Understanding human fetal pancreas development using subpopulation sorting, RNA sequencing and single-cell profiling

Cyrille Ramond[1,2,3*], Belin Selcen Beydag-Tasöz[4*], Ajuna Azad[4*], Martijn van de Bunt[5,6,7], Maja Borup Kjær Petersen[4,8], Nicola L Beer[9], Nicolas Glaser[1,2,3], Claire Berthault[1,2,3], Anna L Gloyn[5,6,9], Mattias Hansson[10], Mark I. McCarthy[5,6,9], Christian Honoré[8], Anne Grapin-Botton[4#], Raphael Scharfmann[1,2,3#]

[1]INSERM U1016, Cochin Institute, Paris, France

[2]CNRS UMR 8104, Paris, France

[3]University of Paris Descartes, Sorbonne Paris Cité, Paris, France

[4]The Novo Nordisk Foundation Center for Stem Cell Biology (DanStem), Faculty of Health Sciences, University of Copenhagen, Denmark

[5]Wellcome Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN, UK

[6]Oxford NIHR Biomedical Research Centre, Churchill Hospital, Old Road, Headington, Oxford, OX3 7LJ UK

[7]Global Research Informatics, Novo Nordisk A/S, Novo Nordisk Park, Måløv, Denmark

[8]Department of Stem Cell Biology, Novo Nordisk A/S, Novo Nordisk Park, Måløv, Denmark

[9]Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Churchill Hospital, Old Road, Headington, Oxford, OX3 7LJ UK

[10]Stem Cell Research, Novo Nordisk A/S, Novo Nordisk Park, Måløv, Denmark.

*: Equal contribution

[#]: Equal contribution

Correspondence

Anne Grapin Botton (anne.grapin-botton@sund.ku.dk) and Raphael Scharfmann: (raphael.scharfmann@inserm.fr)

Abstract

To decipher the populations of cells present in the human fetal pancreas and their lineage relationships, we developed strategies to isolate pancreatic progenitors, endocrine progenitors and endocrine cells. Transcriptome analysis of the individual populations revealed a large degree of conservation among vertebrates in the drivers of gene expression changes occurring at different steps of differentiation, although notably, sometimes, different members of the same gene family are expressed. The transcriptome analysis establishes a resource to identify novel genes and pathways involved in human pancreas development. Single cell profiling further captured intermediate stages of differentiation and enabled us to decipher the sequence of transcriptional events occurring during human endocrine differentiation. Furthermore, we evaluate how well individual pancreatic cells derived *in vitro* from human pluripotent stem cells mirror the natural process occurring in human fetuses. This comparison uncovers a few differences at the progenitor steps, a convergence at the steps of endocrine induction and the current inability to fully resolve endocrine cell subtypes *in vitro.*

Introduction

Due to the limited availability of primary human tissue, much of our knowledge of human organ development is extrapolated from animal models. It is remarkable that the mechanisms of organ formation are conserved enough between vertebrate species so as to enable biologists to control the differentiation of organ-specific cells from human pluripotent stem cells. In the pancreas, directed differentiation informed by the knowledge of development has enabled the field to progress to the production of beta cells with the ambition of transplanting these cells to treat diabetic patients who depending on sub-type, have reduced beta-cell function, or are absent of them entirely (D'Amour et al., 2006). Over the years, the protocols have become more efficient, adaptable to multiple pluripotent stem cell lines and lead to the production of ever-more representative functional cells (Pagliuca et al., 2014; Rezania et al., 2014; Russ et al., 2015; Nostro et al., 2015; Ameri et al., 2017; Cogger et al., 2017). Yet, ~~the~~ *in vitro* generated cells do not fully recapitulate the function of bona fide β cells, notably lacking a tight control of insulin secretion upon glucose stimulation (Johnson, 2016). Accordingly, it is therefore very important at this point to compare the cells we produce *in vitro* to endogenous cell types. Moreover, comparing progenitors and intermediates in the differentiation process may help to pinpoint where the processes diverge, and how we can improve them. Some divergences may originate from previously underappreciated differences between human pancreas development and those model organ vertebrates such as mouse, which are much easier to study.

The pancreas is both a digestive and an endocrine organ. The digestive function is ensured by the acinar cells that secrete digestive enzymes into the pancreatic ducts. The ductal cells also participate in the process, notably by neutralizing stomach acidity. Pancreatic endocrine cells are clustered into islets of Langerhans that are composed of five different types of endocrine cells, α, β, δ, ε and PP, secreting glucagon, insulin, somatostatin, ghrelin and pancreatic polypeptide, respectively.

Pancreas development begins with the invagination of the foregut into dorsal and ventral buds at embryonic day 8 in the mouse and at around 4 Weeks of Development (4WD) in human (Jennings et al., 2013; Larsen and Grapin-Botton, 2017). In both species, pancreatic buds contain multipotent progenitors characterized by the expression of several transcription factors, such as *PDX1*, *NKX6-1* and *SOX9* (Jonsson et al., 1994; Stoffers et al., 1997; Piper et al., 2004; Seymour et al., 2007; Jennings et al., 2013; Cebola et al., 2015). They proliferate and differentiate into all pancreatic lineages (acinar, ductal and endocrine). In the mouse, proliferation is dependent on signals from the mesenchyme and also from cell to cell interactions, notably via the NOTCH pathway, which activates the transcription factor HES1 (Bhushan et al., 2001; Pan and Wright, 2011; Jensen et al., 2000). The function of the NOTCH pathway seems conserved in human (Jeon et al., 2009; Zhu et al., 2016; Jennings et al., 2017).

In mice, endocrine differentiation occurs from multipotent or bipotent endocrine-ductal progenitors and is marked by the expression of the transcriptional factor NEUROG3 (Solar et al., 2009). Many of these mechanisms appear conserved in human, though we know little about the existence of multipotent versus bipotent progenitors (Zhu et al., 2016). Pancreatic endocrine cell differentiation starts at embryonic day 9 in the mouse and at 8WD in human with the expression of the transcription factor NEUROG3 (Gu et al., 2002; Jennings et al., 2013; Salisbury et al., 2014). *NEUROG3* deficiency leads to an important reduction or absence of pancreatic endocrine cell development, both in mouse, human, and in models of human embryonic stem cell (hESC) differentiation towards endocrine cells (Gradwohl et al., 2000; Rubio-Cabezas et al., 2011; McGrath et al., 2015; Zhu et al., 2016).

There are many similarities, but also differences in pancreatic development between rodent and human. While pancreatic endocrine cell development occurs in two waves in rodents, a single wave of endocrine cell differentiation was described in humans (Pictet et al., 1972; Jennings et al., 2013; Salisbury et al., 2014). Another example is represented by the transcription factor NKX2-2, which is expressed in rodents by early pancreatic progenitors upstream of NEUROG3, while its onset is downstream of NEUROG3 in humans (Jennings et al., 2013). Many genes acting downstream of NEUROG3, some of which are direct targets, have been identified in the mouse (Dassaye et al., 2016). Some control endocrine differentiation in all endocrine cell types, while others are specific to one or several subtypes. Important endocrine genes

are also expressed in the human fetal pancreas, including *NEUROD1, NKX2-2, PAX4, PAX6* and *ISL1* (Lyttle et al., 2008; Jeon et al., 2009). Their sequence of activation and their function have been studied in stem cell models of pancreatic differentiation (Liu et al., 2014; Zhu et al., 2016; Petersen et al., 2017). The degree of conservation of gene function and developmental mechanisms between human and other vertebrates require more extensive investigation.

In the field of hematopoiesis, the identification of cell surface markers using flow cytometry has been instrumental in understanding lineage relationships and differentiation mechanisms (Eaves, 2015), an approach we initiated to study human fetal pancreatic cell differentiation. Using a combination of the cell surface markers GP2, ECAD, CD142 and SUSD2 on human fetal pancreatic epithelial cells, we have previously identified distinct endocrine populations at different stages of their development (Ramond et al., 2017). In the present study, we provide an in-depth transcriptional profiling of these populations using RNA sequencing. We additionally improve the purity of the sorted populations by expanding the sorting strategy to include an additional surface marker, CD133, to exclude ductal cells, along with using granularity to distinguish endocrine cells. Using the improved sorting scheme, we isolate highly pure populations from human fetal pancreata and study their composition at single-cell level using qPCR. Our analysis focuses on 9WD pancreata and shows that at a single time point, different steps of the endocrine differentiation path can be captured. Our findings furthermore enable us to benchmark the pancreatic cell types produced *in vitro* from hPSCs using a well-established endocrine-biased differentiation protocol (Rezania et al., 2014).

Results

RNA sequencing profiling of human fetal pancreatic populations

We have previously reported that flow cytometry with specific cell surface markers can be used to purify distinct populations that sequentially differentiate during human fetal pancreas development. Following gating on epithelial cells ($CD45^-CD31^-EPCAM^+GP2^-$), four distinct populations were sorted from human fetal pancreata at 9WD: $ECAD^+CD142^+$ (population A), $ECAD^+CD142^-$ (population B), $ECAD^{low}CD142^-SUSD2^+$ (population C) and $ECAD^{low}CD142^-SUSD2^-$ (population D). We found that *NEUROG3* is first detected in population B but culminates in population C. Endocrine markers, especially *INS*, are first detected in population C and later (from 10WD) in population D (Ramond et al., 2017). Here, we used this sorting strategy to isolate similar populations from three individual human fetal pancreata at 9WD and analyzed them by RNA sequencing (Fig. 1A and B; Fig. S1 for details on the gating strategy). Transcripts encoding the markers that were used for cell sorting, namely *CDH1*

(ECAD), *EPCAM*, *F3* (CD142), *GP2*, *PECAM1* (CD31), *PTPRC* (CD45), and *SUSD2* followed the expected expression pattern, validating the experimental procedure (Fig. S2A).

Principal component analysis on the RNAseq from the four populations revealed that populations A and B clustered together, while populations D and C clustered independently (Fig. 1C). Hierarchical clustering on 1,007 differentially expressed genes revealed that the first branching occurred between populations A-B and populations D-C (Fig. 1D). Moreover, eight clusters of genes with similar expression patterns were identified. Clusters I and II mark genes which were highly expressed in at least three of the populations and had a diversity of expression patterns. Cluster III includes genes expressed at high levels, with minor changes in A vs B and C vs D. Clusters IV and VI–contain known endocrine progenitor and endocrine genes, which were expressed at higher levels in C-D, with some increase already in B for cluster IV. Several of these genes are highlighted in Fig. 1D. In contrast, clusters V and VIII, which contained pancreas progenitor markers (HES1 and REST for example), were enriched in A-B and for VIII, the decrease was particularly strong in C (Fig. 1D). Cluster VII was characterized by genes expressed at higher levels in population D compared to the other populations, some of which are known mature endocrine cell markers (Fig.1D, and Table S1). We additionally identified differential expression of several genes associated to different signaling pathways between the populations, such as *EGFR*, *CXCR4* and *RARG* enriched in population A and B (Table S1).

We then performed a gene set enrichment analysis on populations A-D using gene ontology database. Biological processes related to hormone regulation and secretion appeared specifically enriched in populations C and D compared to populations A and B (Fig. S3 and Table S2). Moreover, biological processes such as regulation of insulin secretion, regulation of hormone secretion, regulation of hormone levels, insulin secretion, hormone transport and hormone secretion were further enriched in population D (Fig. S3 and Table S2). Taken together, our data indicate that population C and D contain endocrine cells.

To understand the composition of each cell population with respect to multipotent pancreatic progenitors, endocrine progenitors and endocrine cells, we generated heatmaps. *PDX1*, a factor expressed in human by pancreatic progenitors at early time points and later on in β, δ and duct cells, was equally detected in all 4 populations (Fig. 2A) (Jennings et al., 2015). It was also the case for *NKX6-1,* a marker of early pancreatic progenitors, which gets later restricted to β-cells (Jennings et al., 2013), (Fig. 2A). The high expression of *SOX9* and *ONECUT1* in populations A and B suggested that they contained pancreatic progenitors (Fig. 2A) (Larsen and Grapin-Botton, 2017). With regards to endocrine progenitors, *NEUROG3* was lowly expressed in population A, increased in population B, highest in population C and lower in D. (Fig. 2B). *NEUROG3* targets *FEV*, *ETV1*, *PAX4*, *ARX*, *ACOT7*, *NKX2-2* and *NEUROD1*

followed similar patterns (Fig. 2B). This data thus indicate that population A is enriched in multipotent pancreatic progenitors, while population C is enriched in endocrine progenitors. Population C and D also expressed the highest level of endocrine cell markers *ISL1*, *CHGA*, *MAFB*, *PAX6* and *PCSK1* (Fig. 2C, Fig. 1D cluster IV and VI) as well as the hormones *GCG*, *SST*, *GHRL* and *INS* (Fig. 2D). *PPY* expression levels appeared enriched in population B and D. *INS* levels were 13 times higher in population D compared to population C (Fig. 2D and Fig. S2B). Additionally, a set of β-cell-specific genes (*MAFA*, *PCSK1*, *IAPP, G6PC2, FFAR1, SLC30A8)* was enriched in population D compared to population C, suggesting that endocrine cells in population D were more mature than in population C (Fig. 1D cluster VII, Fig. 2D). We also observed upregulation of genes with unknown function in pancreas development, such as ELAVL4 and EYA2 in population C. Additional examples of novel genes marking each population can be found in Table S1.

CD133 marks ductal cells

We previously showed by single-cell qPCR that population B contained a majority of duct-like cells (78% positive for *CFTR*) and 16% *NEUROG3*$^{+}$ endocrine progenitors, whereas population C was highly enriched in *NEUROG3*$^{+}$ cells (74%) with rare *CFTR*$^{+}$ cells (2%) (Ramond et al., 2017). Our RNAseq data supported this claim (Fig. S2B) but also showed that compared to the other populations, population D expressed the highest level of INS, with CFTR levels similar to population A. Therefore, we searched for additional markers to deplete *CFTR*$^{+}$ duct-like cells from the populations B and D. In the pancreas, ductal cells express the transmembrane protein CD133 (prominin 1) (Sugiyama et al., 2007, Lardon et al., 2008). FACS analyses indicated that the C population was CD133^{-}, which correlates with the low frequency of duct-like cells in this population (Fig. 3A-B), whereas populations B and D contained cells with different levels of CD133 (Fig. 3A-B). The frequency of CD133^{+} cells was highest in B compared to D population (72±6% vs 31±18%) (Fig. 3B)**.** We next sorted cells from populations B-D based on their CD133 levels. We performed RT-qPCR for the expression of *CFTR,* marking ductal cells, and *CHGA*, *NEUROG3 and NKX2-2*, marking endocrine cells. We found that the CD133^{-} populations expressed lower levels of *CFTR* and were enriched in *CHGA, NEUROG3 and NKX2-2* (Fig. 3C). We concluded that in the human fetal pancreas, CD133 marks duct-like cells and can be used for their depletion from endocrine cell-containing populations.

Granularity parameter isolates fetal endocrine cells in population D CD133^{-}

Pancreatic endocrine cells contain hormone-rich secretory granules in contrast to other pancreatic cell types. Cell granularity can be assessed by flow cytometry in the side scatter (SSC) parameter. This method has previously been used to enrich for endocrine cells from adult pancreas (Pipeleers, 1987; Rui et al., 2017). We therefore asked if granularity could be used to further enrich for hormone-positive cells. In the human fetal pancreas at 10WD, population C (which is CD133^{-}) contained a low frequency (15±7%) of SSC^{hi} cells while population D depleted from duct cells (CD133^{-}) contained a higher frequency (33±14%) of SSC^{hi} cells (Fig. 4A) (Percentages are the mean from three independent pancreata at 10WD)**.** The frequency of SSC^{hi} cells in population D was age-dependent as SSC^{hi} cells first appeared at 9WD and their frequency increased thereafter (Fig. 4B). We sorted SSC^{hi} and SSC^{low} cells from population D$_{CD133-}$ (named as D$_{HI}$ and D$_{LO}$, respectively) and performed RT-qPCR for the expression of endocrine genes (*CHGA*, *INS*, *NEUROD1* and *PAX6*). RT-qPCR results demonstrated that endocrine markers were enriched in population D$_{HI}$ (Fig. 4C). We concluded that as in the adult, endocrine cells in the human fetal pancreas were the most granular.

Single-cell profiling reveals heterogeneity in the sorted human fetal pancreatic populations

To address the heterogeneity within populations B$_{CD133-}$, C, D$_{LO}$ and D$_{HI}$ described earlier, we analyzed sorted cells by single-cell qPCR for the expression of 91 genes (see methods for details and Table S3 for information on primers). Unsupervised hierarchical clustering of the data showed that the major differences in gene expression were found between cells from population B$_{CD133-}$ (*Cluster I*) versus cells from populations C and D$_{HI}$ (*Clusters III-V*) (Fig. 5A and Fig. S4A). Associated to cluster I, we found a mixed population consisting of cells that generally had low levels of gene expression, most likely for technical reasons (*Cluster II*). Most cells of cluster I expressed *ECAD* and key pancreatic progenitor markers such as *ONECUT1, SOX9, MNX1* and *NKX6-1*, whereas *HES1* and *JAG1* were more heterogeneously expressed (Fig. 5A and Fig. S5). Furthermore, a few cells had initiated *NEUROG3* expression (Fig. 5A-B). The YAP/TAZ target gene *CTGF* also displayed a heterogeneous profile, activated in around half of the cells (Fig. S4A and Fig. S5). Though the cells were sorted based on their lack of the surface marker CD142, a subpopulation of the cells expressed the transcript encoding CD142 (Fig. 5A).

The second group of clusters III-V comprised endocrine progenitors (C) and early endocrine cells (D_{HI}). All clusters shared the expression of *NEUROG3, RFX6, CHGA, NEUROD1, NKX2-2, INSM1* and *SUSD2*, with some noise, as commonly seen for low abundance genes when starting from small amounts of RNA. Cells from *cluster III* appear to be early progenitors which do not express *PCSK2* or *MAFB* transcripts and largely no hormone expression (except for *GHRL* in a subpopulation) (Fig. 5A-B). Furthermore, *PDX1, NKX6-1* and *MNX1* transcripts were downregulated compared to pancreatic progenitors. Cluster III was intrinsically heterogeneous, with at least 4 subpopulations that may represent different levels of maturity or distinct subtypes. These included a group of cells likely representing a precursor to α-cells, expressing *ARX, PCSK1, ISL1* and *ETV1*, while another group expressed robust *NKX6-1*, in addition to *PCSK1, PAX6, ETV1* and *PAX4*. Cluster V likely contains more mature endocrine progenitor cells with low levels of *PCSK1* and *PCSK2*, high *CHGA* and *PAX6* levels and heterogeneous hormone expression. Cells from population C were the main contributors to these two clusters. Finally, cluster IV was composed of early endocrine cells (mainly D_{HI} cells) with high levels of endocrine markers, most cells expressing high levels of *INS*. Cells from population D_{LO} did not appear to represent a homogenous population, but was intermingled with the clusters described above (Fig. 5A-B).

Visualizing the data using dimensionality reduction via t-distributed stochastic neighbor embedding (t-SNE, Van Der Maaten and Hinton, 2008) identified similar populations as hierarchical clustering, with population B largely forming a separate cluster from populations C and D_{HI} (Fig. 5B and Fig. S5). The D_{HI} cluster overlapped to some extent with a cluster representing cells of population C, highlighting a high degree of similarity between these populations that are both committed towards endocrine differentiation. T-SNE further confirmed that population D_{LO} is heterogeneous, encompassing cells with profiles similar to B_{CD133-} or D_{HI}. Violin plots highlighted the subpopulation of cells within the B population that had activated *NEUROG3*, and showed that *NEUROD1* was expressed at similar levels in population C and D_{HI}, whereas D_{HI} expressed higher levels of *CHGA*, *PAX6* and *MAFA* in a subset of cells (Fig. S4B).
As the data indicated that cells of different maturation stages were present in the pancreas at 9WD, we next used the algorithm Monocle (Trapnell et al., 2014) to infer a developmental trajectory for the differentiating cells and pinpoint branching points to different cell fates. Through the construction of a minimal spanning tree, cells were linked in a pseudotemporal order with population B as the starting population giving rise to four branches (Fig. 5C and Fig. S6). The pseudotime reflected downregulation of pancreatic progenitor-associated genes (e.g. *SOX9* and *HES1*) and upregulation of endocrine-biased genes (e.g. *NEUROG3*, *CHGA* and *NKX2-2*) (Fig. S6B). The first branch off the main trunk represented polyhormonal endocrine cells (Fig. 6C, Fig. S6A-C), the second branch further diverged into two endocrine populations likely on the

path to β- and α/δ-cells, respectively), and the third branch represented endocrine progenitors (Fig. 6C, Fig. S6A-C). Heat map visualization of temporal gene expression distinguished cells on the β-cell versus α/δ-cell track (Fig. S6D, cells diverging from branching point 1).

Taken together, based on what is known about each marker in mouse pancreas development, the single-cell analysis shows that the FACS strategy enables to discriminate distinct populations of human pancreas progenitors, endocrine progenitors and endocrine cells. In addition, it shows that subpopulations are found within these 3 groups, which likely represent maturation intermediates co-existing at the same time point due to heterochrony of differentiation in individual cells.

Endocrine subtype selection is initiated in endocrine progenitors

We next investigated whether cells differentiating towards different endocrine subtypes expressing different hormones could be captured, and when they diverged. To gain power, we computationally excluded cells of cluster I and II (pancreatic progenitors and low expressors) and performed t-SNE analysis on the remaining endocrine-biased cells (Fig. 6). With this data visualization, the cells distribute into a 3-branch star, each branch ending with different hormonal expression (Fig. 6). One branch contains cells with a gradient of *INS* expression. Another contains cells expressing *GCG*, and the α cell marker *ARX*. Some cells in this branch also express *PPY* and low levels of *INS* and *SST*. The third branch contains cells with the highest level of *SST* in cells at the tip of the branch. To investigate if the co-expression of multiple hormones also was evident at the level of protein expression, we performed immunofluorescence stainings on human fetal pancreas sections at 10WD. We assessed the combination of insulin with glucagon or somatostatin and glucagon with ghrelin, and found rare cells co-expressing each of these three combinations of hormones (Fig. S7). We did not evaluate the expression of PPY.

At the transcriptional level, we furthermore observed co-expression of *NEUROG3* with hormones. In the t-SNE analysis, most endocrine progenitors (C population) express *NEUROG3* transcripts, arguing that endocrine subtype selection is initiated in endocrine progenitors and continues in population D (Fig. 6). The α-cell specifier *ARX* and the β-cell specifier *NKX6-1* are already expressed in subpopulations of C cells in different branches. Even though *PAX4* is commonly described as a β-cell specifier in mouse and human, it is restricted to endocrine progenitors of population C, as is *GHRL* (Fig. 6). *PAX4* and *GHRL* were absent from most cells in population D_{HI}, suggesting them to be endocrine progenitor markers (Fig. 6).

In conclusion, the single-cell gene expression profiling shows the initiation of endocrine subtype specification in endocrine progenitors in human, and reveals unexpected expression patterns for *PAX4* and *GHRL*.

In vitro-produced endocrine cells follow paths similar to those of their *in vivo* counterparts, but do not resolve hormonal subtypes

We have previously shown that ECAD, CD142 and SUSD2 markers also discriminate B and C populations generated *in vitro* from hPSCs (Ramond et al., 2017). To evaluate the degree of similarity between *in vivo* and *in vitro* populations, we furthermore performed single-cell qPCR on *in vitro* differentiated cells. To simulate the biological variability between individual pancreata, we used three different hPSC lines (SA121 hESCs and iPSCs AD2.1 and AD3.1) differentiated until stage 5 day 1 (S5D1) using a protocol adapted from Rezania et al., 2014 (with minor modifications - see methods for details). All three lines efficiently generated NKX6-1^{+} cells and initiated NEUROG3 expression as assessed by flow cytometry, indicating differentiation towards the pancreatic endocrine lineage (Fig. S8). The hPSC-derived cells were then sorted based on ECAD, CD142 and SUSD2 as population B or C and analyzed by single-cell qPCR. Hierarchical clustering of the data showed that cells segregated first by population (B and C) regardless of their *in vivo* or *in vitro* origin, suggesting that similar populations are generated both *in vivo* and *in vitro* (Fig. 7A and Fig. S9). The second clustering distinguished cells from *in vitro* origin from the fetal cells, indicating that there are source-based differences in gene expression. Some of this difference might be driven by lower gene expression levels of cells extracted from the fetal pancreata, however, normalizing gene expression to the housekeeping gene *RPL7* did not significantly alter the clustering (data not shown). Likewise, the four clusters were also generated by t-SNE analysis (Fig. 7B), discriminating populations B and C, and cells generated *in vivo* from those produced *in vitro*. The decreased distance between populations B and C *in vivo* suggests that they are more similar than their *in vitro* counterparts.

In vitro-produced population B was distinct from the *in vivo* reference as they expressed both *RFX6* and *CDX2* (Fig. 7B), which may indicate that they retain a mixed pancreas-duodenum fate or that they are equivalent to earlier pancreas progenitors. Fetal pancreatic progenitors had robust *NKX6-1* and *DLK1* expression, however only a subpopulation of the hPSC-derived cells was *NKX6-1*$^{+}$ and *DLK1*$^{+}$ (Fig. S9). Sporadic expression of *GPM6A, CDKN1A* and *BMP5* was detected among *in vitro* generated cells, whereas these transcripts were absent from fetal cells. Finally, *ZNF453* was robustly expressed in most *in vitro*-derived cells, but only in a small subpopulation *in vivo* (Fig. S9).

