# Investigating how heterogeneity of beta cells according to *Pdx1* expression impacts function using single-cell RNA sequencing

**Introduction**

Diabetes is estimated to affect 4.7 million people in the United Kingdom and over 420 million people globally (1,2). Diabetes involves insulin, a hormone which regulates blood sugar.

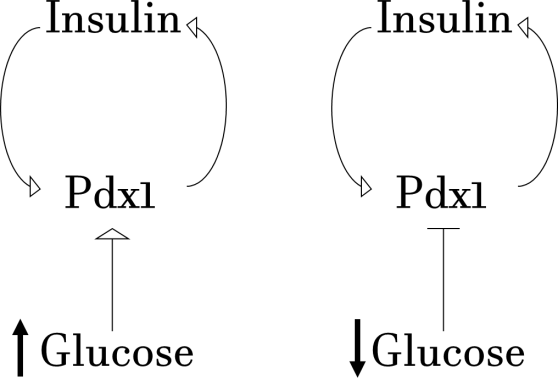
Type 1 diabetes is due to insufficient amounts of insulin being produced by the pancreas, and this can’t be prevented. Type 2 diabetes occurs when the insulin produced is not efficiently used by the body and is often a result of an unhealthy diet and lack of exercise. Both types result in sufferers having a constant state of high blood glucose levels if not treated with exogenous insulin. With the majority of diabetes cases worldwide being type 2, the number of diagnoses globally is expected to grow in the next few years (1).

Production of insulin is carried out by beta cells in the islets of Langerhans which make up 1% of pancreas (3). There are diverse cells that make up the islets of Langerhans, roughly 30% are α-cells, 60% are β-cells, and the remaining 10% are a mixture of δ-cells, γ- or PP cells and ε-cells (4).

Beta cells make up 70-80% of the islet mass indicating the importance of their function (4). These cells release insulin when glucose levels surpass normal concentration in the blood, known as hyperglycaemia. Insulin acts on cells such as adipose and liver to increase uptake of glucose in the blood for energy storage. Alpha cells in the islets act to reverse the effect of beta cells whereby releasing glucagon when there are low levels of glucose in the blood, known as hypoglycaemia

Type 1 diabetes is a chronic autoimmune disease where lymphocytes do not see beta cells as self. This leads to beta cell apoptosis and lack of insulin production (5). Type 2 diabetes occurs due to insulin resistance where insulin receptors on cells do not respond to insulin stimulation. This leads to hyperglycaemia and increased insulin production of beta cells resulting in cell exhaustion (6).

*Pdx1* (Pancreatic duodenal homeobox 1) is a transcription factor that regulates pancreatic cell development and more importantly insulin production (3) (figure 1). Heterogenous mutations of *Pdx1* in humans are noted to lead to early onset of diabetes (7), whereas homogenous mutations lead to pancreatic agenesis in *Pdx1* null mice (8). This highlights the importance of *Pdx1* in development of pancreatic cells and the role in insulin production required for glucose homeostasis.



*Figure 1* Relationship between *Pdx1*, insulin and glucose. Increased glucose concentrations in the blood results in insulin release, production which is regulated by *Pdx1*. Decreased glucose concentration leads to decreased *Pdx1* and insulin production.

In diabetic animal models with type 2 diabetes, downregulation of *Pdx1* were found as well as beta cell impaired function and cell apoptosis due to excess glucose and lipids ( (9,10). *Pdx1* is postulated to self-regulate itself, where *Pdx1* aids in insulin production and in turn insulin regulates *Pdx1* production, as illustrated in Figure 1 (11).

*Pdx1* is expressed during pancreatic development and then specifically in beta cells only post differentiation (12,13). Loss of *Pdx1* in beta cells are found to increase beta cell apoptosis, loss of beta cell identity and transdifferentiation to other cells within the pancreas (12). It would be interesting to investigate how different levels of *Pdx1* can change the behaviour or pathway of beta cells.

Studies which have performed RNA sequencing (RNA-seq) of the whole islet have suggested there is heterogeneity among beta cells (14). Bulk RNA-seq of the whole islet can be used to gain some insight into gene expression, however expression levels of abundant genes between non-diabetic and type 2 diabetic islet cells show only marginal differences (15). Furthermore, the use of bulk RNA sequencing methods have resulted in different studies reporting contradictory findings on genes and pathways involved in diabetes (16).

This is because only the most abundantly expressed genes are detected, and smaller patterns such as cell-type specific genes are obscured, therefore a finer level of resolution is required to increase our understanding of the disease (17).

Despite being a relatively new technology, single-cell RNA sequencing (scRNA-seq) has already had impact in increasing our understanding of transcriptomics across and between individual cells. This relatively new technology has allowed greater variability to be captured within the pancreatic islet, and shown that the six main cell types have hundreds of differences in the genes expressed, with some genes only being expressed in specific cells (17).

Tools designed for analysis of scRNA-seq are emerging and have been developed in recent years, however further progression is needed for optimal analysis and results (18). RNA sequencing of individual cells is more difficult compared to bulk RNA-seq due to limited data input. There can also be issues with the methods creating noise and variability in data, which can be mis-interpreted as genetic variation, thus results need to be validated (18).

