Supplementary information, data S1, Extended Materials and Methods

hESCs culture

Human embryonic stem cell (hESC) lines H1 (WiCell) and H9 were cultured in DMEM/F12 (Life Technologies) supplied with 20% (vol/vol) KnockOut Serum Replacement (KSR, Life Technologies), 1% (vol/vol) MEM Non-Essential Amino Acids Solution (NEAA, Life Technologies), 1% (vol/vol) GlutaMAX Supplement (Life Technologies), 1% penicillin-streptomycin (PS, HyClone), 0.05 mM 2-mercaptoethanol (2-ME, Life Technologies) and 25 ng/ml basic fibroblast growth factor (bFGF, OriGene) on mitomycin-C (Sigma-Aldrich, or Roche)-treated mouse embryonic fibroblast (MEF) feeder layers. The medium was changed daily. Cultures were normally passaged by Dispase II (Roche) at a 1:4-1:6 split ratio every 4-6 days.

Generation of targeting vectors and TALENs

The BAC-based targeting vector for Ngn3 was generated with a modified BAC recombineering method derived from previous description (Liu et al., 2003). The freshly prepared BAC containing *NGN3* gene was electroporated into DY380 *E. coli* cells. To replace the stop coden, two homology arms with the length of ~1kb were inserted into sites flanking the *2A-eGFP-loxp-CAG-neo-loxp* cassette, left arm with BglII and KpnI, right arm with EcoRI and SalI (NEB), respectively. The DY380 cells containing BAC were induced in 42°C water bath for 15 minutes and electroporated with the cassette for replacement released with BglII and SalI. To shorten the right arm of the BAC-based targeting vector, two homology arms with the length of ~1kb were inserted into the sites flanking the *HSV-TK-Amp*^{r} cassette, left arm with SpeI and BamHI, right arm with AgeI and KpnI (NEB), respectively. The DY380 cells containing modified BAC were induced in 42°C water bath for 15 minutes and electroporated with the cassette for shorten released with SpeI and KpnI. Then the BAC-based targeting vector were mega-prepared (Qiagen), linearized with PI-SceI(NEB) and extracted with Phenol–chloroform.

All targeting vectors for other eight gene loci were generated with an improved BAC-recombineering approach (Wang et al., 2006). The plasmid PSC101-BAD-gbaA was electroporated into *E. coli* cells containing BACs. To replace the stop coden for each gene, the *E. coli* cells induced with 0.1% L-arabinose for 0.5h were electroporated with the PCR product of *2A-Tdtm-loxp-CAG-neo-loxp* (or *2A-eGFP-loxp-CAG-neo-loxp*) cassette flanked with 50-bp homology arms for target site. For retrieving, the *E.*

coli cells containing the modified BACs and *PSC101-BAD-gbaA* were induced with 0.1% L-arabinose for 0.5h and electroporated with the PCR product of plasmid PL253SNK flanked with 50-bp homology arms. The plasmid PL253SNK was derived from plasmid PL253 (Liu et al., 2003).The targeting vectors were mini-prepared and retransformed into *E. coli* cells to eliminate BAC. The pure targeting vectors were largely prepared, linearized with proper enzymes and directly precipitated with ethanol. All BACs used in this paper were ordered from Children's Hospital Oakland Research Institute (CHORI) as listed in the following table.

Gene symbol	**BACs used**
PDX1	ch17-423d7
NKX6.1	RP11-726G24
NEUROD1	RP11-880E10
MAFB	RP11-246H10
PAX6	RP11-307I15
FOXA2	RP11-816F2
SOX17	RP11-359F18
INS	rp11-889I17
NGN3	rp11-343J3

DNA sequences of TALENs specific for selected gene loci were synthesized by a high-throughput integrated chip method as the previous described (Wang et al., 2012). The DNA sequences recognized by respective pair of TALENs are listed in following table.

