Single-cell lineage analysis reveals extensive multimodal transcriptional control during directed beta-cell differentiation

Chen Weng1,2,6, Jiajia Xi1,6, Haiyan Li1, Jian Cui1, Anniya Gu1,3, Sisi Lai1,2, Konstantin Leskov1, Luxin Ke1,4, Fulai Jin1,5,6 and Yan Li1

The in vitro differentiation of insulin-producing beta-like cells can model aspects of human pancreatic development. Here, we generate 95,308 single-cell transcriptomes and reconstruct a lineage tree of the entire differentiation process from human embryonic stem cells to beta-like cells to study temporally regulated genes during differentiation. We identify so-called ‘switch genes’ at the branch point of endocrine/non-endocrine cell fate choice, revealing insights into the mechanisms of differentiation-promoting reagents, such as NOTCH and ROCKII inhibitors, and providing improved differentiation protocols. Over 20% of all detectable genes are activated multiple times during differentiation, even though their enhancer activation is usually unimodal, indicating extensive gene reuse driven by different enhancers. We also identify a stage-specific enhancer at the TCF7L2 locus for diabetes, uncovered by genome-wide association studies, that drives a transient wave of gene expression in pancreatic progenitors. Finally, we develop a web app to visualize gene expression on the lineage tree, providing a comprehensive single-cell data resource for researchers studying islet biology and diabetes.

One major hope for diabetes therapy is to generate many functional, transplantable beta cells from patient-derived pluripotent cells. In the past decade, a few in vitro protocols have been developed to differentiate human embryonic stem cells (hESCs) into pancreatic progenitor cells, and further into functional beta-like cells; these cells can be transplanted into mice, mature into glucose-responsive insulin-producing cells in vivo and prevent or reverse diabetes1–5. These studies also established that the in vitro beta-cell differentiation process largely recapitulates key developmental stages in vivo based on the sequential expression of various developmental markers. Therefore, the stem cell system is also a fantastic tool for the study of human pancreatic development.

Current protocols for the directed beta-cell differentiation still have some limitations: (1) the differentiation generates heterogeneous cell populations; (2) differentiation efficiency can be highly variable between different hESC or human induced pluripotent stem cell (hiPSC) lines; (3) the resulting beta-like cells are still not fully equivalent to primary beta cells at molecular and physiological levels6,7,8. Protocol improvement can clearly benefit from better understanding of cell fate choices and beta-cell maturation during differentiation. To overcome the cellular heterogeneity issues, single-cell RNA-sequencing (scRNA-seq) technologies have been used to study the directed beta-cell differentiation, with hESC or hiPSC lines possessing high differentiation propensity towards pancreatic lineage2,6. However, although these recent papers have made a number of important discoveries, they lacked a global picture of molecular events during the cell fate choices along the entire differentiation process, especially those in undesired non-endocrine lineages that may substantially affect the differentiation efficiency in low-efficiency hESC lines.

To fill this gap, here we intentionally performed the differentiation with a relatively low-efficiency hESC line (H1) under a standard seven-stage protocol9, featuring mediocre beta-cell yield and increased cellular heterogeneity. This provided enough cell coverage to analyse both desired and divergent lineages in cell fate determination. We generated transcriptome data from 95,308 single cells with Drop-seq11 and computationally reconstructed the entire lineage tree to systematically reveal the differentiation programmes and lineage branches. We demonstrated that this data resource is highly valuable for the study of islet biology and diabetes.

Results

A time-series single-cell transcriptomic atlas during pancreatic beta-cell differentiation. We profiled the single-cell transcriptomes during the differentiation of H1 cells with a previously published seven-stage protocol (Extended Data Fig. 1a). The functionality of the resulting beta-like cells was confirmed with a glucose-stimulated insulin secretion (GSIS) assay (Extended Data Fig. 1b). We collected data from 87,769 single cells across 12 time points and clustered the cells from every time point into populations (Fig. 1a,c; see quality-control summary in Supplementary Data 1). The cultures were rather homogeneous before stage 4 (S4; pancreatic progenitors) but became much more heterogeneous during later stages (Fig. 1a). Therefore, starting from S4, we performed two rounds of unsupervised clustering: the first round of clustering defined endocrine and non-endocrine cell populations with...
endocrine marker genes \textit{NEUROD1}, \textit{CPE} and \textit{CHGA} (Fig. 1a and Extended Data Fig. 2a,b); the second round of clustering further identified more sub-populations. We identified marker genes for all cell populations in each stage (Supplementary Figs. 1 and 2, Methods and Supplementary Data 1), including surface markers (Supplementary Fig. 3).

The hESC line H1 is not a high-yield line for directed beta-cell differentiation. In the final stage (S7), ~35% (1,982) of the cells are endocrine cells, including a beta-like (S7-I; 421 cells) and an alpha-like (S7-H; 732 cells) population (Fig. 1a and Extended Data Fig. 2d,e). Consistent with another recent single-cell study\textsuperscript{10},\textsuperscript{11} we also observed an enterochromaffin cell (EC) population (S7-G; 829 cells) expressing endocrine marker genes \textit{CHGA}, \textit{CHGB} and \textit{NEUROD1} but neither \textit{INS} nor \textit{GCG}. This population indeed expresses EC markers \textit{TPH1}, \textit{DDC}, \textit{SLC18A1}, \textit{ADRA2A}, \textit{LMX1A} and \textit{CXXC14} (Fig. 1a and Extended Data Fig. 2f). Interestingly, we also observed another small endocrine population (S7-J) marked by \textit{CENPF} and \textit{TOP2A} (Fig. 1a and Extended Data Fig. 2g), suggesting that this is a proliferating endocrine cell population.

We also defined duct-like and pancreatic stellate cell (PSC)-like cells among the non-endocrine cell populations, and a delta-like cell population from S5 endocrine cells (Fig. 1b and Extended Data Fig. 1e). Notably, some believed that PSCs are from mesodermal origin because lineage-tracing studies in mice have confirmed that hepatic stellate cells (HSCs) are from mesenchymal cells, even though direct evidence for PSCs is still lacking\textsuperscript{12}. To confirm the observation of PSC-like cells, we reanalysed an independent scRNA-seq dataset\textsuperscript{10} and also observed PSC-like cells expressing the same marker genes (\textit{COL6A3}, \textit{COL1A2}, \textit{ACTA2} and \textit{COL3A1}). Although our endodermal endocrine differentiation system cannot properly model the mesodermal origins of PSCs, this observation may also reflect the heterogeneous origins of PSCs\textsuperscript{12}. Nevertheless, we illustrated a highly diversified cell atlas from the directed beta-cell differentiation.
Benchmarking the similarity between differentiated and primary islet cell populations. We first examined the molecular similarity between induced endocrine cells and primary human pancreatic alpha or beta cells. We previously reported the transcriptome of 39,905 single human islet cells from nine donors using the same single-cell platform (Drop-seq)\(^1\), offering an opportunity for robust comparison. We therefore pooled the S7 datasets and the published Drop-seq data of primary human alpha and beta cells for principle-component analysis (PCA) after normalization (Fig. 1d). The first PC (PC1) clearly separated the primary alpha and beta cells in this analysis and, as expected, alpha-like and beta-like cells co-segregated with alpha and beta cells, respectively. A majority of known endocrine marker genes were properly expressed in the alpha-like or beta-like cells (Fig. 1d–f). However, the alpha-like and beta-like cells were less separated on PC1, indicating that the expression of marker genes distinguishing alpha and beta cells was not full-blown yet (Fig. 1d). To characterize other cell populations, we also performed canonical correlation analysis (CCA)-based co-embedding analysis to ‘anchor’ the in vitro differentiated cell populations with the primary islet cells\(^4\). As expected, alpha-like, beta-like, delta-like, duct-like and PSC-like cells indeed co-clustered with their islet cell counterparts and properly expressed most of their signature genes (Extended Data Fig. 1c–e). Consistent with a previous study, EC-like cells were only present in the differentiation systems\(^2\) (Extended Data Fig. 1c,d).

The second round of PCA clearly separated induced endocrine cells from their counterparts in primary human islets (Fig. 1d). We found that both alpha-like and beta-like cells expressed comparable levels of hormone genes as primary endocrine cells (Fig. 1e). However, alpha-like cells showed significantly higher expression of the INS gene than primary alpha cells, consistent with previous reports of polyhormonal cells from directed pancreatic differentiations\(^2,3,4\) and during human fetal pancreatic development (Fig. 1e). In contrast, beta-like cells only expressed INS but not GCG. We identified hundreds of differentially expressed genes that distinguished the alpha-like or beta-like cell from the primary endocrine cells (Fig. 1g,h and Supplementary Data 2). For example, PAX4 is overexpressed but MAFA is underexpressed in beta-like cells (Fig. 1h and Extended Data Fig. 1e), which is reminiscent of a previous report that PAX4 may promote beta-cell dedifferentiation and proliferation by repressing MAFA expression\(^4\). Gene-set enrichment analysis (GSEA) revealed that the alpha-like and beta-like cells overexpressed liver-specific genes and endoplasmic reticulum genes but underexpressed secretory vesicle genes, consistent with a general concept that the stem cell-derived endocrine cells are in a relatively immature progenitor state compared to primary cells\(^1\) (Fig. 1i,j).

Reconstruction of a lineage tree for highly branched pancreatic differentiation. Most existing single-cell trajectory inference algorithms only analyse linear or simple branching events within a short time window\(^5\). We found that existing unsupervised trajectory inference methods cannot properly recognize the entire time series of our data. For example, Monocle 3 (ref. 23) called one branch connecting S0–S1 but incorrectly called another branch from S0 directly to S2 cells. This is because many S1-specific genes are inactive in both S0 and S2, leading to higher correlation in S0–S2 (Extended Data Fig. 3a–c). We therefore developed a semi-supervised method to construct a lineage tree to investigate the cell fate choices following the time stamps. Briefly, after unsupervised clustering, we placed cell populations from 12 time points into 12 layers (global inference). In each iteration, we only connected the cell populations from two neighbouring layers (local inference) based on the distances between each other in high-dimensional PC space (Methods and Supplementary Figs. 4 and 5). This strategy computationally reconstructed a lineage tree connecting all major cell populations identified from our data (Fig. 1b), which provides an intuitive map to trace the fate of each cell population over the entire differentiation process. Importantly, we found that, for local inference, our method and Monocle 3 defined very similar trajectories and identified the same trajectory genes (Extended Data Fig. 3d–i), allowing us to investigate both the desired beta-cell lineage and the divergent lineages (shown as side branches) in the lineage tree (Fig. 1b).

From the tree, it is clear that the divergent cell populations in early stages (S0–S3) do not contribute to the undesired cell populations in the final stage (S7; Fig. 1b). Notably, a side branch from S4-C contributed to all the non-endocrine cell populations in S7 (Fig. 1b), including the PSC-like (S7-F) and duct-like (S7-D) cells (Fig. 1a,b). A trajectory in S4 (S4-C → S4-A → S4-E) led to the major endocrine cell populations. The branching during S5 is especially complicated since there are many sub-populations at this stage, including a delta-like SST\(^+\)HHEX\(^\#\) cell population (S5-M; Fig. 1a,b and Supplementary Fig. 2). This is likely because cells were just moved from planar culture to an air–liquid interface during S5 (Extended Data Fig. 1a). However, it appears that prolonged air–liquid interface culture eventually leads to fewer populations: the endocrine cells originated from a trajectory of S5–K, S5–G and S5–I populations based on our modelling (Fig. 1b); other S5 populations may become too small to distinguish from unbiased clustering in S6 and S7 after further differentiation.

