PNAS

Supplementary Information for

Single-cell Transcriptome and Accessible Chromatin Dynamics During Endocrine Pancreas Development

Eliza Duvall, Cecil M. Benitez, Krissie Tellez, Martin Enge, Philip T. Pauerstein, Lingyu Li, Songjoon Baek, Stephen R. Quake, Jason P. Smith, Nathan C. Sheffield, Seung K. Kim, H. Efsun Arda

S.K.K., E-mail: seungkim@stanford.edu
H.E.A., E-mail: efsun.arda@nih.gov

This PDF file includes:

Other supplementary materials for this manuscript include the following:

Supplementary Information Text

1. SI Materials and Methods

A. Animal models. All animal experiments were conducted in accordance with Stanford University IACUC guidelines. *Neurog3*$^{eGFP/+}$ knock-in reporter mice were a kind gift from Dr. Klaus Kaestner (University of Pennsylvania, USA) (1) and were maintained on a CD1 background. *Neurog3-Cre* mice were obtained from Guoqiang Gu (Vanderbilt University, USA) and maintained on a mixed background of C57BL/6 and CD1 (2). *Rosa-mTmG* (3) mice were obtained from the Jackson Laboratories and maintained on a mixed background of C57BL/6 and CD1. *Tg-eGFP; Neurog3*$^{+/+}$ transgenic mice were a kind gift from Drs. Guoqiang Gu and Douglas Melton (4). Timed matings were used to obtain mice at embryonic day (E) E15.5 and E17.5 for experiments; observation of a vaginal plug was considered E0.5 for embryonic staging purposes. Both male and female mice were used in all experiments.

B. Single-cell RNA-Seq Data Processing. Raw reads passing FastQC (5) quality control were aligned to a custom reference genome for mouse genome (mm10), ERCC spike in controls, and three transgenes: *eGFP*, *tdTomato*, and *Cre*. STAR was used to create the custom genome and read alignment (6). The resulting BAM/SAM files were used to create a 'master counts table' using HT-seq (Dataset S1) (7). Cells had an average of 3,044 genes expressed per cell, ranging from 1,237 to 6,047 genes.

C. Unsupervised Single-cell Clustering and Trajectory Analyses. Clustering and trajectory analysis were performed using the single-cell analysis package Monocle 2 (v. 2.4.0) (8). A flowchart summarizing each analysis step is provided in Fig. S2. Before starting the analysis, the transgenes *GFP, Cre* and *Td-tomato* were removed from the master counts table (Dataset S1). Unsupervised clustering aims to cluster the cells based on global gene expression profiles. First step is to choose which genes to use to cluster the cells. Based on the dispersion calculations, we set the `mean_expression` parameter to 1. Before performing dimension reduction, the data was examined using the `plot_pc_variance_explained` function, which plots the percentage of variance explained by each principal component on the normalized expression data. Based on the 'elbow' method, we determined that the first 5 dimensions showed the majority of data variability. Therefore, *t*-distributed stochastic neighbor embedding (*t*-SNE) dimension reduction was performed on the first 5 principal components. We set *num_clusters* to 7 to visualize cell clusters (Fig. S1). The identity of the cell clusters was revealed by mapping marker gene expression levels onto single-cells (Figure 1A-B). Clusters 4 and 6 were combined and labeled "Endocrine 1". At this point, the 14 mesenchymal cells that formed Cluster 5, and genes that were expressed in less than 5 cells were filtered out. To establish pseudotime trajectories, Monocle's `differentialGeneTest` function was used to find genes that vary among the clusters, specified as `fullModelFormulaStr =` "~Cluster". Top 100 genes with the lowest *q*-value were used to order cells, and a pseudotime trajectory was constructed using the `DDRTree` method. To identify gene expression changes between cells aligned along the established pseudotime trajectories, we used Monocle's `differentialGeneTest` function by specifying `fullModelFormulaStr` = "~Pseudotime". We considered genes significant if the rounded *q*-value was less than or equal to 0.05. Gene ontology terms were found for each of the 7 clusters using DAVID v6.8 (Dataset S3) (9).

D. Semi-supervised Single-cell Clustering and Trajectory Analyses. Semi-supervised clustering and trajectory analyses were performed to resolve individual endocrine lineage branching (Fig. S2). The process begins with defining marker genes that represent cell populations, then identifying the genes that co-vary with these markers, and finally ordering the cells based on these co-varying genes. Monocle provides the `CellTypeHierarchy` function for semi-supervised clustering analysis. Since our goal was to resolve the β-, α- and δ-cell branches, we picked marker genes as *Neurog3* for endocrine progenitors, *Ins1* and *Ins2* for β-cells, *Gcg* for α-cells, and *Sst* for δ-cells. We set the expression threshold in each cell for these markers to 100 or more reads. Accordingly, cells that express more than one marker gene are labeled "ambiguous" and cells that do not fit into any marker gene category are labeled as "unknown". The gene list was further filtered to remove genes if detected in less than 5 cells. Top 100 genes that co-varied with the marker genes (400 genes in total) were considered for the clustering and trajectory analysis. Note that the semi-supervised analysis was limited to the 317 cells that were placed after the *Neurog3* peak expression in the unsupervised trajectory, which corresponds to the pseudotime point 6.7. The first iteration separated β-cells in one branch and the majority of α- and δ-cells in a second branch. To split the α- and δ-branches, we again focused on cells of interest, and excluded the cells on

