Title

iPSC-derived pancreatic progenitors are an optimal model system to study T2D regulatory variants active during fetal development of the pancreas

Authors

Agnieszka D'Antonio-Chronowska[1,†], Margaret K.R. Donovan[2,3,†], Kyohei Fujita[1], Bianca M. Salgado[4], Hiroko Matsui[4], Timothy D. Arthur[3,5], Jennifer P. Nguyen[2,3], Matteo D'Antonio[1], Kelly A. Frazer[1,4,*]

Affiliations:

[1] Department of Pediatrics, University of California San Diego, La Jolla, CA, 92093, USA

[2] Bioinformatics and Systems Biology Graduate Program, University of California, San Diego, La Jolla, CA, 92093, USA

[3] Department of Biomedical Informatics, University of California, San Diego, La Jolla, CA, 92093, USA

[4] Institute of Genomic Medicine, University of California San Diego, 9500 Gilman Dr, La Jolla, CA, 92093, USA

[5] Biomedical Sciences Graduate Program, University of California, San Diego, La Jolla, CA 92093, USA;

* Correspondence to: Kelly A. Frazer; Tel: +1 (858) 246-0208; Email: kafrazer@health.ucsd.edu.

† Equal contributions

Abstract

Pancreatic progenitor cells (PPCs), which are multipotent cells that have the potential to give rise to endocrine, exocrine and epithelial cells, provide a powerful model system to examine the molecular characteristics of differentiating, fetal-like pancreas cells and to study the genetics of pancreatic disease. We differentiated ten induced pluripotent stem cell-derived pancreatic progenitor cell (iPSC-PPC) lines and, using single cell RNA-seq and single nuclear ATAC-seq, we found that they show varying levels of cellular maturity, including the presence of early PPC, endocrine and exocrine cells. We evaluated iPSC-PPC as a model system to study type 2 diabetes (T2D) and found that, of 380 T2D risk loci, 208 overlapped regulatory elements active in iPSC-PPC, including 55 fetal-specific. These loci are associated with transcription factor binding site alterations and allele-specific expression, suggesting that iPSC-PPC are an optimal model system to annotate GWAS variants that are not functional in the adult pancreas.

Introduction

The developing fetal pancreas consists of pancreatic progenitor cells (PPCs), which are a multipotent cell type that has the potential to give rise to endocrine cells (clusters of hormone secreting cells, such α, β, γ, and δ cells), exocrine cells, and epithelial cells (i.e. ductal cells) (Cano et al., 2014; Jennings et al., 2015). Recently it has become possible to use pluripotent stem cells to derive PPCs (Pagliuca et al., 2014; Rezania et al., 2014), which provide a virtually unlimited source of cells to examine the molecular characteristics of differentiating, fetal-like pancreas cells. However, additional studies need to be conducted to determine the utility of these cells as a model system to study how regulatory variants influence the developing pancreas or adult pancreatic disease.

Single cell RNA-seq (scRNA-seq) and single nuclear ATAC-seq (snATAC-seq) have been generated for human adult, fetal mouse pancreas and embryonic stem cell-derived PPCs (ESC-PPC) (Baron et al., 2016; Chiou et al., 2019; Enge et al., 2017; Peng et al., 2019; Petersen et al., 2017; Rai et al., 2020; Segerstolpe et al., 2016; Sharon et al., 2019; Veres et al., 2019), which allow the interrogation of gene expression and chromatin accessibility in specific pancreas cell types. As a precursor to mature pancreas-lineage cell types, PPCs capture a snapshot during differentiation that is important for pancreas cell fate decisions. PPCs are characterized as a population of cells that have differentiated beyond the pancreatic foregut (i.e. Stage 3), marked by expression of *PDX1*, and committed to a pancreatic progenitor fate (Stage 4), marked by the co-expression of *PDX1* and *NKX6-1* (Cano et al., 2014; Jennings et al., 2015). However, a recent study characterized the *in vitro* differentiation of embryonic stem cells and showed that PPCs are heterogeneous and, at each analyzed differentiation stage, they are composed of multiple cell types, including the presence of α- and β-cells at early stages, endocrine and early exocrine cells at later stages (Veres et al., 2019). This well-characterized dataset provides a reference for future studies that aim at performing PPC differentiations. It is currently unknown whether using iPSCs, rather than embryonic stem cells, and different differentiation protocols result in cell heterogeneity comparable to this reference dataset. Furthermore, it is still unclear whether different iPSC samples have large differences in differentiation efficiency.

Type 2 Diabetes (T2D) is the most common pancreatic disease, which affects more than 400 million individuals worldwide (W.H.O., 2020). It is the result of the inability of β-cells to control blood glucose homeostasis and is influenced by genetics and environmental factors, including obesity (Kahn et al., 2014). Efforts to identify genetic variants associated with T2D through various genome wide association studies (GWAS) have identified more than 400 T2D signals, where most of these variants fall in non-coding, transcriptional regulatory regions (Mahajan et al., 2018). Interpreting these T2D GWAS is challenging due to ambiguity introduced by cell type-specific function (i.e. the variant functions primarily or solely in specific pancreas cell types) and stage of human development-specific function (i.e. the variant functions primarily or solely during fetal development or during adulthood). While several studies have performed fine-mapping of T2D signals to identify the likely causal variant responsible for T2D (Chiou et al., 2019; Mahajan et al., 2018), the vast majority still remain uncharacterized. Many of these variants may function in specific cell types and can be characterized using scRNA-seq or snATAC-seq on adult pancreas samples. A portion of T2D risk variants may have pancreas developmental-specific functions, as it was shown that environmental factors during fetal development, such as a mother's obesity or malnourishment, affect the correct development of β-cells, resulting in increased risk for T2D (Guenard et al., 2013; Lumey et al., 2015; Nielsen et al., 2014). Thus, to better resolve the genetic basis of T2D, efforts should be made to study cell type-specific variant function during fetal development of the pancreas.

Studying functional genetic variation in fetal pancreas cell types is challenging, because human fetal pancreas specimens are not readily available. Furthermore, only a limited number of human embryonic stem cell lines is available and cannot be used to perform genetic association studies. To address these issues, we used nine induced pluripotent stem cell (iPSC) clones from unrelated individuals (Panopoulos et al., 2017) and a 15-day differentiation protocol to obtain ten induced pluripotent stem cell-derived pancreatic progenitor (iPSC-PPC) samples to use as a model system to characterize molecular phenotypes in human pancreatic cells at an early stage of development. With this pilot study, we found that iPSC-PPC differentiate asynchronously, similar to ESC-PPC derived with a different protocol, and can be used to study the function of T2D risk variants that are not active in the adult pancreas.

Results

Overview of study design

We used iPSC lines from nine unrelated individuals (Panopoulos et al., 2017) (Table S1) and a 15-day differentiation protocol to derive ten iPSC-derived pancreatic progenitor (iPSC-PPC) samples (one iPSC clone was differentiated twice; Figure 1A). As cellular heterogeneity is a central challenge in directed differentiations (D'Antonio-Chronowska et al., 2019), we sought to identify the cell types and their relative proportions comprising the ten iPSC-PPC samples by performing scRNA-seq (Figure 1B). To characterize the cell types represented in the scRNA-seq, we integrated these data with a well-characterized reference set of scRNA-seq data from embryonic stem cell-derived ß cells, which were collected at four stages of differentiation using a 25-day differentiation protocol (hereto referred to as SC-ß cell study; Figures 1A,B) (Pagliuca et al., 2014; Veres et al., 2019). We determined that iPSC-PPC samples are comprised of cells at varying levels of maturity (including early endocrine and early exocrine cells) suggesting that the differentiation was asynchronous (i.e., cells are at different differentiation stages) (Figure 1B). To assess whether iPSC-PPCs can be used as a model system to study regulatory T2D risk variants active during fetal development, we generated snATAC-seq from 24,553 nuclei from 7 iPSC-PPC samples and examined how the T2D signals overlapped the iPSC-PPC fetal peaks compared to peaks obtained from adult pancreas (Mahajan et al., 2018) (Figure 1C). We identified 55 T2D risk variants within fetal-specific regulatory elements, many of which are associated with allele-specific expression (ASE) or alter transcription factor motifs.

Varying levels of cellular maturity explain heterogeneity across iPSC-PPC samples

To perform a baseline assessment of iPSC-PPC differentiation efficiency, we measured the fraction of cells positive for two hallmark PPC transcription factors, PDX1 and NKX6-1, using flow cytometry, and observed that the ten samples differentiated with varying efficiency (Figure S1A,B). The fraction of cells that stained positive for PDX1 ranged from 22.1 to 96.4%, and the fraction that was double-positive stained for PDX1 and NKX6-1 ranged from 9.4 to 91.7% (Figure S1B). As expression of NKX6-1 occurs later than PDX1 expression during development, the lower fraction of double-positive cells could reflect asynchronous differentiation across the ten iPSC-PPC samples resulting in pancreatic progenitor cells at different stages of maturity.

To better understand the cellular heterogeneity of the iPSC-PPCs, we performed scRNA-seq on the ten samples and one iPSC clone (total: 88,188 cells). Using a shared-nearest-neighbor (SNN) graph method, we detected nine distinct cell clusters (Figure 2A, Table S2). While cluster 2 was comprised almost entirely of cells from the iPSC clone (100% of the 8,173 iPSC cells), we found the relative proportion of cells in the other eight clusters varied across the ten iPSC-PPC samples (Figure 2B,C, Figure S2A). Marker genes for stem cells (*POU5F1*) were highly expressed in cluster 2, for endocrine cells (*CHGA* and *INS*) in cluster 7, and for epithelial and endothelial cells (*ESM1*) in cluster 8. Cluster 8 likely includes precursors for acinar and ductal cells, which appear only at later pancreatic development stages (Pictet et al., 1972; Reichert and Rustgi, 2011). Notably, the remaining clusters all expressed *PDX1* and *NKX6-1* at varying levels, with cluster 0 showing the highest co-expression of these two transcription factors (Figure S2B). Given this observation, we asked if the fraction of cells mapping to cluster 0 from each iPSC-PPC sample related to differentiation efficiency, as measured by percent of cells stained double-positive for PDX1 and NKX6-1, and found that the two measurements were highly correlated (r = 0.85, p = 0.0035; Pearson's correlation) (Figure 2D). Overall, these results suggest that the 15-day iPSC-PPC differentiation protocol results in samples consisting of pancreatic cell types at varying levels of maturity and that this accounts for a large fraction of the cellular heterogeneity (i.e. variable fraction of double-positive cells for PDX1 and NKX6-1) across the iPSC-PPC samples.

Cellular composition of human embryonic stem cell-derived pancreatic cells from a reference dataset

We examined if similar cell types in the SC-ß cell study across the four differentiation stages clustered together to understand their utility for characterizing the cell types present in the iPSC-PPC samples, which represent an earlier stage of development (Figure 1A). We integrated 24,321 SC-ß cells representing 11 annotated cell types collected at four distinct stages of differentiation and using a SNN graph method detected six distinct cell populations (Figure 2E). We examined the fraction of each of the 11 annotated SC-ß cell types that mapped to each of the six clusters and to each of the four differentiation stages (Figure 2F-I), as well as the fraction of cells collected at each stage that mapped to each of the six clusters (Figure 2J). We observed that clusters 1 and 5 were comprised of mostly *PDX1*$^+$ progenitor cells collected at stage 3; cluster 2 was comprised of both *NKX6-1*$^+$ progenitors collected at stage 4 and non-endocrine cells collected at stages 5 and 6; clusters 0 and 4 were comprised of a mixture of endocrine cell types collected at stages 4, 5, and 6; and cluster 3 was comprised of mostly replicating cells collected across all four stages (Figure 2E-J). These data show that cells with the same annotation and from the same differentiation stage map to multiple clusters, suggesting that asynchronous differentiation (i.e., cell types at varying levels of maturity) likely occurred in the SC-ß cell study similar to that observed in the iPSC-PPC cells (Figure S1B, S2B).