Every edge in the tree represents the transition between two cell states; we could therefore build a pseudo-time trajectory between any two cell states and project the cells onto the edge. To visualize the dynamic expression of any gene or gene cluster, we used heat map colouring to stain the lineage tree representing their expression level during differentiation (Fig. 2a). We clustered all genes into 64 modules based on their dynamic expression patterns on the tree, each module with different functions (Extended Data Fig. 4 and Supplementary Table 1). For example, modules 50, 52, 56 and 62 contained marker genes for beta-like, alpha-like, PSC-like and duct-like cells, respectively (Fig. 2a and Extended Data Fig. 4). Similarly, gene modules marking early embryogenesis, such as meso/endoderm (module 7), definitive endoderm (module 9) and foregut (module 26), show transient expression patterns on the tree; genes for the desired pancreatic endocrine lineage (module 37), such as PDX1 and NKX6.1, decorated nearly the entire branch of endocrine lineages (Fig. 2a and Extended Data Fig. 4). Interestingly, cell cycle genes (TOP2A, CENPE, MKI67, AURKB and CDK1) were not expressed in the entire endocrine lineage branch, indicating cell cycle exit during the fate choice to differentiate into pancreatic endocrine cells (module 57; Extended Data Fig. 4e)\(^7\). We noticed PPY expression in alpha-like cells (Extended Data Fig. 4b), indicating related lineages, consistent with data from single-cell expression and ATAC-seq (assay for transposase-accessible chromatin followed by sequencing) in primary human islet cells\(^6,22–24\). Similarly, we also observed GHR expression in delta-like cells (Extended Data Fig. 4b). However, it should be noted that our system may not fully reflect alternative lineages, since the differentiation protocol is designed to induce the beta cells.

Stage-specific genes play key roles in beta-cell differentiation and diseases. To characterize the transcriptomic programming during the beta-cell differentiation, we extracted data from all 16 cell populations in the beta-cell lineage (Figs. 1b and 2a), ordered the cells into a linear trajectory and investigated the changes of gene expression along the pseudo-time (Fig. 2b and Supplementary Data 4). Our analysis correctly revealed the sequential expression of the classic developmental markers, including meso-endoderm markers (T, GSC, EOMES and MIXL1), definitive endoderm markers (SOX17, CXCRA4, GATA6 and CER1), gut tube and posterior foregut markers (ONECUT1 and SOX9) and pancreatic progenitor markers (PDX1 and NKX6.1) (Extended Data Fig. 5c). Endocrine
We further clustered 6,872 unimodal (only one peak of expression or 77%) showed dynamic expression patterns during the beta-cell maturing genes (Fig. 5e,f). Finally, as mentioned previously, cell cycle genes were expressed in the trajectory) genes into 26 groups based on the timing of activation: A1–A14 with peak expression before S4 (pancreatic progenitors) and B1–B12 peaked on or after S4 (Fig. 2b). A GSEA of these gene clusters clearly confirmed the developmental events or physiological functions known to associate with their corresponding differentiation stage (Fig. 2c). For example, hESC gene clusters A2 and A3 enriched the signatures of stem cell and blastocyst growth; S1 (definitive endoderm) gene cluster A6 enriched the signatures of primitive streak formation; and S6 and S7 gene clusters B9–B12 enriched the signatures of pancreatic beta cells and regulation of secretion (Fig. 2c). Interestingly, our analysis highlighted a number of RNA-processing terms enriched in A2–A7 genes (expressed

Fig. 2 | Time- and lineage-specific genes in directed beta-cell differentiation. a. Average expression levels of nine representative stage-specific gene modules are projected onto the lineage tree by colour intensity (see Extended Data Fig. 4 for more modules). ME: meso-endoderm; DE: definitive endoderm; FG: foregut; PL: PSC-like; DL: duct-like; NL: beta-like; AL: alpha-like. b. Heat map showing expression patterns of genes along the differentiation trajectory from hESC to beta-like cell. Top: density plot of cell populations in this trajectory along the pseudo-time. Genes are clustered based on their time of expression. PP: pancreatic progenitors. c. GSEA functional enrichment in each gene group (A1–B12) from numbers in the cells represent the number of genes of each gene group belonging to the corresponding GSEA term. Gene numbers of GSEA terms are in parentheses. Colour intensity indicates the enrichment score (q values were calculated using a one-sided binomial test; exact q values are available in Supplementary Data 3). FGF, fibroblast growth factor. d. Enrichment of T2D/glycaemic risk genes in each gene group from b. Significantly enriched gene groups (one-sided binomial test; P < 0.05) are coloured by the unadjusted P value. The numbers of GWAS genes are indicated in each bar. Maturation genes are described in Fig. 1g. The most significant groups were B1 (P = 2.7 × 10⁻¹⁰), B2 (P = 6.1 × 10⁻¹⁰), B3 (P = 1.3 × 10⁻¹⁰), B6 (P = 1.3 × 10⁻¹⁰), B9 (P = 2.8 × 10⁻¹⁰), B11 (P = 3.9 × 10⁻¹⁰), B12 (P = 2.6 × 10⁻¹⁰) and maturing genes (P = 1.5 × 10⁻¹⁰). e. Expression profiles of many stage-specific TFs are correlated with the abundance of their motifs during differentiation. For each TF, the expression levels in seven stages computed from Drop-seq data are shown (top), as well as the TF motif enrichment among the H3K27ac peaks in the seven stages (bottom). TFs with positive (top) and negative (bottom) correlations are shown (Extended Data Fig. 6).
before definite endoderm), including those relevant to mRNA decay, splicing and transportation (Fig. 2c), suggesting a specific role of RNA-processing machinery in hESCs and early development.

We identified 769 time-dependent transcription factors (TFs) during beta-cell differentiation (Supplementary Fig. 6 and Extended Data Fig. 6). To test the time-dependent functions of these TFs, we collected published enhancer maps (defined by analysis of H3K27ac by chromatin immunoprecipitation followed by sequencing (ChIP-seq)) during the differentiation from hESCs to pancreatic progenitors (hESC and S1–S4)\(^2\), as well as in sorted beta-like cells in S5/6 and S7 stages\(^3\). We next used motif analysis to investigate if the stage-specific TFs may contribute to the dynamic enhancer landscapes (Methods). Among the 85 TF motifs examined, 48 (or 56%) show positive correlations \((r > 0.3)\) between TF expression and the enrichment of their motifs in the enhancer DNA across seven stages, including NANOG in hESC, PITX2 and SMAD2 in definitive endoderm, PAX1 in gut tube, GATA5 in foregut and RFX3, ASCL2 and INS1M in later maturing beta-cell stages (Fig. 2c and Extended Data Fig. 6a,b). Interestingly, we also identified 24 TFs (of 85, or 28%) showing negative correlations \((r < -0.3)\) between TF expression and motif enrichment, among which we noticed several well-characterized transcription repressors, including REST, ZIC2, PRDM1, ZFP42 and HES1 (Fig. 2c and Extended Data Fig. 6). It is therefore plausible that at least some of these negatively correlated TFs may prevent enhancer activation via recruiting co-repressors. Taken together, our results strongly support the conclusion that time-dependent TFs create the dynamic enhancer landscapes to govern differentiation.

We next collected the 562 detectable genome-wide association study (GWAS) risk genes in the GWAS catalogue for type II diabetes (T2D) and glycemic traits, including fasting blood glucose/insulin and the homeostatic model assessment (HMOA)-B (Supplementary Data 3), and examined their dynamic expression during beta-cell differentiation. Importantly, many of these risk genes were only transiently expressed during differentiation, with low or no expression in the beta-like cells from the final stage (Extended Data Fig. 7a). We next examined the enrichment of these GWAS genes in the 26 time-dependent gene groups (Fig. 2b) and found enrichment of the signal in groups B9, B11 and B12, which are expressed in maturing endocrine cells. GWAS genes were also enriched among beta-cell maturation genes (with expression in primary beta cells > beta-like cells; Fig. 2d and Supplementary Data 2 and 3). Interestingly, we also observed significant enrichment of GWAS genes in groups B1–3 and B6, which were transiently expressed in S3–S5 (posterior foregut to pancreatic progenitors; Fig. 2d). Since the GWAS catalogue is a highly heterogeneous resource of risk loci, including many small-size studies among different populations, we investigated the 380 fine-mapped T2D loci (99% credible intervals) reported by a recent large meta-analysis\(^4\), which contains 2,019 detectable genes during beta-cell differentiation (Supplementary Data 3). Meta-analysis gene-set enrichment of variant associations (MAGENTA) analysis\(^5\) also revealed enrichment of risk genes among the transiently expressed gene groups A12, B2 and B6 (Methods and Extended Data Fig. 7b). Our results are consistent with a bulk RNA-seq analysis showing that T2D GWAS loci are enriched in genes in the posterior foregut and the final beta-like stages\(^5\). Notably, some transiently expressed T2D GWAS genes, including TCF7L2, were also associated with glycemic traits (Extended Data Fig. 7); these genes warrant extra attention since the intersection between T2D and glycemic associations is more likely to indicate a mechanism associated with islet dysfunction. Taken together, our results highlight an attractive possibility that many diabetes risk loci may modulate beta-cell differentiation instead of functioning in mature tissues.

Switch genes mark cell fate choice and guide the improvement of differentiation protocol. The lineage tree explicitly defines the cell populations at the ‘branch points’ (BPs) as progenitor cells undergoing lineage determination. Two BPs are particularly important in the beta-cell differentiation: BP1 (cell population S4–C) for the choice between non-endocrine and endocrine lineages (Figs. 1b and 3a) and BP2 (cell population S5–I) for the choice between alpha- and beta-cell fates (Fig. 1b and Extended Data Fig. 8a). For BP1, the differentiation moved towards non-endocrine lineage through the S5–E, S5–C and S6–A populations, and towards the endocrine lineage through the S4–A, S4–E, S5–G and S5–I populations (Figs. 1b and 3a). In the PC space, these cell populations clearly formed two bifurcating trajectories (Fig. 3a), allowing us to conveniently identify ‘switch genes’ that are induced or repressed in one lineage but not the other.

At BP1, we identified 1,150 endocrine-specific switch genes and 795 non-endocrine-specific switch genes (Fig. 3a,b). We posited that most switch genes should maintain their lineage specificity in adult islets, and verified the expression of these switch genes in endocrine (alpha, beta, delta and gamma) or non-endocrine cell types (acinar, duct and PSC) using Drop-seq data from primary human islet cells\(^6\) (Fig. 3b). Indeed, 79% (1,539 of 1,945) of the switch genes showed consistent lineage specificity (Fig. 3b). As expected, gene ontology analysis showed that the endocrine-specific switch genes enriched terms relevant to endocrine functions such as ‘secretion regulation’, ‘exocytosis’ and ‘calcium ion regulation’; non-endocrine-specific switch genes enriched ‘digestive system development’ and ‘epithelial cell differentiation’ terms (Fig. 3c), reflecting a cell fate towards pancreatic acinar and duct cells. The non-endocrine-specific switch genes also enriched several housekeeping terms, including ‘cell cycle’, ‘regulation of microtubule’, ‘RNA splicing’ and ‘ribosomal subunits’ (Fig. 3c), suggesting that the non-endocrine cells proliferate and grow more actively than endocrine cells.

The endocrine-specific switch TFs included nearly all the well-documented endocrine fate regulators, including PDX1, NEUROG3, RFX3, RFX6, NEUROD1, PAX4, MAFB, ASCL1 and INSM1 (Fig. 3d). Several other switch TFs, such as PROX1, SIM1 and MLXIPL, are also supported by literature showing potential roles in endocrine differentiation. In addition, we inferred 656 potential TF-target gene pairs (between 12 TFs and 411 genes) among the endocrine switch genes, and 301 TF–target gene pairs (between 18 TFs and 199 genes) among the non-endocrine switch genes using a TF binding site database (TFBSDB)\(^7\) (Supplementary Data 3 and Supplementary Fig. 7). These data may shed light on the regulatory relationship between switch genes. We also performed the same analyses for BP2 and identified switch genes, including TFs that might contribute to alpha- and beta-cell fate choice (Extended Data Fig. 8). As expected, we observed known alpha-cell markers (for example, IRX2 and ARX) and beta-cell markers (for example, PDX1 and PAX4) distinguishing the two cell fates. It will be interesting to explore the functions of other switch genes in lineage determination and endocrine cell functions.

