

A multimodal cross-species comparison of pancreas development

Received: 3 July 2025

Accepted: 25 September 2025

Published online: 22 October 2025



Kaiyuan Yang ^{1,2,18}, Hannah Spitzer ^{3,4,18}, Michael Sterr^{1,2}, Karin Hrovatin ^{3,5}, Sean de la O^{6,7,8}, Xinghao Zhang^{9,10}, Eunike Sawitning Ayu Setyono ^{1,2}, Minhaz Ud-Dean¹¹, Thomas Walzthoeni¹¹, Krzysztof Flisikowski¹², Tatiana Flisikowska ¹², Angelika Schnieke ¹², Katharina Scheibner^{1,2}, James M. Wells ^{9,10,13}, Julie B. Sneddon ^{6,7,8}, Barbara Kessler^{2,14,15}, Eckhard Wolf ^{2,14,15}, Elisabeth Kemter ^{2,14,15}, Fabian J. Theis ^{3,5,16} ✉ & Heiko Lickert ^{1,2,17} ✉

Human pancreas development remains incompletely characterized due to restricted sample access. We investigate whether pigs resemble humans in pancreas development, offering a complementary large-animal model. As pig pancreas organogenesis is unexplored, we first annotate developmental hallmarks throughout its 114-day gestation. Building on this, we construct a pig single-cell multiome pancreas atlas across all trimesters. Cross-species comparisons reveal pigs resemble humans more closely than mice in developmental tempo, epigenetic and transcriptional regulation, and gene regulatory networks. This further extends to progenitor dynamics and endocrine fate acquisition. Transcription factors regulated by *NEUROG3*, the endocrine master regulator, are over 50% conserved between pig and human, many being validated in human stem cell models. Notably, we uncover that during embryonic development, emerging beta-cell heterogeneity coincides with a species-conserved primed endocrine cell (PEC) population alongside *NEUROG3*-expressing cells. Overall, our work lays the foundation for comparative investigations and offers unprecedented insights into evolutionarily conserved pancreas organogenesis mechanisms across animal models.

Mice have been the main mammalian model to investigate the evolutionarily conserved basic principles of development and disease. As human fetal samples are largely inaccessible due to ethical and practical restrictions, our knowledge of pancreas development is mainly built on extensive investigations in mice^{1,2}. In brief, from foregut endoderm, ventral and dorsal pancreatic buds are specified, evaginated, and fused over the midline to form the organ anlage. This multipotent progenitor epithelium undergoes apical-basal polarization, microlumen formation, single-layer stratification, and branching morphogenesis, which is concomitant with tip-trunk patterning and differentiation to form the acinar, ductal, and endocrine compartments of the final pancreas (Fig. 1a)^{1,3–7}. With this blueprint,

the signals and factors orchestrating *in vivo* development are applied to human pluripotent stem cell differentiation *in vitro* to generate endocrine islet cells, such as glucagon-producing alpha cells and insulin-producing beta cells, enabling stem cell-derived islet replacement therapy to treat diabetes^{7–10}. However, these protocols cannot produce fully functional islet cells while eliminating undifferentiated and off-target cell types^{11,12}, partly resulting from the inherent differences in developmental timescales, neighboring tissue interactions, and spatiotemporal gene expression and regulation between mouse and human pancreas differentiation and morphogenesis¹³. These differences highlight the need for cross-species comparisons with additional model systems to uncover

conserved mechanisms of organogenesis and to bridge the translational gap between mice and humans.

Pigs have coevolved with humans over the past 10,000 years, a period when pig domestication coincided with human agricultural civilization¹⁴. Despite diverging from humans earlier than mice (94 vs. 87 million years ago)¹⁵, pigs retain genomic feature similarity to humans compared to the rapidly evolving mouse lineage¹⁶. Moreover, as omnivorous animals, pigs resemble humans in metabolism and physiology. Pig organs share anatomical and functional features with humans, making them a favored option for xenotransplantation with compatible organ size and fewer ethical concerns compared to non-human primates¹⁷. Porcine islets show transcriptional characteristics similar to human islets¹⁸ and represent a potential source for xenotransplantation¹⁹, since pigs and humans have similar glycemic control and identical insulin amino acid sequence. This allowed insulin therapy using pig insulin before recombinant human insulin became available^{20–22}. However, whether these shared traits position pigs as a relevant model for human pancreas development remains unexplored, particularly regarding the molecular and (epi)genetic mechanisms driving embryonic development and organogenesis. Given pigs' anatomical and physiological similarities to humans and their extended gestational period (114 days vs. 21 in mice, 280 in humans), we reasoned that pigs could serve as a large animal model to complement the existing rodent models, bridging the translational gap in understanding pancreas development from mice to humans.

Here, we utilize temporally resolved single-cell multi-omics to compare pancreas development in mice, pigs, and humans. Our analysis demonstrates the complementary potential of the pig model for identifying both species-specific and evolutionarily conserved mechanisms of pancreas organogenesis, morphogenesis, and differentiation.

Results

The developing pig pancreas

To compare pancreas development across species (Fig. 1a), we first defined the unexplored hallmarks of pig pancreas development during the 114-day gestation by identifying major lineage markers known in mouse and human (Fig. 1b–g). The pig pancreas primordium emerged around embryonic day (E)18, when the earliest pancreatic transcription factor (TF) PDX1 was detected in both ventral and dorsal foregut, with the former having a thicker evaginating layer of cells as reported in human embryos (Supplementary Fig. 1a)²³. At E20, pancreas organogenesis initiated with the apparent dorsal and ventral buds (T1, Fig. 1a) that comprised an unpolarized multi-layered progenitor epithelium, harboring glucagon⁺ alpha cells and very few insulin⁺ beta cells (Fig. 1b and Supplementary Fig. 1b). The early emergence of hormone⁺ cell resembles the first wave of Neurogenin-3 (NEUROG3)-mediated endocrinogenesis in mice during the primary (1°) transition⁵, which is absent at the corresponding stage in human when the endocrine transcription factors NEUROG3 and NKX2-2 remain undetectable²³. Upon the fusion of the two pancreatic buds at E30, the expression of NEUROG3 protein and transcripts diminished from the pancreas (Fig. 1c and Supplementary Fig. 1c, h). At E40, concurrent with pancreatic epithelial polarization, stratification and tip-trunk patterning (T2, Fig. 1a), NEUROG3 reappeared, initiating the second wave of endocrinogenesis comparable to the secondary (2°) transition in mice (Fig. 1d and Supplementary Fig. 1d). By E54, CPA1⁺ tip and SOX9⁺ trunk domains were clearly established (Fig. 1e and Supplementary Fig. 1e). SOX9 remained faintly detectable in CPA1⁺ tip cells that were predisposed to acinar fate at E63, while endocrine cell clusters started to appear (Fig. 1f and Supplementary Fig. 1f). From E85 onwards, endocrine (NKX6-1⁺), ductal (SOX9⁺) and acinar (GATA4⁺) compartments continued to segregate, with these TFs showing almost exclusive expression patterns (Fig. 1g and Supplementary Fig. 1g, j). During this time, pig proto-islets appeared (T3, Fig. 1a) and formed an

intermingled islet architecture near birth (Supplementary Fig. 1k), which resembles the postnatal islet architecture in human, but not the typical core (beta cells)-mantle (alpha cells) islet structure known from mice^{24–26}.

Aligning the timing of pancreas development milestones (i.e. time points labeled in Fig. 1a) in human, pig and mouse allowed us to compare the developmental tempo across species (Fig. 1h). Overall, pig pancreas morphogenesis and differentiation speed showed a closer resemblance to humans when compared to mice, particularly during the 2° transition. The formation of the pancreatic anlage in the form of two buds (T1) occupied 10% of the duration of human gestation, 12% in mouse and 17% in pig. In contrast, pancreatic morphogenesis (T2) and islet formation (T3) progressed much faster in mice (42%), contrary to the longer duration in human (82%) and pig (65%), when both species underwent extended acinar terminal differentiation and islet remodeling²³.

To capture the dynamic transcriptional changes of lineage allocation during pig pancreas development, we performed 10X single-cell RNA sequencing (scRNA-seq) on 124,869 cells isolated from pancreata across all three trimesters (Supplementary Fig. 2a and Supplementary Data 1). The resulting dataset was subjected to stringent doublet removal: cells consistently identified as doublets by >3 among the six methods used were removed, and entire clusters with a doublet frequency >70% were excluded (see Methods for details). From this filtered dataset, we extracted pancreatic epithelial cells that co-expressed *CDH1* and *EPCAM* (Supplementary Fig. 2b) and identified eight clusters (Fig. 1i, j, Supplementary Fig. 2c, d and Supplementary Data 2): Ductal (*SOX9/SLC44A4*), Acinar (*CPA1/CEL*), NGN3 (*NEUROG3/TUBB2B*, endocrine progenitors, NGN3 is used as a cluster name to differentiate from NEUROG3 protein and/or mRNA), FEV (*FEV/CHGB*, endocrine precursors with low *NEUROG3* expression), Beta (*INS/G6PC2*), Alpha (*GCG/IRX1*), Delta (*SST/HHEX*), and PP (*PPY/ETV1*) cell clusters. Among the top ranked genes, published human markers were also found in the pig pancreas, such as *MDK* in the NGN3, *DDC* in FEV, *ASB9* in Beta, *TTR* in Alpha, *RBP4* in Delta cell clusters^{27–30}. Diabetes risk related genes were expressed in a cell-type specific manner, e.g. *ABLI1* in Ductal, *TAGLN3* in NGN3, *DIRAS3* in Alpha, *CFAP61* in all endocrine cell clusters^{31–34}. We further identified two molecularly distinct cell clusters: 1) A predicted multipotent progenitor cell (MPC) cluster, which transiently existed during E23–40 and co-expressed key pancreatic progenitor TFs (*PDX1*, *PTF1A*, *SOX9*, *NKX6-1*, *PROX1*). This predicted MPC cluster resolved from E54 onwards, when exocrine and endocrine lineages clearly separated and committed to differentiation; 2) A primed endocrine cell (PEC) cluster first emerged at E23 alongside sparse *NEUROG3*⁺ endocrine progenitors. This population, which persisted throughout all subsequent stages, exhibited features of endocrine cells and expressed genes coding for cytoskeletal components (*TUBA1B*, *TUBB*, *STMN1*) and cell cycle regulators (*H2AFZ*, *PCLAF*).

The emergence of the identified cell clusters in the scRNA-seq atlas was consistent with our immunofluorescence analysis of the pig pancreas, thereby verifying the cell type annotations. For instance, the NGN3 cluster contained only two cells in the scRNA-seq atlas at E33, when the NEUROG3 protein and transcript were almost undetectable in the pancreas (Fig. 1c, i and Supplementary Fig. 1c, h); whereas the *CPA1*⁺ Acinar, *SST*⁺ Delta and *PPY*⁺ PP clusters appeared after E40, when these proteins could be detected in the pancreas (Fig. 1i and Supplementary Fig. 1d, e, i).

Cross-species multiome atlas comparison

To compare the transcriptional features of all epithelial pancreatic cell types across species, we generated atlases of human and mouse pancreas development by integrating published 10X scRNA-seq datasets^{30,35–42}. The human atlas had 188,488 cells from 7–20 week-post-conception (wpc), while the mouse atlas had 135,575 cells covering E8.5–18.5. Cells were clustered, annotated, and the pancreatic

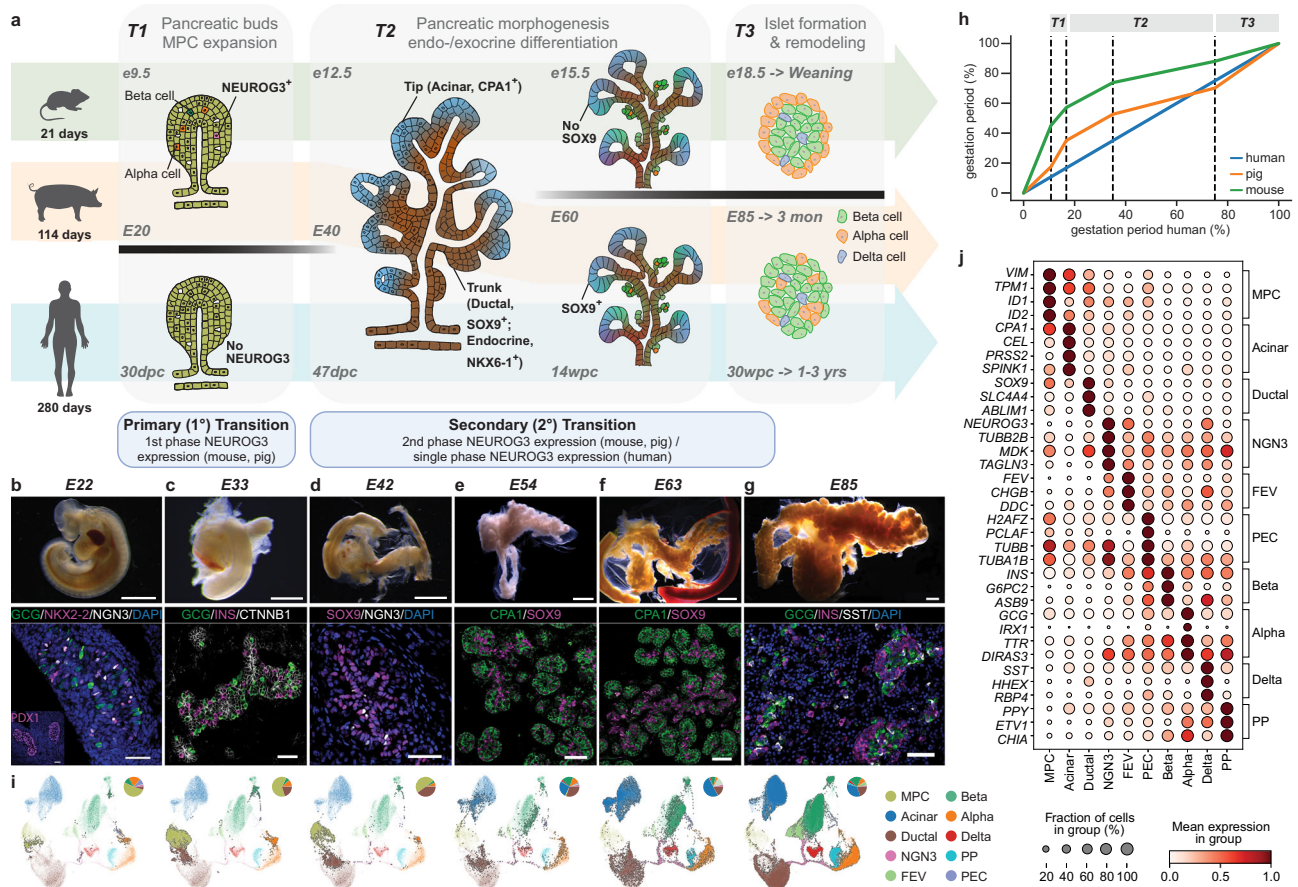


Fig. 1 | A detailed roadmap of pig pancreas organogenesis. a Schematic comparison of pancreas organogenesis in mouse, pig and human^{1,6,7}. Pancreas organogenesis initiates when the pancreatic buds (dorsal bud shown) emerge from the foregut endoderm while predicted multipotent pancreatic progenitor cells (MPC) expand into a multi-layer epithelium (T1). In mouse and pig but not human, neurogenin 3-induced endocrinogenesis marks the primary (1°) transition, generating alpha cells and few beta cells. Pancreatic morphogenesis (T2) occurs after the fusion of the pancreatic buds. The epithelium undergoes polarization, microlumen formation and coalescence into the near single-layer epithelial tree, which is subsequently patterned into trunk and tip domains during branching morphogenesis. This coincides with progenitor differentiation forming exocrine and endocrine compartments, referred to as secondary (2°) transition when the bulk of beta cells emerge. During T3, the pancreatic ductal and acinar cells proliferate and terminally differentiate, while delaminating endocrine cells form proto-islets of various sizes.

b–g Bright-field images of wild-type pig embryos or embryonic pancreas (scale bar 2 mm) and immunofluorescence identification of lineage markers highlighting differentiation events in tissue sections (scale bar 50 μ m) at different time points. Images are representative of 3 samples per time point. **h** Schematic comparison of mouse and pig development speeds relative to human, showing the timing of each pancreas developmental milestone (labeled in **a**) as a percentage of gestation duration marked with dashed lines. **i** Uniform Manifold Approximation Projection (UMAP) plots showing integrated pig pancreas atlas and cluster changes from E22–85. Cells at each developmental stage (see **b–g**) are highlighted and colored by cell type, with a pie chart of relative cell type composition at the upper right corner. **j** Dot plot showing mean gene expression of marker genes for each cluster of the integrated pig pancreas atlas. **i** and **j**: scRNA-seq of pancreatic cells from wild-type and *INS-eGFP* pigs. Detailed sample information is provided in Supplementary Data 1.

epithelial lineage selected as for the pig dataset (Supplementary Fig. 3a–f). All pig pancreatic cell clusters were identified in human and mouse datasets except the PEC cluster, which was absent potentially due to its low abundance in the analyzed samples (Fig. 2a, Supplementary Fig. 3g, h and Supplementary Data 2). PP (*PPY/ETV1*) cells did not separate from the Alpha cluster in mouse (Supplementary Fig. 4a). *FEV*⁺ cells in human were annotated together with the NGN3 cluster as EP (endocrine progenitor) cluster due to high *NEUROG3* expression (Supplementary Fig. 4b). The Epsilon (*GRHL*) cluster was identified in human and mouse datasets, but surprisingly not in the pig pancreas. The Epsilon marker genes from human and mouse did not show significant enrichment in any cell clusters in the pig (Supplementary Fig. 4c).