To further extend the comparison between *in vivo* and *in vitro*-derived cells and to assess their maturity, we combined the single-cell qPCR datasets of the human fetal cells (populations B_{CD133-}, C and D_{HI}) and the *in vitro*-differentiated hPSC-derived cells (populations B and C) (this study), with a previously published single-cell qPCR dataset covering several stages of *in vitro* differentiation along with endocrine cells from human adult islets analyzed with the same set of primers (Petersen et al., 2017). The hPSC-derived cells from this dataset were generated using the same protocol as in the present study, and were isolated from stage 4 (early), end of stage 5 and early stage 6 (mid) and end of stage 6 and 7 (late) of the differentiation protocol using a NEUROG3-GFP reporter hESC-line. First, we evaluated how the sorted hPSC-derived B and C cells aligned with cells from the additional differentiation stages (Fig. 7C). The B population formed a separate cluster close to the stage 4/early population (Fig. 7C), whereas population C was positioned between cells from the early- and mid-stage. We next visualized the complete combined dataset comprising all *in vivo* and *in vitro* samples (Fig 7D). On the t-SNE plot, the X-axis separated cells from immature pancreas progenitors (left) to most differentiated endocrine (right) in both *in vivo* and *in vitro* conditions. The Y axis separated cells produced *in vitro* (up) from *in vivo* differentiation (down) (Fig 7D). The different maturation stages are largely inferred from known sequences of gene activation during mouse pancreas differentiation. The X axis organizes cells at different stages of differentiation that co-exist at stage 5 in the differentiation protocol and also reflects a temporal progression of differentiation/days of *in vitro* differentiation (Fig. 7C-D and Fig. S10). We validate some of the progression at the protein level, including SOX9-only cells, rare cells co-expressing SOX9 and NEUROG3 (Fig. S11A), a vast majority of NEUROG3 cells expressing NKX2-2 (Fig. S11B and C) and a few cells co-expressing NEUROG3 and hormones (Fig. S12A-C). While hPSC-derived cells initiate the expression of β-cell maturity markers *MAFA* and *IAPP*, they do not reach the expression levels seen in the adult β-cells (Fig. S10). Finally, we found that *in vitro*-produced cells do not resolve hormonal subtypes and continue to co-express multiple hormones at the transcriptional level (Fig. 7C). While a subset of cells generated with this differentiation protocol are also polyhormonal at the protein level, the majority express only insulin or glucagon (Rezania et al. 2014; Petersen et al., 2017). As *PCSK1* was often co-expressed with *INS* and *GCG* transcripts, we speculated that these cells might produce C-peptide and GLP-1, a peptide processed by PCSK1 from the *GCG* transcript. By immunofluorescence staining, we could indeed detect GLP-1, with some cells co-expressing GLP-1 and C-peptide/proinsulin (Fig. S12D).

In conclusion, our single cell data combined with the sorting strategy provide a platform for future studies to further delineate lineage allocation of individual endocrine cell types of the human pancreas.

Discussion

The developmental mechanisms of the mouse and zebrafish pancreas have been characterized in depth. Numerous studies have uncovered markers for different cell types, the lineage hierarchies between cells present at different time points, the role of numerous genes, notably transcription factors and extracellular signals and their regulatory mechanisms (Larsen and Grapin-Botton, 2017). These studies and limited comparisons to human development have shown a large conservation of developmental mechanisms between species, but also revealed notable differences in the ratio between different endocrine cell types, in the precise gene used in a family of transcription factors endowed with similar properties (Flasse et al., 2013; Jennings et al., 2015). The limited information we have in human development most likely leads us to underestimate differences. In our previous work, we identified markers that enable the enrichment of human fetal pancreatic progenitors (ECAD^{+}CD142^{+}/population A), acinar cells (GP2^{hi}) and cells at different steps on the endocrine differentiation path initiating *NEUROG3* expression (ECAD^{+}CD142^{-}/population B), activating downstream genes (ECAD^{low}CD142^{-}SUSD2^{+}/population C) and endocrine cells (ECAD^{low}CD142^{-}SUSD2^{-}/population D) (Ramond et al., 2017). Extending our initial analysis limited to a few transcripts and previous transcriptome analysis of the whole human fetal pancreas, our current transcriptome analysis of these populations enables to assess the nature and purity of the isolated populations, the conservation of stage-specific regulators of pancreas development with other vertebrates and identify potentially interesting new players. It also documents the cell types where genes which variants predispose to monogenic primary or syndromic diabetes (HADH, GCK, HNF1A, HNF1B, INS, ISL1; NEUROD1, NEUROG3, PAX4, PAX6, PCSK1, PDX1, RFX6, SPINK1, UCP2) or to type 2 diabetes (NOTCH2, GCK, PAX4, HNF1B, HNF1A) are expressed (Fuchsberger et al., 2016). Their expression at this stage may contribute to diabetes predisposition.

Population A contains mostly pancreas progenitors but the expression of several acinar cell markers (*CPA1&2, NR5A2, RBPJL*) suggests that this population either contains early acinar cells (notably, there is no mature marker such as Amylase), or corresponds to the early multipotent progenitors observed in mice, a population that has the potential to differentiate into endocrine or acinar cells (Zhou et al., 2007). These are not detected in population B, most of which is initiating endocrine differentiation while retaining pancreas progenitor markers. These two progenitor populations are enriched in receptors for multiple signaling pathways. The presence of *EGFR, ERBB3, FGFR2* and *SPROUTY2* in populations A and B suggests that, as in mouse, the EGF and FGF pathways are used in human pancreas progenitors, possibly for their maintenance and proliferation as it was shown in mice (Miralles et al., 1999; Bhushan et al., 2001; Miettinen et al., 2000). Interestingly, the pathway may be conserved between species, but differences in the ligands that activate the pathway were observed. As an example, FGF10 that activates FGFR2b in the mouse is lowly

expressed in the human pancreatic mesenchyme, where it is replaced by FGF7, another activator of the same pathway (Ramond et al., 2017). The receptors for these pathways may be used as alternative sorting markers, as could be CXCR4, enriched in this population and previously used to sort endoderm. The presence of *RARγ* receptor is also suggestive of retinoic acid signaling. This pathway has been shown in mice to initially promote the formation of the pancreatic primordium, which is taking place at earlier stages than those studied here, and endocrine progenitor formation; though *RARα* is the most highly expressed RAR in the mouse epithelium (Tulachan et al., 2003; Molotkov et al., 2005; Martín et al., 2005). *FZD5* enrichment is in agreement with the observation that Wnt signaling promotes pancreas progenitor expansion in mice (Baumgartner et al., 2014; Afelik et al., 2015). *TCF7L2*, a transcription factor mediating canonical Wnt pathway activity is also enriched in A-B. It is a major susceptibility locus for type 2 diabetes and it is therefore possible that its variant affects pancreas development, in addition to previously reported roles in β-cell function (Liu and Habener, 2010). The progenitors in populations A and B are also enriched in signaling molecules such as *GDNF*, a molecule also secreted by multipotent and bipotent progenitors in mice, which promotes pancreatic innervation, and *PDGFC*, of unknown activity in the mouse pancreas (Muñoz-Bravo et al., 2013). The enrichment of *LEFTY1* is enigmatic, possibly a remnant of left-right asymmetry determination at this late stage or an antagonism to NODAL and has not been reported in mice.

Our profiling also detects numerous transcription factors reported to maintain progenitors in mouse (*SOX9, HNF1B, OC1, HES1, REST*), and previously detected in the human pancreatic bud at 6-7WD (Larsen and Grapin-Botton, 2017). It also suggests new transcription factors active in pancreas progenitors such as *HEYL*, a Notch target possibly acting redundantly to *HES1, GRHL2* and *OVOL2*. With regards to genes involved in morphogenesis, the expression of the RhoGAP *STARD13* in these populations suggests a conservation of its role in tip cell morphogenesis (Petzold et al., 2013). Only few markers were specific to the B population, notably the proneural gene *ATOH8*, which represses the *NEUROG3*-induced differentiation program in mouse (Ejarque et al., 2016). Its absence in population C suggests an expression at the early phase of *NEUROG3* expression or even earlier. Population B also exhibits enrichment in several GO terms linked to cell migration, suggesting that these cells have initiated delamination.

Population C is enriched in *NEUROG3*, and also in many of its targets previously described in rodent models such as *PAX6*, *PAX4*, *NEUROD1*, *NKX2-2*, *ACOT7*, *CELSR3, ARX*, *FEV* and *ETV1*. In agreement with the observations in mouse, *CDKN1C* and *RIPPLY3* upregulation may play a role in the cell cycle block observed in endocrine progenitors (Georgia et al., 2006; Osipovich et al., 2014). Several genes with reported activity in neurons are also upregulated in this population, some of which

have an unknown function in pancreas development, such as *ELAVL4, EYA2* or *FGF12/14*.

The profile of population D shows that many components of the secretory machinery of endocrine cells are activated during development. Moreover, the activation of *ACVR1C* suggests the involvement of activin signaling during the late stages of endocrine differentiation, possibly recapitulating the importance of Acvr1c in modulating insulin secretion in mice (Bertolino et al., 2008).

Taken together, our study reveals a large degree of conservation in the expression of genes involved in mouse pancreas development in human. Our RNAseq data also reveal new genes expressed in specific subpopulations during human pancreatic development that will be useful to dissect endocrinogenesis in this organ.

Single-cell qPCR conducted for *NEUROG3, NEUROD1, NKX2-2* and *CFTR* on populations B and C previously showed that these populations were not pure (Ramond et al., 2017). In the present study, we applied an extra step to the sorting scheme to increase purity of the populations, while we additionally performed a more comprehensive single-cell qPCR analysis. While cells in population B_{CD133-} were largely homogeneous in regard to the expression hallmarks of pancreas progenitors (*SOX9*, *ONECUT1* and *HES1*), we also observed heterogeneity. The YAP/TAZ target gene *CTFG* was expressed in only a subpopulation of the pancreatic progenitors, indicating that Hippo signaling is active in these cells (Zhao et al., 2008). Notably, the Hippo signaling pathway has recently been identified as a key regulator of the expansion of human pancreatic progenitors (Cebola et al., 2015). Our data may therefore suggest that at 9WD, there are discrete populations of pancreatic progenitors in the human pancreas with differential proliferative capacities. We further capture the initiation of *NEUROG3* at different levels among pancreatic progenitors, and in a few cells, the initiation of target genes *NEUROD1*, *PAX6* and *CHGA.*

Based on the pseudotemporal ordering of cells proposed by Monocle combined with knowledge on mouse pancreas development and previous studies of human fetal pancreatic tissue, we can infer a sequence of gene activation during differentiation of endocrine cells in the human pancreas. Population C showed a general downregulation of pancreas progenitor markers while *NEUROG3* and its targets *NEUROD1, RUNX1T1, NKX2-2, INSM1, RFX6, PAX4* and *CHGA* were upregulated and the hormones were initiated at low levels, followed shortly after by *PCSK1, GCK, PCSK2, PAX6* and *MAFB*. The onset of hormone transcripts in *NEUROG3*$^{+}$ endocrine progenitors that we previously observed *in vitro* (Petersen et al., 2017), thus appears to recapitulate development *in vivo*. This was not expected from studying hormones and NEUROG3 protein in mice, but would be worth assessing transcriptionally. It would be interesting to know when these proteins are expressed during human fetal development and to which degree they are co-expressed, but the availability of

antibodies against human proteins and the number of combinations that can be used on the same sample is limiting such studies. However, as a proof-of-concept, we show that we can detect different steps of progression at the protein level, including SOX9-only cells, a few cells co-expressing SOX9 and NEUROG3, a vast majority of NEUROG3 cells expressing NKX2-2 and a few cells co-expressing NEUROG3 and hormones. While NKX2-2 is known to be expressed in human α and β-cells (Riedel et al., 2012), we report here its expression at an earlier stage in endocrine progenitors. The elimination of cells with low granularity from population D enabled the isolation of more mature endocrine progenitors along with what we assume to be the most mature endocrine cells present in the human pancreas at 9WD. Comparison with adult islets shows that the fetal cells need more maturation to reach the adult state. For example, although they activate *MAFA*, which is detected from E13.5 in the mouse pancreas, they express little *IAPP*, a highly-expressed β-cell-specific gene (Nishi et al., 1990; Artner et al., 2010).

Our previous observations in hESC-derived cells showed that *Ghrelin* was activated early in endocrine progenitors, as soon as *NEUROG3* was activated (Petersen et al., 2017). This is now confirmed *in vivo*. *Ghrelin* is later repressed as the cells acquire the expression of other hormonal genes. The function of ghrelin at this early stage is worth further investigations. We also previously reported a similarly early activation of *PAX4*, seemingly in all cells initiating *NEUROG3* expression (Petersen et al., 2017). The early expression of *PAX4* has been reported in mouse to mark differentiated β-cells and to be repressed by *ARX* in α-cells. However, the early stage of *PAX4* expression in human and how it becomes repressed in β-cells that do not express *ARX* questions its function and regulation in human. As we have previously reported (Petersen et al., 2017), we confirmed in this study that hPSC-derived endocrine cells co-express transcripts of several hormones. We can now further conclude that this is also the case for a subset of endocrine cells formed during human fetal pancreas development, although it is much less prevalent. While the fetal α-cells (marked by *ARX* expression) co-express transcripts for glucagon, insulin (at low level) and pancreatic polypeptide, the cells on the β-cell path only express insulin transcripts. The cells expressing *SST*, though not numerous, also co-expressed other hormone transcripts such as *INS* and *GCG*. Though multiple hormonal transcripts are detected above background in all *in vitro*-derived cells, we have previously shown that the majority do not express multiple hormonal peptides (Petersen et al., 2017). Different processing of pro-hormones is likely to be involved, as the levels of *PCSK1* and *2* are variable in these cells and many have detectable GLP-1, an alternative peptide processed from the *GCG* transcript in the presence of PCSK1. It is also possible that the level of transcripts is too low to detect the insulin peptide or they may not be translated (Riedel et al., 2012).

The resource we provide and the discovery of the sequence of molecular events sheds light on how cell types are specified in human, and combined with genome editing in stem cell models of human pancreas development, will be of further use to understand the mechanisms by which an ever-expanding number of genes predispose to diabetes. They also provide a reference to assess the quality of protocols and pancreatic cells produced *in vitro* from hPSCs.

Experimental procedures

Human fetal pancreas tissue

Human fetal pancreata were isolated from surgical abortion done by suction aspiration between 9 to 12 weeks of development post conception in compliance with the French bioethics legislation and the guidelines of our institution (Castaing et al., 2001; Capito et al., 2013; Scharfmann et al., 2014). Approval was obtained from Agence de Biomédecine, the French competent authority along with maternal written consent. Fetal ages are displayed as weeks of development post conception, confirmed by and hand and foot morphology.

Maintenance and differentiation of human pluripotent stem cell lines

A human hESC line (SA121) obtained from Takara and two iPSC lines (SB AD2.1 and SB AD3.1) obtained from the StemBANCC consortium were applied for this study (Heins et al., 2004; van de Bunt et al., 2016). These cell lines have been confirmed to be pluripotent by evaluation of pluripotency marker expression, tri-lineage differentiation and karyotyping and tested negative for mycoplasma contamination. All hPSC lines were cultured in mTeSR1 medium (StemCell Technologies) on hESC-qualified matrigel (Corning) and passaged every 3-4 days or when confluent by dissociating to a single cell suspension using TrypLE select (ThermoFisher). Cells were seeded onto freshly coated Matrigel in mTeSR1 with 5 μM Tiger (Rock inhibitor, Sigma-Aldrich) and medium was replenished daily. Differentiation to the pancreatic lineage was conducted essentially as previously described (Petersen et al., 2017; Ramond et al., 2017) based on a protocol published by Rezania et al. (Rezania et al., 2014) with minor modifications. Briefly, cells were dissociated to single cells using TrypLE select and resuspended in mTeSR1 with 5 μM Tiger. Cells were seeded at a concentration of 0.3-0.35x10^6 cells/cm^2 onto growth-factor reduced Matrigel on CellBIND surfaces (Corning). Cells were incubated for 24h before starting the differentiation. Modifications from the original protocol (Rezania et al., 2014) comprise: MCDB131 medium (Life technologies) was used as basal medium throughout the differentiation in place of BLAR medium. Activin A (Peprotech) was used at 100 ng/ml during stage 1 instead of GDF8. CHIR99201 (Axon Medchem) was used at 3 μM and

0.3 μM for the first and second day of stage 1, respectively, instead of MCX-928. Cells were kept in 2D culture throughout the differentiation. The differentiation efficiency with this protocol was reported in Fig. 1 of (Petersen et al., 2017), and is further documented in Fig. S8. Experiments with stem cell lines were approved by the scientific committee of “Region hovedstaden”.

Flow cytometry

Prior to flow cytometry analysis and/or sorting, fetal pancreata and hPSC-derived cultures were dissociated into single-cell suspensions as described in the following: Fetal pancreata were rinsed with Hanks Balanced Salt Solution (HBSS) (Gibco) to remove contaminating blood cells and gently dissected with forceps under a binocular magnifying lens. Pancreata were then incubated for 5 minutes in collagenase V (0.5mg/ml) (Sigma Aldrich) in HBSS with Ca^{2+} and Mg^{2+}. Cells were rinsed in HBSS and then incubated for 5 minutes in trypsin (0.05%) (Gibco). Finally, cells were rinsed in HBSS supplemented with 20% Fetal calf serum (FCS, Eurobio). Differentiated hPSCs were washed with PBS without Ca^{2+} and Mg^{2+} (Invitrogen) and incubated with TrypLE select for 1-3 minutes.

For staining with cell-surface markers, cells were incubated with antibodies for 20 minutes in FACS medium (HBSS + 2% FCS), then rinsed in FACS medium and re-suspended in FACS medium with Propidium Iodide (1/4000) (Sigma Aldrich). For intra-cellular staining of hPSC-derived cultures, dissociated cells were fixed for 20 minutes in 4% PFA and then rinsed with PBS with 1% BSA (MACS Buffer Miltenyi Biotec). Cells were permeabilized for 30 min at 4°C in PBS with 5% donkey serum (Millipore) and 0.2% Triton X-100 (Sigma) and then incubated with primary antibodies in blocking solution (PBS with 0.1% Triton X-100 and 5% donkey serum) overnight at 4°C, rinsed and re-suspended in PBS with 1% BSA. For secondary antibody staining, cells were further incubated with secondary antibodies in blocking buffer for 1h at RT. Finally, cells were washed twice in PBS with 1% BSA and re-suspended in PBS with 1% BSA for analysis.

The following antibodies were used: anti CD45-PerCP/Cy5.5 (1/20, clone 2D1, Biolegend), anti CD31-PerCP/Cy5.5 (1/20, clone WM59, Biolegend), anti CD235a-PerCP/Cy5.5 (1/20, clone HI264, Biolegend), anti EPCAM-Brillant violet 605 (1/20, clone 9C4, Biolegend), anti ECAD-PE-Cy7 (1/20, clone 67A4, Biolegend), anti CD133-APC (1/20, clone 315-2C11, Biolegend), anti GP2-PE (1/5, clone 3G7H9, MBL), anti SUSD2-VioBrightFITC (1/20, W5C5, Miltenyi Biotec), anti CD142-BV711 (1/20, clone HTF-1, BD Biosciences), anti NKX6-1-PE (1/40, clone R11-560, BD Biosciences), sheep anti NEUROG3 (1/300, R&D Systems), mouse anti NKX2-2 (1/200, F4.5A5, DSHB), C-peptide-Alexa Fluor 647 (1/200, clone U8-424, BD Biosciences), Glucagon-BV421 (1/80, clone U16-850, BD Biosciences). Secondary antibodies were from

Jackson ImmunoResearch: anti-sheep Alexa Fluor 647 (1/500), anti-mouse Alexa Fluor 647 (1/500). For each antibody, optimal dilution was determined by titration. An ARIA III (BD Bioscience) and a BD FACS Fusion were used for cell sorting and a FACS LSRFortessa for analysis (BD Bioscience). Data was analyzed using FlowJo 10.4.1 software. Dead cells were excluded from analyses.

Immunohistochemistry

Human fetal pancreatic sections (4–5µm thick) were prepared and processed as previously described (Castaing et al., 2005). The following primary antibodies were used mouse anti-glucagon (1:2,000; Sigma, St. Louis, MO); mouse anti-insulin (1:1,000; Sigma), mouse anti-Ghrelin (1:500; Millipore, clone 1ML-1D7) and rabbit anti-somatostatin (1:500; DAKO). The secondary antibodies were anti-rabbit Alexa Fluor 488 antibodies (1:400, Life Technologies) and anti-mouse Alexa Fluor 594 antibodies (1:400, Jackson ImmunoResearch). The nuclei were stained using the Hoechst 33342 fluorescent stain (0.3 mg/ml, Invitrogen).

For staining of hPSC-derived cultures, cells were washed with PBS and fixed with 4% PFA at RT for 30 min. Cells were then washed with PBS and permeabilized for 10 min with 0.5% Triton-X100 in PBS. Cells were incubated in blocking buffer (0.1M Tris-HCL pH 7.5, 0.15M NaCl, 0.5% TSA Blocking Reagent (Perkin Elmer)) for 30 min followed by overnight incubation at 4°C with primary antibodies diluted in 0.1% Triton X-100 in PBS. The following day cells were washed 3x 5min with PBS and incubated with secondary antibodies in 0.1% Triton X-100 in PBS for 1h at RT. Primary antibodies were: sheep anti-NEUROG3 (1/1000, AF3444, R&D Systems), mouse anti-NKX2-2 (1/200, F4.5A5, DSHB), rabbit anti-SOX9 (1/2000, AB5535, Millipore), rat anti-C-peptide/proinsulin (1/200, GN_ID4, DSHB), mouse anti-glucagon (1/2000, G2654, Sigma), rabbit anti-GLP1 (1/500, ab22625, Abcam).

Bulk RT-qPCR on human fetal pancreatic cell populations

For human fetal pancreas, 50 to 100 cells were sorted in 9µL of RT/pre-amp mix from the One-Step qRT-PCR Kit (Invitrogen). Pre-amplified (20 cycles) cDNAs were obtained according to manufacturer's notice and used for qPCR reaction. The pre-amplified cDNAs were then used for qPCR using the TaqMan protocol. RT, pre-amplification and qPCR were performed using TaqMan primers from Applied Biosystems. The following TaqMan primers were used: *PPIA* (Hs04194521_s1), *CHGA* (Hs00900370_m1), *NEUROD1* (Hs01922995_s1), *INS* (Hs02741908), *PAX6* (Hs01088114_m1), *CFTR* (Hs00357011_m1), *NKX2-2* (Hs00159616_m1) and *NEUROG3* (Hs01875204_s1). RT-qPCR results are presented in arbitrary units (AU) relative to expression of the control gene *PPIA*. qPCR was performed on QuantStudio from ThermoFischer following manufacturer's instructions.

RNA extraction and sequencing of human fetal pancreatic cell populations

RNA sequencing was performed on cells from 3 independent pancreata at 9WD. Cells were FACS sorted into Trizol (Life Technologies 15596–018) according to expression of the surface markers EPCAM, CD45, CD31, GP2, CD142, ECAD and SUSD2 (see Fig. S1 for gating strategy). For each population A, B, C and D, we purified 1228, 2300, 1916 and 5244 cells from pancreas 1; 892, 1600, 993 and 1000 cells from pancreas 2 and, 331, 1021, 433 and 548 cells from pancreas 3. RNA was extracted from each population using Trizol Reagent according to manufacturer's guidelines. RNA was quantified and assessed for quality by Agilent-2100 Bioanalyzer (Agilent, Santa Clara, CA). Library preparation and sequencing was performed at the Oxford Genomics Centre (Wellcome Trust Centre for Human Genetics, University of Oxford). Libraries were prepared using the SMARTer Ultra Low Input HV kit (634820, Clontech). All libraries were multiplexed and sequenced over multiple lanes of Illumina HiSeq4000 as 75-nucleotide paired-end reads to a minimum depth of 20 million read pairs.

Transcript quantification

Raw sequencing reads were aligned to the primary assemblies and 1000 genomes phase 2 decoy sequences for human genome reference GRCh37.p13 with STAR v2.5.1 using GENCODE release 19 as the transcriptome reference (Dobin et al., 2013). Gene level read counts for differential expression analysis were quantified for all protein-coding and long non-coding transcripts present in GENCODE release 19 using featureCounts (Liao et al., 2014). For plotting and filtering purposed, we also normalized the gene counts to transcripts per million (TPM).

Differential expression analysis

Differential expression was performed on all autosomal protein-coding and long non-coding RNA genes. Only genes demonstrating consistent expression (> 1 count per million) in all samples of at least one population were included in the analysis. Gene counts were normalized using voom and genes demonstrating significant expression differences between the populations identifies with the moderated F-test implemented in limma (Ritchie et al., 2015). Significant genes were assigned to the population corresponding to their maximum F-statistic. Results were corrected for multiple testing using the Benjamini-Hochberg procedure, with a false discovery rate < 1% used as the significance threshold.

GO term enrichment

To assess differentially expressed genes of each population for overrepresentation of GO terms, we used the g:Profiler web server with all “Biological Process” (“BP”) terms (Reimand et al., 2016.). GO terms were required to contain a minimum of 3 and maximum of 1000 genes, with a minimum overlap of 3 genes between the GO term and gene sets from each population needed for inclusion. All genes used in the differential expression analysis were used as the background set. P-values were adjusted for multiple testing according to Benjamini-Hochberg.

Single cell qPCR

Live, dissociated cells were stained for and sorted according to expression of the surface markers CD142, ECAD and SUSD2 as described above. Cells derived from human fetal pancreas were additionally stained for EPCAM, CD45, CD31, GP2, CD133 and CD135a to enrich for epithelial cells and exclude endothelial, acinar, ductal and erythrocyte cells, respectively. Single cells were sorted by FACS into individual wells of 96-well PCR plates containing NP40 Detergent Surfact-Amps Solution (Fischer Scientific), SUPERase-In RNase Inhibitor (Ambion), Superscript VILO-cDNA Synthesis reaction mix (Invitrogen) and nuclease-free water. Samples were stored at -80°C until downstream processing. Reverse transcription, specific target amplification and qPCR was performed according to the manufacturer’s instructions (Fluidigm, protocol no. 68000088_G1, appendix B). Briefly, RNA was denatured at 65℃ for 90 seconds, followed by the addition of Superscript VILO^{TM}cDNA Synthesis enzyme mix (Invitrogen) and T4 Gene 32 Protein (NEB) for reverse transcription. cDNA was then pre-amplified for 20 cycles with TaqMan PreAmp Master Mix (Invitrogen) using a pool of all primers, followed by exonuclease I (NEB) treatment. All reactions were performed on a OneStepPlus system (Applied Biosystems). Pre-amplified cDNA was subsequently diluted 5X and mixed with 2X SSo Fast EvaGreen Supermix (Bio-rad laboratories) and DNA Binding Dye Sample Loading Reagent (Fluidigm) before loading onto a primed 96.96 chip (Fluidigm). Primers were mixed with Assay Loading Reagent (Fluidigm) and DNA suspension buffer (Teknova) to a concentration of 5μM. The concentration of each primer was 500nM in the final reaction. qPCR was performed on a BiomarkHD for 30 cycles followed by melt curve generation. Primer sequences can be found in Table S3. A total of 91 primers were used, of which 86 had previously been validated (Petersen et al., 2017).