Other methods such as deep sequencing and multiregional sequencing are alternatives to capture subpopulations of cells, with more genotype information available to be analysed due to being cloned in culture. However, different abilities of replicating in culture can result in strong bias in the proportion of cells present compared to the whole population. Also, adaptation to the culture environment can cause transcription and epigenetic changes in the cells, influencing results. (18). Despite challenges, correlation analysis between single cell average gene expression and bulk-RNA sequencing from the islet of a patient has shown that scRNA data is representative of the islet as a whole (15,17). Therefore, the use of the technology is appropriate for studies within this field.

In addition to identifying the major cell types, scRNA-seq allows us to compare different cells of the same type. Multiple studies have shown there to be genetic heterogeneity within beta cells with distinctly different patterns of gene expression, indicating subpopulations (17,19).

It is estimated that up to 20% of beta cells form a subgroup which express low *Pdx1* and insulin. Real-time PCR analysing differences in gene expression suggests that these low *Pdx1* expressing cells are immature and secrete significantly less insulin than mature cells. A subgroup of these immature cells can then transition to become mature high expressing *Pdx1* cells (20). However, not all immature beta cells progress to express more *Pdx1*, suggesting that this small group of lowly-expressing *Pdx1* cells may have functional importance. As *Pdx1* is linked to insulin production and secretion, understanding changes in gene expression in beta cells in the context of diabetes at the single cell level has potential be utilised to prevent, diagnose and develop new treatment options for diabetes.

A previous study showed the importance of heterogeneity of beta cells, identifying “hubs” which are distributed throughout the islet. Silencing these hub cells caused significantly greater dysfunction in the coordination of the islet responses to glucose levels than when non-hub or “follower” beta cells were silenced (21). Silencing hubs cells also had detrimental impacts to the release of insulin compared to the silencing of non-hub cells, therefore these hub cells are required for insulin secretion (21). The authors suggest that hub cells differ in function through increased proton pumping, ATP synthase activity and ATP generation (21). These findings suggest a likely cause to type 2 diabetes, and its applications to the development of therapeutics may have huge implications in treating diabetes patients.

As hub cells have metabolic signature similar to that which characterises immature beta cells, it has been suggested that there may be a link between the low *Pdx1* expression and hub beta cells (21).

Further highlighting the importance of beta cell heterogeneity, forcing overexpression of *Pdx1* in mice impairs how the islet responds to stimuli, including glucose-stimulated and incretin-stimulated insulin secretion. Similar findings were reported when *Pdx1* expression was silenced across the islet (22). This indicates that *Pdx1* expression level is important in the functionality Islet of Langerhans, and therefore may contribute to diabetes pathogenicity.

Based on findings from previous studies, we hypothesise beta cells that express low levels of *Pdx1* will have different pathways activated compared to beta cells expressing high levels of *Pdx1*. In this study, we aim to use published mouse scRNA-seq data to investigate this heterogeneity to infer its functional importance within the islet.

**Methods**

**Creating Count Matrix**

Single-cell RNA-seq data from mouse pancreatic islets, originally obtained by Qiu et al (23), was downloaded using the European Nucleotide Archive (ENA). Quality control of the sequencing reads was achieved using FastQC (v0.11.5) (24). Trim Galore (v0.4.2) (25) was used to remove adapter sequences and reads with a Phred quality score below 20. The GRCm38.p6 version of the *Mus musculus* genome and the genome annotations GTF file were downloaded from GENCODE. Indexing was performed using STAR (v2.5.2b) (26) to create the reference mouse genome for our study. The good quality single-cell RNA-seq reads were then aligned to this reference sequence, also using STAR. Finally, we extracted the reads per gene values, enabling construction of a count matrix.

# Data Processing

The non-mapped reads were removed from the count matrix. The count values in the matrix were normalised using SCnorm (v1.4.7) (27) using default parameters. Next, we selected for the beta cells from the islet using *Ins2* as a marker gene. *Ins2-*positive cells were defined as having a read count of over 500, as this accounts for insulin staining of cells and corresponds to the expected proportion of beta cells within the mouse islet (28).These *Ins2*- positive cells were separated into three equally sized groups, labelled low, medium and high, based on their *Pdx1* read count value. All Ensembl gene IDs were then converted to gene names using the package biomaRt (v2.38.0) (29,30).

# Data Exploration

Expression levels of *Pdx1* within all insulin positive cells was plotted as a histogram to confirm that the distribution of *Pdx1* within beta cells is varied. *Pdx1* counts within each group were plotted to visualise how expression level changes in the low, medium and high groups using the R package SC3 (v3\_1.10.1) (31). Unsupervised clustering analysis of the normalised count matrix was completed using the mclust package (v5.4.2) in R to investigate how many clusters the dataset optimally separates into (32). Also using SC3, principle component analysis (PCA) was performed and plotted to identify variation within the data and any extreme outliers.