Target genes	**DNA sequences recognized by TALENs**
FOXA2	ATGGTTTCTGCGTGCTttatttatggcttataa ATGTGTATTCTGGCTGC
INS	GCTGCCCCCACCCTGTggctcagggtccagtatgggAGCTGCGGGGGTCTCTG
MAFA	TGAGCCAGGTCTAACTT ctttccaagcgtcc GCTTGTACATACGTTGA
MAFB	AGTCCCGAGAAGTCACC aaggccatctggag ACTCCTGGCTTTCTGA
NEUROD1	ACAAAAGGCAGCCCTTT gggtactactgctgca AAGTGCAAATACTCCA
NKX6.1	GCAACTAAAGTAACCT gttgaaggctctttgta AATAAATCGTGAGTTAC
PAX6	AGAGCCGCTTCAGTTCT acaattgtgtcctgt ATTGTACCACTGGGGA
PDX1	GCCCTCCTACAGCACT ccaccttgggacctgttt AGAGAAGCCGGCTCTTC
SOX17	CCTGTGCCAGATGTTT gttcaatgccattctaaac AGTGTGAGCCTCAACAA

Gene targeting in hES cells

The human embryonic stem cell line H1 were treated with Y27632 (Sigma) overnight and digested to single cell with accutase (Milipore). For targeting *NGN3* locus, the 80μg linearized BAC-based targeting vector was electroporated into $2x10^7$ cells in 0.4 cm cuvettes using Bio-Rad electroporator, in 800μl DMEM/F12 plus

20% KSR, 320v, 200uF. Then the cells were placed onto five 10 cm-dish. For targeting other genes, *NGN3-eGFP* cell line constructed with above method were used as targeting cell line, except *SOX17*, for which wild-type H1 were used. Briefly 10μg linearized targeting vector and 5μg left and right TALENs were nucleofected into $3x10^6$ cells using 4D-Nucleofector™ System (Lonza). Then the cells were placed onto one 10cm diameter dish. Three repeats of nucleofection were carried out for each gene. After three days recovery, drugs were added into the culture medium for 2 weeks selection, G418 50μg/ml for first week and 100μg/ml for second week, ganciclovir 0.2μM. Up to 192 colonies (96 colonies in most cases) for each gene were picked and expanded in 96-well plates. To reduce the laborious work, we used two rounds of long-range genomic PCR analysis. In first round, we treated the colonies in one column as a group for preparation of genomic DNA (Tiangen) and PCR screening (Takara, DR044A). In second round, as long as one group was PCR positive, the colonies in this group would be screened one by one. The PCR positive colonies were expanded for southern blot and differentiation analysis.

HESC Differentiation

Before differentiation, adherent H1 cells (at ~70% confluence) were rinsed with phosphate-buffered saline (PBS, HyClone) and then incubated with Accutase (Millipore) for 6-8 min at 37°C. Released single cells were rinsed twice with DMEM/F12 and spun at 1200 rpm for 3 min. The resulting cells were resuspended in feeder-free hESC culture medium "Essential 8" (E8, Life Technologies), which was supplied with 10μM Y27632 (Sigma-Aldrich), and seeded on 1:30 diluted Matrigel (BD Biosciences)-coated dishes at a density of 20,000 cells/cm^2. The next day, the medium was exchanged for E8 and maintained for one more day prior to differentiation initiation.

Stage 1: Definitive Endoderm (4 days). For the first day, undifferentiated hESCs were exposed to DMEM/F12 supplied with 0.1% bovine serum albumin (BSA, Sigma-Aldrich), 25 ng/ml mWnt3A (R&D Systems), and 100-120 ng/ml Activin-A (Peprotech or Humanzyme). For days 2-3, the cells were cultured in DMEM/F12 with 0.1% BSA and 100-120 ng/ml Activin-A. On day 4, the cells were cultured in DMEM/F12 with 0.1% BSA, 120 ng/ml Activin-A and 0.5 μM Wnt-C59 (Cellagen), which inhibits Wnt signaling.

Stage 2: Primitive Gut Tube (3 days). Cells derived from stage 1 were cultured in DMEM/F12 supplied with 0.5x B27 without vitamin A, 50 ng/ml KGF (Peprotech or Humanzyme) and 1 μM SB525334 (Tocris), also known as Alk5 inhibitor VIII.