To assess the comparability of pancreatic cell types across species, we performed a correlation analysis using the average expression of species-conserved highly variable genes in each cluster, which reflected the global transcriptional programs of each species. Pig and human clusters had an overall stronger correlation ($r = 0.6$ – 0.7),

whereas MPC, Beta, and EP clusters in mouse showed the lowest correlation to human clusters ($r = 0.45$ – 0.54 ; Fig. 2b and Supplementary Fig. 4d). We further examined the significantly upregulated differentially expressed genes (DEGs) of each cluster across species, which can capture the subtle species-specific differences in gene expression. Although the pig genome annotation is not as complete as that of human and mouse, the MPC, Beta, and Ductal clusters of pig and human had more overlapping total DEGs and TFs. In contrast, in the Acinar and Alpha clusters, mouse and human shared more total DEGs, but not TFs (Supplementary Fig. 4e). Each species displayed distinct cell-type-specific expression patterns of DEGs (Fig. 2c, Supplementary Fig. 4f, and Supplementary Data 3). For example, the top TFs *SOX6* and *SOX9* of human MPC cluster showed enriched expression in the Ductal cluster of pig and mouse. The expression patterns of the top TFs in endocrine progenitors were relatively conserved among the three species. Among the top TFs of the human Beta cluster, the maturation TF *MAFA* was detected in pig but absent in mouse beta cells; *PLAGL1*, related to transient neonatal diabetes^{43,44}, appeared in pig beta cells,

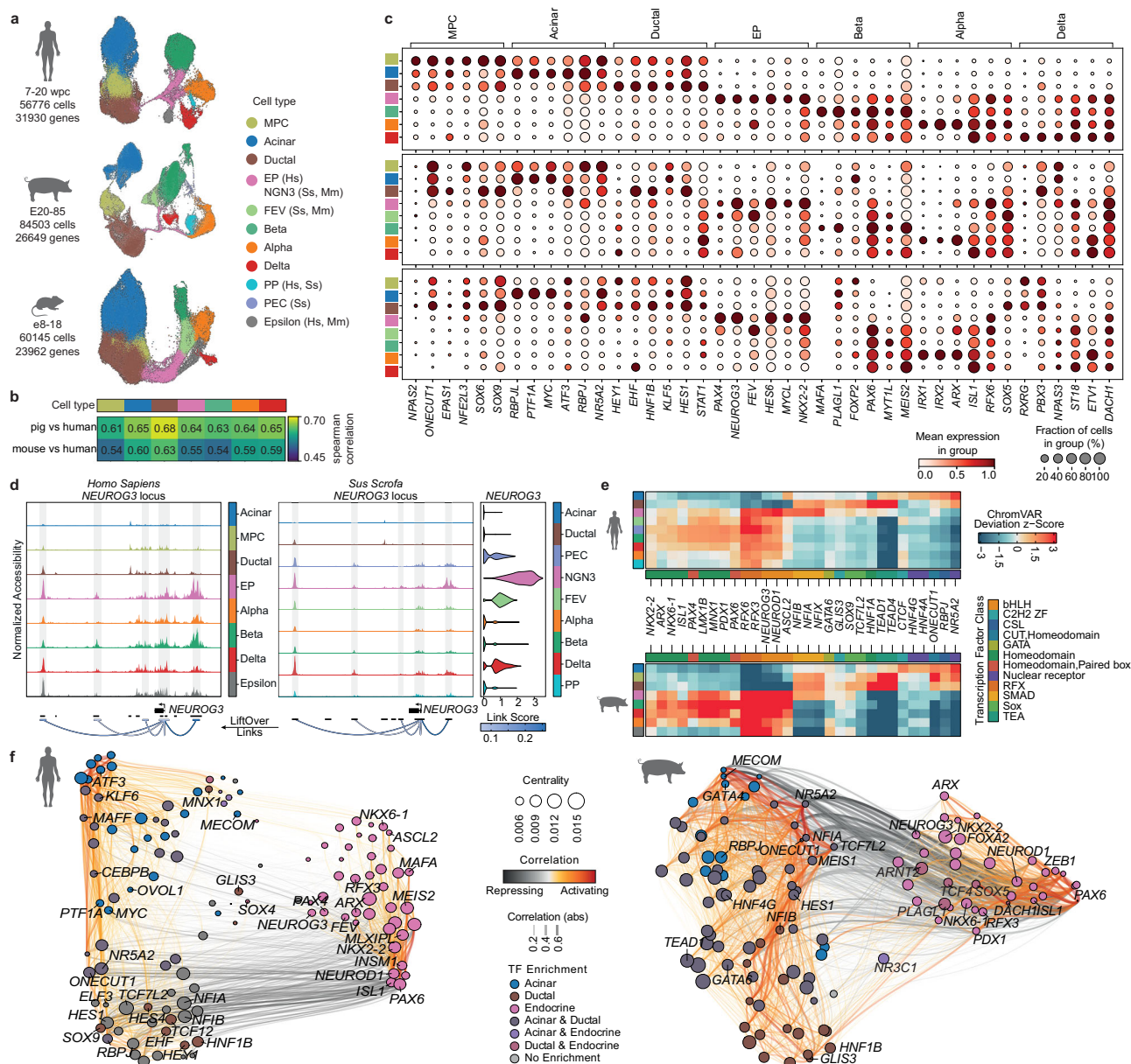


Fig. 2 | Cross-species multiome atlas comparison reveals conservation between species. **a** UMAPs of integrated human, pig, and mouse atlases of pancreas development. Cells were colored by cell type, with the same colors indicating the same cell types across species (Hs, *Homo sapiens*; Ss, *Sus scrofa*; Mm, *Mus musculus*). **b** Spearman correlation of mean normalized gene counts per cluster comparing pig-human and mouse-human pairs (based on human orthologs). Analysis is limited to the 851 genes shared in the intersection of all species' 4000 highly variable genes. **c** Dot plot showing mean expression of cell-type-specific differentially expressed TFs in human (one-vs-rest analysis using edgeR; FDR-corrected p-value < 0.05; Supplementary Data 3). Top genes with the highest logfold change that are expressed in > 20% of cells of the cluster are shown across all clusters (square color = cell type color in **a**) for all three species. Genes are mapped to human orthologs. Dot size represents the fraction of expressing cells per cluster (logarithmic scale). **d** Coverage plots showing pig-human conserved *NEUROG3*

genomic regions. Link positions were converted from pig to human genome assembly with UCSC liftOver tool (<http://genome.ucsc.edu>)¹⁹. Pseudo-bulk accessibility tracks were used to visualize DNA accessibility in a region by averaging signals from all cells within a cluster. **e** Heatmap of top differentially active motifs across cell types computed with chromVAR²¹ using human 12wpc scATAC-seq (top) and pig multiome (bottom) datasets. **f** UMAP visualizations of the inferred TF network from human scGLUE⁴⁸-integrated scRNA/ATAC-seq (left) and pig multiome (right) data. Nodes represent TFs, colored by their highly expressed clusters and sized by network centrality. Edges show TF-target interaction strengths (orange: activating; gray: inhibiting). **a–c**: scRNA-seq of pancreatic cells from wild-type and *INS*-eGFP pigs. **d–f**: Multiome analysis of pancreatic cells from PTF1A-codon-improved-Cre/ROSA-mTmG pigs. Detailed sample information is provided in Supplementary Data 1.

while mainly being expressed in mouse acinar cells. The TF *ETV1* showed enrichment in human delta cells, though in pig and mouse, it was highly expressed in alpha cells.

Conserved cell-type gene expression patterns indicated similarities in higher-order chromatin regulation mediated via cis-regulatory elements (CREs, e.g., enhancers and silencers) to control gene expression⁴⁵. To compare pig and human CREs and upstream

transcriptional regulators of pancreatic cell fate decisions, we additionally performed single-cell 10X multiome sequencing to assess transcriptional changes and chromatin accessibility simultaneously. Samples were collected during the 2^o transition of pig development at E45-85, resulting in 29,072 nuclei (Supplementary Data 1). Cluster label transfer from the pig scRNA-seq atlas identified all cell types, except that the MPC cluster had already diminished, while the Epsilon cells

remained undetectable (Supplementary Fig. 5a, b). We further analyzed single-cell transposase-accessible chromatin with sequencing (scATAC-seq) data from human 12 wpc pancreas³⁰, which had 5592 nuclei. Cell clusters were annotated via label transfer from the human scRNA-seq atlas, and pancreatic epithelial cells were used for further analysis (Supplementary Fig. 5c).

Highly conserved CREs were identified when comparing the sequences between pig and human key lineage regulators, such as the master regulator of endocrinogenesis, *NEUROG3* (Fig. 2d). Further computing cell-type specific motif activities revealed pig-human conserved active TFs towards endocrine and exocrine lineages (Fig. 2e and Supplementary Fig. 5d). Both Acinar and Ductal clusters had active motifs of TFs determining acinar fate differentiation (*NR5A2*, *PTF1A* and *RBPJ*), reflecting the gradual resolution of chromatin and transcription factor profiles in acinar and ductal lineages during terminal differentiation in human and pig. In addition, two *MODY* genes (*HNF4A* and *HNF4G*) that regulate the growth and function of beta cells, were found active in acinar cells. The Ductal cluster conserved TFs contained Hippo effector genes (*TEAD1* and *TEAD4*) and TFs known to be expressed in the embryonic ductal tree, such as *HNF1A*, *TCF7L2*, *SOX9*, *GLIS3*, and *GATA6*. The human-pig conserved TFs in the EP cluster included known regulators of endocrinogenesis (*NEUROG3*, *RFX3*, *RFX6*, and *NEUROD1*), members of the Nuclear Factor 1 family (*NF1A*, *NF1B*, and *NF1X*), and the intestinal stem cell identity TF *ASCL2*⁴⁶. Endocrine cell clusters also had a large panel of TFs shared between pig and human (*PAX6*, *PDX1*, *MX1*, *LMX1B*, *PAX4*, *ISL1*, *NKX6-1*, *ARX*, and *NKX2-2*).

Given the conservation of cell-type-specific active motifs between pig and human, we used Pando⁴⁷ to construct GRNs using human scGLUE⁴⁸-integrated scRNA/ATAC-seq and pig multiome (joint scRNA/ATAC-seq) pancreas datasets. This enabled detailed examination of relationships between pig-human conserved TFs, their potential target gene expression, and regulatory-site accessibility across cell types and species. The generated GRNs revealed groups of pig-human conserved TFs involved in the differentiation and cell state transition of pancreatic lineages (Fig. 2f and Supplementary Fig. 5e). Ductal-specific TF modules (*HNF1B*, *GLIS3*, and *EHF*), as well as acinar-specific modules (*MECOM*, *XBPI*, and *STAT3*), formed interconnected networks with a set of TFs that showed enrichment in both acinar and ductal lineages (*HES1*, *NR5A2*, *ONECUT1*, *TCF7L2*, *MAFF*, *REST*, and *MEIS1*). Among TF modules linked to endocrine differentiation, *SOX4* (not expressed in pigs) was enriched in both ductal and endocrine lineages in humans. Additionally, *NR3C1*, implicated in regulating islet gene programs and conferring genetic risk of type 2 diabetes⁴⁹, was enriched in both the acinar and endocrine lineages in pig, whereas in human, it was only found in endocrine TF modules. Nevertheless, a large group of endocrine TF modules was conserved between pig and human, including *ST18*⁵⁰, *DACH1*⁵¹, *PLAGL1*^{43,44}, and *CDCC88A*⁵², genes linked to beta cell mass and function.

Pig-human conserved endocrinogenesis

To explicate the mechanisms underlying endocrine fate allocation, we focused on the endocrine progenitor clusters stemming from trunk (ductal) cells with low *NEUROG3* expression, progressing towards the cells that clearly diverged into either the alpha or beta cell fate (Fig. 3a–c). To infer developmental trajectories underlying pancreatic lineage acquisition, we applied CellRank with the Palantir Pseudotime kernel⁵³ to the integrated scRNA-seq atlases to estimate cell states, compute cell-cell transition probabilities, and map cell fates. CellRank correctly predicted the differentiated pancreatic cell clusters as the terminal states (Fig. 3a–c and Supplementary Fig. 6a), providing a reliable transition matrix to infer lineage trajectories in all three species. This revealed distinct differentiation programs in endocrine progenitors of pig and human, opposed to mouse (Fig. 3b). Specifically, at the stage of high *NEUROG3* expression, the progenitors in pig and human already segregated toward either the alpha or beta cell fate.

The presence of *FEV*-expressing endocrine precursors was limited to the beta cell lineage alone. Similarly, the same differentiation pattern was identified by an independent analysis with a subset of the human scRNA-seq atlas data³⁰. In contrast, the separation of alpha and beta cell lineages occurred within the *FEV*-expressing endocrine precursors in mice (Fig. 3b and Supplementary Fig. 6b, c).

We next correlated gene expression with CellRank-inferred lineage probabilities and computed putative driver genes of alpha or beta cell fate for each species. Among the orthologous genes mapped across species, pig and human had more overlapping alpha or beta lineage drivers compared to the set shared between mouse and human, regardless of correlation strength (Fig. 3d and Supplementary Data 4). We then identified shared and distinct lineage driver genes with a correlation score >0.7 across species (Supplementary Fig. 6d). To compare expression dynamics of orthologous lineage drivers during endocrine progenitor fate specification, we performed hierarchical clustering in humans to define gene clusters and enriched pathways, then mapped conserved programs in pigs and mice (Fig. 3e, f and Supplementary Data 4). These clusters captured pseudotemporal expression profiles progressing from early endocrine progenitors to fate-committed endocrine cells. For example, in beta cell lineage specification, clusters 3 and 4 showed a conserved expression cascade of Notch signaling-related genes, likely present in early endocrine progenitors. Cluster 2 was enriched with mTOR and RA signaling genes, which showed high expression at both the beginning and end of endocrine progenitor differentiation. Cluster 1 contained genes highly expressed late in differentiation, suggesting their roles in beta cell fate specification. While we observed species-specific differences in extracellular matrix organization, tight junction formation, and cytoskeletal regulation, core features of endocrine differentiation - particularly incretin synthesis and secretion - remained conserved across all species in both beta and alpha cell lineages, demonstrating evolutionary preservation of key differentiation mechanisms.

To evaluate the suitability of pigs as a model of human endocrine development, we performed a comparative analysis of *NEUROG3* TF regulatory networks derived from native human and pig pancreas multiome data and experimentally derived targets from human stem cell models. Using Pando-inferred GRNs from human scGLUE-integrated scRNA/ATAC-seq data and pig multiome data, we extracted *NEUROG3* networks with significant interactions and module activities (Fig. 3g, h and Supplementary Data 5). Both human and pig *NEUROG3* networks contained >100 TFs, including canonical *NEUROG3* targets, such as *NEUROG3*, *NEUROD1*, *NKX6-1*, *NKX2-2*, and *MLXIPL*. We then benchmarked these against targets identified through human stem cell models. In our human embryonic stem cell (hESC) model^{54,55}, we achieved temporal control of *NEUROG3* induction using tetracycline to precisely match physiological levels and timing required for endocrine differentiation. These hESCs were differentiated stepwise via endoderm and then foregut into pancreatic progenitors, followed by an 8-h *NEUROG3* TF induction. Over 90 high-confidence *NEUROG3* downstream TF targets were identified through integrative analysis of time-series mRNA, ATAC and chromatin immunoprecipitation (ChIP) sequencing data (Fig. 3i and Supplementary Data 5). Additionally, we compared these results with *NEUROG3* target genes identified using a human induced pluripotent stem cell (hiPSC) model⁵⁶, in which an epitope-tagged *NEUROG3* was used for cleavage under targets and release using nuclease (CUT&RUN) to identify *NEUROG3*-bound regions in purified hiPSC-derived EPs (Supplementary Data 5). This revealed 59 conserved *NEUROG3* TF targets between native pig and human pancreas GRNs, including 32 targets not detected in either stem cell model, while the stem cell system verified 20 targets absent in the pig in vivo dataset (Fig. 3i and Supplementary Data 5). Several *NEUROG3* targets, such as *EHF*, *DACH1*, *ST18*, and *MAFF*, were conserved in all models. Using temporally controlled *NEUROG3* TF expression, our hESC model further captured