Initial processing of single-cell qPCR data and quality control

Ct values that produced melting curves beyond the validated temperature range and/or Ct values that were greater than or equal to the Limit of Detection (LOD) Ct value were treated as a non-detectable transcript/non-specific amplicon and were set to LOD Ct (LOD Ct = 22 for this primer set). Cells that expressed the housekeeping genes UBC and RPL7 above the LOD Ct value or at very low or high level (greater than the mean Ct value +/- 3x SD) were excluded from further analysis. Ct values were transformed into Log2Ex values (Log2Ex = LOD – Raw Ct). No-template control (NTC) and control bulk cDNA samples were included on each chip for technical quality assessment. Single-cell cDNA from three individual human fetal pancreata at 9WD and hPSC-derived cells from populations B/B_{CD133-} and C was equally distributed between six 96.96 chips. CDNA from three additional individual human fetal pancreata from populations D_{HI} and D_{LO} was equally distributed among three 96.96 chips along with some samples representing B and C cells to enable data merging. In total 864 samples were analyzed by single-cell qPCR distributed on 9 69.69 chips. Among these, 36 samples were controls that were excluded from downstream computational analysis. Based on the expression level for UBC and RPL7 as described above, 181 cells were excluded. Among these where all D_{HI} and D_{LO} samples from one pancreas which all had low/undetectable gene expression. Consequently, 683 cells were included in the final analysis. A detailed table that shows how many cells were processed from each population and pancreata can be found in Table S3.

Single-cell qPCR data analysis

Single-cell qPCR data was analyzed with a custom script in R. Since the sorted populations were run on different 96.96 chips and in two different batches, chip-chip variance was analyzed using the control bulk cDNA samples included on each chip. Due to low technical variance between chips, data from all 9 chips where combined without normalization. Processed single-cell qPCR data for additional hPSC-differentiation stages and human mature islet cells were obtained from a previously published dataset (Petersen et al., 2017). This data was merged with data from the present study after checking for variation based on expression of housekeeping genes and using ComBat to check for batch effects. The 86 primers in common for the two datasets were used for analysis (see Table S3 for the list of primers). For statistical analysis, each cell was treated as a biological replicate. Error bars indicate SD or SEM as indicated in the figure legends. Pseudotemporal ordering of cells from human fetal pancreas was done using Monocle (‘monocle’ package, R), an unsupervised algorithm with a high temporal resolution that enables to visualize bifurcation points or cell fates in pseudotime during developmental progression (Trapnell et al., 2014). Differential expression test was performed on all cells based on the population (B_{CD133-}, C, D_{HI} and D_{LO}) and the top 50 significant genes were used in the Monocle ordering algorithm.

Competing interests

MVDB, MBKP, NB, MH, and CH are or have been employees of Novo Nordisk A/S and may hold shares in the company.

Data availability

Sequence data for this study has been deposited at the European Genome-phenome Archive (EGA), under accession number EGASXXXXXXX.

Funding

This work was supported by the HumEn project funded by the European Commission's Seventh Framework Programme for Research, (agreement No 602587) (RS and AGB), the Foundation Bettencourt Schueller (RS) and the Fondation Francophone pour la Recherche sur le Diabete (FFRD) (RS). The RS laboratory belongs to the Laboratoire d'Excellence consortium Revive and to the Departement Hospitalo-Universitaire (DHU) Autoimmune and Hormonal disease. M, MIMC, CH and RS has also received support from the Innovative Medicines Initiative Joint Undertaking under grant agreement n°115439, resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and EFPIA companies' in kind contribution. This publication reflects only the authors' views and neither the IMI JU nor EFPIA nor the European Commission are liable for any use that may be made of the information contained therein. BSBT is supported by the Copenhagen Bioscience PhD program financed by the Novo Nordisk Foundation (grant number NNF16CC0020994). AA is supported by the Danish National Research Foundation Grant DNRF 116 to the StemPhys project. AGB´s lab received funding from the Novo Nordisk Foundation (grant number NNF17CC0027852). MVDB was supported by a Novo Nordisk postdoctoral fellowship run in partnership with the University of Oxford. ALG is a Wellcome Trust Senior Fellow in Basic Biomedical Science (0951010, 200837).

References

Afelik, S., Pool, B., Schmerr, M., Penton, C. and Jensen, J. (2015). Wnt7b is required for epithelial progenitor growth and operates during epithelial-to-mesenchymal signaling in pancreatic development. *Dev Biol* **399**, 204-217.

Ameri, J., Borup, R., Prawiro, C., Ramond, C., Schachter, K. A., Scharfmann, R. and Semb, H. (2017). Efficient Generation of Glucose-Responsive Beta Cells from Isolated GP2(+) Human Pancreatic Progenitors. *Cell Rep* **19**, 36-49.

Artner, I., Hang, Y., Mazur, M., Yamamoto, T., Guo, M., Lindner, J., Magnuson, M. A. and Stein, R. (2010). MafA and MafB regulate genes critical to beta-cells in a unique temporal manner. *Diabetes* **59**, 2530-2539.

Baumgartner, B. K., Cash, G., Hansen, H., Ostler, S. and Murtaugh, L. C. (2014). Distinct requirements for beta-catenin in pancreatic epithelial growth and patterning. *Dev Biol* **391**, 89-98.

Bertolino, P., Holmberg, R., Reissmann, E., Andersson, O., Berggren, P. O. and Ibáñez, C. F. (2008). Activin B receptor ALK7 is a negative regulator of pancreatic beta-cell function. *Proc Natl Acad Sci U S A* **105**, 7246-7251.

Bhushan, A., Itoh, N., Kato, S., Thiery, J. P., Czernichow, P., Bellusci, S. and Scharfmann, R. (2001). Fgf10 is essential for maintaining the proliferative capacity of epithelial progenitor cells during early pancreatic organogenesis. *Development* **128**, 5109-5117.

Capito, C., Simon, M. T., Aiello, V., Clark, A., Aigrain, Y., Ravassard, P. and Scharfmann, R. (2013). Mouse muscle as an ectopic permissive site for human pancreatic development. *Diabetes* **62**, 3479-3487.

Castaing, M., Peault, B., Basmaciogullari, A., Casal, I., Czernichow, P. and Scharfmann, R. (2001). Blood glucose normalization upon transplantation of human embryonic pancreas into beta-cell-deficient SCID mice. *Diabetologia* **44**, 2066-2076.

Castaing, M., Guerci, A., Mallet, J., Czernichow, P., Ravassard, P. and Scharfmann, R. (2005). Efficient restricted gene expression in beta cells by lentivirus-mediated gene transfer into pancreatic stem/progenitor cells. *Diabetologia* **48**, 709-719.

Cebola, I., Rodríguez-Seguí, S. A., Cho, C. H., Bessa, J., Rovira, M., Luengo, M., Chhatriwala, M., Berry, A., Ponsa-Cobas, J., Maestro, M. A. et al. (2015). TEAD and YAP regulate the enhancer network of human embryonic pancreatic progenitors. *Nat Cell Biol* **17**, 615-626.

Cogger, K. F., Sinha, A., Sarangi, F., McGaugh, E. C., Saunders, D., Dorrell, C., Mejia-Guerrero, S., Aghazadeh, Y., Rourke, J. L., Screaton, R. A. et al. (2017). Glycoprotein 2 is a specific cell surface marker of human pancreatic progenitors. *Nat Commun* **8**, 331.

D'Amour, K. A., Bang, A. G., Eliazer, S., Kelly, O. G., Agulnick, A. D., Smart, N. G., Moorman, M. A., Kroon, E., Carpenter, M. K. and Baetge, E. E. (2006).

Production of pancreatic hormone-expressing endocrine cells from human embryonic stem cells. *Nat Biotechnol* **24**, 1392-1401.

Dassaye, R., Naidoo, S. and Cerf, M. E. (2016). Transcription factor regulation of pancreatic organogenesis, differentiation and maturation. *Islets* **8**, 13-34.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. and Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21.

Eaves, C. J. (2015). Hematopoietic stem cells: concepts, definitions, and the new reality. *Blood* **125**, 2605-2613.

Ejarque, M., Mir-Coll, J., Gomis, R., German, M. S., Lynn, F. C. and Gasa, R. (2016). Generation of a Conditional Allele of the Transcription Factor Atonal Homolog 8 (Atoh8). *PLoS One* **11**, e0146273.

Esni, F., Ghosh, B., Biankin, A. V., Lin, J. W., Albert, M. A., Yu, X., MacDonald, R. J., Civin, C. I., Real, F. X., Pack, M. A. et al. (2004). Notch inhibits Ptf1 function and acinar cell differentiation in developing mouse and zebrafish pancreas. *Development* **131**, 4213-4224.

Flasse, L. C., Pirson, J. L., Stern, D. G., Von Berg, V., Manfroid, I., Peers, B. and Voz, M. L. (2013). Ascl1b and Neurod1, instead of Neurog3, control pancreatic endocrine cell fate in zebrafish. *BMC Biol* **11**, 78.

Fuchsberger, C., Flannick, J., Teslovich, T.M., Mahajan, A., Agarwala, V., Gaulton, K.J., Ma, C., Fontanillas, P., Moutsianas, L., McCarthy, D.J. et al. (2016) The genetic architecture of type 2 diabetes. *Nature* **536**, 41-47.

Georgia, S., Soliz, R., Li, M., Zhang, P. and Bhushan, A. (2006). p57 and Hes1 coordinate cell cycle exit with self-renewal of pancreatic progenitors. *Dev Biol* **298**, 22-31.

Gradwohl, G., Dierich, A., LeMeur, M. and Guillemot, F. (2000). neurogenin3 is required for the development of the four endocrine cell lineages of the pancreas. *Proc Natl Acad Sci U S A* **97**, 1607-1611.

Gu, G., Dubauskaite, J. and Melton, D. A. (2002). Direct evidence for the pancreatic lineage: NGN3+ cells are islet progenitors and are distinct from duct progenitors. *Development* **129**, 2447-2457.

Heins, N., Englund, M. C., Sjöblom, C., Dahl, U., Tonning, A., Bergh, C., Lindahl, A., Hanson, C. and Semb, H. (2004). Derivation, characterization, and differentiation of human embryonic stem cells. *Stem Cells* **22**, 367-376.

Jennings, R. E., Berry, A. A., Gerrard, D. T., Wearne, S. J., Strutt, J., Withey, S., Chhatriwala, M., Piper Hanley, K., Vallier, L., Bobola, N. et al. (2017). Laser Capture and Deep Sequencing Reveals the Transcriptomic Programmes Regulating the Onset of Pancreas and Liver Differentiation in Human Embryos. *Stem Cell Reports* **9**, 1387-1394.

Jennings, R. E., Berry, A. A., Kirkwood-Wilson, R., Roberts, N. A., Hearn, T.,

Salisbury, R. J., Blaylock, J., Piper Hanley, K. and Hanley, N. A. (2013). Development of the human pancreas from foregut to endocrine commitment. *Diabetes* **62**, 3514-3522.

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

Jensen, J., Pedersen, E. E., Galante, P., Hald, J., Heller, R. S., Ishibashi, M., Kageyama, R., Guillemot, F., Serup, P. and Madsen, O. D. (2000). Control of endodermal endocrine development by Hes-1. *Nat Genet* **24**, 36-44.

Jeon, J., Correa-Medina, M., Ricordi, C., Edlund, H. and Diez, J. A. (2009). Endocrine cell clustering during human pancreas development. *J Histochem Cytochem* **57**, 811-824.

Johnson, J. D. (2016). The quest to make fully functional human pancreatic beta cells from embryonic stem cells: climbing a mountain in the clouds. *Diabetologia* **59**, 2047-2057.

Jonsson, J., Carlsson, L., Edlund, T. and Edlund, H. (1994). Insulin-promoter-factor 1 is required for pancreas development in mice. *Nature* **371**, 606-609.

Kim, Y. H., Larsen, H. L., Rué, P., Lemaire, L. A., Ferrer, J. and Grapin-Botton, A. (2015). Cell cycle-dependent differentiation dynamics balances growth and endocrine differentiation in the pancreas. *PLoS Biol* **13**, e1002111.

Lardon, J., Corbeil, D., Huttner, W. B., Ling, Z. and Bouwens, L. (2008). Stem cell marker prominin-1/AC133 is expressed in duct cells of the adult human pancreas. *Pancreas* **36**, e1-6.

Larsen, H. L. and Grapin-Botton, A. (2017). The molecular and morphogenetic basis of pancreas organogenesis. *Semin Cell Dev Biol*

Larsen, H. L., Martín-Coll, L., Nielsen, A. V., Wright, C. V. E., Trusina, A., Kim, Y. H. and Grapin-Botton, A. (2017). Stochastic priming and spatial cues orchestrate heterogeneous clonal contribution to mouse pancreas organogenesis. *Nat Commun* **8**, 605.

Liao, Y., Smyth, G. K. and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923-930.

Liu, H., Yang, H., Zhu, D., Sui, X., Li, J., Liang, Z., Xu, L., Chen, Z., Yao, A., Zhang, L. et al. (2014). Systematically labeling developmental stage-specific genes for the study of pancreatic β-cell differentiation from human embryonic stem cells. *Cell Res* **24**, 1181-1200.

Liu, Z. and Habener, J. F. (2010). Wnt signaling in pancreatic islets. *Adv Exp Med Biol* **654**, 391-419.

Lyttle, B. M., Li, J., Krishnamurthy, M., Fellows, F., Wheeler, M. B., Goodyer, C. G. and Wang, R. (2008). Transcription factor expression in the developing human fetal endocrine pancreas. *Diabetologia* **51**, 1169-1180.

Martín, M., Gallego-Llamas, J., Ribes, V., Kedinger, M., Niederreither, K.,

Chambon, P., Dollé, P. and Gradwohl, G. (2005). Dorsal pancreas agenesis in retinoic acid-deficient Raldh2 mutant mice. *Dev Biol* **284**, 399-411.

McGrath, P. S., Watson, C. L., Ingram, C., Helmrath, M. A. and Wells, J. M. (2015). The Basic Helix-Loop-Helix Transcription Factor NEUROG3 Is Required for Development of the Human Endocrine Pancreas. *Diabetes* **64**, 2497-2505.

Miettinen, P.J., Huotari, M., Koivisto, T., Ustinov, J., Palgi, J., Rasilainen, S., Lehtonen, E., Keski-Oja, J. and Otonkoski, T. (2000). Impaired migration and delayed differentiation of pancreatic islet cells in mice lacking EGF-receptors. *Development* **127**, 2617-2627.

Miralles, F., Czernichow, P., Ozaki, K., Itoh, N. and Scharfmann, R. (1999). Signaling through fibroblast growth factor receptor 2b plays a key role in the development of the exocrine pancreas. *Proc Natl Acad Sci U S A* **96**, 6267-6272.

Molotkov, A., Molotkova, N. and Duester, G. (2005). Retinoic acid generated by Raldh2 in mesoderm is required for mouse dorsal endodermal pancreas development. *Dev Dyn* **232**, 950-957.

Muñoz-Bravo, J. L., Hidalgo-Figueroa, M., Pascual, A., López-Barneo, J., Leal-Cerro, A. and Cano, D. A. (2013). GDNF is required for neural colonization of the pancreas. *Development* **140**, 3669-3679.

Nishi, M., Sanke, T., Nagamatsu, S., Bell, G. I. and Steiner, D. F. (1990). Islet amyloid polypeptide. A new beta cell secretory product related to islet amyloid deposits. *J Biol Chem* **265**, 4173-4176.

Nostro, M. C., Sarangi, F., Yang, C., Holland, A., Elefanty, A. G., Stanley, E. G., Greiner, D. L. and Keller, G. (2015). Efficient generation of NKX6-1+ pancreatic progenitors from multiple human pluripotent stem cell lines. *Stem Cell Reports* **4**, 591-604.

Osipovich, A. B., Long, Q., Manduchi, E., Gangula, R., Hipkens, S. B., Schneider, J., Okubo, T., Stoeckert, C. J., Takada, S. and Magnuson, M. A. (2014). Insm1 promotes endocrine cell differentiation by modulating the expression of a network of genes that includes Neurog3 and Ripply3. *Development* **141**, 2939-2949.

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

Pan, F. C. and Wright, C. (2011). Pancreas organogenesis: from bud to plexus to gland. *Dev Dyn* **240**, 530-565.

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

Petzold, K. M., Naumann, H. and Spagnoli, F. M. (2013). Rho signalling restriction by the RhoGAP Stard13 integrates growth and morphogenesis in the pancreas. *Development* **140**, 126-135.

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.

Pipeleers, D. (1987). The biosociology of pancreatic B cells. *Diabetologia* **30**, 277-291.

Piper, K., Brickwood, S., Turnpenny, L. W., Cameron, I. T., Ball, S. G., Wilson, D. I. and Hanley, N. A. (2004). Beta cell differentiation during early human pancreas development. *J Endocrinol* **181**, 11-23.

Ramond, C., Glaser, N., Berthault, C., Ameri, J., Kirkegaard, J. S., Hansson, M., Honoré, C., Semb, H. and Scharfmann, R. (2017). Reconstructing human pancreatic differentiation by mapping specific cell populations during development. *Elife* **6**,

Reimand, J., Arak, T., Adler, P., Kolberg, L., Reisberg, S., Peterson, H. and Vilo, J. (2016). g:Profiler-a web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Res* **44**, W83-9.

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.

Riedel, M. J., Asadi, A., Wang, R., Ao, Z., Warnock, G. L. and Kieffer, T. J. (2012). Immunohistochemical characterisation of cells co-producing insulin and glucagon in the developing human pancreas. *Diabetologia* **55**, 372-381.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W. and Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**, e47.

Rubio-Cabezas, O., Jensen, J. N., Hodgson, M. I., Codner, E., Ellard, S., Serup, P. and Hattersley, A. T. (2011). Permanent Neonatal Diabetes and Enteric Anendocrinosis Associated With Biallelic Mutations in NEUROG3. *Diabetes* **60**, 1349-1353.

Rui, J., Deng, S., Arazi, A., Perdigoto, A. L., Liu, Z. and Herold, K. C. (2017). β Cells that Resist Immunological Attack Develop during Progression of Autoimmune Diabetes in NOD Mice. *Cell Metab* **25**, 727-738.

Russ, H. A., Parent, A. V., Ringler, J. J., Hennings, T. G., Nair, G. G., Shveygert, M., Guo, T., Puri, S., Haataja, L., Cirulli, V. et al. (2015). Controlled induction of human pancreatic progenitors produces functional beta-like cells in vitro. *EMBO J* **34**, 1759-1772.

Salisbury, R. J., Blaylock, J., Berry, A. A., Jennings, R. E., De Krijger, R., Piper Hanley, K. and Hanley, N. A. (2014). The window period of NEUROGENIN3 during human gestation. *Islets* **6**, e954436.

Scharfmann, R., Pechberty, S., Hazhouz, Y., von Bülow, M., Bricout-Neveu, E., Grenier-Godard, M., Guez, F., Rachdi, L., Lohmann, M., Czernichow, P. et al. (2014). Development of a conditionally immortalized human pancreatic β cell line. *J Clin Invest* **124**, 2087-2098.

Seymour, P. A., Freude, K. K., Tran, M. N., Mayes, E. E., Jensen, J., Kist, R., Scherer, G. and Sander, M. (2007). SOX9 is required for maintenance of the pancreatic progenitor cell pool. *Proc Natl Acad Sci U S A* **104**, 1865-1870.

Solar, M., Cardalda, C., Houbracken, I., Martin, M., Maestro, M. A., De Medts, N., Xu, X., Grau, V., Heimberg, H., Bouwens, L. et al. (2009). Pancreatic exocrine duct cells give rise to insulin-producing beta cells during embryogenesis but not after birth. *Dev Cell* **17**, 849-860.

Stoffers, D. A., Zinkin, N. T., Stanojevic, V., Clarke, W. L. and Habener, J. F. (1997). Pancreatic agenesis attributable to a single nucleotide deletion in the human IPF1 gene coding sequence. *Nat Genet* **15**, 106-110.

Sugiyama, T., Rodriguez, R. T., McLean, G. W. and Kim, S. K. (2007). Conserved markers of fetal pancreatic epithelium permit prospective isolation of islet progenitor cells by FACS. *Proc Natl Acad Sci U S A* **104**, 175-180.

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N.J., Livak, K.J., Mikkelsen, T.S. and Rinn, J.L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudo-temporal ordering of single cells." *Nature Biotechnology* **32**, 381-386.

Tulachan, S. S., Doi, R., Kawaguchi, Y., Tsuji, S., Nakajima, S., Masui, T., Koizumi, M., Toyoda, E., Mori, T., Ito, D. et al. (2003). All-trans retinoic acid induces differentiation of ducts and endocrine cells by mesenchymal/epithelial interactions in embryonic pancreas. *Diabetes* **52**, 76-84.

van de Bunt, M., Lako, M., Barrett, A., Gloyn, A. L., Hansson, M., McCarthy, M. I., Beer, N. L. and Honoré, C. (2016). Insights into islet development and biology through characterization of a human iPSC-derived endocrine pancreas model. *Islets* **8**, 83-95.

Van Der Maaten, L.J.P., and Hinton, G.E. (2008). Visualizing high- dimensional data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605

Zhao, B., Ye, X., Yu, J., Li, L., Li, W., Li, S., Yu, J., Lin, J. D., Wang, C. Y., Chinnaiyan, A. M. et al. (2008). TEAD mediates YAP-dependent gene induction and growth control. *Genes Dev* **22**, 1962-1971.

Zhou, Q., Law, A. C., Rajagopal, J., Anderson, W. J., Gray, P. A. and Melton, D. A. (2007). A multipotent progenitor domain guides pancreatic organogenesis. *Dev Cell* **13**, 103-114.

Zhu, Z., Li, Q. V., Lee, K., Rosen, B. P., González, F., Soh, C. L. and Huangfu, D. (2016). Genome Editing of Lineage Determinants in Human Pluripotent Stem Cells Reveals Mechanisms of Pancreatic Development and Diabetes. *Cell Stem Cell* **18**, 755-768.

Figures:

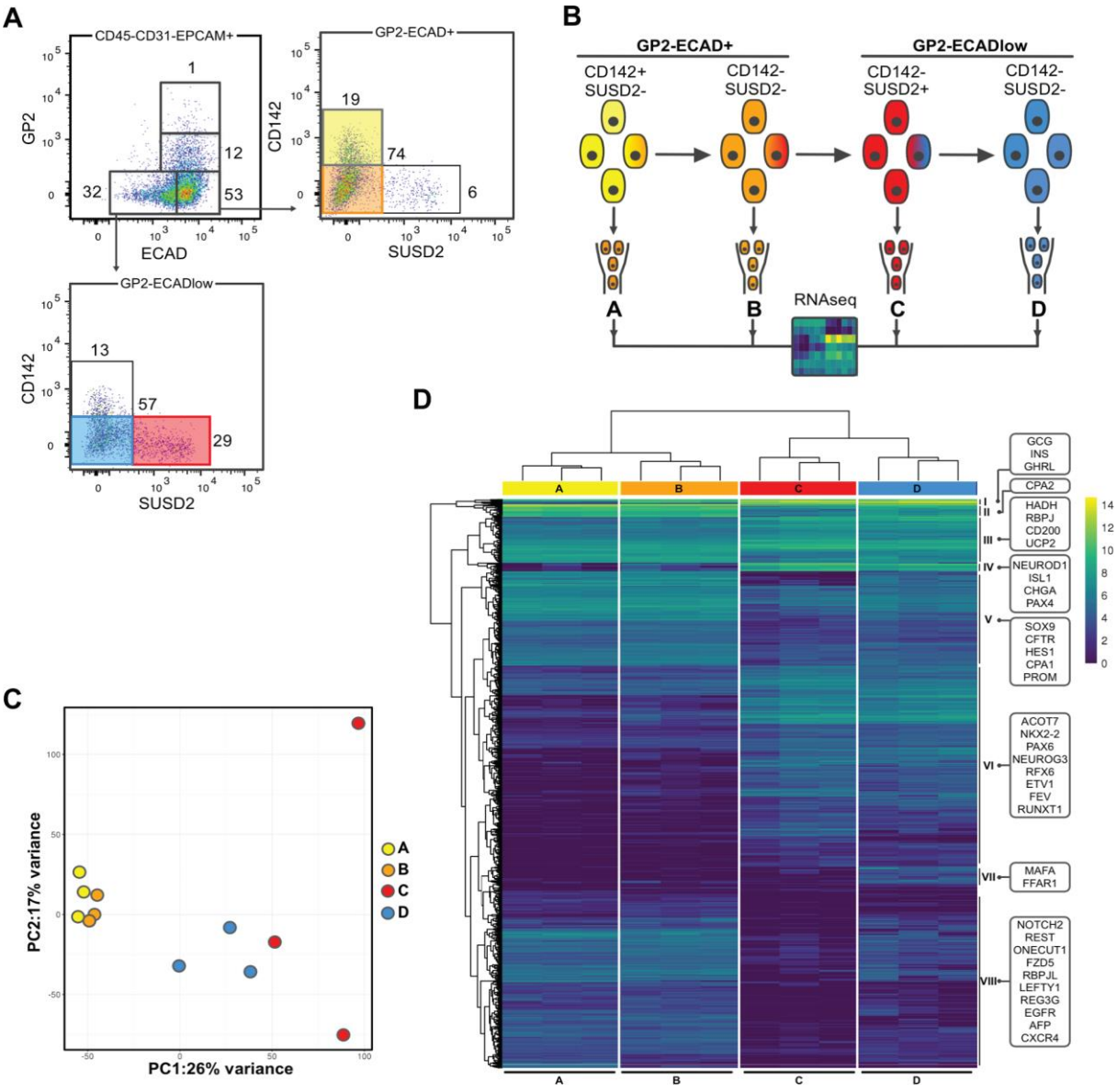


Figure 1: Characterization of sorted-cell populations from human fetal pancreata by RNAseq at 9WD (**A**) Flow cytometry analysis of the expression of GP2 and ECAD on CD45$^-$CD31$^-$EPCAM$^+$ cells at 9WD. CD142 and SUSD2 expression was analyzed in GP2$^-$ECAD$^+$ and GP2$^-$ECAD$^-$. FACS plots are representative of 3 independent pancreata at 9WD (**B**) Scheme of the experimental setup: GP2$^-$ECAD$^+$CD142$^+$SUSD2$^-$ (population A), GP2$^-$ECAD$^+$CD142$^-$SUSD2$^-$ (population B), GP2$^-$ECAD^{low}CD142$^-$SUSD2$^+$ (population C), and GP2$^-$ECAD^{low}CD142$^-$SUSD2$^-$ (population D) were cell sorted and analyzed by RNAseq. All populations were CD45$^-$CD31$^-$EPCAM$^+$ and derived from three independent pancreata at 9WD. The population color code is re-used throughout the article. (**C**) PCA map on the RNAseq from sorted populations (A,

B, C and D) in triplicate at 9WD. (**D**) Heatmap displaying the expression of the 1007 differentially expressed genes in triplicate in populations A, B, C and D in RPKM along with the hierarchical clustering. Clusters are named on the right side and a few gene examples are indicated in boxes. For more extensive examples, Table S1 highlights genes enriched in populations A, B, C or D.

