www.nature.com/cr
www.cell-research.com

Cell Research

ARTICLE

Sequential progenitor states mark the generation of pancreatic endocrine lineages in mice and humans

Xin-Xin Yu[1,2,3,4], Wei-Lin Qiu[1], Liu Yang[1,2,4], Yan-Chun Wang[5], Mao-Yang He[3,6], Dan Wang[1,2,4], Yu Zhang[1,2,4], Lin-Chen Li[2,3,4], Jing Zhang[5], Yi Wang[5] and Cheng-Ran Xu[1,2,3]

The pancreatic islet contains multiple hormone$^+$ endocrine lineages (α, β, δ, PP and ε cells), but the developmental processes that underlie endocrinogenesis are poorly understood. Here, we generated novel mouse lines and combined them with various genetic tools to enrich all types of hormone$^+$ cells for well-based deep single-cell RNA sequencing (scRNA-seq), and gene coexpression networks were extracted from the generated data for the optimization of high-throughput droplet-based scRNA-seq analyses. These analyses defined an entire endocrinogenesis pathway in which different states of endocrine progenitor (EP) cells sequentially differentiate into specific endocrine lineages in mice. Subpopulations of the EP cells at the final stage (EP4^{early} and EP4^{late}) show different potentials for distinct endocrine lineages. ε cells and an intermediate cell population were identified as distinct progenitors that independently generate both α and PP cells. Single-cell analyses were also performed to delineate the human pancreatic endocrinogenesis process. Although the developmental trajectory of pancreatic lineages is generally conserved between humans and mice, clear interspecies differences, including differences in the proportions of cell types and the regulatory networks associated with the differentiation of specific lineages, have been detected. Our findings support a model in which sequential transient progenitor cell states determine the differentiation of multiple cell lineages and provide a blueprint for directing the generation of pancreatic islets in vitro.

Cell Research (2021) 31:886–903; https://doi.org/10.1038/s41422-021-00486-w

INTRODUCTION

Pancreatic endocrine lineages are located in the organized structure of the islet of Langerhans and are composed of α, β, δ, PP and ε cells, which secrete glucagon (GCG), insulin (INS), somatostatin (SST), pancreatic polypeptide (PPY) and ghrelin (GHRL), respectively. Blood glucose homeostasis depends on the functional coordination of β cells and non-β cells within islets, and the autoimmune destruction or dysfunction of β cells leads to type-1 or type-2 diabetes, respectively. Engineered islets assembled from various endocrine lineages induced from pluripotent stem cells in vitro and from the endogenous transdifferentiation of other cell types into β cells in vivo are expected to be effective treatments for patients with diabetes.[1,2] The effective implementation of these therapeutic strategies strongly relies on our understanding of the mechanisms of endocrinogenesis in vivo. However, the precise developmental pathways that produce all endocrine lineages, particularly those found at low abundance, have not been resolved.

Previous studies have uncovered the lineage hierarchy during pancreas organogenesis in mice. *Pdx1*-expressing multipotent pancreatic progenitor cells develop into tip and bipotential trunk cells. Tip cells further differentiate into acinar cells, whereas trunk cells generate ductal or endocrine progenitor (EP) cells.[3] All endocrine lineages are sequentially generated from the Neurogenin3$^+$ (Ngn3$^+$) EP population during embryogenesis.[4–6] Notably, each Ngn3$^+$ cell is unipotent to differentiate into one endocrine cell,[7] which indicates that the differentiation potentials of EPs are heterogeneous. Consistently, recent studies using single-cell RNA sequencing (scRNA-seq) have demonstrated the transcriptomic heterogeneity of EPs during pancreas embryogenesis.[8–12] However, the biological significance of the EP states and their relationship with endocrinogenesis require further investigation.

Although the lineage hierarchy during pancreas organogenesis in humans is considered to be similar to that in mice,[13–15] interspecies discrepancies remain. For example, the order at which endocrine lineages emerge is different; the cells that emerge first are INS-expressing cells in humans but GCG-expressing cells emerge first in mice.[16,17] Of note, decoding the regulatory networks during pancreatic organogenesis in humans would directly guide in vitro generation of endocrine cells from human pluripotent stem cells. However, the elaborate process of human pancreatic lineage differentiation is poorly understood due to the scarcity of samples and the limitations of conventional methods of histological inspection.

Several studies have attempted to elucidate the programs of endocrine lineage differentiation in mice using high-throughput droplet-based scRNA-seq technologies, such as the 10× Genomics

[1]Department of Human Anatomy, Histology, and Embryology, School of Basic Medical Sciences, Peking University, Beijing 100191, China; [2]Peking-Tsinghua Center for Life Sciences, Peking University, Beijing 100871, China; [3]Ministry of Education Key Laboratory of Cell Proliferation and Differentiation, College of Life Sciences, Peking University, Beijing 100871, China; [4]Academy for Advanced Interdisciplinary Studies, Peking University, Beijing 100871, China; [5]Haidian Maternal & Child Health Hospital, Beijing 100080, China and [6]PKU-Tsinghua-NIBS Graduate Program, Peking University, Beijing 100871, China
Correspondence: Cheng-Ran Xu (cxu@pku.edu.cn)
These authors contributed equally: Xin-Xin Yu, Wei-Lin Qiu, Liu Yang, Yan-Chun Wang

Received: 18 September 2020 Accepted: 27 January 2021
Published online: 10 March 2021

platform (10×).[8–12,18–20] Unfortunately, droplet-based approaches have generally been limited to mapping the branched differentiation trajectory of α and β cells, which indicates the insufficiency of these approaches for describing the differentiation pathways of entire endocrine lineages, likely due to their relatively high levels of noise and their low sensitivity for transcripts with low abundance.[21,22] However, the detection of genes expressed at low levels might be critical for identifying the differentiation pathways of cells with slight differences at the transcriptomic level.

Here, to overcome the limitations of droplet-based approaches and define the entire developmental trajectories of pancreatic endocrine lineages, we combined three newly generated mouse lines with various additional mouse lines to enrich low-abundance EP cells and their hormone$^+$ descendants, and we then analyzed these cells using Smart-seq2[23] or modified STRT-seq (mSTRT-seq),[24,25] which are highly sensitive well-based scRNA-seq technologies. We discovered that sequential EP cell states establish developmental windows for the generation of various endocrine lineages. We then optimized the analysis of 10× data using a gene coexpression network (GCN)[26–28] extracted from the Smart-seq2 or mSTRT-seq data to more accurately classify cells, to more precisely describe the cell differentiation pathways, and to determine the temporal differentiation order of all endocrine lineages in mice and humans. Although the paths for the generation of different endocrine lineages are generally conserved between species, significant differences have been found in the proportions of EP and endocrine cells and in gene expression networks associated with cell lineage differentiation.

RESULTS

Genetically labeling all mouse endocrine lineages for scRNA-seq

To comprehensively delineate the developmental trajectory of pancreatic endocrine lineages in mice, we performed scRNA-seq analysis of EP cells and hormone$^+$ lineages from embryonic day (E) 13.5 to postnatal day (P) 3 using the Smart-seq2 method (Fig. 1a). For the enrichment of low-abundance EP cells and their immediate descendant endocrine lineages, we treated *Ngn3-Cre^{ER}*; *Rosa26-RFP*[5] mice with a single pulse of tamoxifen for 1 or 2 days and then collected RFP$^+$ cells from E13.5–E18.5 pancreata by fluorescence-activated cell sorting (FACS) (Fig. 1a; Supplementary information, Fig. S1a).

For the enrichment of low-abundance differentiated ε, δ and PP cells, we generated fluorescently-labeled knock-in *Ghrl-P2A-CFP*, *Sst-P2A-BFP* and *Ppy-P2A-mNeptune* mouse lines (Fig. 1b). These mouse lines were verified by immunofluorescence (IF) of embryonic, neonatal or adult pancreata, which showed that the endogenous fluorescent signals overlapped with the corresponding hormone molecules (Fig. 1c; Supplementary information, Fig. S1b), and by ELISA, which showed no notable differences in SST and PPY secretion into blood serum between the wild-type and genetically-modified adult mice (Supplementary information, Fig. S1c). Together, these results indicated that these mouse lines were suitable for the efficient labeling of GHRL$^+$, SST$^+$ and PPY$^+$ cells. Flow cytometry analyses revealed that few GHRL$^+$, SST$^+$ and PPY$^+$ cells originated at E13.5, E14.5 and E16.5, respectively (Supplementary information, Fig. S1d–f). Moreover, we observed that the percentage of GHRL$^+$ cells markedly decreased after birth, which confirmed previously reported findings[29,30] (Supplementary information, Fig. S1d). We then sorted ε, δ and PP cells at various developmental time points ranging from the earliest generation time to P3 (Fig. 1a; Supplementary information, Fig. S1d–f). In addition, we included single-cell datasets of the second wave of EP and endocrine cells from the *Pdx1-Cre;Rosa26-RFP*, *Ngn3-GFP*, *Ins1-RFP*, *Gcg-Cre;Rosa26-RFP* and *Gcg-P2A-GFP* lines generated in our previous studies[10,31] and generated new datasets from Gcg-P2A-GFP$^+$ pancreata at E17.5 and P3 (Fig. 1a; Supplementary information, Fig. S2a).

During early development, the pancreas is separately generated from dorsal and ventral endoderm domains, and these domains eventually fuse.[32] The dorsal pancreas (DP) contributes most of the pancreatic mass, whereas the ventral pancreas (VP) develops into the head and the uncinate process of the pancreas. Notably, consistent with previous findings, the percentage of Ppy-P2A-mNeptune$^+$ cells in the VP was significantly greater than that in the DP,[33,34] whereas the percentage of Gcg-P2A-GFP$^+$ cells exhibited the opposite pattern (Supplementary information, Fig. S1g). To identify the existence of differences between the endocrinogenesis pathways of the DP and VP, we selectively collected Ngn3-Cre^{ER};Rosa26-RFP$^+$ and Ppy-P2A-mNeptune$^+$ cells and other genetically-labeled cells from the DP and VP (Supplementary information, Fig. S2a).

Single cells at each time point were collected from multiple pooled pancreata as indicated in Supplementary information, Fig. S2a. In total, 3878 cells passed the quality control tests, and an average of one million mapped reads and 6500 genes were detected from each cell (Supplementary information, Fig. S2b, c and Table S1).

Identification of major cell types during mouse endocrinogenesis

After t-distributed stochastic neighbor-embedding (t-SNE) analysis, we excluded contaminating acinar cells (*Rbpjl*$^+$), trunk/ductal cells (*Hes1*$^+$*Ngn3*$^-$), mesenchymal cells (*Col3a1*$^+$) and immune cells (*Ptprc*$^+$) (Supplementary information, Fig. S2d, e). The retained 3730 pancreatic endocrine cells expressing *Ngn3* or *Neurod1* were classified according to the expression patterns of genes identified from a GCN analysis (Materials and Methods; Supplementary information, Fig. S2f), which was based on the assumption that cells of the same type will display strongly correlated gene expression profiles and thereby coexpress the same set of genes.[26–28] Although multihormone-expressing cells were identified, most of the cells were similar to monohormone-expressing cells at the transcriptomic level and were thus classified into certain monohormone-expressing cell clusters (Supplementary information, Fig. S2f, g). However, a few *Ins*$^+$*Ppy*$^+$ cells, which were mainly enriched by the sorting of *Ppy*$^+$ cells from the VP, were sparsely distributed on the t-SNE plot and presented specific transcriptomic profiles in comparison with those of *Ins*$^+$ or *Ppy*$^+$ monohormone cells (Supplementary information, Fig. S2f, g and Tables S1, S3). Because this small group of *Ins*$^+$*Ppy*$^+$ cells negligibly contributed to the endocrine population and might interfere with a subsequent developmental trajectory analysis, we excluded these cells from further analyses. We ultimately identified 10 major cell types, and each cell type consisted of cells from various sources (Fig. 1d–f; Supplementary information, Table S1).

The cluster-1–4 cells expressed a series of transcription factors (TFs), including *Ngn3*, *Fev*, *Pax4* and *Arx*, in a cascade manner and comprised most of the Ngn3-GFP$^+$ cells (Fig. 1d–f; Supplementary information, Fig. S2h). Consistent with our previous study,[10] these clusters represented the sequential developmental stages of EP cells with stage-specific TFs, namely, EP1–4, which indicated that cells at different stages exhibit unique cell states. The cluster-5 cells expressed *Arx*, *Neurod1* and *Pax6*, but no hormone genes (Fig. 1d–f; Supplementary information, Fig. S2h), which indicated that these cells were intermediate progenitors of certain endocrine lineages. The cluster-6–10 cells primarily expressed hormone genes and were clearly identified as ε (cluster-6), β (cluster-7), δ (cluster-8), α (cluster-9) and PP (cluster-10) cells (Fig. 1d–f; Supplementary information, Fig. S2h). Notably, Ngn3-Cre^{ER};Rosa26-RFP$^+$ cells appeared in each cell population on the t-SNE plot (Fig. 1e), which indicated the high efficacy of our genetic tracing system; the hormone$^+$ cells from the *Ngn3-Cre^{ER};Rosa26-RFP* and other reporter lines were intermingled together, which demonstrated that these reporter lines were suitable for labeling each endocrine lineage (Fig. 1e; Supplementary information,

a

Smart-seq2 scRNA-seq: 3,878 mouse cells

Cell source	E13.5	E14.5	E15.5	E16.5	E17.5	E18.5	P0	P3
① Pdx1-Cre;Rosa26-RFP+	38*	68*	19*	-	-	-	-	-
② Ngn3-GFP+	154*,**	126*	55*	108*	115*	-	-	-
③ Ngn3-Cre^{ER};Rosa26-RFP+	99/-	138/235	140/94	211/-	227/111	184/-	-	-
④ Ins1-RFP+	-	-	39*	-	61***	-	84***	88***
⑤ Gcg-Cre;Rosa26-RFP+	-	-	-	-	69***	-	51***	-
⑥ Gcg-P2A-GFP+	-	-	33*	-	50	-	-	62
⑦ Sst-P2A-BFP+	-	5	141	116	132	-	84	117
⑧ Ppy-P2A-mNeptune+	-	-	-	54	73	50	61	69
⑨ Ghrl-P2A-CFP+	4	54	45	59	54	43	35	23