Annotation of iPSC-PPCs cell types using SC-ß cell study as a reference

To utilize the eleven annotated cell types in the SC-ß cell study to further characterize the cell types present in the iPSC-PPC samples, we integrated and clustered cells from both studies (112,509 total cells) and using a SNN graph method detected six distinct clusters (Figure 3A, S3). While the three largest clusters (0, 1 and 2; corresponding to 88.5% of cells) included cells from both studies (Figure 3B, C), the other three predominantly included only iPSC-PPC samples (11.55% of cells, Figure 3C). The largest cluster (cluster 0, 74,303 cells) included cells from all four developmental stages in the SC-ß study that expressed *PDX1* and *NKX6-1* suggesting that this cluster includes cells belonging to a continuum of varying pancreatic progenitor cell maturities (Figure 3D). Cluster 1 included replicating cells and cluster 2 was primarily comprised of endocrine cell types from the SC-ß cell study with only 794 iPSC-PPC cells, which we broadly annotated as endocrine. Cluster 3 was composed of iPSC cells, cluster 4 of early PPC cells and cluster 5 of epithelial cells that are likely precursors of acinar and ductal cells (termed "early exocrine cells" hereafter). We determined the relative proportions of the five annotated iPSC-PPC cell types (PPCs of varying maturity, replicating cells, endocrine cells, early iPSC-PPC, and early exocrine) in the ten samples (Figure 3E-F) and observed varying fractions as anticipated based on their distributions across the nine iPSC-PPC clusters (Figures 2A-D). This analysis shows that iPSC-PPC are composed of pancreatic progenitor cells at different developmental stages, including both early PPCs that express *PDX1* and *NKX6-1* at low levels and early endocrine and exocrine cells, indicating that *in vitro* PPC differentiation from pluripotent stem cells is asynchronous.

Epigenomic analysis of iPSC-PPCs supports asynchronous differentiation

The vast majority of T2D variants are non-coding and have regulatory functions (Mahajan et al., 2018); therefore, to test the utility of iPSC-PPCs to study adult pancreatic disease, we characterized the epigenome of seven iPSC-PPC samples by performing single nuclei ATAC-seq (snATAC-seq). We obtained 29,128 nuclei, which we divided into six distinct clusters (Figure 4A, Figure S4, Figure S5, Table S3). While most clusters included nuclei from all eight samples, clusters 1, 2 and 4 were predominantly composed of nuclei from PPC_134 and PPC_150 (94.6% of nuclei in cluster 1, 94.2% of nuclei in cluster 2 and 93.4% of nuclei in cluster 4, Figure 4B). Since these two samples had the lowest fraction of cells double-positive stained for PDX1 and NKX6-1 (Figure S1), we hypothesized that these clusters identify nuclei at different developmental stages. To test this hypothesis, we performed peak calling on all nuclei, which led us to detect 127,071 peaks, including 7,284 that were overexpressed in one or more clusters, and performed motif enrichment analysis on these cluster-specific peaks (Table S4). We found that cluster 0 was strongly enriched for PDX1, NKX6-1 and ONECUT1 (HNF6) motifs, whereas cluster 1 was associated with the earlier pancreatic development transcription factor SOX17 (definitive endoderm stage), and cluster 2 with transcription factors active later during endocrine differentiation (SOX4, NEUROD1 and RFX3, Figure 4C,D). Cluster 3 included nuclei at early developmental stages, as it was associated with transcription factors active during early embryonic development and in neural crest, including TFAP2A and MEIS2 (Table S5). The two samples with the largest number of nuclei mapping to cluster 3 (PPC_034 and PPC_051, 485 and 566 nuclei, respectively) also had the largest fraction of early iPSC-PPC cells in the scRNA-seq analysis (Figure 3E-F). Cluster 4 was enriched for both early development markers (TFAP2A and MEIS2), epithelial development-associated transcription factors (SOX4 and RFX3), exocrine and ductal transcription factors (PTF1A, GATA4 and SOX9), suggesting that, although the small number of nuclei makes it difficult to functionally characterize this cluster, it likely comprises cells primed to differentiate to acinar and ductal cells. Finally, cluster 5 was enriched for both endocrine (NEUROD1) and exocrine differentiation (PTF1A) markers. Overall, the cell types identified using snATAC-seq are consistent with the results from flow cytometry and scRNA-seq analyses. Furthermore, snATAC-seq shows that there are 7,284 cell type-specific peaks, which may overlap known disease-causing variants in the adult pancreas.

T2D risk variants in fetal-specific regulatory elements

To test if gene regulation in iPSC-PPC provides an advantage in understanding regulatory variants underlying T2D compared to studies performed in adult pancreas, we investigated the associations of 380 signals in fine-mapped T2D loci (Mahajan et al., 2018) with snATAC-seq peaks. For each of these T2D loci, we tested the overlap between each variant with posterior probability of association (PPA) > 0 and 127,071 fetal-like peaks, as well as 103,856 snATAC-seq peaks and 106,460 bulk ATAC-seq peaks from adult pancreatic islets (Rai et al., 2020) (Figure 5A). We found that in 291 loci (76.6%) at least one variant overlapped one ATAC-seq peak, including 55 fetal-specific (18.9%), 83 adult-specific (28.5%) and 153 active in both fetal-like pancreas and adult pancreatic islets (52.6%), indicating that the majority of T2D variants (71.4%) are functional in fetal-like pancreatic cells, including a substantial fraction that is not active in the adult. These results suggest that iPSC-PPCs provide unique insights into study T2D variants that are not functional in the adult pancreas.

T2D variants are associated with transcription factor binding motifs

To further characterize the T2D risk variants active in fetal-like pancreas, we investigated their effects on transcription factor binding. We obtained the genomic sequences around each of the 208 variants that overlap iPSC-PPC ATAC-seq peaks (55 iPSC-PPC-specific and 153 active in both fetal-like and adult pancreas) and tested the changes in transcription factor binding activity between effect and non-effect alleles using FIMO (Grant et al., 2011). We found that 83 T2D risk variants (39.9%) significantly overlapped a transcription factor motif (62 motifs), including 73 (88.0%, corresponding to 55 motifs) that had only one allele significantly associated with one or more transcription factor motif(s) (Figure 5B, C, Table S6). Among the most frequently altered motifs were SP3, WT1, KLF15, IRF3 and KLF6, all of which are known to be associated with response to insulin and/or diabetes (Jung et al., 2013; Kumari et al., 2016; Rubinstein et al., 2004; Webster et al., 1997; White et al., 2006). Overall, these results show that T2D risk variants that occur in regulatory elements that are active in fetal-like pancreatic cells are associated with alterations in the binding affinity of transcription factors involved in diabetes, confirming that iPSC-PPCs are an important model system for studying the effects of genetic variation on diabetes.

T2D variants associated with allele-specific effects in fetal regulatory elements

To confirm the power of iPSC-PPCs as a model system to study the effects of genetic variation on T2D, we investigated the allele specific effects (ASE) of the 208 variants that overlap fetal-like ATAC-seq peaks. We obtained the genotype of these variants in the seven iPSCORE individuals with snATAC-seq and selected only the variants with high-depth read depth (≥35) in heterozygous individuals. Of 77 variants heterozygous in at least one individual and with high read depth, we found six (*ABCC5*, *BPTF*, *ITPR2*, *KCNQ1*, *RBMS1* and *SLC9B1* loci) with significant ASE (FDR-adjusted p-value < 0.1, binomial test, Benjamini-Hochberg method, Figure 5A, Figure 6A, Table S7). Two of these variants (in the *BPTF* and *SLC9B1* loci) were in fetal-specific ATAC-seq peaks, whereas the other four overlapped ATAC-seq peaks active in both fetal-like and adult pancreatic cells. The variant with the strongest ASE (rs61676547, p-value = 1.9×10^{-7}, binomial test, Figure 6B) was located in an intron of *BPTF* and overlapped a fetal-specific ATAC-seq peak. Interestingly, in addition to being a T2D risk variant, it is associated with HDL cholesterol (Klarin et al., 2018), and is not an eQTL for *BPTF* in pancreas or adult islets, although it is an eQTL in cerebellum, cerebellar hemisphere, tibial nerve and aorta (Consortium, 2020; Vinuela et al., 2020), suggesting that its association with *BPTF* function is likely fetal-specific in pancreas. Conversely, the ASE variant associated with the *KCNQ1* locus (rs2237897, p = 2.9 x 10^{-3}, binomial test, Figure 6C) has been associated with increased risk for T2D and diabetic nephropathy in ethnically diverse populations, including admixed Americans and East Asians (Consortium et al., 2014; Mussig et al., 2009; Ohshige et al., 2010; Suzuki et al., 2019; Unoki et al., 2008), as well as increased glucose levels (Palmer et al., 2015) and body max index (Akiyama et al., 2017). The other four ASE variants (Figure S6) are all eQTLs in multiple tissues (Consortium, 2020), but they have not previously been associated with GWAS traits. Interestingly, two of these variants (rs7132434 and rs223490) are eQTLs for multiple genes (*BHLHE41*, *RP11-283G6.3* and *SSPN* for rs7132434; *BDH2*, *CENPE*, *CISD2*, *KRT8P46*, *LRRC37A15P*, *MANBA*, *PABPC1P7*, *RP11-10L12.1*, *RP11-10L12.2*, *RP11-10L12.4* and *UBE2D3*, including *BDH2*, *KRT8P46*, *LRRC37A15P* and *UBE2D3* in pancreas, for rs223490), but not for the genes in their associated T2D risk loci (*ITPR2* and *SLC9B1*, respectively). These results show that, for many loci, iPSC-PPCs are the optimal model system to study the effects of genetic variation on T2D.

Discussion

In this study, we aimed at understanding the utility of iPSC-PPC as a model system to investigate the associations between genetic variation and T2D. We differentiated nine iPSC clones from the iPSCORE cohort (D'Antonio-Chronowska et al., 2019; D'Antonio et al., 2018; DeBoever et al., 2017; Panopoulos et al., 2017) to iPSC-PPC and characterized their transcriptomes and epigenome at single cell resolution. By comparing the iPSC-PPC transcriptomes to a reference set of highly annotated cell types in the SC-ß cell study (Mahajan et al., 2018), we showed that iPSC-PPC differentiate asynchronously, with the majority of cells belonging to a continuum of varying pancreatic progenitor states of maturity, and with approximately 4.7% of the cells being more immature iPSC-PPC, as indicated by the low expression of *PDX1* and *NKX6-1* and 1.5% of the cells being at a more advanced stage (early endocrine and early exocrine). Our study shows that iPSC-PPC have transcriptomic features similar to ESC-PPC even if they are differentiated with a different protocol and are a suitable model system to study the human fetal pancreas.