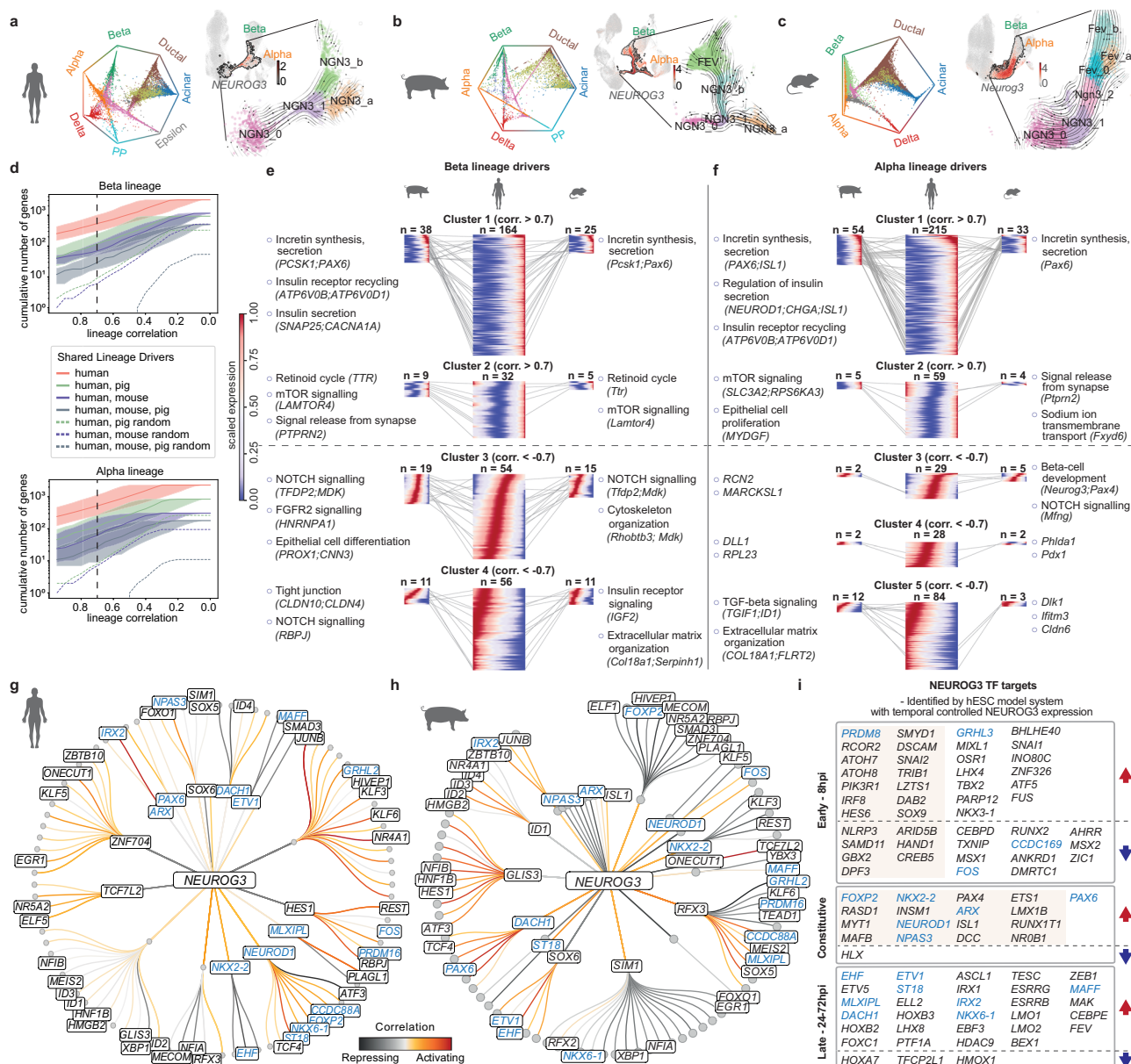


Fig. 3 | Pig-human conserved endocrine fate allocation. **a–c** For each species: circular projection of CellRank-calculated fate probabilities for each cell toward the terminal states (outer labels); and UMAPs detailing endocrine progenitor branching, with the integrated pancreas atlas showing NEUROG3 expression (red) and endocrine progenitor cluster boundaries (black), and an insert showing endocrine progenitor subclusters with overlaid CellRank-inferred trajectories (arrows). **d** Line plots showing the cumulative number of CellRank-derived beta-cell (top) and alpha-cell (bottom) lineage drivers in mouse and pig that overlap with human orthologs, plotted across correlation score thresholds (Supplementary Data 4). Solid lines show significant driver numbers (Benjamini-Hochberg FDR-corrected p-value < 0.05). Shaded regions indicate the number of genes obtained when using the lower and upper bounds of the 95% confidence interval for the corresponding correlation score. **e, f** Heatmaps displaying modeled gene expression patterns for human beta-cell (**e**) and alpha-cell (**f**) lineage driver gene clusters (identified by hierarchical clustering) across pig, human, and mouse along

pseudotemporal trajectories (left to right: 0 → 1). Annotations indicate species-conserved pathways and representative genes for each cluster. (corr., correlation; n, number of conserved genes that are expressed in > 20% of endocrine progenitor subclusters) **g–i** Comparison of NEUROG3 TF targets identified in human/pig pancreas and hESC model (conserved targets in blue). **g, h** Circular GRN graphs showing first- and second-order NEUROG3 targets from human scGLUE-jointed scRNA/ATAC-seq data (**g**) and pig multiomic data (**h**). Nodes represent TFs. Edge color indicates regulatory interaction types (orange, activating; gray, inhibiting); **i** NEUROG3 TF targets in hESC model with inducible NEUROG3 expression. ChIP-seq-identified direct targets are shaded. Differentially expressed TFs comparing cells with/without NEUROG3 expression are indicated by arrows (red, upregulated; blue, downregulated). **a–f** scRNA-seq of pancreatic cells from wild-type and *INS-eGFP* pigs. **h** Multiome analysis of pancreatic cells from *PTF1A*-codon-improved-Cre/*ROSA*-mTmG pigs. Detailed sample information is provided in Supplementary Data 1.

additional conserved early NEUROG3 targets, such as *PRDM8*, *GRHL3* and *FOS*.

Developmental origin of beta cell heterogeneity

We and others have previously reported beta cell heterogeneity in terms of proliferation and maturation regulated by the Wnt/planar cell

polarity (PCP) signaling pathway^{57,58}. However, the timing and mechanisms underlying the origin of beta cell heterogeneity remain unknown. Intriguingly, two heterogeneous beta cell subpopulations formed during the extended period of pig pancreas development. The Beta₀ cluster consisted of cells appearing at all stages, whereas the Beta₁ cluster emerged mainly during the second wave of

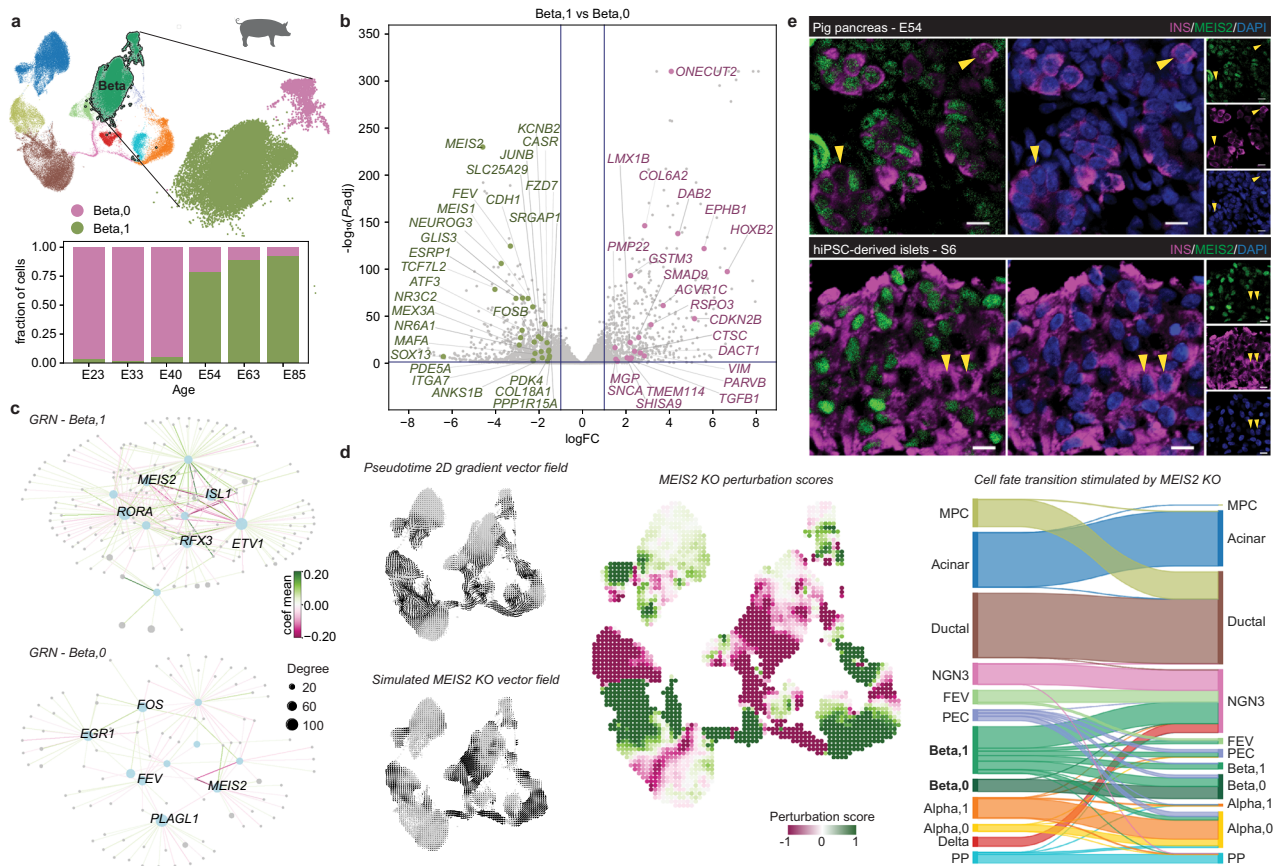


Fig. 4 | Beta cell heterogeneity arises during pancreas development. **a** Detailed view of pig beta cells. (Top) UMAP of pig pancreas atlas with an inset showing a UMAP of the beta cells, colored by subclusters Beta,0 and Beta,1. (Bottom) Distribution of beta cell subclusters at each sampling age. **b** Volcano plot of differentially expressed genes between Beta,0 and Beta,1 (edgeR, FDR-corrected p -value < 0.05 ; Supplementary Data 6). Genes mentioned in core transcriptional regulatory circuits and identified in pathway analysis are highlighted. **c** CellOracle inferred GRN plots showing networks of the top 5 regulons from Beta,0 and Beta,1 clusters. **d** *MEIS2* in silico KO simulation with CellOracle. (Left) UMAPs overlaid with

the Palantir-pseudotime (top) or *MEIS2* KO simulation (bottom) vector fields. (Middle) UMAP colored by *MEIS2* KO perturbation score. (Right) Sankey diagram showing the effect of *MEIS2* KO on cell fate transitions. **e** Immunofluorescence identification of *MEIS2* positive and negative beta cells in wild-type pig pancreas and hiPSC-derived islet sections, scale bar 10 μ m. Images are representative of 3 pig pancreas samples and 3 independent hiPSC differentiations. **a–d**: scRNA-seq of pancreatic cells from wild-type and *INS*-eGFP pigs. Detailed sample information is provided in Supplementary Data 1.

endocrinogenesis (Fig. 4a). Comparing the two beta subclusters unveiled unique differentially expressed gene sets (Fig. 4b and Supplementary Data 6). Beta,1 cells were enriched in TFs of the reported core transcriptional regulatory circuits for beta cells⁵⁹, such as *NEUROG3*, *FEV*, *TCF7L2*, *MEIS1*, *MEIS2*, *SOX13*, *GLIS3*, *NR3C2*, and *MAFA*. Further gene set enrichment analysis identified active pathways related to epithelial differentiation, extracellular matrix, and cell junction organization (Supplementary Fig. 5a and Supplementary Data 6). In contrast, Beta,0 cells had elevated expression of cell cycle regulators and components of Wnt/planar cell polarity (PCP), TGF β , and synaptic transmission pathways, suggesting a distinct beta cell phenotype compared to Beta,1 cells.

To gain insight into the regulatory networks shaping the gene expression features of the beta cell subpopulations, we used CellOracle⁶⁰ to infer cell-type-specific GRN modules. A custom-assembled base GRN from our pig multiome dataset was applied to construct the GRN configurations in the scRNA-seq data. This resulted in two unique TF networks between Beta,0 and Beta,1 clusters with clearly distinct top 5 regulons (Fig. 4c and Supplementary Data 7). Notably, all these regulons were *NEUROG3* targets identified in pig, whereas in Beta,0 cells, only secondary targets were observed (Fig. 3e). *PLAGL1* is a zinc-finger TF implicated in cell-cycle control, ECM organization, and risk of neonatal diabetes^{43,44,61–63}. In pig Beta,0 cells, *PLAGL1* formed a network linking genes related to cell migration,

adhesion, and cell-cycle regulation (*TMEM176A*, *MFAP4*, *NME2*). In both clusters, *MEIS2* positively targeted beta cell identity genes (*G6PC2*, *PDX1*, *CHGA*), albeit interconnecting with different regulons.

MEIS2 encodes a homeobox TF in the three amino acid loop extension (TALE) family and acts together with PBX and HOX TFs to form dimeric or trimeric complexes to enhance DNA binding specificity and affinity for target gene regulation⁶⁴. *MEIS2* was detected in human embryonic²⁹ and adult beta cells⁶⁵ and shown to regulate *PAX6* expression during pancreas development⁶⁶. To understand the role of *MEIS2* in beta cell differentiation, we performed in silico perturbation using CellOracle to mimic a *MEIS2* knockout (KO, Fig. 4d). In the *MEIS2* KO simulation, Beta,1 cell differentiation was blocked, with only minimal effects on the Beta,0 cluster. This was further validated by Markov simulation to estimate cell distribution changes, showing that *MEIS2* KO reverted Beta,1 cells to the FEV and NGN3 progenitor states.

Next, we assessed whether *MEIS2* could serve as a pig-human conserved marker to distinguish beta cell subpopulations. In the pig scRNA-seq atlas, *MEIS2*⁺ beta cells mostly appeared in the Beta,1 cluster, while *MEIS2*⁺ beta cells were primarily found in the Beta,0 cluster, both subpopulations sharing characteristics matching their respective beta subclusters (Fig. 5b, c). Remarkably, heterogeneous human beta subpopulations in the scRNA-seq atlas could also be identified based on *MEIS2* expression. Human *MEIS2*⁺ beta cells resembled pig *MEIS2*⁺ beta cells and Beta,1 cluster, showing enrichment of genes involved in

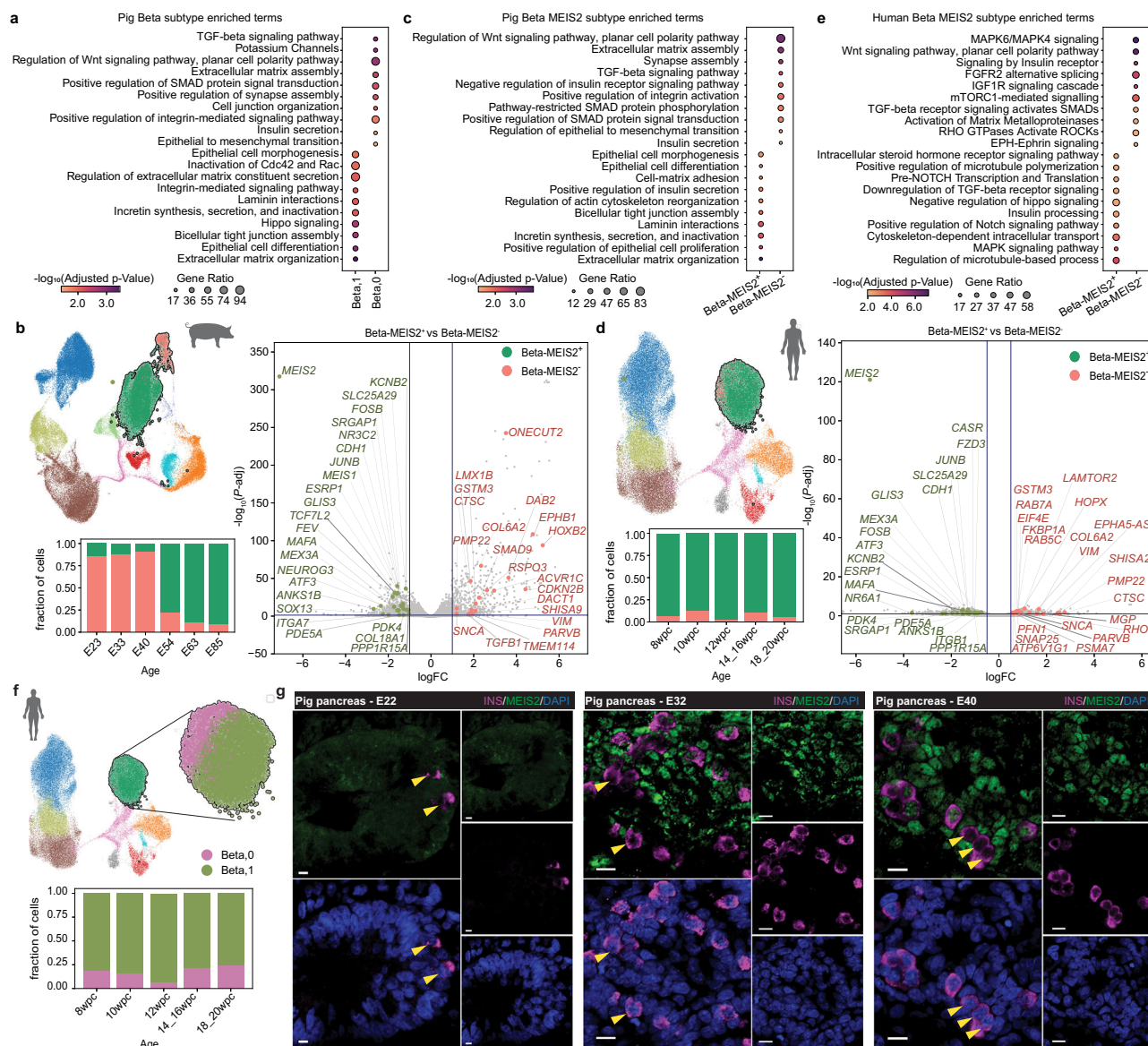


Fig. 5 | Comparison of beta subpopulations in human and pig. **a** Dot plot of enriched pathways identified by Enrichr using the DEGs of pig Beta,0 and Beta,1 clusters. **b** (Left) UMAP showing pig Beta clusters separated according to *MEIS2* mRNA expression with bar plot showing *MEIS2*⁺ and *MEIS2*⁻ beta cell distribution at each sampling age. (Right) volcano plot of differentially expressed genes between pig *MEIS2*⁺ and *MEIS2*⁻ beta cells. **c** Dot plot of enriched pathways identified by Enrichr using the DEGs of pig *MEIS2*⁺ and *MEIS2*⁻ beta cells. **d** (Left) UMAP showing human Beta clusters separated according to *MEIS2* mRNA expression with bar plot showing *MEIS2*⁺ and *MEIS2*⁻ beta cell distribution at the corresponding developmental stage. (Right) volcano plot of differentially expressed genes between human

MEIS2⁺ and *MEIS2*⁻ beta cells. **e** Dot plot of enriched pathways identified by Enrichr using the DEGs of human *MEIS2*⁺ and *MEIS2*⁻ beta cells. **f** UMAP showing human Beta clusters separated according to Louvain clustering, with a bar plot showing two beta subcluster distributions at the corresponding developmental stage.

g Immunofluorescence identification of *MEIS2* positive and negative beta cells in wild-type pig pancreas sections from E22, 32, and 40, scale bar 10 μ m. Images are representative of 3 pig pancreas samples per time point. All p-values from Enrichr and edgeR analyses are adjusted by Benjamini-Hochberg FDR method (Supplementary Data 6). **a–c:** scRNA-seq of pancreatic cells from wild-type and *INS*-eGFP pigs. Detailed sample information is provided in Supplementary Data 1.

microtubule and cell-matrix organization. Conversely, human *MEIS2* beta cells had active Wnt/PCP and TGF β signaling that were observed in pig *MEIS2* beta cells and Beta,0 cluster (Fig. 5d, e). Louvain clustering further identified a distinct beta subcluster composed of the majority of human *MEIS2* cells (Fig. 5f). The presence of *MEIS2*-positive and negative beta cells was confirmed in both pig pancreases (E20–54) and hiPSC-derived islets (Figs. 4e and 5g).