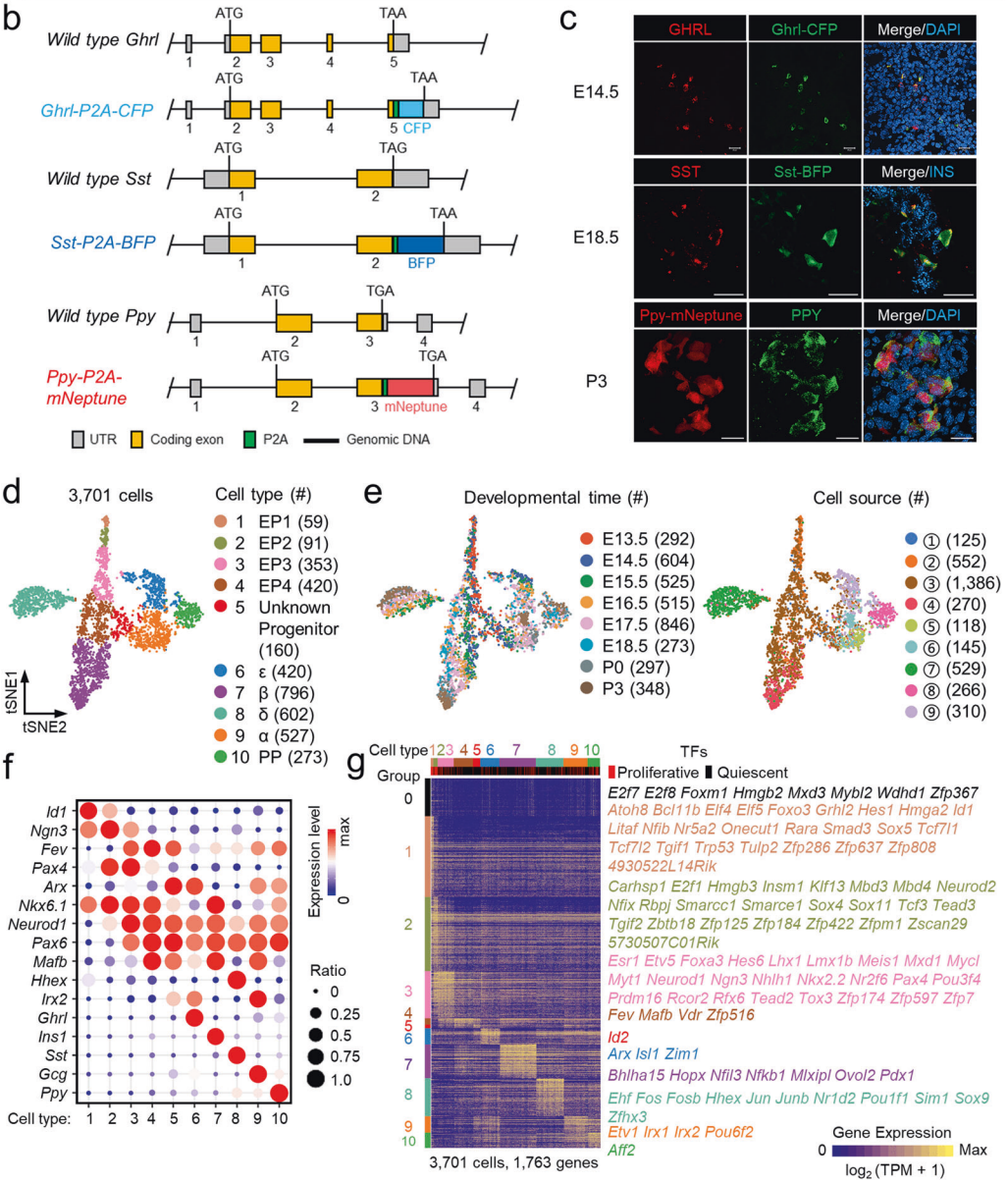

Fig. 1 Identification of cell types present during mouse endocrinogenesis by Smart-seq2 scRNA-seq analysis. a Overview of 3878 pancreatic cells from the analysis of Smart-seq2 scRNA-seq data obtained in this study. The numbers indicate the cell counts obtained from various mouse lines (cell sources) labeled with circled numbers at different developmental times. The cell numbers before and after the slashes indicate Ngn3-Cre^{ER};Rosa26-RFP^{+} cells from mice with a single tamoxifen injection 1 and 2 days before harvest, respectively. *, ** and *** indicate the GEO datasets GSE115931, GSE84324 and GSE87375, respectively. **b** Generation of *Ghrl-P2A-CFP*, *Sst-P2A-BFP* and *Ppy-P2A-mNeptune* strains by inserting P2A and fluorescent protein DNA sequences before the stop codon. P2A is a self-cleaving peptide that enables the fluorescence downstream of P2A to be exploited as a marker of endogenous hormone genes. **c** IF staining of GHRL in *Ghrl-P2A-CFP* pancreas at E14.5, SST in *Sst-P2A-BFP* pancreas at E18.5, and PPY in *Ppy-P2A-mNeptune* pancreas at P3. Scale bars, 20 μm. **d** t-SNE plot showing 10 distinct cell types. Each dot represents a single cell. **e** t-SNE plots showing the developmental time (left) and cell source (right) of the cells in **d**. The circled numbers are the same as those in **a**. **f** Average expression levels of canonical markers for each cell type. The color of each dot represents the average expression levels from low (blue) to high (red), and the size of each dot represents the ratio of the positive cells for each gene. **g** Heatmap of cell type-featured genes. Single cells are displayed in columns, and genes are displayed in rows. The colors ranging from blue to yellow indicate low to high levels of relative gene expression. Cell cycle-related genes were extracted as group 0.

Table S1). Moreover, the same type of cells from the DP and VP did not show significant transcriptomic differences and were intermingled on the t-SNE plot (Supplementary information, Fig. S2i, j). After excluding cell cycle-related genes (group 0), we performed a differential expression analysis to identify cell type-featured genes (groups 1–10) (Fig. 1g; Supplementary information, Table S2).

To obtain an unbiased and comprehensive distribution of EP cells and endocrine lineages, we performed a high-throughput scRNA-seq analysis using the 10× Genomics platform to assess the DP and VP at E16.5–E18.5 and integrated published 10× datasets of pancreata at E12.5–E18.5[8,10,12,20] (Supplementary information, Fig. S3a). Among the 103,400 cells that passed quality control, we detected an average of 12,000 unique molecular identifiers (UMIs) and 3000 genes in each cell (Supplementary information, Fig. S3b, c and Table S1). After excluding doublets, including multihormonal cells that might be caused by contamination with doublets or cell debris, cells that expressed nonendocrine markers, and the first wave of α cells,[10] we identified 34,129 cells as EP cells and endocrine lineages based on the expression of marker genes (Materials and Methods; Supplementary information, Fig. S3d–h and Table S1). Our 10× analyses classified ten major cell types (Supplementary information, Fig. S3i, j) that corresponded to those identified in our Smart-seq2 analyses but displayed a lower number of cell type-featured genes, although most of the genes identified by the 10× analysis overlapped with those identified by Smart-seq2 technology (Fig. 1g; Supplementary information, Fig. S3k, l and Table S2). Taken together, the results indicate that these two scRNA-seq methods independently identified ten major cell types present during mouse pancreatic endocrinogenesis.

Identification of endocrine lineage allocation pathways

To determine the developmental trajectory of the endocrine lineages, we performed a three-dimensional (3D) force-directed layout (FDL) analysis of all 3701 Smart-seq2 single-cell datasets (Fig. 2a; Supplementary information, Fig. S4a). On the 3D FDL plot, the EP1 population, which developed along the EP1–EP4 trajectory, was considered as the starting point for endocrinogenesis.[10] Curiously, the EP3 population represented the first branching node, which led to the development of EP4 and ε cells (Fig. 2a). Subsequently, ε cells underwent a developmental trajectory that eventually branched into α and PP cells (Fig. 2a). Because both ε cells and cluster-5 cells were directly upstream of α and PP cells on their developmental trajectories, we tentatively designated ε cells as α/PP-progenitor-I (α/PP-Pro-I) and the cluster-5 population as α/PP-Pro-II (Fig. 2a).

Surprisingly, EP4 cells were divided into two branches on the trajectory, and each branch was linked to a subgroup of β cells (Fig. 2a, b). One branch primarily consisted of E13.5–E15.5 cells, whereas the other mainly included E16.5–P3 cells; therefore, the two branches were designated as the EP4-β^{early} and EP4-β^{late} branches, respectively (Fig. 2a, b). This separation might be associated with a temporal shift in the chromatin states between early and late EP cells.[9] A comparative analysis of $EP4^{early}$ and $EP4^{late}$ cells identified 114 genes that were highly expressed in $EP4^{early}$ cells, including *Arx*, which encodes a critical TF for α-cell differentiation[35] (Fig. 2c; Supplementary information, Tables S3, S4), and 40 genes that were up-regulated in $EP4^{late}$ cells, including *Hhex*, which encodes a critical TF for δ-cell differentiation[36] (Fig. 2c; Supplementary information, Tables S3, S4). The expression patterns of *Arx* and *Hhex* in $EP4^{early}$ and $EP4^{late}$ cells were verified by single-cell RT-qPCR (scRT-qPCR) (Fig. 2d). Moreover, we found that α/PP-Pro-II cells were continuously generated along the entire developmental path of $EP4^{early}$ cells and served as a node for the generation of α and PP cells (Fig. 2a), whereas the δ-cell branch was generated from a node located in the $EP4^{late}$ pathway and adjacent to EP3 cells (Fig. 2a). Altogether, these findings suggest that $EP4^{early}$ and $EP4^{late}$ cells have distinct developmental potentials. A comparative analysis of β^{early} and β^{late} cells revealed that β^{early} cells highly expressed *Pax4*, which regulates the β-over-α cell fate choice,[35] whereas β^{late} cells highly expressed genes related to metabolic processes, as indicated by gene ontology analysis (Supplementary information, Fig. S4b, c and Tables S3, S4). In addition, an RNA velocity analysis[37] showed that β^{early} cells tended to develop into β^{late} cells (Fig. 2b). All these data suggest that β^{late} cells are in a more mature state in comparison with β^{early} cells.

To examine the temporal order of the emergence of various endocrine lineages during endocrinogenesis, we assessed *Ngn3-Cre^{ER};Rosa26-RFP^{+}* cells, which were collected from fetal pancreata exposed to a single-pulse injection of tamoxifen one day before harvest, at different developmental stages on the developmental pathway (Fig. 2e). EP1–3, $EP4^{early}$, α/PP-Pro-II cells and small fractions of α/PP-Pro-I (ε), α and β^{early} cells appeared at E13.5. At E14.5–E15.5, the proportions of α, β and α/PP-Pro-I (ε) cells increased. Starting at E16.5, the proportion of EP1–3 cells decreased, and $EP4^{early}$ cells were replaced by $EP4^{late}$ cells; these changes were associated with the generation of δ and β^{late} cells. We observed that δ-cell production occurred primarily after E15.5, although a few δ cells were detected earlier (Fig. 2e; Supplementary information, Fig. S2a). Additionally, starting at E16.5, the proportion of newly generated α/PP-Pro-II cells was greatly reduced, and PP cells began to appear (Fig. 2e; Supplementary information, Fig. S2a). Therefore, these analyses revealed the temporal order of the generation of hormone^{+} cells during endocrinogenesis.

To perform an unbiased analysis of the cell lineage components and pancreatic lineage differentiation pathways, we analyzed 10× scRNA-seq datasets of endocrine cells from E12.5 to E18.5 (Supplementary information, Fig. S3a). However, using the standard analytic pipeline for 10× datasets (Materials and Methods), we found that the clarity of the differentiation pathways was limited by a significant number of scattered cells nearby (Fig. 2f), which was mainly due to the limitation of the 10× method in discriminating informative transcripts (signal) from uninformative transcripts (noise).[21] Each gene in the GCN is linked to many genes, and genes in the GCN can reasonably be considered informative transcripts in their respective cell types. Therefore, we applied the genes in the endocrine lineage-related GCNs derived from Smart-seq2 data as "informative transcripts" to reanalyze the 10× data. Surprisingly, after adjusting the cell classification (Supplementary information, Fig. S3m), we clearly defined endocrine lineage differentiation pathways similar to those found in our Smart-seq2 analyses (Fig. 2g). We subsequently presented cells at different developmental stages in the pathway and found that the temporal order of endocrinogenesis was similar to that observed in our Smart-seq2 analyses (Fig. 2h). However, the EP4-β^{early} and EP4-β^{late} branches could not be defined using the 10× method, although the EP4 population displayed biased expression of *Arx* and *Hhex* at the earlier and later developmental time points, respectively (Fig. 2h, i), which indicated that the 10× method lacks the sensitivity necessary to distinguish EP4 subpopulations. Notably, PP cells clearly branched out from the nodes of the α/PP-Pro-I (ε) and α/PP-Pro-II populations on the FDL plot starting at E16.5 (Fig. 2h). Although the VP had a higher proportion of PP cells and a lower proportion of α cells in comparison with the DP, the developmental trajectories of the VP and DP were identical (Fig. 2h; Supplementary information, Fig. S1g).

Therefore, by combining our analyses of Smart-seq2 and 10× data, we deciphered the precise temporal differentiation pathways of all pancreatic endocrine lineages (Fig. 2j).

Genetic tracing to verify the temporal order of the branched paths

Our new model showed that the EP states mark the locations of branch nodes where various endocrine lineages are generated along the endocrinogenesis pathway (Fig. 2j). A single-cell

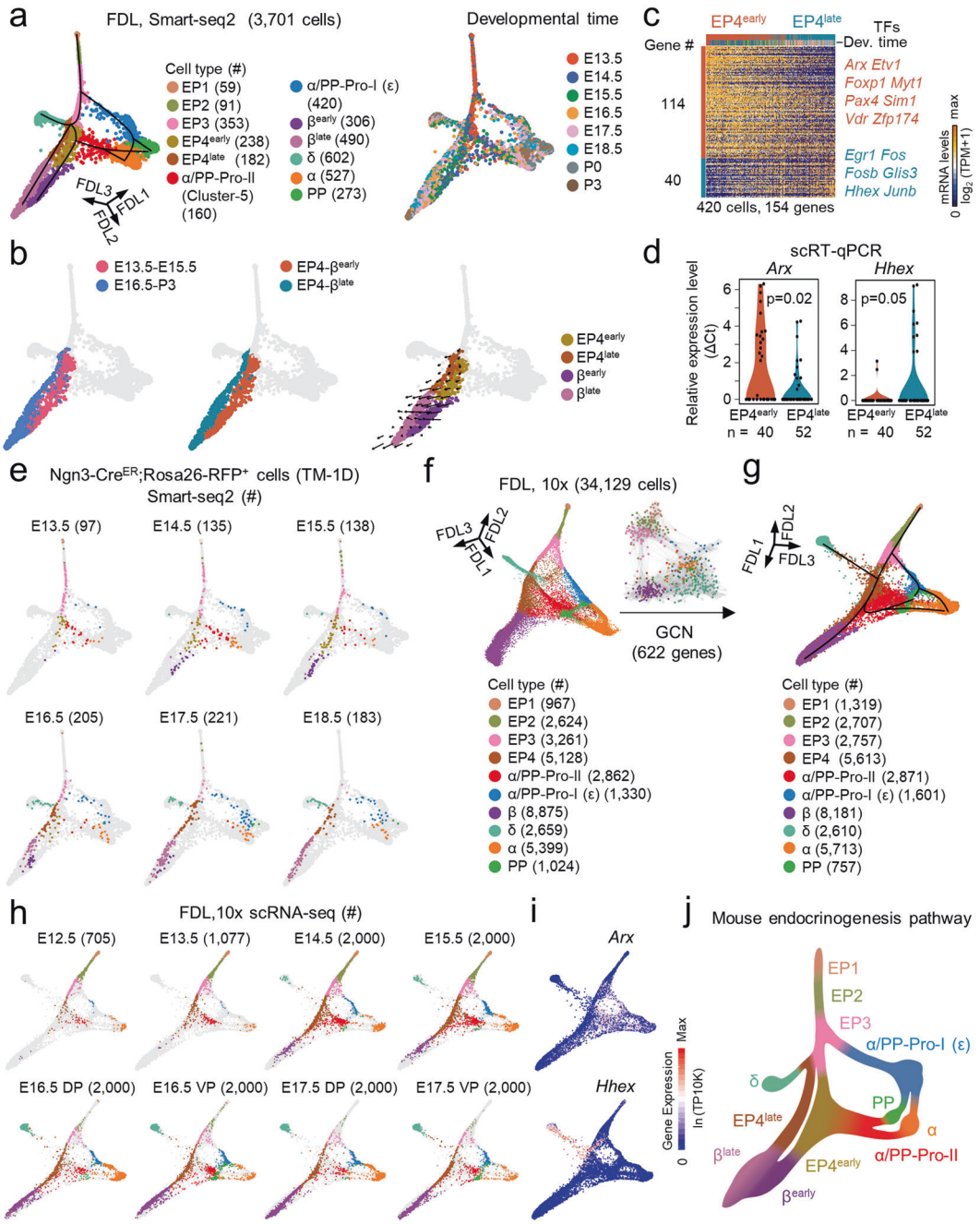

Fig. 2 Identification of allocation pathways for mouse endocrine lineages by scRNA-seq analyses. a 3D FDL analysis of EP and endocrine cells based on Smart-seq2 scRNA-seq data. Each dot represents a single cell. The cell types (left) and developmental time (right) are color coded. The curves indicate the pathways of endocrine lineage development. **b** The EP4 and β cells at E13.5–E15.5 and E16.5–P3 on the 3D FDL from **a** (left) and the EP4-β^{early} and EP4-β^{late} cell branches on the 3D FDL from **a** (middle). The arrows on the FDL plot indicate the developmental direction estimated by RNA velocity analysis (right). Each dot represents a single cell. **c** Heatmap of differentially expressed genes between $EP4^{early}$ and $EP4^{late}$ cells. Single cells are displayed in columns, and genes are displayed in rows. **d** The expression levels of *Arx* and *Hhex* were verified by scRT-qPCR with normalization to *Gapdh* expression. *n*, cell counts. The dots represent cells. Wilcoxon rank-sum test. **e** Temporal distribution of the Ngn3-Cre^{ER};Rosa26-RFP^{+} cells at E13.5–E18.5 from mice with a single tamoxifen injection 1 day (TM-1D) before harvest on the pathways from **a**. The colors denote the cell types in **a**. **f**, **g** 3D FDL analysis of EP and endocrine cells using 10× scRNA-seq data through application of the top 2000 highly variable genes in the 10× dataset (**f**) or the 622 genes in the GCN derived from the Smart-seq2 dataset (**g**). The curves in **g** indicate the pathways of endocrine lineage development. **h** Temporal distribution of single cells generated by 10× scRNA-seq using the pathways from **g**. The colors denote the cell types in **g**. **i** *Arx* and *Hhex* expression levels projected onto the 3D FDL plot from **g**. The color of each dot represents the expression level from low (blue) to high (red). **j** Summary of temporal endocrine lineage allocation pathways in mice.