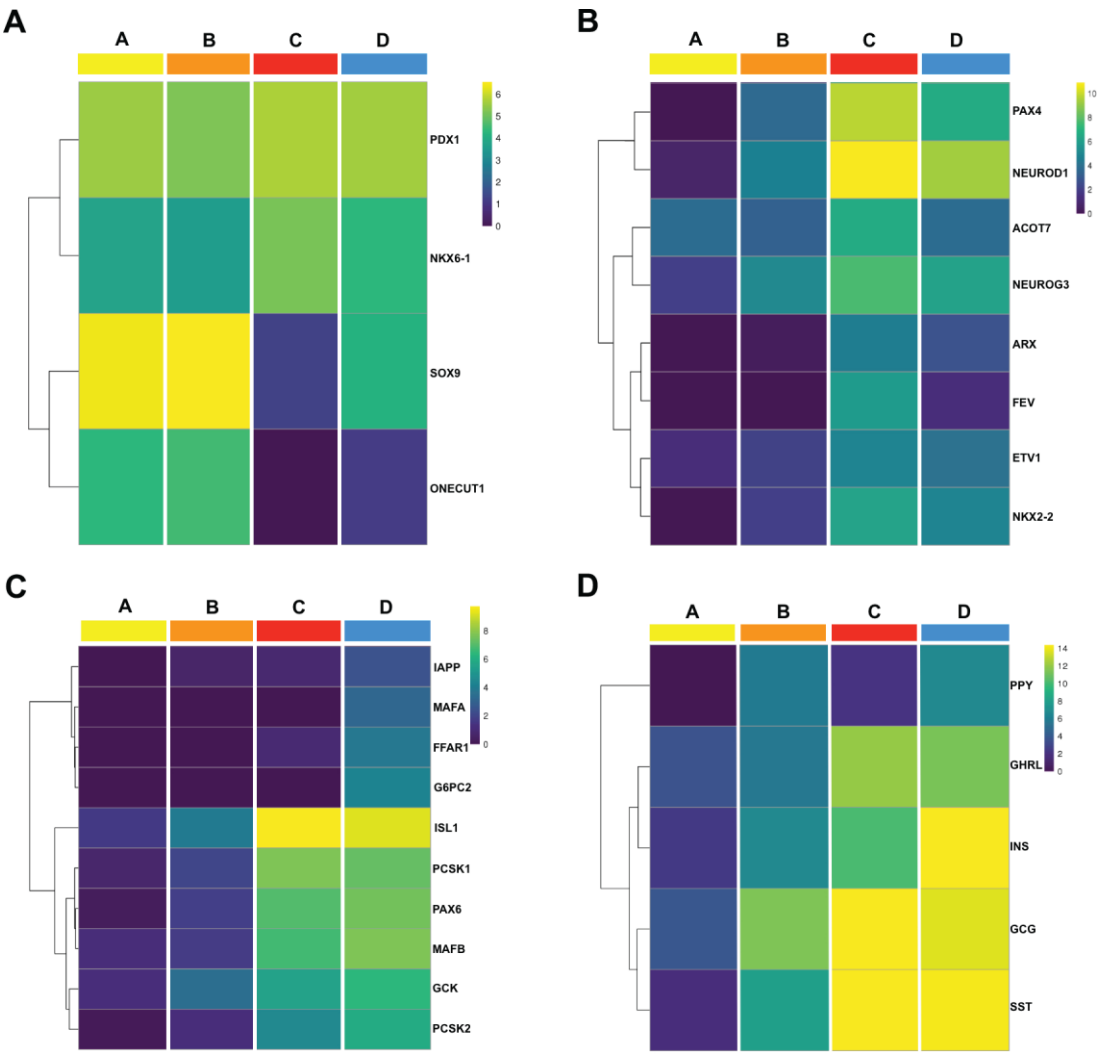


Figure 2: Characterization of populations A, B, C and D at 9WD Heatmaps displaying the expression of (**A**) multipotent genes (*PDX1, NKX6-1*, *SOX9* and *ONECUT1)*; (**B**) progenitor genes and early endocrine markers (*PAX6, PAX4, NEUROD1, ACOT7, NEUROG3, ARX*, *FEV*, *ETV1* and *NKX2-2*) enriched in population C; (**C**) Endocrine genes (*IAPP, MAFA, FFAR1, G6PC2, ISL1, PCSK1, MAFB, GCK* and *PCSK2*) enriched in population D; and (**D**) pancreatic hormones (*PPY*, *GHRL, INS, GCG* and *SST*) in sorted population A, B, C and D. Heatmaps are displayed using a log2 expression scale.

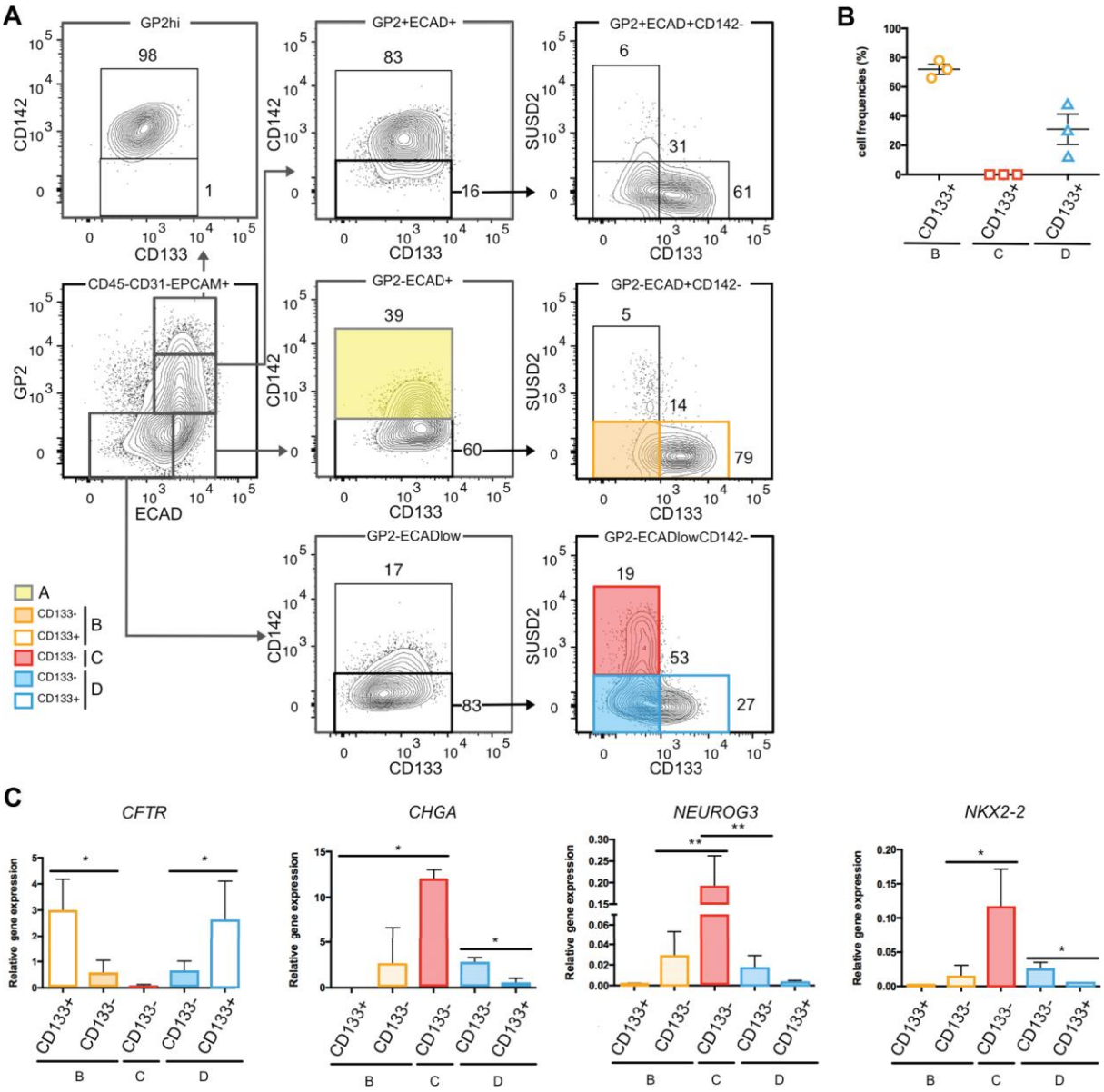


Figure 3: CD133 marks human fetal pancreatic ductal cells (**A**) Flow cytometry analysis displaying the expression of CD133 in population B, C and D at 11WD. FACS plots are representative of 3 independent pancreata at 11WD. (B) Cell frequencies of B CD133$^+$, C CD133$^+$ and D CD133$^+$ from three independent pancreata at 10WD) (**C**) Expression of *CFTR, CHGA, NEUROG3* and *NKX2-2* (9-11WD) by RT-qPCR in population B CD133$^+$, B CD133$^-$, C CD133$^-$, D CD133$^-$ and D CD133$^+$. *p<0.05, **p<0.001 t test (data represents mean of 3 independent pancreata ± SEM).

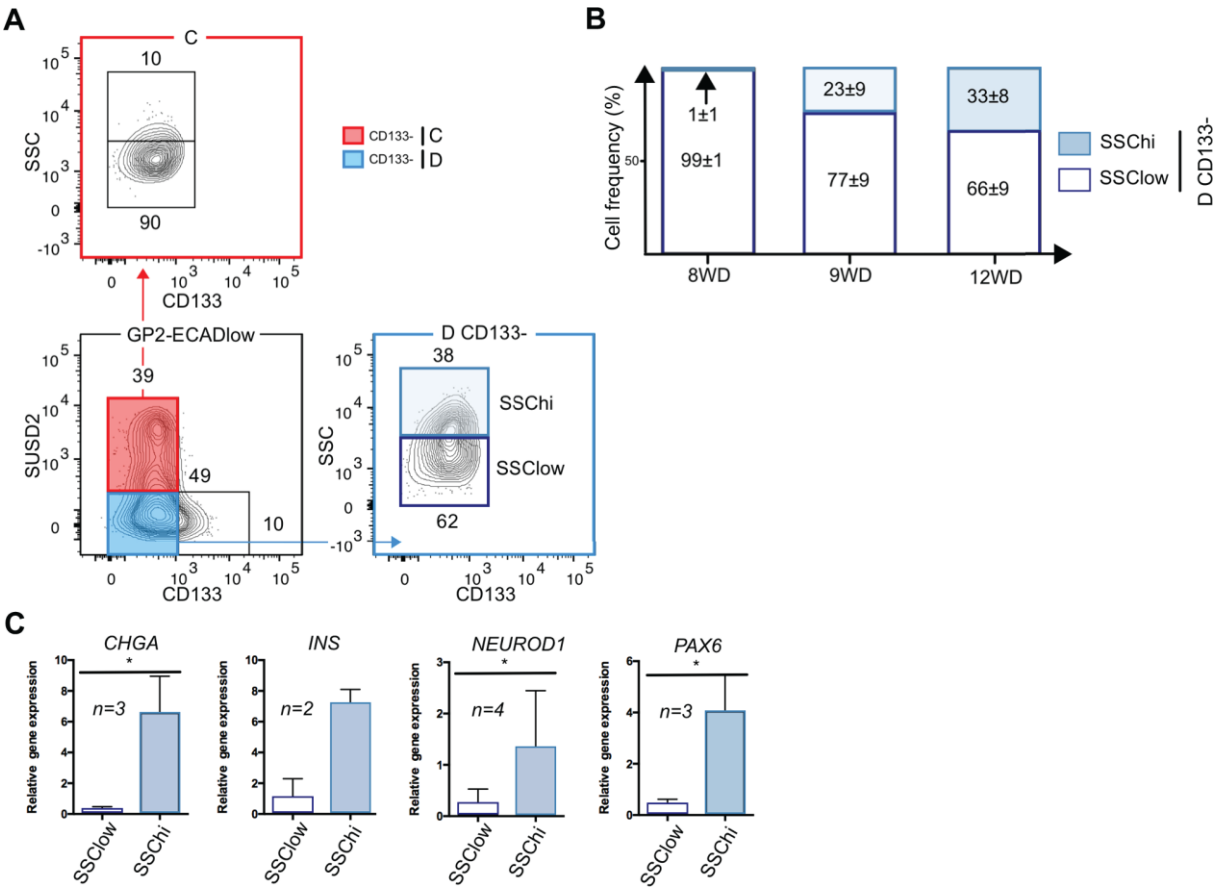


Figure 4: Endocrine cells in population D are the most granular (**A**) Flow cytometry displaying the granulometry (SSC) in C and D_{CD133-} at 10WD. FACS plots are representative of 3 independent pancreata at 10WD (**B**) Frequency of SSC^{hi} and SSC^{low} in population D_{CD133-} at 8, 9 and 12 WD (data represent mean of 3 independent pancreata at each stage ± SEM). (**C**) Expression of *CHGA, INS*, NEUROD1 and PAX6 at (10-12WD) by RT-qPCR in population D_{CD133-}, SSC^{low} or SSC^{hi}. *p<0.05, t test (data represent mean of 3 independent pancreata ± SEM).

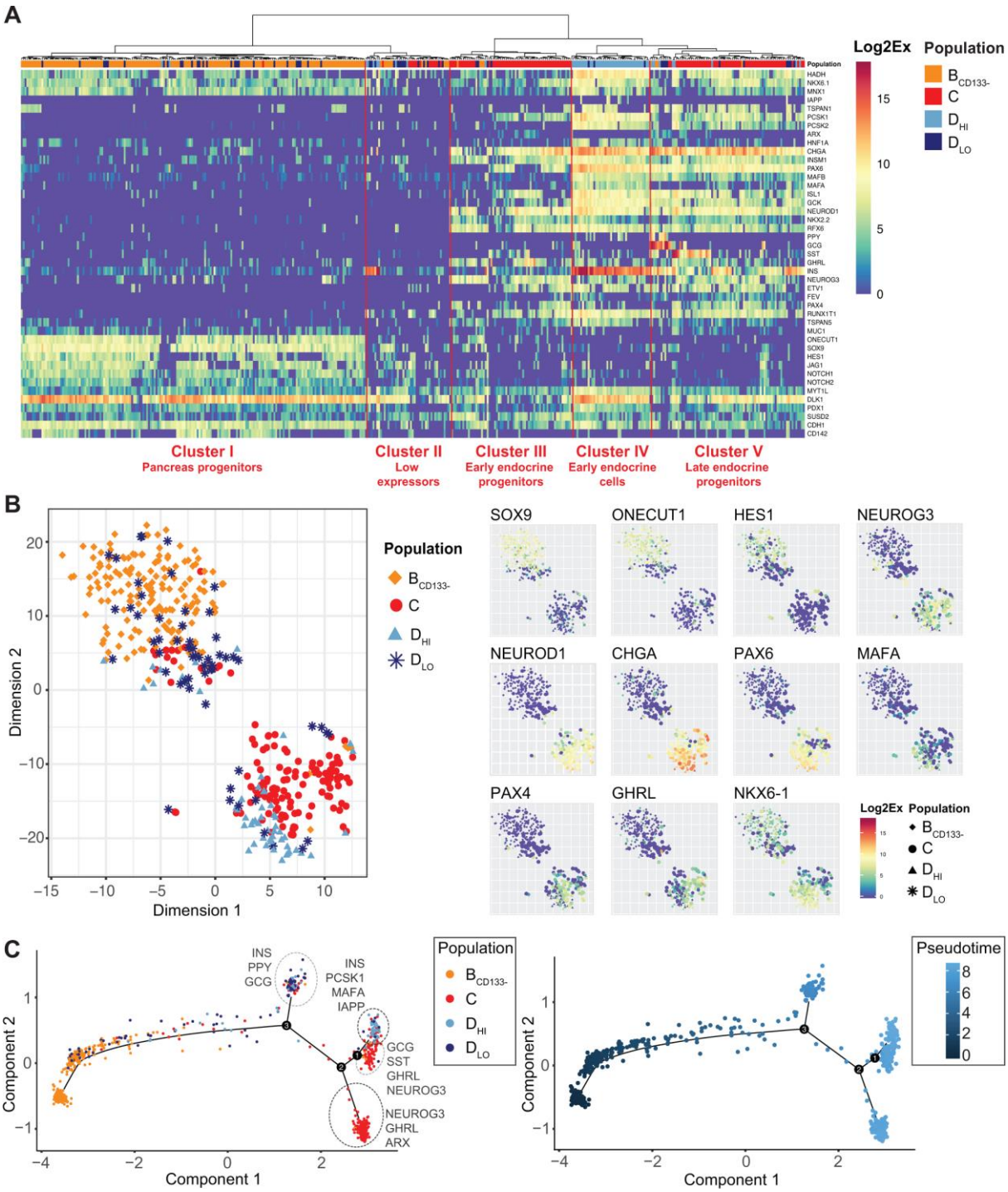


Figure 5: Single-cell profiling of sorted human fetal pancreatic cells

(**A**) Heatmap showing gene expression for a selected set of genes in individual cells sorted from human fetal pancreas based on cell-surface markers and granularity discriminating populations B_{CD133-}, C, D_{HI} and D_{LO}. The genes with greatest variance

are shown. A heatmap with all genes is provided in Fig. S4A. Cells of population B_{CD133-} and C are derived from three individual pancreata at 9WD; cells from D_{HI} and D_{LO} are derived from two individual pancreata at 9WD. (**B**) t-SNE plot of single-cell qPCR data from human fetal pancreas colored by population (left panel) or according to expression level of selected genes (right panel). See also Fig S5. **(C)** Developmental trajectory for endocrine differentiation in human fetal pancreas using pseudotemporal ordering with Monocle. Cells on the trajectory are colored according to population (left panel) or pseudotime (right panel). Specific genes characteristic for each branch are indicated on the left panel. See also Fig. S6.

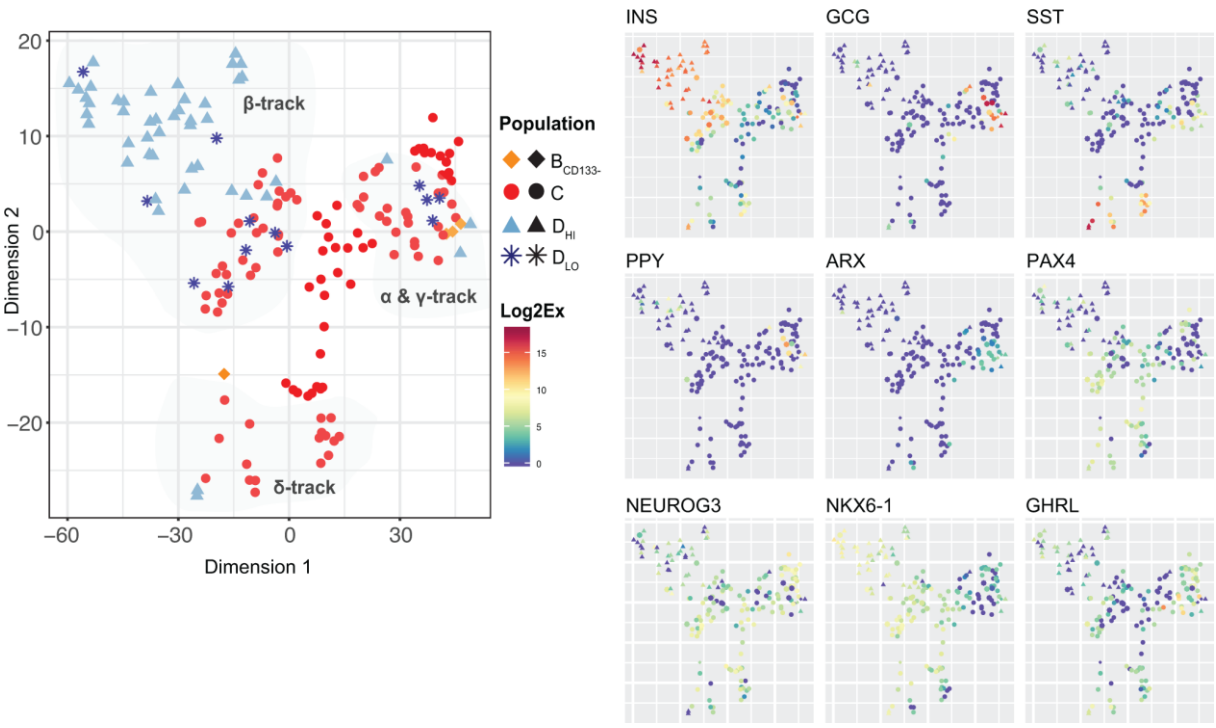


Figure 6: Gene expression profile for human fetal endocrine progenitors and early endocrine cells

T-SNE plot of single cell gene expression data for the endocrine-biased cluster (Cluster III-V) formed by the hierarchical clustering shown in Fig 5A (left panel). Cells are colored according to population (left panel) or gene expression level (right panel) for selected genes.