the β-cell branch to create a new `CellDataSet` (cds) object in Monocle. In this new `cds()` object, cells were relabeled as α-, δ-, and *Neurog3*^{pos} cells based on marker gene expression. Trajectory analysis was performed as described earlier. The final iteration established trajectories with α- and δ-cells separated on own branches. Similar to unsupervised clustering, Monocle's `differentialGeneTest` (by specifying `fullModelFormulaStr` = "~Pseudotime") function was used to identify genes whose expression changes significantly during each endocrine lineage specification. For differential gene expression analysis, cells with pseudotime point > 5.7 and ≤ 6.7 were also included (peak *Neurog3* expression) to visualize the cell fate transitions beginning from the *Neurog3*^{pos} progenitors. Hence, three differential gene tests were performed to determine transcriptome changes from *Neurog3*^{pos} progenitor cells to each of the three endocrine lineages. Results from differential expression analyses were filtered to include genes with a *q*-value less than 0.1 and those in the top 50% of normalized base mean expression among cells within each branch. All differentially expressed genes lists were further narrowed to only include transcription factors (TFs) for a total of 145 TFs. These TFs are visualized in a heatmap where all cells were aligned in pseudotime order (Fig. S4).

E. Analysis and Classification of *Neurog3*^{pos} Progenitors. The master read counts table (Dataset S1) was subset to select *Neurog3*^{pos} cells. We defined *Neurog3*^{pos} cells as any cell with at least 10 read counts for *Neurog3*, resulting in 214 cells. The semi-supervised clustering approach was used to label and cluster cells based on either *Neurog3* or *Chga* expression (see SI Methods Section D). Top 100 genes that co-varied with the marker genes (200 genes in total) were considered for the clustering and trajectory analysis. *t*-SNE dimension reduction was performed on the first two principal components, and *num_clusters* was set to 3. Based on the *Neurog3* levels, the clusters were named High, Medium, and Low. A trajectory was established by finding differentially expressed genes among the High, Medium, Low clusters, using Monocle's `differentialGeneTest` function by specifying `fullModelFormulaStr` = "~Cluster". Top 100 genes with the lowest *q*-value were used to order cells, and a pseudotime trajectory was constructed using the `DDRTree` method. The trajectory was colored based on embryonic day (Figure 2F) or cluster (Figure 2H). To count hormone expressing cells, we analyzed the read counts of *Ins1*, *Ins2*, *Gcg* and *Sst* in each *Neurog3*^{pos} cell. Any detectable expression (i.e. size-factor normalized counts > 0) was counted. The cells were then categorized as expressing zero, one, two or three hormones (*Ins1* and *Ins2* reads were combined and presented as *Ins*).

F. Expression Specificity Scores, TF-Cell Type/State Network. We derived expression specificity scores for TFs that are differentially expressed during endocrine cell lineage specification. We have previously used this method to reveal cell type-specific gene expression in human pancreas cells (10). ESS was calculated as follows:

$$x_i \text{ is the expression of the gene in cell state } i$$
$$n \text{ is the number of cell states}$$
$$ESS = \frac{median\ (x_i)}{\sum_{i=1}^{n} median(x_i)}$$

A cell state is defined here as population of cells that are quantitatively distinct based on their transcriptome. Two cell states (early progenitor, late progenitor) and four cell types (duct, β-, α- and δ-cells) were used to determine the expression specificity score of each TF. The duct cells were categorized as cells with pseudotime values < 3 (53 cells) based on the unsupervised trajectory analysis. Early progenitor state has cells with pseudotime values between 3 and 6.7 (111 cells). Late progenitor state cells have a pseudotime value greater than 6.7 and include those that were not assigned to an endocrine lineage (121 cells). The hormone producing cells consist of those assigned to their respective endocrine cell branch (90 cells in the β-lineage, 76 cells in the α-lineage, and 30 cells in the δ-lineage). To obtain $\boldsymbol{x_i}$, we used the size-factor normalized single-cell RNA-Seq counts as gene expression values. Thus, a TF with an ESS of zero would indicate no expression in that cell type/state, and an ESS of 1 would indicate exclusive expression, i. e. the TF is only expressed in that cell state. We obtained the list of differentially expressed TFs by overlapping the gene the list with a curated TF list described in (11), yielding 145 TFs. The TF list was further narrowed to 87 by only including those that were detected in at least 50% of the cells in that cell type/state (Dataset S5). The network was generated by Cytoscape (version: 3.8.2) (12). The color and thickness of the network edges (connections) directly corresponds with the expression specificity score (ESS) of the TF in the interacting cell type/state.