quantitative analysis revealed that *Ngn3* expression increased at EP1, peaked at EP2, decreased at EP3, and ceased at EP4[10] (Fig. 1f; Supplementary information, Figs. S2h and S4a). On the 3D FDL plot, α/PP-Pro-I (ε) cells adjacent to EP3 cells highly expressed *Ngn3*, which indicated that ε cells were derived from EP3 cells rather than EP4 cells (Figs. 2a and 3a). In addition, GFP fluorescence persisted after *Ngn3* expression is turned off in the *Ngn3-GFP* fetal pancreas, which allowed us to track the immediate

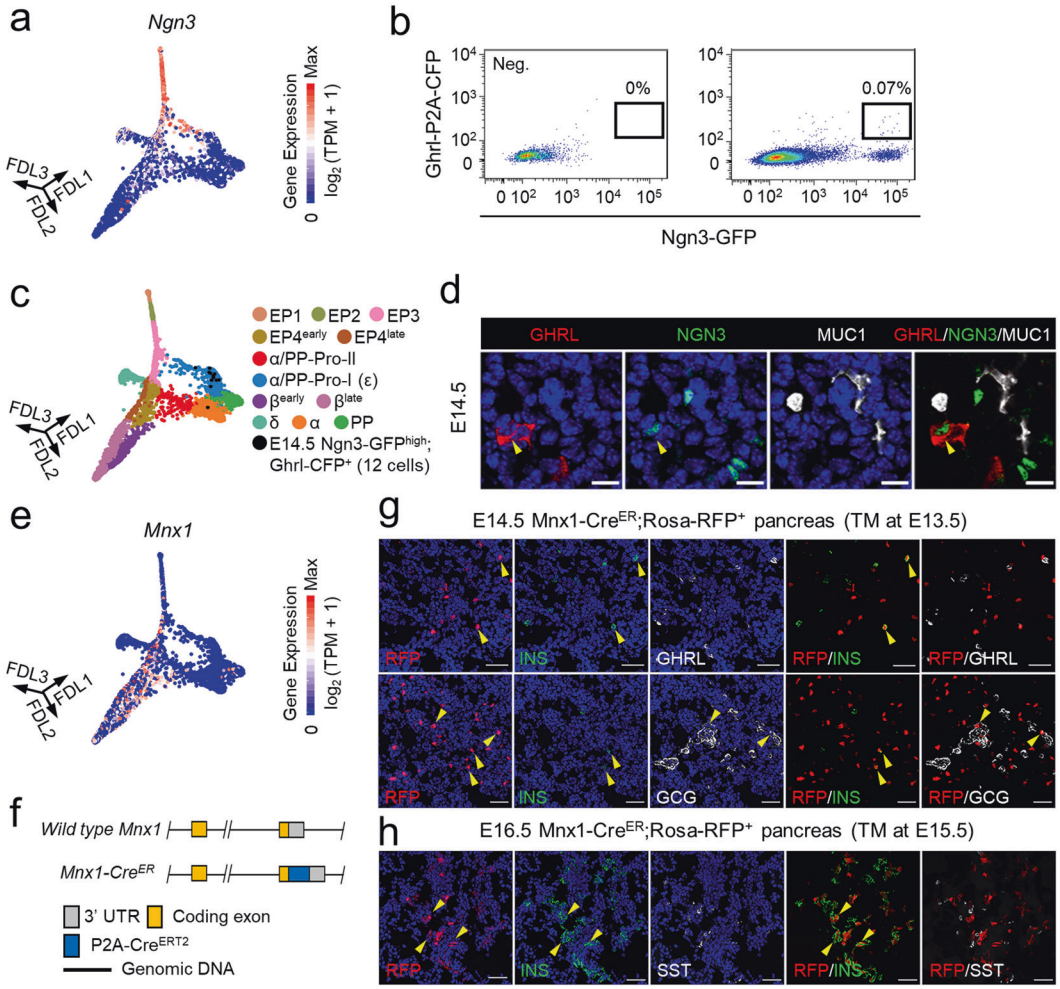

Fig. 3 Genetic tracing to verify the temporal order of the mouse endocrinogenesis pathways. a Expression level of *Ngn3* projected onto the 3D FDL plot in Fig. 2a. The color of each dot represents the expression level from low (blue) to high (red). **b** FACS gating strategies for purifying Ngn3-GFP^{high};Ghrl-CFP^{+} cells at E14.5. Pancreatic tissues from wild-type embryos were used as negative controls (left panel). **c** Ngn3-GFP^{high};Ghrl-CFP^{+} single cells at E14.5 projected onto the 3D FDL plot in Fig. 2a. **d** DAPI and IF staining of GHRL, NGN3 and MUC1 in paraffin sections of pancreas at E14.5. Scale bars, 10 μm. The arrowhead indicates a GHRL^{+}NGN3^{+} cell. **e** Expression level of *Mnx1* projected onto the 3D FDL plot in Fig. 2a. **f** Strategy for the generation of the *Mnx1-Cre^{ER}* strain by inserting P2A and CreERT2 sequences before the stop codon. **g** DAPI, RFP and IF staining of INS, GHRL and GCG in frozen sections of Mnx1-Cre^{ER};Rosa26-RFP^{+} pancreas at E14.5 with a tamoxifen (TM) injection at E13.5. Scale bars, 40 μm. The arrowheads indicate INS^{+}RFP^{+} or GCG^{+}RFP^{+} cells. **h** DAPI, RFP and IF staining of INS and SST in frozen sections of Mnx1-Cre^{ER};Rosa26-RFP^{+} pancreas at E16.5 with a TM injection at E15.5. Scale bars, 40 μm. The arrowheads indicate INS^{+}RFP^{+} cells.

progeny of cells with high *Ngn3* expression. *Ngn3-GFP* and *Ghrl-CFP* mouse strains were crossed to generate Ngn3-GFP^{high};Ghrl-CFP^{+} cells for Smart-seq2 scRNA-seq (Fig. 3b). When projected onto the 3D FDL plot, these cells were located on the α/PP-Pro-I cell developmental pathway (Fig. 3c). In addition, IF of NGN3 and GHRL in E14.5 pancreas confirmed the existence of NGN3^{+}GHRL^{+} cells (Fig. 3d). These genetic tracing results demonstrate that EP3 cells with a high level of *Ngn3* expression have the potential to develop into α/PP-Pro-I (ε) cells (Fig. 2a, j).

Our previous study found that EP4 cells specifically expressed a group of genes distinct from those expressed by EP1–EP3 cells, including *Mnx1*[10] (Fig. 3e), which is required for β cell generation and fate maintenance.[38] However, the fate of *Mnx1*-expressing EP cells has not been traced. We crossed a new *Mnx1-Cre^{ER}* mouse strain (Fig. 3f) with the *Rosa26-RFP* strain. One day after a single injection of tamoxifen at E13.5 or E15.5, the pancreata of E14.5 or E16.5 embryos were collected for IF staining of hormones in frozen sections. At E14.5, RFP^{+} cells expressing INS or GCG instead of GHRL were observed (Fig. 3g). At E16.5, many RFP^{+} cells were INS^{+} but not SST^{+} cells (Fig. 3h). Therefore, *Mnx1*-expressing EP4 cells can generate β and α cells, but they are not the progenitors of ε and δ cells. These results indicate that, along the EP1–EP4 path, the nodes of the ε and δ branches are upstream of the location where β cells are generated.

Next, we sought to analyze and verify cell differentiation branches along the entire allocation pathways of endocrine lineages defined by the Smart-seq2 method, which is more sensitive with regard to transcript detection in comparison with the 10× method (Supplementary information, Fig. S3l).

Subpopulations and developmental pathways of α/PP-Pro-I (ε) cells

To determine the differentiation process of α/PP-Pro-I (ε) cells, we performed Monocle2[39] and principal component analyses (PCA) of EP3, EP4 and α/PP-Pro-I cells. EP3 and EP4 cells formed a linear pathway, whereas α/PP-Pro-I cells branched out from the EP3 population (Fig. 4a). An RNA velocity analysis indicated the developmental direction from EP3 cells to α/PP-Pro-I cells (Fig. 4a). A differential expression analysis identified 1112 cell type-featured genes (Fig. 4b, c; Supplementary information, Fig. S5a and

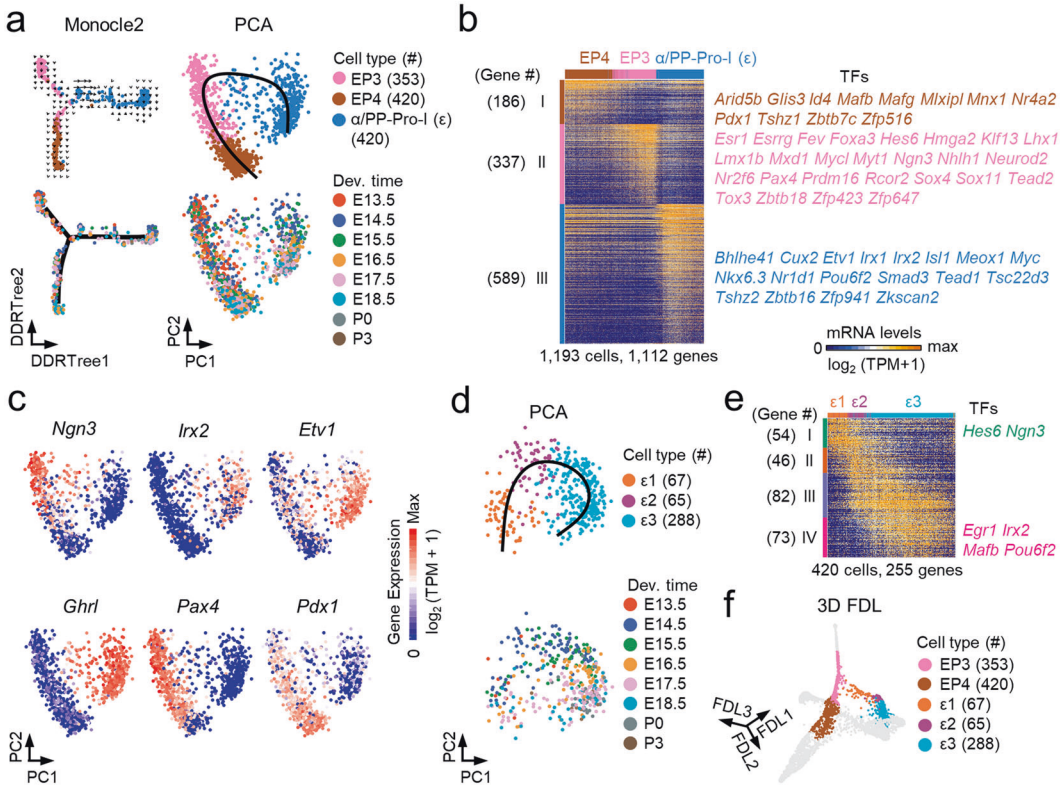

Fig. 4 Differentiation pathways of EP3 cells in mice. a Monocle2 and PCA plots of EP3, EP4 and α/PP-Pro-I (ε) cells from the Smart-seq2 dataset. Each dot represents a single cell. The cell types (top) and developmental time (bottom) are color coded. The principal tree on the Monocle2 plot and the simultaneous principal curve on the PCA plot indicate the developmental pathways of EP3 cells. The arrows on the Monocle2 plot indicate the developmental direction estimated by RNA velocity analysis. **b** Heatmap of differentially expressed genes in EP3, EP4 and α/PP-Pro-I (ε) cells. **c** Expression levels of marker genes projected onto the PCA plot from **a**. **d** PCA plot of α/PP-Pro-I (ε) cells showing three subtypes of ε cells (top) and the corresponding developmental time (bottom). **e** Heatmap of coexpressed genes in ε1–3 cells. **f** ε1–3, EP3 and EP4 cells projected onto the 3D FDL plot from Fig. 2a.

Tables S3, S4). Cluster III included 589 genes that were mainly expressed in α/PP-Pro-I cells, including the α/PP lineage-associated TFs *Irx2* and *Etv1*[40] but not the β/δ lineage-associated TFs *Pax4* and *Pdx1* (Fig. 4b, c), which is consistent with their predicted differentiation potential, as indicated in Fig. 2a.

The α/PP-Pro-I cells were further divided into three subpopulations (ε1–ε3) based on the hierarchical clustering analysis (Fig. 4d, e; Supplementary information, Tables S3, S4). Several TFs associated with ε1–ε3 development were identified (Fig. 4e; Supplementary information, Fig. S5b). Along the ε1–ε3 path, the EP-featured genes *Ngn3* and *Hes6* were down-regulated, whereas the endocrine lineage-related TFs *Irx2* and *Mafb*[41,42] were up-regulated (Supplementary information, Fig. S5b). The ε1 subpopulation was located close to the EP3 population on the 3D FDL plot (Fig. 4f) and primarily included cells at earlier developmental times, whereas the ε3 subpopulation included cells at later times (Fig. 4d). Taken together, these analyses identified the developmental pathway of α/PP-Pro-I cells.

Distinct differentiation potentials of EP4^{early} and EP4^{late} populations

We subsequently focused on the differentiation pathways of EP4^{early} and EP4^{late} cells. Monocle2 and PCA analyses identified the differentiation trajectory of EP4^{early} cells toward either β^{early} cells or toward α cells via the α/PP-Pro-II population (Fig. 5a). Four gene clusters (I–IV), which included many TFs that are essential for pancreatic endocrine differentiation, were differentially expressed in four populations (Fig. 5b; Supplementary information, Fig. S5c and Tables S3, S4). In comparison with other populations, the branch composed of α/PP-Pro-II and α cells showed higher expression levels of *Irx2* and the EP marker *Fev* at E14.5 (Fig. 5c). We performed single-molecule fluorescent in situ hybridization (smFISH) to identify α/PP-Pro-II and α cells (*Fev*$^{+}$*Irx2*$^{+}$), which were distinguishable from the EP3/4 (*Fev*$^{+}$*Irx2*$^{-}$), α/PP-Pro-I (*Fev*$^{-}$*Irx2*$^{+}$) and β cell (*Irx2*$^{-}$*Adra2a*$^{+}$) lineages (Fig. 5c, d).

Unlike the broad connection between α/PP-Pro-II and EP4^{early} cells on the PCA plot (Fig. 5a), δ cells branched out from a restricted region located at the boundary between the EP3 and EP4^{late} populations along the developmental pathway (Figs. 2a and 5e), which suggested that EP cells only provided a transient developmental window in which the generation of δ cells was possible. In comparison with β cells (Cluster-III), δ cells expressed more lineage-specific genes (Cluster-I), including many TFs (Fig. 5f, g; Supplementary information, Tables S3, S4), which indicated that the differentiation of δ cells from EP4^{late} cells was accompanied by marked changes in gene expression. *Hhex*, which promotes δ-cell differentiation,[36] was enriched in the δ-cell branch (Fig. 5g). In addition to *Mnx1*, TFs expressed in EP4^{late} and β cells included *Nkx6.1*, *Mafb*, and *Nkx2.2* (Fig. 5g), which have been demonstrated to play key roles in regulating β-cell development.[41–45]

In summary, our analyses revealed that the EP4^{early} and EP4^{late} cell states are associated with the generation of different endocrine cell types following distinct developmental trajectories.