While GWAS studies involving hundreds of thousands of individuals have identified 380 genomic loci associated with T2D (Mahajan et al., 2018), it is still complicated to determine the causal variant in each locus and to characterize the molecular mechanisms underlying the associations between each variant and the disease. The vast majority of T2D risk variants map to non-coding sequences, indicating that they likely affect the function of regulatory elements, rather than directly influencing transcription (Gaulton et al., 2015; Geusz et al., 2020; Greenwald et al., 2019; Pasquali et al., 2014; Varshney et al., 2017). Many T2D risk variants overlap regulatory elements that are functional in pancreatic endocrine cells and their function and disease association have been thoroughly characterized (Chiou et al., 2019; Mahajan et al., 2018; Spracklen et al., 2020; Vujkovic et al., 2020; Xue et al., 2018). Variants without evidence of overlapping active regulatory elements in adult pancreatic islets may be functional only during pancreatic embryonic development. Supporting this hypothesis, it was recently shown that the T2D risk variants associated with *LAMA1* and *CRB2* overlap fetal-specific enhancers, suggesting that they affect pancreatic endocrine cells development, which may predispose to T2D later in life (Geusz et al., 2020). Here, we detected 208 T2D loci that overlap pancreatic regulatory elements that are active in iPSC-PPC, including 55 that are active only in iPSC-PPC and not in the adult pancreas. Therefore, iPSC-PPC are an optimal model system to study these 55 fetal-specific T2D risk variants. We further investigated the fetal-specific *BPTF* locus and found that the variant with strongest posterior probability of being causal that overlaps a fetal-specific regulatory element (rs61676547) is not an eQTL in adult pancreas or pancreatic islets. Although our study did not have enough power for a QTL analysis, we found that rs61676547 had a strong allele-specific expression in iPSC-PPC, suggesting that it is likely functional only in the fetal pancreas.

The fetal origin of adult disease hypothesis suggests that *In Utero* exposure to certain risk factors affect embryonic development and results in increased risk of specific disease in the adult (Barker, 1995). Indeed, it was shown that a mother's malnourishment or obesity during pregnancy is associated with increased T2D risk in her progeny (Lumey et al., 2015), which likely results from alterations in beta cell mass at birth (Nielsen et al., 2014). By showing that 55 T2D risk loci are fetal-specific, our results suggests that the fetal origin of adult disease hypothesis needs to be further investigated and indicate that iPSC-PPC are an optimal model system to annotate GWAS variants that are not functional in the adult pancreas. Ultimately, by expanding iPSC-PPC samples to hundreds of genetic backgrounds, iPSC-PPC will become an important resource to characterize genetic associations that are present only in the fetal pancreas and may affect an individual's predisposition to T2D, as well as other pancreatic diseases.

Methods

iPSCORE subject information

We obtained 9 iPSC lines from the iPSCORE collection (Panopoulos et al., 2017) (Table S1). These lines were reprogrammed from skin fibroblasts collected from 9 unrelated subjects (8 female, 1 male), who ranged in age at time of donation from 21 to 65 years old, and represent three 1000 Genomes Project super populations: European American (7), Asian American (1), and African American (1). From each subject, whole blood samples were collected and used to generate and process whole genome sequence (WGS) data as previously described (D'Antonio et al., 2018; DeBoever et al., 2017). Briefly, reads were aligned against human genome b37 with decoy sequences (Genomes Project et al., 2015) using BWA-mem and default parameters (Li and Durbin, 2009). We applied the GATK best-practices pipeline for variant calling that includes indel-realignment, base-recalibration, genotyping using HaplotypeCaller, and finally joint genotyping using GenotypeGVCFs (DePristo et al., 2011; McKenna et al., 2010; Van der Auwera et al., 2013). The recruitment of these individuals was approved by the Institutional Review Boards of the University of California, San Diego and The Salk Institute (Project no. 110776ZF).

iPSC-PPC Derivation

iPSC-PPCs were derived using STEMdiff™ Pancreatic Progenitor Kit (StemCell Technologies) following manufacture's recommendations except as noted below. One iPSC line (from subject 90e8222f-2a97-4a3c-9517-fbd7626122fd) was independently differentiated twice (PPC_029 and PPC_036) resulting in a total of 10 derived iPSC-PPC samples.

Expansion of iPSC: One vial from each of 9 iPSC lines was thawed into mTeSR1 medium containing 10 µM ROCK Inhibitor (Selleckchem) and plated on one well of a 6-well plate coated with matrigel. iPSCs were grown until they reached 80% confluency, then were passaged using 2mg/ml solution of Dispase II (ThermoFisher Scientific). To obtain a sufficient number of iPSCs for differentiation, iPSCs were passaged twice: 1) cells from the first passage were plated on three wells of a 6-well plate (ratio 1:3); and 2) cells from the second passage were plated on six wells of a 6-well plate (ratio 1:2).

Monolayer plating (Day 0; D0): When the confluency of iPSC cells in the six wells of a 6-well plate reached 80%, cells were dissociated into single cells using Accutase (Innovative Cell Technologies Inc.). Single iPSC cells were resuspended at the concentration of 1.85 x10^6 cells/ml in mTeSR containing 10μM ROCK inhibitor and plated on six wells of a 6-well. Cells were grown for approximately 16 to 20 hours to achieve a uniform 90-95% confluency (3.7x10^6 cells/well; about 3.9x10^5 cells/cm^2).

Differentiation: Differentiation of the confluent iPSC monolayers were initiated by addition of the STEMDiff Stage Endoderm Basal medium supplemented with Supplement MR and Supplement CJ (2ml/well) (D1). All following media changes were performed every 24 hours following initiation of differentiation (2ml/well). On D2 and D3, the medium was changed to fresh STEMDiff Stage Endoderm Basal medium supplemented with Supplement CJ. On D4, the medium was changed to STEMDiff Pancreatic Stage 2-4 Basal medium supplemented with Supplement 2A and Supplement 2B. On D5 and D6, the medium was changed to STEMDiff Pancreatic Stage 2-4 Basal medium supplemented with Supplement 2B. On D7, D8 and D9, the medium was changed to STEMDiff Pancreatic Stage 2-4 Basal medium supplemented with Supplement 3. On D10, D11, D12, D13 and D14, the medium was changed to STEMDiff Pancreatic Stage 2-4 Basal medium supplemented with Supplement 4.

Harvest: On D15 cells were dissociated using Accutase, collected and counted, and either processed fresh (scRNA-seq) or cryopreserved (scRNA-seq and snATAC-seq).

Flow Cytometry

Each of the 10 iPSC-PPC differentiations were analyzed for co-expression of two pancreatic precursor markers, PDX1 and NKX6-1, using flow cytometry. Specifically, at least 2x10^6 iPSC-PPC cells were fixed and permeabilized using the Fixation/Permeabilization Solution Kit with BD GolgiStopTM (BD Biosciences) following manufacturer recommendations. After the last centrifugation, cells were resuspended in 1X BD Perm/WashTM Buffer at the concentration of 1 x 10^7/ml. For each flow cytometry staining, 2.5 x 10^5 cells were stained with PE Mouse anti-PDX1 Clone-658A5 (BD Biosciences; 1:10) and Alexa Fluor® 647 Mouse anti-NKX6.1 Clone R11-560 (BD Bioscience; 1:10) or with appropriate class control antibodies, PE Mouse anti-IgG1 κ R-PE Clone MOPC-21 (BD Biosciences) and Alexa Fluor® 647 Mouse anti IgG1 κ Isotype Clone MOPC-21 (BD Biosciences). Cells were stained for 75 minutes at room temperature, washed three times, resuspended in PBS containing 1% BSA and 1% Formaldehyde, and immediately acquired using FACS Canto II flow cytometer and analyzed using FlowJo software V 10.4. The fractions of PDX1 and NKX6-1-positive cells varied across the analyzed iPSC-PPC lines, where percent PDX1/NKX6-1 double-positive expression ranged from 20.5 – 91.7% (mean = 72.31%; median = 76.6%).

Generation of scRNA-seq

Library Generation: One 1 iPSC line (from subject: iPSC_PPC034) and 10 iPSC-PPC samples were used for scRNA-seq generation (Table S1). Fresh cells from the iPSC line and from seven iPSC-PPC samples were captured individually at D15. Four of these same iPSC-PPC samples were also captured as cryopreserved cells (immediately after thawing) along with three iPSC-PPC samples that were only captured as cryopreserved cells. Cells from four cryopreserved iPSC-PPC samples were pooled (RNA_Pool_1), and cells from the other 3 iPSC-PPC samples were pooled (RNA_Pool_2) prior to capture (Table S1). All single cells were captured using the 10x Chromium controller (10x Genomics) according to the manufacturer's specifications and manual (Manual CG000183, Rev A). Cells from each scRNA-seq sample (1 iPSC, 7 fresh iPSC-PPCs, RNA_Pool_1, and RNA_Pool_2) were loaded on an individual lane of a Chromium Single Cell Chip B. Libraries were generated using Chromium Single Cell 3' Library Gel Bead Kit v3 (10x Genomics) following manufacturer's manual with small modifications. Specifically, the purified cDNA was eluted in 24μl of Buffer EB, half of which was used for the subsequent step of the library construction. cDNA was amplified for 10 cycles and libraries were amplified for 8 cycles.

Sequencing: Libraries produced from fresh and cryopreserved cells were sequenced on a HiSeq 4000 using custom programs (fresh: 28-8-175 Pair End and cryopreserved: 28-8-98 Pair End). Specifically, 8 libraries generated from fresh samples (1 iPSC and 7 iPSC-PPC samples) were pooled together and loaded evenly on 8 lanes and sequenced to an average depth of 163 million reads. Two libraries from seven cryopreserved lines (RNA_Pool_1 and RNA_Pool_2) were each sequenced on an individual lane to an average depth of 264 million reads. In total, we captured 98,474 cells.

Processing scRNA-seq data

Raw data processing. We retrieved FASTQ files for 10 scRNA-seq samples (one iPSC, seven fresh iPSC-PPCs, one RNA_Pool_1, and one RNA_Pool_2) and used CellRanger V2.1 (https://support.10xgenomics.com/) with default parameters and Gencode V19 gene annotations to generate single-cell gene counts and BAM files for each individual sample (Table S1). We then applied the BCFtools (Li et al., 2009) variant calling command *mpileup* to each BAM file to call single-nucleotide polymorphisms (SNPs).

Demultiplexing. To reassign pooled cells iPSC-PPCs back to the original subject (RNA_Pool_1 and RNA_Pool_2; Table S1), we obtained the BAM files for each scRNA-seq sample and a VCF file containing SNPs (called from WGS) that are bi-allelic and located at UTR or exon regions on autosomes as annotated by GENCODE calls from each of the seven subjects. The two files (BAM and VCF) were used as input to Demuxlet (Kang et al., 2018), which outputted the subject identities of each single cell based on genotype.

Quality Control of demultiplexing. To verify if Demuxlet assigned the correct subject identity of each cell, we performed two validation experiments. First, we simulated a pooled scRNA-seq experiment from cells with known subject identities. To generate the simulated pool sample, we downsampled FASTQ files from 6 independent scRNA-seq samples from 6 subjects (Table S1) to 40 million reads each (total 240 million reads). We then reprocessed the downsampled FASTQ files using CellRanger V2.1 (https://support.10xgenomics.com/) to generate single-cell gene counts and BAM files for the simulated pool sample. Assignment of subject identity of the pooled cells using Demuxlet (882,312 SNPs at exon or UTR regions) resulted in 99.99% of the pooled cells being correctly assigned to their true subject identity. Second, to verify the accuracy of Demuxlet results, we examined the concordance rate of called SNPs between WGS and three scRNA-seq samples (RNA_Pool_1, RNA_Pool_2, and simulated pool sample). The examined SNPs were selected based on the following criteria: 1) at UTR or exon regions on autosomes as annotated by GENCODE, 2) at peaks of at least 20X depth and 100bp long, and exist in all samples, and 3) are homozygous alternate in at least one sample and not heterozygous in the remaining samples, resulting in 261 SNPs analyzed. Comparing WGS SNPs to scRNA-seq SNPs, we observed 96.1% to 100.0% concordance rate in all samples (Table S8).