G. TF motif enrichment analysis

HOMER's findMotifsGenome.pl function with '`size 500 –len 6,8`' options was used to find enriched TF motifs in each DOR group (13). HOMER's de novo motif discovery analysis outputs a position weight matrix (PWM) for each significant motif. These PWMs were queried in the CisBP database (11) to find transcription factors associated with the significant motifs.

I. BaGFoot Analysis and Integration of Gene Expression

BaGFoot footprint analysis was performed as described in (14). Narrow peaks were called for all ATAC-seq samples using MACS2 (15) and merged to generate a set of consensus peaks for BaGFoot. Peaks overlapping with blacklisted regions (downloaded from http://mitra.stanford.edu/kundaje/akundaje/release/blacklists/mm10-mouse/mm10.blacklist.bed.gz) (78) were removed from the analysis. 662 mouse TF motifs were curated from TRANSFAC (16) , JASPAR (17) and UniPROBE (18). In addition, we included 19 de novo motifs derived from our ATAC-seq data by HOMER motif analysis. ATAC-seq sample replicates were grouped as follows: the duct dataset consisted of duct-het, duct-null, and Neurog3-null samples, the Neurog3 dataset consisted of Neurog3-het and Neurog3-Tg samples, and the endocrine dataset consisted of Endo-het and Endo-Tg samples. Each group were compared pairwise to detect TF footprint activity at motif locations. BaGFoot results are presented in "bag plots", where each data point represents a TF motif. In a bag plot, the bag area contains 50% of the data (similar to the box in the box plot), the fence contains 97%-100% of the data points (similar to the whiskers in a box plot) (19). Any data point outside the fence is an outlier. Most TF motifs are not expected to be different between two conditions, and thus are localized around the origin. The significant motifs were statistically determined by Hotelling's T-squared test and were labeled as outliers. Based on the BaGFoot results, we compiled a list of outlier TFs (and their paralogs) to analyze their expression levels in the scRNA-Seq data. 481 cells were divided into duct, progenitor, and endocrine cell types to obtain average expression levels for outlier TFs. Cells were assigned to one of these three cell types based on their placement from the pseudotime trajectory analyses. Endocrine cells are a combination of cells aligned on the β-, α-, and δ-branch (Dataset S8). The TFs whose expression was detected in at least 25 cells within each cell group were listed in Figure 6F. Those detected in fewer than 25% of the cells were shown in Fig. S5.