# Differential Gene Expression Analysis

Differentially expressed genes between high and low *Pdx1* cells were identified using DESeq2 (v1.22.2) (33). Only significantly different genes with an adjusted p-value < 0.05 were considered. To group genes according to the similarity of their function, gene functional classification was performed on genes that were significantly upregulated in high *Pdx1* cells and genes which were significantly downregulated in high *Pdx1* cells by entering the Ensembl gene IDs into DAVID (v6.8) (34,35) . DAVID functional annotation was then carried out to identify the gene ontology (GO) terms associated with each list, filtering specifically for biological processes (GOTERM\_BP\_ALL, GOTERM\_BP\_DIRECT), and clusters them together based on being part of the same functional pathway. Both of these analyses were performed using a range of stringency parameters.

# Data Availability

Single-cell RNAseq data accession number: [GSE87375](http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE87375). Reference mouse genome data and annotations are available from: <https://www.gencodegenes.org/mouse/>

# Code Availability

All codes used are provided in [supplementary information](file://localhost/E:/Report/supplementary_info.docx).

**Results**

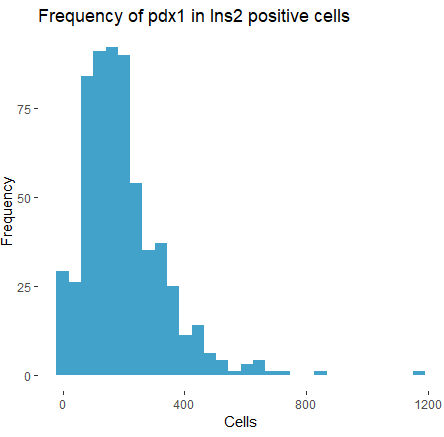
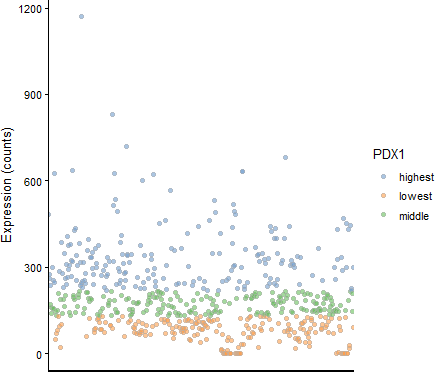
**Data Exploration of Beta Cells**

Within beta cells, we confirmed that there is a varied distribution of *Pdx1* expression levels (figure 2a), concordant with previous research (20). Figure 2b shows how we separated the beta cells into three equal groups based on their *Pdx1* expression level.

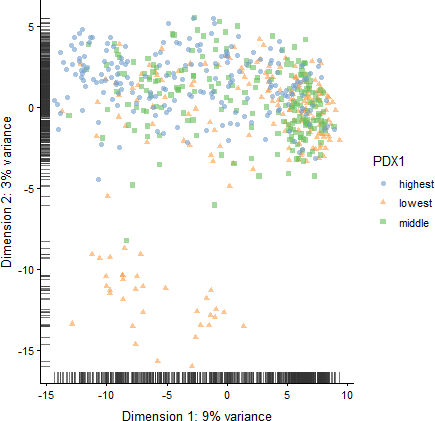
Clustering analysis on the top two principle components suggests that three clusters may be the optimum to explain the data (figure 2c). This could potentially correspond to *Pdx1* expression level. Principle component analysis does show there is variation between the different beta cells and that no extreme outlier cells were identified (figure 2d). The lowest *Pdx1*-expressing cells appear distinctly different from the medium and high groups.

However, the cells do not form three distinct clusters corresponding to *Pdx1* group, as the medium and top thirds are similar. Therefore, to ensure that the cells we are comparing are biologically distinct from each other, we selected the top and lowest thirds for comparative analysis. We chose not to split the distribution in half because cells either side of the boundary would be very similar to each other, potentially confounding our downstream analysis.

1. *b)*

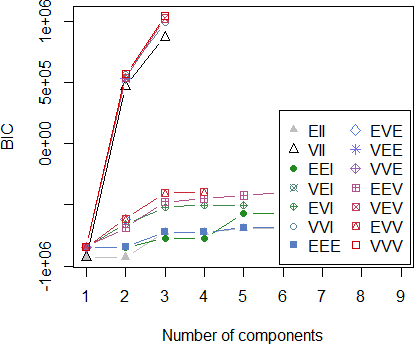






*d)*

*c)*

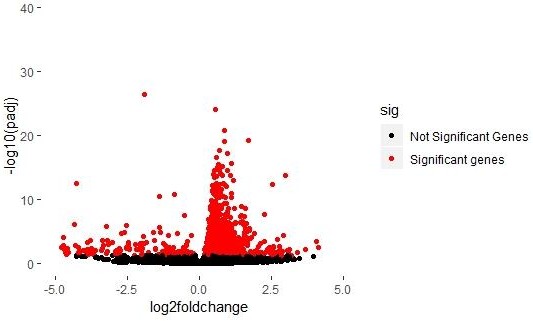


*Figure 2*

* 1. Expression levels of *Pdx1* within insulin positive (beta) cells, showing a varied and skewed distribution.
  2. Expression levels of *Pdx1* within the three groups of cells we created, confirming that they have been divided correctly.
  3. Unsupervised clustering analysis from mclust, plotting number of clusters against the Bayesian Information Criterion measurement (BIC). The highest BIC is returned when three clusters are used, and this plateaus as the number of clusters is increased further, suggesting this may be the optimum to describe our normalised gene expression data.
  4. Top two principle components identified in PCA of the normalised count data, showing that middle and high expressing *Pdx1* cells are more similar than when comparing cells which express *Pdx1* at a high and low level.