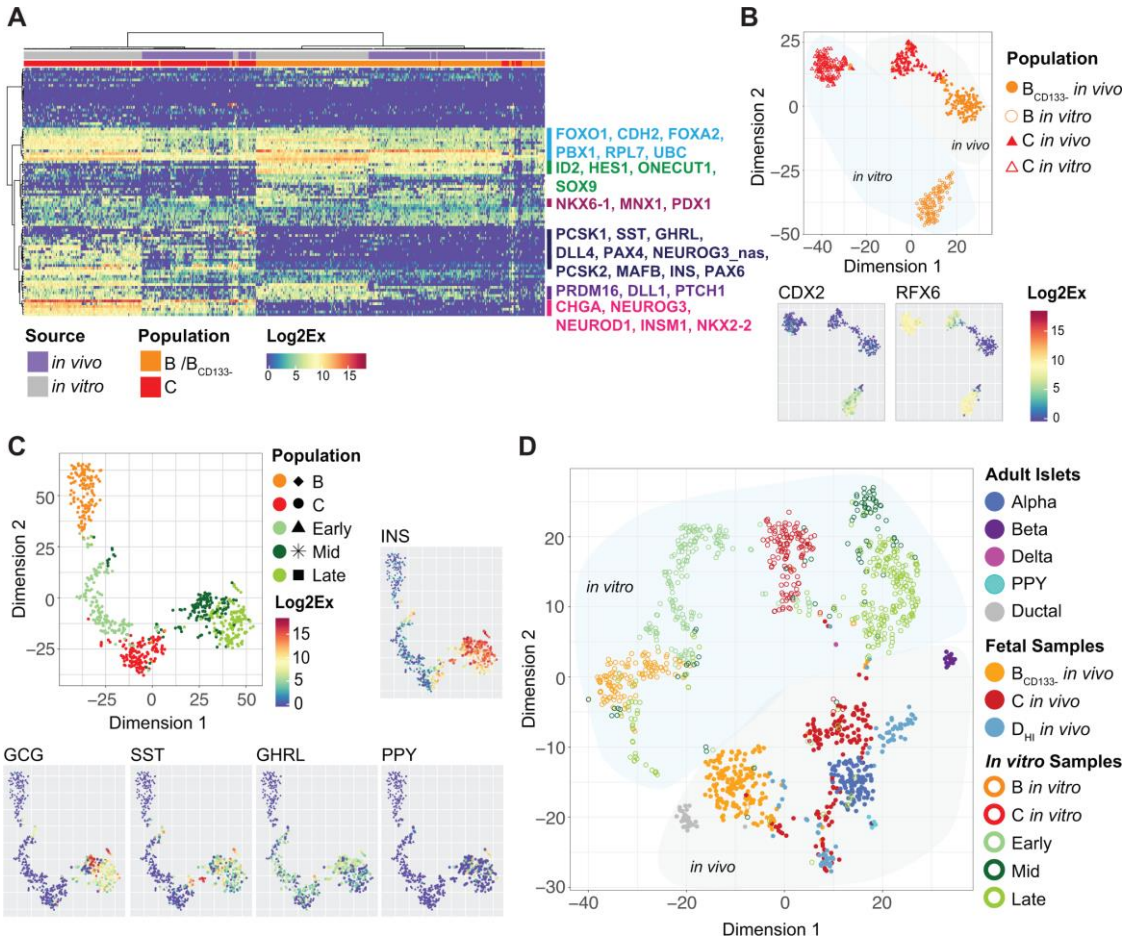
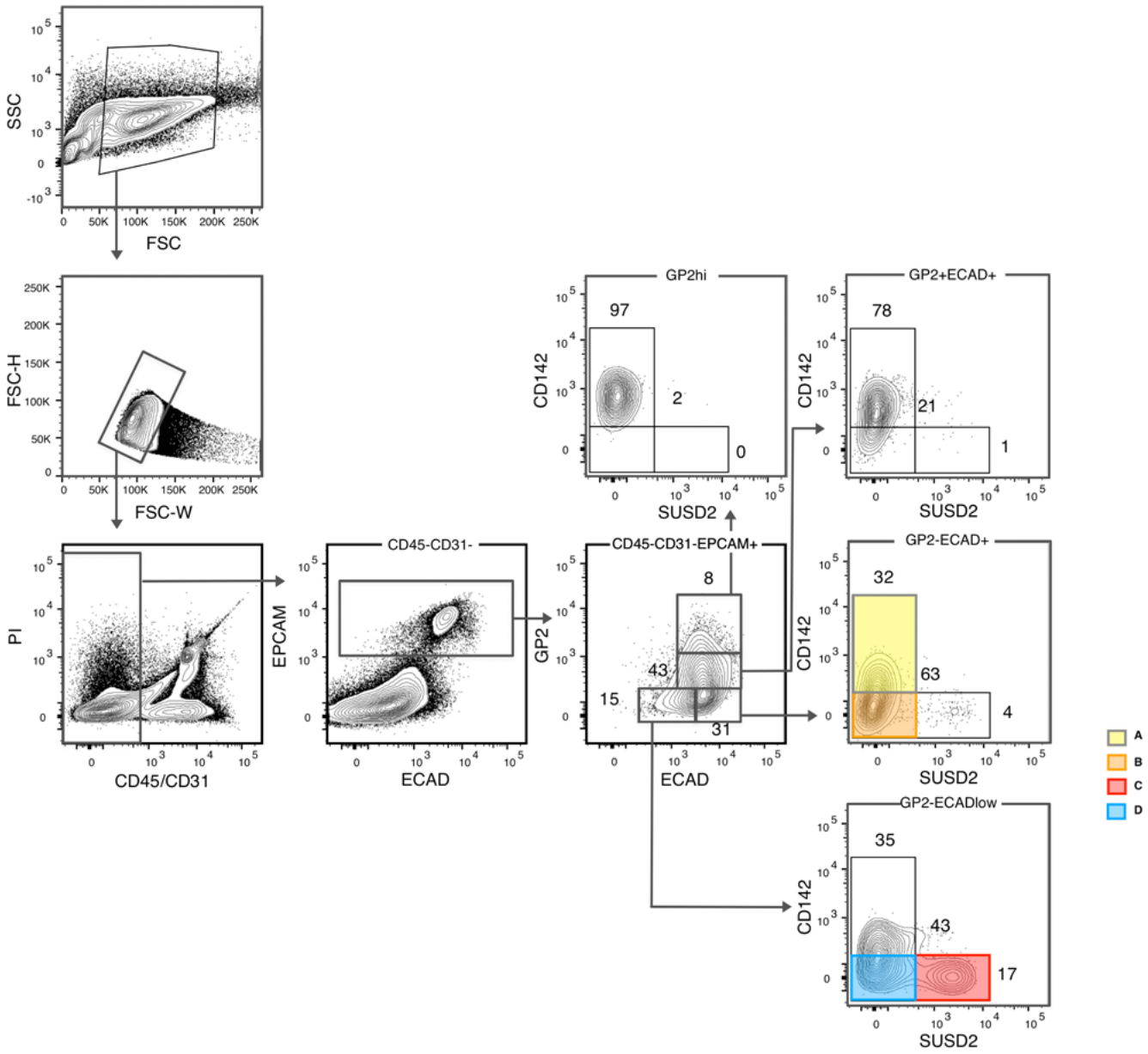
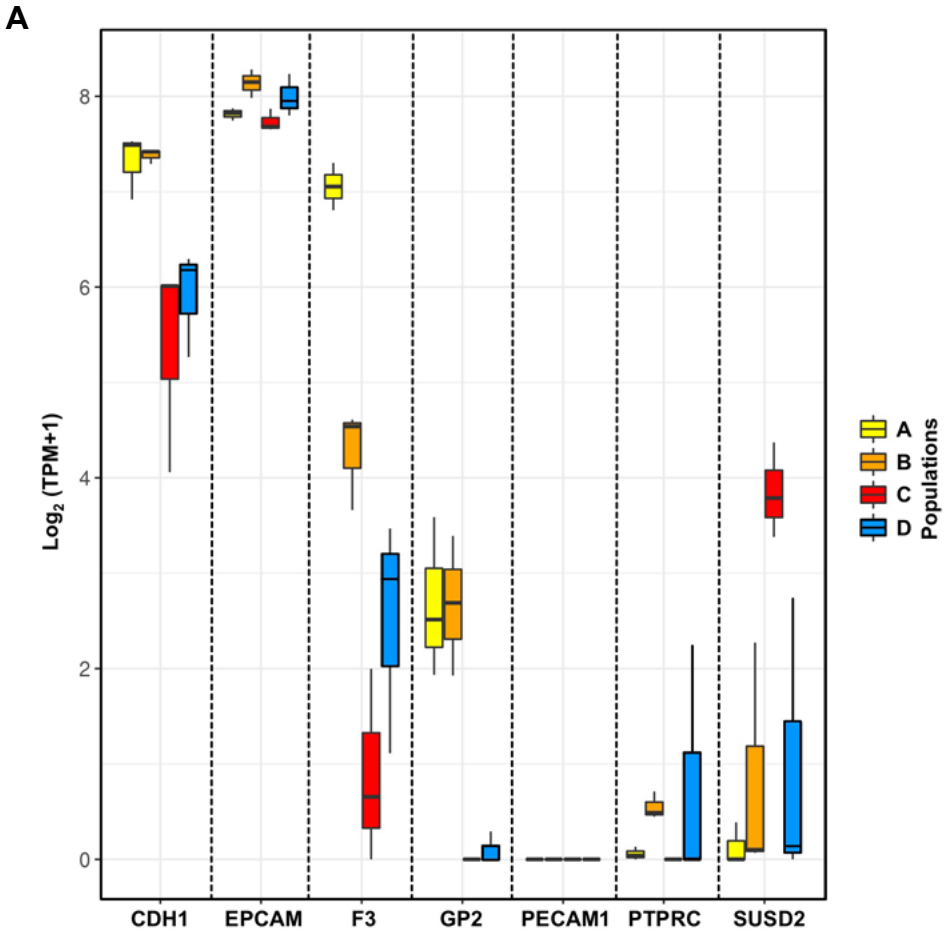


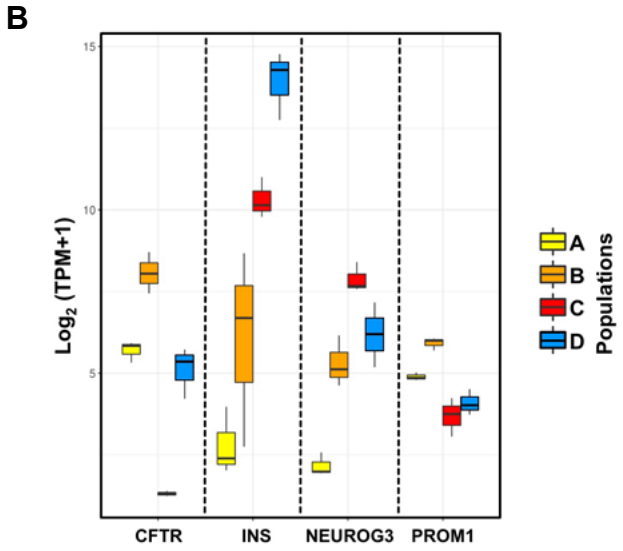
Figure 7: Comparison of the single-cell expression profile of pancreatic cells generated *in vitro* to *in vivo* fetal and adult pancreata

(**A**) Heatmap showing the gene expression level in single cells of populations B and C isolated from human fetal pancreas (*in vivo*) or from hPSC-derived cultures (*in vitro*). Selected genes are indicated. A heatmap with all gene annotations is provided in Fig. S9. (**B**) T-SNE plot of single-cell qPCR data for populations B and C isolated from human fetal pancreas (*in vivo*) or from hPSC-derived cultures (*in vitro*). Gene expression level in individual cells for *RFX6* and *CDX2* is indicated below. (**C**) T-SNE plot combining data from the present study on sorted hPSC-derived cells (populations C and D) with data from a previously published dataset comprising endocrine-biased hPSC-derived cells collected at different time points (stage 4 day 1 and 3: “early”, stage 5 day 3 + stage 6 day 2: “mid” and stage 6 day 7 + stage 7 day 7: “late”). Expression level of hormonal genes are mapped onto the t-SNE plot as indicated. (**D**) T-SNE plot combining the data from C with data on the sorted cell populations from human fetal pancreas (populations B, C and D_{HI}) from the present study and data on adult human islet cells (from the same previously published dataset as used in C). See also Fig. S10.

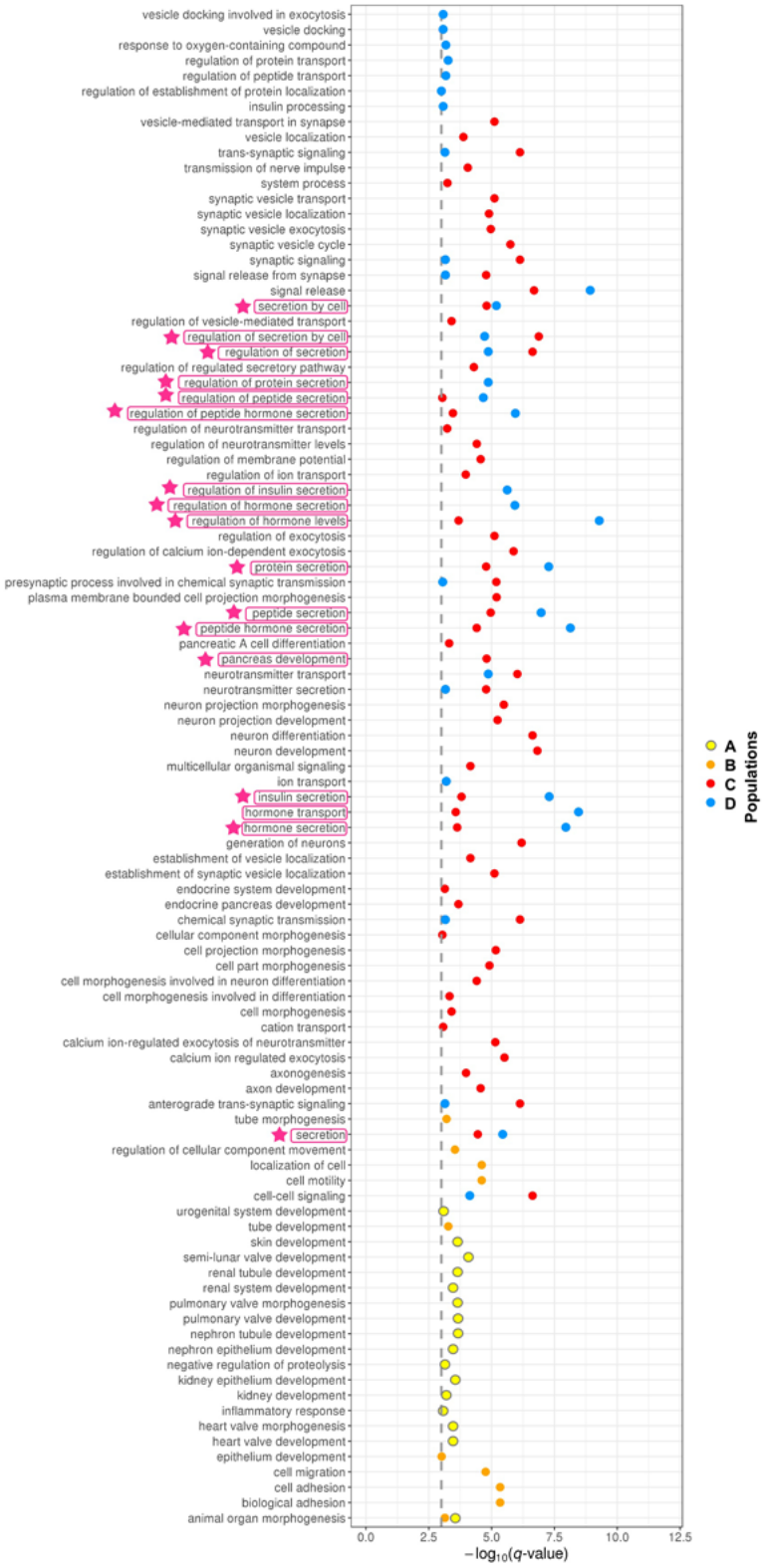


Supplementary Figure 1: Gating strategy for the expression of GP2, ECAD, CD142 and SUSD2 Human fetal pancreata at 9WD were stained for CD45, CD31, EPCAM, ECAD, GP2, CD142 and SUSD2. Doublet cells were excluded from the analysis with FSC-H and FSC-W (middle top plot). Propidium iodide (PI) was used to exclude dead cells as shown in the right top plot in the diagonal. GP2 and ECAD expression was analyzed in the CD45-CD31-EPCAM+ population. CD45-CD31-EPCAM- population was used as negative control to set up the GP2-ECAD+ and GP2+ECAD+ gates. GP2+ECAD+ population was used to set up the gate for ECAD levels. CD142 and SUSD2 expression was analyzed in GP2hiECAD+, GP2+ECAD+ and GP2-ECADlow. GP2hiECAD+ population was used as a positive control for the expression of CD142 and as a negative control for the expression of SUSD2. This gating strategy was applied to each pancreatic stage.

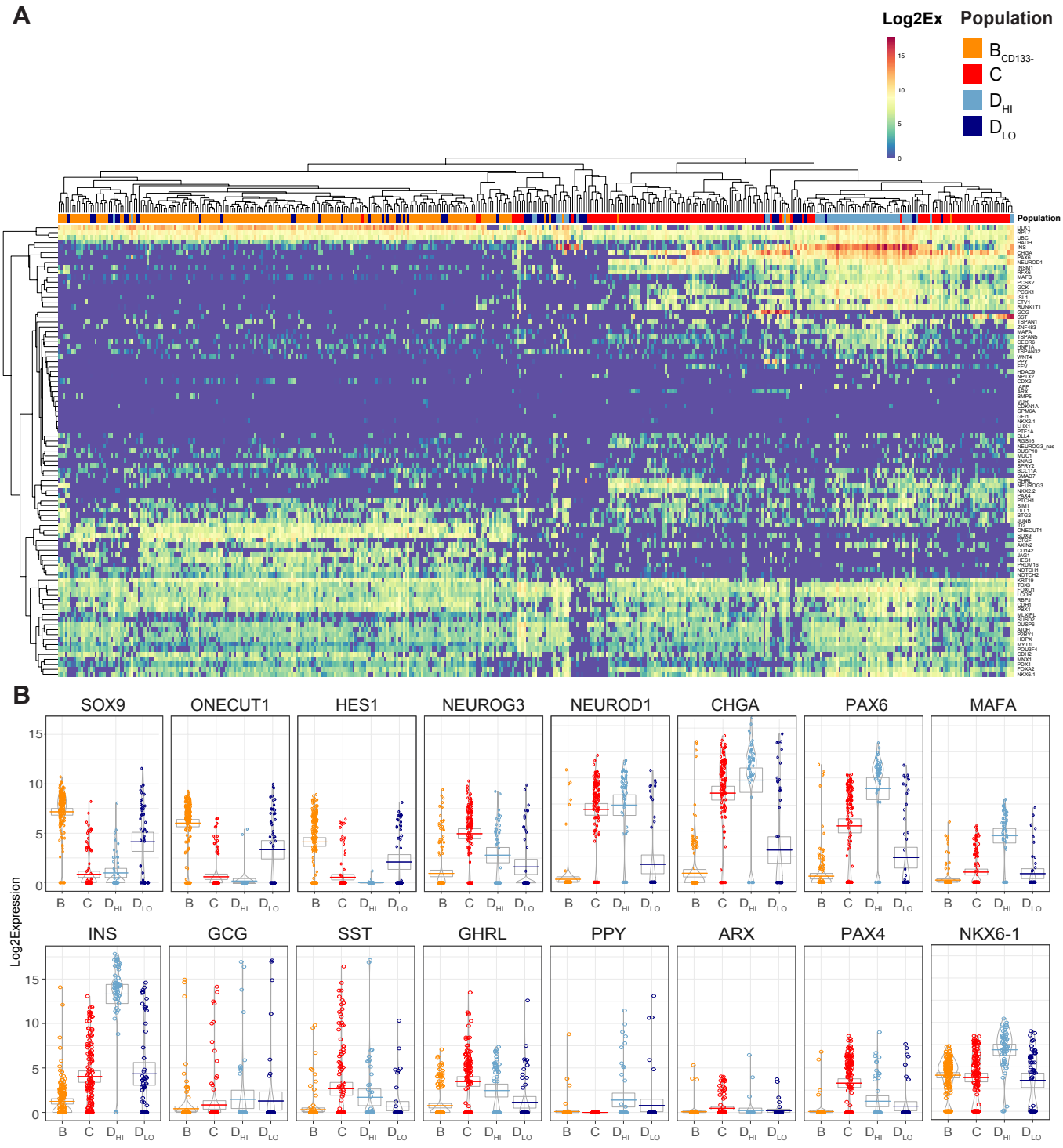




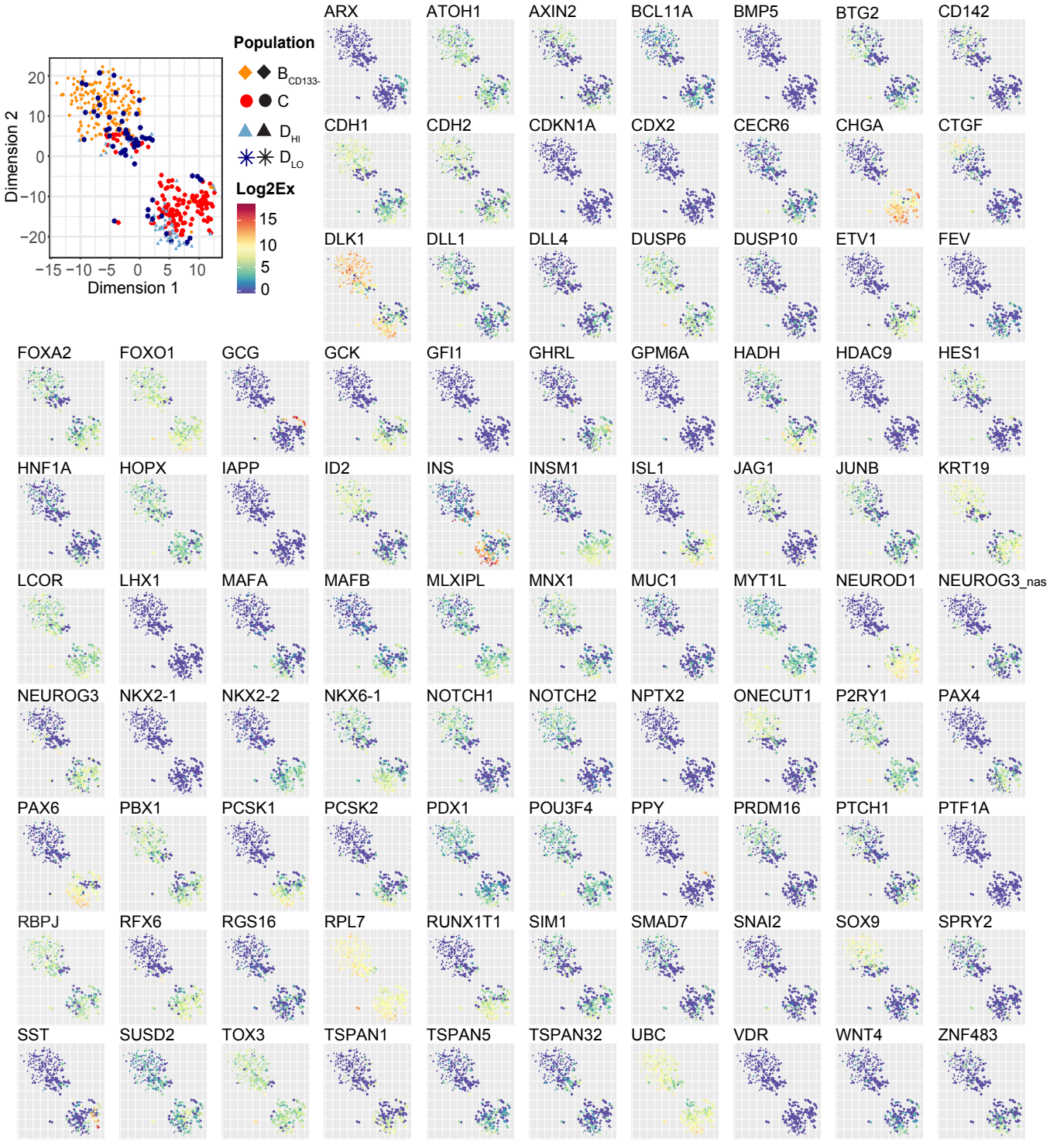
Supplementary Figure 2: Expression of membrane protein genes, ductal and endocrines genes in population A-D at 9WD
(A) Expression of genes encoding membrane proteins (CDH1, EPCAM, F3, GP2, PECAM1, PTPRC and SUSD2) used for the cell sorting of population A, B, C and D. CDH1 codes for ECAD, F3 for CD142, PECAM1 for CD31, PTPRC for CD45. **(B)** Expression of CFTR, NEUROG3 and PROM1 in populations A, B, C and D



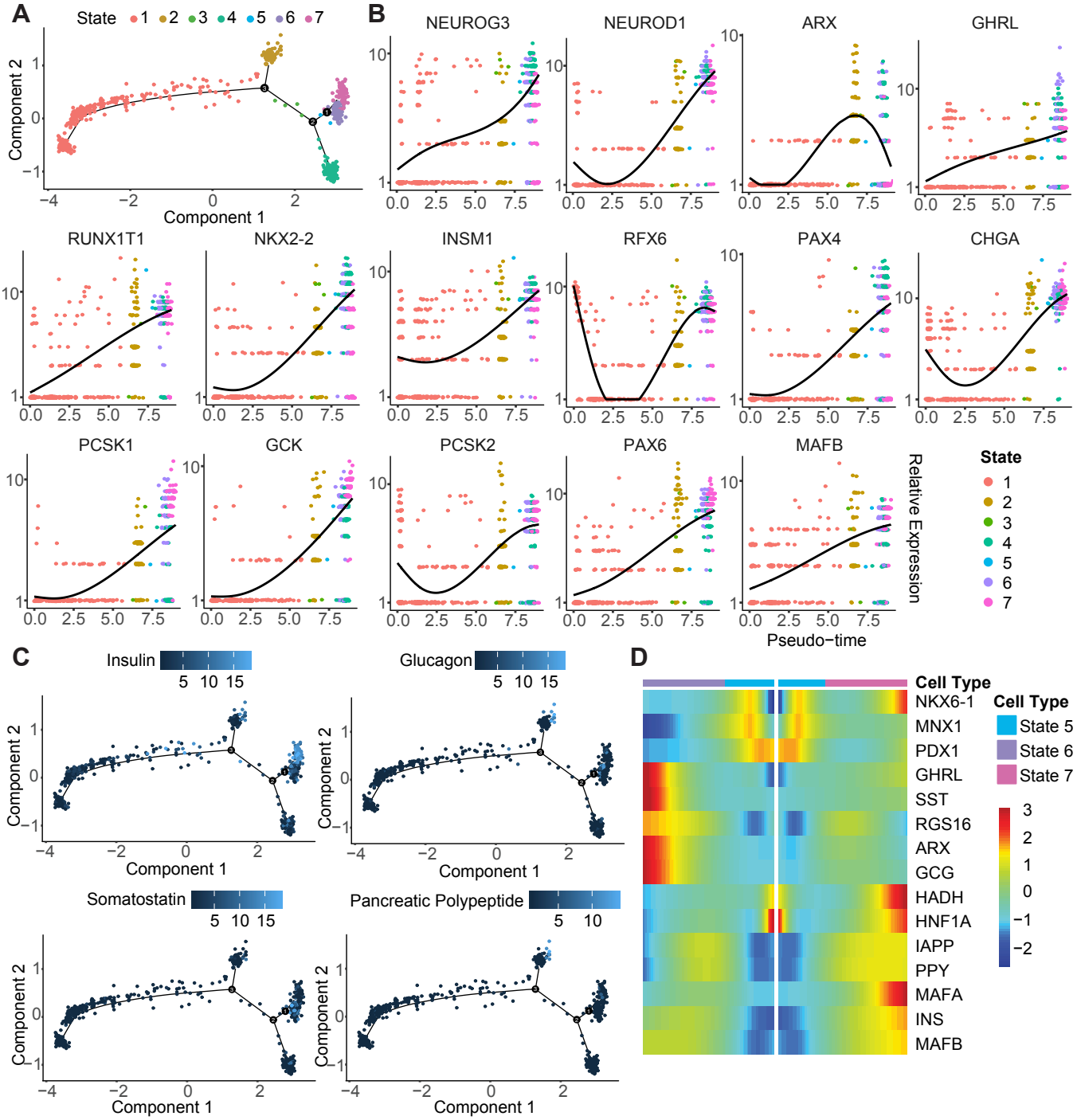
Supplementary Figure 3: Gene set enrichment analysis on population A, B, C and D GSEA analysis on populations A, B, C and D at 9WD using Gene ontology database (FDR <1%).