Embryonic endocrine progenitor heterogeneity

Having observed distinct beta cell subpopulations emerging during development, we sought to investigate whether these cells may

originate from different endocrine progenitors. We first examined the CellRank-inferred developmental trajectories that faithfully mapped cell fate transitions in the pig scRNA-seq atlas. The NGN3 cluster was predicted to generate major beta cell clusters as expected. Additionally, the PEC cluster was identified as another intermediate state upstream of a beta cell subpopulation (Fig. 6a). To compare the gene expression programs along the trajectories from NGN3 or PEC towards the respective beta cell subtypes, we applied tradeSeq⁶⁷ to detect differential gene expression patterns between lineages (Fig. 6b and Supplementary Data 8). Two distinguishable gene clusters were identified comparing lineages NGN3-to-Beta,1 and PEC-to-Beta,1. Gene

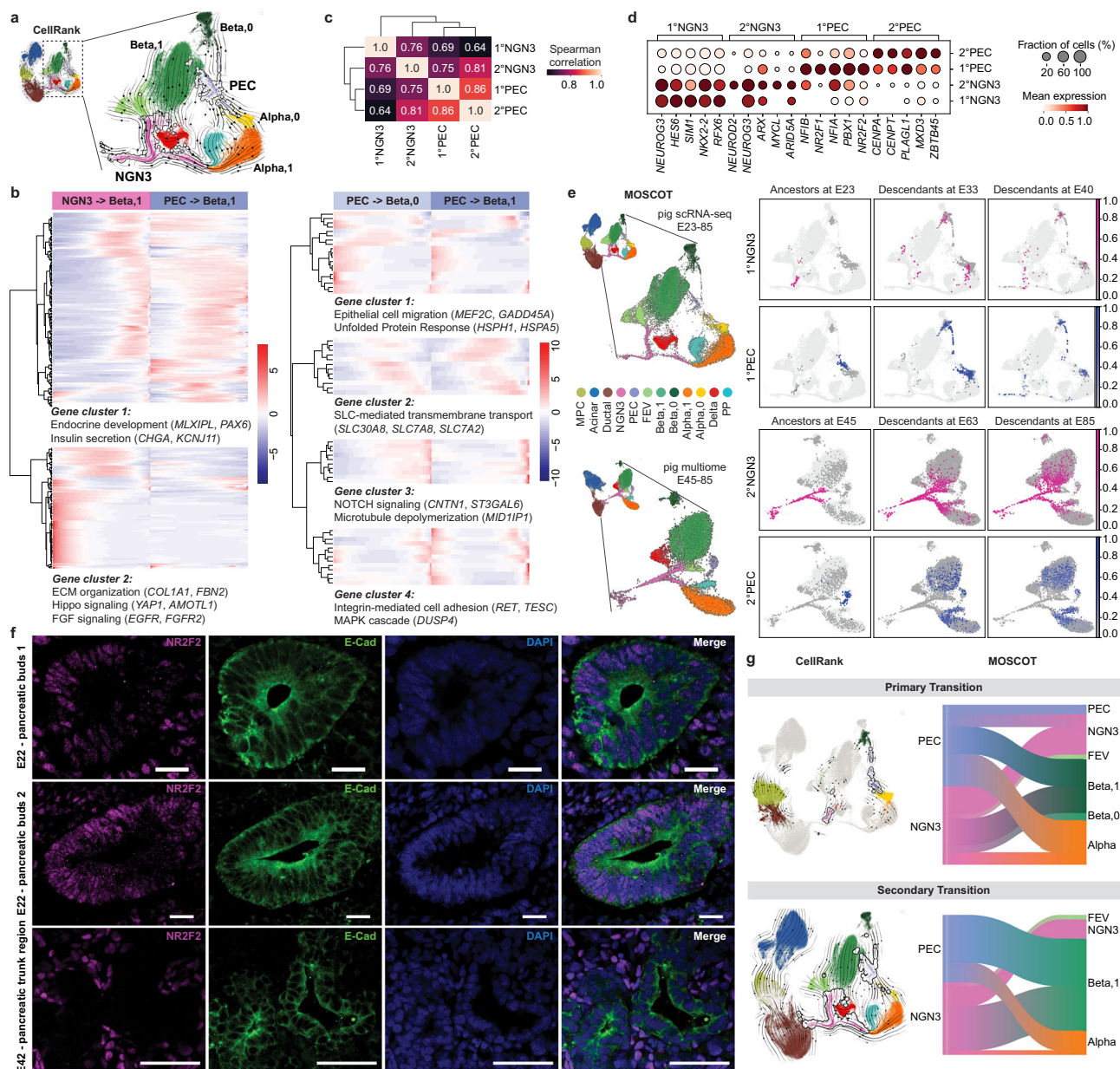


Fig. 6 | Existence of subtypes of endocrine progenitors during development. **a** UMAP showing CellRank-inferred trajectories of NGN3 and PEC as intermediate states upstream of endocrine cells. **b** Heatmaps showing modeled gene expression patterns along pseudotime for differentially expressed gene groups identified by tradeSeq analysis (Benjamini-Hochberg FDR-adjusted p -value < 0.05 ; Supplementary Data 8). Comparisons show transcriptional trajectories from NGN3 cluster to beta cell subclusters (left) versus PEC cluster to beta cell subclusters (right). Annotations indicate enriched pathways and signature genes for each gene group. **c** Cluster correlation analysis using Spearman correlation of mean counts per cluster for each highly variable gene ($n = 4000$). **d** Dot plot of the top 5 differentially expressed TFs of cells in 1° transition (E23-33) and 2° transition (E40-85) NGN3 and PEC clusters (edgeR, FDR-corrected p -value < 0.05 ; Supplementary Data 9). Dot size is relative to the fraction of cells within a cluster expressing the gene. **e** MOSCOT-inferred descendants of NGN3 and PEC cluster from age E23 (using scRNA-seq data, top) and from age E45 (using multiome data, bottom). The left

panel shows the respective UMAP of the endocrine cells from the scRNA-seq (top) and multiome (bottom) atlas used for the MOSCOT calculation.

f Immunofluorescence identification of NR2F2⁺ cells in E22 pancreatic epithelium labeled by E-Cad (top and middle panels). NR2F2 protein is not detected in E42 pancreatic ductal domain (bottom panel). Scale bar 50 μ m. Wild-type pig samples are used for this figure. Images are representative of 3 pig pancreas samples per time point. **g** CellRank and MOSCOT (sankey diagrams showing the summarized MOSCOT-inferred ancestors and descendants) prediction of NGN3 and PEC clusters as intermediated states in the 1° transition (E23-33) and 2° transition (E40-85). (Left) UMAPs overlaid with CellRank-inferred trajectories (arrows); (Right) Sankey diagrams showing the summarized MOSCOT-inferred ancestors and descendants in (e, a-d, e (top), g): scRNA-seq of pancreatic cells from wild-type and *INS*-eGFP pigs. e (bottom) and g: Multiome analysis of pancreatic cells from *PTF1A*-codon-improved-Cre/*Rosa*-mTmG pigs. Detailed sample information is provided in Supplementary Data 1.

cluster 1 was enriched with genes regulating endocrine development (*MLXIPL*, *PAX6*) and insulin secretion (*CHGA*, *KCNJ11*). It showed overall restricted expression in lineage NGN3-to-Beta,1 as opposed to the extended expression in lineage PEC-to-Beta,1. Gene cluster 2 included genes related with ECM organization, Hippo, and FGF signaling, which

had limited expression in lineage PEC-to-Beta,1. We further compared the two lineages from PEC cluster to either Beta,0 or Beta,1. The identified gene groups comprised known factors involved in beta cell differentiation, such as epithelial cell differentiation, insulin secretion, NOTCH signaling, and integrin-mediated cell adhesion. Surprisingly,

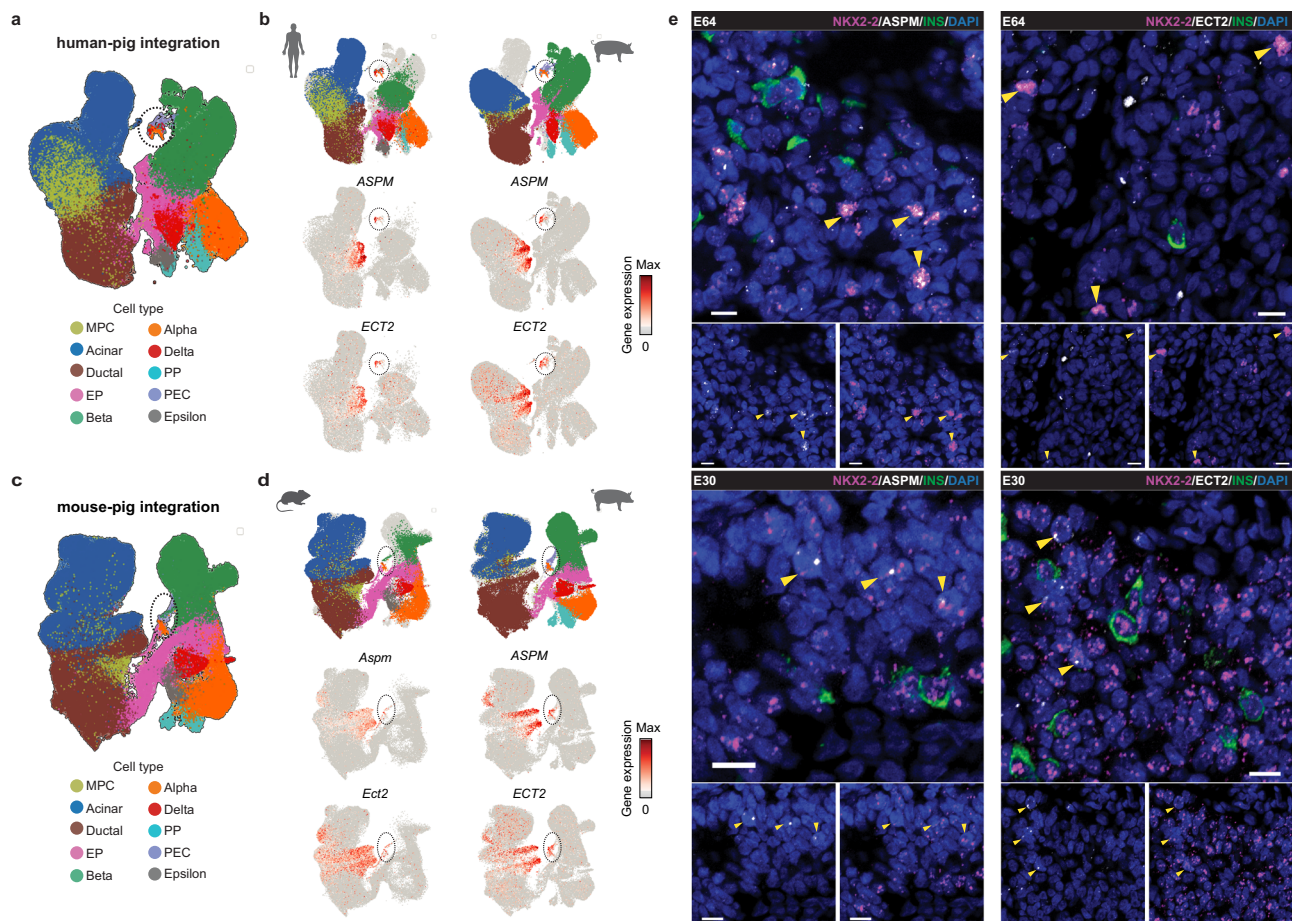


Fig. 7 | PEC cell confirmation by cross-species integration and RNAscope analysis. **a** UMAP of sysVI-integrated human and pig scRNA-seq pancreas atlases. The black circle indicates cells co-clustering with the pig PEC cluster. **b** UMAPs of split views of human (left) and pig (right) data overlaid with the sysVI-integrated embedding, and conserved PEC marker gene (*ASPM* and *ECT2*) expression in each species. **c** UMAP of sysVI-integrated mouse and pig scRNA-seq pancreas atlases. The black circle indicates cells co-clustering with the pig PEC cluster. **d** UMAPs of split views of mouse (left) and pig (right) data overlaid with the sysVI-integrated

embedding, and conserved PEC marker gene (*ASPM* and *ECT2*) expression in each species. **e** Detection of *NKX2-2*⁺ cells co-expressing *ASPM* (left) or *ECT2* (right) in wild-type pig pancreas samples (E30 and E64) by RNAscope assay. Yellow arrows indicate double-positive cells. Scale bar 10 μ m. Images are representative of 3 pig pancreas samples per time point. **a–d**: scRNA-seq of pancreatic cells from wild-type and *INS-eGFP* pigs. Detailed sample information is provided in Supplementary Data 1.

these genes showed subtle differences in the gene expression timing and duration, yet without discrete patterns along the two trajectories.

Despite slight deviations from the NGN3 trajectory, PEC-to-Beta gene expression programs favored beta cell differentiation. Hypothesizing that the PEC cluster contains endocrine progenitors, we compared the transcriptional features between PEC and NGN3 clusters. Given the apparent differences in spatial patterning and cell organization of the pancreas during development, we performed correlation analysis of the PEC and NGN3 clusters divided by 1^o transition (E23–33) and 2^o transition (E40–85), namely following clusters: 1^oNGN3, 1^oPEC, 2^oNGN3 and 2^oPEC. Pairwise correlation scores (0.64–1) revealed substantial transcriptional similarity among all four subgroups (Fig. 6c). However, differential expression analysis uncovered distinct patterns in key transcription factors of each cluster (Fig. 6d, Supplementary Fig. 7a, b and Supplementary Data 9). While 1^oPEC and 2^oPEC showed only marginal expression of canonical endocrine regulators (*NEUROG3*, *NKX2-2*, *NEUROD2*), they were enriched for endocrine-related factors (*NFIA*⁶⁸, *PBX1*⁶⁹ in 1^oPEC; *PLAGL1*⁴⁴ in 2^oPEC), suggesting a partially active endocrine program. Notably, we identified *NR2F2* as a key transcription factor highly expressed in the 1^oPEC cluster, though it was also detected in broader early progenitors and mesenchymal cells (Fig. 6f). *NR2F2* (also known as *COUP-TFII*) encodes a member of the steroid/thyroid hormone superfamily of nuclear receptors. It has been

detected in *PDX1*⁺ cells at E11.5 in mice and may play a role in regulating beta cell mass^{70,71}. We further confirmed 1^oPEC presence in the pancreatic primordium by detecting *NR2F2*⁺ cells within the multilayered epithelium of E23 pancreatic buds (Fig. 6f).

To verify the differentiation trajectories linking PEC cluster to beta cell subpopulations, we applied our recently developed Multi-Omics Single-Cell Optimal Transport (MOSCOT)⁷² framework, which uses optimal transport theory to reconstruct developmental trajectories across real developmental time points and multiple modalities. Applying MOSCOT to both pig scRNA-seq and multiome time-series atlases recovered the differentiation lineages identified by CellRank, showing that, in addition to NGN3, PEC at both early and late stage was coupled to a subset of endocrine cells as its descendants (Fig. 6e, g).