We also sought to validate the expression pattern of the switch genes in vivo. Despite the difficulty in studying human embryos, several studies have investigated the in vivo beta-cell development in mouse embryos focusing on the time window (embryonic days E13.5 to E18.5) after endocrine progenitors appear\(^8\). We therefore compared our list of human switch genes to a list of mouse switch genes reported by Scavuzzo et al., which contains 716 endocrine-specific and 882 non-endocrine-specific genes\(^9\). A total of 414 switch genes (186 endocrine and 228 non-endocrine) were common between our list and that of mouse data, significantly more than expected \((P < 2.2\times10^{-16})\), Fisher’s exact test; Fig. 3e and Supplementary Data 3). The common switch genes included well-known endocrine (RFX6, PDX1 and NEUROD1) and non-endocrine (HESI1 and ONECUT1) TFs. We reasoned that these consistent genes are more likely to play key roles in the endocrine and non-endocrine lineage choice.
Intriguingly, we noticed a few endocrine and non-endocrine switch genes that were also activated in another stage or BP. For example, the non-endocrine switch gene GATA6 is activated in S1, and indeed is a well-known marker gene for definitive endoderm (Fig. 3f). Another example is TGFB2, which marks the non-endocrine branch in BP1 but is also activated during the final stage of beta-cell differentiation (S6–S7) and, therefore, is also a beta-cell-specific switch gene at BP2 (Fig. 3g). Interestingly, a recent study showed that although inhibiting transforming growth factor (TGF)-β at an earlier pancreatic endocrine stage (S5) helps the differentiation, permitting TGF-β signalling at the final stage is important for beta-like cell maturation, suggesting that omission of TGF-β inhibitor (ALK5i) during S7 can be beneficial to beta-cell maturation. These examples suggest that the same gene or pathway may have time-dependent functions along the differentiation process (multimodal expression is discussed further below).

Finally, we followed up the non-endocrine switch gene HES1, which has a reported role in the inhibition of endocrine lineage as a downstream target of NOTCH signalling. Consistently, all three human NOTCH genes were expressed at higher levels in non-endocrine lineages. HES1 deletion line indeed showed significantly higher endocrine differentiation efficiency in S5 (Fig. 4c). From the lineage tree, we noticed that HES1 expression started in S5 after BP1 in the non-endocrine branch (Fig. 4b). HES1 showed significantly higher endocrine differentiation efficiency in all three human NOTCH genes were expressed at higher levels in non-endocrine lineages. HES1 deletion line indeed showed significantly higher endocrine differentiation efficiency in S5 (Fig. 4c). From the lineage tree, we noticed that HES1 expression started in S5 after BP1 in the non-endocrine branch (Fig. 4b). HES1 showed significantly higher endocrine differentiation efficiency in all three human NOTCH genes were expressed at higher levels in non-endocrine lineages. HES1 deletion line indeed showed significantly higher endocrine differentiation efficiency in S5 (Fig. 4c). From the lineage tree, we noticed that HES1 expression started in S5 after BP1 in the non-endocrine branch (Fig. 4b). HES1 showed significantly higher endocrine differentiation efficiency in all three human NOTCH genes were expressed at higher levels in non-endocrine lineages. HES1 deletion line indeed showed significantly higher endocrine differentiation efficiency in S5 (Fig. 4c). From the lineage tree, we noticed that HES1 expression started in S5 after BP1 in the non-endocrine branch (Fig. 4b). HES1 showed significantly higher endocrine differentiation efficiency in all three human NOTCH genes were expressed at higher levels in non-endocrine lineages. HES1 deletion line indeed showed significantly higher endocrine differentiation efficiency in S5 (Fig. 4c). From the lineage tree, we noticed that HES1 expression started in S5 after BP1 in the non-endocrine branch (Fig. 4b). HES1 showed significantly higher endocrine differentiation efficiency in all three human NOTCH genes were expressed at higher levels in non-endocrine lineages. HES1 deletion line indeed showed significantly higher endocrine differentiation efficiency in S5 (Fig. 4c). From the lineage tree, we noticed that HES1 expression started in S5 after BP1 in the non-endocrine branch (Fig. 4b). HES1 showed significantly higher endocrine differentiation efficiency in all three human NOTCH genes were expressed at higher levels in non-endocrine lineages. HES1 deletion line indeed showed significantly higher endocrine differentiation efficiency in S5 (Fig. 4c). From the lineage tree, we noticed that HES1 expression started in S5 after BP1 in the non-endocrine branch (Fig. 4b). HES1 showed significantly higher endocrine differentiation efficiency in all three human NOTCH genes were expressed at higher levels in non-endocrine lineages. HES1 deletion line indeed showed significantly higher endocrine differentiation efficiency in S5 (Fig. 4c). From the lineage tree, we noticed that HES1 expression started in S5 after BP1 in the non-endocrine branch (Fig. 4b). HES1 showed significantly higher endocrine differentiation efficiency in all three human NOTCH genes were expressed at higher levels in non-endocrine lineages. HES1 deletion line indeed showed significantly higher endocrine differentiation efficiency in S5 (Fig. 4c). From the lineage tree, we noticed that HES1 expression started in S5 after BP1 in the non-endocrine branch (Fig. 4b). HES1 showed significantly higher endocrine differentiation efficiency in all three human NOTCH genes were expressed at higher levels in non-endocrine lineages. HES1 deletion line indeed showed significantly higher endocrine differentiation efficiency in S5 (Fig. 4c). From the lineage tree, we noticed that HES1 expression started in S5 after BP1 in the non-endocrine branch (Fig. 4b). HES1 showed significantly higher endocrine differentiation efficiency in all three human NOTCH genes were expressed at higher levels in non-endocrine lineages. HES1 deletion line indeed showed significantly higher endocrine differentiation efficiency in S5 (Fig. 4c).
the differentiation efficiency. To test this hypothesis, we compared the yield of beta-like cells (measured with C-peptide and NKX6.1) between the three protocols γ-S4, γ-S5 and γ-S6 (standard protocol), which added γ-secretase inhibitor during S4, S5 and S6, respectively (Fig. 4d). As expected, the γ-S5 protocol showed the best results (Fig. 4d,e), consistent with the timing of HES1 expression; γ-secretase inhibition in both S5 and S6 did not provide additional benefits. These results demonstrate that the switch genes can guide the development of an improved differentiation protocol.

**ROCKII inhibition promotes endocrine lineage via suppression of non-endocrine cell fate.** A chemical screen found that inhibition of Rho-associated coiled-coil-containing kinase II (ROCKII) can promote pancreatic beta-cell differentiation\(^{51}\). However, it remains unclear how ROCKII inhibition leads to better differentiation outcomes. We therefore modified the differentiation protocol by adding ROCKII inhibitor (Y-27632) after S4 (Ro.In\(^{−}\) protocol) and performed Drop-seq again in S5–7. Consistent with previous reports, we observed a significantly increased fraction of endocrine cell populations in all stages after ROCKII inhibition, despite the potential biases of using scRNA-seq data to quantify cell-type composition (Fig. 5d and Extended Data Fig. 9a).

We reconstructed the lineage tree with the new S5–S7 Ro.In\(^{−}\) Drop-seq data and compared it to the tree from the standard protocol (Ro.In\(^{+}\); Fig. 5a–c). The new protocol also generated alpha-like, beta-like and EC populations. Due to the improved yield of endocrine cells, we also observed a small cluster of delta-like cells from the Y-27632-treated S7 cells. We compared the expression of 14 marker genes on the two lineage trees and found that they were all turned on at the right time and on the expected branches (Extended Data Fig. 9c). Most dynamic genes in the beta-cell differentiation trajectory maintained the same dynamic pattern in both protocols (Extended Data Fig. 9b). In the PC space, the beta-like cells from both Ro.In\(^{−}\) and Ro.In\(^{+}\) protocols occupied the same area and were distinct from the primary human beta cells (Extended Data Fig. 9d). Furthermore, similar to what we showed in Fig. 1e, alpha-like cells from the new protocol also transcribed excessive levels of INS, while beta-like cells from both protocols transcribed comparable levels of GCG to primary pancreatic beta cells (Extended Data Fig. 9e). Taken together, although ROCKII inhibitor clearly improves the yield of endocrine cells, we did not find evidence of improved endocrine cell quality.

We next examined the effects of ROCKII inhibition of non-endocrine lineage. Under the standard protocol, cell cycle genes are expressed in the non-endocrine lineage but undetected in the endocrine lineage (Fig. 5e). With ROCKII inhibition (Ro.In\(^{+}\)), we noticed a significant down-regulation of cell cycle genes in the non-endocrine cell populations (Fig. 5e), suggesting reduced
Fig. 5 | ROCKII inhibitor promotes endocrine differentiation by suppressing the non-endocrine lineage. a, Unsupervised clustering of single cells for each time point with ROCKII inhibitor (Ro.In⁺) treatment. Cell populations are marked by dashed lines and named in alphabetical order. b, Reconstruction of the lineage tree under the Ro.In⁺ condition. Each dot represents a cell population from a. Cell populations connected by red lines constitute the lineage to beta-like cells. c, A summary of cell numbers in each stage. d, Pie charts comparing the fraction of various cell populations at the final stage between standard and Ro.In⁺ protocols. e, Visualization of average expression of cell cycle genes under the two protocols. f-h, Volcano plots of differentially expressed genes between two protocols in non-endocrine cells in S5–7. Differential expression tests are performed based on negative binomial distribution (Methods). Multiple comparisons are not adjusted. i, Protein interaction network of the top 65 Ro.In⁺ repressed genes. Non-endocrine-specific switch genes are coloured in red. j, Expression profiles of the top 65 Ro.In⁺ repressed genes under the standard protocol. Heat maps comparing the expression of these genes in the non-endocrine and endocrine trajectories after BPI. Drop-seq data showing that nearly all these genes are non-endocrine specific in primary human islets (right). k, Visualization of eight genes that were repressed by ROCKII inhibitor.

cell proliferation in non-endocrine cells. We further systematically identified the differentially expressed genes in non-endocrine cells between Ro.In⁻ and Ro.In⁺ protocols. Surprisingly, in all stages (S5–S7), the top genes repressed by ROCKII inhibition (Ro.In⁺) were ribosomal protein genes (Fig. 5f–i). Interestingly, nearly all of these repressed genes were also non-endocrine-specific switch genes under the standard protocol (Ro.In⁻; Fig. 5j,k). Consistently, we also verified that these genes were only active in non-endocrine cells from primary human islets (Fig. 5l). Since the down-regulation of cell cycle genes and ribosomal proteins suggested reduced cell growth or proliferation, we concluded that ROCKII inhibition promotes endocrine differentiation by suppressing the non-endocrine cell fate. Notably, ROCK2 was expressed at higher levels in the non-endocrine lineage (Extended Data Fig. 9g).

Interestingly, several recent beta-cell differentiation studies already included the ROCKII inhibitor in their differentiation protocols[10,42]. However, these studies added the ROCKII inhibitor (Y-27632) in S3 and S4 and withdrew in S5, in contrast to our protocol. We therefore performed parallel experiments to test various ROCKII inhibition protocols (Extended Data Fig. 9f–h). At the end of S4, we did not detect significant differences in pancreatic endoderm markers (PDX⁺/NKX6.1⁺) with or without ROCKII inhibitor, probably because the percentage of PE cells in all groups was already high (~80%). At the end of S5, all conditions with ROCKII inhibition had more pancreatic endocrine cells (CHGA⁺/NKX2.2⁺), including the condition with ROCKII inhibition during S3–4 (condition 1). At the end of S6, all conditions with ROCKII inhibition (conditions 1–5) still outperformed (more C-peptide⁺/NKX6.1⁺
Over 2,200 genes have multiple waves of activation driven by stage-specific enhancers. From the lineage tree analyses, we noticed several gene modules that were active at more than one differentiation stage or branch (Extended Data Fig. 4), suggesting multiple functions (pleiotropy) during development. We were most interested in genes with multiple waves of activation during the beta-cell differentiation process and, therefore, developed an algorithm to systematically identify them (Methods). Unexpectedly, a large number of genes (2,245 or ~20% of all detectable) showed at least two waves of expression along the beta-cell differentiation trajectory, suggesting extensive gene reuse during development (Fig. 6a and Supplementary Data 5). We termed these genes ‘double-wave’ genes. To suggest the functional relevance of the double-wave genes, we identified 101 double-wave genes that were also endocrine-specific switch genes, that is, these genes were reactivated only in the endocrine lineages during differentiation (Fig. 6c). Further GSEA showed that these double-wave switch genes were enriched with lipid and sterol biosynthetic terms (Fig. 6d), possibly relevant to their roles in regulating insulin secretion43,44.

