work comes from the following observation. In (Fig. 2) MAFA^{WT} produces a significantly larger amount of insulin after 3 days of Dox treatment. This however does not last, as the functionality drops precipitously in a few days. If a β cell could be engineered to produce higher levels of insulin consistently this would significantly benefit any future transplant recipients.

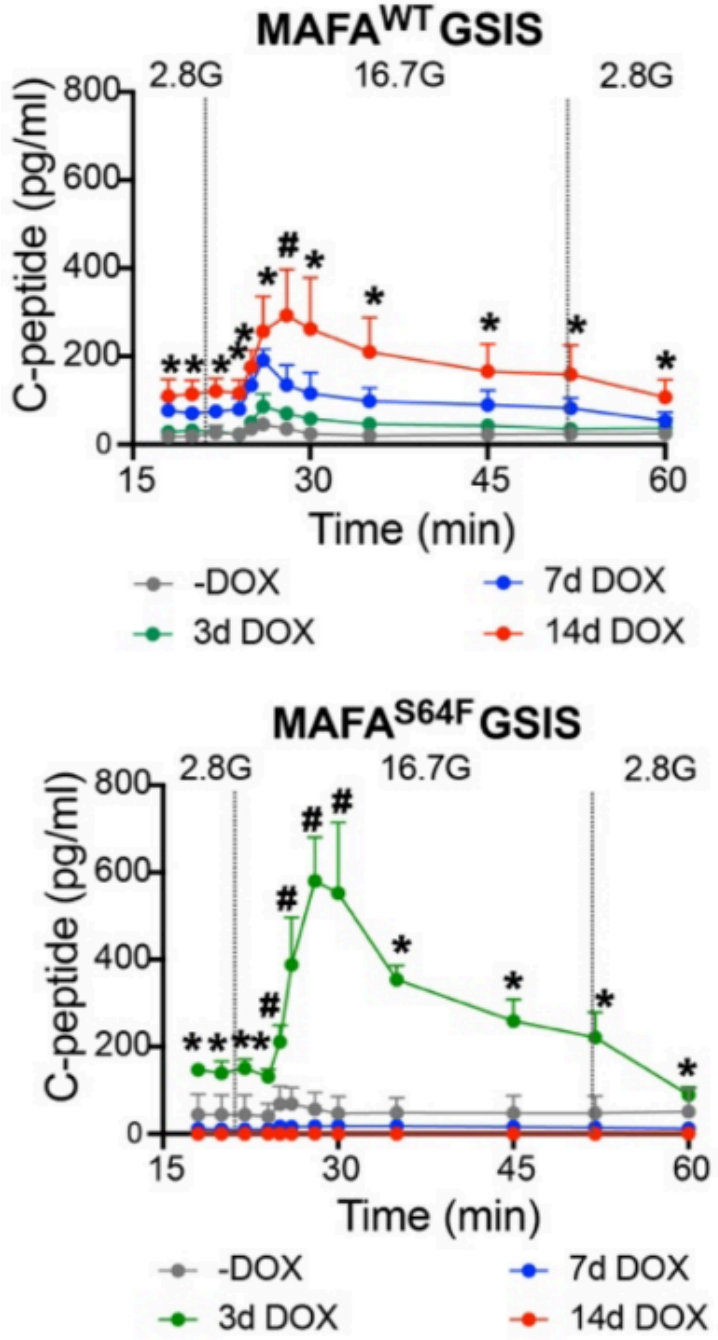

Figure 2| glucose-stimulated insulin secretion (GSIS) results after Doxycycline treatment leading to MAFA overexpression in both the MAFA^{WT} and MAFA^{S64F} for the different data points.

To gain insight into transcriptional level changes induced by $MAFA^{WT}$ and $MAFA^{S64F}$ expression I mainly work with the Bulk RNA dataset described in the (Fig. 3) below. The data used in this project is gathered from the $MAFA^{WT}$ and $MAFA^{S64F}$ lines as well as their respective controls for three time points. Timepoint 0 denotes an initial reading before Dox treatment and subsequent MAFA production begins. Time point 1 is taken after 7 days of Dox treatment and Time point 2 after 14 days. I go on to identify differences between the two cell lines and the temporal differences that emerge with continued Dox treatment.

Finally, I perform a verification step to check the transcription factors I identify against primary human islet data published by the Millman Lab[6]. This is to verify that these transcription factors appear in β cells, as this cannot be known from using a Bulk RNA dataset.

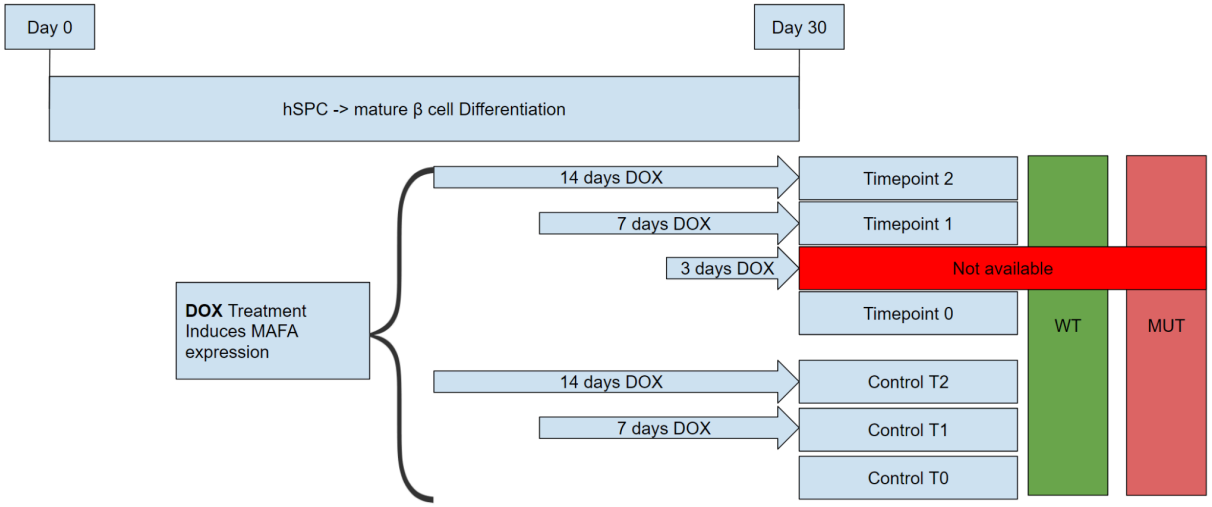

Figure 3| Schema depicting the experiment and how $MAFA^{WT}$ and $MAFA^{MUT}$ (S64F) data at day 0, day -7, and day -14 was generated.

Methods and Materials

The overarching goal of this work was to learn about the changes in biological systems in stem cell-derived β cells caused by the overexpression of MAFA.
I primarily used a Bulk RNA dataset composed of two cell lines, augmented by prior knowledge from OmniPath and cross checked against human β cell data.

Bioinformatics Methods

I used a variety of bioinformatics algorithms and existing software packages.

Resource	Source	Identifier
Deposited data		
Bulk RNA	Generated for this thesis by Veronica Cochrane 2023[5]	Unpublished
scRNA	Augsornworawat P, et al, 2023[6]	Gene Expression Omnibus (GEO) accession code GSE199636
Experimental models		
Unpublished	Na	Na
Software and Algorithms		
anndata v 0.10.2	Virshup I, et al 2021[7]	https://github.com/scverse/anndata
decoupler v 1.6.0	Badia-i-Mompel P, et al 2022[8]	https://decoupler-py.readthedocs.io/en/latest/
matplotlib	Team TMD, 2023[9]	https://pypi.org/project/matplotlib/
omnipath v 1.0.7	Türei D, et al 2021[10]	https://github.com/saezlab/omnipath
pydeseq2 v 0.4.1	Muzellec B, et al 2023[11]	https://github.com/owkin/PyDESeq2
scanpy v 1.9.5	Wolf FA, et al 2018[12]	https://pypi.org/project/scanpy/
scikit-learn v 1.3.1	Olivier Grisel, et al 2024[13]	https://github.com/scikit-learn/scikit-learn
scipy v 1.11.3	Virtanen P, et al 2020[14]	https://github.com/scipy/scipy

Table 1| List of various resources used in this paper, their sources, and links to where to find them.

The utilized conda environment .yaml file containing all used packages. See in the Code Availability section for a link to my github.

Input Data

Bulk RNA

The main datasets input into this analysis were the raw count matrix as well as a Metadata file annotating the various samples and treatments for both the $MAFA^{WT}$ and the $MAFA^{S64F}$ lines respectively. Both are read in with pandas[15] and then combined in an AnnData[7] object for further processing. The Annotated Data (AnnData) python package is useful for handling multilayered annotated data matrices in an efficient and organized manner.

OmniPath

To augment the Bulk RNA data I employed prior knowledge in the form of two networks from the collection OmniPath[10]. This gives me access to PROGENy[16], a network of pathways and their involved genes, as well as CollecTRI[17], a network of transcription factors and their targets.

Human Single Cell RNAseq Data for β Cells

I used an external single cell human β cell dataset as a ground truth to compare my results to. The authors (Millman et al)[6] of this Paper identified deficiencies in lineage specification in human stem cell-derived islets, using Single-nucleus multiomics (RNA + ATAC) data for the stem cell-derived β cells and human primary islets.

This verification step to support the relevance of my findings became necessary because the other analyses in this work were based on a Bulk RNA dataset which lacked the specificity to show that what is found is indeed from a β cell. Later I refer to the primary human islets as the Millman comparison dataset when I check whether the transcription factors I identify are expressed in β cells.

Overview of Technical Implementation

Following are the main analytical steps summarizing this project. A detailed explanation ensues in the subsequent chapters.
I filtered and denoised the Bulk RNA dataset by removing low count genes that do not occur at least once in each sample. After I transformed it into an Annotated Data object, I pre-processed the data by running a Differential Expression Analysis to find changes in gene expression over time. With results from this step I moved on to identify transcription factor activity scores using a multivariate linear model. I used functional annotation to identify enriched clusters based on the most activated and inactivated transcription factors identified earlier.
I then inferred pathway activity for major cellular pathways based on the initial results of the Differential Expression Analysis. Finally, I cross checked whether the transcription factors I identified, and the pathways they are in, are actually expressed in human β cell populations.
A flow diagram showing the input data to and order of various analysis stepsis shown in (Fig. 4).