Identification of doublets. To detect occurrences of two cells co-encapsulated in a single droplet (i.e. doublet cells), we applied the Scrublet (Wolock et al., 2019) pipeline to the 10 scRNA-seq samples (2 pooled and 8 individual) (Table S1). Identification of predicted doublets were based on cells with doublet scores greater than or equal to an automated threshold set by the tool. Scrublet predicted <1% of cells to be doublets in all scRNA-seq samples but PPC_036 (Table S9). As the percentage of doublets per scRNA-seq was negligible, we concluded no predicted doublets should be excluded from downstream analyses.

Data Processing. To merge the 10 scRNA-seq samples (1 iPSC, 7 fresh iPSC-PPCs, RNA_Pool_1, and RNA_Pool_2), we first aggregated the samples that were sequenced as a single batch using CellRanger V2.1 (https://support.10xgenomics.com/) *aggr*, including the 1 iPSC and 7 fresh iPSC-PPC scRNA-seq samples. We next log normalized each sample (aggregated sample, RNA_Pool_1, and RNA_Pool_2), then applied Seurat's standard integration workflow to account for batch effects, including the *FindIntregrationAnchors* and *IntegrateData* functions. To filter low quality cells from downstream analyses, we removed cells with fewer than 500 genes/cell. After filtering, 88,188 cells remained for downstream analyses. Variable genes were then identified using a threshold of 0.5 for the standardized log dispersion. Principal component analysis (PCA) was then performed on the variable genes and significant PCs. Clustering was performed using a shared-nearest-neighbor (SNN) graph of the significant PCs and single cells were visualized using Uniform Manifold Approximation and Projection (UMAP).

Integrating scRNA-seq from ESC-B time course and iPSC-PPCs

Processed scRNA-seq data and metadata (cell annotations) from four differentiation stages of SC-ß cells (Mahajan et al., 2018): Stage 3 (Day 6; 7,982 cells), Stage 4 (Day 13; 6,960 cells), Stage 5 (Day 18; 4,193 cells), and Stage 6 (Day 25; 5,186 cells) were downloaded from GEO (GSE114412). Briefly, as described in the original publication, the ES-ß scRNA-seq data was filtered by removing genes with fewer than three counts, mitochondrially encoded, or under-annotated, removing cells with fewer than 750 UMI counts, and normalizing the sum of counts per cell to 10,000, resulting in 24,321 cells. To integrate scRNA-seq across two datasets: 1) processed iPSC and iPSC-PPC samples; and 2) processed differentiating SC-ß cells (SC-ß study cells), both datasets were log normalized, then Seurat's standard integration workflow was applied using *FindIntregrationAnchors* and *IntegrateData* functions. The integrated data was then scaled (*ScaleData*), variable genes were identified using a threshold of 0.5 for the standardized log dispersion, and PCA was performed on the variable genes and significant PCs. Clustering was performed using a SNN graph of the significant PCs and cells were visualized using UMAP.

Annotation of iPSC-PPC cells

To annotate the 88,188 iPSC-PPC cells, we examined each cluster from the integrated dataset (iPSC-PPC study cells and SC-ß cells, Figure 3A,B), independently. To characterize iPSC-PPC cells, we used the annotations from the SC-ß study as reference and confirmed each annotation by investigating marker gene expression. For clusters where only iPSC-PPC cells mapped (cluster 3, 4, and 5), we calculated the top significantly differentially overexpressed genes from each cluster versus all other clusters using Seurat's *FindMarker* function. These clusters were then annotated based on expression of markers characteristic of distinct cell types.

Generation of snATAC-seq

Library Generation: A total of 7 iPSC-PPC samples were used for snATAC-seq generation (Table S1). Cells from seven cryopreserved iPSC-PPCs samples were captured for snATAC-seq immediately after thawing. Cells from four cryopreserved iPSC-PPC samples were pooled (ATAC_Pool_1) and cells from the other 3 iPSC-PPC samples were pooled (ATAC_Pool_2) prior to capture (Table S1). Nuclei from two pools were isolated according to the manufacturer's recommendations (Manual CG000169, Rev B), transposed, and captured as independent samples according to the manufacturer's recommendations (Manual CG000168, Rev B). All single nuclei were captured using the 10x Chromium controller (10x Genomics) according to the manufacturer's specifications and manual (Manual CG000168, Rev B). Cells for each sample were loaded on the individual lane of a Chromium Chip E. Libraries were generated using Chromium Single Cell ATAC Library Gel & Bead Kit (10x Genomics) following manufacturer's manual (Manual CG000168, Rev B). Sample Index PCR material was amplified for 11 cycles.

Sequencing: Libraries were sequenced using a custom program (50-8-16-50 Pair End) on HiSeq 4000. Specifically, two libraries from seven cryopreserved iPSC-PPC samples (ATAC_Pool_1 and ATAC_Pool_2) were each sequenced on an individual lane.

Processing snATAC-seq data

Raw data processing. For two snATAC-seq samples (ATAC_Pool_1 and ATAC_Pool_2; Table S1), we retrieved FASTQ files and used CellRanger V2.1 (https://support.10xgenomics.com/) to align files to the hg19 genome using *cellranger-atac count* with default parameters. NarrowPeaks were called using the MACS2 command *macs2 callpeak --keep-dup all --nomodel --call-summits* (Feng et al., 2012) on the BAM files from the two pooled samples and detected 289,697 peaks. Using the peaks called by MACS2, each snATAC-seq sample was then reanalyzed using *cellranger-atac reanalyze* to generate single-nuclei peak counts and BAM files for each sample. To integrate the two snATAC-seq datasets for downstream analyses, we performed *cellranger-atac aggr* and obtained 31,587 nuclei.

Demultiplexing. To reassign pooled nuclei back to the original subject from two snATAC-seq samples (ATAC_Pool_1 and ATAC_Pool_2; Table S1), we applied Demuxlet (Kang et al., 2018) to the two samples.

Filtering snATAC-seq results: To analyze the snATAC-seq nuclei and determine cell types, we used the Seurat and Signac pipelines (Butler et al., 2018). We performed dimensionality reduction using the LSI method in the *RunSVD* function, obtained UMAP coordinates using the *RunUMAP* function and performed clustering using the *FindClusters* function with *resolution = 0.013* (Figure S4A). Since we observed that many nuclei from clusters 0 and 1 mapped to other clusters, we further investigated these two clusters independently and removed nuclei not correctly mapped, thereby reducing the noise in downstream analyses. We used the following approach:

1- We obtained all nuclei mapping to cluster 0 using the *subset* function, re-performed dimensionality reduction, UMAP and clustering;

2- We visually inspected the UMAP plots and removed two sub-clusters (3 and 4, Figure S4B, C, D), which had nuclei mapping to other clusters in the original UMAP space;

3- From each of the remaining sub-clusters, we removed all nuclei mapped to other sub-clusters.

Similarly, we reanalyzed cluster 1 and removed its clusters 1, 2 and 4 (Figure S4E, F, G). We obtained 29,128 nuclei, which we used for dimensionality reduction, UMAP and clustering with *resolution = 0.2*. We identified six clusters (Figure 4A).

Motif enrichment analysis to annotate snATAC-seq clusters and identify cell types

For all nuclei included in each cluster, we performed differential peak expression analysis using the *FindMarkers* function in Seurat with parameters *only.pos = FALSE, min.pct = 0.1, logfc.threshold = 0.1* to identify both overexpressed and downregulated peaks associated with each cluster (Butler et al., 2018). To detect transcription factor binding sites associated with each cluster, we performed Homer *findMotifsGenome.pl* using overexpressed peaks as input regions, downregulated peaks as background and HOCOMOCO V.11 motifs (Kulakovskiy et al.,

2018) as known transcription factor binding sites (478 motifs from 401 transcription factors). For each cluster, p-values detected using Homer were corrected for FDR using Benjamini-Hochberg's method (Table S5).

Processing T2D loci

We obtained the genomic coordinates of each of 380 fine-mapped T2D loci from Mahajan et al. (Mahajan et al., 2018) . For each variant with PPA > 0 in each locus, we extracted all its iPSC-PPC snATAC-seq peaks, as well as snATAC-seq and bulk ATAC-seq peaks from adult pancreatic islets (Rai et al., 2020), using *bedtools intersect* (Quinlan and Hall, 2010). To identify peaks active only in fetal-like pancreas, only in adult pancreas or both, we merged all peaks using *bedtools merge*. For each locus, we selected the variant with highest PPA that overlapped ATAC-seq peaks for downstream analyses.

Calculating the effects of T2D variants on transcription factor binding

For each of the 208 variants overlapping fetal ATAC-seq peaks, we obtained their sequence up to 20 bp upstream and downstream using *bedtools getfasta* on the hg19 genome and created two fasta sequences (one with the effect variant and one with the non-effect variant). Next, we used FIMO (Grant et al., 2011) using HOCOMOCO V.11 transcription factor motifs (Kulakovskiy et al., 2018) to find associations between each allele and transcription factor binding. We considered all motifs that had q-value (Benjamini-Hochberg's method, calculated by FIMO) < 0.05 and score > 4 as significant.

Allele-specific effects of T2D variants on fetal-like ATAC-seq peaks

For each of the 208 variants overlapping fetal ATAC-seq peaks, we obtained their genotype in the seven sequenced iPSCORE individuals using BCFtools (Li et al., 2009) and found 161 heterozygous in at least one individual. We created a BAM file including all the reads overlapping each of these variants using Samtools and filtered only the heterozygous individuals by intersecting the barcodes ("CB:Z" column in the BAM files) with the barcode-to-subject assignments detected using Demuxlet. To obtain the genotype of all the reads at each variant, we used *Samtools mpileup*. For each variant with read depth ≥ 35, we calculated allele-specific effects using *binom.test* function in R with option *p = 0.5* and corrected p-values for FDR using Benjamini-Hochberg's method (*p.adjust* function in R).

Data availability

iPSC-PPC scRNA-seq and snATAC-seq data was submitted to GEO: GSE152610.