Distinct pathways generate α and PP cells

The α/PP-Pro-I (ε) and α/PP-Pro-II populations were presumed to be intermediate progenitors of α and PP cells. Surprisingly, 469 genes were differentially expressed between ε3 and α/PP-Pro-II cells (Fig. 6a; Supplementary information, Tables S3, S4). Among these genes, *Fev* was highly expressed in α/PP-Pro-II cells, whereas

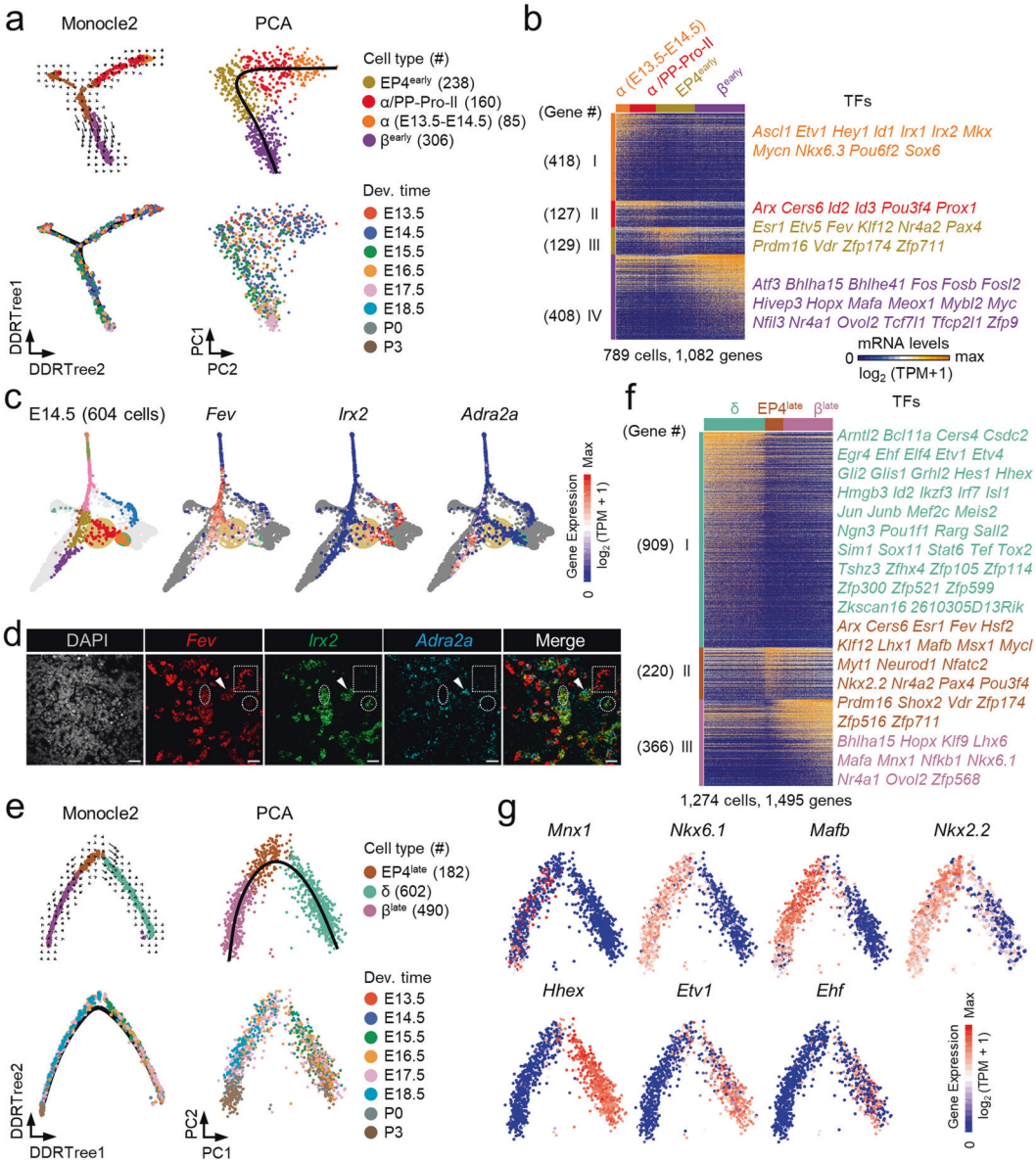

Fig. 5 Stage-dependent differentiation pathways of EP4 cells in mice. a Monocle2 and PCA plots of EP4^{early}, α/PP-Pro-II, α (E13.5–E14.5) and β^{early} cells from the Smart-seq2 dataset. The principal tree on the Monocle2 plot and the simultaneous principal curve on the PCA plot indicate the developmental pathways of EP4^{early} cells. The arrows on the Monocle2 plot indicate the developmental direction estimated by RNA velocity analysis. **b** Heatmap of differentially expressed genes in EP4^{early}, α/PP-Pro-II, α (E13.5–E14.5) and β^{early} cells. **c** Expression levels of *Fev*, *Irx2* and *Adra2a* in cells at E14.5 projected onto the 3D FDL plot in Fig. 2a. The yellow and green shadows represent α/PP-Pro-II and α cells, respectively. **d** smFISH of *Fev*, *Irx2* and *Adra2a* in paraffin sections of pancreas at E14.5. Scale bars, 20 μm. The oval contains the *Fev*$^+$*Irx2*$^+$ region, the box contains the *Fev*$^+$*Irx2*$^-$ region, the circle contains the *Fev*$^-$*Irx2*$^+$ region, and the triangle indicates the *Irx2*$^-$*Adra2a*$^+$ region. **e** Monocle2 and PCA plots of EP4^{late}, δ and β^{late} cells from the Smart-seq2 dataset. The principal tree on the Monocle2 plot and the simultaneous principal curve on the PCA plot indicate the developmental pathways of EP4^{late} cells. The arrows on the Monocle2 plot indicate the developmental direction estimated by RNA velocity analysis. **f** Heatmap of differentially expressed genes in EP4^{late}, δ and β^{late} cells. **g** Expression levels of marker genes projected onto the PCA plot from **e**.

Ghrl was enriched in ε3 cells, and these two genes were also heterogeneously expressed in α and PP cells (Fig. 6b, c). We presumed that *Ghrl*$^+$ α and PP cells originated from α/PP-Pro-I cells, whereas *Fev*$^+$*Ghrl*$^-$ α and PP cells were predominantly generated from α/PP-Pro-II cells (Supplementary information, Fig. S6a). In the FDL plot, *Ghrl*$^+$ and *Fev*$^+$*Ghrl*$^-$ cells were near α/PP-Pro-I ε3 cells and α/PP-Pro-II cells, respectively (Fig. 6b, d). At E13.5, when α/PP-Pro-I cells had just begun to appear (Fig. 2e, h), newly generated α cells should be descended from α/PP-Pro-II cells. As expected, α cells at E13.5 were located in the region near α/PP-Pro-II cells on the FDL plot (Supplementary information, Fig. S6b). We then performed IF to confirm the existence of GHRL$^+$GCG$^+$ cells in the pancreas at E14.5 (Fig. 6e). Additionally, scRNA-seq analysis of the sorted Ghrl-CFP$^+$;Gcg-GFP$^+$ cells at E15.5 showed that these cells were located in the region of ε3 cells (Fig. 6f). To verify the differentiation trajectory of *Ghrl*-expressing cells, we crossed the *Ghrl-Cre* mouse strain with the *Rosa26-RFP* strain and sorted RFP$^+$ cells from both the DP and VP at E17.5 for scRNA-seq. On the FDL plot, we found that the majority of the progeny of *Ghrl*$^+$ cells were α and PP cells, which were located near the α/PP-Pro-I ε3 region rather than the α/PP-Pro-II region (Fig. 6b, g). This result is consistent with a previous finding that

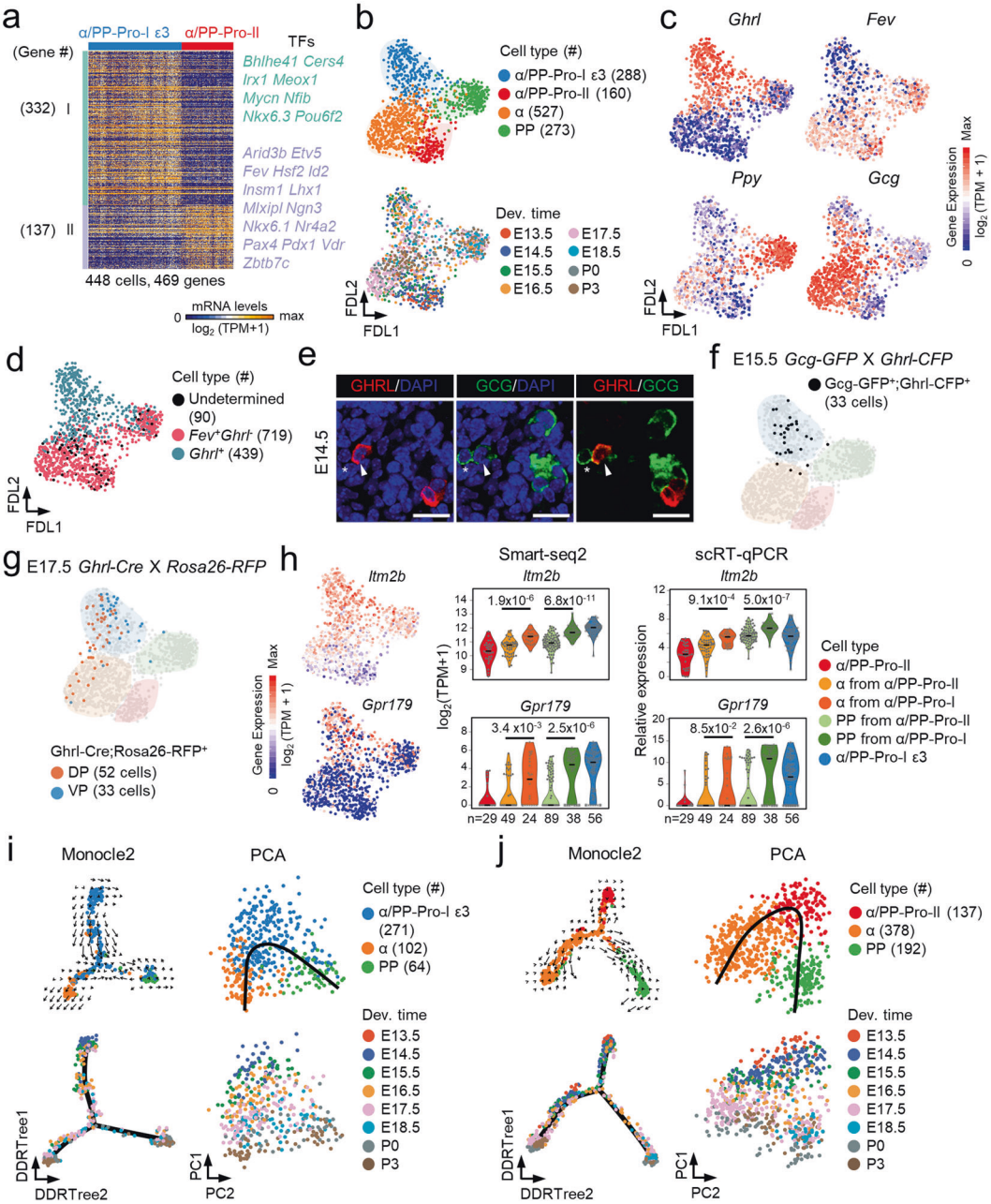

Fig. 6 Distinct pathways generate α and PP cells in mice. a Heatmap of differentially expressed genes between α/PP-Pro-I ε3 and α/PP-Pro-II cells. Single cells are displayed in columns, and genes are displayed in rows. **b** FDL plot of α/PP-Pro-I ε3, α/PP-Pro-II, α and PP cells from the Smart-seq2 dataset. Each dot represents a single cell. The cell types (top) and developmental time (bottom) are color coded. **c** Expression levels of *Ghrl*, *Fev*, *Ppy* and *Gcg* projected onto the FDL plot from **b**. **d** Two origins of α and PP cells from *Fev*$^{+}$*Ghrl*$^{-}$ or *Ghrl*$^{+}$ progenitors shown in the FDL plot in **b**. **e** DAPI and IF staining of GHRL and GCG in paraffin sections of pancreas at E14.5. Scale bars, 20 μm. The arrowhead signifies GHRL^{+}GCG^{+} cells, and the asterisk signifies GHRL^{-}GCG^{+} cells. **f** E15.5 Gcg-GFP^{+}Ghrl-CFP^{+} single cells projected onto the FDL plot from **b** to show the distribution between α/PP-Pro-I ε3 and α cells. The shadows represent the cell types in **b**. **g** E17.5 Ghrl-Cre;Rosa26-RFP^{+} single cells from the DP and VP projected onto the FDL plot in **b**. Several cells located in other regions were excluded from this analysis. The shadows represent the cell types in **b**. **h** Expression of *Itm2b* and *Gpr179* in α/PP-Pro-I ε3, α/PP-Pro-II, α and PP cells. The expression levels [$\log_2$(TPM + 1)] are projected onto the FDL plot in **b** (left) and are shown in violin plots (middle). The expression levels were verified by scRT-qPCR with normalization to *Gapdh* expression (right). *n*, cell counts. The numbers in the violin plots represent the *P*-values calculated using the Wilcoxon rank-sum test. The line in each violin plot represents the median, and the dots represent cells. **i, j** Monocle2 and PCA plots of α/PP-Pro-I ε3, α and PP cells (**i**), as well as α/PP-Pro-II, α and PP cells (**j**), from the Smart-seq2 dataset. Each dot represents a single cell. The principal tree on the Monocle2 plot and the simultaneous principal curve on the PCA plot indicate the developmental pathways of α/PP-Pro-I ε3 cells (**i**) and α/PP-Pro-II cells (**j**). The arrows on the Monocle2 plot indicate the developmental direction estimated by RNA velocity analysis.