To determine whether PEC-like cells exist during human and mouse pancreas development, we performed cross-species integration of pig-human and pig-mouse scRNA-seq datasets using sysVI⁷³. This machine learning pipeline enables cross-species integration of scRNA-seq datasets while retaining high biological preservation. The sysVI-integrated embedding confirmed correct cell type alignment (Fig. 7a–d), as supported by established pancreas marker gene expression (Supplementary Fig. 7c, d). In both human and mouse datasets, we identified cell populations (initially annotated as diverse endocrine cell types) that co-clustered with pig PEC cells, indicating

transcriptionally shared cellular states (Fig. 7a, c). Notably, mouse PEC-like cells showed connectivity to a subset of cells from the mouse NGN3 cluster, which was not observed in human and pig. We further identified *ASPM* and *ECT2* as cross-species markers, expressed in pig PEC cells and their counterparts in both human and mouse (Fig. 7b, d and Supplementary Data 14). *ASPM* encodes protein abnormal spindle-like microcephaly-associated, which was identified as a novel Wnt and stemness regulator⁷⁴. *ECT2* encodes protein epithelial cell transforming 2, a guanine nucleotide exchange factor for Rho-like GTPases⁷⁵. While these markers were also detected in some ductal and acinar cells, PECs showed a clear endocrine identity as evidenced by their strong transcriptional similarity to the NGN3 cluster (Fig. 6c). To validate this population, we selected *NKX2-2*, a pan-endocrine marker, in combination with *ASPM* or *ECT2* to distinguish PECs from ductal and acinar cells. Using RNAscope, we identified *NKX2-2*⁺ cells co-expressing *ASPM* or *ECT2*, confirming their presence in ductal and acinar regions of the pancreas across pig developmental stages (Fig. 7e).

Discussion

Through cross-species comparative multiomics integrating transcriptomic and chromatin accessibility profiling of pancreas development in mice, pigs, and humans, our work reveals both species-divergent and evolutionarily conserved gene regulatory mechanisms governing pancreatic lineage differentiation. First, the resemblance in developmental tempo of pig and human gestation provides a temporally aligned framework to study extended pancreas organogenesis events that are compressed in the mouse. Second, we observe a pig-human conservation in epigenetic and transcriptional regulation, particularly in the endocrine lineage. The high conservation of transcription factors downstream of *NEUROG3* (over 50% between pig and human) suggests a core program for endocrine fate acquisition in larger mammals. This aligns with both the recent human study³⁰ and the observed similarities in postnatal islet characteristics, fasting C-peptide and glucose levels between pigs and humans^{18,21}. By leveraging the temporally resolved single-cell multiomic pig pancreas atlas covering all three trimesters, we identified a unique primed endocrine cell (PEC) population, potentially representing a progenitor state that is distinct from the classic NGN3 endocrine progenitor cluster. Transcriptionally matched PEC-like cells were also identified in human and mouse, with murine PEC-like cells showing a possible link to the NGN3 cluster. Concurrent with pancreas morphogenesis, we discovered that both NGN3 and PEC clusters emerged as heterogeneous populations and were predicted to hold dynamic lineage potential over time. This conserved PEC population is intriguing as it may suggest a potential *NEUROG3*-independent pathway for endocrinogenesis, which could offer an explanation for the persistence of endocrine cells in some human patients carrying homozygous *NEUROG3* mutations^{76,77}. However, definitive validation of this pathway and the differentiation capacity of PECs remains a crucial goal for future work. We further identified heterogeneous beta cells, suggesting the acquisition of islet cell heterogeneity may originate in embryonic development. In summary, our study presents the pig as a valuable complementary model to existing systems and enhances the understanding of in vivo pancreas development through a comprehensive multimodal comparison across species. These resources can be harnessed to refine stem cell and organoid models, offering unique opportunities to 1) address open questions in human biology and disease, and 2) bridge translational gaps from rodents to large animals (pigs) and ultimately to humans.

Methods

Pig pancreas collection

Pigs were housed at the designated pathogen-free pig facilities in LMU or TUM. All animals received standard diet and water *ad libitum* as well as standard vaccination. After hormonal estrus cycle synchronization,

pigs in heat were artificially inseminated or mated and ultrasonic confirmation of pregnancy was performed 21 days post insemination. At selected gestation stages, pregnant pigs were euthanized, and fetuses were collected. Pancreases were extracted and used for further analysis.

Immunofluorescence analysis

Pig pancreases or hiPSC-derived islets were fixed in pre-chilled 4% paraformaldehyde overnight (pancreas) or for 30 min (islets) then dehydrated in a progressive sucrose gradient at 4 °C. Samples were embedded in tissue freezing medium (Leica Biosystems), sectioned at 12 µm, and mounted onto Superfrost® Plus slides (Thermo Fisher Scientific) and stored at -80 °C.

Immunostaining. Sections were washed with PBS for 30 min, permeabilized in 0.1% Triton X-100 (Sigma-Aldrich) in 0.1M Glycine (Sigma-Aldrich) for 30 min, blocked in blocking solution (PBS + 0.1% Tween-20 + 10% FCS + 0.1% BSA + 3% donkey serum) for 1 h at room temperature (RT), and incubated with primary antibodies diluted in blocking solution (Supplementary Data 10) overnight at 4 °C. Sections were washed in PBS for 15 min twice, incubated with secondary antibodies diluted in blocking solution (Supplementary Data 10) for 3 h at 4 °C, counterstained with DAPI for 20 min at RT, and mounted with the ProLong™ Diamond Antifade Mountant (Thermo Fisher Scientific).

Integrated mRNA and protein co-detection assay. Pig pancreas sections were used for mRNA and protein co-detection assay according to the RNAscope® Multiplex Fluorescent v2 Assay combined with Immunofluorescence - Integrated Co-Detection Workflow (ACD, Bio-technique). Briefly, sections were washed in PBS, baked at 60 °C, and dehydrated in a progressive ethanol gradient. Sections were treated with hydrogen peroxide and 1x Co-Detection Target Retrieval solution and incubated with primary antibody (Supplementary Data 10) diluted in Co-Detection Antibody Diluent overnight at 4 °C. Sections were fixed in 10% neutral buffered formalin for 30 min at RT and treated with RNAscope® Protease III. Sections were then hybridized with probes against pig *NEUROG3* (ACD, Cat#498781), *NKX2-2* (Cat#1570601-C1), *ASPM* (Cat#1734071-C3), and *ECT2* (Cat#1734091-C3) for 2 h at 40 °C and signals developed following the RNAscope® Multiplex Fluorescent Reagent Kit v2 Manual. Sections were incubated with secondary antibodies for 30 min at RT, counterstained with DAPI for 20 min at RT, and mounted with the ProLong™ Diamond Antifade Mountant (ThermoFisher). All slides were stored at 4 °C until imaging by confocal microscope (ZEISS LSM 880 with Airyscan). Images were acquired and processed using Zeiss Zen 2.3 lite (Blue edition, ZEISS).

Improved genome annotation for *Sus scrofa*

Pig pancreas scRNA-seq data was previously aligned using Ensembl *Sus scrofa* gene annotation (version 94), however showed incomplete gene annotation missing certain genes of interest (e.g. *MAFA*, *FEV*, *PTF1A*). Our analysis found reads downstream of the gene bodies which were not covered by the annotation. Beiki et al.⁸² generated an improved pig genome annotation (iso-seq annotation) by integrating poly(A) selected single-molecule long-read isoform sequencing (Iso-seq) and Illumina (short read) RNA sequencing (RNA-seq). Therefore, we further improved the annotation by combining iso-seq annotation with the Ensembl annotation version 101. We annotated the transcripts of iso-seq annotation using gffcompare (v0.12.1)⁸³ with referencing the annotation from Ensembl version 101. The Ensembl annotation was filtered by removing “pseudogene” and “processed_pseudogene” biotypes genes. We added the associated same strand transcripts from the iso-seq annotation to the Ensembl annotation and extended the gene bodies if the added transcripts required it. If a transcript was added to a gene, the gene ID was altered by adding the suffix -iso. For each gene, we then added an additional “extension gene” downstream of the gene

body. The length of the extension gene was defined by the maximum possible region of 1–10 kb (using 1 kb steps) that does not intersect with any other same strand gene. The extension genes were named by the name of the original genes and adding the suffix -ext'x'kb, whereby 'x' corresponds to the integer length of the addition from 1–10. Final gene counts were obtained by summing the -iso and -ext'x'kb (if available) versions of each gene. Supplementary Data 11 shows UMI counts for 37 genes of interest before ("ensembl") and after ("iso") gene extension as examples.

Single-cell sequencing

Biological replication. For next-generation sequencing experiments, a minimum of three biological replicates (individual embryos/fetuses per pregnancy) at each developmental stage were sampled where possible. Modified approaches were used for two exceptions: At E40, only two embryos were available for analysis; At E22/23 and E33, pancreatic tissues were extremely small in size. To obtain sufficient cell numbers for reliable single-cell sequencing, all available pancreases at each age were combined for processing, as summarized in Supplementary Data 1 and Supplementary Fig. 2a, c.

Single cell suspension preparation. Freshly dissected pancreases were minced into fine pieces and digested with collagenase V in HBSS with Ca/Mg (0.5 mg/mL for E20–50, 1 mg/mL for E60–85, Sigma-Aldrich) for 5 min followed by dissociation with TrypLE™ Express (Gibco) for 10–15 min at 37 °C. Cell suspension was filtered through a 40 µm cell strainer.

Pancreatic epithelial cell enrichment. This procedure was not applied to E22/23 and E33 samples due to their small size; all cells from these samples were used for scRNA-seq without enrichment. Pancreases from transgenic reporter pigs and wild-type littermates at E40–85 were used to collect pancreatic epithelial cells. For enrichment, cell suspension was stained with EpCAM-PE-Cy7 antibody (1:200, Invitrogen) in 1% BSA + PBS for 30 min at 4 °C. Cells were then washed and filtered into a tube through a 35 µm filter. 7-AAD (Invitrogen) or Sytox blue (Invitrogen) was used to distinguish dead cells. The resulting cell suspension was loaded onto a FACSAria III (BD) for sorting. The transgenic reporter pigs were used for specific cell-type enrichment: 1) *INS-eGFP* pigs⁷⁸: beta cells were enriched via insulin promoter-driven GFP expression. 2) *PTF1A*-codon-improved-Cre⁷⁹ x *ROSA-mTmG*⁸⁰ pigs: *PTF1A*-codon-improved-Cre/*ROSA-mTmG* embryos were identified with epifluorescent microscope by GFP signal. The pancreases from these embryos were used to enrich pancreatic epithelial cells, particularly acinar cell populations, with GFP signal. Supplementary Data 1 provided an overview of all samples obtained through different enrichment strategies. Supplementary Fig. 2c provided a summary of pancreas number and cell counts of identified cell types in each sample. Supplementary Fig. 8 provided representative gating strategies.

scRNA. Single-cell suspensions were processed for scRNA-seq with a targeted cell recovery of 10,000. 10X Genomics' Single Cell Gene Expression protocols were followed according to the manufacturer's specifications and guidelines. Libraries were pooled and sequenced by a HiSeq4000 or NovaSeq 6000 platform following the recommendations from 10X Genomics. With Cell Ranger pipeline (v3.1.0), samples were demultiplexed to produce a pair of FASTQ files for each sample. Reads containing sequence information were aligned to the improved pig genome annotation and pre-processed for downstream analyses.

Multiome (scRNA/ATAC). For nuclei isolation and library construction, a low-input nuclei isolation protocol adapted from 10X Genomics was performed. In brief, sorted cells were washed once with 1 mL PBS + 1% BSA, counted, centrifuged, and the supernatant was aspirated. Subsequently, the washed cell pellet was resuspended in chilled

lysis buffer with 0.5x detergent concentration (50 µL per sample) and placed on ice for 5 min. Then wash buffer (500 µL per sample) was added and nuclei were centrifuged. To gradually change from wash to diluted nuclei buffer, cells were washed once in a 1:1 mixture of wash buffer and diluted nuclei buffer and subsequently one with pure diluted nuclei buffer. The washed isolated nuclei were then resuspended in 7–10 µL diluted nuclei buffer and were directly added to the transposition reaction after quality control and counting. In all following steps, 10X Genomics' Single Cell Multiome ATAC and gene-expression protocols were followed according to the manufacturer's specifications and guidelines. The final libraries were sequenced on the Illumina NovaSeq 6000 platform following the recommendations from 10X Genomics. Raw reads were aligned to the improved pig genome annotation and pre-processed using the 10X Genomics Cell Ranger ARC pipeline (v 2.0.0) for downstream analyses.

Single-cell data analysis

Preprocessing of 10X scRNA-seq raw data. Pig pancreas samples (Supplementary Data 1) and published datasets of human and mouse^{30,35–41} were preprocessed using Scanpy⁸⁴ (v1.8.2). Filtering of low-quality cells Each sample was assessed using Scanpy's quality control measures and sample-specific minimum number of genes per cell, minimum number of counts per cell, and maximum number of counts per cell were set to filter out low-quality cells (Supplementary Data 12). In addition, all cells with a mitochondrial fraction > 0.15 were excluded, as well as all genes that were expressed in less than 20 cells. Read counts and gene counts across clusters of pig scRNA-seq data were shown in Supplementary Data 13. The filtered gene matrices from Goncalves et al.³⁷ were not filtered. Normalization Gene counts were normalized using Scanpy⁸⁵ (v1.22.1) for data from each lab separately. For this, we first performed a total counts normalization of each cell counts to 1,000,000, then performed a log transformation using natural log and pseudocount 1, further calculated a neighborhood graph using the first 50 principal components and number of neighbors k = 15. Clusters were obtained using louvain clustering⁸⁶ with resolution r = 0.5. We then used Scanpy to estimate size factors with the louvain clusters as input clusters and minimum mean average count of genes to be used for normalization set to 0.1. The size factors were then used for normalization of raw gene counts (summed -iso and -ext'x'kb gene counts for pig samples). For downstream integration, log-transformed counts using natural log and pseudocount 1 were calculated. Ambient gene calculation Ambient genes were estimated based on expression in empty droplets using DropletUtils⁸⁷ (v1.14.2). Genes with an ambient expression score larger than 0.005 were considered ambiently expressed genes. Ambient genes were generalized to datasets where raw data was not available. Highly variable gene calculation Highly variable genes (HVGs) were calculated per batch⁸⁸ to select HVGs unaffected by batch variance. Per-batch HVGs were obtained with Scanpy, using the Cell Ranger⁸⁹ flavor, and were ranked first by the number of batches in which the genes were highly variable and second by the mean dispersion across batches. Finally, the top genes in this ranking were selected as highly variable genes. Integration Pig samples were integrated with Scanorama⁹⁰ (v1.7.1) using 8,000 HVGs (excluding ambient genes), resulting in a 100-dimensional latent embedding. Human and mouse samples were each integrated with scVI⁹¹ (v0.16.1) using 2,000 HVGs, resulting in a 20-dimensional latent embedding. Clustering and annotation Clustering was performed on the k-nearest neighbor (KNN) graph (k = 15) calculated from the integrated embedding using louvain clustering⁸⁶ (leiden⁹² for human and mouse data) with resolution 1. The pig integrated embedding was reduced to 50 principal components before calculating the neighborhood graph. Clusters were annotated using differentially expressed marker genes. For some analysis several clusters (NGN3, Beta) were subclustered using a higher clustering resolution. Re-normalization after integration Scanpy normalization might be based on batch-specific clusters when a strong

batch effect is present, leading to non-comparable counts across samples. We ran a check on all species datasets by visualizing the mean number of counts per cluster (based on integrated embedding) and sample before and after normalization. For human and mouse samples we observed strong sample-specific counts after normalization. To obtain comparable counts across samples, we re-normalized these samples using Scran and the clusters obtained from the integrated embedding. Doublet detection To obtain robust doublet estimates, we used a combination of scrublet⁹³ (v0.2.3), DoubletDetection⁹⁴ (v4.2), scds⁹⁵ (v1.10.0), scDblFinder⁹⁶ (v1.11.4), DoubletFinder⁹⁷ (v2.0.3) (default parameters, expected doublet rate 0.8) to detect doublets. Cells consistently detected by three or more methods as doublets were excluded from further analysis. In addition, clusters with a doublet frequency larger than 70% were entirely excluded.

Reanalysis of human fetal scATAC-seq dataset. The 12wpc human fetal scATAC-seq data³⁰ were read into Signac⁹⁸ for preprocessing. Peaks from standard chromosomes and additionally called using MACS2 were used. Gene annotation from EnsDb (EnsDb.Hsapiens.v86) was added. Quality control metrics were computed to filter low quality cells (Supplementary Data 12). Data was normalized by term frequency-inverse document frequency (TF-IDF) normalization. Dimension reduction was done by running singular value decomposition (SVD) on the TF-IDF matrix, using the peaks selected by the function FindTopFeatures. Graph-based clustering and non-linear dimension reduction for visualization was performed on the KNN graph ($k = 30$) calculated from the low-dimensional embedding using SLM⁸⁶ algorithm. Gene activity matrix was created by the GeneActivity function. Doublets were called by scDblFinder⁹⁶ and excluded for further analysis. For cell cluster label transfer, we first extracted the 12wpc data subset from the integrated human scRNA-seq data. The scATAC-seq data gene activity was used as an approximation of a gene expression matrix and was integrated with the 12wpc scRNA-seq data following the standard scANVI⁹⁹ (v0.20.3) workflow to enable label transfer. Cicero was used to compute pairwise co-accessibility scores for each peak, which were further grouped into cis-co-accessible networks. The co-accessible links along with DNA accessibility information were visualized by CoveragePlot.