Acknowledgements

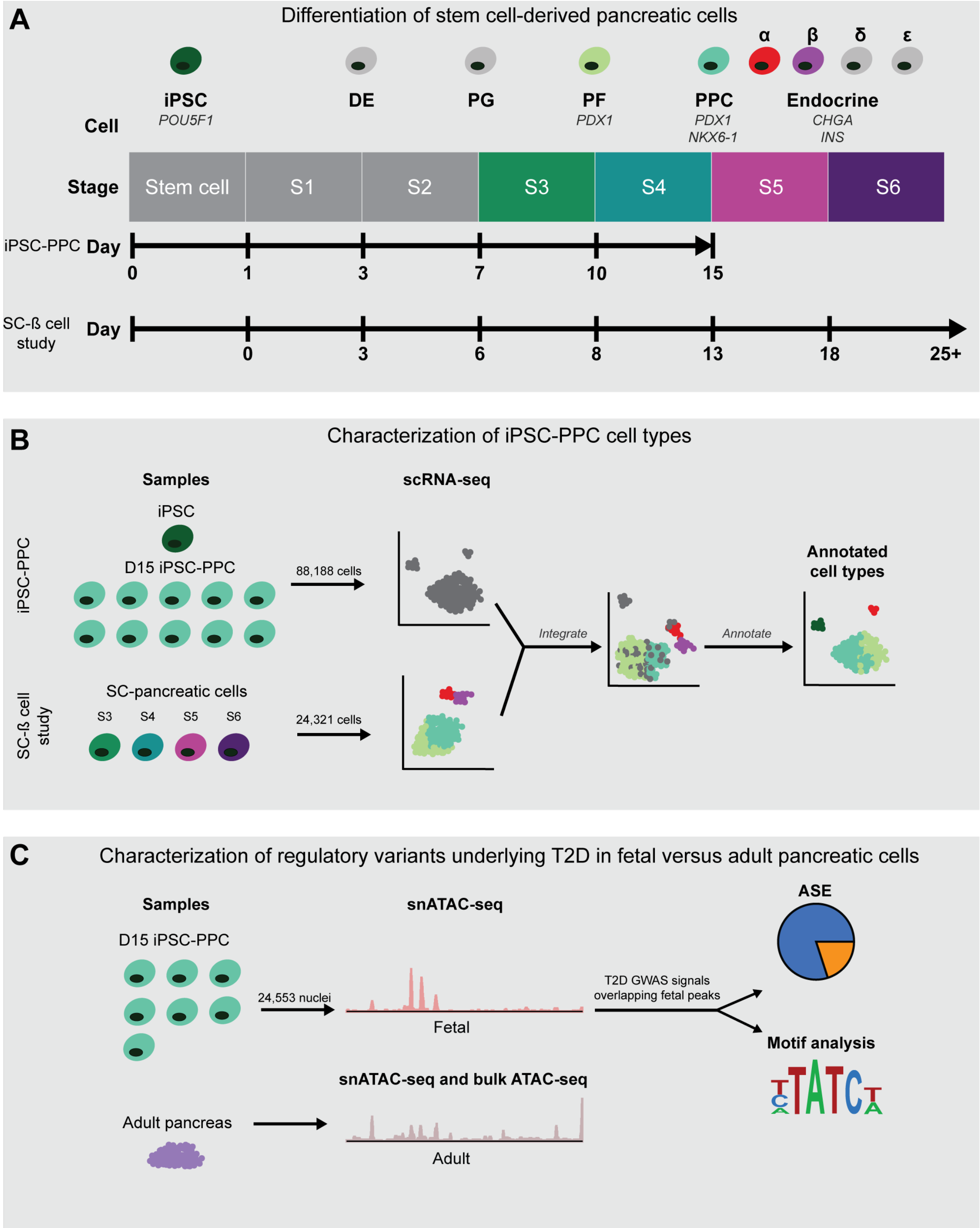
This work was supported in part by a California Institute for Regenerative Medicine (CIRM) grant GC1R-06673 and NIH grants HG008118, HL107442, DK105541, and DK112155. JPN, TDA and MKRD were supported by the National Library of Medicine Training Grant T15LM011271. BS was supported by CIRM Bridges Stem Cell Research and Therapy Training Grant, Award #EDUC2-08375.

Author information

KAF, ADC, MKRD, and MD conceived the study. ADC, KF and BMS performed iPSC-PPC differentiations. ADC generated the molecular data. MKRD, HM, MD, JPN and TDA performed computational analysis. MKRD and JPN performed scRNA-seq data processing and analyses. HM, MD and JPN performed the snATAC-seq data processing and analyses. KAF oversaw the study. KAF, MKRD and MD prepared the manuscript.

Figures

Figure 1: Study design

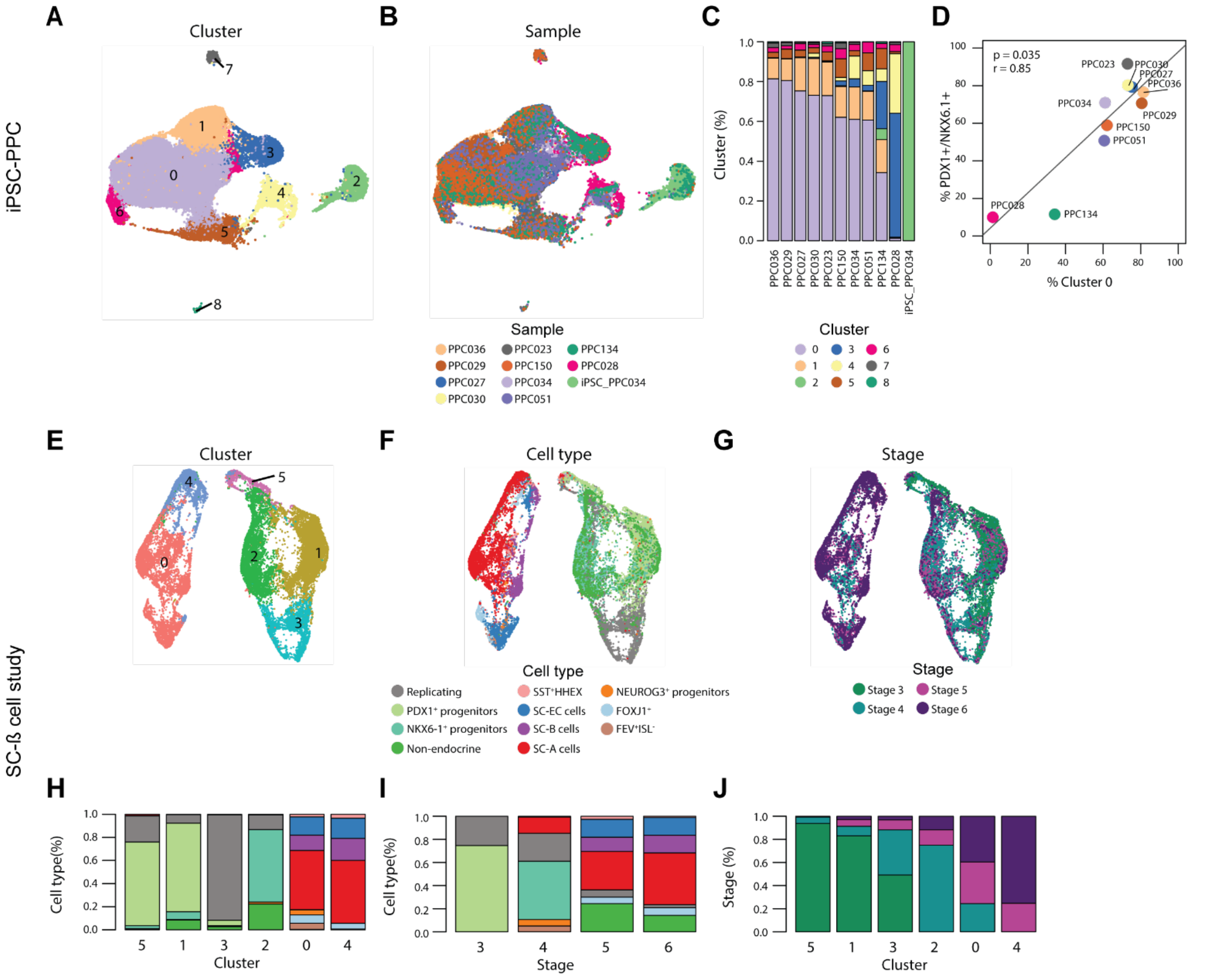


A. Cartoon comparing the differentiation protocols for iPSC-PPC (15 days) and SC-ß cells (SC-ß cell study; 25 days). Across the two derivation methods, the various cell stages/types of differentiating stem cells as they mature to PPCs and more mature endocrine cells are highlighted. Marker genes for cell stages/types are also noted.

B. Cartoon depicting annotation of cell types in the iPSC-PPC samples. scRNA-seq from ten iPSC-PPCs and one iPSC clone and from the SC-ß cell study were integrated in order to identify which cell types the unannotated iPSC-PPCs mapped to in the reference SC-ß cell study.

C. Cartoon depicting the advantage of using iPSC-PPCS in conjunction with adult pancreatic cells to functionally examine T2D GWAS regulatory variants. First, we identified peaks unique to or overlapping both fetal iPSC-PPC (snATAC-seq) and adult pancreas cells (snATAC-seq and bulk ATAC-seq). Then, we identified T2D GWAS signals that overlapped fetal peaks (unique to fetal or fetal and adult) and examined ASE and performed motif analyses.

Figure 2: Cellular heterogeneity in stem cell derived pancreatic precursor samples



A. UMAP plot of clustered scRNA-seq data from 88,188 single cells from one iPSC and 10 iPSC-PPC samples (i.e. iPSC-PPC study cells). Each point represents a single cell color-coded by its assigned cluster.

B. UMAP plot of clustered scRNA-seq data from 88,188 single cells from one iPSC and 10 iPSC-PPC samples (i.e. iPSC-PPC study cells). Each point represents a single cell color-coded by its unique differentiation identifier (UDID).

C. Stacked bar plot showing the fraction of cells from each sample (UDID) assigned to each cluster. Color-coding is the same as in Figure 2A. Differentiations PPC029 and PPC036 were from the same iPSC clone.

D. Scatter plot showing the association of the fraction of each sample's cells assigned to cluster 0 (high PDX1 and NKX6-1 expressing) with the fraction of cells positively stained for both PDX1 and NKX6-1.

E. UMAP plot of clustered scRNA-seq data from 24,321 SC-ß cells collected from four stages of differentiation (i.e. SC-ß study cells). Each point represents a single cell color-coded by its assigned cluster.

F. UMAP plot of clustered scRNA-seq data from 24,321 SC-ß cells collected from four stages of differentiation (i.e. SC-ß study cells). Each point represents a single cell color-coded by its annotated cell type.

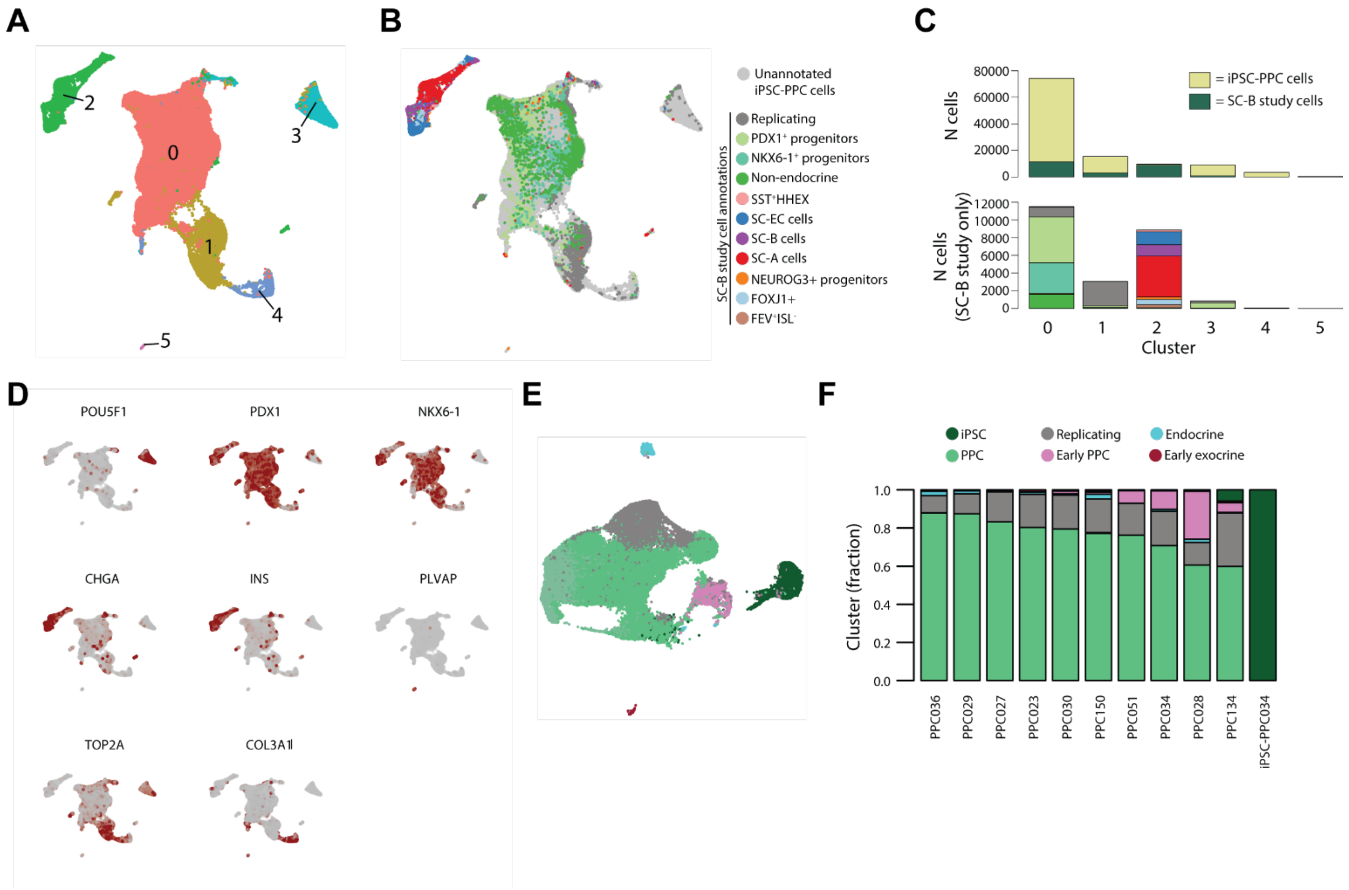
G. UMAP plot of clustered scRNA-seq data from 24,321 SC-ß cells collected from four stages of differentiation (i.e. SC-ß study cells). Each point represents a single cell color-coded by the stage the cell was obtained (color codes are the same as in Veres et al. (Veres et al., 2019)).