Ghrl-expressing cells contribute a significant number of α and PP cells to adult islets.[46] Additionally, the α and PP cells presumably derived from α/PP-Pro-I cells displayed higher expression of *Itm2b* and *Gpr179* than those derived from α/PP-Pro-II cells, which was confirmed by scRT-qPCR (Fig. 6h; Supplementary information, Table S3). We then performed Monocle2 and PCA analyses to describe the differentiation pathways of α/PP-Pro-I ε3 and α/PP-Pro-II cells to α and PP cells, respectively, and identified genes that

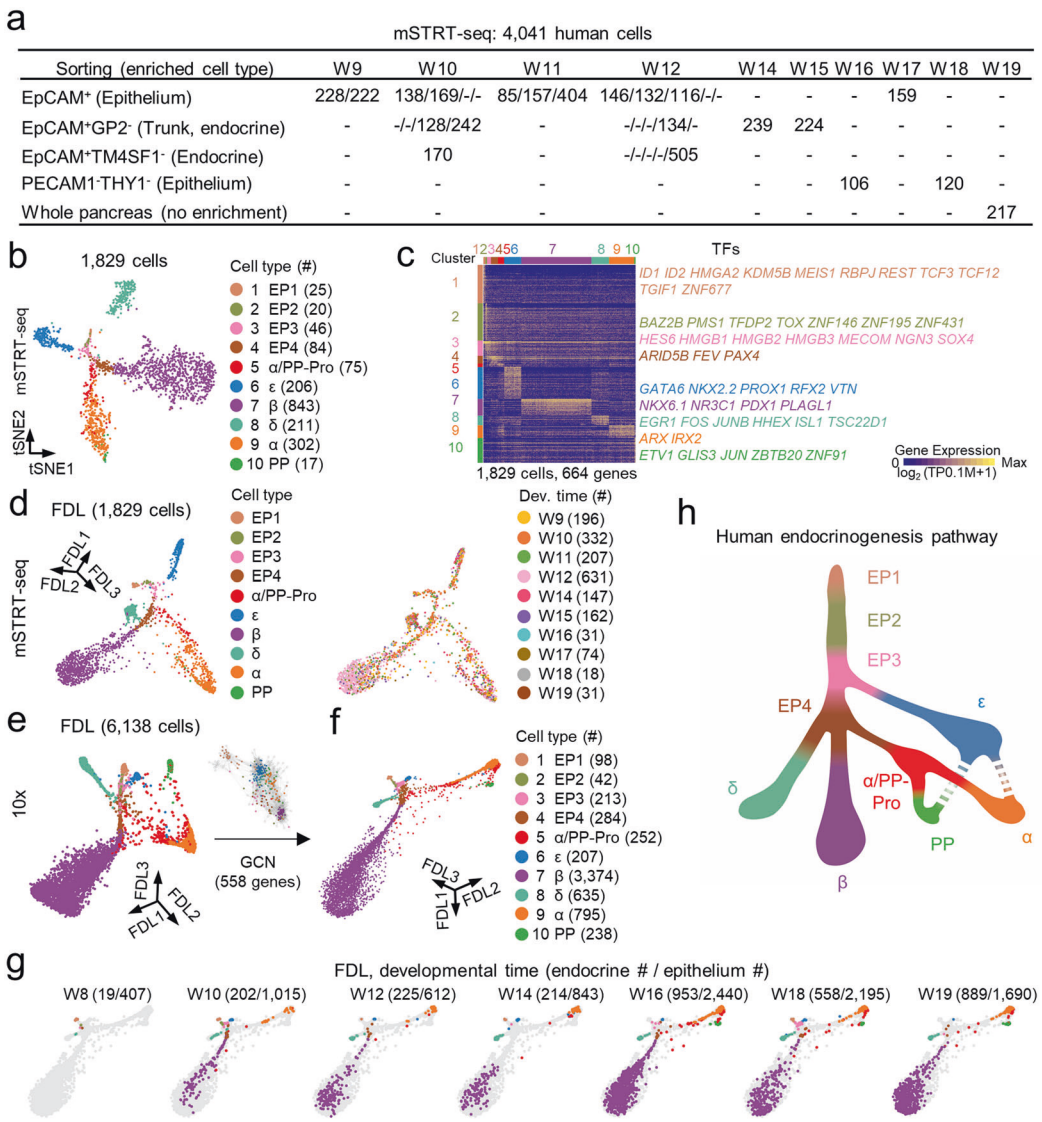

Fig. 7 Identification of cell types and allocation pathways of human endocrinogenesis. a Overview of 4041 pancreatic cells generated by mSTRT-seq in this study. The numbers indicate the cell counts obtained using different enrichment strategies at different developmental time points in each experimental batch. **b** The t-SNE plot shows 10 distinct cell types in EP and endocrine cells generated by mSTRT-seq. Each dot represents a single cell. The cell types are color coded. **c** Heatmap of cell type-featured genes. **d** 3D FDL of EP and endocrine cells generated by mSTRT-seq. The cell types (left) and developmental time (right) are color coded. **e**, **f** 3D FDL analysis of EP and endocrine cells generated by 10× scRNA-seq through application of the top 2000 highly variable genes in the 10× dataset (**e**) or the 558 genes in the GCN derived from the mSTRT-seq dataset (**f**). **g** Temporal distribution of the cells generated by 10× scRNA-seq on the pathways from **f**. The colors denote the cell types in **f**. **h** Summary of endocrine lineage allocation pathways in the human fetal pancreas.

were differentially expressed during these processes (Fig. 6i, j; Supplementary information, Fig. S6c–f and Tables S3, S4).

Altogether, through genetic tracing combined with single-cell transcriptomic analysis, we identified the heterogeneity of early α and PP cells, which might reflect their different origins.

Identification of cell types in human fetal pancreas

The peak of NGN3$^+$ cell generation occurs between 10 weeks post conception (W10) and W14 during human embryonic development.[47] To investigate the endocrine lineage differentiation pathways in humans, we performed scRNA-seq using pancreatic cells from human embryos at W9–W19 using mSTRT-seq method[24,25] (Fig.7a; Supplementary information, Table S1), which is similar to the Smart-seq2 protocol with regard to cDNA production but also includes a unique barcoding step during reverse transcription as well as sample pooling during library construction to overcome the low-throughput shortcoming of Smart-seq2, even though it can only detect 3′ transcripts. Our pilot experiments indicated that the percentage of epithelia was relatively low in embryos at earlier stages, and we thus used the epithelial marker EpCAM to positively enrich pancreatic epithelia by FACS (Fig. 7a). For some samples, we also used the multipotent progenitor and acinar cell marker GP2[48] to deplete these cells and enrich the trunk and endocrine lineages (EpCAM$^+$GP2$^-$) (Fig. 7a). A pilot 10× scRNA-seq analysis of a whole W12 human pancreas revealed that *TM4SF1*, which encodes a member of the transmembrane 4 superfamily, was expressed in nonendocrine epithelial lineages (Supplementary information, Fig. S7a). We therefore used an antibody against this marker to deplete nonendocrine epithelial cells and enrich endocrine lineages (EpCAM$^+$TM4SF1$^-$) at W10 and W12 (Fig. 7a; Supplementary information, Fig. S7b). At later stages, as the percentage of epithelial cells increased, we simply removed the endothelial cells and mesenchymal cells by FACS using PECAM1 and THY1 to

enrich epithelia from pancreatic tissue at W16 and W18, or unbiasedly sorted the cells from a pancreas at W19 (Fig. 7a). After quality control, we obtained a total of 4041 cells, with an average of 0.1 million UMIs and more than 4000 genes (Supplementary information, Fig. S7c, d and Table S1). After removing other cell lineages based on their marker gene expression, we identified 3717 cells as pancreatic epithelia (Supplementary information, Fig. S7e, f). We further divided the pancreatic epithelia into 11 cell types (Supplementary information, Fig. S7g–i) and identified cell type-featured genes (Supplementary information, Fig. S7j and Table S2). We used Seurat v3[49] to integrate the human mSTRT-seq datasets and mouse Smart-seq2 datasets, which revealed that the cell types identified in mice were conserved in humans (Supplementary information, Fig. S7k).

We also performed 10× scRNA-seq using human pancreatic cells at W8–W19, enriched tip and acinar cells ($EpCAM^+GP2^+$) at W12, and trunk and endocrine cells ($EpCAM^+GP2^-$) at W12 (Supplementary information, Fig. S8a). After quality control and the exclusion of nonepithelial lineages and multihormonal cells (Supplementary information, Fig. S8b–g and Table S1), we obtained 16,999 pancreatic epithelia, which were also classified into 11 cell types (Supplementary information, Fig. S8h, i). Compared with those obtained with the mSTRT-seq method, we identified a lower number of cell type-featured genes using this 10× dataset (Supplementary information, Figs. S7j, S8j and Table S2), although most of the genes overlapped with the genes identified by mSTRT-seq (Supplementary information, Fig. S8k).

Conserved endocrinogenesis pathways in humans

We performed a 3D FDL analysis of the mSTRT-seq dataset to define the differentiation pathways of human pancreatic endocrine and exocrine lineages. Additionally, we used the GCNs extracted from mSTRT-seq data to optimize the cell clustering and differentiation pathways based on the 10× analyses (Supplementary information, Fig. S8l, m). We also compared the developmental pathways of pancreatic endocrine and exocrine lineages in humans and mice (Supplementary information, Fig. S8m, n). On the developmental pathways defined by the unbiased 10× datasets, we observed that the human endocrine lineages were strikingly separated from the trunk cells with a connection of scattered *NGN3*$^+$ EP cells (Supplementary information, Fig. S8m, o), which was different from the continuous connection of *Ngn3*$^+$ EP cells in mice (Supplementary information, Fig. S8n, p). In addition, the proportion of EP cells in human pancreatic epithelia was markedly lower than that in mice (Supplementary information, Fig. S8q).

We then focused on deciphering human endocrine lineage differentiation pathways. We first examined the heterogeneity of EP cells. Curiously, as we found in our analysis of mice[10] (Fig. 1d–f), human EP cells could also be classified into four stages with a cascading gene expression pattern (Fig. 7b, c; Supplementary information, Fig. S9a, b and Tables S2, S5, S6). However, unlike mice (Fig. 2a, b), EP4 cells were not heterogeneous during human endocrinogenesis, which might be due to an insufficient number of EP cells. By 3D FDL analysis, we then defined the developmental pathways of all endocrine lineages in humans, which were similar to those in mice. EP3 cells generated ε cells, and early EP4 cells on the trajectory differentiated into δ and α/PP-Pro cells, and late EP4 cells developed into β cells (Fig. 7d). Moreover, ε cells also exhibited heterogeneity in their developmental states (Supplementary information, Fig. S9c, d and Tables S5, S6).

Again, using the GCNs obtained from mSTRT-seq data, we adjusted the cell classification of endocrine lineages and obtained an optimized cell differentiation pathway from our 10× data, which was similar to that described by mSTRT-seq (Fig. 7e, f; Supplementary information, Fig. S8r, s). By presenting the 10× cells according to their developmental stages, we determined the temporal order of endocrinogenesis (Fig. 7g). At W8, only sporadic endocrine cells began to differentiate, and ε, β, δ and α cells appeared at W10. At W16, PP cells were produced in significant numbers (Fig. 7g). However, in this study, the capacity for ε cells to differentiate into α and PP cells could not be confirmed due to the lack of late-stage human embryos. Curiously, during endocrinogenesis in humans, β cells are predominantly produced. The proportion of β cells in human fetal pancreatic endocrine cells (~70%) was higher than that in the mouse fetal pancreas (~45%) (Supplementary information, Fig. S9e). However, at the adult stage, mouse islets contain a higher percentage of β cells (60%–80%) than human islets (50%–70%).[50] In summary, humans and mice exhibit similar islet cell differentiation trajectories, but they display clear differences in the proportions of the EP and endocrine lineages (Fig. 7h).

Different GCNs during human and mouse islet differentiation

To comprehensively evaluate the conservation of endocrinogenesis programs between humans and mice, we compared the GCNs of the two species during the key steps of islet lineage differentiation. We first performed PCA and FDL analyses of EP3, EP4 and ε cells and clearly observed the paths for the differentiation of EP3 to EP4 and ε cells (Fig. 8a). Differential expression analysis identified 1155 cell type-featured genes among EP3, EP4 and ε cells (Fig. 8b; Supplementary information, Tables S5, S6). Similar to our observations in mice, ε cells displayed exclusive expression of a large number of genes (Fig. 8b). Similarly, we defined the pathways and differentially expressed genes during the differentiation from EP4 to α/PP-Pro, δ and β cells (Fig. 8c, d; Supplementary information, Tables S5, S6) and the differentiation from α/PP-Pro to α and PP cells (Fig. 8e, f; Supplementary information, Tables S5, S6).

Coexpressed genes are often functionally related and participate in the same biological processes, such as the regulation of cell fate determination.[26–28] Hence, we expect that the comparison of the GCNs between humans and mice will identify candidate species-specific regulators of cell fate determination during endocrinogenesis. We identified an up-regulated gene module and a down-regulated gene module in the GCN during the transition from EP3/4 to ε cells in both humans and mice, and a considerable fraction of the GCN genes, including key TFs, were species-specific (Materials and Methods, Fig. 9a; Supplementary information, Fig. S10a and Table S7). Previous studies have shown that the TF *Pax6* was specifically expressed in mouse ε cells but not in human ε cells.[51] Consistent with this finding, the GCN analysis showed that *PAX6* was specifically down-regulated during the differentiation of EP3/4 cells to ε cells in humans, and a similar pattern was found for the TF *MAFB* (Fig. 9a; Supplementary information, Fig. S10a). The TF *Irx2* was shown to be up-regulated in mouse ε cells (Fig. 9a; Supplementary information, Table S7), and smFISH confirmed that *Irx2* was expressed in mouse ε cells but not in human ε cells (Fig. 9a–c; Supplementary information, Table S7). Subsequently, we compared the GCNs between species during the β, α and δ cell differentiation steps and identified many species-specific GCN-related genes (Fig. 9a; Supplementary information, Fig. S10b–e and Table S7). We verified the expression of several of these genes by IF. Notably, S100A10 was expressed in mouse but not human β cells (Fig. 9d, e; Supplementary information, Table S7); ERO1B was specifically expressed in human but not mouse δ cells (Fig. 9f, g; Supplementary information, Table S7), whereas MEF2C was expressed in mouse but not human δ cells (Fig. 9h, i; Supplementary information, Table S7). In addition to endocrine lineages, the differentiation of early pancreatic progenitor and exocrine lineages also involved different GCNs between species (Supplementary information, Fig. S10f–h and Table S7). These findings provide evidence that humans and mice might use distinct genetic networks to regulate lineage differentiation.

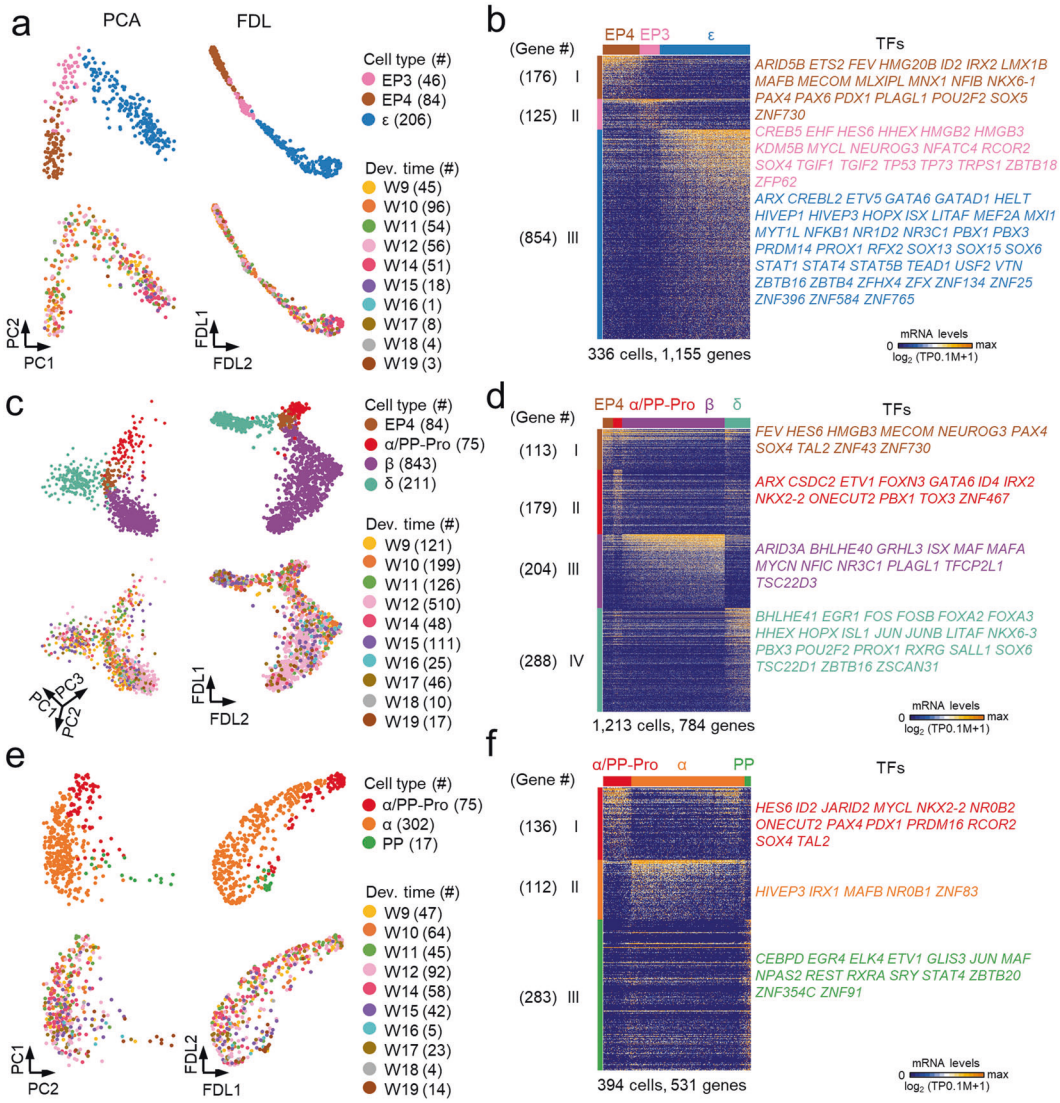

Fig. 8 Differentiation pathways of endocrine lineages in the human fetal pancreas. a PCA and FDL plots of EP3, EP4 and ε cells from the mSTRT-seq dataset. Each dot represents a single cell. The cell types (top) and developmental time (bottom) are color coded. **b** Heatmap of differentially expressed genes in EP3, EP4 and ε cells. **c** PCA and FDL plots of EP4, α/PP-Pro, β and δ cells from the mSTRT-seq dataset. **d** Heatmap of differentially expressed genes in EP4, α/PP-Pro, β and δ cells. **e** PCA and FDL plots of α/PP-Pro, α and PP cells from the mSTRT-seq dataset. **f** Heatmap of differentially expressed genes in α/PP-Pro, α and PP cells.