We next asked if the double-wave expression pattern was encoded in temporal epigenetic dynamics. Again, we used the published H3K27ac ChIP–seq data to define active promoters and enhancers in these stages. Firstly, most double-wave gene promoters maintained similar levels of H3K27ac occupancy across different stages (Extended Data Fig. 10a). However, after classifying the double-wave genes into five groups with expression valleys at stages S1–S5/6, we observed fluctuations of average promoter H3K27ac occupancy consistent with the timing of expression valleys (Fig. 6b and Supplementary Data 5), indicating that the double-wave expression pattern is indeed at transcription level governed by enhancer activity. In line with a recent finding that WNT inactivation promotes endocrine differentiation and demonstrated the value of integrating time-series transcriptomic and epigenomic data in the study of disease aetiology.

Discussion

Current protocols for directed beta-cell differentiation are not perfect. The differentiation results in a heterogeneous cell population; the yield of desired beta-like cells varies between different hESC or hiPSC lines. Here we performed a comprehensive single-cell lineage analysis during the entire directed beta-cell differentiation process starting from hESCs. Compared to two recent single-cell studies in the beta-cell differentiation system6,10, our study used a different seven-stage protocol1, included a full spectrum of differentiation stages, and, importantly, intentionally used a less efficient hESC line to allow better coverage of undesired cell lineages. Consequently, we reconstructed a highly branched lineage tree to explicitly trace the fates of all major cell populations during the entire differentiation process. Notably, although we used a different hESC line with a different protocol, the endocrine lineage in our data contains the same cell population as previously reported, including alpha-like, beta-like and EC-like cells, suggesting that the stem cell-based in vitro differentiation is a well-reproducible system in modeling human pancreatic development1. Our data indicated that the bottleneck of our differentiation system was the decision between endocrine and non-endocrine lineages. It remains unclear whether other hESC or hiPSC lines share the same bottleneck. Further studies are necessary to address the issue about the variation of differentiation propensity between different stem cell lines.

We defined endocrine and non-endocrine switch genes from the lineage tree and demonstrated that these data could reveal the mechanisms of differentiation-promoting chemicals and lead to protocol improvements. Firstly, based on the timing of HES1...
expression, we determined that the best time to start NOTCH inhibition (with γ-secretase inhibitor) is the 3-d window immediately after transferring pancreatic progenitor cells into the air–liquid interface (S5), which is earlier than the standard protocol1. Notably, an alternative six-stage differentiation protocol in suspension adds γ-secretase inhibitor for 7 d between the pancreatic progenitor (PDX1–NKKX6-1+) and endocrine cell (NKKX6-1*–C-peptide*) stages1, which should include the critical transition period between pancreatic progenitors to endocrine precursors as we defined. In another example, we found that ROCKII inhibition increases the proportion of endocrine cells by suppressing the non-endocrine lineage. We found that cell cycle genes are silenced in endocrine cells but remain expressed in the non-endocrine lineage according to the standard protocol, especially in PSC-like cells; ROCKII inhibition significantly down-regulated the cell cycle genes in non-endocrine cells. This is reminiscent of a previous report demonstrating that ROCK inhibition can induce apoptosis and decrease the spreading of HSCs20. We speculate that a similar mechanism might exist in PSC-like cells too. ROCKII inhibition also led to marked down-regulation of many ribosomal protein genes, also suggesting a slow growth. In addition, it is interesting that these ribosomal protein genes are specifically expressed in non-endocrine cells, raising the possibility that ribosome heterogeneity may also contribute to islet differentiation11,12. Lastly, a recent study also showed that depolymerized cytoskeleton leads to endocrine induction13, which may also explain why ROCKII inhibition enhances beta-cell differentiation. Despite these findings, it should be noted that further studies are still necessary to determine if the transcriptional changes are the reasons for or the consequences of non-endocrine lineage suppression.

Most detectable genes are dynamically regulated during differentiation. We showed a significant enrichment of T2D GWAS signal among the genes transiently expressed during pancreatic progenitor stage, highlighting a developmental mechanism affecting the disease
Fig. 7 | A stage-specific enhancer within the TCF7L2 diabetes GWAS locus drives transient gene expression. a, Visualization of TCF7L2 expression on the lineage tree. b, Top: LD at the TCF7L2 locus; arrows indicate the locations of a few T2D GWAS SNPs. The enhancers are highlighted in blue and the promoters are highlighted in yellow. c, Schematic of the primer design validating the deletion of the TCF7L2 E4 enhancer. We expected a 1.16-kb product before deletion (wild-type (WT) band) and a 1.98-kb product after deletion (KO band). d, Representative gel image of validation heterozygous, homozygous and WT clones (repeated three times). e, The effect of E4 enhancer deletion on TCF7L2 expression. Error bars represent the s.e.m. One-sided Student’s t-test was performed. Data are representative of \( n = 4 \) biologically independent experiments. Unadjusted \( P \) values are shown. qPCR, quantitative PCR. f, Flow cytometry comparing the outcomes of differentiating WT and E4 deletion (\( TCF7L2^{E4-/-} \)) cells at the end of S5–7.
propensity. It should be noted, however, that our analyses (including MAGENTA analysis) assumed the one-dimensional proximity between GWAS signal and causal genes; alternative approaches using expression quantitative trait loci or three-dimensional genome data (for example, Hi-C) should improve the identification of distal GWAS target genes when those data are available in the pancreatic developmental system. Most strikingly, our data also revealed that a large percentage of (~20%) all detectable genes have multiple waves of activation. We also showed evidence that different stage-specific enhancers are driving the gene reactivation during development. Interestingly, we discovered that a stage-specific enhancer (E4) located in the TCF7L2 diabetes GWAS locus governs endocrine differentiation through a transient wave of TCF7L2 expression. Taken together, our comprehensive single-cell analysis provides a valuable data resource for the study of islet biology and diabetes.

Methods

Directed differentiation of pancreatic beta cells from hESCs. H1 (WA01) and H9 (WA09) cells were cultured on a Matrigel (Corning, 354277)-coated plate with mTeSR1 medium (Stemcell Technology, 58570) and sub-passaged every 5–7 d. Differentiation was performed following a previous publication. Briefly, H1 cells were dissociated with gentle cell dissociation reagent into singlets. Cells were resuspended in mTeSR1 medium with ROCKII inhibitor (Y-27632; Abcam, ab120129) at a 10–μM concentration and seeded in a Matrigel-coated plate. On the consecutive day, differentiation started with a pancreatic differentiation kit (Stemcell Technology, 5120) from 5–4 following the manufacturer’s manual. At the beginning of 55, cells are dissociated and seeded onto inserts (Fisher Scientific, 877115). From 55–7, cells were cultured in basal medium (MCDB131; Thermofisher, 10372-019) with 1.5 g l−1 sodium bicarbonate (Millipore Sigma, S5761), 1× Glutamax (Thermo Fisher, 35050061), 20 mM glucose (Sigma-Aldrich, G8769 (100 ml)), 2% fatty acid-free BSA (Proliant, 68700) with stage-specific supplements. In S5 (3 d), basal medium was supplemented with 100 nM LDN193189 (Stemgent, 040019), 1× ITS-X, 1× S6 (8 d), basal medium was supplemented with 100 nM LDN193189, 1:200 of heparin (Sigma, H3149). In S7 (7 d), basal medium was supplemented with 100 nM LDN193189, 1:200 ITS-X, 1× T3, 10 μM ALK5 inhibitor II (Enzo Life Sciences, ALX-270-445), 10 μM zinc sulfate (Sigma, 20525) and 10 μg ml−1 of heparin (Sigma, H3149). In S6 (8 d), basal medium was supplemented with 100 nM LDN193189, 1:200 ITS-X, 1× T3, 10 μM ALK5 inhibitor II, 10 μM zinc sulfate and 10 ng ml−1 of heparin (Sigma, A9165), 1 mM N-acetylcysteine (Sigma, A9165), 10 μM T30x (EMD Millipore, 648471), 2 μM R428 (Selleckchem, S2841). For the test with γ-secretase inhibitor, γ-secretase inhibitor was added additionally at 100 nM in S5. For the test with RockII inhibitor (Ro.In+ protocol), RockII inhibitor was added in S5–7 at 10 μM.

Flow cytometry and cell sorting. The hESCs or the differentiated cells were washed once with 1× DPBS and dissociated with Accutase (Innovative Cell Technologies, AT-104) to release single cells, which were stained with fixable viability dye (Thermo Fisher, L34967) and then fixed for staining intracellular or nuclear markers with a staining kit (ebioscience, 00-5523-00). In instances where primary antibody was directly conjugated, the cells were resuspended in 2% BSA in 1× PBS after staining and preparation for flow cytometry analysis; otherwise, cells were subjected to staining with the respective secondary antibody, then washed and resuspended for flow cytometry analysis.

Glucose-stimulated insulin secretion assay. Groups of five to six clusters of S7 H1-derived cells were rinsed twice with Krebs buffer (129 mM NaCl, 4.8 mM KCl, 2.5 mM CaCl2, 1.2 mM MgSO4, 1 mM NaH2PO4, 1.2 mM KH2PO4, 5 mM NaHCO3, 10 mM HEPES and 0.1% BSA) and then pre-incubated in Krebs buffer containing 3.3 mM glucose for 2 h. Cells were subsequently incubated in Krebs buffer containing 3.3 mM glucose for 60 min, washed twice with Krebs buffer and incubated in Krebs buffer containing 16.7 mM glucose for another 60 min; this stimulation of 3.3 mM glucose and 16.7 mM glucose was repeated for two more rounds. Cells were then incubated in Krebs buffer containing 3.3 mM glucose for 60 min, washed again and then incubated in Krebs buffer containing 3.3 mM glucose and 30 mM KCl for 60 min. Supernatant of each stimulation was collected and stored at −70°C for human insulin ELISA (Mercodia, 11-1031-01) measurement.

Quantitative PCR with reverse transcriptase. The mRNA of cells at a desired stage was purified by Zymo Quick-RNA MicroPrep kit (Zymo, R1050), and 1 μg of mRNA was then purified with RNase-free DNase I (NEB, M0303S) and reverse transcribed into cDNA with M-MLV Reverse Transcriptase (Thermo Fisher, 28025013). Quantitative real-time PCR was then performed with PrefCta SYBR Green Supermix Reaction Mixes (Quantabio, 95054-00) on a Bio-Rad qPCR machine. Data were analysed using Bio-Rad CFX Manager software and normalized to undifferentiated H1 cells using the ΔΔCt method. Detailed information on primers is included in Supplementary Table 1.

Deletion of TCF7L2 enhancer E1 in hESCs with single-guide RNAs. To delete the TCF7L2 enhancer E1 in H1 cell line, we made CARGO constructs to introduce two single-guide RNAs (sgRNAs) for each cell. The sgRNAs were gifts from the laboratory of J. Wysocka. All sgRNAs were designed on the CHOPCHOP CRISPR-Cas9 target online predictor (https://chopchop.cbu.uibk.at/). We designed three sgRNAs for the left side and the right side of the targeted enhancer E4 (Fig. 7), respectively. The sgRNA sequences are listed in Supplementary Table 1. H1 cells were transfected using electroporation by Nucleofector 2b (one pulse, 40 V, 230 μF, 500 Ohm, BTX). A total of 5 μl of plasmids (sgRNA expression plasmid + 1:1) were used in each electroporation for 2 million cells. Cells were then cultured for one more day for recovery and then dissociated and digested for single-colony isolation. Cells then grew into viable colonies for about 10 d, and then the colonies were picked into a 24-well plate. Colonies were then expanded and identified by PCR. PCR validation primers are listed in Supplementary Table 1.