Stage 3: Posterior Foregut (4 days). Cells were cultured in DMEM-H (Life Technologies) supplied with 0.5x

B27 without vitamin A, 2 μM all-trans retinoic acid (RA, Sigma-Aldrich), 250 ng/ml Noggin (Peprotech or Humanzyme), and 0.25 μM SANT-1(Sigma-Aldrich), an inhibitor of SHH signaling.

Stage 4: Pancreatic progenitor and endocrine progenitor (3-5 days). Cells were cultured for three to five days in DMEM-H supplied with 1% B27 without vitamin A, 250 ng/ml Noggin and 50 nM TPB ((2S,5S)-(E,E)-8-(5-(4-(trifluoromethyl)phenyl)-2, 4-pentadienoylamino) benzolactam) (EMD Chemicals Inc). The treatment time course varied according to the cultures and cell lines.

Stage 5: Hormone-producing cells (>5 days). Cells were cultured for more than 5 days in DMEM-H with 1% B27 without vitamin A, 250 ng/ml Noggin, 10 ng/ml human LIF (Millipore), and Alk5 inhibitor II (Tocris).

Flow Cytometry and Cell Sorting

Single-cell suspension from cell cultures: Differentiated hESC cultures were rinsed with PBS and then incubated with 0.25% trypsin-EDTA (Life Technologies) at 37°C for 1-3 minutes. The trypsin was neutralized with MEF medium. The dissociated cells were rinsed twice in PBS or DMEM/F12 medium for further analysis.

Single-cell suspension from tissues: fetal pancreatic tissues were washed twice in cold PBS and minced into 1 mm^3 pieces with a sterile scalpel. The tissues were then incubated in PRMI 1640 supplemented with 100-400 U/ml Collagenase IV (Life Technologies), 1.2 U/ml Dispase II (Roche), DNase I (0.02%, (wt/vol)) and 0.5% fetal bovine serum (FBS, Hyclone) at 37°C for 30 minutes. Tissue masses were gently dispersed by pipetting every 5 minutes. The released single cells were transferred into PRMI 1640 with 0.5% FBS and washed twice in PBS with 0.5% BSA and 2 mM EDTA. The residual tissue masses were collected for repeated digestion as above. Finally, a uniform single-cell suspension was obtained by using a 40 μm Cell Strainer (BD Biosciences). Collected singles were kept on ice in PBS with 0.5% BSA and 2 mM EDTA for future analysis.

Flow Cytometry analysis: For the flow cytometry analysis of living cells, single cells were resuspended in 300 μL of basal culture medium. The data were acquired by BD FACSCalibur. For surface marker staining, approximately ~10^6 single cells were resuspended in 150 μL DMEM/F12 containing 0.1% BSA (washing buffer). The cells were then incubated at 4°C for 30 minutes with fluorochrome-conjugated antibodies. Stained cells were washed twice in washing buffer prior to analysis on the BD FACSCalibur. For intracellular antibody staining, single cells were fixed in 200 μL of BD Cytofix/Cytoperm Buffer (BD Biosciences) at 4°C for 20 minutes followed by three washes in BD Perm/Wash Buffer. Fixed cells were incubated in 150 μL of

primary antibody buffer at 4°C for 1 hour, followed by 30 minutes in a secondary antibody buffer after being rinsed twice in Perm/Wash Buffer. Stained cells were washed twice in Perm/Wash Buffer prior to analysis on the BD FACSCalibur. The acquired data were analyzed by FlowJo 7.6.

FACS: Fluorescence-activated cell sorting (FACS) was performed on BD FACS Aria IIu.

MACS: Magnetic-activated cell sorting (MACS) experiments were performed by using reagents from Miltenyi Biotec according to the manufacturer's instructions. In brief, dissociated cells were first stained with phycoerythrin (PE)-conjugated antibodies SUSD2-PE in MACS buffer (PBS with 0.5% (wt/vol) BSA and 2 mM EDTA) and then incubated with Anti-PE (PE) MicroBeads in MACS buffer after a rinse. Cell separations were performed using MS MACS cell separation columns on a MiniMACS Separator. Cells flowing through with buffer were named the "unbound fraction", while cells recovered by eluting the column were in the "bound fraction". Sorted cell fractions were used directly for staining and flow cytometry analyses or culture. For further culture, cells were replated on Matrigel-coated dishes in DMEM-H supplied with 1% B27 without vitamin A and 10 μM Y27632 overnight to attach to the dish bottom. Cells were then gently rinsed with PBS and maintained in stage 5 medium of hESC differentiation for more than 5 days prior to staining analysis.