DISCUSSION

It has long been known that pancreatic endocrine lineages differentiate from EP cells, but the developmental pathways and intrinsic regulatory logic of endocrine lineage allocation have not been resolved due to the complexity of the differentiation pathways, the existence of multiple intermediate cell states, and the subtle transcriptomic differences among the relevant cell lineages and cell states. In this study, we generated new knock-in mouse lines to enrich the rare ε, δ and PP cell lineages, and applied various additional genetic tools. Then, using the high-quality Smart-seq2 approach and high-throughput 10× method, we successfully defined the temporal developmental pathways of all endocrine lineages during mammalian pancreatic organogenesis (Fig. 2j). Most of our significant findings were carefully verified by IF, smFISH, scRT-qPCR and several independent genetic tracing experiments.

EP cells contain several intermediate stages/states, which are thought to represent progressively more mature cell states.[9,10] However, our findings reveal that EP states represent the differential potentials of certain endocrine lineages. For example, EP3 cells retained the potential to differentiate into α/PP-Pro-I (ε) and EP4 cells. Intriguingly, we divided the EP4 cells in mice into $EP4^{early}$ and $EP4^{late}$ subpopulations, which tended to generate α/PP-Pro-II/$β^{early}$ and δ/$β^{late}$ cells, respectively. Additionally, we observed that α/PP-Pro-I (ε) and α/PP-Pro-II cells independently generated α and PP cells. Notably, β cells were located at the end of the EP differentiation pipeline, which suggested that EP cells might eventually become β cells if they do not differentiate into other endocrine cell types. Consistent with our model, previous studies have observed that specific endocrine lineages are generated in an asynchronous manner during endocrinogenesis,[52,53] and α cells appear earlier than β cells.[19] Based on these findings, we hypothesize that the various states of EP cells create different permissive windows for the generation of certain endocrine lineages. The switch of cell states from EP4-$β^{early}$ to EP4-$β^{late}$ might be regulated by changes in the niche-like environment during pancreatic organogenesis.[8,54,55]

As mentioned above, the genes in GCNs are usually functionally related to important biological processes. Hence, mutations in GCN-related genes might cause pancreas agenesis and diabetes. For

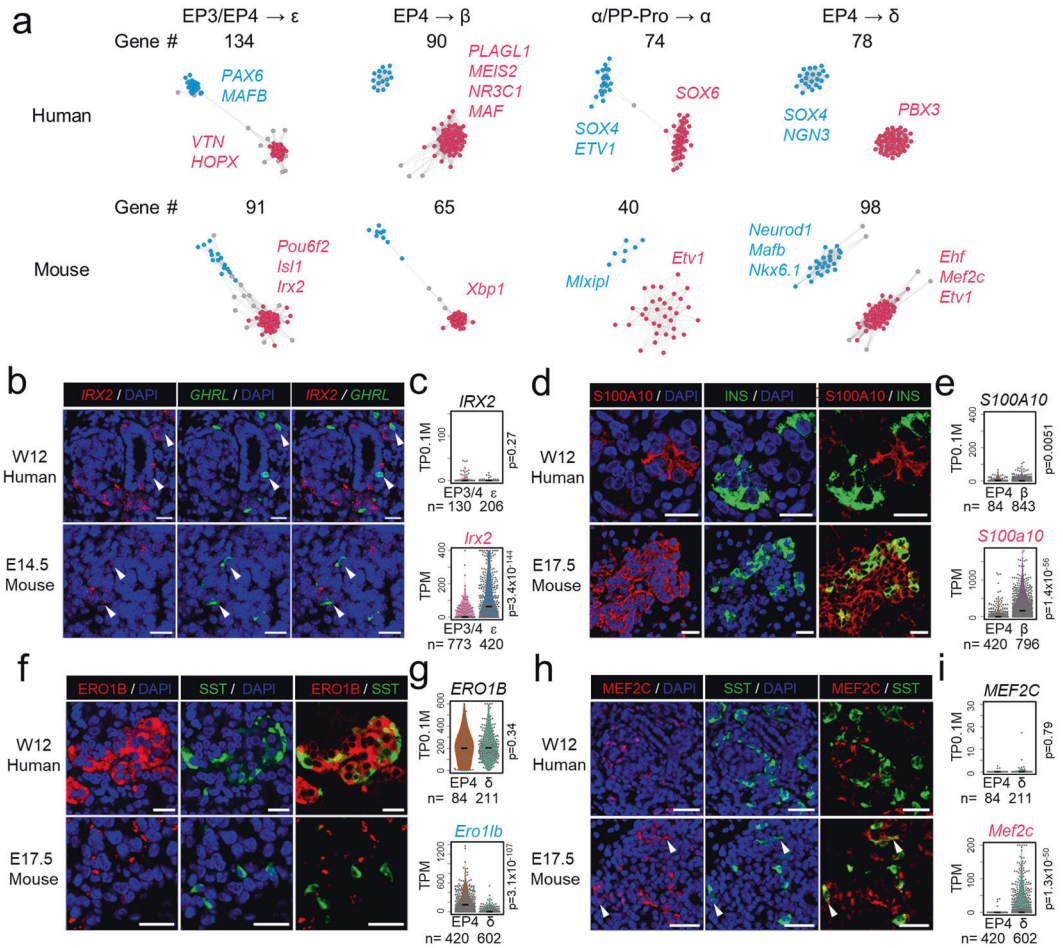

Fig. 9 GCN comparison between human and mouse fetal pancreases. a GCNs during key steps of islet lineage differentiation in humans and mice. The dots in red and blue indicate the up- and down-regulated genes, respectively. The species-specific up- (red) and down-regulated (blue) TFs are listed next to the network graphs. **b** smFISH of *GHRL* and *IRX2* in paraffin sections of human pancreas at W12 and mouse pancreas at E14.5. Scale bars, 20 μm. The arrow heads indicate *GHRL*$^+$ cells. **c** Violin plots show the expression level of *IRX2* in humans (top) and mice (bottom). **d**–**i** IF staining of S100A10 and INS (**d**), ERO1B and SST (**f**), and MEF2C and SST (**h**) in paraffin sections of human pancreas at W12 and mouse pancreas at E17.5. Scale bars, 20 μm. The arrowheads in **h** indicate SST$^+$MEF2C$^+$ cells in mice. Violin plots show the expression levels of *S100A10* (**e**), *ERO1B* (**g**) and *MEF2C* (**i**) in humans (top) and mice (bottom). The numbers above the violin plots represent the *P*-values calculated using the Wilcoxon rank-sum test. The gene symbols of species-specific up- and down-regulated genes are highlighted in red and blue, respectively. The line in each violin plot represents the median, and the dots represent cells. *n*, cell counts.

instance, mutations of TFs that play key roles in pancreatic development, such as PDX1, MNX1 and NEUROD1, lead to diabetes.[14,56–59] An integrated analysis of scRNA-seq data of human pancreatic development and genome-wide association study (GWAS) data of diabetes-associated genes (https://www.ebi.ac.uk/gwas/) indicated two groups of GCN-related genes associated with type-1 and type-2 diabetes (Supplementary information, Table S8).

A recent study identified a *Procr*$^+$ endocrine progenitor population in adult mouse islets that retains the potential to generate all endocrine cells and indicated that these adult *Procr*$^+$ cells developed from *Ngn3*$^+$*Procr*$^+$ cells in the fetal pancreas.[60] However, we did not identify *Ngn3*$^+$ cells coexpressing the *Procr* gene in our Smart-seq2 scRNA-seq data analysis (Supplementary information, Fig. S2f, g). In addition, reanalysis of the 10× data of *Ngn3*$^+$ cells enriched from an *Ngn3-Venus* fusion reporter mouse line[12] did not allow identification of the *Ngn3*$^+$*Procr*$^+$ population (Supplementary information, Fig. S10i, j). Considering that droplet-based methods inevitably lead to doublet contamination, the existence of *Ngn3*$^+$*Procr*$^+$ cells requires further verification. Hence, the current analyses do not support the notion that *Ngn3*$^+$ cells develop into *Procr*$^+$ endocrine progenitor cells through an alternative trajectory.

This study excluded hormone-expressing cells generated from the first wave of endocrinogenesis because their contribution to the number of pancreatic endocrine cells might be negligible. Lineage tracing is required to reveal the ultimate fate of the first-wave hormone-expressing cells. Notably, there is no evidence on the first wave of generation of endocrine cells in humans.[61]

We also compared regulatory networks during endocrinogenesis between humans and mice. Our analyses showed that although the branched pathways of the species were similar, the GCNs governing lineage differentiation were significantly different. Therefore, our work represents a unique and valuable resource that provides insights into the regulatory mechanisms underlying mouse and human pancreatic development, as well as key guidelines for the generation of whole islet tissue in vitro.

MATERIALS AND METHODS

Mice

Ngn3-Cre,[62] *Ngn3-Cre*^{ER},[5] *Ghrl-Cre*,[46] *Mnx1-Cre*^{ER}, *Rosa26-RFP*, *Gcg-P2A-GFP*,[10] *Sst-P2A-BFP*, *Ppy-PA2-mNeptune* and *Ghrl-P2A-CFP* transgenic mouse lines were used to obtain pancreatic cells at various developmental time points. The *Mnx1-Cre*^{ER} mouse line

was purchased from Shanghai Model Organisms Center, Inc. The day of vaginal plug appearance was considered E0.5. All the animals were housed in specific pathogen-free animal facilities with a 12-h light-dark cycle at Peking University. The mice were handled according to the rules established by the ethics committee for animal care.

Mnx1-Cre^{ER}, *Sst-P2A-BFP*, *Ppy-P2A-mNeptune*, and *Ghrl-P2A-CFP* transgenic mouse lines were created by inserting the P2A-CreERT2 or P2A-fluorescent protein cassette upstream of the translation stop codon through CRISPR/Cas9-mediated homologous recombination via a previously reported method.[10] The sgRNA sequences used to target *Mnx1*, *Sst*, *Ppy* and *Ghrl* were TCCGCCCTGGAGGCAACTACTGG, ACAACAATATTAAAGCTAAC, CCCTGCACCAGCCCCCCAGT, and CGCCAGCTGACAAGTAACCA, respectively.

For the induction of Cre expression in *Ngn3-Cre^{ER}* and *Mnx1-Cre^{ER}* embryos, pregnant mice were intraperitoneally injected with a single pulse of tamoxifen (Sigma T5648) at a dosage of 0.1 mg/g body weight (20 mg/mL stock solution in corn oil) 1 or 2 days prior to sacrifice.

Human fetal pancreas

Human fetal pancreata were obtained from women who underwent elective pregnancy termination and provided written informed consent at Haidian Maternal & Child Health Hospital in Beijing, China. The operations were performed in accordance with protocols approved by the Peking University Institutional Review Board (PU-IRB) (certificate# IRB00001052-18083). Within 3 h after voluntary termination of pregnancy, the fetal pancreas was dissected from the human embryo which was placed in ice-cold phosphate-buffered saline, and the sex was not deliberately selected. The fetal ages were estimated based on the time since the last menstrual period combined with the crown-rump length. The sex was determined based on the expression of the Y chromosome-associated genes *SRY* and *RPS4Y1*.

Single cell isolation

The pancreata of mice were dissociated as previously described.[10] Briefly, pancreata at E13.5–E14.5 were digested with 0.25% trypsin (Sigma, T4799) at 37 °C for 5 min, whereas pancreata at E15.5–P3 were digested with 0.5 mg/mL collagenase P (Roche, 11213873001) for 2 min and then treated with 0.25% trypsin. Pancreata from the human fetal samples were first cut into small pieces and then incubated in a mixture of 0.5 mg/mL collagenase XI (Sigma, C7657), 1 mg/mL dispase II (Sigma, D4693) and 0.1 mg/mL DNase I (Sigma, DN25) in 1× Hank's balanced salt solution (HBSS) with Ca^{2+} and Mg^{2+} at 37 °C for 5–7 min with gentle shaking. The pellet was then incubated with TrypLE Express (Thermo Fisher, 12605028) and DNase I at 37 °C for 3–5 min with gentle shaking. The cell suspensions were filtered through a 35-μm cell strainer (Corning, 352235) prior to use in subsequent experiments.

Flow cytometry

The cells from mice were directly sorted using specific cell type-labeled fluorescent proteins and a BD Aria SORP flow cytometer. The cells from human pancreata were stained with anti-human EpCAM conjugated to APC (BioLegend, 324208) and anti-human GP2 conjugated to PE (MBL, D277-5) or anti-human TM4SF1 conjugated to Alexa Fluor 488 (NOVUS, FAB8164G), anti-human CD31 (PECAM) conjugated to PerCP/Cy5.5 (BioLegend, 303132), and anti-human CD90 (THY1) conjugated to FITC (BioLegend, 328108) for 15 min on ice and then washed twice with FACS buffer (HBSS containing 1% fetal bovine serum, pH 7.4). The cell suspensions were sorted with a BD Aria SORP flow cytometer to enrich specific cell types. Cell suspensions of human fetal pancreas for 10× Genomics scRNA-seq were stained with 1 μg/mL DAPI (Sigma, D9564) to remove dead cells.

scRNA-seq

Smart-seq2 was performed following a previously described procedure[10] based on the Smart-seq2 method.[23] mRNA was reverse transcribed into cDNA for library construction using a TruePrep DNA Library Prep Kit (Vazyme, TD502). The libraries were sequenced using an Illumina HiSeq 2500 system to obtain 50-bp single-end reads. FACS-purified cells were mouth-pipetted or directly sorted into 96-well plates with a single-cell sorting setting.

The mSTRT-seq procedure was modified from the STRT-seq method.[24,25] The cells were lysed in lysis buffer containing a unique barcode for each cell. cDNA from 48–96 cells was pooled together for subsequent library construction using a Kapa Hyper Prep Kit (Kapa, KK8505). The libraries were sequenced as 150-bp paired-end reads using an Illumina HiSeq 4000 system. FACS-purified cells were mouth-pipetted or directly sorted into 96-well plates with a single-cell sorting setting.

For droplet scRNA-seq, the procedure was performed according to the manufacturer's instructions using the Single Cell 3′ Reagent Kit v2 from 10× Genomics. Ngn3-Cre;Rosa26-RFP^{+} cells from the DP and VP of pancreata at E16.5 and E17.5 were sorted and loaded on a Chromium controller. A fraction of cells from the DP or VP of wild-type pancreata at E16.5 was added to the corresponding Ngn3-Cre;Rosa26-RFP^{+} cells at a 1:2 ratio. The sorted EpCAM^{+}GP2^{-} and EpCAM^{+}GP2^{+} cells from two human samples at W12 were pooled together and then loaded onto the Chromium controller. The libraries were sequenced using an Illumina HiSeq 4000 system to obtain 150-bp paired-end reads. The same human pancreata at W16, W18 and W19 were used for both mSTRT-seq and 10× scRNA-seq.