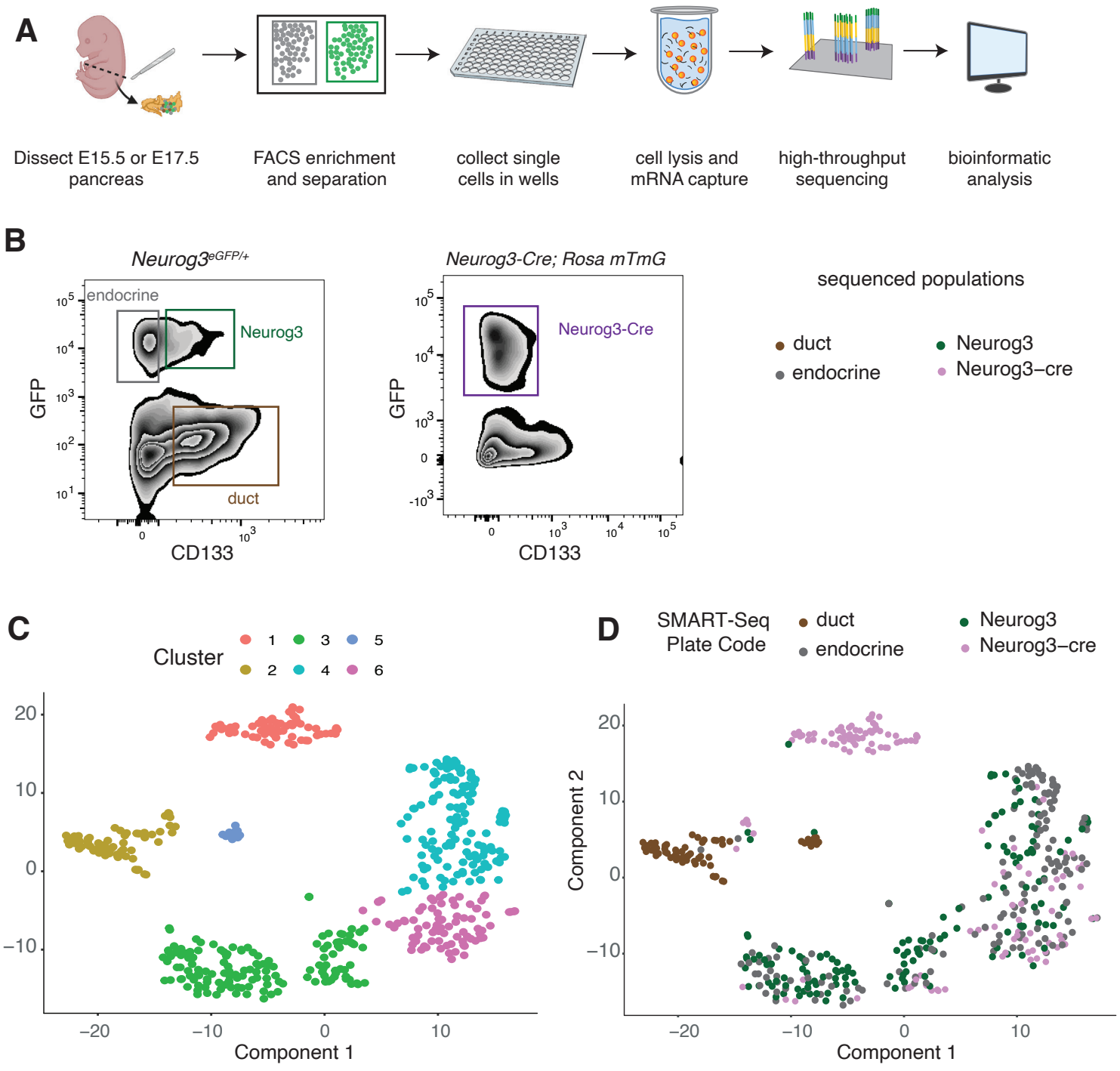
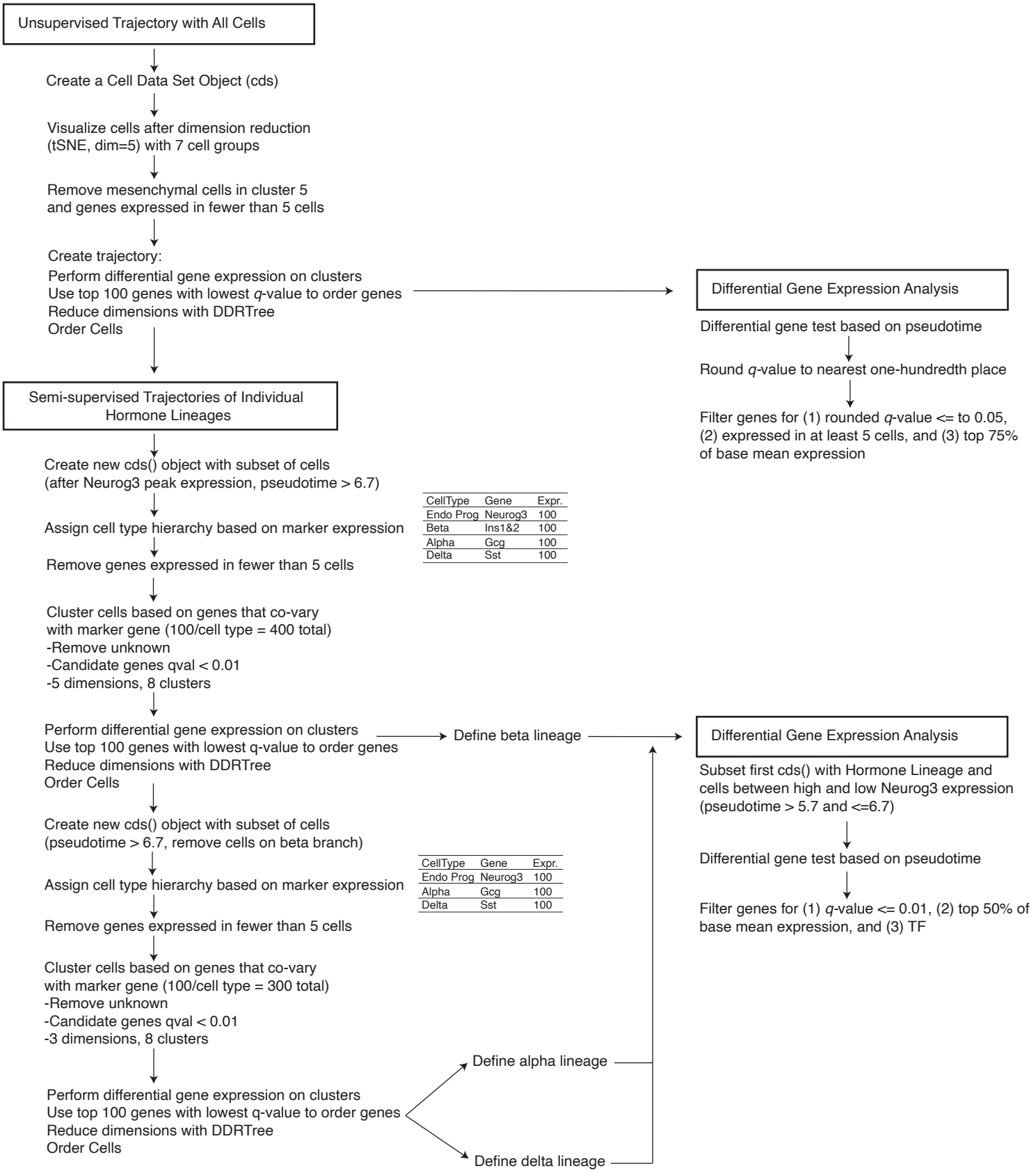