HES1 gene knockout in hESCs. We designed two sgRNAs on the first and second exons on the coding region. Two sgRNAs were assembled in tandem on the lentiviruses (sv plasmid (no. 52961) for lentivirus packaging. Around 24 h before transfection, 80 million HEK 293T cells were split into ten 10-cm plates until the cells reached 60% confluence the next day. Each plate was co-transfected with (1) 9 μg lentiviruses E1 plasmid expressing sgRNAs, (2) 9 μg delta V8.91 and (3) 3 μg PMCV-VSVG. For each plate, plasmids and 21 μg polyethylenimine were premixed in 500 μl OptiMEM (Invitrogen) and incubated at room temperature for 10 min. Meanwhile, the culture medium for HEK 293T cells was switched with 1 ml of fresh Opti-MEM. Polyethylenimine–plasmid mixtures were then added to the cells after 10 min of incubation. Six hours after transfection, medium was changed with 10 ml of fresh complete medium. Three days after transfection, culture medium containing viral particles was collected and filtered through 0.45-μm Milipore filters. The viral supernatant was concentrated by centrifugation for 90 min at 25,000 rpm in 4°C. Virus pellets were washed once with ice-cold PBS before resuspension in PBS. The hESC (H1) cells were transduced by concentrated viral solution and cultured for 24 h. Three days after transfection, cells were selected with 1 μg ml−1 puromycin for 2 d. After selection, cells were passed and prepared for differentiation.

Drop-seq library preparation. We collected 600,000 cells from each time point for Drop-seq (12 Drop-seqs at 12 time points in total) encompassing seven developmental stages. We examined more than one time point for early stages because of the quick transcriptome change across time (Extended Data Fig. 1). We performed Drop-seq using the protocol as previously described15,16. Briefly, three pump-controlled syringes with cell suspension (200,000 cells per ml), barcoded beads in lysis buffer (500,000 beads per ml) and droplet generation oil were connected to a microfluidic device under microscope supervision. During droplet generation, we set the cell and bead flow speed at 4,000 μl h−1 and the oil speed at 15,000 μl h−1. The droplets were collected into 50-ml Falcon tubes (usually less than 5 ml). Under this setting, most droplets had at most one bead or one cell. Following droplet barcoding, we performed first-strand cDNA synthesis on beads following the SMART-PCR protocol. Finally, the resulting full-length cDNA library was prepared using Nextera DNA library kit for sequencing.

Drop-seq read processing. We processed raw reads following the instructions described in the original Drop-seq analytical pipeline15. The sequenced Drop-seq libraries yielded 50-bp paired-end reads (PE50). We then extracted bases 1–20 from read 1 for cell and molecular barcodes. Read 2 was trimmed at the 3′ end to remove poly-A tails with at least six bases and trimmed at 5′ if the adaptor sequences appeared. Clean reads were then aligned to hg18 using STAR with default settings. We only retained uniquely mapped reads on gene exons and 3′/5′ untranslated regions appeared. We then filtered out PCR duplicates with the same chromosome positions, cell barcodes and unique molecular identifiers (UMIs). We then grouped the reads by cell barcode and generated the digital UMI-count matrix after accounting transcripts for each gene across all cell barcodes.

Identification of cell populations by unsupervised clustering. Our heterogeneity analysis showed that the cellular variation increased substantially from early to late stages. We reasoned that an unsupervised clustering using cells from all time points together would be less sensitive, especially in identifying cell clusters in early stages because cells from later stages would account for the majority of the variances. Therefore, we did unsupervised clustering for cells from each time point separately and performed a two-step unsupervised clustering, where a low-resolution clustering was firstly used to separate major populations followed by a secondary high-resolution clustering that more sensitively identifies sub-populations.
Before clustering analysis, we set a series of criteria to filter out low-quality cells. Firstly, we filtered out cells in which we observed fewer than 500 transcripts or 200 genes. Secondly, any cells expressing more than 2,500 genes were considered as potential doublets and were also filtered, which removed genes whose expression was too high to be normal. Third, we removed genes whose expression was too low and the mitochondrial genes accounted for more than 5% of the transcriptome. Lastly, we filtered out genes that were expressed by three or fewer than three cells.

We then set rules to pick 'informative genes' for clustering as described in the original Drop-seq study. Briefly, we first ranked the top 10,000 genes based on average expression level among all cells; we then grouped them into ten bins with 1,000 genes each. The coefficient of variation (CV) was calculated for each gene within each bin. From each bin, we picked the top 50 genes with the highest CV as informative genes. Altogether, we picked 500 informative genes for clustering analysis. We used Seurat package for clustering analysis with default parameters. In Seurat, PCA was performed with the 500 informative genes as described above. Using PC1 to PC10, cells were embedded in a k-nearest neighbour graph. The smart local moving algorithm was applied to group cells into communities. PCl to PC10 were used as input to visualize cell clusters in two-dimensional r-distributed stochastic neighbour embedding space.

Human islet Drop-seq processing and data analysis. We previously performed single-cell transcriptome analysis on human islets from nine donors. Detailed procedures were described previously. Briefly, we generated transcriptome data for 39,905 single cells. All samples contained 80–90% endocrine cells, where 90% were alpha or beta cells. Three pieces of information were used for the current study. First, the single-cell transcriptome data of alpha and beta cells were pooled to build the current data and renormalized for comparative analysis with alpha-like and beta-like cells. Second, we previously did a differential expression analysis across four endocrine cell types (alpha, beta, delta and pancreatic progenitor cells), which identified hundreds of beta- and alpha-specific genes. We revisited those genes for the enrichment analysis. Third, for switch gene analysis in primary islet cells, alpha, beta, delta and progenitor time points. We chose 0.25 as the cut-off for log fold change to define cluster-specific gene markers. For cells from two adjacent time points, cells were clustered using Seurat as described above. Using PC1 to PC10, cells were embedded in a k-nearest neighbour graph. The smart local moving algorithm was applied to group cells into communities. PCl to PC10 were used as input to visualize cell clusters in two-dimensional r-distributed stochastic neighbour embedding space.

Identification of molecular signatures. We used the Seurat ‘FindMarker’ function to find marker genes of each cell cluster within each dataset (cells from the same time point were grouped as one dataset). Briefly, the ‘bimod’ test or likelihood-ratio test based on zero inflated data for single-cell gene expression was performed between the tested cell clusters and all the other cells. For any given cell cluster, only genes that were expressed in more than 25% of cells were considered. We then chose 0.25 as the cut-off for log fold change to define cluster-specific gene markers. Cells from S5, S6 and S7 were clearly separated as dual populations (general endocrine and non-endocrine cells). Therefore, we used two-step identification, where a low-resolution analysis was first performed to call gene markers of general endocrine and non-endocrine populations. Second, high-resolution gene markers were identified to distinguish sub-populations within endocrine and non-endocrine populations.

Surface marker genes are highlighted in particular because of the potential application of population enrichment. We integrated the human cell surfaceome from a published database by da Cunha et al. into our signature gene analysis. Any annotated surface gene that showed up as a cell cluster marker gene is highlighted in Supplementary Data 1 and Supplementary Fig. 3.

Benchmarking the similarity between differentiated and primary islet cell populations. To compare a specific cell population in S7 with the primary counterpart in human donors, we combined co-embedding analysis, PCA, signature gene visualization and differential expression analysis. First, to identify the correspondences between the in vitro differentiated populations and the potential counterparts in primary islets, we used Seurat v3 to perform the co-embedding. Briefly, we applied CCA to anchor the shared sub-populations across datasets. As CCA maximizes the correlation between the two datasets, it enabled the robust identification of the relative similarity. Furthermore, to reveal the remaining differences, single-cell data from S7-1 (beta-like cluster), S7-H (alpha-like cluster), islet beta cells and islet alpha-like cells were pooled and renormalized. Highly variable genes were reselected as described and used for PCA analysis. Second, the expression levels of well-known signature genes for each cell type from the literature were then visualized using bubble plots (Fig. 1 and Extended Data Fig. 1). In the plot, for any given gene, the expression levels of all single cells in one population were averaged. Averaged expression levels were then scaled across different populations. The z-scores of scaled average expression levels were then visualized by colour intensity. Similarly, for any given gene, the percentage of cells with detectable expression was also scaled across groups and denoted by dot size. Third, differential expression analysis was performed to identify genes that were amplified or repressed in induced cells. Negative binomial distribution was used in differential expression analysis for models with dispersion in count data. Here, we assumed that, for any gene in each cell, the transcripts number or UMI, could be modelled using negative binomial distribution (equation 1).

\[
\log \text{UMI} = \beta_0 + \beta_1 C + \log(f)
\] (1)

Umi is the expected value of UMI; \(\beta_0\) is the intercept, and \(\beta_1\) is the slope for C and D, respectively; \(C\) represents the cell population, which is a categorical variable; \(D\) is the size factor, which was used to normalize the single-cell transcriptome. It mainly corrects the sequencing depth of each cell (total transcript count of a cell). However, the size factor can be biased due to the dropout zeros and therefore requires further correction. In this study, we calculated \(f\) using the ‘computeSumFactors’ function in the Bioconductor package scran.

GSEA functional enrichment analysis. Gene function enrichment analysis was performed by integrating MSigDB (v.5.2). All functional term lists were read into R using the package ‘gagé’. For any given group of genes, a binomial test (binom. test) was performed iteratively through all annotated functional terms. \(P\) values for enrichment were further adjusted using the qvalue package. Enrichment terms were ranked by \(q\) value. The most representative top terms were selected and visualized using heat maps. On each heat map, \(q\) values of enriched terms were visualized by colour intensity as enrichment scores.

Reconstruction of the lineage tree. To reconstruct the lineage tree for a time-course-directed differentiation process, we developed an algorithm that adopts a bottom-up strategy from ‘sub-transition network’, that is, a graph between two adjacent time points. We first isolated single-cell data for a pair of given adjacent data points along the whole differentiation process from hESC to insulin-producing cells (S7; Extended Data Fig. 1). The overall assumption is that most cells from one time point are supposed to either develop into the cells at the same or next adjacent time point or be terminated by apoptosis for instance. We assumed it was less possible for cells from one time point to randomly jump into cells at a far-away time point. Therefore, we took advantage of the timeline information to constrain the possible networks into an experimental related time frame. We first built the relationships for cells from two adjacent time points (sub-transition network). To be exact, there were 11 data pairs for sub-transition-network reconstruction, including H1–S0, S0–S1–D1, S1–D1–S2, D1–S2–D2, S2–D2–S3, S2–D3–S3, S3–S4, S4–S5, S5–S6 and S6–S7. The entire lineage tree was then reconstructed by stacking together all sub-transition networks.

To build each sub-transition network, we applied a cluster-based connectivity analysis with an in-house decision-making algorithm (Supplementary Fig. 4). Another recent study has also used a similar approach to infer lineage trees. First, we isolated single-cell data for a pair of given adjacent data points. For each of the two time points, cells were clustered using Seurat as described above. With the cluster identity, cells were then pooled together and renormalized. The high-dimensional transcripts space was then reduced into a ten-dimensional PC space. The centroid of each cluster was calculated in ten-dimensional PC space. The Euclidean distance between any two clusters was calculated as pairwise cell-type specificity of any gene is provided by the function as the significance of \(\lambda\). The \(P\) values of all genes were further adjusted with Bioconductor package ‘qvalue’ to obtain \(q\) values. We also corrected for the log fold change and applied the average transcript counts. Differentially expressed genes (\(q\) value <0.05) were ranked by fold change. The statistics of differentially expressed genes are included in Supplementary Data 2.

GSEA functional enrichment analysis. Gene function enrichment analysis was performed by integrating MSigDB (v.5.2). All functional term lists were read into R using the package ‘gagé’. For any given group of genes, a binomial test (binom. test) was performed iteratively through all annotated functional terms. \(P\) values for enrichment were further adjusted using the qvalue package. Enrichment terms were ranked by \(q\) value. The most representative top terms were selected and visualized using heat maps. On each heat map, \(q\) values of enriched terms were visualized by colour intensity as enrichment scores.