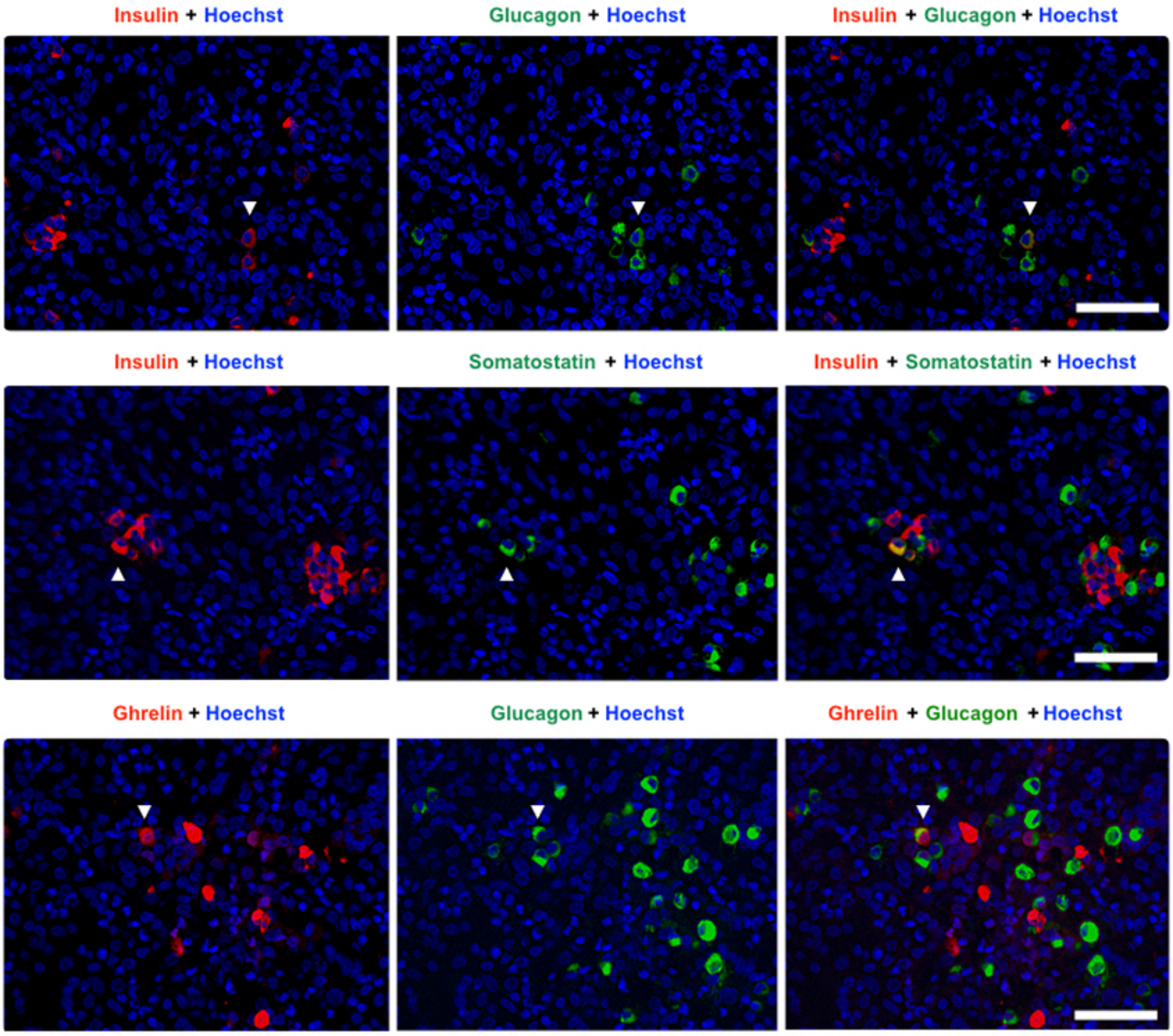
Supplementary Figure 4: Unsupervised hierarchical clustering of single-cell qPCR data from sorted human fetal pancreatic cells
(A) Unsupervised hierarchical clustering of sorted human fetal pancreatic cells at 9WD, showing all genes for populations B_{CD133-}, C, D_{HI} and D_{LO}. **(B)** Pirate plots showing the expression levels of specific genes in the four sorted populations. Each plot is a combination of box plots (25%-75%), means, violin plots and individual data points.



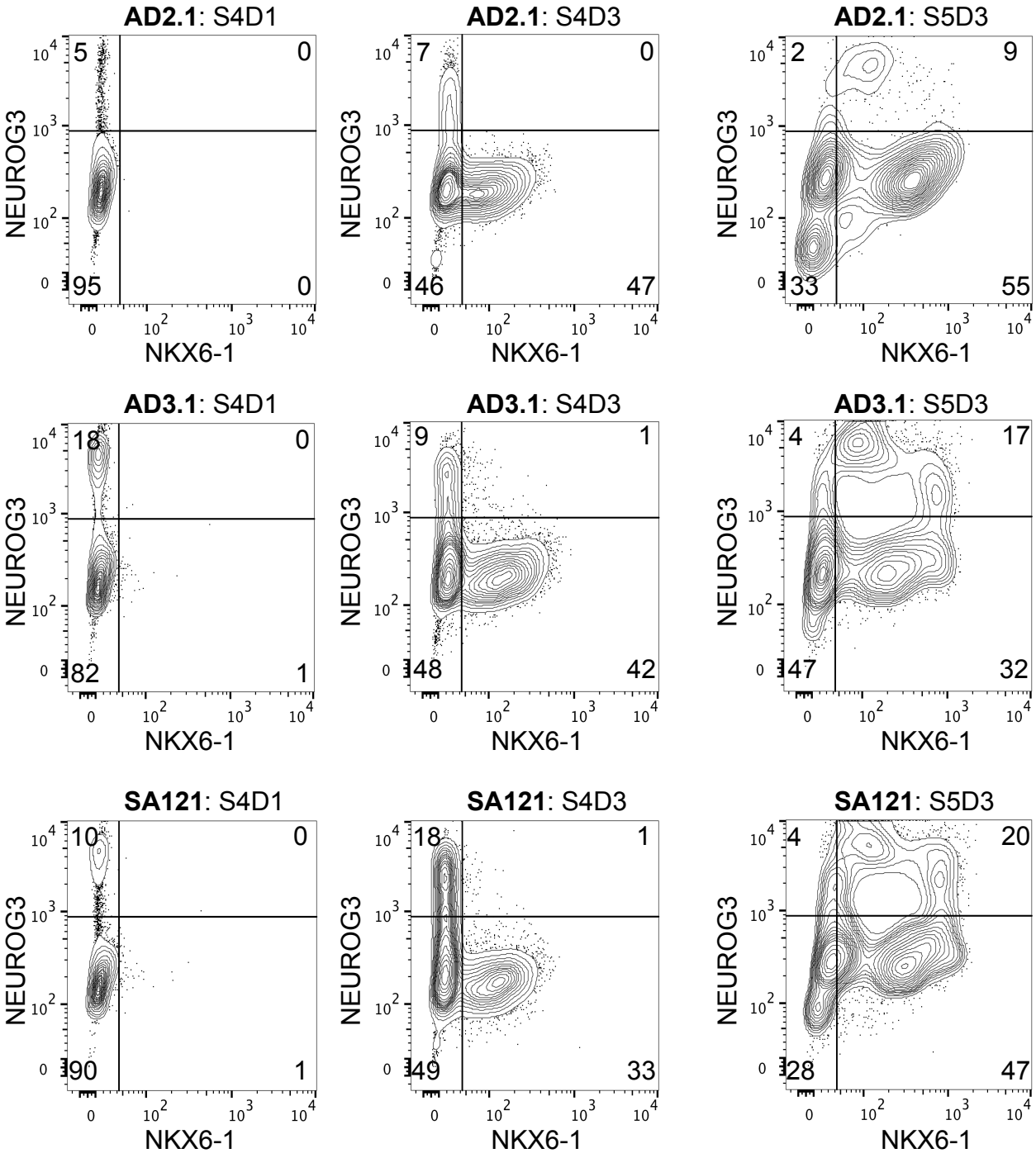
Supplementary Figure 5: Extended gene expression profiling of individual human fetal pancreas cells. t-SNE plots (corresponding to Fig. 5B) colored according to gene expression level of the indicated genes.



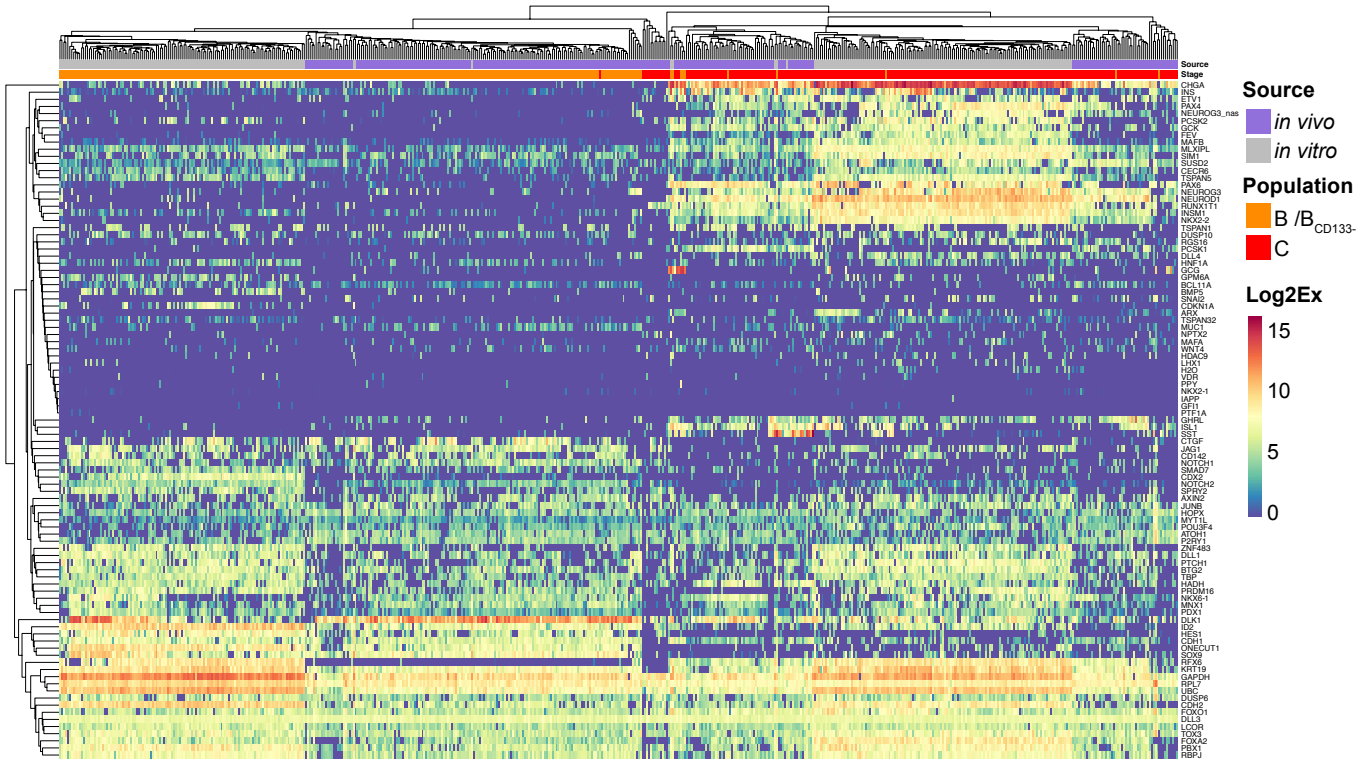
Supplementary Figure 6: Pseudotemporal ordering of single-cell gene expression data from human fetal pancreas. (A) Developmental trajectory of human fetal pancreatic cells (corresponding to Fig. 5C) colored by states. **(B)** Gene expression plots showing the pseudotemporal development of key genes involved in pancreas development. Gene expression level is shown on the y-axis; pseudotime on the x-axis. Each data point represents a single cell and is colored according to state on the trajectory shown in A. **(C)** Developmental trajectory of human fetal pancreatic cells (corresponding to Fig. 5C) colored by gene expression levels of selected hormonal genes. **(D)** Heat map showing pseudotemporal development of gene expression for the two cell fates derived from branching point 1 on the developmental trajectory shown in A. Cells at this branching point differentiates towards either State 6 or State 7.



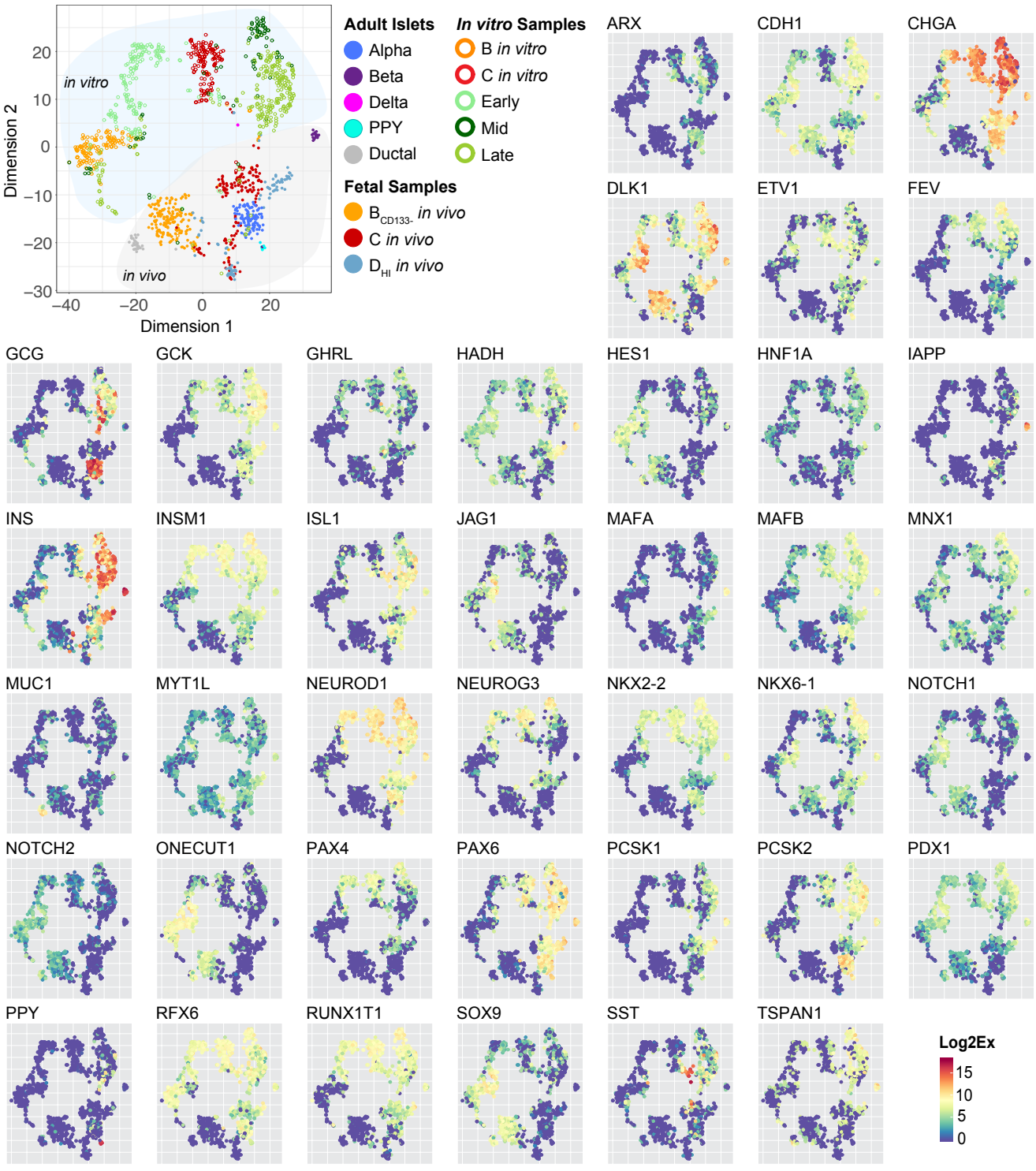
Supplementary Figure 7: Co-expression of pancreatic hormones in human fetal pancreas. Immunofluoresence staining for insulin, glucagon, somatostatin and ghrelin on pancreatic section at 10WD. Scale bar: 50µm. Arrowheads indicate double hormone positive cells.



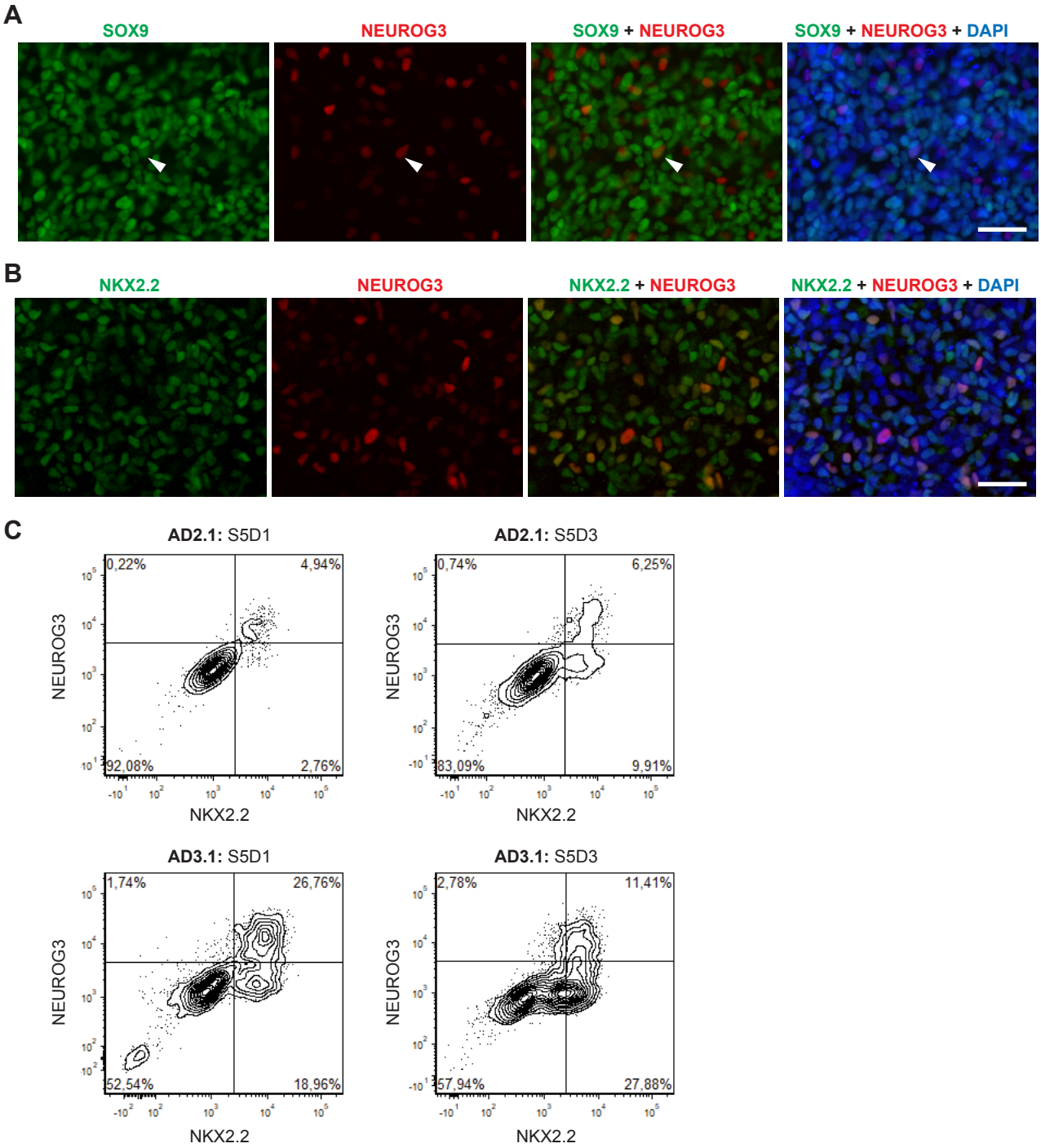
Supplementary Figure 8: Generation of endocrine progenitors from hPSCs. Representative flow cytometry plots showing the percentage of cells expressing NEUROG3 and NKX6-1 protein at S4D1, S4D3 and S5D3 of the differentiation protocol for each of the hPSC-lines SA121, AD2.1 and AD3.1 used for sorting based on cell-surface markers and subsequent single-cell qPCR analysis.



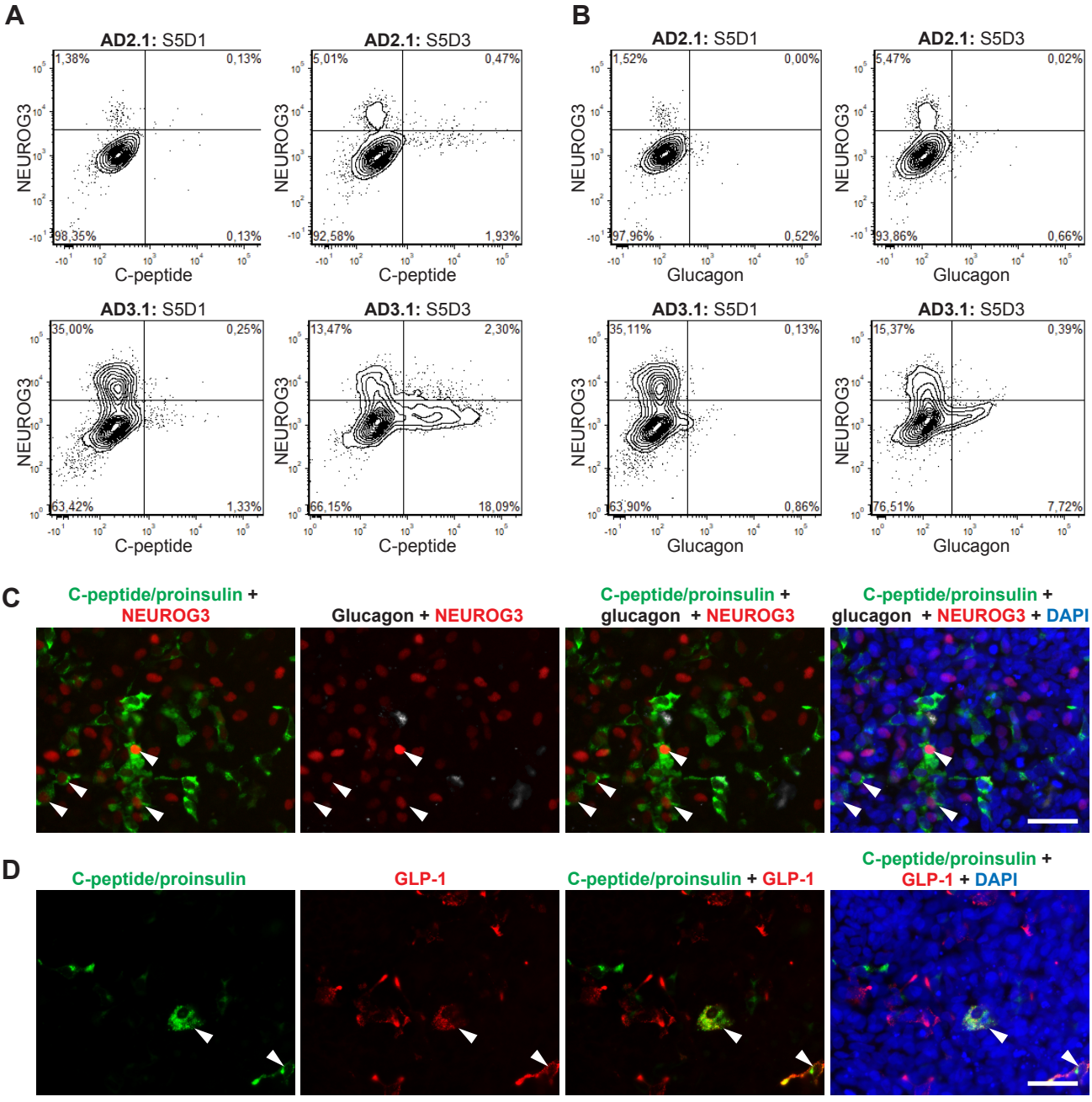
Supplementary Figure 9: Extended comparison of the single-cell expression profile of pancreatic cells generated *in vitro* to *in vivo* fetal and adult pancreas Hierarchical clustering of single-cell qPCR data showing the expression of all genes for populations B and C sorted from human fetal pancreas ("*in vivo*") or hPSC-derived pancreatic cells ("*in vitro*"). Refers to Fig. 7A and B.



Supplementary Figure 10: Gene expression in single pancreatic cells from human fetal pancreas, adult human islets or generated *in vitro* from hPSCs. t-SNE plots corresponding to Fig. 7D colored according to gene expression level of the indicated genes.



Supplementary Figure 11: Characterization of SOX9 and NKX2-2 protein co-expression in hPSC-derived endocrine progenitor subpopulations. (A) Representative image of co-staining for SOX9 (green) and NEUROG3 (red) at S5D1 of in vitro differentiated cultures (hiPSC line AD2.1). Arrowhead indicates a SOX9+/NEUROG3+ cell. Nuclei are counterstained with DAPI. Scale bar: 50 µm. **(B)** Representative image of co-staining for NKX2-2 (green) and NEUROG3 (red) at S5D1 of in vitro differentiated cultures (hiPSC line AD3.1). Nuclei are counterstained with DAPI. Scale bar: 50 µm. **(C)** Flow cytometry analysis for co-expression of NKX2-2 and NEUROG3 at S5D1 and S5D3 of in vitro differentiated cultures (hiPSC lines AD2.1 is shown in the top panel and AD3.1 in the bottom panel).



Supplementary Figure 12: Characterization of hormone expression at the protein level in hPSC-derived endocrine progenitor subpopulations. (A) Flow cytometry analysis for co-expression of NEUROG3 and C-peptide or **(B)** NEUROG3 and glucagon, at S5D1 and S5D3 of in vitro differentiated cultures (hiPSC lines AD2.1 is shown in the top panel and AD3.1 in the bottom panel). **(C)** Representative image of co-staining for C-peptide/proinsulin (green), NEUROG3 (red) and glucagon (white) at S5D1 of in vitro differentiated cultures (hiPSC line AD2.1). Arrowheads indicate co-expression of C-peptide and NEUROG3. Scale bar: 50 µm. **(D)** Representative image of co-staining for C-peptide/proinsulin (green) and GLP-1 (red) at S5D1 of in vitro differentiated cultures (hiPSC line AD3.1). Arrowhead indicates co-expression of C-peptide/proinsulin and GLP1. Scale bar: 50 µm.