All antibodies used above were listed in Table S3.

Immunofluorescence and imaging

Cell culture staining: cell cultures were washed twice in PBS and fixed in 4% paraformaldehyde (DGCS biotech) at room temperature for 15 minutes. Fixed cells were blocked in PBST (PBS containing 0.2% Triton 100 (DGCS) and 0.5% normal donkey serum (Jackson Immuno Research Laboratories)) at room temperature for 1 hour. Primary and secondary antibodies were diluted in PBST. Cells were incubated with primary antibodies at 4°C overnight, followed with three rinses and an incubation with secondary antibodies at room temperature for 1 hour. The stained cells were rinsed with PBS and then incubated with DAPI (Sigma-Aldrich) for 2 minutes to stain the nuclei. Cells were then washed three times in PBS prior to imaging. For confocal imaging, cells were usually mounted with mounting medium (ZS Biotech Inc.) and covered with a cover glass (ZS Biotech Inc.)

Tissue sectioning and staining: pancreatic tissues were fixed with 4% paraformaldehyde at 4°C for 2 hours, followed by three PBS washes at 4°C (which lasted a few seconds, 10 min and 2 h). The tissues were then

incubated in 30% (w/vol) sucrose solution at 4°C overnight. The tissues were embedded in Optimal Cutting Temperature Compound (O.C.T) (Tissue-Tek), frozen in liquid nitrogen and sectioned at 10 μm using a Cryostat (Leica). Section staining was performed by using the same procedure as in "Cell culture staining" without the fixation step.

Imaging: regular images were acquired using ECLIPSE TE2000 (NIKON). For confocal imaging, hESC differentiation was initiated in specific Lab-Tek II 8-Well Chambers (NUNC) or terminal cell cultures were sometimes replated on chambers for staining. Images were acquired using LSM 710&NLO (Zeiss). For real-time cell tracing, day 11 cell cultures of NEUROD1-DR cell line were traced with a High Speed Live Cell Confocal Imaging System (PE UltraView VOX, PerkinElmer Inc.). The tracking time course lasted 12 hours with intervals of 15 minutes.

Quantification of staining cultures or sections: To quantify the correlation between reporter gene expression and the corresponding endogenous genes, we combined immunostaining with confocal imaging. Because reporter proteins (especially TdTomato proteins) were sensitive to fixation and lasers, the expression of reporter gene was usually enhanced with an eGFP antibody or RFP antibody, while expression of endogenous genes were visualized with the corresponding antibodies (see Table S3). Co-localization counting was based on at least six randomly selected views from three independent experiments for each reporter. More than 2000 cells were assessed on average for one reporter cell line. The result is shown in Table S1**.** To quantify the correlation between the expression of SUSD2 and other genes in the fetal pancreas, co-localization counting was based on at least 12 random selected views in 6 sections from different fetal pancreas locations. More than 2000 total cells were assessed on average.

All antibodies used above were listed in Supplementary Table S3.

Quantitative RT-PCR, Genomic PCR

Regular PCR: 2×EasyTaq PCR SuperMix (TransGene) was used according to the manufacturer's instructions.

Genomic PCR: Genomic DNA was obtained by using a DNeasy Blood & Tissue Kit (QIAGEN). LA Taq (TaKaRa) was used for long-range genomic PCR according to the manufacturer's instructions.

Quantitative RT- PCR: Total RNA was extracted by using the Qiagen RNeasy Plus Mini Kit (Qiagen) or RNeasy Plus Micro Kit (Qiagen). Approximately 500 ng of RNA template was reverse-transcribed into cDNA

by using Random Primers with Transcript II FirstStrand cDNA Synthesis SuperMix (Transgene AH301). PCR was conducted with Power SYBR Green PCR Master Mix (Life Technologies) according to the user guide and performed in a Stratagene Mx3000P q-PCR System (SABiosciences). All quantified gene values were normalized against housekeeping gene GAPDH.