Figure S1.
A) Illustration of the scRNA-Seq workflow performed in this study. See methods for details.
B) FACS plots showing the gating strategy and cell populations collected for single-cell sequencing.
C) t-SNE plot showing single cell clusters after unsupervised clustering approach.
D) Same plot as (C), colored by cell populations indicated in (B).



CellType	Gene	Expr.
Endo Prog	Neurog3	100
Beta	Ins1&2	100
Alpha	Gcg	100
Delta	Sst	100

CellType	Gene	Expr.
Endo Prog	Neurog3	100
Alpha	Gcg	100
Delta	Sst	100

Figure S2: Flowchart detailing the scRNA-Seq analysis using Monocle.

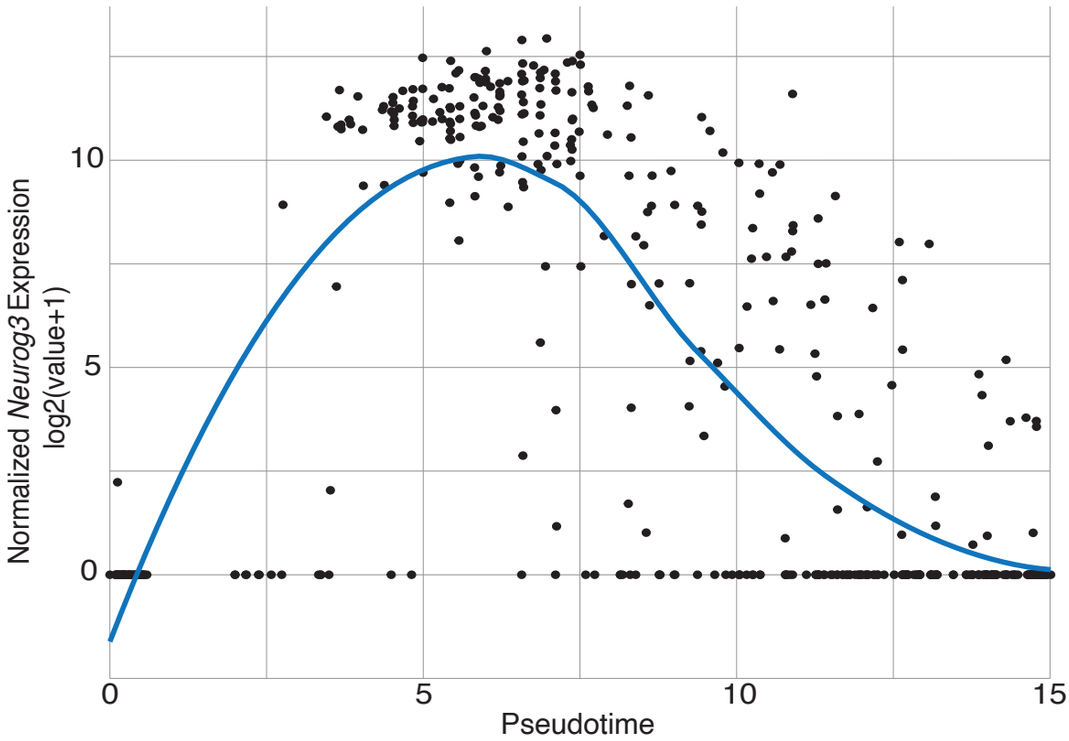


Figure S3. Distribution of Neurog3 transcript levels in single cells, ordered by the pseudotime defined in Figure 1C.