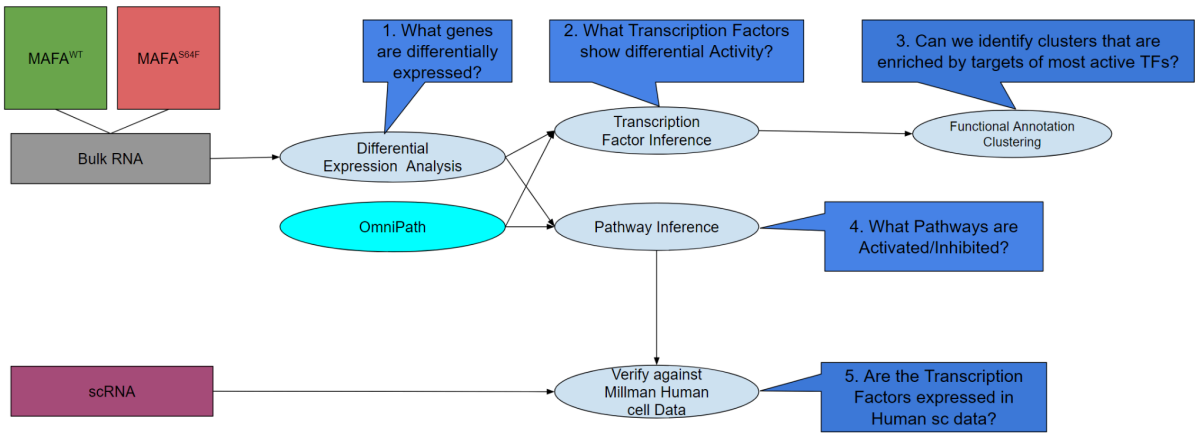

Figure 4| Schema explaining the order of Analyses, the input data sets, and the Questions I am trying to answer

Differential Expression Analysis

Quality Control and Filtering

After combining the raw gene counts and Metadata, I filtered and denoised using the **decoupler** python package[8]. To clean up this data I removed genes that have too few counts in any given sample, have too few counts across all samples, and those that aren't counted in all samples. The parameters to do so were selected as follows.

The **min_count** is the minimum number of counts required per gene for at least some samples. This threshold of counts was set with the assumption that any measurement with less than 10 counts appearing in any sample is likely background and thus not of a high enough concentration to be greatly influential in the cell overall. With the **min_total_count** variable I denoted the limit of how many counts for that gene were required across all samples to be included. The **min_prop** parameter is the proportion of samples the gene must be present in. Using a proportion of 1 as I later do equates to a gene needing at least 1 count in all samples to be included. This was made under the assumption that no gene with a statistically significant effect will be completely absent in any sample.

The thresholds set by **min_total_count** and **min_proportion** set the 2 dashed gray cutoffs in (Fig. 5) respectively. All subsequent analyses were performed on the data represented in the upper right quadrant.

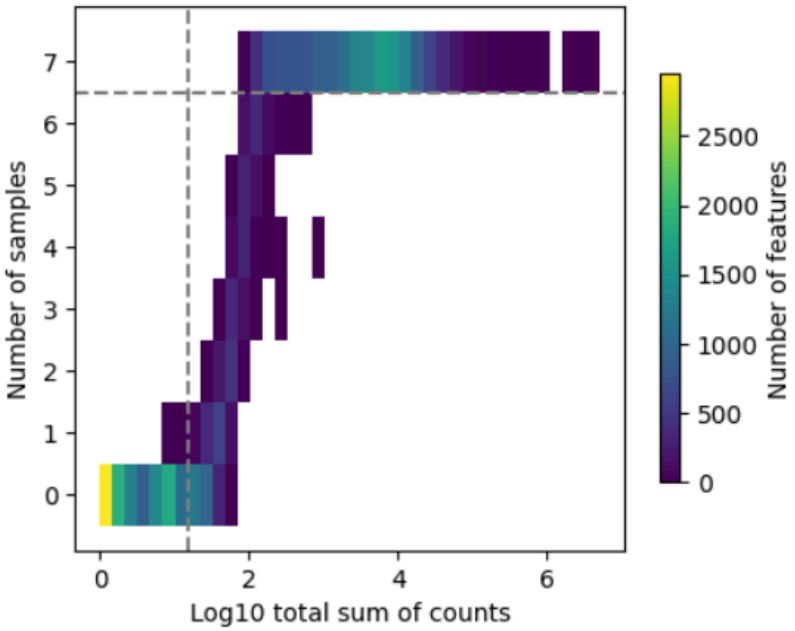

Figure 5| This heatmap shows the data retained after the filtering and denoising process. The colors indicate the absolute number of genes that fall in that window of the log10 total sum of counts.

Analysis

The primary objective was to curate a dataset that facilitates a multidimensional analysis across different time points and cell lines. To enable temporal comparisons, I needed to identify the genes exhibiting differential expression patterns. Additionally,

I aimed to elucidate the expression disparities between the $MAFA^{WT}$ and the $MAFA^{MUT}$ cell lines.

To determine the differential expression of genes across samples for the Bulk RNA dataset I utilized the PyDESeq2 package.[11] This is a python implementation of the original DESeq2[18] method which was written in R. Subsequently I refer to the python package used as DESeq2.

I built a DESeq2 object from my AnnData dataset using un-normalized raw counts as the model internally corrects for library size. The DESeq2 method first estimates size factors to normalize the raw count data and account for differences in sequencing depth between samples. This was done by calculating the geometric mean of each gene across samples and finding a scaling factor for each sample that minimizes the log-fold changes between samples. It then estimates gene-wise dispersion with an empirical Bayes approach to shrink the dispersion estimates towards a fitted dispersion-mean trend. With the normalized counts and dispersion estimates, DESeq2 fits a negative binomial generalized linear model for each gene. The negative binomial distribution was used to account for the overdispersion often observed in RNA-seq count data. A Wald test was run to assess the significance of the log2 fold changes between conditions for each gene. The Wald test statistic was calculated as the log2 fold change divided by its standard error. To account for multiple testing across thousands of genes, DESeq2 adjusts the p-values using the Benjamini-Hochberg procedure to control the false discovery rate (FDR). A subset of the results of running DESeq2 is shown in (Table. 2).

Gene Name	baseMean	L2FC	lfcSE	stat	pvalue	padj
KCNIP3	386.264954	3.018724	0.358913	8.410741	4.07E-17	2.14E-13
CYP1B1	141.070908	2.164694	1.23261	1.756187	7.91E-02	6.72E-01
SMIM11A	46.788525	1.717207	0.650525	2.639724	8.30E-03	2.71E-01
DHRS2	80.737129	1.715949	0.320172	5.359464	8.35E-08	7.18E-05
PCDHGB5	75.252068	1.658963	0.176289	9.410484	4.94E-21	7.78E-17
…	…	...	…	…	…	…
AC131392.1	149.144211	-1.409846	0.536303	-2.628821	8.57E-03	2.77E-01
HTR1A	117.530663	-1.430303	0.267636	-5.344202	9.08E-08	7.18E-05
ZNF100	99.318161	-1.562824	0.236234	-6.615582	3.70E-11	8.33E-08
SLC5A12	31.017942	-1.576856	0.381051	-4.13817	3.50E-05	7.56E-03
TCEAL5	163.870865	-2.03797	0.511908	-3.981128	6.86E-05	1.29E-02

Table 2| This Table shows a subset of the dataframe output from the DESeq2 method. This was filtered by L2FC descending. The top 5 results show high log2FoldChange scores and are upregulated at this time point. Whereas the bottom five genes are downregulated.

The results of this differential expression analysis are later visualized as Volcano plots with the Log2FoldChange on the x-axis and the -log10 transformed p-values on the y-axis using the decoupler[8] and matplotlib[9] packages. To be able to keep visual comparisons easy the x-axis was limited to a range of (-10, 10) and the y-axis to a

range of (0, 300). The threshold for significance was set at 0.5 Log 2 Fold Change and 0.05 for the -log10(pvalues).

Transcription Factor Inference

To discover which transcription factors are expressed at different timepoints, I used a prior knowledge network of which TFs act on which genes. I used the CollecTRI network provided by OmniPath.

The CollecTRI network is composed of transcription factors, their target genes and the weighted links between them. These weights were computed by aggregating a vast dataset from various sources and assigning each link a mode of regulation. The weights have been normalized at the transcription factor level, where for each transcription factor and the binding weights of all its target genes were divided by the maximum binding weight observed for that particular transcription factor. The weights are normalized at the gene level, where for each gene, the binding weights of all regulating transcription factors were divided by the maximum binding weight observed for that gene. Subsequently, the links whose binding weights fell within the lowest 30th percentiles of the overall weight distribution were filtered out.

To investigate the transcription factors potentially affected by different MAFA expressions, I employed a Multivariate Linear Model (MLM) shown below (Fig. 6).

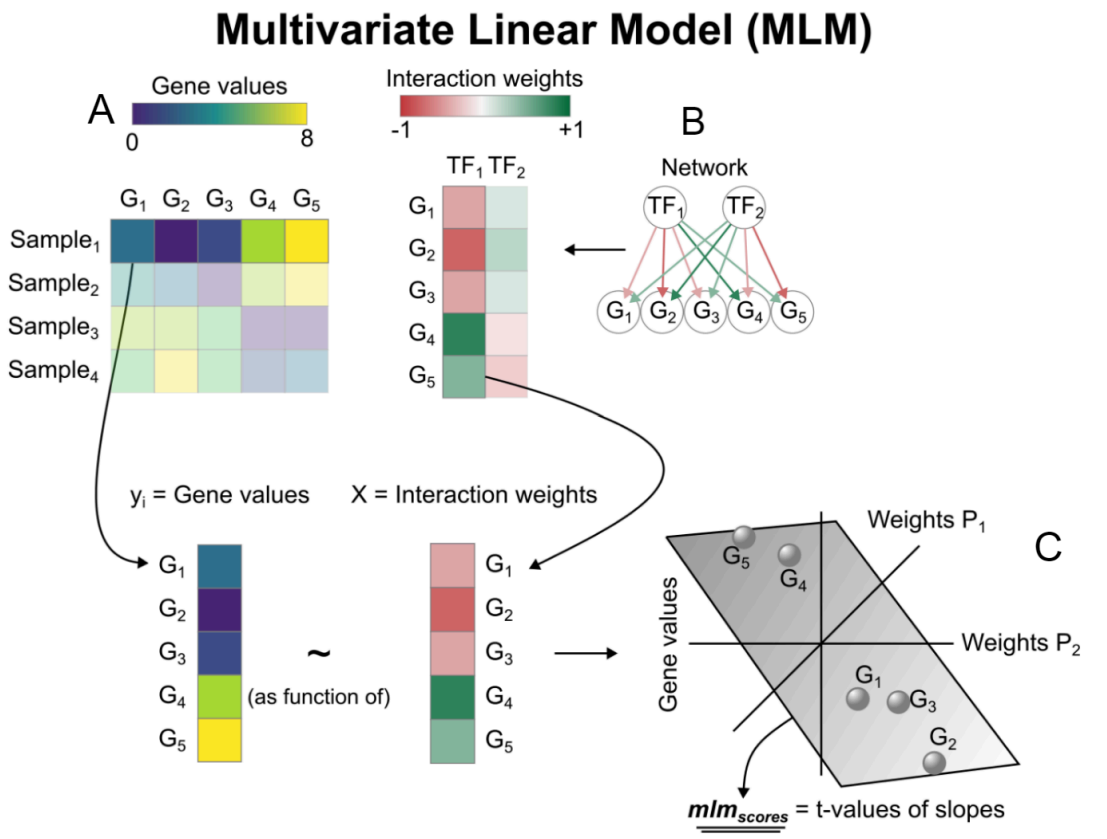

Figure 6| A diagram showing the inputs and method for calculating a predicted score with a Multivariate Linear Model.9 ***A****, Gene values for the samples I compare from the AnnData object.* ***B****, CollecTRI interaction weights between a transcription factor and the genes represented as a matrix* ***C****, The prediction from the model (the slope of a plane).*

A Multivariate Linear Model is an aggregated linear combination where the general form can be written as follows.

$$Y = XB + U$$

Here **Y** is the vector for each gene of the ‘stat’ value, which is the result of the Wald test performed by DESeq2. This was calculated by dividing a Genes log2 fold change by its standard error.
This method requires a reference network of transcription factors and their target genes. **X** is a matrix with columns of these transcription factors and rows of genes. The interaction weights in this matrix represent the edge weights of the CollecTRI network
U is an optional matrix of errors or noise which is left empty in this case.
B is the slope of the plane that I seek to predict as shown in **C** in (Fig. 6).