Integration of human fetal scRNA-seq and scATAC-seq data. The unmatched modalities were integrated using GLUE⁴⁸ v0.3.2. The RNA modality input, i.e. the 12wpc data subset was extracted from the integrated human scRNA-seq data. The ATAC modality input was processed as described in section “*Reanalysis of human fetal scATAC-seq dataset*”. We then constructed a guidance graph that contains omics features as nodes (i.e., genes for scRNA-seq, and peaks for scATAC-seq) and prior regulatory interactions as edges. We used the default implementation that links an ATAC peak to a gene if it overlaps either the gene body or promoter region. This graph was utilized by GLUE to orient the multi-omics alignment. To match cells from both modalities, we performed minimum cost maximum flow bipartite matching on the joint embedding derived from GLUE^{47,100}. The cost graph was inferred using `get_cost_knn_graph` with `knn_k=15`, `null_cost_percentile = 99` and `capacity_method = 'uniform'`. Using the bipartite matches, we matched each ATAC cell to an RNA cell. In cases where no ATAC match was found for an RNA cell, only RNA information was used. The latent vector of the cell was calculated as the average latent vector of the matched cells. Gene activities were further denoised with MAGIC¹⁰¹ by smoothing over nearby cells in the joint embedding as proposed and benchmarked in ArchR¹⁰². The Python implementation of magic (v3.0.0) was used to smooth gene activities over the k-nearest neighbors graph of the joint embedding with $k = 15$ neighbors, `decay = 1` and k-nearest neighbors autotune parameter `ka = 4`. The integrated and imputed dataset was used for gene regulatory network construction.

Preprocessing of pig 10X multiome raw data. Multiome data was preprocessed similarly to scRNA-seq data as described above using Scanpy (v1.9.1) and Muon¹⁰³ (v0.1.2). After summing the `-iso` and `-ext'x'kb` (if available) counts of each gene to generate final counts, DropletUtils⁸⁷ (v1.14.2) was used with default parameters to estimate ambient gene expression probabilities.

Filtering of low-quality cells. Each sample was assessed using standard quality control measures and sample-specific maximum mitochondrial gene fraction, minimum number of genes per cell, minimum number of counts per cell, and maximum number of counts per cell were set to filter out low quality cells (Supplementary Data 12). To further filter out cells with low ATAC-seq quality, Muon was used to calculate ATAC-specific quality metrics. Sample-specific thresholds were identified for minimum and maximum number of counts, minimum and maximum TSS enrichment score as well as minimum and maximum nucleosome signal (Supplementary Data 12).

Doublet detection. Similar to scRNA-seq, we used a combination of scrublet⁹³ (v0.2.3), DoubletDetection⁹⁴ (v4.2), scds⁹⁵ (v1.10.0), scDblFinder⁹⁶ (v1.11.4), DoubletFinder⁹⁷ (v2.0.3) (default parameters, expected doublet rate 0.8) and SOLO¹⁰⁴ (as implemented in scvi-tools v0.19.0) to detect doublets based on the gene expression modality. In addition, scDblFinder and its implementation of AMULET¹⁰⁵ were used to identify doublets on the ATAC-seq modality. Cells consistently detected by three or more methods as doublets were excluded from further analysis.

Generation and quantification of common peak set. To merge the ATAC-seq data from individual samples, we followed the respective vignette on the Signac⁹⁷ website. In brief, peaks from all samples were merged using the “reduce” function of the GenomicRanges (v1.46.1) package and only peaks on standard chromosomes were kept. Next, for each sample, fragment counts were determined using Signac and stored, together with gene expression data in a Seurat object, which were subsequently merged into a single object.

Normalization of ATAC-seq counts. Signac was used to run TF-IDF normalization on ATAC-seq counts with default parameters. TF-IDF normalized count matrix was then imported into Muon.

Normalization of gene expression counts. Prior to normalization, data from individual samples was merged. SCTransform (v0.3.3) was used for normalization using settings `vst.flavor = "v2"` and `clip.range = c(-sqrt(n), sqrt(n))`, where n ($n = 33898$) represented the number of cells.

Highly variable genes. The top 4000 highly variable genes were identified using the `devianceFeatureSelection` function from the scry package¹⁰⁶ (v1.6.0) with default parameters.

Label transfer from scRNA-seq. To ensure consistent cell type labels between scRNA-seq and Multiome data, we employed a k-nearest neighbors classifier to transfer cell type annotations from the scRNA-seq reference. We first trained an scVI model (scvi-tools v0.20.0), with parameters `n_hidden=1024`, `n_latent=50`, `n_layers=2`, `gene_likelihood = 'nb'`, `dispersion = 'gene-batch'`, sample names as batch key, and the technology (i.e. scRNA-seq and Multiome) as categorical covariate, to generate a shared latent space for the scRNA-seq reference and the Multiome gene expression data. Next, we used the k-nearest neighbors classifier ($k = 5$), as implemented in scikit learn (v0.24.2), to predict cell type labels in the shared latent space.

Integration. To integrate the gene expression modality of the different Multiome samples, we used harmony¹⁰⁷ (v0.0.9) with sample names

as batch key. To integrate the chromatin accessibility modality, we used PoissonVI¹⁰⁸, with parameters $n_hidden=1024$, $n_latent=50$, $n_layers=3$, sample names as batch key, transferred labels as labels key and the developmental stage as categorical covariate.

Clustering and annotation. Clustering was performed on the k-nearest neighbor (KNN) graph ($k=21$, metric = ‘minkowski’) calculated from the harmonypy integrated embedding using leiden⁹² clustering with resolution 2. To further separate subtypes of endocrine progenitors, the respective clusters were subclustered with a resolution of 1. The resulting clusters were then annotated to match the transferred cell type labels.

Differential gene expression analysis using edgeR. We calculated differentially expressed genes between cell clusters by pseudo-bulking samples with edgeR¹⁰⁹ (v4.0.16). We calculated pseudo-bulk expression by summing normalized raw counts of cells from one sample and cell type. Pseudo-bulks with fewer than 30 cells were excluded. To ensure that ambiently expressed genes were not erroneously predicted as differentially expressed genes, we balanced the pseudo-bulks on the sample level. For differential testing, we modeled gene expression using a generalized linear model¹¹⁰ with cell type and sample as covariates. Significantly differentially expressed genes were identified using a likelihood-ratio test for the coefficients of interest ($q < 0.05$, corrected for multiple testing with the Benjamini-Hochberg method¹¹¹ at $\alpha=0.05$). Gene set enrichment analysis was done with Erichr¹¹² to identify pathway enrichment signatures of the respective cell cluster.

Cross-species comparison using cluster correlation. We compared cell-type gene expression profiles across species. For this, we calculated the mean normalized gene count per cluster. Then, we mapped pig and mouse genes to human gene symbols using orthologues from Ensembl BioMart¹¹³. If there were multiple genes mapping to one human gene, we used the summed mean normalized gene counts. For the correlation, we considered only the intersection of the top 4000 highly variable genes from all three species, resulting in 851 genes. We used Spearman's rank correlation¹¹⁴ to compare gene expression values for each cluster across species.

Trajectory inference of scRNA-seq data with CellRank. We estimated developmental trajectories and cell fates using CellRank⁵³ (v1.5.1). For this, we estimated a pseudo-time for every cell using Palantir¹¹⁵ (v1.0.1), with a highest *PDX1*-expressing MPC cell as root cell from E22 for pig, 49dpc for human and e9.75 for mouse, respectively. Using the CellRank pseudo-time kernel, we calculated terminal states for all clusters and endocrine progenitor clusters (NGN3 and FEV clusters in pig and mouse data, EP in human data).

CellRank lineage driver estimation and cross-species comparisons. To compare endocrine development across species, we computed putative lineage drivers by calculating Pearson's correlation of each gene with CellRank fate probabilities for each terminal state. We mapped pig and mouse genes to human gene symbols using orthologues from Ensembl BioMart¹¹³. If there were multiple genes mapping to one human gene, we used the maximum absolute lineage correlation score. We then only considered genes that were present in all three species and had orthologues, resulting in 12,437 genes. To compare lineage correlation scores, we scaled the scores per species by the maximum of the 0.01 and 0.99 quantile and clip values to -1 and 1. Finally, we calculated the lineage driver genes (Benjamini-Hochberg FDR-corrected $p > 0.05$, scaled correlation > 0.7) of this mapped subset for every species and lineage and compared overlaps across species.

Gene module analysis of CellRank lineage drivers. To compare gene expression dynamics during alpha/beta cell development in mouse, pig and human, we first extracted orthologous alpha/beta lineage drivers expressed in at least 20% human cells and performed hierarchical clustering with AgglomerativeClustering from scikit-learn¹¹⁶ to identify gene groups for lineage drivers with positive or negative correlation, respectively. Gene set enrichment analysis was performed with GSEapy enrichr module¹¹² to identify pathway enrichment signatures for each identified gene group. The expression patterns of these gene groups were analyzed in pig and mouse data.

Differential gene expression analysis between lineages by tradeSeq. We used tradeSeq116⁶⁷ (v1.14.0) to calculate differentially expressed genes between lineages, i.e., NGN3-to-Beta,0 versus (vs) PEC-to-Beta,0; PEC-to-Beta,0 vs PEC-to-Beta,1. CellRank computed trajectories for these lineages was extracted and used for tradeSeq downstream analysis. We fitted a negative binomial generalized additive model (NB-GAM) using tradeSeq for each of the top 2000 highly variable genes and each lineage. We then identified genes with significantly different expression patterns between lineages using the PatternTest function. To exclude the genes that were already differentially expressed at the initial or terminal states, we inverted the gene rank from the edgeR differential expression (DE) analyses of NGN3 vs PEC and Beta,0 vs Beta,1. We further scored each gene by combining the inverted DE rank and the rank in the PatternTest results (transientScore). This allowed us to identify genes with similar expression at the initial or terminal state but showed significantly different expression pattern along the lineage ($q < 0.05$, corrected for multiple testing with the Benjamini-Hochberg method¹¹¹ at $\alpha=0.05$). Gene set enrichment analysis was performed with GSEapy enrichr module¹¹² to identify pathway enrichment signatures for each identified gene group.

Cross-species integration by sysVI. To integrate scRNA-seq dataset from different species, we used sysVI⁷³ model that combines machine learning with conditional variational autoencoders (cVAE) for integrating datasets with substantial batch effects while better preserving biological variation. We subset the pig and human datasets to the intersection of top orthologous HVGs (2535 genes) as recommended. The two datasets were concatenated with “species” defined as batch_key covariate and “samples” defined as categorical_covariate_keys covariate. The default implementation of sysVI with multimodal variational mixture of posteriors prior (VampPrior) combined with latent cycle-consistency loss was used for the integration.

Gene regulatory network inference using CellOracle. Using CellOracle⁶⁰ (v0.12.0), we first constructed a base GRN using the scATAC part of the multiome data. The co-accessible peak information was extracted to generate the active gene regulatory element data, which contained the open accessible genomic regions and cis-regulatory connection data. We then annotated transcription start sites (TSS) to generate the active promoter/enhancer genomic region data. These data were integrated and peaks with weak co-accessibility scores removed, resulting in the final pig base GRN. Pig scRNA-seq data was reduced to 25,000 cells with 3028 genes (top 2000 HVGs + all TFs). An Oracle object was built by combining the gene expression counts, clustering information, CellRank trajectory with the base GRN. After KNN imputation, cluster-specific GRN for all clusters was calculated with the get_links function. To remove the weak edges and insignificant edges, we filtered the network edges by keeping the top 2000 edges ranked by edge strength with a p-value < 0.001 before network structure analysis. Network (centrality) scores were calculated using the links.get_network_score function. The top genes ranked by betweenness centrality were selected for Beta,0 and Beta,1 cluster to

plot the GRNs with NetworkX¹¹⁷. To simulate *MEIS2* KO in silico, we first used the Oracle object and the filtered GRNs of all clusters to make the regression models (a regularized linear machine-learning model) for simulation. *MEIS2* expression was set to 0, and the global future gene expression shift after perturbation was then calculated.

Gene regulatory network inference using Pando

Metacells. To overcome the sparsity of single-cell data, we used SEACells¹¹⁸ (v0.3.2) with default parameters to identify metacells ($n = 387$), representing cell states in the integrated gene expression latent space.

Peak to gene linking. To identify putative regulatory elements, we used the Signac (v1.9.0) LinkPeaks function with default parameters to calculate the correlation between chromatin accessibility and gene expression of nearby highly variable genes.

GRN inference. We then used all peaks with significant links to construct a GRN using Pando⁴⁷ (v1.0.3) for both pig multiome data and human integrated scRNA/ATAC-seq data. In brief, we used the motif collection and find_motifs function from the Pando package to identify transcription factor motifs within the peaks. We then inferred the GRN considering only highly variable transcription factors found in the motif collection and peaks within 1×10^6 bp around their TSS, using the following parameters: peak_to_gene_method = 'Signac', aggregate_peaks_col = 'SEACell', tf_cor = 0.05, method = 'xgb'. Next, we constructed transcription factor modules within the GRN using the find_modules function with a p-value threshold of 0.05, a R^2 threshold of 0.15 (0.05 for human data), a minimum number of variables of 50 (10 for human data) and a minimum number of genes per module of 5. The GRN was visualized as a UMAP embedding of the TFs based on co-expression and regulatory relationship as measured by the inferred coefficients. Nodes were sized by the PageRank centrality of each TF and colored according to the enrichment of TF expression. Coverage tracks and peak-to-gene links were visualized using Signac. UCSC lift-Over tool (<http://genome.ucsc.edu>)¹¹⁹ was used to lift link coordinated from pig to human reference genome. ALRA¹²⁰ was used to impute pig gene expression and the calculated values were shown as violin plots next to the coverage plots.

Motif activity analysis with ChromVAR. We computed motif activities for human and pig scATAC-seq data using chromVAR¹²¹ (v1.24.0). We first identified motif matches of the human_pwns_v2 motif collection from the chromVARmotifs package using Signac. Motif class information was derived from CIS-BP Database Build 2.00. Next, we used the RunChromVAR function with default parameters to calculate per-cell motif activity scores. We then tested for differential activity scores between cell types using the FindAllMarkers function with mean.fxn = rowMeans, to compute the average difference in z-scores. Motifs with an adjusted p-value < 0.01 and an average difference > 1 were considered differentially active. To further select meaningful motifs for plotting in the heatmaps in the main figures, we computed the Pearson correlation between transcription factor gene expression z-scores and chromVAR motif z-scores in the pig multiome data and kept only motifs/transcription factors with a correlation coefficient > 0.1 . Heatmaps show the chromVAR z-scores stored in the data slot of the chromVAR assay.

Trajectory inference of scRNA-seq and Multiome data with MOSCOT. To confirm trajectory results obtained using CellRank, we additionally used MOSCOT⁷² (v0.3.3) to infer endocrine lineages using real time points. We ran MOSCOT on the porcine scRNA-seq data using time points E23, 33, and 40 (using geodesic distances, tau_a = 0.999, tau_b = 0.99999). We constructed an approximation of the geodesic distance by using an approximation of the Heat Kernel¹²² and

constructed a KNN-graph with $k = 15$ based on a 50-dimensional PCA embedding. For time points E45, 63, 85 we ran MOSCOT on the multiome data with a joint embedding created by concatenating the scaled harmony embedding of the scRNA-seq data and scaled atac_poisson embedding. The inferred cell transitions are visualized using a Sankey diagram, excluding transitions that come from $< 5\%$ of cells of one cell type.

Human pluripotent stem cell culture and differentiation

hESC culture. The NEUROG3^{-/-} hESC line was previously created using CRISPR/Cas9 to disrupt endogenous expression with a frame-shift INDEL⁵⁵. hESCs were maintained in mTeSR (StemCell Technologies) on hESC-qualified Matrigel (BD Biosciences) coated plates under standard culture conditions (37 °C, 5% CO₂ and 95% humidity). Cells were routinely passaged every four days with Dispase (Invitrogen). High-titer lentivirus with inducible NEUROG3 vectors was added to the media of newly plated hESCs. After 24 h, the media was replaced with mTeSR containing selective antibiotic G418 (500 mg/mL, Sigma). All transduced cell lines were maintained under selection.

hESC differentiation. hESCs were dispersed with Accutase (StemCell Technologies), washed, collected, resuspended in mTeSR containing 10 mM ROCK inhibitor (Y-27632, Tocris Bioscience), and plated at a concentration of 1×10^5 cells/cm² on Matrigel-coated, 24-well plates (Nunc, Delta treated). When cells reached 75% confluency, differentiation was initiated. Day 0 medium was RPMI 1640 supplemented with non-essential amino acids, 100 ng/mL Activin A (Cell Guidance Systems) and 50 ng/mL BMP4 (R&D Systems). Days 1–2 medium contained 0.2% tetracycline-free FBS (Hyclone) but not BMP4. Days 3–4 medium was RPMI 1640 containing 2% FBS, 50 ng/mL FGF7 (R&D Systems), and 50 ng/mL Noggin (R&D Systems). Days 5–8 medium was high-glucose-DMEM (Gibco) containing 50 ng/mL Noggin, 2 mM all-trans retinoic acid (Stemgent), and 1% (0.5x) B27 without vitamin A (Gibco). Days 9–12 medium was high-glucose-DMEM supplemented with 1% B27 and 25 ng/mL Noggin.

hiPSC culture. hiPSCs were maintained in StemMACS™ iPS-Brew XF (iPS-Brew) culture medium (Miltenyi Biotec) on Geltrex (Gibco) coated dishes under standard culture conditions (37 °C, 5% CO₂ and 95% humidity). Cells were passaged every 4–5 days by single-cell dispersion using Accutase (Sigma-Aldrich). For aggregate suspension cultures, hiPSCs were detached with Accutase and seeded at a concentration of 0.8×10^6 cells/mL in iPS-Brew supplemented with Y-27632 to a 30-mL spinner flask (Reprocell) on a magnetic stirrer (Cultistir, Able) set at 60 rpm in a humidified 5% CO₂ 37 °C incubator. The aggregates were split every 3–4 days with Accutase.

hiPSC differentiation. To initiate differentiation, hiPSC aggregates were dispersed into single-cell suspension and seeded at 0.8×10^6 cells/mL in a 30-mL spinner flask. Cells were cultured for 72 h in iPS-Brew and then differentiated towards pancreatic islets with a 6-stage protocol (detailed in Supplementary Data 10)¹²³. Samples were collected on differentiation stage 5 day 7, stage 6 day 7 and stage 6 day 14.