# Statistical Results

DESeq analysis found 54446 genes to be differentially expressed between the low *Pdx1* and high *Pdx1* cell groups, of that 1628 were found to be statistically significant (figure 3). 1479 of these genes were upregulated in high *Pdx1* cells compared to low *Pdx1* cells, but only 149 were downregulated in high *Pdx1* cells compared to low *Pdx1* cells.

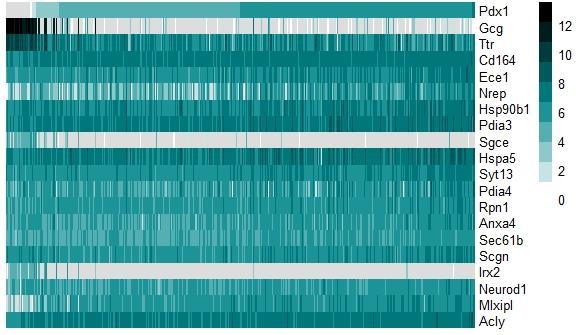


*Figure 3* Volcano plot showing the log2 fold change of gene expression between low *Pdx1* and high *Pdx1* cells against -log10(adjusted p-value), where red colouration represents statistical significance (adjusted p-value < 0.05).

The heatmaps show how the read counts for each gene vary from low to high *Pdx1* for the top 20 significant genes with the lowest adjusted p-value shows (figure 4a) and the top 20 significant genes with the largest absolute log2 fold change (figure 4b). *Pdx1* is the most significantly different gene when comparing expression levels between low and high *Pdx1* cells. Out of all the statistically significantly different genes identified from DESeq, there are only 4 genes, *Gcg, Ttr, Sgce* and *Irx2*, which are highly expressed in low *Pdx1* beta cells but have lower levels of expression in high *Pdx1* cells (figure 4a). *Ins2* expression was not found to be significantly different between low and high *Pdx1* cells with an adjusted p-value of 0.4689.

Out of the 1628 statistically significant genes, *Pdx1* ranked 42 in this list with a log2 fold increase of 2.15 in high *Pdx1* beta cells. In figure 4b, higher count values can be clearly seen in high *Pdx1* cells (on the right side of the heatmap) compared to low *Pdx1* cells (on the left side of the heatmap). The gene with the largest log2 fold change was *Styk1*, with a log2 fold increase of 4.12 in high *Pdx1* cells.

*a)*



*b)*

*Figure 4* Heatmap plots showing differences in gene expression between low *Pdx1* and high *Pdx1* cells, where a darker colour represents higher expression and grey represents 0 expression. *Pdx1* expression increases from left to right. Labels used are corresponding MGI symbol gene IDs. Created using R package pheatmap (v1.0.12)

1. Top 20 significant genes with the lowest adjusted p-value
2. Top 20 significant genes (adjusted p-value <0.05) with the largest absolute log2 fold change value

# Pathway analysis

Upregulated genes in high *Pdx1* beta cells

Using DAVID gene functional classification, 21 clusters were found in significantly upregulated genes in high *Pdx1* cells using the highest stringency parameter. 1124 upregulated genes were not able to be linked to a pathway. The top 5 clusters were found to be RNA splicing genes, subunits of RAB proteins, ribonucleoproteins, proteasome subunits for the cleavage of peptide bonds and mitochondrial proton-transporting ATP synthase complexes*,* with enrichment scores of 21.13, 18.76, 18.66, 18.19 and 17.17 respectively (figure 5a).

The top 5 out of a total of 187 clusters for DAVID functional annotation of biological processes using the highest stringency were reported to be RNA splicing, biosynthetic process genes, regulation of peptide hormone and secretion, genes involved in metabolic processes and regulation of RNA and transcription (figure 6a). These had respective enrichment values of 7.98, 7.33, 5.99, 5.66 and 5.37. When using the low stringency setting, this introduced other pathways in the clusters in the genes found, interestingly including 190 genes involved in cellular responses to stress.

Downregulated genes in high *Pdx1* beta cells

Using the highest or high stringency settings for DAVID gene functional classification, no clusters could be found on the significantly downregulated genes in high *Pdx1* cells. Setting the stringency parameter to medium enabled 3 clusters to be identified. These clusters were composed of genes with a variety of functions, and had enrichment scores of 2.50, 1.82 and

0.48 (figure 5b). 111 downregulated genes couldn’t be associated with any pathways.

Functional annotation using the highest stringency found 90 clusters, the top 5 clusters of which were associated with regulation of protein kinase activity, synaptic signalling, positive regulation of protein kinase activity, phosphorylation and junction assembly. The enrichment scores were 3.96, 3.56, 3.36, 3.24 and 3.10 (figure 6b).