I was able to use this slope as a single numeric metric for how active a transcription factor is based on how many of its targets are upregulated or downregulated. If the majority of the target genes associated with a particular transcription factor exhibit upregulated expression levels, it would result in a positive slope or trend. This positive slope would indicate that the transcription factor is actively engaged.

Functional Annotation Clustering

I used Functional Annotation clustering to find out which clusters of biological annotations the targets of the transcription factors I had identified are involved in.

Data preparation

With the predicted activity scores for transcription factors and the accompanying p-values (pvalue) I narrowed down a list of interesting candidates. I filtered only those transcription factors that both have expression (Log 2 Fold Change) and activity that are predicted to be overexpressed and activated, or underexpressed and inactivated respectively.
The p-adjusted (padj) value from the differential expression analysis must also be above a given threshold.
The thresholds for the steps described above are as shown below in (Table. 3).

Threshold for **padj** (Wald Test)	0.05
Threshold for **Log 2 Fold Change**	0.5
Transcription Factor Activity **pvalue**	0.05

Table 3| Filtering parameters to narrow down transcription factors

I combined the list of remaining gene names with their corresponding Ensembl[19] (ENSG) identifiers to determine the target genes regulated by each of the remaining transcription factors. I split the targets by negative and positive Log 2 Fold Change into 2 lists.

Clustering

With these lists of positive and negatively active targets I utilized the tool DAVID[20] for an automated functional enrichment analysis and in depth look at the affected pathways. Using DAVID I dove into KEGG[21] pathways relevant to β cells and insulin production. KEGG (Kyoto Encyclopedia of Genes and Genomes) is a database resource that integrates genomic, chemical and systemic functional information. I used it to better understand the higher level functions of various cellular systems relating to insulin production and metabolism. I plotted the number of genes identified in each KEGG pathway and compared results for each of the Time Points to identify changes in enriched pathways.
DAVID takes in an input list of genes and finds similar annotations where this collection of genes occurs and clusters those together. A cluster receives a higher enrichment score if more of the input genes occur in it.

Pathway Inference

I used Pathway Inference to identify whether a change in gene expression levels has a net activating or inactivating effect on various pathways.

To investigate the pathways potentially affected by different MAFA expressions, I employed a Multivariate Linear Model (MLM). However, for this analysis I used a different reference network because I needed to know which genes were involved in various pathways, not just which transcription factors activated or inhibited them.

This interaction network is again represented as a matrix as described in the transcription factor inference but here has pathways for columns and rows of genes. The interaction weights in this matrix represent the edge weights of the network and come from the “human” dataset of pathways and targets PROGENy provided by OmniPath. For each pathway, I extracted the top 500 statistically significant interactions from the network. These interaction scores have been scaled to a mean of 0 and a standard deviation between -1 and 1. This normalization step accounts for variations in the strength of gene expression signatures, enabling a fair comparison of relative enrichment scores across different pathways and samples simultaneously.

The resulting slope is a single numeric metric for how active a pathway is based on how many genes in a given pathway are upregulated/downregulated. If most genes in a pathway are downregulated, it will result in a negative slope, leading me to conclude that this pathway is inactivated.

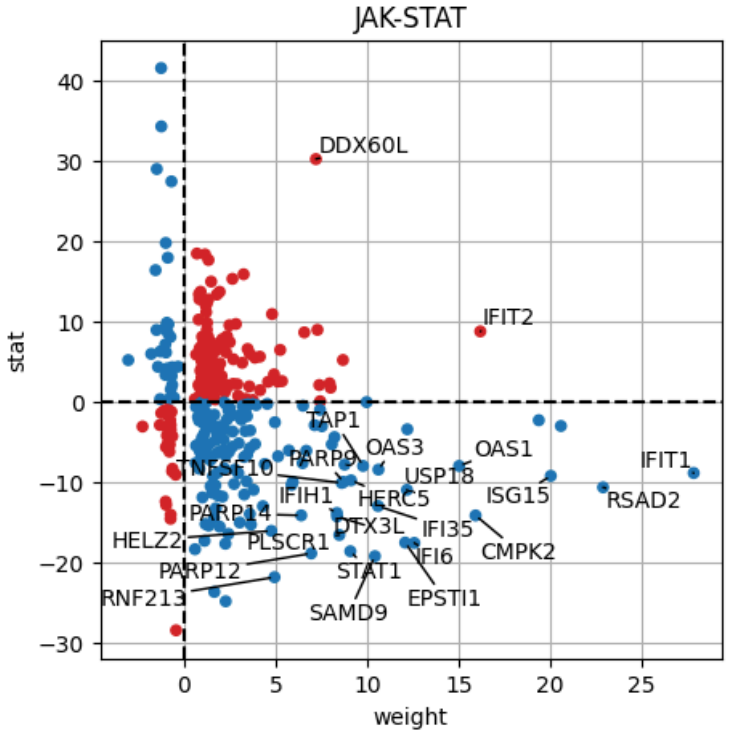

Figure 8| Cell line MAFAMUT comparing time point 2 to time point 0. This shows the targets of the JAK-STAT pathway and their stat and weight as used by the Multivariate Linear Model (MLM).

To look more closely into how these activity scores are calculated I plotted the weights and the ‘stat’ of each gene included in each pathway. In (Fig. 8) above, it is

visible that there are many genes involved in the JAK-STAT pathway and they are split in color to denote their effect. Due to there being more target genes with negative stat and a positive weight (2nd Quadrant), and the majority with negative weights have a positive stat (4th Quadrant). This observation led the model to predict a low activity for this pathway.

Verify Against Millman Human Cell Data

I compared the transcription factors I have identified as activated or inhibited to human cells to see if these transcription factors really appear in human β cells.

After performing functional annotation clustering on the post-filtered transcription factors, I sought to determine whether these regulatory factors could be detected in primary human islets and stem cell-derived islets. This comparative analysis aimed to validate the relevance of the identified transcription factors in the context of both native and stem cell-derived islet cell populations. To achieve this, I utilized a single-cell RNA-sequencing (scRNA-seq) dataset generated by the Millman Lab.

After a manual annotation of this data by Sara Jimenez, I visualized the expression of these transcription factors. This was done by first labeling cell types to be able to focus on what is expressed in which cell types. This is made possible with single cell sequencing data and the use of marker genes to identify and separate these cell types.
Also included is the data for whether the transcription factor was upregulated or downregulated as found by the Differential Expression Analysis I conducted. Whether the responsible transcription factor was inferred to act as an activator or repressor was taken into account and also annotated to the resulting visualization.

Results

The variance in differential expression of the Controls is in line with expectations

In order to determine the baseline natural variance between samples I performed a Differential Expression Analysis finding that natural variance is significantly lower than the signal.
I performed this analysis to find differential expressions across all Time Points and between cell lines. In (Fig. 9) I show the top 25 Log 2 Fold change values and their standard error for three such scenarios.

I verify that the control data displays a consistent and low level of variance.
Shown in (Fig. 9.A) is the Control $MAFA^{WT}$ at Time Point 0 compared to itself at Time Point 1. This is to see if there is any change in gene expression over time. As expected there is very little differential expression which should be brought about solely by natural variance. The highest Log 2 Fold Change (L2FC) values fall between 3 and -2.

I also compare expression differences between the $MAFA^{WT}$ and Control $MAFA^{WT}$ cell lines at Time Point 0 before the induction of MAFA begins. This shows that the starting point for both cell lines is as close to identical as variance will allow. The L2FC values from (Fig. 9.B) fall in the same range as those shown in (Fig. 9.A). I thereby conclude that the difference between the two cell lines is similar to the difference between Time Points in the Control.The similar level of L2FC values in (Fig. 9.A) and (Fig. 9.B) are in line with the expectation of little change occurring at all in the control, and that both lines have comparable starting expression levels.

However, when looking at (Fig. 9.C) both higher absolute L2FC values and lower standard error spread than those shown in (Fig. 9.A) or (Fig. 9.B). The L2FC values being roughly twice as large as those in the comparison to the Control leads me to conclude that there is clearly a detectable difference in expression levels in the $MAFA^{WT}$ cell line that is distinguishable from the background variance. The other unshown scenarios provide similar levels of evidence to this point

After establishing a baseline of natural variation in gene differential expression and confirming that the Control $MAFA^{WT}$ differential expression does not change significantly over time, the observed significant changes in differential gene expression in the $MAFA^{WT}$ sample over time warrant further analysis. With no anomalies detected in the initial assessments, the analysis can proceed to investigate the underlying causes and implications of these significant expression changes.

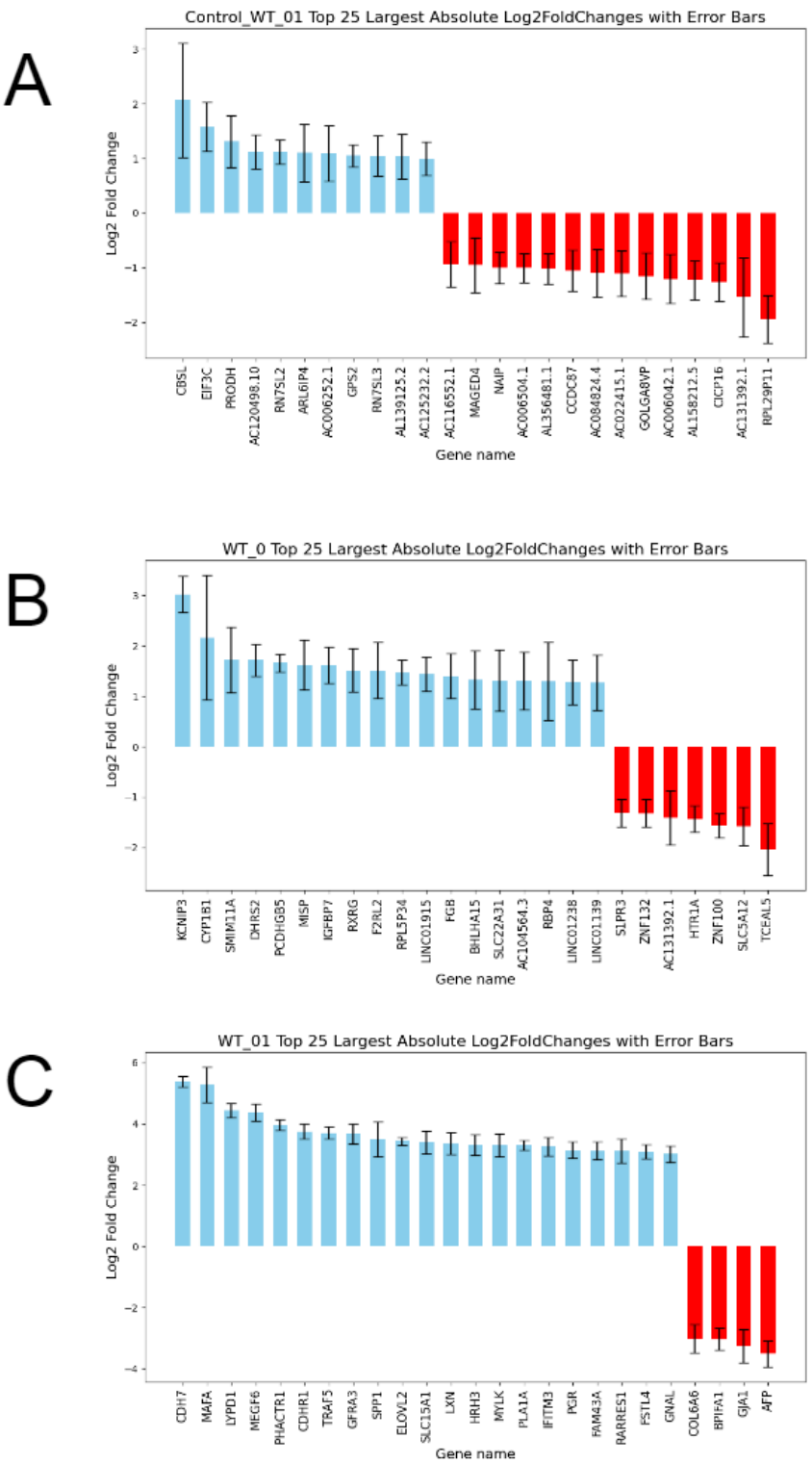

Figure 9| A comparison of the genes with the top 25 Log 2 Fold Change scores are shown with the standard error as computed as part of DESeq2. : A, Control MAFA^{WT} cell line direct comparison between Time Point 1 vs. Time Point 0. B, MAFA^{WT} Time Point 0 vs. Control MAFA^{WT} Time Point 0. C, Direct comparison of MAFA^{WT} Time Point 1 vs. Time Point 0.