Immunofluorescence and microscopy

To validate the newly generated mouse lines, pancreata at E14.5 from *Ghrl-P2A-CFP* embryos, pancreata at E18.5 and P60 from *Sst-P2A-BFP* mice, and pancreata at P3 and P60 from *Ppy-P2A-mNeptune* mice were fixed with 4% paraformaldehyde at 4 °C for 8–12 h and cryoprotected in 30% sucrose solution. To obtain paraffin sections, the fetal pancreata were fixed with 4% paraformaldehyde at 4 °C for 8–12 h and dehydrated with ethyl alcohol according to standard procedures. The tissue samples were cut into 5-μm-thick sections. The sections were stained with primary antibodies against GHRL (1:500, Abcam, ab209790), GHRL (1:500, R&D Systems, MAB8200), SST (1:2000, ImmunoStar, 20067), SST (1:400, Santa Cruz Biotechnology, sc-74556), PPY (1:500, Abcam, ab77192), INS (1:500, Abcam, ab7842), GCG (1:200, Millipore, AB932), GCG (1:200, R&D Systems, MAB1249), NGN3 (1:40, DSHB, F25A1B3), MUC1 (1:500, Abcam, ab15481), S100A10 (1:100, Thermo Fisher, PA5-95505), ERO1LB (1:100, Thermo Fisher, PA5-25142) and MEF2C (1:100, Proteintech, 10056-1-AP). The sections stained with antibodies against NGN3, MUC1, S100A10, ERO1LB and MEF2C were boiled in antigen unmasking solution (H-3300, Vector labs) for 10 min in a microwave oven for antigen recovery. The corresponding secondary antibodies were donkey anti-rabbit conjugated to Alexa Fluor 594 (Thermo Fisher, A21207), donkey anti-goat conjugated to Alexa Fluor 488 (Thermo Fisher, A11055), donkey anti-mouse conjugated to Alexa Fluor 647 (Thermo Fisher, A31571), donkey anti-rat conjugated to Alexa Fluor 488 (Thermo Fisher, A21208) and donkey anti-rabbit IgG (H + L) conjugated to Alexa Fluor 647 (Jackson ImmunoResearch, 711-607-003). Images were obtained using a Zeiss LSM 710 NLO and DuoScan System or a Leica TCS SP8.

scRT-qPCR

Prior to sequencing, some of the cDNA obtained from individual cells was diluted and subjected to RT-qPCR with AceQ qPCR SYBR Green Master Mix (Vazyme, Q121-02) and a Roche LightCycler 480 Instrument II for the detection of *Arx* (forward, 5′-TCCGGATACCCC ACTTAGCTT-3′, reverse, 5′-GACGCCCCTTTCCTTTAAGTG-3′), *Hhex* (forward, 5′-CGGACGGTGAACGACTACAC-3′, reverse, 5′-CGTTGGA

GAACCTCACTTGAC-3′), *Itm2b* (forward, 5′-AACATTAAGGCCGGGA CCTAC-3′, reverse, 5′-AGTTACTGGCTTCCCGCTTC-3′) and *Gpr179* (forward, 5′-ATCAGAGGACTCCAGGGATCT-3′, reverse, 5′-CGCCCC ACGGACTTCATATT-3′). *Gapdh* (forward, 5′-ATGGTGAAGGTCGGT GTGAAC-3′, reverse, 5′-GCCTTGACTGTGCCGTTGAAT-3′) was used as an internal control for normalization.

smFISH and microscopy

Pancreata from mouse embryos at E14.5 and human fetal samples at W12 were fixed with 10% neutral-buffered formalin (Solarbio, G2162) at room temperature for 24 h. Hybridization of 5-μm paraffin sections was performed using the RNAscope® Multiplex Fluorescent Reagent Kit v2 (ACDBio) following the manufacturer's instructions. Probes against mouse *Fev* (413241-C3), mouse *Irx2* (519901-C1), mouse *Adra2a* (425341-C4), mouse *Ghrl* (415301-C2), human *GHRL* (455131-C2) and human *IRX2* (450531-C1) were used for detection. Fluorescent images were acquired using a Zeiss LSM 710 NLO and DuoScan System.

ELISAs

Six-week-old male *Sst-P2A-BFP* and *Ppy-P2A-mNeptune* mice and their corresponding wild-type littermates were fasted for 16 h. The next morning, the mice were intraperitoneally injected with glucose at a dosage of 2 mg/g body weight. After 30 min, blood was drawn from the retro-orbital plexus. The serum SST and PPY levels were measured using the Mouse Somatostatin ELISA Kit (Elabscience, E-EL-M1086) and Mouse Pancreatic Polypeptide ELISA Kit (Elabscience, E-EL-M0878), respectively, following the manufacturer's instructions.

Statistical analysis

Using data from the FACS analyses and ELISAs performed with independent biological replicates, the standard error of the mean (SEM) and *P*-value from an unpaired two-tailed *t*-test were calculated ($P < 0.05$ was considered significant). Unpaired two-tailed Wilcoxon rank-sum tests were used to analyze the scRT-qPCR experiments. The statistical criteria used for the single-cell transcriptomic analyses are detailed in the data analysis sections.

Quantification of gene expression from Smart-seq2 data

We aligned the sequenced reads to the mouse reference genome (mm10) with TopHat (v2.1.0)[63] using the parameters "-o out_dir -G gtf --transcriptome-index trans_index bowtie2_index input_fastq". We quantified the reads assigned to each gene using HTSeq (v0.6.0)[64] with the parameters "-f bam -r pos -s no -a 30". The read count was normalized to transcripts per million (TPM). To avoid transcriptomic perturbations due to individual highly expressed genes, we excluded genes that represented > 10% of all transcripts in at least 10 single-cell samples (*Ins1*, *Ins2*, *Gcg*, *Sst*, *Ghrl*, *Ppy*, *Iapp*, *Pyy* and *Yam1*) during the total transcript count calculation process. Smart-seq2 samples with > 0.2 million mapped reads and > 4000 detected genes were used for further analyses.

Quantification of gene expression from mSTRT-seq data

The 3′ and informative sequences of transcripts were included in Read1, and the corresponding 8-bp cell-specific barcode and 8-bp UMI sequences were included in Read2. We appended the UMI sequence to the identifier of Read1. After trimming the polyA sequences, Read1 sequences were split for each single-cell sample based on the cell-specific barcode sequences and aligned to the human reference genome (hg38) using TopHat (v2.1.0)[63] and the same parameters used for processing the Smart-seq2 data. The gene ID was annotated as an XT tag for each read in the bam file using featureCounts (v1.5.3).[65] The bam file was then sorted and indexed with SAMtools (v1.3.1).[66] We quantified the UMI using umi_tools (v0.5.0)[67] with the parameters "count --per-gene --gene-tag=XT --method unique -I indexed_sorted_bam -S out_file". The total UMI count for each mSTRT-seq sample was approximately 0.1 million, and we thus normalized the UMI counts to transcripts per 0.1 million (TP0.1 M). During the calculation of total transcripts, we ignored the extremely highly expressed hormone genes (*INS*, *GCG*, *SST*, *PPY* and *GHRL*) and mitochondrial genes. To reduce the index-switching effect of the HiSeq 4000 platform, we subtracted 2.5% of the maximum TP0.1 M value for each gene. The mSTRT-seq samples with > 50,000 mapped reads and > 1000 detected genes were used for further analyses.

Quantification of gene expression from 10× Genomics data

The sequencing data were processed using Cell Ranger (v2.0.2, https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/what-is-cell-ranger) with the default settings. The "raw_gene_bc_matrices" generated from Cell Ranger were imported and integrated into a Seurat object. The UMI count was then transformed to ln(transcripts per 10,000) [ln (TP10K)] using the *NormalizeData* function in the Seurat package (v3.0.2).[49] 10× Genomics samples with total UMI count > 3000, > 500 detected genes, and a mitochondrial UMI proportion < 20% were used for the subsequent analyses.

Cell type identification of Smart-seq2 and mSTRT-seq data

For the mouse Smart-seq2 data, variable genes were identified based on ERCC spike-ins, as previously described.[68] To reduce the noise in our scRNA-seq data, we applied GCNs to further filter variable genes. We calculated a pairwise ρ_p matrix[69,70] for the variable genes based on the $\log_2(\mathrm{TPM}+1)$ values and filtered the variable genes using the following criteria: (i) coexpression with at least 10 other variable genes ($\rho_p > 0.25$), (ii) detected in at least 0.5% of samples, and (iii) not detected in at least 10% of samples. The pairwise ρ_p matrix of filtered variable genes was used to construct an undirected weighted graph with the *graph.adjacency* function in the igraph package (v1.2.4.1).[71] The filtered variable genes were considered graph vertices, and two genes were connected with an edge if $\rho_p > 0.25$. The gene modules were defined using the *cluster_walktrap* function in the igraph package. Genes in cell cycle-related modules (including cell cycle regulators such as cyclins and Cdks) were excluded in the cell type identification process.

For the cell type identification shown in Fig. 1d and Supplementary information, Fig. S2f, genes in the exocrine-related module (including exocrine enzymes such as *Cel* and *Prss1*, which might have been detected due to ambient RNA contamination) were excluded. The remaining coexpressed genes were employed for PCA using a $\log_2(\mathrm{TPM}+1)$ matrix and RunPCA in the Seurat package. t-SNE and Louvain clustering were performed using RunTSNE and FindNeighbors/FindClusters, respectively, in the Seurat package. Cell clusters were annotated based on marker gene expression levels as shown in Fig. 1d, f and Supplementary information, Fig. S2d–g. The identified Ins^+Ppy^+ cells were excluded from this study (Supplementary information, Fig. S2f, g). Cell type-featured genes were identified using the *FindAllMarkers* function in the Seurat package with the parameters "only.pos = TRUE, min.pct = 0.5, logfc.threshold = 1". We referred to AnimalTFDB[72] for TF information. Proliferative cells were identified by hierarchical clustering with cell cycle-related genes identified in the GCN.

For human mSTRT-seq data, variable genes were identified using the *M3DropFeatureSelection* function in the M3Drop package (v1.10.0).[73] The GCN algorithm and Louvain clustering were used as described for the mouse Smart-seq2 data analyses. We identified epithelia (Supplementary information, Fig. S7e, f) and divided them into endocrine, trunk, tip and acinar cell types. To identify the initially specified endocrine cells, we performed clustering of the trunk and endocrine cells together and considered the $NGN3^+$ cells in a subcluster of trunk cells as endocrine cells. We then classified the endocrine cells as EP, β, δ, ε, α/PP-Pro, α and PP cells. We further defined the boundary of α/PP-

Pro, α and PP cells based on the expression patterns of *GCG* and *PPY*. The EP cells were divided into four stages by hierarchical clustering. The cell type-featured genes and proliferative cells were identified as described for the mouse Smart-seq2 data analyses.

Cell type identification of 10× Genomics data

Using the 10× data, epithelial cells were identified by two rounds of clustering. In the first round, the top 2000 variable genes were identified using the *FindVariableFeatures* function in the Seurat package. After excluding the cell cycle-related genes defined in our previous work,[10] the retained variable genes were employed for PCA and Louvain clustering. The cell types of these identified clusters were then annotated based on the expression of marker genes: endocrine cells (*Neurod1*$^+$), acinar cells (*Ctrb1*$^+$), pancreatic progenitors (*Epcam*$^+$;*Neurod1*$^-$;*Ctrb1*$^-$), mesenchyme cells (*Col3a1*$^+$), neurons (*Ascl1*$^+$), endothelial cells (*Pecam1*$^+$), immune cells (*Ptprc*$^+$) and erythrocytes (*Hba-a1*$^+$). Cell type-featured genes were then identified using the *FindAllMarkers* function in the Seurat package with the parameters "only.pos = TRUE, logfc.threshold = 0.6". These cell type-featured genes were applied for the correction of batch effects using the *fastMNN* function in the scran package (v1.12.1).[74] Based on the batch correction result, the second round of Louvain clustering was performed. Cell clusters showing expression of marker genes of any two cell types among the mesenchyme, neuron, immune, endothelium, endocrine epithelium and nonendocrine epithelium were considered doublets and discarded in the downstream analyses. In addition, erythrocytes and *Hba-a1*$^+$ cells in nonerythrocyte clusters were excluded. The types of the remaining cells were then reannotated based on the gene expression pattern of markers, as shown in Supplementary information, Fig. S3d, e. Subsequently, the sparsely distributed multihormone$^+$ cells on the t-SNE plot were ignored in this study (Supplementary information, Fig. S3f, g).

To define the cell types of endocrine lineages, we applied the top 2000 variable genes (excluding cell cycle-related genes) or genes in the GCN identified from the Smart-seq2 data analysis. First, we performed PCA and batch effect correction using the *fastMNN* function in the scran package. Louvain clustering was then performed using the corrected PCA result, and the cell types were annotated according to marker gene expression patterns, as shown in Supplementary information, Fig. S3i, j. To distinguish the first-wave α-cells (α-1st cells) and the second-wave α-cells (α-2nd cells) in the 10× datasets, we then identified the differentially expressed genes between α-1st and α-2nd cells based on Smart-seq2 datasets and used these genes to perform the hierarchical clustering analysis of the 10× datasets (Supplementary information, Fig. S3h). After excluding the α-1st cells (Fig. 2f, g; Supplementary information, Fig. S3i, m), cell type-featured genes were identified using the *FindAllMarkers* function in the Seurat package with the parameters "only.pos = TRUE, min.pct = 0.5, logfc.threshold = 0.6". GO enrichment analysis was performed with GOstats (v2.46.0).[75] Proliferative cells were identified by hierarchical clustering with the cell cycle-related genes identified in the Smart-seq2 data.

Developmental trajectory inference

FDL and DDRTree algorithms were employed to infer the developmental trajectories. For the FDL analysis, we constructed a shared nearest neighbor (SNN) matrix using the *FindNeighbors* function in the Seurat package based on the PCA results and converted the SNN matrix into a graph using the *graph.adjacency* function in the igraph package. The FDL results were then generated using the *layout_with_fr* function in the igraph package. We inferred the lineage structure of the FDL results using slingshot (v1.2.0).[76] To avoid loop structures, we performed the slingshot analysis using EP1–4, β, δ, α/PP-Pro-II, α and PP cells, as well as EP1–3, ε, α and PP cells. The three-dimensional FDL results were visualized using rgl (v0.100.26) (https://cran.r-project.org/package=rgl). The DDRTree was constructed using Monocle (v2.8.0).[39]

RNA velocity analysis

The RNA velocity analysis was performed following the Velocyto. R (v0.6)[37] pipeline. Briefly, we used Velocyto to count the spliced and unspliced reads of each gene from the aligned data. The spliced and unspliced read counts were then used to estimate the RNA velocity, which was then projected onto the Monocle or FDL results.

Cell type integration between humans and mice

"One to one" orthologous genes in humans and mice annotated with the Ensembl genome annotation system (http://www.ensembl.org/index.html) were used for cross-species analyses. We used the standard workflow of Seurat v3 to integrate the human mSTRT-seq dataset and mouse Smart-seq2 dataset from this study and our previous work.[49] Briefly, two datasets were projected into a shared subspace by canonical correlation analysis, and the L2-norm of the canonical correlation vectors was calculated. The anchor correspondences between the two datasets were identified with mutual nearest neighbors (MNNs). Finally, the two datasets were then integrated based on anchor correspondences and the t-SNE was performed with *RunTSNE* in the Seurat package.[49]

Comparisons of the GCNs of humans and the GCNs of mice

We compared the GCNs using one-to-one homologous and nonhomologous genes during the key steps of pancreatic lineage differentiation between humans and mice to identify species-specific regulators. The analysis process consists of five steps.

(1) The compared cell populations were sampled to obtain the same sample size and thus avoid analytical bias due to sample size imbalance.
(2) The one-to-one homologous genes were filtered to make the GCNs from mouse Smart-seq2 data comparable to those from human mSTRT-seq data. mSTRT-seq involves a cDNA production procedure similar to that used in Smart-seq2. However, unlike Smart-seq2, which detects full-length transcripts, mSTRT-seq only detects 3′ transcripts. Therefore, the transcripts detected by Smart-seq2 contain the transcripts detected by mSTRT-seq. To ensure the comparability of GCNs obtained from mSTRT-seq with those obtained from Smart-seq2, we narrowed the comparison scope to coexpressed genes in the mSTRT-seq dataset. To obtain a complete list of coexpressed genes from the mSTRT-seq dataset, we identified coexpressed genes in all cell types (epithelium and other cell types in the pancreas), epithelial cells, or endocrine cells, respectively. The union of these three coexpressed gene sets was considered a comparable gene set. Nonhomologous protein-coding genes, which were detected in at least 25% of the samples of at least one cell type, were also included in this study.
(3) The candidate GCN genes associated with each key step of pancreatic lineage differentiation were identified using the following steps: (i) The differentially expressed genes in each step of pancreatic lineage differentiation in humans and mice were identified and combined to obtain the seed orthologous gene set. (ii) Using the *moduleEigengenes* function in the WGCNA package (v1.68),[77] we generated module eigengenes from the seed orthologous gene sets of humans and mice. This module eigengene can represent the developmental pseudotime during the process of cell lineage differentiation. (iii) Using the *signedKME* function in the WGCNA package, we calculated the kME score corresponding to the module eigengene for each gene in

the comparable gene set in humans and mice. The kME score can evaluate whether a gene is related to the developmental pseudotime. (iv) The genes with an absolute kME score value > $threshold_{kME\_1}$ (independently determined based on the analyzed cell types, ranging from 0.45 to 0.55) in humans or mice were considered candidate GCN orthologous genes.