Gene functional classification analyses failed to show what similarities were used to cluster the groups of genes identified functionally alike. Therefore, we kept labels for these groups as default output options to avoid mislabelling. The highest gene stringency was used where possible to maintain consistency and increase confidence in our results.

*a)* **Functional classification on upregulated genes** *b)*

**Functional classification on downregulated genes**

Gene Group 5

Gene Group 3

Gene Group 4

Gene Group 3 Gene Group 2

Gene Group 2

Gene Group 1

Gene Group 1

Enrichment score Enrichment score

0

5

10

15

20

25

0

1

2

3

4

*Figure 5* Enrichment of the gene groups identified to be similar in function using DAVID gene functional classification analysis

1. Top 5 groups of genes identified using highest stringency on significantly upregulated genes in high *Pdx1*

cells

1. The only 3 gene groups that could be labelled with medium stringency in functional using medium stringency on significantly downregulated genes in high *Pdx1* cells. No genes could be identified as similar using higher stringency levels.
   1. **Functional annotation on upregulated genes**
   2. **Functional annotation on downregulated genes**

Positive regulation of RNA biosynthetic process

Nicotinamide nucleotide metabolic process

Focal adhesion

assembly

Positive regulation of protein phosphorylation

Regulation of peptide hormone secretion

Positive regulation of protein kinase activity

Nucleoside biosynthetic process

Synaptic signaling

Splicing

Regulation of protein

kinase activity

Enrichment score Enrichment score

0

2

4

6

8

10

0

1

2

3

4

5

6

*Figure 6* Enrichment of the top 5 clusters of genes identified to form part of the same part of a functional pathway using DAVID functional annotation analysis. Labels used are the first GO term for that cluster, which are the most statistically significant

1. Significantly upregulated genes in high *Pdx1* cells using highest stringency
2. Significantly downregulated genes in high *Pdx1* cells using highest stringency

**Discussion**

Beta cells are known to be heterogeneous in the genes they express, but the basis on separating the cells varies from study to study. We approached this by separating beta cells based on their level of *Pdx1* expression. No outliers were detected in PCA, which suggests there was no contamination in the experimental design and minimal confounding variables such as batch effects. This validates our downstream analysis. As the values of the top two principle components were fairly low, this implies that there is not one clear factor driving separation between the samples. Therefore, attempts to explain the variance of the data in general proves to be difficult. Future research could investigate other factors in addition to *Pdx1* which may impact gene expression of beta cells.

*Pdx1* is the most significantly different expressed gene and shows a gradient in the heatmap (figure 4a), confirming variability of expression across beta cells. Several genes are highly expressed in either low or high *Pdx1* cells, further indicating that there is a relationship between *Pdx1* level and gene expression. The majority of genes have higher levels of expression in high *Pdx1* cells (figure 4b), indicating that these mature beta cells are more active and carry out a greater range of functions.

*Gcg, Ttr, Sgce* and *Irx2* are the only genes which are shown to be highly expressed in low *Pdx1* cells but the expression of which decreases in high *Pdx1* cells. *Sgce* encodes a type of sarcoglycan, which are involved in glucose homeostasis. Silencing its expression has been found to damage insulin-stimulated glucose uptake (36). Therefore, it is surprising that mature beta cells, which secrete insulin upon glucose exposure, have lower expression levels of this gene compared to immature beta cells. As this important gene is expressed in low *Pdx1* cells, this provides support that heterogeneity of beta cells can further improve our understanding of diabetes.

Particularly highly expressed in low *Pdx1* cells is *Gcg*, which is translated to produce glucagon, a hormone secreted by alpha cells to inhibit the action of insulin and increase levels of glucose during hypoglycaemia (37). *Gcg* and *Irx2* are genes which have been previously discovered using scRNA-seq to be specifically expressed in alpha cells, so we would not expect them to be expressed in beta cells at all (15). There are several possible explanations for this. Firstly, despite our efforts to select only for beta cells by implementing a threshold of insulin expression, contamination of alpha cells may have occurred, and these were mistakenly classified as low *Pdx1*-expressing beta cells. Alternatively, it could be that there is a subset of immature islet cells which have differentiated enough to secrete insulin, but still retain some of the alpha-specific genes.

This is not surprising as several literatures have noticed that beta and alpha cells are very similar during differentiation (38). The point of where beta cells and alpha have different identities and no longer share transcriptomics respective to their identities remains unknown. Our findings indicate that perhaps not all beta cells are independent to alpha cells after differentiation and this only occurs in a subset of beta cells. Another finding from our results could be that these subset of lowly expressing *Pdx1* beta cells are yet to differentiate and are still producing insulin.

Previous studies have found loss of *Pdx1* in beta cells resulted in characterising alpha cell marker genes glucagon and MafB (39). Interestingly, forcing *Pax4* – a transcription factor for insulin production in beta cells – expression in alpha cells causing these cells to convert to beta-like-cells. However, the behaviour of *Pdx1* was not investigated in the study (40). We speculate from previous studies and our findings that beta cells expressing low *Pdx1* retains the identities of alpha cells and perhaps can differentiate into alpha. Further works are

needed to verify this hypothesis and investigate what level of *Pdx1* expression correlates with alpha-specific genes to no longer be expressed.