The Doxycycline treatment does not induce detectable differential expression

A Doxycycline treatment is used to induce the overexpression of MAFA in both the Wild Type and the S64F mutant cell lines. The same treatment was given to the Control cell lines, though these do not contain the same cassette that subsequently produces MAFA. A question posed from the onset of this project was whether the Dox treatment caused any side effects and caused differential expression of any other genes.

To show that there is no significant side effect from the Doxycycline treatment the differential gene expression of the Control lines were analyzed showing that any overexpression/underexpression present is likely due to natural variation.

Shown in (Fig. 10) as volcano plots are the results of the differential gene expression analysis. On the x-axis are the Log 2 Fold Change values of genes. The y-axis measures the -10log(pvalues). The axes have been scaled to a L2FC range of (-10, 10) and a -10log(pvalues) range from 0 to 300. This was done to have consistent axes when visualizing all data points.
The top row in this (Fig. 10) shows the two timepoints of the Control $MAFA^{WT}$ cell line both compared to the same cell line at Time Point 0. The bottom row shows the same Time Points but for the Control $MAFA^{MUT}$ line. Visibly apparent is the small spread of values clustered at the center bottom of the plot.
This is in stark contrast to any of the explosions of color shown in (Fig. 11). Shown in this Figure are the non-Control $MAFA^{WT}$ and $MAFA^{MUT}$ cell lines. There are both bright red and blue fountains of genes to either side showing differential expression for many genes at far larger L2FC values with higher -10log(pvalues)as well.

Looking at the differential expression found when looking into differences over time in the Controls in (Fig. 11), I conclude that any overexpression/underexpression present is likely due to natural variation or noise. Perhaps it is due to a very, very small effect from the Doxycycline (DOX) treatment. Though this effect is so small it is likely negligible.
Additional evidence to this conclusion is provided when filtering activated/inhibited transcription factors for Functional annotation later. After this filtering step as described in the Data Preparation part of the Methods section, I was left with an empty list for the Controls.

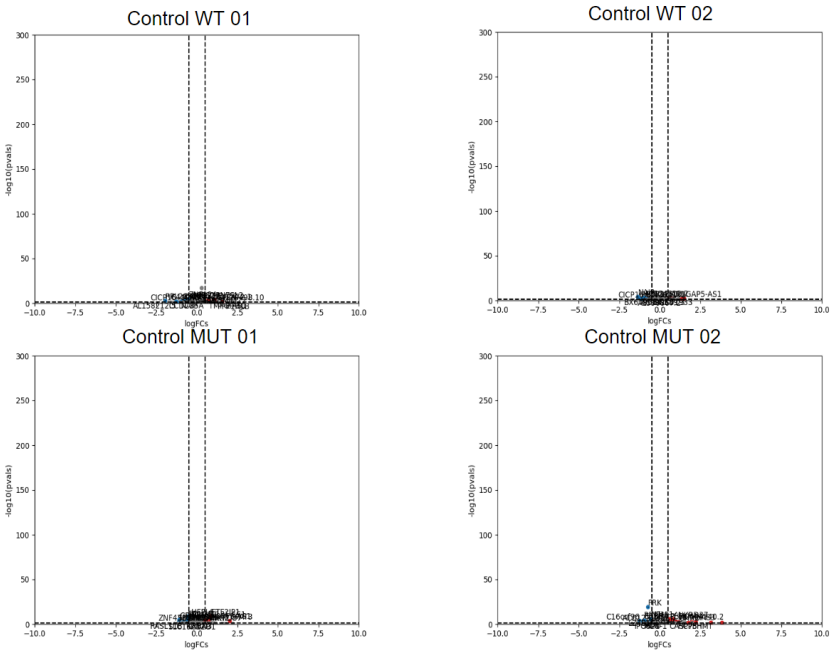

Figure 10| Volcano plots showing differential gene expression. top row: Control MAFA^{WT} Time Points 1,2 with Timpoint 0 as control, bottom row: Control MAFA^{MUT} Time Points 1,2 with Timpoint 0 as control. The Log 2 Fold Change is the X axis, -10log(pvalues) are the y axis. The right hand side colored red are genes that are overexpressed, blue coloring on the left is underexpressed.

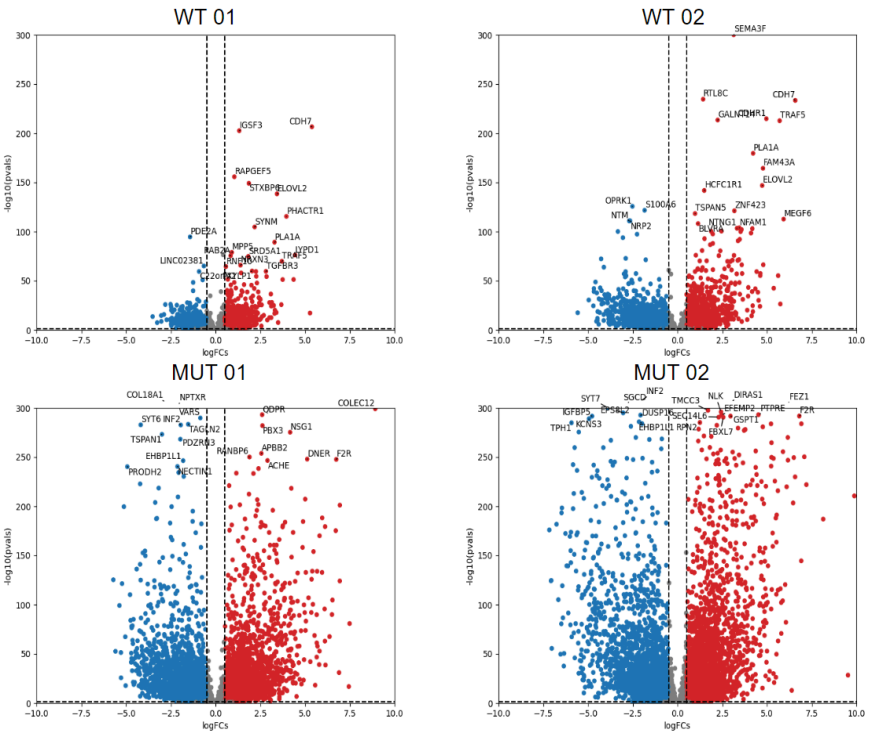

Figure 11| Volcano plots top row: MAFA^{WT} Time Points 1,2 with Timpoint 0 as control, bottom row: MAFA^{MUT} 1,2 with Timpoint 0 as control. The Log 2 Fold Change is the X axis, -10log(pvalues) are the y axis. The right hand side colored red are genes that are overexpressed, blue coloring on the left is underexpressed.

MAFA^{MUT} cell line displayed significantly higher Differential Gene Expression

Another question of interest, central to this work is the difference between the MAFA^{WT} and the MAFA^{MUT} cell lines. In order to determine the possible root cause for the differences observed between cell lines, a differential gene expression analysis was performed, displaying significantly higher differential expression levels.

From (Fig. 12) the observation can be made that there is a far greater number of both positively and negatively, differentially expressed genes. This is evident from both the comparisons of Time Point 1 and 2. This increase also seems to grow from Time Point 1 to Time Point 2.

I looked for an explanation in the activated transcription factors shown in (Fig. 13)**.** The transcription factors identified here vary between Time Points with no clear pattern emerging. The predicted activities are remarkably similar across the board given the variance in differential expression levels shown in (Fig. 12).

The MAFA^{MUT} line shows significantly larger over and underexpression of genes than the MAFA^{WT} but further analysis aided by more data is required. I can give no clear answer to either a potential cause for this disparity in differential expression.

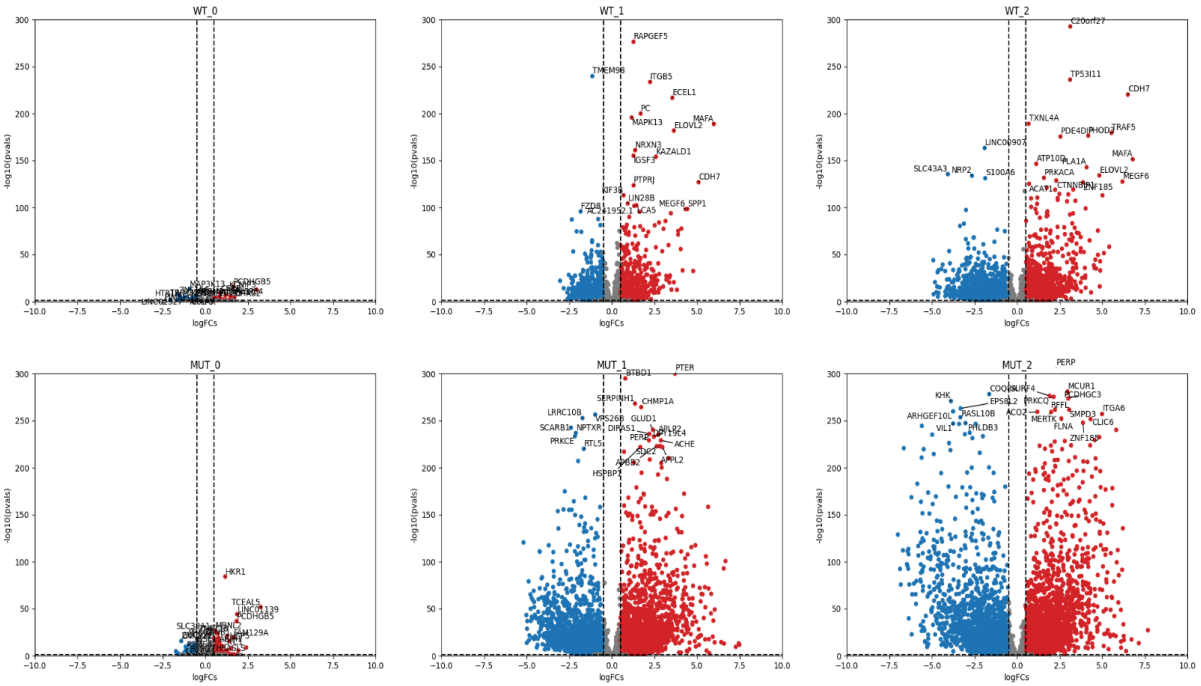

Figure 12| Volcano plots with MAFAWT Timepoints 0-1-2 on the top row, and MAFAMUT 0-1-2 on the bottom row. The Log 2 Fold Change is the X axis, -10log(pvalues) are the y axis. The right hand side colored red are genes that are overexpressed, blue coloring on the left is underexpressed.