H. Stacked bar plot showing the fraction of cells from each of the 11 annotated cell types that map into each cluster (0-5). Color-coding corresponds to cell type annotations from Figure 2F.

I. Stacked bar plot showing the fraction of cells from each of 11 annotated cell types that map into each stage (3, 4, 5 and 6). Color-coding corresponds to cell type annotations from Figure 2F.

J. Stacked bar plot showing the fraction of cells from each stage (3, 4, 5 and 6) that map into each cluster (0-5). Color-coding corresponds to stage annotations from Figure 2G.

Figure 3: Annotation of iPSC-PPCs using a reference SC-ß cell study dataset



A. UMAP plot of clustered scRNA-seq data from 112,509 integrated cells from the iPSC-PPC and SC-ß cell studies. Each point represents a single cell color-coded by its assigned cluster.

B. UMAP plot of 112,509 integrated cells. Each point represents a single cell color-coded by its cell type annotation defined in the SC-ß study. Unannotated iPSC-PPC cells are filled in a light shade of grey.

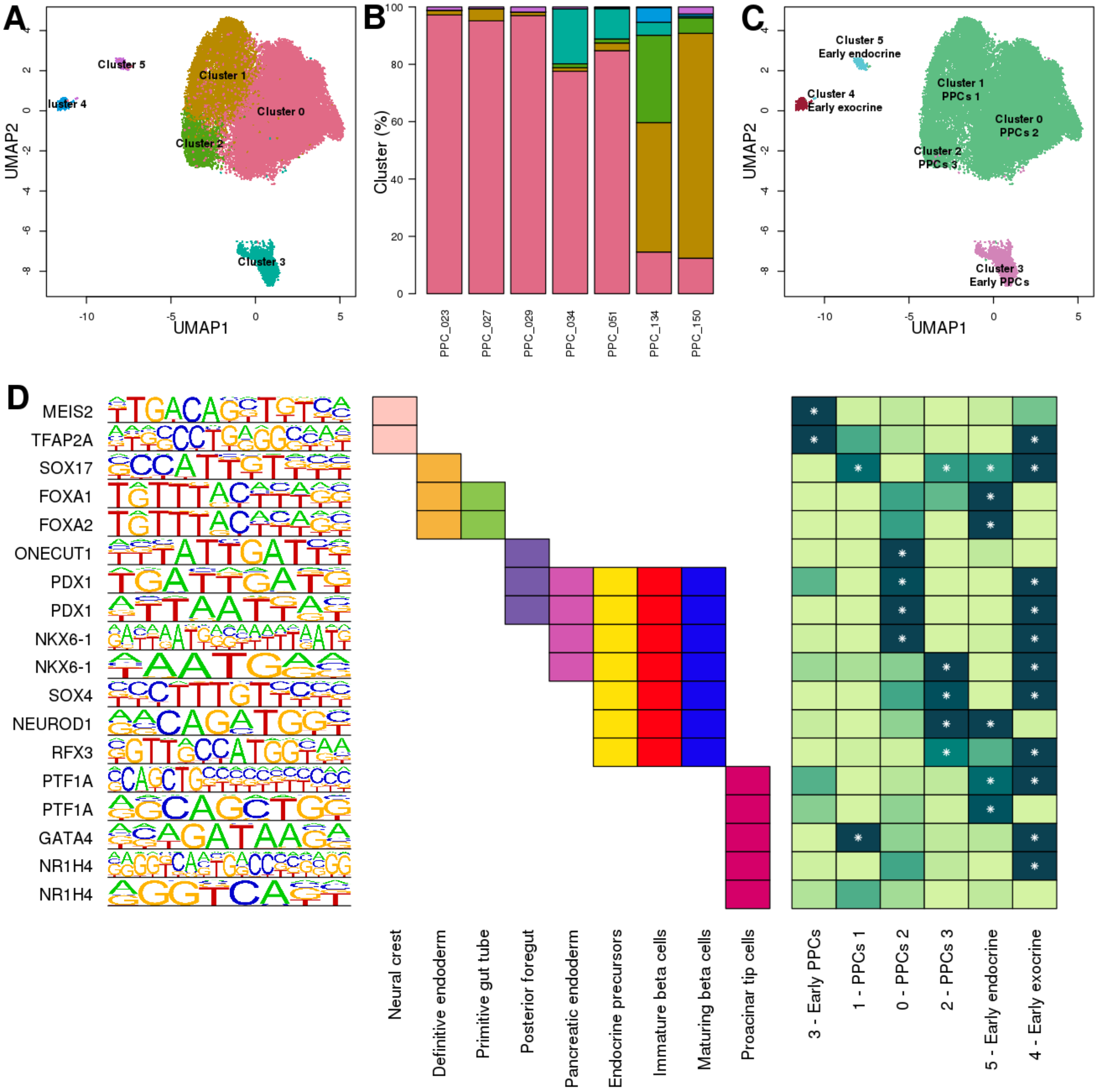
C. Stacked bar plot showing the fraction of: 1) iPSC-PPC cells and SC-ß study cells assigned to each cluster (*top*); and 2) SC-ß study cell types assigned to each cluster (*bottom*).

D. UMAP plots showing the expression of marker genes for stem cells (*POU5F1*), PPCs (*PDX1*, *NKX6-1*), endocrine cells (*CHGA*, *INS*), early exocrine/ductal cells (*PLVAP*), replicating cells (*TOP2A*), and early PPC (*COL3A1*).

E. UMAP plot of 88,188 cells from 10 iPSC-PPC differentiations and one iPSC. Each point represents a single cell color-coded by its assigned cell type based on expression of marker genes.

F. Stacked bar plot showing for each iPSC-PPC sample the fraction of each assigned cell type.

Figure 4: Epigenomic analysis of iPSC-PPCs



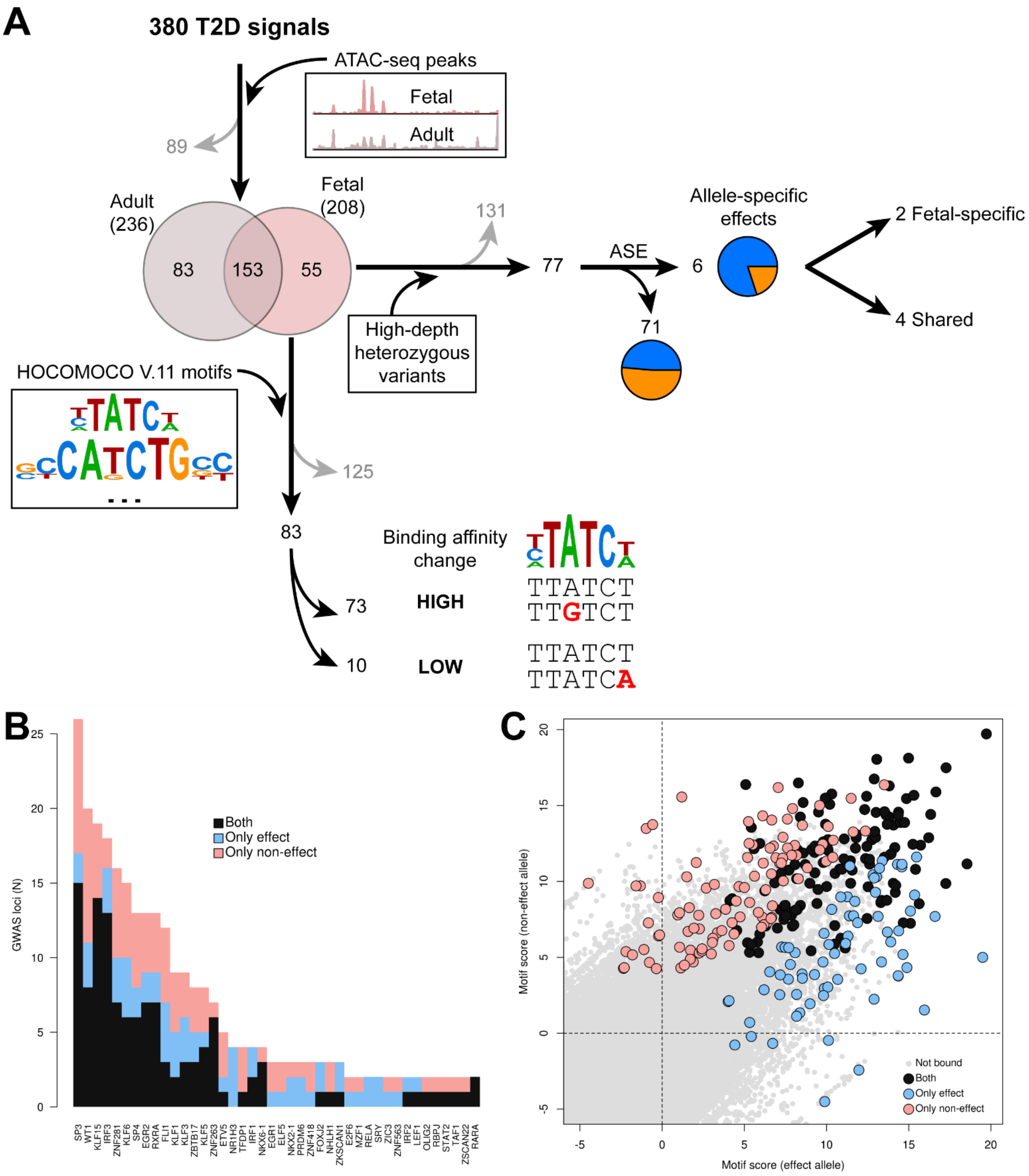
A. UMAP plot of 29,128 iPSC-PPC nuclei. Each point represents a single nucleus color-coded by its assigned cluster.