(4) GCNs of humans and mice were constructed through the following steps. We applied the candidate GCN orthologous genes for pairwise ρ_p matrix calculation in each step of pancreatic lineage differentiation in humans and mice. Only genes that were coexpressed ($\rho_p > 0.2$) with three or more genes were reserved. To visualize the GCNs, we used the *graph.adjacency* function in the igraph package to convert the ρ_p matrix of the reserved genes into a graph and generated a network map using the *plot.igraph* function in the igraph package.

(5) Species-specific up- or down-regulated genes were identified. We considered the genes in the GCN with higher (or lower) kME scores than a positive (or negative) threshold$_{kME_2}$ (independently determined based on the analyzed cell types, ranging from 0.1 to 0.3) as up-regulated (or down-regulated) genes. The genes that were up-regulated or down-regulated in only one species were considered species-specific up-regulated or down-regulated genes.

DATA AVAILABILITY

The mouse RNA-seq data from this publication have been deposited to the Gene Expression Omnibus (GEO) and assigned the identifier GSE139627. The human RNA-seq expression matrix data from this publication have been deposited to the OMix (https://bigd.big.ac.cn/omix/) and assigned the identifier OMIX236.

ACKNOWLEDGEMENTS

We thank Drs. Chris Wright, Zemin Zhang, Ge Gao and members of the Xu laboratory for the advice and comments. We thank Nurse Jianrong Tian for the assistance with human embryo collection. We also thank the Peking-Tsinghua Center for Life Science High Performance Computing Platform, the Flow Cytometry Core at the National Center for Protein Sciences at Peking University, particularly Ms. Fei Wang, Ms. Yinghua Guo and Ms. Hongxia Lv for the technical help, and the Core Facilities at the School of Life Sciences of Peking University, particularly Ms. Siying Qin for the technical help and Mr. Ming Du for the assistance with imaging. This work was supported by the National Key R&D Program of China (2019YFA0801500 to C.-R.X.), the Ministry of Science and Technology of China (2015CB942800 to C.-R.X.), the National Natural Science Foundation of China (32030034, 91753138, and 31521004 to C.-R.X. and 32000566 to X.-X.Y.), funding from the Peking-Tsinghua Center for Life Sciences to C.-R. X, and the China Postdoctoral Science Foundation (BX20190009 to X.-X.Y. and 2020TQ0018 to W.-L.Q.).

AUTHOR CONTRIBUTIONS

C.-R.X. conceived the project; C.-R.X., X.-X.Y. and W.-L.Q. designed the experiments; X.-X.Y., L.Y., Y.-C.W., M.-Y.H., D.W., Y.Z., L.-C.L., J.Z. and Y.W. performed the experiments; Y.-C.W., J.Z. and Y.W. collected the human fetal samples; W.-L.Q., X.-X.Y. and C.-R.X. analyzed the data; X.-X.Y., W.-L.Q. and C.-R.X. wrote the manuscript.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41422-021-00486-w.

Competing interests: The authors declare no competing interests.

REFERENCES

1. Zhou, Q. & Melton, D. A. Pancreas regeneration. *Nature* **557**, 351–358 (2018).
2. Aguayo-Mazzucato, C. & Bonner-Weir, S. Pancreatic beta cell regeneration as a possible therapy for diabetes. *Cell Metab.* **27**, 57–67 (2018).
3. Shih, H. P., Wang, A. & Sander, M. Pancreas organogenesis: from lineage determination to morphogenesis. *Annu. Rev. Cell Dev. Biol.* **29**, 81–105 (2013).
4. Gradwohl, G., Dierich, A., LeMeur, M. & Guillemot, F. neurogenin3 is required for the development of the four endocrine cell lineages of the pancreas. *Proc. Natl. Acad. Sci. USA* **97**, 1607–1611 (2000).
5. Gu, G., Dubauskaite, J. & Melton, D. A. Direct evidence for the pancreatic lineage: NGN3+ cells are islet progenitors and are distinct from duct progenitors. *Development* **129**, 2447–2457 (2002).
6. Heller, R. S. et al. Genetic determinants of pancreatic epsilon-cell development. *Dev. Biol.* **286**, 217–224 (2005).
7. Desgraz, R. & Herrera, P. L. Pancreatic neurogenin 3-expressing cells are unipotent islet precursors. *Development* **136**, 3567–3574 (2009).
8. Byrnes, L. E. et al. Lineage dynamics of murine pancreatic development at single-cell resolution. *Nat. Commun.* **9**, 3922 (2018).
9. Scavuzzo, M. A. et al. Endocrine lineage biases arise in temporally distinct endocrine progenitors during pancreatic morphogenesis. *Nat. Commun.* **9**, 3356 (2018).
10. Yu, X. X. et al. Defining multistep cell fate decision pathways during pancreatic development at single-cell resolution. *EMBO J.* **38**, e100164 (2019).
11. van Gurp, L. et al. A transcriptomic roadmap to alpha- and beta-cell differentiation in the embryonic pancreas. *Development* **146**, dev173716 (2019).
12. Bastidas-Ponce, A. et al. Comprehensive single cell mRNA profiling reveals a detailed roadmap for pancreatic endocrinogenesis. *Development* **146**, dev173849 (2019).
13. Jennings, R. E., Berry, A. A., Strutt, J. P., Gerrard, D. T. & Hanley, N. A. Human pancreas development. *Development* **142**, 3126–3137 (2015).
14. Baeyens, L. et al. (Re)generating human beta cells: status, pitfalls, and perspectives. *Physiol. Rev.* **98**, 1143–1167 (2018).
15. Petersen, M. B. K., Goncalves, C. A. C., Kim, Y. H. & Grapin-Botton, A. Recapitulating and deciphering human pancreas development from human pluripotent stem cells in a dish. *Curr. Top. Dev. Biol.* **129**, 143–190 (2018).
16. Jeon, J., Correa-Medina, M., Ricordi, C., Edlund, H. & Diez, J. A. Endocrine cell clustering during human pancreas development. *J. Histochem. Cytochem.* **57**, 811–824 (2009).
17. Piper, K. et al. Beta cell differentiation during early human pancreas development. *J. Endocrinol.* **181**, 11–23 (2004).
18. Liu, J. et al. Neurog3-independent methylation is the earliest detectable mark distinguishing pancreatic progenitor identity. *Dev. Cell* **48**, 49–63 (2019).
19. Sharon, N. et al. A peninsular structure coordinates asynchronous differentiation with morphogenesis to generate pancreatic islets. *Cell* **176**, 790–804 (2019).
20. Krentz, N. A. J. et al. Single-cell transcriptome profiling of mouse and hESC-derived pancreatic progenitors. *Stem Cell Rep.* **11**, 1551–1564 (2018).
21. Yu, X. X. & Xu, C. R. Understanding generation and regeneration of pancreatic beta cells from a single-cell perspective. *Development* **147**, dev179051 (2020).
22. Ziegenhain, C. et al. Comparative analysis of single-cell RNA sequencing methods. *Mol. Cell* **65**, 631–643 (2017).
23. Picelli, S. et al. Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014).
24. Dong, J. et al. Single-cell RNA-seq analysis unveils a prevalent epithelial/mesenchymal hybrid state during mouse organogenesis. *Genome Biol.* **19**, 31 (2018).
25. Islam, S. et al. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res.* **21**, 1160–1167 (2011).
26. van Dam, S., Vosa, U., van der Graaf, A., Franke, L. & de Magalhaes, J. P. Gene co-expression analysis for functional classification and gene-disease predictions. *Brief Bioinform.* **19**, 575–592 (2018).
27. Iacono, G., Massoni-Badosa, R. & Heyn, H. Single-cell transcriptomics unveils gene regulatory network plasticity. *Genome Biol.* **20**, 110 (2019).
28. Pollen, A. A. et al. Establishing cerebral organoids as models of human-specific brain evolution. *Cell* **176**, 743–756 (2019).
29. Wierup, N., Yang, S., McEvilly, R. J., Mulder, H. & Sundler, F. Ghrelin is expressed in a novel endocrine cell type in developing rat islets and inhibits insulin secretion from INS-1 (832/13) cells. *J. Histochem. Cytochem.* **52**, 301–310 (2004).
30. Wierup, N., Svensson, H., Mulder, H. & Sundler, F. The ghrelin cell: a novel developmentally regulated islet cell in the human pancreas. *Regul. Pept.* **107**, 63–69 (2002).
31. Yu, X. X. et al. Dynamics of chromatin marks and the role of JMJD3 during pancreatic endocrine cell fate commitment. *Development* **145**, dev163162 (2018).
32. Pan, F. C. & Wright, C. Pancreas organogenesis: from bud to plexus to gland. *Dev. Dyn.* **240**, 530–565 (2011).
33. Wang, X. et al. Quantitative analysis of pancreatic polypeptide cell distribution in the human pancreas. *PLoS One* **8**, e55501 (2013).
34. Herrera, P. L. et al. Embryogenesis of the murine endocrine pancreas; early expression of pancreatic polypeptide gene. *Development* **113**, 1257–1265 (1991).
35. Collombat, P. et al. Opposing actions of Arx and Pax4 in endocrine pancreas development. *Genes Dev.* **17**, 2591–2603 (2003).

36. Zhang, J., McKenna, L. B., Bogue, C. W. & Kaestner, K. H. The diabetes gene Hhex maintains delta-cell differentiation and islet function. *Genes Dev.* **28**, 829–834 (2014).
37. La Manno, G. et al. RNA velocity of single cells. *Nature* **560**, 494–498 (2018).
38. Pan, F. C., Brissova, M., Powers, A. C., Pfaff, S. & Wright, C. V. Inactivating the permanent neonatal diabetes gene Mnx1 switches insulin-producing beta-cells to a delta-like fate and reveals a facultative proliferative capacity in aged beta-cells. *Development* **142**, 3637–3648 (2015).
39. Qiu, X. et al. Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods* **14**, 979–982 (2017).
40. Li, J. et al. Single-cell transcriptomes reveal characteristic features of human pancreatic islet cell types. *EMBO Rep.* **17**, 178–187 (2016).
41. Artner, I. et al. MafB is required for islet beta cell maturation. *Proc. Natl. Acad. Sci. USA* **104**, 3853–3858 (2007).
42. Nishimura, W. et al. A switch from MafB to MafA expression accompanies differentiation to pancreatic beta-cells. *Dev. Biol.* **293**, 526–539 (2006).
43. Henseleit, K. D. et al. NKX6 transcription factor activity is required for alpha- and beta-cell development in the pancreas. *Development* **132**, 3139–3149 (2005).
44. Nelson, S. B., Schaffer, A. E. & Sander, M. The transcription factors Nkx6.1 and Nkx6.2 possess equivalent activities in promoting beta-cell fate specification in Pdx1+ pancreatic progenitor cells. *Development* **134**, 2491–2500 (2007).
45. Sussel, L. et al. Mice lacking the homeodomain transcription factor Nkx2.2 have diabetes due to arrested differentiation of pancreatic beta cells. *Development* **125**, 2213–2221 (1998).
46. Arnes, L., Hill, J. T., Gross, S., Magnuson, M. A. & Sussel, L. Ghrelin expression in the mouse pancreas defines a unique multipotent progenitor population. *PLoS One* **7**, e52026 (2012).
47. Salisbury, R. J. et al. The window period of NEUROGENIN3 during human gestation. *Islets* **6**, e954436 (2014).
48. Ramond, C. et al. Reconstructing human pancreatic differentiation by mapping specific cell populations during development. *Elife* **6**, e27564 (2017).
49. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902 (2019).
50. Dolensek, J., Rupnik, M. S. & Stozer, A. Structural similarities and differences between the human and the mouse pancreas. *Islets* **7**, e1024405 (2015).
51. Andralojc, K. M. et al. Ghrelin-producing epsilon cells in the developing and adult human pancreas. *Diabetologia* **52**, 486–493 (2009).
52. Miyatsuka, T., Li, Z. & German, M. S. Chronology of islet differentiation revealed by temporal cell labeling. *Diabetes* **58**, 1863–1868 (2009).
53. Johansson, K. A. et al. Temporal control of neurogenin3 activity in pancreas progenitors reveals competence windows for the generation of different endocrine cell types. *Dev. Cell* **12**, 457–465 (2007).
54. Bankaitis, E. D., Bechard, M. E. & Wright, C. V. Feedback control of growth, differentiation, and morphogenesis of pancreatic endocrine progenitors in an epithelial plexus niche. *Genes Dev.* **29**, 2203–2216 (2015).
55. Cozzitorto, C. et al. A specialized niche in the pancreatic microenvironment promotes endocrine differentiation. *Dev. Cell* **55**, 150–162 (2020).
56. Stoffers, D. A., Zinkin, N. T., Stanojevic, V., Clarke, W. L. & Habener, J. F. Pancreatic agenesis attributable to a single nucleotide deletion in the human IPF1 gene coding sequence. *Nat. Genet.* **15**, 106–110 (1997).
57. Bonnefond, A. et al. Transcription factor gene MNX1 is a novel cause of permanent neonatal diabetes in a consanguineous family. *Diabetes Metab.* **39**, 276–280 (2013).
58. Flanagan, S. E. et al. Analysis of transcription factors key for mouse pancreatic development establishes NKX2-2 and MNX1 mutations as causes of neonatal diabetes in man. *Cell Metab.* **19**, 146–154 (2014).
59. Rubio-Cabezas, O. et al. Homozygous mutations in NEUROD1 are responsible for a novel syndrome of permanent neonatal diabetes and neurological abnormalities. *Diabetes* **59**, 2326–2331 (2010).
60. Wang, D. et al. Long-term expansion of pancreatic islet organoids from resident Procr(+) progenitors. *Cell* **180**, 1198–1211 (2020).
61. Jennings, R. E. et al. Development of the human pancreas from foregut to endocrine commitment. *Diabetes* **62**, 3514–3522 (2013).
62. Schonhoff, S. E., Giel-Moloney, M. & Leiter, A. B. Neurogenin 3-expressing progenitor cells in the gastrointestinal tract differentiate into both endocrine and non-endocrine cell types. *Dev. Biol.* **270**, 443–454 (2004).
63. Kim, D. et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013).
64. Anders, S., Pyl, P. T. & Huber, W. HTSeq — a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
65. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
66. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
67. Smith, T., Heger, A. & Sudbery, I. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res.* **27**, 491–499 (2017).
68. Brennecke, P. et al. Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Methods* **10**, 1093–1095 (2013).
69. Quinn, T. P., Richardson, M. F., Lovell, D. & Crowley, T. M. propr: an R-package for identifying proportionally abundant features using compositional data analysis. *Sci. Rep.* **7**, 16252 (2017).
70. Skinnider, M. A., Squair, J. W. & Foster, L. J. Evaluating measures of association for single-cell transcriptomics. *Nat. Methods* **16**, 381–386 (2019).
71. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *Int. J. Complex Syst.* **1695**, 1–9 (2006).
72. Zhang, H. M. et al. AnimalTFDB: a comprehensive animal transcription factor database. *Nucl. Acids Res.* **40**, D144–D149 (2012).
73. Andrews, T. S. & Hemberg, M. M3Drop: dropout-bas d feature selection for scRNASeq. *Bioinformatics* **35**, 2865–2867 (2019).
74. Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* **36**, 421–427 (2018).
75. Falcon, S. & Gentleman, R. Using GOstats to test gene lists for GO term association. *Bioinformatics* **23**, 257–258 (2007).
76. Street, K. et al. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* **19**, 477 (2018).
77. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).