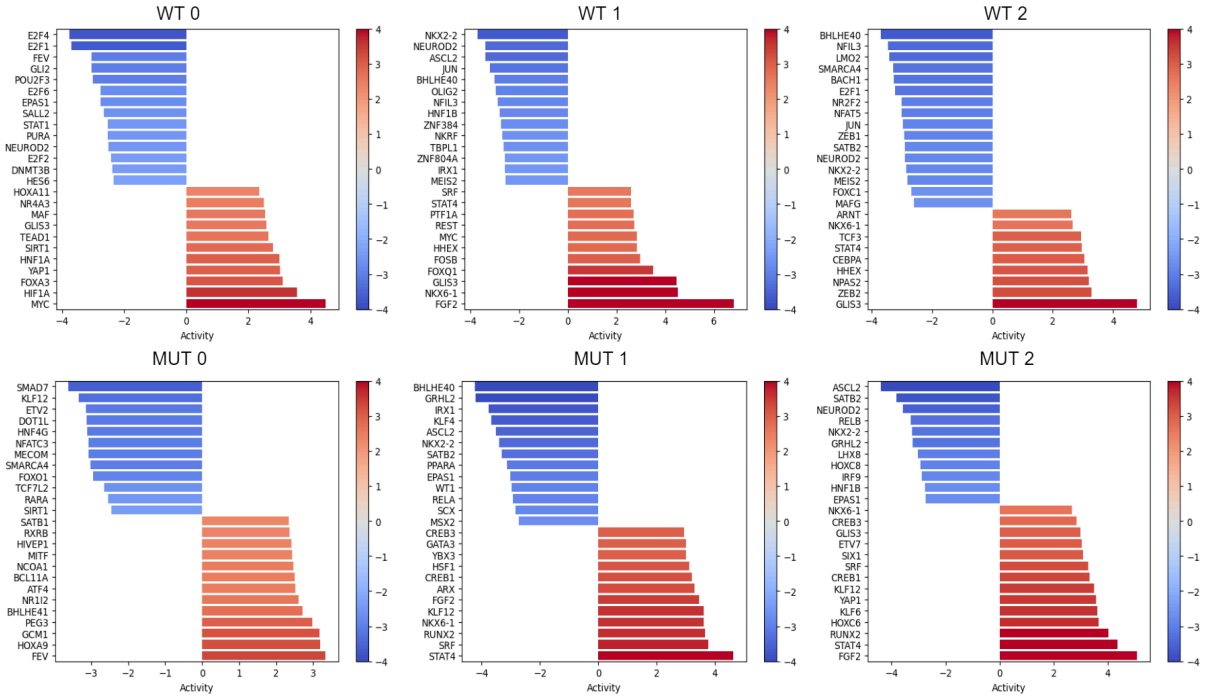

Figure 13| The transcription factor activities inferred from the multivariate linear model using Collectri. The Activity score being shown here is the t-value of the slope predicted by the MLM. The color scale on the right hand side is scaled from -4 to 4.

The Multivariate Linear Model is more useful for Transcription Factor inference with higher number of target genes

To infer transcription factor activity I employ an MLM, yielding results with a caveat.

The resultant predicted activity scores are ultimately based on a linear model composed of the input Bulk RNA data as well as the reference network CollecTRI. This reference network includes transcription factors with a wide range of targets. In (Fig. 14) there are shown the weight and wald statistic stat for each target for the transcription factors PPARGC1B and E2F1. There is a limited number of targets apparent in PPARGC1B with most being only slightly negative. On the right of the Figure in E2F1 there are dozens of targets. This disparity between the number of targets between some transcription factors is potentially biased by how well studied they are and can skew some results.
The linear model predicting a negative activity for PPARGC1B is not wrong, it just isn't as supported as any conclusion derived from the more datarich E2F1. Having a small number of data points in a linear model can lead to high activity scores predicted for networks based on a small subset of genes.
Importantly this does not indicate that the results are inaccurate. Some transcription factors simply have less target genes than another, and when those are differentially expressed the resulting prediction could be spot on. It is however something to keep in mind especially when interpreting results.

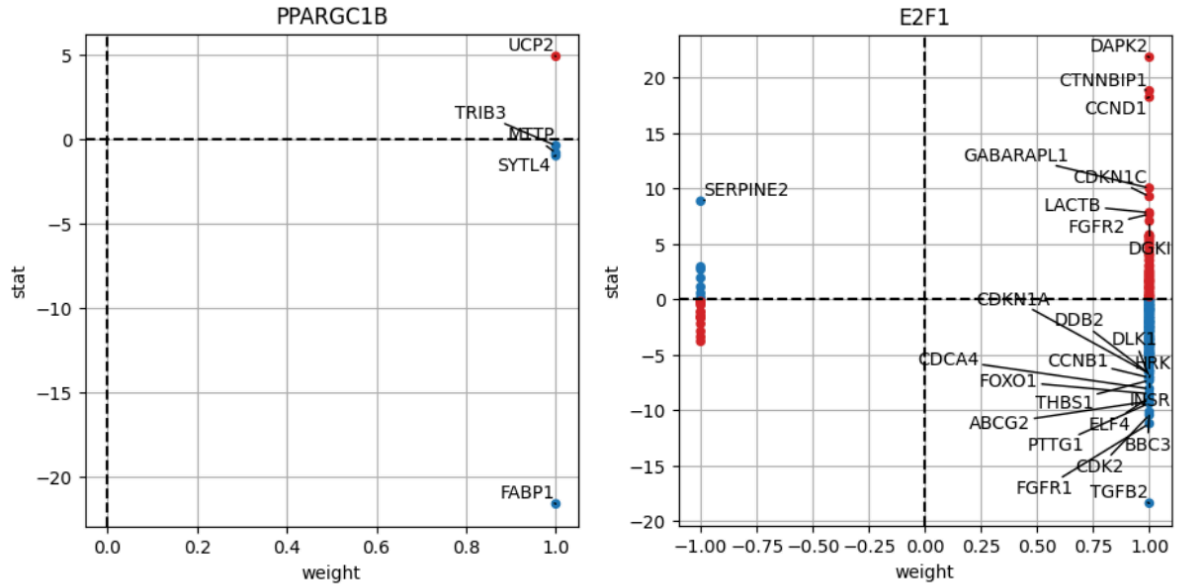

Figure 14| For the MAFA^{WT} Time Point 2 compared to Time Point 0, the targets of the transcription factors are shown: PPARGC1B on the left and E2F1 on the right. Their stat is shown on the y-axis and weight on the x-axis as used by the Multivariate Linear Model (MLM).

The Control Problem for Pathway Inference Analysis

To explore differences between the $MAFA^{WT}$ and $MAFA^{MUT}$ cell lines and to better understand the broader biological systems therein, I conducted a Pathway Inference Analysis finding that predicted pathway activity for the Controls is higher than expected.

In (Fig. 15) the pathway with the highest activity score for each Timepoint in the $MAFA^{WT}$ cell line is p53. A similar pattern emerges in the $MAFA^{MUT}$ line with Hypoxia being the dominant pathway. Given that both of these are pathways linked to cellular stress this was not the expected result I hoped to find. With the exception of the $MAFA^{MUT}$ at Time Point 0 all others show the JAK-STAT pathway as the most repressed with a rather negative score.

To find out if some of what I saw causing the large activity scores was coming from the controls I made direct comparisons between Time Points 1 and 2, and 0. These results can be seen in (Fig. 16). I see very large predicted activity scores with the JAK-STAT pathway reaching scores of over 6 in the Mut line.

Given that earlier I saw very low differential expression in all timepoints of the Control cell lines these results speak to an artifact in how the MLM functions with lower counts. I also see a strong contrast between the most active pathways like JAK-STAT in the controls and the repressed pathways I see in the $MAFA^{WT}$ and $MAFA^{MUT}$ cell lines. If JAK-STAT is predicted to be hyper expressed in the Controls, then when compared to the $MAFA^{WT}$ cell line it appears strongly inactivated.

In (Fig. 17) I observe the activity scores for direct comparisons from Time Points 1 and 2 using 0 as control. There are generally lower predicted activity scores than in the previous (Fig. 16) and (Fig. 15). This is especially the case for $MAFA^{WT}$ where the scores here fall between -2 and 2. Also notable is the fact that there is a general trend of more inhibition of pathways though I can offer no theory for why this is.

(Fig. 18) shows a heatmap of activity scores predicted for the pathways. The disparity between the predicted extremely positive activity for JAK-STAT in the Control $MAFA^{MUT}$ samples, against the inactivity predicted for all other samples is readily apparent. Visible in the dendrogram on the left hand side the Control samples are clustered closer to each other than to the same Time Point in either $MAFA^{WT}$ or $MAFA^{MUT}$ which is what I expect to find. The difference between the Control $MAFA^{MUT}$ and the Control $MAFA^{WT}$ is perplexing.

The combination of small differential expression changes perhaps contributing to a large predicted activity and the difference in predicted pathway activity scores led me to focus my analysis on differences over the time course of the cell lines by comparing to Time Point 0 instead of to the respective Controls.