B. Stacked bar plot showing the fraction of each iPSC-PPC sample's nuclei assigned to each cluster.

C. UMAP plot of the 29,128 iPSC-PPC nuclei. Cell types were determined based on the enrichment for pancreas development transcription factor binding sites at peaks specific for each cluster defined in panel A. Colors correspond to the cell types defined in Figure 3E-F.

D. Transcription factor binding site enrichment for transcription factors that drive different stages of pancreatic development. For each transcription factor, its consensus motif (left), the developmental stage where it is active and the enrichment (calculated using Homer) are indicated. Colors represent enrichment values. Significant associations (Benjamini-Hochberg-corrected p-values) are shown with stars.

Figure 5: Associations between T2D loci and ATAC-seq peaks

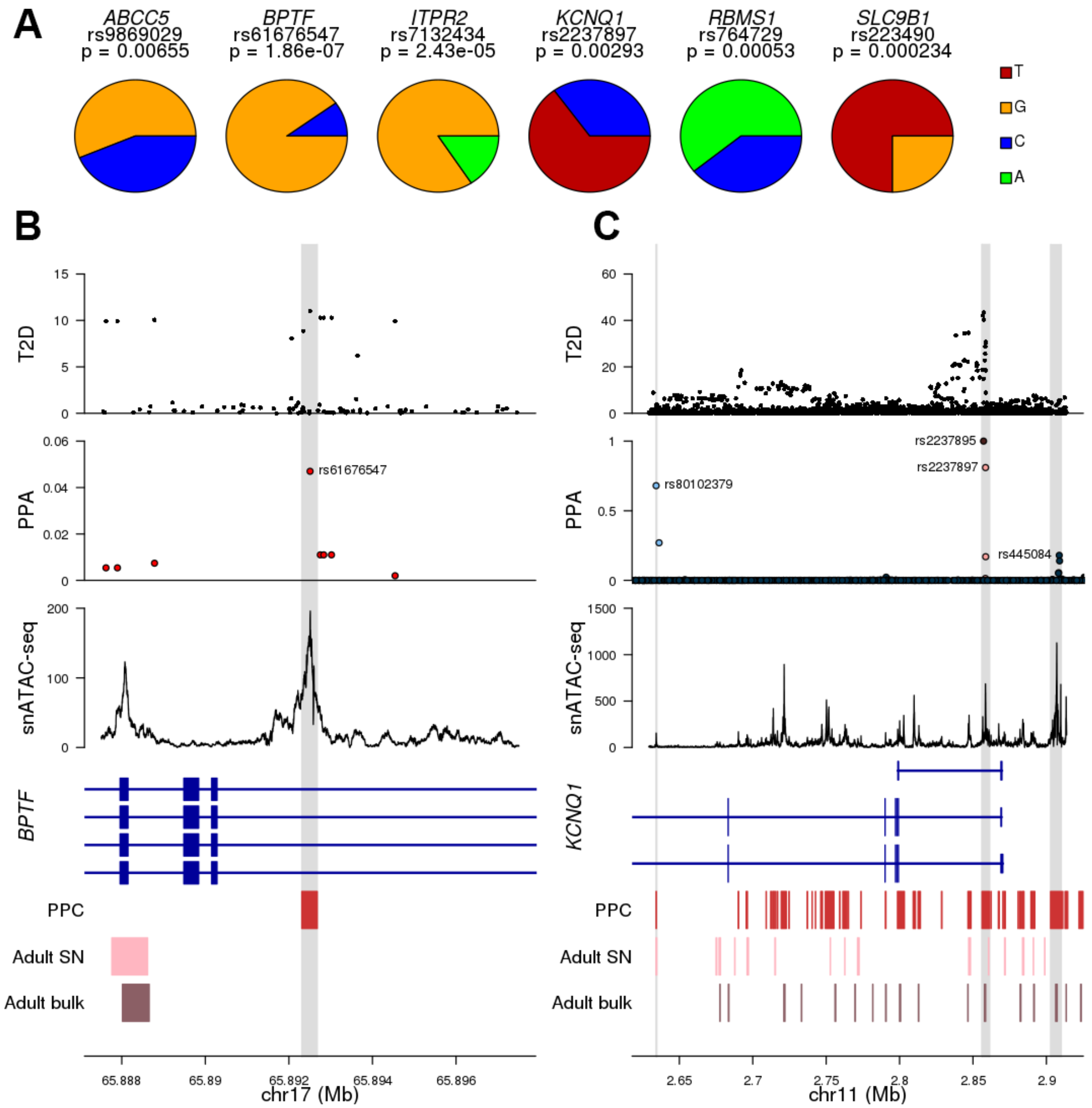


A. Functional analysis of 380 T2D signals in fetal-like pancreatic cells. We intersected each of the 380 fine-mapped T2D signals with fetal-like and adult ATAC-seq peaks and identified 208 T2D variants active in fetal-like pancreas, of which 153 were shared with adult peaks and 55 were fetal-specific. For each of these 208 variants, we tested: 1) their effect on transcription factor binding affinity, and found that 38 variants were associated with large changes; and 2) their allele-specific effects, and observed that six heterozygous variants were associated with significant differences between their effect and non-effect alleles.

B. For each transcription factor with binding sites overlapping at least two of the 83 fetal T2D risk loci, we show the number of T2D risk loci for which only the non-effect allele is associated with the transcription factor motif (pink), only the effect allele is associated (blue) or both (black) are associated.

C. Scatterplot showing the associations between the motif scores calculated using FIMO in the sequences harboring the effect variant (X axis) and the non-effect variant (Y axis). Colors are as in panel B.

Figure 6: Allele-specific effects of T2D variants



A. For each of the six loci with significant ASE, shown are their associated variant, the ASE p-value (binomial test) and a pie chart showing the distributions between effect and non-effect alleles (Table S6).

B. For two loci (*BPTF* and *KCNQ1*), shown are (top to bottom): the GWAS signal in T2D; the posterior probability of association (PPA) for all the fine-mapped variants in the locus; the read depth in iPSC-PPC snATAC-seq; the exon structure of all transcripts (Gencode V.19); the position of all iPSC-PPC, adult snATAC-seq and adult bulk ATAC-seq peaks. Peaks overlapping with the lead variants are shown in gray. For *KCNQ1*, PPA is shown for four distinct signals with their lead variant overlapping ATAC-seq peaks.

References

Akiyama, M., Okada, Y., Kanai, M., Takahashi, A., Momozawa, Y., Ikeda, M., Iwata, N., Ikegawa, S., Hirata, M., Matsuda, K.*, et al.* (2017). Genome-wide association study identifies 112 new loci for body mass index in the Japanese population. Nat Genet *49*, 1458-1467.

Barker, D.J. (1995). Fetal origins of coronary heart disease. BMJ *311*, 171-174.

Baron, M., Veres, A., Wolock, S.L., Faust, A.L., Gaujoux, R., Vetere, A., Ryu, J.H., Wagner, B.K., Shen-Orr, S.S., Klein, A.M.*, et al.* (2016). A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. Cell Syst *3*, 346-360 e344.

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol *36*, 411-420.

Cano, D.A., Soria, B., Martin, F., and Rojas, A. (2014). Transcriptional control of mammalian pancreas organogenesis. Cell Mol Life Sci *71*, 2383-2402.

Chiou, J., Zeng, C., Cheng, Z., Han, J.Y., Schlichting, M., Huang, S., Wang, J., Sui, Y., Deogaygay, A., Okino, M.L.*, et al.* (2019). Single cell chromatin accessibility reveals pancreatic islet cell type- and state-specific regulatory programs of diabetes risk. bioRxiv.

Consortium, G.T. (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science *369*, 1318-1330.

Consortium, S.T.D., Williams, A.L., Jacobs, S.B., Moreno-Macias, H., Huerta-Chagoya, A., Churchhouse, C., Marquez-Luna, C., Garcia-Ortiz, H., Gomez-Vazquez, M.J., Burtt, N.P.*, et al.* (2014). Sequence variants in SLC16A11 are a common risk factor for type 2 diabetes in Mexico. Nature *506*, 97-101.

D'Antonio-Chronowska, A., Donovan, M.K.R., Young Greenwald, W.W., Nguyen, J.P., Fujita, K., Hashem, S., Matsui, H., Soncin, F., Parast, M., Ward, M.C.*, et al.* (2019). Association of Human iPSC Gene Signatures and X Chromosome Dosage with Two Distinct Cardiac Differentiation Trajectories. Stem Cell Reports *13*, 924-938.

D'Antonio, M., Benaglio, P., Jakubosky, D., Greenwald, W.W., Matsui, H., Donovan, M.K.R., Li, H., Smith, E.N., D'Antonio-Chronowska, A., and Frazer, K.A. (2018). Insights into the Mutational Burden of Human Induced Pluripotent Stem Cells from an Integrative Multi-Omics Approach. Cell Rep *24*, 883-894.

DeBoever, C., Li, H., Jakubosky, D., Benaglio, P., Reyna, J., Olson, K.M., Huang, H., Biggs, W., Sandoval, E., D'Antonio, M.*, et al.* (2017). Large-Scale Profiling Reveals the Influence of Genetic Variation on Gene Expression in Human Induced Pluripotent Stem Cells. Cell Stem Cell *20*, 533-546 e537.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M.*, et al.* (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet *43*, 491-498.

Enge, M., Arda, H.E., Mignardi, M., Beausang, J., Bottino, R., Kim, S.K., and Quake, S.R. (2017). Single-Cell Analysis of Human Pancreas Reveals Transcriptional Signatures of Aging and Somatic Mutation Patterns. Cell *171*, 321-330 e314.

Feng, J., Liu, T., Qin, B., Zhang, Y., and Liu, X.S. (2012). Identifying ChIP-seq enrichment using MACS. Nat Protoc *7*, 1728-1740.

Gaulton, K.J., Ferreira, T., Lee, Y., Raimondo, A., Magi, R., Reschen, M.E., Mahajan, A., Locke, A., Rayner, N.W., Robertson, N.*, et al.* (2015). Genetic fine mapping and genomic annotation defines causal mechanisms at type 2 diabetes susceptibility loci. Nat Genet *47*, 1415-1425.

Genomes Project, C., Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A.*, et al.* (2015). A global reference for human genetic variation. Nature *526*, 68-74.

Geusz, R.J., Wang, A., Chiou, J., Lnacman, J.J., Wetton, N., Kefalopoulu, S., Wang, J., Qiu, Y., Yan, J., Aylward, A.*, et al.* (2020). Pancreatic progenitor epigenome maps prioritize type 2 diabetes risk genes with roles in development. BioRxiv.

Grant, C.E., Bailey, T.L., and Noble, W.S. (2011). FIMO: scanning for occurrences of a given motif. Bioinformatics *27*, 1017-1018.

Greenwald, W.W., Chiou, J., Yan, J., Qiu, Y., Dai, N., Wang, A., Nariai, N., Aylward, A., Han, J.Y., Kadakia, N.*, et al.* (2019). Pancreatic islet chromatin accessibility and conformation reveals distal enhancer networks of type 2 diabetes risk. Nat Commun *10*, 2078.

Guenard, F., Deshaies, Y., Cianflone, K., Kral, J.G., Marceau, P., and Vohl, M.C. (2013). Differential methylation in glucoregulatory genes of offspring born before vs. after maternal gastrointestinal bypass surgery. Proc Natl Acad Sci U S A *110*, 11439-11444.

Jennings, R.E., Berry, A.A., Strutt, J.P., Gerrard, D.T., and Hanley, N.A. (2015). Human pancreas development. Development *142*, 3126-3137.