ChIP, RNA and ATAC-seq and data analysis

ChIP-seq. Samples were collected on differentiation day 9 post 8 h NEUROG3 induction. Cells were cross-linked in 1% formaldehyde in PBS for 12 min at RT and were quenched by 0.125 M glycine. Nuclei were pelleted in lysis buffer (10 mM Tris-HCl pH 8.0, 10 mM NaCl, 0.2% NP-40). For chromatin fragmentation, cells were resuspended in Nuclear Lysis Buffer (20 mM Tris-HCl pH 8.0, 0.1% SDS, 2 mM EDTA) and sonicated in Diagenode Sonicator for 9 cycles of 20 s on, 60 s off at 4 °C. The desired amount of fragmented chromatin was supplemented with Nuclear Lysis Dilution Buffer (20 mM Tris-HCl pH 8.0, 0.1% SDS, 2 mM EDTA, 150 mM NaCl, 1% Triton X-100) and precleared with

blocked Protein G magnetic beads (Thermo Fisher Scientific) with rotation at 4 °C for 3 h. 1% precleared samples were saved as input and each 25 µg sample was incubated with 20 µg Protein G magnetic beads preloaded with 5 µg of NEUROG3 antibody (R&D) overnight with rotation at 4 °C. Beads were then washed sequentially at 4 °C using: (1) Washing A (150 mM NaCl, 20 mM Tris-HCl pH 8.0, 2 mM EDTA, 0.1% SDS, 1% Triton X-100, 0.1% sodium deoxycholate), (2) Serial Washing B (20 mM Tris-HCl pH 8.0, 2 mM EDTA, 0.1% SDS, 1% Triton X-100, 0.1% sodium deoxycholate) with 500 mM, 1 M and 2 M NaCl, (3) Washing C (50 mM Tris-HCl pH 8.0, 2 mM EDTA, 500 mM LiCl, 1% NP-40, 0.5% sodium deoxycholate) and final 2 times wash with TE (10 mM Tris-HCl pH 8.0, 10 mM EDTA). Protein-DNA complexes were eluted from the beads in Elution Buffer (50 mM Tris-HCl pH 8.0, 10 mM EDTA, 1% SDS) at 65 °C for 30 min. Cross-links were reversed at 65 °C overnight, and DNA was purified with phenol:chloroform (1:1), chloroform and ethanol for library construction and sequencing at CCHMC DNA sequencing and Genotyping Core Facility.

ATAC-seq. Samples were collected on differentiation day 9 and day 12 with or without 8 or 24 h NEUROG3 induction. Cells were dissociated with Accutase into a single-cell suspension. About 50,000 cells were collected for lysis and transposition based on the Omni-ATAC protocol. Briefly, cells were lysed with Lysis buffer (10 mM Tris-HCl pH 8.0, 10 mM NaCl, 3 mM MgCl₂, 0.1% NP-40, 0.1% Tween-20, 0.01% Digitonin), washed with washing buffer (10 mM Tris-HCl pH 8.0, 10 mM NaCl, 3 mM MgCl₂, 0.1% Tween-20), and incubated with Transposition buffer (Nextra) at 37 °C for 30 min rotated at 1,000 rpm. DNA was purified (Qiagen MinElute) for library construction and sequencing at the Kottyan lab, CCHMC.

RNA Isolation and RNA-seq. Samples were collected on differentiation days 9, 10, 11, and 12 with or without 8 h NEUROG3 induction. All RNA samples were column-purified using a NucleoSpin RNA kit (Macherey-Nagel) with an on-column DNase digestion according to the manufacturer's protocol. Frozen RNA samples were sent for library construction and sequencing at Novogene.

NGS data processing. ChIP-seq was performed at a depth of 30 M reads per sample. ATAC-seq was performed at a depth of 20 M reads per sample. RNA-seq was performed at 150 bp paired-end with a depth of 30 M reads per sample. Fastq read files for each sample were obtained and then aligned using the Computational Suite for Bioinformaticians and Biologists version 3.0 (CSBB-v3.0) to remove low quality bases and potential adapter contamination. Open chromatin region and transcription factor binding sites (NEUROG3) were then called, filtered, and annotated to the nearest gene using MACS2¹²⁴ (v2.1.0) and HOMER¹²⁵ (v4.11). Motif analysis annotations with genomic features and peak overlap were performed using HOMER, ChIPseeker¹²⁶ (v1.38.0) and ChIPpeakAnno¹²⁷ (v3.36.1). Raw transcript counts and normalized transcripts per million (TPM) values were obtained and analyzed for differential expression with DESeq2¹²⁸ (v1.42.1). For differential expression, statistical and biological significance was set at FDR < 0.05, log-fold-change > 1, with a minimum of 1 count from the triplicates of the total 12 samples (Supplementary Data 1). Directed GRN was constructed with the integration of ChIP, ATAC and RNA-seq data using mLASSO-STARS¹²⁹ based algorithm. Nodes of GRN were filtered with differentially expressed genes and protein function, whereas edges were pruned with interaction types and weights. Graph analysis on the constructed GRN was performed for cliques, coefficients, shortest paths, centrality, and community. The results were compared across sampling time points. Dynamic simulation and pseudo-interruption were conducted based on graph diffusion with the hyperbolic tangent as the activation function and as the filtration.

Data sources

Published datasets are available publicly or accessible upon reasonable request from the corresponding authors of the original publications. Human data were from Yu et al.⁴¹ (OMix (<https://bigd.big.ac.cn/omix/>) identifier OMIX236), Goncalves et al.³⁷ (European Genome-Phenome Archive (EGA, <https://ega-archive.org/>) ID# EGAD00001007506), de la O et al.³⁰ (database of Genotypes and Phenotypes (dbGaP) accession phs002693.v1.p1). Mouse data were from Bastidas-Ponce et al.³⁵ (Gene Expression Omnibus (GEO) accession GSE132188), Byrnes et al.³⁶ (GEO accession GSE101099), Han et al.³⁸ (GEO accession GSE136689), Krentz et al.³⁹ (GEO accession GSE120522), Yu et al.⁴² (GEO accession GSE115931).

Ethics

The use of pigs in this study was approved by the Committee on Animal Health and Care of the local government body of the state of Upper Bavaria in Germany for the wild-type and *INS-eGFP*⁷⁸ German Landrace pigs from the Ludwig Maximilian University of Munich (LMU, Permission No. 55.2-2532.Vet_02-17-136) and the *PTF1A*-codon-improved Cre (iCre)⁷⁹ and *ROSA-mTmG*⁸⁰ pigs from the Technical University of Munich (TUM, Permission No. 55.2-2532.Vet_02-18-33). Experiments were conducted in accordance with the German Animal Welfare Act and Directive 2010/63/EU on the protection of animals used for scientific purposes.

The parent human embryonic stem cell (hESC) line WA01 (H1) was obtained from WiCell. All experiments with hESCs were approved by the Cincinnati Children's Hospital ESCRO committee (Protocol# EIPDB2713).

The human-induced pluripotent stem cell (hiPSC) line HMGUi001-A⁸¹ was generated at Helmholtz Munich. All experiments with hiPSCs were approved by the Ethics Committee of the Technical University Munich (219/20 S, 290/20 S).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The data generated in this study have been deposited in NCBI's Gene Expression Omnibus. The pig pancreas data is accessible through GEO Series accession number GSE262280. The hESC datasets are accessible through GSE261950, GSE261951, and GSE261952. Sequencing data from pig pancreas were aligned using the Sscrofa11.1 assembly of the pig genome (https://www.ebi.ac.uk/ena/browser/view/GCA_000003025.6) and the improved annotation based on the Ensembl annotation version 101 (see Methods; Improved pig gene annotation file is available at: <https://github.com/theislab/pig-embryo-ana>). Sequencing data from hESC were aligned using the GRCh37/hg19 reference genome and Ensembl gene annotation (https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001405.13/). Any other data supporting the findings of this study are available from the corresponding authors on reasonable request.

Code availability

Jupyter notebooks to reproduce the analysis and figures are available at: <https://github.com/theislab/pig-embryo-ana>¹³⁰.

References

- Nair, G. & Hebrok, M. Islet formation in mice and men: lessons for the generation of functional insulin-producing β -cells from human pluripotent stem cells. *Curr. Opin. Genet. Dev.* **32**, 171–180 (2015).
- Zong, W. et al. Transcriptomic congruence analysis for evaluating model organisms. *Proc. Natl Acad. Sci. USA* **120**, e2202584120 (2023).

3. Mastracci, T. L. & Sussel, L. The endocrine pancreas: insights into development, differentiation, and diabetes. *Wiley Interdiscip. Rev. Dev. Biol.* **1**, 609–628 (2012).
4. Zorn, A. M. & Wells, J. M. Vertebrate endoderm development and organ formation. *Annu Rev. Cell Dev. Biol.* **25**, 221–251 (2009).
5. Pan, F. C. & Wright, C. Pancreas organogenesis: from bud to plexus to gland. *Dev. Dyn.* **240**, 530–565 (2011).
6. Hick, A. C. et al. Mechanism of primitive duct formation in the pancreas and submandibular glands: a role for SDF-1. *BMC Dev. Biol.* **9**, 66 (2009).
7. 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).
8. Melton, D. A. Applied developmental biology: making human pancreatic beta cells for diabetics. *Curr. Top. Dev. Biol.* **117**, 65–73 (2016).
9. Siehler, J., Blochinger, A. K., Meier, M. & Lickert, H. Engineering islets from stem cells for advanced therapies of diabetes. *Nat. Rev. Drug Discov.* **20**, 920–940 (2021).
10. Migliorini, A., Nostro, M. C. & Sneddon, J. B. Human pluripotent stem cell-derived insulin-producing cells: A regenerative medicine perspective. *Cell Metab.* **33**, 721–731 (2021).
11. Augsornworawat, P. et al. Single-nucleus multi-omics of human stem cell-derived islets identifies deficiencies in lineage specification. *Nat. Cell Biol.* **25**, 904–916 (2023).
12. Zhu, H. et al. Understanding cell fate acquisition in stem-cell-derived pancreatic islets using single-cell multiome-inferred regulomes. *Dev. Cell* **58**, 727–743.e711 (2023).
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. Ottoni, C. et al. Pig domestication and human-mediated dispersal in western Eurasia revealed through ancient DNA and geometric morphometrics. *Mol. Biol. Evol.* **30**, 824–832 (2013).
15. Kumar, S., Stecher, G., Suleski, M. & Hedges, S. B. Timetree: a resource for timelines, timetrees, and divergence times. *Mol. Biol. Evol.* **34**, 1812–1819 (2017).
16. Wernersson, R. et al. Pigs in sequence space: a 0.66X coverage pig genome survey based on shotgun sequencing. *BMC Genomics* **6**, 70 (2005).
17. Lunney, J. K. et al. Importance of the pig as a human biomedical model. *Sci. Transl. Med.* **13**, eabd5758 (2021).
18. Kim, S. et al. Molecular and genetic regulation of pig pancreatic islet cell development. *Development* **147** <https://doi.org/10.1242/dev.186213> (2020).
19. Pilz, J. et al. Functional maturation and longitudinal imaging of intraportal neonatal porcine islet grafts in genetically diabetic pigs. *Am. J. Transplant.* **24**, 1395–1405 (2024).
20. Bakhti, M., Bottcher, A. & Lickert, H. Modelling the endocrine pancreas in health and disease. *Nat. Rev. Endocrinol.* **15**, 155–171 (2019).
21. Renner, S. et al. Porcine models for studying complications and organ crosstalk in diabetes mellitus. *Cell Tissue Res* **380**, 341–378 (2020).
22. Zettler, S. et al. A decade of experience with genetically tailored pig models for diabetes and metabolic research. *Anim. Reprod.* **17**, e20200064 (2020).
23. Jennings, R. E. et al. Development of the human pancreas from foregut to endocrine commitment. *Diabetes* **62**, 3514–3522 (2013).
24. Sarkar, S. A. et al. Global gene expression profiling and histochemical analysis of the developing human fetal pancreas. *Diabetologia* **51**, 285–297 (2008).
25. Robb, P. The development of the islets of Langerhans in the human foetus. *Q. J. Exp. Physiol. Cogn. Med. Sci.* **46**, 335–343 (1961).
26. 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).
27. van Gurp, L. et al. Generation of human islet cell type-specific identity gene sets. *Nat. Commun.* **13**, 2020 (2022).
28. Dorrell, C. et al. Transcriptomes of the major human pancreatic cell types. *Diabetologia* **54**, 2832–2844 (2011).
29. Olaniru, O. E. et al. Single-cell transcriptomic and spatial landscapes of the developing human pancreas. *Cell Metab.* **35**, 184–199.e185 (2023).
30. de la O., S. et al. Single-Cell Multi-Omic Roadmap of Human Fetal Pancreatic Development. *bioRxiv* <https://doi.org/10.1101/2022.02.17.480942> (2022).
31. Chen, M., Yan, J., Han, Q., Luo, J. & Zhang, Q. Identification of hub-methylated differentially expressed genes in patients with gestational diabetes mellitus by multi-omic WGCNA basing epigenome-wide and transcriptome-wide profiling. *J. Cell Biochem* **121**, 3173–3184 (2020).
32. Prashanth, G., Vastrad, B., Tengli, A., Vastrad, C. & Kotturshetti, I. Investigation of candidate genes and mechanisms underlying obesity associated type 2 diabetes mellitus using bioinformatics analysis and screening of small drug molecules. *BMC Endocr. Disord.* **21**, 80 (2021).
33. Taneera, J. et al. Identification of novel genes for glucose metabolism based upon expression pattern in human islets and effect on insulin secretion and glycemia. *Hum. Mol. Genet* **24**, 1945–1955 (2015).
34. Berglund, D. et al. Differential methylation in inflammation and type 2 diabetes genes in siblings born before and after maternal bariatric surgery. *Obes. (Silver Spring)* **24**, 250–261 (2016).
35. Bastidas-Ponce, A. et al. Comprehensive single cell mRNA profiling reveals a detailed roadmap for pancreatic endocrinogenesis. *Development* **146**, dev173849 (2019).
36. Byrnes, L. E. et al. Lineage dynamics of murine pancreatic development at single-cell resolution. *Nat. Commun.* **9**, 3922 (2018).
37. Goncalves, C. A. et al. A 3D system to model human pancreas development and its reference single-cell transcriptome atlas identify signaling pathways required for progenitor expansion. *Nat. Commun.* **12**, 3144 (2021).
38. Han, L. et al. Single cell transcriptomics identifies a signaling network coordinating endoderm and mesoderm diversification during foregut organogenesis. *Nat. Commun.* **11**, 4158 (2020).
39. Krentz, N. A. J. et al. Single-Cell Transcriptome Profiling of Mouse and hESC-Derived Pancreatic Progenitors. *Stem Cell Rep.* **11**, 1551–1564 (2018).
40. Liu, H. et al. Chemical combinations potentiate human pluripotent stem cell-derived 3D pancreatic progenitor clusters toward functional beta cells. *Nat. Commun.* **12**, 3330 (2021).
41. Yu, X. X. et al. Sequential progenitor states mark the generation of pancreatic endocrine lineages in mice and humans. *Cell Res* **31**, 886–903 (2021).
42. Yu, X. X. et al. Defining multistep cell fate decision pathways during pancreatic development at single-cell resolution. *EMBO J* **38** <https://doi.org/10.15252/emboj.2018100164> (2019).
43. Kamiya, M. et al. The cell cycle control gene ZAC/PLAGL1 is imprinted—a strong candidate gene for transient neonatal diabetes. *Hum. Mol. Genet* **9**, 453–460 (2000).
44. Ma, D. et al. Impaired glucose homeostasis in transgenic mice expressing the human transient neonatal diabetes mellitus locus, TNDM. *J. Clin. Invest* **114**, 339–348 (2004).
45. Levine, M. & Davidson, E. H. Gene regulatory networks for development. *Proc. Natl Acad. Sci. USA* **102**, 4936–4942 (2005).