Table S1: List of the 1000 genes differentially expressed in populations A, B, C and D

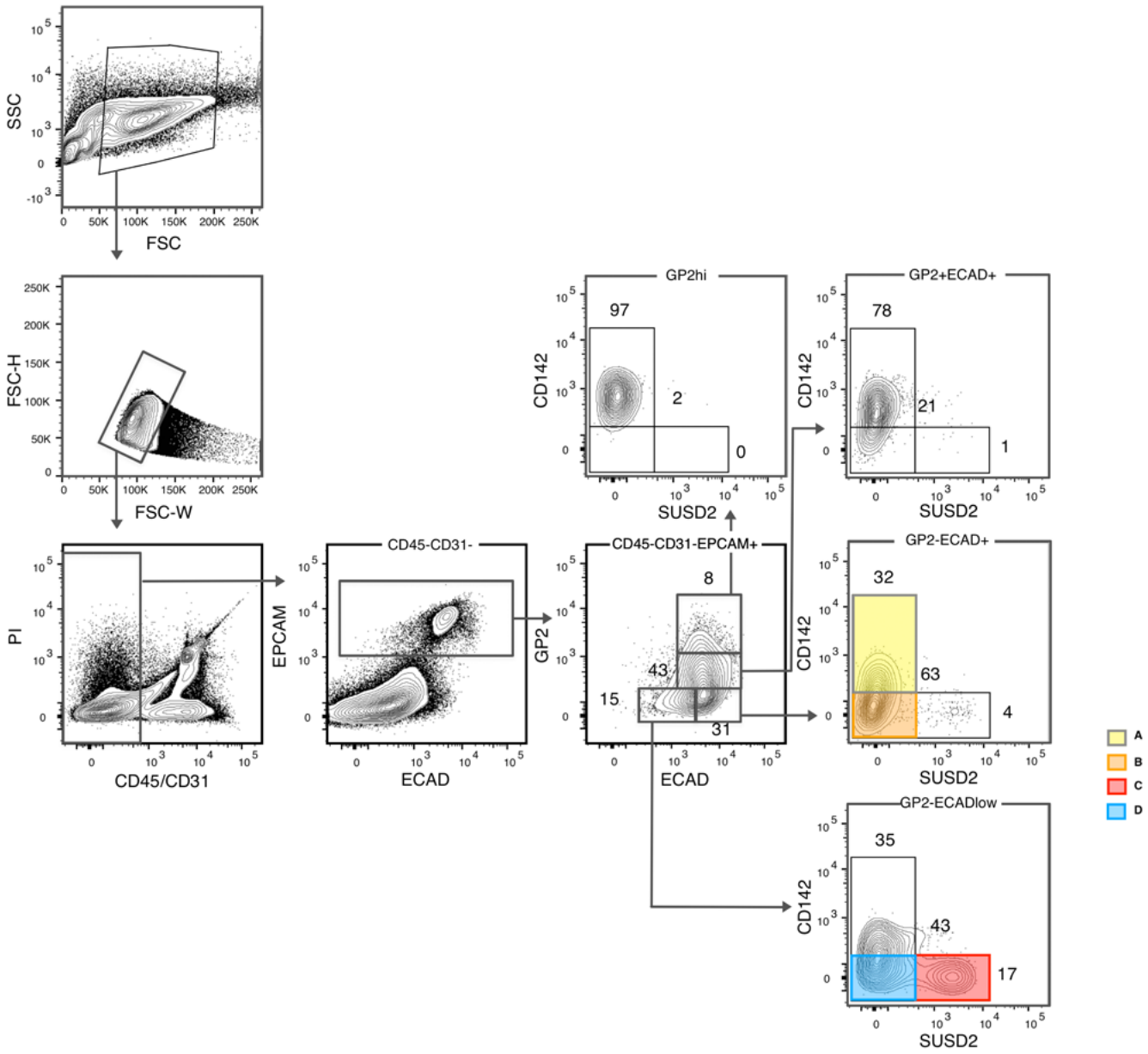
Click here to Download Table S1

Table S2: Gene ontology functional classification

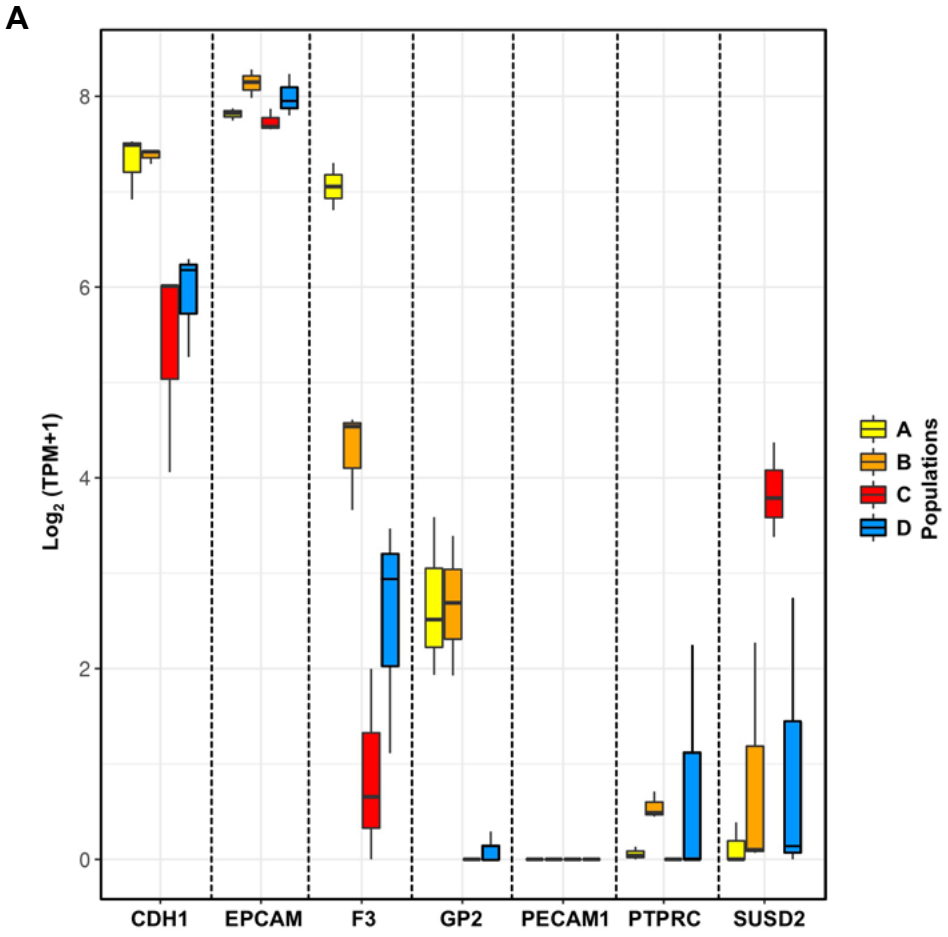
Click here to Download Table S2

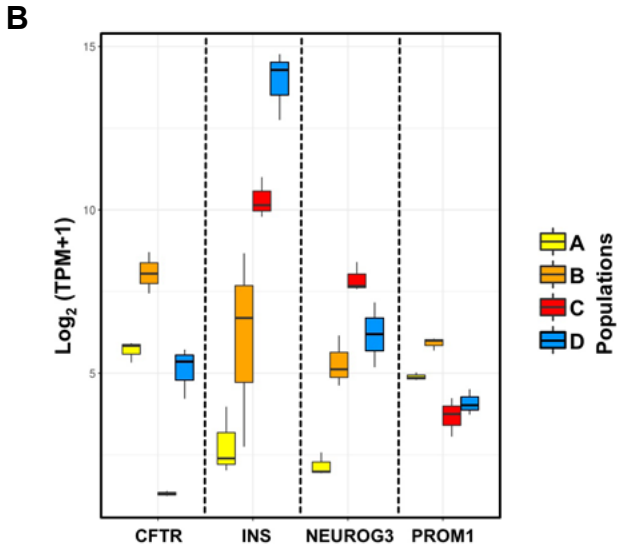
Table S3: Single cell qPCR processed data (Log2 expression values), primer sequences and information on number of cells analyzed

Click here to Download Table S3

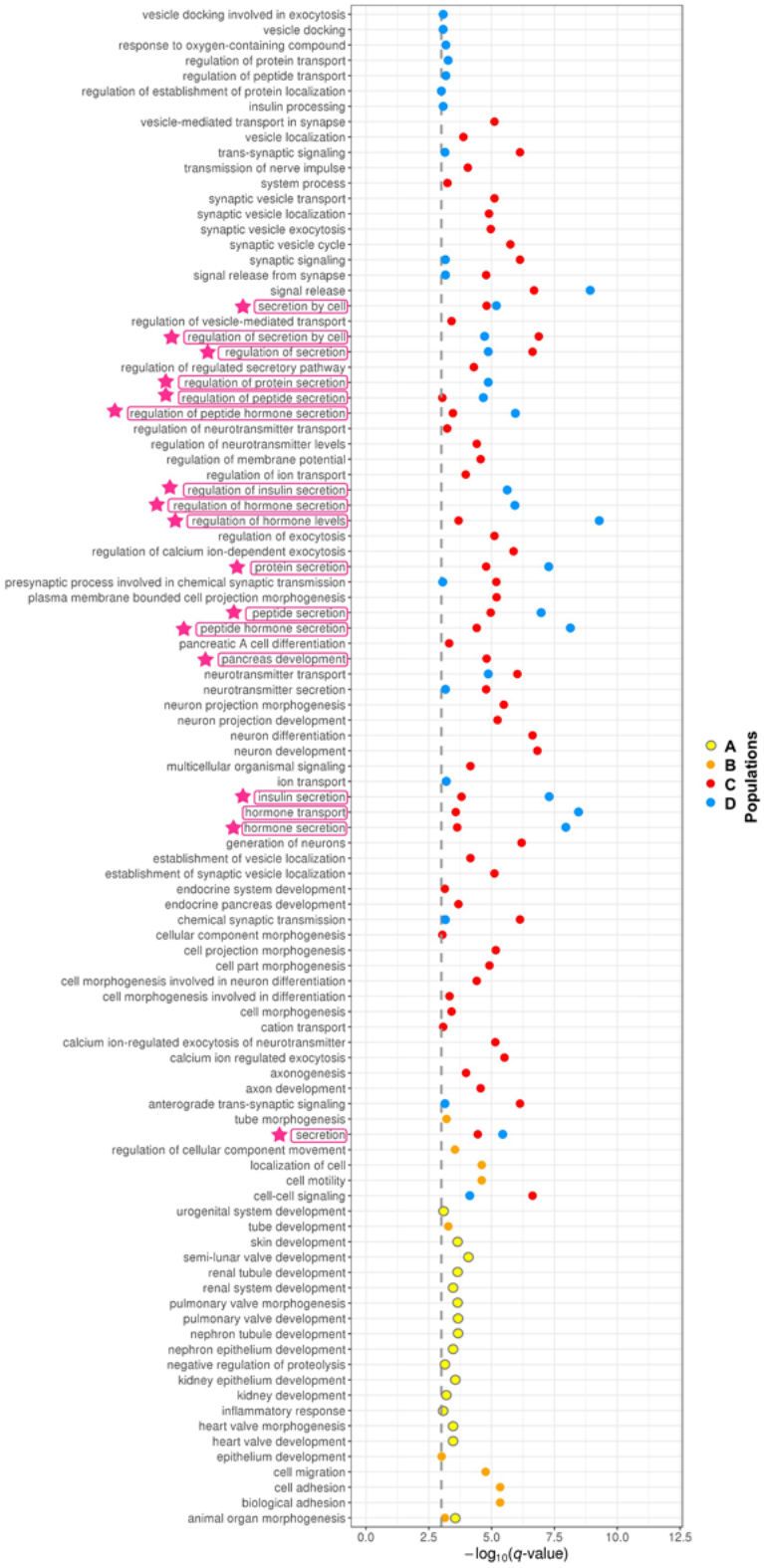


Supplementary Figure 1: Gating strategy for the expression of GP2, ECAD, CD142 and SUSD2 Human fetal pancreata at 9WD were stained for CD45, CD31, EPCAM, ECAD, GP2, CD142 and SUSD2. Doublet cells were excluded from the analysis with FSC-H and FSC-W (middle top plot). Propidium iodide (PI) was used to exclude dead cells as shown in the right top plot in the diagonal. GP2 and ECAD expression was analyzed in the CD45-CD31-EPCAM+ population. CD45-CD31-EPCAM- population was used as negative control to set up the GP2-ECAD+ and GP2+ECAD+ gates. GP2+ECAD+ population was used to set up the gate for ECAD levels. CD142 and SUSD2 expression was analyzed in GP2hiECAD+, GP2+ECAD+ and GP2-ECADlow. GP2hiECAD+ population was used as a positive control for the expression of CD142 and as a negative control for the expression of SUSD2. This gating strategy was applied to each pancreatic stage.

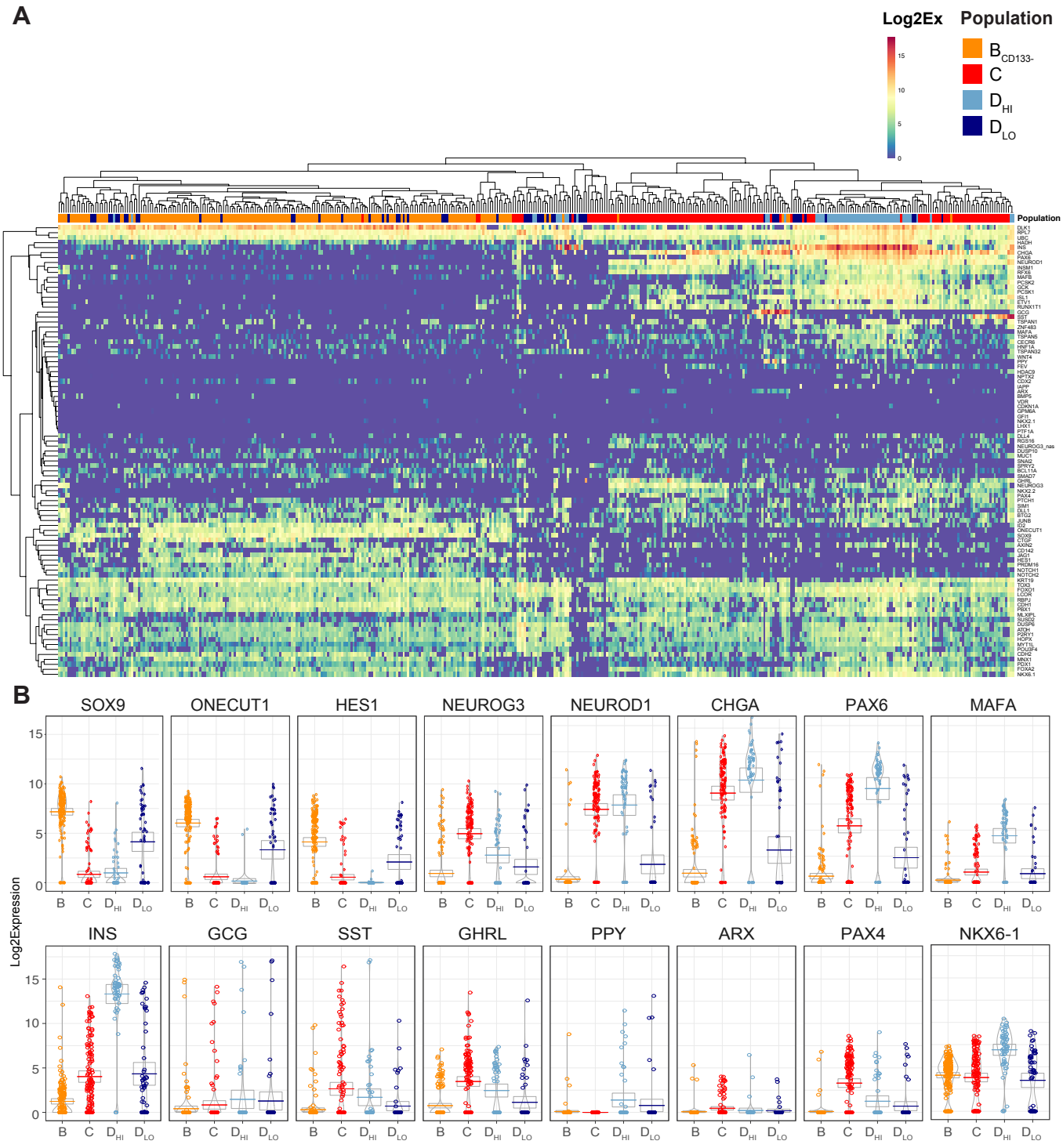




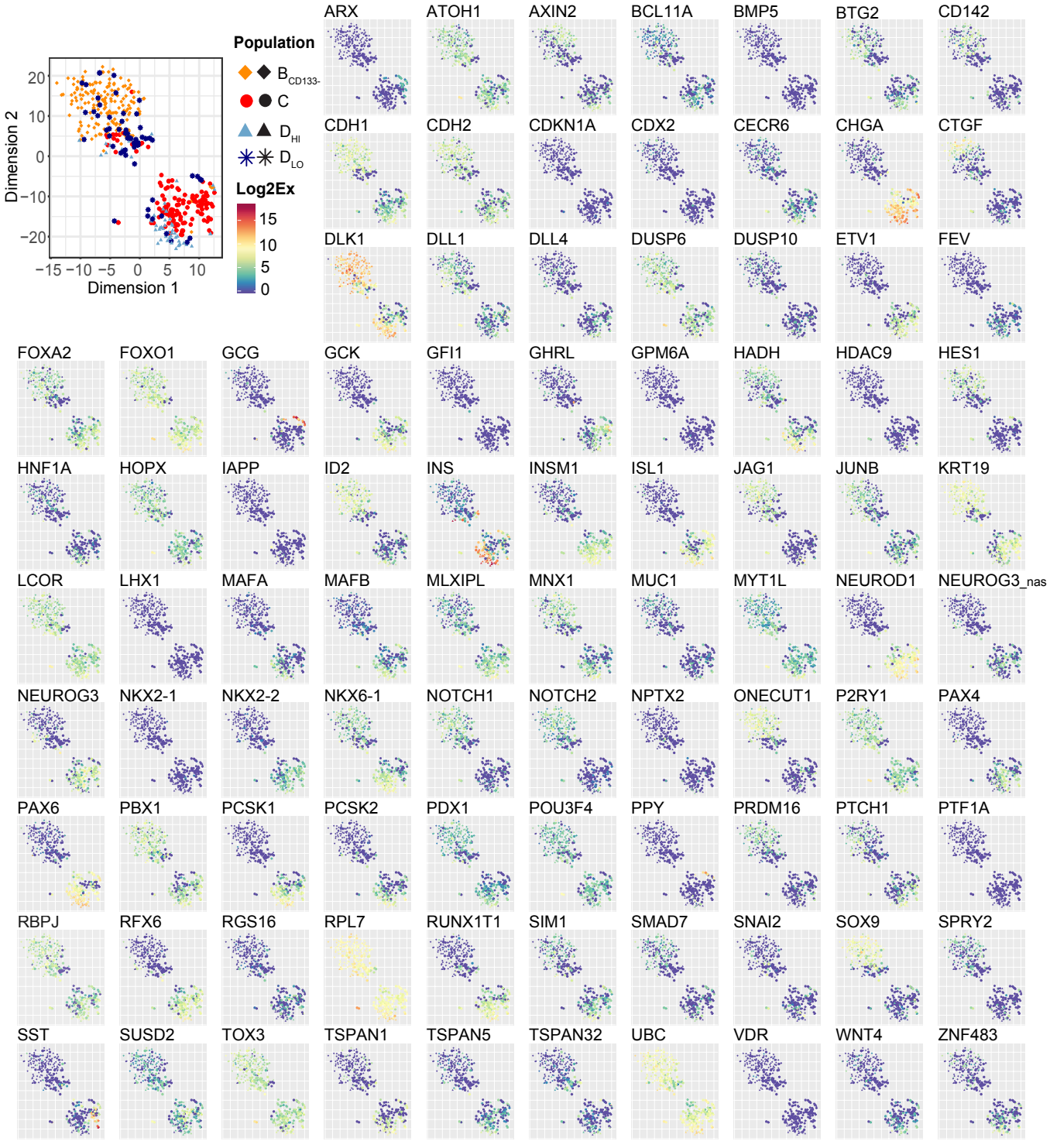
Supplementary Figure 2: Expression of membrane protein genes, ductal and endocrines genes in population A-D at 9WD
(A) Expression of genes encoding membrane proteins (CDH1, EPCAM, F3, GP2, PECAM1, PTPRC and SUSD2) used for the cell sorting of population A, B, C and D. CDH1 codes for ECAD, F3 for CD142, PECAM1 for CD31, PTPRC for CD45. **(B)** Expression of CFTR, NEUROG3 and PROM1 in populations A, B, C and D



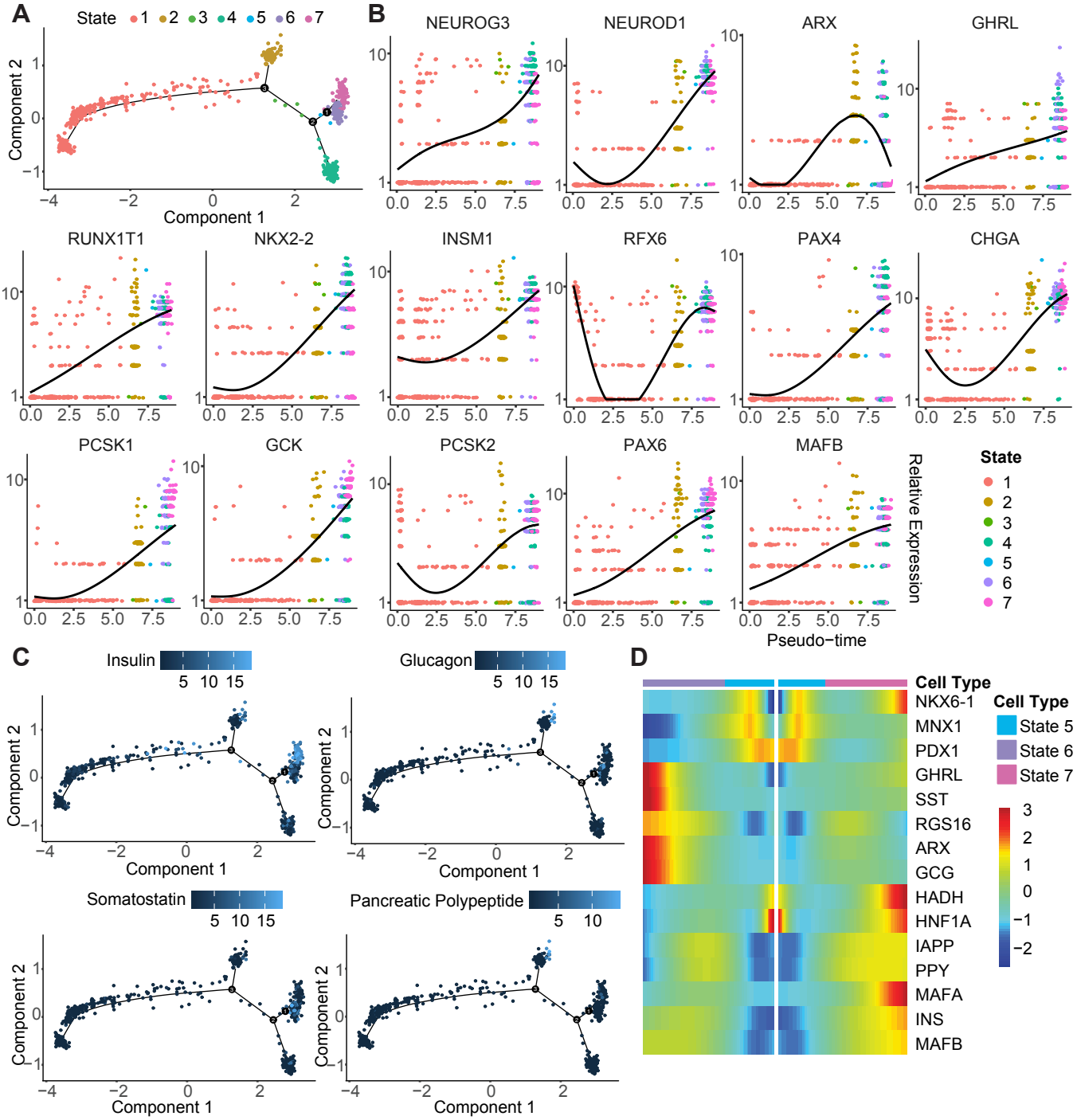
Supplementary Figure 3: Gene set enrichment analysis on population A, B, C and D GSEA analysis on populations A, B, C and D at 9WD using Gene ontology database (FDR <1%).