Reconstruction of the lineage tree. To reconstruct the lineage tree for a time-course-directed differentiation process, we developed an algorithm that adopts a bottom-up strategy from ‘sub-transition network’, that is, a graph between two adjacent time points. We first isolated single-cell data for a pair of given adjacent data points along the whole differentiation process from hESC to insulin-producing cells (S7; Extended Data Fig. 1). The overall assumption is that most cells from one time point are supposed to either develop into the cells at the same or next adjacent time point or be terminated by apoptosis for instance. We assumed it was less possible for cells from one time point to randomly jump into cells at a far-away time point. Therefore, we took advantage of the timeline information to constrain the possible networks into an experimental related time frame. We first built the relationships for cells from two adjacent time points (sub-transition network). To be exact, there were 11 data pairs for sub-transition-network reconstruction, including H1–S0, S0–S1–D1, S1–D1–S2, D1–S2–D2, S2–D2–S3, S2–D3–S3, S3–S4, S4–S5, S5–S6 and S6–S7. The entire lineage tree was then reconstructed by stacking together all sub-transition networks.

To build each sub-transition network, we applied a cluster-based connectivity analysis with an in-house decision-making algorithm (Supplementary Fig. 4). Another recent study has also used a similar approach to infer lineage trees. First, we isolated single-cell data for a pair of given adjacent data points. For each of the two time points, cells were clustered using Seurat as described above. With the cluster identity, cells were then pooled together and renormalized. The high-dimensional transcripts space was then reduced into a ten-dimensional PC space. The centroid of each cluster was calculated in ten-dimensional PC space. The Euclidean distance between any two clusters was calculated as pairwise cell-type specificity of any gene is provided by the function as the significance of \(\lambda\). The \(P\) values of all genes were further adjusted with Bioconductor package ‘qvalue’ to obtain \(q\) values. We also corrected for the log fold change and applied the average transcript counts. Differentially expressed genes (\(q\) value <0.05) were ranked by fold change. The statistics of differentially expressed genes are included in Supplementary Data 2.
reconstructed a pseudo-time trajectory between two connected nodes. We pooled the cells from all nodes, followed by normalization and dimension reduction by PCA. On the ten-dimensional PC space, the centroid of node \( r \) and node \( \lambda' \) were calculated. Two nodes were connected by a directed edge if all the single cells from nodes \( r \) and \( \lambda' \) were projected onto \( \mathbf{t} \). The relative position on the \( \mathbf{t} \) was taken as ‘relative pseudo-time’. We did that iteratively for all the connected nodes and calculated relative pseudo-times. Finally, by stacking all the nodes together, we calibrated the relative pseudo-time into the full graph \( G \). After projecting all cells on all the edges of graph \( G \), we divided the full graph \( G \) into 1,475 bins. The number of bins on any edge was determined by the distance on PC space. The longer the distance, the more bins would be taken. Each bin contained at least 50 cells. Next, graph \( G \) was visualized as in Fig. 2a. The distance on the \( y \) axis indicates the relative distance on PC space between two nodes; the \( x \) axis is only used to distinguish different nodes. For any given gene, the average expression level was normalized by the total expression in the bin. The normalized average expression level of each gene was then visualized by heat map colouring on the \( G \) map.

Classification of gene modules on the lineage tree. Among all 17,978 detected genes, we first selected a training gene set for gene module clustering. For any given gene, we averaged the top 20 highest expression bins. This value was used to filter out the genes that were never highly expressed in any branch at any time. This resulted in 11,838 genes that showed high expression at a certain time point. We then took the top 2,000 highly variable genes (of 11,838 genes) based on the CV of their time series. We then used the first calculated correlation between any two genes across all 1,475 bins on the G map. With the dimension matrix, we further performed \( k \)-medoid clustering to partition 2,000 genes into 64 gene modules. We chose 64 modules because our analysis called a total of 56 cell populations across all time points (Fig. 1a), and we reasoned that the number of gene modules would be larger than the number of cell populations. We initially tested 32, 64 and 128 clusters and determined that 64 appeared to be the preferred choice based on the elbow method.61 With the predefined gene modules by the training set, we further mapped the rest of the genes (9,883) using the \( k \)-nearest neighbour (KNN, \( k = 5 \) ) algorithm. Finally, we averaged the genes in the same module for each bin and manually ordered the 64 gene modules based on the apparent enrichment level (Fig. 2a and Extended Data Fig. 4). GSEA was performed for each gene module.

Trajectory analysis of beta-specific lineage. To specifically characterize the beta-cell differentiation, we pooled the single cells from the clusters on the main lineage highlighted in Fig. 1b and performed renormalization followed by PCA dimension reduction. On the ten-dimensional PC space, we reconstructed the full pseudo-time trajectory by only connecting all the nodes on the main lineage. We further divided the full main lineage into 80 bins, and, for any given gene, the averaged expression level was normalized by the total expression in each bin. Considering the potential non-linear nature of the time-dependent gene expression changes, we used polynomial regression to model the gene expression dynamic across the 80 bins. The \( P \) value of the goodness of fit for each gene was calculated and adjusted using the \( q \) value package. Additionally, we further aggregated the 80 bins into ten big bins to estimate the expression fold change across time. We measured the fold change between the highest big bin versus lowest big bin. We firstly filtered out only those genes whose average expression in the highest bin was lower than 0.01 (that is, expression that can be detected in less than 1% of cells even within the most highly expressed bin). Then, we identified time-dependent dynamic genes that fitted the regression model well (\( q \) value <0.01) and also showed more than a twofold expression change between the lowest and highest big bin (Fig. 2 and Extended Data Fig. 5). The rest of the genes were classified as constant or stochastically expressed genes, which do not show time-dependent expression change. To further stratify the timing of the time-dependent dynamic genes, we firstly selected the top 3,000 genes with the lowest \( q \) value of model fitness for a hierarchical clustering, which resulted in group A and B. Group A included genes that are generally up-regulated across time, whereas B includes down-regulated genes. In group A and B, we did a secondary hierarchical clustering that eventually resulted in gene groups A1–A14 and B1–B12, respectively. Next, we mapped back the remaining genes into each group by finding the highest correlation. Using genes in each group, we then performed GSEA as described above. We also compared the results from our trajectory analysis to those using Monocle 3 (ref. 7) and Extended Data Fig. 3). We found that Monocle 3 could not properly recognize the entire time series of our data. For example, the cell identity undergoes drastic changes from S0 (hESC) to S1 (definite endoderm) and S2 (gut tube). The correlation between S1_D2 and S2_D1 was low because many definite endoderm marker genes are activated in S1 (including LEFTY1, NODAL and GCE4) and quickly turned off in S2. On the other hand, since the S1-specific genes were not expressed in both S0 and S2, the correlation between S0 and S2_D1 was higher. For this reason, Monocle 3 (ref. 7) called one branch connecting S0 to S1 but incorrectly called another branch from S0 directly to S2_D1 cells.

T2D/glycaemic GWAS genes in the major beta-cell differentiation trajectory. Among the time-dependent dynamic genes, we particularly interrogated the reported genes associated with T2D and glycaemic traits. We collected 698 genes reported from the GWAS catalogue with the traits: ‘T2D’, ‘fasting blood glucose’, ‘fasting blood glucose (body mass index (BMI) interaction)’, ‘fasting blood glucose adjusted for BMI’, ‘fasting insulin (BMI interaction)’, ‘fasting blood insulin adjusted for BMI’, ‘fasting blood insulin (BMI interaction)’, ‘fasting blood insulin adjusted for BMI’, ‘fasting blood proinsulin levels’, ‘fasting insulin (dietary factor interaction)’ and ‘HMOA-B’ (Supplementary Data 3). These genes were used for the analysis in Fig. 2d. Using all detected 17,978 genes as background, we performed a binomial test for the enrichment of diabetes or obesity GW AS genes in each branch group. In Fig. 2d, gene groups with a \( P \) value >0.05 show the colour white, whereas gene groups with a \( P \) value <0.05 are coloured by the enrichment value \( P \).

We also performed the MAGENTA analysis6 using the fine-mapping results from a large meta-analysis of T2D GWAS studies6 (Extended Data Fig. 7). Briefly, for every gene in the genome, we assigned a set of SNPs laying within 110 kb upstream and 40 kb downstream of the gene’s most extreme transcription boundaries. Each gene was then assigned a score, defined as the most significant posterior probability among the SNPs within the extended gene boundaries. The top 10% of genes with the highest score were used for enrichment analysis among gene groups (A1–B12; Fig. 2b). The significance of enrichment of each gene group was measured for gene modules and used for a combinatory analysis with available 50–54 H3K27ac ChIP–seq data. Starting from those genes, we then selected the ones with an available D1 by the HOCOMOCO v2 tool and finally obtained that 60% of the 128 ChIP–seq data points overlapped with the published H3K27ac ChIP–seq data11,62 and performed peak calling with MACS2 software. For each investigated TF motif, we scanned on H3K27ac ChIP–seq peaks of each stage by the motif matrix using FIMO (http://meme-suite.org/doc/fimo.html)65. A significantly enriched motif in a given peak was defined by a \( q \) value <0.01. For each TF, the number of enriched motifs across the entire genome was then normalized by the number of peaks as ‘motif density’ (Fig. 2). The single-cell expression level was then aggregated into each stage as pseudo-bulk RNA expression data for S0, S1, S2, S3 and S4, corresponding to each stage in H3K27ac data. Finally, we calculated the Pearson’s correlation (\( r \)) between RNA expression and motif density. TFs with \( r \) were defined as positively correlated, \( r \) as negatively correlated, and \( r \) as undetermined.

Identification of switch-like genes associated with bifurcating cell fate choices. To examine the lineage bifurcation from the BP, we first pooled the involved clusters together, including S4_C, S4_A, S4_E, S5_E, S5_C, S5_S, S5_S1 and S6_A, and performed renormalization followed by dimension reduction using PCA. Next, we reconstructed the pseudo-time trajectory for endocrine lineage (S4_C, S4_A, S4_E, S5_G and S5_I) and non-endocrine lineage (S4_C, S5_E, S5_C and S6_A). For endocrine and non-endocrine pseudo-trajectories, we further divided the lines into two trajectories from the branch points. For any given gene, we measured the average fold change along two lines and invoked a paired Wilcoxon rank-sum test. We defined a lineage switch-like gene when it met the following criteria: (1) an average fold change more than two and (2) a paired Wilcoxon rank-sum test adjusted \( P \) value <0.01. According to this rationale, we identified both non-endocrine-specific (or non-endocrine lineage switch-like) genes and as endocrine-specific (or endocrine lineage switch-like) genes, as well as endocrine-specific (or endocrine lineage switch-like) genes, as well as endocrine-specific (or endocrine lineage switch-like) genes, as well as endocrine-specific (or endocrine lineage switch-like) genes, as well as endocrine-specific (or endocrine lineage switch-like) genes, as well as endocrine-specific (or endocrine lineage switch-like) genes, as well as endocrine-specific (or endocrine lineage switch-like) genes, as well as endocrine-specific (or endocrine lineage switch-like) genes, as well as endocrine-specific (or endocrine lineage switch-like) genes, as well as endocrine-specific (or endocrine lineage switch-like) genes, as well as endocrine-specific (or endocrine lineage switch-like) genes, as well as endocrine-specific (or endocrine lineage switch-like) genes, as well as endocrine-specific (or endocrine lineage switch-like) genes, as well as endocrine-specific (or endocrine lineage switch-like) genes, as well as endocrine-specific (or endocrine lineage switch-like) genes.
Rezeania, A. et al. Maturation of human embryonic stem cell-derived pancreatic progenitors into functional islets capable of treating pre-existing diabetes in mice. *Diabetes* **61**, 2016–2029 (2012).

Riedel, M. J. et al. Immunohistochemical characterisation of cells co-producing insulin and glucagon in the developing human pancreas. *Diabetologia* **55**, 372–381 (2012).

Hu He, K. H. et al. In vivo conditional Pax4 overexpression in mature islet beta cells prevents stress-induced hyperglycemia in mice. *Diabetes* **60**, 1705–1715 (2011).

Saendens, W., Cannoodt, R., Todover, H. & Saeya, Y. A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **37**, 547–554 (2019).

Cao, J. et al. The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496–502 (2019).