Primer sequences were listed in Supplementary Table S4.

Southern Blotting

Genomic DNA (15-20 μg) was digested for 8 hours at 37°C, separated on a 0.7-0.8% agarose gel and then transferred to a Hybond-N+ nylon membrane (Amersham). Afterwards, digoxigenin (DIG)-labeled DNA probes in DIG Easy Hyb buffer (Roche) were incubated with the membrane at 42°C for 16 hours. After a gentle rinse, the membrane was then incubated with alkaline phosphatase-conjugated anti-DIG antibody (approximately 1:10,000, Roche). The membrane was additionally incubated with CSPD (Roche) and exposed on film. The probes and primers for amplifying these probes were listed in Table S5.

RNA-seq sample preparation and analysis

2μg Total RNAs were extracted from cells with RNeay Plus Mini Kit (QIAGEN). MRNA was then purified for reverse transcription. A poly-A tail was added to the 3' end of first-strand cDNAs by terminal deoxynucleotidyl transferase, and then the cDNAs were amplified by 12 PCR cycles. The cDNA libraries were constructed following the standard operating procedure from Illumine with the Illumina Paired-End DNA Sample Prep Kit. Briefly, the cDNAs were sheared to 200-300bp fragments. After END-repairing and A-tailing, the adapters were added to the both ends of the cDNA for further amplification. The products were sequenced on an Illumina Hiseq2000 platform. We aligned the raw reads from each sample to the human reference genome (NCBI Build 37, hg19) by Tophat software. The transcriptome reconstruction was built by the mapping reads with normal and reverse and NCBI (RefSeq Genes hg19) database. We estimated the expression abundance of all genes by statistics mapping reads and used the Reads Per Kb per Million reads (RPKM) to normalize expression abundance within cells. We used the improved methods to calculate the different expression of genes and transcripts and test the statistical significance of two cell changes from NGN3-eGFP^{+} to NGN3-eGFP^{-} with a criterion of 1)fold change (FC)>2 or FC<0.5 and 2)P-value<0.05. We used the heatmap package of R to perform the heatmap based on RPKM of all genes with log10. We used all DGE（Differential Gene Expression）to map the each term of Gene Ontology Database, statistics the number

of mapping gene, and adopt hypergeometric distribution(corrected p value≤0.05,with Bonferroni correction) to find the enrichment of GO term.

Kidney capsule transplantation

6-to-8-week-old male SCID-Beige mice were purchased from Vital River Inc. Kidney capsule transplantation assay was conducted as previously described (Cell Stem Cell. 2013 Aug 1;13(2):230-6). Briefly, hESC-derived SUSD2-enriched cells and SUSD2- negatively-enriched cells were reaggregated, respectively, and cultured on 8 μm pore size filters (Millipore) in the DMEM medium supplemented with B27 for 24hours. Then, the solidified cell reaggregates were transplanted under the kidney capsule of SCID-BG mice. 19 weeks post-transplantation, the recipient mice were sacrificed and the engraftments were removed for further analysis.

References (gene targeting)

1. Kelly OG, Chan MY, Martinson LA *et al.*. Cell-surface markers for the isolation of pancreatic cell types derived from human embryonic stem cells. *Nat Biotechnol* 2011; **29**:750-756.
2. Liu P, Jenkins NA, Copeland NG. A highly efficient recombineering-based method for generating conditional knockout mutations. *Genome Res* 2003; **13**:476-484.
3. Wang J, Sarov M, Rientjes J *et al.*. An improved recombineering approach by adding RecA to lambda Red recombination. *Mol Biotechnol* 2006; **32**:43-53.
4. Wang Z, Li J, Huang H *et al.*. An integrated chip for the high-throughput synthesis of transcription activator-like effectors. *Angew Chem Int Ed Engl* 2012; **51**:8505-8508.
5. Xie R, Everett LJ, Lim HW *et al.*. Dynamic chromatin remodeling mediated by polycomb proteins orchestrates pancreatic differentiation of human embryonic stem cells. *Cell Stem Cell* 2013; **12**:224-237.