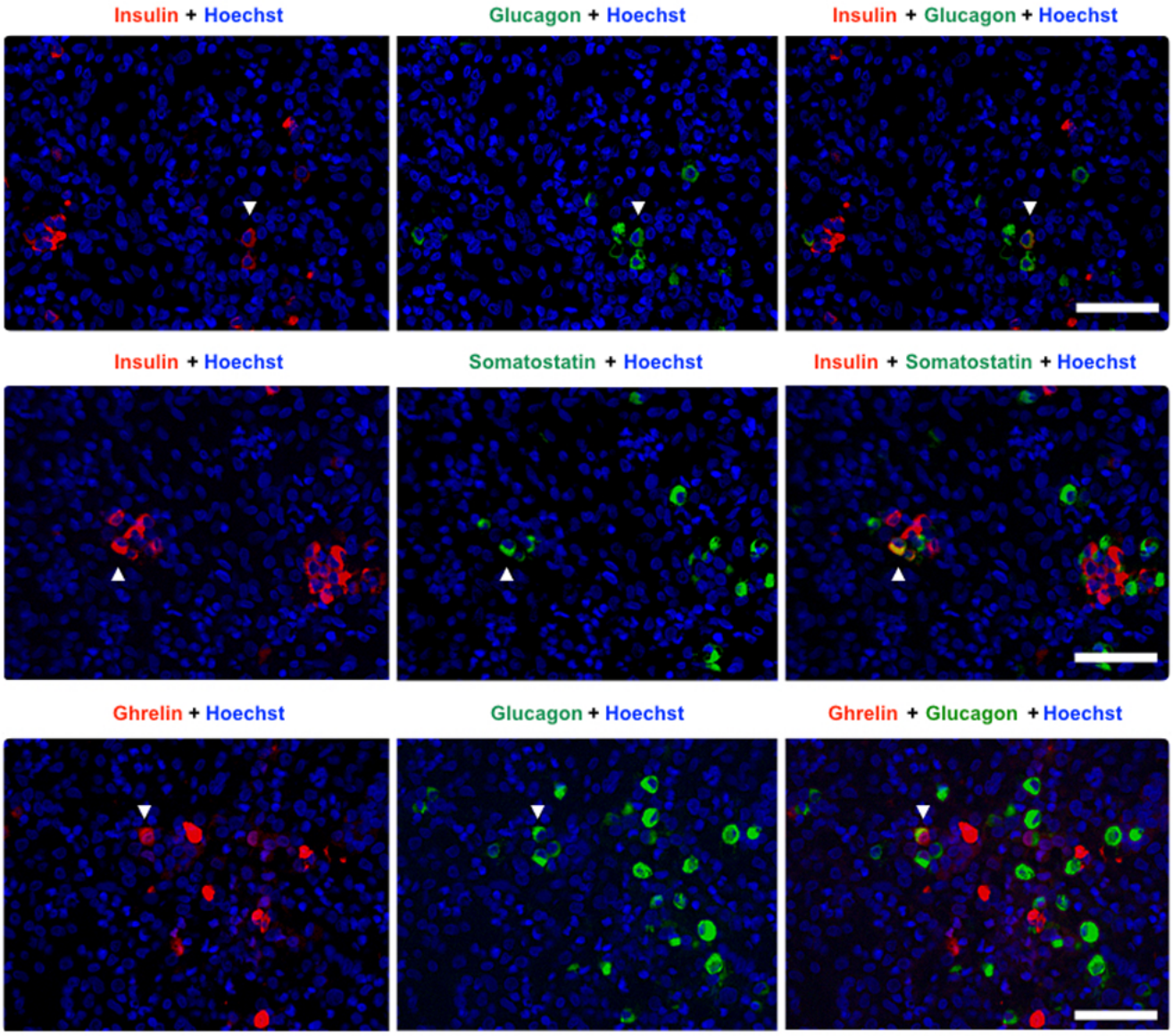
Supplementary Figure 4: Unsupervised hierarchical clustering of single-cell qPCR data from sorted human fetal pancreatic cells
(A) Unsupervised hierarchical clustering of sorted human fetal pancreatic cells at 9WD, showing all genes for populations B_{CD133-}, C, D_{HI} and D_{LO}. **(B)** Pirate plots showing the expression levels of specific genes in the four sorted populations. Each plot is a combination of box plots (25%-75%), means, violin plots and individual data points.



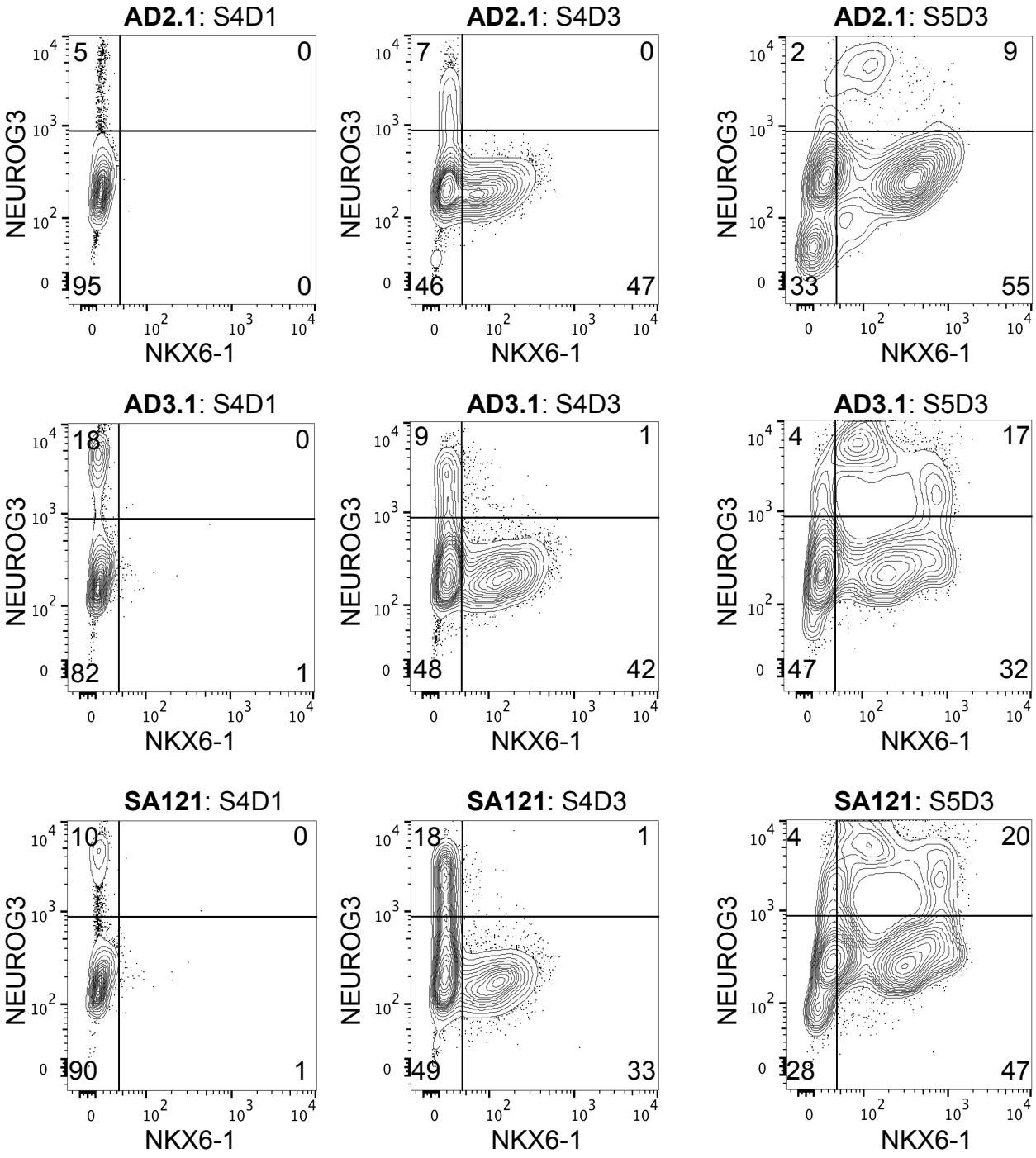
Supplementary Figure 5: Extended gene expression profiling of individual human fetal pancreas cells. t-SNE plots (corresponding to Fig. 5B) colored according to gene expression level of the indicated genes.



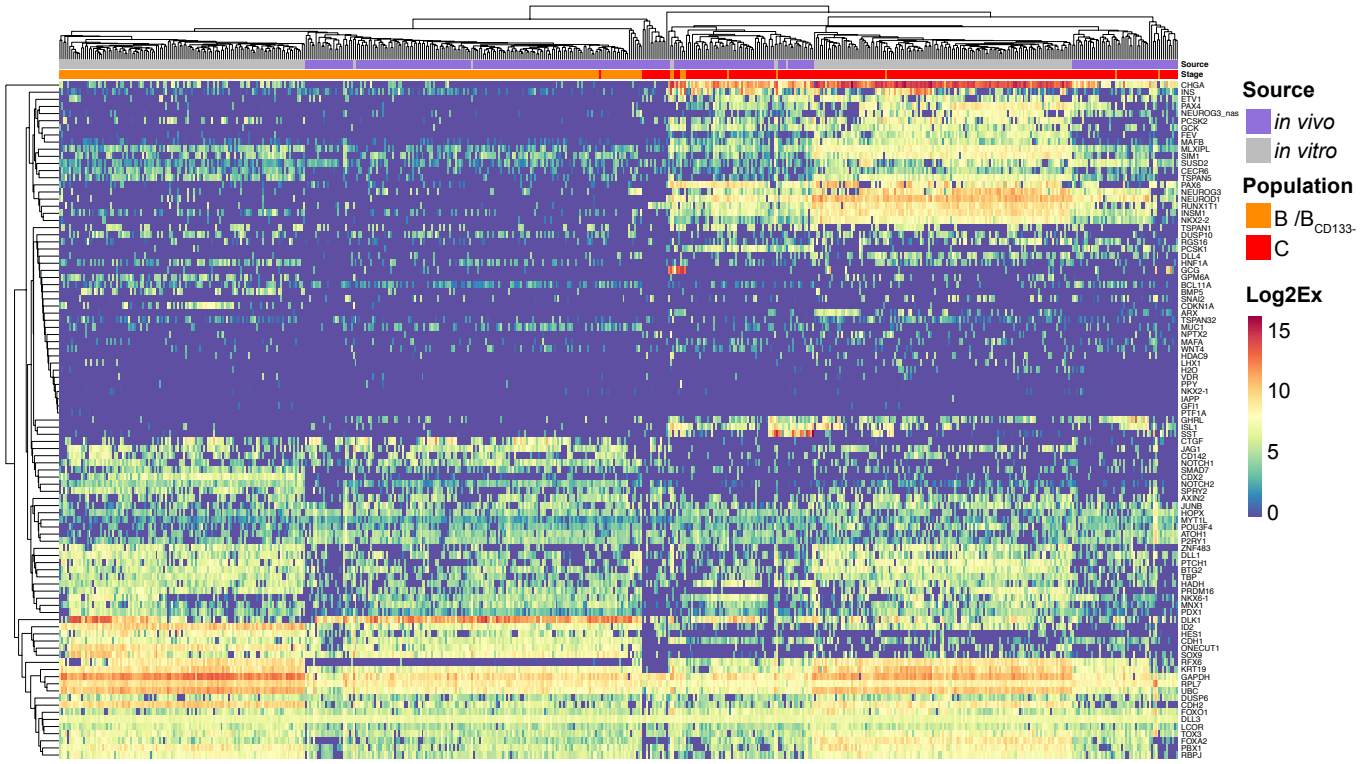
Supplementary Figure 6: Pseudotemporal ordering of single-cell gene expression data from human fetal pancreas. (A) Developmental trajectory of human fetal pancreatic cells (corresponding to Fig. 5C) colored by states. **(B)** Gene expression plots showing the pseudotemporal development of key genes involved in pancreas development. Gene expression level is shown on the y-axis; pseudotime on the x-axis. Each data point represents a single cell and is colored according to state on the trajectory shown in A. **(C)** Developmental trajectory of human fetal pancreatic cells (corresponding to Fig. 5C) colored by gene expression levels of selected hormonal genes. **(D)** Heat map showing pseudotemporal development of gene expression for the two cell fates derived from branching point 1 on the developmental trajectory shown in A. Cells at this branching point differentiates towards either State 6 or State 7.



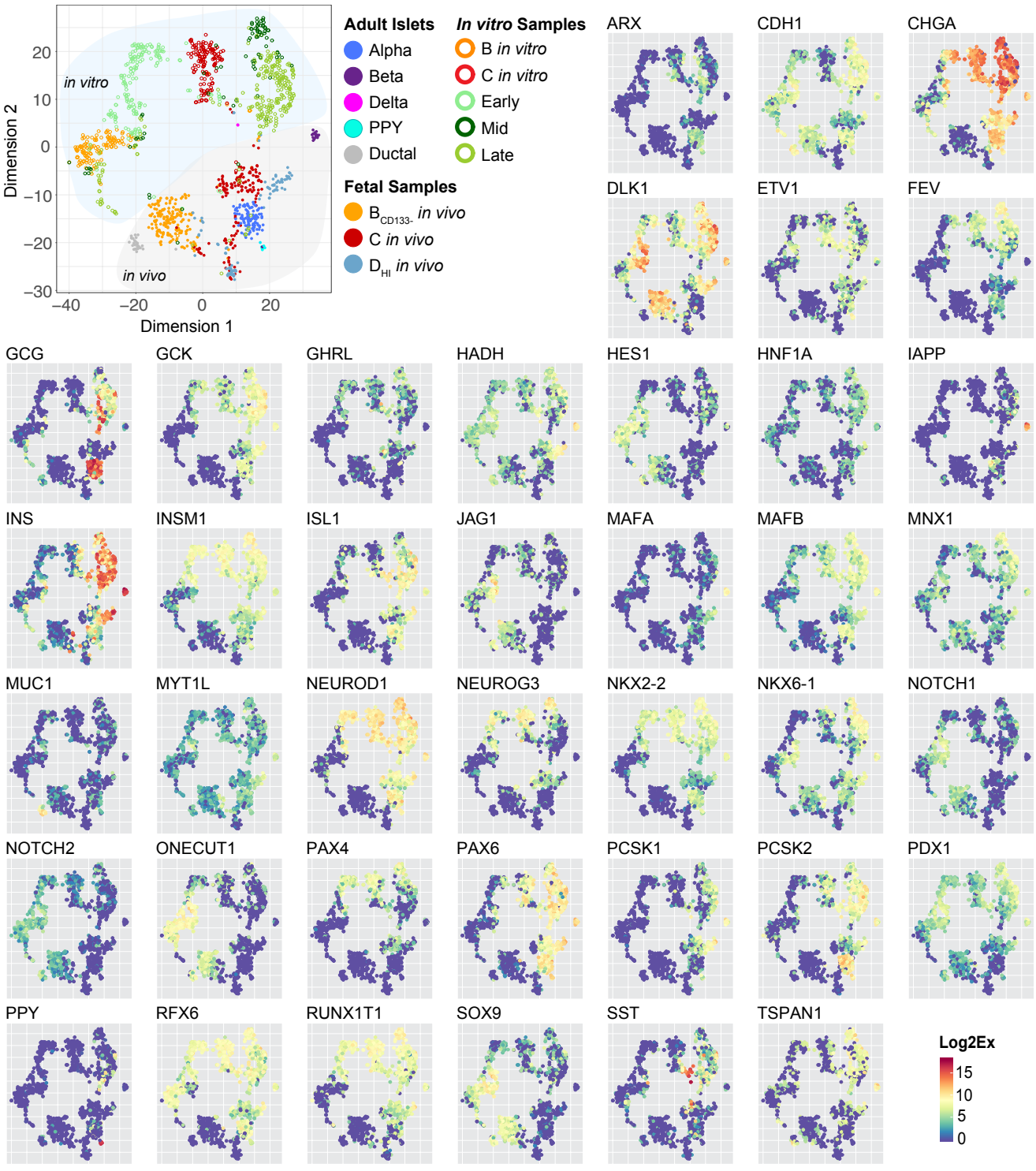
Supplementary Figure 7: Co-expression of pancreatic hormones in human fetal pancreas. Immunofluoresence staining for insulin, glucagon, somatostatin and ghrelin on pancreatic section at 10WD. Scale bar: 50µm. Arrowheads indicate double hormone positive cells.



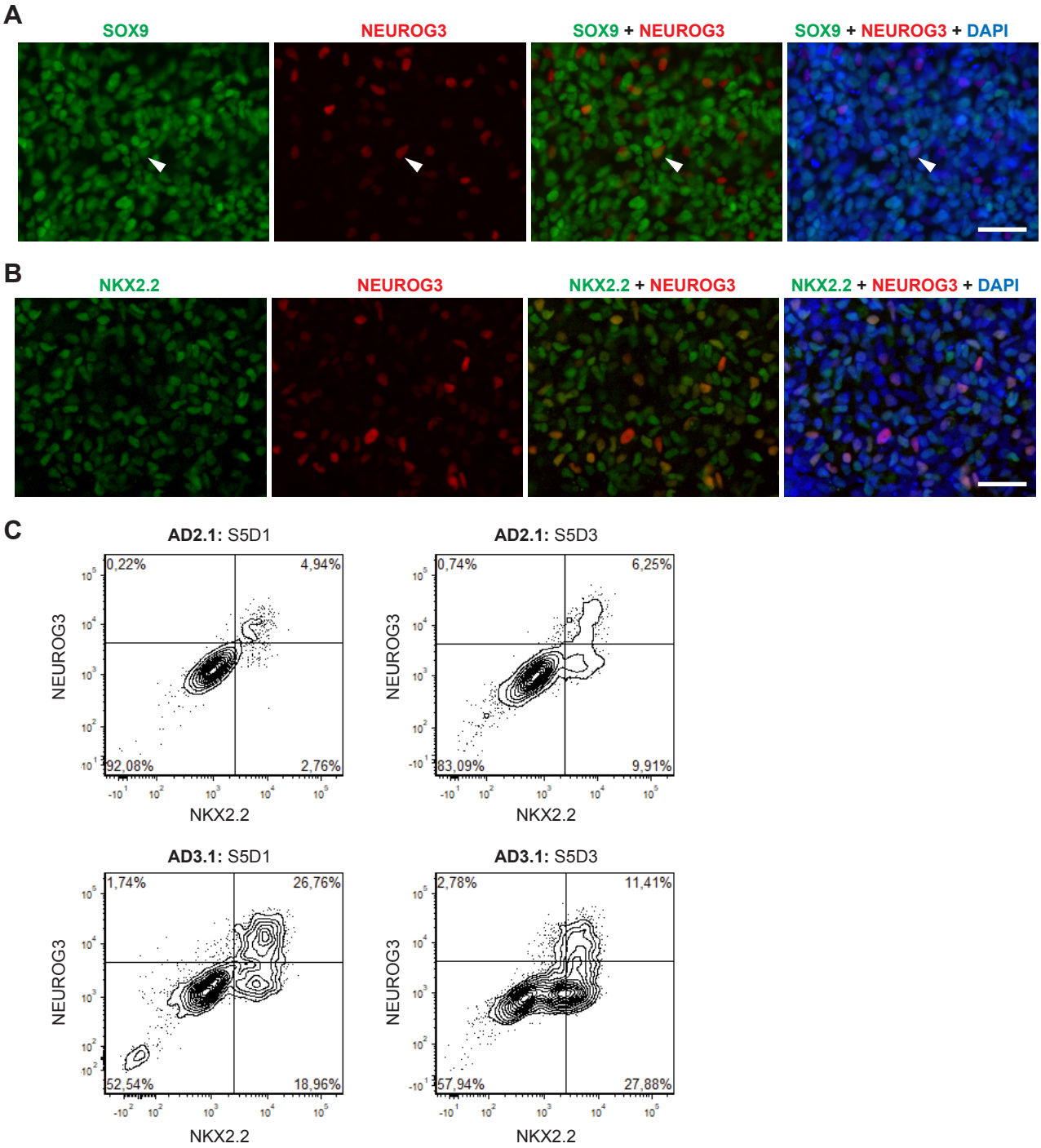
Supplementary Figure 8: Generation of endocrine progenitors from hPSCs. Representative flow cytometry plots showing the percentage of cells expressing NEUROG3 and NKX6-1 protein at S4D1, S4D3 and S5D3 of the differentiation protocol for each of the hPSC-lines SA121, AD2.1 and AD3.1 used for sorting based on cell-surface markers and subsequent single-cell qPCR analysis.



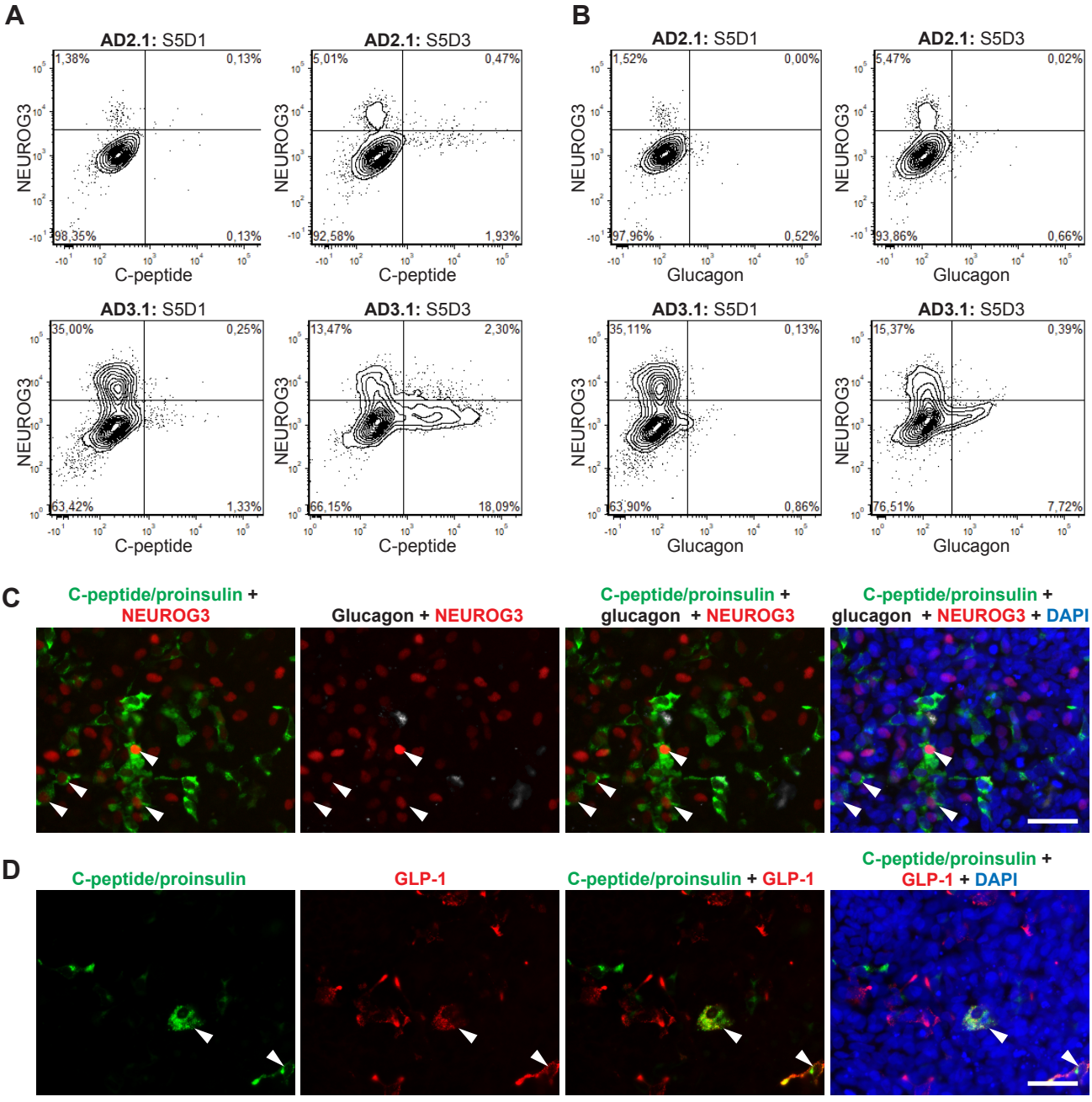
Supplementary Figure 9: Extended comparison of the single-cell expression profile of pancreatic cells generated *in vitro* to *in vivo* fetal and adult pancreas Hierarchical clustering of single-cell qPCR data showing the expression of all genes for populations B and C sorted from human fetal pancreas ("*in vivo*") or hPSC-derived pancreatic cells ("*in vitro*"). Refers to Fig. 7A and B.



Supplementary Figure 10: Gene expression in single pancreatic cells from human fetal pancreas, adult human islets or generated *in vitro* from hPSCs. t-SNE plots corresponding to Fig. 7D colored according to gene expression level of the indicated genes.



Supplementary Figure 11: Characterization of SOX9 and NKX2-2 protein co-expression in hPSC-derived endocrine progenitor subpopulations. (A) Representative image of co-staining for SOX9 (green) and NEUROG3 (red) at S5D1 of in vitro differentiated cultures (hiPSC line AD2.1). Arrowhead indicates a SOX9+/NEUROG3+ cell. Nuclei are counterstained with DAPI. Scale bar: 50 µm. **(B)** Representative image of co-staining for NKX2-2 (green) and NEUROG3 (red) at S5D1 of in vitro differentiated cultures (hiPSC line AD3.1). Nuclei are counterstained with DAPI. Scale bar: 50 µm. **(C)** Flow cytometry analysis for co-expression of NKX2-2 and NEUROG3 at S5D1 and S5D3 of in vitro differentiated cultures (hiPSC lines AD2.1 is shown in the top panel and AD3.1 in the bottom panel).



Supplementary Figure 12: Characterization of hormone expression at the protein level in hPSC-derived endocrine progenitor subpopulations. (A) Flow cytometry analysis for co-expression of NEUROG3 and C-peptide or **(B)** NEUROG3 and glucagon, at S5D1 and S5D3 of in vitro differentiated cultures (hiPSC lines AD2.1 is shown in the top panel and AD3.1 in the bottom panel). **(C)** Representative image of co-staining for C-peptide/proinsulin (green), NEUROG3 (red) and glucagon (white) at S5D1 of in vitro differentiated cultures (hiPSC line AD2.1). Arrowheads indicate co-expression of C-peptide and NEUROG3. Scale bar: 50 µm. **(D)** Representative image of co-staining for C-peptide/proinsulin (green) and GLP-1 (red) at S5D1 of in vitro differentiated cultures (hiPSC line AD3.1). Arrowhead indicates co-expression of C-peptide/proinsulin and GLP1. Scale bar: 50 µm.

Table S1: List of the 1000 genes differentially expressed in populations A, B, C and D

Click here to Download Table S1

Table S2: Gene ontology functional classification

Click here to Download Table S2

Table S3: Single cell qPCR processed data (Log2 expression values), primer sequences and information on number of cells analyzed

Click here to Download Table S3