Jung, D.Y., Chalasani, U., Pan, N., Friedline, R.H., Prosdocimo, D.A., Nam, M., Azuma, Y., Maganti, R., Yu, K., Velagapudi, A.*, et al.* (2013). KLF15 is a molecular link between endoplasmic reticulum stress and insulin resistance. PLoS One *8*, e77851.

Kahn, S.E., Cooper, M.E., and Del Prato, S. (2014). Pathophysiology and treatment of type 2 diabetes: perspectives on the past, present, and future. Lancet *383*, 1068-1083.

Kang, H.M., Subramaniam, M., Targ, S., Nguyen, M., Maliskova, L., McCarthy, E., Wan, E., Wong, S., Byrnes, L., Lanata, C.M.*, et al.* (2018). Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. Nat Biotechnol *36*, 89-94.

Klarin, D., Damrauer, S.M., Cho, K., Sun, Y.V., Teslovich, T.M., Honerlaw, J., Gagnon, D.R., DuVall, S.L., Li, J., Peloso, G.M.*, et al.* (2018). Genetics of blood lipids among ~300,000 multi-ethnic participants of the Million Veteran Program. Nat Genet *50*, 1514-1523.

Kulakovskiy, I.V., Vorontsov, I.E., Yevshin, I.S., Sharipov, R.N., Fedorova, A.D., Rumynskiy, E.I., Medvedeva, Y.A., Magana-Mora, A., Bajic, V.B., Papatsenko, D.A.*, et al.* (2018). HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. Nucleic Acids Res *46*, D252-D259.

Kumari, M., Wang, X., Lantier, L., Lyubetskaya, A., Eguchi, J., Kang, S., Tenen, D., Roh, H.C., Kong, X., Kazak, L.*, et al.* (2016). IRF3 promotes adipose inflammation and insulin resistance and represses browning. J Clin Invest *126*, 2839-2854.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754-1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078-2079.

Lumey, L.H., Khalangot, M.D., and Vaiserman, A.M. (2015). Association between type 2 diabetes and prenatal exposure to the Ukraine famine of 1932-33: a retrospective cohort study. Lancet Diabetes Endocrinol *3*, 787-794.

Mahajan, A., Taliun, D., Thurner, M., Robertson, N.R., Torres, J.M., Rayner, N.W., Payne, A.J., Steinthorsdottir, V., Scott, R.A., Grarup, N.*, et al.* (2018). Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. Nat Genet *50*, 1505-1513.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M.*, et al.* (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res *20*, 1297-1303.

Mussig, K., Staiger, H., Machicao, F., Kirchhoff, K., Guthoff, M., Schafer, S.A., Kantartzis, K., Silbernagel, G., Stefan, N., Holst, J.J.*, et al.* (2009). Association of type 2 diabetes candidate polymorphisms in KCNQ1 with incretin and insulin secretion. Diabetes *58*, 1715-1720.

Nielsen, J.H., Haase, T.N., Jaksch, C., Nalla, A., Sostrup, B., Nalla, A.A., Larsen, L., Rasmussen, M., Dalgaard, L.T., Gaarn, L.W., *et al.* (2014). Impact of fetal and neonatal environment on beta cell function and development of diabetes. Acta Obstet Gynecol Scand *93*, 1109-1122.

Ohshige, T., Tanaka, Y., Araki, S., Babazono, T., Toyoda, M., Umezono, T., Watada, H., Suzuki, D., Iwamoto, Y., Kawamori, R., *et al.* (2010). A single nucleotide polymorphism in KCNQ1 is associated with susceptibility to diabetic nephropathy in japanese subjects with type 2 diabetes. Diabetes Care *33*, 842-846.

Pagliuca, F.W., Millman, J.R., Gurtler, M., Segel, M., Van Dervort, A., Ryu, J.H., Peterson, Q.P., Greiner, D., and Melton, D.A. (2014). Generation of functional human pancreatic beta cells in vitro. Cell *159*, 428-439.

Palmer, N.D., Goodarzi, M.O., Langefeld, C.D., Wang, N., Guo, X., Taylor, K.D., Fingerlin, T.E., Norris, J.M., Buchanan, T.A., Xiang, A.H., *et al.* (2015). Genetic Variants Associated With Quantitative Glucose Homeostasis Traits Translate to Type 2 Diabetes in Mexican Americans: The GUARDIAN (Genetics Underlying Diabetes in Hispanics) Consortium. Diabetes *64*, 1853-1866.

Panopoulos, A.D., D'Antonio, M., Benaglio, P., Williams, R., Hashem, S.I., Schuldt, B.M., DeBoever, C., Arias, A.D., Garcia, M., Nelson, B.C., *et al.* (2017). iPSCORE: A Resource of 222 iPSC Lines Enabling Functional Characterization of Genetic Variation across a Variety of Cell Types. Stem Cell Reports *8*, 1086-1100.

Pasquali, L., Gaulton, K.J., Rodriguez-Segui, S.A., Mularoni, L., Miguel-Escalada, I., Akerman, I., Tena, J.J., Moran, I., Gomez-Marin, C., van de Bunt, M., *et al.* (2014). Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. Nat Genet *46*, 136-143.

Peng, J., Sun, B.F., Chen, C.Y., Zhou, J.Y., Chen, Y.S., Chen, H., Liu, L., Huang, D., Jiang, J., Cui, G.S., *et al.* (2019). Single-cell RNA-seq highlights intra-tumoral heterogeneity and malignant progression in pancreatic ductal adenocarcinoma. Cell Res *29*, 725-738.

Petersen, M.B.K., Azad, A., Ingvorsen, C., Hess, K., Hansson, M., Grapin-Botton, A., and Honore, C. (2017). Single-Cell Gene Expression Analysis of a Human ESC Model of Pancreatic Endocrine Development Reveals Different Paths to beta-Cell Differentiation. Stem Cell Reports *9*, 1246-1261.

Pictet, R.L., Clark, W.R., Williams, R.H., and Rutter, W.J. (1972). An ultrastructural analysis of the developing embryonic pancreas. Dev Biol *29*, 436-467.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841-842.

Rai, V., Quang, D.X., Erdos, M.R., Cusanovich, D.A., Daza, R.M., Narisu, N., Zou, L.S., Didion, J.P., Guan, Y., Shendure, J., *et al.* (2020). Single-cell ATAC-Seq in human pancreatic islets and deep learning upscaling of rare cells reveals cell-specific type 2 diabetes regulatory signatures. Mol Metab *32*, 109-121.

Reichert, M., and Rustgi, A.K. (2011). Pancreatic ductal cells in development, regeneration, and neoplasia. J Clin Invest *121*, 4572-4578.

Rezania, A., Bruin, J.E., Arora, P., Rubin, A., Batushansky, I., Asadi, A., O'Dwyer, S., Quiskamp, N., Mojibian, M., Albrecht, T., *et al.* (2014). Reversal of diabetes with insulin-producing cells derived in vitro from human pluripotent stem cells. Nat Biotechnol *32*, 1121-1133.

Rubinstein, M., Idelman, G., Plymate, S.R., Narla, G., Friedman, S.L., and Werner, H. (2004). Transcriptional activation of the insulin-like growth factor I receptor gene by the Kruppel-like factor 6 (KLF6) tumor suppressor protein: potential interactions between KLF6 and p53. Endocrinology *145*, 3769-3777.

Segerstolpe, A., Palasantza, A., Eliasson, P., Andersson, E.M., Andreasson, A.C., Sun, X., Picelli, S., Sabirsh, A., Clausen, M., Bjursell, M.K., *et al.* (2016). Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes. Cell Metab *24*, 593-607.

Sharon, N., Chawla, R., Mueller, J., Vanderhooft, J., Whitehorn, L.J., Rosenthal, B., Gurtler, M., Estanboulieh, R.R., Shvartsman, D., Gifford, D.K., *et al.* (2019). A Peninsular Structure Coordinates Asynchronous Differentiation with Morphogenesis to Generate Pancreatic Islets. Cell *176*, 790-804 e713.

Spracklen, C.N., Horikoshi, M., Kim, Y.J., Lin, K., Bragg, F., Moon, S., Suzuki, K., Tam, C.H.T., Tabara, Y., Kwak, S.H., *et al.* (2020). Identification of type 2 diabetes loci in 433,540 East Asian individuals. Nature *582*, 240-245.

Suzuki, K., Akiyama, M., Ishigaki, K., Kanai, M., Hosoe, J., Shojima, N., Hozawa, A., Kadota, A., Kuriki, K., Naito, M.*, et al.* (2019). Identification of 28 new susceptibility loci for type 2 diabetes in the Japanese population. Nat Genet *51*, 379-386.

Unoki, H., Takahashi, A., Kawaguchi, T., Hara, K., Horikoshi, M., Andersen, G., Ng, D.P., Holmkvist, J., Borch-Johnsen, K., Jorgensen, T.*, et al.* (2008). SNPs in KCNQ1 are associated with susceptibility to type 2 diabetes in East Asian and European populations. Nat Genet *40*, 1098-1102.

Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J.*, et al.* (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics *43*, 11 10 11-11 10 33.

Varshney, A., Scott, L.J., Welch, R.P., Erdos, M.R., Chines, P.S., Narisu, N., Albanus, R.D., Orchard, P., Wolford, B.N., Kursawe, R.*, et al.* (2017). Genetic regulatory signatures underlying islet gene expression and type 2 diabetes. Proc Natl Acad Sci U S A *114*, 2301-2306.

Veres, A., Faust, A.L., Bushnell, H.L., Engquist, E.N., Kenty, J.H., Harb, G., Poh, Y.C., Sintov, E., Gurtler, M., Pagliuca, F.W.*, et al.* (2019). Charting cellular identity during human in vitro beta-cell differentiation. Nature *569*, 368-373.

Vinuela, A., Varshney, A., van de Bunt, M., Prasad, R.B., Asplund, O., Bennett, A., Boehnke, M., Brown, A.A., Erdos, M.R., Fadista, J.*, et al.* (2020). Genetic variant effects on gene expression in human pancreatic islets and their implications for T2D. Nat Commun *11*, 4912.

Vujkovic, M., Keaton, J.M., Lynch, J.A., Miller, D.R., Zhou, J., Tcheandjieu, C., Huffman, J.E., Assimes, T.L., Lorenz, K., Zhu, X.*, et al.* (2020). Discovery of 318 new risk loci for type 2 diabetes and related vascular outcomes among 1.4 million participants in a multi-ancestry meta-analysis. Nat Genet *52*, 680-691.

W.H.O., W.H.O. (2020). Diabetes Fact Sheet. https://wwwwhoint/news-room/fact-sheets/detail/diabetes.

Webster, N.J., Kong, Y., Sharma, P., Haas, M., Sukumar, S., and Seely, B.L. (1997). Differential effects of Wilms tumor WT1 splice variants on the insulin receptor promoter. Biochem Mol Med *62*, 139-150.

White, N.R., Mulligan, P., King, P.J., and Sanderson, I.R. (2006). Sodium butyrate-mediated Sp3 acetylation represses human insulin-like growth factor binding protein-3 expression in intestinal epithelial cells. J Pediatr Gastroenterol Nutr *42*, 134-141.

Wolock, S.L., Lopez, R., and Klein, A.M. (2019). Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. Cell Syst *8*, 281-291 e289.

Xue, A., Wu, Y., Zhu, Z., Zhang, F., Kemper, K.E., Zheng, Z., Yengo, L., Lloyd-Jones, L.R., Sidorenko, J., Wu, Y.*, et al.* (2018). Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes. Nat Commun *9*, 2941.