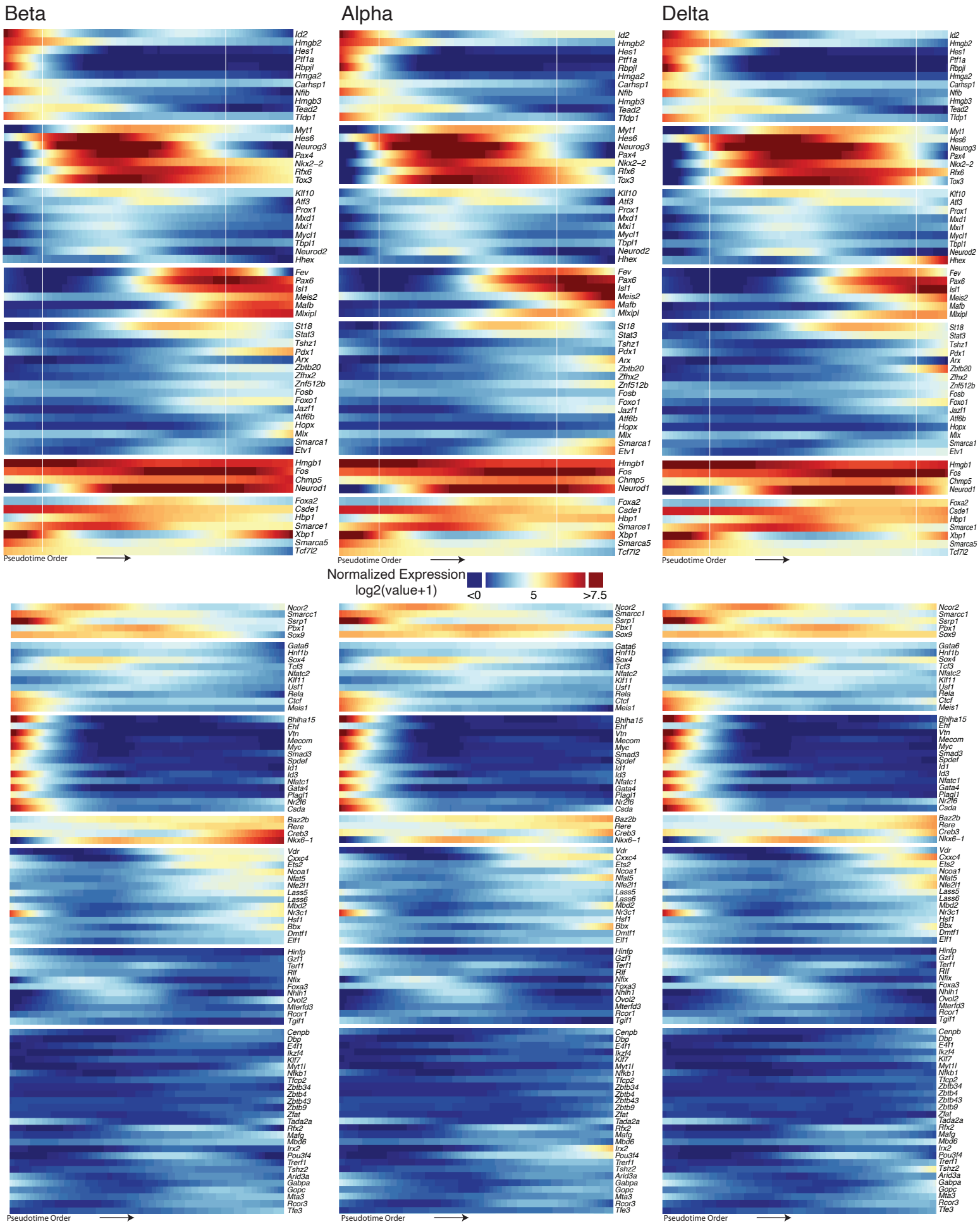

Beta
Alpha
Delta
Id2
Hmgb2
Hes1
Ptf1a
Rbpjl
Hmga2
Carhsp1
Nfib
Hmgb3
Tead2
Tfdp1
Myt1
Hes6
Neurog3
Pax4
Nkx2–2
Rfx6
Tox3
Klf10
Atf3
Prox1
Mxd1
Mxi1
Mycl1
Tbpl1
Neurod2
Hhex
Fev
Pax6
Isl1
Meis2
Mafb
Mlxipl
St18
Stat3
Tshz1
Pdx1
Arx
Zbtb20
Zfhx2
Znf512b
Fosb
Foxo1
Jazf1
Atf6b
Hopx
Mlx
Smarca1
Etv1
Hmgb1
Fos
Chmp5
Neurod1
Foxa2
Csde1
Hbp1
Smarce1
Xbp1
Smarca5
Tcf7l2
Pseudotime Order
Normalized Expression
log2(value+1)
<0
5
>7.5
Ncor2
Smarcc1
Ssrp1
Pbx1
Sox9
Gata6
Hnf1b
Sox4
Tcf3
Nfatc2
Klf11
Usf1
Rela
Ctcf
Meis1
Bhlha15
Ehf
Vtn
Mecom
Myc
Smad3
Spdef
Id1
Id3
Nfatc1
Gata4
Plagl1
Nr2f6
Csda
Baz2b
Rere
Creb3
Nkx6–1
Vdr
Cxxc4
Ets2
Ncoa1
Nfat5
Nfe2l1
Lass5
Lass6
Mbd2
Nr3c1
Hsf1
Bbx
Dmtf1
Elf1
Hinfp
Gzf1
Terf1
Rlf
Nfix
Foxa3
Nhlh1
Ovol2
Mterfd3
Rcor1
Tgif1
Cenpb
Dbp
E4f1
Ikzf4
Klf7
Myt1l
Nfkb1
Tfcp2
Zbtb34
Zbtb4
Zbtb43
Zbtb9
Zfat
Tada2a
Rfx2
Mafg
Mbd6
Irx2
Pou3f4
Trerf1
Tshz2
Arid3a
Gabpa
Gopc
Mta3
Rcor3
Tle3
Pseudotime Order
Normalized Expression
log2(value+1)
<0
2
4
>5.5