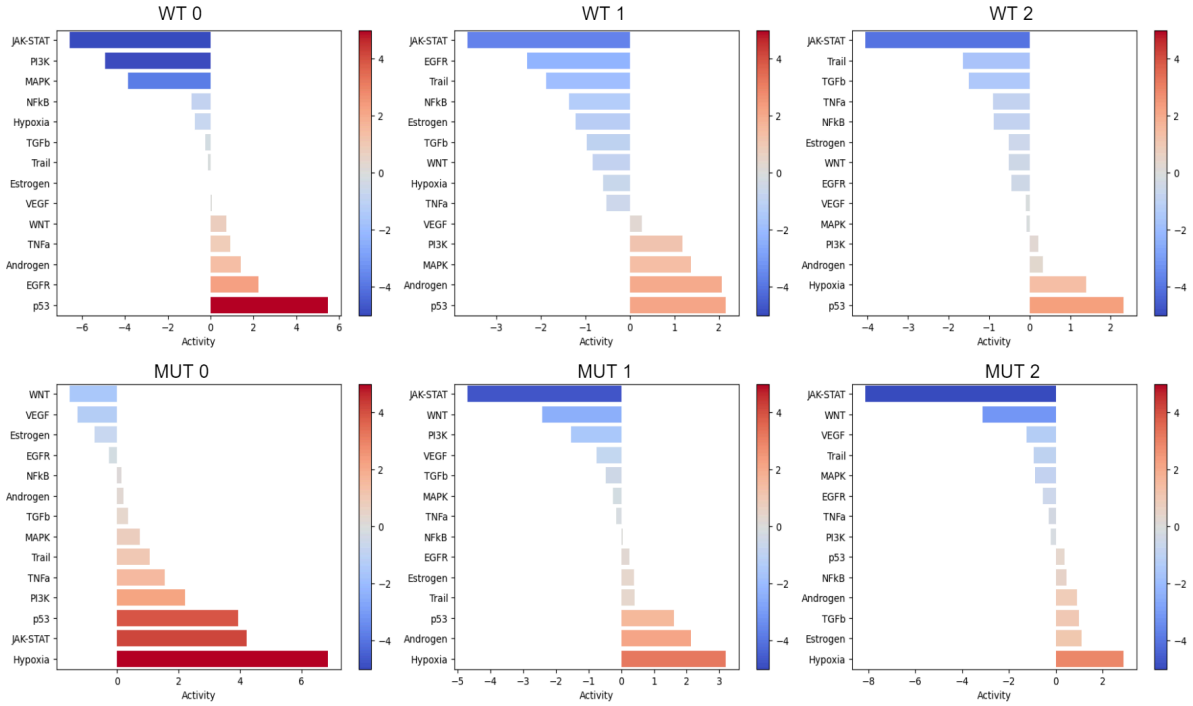

Figure 15| Shown are the inferred pathway activities for each Timepoint compared to its control. Top row: MAFA^{WT} Bottom row: MATA^{MUT}

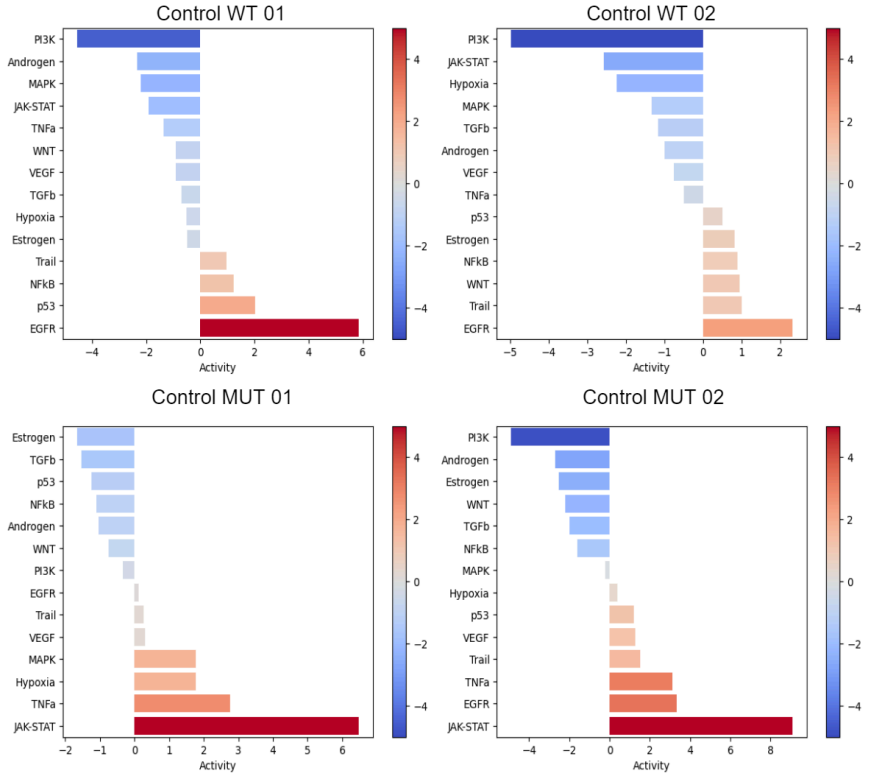

Figure 16| Inferred pathway activities for the control cell lines each Time Point compared to Time Point 0. Top row: Control MAFA^{WT} with Time Points 1 and 2 compared to 0. Bottom row: Control MATA^{MUT} Time Points 1 and 2 compared to 0.

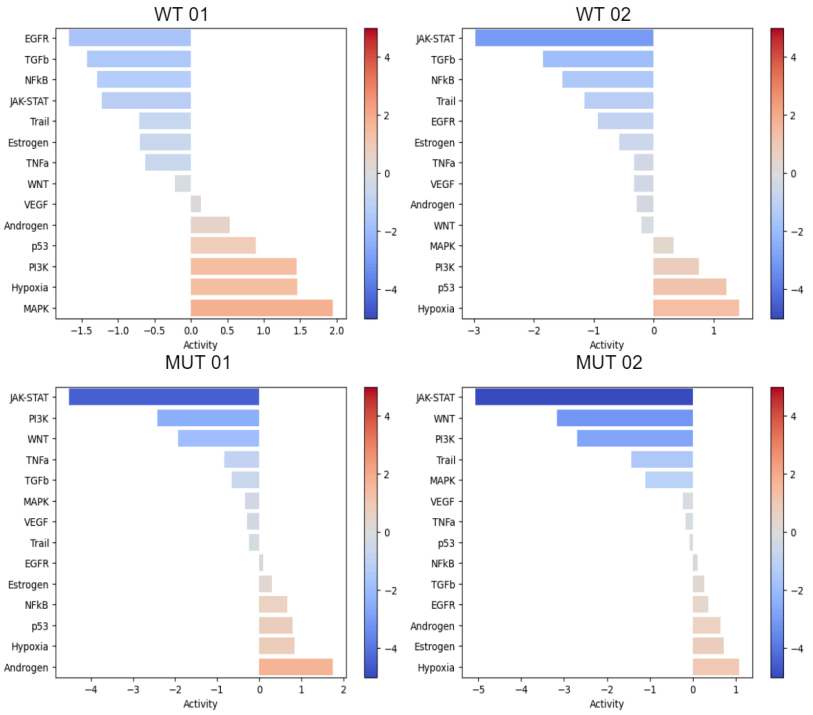

Figure 17| Inferred pathway activities for each Time Point compared to its Time Point 0. Top row: $MAFA^{WT}$ with Time Points 1 and 2 compared to 0. Bottom row: $MATA^{MUT}$ Time Points 1 and 2 compared to 0.

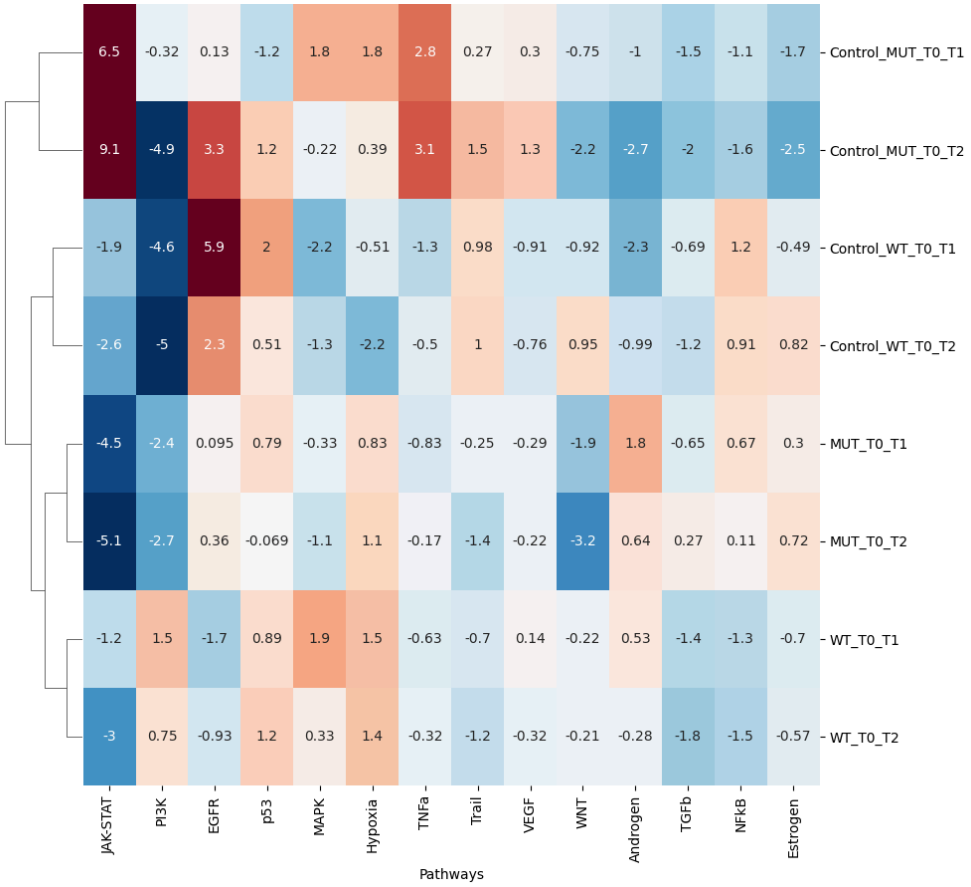

Figure 18| Heatmap showing the predicted activity scores for pathways at different Time Point comparisons.

JAK-STAT Pathway has higher gene count in MAFA^{MUT}

In order to gather more evidence for the differences in activated pathways in MAFA^{WT} and MAFA^{MUT}, I utilize DAVID to look at KEGG pathways and find that the JAK-STAT has more supporting genes in MAFA^{MUT}.

I expected to find similar pathways being activated as shown previously in (Fig. 18). However the predicted repression of the JAK-STAT pathway in the MAFA^{MUT} being greater than that of the MAFA^{WT,} is not seen reflected in the number of counted genes here.
In (Fig. 19) there are ~50% more genes counted for the JAK-STAT pathway in the MAFA^{MUT} than for the MAFA^{WT} at the first Time Point. The increase in gene count does not equate to an increase in activity, it can also be an increase in repression which we see here. There is also a three times larger count of apoptosis genes in the MAFA^{MUT} than for the MAFA^{WT}.

The JAK-STAT pathway is highly conserved and regulates many pathways in a variety of cells. A dysfunction in this pathway can cause apoptosis and downstream p53 signaling. Evidence from mouse models links a requirement of STAT1 expression in β cells for the onset of Type 1 Diabetes and could be a mechanism for β-cell loss. “Emerging evidence shows that the highly conserved and potent JAK/STAT signaling pathway is required for normal homeostasis”(Gurzov et al, 2016[22]).

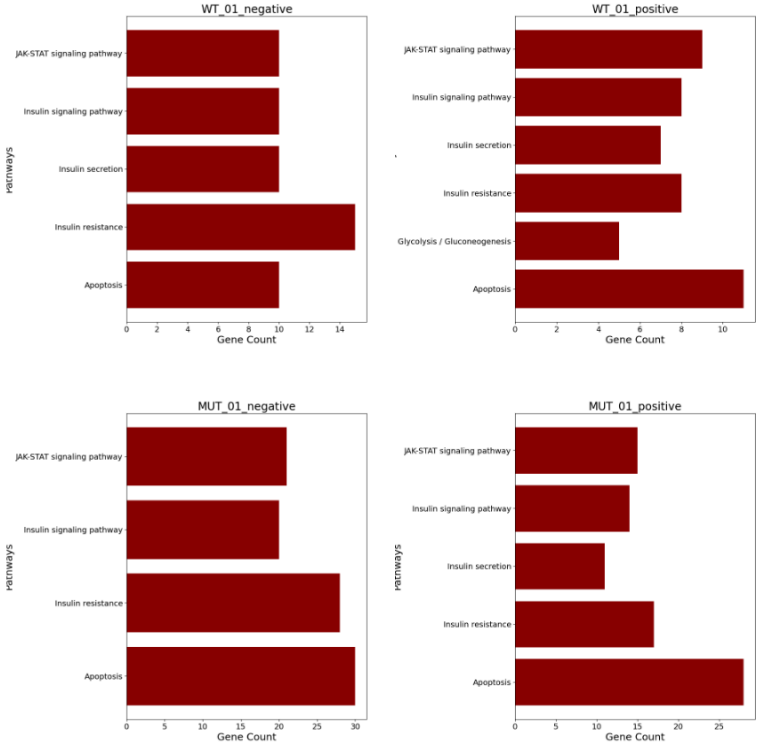

Figure 19| Horizontal bar plots showing relevant Kegg pathways x-axis gene counts in that pathway, y-axis description of Kegg pathway. List of the shown relevant pathways: ['Insulin signaling pathway', 'Insulin resistance', 'Insulin secretion', 'Glycolysis / Gluconeogenesis', 'JAK-STAT signaling pathway', 'Apoptosis']

Transcription Factors Identified prior do not match up with expression in Primary Human Islets

To verify that the transcription factors I identified above actually appear in β cells I utilize single cell RNAseq data from the Millman Lab to compare against.