As hub cells and immature cells expressed low levels of *Pdx1*, we suspected that both cell types were similar or carried the same function. However, other than expressing similar levels of *Pdx1*, we did not find immature beta cells to have increased proton pumping, ATP synthase activity and ATP generation (21). This suggests that hub cells and immature beta cells are two distinctively different cell types and carry functions independent to themselves.

Despite being the most statistically significant, in terms of log2 fold change, immature and mature beta cells vary by many more genes than they do by their *Pdx1* expression levels, as the log2 fold change for *Pdx1* was ranked 42nd. *Styk1* was the gene with the largest log2 fold change value of 4.12. This gene has been shown to be upregulated in patients with type 2 diabetes (41). As *Styk1* and *Pdx1* levels are correlated, it would be interesting to see how the levels of these two genes play a role in the development of diabetes.

*Pdx1* was the most statistically significant gene in terms of difference of expression levels, however *Ins2* expression was not statistically significant. This is surprising as *Pdx1* is a transcription factor of insulin, so we would expect them to be correlated (figure 1). A possible reason for this could be that the threshold value we used to select for *Ins2* positive cells was too high, thus meaning we are comparing beta cells with fairly similar *Ins2* expression levels. Furthermore, this data is just from one study, so it could possibly be due to a study design issue. As scRNA-seq only obtains a snapshot of the genes expressed in a cell at any one time, by chance or experimental conditions there could have been low level of insulin production across all beta cells. Alternatively, this could be due to using the DESeq tool for our analysis. This could be investigated further by using this data in addition to other biological samples with different threshold values and differential expression analysis tools.

From functional classification and functional annotation analyses using DAVID, the functions of upregulated genes in high *Pdx1* cells are involved in key metabolic processes required for normal beta cell function. As gene functional classification analysis found the top gene cluster to be involved in RNA splicing, this indicates that the mature beta cells may have intensified transcription activity (figure 5a).Another upregulated gene cluster identified in high *Pdx1* cells were members of RAB GTPases. This family of proteins has been found to regulate insulin secretion in beta cells, so this further confirms that mature beta cells are expressing higher levels of insulin (42).

In contrast, functional annotation and classification found low *Pdx1* cells to have secretory processes downregulated. The genes which were upregulated in these cells involved signalling processes, suggesting a role of communication in immature beta cells. This potentially explains why forcing overexpression of *Pdx1*, and therefore disrupting maturity and function of beta cells, is detrimental to the whole islet’s function (22).

Gene functional classification analysis using DAVID on the upregulated genes in high *Pdx1* cells resulted in much higher enrichment scores than for the downregulated genes, with a maximum value of 21.13 compared to 2.50 (figure 5). Even when taking into account the different stringency levels used, this is to be expected, as there were many more upregulated genes (1479 compared to only 149 downregulated genes). This also explains why many more clusters in total were found in upregulated genes, both using the functional classification and functional annotation tools, than reported for downregulated genes. For upregulated genes, enrichment scores were much higher using functional classification than functional annotation, with respective maximum scores of 21.13 and 7.98. This suggests that genes which have increased levels of expression in mature beta cells are involved in similar

functions, however they are not all present within the same pathway. The enrichment scores for functional annotation in general are lower than one may have predicted, however this is due to the highest stringency parameter being used, which increases the certainty of our findings (figure 6). When using the lowest stringency parameter for functional classification, this allowed identification of upregulated genes in high *Pdx1* cells which are associated with stress. This is in line with previous work (43), however as these genes only appear at lower stringency levels, we can’t be as certain about them than we can for clusters which are present at a higher stringency.

By using scRNA-seq, we were able to compare across beta cells with different *Pdx1* levels. Had we used bulk RNA-seq, we would have the issue of not seeing trends across genes from low to high pdx1 levels. However, an issue with scRNA-seq is low read count depth. A higher sequencing depth study would allow us to perform pre-ranked GSEA, which considers the importance of the differentially expressed genes for a specific function. We could compare genes identified to be differentially expressed to the gene ontology for regulation of insulation secretion in mice. This would require a rank value to be calculated for all differentially expressed genes, which is calculated by:

±𝑙𝑜𝑔2𝑓𝑜𝑙𝑑 𝑐ℎ𝑎𝑛𝑔𝑒 × (−𝑙𝑜𝑔10(𝑎𝑑𝑗𝑢𝑠𝑡𝑒𝑑 𝑝𝑣𝑎𝑙𝑢𝑒))

Increasing read depth could be done in future work by combining results at the individual cell level obtained by multiple studies.

Furthermore, concerns over the use of DESeq for scRNA-seq data have been raised as it doesn’t cope well with lots of zero values, which result from low read depth (44). However, the number of tools available for differential expression analysis for scRNA-seq data is limited. As more software specifically for the optimisation of single cell data is developed and released, these could be used to validate our findings.