46. Schuijers, J. et al. Ascl2 acts as an R-spondin/Wnt-responsive switch to control stemness in intestinal crypts. *Cell Stem Cell* **16**, 158–170 (2015).
47. Fleck, J. S. et al. Inferring and perturbing cell fate regulomes in human brain organoids. *Nature* **621**, 365–372 (2023).
48. Cao, Z. J. & Gao, G. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nat. Biotechnol.* **40**, 1458–1466 (2022).
49. van Raalte, D. H. et al. Glucocorticoid receptor gene polymorphisms are associated with reduced first-phase glucose-stimulated insulin secretion and disposition index in women, but not in men. *Diabet. Med.* **29**, e211–e216 (2012).
50. Henry, C., Close, A. F. & Buteau, J. A critical role for the neural zinc factor ST18 in pancreatic beta-cell apoptosis. *J. Biol. Chem.* **289**, 8413–8419 (2014).
51. Kalousova, A. et al. Dachshund homologues play a conserved role in islet cell development. *Dev. Biol.* **348**, 143–152 (2010).
52. Wang, H. et al. The signaling protein GIV/Girdin mediates the Nephroin-dependent insulin secretion of pancreatic islet beta cells in response to high glucose. *J. Biol. Chem.* **299**, 103045 (2023).
53. Lange, M. et al. CellRank for directed single-cell fate mapping. *Nat. Methods* **19**, 159–170 (2022).
54. Zhang, X. et al. A comprehensive structure-function study of neurogenin3 disease-causing alleles during human pancreas and intestinal organoid development. *Dev. Cell* **50**, 367–380.e367 (2019).
55. McGrath, P. S., Watson, C. L., Ingram, C., Helmrath, M. A. & Wells, J. M. The basic helix-loop-helix transcription factor neurog3 is required for development of the human endocrine pancreas. *Diabetes* **64**, 2497–2505 (2015).
56. Schreiber, V. et al. Extensive NEUROG3 occupancy in the human pancreatic endocrine gene regulatory network. *Mol. Metab.* **53**, 101313 (2021).
57. Bader, E. et al. Identification of proliferative and mature beta-cells in the islets of Langerhans. *Nature* **535**, 430–434 (2016).
58. Yoshihara, E. et al. Immune-evasive human islet-like organoids ameliorate diabetes. *Nature* **586**, 606–611 (2020).
59. Alvarez-Dominguez, J. R. et al. Circadian entrainment triggers maturation of human in vitro islets. *Cell Stem Cell* **26**, 108–122.e110 (2020).
60. Kamimoto, K. et al. Dissecting cell identity via network inference and in silico gene perturbation. *Nature* **614**, 742–751 (2023).
61. Docherty, L. E. et al. Clinical presentation of 6q24 transient neonatal diabetes mellitus (6q24 TNDM) and genotype-phenotype correlation in an international cohort of patients. *Diabetologia* **56**, 758–762 (2013).
62. Spengler, D. et al. Regulation of apoptosis and cell cycle arrest by Zac1, a novel zinc finger protein expressed in the pituitary gland and the brain. *EMBO J.* **16**, 2814–2825 (1997).
63. Varrault, A. et al. Identification of Plagl1/Zac1 binding sites and target genes establishes its role in the regulation of extracellular matrix genes and the imprinted gene network. *Nucleic Acids Res.* **45**, 10466–10480 (2017).
64. Schulte, D. & Geerts, D. MEIS transcription factors in development and disease. *Development* **146** <https://doi.org/10.1242/dev.174706> (2019).
65. Vanheer, L. et al. Predicting the Key Regulators of Cell Identity in Human Adult Pancreas. *bioRxiv* <https://doi.org/10.1101/2020.09.23.310094> (2020).
66. Zhang, X. et al. Pax6 is regulated by Meis and Pbx homeoproteins during pancreatic development. *Dev. Biol.* **300**, 748–757 (2006).
67. Van den Berge, K. et al. Trajectory-based differential expression analysis for single-cell sequencing data. *Nat. Commun.* **11**, 1201 (2020).
68. Scavuzzo, M. A. et al. Pancreatic Cell Fate Determination Relies on Notch Ligand Trafficking by NFIA. *Cell Rep.* **25**, 3811–3827.e3817 (2018).
69. Kim, S. K. et al. Pbx1 inactivation disrupts pancreas development and in *lplf1*-deficient mice promotes diabetes mellitus. *Nat. Genet.* **30**, 430–435 (2002).
70. Zhang, P. et al. Expression of COUP-TFII in metabolic tissues during development. *Mech. Dev.* **119**, 109–114 (2002).
71. Boutant, M. et al. COUP-TFII controls mouse pancreatic beta-cell mass through GLP-1-beta-catenin signaling pathways. *PLoS One* **7**, e30847 (2012).
72. Klein, D. et al. Mapping cells through time and space with moscot. *Nature* **638**, 1065–1075 (2025).
73. Hrovatin, K. et al. Integrating single-cell RNA-seq datasets with substantial batch effects. *bioRxiv* <https://doi.org/10.1101/2023.11.03.565463> (2024).
74. Hsu, C. C. et al. The differential distributions of ASPM isoforms and their roles in Wnt signaling, cell cycle progression, and pancreatic cancer prognosis. *J. Pathol.* **249**, 498–508 (2019).
75. Zheng, Y. Dbl family guanine nucleotide exchange factors. *Trends Biochem. Sci.* **26**, 724–732 (2001).
76. Pauerstein, P. T. et al. Dissecting human gene functions regulating islet development with targeted gene transduction. *Diabetes* **64**, 3037–3049 (2015).
77. Rubio-Cabezas, O. et al. Neurogenin 3 is important but not essential for pancreatic islet development in humans. *Diabetologia* **57**, 2421–2424 (2014).
78. Kemter, E. et al. INS-eGFP transgenic pigs: a novel reporter system for studying maturation, growth and vascularisation of neonatal islet-like cell clusters. *Diabetologia* **60**, 1152–1156 (2017).
79. Kalla, D. et al. The Missing Link: Cre Pigs for Cancer Research. *Front Oncol.* **11**, 755746 (2021).
80. Li, S. et al. Dual fluorescent reporter pig for Cre recombination: transgene placement at the ROSA26 locus. *PLoS One* **9**, e102455 (2014).
81. Wang, X. et al. Genome-wide analysis of PDX1 target genes in human pancreatic progenitors. *Mol. Metab.* **9**, 57–68 (2018).
82. Beiki, H. et al. Improved annotation of the domestic pig genome through integration of Iso-Seq and RNA-seq data. *BMC Genomics* **20**, 344 (2019).
83. Pertea, G. & Pertea, M. GFF Utilities: GffRead and GffCompare. *F1000Res* **9** (2020).
84. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
85. Lun, A. T., Bach, K. & Marioni, J. C. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.* **17**, 75 (2016).
86. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech.: Theory Exp.* **2008**, P10008 (2008).
87. Lun, A. T. L. et al. EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome Biol.* **20**, 63 (2019).
88. Luecken, M. D. et al. Benchmarking atlas-level data integration in single-cell genomics. *Nat. Methods* **19**, 41–50 (2022).
89. Zheng, G. X. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
90. Hie, B., Bryson, B. & Berger, B. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nat. Biotechnol.* **37**, 685–691 (2019).
91. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **15**, 1053–1058 (2018).

92. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9**, 5233 (2019).
93. Wolock, S. L., Lopez, R. & Klein, A. M. Scrublet: computational identification of cell doublets in single-cell transcriptomic data. *Cell Syst.* **8**, 281–291 e289 (2019).
94. Gayoso, A. & Shor, J. Doublet Detection. *Zenodo*. <https://doi.org/10.5281/zenodo.2678042> (2018).
95. Bais, A. S. & Kostka, D. scds: computational annotation of doublets in single-cell RNA sequencing data. *Bioinformatics* **36**, 1150–1158 (2020).
96. Germain, P. L., Lun, A., Garcia Meixide, C., Macnair, W. & Robinson, M. D. Doublet identification in single-cell sequencing data using scDblFinder. *F1000Res* **10**, 979 (2021).
97. McGinnis, C. S., Murrow, L. M. & Gartner, Z. J. DoubletFinder: doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. *Cell Syst.* **8**, 329–337.e324 (2019).
98. Stuart, T., Srivastava, A., Madad, S., Lareau, C. A. & Satija, R. Single-cell chromatin state analysis with Signac. *Nat. Methods* **18**, 1333–1341 (2021).
99. Gayoso, A. et al. A Python library for probabilistic analysis of single-cell omics data. *Nat. Biotechnol.* **40**, 163–166 (2022).
100. Stark, S. G. et al. SCIM: universal single-cell matching with unpaired feature sets. *Bioinformatics* **36**, i919–i927 (2020).
101. van Dijk, D. et al. Recovering gene interactions from single-cell data using data diffusion. *Cell* **174**, 716–729.e727 (2018).
102. Granja, J. M. et al. ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nat. Genet.* **53**, 403–411 (2021).
103. Bredikhin, D., Kats, I. & Stegle, O. MUON: multimodal omics analysis framework. *Genome Biol.* **23**, 42 (2022).
104. Bernstein, N. J. et al. Solo: Doublet Identification in Single-Cell RNA-Seq via Semi-Supervised Deep Learning. *Cell Syst.* **11**, 95–101.e105 (2020).
105. Thibodeau, A. et al. AMULET: a novel read count-based method for effective multiplet detection from single nucleus ATAC-seq data. *Genome Biol.* **22**, 252 (2021).
106. Townes, F. W., Hicks, S. C., Aryee, M. J. & Irizarry, R. A. Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. *Genome Biol.* **20**, 295 (2019).
107. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
108. Martens, L. D., Fischer, D. S., Yezpez, V. A., Theis, F. J. & Gagneur, J. Modeling fragment counts improves single-cell ATAC-seq analysis. *Nat. Methods* **21**, 28–31 (2024).
109. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
110. Dunn, P. K. & Smyth, G. K. *Generalized linear models with examples in R*. Vol. 53 (Springer, 2018).
111. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.: Ser. B (Methodol.)* **57**, 289–300 (1995).
112. Xie, Z. et al. Gene Set Knowledge Discovery with Enrichr. *Curr. Protoc.* **1**, e90 (2021).
113. Kinsella, R. J. et al. Ensembl BioMart: a hub for data retrieval across taxonomic space. *Database* **2011** <https://doi.org/10.1093/database/bar030> (2011).
114. Mukaka, M. M. Statistics corner: A guide to appropriate use of correlation coefficient in medical research. *Malawi Med J.* **24**, 69–71 (2012).
115. Setty, M. et al. Characterization of cell fate probabilities in single-cell data with Palantir. *Nat. Biotechnol.* **37**, 451–460 (2019).
116. Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
117. Hagberg, A., Swart, P. J. & Schult, D. A. *Exploring network structure, dynamics, and function using NetworkX*, <<https://www.osti.gov/biblio/960616>> (2008).
118. Persad, S. et al. SEACells infers transcriptional and epigenomic cellular states from single-cell genomics data. *Nat. Biotechnol.* **41**, 1746–1757 (2023).
119. Lander, E. S. et al. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
120. Linderman, G. C. et al. Zero-preserving imputation of single-cell RNA-seq data. *Nat. Commun.* **13**, 192 (2022).
121. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. Methods* **14**, 975–978 (2017).
122. Solomon, J. et al. Convolutional wasserstein distances: efficient optimal transportation on geometric domains. *ACM Trans. Graph.* **34**, Article 66 <https://doi.org/10.1145/2766963> (2015).
123. Velazco-Cruz, L. et al. Acquisition of Dynamic Function in Human Stem Cell-Derived beta Cells. *Stem Cell Rep.* **12**, 351–365 (2019).
124. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
125. Duttke, S. H., Chang, M. W., Heinz, S. & Benner, C. Identification and dynamic quantification of regulatory elements using total RNA. *Genome Res* **29**, 1836–1846 (2019).
126. Wang, Q. et al. Exploring Epigenomic Datasets by ChIPseeker. *Curr. Protoc.* **2**, e585 (2022).
127. Zhu, L. J. et al. ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. *BMC Bioinform.* **11**, 237 (2010).
128. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
129. Miraldi, E. R. et al. Leveraging chromatin accessibility for transcriptional regulatory network inference in T Helper 17 Cells. *Genome Res* **29**, 449–463 (2019).
130. Yang, K. et al. A multimodal cross-species comparison of pancreas development. *figshare* <https://doi.org/10.6084/m9.figshare.30069532> (2025).

Acknowledgements

The authors thank S. Schirge, A. Bastidas-Ponce, L. Appel, I. Kunze, K. Diemer (IDR, Helmholtz Munich), C. Blechinger, M. Krätzl (CiMM, Ludwig Maximilian University of Munich) and D. Kalla (Technical University of Munich) for their excellent technical assistance; D. Kechele (Cincinnati Children's Hospital Medical Center) for great support in organizing and depositing next generation sequence data; M. Bakhti and A. Böttcher (IDR, Helmholtz Munich) for helpful discussions; A. Grapin-Botton, J. Stratmann and C.A. Gonçalves (Max Planck Institute of Molecular - Cell Biology and Genetics) for generously sharing human fetal pancreas data; D. Klein, M. Lange, V. Bergen, and A. Moinfar (ICB, Helmholtz Munich) for bioinformatic analysis advice; the animal caretakers in the pig facilities in Ludwig Maximilian University of Munich and Technical University of Munich. The authors of this study received funding from the European Research Council (grant 101054564 – to H.L. and grant 101054957 – to F.J.T), the German Research Foundation (LI 1006/3-1 – to H.L., TH 900/13-1 – to F.J.T, WO 685/22-1 – to E.W., Transregio Research Unit 127 and 205 funding – to E.K. and E.W., Collaborative Research Centre 1321 funding to A.S. and T.F.), the Juvenile Diabetes Research Foundation (grant 3-SRA-2023-1420-S-B – to H.L, E.W., F.J.T, grant 2-SRA-2019-773-S-B – to J.B.S.), the German Federal Ministry of Education and Research to the German Centre for Diabetes Research (grant 82DZD00802 – to E.W.), the National Institutes of Health (grant R01DK118421 – to J.B.S., grant R25GM056847-23 – to S.O., grant R01DK118421-02S1 – to S.O.), the Kraft Family Fellowship to the UCSF Diabetes Center (S.O.).

Author contributions

K.Y. and H.L. designed the study. H.S. and F.J.T. designed bioinformatic analysis plans. K.Y., H.S., and H.L. wrote the manuscript. B.K., E.W., and E.K. advised on the study design and provided pig pancreas samples. K.Y. and M.S. prepared samples for single-cell genomics. K.Y., H.S., M.S., and K.H. performed bioinformatic analyses. S.O. and J.B.S. provided human fetal data and analysis advice. X.Z. designed and performed hESC differentiation experiments and analyzed the data, supervised by J.M.W., K.Y., E.S., and K.S. designed and performed hiPSC differentiation experiments and analyzed the data. M.U. and T.W. improved pig genome annotations and sequenced the libraries. K.F. and T.F. provided pig pancreas samples, supervised by A.S. H.L. provided financial support. F.J.T. and H.L. supervised the study. All authors reviewed and edited the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

F.J.T. consults for Immunai Inc., Singularity Bio B.V., CytoReason Ltd, Cellarity, Curie Bio Operations, LLC, and has an ownership interest in Dermagnostix GmbH and Cellarity. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-025-64774-4>.

Correspondence and requests for materials should be addressed to Fabian J. Theis or Heiko Lickert.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2025

¹Institute of Diabetes and Regeneration Research (IDR), Helmholtz Munich, Neuherberg, Germany. ²German Center for Diabetes Research (DZD), Neuherberg, Germany. ³Institute of Computational Biology (ICB), Helmholtz Munich, Neuherberg, Germany. ⁴Institute for Stroke and Dementia Research, University Hospital, Ludwig Maximilian University of Munich, Munich, Germany. ⁵Department of Mathematics, Technical University of Munich, Munich, Germany. ⁶Department of Cell and Tissue Biology, University of California, San Francisco, USA. ⁷Diabetes Center, University of California, San Francisco, USA. ⁸Eli and Edythe Broad Center of Regeneration Medicine and Stem Cell Research, University of California, San Francisco, USA. ⁹Division of Developmental Biology, Cincinnati Children's Hospital Medical Center, Cincinnati, USA. ¹⁰Center for Stem Cell and Organoid Medicine (CuSTOM), Cincinnati Children's Hospital Medical Center, Cincinnati, USA. ¹¹Core Facility Genomics, Helmholtz Munich, Neuherberg, Germany. ¹²Chair of Livestock Biotechnology, Department of Molecular Life Sciences, School of Life Sciences, Technical University of Munich, Freising, Germany. ¹³Division of Endocrinology, Cincinnati Children's Hospital Medical Center, Cincinnati, USA. ¹⁴Chair for Molecular Animal Breeding and Biotechnology Gene Center, Ludwig Maximilian University of Munich, Munich, Germany. ¹⁵Center for Innovative Medical Models (CiMM), Ludwig Maximilian University of Munich, Munich, Germany. ¹⁶School of Life Sciences Weihenstephan, Technical University of Munich, Freising, Germany. ¹⁷School of Medicine, Technical University of Munich, Munich, Germany. ¹⁸These authors contributed equally: Kaiyuan Yang, Hannah Spitzer. ✉ e-mail: fabian.theis@helmholtz-munich.de; heiko.lickert@helmholtz-munich.de