I plot the measured counts in heatplots looking for which transcription factors I can find in human β cells. The cell types distinguished and clustered with the help of marker genes are shown as clusters on the left hand side of (Fig 21).
The colors from this match the top bar legend above the heatplots to group data from that type of cell together.
Whether they were up or down regulated in (DGE) my results are shown in light blue and gray on the first left hand column. Whether the transcription factor was inferred to be activated or repressed in my data is shown in red and green in the second left hand column (Activity).

I am looking for the transcription factors I identified from my analysis with Bulk RNA data being expressed in β cells in the primary human islet. In (Fig 21) I show the expression of several of the transcription factors I identified as most active for both Time Points 1 and 2 with respect to Time Point 0 in Primary human cells.
For Time Point 1 I find that many of my transcription factors are not highly expressed in the primary human islets. Several of the transcription factors I had identified are expressed more in β cells like GLIS3 or NR3C1. However these are also expressed in several other cell types. Others like ARID5B are found expressed in almost every cell type. Whereas many TFs like ASCL-1 and IRX1 are hardly expressed in any cell types.

In the second Time Point I can again identify GLIS3 as being expressed primarily in β cells. While SOX5 shows higher expression in Alpha cells, again many transcription factors do not show high activity in β cells.
This raises the interesting question about why this list of transcription factors emerged in my previous analysis. Perhaps this is due to them being specifically activated as a result of the MAFA overexpression, or as a differentiating feature of stem cells.

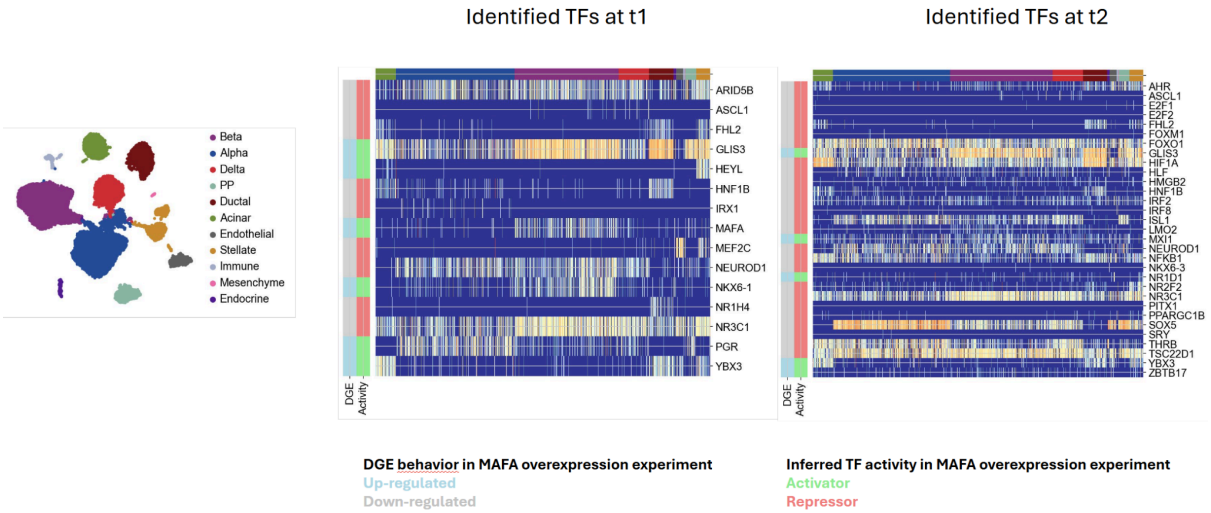

Figure 20| Transcription factor regulation in Primary human cells shown for both a comparison of Time Point 1 to Timepoint 0 and Time Point 2 to Time Point 0.

In (Fig. 22) I identify the same transcription factors in stem cell-derived human islets to test whether the expression remains consistent.
For both Time Point 1 and Time Point 2 the same genes appear, with the same expression as in the primary human islets. GLIS3 and SOX5 are both highly expressed across differing cell types. Many others show very little expression across all cell types which is labeled as dark blue.
When the same TFs show similar expression patterns for the same cell types as in the Primary Human islets, this supports the idea that the difference in expression profiles stems from the overexpression of MAFA rather than the difference in nature of stem cells.

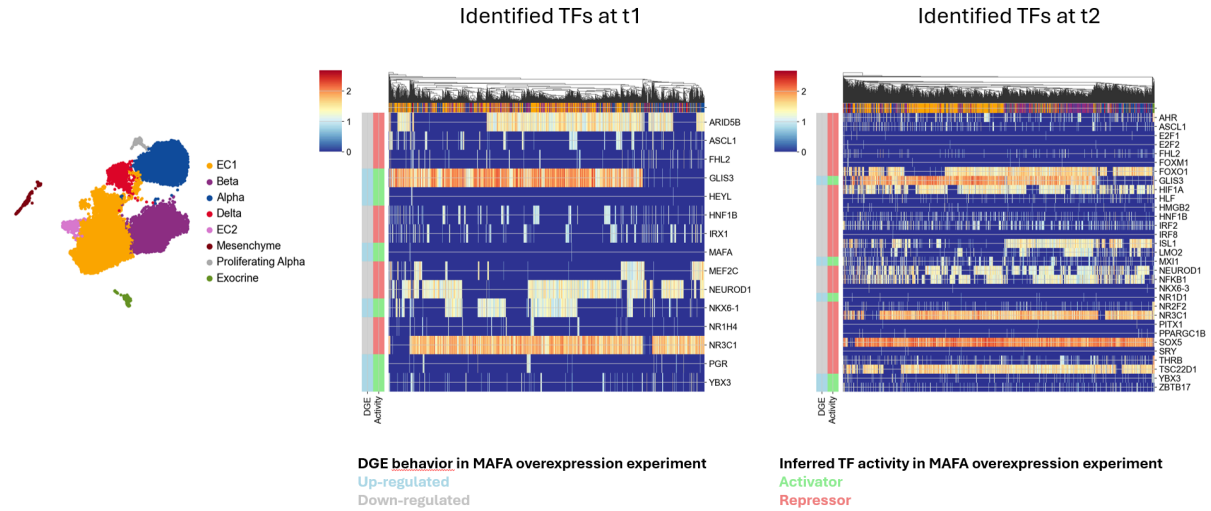

Figure 21| Transcription factor regulation in stem cell-derived islets shown for both a comparison of Time Point 1 to Time Point 0 and Time Point 2 to Time Point 0.

Discussion

While I was able to build a complete analytics process and showed the following:

- A better understanding of the insulin production process and the involved biological mechanism through the course of this project
- The Control cell lines have consistent natural variation and can be differentiated from the overexpressed cell lines.
- There are no significant side effects from Dox treatment
- There is a noticeable differential expression disparity between the $MAFA^{WT}$ and $MAFA^{MUT}$ cell lines.

I was not, however, able to positively identify a root cause for the instability in the mutant without more data. The lack of supporting data also inhibited the conclusive answering of the initial, and lofty, goal of pointing to a stabilizing factor/s for stem cells.

In this work I created a blueprint for the functional analysis of bulk RNA data. There are many figures that can be used to compare against in the future, as well as some jumping off points for a new analysis.
I identified topics that further research projects can expand upon, like the high predicted activity for JAK-STAT in the Control line for the $MAFA^{MUT}$. This behavior, not seen in the other Control line is strange, and warrants further investigation. Future work with stem cell-derived β cells can compare their results to the list of transcription factors identified here at the given time points. The large count of suppressing genes in the $MAFA^{MUT}$ at Time Point 1 when compared to the $MAFA^{WT}$ can also be a point of future study. The link between JAK-STAT and β cell loss particularly through apoptosis could play a role in the instability of the $MAFA^{MUT}$.

If not for time constraints I would have liked to try a deconvolution analysis to find cell type proportions. I received the advice that deconvolution is very dependent on the reference cells, which I did not have, and decided not to pursue this. In the future someone using a single cell dataset could compare their clustering to a deconvolution done with this bulk RNA dataset.

The original research goal of finding factors that might contribute to stability of insulin production in the $MAFA^{MUT}$ and potentially stem cells in general was based on the availability of a single cell dataset. Due to delays in collecting the data the scope had to be adjusted to what was possible to answer with the bulk RNA dataset on hand.

Utilizing single cell data would allow for the exploration of new questions which I was unable to answer with the dataset I had available. In the future it would be interesting to compare my functional analysis results to a pseudo-bulk analysis from the single cell data. With single cell data more options for analysis become available.
RNA velocity can be used to quantitatively visualize and analyze the directionality of cell trajectories over time. With this, one could more closely examine whether the expression of $MAFA^{WT}$ or $MAFA^{MUT}$ differentiates a cell closer to or further away from a human β cell. Another possible analysis with single cell data would be to infer a Gene Regulatory Network (GRN), and use that to get a better systems level understanding of the gene regulation underlying cell states and transitions.

If additional Time Points were available in future data it would also be possible to get better correlations between the expression and activity of transcription factors. This analysis would allow for a scoring and ranking of transcription factors that are both differentially expressed and predicted to be active. This was difficult to do in this analysis because of the normalization required in DESeq2 as well as the decision to look more closely at the direct comparisons in a cell line rather than comparing to the control. The result of this is that I have 2 Timepoints (T1, T2) that are compared to the Time Point 0 as the control. These 2 comparisons have several samples each, but by normalizing them it becomes a single datapoint for each Timepoint. Limited to just 2 points any correlation will always be either 1 or -1. The addition of several more Time Points would greatly expand the reliability and power of such an analysis.

With more time I would have liked to do a regression with the comparisons of the $MAFA^{WT}$ and $MAFA^{MUT}$ to the Control cell lines. This would have been 3 points, which is not ideal but would have allowed for tracking of changes in Activity over time as an interesting metric to compare the Wild Type and the Mutant.
In my work I use DAVID only as a quick way to access KEGG because that was the simplest to interpret with limited time. With more time, a deeper look into various other functional annotations that DAVID offers could uncover additional biologically relevant annotations for different levels of analysis.

The nature of the dataset restricted the possible analysis I could conduct nevertheless I hope this may act as a stepping stone for future work.