Figure S4. Heat maps showing dynamic expression level changes through endocrine cell differentiation in beta-, alpha-, or delta-branch. 145 TFs were found to be differentially expressed along this pseudotime order obtained from semi-supervised clustering analysis. For visualization purposes, the TF list is split into those with high (top) or low (bottom) expression. Color scale indicates log2 transformed normalized expression values after LOESS smoothing.

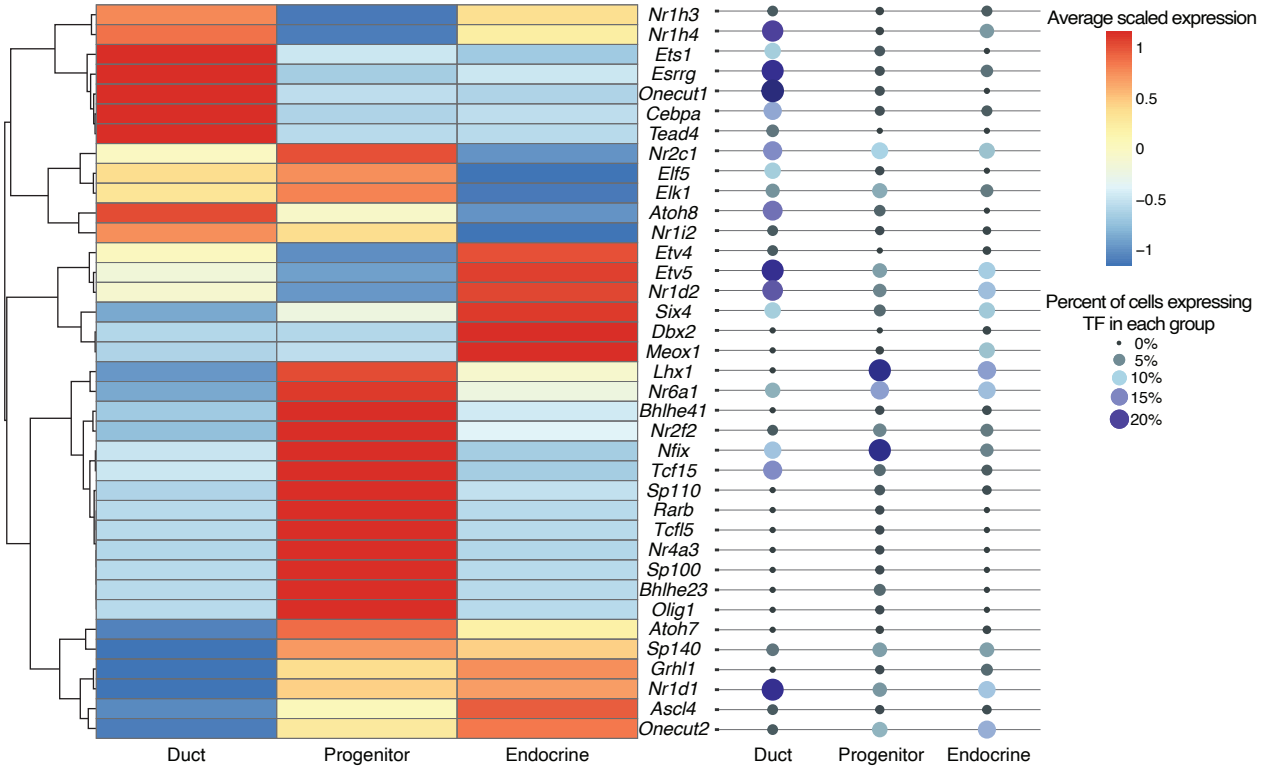


Figure S5. Heat map shows average expression levels of outlier TFs in duct, progenitor, or endocrine cells. TFs are ordered by hierarchical clustering; expression levels are scaled to each row. Each TF is detected in less than 25% of cells for each group.