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

Baron, M. et al. A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure. *Cell Syst.* **3**, 346–360 e344 (2016).

Segerstolpe, A. et al. Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. *Cell Metab.* **24**, 593–607 (2016).

Lawlor, N. et al. Single-cell transcriptomes identify human islet cell signatures and reveal cell-type-specific expression changes in type 2 diabetes. *Genome Res.* **27**, 208–222 (2017).

The, J. et al. Single-cell chromatin accessibility reveals pancreatic cell type- and state-specific regulatory programs of diabetes risk. Preprint at bioRxiv https://doi.org/10.1101/693671 (2019).

Chen, Q. & Hu, G. Post-transcriptional regulation of the pluripotent state. *Curr. Opin. Genet. Dev.* **46**, 15–23 (2017).

Wang, A. et al. Epigenetic priming of enhancers predicts developmental competence of hESC-derived endodermal lineage intermediates. *Cell Stem Cell* **16**, 386–399 (2015).

Alvarez-Dominguez, J. R. et al. Circadian entrainment triggers maturation of human in vitro islets. *Cell Stem Cell* **26**, 108–120 (2020).

Mahajan, A. et al. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat. Genet.* **50**, 1505–1513 (2018).

Segre, A. V. et al. Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. *PLoS Genet.* **6**, e1001058 (2010).

Russo-Alcantara, M. et al. Patterns of differential gene expression in a cellular model of human islet development, and relationship to type 2 diabetes predisposition. *Diabetologia* **61**, 1614–1622 (2018).

Plaisier, C. L. et al. Causal mechanistic regulatory network for glioblastoma deciphered using systems genetics network analysis. *Cell Syst.* **3**, 172–186 (2016).

van Cuijper, I. L. et al. A transcriptomic roadmap to alpha- and beta-cell differentiation in the embryonic pancreas. *Development* **146**, dev173716 (2019).

Krentz, N. A. J. et al. Single-cell transcriptome profiling of mouse and hESC-derived pancreatic progenitors. *Cell Stem Cell* **11**, 1551–1564 (2018).

Byrnes, L. E. et al. Lineage dynamics of murine pancreatic development at single-cell resolution. *Nat. Commun.* **9**, 3922 (2018).

Scavuzzo, M. A. et al. Endocrine lineage biases arise in temporally distinct endocrine progenitors during pancreatic morphogenesis. *Nat. Commun.* **9**, 3556 (2018).

Velazco-Cruz, L. et al. Acquisition of dynamic function in human stem cell-derived beta cells. *Cell Stem Cell* **12**, 351–365 (2019).

Murtagh, L. C., Stanger, B. Z., Kwan, K. M. & Melton, D. A. Notch signalling controls multiple steps of pancreatic differentiation. *Proc. Natl Acad. Sci. USA* **100**, 14920–14925 (2003).

Alepisht, A. et al. Notch signalling controls pancreatic cell differentiation. *Nat. Rev. Endocrinology* **8**, 139–150 (2012).

Jensen, I. et al. Control of endodermal endocrine development by Hes-1. *Nat. Genet.* **24**, 36–44 (2000).

Ghazizadeh, Z. et al. ROCKII inhibition promotes the maturation of human pancreatic beta-like cells. *Nat. Commun.* **8**, 298 (2017).

Hogrebe, N. I., Augoornworawat, P., Maxwell, K. G., Velazco-Cruz, L. & Milliman, I. R. Targeting the cytoskeleton to direct pancreatic differentiation of human pluripotent stem cells. *Nat. Biotechnol.* **38**, 460–470 (2020).

Tschiu, M. et al. Cholesterol biosynthesis pathway intermediates and inhibitors regulate glucose-stimulated insulin secretion and secretory granule formation in pancreatic beta cells. *Endocrinology* **151**, 4705–4716 (2010).

Xia, F. et al. Inhibition of cholesterol biosynthesis impairs insulin secretion and voltage-gated calcium channel function in pancreatic beta cells. *Endocrinology* **149**, 5136–5145 (2008).

Bailey, K. A. et al. Evidence of non-pancreatic beta cell-dependent roles of Tfnf2 in the regulation of glucose metabolism in mice. *Hum. Mol. Genet.* **24**, 1646–1654 (2015).
Articles

64. Zhang, Y. et al. Model-based analysis of ChIP-seq (MACS). *Genome Biol.* 9, R137 (2008).
65. Bailey, T. L. et al. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* 37, W202–W208 (2009).
66. MacArthur, J. et al. The new NHGRI-EBI catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* 45, D896–D901 (2017).
67. Wilson, D., Charoensawan, V., Kummerfeld, S. K. & Teichmann, S. A. DBD—taxonomically broad transcription factor predictions: new content and functionality. *Nucleic Acids Res.* 36, D88–D92 (2008).

Acknowledgements

This work was supported by grants from the National Institutes of Health (NIH; R01DK113185 to Y.L. and R01HG009658 to F.J.) and Mt. Sinai Health Care Foundation (OSA510114 to Y.L. and OSA510113 to F.J.), and a pilot award from the Clinical and Translational Science Collaborative at Case Western Reserve University (VSN639001 to Y.L.). A.G. is supported by a NIH MSTP training grant (T32 GM007250) and a Functional Genomics Training Program grant (T32 GM135081). We also gratefully acknowledge the insightful discussions with M. Scavuzzo, X. Liu, S. Zhang, L. Lu and L. Wang.

Author contributions

Y.L. and F.J. conceived the project. C.W., J.X. and H.L. performed the experiments. J.C., S.L., A.G., K.L. and L.K. contributed to the experiments. C.W. and J.X. carried out the data analysis. C.W., X.I., F.J. and Y.L. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s42255-020-00314-2.
Supplementary information is available for this paper at https://doi.org/10.1038/s42255-020-00314-2.
Correspondence and requests for materials should be addressed to F.J. or Y.L.
Peer review information Primary Handling Editors: Elena Bellafante; Christoph Schmitt.
Reprints and permissions information is available at www.nature.com/reprints.
Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020
Extended Data Fig. 1 | Schematics of experimental design and benchmark of differentiated cell maturity. a, On the top is the timeline of the directed beta-cell differentiation process that takes 31 days encompassing 7 stages. Each stage recapitulates an in vivo developmental stage respectively. The key chemicals added at different time points are listed. Cells from each of the 12 time points are collected for Drop-seq experiments. For instance, S1_D1 stands for stage 1 day 1. b, Glucose stimulated insulin secretion (GSIS) assay for differentiated Stage7 cells. Three rounds of glucose challenges were performed before final treatment with KCl (detail in Methods). n=3 biologically independent replicates. Error bars show +/−1 standard error of mean (SEM). One-sided Student’s t-test was performed. *** indicates p-value < 0.01. The exact p-values (multiple test unadjusted) for 1st, 2nd and 3rd glucose challenge are 1.4*10^-4; 2.8*10^-3 and 3.2*10^-4 respectively. c, CCA-based co-embedding analysis between the differentiated cells and the primary islet dataset. d, Cell type specific marker genes visualized on co-embedded UMAP. Color intensity indicates relative expression level (Z score). e, Bubble plot of endocrine and non-endocrine signature genes in different populations (Extended for Fig. 1f). Color intensity indicates the Z score of averaged expression levels; bubble size indicates the percentage of expressing cells.
Extended Data Fig. 2 | See next page for caption.
Extended Data Fig. 2 | Unsupervised clustering of sub-populations from Stage 4 to Stage 7. a, Two step clustering for S4, S5, S6, S7 respectively. For each stage, a low-resolution clustering for all cells is shown on the left, where two big populations, endocrine (green) and non-endocrine (grey), are clearly separated. On the right are the secondary high-resolution sub-clusters for non-endocrine (right top of each stage) and endocrine population (right bottom of each stage) respectively. Below the tSNEs is the endocrine/non-endocrine population summary for each stage (b) Endocrine marker genes examination (NEUROD1, CPE and CHGA) on first low-resolution tSNE map for each stage. Color intensity indicates relative expression level (z score). c, Summary of the sub-populations of stage 7. d-g, Marker gene examinations on stage 7 endocrine tSNE plot. Color intensity indicates relative expression level (z score) with INS for beta-like population (d); GCG for alpha-like population (e); EC-like population markers (f); and proliferation markers (g).
Extended Data Fig. 3 | See next page for caption.
Extended Data Fig. 3 | Comparison of trajectory inference algorithms. a-b. Global lineage inference connecting cells from all 12 time points in the same UMAP embedding, using the method in this manuscript in (a) and using Monocle3 in (b). Note that starting from H1/S0, Monocle3 created a branch directly to Stage 2 Day1 (S2_D1), bypassing the time points in Stage 1. c. Differentially expressed genes among H1(ES), S1_D2 and S2_D1. Color intensity indicates the expression level. TPM: transcripts per million UMI. d-g. Our method and Monocle3 inferred the consistent local differentiation trajectories. Local trajectory inference of S4 and S5 starting from Branch point 1 (BP1), using our method is shown in (d) and using Monocle3 in (e). Local trajectory inference of alpha-like and beta-like lineages starting from branch point 2 (BP2), using our method is shown in (f) and using Monocle3 in (g). h, Side-by-side comparisons of switch genes at BP1 in Fig. 3b between on our trajectories (Left panels) and on Monocle3 trajectories (Right panels). i, Similar to (g), side-by-side comparison between our method and Monocle3 for BP2 switch genes reported in Extended Data Fig. 8b.)
Extended Data Fig. 4 | See next page for caption.
Extended Data Fig. 4 | Lineage specific gene modules on full developmental tree. a, 64 gene modules extended for Fig. 2a. Each gene module contains a couple of hundreds of genes that share a similar expression pattern of time & lineage specificity on the full developmental tree. Three representative genes in each module are shown. For full list see Supplementary Data 3. ME: meso-endoderm; DE: definitive endoderm; FG: foregut. PL: PSC-like; DL: Duct-like; EC: enterochromaffin cells; βL: beta-like; αL: α-like. b, Expression of islet cell type specific marker genes are illustrated on the full tree. The color intensity indicates the relative expression level. INS: beta cell marker; GCG: alpha cell marker; SST: delta cell marker; PPY: gamma cell marker and GHRL: epsilon cell marker (c) Expression of representative cell cycle genes are illustrated on the full tree. Color intensity indicates the relative expression level.
Extended Data Fig. 5 | See next page for caption.
Extended Data Fig. 5 | Examples of time dependent dynamic genes on β-like specific lineage. a. Characterization of genes along β-like specific lineage highlighted in Fig. 1b. Left pie-chart: Summary of genes in each category. Out of all the genes, 11,876 genes with detectable expression level. 9,117 genes show time dependent expression pattern. Right histogram: histogram showing the number of genes with different variance across time. Dashed line indicates 2-fold relative expression change along trajectory. b-f. Show the examples of time dependent dynamic genes from ES cell to beta cell lineage. b, shows the cell cycle genes. c, shows the developmental genes. (d) shows classic beta cell signature genes. e, f, shows calcium channel and sodium channel genes respectively. The color intensity indicates the relative expression level (z score).
Extended Data Fig. 6 | See next page for caption.
Extended Data Fig. 6 | Time dependent TF motif analysis on β-like specific lineage. a, Summary of the motif analysis flow. 769 transcription factors show time dependent expression pattern. 195 TFs out of those have reported DNA binding motif matrix that can be used for motif enrichment analysis. 85 out of 195 shows motif enrichment at least in one of the 7 stage H3K27ac peaks genome wide. For these 85 TFs, the single cell expression level was aggregated into each stage as a pseudo-bulk RNA expression data; TF motif density among enhancers from each stage of H3K27ac can be then correlated to the TF expression. We defined positive correlation when \( r > 0.3 \) (Pearson’s correlation coefficient), and defined negative correlation when \( r < -0.3 \). 48 TFs show positive correlation between expression pattern and motif density pattern. 24 TFs show negative correlation. 13 TFs with no significant correlation. b. The details of all correlated 72 TFs. For each TF, Top panel shows the relative expression change along the pseudo-time computed from single-cell RNA data. Bottom shows the TF motif enrichment on genome-wide H3K27ac peaks across 7 stages. Stage labels are shown in the middle, shared by top and bottom panels. Green: positive correlation; Red: Negative correlation.
Extended Data Fig. 7 | See next page for caption.
Extended Data Fig. 7 | Time dependent dynamic GWAS genes on β-like specific lineage. a, Heatmap showing top time dependent dynamic T2D/glycemic GWAS genes from ES cell to beta cell lineage highlighted in Fig. 1b. The color intensity indicates the relative expression level. The left color bar indicates the traits associated with a certain gene. Gene loci associated with both T2D and glycemic traits are highlighted in bold with asterisk. On the right: bubble plot showing the expression level in human primary beta cells. The size of the bubble indicates the percentage of cells with detectable expression. Bubble color indicates the averaged expression level. Full list of dynamic T2D/glycemic GWAS genes in Supplementary Data 3. b, MAGENTA enrichment for stage-specific gene groups (defined in Fig. 2b) by T2D GWAS 99% credible variants. The significance of enrichment of each gene group was measured by nominal p-value after randomly selecting the same number of genes in the genome. (For more details, see Methods). Multiple tests are adjusted by qvalue.
Extended Data Fig. 8 | Molecular profiling of lineage specification between α-like and β-like lineages. a, PCA analysis of pooled populations for the bifurcation of β-like and α-like lineages, including S5-I, S6-E, S7-H for α-like and S5-I, S6-D, S7-I for β-like, respectively. α-like in yellow, β-like in green, cells of BP2 (S5I) is in grey. b, Heatmap showing β-like vs α-like lineage switch genes. 282 genes are specifically up-regulated in β-like lineage. 58 genes are specifically up-regulated in α-like lineage. c, Fold change between primary β vs. primary α cells computed from human islet Drop-Seq for each gene in (b). d, GSEA functional enrichment in β-like and α-like lineage genes respectively. Color intensity indicates enrichment score (One-sided binomial test, adjusted q value). Gene numbers of all enriched functional terms are shown. The exact q values for the presented 8 terms (top-down) are: 2.6*10^{-2}, 1.1*10^{-15}, 9.6*10^{-9}, 3.8*10^{-9}, 7*10^{-5}, 4.9*10^{-5}, 4.2*10^{-5}. e, The β-like and α-like lineage switch transcription factors. f, Fold change between primary β vs. primary α cells computed from human islet Drop-Seq for each lineage transcription factor in (e).
Extended Data Fig. 9  |  Assessing the effects of ROCKII inhibition on the endocrine differentiation and protocol optimization.  