Conclusion

Based on a bulk RNA dataset I sought to identify differences between the $MAFA^{WT}$ and $MAFA^{MUT}$ cell lines and what might contribute to the instability of the $MAFA^{S64F}$ mutant.

Following best practice I verify the solidity of my dataset by checking that the Control cell lines have consistent natural variation and can be differentiated from the overexpressed cell lines. I also verify that there are no significant side effects induced by the doxycycline treatment employed to induce MAFA overexpression, ensuring that any observed effects can be attributed to the overexpression itself rather than confounding factors.
I identified that the differential gene expression profile exhibited by the $MAFA^{MUT}$ cell line displays a significantly higher level of differential expression compared to other $MAFA^{WT}$.
I found that the small number of target genes in some pathway networks from CollecTRI need to be accounted for because the predicted activity score may be “correct” but this difference in network size lowers the confidence in that result.
Unexpectedly, the pathway activity inference uncovers that the Control cell lines show large activation scores for pathways despite low differential gene expression.
The $MAFA^{MUT}$ cell line shows a higher gene count in the JAK-STAT pathway than in the $MAFA^{WT}$. This pathway's role in the apoptosis of β cells is documented and suspected to play a role in Type 1 diabetes.
The final cross check with the single cell dataset revealed an apparent lack of expression of previously identified transcription factors within the β cell population of primary human islets.

While the issues brought about by the limited dataset proved insufficient to support the original goal for this project fully, these findings provide valuable insights and lay the groundwork for future investigations. The challenges encountered highlight the complexities inherent in biological systems and the need for continued research efforts. As Albert Einstein aptly stated, "If we knew what it was we were doing, it would not be called research, would it"(Albert Einstein, 1959)?
This experience has been a great learning opportunity and has better equipped me with the skills and perspective necessary for future explorations.

Code Availability

https://github.com/theislab/MAFA_multiome_wip
https://github.com/AlexanderFastner/MAFA_functional_analysis

List of Figures and Tables

List Of Figuress

- Figure 1| The differentiation process and the timeline for Dox treatment. To ensure the same maturity of the cells all samples were taken on the same end date, having undergone the treatment for differing durations.
- Figure 2| glucose-stimulated insulin secretion (GSIS) results after Doxycycline treatment leading to MAFA overexpression in both the MAFAWT and MAFAS64F for the different data points.
- Figure 3| Schema depicting the experiment and how $MAFA^{WT}$ and $MAFA^{MUT}$ (S64F) data at day 0, day -7, and day -14 was generated.
- Figure 4| Schema explaining the order of Analyses, the input data sets, and the Questions I am trying to answer
- Figure 5| This heatmap shows the data retained after the filtering and denoising process. The colors indicate the absolute number of genes that fall in that window of the log10 total sum of counts.
- Figure 6| A diagram showing the inputs and method for calculating a predicted score with a Multivariate Linear Model.9 A, Gene values for the samples I compare from the AnnData object. B, CollecTRI interaction weights between a transcription factor and the genes represented as a matrix C, The prediction from the model (the slope of a plane).
- Figure 8| Cell line MAFAMUT comparing time point 2 to time point 0. This shows the targets of the JAK-STAT pathway and their stat and weight as used by the Multivariate Linear Model (MLM).
- Figure 9| A comparison of the genes with the top 25 Log 2 Fold Change scores are shown with the standard error as computed as part of DESeq2. : A, Control $MAFA^{WT}$ cell line direct comparison between Time Point 1 vs. Time Point 0. B, $MAFA^{WT}$ Time Point 0 vs. Control $MAFA^{WT}$ Time Point 0. C, Direct comparison of $MAFA^{WT}$ Time Point 1 vs. Time Point 0.
- Figure 10| Volcano plots showing differential gene expression. top row: Control $MAFA^{WT}$ Time Points 1,2 with Timpoint 0 as control, bottom row: Control $MAFA^{MUT}$ Time Points 1,2 with Timpoint 0 as control. The Log 2 Fold Change is the X axis, -10log(pvalues) are the y axis. The right hand side colored red are genes that are overexpressed, blue coloring on the left is underexpressed.
- Figure 11| Volcano plots top row: $MAFA^{WT}$ Time Points 1,2 with Timpoint 0 as control, bottom row: $MAFA^{MUT}$ 1,2 with Timpoint 0 as control. The Log 2 Fold Change is the X axis, -10log(pvalues) are the y axis. The right hand side colored red are genes that are overexpressed, blue coloring on the left is underexpressed.
- Figure 12| Volcano plots with MAFAWT Timepoints 0-1-2 on the top row, and MAFAMUT 0-1-2 on the bottom row. The Log 2 Fold Change is the X axis,

-10log(pvalues) are the y axis. The right hand side colored red are genes that are overexpressed, blue coloring on the left is underexpressed.
- Figure 13| The transcription factor activities inferred from the multivariate linear model using Collectri. The Activity score being shown here is the t-value of the slope predicted by the MLM. The color scale on the right hand side is scaled from -4 to 4.
- Figure 14| For the MAFAWT Time Point 2 compared to Time Point 0, the targets of the transcription factors are shown: PPARGC1B on the left and E2F1 on the right. Their stat is shown on the y-axis and weight on the x-axis as used by the Multivariate Linear Model (MLM).
- Figure 15| Shown are the inferred pathway activities for each Timepoint compared to its control. Top row: $MAFA^{WT}$ Bottom row: $MATA^{MUT}$
- Figure 16| Inferred pathway activities for the control cell lines each Time Point compared to Time Point 0. Top row: Control $MAFA^{WT}$ with Time Points 1 and 2 compared to 0. Bottom row: Control $MATA^{MUT}$ Time Points 1 and 2 compared to 0.
- Figure 17| Inferred pathway activities for each Time Point compared to its Time Point 0. Top row: $MAFA^{WT}$ with Time Points 1 and 2 compared to 0. Bottom row: $MATA^{MUT}$ Time Points 1 and 2 compared to 0.
- Figure 18| Heatmap showing the predicted activity scores for pathways at different Time Point comparisons.
- Figure 19| Horizontal bar plots showing relevant Kegg pathways x-axis gene counts in that pathway, y-axis description of Kegg pathway. List of the shown relevant pathways: ['Insulin signaling pathway', 'Insulin resistance', 'Insulin secretion', 'Glycolysis / Gluconeogenesis', 'JAK-STAT signaling pathway', 'Apoptosis']
- Figure 20| Transcription factor regulation in Primary human cells shown for both a comparison of Time Point 1 to Timepoint 0 and Time Point 2 to Time Point 0.
- Figure 21| Transcription factor regulation in stem cell-derived islets shown for both a comparison of Time Point 1 to Time Point 0 and Time Point 2 to Time Point 0.

List Of Tables

- Table 1| List of various resources used in this paper, their sources, and links to where to find them.
- Table 2| This Table shows a subset of the dataframe output from the DESeq2 method. This was filtered by L2FC descending. The top 5 results show high log2FoldChange scores and are upregulated at this time point. Whereas the bottom five genes are downregulated.
- Table 3| Filtering parameters to narrow down transcription factors

Bibliography

1. Imperatore G, Boyle JP, Thompson TJ, et al. Projections of Type 1 and Type 2 Diabetes Burden in the U.S. Population Aged <20 Years Through 2050. *Diabetes Care*. 2012;35(12):2515-2520. doi:10.2337/dc12-0669
2. Warshauer JT, Bluestone JA, Anderson MS. New Frontiers in the Treatment of Type 1 Diabetes. *Cell Metab*. 2020;31(1):46-61. doi:10.1016/j.cmet.2019.11.017
3. Banting FG, Best CH, Collip JB, Campbell WR, Fletcher AA. Pancreatic Extracts in the Treatment of Diabetes Mellitus. *Can Med Assoc J*. 1922;12(3):141-146.
4. Zhu H, Wang G, Nguyen-Ngoc KV, et al. Understanding cell fate acquisition in stem-cell-derived pancreatic islets using single-cell multiome-inferred regulomes. *Dev Cell*. 2023;58(9):727-743.e11. doi:10.1016/j.devcel.2023.03.011
5. Cochrane VA, Hebrok M. Stem cell-derived islet therapy: is this the end of the beginning? *Nat Rev Endocrinol*. 2023;19(12):681-682. doi:10.1038/s41574-023-00910-8
6. Augsornworawat P, Hogrebe NJ, Ishahak M, et al. Single-nucleus multi-omics of human stem cell-derived islets identifies deficiencies in lineage specification. *Nat Cell Biol*. 2023;25(6):904-916. doi:10.1038/s41556-023-01150-8
7. Virshup I, Rybakov S, Theis FJ, Angerer P, Wolf FA. anndata: Annotated data. Published online December 19, 2021. doi:10.1101/2021.12.16.473007
8. Badia-i-Mompel P, Vélez Santiago J, Braunger J, et al. decoupleR: ensemble of computational methods to infer biological activities from omics data. Kuijjer ML, ed. *Bioinforma Adv*. 2022;2(1):vbac016. doi:10.1093/bioadv/vbac016
9. Team TMD. Matplotlib: Visualization with Python. Published online September 15, 2023. doi:10.5281/ZENODO.8347255
10. Türei D, Valdeolivas A, Gul L, et al. Integrated intra‐ and intercellular signaling knowledge for multicellular omics analysis. *Mol Syst Biol*. 2021;17(3):e9923. doi:10.15252/msb.20209923
11. Muzellec B, Teleńczuk M, Cabeli V, Andreux M. PyDESeq2: a python package for bulk RNA-seq differential expression analysis. Ponty Y, ed. *Bioinformatics*. 2023;39(9):btad547. doi:10.1093/bioinformatics/btad547
12. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol*. 2018;19(1):15. doi:10.1186/s13059-017-1382-0
13. Olivier Grisel, Andreas Mueller, Lars, et al. scikit-learn/scikit-learn: Scikit-learn 1.5.0. Published online May 21, 2024. doi:10.5281/ZENODO.591564
14. Virtanen P, Gommers R, Oliphant TE, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods*. 2020;17(3):261-272. doi:10.1038/s41592-019-0686-2
15. The pandas development team. pandas-dev/pandas: Pandas. Published online April 10, 2024. doi:10.5281/ZENODO.10957263
16. Schubert M, Klinger B, Klünemann M, et al. Perturbation-response genes reveal signaling footprints in cancer gene expression. *Nat Commun*. 2018;9(1):20. doi:10.1038/s41467-017-02391-6
17. Müller-Dott S, Tsirvouli E, Vazquez M, et al. Expanding the coverage of regulons from high-confidence prior knowledge for accurate estimation of transcription factor activities. *Nucleic Acids Res*. 2023;51(20):10934-10949. doi:10.1093/nar/gkad841
18. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15(12):550. doi:10.1186/s13059-014-0550-8
19. Yates AD, Achuthan P, Akanni W, et al. Ensembl 2020. *Nucleic Acids Res*. Published online November 6, 2019:gkz966. doi:10.1093/nar/gkz966
20. Sherman BT, Hao M, Qiu J, et al. DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Res*. 2022;50(W1):W216-W221. doi:10.1093/nar/gkac194

21. Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*. 2000;28(1):27-30. doi:10.1093/nar/28.1.27
22. Gurzov EN, Stanley WJ, Pappas EG, Thomas HE, Gough DJ. The JAK / STAT pathway in obesity and diabetes. *FEBS J*. 2016;283(16):3002-3015. doi:10.1111/febs.13709