Following on from this study, more research is needed to understand the importance of heterogeneity in the islet using scRNA-seq. Transcriptional changes in non-beta cells within the islet have also been linked to diabetes pathogenesis (16). Future studies could incorporate beta cell heterogeneity with the variation of other islet cells to improve our understanding of islet function and hence the disease even further. The distribution of beta cells varies according to disease state with loss of beta cells in the rear of the pancreas associated with type 2 diabetes (45), whereas greater loss at the head is linked to type 2 diabetes (46). Therefore, it would be interesting to find out using spatial proteomic and genomic technologies if the distribution of low *Pdx1*-expressing beta cells within the islet is related to the disease phenotype too. Other works are needed to verify whether immature beta cells are a rare subtype able to produce insulin in type 1 diabetes (47). Finally, to be of medical benefit to humans, this study should be replicated in beta cells from the human islet. Gene expression in human and mouse beta cells with and without diabetes has been found to be generally conserved, however there are some species-specific gene enrichments (15). Thus, care needs to be taken when extrapolating results from mice to humans.

Overall, our findings indicate that there may be specific functions or pathways for immature beta cells which are different to those mature beta cells are involved in, confirming our hypothesis. High *Pdx1* beta cells appear more functionally and metabolically active than immature beta cells. Genes highly expressed in immature beta cells are involved in signalling, suggesting their importance as cellular communicators within the islet.

Upregulated genes as well as significant genes were found to be markers of alpha cells, suggesting that immature beta cells are similar to alpha cells.