**a** The tSNE plots of S5 - S7 cells differentiated with or without ROCKII inhibitor (Ro.In). Green: endocrine cells; grey: non-endocrine cells. 

**b** Heatmap of 6872 unimodal time-dependent genes on ES to beta-like lineage on two protocols. 

**c** Marker genes visualized on lineage tree of both standard and Ro.In+ protocol. The expression patterns of marker genes are consistent in two protocols. 

- **NEUROG3** for endocrine precursors; **CHGA**, **CHGB**, **CPE**, **PAX6**, **NEUROD1** for endocrine cells. 
- **FOXA2**, **PDX1**, **INSM1**, **INS** for β-like cells. 
- **GCG**, **ARX** for α-like cells. 
- **SST**, **HHEX** for delta-like cells. 

**d** PCA analysis for three populations: primary β cell (Grey); β-like cell in standard protocol(S7-1); β-like cell in Ro-in+ protocol(S7-R-F). 

**e** Hormone gene transcripts proportion. Top panel is insulin transcripts proportion across 6 populations: islet β; β-like in standard protocol; β-like in Ro-in+ protocol; islet α; α-like in standard protocol; α-like in Ro-in+ protocol. Bottom panel is glucagon proportion across those 6 cell populations. (n= 3953, 421, 1676, 5700, 732, 655 single cells for islet-β, β-like, β-like-Ro.In+, islet-α, α-like, α-like-Ro.In+ respectively). Boxplots are shown within the violin plots, where the middle line is the median, the lower and upper hinges correspond to the first and third quartiles, the upper whisker extends from the hinge to the largest value no further than 1.5xIQR from the hinge, the lower whisker extends from the hinge to the smallest value at most 1.5xIQR of the hinge (IQR is the inter-quartile range). 

**f** Schematics of six differentiation protocols with or without ROCKII inhibition at different stages. 

**g** Visualization of ROCK2 expression in the lineage tree. 

**h** Flow cytometry data showing the yield of pancreatic progenitors (NKX6.1+/PDX1+), endocrine cells (NKX2.2+/CHGA+) and β-like cells (C-peptide+/NKX6.1+) at the end of stage 4, Stage 5, and Stage 6 with different protocols in (f). The standard protocol without ROCKII inhibition (condition 6) was included as control.
Extended Data Fig. 10 | H3K27ac ChIP-seq on the promoters and distal enhancers of double wave genes. 

a. Density heatmap displaying the H3K27ac on TSSs of double wave genes across 7 stages. ChIP-seq signal is plotted as reads per million per base pair in TSS ± 5kb region.

b. Density heatmap displaying the H3K27ac on putative distal enhancers within 250kb from TSSs of double wave genes across 7 stages. ChIP-seq signal is plotted as reads per million per base pair in enhancer peak center ± 5kb region.

c. Eight T2D GWAS fine-mapping signals at TCF7L2 loci reported by Mahajan et al. Black vertical lines in the top tracks are 99% credible variants in each signal; red lines are the leading SNPs.
Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a

- Confirmed

☐ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement

☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly

☐ The statistical test(s) used AND whether they are one- or two-sided

☐ Only common tests should be described solely by name; describe more complex techniques in the Methods section.

☐ A description of all covariates tested

☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons

☐ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)

☐ For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted Give P values as exact values whenever suitable.

☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings

☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes

☐ Estimates of effect sizes (e.g. Cohen’s d, Pearson’s r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection: Flow.io software

Data analysis: https://github.com/l1nlabBioinfo/LineageTreeBuilder

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. Github). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Raw and processed Drop-Seq data are accessible at GEO: GSE143783.

GWAS data were downloaded from the GWAS catalog (https://www.ebi.ac.uk/gwas/).

The list of transcription factors (TFs) was obtained from the transcriptional factor prediction database (TFDB) (http://www.transcriptionfactor.org/).

The list of cell surface proteins was downloaded from a previous study (https://doi.org/10.1073/pnas.0907939106).

All these gene lists used in this study are included in Supplementary Data 3.

Other data that support the plots within this paper and other findings of this study are available from the corresponding author upon reasonable request.

All figures are associated with raw data and preprocessed data.
Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- Life sciences
- Behavioural & social sciences
- Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nt-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

**Sample size**
Sample sizes for Dropseq were determined based on previous experience with human islet Dropseq (PMID: 30865899). A minimum requirement of ~1000 cells per experiment was used to allow high confidence identification of relatively rare cell subpopulations (PMID: 31068996). No statistical tests or power analyses were used to pre-determine sample size.

**Data exclusions**
For single-cell data, standard filtering was used following Seurat3. Briefly, Genes that are expressed in less than 3 cells are filtered. Cells that express less than 200 genes or over 7500 are filtered. Cells that have more than 5% mitochondrial counts are filtered. No other data were excluded.

**Replication**
For the differentiation process analysis, the major reproducibility examination is performed by two comparable protocols (W/ and W/O ROCKII inhibitors). All dropseq included thousands of cells as independent biological replicates. Flow cytometry panels presented are representative over more than 3 reproducible results. qPCR is performed by 4 independent biological replicates.

**Randomization**
Due to the experiment design, the cells collected for Dropseq is by time series from ES cells to Stage 7.

**Blinding**
Blinding is not applicable. The time series single-cell genomics is quantitative, no subjective group allocation is involved.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

| n/a | Involved in the study |
|-----|-----------------------|
| ☐   | Antibodies            |
| ☒   | Eukaryotic cell lines |
| ☒   | Palaeontology         |
| ☒   | Animals and other organisms |
| ☒   | Human research participants |
| ☐   | Clinical data         |

### Methods

| n/a | Involved in the study |
|-----|-----------------------|
| ☒   | ChIP-seq              |
| ☒   | Flow cytometry        |
| ☒   | MRI-based neuroimaging |

**Antibodies**

**Antibodies used**

- PE Mouse anti-PDX1 antibody (BD Biosciences, Catalog No 562161);
- PE mouse anti-NX6.1 antibody (BD Biosciences, Catalog No 563023);
- Alexa Fluor®647 Mouse anti-NX6.1 antibody (BD Biosciences, Catalog No B0856338);
- Alexa Fluor®647 Mouse anti-NX2.2 antibody (BD Biosciences, Catalog No B08564729);
- PE Mouse anti human Chromogranin A antibody (BD Biosciences, Catalog No 564563);
- Rat Anti C-peptide antibody (Developmental Studies hybridoma Bank, University of Iowa, Catalog No GN-ID4);
- A FC Goat anti-Rat IgG antibody (BIOlegend, Catalog No 405407).

**Validation**

- PE Mouse anti-PDX1 antibody (BD Biosciences, Catalog No 562161, clone 658A5, 1:50); Manufacturer validation: flow cytometry in Human sample test, Relevant use in literature: Babu DA, et al. 2007, Molecular Genetics and Metabolism.
- PE mouse anti-NX6.1 antibody (BD Biosciences, Catalog No 563023, clone R11-560, 1:50); Manufacturer validation: Human and mouse species with flow cytometry, Relevant use in literature: D’Amour KA, et al. 2006, Nature Biotechnology.
- Alexa Fluor®647 Mouse anti-NX6.1 antibody (BD Biosciences, Catalog No B0856338, clone R11-560, 1:50); Manufacturer validation: human and mouse species with flow cytometry, immunofluorescence, Bioimaging, Relevant use in literature: D’Amour KA, et al. 2006, Nature Biotechnology.
Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)  H1 cells (WA01) was obtained from WiCell company, cell stocks were created and maintained in U’s Lab.

Authentication  H1 cells were authenticated using hPSC genetic analysis kit (catalog No 07550) from STEMCELL Technology

Mycoplasma contamination  H1 cells was tested negative for mycoplasma contamination

Commonly misidentified lines

(See ATCC register)

No

Flow Cytometry

Plots

Confirm that:
- The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
- The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a ‘group’ is an analysis of identical markers).
- All plots are contour plots with outliers or pseudocolor plots.
- A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation  The ES cells or the differentiated cells were washed once with 1 x DPBS and dissociated with Accutase [Innovative cell technologies #A1-104] to release single cells, which were stained with fixable viability dye (Thermofisher #LM969), then fixed for staining intracellular or nuclear markers with staining kit (eBioscience #00-5523-00). In the case where primary antibody is directly conjugated, the cells will be resuspended 2% BSA in 1xPBS after staining and ready for flow cytometry analysis; otherwise, cells will be subjected for staining with respective secondary antibody, then washed and resuspended for flow cytometry analysis.

Instrument  BD LSR-II

Software  FlowJo 7.6.1

Cell population abundance  The abundance of positive cell population was generally above 5%, as indicated in each plot in the paper, among 20,000 to 30,000 total events considered for the relevant gating. The purity of sample is high, because the sample is originated from one cell line, and then cultured and differentiated towards human pancreatic beta-like cells following the published protocol. The purity of sample was determined by flow cytometry as well as by single cell RNA-seq method.

Gating strategy  The stained samples were always matched with staining control, which was a parallel sample stained with viability dye and isotype control antibody but not with desired primary antibody, to determine the positive and negative population. In SSC-A/ SSC-C chart, the left lower corner of SSC/SSC chart was considered as cell debris, the majority events which did not hit the walls of chart were considered as cells for following gating. Resulting cells were then gated for singlet according to their FSC-H v.s. FSC-A ratio, the doublet cells were excluded. The desired population was subject to live cell gating subsequently, as the dead cells were stained positive with fixable viability dye (Thermofisher: 123101) and excluded. Among live cells, the population only appeared in staining control was defined as negative population, while the population only appeared in tested sample and showed strong fluorochrome signals were defined as positive population.

Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.