Table S1. ATAC-seq samples generated in this study

SampleID	Genotype	CellType	FACS gating
Tn5_EA20	*Neurog3*$^{eGFP/eGFP}$	Neurog3 null	CD133+GFP+
Tn5_EA23	*Neurog3*$^{eGFP/+}$	Duct	CD133+GFP-
Tn5_EA32	*Neurog3*$^{eGFP/eGFP}$	Neurog3 null	CD133+GFP+
Tn5_EA33	*Neurog3*$^{eGFP/eGFP}$	Duct	CD133+GFP-
Tn5_EA34	*Neurog3*$^{eGFP/+}$	Neurog3	CD133+GFP+
Tn5_EA35	*Neurog3*$^{eGFP/+}$	Duct	CD133+GFP-
Tn5_EA36	*Neurog3*$^{eGFP/+}$	Endocrine	CD133-GFP+
Tn5_EA37	*Neurog3*$^{eGFP/eGFP}$	Neurog3 null	CD133+GFP+
Tn5_EA38	*Neurog3*$^{eGFP/eGFP}$	Duct	CD133+GFP-
Tn5_EA39	*Neurog3*$^{eGFP/+}$	Neurog3	CD133+GFP+
Tn5_EA40	*Neurog3*$^{eGFP/+}$	Duct	CD133+GFP-
Tn5_EA41	*Neurog3*$^{eGFP/+}$	Endocrine	CD133-GFP+
Tn5_EA57	*Tg-Neurog3*	Neurog3	CD133+GFP+
Tn5_EA59	*Tg-Neurog3*	Endocrine	CD133-GFP+
Tn5_EA60	*Tg-Neurog3*	Neurog3	CD133+GFP+

Dataset S1. Gene transcript counts obtained from single cells sequenced in this study.

Dataset S2. Differentially expressed genes during endocrine cell differentiation, based on the pseudotime established in Figure 1.

Dataset S3. GO Term results of the gene clusters identified in Figure 1D.

Dataset S4. Differentially expressed genes during α-, β- and δ-cell lineage specification, as identified by Monocle analysis.

Dataset S5. ESS of genes in each cell state. Number of cells in which a given gene transcript is detected is also reported.

Dataset S6. Genomic coordinates of differentially accessible open chromatin regions identified in this study.

Dataset S7. FPD and FA scores for each TF based on BaGFoot analysis.

Dataset S8. Average expression of TF transcripts detected in single cells. This data was used to integrate with BaGFoot results.

SI References

1. C. S. Lee, N. Perreault, J. E. Brestelli, K. H. Kaestner, Neurogenin 3 is essential for the proper specification of gastric enteroendocrine cells and the maintenance of gastric epithelial cell identity. *Genes Dev.* **16**, 1488–1497 (2002).

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

3. M. D. Muzumdar, B. Tasic, K. Miyamichi, L. Li, L. Luo, A global double-fluorescent Cre reporter mouse. *Genesis* **45**, 593–605 (2007).

4. G. Gu, *et al.*, Global expression analysis of gene regulatory pathways during endocrine pancreatic. *Development* **131**, 165–179 (2004).

5. S. Andrews, FastQC: A Quality Control Tool for High Throughput Sequence Data [Online] (2015).

6. A. Dobin, *et al.*, STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, bts635 (2012).

7. S. Anders, P. T. Pyl, W. Huber, HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).

8. X. Qiu, *et al.*, Reversed graph embedding resolves complex single-cell trajectories. *Nat Methods* **14**, 979–982 (2017).

9. D. W. Huang, B. T. Sherman, R. A. Lempicki, Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**, 44–57 (2009).

10. H. E. Arda, *et al.*, A chromatin basis for cell lineage and disease risk in the human pancreas. *Cell systems* **7**, 310–322 (2018).

11. M. T. Weirauch, *et al.*, Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431–1443 (2014).

12. P. Shannon, *et al.*, Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498–2504 (2003).

13. S. Heinz, *et al.*, Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).

14. S. Baek, I. Goldstein, G. L. Hager, Bivariate Genomic Footprinting Detects Changes in Transcription Factor Activity. *Cell Rep* **19**, 1710–1722 (2017).

15. J. Feng, T. Liu, B. Qin, Y. Zhang, X. S. Liu, Identifying ChIP-seq enrichment using MACS. *Nature Protocols* **7**, 1728–1740 (2012).

16. V. Matys, *et al.*, TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res* **34**, D108-110 (2006).

17. A. Mathelier, *et al.*, JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* **44**, D110-115 (2016).

18. D. E. Newburger, M. L. Bulyk, UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. *Nucleic Acids Res* **37**, D77-82 (2009).

19. P. J. Rousseeuw, I. Ruts, J. W. Tukey, The Bagplot: A Bivariate Boxplot. *The American Statistician* **53**, 382–387 (1999).