**References**

* 1. Diabetes. Available at:<https://www.who.int/news-room/fact-sheets/detail/diabetes>. Accessed Mar 19, 2019.
  2. Diabetes Prevalence 2018. Available at: [https://www.diabetes.org.uk/professionals/position-statements-reports/statistics/diabetes- prevalence-2018](https://www.diabetes.org.uk/professionals/position-statements-reports/statistics/diabetes-prevalence-2018). Accessed Mar 19, 2019.
  3. Kim SK, Hebrok M. Intercellular signals regulating pancreas development and function. Genes Dev 2001 Jan 15,;15(2):111-127.
  4. Da Silva Xavier G. The Cells of the Islets of Langerhans. Journal of clinical medicine 2018 Mar 12,;7(3):54.
  5. WU J, YANG X, CHEN B, XU X. Pancreas β cell regeneration and type 1 diabetes (Review). Experimental and Therapeutic Medicine 2015 Mar;9(3):653-657.
  6. Wu Y, Ding Y, Tanaka Y, Zhang W. Risk Factors Contributing to Type 2 Diabetes and Recent Advances in the Treatment and Prevention. International journal of medical sciences 2014;11(11):1185-1200.
  7. Clocquet AR, Egan JM, Stoffers DA, Muller DC, Wideman L, Chin GA, et al. Impaired insulin secretion and increased insulin sensitivity in familial maturity-onset diabetes of the young 4 (insulin promoter factor 1 gene). Diabetes 2000 Nov;49(11):1856-1864.
  8. Jonsson J, Carlsson L, Edlund T, Edlund H. Insulin-promoter-factor 1 is required for pancreas development in mice. Nature 1994 Oct 13,;371(6498):606-609.
  9. Fujimoto K, Polonsky KS. Pdx1 and other factors that regulate pancreatic β-cell survival. Diabetes, obesity & metabolism 2009 Nov;11 Suppl 4:30-37.
  10. Chang-Chen K, Mullur R, Bernal-Mizrachi E. β-cell failure as a complication of diabetes. Rev Endocr Metab Disord 2008 Dec;9(4):329-343.
  11. Johnson JD, Bernal-Mizrachi E, Alejandro EU, Han Z, Kalynyak TB, Li H, et al. Insulin protects islets from apoptosis via Pdx1 and specific changes in the human islet proteome. Proc Natl Acad Sci U S A 2006 Dec 19,;103(51):19575-19580.
  12. Gao T, McKenna B, Li C, Reichert M, Nguyen J, Singh T, et al. Pdx1 Maintains β Cell Identity and Function by Repressing an α Cell Program. Cell Metabolism 2014 Feb 4,;19(2):259-271.
  13. Gannon M, Ables ET, Crawford L, Lowe D, Offield MF, Magnuson MA, et al. pdx-1 function is specifically required in embryonic beta cells to generate appropriate numbers of endocrine cell types and maintain glucose homeostasis. Dev Biol 2008 Feb 15,;314(2):406- 417.
  14. Benninger RKP, Hodson DJ. New Understanding of β-Cell Heterogeneity and In Situ Islet Function. Diabetes 2018 04;67(4):537-547.
  15. Xin Y, Kim J, Okamoto H, Ni M, Wei Y, Adler C, et al. RNA Sequencing of Single Human Islet Cells Reveals Type 2 Diabetes Genes. Cell Metab 2016 10 11,;24(4):608-615.
  16. Lawlor N, George J, Bolisetty M, Kursawe R, Sun L, Sivakamasundari V, et al. Single- cell transcriptomes identify human islet cell signatures and reveal cell-type-specific expression changes in type 2 diabetes. Genome Res 2017 02;27(2):208-222.
  17. Segerstolpe Å, Palasantza A, Eliasson P, Andersson E, Andréasson A, Sun X, et al. Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes. Cell Metab 2016 10 11,;24(4):593-607.
  18. Wang Y, Navin NE. Advances and applications of single-cell sequencing technologies. Mol Cell 2015 May 21,;58(4):598-609.
  19. Muraro MJ, Dharmadhikari G, Grün D, Groen N, Dielen T, Jansen E, et al. A Single-Cell Transcriptome Atlas of the Human Pancreas. Cell Syst 2016 10 26,;3(4):385-394.
  20. Szabat M, Luciani DS, Piret JM, Johnson JD. Maturation of adult beta-cells revealed using a Pdx1/insulin dual-reporter lentivirus. Endocrinology 2009 Apr;150(4):1627-1635.
  21. Johnston NR, Mitchell RK, Haythorne E, Pessoa MP, Semplici F, Ferrer J, et al. Beta Cell Hubs Dictate Pancreatic Islet Responses to Glucose. Cell Metab 2016 09 13,;24(3):389- 401.
  22. Nasteska, Daniela Rutter, Guy A., Zhou, Qiao Hodson, David J. Loss of beta-cell heterogeneity disrupts normal islet function.
  23. Qiu W, Zhang Y, Feng Y, Li L, Yang L, Xu C. Deciphering Pancreatic Islet β Cell and α Cell Maturation Pathways and Characteristic Features at the Single-Cell Level. Cell Metab 2017 May 02,;25(5):1194-1205.
  24. Andrews S. FastQC: a quality control tool for high throughput sequence data. 2010; Available at:<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>.
  25. Krueger Felix. Trim Galore. 2012; Available at:<http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/>.
  26. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 2013 Jan 01,;29(1):15-21.
  27. Bacher R, Chu L, Leng N, Gasch AP, Thomson JA, Stewart RM, et al. SCnorm: robust normalization of single-cell RNA-seq data. Nat Methods 2017 Jun;14(6):584-586.
  28. Da Silva Xavier G. The Cells of the Islets of Langerhans. J Clin Med 2018 -3-12;7(3).
  29. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, et al. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics 2005 Aug 15,;21(16):3439-3440.
  30. Huber W, Durinck S, Spellman PT, Birney E. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nature Protocols 2009 Jul;4(8):1184-1191.
  31. Kiselev VY, Kirschner K, Schaub MT, Andrews T, Yiu A, Chandra T, et al. SC3: consensus clustering of single-cell RNA-seq data. Nature Methods 2017;14(5):483-486.
  32. Scrucca L, Fop M, Murphy TB, Raftery AE. mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. R J 2016 Aug;8(1):289-317.
  33. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol 2010;11(10):106.
  34. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 2009;4(1):44-57.
  35. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 2009 Jan;37(1):1-13.
  36. Groh S, Zong H, Goddeeris MM, Lebakken CS, Venzke D, Pessin JE, et al. Sarcoglycan complex: implications for metabolic defects in muscular dystrophies. J Biol Chem 2009 Jul 17,;284(29):19178-19182.
  37. NCBI Gene. Available at:<https://www.ncbi.nlm.nih.gov/gene/>. Accessed 20/03/, 2019.
  38. Herrera PL. Adult insulin- and glucagon-producing cells differentiate from two independent cell lineages. Development 2000 Jun;127(11):2317-2322.
  39. Gao T, McKenna B, Li C, Reichert M, Nguyen J, Singh T, et al. Pdx1 Maintains β Cell Identity and Function by Repressing an α Cell Program. Cell Metabolism 2014 Feb 4,;19(2):259-271.
  40. Yang Y, Thorel F, Boyer DF, Herrera PL, Wright CVE. Context-specific α-to-β-cell reprogramming by forced Pdx1 expression. Genes Dev 2011 -8-15;25(16):1680-1685.
  41. Mao J, Ai J, Zhou X, Shenwu M, Ong M, Blue M, et al. Transcriptomic profiles of peripheral white blood cells in type II diabetes and racial differences in expression profiles. BMC Genomics 2011 Dec 23,;12 Suppl 5.
  42. Sugawara K, Shibasaki T, Mizoguchi A, Saito T, Seino S. Rab11 and its effector Rip11 participate in regulation of insulin granule exocytosis. Genes Cells 2009 Apr;14(4):445-456.
  43. Robertson RP. Chronic oxidative stress as a central mechanism for glucose toxicity in pancreatic islet beta cells in diabetes. J Biol Chem 2004 Oct 08,;279(41):42351-42354.
  44. Lun ATL, Bach K, Marioni JC. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. Genome Biol 2016 Apr 27,;17:75.
  45. Poudel A, Savari O, Striegel DA, Periwal V, Taxy J, Millis JM, et al. Beta-cell destruction and preservation in childhood and adult onset type 1 diabetes. Endocrine 2015 Aug;49(3):693-702.
  46. Wang X, Misawa R, Zielinski MC, Cowen P, Jo J, Periwal V, et al. Regional differences in islet distribution in the human pancreas--preferential beta-cell loss in the head region in patients with type 2 diabetes. PLoS ONE 2013;8(6):67454.
  47. Oram RA, Jones AG, Besser REJ, Knight BA, Shields BM, Brown RJ, et al. The majority of patients with long-duration type 1 diabetes are insulin microsecretors and have functioning beta cells. Diabetologia 2014 Jan;57(1):187-191.